Oracle® Data Miner

User's Guide

Release 4.1

E58114-03

June 2015

Oracle Data Miner User's Guide, Release 4.1

E58114-03

Primary Author: Moitreyee Hazarika

Contributing Author: Kathy Taylor

Contributor: Mark Kelly, Margaret Taft

Contents

3 Data Miner Projects

4 Workflow

5 Data Nodes

6 Using the Oracle Data Miner GUI

7 Transforms Nodes

8 Model Nodes

9 Evaluate and Apply Nodes

10 Predictive Query Nodes

11 Text Nodes

12 Testing and Tuning Models

13 Data Mining Algorithms

Preface

This document contains an overview of Oracle Data Miner 4.1. Oracle Data Miner is packaged with Oracle SQL Developer 4.1 and later.

Audience

This document is intended for data analysts whose primary goal is to perform traditional data mining (build, test, and apply models) using data that resides in an Oracle Database. They are not programmers or database administrators.

Documentation Accessibility

For information about Oracle's commitment to accessibility, visit the Oracle Accessibility Program website at `http://www.oracle.com/pls/topic/lookup?ctx=acc&id=docacc`.

Access to Oracle Support

Oracle customers that have purchased support have access to electronic support through My Oracle Support. For information, visit `http://www.oracle.com/pls/topic/lookup?ctx=acc&id=info` or visit `http://www.oracle.com/pls/topic/lookup?ctx=acc&id=trs` if you are hearing impaired.

Related Documents

For more information, see the following documents in the Documentation Library for Oracle Database 12*c* Release 1 (12.1):

- *Oracle Data Mining Concepts*
- *Oracle Data Mining User's Guide*

If you are connected to an earlier version of the database, see the corresponding documents for that database version.

Conventions

The following text conventions are used in this document:

Convention	Meaning
boldface	Boldface type indicates graphical user interface elements associated with an action, or terms defined in text or the glossary.

Convention	Meaning
italic	Italic type indicates book titles, emphasis, or placeholder variables for which you supply particular values.
`monospace`	Monospace type indicates commands within a paragraph, URLs, code in examples, text that appears on the screen, or text that you enter.

1 Oracle Data Miner

Oracle Data Miner is an extension to Oracle SQL Developer. Oracle Data Miner is a graphical user interface to Oracle Data Mining, a feature of Oracle Database. Data analysts can use the intuitive Oracle Data Miner graphical user interface (GUI) to discover hidden patterns, relationships, and insights in their data. With Oracle Data Miner, everything occurs in an Oracle Database—in a single, secure, scalable platform for advanced business intelligence. Oracle Data Miner eliminates data movement and duplication, maintains security, and minimizes latency time from raw data to valuable information. Enterprises can use Oracle Data Miner for knowledge discovery to better compete on analytics.

Oracle Data Miner helps you perform the data preparation and model building required by the data mining process.

See Also: "About the Data Mining Process"

1.1 About the Data Mining Process

Data mining is the process of extracting useful information from masses of data by extracting patterns and trends from the data.

Data mining requires a problem definition, collection and cleansing of data, and model building. Much of the time spent in a typical data mining project is devoted to understanding and processing of data.

See Also: *Oracle Data Mining Concepts* for more information about the data mining process.

1.2 New Features and Changes in SQL Developer 4.1

Oracle Data Miner 4.1 has been enhanced with new features, along with some general enhancements:

See Also:

- *Oracle Data Miner Release Notes* for more information about the new features in SQL Developer 4.1.
- *Oracle Data Miner Release Notes* for more information about the general enhancements in Oracle Data Miner 4.1.

1.3 Overview of Oracle Data Miner

Oracle Data Miner is an extension to Oracle SQL Developer. Oracle Data Miner is a graphical user interface to Oracle Data Mining, a feature of Oracle Database. Oracle Data Miner enables users to build descriptive and predictive models to:

- Predict customer behavior
- Target best customers
- Discover customer clusters, segments, and profiles
- Identify customer retention risks
- Identify promising selling opportunities
- Detect anomalous behavior

Oracle Data Miner provides an Application Programming Interface (API) that enables programmers to build and use models.

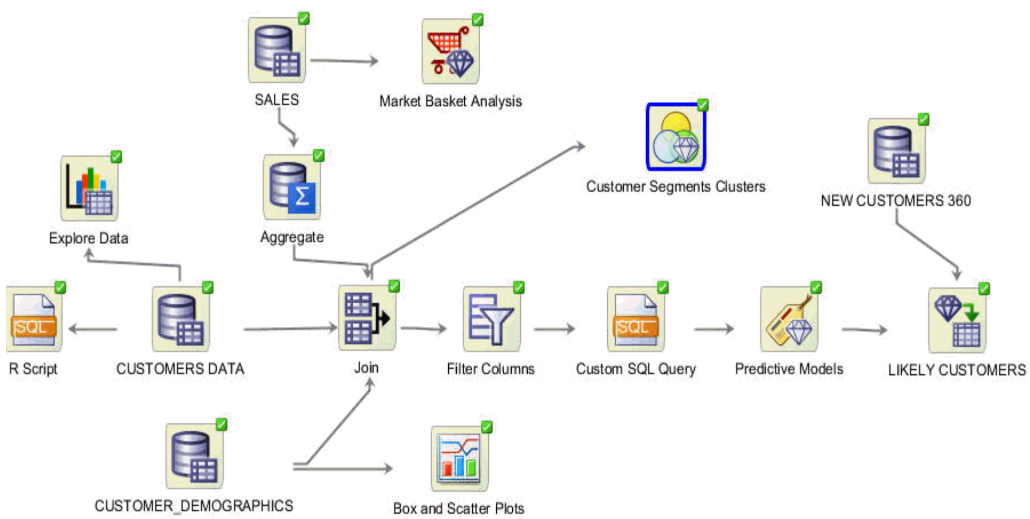
Oracle Data Miner workflows capture and document the analytical methodology of the user. It can be saved and shared with others to automate advanced analytical methodologies.

The Oracle Data Miner GUI is an extension to Oracle SQL Developer 3.0 or later that enables data analysts to:

- Work directly with data inside the database
- Explore the data graphically
- Build and evaluate multiple data mining models
- Apply Oracle Data Miner models to new data
- Deploy Oracle Data Miner predictions and insights throughout the enterprise

Figure 1–1 shows a sample workflow of Oracle Data Miner.

Figure 1–1 Sample Data Miner Workflow



Sample Oracle Data Miner workflow that identifies valuable customers.

****

Oracle Data Miner creates predictive models that application developers can integrate into applications to automate the discovery and distribution of new business intelligence—predictions, patterns, and discoveries—throughout the enterprise.

> **See Also:** "Oracle Data Miner Documentation" to access the documentation.

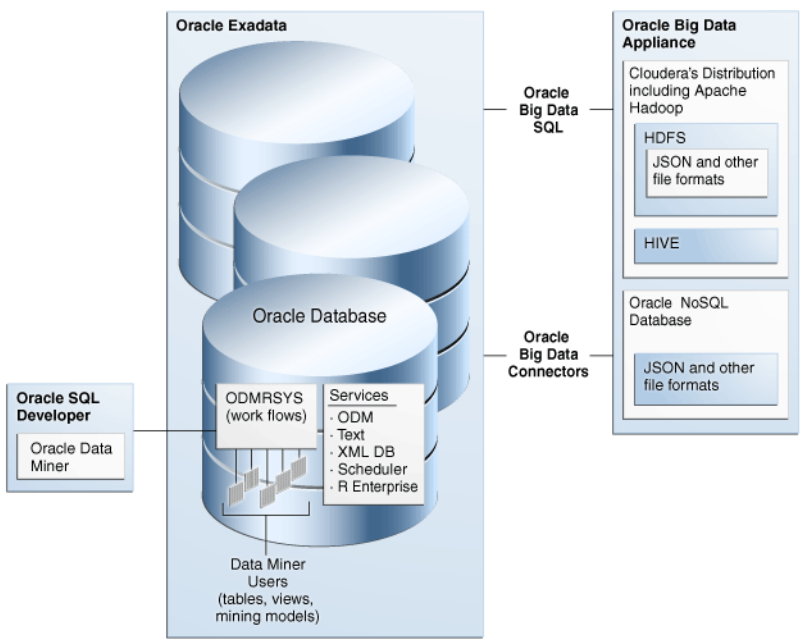
1.4 Architecture of Oracle Data Miner

Before you install Oracle Data Miner, you must understand the architecture of Oracle Data Miner. Oracle Data Miner consists of a server and one or more clients:

- Oracle Data Miner, the client, is an integrated feature of Oracle SQL Developer 3.0 or later.
- Oracle Database 12*c* Release 1 (12.1) Enterprise Edition (includes Oracle Database Personal Edition) or an earlier version of Oracle Database Enterprise Edition is the server. In addition to the database, Oracle Data Miner requires the installation of a Data Miner repository account. The repository is a separate account in the database named ODMRSYS. This repository is shared by all user accounts in the database that have been granted the appropriate privileges to use the Data Miner repository.

Figure 1–2 shows the components of Oracle Data Miner.

Figure 1–2 Oracle Data Miner Components



This figure shows SQL Developer as a client application that executes SQL and PL/SQL with the Oracle Database. Oracle Data Miner is one component of SQL Developer. The ODMRSYS repository in the Database stores metadata and APIs that support Oracle Data Miner operations and administration. The Oracle Data Miner repository supports multiple users. Oracle Data Miner uses the services of Oracle Data Mining, Oracle XML DB, Oracle Scheduler, and Oracle Text within Oracle Database Enterprise Edition.

****

Oracle Database Enterprise Edition includes these services that are critical to the support of Oracle Data Miner:

- Oracle Data Miner: A component of the Oracle Advanced Analytics option to Oracle Database Enterprise Edition. Oracle Data Miner provides model building, testing, and scoring capabilities for Oracle Data Miner.
- Oracle XML DB: Provides services to manage the Oracle Data Miner repository metadata, such as the details about the workflow specifications.
- Oracle Scheduler: Provides the engine for scheduling the Oracle Data Miner workflows.
- Oracle Text: Provides services necessary to support text mining.

1.5 Snippets in Oracle Data Miner

SQL Developer provides Snippets to help you write PL/SQL programs. Snippets are code fragments, such as SQL functions, Optimizer hints, and miscellaneous PL/SQL programming techniques. Some snippets are just syntax, and others are examples. You can insert and edit snippets when you are using the SQL Worksheet or creating or editing a PL/SQL function or procedure using SQL Worksheet or SQL Query node.

- Open SQL worksheet: To open SQL Worksheet, go to Connections for SQL Developer, right-click the connection to use, and select **Open SQL Worksheet.** See the *SQL Developer Online Help* for more information about SQL Worksheet.
- Insert snippet: To insert a snippet into your code in a SQL Worksheet or in a PL/SQL function or procedure, drag the snippet from the snippets window and drop it into the desired place in your code. Then edit the syntax so that the SQL function is valid in the current context. To see a brief description of a SQL function in a tool tip, hold the pointer over the function name. Oracle Data Miner provides snippets for the EXPLAIN, PREDICT, and PROFILE procedures in DBMS_PREDICTIVE_ANALYTICS and for the Data Mining functions for scoring data using prediction, clustering, or feature extraction.

  See Also: "Predictive Analytics Snippets" for more information about how to use snippets.

1.5.1 Predictive Analytics Snippets

To view the list of Predictive Analytics functions, and to use a snippet:

1. Open Oracle SQL Developer.
2. Select the connection that you are using for Oracle Data Miner.
3. From the SQL Developer menu, go to **View** and then select **Snippets.**
4. From the drop-down list, select **Predictive Analytics.**

   The Predictive Analytics group of snippets includes the following snippets:

   - Explain: Use `DBMS_PREDICTIVE_ANALYTICS.EXPLAIN()` to rank attributes in order of influence when explaining a target column.
   - Predict: Use `DBMS_PREDICTIVE_ANALYTICS.PREDICT()` to predict the value of a target column based on values in the input data.
   - Prediction Anomaly Function: Use the anomaly detection predictive query to predict the anomalous customers.

- Prediction Classification Function: Makes predictions using dynamic classifications.
- Prediction Cluster Function: Predicts the cluster a customer belongs to.
- Prediction Feature Set Function: Predicts feature sets to provide a general characterization of the underlying customer data.
- Prediction Regression Function: Predicts the age of customers who are likely to use an affinity card.
- Profile: Use `DBMS_PREDICTIVE_ANALYTICS.PROFILE()` to generate rules that identify the records that have the same target value.

5. To use a snippet, drag the snippet to the SQL Worksheet or to a place in a PL/SQL program.

> **Note:** The Explain, Predict, and Profile snippets have one or more commented-out **DROP** statements, such as:
>
> `--DROP TABLE mining_explain_result;`
>
> If you run one of these snippets more than once, remove the comment characters for the **DROP** statement.

If you drag the Explain snippet to SQL Worksheet, you see:

```
--Available in Oracle Enterprise DB 10.2 and later

--Ranks attributes in order of influence to explain a target column.

--For more info go to: http://www.oracle.com/pls/db112/vbook_
subject?subject=dma

--Remove comment on the Drop statement if you want to rerun this script

--DROP TABLE mining_explain_result;

--Perform EXPLAIN operation

BEGIN

    DBMS_PREDICTIVE_ANALYTICS.EXPLAIN(

         data_table_name => '"CUSTOMERS"',

         explain_column_name => '"CUST_GENDER"',

         result_table_name => 'mining_explain_result',

         data_schema_name => '"SH"');

END;

/

--output first 10 rows from resulting table mining_explain_result

COLUMN ATTRIBUTE_NAME FORMAT A30

COLUMN ATTRIBUTE_SUBNAME FORMAT A30

COLUMN EXPLANATORY_VALUE FORMAT 0D999999

COLUMN RANK FORMAT 999999

select * from mining_explain_result where rownum < 10;
```

When you run this code, you get the following results (in Script Output):

```
anonymous block completed

ATTRIBUTE_NAME      ATTRIBUTE_SUBNAME      EXPLANATORY_VALUE        RANK

----------------- ------------------------------ ----------------- ------

CUST_LAST_NAME                             0.151359                 1

CUST_MARITAL_STATUS                        0.015043                 3

CUST_INCOME_LEVEL                          0.002592                 4

CUST_CREDIT_LIMIT                          0.000195                 5

CUST_EMAIL                                 0.000000                 6

CUST_TOTAL                                 0.000000                 6

CUST_TOTAL_ID                              0.000000                 6

CUST_FIRST_NAME                            0.000000                 6

9 rows selected
```

See Also:

- *Oracle Database PL/SQL Packages and Types Reference* for more information.
- *Oracle Data Mining User's Guide* and *Oracle Database SQL Language Reference* for detailed information about these functions.
- "Snippets in Oracle Data Miner" for information about how to write code using some of these snippets.

1.6 About the Oracle Data Miner Repository Installation

The Oracle Data Miner Repository resides in the database that the Oracle Data Miner client connects to. The repository stores metadata about workflows.

The Oracle by Example tutorial Setting Up Oracle Data Miner 4.1 describes how to install the Data Miner Repository and grants rights to an account. The tutorial describes the following steps:

1. Create a Data Miner User Account using SQL Developer.
2. Create a SQL Developer connection for the Data Miner User.
3. Install the Data Miner Repository.

   By default, sample data is installed. You can deselect loading these tables and views.

   You can drop the repository through GUI, if necessary.

If you have dropped a Repository or if you do not want to use the OBE, you can reinstall the repository.

When you install the repository, rights are automatically granted to the user account that you connect with.

To grant rights to another account, define a database connection for the account and open the account in the Data Miner navigator. The GUI tells you that the account does not have the correct grants. Click **OK** for the grants to be created. You must log in using an administrative (SYS) account.

See Also:

- "About Dropping the Oracle Data Miner Repository" for information about how to drop the repository using the GUI.
- "Installing the Data Miner Repository Using the GUI" for information about how to reinstall the repository.
- `http://www.oracle.com/webfolder/technetwork/tutorials/obe/db/12c/r1/dm/ODM12c-SetUp.html`

1.6.1 Installing the Data Miner Repository Using the GUI

To install the repository:

1. Double-click a connection. If the repository is not installed, the *Repository not Installed* warning message is displayed. You also see this message if you try to create a project when the repository is not installed.

   Click **Yes** to install a Data Miner repository in the database to which you are connected.

   > **Note:** If you have several database connections, you need to install the repository using only one connection. You must grant privileges to the other connections.

2. Log in using an administrative account for the database that you are connected to.
3. The Repository Installation Settings dialog box opens.
   - Name of the repository—ODMRSYS. You cannot change this name.
   - Select a default tablespace and a temporary tablespace, and click **OK.**

     If you are connected to Oracle Database 11*g* Release 2 (11.2.0.4) or later, permanent tablespaces are filtered so that only Oracle ASM tablespaces are displayed. (The temporary tablespace does not have to be Oracle ASM.)

     If you are connected to Oracle Database 11*g* Release 2 (11.2.0.3) or earlier, no filtering takes place.
4. The Install Data Miner Repository dialog box opens. By default, demo data is installed used by the Oracle By Example tutorials. Click **Start** to begin the installation.

   > **Note:** The sample data requires the SH schema.

5. The installation may take several minutes. After the installation completes, you get a message indicating that the task completed successfully. You can examine the log of the task by clicking **Show Log.** Click **Close** when done.
6. If you have more than one Connection, you must grant privileges to each connection as follows:
   a. Click the connection.
   b. In the Required Privileges Missing dialog box, click **Yes** to grant privileges.
   c. Log in using an administrative account for the database that you connect to.

d. In the Data Miner Grants dialog box, click **Start** to grant privileges and to install sample data.

e. After the tasks completes, a message is displayed indicating that the task has completed successfully. You can examine the log of the task by clicking **Show Log.**

f. Click **Close** when done.

See Also: *Oracle Data Miner Installation and Administration Guide* for information about how to install the Data Miner Repository using scripts.

1.7 About Dropping the Oracle Data Miner Repository

If you plan to stop using Oracle Data Miner, you should drop the Repository. You may have to drop the repository when you upgrade from one version of Oracle Database to another.

When you drop the Repository, all workflows and internal tables are dropped. Models created by Oracle Data Miner are dropped. Tables created by the Create Table or View node are not dropped.

Note: **Drop Repository** cannot be undone. When you drop a repository, the repository and all internal user objects created by Data Miner are permanently removed. The objects removed include models, tables or views generated by the Create Table node, hidden tables used store results (model viewers), and text specification objects used to support text transformations. In addition all Data Miner workflows are dropped. To save a workflow, export it and later import it.

Before you drop the Repository, check that no projects or workflows are open. If any connections are open, the process may fail.

For **Drop Repository** to complete successfully, no sessions with the role of ODMRUSER can be active. Active ODMRUSER sessions can result in object locks that block dropping of the repository.

During the process of dropping the repository:

- The database does not allow any new connections to be established during this process.
- All sessions with the ODMRUSER role are automatically disconnected.
- All workflows and internal tables are dropped. Models created by Oracle Data Miner are dropped. Tables created by the Create Table or View node are dropped.

Note: After you drop the repository, you must install the repository again before you can perform any data mining.

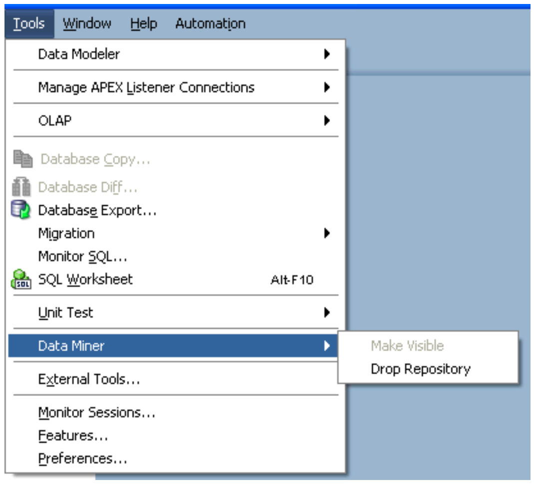
See Also:

- "Exporting a Workflow Using the GUI"
- "Importing a Workflow Using the GUI"

1.7.1 Dropping the Data Miner Repository Using GUI

There is one Oracle Data Miner repository per database. If you have several database connections, it is necessary to drop the repository for one connection only. To drop the repository:

1. In SQL Developer UI, go to **Tools** and click **Data Miner.** Then click **Drop Repository.**



2. The Drop Repository dialog box opens. Select the connection where you want to drop the repository. You can add and edit connections here, if required.
3. Log in as SYS, and click **OK.**
4. A list os sessions that will be disconnected is displayed. Click **OK** to continue.
5. The **Drop Data Miner Repository** dialog box opens. Click **Start** to begin the drop of the repository. This process may take several minutes. Specific messages indicate which steps are being performed.

   After the operation completes, examine the log files to see which operations were performed.

> **See Also:** *Oracle Data Miner Installation and Administration Guide* for information about how to drop the repository using scripts.

1.8 About Oracle Data Miner Repository Migration

After you download a patch or a new version or you update the database that you connect to, you are notified that upgrade is necessary when you open a connection used for data mining. The GUI issues a message describing the problem and asks you if it should perform necessary migration. If you answer yes, you are prompted for the administrative (SYS) password.

Migration may require conversion of the repository. If migration is required, then the Migrating the Oracle Data Miner Repository Using the GUI dialog box opens.

1.8.1 Migrating the Oracle Data Miner Repository Using the GUI

Migration of the Oracle Data Miner repository arises in the following cases:

- A version of SQL Developer before 4.0 is installed on Oracle Database 11*g* Release 2 (11.2.0.4) or later, and you want to connect using a SQL Developer 4.0 or later client, that is, you upgrade the client.

- SQL Developer 4.0 installed on a version of Oracle Database 11*g* Release 2 (11.2.0.3) or earlier, and the database is upgraded to Oracle Database 11*g* Release 2 (11.2.0.4) or later.

Either of these conditions is detected when you open the connection in the Data Miner Navigator. If the ODMRSYS default permanent tablespace is not an ASM tablespace, then a dialog is displayed requesting an ASM tablespace. The ASM tablespace does not replace the existing default permanent tablespace already specified for ODMRSYS, but instead, it converts the `workflow_data` column that stores the workflow XML data in the `ODMR$_WORKFLOWS` table. The `workflow_data` is usually the largest data component stored in ODMRSYS. This approach reduces the amount of time required to perform the migration.

The Repository Migration does not request temporary tablespace.

> **See Also:** *Oracle Data Miner Installation and Administration Guide* for information about how to migrate the repository using scripts.

1.9 How to Use Oracle Data Miner

You can learn how to use Oracle Data Miner in the following ways:

- By using Oracle By Example for Oracle Data Miner 4.0 tutorials
- By experimenting on your own using with Sample Data
- By using the Oracle Data Miner Online Help
- By asking questions on the Oracle Data Mining Forum
- By referring to Oracle Data Miner Documentation for reference information about data mining in general and Oracle Data Miner in particular.

1.9.1 Oracle By Example for Oracle Data Miner 4.0

The Oracle By Example tutorials teach you how to install and use Oracle Data Miner:

- **Setting Up Oracle Data Miner 4.0**: This tutorial covers the process of setting up Oracle Data Miner for use within Oracle SQL Developer 4.0 connected to Oracle Database 12*c*.
- **Using Oracle Data Miner 4.0:** This tutorial covers the use of Oracle Data Miner 4.0 to perform data mining on Oracle Database 12*c*. In this lesson, you examine and solve a data mining business problem by using the Oracle Data Miner graphical user interface (GUI). The Oracle Data Miner GUI is included as an extension of Oracle SQL Developer, version 4.0.
- **Using Feature Selection and Generation with GLM**: This tutorial covers the use of Oracle Data Miner 4.0 to leverage enhancements to the Oracle implementation of Generalized Liner Models (GLM) for Oracle Database 12*c*. These enhancements include support for Feature Selection and Generation.
- **Text Mining with an Expectation Maximization Clustering Model**: This tutorial covers the use of Oracle Data Miner 4.0 to leverage new text mining enhancements while applying a clustering model. In this lesson, you learn how to use the Expectation Maximization (EM) algorithm in a clustering model.
- **Using Predictive Queries With Oracle Data Miner 4.0**: This tutorial covers the use of Predictive Queries against mining data by using Oracle Data Miner 4.0.
- **Using the SQL Query Node in an Oracle Data Miner Workflow**: This tutorial covers the use of the new SQL Query Node in an Oracle Data Miner 4.0 workflow.

The tutorials are in Oracle Learning Library under the Oracle Data Miner 12*c* OBE Series at `http://apex.oracle.com/pls/apex/f?p=44785:24:::NO:24:P24_CONTENT_ID,P24_PREV_PAGE:8297,1`.

1.9.2 Sample Data

Sample Data is loaded in your account when you install Oracle Data Miner. You also have access to sample data in `SH` and other schemas. See *Oracle Data Miner Installation and Administration Guide* for information about how to add tables and views from `SH` and other schemas to you account.

1.9.3 Oracle Data Miner Online Help

The online help specific to Oracle Data Miner is in the help folder **Oracle Data Miner Concepts and Usage**.

To view or search the online help for Oracle Data Miner click **Help** and then click **Table of Content.** Then expand the Table of Content and go to **Oracle Data Miner Concepts and Usage** on the Contents tab of Help Center.

To get help for a specific dialog box, click the **Help** button or press F1. To get help for objects in a workflow, select the object and press **F1.**

Online help contains reference topics and the topics that describe how the GUI works. To see reference topics, either expand the help contents in the online help or search in the online help.

> **See Also:** "Search the Online Help"

1.9.3.1 Search the Online Help

To search the online help, use the search box at the top of Help Center:

Type the word or words that you want to search for in the search box and press **Enter**.

Select one of the following search options: case sensitive (Match case) or case insensitive; and whether to match topics based on all specified words, any specified words, or a Boolean expression.

Search performs a full-text search of all Oracle Data Miner online help topics, including Oracle Data Miner Release Notes.

To cancel a search, click ■.

1.9.4 Oracle Data Mining Forum

Participate in the Oracle Data Miner Discussion Forum `http://forums.oracle.com/community/developer/english/business_intelligence/data_warehousing/data_mining` to discuss using Oracle Data Miner and data mining.

1.9.5 Oracle Data Miner Documentation

Oracle Data Miner is the graphical user interface for Oracle Data Miner. Oracle Data Miner is a component of the Oracle Advanced Analytics option to Oracle Database Enterprise Edition.

Oracle Data Miner documentation is included in the Oracle Database Documentation Library for the version of the database that you have installed. Documentation Libraries are posted at the **Documentation** site at `http://docs.oracle.com` in the **Database** section. To go directly to the Business Intelligence and Data Warehousing documentation, use `http://www.oracle.com/pls/topic/lookup?ctx=db112&id=dwbitab` if you connect to Oracle Database 11*g* Release 2 (11.2) or `http://www.oracle.com/pls/topic/lookup?ctx=db121&id=dwbitab` if you connect to Oracle database 12*c* Release1 (12.1).

See Also:

- *Oracle Data Mining Concepts* for more information about Oracle Data Miner.
- Documentation Library for:
  - Oracle Database 12c Release 1 (12.1) at: `http://docs.oracle.com/database/121/nav/portal_6.htm`
  - Oracle Database 11g Release 2 (11.2) at: `http://docs.oracle.com/cd/E11882_01/nav/portal_6.htm`

Oracle Data Miner documentation is on the Business Intelligence and Data Warehousing page under the heading *Oracle Advanced Analytics*. See "Related Documentation" in the preface of *Oracle Data Mining Concepts* for other documentation.

2 Connections for Data Mining

Oracle Data Mining takes place in an Oracle Database. All data mining objects including input tables and models reside in a database.

This chapter contains the following topics:

- About the Data Miner Tab
- Creating a Connection

2.1 About the Data Miner Tab

In Oracle SQL Developer, the **Data Miner** tab lists the following:

- Connections

> **Note:** Connections are listed both in the **Connections** tab and in the **Data Miner** tab.

- Projects
- Workflows

2.1.1 Prerequisites for Data Mining

Before you start data mining, ensure the following:

- Specify a connection to an account in an Oracle Database 11*g* Release 2 (11.2.0.1) or Oracle Database 12*c* Release 1 (12.1). The account that you connect to must have all grants required by Oracle Data Mining.
- Define at least one connection to indicate the account where workflow nodes run.

2.1.2 Viewing the Data Miner Tab

If the **Data Miner** tab is not visible in the Oracle SQL Developer window, then:

- Go to **Tools** and click **Data Miner.** Then, select **Make Visible.**
- Go to **View** and click **Data Miner.** Then, select **Data Miner Connections.**

This docks the Data Miner pane in the Oracle SQL Developer window.

2.2 Creating a Connection

You can create connections to Oracle Database from the Connections tab and from the Data Miner tab.

See Also:

- "Creating Connections from the Connections Tab"
- "Creating Connections from the Data Miner Tab"

2.2.1 Creating Connections from the Connections Tab

To create connections from the **Connections** tab:

1. In Oracle SQL Developer, go to **View** and click **Connections.** This docks the Connections pane in the Oracle SQL Developer window.
2. In the Connections pane, right-click **Connections** and click **New Connections.** Alternately, you can click  to add a new connection.
3. In the New/Select Database Connection dialog box, enter the following for the new connection:
   - **Connection Name:** The name of the connection.
   - **Username:** The Data Miner user name. See *Oracle Data Miner Installation and Administration Guide* for more information about Data Miner users.
   - **Password:** The Data Miner user password.
   - **Save Password:** Select this option to save your password.
4. In the Oracle tab, provide the following details about the Oracle Database:
   - **Connection Type**
   - **Role**
   - **Hostname**
   - **Port**
   - **SID:** This is the Oracle Service Identifier (SID), a name by which the Oracle Database instance is uniquely identified from other database instance.
   - **Service Name:** Select this option if you are connecting to a database that is installed in a cluster such as Oracle Real Application Clusters.
5. Select one or all of the following options:
   - **OS Authentication**
   - **Kerberos Authentication**
   - **Proxy Connection**
6. To connect to the database, click **Connect.**
7. To save the connection, click **Save.**
8. To test the connection, click **Test.**

2.2.2 Creating Connections from the Data Miner Tab

To create connections from the **Data Miner** tab:

1. In Oracle SQL Developer, if the **Data Miner** tab is not visible, then you can dock it in the Oracle SQL Developer window.
2. In the **Data Miner** tab, right-click **Connections.** The Select Connection dialog appears.
3. To create a new connection, click . The New/Select Database Connection is displayed.
4. Follow step 3 through step 8, to complete the task of creating a connection.

See Also:

- "Viewing the Data Miner Tab" for more information about making the Data Miner tab visible.
- "Creating Connections from the Connections Tab" for more information about creating connections.

2.2.3 Managing a Connection

After you create a connection, you can manage it from the Data Miner Connections context menu. To perform the following tasks, right-click the connection name in the **Data Miner** tab and click:

- **Disconnect:** To disconnect the connection from the database.

> **Note:** The connection still exists. You can reconnect by double-clicking the connection, and entering the password.

- **Add Connection:** To create a new connection.
- **New Project:** To create a new project.
- **Remove:** To remove a connection.
- **Properties:** To open the **New/Select Database Connection** dialog box. You can view and edit connection properties here.

See Also:

- "Creating Connections from the Connections Tab"
- "Creating Connections from the Data Miner Tab"
- "Creating a Project" for more information about project creation.
- "Removing a Connection" for more information about connection removal and deletion.

2.2.3.1 Viewing and Editing Connection Properties

To view connection properties:

1. In the **Data Miner** tab, right-click the connection name and select **Properties.**
2. In the **New/Select Database Connection** dialog box, you can view the connection properties and edit them, as required.
3. Click **OK.**

2.2.3.2 Removing a Connection

To remove a connection:

1. In the **Data Miner** tab, right-click the connection name and select **Remove.** This removes the connections from the Data Miner tab.

> **Note:** Removing the connection from the **Data Miner** tab does not remove the connection permanently. To permanently remove the connection, delete it from **Connections** tab.

2. Go to the **Connections** tab and right-click the connection.
3. Click **Delete.** This permanently removes the connection.

> **Note:** When you delete a connection, the projects and workflows accessible through the connection are not deleted.

2.2.4 Connecting to a Database

To connect to Oracle Database:

1. In the **Data Miner** tab, right-click the connection that you want to connect and click **Add Connections.** If the **Data Miner** tab is not visible, then dock it in the Oracle SQL Developer window.
2. In the **Select Connections** dialog box, select the connection from the **Connections** drop-down list. If you are disconnected from the connection, you may have to provide a password.
3. Click **OK.**

See Also:

- "Provide Password"
- "Viewing the Data Miner Tab" for more information about adding connections to the database.

2.2.4.1 Provide Password

If you do not select the **Save Password** option when defining the connection, then you must provide the password when you connect.

After you log in, Oracle Data Miner monitors the connection. You can now perform data mining operations.

3

Data Miner Projects

A Data Miner project resides in a database connection. The project contains the workflows created for the project. You must create at least one project before you create any workflow.

This chapter contains the following topics:

- Creating a Project
- Managing a Project

3.1 Creating a Project

You must create at least one project in a connection. You can also create multiple projects in a connection. To create a project:

1. Connect to a database.
2. In the Data Miner tab, expand **Connections** and right-click the connection in which you want to create the project. Click **New Project.**

   > **Note:** If the Data Miner tab is not visible, you can dock it in Oracle SQL Developer window.

3. In the New Project dialog box, enter the following details:
   - **Project Name—**Enter a project name. Ensure that it conforms to the project naming conventions.
   - **Comment—**You can enter any comment in 4000 or fewer characters. This is an optional field.
4. Click **OK.**

After the project is created, you can rename it, edit any comments, delete it, or copy it to an other connection.

See Also:

- "Managing a Project" for more information about project related tasks.
- "Viewing the Data Miner Tab" for information about how to make the Data Miner tab visible.
- "Project Name Restrictions" for information about project naming conventions.
- "Connecting to a Database" for more information about how to connect to the Oracle Database.

3.1.1 Project Name Restrictions

A project name must meet the following conditions:

- The name must be unique within the project.
- The character count for the workflow name should be between 1 and 30. It should not exceed 30 characters.
- The name cannot contain a slash (/).

3.2 Managing a Project

In the Data Miner tab, right-click the project that you want to manage. The following operations are available to manage the project through the context menu:

- **New Workflow:** To create a new workflow.
- **New Project:** To create a new project.
- **Delete:** To delete a project.
- **Properties:** To add or modify comments for the project. You can edit the comments for an existing project from here.
- **Rename:** To rename the project.
- **Import Workflow:** To import a workflow in to this project, that was previously exported.

See Also:

- "Creating a Workflow"
- "Creating a Project"
- "Deleting a Project"
- "Project Name Restrictions"
- "Importing a Workflow Using the GUI"

3.2.1 Deleting a Project

To delete a project, right-click the project and click **Delete.** The following are deleted when you delete a project:

- All workflows contained in the project.
- All objects generated by those workflows such as models or apply results.

The database objects, such as tables referenced by workflows in the project, are not removed when a project is deleted.

Note: You **cannot** undo a project deletion.

3.2.2 Expanding a Project

To expand a project:

- Click the plus sign to the left of the project name
- Double-click the project name

The list of workflows created in the project is displayed.

4 Workflow

A Data Miner workflow is a mechanism to define data mining operations such as model build, test, and apply. These operations are defined in nodes, which comprise the workflow. You can create multiple workflows in a single project. This chapter provides the following information:

- About Workflows
- About Nodes
- Working with Workflows
- Working with Nodes
- About Parallel Processing

4.1 About Workflows

A workflow must contain one or more sources of data, such as a table or a model. For example, to build a Naive Bayes model, you must first identify the input with a Data Source node. Then you create a Classification Node to build and test the model.

Using a workflow, you can:

- Build, analyze, and test data mining process
- Identify and examine data sources
- Build and apply models
- Create predictive queries

This section contains of following topics:

- Workflow Sequence
- Workflow Terminology
- Workflow Thumbnail
- Components
- Properties

4.1.1 Workflow Sequence

A workflow is built and run in the following sequence:

1. Create a blank workflow.
2. Add nodes to the workflow.

3. Connect the nodes in a workflow.
4. Run the nodes.
5. Examine the results.
6. Repeat the steps as required.

See Also:

- "Creating a Workflow"
- "Add Nodes or Create Nodes"
- "Link Nodes"

4.1.2 Workflow Terminology

A workflow is a directed graph consisting of nodes that are connected to one another. There are specific terms used to indicate the nodes in relation to the workflow.

For example, consider a workflow composed of two nodes, N1 and N2. Assume that N1 is connected to N2:

- **Parent node** and **child node:** The node N1 is the *parent* node of N2, and the node N2 is a *child* of N1. A parent node provides information that the child node needs when it runs. You must build a model before you apply to new data.
- **Descendants** and **ancestors:** The node N2 is called the *descendant* of N1 if there is a workflow connection starting from N1 that eventually connects to N2. N1 is called the *ancestor* of N2. N1 is always closer to the root node than N2.
- **Root nodes:** The nodes that have no parent nodes are called *root* nodes. All workflows have at least one root node. A parent node can have multiple root nodes.

Note: A parent node is closer to a root node than its child node.

- **Siblings:** If a node has several child nodes, then the child nodes are referred to as *siblings*.
- **Upstream:** Parent nodes are called *upstream* of child nodes.

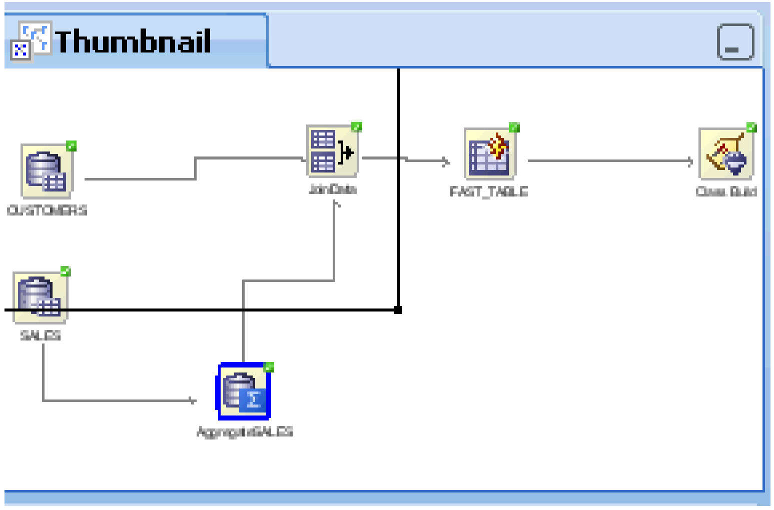
4.1.3 Workflow Thumbnail

The thumbnail view of the workflow provides an overall view of the workflow. Thumbnails are most useful for large workflows.

To open the Thumbnail viewer, go to **View** and click **Thumbnail.** Alternately, press Ctrl+Shift+T to open the Thumbnail viewer.

In a Thumbnail view, you can:

- Navigate to a specific node in the workflow
- Change the zoom level of the workflow
- Change the node that is currently viewed in the workflow



Thumbnail view of a workflow.

****

The Thumbnail view includes a rectangle. Click inside the viewer to move the rectangle around, thereby changing the focus of the display in the workflow editor. You can also resize the rectangle for a finer level of control.

When viewing models in the Thumbnail view, if the model generates trees, then the Thumbnail view automatically opens to provide an overall view to display the shape of the tree. The Thumbnail view also shows your location in the overall model view.

4.1.4 Components

The Components pane lists the components that you can add to workflows. To add a component to a workflow, drag and drop the component from the Components pane to the workflow.

The Components pane opens automatically when you load a workflow. If the Components pane is not visible, then go to **View** and click **Components.**

The Components pane includes the following:

- My Components
- Workflow Editor
- All Pages

4.1.4.1 My Components

My Components has two tabs:

- **Favorites:** To add a component to the Favorites list, right-click the component and select **Add to Favorites.**
- **Recently Used:** Lists recently used components.

4.1.4.2 Workflow Editor

The Workflow Editor contains nodes that are categorized into sections. You can create and connect the following nodes in a workflow:

- Data Nodes

- Transforms Nodes
- Model Nodes
- Evaluate and Apply Nodes
- Predictive Query Nodes
- Text Nodes
- Link Nodes

4.1.4.3 All Pages

All Pages lists all available components in one list.

4.1.5 Workflow Properties

The Workflow Properties tab enables you to add or change comments associated with the selected workflow.

> **See Also:** "Properties" for an overview of the Properties functionality.

4.1.6 Properties

Properties enables you to view and change information about the entire workflow or a node in a workflow.

> **Note:** In the earlier releases of Oracle Data Miner, the Properties tab was known as the Property Inspector.

To view the properties of a node, right-click the node and select **Go to Properties.**

If you select either the workflow or one of its node, the respective properties are displayed in the Properties pane. For example, if you select a Data Source node, the Data Source node properties are displayed in the Properties pane.

You can perform multiple tasks for a node and a workflow from:

- **Properties context menu:**
- **Properties pane:**

> **See Also:**
>
> - "Performing Tasks from Workflow Context Menu"
> - "Managing Workflows and Nodes in the Properties Pane"

4.2 Working with Workflows

You can perform the following tasks with a workflow. However, you must ensure that the Workflow Prerequisites are met.

- Creating a Workflow
- Deleting a Workflow
- Renaming a Workflow
- Managing a Workflow

- Exporting a Workflow Using the GUI
- Importing a Workflow Using the GUI
- Deploying Workflows

4.2.1 Workflow Prerequisites

Before you can perform any task with a workflow, ensure that the following workflow prerequisites are met:

- Create and establish a connection to the database.
- Create a Project under which the workflow is created.

See Also:

- "Creating a Connection"
- "Creating a Project"

4.2.2 Creating a Workflow

Before you create a workflow, ensure that you meet the workflow prerequisite conditions.

To create a blank workflow:

1. In the Data Miner tab, expand **Connections** in the left pane.

   If the Data Miner tab is not open, then click **Tools** and select **Data Miner.**
2. Select the project under which you want to create the workflow.
3. Right-click the project and select **New Workflow** from the context menu. The Create Workflow dialog box opens.
4. In the **Name** field, enter a unique name for the workflow.
5. Click **OK.** This creates a blank workflow.

See Also::

- "Workflow Prerequisites"
- "Workflow Name Restrictions"

4.2.2.1 Workflow Name Restrictions

A workflow name must meet the following conditions:

- The character count for the workflow name should be between 1 and 30. It should not exceed 30 characters.
- The workflow name must not have a slash (/).
- The workflow name must be unique within the project.

4.2.3 Deleting a Workflow

To delete a workflow:

1. In the Data Miner tab, expand **Projects** and select the workflow that you want to delete.
2. Right-click and click **Delete.**

4.2.4 Renaming a Workflow

To change the name of a workflow or project:

1. Right-click the name of the workflow or project in the **Data Miner** tab and select **Rename.**
2. The **Rename Workflow** dialog box or **Rename Project** dialog box opens. Enter the new name in the **Rename To** field.

   For a project, the new name must satisfy the project name restrictions. For a workflow the new name must satisfy the workflow name restrictions.
3. Click **OK.**

See Also:

- "Project Name Restrictions"
- "Workflow Name Restrictions"

4.2.5 Loading a Workflow

After creating a workflow, you can perform the following tasks:

- Load a workflow: In the **Data Miner** tab, expand the project and double-click the workflow name. The workflow opens in a new tab.
- Close a workflow: Close the tab in which the workflow is loaded and displayed.
- Save a workflow: Go to **File** and click **Save.** If a workflow has any changes, they are saved when you exit.

4.2.6 Managing a Workflow

After you create a workflow, the workflow is listed under **Projects** in the **Data Miner** tab.

To perform any one of the following tasks, right-click the workflow under **Projects** and select an option from the context menu:

- **New Workflow:** To create a new workflow in the current project.
- **Delete:** To delete the workflow from the project.
- **Rename:** To rename the workflow.
- **Export:** To export a workflow.
- **Import:** To import a workflow.

See Also:

- "Creating a Workflow"
- "Exporting a Workflow Using the GUI"

4.2.6.1 Exporting a Workflow Using the GUI

To export a workflow:

1. In the **Data Miner** tab, expand **Connection** and click the project under which the workflow is created.
2. Right-click the workflow that you want to export and click **Export.** The **Save** dialog box opens.

3. In the **Save** dialog box, navigate to the location where you want to export the workflow and click **Save.** The workflow is saved as an XML file.

> **Note:** The default directory to save the workflow is the system default directory for saving documents. You can change this directory.

4.2.6.2 Importing a Workflow Using the GUI

You can import workflows into a project. Before you import a workflow, ensure that the import requirements of a workflow are met.

To import a workflow:

1. In the **Data Miner** tab, right-click the project where you want to import the workflow and click **Import Workflow.** The **Open** dialog box opens.
2. In the **Open** dialog box, navigate to the location where the workflow XML file is.
3. Click **Open.** The system verifies that the specified file contains a workflow and determines the version number of the workflow. The **Import Workflow** dialog box opens. You can select and import only one workflow at a time.
4. In the **Import Workflow** dialog box, specify a name for the new workflow and select any one of the following options to handle naming conflicts:
   - **Rename Model and Output Table name if necessary**
   - **Keep Existing Model and Output Table name even if conflict exists**

   See Also: "Import Requirements of a Workflow"

4.2.6.2.1 Import Workflow

The **Import Workflow** dialog box enables you to specify the name of the workflow that is imported and how to handle naming conflicts. The default workflow name is the file name of the workflow. It may be necessary to:

- Change the name slightly to avoid conflicts.
- Edit the name so that it is a valid workflow name. The name of the new workflow must be unique in the new connection.

You must also select how to handle naming conflicts in names of workflow nodes and output tables. The default is **Rename Model and Output Table names if necessary**.

You can select **Keep Existing Model and Output table names even if conflict exists.** If you select this option, you must resolve any conflicts yourself.

To start the import, click **OK.**

4.2.6.3 Import Requirements of a Workflow

The import requirements of a workflow are:

- All the tables and views used as data sources in the exported workflow must be in the new account.
- The tables or views must have the same name in the new account as they did in the old account.
- It may be necessary to redefine the Data Source node.

- If the workflow includes Model nodes, then the account where the workflow is imported must contain all models that are used. The Model node may have to be redefined in the same way that Data Source nodes are.
- You must have permission to run the workflow.
- The workflow must satisfy the compatibility requirements.

See Also:

- "Data Table Names"
- "Workflow Compatibility"

4.2.6.4 Data Table Names

The account from which the workflow is exported is encoded in the exported workflow.

Example:

Assume that a workflow is exported from the account `DMUSER` and contains the Data Source node with data `MINING_DATA_BUILD`.

If you import the schema into a different account, that is, an account that is not `DMUSER`, and try to run the workflow, then the Data Source node fails because the workflow looks for `DMUSER.MINING_DATA_BUILD_V`.

To resolve this issue:

1. Right-click the Data Source node `MINING_DATA_BUILD_V` and select **Define Data Wizard.**

   A message appears indicating that `DMUSER.MINING_DATA_BUILD_V` does not exist in the available tables or views.

2. Click **OK** and select `MINING_DATA_BUILD_V` in the current account. This resolves the issue.

4.2.6.5 Workflow Compatibility

The compatibility requirements for importing and exporting workflows are:

- Workflows exported using a version of Oracle Data Miner earlier than Oracle Database 11*g* Release 2 (11.2.0.1.2) cannot always be imported into a later version of Oracle Data Miner.
- Workflows exported from an earlier version of Oracle Data Miner may contain invalid XML. If a workflow XML file contains invalid XML, then the import is terminated.

To check the Oracle Data Miner version:

1. Go to **Help** and click **Version.**
2. Click the **Extensions** tab.
3. Select **Data Miner.**

4.2.6.6 Building and Modifying Workflows

The Workflow Editor is the tool to modify workflows. The Workflow Editor is supported by:

- Properties

- Components
- Workflow Thumbnail

4.2.6.7 Missing Tables or Views

The schema that is imported into a workflow may not have all the tables or views used by the workflow. When Data Miner detects this problem, it generates the `Table Selection Failure` message indicating that the table is missing. You have the choice to select another table.

To select another table, click **Yes.** The Define Data Source Wizard opens. Use the wizard to select a table and attributes.

4.2.6.8 Managing Workflows using Workflow Controls

Several controls are available as icon in the top border of a workflow, just below the name of the workflow. You can perform the following tasks:

- Zoom in and zoom out workflow nodes: Click the and icon respectively.
- Control node size: Click the percent drop-down list and select the size. Default size is 100%.
- View event log: Click the icon.
- Run selected nodes: Click the icon. When you select the node to be run, the triangle turns green.
- Cancel a running workflow: Click the icon.

4.2.6.9 Managing Workflows and Nodes in the Properties Pane

In the **Properties** pane, you can view and change information about the entire workflow or a node in a workflow. To view the properties for a node:

1. Right-click the node and select **Go to Properties.** The corresponding Properties pane opens. For example, if you select a Data Source Node, in the Properties pane, the details of the Data Source node is displayed.
2. Use the **Search** field to find items in Properties.
3. Click the icon to open the editor for the item that you are viewing.
4. Use the options in the context menu to perform additional tasks.

See Also: "Performing Tasks from Workflow Context Menu"

4.2.6.10 Performing Tasks from Workflow Context Menu

To view a workflow, double-click the name of the workflow in the Data Miner tab. The workflow opens in a tab located between the Data Miner tab and the Components pane. If you open several workflows, each workflow opens in a different tab. Only one workflow is active at a time.

To display the workflow context menu, right-click a workflow tab. The following options are available in the context menu:

- **Close:** Closes the selected tab.
- **Close All:** Closes all workflow tabs.
- **Close Other:** Closes all tabs except the current one.

- **Maximize:** Maximizes the workflow. The Data Miner tab, Components pane, and other items are not visible. To return to previous size, click the selection again.
- **Minimize:** Minimizes the Properties tab to a menu. To return to the previous size, right-click the Properties tab, and click **Dock.**
- **Split Vertically:** Splits the active editor or viewer into two documents. The two documents are split vertically. Click the selection again to undo the split.
- **Split Horizontally:** Splits the active editor or viewer into two documents. The two documents are split horizontally. Click the selection again to undo the split.
- **New Document Tab Group:** Adds the currently active editor or viewer into its own tab group. Use Collapse Editor Tab Groups to undo this operation.
- **Collapse Document Tab Groups:** Collapses all the editors or viewers into one tab group. Only displayed after New Editor Tab Group.
- **Float:** Converts the Properties tab into a movable pane.
- **Dock:** Sets the location for floating window. Alternately, you can press Alt+Shift+D.
- **Clone:** Creates a new instance of Properties.
- **Freeze Content:** Freezes the content. To unfreeze Properties, click again on this selection.

4.2.7 Running a Workflow

The View Event Log enables you to view the progress of running of a workflow. You can also view the time taken for workflow creation.

Oracle Data Miner uses Oracle Scheduler to schedule running of workflows. For an overview of Oracle Scheduler, see the *Oracle Database Administrator's Guide*

Model builds can be very resource-intensive. The `MAX_NUM_THREADS` parameter in the Oracle Data Miner server controls the number of parallel builds. `MAX_NUM_THREADS` specifies the maximum number of model builds across all workflows running in the server.

Default Value=`10`. Therefore, by default, 10 models can occur concurrently across all workflows. There is the `MAX_NUM_THREADS` parameter in the repository table ODMRSYS. ODMR$REPOSITORY_PROPERTIES, where you can specify the value.

If you increase the value of `MAX_NUM_THREADS`, do it gradually. Workflows can appear to be running even though the network connection is lost.

To control the parallel model build behavior, the following parameters are used:

- `THREAD_WAIT_TIME`. The default is 5. When `MAX_NUM_THREADS` is reached, further Build process will be put on queue until parallel model build count < `MAX_NUM_THREADS`. This setting (in seconds) determines how often to check for parallel model build count.
- `MAX_THREAD_WAIT`. The default is NULL. The timeout (in seconds) for Build process that has been put on queue. If NULL, no timeout will occur.

See Also:

- "View an Event Log"
- "Network Connection Interruption"

4.2.7.1 Network Connection Interruption

If a network connection to the Data Miner server is interrupted while running a workflow, the Workflow Editor may indicate that the workflow is still running indefinitely. On the other hand, the Workflow Jobs window may indicate that the connection is down.

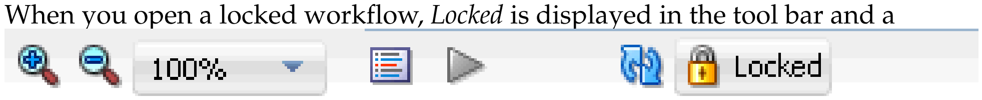
Data Miner issues a message indicating that the connection was lost. If the connection is not recovered in a reasonable period, close and open the Workflow Editor again.

Note: Use the Workflow Jobs window to monitor connections.

4.2.7.2 Locking and Unlocking Workflows

A workflow is locked under the following conditions:

- When a node is running, the workflow is locked, and none of its nodes can be edited. Existing results may be viewed while the workflow is running. You can also go to a different workflow and edit or run it.
- When a user opens a workflow, the workflow is locked so that other users cannot modify it.
- When a workflow is running, an animation in the tool bar (circular arrow loop) shows that the workflow is running.
- When you open a locked workflow, *Locked* is displayed in the tool bar and a

- If no one else has the workflow locked, and you are given the lock, then lock icon is removed from the tool bar.
- **Unlocking a locked workflow:** If a workflow seems to complete, but is still locked, click **Locked** to unlock the workflow.
- **Refreshing Workflow:** Once a locked workflow has stopped running, you can refresh the workflow by clicking the  icon. You can also try to obtain the lock yourself by clicking on the lock.

4.2.8 Deploying Workflows

Data Miner 4.1 enables you to generate a script from a workflow that recreates all the objects generated by that workflow. In earlier releases of Data Miner, you could only generate a script for Transformation nodes.

A script that recreates all objects generated by that workflow, also enables you to replicate the behavior of the entire workflow. Such scripts provide a basis for application integration or a lightweight deployment of the workflow, that does not require Data Miner repository installation and workflows in the target and production system.

Data Miner 4.1 provides the following two types of deployment:

- Deploying Workflows using Data Query Scripts
- Deploying Workflows using Object Generation Scripts

4.2.8.1 Deploying Workflows using Data Query Scripts

Any node that generates data has the **Save SQL** option in its context menu.

The **Save SQL** option generates a SQL script. The generated SQL can be created using the SQL*Plus script or as a standard SQL script. The advantage of the script format is the ability to override table and model references with parameters.

The **Save SQL** option enables you to select the kind of script to generate, and the location to save the script.

See Also: Save SQL

4.2.8.2 Deploying Workflows using Object Generation Scripts

The Object Generation Script generates scripts that create objects. The scripts that generate objects are:

- Scripts Generated
- Script Variable Definitions

4.2.8.3 Running Workflow Scripts

You can run the generated scripts in the following ways:

- As Oracle Enterprise Manager Jobs, either as SQL Script or an operating system command.
- As Oracle Scheduler Jobs:
  - External Jobs that calls the SQL Script
  - Environment: Using Oracle Enterprise Manager for runtime management or using PL/SQL
- Run the scripts using SQL*PLus or SQL Worksheet. To run a script, ensure that Oracle Data Miner repository is installed.

See Also:

- "Runtime Requirements"
- "Running Scripts using SQL*Plus or SQL Worksheet"
- "Oracle Enterprise Manager Jobs"

4.2.9 Workflow Script Requirements

Workflow script requirements include the following specific requirements:

- Script File Character Set Requirements
- Script Variable Definitions
- Scripts Generated

4.2.9.1 Script File Character Set Requirements

Scripts can contain characters based on character sets that will not be handled well unless the script file is generated using UTF8 character set.

Note: Ensure that all script files are generated using UTF8 character set.

4.2.9.2 Script Variable Definitions

The scripts have variable definitions that provide object names for the public objects created by the scripts. The Master Script is responsible for calling all underlying scripts in order. So, the variable definitions must be defined in the Master Script.

The variable `Object Types` enables you to change the name of the object names that are input to the scripts, such as tables or views, and models. By default, these names are the original table or view, and model names.

All generated scripts should be put under the same directory.

4.2.9.3 Scripts Generated

In this kind of deployment, several general scripts are generated:

- **Master Script:** Starts all the required scripts in the appropriate order. The script performs the following:
  - Validates if the version of the script is compatible with the version of the Data Miner repository that is installed.
  - Creates a workflow master table that contains entries for all the underlying objects created by the workflow script.
  - Contains generated documentation covering key usage information necessary to understand operation of scripts.
- **Cleanup Script:** Drops all objects created by the workflow script. The Cleanup Script drops the following objects:
  - Hidden objects, such as the table generated for Explore Data.
  - Public objects, such as Model Names created by Build Nodes.
  - Tables created by a Create Table Node.
- **Workflow Diagram Image:** Creates an image (.png file) of the workflow at the time of script generation. The entire workflow is displayed.

The other scripts that are generated depend on the nodes in the chain. Table 4–1 lists the nodes and their corresponding script functionality.

Table 4–1 Nodes and Script Functionality

Node	Script Functionality
▪ Data Source node ▪ Transform node ▪ Aggregate node ▪ Join node ▪ Filter Rows node ▪ Sample node ▪ Model Details Node ▪ Apply node ▪ Apply Text node ▪ Filter Columns Details node	Creates a view reflecting the output of the node.

Table 4–1 (Cont.) Nodes and Script Functionality

Node	Script Functionality
Filter Column node	Creates a view reflecting the output of the Filter Column node such as other Transform type nodes. If the `Attribute Importance` setting is specified, a table is generated containing the AI result (private).
Build Text node	For each text transformation, the following objects are created: ■ Feature Table Name ■ Oracle Text Policy Object Creates a view reflecting the output of the Build Text node. This is essentially the same output as an Apply Text node
Classification Build node	A model is created for each model build specification. A master test result table is generated to store the list of test result tables generated per model. GLM Model Row Diagnostics Table is created if row diagnostics is turned on. Each Model Test has one table generated for each of the following test results: Performance, Performance Matrix, ROC (for binary classification only). Each Model Test has one table for each of the following tests per target value (up to 100 maximum target values): List and Profit.
Regression Build node	A model is created for each model build specification. GLM Model Row Diagnostics Table is created if row diagnostics is turned on. A master test result table is generated to store the list of test result tables generated per model. Each Model Test will have one table generated for each of the following test results: Performance and Residual.
■ Clustering Build ■ Anomaly Detection Build ■ Feature Extraction Build ■ Association Build	A model is created for each model build specification.
Test Node (Classification)	A master test result table is generated to store the list of test result tables generated per model. GLM Model Row Diagnostics Table is created if row diagnostics is turned on. Each Model Test has one table generated for each of these test results: Performance, Performance Matrix, ROC (for binary classification only). Each Model Test has one table for each of Lift and Profit per target value up to 100 maximum target values.
Test Node (Regression)	GLM Model Row Diagnostics Table is created if row diagnostics is turned on. A master test result table is generated to store the list of test result tables generated per model. Each Model Test has one table generated for each of Performance and Residual.

Table 4–1 (Cont.) Nodes and Script Functionality

Node	Script Functionality
▪ Model Node ▪ Text Reference Node	No scripts are generated. These nodes are just reference nodes to metadata.

4.2.9.4 Running Scripts using SQL*Plus or SQL Worksheet

If you run generated scripts using either SQL*Plus or SQL Worksheet and require input from users, then use the following command before the generated SQL:

```
set define off
```

You can either run this new line separately before running the generated SQL. You can also generate the new line along with the generated SQL.

4.2.10 Oracle Enterprise Manager Jobs

Oracle Enterprise Manager (OEM) allows Database Administrators to define Jobs through the OEM application. The Job is run through OEM instead of Oracle Scheduler. The Job can be run manually from OEM. The running of the job can be monitored. The result is either a success or a reported failure.

- The Job definitions can directly open the generated scripts files.
- The Job definition should define the master script invocation as a script file using a full file path.
- The Job can be run on a schedule or on demand.
- All script that run within the master script must have fully qualified path names.

For information about Enterprise Manager, see the Oracle Enterprise Manager documentation set, *Oracle Database 2 Day DBA,* or any of the *Oracle Database 2 Day +* books for the database to which you are connected.

4.2.11 Runtime Requirements

The generated scripts require access to Data Miner repository objects. The Data Miner repository must be installed on the system where the scripts are run. The script checks the repository version and ensures that the repository is the same version or greater than the version of the source system.

4.3 About Nodes

A workflow consists of one or more nodes that are connected by a link. Nodes define the data mining operations in a workflow. The node categories, listed in Table 4–2, are available in the Components pane.

4.3.1 Node Name and Node Comments

Every node must have a name and may have a comment.

Name assignment is fully validated to assure uniqueness.

Oracle Data Miner generates a default node name. When a new node of particular type is created, its default name is based on the node type, such as *Class Build* for a Classification node. If a node with that name already exists, then the name is appended with 1 to make it unique. So if *Class Build* exists, then the node is named *Class Build 1.* If *Class Build 1* exists, then 2 is appended so that the third Classification node is named *Class Build 2*, and so on. Each node type is handled independently. For example, Filter Columns node have its own sequence, different from the sequence of Classification nodes.

You can change the default name to any name that satisfies the following requirements:

- Node names must be unique within the workflow.
- Node names must not contain any / (slash) characters.
- Node names must be at least one character long. They cannot be 31 or more characters long.

To change a node name, either change it in the Details tab of Properties or select the name in the workflow and type the new name.

Comments for the node are optional. If you provide any comments, it must be not more than 4000 characters long.

4.3.2 Node Types

Table 4–2 lists the different categories of nodes:

Table 4–2 Types of Nodes

Type	Description
Model Nodes	They specify models to build or models to add to a workflow.
Evaluate and Apply Nodes	They evaluate and apply models.
Data Nodes	They specify data for mining operation, data transformation or to save data to a table.
Transforms Nodes	They perform one or more transformation on the table or tables identified in a Data node.
Predictive Query Nodes	They create predictive results without the need to build models. The predictive queries automatically generate refined predictions based on data partitions.
Text Nodes	They prepare data sources that contain one or more text columns so that the data can be used in Build and Apply models.
Link Nodes	They provide a way to link or connect nodes.

4.3.3 Node States

A node is always associated with a state which indicates its status. Table 4–3 lists the states of a node.

Table 4–3 Node States

Node States	Description	Graphical Indicator
`Invalid`	Indicates that the node has not been completely defined and is not ready for execution. Most nodes must be connected to be valid. That is, they must have the input defined. A Data node does not need to be connected to be valid.	
`Error`	Indicates that the node attempted to run but encountered an error. For nodes that perform several tasks such as build several models, any single failure sets the status of the node to `Error`. You must correct all problems to clear the `Error` state. Any change clears the `Error` state to `Ready`, if the problem is a server runtime failure not attributable to standard specification validations. To find out if the problem is really fixed, run the node.	
`Ready`	Indicates that the node is properly defined and is ready for execution. Nodes in `Ready` state are also known as `Valid`.	No graphical indicator
`Complete`	Indicates that the node execution is successfully completed.	
`Warning`	Indicates that the node execution is complete but not with the expected result.	

4.4 Working with Nodes

You can perform the following tasks with any nodes:

- Add Nodes or Create Nodes
- Link Nodes
- Copy Nodes
- Edit Nodes
- Validate Parents
- Run Nodes
- Refresh Nodes
- Performing Tasks from the Node Context Menu

4.4.1 Add Nodes or Create Nodes

You create nodes or add nodes to the workflow. For the node to be ready to run, you may have to specify information such as a table or view for a Data Source node or the target for a Classification node. To specify information, you Edit Nodes. You must connect nodes for the workflow to run. For example, you specify input for a Build node by connecting a Data node to a Build node.

To add nodes to a workflow:

1. Load the workflow.
2. In the **Components** pane, go to the **Workflow Editor** and expand the section for node. You can create and connect the following types of nodes in the workflow:
   - Model Nodes
   - Data Nodes

- Transforms Nodes
- Evaluate and Apply Nodes
- Predictive Query Nodes
- Text Nodes
- Link Nodes

4.4.2 Edit Nodes

You can edit a node by:

- Using the Editor dialog box.
- Using Properties.

See Also:

- "Editing Nodes through Editor Dialog Box"
- "Editing Nodes through Properties"

4.4.2.1 Editing Nodes through Editor Dialog Box

To display the Editor dialog box for a node:

1. Double-click the node or right-click the node and select **Edit.**
2. The Editor dialog box opens. Make the edits to the node as applicable, and click **OK.**

For some nodes, such as Data Source node, an Editor dialog box is automatically displayed either when the node is dropped on the workflow or when an input node is connected to a node.

4.4.2.2 Editing Nodes through Properties

The Properties pane of a node is the pane that opens when a node in a workflow is selected. You can dock this pane.

To edit a node through Properties pane:

1. Click the node that you want to edit. The Properties pane for the node is displayed in the lower right pane.
2. Click the  icon to edit the node. For all other edits, click the respective sections in the Properties pane.

   All nodes have one common section named **Details** in the Properties pane. This section contains the node name and comment. The other section of Properties depend on the type of node.
3. The corresponding Edit dialog box opens. Make the edits to the node as applicable, and click **OK.**

4.4.3 Copy Nodes

You can copy one or mode nodes and paste them into the same workflow or into a different workflow. Copy and paste does not carry with it any mining model or results that belong to the original nodes. Since model names must be unique, unique names are assigned to models, using original name as a starting point.

For example, when a copied Build node is pasted into a workflow, the paste operation creates a new Build node with settings identical to those of the original node. However, models or test results do not exist until the node runs.

> **See Also:** "Copy"

4.4.4 Validate Parents

See "Validate Parents" for more information.

4.4.5 Run Nodes

You perform the tasks specified in the workflow by running one or more nodes. If a node depends on outputs of one or more parent nodes, the parent node runs automatically only if the outputs required by the running node are missing. You can also run one or more nodes by selecting the nodes and then clicking in the toolbar of the workflow.

> **Note:** Nodes cannot be run always. If any ancestor node is in the `Invalid` state and its outputs are missing, the child nodes that depend on it cannot run.

Each node in a workflow has a state that indicates its status, as described in Node States.

You can run a node by clicking the following options in the context menu:

- Run: Runs any tasks that have not already run. For example, if you add a new model to a Build node, Run builds the new model and does not rerun any models that are complete.
- Force Run: Reruns the node. In a Model Build node, all models are built, even the ones that are complete.

4.4.6 Refresh Nodes

Nodes such as Data Source node, Update Table node, and Model node rely on database resources for their definition. It may be necessary to refresh a node definition if the database resources change.

To refresh a node:

1. Click the node that you want to refresh.
2. In the right pane, go to **Properties**, and click the applicable section:
   - **Data:** For Data Source node
   - **Model:** For a Model node
   - **Columns:** For an Update Table node
3. Click the icon. Data Miner connects to the database and validates the attributes or models. The status of the attribute or model is set to either:
   - `Valid:` Indicated by the node without any mark on its top right corner.
   - `Invalid:` Indicated by the mark or mark on the top right corner of the node.

Other possible validation error scenarios include:

- **Source table not present:** Error message states that the source table or attribute is missing. You have the option to:
  – Select another table or replace a missing attribute.
  – Leave the node in its prior state by clicking **Cancel.**
- **Model is invalid:** Click the Model refresh button on the Properties Model tool bar to make the node valid. For Model nodes, the node becomes `Invalid` if a model is found to be invalid. The Model node cannot be run until the node is made `Valid`.
- **Model is missing:** If the Model node is run and the server determines that the model is missing, the node is set to `Error`. You can rerun the Model after the missing Model is replaced.

4.4.7 Link Nodes

A workflow is a directional graph. That is, it consists of nodes that are connected in order. Each link connects a source node to a target node. To create a workflow, you connect or link nodes. When you connect two nodes, you indicate a relationship between the nodes. For example, to specify the data for a model build, link a Data Source node to a Model Build node.

You can link nodes using the following options:

- Connect option in content menu
- Link Nodes in Components Pane
- Node Connection Dialog Box
- Connect Option in Diagram Menu

4.4.7.1 Link Nodes in Components Pane

To connect two nodes using the **Linking Nodes** option:

1. In the Components pane, expand the **Linking Nodes** section.
2. Click **Link.** Move the cursor to the source node and click. From the source node, drag the link that appears to the target node, and click again.
   - If the link is valid, then clicking the target node creates the link.
   - If the link is invalid, then the line does not terminate. To end the link process without completing a link, either press ESC or click in the Components pane.

4.4.7.2 Node Connection Dialog Box

To link two nodes using the Node Connection dialog box:

1. In the Components pane, expand the **Linking Nodes** section.
2. Click the **Link** icon and press ENTER. This opens the Select Source And Destination For A New Link dialog box.
3. Select the source node in the **Source List** and the target node in the **Destination List.**

   The **Source List** lists all the nodes in the workflow. The **Destination List** lists the allowed target nodes for the selected node.
4. Click **OK.**

4.4.7.3 Connect Option in Diagram Menu

To use the Connect option in Diagram menu:

1. Select the source node in the workflow.
2. In Data Miner menu bar, click **Diagram** and select **Connect.**
3. Move the cursor from the source node to the target node and click again.
   - If the link is valid, then clicking the target node creates the link.
   - If the link is invalid, then the line does not terminate. To end the link process without completing a link, press ESC.

4.4.7.4 Deleting a Link

To delete an existing link, select the link and press the DELETE key.

4.4.7.5 Cancelling a Link

To cancel a Link while you are linking it, press the ESC key or select another item in the Components pane.

4.4.8 Node Position

You can change the position of workflow nodes in these ways:

- Drag: To drag a node, click the node. Without releasing the mouse button, drag the node to the desired location. Links to the node are automatically repositioned.
- Adjust with arrow keys: To adjust the position by small increments, select the node. Then press and hold Shift + CONTROL keys. Use the arrow keys to move the node.

4.4.9 Performing Tasks from the Node Context Menu

The options in the context menu depend on the type of the node. Right-click a node to display the context list for the node. The context menu includes the following:

- Connect
- Edit
- Validate Parents
- Run
- Force Run
- View Data. Applicable for Data nodes. See "Data Source Node Viewer" for more information.
- Save SQL
- Deploy
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste

- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- View Models
- View Test Results
- View Test Results
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed only if there is a validation error.
- Go to Properties
- Navigate

4.4.9.1 Connect

The Connect option enables you to link nodes without using the options in Components pane.

To connect nodes:

1. Right-click a node and click **Connect.** Alternately, go to **Diagram** and click **Connect.**
2. Use the cursor to draw a line from this node to the target node.
3. Click to establish the connection. Note the following:
   - You can create only valid connections.
   - You can create only one connection between any two nodes.
   - You can remove a connection by pressing the ESC key.

   See Also: "Connect"

4.4.9.2 Edit

Nodes have default algorithms and settings. When you edit a node, the default algorithms and settings are modified. You can edit a node in any one of the following ways:

- Editing Nodes through Editor Dialog Box
- Editing Nodes through Properties

4.4.9.3 Validate Parents

Data Miner validates all parent nodes of the current node. To validate parent nodes of a node, right-click the node and select **Validate Parents.**

You can validate parent nodes when the node is in `Ready, Complete` and `Error` state. All parent nodes must be in completed state.

4.4.9.4 Run

The Data Miner server runs workflows asynchronously. The client does not have to be connected. The tasks specified in the workflow are performed by running one or more nodes in the workflow:

- To run one node: Right-click the node and select **Run.**

- To run multiple nodes simultaneously: Select the nodes by holding down the Ctrl key and click each individual node. Then right-click any selected node and select **Run.**

If a node depends on outputs of one or more parent nodes, the parent node runs automatically only if the outputs required by the running node are missing.

4.4.9.5 Force Run

The **Force Run** option reruns one or more nodes that are complete. **Force Run** deletes any existing models before building them once again.

To select more than one node, click the nodes while holding down the `Ctrl` key.

You can **Force Run** a node at any location in a workflow. Depending on the location of the node in the workflow, you have the following choices for running the node using **Force Run:**

- **Selected Node**
- **Selected Node and Children** (available if the node has child nodes)
- **Child Node Only** (available if the node one or more child nodes)
- **Selected Node and Parents** (available if the node has parent nodes)

4.4.9.6 Deploy

You can deploy a node or workflow by creating SQL scripts that perform the tasks specified in the workflow.The scripts generated by **Deploy** are saved to a directory.

Note: You must run a node before deploying it.

You can generate a script that replicates the behavior of the entire workflow. Such a script can serve as the basis for application integration or as a light-weight deployment than the alternative of installing the Data Miner repository and workflows in the target and production system.

To deploy a workflow or part of a workflow:

1. Right-click a node and select **Deploy.**
2. Select any one of the deployment options:
   - **Selected node and dependent nodes**
   - **Selected node, dependent nodes, and child nodes**
   - **Selected node and connected nodes**
3. After selecting the deployment option, the **Generate SQL Script** wizard opens. In the wizard, enter details for the following:
   - **Target Database**
   - **Script Directory**
   - **Select Script Directory**

The SQL that is saved consists of SQL generated by the current node and all the parent nodes that are data providers. The lineage ends when it encounters nodes that represent persisted items such as tables or models.

See Also:

- "Deploying Workflows using Object Generation Scripts"
- "Running Workflow Scripts"

4.4.9.6.1 Target Database

Select the version of Oracle Database where the scripts will run from the **Target Database Version** drop down list and click **Next.** The default version is `11.2.0.1.0`.

4.4.9.6.2 Script Directory

Select the directory where the scripts are stored. You can browse for the directory and create a directory for the scripts. Click **Finish.**

See Also: "Select Script Directory"

4.4.9.6.3 Select Script Directory

Click **Browse.** You can either browse to an existing directory or click  icon to create a new directory for the scripts.

- If you create a new directory, click **OK.**
- Click **Select** to select the directory.

4.4.9.7 Cut

Removes the selected object, which could be a node or connection.

You can also delete objects by selecting them and pressing DELETE on your keyboard.

4.4.9.8 Copy

You can copy one or more nodes and paste them into the same workflow or a different workflow. To copy and paste nodes:

1. Select the nodes to copy. To select several nodes, hold down the Ctrl key when you click the nodes.

   The selected node is highlighted. In this example `Classification` is selected. The other node is not selected.

2. Right-click and select **Copy** from the context menu. Alternately, you can press Ctrl+C to copy the selected nodes.

> **Note:** Copying and pasting nodes do not carry any mining models or results from the original node.

4.4.9.9 Paste

The **Paste** option pastes the copied object in the workflow. To paste an object, right-click the workflow and click **Paste.** Alternately, you can press Ctrl+V.

Note: Node names and model names are changed to avoid naming collisions. To preserve names, use Extended Paste.

See Also: "Copy"

4.4.9.10 Extended Paste

The **Extended Paste** option enables you to preserve node and model names. The default behavior for Paste is to change node names and model names to avoid naming collisions.

To go to the Extended Paste option, right-click the workflow and click **Extended Paste.** Alternately, you can press Control+Shift+V.

Note: If model names are not unique, models may be overwritten when they are rebuilt.

See Also: "Copy"

4.4.9.11 Generate Apply Chain

Generate Apply Chain creates a new node that contains the specification of a node that performs a transformation. If you have several transformations performed in sequence, for example, `Sample` followed by a `Custom transform`, then you must select **Generate Apply Chain** for each transformation in the sequence.You must connect the individual nodes and connect them to an appropriate data source.

Generate Apply Chain helps you create a sequence of transformations that you can use to ensure that new data is prepared in the same way as existing data. For example, to ensure that Apply data is prepared in the same way as Build data, use this option.

The Generate Apply Chain option is not valid for all nodes. For example, it does not copy the specification of a Build node.

4.4.9.12 Select All

This option enables you to select all the nodes in the workflow. The selected nodes and links are highlighted in a dark blue border.

4.4.9.13 Toolbar Actions

This option enables you to select actions in the toolbar from the context menu. Current actions are Zoom In and Zoom Out.

See Also: "Managing Workflows using Workflow Controls"

4.4.9.14 Show Event Log

This option opens the View an Event Log dialog box. The Event Log displays information about events in the current connection. It displays errors, warnings, and information messages.

4.4.9.15 Show Runtime Errors

This option is displayed only when running of the node fails at runtime. The Event Log opens with a list of errors. Select the error to see the exact message and details.

See Also: "View an Event Log"

4.4.9.16 Show Validation Errors

This option is displayed only when there are validation errors. For example, if an Association node is not connected to a Data Source node, select **Show Validation Errors** to view the validation error `No build data input node connected`.

You can also view validation errors by moving the mouse over the node. The errors are displayed in a tool tip.

4.4.9.17 Save SQL

The **Save SQL** option generates a SQL script for the selected node:

1. Right-click the node and click **Save SQL.**
2. Select any one of the options to save the generated SQL script:
   - **SQL to Clipboard**
   - **SQL to File**
   - **SQL Script to Clipboard**
   - **SQL Script to File**

   When you save to a file, the system provides a default location. You can browse to change this location. You can also create a folder for scripts.

   The saved SQL includes SQL generated by the current node and all of its parent nodes that are data providers. The SQL lineage ends when it encounters a node that represents persisted objects, such as tables or models.

   The generated script does not generate all behavior of the node. The script does not create any objects. For example, if you select **Save SQL** for a Create Table node, it does not generate a script to create the table. Instead, it generates a script to query the created table.

See Also: "Deploying Workflows using Data Query Scripts"

4.4.9.18 Go to Properties

The Go to Properties option opens the **Properties** pane for the selected node.

See Also: "Managing Workflows and Nodes in the Properties Pane"

4.4.9.19 Navigate

This option is enabled if there are links to other nodes.

Navigate displays the collection of links available from this node. Picking one of the links selects the link and the selected link is highlighted. The link itself has context menu options as well so you can right click and continue with the Navigate option. You can also use the arrow keys to progress to the next node.

4.4.9.20 View Data

After a Data node is run, the Data Source Node Viewer opens.

4.4.9.21 View Models

Select a model from the list to open the model viewer for that model. A model must be built successfully before it can be viewed.

4.4.9.22 View Test Results

For Classification and Regression models only, select a model and view test results for the model.

4.4.9.23 Compare Test Results

For Classification and Regression models, this option displays the test results for all successfully built models to allow you to pick the model that best solves the problem.

4.5 About Parallel Processing

Oracle Data Miner uses the specifications in a workflow to create SQL queries. These queries are passed and run in the Oracle Database.

In Parallel Query or Parallel Processing, multiple processes work simultaneously to run a single SQL statement. By dividing the work among multiple processes, the Oracle Database can run the statement more quickly. For example, suppose four processes handle four different quarters in a year instead of one process handling all four quarters by itself.

The benefits of Parallel Processing:

- Reduces response time for data-intensive operations on large databases such as data warehouses.
- Enhances performance of symmetric multiprocessing (SMP) as statement processing are split up among multiple systems. Certain types of OLTP and hybrid systems also benefit from parallel processing.

In Oracle RAC systems, the service placement of a specific service controls parallel processing. Specifically, parallel processes run on the nodes on which the service is configured. By default, Oracle Database runs parallel processes only on an instance that offers the service used to connect to the database. This does not affect other parallel operations such as parallel recovery or the processing of GV$ queries.

Parallel processing must be configured by a Database Administrator (DBA). For more information on parallel processing in Oracle Database, see:

- *Oracle Database Data Warehousing Guide* or *Oracle Database VLDB and Partitioning Guide* for more information about parallel processing
- *Oracle Real Application Clusters Administration and Deployment Guide* for considerations about parallel processing in Oracle RAC environments

4.5.1 Parallel Processing Use Cases

This section lists some common use cases for parallel processing.

- Building Models in Parallel
- Making Transformation Run Faster, using Parallel Processing and Other Methods

- Running Graph Node in Parallel
- Running a Node in Parallel to Test Performance

4.5.1.1 Building Models in Parallel

Premise of the use case:

- If the input source for a model is a table defined with parallel and no intervening workflow nodes generate a table that changes this state, then models are built in parallel without any changes to the workflow settings.
- If the input source is not already defined in parallel, you can still build the model in parallel:
  - For Classification and Regression models: Turn on Parallel Processing for the workflow. The setting for Classification and Regression will be to Split input into Tables with the **Parallel** option.
  - For all models: Turn on Parallel Processing for the workflow. Insert a Create Table node before the Build node. Use the created table as input to the models.

4.5.1.2 Making Transformation Run Faster, using Parallel Processing and Other Methods

You can make transformations run faster by using Parallel Processing and other techniques by:

- **Turning on Parallel Processing for the workflow:** All nodes that have a form of sample input could have some benefit. If the sample size is small, then the Oracle Database may not generate parallel queries. But a complex query could still trigger parallel processing.
- **Adding a Create Table node:** You can add a Create Table node after expensive transformations to reduce repetitive querying costs.
- **Adding an index to Create Table node:** You can add an index to a Create Table node to improve downstream join performance.

4.5.1.3 Running Graph Node in Parallel

You can run Graph nodes in parallel. Even if no other nodes are parallel, performance may improve.

To run Graph node in parallel:

1. Set parallel processing for the entire workflow.
2. Turn off parallel processing for all nodes except for the Graph nodes.
3. Run the Graph nodes. Graph node sample data is now generated in parallel. If the Graph node sample is small or the query is simple, the query may not be made parallel.

See Also: "Setting Parallel Processing for a Node or Workflow"

4.5.1.4 Running a Node in Parallel to Test Performance

You can run parallel processing on a node just once to see if parallel processing results in improved performance. To run parallel processing:

1. Set parallel processing for the entire workflow.
2. Run the workflow. Note the performance.

3. Now, turn off parallel processing for the workflow.

    See Also: "Setting Parallel Processing for a Node or Workflow"

4.5.2 Oracle Data Mining Support for Parallel Processing

Model scoring is done in parallel for all algorithms and data in Oracle Database 12.1.

All algorithms do not support parallel build. For Oracle Database 12.1, the following algorithms support parallel build:

- Decision Trees
- Naive Bayes
- Minimum Description Length
- Expectation Maximization

All other algorithms support serial build only.

4.6 Setting Parallel Processing for a Node or Workflow

By default, parallel processing is set to `OFF` for any node type. Even if parallel processing is set to `ON` with a preference, the user can override the section for a specific workflow or node.

To set parallel processing for a node or workflow:

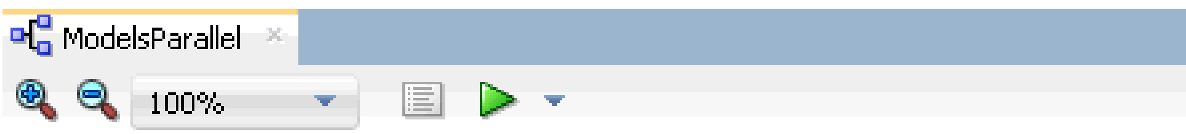
1. Right-click the node and select **Parallel Query** from the context menu. The **Edit Selected Node Settings** dialog box opens.
2. Click **OK.**

4.6.1 Edit Selected Node Settings

The Edit Selected Node Settings dialog box lists all the nodes that comprise the workflow. Here, you can:

- Set parallel processing:
    - For selected nodes: Click the check box in the Parallel column against the nodes for which you want to run parallel query.
    - For all nodes: Click **All** to select parallel processing for all nodes in the workflow.
    - For none: Click **None** to deselect all nodes and turn off parallel processing.
- Specify degree of parallelism: Select the node and click .

    If you specify parallel processing for at least one node, this indication appears in the workflow title bar:



    Parallel Query is either On for Selected nodes, On (for All nodes), or Off. You can click the Parallel Query box to open the Edit Selected Node Settings dialog box.

- Edit preferences for parallel processing: Click to edit the preferences for parallel processing.

  See Also:

  - "Edit Node Parallel Settings"
  - "Editing Parallel Processing Preferences"

4.6.2 Edit Node Parallel Settings

The default degree of parallelism is System Determined. To specify a value:

1. In the Edit Node Parallel Setting dialog box, click the option **Parallel Query On** to turn on parallel processing for the selected node.
2. Click the option **Degree Value** and select a higher value. The default value is `1`.
3. Click **OK.** The settings are displayed in the **Parallel** and **Degree of Parallel** columns for the node.

4.6.3 Editing Parallel Processing Preferences

To edit parallel query preferences:

1. In the **Edit Selected Node Settings** dialog box, select the node and click .
2. Select any one of the following options:
   - **Edit Node Default Settings:** To edit the node settings. The Preferences dialog box opens, listing all the nodes. You can select and edit nodes here.
   - **Change Settings to Default:** To revert the node settings to default.
3. Click **OK.**

5 Data Nodes

Data nodes specify data for a mining operation, to transform data, or to save data to a table. The input for Oracle Data Mining operations is a table or view in Oracle Database or an Oracle Data Miner node that is part of a data flow.

The Data nodes are available in the Data section in the **Components** pane.

The following Data nodes enable you to specify data in a workflow and to create and modify tables:

- Create Table or View Node
- Data Source Node
- Explore Data Node
- Graph Node
- SQL Query Node
- Update Table Node

5.1 Create Table or View Node

The Create Table or View node is a type of node that enables you to save the results in a table in the schema to which you are connected. For example, use a Create Table or View node to save the results of an Apply node to a table.

The Create Table node automatically uses compression when possible.

The Create Table node can run in parallel.

The benefits of Create Table or View node are:

- **Data Persistence:** A Create Table or View node saves the data that flow into the node as a view or table in the database. If you create a table, then the actual data persists. If you create a view, then the SQL definition (full lineage) persists. Output from this node is the data provided by the view or table.
- **Performance Improvement:** If you perform one or more complex transformations, such as *Joins* and *Aggregations*, and save the result of the transformations as a table, then the subsequent operations are faster. For example, you can perform an *Aggregation* and a *Join*, create a table that contains the results of the transformation, and then use the table as input for building the model. Therefore, you will create a table from the Join node. The Classification models are built against this table.

See Also:

- "Create Table Node and Compression"
- "About Parallel Processing"

5.1.1 Working with the Create Table or View Node

You can perform the following tasks with a Create Table or View node:

- "Create a Create Table or View Node"
- "Create Table Node and Compression"
- "Edit Create Table or View Node"
- "View Create Table or View Data"
- "Create the Table or View Context Menu"

5.1.1.1 Create a Create Table or View Node

You create a Create Table or View node to save a data flow to a table or view. You can connect a Create Table or View node to any node that create a data flow, such as an Apply node.

To create a Create Table or View node:

1. In the Components pane, go to the **Workflow Editor** and expand **Data**.

   If the Components pane is not visible, then in the SQL Developer menu bar, go to **View** and click **Components.** Alternately, press Ctrl+Shift+P to dock the **Components** pane.
2. In the **Data** section, click the **Create Table or View** icon.
3. Drag and drop the **Create Table or View** node from the **Components** pane to the **Workflow** pane. This adds the Create Table or View node to the workflow.
4. Right-click the node from which to create the table, and click **Connect** in the context menu.
5. Draw a line from the selected node to the **Create Table or View** node and click again.
6. You can accept the default settings of the **Create Table or View** node or edit the default settings. Click **Edit** in the context menu.
7. To create the table, right-click the **Create Table or View** node and select **Run** from the context menu.

   The table is automatically compressed, if possible.
8. After running the node, right-click the node and select **View Data** to view the results.

See Also:

- "Edit Create Table or View Node"
- "Create Table Node and Compression"
- "View Create Table or View Data"

5.1.1.2 Create Table Node and Compression

Table compression does the following:

- Saves disk space
- Reduces memory use in the buffer cache
- Speeds up the running of queries during reads

When Oracle Data Miner creates a table, if Oracle Data Miner determines that the data does not have nested columns `DM_NESTED_*`, then it uses compression to create the table. Oracle Data Miner creates tables in the Create Table node, and creates split data sets for testing in the Classification and Regression model build.

For the Create Table node, Oracle Data Miner also verifies that a primary key is not defined and no indexes are defined.

5.1.1.3 Edit Create Table or View Node

You can modify the operation of the Create Table or View node. To edit a Create Table or View node:

1. Double-click the node or right-click and select **Edit.** The **Edit Create Table or View** node dialog box opens.
2. In the **Edit Create Table or View** node dialog box, you can make the following changes:
   - **Name:** Displays the default table name. You can change the default name of the table or view to any unique and valid name.
   - **Type:** By default, the object type is `Table.` You can change the default to `View` by selecting the appropriate option.
   - **Auto Input Column Selection:** Deselect the check box to manually select and edit input columns. You can perform the following tasks:
     – Delete columns: Select the column and click .
     – Edit columns: Select the column and click . The **Select Column** dialog box opens.

     Note: If JSON data is present, select the column and click  to specify Data Guide settings in the **Edit Data Guide** dialog box.

     – Edit the data type entry in the **Target Type** column for JSON data only. Click the data type in the Target column to select an option from the in-place drop-down list.
     – Specify Key, Index and Alias for each column.

     Note: If you create a table that join another object, adding an index will speed up the join.

   - **JSON Settings:** Click **JSON Settings** to specify node settings that determines how Data Guides are generated. This option is applicable only for JSON data.
3. Click **OK.**

See Also:

- "Select Columns"
- "Edit Data Guide"
- "JSON Settings"

5.1.1.4 Select Columns

By default, all columns are selected. You must select at least one attribute for the Create Table or View definition to be complete. You can perform the following tasks:

- Remove columns: Select the attribute in the **Selected Attribute** section and move it to the **Available Attribute** section. Click **OK.**
- Add columns: Select the attribute in the **Available Attribute** section and move it to the **Selected Attribute** section. Click **OK.**

5.1.1.5 View Create Table or View Data

After the node runs successfully, right-click the node and select **View Data.** The data is displayed in the **Data Source Node** viewer.

See Also: "Data Source Node Viewer"

5.1.1.6 Create the Table or View Context Menu

To view the context menu, right-click the **Create Table or View** node. You can perform the following tasks:

- Connect
- Edit. See "Edit Create Table or View Node" for more information.
- Run. See "Running a Data Source Node" for more information about how the node is run.
- Force Run
- View Data. See "Data Source Node Viewer" for more information.
- Deploy
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors
- Show Validation Errors
- Go to Properties
- Navigate

5.1.1.7 Create the Table or View Node Properties

In the **Properties** pane, you can examine and change the characteristics or properties of a **Create Table or View** node. To manage the Create Table or View node from **Properties** pane:

1. Right-click the node and click **Go to Properties.**
2. The **Properties** pane opens in the lower right panel in the SQL Developer window. The name of the table or view identifies the Create Table or View property. The **Properties** pane consists of the following:
   - Table
   - Columns
   - Cache
   - Details

   See Also: "Performing Tasks from Workflow Context Menu"

5.1.1.7.1 Table Displays the name of the Table or View. You can perform the following tasks:

- Change the name of the Table or View: If you change the default name of the table, name of the node changes to match the name of the table. For example, if you change the default name of the table to `PREDICTIONS`, the name of the Create Table or View node also changes to `PREDICTIONS`.
- Change the object type from table to view: The default type is `Table`. To create a view, click **View.**

5.1.1.7.2 Columns Displays the columns of the table. The default setting enables automatic behavior. If you deselect **Auto Input Column Selection,** then you can manually select and edit the columns in the table. You can:

- Delete columns: Select the column to delete and click .
- Edit columns: Select the column and click . Edit the required changes in:
  – **Select Column** dialog box
  – **Edit Data Guide** dialog box for JSON data

  See Also:
  - "Automatic Behavior"
  - "Select Columns"
  - "Edit Data Guide"

5.1.1.7.3 Automatic Behavior If **Auto Input Column Selection** is selected, then the possible scenarios are:

- Scenario 1: When input is connected
  All columns in the input node are selected. The node becomes valid. It assumes that at least one column is included in the specification.
- Scenario 2: When input is disconnected
  All columns are automatically removed. The node becomes invalid.
- Scenario 3: When the input node is edited
  The following are the possible edit scenarios:

- Add column: The column is added to the Create Table or View node, if it is compatible.
- Remove column: The column is removed from the Create Table or View node.
- Edit column: If a column is edited, any change in the column data type could trigger an invalid state in the node.

Note: If the **Auto Input Column Selection** option is deselected, you must manually add and remove column specifications from the Create Table or View node.

5.2 Data Source Node

A Data Source node defines source data for the workflow. For example, a Data Source node specifies the build data for a model. Any table or view accessible to the user can be selected as the source. This node does not allow any input node connections. Data Source nodes rely on database resources for their definitions. It may be necessary to refresh a node definition if the database resources change. For example, if the resources are deleted or re-created.

Data Source nodes can run in parallel.

Only certain data types are allowed in a Data Source node. Columns with other data types are excluded.

See Also:

- "About Parallel Processing"
- "Supported Data Types for Data Source Nodes" for the list of supported data types
- "Support for Date and Time Data"
- "Refresh Nodes"

5.2.1 Supported Data Types for Data Source Nodes

Most of the basic Oracle data types are supported by the Data Source node. Object-based data types can be included, but each object type has to be well understood. Object data types require storage clauses to be defined at appropriate levels with the object hierarchy.

These data types are fully supported:

- `VARCHAR2`
- `CHAR`
- `FLOAT`
- `NUMBER`
- `CLOB`
- `NESTED_NUMERICALS`
- `NESTED_CATEGORICALS`

These data types are supported by Oracle Data Mining 12*c* Release 1 (12.1):

- `BINARY_DOUBLE`

- BINARY_FLOAT
- DM_NESTED_BINARY_DOUBLES
- DM_NESTED_BINARY_DOUBLES
- BLOB, for Text only

The BINARY data types and BLOB are supported by Oracle Data Miner if you are connected to Oracle Data Mining 12*c* Release 1 (1.2) or later.

These data types for date and time are partially supported:

- DATE
- TIMESTAMP
- TIMESTAMP_WITH_LOCAL_TIMEZONE
- TIMESTAMP_WITH_TIMEZONE
- TIMESTAMP_WITH_LOCAL_TIMEZONE

Oracle Database 12.1.0.2 supports JSON data in the following data types. Oracle Data Miner derives pseudo JSON data types from these data types:

- VARCHAR2
- CLOB
- BLOB
- RAW
- NCLOB
- NVARCHAR2

See Also: "Support for Date and Time Data"

5.2.2 Support for Date and Time Data

The *Transform* transformation partially supports the data and time data types DATE, TIMESTAMP, TIMESTAMP_WITH_LOCAL_TIMEZONE, and TIMESTAMP_WITH_TIMEZONE, as follows:

- Attributes with data and time data types can be binned using Equal Width or Custom binning.
- You can apply Statistic or Value missing values treatments to attributes with the data and time data types.

Attributes with data and time types are analyzed by the Explore Data node.

Data and time data types cannot be used for other functions in Oracle Data Miner. In particular, date and time data types cannot be the targets of model builds.

See Also:

- "Transforms Nodes"
- "Explore Data Node"

5.2.3 Working with the Data Source Node

You can perform the following tasks with Data Source nodes:

- Create a Data Source Node

- Edit a Data Source Node
- Define a Data Source Node

5.2.3.1 Create a Data Source Node

You create a Data Source node after creating a workflow. To create a Data Source node and attach data to it:

1. In the **Components** pane, go to the **Workflow Editor** and expand **Data.**

   If the **Components** pane is not visible, then in the SQL Developer menu bar, go to **View** and click **Components.** Alternately, press Ctrl+Shift+P to dock the **Components** pane.

2. In the **Data** section, click **Data Source** node icon.
3. Drag and drop the Data Source node from the **Components** pane to the **Workflow** pane. This adds the Data Source node to the workflow. The **Define Data Source** dialog box opens.
4. In the **Define Data Source** dialog box, you can select a table or view. By default, the tables in your schema are listed. You can add tables from other schemas to which you have access, in the **Edit Schema List** dialog box.
5. Click **Next.**
6. In the **Define Data Source - Select Columns** dialog box, add or remove attributes to the table.
7. Click **Finish.**
8. In the **Select Table** window, select the table or view to use. Click **OK.** The **Properties** pane displays information about the table or view that you selected. The node can now be run.

   The default name of the node is the name of the table or view that you select. If `SH.CUSTOMERS` as the table, the node is named `CUSTOMERS`.

   See Also:

   - "Define a Data Source Node"
   - "Edit Schema List"

5.2.3.1.1 Edit Schema List

By default, tables and views from the schema to which you are connected are listed.

To add other schemas:

- If this is the first time that you are adding schemas, then click **Add Schemas.**
- If you have already added schemas, then click **Edit Schemas.**

The **Edit Schema List** dialog box opens. **Available Schemas** is a list of the schemas to which you have access. To display tables and views in one of these schemas, move the name to **Selected Schemas.**

For example, to display tables in the SH schema, move SH to the Selected Schemas list. Click **OK**.

You return to the Define Data Source wizard. Select **Include Tables from Other Schemas**. Tables and views in the added schemas are listed. For example, if you selected SH, you now see tables, such as `SH.CUSTOMERS`, in the Available Tables/Views list. You can select these tables as data sources.

5.2.3.2 Define a Data Source Node

You can use this interface to:

- Define table and attributes for a new Data Source node.
- Edit an existing data source node.

If you import a workflow that uses tables or views that are not available, then you can use this wizard to define a data source to replace the missing one. The wizard also detects input columns with JSON data types.

The wizard has two steps:

1. **Select Table:** Here, the tables and views to which you have access, are displayed. The schemas are the schemas to which you are connected, along with the schemas added using the **Edit Schema List.** Select the table or view types. The columns and data are displayed in the lower pane in the following tabs:
   - **Columns:** Lists the columns of the selected table in a grid. For each column, Data Type, Mining Type, Length, Precision, Scale, and Column ID are displayed.
   - **Data:** Displays the data in the table arranged according to Column ID.
     - To include all attributes in the table, click **Finish.**
     - To manually exclude certain attributes in the table, click **Next.**
     - To change the table selection for an existing node, click **Edit Data Source Node.**
2. **Select Columns:** By default, the **Define Data Source** wizard includes all columns in the Table or View. You can perform the following tasks:
   - Include a column by moving the attribute from **Available Attribute** to **Selected Attributes** section.

     A drop-down list is available for JSON data. If there is an input table with JSON data, then the wizard detects the column as a JSON column and displays it in the Data Type column. If the wizard cannot detect JSON data, then you can manually change the data type for the column by clicking the drop-down list.
   - **Specify Data Guide Settings:** Applicable only for JSON data type. Select the attribute and click  to specify Data Guide settings in the **Edit Data Guide** dialog box.
   - **JSON Settings:** Click **JSON Settings** to specify node settings that determines how Data Guides are generated. This option is applicable only for JSON data.

     See Also:

     - "Edit a Data Source Node"
     - "Edit Schema List"
     - "Edit Data Guide"
     - "JSON Settings"

5.2.3.3 Edit Data Guide

The **Edit Data Guide** dialog box enables users to specify how data guide should be generated for a selected JSON type column. The dialog box has two tabs:

- **Structure:** Displays the JSON data structure for the selected column

- **Data:** Displays the content of the imported or generated Data Guide table

You can perform the following tasks:

- **Generate Data Guide:** You can generate a new data guide table whenever the node is run or rerun. To generate a data guide, select any of the following options from the **Data Guide Generation** drop-down list:
  - **Default:** If the option **Generate Data Guide if necessary** in the **JSON Settings** dialog box is selected, then the data guide table is generated whenever the node is run or rerun. Otherwise, no data guide is generated.
  - **On:** Select this option to generate a new data guide table whenever the node is run or rerun.
  - **Off:** Select this option to *not* generate a new data guide table whenever the node is run or rerun.
  - **Import from Workflow:** Select this option to import a data guide from an existing JSON type column defined in a Data Source node or Create Table node within the same workflow or from a different workflow using the **Select Data Guide** dialog box.
  - **Import from file:** Select this option to import a data guide from a CSV file. The process validates the data guide, that is, it verifies the correctness of the column header, JSON path format, JSON types, and so on during an import operation.
- **Remove Data Guide:** To delete the current Data Guide table, click .
- **Export Data Guide:** Select this option to export the current data guide table to a CSV file. If you find the generated data guide does not fully represent the underlying JSON data, then you have the option to export the data guide and add any missing JSON paths. You can then import the data guide back to replace the generated one.

See Also:

- "Select Data Guide"
- "JSON Settings"

5.2.3.3.1 Select Data Guide

The Select Data Guide dialog box enables you to import a data guide table from an existing JSON type column defined in a Data Source node or Create Table node within the same workflow or from a different workflow. Only completed nodes with generated JSON schema are displayed. To import a data guide table:

1. In the **Show** field, select a workflow from the drop-down list.
2. Select the nodes to import.
3. Click **OK.**

5.2.3.4 JSON Settings

In the **JSON Parsing Settings** dialog box, you can specify node settings that determine how data guides are generated. A data guide is used whenever a JSON structure is present in the UI, for example, JSON Query node. Because the data guide table generation could be time consuming, especially for large JSON data, the following settings offers some control on the table generation:

- **Generate Data Guide if necessary:** By default, this option is selected. It generates a data guide for the JSON type columns. Deselect this option if JSON data is not used in the product. Therefore, no Data Guide will be generated.
- **Sampling:** Defines how many JSON documents stored in a column are to be processed, to generate the data guide table. The JSON document contains the entire content of the JSON column for a given row.
  - **Max number of documents:** `2000` (Default). Use the arrows to change this setting.
- **Limit Document Values to Process:** Defines how many JSON values (Number, String, Boolean) are to be parsed in the document to generate the data guide table.
  - **Max number per document:** `10,000` (Default). Use the arrows to change this setting.

5.2.3.5 Edit a Data Source Node

To edit an existing Data Source node:

1. Double-click or right-click the node and select **Edit.** The Edit Data Source Node opens.
2. In the **Edit Data Source Node** dialog box, you can change the attributes selected in the current Data Source. You can perform the following tasks:
   - **Change attribute selection:** To change the attribute selection, move the attributes from the **Available Attributes** pane to the **Selected Attributes** pane by using the arrows. For example, to remove ATTRIBUTE1 from the data source, move ATTRIBUTE1 from the **Selected Attributes** list to the **Available Attributes** list. After you are done, click **OK.**
   - **Select a different table:** To select a different table or view, click **Edit.** The **Define Data Source** dialog box opens.

   See Also: "Define a Data Source Node"

5.2.3.6 Running a Data Source Node

To run a valid Data Source node, right-click the node and select **Run**.

The Oracle Data Miner server generates a sample of the selected table or view. The size of the table and type of sampling used is determined by the sample settings of the node.

If the node is complete but is required to provide data to a child node that is being run, then the node is run for validation to ensure that the columns and table still exist. If there are any errors, then the node state is set to `Error` and affected attributes are set to `Invalid`.

5.2.3.7 Data Source Node Context Menu

To see the context menu, right-click a Data Source node. These selections are available in the context-menu:

- Connect
- Edit: Opens the **Edit a Data Source Node** dialog box.
- Attributes: Opens the **Select Attributes** dialog box.
- Define a Data Source Node

- Run: Runs the node, as described in Running a Data Source Node.
- Force Run
- View Data: Opens the **Data Source Node Viewer** dialog box.
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query: See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed only if there are validation errors.
- Go to Properties
- Navigate

5.2.3.7.1 Select Attributes The **Select Attributes** dialog box enables you to move attributes between the **Available Attributes** list and the **Selected Attributes** list. To deselect an attribute, move it to the **Available Attributes** list.

You can search the **Available Attributes** list for attributes.

Use the shuttle controls to move attributes between the lists.

When you have finished selecting attributes, click **OK**.

5.2.4 Data Source Node Viewer

To view data, right-click the node and select **View Data** from the context menu. The Data Viewer opens.

Note: You can view data only when the Data Source node is in the `Valid` state.

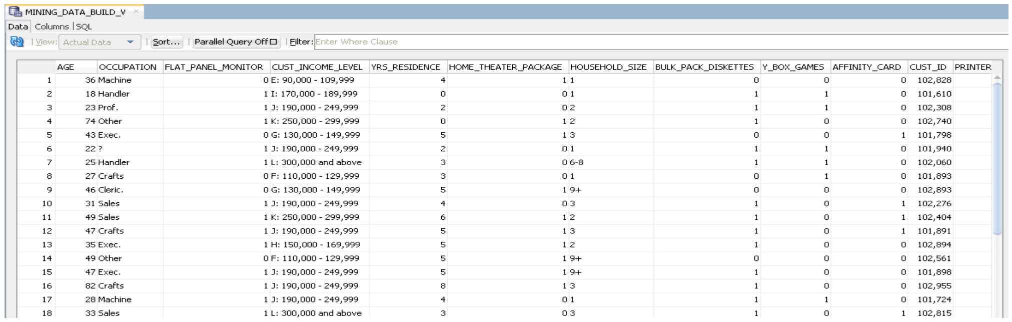
The data viewer has these tabs:

- Data
- Graph
- Columns
- SQL

5.2.4.1 Data

The **Data** tab displays a sample of the data. The Data Viewer provides a grid display of the rows of data either from the sampling defined in Cache, or pulled through the lineage of nodes back to the source tables.

The **Data** tab for `MINING_DATA_BUILD_V`:



	AGE	OCCUPATION	FLAT_PANEL_MONITOR	CUST_INCOME_LEVEL	YRS_RESIDENCE	HOME_THEATER_PACKAGE	HOUSEHOLD_SIZE	BULK_PACK_DISKETTES	Y_BOX_GAMES	AFFINITY_CARD	CUST_ID	PRINTER
1	36	Machine	0	E: 90,000 - 109,999	4	1	1	0	0	0	102,828	
2	18	Handler	1	I: 170,000 - 189,999	0	0	1	1	1	0	101,610	
3	23	Prof.	1	J: 190,000 - 249,999	2	0	2	1	1	0	102,308	
4	74	Other	1	K: 250,000 - 299,999	0	1	2	1	0	0	102,740	
5	43	Exec.	0	G: 130,000 - 149,999	5	1	3	0	0	1	101,798	
6	22	?	1	J: 190,000 - 249,999	2	0	1	1	1	0	101,940	
7	25	Handler	1	L: 300,000 and above	3	0	6-8	1	1	0	102,060	
8	27	Crafts	0	F: 110,000 - 129,999	3	0	1	0	1	0	101,893	
9	46	Cleric.	0	G: 130,000 - 149,999	5	1	9+	0	0	0	102,893	
10	31	Sales	1	J: 190,000 - 249,999	4	0	3	1	0	1	102,276	
11	49	Sales	1	K: 250,000 - 299,999	6	1	2	1	0	1	102,404	
12	47	Crafts	1	J: 190,000 - 249,999	5	1	3	1	0	1	101,891	
13	35	Exec.	1	H: 150,000 - 169,999	5	1	2	1	0	0	102,894	
14	49	Other	0	F: 110,000 - 129,999	5	1	9+	0	0	0	102,561	
15	47	Exec.	1	J: 190,000 - 249,999	5	1	9+	1	0	0	101,898	
16	82	Crafts	1	J: 190,000 - 249,999	8	1	3	1	0	0	102,955	
17	28	Machine	1	J: 190,000 - 249,999	4	0	1	1	1	0	101,724	
18	33	Sales	1	L: 300,000 and above	3	0	3	1	0	1	102,815	

A partial view of the **Data** tab for a `MINING_DATA_BUILD_V`

****

You can perform the following tasks:

- **Refresh:** Click  to refresh the data.
- **View:** Click **View** and select either **Actual Data** or **Cached Data.**
  Cached data is available only if you cache data in the **Cache** section of the **Properties** pane.
- **Sort:** Click **Sort** to sort data according to applicable criteria. Displays the **Select Column to Sort By** dialog box.
- **Filter:** In the **Filter** field, enter a `WHERE` clause to select data.

> **See Also:**
>
> - "Cache"
> - "Select Column to Sort By"

5.2.4.1.1 Select Column to Sort By

This dialog box enables you to:

- Select multiple columns to sort
- Determine the column ordering
- Determine ascending or descending order by column
- Specify `Nulls First` so that null values appear before real data values

The sort order is preserved until you clear it.

Column headers are also sort-enabled to provide a temporary override to the sort settings.

5.2.4.2 Graph

This tab enables you to create graphs of numeric data.

> **See Also:** "Graph Node" for more information.

5.2.4.3 Columns

The Column tab is a list of all the columns that are output from the node. For each column, the Name, Data Type, Mining Type, Length, Precision and Scale (for floating point), and Column ID are displayed.

- If the node has not run, the table or view structure provided by the database is displayed.
- If the node has run successfully, then the structure of the sample table is displayed, based on the sampling defined when the node was specified.

There are several filtering options that limit the columns displayed. The filter settings with the `(or)/(and)` suffixes allow you to enter multiple strings separated by spaces. For example, if the `Name/Data Type/Mining Type(or)` is selected, then the filter string `A B` produces all columns where the name or the data type or the mining type starts with the letter *A* or *B*.

5.2.4.4 SQL

The SQL tab provides the SQL Details text area. The text area displays the SQL code that generates the data provided by the actual view shown in the Data tab.

The SQL can be a stacked expression that includes SQL from parent nodes, depending on what lineage is required to access the actual data.

You can copy the SQL and run it in a suitable SQL interface. **Select All** (Ctrl+A) and **Copy** (Ctrl+C) are enabled.

The Search control is a standard search control that highlights matching text and searches forward and backward.

5.2.5 Data Source Node Properties

Data Source node properties enables you to examine and change characteristics of a Data Source node. To open the Properties pane, right-click the node and select **Go to Properties** from the context menu. The **Properties** pane for a Data Source node has these sections:

- Data
- Cache
- Details

> **See Also:** "Properties"

5.2.5.1 Data

The Data section consists of the following:

- **Source Table**: Displays the name of the source table or view for the Data Source node. If no source table is associated with the node, click **...** to the right of **Source Table**. A list of tables and views that are accessible from the data mining account is displayed. You can select the table or view. You can also use this process to change the table or view.
- **Data**: Displays the attributes in a grid. For each attribute, the name, the alias, and the data type are displayed. You can perform the following tasks:
  - Create an alias for an attributed by entering the alias in the appropriate cell.
  - Filter attributes.

- Delete attributes. Select the attribute and click .
- Edit attributes. Select the attribute and click .
- **Select Attributes** to include in the Data Source.
- Refresh the node. Click .

See Also:

- "Filter"
- "Select Attributes"
- "Refresh Nodes"

5.2.5.2 Cache

The **Cache** section provides the option to generate a cache for output data. You can change this default using the transform preference.

You can perform the following tasks:

- **Generate Cache of Output Data to Optimize Viewing of Results:** Select this option to generate a cache.
  The default setting is to *not* generate a cache.
  - **Sampling Size:** You can select caching or override the default settings.
    Default sampling size is `Number of Rows`
    Default Value=`2000`

See Also: "Transforms"

5.2.5.3 Details

The **Details** section displays the name of the node and any comments about it. You can change the name and comments in the fields here:

- **Node Name**
- **Node Comments**

5.3 Explore Data Node

The Explore Data node provides the profile of any input data source. Explore Data Statistics can either be based on all data or on a sample of the data. The Explore Data node enables you to do the following:

- View common statistics and a histogram for each column. Optionally, you can select a **Group By** attribute and have multivariate view of histograms generated.
- Generate an output flow containing the results of all the statistical analysis. You can connect any source of data input to an Explore Data node. For example, you can attach an Explore Data node to an Apply node.
- Analyze attributes with all supported data types and the following data types for date and time:
  - `DATE`
  - `TIMESTAMP`
  - `TIMESTAMP_WITH_TIMEZONE`
  - `TIMESTAMP_WITH_LOCAL_TIMEZONE`

- Run Explore Data node in parallel.
- Create graphs.
- Export the statistic generated by the Explore Data node using SQL Developer.

See Also:

- "About Parallel Processing"
- "Graph Node"
- "Export Explore Node Calculations"

5.3.1 Working with the Explore Data Node

You can perform the following tasks with the Explore Data node:

- Create an Explore Data Node
- Edit the Explore Data Node
- Perform Tasks from the Explore Data Node Context Menu
- Export Explore Node Calculations

5.3.1.1 Create an Explore Data Node

You create an Explore Data node and connect it to a data source to analyze the data in the data source. You can connect an Explore Data node to any data source.

To create an Explore Data node:

1. In the **Components** pane, go to the **Workflow Editor** and expand **Data.**
   If the **Components** pane is not visible, then in the SQL Developer menu bar, go to **View** and click **Components.** Alternately, press Ctrl+Shift+P to dock the **Components** pane.
2. In the Data section, click **Explore Data.**
3. Drag and drop the **Explore Data** node from the **Components** pane to the workflow pane. This adds the Explore Data node to the workflow.
4. Right-click the node in the workflow that you want to analyze and click **Connect** in the context menu.
5. Draw a line from the node to analyze to the **Explore Data** node and click again.
6. To generate statistics and analyze data, right-click the Explore Data node, and click **Run.**
7. After running the node, right-click the node and click **View Data.** The data is displayed in the **Explore Data Node Data Viewer.**

By default, all attributes in the data source are analyzed. You can specify specific attributes to analyze.

See Also:

- "Explore Data Node Data Viewer"
- "Edit the Explore Data Node"

5.3.1.2 Edit the Explore Data Node

You can edit the Explore Data node by either double-clicking the node or by right-clicking the node and selecting **Edit.**

The Edit Explore Data Node editor has these tabs:

- Input (Explore)
- Statistics (Explore)

5.3.1.2.1 Input (Explore)

Click the **Input** tab to specify the attributes to analyze. By default all attributes in the data source are analyzed.

The default is to select no **Group By** attribute. **Group By** displays a sorted list of attributes that are limited to data types that can be binned by Explore Data. Selecting a **Group By** attribute enables you to analyze data based on the selected attribute. For example, suppose that the data contains AGE and GENDER as attributes. If you select AGE as the **Group By** attribute, then the histograms for GENDER show the age composition of each value of GENDER.

Change the list of attributes to analyze in either of these ways

- Right-click the Explore Data node and select **Edit**. The **Select Attributes** dialog box opens.
- Select the Explore Data node. The **Explore** tab in the Explore Data node **Properties** pane enables you to change the list of attributes.

> **See Also:**
>
> - "Select Attributes"
> - "Explore Data Node Properties"

5.3.1.2.2 Select Attributes

By default, all attributes of the data source are selected. If you do not want to view statistics for an attribute, move it from the **Selected Attributes** list to the **Available Attributes** list.

You can sort the lists by Name or by Data Type.

When you have finished, click **OK**.

5.3.1.2.3 Statistics (Explore)

Click the **Statistics** tab to specify the statistics to calculate. You can also change statistics in the **Statistics** section of Explore node **Properties** pane.

The statistics tab contains a list of statistics available. For each statistic, there is a brief definition and an indication of the cost to calculate that statistic.

If you change any of the default selections, click **Restore Defaults** to change all selections to the default selections.

You can search for an individual statistic by name.

By default, Oracle Data Miner calculates these statistics:

- Average
- Distinct percent
- Maximum
- Median
- Minimum
- Mode sampled
- Percent NULLs
- Standard deviation

- Variance

The default statistics have a low or medium cost to calculate.

Skewness and kurtosis have a high cost to calculate. They are not selected by default. You can select them if necessary.

Calculating mode using a sample is a low cost operation. Calculating the mode using all available data is a very high cost operation. To calculate mode using sample data or all available data, click **Mode (Sampled)** in the **Edit Explore Data Node** dialog box.

See Also:

- "Mode (Sampled)"
- "Export Explore Node Calculations" for information about exporting statistics to a spreadsheet.

5.3.1.2.4 Mode (Sampled) The default is to calculate mode using a sample of 2000 records. You can calculate the mode using all available data. This calculation has a very high cost.

To use all available data, click the **Available Data** option and click **OK**.

5.3.1.3 Perform Tasks from the Explore Data Node Context Menu

The context menu for an Explore Data node has these selections:

- Connect
- Edit: Opens the **Select Attributes** dialog box.
- Validate Parents
- Run
- Force Run
- View Data: Opens the Data Source Node Viewer.
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query: See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed only if there are validation errors.
- Go to Properties
- Navigate

5.3.1.4 Export Explore Node Calculations

You can export the statistics calculated by an Explore node to a Microsoft Excel spreadsheet. To export:

1. When an Explore node runs, it writes the statistics that it calculates to a database table. The name of the table is in the **Output** section of the Explore Data node **Properties** pane. Suppose that the name of the table is `OUTPUT_8_3`.
   If the **Properties** pane of the node is not visible, then right-click the Explore node and select **Go to Properties.**
2. In SQL Developer, go to the **Connections** tab. Expand the connection that you used for data mining.
3. Expand **Tables.** Find `OUTPUT_8_3` and right-click `OUTPUT_8_3` .
4. Select **Export** in the context menu. The **Export Wizard** opens.
5. In the **Export Wizard**:
   - Deselect **Export DDL.**
   - In the **Export Data** section, select the version of Microsoft Excel to which you want to export from the **Format** drop-down list.
   - Specify a file name. Alternately, you can accept the default.
   - Click **Next.**
6. Click **Finish.** SQL Developer exports the table to the spreadsheet.

The spreadsheet contains the statistics. It contains the names of the histograms generated by the Explore Data node.

To export an individual histogram, right-click the histogram and save the graphic.

5.3.2 Explore Data Node Data Viewer

After the node runs successfully, you can view the data. To view the data, right-click the node and select **View Data**. The viewer opens in a new tab.

The viewer enables you to see the statistics and other analyses performed by the Explore Data node.

The viewer displays the information in the following tabs:

- Statistics
- Columns
- Data
- SQL

5.3.2.1 Statistics

The **Statistics** tab show the statistics calculated by the Explore Data node.

Attributes are listed the Statistics grid. For each attribute, the name, data type, histogram, and summary of the statistics are displayed.

To view a large version of the histogram for an attribute, select the attribute. The histogram is displayed below the grid. If the attribute has null values, then the histogram has a separate bin labeled `Null bin`. The histogram may also contain an `Others` bin consisting of values that are not `NULL` but are not in any other bin.

For more information about the histogram, go to the large histogram and right-click the attribute name. These choices are available:

- **Copy to Clipboard:** Copies the histogram to the Microsoft Windows clipboard. You can paste this histogram in to any rich editor, such as a word processor or image editor.
- **Save Image As:** Exports the image to a PNG file that can be stored in the file system.
- **View Data:** Displays the data used to create the histogram in a pop up window. You can search for attribute name, attribute value, or attribute percent. Click **Close** when you have finished.

Histograms created by Oracle Data Miner use a sample of the data and can be slightly different each time they are displayed.

For all data types, a histogram is created. The number of distinct values, the percentage of `NULL` values, and the number of distinct values are displayed.

Additional statistics calculated depend on the data type of the attribute:

- For character attributes `VARCHAR2`, the Mode is calculated.
- For numeric attributes `NUMBER` or `FLOAT`, the following are calculated:
  - Average
  - Minimum value
  - Maximum value
  - Standard deviation
  - Skewness (a measure of the asymmetry of the distribution)
  - Kurtosis (a measure of flatness or peakedness of the curve describing a frequency distribution in the region about its mode)

5.3.3 Explore Data Node Properties

To open the **Properties** pane, right-click the node and select **Go to Properties** from the context menu.

The Explore Data node **Properties** pane has these sections:

- Input (Properties)
- Statistics (Properties)
- Output
- Histogram
- Sample
- Details

5.3.3.1 Input (Properties)

The **Input** section lists the attributes that are analyzed.

Attributes are listed in a grid. For each attribute, the name and the data type are displayed. To sort the list of attributes by name or data type, click the heading in the grid.

You can specify a **Group By** attribute. Select an attribute from the list.

You can either choose the default settings or select columns:

- To select columns, deselect **Auto Input Columns Selection.**
- To delete an attribute, select it and click .
- To edit an attribute, click .

The **Select Attributes** dialog box opens.

See Also:

- "Edit the Explore Data Node" for more information about the Group By attribute.
- "Select Attributes"

5.3.3.2 Statistics (Properties)

The **Statistics** section lists the calculated statistics.

See Also: "Statistics (Explore)"

5.3.3.3 Output

The **Output** section lists the columns in the data source. The names and data type of each column is listed in a grid. You can search the grid by name (the default) or by data type.

To clear the search, click .

5.3.3.4 Histogram

You can use bins to create histograms. This tab lists the default number of bins for the following types of bins:

- Numerical Bins
- Categorical Bins
- Date Bins
- Null Values
- Other Values

The default number of bins for all of these bin types is 10. The default specifies a maximum number of bins.

5.3.3.5 Sample

The data is sampled to support data analysis. The default is to use a sample. The **Sample** tab has the following selections:

- **Use All Data:** By default, **Use All Data** is deselected.
- **Sampling Size:** The default is `Number of Rows` with a value of 2000. You can change sampling size to `Percent`.
  The default is 60 percent.

5.4 Graph Node

A Graph node creates a two-dimensional graph of numeric data. A Graph node is not a data provider, which means that it cannot be connected to another node.

Note: You cannot generate code from a Graph node.

You can perform the following tasks:

- Create and edit one or more graphs of different types such as line, scatter, bar, histogram, and box.
- Create graphs with actual data and sample data.
- Run a Graph node in parallel.

See Also: "About Parallel Processing"

5.4.1 Types of Graphs

The Graph node enables you to create two-dimensional graphs of numeric data in several ways. You can create the following types of graphs:

- **Line plot:** Uses a line to connect the data points. Line plots are useful for identifying if two variable are correlated. Oracle Data Miner supports two-dimensional line plots.
- **Bar plot:** Compares values. The height of the bar represents the measured value or frequency.
- **Histogram:** Shows frequencies, as adjacent rectangles, erected over discrete intervals (bins), with an area equal to the frequency of the observations in the interval.
- **Scatter plot:** Display values in two discrete sets. One variable determines the position on the horizontal axis and the other variable determines the position on the vertical axis.
- **Box plot or Box graph:** Graphically depicts groups of numeric data using the quantiles of the data.

Graph nodes require you to specify one or more axes. Axes must consists of numeric data.

See Also:

- "Box Plot or Box Graph"
- "Supported Data Types for Graph Nodes"

5.4.1.1 Box Plot or Box Graph

A box plot graphically depicts groups of numeric data using the quantiles of the data. The bottom and top of the box are the first and third quartiles, and the band inside the box is the second quartile (the median). In the type of box plot created by Data Miner, the whiskers show the minimum and maximum values of all the data.

5.4.2 Supported Data Types for Graph Nodes

A Graph node supports the following data types:

- `NUMBER`
- `FLOAT`

Date and Time data types are also supported:

- DATE
- TIMESTAMP
- TIMESTAMP_WITH_TIMEZONE
- TIMESTAMP_WITH_LOCAL_TIMEZONE

5.4.3 Working with Graph Nodes

You can create a Graph node and perform additional tasks through the context menu.

- Create a Graph Node
- Graph Node Context Menu

5.4.3.1 Create a Graph Node

You can create a Graph node to visualize numeric data and relationships between numeric variables.

To create a Graph node:

1. In the **Components** pane, go to Workflow Editor.
   If the **Components** pane is not visible, then go to **View** and click **Components.**
2. In the **Workflow Editor,** expand **Data** and click **Graph.**
3. Drag and drop the Graph node from the **Components** pane to the **Workflow** pane. The Graph node is added to the workflow.
4. Right-click the node that contains the data for which you want to create the graph. Select **Connect** from the context menu.
5. Draw a line to the Graph node and click again.
6. To create a graph using sample data, then right-click the graph node and click **Run.**
7. Double-click the Graph node or right-click the node and select **Edit** from the context menu.
   - If no graphs are defined, then the **New Graph** dialog box opens. You can define a graph here.
   - If a graph is defined, then in the Graph Node Editor, you can add new graphs or edit existing graphs.
8. After running of the graph node is complete, the graph is automatically displayed in the **Graph Node Editor.** You can edit the graph. You can also define new graphs, modify settings and attributes of graphs, and delete graphs.

   See Also:

   - "New Graph"
   - "Graph Node Editor"

5.4.3.2 Graph Node Context Menu

To view the context menu, right-click a Graph node. The context menu provides the following options:

- Connect
- Edit. Opens the **Edit a Data Source Node** dialog box.

- Validate Parents
- Run. See "Running Graph Node" for more information.
- Force Run
- Save SQL: This option is disabled. It indicates that you cannot generate SQL query for this node.
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Validation Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed if there are validation errors.
- Go to Properties
- Navigate

5.4.3.2.1 Running Graph Node

To run the Graph node, right-click the Graph node and click **Run.** Run the Graph node to generate sample data. If you do not generate sample data, graphs are created using the data provided to the node.

> **Note:** If you import workflow that contains Graph nodes, you must run the Graph node to view the graphs.

5.4.3.2.2 Show Graph

The **Show Graph** option opens the **Graph Node Editor**. All graphs are displayed in the Graph Node Editor.

5.4.4 New Graph

The **New Graph** dialog box defines a graph with a default name. You can change the name of the graph here.

Select the type of graph to create and follow the steps to define that type of graph:

- Line or Scatter
- Bar
- Histogram
- Box

The graph that you defined is displayed in a container. After you have created a graph, you can edit the definition. You can also add graphs, delete graphs, view the data used to construct a graph, and save the graph as a graphic.

See Also:

- "Graph Node Editor"
- "Types of Graphs"
- "Supported Data Types for Graph Nodes"

5.4.4.1 Line or Scatter

To create a Line or Scatter:

1. Click **Line** to create a Line graph. The Line graph is the default type.
   To create scatter graph, click **Scatter.**
2. For a line graph or a scatter graph, enter the following details:
   - **Title:** This is the title of the graph. You can use the default name or you can provide a different name.
   - **Comment:** Type a description of the graph. This is an optional field.
   - For line graph settings, specify the following information:
     – **X-Axis:** Select an attribute for the x-axis of the graph.
     – **Y-Axis:** Select an attribute for the y-axis of the graph.
     – **Group By:** Click this option to select an optional **Group By** attribute. Use this option to create a series based on the Group By attribute values. Select an attribute for Group By. To specify how the grouping is performed, click **Settings.**
3. Click **OK.**

5.4.4.2 Bar

A bar graph plots the values of a selected attribute (x-axis) against frequency counts (y-Axis). To specify a bar graph:

1. Click **Bar.**
2. Specify the following information for Bar Graph settings:
   - **Title:** This is the name of the graph. You can use the default name or you can enter a different name.
   - **Comment:** Provide a description for the bar graph. This is an optional field.
   - Specify the following Bar Graph settings:
     – **X-Axis:** Select one attribute to be plotted along the x-axis of the graph. This attribute is handled using either Top N, or Binned (the default). To specify handling, click **Settings.**
     – **Y-Axis:** Select one attribute to be plotted along the y-axis of the graph. Specify the value for frequency count as statistic. For example, Average, Min, Max Median, Count. The statistic that you choose summarizes the values in the bin.
     – **Group By:** Click this option to select an optional Group By attribute. This attribute groups the values in the bin (stack the bar).

Select an attribute for Group By. To specify how the grouping is performed, click **Settings.**

3. Click **OK.**

See Also: "Settings (Graph Node)"

5.4.4.3 Histogram

A histogram plots the values of a selected attribute along the x-axis against frequency counts along the y-axis. To create a histogram:

1. Click **Histogram.**
2. Enter the following information for the histogram:
   - **Title:** This is the name of the histogram. You can use the default name or you can enter a different name.
   - **Comment:** Enter a description of the histogram. This is an optional field.
   - Specify the following settings for the Histogram:
     - **X-Axis:** Select an attribute to plot along the x-axis of the graph. This attribute is handled using either Top N, or Binned (the default). To specify handling, click **Settings.**
     - **Group By:** Click **Group By** to select an optional Group By attribute. This attribute groups values in the bin (stack the bar).
       To specify how Group By is performed, click **Settings.**
       By default, a histogram with a Group By attribute is displayed as a Stacked Bar Graph.
       Click , the Stacked Bar Graph icon, and select  the Dual Y Bar Graph icon.
       If you specify an invalid value, a message appears explaining the problem.
3. Click **OK.**

See Also: "Settings (Graph Node)"

5.4.4.4 Box

A box plot summarizes binned data of a selected attribute (X axis).

1. To specify a box plot or box graph, click **Box.**
2. Enter the following details of the Box Graph:
   - **Title:** This is the name of the Box Graph. You can use the default name or you can enter a different name.
   - **Comment:** Enter a description of the Box Graph. This is an optional field.
   - Specify the following settings for the box graph:
     - **X-Axis:** Select an attribute to be plotted along the x-axis of the graph. This attribute is binned using either Top N, or Binned (the default).
       To specify handling, click **Settings.**
     - **Group By:** Click **Group By** to select an optional group by attribute. This attribute provides values on the y-axis.
       To specify how the grouping is performed, click **Settings.**
       If you specify an invalid value, a message appears explaining the problem.
3. Click **OK.**

See Also:

- "Settings (Graph Node)"
- "Box Plot or Box Graph"

5.4.4.5 Settings (Graph Node)

Depending on the data type of the attribute selected, one of these dialog boxes is displayed when you click **Settings**:

- Select Values to Display
- Axis Treatment

5.4.4.5.1 Select Values to Display If you specify a categorical attribute for an axis or the Group By attribute, click **Settings** to display the **Select Values to Display** dialog box.

The values of the attribute are listed in the **Values** column. Select values in one of the following:

- **All**
- **None**
- **Default**

  By default, the most frequent values using Top N is selected.
- Select specific values by clicking the check box for the values.

You can search for values using the search box.

When you have finished, click **OK** to return to the definition of the graph.

5.4.4.5.2 Axis Treatment If you specify a numeric attribute for an axis or Group By attribute, click **Settings** to display the **Axis Treatment Settings** dialog box.

The options are:

- **Raw Values** (as is)
- **Binned automatically**: Use Equal width binning for the axis values. The default number of bins is 10. You can change this value.
  The default is to *not* show null values. To show null values, click the check box.

Click **OK** to return to the definition of the graph.

5.4.5 Graph Node Editor

If at least one graph is defined for a Graph node, double-click the Graph node to open the Graph Node Editor.
If no graphs are defined, then the **New Graph** dialog box opens.

The editor displays all defined graphs. Each graph is in a container. The graph containers are laid out in a grid.

You can modify an existing graph or add graphs to the node. You can also zoom in to view details of a graph.

These icons are applicable to all graphs in the node:

- To add a new graph, click to open the **New Graph** dialog box.
- To refresh the display, click .
- To select data used to create the graph, click **View**. The options are:

– **Actual Data**: Displays the default unless you have run the Graph node. In this case, the node creates the graph using whatever data is provided.

– **Sample Data**: Available only if you have run the Graph node to generate sample data.

The name of the new or existing graph is displayed at the top of the container for the graph, along with these controls:

- To adjust the size of the graph (zoom in and out), click .
- To edit the current graph, click . When you edit or add a graph, the results are automatically displayed in the editor.
- To delete the current graph, click .
- To examine specific values of the graph, zoom the graph.

See Also:

- "New Graph"
- "Zoom Graph"

5.4.5.1 Zoom Graph

To examine details of a graph, zoom in on the selected values. For example, to see more detail for selected x-axis values, draw a selection box with the mouse that encloses the values. The display shows values in the selection box in more detail and the axis expands. You can zoom in multiple times. When you have finished viewing the selected values, click the graph once for each time that you zoomed in to return to the original graph.

5.4.5.2 Viewing Data used to Create Graph

To view the data used to create the graph or to save the graph as a graphic:

1. Right-click the graph and select an one of the following options:
2. Select any one of the following options:
   - **Copy to Clipboard:** Copies the graph to the Microsoft Windows clipboard
   - **Save Image As:** Saves the graph in a PNG file format
   - **View Data:** Displays the data used to create the graph

5.4.6 Edit Graph

When you edit a graph, you can do the following:

- Change the attributes of the existing graph type. For example, if you edit a line graph, you can change the x-axis and y-axis or add a Group By attribute. You can change the axis treatment or values to display if the graph uses these items.
- Specify a different graph type. You can change the type of graph. For example, you can change a line graph to a histogram. To change the type of the graph, click the button at the top of the window and specify the required information:
  – Line or Scatter
  – Bar
  – Histogram
  – Box

5.4.7 Graph Node Properties

Graph node properties is identified by the node name that you are viewing. In the Properties pane, you can examine and change characteristics of a Graph node.

To open **Properties** pane, right-click the node and select **Go to Properties**.

The Properties pane of a Graph node consists of these sections:

- Cache (Graph Node)
- Details
- Data: Indicates the Cache Settings to use.

See Also: "Properties"

5.4.7.1 Cache (Graph Node)

After you generate sample data, the option **Generate Cache of Output Data to Optimize Viewing Results** is selected.

The default value is 2000 records. You can change this value.

5.5 SQL Query Node

A simple use case for SQL Query node is to write a SQL query that performs special data preparation. Use the SQL Query node to provide input for a model build. A SQL Query node enables you to do the following:

- Manually enter a SQL query with the least amount of constraints possible.
- Incorporate a broader array of methodologies as part of an Oracle Data Miner workflow. You can insert any SQL query as a source of data or as a transformation by using existing data.
- Run Oracle R Enterprise scripts that are registered for the database.
- Run parallel processes or parallel queries.

See Also:

- "Input for SQL Query Node"
- "SQL Query Restriction"
- "Oracle R Enterprise Script Support"
- "About Parallel Processing"

5.5.1 Input for SQL Query Node

A SQL Query node requires the following input:

- Zero to many data provider nodes, such as Data Source nodes and Transform nodes.
- Zero to many input model provider nodes, such as Model Build nodes and Model nodes.

Data provider nodes are used as follows:

- If there are no data provider nodes, then you define an originating Data Source node using an `SQL SELECT` statement that is not constrained by any input sources that must be defined within Oracle Data Miner. The statement can contain its own internal table references, such as:

```
Select * from a, b where a.id = b.id
```

When there are no data provider nodes, the source tables or views are hidden from Data Miner. Code generation cannot parameterize such tables in the generated SQL script.

- If there are one or more data provider nodes, then you can reference each data flow within the expression builder interface. You can continue to expose all data sources within the Oracle Data Miner workflow.

When a model provider node is connected as input, it enables you to see the list of model names contained in the node. This is useful in creating SQL that requires a model name.

Oracle Data Miner includes snippets that help you write SQL queries.

See Also: "Snippets in Oracle Data Miner"

5.5.2 SQL Query Restriction

SQL is limited to a query that returns a flow of data. A SQL Query node can provide data to any node that requires data, such as a node that builds a model.

5.5.3 Oracle R Enterprise Script Support

Oracle R Enterprise, a component of the Oracle Advanced Analytics option, makes the open source R statistical programming language and environment ready for the enterprise and big data. Oracle R Enterprise integrates R with Oracle Database. It is designed for problems that involve large amounts of data.

R users can develop, refine, and deploy R scripts that leverage the parallelism and scalability of the database to perform predictive analytics and data analysis.

The SQL Query node in Oracle Data Miner 4.1 provides a simplified interface for integrating R scripts that have been registered with the database. This enables R developers to provide useful scripts for analyzing data.

Note: Oracle R Enterprise must be installed with the server in the same database to which Oracle Data Miner connects to. See *Oracle R Enterprise Installation and Administration Guide* for more information.

You can run embedded R scripts that use theses interfaces:

- `rqEval`
- `rqTableEval`
- `rqRowEval`
- `rqGroupEval`

5.5.3.1 Oracle R Enterprise Database Roles

The Oracle R Enterprise database roles are added to the `OMDRUSER` role. The `OMDRUSER` role includes both these two roles:

- `RQUSER`

- RQADMIN

If the RQUSER and RQAMIN roles are not available in the database configuration at the time of Oracle Data Miner repository installation, then the roles must be added by the DBA manually to the ODMRUSER role after Oracle R Enterprise is installed.

You must register R scripts before you can use them. You can register scripts using SQL*Plus or SQL Worksheet, using the SYS connection.

Registered R Scripts are listed on the **R Scripts** tab of the SQL Query node. There are also R code snippets, which are code fragments that help you write scripts. Some snippets are just syntax, and others are examples. You can insert and edit snippets when you are using SQL Worksheet or creating or editing R code using SQL Query node.

5.5.4 Working with SQL Query Node

You can create a SQL Query node and also perform related tasks through the context menu and the SQL Query Node editor:

- Create a SQL Query Node
- SQL Query Node Editor
- SQL Query Node Context Menu

5.5.4.1 Create a SQL Query Node

You create a SQL Query node to write a SQL query. To create a SQL Query node:

1. In the **Components** pane, go to **Workflow Editor.**
   If the **Components** pane is not visible, then go to **View** and click **Components.**
2. Create zero or more data provider nodes, such as Data Source nodes or Transform nodes.
3. Create zero or more model provider nodes, such as Model Build nodes or Model nodes.
4. In the Workflow Editor, expand **Data,** and click **Create SQL Query.**
5. Drag and drop the SQL Query node in the workflow pane. The SQL Query node is added to the workflow.
6. Right-click the data provider and or model provider nodes. For each node, select **Connect** from the context menu.
7. Draw a line to the SQL Query node and click again. Ensure that you connect all the required nodes.
8. Open **SQL Query Node Editor** by double-clicking the SQL Query node or by selecting **Edit** from the context node.
9. Write the SQL query, and validate or preview it.
10. Click **OK.**

See Also: "SQL Query Node Editor"

5.5.4.2 SQL Query Node Editor

The SQL Query node Editor enables you to define and validate a SQL query.

The editor provides specifications that can be dragged and dropped or double-clicked into the query build text area.

The tabs on the left of the window provide help in writing a query:

- **Source** is a list of input nodes that includes both Data Provider nodes and Model Provider nodes. When you select an item in the Source list, information appears in the Message box.
- **Snippets** are standard SQL Developer Snippets (SQL code fragments) arranged by the type of calculation.

  The snippet category Predictive Query helps write queries using the `DBMS_PREDICTIVE_ANALYTICS PL/SQL` package.

  For information about snippets, see "Using Snippets to Insert Code Fragments" in *SQL Developer Concepts and Usage*.

  For a brief description of snippet functionality, move the cursor over the name. The information is displayed in a tool tip.
- **PL/SQL Functions** is a list of PL/SQL functions by user schema.
- **R Scripts** is a list of registered R Scripts; this tab is displayed if Oracle R Enterprise is installed.

Write SQL in the text area.

Below the text area are:

- **Columns**—List the columns and data types.
- **Preview**—Displays a small number of the rows returned by the query.

When you click **OK** or **Validate**, the following are verified:

- The query generates at least one column of output.
- All data types are data types that Oracle Data Miner supports. Most scalar data types are supported. The most common data types that are not supported are user custom object types.

  If unsupported data types are found, then a table is created and an error message is displayed above the column panel. All unacceptable data types in the column list are marked with an invalid icon next to the column name.
- If no validation errors occur during the parsing of the query, then the **Columns** tab shows data types of the columns and the Data tab shows a small sample of results.

  If a validation error occurs, then the validation panel is displayed for the **Column** and **Data** tabs.

5.5.4.3 SQL Query Node Context Menu

To see the context menu, right-click a Data Source node. The following options are available in the context menu:

- Connect
- Edit. Opens the SQL Query Node Editor.
- Validate Parents
- Run
- Force Run
- Save SQL
- Deploy

- Show Graph
- Generate Apply Chain
- Cut
- Cut
- Extended Paste
- Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed if there are validation errors.
- Go to Properties
- Navigate

5.5.5 SQL Query Node Properties

To view the **Properties** pane, right-click the node and select **Go to Properties.**

SQL Query node Properties contains these sections:

- SQL: Shows the SQL query. Click to edit the query.
- Output: Shows the columns and data types in the output of the query.
- Cache
- Details

5.6 Update Table Node

An Update Table node updates an existing table with selected columns from the data input to the node. Input columns are mapped to the existing table columns.

Update Table node can run in parallel.

Update Table nodes rely on database resources for their definitions. It may be necessary to refresh a node definition if the database resources change. For example, if the resources are deleted or re-created.

- Input and Output for Update Table Node
- Data Types for Update Table Node
- Working with Update Table Node
  - Create Update Table Node
  - Edit Update Table Node
  - Update Table Node Context Menu
  - Table (Update Table)
  - Columns (Update Table)

See Also:

- "Refresh Nodes"
- "About Parallel Processing"

5.6.1 Input and Output for Update Table Node

The input for an Update Table node is any node that produces a data flow. You can connect only one node to an Update Table node.

The output of an Update Table node is a data flow. You can use the data flow as a data source. To save the output as a table or view, use a Create Table or View node.

See Also: "Create Table or View Node"

5.6.2 Data Types for Update Table Node

Update Table node supports certain data types. You can also support other types, such as B, and also supports these additional types with manual mapping.

You can manually map columns that do not have exact data type matches but the column data types have reasonable, safe implicit conversion default. For example, you can map `BINARY_DOUBLE` to `NUMBER`, or `NVARCHAR2` to `VARCHAR2`. There could be some loss of data for such mappings:

- Mapping `BINARY_DOUBLE` or `BINARY_FLOAT` to `NUMBER` could result in the loss of precision
- Mapping `NVARCHAR2` and `NCHAR` to `VARCHAR2` can result in the loss of data because `NVARCHAR2` and `NCHAR` are based on a potentially different character set than `VARCHAR2`. For mappings to work, your database must be set up so that `NVARCHAR2` and `NCHAR` are based on the same character set as `VARCHAR2`.

See Also: "Supported Data Types for Data Source Nodes"

5.6.3 Working with Update Table Node

You can create an Update Table node and perform related tasks through the context menu and the editor, as listed in the following topics:

- Create Update Table Node
- Edit Update Table Node
- Update Table Node Context Menu
- Table (Update Table)
- Columns (Update Table)

5.6.3.1 Create Update Table Node

You create an Update Table node to update an existing table with data from selected columns in the table. You can connect an Update Table node to any node that creates a data flow, such as an Apply node.

To create an Update Table node:

1. In the **Components** pane, go to **Workflow Editor.** If the **Components** pane is not visible, then go to **View** and click **Components.**
2. In **Workflow Editor,** expand **Data,** and click **Update Table.**

3. Drag and drop the Update Table node in the workflow pane. This adds the Update Table node to the workflow.
4. Move the mouse to the node in the workflow that produces the data flow to update. Right-click the node, and select **Connect** from the context menu.
5. Draw a line to the **Update Table** node and click again.
6. The **Edit Update Table Node** dialog box opens. You can define the characteristics of the Update Table node.
7. You can do either of the following:
   - Accept the default settings for the Table Update node.
   - Edit the default settings in Edit Update Table node.
8. To update the table, right-click the Update Table node, and select **Run.**
9. After running of the Update Table node is complete, you can view the results. Right-click the node and select **View Data.**

   See Also:

   - "Update Table Node Automatic Behavior"
   - "Edit Update Table Node"

5.6.3.1.1 Update Table Node Automatic Behavior

If the **Auto Input Column Selection** option is selected, then the behavior under the following scenarios are:

- When input is connected—If the Update Table node has a selected table to update, then any column that matches the existing columns in the table are mapped to that column automatically. The node becomes valid, if at least one column was included in the specification.
- When input is disconnected—All columns are automatically removed. The node becomes invalid.
- When the input node is edited:
  – If a column is added to the input node, the column is added to the update node, if there is a matching column in the existing table.
  – If a column is removed, then the column is removed from the node.
  – If a column is edited, then there are two possibilities:
    * If the edited column no longer matches an existing column, then it is removed.
    * If the edited column matches an existing column, then it is added.

If **Auto Input Column Selection** is not selected, then you must manually add and remove column specifications from the Update Table node.

5.6.3.2 Edit Update Table Node

The **Edit Update Table** dialog box provides the specification for the Update Table node. To edit the Update Table node:

- **Name:** Displays the name of the table. You can select an existing table by clicking **Browse Table** and select an existing table. Or you can create a new table by clicking **New.** S

- **Auto Input Columns Selection:** This option is selected by default. To select other columns, deselect this option.
  If you deselect this option, then click to select an input column. Use the arrows to move attributes from **Available Attributes** to **Selected Attributes.**
- **Drop Existing Rows:** If this option is selected, then existing rows in the table are dropped before the table is updated. By default, the option is not selected.
- Columns are listed in the **Data** grid.

> **See Also:**
> - "Create Table (Update Table)"
> - "Update Table Node Automatic Behavior"
> - "Edit Columns (Update Table)"

5.6.3.2.1 Create Table (Update Table)

You can either accept the name for the new table or select a different name. All of the attributes of the attached table are listed.

- To delete an attribute, click .
- To edit an attribute, click .

The **Select Attributes (Update Table)** dialog box opens.

5.6.3.2.2 Select Attributes (Update Table)

By default, all columns are selected. If you do not want to include a column in the data, then move the attribute from **Selected Attributes** to **Available Attributes.**

Click **OK**.

5.6.3.2.3 Edit Columns (Update Table)

This tab displays the columns of the table to be updated. For each column, the Input Name, Target Name and Target (data) type are displayed.

By default, the option **Auto Input Columns Selection** is selected. If you deselect **Auto Input Columns Selection,** then you must manually select input columns by clicking .

Use the arrows to move the attributes from **Available Attributes** to **Selected Attributes.**

> **See Also:** "Update Table Node Automatic Behavior"

5.6.3.3 Update Table Node Context Menu

To see the context menu, right-click a Data Source node. The following options are available in the context menu:

- Connect
- Edit. Opens the **Edit Update Table Node** dialog box.
- Validate Parents
- Run
- View Data. Opens the **Update Table Node Data Viewer** dialog box.
- Show Graph
- Generate Apply Chain

- Force Run
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed only if there are validation errors.
- Go to Properties
- Navigate

5.6.4 Update Table Node Data Viewer

After the node runs successfully, you can view the data. To view the data, right-click the node and select **View Data.** The data viewer is the same as Data Source Node Viewer.

5.6.5 Update Table Node Properties

To view the **Properties** pane, right-click the node and select **Go to Properties** from the context menu.

Update Table Properties consists of these sections:

- Table (Update Table)
- Columns (Update Table)
- Cache
- Details

5.6.5.1 Table (Update Table)

This tab displays the name of the table to update. The default is to *not* drop existing rows. To drop existing rows, select **Drop Existing Rows.**

5.6.5.2 Columns (Update Table)

This tab displays the columns of the table that are to be updated. For each column, Input Name, Target Name and Target (data) type are displayed.

By default, the **Auto Input Columns Selection** is selected.

If you deselect the **Auto Input Columns Selection,** then you must manually select input columns by clicking . Use the arrows to move the attributes from **Available Attributes** list to **Selected Attributes** list.

If data on the server changes, then it may be necessary to refresh the nodes. To refresh, click .

See Also:

- "Refresh Nodes"
- "Update Table Node Automatic Behavior"

6

Using the Oracle Data Miner GUI

The Oracle Data Miner graphical user interface (GUI) is based on the GUI for SQL Developer 4.0. See *SQL Developer User Interface.* In SQL Developer, click **Help** and then click **SQL Developer Concepts and Usage** for an overview of the SQL Developer GUI. The following topics describe the common procedures and menus for the GUI that are specific to data mining and the Oracle Data Miner:

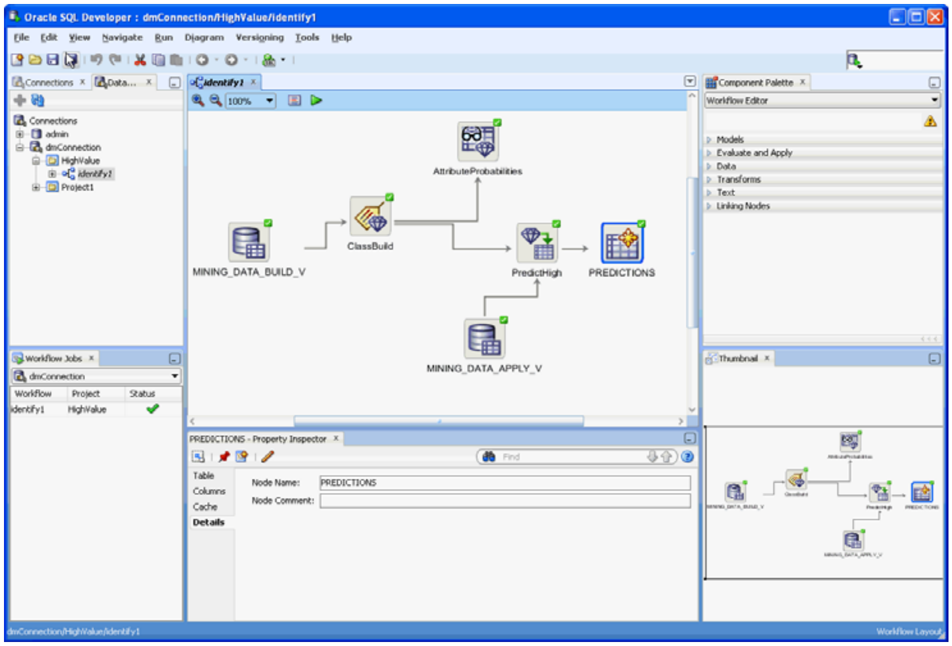
- Graphical User Interface Overview
- Oracle Data Miner Functionality in the Menu Bar
- Workflow Jobs
- Projects
- Import Data (Oracle Data Miner)
- Data Miner Preferences
- Miscellaneous (Search, Filtering, and Saving charts and graphs)

6.1 Graphical User Interface Overview

The Oracle Data Miner window generally uses the left side for navigation to find and select objects and the right side to display information about selected objects.

> **Note:** This text explains the default interface. However, you can customize many aspects of the appearance and behavior of Oracle Data Miner by setting the Oracle Data Miner preferences.

Here is a simple workflow:



Oracle Data Miner workflow from Getting Started cue card set in GUI.

****

- About the Data Miner Tab: Positioned in the left pane. You manage database connections here.
- Workflow Jobs: Positioned in the left pane. You can view the status for running tasks.
- Components: Positioned in the right pane. You select the nodes of the workflow here.
- Menu bar: Positioned at the top of the window.
- Workflows are displayed in the middle pane of the window.
- Properties: You can put this pane any place you want. It is useful to position the **Properties** pane just below the workflow that you are viewing.

Some settings are sticky settings. If you change a sticky setting, the new value becomes the default value.

See Also:

- "Data Miner Preferences"
- "Oracle Data Miner Functionality in the Menu Bar"

6.2 Oracle Data Miner Functionality in the Menu Bar

The menus at the top of the Oracle SQL Developer window contain standard entries, and entries for features specific to Oracle Data Miner.

You can use keyboard shortcuts to access menus and menu items. For example Alt+F for the File menu and Alt+E for the Edit menu, or Alt+H, then Alt+S for Full Text Search of the Help topics. You can also display the File menu by pressing the F10 key.

The icons just below the menus perform a variety of actions, including the following:

- **New:** Opens the New Gallery to define new database objects, such as a new database connection.
- **Open:** Opens a file.
- **Save:** Saves any changes to the currently selected object.
- **Save All:** Saves any changes to all open objects.
- **Undo:** Reverses the last operation.

  There are several forms of Undo, including Undo Create to undo the most recent create node, Undo Edit Node to undo the latest edit node, and so on.
- **Redo:** Does the latest undo operation again.

  There are several forms of Redo, including Redo Create to redo the most recent create node, Redo Edit Node to redo the latest edit node, and so forth.
- **Back:** Moves to the pane that you most recently visited. Use the drop-down arrow to specify the tab view.
- **Forward:** Moves to the pane after the current one in the list of visited panes. Or use the drop-down arrow to specify a tab view.

These menus contain functionality specific to Oracle Data Miner:

- View Menu
- Tools Menu
- Diagram Menu
- Help menu: Oracle Data Miner Online Help

6.2.1 View Menu

The following options are available in the View menu of Oracle Data Miner:

- Components
- Properties
- **Data Miner**

  Select one of these Data Miner interfaces to open it:

  – **Data Miner Connections**. Also known as **Data Miner** tab.

  – Workflow Jobs

  For example, the option **Data Miner Connections** under **Data Miner** in **View** menu option, opens the Data Miner tab and related windows.

  Depending on the state of the GUI, View Data Miner also enables you to:

  – **Make Visible**

  – **Drop Repository**

  See Also: "About the Data Miner Tab"

6.2.1.1 Structure Window

The **Structure** window shows the structure of a workflow or certain model viewers that are currently active.

The nodes in a tree or workflow are listed in a flat list, which does not show parent or child relationships. The links are the keys that tie the nodes together.

When you view nodes and links in the **Structure** window, the workflow editor reacts by immediately making the selected items visible. This property is useful when you are navigating a complex workflow or tree.

See Also:

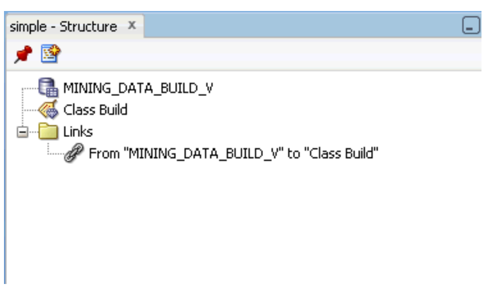
- "Structure Window and Workflow"
- "Structure Window and Model Viewers"
- "Structure Window Controls"

6.2.1.1.1 Structure Window and Workflow Suppose that you have a two-node workflow `simple` consisting of a Data Source node and a Classification Build node.

A two-node workflow consisting of a Data Source node and a Classification Build node.

****

The view of the workflow `simple` in the **Structure** window:



The **Structure** window of a two-node workflow that consists of a Data Source node and a Classification node.

****

If you select an item in the **Structure** window, the corresponding item is selected in the workflow. For example, if you go to the `Links` folder and select `From "DATA_MINING_BUILD_V" to "Class Build"`, the link between the two nodes in `simple` is highlighted.

6.2.1.1.2 Structure Window and Model Viewers These model viewers include a Tree tab:

- *k*-Means and O-Cluster model viewer for the Clustering models
- Decision Tree model viewer for the DT Classification model

The Tree tab of the model viewer illustrates how the rules generated by the model are related. These trees are sometime large and complex. You can use the Structure window to navigate among the nodes of the tree, just as you navigate among the nodes of a workflow. In the Structure window, select a node or an item in the Link folder. The node or link in the model viewer is automatically highlighted.

6.2.1.1.3 Structure Window Controls The Structure window supports these controls:

- To freeze the Structure window on the current view, click .

  A window that is frozen does not track the active selection in the active window.

- To open a new instance of the Structure window, click .

  The new view appears as a new tab in the Structure window.

6.2.2 Tools Menu

The following options are available in the Tools menu of Oracle Data Miner:

- Data Miner
- Data Miner Preferences

6.2.2.1 Data Miner

The following options are available in the **Data Miner** option under **Tools** menu:

- **Make Visible:** Opens the **Data Miner** tab and **Workflow Jobs.**
- **Drop Repository:** Drops the Data Miner repository.

These actions are usually performed as part of the Oracle Data Miner installation.

> **See Also:** *Oracle Data Miner Installation and Administration Guide*

6.2.2.2 Data Miner Preferences

You can set preferences for Oracle Data Miner. To set preferences, click **Tools** and then **Preferences.** Under **Preferences,** click **Data Miner.** Data Miner preferences are divided into several sets:

- Node Settings: Consists of settings for model nodes, parallel processing, and the Transforms node
- Viewers: Consists of settings for Data Viewers and Model Viewers
- Workflow Editor
- Workflow Import/Export directory
- Workflow Jobs

6.2.2.2.1 Node Settings Node settings specify the behavior of workflow nodes:

- Models
- Transforms

6.2.2.2.2 Models Model settings specify properties for:

- Apply
- Model Build
- Model Details
- Test
- Text

6.2.2.2.3 Apply These preferences specify how Apply nodes operate.

Automatic Apply Settings Default is either Automatic or Manual. By default, the settings are set to Automatic.

> **See Also:** "Automatic Settings"

Automatic Data Settings Default is either Automatic, the default, or Manual.

Default Column Order is either Data Columns First or Apply Columns First. Default Columns First is set as the default.

6.2.2.2.4 Model Build

Specify default values for model build options. Defaults depend on the mining function:

- Association
- Classification
- Clustering
- Feature
- Regression

6.2.2.2.5 Association

The default maximum distinct count for item values is 10. Change the default to a different integer.

6.2.2.2.6 Classification

By default, a Classification node automatically generates four models, one each using:

- Decision Tree
- General Linear Model
- Naive Bayes
- Support Vector Machine

All four models have the same input data, the same target, and the same case ID (if a case ID is specified).

If you do not want to build models using one of the default algorithms, then deselect that algorithm. You can still add models using the deselected algorithm to a Classification node.

By default, the node generates these test results for tuning:

- Performance Metrics
- Performance Matrix (Confusion Matrix)
- ROC Curve (Binary only)
- Lift and Profit. The default is set to the top 5 target values by frequency. You can edit the default setting. By default, the node does not generate selected metrics for Model tuning. You can select the metrics for Model tuning.

You can deselect any of the test results. For example, if you deselect Performance Matrix, a Performance Matrix is not generated by default.

By default, split data is used for test data. Forty percent of the data is used for testing, and the split data is created as a table. You can change the percentage used for testing and you can create the split data as a view instead of a table. If you create a table, then you can create it in parallel. You can use all of the build data for testing, or you can use a separate test source.

See Also:

- "Testing Classification Models"
- "Tuning Classification Models"

6.2.2.2.7 Clustering

By default, a Clustering node builds two models, one each using O-Cluster and *k*-Means.

If you are connected to Oracle Database 12*c* Release 1 (12.1) or later, then a Clustering node also builds an Expectation Maximization model, so that three models are built.

All Clustering models in the node have the same input data and the same Case ID, if one is specified.

If you do not want to build models using one of the algorithms by default, then deselect that algorithm. A user will still be able to add models using the deselected algorithm to a Clustering node.

6.2.2.2.8 Feature

By default, a Feature Extraction node builds a Nonnegative Matrix Factorization model.

If you are connected to Oracle Database 12*c* Release 1 (12.1) or later, then the node also builds a Principal Component Analysis model. You can specify that the node builds a Singular Value Decomposition model.

If you do not want to build models using one of the default algorithms, then deselect that algorithm. You can still add models using the deselected algorithm to a Classification node.

6.2.2.2.9 Regression

By default, a Regression node builds two models, one each using General Linear Model and Support Vector Machine.

All models in the node have the same input data, target, and case ID, if a case ID is specified.

If you do not want to build models using one of the default algorithms, then deselect that algorithm. You can still add models using the deselected algorithm to a Regression node.

By default, two test results, Performance Metrics and Residuals, are calculated.

By default, split data is used for the test data. The split is 40 percent and the split data is created as a view.

6.2.2.2.10 Model Details

By default, a Model Details node uses automatic settings. You can deselect automatic settings for a specific model details node, or you can deselect automatic settings for all model details nodes.

6.2.2.2.11 Test

By default, a Test node uses Automatic Setting. You can:

- Deselect Automatic Settings for a particular Test node.
- Deselect Automatic Settings for all Test nodes.

6.2.2.2.12 Parallel Processing

You can specify default settings for parallel processing for nodes.

- To turn on parallel processing:
  - Click the corresponding check box in the **Parallel** column for the node type.
  - Click **All** to set parallel processing for all node types.
  - Click **None** to set parallel processing for none of the node types.
- To specify degrees of parallelism for a node type:
  - Select the node type and click . The default degree of parallelism is system determined.
  - To specify a value, select **Degree Value.** The Default value is 1. The specified value is displayed in the **Degree of Parallel** column for the node type.

If you specify parallel processing for a node type, then the query generated by the node may not run in parallel.

You can override the preferences for a specific node.

See Also: "About Parallel Processing"

6.2.2.2.13 Text

These values describe how Text is handled during a model build.

The default categorical cutoff value is `200`

The default transformation is `Token`.

Token transformation settings use these defaults:

- **Language:** English with stemming (deselected)
- **Stoplist:** Default Stoplist
- **Maximum number of tokens:** 3000

Theme transformations use these defaults:

- **Language:** English with stemming (deselected)
- **Stoplist:** Default Stoplist
- **Maximum Number of themes across all documents:** 3000

6.2.2.2.14 Transforms

The preference that applies to all transformations is:

- **Generate Cache Sample Table to Optimize Viewing:**
  The default setting is to *not* generate a cache sample table. Generating a sample cache table is useful if you are processing large amounts of data.

For individual transformations, the options are:

- Filter Columns
- Filter Columns Details

- Join
- Sampling

6.2.2.2.15 Filter Columns

The preferences specify the behavior of the Filter Columns transformation. You can specify the following **Data Quality** criteria:

- **% Nulls less than or equal:** Indicates the largest acceptable percentage of `NULL` values in a column of the data source. The default value is `95%`.
- **% Unique less than or equal:** Indicates the largest acceptable percentage of values that are unique to a column of the data source. The default value is `95%`.
- **% Constant less than or equal:** Indicates the largest acceptable percentage of constant values in a column of the data source.

Attribute Importance is not selected by default. You can specify these settings:

- **Importance Cutoff:** A number between 0 and 1.0. The default cutoff is 0.
- **Top N:** The maximum number of attributes. The default is 100.

The column filter by default uses sampling to determine data quality and attribute importance. The default is to use a sample size of 2000 records. You can turn off sampling, that is use all of the data, or increase the sample size.

6.2.2.2.16 Filter Columns Details

The default setting is to select Automatic Settings.

> **See Also:** "Filter Columns Details"

6.2.2.2.17 Join

These preferences control how the Join transformation works:

- **Automatic Key Column Deletion:**
  - Automatic
  - Manual
  - Default
- **Automatic Data Column Default:**
  - Automatic
  - Manual
  - Default

6.2.2.2.18 Sampling

There are two preferences for Sampling:

- **Sampling Type**: By default, the Sampling Type is `Random` with the Seed 12,345. You can change the value of Seed.

  Sampling Type can be changed to `TopN`.
- **Sampling Size**: This is either the number of rows or the percentage for the sampling size. The default size is either 2000 rows or 60%.

6.2.2.2.19 Viewers

Viewers settings specify the behavior of:

- Data Viewers
- Model Viewers

6.2.2.2.20 Data

There are preferences for:

- Explore Data Viewer
- Graphical Settings

6.2.2.2.21 Explore Data Viewer

Precision is the maximum number of significant decimal digits, where the most significant digit is the left-most nonzero digit, and the least significant digit is the right-most known digit. Precision is an integer greater than or equal to zero.

These preferences specify data precision for data viewed in the Explore Data node.

The default precision for both percentage based values and numeric values is 4. You can change either or both of these values.

6.2.2.2.22 Graphical Settings

Specify preferences for **Depth Radius** and **Chart Style.**

The default Depth Radius is 0.

The default Chart Style is Nautical.

6.2.2.2.23 Model

These general preferences apply to all Model viewers.

- **Precision Level Settings** specify precision:
  - For Percentage Based Values, the default precision is 4.
  - For Numerical Values, the precision is 8.
- **Fetch Size Settings** specify the number of items fetched. The default fetch size for:
  - Association Rule Model: 1000
  - Clustering Rules Model: 20
  - All Other Models: 10,000

There are additional preferences for the tree displays:

- Cluster Tree
- Decision Tree

> **See Also:** "Explore Data Viewer" for more information about precision definition.

6.2.2.2.24 Cluster Tree

- Default Node display: A detailed header
- Default Layout: Vertical

There are also settings for Cluster Tree Node.

6.2.2.2.25 Tree Node

For each Tree Node in a cluster tree:

- The maximum display length for the **Centroid Attribute Name** is 25
- The maximum display length for **Centroid Value** is 25

6.2.2.2.26 Decision Tree

- **Default Node Display:** Histogram and Detailed Header.
- **Sort Target Values By:** Root Node order. You can also sort by Confidence.
- **Default layout:** Vertical. You can also choose Horizontal.

There are also settings for Decision Tree Tree Node.

6.2.2.2.27 Tree Node

For each Tree Node in a Decision tree, the length of the target value is 25 characters.

6.2.2.2.28 Workflow Editor

In the **Workflow Editor,** the following are the options and their default settings:

- **Node Assist:** Selected by default. Wizards are automatically displayed when a node is created or connected. For example, if you add a Data Source node to a workflow, then the Data Source Editor is automatically opened. You can deselect this option.
- **Link Style:** Direct. In the Direct link style, links are straight lines from one node to another with short segments. Direct link style produces a more compact, direct diagram layout. You can select Orthogonal link style.
- **Alternate Link Routing:** Deselected by default.

6.2.2.2.29 Workflow Import/Export Workflow Directory is the default directory to import workflows from and to export works. The default value for Workflow Directory is the default for the operating system where Oracle Data Miner is installed. For example, Workflow Directory is `My Documents` for Microsoft Windows operating systems.

To change the Workflow Directory, enter the name of the new directory or click **Browse** to browse for the directory. After you specify the directory name, click **OK**.

6.2.2.2.30 Workflow Jobs

Preferences for the Workflow Jobs specify which connection is displayed and how long the workflow status is displayed:

- **Automatically Display Connection Selected in Navigator:** By default, this option is selected. You can deselect it. If you deselect this option, then you must explicitly select a connection.
- **Don't Display Jobs Older Than 24 Hours:** By default, this option is selected. You can change this option.

6.2.3 Diagram Menu

The **Diagram** menu is available when a workflow is open. Use the options on this menu to arrange workflow nodes.

The **Diagram** menu has these options:

- Connect
- Align
- Distribute
- Bring to Front
- Send to Back
- Zoom
- Publish Diagram

6.2.3.1 Connect

Use this option to select a node in the workflow.

- To connect a node: Select **Diagram** and click **Connect.** Link the selected node to another node by drawing a line from the selected node. You can only make valid selections.
- To cancel a connection: To cancel a line that is not connected to anything, press Esc. This is the same operation as using the **Connect** option from the node context menu.

6.2.3.2 Align

You can use this option to align a set of elements and normalize the size of elements.

Horizontal Alignment

None (Default): Performs no horizontal alignment.
Top: Aligns the top edges of the selected elements.
Middle: Aligns the horizontal centers of the selected elements.
Bottom: Aligns the bottom edges of the selected elements.

Vertical Alignment

None (Default): Performs no vertical alignment.
Left: Aligns the left edges of the selected elements.
Middle: Aligns the vertical centers of the selected elements.
Right: Aligns right edges of the selected elements.

Size Adjustments

Same Width: Changes the width of the selected elements to the average width of all the selected elements.
Same Height: Changes the height of the selected elements to the average height of all the selected elements.

6.2.3.3 Distribute

You can use this option to evenly distribute (horizontally and vertically) selected elements in a diagram.

- **Horizontal Distribution:** Changes the left or right distribution of the selected diagram elements.
- **Vertical Distribution:** Changes the up or down distribution of the selected diagram objects.

6.2.3.4 Bring to Front

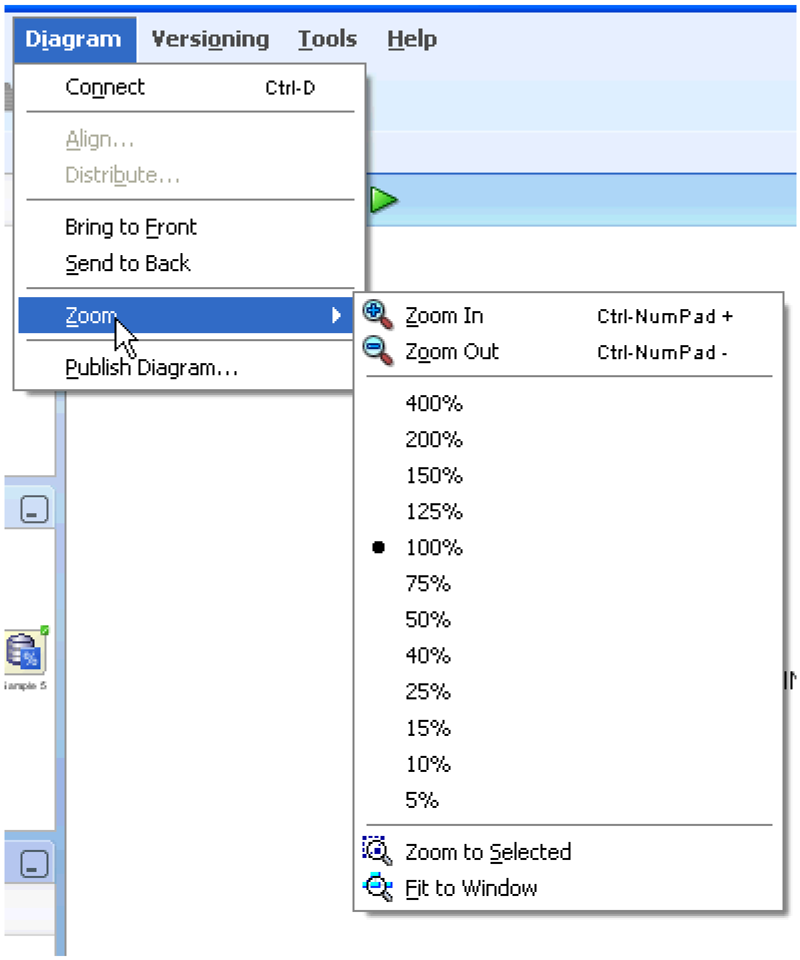
You can use this option to move the selected node in front of all other nodes.

6.2.3.5 Send to Back

You can use this option to move the selected node behind all other nodes.

6.2.3.6 Zoom

To zoom the view, go to **Diagram** and click **Zoom**. This menu is displayed:



The zoom setting under the **Diagram** menu. The default zoom setting is 100 percent.

****

The default zoom setting is 100%. To return to the default, select 100%.

- To zoom in or zoom out of an entire workflow: Click  and  respectively, or set a specific percentage.
- To zoom in on a specific node or nodes: Select the node and then go to **Diagram** and click **Zoom.**
- To fit the entire workflow in the window: Select **Fit to Window.**

6.2.3.7 Publish Diagram

You can use this option to save the current workflow as a graphic file in your system.

To publish a diagram:

1. Open the workflow that you want to save as a graphic file.
2. Go to **Diagram** and select **Publish.**
3. Browse for a location to save the graphic file.
4. Select the graphic file format. The supported file formats are:
   - PNG (default)
   - SVG
   - SVGZ
   - JPEG
5. Click **OK.** The workflow is saved as a graphic file.

6.3 Workflow Jobs

Workflow Jobs displays all running and recently run tasks, arranged according to connection. A task consists of running of a selected node and any ancestor nodes that must be run before running the node.

The following conditions apply for Workflow Job display:

- Workflow Jobs displays the most recent run of a workflow.
- When two or more tasks are active, the Workflow Jobs window is automatically displayed.
- By default, the Workflow Jobs automatically displays the connection selected in the tab. To view tasks in a different connection, select a connection from the list at the top of the tab.
- Workflow Jobs specifies for how long a task is displayed.

You an perform multiple tasks from the Workflow Jobs context menu. Right-click a line in the grid or below the lines in the grid.

> **See Also:** "Workflow Jobs Context Menu"

6.3.1 Viewing Workflow Jobs

To view Workflow Jobs:

1. In the SQL Developer menu bar, go to **View** and click **Data Miner.**
   If the **Data Miner** tab is not visible, then go to **Tools** and click **Data Miner.** Under **Data Miner,** click **Make Visible.**
2. Under **Data Miner,** click **Workflow Jobs.**

6.3.2 Working with Workflow Jobs

You can perform the following tasks with Workflow Jobs:

- View a particular task: Select the connection in which the task is running. Connections are listed in a drop-down list just above the Workflow Jobs Grid.
- View an Event Log
- Terminate a running workflow: Click .
- View a log: Right-click a process in the Workflow Jobs and click **View Log.**

6.3.3 Workflow Jobs Grid

The Workflow Jobs grid displays the following:

- Workflow name
- Project name
- Status. The values are:
  - ACTIVE: Indicates that the workflow has been executed. Indicated by .
  - INACTIVE: Indicates that the workflow is idle.
  - STOPPED: Indicates that the workflow has been stopped.
  - SCHEDULED: Indicates that the workflow is scheduled to run.
  - FAILED: Indicates that the workflow execution has failed. Indicated by .

6.3.4 View an Event Log

To view a log of data mining events in the selected connection, right-click an entry in the workflow jobs and select **Show Event Log**. You can also click at the top of a workflow, just under the tab.

By default, all errors are displayed. You can display errors or informational events. The total number of events and the number of events displayed is at the top of the list. For example, `Events: 2 of 90` means that 2 of the 90 events are displayed.

Each error or warning has a message and details associated with it. Select an event. The message and details are displayed in the lower pane of the Event Log window.

For each data mining event in the selected connection, the following are displayed:

- **Event**: In Oracle Data Miner, events indicate the beginning and end of actions, such as `START(WORKFLOW)` and `END(WORKFLOW)`. Each node is processed sequentially.
- **Job**: The name of the job that processes the event. These jobs are internal to Oracle Data Miner.
- **Node**: The name of the node that is being processed. Not all events are associated with a node.
- **Sub-node**: An internal step during node processing. For example, an Anomaly Build node has a sub-node that builds the model.
- **Time**: Start time of the event.
- **Message**: A message about the event. If the event did not encounter problems, then there is no message.

  To see more information about the message, including message details, select the event. The message and the details are displayed in the pane below the list of events. Not all events have messages or message details associated with them.
- **Duration**: The amount of time that is spent on processing the event. The duration is displayed for `END` events. The duration is displayed in days, hours, minutes and seconds.

All errors are shown. You can select the type of event to display by clicking the icons above the list of events:

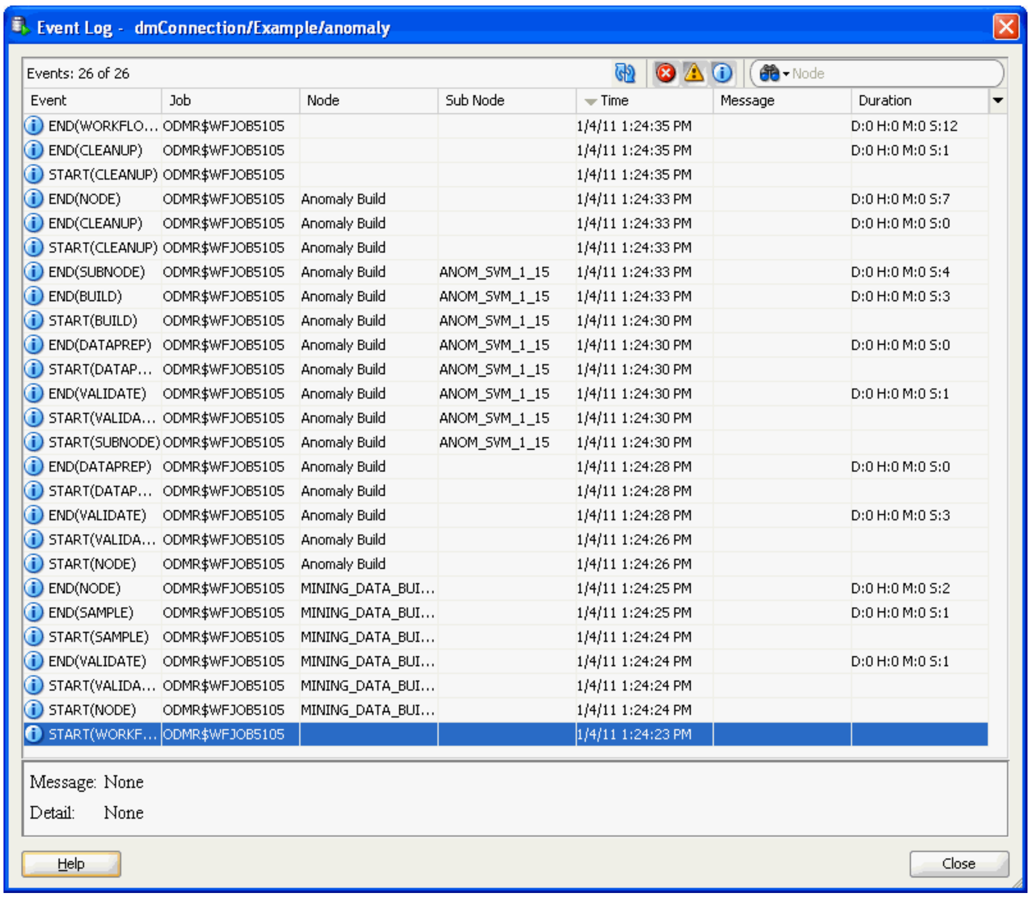
- To display errors, and events that failed, click .

  The default is to display errors only.

- To display warnings, click .
- To display Informational Messages, such as the start and end of operations, click .
- To refresh, click .
- To search for events, click the down arrow next to . You can search by `Node` (default) or by `Any` to search for anything that is not a node.

6.3.4.1 Informational Messages

Informational messages show the beginning and end of internal operations performed when a workflow runs. The following informational messages describe a successful build of an Anomaly Detection Model. The workflow took 5 minutes and 12 seconds to run. There were 26 events.



A sample informational message about a successful build of an Anomaly Detection Model. The workflow took 5 minutes and 12 seconds to run. There were 26 events. The message shows the beginning and end of the internal operations performed when the workflow is run.

****

6.3.5 Workflow Jobs Context Menu

To view the context menu, right-click a line in the Workflow Jobs grid or in a blank area of the grid. The context menu options are:

- Go to Workflow: Displays the workflow for this result.

- View an Event Log
- Sort by Most Recent: Sorts the entries.
- Preferences: Displays the Workflow Jobs.

6.4 Projects

Projects reside in a connection. Projects contain all the workflows created in the data mining process.

You must create at least one project in a connection.

See Also: "Data Miner Projects"

6.5 Miscellaneous

This section describes several common controls and tasks:

- Filter: Describes general information about filters.
- Import Data (Oracle Data Miner)
- Filter Out Objects Associated with Oracle Data Mining
- Copy Charts, Graphs, Grids, and Rules

6.5.1 Filter

To filter your search:

- To display only those items that you are interested in, click . In other words, you can filter out items that you are not interested in, by using this option. There are several filter options and a default option. To select a different filter option, click the triangle beside the binoculars icon.
- To clear the search, click .

6.5.2 Import Data (Oracle Data Miner)

Use the SQL Developer Import Data wizard to import data into a database table. You can create a table and import data into it, or you can import data into an existing table. In other words, you can import delimited data in an operating system file to the database.

To import data:

1. In SQL Developer, go to **Connections.**
2. Expand the connection that you are using for data mining.
3. Right-click the **Tables** node or a table name and select **Import Data.**
4. Specify the files from which to import the data. The file can be one of the following formats: XLS, TXT, CSV.
5. Click **OK.**

6.5.3 Filter Out Objects Associated with Oracle Data Mining

The Filter option filters out Oracle Data Miner (DR) objects and Oracle Data Mining (DM) objects. To filter the tables associated with data mining:

1. In SQL Developer, go to **Connections**.
2. Expand the user account that you use for data mining and select **Tables.**
3. In the tool bar for **Connections,** click  .
4. Select these filter criteria:

```
NAME NOT LIKE DR$%
NAME NOT KLIKE DM$%
NAME NOT LIKE ODMR$%
```

5. Click **OK** to filter the tables.

6.5.4 Copy Charts, Graphs, Grids, and Rules

You can export charts, graphs, grids, and also Cluster and Decision Tree Rules for use in external documents by copying them to the Microsoft Windows clipboard or saving them to a file. You can:

- Copy charts to the clipboard or save to a file: To copy charts, right-click the chart and select any one of the following options in the content-menu:
  - **Copy to Clipboard**
  - **Save Image As**
  - **View Data**
- Copy data grids to the clipboard: You can copy one or multiple rows in the **Graph Data** dialog box. To copy data grids:
  1. Select the row that you want to copy in the **Graph Data** dialog box, while pressing the Ctrl key. To select multiple rows, click the rows while pressing the Ctrl key and Shift keys together.
  2. Then copy the selected rows by pressing the Ctrl + C keys.
  3. To paste the copied rows in the clipboard, press Ctrl + V keys.
- Copy and view data content of charts: To copy data content of charts, right-click and select any one of the following options in the content menu:
  - **Copy Image to Clipboard**
  - **Save Image As**
- Copy Cluster and Decision Tree Rules to the clipboard or to a file: To copy Cluster and Decision Tree Rules to a clipboard or a file, use the **Save Rules** option in the Model viewer.

7 Transforms Nodes

A Transforms node performs one or more transformations on the table or tables identified in a Data node. Transformations are available in the **Transforms** section of the Components pane. The Evaluate and Apply Data nodes must be prepared in the same way that build data was prepared.

Transform nodes include s:

- Aggregation
- Data Viewer
- Expression Builder
- Filter Columns
- Filter Columns Details
- Filter Rows
- Join
- JSON Query
- Sample
- Transform

7.1 Aggregation

Aggregation is the process of consolidating multiple values into a single value. For example, you could aggregate sales in several states into sales for a region that consists of several states.

To preform aggregation use an Aggregate node. These topics describe Aggregate nodes:

- Creating Aggregate Nodes
- Editing Aggregate Nodes
- Aggregate Node Properties
- Aggregate Node Context Menu

See Also: "About Parallel Processing"

7.1.1 Creating Aggregate Nodes

To define an aggregation, you must identify a Data Source node and columns to aggregate.

1. Identify or create the node to aggregate. The node can be any node that provides a data flow, including Data Source nodes.
2. If the **Components** window is not open, then go to **View** and click **Components**. Expand the **Transforms** section.
3. Click **Aggregate.** Move the cursor to the workflow and click again.
4. Connect the Data Source node to the Aggregate node:
   a. Right-click the Data Source node and click **Connect.**
   b. Draw a line to the Aggregate node and click again.
5. Right-click the Aggregate node and click **Edit**.
6. Right-click the Aggregate node and click **Run**. Monitor the running of the node in the Workflow Jobs. If the Workflow Jobs is not open, then go to **View** and click **Data Miner.** Under **Data Miner,** click **Workflow Jobs**.
7. After the running of the node is complete, right-click the Aggregate node and select **View Data** to view the results of the aggregation.

See Also: "Editing Aggregate Nodes"

7.1.2 Editing Aggregate Nodes

You can define and edit aggregation elements of an Aggregate node in the **Edit Aggregate Node** dialog box.

To edit an Aggregate node:

1. Double-click or right-click the node and click **Edit.**
2. To select the Group By columns or Group By expression, click **Edit.** The **Edit Group By** dialog box opens.
3. You can define the following:
   - To use the Aggregation Wizard, click . The **Define Aggregation** wizard opens. You can add aggregations one by one.
   - To edit an already defined aggregation column, select the aggregation element and click . The **Edit Aggregate Element** dialog box opens.
   - To delete aggregation columns, click .
   - To add aggregation columns, click . The **Add Column Aggregation** dialog box opens.
   - To add custom aggregations (an expression), click . The **Add Custom Aggregation** dialog box opens.
4. When defining the aggregations is complete, click **OK.**

See Also:

- "Define Aggregation"
- "Edit Aggregate Element"
- "Add Column Aggregation"
- "Add Custom Aggregation"
- "Edit Group By"

7.1.2.1 Edit Group By

Default Type: `Column`
You can change the Type to: `Expression`

- If Type is `Column,` then select one or more columns in the **Available Attributes** list. You can search the list by name or by data type. Move the selected columns to the **Selected Attributes** list using the arrows.
- If the Type is `Expression,` then type an appropriate expression in the Expression box.
  Click **Validate** to validate the expression.

After you are done, click **OK.**

7.1.2.2 Define Aggregation

You can define aggregations using the **Define Aggregations** wizard.

To define aggregations:

1. Define the **Function** to use for aggregation. Available functions depend on the data type of the column that you are aggregating.
   For example, you can select `SUM` if you plan to aggregate one or more columns of numeric values.

   Click **Next**.
2. Select one or more **Columns** to aggregate. You must select columns with a data type compatible with the function that you selected.
   For example, if the function is `SUM,` you must select columns that have numeric data types.

   Click **Next**.
3. Optionally, select a **Sub-Group By** column for aggregation. Specifying a Sub-Group By column creates a nested table.
   For example, you could use sub-group by to calculate amount sold per product per customer. The nested table contains columns with data type `DM_NESTED_NUMERICALS`.

   You can select a **Sub Group By** expression by changing `Type` to `Expression.` If you define an Expression, click **Validate** to validate the expression.

   Click **Next**.
4. Review the default names for the columns. You can change the names.
5. Review the definitions if necessary. You can click **Back** to make changes.
6. After you are done, click **Finish.**

7.1.2.3 Edit Aggregate Element

You can define or modify the individual elements of an aggregation. To define or modify the individual element:

1. In **Output,** you can provide a name. To provide a name, deselect **Auto Generate** and enter a name. By default, Auto Generate is selected.
   Output is the name of the column that holds the results of the aggregation.
2. Select or change the Column that is being aggregated.
3. Select the function to apply to the column. The functions available depend on the data type of the column.

4. Click **Edit** to define a new Sub-Group By column. The **Edit Group By** dialog box opens.
5. Once done, click **OK.**

> **See Also:** "Define Aggregation" for more information about Sub-Group By.

7.1.2.4 Add Column Aggregation

You can define how a column is aggregated.

To add an attribute:

1. Click .
2. To provide a name, deselect **Auto Generate** and type in the name. By default, **Auto Generate** is selected.
   Output is the name of the column that holds the results of the aggregation.
3. Select **Columns** to aggregate from the list.
4. Select the **Function** to apply to the column. The functions available depend on the data type of the column. For example, you can specify average (AVG) for a numeric value.
5. To define a Sub Group By column, click **Edit.** The **Edit Group By** dialog box opens. It is not necessary to define a Sub Group By column.
6. After you are done, click **OK.**

> **See Also:**
>
> - "Add Custom Aggregation" for information about how to create a custom aggregation.
> - "Define Aggregation" for more information about Sub-Group By.
> - "Edit Group By"

7.1.2.5 Add Custom Aggregation

To add a custom aggregation, click and follow these steps:

1. **Output** is the name of column that holds the results of the aggregation. Specify a name.
2. **Expression** is the expression to add. To define an expression, click to open Expression Builder.

   This expression calculates all products bought by a customer and casts the result to a nested data type:

   ```
   CAST (COLLECT (TO_CHAR (PROD_ID)) AS ODMR_NESTED_VARCHAR2)
   ```

3. To define a **Sub Group By** column, click **Edit**. The **Edit Group By** dialog box opens.
   It is not necessary to define a **Sub Group By** column.
4. Click **Validate** to validate the expression.
5. After you are done, click **OK**.

See Also:

- "Define Aggregation"
- "Edit Group By"

7.1.3 Aggregate Node Properties

To view the **Properties** pane, select the node. If the **Properties** pane is closed, then go to **View** and click **Properties.** Alternately, right-click the node and select **Go to Properties** from the context menu.

The Aggregate Node **Properties** pane has these sections:

- Columns, the columns for the aggregation.
- Cache
- Details

See Also: "Editing Aggregate Nodes" for more information about how to edit the columns.

7.1.3.1 Cache

The default setting is to *not* generate the cache to optimize the viewing of results.

You can generate the cache. If you generate the cache, then specify the sampling size as either:

- Number of rows. The default is 2000 rows
- Percent. The default is 60 percent

7.1.4 Aggregate Node Context Menu

The Aggregate node context menu contains these entries:

- Connect
- Edit. See "Editing Aggregate Nodes" for more information.
- Validate Parents
- Run
- View Data. See "Data Viewer" for more information.
- Show Graph
- Force Run
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Parallel Query. See "About Parallel Processing" for more information.

- Select All
- Show Event Log
- Show Runtime Errors. Displayed only if there are errors.
- Show Validation Errors. Displayed if there are validation errors.
- Navigate

7.2 Data Viewer

You can view data when the Transform node is in a valid state.

To view data, right-click the node and select **View Data** from the context menu. The data viewer opens.

The data viewer has these tabs:

- Data
- Columns
- SQL

7.2.1 Data

The Data tab displays a sample of the data. The Data Viewer provides grid display of rows of data either from the sampling defined in the cache or pulled through the lineage of nodes back to the source tables.

The display is controlled with:

- **Refresh:** To refresh the display, click .
- **View:** Enables you to select either **Cached Data** or **Actual Data.**
- **Sort:** Displays the **Select Column to Sort By** dialog box.
- **Filter:** Enables you to type in a `WHERE` clause to select data.

7.2.1.1 Select Column to Sort By

The **Select Column to Sort By** dialog box enables you to:

- Select multiple columns to sort.
- Determine the column ordering.
- Determine ascending or descending order by column.
- Specify `Nulls First` so that null values appear ahead of real data values.

The sort order is preserved until you clear it.

Column headers are also sort-enabled to provide a temporary override to the sort selection.

7.2.2 Graph

The **Graph** tab enables you to create graphs from numeric data.

See Also: "Graph Node"

7.2.3 Columns

The Column tab is a list of all the columns that are output from the node. The display in the tab depends on the following conditions:

- If the node has not run, then the Table or View structure provided by the database is displayed.
- If the node has run successfully, then the structure of the sampled table is displayed. This is based on the sampling defined when the node was specified.

For each column, the following are displayed:

- Name
- Data type
- Mining type
- Length
- Precision
- Scale (for floating point)
- Column ID

There are several filtering options that limit the columns displayed. The filter settings with the `(or)/(and)` suffixes allow you to enter multiple strings separated by spaces. For example, if the `Name/Data Type/Mining Type(or)` is selected, then the filter string `A B` produces all columns where the name or the data type or the mining type starts with the letter A or B.

7.2.4 SQL

In the **SQL** tab, the SQL Details text area displays the SQL code that generated the data provided by the actual view displayed in the **Data** tab.

The SQL can be a stacked expression that includes SQL from parent nodes, depending on what lineage is required to access the actual data.

You can perform the following tasks:

- Copy and run the SQL query in a suitable SQL interface. The following options are enabled:
  - **Select All** (Ctrl+A)
  - **Copy** (Ctrl+C)
- Search texts. The Search control is a standard search control that highlights matching text and searches forward and backward.

7.3 Expression Builder

Expression Builder helps you to enter and validate SQL expressions, such as constraints for filters. An expression is a SQL statement or clause that transforms data or specifies a restriction. Expression Builder displays the available columns, provides a selection of functions and commonly used operators, and validates expressions.

To build and validate expressions in Expression Builder:

1. Click  in the **Add Custom Transform** dialog box. The **Expression Builder** dialog box opens.

2. The **Expression Builder** dialog box has the following components:

   - **Attributes** tab: Lists attributes (columns) in the source data. To insert an attribute into the query that you are creating in the Expression box, or to replace characters that you selected, double-click the attribute at the current character position.
   - **Functions** tab: Lists the commonly used SQL Functions, divided into folders. Double-click a folder to view the functions listed there. To insert a function into the expression at the current character position or to replace characters that you selected, double-click the function.
   - **Expression** box: The expression that you create is displayed in the Expression box. You can create an expression in any one of the following ways:
     - Type the expression in the Expression box directly.
     - Add **Attributes** and **Functions** by double-clicking them in the **Attributes** tab and the **Functions** tab respectively.

     To add an operator into the expression, click the operator.
   - Commonly used operators are listed in below the Expression box. Click the appropriate operator, indicated by its symbols. If you prefer, you can enter an operator directly into the Expression box. Table 7–1 lists the operators that you can enter:

Table 7–1 Commonly Used Operators

To Enter this Operator	Click
Less than	<
Greater than	>
Less than or equal to	... for the symbol <=
Greater than or equal to	... for the symbol >=
Not equal to	!=
Equal to	=
Or (Logical Or)	...
And	...
Left parenthesis	(
Right parenthesis	)
Parallel symbol	||
Add	+
Minus	-
Multiply	*
Divide	/
Percent	%

   - **Validation Results** text area (read-only): Displays the validation results.
   - **Validate:** Click **Validate** to validate the expression in the Expression box. The results appear in Validation Results.

3. When you have finished creating the expression, click **OK.**

7.3.1 Functions

Expression Builder includes a variety of functions that can be applied to character, numeric, and date data. There are functions that support most of the common data preprocessing required for data mining, including missing values treatment. To see a list of the available functions, expand the category of interest.

Functions are divided into the following categories:

- **Character:** Includes concatenate, trim, length, substring, and others.
- **Conversion:** Converts to character, date, number, and others.
- **Date:** Calculates next day, inserts time stamp, truncates, rounds, and performs other date operations.
- **Numeric:** Includes functions such as absolute value, ceiling, floor, trigonometric functions, hyperbolic functions, logarithms, and exponential values.
- **Analytical:** Performs analytical functions.
- **NULL Value Substitution:** For dates, characters, and numbers.

The notation for the functions is that of SQL.

See Also: *Oracle Database SQL Language Reference*

7.4 Filter Columns

Filter Columns filters out columns so that the columns are not used in subsequent workflow calculations. For example, you might want to filter out or ignore columns that have more than 94 percent Null values.

Optionally, you can identify important attributes.

Filter Columns requires analysis after it runs. The transformation makes recommendations. You can decide which recommendation to accept.

Filter Columns can run in parallel.

These topics describe Filter Columns nodes:

- Creating Filter Columns Node
- Editing Filter Columns Node
- Filter Columns Node Properties
- Filter Columns Node Context Menu

See Also:

- "Attribute Importance"
- "About Parallel Processing"

7.4.1 Creating Filter Columns Node

Before you define filter columns, you must identify a Data Source node and decide whether to find important attributes. To define Filter Columns:

1. Identify or create the node to filter. The node can be any node that provides a data flow, including Data Source nodes.

2. In the **Component** pane, expand the **Transform** section.
   If the **Components** pane is not open, then go to **View** and click **Components.**
3. Click **Filter Columns.** Move the cursor to the workflow and click again.
4. Connect the Data Source node to the Filter Columns node:
   a. Right-click the Data Source node, and select **Connect.**
   b. Draw a line to the Filter Columns node and click again.
5. Right-click the filter columns node and click **Edit**.
6. Right-click the Filter Columns node and click **Run**. Monitor the running of the node in Workflow Jobs.
   If the Workflow Jobs is not open, then in **View,** go to **Data Miner** and click **Workflow Jobs**.
7. After the running of the node is complete, right-click the Filter Columns node and select **View Data** to view the results of the filtered columns.

After you examine the hints, you may want to filter out additional columns.

See Also: "Editing Filter Columns Node"

7.4.2 Editing Filter Columns Node

You can define or edit the filter performed by the Filter Columns node. You can perform the following tasks:

- Exclude Columns when you first edit the Filter Columns node.
- Edit or view Define Filter Columns Settings when you first edit the Filter Columns node.
- Calculate important attributes. Click **Settings** to enable Attribute Importance.
- Evaluate hints and decide which columns to filter out. Additional information is available in the form of Hints after the Filter Columns node runs.

See Also: "Performing Tasks After Running Filter Columns Node"

7.4.2.1 Exclude Columns

By default, all columns are selected for output, that is, all columns are passed to the next node in the workflow.

- To exclude a column, click . The arrow is crossed out, indicated by . The excluded column is ignored, that is, it is not passed on.
- To view or change settings, click **Settings.** The **Define Filter Columns Settings** dialog box opens.

7.4.2.2 Define Filter Columns Settings

You can create and edit Filter Columns settings here. There are three kinds of settings:

- **Data Quality:** Allows Filter Columns settings in terms of percent age of null values, percentage of values that are unique, and percentage of constants. The default values for Data Quality are specified in preferences. You can change the default.You can specify the following Data Quality criteria:
  - **% Nulls less than or equal:** Indicates the largest acceptable percentage of Null values in a column of the data source. You may want to ignore columns that have a larger percentage of Null values. The default value is 95 percent.

- **% Unique less than or equal:** Indicates the largest acceptable percentage of values that are unique in a column of the data source. If a column contains many unique values, then it may not contain useful information for model building. The default value is 95 percent.
- **% Constant less than or equal:** Indicates the largest acceptable percentage of constant values in a column of the data source. If most of the values in a column is the same, the column may not be useful for model building.

- **Attribute Importance:** Enables you to build an Attribute Importance model to identify important attributes.
  By default, this setting is turned OFF. Filter Columns does not calculate Attribute Importance.
- **Sampling:** Enables Filter Column settings according to the default size for random sample for calculating statistics. The default values for sampling are specified in preferences. You can change the default or even turn off sampling. The default sample size is 2000 records.

See Also:

- "Filter Columns"
- "Attribute Importance" for information about how to find important attributes.

7.4.2.3 Performing Tasks After Running Filter Columns Node

After running the Filter Columns Node, you can perform the following tasks:

- View hints: To view hints, double-click the Filter Columns node. The **Edit Column Filter Details Node** dialog box displays hints, to indicate attributes that did not pass the data quality checks. For more information, click .
  - Summary information is displayed about data quality.
  - Values are indicated graphically in the Data Viewer.

  If you specified **Attribute Importance:**
  - Hints indicate attributes that do not have the minimum importance value.
  - The importance of each column is displayed.
- Exclude columns: Go to the Output column for the attribute and click . The icon in the Output column changes to . The selected columns are ignored or excluded, which means that the columns are not for subsequent nodes. It is not necessary to run the nodes again.
- Accept recommendations:
  - For several recommendations, select the attributes and click .
  - For all recommendations, type Ctrl+A and click .
- Apply recommended output settings: Attributes that have Hints are not passed on. Attributes that have no hints are not changed and they are passed on.
- Create a Table or View node: The output of this node is a data flow. To create a table containing the results, use a Create Table or View Node.

See Also:

- "Columns Filter Details Report"
- "Create Table or View Node"

7.4.2.4 Columns Filter Details Report

After the node runs, the Columns Filter Details Report is generated in the **Edit Column Details** dialog box. Columns of the grid summarize the data quality information.

The default settings show both **Attribute Importance** and **Data Quality.**

- When **Attribute Importance** is selected, the following are displayed:
  - Rank
  - Importance
- When **Data Quality** is selected, the following columns are displayed:
  - % Null
  - % Unique
  - % Constant

The **Hints** column in the grid indicates the columns in the data set that do not pass data quality or do not meet the minimum importance value.

Bar graphs give a visual indication of values.

For example, if the percentage of Null values is larger than the value specified for **% Nulls less than or equal**, a hint is generated that indicates that the percentage of Null values is exceeded. If the percent of `NULL` values is very large for a column, you may want to exclude that column.

7.4.2.5 Attribute Importance

If a data set has many attributes, it is likely that not all attributes contribute to a predictive model. Some attributes may simply add noise, that is, they actually detract from the predictive value of the model. Oracle Data Miner ranks the attributes by significance in determining the target value. You can then filter out attributes that are not important in determining the target value.

Using fewer attributes does not necessarily result in loss of predictive accuracy. Using too many attributes, can affect the model and degrade its performance and accuracy. Mining using the smallest number of attributes can save significant computing time and may build better models.

The following are applicable for Attribute Importance:

- Attribute Importance is most useful with Classification.
- The target for Attribute Importance in Filter Column should be the same as the target of the Classification model that you plan to build.
- Attribute Importance calculates the rank and importance for each attribute.
  - The rank of an attribute is an integer.
  - The Importance of an attribute is a real number, which can be negative.

Specify these values for attribute importance:

- **Target:** The value for which to find important attributes. Usually the target of a classification problem.
- **Importance Cutoff:** A number between 0 and 1.0. This value identifies the smallest value for importance that you want to accept. If the importance of an attribute is a negative number, then that attribute is not correlated with the target, so the cutoff should be nonnegative. The default cutoff is 0. The rank or importance of an attribute enables you to select the attribute to be used in building models.
- **Top N:** The maximum number of attributes. The default is 100.
- Select a **Sample** technique for the Attribute Importance calculation. The default is system determined. You can also select Stratified or Random.

  System determined has a stratified cutoff value with a default value of 10.

  – If the distinct count of the selected column is greater than the cutoff value, then use random sampling.

  – If the distinct count of the selected column is less than or equal to the cutoff value, then use stratified sampling.

  Certain combinations of target and sampling may result in performance problems. You are given a warning if there is a performance problem.

7.4.2.5.1 Attribute Importance Viewer

To view an Attribute Importance model, build a Filter Columns node with **Attribute Importance** selected. Right-click the node and select **View Data.** Results are displayed in a new **Filter Columns Details** tab. The viewer has these tabs:

- **Attribute Importance:** Lists those attributes with Importance greater than or equal to 0. Attributes are listed in order of rank from lowest rank (most important) to highest rank. The tab also displays the data type of each attribute. A blue bar indicates the rank. You can sort any of the columns by clicking the column heading.

  – To filter columns, that is, to limit the number of columns displayed, use .

  – To clear filter definition, click . You can also search by name, type, rank or importance.

- **Data:** Lists the important attributes in order of importance, largest first. For each attribute rank and importance, values are listed. Only those attributes with an importance value greater than or equal to 0 are listed.
- **Columns:** Displays the columns created by Attribute Importance, that is attribute name, rank, and importance value.
- **SQL:** This is the SQL that generates the details.

7.4.2.6 Specifying Values for Attributes Importance

Specify the following values for Attribute Importance:

- **Target:** The value for which to find important attributes. Usually the target of a classification problem.
- **Importance Cutoff:** A number between 0 and 1.0. This value identifies the smallest value for importance that you want to accept. If the importance of an attribute is a negative number, then that attribute is not correlated with the target, so the cutoff should be nonnegative. The default cutoff is 0. The rank or importance of an attribute enables you to select the attribute to be used in building models.

- **Top N:** The maximum number of attributes. The default is 100.
- Select a **Sample technique** for the Attribute Importance calculation. The default is system determined. You can also select Stratified or Random.
  - The system determined technique has a stratified cutoff value with a default of 10.
    - If the distinct count of the selected column is greater than the cutoff value, then use random sampling.
    - If the distinct count of the selected column is less than or equal to the cutoff value, then use stratified sampling

Certain combinations of target and sampling may result in performance problems. You are given a warning if there is a performance problem.

7.4.3 Filter Columns Node Properties

To view the **Properties** pane:

1. Select the Filter Column node.
2. The **Properties** pane of the node is displayed in the **Properties** tab. If the **Properties** tab is not visible, then go to **View** and click **Properties.** Alternately, right-click the node and select **Go to Properties.**

Filter Columns Node Properties has these sections:

- Columns: Displays the columns of the data source. After the node runs, hints are displayed.
- Filters: The specifications created with Define Filter Columns Settings.
- Cache
- Details

7.4.4 Filter Columns Node Context Menu

The Filter Columns node context menu contains these entries:

- Connect
- Edit. See "Editing Filter Columns Node" for more information.
- Validate Parents
- Run
- View Data. See "Data Viewer" for more information.
- Force Run
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste

- Parallel Query. See "About Parallel Processing" for more information.
- Select All
- Show Event Log
- Show Runtime Errors. Displayed if there are errors.
- Show Validation Errors. Displayed if there are validation errors.
- Navigate

7.5 Filter Columns Details

Filter Columns Details creates a data flow that consists of the result of Attribute Importance. For each attribute, the rank and importance values are listed.

Note: Filter Columns Details must be connected to a Filter Columns Node that has Attribute Importance selected in Settings. Otherwise, the Filter Columns Details node is Invalid.

Filter Columns Details can run in parallel.

This section consists of the following topics:

- Creating the Filter Columns Details Node
- Editing the Filter Columns Details Node
- Filter Columns Details Node Properties
- Filter Columns Details Node Context Menu

See Also: "About Parallel Processing"

7.5.1 Creating the Filter Columns Details Node

Before creating the Filter Columns Details node, you must identify a Filter Columns node, where **Attribute Importance** is selected in the Settings.

To create a Filter Columns Details node:

1. Identify or create the node to filter. You can only connect a Filter Columns node where Attribute Importance was calculated.
2. In the **Components** pane, expand the **Transforms** section.
   If the **Components** pane is not open, then go to **View** and click **Components.**
3. Click **Filter Columns Details.** Move the cursor to the workflow and click again.
4. Connect the Filter Columns node to the Filter Columns Details node:
   a. Right-click the Filter Columns node, and select **Connect.**
   b. Draw a line to the Filter Columns Details node and click again.
5. You can right-click the Filter Columns Details node and select **Edit**. In this release there are no options to select.
6. Right-click the Filter Columns Details node and select **Run**. Monitor the running of the node in the Workflow Jobs.
   If the Workflow Jobs is not open, then go to **View** and click **Data Miner.** Under Data Miner, click **Workflow Jobs**.

7. After the running of the node is complete, right-click the Filter Columns Details node and select **View Data** to view the results.

The output of this node is a data flow. To create a table containing the results, use a Create Table or View Node.

> **Note:** Filter Columns Details consists of the results of Attribute Importance only. It does not contain any information about Data Quality.

7.5.2 Editing the Filter Columns Details Node

The **Attribute Importance** option is the only option available.

7.5.3 Filter Columns Details Node Properties

To view the properties of Filter Columns Details node:

1. Select the Filter Columns Detail node.
2. The properties of the node is displayed in the **Properties** tab. If the **Properties** tab is not visible, then go to **View** and click **Properties.** Alternately, right-click the node and select **Go to Properties.**

The Filter Columns Node Properties has these sections:

- Output: The only valid value is Attribute Importance, which is the default.

  A grid lists the data types of `ATTRIBUTE_NAME`, `RANK`, and `IMPORTANCE_VALUE`.
- Cache
- Details

7.5.4 Filter Columns Details Node Context Menu

The filter columns details node context menu contains these options:

- Connect
- Edit. See "Editing the Filter Columns Details Node" for more information.
- Validate Parents
- Run
- View Data. See "Attribute Importance Viewer" for more information.
- Force Run
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Parallel Query. See "About Parallel Processing" for more information.

- Select All
- Show Event Log
- Show Runtime Errors. Displayed if there are errors.
- Show Validation Errors. Displayed if there are validation errors.
- Navigate

7.6 Filter Rows

A Filter Rows node enables you to select rows by specifying a SQL statement that describes the rows.
For example, to select all rows where CUST_GENDER is F, specify:

```
CUST_GENDER = 'F'
```

You can either write the SQL expression directly or use Expression Builder.

Filter Rows can run in parallel

This section consists of the following topics:

- Creating a Filter Rows Node
- Edit Filter Rows
- Filter Rows Node Properties
- Filter Rows Node Context Menu

See Also: "About Parallel Processing"

7.6.1 Creating a Filter Rows Node

Prerequisite:

Identify a Data Source node. Identify or create the node to filter. The node can be any node that provides a data flow, including Data Source nodes.

To define a Filter Rows node:

1. In the **Components** pane, expand the **Transform** section.
   If the **Components** pane is not open, then go to **View** and click **Components**.
2. Click **Filter Rows.** Move the cursor to the workflow and click again.
3. Connect the Data Source node to the Filter Rows node:
   a. Move the cursor to the Data Source node.
   b. Right-click the Data Source node, and select **Connect**.
   c. Drag the line to the Filter Rows node and click again.
4. Right-click the Filter Rows node and select **Edit**. Use the **Edit Filter Rows** dialog box to define the filter.
5. Right-click the Filter Rows node and select **Run**. Monitor the running of the node in the Workflow Jobs.
   If the Workflow Jobs is not open, then go to View and click **Data Miner.** Under Data Miner, click **Workflow Jobs.**
6. After the running of the node is complete, right-click the Filter Rows node and select **View Data** to view the results of the Filter Rows.

7.6.2 Edit Filter Rows

The **Edit Filter Rows** dialog box defines or edits the filter performed by the Filter Rows node.

The Edit Filter Rows dialog has two tabs:

- Filter
- Columns

7.6.2.1 Filter

The filter is one or more SQL expressions that describe the rows to select.

To create or edit a filter:

1. Open the Expression Builder by clicking .
2. Write the SQL query to use for filtering.
3. After specifying an expression, you can delete it. Select it and click .
4. After you are done, click **OK.** Data Miner validates the expression.

7.6.2.2 Columns

This tab lists the output columns. You can Filter in several ways.

Click **OK** when you have finished. Data Miner validates the expression.

7.6.3 Filter Rows Node Properties

To view the properties of the Filter Rows node:

1. Select the Filter Rows node.
2. The properties of the node is displayed in the **Properties** tab. If the **Properties** tab is not visible, then go to **View** and click **Properties.** Alternately, right-click the node and select **Go to Properties.**

The Filter Rows node **Properties** tab has these sections:

- Filter: The SQL expression created with Edit Filter Rows. You can modify the expression in the Properties by clicking .
- Columns: The output data columns. For each column, name, alias (if any), and data types are listed.
- Cache
- Details

7.6.4 Filter Rows Node Context Menu

The Filter Rows node context menu contains these options:

- Connect
- Edit. See "Edit Filter Rows" for more information.
- Validate Parents
- Run
- View Data. See "Data Viewer" for more information.
- Force Run

- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Parallel Query. See "About Parallel Processing" for more information.
- Select All
- Show Event Log
- Show Runtime Errors. Displayed if there are errors.
- Show Validation Errors. Displayed if there are validation errors.
- Navigate

7.7 Join

A Join node combines data from two or more Data Source nodes into a new data source.

Technically, a Join node is a query that combines rows from two or more tables, views, or materialized views. For example, a Join node combines tables or views (specified in a `FROM` clause), selects only rows that meet specified criteria (`WHERE` clause), and uses projection to retrieve data from two columns (`SELECT` statement).

Join can run in parallel.

This section contains the following topics:

- Create a Join Node
- Edit a Join Node
- Join Node Properties
- Join Node Context Menu

See Also:

- *Oracle Database Concepts Guide*
- "About Parallel Processing"

7.7.1 Create a Join Node

A Join node requires you to specify two or more Data Source nodes and at least one output column.

Joins are sometimes very slow. If you materialize the join input as an indexed table, then the join may be much faster.

The output of Join is a data flow. To materialize it as a table or view, connect it to a Create Table or View node.

Follow these steps to join two or more Data Source nodes:

1. Identify or create at least two nodes to join. The nodes can be any nodes that provide a data flow, including Data Source nodes.
2. In the **Components** pane, expand the **Transform** node.
   If the **Components** pane is not open, then go to **View** and click **Components**.
3. Drag and drop the **Join** node from the **Components** pane to the **Workflow** pane.
   This adds the Join node to the workflow.
4. Connect the Data Source nodes to be joined to the Join node:
   a. Move the cursor to one of the nodes to join.
   b. Right-click the node, and select **Connect.**
   c. Draw a line to the Join node and click again.
   d. Repeat until all nodes to join are connected to the Join node.
5. Right-click the Join node and select **Edit.** Use the **Edit Join Node** option to define the Join node.
6. Right-click the Join node and select **Run**. Monitor the running of the node in the Workflow Jobs.
   If the Workflow Jobs is not open, then go to **View** and click **Data Miner.** Under Data Miner, click **Workflow Jobs**.
7. After the running of the node is complete, right-click the Join node and select **View Data** to view the results of the join.

You can also define the join and view results through the Join Node Properties.

See Also:

- "Create Table or View Node"
- "Edit a Join Node"
- "Join Node Properties"

7.7.2 Edit a Join Node

You can define a Join node in one of the following ways:

- Either double-click the Join node, or right-click the node and select **Edit**. Click the **Join** tab.
- Select the node. In the **Properties** pane, select the **Join** tab. Click  .

In either case, the **Edit Join Node** dialog box opens.

7.7.2.1 Edit Join Node

Click the **Join** tab if it is not displayed. In the **Edit Join Node** dialog box, you can perform the following tasks:

- To add a new Join column, click . The E**dit Join Column** dialog box opens.
  1. In the **Edit Join Column** dialog box, select the Data Sources—Source 1 and Source 2. You can search columns in either **Source by Name** or by **Data Type.**
  2. Select one entry in Source 1 and the corresponding entry in Source 2.
  3. Click **Add.** Data Miner selects an appropriate Join Type.
     Column 1 (from Source 1), Column 2 (from Source 2), and the Join type are displayed in a grid. You can search this grid by Column 1, Column 2, or Join Type.

4. After you are done, click **OK.**

- To select the columns in the Join, click the **Columns** tab to display **Edit Columns** dialog box.
- To define a filter for the Join, select the **Filter** tab and enter an appropriate SQL expression. You can also use SQL Worksheet (part of SQL Developer) to write a filter.

If there are problems with the join, for example, if one of the data nodes is no longer connected to the Join node, then an information indicator is displayed as follows:

Click the **Resolve Issues.** This opens the **Resolve** dialog box.

The Resolve dialog box displaying the resolved issues.

****

7.7.2.2 Edit Columns

The default setting is to use **Automatic Settings** for the list of columns displayed. To select columns, go to the **Columns** tab of Edit Joins Details in one of the following ways:

- Right-click the Join node and select **Edit**. Click the **Columns.**
- Select the Join node. In the **Properties** pane, click the **Columns.**

To make changes, deselect **Automatic Settings.** You can perform the following tasks:

- Edit the list of columns: Open the **Edit Output Data Column** dialog box and click .
- Delete a column from the output: Select the column and click .

If the node was run, you must run it again.

7.7.2.3 Edit Output Data Column

The default setting is to include all columns from both tables in the output.

To remove a column from the output:

1. Move the columns from the **Selected Attributes** list to the **Available Attributes** list.
2. Click **OK.**

7.7.2.4 Resolve

When a Data Source node is disconnected from a Join node, all the join specifications for that node are retained and are marked as Invalid. Before you run the Join node, you must resolve the issues.

The **Resolve** dialog box provides two ways to resolve the join issues:

- **Remove**: Removes all invalid entries from all specifications (Apply and Data).
- **Resolve**: Displays a grid that enables you to associate an unassigned node with a missing node. The missing nodes are listed in the grid, and an action is suggested.

7.7.3 Join Node Properties

To view the properties of a Join node:

1. Select the Join Node.
2. The properties of the node is displayed in the **Properties** pane. If the **Properties** pane is not visible, then go to **View** and click **Properties.** Alternately, right-click the node and select **Go to Properties.**

The **Properties** pane for a Join node has these sections:

- **Join:** Defines the Join.
- **Columns:** Displays the output columns of the Join.
  For each column, name, node, alias (if any) and data type are displayed. Up to 1000 columns are displayed.
- **Filter** results by defining filter conditions using the **Expression Builder.** Open the **Expression Builder** by clicking .
- Cache
- Details

See Also:

- "Edit Join Node"
- "Edit Columns"
- "Expression Builder"

7.7.4 Join Node Context Menu

The Join node context menu contains these entries:

- Connect
- Edit. See "Edit a Join Node" for more information.
- Validate Parents
- Run
- View Data. See "Data Viewer" for more information.
- Force Run
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Extended Paste
- Paste
- Parallel Query. See "About Parallel Processing" for more information.
- Select All
- Show Event Log
- Show Runtime Errors. Displayed if there are errors.

- Show Validation Errors. Displayed if there are validation errors.
- Navigate

7.8 JSON Query

JSON or JavaScript Object Notation is a data format, that enables users to store and communicate sets of values, lists, and key-value mappings across systems. The support for the JSON data format in Oracle Data Miner (SQL Developer 4.1) is facilitated by the JSON Query node.

The JSON Query node projects the JSON data format to the relational format. It supports only one input data provider node, such as Data Source node. You can perform the following tasks in the JSON Query node:

- Select any JSON attributes in the source data to project it as relational data.
- Select relational columns in the source data to project it as relational data.
- Define aggregation columns on JSON data.
- Preview output data.
- Construct a JSON Query based on user specifications.

> **Note:** JSON Query Node is supported when SQL Developer 4.1 is connected to the Oracle Database 12.1.0.2.

7.8.1 Create JSON Query Node

A JSON Query node should be connected to an input provider node, such as a Data Source node. To run the nodes successfully, the input provide node must contain JSON data.

To create a JSON Query node:

1. In the **Components** pane, go to the **Workflow Editor** and expand **Transform.**

   If the **Components** pane is not visible, then in the SQL Developer menu bar, go to **View** and click **Components.** Alternately, press Ctrl+Shift+P to dock the **Components** pane.

2. In the **Transforms** section, click **JSON Query Node.**
3. Drag and drop the JSON Query node from the **Components** pane to the **Workflow** pane. This adds the JSON Query node to the workflow.

   > **Note:** Ensure that in the workflow there is a Data Source node, which contains JSON data.

4. Right-click the node, for example a Data Source node, from which to create the connection and click **Connect** in the context menu.
5. Draw a line from the selected node to the JSON Query node and click again. This connects the JSON Query node to the Data Source node.

7.8.2 JSON Query Node Editor

In the **Edit JSON Query Node** dialog box, you can only operate on the input columns that are pseudo JSON types. To open the **Edit JSON Query Node** dialog box:

- Double-click the JSON Query node, or
- Right-click the node and click **Edit.**

The Edit JSON Query node consists of the following tabs:

- JSON
- Additional Output
- Aggregate
- Preview

7.8.2.1 JSON

You can select JSON data in the **JSON** tab. In the **Columns** drop-down list, only the input columns containing JSON data (pseudo JSON data types) are listed. Select an input column from the drop-down list. The data structure of the selected input column is displayed in the Structure tab.

The **JSON** tab consists of the following:

- Structure
- Data

7.8.2.1.1 Structure In the **Structure** tab, the JSON data structure for the selected column is displayed. The structure or the Data Guide table must be generated in the parent source node, for example, Data Source node. If the structure is not found, then a message is displayed to communicate the same.

The following information about the data structure is displayed:

- **JSON Attribute:** Displays the generated JSON structure in a hierarchical format. You can select one or more attribute to import. When you select a parent attribute, all child attributes are automatically selected.
- **JSON Data Type:** Displays the JSON data types of all attributes, derived from JSON data.
- **Unnest:** All attributes inside an array are unnested in a relational format. By default, the **Unnest** option is enabled. When the **Unnest** option for an array attribute is disabled, then:
  - The child attributes are displayed but cannot be selected.
  - If the array attribute is selected for output, then the output column will contain JSON representation of the array.

7.8.2.1.2 Data The **Data** tab displays the JSON data that is used to create the JSON structure. In the text panel, the data is displayed in read-only mode. You can select text for copy and paste operations.

You can query the data to be viewed. To query data, click .

7.8.2.2 Additional Output

In the **Additional Output** tab, you can select relational columns in the source data for the output. The input columns that are used by the aggregation definitions in the **Aggregate** tab are automatically added to the list for output.

You can perform the following tasks here:

- **Add relational columns:** Click to add relational columns in the Edit Output Data Column Dialog.
- **Delete relational columns:** Select the relational columns to delete and click .

7.8.2.2.1 Edit Output Data Column Dialog

In the **Edit Output Data Column** dialog box, all the relational columns available in the data source are listed. You can select one or more columns to be added to the output. To add columns:

1. In the **Available Attributes** list, select the columns that you want to include in the output.
2. Click the right arrow to move the attributes to the **Selected Attributes** list. To exclude any columns from the output, select the attribute and click the left arrow.
3. Click **OK.** This includes the columns in the output, and they are listed in the Additional Output tab.

7.8.2.3 Aggregate

You can define aggregation column definitions on JSON attributes in the **Aggregate** tab. The **Aggregate** tab displays the information in two sections:

- **Group By Attribute** section: Here, the **Group By** attributes are listed, along with the attribute count. You can perform the following tasks:
  - **View JSON Paths:** Click **JSON Paths** to display the attribute name with context information. For example, `$."customers"."cust_id"`.
    If not enabled, then only the attribute name is displayed.
  - **Edit and Add Attributes:** Click to add **Group By** attributes in the Edit Group By dialog box.
  - **Delete Attributes:** Select the attributes to delete and click .
- **Aggregate Attributes** section: Here, the aggregation columns are displayed along with the column count.
  - **View JSON Paths:** Click **JSON Paths** to display the attribute name with context information. For example, `$."customers"."cust_id"`.
    If not enabled, then only the attribute name is displayed.
  - **Define Aggregations Columns:** Click to define an aggregation column in the Add Aggregations box.
  - **Delete Aggregation column:** Click to delete selected columns.

7.8.2.3.1 Add Aggregations

In the **Add Aggregation** dialog box, you can define functions for JSON attributes. The dialog box displays JSON structures in a hierarchical view. You can select multiple attributes and then apply an aggregation function to it.

Note: Object and Array type attributes cannot be selected.

You can perform the following tasks:

- Define aggregation functions:
  1. Select the JSON attributes. You can select multiple attributes by pressing the Ctrl key and clicking the attributes for which you want to define functions.
  2. Click to select and apply a function for the selected attributes. The applicable functions are listed. Select the function that you want to apply.

     Alternately, click the corresponding row in the **Function** column. The applicable functions are listed in a drop-down list. Select the function that you want to apply. Using this option you can define function for only one attribute at a time.
  3. Click **OK.**
- **Clear Aggregation Definition:** Select the attribute and click . The defined function along with the output and Sub Group By entries are deleted.
- **Edit Sub Group By Elements:** Select the attribute and click . The **Edit Sub Group By** dialog box opens.
- **Search:** Click to find attribute based on partial attribute name.

7.8.2.3.2 Edit Sub Group By In the **Edit Sub Group By** dialog box, you can add Sub Group By attributes to the selected JSON attribute. To add attributes:

1. In the upper pane, expand the **Available Attributes** folder.
2. Select the attributes that you want to add as Sub Group By attributes. The selected attributes are listed in the lower pane, which also displays the count of attributes that are added.
3. Click **OK.**

7.8.2.3.3 Edit Group By The **Edit Group By** dialog box displays the relational columns above the JSON attribute collection. You can add relational columns as part of the top level Group By. To add relational columns:

1. In the upper pane, expand the **Available Attributes** folder.
2. Select the columns that you want to add. The selected columns are listed in the lower pane.
3. Click **OK.**

7.8.2.4 Preview

You can preview the node output in the **Preview** tab. The output is displayed in two tabs:

- Output Columns
- Output Data

7.8.2.4.1 Output Columns In the **Output Column** tab, the columns in the header are displayed in a grid format. Click **JSON Paths** to view source attribute name.

- If you click **JSON Paths,** then the source attribute name along with the contextual information is displayed. For example, `$."customers"."cust_id"`.
- If you do not click **JSON Paths,** then only the attribute name is displayed. For example, `cust_id`.

The following details of the columns are displayed in the **Output Columns** tab:

- **Name:** Displays the output column names
- **Data Type:** Displays the data type of the output column
- **Data Source:** Displays the source of the attribute name
- **JSON Paths:** Displays the attribute source
- **Aggregate:** Displays the aggregation function used for the aggregations
- **Group By:** Displays the Group By attributes
- **Sub Group By:** Displays the Sub-Group By attributes used in the aggregations

7.8.2.4.2 Output Data In the **Output Data** tab, the query result of the top N rows is displayed. The query reflects the latest user specifications. The query result is displayed in a grid format.

7.8.3 JSON Query Node Properties

If the **Properties** pane is closed, then go to **View** and click **Properties.** Alternately, right-click the node and select **Go to Properties** from the context menu.

The JSON Query node **Properties** pane has these sections:

- Output
- Cache
- Details

7.8.3.1 Output

The **Output** section in the **Properties** pane displays the **Output Columns** in read-only mode.

7.8.3.2 Cache

The **Cache** section provides the option to generate cache for the output data. To generate cache output:

1. Select **Generate Cache of Output Data to Optimize Viewing of Results** to generate cache output.
2. In the **Sampling Size** field, select an option:
   - **Number of Rows** (default): The default sampling size is 2000. Use the arrows to set a different number.
   - **Percent:** Move the pointer to set the percentage.

7.8.3.3 Details

The **Details** section displays the name of the node and any comments about it. You can change the name and add comments in the fields here:

- **Node Name**
- **Node Comments**

7.8.4 JSON Query Node Context Menu

The context menu for a JSON Query Node has these selections:

- Connect
- Run
- Force Run
- Edit. See "JSON Query Node Editor" for more information.
- View Data
- Generate Apply Chain
- Show Event Log
- Validate Parents
- Deploy
- Save SQL
- Cut
- Copy
- Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Go to Properties
- Navigate

7.9 Sample

A Sample node enables you to sample data in one of the following ways:

- **Random Sample**: A sample in which every element of the data set has an equal chance of being selected.
- **Top N Sample**: The default sample that selects the first *N* values.
- **Stratified Sample**: A sample is created as follows:
  - First, the data set is divided into disjoint subsets or strata.
  - Then, a random sample is taken from each subsets.

  This technique is used when the distribution of target values is skewed greatly. For example, the response to a marketing campaign may have a positive target value of 1% of the time or less.

Sampling nested data is best done with a Case ID.

The Sample node can run in parallel.

This section has the following topics:

- Sample Nested Data
- Creating a Sample Node
- Edit Sample Node
- Sample Node Properties
- Sample Node Context Menu

See Also:

- "Sample Nested Data"
- "About Parallel Processing"

7.9.1 Sample Nested Data

Sampling nested data may require a case ID. If you do not specify a case ID, the sample operation may fail for a nested column that is very dense and deep. Failure may occur if the amount of nested data per row exceeds the maximum of 30,000 for a specific column or row.

A case ID also allows Data Miner to perform stratified sorts on data that is dense and deep.

7.9.2 Creating a Sample Node

To specify a sample, you must identify a Data Source node and the details of the sample.

1. Identify or create the node to sample. The node can be any node that provides a data flow, including Data Source nodes.
2. In the **Components** pane, expand the **Transforms** section.
   If the **Components** pane is not open, then go to **View** and click **Components**.
3. Click **Sample.** Move the cursor to the workflow and click again.
4. Connect the Data Source node to the Sample node:
   a. Move the cursor to the Data Source node.
   b. Right-click the Data Source node and select **Connect** from the context menu.
   c. Drag the line to the Sample node and click again.
5. Either double-click the Sample node, or right-click the Sample node, and click **Edit**. The **Edit Sample Node** dialog box opens.
6. Define the sample in the **Edit Sample Node** dialog box.
7. Right-click the Sample node and click **Run**. Monitor the running of the node in the Workflow Jobs.
   If the **Workflow Jobs** is not open, then go to **View** and click **Data Miner.** Under Data Miner, click **Workflow Jobs**.
8. After the running of the node is complete, right-click the Sample node and select **View Data** to view the results of the sample.

See Also: "Edit Sample Node"

7.9.3 Edit Sample Node

The settings describe the type of sample to create and the size of the sample. In the **Edit Sample Node** dialog box, you can define and edit a sample.

To edit the settings for the Sample node:

1. Open the **Edit Sample Node** dialog box:
   - Either double-click the Sample node, or right-click the Sample node, and select **Edit.**
   - Select the node and go to the **Settings** tab in the Sample node **Properties** pane.

2. In the **Edit Sample Node** dialog box, you can provide and edit the following details:
   - **Sample Size:** This is the number of rows in the sample. You can specify the number of rows in terms of:
     – Number of rows (default)
     – Percent. The default is 60 percent.
   - **Rows:** This is the number of rows in the sample. You can change the default value, and enter a different value.
     The default is 2000.
   - **Sample Type:** The options are:
     – Random (Default)
     – Top N
     – Stratified

   See Also: "Sample Node Properties"

7.9.3.1 Random

For a random sample, specify the following:

- **Seed**: The default seed is 2345
  You can specify a different integer.
- **Case ID** (optional): Select a case ID from the drop-down list.

If you specify a seed and a case ID, then the sample is reproducible.

7.9.3.2 Top N

No other specifications are available for Top N.

7.9.3.3 Stratified

For a stratified sample, specify the following:

- **Column**: Select the column for stratification.
- **Seed**: Default seed=12345
  You can specify a different integer.
- **Case ID** (optional): Select a case ID from the drop-down list.

  If you specify a seed and a case ID, the sample is reproducible.
- **Distribution**: Specify how the sample is to be created. There are three options:
  – **Original**: The distribution of the selected column in the sample is the same as the distribution in data source.
  For example, if the column GENDER has M as the value for 95 percent of the cases, then in the sample, the value of GENDER is M for 95% of the cases.
  – **Balanced**: The distribution of the values of the column is equal in the sample, regardless of the distribution in the data source.
  If the column is GENDER, and GENDER has two values M and F, then 50% of the time the value of GENDER is M.

- **Custom**: You define how the values of the columns are distributed in the sample. You must run the node once before you define the custom distribution. Click **Edit** to open the **Custom Balance** dialog box.

The **Stratified** dialog box displays a histogram of the values of the selected column at the bottom of the window.
To see more details, click **View** to display the **Custom Balance** dialog box.

See Also: "Custom Balance"

7.9.3.4 Custom Balance

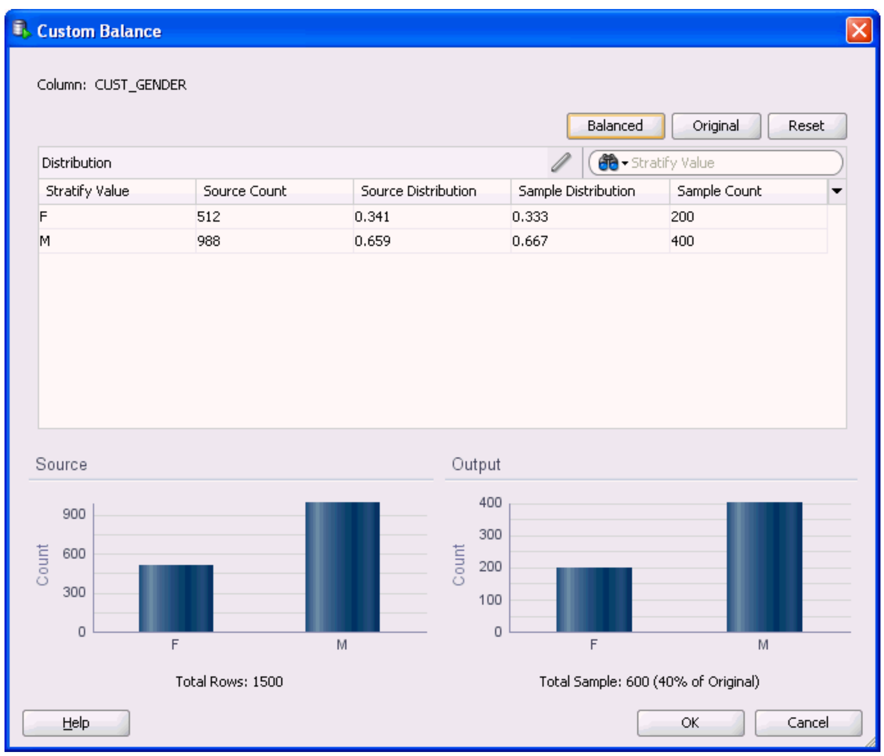
The **Custom Balance** dialog box enables you to specify exactly how the selected column is balanced.

You must run the node to collect statistics before you create custom balance. After you run the node, edit it, select **Custom Distribution,** and click **View**. The **Custom Balance** dialog opens:

You can either create custom entries for each value of the stratify attribute, or you can click **Original** or **Balanced** to provide a starting point. You can click **Reset** to reset to the original values.

To create a custom value, select the attribute to change and click  .

Change the value in the **Sample Count** column to the custom value. Press Enter. The new sample is displayed as output in the lower part of the screen. Change as many values as required. Click **OK** when you have finished.



The **Custom Balance** dialog box showing details for a stratified sample.

****

7.9.4 Sample Node Properties

To view the properties of a Sample node:

1. Select the Sample node.
2. The properties of the node are displayed in the **Properties** pane.
   If the **Properties** pane is not visible, then go to **View** and click **Properties.**
   Alternately, right-click the node and select **Go to Properties.**

The Sample node **Properties** pane has these sections:

- **Settings:** You can specify the following:
  - **Sample Size:** Select a sample size in terms of:
    * Percent. Default=60%
    * Number of Rows. The default number of rows is 2000.
  - **Sample Type:** The options are:
    * Random (default)
    * Stratified
    * Top N
  - **Seed:** The default seed is 12345. You can specify a different integer.
  - **Case ID.** This is an optional field. Select a case ID from the drop-down list.
    If you specify a seed and a case ID, then the sample is reproducible.
- Cache
- Details

See Also: "Edit Sample Node"

7.9.5 Sample Node Context Menu

The Sample node context menu has these options:

- Connect
- Edit. See "Edit Sample Node" for more information.
- Validate Parents
- Run
- View Data. See "Data Viewer" for more information.
- Force Run
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Extended Paste
- Paste
- Parallel Query. See "About Parallel Processing" for more information.

- Select All
- Show Event Log
- Show Runtime Errors. Displayed if there are errors.
- Show Validation Errors. Displayed if there are validation errors.
- Navigate

7.10 Transform

A Transform node can use either sampled data or all data to calculate statistics. You can use these statistics as a guide to defining one of several transformation.

These topics describe Transform nodes:

- Transforms Overview
- Support for Date and Time Data Types
- Creating Transform Node
- Edit Transform Node
- Add Transform
- Transform Node Properties
- Transform Node Context Menu

7.10.1 Transforms Overview

A Transform node can use either sampled data or all data to calculate statistics. You can use these statistics as a guide to defining one of several transformation.

The following transformations are supported:

- Binning
- Custom
- Missing Values
- Normalization
- Outlier

The transformations available depend on the data type of the attribute. For example, normalization cannot be performed on character data.

You define transformation on a per-column basis. After you have defined a transformation, you can transform several columns in the same way.

The Transform node can run in parallel.

To use a Transform node, connect it to a data flow, that is a Data Source node or some other node such as a filtering node that produces attributes. Then select the attributes to transform.

See Also: "About Parallel Processing"

7.10.1.1 Binning

Binning converts the following:

- A continuous variable into a categorical variable.

- A continuous value into a continuous value. For example, age can be converted to 10 groups from 1 to 10.
- A categorical value with many values into a categorical variable with fewer variables.

For example, salary is a continuous variable. If you divide salary into 10 bins, you convert salary into a categorical variable with 10 values, where each value represents a salary range.

You can bin both numbers and character types `VARCHAR2` and `CHAR`.

7.10.1.1.1 Recode Oracle Data Miner does not support recode transformations. However, you can use custom binning to perform a recode transformation. For example, to recode the US states ME, NH, VT, CT, MA, and RI to the value NE, create a custom bin that puts the 5 states into a bin named NE.

7.10.1.2 Custom

Custom enables you to compute a new value for a field based on combinations of existing attributes and common function. Use **Expression Builder** to create the new attribute.

> **See Also:** "Expression Builder"

7.10.1.3 Missing Values

The Missing Values Transform enables you to specify how missing values are treated.

A data value can be missing for a variety of reasons:

- If the data value was not measured, that is, it has a Null value.
- If the data value was not answered.
- If the data value was unknown.
- If the data value was lost.

Data Mining algorithms vary in the way they treat missing values:

- Ignore the missing values, and then omit any records that contain missing values.
- Replace missing values with the mode or mean.
- Infer missing values from existing values.

7.10.1.4 Normalization

Normalization consists of transforming numeric values into a specific range, such as [–1.0,1.0] or [0.0,1.0] such that `x_new = (x_old-shift)/scale`. Normalization applies only to numeric attributes.

Oracle Data Miner enables you to specify the following kinds of normalization:

- **Min Max**: Normalizes each attribute using the transformation `x_new = (x_old-min)/(max-min)`
- **Linear Scale**: Normalizes each attribute using the transformation `x_new = (x_old-shift)/scale`
- **Z-Score**: Normalizes numeric attributes using the mean and standard deviation computed from the data. Normalizes each attribute using the transformation `x_new = (x-mean)/standard deviation`

- **Custom**: The user defines normalization.

Normalization provides transformations that perform min-max normalization, scale normalization, and z-score normalization.

Note: You cannot normalize character data.

7.10.1.5 Outlier

An Outlier is a data value that is not in the typical population of data, that is, extreme values. In a normal distribution, outliers are typically at least 3 standard deviations from the mean.

You specify a treatment by defining what constitutes an outlier (for example, all values in the top and bottom 5 percent of values) and how to replace outliers.

Note: Usually, you can replace outliers with null or edge values.

For example:

Mean of an attribute distribution=10

Standard deviation=5

Outliers are values that are:

- Less than -5 (The mean minus 3 times the standard deviation)
- Greater than 25 (The mean plus three times the standard deviation)

Then, in this case you can either replace the outlier -10 with Null or with 5.

7.10.2 Support for Date and Time Data Types

The Transform Node provides limited support for these data types for date and time:

- `DATE`
- `TIMESTAMP`
- `TIMESTAMP_WITH_TIMEZONE`
- `TIMESTAMP_WITH_LOCAL_TIMEZONE`

You can bin date and time attributes using Equal Width or Custom binning. You can apply Statistic and Missing Value transformation with either Statistic or Value treatment.

7.10.3 Creating Transform Node

To specify a transform, you must identify a Data Source node or any other node which provides data, such as the Create Table Node, and the details of the transform.

1. Identify or create the node to transform. The node can be any node that provides a data flow, including Data Source nodes.
2. In the **Components** pane, expand the **Transforms** section.
   If the **Components** pane is not open, then go to **View** and click **Components**.
3. Click **Transform.** Move the cursor to the workflow and click again.
4. Connect the Data Source node to the Transform node:

a. Move the cursor to the Data Source node.

b. Right-click the Data Source node and select **Connect.**

c. Drag the line to the Transform node and click again.

5. Either double-click the Transform node or right-click it node and select **Edit**. Use the **Edit Transform Node** dialog box to define the transform.

6. Right-click the Transform node and select **Run**. Monitor the running of the node in Workflow Jobs.
If **Workflow Jobs** is not open, then go to **View** and click **Data Miner.** Under Data Miner, click **Workflow Jobs.**

7. After the running of the node is complete, right-click the Transform node and select **View Data** to view the results of the transform.

See Also: "Edit Transform Node"

7.10.4 Edit Transform Node

You can define and edit the Transform node using the **Edit Transform Node** dialog box. This dialog box has two tabs:

- **Transformations**
- **Statistics**

In the **Transformation** tab, the statistics for each column is displayed. If you do not want to see statistics, then deselect **Show Statistics.**

Note: You must run the node to see statistics.

You can perform the following tasks in the Transformations tab:

- **Define Transformation:** Select one or more original columns, that is, columns that are not transformed. Click .

  The **Add Transform** dialog box opens if you selected one or fewer columns. Otherwise, the **Apply Transform Wizard** opens.

- **Define Custom Transformation:** Select one or more original columns, that is, columns that are not transformed. Click .

  The **Add Custom Transform** dialog box opens. You can add custom transformations here.

  The default behavior is to ignore the original column and to use the transformed column as output. The values displayed in the **Output** column are indicated by:

  – : For a column that is included

  – : For a column that is ignored

- **Change Values in the Output Column:** Click the icon displayed in the **Output** column to edit the values in the **Add Transform** dialog box.

- **Edit Transformed Column:** You can edit transformed columns only. For example, you can edit AGE_BIN, but not AGE. To edit transforms, select one or more transformed columns and click . The **Edit Transform** dialog box opens if you selected one or fewer columns.

- **Delete Transform:** Select one or more transformed columns and click .

- **Filter Column:** To limit the number of columns displayed, click . You can search by:
  - Output Column
  - Transform
  - Source Column
- **Clear filter definition:** To clear filter definitions, click .
- **View effects of transform:** To view the effect of a transformation:
  - Run the node.
  - After running of the node is complete, double-click the node.
  - Select the transformed column to see histograms that compare the original column with the transformed column.

When a column has a transformation applied to it, a new row in the list of columns is generated. Because each column must have a name, the name of the new row is based on the name of the old column and the type of transformation performed. Typically users want to transform a column and only have the *output* of the transform node contained in the new column. The original column has an option set to prevent it from being passed as one of the output columns. For example, if you bin AGE creating AGE_BIN, then AGE is not passed on and AGE_BIN is passed on.

See Also:

- "Add Transform"
- "Edit Transform"
- "Add Custom Transform"
- "Apply Transform Wizard"

7.10.4.1 Add Transform

To add a transformation:

1. In the **Edit Transform Node** dialog box, click . The **Add Transform** dialog box opens.
   To add a custom transformation, click .
2. In the **Transform Type** field, select a transform type, that is, the type of transformation that you want to define. Default type is Binning.
   The field in the Add Transform dialog box depend on the transform type that you select:
   - Binning. For binning transformation, enter the details as applicable.
   - Missing Values
   - Normalization
   - Outlier
   - Use Existing Column
3. After you are done, click **OK.**

See Also:

- "Bin Assignment"
- "Add Custom Transform"

7.10.4.1.1 Binning

Binning is a transformation type that you can use to:

- Transform a continuous variable to a discrete one.
- Transform a variable with large number of discrete values to one with a smaller number of discrete values.

The default transformation type is `Binning`.

The types of binning supported depend on the data type of the column:

- For numeric data type `NUMBER`, the supported types of binning are:
  - Bin Equal Width (Number) (Default)
  - Bin Quantile
  - Custom
- For categorical data type `VARCHAR2`, the supported types of binning are:
  - Bin Top N (Default)
  - Custom
- For date and time data types `DATE`, `TIMESTAMP`, `TIMESTAMP WITH LOCAL TIMEZONE`, and `TIMESTAMP WITH TIMEZONE`, the supported types of binning are:
  - Bin Equal Width (Number) (Default)
  - Custom

Note: The number of bins must be two.

7.10.4.1.2 Bin Equal Width (Number)

This selection determines bins for numeric attributes by dividing the range of values into a specified number of bins of equal size. Edit the following fields:

- **Bin Count:** You can change the Bin Count to any number greater than or equal to 2. The default count is set to 10.
- **Bin Label:** Select a different Bin Label scheme from the list. The default is set to `Range`.

Click **OK** when you have finished.

7.10.4.1.3 Bin Quantile This selection divides attributes into bins so that each bin contains approximately the same number of cases. Edit the following fields:

- **Bin Count:** You can change the Bin Count to any number greater than or equal to 2. The default count is set to `10`.
- **Bin Label:** You can select a different Bin Label scheme from the list. The default is set to `Range`.

Click **OK** when you have finished.

7.10.4.1.4 Bin Top N

The Bin Top N type bins categorical attributes. The bin definition for each attribute is computed based on the occurrence frequency of values that are computed from the data.

Specify `N`, the number of bins. Each of the bins `bin_1, …, bin_N` contains the values with the highest frequencies. The last `bin_N` contains all remaining values.

You can change the **Bin Count** to any number greater than or equal to 3. The default count is set to `10`.

Click **OK** when you have finished.

7.10.4.1.5 Custom

Custom binning enables you to define custom bins.

To define bins, click **Bin Assignment** and then modify the default bins.

After you generate default bins, you can modify the generated bins in several ways:

- **Edit Bin Name:** If it is a Range label.
- **Delete Bins:** Select it and click .
- **Add Bins:** Click .
- **Edit Bins:** Select a bin and click .

> **See Also:**
>
> - "Bin Assignment"
> - "Add Bin"
> - "Edit Bin"

7.10.4.1.6 Bin Assignment

Select the following options:

- **Binning Type:** The default type depends on the data type of the attribute that you are binning:
  - If the data type of the attribute is Number, then the default binning type is **Bin Equal Width.**
  - If the data type of the attribute is Character, then the default binning type is **Bin Top N.**

  You can change the binning type for numbers.
- **Bin Count:** The default count is `10`. You can change this to any integer greater than `2`.
- **Bin Labels:** The default label for numbers is `Range`. You can change the bin label to `Number`.
- **Transform NULLs:** If the **Transform NULLs** check-box is selected for a binning transformation that produces NUMBER data type, then null values are placed into the last bin. For example, if the AGE column has null values and Equal Width Binning was requested with Bin Labels value equal to Number, and the number of bins is 10, then null values will be in bin number 11.
  For this option, the following conditions apply:

- When deselected, null values are excluded from the generated transformation SQL.

> **Note:** Applicable only for those binning transformations that produce VARCHAR2 data type after transformation.

- This field is not editable for those binning transformations which produce numeric data type after transformation.
- For legacy workflows, this field is selected by default, and the corresponding field contains the value `Null bin`.

Click **OK** when you have finished. You return to the Custom display where you modify the generated bins.

7.10.4.1.7 Edit Bin

The way to edit bins depends on the data type of the attribute:

- For numbers: Edit lower bounds in the grid. You cannot edit a bin without a lower bound. You cannot add a value that is less than the prior bin lower bound value or greater than the following bin lower bound value.
- For characters: The **Edit Custom Categorical Bins** dialog box has two columns:
  - Bins: You can add bins, delete a selected bin and change the name of the selected bin.
  - Bin Assignment: You can delete values for the selected Bin.

Click **OK** when you have finished editing bins. If you are editing custom bins for categorical, first click **OK** twice (once to closed the **Edit Custom Categorical Bins** dialog box).

7.10.4.1.8 Add Bin

You can add bins for:

- **Categoricals:** Open the **Edit Custom Categorical Bins** and click . The new bin has a default name that you can change. Add values to the bin in the **Bin Assignment** column.
- **Numericals:** Select a bin and click . You can change the name of the bin and add a range of values.

7.10.4.1.9 Missing Values

Missing Values is a transformation type that replaces missing values with an appropriate value.

To specify a Missing Values transformation:

1. In the **Transform Type** field, select the option **Missing Values.**
2. In the **Missing Values** field, select an option:
   - **Statistic:** Replaces the missing value with a statistical measure. Statistic is the default treatment for missing values. The applicable Statistic type depends on the data type of the column:
     - For numeric columns, you can replace missing values with `Mean` (default), `Median, Minimum, Maximum.`

  - For categorical columns, you can replace missing values with `Mode` (default).
- **Value:** Replaces Missing Values with the specified value. Oracle Data Miner provides a default value that you can change.
  - If statistics are not available, then the default value is `0`.
  - If statistics are available, then the default value is:
    `Mean` for numerical columns
    `Mode` for categorical columns

  Both these treatments can be applied to attributes that have a date or time data type `DATE`, `TIMESTAMP`, `TIMESTAMP_WITH_LOCAL_ TIMEZONE`, and `TIMESTAMP_WITH_TIMEZONE`.

3. After you are done, click **OK.**

7.10.4.1.10 Normalization

Normalization consists of transforming numeric values into a specific range, such as [–1.0,1.0] or [0.0,1.0] so that x_new = (x_old-shift)/scale. Normalization usually results in values whose absolute value is less than or equal to 1.0.

Note: Normalization applies only to numeric columns. Therefore, you can normalize numeric attributes only.

To normalize a column:

1. In the **Transform Type** field, select the option **Normalization.**
2. In the **Normalization Type** field, select a type from the drop-down list. Oracle Data Miner supports these types of normalization:
   - **Min Max:** Normalizes the column using the transformation `x_new = (x_old-min)/ (max-min)`. The default is **min-max.**
   - **Z-score:** Normalizes numeric columns using the mean and standard deviation computed from the data. Normalizes each column using the transformation `x_new = (x-mean)/standard deviation.`
   - **Linear Scale:** Normalizes each column using the transformation `x_new = (x-0)/ max(abs(max), abs(min)).`
   - **Manual:** Defines normalization by specifying the shift and scale for the transformation `x_new = (x_old-shift)/scale.` If you select **Manual,** then specify the following:
     - **Shift**
     - **Scale**
3. After you are done, click **OK.**

7.10.4.1.11 Outlier

An outlier is a data value that does not come from the typical population of data. In other words, it is an extreme value. In a normal distribution, outliers are typically at least 3 standard deviations from the mean. Outliers are usually replaced with values that are not extreme, or they are replaced with `Null`.

Note: You can define outlier treatments for numeric columns only.

To define an Outlier transformation:

1. In the **Transform Type** field, select the option **Outlier.**
2. In the **Outlier Type** field, select any one of the following options:
   - **Standard Deviation:** This is the default Outlier type. For this outlier type, enter a Standard Deviation to define the Outlier in the following field:
     – **Multiples of Sigma:** This is the number of standard deviations that define an outlier.
     The default is `3`, that is, 3 standard deviations.
     3 Standard Deviation means that an outlier is a value less than `mean - 3 * Standard Deviation` or greater than `mean + 3* Standard Deviation.`
   - **Percent:** Enables you to specify that outliers are values in a bottom percentage and a top percent. The default is to specify that outliers are in the bottom 5 percent or in the top 5 percent. You can change the defaults by entering values in these fields:
     – **Lower Percent Value**
     – **Upper Percent Value**
   - **Value:** Enables you to specify a lower value and an upper value so that outliers are those values less than the lower value or greater than the upper value.
   You can change these values, but the upper value must be bigger than the lower value.
     – **Lower Value:** If statistics are available, then the default is `-3* standard deviation.`
     If statistics are not available, then the default is `0`.
     – **Upper Value:** If statistics are available, then the default is `+3* standard deviation.`
     If statistics are not available, then the default is `1`.
3. In the **Replace With** field, select an option to specify how to replace outliers. The options are:
   - **Null** (Default)
   - **Edge Value**

   For example:
   If the mean of a column distribution is 10, and
   If standard deviation is 10
   Then, outliers can be:
     – Values that are less than `-5`, that is, `Mean-3*Standard Deviation`
     – Values that are greater than `25`, that is, `Mean+3*Standard Deviation`

   Outlier=`-10`. You can replace `-10` with `Null` or with `-5`, which is the edge value.
4. After you are done, click **OK.**

7.10.4.1.12 Use Existing Column

This selection is not displayed until at least one transformation exists.

This selection is used when you add, or edit multiple transformation.

> **See Also:** "Add or Edit Multiple Transforms"

7.10.4.1.13 Add or Edit Multiple Transforms

You can define or edit transformations for several column at a time. You can also apply an existing transformation to one or more columns.

To add or edit transformations for multiple transforms:

1. Double-click the Transform node. The **Transformation Editor** opens.
2. To define the same transform for several columns, select the columns. You can select columns with different but compatible data types. For example, CHAR and VARCHAR are characters, and are compatible data types. If there are no transformations that apply to all the columns, you get a message. Click .

   The **Apply Transform Wizard** opens.

   a. Select the **Transform** type to apply to all columns.
   b. Provide the specific details related to the transform type that you have selected.
   c. Click **Next.**
   d. Click **Generate Statistic.**
   e. Click **Finish.**
3. If you have already transformed a column, then you can define the same transformation for several other columns.

   Suppose you have binned AGE creating `AGE_BIN`. To bin several columns in the same way, select AGE and the columns that you want to bin in the same way. Click .

   The **Apply Transform Wizard** opens.

   a. For Transform Type, select **<Use Existing>.** AGE_BIN is listed as the Transformed column. You cannot change any other values.
   b. Click **Next.** You can change the names of the output columns.
   c. Select **Generate Statistic on Finish.**
   d. Click **Finish.**
4. To edit several transformations at the same time, select the transformations and click .

   The **Apply Transform Wizard** opens. Edit the transformation and click **Finish**.

7.10.4.2 Add Custom Transform

In the **Add Custom Transform** dialog box, you can define a custom transformation. The default name for the new attribute is `EXPRESSION`. You can change this name.

In the **Add Custom Transform** dialog box, you can perform the following tasks:

- Add an expression: Click . The **Expression Builder** opens. Use expression build to define an expression.
  - Validate the expression.

- Click **OK.**

- Edit a custom transformation.
- Delete a custom Transformation: Click .

See Also: "Edit Custom Transform"

7.10.4.3 Apply Transform Wizard

The **Apply Transform** wizard enables you to define or edit transformations for several columns at the same time. The first step of the wizard is similar the Add Transform dialog box.

You cannot select custom transformations.

1. In the **Choose Transformation** section:
   - **Transform:** Select the transform type.
   - Provide the details related to the selected transformation type.
2. Click **Next.**
3. In the **Choose Columns** section, specify names for the transformed columns. You can accept the names or change them. The available transformations are those transformations that can be performed on all selected columns. This is an optional section.
4. Click **Finish.**

See Also: "Add Transform"

7.10.4.3.1 Define Columns

The second step of the wizard enables you to specify names for the transformed columns. You can accept the names or change them.

The default is *not* to generate statistics on finish. Select the check box to generate statistics.

Click **Finish** when you have finished.

7.10.4.4 Edit Transform

The **Edit Transform** dialog box is similar to the **Add Transform** dialog box.

If the node has run, the **Edit** dialog displays information about both the untransformed column and transformed version:

- The **Histogram** tab shows histogram for both the untransformed attribute and the transformed in two sets of histograms. On the left side of the tab are histograms for the untransformed column. On the right side of the tab are histograms for the transformed column.
- The **Statistics** tab shows statistics for the transformed data and for the original data.

Note: When you transform data, the transformed data may have a different data type from the type of the original data. For example, AGE has type `NUMBER` and AGE_BIN has type `VARCHAR2`.

See Also: "Add Transform"

7.10.4.5 Edit Custom Transform

To edit an expression:

1. Select the attribute and click . The **Expression Builder** opens.
2. Use the **Expression Builder** to modify the expression.
3. Validate the expression.
4. Click **OK.**

To delete an expression, click .

See Also: "Expression Builder"

7.10.5 Transform Node Properties

If **Properties** pane is closed, then go to **View** and click **Properties.** Alternately, right-click the node and select **Go to Properties** from the context menu.

The Transform node **Properties** pane has these sections:

- **Transform:** Specifies how the transformations are defined. You can modify these values.

  The transformations are summarized in a grid. For each column name (data) type, transform, and output are displayed. If you bin `AGE` creating `AGE_BIN`, then `AGE` is not used as output, that is, is not passed to subsequent nodes.
- **Histogram:** Specifies the number of bins used in histograms. You can specify a different number of bins for histograms created for numeric, categorical, and date data types. The default is `10` bins for all data types.
- Sample
- Cache
- Details

See Also: "Edit Transform Node"

7.10.6 Transform Node Context Menu

The Transform node context menu contains these options:

- Connect
- Edit. See "Edit Transform Node" for more information.
- Validate Parents
- Run
- View Data. See "Data Viewer" for more information.
- Force Run
- Deploy
- Show Graph
- Generate Apply Chain
- Cut

- Copy
- Extended Paste
- Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed if there are errors.
- Show Validation Errors. Displayed if there are validation errors.
- Navigate

8 Model Nodes

Model nodes specify the models to build and the models to add to the workflow. The **Models** section in the **Components** pane contains the Models nodes. The models in the **Components** pane are:

- Anomaly Detection Node
- Association Node
- Classification Node
- Clustering Node
- Feature Extraction Node
- Model Node
- Model Details Node
- Regression Node

8.1 Types of Models

The types of models available are:

- Anomaly Detection Node: Builds Anomaly Detection models using a one-class Support Vector Machine (SVM).
- Association Node: Builds models for market basket analysis.
- Classification Node: Builds and tests classification models with the same target, case ID, cost, and split settings, where relevant. The models use the classification algorithms: Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), and Generalized Linear Model (GLM).
- Clustering Node: Builds clustering models using the clustering algorithms: *k*-Means, O-Cluster, and Expectation Maximization (EM). EM requires Oracle Database 12*c* Release 1 (12.1) or later.
- Feature Extraction Node: Builds feature extraction models using the feature extraction algorithms: nonnegative matrix factorization, principal components analysis (PCA), and singular value decomposition (SVD). PCA and SVD require Oracle Database 12*c* Release 1 (12.1) or later.
- Model Node: Adds models to a workflow that were not built in the current workflow. This node has no input data.
- Model Details Node: Extracts model details from a model build node, a Model node, or any node that produces a model.

- Regression Node: Builds and tests a collection of Regression models with the same target, case ID, cost, and split settings, where relevant. The models use the regression algorithms: SVM and GLM.

8.2 Automatic Data Preparation (ADP)

Data used for building a model must be properly prepared. Different algorithms have different input requirements. For example, Naive Bayes requires binned data.

ADP transforms the build data according to the requirements of the algorithm, embeds the transformation instructions in the model, and uses the instructions to transform the test or scoring data when the model is applied.

If you are connected to Oracle Database 12*c*, ADP prepares text data.

See Also: *Oracle Database PL/SQL Packages and Types Reference*

8.2.1 Numerical Data Preparation

Here are some examples of how ADP prepares numerical data:

- For algorithms that require binned data (such as Naive Bayes), ADP performs supervised binning. Supervised binning is a special binning approach that takes into account the target to find good cut-points in the predictor.
- For algorithms that require normalized data (such as Support Vector Machines), the numerical data is normalized.
- For algorithms that can handle untransformed data (such as Decision Tree), you can use the numerical data to find splitters in the tree with an approach similar to supervised binning.

8.2.2 Manual Data Preparation

Manual data preparation is complicated to perform because you must understand the requirements of each algorithms, and you must carry the transformations so that you can properly prepare the test data or scoring data.

You must perform manual binning for data which has business meaning, such as recoding a numeric column of ages to desired ranges like YOUTH, ADULT and so on. Otherwise, automatic data preparation is recommended.

8.3 Data Used for Model Building

Oracle Data Miner does not necessarily use all the columns in a data source when it builds models. Build Model nodes use a set of heuristics to determine whether to exclude columns from the model building process or change the mining type from numerical to categorical only:

- There are several reasons for not using a particular column for model building. If a column does not contain useful information, it is usually not used.

  The exact list of attributes used as input to build the model depends on the algorithm used to build the model. If an algorithm does not support a data type, then Oracle Data Miner does not use attributes with that data type as input.

  For models that have targets, such as Classification models, the target cannot be text.
- The same mining types are used for all models.

If you are connected to Oracle Database 12*c* Release 1 (12.1), then you can specify characteristics of Text attributes when you edit the build node.

See Also:

- "Viewing and Changing Data Usage" for more information about attributes that are used to build the models and the mining type of each attribute.
- "Specifying Text Characteristics"

8.3.1 Viewing and Changing Data Usage

You can view and change data usage for models in:

- Input Tab of Build Editor
- Advanced Settings

8.3.1.1 Input Tab of Build Editor

To edit a Build node:

1. Double-click the node or right-click the node and select **Edit.**
2. Click the **Input** tab. In the **Input** tab, the field **Determine inputs automatically (using heuristics)** is selected by default for all models. Oracle Data Miner decides which attributes to use for input and characteristics of the attributes. Oracle Data Miner also determines mining type and specifies that auto data preparation is performed for all attributes.
   After the model is run, Oracle Data Miner generates **Rules** describing changes that it made, such as excluding an attribute or changing the mining type. To see detailed information about the heuristics, click **Show.**

Note: You cannot view and edit data usage for an Association model using these steps.

See Also:

- "Automatic Input"
- "Manual Input"

8.3.1.1.1 Automatic Input

Oracle Data Miner does not use attributes that do not provide useful information. For example, attributes that are almost constant may not be suitable for input.

After the node runs, rules describe the heuristics used. Click **Show** to see detailed informations.

8.3.1.1.2 Manual Input

To specify inputs manually, deselect **Determine inputs automatically (using heuristics).** You can make the following changes:

- To ignore an attribute: If you do not want to use an attribute as input, go to the **Input** column and click the output icon. Select the ignore icon and click **OK.** The attribute will not be used. It will be ignored.

Similarly, to use an attribute that you have ignored, click in the Input column and select . The attribute is used in model build.

- To change mining type of an attribute: Go to the **Mining Type** column and select an option from the drop-down list:
  - Numerical
  - Categorical

  Text mining types are Text and Text Custom. Select **Text Custom** to create a column- level text specification.

- To manually prepare data: By default, Automatic Data Preparation (ADP) is performed on all attributes. If you do not want Automatic Data Preparation performed for an attribute, then deselect the corresponding check box for that attribute in the **Auto Prep** column. If you turn off **Auto Prep,** then you are responsible for data preparation for that attribute.

Note: If the mining type of an attribute is Text or Text Custom, then you cannot deselect ADP.

See Also: "Automatic Data Preparation (ADP)"

8.3.1.2 Advanced Settings

To view which columns are selected by Oracle Data Miner and what mining type is assigned to each selected column, follow these steps:

Note: You cannot view and edit data usage for an Association Model using these steps.

1. Connect the Data Source node to the Model node.
2. Right-click the Model node and select **Run.**
3. Open the **Advanced Settings** dialog box in one of these ways:
   - After the model build completes, right-click the Model node and select **Edit.**

     The **Edit** dialog box opens. Click **Advanced.**
   - After the model build completes, right-click the Model node and select **Advanced Settings.**
4. The A**dvanced Settings** has two grids:
   - The **Model Settings** grid: The grid at the top lists the models built by the node.
   - The lower part of the dialog box is a tabbed display of the following:
     - **Data Usage:** The **Data Usage** tab displays information about which columns are selected for Model build, the mining type used for model building for each column, Data Type, Input, Auto Data Prep, and Rules. To view the details about the rules (Heuristics), click **Show.**
     - **Algorithm Settings**
     - **Performance Settings**

5. To view which columns are used as input for the Model build, select the model. In the **Data Usage** tab, the attributes used in model build rules indicate the heuristics applied to the attribute. For example, the mining type may be changed. For details, click **Show**.

6. You can change data usage information on a per-model basis, or you can change the data usage for several models at the same time.

   To change data usage for several models, select the models by pressing the Ctrl key and clicking it simultaneously. Make the changes and click **OK.** The changes are made to the data usage for all selected models.

> **Note:** You can also turn **Auto Data Prep** *off*. This is not recommended. If you turn Auto Data Prep *off*, then you must ensure that the input is properly prepared for each algorithm.

See Also: "Advanced Settings Overview"

8.3.2 Specifying Text Characteristics

If you are connected to Oracle Database 12*c* Release 1 (12.1), the **Text** tab in the **Edit Model Build** dialog box enables you to specify text characteristics.

If you specify text characteristics on the **Text** tab, then you are not required to use the Text nodes.

> **Note:** If you are connected to Oracle Database 11*g* Release 2 (11.2) or earlier, then use Text nodes. The **Text** tab is not available in Oracle Database 11*g* Release 2 and earlier.

Text is available for any of the following data types: `CHAR, VARCHAR2, BLOB, CLOB, NCHAR,` or `NVARCHAR2.`

To examine or specify text characteristics for data mining, either double-click the build node or right-click the node and select **Edit** from the context menu. Click the **Text** tab.

The **Text** tab enables you to modify the following:

- **Categorical cutoff value:** Enables you to control the cutoff used to determine whether a column should be considered as a Text or Categorical mining type. The cutoff value is an integer. It must be 10 or greater and less than or equal to 4000. The default value is 200.
- **Default Transform Type:** Specifies the default transformation type for column-level text settings. The values are:
  - **Token** (Default): If Token is selected, the **Default Settings** are as follows:
    - **Languages:** Specifies the languages used in the documents. The default is `English.` To change this value, select an option from the drop-down list. You can select more than one language.
    - **Stemming:** By default, this option is not selected.
      Not all languages support stemming. If the language selected is English, Dutch, French, German, Italian, or Spanish, then stemming is automatically enabled.
      If Stemming is enabled, then stemmed words are returned for supported languages. Otherwise the original words are returned.

  * **Stoplists:** Specifies the stoplist to use. The default setting is to use the default stoplist. You can add stoplists or edit stoplists.
    If you select more than one language and the selected stoplist is `Default,` then the default stop words for languages are added to the default stoplist from the repository. No duplicate stop words are added.
  * **Tokens:** Specifies the maximum number of tokens across all documents. The default is 3000.
- **Theme:** If Theme is selected, then the **Default Settings** are as follows:
  * Language: Specifies the languages used in the documents. The default is Arabic. To change this value, select one from the drop-down list. You can select more than one language.
  * Stoplists: Specifies the stoplist to use. The default setting is to use the default stoplist. You can add stoplists or edit stoplists.
    If you select more than one language and the selected stoplist is `Default,` then the default stop words for languages are added to the default stoplist (from the repository). No duplicate stop words are added.
  * Themes: Specifies the maximum number of themes across all documents. The default is 3000.

- Click **Stoplists** to open the **Stoplist Editor.** You can view, edit, and create stoplists.

  You can use the same stoplist for all text columns.

  See Also:

  - "Oracle Text Concepts"
  - "Stoplist Editor"
  - "Text Nodes"

8.4 Model Nodes Properties

You can view the Properties of a Model Build node in any one of the following ways:

- Select the node and go to **View** and click **Properties.** Click the **Properties** tab if necessary.
- Right-click the node and select **Go to Properties** from the context menu.

Properties for Model nodes have the following sections:

- Models
- Build
- Test (Classification and Regression nodes only)
- Details

In earlier releases, **Properties** was called Property Inspector.

8.4.1 Models

The **Models** section displays a list of the models defined in the node. The default setting is to build one model for each algorithm supported by the node.

For each model, the name of the model, build information, the algorithm, and comments are listed in a grid. The Build column shows the time and date of the last successful build or if the model is not built or did not build successfully.

You can add, delete, or view models in the list. You can also indicate in which models are passed to subsequent nodes or not.

- To delete a model from the list, select it and click .
- To add a model, click . The **Add Model** dialog box opens.
- To view a model that was built successfully, select the model and click .

You can tune classification models from **Properties** pane.

See Also:

- "Output Column"
- "Add Model"
- "Tuning Classification Models"

8.4.1.1 Output Column

The Output column in the **Model Settings** grid controls passing models to subsequent nodes. The default setting is to pass all models to subsequent nodes.

- To ignore a model, that is, to *not* pass it to subsequent nodes, click . The Output icon is replaced with the Ignore icon .
- To cancel the *ignore,* click the **Ignore** icon again. It becomes the output icon.

8.4.1.2 Add Model

In the **Add Model** dialog box, you can add a model to a node.

To add a model to a note:

1. In the **Algorithm** field, select an algorithm from the drop-down list. For example, if you add a model to a clustering node, then the available algorithm are *k*-Means and O-Cluster.
   A default model name is displayed. You can change the default model.
2. In the **Comments** field, add your comments, if any. This is an optional field.
3. Click **OK.**

8.4.2 Build

For models that have a target (Classification and Regression), the targets are listed. All models in a node have the same target. This section displays the following:

- **Target:** Displays the target. To change the target, select a new target from the drop-down list.
- **Case ID:** Displays the case ID of the model defined in this node. All the models in the node have the same case IDs. To edit the case IDs, select a different case ID from the drop-down list.
- **Transaction ID:** Displayed for Association models only. To change the transaction ID, click **Edit.**
- **Item ID:** Displayed for Association models only. To change the value, select an option from the drop-down list.
- **Item Value:** Displayed for Association models only. To change the value, select an option from the drop-down list.

8.4.3 Test

The Test section is displayed for Classification and Regression models. They are the only models that can be tested.

The Test section defines how tests are done. By default, all models are tested. All models in the node are tested in the same way.

See Also:

- "Classification Node Properties"
- "Regression Node Properties"

8.4.4 Details

The Details section shows the node name and comments about the node. You can change the name of the node and edit the comments from this section. The new node name and comments must meet the requirements in Node Name and Node Comments.

8.5 Anomaly Detection Node

An Anomaly Detection node builds one or more models that detect rare occurrences, such as fraud, using the One-Class SVM algorithm.

There are two ways to detect anomalies:

- Build and apply an Anomaly Detection model.
- Use an Anomaly Detection Query, one of the Predictive Query nodes.

An Anomaly Detection build can run in parallel.

The following topics describe Anomaly Detection Nodes:

- Default Behavior for Anomaly Detection Node
- Create Anomaly Detection Node
- Data for Anomaly Detection Build
- Edit Anomaly Detection Node
- Advanced Model Settings
- Anomaly Detection Node Properties
- Anomaly Detection Node Context Menu

See Also:

- "Anomaly Detection"
- "About Parallel Processing"
- "Anomaly Detection Query"

8.5.1 Default Behavior for Anomaly Detection Node

An Anomaly Detection node builds one model using the one-class SVM algorithm by default. All models in the node have the same case ID.

8.5.2 Create Anomaly Detection Node

First create a workflow and then identify or create a Data Source node. The input for a Model node is any node that generates data as an output, including Transform nodes and Data nodes.

Note: If the data includes text columns, then prepare the text columns using a Build Text node. If you are connected to Oracle Database 12*c*, then use ADP.

To create an Anomaly Detection node:

1. In the **Components** pane, click **Workflow Editor.** If the **Components** pane is not visible, then go to **View** and click **Components.**
2. In the **Workflow Editor,** expand **Models,** and click **Anomaly Detection.**
3. Drag and drop the Anomaly Detection node in to the **Workflow** pane.

Note: The GUI displays that the node has no data associated with it.

---

4. Move to the node that provides data for the build. Right-click and click **Connect.** Drag the line to the Anomaly Detection node and click again.
5. You can also specify a case ID, edit the data usage, and change the algorithm settings. To perform any of these tasks, right-click the node and select **Edit**.
6. The node is now ready to build. Right-click the node and click **Run.**

See Also: "Edit Anomaly Detection Node"

8.5.3 Edit Anomaly Detection Node

In the **Edit Anomaly Detection Node** dialog box, you can specify or change the characteristics of the models to build.

To open the **Edit Anomaly Detection Node** dialog box, either double-click an Anomaly Detection node, or right-click an Anomaly Detection node and select **Edit**.

The **Edit Anomaly Detection Node** dialog box has the following tabs:

- Build (AD)
- Input
- Text

See Also:

- "Viewing and Changing Data Usage" for more information about the **Input** tab.
- "Specifying Text Characteristics" for more information about the **Text** tab.

8.5.3.1 Build (AD)

The **Build** tab for Anomaly Detection lists the models to be built and the case ID. Specify the following:

1. Select the **Case ID**. Select one attribute from the **Case ID** list. This attribute must uniquely identify a case.

   > **Note:** A case ID is not required. However, a case ID helps ensure build and test repeatability.
   >
   > A case ID is required to generate GLM diagnostics.

   If you specify a case ID, then all models in the node have the same case ID.

2. In the **Models Settings** list, specify the models you want to build. You can also perform the following tasks:

   - To add a model, click . The **Add Model** dialog box opens.
   - To edit a model, select the model and click . The **Advanced Model Settings** dialog box opens.
   - To delete a model, select the model and click .
   - To copy an existing model, select the model and click .

3. To complete the node definition, **OK.**

   See Also:

   - "Add Model (AD)"
   - "Advanced Model Settings"

8.5.3.2 Input

The **Input** tab specifies the input for model build.

Determine inputs automatically (using heuristics) is selected by default for all models. Oracle Data Miner decides which attributes to use for input. For example, attributes that are almost constant may not be suitable for input. Oracle Data Miner also determines mining type and specifies that auto data preparation is performed for all attributes.

After the node runs, rules are displayed explaining the heuristics. Click **Show** for detailed information.

You can change these selections. To do so, deselect **Determine inputs automatically (using heuristics)**.

> **See Also:** "Data Used for Model Building"

8.5.4 Data for Anomaly Detection Build

Oracle Data Miner uses heuristics to determine the attributes of the input data used for model build and also to determine the mining type of each attribute.

> **See also:** "Data Used for Model Building"

8.5.5 Advanced Model Settings

To change or view advanced settings, right-click the node and select **Advanced Settings.**

The **Advanced Settings** dialog box lists all the models in the **Model Settings** section in the upper pane. You can add and delete models from the node.

- To delete a model, select it and click .
- To add a model, click . The **Add Model (AD)** dialog box opens.
- To modify data usage of a model, select the model in the upper pane. Make the necessary modifications in the **Data Usage** tab.
- To modify the default algorithm, select the model in the upper pane. Make the necessary changes in the **Algorithm Settings** tab.

See Also:

- "Advanced Settings Overview"
- "Algorithm Settings for AD"
- "Viewing and Changing Data Usage"

8.5.6 Anomaly Detection Node Properties

The Anomaly Detection node Properties enables you to view and change information about the models defined in the node.

To view the properties for an Anomaly Detection node, select the node.

If the **Properties** pane is closed, then go to **View** and click **Properties.** Alternately, right-click the node and select **Go to Properties** from the context menu.

Anomaly Detection **Properties** pane has the following sections:

- Models (AD)
- Build (AD)
- Details

See Also: "Properties"

8.5.6.1 Models (AD)

The **Models** section displays a list of the models defined in the node. The default is to build one model.

For each model, the name of the model, the build information, the algorithm, and comments are listed in a grid. The Build column shows the time and date of the last successful build or if the model is not built or did not build successfully.

You can add, delete, or view models in the list. You can also indicate in the which models are passed to subsequent nodes or not, as described in Output Column (AD).

- To delete a model, select it and click .
- To add a model, click . The **Add Model (AD)** model dialog box opens.
- To view a model, click . The appropriate model viewer opens.
- To duplicate a model, select a model to duplicate and click .

8.5.6.2 Output Column (AD)

The Output column in the Model Settings grid controls passing models to subsequent nodes. The default is to pass all models to subsequent nodes.

- To ignore a model, click . The Output icon is replaced with the Ignore icon.
- To cancel the ignore, click the Ignore icon again. The icon changes to the Output icon.

8.5.6.3 Add Model (AD)

In the **Add Model** dialog box, you can add or change a model for the node. The algorithm is already selected for you. To add a model:

1. In the **Algorithm** field, the selected algorithm is displayed. You can change this and select a different algorithm from the drop-down list.
2. In the **Name** field, enter a name for the model.
3. In the **Comments** field, add your comments, if any. This is an optional field.
4. Click **OK.**

8.5.6.4 Build (AD)

The **Build** section displays the case ID for the models defined in this node. All of the models in the node have the same case ID.

To change the case ID, select a different attribute from the list.

8.5.7 Anomaly Detection Node Context Menu

Right-click an Anomaly Detection node. The following options are available in the context menu:

- Connect
- Edit. Opens Edit Anomaly Detection Node.
- Advanced Settings. Opens Advanced Model Settings.
- Validate Parents
- Run
- Force Run
- View Models. Opens Anomaly Detection Model Viewer for the selected model.
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed only if there are validation errors.
- Navigate

8.6 Association Node

The Association node defines one or more Association models. To specify data for the build, connect a Data Source node to the Association node.

All models in an Association node have the same input data.

Note: The data for an Association model must be in transactional format.

Association models could generate a very large number of rules with low confidence and support, or they could generate no rules at all.

An Association build can run in parallel.

This section contains the following topics:

- Behavior of the Association Node
- Create Association Node
- Edit Association Build Node
- Advanced Settings for Association Node
- Association Build Properties
- Association Node Context Menu

See Also:

- "About Parallel Processing"
- "Troubleshooting AR Models"

8.6.1 Behavior of the Association Node

By default, an Association node builds one model using the Apriori algorithm. The Apriori algorithm assumes the following:

- The data is transactional.
- The data has many missing values. The apriori algorithm interprets all missing values as sparse data, and it has its own mechanisms for handling sparse data.

All models in the node have the same case ID, item ID, and item value. The case ID can be two columns. For example, the data sources `SH.SALES`, `CUST_ID` and `TIME_ID` combined can be the case ID.

No automatic data preparation is done for an Association node. If you select a value for **Item Value** that is different from the default `<Existence>`, you might have to prepare the data.

See Also: "Data for AR Models"

8.6.2 Create Association Node

First, create a workflow and identify or create a data source.

Note: The data used to build an Association model must be in transactional format.

To create an Association node:

1. In the **Components** pane, select **Workflow Editor.**
   If the **Components** pane is not visible, then go to **View** and click **Components.**
2. In the **Workflow Editor,** expand **Models,** and click **Association.**

3. Drag and drop the node from the **Components** pane to the **Workflow pane.** The GUI shows that the node has no data associated with it and it cannot be built.
4. Move to the node that provides data for the build. Right-click the node and click **Connect.** Drag the line to the Association node and click again.
5. The **Edit Association Build Node** opens.
6. For an Association node, specify the following:
   - **Transaction ID:** Click to insert one or more Transaction IDs.
   - **Item ID:** Select an option from the drop-down list.
   - **Value:** Existence (default)
7. Click **OK.**
8. After you finish the node definition, the node is ready for build. Right-click the node and click **Run.**

8.6.3 Edit Association Build Node

The Association Build Node editor enables you to specify or change the characteristics of the models to build. To open the **Edit Association Build Node** dialog box, either double-click an Association node, or right-click an Association node and select **Edit**.

Specify the following for all model nodes:

- **Transaction IDs:** These are a combination of attributes that uniquely identifies a transaction. To specify a transaction ID, click . The **Select Columns (AR)** opens. Move one or more attributes from the **Available Attributes** list to the **Selected Attributes** list. Click **OK**.
- **Item ID:** Identifies an item. Select an attribute from the list.
- **Item Value:** `Existence` (default). You can select an attribute from the drop-down list. This is an optional field.

  The item value column may specify information such as the number of items (for example, three apples) or the type of the item (for example, Macintosh apples).

  If you select an attribute from the list, then the attribute must have less than 10 distinct values. The default value for the maximum distinct count is 10. You can change the value in Model Build Preferences for Association.

  Note: If you specify an attribute for Item Value, then you might have to prepare the data.

You can perform the following tasks:

- Add a model: Click . The **Add Model** dialog box opens.
- Delete a model: Select the model and click .
- Edit a model: Select the model and click . The **Advanced Settings for Association Node** dialog box opens. Here, you can specify Model settings or Algorithm settings.
- Copy an existing model: Select the model and click .

At this point, you can click **OK** to finish the model definition.

See Also:

- "Add Model (AR)"
- "Advanced Settings for Association Node"
- "Model Build"

8.6.3.1 Select Columns (AR)

To select attributes:

1. Select one or more attributes in the **Available Attributes** list.
2. Use the arrows between the lists to move the selections to the **Selected Attributes** list.
3. Click **OK.**

8.6.4 Advanced Settings for Association Node

You can access the advanced settings for an Association node:

- Click in the **Edit Association Build Node** dialog box.
- Right-click the node and click **Advanced Settings.**

This dialog box enables you to add or delete models and to modify the default algorithm settings for each model.

The upper pane of the dialog box lists all the models in the node. You can add and delete models.

- To delete a model, select it and click .
- To add a model, click . The **Add Model (AR)** dialog box opens.
- To change algorithm settings, select a model in the upper pane. In the Algorithm Settings tab, you can change maximum rule length, minimum confidence, and minimum support.

Note: It is possible for an Association model to generate a very large number of rules or no rules at all.

See Also:

- "Advanced Settings Overview"
- "Algorithm Settings"

8.6.5 Association Node Context Menu

Right-click an Association node. The following options are available in the Association node context menu:

- Connect
- Edit. Opens the Edit Association Build Node.
- Advanced Settings. Opens Algorithm Settings for AR.
- Run
- Force Run

- View Models. Opens AR Model Viewer for the selected model.
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed only if there are validation errors.
- Navigate

8.6.6 Association Build Properties

The Association Build node properties enables you to view and change information about the models defined in the node.

To view the properties of a node, click the node.

If the **Properties** pane is closed, then go **View** and click **Properties.** Alternately, right-click the node and select **Go to Properties.**

Association Build node **Properties** pane has the following sections:

- Models (AR)
- Build (AR)
- Details

See Also: "Properties"

8.6.6.1 Models (AR)

The **Models** section displays a list of the models defined in the node. The default is to build one model.

For each model, the name of the model, build information, the algorithm, and comments are listed in a grid. The Build column shows the time and date of the last successful build or if the model is not built or did not build successfully.

You can add, delete, or view models in the list. You can also indicate which models are passed to subsequent nodes or not, as described in Output Column (AR).

- To delete a model from the list, select it and click .
- To add a model, click . The Add Model (AR) dialog box opens.
- To view a model that is built successfully, click . The appropriate model view opens.
- To make a copy of a model, select the model and click .

8.6.6.2 Add Model (AR)

The algorithm is already selected for you. To add a model to the list:

1. Accept or change the model name.
2. In the **Comments** field, add comments, if any. This is optional.
3. Click **OK.** This adds the new model to the list. The new model has the same build characteristics as existing models. It also has the default values for advanced settings.

8.6.6.3 Output Column (AR)

The Output column in the **Model Settings** grid controls the passing of models to subsequent nodes. By default, all models are passed to subsequent nodes. You can perform the following tasks:

- To ignore a model, click . The icon changes to .
- To cancel an ignored model, click the ignore icon  again. The icon changes to the Output icon.

8.6.6.4 Build (AR)

All models in the node have the same transaction ID, item ID and item value. The **Build** section displays those for the models defined in the node:

- **Transaction IDs:** Click **Edit** to change the transaction ID.
- **Item ID:** You can select a different item ID from the drop-down list.
- **Item Value:** You can select a different item value from the drop-down list.

8.7 Classification Node

There are two ways to make classification predictions:

- By building and testing a classification model. This can be done by using a classification node, and then applying the model to the new data to make classifications.
- By using a prediction query, which is one of the predictive queries.

The Classification node defines one or more classification models to build and to test. To specify data for the build, connect a Data Source node to the Classification node.

The models in a Classification node all have the same target and case ID.

You can only specify one target.

A Classification build can run in parallel.

The section consists of the following topics:

- Default Behavior for Classification Node
- Create a Classification Node
- Data for the Classification Build
- Edit Classification Build Node
- Classification Node Properties
- Advanced Settings for Classification Models
- Classification Build Node Context Menu

See Also:

- "About Parallel Processing"
- "Prediction Query"

8.7.1 Default Behavior for Classification Node

The default behavior for Classification node is described related to the following:

- Algorithms used: For a binary target, the Classification node builds models using the following four algorithms:
  - GLM Classification Models
  - SVM Classification Models
  - Decision Tree Algorithm
  - Naive Bayes

  If the target is not binary, then GLM is not built by default. You can explicitly add a GLM model to the node.
  The models must have the same build data and same target.

  > **Note:** If do not want to create a particular model, then delete the model from the list of models. The blue check mark to the left side of the model name selects models to be used in subsequent nodes. It does not select models to build.

- Testing of models: By default, the models are all tested. The test data is created by randomly splitting the build data into a build data set and a test data set. The default ratio for the split is 60:40. That is, 60 percent build and 40 percent test. Oracle Data Miner uses compression when it creates the build and test tables when appropriate.
- Connecting nodes: You can connect both a build Data Source node and a test Data Source node to the Build node.
- Testing models: You can test Classification models using a Test node along with separate test data.
- Interpreting test results
- Tuning models: After testing a classification, you can tune each model.
- Case ID: The case ID is optional. However, if you do not specify a case ID, then the processing will be slower.

See Also:

- "Create Table Node and Compression"
- "Classification Model Test Viewer" for more information about how to interpret test result.
- "Tuning Classification Models"
- "Test Node"

8.7.2 Create a Classification Node

Before creating a Classification node, first create a workflow. Then, identify or create a data source for the Classification node.

To create a classification node and to attach data to it:

1. In the **Components** pane, go to **Workflow Editor.** If the **Components** pane is not visible, then go to **View** and click **Components.**
2. In the **Workflow Editor,** expand **Models,** and click **Classification.**
3. Drag and drop the node from the **Components** pane to the **Workflow** pane. The GUI shows that the node has no data associated with it. Therefore, it cannot be built.
4. Move to the node that provides data for the build. Right-click and click **Connect.** Drag the line to the Classification node and click again.
5. The **Edit Classification Build Node** dialog box opens. You must specify a target. All models in the node have the same target. The target cannot be text.
6. To specify a separate Data Source node for test, connect a second Data Source node to the build node. This is optional.
7. After you finish the edit operation and connect the optional test data source, the node should be ready to build. Right-click the node and select **Run** from the menu.

   If you specified a test data source, when the node runs, then the connection from the build data source is labeled **Build** and the connection from the test data source is labeled **Test**.

8.7.3 Data for the Classification Build

Oracle Data Miner uses heuristics to determine the attributes of the input data used for model build and also to determine the mining type of each attribute.

See Also: "Data Used for Model Building"

8.7.4 Edit Classification Build Node

To open the **Edit Classification Build Node** dialog box, either double-click a Classification Node, or right-click a Classification node and select **Edit.**

The **Edit Classification Node** dialog has the following three tabs:

- Build (Classification)
- Input
- Text

See Also:

- "Specifying Text Characteristics" for more information about the **Text** tab.
- "Viewing and Changing Data Usage" for more information about the **Input** tab.

8.7.4.1 Build (Classification)

The Build node enables you to specify or change the characteristics of the models to build. To edit the characteristics of the models to build, follow these steps:

1. In the **Target** field, select the target from the drop-down list. The list consist of attributes from the table or view specified in the Data Source node that is connected to the build node.

   You must specify a target. All models in the node have the same target.

2. In the **Case ID** field, select one attribute from the drop-down list. This attribute must uniquely identify a case.
   If you specify a case ID, all models in the node will have the same case ID.

> **Note:** If you do not specify a case ID, then the processing will be slower because a table must be generated.
>
> The case ID is *required* to generate GLM diagnostics.
>
> A case ID is *required* if a column in the input data is a nested column. That is, very dense and deep (lots of name-value pairs). If there is no case ID, then the sorting operations may fail.

3. In the **Models Settings** section, select which models you want to build. For a Classification node with a binary target, Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and Generalized Linear Models (GLM) models are specified by default.

   - To delete a model, select the model and click .
   - To edit a model, select the model and click .
   - To add models, click .
   - To copy an existing model, select the model to be copied and click .

By default, the model is tested using a test data set created by splitting the build data set. If you do not want to test the model in this way, go to the **Classification Node Test** section in Classification node **Properties** pane. You can instead use a Test Node and a test data source to test the model.

> **See Also:**
>
> - "Advanced Settings for Classification Models"
> - "No Case ID"

8.7.4.1.1 No Case ID

If a case ID is not supplied, Oracle Data Miner creates a table for the all the input data that contains a generated case ID using the row number. This table is used as the source to create the build and test random sample views. The generated case ID is constant for all queries. This ensures that consistent test results are generated.

8.7.4.2 Delete Models

To delete a model from the list, select it and click .

8.7.4.3 Add Models

To add a model to the list, click . The **Add Model (Classification)** dialog box opens.

8.7.4.3.1 Add Model (Classification)

In the **Add Models** dialog box:

1. In the **Algorithm** field, select an algorithm.

2. In the **Name** field, a default name is displayed. You can use the default or rename the model.
3. In the **Comments** field, you can enter comments, if any. This is an optional field.
4. Click **OK** to add the model to the node.

8.7.5 Advanced Settings for Classification Models

The **Advanced Settings** dialog box enables you to inspect and change the following:

- Data usage
- Algorithm settings
- Performance settings

To change or view advanced settings, click  in the **Edit Classification Build Node** dialog box dialog box. Alternately, right-click the Classification Build node and click **Advanced Settings.**

The Advanced Settings dialog box lists all of the models in the node in the upper pane. You can add models and delete models in the upper pane of the dialog box.

In the lower pane, you can view or edit the following for the model selected in the upper pane:

- Data Usage
- Algorithm Settings

  The settings that can be changed depend on the algorithm:

  – Decision Tree Algorithm Settings
  – GLM Classification Algorithm Settings
  – Naive Bayes Algorithm Settings
  – SVM Classification Algorithm Settings
- Performance Settings

See Also:

- "Advanced Settings Overview"
- "Add Models"
- "Delete Models"
- "Edit Classification Build Node"

8.7.6 Classification Node Properties

The Classification node properties enables you to view and change information about model build and test.

Specify a target before building Classification models. You can specify a case ID. If you do not specify a case ID, then the processing will be slower.

If you are unable to view Properties, then go to **View** and click **Properties.** Alternately, right-click the node and click **Go to Properties.**

Classification node **Properties** pane has these sections:

- Classification Node Models

- Classification Node Build
- Classification Node Test
- Details

See Also: "Properties"

8.7.6.1 Classification Node Models

The Classification node lists the models that are built when the node runs. By default, the Classification Build node creates three classification models. Each one uses a different classification algorithm:

- Support Vector Machine (SVM)
- Naive Bayes (NB)
- Decision Tree (DT)
- Generalized Linear Models (GLM). This algorithm is used as default, only if the target is binary. For multi-class targets, you can also specify the GLM algorithm if you add a model.

Model Setting lists the models that are built.

You can perform the following tasks:

- **Add:** To add a model, click . The **Add Model** dialog box opens.
- **Delete:** To delete the models, select it and click .
- **Compare Test Results:** If models were tested, then you can compare test results by selecting two or more models and clicking .
- **View:** If a model built successfully, then you can view the model by selecting the model and clicking . The Model viewer depends on the algorithm used to create the model.
  - GLM Classification Model Viewer
  - Decision Tree Model Viewer
  - Naive Bayes Model Viewer
  - SVM Classification Model Viewer
- **Duplicate:** To copy a model, select the model and click .
- **Tune Models:** To tune models, select the model and click .

You can also indicate which models are passed to subsequent nodes or not.

See Also:

- "Add Model (Classification)"
- "Tuning Classification Models"
- "Classification Node Output Column"

8.7.6.1.1 Classification Node Output Column The Output column in the Model Settings grid controls the passing of models to subsequent nodes. By default, all models are passed to subsequent nodes.

- To ignore a model, that is, not to pass it to subsequent nodes, click . The icon changes to the Ignore icon

- To cancel the ignore, click the Ignore icon again. It changes to the output icon.

8.7.6.2 Classification Node Build

The Build section displays the target and the case ID. The Build node must be connected to a Data Source node. You can perform the following tasks:

- **Target:** You can select a target from the **Target** drop-down list.
- **case ID:** To change or select a case ID, select one attribute from the case ID drop-down list. This attribute uniquely identifies a case.
  case ID is an optional field. If you do not select a case ID, then the processing will be slower.

8.7.6.3 Classification Node Test

The Test section specifies the data used for test and which tests to perform.

You can set the following settings:

- **Perform Test:** Select this option to test the Classification Node. The default setting is to test all models built using the test data that is created by randomly splitting the build data into two subsets. By default, the following tests are performed:
  - Performance Metrics
  - Performance Matrix
  - ROC Curve (Binary Class only)
  - Lift and Profit
    Lift and profit for the top 5 target classes by frequency. Click **Edit.** The **Target Values Selection** dialog box opens.
  - Generate selected Test Results for Tuning: If you plan to tune the models, then you must test the models in the build node, not in a Test node.
- **Test Data:** Select any one of the following options, by which Test Data is created:
  - Use all Mining Build Data for Testing
  - Use Split Build Data for Testing
    Split for Test (%)
    Create Split as: `Table` (default)
  - Use a Test Data Source for Testing: Select this option to connect the Test Data Source to the Build node, after you connect the Build data.

    Note: Another way to test a model is to use a Test Node.

    See Also:

    - "Testing Classification Models"
    - "Target Values Selection"

8.7.6.3.1 Target Values Selection

The **Target Values Selection** dialog box displays the number of target values selected. The default option Automatic is to use the top five target class values by frequency. You can change the number of target values by changing the frequency count. You can also select the option Use Lowest Occurring.

- **Automatic:** By default, use the top five target class values by frequency.

  – **Frequency Count:** You can change the number of target values by changing the values in this value.
  – **Use Lowest Occurring**
  – **Use Highest Occurring**
- **Custom:** Select this option to specify specific target values. Then, move the values from **Available Values** to **Selected Values.**

8.7.7 Classification Build Node Context Menu

To view the context menu, right-click a Classification node. The following options are available in the context menu:

- Connect
- Edit. Opens the Edit Classification Build Node.
- Advanced Settings. Opens Advanced Settings for Classification Models.
- Validate Parents
- Run
- Force Run
- View Models
- View Test Results
- Compare Test Results
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed only if there are validation errors.
- Navigate

8.7.7.1 View Test Results

Select a model and then view the test results for the model.

See Also: "Compare Classification Test Results"

8.7.7.2 Compare Test Results

You can compare all successfully built models in the node by comparing the text results.

See Also: "Compare Test Results Viewer"

8.8 Clustering Node

A Clustering node builds clustering models using the *k*-Means, O-Cluster, and Expectation Maximization algorithms.

Note: Expectation Maximization models require Oracle Database 12*c* Release 1 (12.1) or later.

There are two ways to cluster data:

- By building a Clustering model: Use a Classification node. Then apply the model to new data to create clusters.
- By using a Clustering Query, which is one of the predictive queries.

A Clustering build can run in parallel.

This section contains the following topics:

- Default Behavior for Clustering Node
- Create Clustering Build Node
- Data for Clustering Build
- Edit Clustering Build Node
- Advanced Settings for Clustering Models
- Clustering Build Node Properties
- Clustering Build Node Context Menu

Tip: "About Parallel Processing"

8.8.1 Default Behavior for Clustering Node

A Clustering node builds three models using the following algorithms:

- k-Means Algorithm (KM)
- Orthogonal Partitioning Clustering (OC)
- Expectation Maximization (EM). For EM, Oracle Database 12*c* Release 1 (12.1) is required.

A case ID is optional.

The models all have the same build data.

Note: If do not want to create a model, then delete the model from the list of models. The blue check mark to the left of the model name selects models to be used in subsequent nodes, such as Apply. It does *not* select models to build.

8.8.2 Create Clustering Build Node

Before creating a Clustering Build node, first create a workflow. Then, identify or create a Data Source node.

To create a Clustering node and attach data to it:

1. In the **Components** pane, go to **Workflow Editor.**
   If the **Components** pane is not visible, then go to **View** and click **Components.**
2. In the **Workflow Editor,** expand **Models,** and click **Clustering.**
3. Drag and drop the node from the **Components** pane to the **Workflow** pane.
   The GUI shows that the node has no data associated with it. Therefore, it cannot be built.
4. Move to the node that provides data for the build. Right-click the node and click **Connect.** Drag the line to the Classification node and click again.
5. Right-click the Clustering node and click **Run.** The node runs and builds the models.

8.8.3 Data for Clustering Build

Oracle Data Miner uses heuristics to:

- Determine the attributes of the input data used for building models
- Determine the mining type of each attribute

See Also: "Data Used for Model Building"

8.8.4 Edit Clustering Build Node

To open the **Edit Clustering** dialog box, double-click a Clustering node. Alternately, you can right-click a Clustering node and select **Edit**.

The **Edit Clustering Node** dialog box has three tabs:

- Build (Clustering)
- Input
- Text

See Also:

- "Viewing and Changing Data Usage" for more information about the **Input** tab.
- "Specifying Text Characteristics" for more information about the **Text** tab.

8.8.4.1 Build (Clustering)

The **Build** tab enables you to specify or change the characteristics of the models to build. To edit the characteristics of the models to build, follow these steps:

1. In the **Case ID** field, select an attribute from the drop-down list.This attribute must uniquely identify a case.

   Note: A case ID is not required. However, a case ID helps ensure build and test repeatability.

   If you specify a case ID, then all models in the node have the same case ID.
2. In the **Models Settings** list, select the models you want to build. For a Clustering node, you can build models using the following algorithms:
   - k-Means (KM)

- Orthogonal Partitioning Clustering (OC)
- Expectation Maximization (EM).
  For this algorithm, Oracle Database 12*c* release 1 (12.1) is required.

You can perform the following tasks:

- Delete: To delete any models, select the models and click .
- Add: To add a model, click .
- Copy: To copy a model, select the model and click .

See Also:

- "Specifying Text Characteristics"
- "Delete Model (Clustering)"

8.8.4.2 Delete Model (Clustering)

To delete a model from the list, select it and click .

8.8.4.3 Add Model (Clustering)

To add a model to the list, click . The **Add Model** dialog box opens.

In the **Add Model** dialog box:

1. In the **Algorithm** field, select an algorithm, either KM, OC or EM. For EM Oracle Database 12*c* Release 1 (12.1) is required.
   - k-Means
   - Orthogonal Partitioning Clustering
   - Expectation Maximization. For this option, Oracle Database 12*c* Release 12.1 is required.
2. In the **Name** field, a default name is displayed. You can use the default name or rename the model.
3. In the **Comment** field, enter comments, if any. This is an optional comment.
4. Click **OK** to add the model to the node.

8.8.5 Advanced Settings for Clustering Models

To access advanced settings, click in the **Edit Clustering Build Node** dialog box. Alternately, right-click the node and select **Advanced Settings.** The **Advanced Settings** dialog box list all the models in the upper pane.

You can perform the following tasks:

- Inspect and change the data usage and algorithm
- Add models to the node
- Delete models from the node

In the lower pane, you can view and edit the following, for the model selected in the upper pane:

- Data usage on the **Data Usage** tab
- Algorithm settings on the **Algorithm Settings**tab

  The settings that can be changed depend on the algorithm:

- KM Algorithm Settings
- OC Algorithm Settings
- EM Algorithm Settings

See Also: "Advanced Settings Overview"

8.8.6 Clustering Build Node Properties

Clustering Build node properties enables you to view and change information about model build. If you are unable to view the **Properties** pane, then go to **View** and click **Properties.** Alternately, right-click the node and click **Go to Properties.**

The Clustering Build node properties has these sections:

- Models (Clustering)
- Build (Clustering)
- Details

8.8.6.1 Models (Clustering)

Models lists the models that are built when the nodes are run. the default is to build two clustering models using the KM, OC, and EM algorithms.

The **Model Settings** grid lists the models in the node. You can perform the following tasks:

- Search models
- Add Model (Clustering)
- Delete Model (Clustering)
- View Models
- Indicate which models are passed on to subsequent nodes.

See Also: "Clustering Node Output Column"

8.8.6.1.1 Clustering Node Output Column The Output column in the **Model Settings** grid controls the passing of models to subsequent nodes. By default, all models are passed to subsequent nodes.

- To ignore a model. that is, to *not* pass it to subsequent nodes, click . The Output icon changes to .
- To cancel the ignore, click the Ignore icon again. The icon changes to the Output icon.

8.8.6.1.2 View Models If a model built successfully, then you can view the model by selecting the model and clicking . The **Model Viewer** opens. The **Model Viewer** depends on the algorithm used to create the model:

- KM Model Viewer
- OC Model Viewer

8.8.6.2 Build (Clustering)

Displays the optional case ID for the clustering models. To change the case ID, select an attribute from the list.

8.8.7 Clustering Build Node Context Menu

To view the Clustering Build node context menu, right-click the node. The following options are available in the context menu:

- Connect
- Edit. Opens the Edit Association Build Node.
- Validate Parents
- Advanced Settings. Opens Advanced Settings for Association Node.
- Run
- View Models. Opens the appropriate viewer (KM Model Viewer or OC Model Viewer) for the selected model.
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log. Displayed only if the running of the node fails.
- Navigate

8.9 Feature Extraction Node

A Feature Extraction node uses the Nonnegative Matrix Factorization (NMF) algorithm, to build models. There are two ways to extract features:

- Build a feature extraction model, using a Feature Extraction node.
- Use a Feature Extraction Query, one of the predictive queries.

If Oracle Data Miner is connected to Oracle Database 12*c* Release 1 (12.1), a Feature Extraction node uses the PCA and SVD algorithms to build models.

> **Note:** Principal Components Analysis and Singular Value Decomposition models require Oracle Database 12*c* Release 1 (12.1).

A Feature Extraction Build can run in parallel.

This section contains the following topics:

- Default Behavior for Feature Extraction Node
- Create Feature Extraction Node
- Data for Feature Extraction Build
- Edit Feature Extraction Build Node
- Advanced Settings for Feature Extraction
- Feature Extraction Node Properties

- Feature Extraction Node Context Menu

See Also:

- "Singular Value Decomposition and Principal Components Analysis"
- "About Parallel Processing"

8.9.1 Default Behavior for Feature Extraction Node

By default, a Feature Extraction node builds one model using the Non-Negative Matrix Factorization (NMF) algorithm.

If you are connected to Oracle Database 12*c*, the node builds two models by default:

- NMF model
- PCA model

You can add SVD models.

All models in the node use the same build data and have the same case ID, if you specify a case ID.

See Also:

- "Nonnegative Matrix Factorization"
- "Singular Value Decomposition and Principal Components Analysis"

8.9.2 Create Feature Extraction Node

Before creating a Feature Extraction node, first, create a workflow. Then, identify or create a Data Source node.

To create a Feature Extraction node:

1. In the **Components** pane, go to **Workflow Editor.**
   If the **Components** pane is not visible, then go to **View** and click **Components.**
2. In the **Workflow Editor,** expand **Models,** and click **Feature Extraction.**
3. Drag and drop the node from the **Components** pane to the **Workflow** pane. A **Feature Build** node is added to the workflow.
   The GUI shows that the node has no data associated with it. Therefore, it cannot be built.
4. Move to the node that provides data for the build. Right-click and click **Connect.** Drag the line to the Feature Extraction node and click again.
5. You can edit the node. To edit the node, right-click the node and click **Edit.** The **Edit Feature Extraction Build Node** dialog box opens.
6. The node is ready to build. Right-click the node and click **Run.**

8.9.3 Data for Feature Extraction Build

Oracle Data Miner uses heuristics to:

- Determine the attributes of the input data used for model build
- Determine the mining type of each attribute

See Also: "Data Used for Model Building"

8.9.4 Edit Feature Extraction Build Node

To edit a Feature Build node, either double-click a Feature Build node, or right-click the node and select **Edit**. The **Edit Feature Extraction Build Node** dialog box opens. The same dialog box opens when you drop a Feature Build node on a workflow.

The Edit Feature Extraction Build dialog box has three tabs:

- Build (Feature Extraction)
- Input
- Text

See Also:

- "Viewing and Changing Data Usage" for more information about the **Input** tab.
- "Specifying Text Characteristics" for more information about the **Text** tab.

8.9.4.1 Build (Feature Extraction)

In the **Build** tab, you can perform the following tasks:

- **Case ID:** Specify case ID for Feature Extraction is optional. Specify one by selecting an attribute from the drop-down list.
- **Add Model:**
- **Delete:** To delete a model, select the model and click .
- **Copy:** To copy an existing model, select the model and click .

See Also: "Add Model (Feature Extraction)"

8.9.4.2 Add Model (Feature Extraction)

To add a model, click . The **Add Model (Feature Extraction)** dialog box opens.

1. In the **Algorithm** field, select an algorithm. The default algorithm is NMF.
2. In the **Name** field, the default name is displayed. You can accept the default name or change it.
3. In the **Comments** field, enter comments, if any. This is an optional field.
4. Click **OK.** The model is added to the list. The new model has the same build characteristics as existing models. The new model has the default values for advanced settings.

8.9.5 Advanced Settings for Feature Extraction

The advanced settings selection enables you to inspect and change the data usage and algorithm settings for each model in the node.

- Inspect and change data usage.
- Change algorithm settings for each model in the node.

To change or view advanced settings, click in the **Edit Feature Extraction Build Node** dialog box. Alternately, right-click the node and select **Advanced Settings.** The advanced settings selection enables you to inspect and change the data usage and algorithm settings for each model in the node.

In the upper pane, all models are listed. You can perform the following tasks:

- **Delete:** To delete a model, select it and click .
- **Add:** To add a model, click .

In the lower pane, you can view or edit the following for the model selected in the upper pane:

- Data Usage
- Algorithm Settings

  The settings depend on the algorithm:

  - NMF Algorithm Settings
  - PCA Algorithm Settings
  - SVD Algorithm Settings

  PCA and SVD are available if Oracle Data Miner is connected to Oracle Database 12*c* Release 1 (12.1).

  See Also:

  - "Advanced Settings Overview"
  - "Edit Feature Extraction Build Node"

8.9.6 Feature Extraction Node Properties

The Feature Extraction node properties enables you to view and change information about the models defined in the node.

To view the properties for a node, click the node. If the **Properties** pane is closed, then go to **View** and click **Properties.** Alternately, right-click the node and click **Go to Properties.**

The Feature Extraction node **Properties** pane has the following sections:

- Models
- Build (Feature Extraction)
- Details

See Also: "Properties"

8.9.6.1 Build (Feature Extraction)

The Build section displays the case ID for the models defined in this node. All the models in the node have the same case ID.

A case ID is not required.

To edit the case ID, select a different attribute from the list.

8.9.7 Feature Extraction Node Context Menu

To view the Feature Extraction node context menu, right-click the node. The following options are available in the context menu:

- Connect

- Edit. Opens the **Edit Feature Extraction Build Node** dialog box.
- Advanced Settings. Opens **Advanced Settings for Feature Extraction** dialog box.
- Validate Parents
- Run
- Force Run
- View Models. Opens **NMF Model Viewer** for the selected model.
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed only if there are validation errors.
- Navigate

8.10 Model Node

A Model node enables you to add models to a workflow that were not built in the workflow. For example, you can specify a model that was built using either of the ODM APIs.

The models in a Model node must satisfy the model constraints.

The Model node takes no input. A Model node can be an input to any node that accepts models, such as the Apply and Test nodes, at least for some function types. For example, if a model node contains Classification or Regression models, it can be input to a test node. Test data must be prepared in the same way that the build data was prepared.

This section about Model nodes contains the following topics:

- Create a Model Node
- Edit Model Selection
- Model Node Properties
- Model Node Context Menu

Model nodes rely on database resources for their definition. It may be necessary to refresh a node definition if the database resources change, for example, if the resources are deleted or re-created.

See Also:

- "Refresh Nodes"
- "Model Constraints"

8.10.1 Create a Model Node

To add a model node to a workflow and add models to the model node:

1. Open the **Components** pane and select the **Workflow Editor.** If the Components pane is not visible, then go to **View** and click **Components.**
2. In the **Workflow Editor,** expand **Models,** and click **Model.**
3. Drag and drop the node from the **Components** pane to the **Workflow** pane.
4. The **Edit Model Selection** dialog box opens automatically. The models in the Model node must have the same mining function and the same target (Classification and Regression models only).

See Also: "Edit Model Selection"

8.10.2 Edit Model Selection

In the **Edit Model Section** dialog box, you can select one or more models to include in the Model node or to remove models from the Model node. To edit the models in the node, double-click the Model node or right-click the Model node and select **Edit**.

Note: All the models in a model node must satisfy the Model Constraints.

You can perform the following tasks:

- Select models from the **Available Compatible Models** list and move them to the **Selected Models** list using the controls between the lists.
  The selected models are checked for compatibility. The models in a model node must satisfy the model constraints.
  The selected models are part of the model node. You can view the models using the Model node properties.
- Include models from other schemas. To include models, select **Include Models from Other Schemas.**
- Filter the Available Compatible Models list in the following ways:
  - Select a model function from the Model Function list. The options are:
    * All
    * Anomaly Detection
    * Association Rules
    * Regression
    * Clustering
    * Feature Extraction
  - Sort the models by name, function, algorithm, target, target data type, creation date, or comments. To sort, click the column header in the list of available models.
- Add and remove models:
  - Add models by moving them from **Available Compatible Models** list to the **Selected Models** list.

– Remove models by moving them from the **Selected Modes** list to the **Available Compatible Models** list. You can also remove models using the Models tab.

See Also:

- "Model Node Properties"
- "Model Constraints"

8.10.2.1 Model Constraints

A Model node consists of models that are similar. The models in a Model node must satisfy the following;

- All models must have the same function type (Classification, Regression, Clustering, Anomaly Detection, Association Rules, or Feature Extraction). You cannot include models that have different function types.

  You can add models that are built using different algorithms if the models have the same function type.

- Classification or Regression models must have the same target attribute. The target attributes must all have the same data type.

  `CHAR` and `VARCHAR2` are considered to be the same data type for Classification models.

- Classification models must have the same list of target values.

8.10.3 Model Node Properties

To access the properties of the Model node, click the node. If the **Properties** pane is closed, then go to **View** and click **Properties.** Alternately, right-click the node and click **Go to Properties.**

In the Model node Properties pane, you can:

- Add models to the Model node
- Delete models from the Model node
- View models in the Model node

The **Properties** pane for a model node source node has the following sections:

- Models (Model Node)
- Details

See Also: "Properties"

8.10.3.1 Models (Model Node)

The Models section shows the mining function that the models use and lists all the models included in the node in a grid.

You can search for models, add models to the node, and delete models.

You can perform the following tasks:

- **Add Models:** To add models:

  1. Click . The **Edit Model Selection** dialog box opens.

2. In the **Edit Model Selection** dialog box, select the models to add to the node. You can add models from other schemas too. However, any models that you add must be compatible with the models already in the node.
3. Click **OK.** This adds the models to the node. You can go to the **Properties** pane for the Model node to view the models.

- **Delete Models:** To delete a model, select it and click .
- **View Models:** To view a model, select it and click .
- **Refresh models:** To refresh models, click . If data on the server changes, it may be necessary to refresh the node.

  See Also: "Refresh Nodes"

8.10.4 Model Node Context Menu

To view the Model node context menu, right-click the node. The following options are available in the context menu:

- Connect
- Edit. Opens the **Edit Model Selection** dialog box.
- Run.Validates that the models specified in the node exist.
- View Models
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed only if there are validation errors.
- Navigate

8.11 Model Details Node

The Model Details nodes are the most useful for application developers. The Model Details node performs the following functions:

- Extracts model details from a Model Build node, a Model node or any node that outputs a model.
- Reveals information about model attributes and their treatment by the algorithm. The output depends on the type of models selected and the specific type of model details you specify.
- The output of the Model Details node is a data flow. To enable the data to persist, use a Create Table or View node.

A Model Details node can run in parallel.

This section on Model Detail node contains the following topics:

- Model Details Node Input and Output
- Create Model Details Node
- Edit Model Details Node
- Model Details Automatic Specification
- Model Details Node Properties
- Model Details Node Context Menu
- Model Details Per Model

See Also:

- "About Parallel Processing"
- "Create Table or View Node"

8.11.1 Model Details Node Input and Output

The input for a Model Details node is one or more of the following:

- Build node (any model type)
- Model Node

All models in Build nodes or Model nodes must have the same mining function type. For example, if one is a Classification model, then all of them must be Classification models.

The output for a Model Details node is a data flow based on the model detail specifications. To enable the data to persist, use a Create Table or View node.

See Also:

- "Create Table or View Node"
- "Default Model and Output Type Selection"

8.11.2 Create Model Details Node

To create a Model Details node, follow these steps:

1. Identify the input node or nodes for model details. The input node must be one or more of the following:
   - Any Model Build node
   - Any Model node

   Note: All the models selected must have the same mining function type. For example, if one of the nodes is a Classification node, then all other nodes must build Classification models.

2. In the **Components** pane, go to **Workflow Editor.** If the **Components** pane is not visible, then go to **View** and click **Components.**
3. In the **Workflow Editor,** expand **Models,** and click **Model Details.**
4. Drag and drop the node from the **Components** pane to the **Workflow** pane. You may want to place the Model Details node close to the Build or Model nodes that provide input.

5. Move the cursor to the workflow. Right-click one of the input nodes, and select **Connect.** Drag the link to the Model Details node. Repeat if you must connect more nodes.
6. The default specification for model details depends on the model. To use the default specification, right-click the Model Details node and click **Run**.
7. To change the specification for the Model Details node, right-click the node and select **Edit.** Alternately, you can change the specifications in the **Properties** pane of the node.

   See Also:

   - "Model Details Automatic Specification"
   - "Edit Model Details Node"

8.11.3 Edit Model Details Node

The Model Details Node editor enables you to view or specify the models details provided by the node. To open Edit Model Details Node, double-click a Model Details node. Alternately, right-click a Model Details node and select **Edit**.

You can perform the following tasks:

- **Automatic Specification:** If this option is selected (the default), then the system determines the specification. You cannot change the output types, algorithm types, or selected models.
- **Function:** Displays the function type of the input nodes connected, for example, if a Classification node is connected to Model Details, the function is Classification. If no input nodes are connected, then it is undefined.
- **Model Type:** It is a list of the algorithms available, including `All`.
- **Output:** Displays the default output of Model Details for the algorithm.
- **Column:** Click columns to view the list of the columns (name and data type) for the selected output type.
- **Add:** To add model type or edit output type, deselect **Automatic Specification**. To add another model type, select the model type and click . The **Edit Model Selection Details** dialog box opens. You can accept the default specifications or edit them.

  See Also:

  - "Model Details Automatic Specification"
  - "Edit Model Selection Details"

8.11.3.1 Edit Model Selection Details

The top pane of the **Edit Model Selection Details** dialog box contains general information:

- **Function:** Displays the function type of the input nodes connected, for example, if a Classification node is connected to Model Details, the function is Classification. If no input nodes are connected, then it is undefined.
- **Model Type:** Displays algorithms. If there are models already selected (listed in the Selected Models grid), then the Model Type field is disabled to match the already selected models. If you move all models out of the Selected Models grid,

the Model Type field is enabled again. If the Model Type is enabled, then you can select models. The default is `All Models`.

- **Output Type:** Displays the list of possible output types (model queries) that are available for the specified model types. The values for each algorithm selection are as follows:
  - Decision Tree (initial default): Full Tree (default), Full Tree XML, Leaf Nodes, Model Signature
  - SVM Classification: Coefficients (Default), Model Signature
  - SVM Regression, Coefficients (Default), Model Signature
  - Naive Bayes: Pair Probabilities (Default), Model Signature
  - Association Rules: Rules (Default), Global Details, Itemsets
  - Anomaly Detection: Coefficients (Default), Model Signature
  - GLM Classification: Statistics (Default), Row Diagnostics, Model Signature, Global Details
  - GLM Regression: Statistics (Default), Row Diagnostics, Model Signature, Global Details
  - KM or OC Clustering: Full Tree (Default), Rules, Attribute Histograms, Centroid, Model Signature
  - Expectation Maximization (EM): Full Tree (Default), Attribute Histograms, Centroid Components, Global Details, Model Signature, Projections, Rules.

    EM requires Oracle Database 12*c* Release 1 (12.1) or later.
  - NMF: Features Transactional (Default), Model Signature
  - SVD: Features Transactional (Default), Global Details, Model Signature, Projections, Singular Values

    SVD requires Oracle Database 12*c* Release 1 (12.1) or later.
  - PCA: Features Transactional (Default), Eigen Values, Global Details, Model Signature, Projections

    PCA requires Oracle Database 12*c* Release 1 (12.1) or later.

  Output values are also available for multiple model types. For example, you can select Centroid for all clustering models.
- **Columns**: Click to see a list of the columns (name and data type) for the selected output type.

The lower portion of the dialog box displays the following:

- **Available Compatible Models:** Lists the available models, that is, models that match the algorithm selection. The grid, for each model, displays the model Name, the input node for the model, and the algorithm used to build the model.
- **Selected Models:** Lists the selected models. The grid, for each model, displays the model name, the input node for the model, and the algorithm used to build the model.

8.11.4 Model Details Automatic Specification

How specifications change automatically depends on whether automatic selection is on or off:

- By default, **Automatic Specification** is selected. Automatic Specification results in the following behavior:
  - When the first input node is connected to a Model Details node, the input node is searched for models in a default order of priority. For the first model type found, all the nodes matching models are added to the Model Details Specification along with the default Output Type.
  - On subsequent connections, the models that match the type in the Model Details node are automatically added. A message is displayed telling you that models are being added automatically.
  - When an input node is disconnected, all model specifications provided by that node are automatically removed from the Model Details node.
  - When an input node is edited, any models added are automatically added to the Model Details node if the added model matches the model type contained in the node. If models are deleted from an input node, then they are deleted from the Model Details node.
  - When a parent node is edited so that all models are removed, the model node is set to undefined. When a new model is added to the parent node, the model node remains undefined because it is too unpredictable about what model and output type would be selected by default given that there may be many parent nodes connected to a model node.
  - When an input node is edited and the model is changed so that it is no longer consistent with its specification in the model details node, the model specification is removed.
- **Automatic Specification** is `Off` or deselected, and it results in the following behavior:
  - Models are not added automatically.
  - You must edit the Model Details node.
  - Validations are performed as usual, so models that are now inconsistent or missing are marked as invalid. Also, if models are missing and a node is added that contains a match with that model, then it is made valid and associated to the new node.
  - You must manually fix or remove invalid model references.

See Also:

- "Edit Model Details Node"
- "Default Model and Output Type Selection"

8.11.4.1 Default Model and Output Type Selection

The specification that is automatically added depends on the mining function of the model as follows:

- Classification
  - Decision Tree: Full Tree
  - GLM: Statistics
  - NB: Probabilities
  - SVM: LINEAR KERNEL ONLY Coefficients

- Clustering
  - KM: Full Tree
  - OC: Full Tree
  - EM: Full Tree
- Regression
  - GLM: Statistics
  - SVM: LINEAR KERNEL ONLY Coefficients
- Anomaly Detection
  - SVM: LINEAR KERNEL ONLY Coefficients
- Association
  - Apriori: Rules
- Feature Extraction
  - NMF, SVD, or PCA: Features transactional

8.11.5 Model Details Node Properties

The Model Details node properties enables you to view the model for which details were extracted. To display the **Properties** pane for a node, click the node. If the **Properties** pane is closed, then go to **View** and **Properties.** Alternately, right-click the node and click **Go to Properties.**

Model Details node **Properties** has the following sections:

- Models (Model Details)
- Output (Model Details)
- Cache (Model Details)
- Details

See Also: "Properties"

8.11.5.1 Models (Model Details)

The **Model** section lists the models that you want to save details about. You can add and remove models from the list.

8.11.5.2 Output (Model Details)

The **Output** tab lists the columns produced by the Model Details node. For each column, the alias (if any) and the data type are displayed.

See Also: "Model Details Automatic Specification"

8.11.5.3 Cache (Model Details)

The default is to *not* generate cache to optimize the viewing of results. You can generate cache. If you generate cache, then you can specify the sampling size. The default sampling size is 2000 rows.

8.11.6 Model Details Node Context Menu

To view the Model Details node context menu, right-click the node. The following options are available in the context menu:

- Connect
- Edit. Opens Edit Model Details Node.
- Validate Parents
- Run
- Force Run
- View Data (Model Details)
- Generate Apply Chain
- Deploy
- Show Graph
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed only if there are validation errors.
- Navigate

8.11.6.1 View Data (Model Details)

To view the complete Model Details output, right-click the node and select **View Data**.

The output is displayed in a multitab display:

- **Data**

  The data that constitutes the model details. What the data represents depends on the model. For example, the data could represent a tree or rules.

  You can sort and filter the columns of this tab.
- **Columns**

  Data Type and Mining Type of the columns in the output.
- **SQL**

  SQL used to generate the model details.

See Also: "Model Details Per Model"

8.11.7 Model Details Per Model

The exact data displayed in a Model Details node depends on the particular models.

All models that can be applied (scored) can have model signature as output.

See Also:

- "Default Model and Output Type Selection" for more information about the default model details.
- "Edit Model Selection Details" for more information about the possible selections for the Model Details node for each algorithm.

8.12 Regression Node

There are two ways to make regression predictions:

- By building and testing a Regression model: Use a Regression node, and then apply the model to new data to make classifications.
- By using a Prediction Query, which is one of the predictive queries.

The Regression node defines one or more Regression models to build and to test. To specify data for the build, connect a Data Source node to the Regression node. You can also connect a second data source to the Regression build node to specify test data.

The models in a Regression Node all have the same target and case ID.

You can only specify one target.

A Regression build can run in parallel.

This section consists of the following topics:

- Default Behavior for Regression Node
- Create a Regression Node
- Edit Regression Build Node
- Data for a Regression Build
- Advanced Settings for Regression Models
- Regression Node Properties
- Regression Node Context Menu

See Also:

- "Prediction Query"
- "About Parallel Processing"

8.12.1 Default Behavior for Regression Node

For a binary target, the Regression node builds four models using the following algorithms:

- Generalized Linear Model (GLM)
- Support Vector Machine (SVM)

The models have the same build data and the same target.

By default, the models are all tested. The test data is created by randomly splitting the build data into a build data set and a test data set. The default ratio for the split is 60 percent build and 40 percent test. When possible Data Miner uses compression when creating the test and build data sets.

You can instead use all the build data as test data.

To use separate test data, connect a test data source to the build node or use a Test node.

After you test models, you can view test results.

You can compare test results for two or more Regression models using the Compare Test Results selection of the context menu.

The case ID is optional. However, if you do not specify a case ID, then the processing will be slower.

See Also:

- "Create Table Node and Compression"
- "Regression Model Test Viewer"
- "Test Node"

8.12.2 Create a Regression Node

Before creating a Regression node, first, create a workflow. Then, identify or create a data source.

To create a regression node and attach data to it:

1. In the **Components** pane, go to **Workflow Editor.** If the **Components** pane is not visible, then go to **View** and click **Components.**
2. In the **Workflow Editor,** expand **Models,** and click **Regression.**
3. Drag and drop the node from the **Components** pane to the **Workflow** pane.
   The GUI shows that the node has no data associated with it. Therefore, it cannot be built.
4. Move to the node that provides data for the build. Right-click, and click **Connect.** Drag the line to the Regression node and click again.
5. The **Edit Regression Build Node** dialog box opens. You must specify a target (all models in the node have the same target). A target cannot be text.
6. To specify a separate Data Source node for test, connect a second Data Source node to the Build node. This is optional.
7. After you finish editing the node, and connecting the optional test Data Source node, the node should be ready to build. Right-click the node and click **Run.**

   If you specified a test Data Source node when the node runs, then the connection from the build data source is labeled **Build** and the connection from the test data source is labeled **Test**.

8.12.3 Data for a Regression Build

Oracle Data Miner uses heuristics to:

- Determine the attributes of the input data used for model build.
- Determine the mining type of each attribute.

See Also: "Data Used for Model Building"

8.12.4 Edit Regression Build Node

To open the **Edit Regression Build Node** dialog box, double-click a Regression Build node, or right-click a Regression Build node and select **Edit**.

The **Edit Build** dialog box has three tabs:

- Build (Regression)
- Input
- Text

See Also:

- "Viewing and Changing Data Usage" for more information about the **Input** tab.
- "Specifying Text Characteristics" for more information about the **Text** tab.

8.12.4.1 Build (Regression)

The **Build** tab enables you to specify or change the characteristics of the models to build. To edit the characteristics of the model to build, follow these steps:

1. In the **Target** field, select a target from the drop-down list. The list consist of attributes from the table or view specified in the Data Source node that is connected to the build node.

   You must specify a target. All models in the node have the same target.

2. In the **Case ID** field, select one attribute from the drop-down list. This attribute must uniquely identify a case.

   Note: A case ID is not required. However, if you do not specify a case ID, then the processing will be slower.

   A case ID is *required* to generate GLM diagnostics.

   If you specify a case ID, all models in the node have the same case ID.

3. In the **Models Settings** list, select which models you want to build. You can build Support Vector Machine (SVM) and Generalized Linear Models (GLM). You can delete any of these models by selecting the model and clicking

   - To delete any model, select the model and click .
   - To add models, click . The **Add Model (Regression)** dialog box opens.
   - To edit a model, click . The **Advanced Settings for Regression Models** dialog box opens.
   - To copy an existing model, select the model and click

4. Click **OK.**

The default is to test the model using a test data set created by splitting the build data set. If you do not want to test the model in this way, go to the Test section in of Regression node **Properties** pane. You can instead use a Test Node and a test data source to test the model.

See Also:

- "No Case ID"
- "Advanced Settings Overview"
- "Test (Regression)"

8.12.4.2 Add Model (Regression)

To add a model to the node:

1. In the **Algorithm** field, select an algorithm.
2. In the **Name** field, a default name is displayed. You can use the default or rename the model.
3. In the **Comment** field, add comments if any. This is an optional field.
4. Click **OK**. The new model is added to the node.

8.12.5 Advanced Settings for Regression Models

The **Advanced Settings** dialog box enables you to:

- Inspect and change data usage and algorithm settings for each model in the node
- Add and delete models

To change or view Advanced Settings, click in the **Edit Regression Build Node** dialog box. Alternately, right-click the node and select **Advanced Settings.**

The upper panes lists all the models in the node. You can perform the following functions:

- **Delete:** To delete a model, select the model and click
- **Add:** To add a model, click . The **Add Model (Regression)** dialog box opens.

In the lower pane, you can view and modify data usage and algorithm settings for the model selected in the upper pane. You can edit the following:

- Data Usage
- Algorithm Settings

  The settings that can be changed depend on the algorithm:

  - GLM Regression Algorithm Settings
  - SVM Regression Algorithm Settings

8.12.6 Regression Node Properties

The Regression node properties enables you to view and change information about the model builds. To view the **Properties** pane of a Regression node, click the node. If **Properties** pane is closed, then go to **View** and **Properties.** Alternately, right-click the node and click **Go to Properties.**

Before building Regression models, ensure the following:

- Specify a Target.
- Specify a case ID. This is optional. However, if you do not specify a case ID, then the processing will be slower.

This section contains the following topics:

- Models (Regression)
- Build (Regression)
- Test (Regression)
- Details

See Also: "Properties"

8.12.6.1 Models (Regression)

The **Model** section lists the models that are built. By default, three Regression models are built using three different algorithms (SVM, NB, and DT). You can also specify the GLM algorithm if you add a model.

You can perform the following tasks:

- **Delete:** To delete a model, select the model and click .
- **Add:** To add a model, click
- **Compare Test Results:** If models were tested, then you can compare test results. Select two or more models and click
- **View Model:** If a model built successfully, then you can view the model. Select the model and click . The corresponding viewer opens.
- **Indicate Models Status:** Indicates whether models are passed to subsequent nodes.

See Also:

- "Support Vector Machine"
- "Naive Bayes"
- "Decision Tree"
- "Generalized Linear Models"
- "Output Column"

8.12.6.1.1 Output Column

The Output column in the **Model Settings** grid controls the passing of models to subsequent nodes. By default, all models are passed to subsequent nodes. To ignore a model (that is, to *not* pass it to subsequent nodes, click

- To ignore a model, that is, to *not* pass it to subsequent nodes, click . The icon changes to , the Ignore icon.
- To cancel the ignore, click the Ignore icon again. It changes to the Output icon.

8.12.6.2 Build (Regression)

The Build section displays the following:

- **Target:** The Build node must be connected to a Data Source node. You then select the target from the target list. To change the target, select a different target from the drop-down list.
- **Case ID:** Select an attribute from the drop-down list. This attribute must uniquely identify a case. The case ID is optional. If no case ID is selected, then `<None>` is displayed. However, if no case ID is specified, then the processing will be slower.

8.12.6.3 Test (Regression)

The **Test** section specifies the data used for testing and the tests performed. By default, all models that are built using test data are tested. The test data is created randomly splitting the build data.

The following settings are available in the Test section:

- **Perform Test:** By default, all models that are built using test data are tested. The test data is created randomly splitting the build data.
  The default test results are:
  - **Performance Metrics**
  - **Residuals**

    You can deselect both.
- **Test Data:** Test Data is created is one of the following ways:
  - **Use all of the Mining Build Data for Testing**
  - **Use Split Build Data for Testing**
    Split for Test (%)
    Create Split as: `View` (default). The split creates a view that is not parallel.
  - **Use a Test Data Source for Testing**: Select this option to provide a separate test Data Source and connect the test data source to the build node after you connect the build data. Alternately, you can test a model by using a Test node.

  See Also:
  - "Testing Regression Models"
  - "Test Node"

8.12.7 Regression Node Context Menu

To view the Regression node context menu, right-click the node. The following options are available in the context menu:

- Connect
- Edit. Opens the **Edit Regression Build Node** dialog box.
- Advanced Settings. Opens the **Advanced Settings for Regression Models** dialog box.
- Validate Parents
- Run
- Force Run
- View Models
- View Test Results
- Compare Test Results
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All

- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors, displayed only if there is an error
- Show Validation Errors, displayed only if there are validation errors
- Navigate

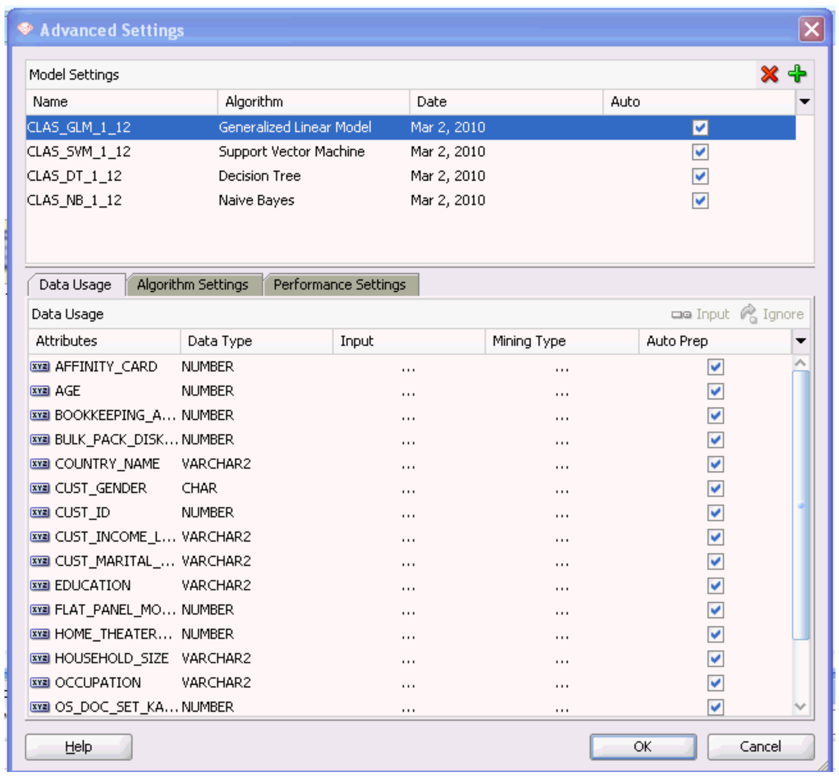
8.13 Advanced Settings Overview

The **Advanced Settings** dialog box enables you to edit data usage and other model specifications, add and remove models from the node.

You can open the **Advanced Settings** dialog box in one of these ways:

- Right-click any model node and click **Advanced Settings** from the context menu.
- Double-click the node to open the editor. Then click  .

The dialog box has two panes, as illustrated in this example of Advanced Settings for a Classification Build node:



The Advanced Settings dialog box for a Classification Build node.

****

The Upper Pane of Advanced Settings enables you to delete models and add models. You can also select models in the upper pane to change data usage. The Lower Pane of Advanced Settings, which has one, two, or three tabs, enables you to edit model specifications.

8.13.1 Upper Pane of Advanced Settings

The upper pane of the Advanced Settings lists all of the models in the node. The Model Settings grid provides the following information about each model:

- Model Name
- Algorithm
- Date of Last Build
- Auto
- Data Usage
- Column Excluded By...

To view the input and mining type for attributes, select the model in the upper pane and deselect **Auto.** If Auto is selected (the default), then the system automatically determines the attributes used to build the model.

Data Miner does not necessarily select all attributes to use in building a model. For example, if most of values of an attribute are the same, then the attribute is not selected.

To see which attributes are selected, deselect **Auto.** Select a model. The lower pane indicates the selected attributes with a check mark in the Input column.

If Auto is not selected, you can override the system's choices in the **Data Usage** tab. If Auto is not selected you can also view Input and Mining Type. This enables you to see which Attributes are used for model build, and to change them if necessary.

The **Model Settings** grid enables you to delete or add models to the node.

- **Delete:** To delete a model, select the model and click .
- **Add:** To add a model to the node, click . The **Add Models** dialog box for the node opens. In the **Add Models** dialog box, select an algorithm, either accept the default name or specify a different name, and add optional comments.

  See Also:

  - "Viewing and Changing Data Usage" for a description of how to change data usage for several columns at the same time.
  - "Data Usage"

8.13.2 Lower Pane of Advanced Settings

Select a model in the upper pane. The lower pane of the Advanced Settings displays the related information in the following tabs:

- Data Usage: For all models except Association
- Algorithm Settings: For all models
- Performance Settings: For Classification models only

These tabs display the specification used to build the selected model. You can change the specification.

8.13.2.1 Data Usage

The **Data Usage** tab is not supported for the Association node. To modify any values, to see which attributes are not used as input, or to see mining types, select **View** in the lower pane.

You can change data usage information for several models at the same time.

The Data Usage tab contains the data grid. The data grid lists all attributes in the data source. For each attribute, the grid lists displays the following:

- Name: This is the name of the attribute.
- Data Type: This is the Oracle Database data type of the attribute.
- Input: Indicates if the attribute is used to build the model. To change the input type, click **Automatic**. Then click the icon and select the new icon. For models that have a target, such as Classification and Regression models, the target is marked with a red target icon.
  - The  icon indicates that the attribute is used to build the model.
  - The  icon indicates that the attribute is ignored, that is, it is not used to build the model.
- Mining Type: This is the logical type of the attribute, either Numerical (numeric data), Categorical (character data), nested numerical, or nested categorical, text or custom text. If the attribute has a type that is not supported for mining, then the column is blank. Mining type is indicated by an icon. Move the cursor over the icon to see what the icon represents.
  To change the mining type, click **Automatic** and then click the type for the attribute. Select a new type from the list. You can change mining types as follows:
  - Numerical can be changed to Categorical. Changing to *Categorical* casts the numerical value to string.
  - Categorical.
  - Nested Categorical and Nested Numerical cannot be changed.
- Auto Prep: If **Auto Prep** is selected, then automatic data preparation is performed on the attribute. If Auto Prep is not selected, then no automatic data preparation is performed for the attribute. In this case, you are required to perform any data preparation, such as normalization, that may be required by the algorithm used to build the model. No data preparation is done (or required) for target attributes. The default is to perform automatic data preparation.
- Rules: After a model runs, Rules describe the heuristics used. For details, click **Show.**

There are two types of reasons for *not* selecting an attribute as input:

- The attribute has a data type that is not supported by the algorithm used for model build.

  For example, O-Cluster does not support nested data types such as `DM_NESTED_NUMERICALS`. If you use an attribute with type `DM_NESTED_NUMERICALS` to build a O-Cluster model, then the build fails.
- The attribute does not provide data useful for mining. For example, an attribute that has constant or nearly constant values.

  If you include attributes of this kind, then the model has lower quality than if you exclude them.

See Also:

- "Viewing and Changing Data Usage"
- "Automatic Data Preparation (ADP)"

8.13.2.2 Algorithm Settings

The **Algorithm Settings** section displays the values of algorithm settings. The settings are determined by the algorithm used to build the model.

8.13.2.3 Performance Settings

The performance settings are available for Classification models only.

The **Performance Settings** tab defines the performance objective for Classification model build. To view or change performance settings for a model, select the model in the upper pane. Weights are listed in the Weights grid. Select one of these settings:

- **Balanced:** (default) Attempts to achieve the best overall accuracy across all the target class values. This is done in different ways depending on the algorithm selected. Generally, it requires the model build process to be biased using weight values that provide extra weight to target values that occur less frequently.
- **Natural:** Enables the model to build without any bias, so that the model uses its natural view of the data to build an accurate model. In this case, rare target class values are probably not going to be predicted as frequently as they would predict the model that was built using the balanced option.
- **Custom**: Enables you to enter a set of weights for each target value. One way to get started defining custom weights is to click **Balanced** or **Natural**, just above the Weights grid. Either of these options generate weights similar to those that would result in either Balanced or Natural performance. You can then change these weights to different values.

To save the values, click **OK.**

See Also: "Lift Detail"

8.14 Mining Functions

Mining functions represent a class of mining problems that can be solved using data mining algorithms. When creating a data mining model, you must first specify the mining function and then choose an appropriate algorithm to implement the function if one is not provided by default.

Oracle Data Mining supports these mining functions:

- Classification
- Regression
- Anomaly Detection
- Clustering
- Association
- Feature Extraction and Selection

See Also: *Oracle Data Mining Concepts* for more information.

8.14.1 Classification

Classification is a data mining function that assigns items in a collection to target categories or classes, that is, items are classified according to target categories. The goal of classification is to accurately predict the target class for each case in the data. For example, a Classification model could be used to identify loan applicants as low, medium, or high credit risks.

The target categories for a classification are discrete and not ordered. The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating.

The following topics describe the classification:

- Building Classification Models
- Applying Classification Models
- Classification Algorithms

8.14.1.1 Building Classification Models

A Classification model is built from historical data for which the classifications are known. To build (train) a Classification model, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model. The model can then be applied to a different data set in which the class assignments are unknown.

Algorithm settings control model build. Settings depend on the algorithm.

Use a Build Node to build one or more Classification models.

Classification models are tested by default.

See Also:

- "Testing Classification Models"
- "Tuning Classification Models"

8.14.1.2 Comparing Classification Models

You can compare Classification models by comparing the test metrics for the models.

See Also:

- "Test Metrics for Classification Models"
- "Compare Classification Test Results"

8.14.1.3 Applying Classification Models

Scoring or applying a Classification model results in class assignments and the probability that the assignment is the correct one. For example, a model that classifies customers as low, medium, or high value would also predict the probability that the classification is correct.

Use an Apply Node to score a Classification model, that is to apply the model to new data.

8.14.1.4 Classification Algorithms

Oracle Data Mining provides the following algorithms for classification:

- **Decision Tree**

  Decision Tree automatically generates rules, which are conditional statements that reveal the logic used to build the tree.

- **Naive Bayes**

Naive Bayes uses Bayes' Theorem, a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data.

- **Generalized Linear Models (GLM)**

  Generalized Linear Models is a popular statistical technique for linear modeling. Oracle Data Mining implements GLM for binary classification and for regression.

  GLM provides extensive coefficient statistics and model statistics, and row diagnostics. GLM also supports confidence bounds, which are the upper and lower boundaries of an interval in which the predicted value is likely to lie.

- **Support Vector Machine (SVM)**

  Support Vector Machine is a powerful, state-of-the-art algorithm based on linear and non-linear regression. Oracle Data Mining implements SVM for binary and multiclass classification.

  Oracle Data Mining implements SVM for binary and multiclass classification.

8.14.2 Regression

Regression is a data mining function that predicts a number. Profit, sales, mortgage rates, house values, square footage, temperature, or distance could all be predicted using regression techniques. For example, a Regression model could be used to predict the value of a house based on location, number of rooms, lot size, and other factors.

This section on Regression contains the following topics:

- Building Regression Models
- Applying Regression Models
- Regression Algorithms

Regression models are tested by default.

See Also: "Testing Regression Models"

8.14.2.1 Building Regression Models

Use a Build Node to build one or more Regression models. Algorithm settings control the model build. Settings depend on the algorithm.

A regression task begins with a data set in which the target values are known. For example, a Regression model that predicts house values could be developed based on observed data for many houses over a period of time. In addition to the value, the data might track the age of the house, square footage, number of rooms, taxes, school district, proximity to shopping centers, and so on. House value would be the target, the other attributes would be the predictors, and the data for each house would constitute a case.

In the model build (training) process, a regression algorithm estimates the value of the target as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown.

8.14.2.2 Applying Regression Models

Scoring, or applying, a Regression model results in class assignments and the probability that the assignment is correct for each case. For example, a model that predicts a value for each cased also predicts the probability that the value is correct.

Use an Apply Node to score a Regression model, that is to apply the model to new data.

8.14.2.3 Regression Algorithms

Oracle Data Mining provides the following algorithms for regression:

- **Generalized Linear Models (GLM)**

  Generalized Linear Models is a popular statistical technique for linear modeling. Oracle Data Mining implements GLM for binary classification and for regression.

  GLM provides extensive coefficient statistics and model statistics, and row diagnostics. GLM also supports confidence bounds.

- **Support Vector Machines (SVM)**

  Support Vector Machine is a powerful, state-of-the-art algorithm based on linear and non-linear regression.

  SVM regression supports two kernels: the Gaussian Kernel for non-linear regression, and the Linear Kernel for linear regression. SVM also supports active learning.

8.14.3 Anomaly Detection

Anomaly Detection is a kind of classification problem. Standard classification algorithms require the presence of both positive and negative examples (counterexamples) for a target class. One-Class Support Vector Machine (SVM) classification requires only the presence of examples of a single target class.

The model learns to discriminate between the known examples of the positive class and the unknown negative set of counter examples. The goal is to estimate a function that is positive if an example belongs to a set and negative or zero, if the example belongs to the complement of the set.

> **Note:** Solving a one-class classification problem can be difficult. The accuracy of one-class classifiers cannot usually match the accuracy of standard classifiers built with meaningful counterexamples.

This section about Anomaly Detection models contains the following topics:

- Building Anomaly Detection Models
- Applying Anomaly Detection Models

8.14.3.1 Building Anomaly Detection Models

Oracle Data Mining uses SVM as the one-class classifier for Anomaly Detection (AD). When SVM is used for anomaly detection, it has the classification mining function but no target.

To build an AD model, use an Anomaly Detection node connected to an appropriate data source.

See Also:

- "Support Vector Machine"
- "Anomaly Detection Node"

8.14.3.2 Applying Anomaly Detection Models

One-class SVM models, when applied, produce a prediction and a probability for each case in the scoring data. If the prediction is `1,` then the case is considered typical. If the prediction is `0,` then the case is considered anomalous. This behavior reflects the fact that the model is trained with normal data.

8.14.4 Clustering

Clustering finds natural groupings of data objects, that is objects that are similar in some sense to one another. The members of a cluster are more like each other than they are like members of other clusters. The goal of clustering analysis is to find high-quality clusters such that the inter-cluster similarity is low and the intra-cluster similarity is high.

The following topics discuss clustering:

- Using Clusters
- Calculating Clusters
- Algorithms for Clustering

8.14.4.1 Using Clusters

You can use Clustering, like Classification, to segment the data. Unlike classification, Clustering models segment data into groups that were not previously defined. Classification models segment data by assigning it to previously defined classes, which are specified in a target. Clustering models do not use a target.

Clustering is useful for exploring data. If there are many cases and no obvious groupings, then you can use clustering algorithms to find natural groupings. Clustering can also serve as a useful data preprocessing step to identify homogeneous groups on which to build supervised models.

Clustering can also be used for anomaly detection. After the data has been segmented into clusters, you might find that some cases do not fit well into any clusters. These cases are anomalies or outliers.

Clusters are not necessarily disjoint; an item can be in several clusters.

8.14.4.2 Calculating Clusters

Oracle Data Mining performs hierarchical clustering. The leaf clusters are the final clusters generated by the algorithm. Clusters higher up in the hierarchy are intermediate clusters.

8.14.4.3 Algorithms for Clustering

Oracle Data Mining supports these algorithms for clustering:

- k-Means Algorithm
- O-Cluster Algorithm
- Expectation Maximization. Requires Oracle Database 12*c* Release 1 (12.1) or higher.

8.14.5 Association

Association is a data mining function that discovers the probability of the items in a collection that happen at the same time. The relationships between items that happen at the same time are expressed as association rules.

Association rules are often used to analyze sales transactions. For example, it might be noted that customers who buy cereal at the grocery store often buy milk at the same time. In fact, association analysis might find that 85 percent of the checkout sessions that include cereal also include milk.

This application of association modeling is called market-basket analysis. It is valuable for direct marketing, sales promotions, and for discovering business trends. Market-basket analysis can also be used effectively for store layout, catalog design, and cross-sell.

Association modeling has important applications in other domains as well. For example, in e-commerce applications, association rules may be used for web page personalization. An association model might find that a user who visits pages A and B is 70 percent likely to also visit page C in the same session. Based on this rule, a dynamic link could be created for users who are likely to be interested in page C.

Association modeling analyzes data that consists of Transactions.

8.14.5.1 Transactions

Association is transaction-based. A case consists of a transaction such as a market-basket or web session. The collection of items in the transaction is an attribute of the transaction. Other attributes might be the date, time, location, or user ID associated with the transaction.

In transactional data, a collection of items is associated with each case. However, in most cases, only a tiny subset of all possible items are present in a given transaction. The items in the market-basket represent only a small fraction of the items available for sale in the store.

When an item is not present in a collection, it may have a null value or it may be missing. Many of the items may be missing or null, because many of the items that could be in the collection are probably not present in any individual transaction.

8.14.6 Feature Extraction and Selection

Sometimes too much information can reduce the effectiveness of data mining. Some columns of data attributes assembled for building and testing a model may not contribute meaningful information to the model. Some may actually detract from the quality and accuracy of the model.

Irrelevant attributes add noise to the data and affect model accuracy. Irrelevant attributes increases the size of the model and the time and system resources needed for model building and scoring.

Feature Selection selects the most relevant attributes.

Feature Extraction combines attributes into a new reduced set of features. Feature Selection selects the most relevant attributes.

See Also:

- "Feature Selection"
- "Feature Extraction"

8.14.6.1 Feature Selection

Finding the most significant predictors is the goal of some data mining projects. For example, a model might seek to find the principal characteristics of clients who pose a high credit risk.

Attribute importance is also useful as a preprocessing step in classification modeling. Decision Tree and Generalized Linear Models benefit from this type of preprocessing. Oracle Data Mining implements Feature Selection for optimization within both of these algorithms

Oracle Data Miner provides the **Attribute Importance** setting in the Filter Columns transformation to identify important features using the Oracle Data Mining importance function.

8.14.6.2 Feature Extraction

Feature Extraction is an attribute reduction process. Unlike Feature Selection, which ranks the existing attributes according to their predictive significance, Feature Extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes.

The Feature Extraction process results in a much smaller and richer set of attributes. The maximum number of features may be user-specified or determined by the algorithm. By default, it is determined by the algorithm.

Oracle Data Mining supports these algorithms for Feature Extraction:

- Nonnegative Matrix Factorization
- Singular Value Decomposition and Principal Components Analysis. Requires Oracle Database 12*c* Release 1 (12.1) or higher.

9 Evaluate and Apply Nodes

Oracle Data Mining enables you to test Classification and Regression models. A Test node is one of several ways to test a model. After you build a model, you apply the model to new data using an Apply node. Evaluate and Apply data must be prepared in the same way that build data was prepared.

The Evaluate and Apply nodes are:

- Apply Node
- Test Node

See Also: "Evaluate and Apply Data"

9.1 Apply Node

The Apply node takes a collection of models and returns a single score.The Apply node produces a query as the result. The result can further transformed or connected to a Create Table or View node to save the data as a table.

To make predictions using a model, you must apply the model to new data. This process is also called **scoring** the new data.

An Apply node generates the SQL for Scoring using one or more models. The SQL includes pass-through (supplemental) attributes and columns generated using Scoring functions.

Note: You cannot apply Association or Attribute Importance models.

This section on Apply node contains the following topics:

- Apply Preferences
- Apply Node Input
- Apply Node Output
- Creating an Apply Node
- Apply and Output Specifications
- Edit Apply Node
- Apply Node Properties
- Apply Node Context Menu

See Also: "About Parallel Processing"

9.1.1 Apply Preferences

To apply preferences to an Apply Node:

1. In **Tools** menu option, click **Preferences.**
2. In the **Preferences** dialog box, click **Data Miner.** You can view and change preferences for Apply operations. The default preferences for Data Miner are:
   - **Automatic Apply Settings**
   - **Data Columns First**
3. Click **OK.**

9.1.2 Apply Node Input

An Apply node requires the following input:

- One or more of the following:
  – Model node
  – Model Build node

  You must specify at least one model to apply. You can apply several models at the same time.
- Any node that generates data as an output such as a Data node, a Transforms node, or an appropriate Text node.

  Only one input node is permitted.

  When you apply a model to new data, the new data must be transformed in the same way as the data used to build the model.

> **Note:** You cannot apply Association or Attribute Importance models.

9.1.3 Apply Node Output

An Apply node generates a data flow based on the Apply and Output Specifications.

9.1.4 Creating an Apply Node

Before creating an Apply node, you must connect a Data node and a Model node or Build node to the Apply node. To create an Apply node:

1. Open the **Components** pane and select **Workflow Editor.**
   If the **Components** pane is not visible, then go to **View** and click **Components.**
2. Either identify Apply Data or create a Data Source Node containing the Apply Data. Ensure that the Apply Data is prepared in the same way as the Build Data.
3. Create a Model Node, a Model Build node (such as a Classification Node), or a combination of these nodes. At least one model must be successfully built before it can be applied.
   You cannot apply Association models.
4. In the **Workflow Editor,** expand **Evaluate and Apply**, and click **Apply.**
5. Drag and drop the Apply node in the **Workflow** pane.
6. Link the Data node, Model nodes, and Build nodes to the Apply Node.

See Also:

- "Evaluate and Apply Data"
- "Edit Apply Node"

9.1.5 Apply and Output Specifications

There are two ways to create Apply and Output specifications:

- Use the Automatic Settings.
- Create specifications using the Edit Apply Node.
- Add Additional Output to help identify output.

9.1.5.1 Automatic Settings

By default, **Automatic Settings** is used.

9.1.5.2 Edit Apply Node

To edit or view an Apply specification, either double-click the Apply node or right-click the Apply node and select **Edit**. The **Edit Apply Node** dialog box opens.

The **Edit Apply Node** dialog box has two tabs:

- Predictions: Defines the Apply Scoring specifications.

  An Apply specification consists of several Output Apply columns. The column names are generated automatically.

  – You can specify names. The names must not be more than 30 characters.

  – You can then select a model from the list of models in all Input nodes and an Apply function.
  The Apply functions that you can select depend on the selected models.

- Additional Output: Specifies pass-through columns from the Input node. You can select as many columns as you want. You can specify that these selected columns are displayed before the Apply columns (the default) or after the Apply columns.

  These columns are often used to identify the Apply output. For example, you can use the Case ID column to identify the Apply output.

  The default is to *not* specify any additional output.

  The **Default Column Order,** at the bottom of the **Edit Apply Node** dialog box, is `Data Columns First` in the output. You can change this to `Apply Columns First`.

  See Also:

  - "Apply Functions Parameters"
  - "Apply Functions"
  - "Default Apply Column Names"

9.1.5.2.1 Predictions

To define specific Apply settings or to edit the default settings, deselect **Automatic Settings.** You then add new Apply functions or edit existing ones.

You can edit settings in several ways:

- **Add a setting:** Click to open the **Add Output Apply Column** dialog box.

- **Edit an existing setting:** Select the setting and click . The **Edit Output Data Column** dialog box opens.
- **Delete a specification:** Select it and click .
- **Define Apply Columns:** Click . In the **Define Apply Columns Wizard**, click the Define Apply Columns icon.

See Also:

- "Add Output Apply Column Dialog"
- "Apply Functions"
- "Edit Output Data Column Dialog"
- "Define Apply Columns Wizard"

9.1.5.2.2 Apply Functions

The Apply functions that you can choose depend on the models that you apply.

Note: Certain Apply functions are available only if you are connected to Oracle Database 12c.

The Apply functions, arranged according to Model node are:

- **Anomaly Detection Models**
  - **Prediction**: An automatic setting that returns the best prediction for the model. The data type returned depends on the target value type used during the build of the model. For Regression models, this function returns the expected value. The function returns the lowest cost prediction using the stored cost matrix if a cost matrix exists. If no stored cost matrix exists, then the function returns the highest probability prediction.
  - **Prediction Details**: Returns prediction details. The return value describes the attributes of the prediction. For Anomaly Detection, the returned details refer to the highest probability class or the specified class value.

    Note: Prediction Details requires a connection to Oracle Database 12 c.

    The defaults for Predictions Details are:

    * **Target Value**: *Most Likely*
    * **Sort by Weights**: Absolute value
    * **Maximum Length of Ranked Attribute List**: 5

    Prediction Details output is in XML format (XMLType data type). You must parse the output to find the data that you need.
  - **Prediction Probability**: An automatic setting that returns the probability associated with the best prediction.
  - **Prediction Set**: Returns a varray of objects containing all classes in a multiclass classification scenario. The object fields are named `PREDICTION`, `PROBABILITY`, and `COST`. The data type of the `PREDICTION` field depends on the target value type used during the build of the model. The other two fields are

both Oracle NUMBER. The elements are returned in the order of best prediction to worst prediction.

- **Clustering Models**
  - **Cluster Details**: The return value describes the attributes of the highest probability cluster or the specified cluster ID. If you specify a value for TopN, then the function returns the *N* attributes that most influence the cluster assignment (the score). If you do not specify TopN, then the function returns the five most influential attributes.

    Note: Cluster Details requires a connection to Oracle Database 12*c*.

    The defaults for Predictions Details are as follows:

    * **Cluster ID**: *Most Likely*
    * **Sort by Weight**: Absolute value
    * **Maximum Length of Ranked Attribute List**: 5

    The returned attributes are ordered by weight. The weight of an attribute expresses its positive or negative impact on cluster assignment. A positive weight indicates an increased likelihood of assignment. A negative weight indicates a decreased likelihood of assignment.

    Cluster Details output is in XML format (XMLType data type). You must parse the output to find the data that you need.
  - **Cluster Distance**: Returns a cluster distance for each row in the selection. The cluster distance is the distance between the row and the centroid of the highest probability cluster or the specified cluster ID.

    Note: Custer Distance requires connection to Oracle Database 12*c*.

    The defaults for Predictions Details are as follows:

    * **Cluster ID**: *Most Likely*
  - **Cluster ID**: An automatic setting that returns the NUMBER of the most probable cluster ID. If the cluster ID has been renamed, then a VARCHAR2 is returned instead.
  - **Cluster Probability**: An automatic setting that returns a measure of the degree of confidence of membership (NUMBER) of an input row in a cluster associated with the specified model.
  - **Cluster Set**: Returns a varray of objects containing all possible clusters that a given row belongs to given the parameter specifications. Each object in the varray is a pair of scalar values containing the cluster ID and the cluster probability. The object fields are named CLUSTER_ID and PROBABILITY, and both are Oracle NUMBER Clustering models only.
- **Feature Extraction Models**
  - **Feature ID**: Returns an Oracle NUMBER that is the identifier of the feature with the highest value for the row.
  - **Feature Set**: An automatic setting that is similar to Cluster Set.

- **Feature Value**: Returns the value of a given feature. If you omit the feature ID argument, then the function returns the highest feature value.
- **Feature Details**: The return value describes the attributes of the highest value feature or the specified feature ID. If you specify a value for TopN, the function returns the *N* attributes that most influence the feature value. If you do not specify TopN, the function returns the 5 most influential attributes.

  Note: Feature Extraction Model requires connection to Oracle Database 12*c.*

  The returned attributes are ordered by weight. The weight of an attribute expresses its positive or negative impact on the value of the feature. A positive weight indicates a higher feature value. A negative weight indicates a lower feature value.

  The defaults for Predictions Details are as follows:

  * **Feature ID**: *Most Likely*
  * **Sort by Weight**: Absolute value
  * **Maximum Length of Ranked Attribute List**: 5

  Feature Details output is in XML format (XMLType data type). You must parse the output to find the data that you need.

- **Classification and Regression Models**
  - **Prediction**: An automatic setting that returns the best prediction for the model. The data type returned depends on the target value type used during the build of the model.
    * For Regression models, this function returns the expected value.
    * For Classification models, the returned details refer to the highest probability class or the specified class value.

      The function returns the lowest cost prediction using the stored cost matrix if a cost matrix exists. If no stored cost matrix exists, then the function returns the highest probability prediction.
  - **Prediction Bounds**: For generalized linear models, it returns an object with two NUMBER fields `LOWER` and `UPPER`. If the GLM was built using Ridge Regression, or if the Covariance Matrix is found to be singular during the build, then this function returns `NULL` for both fields.
    * For a Regression mining function, the bounds apply to value of the prediction.
    * For a Classification mining function, the bounds apply to the probability value.
  - **Prediction Bounds Lower**: Same as Prediction Bounds but only returns the lower bounds as a scalar column. Automatic Setting for GLM models.
  - **Prediction Bounds Upper**: Same as Prediction Bounds but only returns the upper bounds as a scalar column. Automatic Setting for GLM models.
  - **Prediction Details**: Requires connection to Oracle Database 12c except for Decision Tree.

    The defaults for Predictions Details for Classification are as follows:

* **Target Value**: *Most Likely*
* **Sort by Weights**: Absolute value
* **Maximum Length of Ranked Attribute List**: 5

The defaults for Predictions Details for Regression are as follows:

* **Sort by Weights**: Absolute value
* **Maximum Length of Ranked Attribute List**: 5

DT Prediction Details: Returns a string containing model-specific information related to the scoring of the input row. In Oracle Data Miner releases earlier than 4.0, the return value is in the form `<Node id = "integer"/>`.

Note: DT Prediction Details requires a connection to Oracle Database 11*g* Release 2 (11.2)

- **Classification**
  - **Prediction Costs**: Returns a measure of cost for a given prediction as a `NUMBER`. Classification models only. Automatic Setting for DT models.
  - **Prediction Probability**: Returns the probability associated with the best prediction.
    The Automatic Setting for is *Most Likely.*
  - **Prediction Set**: Returns a varray of objects containing all classes in a multiclass classification scenario. The object fields are named `PREDICTION`, `PROBABILITY`, and `COST`. The data type of the `PREDICTION` field depends on the target value type used during the build of the model. The other two fields are both Oracle NUMBER. The elements are returned in the order of best prediction to worst prediction.

See Also: "Apply Functions Parameters"

9.1.5.2.3 Apply Functions Parameters

The Apply Function parameters that can be specified:

- **Cluster ID:** The default is *Most Probable.* No other parameters are supported.
- **Cluster Probability:** The default is *Most Probable.*
  You can also select a specific cluster ID or specify `NULL` or `Most Likely` to return the bounds for the most likely cluster.
- **Cluster Set:** The default is *All Clusters.*
  You can also specify either or both of the following:
  - `TopN`: Where *N* is between one and the number of clusters. The optional TopN argument is a positive integer that restricts the set of features to those that have one of the top *N* values. If there is a tie at the *N*th value, then the database still returns only *N* values. If you omit this argument, then the function returns all features.
  - `Probability Cutoff`: It is a number strictly greater than zero and less than or equal to 1. The optional cutoff argument restricts the returned features to only those that have a feature value greater than or equal to the specified cutoff. To filter only by cutoff, specify Null for TopN and the desired cutoff for cutoff.
- **Feature ID:** The default is *Most Probable.* No other values are supported.

- **Feature Set:** The default is All Feature IDs. You can also specify either or both of the following:
  - `TopN`: Where *N* is between 1 and the number of clusters. The optional TopN argument is a positive integer that restricts the set of features to those that have one of the top *N* values. If there is a tie at the *N*th value, then the database still returns only *N* values. If you omit this argument, then the function returns all features.
  - `Probability Cutoff`: It is a number strictly greater than zero and less than or equal to one. The optional cutoff argument restricts the returned features to only those that have a feature value greater than or equal to the specified cutoff. To filter only by cutoff, specify Null for TopN and the desired cutoff.
- **Feature Value:** The default is `Highest Value`.
  You can also select a specific feature ID value or specify anyone of the following value to return the bounds for the most likely feature:
  - `NULL`
  - `Most Likely`
- **Prediction:** The default is `Best Prediction` to consider the cost matrix.
- **Prediction Upper Bounds** or **Prediction Lower Bounds:** The default is *Best Prediction* with Confidence Level 95%.
  You can change Confidence Level to any number strictly greater than zero and less than or equal to one. For Classification models only, you can use the **Target Value Selection** dialog box option to pick a specific target value. You can also specify Null or *Most Likely* to return the bounds for the most likely target value.
- **Prediction Costs:** The default is `Best Prediction`.
  Applicable for Classification models only. You can use the T**arget Value Selection** option to pick a specific target value.
- **Prediction Details:** Only value is the details for the Best Prediction.
- **Prediction Probability:** The default is *Best Prediction.*
  Applicable for Classification models only. You can use the **Target Value Selection** option to pick a specific target value.
- **Prediction Set:** The default is `All Target Values`.
  You can also specify one or both of the following:
  - `bestN`: Where *N* is between one and the number of targets. The optional `bestN` argument is a positive integer that restricts the returned target classes to the *N* having the highest probability, or lowest cost if cost matrix clause is specified. If multiple classes are tied in the *N*th value, then the database still returns only *N* values.
    To filter only by cutoff, specify `Null` for this parameter.
  - `Probability Cutoff`: Is a number strictly greater than zero and less than or equal to one. The optional cutoff argument restricts the returned target classes to those with a probability greater than or equal to (or a cost less than or equal to if cost matrix clause is specified) the specified cutoff value.
    You can filter solely by cutoff by specifying `Null` for this value.

9.1.5.2.4 Default Apply Column Names

The syntax of the default Apply column name is:

```
"<FUNCTION ABBREVIATION>_<MODEL NAME><SEQUENCE>
```

`SEQUENCE` is used only if necessary to avoid a conflict. A sequence number may force the model name to be partially truncated.

`FUNCTION ABBREVIATION` is one of the following:

- Cluster Details: `CDET`
- Cluster Distance: `CDST`
- Cluster ID: `CLID`
- Cluster Probability: `PROB`
- Cluster Set: `CSET`
- Feature Details: `FDET`
- Feature ID: `FEID`
- Feature Set: `FSET`
- Feature Value: `FVAL`
- Prediction: `PRED`
- Prediction Bounds: `PBND`
- Prediction Upper Bounds: `PBUP`
- Prediction Lower Bounds: `PBLW`
- Prediction Costs: `PCST`
- Prediction Details: `PDET`
- Prediction Probability: `PROB`
- Prediction Set: `PSET`

Specific target, feature, or cluster default names are abbreviated in one of two ways.

- The first approach attempts to integrate the value of the target, feature, or cluster into the column name. This approach is used if the maximum value of the target, cluster, or feature does not exceed the remaining character spaces available in the name. The name must be 30 or fewer characters.
- The second approach substitutes the target, cluster, or feature with a sequence ID. This approach is used if the first approach is not possible.

9.1.5.2.5 Add or Edit Apply Output Column

The **Add Apply Output** dialog box or the **Edit Apply Output** dialog box enables you to add manually or edit a single column Apply definition. You can edit or add Apply definitions one at a time.

Before you add or edit columns, you must deselect **Automatic Settings.**

You can perform the following tasks:

- **Add an Apply Output column:** Click .
- **Edit an Apply Output column:** Click .When you edit a column, only the Function selection box and its parameters can be edited.

The following controls are available:

- **Column**: Name of column to be generated.
- **Auto**:

- If selected, you cannot edit the name of the column.
- If deselected, auto naming is disabled and you can rename the column. Column names are validated to ensure that they are unique.

- **Node**: List of Model Input nodes connected to node. If there is only one Input node, then it is selected by default.
- **Model**: List of models for the selected node. If there is only one model, then it is selected by default.
- **Function**: List of model scoring functions for the selected model.
- **Parameters**: Displays `0` or more controls necessary to support the parameter requirements of the selected function.

When you have finished defining the output column, click **OK**.

See Also:

- "Default Apply Column Names"
- "Apply Functions"
- "Default Apply Column Names"

9.1.5.2.6 Add Output Apply Column Dialog

The default is to automatically name the output column.

To add a column:

1. In the **Column** field, provide a name.
2. Deselect **Auto.**
3. In the **Node** field, select one of the node connected to the Apply node. The type of the node that you select determines the choices in the Model and Function fields.
4. In the **Model** field, select a model.
5. In the **Function** field, select a function.
6. After you are done, click **OK.**

See Also: "Apply and Output Specifications"

9.1.5.3 Define Apply Columns Wizard

The **Define Apply Column** wizard has two steps:

1. Models
2. Output Specifications

9.1.5.3.1 Models

In the Model section:

1. Select the Models for which you want to define output specification.
2. Click **Next.**

9.1.5.3.2 Output Specifications

In Output Specifications, the possible output specifications are listed with the defaults selected.

1. Make the changes as applicable.
2. Click **Finish** to complete the definition.

See Also: "Apply and Output Specifications"

9.1.5.4 Additional Output

Additional Output consists of columns that are passed unchanged through the Apply operation.

See Also: "Additional Output"

9.1.6 Evaluate and Apply Data

Test and Apply data for a model must be prepared in the same way that build data for the model was prepared. To properly prepare test and apply data, duplicate the transformation chains of build data for test and apply data by copying and pasting build Transforms nodes.

9.1.7 Edit Apply Node

The **Edit Apply Node** dialog box has the following tabs:

- Predictions
- Additional Output

The default value for **Default Column Order** is `Data Columns First`, which means that any data columns that you add come first in the output. You can change this to `Apply Columns First`.

9.1.7.1 Apply Columns

To create an Apply specification, deselect **Automatic Settings**. By default, Automatic Settings is selected.

You can perform the following tasks:

- To define Apply columns, click .
  The **Define Apply Column** wizard opens.
- To add an Output Apply column, click .

  The **Add Output Apply Column** dialog box opens.
- To delete an Output Apply column, click .
- To edit an Output Apply column specification, select the specification. Click .
  The **Add or Edit Apply Output Column** dialog box opens.

See Also:

- "Add Output Apply Column Dialog"
- "Add or Edit Apply Output Column"
- "Add Output Apply Column Dialog"

9.1.7.2 Additional Output

In the **Additional Output** tab, you can specify pass-through attributes from Data Source nodes.

To add a columns:

1. Click . The **Edit Output Data Column** dialog box opens.
2. In the **Edit Apply Node** dialog box, the **Default Column Order** is `Data Columns First`. You can change this to `Apply Columns First`.
3. After you are done, click **OK.**

> **See Also:** "Edit Output Data Column Dialog"

9.1.7.2.1 Edit Output Data Column Dialog

By default, no data columns are specified. To specify or include data columns:

1. Move the attributes from the **Available Attributes** list to the **Selected Attributes** list.
2. Click **OK.** Data columns are passed through the apply operation unchanged. Certain attributes, such as the case ID, can be useful in interpreting apply output.

9.1.8 Apply Node Properties

To view properties of the Apply node, right-click the node and click **Go to Properties.** Alternately, go to **View** and click **Properties.**

Apply Node Properties has the following sections:

- **Predictions**: Displays the Output Apply columns defined on the Apply Columns. You can edit these details. The option **Automatic Selection** is selected if selections were not modified.

  For each Output Apply Column, the Name, Function, Parameters, and Node are listed.
- **Additional Output:** Lists the Output Data Columns that are passed through. For each column, the Name, Alias (if any), and Data Type are listed.
- Cache
- **Details:** Displays the name of the node and comments.

9.1.9 Apply Node Context Menu

To view the Apply node context menu, right-click the node. The following options are available in the context menu:

- Connect
- Edit. Opens Edit Apply Node.
- Validate Parents
- Run
- View Data. Opens Apply Data Viewer.
- Force Run
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy

- Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log. Displayed only if the running of the node fails.
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed only if there is an error.
- Navigate

9.1.9.1 Apply Data Viewer

The **Apply Data** viewer opens in a new tab. The viewer has these tabs:

- **Data:** Displays rows of data. The default is to view the cache data. You can perform the following tasks:
  - View actual data.
  - Sort data.
  - Filter data with a SQL expression.
  - Refresh the display. To refresh, click .
- **Columns:** Lists the columns in the apply.
- **SQL:** Lists the SQL used to generate the Apply Output.

9.2 Test Node

Oracle Data Mining enables you to test Classification and Regression models. You cannot test other kinds of models.

> **Note:** All of the models tested in a node must be either Classification or Regression Models. You cannot test both kinds of models in the same test node.

A Test node can test several models using the same test set.

If Automatic Settings are on, the test node specification is generated when you connect the input nodes.

A Test node can run in parallel.

This section on Test node consists of the following topics:

- Support for Testing Classification and Regression Models
- Test Node Input
- Automatic Settings
- Creating a Test Node
- Edit Test Node
- Compare Test Results Viewer
- Test Node Properties
- Test Node Context Menu

See Also:

- "Automatic Settings"
- "About Parallel Processing"
- "Test Node Input"

9.2.1 Support for Testing Classification and Regression Models

Oracle Data Miner supports testing of a Classification or Regression models in the following ways:

- Test the model as part of the Build node using any one of the following ways:
  - Split the Build data into build and test subsets.
  - Use all of the Build data as test data.
  - Connect a second Data Source node, the test Data Source node to the Build node
- Test the model in a Test node. In this case, the test data can be any table that is compatible with the Build data.
- After you have tested a Classification Model, you can tune it.

Note: You cannot tune Regression models.

See Also:

- "Evaluate and Apply Data"
- "Tuning Classification Models"
- "Test"

9.2.2 Test Node Input

A Test node has the following input:

- At least one node that identifies one or more models. The nodes can be a Model node, a Classification node, or a Regression node. A Model node must contain either Classification or Regression model, but not both.
- Any node that generates data as an output such as a Data node, a Transform node, or an appropriate Text node. This node contains the test data.
- It is recommended that a case ID is specified. If you do not specify a Case ID, then the processing will take longer.

You can test several Classification or several Regression Models at the same time. The models to be tested can be in different nodes. The models to be tested must satisfy these conditions:

- The nodes that contain the models must have the same function type. That is, they must be all Classification Build nodes or all Regression Build nodes.

  Classification Models must also have the same list of target attribute values.
- The models must have the same target attribute with the same data type.
- The Data Source node for Test must contain the target of the models.

- The test data must be compatible with the models. That is, it should have been transformed in the same way as the data used to build the model.

9.2.3 Automatic Settings

By default, the option **Automatic Settings** are selected for a Test node. Automatic Settings result in the following behavior:

- When a model input node is connected, all models are added to the specification.
- When a model input node is disconnected, all models are removed from the specification. The test node may become invalid.
- When a model input node is edited in the following ways, the resultant behavior is as follows:
  - If models are added, then model specifications are automatically added to the Test node.
  - If models are removed, then the specifications are removed from the Test node.
  - If models are changed, then the following is done:
    * The Test node is updated to ensure the algorithm is consistent.
    * If the target changes and there is only one node as input to the Test node, then the node is updated to reflect the new target and keep all the models. Also, the test input data is validated to ensure that it still has the new column target.
    * If there are multiple Model nodes as input to the Test nodes, then the models with the changed target are automatically removed.

If **Automatic Settings** is deselected, then you must edit the node to reflect all changes to the input. Models are validated if they are added.

9.2.4 Creating a Test Node

Before you create a Test node, you must connect or a Data Source node and a Model node or Build node to the Test node. To create a Test node:

1. In the **Components** pane, click **Workflow Editor.**
   If the **Components** pane is not visible, then go to **View** and click **Components.**
2. Either identify or create a Data Source Node containing the test data.
   Ensure that the test data is prepared in the same way as the build data.
3. Select at least one Model Node, Classification Node, or Regression Node.
   A model must be successfully built before it can be tested.

   > **Note:** You can test either Classification or Regression models but not both kinds of models in one Test node.

4. In the **Workflow Editor,** expand **Evaluate and Apply,** and click **Test.**
5. Drag and drop the Test node from the **Workflow Editor** to the workflow pane.
6. Link the Data node, the Model node or Build node to the Test Node.
7. Characteristics of the Test node are set by default. You can also edit the node.

See Also:

- "Link Nodes"
- "Edit Test Node"
- "Evaluate and Apply Data"

9.2.5 Edit Test Node

To edit the Test Node, right-click the node and select **Edit** or double-click the node. The **Edit Test Node** dialog box opens.

The **Edit Test Nod**e dialog box displays the following:

- Function (CLASSIFICATION or REGRESSION)
- Target and the Target type (data type of the target)
- The Case ID (if there is one)

  It is recommended that you specify a case ID. If you do not specify a case ID, processing will be slower. The case ID that you specify for the Text node should be the same as the case ID specified for the Build node.

By default, the option **Automatic Settings** is selected.

You can perform the following tasks:

- Compare test results and view individual models even when Automatic Settings is selected. The models tested are listed in the Selected Models grid.
- Make change to the list of the models. Deselect **Automatic Settings** and make changes in the Selected Models grid.

See Also:

- "No Case ID"
- "Automatic Settings"

9.2.5.1 Selected Models

For each model, the grid lists the following:

- The model name
- The node containing the model
- Test status of the model
- The algorithm that is used to build the model

You can perform the following tasks:

- **View Models:** You can view models that are successfully built. Select the model in the grid and click .
- **Compare Test Results:** Click . The test results are displayed in Compare Test Results Viewer.
- **Add Model:** Click . You can only add models that have the same function. Before adding a model, deselect **Automatic Settings.**
- **Delete Model:** Select it and click .
  Before deleting a model, deselect **Automatic Settings.**

9.2.5.2 Select Model

The **Select Model** dialog box lists the models that are available for testing. To select models:

1. Move the Models from **Available Models** to **Selected Models.**
2. Click **OK.**

9.2.6 Compare Test Results Viewer

The **Compare Test Result** viewer displays test results for one or more models in the same node. The following test results are displayed:

- Compare Classification Test Results
- Compare Regression Test Results

See Also:

- "Classification Model Test Viewer"
- "Regression Model Test Viewer"

9.2.7 Test Node Properties

To view properties, right-click the node and select **Go to Properties.** If the **Properties** pane is closed, then go to **View** and click **Properties.**

The Test node **Properties** pane has these sections:

- **Models:** Lists the models to test in the Selected Models grid.
- Test
- **Details:** Displays the name of the node and comments.

9.2.7.1 Models

The **Models** tab lists the models to test in the Selected Models grid.

9.2.7.2 Test

The Test section describes how testing is performed. Test contains these information:

- **Function:** CLASSIFICATION or REGRESSION.
- **Target:** The name of the target.
- **Data Type:** The data type of the target
- For CLASSIFICATION, these test results are calculated by default:
  - Performance Metrics
  - ROC Curve (Binary Target Only)
  - Lift and Profit

  You can deselect **Metrics.**

  By default, the top 100 target values by frequency is specified. To change this value, click **Edit.** Edit the value in the **Target Values Selection** dialog box.
- For REGRESSION, Accuracy Matrix and Residuals are selected. You can deselect **Metrics.**

– The Performance Metrics are the metrics displayed on the Performance tab of the Test Viewer.

– Residuals are displayed on the Residual tab of the Test Viewer.

See Also:

- "Test Metrics for Classification Models"
- "Performance (Regression)"
- "Residual"

9.2.7.2.1 Target Values Selection

The **Target Values Selection** dialog box displays the number of target values selected. The default setting is `Automatic`.

It uses the top 10 Target Class Values by frequency. You can change the number of target values by changing `Frequency Count`. You can also select `Use Lowest Occurring`.

To select custom values:

1. Select **Custom.**
2. Move the values from **Available Values** to **Selected Values.**
3. After you are done, click **OK.**

9.2.8 Test Node Context Menu

To view the Test node context menu, right-click the node. The following options are available in the context menu:

- Connect
- Edit. Opens Edit Test Node.
- Validate Parents
- Run
- Force Run
- View Models
- View Test Results
- Compare Test Results. Opens the Compare Test Results Viewer
- Generate Apply Chain
- Cut
- Copy
- Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Navigate
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed only if there are validation errors.

- Show Event Log. Displayed only if there is an error.

10 Predictive Query Nodes

Predictive Query nodes enable you to score data dynamically without a predefined model. Predictive Queries use in-database scoring technology.

Note: Predictive Query Nodes require Oracle 12*c* Release 1 or later.

Scoring using Predictive Query nodes has the following limitations:

- The transient models created during the running of Predictive Query node are not available for inspection or fine tuning.
- If it is necessary to inspect the model, correlate scoring results with the model, specify special algorithm settings, or run multiple scoring queries that use the same model, then a predefined model must be created.

The output of a Predictive Query is the output of an Apply operation.

There are several Predictive Query nodes:

- Anomaly Detection Query
- Clustering Query
- Feature Extraction Query
- Prediction Query

See Also: "Apply Node Output"

10.1 Anomaly Detection Query

An Anomaly Detection Query node analyses the input for anomalies. That is, it detects unusual cases in data.

Note: Predictive Query Nodes require Oracle 12*c* Release 1 or later.

Anomaly Detection Query can run in parallel.

This section about Anomaly Detection consists of the following topics:

- Create an Anomaly Detection Query Node
- Edit an Anomaly Detection Query
- Run Predictive Query Node
- View Data for a Predictive Query

- Anomaly Detection Query Properties
- Anomaly Detection Query Context Menu

See Also:

- "About Parallel Processing"
- "Anomaly Detection"

10.1.1 Create an Anomaly Detection Query Node

To create an Anomaly Detection Query in an existing workflow:

1. In the **Components** pane, go to **Workflow Editor.** If the **Components** pane is not visible, then go to **View** and click **Components.**
   Alternately, press Ctrl+Shift+P to dock the **Components** pane.
2. Create the Data Source node containing the input data.
3. Expand the **Predictive Query** section in the Components pane.
4. Drag and drop the Anomaly Detection Query node from the Components pane to the workflow pane.
5. Connect the Data Source node to the Anomaly Detection Query Node.
6. Edit the Anomaly Detection Query node.
7. Run the Predictive Query node and view the data for a Predictive Query node.
8. To save the results of the query, use a Create Table or View node.

See Also:

- "Create a Data Source Node"
- "Create Table or View Node"
- "Edit an Anomaly Detection Query"
- "Run Predictive Query Node"
- "View Data for a Predictive Query"

10.1.2 Edit an Anomaly Detection Query

To edit an Anomaly Detection Query Node, either double-click the node or right-click the node and select **Edit.** In the **Select Columns** window, specify the following information on the **Anomaly Predictions** tab:

1. Double-click the Anomaly Detection Query node or right-click the node and click **Edit.** The **Edit Anomaly Detection Query Node** dialog box opens.
2. In the **Anomaly Detection Query Node** dialog box, enter the details in the following tabs:
   - **Anomaly Predictions** tab:
     - In the **Case ID** field, select a case ID from the drop-down list (Optional). It is recommended that you specify a case ID to uniquely define each record. The case ID helps with model repeatability and is consistent with good data mining practices.

- You can edit the outputs of an Anomaly Prediction (optional). Oracle Data Miner automatically determines the output for the query. You can modify the output.

- **Partition** tab: You can perform the following tasks.
  - Add one or more Partition attributes. This is optional. Selecting a Partition attribute directs the predictive query to build a virtual model for each unique partition.
  - Select partitions. To select partitions, click the **Partition** tab and click . Use the **Add Partitioning Columns** dialog box to select the partitions. You can also specify partitioning expressions.
- **Input** tab: You can perform the following tasks:
  - Modify Input.
  - Add and modify input. Click **Input** to add and modify input.
  - Remove input.
  - Change the mining type.
- **Additional Output** tab: You can add output (optional). By default, all target columns, the Case ID column, and partitioning columns are automatically added to additional output. To make changes, click **Additional Output.**

3. Click **OK.**

See Also:

- "Add Additional Output"
- "Add Partitioning Columns"
- "Edit Anomaly Prediction Output"

10.1.2.1 Edit Anomaly Prediction Output

Oracle Data Miner automatically selects output for query. The default output is listed in the Anomaly Prediction Outputs section in the **Anomaly Predictions** tab. The defaults are:

- Prediction
- Prediction Details
- Prediction Probability

You can select Prediction Set and edit parameters of the output functions. You can perform the following tasks:

- **Delete:** To delete an output, select the output and click .
- **Add:** To add an output, click . Use the **Add Anomaly Function** dialog box to select an output.
- **Edit:** To edit an output, either double-click the function or select the function and click . Use **Edit Anomaly Function** dialog box to make changes.

The output of a Predictive Query is the output of an Apply (Scoring) operation.

See Also:

- "Apply Functions"
- "Add Anomaly Function"
- "Edit Anomaly Function Dialog"

10.1.2.2 Add Anomaly Function

To add Anomaly Function:

1. In the **Function** field, select a function from the drop-down list. The options are:
   - Prediction
   - Prediction Probability
   - Prediction Details
   - Prediction Set
2. To specify a default name instead of using the default name, deselect **Auto.** This turns off automatic selection.
3. Click **OK.**

10.1.2.3 Edit Anomaly Function Dialog

To edit an Anomaly Function:

1. Select the change that you want to make to the function.
2. To specify a name instead of using the default name, deselect **Auto.** This turns off automatic selection.
3. Click **OK.**

10.1.3 Anomaly Detection Query Properties

The Anomaly Detection Query Properties enables you to view and change information about this Predictive Query node. If Properties is closed, then go to **View** and click **Properties.** Alternately, right-click the node and click **Go to Properties.**

The Anomaly Detection Query Node Properties has these sections:

- Anomaly Predictions
- Partition
- Additional Output (Anomaly Query)
- Cache
- Details

10.1.3.1 Anomaly Predictions

Displays the predictions produced by the query.

See Also: "Edit an Anomaly Detection Query"

10.1.3.2 Additional Output (Anomaly Query)

Displays the output specified.

See Also: "Edit Anomaly Prediction Output"

10.1.4 Anomaly Detection Query Context Menu

To view the Anomaly Detection Query node context menu, right-click the node. The following options are available in the context menu:

- Connect
- Edit. Opens **Edit an Anomaly Detection Query** dialog box.
- Run
- Force Run
- View Data. It performs two functions:
  - Runs the node on a small sample of the data.
  - Opens the **View Data for a Predictive Query** dialog box.
- Save SQL
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed if there are validation errors.
- Go to Properties
- Navigate

10.2 Clustering Query

A Clustering Query node returns the clusters in the input.

Note: Predictive Query nodes require Oracle 12*c* Release 1 or later.

A Clustering Query can run in parallel.

This section about Clustering Query node contains the following topics:

- Create a Clustering Query
- Edit a Clustering Query
- Run Predictive Query Node
- View Data for a Predictive Query
- Clustering Query Properties

- Clustering Query Context Menu

See Also:

- "About Parallel Processing"
- "Clustering"

10.2.1 Create a Clustering Query

To create a Clustering Query in an existing workflow:

1. In the **Components** pane, go to **Workflow Editor.** If the **Components** pane is not visible, then go to **View** and click **Components.**
   Alternately, press Ctrl+Shift+P to dock the **Components** pane.
2. Create a Data Source node containing the input data.
3. Expand the **Predictive Queries** section in the Components pane.
4. Drag and drop a Clustering Query node to the workflow.
5. Connect the Data Source node to the Clustering Query node.
6. Edit the Clustering Query node.
7. Run the Predictive Query node and view the data.
8. To save the results of the query, use a Create Table or View node.

See Also:

- "Create a Data Source Node"
- "Create Table or View Node"
- "Edit a Clustering Query"
- "Run Predictive Query Node"
- "View Data for a Predictive Query"

10.2.2 Edit a Clustering Query

To edit a Clustering Query node:

1. Double-click the Clustering Query node or right-click the node and click **Edit.** The **Edit Clustering Query Node** dialog box opens.
2. In the **Edit Clustering Query Node** dialog box, enter the details in the following tabs:
   - **Cluster Predictions** tab: In the Case ID field, select a case ID from the drop-down list. The case ID is optional. It is recommended that you specify a Case ID to uniquely define each record. The case ID helps with model repeatability and is consistent with good data mining practices.
     – In the **Case ID** field, select a case ID from the drop-down list. The case ID is optional. It is recommended that you specify a case ID to uniquely define each record. The case ID helps with model repeatability and is consistent with good data mining practices.
     – In the **Number of Clusters to Compute** field, specify the number to compute. Default is `10`.

  - Edit Cluster Prediction Outputs. Oracle Data Miner automatically determines the output for the query. You can modify the output.
- **Partition** tab: You can perform the following tasks:
  - Add one or more Partition attributes. This is optional. Selecting a Partition attribute directs the predictive query to build a virtual model for each unique partition.
  - Select Partitions: To select partitions, click the **Partitions** tab. Then, click . Then use the Add Partitioning Columns to select the partitions. You can also specify partitioning expressions.
- **Input** tab: You can modify Input. This is optional. You can add or remove inputs and change the mining types of inputs. Click **Input**.
- **Additional Output** tab: You can add outputs (optional). By default, all target columns, the Case ID column, and partitioning columns are automatically added to Additional Output. To make changes, click **Additional Output.**

3. Click **OK.**

See Also:

- "Add Partitioning Columns"
- "Add Additional Output"
- "Modify Input"

10.2.2.1 Edit Cluster Prediction Outputs

The output of a Predictive Query is the output of an Apply (scoring) operation. See "Apply Functions" for more information about Apply output.

Oracle Data Miner automatically selects output for query. The default outputs are listed in the Cluster Prediction Outputs section in the Cluster Predictions tab. The defaults are:

- Cluster Details
- Cluster Distance
- Cluster ID
- Cluster Probability

You can also select Cluster Set, and edit parameters of the output functions. You can perform the following tasks:

- **Delete:** To delete an output, select the output and click .
- **Add:** To add an output, click . Use **Add Cluster Function** dialog box to select an output.
- Edit: To edit an output, either double-click the function or select the function and click . Use the **Edit Cluster Function** dialog box to make changes.

10.2.2.2 Add Cluster Function

In the Add Cluster Function dialog box, you can add cluster functions. To add Cluster function:

1. In the **Function** field, select a function from the drop-down list. The options are:
   - Cluster ID

- Cluster Probability
- Cluster Details
- Cluster Distance
- Cluster Set

2. To specify a default name instead of using the default name, deselect **Auto.** This turns off automatic selection.
3. Click **OK.**

10.2.2.3 Edit Cluster Function

To edit Cluster function:

1. Select the change that you want to make to the function.
2. To specify a name instead of using the default name, deselect **Auto.** This turns off automatic selection.
3. Click **OK.**

10.2.3 Clustering Query Properties

The Clustering Query properties enables you to view and change information about this Predictive Query node. If the **Properties** pane is closed, then go to **View** and click **Properties.** Alternately, right-click the node and click **Go to Properties.**

The Clustering Query **Properties** pane has these sections:

- Cluster Predictions
- Partition
- Additional Output (Clustering Query)
- Cache
- Details

10.2.3.1 Cluster Predictions

Displays the predictions produced by the query.

See Also: "Edit a Clustering Query"

10.2.3.2 Additional Output (Clustering Query)

Displays the output specified.

See Also: "Edit Cluster Prediction Outputs"

10.2.4 Clustering Query Context Menu

To view the Clustering Query node context menu, right-click the node. The following options are available in the context menu:

- Connect
- Edit. Opens **Edit a Clustering Query** dialog box.
- Validate Parents
- Run

- Force Run
- View Data. It perform two functions:
  - Runs the node on a small sample of data.
  - Opens the **View Data for a Predictive Query** dialog box.
- Save SQL
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed if there are validation errors.
- Go to Properties
- Navigate

10.3 Feature Extraction Query

A Feature Extraction Query extracts features from the input.

Note: Predictive Query nodes require Oracle 12*c* Release 1 or later.

A Feature Extraction Query node can run in parallel.

This section on Feature Extraction Query node contains the following topics:

- Create a Feature Extraction Query
- Edit Feature Extraction Query
- Run Predictive Query Node
- View Data for a Predictive Query
- Feature Extraction Query Properties
- Feature Extraction Query Context Menu

See Also:

- "About Parallel Processing"
- "Feature Extraction and Selection"

10.3.1 Create a Feature Extraction Query

To create a Feature Extraction Query to an existing workflow:

1. In the **Components** pane, go to **Workflow Editor.** If the **Components** pane is not visible, then go to **View** and click **Components.**
   Alternately, press Ctrl+Shift+P to dock the **Components** pane.
2. Create a Data Source node containing the input data.
3. Expand the Predictive Queries section in the **Components** pane.
4. Drag and drop a Feature Extraction Query node from the **Components** pane to the **Workflow** pane.
5. Connect the Data Source node to the Feature Extraction Query node.
6. Edit the Feature Extraction Query node.
7. Run the Predictive Query node and view the data.
8. To save the results of the query, use a Create Table or View node.

See Also:

- "Create a Data Source Node"
- "Create Table or View Node"
- "Edit Feature Extraction Query"
- "Link Nodes"
- "Run Predictive Query Node"
- "View Data for a Predictive Query"

10.3.2 Edit Feature Extraction Query

To edit a a Feature Extraction Query Node:

1. Double-click the Feature Extraction Query node or right-click the node and click **Edit.** The **Edit Feature Extraction Query Node** dialog box opens.
2. In the **Feature Extraction Query Node** dialog box, enter the details in the following tabs:
   - **Feature Predictions** tab:
     – In the **Case ID** field, select a case ID from the drop-down list. The case ID is optional. It is recommended that you specify a case ID to uniquely define each record. The case ID helps with model repeatability and is consistent with good data mining practices.
     – In the **Number of Features to Extract** field, specify an input. The default is `10`.
     – Edit Feature Prediction Outputs. Oracle Data Miner automatically determines the output for the query. You can modify the output.
   - **Partitions** tab:
     – Add one or more Partition attributes. This is optional. Selecting a Partition attribute directs the predictive query to build a virtual model for each unique partition.
     To add partitions, click in the Partitions tab. Use the Add Partitioning Columns dialog box.

- Specify partitioning expressions.

- **Input** tab: You can perform the following tasks:
  - Add input
  - Modify input
  - Change the mining type
- **Additional Output** tab: You can add output. This is optional. By default, all target columns, the Case ID column, and partitioning columns are automatically added to Additional Output. To make changes, click **Additional Output.**

3. Click **OK.**

See Also:

- "Add Additional Output"
- "Add Partitioning Columns"
- "Modify Input"

10.3.2.1 Edit Feature Prediction Outputs

The output of a Predictive Query is the output of an Apply (scoring) operation. Oracle Data Miner automatically selects output for query. The default output is Feature Set. You can also select feature ID, feature details, and feature value. You can edit the parameters of the functions and perform the following tasks:

- **Delete:** To delete an output, select the output and click .
- **Add:** To add an output, click . Use the **Add Feature Function** dialog box to select an output.
- **Edit:** To edit an output, either double-click the function or select the function and click . Use the **Edit Feature Function** dialog box to make changes.

See Also:

- "Apply Functions"
- "Add Feature Function"
- "Edit Feature Function"

10.3.2.2 Add Feature Function

To add a Feature Function:

1. In the **Function** field, select a function from the drop-down list. The options are:
   - Feature ID
   - Feature Value
   - Feature Details
   - Feature Set
2. To specify a default name instead of using the default name, deselect **Auto.** This turns off automatic selection.
3. Click **OK.**

10.3.2.3 Edit Feature Function

To edit a Feature Function:

1. Select the change that you want to make to the function.
2. To specify a name instead of using the default name, deselect **Auto.** This turns off automatic selection.
3. Click **OK.**

10.3.3 Feature Extraction Query Properties

The Feature Extraction Query properties enables you to view and change information about this Predictive Query node. If the **Properties** pane is closed, then go to **View** and **Properties.** Alternately, right-click the node and click **Go to Properties.**

The Feature Extraction Query Properties has these sections:

- Feature Predictions
- Partition
- Additional Output (Feature Extraction Query)
- Cache
- Details

10.3.3.1 Feature Predictions

Displays the predictions produced by the query.

> **See Also:** "Edit Feature Extraction Query"

10.3.3.2 Additional Output (Feature Extraction Query)

Displays the output specified.

> **See Also:** "Edit Feature Prediction Outputs"

10.3.4 Feature Extraction Query Context Menu

To view the Feature Extraction Query node context menu, right-click the node. The following options are available in the context menu:

- Connect
- Edit. Opens **Edit Feature Extraction Query** dialog box.
- Validate Parents
- Run
- Force Run
- View Data. It perform two functions:
  - Runs the node on a small sample of the data.
  - Opens the **View Data for a Predictive Query** dialog box.
- Save SQL
- Deploy
- Show Graph

- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed if there are validation errors.
- Go to Properties
- Navigate

10.4 Prediction Query

A Prediction Query node performs classification and regression using the input. The data type of the target determines whether classification or regression is performed.

A Prediction Query can run in parallel.

Note: Predictive Query nodes require Oracle 12*c* Release 1 or higher.

This section on Prediction Query node contains the following topics:

- Create a Prediction Query
- Edit a Prediction Query
- Run Predictive Query Node
- View Data for a Predictive Query
- Prediction Query Properties
- Prediction Query Node Context Menu

See Also:

- "About Parallel Processing"
- "Classification"
- "Regression"

10.4.1 Create a Prediction Query

To create a Prediction Query in an existing workflow:

1. In the **Components** pane, go to **Workflow Editor.** If the **Components** pane is not visible, then go to **View** and click **Components.**
   Alternately, press Ctrl+Shift+P to dock the **Components** pane.
2. Create a Data Source node containing the input data.
3. Expand the Predictive Queries section in the **Components** pane.

4. Drag and drop the node from the **Components** pane to the **Workflow** pane.
5. Connect the Data Source Node to the Prediction Query Node.
6. Edit the Prediction Query node.
7. Run the Predictive Query node and view the date.
8. To save the results of the query, use a Create Table or View node.

   See Also:

   - "Create a Data Source Node"
   - "Create Table or View Node"
   - "Edit a Prediction Query"
   - "Link Nodes"
   - "Run Predictive Query Node"
   - "View Data for a Predictive Query"

10.4.2 Edit a Prediction Query

To edit a Prediction Query node:

1. Double-click the Prediction Query node or right-click the node and click **Edit.** The **Edit Prediction Query Node** dialog box opens.
2. In the **Edit Prediction Query Node** dialog box, enter the details in the following tabs:
   - **Predictions** tab:
     – In the **Case ID** field, select a case ID from the drop-down list. The case ID is optional. It is recommended that you specify a case ID to uniquely define each record. The case ID helps with model repeatability and is consistent with good data mining practices.
     – In the **Targets** section, you can add one or more targets. To add target, click . Use the **Add Target** dialog box to define targets.
     – In the **Targets** section, change the Mining Type of a target, if necessary. Each attribute has an associated data type and mining type. The mining type defines how the attribute is treated in the predictive query.
     For Regression query analysis, the mining type must be `Numerical`.
     For Classification query analysis, the mining type must be `Categorical`.
     If you make changes, click **OK.**
     – In the **Prediction Output** section, you can edit the prediction output. Oracle Data Miner automatically determines the output for each target. You can modify the output.
   - **Partition** tab: You can perform the following tasks:
     – Add one or more Partition attributes. This is optional. Selecting a Partition attribute directs the predictive query to build a virtual model for each unique partition.
     To add partitions, click . Use the Add Partitioning Columns dialog box to select partitions.
     – Specify partitioning expressions.
   - **Input** tab: You can perform the following tasks:

  - Modify input
  - Add or remove inputs
  - Change mining types of inputs
- **Additional Output** tab: You can add outputs (optional). By default, all target columns, the Case ID column, and partitioning columns are automatically added to Additional Output. To make changes, click **Additional Output.**

3. Click **OK.**

Tip:

- "Add Additional Output"
- "Add Partitioning Columns"
- "Add Target"
- "Edit Prediction Output"
- "Modify Input"

10.4.2.1 Add Target

You must add at least one target. Targets can have different mining types. To add a target:

1. Select one or more attributes in the **Available Attributes** list to serve as prediction targets. The targets do not have to have the same data type.

   You can select nested attributes as targets if they are of type `ODMR_NETSTED_*`. For example, you could use a join to create a nested attribute consisting of all products purchased by a customer; this attribute would be nested and have the data type `ODMR_NESTED_VARCHAR2`.

2. Move the target columns to the **Selected Attributes** list using the arrows.
3. Click **OK**. The selected attributes are added to the Targets list.

10.4.2.2 Add Partitioning Columns

Partitioning columns result in building a virtual model for each unique partition. Because the virtual model uses data only from a specific partition, it can potentially predict cases more accurately than if you did not select a partition.

In addition to selecting attributes, you can specify partitioning expressions. Partitioning expressions are concatenated and the result expression is the same for all predictive functions.

1. Select one or more attributes in the **Available Attributes** list to serve as partitions.
2. Move the selected columns to the **Selected Attributes** list using the arrows.
3. Click **OK**. The attributes are moved to the Partition list.

Optionally, you can add partitioning expressions.

See Also: "Add Partitioning Expressions"

10.4.2.2.1 Add Partitioning Expressions To specify a partitioning expression, click . Use **Expression Builder** to create an expression.

Suppose one of the partitions is AGE. Here is a sample partitioning expression:

```
CASE WHEN AGE < 20 THEN 1
     WHEN AGE >=20 AND AGE < 40 THEN 2
     WHEN AGE >=40 AND AGE < 60 THEN 3
     ELSE 4
  END
```

Suppose this expression is named `Expression_1`. After you run the node, the output includes a column titled `Expression_1`. This column will contain the value `1` if `AGE` is less than `20`, `2` if `AGE` is `20` or larger but less than `40`, and so forth.

> **See Also:** "Expression Builder"

10.4.2.3 Edit Prediction Output

The output of a Predictive Query is the output of an Apply (Scoring) operation. Oracle Data Miner automatically selects output for the target. The default output is listed in the **Prediction Outputs** section in the **Predictions** tab.

- For Classification, the default output are:
  - Prediction
  - Prediction Details
  - Prediction Probability

  You can also select Prediction Set.
- For Regression, the default outputs are:
  - Prediction
  - Prediction Details

For Classification or Regression, you can edit the parameters of functions and the output for each target one at a time. You can perform the following tasks:

- To edit Prediction Output Function, select a target and click . You edit output for each target one at a time. Use the **Edit Prediction Function** dialog box to make changes.
- To delete an output, select the output and click .
- To add an output, select the target in the Targets section and click . Use the **Add Prediction Output Function** dialog box to select an output.

> **See Also:**
> - "Apply Functions"
> - "Add Prediction Output Function"
> - "Edit Prediction Function Dialog"

10.4.2.3.1 Add Prediction Output Function

To add Prediction Output Function:

1. In the **Function** field, select a function from the drop-down list.
2. To specify a default name instead of using the default name, deselect **Auto.** This turns off automatic selection.
3. Click **OK.**

10.4.2.3.2 Edit Prediction Function Dialog

To edit Prediction Function:

1. Select the change that you want to make to the function.
2. To specify a name instead of using the default name, deselect **Auto.** This turns off automatic selection.
3. Click **OK.**

10.4.2.4 Modify Input

The **Input** tab shows all columns that are used as input for the Predictive Query. Target (for Prediction Query) and Case ID columns are identified with special icons. The Rule column in the grid explains why an attribute is not used.

By default, **Determine inputs automatically (using heuristics)** is selected. After you run the node, you can click the link to View Heuristic Results Details. To change the inputs, deselect **Determine inputs automatically (using heuristics)**. You can perform the following tasks:

- Override defaults, and add or remove input columns.
- Change Mining Types: To change the mining type of a column, click the **Mining Type** entry for the column, and select a new mining type from the drop-down list.
- Ignore columns: If you do not want to use a column as input, click the Input entry for the attribute and select from the drop-down list. It ignored the selected column, and not used for input.
  To use a column, select from the drop-down list.
- Search column: To search for columns, use the **Find** field.

10.4.2.4.1 View Heuristic Results Details

Provides detailed information about automatic changes made to the input. For example, the mining type is changed to `Categorical` when the number of unique values is less than the threshold value of `5.` A column that has a constant value is excluded (not used as input).

10.4.2.5 Add Additional Output

The **Output** tab shows the columns that will be used in the output to identify the prediction data. By default, all target columns for Prediction Query, the Case ID column, and partitioning columns are automatically added to Additional Output. You can perform the following tasks:

- Add Additional Output: To add additional output, click **Automatic** to turn off automatic selection. Then click . Use the **Add Supplemental Dialog** to add columns to the output.
- Remove Output Columns: To remove columns, select the column and click .

10.4.2.5.1 Add Supplemental Dialog

Enables you to move columns from the **Available Attributes** list to the **Selected Attributes** list.

10.4.3 Run Predictive Query Node

To run Predictive Query nodes, right-click the node and select either **Run** or **View Data.** Virtual models may take a while to be formulated. The View Data option generates a small sample output of the query.

Regardless of how you run it, select **View Data** to view the results of the query.

See Also: "View Data for a Predictive Query"

10.4.4 View Data for a Predictive Query

To view data for a Predictive Query node:

1. Right-click the node and select View Data. For Predictive Query Nodes, **View Data** displays the output from a node that has run or displays the results when the query is applied to a small subset of the data.
2. You can either sort or filter the data.
3. Click **OK.**

You can view prediction details to see Prediction Details in a separate window.

See Also:

- "Data Source Node Viewer"
- "View Prediction Details"

10.4.4.1 View Prediction Details

To view the prediction details:

1. Click the details that you want to view.
2. Then, click . The details are displayed in the **View Value** dialog box. You can also search for specific values.

10.4.5 Prediction Query Properties

The Prediction Query properties enables you to view and change information about this Predictive Query node. If the **Properties** pane is closed, then go to **View** and click **Properties.** Alternately, right-click the node and click **Go to Properties.**

The Prediction Query Node Properties has these sections:

- Predictions (Prediction Query)
- Partition
- Additional Output (Prediction Query)
- Cache
- Details

10.4.5.1 Predictions (Prediction Query)

Displays the predictions produced by the query.

See Also: "Modify Input"

10.4.5.2 Partition

Displays the columns selected to be used for partitioning. You can also delete partitioning columns and edit expressions for partitioning columns.

See Also: "Add Partitioning Columns"

10.4.5.3 Additional Output (Prediction Query)

Displays the output specified.

See Also: "Add Prediction Output Function"

10.4.6 Prediction Query Node Context Menu

To view the Prediction Query node context menu, right-click the node. The following options are available in the context menu:

- Connect
- Edit. Opens **Edit a Prediction Query** dialog box.
- Validate Parents
- Run
- Force Run
- View Data. It perform two functions:
  - Runs the node on a small sample of the data.
  - Opens the **View Data for a Predictive Query** dialog box.
- Save SQL
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed only if there is an error.
- Show Validation Errors. Displayed if there are validation errors.
- Go to Properties
- Navigate

11

Text Nodes

Text nodes are available in the Text section of the Components pane. Oracle Data Miner supports the following Text nodes:

- Apply Text Node
- Build Text
- Text Reference

If you are connected to Oracle Database12*c*, then you can use Automatic Data Preparation (ADP) to prepare text data using the Text tab to specify data usage.

Note: The Oracle Text Knowledge Base is required for text processing.

To install the Knowledge Base, you must install the Oracle Database Examples. For directions on how to install the examples, see Oracle Database Examples Installation Guide for the release of Oracle Database that you connect to.

See Also:

- "Specifying Text Characteristics"
- "Text Mining in Oracle Data Mining"
- "Oracle Text Concepts"
- *Oracle Database Examples Installation Guide* for information about hot to install the examples.

11.1 Oracle Text Concepts

Oracle text concepts include:

- **Theme:** A theme is a topic associated with a given document. A document can have many themes. A theme does not have to appear in a document. For example, a document containing the words `San Francisco` may have `California` as one of its themes.
- **Stopword:** A stopword is a word that is not indexed during text transformations. A stopword is usually a low information word. In English *a*, *the*, *this*, or *with* are usually stopwords.

- **Stoplist:** A stoplist is a list of stopwords. Oracle Text supplies a stoplist for every language. By default during indexing, the system uses the Oracle Text default stoplist for your language. You can edit the default stoplist or create a new one.

> **Note:** In Oracle Data Miner, stoplists are shared across all transformations and are not owned by a specific transformation.

- **Stoptheme:** A stoptheme is a theme to be skipped over during indexing. Stopthemes are specified by adding them to stoplists.

Oracle Text uses stopwords and stopthemes to indicate text that can be safely ignored during text mining.

The Oracle Text Lexer breaks source text into tokens or themes—usually words—in accordance with a specified language. To extract tokens, the Lexer uses parameters as defined by a lexer preference. These parameters include:

- Definitions for the characters that separate tokens. For example, whitespace.
- Conditions to convert text to all uppercase or not.
- Text analysis text to create theme tokens. This is done when theme indexing is enabled.

See Also: *Oracle Text Reference*

11.2 Text Mining in Oracle Data Mining

Text must undergo a transformation process before it can be mined. After the data has been properly transformed, a case table can be used for building, testing, or scoring data mining models. The case table must be a relational table. It cannot be created as a view.

A Source table for Oracle Data Mining can include one or more columns of text. A text column cannot be used as a target.

The following Oracle Data Mining algorithms support text:

- Anomaly Detection (one-class Support Vector Machine)
- Classification algorithms: Naive Bayes, Generalized Linear Models, and Support Vector Machine

  Decision Tree when you connect to Oracle Database 12*c*
- Clustering algorithms: *k*- Means and Expectation Maximization
- Feature Extraction algorithms: Nonnegative Matrix Factorization, Singular Value Decomposition, and Principal Components Analysis
- Regression algorithms: Generalized Linear Models and Support Vector Machine

---

Note: These algorithms do *not* support text:

- O-Cluster
- Decision Tree when you connect to Oracle Database 11*g*
- Association (Apriori)

Any text attributes are automatically filtered out for model builds when you use O-Cluster or Decision Tree connected to Oracle Database 11*g*.

See Also: "Data Preparation for Text"

11.2.1 Data Preparation for Text

Data Preparation for text depends on which version of Oracle Database that you connect to:

- Text Processing in Oracle Data Mining 12c Release 1 (12.1)
- Text Processing in Oracle Data Mining 11g Release 2 (11.2) and Earlier

11.2.1.1 Text Processing in Oracle Data Mining 12*c* Release 1 (12.1)

In Oracle Data Mining 12*c* Release 1 (12.1) and earlier, if unstructured text data is present, then text processing includes text transformation before text mining. Oracle Data Mining includes significant enhancements in text processing that simplify the data mining process (model build, deployment, and scoring) when unstructured text data is present in the input. Some points about unstructured text and text transformation:

- Unstructured text includes data items such as web pages, document libraries, Microsoft Power Point presentations, product specifications, email messages, comment fields in reports, and call center notes.
  - `CLOB` columns and long `VARCHAR2` columns are automatically interpreted as unstructured text by Oracle Data Mining.
  - Columns of short `VARCHAR2`, `CHAR`, `BLOB`, and `BFILE` can be specified as unstructured text.
- To transform unstructured text for mining, Oracle Data Mining uses Oracle Text utilities and term weighting strategies.
- Text terms are extracted and given numeric values in a text index.
- Text transformation process is configurable for models and individual attributes. You can specify data preparation for text nodes when you define a model node.

After text transformation, the text can be mined with a data mining algorithm.

Note: If you connect to Oracle 12*c* Release 1 or higher, then it is not always necessary to use the Text nodes, Apply Text Node, Build Text, and Text Reference.

See Also: "Specifying Text Characteristics"

11.2.1.2 Text Processing in Oracle Data Mining 11*g* Release 2 (11.2) and Earlier

In Oracle Data Mining 11*g* Release 2 (11.2) and earlier, before text mining is done, it must undergo the following processes:

- **Extraction** or **Feature Extraction:** This is a special preprocessing step, where the text is broken down into units (terms) that can be mined. Text terms can be keywords or other document-derived features.
- **Text preparation:** Text preparation uses a Build Text node to transform text columns. Build Text does not support HTML or XML documents. It also does not support any binary data types.

Oracle Data Miner uses the facilities of Oracle Text to preprocess text columns.

> **Note:** You must preprocess text using the Text nodes, Apply Text Node, Build Text, and Text Reference.

11.3 Apply Text Node

The Apply Text node enables you to apply existing text transformations from either a Build Text node or from a Text Node to new data. This ensure that the apply data is transformed in the same way that the build data was transformed.

Apply Text can run in parallel. See "About Parallel Processing" for more information.

This section about Apply Text node contains the following topics:

- Default Behavior for the Apply Text Node
- Create an Apply Text Node
- Edit Apply Text Node
- Apply Text Node Properties
- Apply Text Node Context Menu

11.3.1 Default Behavior for the Apply Text Node

Apply Text node applies existing text transformations from either a Build Text or Text Node to new data. This ensures that the apply data is transformed in the same way that the build data was transformed.

> **Note:** All models in the node must have the same case ID.

11.3.2 Create an Apply Text Node

Before creating the Apply Text node, first create a workflow. Then create a Data Source node.

To create an Apply Text node:

1. In the **Components** pane, go to **Workflow Editor.** If the **Components** pane is not visible, then go to **View** and click **Components.** Alternately, press Ctrl+Shift+P to dock the **Components** pane.
2. In the **Workflow Editor,** expand **Text** and click **Apply Text.**
3. Drag and drop the Apply Text node from the **Components** pane to the **Workflow** pane.

The GUI shows that the node has no data associated with it. Therefore, it cannot be run.

4. Move to the node that provides data for Apply Text. Right-click and select **Connect.** Drag the line to the Build Text node and click again.

> **Note:** The Apply data must be compatible with the Build data.

5. Move to the Build Text node or Text node that indicates how the text columns were prepared. For example, go to the Build Text node for the model to be applied. Link the Build Text node to the Apply node.
6. To view or modify the transformation details, right-click the Apply Text node and select **Edit.** This opens the **Edit Apply Text Node** dialog box.
7. The node is ready to run. Right-click the node and select **Run.**

See Also: "Edit Apply Text Node"

11.3.3 Edit Apply Text Node

The **Edit Apply Text Node** dialog box enables you to view the text transformations performed on the Build data. The Apply data must be prepared in the same way as the Build data.

To open the **Edit Apply Text Node** dialog box right-click the Apply Text node and select **Edit** or just double-click the node.

1. Right-click the Apply Text node and click **Edit.** Alternately, you can double-click the Apply Text node.
2. The Edit Apply Text Node has two panes:
   - In the top pane, you can perform the following tasks:
     – **Case ID:** Specify a case ID in this field. This is optional.
     – **View Attributes:** Select **All** or **Text and Transformed** from the drop-down list. For each attribute, the following are displayed:
     Type: The data type of the attribute. The type of an attribute that has a text transform applied is `DM_NESTED_NUMERICALS`.
     Source: The source column for a transformed column.
     Transform: The type of text transform—Token or Theme.
     Output: Indicates if the attribute is passed on to subsequent nodes. By default, all nodes are passed on.
     – **View Stoplist.** To view the stoplist, select a transformed column and click **View Stoplist**. The **Stoplist Editor** starts. You can view the items in the stoplist.
     – **View and Edit text transform:** To view the definition of a text transformation, select a transformed attribute and click . This opens the **Add/Edit Text Transform** dialog box.
     – **Exclude attributes:** To exclude attributes, select the option and in the Output column of the grid, click . The icon changes to . The excluded attribute is not passed on to subsequent operations. You may want to exclude the non-transformed version of a text column.
     – **Include attributes:** To include an attribute, click the icon again. The icon changes to , indicating that it is included.

- In the lower pane, you can view the text transformation after running the node.

3. Click **OK.**

See Also:

- "Add/Edit Text Transform"
- "Stoplist Editor"
- "View the Text Transform (Apply)"

11.3.3.1 View the Text Transform (Apply)

You can view the effects of the text transformation defined in the Build Text node or the Text node in the View Text Transform window. To access the View Text Transform window:

1. Open the **Edit Apply Text Node** dialog box by double-clicking the Apply Text node or right-clicking the node and select **Edit.**
2. In the Upper pane, select the name of the transformed attribute.
3. In the lower pane:
   - Click **Tokens** or **Themes**. A grid displays all of the tokens or themes in the document and the frequency of each token or theme. Use the search field to search for topics or themes by name, the default, or by frequency.
   - Click the **Output** to view the tokens or themes for a sample of the attributes. Output Sample lists a sample of the attributes listed by case ID, if you specified one, or by row ID if you did not specify a case ID. You can search the IDs.
     Click an ID. The original text from the untransformed attribute is displayed along with the tokens or themes identified in it, along with the frequency of each token or theme.
4. Click **OK.**

11.3.4 Apply Text Node Properties

The Apply Text node Properties enables you to view and change information about the models defined in the node.

To view the Properties for a node, click the node. If Properties is closed, then go to **View** and click **Properties.** Alternately, right-click the node and click **Go to Properties.**

The Apply Text Properties has the following sections:

- Transforms
- Cache
- Sample
- Details

See Also: "Properties"

11.3.4.1 Transforms

Displays transformations defined in the **Edit Build Text Node** dialog box. You can edit transformations from this tab.

See Also: "Edit Build Text Node"

11.3.4.2 Details

Displays the node name and comments about the node. You can change the name of the node and edit the comments from this tab. The new node name and comments must satisfy the node name and node comments requirements.

See Also: "Node Name and Node Comments"

11.3.5 Apply Text Node Context Menu

To view the Apply Text node context menu, right-click the node. The following options are available in the context menu:

- Connect
- Edit. Opens Edit Apply Text Node dialog box.
- Validate Parents
- Run
- Force Run
- View Data
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed if there are errors.
- Show Validation Errors. Displayed if there are validation errors.
- Navigate

11.4 Build Text

Build Text node prepares a data source that has one or more Text columns. You can use the data to build models.

Build Text can run in parallel.

This section about Build Text node consists of the following topics:

- Default Behavior for the Build Text Node
- Create Build Text Node
- Edit Build Text Node

- Build Text Node Properties
- Build Text Node Context Menu

See Also: "About Parallel Processing"

11.4.1 Default Behavior for the Build Text Node

The Build Text node enables you to define a text transformation for each test column. You can use the transformed columns to build models using any algorithm that supports text.

Note: O-Cluster and Decision Tree do *not* support text.

A Build Text node builds one model using the NMF algorithm by default. The transformed column or columns are passed to subsequent nodes and the non-transformed columns are not passed on.

All models in the node have the same case ID.

11.4.2 Create Build Text Node

Before creating a Build Text node, create a workflow. Then identify or create a Data Source node.

To create a Build Text node:

1. In the **Components** pane, go to **Workflow Editor.** If the **Components** pane is not visible, then go to **View** and click **Components.** Alternately, press Ctrl+Shift+P to dock the **Components** pane.
2. In the **Workflow Editor,** expand **Text** and click **Build Text.**
3. Drag and drop the node from the **Components** pane to the **Workflow** pane. The GUI shows that the node has no data associated with it. Therefore, it cannot be run.
4. Move to the node that provides data for the Build Text node. Right-click and select **Connect.** Drag the line to the Build Text node and click again.
5. Either accept the default settings or edit the text details. To edit the transformation details, right-click the node and select **Edit.** The **Edit Build Text Node** dialog box opens.
6. The node is ready to run. Right-click the node and select **Run.**

See Also: "Edit Build Text Node"

11.4.3 Edit Build Text Node

The **Edit Build Text Node** dialog box enables you to define transformations for text columns. The transformed text columns can be used in data mining.

To open the **Edit Build Text Node** dialog box:

1. Right-click the Build Text node and select **Edit.** Alternately, double-click the node. The **Edit Build Text Node** dialog box opens.
2. The **Edit Build Text Node** dialog box has two panes.
   - In the top pane, you can perform the following tasks:

- Specify the case ID (optional).
- Open the **Stoplist Editor.**
- **View attributes:** Select **All** or **Text and Transformed** from the drop-down list. For each attribute, the following are displayed:
  Type: The data type of the attribute. The type of an attribute that has a text transform applied is `DM_NESTED_NUMERICALS`.
  Source: The source column for a transformed column.
  Transform: The type of text transform—Token or Theme.
  Output: Indicates if the attribute is passed on to subsequent nodes. By default, all nodes are passed on.
- **Define transformation:** To define a transformation, select a text attribute and click . Define the text transformation in the **Add/Edit Text Transform** dialog box. Repeat the step for each text attribute.
- **Edit Transformation:** Select the transformed attribute and click .
- **Delete Transformation:** Select the transformed attribute and click .
- **Exclude attribute:** To exclude attributes, select it and in the Output column of the grid, click . The icon changes to . The excluded attribute is not passed on to subsequent operations. You may want to exclude the non-transformed version of a text column.
- **Include attribute:** To include an attribute, click the icon again. The icon changes to , indicating that it is included.

- In the lower pane, you can view the test transformation after the node is run.

3. Click **OK.**

See Also:

- "Add/Edit Text Transform"
- "View the Text Transform"
- "Stoplist Editor"

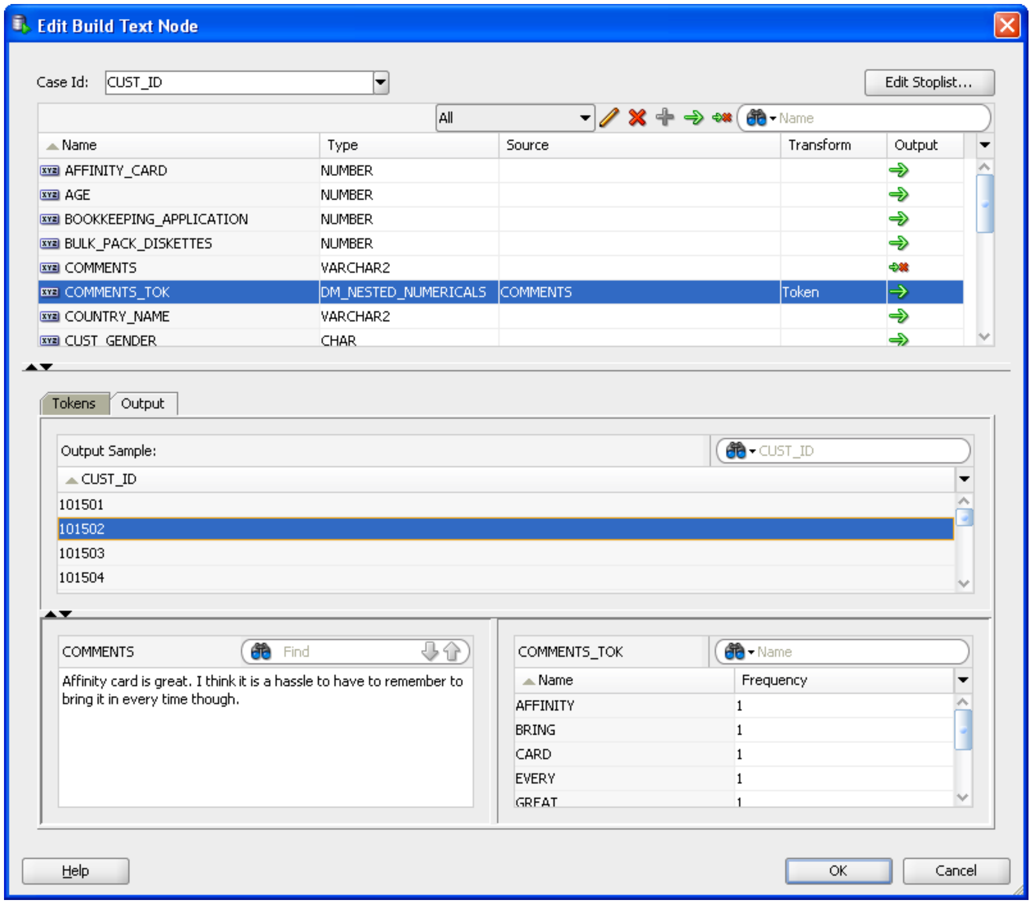
11.4.3.1 View the Text Transform

To view the effects of a text transformation:

1. Open the **Edit Build Text Node** dialog box by double-clicking the Build Text node or right-clicking the node and select **Edit.**
2. In the upper pane, select the name of the transformed attribute.
3. In the lower pane:
   - Click **Tokens** or **Themes.** A grid displays all the tokens or themes in the document and the frequency of each token or theme. Use the search field to search for topics or themes by name, the default, or by frequency.
   - To add a token or theme to the stoplist, select it and click **Add to Stoplist**.
   - Click **Output** to view the tokens or themes for a sample of the attributes. Output Sample lists a sample of the attributes listed by case ID, if you specified one, or by row ID if you did not specify a case ID. You can search the IDs.

- Click an ID. The original text from the non-transformed attribute is displayed along with the tokens or themes identified in it, along with the frequency of each token or theme.

4. Click **OK.**

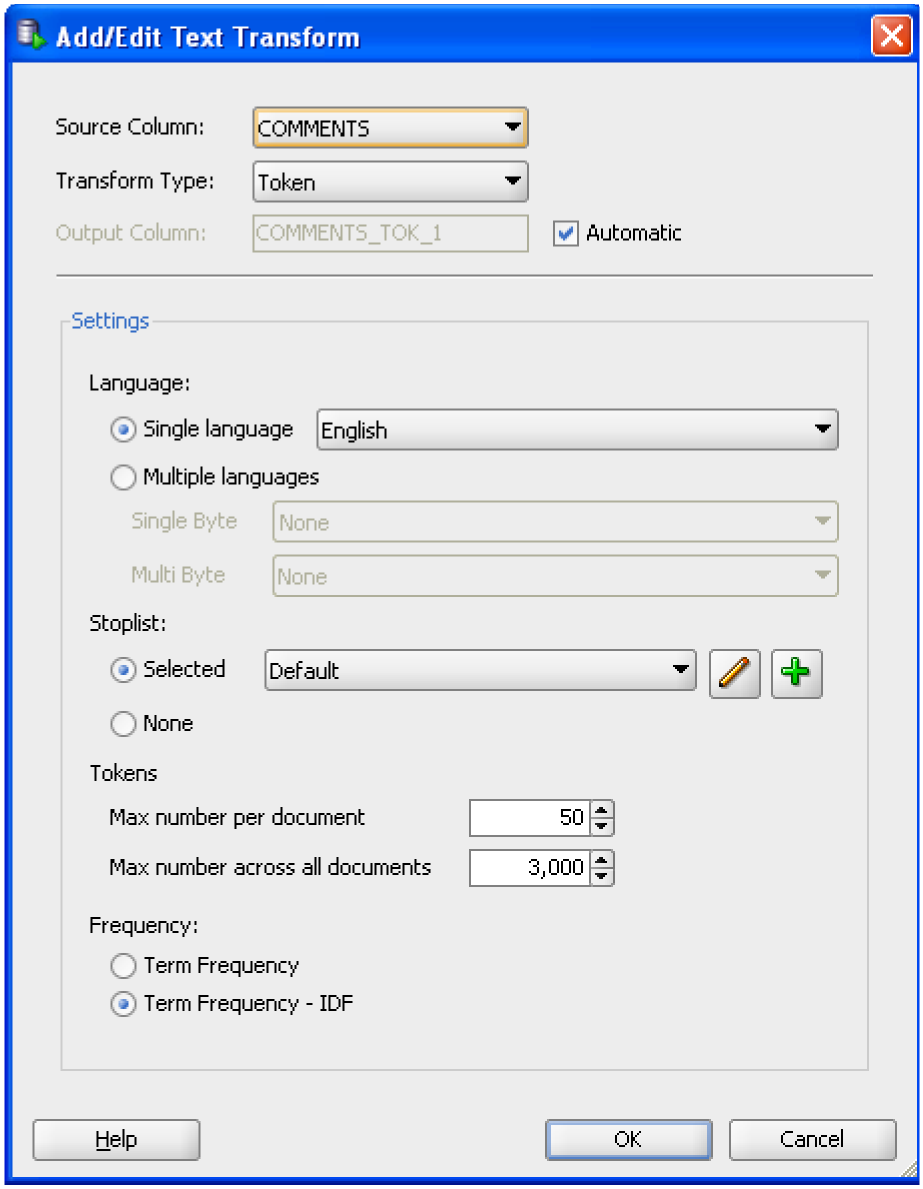


The **Edit Build Text** dialog box.

****

11.4.3.2 Add/Edit Text Transform

The **Add/Edit Text Transform** dialog box can be opened from the **Edit Build Text Node** dialog box. To open or edit a text transform, click The default values for the transformation are illustrated in this graphic:



The Add/Edit Text Transform dialog box displaying the default values for text transformation.

****

- **Source Column:** This is the name of the column to be transformed.
- **Transform Type:** This is either Token (the default) or Theme.
- **Output Column:** This is the name of the new column. The default name is the source column name with either TOK (for Token) or THM (for Theme) appended, depending on the transformation type. To specify the output column name, deselect **Automatic** and enter a name in the **Output Column** field.

In the **Settings** section, specify characteristics of the text and the transform:

- **Language**: Select any one of the following options:
  - **Single Language:** By default, a single language is specified. English is the default language. You can select a different language.

- **Multiple Language:** Select this option to specify multiple language. For example, to specify Single Byte languages, such as Arabic, Turkish, Thai, and European languages, select them from the Single Byte list.
  To specify Multibyte languages, such as Chinese (simplified or traditional), Japanese or Korean, select them from the Multibyte languages.

- **Stoplist**: Oracle Text provides default stoplists for several single languages. If there is a default stoplist, it is selected. For several languages, the default is no stoplist. You can select any stoplist that was previously created for this attribute from the drop-down list. You can perform the following tasks:
  - **Edit a Stoplist:** To edit a stoplist, click . The **Stoplist Editor** opens.
  - **Add a Stoplist:** To add a stoplist, click . The **Stoplist Editor** opens.

- **Token:** If you select Token, the defaults are:
  - Maximum number per document: `50` (default)
  - Maximum number across all document: `3000` (default)

  You can change these values. The tokens per document and across all documents cutoffs are for rankings, not for an absolute count of tokens. You could have more than 3000 tokens across all documents if there were ties.

- **Theme:** If you select theme, the defaults are:
  - Maximum number per documents: `50` (default)
  - Maximum number across all document: `3000` (default)

  You can change these values. The themes per document and across all documents cutoffs are for rankings, not for absolute count of themes. You could have more then 3000 themes across all documents if there were ties.

  Theme incudes a **Theme Type** specification. The default is `Single.` You can select `Full.`

- **Frequency**: The default is `Term Frequency.` You can select `Term Frequency IDF.`

  Note: Frequency is a sticky setting. If you change it, then the changed value becomes the default.

  Term Frequency uses the term frequency in the document itself. It does not take collection information into account.

  Term Frequency IDF is the traditional TF-IDF. It takes into account information from the document (Term Frequency) and collection-level information (IDF plus the terms to use if a maximum overall number of terms for the collection is set).

  TF-IDF (Term Frequency–Inverse Document Frequency) is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection. The importance increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the collection.

11.4.3.3 Stoplist Editor

In the Stoplist Editor, you can either edit an existing stoplist, or you can create a new stoplist. Stoplists are shared among all workflows. You can edit any stoplist in this dialog box, not just the ones associated with transformations defined in this node.

To access the stoplist editor, open the **Edit Build Text Node** by double-clicking a Build Text node. To view, edit, and create a stoplist:

1. Click **Edit Stoplist**.
2. The Stoplist Editor opens. All stoplists for all transformations are listed.
3. To add a stoplist, click . The **New Stoplist Editor** wizard opens.
4. To modify an existing stoplist, select the stoplist from the Custom Stoplists list.

   The items in the stoplist are listed in the bottom pane.

   - To delete an item from the stoplist, select the item and click .
   - To add stopwords or stopthemes to the selected list, click . The **Add Stopwords/Stopthemes** dialog box opens.
5. To delete a stoplist, select it in the **Custom Stoplists** list and click .

See Also:

- "Add Stopwords/Stopthemes"
- "New Stoplist Editor"

11.4.3.3.1 New Stoplist Editor

In the **New Stoplist Editor** wizard, you can perform the following tasks:

- Create new stoplists. To create a stoplist, click . The **New Stoplist Wizard** starts. The wizard has two steps:
  – Stoplist Definition
  – Review
- Remove words from an existing stoplist.
- Combine several stoplists to create a new one. For example, if the document is in both French and English, then you can combine the French and English stoplists.
- Create an empty stoplist to which you must add all stopwords and stopthemes.

11.4.3.3.2 Stoplist Definition

Follow these steps to define a stoplist:

1. Either accept the provided name or enter a different name.
2. The default selection **Extends following stoplist(s)** enables you to create a new stoplist by combining and modifying existing stoplists.

   Select one or more stoplists to extend. If you select several stoplists, then they are combined.
3. To create a completely new stoplist, select **Empty**, and select a language for the stoplist. The default is English.
4. Click **Next**.

11.4.3.3.3 Review

Add or remove stopwords and stopthemes.

1. To add items to the stoplist, click . The **Add Stopwords/Stopthemes** dialog box opens.
2. To remove items from the stoplist, select it and click .
3. Click **Finish** when done.

11.4.3.3.4 Add Stopwords/Stopthemes

This dialog box adds stopwords and stopthemes to a stoplist.

1. Enter the stopwords, separated by commas.
2. Click **OK** when you have finished.

11.4.4 Build Text Node Properties

The Build Text node Properties enables you to view and change information about the models defined in the node. If the **Properties** pane is closed, then go to **View** and click **Properties.** Alternately, right-click the node and click **Go to Properties.**

The Build Text node **Properties** pane has the following sections:

- Transforms
- Sample
- Cache
- Details

See Also: "Properties"

11.4.4.1 Transforms

Displays transformations defined in the **Edit Build Text Node** dialog box. You can edit transformation in this tab.

11.4.5 Build Text Node Context Menu

Right-click a Build Text node. The following selections are displayed:

- Connect
- Edit. Edits the text apply. Opens the Edit Build Text Node dialog box.
- Validate Parents
- Run
- Force Run
- View Data
- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed if there are errors.

- Show Validation Errors. Displayed if there are validation errors.
- Navigate

11.5 Text Reference

A Text Reference node enables you to reference text transformations defined in a Build Text node in the current workflow or in a different workflow.

For example, if you have one workflow that builds a Text model (that is, a workflow that includes a Build Text node) and you want to create a separate workflow that applies the model created in the first workflow, then you can use a Text Reference to provide the text transformation information required by Apply Text.

This section contains the following topics:

- Create a Text Reference Node
- Edit Text Reference Node
- Text Reference Node Properties
- Text Reference Node Context Menu

11.5.1 Create a Text Reference Node

Before creating a Text Reference node, create a workflow. Then, identify or create a data source.

To create a Build Text node:

1. In the **Components** pane, go to **Workflow Editor.** If the **Components** pane is not visible, then go to **View** and click **Components.**
   Alternately, press Ctrl+Shift+P to dock the **Components** pane.
2. In the Workflow Editor, expand **Text** and click **Text Reference.**
3. Drag and drop the node from the **Components** pane to the **Workflow** pane.
   The GUI shows that the node does not have a reference to a Build Text node defined.
4. Go to the **Edit Text Reference Node** dialog box to select a Build Text node to reference.
5. The node is ready to be used. Connect it to an Apply Text node. The Text Reference node is used instead of a Build Text node.

See Also:

- "Edit Text Reference Node"
- "Apply Text Node"

11.5.2 Edit Text Reference Node

The **Edit Text Reference Node** dialog box enables you to select a Build Text node so that you can use its transformations in the current location in the current workflow. To open the Edit Text Reference Node:

1. Right-click the node and select **Edit.** Alternately, double-click the node. The E**dit Text Reference Node** dialog box has two panes.
2. In the upper pane, click **Select.** The **Select Build Text Node** dialog box opens.

3. After you select a Build Text node, you can view tokens or themes for any transformed nodes. Select a transformed node in the upper pane.
4. In the bottom pane, the Tokens and Themes and their frequencies are displayed. You can search by token or theme (the default) or by frequency
5. Click **OK.**

See Also: "Select Build Text Node"

11.5.2.1 Select Build Text Node

In the **Select Build Text Node** dialog box, you can select a Build Text node that is either in the current workflow, the default) or in all workflows. **Show** specifies the list of Build Text nodes to select from.

1. In the **Show** field, select either **All Workflows** or **Current Workflow** (default).
2. In the **Search** field, you can search for Build Text nodes by project (the default), workflow, or node.
3. Select a Build Text node from the **Available Nodes** grid. For each Build Text node the grid shows project, workflow, and status.
4. Click **OK.**

Note: You cannot select a Text node that is not complete.

11.5.3 Text Reference Node Properties

The Build Text node Properties enables you to view and change information about the models defined in the node. To view the properties for a node, click the node. If the **Properties** pane is closed, then go to **View** and click **Properties.** Alternately, right-click the node and click **Go to Properties.**

The Build Text **Properties** pane has the following sections:

- Transforms
- Details

See Also: "Properties"

11.5.3.1 Transforms

Displays transforms selected in **Edit Text Reference Node** dialog box. You can select a different Build Text node from the Properties pane.

11.5.4 Text Reference Node Context Menu

Right-click a Text Reference node. The following selections are displayed:

- Connect
- Edit. Edits the text apply. Opens Edit Text Reference Node.
- Validate Parents
- Run
- Force Run
- View Data

- Deploy
- Show Graph
- Generate Apply Chain
- Cut
- Copy
- Paste
- Extended Paste
- Select All
- Parallel Query. See "About Parallel Processing" for more information.
- Show Event Log
- Show Runtime Errors. Displayed if there are errors.
- Show Validation Errors. Displayed if there are validation errors.
- Navigate

12

Testing and Tuning Models

Testing a model enables you to estimate how accurate the predictions of a model are. You can test Classification models and Regression models, and tune Classification models.

This section contains the following topics:

- Testing Classification Models
- Tuning Classification Models
- Testing Regression Models

12.1 Testing Classification Models

Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a Classification project is typically divided into two data sets:

- One for building the model
- One for testing the model

The test data must be compatible with the data used to build the model and must be prepared in the same way that the build data was prepared.

These are the ways to test Classification and Regression models:

- By splitting the input data into build data and test data. This is the default. The test data is created by randomly splitting the build data into two subsets. 40 percent of the input data is used for test data.
- By using all the build data as test data.
- By attaching two Data Source nodes to the build node.
  – The first data source that you connect to the build node is the source of the build data.
  – The second node that you connect is the source of the test data.
- By deselecting **Perform Test** in the **Test** section of the **Properties** pane and using a Test node. The **Test** section define how tests are done. By default, all Classification and Regression models are tested.

Oracle Data Miner provides test metrics for Classification models so that you can evaluate the model.

After testing, you can tune the models.

See Also:

- "Tuning Classification Models"
- "Classification Model Test and Results Viewers"
- "Test Node"
- "Test Metrics for Classification Models"

12.1.1 Test Metrics for Classification Models

Test metrics assess how accurately the model predicts the known values. Test settings specify the metrics to be calculated and control the calculation of the metrics. By default, Oracle Data Miner calculates the following metrics for Classification models:

- Performance measurements, Predictive Confidence, Average Accuracy, Overall Accuracy, and Cost
- Performance Matrix, also know as Confusion Matrix
- Receiver Operating Characteristics
- Lift
- Profit and ROI

You can change the defaults using the preference setting.

To view test results, first test the model or models in the node:

- If you tested the models using the default test in the Classification node:
  1. Run the Classification node.
  2. Right-click the node and select **View Test Results.**
  3. To view the model, select the model that you are interested in. The **Classification Model Test View** opens.
  4. To compare the test results for all models in the node, select **Compare Test Results.**
- If you tested the models using a Test node:
  1. Run the Test node.
  2. Right-click the node and select **View Test Results.**
  3. To view the model, select the model that you are interested in. The **Classification Model Test View** opens.
  4. To compare the test results for all models in the node, select **Compare Test Results.**

See Also:

- "Data Miner Preferences"
- "Classification Model Test Viewer"

12.1.1.1 Performance

The following performance measures are calculated:

- Predictive Confidence
- Average Accuracy

- Overall Accuracy
- Cost

You can view these values separately and view all of them at the same time. To view the performance measures:

1. Select the measure that you want to display. Alternately, click **All Measures** from the **Measures** list.
2. Use the **Sort By** lists to sort the measures. You can sort by:
   - Name (default)
   - Measures
   - Creation date.

   The sort can be by descending order (default) or ascending order.

12.1.1.1.1 Predictive Confidence Predictive Confidence provides an estimate of accurate the model is. Predictive Confidence is a number between 0 and 1. Oracle Data Miner displays Predictive Confidence as a percentage. For example, the Predictive Confidence of 59 means that the Predictive Confidence is 59 percent (0.59).

Predictive Confidence indicates how much better the predictions made by the tested model are than predictions made by a naive model. The naive model always predicts the mean for numerical targets and the mode for categorical targets.

Predictive Confidence is defined by the following formula:

```
Predictive Confidence = MAX[(1-Error of model/Error of Naive Model),0]X100
```

Where:

Error of Model is (1 - Average Accuracy/100)

Error of Naive Model is (Number of target classes - 1) / Number of target classes

- If the Predictive Confidence is `0`, then it indicates that the predictions of the model are no better than the predictions made by using the naive model.
- If the Predictive Confidence is `1`, then it indicates that the predictions are perfect.
- If the Predictive Confidence is `0.5`, then it indicates that the model has reduced the error of a naive model by 50 percent.

12.1.1.1.2 Average Accuracy Average Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data. The formula to calculate the Average Accuracy is:

```
Average Accuracy = (TP/(TP+FP)+TN/(FN+TN))/Number of classes*100
```

Where:

- TP is True Positive.
- TN is True Negative.
- FP is False Positive.
- FN is False Negative.

The average per-class accuracy achieved at a specific probability threshold that is greater than the accuracy achieved at all other possible thresholds.

12.1.1.1.3 Overall Accuracy Overall Accuracy refers to the percentage of correct predictions made by the model when compared with the actual classifications in the test data. The formula to calculate the Overall Accuracy is:

```
Overall Accuracy = (TP+TN)/(TP+FP+FN+TN)*100
```

Where:

- TP is True Positive.
- TN is True Negative.
- FP is False Positive.
- FN is False Negative.

12.1.1.1.4 Cost In a classification problem, it may be important to specify the costs involved in making an incorrect decision. By doing so, it can be useful when the costs of different misclassifications vary significantly.

For example, suppose the problem is to predict whether a user is likely to respond to a promotional mailing. The target has two categories: YES (the customer responds) and NO (the customer does not respond). Suppose a positive response to the promotion generates $500 and that it costs $5 to do the mailing. Then, the scenarios are:

- If the model predicts YES, and the actual value is YES, then the cost of misclassification is $0.
- If the model predicts YES, and the actual value is NO, then the cost of misclassification is $5.
- If the model predicts NO, and the actual value is YES, then the cost of misclassification is $500.
- If the model predicts NO, and the actual value is NO, then the cost of misclassification is $0.

Algorithms for classification use the cost matrix during scoring to propose the least expensive solution.

If you do not specify a cost matrix, then all misclassifications are counted as equally important.

If you are building an SVM model, then you specify costs using model weights instead of a cost matrix.

12.1.1.2 Performance Matrix

A Performance Matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. It is calculated by applying the model to a hold-out sample (the test set, created during the split step in a classification activity) taken from the build data. The values of the target are known. The known values are compared with the values predicted by the model. Performance Matrix does the following:

- Measures the likelihood of the model to predict incorrect and correct values
- Indicates the types of errors that the model is likely to make

The columns are predicted values and the rows are actual values. For example, if you are predicting a target with values 0 and 1, then the number in the upper right cell of the matrix indicates the false-positive predictions, that is, predictions of 1 when the actual value is 0.

12.1.1.3 Receiver Operating Characteristics (ROC)

Receiver Operating Characteristics (ROC) analysis is a useful method for evaluating Classification models. ROC applies to binary classification only. ROC is plotted as a curve. The area under the ROC curve measures the discriminating ability of a binary Classification model. The correct value for the ROC threshold depends on the problem that the model is trying to solve.

ROC curves are similar to lift charts in that they provide a means of comparison between individual models and determine thresholds that yield a high proportion of positive results. An ROC curve does the following:

- Provides a means to compare individual models and determine thresholds that yield a high proportion of positive results.
- Provides insight into the decision-making ability of the model. For example, you can determine how likely the model is to accurately predict the negative or the positive class.
- Compares predicted and actual target values in a Classification model.

See Also: "How to Use ROC"

12.1.1.3.1 How to Use ROC Receiver Operating Characteristics (ROC) supports what-if analysis. You can use ROC to experiment with modified model settings to observe the effect on the Performance Matrix. For example, assume that a business problem requires that the false-negative value be reduced as much as possible within the confines of the requirement that the number of positive predictions be less than or equal to some fixed number. You might offer an incentive to each customer predicted to be high-value, but you are constrained by a budget with a maximum of 170 incentives. On the other hand, the false negatives represent missed opportunities, so you want to avoid such mistakes.

To view the changes in the Performance Matrix:

1. Click **Edit Custom Operating Point** at the upper right corner. The **Specify Custom Threshold** dialog box opens.
2. In the **Specify Custom Threshold** dialog box, mention the desired settings and view the changes in the **Custom Accuracy** field.

As you change the Performance Matrix, you are changing the probability that result in a positive prediction. Typically, the probability assigned to each case is examined and if the probability is 0.5 or higher, then a positive prediction is made. Changing the cost matrix changes the positive prediction threshold to some value other than 0.5, and the changed value is displayed in the first column of the table beneath the graph.

12.1.1.4 Lift

Lift measures the degree to which the predictions of a Classification model are better than randomly-generated predictions. Lift applies to binary classification and non-binary classifications.

Lift measures how rapidly the model finds the actual positive target values. For example, lift enables you to figure how much of the customer database you must contact to get 50 percent of the customers likely to respond to an offer.

The x-axis of the graph is divided into quantiles. To view exact values, place the cursor over the graph. Below the graph, you can select the quantile of interest using **Selected Quantile.** The default quantile is quantile `1`.

To calculate lift, Oracle Data Mining does the following:

- Applies the model to test data to gather predicted and actual target values. This is the same data used to calculate the Performance Matrix.
- Sorts the predicted results by probability, that is, the confidence in a positive prediction.
- Divides the ranked list into equal parts, quantiles. The default is `100`.
- Counts the actual positive values in each quantile.

You can graph the lift as either Cumulative Lift or as Cumulative Positive Cases (default). To change the graph, select the appropriate value from the **Display** list. You can also select a target value in the **Target Value** list.

12.1.1.5 Profit and ROI

Profit uses user-supplied values for startup cost, incremental revenue, incremental cost, budget, and population to maximize the profit.

Oracle Data Miner calculates profit as follows:

```
Profit = -1 * Startup Cost + (Incremental Revenue * Targets Cumulative -
Incremental Cost * (Targets Cumulative + Non Targets Cumulative)) * Population /
Total Targets
```

Profit can be positive or negative, that is, it can be a loss.

To view the profit predicted by this model, select the **Target Value** that you are interested in. You can change the **Selected Population%**. The default is 1 percent.

Return on Investment (ROI) is the ratio of money gained or lost (whether realized or unrealized) on an investment relative to the amount of money invested. Oracle Data Mining uses this formula:

```
ROI = ((profit - cost) / cost) * 100

where profit = Incremental Revenue * Targets Cumulative, cost = Incremental Cost *
(Targets Cumulative + Non Targets Cumulative)
```

See Also:

- "Profit and ROI Example"
- "Profit and ROI Use Case"
- "Profit Setting Dialog"

12.1.1.5.1 Profit and ROI Example This example illustrates how profit and ROI are calculated.

To calculate profit:

1. Profit is calculated for a quantile. In this example, profit and ROI for quantile 20 is calculated.
2. Find the value of Targets Cumulative for Quantile 20 by looking at the lift chart data. Suppose that this value is `18`.
3. Suppose that the value of Non Targets Cumulative for Quantile 20 is `2`. Find this value by looking at the lift chart.
4. Calculate Total Targets which is Targets Cumulative at last Quantile plus Non Targets Cumulative at last Quantile. Suppose that this value is `100`.

5. These values are all user provided. You must provide values based on the business problem:
   - Startup cost = 1000
   - Incremental revenue = 10
   - Incremental cost = 5
   - Budget = 10000
   - Population = 2000
6. Calculate profit using this formula

   ```
   Profit = -1 * Startup Cost + (Incremental Revenue * Targets Cumulative -
   Incremental Cost * (Targets Cumulative + Non Targets Cumulative)) * Population
   / Total Targets
   ```

   Substituting the values in this example results in

   ```
   Profit = -1 * 1000 + (10 * 18 - 5 * (18 + 2) * 2000 / 100 = 600
   ```

To calculate ROI, use the formula

```
ROI = ((profit - cost) / cost) * 100

profit = Incremental Revenue * Targets Cumulative, cost = Incremental Cost *
(Targets Cumulative + Non Targets Cumulative)
```

Substituting the values in this example results in

```
ROI = ((180 - 100) / 100) * 100 = 80
```

12.1.1.5.2 Profit and ROI Use Case This use case depicts how to interpret results for profit and ROI calculations.

Suppose you run a mail order campaign. You will mail each customer a catalog. You want to mail catalogs to those customers who are likely to purchase things from the catalog.

Here is the input data from Profit and ROI Example

- Startup cost = 1000. This is the total cost to start the campaign.
- Incremental revenue = 10. This is estimated revenue that results from a sale or new customer.
- Budget = 10000. This is the total amount of money that you can spend.
- Population = 2000. This is the total number of cases.

Therefore, each quantile contains 20 cases:

```
total population /number of quantiles = 2000/100 = 20
```

The cost to promote a sale in each quantile is (Incremental Cost * number of cases per quantile) = $5 * 20 = $100).

The cumulative costs per quantile are as follows:

- Quantile 1 costs $1000 (startup cost) + $100 (cost to promote a sale in Quantile 1) = $1100.
- Quantile 2 costs $1100 (cost of Quantile 1) + $100 (cost in Quantile 2).

- Quantile 3 costs $1200.

If you calculate all of the intermediate values, then the cumulative costs for Quantile 90 is $10,000 and for Quantile 100 is $11,000. The budget is $10,000. If you look at the graph for profit in Oracle Data Miner, then you should see the budget line drawn in the profit chart on the 90th quantile.

In the Profit and ROI Example, the calculated profit is $600 and ROI is 80 percent, which means that if you mail catalogs to first 20 quantiles of the population (400), then the campaign will generate a profit of $600 (which has ROI of 80 percent).

If you randomly mail the catalogs to first 20 quantiles of customers, then the profit is

```
Profit = -1 * Startup Cost
         + (Incremental Revenue * Targets Cumulative - Incremental Cost
           * (Targets Cumulative + Non Targets Cumulative))
           * Population / Total Targets
Profit = -1 * 1000 + (10 * 10 - 5 * (10 + 10)) * 2000 / 100 = -$1000
```

In other words, there is no profit.

12.1.2 Classification Model Test and Results Viewers

This section about Classification models includes:

- Classification Model Test Viewer
- Compare Classification Test Results

12.1.2.1 Classification Model Test Viewer

Open the test viewer by selecting either **View Test Results** or **Compare Test Results** in the context menu for a Classification node or a Test node that tests Classification models. Select the results to view.

The Classification model test viewer shows the following tabs:

- Performance
- Performance Matrix
- ROC
- Lift
- Profit

See Also: "Test Metrics for Classification Models"

12.1.2.1.1 Performance

The **Performance** tab provides an overall summary of the performance of each model generated. It displays test results for several common test metrics. It displays the following:

- All Measures (default). The **Measure** list enables you to select the measures to display. By default, all measures are displayed. The selected measures are displayed as graphs. If you are comparing test results for two or more models, then different models have graphs in different colors.
- Predictive Confidence
- Average Accuracy
- Overall Accuracy
- Cost, if you specified costs or the system calculated costs

In the **Sort By** fields. you can specify the sort attribute and sort order. The first list is the sort attribute: measure, creation date, or name (the default). The second list is the sort order: ascending or descending (default).

Below the graphs, the Models table supplements the information presented in the graph. You can minimize the table using the splitter line. The **Models** table summarizes the data in the histograms:

- Name, the name of the model along with color of the model in the graphs
- Predictive Confidence percent
- Overall Accuracy percent
- Average Accuracy percent
- Cost, if you specified cost (costs are calculated by Oracle Data Miner for decision trees)
- Algorithm (used to build the model)
- Creation date

By default, results for the selected model are displayed. To change the list of models, click and deselect any models for which you do not want to see results. If you deselect a model, then both the histogram and the summary information are removed.

> **See Also:** "Test Metrics for Classification Models"

12.1.2.1.2 Performance Matrix The Performance Matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. You can either view the detail for a selected model, or you can compare performance matrices for all models.

- Click **Show Detail** to view test results for one model.
- Click **Compare Models** to compare test results.

12.1.2.1.3 Show Detail First, select a model. If you are viewing test results for one model, the details for that model are displayed automatically.

- In the top pane, Average Accuracy and Overall Accuracy are displayed with a grid that displays the correct predictions for each target value. Cost information is displayed if you have specified costs.
- In the bottom pane, a Performance Matrix with rows showing actual values and columns showing predicted values is displayed for the selected model. The percentage correct and cost are displayed for each column.

Select **Show totals and cost** to see the total, the percentage correct, and cost for correct and incorrect predictions.

12.1.2.1.4 Compare Models This option compares performance information for all models in the node that were tested.

- The top pane lists the following for each model:
  - Percentage of correct predictions
  - Count of correct predictions
  - Total case count
  - Cost information

To see more detail, select a model.

- The bottom pane displays the target value details for the model selected in the top pane. Select the measure.
  - **Correct Predictions** (default): Displays correct predictions for each value of the target attribute
  - **Costs:** Displays costs for each value of the target

12.1.2.1.5 ROC
Receiver Operating Characteristics (ROC) compares predicted and actual target values in a binary Classification model.

To edit and view an ROC:

1. Select the **Target** value. The ROC curves for that value are displayed.
2. Click **Edit Custom Operating Point** to change the operating point. The ROC graph displays a line showing ROC for each model. Points are marked on the graph indicating the values shown in the key at the bottom of the graph. Below the graph, the ROC Summary results table supplements the information presented in the graph. You can minimize the table using the splitter line.
   - The **Models** grid, in the lower pane, contains the following summary information:
     - Name
     - Area under the curve
     - Maximum Overall Accuracy Percentage
     - Maximum Average Accuracy Percentage
     - Custom Accuracy Percentage
     - Model Accuracy Percentage
     - Algorithm
     - Creation Date and Time
   - Select a model and click to see the **ROC Detail Dialog** box, which displays the statistics for probability thresholds.
   - To change the list of models, click to open the **Edit Test Result Selection** dialog box. By default, results for all models in the node are displayed.

   See Also: "How to Use ROC"

12.1.2.1.6 Edit Test Result Selection
Deselect the check box for those models for which you do *not* want to see results. If you deselect a model, then both the ROC curve and the details for that model are not displayed.

Click **OK** when you have finished.

12.1.2.1.7 ROC Detail Dialog
The ROC Detail Dialog displays statistics for probability thresholds. For each probability threshold, the following are displayed:

- True Positive
- False Negative
- False Positive
- True Negative

- True Positive Fraction
- False Positive Fraction
- Overall Accuracy
- Average Accuracy

Click **OK** to dismiss the dialog box.

12.1.2.1.8 Lift

The lift graph displays at least three lines:

- A line showing the lift for each model
- A red line for the random model
- A vertical blue line for threshold

The x-axis of the graph is divided into quantiles. The graph shows the lift from the model (or models) and also shows the lift from a naive model (Random) and the ideal lift.

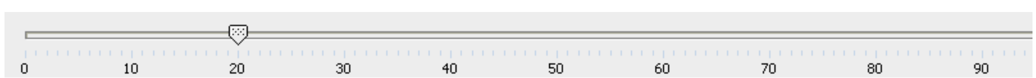
The Lift viewer compares lift results for a given target value in two or more models. It displays either the Cumulative Positive Cases or the Cumulative Lift.

If you are comparing the lift for two or more models, then the lines for different models are in different colors. The table below the graph shows the name of the model and the color used to display results for that model.

The viewer has the following controls:

- **Display:** Selects the display option, either **Cumulative Positive Cases** (default) or **Cumulative Lift.**
- **Target Value:** Selects the target value for comparison. The default target value is the least frequently occurring target value.

The threshold is a blue vertical line used to select a quantile. As the threshold moves, the details for each test result in the Lift Detail table changes to the point on the Lift Chart that corresponds to the selected quantile. You move the threshold by dragging the indicator on the quantile line. Here is the quantile set to 20:



The Quantile line depicting the indicator that the quantile is set to 20.

****

Below the graph, a data table supplements the information presented in the graph. You can minimize the table using the splitter line.

The table has the following columns:

- Name, the name of the model along with color of the model in the graph
- Lift Cumulative
- Gain Cumulative Percentage
- Percentage Records Cumulative
- Target Density Cumulative
- Algorithm

- Creation Date (date and time)

Above the Models grid is the Lift Detail Dialog icon. Select a model and click the icon to open the **Lift Detail** dialog box, which displays lift details for 100 quantiles.

To change the list of models, click and deselect any models for which you do not want to see results. If you deselect a model, then both the lift curve and the detail information for that model are not displayed. By default, results for all models in the node are displayed.

See Also: "Test Metrics for Classification Models"

12.1.2.1.9 Lift Detail The Lift Detail Dialog displays statistics for each quantile from 1 to 100. Click **OK** to dismiss the dialog box.

Threshold probability does not always reflect standard probability. For example, the Classification node enables you to specify three different performance settings:

- **Balanced**: Apply balance weighting to all target class values.
- **Natural**: Do not apply any weighting.
- **Custom**: Apply user- created custom weights file.

The default for Classification models is **Balanced**. Balanced is implemented by passing weights or costs into the model, depending on the algorithm used.

The threshold probability actually reflects cost rather than standard probability.

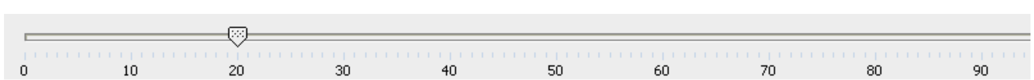
To see the difference between **Balanced** and **Natural:**

1. Create a Classification model.
2. Select the performance setting options and view the lift details:
   - **Natural:** The threshold probability values are the greatest probabilities for each quantile.
   - **Balanced:** The threshold reflects cost. You see the lowest cost value for each quantile.

12.1.2.1.10 Profit The Profit graph displays at least three lines:

- A line showing the profit for each model
- A line indicating the budget
- A line indicating the threshold

The threshold is a blue vertical line used to select a quantile. As the threshold moves, the details for each test result in the Lift Detail table changes to the point on the Lift Chart that corresponds to the selected quantile. You can move the threshold by dragging the indicator on the quantile line. Here is the quantile set to 20:



The Quantile line depicting the indicator that the quantile is set to 20.

****

To specify the values for profit, click **Profit Settings** to open the Profit Setting Dialog.

If you are comparing the profit for two or more models, then the lines for different models are different colors. The table below the graph shows the name of the model and the color used to display results for that model.

The bottom pane contains the Models grid and supplements the information presented in the graph. You can minimize the table using the splitter line.

The table has the following columns:

- Name, the name of the model along with color of the model in the graphs
- Profit
- ROI Percentage
- Records Cumulative Percentage
- Target Density Cumulative
- Maximum Profit
- Maximum Profit Population Percentage
- Algorithm
- Creation Date (and time)

Above the Models grid is the Browse Detail icon. Select a model and click  to see the **Profit Detail Dialog** box which displays statistics for each quantile from 1 to 100.

To change the list of models, click  and deselect any models for which you do not want to see results. If you deselect a model, then both the profit curve and the detail information for that model are not displayed. By default, results for all models in the node are displayed.

See Also:

- "Profit and ROI"
- "Profit and ROI Example"
- "Profit and ROI Use Case"
- "Test Metrics for Classification Models"

12.1.2.1.11 Profit Detail Dialog Displays statistics about profit for quantiles 1 to 100. Click **OK** to dismiss the dialog box.

12.1.2.1.12 Profit Setting Dialog Click **Profit Settings** to change the following values:

- **Startup Cost**: The cost of starting the process that creates the profit. The default is `1`.
- **Incremental Revenue**: Incremental revenue earned for each correct prediction. The default is `1`.
- **Incremental Cost**: The cost of each additional item. The default is `1`.
- **Budget**: A total cost that cannot be exceeded. The default value is `1`.
- **Population**: The number of individual cases that the model is applied to. The default is `100`.

Click **OK** to save the settings.

12.1.2.2 Compare Classification Test Results

To compare test results for all of the models in a Classification Build node:

- If you tested the models when you ran the Classification node: Right-click the Classification node that contains the models and select **Compare Test Results.**
- If you tested the Classification models in a Test node: Right-click the Test node that tests the models and select **Compare Test Results.**

The **Classification Model Test Viewer** that compares the test results opens. The comparison enables you to select the model that best solves a business problem.

The graphs on the **Performance** tab for different models are in different colors. On the other tabs, the same color is used for the line indicating measures such as lift.

The color associated with each model is displayed in the bottom page of each tab.

12.1.2.2.1 Compare Test Results Compare Test Results for Classification displays these subtabs:

- **Performance**: Compares performance results in the top pane for the models listed in the bottom panel.

  To edit the list of models, click above pane that lists the models. This opens the **Edit Test Selection (Classification and Regression)** dialog box. By default, test results for all models are compared.
- **Performance Matrix**: Displays the Performance Matrix for each model. You can display either Compare models (a comparison of the performance matrices) or Details (the Performance Matrix for a selected model).
- **ROC**: Compares the ROC curves for the models listed in the lower pane.

  To see information for a curve, select a model and click.

  To edit the list of models, click above pane that lists the models. This opens the **Edit Test Selection (Classification and Regression)** dialog box.
- **Lift**: Compares the lift for the models listed in the lower pane. For more information about lift, see Lift.

  To edit the list of models, click above pane that lists the models. This opens the **Edit Test Selection (Classification and Regression)** dialog box.
- **Profit**: Compares the profit curves for the models listed in the lower pane.

  To edit the list of models, click above pane that lists the models. This opens the **Edit Test Selection (Classification and Regression)** dialog box.

  See Also:

  - "Profit and ROI"
  - "Performance"
  - "Performance Matrix"
  - "Receiver Operating Characteristics (ROC)"

12.1.2.2.2 Edit Test Selection (Classification and Regression) By default, test results for all successfully built models in the build node are selected. If you do not want to view test results for a model, then deselect the model.

Click **OK** when you have finished.

12.2 Tuning Classification Models

When you tune a model, you select one of several ways to create a derived cost matrix. The derived cost matrix is used for any subsequent test and apply operations. Each tuning dialog box has a different goal in how the tuning is performed.

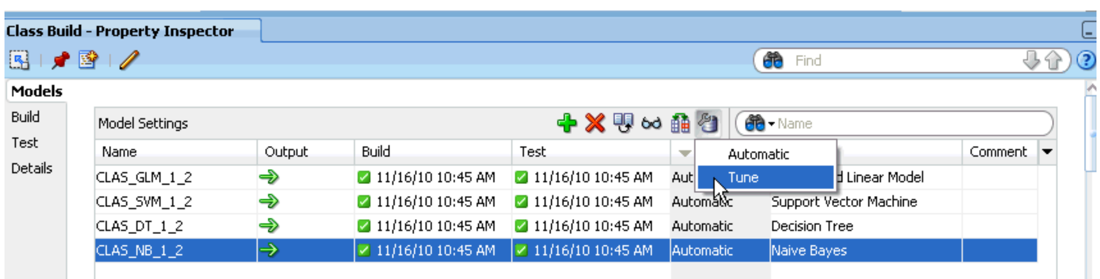
Note: To tune models, you must test the models in the same node that you build them.

If necessary, you can remove tuning and then re-run the node.

To tune a model:

1. Open the **Properties** pane for the Build node. Right-click the node and select **Go to Properties.**
2. Go to the Test section. Select **Generate Select Test Results for Model Tuning**. The model test operation then generates the unbiased test results (no cost matrix used) for the corresponding test results. The unbiased test results are determined by the **Tune Settings** dialog box to initialize tuning options. For example, if only ROC is selected for Test Results, then the test operation generates the regular ROC result and the unbiased ROC result.
3. Run the Build node. You must test the models in the Build node. This is the default behavior of the Classification Build node.
4. In **Properties** pane for the Build node, go to the **Models** section. Select the models that you want to tune and click the tune icon  in the menu bar.

   Select **Tune** from the drop-down list.



The Properties pane of a Build node, that shows the Tune option selected for a particular model. Properties was known as Property Inspector.

****

5. The **Tune Settings** dialog box opens with all the available test results.

   You can tune a model using one technique. For example, you can tune using costs or using lift, but not using both costs and lift at the same time.
6. If you are tuning more than one model, then select a model from the **Models** list in the bottom pane of the dialog box. After you tune the first model, return to this pane and select another model.
7. Click the tab for the test result to tune. The tabs are:
   - Cost
   - Benefit
   - ROC

- Lift
- Profit

8. When you finish tuning a model, click **Tune** in the pane on the right to generate tuning. In the **Models** list in the bottom pane, the Tune setting changes from Automatic to the new setting.
9. Tune as many models in the node you want. Go to other tuning tabs and perform tuning from there. When you have finished tuning, click **OK**.
10. All models that have tuning specifications changed during the session have their test results marked as not run. When you run the node again:
    - The new cost matrix is generated and inserted into the model.
    - A new test result is generated showing full test result information for the behavior of the current model.
11. Run the tuned model. After running of the model is complete, the Models section of **Properties** indicates how each model was tuned. For example, if you tune a model by changing costs, then the Tune entry for that model is Tune - Cost.
12. Right-click the Build node and select **View Test Results** for the tuned model to see the effects of the tuning.

You may have to repeat the tuning steps several times to get the desired results. If necessary, you can remove tuning for a model.

See Also: "Remove Tuning"

12.2.1 Remove Tuning

To remove tuning for a model:

1. Right-click the node and select **Go to Properties.**
2. Go to the **Models** section and click .
3. Select **Automatic.**
4. Run the node.

12.2.2 Cost

The **Cost** tab of **Tune Settings** enables you to specify costs for target for scoring purposes.

By default the cost matrix is initially generated based on all the known target values in the Build Data Source. The cost matrix is set to cost values of 1 to start with.

To specify costs:

1. Open the **Properties** pane for the Build node. Right-click the node and select **Go to Properties.**
2. In the **Test** section, select **Generate Select Test Results for Model Tuning** and run the node.
3. In the **Models** section, select the models that you want to tune and click .
4. Select **Tune** from the drop-down list. The **Tune Settings** dialog box opens.
5. In the **Tune Settings** dialog box, go to the **Cost** tab.

6. If you are tuning more than one model, then select a model from the **Models** list in the bottom pane. After you tune the first model, return to this pane and select another model.
7. Select the target value for which to specify costs.
8. Select the appropriate option:
   - **False Positive: Incorrectly identifying a case as a target.** (Default)
   - **False Negative: Incorrectly identifying a case as a non-target.**
9. In the **Weight** field, specify a weight for the cost.
10. Click **Apply** to add the cost that you just specified to the cost matrix.
11. Define costs for all target values that you are concerned about.
12. To apply the matrix, click **Tune** in the upper right pane.
13. Click **Derived Matrix** to view the cost matrix that you created. Examine the derived cost matrix. You can continue tuning by changing any selections that you made.
14. When you have finished, click **OK** to accept the tuning. Click **Cancel** to cancel the tuning

To cancel the tuning, click **Reset**. Tuning returns to Automatic.

To see the impact of the tuning, rerun the model node.

See Also:

- "Costs and Benefits"
- "Tuning Classification Models"

12.2.2.1 Costs and Benefits

In a classification problem, it is often important to specify the cost or benefit associated with correct or incorrect classifications. By doing so, it can be valuable when the cost of different misclassification varies significantly.

You can create a cost matrix to bias the model to minimize the cost or maximize the benefit. The cost/benefit matrix is taken into consideration when the model is scored.

See Also:

- "Costs"
- "Benefits"

12.2.2.1.1 Costs For example, suppose the problem is to predict whether a customer is likely to respond to a promotional mailing. The target has two categories: YES (the customer responds) and NO (the customer does not respond). Suppose a positive response to the promotion generates $500 and that it costs $5 to do the mailing. After building the model, you compare the model predictions with actual data held aside for testing. At this point, you can evaluate the relative cost of different misclassifications:

- If the model predicts YES and the actual value is YES, then the cost of misclassification is $0.
- If the model predicts YES and the actual value is NO, then the cost of misclassification is $5.

- If the model predicts NO and the actual value is YES, then the cost of misclassification is $495.
- If the model predicts NO and the actual value is NO, then the cost is $0.

12.2.2.1.2 Benefits Using the same costs, you can approach the relative value of the outcomes from a benefits perspective. When you correctly predict a YES (a responder), the benefit is $495. When you correctly predict a NO (a non-responder), the benefit is $5.00 because you can avoid sending out the mailing. Because the goal is to find the lowest cost solution, benefits are represented as negative numbers.

12.2.3 Benefit

The **Benefit** tab enables you to:

- Specify a benefit for each value of the target. The values specified are applied to the cost benefit matrix. Specifying benefits is useful when there are many target values.
- Indicate the most important values.

To tune a model using the **Benefit** tab:

1. Open the **Properties** pane for the Build node. Right-click the node and select **Go to Properties.**
2. In the **Test** section, select **Generate Select Test Results for Model Tuning** and run the node.
3. In the **Models** section, select the models that you want to tune and click .
4. Select **Tune** from the drop-down list. The **Tune Settings** dialog box opens in a new tab.
5. In the **Tune Settings** dialog box, click **Benefit.**
6. If you are tuning more than one model, then select a model from the **Models** list in the bottom pane. After you tune the first model, return to this pane and select another model.
7. Select the target value for tuning from the **Target Value** list.
8. Specify benefit values for the target value selected. Benefit values can be positive or negative. If there is more benefit from a target value, then the benefit value should be higher than other benefit values. The default benefit value for each target value is `0`.

   Enter the benefit value for the selected target in the **Benefit** box and click **Apply** to update the Cost Benefits matrix.
9. When you have finished specifying benefit values, click **Tune** in the right-hand column.
10. Click **View** to see the derived cost matrix.
11. When you have finished, click **OK** to accept the tuning, or click **Cancel** to cancel the tuning.

See Also:

- "Costs and Benefits"
- "Tuning Classification Models"

12.2.4 ROC

ROC is only supported for binary models. The **ROC Tuning** tab adds a side panel to the standard ROC Test Viewer. The following information is displayed:

- Performance Matrix in the upper right pane, enables you to display these matrices:
  - Overall Accuracy: Cost matrix for the maximum Overall Accuracy point on the ROC chart.
  - Average Accuracy: Cost matrix for the maximum Average Accuracy point.
  - Custom Accuracy: Cost matrix for the custom operating point.

    You must specify a custom operating point for this option to be available.
  - Model Accuracy: The current Performance Matrix (approximately) of the current model.

    You can use the following calculation to derive Model Accuracy from the ROC result provided:

    If there is no embedded cost matrix, then find the 50 percent threshold point or the closest one to it. If there is an embedded cost matrix, then find the lowest cost point. For a model to have an embedded cost matrix, it must have either been tuned or it has a cost matrix or cost benefit defined by the default settings of the Build node.
- The **Performance Matrix** Grid shows the Performance Matrix for the option selected.
- Click **Tune** to:
  - Select the current performance option as the one to use to tune the model.
  - Derive a cost matrix from the ROC result at that probability threshold.

  Tune Settings, in the lower part of this panel, is updated to display the new matrix.
- Click **Clear** to clear any tuning specifications and set tuning to Automatic. In other words, no tuning is performed.

See Also:

- "Receiver Operating Characteristics"
- "Select Custom Operating Point"

12.2.4.1 ROC Tuning Steps

To perform ROC tuning:

1. Open the **Properties** pane for the Build node. Right-click the node and select **Go to Properties.**
2. In the **Test** section, select **Generate Select Test Results for Model Tuning** and run the node. See "Tuning Classification Models" for more information.
3. In the **Models** section, select the models that you want to tune and click .
4. Select **Tune** from the drop-down list. The **Tune Settings** dialog box opens in a new tab.
5. In the **Tune Settings** dialog box, go to the **ROC** tab.

6. If you are tuning more than one model, select a model from the **Models** list in the bottom pane. After you tune the first model, return to this pane and select another model.
7. Select a target value. In the case of ROC, there are only two values.
8. Select a custom operating point if you do no want to use the default point. See "Select Custom Operating Point" for more information.
9. Select the kind of Performance Matrix to use:
   - **Overall Accuracy** (default)
   - **Average Accuracy**
   - **Custom Accuracy.** Fill in the values for the Performance Matrix if you select this option.
   - **Model Accuracy**
10. Click **Tune.** New Tune settings are displayed in the same panel as the Performance Matrix. Examine the Derived Cost Matrix. You can continue tuning by changing any selections that you made.
11. When you have finished, click **OK** to accept the tuning, or click **Cancel** to cancel the tuning.
    - To reset the tuning, click **Reset.**
    - To see the impact of the tuning, run the Model node.

12.2.4.1.1 Select Custom Operating Point The **Specify Custom Threshold** dialog box enables you to edit the custom operating point for all the models in the node.

- To change the **Hit Rate** or **False Alarm,** click the appropriate option and adjust the value that you want to use.
- Alternatively, you can specify the **False Positive** or **False Negative** ratio. To do this, click the appropriate option and specify the ratio.

Click **OK** when you have finished.

12.2.4.2 Receiver Operating Characteristics

Receiver Operating Characteristics (ROC) is a method for experimenting with changes in the probability threshold and observing the resultant effect on the predictive power of the model.

- The horizontal axis of an ROC graph measures the False Positive Rate as a percentage.
- The vertical axis shows the True Positive Rate.
- The top left corner is the optimal location in an ROC curve, indicating a high TP (True Positive) rate versus low FP (False Positive) rate.
- The area under the ROC curve measures the discriminating ability of a binary Classification model. This measure is especially useful for data sets with an unbalanced target distribution (one target class dominates the other). The larger the area under the curve, the higher the likelihood that an actual positive case is assigned a higher probability of being positive than an actual negative case.

ROC curves are similar to lift charts in that they provide a means of comparison between individual models, and then determine thresholds that yield a high proportion of positive hits. ROC was originally used in signal detection theory to gauge the true hit versus false alarm ratio when sending signals over a noisy channel.

12.2.5 Lift

Lift measures the degree to which the predictions of a Classification model are better than randomly generated predictions.

To tune a model using Lift:

1. Open the **Properties** pane for the Build node. Right-click the node and select **Go to Properties**.
2. In the **Test** section, select **Generate Select Test Results for Model Tuning** and run the node.
3. In the **Models** section, select the models that you want to tune and click .
4. Select **Tune** from the drop-down list. The **Tune Settings** dialog box opens in a new tab.
5. In **Tune Settings** dialog box, go to the **Lift** tab.
6. If you are tuning more than one model, then select a model from the Models list in the bottom pane. After you tune the first model, return to this pane and select another model.
7. Select the target value for tuning from the **Target Value** list.
8. Decide whether to tune using the Cumulative Positive Cases chart, the default or the Cumulative Lift Chart. Select the chart from the **Display** list.

   Either chart displays several curves: the lift curve for the model that you are tuning, ideal lift, and random lift, which is the lift from a model where predictions are random.

   The chart also displays a blue vertical line that indicates the threshold, the quantile of interest.
9. Selected a quantile using the slider in the quantile display below the lift chart. As you move the slider, the blue vertical bar moves to that quantile, and the tuning panel is updated with the Performance Matrix for that point.
10. Click **Tune**, below the Performance Matrix. New tune settings are displayed in the same panel as the Performance Matrix. Examine the Derived Cost Matrix. You can continue tuning by changing any selections that you made.
11. When you have finished, click **OK** to accept the tuning, or click **Cancel** to cancel the tuning.

    - To reset the tuning, click **Reset.**
    - To see the impact of the tuning, run the Model node.

    See Also:

    - "Tuning Classification Models"
    - "About Lift"

12.2.5.1 About Lift

Lift is the ratio of positive responders in a segment to the positive responders in the population as a whole. For example, if a population has a predicted response rate of 20 percent, but one segment of the population has a predicted response rate of 60 percent, then the lift of that segment is 3 (60 percent/20 percent). Lift measures the following:

- The concentration of positive predictions within segments of the population and specifies the improvement over the rate of positive predictions in the population as a whole.
- The performance of targeting models in marketing applications. The purpose of a targeting model is to identify segments of the population with potentially high concentrations of positive responders to a marketing campaign.

The notion of lift implies a binary target: either a Responder or a Non-responder, which means either YES or NO. Lift can be computed for multiclass targets by designating a preferred positive class and combining all other target class values, effectively turning a multiclass target into a binary target. Lift can be applied to both binary and non-binary classifications as well.

The calculation of lift begins by applying the model to test data in which the target values are already known. Then, the predicted results are sorted in order of probability, from highest to lowest Predictive Confidence. The ranked list is divided into quantiles (equal parts). The default number of quantiles is `100`.

12.2.6 Profit

The **Profit** tab provides a method for maximizing profit.

To tune a model:

1. Open **Properties** for the Build node. Right-click the node and select **Go to Properties.**
2. In the **Test** section, select **Generate Select Test Results for Model Tuning** and run the node.
3. In the **Models** section, select the models that you want to tune and click .
4. Select **Tune** from the drop-down list. The **Tune Settings** dialog box opens in a new tab.
5. In the **Tune Settings** dialog box, go to the **Profit** tab.
6. If you are tuning more than one model, select a model from the **Models** list in the bottom pane. After you tune the first model, return to this pane and select another model.
7. Select the target value for tuning from the **Target Value** list.
8. Click **Profit Settings** and specify the values in the **Profit Settings** dialog box.
9. After you specify Profit Settings, the graph reflects the values that you specified.
10. Use the slider below the chart to adjust the Threshold (blue vertical line).
11. Click **Tune**, below the Performance Matrix. New tune settings are displayed in the same panel as the Performance Matrix. Examine the Derived Cost Matrix. You can continue tuning by changing any selections that you made.
12. When you have finished, click **OK** to accept the tuning, or click **Cancel** to cancel the tuning.
    - To reset tuning, click **Reset.**
    - To see the impact of the tuning, run the Model node.

See Also:

- "Profit"
- "Tuning Classification Models"
- "Profit Setting"

12.2.6.1 Profit Setting

The default values for Startup Cost, Incremental Revenue, Incremental Cost, and Budget are all `1.` The default value for Population is `100.` Change these values to ones appropriate for your business problem.

Click **OK**.

12.2.6.2 Profit

Profit provides a method for maximizing profit. You specify the following information. Oracle Data Miner uses the following information to create a cost matrix that optimizes profit:

- Startup cost
- Incremental revenue
- Incremental cost
- Budget
- Population

12.3 Testing Regression Models

Regression models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a regression project is typically divided into two data sets:

- One for building the model
- One for testing the model

The test data must be compatible with the data used to build the model and must be prepared in the same way that the build data was prepared.

The ways to test Classification and Regression models:

- By splitting the input data into build data and test data. This is the default. The test data is created by randomly splitting the build data into two subsets. 40 percent of the input data is used for test data.
- By using all the build data as the test data.
- By attaching two Data Source nodes to the build node.
  - The first data source, that you connect to the build node, is the source of build data.
  - The second node that you connect is the source of test data.
- By deselecting **Perform Test** in the **Test** section of the **Properties** pane and then using a Test Node. By default, all Classification and Regression models are tested.

Test settings specify which metrics to calculate and control the calculation of the metrics.

Oracle Data Mining provides several kinds of information to assess Regression models:

- Residual Plot
- Regression Statistics
- Regression Model Test Viewer
- Compare Regression Test Results

To view test results, first test the model or models in the node:

- If you tested the models using the default test in the Regression node, then run the node and then right-click the node. Select **View Test Results** and select the model that you are interest in. The **Regression Model Test Viewer** starts. To compare the test results for all models in the node, select **Compare Test Results**.
- If you tested the models using a Test node, then run the Test node and then right-click the node. Select **View Test Results** and select the model that you are interested in. The **Regression Model Test Viewer** starts. To compare the test results for all models in the node, select **Compare Test Results**.

You can also compare test results by going to the **Models** section of the **Properties** pane of the Build node where you tested the models and click .

12.3.1 Residual Plot

The residual plot is a scatter plot of the residuals. Each residual is the difference between the actual value and the value predicted by the model. Residuals can be positive or negative. If residuals are small (close to 0), then the predictions are accurate. A residual plot may indicate that predictions are better for some classes of values than others.

12.3.2 Regression Statistics

Oracle Data Mining calculates the following statistics to help assess the overall quality of Regressions models:

- **Root Mean Squared Error:** The square root of the average squared distance of a data point from the fitted line.
- **Mean Absolute Error:** The average of the absolute value of the residuals (error). The Mean Absolute Error is very similar to the Root Mean Square Error but is less sensitive to large errors.

12.3.3 Regression Model Test Viewers

The help for Regression models includes:

- Regression Model Test Viewer
- Compare Regression Test Results

12.3.3.1 Regression Model Test Viewer

To view information in the Regression Model Test Viewer:

1. Right-click a Regression node or a Test node (that tests Regression Models) and select **View Test Results** or **Compare Test Results.**
2. The **Regression Model Test Viewer** opens, and displays the following tabs:

- [Performance (Regression)]
- [Residual]

3. Click **OK.**

> **See Also:** "Regression Statistics" for an overview of the test metrics that Oracle Data Miner calculates.

12.3.3.1.1 Performance (Regression) The **Performance** tab displays the test results for several common test metrics.

The **Performance** tab for Regression model show these measures for all models:

- **All Measures** (default). The **Measure** list enables you to select the measures to display. By default, All Measures are displayed. The selected measures are displayed as graphs. If you are comparing test results for two or more models, then the different models have graphs in different colors.
- **Predictive Confidence:** Measures how much better the predictions of the mode are than those of the naive model. Predictive Confidence for regression is the same measure as Predictive Confidence for classification.
- **Mean Absolute Error**
- **Root Mean Square Error**
- **Mean Predicted Value:** The average of the predicted values.
- **Mean Actual Value:** The average of the actual values.

Two **Sort By** lists specify sort attribute and sort order. The first **Sort By** list contains Measure, Creation Date, or Name (the default). The second **Sort By** list contains the sort order: ascending or descending (default).

The top pane displays these measures as histograms.

The bottom pane contains the Models grid that supplements the information presented in the graphs. You can minimize the table using the splitter line.

The Models grid has the following columns:

- Name, the name of the model along with color of the model in the graphs.
- Predictive Confidence
- Mean Absolute Error
- Root Mean Square Error
- Mean Predicted Value
- Mean Actual Value
- Algorithm
- Creation Date (and time)

By default, results for the selected model are displayed. To change the list of models, click  and deselect any models for which you do not want to see results. If you deselect a model, both the histograms and the detail information for that model are not displayed.

> **See:** "Regression Statistics"for more information about Mean Absolute Error and Root Mean Square Error.

12.3.3.1.2 Residual The **Residual Plot** tab show the Residual Plot on a per-model basis. By default, the residual plots are displayed as graph.

- To see numeric results, click .
- To change the display back to a graph, click .
- To see the plot for another model, select the model from the **Show** list and click **Query**.

You can control how the plot is displayed in several ways:

- Select the information displayed on the y-axis and on the x-axis. The default depictions are:
  - X axis: Predicted Value
  - Y axis: Residual

  To change this, select information from the lists.
- The default sample size is `2000`. You can make this value larger or smaller.
- You can compare plots side by side. The default is to not compare plots.

If you change any of these fields, then click **Query** to see the results.

To compare plots side by side, select the model to compare with the current model from the **Compare** list and click **Query**. The residual plots are displayed side by side.

The bottom pane show the **Residual result summary table**. The table contains the Models grid which supplements the information presented in the plots. You can minimize the table using the splitter line.

The table has the following columns:

- Model, the name of the model along with color of the model in the graphs
- Predictive Confidence
- Mean Absolute Error
- Root Mean Square Error
- Mean Predicted Value
- Mean Actual Value
- Algorithm
- Creation Date (and time)

By default, results for all models in the node are displayed. To change the list of models, click to open the **Edit Test Selection (Classification and Regression)** dialog box.

12.3.3.2 Compare Regression Test Results

To compare test results for all the models in a Regression Build node:

- If you tested the models when you ran the Regression node:
  - Right-click the Regression node that contains the models.
  - Select **Compare Test Results.**
- If you tested the Regression models in a Test node:
  - Right-click the Test node that tests the models.

– Select **Compare Test Results.**

12.3.3.2.1 Compare Test Results When you compare test results for two or more Regression models, each model has a color associated with it. This color indicates the results for that model. For example, if model M1 has purple associated with it, then the bar graphs on the **Performance** tab for M1 is displayed in purple.

By default, test results for all models in the node are compared. If you do not want to compare all test results, then click . The **Edit Test Results Selection** dialog box opens. Deselect results that you do not want to see. Click **OK** when you have finished.

Compare Test Results opens in a new tab. Results are displayed in two tabs:

- **Performance** tab: The following metrics are compared on the **Performance** tab:
  - Predictive Confidence for Classification Models
  - Mean Absolute Error
  - Mean Predicted Value

  By default, test results for all models are compared. To edit the list of models, click above pane that lists the models to open the **Edit Test Selection (Classification and Regression** dialog box.

- **Residual** tab: Displays the residual plot for each model.
  - You can compare two plots side by side. By default, test results for all models are compared.
  - To edit the list of models, click above pane that lists the models to open the **Edit Test Selection (Classification and Regression** dialog box.

    See Also:

    - "Regression Statistics"
    - "Predictive Confidence" for more information about Mean Absolute Error.
    - "Edit Test Selection (Classification and Regression)"

13 Data Mining Algorithms

The following algorithms are supported by Oracle Data Miner:

- Anomaly Detection
- Association
- Decision Tree
- Expectation Maximization, requires Oracle Database 12*c*
- Generalized Linear Models
- k-Means
- Naive Bayes
- Nonnegative Matrix Factorization
- Orthogonal Partitioning Clustering
- Singular Value Decomposition and Principal Components Analysis, requires Oracle Database 12*c*
- Support Vector Machine

Settings Information contains the setting information common to most model viewers.

Oracle Data Mining Concepts provides overview information about algorithms, data preparation, and scoring. See the manual for the database version that you connect to, as described in Oracle Data Miner Documentation.

13.1 Anomaly Detection

Anomaly Detection (AD) identifies cases that are unusual within data that is apparently homogeneous. Anomaly detection is an important tool for fraud detection, network intrusion, and other rare events that may have great significance but are hard to find.

Oracle Data Mining uses Support Vector Machine (SVM) as the one-class classifier for Anomaly Detection (AD). When SVM is used for anomaly detection, it has the classification mining function but no target.

See Also: "Support Vector Machine Algorithms"

There are two ways to search for anomalies:

- By building and applying an Anomaly Detection model. To build an AD model, use an Anomaly Detection Node connected to an appropriate data source.

- By using an Anomaly Detection Query, one of the Predictive Query nodes.

  See Also: "Applying Anomaly Detection Models"for information about how to use AD models to make predictions

13.1.1 Applying Anomaly Detection Models

One-class SVM models, when applied, produce a prediction and a probability for each case in the scoring data.

- If the prediction is 1, then the case is considered Typical.
- If the prediction is 0, then the case is considered Anomalous.

This behavior reflects the fact that the model is trained with normal data.

13.1.2 Anomaly Detection Viewers and Algorithm Settings

This section discusses the Anomaly Detection Model viewer, the procedure to view the AD Model Viewer and the algorithm settings related to Anomaly Detection. The topics discussed are:

- Anomaly Detection Model Viewer
- Algorithm Settings for AD

13.1.2.1 Anomaly Detection Model Viewer

You can view an Anomaly Detection model in one of the following ways:

- Method 1
  1. Right-click the node where the model was built.
  2. Select **Go to Properties.**
  3. In the **Models** section in Properties, click .
- Method 2
  1. Select the workflow node where the model was built.
  2. Right-click and click **View Models.**
  3. Select the model to view. The Anomaly Detection Model Viewer opens in a new tab. The default name of an Anomaly Detection Models has `AD` in the name.

The information displayed in the model viewer depends on which kernel was used to build the model.

- If the Gaussian Kernel is used, then there is one tab, **Settings.**
- If the Linear kernel is used, then there are three tabs: **Coefficients**, **Compare**, and **Settings**.

Anomaly Detection model is a special kind of Support Vector Machine Classification model.

See Also:

- "AD Model Viewer for Linear Kernel"
- "SVM Classification Model Viewer"
- "AD Model Viewer for Gaussian Kernel"

13.1.2.1.1 AD Model Viewer for Gaussian Kernel The model viewer for an AD model with Gaussian Kernel has the following tabs:

- Summary (AD): Contains Model and Algorithm settings.
- Input (AD): Contains attributes used to build the model.

13.1.2.1.2 Settings (AD) The AD Settings Viewer has two tabs:

- Summary (AD)
- Input (AD)

13.1.2.1.3 Summary (AD) **General** settings describe the characteristics of the model, including:

- Owner
- Name
- Type
- Algorithm
- Target Attribute
- Creation Date
- Duration of Model Build
- Comments

> **See Also:** "General Settings"

Algorithm settings control the model build. Algorithm setting are specified when the Build node is defined.

> **See Also:** "Algorithm Settings for AD"

13.1.2.1.4 Input (AD) A list of the attributes used to build the model. For each attribute the following information is displayed:

- **Name**: The name of the attribute.
- **Data Type**: The data type of the attribute
- **Mining Type**:
  - Categorical
  - Numerical
- **Data Prep**: `YES` indicates that data preparation was performed.

When you select an attribute in the **Attributes** list, the transformation properties viewer displays the transformation, created by either the user or Automatic Data Preparation, in the **Model transformations** list.

To see the reverse transformation, click **Show Reverse Expression**. Transformations are displayed in SQL notation. Not all transformations have a reverse. Transformations and Reverse transformations are not always displayed.

13.1.2.1.5 Anomaly Detection Algorithm Settings Anomaly Detection models are built using a special version of the SVM classification, one-class SVM. The algorithm has these default settings:

- Kernel function: The default is System Determined. After the model is built, the kernel function used (Linear or Gaussian) is displayed.
- Tolerance value: The default is `0.001`
- Specify complexity factors: The default is `Do Not`
- Specify the rate of outliers: The default is `0.1`
- Active learning: `ON`
- Automatic Data Preparation: `ON`

See Also: "Algorithm Settings for AD"

13.1.2.1.6 AD Model Viewer for Linear Kernel

The model viewer for an AD model with a Linear Kernel has these tabs:

- Coefficients (SVMC Linear)
- Compare (SVMC Linear)
- Settings (AD)

13.1.2.2 Algorithm Settings for AD

The algorithm for Anomaly Detection is one-class SVM. The kernel setting is one of the following:

- System Determined (Default)
- Gaussian
- Linear

The settings that you can specify for any version of the Support Vector Machine (SVM) algorithm depend on which of the SVM Kernel Functions that you select:

- AD Algorithm Settings for Linear or System Determined Kernel
- AD Algorithm Settings for Gaussian Kernel

Note: After the model is built, the kernel function used (Linear or Gaussian) is displayed in **Kernel Function in Algorithm Settings**.

13.1.2.2.1 AD Algorithm Settings for Linear or System Determined Kernel

If you specify a linear kernel or if you let the system determine the kernel, then you can change the following settings:

- Tolerance Value
- Complexity Factor
- Rate of Outliers
- Active Learning

13.1.2.2.2 AD Algorithm Settings for Gaussian Kernel

If you specify the Gaussian Kernel, then you can change the following settings:

- Tolerance Value
- Complexity Factor

- Rate of Outliers
- Active Learning
- Standard Deviation (Gaussian Kernel)
- Cache Size (Gaussian Kernel)

13.1.2.2.3 Rate of Outliers The rate of outliers is the approximate rate of outliers (negative predictions) produced by a one-class SVM model on the training data. This rate indicates the percent of suspicious records.

The rate is a number greater than 0 and less than or equal to 1. The default value is `0.1`.

If you do not want to specify the rate of outliers, then deselect **Specify the rate of outliers**.

13.2 Association

Association is an unsupervised mining function for discovering association rules, that is predictions of items, that are likely to be grouped together. Oracle Data Mining provides one algorithm, Association Rules (AR).

To build an AR model, use an Association Node.

> **See Also:** "Troubleshooting AR Models"if the model has 0 rules or a very large number of rules

Data for Association Rules (AR) models is usually in transactional form, unlike the data for other kinds of models.

Oracle Data Mining does not support applying (scoring) AR models.

These topics describe AR models:

- Calculating Associations
- Data for AR Models
- Troubleshooting AR Models
- AR Model Viewers and Algorithm Settings

13.2.1 Calculating Associations

An association mining problem can be broken down into two subproblems:

1. Find all combinations of items in a set of transactions that occur with a specified minimum frequency. These combinations are called frequent Itemsets.
2. Calculate Association Rules that express the probability of items occurring together within frequent itemsets.

Apriori calculates the probability of an item being present in a frequent itemset, given that other items are present.

13.2.1.1 Itemsets

An itemset is any combination of two or more items in a transaction.

The maximum number of items in an itemset is user-specified.

- If the maximum is 2, then all the item pairs are counted.

- If the maximum is greater than 2, then all the item pairs, all the item triples, and all the item combinations up to the specified maximum are counted.

Association rules are calculated from itemsets. Usually, it is desirable to only generate rules from itemsets that are well-represented in the data. Frequent itemsets are those that occur with a minimum frequency specified by the user.

The minimum frequent itemset support is a user-specified percentage that limits the number of itemsets used for association rules. An itemset must appear in at least this percentage of all the transactions if it is to be used as a basis for Association Rules.

13.2.1.2 Association Rules

The Apriori algorithm calculates rules that express probable relationships between items in frequent itemsets. For example, a rule derived from frequent itemsets containing A, B, and C might state that if A and B are included in a transaction, then C is likely to also be included.

An association rule is of the form IF antecedent THEN consequent. An association rule states that an item or group of items, the antecedent, implies the presence of another item, the consequent, with some probability. Unlike Decision Tree rules, which predict a target, Association Rules simply express correlation.

Association rules have Confidence and Support:

- **Confidence** of an association rule indicates the probability of both the antecedent and the consequent appearing in the same transaction. Confidence is the conditional probability that the consequent occurs given the occurrence of the antecedent. In other words, confidence is the ratio of the rule support to the number of transactions that include the antecedent.
- **Support** of an association rule indicates how frequently the items in the rule occur together. Support is the ratio of transactions that include all the items in the antecedent and consequent to the number of total transactions.

13.2.2 Data for AR Models

Association Rules is normally used with transactional data, but it can also be applied to single-record case data (similar to other algorithms).

Association does not support text.

> **See Also:** "Text Mining in Oracle Data Mining"

Native transactional data consists of two columns:

- Case ID, either categorical or numerical
- Item ID, either categorical or numerical

Transactional data may also include a third column:

- Item value, either categorical or numerical

A typical example of transactional data is market basket data. In market basket data, a case represents a basket that might contain many items. Each item is stored in a separate row, and many rows may be needed to represent a case. The Case ID values do not uniquely identify each row. Transactional data is also called Multirecord Case Data.

When building an Association model, specify the following:

- **Item ID:** This is the name of the column that contains the items in a transaction.

- **Item Value:** This is the name of a column that contains a value associated with each item in a transaction. The Item Value column may specify information such as the number of items (for example, three apples) or the type of the item (for example, Macintosh Apples).

  The default value for **Item Value** is `<Existence>`. That is, one or more item identified by **Item ID** is in the basket.

  If you select a specific value for **Item Value**, you may have to perform appropriate data preparation. The maximum number of distinct values of Item Value is 10. If the specific value for Item Value is greater than 128, then bin the attribute specified in Item Value using a Transform node.

  For more information, see the discussion of Market Basket data in the *Oracle Data Mining User's Guide*.

13.2.2.1 Support for Text (AR)

In Oracle Data Miner, Association does *not* support text.

If the Oracle Data Mining API, supports text, but using test for Association is not recommended.

13.2.3 Troubleshooting AR Models

AR models may generate many rules with very low Support and Confidence. If you increase Support and Confidence, then you reduce the number of rules generated.

Usually, Confidence should be greater than or equal to Support.

If a model has no rules, then the following message is displayed in the **Rules** tab of the Model Viewer:

```
Model contains no rules. Consider rebuilding model with lower confidence and
support settings.
```

See Also: "Algorithm Settings for AR"to learn how to change these values

13.2.3.1 Algorithm Settings for AR

To change algorithm settings for an Association node, right-click the node, and select **Advanced Settings**.Then select the model. The following settings are displayed in the Algorithm settings tab:

1. Right-click the node.
2. Select **Advanced Settings**.
3. Select the model. The following settings are displayed in the Algorithm Settings tab:
   - Maximum rule length. The default is `4`
   - Minimum Confidence. The default is `10%`
   - Minimum Support. The default is `1%`
4. If no rules were generated, then:
   - First, try decreasing Minimum Support.
   - If that does not work, decrease the minimum Confidence value. It may be necessary to specify a much smaller value for either of these values.

5. After you have finished, click **OK.**
6. Run the node.

> **See Also:** "Algorithm Settings"

13.2.4 AR Model Viewers and Algorithm Settings

This section discusses the Association Rules Model viewer, the procedure to view the AR Model Viewer and the Algorithm Settings related to AR. The topics discussed are:

- Data for AR Models
- Algorithm Settings
- AR Model Viewer

13.2.4.1 AR Model Viewer

You can view an AR model in one of the following ways:

- Method 1
  1. Right-click the node where the model was built.
  2. Select **Go to Properties.**
  3. In the **Models** section in **Properties**, click .
- Method 2
  1. Select the workflow node where the model was built.
  2. Right-click and click **View Models.**
  3. Select the model to view.

The AR model viewer opens in a new tab. The default name of an Association model has `ASSOC` in the name. The AR Model viewer has these tabs:

- AR Rules (displayed by default)
- Itemsets
- Settings (AR))

13.2.4.1.1 AR Rules An Association Rule states that an item or group of items implies the presence of another item. Each rule has a probability. Unlike Decision Tree rules, which predict a target, Association Rules simply express correlation.

If an attribute is a nested column, then the full name is displayed as `COLUMN_NAME.SUBNAME`. For example, `GENDER.MALE`. If the attribute is a normal column, then just the column name is displayed.

Oracle Data Mining supports Association Rules that have one or more items in the antecedent (IF part of a rule) and a single item in the consequent (THEN part of the rule). The antecedent may be referred to as the *condition*, and the consequent as the *association.*

Rules have Confidence, Support and Lift.

> **See Also:** "AR Rules Display"

The **Rules** tab is divided into two sections: Filtering and Sorting in the upper section, and AR Rules Grid in the lower section. Sorting or filtering that is defined using the settings in the upper section apply to all the rules in the model. Sorting or filtering defined using the settings in the lower section apply to the grid display only.

You can perform the following functions in the **Rules** tab:

- **Sort By:** You can specify the order for display of the rules. You can sort rules by:
  - Lift, Confidence, Support, or Length
  - Ascending or Descending order

  For more sort options, click **More.** You can specify up to four levels of sorting, and you can specify the order for each level.

  > **See Also:** "Sorting"

- **Filter:** You can filter rules. To see filtering options, click **More.** You can specify the following in rules:
  - Minimum Lift
  - Minimum Support
  - Minimum Confidence
  - Maximum Items in Rules
  - Minimum Items in Rules

  > **See Also:** "Filtering"

- **Fetch Size:** Association models often generate many rules. Specify the number of rules to examine by clicking **Fetch Size.** The default is `1000`.
- **Query:** You can query the database using the criteria that you specify. For example, if you change the default sorting order, specify filtering, or change the fetch size, then click **Query.**

13.2.4.1.2 AR Rules Grid The lower part of the **Rules** tab displays the retrieved rule in a grid. The following is displayed above the grid:

- **Available Rules**: The total number of rules in the model.
- **Rules Retrieved**: The number of rules retrieved by the query, that is, the number of rules retrieved subject to filtering.
- **Rule Content**: For maximum information, select Name, Subname, and Value; you can select fewer characteristics from the menu. This selection applies to the rules in the grid only. Rule content is smart in the sense that it sets this value to whatever is more appealing given the nature of the model.
- **Search:** To search the rules, use the search box, identified by .

  The drop-down list enables you to search rules by **All** (the default), by **Antecedent,** or by **Consequent.** If you search by Antecedent for 117, then all rules that have 117 in the antecedent are displayed.

  To clear a search, click .

  > **See Also:** "AR Rules Display"for more information about how the rules are displayed

13.2.4.1.3 AR Rules Display For each rule, the Rules grid displays the following information:

- ID: Identifier for the rule, a string of integers.
- Condition
- Association
- Lift: A bar is included in the column. The size of the bar is set to scale to the largest lift value provided by the model by any rule.

  See Also: "Lift for AR Rules"for more information about Lift

- Confidence:

  See Also: "Confidence for AR Rules"

- Support:

  See Also: "Support for AR Rules"

- Length
- Antecedent Support
- Condition Support

You can perform the following tasks:

- Sort: You can sort the items in the grid by clicking on the title of the column. This sorting applies to the grid only.
- View details: To see details for a rule, click the rule and examine the Rule Details.
- Determine validity of a rule: To determine if a rule is valid, you must use support and confidence plus lift, as described in Lift for AR Rules.

For more information, including examples of these statistics, see the discussion of evaluating association rules in *Oracle Data Mining Concepts*.

> **See Also:** *Oracle Data Mining Concepts* for more information, including examples of these statistics.

13.2.4.1.4 Lift for AR Rules Both Support for AR Rules and Confidence for AR Rules must be used to determine if a rule is valid. However, there are times when both of these measures may be high, and yet produce a rule that is not useful.

Lift indicates the strength of a rule over the random co-occurrence of the Antecedent and the Consequent, given their individual support. It provides information about the improvement, the increase in probability of the consequent given the antecedent. Lift is defined as follows:

```
(Rule Support) /(Support(Antecedent) * Support(Consequent))
```

Lift can also be defined as the Confidence of the combination of items divided by the support of the consequent.

13.2.4.1.5 Confidence for AR Rules The Confidence of a rule indicates the probability of both the Antecedent and the Consequent appearing in the same transaction. Confidence is the conditional probability of the Consequent, given the Antecedent.

AR rules are of the form `IF antecedent THEN consequent`.

13.2.4.1.6 Support for AR Rules The Support of a rule indicates how frequently the items in the rule occur together. Support is the ratio of transactions that include all the items in the Antecedent and Consequent to the number of total transaction.

AR rules are of the form `IF antecedent THEN consequent`.

13.2.4.1.7 Rule Details The information in the rule grid is displayed in a readable format in the **Rule Detail** list.

13.2.4.1.8 Sorting The default sort is:

1. Sort by Lift in descending order
2. Sort by Confidence in descending order
3. Sort by Support in descending order
4. Sort by rule length in descending order

The sorting specified here applies to all rules in the model.

13.2.4.1.9 Filtering To see all the filtering options, click **More**.

You can specify the following:

- **Filter**: Filter rules are based on values of rule characteristics. You can specify the following:
  - Minimum lift
  - Minimum support
  - Minimum confidence
  - Maximum items in rule
  - Minimum items in rule
- **Fetch Size:** It is the maximum number of rows to fetch. The default is `1000`. Smaller values result in faster fetches.
- Define Item Filters to reduce the number of rules returned.

To define a filter, select **Use Filter**. After you define the filter, click **Query**.

13.2.4.1.10 Item Filters Item filters enable you to see only those rules that contain what you are interested in. A rule filter must consider the item as being required for the Association, Condition, or Both. The rule filter uses `OR` logic for each side of the Rule (Association Collection, Condition Collection). However, the rule filter performs an `AND` rule across the collection. So, for a Rule to be returned, it must have at least one Association item AND one Condition item.

You can manage Item Filters using these controls:

- To open the Add Item Filter dialog box, click .
- To delete selected item filters, click .
- To change the Filter column of selected rows to Both, click . Both implies Association and Condition.
- To change the Filter column of selected rows to Condition, click .
- To change the Filter column of selected rows to Association, click .

13.2.4.1.11 Add Item Filter To open the Add Item Dialog, click .

The exact information that is displayed depends on the model. For example, if data has different values for the model that you are viewing, then there is a **Values** column.

Click **More** to see all possibilities:

- Specify sorting for item filters: the default is to sort by Attribute Descending and then by Support Ascending.
- Specify a name for the filter.
- Change the Fetch Size from the default of `100,000`.
- If you made any changes, then click **Query** to retrieve the attribute or value pairs.
- Filter the retrieved items by name or value.
- Select one or more item in the grid.
- Select how to use items when filtering rules.

Click **OK** when you have finished defining the filter.

13.2.4.1.12 Itemsets Rules are calculated from itemsets. The **Itemsets** tab displays information about the itemsets.

If an attribute is a nested column, then the full name is displayed as `COLUMN_NAME.SUBNAME`. For example, `GENDER.MALE`. If the attribute is a normal column, then just the column name is displayed.

Itemsets have Support. Each itemset contains one or more items.

- Sort Itemsets: **Sort By** specifies the order of itemsets. You can sort itemsets by:
  - ID
  - Number of Items
  - Support in Ascending Order
  - Support in Descending Order

  By default, itemset is sorted by **Support in Descending order.** For more sort options, click **More.** To change the sort order, make changes and then click **Query.**
- Filter Itemsets
- View Itemset details. Click the itemset to view the details.

The **Itemsets** tab displays the following information:

- **Available Itemsets**: The total number of itemsets in the model.
- **Itemsets Retrieved**: The number of itemsets retrieved by the query. That is, the number of itemsets retrieved subject to filtering.
- **Itemset Content**: For maximum information, select all three—Name, Subname, and Value. You can select a few characteristics from the menu.

> **See Also:** "Itemsets Display"for more information displayed for each itemset

To see details of a itemset, click the itemset and examine Rule Details.

Other Tabs: The AR Model viewer has these other tabs:

- AR Rules

- Settings (AR))

13.2.4.1.13 Itemsets Display

For each itemset, the Itemsets grid displays the following information:

- ID: Identifier for the itemset, a string of integers
- Items
- Support. A bar in the column illustrates the relative size of the support.
- Number of Items in the itemset

13.2.4.1.14 Itemset Details

To see itemset details, select one or more itemsets in the itemsets grid. The information in the itemsets grid is displayed in a more readable format.

13.2.4.1.15 Settings (AR))

The **Settings** tab has the following tabs:

- Summary

> **Note:** AR models are not scoreable, that is, they cannot be applied to new data. Models that are not scoreable do not have an **Attributes** tab in the model viewer.

Other Tabs: The AR Model viewer has these other tabs:

- Itemsets
- AR Rules

13.2.4.1.16 Summary

This tab show information about the model. The summary tab show two kinds of information:

- **General:** Lists the following:
  - Type of Model
  - Owner of the model (Classification, Regression, and so on)
  - Model Name (the Schema where the model was built)
  - Creation Date
  - Duration of the Model build (in minutes)
  - Model size in MB
  - Comments

  See Also: "General Settings"

- **Algorithm:** Lists the following:
  - Automatic Preparation (on or off),
  - Minimum Confidence
  - Minimum Support

  To change these values, right-click the model node and select **Advanced Settings.**

  See Also: "Algorithm Settings for AR"if the model has no rules

13.2.4.1.17 Algorithm Settings

Association (AR) supports these settings:

- **Maximum Rule Length**: The maximum number of attributes in each rule. This number must be an integer between 2 and 20. Higher numbers of rules result in slower builds. The default value is 4.

  You can change the number of attributes in a rule, or you can specify no limit for the number of attributes in a rule. Specifying many attributes in each rule increases the number of rules considerably. A good practice is to start with the default and increase this number slowly.

- **Minimum Confidence**: Confidence indicates how likely it is that these items to occur together in the data. Confidence is the conditional probability that the consequent will occur given the occurrence of the antecedent.

  See Also: "Association Rules"

  Confidence is a number between 0 and 100 indicating a percentage. High confidence results in a faster build. The default is 10 percent.

- **Minimum Support**: A number between 0 and 100 indicating a percentage. Support indicates how often these items occur together in the data.

  See Also: "Association Rules"

  Smaller values for support results in slower builds and requires more system resources. The default is 1 percent.

13.3 Decision Tree

The Decision Tree algorithm is a Classification algorithm that generates rules. Oracle Data Mining supports of the Decision Tree (DT) algorithm. This section contains the following topics:

- Decision Tree Algorithm
- Build, Test, and Apply Decision Tree Models
- Decision Tree Model Viewer and Algorithm Settings

13.3.1 Decision Tree Algorithm

The Decision Tree algorithm is based on conditional probabilities. Unlike Naive Bayes, Decision Trees generate rules. A rule is a conditional statement that can easily be used by humans and easily used within a database to identify a set of records.

See Also: "Decision Tree Rules"

The Decision Tree algorithm:

- Creates accurate and interpretable models with relatively little user intervention. The algorithm can be used for both binary and multiclass classification problems. The algorithm is fast, both at build time and apply time. The build process for the Decision Tree Algorithm is parallelized. Scoring can be parallelized irrespective of the algorithm.
- Predicts a target value by asking a sequence of questions. At a given stage in the sequence, the question that is asked depends upon the answers to the previous

questions. The goal is to ask questions that together uniquely identify specific target values.

Decision Tree scoring is especially fast. The tree structure, created in the model build, is used for a series of simple tests, (typically 2-7). Each test is based on a single predictor. It is a membership test: either IN or NOT IN a list of values (categorical predictor); or LESS THAN or EQUAL TO some value (numeric predictor).

During the model build, the Decision Tree algorithm must repeatedly find the most efficient way to split a set of cases (records) into two child nodes. Oracle Data Mining offers two homogeneity metrics, gini and entropy, for calculating the splits. The default metric is gini.

13.3.1.1 Decision Tree Rules

Rules provide model transparency, a window on the inner workings of the model. Rules show the basis for the prediction of the model. Oracle Data Mining supports a high level of model transparency.

Confidence and Support are used to rank the rules generated by the Decision Tree Algorithm:

- **Support:** It is the number of records in the training data set that satisfy the rule.
- **Confidence:** It is the likelihood of the predicted outcome, given that the rule has been satisfied.

13.3.2 Build, Test, and Apply Decision Tree Models

The Decision Tree manages its own data preparation internally. It does not require pre-treatment of the data. The Decision Tree is not affected by Automatic Data Preparation.

The Decision Tree interprets missing values as missing at random. The algorithm does not support nested tables and thus does not support sparse data.

Building Decision Tree model

To build a Decision Tree model, use a Classification Node. In Oracle Data Mining 12*c* Release 1(12.1) or later, Decision Tree supports nested data. The Decision Tree supports text for Oracle Database 12*c*, but not for earlier releases.

Testing Decision Tree model

By default, a Classification Node tests all models that it builds. The test data is created by splitting the input data into build and test subsets. You can also test a Decision Tree model using a Test Node.

See Also: "Testing Classification Models"

Tuning Decision Tree model

After you build and test a Decision Tree model, you can tune it.

See Also: "Tuning Classification Models"

Applying Decision Tree model

To apply a model, use an Apply Node.

13.3.3 Decision Tree Model Viewer and Algorithm Settings

This section discusses the Decision Tree model viewer, the procedure to view the Decision Tree Model viewer and the algorithm settings related to Decision Tree. Topic include:

- Decision Tree Model Viewer
- Decision Tree Algorithm Settings

13.3.3.1 Decision Tree Model Viewer

You can view a Decision Tree model in one of the following ways:

- Method 1
  1. Right-click the node where the model was built.
  2. Select **Go to Properties.**
  3. In the **Models** section in Properties, click .
- Method 2
  1. Select the workflow node where the model was built.
  2. Right-click and click **View Models.**
  3. Select the model to view.

  The Decision Tree viewer opens in a new tab. The default name of a Decision Tree model has `DT` in the name.

The Tree viewer has two tabs:

- **Tree:** This tab is displayed by default. Use the Structure Window to navigate and analyze the tree. It is split horizontally into two panes:
  – The upper pane displays the tree. The root node is at the top of the pane. The following information is displayed for each node of the tree:
    * **Node number.** `0` is the root node.
    * **Prediction**, the predicted target value.
    * **Support** for the prediction.
    * **Confidence** for the prediction.
    * A histogram shows the distribution of target values in the node.
    * **Split,** the attribute used to split the node (Leaf nodes do not have splits).
  – The lower pane displays rules. To view the rule associated with a node or a link, select the node or link. The rule is displayed in the lower pane. The following information is displayed in the lower pane:
    * Rule
    * Surrogates
    * Target Values
- Settings (DT)

Icons and menus at the top of the upper pane control how the tree and its nodes are displayed. You can perform the following tasks:

- Zoom in or zoom out for the tree. You can also select a size from the drop-down list. You can also fit the tree to the window.

- Change the layout to horizontal. The default Layout Type for the tree is vertical.
- Hide the histograms displayed in the node.
- Show less detail.
- Expand all nodes.
- Save Rules.

13.3.3.1.1 Save Rules

To save the Decision Tree or Clustering Rules:.

1. Click **Save Rules** on the far right of the upper tab.
   By default, rules are saved for leaf nodes only to the Microsoft Windows Clipboard. You can then paste the rules into any rich document, such as a Microsoft Word document.
   You can deselect **Leaves Only** to save all rules.
2. To save rules to a file, click **Save to File** and specify a file name.
3. Select the location of the file in the **Save** dialog box. By default, the rules are saved as an HTML file.
4. Click **OK.**

13.3.3.1.2 Settings (DT)

The **Settings** tab has these tabs:

- DT Summary
- DT Inputs
- DT Target Values
- **Cost Matrix/Benefit** tab: If you tune the model, then this tab displays the cost matrix created by tuning.

13.3.3.1.3 DT Summary

This tab displays the following information about the model:

- **General:** Describes the following:
  - Type of model
  - Owner of the model (the Schema where the model was built)
  - Model Name
  - Creation Date
  - Duration of the model Build (in minutes)
  - Size of the model (in MB)
  - Comments (if the model has any comments)

  See Also: "General Settings"

- **Algorithm** settings are Decision Tree Algorithm Settings.

13.3.3.1.4 DT Inputs

This tabs shows information about those attributes used to build the model.

Oracle Data Miner does not necessarily use all of the attributes in the build data. For example, if the values of an attribute are constant, then that attribute is not used.

For each attribute used to build the model, this tab displays:

- **Name**

- **Data type**
- **Mining Type:** Categorical, Numerical, or text.
- **Target:** The in this column indicates that the attribute is a target.
- **Data Prep**
  - `YES`: Indicates that data preparation was performed.

    If Data Prep is `Yes`, then select the column (click it). The data preparation is displayed in Data Preparation. To see the reverse transformation for data preparation, select **Show Reverse Transformation.** Transformations are displayed in SQL notation. Not all transformations have a reverse. Transformations and Reverse transformations are not always displayed.
  - `NO`: Indicates that no data preparation was performed

13.3.3.1.5 DT Target Values Displays the target attributes, their data types, and the values of each target attribute.

13.3.3.2 Decision Tree Algorithm Settings

The Decision Tree algorithm supports these settings:

- **Homogeneity Metric**:
  - Gini (default)
  - Entropy
- **Maximum Depth**: The maximum number of levels of the tree. The default is 7. The value must be an integer in the range 2 to 20.
- **Minimum Records in a Node**: The minimum number of records in a node. The default is 10. The value must be an integer greater than or equal to 0.
- **Minimum Percent of Records in a Node**: The default is 0.05. The value must be a number in the range 0 to 10.
- **Minimum Records for a Split**: The minimum number of records for a split. The default is 20. The value must be an integer greater than or equal to 0.
- **Minimum Percent of Records for a Split**: The default is 0.1. The value must be a number in the range 0 to 20.

13.4 Expectation Maximization

Expectation Maximization (EM) is a density estimation technique. Oracle Data Mining implements EM as a distribution-based clustering algorithm that uses probability density estimation.

In density estimation, the goal is to construct a density function that captures how a given population is distributed. The density estimate is based on observed data that represents a sample of the population.

Note: Expectation Maximization requires Oracle Database 12*c*.

Dense areas are interpreted as components or clusters. Density-based clustering is conceptually different from distance-based clustering (such as *k*-Means), where emphasis is placed on minimizing intercluster and maximizing the intracluster distances.

The shape of the probability density function used in EM effectively predetermines the shape of the identified clusters. For example, Gaussian density functions can identify single peak symmetric clusters. These clusters are modeled by single components. Clusters of a more complex shape need to be modeled by multiple components. The EM algorithm assigns model components to high-level clusters by default.

13.4.1 Build and Apply an EM Model

To build an EM model, use a Clustering Node. To build an EM model, you must be connected to Oracle Database 12*c*.

To apply an EM model, use an Apply Node.

13.4.2 EM Model Viewer and Algorithm Settings

This section discusses the Expectation Maximization Model Viewer, the procedure to view the EM Model viewer and the algorithm settings related to EM. The topics include:

- EM Model Viewer
- EM Algorithm Settings

13.4.2.1 EM Model Viewer

You can view and examine an EM Model in an EM Model Viewer. You can view the model in one of the following ways:

- Method 1
  1. Right-click the node where the model was built.
  2. Select **Go to Properties.**
  3. In the **Models** section in Properties, click .
- Method 2
  1. Select the workflow node where the model was built.
  2. Right-click and click **View Models**.
  3. Select the model to view.

  The EM model viewer opens in a new tab. The default name of an EM model has `EM` in the name.

The Tree tab is displayed by default. The EM model viewer has these tabs:

- EM, KM, and OC Tree
- Cluster (Viewer)
- EM Component
- Cluster Model Settings (Viewer)
- EM, KM, and OC Compare
- EM Details

13.4.2.1.1 EM Component

The **Component** tab provides detailed information about the components of the EM model.

The tab is divided into several panes.

The top pane specifies the cluster to view:

- **Component:** It is the integer that identifies the cluster. The default value is 1.
- **Prior:** It is the priority for the specified component.
- **Filter by Attribute Name:** Enables you to display only those attributes of interest. Enter the attribute name, and click **Query**.
- **Fetch Size:** It is the number of records fetched. The default is 2,000.

The middle pane displays information about the attributes in the specified component:

- You can search for a specified **Attribute** using the search box.
- The attributes are displayed in a grid. The grid lists Attribute (name), Distribution (as a histogram), and Mean and Variance (numerical attributes only).

  To sort any of these columns, click the column title.
- To see a larger version of the histogram for an attribute and information about the distribution, select the attribute. The histogram is displayed in the bottom pane.

The bottom pane displays a large version of the selected histogram, data, and projections (if any):

- The **Chart** tab contains a larger version of the histogram of the selected attribute.
- The **Data** tab shows the frequency of the histogram bins.
- The **Projections** tab lists projects in a grid, listing Value and Coefficient for each Attribute Subname.

13.4.2.1.2 EM Details The **Details** tab shows global details for the EM model. The following information is displayed:

- Log Likelihood Improvement
- Number of Clusters
- Number of Components

13.4.2.2 EM Algorithm Settings

EM supports these settings:

- **Number of Clusters** is the maximum number of leaf clusters generated by the algorithm. EM may return fewer clusters than the number specified, depending on the data. The number of clusters returned by EM cannot be greater than the number of components, which is governed by algorithm-specific settings. Depending on these settings, there may be fewer clusters than components. If component clustering is disabled, then the number of clusters equals the number of components.

  The default is system determined. To specify a specific number of clusters, click **User specified** and enter an integer value.
- **Component Clustering** is selected by default.

  Component Cluster Threshold specifies a dissimilarity threshold value that controls the clustering of EM components. Smaller values may produce more clusters that are more compact while large values may produce fewer clusters that are more spread out. The default value is 2.
- **Linkage Function** enables the specification of a linkage function for the agglomerative clustering step. The linkage functions are:

- **Single** uses the nearest distance within the branch. The clusters tend to be larger and have arbitrary shapes.

  Single is the default.
- **Average** uses the average distance within the branch. There is less chaining effect and the clusters are more compact.
- **Complete** uses the maximum distance within the branch. The clusters are smaller and require a strong component overlap.

- **Approximate Computation** indicates whether the algorithm should use approximate computations to improve performance.

  For EM, approximate computation is appropriate for large models with many components and for data sets with many columns. The approximate computation uses localized parameter optimization that restricts learning to parameters that are likely to have the most significant impact on the model.

  The values for approximate Computation are:

  - **System Determined** (Default)
  - **Enable**
  - **Disable**

- **Number of Components** specifies the maximum number of components in the model. The algorithm automatically determines the number of components, based on improvements in the likelihood function or based on regularization, up to the specified maximum.

  The number of components must be greater than or equal to the number of clusters.

  The default number of components is 20.

- **Max Number of Iterations** specifies the maximum number of iterations in the EM core algorithm. Maximum number of iterations must be greater than or equal to 1. This setting applies to the input table/view as a whole and does not allow per attribute specification.

  The default is 100.

- **Log Likelihood Improvement** specifies the percentage improvement in the value of the log likelihood function required to add a new component to the model.

  The default value is 0.001

- **Convergence Criterion** specifies the convergence criterion for EM. The convergence criteria are:

  - **System Determined** (Default)
  - **Bayesian Information Criterion**
  - **Held-aside data set**

- **Numerical Distribution** specifies the distribution for modeling numeric attributes. The options are the following distributions:

  - **Bernoulli**
  - **Gaussian**
  - **System Determined** (Default)

When the Bernoulli or Gaussian distribution is chosen, all numerical attributes are modeled using the same distribution. When the distribution is system-determined, individual attributes may use different distributions (either Bernoulli or Gaussian), depending on the data.

- **Gather Class Statistics** enables or disables the gathering of descriptive statistics for clusters (centroids, histograms, and rules). Disabling the cluster statistics will result in smaller models and will reduce the model details calculated.

  The default is to enable or select the **Gather Class Statistics.**

  – If you disable Gather Class Statistics, then you will *not* be able to view models.

  – If you enable Gather Class Statistics, then you can specify **Min Percent of Attribute Rule Support**.

  Min Percent of Attribute Rule Support specifies the percentage of the data rows assigned to a cluster that must be present in an attribute to include that attribute in the cluster rule. The default value is 0.1.

- **Data Preparation and Analysis** specifies settings for data preparation and analysis. To view or change the selections, click **Settings**.

  See Also: "EM Data Preparation and Analysis Settings"

Click **OK** after you are done.

13.4.2.2.1 EM Data Preparation and Analysis Settings

This dialog box enables you to view or change these settings:

- **Max Number of Correlated 2D Attributes** specifies the maximum number of correlated two-dimensional attributes that will be used in the EM model. Two-dimensional attributes correspond to columns that have a simple data type (not nested).

  The default is 50.

- **Number of Projections per Nested Column** specifies the number of projections that will be used for each nested column. If a column has fewer distinct attributes than the specified number of projections, then the data will not be projected. The setting applies to all nested columns.

  The default is 50.

- **Number of Quantile Bins (Numerical Columns)** specifies the number of quantile bins that will be used for modeling numerical columns with multivalued Bernoulli distributions.

  The default is system determined.

- **Number of Top-N Bins (Categorical Columns)** specifies the number of top-N bins that will be used for modeling categorical columns with multivalued Bernoulli distributions.

  The default is system determined.

- **Number of Equi-Width Bins (Numerical Columns)** specifies the number of equi-width bins that will be used for gathering cluster statistics for numerical columns.

  The default is 11.

- **Include uncorrelated 2D Attributes** specifies whether uncorrelated two-dimensional attributes should be included in the model or not. Two-dimensional attributes correspond to columns that are not nested.

  The values are:

  – **System Determined** (Default)

  – **Enable**

  – **Disable**

When you have finished making changes, click **OK**.

13.5 Generalized Linear Models

Generalized Linear Models (GLM) is a statistical technique for linear modeling. Oracle Data Mining supports GLM for both Regression and Classification. The following topics describe GLM models:

- Generalized Linear Models Overview
- GLM Classification Models
- GLM Regression Models
- GLM Model Viewers and Algorithm Settings

13.5.1 Generalized Linear Models Overview

Generalized Linear Models (GLM) include and extend the class of linear models referred to as Linear Regression.

Oracle Data Mining includes two of the most popular members of the GLM family of models with their most popular link and variance functions:

- Linear Regression with the identity link and variance function equal to the constant 1 (constant variance over the range of response values).
- Logistic Regression with the logistic link and binomial variance functions.

In Oracle Database 12*c*, GLM Classification and Regression are enhanced to implement Feature Selection and Feature Generation. This capability, when specified, can enhance the performance of the algorithm and improve the accuracy and interpretability.

> **See Also:** "Data Preparation for GLM" for information about how to use ADP with GLM

13.5.1.1 Linear Regression

Linear Regression is the GLM Regression algorithm supported by Oracle Data Mining. The algorithm assumes no target transformation and constant variance over the range of target values.

13.5.1.2 Logistic Regression

Binary Logistic Regression is the GLM classification algorithm supported by Oracle Data Mining. The algorithm uses the logit link function and the binomial variance function.

13.5.1.3 Data Preparation for GLM

Oracle recommends that you use Automatic Data Preparation with GLM.

13.5.2 GLM Classification Models

You can perform the following tasks with GLM Classification model:

- **Build and Test GLM Classification Model:** To build and test a GLM Classification (GLMC) model, use a Classification Node. By default, the Classification Node tests the models that it builds. Test data is created by splitting the input data into build and test subsets. You can also test a model using a Test Node.

  See Also: "Testing Classification Models"

- **Tune GLM Classification Model:** After you build and test a GLM classification model, you can tune it.

  See Also: "Tuning Classification Models"

- **Apply GLM Classification Model:** To apply the GLM Classification model, use an Apply Node.

13.5.3 GLM Regression Models

You can perform the following tasks with GLM Regression models:

- **Build and Test GLM Regression model:** To build and test a GLM Regression (GLMR) model use a Regression Node. By default, a Regression Node tests the models that it builds. Test data is created by splitting the input data into build and test subsets. You can also test a model using a Test Node.

  See Also: "Testing Regression Models" for more information about testing

- **Apply GLM Regression model:** To apply a GLM Regression model, use an Apply Node.

13.5.4 GLM Model Viewers and Algorithm Settings

This section discusses the Generalized Linear Model viewer, the procedure to view the GLM viewer and the algorithm settings related to the GLM model. The topics are:

- GLM Classification Model Viewer
- GLM Classification Algorithm Settings
- GLM Regression Model Viewer
- GLM Regression Algorithm Settings

13.5.4.1 GLM Classification Model Viewer

The GLM Classification (GLMC) model viewer displays characteristics of a GLMC model. GLMC is also known as Logistic Regression. To view a GLMC model, use one of these methods:

Method 1

- Method 1

  1. Right-click the node where the model was built.
  2. Select **Go to Properties**.
  3. In the **Models** section in **Properties**, click .

- Method 2
  1. Select the workflow node where the model was built.
  2. Right-click and click **View Models.**
  3. Select the model to view.

  The GLMC viewer opens in a new tab. The default name of a GLM model has `GLM` in the name. The **Detail** tab is displayed by default.

The viewer has these tabs:

- **Details.** Described in GLMC Details.
- **Coefficients.** Described in GLMC Coefficients.
- **Compare.** Described in GLMC Compare.
- **Diagnostics.** Described in GLMC Diagnostics. Diagnostics are not generated by default.
- **Settings.** Described in GLMC Settings.

13.5.4.1.1 GLMC Details

Model Details list global metrics for the model as a whole. The metrics display has two columns: **Name** of the metric and **Value** of the metric. The following metrics are displayed:

- Akaike's criterion (AIC) for the fit of the intercept only model
- Akaike's criterion model for the fit of the intercept and the covariates (predictors) mode
- Dependent mean
- Likelihood ratio chi-square value
- Likelihood ratio chi-square probability value
- Likelihood ratio degrees of freedom
- Model converged (Yes or No)
- -2 log likelihood of the intercept only model
- -2 log likelihood of the mode
- Number of parameters (number of coefficients, including the intercept)
- Number of rows
- Correct Prediction percentage
- Incorrectly predicted percentage of rows
- Tied cases prediction, that is, cases where no prediction can be made
- Pseudo R-Square Cox and Snell
- Pseudo R-Square Nagelkerke
- Schwartz's Criterion (SC) for the fit of the intercept-only model
- Schwartz's Criterion for the fit of the intercept and the covariates (predictors) model
- Termination (normal or not)
- Valid covariance matrix (Yes or No)

> **Note:** The exact list of metrics computed depends on the model settings.

Other Tabs: The viewer has these other tabs:

- GLMC Coefficients
- GLMC Compare
- GLMC Diagnostics (if generated)
- GLMC Settings

13.5.4.1.2 GLMC Coefficients

The C**oefficient** tab enables you to view GLM coefficients. The viewer supports sorting to control the order in which coefficients are displayed and filtering to select the coefficients to display.

The default is to sort coefficients by absolute value. If you deselect **Sort by absolute value**, click **Query**.

The default fetch size is 1000 records. To change the fetch size, specify a new number of records and click **Query**.

> **Note:** After you change any criteria on this tab, click **Query** to query the database. You must click **Query** even for changes such as selecting or deselecting sort by absolute value or changing the fetch size.

The relative value of coefficients is shown graphically as a bar, with different colors for positive and negative values. If a coefficient is close to 0, then the bar may be too small to display.

> **See Also:** "Sort and Search GLMC Coefficients" for information about sorting and searching the grid

- **Target Value**: Select a specific target value and see only those coefficients. The default is to display the coefficients of the value that occurs least frequently. It is possible for a target value to have no coefficients; in that case, the list has no entries.
- **Sort by absolute value**: The default is to sort the list of coefficients by absolute value; you can deselect this option.
- **Fetch Size**: The number of rows displayed. The default is 1000. To figure out if all coefficients are displayed, choose a fetch size that is greater than the number of rows displayed.

Coefficients are listed in a grid. If no items are listed, then there are no coefficients for that target value. The coefficients grid has these columns:

- **Attribute**: Name of the attribute
- **Value**: Value of the attribute
- **Coefficient**: The linear coefficient estimate for the selected target value is displayed. A bar is shown in front of (and possibly overlapping) each coefficient. The bar indicates the relative size of the coefficient. For positive values, the bar is light blue; for negative values, the bar is red. (If a value is close to 0, then the bar may be too small to be displayed.)

- **Standardized coefficient:** The coefficient rescaled by the ratio of the standard deviation of the predictor to the standard deviation of the target.

  The standardized coefficient places all coefficients on the same scale, so that you can, at a glance, tell the large contributors from the small ones.
- Standard error
- **Exp (Coefficient).** This is the exponent of the coefficient.
- **Standard Error** of the estimate.
- Wald Chi Square
- Probability of greater than Chi Square
- **Test Statistic**: For linear Regression, the t-value of the coefficient estimate; for Logistic Regression, the Wald chi-square value of the coefficient estimate
- **Probability** of the test statistic. Used to analyze the significance of specific attributes in the model
- **Variance Inflation Factor**
  - `0` for the intercept
  - `Null` for Logistic Regression
- **Lower Coefficient Limit**, lower confidence bound of the coefficient
- **Upper Coefficient Limit**, upper confidence bound of the coefficient
- **Exp (Coefficient)**
  - Exponentiated coefficient for Logistic Regression
  - Null for Linear Regression
- **Exp (Lower Coefficient Limit)**
  - exponentiated coefficient of the lower confidence bound for Logistic Regression
  - Null for Linear Regression
- **Exp (Upper Coefficient Limit**)
  - Exponentiated coefficient of upper confidence bound for Logistic Regression
  - Null for Linear Regression

Note: Not all statistics are necessarily returned for each coefficient.

Statistics are null if any of the following are true:

- The statistic does not apply to the mining function. For example, Exp (coefficient) does not apply to Linear Regression.
- The statistic cannot be calculated because of limitations in system resources.
- The value of the statistics is infinity.
- If the model was built using Ridge Regression, or if the covariance matrix is found to be singular during the build, then coefficient bounds (upper and lower) have the value NULL.

Other Tabs: The viewer has these other tabs:

- GLMC Details
- GLMC Compare
- GLMC Diagnostics (if generated)
- GLMC Settings

13.5.4.1.3 Sort and Search GLMC Coefficients You can sort the numerical columns by clicking the title of the column. For example, to arrange the coefficients in increasing order, click **Coefficients** in the grid.

Use to search for items. The default is to search by **Attribute** (name).

There are search options that limit the columns displayed. The filter settings with the `(or)/(and)` suffixes enable you to enter multiple strings separated by spaces. For example, if you select `Attribute/Value/Coefficient(or)`, the filter string `A .02` produces all columns where the Attribute or the Value Type starts with the letter *A* or the Coefficient starts with 0.02.

To clear a search, click .

13.5.4.1.4 GLMC Compare GLM Classification Compare viewer is similar to the SVM Coefficients Compare viewer except that the GLM model can only be built for binary classification models. Only two target class values would be available to compare.

> **See Also:** "GLMC Compare" for details

Other Tabs: The viewer has the following tabs:

- GLMC Details
- GLMC Coefficients
- GLMC Diagnostics
- GLMC Settings

13.5.4.1.5 GLMC Diagnostics The **Diagnostics** tab for GLM Classification displays diagnostics for each Case ID in the build data. You can filter the results.

> **Note:** Diagnostics are not generated by default. To generate diagnostics, you must specify a Case ID and select **Generate Row Diagnostics** in **Advanced Settings**.

The following information is displayed in the Diagnostics grid:

- **CASE_ ID**
- **TARGET_VALUE** for the row in the training data
- **TARGET_VALUE_PROB**, probability associated with the target value
- **HAT**, value of diagonal element of the HAT matrix
- **WORKING_RESIDUAL**, the residual with the adjusted dependent variable
- **PEARSON_RESIDUAL**, the raw residual scaled by the estimated standard deviation of the target
- **DEVIANCE_RESIDUAL**, contribution to the overall goodness of the fit of the model

- **C**, confidence interval displacement diagnostic
- **CBAR**, confidence interval displacement diagnostic
- **DIFDEV**, change in the deviance due to deleting an individual observation
- **DIFCHISQ**, change in the Pearson chi-square

Other Tabs: The viewer has these other tabs:

- GLMC Details
- GLMC Coefficients
- GLMC Compare
- GLMC Settings

13.5.4.1.6 GLMC Settings

The **Settings** tab has these other tabs:

- GLMC Data Usage:
- GLMC Summary
- GLMC Inputs
- GLMC Target Values
- **Cost Matrix/Benefit:** If you tune the model, then this tab displays the cost matrix created by tuning.

Other Tabs: The viewer has these other tabs:

- GLMC Details
- GLMC Coefficients
- GLMC Compare
- GLMC Diagnostics

13.5.4.1.7 GLMC Data Usage:

Describes data usage for the model.

> **See Also:** "Viewing and Changing Data Usage"

13.5.4.1.8 GLMC Summary

General settings describe the characteristics of the model, including:

- Name
- Type
- Algorithm
- Target Attribute
- Creation Date
- Duration of Model Build
- Comments

> **See Also:** "General Settings"

Algorithm settings control model build. Algorithm setting are specified when the Build node is defined.

See Also: "GLM Classification Algorithm Settings" for a list of model settings

After a model is built, values calculated by the system are displayed on this tab. For example, if you select **System Determined** for **Enable Ridge Regression**, this tab shows if Ridge Regression was enabled and what ridge value was calculated.

Other Tabs: The **Settings** tab has these other tabs:

- GLMC Inputs
- GLMC Target Values

13.5.4.1.9 GLMC Inputs A list of the attributes used to build the model. For each attribute the following information is displayed:

- **Name**: The name of the attribute.
- **Data type** The data type of the attribute.
- **Mining type**: Categorical or Numerical.
- **Target**: The icon indicates that the attribute is a target attribute.
- **Data Prep**: `YES` indicates that data preparation was performed on the attribute.

When you select an attribute in the **Attributes** list, the transformation properties viewer displays the imbedded transformation, created by either the user or Automatic Data Preparation, in the **Model transformations** list. To see the reverse transformation, click **Show Reverse Expression**. Transformations are displayed in SQL notation. Not all transformations have a reverse. Transformations and Reverse transformations are not always displayed.

Other Tabs: The **Settings** tab has these other tabs:

- GLMC Summary
- GLMC Target Values

13.5.4.1.10 GLMC Target Values Displays the target attributes, their data types, and the values of each target attribute.

Other Tabs: The **Settings** tab has these other tabs:

- GLMC Summary
- GLMC Inputs

13.5.4.2 GLM Classification Algorithm Settings

GLM supports these settings for classification:

- **Generate Row Diagnostics:** By default, Generate Row Diagnostic is deselected. To generate row diagnostics, you must select this option and also specify a Case ID.

  If you do not specify a Case ID, then this setting is not available.

  You can view Row Diagnostics on the **Diagnostics** tab in the model viewer. To further analyze row diagnostics, use a Model Details node to extract the row diagnostics table.

- **Confidence Level**: A positive number that is less than 1.0. This value indicates the degree of certainty that the true probability lies within the confidence bounds computed by the model. The default confidence is 0.95.

- **Reference Class name**: The Reference Target Class is the target value used as a reference in a binary Logistic Regression model. The probabilities are produced for the other (non-reference) classes. By default, the algorithm chooses the value with the highest prevalence (the most cases). If there are ties, then the attributes are sorted alpha-numerically in ascending order. The default for Reference Class name is System Determined, that is, the algorithm determines the value.

  See Also: "Choose Reference Value (GLMC)" for more information about how to select a specific value

- **Missing Values Treatment**: The default is Mean Mode, that is, use mean for numeric values and mode for categorical values. You can also select Delete Row to delete any row that contains missing values. If you delete rows with missing values, then the same missing values treatment (delete rows) must be applied to any data that the model is applied to.
- **Specify Row Weights Column**: By default, Row Weights Column is *not* specified. The Row Weights Column is a column in the training data that contains a weighting factor for the rows.

  Row weights can be used as a compact representation of repeated rows, as in the design of experiments where a specific configuration is repeated several times.

  Row weights can also be used to emphasize certain rows during model construction. For example, to bias the model toward rows that are more recent and away from potentially obsolete data.

  To specify a Row Weights column, click the check box and select the column from the list.
- **Ridge Regression**: By default, Ridge Regression is system determined (not disabled) in both Oracle Database 11*g* and 12*c*.

  Note: The Ridge Regression setting in both Oracle Database 11*g* and Oracle Database 12*c* should be consistent (system determined).

  If you select Ridge Regression, then Feature Selection is automatically deselected.

  Ridge Regression is a technique that compensates for multicollinearity (multivariate regression with correlated predictors). Oracle Data Mining supports Ridge Regression for both regression and classification mining functions.

  To specify options for Ridge Regression, click **Option** to open the Ridge Regression Option Dialog (GLMC).

  When Ridge Regression is enabled, fewer global details are returned. For example, when Ridge Regression is enabled, no prediction bounds are produced.

  Note: If you are connected to Oracle Database 11*g* Release 2 (11.2) and you get the error `ORA-40024` when you build a GLM model, enable Ridge Regression and rebuild the model.

- **Feature Selection:** By default, Feature Selection is deselected. This setting requires connection to Oracle Database 12*c*. To specify Feature Selection or view or specify Feature Selection settings, click **Option** to open the Feature Selection Option Dialog.

If you select Feature Selection, then Ridge Regression is automatically deselected.

> **Note:** The Feature Selection setting is available only in Oracle Database 12*c*.

- **Approximate Computation**: Specifies whether the algorithm should use approximate computations to improve performance. For GLM, approximation is appropriate for data sets that have many rows and are densely populated (not sparse).

  The values for Approximate Computation are:

  – **System Determined** (Default)

  – **Enable**

  – **Disable**

13.5.4.2.1 Feature Selection Option Dialog

The setting requires connection to Oracle Database 12*c*.

If you select Feature Selection, then Ridge Regression is automatically deselected. This dialog box enables you to specify Feature Selection for a GLMC or a GLMR model:

- **Feature Selection Criteria**: The default setting is system determined. You can select one of the following:

  – Akaike Information

  – Schwarz Bayesian Information

  – Risk Inflation

  – Alpha Investing

- **Max Number of Features**: The default setting is system determined.

  To specify several features, click the **User specified** option and enter an integer number of features.

- **Feature Identification**: The default setting is system determined.

  You can also choose:

  – Enable Sampling

  – Disable Sampling

- **Feature Acceptance**: The default setting is system determined.

  You can also choose:

  – Strict

  – Relaxed

- **Prune Model**: By default, Enable is selected. You can also select Disable.

- **Categorical Predictor Treatment**: By default, Add One at a Time is selected. You can also select Add All at Once.

  If you accept the default, that is Add One at a Time, then Feature Generation is *not* selected. If you select Feature Generation, then the default is Quadratic Candidates. You can also select Cubic Candidates.

13.5.4.2.2 Choose Reference Value (GLMC) To select a value, click **Edit**. In the Choose Reference Value dialog, select **Custom.** Then, select one of the values in the target values list. Click **OK**.

13.5.4.2.3 Ridge Regression Option Dialog (GLMC) You can use the system-determined **Ridge Value** or you can supply your own. By default, the system determined value is used.

Click **OK**.

13.5.4.3 GLM Regression Model Viewer

The GLM Regression (GLMR) model viewer displays characteristics of a GLMR model. GLMR is also known as Linear Regression.

To view a GLMR model, use one of these methods:

- Right-click the node where the where the model was built and select **Go to Properties** from the context menu. In the **Models** section of **Properties**, select the model and click .
- Select the workflow node where the model was built and right-click. Select **View Models** from the context menu and then select the model to view.

The default name of a GLM model has `GLM` in the name.

The GLMR viewer opens in a new tab.

The **Detail** tab is displayed by default.

The GLM Regression Model Viewer has these tabs:.

- **Details** described in GLMR Details
- **Coefficients** described in GLMR Coefficients
- **Diagnostics** described in GLMR Diagnostics (The default is to not generate diagnostics.)
- **Settings** described in GLMR Settings

13.5.4.3.1 GLMR Coefficients The **Coefficient** tab enables you to view GLM coefficients. The viewer supports sorting to control the order in which coefficients are displayed and filtering to select the coefficients to display.

> **See Also:** "GLMR Settings" for more information about confidence for GLM regression models

By default, coefficients are sorted by absolute value. You can deselect or select **Sort by absolute value** and click **Query**.

The default fetch size is 1,000 records. To change the fetch size, specify a new number of records and click **Query**.

> **Note:** After you change any criteria on this tab, click **Query** to query the database. You must click **Query** even for changes such as selecting or deselecting sort by absolute value or changing the fetch size.

Sort and Search GLMC Coefficients describes sorting and searching the grid.

The relative value of coefficients is shown graphically as a bar, with different colors for positive and negative values. If a coefficient is close to 0, then the bar may be too small to display.

- **Sort by absolute value**: Sort the list of coefficients by absolute value.
- **Fetch Size**: The number of rows displayed. To figure out if all coefficients are displayed, choose a fetch size that is greater than the number of rows displayed.

Coefficients are listed in a grid. If no items are listed, then there are no coefficients for that target value. The coefficients grid has these columns:

- **Attribute**: Name of the attribute
- **Value**: Value of the attribute
- **Coefficient**: The linear coefficient estimate for the selected target value is displayed. A bar is shown in front of (and possible overlapping) each coefficient. The bar indicates the relative size of the coefficient. For positive values, the bar is light blue; for negative values, the bar is red. (If a value is close to 0, then the bar may be too small to be displayed.)
- **Standard Error** of the estimate
- Wald Chi Squared
- Pr > Chi Square
- Upper coefficient limit
- Lower coefficient limit

Note: Not all statistics are necessarily returned for each coefficient.

Statistics are null if any of the following are true:

- The statistic does not apply to the mining function. For example, exp_coefficient does not apply to Linear Regression.
- The statistic cannot be calculated because of limitations in the system resources.
- The value of the statistics is infinity.
- If the model was built using Ridge Regression, or if the covariance matrix is found to be singular during the build, then coefficient bounds (upper and lower) have the value NULL.

Other Tabs: The viewer has these other tabs:

- GLMR Details
- GLMR Diagnostics
- GLMR Settings

13.5.4.3.2 GLMR Details The Model Details list global metrics for the model as a whole. The metrics display has two columns: Name of the metric and Value of the metric. The following metrics are displayed:

- Adjusted R-Square
- Akaike's information criterion
- Coefficient of variation

- Corrected total degrees of freedom
- Corrected total sum of squares
- Dependent mean
- Error degrees of freedom
- Error mean square
- Error sum of squares
- Model F value statistic
- Estimated mean square error
- Hocking Sp statistic
- JP statistic (the final prediction error)
- Model converged (Yes or No)
- Model degrees of freedom
- Model F value probability
- Model mean square
- Model sum of squares
- Number of parameters (the number of coefficients, including the intercept)
- Number of rows
- Root mean square error
- R-square
- Schwartz's Bayesian Information Criterion
- Termination
- Valid covariance matrix computed (Yes or No)

13.5.4.3.3 GLMR Diagnostics The **Diagnostics** tab displays diagnostics for each Case ID in the build data. You can filter the results.

> **Note:** Diagnostics are not generated by default. To generate diagnostics, you must and specify a Case ID and select **Generate Row Diagnostics**.

The following information is displayed in the Diagnostics grid:

- **CASE_ID**
- **TARGET_VALUE** for the row in the training data
- **PREDICTED_VALUE**, value predicted by the model for the target
- **HAT**, value of the diagonal element of the HAT matrix
- **RESIDUAL**, the residual with the adjusted dependent variable
- **STD_ERR_RESIDUAL**, Standard Error of the residual
- **STUDENTIZED_RESIDUAL**
- **PRED_RES**, predicted residual

- **COOKS_D**, Cook's D influence statistic

Other Tabs: The viewer has these other tabs:

- GLMR Details
- GLMR Coefficients
- GLMR Settings

13.5.4.3.4 GLMR Settings

The **Settings** tab has these tabs:

- GLMR Summary
- GLMR Inputs

Other Tabs: The viewer has these other tabs:

- GLMR Details
- GLMR Coefficients
- GLMR Diagnostics

13.5.4.3.5 GLMR Summary

General settings describe the characteristics of the model, including owner, name, type, algorithm, target attribute, creation date duration of model build, and comments.

> **See Also:** "General Settings"

Algorithm settings control model build; algorithm setting are specified when the build node is defined.

> **See Also:** "GLM Regression Algorithm Settings" for a list of algorithm settings

After a model is built, values calculated by the system are displayed on this tab. For example, if you select **System Determined** for **Enable Ridge Regression**, this tab shows if Ridge Regression was enabled and what ridge value was calculated.

Other Tabs: The **Settings** tab has this other tab:

- GLMR Inputs

13.5.4.3.6 GLMR Inputs

A list of the attributes used to build the model. For each attribute the following information is displayed:

- **Name**: The name of the attribute.
- **Data type** The data type of the attribute
- **Mining type**: Categorical or numerical
- **Target**: A check mark indicates that the attribute is a target attribute.
- **Data Prep**: `YES` indicates that data preparation was performed.

When you select an attribute in the **Attributes** list, the transformation properties viewer displays the imbedded transformation, created by either the user or Automatic Data Preparation, in the **Model transformations** list. To see the reverse transformation, click **Show Reverse Expression**. Transformations are displayed in SQL notation. Not all transformations have a reverse. Transformations and Reverse transformations are not always displayed.

Other Tabs: The **Settings** tab has this other tab:

- GLMR Summary

13.5.4.4 GLM Regression Algorithm Settings

GLM supports these settings for regression:

- **Generate Row Diagnostics** is set to `OFF` by default. To generate row diagnostics, you must select this option and also specify a Case ID.

  If you do not specify a Case ID, then this setting is not available.

  You can view **Row Diagnostics** on the **Diagnostics** tab when you view the model. To further analyze row diagnostics, use a Model Details node to extract the row diagnostics table.

- **Confidence Level**: A positive number that is less than 1.0. This level indicates the degree of certainty that the true probability lies within the confidence bounds computed by the model. The default confidence is 0.95.

- **Missing Values Treatment**: The default is Mean Mode.That is, use Mean for numeric values and Mode for categorical values.
  You can also select **Delete Row** to delete any row that contains missing values. If you delete rows with missing values, then the same missing values treatment (delete rows) must be applied to any data that the model is applied to.

- **Specify Row Weights Column**: The Row Weights Column is a column in the training data that contains a weighting factor for the rows. By default, Row Weights Column is *not* specified. Row weights can be used:

  – As a compact representation of repeated rows, as in the design of experiments where a specific configuration is repeated several times.

  – To emphasize certain rows during model construction. For example, to bias the model toward rows that are more recent and away from potentially obsolete data

- **Ridge Regression**: Ridge Regression is a technique that compensates for multicollinearity (multivariate regression with correlated predictors). Oracle Data Mining supports Ridge Regression for both regression and classification mining functions.

  By default, Ridge Regression is system determined (not disabled) in both Oracle Database 11*g* and Oracle Database 12*c*. If you select Ridge Regression, then Feature Selection is automatically deselected.

  To specify options for Ridge Regression, click **Option** to open the Ridge Regression Option Dialog (GLMR).

  When Ridge Regression is enabled, fewer global details are returned. For example, when Ridge Regression is enabled, no prediction bounds are produced.

  > **Note:** If you are connected to Oracle Database 11*g* Release 2 (11.2) and you get the error `ORA-40024` when you build a GLM model, enable Ridge Regression and rebuild the model.

- **Feature Selection:** This setting requires connection to Oracle Database 12*c*. By default, Feature Selection is deselected. To specify Feature Selection or view or specify Feature Selection settings, click **Option** to open the Feature Selection Option Dialog.

  If you select Feature Selection, then Ridge Regression is automatically deselected.

Note: The Feature Selection setting is available only in Oracle Database 12*c*.

- **Approximate Computation**: Specifies whether the algorithm should use approximate computations to improve performance. For GLM, approximation is appropriate for data sets that have many rows and are densely populated (not sparse).

  Values for Approximate Computation are:

  – **System Determined** (Default)

  – **Enable**

  – **Disable**

13.5.4.4.1 Ridge Regression Option Dialog (GLMR) You can use the System Determined Ridge Value or you can supply your own. By default, the System Determined value is used. **Produce Variance Inflation Factor (VIF)** is not selected by default. You can select it.

Click **OK**.

13.5.4.4.2 Choose Reference Value (GLMR) To select a value:

1. Click **Edit.**
2. In the **Choose Reference** dialog box, click **Custom.**
3. Select one of the values in the **Target Values** field.
4. Click **OK.**

13.6 k-Means

The *k*-Means (KM) algorithm is a distance-based clustering algorithm that partitions the data into a predetermined number of clusters, provided there are enough distinct cases.

Distance-based algorithms rely on a distance metric (function) to measure the similarity between data points. The distance metric is either Euclidean, Cosine, or Fast Cosine distance. Data points are assigned to the nearest cluster according to the distance metric used.

Use a Clustering Node to build KM models.

Use an Apply Node to apply a KM model to new data.

The following topics describe KM models:

- k-Means Algorithm
- KM Model Viewer and Algorithm Settings

13.6.1 k-Means Algorithm

Oracle Data Mining implements an enhanced version of the *k*-Means algorithm with the following features:

- The algorithm builds models in a hierarchical manner. The algorithm builds a model top down using binary splits and refinement of all nodes at the end. In this sense, the algorithm is similar to the bisecting *k*-Means algorithm. The centroid of the inner nodes in the hierarchy are updated to reflect changes as the tree evolves. The whole tree is returned.

- The algorithm grows the tree, one node at a time (unbalanced approach). Based on a user setting, the node with the largest variance is split to increase the size of the tree until the desired number of clusters is reached. The maximum number of clusters is specified as a build setting.
- The algorithm provides probabilistic scoring and assignment of data to clusters.
- The algorithm returns the following, for each cluster:
  - A centroid (cluster prototype). The centroid reports the mode for categorical attributes, or the mean and variance for numerical attributes.
  - Histograms (one for each attribute),
  - A rule describing the hyper box that encloses the majority of the data assigned to the cluster.

The clusters discovered by enhanced *k*-Means are used to generate a Bayesian probability model that is then used during scoring (model apply) for assigning data points to clusters. The *k*-Means algorithm can be interpreted as a mixture model where the mixture components are spherical multivariate normal distributions with the same variance for all components.

> **Note:** The *k*-Means algorithm samples one million rows. You can use the sample to build the model.

13.6.2 KM Model Viewer and Algorithm Settings

This section discusses the *k*-Means (KM) Model viewer, the procedure to view the KM Model viewer and the algorithm settings related to KM. Topics are:

- KM Model Viewer
- KM Algorithm Settings

13.6.2.1 KM Model Viewer

The KM Model Viewer lets you examine a KM model. You can view a KM model in one of the following ways

- Method 1
  1. Right-click the node where the model was built.
  2. Select **Go to Properties.**
  3. In the **Models** section in **Properties**, click .
- Method 2
  1. Select the workflow node where the model was built.
  2. Right-click and click **View Models.**
  3. Select the model to view.

  The KM model viewer opens in a new tab. The default name of a *k*-Means model has `KM` in the name. The **Tree** tab is displayed by default.

The KM model viewer has these tabs:

- EM, KM, and OC Tree
- Cluster (Viewer)
- EM, KM, and OC Compare
- KM Settings

13.6.2.1.1 EM, KM, and OC Tree The tree viewer for EM, KM, and OC operate in the same way.

The Tree Viewer is the graphical tree for hierarchical clusters. When you view the tree:

- The Workflow Thumbnail opens to give you a view of the entire tree.
- The Structure Window helps you navigate and analyze the tree.

You can compare the attributes in a given node with the attributes in the population using EM, KM, and OC Compare.

Viewing Information:

To view information about a particular node:

1. Select the node.
2. In the lower pane, the following are displayed in each of these tabs:
   - **Centroid:** Displays the centroid of the cluster
   - **Cluster Rule:** Displays the rule that all elements of the cluster satisfy.

Display Control:

The following control the display of the tree as a whole:

- **Zoom in:** Zooms in to the diagram, providing a more detailed view of the rule.
- **Zoom out:** Zooms out from the diagram, providing a view of much or all of the rule.
- **Percent size:** Enables you select an exact percentage to zoom the view.
- **Fit to Window:** Zooms out from the diagram until the whole diagram fits within the screen.
- **Layout Type:** Enables you to select horizontal layout or vertical layout; the default is vertical.
- **Expand:** All nodes shows branches of the tree.
- **Show more detail:** Shows more data for each tree node. Click again to show less detail.
- **Top Attributes:** Displays the top N attributes. N is 5 by default. To change N, select a different number from the list.
- **Refresh:** Enables you to apply the changed **Query Settings**.
- **Query Settings:** Enables you to change the number of top settings. The default is 10. You can save a different number as the new default.
- Save Rules.

13.6.2.1.2 Cluster (Viewer) The **Cluster** tab for EM, KM, and OC operate in the same way.

The **Cluster** tab enables you to view information about a selected cluster. The viewer supports filtering so that only selected probabilities are displayed.

The following information is displayed:

- **Cluster**: The ID of the cluster being viewed. To view another cluster, select a different ID from the menu. You can view **Leaves Only** (terminal clusters) by selecting Leaves Only. **Leaves Only** is the default.
- **Fetch Size:** Default is 20. You can change this value.

  If you change Fetch Size, click **Query** to see the new display.

The grid lists the attributes in the cluster. For each attribute, the following information is displayed:

- **Name** of the attribute.
- **Histogram** of the attribute values in the cluster.
- **Confidence** displayed as both a number and with a bar indicating a percentage. If confidence is very small, then no bar is displayed.
- **Support**, the number of cases.
- **Mean**, displayed for numeric attributes.
- **Mode**, displayed for categorical attributes.
- **Variance**

To view a larger version of the histogram, select an attribute; the histogram is displayed below the grid. Place the cursor over a bar in the histogram to see the details of the histogram including the exact value.

You can search the attribute list for a specific attribute name or for a specific value of mode. To search, use the search box.

The drop-down list enables you to search the grid by **Attribute** (the default) or by **Mode.** Type the search term in the box next to .

To clear a search, click .

Other Tabs: The NB Model Viewer has these other tabs:

- EM, KM, and OC Tree
- EM, KM, and OC Compare
- KM Settings

13.6.2.1.3 EM, KM, and OC Compare The **Compare** tab for EM, KM, and OC operate in the same way. The **Compare** tab enables you to compare two clusters in the same model. The display enables you to select the two clusters to compare.

You can perform the following tasks:

- **Compare Clusters:** To select clusters to compare, pick them from the lists. The clusters are compared by comparing attribute values. The comparison is displayed in a grid. You can use **Compare** to compare an individual cluster with the population.

  See Also: "Compare Cluster with Population"

- **Rename Clusters:** To rename clusters, click **Edit.** This opens the Rename Cluster dialog box. By default, only **Show Leaves** is displayed. To show all nodes, then deselect **Leaves Only.** The default Fetch Size is 20. You can change this value.

- **Search Attribute:** To search an attribute, enter its name in the search box. You can also search by rank.
- **Create Query:** If you make any changes, click **Query.**

For each cluster, a histogram is generated that shows the attribute values in the cluster. To see enlarged histograms for a cluster, click the attribute that you are interested in. The enlarged histograms are displayed below the attribute grid.

In some cases, there may appear to be Missing Histograms in a Cluster.

13.6.2.1.4 Compare Cluster with Population

To see how an individual cluster compares with the population:

1. Click **Compare**.
2. Deselect **Leaves Only**.
3. Select the root node as Cluster 1. This is cluster **1**, if the clusters have not been renamed. The distribution of attribute values in Cluster 1 represents the distribution of values in the population as a whole. Select the cluster that you want to compare with the population as Cluster 2.
4. You can now compare the distribution of values for each attribute in the cluster selected as Cluster 2 with the values in Cluster 1.

13.6.2.1.5 Missing Histograms in a Cluster

If clusters are built using sparse data, then some attribute values are not present in the records assigned to a cluster.

In this case, a cluster comparison shows the centroid and histogram values for the cluster where the attribute is present and leaves blanks for the cluster where the attribute is present.

13.6.2.1.6 Rename Cluster

The title bar of the dialog box shows the cluster to rename. Cluster ID is a number. You can change it to a string. Enter in the new name and click **OK**.

> **Note:** Two different clusters cannot have the same name.

13.6.2.1.7 KM Settings

The **Settings** tab displays information about how the model was built:

- Cluster Model Summary
- Cluster Model Input on a separate tab

Other Tabs: The KM Model Viewer has these other tabs:

- EM, KM, and OC Tree
- Cluster (Viewer)
- EM, KM, and OC Compare

13.6.2.1.8 Cluster Model Settings (Viewer)

The **Settings** tab of the model viewer contains two tabs: Cluster Model Summary and Cluster Model Input.

13.6.2.1.9 Cluster Model Summary

The **Summary** tab lists the following:

- **General** settings lists the following information:
  – Type of Model (Classification, Regression, and so on)

– Owner of the model (the Schema where the model was built)
– Model Name
– Creation Date
– Duration of the Model Build (in minutes)
– Size of the model (in MB)
– Comments

See Also: "General Settings"

- **Algorithm** settings list the algorithm and algorithm settings used to build the model.

See Also:

- "EM Algorithm Settings"
- "KM Algorithm Settings"
- "OC Algorithm Settings"

13.6.2.1.10 Cluster Model Input The **Input** tab is displayed for models that can be scored only. A list of the attributes used to build the model. For each attribute the following information is displayed:

- **Name**: The name of the attribute.
- **Data Type:** The data type of the attribute.
- **Mining type**: Categorical or numerical.
- **Data Prep**: `YES` indicates that data preparation was performed.

When you select an attribute in the **Attributes** list, the transformation properties viewer displays the imbedded transformation, created by either the user or Automatic Data Preparation, in the **Model transformations** list.

To see the reverse transformation, click **Show Reverse Expression**. Transformations are displayed in SQL notation. Not all transformations have a reverse. Transformations and Reverse transformations are not always displayed.

13.6.2.2 KM Algorithm Settings

The *k*-Means (KM) algorithm supports these settings:

- **Number of Clusters** is the maximum number of leaf clusters generated by the algorithm. The default is 10. *k*-Means usually produces the exact number of clusters specified, unless there are fewer distinct data points.
- **Growth Factor** is a number greater than 1 and less than or equal to 5. This value specifies the growth factor for memory allocated to hold cluster data. Default is 2.
- **Convergence Tolerance** must be between 0.001 (slow build) and 0.1 (fast build). The default is 0.01. The tolerance controls the convergence of the algorithm. The smaller the value, the closest to the optimal solution at the cost of longer run times. This parameter interacts with the number of iterations parameter.
- **Distance Function** specifies how the algorithm calculates distance. The default distance function is Euclidean.The other distance functions are:
  – Cosine

  - Fast Cosine
- **Number of Iterations** must be greater than or equal to 1. The default is 30. This value is the maximum number of iterations for the *k*-Means algorithm. In general, more iterations result in a slower build. However, the algorithm may reach the maximum, or it may converge early. The convergence is determined by whether the **Convergence Tolerance** setting is satisfied.
- **Min Percent Attribute Support** is not an integer. The range of the value for Min Percent Attribute Support is:
  - Greater than or equal to 0, and
  - Less than or equal to 1.

    The default value is 0.1. The default value enables you to highlight the more important predicates instead producing a long list of predicates that have very low support.

  You can use this value to filter out rule predicates that do not meet the support threshold. Setting this value too high can result in very short or even empty rules.

  In extreme cases, for very sparse data, all attribute predicates may be filtered out so that no rule is produced. If no rule is produced, then you can lower the support threshold and rebuild the model to make the algorithm produce rules even if the predicate support is very low.
- **Number of Histogram Bins** is a positive integer; the default value is 10. This value specifies the number of bins in the attribute histogram produced by *k*-Means. The bin boundaries for each attribute are computed globally on the entire training data set. The binning method is equiwidth. All attributes have the same number of bins except attributes with a single value that have only one bin.
- **Split Criterion** is either Variance or Size. The default is `Variance.` The split criterion is related to the initialization of the *k*-Means clusters. The algorithm builds a binary tree and adds one new cluster at a time. Size results in placing the new cluster in the area where the largest current cluster is located. Variance places the new cluster in the area of the most spread out cluster.

13.7 Naive Bayes

The Naive Bayes (NB) algorithm is used to build Classification models. You can build, test, apply, and tune a Naive Bayes model.

- To build an NB model, use a Classification Node. By default, a Classification Node tests all models that it builds. The test data is created by splitting the input data into build and test subsets.
- To test an NB model, you can also use a Test Node.

  See Also: "Testing Classification Models"

- To apply an NB model to new data, use an Apply Node.
- To tune an NB model, see "Tuning Classification Models" for more information. After building and testing an NB model, you can tune the NB model.

The following topics describe Naive Bayes:

- Naive Bayes Algorithm
- Naive Bayes Viewers and Algorithm Settings

13.7.1 Naive Bayes Algorithm

The Naive Bayes (NB) algorithm is based on conditional probabilities. It uses Bayes' Theorem, which calculates a probability by counting the frequency of values and combinations of values in the historical data.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred.

Assumption:

Naive Bayes makes the assumption that each predictor is conditionally independent of the others. For a given target value, the distribution of each predictor is independent of the other predictors. In practice, this assumption of independence, even when violated, does not degrade the model's predictive accuracy significantly, and makes the difference between a fast, computationally feasible algorithm and an intractable one.

Sometimes, the distribution of a given predictor is clearly not representative of the larger population. For example, there might be only a few customers under 21 in the training data, but in fact, there are many customers in this age group in the wider customer base. To compensate, you can specify prior probabilities (priors) when training the model.

There are several Advantages of Naive Bayes.

13.7.1.1 Advantages of Naive Bayes

The advantages of Naive Bayes model are:

- The Naive Bayes algorithm provides fast, highly scalable model building and scoring. It scales linearly with the number of predictors and rows.
- The Naive Bayes build process is parallelized. Scoring can also be parallelized irrespective of the algorithm.
- Naive Bayes can be used for both binary and multiclass classification problem

13.7.2 Naive Bayes Viewers and Algorithm Settings

This section discusses the Naive Bayes (NB) Model viewer, the procedure to view the NB Model viewer and the algorithm settings related to NB model. Topics are:

- Naive Bayes Model Viewer
- Naive Bayes Test Viewer
- Summary (NB)

13.7.2.1 Naive Bayes Model Viewer

The NB Model Viewer lets you examine an NB model. You can view an NB model using any one of the following methods.

- Method 1
  1. Right-click the node where the model was built.
  2. Select **Go to Properties.**
  3. In the **Models** section in **Properties**, click .
- Method 2
  1. Select the workflow node where the model was built.

2. Right-click and click **View Models.**
3. Select the model to view.

   The model viewer opens in a new tab. The default name of a Naive Bayes model has `NB` in the name.

The NB model viewer has these tabs:

- Probabilities (NB)
- Compare (NB)
- Settings (NB)

13.7.2.1.1 Probabilities (NB) The **Probabilities** tab lists the probabilities calculated during model build. You can sort and filter the order in which probabilities are displayed.

The relative value of probabilities is shown graphically as a bar, with a blue bar for positive values and red bar for negative values. For numbers close to zero, the bar may be too small to be displayed.

Select **Target Value.** The probabilities associated with the selected value are displayed. The default is to display the probabilities for the value that occurs least frequently.

Probabilities are listed in the Grid.

Other Tabs: The NB Model Viewer has these other tabs:

- Compare (NB)
- Settings (NB)

13.7.2.1.2 Grid If no items are listed, then there are no values that satisfy the criteria that you specified:

- **Row Count**: The number of rows displayed.
- **Grid Filter**: Use the Grid Filter to filter the information in the grid.

The probabilities grid has these columns:

- **Attribute**: Name of the attribute
- **Value**: Value of the attribute
- **Probability**: The probability for the value of the attribute. Probability is displayed as both a number and with a bar indicating a percentage. If the probability is very small, then no bar is displayed.

13.7.2.1.3 Fetch Size This value limits the number of rows returned regardless of Filter or Server settings. The default fetch size is 1000. Change the Fetch Size by clicking the up or down arrows. If you change the Fetch Size, click **Query**.

13.7.2.1.4 Grid Filter The filter control  enables you to filter the items that are displayed in the grid. The filtering is done as you type in the filter search box.

To see the filter categories, click the down arrow next to the binoculars icon. The following categories are supported for probabilities:

- **Attribute:** Filters the Attribute (name) column. This is the default category. For example, to display all entries with CUST in the attribute name, enter CUST in the search box.
- **Value:** Filters the value column.

- **Probability:** Filters the probability column.
- **All (And)**: Enter in one or more strings and their values are compared against the Attribute and Value columns using the AND condition. For example, enter `CUST M` to display rows where the attribute name contains `CUST` and the value is `M`.
- **All (Or)**: Works the same as **All (And)** except that the comparison uses an OR condition.

The Grid Filter for Compare lists similar categories:

- **Name:** Filters by attribute name (Default).
- **Value:** Filters the value column.
- **Attribute/Value/Propensity (or):** Filters for values in any of the attribute, value, and propensity columns.
- **Attribute/Value/Propensity (and):** Filters for values in any of the attribute, value, and propensity columns.
- **Propensity for Target Value 1:** Filters the propensity values for Target Value 1.
- **Propensity for Target Value 2:** Filters the propensity values for Target Value 2.

After you enter one or more strings into the filter search box, ✖ is displayed. Click this icon to clear the search string.

13.7.2.1.5 Compare (NB) The **Compare** tab enables you to compare results for two different target values. Select the two target values.

The default values for Target Value 1 and Target Value 2 are displayed. You can do the following:

- Change the Target Values. The Target Values that you select must be different.
- Use the Grid Filter to display specific values.
- Change the Fetch Size.
- Sort the grid columns. The grid for compare has these columns:
  - **Attribute**: Name of the attribute
  - **Value**: Value of the attribute
  - Propensity for target value 1
  - Propensity for target value 2

  For both propensities, a histogram bar is displayed. The maximum value of propensity is 1.0. The minimum is -1.0.

  Propensity shows which of the two target values has a more predictive relationship for a given attribute value pair. Propensity can be measured in terms of being predicted for or against a target value, where prediction against is shown as a negative value.

Other Tabs: The NB Model Viewer has these other tabs:

- Probabilities (NB)
- Settings (NB)

13.7.2.1.6 Settings (NB) The **Settings** tab displays information about how the model was built:

- Settings (NB)

- Input (NB)
- Target Values (NB)
- Weights (NB), displayed for tuned models only

Other Tabs: The NB Model Viewer has these other tabs:

- Compare (NB)
- Probabilities (NB)

13.7.2.1.7 Settings (NB)

The **Settings** tab shows information about the model.

The **Settings** tab has these tabs:

- Summary (NB)
- Input (NB)
- Weights (NB)
- Target Values (NB)

13.7.2.1.8 Summary (NB)

The **Summary** tab describes all model settings. Model settings describe characteristics of model building. The Settings are divided into:

- General Settings
- Naive Bayes Algorithm Settings

13.7.2.1.9 Naive Bayes Algorithm Settings

This section identifies the algorithm and whether Automatic Data Preparation (ADP) is `ON` or `OFF`.

These settings are specific to Naive Bayes:

- **Pair wise Threshold**: The minimum percentage of pair wise occurrences required for including a predictor in the model. The default is 0.
- **Singleton Threshold**: The minimum percentage of singleton occurrences required for including a predictor in the model. The default is 0.

13.7.2.1.10 Input (NB)

The **Input** tab is displayed for models that can be scored only.

A list of the attributes used to build the model. For each attribute the following information is displayed:

- **Name**: The name of the attribute.
- **Data Type:** The data type of the attribute.
- **Mining Type**: Categorical or Numerical.
- **Target**: The  icon indicates that the attribute is a target attribute.
- **Data Prep**: `YES` indicates that data preparation was performed.

When you select an attribute in the **Attributes** list, the transformation properties viewer displays the imbedded transformation, created by either the user or Automatic Data Preparation, in the **Model transformations** list.

To see the reverse transformation, click **Show Reverse Expression**. Transformations are displayed in SQL notation. Not all transformations have a reverse. Transformations and Reverse transformations are not always displayed.

13.7.2.1.11 Weights (NB)

Weights that are calculated by the system for each target value are displayed on the **Weights** tab. If you tune the model, the weights may change.

13.7.2.1.12 Target Values (NB) Displays the following:

- Target Attributes
- Data Types
- Values of each Target Attributes

13.7.2.2 Naive Bayes Test Viewer

By default, any Classification or Regression model is automatically tested. A Classification model is tested by comparing the predictions of the model with known results. Oracle Data Miner keeps the latest test result.

To view the test results for a model, right-click the Build node and select **View Results**.

> **See Also:** "Testing Classification Models" for more information about Test Viewers

13.8 Nonnegative Matrix Factorization

Nonnegative Matrix Factorization (NMF) is the unsupervised algorithm used by Oracle Data Mining for feature extraction.

- To build an NMF model, use a Feature Extraction Node.
- To apply an NMF model to new data, use an Apply Node.

These topics describe NMF:

- Using Nonnegative Matrix Factorization
- How Does Nonnegative Matrix Factorization Work
- NMF Model Viewer and Algorithm Settings

13.8.1 Using Nonnegative Matrix Factorization

Nonnegative Matrix Factorization (NMF) is useful when there are many attributes and the attributes are ambiguous or have weak predictability. By combining attributes, NMF can produce meaningful patterns, topics, or themes.

NMF is especially well-suited for text mining. In a text document, the same word can occur in different places with different meanings. For example, *hike* can be applied to the outdoors or to interest rates. By combining attributes, NMF introduces context, which is essential for predictive power:

```
"hike" + "mountain" -> "outdoor sports"
"hike" + "interest" -> "interest rates"
```

13.8.2 How Does Nonnegative Matrix Factorization Work

Non-Negative Matrix Factorization (NMF) uses techniques from multivariate analysis and linear algebra. NMF decomposes multivariate data by creating a user-defined number of features. Each feature is a linear combination of the original attribute set. The coefficients of these linear combinations are nonnegative.

NMF decomposes a data matrix V into the product of two lower rank matrices W and H so that V is approximately equal to W times H. NMF uses an iterative procedure to modify the initial values of W and H so that the product approaches V. The procedure terminates when the approximation error converges or the specified number of iterations is reached.

When applying to a model, an NMF model maps the original data into the new set of attributes (features) discovered by the model.

13.8.3 NMF Model Viewer and Algorithm Settings

This section discusses the Nonnegative Matrix Factorization (NMF) Model viewer, the procedure to view the NMF Model viewer and the algorithm settings related to NMF. The topics are:

- NMF Model Viewer
- NMF Algorithm Settings

13.8.3.1 NMF Model Viewer

View an NMF model in one of the following ways:

- Method 1
  1. Right-click the node where the model was built.
  2. Select **Go to Properties.**
  3. In the **Models** section in **Properties**, click .
- Method 2
  1. Select the workflow node where the model was built.
  2. Right-click and click **View Models.**
  3. Select the model to view. The model viewer opens in a new tab. The **Settings** tab is displayed by default.

The NMF Model Viewer has these tabs:

- Coefficients (NMF)
- Settings (NMF)

13.8.3.1.1 Coefficients (NMF)

For a given Feature ID, the coefficients are displayed in the Coefficients grid. The title of the grid **Coefficients *x* of *y*** displays the number of rows returned out of all the rows available in the model.

By default, Feature IDs are integers.

Fetch Size limits the number of rows returned. The default is 1000 or the value specified in the **Preference** settings for Model Viewers.

You can perform the following tasks:

- Rename (NMF)
- Filter (NMF)

The **Coefficients** grid has these columns:

- **Attribute**, attribute name
- **Value**, value of attribute
- **Coefficient**. The value is shown as a bar with the value centered in the bar. Positive values are light blue. Negative values are red.

13.8.3.1.2 Rename (NMF)

You can rename the selected Feature ID.

1. Enter in the new name in the **Feature ID** field.

2. Click **OK.**

> **Note:** Different features should have different names.

13.8.3.1.3 Filter (NMF) To view the filter categories, click .

The filter categories are

- **Attribute** (Default): Search for an attribute name.
- **Value:** This is the value column.
- **Coefficient:** This is the coefficient column

To create a filter, enter a string in the text box. After a string has been entered,  icon is displayed. To clear the filter, click the icon.

13.8.3.1.4 Settings (NMF) The **Settings** tab has these tabs:

- Summary (NMF)
- Inputs (NMF)

13.8.3.1.5 Summary (NMF)

- **General** Settings lists the following:
  - Type of model (Classification, Regression, and so on)
  - Owner of the model (the Schema where the model was built)
  - Model Name
  - Creation Date
  - Duration of the model build (in minutes)
  - Size of the model (in MB)
  - Comments

  See Also: "General Settings"

- **Algorithm** Settings lists the following:
  - The name of the algorithm used to build the model.
  - The algorithm settings that control the model build.

  See Also: "NMF Algorithm Settings"

13.8.3.1.6 Inputs (NMF) This tabs shows information about those attributes used to build the model.

Oracle Data Miner does not necessarily use all of the attributes in the build data. For example, if the values of an attribute are constant, then that attribute is not used.

For each attribute used to build the model, this tab displays:

- **Name**: The name of the attribute.
- **Data Type:** The data type of the attribute
- **Mining Type**: Categorical or Numerical

- **Data Prep**: YES indicates that data preparation was performed. If Data Prep is YES, then select the column (click it). The data preparation is displayed in Data Preparation at the bottom of the tab.
  To see the reverse transformation for data preparation, select **Show Reverse Transformation.** Transformations are displayed in SQL notation. Not all transformations have a reverse. Transformations and Reverse transformations are not always displayed.

13.8.3.2 NMF Algorithm Settings

The Nonnegative Matrix Factorization (NMF) algorithm supports these settings:

- **Convergence Tolerance:** Indicates the minimum convergence tolerance value. The default is 0.5.
- **Automatic preparation:** ON (Default). Indicates automatic data preparation.
- **NMFS_NONNEGATIVE_SCORING:** Enabled or Disabled. The default is Enabled (NMFS_NONNEG_SCORING_ENABLE).
- **Number of features**: The default is to *not* specify the number of features. If you do not specify the number of features, then the algorithm determines the number of features.

  To specify the number of features, then select **Specify number of features** and enter the integer number of features. The number of features must be a positive integer less than or equal to the minimum of the number of attributes and to the number of cases. In many cases, 5 or some other number less than or equal to 7 gives good results.
- **Number of iterations:** Indicates the maximum number of iterations to be performed. The default is 50.
- **Random Seed:** It is the random seed for the sample. The default value is -1.The seed can be changed. If you plan to repeat this operation to get the same results, ensure to use the same random seed.

13.9 Orthogonal Partitioning Clustering

Orthogonal Partitioning Clustering (O-Cluster) is a clustering algorithm that is proprietary to Oracle. The requirements to build and apply the O-Cluster algorithm:

- To build OC models, use a Clustering Node.
- To apply an OC model to new data, use an Apply Node.

The following topics describe O-Cluster:

- O-Cluster Algorithm
- OC Model Viewer and Algorithm Settings

13.9.1 O-Cluster Algorithm

The O-Cluster (OC) algorithm creates a hierarchical grid-based clustering model. That is, it creates axis-parallel (orthogonal) partitions in the input attribute space. The algorithm operates recursively. The resulting hierarchical structure represents an irregular grid that tessellates the attribute space into clusters. The resulting clusters define dense areas in the attribute space.

The clusters are described by intervals along the attribute axes and the corresponding centroids and histograms. The sensitivity parameter defines a baseline density level. Only areas with a peak density above this baseline level can be identified as clusters.

The clusters discovered by O-Cluster are used to generate a Bayesian probability model that is then used during scoring (model apply) for assigning data points to clusters. The generated probability model is a mixture model where the mixture components are represented by a product of independent normal distributions for numerical attributes and multinomial distributions for categorical attributes.

O-Cluster goes through the data in chunks until it converges. There is no explicit limit on the number of rows processed.

O-Cluster handles missing values naturally as missing at random. The algorithm does not support nested tables and thus does not support sparse data.

Note: OC does not support text.

13.9.2 OC Model Viewer and Algorithm Settings

This section discusses the O-Cluster Model viewer, the procedure to view the OC Model viewer and the algorithm settings related to OC. The topics are:

- OC Model Viewer
- OC Algorithm Settings

13.9.2.1 OC Model Viewer

The OC Model Viewer lets you examine an OC model. You can view an OC model using one of the following methods:

- Method 1
  1. Right-click the node where the model was built.
  2. Select **Go to Properties.**
  3. In the **Models** section in **Properties**, click .
- Method 2
  1. Select the workflow node where the model was built.
  2. Right-click and click **View Models.**
  3. Select the model to view. The OC model viewer opens in a new tab. The **Tree** tab is displayed by default. The default name of an O-Cluster model has `OC` in the name.

The OC Model viewer has these tabs:

- EM, KM, and OC Tree (The tree display for O-Cluster is the same as the tree display for KM).
- Cluster (Viewer) (The detail display for O-Cluster is the same as the detail display for KM).
- EM, KM, and OC Compare (The compare display for O-Cluster is the same as the compare display for KM).
- Settings (OC)

13.9.2.1.1 Detail (OC) The **Details** tab enables you to view details for a cluster. You can discover what values of an attribute are in the selected cluster. The viewer supports filtering so that only selected probabilities are displayed.

The following information is displayed:

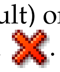
- **Cluster**: The ID of the cluster being viewed. You can change the cluster by selecting a different ID. Select **Leaves Only** to view terminal clusters only.
- **Fetch Size**: The number of columns selected. The default is 50. You can change the Fetch Size. If you change the Fetch Size, click **Query**.

The grid lists the attributes in the cluster. For each attribute, the following information is displayed:

- **Attribute:** An attribute is a predictor in a predictive model or an item of descriptive information in a descriptive model. Data attributes are the columns of data that are used to build a model. Data attributes undergo transformations so that they can be used as categoricals or numericals by the model. Categoricals and numericals are model attributes.
- **Histogram:** The attribute values in the selected cluster are displayed as a histogram.

  To view a larger version of the histogram, select an attribute. The histogram is displayed below the grid. Place the cursor over a bar in the histogram to see the details of the histogram including the exact value.
- **Confidence:** Displayed as both a number and with a bar indicating a percentage. If the confidence is very small, then no bar is displayed.
- **Support:** The number of cases.
- **Mean:** Displayed for numeric attributes.
- **Mode:** Displayed for categorical attributes.
- **Variance**

You can perform the following tasks:

- Sort the attributes in the cluster. To sort, click the appropriate column heading in the grid. For example, to sort by attribute name, click **Attribute**. The attributes are sorted by:
  - Confidence
  - Support
  - Mean
  - Mode
  - Variance
  - Attribute name
- Search the attribute list for a specific attribute name or for a specific value of mode. To search, use the search box next to .
- Search the grid by Attribute. The drop down list enables you to search the grid by Attribute (the default) or by Mode. Enter the search term in the search field. To clear a search, click .

Other Tabs: The OC Model Viewer has this other tab:

- Settings (OC)

13.9.2.1.2 Settings (OC) The **Settings** tab displays information about how the model was built:

- Summary (OC) is on the **Model Settings** tab.
- Inputs (OC) is on a separate tab.
- Detail (OC). This is an additional tab.

13.9.2.1.3 Summary (OC)

- **General** settings lists the following:
  - Type of model (Classification, Regression, and so on)
  - Owner of the model (the Schema where the model was built)
  - Model Name
  - Creation Date
  - Duration of the model build (in minutes)
  - Size of the model (in MB)
  - Comments

  See Also: "General Settings"

- **Algorithm** settings lists the following:
  - The name of the algorithm.
  - The settings that control the model build. Algorithm settings are specified when the build node is defined.

  See Also: "OC Algorithm Settings"

13.9.2.1.4 Inputs (OC) The **Inputs** tab is displayed for models that can be scored only.

When you select an attribute in the **Attributes** list, the transformation properties viewer displays the imbedded transformation, created by either the user or Automatic Data Preparation, in the **Model transformations** list.

To see the reverse transformation, click **Show Reverse Expression**. Transformations are displayed in SQL notation. Not all transformations have a reverse. Transformations and Reverse transformations are not always displayed.

13.9.2.2 OC Algorithm Settings

The O-Cluster (OC) algorithm supports these settings:

- **Number of Clusters:** It is the maximum number of leaf clusters generated by the algorithm. The default is 10.
- **Buffer Size:** It is the maximum size of the memory buffer, in logical records, that can be used by the algorithm. The default is 50,000 logical records.
- **Sensitivity:** It is a number between 0 (fewer clusters) and 1 (more clusters). The default is 0.5.This value specifies the peak density required for separating a new cluster. This value is related to the global uniform density.

13.10 Singular Value Decomposition and Principal Components Analysis

Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) are unsupervised algorithms used by Oracle Data Mining for feature extraction.

Unlike Nonnegative Matrix Factorization, SVD and PCA are orthogonal linear transformations that are optimal for capturing the underlying variance of the data. This property is extremely useful for reducing the dimensionality of high-dimensional data and for supporting meaningful data visualization.

> **Note:** Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) require Oracle Database 12*c*.

In addition to dimensionality reduction, SVD and PCA have several other important applications, such as, data denoising (smoothing), data compression, matrix inversion, and solving a system of linear equations. All these areas can be effectively supported by the Oracle Data Mining implementation SVD/PCA.

SVD is implemented as a feature extraction algorithm. PCA is implemented as a special scoring method for the SVD algorithm.

> **See Also:** For more information about SVD and PCA models and algorithm settings:
>
> - "Build and Apply SVD and PCA Models"
> - "PCA Model Viewer and Algorithm Settings"
> - "SVD Model Viewer and Algorithm Settings"

13.10.1 Build and Apply SVD and PCA Models

To build an SVD or PCA model, use a Feature Extraction Node. A Feature Extraction model creates a Feature Build Node. If you are connected to Oracle Database 12*c*, then a Feature Build node creates one NMF model and one PCA model. You can add an SVD model.

To apply an SVD or PCA model, use an Apply Node.

13.10.2 PCA Model Viewer and Algorithm Settings

This section discusses the PCA Model viewer, the procedure to view the PCA Model viewer and the algorithm settings related to PCA. The topics are:

- PCA Model Viewer
- PCA Algorithm Settings

13.10.2.1 PCA Model Viewer

The PCA Model Viewer lets you examine a successfully built PCA model. You can view a PCA model by using any one of the following methods.

- Method 1
  1. Right-click the node where the model was built.
  2. Select **Go to Properties.**
  3. In the **Models** section in **Properties**, click .
- Method 2

1. Select the workflow node where the model was built.
2. Right-click and click **View Models.**
3. Select the model to view. The model viewer opens in a new tab. The default name of a PCA model has `PCA` in the name.

The model viewer has these tabs:

- Coefficients (PCA)
- PCA Scree Plot
- PCA Details
- Settings (PCA)

13.10.2.1.1 Coefficients (PCA)

For a given Feature ID, the coefficients are displayed in the Coefficients grid. The title of the grid Coefficients x of y displays number of rows returned out of all the rows available in the model. By default, Feature IDs are integers (1, 2, 3, …). The Eigenvalue for the selected Feature ID is displayed as a read-only value.

You can perform the following tasks:

- Filter (PCA)
- Rename (PCA)

The **Coefficients** grid has these columns:

- **Attribute**
- **Singular Value**

  The value is shown as a bar with the value centered in the bar. Positive values are light blue; negative values are red.

  The default is **Sort by absolute value**; if you deselect this option, click **Query**.

13.10.2.1.2 Rename (PCA) You can rename the selected Feature ID.

1. Enter in the new name in the Feature ID field.
2. Click **OK.**

> **Note:** Different features should have different names.

13.10.2.1.3 Filter (PCA) To view the filter categories, click .

The filter categories are

- Attribute, the default; search for an attribute name.
- Singular Value, the Singular value column

To create a filter, enter a string in the text box. After a string has been entered,  is displayed. To clear the filter, click it.

13.10.2.1.4 PCA Scree Plot

In the PCA Screen Plot:

- Features are plotted along the X-axis.

- Cutoff is plotted along the Y-axis.
- Variance is plotted as a red line.
- Cumulative percent is plotted as a blue line.

A grid below the graph shows Eigenvalue, Variance, and Cumulative Percent Variance for each Feature ID.

13.10.2.1.5 PCA Details This tab displays the value for these global details of the SVD model:

- Number of Components
- Suggested Cutoff

13.10.2.1.6 Settings (PCA) The **Settings** tab has these tabs:

- Summary (PCA)
- Inputs (PCA)

13.10.2.1.7 Summary (PCA)

- **General** settings lists the following:
  - Type of model (Classification, Regression, and so on
  - Owner of the model (the schema where the model was built)
  - Model Name
  - Creation Date
  - Duration of the model build (in minutes)
  - Size of the model (in MB)
  - Comments

  See Also: "General Settings"

- **Algorithm** settings lists the following:
  - The name of the algorithm used to build the model.
  - The algorithm settings that control the model build.

  See Also: "PCA Algorithm Settings"

13.10.2.1.8 Inputs (PCA) This tabs shows information about those attributes used to build the model.

Oracle Data Miner does not necessarily use all of the attributes in the build data. For example, if the values of an attribute are constant, then that attribute is not used.

For each attribute used to build the model, this tab displays:

- **Name**: The name of the attribute.
- **Data Type:** The data type of the attribute.
- **Mining Type**: Categorical or Numerical.
- **Data Prep**: `YES` indicates that data preparation was performed.

When you select an attribute in the **Attributes** list, the transformation properties viewer displays the imbedded transformation, created by either the user or Automatic Data Preparation, in the **Model transformations** list.

To see the reverse transformation, click **Show Reverse Expression**. Transformations are displayed in SQL notation. Not all transformations have a reverse. Transformations and Reverse transformations are not always displayed.

13.10.2.2 PCA Algorithm Settings

The PCA algorithm supports these settings:

- **Number of features**: The default is System Determined. To specify a value, select **User specified** and enter in an integer value.
- **Approximate Computation**: The default is System Determined. You can select either Enable or Disable. Approximate computations improve performance.

  See Also: "SVD Algorithm Settings" for more information about Approximate Computation

- **Projections**: The default is to *not* select Projections.
- **Number of Features:** The default is System Determined. You can specify a number.
- **Scoring Mode:** It is the scoring mode to use, either Singular Value Decomposition Scoring or Principal Components Analysis Scoring. The default is `Principal Components Analysis Scoring (PCA scoring)`.
  - When the build data is scored with SVD, the projections will be the same as the U matrix.
  - When the build data is scored with PCA, the projections will be the product of the U and S matrices.
- **U Matrix Output:** Whether or not the U matrix produced by SVD persists. The U matrix in SVD has as many rows as the number of rows in the Build data. To avoid creating a large model, the U matrix persists only when U Matrix Output is enabled. When U Matrix Output is enabled, the Build data must include a Case ID. The default is `Disable`.

13.10.3 SVD Model Viewer and Algorithm Settings

This section discusses the SVD Model viewer, the procedure to view the SVD Model viewer and the algorithm settings related to SVD. The topics are:

- SVD Model Viewer
- SVD Algorithm Settings

13.10.3.1 SVD Model Viewer

The SVD Model Viewer lets you examine a successfully built SVD model. You can view an SVD model by using one of the following methods:

- Method 1
  1. Right-click the node where the model was built.
  2. Select **Go to Properties.**
  3. In the **Models** section in **Properties**, click .

- Method 2
  1. Select the workflow node where the model was built.
  2. Right-click and click **View Models.**
  3. Select the model to view. The model viewer opens in a new tab. The default name of an SVD model has `SVD` in the name.

The model viewer has these tabs:

- Coefficients (SVD)
- SVD Singular Values
- SVD Details
- Settings (SVD)

13.10.3.1.1 Coefficients (SVD)

For a given Feature ID, the coefficients are displayed in the Coefficients grid. The title of the grid **Coefficients *x* of *y*** displays the number of rows returned out of all the rows available in the model. By default, Feature IDs are integers.

The Eigenvalue for the selected Feature ID is displayed as a read-only value.

Fetch Size limits the number of rows returned. The default is 1,000 or the value specified in the **Preference** settings for Model Viewers.

You can perform the following tasks:

- Rename (SVD)
- Filter (SVD)

The **Coefficients** grid has these columns:

- **Attribute**, Attribute name
- **Singular Value**

  The value is shown as a bar with the value centered in the bar. Positive values are light blue; negative values are red.

  The default is **Sort by absolute value**. To sort by signed value, deselect the option and then click **Query**.

13.10.3.1.2 Rename (SVD) You can rename the selected Feature ID. Enter in the new name and click **OK**. Different features should have different names.

13.10.3.1.3 Filter (SVD) To view the filter categories, click .

The filter categories are:

- Attribute, the default; search for an attribute name
- Singular Value, the singular value column

To create a filter, enter a string in the text box. After a string has been entered,  is displayed. To clear the filter, click it.

13.10.3.1.4 SVD Singular Values A grid shows the Singular Value for each Feature ID.

13.10.3.1.5 SVD Details This tab displays the value for these global details of the SVD model:

- Number of Components
- Suggested Cutoff

13.10.3.1.6 Settings (SVD)

The **Settings** tab has these tabs:

- Summary (SVD)
- Inputs (SVD)

13.10.3.1.7 Summary (SVD)

- **General** settings list the following:
  - Type of model (Classification, Regression, and so on)
  - Owner of the model (the Schema where the model was built)
  - Model Name
  - Creation Date
  - Duration of the model build (in minutes)
  - Size of the model (in MB)
  - Comments

  See Also: "General Settings"

- **Algorithm** settings list the following:
  - The name of the algorithm used to build the model.
  - The algorithm settings that control model build.

  See Also: "SVD Algorithm Settings"

13.10.3.1.8 Inputs (SVD)

This tabs shows information about those attributes used to build the model.

Oracle Data Miner does not necessarily use all of the attributes in the build data. For example, if the values of an attribute are constant, then that attribute is not used.

For each attribute used to build the model, this tab displays:

- **Name**: The name of the attribute.
- **Data Type:** The data type of the attribute.
- **Mining Type**: Categorical or Numerical.
- **Data Prep**: `YES` indicates that data preparation was performed.

  When you select an attribute in the **Attributes** list, the transformation properties viewer displays the imbedded transformation, created by either the user or Automatic Data Preparation, in the **Model transformations** list.

  To see the reverse transformation, click **Show Reverse Expression**. Transformations are displayed in SQL notation. Not all transformations have a reverse. Transformations and Reverse transformations are not always displayed.

13.10.3.2 SVD Algorithm Settings

The SVD algorithm supports the following settings:

- **Approximate Computation**: Specifies whether the algorithm should use approximate computations to improve performance. For SVD, approximation is often appropriate for data sets with many columns. An approximate low-rank decomposition provides good solutions at a reasonable computational cost. If you disable approximate computations for SVD, then approximation depends on the characteristics of the data. For data sets with more than 2500 attributes (the maximum number of features allowed) only approximate decomposition is possible. If approximate computation is disabled for a data set with more than 2500 attributes, then an exception is raised.

  Values for Approximate Computation are

  – **System Determined** (Default)

  – **Enable**

  – **Disable**

- **Automatic Preparation:** ON or OFF. The default is ON.
- **Number of Features:** System Determined (Default). You can specify a number.
- **Scoring Mode:** It is the scoring mode to use, either Singular Value Decomposition Scoring or Principal Components Analysis Scoring. The default is `Singular Value Decomposition Scoring (SVD scoring).`

  – When the build data is scored with SVD, the projections will be the same as the U matrix.

  – When the build data is scored with PCA, the projections will be the product of the U and S matrices.

- **U Matrix Output** Whether or not the U matrix produced by SVD persists. The U matrix in SVD has as many rows as the number of rows in the build data. To avoid creating a large model, the U matrix persists only when U Matrix Output is enabled. When U Matrix Output is enabled, the build data must include a Case ID. The default is Disable.

13.11 Support Vector Machine

You can use the Support Vector Machine (SVM) algorithm to build Classification, Regression, and Anomaly Detection models. The following topics explain Support Vector Machine:

- Support Vector Machine Algorithms
- Building and Testing SVM Models
- Applying SVM Models
- SVM Model Viewers and Algorithm Settings

13.11.1 Support Vector Machine Algorithms

The Support Vector Machines (SVM) algorithms are a suite of algorithms that can be used with variety of problems and data. By changing one kernel for another, SVM can solve a variety of data mining problems. Oracle Data Mining supports two kernel functions:

- Linear
- Gaussian

See Also: "SVM Kernel Functions"

The key features of SVM are:

- SVM can emulate traditional methods, such as Linear Regression and Neural Nets, but goes far beyond those methods in flexibility, scalability, and speed.
- SVM can be used to solve the following kinds of problems: Classification, Regression, and Anomaly Detection.

  Oracle Data Mining uses SVM as the one-class classifier for anomaly detection. When SVM is used for anomaly detection, it has the classification mining function but no target. Applying a One-class SVM model results in a prediction and a probability for each case in the scoring data. If the prediction is 1, then the case is considered typical. If the prediction is 0, then the case is considered anomalous.

  See Also:
  - "SVM Kernel Functions"
  - "How Support Vector Machines Work" for more information about SVM and *Oracle Data Mining Concepts.*

13.11.1.1 How Support Vector Machines Work

Data records with *n* attributes can be considered as points in *n*-dimensional space. SVM attempts to separate the points into subsets with homogeneous target values. Points are separated by hyperplanes in the linear case, and by non-linear separators in the non-linear case (Gaussian). SVM finds those vectors that define the separators giving the widest separation of classes (the support vectors). This is easy to visualize if $n = 2$; in that case, SVM finds a straight line (linear) or a curve (non-linear) separating the classes of points in the plane.

SVM solves regression problems by defining an *n*-dimensional tube around the data points, determining the vectors giving the widest separation.

13.11.1.2 SVM Kernel Functions

The Support Vector Machine (SVM) algorithm supports two kernel functions: Gaussian and Linear. The choice of kernel function depends on the type of model (classification or regression) that you are building and on your data.

When you choose a Kernel function, select one of the following:

- System Determined (Default)
- Gaussian
- Linear

For Classification models and Anomaly Detection models, use the Gaussian kernel for solving problems where the classes are not linearly separable, that is, the classes cannot be separated by lines or planes. Gaussian kernel models allow for powerful non-linear class separation modeling. If the classes are linearly separable, then use the linear kernel.

For Regression problems, the linear kernel is similar to approximating the data with a line. The linear kernel is more robust than fitting a line to the data. The Gaussian kernel approximates the data with a non-linear function.

13.11.2 Building and Testing SVM Models

This section describes how to build and test SVM models. You specify building a model by connecting the Data Source node that represents the build data to an appropriate Build node.

By default, a Classification or Regression node tests all the models that it builds. By default, the test data is created by splitting the input data into build and test subsets. Alternatively, you can connect two data sources to the build node, or you can test the model using a Test Node.

> **See Also:** "Testing Classification Models" or "Testing Regression Models" for more information

You can build three kinds of SVM models:

- SVM Classification Models
- SVM Regression Models
- SVM Anomaly Detection Models

13.11.2.1 SVM Classification Models

SVM Classification (SVMC) is based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. SVM finds the vectors (support vectors) that define the separators giving the widest separation of classes.

SVMC supports both binary and multiclass targets.

To build and test an SVMC model, use a Classification Node. By default, the SVMC Node tests the models that it builds. Test data is created by splitting the input data into build and test subsets. You can also test a model using a Test Node.

After you test an SVMC model, you can tune it.

> **See Also:** "Tuning Classification Models"

SVMC uses SVM Weights to specify the relative importance of target values.

13.11.2.1.1 SVM Weights SVM models are automatically initialized to achieve the best average prediction across all classes. If the training data does not represent a realistic distribution, then you can bias the model to compensate for class values that are under-represented. If you increase the weight for a class, then the percentage of correct predictions for that class should increase.

13.11.2.2 SVM Regression Models

SVM uses an epsilon-insensitive loss function to solve regression problems. SVM Regression (SVMR) tries to find a continuous function such that the maximum number of data points lie within the epsilon-wide insensitivity tube. Predictions falling within epsilon distance of the true target value are not interpreted as errors.

The epsilon factor is a regularization setting for SVMR. It balances the margin of error with model robustness to achieve the best generalization to new data.

To build and test an SVMR model, use a Regression Node. By default, the Regression Node tests the models that it builds. Test data is created by splitting the input data into build and test subsets. You can also test a model using a Test Node.

See Also: "Testing Regression Models" or more information about testing an SVMR model

13.11.2.3 SVM Anomaly Detection Models

Oracle Data Mining uses one-class SVM for anomaly detection (AD). There is no target for anomaly detection.

To build an AD model, use an Anomaly Detection Node connected to an appropriate data source.

See Also: "Anomaly Detection"

13.11.3 Applying SVM Models

You apply a model to new data to predict behavior. Use an Apply Node to apply an SVM model.

You can apply all three kinds of SVM models.

See Also: "Applying One-Class SVM Models" for more information about interpreting the apply results for One-Class SVM (anomaly detection) models

13.11.3.1 Applying One-Class SVM Models

One-class SVM models, when applied, produce a prediction and a probability for each case in the scoring data. This behavior reflects the fact that the model is trained with normal data.

- If the prediction is 1, then the case is considered typical.
- If the prediction is 0, then the case is considered anomalous.

13.11.4 SVM Model Viewers and Algorithm Settings

This section discusses the SVM Model viewer, the procedure to view the SVM Model viewer and the algorithm settings related to SVM. The topics are:

- SVM Classification Model Viewer
- SVM Classification Test Viewer
- SVM Classification Algorithm Settings
- SVM Regression Model Viewer
- SVM Regression Test Viewer
- SVM Regression Algorithm Settings

13.11.4.1 SVM Classification Model Viewer

The SVM Model Viewer lets you examine an SVM Classification model. You can view an SVMC model by using one of the following methods:

- Method 1
  1. Right-click the node where the model was built.
  2. Select **Go to Properties.**
  3. In the **Models** section in **Properties**, click .
- Method 2

1. Select the workflow node where the model was built.
2. Right-click and click **View Models.**
3. Select the model to view. The model viewer opens in a new tab. The information displayed in the model viewer depends on which kernel was used to build the model.

   The information displayed in the model viewer depends on which kernel was used to build the model.

   – If the Gaussian kernel was used, then there is one tab, **Settings**.

   – If the Linear Kernel was used, then there are three tabs, **Coefficients**, **Compare**, and **Settings**.

The tabs displayed in a SVMC model viewer depend on the kernel used to build the model:

- SVMC Model Viewer for Models with Linear Kernel
- SVMC Model Viewer for Models with Gaussian Kernel

13.11.4.1.1 SVMC Model Viewer for Models with Linear Kernel

If the SVMC model has a Linear kernel, then the viewer has these tabs:

- Coefficients (SVMC Linear)
- Compare (SVMC Linear)
- Settings (SVMC)

13.11.4.1.2 SVMC Model Viewer for Models with Gaussian Kernel

If the SVMC model has a Gaussian kernel, then the viewer has these tabs:

- Summary (SVMC)
- Inputs (SVMC)
- Weights (SVMC)
- Target Values (SVMC)

13.11.4.1.3 Coefficients (SVMC Linear)

Support Vector Machine Models built with the Linear Kernel have coefficients; the coefficients are real numbers. The number of coefficients may be quite large.

The **Coefficients** tab enables you to view SVM coefficients. The viewer supports sorting to specify the order in which coefficients are displayed and filtering to select which coefficients to display.

Coefficients are displayed in the Coefficients Grid (SVMC). The relative value of coefficients is shown graphically as a bar, with different colors for positive and negative values. For numbers close to zero, the bar may be too small to be displayed.

13.11.4.1.4 Coefficients Grid (SVMC)

The coefficients grid has these controls:

- **Target Value**: Select a specific target value and see the coefficients associated with that value. The default is to display the coefficients for the value that occurs least frequently.
- **Sort By Absolute Value**: If selected, coefficients are sorted by absolute value. If you sort by absolute value, then a coefficient of -2 comes before a coefficient of 1.9.The default is to sort by absolute value.

- **Fetch Size**: The number of rows displayed. To figure out if all the coefficients are displayed, choose a fetch size that is greater than the number of rows displayed.

You can search for attributes by name. Use . If no items are listed in the grid, then there are no coefficients for the selected target value. The coefficients grid has these columns:

- **Attribute**: Name of the attribute.
- **Value**: Value of the attribute. If the attribute is binned, then this may be a range.
- **Coefficient:** The probability for the value of the attribute.

  The value is shown as a bar with the value centered in the bar. Positive values are light blue; negative values are red.

13.11.4.1.5 Compare (SVMC Linear) Support Vector Machine Models built with the Linear kernel allow the comparison of target values. For selected attributes, Data Miner calculates the propensity (that is, the natural inclination or preference) to favor one of two target values. For example, propensity for *target value 1* is the propensity to favor *target value 1*.

> **See Also:** "Propensity"

To compare target values:

1. Select how to display information:
   - **Fetch Size**: The default fetch size is 1000 attributes. You can change this number.
   - **Sort by absolute value** is the default. You can deselect this option.
2. Select two distinct target values to compare:
   - **Target Value 1**: Select the first target value.
   - **Target Value 2**: Select the second target value.
3. Click **Query**. If you have not changed any defaults, then this step is not necessary.

The following information is displayed in the grid:

- **Attribute**: The name of the attribute.
- **Value**: Value of the attribute
- **Propensity for** *Target_Value_1*: Propensity to favor **Target Value 1**.
- **Propensity for** *Target_Value_2*: Propensity to favor **Target Value 2**.

You can Search the grid in several ways:

13.11.4.1.6 Search Use to search the grid.

You can search by name (the default), by value, and by propensity for Target Value 1 or propensity for Target Value 2.

- To select a different search option, click the triangle beside the binoculars.
- To clear a search, click .

13.11.4.1.7 Propensity Propensity is intended to show for a given attribute/value pair, which of the two target values has more predictive relationship. Propensity can be measured in terms of being predicted for or against a target value. If propensity is against a value, then the number is negative.

13.11.4.1.8 Settings (SVMC) The **Settings** tab displays information about how the model was built:

- Summary (SVMR) tab: Contains Model and Algorithm settings
- Inputs (SVMC) tab: Contains attributes used to build the model.
- Target Values (SVMC) tab: Contains targets.
- Cost Matrix/Benefit tab: If you tune the model, then this tab displays the cost matrix created by tuning.

13.11.4.1.9 Summary (SVMC)

- **General** settings list the following:
  - Type of model (Classification, Regression, and so on)
  - Owner of the model (the Schema where the model was built)
  - Model Name
  - Creation Date
  - Duration of the model build (in minutes)
  - Size of the model (in MB)
  - Comments.

  See Also: "General Settings"

- **Algorithm** settings list the following:
  - The name of the algorithm used to build the model.
  - The algorithm settings that control the model build.

  See Also: "SVM Classification Algorithm Settings"

13.11.4.1.10 Settings (SVMC Linear) The **Settings** tab has these tabs:

- Summary (SVMC)
- Inputs (SVMC)
- Target Values (SVMC)

13.11.4.1.11 Inputs (SVMC) A list of the attributes used to build the model. For each attribute the following information is displayed:

- **Name**: The name of the attribute.
- **Data Type:** The data type of the attribute.
- **Mining Type**: Categorical or Numerical.
- **Target**: The  icon indicates that the attribute is a target attribute.
- **Data Prep**: `YES` indicates that data preparation was performed.

  When you select an attribute in the **Attributes** list, the transformation properties viewer displays the imbedded transformation, created by either the user or Automatic Data Preparation, in the **Model transformations** list.

To see the reverse transformation, click **Show Reverse Expression**. Transformations are displayed in SQL notation. Not all transformations have a reverse. Transformations and Reverse transformations are not always displayed.

13.11.4.1.12 Target Values (SVMC) Shows the values of the target attributes.

- Click to search for target values.
- Click to clear search.

13.11.4.1.13 Weights (SVMC) In Support Vector Machines classifications, weights are a biasing mechanism for specifying the relative importance of target values (classes). SVM models are automatically initialized to achieve the best average prediction across all classes. However, if the training data does not represent a realistic distribution, then you can bias the model to compensate for class values that are underrepresented. If you increase the weight for a class, then the percentage of correct predictions for that class should increase.

13.11.4.1.14 Algorithm Settings for SVMC For Classification, the SVM algorithm has these settings:

- **Algorithm Name:** Support Vector Machine
- **Kernel Function:** Gaussian or Linear
- **Tolerance value:** The default is 0.001
- **Specify complexity factor:** By default, it is not specified.
- **Active Learning:** ON
- **Standard Deviation** (Gaussian Kernel only)
- **Cache Size** (Gaussian Kernel only)

> **See Also:** "SVM Classification Algorithm Settings"for more information about these settings

13.11.4.2 SVM Classification Test Viewer

By default, any Classification or Regression model is automatically tested. A Classification model is tested by comparing the predictions of the model with known results. Oracle Data Miner keeps the latest test result.

To view the test results for a model, right-click the build node and select **View Results**.

> **See Also:** "Testing Classification Models" for more information about the Test Viewers

13.11.4.3 SVM Classification Algorithm Settings

The settings that you can specify for the Support Vector Machine (SVM) algorithm depend on the Kernel function that you select. See "SVM Kernel Functions" for information about how to select a Kernel Function.

The meaning of the individual settings is the same for both Classification and Regression.

To edit settings SVM Classification algorithm settings:

1. You can edit the settings by using one of the following options:
    - Right-click the Classification node and select **Advanced Settings.**

- Right-click the Classification node and select **Edit.** Then, click **Advanced.**

2. In the **Algorithm Settings** tab, the settings are available. Select the Kernel Function. The options are:

   - **System determined** (Default). After the model is built, the kernel used is displayed in the settings in the model viewer.

     See Also: "Algorithm Settings for Linear or System Determined Kernel (SVMR)" for remaining settings

   - **Linear**. If SVM uses the linear kernel, then the model generates coefficients.

     See Also: "Algorithm Settings for Linear or System Determined Kernel (SVMR)" for remaining settings

   - **Gaussian** (a non-linear function).

     See Also: "Algorithm Settings for Gaussian Kernel (SVMC)" for remaining settings

3. Click **OK** after you are done.

13.11.4.3.1 Algorithm Settings for Linear or System Determined Kernel (SVMC) If you specify a linear kernel or if you let the system determine the kernel, then you can change the following settings:

- Tolerance Value
- Complexity Factor
- Active Learning

13.11.4.3.2 Algorithm Settings for Gaussian Kernel (SVMC) If you specify the Gaussian kernel, then you can change the following settings:

- Tolerance Value
- Complexity Factor
- Active Learning
- Standard Deviation (Gaussian Kernel)
- Cache Size (Gaussian Kernel)

13.11.4.4 SVM Regression Model Viewer

The SVM Regression Model Viewer lets you examine an SVMR model. You can view an SVMR model by using one of the following methods:

- Method 1
  1. Right-click the node where the model was built.
  2. Select **Go to Properties.**
  3. In the **Models** section in **Properties**, click .
- Method 2
  1. Select the workflow node where the model was built.
  2. Right-click and click **View Models.**

3. Select the model to view.

The information displayed in the model viewer depends on which kernel was used to build the model.

- If the Gaussian kernel was used, then there is one tab, **Settings**.
- If the Linear Kernel was used, then there are three tabs: **Coefficients**, **Compare**, and **Settings**.

The tabs displayed in a SVMC model viewer depend on the kernel used to build the model:

- SVMR Model Viewer for Models with Linear Kernel
- SVMR Model Viewer for Models with Gaussian Kernel

13.11.4.4.1 SVMR Model Viewer for Models with Linear Kernel

If the SVMC model has a linear kernel, then the viewer has these tabs:

- Coefficients (SVMR)
- Settings (SVMR)

13.11.4.4.2 SVMR Model Viewer for Models with Gaussian Kernel

If the SVMC model has a Gaussian kernel, then the viewer has these tabs:

- Summary (SVMR)
- Inputs (SVMR)

13.11.4.4.3 Coefficients (SVMR)

Support Vector Machine Models built with the Linear Kernel have coefficients. The coefficients are real numbers. The number of coefficients may be quite large.

The **Coefficients** tab enables you to view SVMR coefficients. The viewer supports sorting to specify the order in which coefficients are displayed and filtering to select which coefficients to display.

The coefficients are displayed in the SVMR Coefficients Grid. The relative value of the coefficients is shown graphically as a bar, with different colors for positive and negative values. For numbers close to zero, the bar may be too small to be displayed.

13.11.4.4.4 SVMR Coefficients Grid

Information about coefficients is organized as follows:

- **Sort by absolute value**: The default is to sort by absolute value. For example 1 and -1 have the same absolute value. If you change this value, then you must click **Query**.
- **Fetch Size:** It is the maximum number of rows to fetch; the default is 1,000. Smaller values result in faster fetches. If you change this value, then you must click **Query**.
- **Coefficients**: The number of coefficients displayed; for example, *95 out of 95*, indicating that there are 95 coefficients and all 95 of them are displayed.

You can perform the following tasks:

- Search: Use to search for items. You can search by:
  - Attribute name (Default)
  - Value

- Coefficient
- All (AND): If you search by this criteria, then you search for items that satisfy all criteria specified. For example, a search for *ED Bac* finds all attributes where both values appear.
- All (Or): If you search by this criteria, then you search for attributes that include at least one value

- Clear search: To clear a search, click .
- To select a different search option, click the triangle beside the binoculars.

Coefficients are listed in a grid. The coefficients grid has these columns:

- **Attribute**: Name of the attribute
- **Value**: Value of the attribute
- **Coefficient**: The value of each coefficient for the selected target value is displayed. A bar is shown in front of (and possible overlapping) each coefficient. The bar indicates the relative size of the coefficient. For positive values, the bar is light blue; for negative values, the bar is red. If a value is close to 0, then the bar may be too small to be displayed.

13.11.4.4.5 Inputs (SVMR) A list of the attributes used to build the model. For each attribute, the following information is displayed:

- **Name**: The name of the attribute.
- **Data Type:** The data type of the attribute.
- **Mining Type**: Categorical or Numerical.
- **Target**: The  icon indicates that the attribute is a target attribute.
- **Data Prep**: `YES` indicates that data preparation was performed.

  When you select an attribute in the **Attributes** list, the transformation properties viewer displays the imbedded transformation, created by either the user or Automatic Data Preparation, in the **Model transformations** list.

  To see the reverse transformation, click **Show Reverse Expression**. Transformations are displayed in SQL notation. Not all transformations have a reverse. Transformations and Reverse transformations are not always displayed.

13.11.4.4.6 Settings (SVMR) The **Settings** tab displays information about how the model was built:

- Summary (SVMR) tab: Contains the Model and Algorithm settings.
- Inputs (SVMR) tab: Contains the attributes used to build the model.

13.11.4.4.7 Summary (SVMR)

- **General** settings list the following:
  - Type of model (Classification, Regression, and so on)
  - Owner of the model (the Schema where the model was built)
  - Model Name
  - Creation Date
  - Duration of the model build (in minutes)

- Size of the model (in MB)
- Comments.

See Also: "General Settings"

- **Algorithm** settings list the following:
  - The name of the algorithm used to build the model.
  - The algorithm settings that control the model build.

See Also: "SVM Regression Algorithm Settings"

13.11.4.5 SVM Regression Test Viewer

By default, any Classification or Regression model is automatically tested. A Classification model is tested by comparing the predictions of the model with known results. Oracle Data Miner keeps the latest test result.

To view the test results for a model, right-click the Build node and select **View Results**.

See Also: "Testing Regression Models"

13.11.4.6 SVM Regression Algorithm Settings

The settings that you can specify for the Support Vector Machine (SVM) algorithm depend on the Kernel function that you select.

See Also: "SVM Kernel Functions" for information about how to select a Kernel Function

The meaning of the individual settings is the same for both classification and regression.

To edit settings SVM Regression Algorithm settings:

1. You can edit the settings by using one of the following options:
   - Right-click the Classification node and select **Advanced Settings.**
   - Right-click the Classification node and select **Edit.** Then click **Advanced.**
2. In the **Algorithm Settings** tab, the settings are available. Select **Kernel Function**. The options are:
   - **System determined** (Default). After the model is build, the kernel used is displayed in the settings in the model viewer.

     See Also: "Algorithm Settings for Linear or System Determined Kernel (SVMR)" for remaining settings

   - **Linear**. If SVM uses the linear kernel, then the model generates coefficients.

     See Also: "Algorithm Settings for Linear or System Determined Kernel (SVMR)" for remaining settings

   - **Gaussian** (a non-linear function).

     See Also: "Algorithm Settings for Gaussian Kernel (SVMC)" for the remaining settings

3. Click **OK** after you are done.

13.11.4.6.1 Algorithm Settings for Linear or System Determined Kernel (SVMR) If you specify a linear kernel or if you let the system determine the kernel, then you can change the following settings for an SVM Regression model:

- Tolerance Value
- Complexity Factor
- Active Learning

13.11.4.6.2 Algorithm Settings for Gaussian Kernel (SVMR) If you specify the Gaussian kernel, then you can change the following settings for an SVM Regression model:

- Tolerance Value
- Complexity Factor
- Active Learning
- Standard Deviation (Gaussian Kernel)
- Cache Size (Gaussian Kernel)

13.11.4.6.3 Active Learning Active Learning is a methodology optimizes the selection of a subset of the support vectors that maintain accuracy while enhancing the speed of the model. The Key features of Active Learning are:

- Increases performance for a linear kernel. Active learning both increases performance and reduces the size of the Gaussian kernel. This is an important consideration if memory and temporary disk space are issues.
- Forces the SVM algorithm to restrict learning to the most informative examples and not to attempt to use the entire body of data. Usually, the resulting models have predictive accuracy comparable to that of the standard (exact) SVM model.

You should not disable this setting

Active Learning is selected by default. It can be turned off by deselecting **Active Learning**.

13.11.4.6.4 Automatic Data Preparation Most algorithms require some form of data transformation. During the model building process, Oracle Data Mining can automatically perform the transformations required by the algorithm. You can supplement the automatic transformations with additional transformations of your own, or you can manage all the transformations yourself.

In calculating automatic transformations, Oracle Data Mining uses heuristics that address the common requirements of a given algorithm. This process results in reasonable model quality mostly.

13.11.4.6.5 Cache Size (Gaussian Kernel) If you select the Gaussian kernel, then you can specify a cache size for the size of the cache used for storing computed kernels during the build operation. The default size is 50 megabytes.

The most expensive operation in building a Gaussian SVM model is the computation of kernels. The general approach taken to build is to converge within a chunk of data at a time, then to test for violators outside of the chunk. The build is complete when there are no more violators within tolerance. The size of the chunk is chosen such that the associated kernels can be maintained in memory in a Kernel Cache. The larger the chunk size, the better the chunk represents the population of training data and the fewer number of times new chunks must be created. Generally, larger caches imply faster builds.

13.11.4.6.6 Complexity Factor The default is to *not* specify a complexity factor.

You specify the complexity factor for an SVM model by selecting **Specify the complexity factors**.

The complexity factor determines the trade-off between minimizing model error on the training data and minimizing model complexity. Its responsibility is to avoid over-fit (an over-complex model fitting noise in the training data) and under-fit (a model that is too simple).

A very large value of the complexity factor places an extreme penalty on errors, so that SVM seeks a perfect separation of target classes. A small value for the complexity factor places a low penalty on errors and high constraints on the model parameters, which can lead to under-fit.

If the histogram of the target attribute is skewed to the left or to the right, then try increasing complexity.

The default is to specify no complexity factor, in which case the system calculates a complexity factor. If you do specify a complexity factor, then specify a positive number. If you specify a complexity factor for Anomaly Detection, then the default is 1.

13.11.4.6.7 Standard Deviation (Gaussian Kernel) If you select the Gaussian kernel, then you can specify the standard deviation of the Gaussian kernel. This value must be positive. The default is to not specify the standard deviation.

For anomaly detection, if you specify standard deviation, then the default is 1.

13.11.4.6.8 Tolerance Value Tolerance value is the maximum size of a violation of convergence criteria such that the model is considered to have converged. The default value is 0.001. Larger values imply faster building but less accurate models.

13.12 Settings Information

This section contains topics about settings that are common to most algorithms:

- General Settings
- Other Settings
- Epsilon Value
- Automatic Data Preparation (ADP)

13.12.1 General Settings

The **Settings** tab of a model viewer displays settings in two categories:

- General Settings, described in this topic
- Algorithms Settings that are specific to the selected algorithm

These General Settings are provided for all algorithms:

- **Type**: The mining function for the model: anomaly detection, association rules, attribute importance, classification, clustering, feature extraction, or regression.
- **Owner**: The data mining account (schema) used to build the model.

- **Model Name**: The name of the model.
- **Target Attribute**: The target attribute; only Classification and Regression models have targets.
- **Creation Date**: The date when the model was created in the form `MM/DD/YYYY`
- **Duration**: Time in minutes required to build model.
- **Size**: The size of the model in megabytes.
- **Comment**: For models not created using Oracle Data Miner, this option displays comments embedded in the models. To see comments for models built using Oracle Data Miner, go to **Properties** for the node where the model is built.

  Models created using Oracle Data Miner may contain BALANCED, NATURAL, CUSTOM, or TUNED. Oracle Data Miner inserts these values to indicate if the model has been tuned and in what way it was tuned.

13.12.2 Other Settings

Other Settings include:

- **Limit Number of Attributes in Each Rule**: By default, this option is selected. The maximum number of attributes in each rule. This number must be an integer between 2 and 20. Higher numbers of rules result in slower builds. You can change the number of attributes in a rule, or you can specify no limit for the number of attributes in a rule. It is a good practice to start with the default and increase this number slowly.
  - To specify no limit, deselect this option.
  - Specifying many attributes in each rule increases the number of rules considerably.
  - The default is 3.
- **Automatic preparation:** `ON` or `OFF`. `ON` signifies that Automatic Data Preparation (ADP) is used for normalization and outlier detection. The SVM algorithm automatically handles missing value treatment and the transformation of categorical data. Normalization and outlier detection must be handled by ADP or prepared manually. The default is `ON`.
- **Minimum Support**: A number between 0 and 100 indicating a percentage. Smaller values for support results in slower builds and requires more system resources. The default is 5%.
- **Minimum Confidence**: Confidence in the rules. A number between 0 and 100 indicating a percentage. High confidence results in a faster build. The default is 10%.

13.12.3 Epsilon Value

You specify the epsilon value for an SVM model by clicking the option **Yes** in answer to the question `Do you want to specify an epsilon value?`. The epsilon value must be either greater than 0 or undefined.

SVM makes a distinction between small errors and large errors. The difference is defined by the epsilon value. The algorithm calculates and optimizes an epsilon value internally, or you can supply a value.

- If the number of support vectors defined by the model is very large, then try a larger epsilon value.

- If there are very high cardinality categorical attributes, then try decreasing epsilon.

By default, no epsilon value is specified. In such a case, the algorithm calculates an epsilon value.

Index

S

T

U

V

W