

Machine Learning Process

ORACLE®

Copyright © 2022, Oracle and/or its affiliates. All rights reserved.

This software and related documentation are provided under a license agreement containing restrictions on use and disclosure and are protected by intellectual property laws. Except as expressly permitted in your license agreement or allowed by law, you may not use, copy, reproduce, translate, broadcast, modify, license, transmit, distribute, exhibit, perform, publish, or display any part, in any form, or by any means. Reverse engineering, disassembly, or decompilation of this software, unless required by law for interoperability, is prohibited.

The information contained herein is subject to change without notice and is not warranted to be error-free. If you find any errors, please report them to us in writing.

If this is software or related documentation that is delivered to the U.S. Government or anyone licensing it on behalf of the U.S. Government, then the following notice is applicable:

U.S. GOVERNMENT END USERS: Oracle programs, including any operating system, integrated software, any programs installed on the hardware, and/or documentation, delivered to U.S. Government end users are "commercial computer software" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, use, duplication, disclosure, modification, and adaptation of the programs, including any operating system, integrated software, any programs installed on the hardware, and/or documentation, shall be subject to license terms and license restrictions applicable to the programs. No other rights are granted to the U.S. Government.

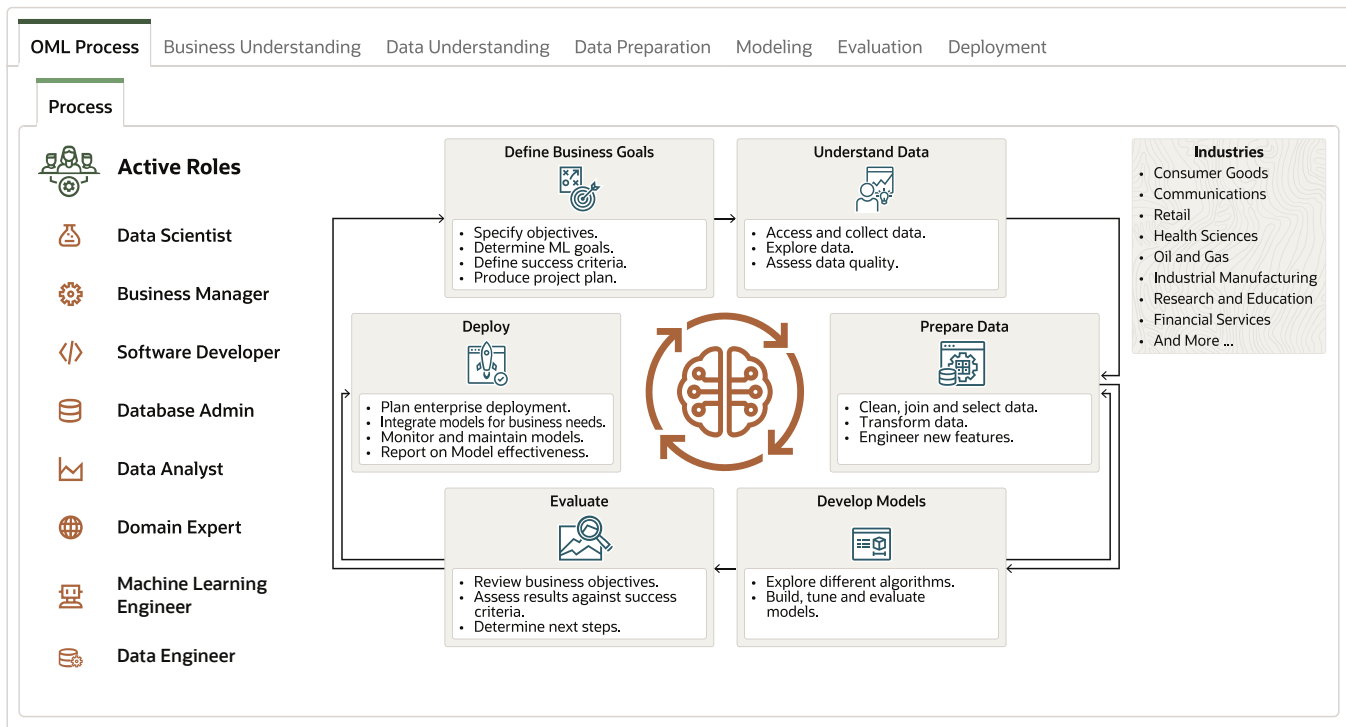
This software or hardware is developed for general use in a variety of information management applications. It is not developed or intended for use in any inherently dangerous applications, including applications that may create a risk of personal injury. If you use this software or hardware in dangerous applications, then you shall be responsible to take all appropriate fail-safe, backup, redundancy, and other measures to ensure its safe use. Oracle Corporation and its affiliates disclaim any liability for any damages caused by use of this software or hardware in dangerous applications.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark of The Open Group.

This software or hardware and documentation may provide access to or information about content, products, and services from third parties. Oracle Corporation and its affiliates are not responsible for and expressly disclaim all warranties of any kind with respect to third-party content, products, and services unless otherwise set forth in an applicable agreement between you and Oracle. Oracle Corporation and its affiliates will not be responsible for any loss, costs, or damages incurred due to your access to or use of third-party content, products, or services, except as set forth in an applicable agreement between you and Oracle.

Oracle Machine Learning Process



Welcome to the Machine Learning poster. This poster will walk you through the cross-industry standard process for data mining (CRISP-DM), which is a widely adopted process to guide the implementation of machine learning projects. The poster also depicts how the Oracle Machine Learning products help at each step of the process as well as the primary actors involved at each step.

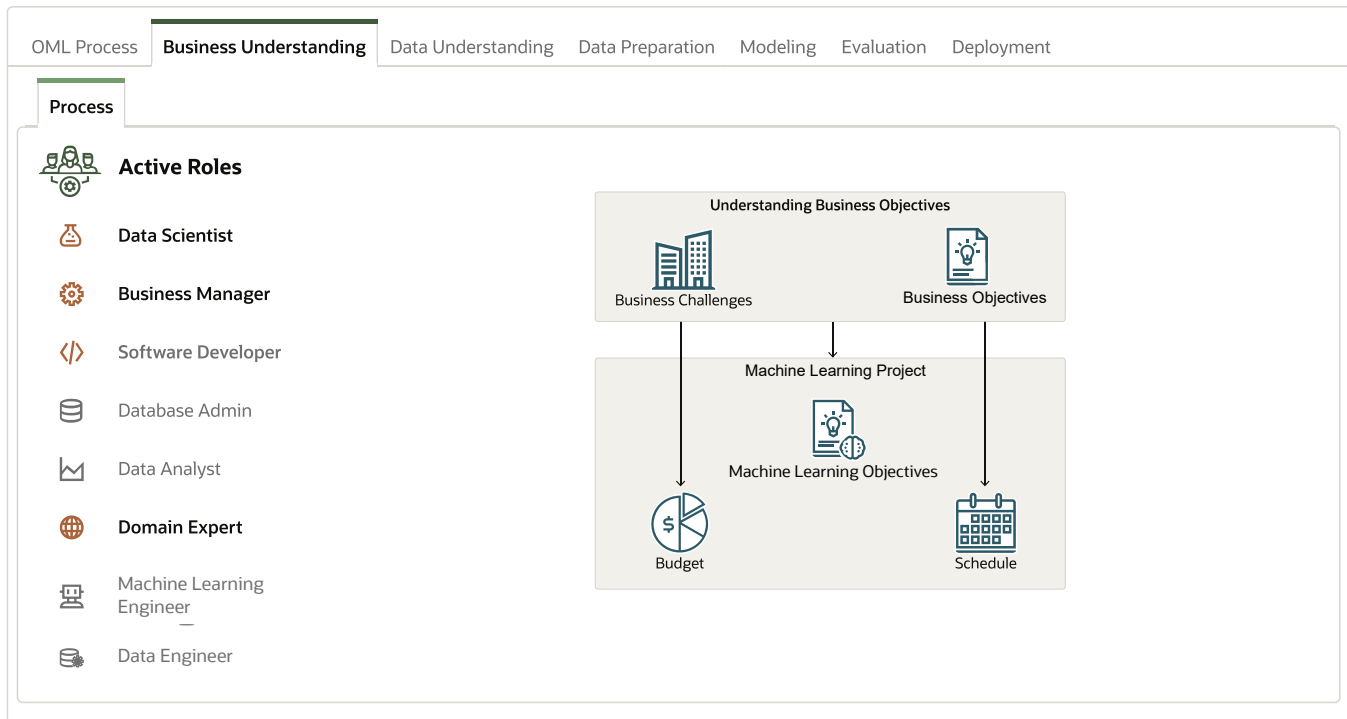
In the poster we show CRISP-DM as a linearly progressing process, however, in practice this is rarely the case. At many steps, based on the output, you may need to revert to the preceding step or steps.

For instance, during modelling, you may find that the model is not performing well on test data but does well on the training data, which could be because of model overfitting. This may necessitate reverting to the data preparation step for either better feature selection or to increase the size of training data set. It could also mean needing to supplement with additional data (predictors), do additional feature engineering, or explore data quality issues.

This poster depicts:

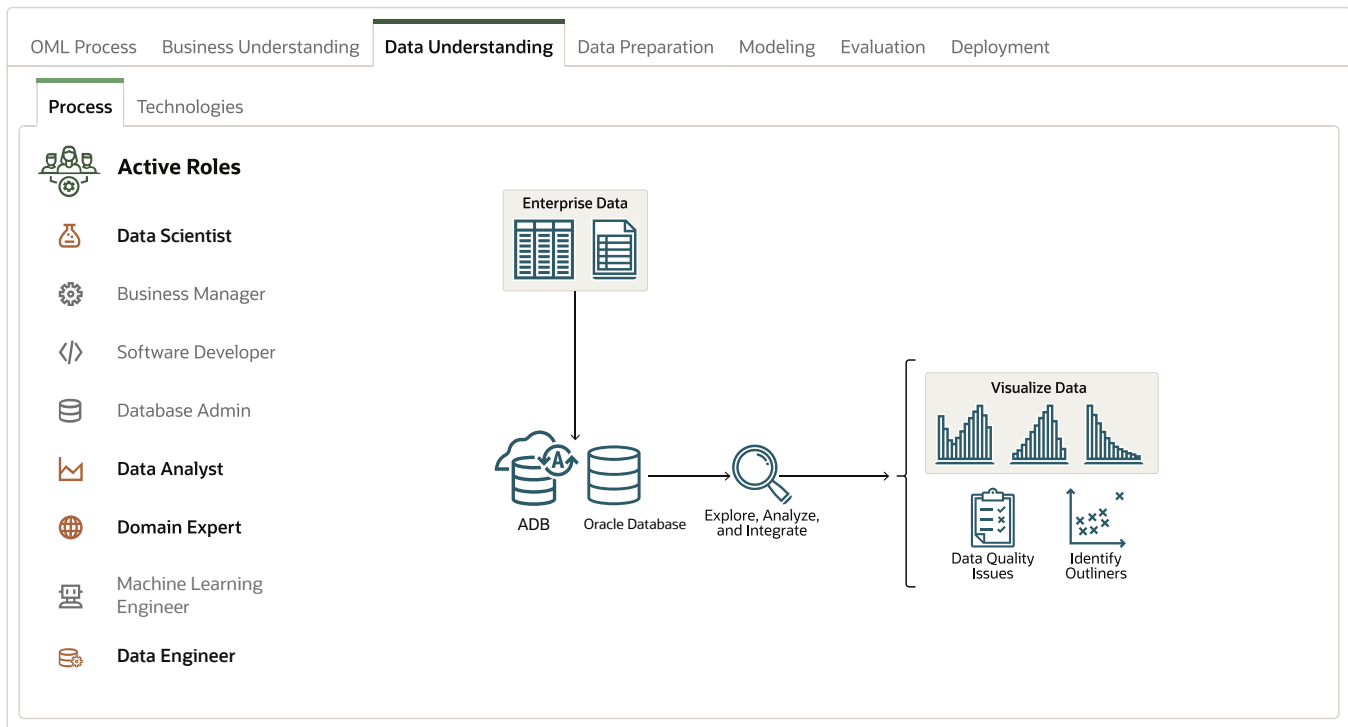
- CRISP-DM process steps for machine learning projects
- Oracle Machine Learning technologies that apply to a step
- Roles involved in each step

Business Understanding Process



















A machine learning project, like any other technology implementation project, is triggered by business needs of the enterprise. Business leaders within an enterprise determine the business challenges and define business-level goals, objectives, and requirements the project must address. In a consultative process, a data scientist then analyzes the business requirements and objectives, and translates these into machine learning objectives. A budget and schedule is typically also put in place.

Data Understanding Process

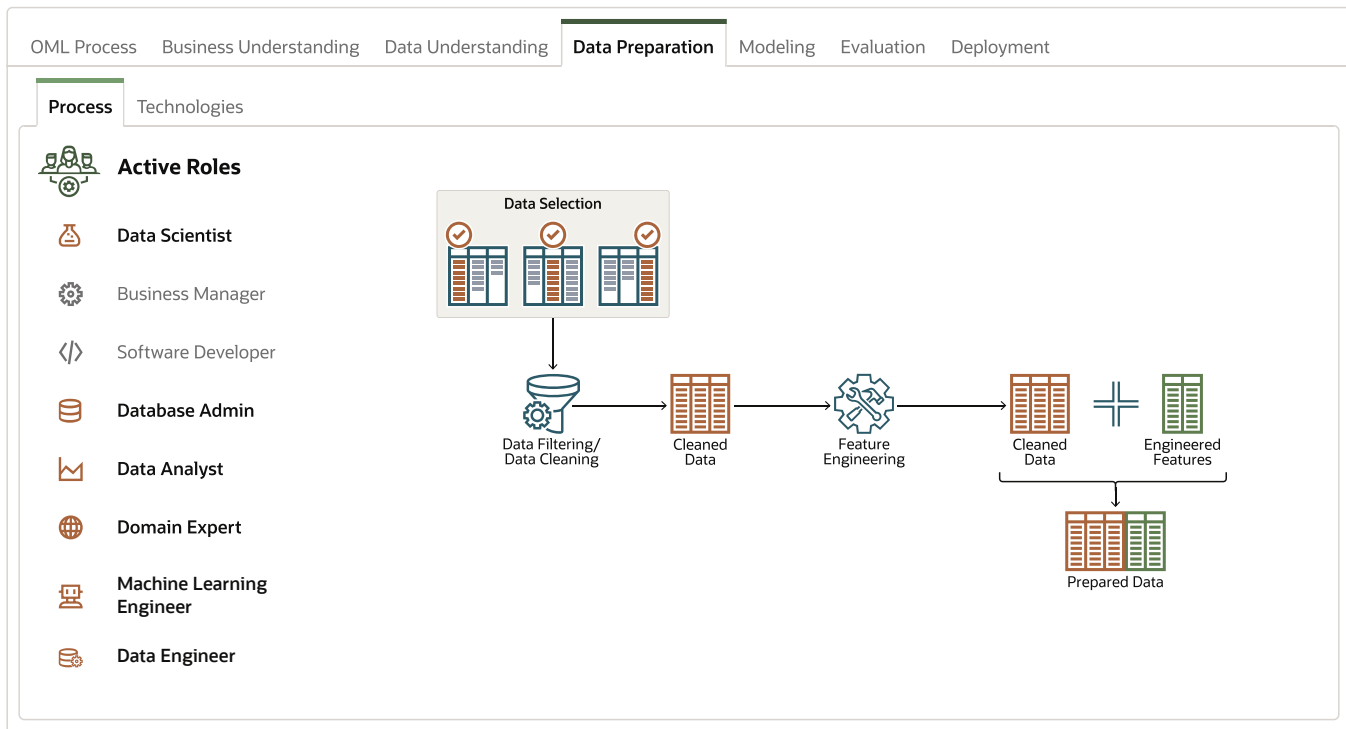


Once the machine learning objectives are defined, you acquire, integrate and analyze the business data. Data may exist in many places. Ideally, most data resides conveniently in Oracle Database, Oracle Database Cloud Service, or Oracle Autonomous Database (ADB). Other data may reside in Cloud Storage, flat files, or other repositories. Oracle can provide access to such data so that it can be explored, integrated, and analyzed through Oracle Database.

Data Understanding Technologies















OML Process		Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Process	Technologies						
Active Roles		Autonomous Database	Database Cloud Service	On-Premises Database	Additional Oracle Stack Tools		
 Data Scientist	OML for SQL 				<ul style="list-style-type: none"> • Oracle Analytics Cloud • OCI Data Catalog 		
 Business Manager		✓	✓	✓			
 Software Developer	OML for R 	✓	✓	✓			
 Database Admin							
 Data Analyst	OML for Python 	✓		✓			
 Domain Expert							
 Machine Learning Engineer	Data Miner 	✓	✓	✓			
 Data Engineer	OML Notebooks 	✓					

Data Preparation Process

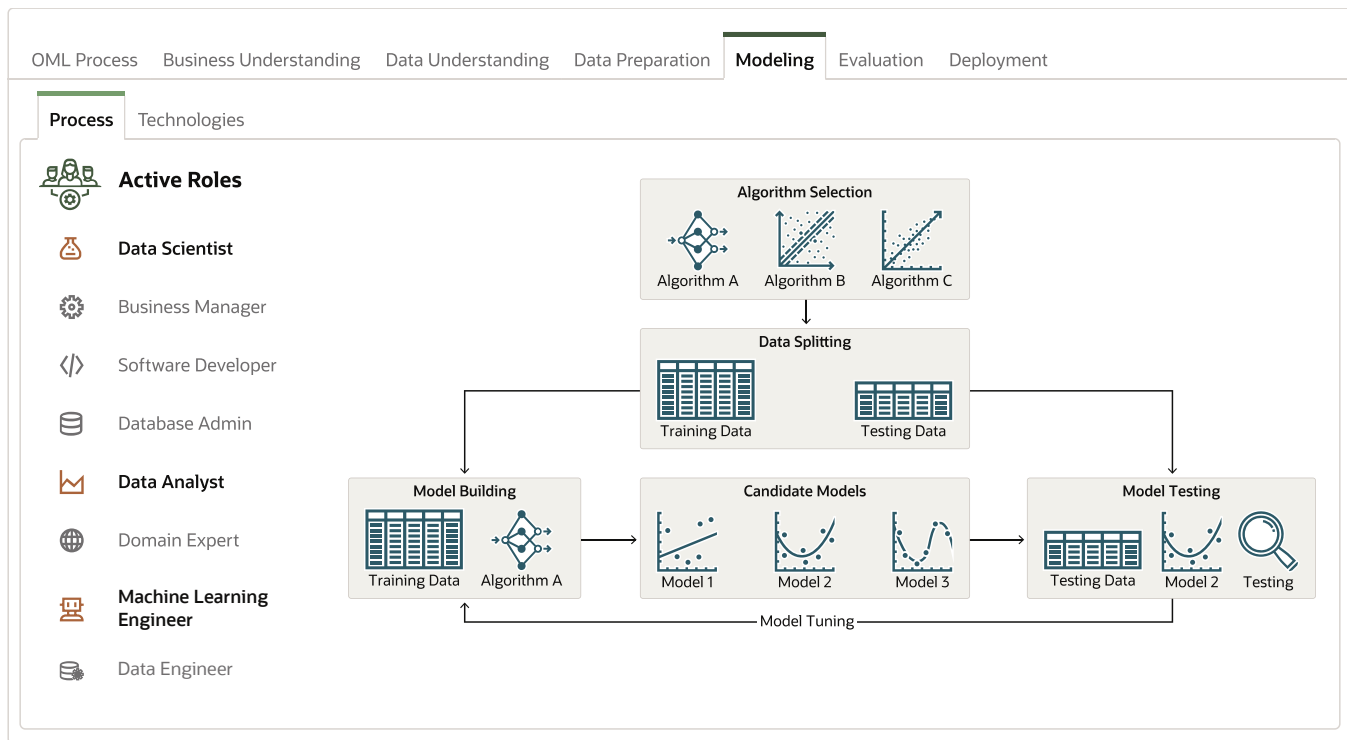


Data preparation is the step that follows understanding and analyzing the data. In this step, you use the outcome of the previous step to fill in any gaps in the data. If the data contains outlier values, you may want to treat or remove those, and you may want to exclude attributes (a.k.a. features or predictors) that are not significant. It is likely that new attributes need to be engineered, such as computing a person's age from their date of birth.

Data Preparation Technologies

Process		Technologies			Additional Oracle Stack Tools
		Autonomous Database	Database Cloud Service	On-Premises Database	
 Active Roles					<ul style="list-style-type: none"> • Oracle Analytics Cloud • Oracle Data Integrator
	Data Scientist	OML for SQL 	✓	✓	✓
	Business Manager				
	Software Developer	OML for R 	✓	✓	✓
	Database Admin				
	Data Analyst	OML for Python 	✓		✓
	Domain Expert				
	Machine Learning Engineer	Data Miner 	✓	✓	✓
	Data Engineer	OML Notebooks 	✓		

Modeling Process



After the data has been prepared, it is ready for the next step, modeling. In this step, one or more algorithms are used to build models that reflect the patterns extracted from the data according to each algorithm's behavior. Each algorithm has various settings, or hyperparameters, that affect algorithm performance and resulting model quality. In the case of supervised learning, where there are known outcomes by which the algorithm determines patterns, data scientists will assess the accuracy of the competing algorithms using a variety of evaluation metrics. In the case of unsupervised learning, where there is not a known outcome, patterns are discovered and various techniques may be applied to compare and select useful models.

Modeling Technologies

OML Process Business Understanding Data Understanding Data Preparation **Modeling** Evaluation Deployment

Process **Technologies**



Active Roles



Data Scientist



Business Manager



Software Developer



Database Admin



Data Analyst



Domain Expert



Machine Learning Engineer



Data Engineer

Autonomous Database



Database Cloud Service



On-Premises Database



Additional Oracle Stack Tools

- Oracle Analytics Cloud
- Oracle Data Integrator

OML for SQL



OML for R



OML for Python



Data Miner



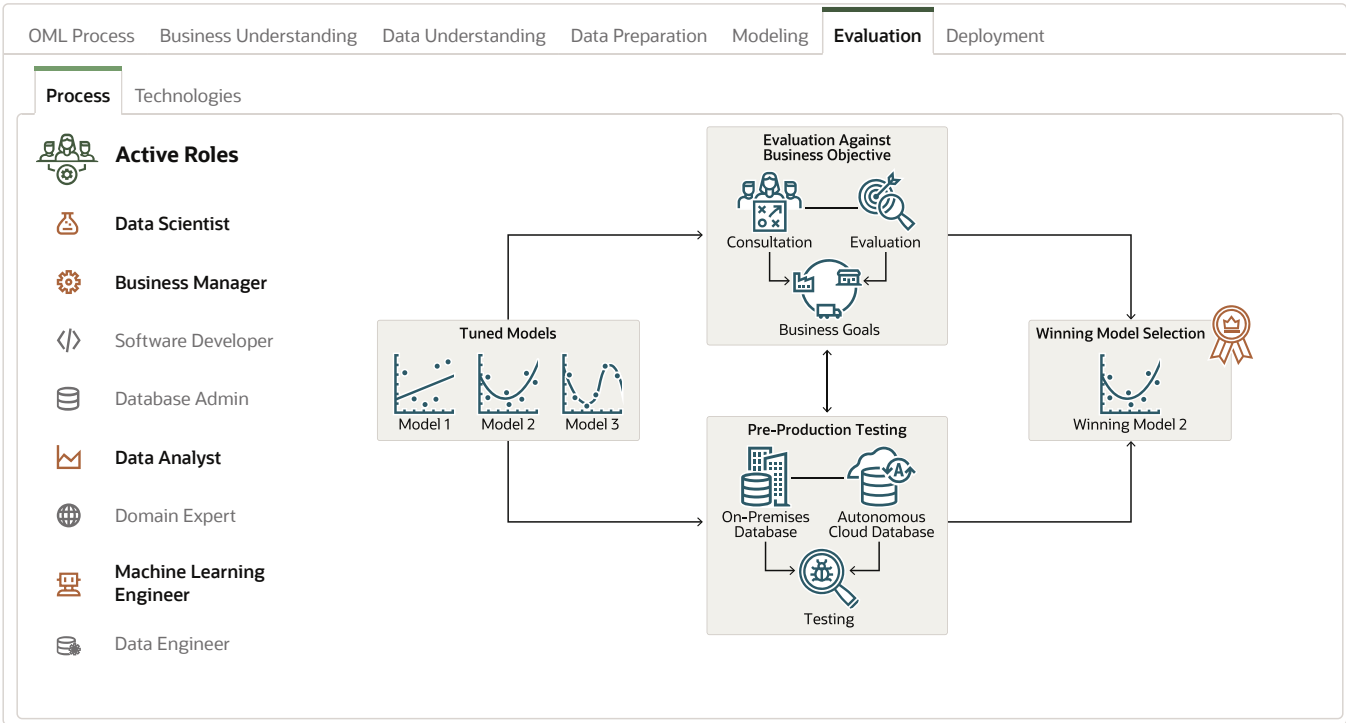
OML Notebooks



OML AutoML UI



Evaluation Process

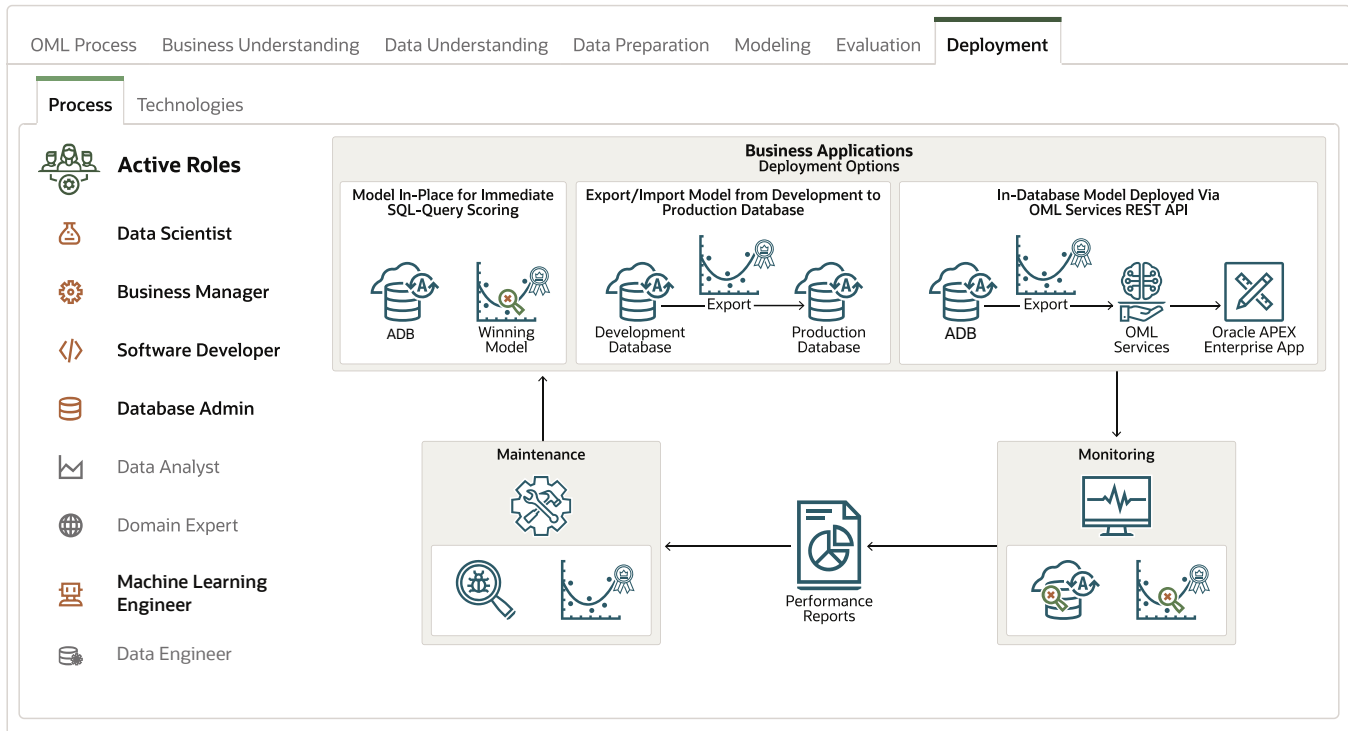


Evaluation Technologies

OML Process		Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment	
Process	Technologies							
Active Roles	Autonomous Database	Database Cloud Service	On-Premises Database	Additional Oracle Stack Tools				
Data Scientist	 OML for SQL 	✓	✓	✓	• None			
Business Manager	OML for R 	✓	✓	✓				
Software Developer	OML for Python 	✓	✓	✓				
Database Admin	Data Miner 	✓	✓	✓				
Data Analyst	OML Notebooks 	✓						
Domain Expert	OML AutoML UI 	✓						
Machine Learning Engineer								
Data Engineer								

The modeling step may generate multiple machine learning models that address the business objectives and success criteria posed at the beginning of the project. However, some of these models may do better than others. In the evaluation step of the CRISP-DM process you zero in on the model of choice by evaluating it against the business objectives. In addition, in this step, you typically also test the overall solution in a pre-production environment.

Deployment Process



Deployment is the last stage of the CRISP-DM process. At this stage, you operationalize the machine learning solution by deploying it in a production environment.

Deployment can take on many forms. The most common is to take a predictive model and score data, either in batch or interactively through an application or dashboard. While some uses require only the prediction, others may also require prediction details, i.e., what factors contributed to the prediction. In other cases, the model itself contains information that can be surfaced to end users, e.g., overall important factors (feature coefficients) that determine predictions, rules of a decision tree that identify customer segments, cluster centroid definitions that represent each cluster, or rules from market basket analysis describing which items are frequently purchased together.

Deployment may involve using a specific model produced from an earlier phase, or may include the automatic rebuilding of models either on a set schedule or triggered when model monitoring determines that predictions are losing accuracy.

Note: The deployment mechanisms are applicable to Oracle Database, Database Cloud Service (DBCS), and Oracle Autonomous Database or ADB. The graphic uses ADB as an example.

Deployment Technologies

		OML Process	Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment	
Process	Technologies								
Active Roles Data Scientist Business Manager Software Developer Database Admin Data Analyst Domain Expert Machine Learning Engineer Data Engineer		Autonomous Database	Database Cloud Service	On-Premises Database	Additional Oracle Stack Tools • Oracle Analytics Cloud				
	OML for SQL								
	OML for R								
	OML for Python								
	Data Miner								
	OML Notebooks								
	OML AutoML UI								
	OML Services								
	Oracle APEX								