# Purpose

This tutorial covers the use of Oracle Data Miner to perform data mining against Oracle Database 19c. In this workshop, 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 19.2.1.247.

Oracle SQL Developer is a free graphical tool for database development. With SQL Developer, you can browse database objects, run SQL statements and SQL scripts, and edit and debug PL/SQL statements. Starting with SQL Developer, version 3.0, you can also access the Oracle Data Miner GUI, which provides a tightly integrated interface to Oracle Data Mining features.

# Time to Complete

Approximately 30 mins.

# Overview

Data mining is the process of extracting useful information from masses of data by extracting patterns and trends from the data. Data mining can be used to solve many kinds of business problems, including:

* Predict individual behavior, for example, the customers likely to respond to a promotional offer or the customers likely to buy a specific product (Classification)
* Find profiles of targeted people or items (Classification using Decision Trees)
* Find natural segments or clusters (Clustering)
* Identify factors more associated with a target attribute (Attribute Importance)
* Find co-occurring events or purchases (Associations, sometimes known as Market Basket Analysis)

**Problem Definition and Business Goals**

When performing data mining, the business problem must be well-defined and stated in terms of data mining functionality. For example, retail businesses, telephone companies, financial institutions, and other types of enterprises are interested in customer “segmentation” – that is, the act of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits.

The statement “I want to use data mining to study the retail data” is much too vague. From a business point of view, we need specific questions that needs to be answered from the data available to us. The questions must be relevant to the business and should be able to positively influence the business revenues or returns.

This is a transactional level data set, which contains all the transactions occurring between 01/12/2018 and 09/12/2019 for a US-based retail store.

Data is stored in multiple sources (RDBMS, XML and JSON) and we need to combine them into a single data source (CSV) to gather the insights

The company mainly sells unique all-occasion gifts.

Many customers of the company are wholesalers.

# Prerequisites

Before starting this tutorial, you should:

|  |  |
| --- | --- |
| **1.** | Have access to or have Installed Oracle Database 19c Enterprise Edition High Performance, Release 19.8.0.0.0 with Data Mining Option. |
| **2.** | Have access to or have installed Oracle SQL Developer, version 3.0, or later.  Note: SQL Developer does not have an installer. To install SQL Developer, just unzip the free SQL Developer download to any directory on your PC. |
| **3 .** | Have set up Oracle Data Miner 19c for use within Oracle SQL Developer.  Note: If you have not already set up Oracle Data Miner, complete the following lesson: [Setting Up Oracle Data Miner 19c Release 2](https://docs.oracle.com/en/database/oracle/oracle-database/19/tutorial-install-dm-repo/). |

# **Create a Data Miner Project**

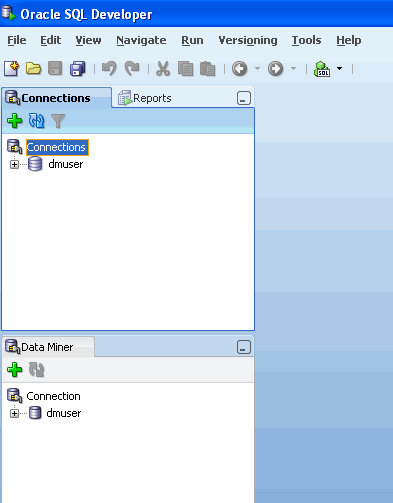
Before you create a Data Miner Project and build a Data Miner workflow, it is helpful to examine some of the Data Miner interface components within SQL Developer. You can then structure your working environment to provide simplified access to the necessary Data Miner features.

**Identifying SQL Developer and Data Miner Interface Components**

After setting up Oracle Data Miner for use within SQL Developer, different interface elements may be displayed, including both SQL Developer and Data Miner components.

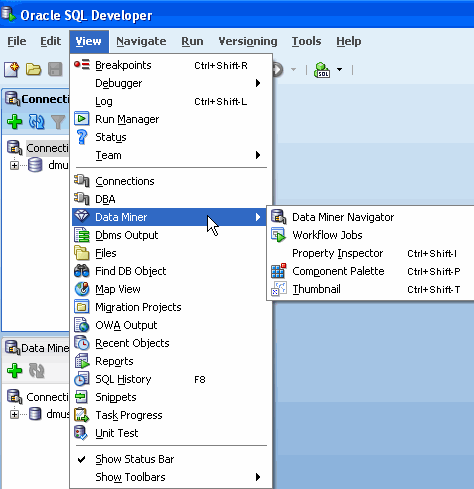
In the following example, several display elements are open, including the:

* The SQL Developer Connections tabs
* The SQL Developer Reports tabs
* The Data Miner tab



Notes:

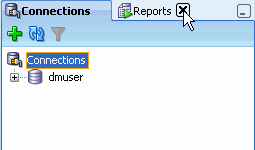
* The layout and contents of your SQL Developer window may be different than the example shown above.
* SQL Developer and Data Miner interface elements open automatically when needed.
* Additional Data Miner interface elements include the Workflow Jobs, Property Inspector, Component Palette, and Thumbnail tabs. You can open any one them manually from the main menu by using **View** > **Data Miner** > as shown here:

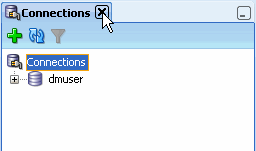


In order to simplify the interface for Data Mining development, you can dismiss the SQL Developer specific interface elements by clicking on the respective Close **[x]** icons for each tab or window.

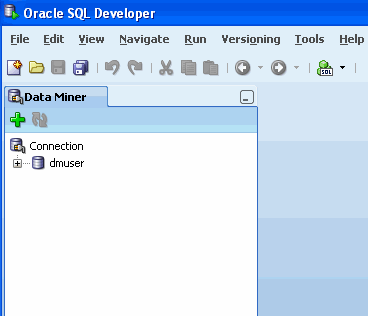
For example, close both of the SQL Developer tabs mentioned above:

1. SQL Developer **Reports** tab
2. SQL Developer **Connections** tab

1. 

2. 

Now, the SQL Developer interface should look like this:



**DataMiner > Workflow Jobs**.

Note: You can re-open the SQL Developer Connections tab (and other interface elements) at any time by using the View menu.

## Step1: Start creating a Data Miner Project

Before you begin working on a Data Miner Workflow, you must create a Data Miner Project, which serves as a container for one or more Workflows.

In the tutorial **Setting Up Oracle Data Miner 11g Release 2**, you learned how to create a database account and SQL Developer connection for a data mining user named **dmuser**. This user has access to the sample data that you will be mining.

Note: If you have not yet set up Oracle Data Miner, or have not created the data mining user, you must first complete the tasks presented in the tutorial [Setting Up Oracle Data Miner 19c Release 2](https://docs.oracle.com/en/database/oracle/oracle-database/19/tutorial-install-dm-repo/),

To create a Data Miner Project, perform the following steps:

|  |  |
| --- | --- |
| **1.** | In the Data Miner tab, right-click the data mining user connection that you previously created, and select **New Project**, as shown here:  Screenshot for Step |
| **2.** | In the Create Project window, enter a project name (in this example Regression) and then click **OK**.    Note: You may optionally enter a comment that describes the intentions for this project. This description can be modified at any time.  Result: The new project appears below the data mining user connection node. |

# Build a Data Miner Workflow

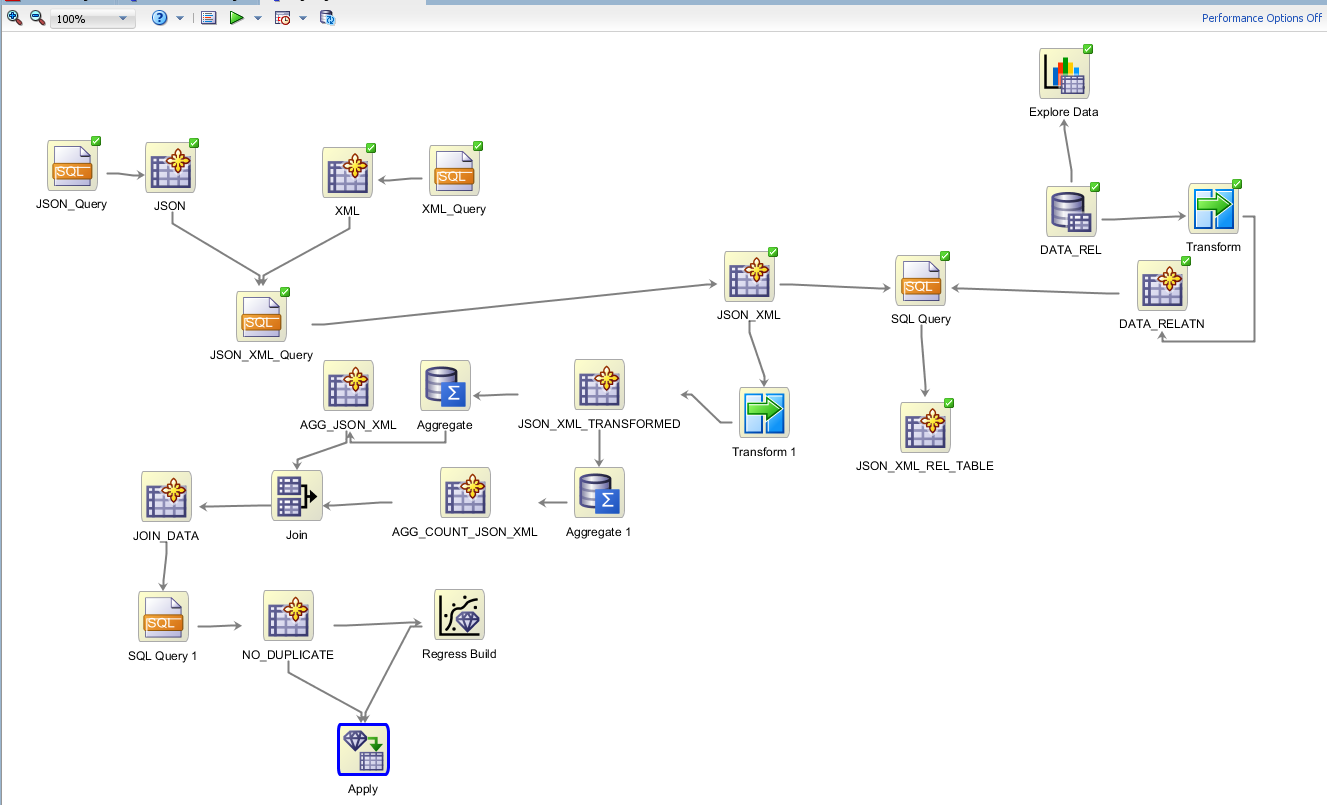
A Data Miner Workflow is a collection of connected nodes that describe a data mining processes.

A workflow:

* Provides directions for the Data Mining server. For example, the workflow says, "Build a model with these characteristics." The data-mining server builds the model with the results returned to the workflow.
* Enables you to interactively build, analyze, and test a data mining process within a graphical environment.
* Might be used to test and analyze only one cycle within a particular phase of a larger process, or it may encapsulate all phases of a process designed to solve a particular business problem.

## What Does a Data Miner Workflow Contain?

Visually, the workflow window serves as a canvas on which you build the graphical representation of a data mining process flow, like the one shown here:



Notes:

* Each element in the process is represented by a graphical icon called a node.
* Each node has a specific purpose, contains specific instructions, and may be modified individually in numerous ways.
* When linked together, workflow nodes construct the modeling process by which your particular data mining problem is solved.

As you will learn, any node may be added to a workflow by simply dragging and dropping it onto the workflow area. Each node contains a set of default properties. You modify the properties as desired until you are ready to move onto the next step in the process.

## Step 2: Sample Data Mining Scenario

In this topic, you will create a data mining process that predicts the revenue of customers.

To accomplish this goal, you build a workflow that enables you to:

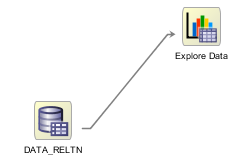
* Build and compare several Classification models
* Select and run the models that produce the most actionable results

To create the workflow for this process, perform the following steps.

## Step 3: Create a Workflow and Add data for the workflow

|  |  |
| --- | --- |
| **1.** | Right-click your project (Retail\_Data\_Analysis) and select **New Workflow** from the menu.    Result: The Create Workflow window appears. |
| **2.** | In the Create Workflow window, enter **Predicting\_Customer\_Value** as the name and click **OK**.    Result:   * In the middle of the SQL Developer window, an empty workflow canvas opens with the name that you specified. * On the right-hand side of the interface, the Component Palette tab of the Workflow Editor appears (shown below with a red border). * In addition, three other Oracle Data Miner interface elements are opened:   + The Thumbnail tab   + The Workflow Jobs tab   + The Property Inspector tab |
| **3.** | The first element of any workflow is the source data. We will extract data from a JSON table and a XM table. Here, we cannot directly add a data source. We will use a query editor to read the tables in a relational table format. You add a Data Source node to the workflow, and select the JSON\_PURCHASEORDER and XML\_PURCHASEORDER tables as the data source.  A. In the Component Palette, click the **Data** category. A list of data nodes appear, as shown here:    B |
|  | 1. We will add a RDBMS regular table data, using a data source. This is the JSON and XML combined table ()   Drag and drop the **Data Source** node onto the Workflow pane.  Result: A Data Source node appears in the Workflow pane and the Define Data Source wizard opens.    In Step 1 of the wizard:   1. Select FINAL\_JSON\_XML\_DATA from the Available Tables/Views list, as shown here:     Note: You may use the two tabs in the bottom pane in the wizard to view and examine the selected table. The Columns tab displays information about the table structure, and the Data tab shows a subset of data from the selected table or view.   1. Click Next to continue.   5. In Step 2 of the wizard, you may remove individual columns that you do not need in your data source. In our case, we will keep all of the attributes that are defined in the table.  At the bottom of the wizard window, click Finish.    Result: As shown below, the data source node name is updated with the selected table name, and the properties associated with the node are displayed in the Property Inspector, located below the Component Palette pane.    Notes:   * You can resize nodes in the workflow canvas by entering or selecting a different value from the Zoom options. Notice that 75% has been selected from the Zoom pull-down list. * You can add descriptive information about any node by using the Details tab in the Property Inspector. * The Thumbnail tab also provides a smaller display of the larger workflow window. As you drag nodes around the workflow window, the thumbnail view automatically adjusts.  Examine the Source Data You can use an Explore Data node to examine the source data. Although this is an optional step, Oracle Data Miner provides this tool to enable you to verify if the selected data meets the criteria to solve the stated business problem.  Follow these steps:   1. Drag and drop the Explore Data node from the Component Palette to the Workflow, like this:     Result: A new Explore Data node appears in the workflow pane, as shown here. (As before, a node name is automatically generated.)  Notes:   * A yellow Information (!) icon in the border around any node indicates that it is not complete. Therefore, at least one addition step is required before the Explore Data node can be used. * In this case, a data source node must be "linked" to the Explore Data node to enable further exploration of the source data.  1. To link the data source and explore data nodes, use the following instructions:   A. Right-click the data source node (DATA\_RELTN), select **Connect** from the pop-up menu, and then drag the pointer to the Explore Data node, as shown here: |

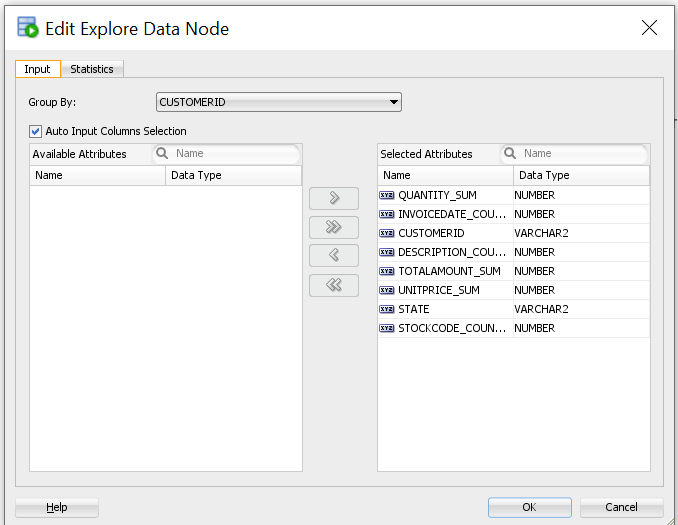
1. Then, click the**Explore Data** node to connect the two nodes. The resulting display looks like this:



1. Next, select a "Group By" attribute for the data source.

A. Double-click the Explore Data node to display the Select Attributes window.

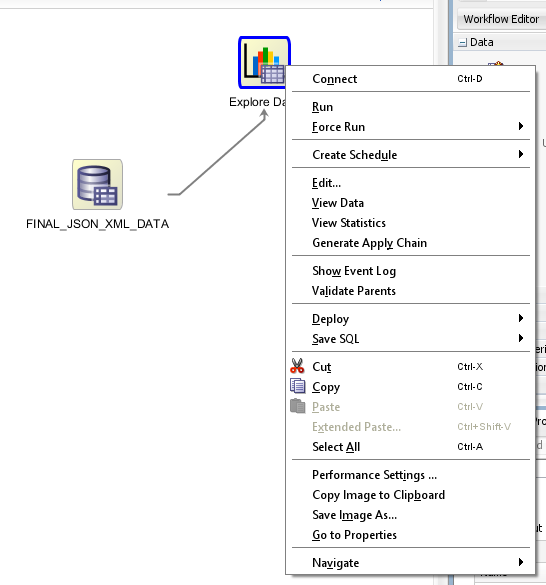
B. In the Group By list, select the CUSTOMER\_ID attribute, as shown here:



C. Then, click OK.

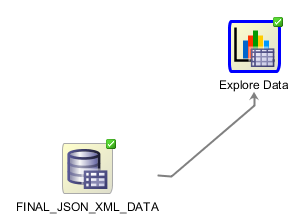
Note: The Select Attributes window also allows you to remove (or re-add) any attributes from the source data.

4. Next, right-click the Explore Data node and select Run.



Result:

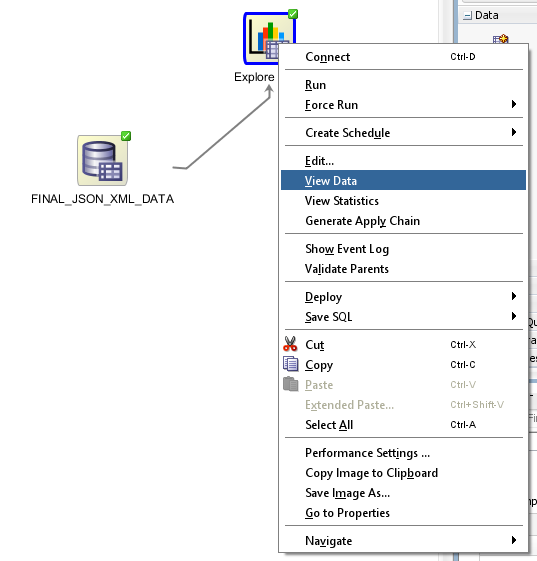
* Data Miner displays status information in the Workflow Jobs tab while processing the node.
* When the update is complete, the data source and explore data nodes show a green check mark in the borders, like this:



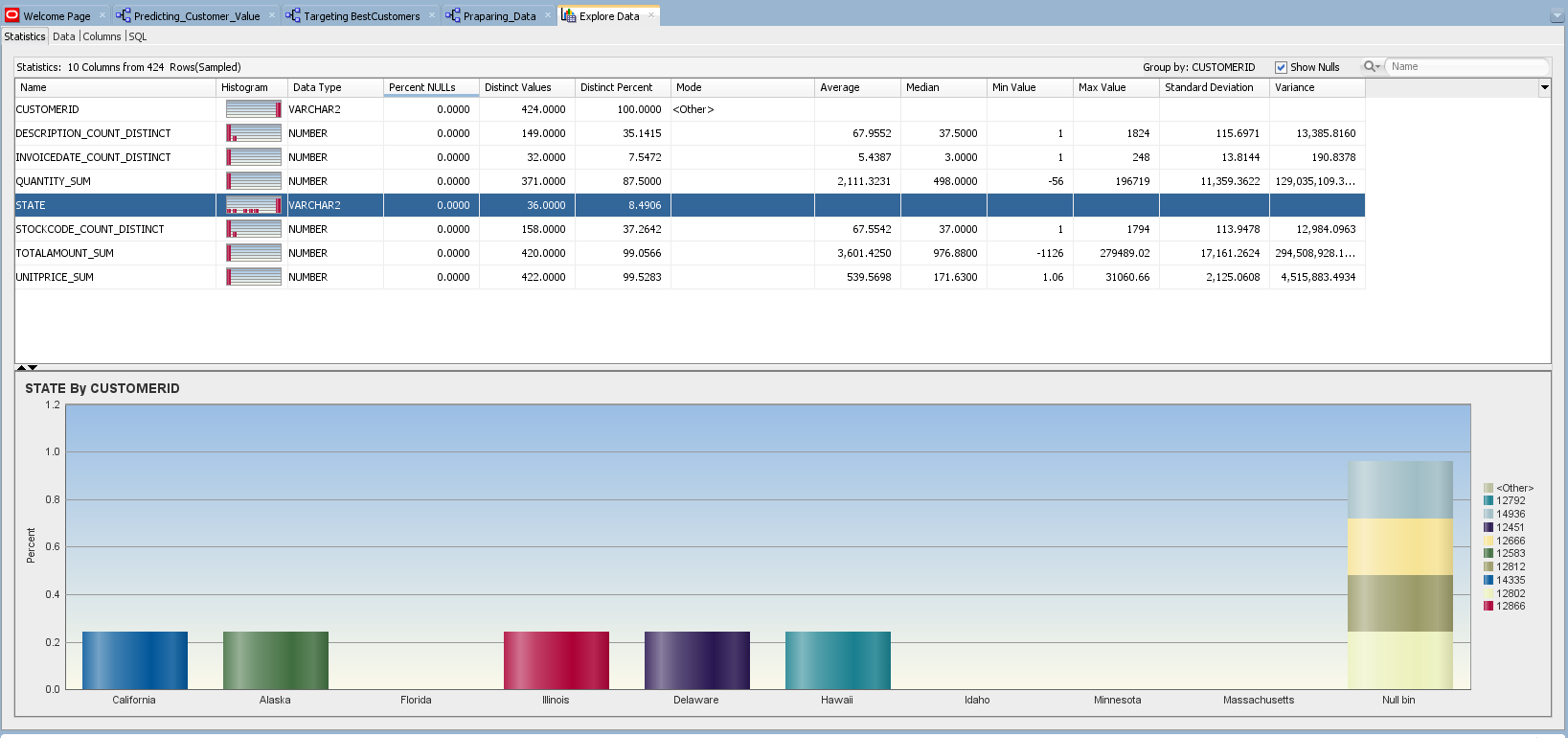
Note: When you run any process from the workflow canvas, the steps that you have specified are executed by the Oracle Data Miner Server.

5. To see results from the Explore Data node, perform the following:

A. Right-click the Explore Data node and select View Data from the menu



Result: A new tab opens for the data profile node, as shown below



Notes:

* Data Miner calculates a variety of information about each attribute in the data set, as it relates to the "Group By" attribute that you previously defined, including a Histogram, Distinct Values, Mode, Average, Min and Max value, Standard Deviation, Variance, Skewness, and Kurtosis.
* The display enables you to visualize and validate the data, and also to manually inspect the data for patterns or structure.

B. Select any of the attributes in the Name list to display the associated histogram in the bottom window.

C. When you are done examining the source data, dismiss the Explore Data tab by clicking the Close icon (X).

Next, you move from a high-level manual analytic exercise to using the power of database data mining.

6. Save the workflow by clicking the Save All icon in main toolbar.



## Step 4: Build the Models

In this topic, you build the selected models against the source data. This operation is also called “training” a model, and the model is said to “learn” from the training data.

A common data mining practice is to build (or train) your model against part of the source data, and then to test the model against the remaining portion of your data. By default, Oracle Data Miner this approach.

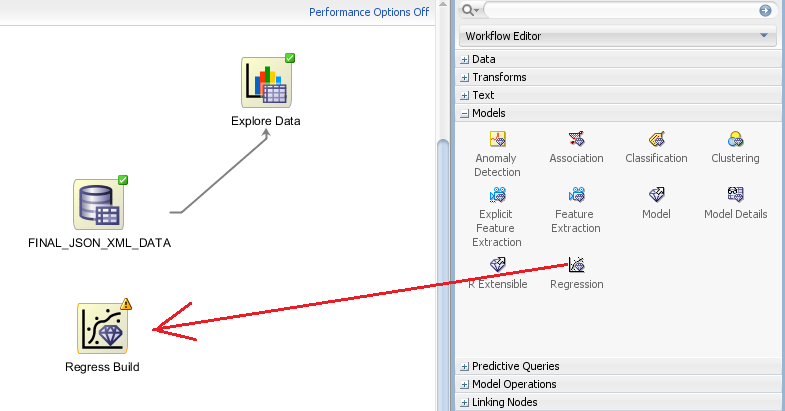
The models have the same build data and the same target.

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.

In the Workflow Editor, expand Models, and click Regression.

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.



Right-click, and click Connect the data to regression node. Drag the line to the Regression node and click again

The **Edit Regression Build Node** dialog box opens.

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:

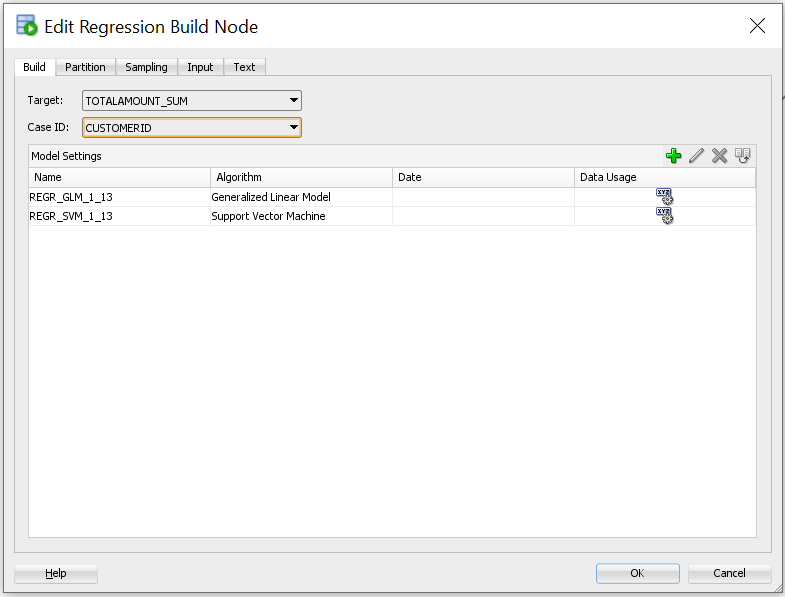
* In the Target field, select a target from the drop-down list. The list consist of attributes from the table or view specified in the Data Source node that is connected to the build node. For our use case, we select the TOTALAMOUNT\_SUM column

You must specify a target. All models in the node have the same target.

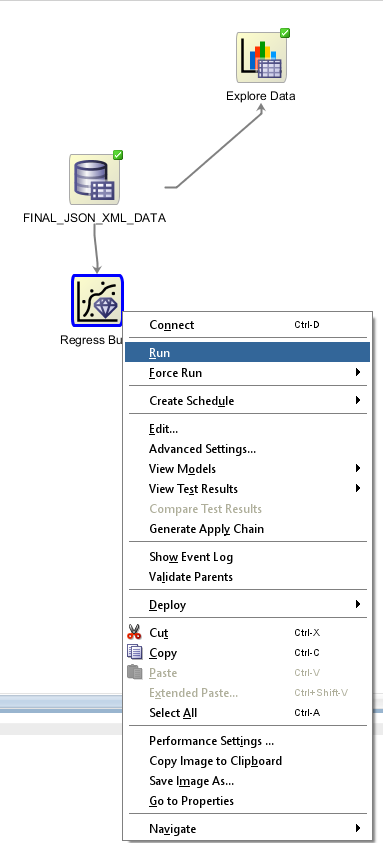
* In the Case ID field, select one attribute from the drop-down list. This attribute must uniquely identify a case. We will select the column CUSTOMERID.

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 **x**. We will keep both the models.



* Right Click on the **Regression** node and select **Run**

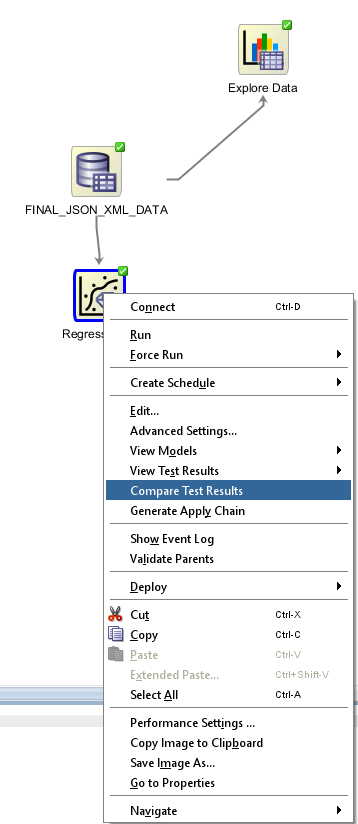


## Step 5: Compare the Models

After you build/train the selected models, you can view and evaluate the results for all of the models in a comparative format. Here, you compare the relative results of all four classification models.

Follow these steps:

Right-click the classification build node and select **Compare Test Results** from the menu.



Results: A Class Build display tab opens, showing a graphical comparison of the four models, as shown here:

The **Performance** tab displays the test results for several common test metrics. The **Performance tab** for **Regression** model show these measures for all models:

By default, All Measures are displayed. The selected measures are displayed as graphs. Since you are comparing test results for two models, then the different models have graphs in different colors.

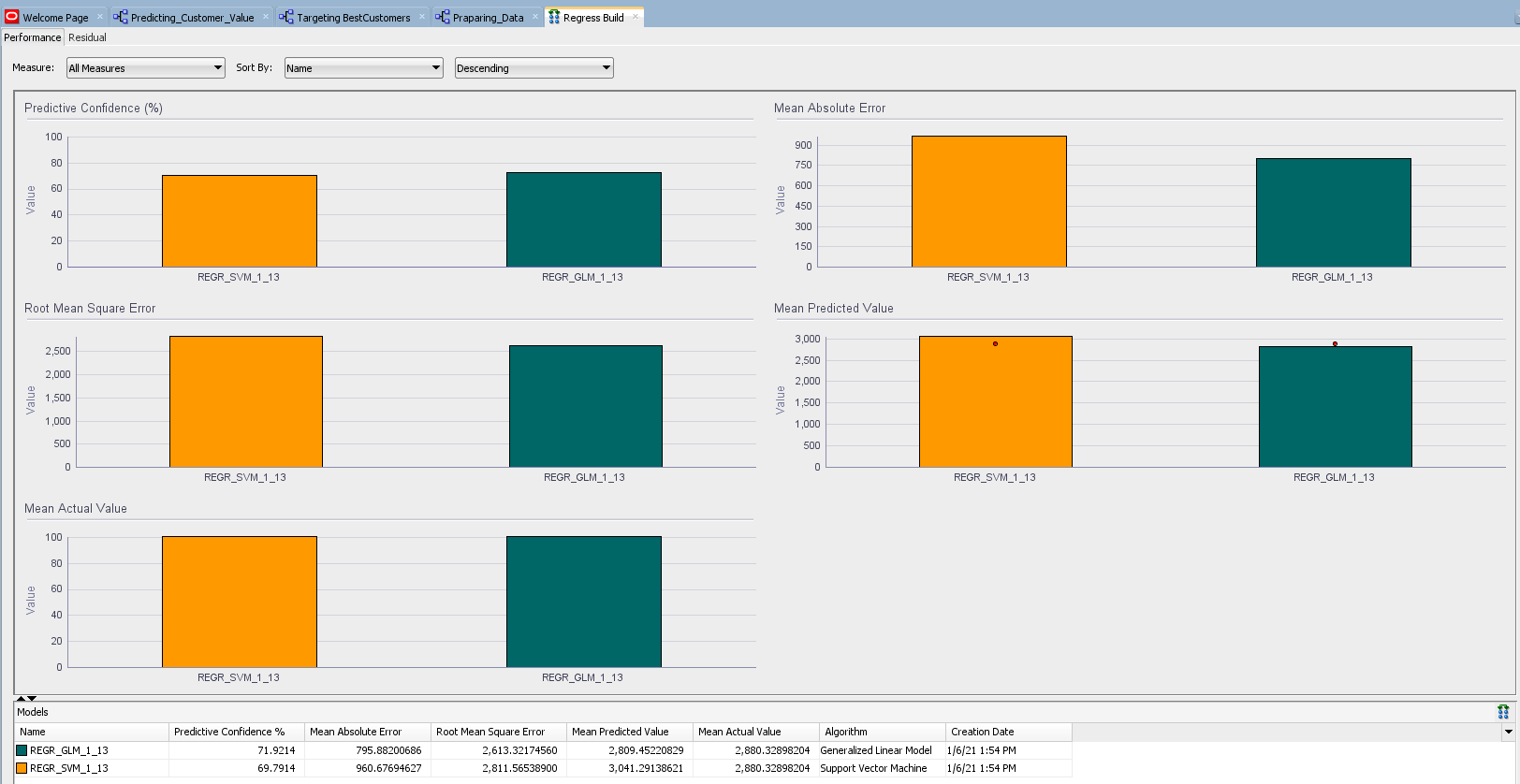
**Predictive Confidence**: Predictive Confidence provides an estimate of accurate the model is. Predictive Confidence is a number between 0 and 1. Oracle Data Miner displays Predictive Confidence as a percentage. For example, the Predictive Confidence of 59 means that the Predictive Confidence is 59 percent (0.59). Measures how much better the predictions of the mode are than those of the naive model.

**Mean Absolute Error**: The Mean Absolute Error (MAE) is the average of the absolute value of the residuals (error).

**Root Mean Square Error**: The Root Mean Squared Error (RMSE) is the square root of the average squared distance of a data point from the fitted line.

**Mean Predicted Value**: The average of the predicted values.

**Mean Actual Value**: The average of the actual values.



The **Residual Plot tab** shows the Residual Plot for each model. You can compare two plots side by side.

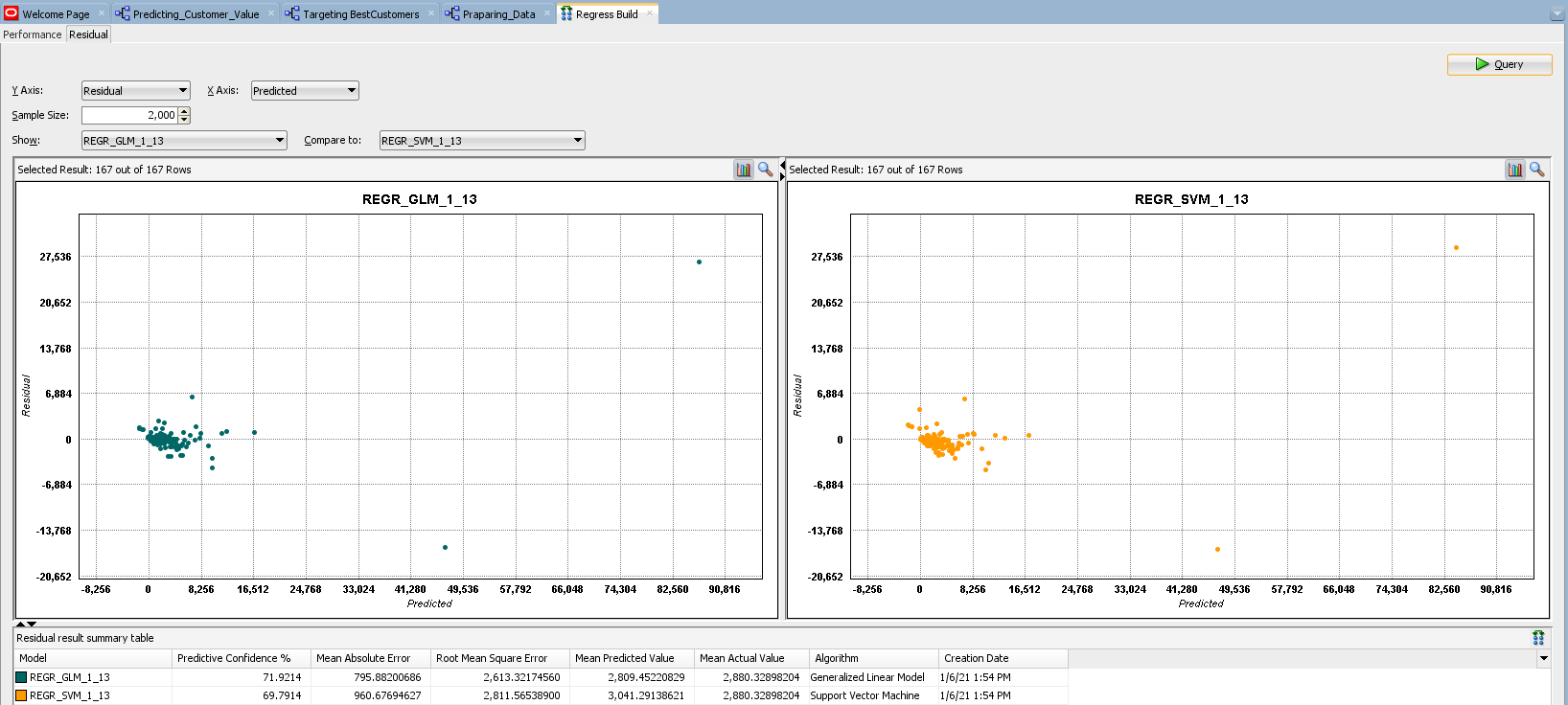
The **residual plot** is a scatter plot of the residuals. Each residual is the difference between the actual value and the value predicted by the model. Residuals can be positive or negative. If residuals are small (close to 0), then the predictions are accurate. A residual plot may indicate that predictions are better for some classes of values than others.

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:

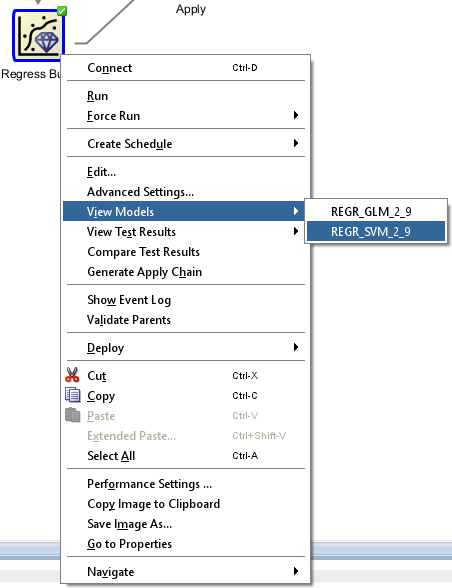
* Model, the name of the model along with color of the model in the graphs
* Predictive Confidence
* Mean Absolute Error
* Root Mean Square Error
* Mean Predicted Value
* Mean Actual Value
* Algorithm
* Creation Date (and time)

By default, results for all models in the node are displayed

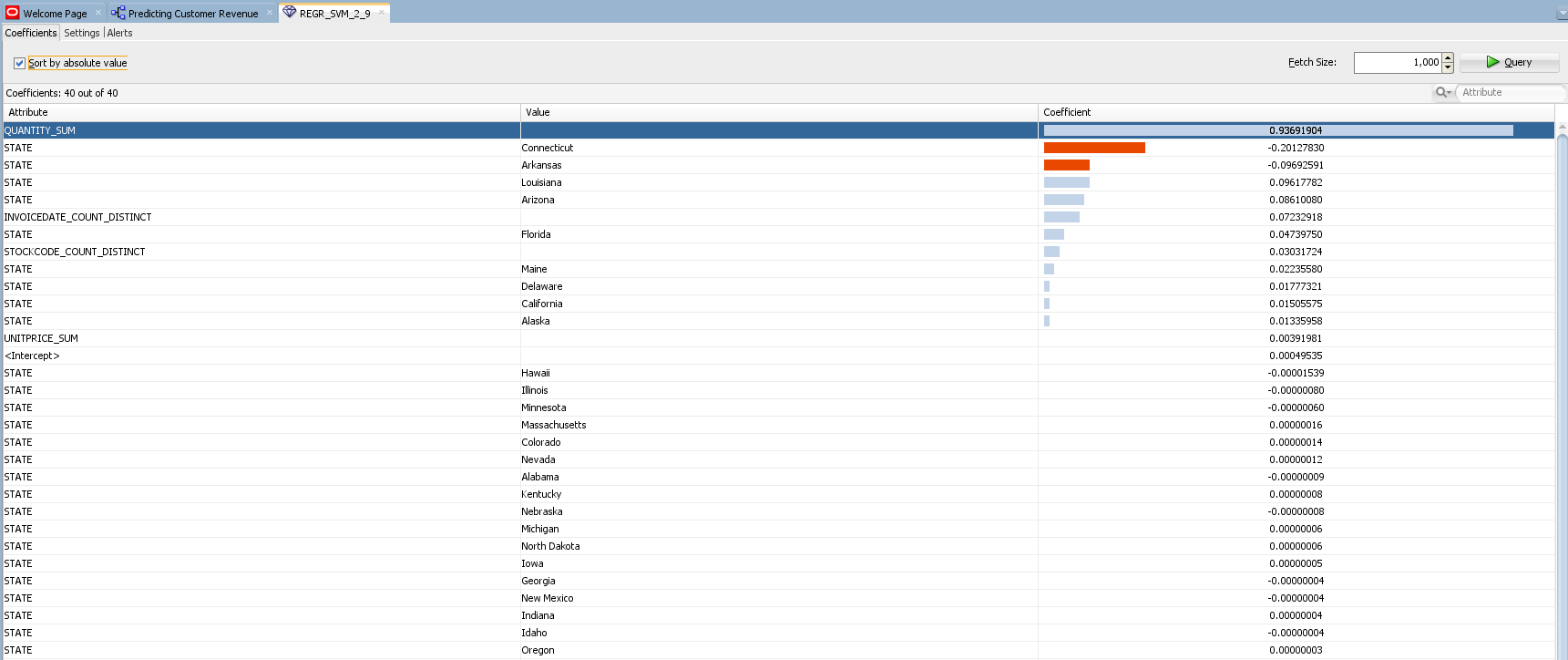


## Step 6: Select and Examine a Specific Model

Using the analysis performed in the past topic,



The **Coefficients** Tab shows the Regression Coefficients. When you build a multivariate linear regression model, the algorithm computes a coefficient for each of the predictors used by the model. The coefficient is a measure of the impact of the predictor x on the target y.



## Step 7: Apply the Model

In this topic, you apply the Regression model and then create a table to display the results. You "apply" a model in order to make predictions - in this case to predict which customers are likely to buy insurance.

To apply a model, you perform the following steps:

First, specify the desired model (or models) in the Class Build node.

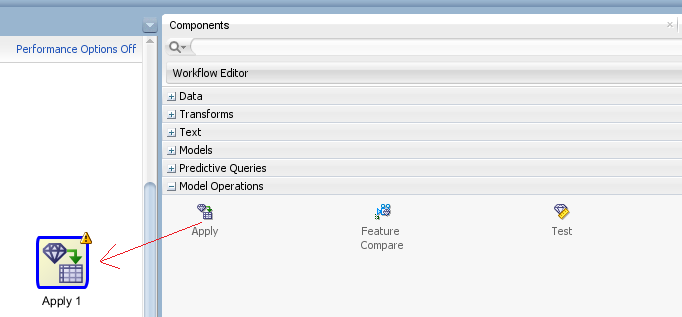
Then, you add a new Data Source node to the workflow. (This node will serve as the "Apply" data.)

Next, add an Apply node to the workflow.

Finally, you link both the Class Build node and the new Data Source node to the Apply node.

Follow these steps to apply the model and display the results:

In the Workflow Editor, expand **Model Operations**, and click **Apply**. Drag and drop the **Apply** node in the Workflow pane.

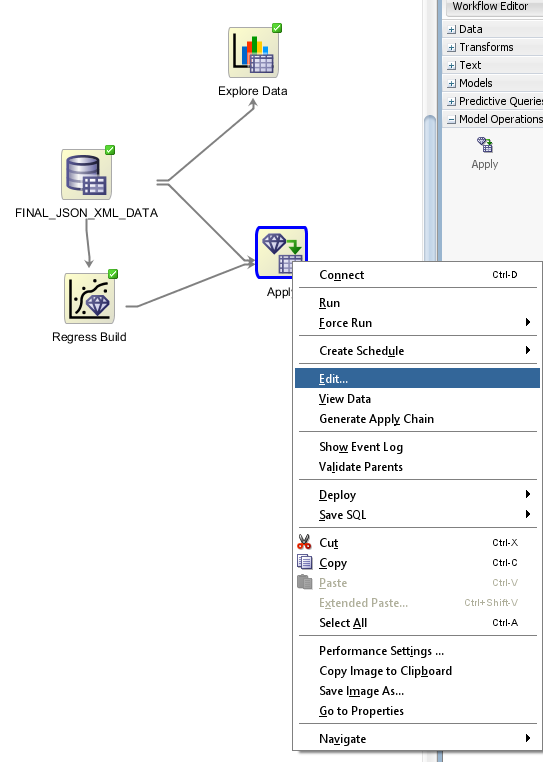
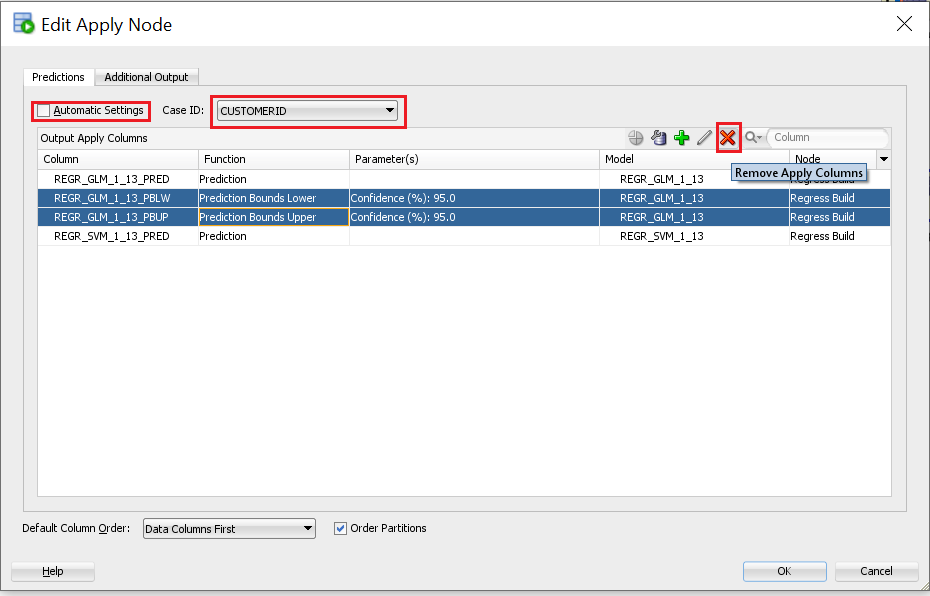


Link the Data node, Model nodes, and Build nodes to the Apply Node. Since we do not have a separate dataset for test data, we will connect the same dataset to the apply node. Alternatively, you can use the sample node to create a separate train data and the remaining data can be used as test data.

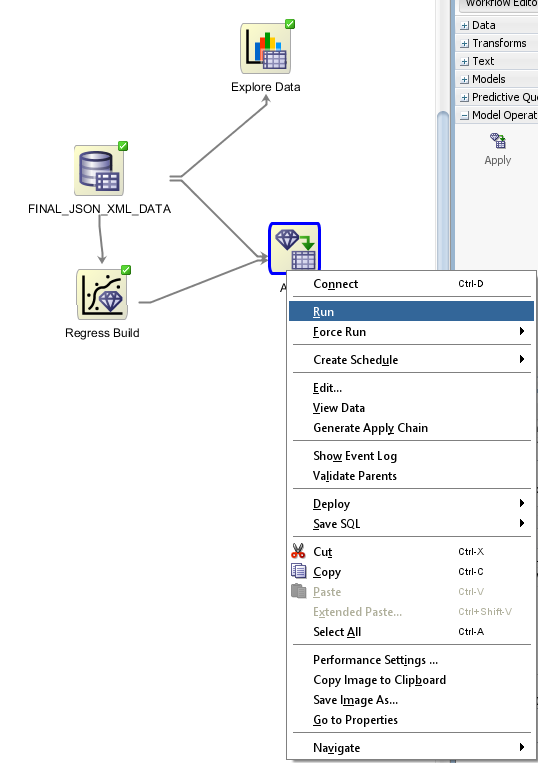
The Apply node takes a collection of models and returns a single score. The Apply node produces a query as the result. The result can further transformed or connected to a Create Table or View node to save the data as a table.

To make predictions using a model, you must apply the model to new data. This process is also called scoring the new data.

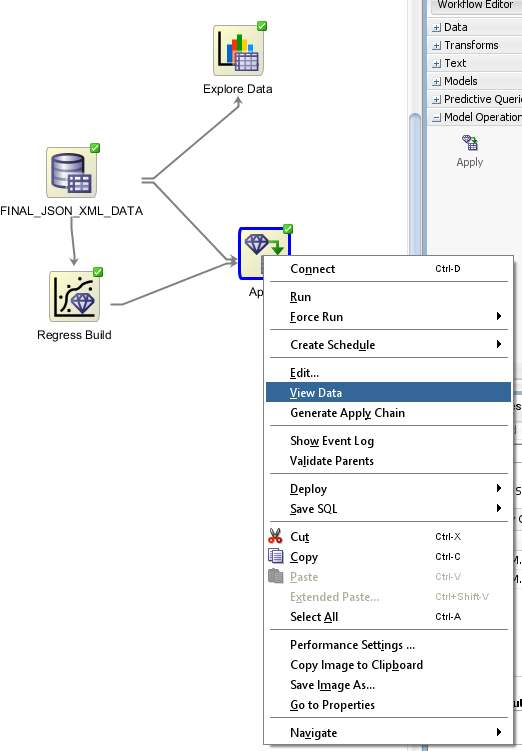
Right Click and select **Edit** in the context menu. In the **Edit Apply Node** wizard, uncheck automatic settings. Select CUSTOMERID as the Case ID. Now delete the columns Prediction Upper Bounds (PBUP) and Prediction Lower Bounds (PBLW) using the **x** symbol as shown in the image below. Click **Ok**.

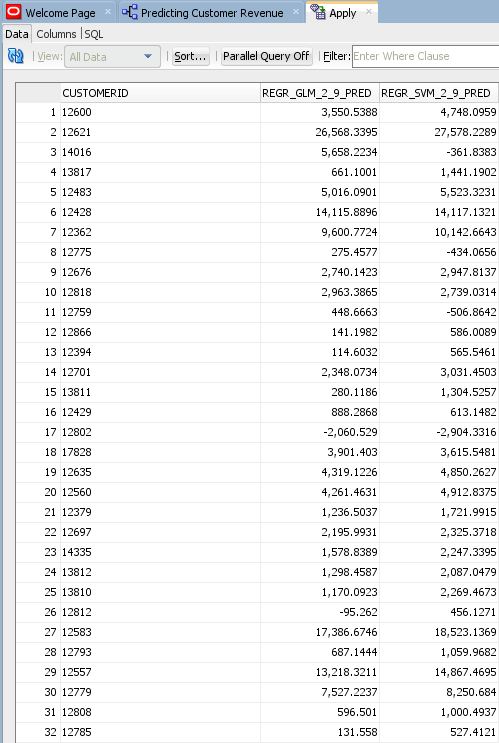
Right Click on Apply node and select Run to execute the node.



After Successful execution, right click on Apply node and select View Data.



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.



# Summary

In this workshop, you examined and solved a "Regression" prediction data mining business problem by using the Oracle Data Miner graphical user interface, which is included as an extension to SQL Developer.

In this tutorial, you have learned how to:

* Identify Data Miner interface components
* Create a Data Miner project
* Build a Workflow document that uses Regression models to predict customer behavior.