THE POWER TO KNOW.

# SAS® Enterprise Miner™ 12.1: JMP Extension Nodes

SAS® Documentation

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2012. *SAS® Enterprise Miner™ 12.1: JMP Extension Nodes*. Cary, NC: SAS Institute Inc.

**SAS® Enterprise Miner™ 12.1: JMP Extension Nodes**



SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, April 2012

# Contents

*Chapter 1*

# Overview of the JMP Extension Node Package for SAS Enterprise Miner



## SAS Enterprise Miner Extension Nodes

Extension nodes provide a mechanism for extending the functionality of a SAS Enterprise Miner installation. Using the SEMMA process, you can develop extension nodes to perform any essential data mining activity. (SEMMA stands for sample, explore, modify, model, or assess.) Although the SAS Enterprise Miner nodes that are distributed by SAS are typically designed to satisfy the needs of a diverse audience, extension nodes provide a means to develop custom solutions.

Developing an extension node is conceptually simple. An extension node consists of the following:

- one or more SAS source code files stored in a SAS library or in external files that are accessible by the SAS Enterprise Miner server
- an XML file defining the properties of the node
- two graphic images stored as .gif files.

When properly developed and deployed, an extension node integrates into the SAS Enterprise Miner workspace so that, from the perspective of the end user, it is indistinguishable from any other node in SAS Enterprise Miner. From a developer's perspective, the only difference is the storage location of the files that define an extension node's functionality and appearance. Any valid SAS language program statement can be used in the source code for an extension node, so an extension node's functionality is virtually unlimited.

## The Basics

There are four extension nodes that run JMP platforms and display results in SAS Enterprise Miner.

- **JMP Boosted Tree** — runs the **Boosted Tree** option of the JMP Partition platform to produce a gradient boosting model. This node requires JMP Pro.

- **JMP Bootstrap Forest** — runs the **Bootstrap Forest** option of the JMP Partition platform. Bootstrap forests are a variation of random forests, a special type of bagging where predictors are selected at random for each split. This node requires JMP Pro.
- **JMP Data Exploration** — opens a SAS Enterprise Miner data set and runs the JMP Graph Builder. By default, it displays a graph of the distribution of each target. But, you can interactively adjust the graphs to show, for example, bivariate relationships between targets and predictors. Or, you can create a new graph or analysis within JMP. For large data sets, you can specify that the **JMP Data Exploration** node should use a sample of the input data set.
- **JMP Neural** — runs the JMP Neural platform, which produces a neural network model. The node has a built-in gradient boosting option. Specify a positive value for **Component Models** to use boosting. This node requires JMP Pro.

All nodes are compatible with SAS Enterprise Miner data partitions, but only the modeling nodes actually honor partitions. The **Data Exploration** node displays unpartitioned results by default. The **JMP Neural** node honors the SAS Enterprise Miner partition, but if you disable partitioning, JMP will always perform **holdback** or **K-fold** validation, determined by the user setting in **Fitting Options**.

## Requirements and Known Issues

The JMP Extension Node package requires the following:

- SAS Enterprise Miner, version 6.2 or later. The modeling nodes are single machine tools. Currently, they do not work in a client/server, multi-machine deployment of SAS Enterprise Miner.
- JMP Pro, Version 9.0.3. There is no error checking to make sure you have JMP Pro 9 on your system. Errors might occur if you use a standard version of JMP, rather than JMP Pro.

When using the JMP Extension Node package, you should be aware of the following issues:

- Modeling Nodes
  - The modeling nodes, JMP Boosted Tree, JMP Bootstrap Forest, and JMP Neural, require JMP Pro. They fail with error messages when standard JMP is used.
  - The nodes work in a SAS Enterprise Miner client/server configuration, provided that JMP Pro, the SAS Server, and the SAS Enterprise Miner client are all on the same machine. This includes SAS Enterprise Miner desktop, SAS Enterprise Miner classroom, SAS Enterprise Miner workstation, or a three-tier, single-machine install of SAS Enterprise Miner. The modeling nodes do not currently run in a multi-machine deployment of SAS Enterprise Miner.
  - The modeling nodes accept a single target variable only. If there are more targets, an error condition is triggered.
- SAS Syntax Errors
  - SAS syntax errors can occur if the input data set has many predictors, predictors with long names, or both. If you see syntax errors along with a **Truncated Record** warning in the SAS log, set a high value for the LRECL option in your project start-up code. Here is an example: `OPTIONS LRECL=5000;`.

    - SAS syntax errors can occur with long variable names (approximately 25 characters or more) and string data values that have embedded quotation marks, such as "isn't". If you encounter syntax errors, try using shorter variable names, and transform any input strings that have embedded quotation marks or apostrophes.

- Results

    - Under certain conditions, there can be discrepancies between the results reported by JMP and those reported by SAS Enterprise Miner. To avoid these discrepancies, set the **Number of Terms** property to its default value of **1** for Bootstrap Forest models, impute missing data values, or both. Specifically, if you have missing predictor values, use the default value for the **Number of Terms** property. If you need to change the **Number of Terms** property, then impute missing predictor values before you run the node.

*Chapter 2*

# Using the JMP Extension Node Package

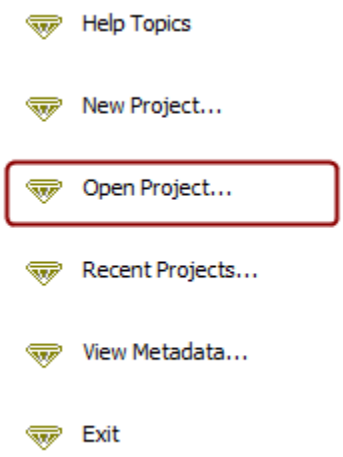


## Getting Started

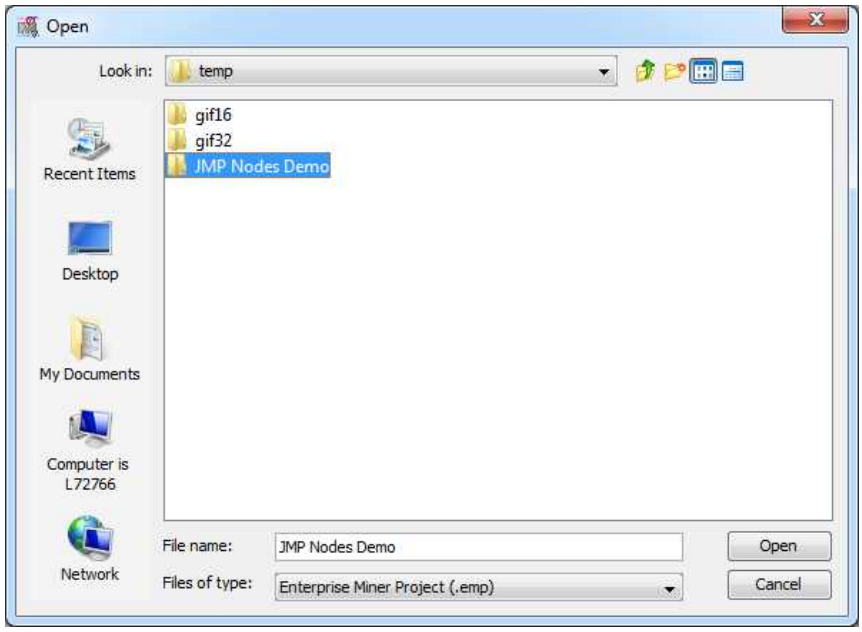
Use the provided SAS Enterprise Miner project, named JMP Nodes Demo, to explore and familiarize yourself with the nodes

Follow these steps to use the sample project:

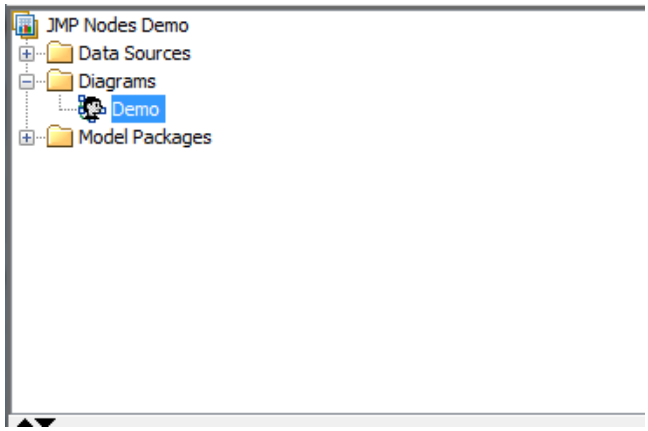
1. Start SAS Enterprise Miner.
2. From the SAS Enterprise Miner home screen, select **Open Project**.



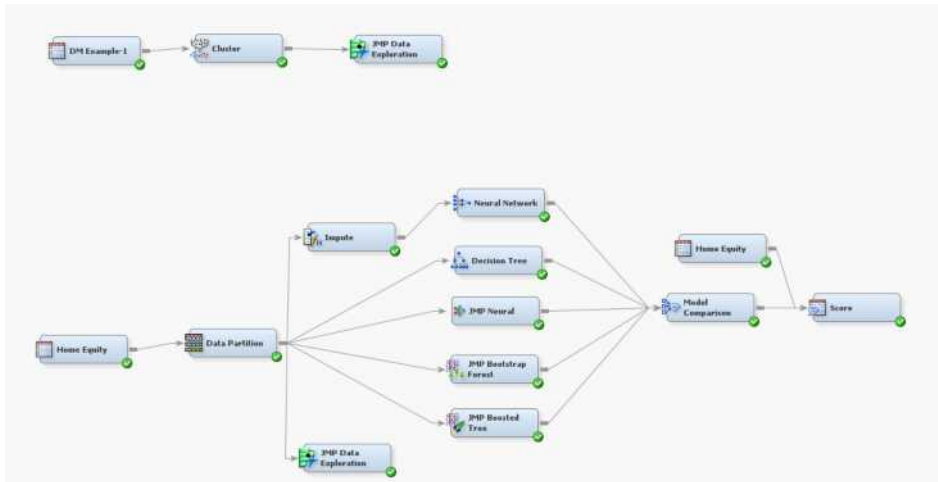

3. Navigate to the folder that contains the unzipped project.
4. Select the **JMP Nodes Demo** folder and select **Open** to load the project.




5. Open the **Demo** diagram that is listed in the Project Panel.



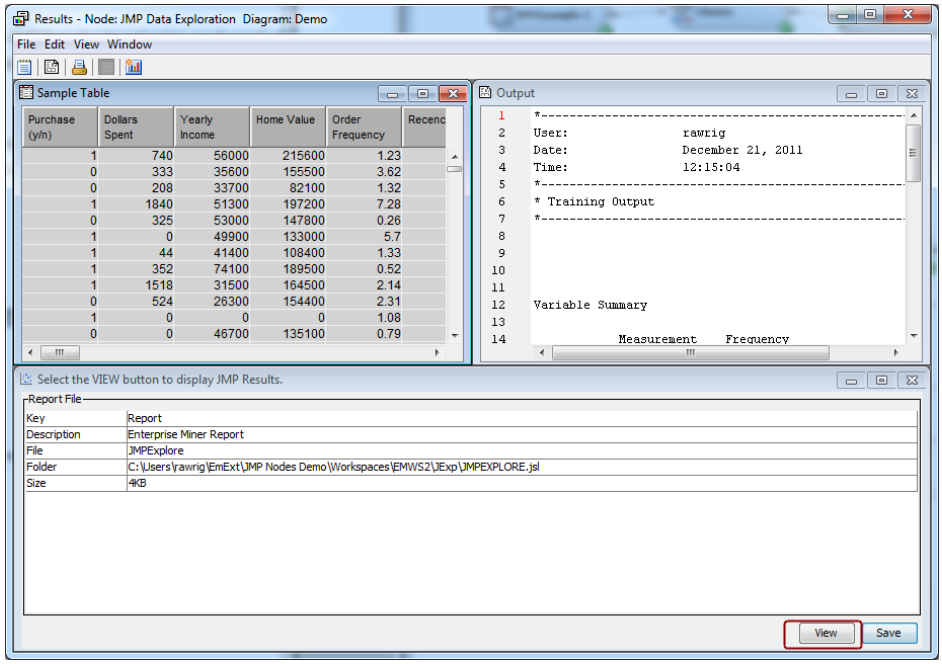

6. The full process flow diagram is presented in the image below. This process flow diagram contains two process flows, the **DM Example-1** branch and the **Home Equity** branch, named for the first node in each process flow. You will examine the results of each branch in the sections that follow.



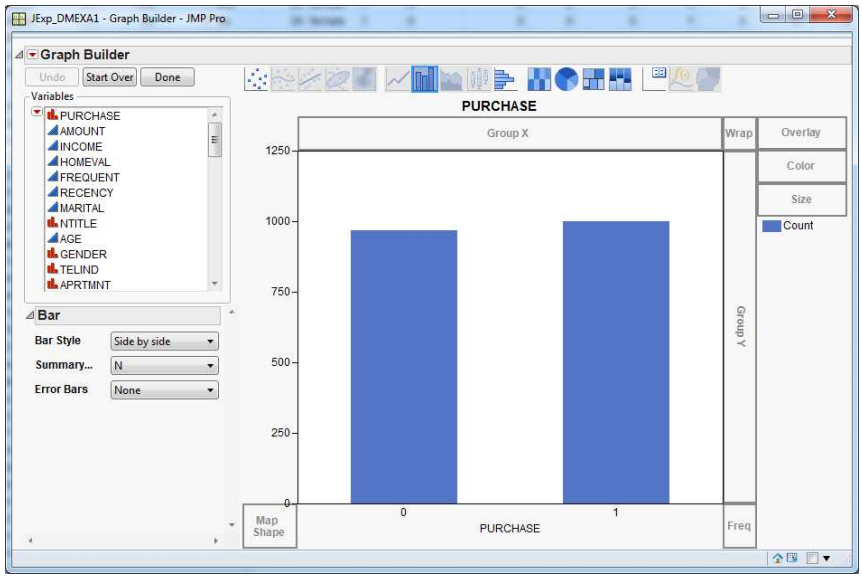
# Cluster Profiling

To explore clusters, use the **DM Example-1** branch, which starts with the **DM Example-1** node and ends with the **JMP Data Exploration** node. The **JMP Data Exploration** node explores clusters using the JMP Graph Builder and the mapping capabilities provided by JMP. Follow the steps below to examine the results of the **DM Example-1** branch.

1. Select the **JMP Data Exploration** node. Right-click the node and select **Run**. In the Confirmation window, select **Yes**.
2. After the process flow diagram has successfully run, select **Results** in the Run Status window.
3. In the Results window, click the **View** button.

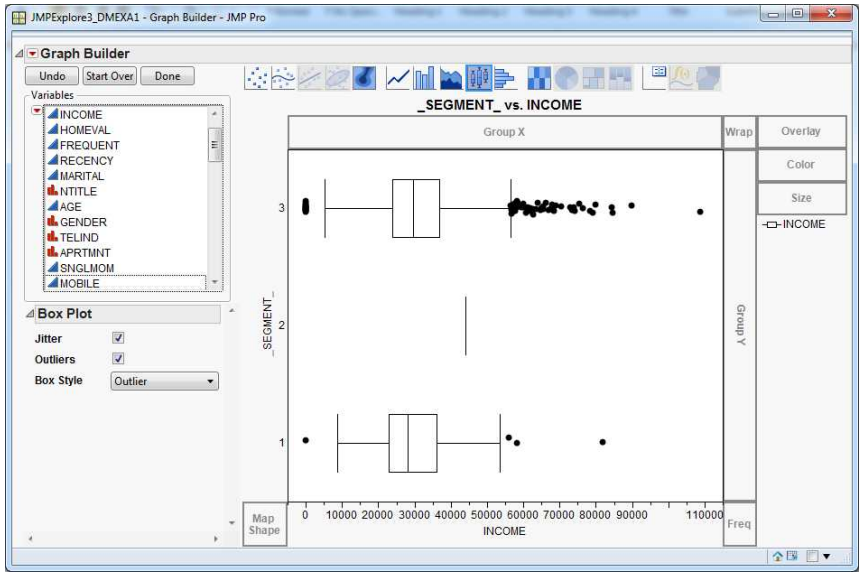



4. By default, the target variable is plotted in the Graph Builder. In this example, the target variable is PURCHASE.

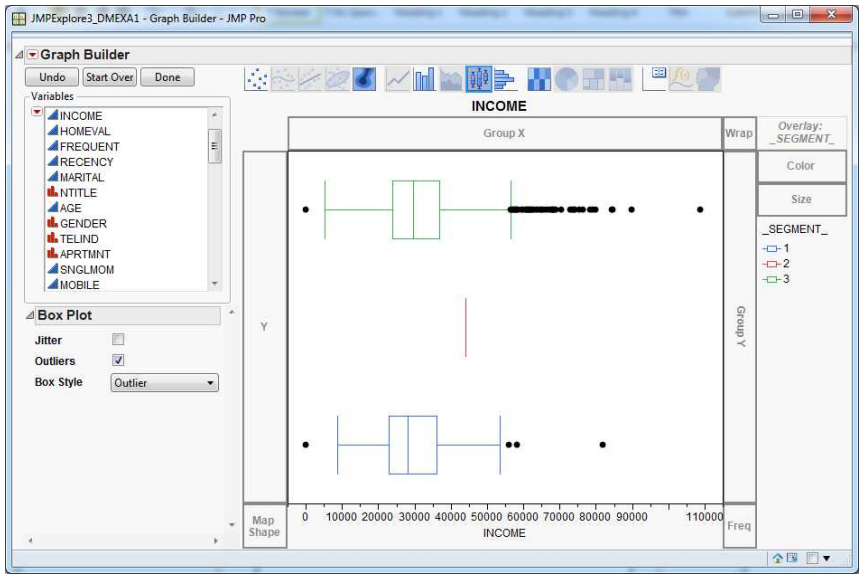



5. To create a segment plot, first click the **Start Over** button in the upper left corner of the Graph Builder window.

6. In the **Variables** list, select **_Segment_** and drag it to the Y-axis. Next, select the variable **INCOME** and drag it to the X-axis. You should see a cluster profile plot that is similar to the one below.

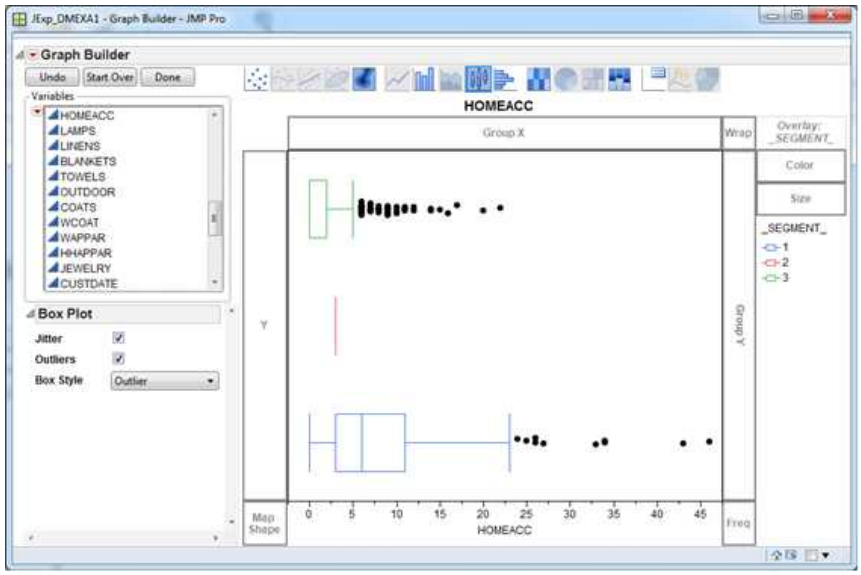



Alternatively, if you select the **Overlay** option in the upper right corner of the Graph Builder window, the _SEGMENT_ variable is used as an overlay variable.




Both versions of the chart suggest that income is not a strong differentiator of the clusters. Segments 1 and 3 have similar interquartile ranges. Segment 2 has a higher median than the other segments, but it represents only one customer.

The Home Furniture variable (HOMEACC) is a better differentiator of segments 1 and 3. Segment 1 has a median of 6 home accessories, whereas segment 3 has a median of 0.

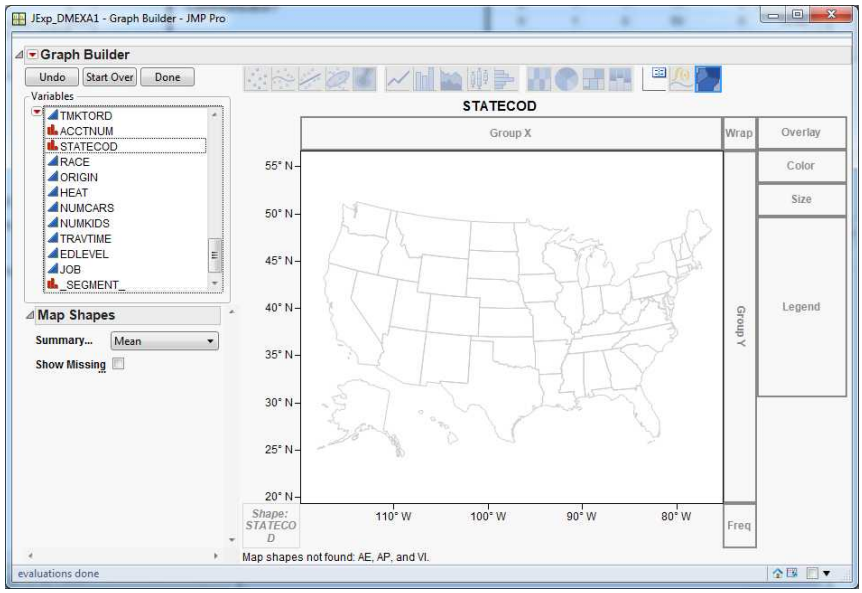



## JMP Map Capabilities

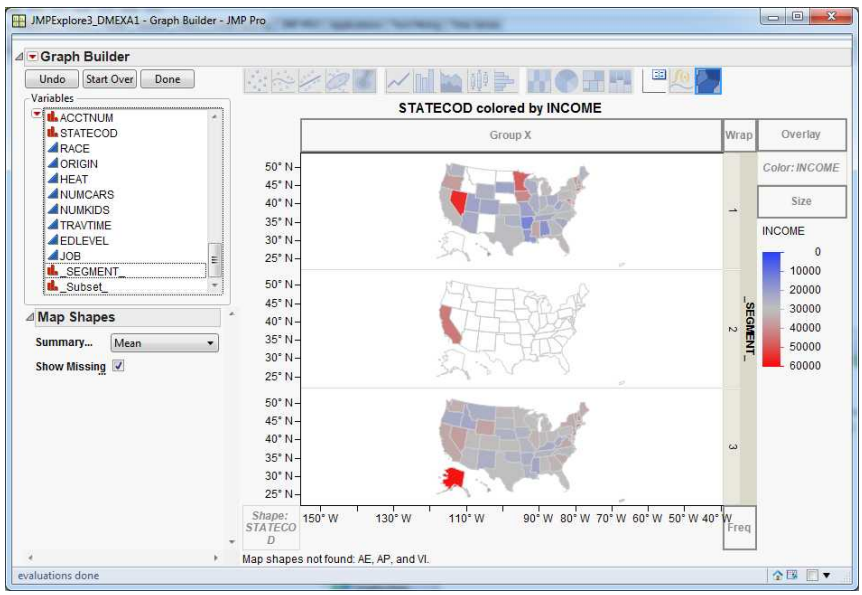
The JMP Graph Builder enables you to overlay data onto a map. In this example, you plot average income on a state-by-state basis over a map of the USA.

1. Select the **JMP Data Exploration** node. Right-click the node and select **Run**. In the Confirmation window, select **Yes**.

2. After the process flow diagram has successfully run, select **Results** in the Run Status window.
3. In the Results window, click the **View** button.
4. By default, the target variable is plotted in the Graph Builder. In this example, the target variable is PURCHASE.
5. To create a custom graph, click the **Start Over** button in the upper left corner of the Graph Builder window.
6. In the **Variables** list, select **STATECOD** and drag it to the middle of the graph. You should see a state map for the USA.

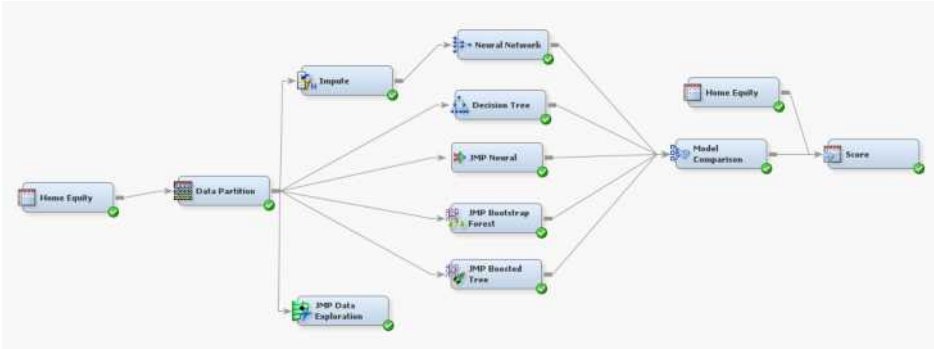



7. Right-click the graph and select **Map Shapes** ⇨ **Show Missing Shapes**.
8. In the **Variables** list, select **Income** and drag it to the **Color** drop zone. Next, drag the variable **_SEGMENT_** to the **Group Y** drop zone. This creates a graph of income distribution by cluster and state.

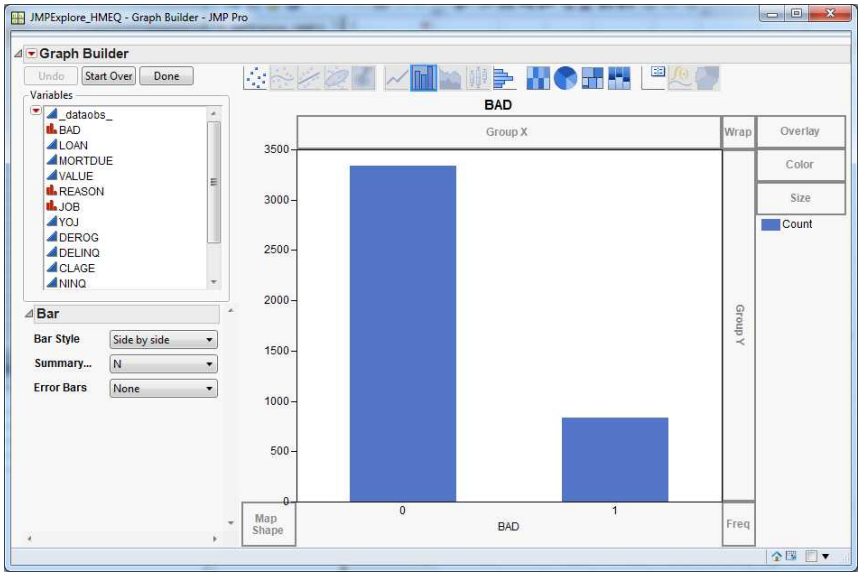


# Predictive Modeling

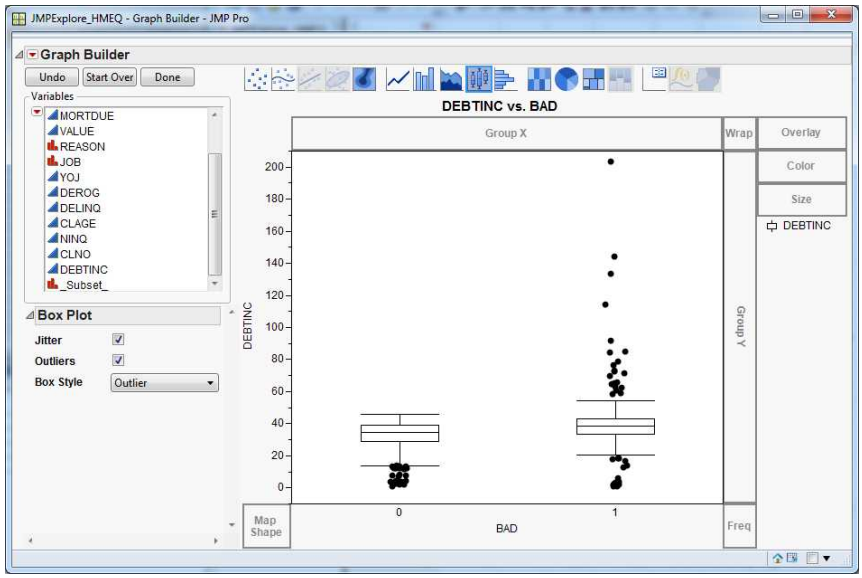
Use the **Home Equity** branch to explore the predictive modeling capabilities of JMP.




1. Right-click the **JMP Data Exploration** node in the process flow diagram and select **Run**. In the Confirmation window, select **Yes**.
2. After the process flow diagram has successfully run, select **Results** in the Run Status window.
3. In the Results window, click the **View** button. The target distribution is plotted by default.

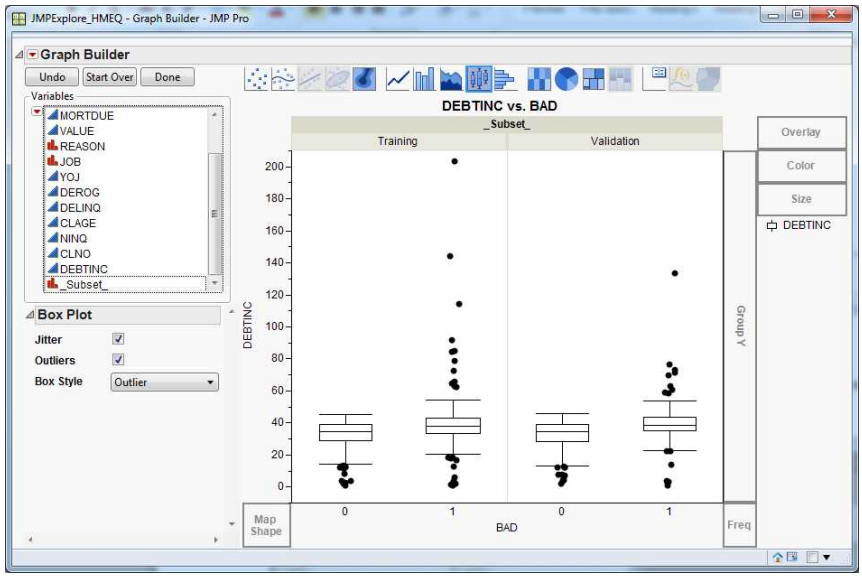



4. In the **Variables** list, select **DEBTINC** and drag it to the **Y** drop zone.
5. Right-click the graph and select **Box Plot**. The box plot shows how the debt-to-income ratio varies by loan status.



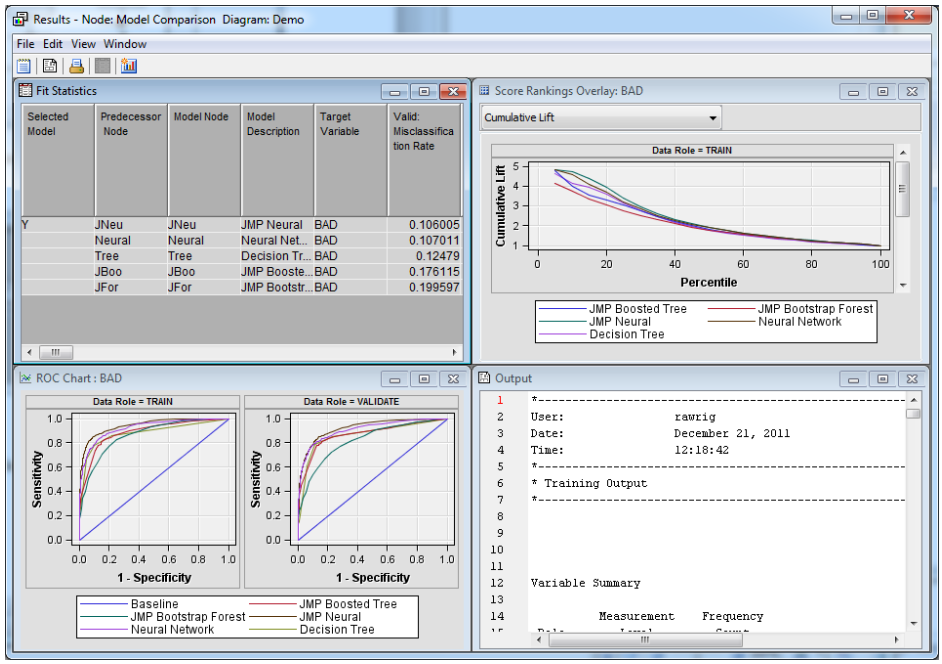

Note that there are a lot of outliers with a high debt-to-income ratio for the delinquent segment, where **BAD** equals **1**.

6. Suppose that you want to check whether the relationship between DEBTINC and the target variable varies by partition. In the **Variables** list, select **_Subset_** and drag it to the **Group X** drop zone. This separates the data by partition, which is **Training** and **Validation** for this example.




7. For both partitions, there are more customers with a high debt-to-income ratio in the delinquent segment. Close the Graph Builder and the Results windows.

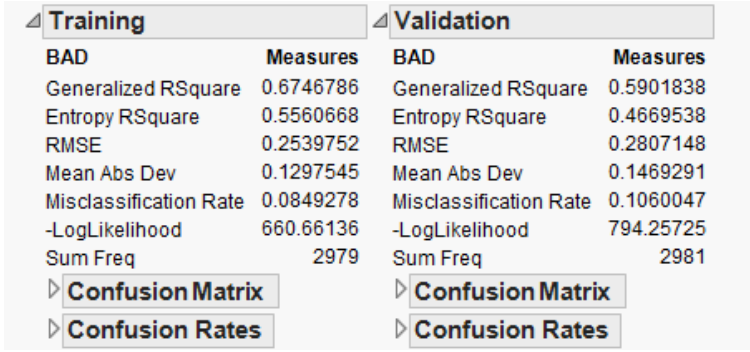
8. Right-click the **Model Comparison** node and select **Run**. In the Confirmation window, select **Yes**. The **Model Comparison** node evaluates the models created by its five predecessor nodes.



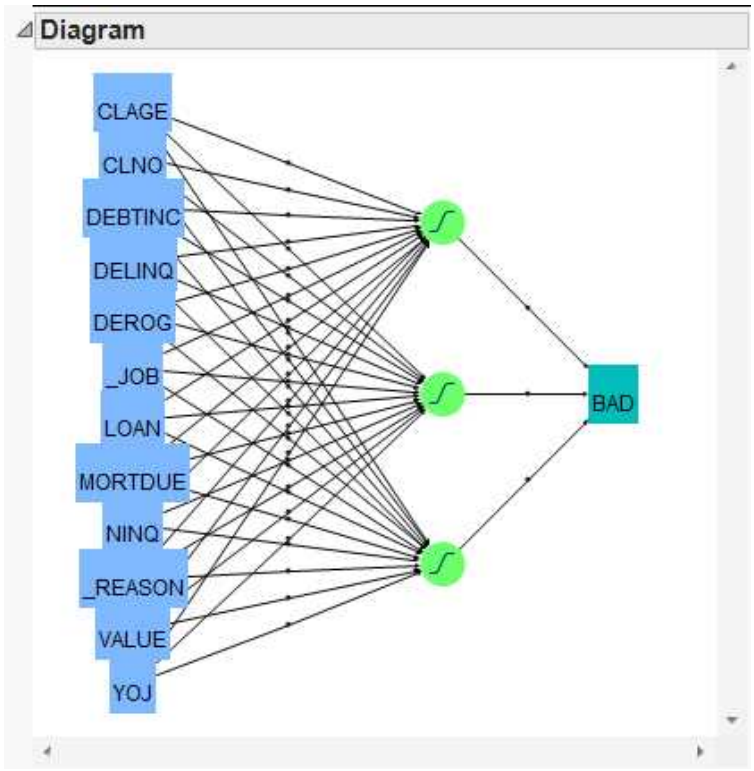

9. After the process flow diagram has successfully run, select **Results** in the Run Status window.

10. Based on the misclassification rate, the best two models are those created by the **JMP Neural** node and the **Neural Network** node. Close the Results window.

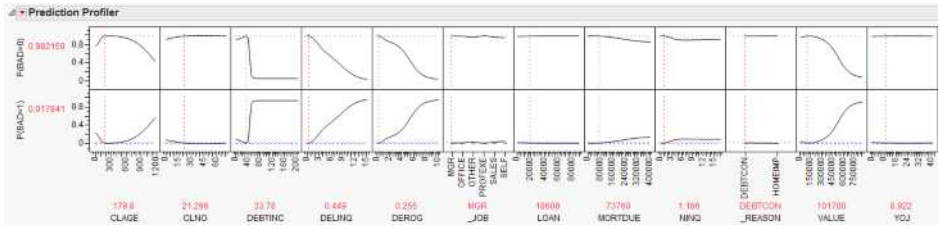
11. Right-click the **JMP Neural** node and select **Results**. In the Results window, click **View**. Classification results are shown at the beginning of the Interactive Report.




Also, a network structure diagram is included in the results window. For this example, there is a single hidden layer with three nodes.

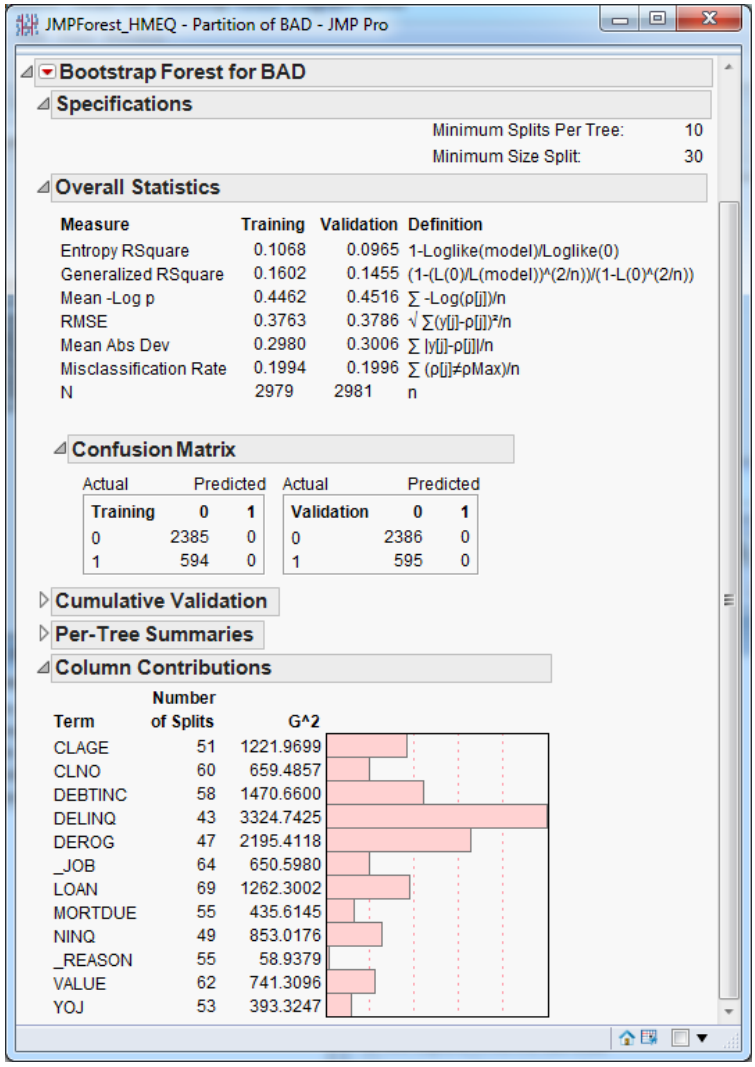



The network structure diagram is not informative on its own, but you can use the JMP **Prediction Profiler** to interactively explore how each predictor relates to the predicted values. For example, drag the dotted vertical line in the **DEBTINC** column to see how the debt-to-income ratio affects the probabilities for the target variable.




For debt-to-income ratios below 40, the odds of default are very low. Conversely, for debt-to-income ratios above 50, the odds of default are very high. Close the Results window.

12. The Results windows for the **JMP Bootstrap Forest** and **JMP Boosted Tree** nodes have a layout similar to the results of the **JMP Neural** node. Both Results windows include predictor (column) contributions. You should explore these results on your own. A portion of the **JMP Bootstrap Forest** node results is shown below.




Close any Results windows that you have open.

13. Right-click the **Score** node and select **Run**. In the Confirmation window, select **Yes**.
14. After the process flow diagram has successfully run, select **Results** in the Run Status window.
15. In the Results window, maximize the Optimized SAS Code window. The Optimized SAS Code window displays the score code for the best model, as determined by the **Model Comparison** node. In this example, that is the **JMP Neural** node.