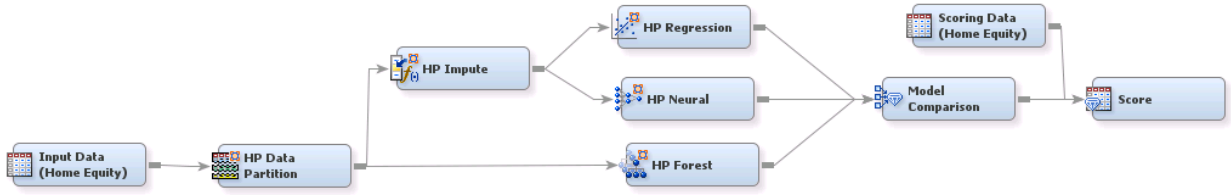
Predictive Modeling in SAS® Enterprise Miner™



Data:
The data set for this example is SAMPSIO.HMEQ. To create the data set, click **Help >> Generate Sample Data Sources** and select the **Home Equity** data source in SAS Enterprise Miner. The target is the binary variable BAD, which indicates home loans that were defaulted on.

Goal:
The goal is to create a model to predict which home loans will be bad loans (that is, will be defaulted on).

Flow:
The Home Equity node (an Input Data node) is connected to the Data Partition node. It is important to have a Data Partition node in every predictive modeling process flow diagram. The training data are used for building models, and the validation data are used to choose the best model while avoiding overfitting the model to the training data. The Data Partition node is changed from the default so that 70% of the data is used for training and 30% is used for validation. When you choose the split between training and validation, it is important to consider the properties of the data, such as how rare the target event is. Having too few targets in either the training data or the validation data can lead to poor models.

After you configure the Data Partition node, you can create a suite that has as many modeling nodes as you want. This example includes only the Regression, Neural Network, and Decision Tree nodes with default parameters, but you can add other modeling nodes. In addition, you can use multiple Regression, Neural Network, and Decision Tree nodes, each with different settings (instead of default settings), in order to explore more models.

There are some important considerations to take into account when you use various modeling nodes. Some nodes, such as Decision Tree, can interpret and include missing values in the model. Other nodes, such as Regression and Neural Network, should be preceded by an Impute node so that missing values are imputed in order to prevent excluding observations from the analysis.

The Impute node replaces missing values based upon certain rules, which you can select. The default is to replace a missing nominal variable with the most common level, and to replace a missing interval variable with the mean.

It is also often good practice to include a Variable Selection node (found on the Explore tab) before the Neural Network node. The number of variables in the Home Equity data is not large, but when a data set has a large number of variables, you can include a Variable Selection node where you select the most important variables to help the Neural Network node run more efficiently.

After you have created all the models that you want, all the modeling nodes are input into the Model Comparison node. The Model Comparison node compares each of the input models and chooses the champion model that the diagram will use to score future data. Under the **Selection Statistic** property for the Model Comparison node, you can select which criterion to use to determine the best model. The criterion will be specific to your goals.

The Model Comparison node is connected to the Score node. In addition, a scoring data set whose **Role** is set to **Score,** needs to be connected to the Score node. The Score node will take the new scoring data and apply the model chosen by the Model Comparison node to the data. In this way, new data can be scored after models have been trained. For this example, the scoring data are the same as the input data (Home Equity data), but this should not be the case in real-world scenarios.