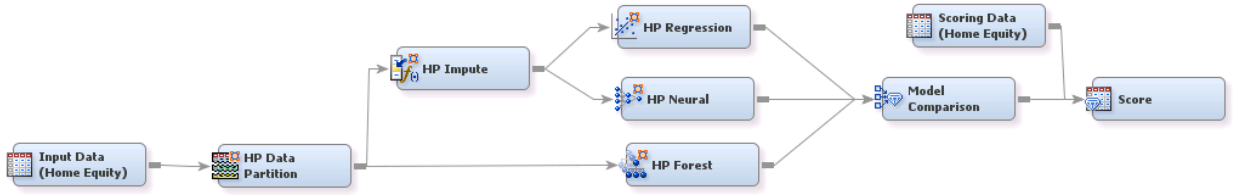
**High-Performance Predictive Modeling Using SAS® Enterprise Miner™**




**Data:**
The Home Equity (MHEQ) data set in the SAS library SAMPSIO is used to create the data source. You can generate the data by clicking **Help >> Generate Sample Data Sources >> Home Equity**. The target is the binary variable BAD, which takes a value 1 if a home loan was defaulted on, and 0 otherwise.

**Goal:**
The goal is to create a model for a binary target that predicts which home loans are likely to be defaulted. This flow uses high-performance nodes in SAS Enterprise Miner. If you set up a grid environment, the high-performance nodes run in a distributed computing environment, enabling you to use a large number of observations and variables to train and assess a predictive in a shorter period of time. These high-performance nodes are built on multithreaded procedures, so you might see performance gain even when you run these nodes on single-machine mode.

**Flow:**
The Input Data (Home Equity) node is connected to the HP Data Partition node which partition the data into two parts: 70% training and 30% validation. The training data is used to build a model, and the validation data is used to avoid overfitting in the final model. Properties of the data, such as rareness of the target event, are important when you split the data into training and validation sets. Accumulation of the rare target events in the training data or in the validation data can lead to poor models. However, target events are not rare in the home equity data.

Three models are built by using the default settings in HP Regression, HP Neural, and HP Forest nodes. You can add as many high-performance modeling nodes as your data mining task requires. As with other SAS Enterprise Miner nodes, you can also change the default options of these nodes to create additional models.

Notice that the decision tree model (in the HP Forest node) is developed before missing value imputation (in the HP Impute node), whereas regression (in the HP Regression node) and neural network analysis (in the HP Neural node) are done after missing data imputation. This is because the decision tree modeling can handle missing values well, whereas regression and neural network models cannot. Several different options are available in the HP Impute node for missing data imputation. This analysis uses the default options which set the missing values of classification variables to the most common level of the variable, and set the missing values of the each interval variable to its mean.

If the number of predictor variables is large, it is often recommended to perform variable selection before fitting a neural networks model in order to help the HP Neural node to run more efficiently. The number of variables is not large for the Home Equity data, so this step is not needed.

The Model Comparison node compares the three developed models and selects the champion model. You can use the **Selection Statistic** property of the Model Comparison node to choose a criterion that determines the champion model. This analysis uses the default criterion (lowest validation misclassification) to compare the three models.

The Score node uses the champion model that is chosen by the Model Comparison node to fit to the scoring data. In this example, the scoring data is same as the input data (Home Equity).