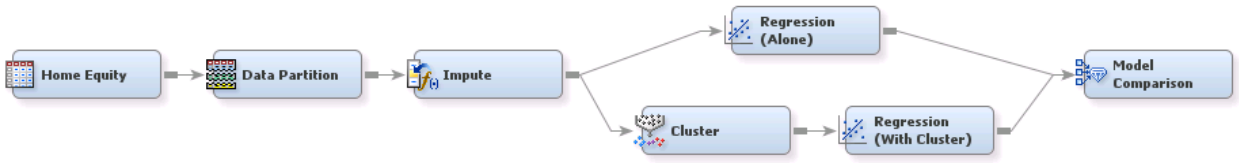
Using the Cluster Node in a Predictive Model in SAS® Enterprise Miner™



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 use the Cluster node to create segments that are used as input to a predictive model and to compare this model with a similar model that does not have this input.

Flow:
The Home Equity node (an Input Data node) is connected to the Data Partition node. Because the goal of the example is a predictive model, it is important to include a Data Partition node to prevent overfitting the models to the data. The Data Partition node specifies that 70% of the data be used for training and 30% be used for validation.

The Data Partition node is connected to the Impute node which fills in missing values in the data. Imputation is important for modeling when you use methods such as regression or a neural network. In this example, the data have no missing values, but the Impute node is included as a good practice. Imputing the data before using the Cluster node can change the results of the Cluster node. When you use the Cluster node to explore the data, an Impute node should not precede the Cluster node. But the Impute node precedes the Cluster node in this example because of the regression modeling nodes that follow it.

An alternative to using the Impute node is to use an additional Cluster node for the purpose of imputation. Under the **Scoring Imputation Method** property, you can select **Seed of Nearest Cluster** to indicate that you want the Cluster node to impute the value of the nearest cluster seed for any missing values.

The Regression (Alone) node is the default Regression node with two changes. The **Selection Model** property has been changed to **Stepwise**, and the **Regression Type** property has been changed to **Logistic Regression**. The Regression (Alone) node serves as a baseline with which to compare other regression models.

The Cluster node can assign segments to observations in ways that consider higher-level effects in the data than regression can. So, it can be useful to use the output of the Cluster node as input into a regression model. The Cluster node has two of the default values changed: the **Final Maximum** property is set to 8, and the **Cluster Variable Role** property is set to **Input** instead of **Segment**. The **Input** value indicates that subsequent model nodes will use the cluster into which an observation was placed as input for modeling.

An important property in the Cluster node is **Internal Standardization**, whose default is **Standardization**. The **Internal Standardization** property scales the columns of the data, so that the scale of the units does not affect the model. For example, a column in miles might have less variability than the same information recorded in feet. In this case, the **Internal Standardization** property can help improve the model. If the data were such that some columns are meant to have more variance than others, then you might consider changing the default for this property. In this diagram, the default is used.

The Regression (With Cluster) node uses the same properties as the Regression (Alone) node. The only difference is that the Regression (With Cluster) node has access to segments that the Cluster node creates as an additional input.

Both the Regression (Alone) and Regression (With Cluster) nodes are input into a Model Comparison node. By running the diagram and then viewing the results of the Model Comparison node, you can see that the branch that includes the Cluster and Regression (With Cluster) nodes performed best. The Cluster node was able to capture higher-level interactions in the data and pass that information on to the regression model.