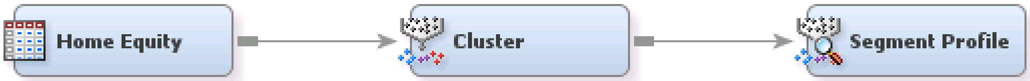
Using the Cluster and Segment Profile Nodes in SAS® Enterprise Miner™



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 and the Segment Profile node to explore the data.

Flow:

The Home Equity node (an Input Data node) is connected to the Cluster node. The Cluster node has many properties that you can change; all these details are fully described in the SAS Enterprise Miner help. In this example, the **Cluster Variable Role** is set to **Segment** (the default) because the next node is the Segment Profile node, which requires a variable with the role of **Segment**. Other role possibilities are **Input**, **Target**, and **ID**.

The **Final Maximum** option has also been changed to 8 from the default of 20. This property sets the maximum number of clusters that are ultimately created by the Cluster node. Setting this value too high might make it difficult for you to gain insight into the data because too many clusters are created. However, setting the value too low might not adequately separate the data. The choice for this value depends on the data set.

After you run the Cluster node, the **Mean Statistics** table in the **Results** includes statistics that describe each of the created clusters. The **Segment Size** chart shows you how large each segment is. Clicking on a segment in this chart highlights the particular row in the **Mean Statistics** table. To see the decisions that the Cluster node used to segment the data, click **View >> Cluster Profile >> Tree**.

The Cluster node is connected to the Segment Profile node, which requires the data to have a variable with the role of **Segment** in order to produce the profile of the segments in the data.

After you run the Segment Profile node, you can view in the **Results** a variety of statistics about the various segments in the data. In the **Variable Worth** plots, you can scroll through charts (one for each segment). These charts show how much each variable affects the selected segment.

In the **Profile** plots, you can see a further breakdown based on the variables. For each segment, there is a sequence of charts for each variable, listed in decreasing order of importance. For each chart, the blue bars represent the distribution of the variable among observations in the segment, and the red outlines represent the overall distribution among the entire data set.

For example, consider segment 2, which is represented in the top row of charts. The second chart is of the variable MortDue. In segment 2, the distribution of the segment is shifted to the right, whereas the red outlines show that the overall distribution is more to the left. In segment 8, MortDue is the third chart instead of the second. In this case, the distribution of the segment over observations is much more in line with the overall data set.