Predictive Modeling with SAS® Enterprise Miner™

Practical Solutions for Business Applications

Predictive Analytics

- Study the three main predictive modeling tools: Decision Tree, Neural Network, and Regression
- Examine the SAS code generated by each node and show the correspondence between the theory and the results produced by Enterprise Miner.
- Give intuitive explanations of the way that various nodes such as Decision Tree, Neural Network, Regression, and Variable Selection operate and how different options such as Model Selection Criteria and Model Assessment are implemented.

Chapter 3: Variable Selection and Transformation of Variables

Objectives

- Study the **Variable Selection** and **Transform Variables** nodes in detail.
- **Study the Selection** node
- Study how to make variable selection using the **Variable Clustering** and **Decision Tree** nodes.

Variable Selection

- In predictive modeling and data mining, we are often confronted with a large number of inputs (explanatory variables).
- The number of potential inputs to choose from may be 2000 or higher.
- Some of these inputs may not have any relation to the target.
- An initial screening can eliminate irrelevant variables and keep the number of inputs to a manageable size.
- A final selection can then be made from the remaining variables using either a stepwise regression or a decision tree.

The **Variable Selection** node performs the initial selection as well as a final variable selection:

• Case 1: The target is continuous (interval-scaled) and the inputs are numeric, consisting of interval-scaled variables.

• Case 2: The target is continuous and the inputs are categorical and nominal-scaled (These are sometimes referred to as class variables).

• Case 3: The target is binary and the inputs are numeric interval-scaled.

• Case 4: The target is binary and the inputs are categorical and nominal-scaled.

• Case 5: If the target is continuous and inputs are mixed, you have to specify how you want the Variables Selection node to handle categorical variables and continuous variables by appropriate property settings. The options considered in Case 1 and Case 2 are available even if you have mixed inputs. I did not include explicit examples for this case in this section because most of the examples included in subsequent chapters are with mixed inputs.

• Case 6: If the target is binary and inputs are mixed, you have to specify how you want the Variables Selection node to handle categorical variables and continuous variables by appropriate property settings. The options considered in Case 3 and Case 4 are available even if you have mixed inputs.

Variable Clustering Node

Selection of the Best Variable from Each Cluster

- **Variable Clustering** node is used for variable (input) selection.
- The variable selection done by the **Variable Clustering** node differs from the variable selection done by the **Variable Selection** node.
- The **Variable Selection** node selects the inputs on the basis of the strength of their relationship with the target (dependent) variable, whereas the **Variable Clustering** node selects the variables without Selecting the Cluster Components reference to the target variable.

Variable Selection Using the Decision Tree Node

- In this section, we use the **Decision Tree** node to select the inputs which are most useful in creating *groups of customers* called segments or leaf nodes.
- The target variable played no role in variable clustering, whereas it plays an important role in the **Decision Tree** node.
- When you have a very large number of variables, you can first eliminate obviously irrelevant variables using the **Variable Selection** node, and then use the **Decision Tree** node for further variable selection.

Transformation of Variables

- The **Transform Variables** node can make a variety of transformations of interval-scaled variables.
- There are two types of transformations for the categorical (class) variables:
  - Group Rare Levels
  - Dummy Indicators Transformation for the categories

- If you want to transform variables prior to variable selection, you can set up the path as **Input Data** node →
- **Data Partition** node → **Transform Variables** node → **Variable Selection** node, as shown in the upper path in
- If you do not want to reject inputs on the basis of linear relationship alone, you can use this process to capture non-linear relationships between the inputs and the target. However, with large data sets this process may become quite tedious.
- When there are a large number of inputs (2000 or more), it may be more practical to eliminate irrelevant variables first, and then define transformations on important variables only. The process flow for this is **Input Data** node → **Data Partition** node → **Variable Selection** node→ **Transform Variables** node (shown in the
- This section examines the results of both of these sequences with alternative transformations using a small data set.

Summary

- This chapter showed how the **Variable Selection** node selects categorical and continuous variables to be used as inputs in the modeling process when the target is binary or continuous.
- Two criteria are available for variable selection when the target is binary: R-Square and Chi-Square.
- The R-Square criterion can be used with binary as well as continuous targets, but the Chi-Square criterion can be used only with binary targets.
- The **Variable Selection** node transforms inputs into binned variables, called AOV16 variables, which are useful in capturing non-linear relationships between inputs and targets.
- The **Variable Selection** node also simplifies categorical variables by collapsing the categories.
- The **Variable Clustering** node can be used for variable selection by grouping the variables that look similar in the data set, and selecting a representative variable from each group. It can also be used to combine variables that look similar.
- The **Decision Tree** node can also be used for variable selection. However, if you have a large number of inputs, you should make a preliminary selection by using the **Variable Selection** node, and then use the **Decision Tree node** for further selection**.**