Predictive Modeling with SAS® Enterprise Miner™

Practical Solutions for Business Applications

Predictive Analytics-Course Summary

- 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 1: Research Strategy

- Defining the target population
- Defining the dependent variable
- Collecting data
- Cleaning the data
- Selecting an appropriate model.

Review of Some Definitions

- A *categorical variable* is one for which the measurement scale consists of a set of categories.
- Categorical variables for which levels (categories) do not have a natural ordering are called *nominal*.
- Categorical variables that do have a natural ordering of their levels are called *ordinal*.
- An *interval variable* is one that has numerical distances between any two levels of the scale.

- Interval-scaled variables are sometimes called *continuous*. Continuous variables are treated as interval variables.
- Therefore the terms *interval-scaled* and *continuous are* interchangeably.

Defining the Target

- The first step in any data mining project is to define and measure the target variable to be predicted by the model that emerges from your analysis of the data.
- This section presents examples of this step applied to five different business questions.

Target: Study the five different business questions (See Chapter 1 of your Text)

- Predicting Response to Direct Mail
- Identify the target variable
- Classify the target variable
- Discuss decision optimization

- Predicting Risk in the Auto Insurance Industry (Risk Model)
- Identify the target variable
- Classify the target variable
- Discuss decision optimization

- Predicting Rate Sensitivity of Bank Deposit Products

➢Identify the target variable
➢Classify the target variable
➢Discuss decision optimization

- Predicting Customer Attrition

➢Identify the target variable
➢Classify the target variable
➢Discuss decision optimization

- Predicting a Nominal Categorical (Unordered Polychotomous)Target

➢Identify the target variable
➢Classify the target variable
➢Discuss decision optimization

Sources of Modeling Data

- Data can be based on an experiment carried out by conducting a marketing campaign on a well-designed sample of customers drawn from the target population.
- Data can be based on a sample drawn from the results of a past marketing campaign and not from the target population.
- Before launching a modeling project, you must verify that the sample is a good representation of the target universe.
- You can do this by comparing the distributions of some key variables in the sample and the target universe.
- For example, if the key characteristics are age and income, then you should compare the age and income distribution between the sample and the target universe.

Observation Weights

- If the distributions of key characteristics in the sample and the target population are different, sometimes observation weights are used to correct for any bias.
- In order to detect the difference between the target population and the sample, you must have some prior knowledge of the target population.
- Another source of bias is often deliberately introduced. This bias is due to over-sampling of rare events.
- For example, in response modeling, if the response rate is very low, you must include all the responders available and only a random fraction of non-responders.
- The bias introduced by such over-sampling is corrected by adjusting the predicted probabilities with prior probabilities.

Pre-Processing the Data

Pre-processing has several purposes:

- eliminate obviously irrelevant data elements, e.g., name, social security number, street address, etc., that clearly have no effect on the target variable
- convert the data to an appropriate measurement scale, especially converting categorical (nominal scaled) data to interval scaled when appropriate
- eliminate variables with highly skewed distributions
- eliminate inputs which are really target variables disguised as inputs
- impute missing values

Alternative Modeling Strategies

- The choice of modeling strategy depends on the modeling tool and the number of inputs under consideration for modeling.
- Here are examples of two possible strategies when using the **Regression** node.
- Regression with a Moderate Number of Input Variables
- Regression with a Large Number of Input Variables

Regression with a Moderate Number of Input Variables

Pre-process the data:

- Eliminate obviously irrelevant variables.
- Convert nominal-scaled inputs with too many levels to numeric interval-scaled inputs, if appropriate.
- Create composite variables (such as average balance in a savings account during the six months prior

to a promotion campaign) from the original variables if necessary. This can also be done with SAS

- Enterprise Miner using the **SAS Code** node.

Next, use SAS Enterprise Miner to perform these tasks:

- Impute missing values.
- Transform the input variables.
- Partition the modeling data set into train, validate, and test (when the available data is large enough) samples. Partitioning can be done prior to imputation and transformation, because SAS Enterprise Miner automatically applies these to all parts of the data.
- Run the **Regression** node with the Stepwise option.

Regression with a Large Number of Input Variables

Pre-process the data:

- Eliminate obviously irrelevant variables.
- Convert nominal-scaled inputs with too many levels to numeric interval-scaled inputs, if appropriate.
- Combine variables if necessary.

Next, use SAS Enterprise Miner to perform these tasks:

- Impute missing values.
- Make a preliminary variable selection. (Note: This step is not included in Section 1.6.1.)
- Group categorical variables (collapse levels).
- Transform interval-scaled inputs.
- Partition the data set into train, validate, and test samples.
- Run the **Regression** node with the Stepwise option.