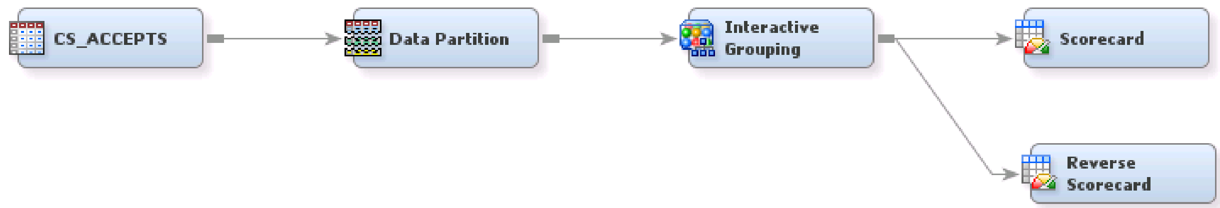
Scorecard and Reverse Scorecard Using Credit Scoring for SAS Enterprise Miner



Data:

The data set for this example is SAMPSIO.CS_ACCEPTS, which contains 3,000 observations. The target variable is GB, popularly known as a good/bad flag. This binary target indicates whether a new customer defaulted during a certain time window, typically one year.

The rest of the input variables are a set of demographics (age, number of children, region, income, and so on) and a summary of information from the credit bureau or the credit institution (number of loans with the bank, number of credit cards, type of credit product, and so on). The data also contain a frequency variable (_freq_) that serves as a weight variable for any statistical procedures.

Many training data sources used in Credit Scoring models are oversampled, which means that the distribution of events versus non-events in the training table is different from the true population distribution. Oversampled data should have a frequency variable that adjusts the distribution back in line with the true population. For the CS_ACCEPTS data set, the training table prior to adjustment the counts based on the _freq_ variable represents a 50/50 split of good to bad loans. Once adjusted by the frequency variable, the table now represents a 30/1 distribution. For this example, specify the Role of the _freq_ variable to **Frequency** and the Level as **Binary**.

Goal:

Scorecards are the standard model for credit scoring because they are easy to interpret and their output can be easily used to score new applications.

The SAS Enterprise Miner flow diagram of this template shows the basic steps to build a scorecard and a reverse scorecard. In the output of a scorecard, the higher the score, the less likely a client is to default. On a reverse scorecard, the higher the score, the more likely a client is to default.

Flow:

Data Partition Node

Use a Data Partition node to divide the data into training data and validation data. This is common practice in modeling to avoid overfitting. A good split for this data is 70% for training and 30% for validation.

Interactive Grouping Node

The Interactive Grouping node bins the original input variables, calculates the associated weight of evidence (WOE) values, and selects the most meaningful variables based on Variable Importance.

In many cases, business rules dictate a certain distribution for the WOE values of a binned variable. Use the **Interactive Mode** option to inspect the distribution of WOE values across the bins for each variable.

This mode enables you to visually examine all trends in WOE distribution and modify your grouping if necessary.

Find the WOE and associated statistics for each group on the Coarse Detail tab of interactive training. Use the column Manual WOE to adjust WOE values based on industry knowledge or business decisions. It is a best practice to correct any unusual trend of WOE in the Interactive Grouping node directly, rather than try to account for it at the Scorecard node (next step).

Scorecard Node

The Interactive Grouping node is typically followed by a Scorecard node. By default, the Scorecard trains a logistic regression model by using the WOE variables from the previous step as input variables and produces parameter estimates of the logistic regression. Then it calculates and scales score points for each level or group based on the WOE values and corresponding parameter estimates.

By default, points are assigned so that the higher the score, the lower the odds that default or event will happen. For example, a total score of 300 implies lower odds of a default, than a total score of 200 implies.

In certain situations, it is more convenient to have a reverse scorecard, in which a higher score implies that the odds of an event increases with a higher score. For example, when a bank creates a scorecard to model the event of customers accepting a promotional offer, it wants the high values of scorepoints to be associated with high likelihood of accepting the offer. Specify the **Reverse Scorecard** option as Yes to calculate the total score so that the higher the score, the more likely a customer will accept the offer.

Other Credit Scoring Examples:

Add Reject Inference to solve sample bias