Summary

Predictive Modeling Fundamentals

Predictive modeling tries to find good rules or formulas for predicting the value of the target variable in a data source from the values of the input variables in the data source. The primary goal of predictive modeling is to predict outcomes for new data. Other goals of predictive modeling include developing models that maximize predictive power and meet the needs of your organization.

Common terms for elements in predictive modeling include the following: cases, which are the observations or examples in the data used to develop the predictive model; input variables, also known as inputs or predictors; and the target variable, which is the outcome to be predicted.

There are two basic steps of predictive modeling. The first step is to build a model on historic data in which the target classification is known; this process is known as supervised classification. The second step is to apply the predictive model to cases in which the target classification is unknown; this process is known as generalization, or scoring.

Predictive modeling has many applications including target marketing, attrition prediction, credit scoring, and fraud detection.

View the "Examining the Code for Generating Descriptive Statistics and Frequency Tables" demonstration that shows how to use SAS to explore a data set by generating descriptive statistics and frequency tables.

Predictive Modeling Challenges

When developing predictive models, you will likely face a number of data challenges, including issues with observational data, mixed measurement scales, high dimensionality, and rare target events. These challenges can skew your results. In addition, modelers typically face analytical challenges, including the nonlinear and non-additive effects of using high numbers of variables, as well as selecting a model with the right amount of complexity to represent the true relationships between inputs and the target.

When the ratio of events to non-events is very small, separate sampling is used to oversample the rare target event − that is, to create a sample that disproportionately over-represents the event cases. In separate sampling, a target-based sample is created by drawing samples separately based on the target outcome − that is, whether it is a non-event or an event.

The optimism principle states that when you assess the accuracy of a predictive model on the same data that was used to fit the model, you tend to get better assessment statistics than when you assess the model on other data. This bias is known as the optimism bias. To avoid an optimistically biased assessment and create a predictive model that generalizes well, you need to assess the performance of the model on new data that was not used to fit the original model. This approach is called honest assessment.

The simplest way to do an honest assessment of how well your model generalizes to new data is to split the data. You split your data into two data sets: a training data set and a validation data set. You fit the model to the training data. In other words, you use the training data to "train" the model. The validation data can be referred to as the holdout portion. You use the validation data to assess and compare models. To ensure that the training and validation sets have an equal percentage of events, you can use stratified random sampling. In stratified random sampling, you divide the observations into non-overlapping groups (or strata) and then select a sample from each group. In the scenarios in this course, the strata are the target values and the resulting samples are the training and validation data sets.

View the "Splitting the Data" demonstration in SAS that shows how to use PROC SURVEYSELECT to select the records for the training and validation data sets − that is, to split the data. To create a stratified sample, the data must be sorted by the stratum variable.