

4.5 Chapter Summary

One of the most important steps in predictive modeling is to assess the performance of the model. To correct for the optimistic bias, a common strategy is to holdout a portion of the development data for assessment. The LOGISTIC procedure can then be used to score the data set used for assessment. Statistics that measure the predictive accuracy of the model include sensitivity and positive predicted value. Graphics such as the ROC curve, the gains chart, and the lift chart can also be used to assess the performance of the model.

If the assessment data set was obtained by splitting oversampled data, then the assessment data set needs to be adjusted. This can be accomplished by using the sensitivity, specificity, and the prior probabilities.

In predictive modeling, the ultimate use of logistic regression is to allocate cases to classes. To determine the optimal cutoff probability, the plug-in Bayes rule can be used. The information you need is the ratio of the costs of false negatives to the cost of false positives. This optimal cutoff will minimize the total expected cost (or maximize the total expected profit).

The profit itself can be used as an assessment statistic, presuming that some reasonable estimate of the appropriate financial figures can be found.

When the target event is rare, the cost of a false negative is usually greater than the cost of a false positive. For example, the cost of not soliciting a responder is greater than the cost of sending a promotion to someone who does not respond. Such considerations dictate cutoffs that are usually much less than .50, the cutoff that maximizes accuracy.

A popular statistic that summarizes the performance of a model across a range of cutoffs is the Kolmogorov-Smirnov statistic. However, this statistic is not as powerful in detecting location differences as the Wilcoxon-Mann-Whitney test. Furthermore, the Wilcoxon-Mann-Whitney test statistic is equivalent to the area under the ROC curve (the c statistic). Thus, the c statistic should be used to assess the performance of a model across a range of cutoffs.

General form of PROC NPARIWAY:

```
PROC NPARIWAY DATA=SAS-data-set <options>;  
  CLASS variable;  
  VAR variable;  
RUN;
```