Chapter 1. Predictive Modeling

1.1 Introduction

Supervised Classification

A predictive model maps the vector of input variables to the target. The target is the outcome to be predicted. The cases are the units on which the prediction is made.

In supervised classification, the target is a class label. A predictive model assigns, to each case, a score that measures the propensity that the case belongs to a particular class. The term *supervised* is used when the class label is known for each case. If the label is known, then why build a prediction model?

Generalization

The prediction model is used on new cases where the values of the input variables are known, but the class labels are unknown. The principal aim of predictive modeling is generalization. *Generalization* means the ability to predict the outcome on novel cases.

Applications

There are many business applications of predictive modeling.

- Target Marketing
- Attrition Prediction
- Credit Scoring
- Fraud Detection

1.2 Analytical Challenges

Opportunistic Data

- Operational / Observational
- Massive
- Errors and Outliers
- Missing Values

The data that are typically used to develop predictive models can be characterized as opportunistic. The data were collected for operational purposes unrelated to statistical analysis. Such data are usually massive, dynamic, and dirty. Preparing data for predictive modeling is often very difficult.

Mixed Measurement Scales

When there are large numbers of input variables, there is usually a variety of measurement scales represented. The input variable may be intervally scaled, binary, nominally scaled, ordinally scaled, or counts.

High Dimensionality

The *dimension* refers to the number of input variables. The number of variables often has a greater effect on computational performance than the number of cases. High dimensionality limits the ability to explore and model the relationships among the variables. This is known as the *curse of dimensionality*.

Rare Target Event

In predictive modeling, the event of interest is often rare. One widespread strategy for predicting rare events is to build a model on a sample that disproportionally over-represents the event cases. Such an analysis introduces biases that need to be corrected so that the results are applicable to the population.

Nonlinearities and Interactions

Predictive modeling is a multivariate problem. Each important dimension might affect the target in complicated ways. Moreover, the effect of each input variable might depend on the values of other input variables. The curse of dimensionality makes this difficult to untangle.

Model Selection

Predictive modeling typically involves choices from among a set of models. A common pitfall is to overfit the data (to use too complex a model). An overly complex model might be too sensitive in the sample data set and not generalize well to new data. However, using too simple a model can lead to *underfitting*, where true features are disregarded.