1.3 Chapter Summary

The principal objective of predictive modeling is to predict the outcome on new cases. Some of the business applications include target marketing, attrition prediction, credit scoring, and fraud detection. One challenge in building a predictive model is that the data usually was not collected for purposes of data analysis. Therefore, it is usually massive, dynamic, and dirty. For example, the data usually has a large number of input variables. This limits the ability to explore and model the relationships among the variables. Thus, detecting interactions and nonlinearities becomes a cumbersome problem.

When the target is rare, a widespread strategy is to build a model on a sample that disproportionally over-represents the events. The results will be biased, but they can be easily corrected to represent the population.

A common pitfall in building a predictive model is to overfit the data. An overfitted model will be too sensitive to the nuances in the data and will not generalize well to new data. However, a model that underfits the data will systematically miss the true features in the data.