15.7 Analytics and Data Mining

Section 1.5 noted that advances in computing processing during the past 40 years have brought statistical applications to the business desktop. More recently, those advances have combined with breakthroughs in data storage technologies to create the new field of analytics.

Analytics combines computer systems and statistics to analyze enterprise data, particularly collections of historical data. Analytics appears in many forms and as part of many types of modern managerial applications, including business intelligence, metrics, and current customer relationship management systems. Portfolio risk analysis, discussed in Section 5.2, illustrates, in miniature, one common application of analytics. Unlike the actual analysis done in Section 5.2, the “analytical” version of portfolio risk analysis would use historical data and have the ability to perform multiple what-if substitutions of investments.

Analytics may prove to be an important part of tomorrow’s business decision making. Tom Davenport and Jeanne Harris, the authors of two recent books about this field, including Competing on Analytics: The New Science of Winning, argue that using sophisticated quantitative analyses should be part of the competitive strategy of an organization (see references 1 and 2).

Data Mining

One analytic technique that has gained widespread use is data mining. Data mining can be defined as the useful combination of database technologies and statistics. Data mining allows statistical methods to be applied to very large data sets that contain many variables, each with many values.

Data Mining for Predictive Modeling

Many data mining applications extend the methods of predictive modeling discussed in this chapter and Chapters 13 and 14. As with regression modeling, in order to validate any data mining analysis, where possible, you should split the data into a training sample that is used to develop models for analysis and a validation sample that is used to determine the validity of the models developed in the training sample.

Data mining-based predictive applications typically differ from the regression methods discussed in this book in the following ways:

- They use many variables from many files in a corporate database instead of using one data file of a fixed number of variables.
- They use a semi-automated process to search for possible independent variables (from a much wider set of candidate variables) instead of using a fixed list of previously identified independent variables.
- They make extensive use of historical data sets.
- They can increase the chance of encountering a pitfall in regression (see Sections 13.9 and 15.5) when used in an unknowing manner.

One recent application of data mining for predictive modeling was the consumer research that uncovered a correlation between the number of Web searches for a new feature film, new video game, or new song and the opening weekend revenue for a new film, the first-month sales for a new video game, and the rank of the new song on the Billboard Hot 100 chart (see reference 3). This research would not have been possible without the database technology that maintains the Yahoo! U.S. Web search query logs that were used as the source of data for the analysis.

Predictive modeling applications of data mining are also used to assist in a wide variety of business decision-making processes. The following are some of these applications, by business field:

- **Banking and financial services** Predict which applicants will qualify for a specific type of mortgage and which applicants may default on their mortgage (mortgage acceptance and default). Predict which customers will not change their financial services company (retention).
- **Retailing and marketing** Predict which customers will best respond to promotions (promotion planning). Predict which customers will remain loyal to a product or service (brand loyalty). Predict which customers are ready to purchase a product at a certain
time (purchasing sequence). Predict which product a consumer will purchase given the previous purchase of another product (purchase association).

- **Quality and warranty management** Predict the type of product that will fail during a warranty period (product failure analysis). Detect the type of individual who might be involved in fraudulent activities concerning the warranty of the product (warranty fraud)

- **Insurance** Predict the characteristics of a claim and an individual that indicate a fraudulent claim (fraud detection). Predict the characteristics of an individual who will file a specific type of claim (claim submission).

**Data Mining for Exploratory Data Analysis** Data mining has other applications besides predictive modeling. Data mining can be an exploratory data analysis tool that allows business decision makers to examine the basic features of a data set, using tables, descriptive statistics, and graphic displays. **Dashboards**, so called because of their similarity to an automotive dashboard, are one example of this type of data mining. For example, the Command Center application developed by The New York Jets football organization (see reference 6), helps manages all activities in their stadium as they occur on game day. Using the Command Center, team managers can instantaneously track attendance and concession and merchandise sales, comparing those statistics with historical averages or last-game values as well as determine such things as the elapsed time between when a fan first enters a parking lot and then first enters the stadium.

**Statistical Methods in Data Mining**

Besides the regression methods of Chapters 13 and 14 and this chapter, data mining uses the following methods, also discussed in this book:

- Bar charts
- Pareto charts
- Multidimensional contingency tables
- Descriptive statistics such as the mean, median, and standard deviation
- Boxplots

Data mining and analytics in general also use statistical methods that are beyond the scope of this introductory-level book. Those methods include the following:

- Classification and regression trees (CART)
- Chi-square automatic interaction detector (CHAID)
- Neural nets
- Cluster analysis
- Multidimensional scaling

**Classification and regression trees (CART)** is an example of a decision tree (see Section 4.2) algorithm that splits the data set into groups based on the values of independent or explanatory ($X$) variables. The CART algorithm goes through a search process to optimize the split for each independent or explanatory ($X$) variable chosen. Often, the tree has many stages or nodes and a decision needs to be made as to how to prune (cut back) the tree.

**Chi-square automatic interaction detector (CHAID)** also uses a decision tree (see Section 4.2) algorithm that splits the data set into groups based on the values of independent or explanatory ($X$) variables. Unlike CART, CHAID allows a variable to be split into more than two categories at each node of the tree.

**Neural nets** have the advantage of using complex nonlinear regression functions to predict a response variable. Unfortunately, this method’s use of a complex nonlinear function can make the results of a neural net difficult to interpret.

**Cluster analysis** is a dimension-free procedure that attempts to subdivide or partition a set of objects into relatively homogeneous groups. These homogenous groups are developed so that objects within a group are more alike other objects in the group than they are to objects outside the group.
Multidimensional scaling uses a measure of distance to develop a mapping of objects usually within a two-dimensional space so that the characteristics separating the objects can be interpreted. Typically, multidimensional scaling attempts to maximize the goodness of fit of the actual distance between objects with the fitted distances.

**Figure 15.17**
Classification and regression tree (CART) results for predicting the proportion of credit card holders who would upgrade to a premium card (see next page for discussion)
For more detailed discussion of the statistical methods used in data mining, see references 4, 5, and 7.

**Data Mining Using JMP**

There are data analytics applications just as there are business statistics applications. One data analytics application, JMP from the SAS Institute, brings data analytics to the desktop in a form recognizable to users of Excel and Minitab.

To illustrate using JMP for classification and regression tree (CART) analysis, return to the Section 14.7 example in which a logistic regression model was used to predict the proportion of credit card holders who would upgrade to a premium card. Using JMP, one can create the classification and regression tree (CART) results shown in Figure 15.17.

Observe from Figure 15.17 that the first split of the data is based on whether the cardholder has additional cards. Then, in the next row, the two categories Extra Cards(Yes) and Extra Cards(No) are split again, using the annual purchase amount as the basis of this second split. In the Extra Cards(Yes) category, the split is between those who charge more than $49,738.80 per year and those who charge less than $49,738.80 per year. In the Extra Cards(No) category, the split is between those who charge more than $35,389.90 per year and those who charge less than $35,389.90 per year.

These results show that customers who have extra cards and have charged over $49,738.80 per year are much more likely to upgrade to a premium card. (Least likely to upgrade to a premium card are customers who have only a single charge card and have charged less than $35,389.90.) Therefore, the credit card company might want to focus future premium-card upgrade marketing efforts at customers who have already have additional cards and charge more than $49,738.80 per year. The $r^2$ of 0.516, shown in the summary box below the plot, means that 51.6% of the variation in whether a cardholder upgrades can be explained by the variation in whether the cardholder has additional cards and the amount the cardholder charges per year.

**Using JMP 9 for Classification and Regression Trees (CART)**

Use Partition to perform a classification and regression tree (CART) analysis. For example, to perform the analysis of the likelihood of upgrading to a premium card shown in Figure 15.17, open JMP 9 and select **File ➔ Open**. In the Open Data File dialog box:

1. Select **Excel 97-2003 Files (*.xls)** from the drop-down list to the right of the **File name** box.
2. Navigate to the folder containing **LogPurch.xls** and then select that file to enter its name in the **File name** box.
3. Click **Open**.

JMP opens a new DATA window that contains the credit card data. Double-click the **Upgraded** column (the first column). In the Upgraded - JMP (column properties) dialog box (shown on page 5):

4. Select **Character** from the **Data Type** pull-down list.
5. Select **Value Labels** from the **Column Properties** pull-down list.
6. In the expanded dialog box, enter 0 in the **Value** box, No in the **Label** box, and click **Add**. Then enter 1 in the **Value** box and Yes in the **Label** box and click **Add** a second time.
7. Click **OK**.

JMP displays the words No and Yes for the values 0 and 1.

8. Double-click the **Extra Cards** column. In the Extra Cards - JMP dialog box, repeat steps 4 through 7.
9. Select **Analyze Modeling Partition**.
In the Partition JMP dialog box (shown below):

10. Select Upgraded in the Select Columns list and click Y, Response to enter Upgraded in the box by the Y, Response button.
11. Select Purchases and click X, Factor to enter Purchases in the box by the X, Factor button.
12. Select Extra Cards and click X, Factor to enter Extra Cards in the box by the X, Factor button.
13. Click OK.

JMP opens a new DATA - Partition of Upgraded - JMP window. Click Split to split the data based on whether the cardholder has additional cards. Click Split two more times to create the analysis tree shown in Figure 15.17.

REFERENCES

7. Tan, P.-N., M. Steinbach, and V. Kumar, Introduction to Data Mining (Boston: Addison-Wesley, 2006).