MAKING SENSE OF DATA II

A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications

GLENN J. MYATT

WAYNE P. JOHNSON

WILEY

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Myatt, Glenn J., 1969-

Making sense of data II: a practical guide to data visualization, advanced data mining methods, and applications/Glenn J. Myatt, Wayne P. Johnson.

p. cm.

Making sense of data 2

Includes bibliographical references and index.

ISBN 978-0-470-22280-5 (pbk.)

1. Data mining. 2. Information visualization. I. Johnson, Wayne P. II. Title. III. Title: Making sense of data 2.

QA76.9.D343M93 2008

005.74--dc22

2008024103

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CHAPTER 4

PREDICTIVE ANALYTICS

4.1 OVERVIEW

4.1.1 Predictive Modeling

Predictive analytics refers to a series of techniques concerned with making more informed decisions based on an analysis of historical data. These methods are used throughout the industries of science and business. For example, pharmaceutical companies use these techniques to assess the safety of potential drugs before testing them in human clinical trials, and marketing organizations use these models to predict which customers will buy a specific product.

This section will use the following example to illustrate the process of building and using predictive models. A telecommunications company wishes to identify and prioritize all customers they believe are likely to change to another service provider. In an attempt to avoid losing business, the company will offer new products or new service options to these customers. Over the years, the organization collected customer data monthly that included decisions to switch to another service provider. Table 4.1 is an example of the information collected on customers, where each observation relates to a specific customer in one month. The data includes a variable, *churn*, where a 1 indicates a switch to a new service provider and a 0 indicates a continuation of service. The table also includes the specific month the observation was recorded (*month*), the age of the customer (*age*), the annual income of the customer (*income*), the number of months they have been a customer (*customer length*), the gender of the customer (*gender*), the number of calls made that month (*monthly calls*), and the number of calls made to the customer service department (*service requests*). This data will be used to build a model attempting to understand any relationships between (1) the customer data collected and (2) whether the customer changes service provider.

The telecommunications company would like to use the model in their marketing department to identify customers likely to switch. Since the company has limited resources, it would also like to prioritize the customers most likely to switch. The company has data for the current month, which is shown in Table 4.2. The data for the current month is the same type of data used to build the model, except that *churn* is currently unknown. The model created from the historical data can then be used with the new data to make a series of predictions. Table 4.3 is an example of a table where two columns have been added for the values generated by the

TABLE 4.1 Example of Telecommunications Data Used to Build a Model

ID	Month	Age	Income	Customer length	Gender	Monthly calls	Service requests	Churn
A	January	45	$72k	36	Female	46	1	0
B	March	27	$44k	24	Male	3	5	1
C	July	51	$37k	47	Male	52	0	0
D	February	17	0	16	Female	62	1	1
E	December	45	$63k	63	Female	52	0	0
F	October	24	$36k	24	Male	72	1	0
G	March	39	$48k	5	Male	36	0	0
H	June	46	$62k	17	Male	1	0	1
...	...	...	...	...	...	...	...	...

TABLE 4.2 Data Collected on Customers for the Current Month

ID	Month	Age	Income	Customer length	Gender	Monthly calls	Service requests	Churn
a	May	52	$84k	52	Female	52	0	?
b	May	26	$28k	14	Male	12	2	?
c	May	64	$59k	4	Male	31	1	?
...	...	...	...	...	...	...	...	...

model. *Predicted churn* indicates whether that particular customer is likely to switch services (*predicted churn* = 1) or not (*predicted churn* = 0) this month. *Churn probability* reflects the likelihood, or probability, that a particular customer will be assigned a *predicted churn* value of 1. This table can then be sorted by the probability column, allowing the marketing team to focus on those customers most likely to switch.

Figure 4.1 illustrates the general process of generating and using a prediction model. Data is used to generate a model that predicts a specific response variable from a set of independent variables. In the telecommunications example, the independent variables are *age*, *income*, *gender*, *customer length*, *monthly calls*, and *service requests*. These independent variables are used in concert with the response variable,

TABLE 4.3 Customers Predicted to Change Services this Month, and a Measure of the Likelihood of Switching

ID	Month	Age	Income	Customer length	Gender	Monthly calls	Service requests	Predicted churn	Churn probability
a	May	52	$84k	52	Female	52	0	0	0.33
b	May	26	$28k	14	Male	12	2	1	0.74
c	May	64	$59k	4	Male	31	1	1	0.88
...	...	...	...	...	...	...	...	...	...

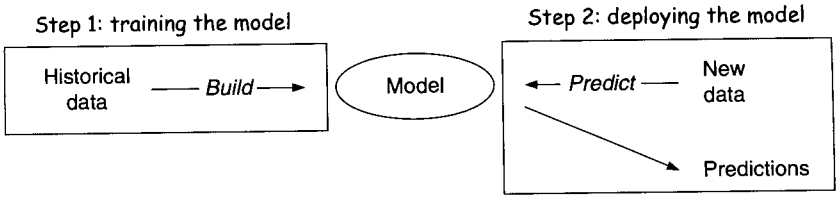

Figure 4.1 Process of generating and using a prediction model

or *churn* in this example. The model captures relationships between the independent variables and the response variable. Models are built in an initial training step where the model learns from the historical data. Once a model has been built, a new data set containing the same independent variables can be used with the model to calculate predictions from the captured relationships. In Fig. 4.1, this is identified as the second step, where the model is deployed and being used to predict a value for a response variable.

Building predictive models assumes the same exploratory data analysis as discussed in the previous chapters has been done. (The first three chapters discussed the process and methods such as data visualization, descriptive statistics, and clustering that help understand the content.)

For the model to capture important relationships, the data set must have a balanced mixture of positive and negative examples. In the telecommunications example, this means including observations where the customer did and did not switch. It will be hard to build an objective model without a good proportion of both cases because the model may falsely identify certain characteristics that are merely common to all customers, not specifically to those that switch services.

Preparing and selecting which variables should be used with a model is a critical phase in the process. Prioritizing the variables to be considered is important, especially when the data set has many variables. One way to prioritize is to look at the relationships between the potential independent variables and the target response variable to see if there is any relationship. Using methods such as a hypothesis test, a chi-square test, or a matrix of correlation coefficients can help to prioritize the variables. There should be no strong relationships between the independent variables included in the model because related variables are redundant, or worse, they may introduce problems when building models. Knowledge of the specific subject matter and an understanding of how the model is going to be deployed is also often critical to making choices about what variables to include. For example, if collecting or measuring a particular variable's values is too costly, then this variable should be excluded despite its potential utility.

The next major factor that affects the model-building approach is the type of response variable involved, that is, whether the variable is categorical or continuous. If the response variable is categorical, then a *classification* modeling approach should be used. These include logistic regression, discriminant analysis, naive Bayes, k-nearest neighbors (kNN), classification trees, and neural networks. If the response variable is continuous, then a *regression* modeling approach should be considered. These include linear regression, k-nearest neighbors, regression trees,

TABLE 4.4 Summary of Different Modeling Approaches

Method	Model type	Independent variables	Comments
Linear regression	Regression	Any numeric	Assumes a linear relationship Easy to explain Quick to build
Discriminant analysis	Classification	Any numeric	Assumes the existence of mutually exclusive groups with common variances
Logistic regression	Classification	Any numeric	Will calculate a probability Easy to explain
Naive Bayes	Classification	Only categorical	Requires a lot of data
Neural networks	Regression or classification	Any numeric	Black box model
kNN	Regression or classification	Any numeric	Difficult to explain results Handles noise well Handles nonlinear relationships
CART	Regression or classification	Any	Explanation of reasoning through use of a decision tree

and neural networks. A few approaches may be used with categorical and continuous responses. All approaches have advantages and disadvantages, and the different approaches often require the data to conform to certain assumptions. The major approaches are summarized in Table 4.4.

In selecting which approach to use, other practical considerations may need to be considered. In the telecommunication example, the marketing department needed a classification model to predict the binary response variable *churn*; however, the department also wanted to prioritize the results so they can focus their efforts on customers most likely to switch. In this situation, the logistic regression method might be a good candidate since it generates a prediction for the binary variable *churn* and calculates the probability of switching.

Different modeling approaches operate in different ways and selecting the best method requires an understanding of the data and how the different methods operate. Figure 4.2 illustrates different types of approaches. In chart A, a single independent variable is plotted against a single response variable. Because a linear relationship exists between the two variables, they can be modeled using a linear regression method. In contrast, chart B indicates a nonlinear relationship where the data either needs to be transformed or needs to be used with a method that is capable of modeling these relationships. In chart C, the data used to build the model was divided into regions based on specific values or ranges of the independent variables. In this example, two independent variables are used to illustrate how the data can be divided into regions (no response variable is shown). When making a prediction, the approach assigns an observation into a specific region based on values or ranges for the independent variables, and a prediction is made based on training data in the region, such

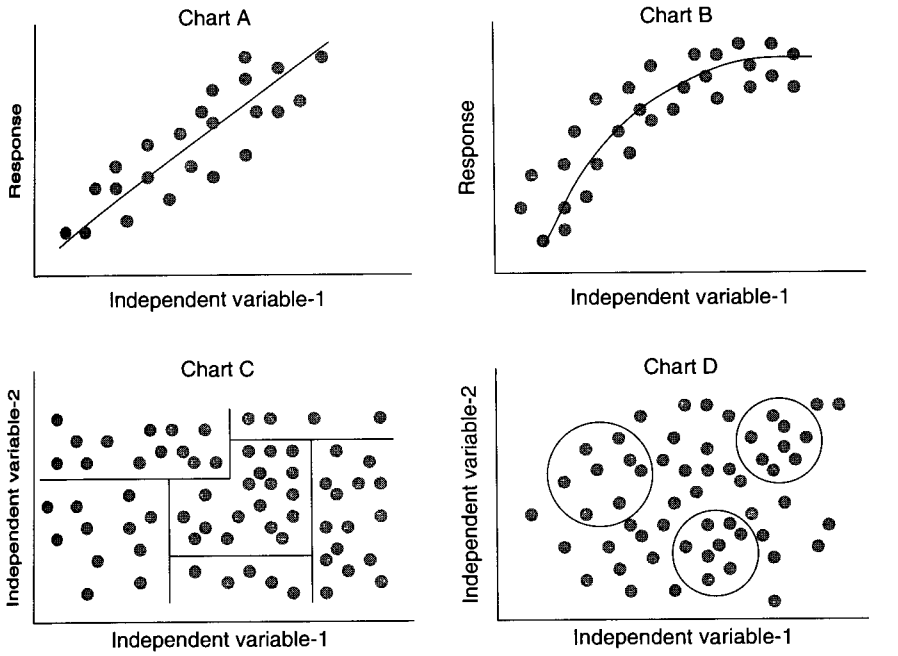

Figure 4.2 Illustrations of different regression modeling approaches

as the average value. Techniques such as regression trees make predictions in this manner. Another approach is illustrated in chart D, where similar observations from the training data to the observation to be predicted are identified and a prediction is made based on the average response values from this set. *k*NN uses this approach to make predictions.

Figure 4.3 illustrates a number of classification approaches. In chart A, the independent variable space is characterized by grouping similar observations. A prediction is made by deciding what similar observations are present in the training set, and then using the mode response value from these observations as the predicted classification. Again, the *k*NN method is an example of this approach. Similarly, the training data could have been divided into specific regions based on a good classification

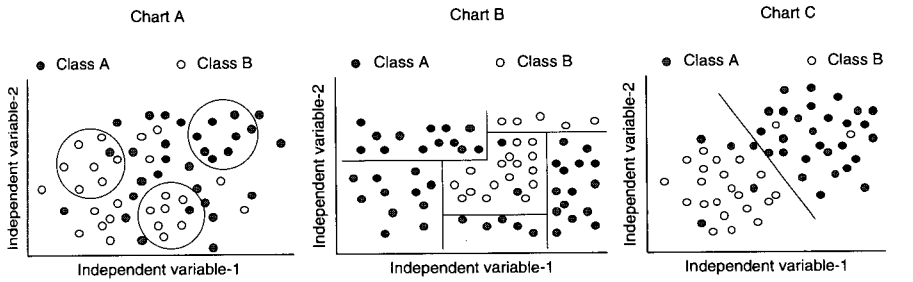

Figure 4.3 Illustration of different classification modeling approaches

of the data, and predictions may be made by determining where in this space a test observation should be assigned. Classification trees are an example of this approach. Another approach is to determine whether there are boundaries between the classes, as illustrated in chart C. These boundaries allow for classification, and methods such as discriminant analysis will model the data based on this type of approach.

Modeling the data using different approaches and fine-tuning the parameters used within each method will lead to models with greater predictive accuracy. Understanding how the individual approaches operate will help in this optimization process. Building multiple models and assessing them with a consistent methodology will help to select the best one for the particular problem being addressed. Care must be taken to build the model such that the relationships are general enough to make good predictions beyond the specific examples in the training data. In addition, the bounds of the model should be characterized based on the types and ranges of data used to build the model or based on general population characteristics or assumptions.

This chapter describes a series of multivariate approaches for building and testing predictive models. Section 4.1.2 describes methods for dividing data sets into sets for training and testing the model, thus ensuring objective testing of the model. Different metrics for assessing models are provided in Sections 4.1.3–4.1.7. These methods differ based on the type of the response variable, that is, whether the variable is continuous, categorical, or binary. Prior to building any model, it is important to understand and select variables to use as independent variables in the model. In Section 4.2, a technique referred to as *principal component analysis* is described. This technique helps in selecting variables or determining derived variables to use in any model. Sections 4.3–4.7 describe a series of widely used modeling approaches, including multiple linear regression, discriminant analysis, logistic regression, and naive Bayes.

4.1.2 Testing Model Accuracy

In order to assess which predictive data mining approach is most promising, it is important to assess the various options in a way that is objective and consistent. Evaluating the different approaches also helps set expectations about performance levels for a model ready to be deployed. In evaluating a predictive model, different data sets should be used to build the model and to test the performance of the model. Using different data ensures that the model has not overfitted the training data. The following approaches are commonly used to achieve this:

- *Test data*: Before the data set is used to train any models, a set of data selected randomly is set aside for the sole purpose of testing the quality of the results, such as one-third of the data set. These observations will not be used in building the model, but they will be used with any built model to test the model's predictive performance. It should be noted that, in the ideal case, the test set is only used for model assessment. However in practical situations, there may not be enough data available.
- *Cross-validation*: The same set of observations can be used for both training and testing a model, but not at the same time. In the cross-validation approach, a percentage of the data set is assigned for test purposes. Then, multiple training

and validation set combinations are identified to ensure that all observations will be a member of one validation set, and hence there will be a prediction for each observation. The assignment to the training and validation sets is random, and all validation sets will be mutually exclusive and approximately the same size. As an example, if the validation set size percentage is 10%, one-tenth of the data set will be set aside for testing and the remaining nine-tenths used for training the model. Under this scenario, 10 models must be built to ensure that all observations will be tested.

4.1.3 Evaluating Regression Models' Predictive Accuracy

The following section discusses methods used to assess the accuracy of a regression model, that is, a model built to predict a continuous variable. One effective way to visualize the accuracy of the model is to draw a scatterplot of the actual response values against the predicted response values. In Fig. 4.4 a model was built with the actual response values (y) plotted against the predicted values ($\hat{y}$). The reference line indicates where values would lie if the model made a perfect prediction, that is, when the predicted values are equal to the actual values. A good model has points close to the line, like the model displayed in Fig. 4.4.

A number of methods can be used to assess the model. The *error* or *residual* refers to the difference between the actual response value (y_i) and the predicted response value ($\hat{y}_i$). To quantify the error over the entire test set, the squared or absolute error is used, thus avoiding using a negative number which would bias the overall error evaluation. The *mean square error* and the *mean absolute error* sum these errors and divide the sum by the number of observations (n). Both values provide a good indication of the overall error level of the model.

$$\text{mean square error} \quad \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}$$

$$\text{mean absolute error} \quad \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}$$

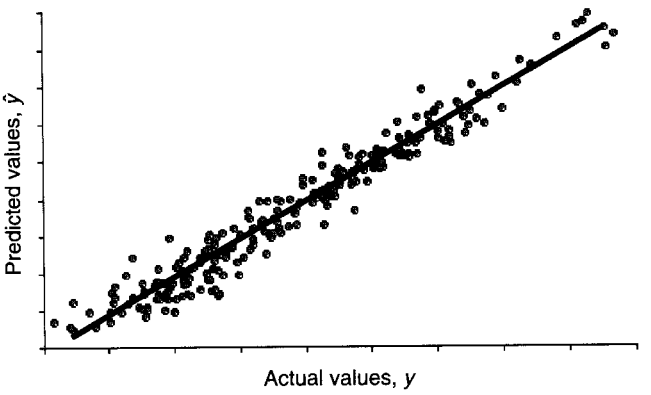

Figure 4.4 Scatterplot showing the actual values plotted against predicted values

Two additional approaches, *relative square error* and *relative absolute error*, normalize the overall error based on using the mean value ($\bar{y}$) as a simple prediction:

$$\text{relative square error} \quad \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\text{relative absolute error} \quad \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{\sum_{i=1}^{n}|y_i - \bar{y}|}$$

The *correlation coefficient* is a measure of the linear relationship between two continuous variables. In this situation, the variables are the actual values and the predicted values. The resulting values are always between -1 and $+1$ where strong positive linear relationships are signified by values close to $+1$, strong negative linear relationships are close to -1, and values close to 0 indicate a lack of any linear relationship. When this value is squared, the resulting range will be between 0 and 1. The equation uses the average value of both the actual values ($\bar{y}$), and the predicted values ($\bar{\hat{y}}$), as well as the standard deviation of the actual value (s_y), and the standard deviation of the predicted values ($s_{\hat{y}}$).

$$\text{correlation coefficient} \quad \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{(n-1)s_y s_{\hat{y}}}$$

Figure 4.5 displays the results from three models: models A, B, and C. The figure shows three different scatterplots of a model's predictions against the actual values. The closer the predicted values are to the actual values for the entire data set, the better the model is. Model A is the least predictive of the three, with model B providing a greater level of prediction, and model C showing the best level of accuracy.

The metrics described for assessing predictive accuracy are calculated for the three models and shown in Table 4.5. The first four values (mean square error, mean absolute error, relative square error, and relative absolute error) all have values that decrease with improved predictive accuracy. The correlation coefficient and the square correlation coefficient have values approaching 1 as the model accuracy improves.

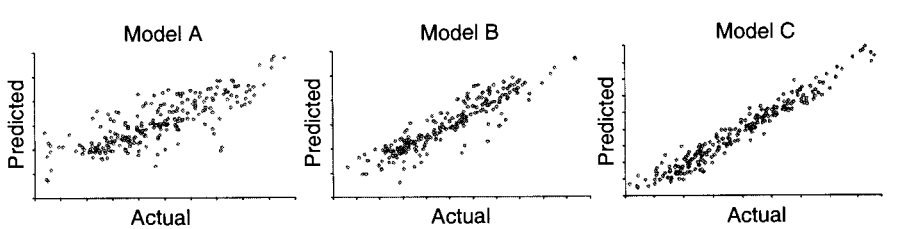

Figure 4.5 Scatterplots of predicted values vs actual values for three models

TABLE 4.5 Comparison of Model Accuracy for Three Models

	Model A	Model B	Model C
Mean square error	1.42	0.622	0.176
Mean absolute error	0.874	0.579	0.333
Relative square error	0.342	0.161	0.051
Relative absolute error	0.52	0.346	0.212
Correlation coefficient	0.811	0.916	0.974
Square correlation coefficient	0.658	0.839	0.949

4.1.4 Evaluating Classification Models' Predictive Accuracy

The overall *accuracy* of a classification prediction model can be estimated by comparing the actual values against those predicted, as long as there are a reasonable number of observations in the test set. An accuracy estimate is calculated using the number of correctly classified observations divided by the total number observations. This results in a number between 0 and 1, where values close to 1 indicates a high accuracy level. The *error rate* or *misclassification rate* is calculated using the number of observations incorrectly classified or one minus accuracy. A confusion matrix, or contingency table, is an effective way of viewing the accuracy of a classification model. For example, Fig. 4.6 shows a table illustrating the results of a classification model. The model's response is the categorical variable *cylinders* which can take five values: 3, 4, 5, 6, and 8. The actual values are shown on the x-axis and the predicted values are shown on the y-axis. In this example, four observations are predicted as 3; however, only three values are correctly predicted. A single value is incorrectly predicted as 3, when in fact it is 4.

The total number of correctly classified observations can be determined by summing the counts on the diagonal. In Fig. 4.6, that would include 3, 170, 0, 45, and 103, which equal 321. To calculate the overall accuracy, the correct 321

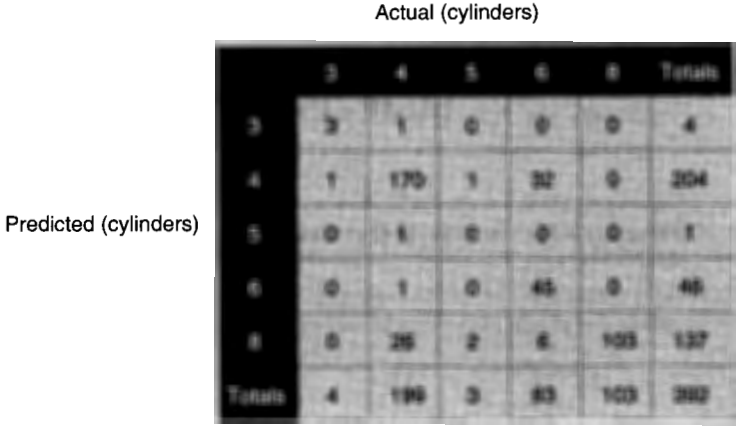

Figure 4.6 Contingency table showing predicted values against actual values

observations should be divided by the 392 total number of observations, which equals 0.82. One minus this accuracy level is the error rate, or 0.18. Good classification models have high values along the diagonals in the contingency table.

4.1.5 Evaluating Binary Models' Predictive Accuracy

In many situations, prediction models are built for a binary response variable. For example, a model may be built to predict whether an insurance application is fraudulent or not. The ability to predict a fraudulent case may be more important than predicting a nonfraudulent case, so it makes sense to look at the model results in more detail. In this situation, models that minimize false negatives should be selected over those that maximize accuracy.

In the following section, the results from prediction models with a binary response are assessed in greater detail. Counts for the following four properties are initially required:

- *True positive (TP)*: The number of observations predicted to be true (1) that are in fact true (1).
- *True negative (TN)*: The number of observations predicted to be false (0) that are in fact false (0).
- *False positive (FP)*: The number of observations that are incorrectly predicted to be positive (1), but which are in fact negative (0).
- *False negative (FN)*: The number of observations that are incorrectly predicted to be negative (0), but which are in fact positive (1).

These four alternatives are illustrated in the contingency table, or confusion matrix, shown in Table 4.6.

The following values can be calculated to assess the quality of a binary classification prediction model:

- *Accuracy*: The overall accuracy of the model can be calculated based on the number of correctly classified examples divided by the total number of observations,

$$\frac{TP + TN}{TP + FP + FN + TN}$$

TABLE 4.6 Contingency Table Showing the Four Possible Situations

		Actual response	
		Positive (1)	Negative (0)
Prediction	Positive (1)	TP	FP
	Negative (0)	FN	TN

- *Error rate*: The error rate, or misclassification rate, is 1 minus the accuracy value,

$$1 - \frac{TP + TN}{TP + FP + FN + TN}$$

- *Sensitivity*: This is the *true positive rate*, also referred to as the *hit rate*, or *recall*. It is calculated using the number of observations identified as true positives, divided by the actual number of positive observations (TP + FN),

$$\frac{TP}{TP + FN}$$

- *Specificity*: This is the number of negative observations that are correctly predicted to be negative, or the *true negative rate*. It is calculated using the number of correctly predicted negative observations, divided by the total number of actual negative observations (TN + FP),

$$\frac{TN}{TN + FP}$$

- *False positive rate*: This value is the same as 1 minus the sensitivity and is calculated using the number of incorrectly predicted negative observations divided by the actual number of negative observations (FP + TN),

$$\frac{FP}{FP + TN}$$

- *Positive predictive value*: This value is also called *precision*, and it is the number of correctly predicted positive observations divided by the total number of predicted positive observations (TP + FP),

$$\frac{TP}{TP + FP}$$

- *Negative predictive value*: This value is the total number of correctly predicted negative observations divided by the number of negative predictions (TN + FN),

$$\frac{TN}{TN + FN}$$

- *False discovery rate*: This value is the number of incorrectly predicted positive observations divided by the number observations predicted positive (FP + TP),

$$\frac{FP}{FP + TP}$$

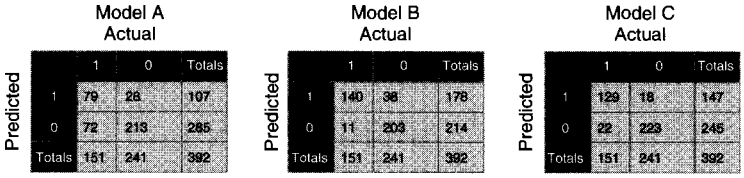

Figure 4.7 Summary of three different models

TABLE 4.7 Comparison of Different Metrics Across Three Models

	Model A	Model B	Model C
Accuracy	0.75	0.88	0.90
Error	0.26	0.13	0.10
Sensitivity	0.52	0.93	0.86
Specificity	0.88	0.84	0.93
False positive rate	0.12	0.16	0.07
Positive predictive value	0.74	0.79	0.88
Negative predictive value	0.75	0.95	0.91
False discovery rate	0.26	0.21	0.12

Figure 4.7 shows the results from three binary classification models: models A, B, and C. These models show the number of correctly, as well as incorrectly, classified observations, including false positives and false negatives.

Table 4.7 presents an assessment of the three models using the metrics detailed in this section. The overall accuracy and error rate of the models are summarized in the accuracy and error metric. In general, model C is most accurate, followed by model B, and then model A. The metrics also assess how well the models specifically predict positives, with model B performing the best based on the sensitivity score. Model C has the highest specificity score, indicating that this model is the best of the three at predicting negatives.

These different metrics are used in different situations, depending on the goal of the specific project.

4.1.6 ROC Charts

A receiver operating characteristics, or ROC, curve provides an assessment of one or more binary classification models. This chart plots the true positive rate or sensitivity on the y-axis and the false positive rate or 1 minus specificity on the x-axis. Usually a diagonal line is plotted as a baseline, that is, where a random prediction would lie. For classification models that generate a single value, a single point can be plotted on the chart. A point above the diagonal line indicates a degree of accuracy that is better than a random prediction. Conversely, a point below the line indicates that the prediction is worse than a random prediction. The closer the point is to the upper top left point in the chart, the better the prediction.

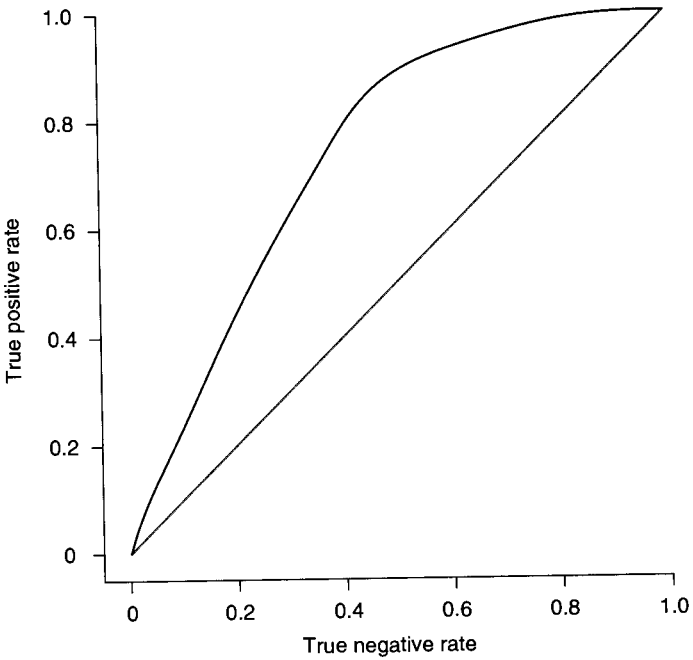

Figure 4.8 ROC chart for a model

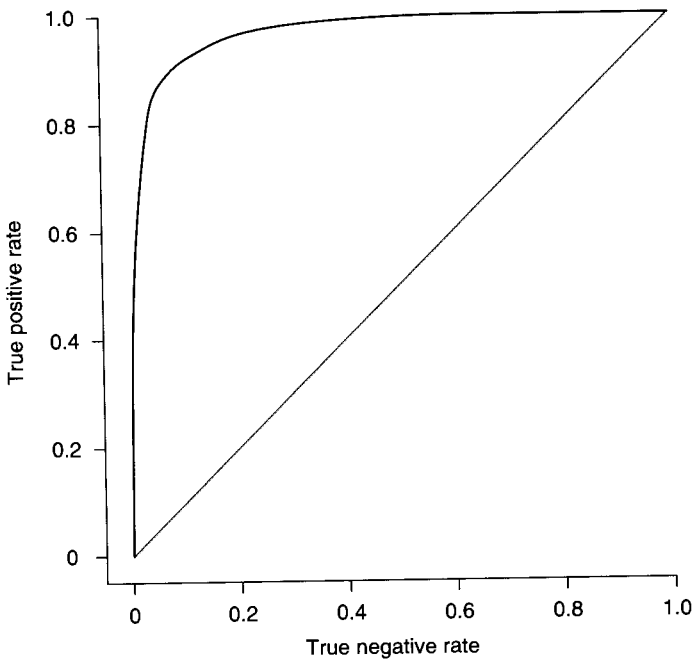

Figure 4.9 ROC chart of a good model

When a classification model generates a numeric value, such as a probability, then a classification can be made by specifying a cutoff threshold. Those numeric predictions above the cutoff are predicted positive, and those below are predicted to be negative. By building multiple models using different threshold cut-offs, a curve can be generated. For example, Fig. 4.8 presents an ROC curve for a model, and Fig. 4.9 shows an ROC curve for a model with a higher level of performance. The area under the curve (AUC) can be used to assess the model's accuracy.

4.1.7 Lift Chart

Many predictive analytics applications require the prediction of a binary response variable. For example, a direct mailing company may wish to predict which households will respond to a specific direct mailing campaign. Those that respond correspond to the positive outcome and those that do not respond correspond to the negative outcome. A predictive model can be built to generate the probability that

TABLE 4.8 Ordered Table of Cumulative Percentages of Observations and Positives

Actual	Prediction probability	Cumulative percent of all observations	Cumulative percentage of positives
1	1	0.3%	0.4%
1	1	0.5%	0.9%
1	1	0.8%	1.3%
1	1	1.0%	1.7%
1	1	1.3%	2.2%
...	...	...	...
1	0.997	25.0%	42.2%
1	0.997	25.3%	42.7%
1	0.997	25.5%	43.1%
1	0.997	25.8%	43.5%
1	0.997	26.0%	44.0%
...	...	...	...
1	0.839	50.0%	81.5%
1	0.837	50.3%	81.9%
1	0.836	50.5%	82.3%
1	0.835	50.8%	82.8%
1	0.834	51.0%	83.2%
...	...	...	...
0	0.0507	75.0%	99.6%
0	0.0499	75.3%	99.6%
0	0.0495	75.5%	99.6%
0	0.0494	75.8%	99.6%
0	0.0478	76.0%	99.6%
...	...	...	...

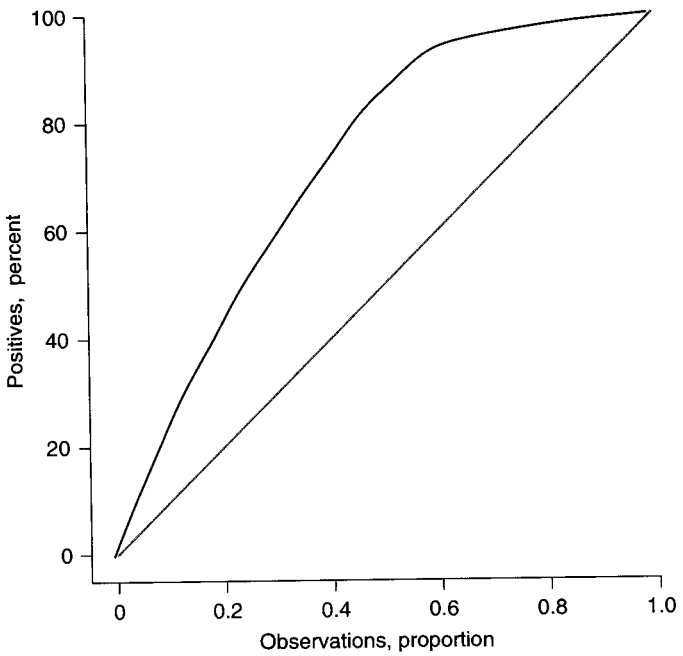

Figure 4.10 Lift chart

a customer will respond. This number would allow the direct mailing company to prioritize customers from a long list of potential households.

A *lift chart* can help to understand how to use a particular model. The chart is based on the original data along with the probability of a positive event generated from a model built from historical data. If this table is ordered according to this probability, those observations with the greatest likelihood of being positive will be at the top of the list. For example, Table 4.8 shows part of a table where a predictive model has been generated for a binary response variable. The actual response is in the column *actual*, and the predicted probability that an observation is true is in the column *probability*. The values in *cumulative percentage of observations* and *cumulative percentage of positives* have been calculated. These last two columns are plotted as a lift chart, as shown in Fig. 4.10. A diagonal line has been included, representing a random outcome.

The lift chart in Fig. 4.10 shows that using the top 50% of the ranked observations will result in approximately 80% of the total positives. For example, if this data represents the direct mailing example, then targeting only the top-ranked 50% of households will result in reaching most of those who will respond. The lift can be calculated at any point, using the target response and dividing it by the average response. In this example, at 50%, the target response is 80%, with an average response at 50%, which gives a lift of 1.6, which is 1.6 times better than not using the model.

4.2 PRINCIPAL COMPONENT ANALYSIS

4.2.1 Overview

Often, data mining projects involve a data set with a large number of continuous variables. If a data set has too many variables, it can be difficult to understand. In addition, the use of all variables in any analysis may introduce a host of logistical and accuracy problems. For example, using a large number of variables may increase the time to compute a particular model beyond an acceptable threshold. Using too many independent variables in a particular model may also impair the accuracy or reliability of a model by overfitting the data. There may even be logistical problems in collecting the values of many variables when the model is deployed.

Principal component analysis provides a method for understanding the meaning of a data set by extracting a smaller series of important components that account for the variability in the data. Each of these factors or *principal components* considers a subset of the variables to be important. For example, a data set containing information on home loans may contain a host of information about an individual, such as salary information, current home price, credit card debt, credit score, and so on. These variables could then be analyzed using principal component analysis. The analysis may group the variables in a number of different ways. For example, variables such as current home price and salary information may be grouped together in a larger group indicating "wealth indicators," whereas credit card debt and credit score may be grouped together as "credit rating" factors.

The use of principal component analysis offers the following benefits:

- *Data set insight*: The process of generating and interpreting the results of a principal component analysis can play an important role in becoming familiar with a data set, and even questioning assumptions about the data. This process may help uncover major factors underlying the data.
- *Reducing the number of variables in the model*: Identifying a smaller set of variables is often helpful, and one approach is to select variables from each important principal component. Alternatively, new variables for each of the important principal components can be generated from the original variables and used as independent variables directly in any modeling exercise.

4.2.2 Principal Components

Each principal component represents a weighted combination of the observed variables. Although all variables in a dataset are combined in a specific principal component, the weights reflect the relative importance of each variable within each principal component. For example, Fig. 4.11 displays a series of weights for five principal components (PC1–PC5) extracted from a dataset of five variables [*age* (*years*), *weight* (*lbs*), *height* (*inches*), *abdomen* (*cm*), *ankle* (*cm*)]. Each *weight*, or *loading*, within a principal component reflects the relative importance of the variable, and these values fall within the range of -1 to $+1$. For example, in the second principal component, PC2, *age* (*years*) has a strong negative score of -0.781 whereas *weight* (*lbs*) is given a score close to 0.

Principal components	PC1	PC2	PC3	PC4	PC5
Age, year	–0.0186	–0.781	–0.542	0.291	0.105
Weight, lb	0.599	–0.0858	0.0951	–0.28	0.739
Height, in	0.374	0.464	–0.774	–0.104	–0.189
Abdomen, cm	0.521	–0.386	0.223	–0.361	–0.632
Ankle, cm	0.48	0.134	0.221	0.834	–0.0853

Figure 4.11 Five principal components

Principal component analysis produces the same number of components as variables. However, each principal component accounts for a different amount of the variation in the data set. In fact, only a small number of principal components usually account for the majority of the variation in the data. The first principal component accounts for the most variation in the data. The second principal component accounts for the second highest amount of variation in the data, and so on.

Principal component analysis attempts to identify components that are independent of one another; that is, they are not correlated. The first principal component accounts for the largest amount of variation in the data. The second principal component is not correlated to the first; that is, it is *orthogonal* to the first principal component as well as accounting for the second largest remaining variation in the data. The other principal components are generated using the same criteria.

4.2.3 Generating Principal Components

Like most data analysis exercises, principal component analysis starts with a data table comprising a series of observations. Each observation is characterized by a number of variables. The first step in generating the principal components is to construct either a correlation matrix or a covariance matrix. If a covariance matrix is used, the original data may need to be normalized to ensure all variables are on a consistent range. If a correlation matrix is used, this matrix is generated by computing a correlation coefficient (r) for each pair of variables (see Section 3.2.5). For example, Fig. 4.12 shows a correlation matrix formed from a series of 13 variables: *age (year)*, *weight (lbs)*, and so on. The variable *age (years)* is correlated with each other variable, shown in the first row and first column. For example, the correlation coefficient between *age (years)* and *weight (lbs)* is -0.0125.

	Age, year	Weight, lb	Height, in	Neck, cm	Chest, cm	Abdomen, cm	Hip, cm	Thigh, cm	Knee, cm	Ankle, cm	Biceps, cm	Forearm, cm	Wrist, cm
Age, year	1	–0.0161	–0.246	0.119	0.182	0.243	–0.0581	–0.216	0.0172	–0.11	–0.0441	–0.0851	0.218
Weight, lb	–0.161	1	0.513	0.81	0.891	0.874	0.933	0.852	0.843	0.581	0.785	0.683	0.725
Height, in	–0.246	0.513	1	0.325	0.224	0.167	0.397	0.35	0.513	0.395	0.319	0.322	0.397
Neck, cm	0.119	0.81	0.325	1	0.769	0.728	0.708	0.669	0.648	0.434	0.709	0.661	0.731
Chest, cm	0.182	0.891	0.224	0.769	1	0.91	0.825	0.708	0.698	0.447	0.707	0.599	0.644
Abdomen, cm	0.243	0.874	0.167	0.728	0.91	1	0.861	0.737	0.71	0.407	0.656	0.53	0.602
Hip, cm	–0.0581	0.933	0.397	0.708	0.825	0.861	1	0.881	0.809	0.521	0.722	0.603	0.626
Thigh, cm	–0.216	0.852	0.35	0.669	0.708	0.737	0.881	1	0.777	0.504	0.744	0.604	0.544
Knee, cm	0.0172	0.843	0.513	0.648	0.698	0.71	0.809	0.777	1	0.585	0.654	0.579	0.656
Ankle, cm	–0.11	0.581	0.395	0.434	0.447	0.407	0.521	0.504	0.585	1	0.449	0.429	0.545
Biceps, cm	–0.0441	0.785	0.319	0.709	0.707	0.656	0.722	0744	0.654	0.449	1	0.701	0.614
Forearm, cm	–0.0851	0.683	0.322	0.661	0.599	0.53	0.603	0.604	0.579	0.429	0.701	1	0.598
Wrist, cm	0.218	0.725	0.397	0.731	0.644	0.602	0.626	0.544	0.656	0.545	0.614	0.598	1

Figure 4.12 Correlation matrix

Variance explained:

Principal components	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Eigenvalues	8.05	1.46	0.889	0.661	0.58	0.318	0.284	0.262	0.199	0.143	0.0773	0.0603	0.0181
Percentage	61.9%	11.2%	6.84%	5.08%	4.46%	2.45%	2.18%	2.02%	1.53%	1.1%	0.595%	0.464%	0.139%

Loadings:

Age, year	0.00574	0.726	−0.427	−0.079	0.073	−0.359	0.033	−0.108	0.14	−0.298	−0.0922	−0.138	−0.0416
Weight, lb	0.345	−0.0213	0.0288	−0.126	0.0993	0.0773	−0.122	−0.073	−0.0883	0.00318	−0.00717	−0.0544	−0.904
Height, in	0.166	−0.488	−0.517	−0.156	0.555	−0.0241	−0.157	−0.241	−0.0201	−0.156	−0.0999	0.0648	0.177
Neck, cm	0.299	0.137	−0.0323	0.273	0.126	0.555	0.0319	−0.0149	0.683	−0.0243	−0.133	0.0648	0.0649
Chest, cm	0.311	0.237	0.144	−0.11	0.0093	0.173	−0.343	−0.206	−0.201	−0.412	−0.483	−0.338	0.257
Abdomen, cm	0.305	0.281	0.179	−0.275	0.029	0.0561	−0.24	−0.0454	−0.123	−0.0597	0.221	0.751	0.162
Hip, cm	0.325	−0.0225	0.188	−0.268	0.0351	−0.0152	−0.0153	0.126	−0.188	−0.312	0.56	−0.529	0.207
Thigh, cm	0.306	−0.144	0.319	−0.151	−0.0454	−0.0976	0.326	0.204	0.142	−0.484	−0.582	0.066	0.0629
Knee, cm	0.306	−0.0769	−0.127	−0.26	0.029	−0.428	0.199	0.372	0.323	0.583	0.0896	0.0236	0.0291
Ankle, cm	0.221	−0.229	−0.369	−0.11	−0.807	0.0075	−0.152	−0.226	0.113	−0.101	0.0127	0.0121	0.0318
Biceps, cm	0.294	−0.0187	0.172	0.332	0.0268	−0.261	0.482	−0.651	−0.103	0.119	0.137	0.044	0.0467
Forearm, cm	0.264	−0.075	0.058	0.663	0.00797	−0.399	−0.491	0.255	−0.0166	−0.121	−0.00141	0.0215	0.0239
Wrist, cm	0.276	0.116	−0.418	0.257	−0.0514	0.322	0.374	0.379	−0.517	0.0492	−0.04	0.0811	0.0404

Figure 4.13 Extracted principal components along with eigenvalues

The next step is to extract the principal components from this correlation matrix (or covariance matrix), along with the amount of variation explained by each principal component. This is achieved by extracting *eigenvectors* and *eigenvalues* from the matrix (Strang, 2006). The eigenvector is a vector of weights or loadings. The eigenvalues represent the amount of variation explained by each factor. The principal components are then sorted according to the amount of variation they account for. Figure 4.13 illustrates a complete principal component analysis for the same 13 variables as used in Fig. 4.12. An eigenvalue is shown for each principal component along with a percentage of the variance explained by each component. The first principal component (PC1) accounts for 61.9% of the variation in the data, the second principal component (PC2) accounts for 11.2% of the variation in the data, and so on. The weights are also shown for each principal component. For the first principal component, a loading of 0.00574 is assigned to the *age* (*years*) variable, a loading of 0.345 is assigned to the *weight* (*lbs*) variable, and so on. The absolute value of each of the variable's loading values within each principal component represents its relative importance.

Additionally, a new variable can be generated from the original variables' values and the weights of the principal component. For each original data point, the mean for the variable must initially be subtracted from the value (mean centered) and then multiplied by the variable's weight. These values are then summed to create a new variable for each selected principal component. Figure 4.14 shows a scatterplot representing a derived score from principal component 1 against a derived score for principal component 2.

4.2.4 Interpretation of Principal Components

Since the objective of principal component analysis is to identify a small number of factors, the first step is to determine the specific number of principal components to use. Figure 4.15 shows a plot of the variance explained by each principal component, usually referred to as a *scree* plot. The first principal component accounts for the majority of the variation. The ideal number of factors is usually the number just prior to where on the graph the tail levels off. Selecting principal components after this point would add little additional information. In this example, the cutoff should be either at PC2 or PC3.

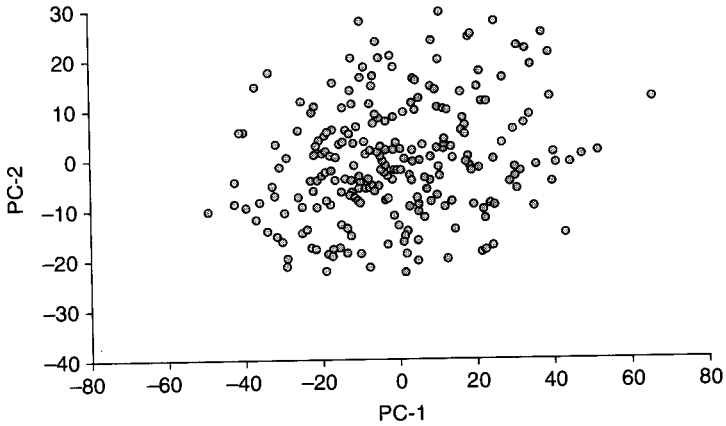

Figure 4.14 Scatterplot of two derived scores from the principal components

Having selected the number of principal components, a process called *rotation of the factors* can help interpret them, if the original analysis is unsatisfactory. This is achieved through a redistribution with these newly rotated principal components now containing loadings towards either $+1$ or -1, with fewer loading values in between. This process also redistributes the amount of variance attributable to each principal component. Methods such as *varimax* (Kaiser, 1958) will perform an optimization on the principal components to accomplish factor rotation. In Fig. 4.16, three

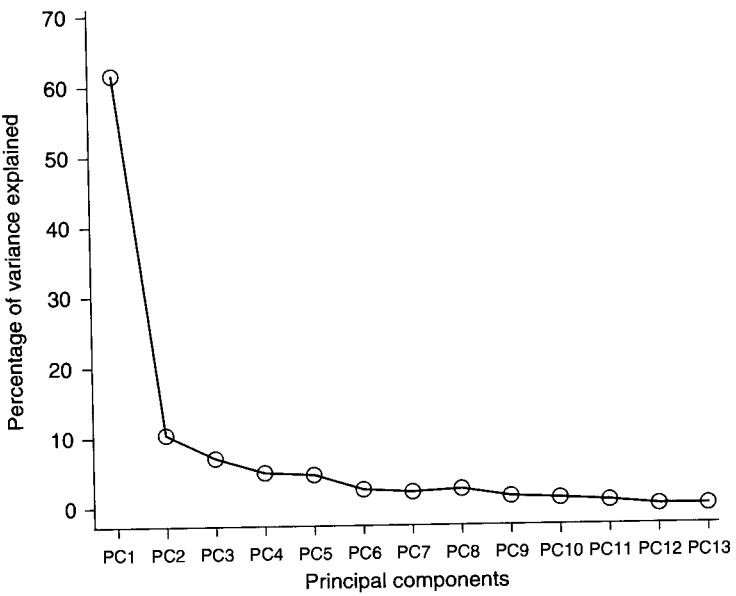

Figure 4.15 Scree plot for the percentage of the variance explained by the principal components

Variance explained:

Principal components	PC1	PC2	PC3
Eigenvalues	8.05	1.46	0.889
Percentage	77.4%	14%	8.55%

Loadings:

Age, year	0.00574	0.726	−0.427
Weight, lb	0.345	−0.0213	0.0288
Height, in	0.166	−0.468	−0.517
Neck, cm	0.299	0.137	−0.0323
Chest, cm	0.311	0.237	0.144
Abdomen, cm	0.305	0.281	0.179
Hip, cm	0.325	−0.0225	0.188
Thigh, cm	0.306	−0.144	0.319
Knee, cm	0.306	−0.0789	−0.127
Ankle, cm	0.221	−0.229	−0.369
Biceps, cm	0.294	−0.0187	0.172
Forearm, cm	0.264	−0.075	0.058
Wrist, cm	0.276	0.116	−0.418

Rotated factors:

Principal components	PC1 (rot)	PC2 (rot)	PC3 (rot)
Age, year	−0.00594	0.84	0.0532
Weight, lb	0.33	−0.0217	−0.105
Height, in	−0.074	−0.097	−0.706
Neck, cm	0.294	0.142	−0.0532
Chest, cm	0.374	0.128	0.132
Abdomen, cm	0.385	0.145	0.185
Hip, cm	0.359	−0.112	0.0239
Thigh, cm	0.36	−0.285	0.0674
Knee, cm	0.237	0.0156	−0.244
Ankle, cm	0.0606	0.0216	−0.483
Biceps, cm	0.325	−0.101	0.0242
Forearm, cm	0.253	−0.0853	−0.0839
Wrist, cm	0.154	0.338	−0.356

Figure 4.16 The original loadings along with the rotated factors

principal components were selected and rotated. In this example, by comparing the original values to the new rotated values, the second principal component, *age* (*years*), is now closer to +1 while the others, with the exception of *wrist* (*cm*), are closer to 0. Once a set of selected and rotated principal components is identified, the final step is to name them, using the weights as a guide to assist the analysis. Section 5.5 provides an example of the use of principal component analysis.

4.3 MULTIPLE LINEAR REGRESSION

4.3.1 Overview

Multiple linear regression analysis is a popular method used in many data mining projects for building models to predict a continuous response variable. This model defines

the linear relationship between a series of independent variables and a single response variable. It can be used to generate models, for example, to predict sales from transactional data, or to predict credit scores from information in a person's credit history.

The use of multiple linear regression analysis has a number of advantages, including:

- *Easy to understand*: Multiple linear regression models are easy to understand and interpret, because they are represented as a weighted series of independent variables. They can be effective in predicting new data as well as explaining what variables are influential within the data set.
- *Detect outliers*: In addition to this method's use as a prediction model, it can also help identify outliers, that is, those observations that do not follow a linear trend observed by the other entries.
- *Fast*: The generation of a multiple linear regression equation is fast, and it enables the rapid exploration of alternative variables since multiple models can be quickly built using different combinations of variables to determine an optimal model.

Despite being an effective method for prediction from and explanation of a data set, multiple linear regression analysis has a number of disadvantages, including:

- *Sensitivity to noise and outliers*: Multiple linear regression models are sensitive to noisy data as they try to find a solution that best fits all data, including the outliers. Outliers are erroneous pieces of data that can have especially undesirable consequences as the model tries to fit the potentially erroneous values.
- *Only linear relationships handled*: These models cannot model nonlinear datasets; however, the calculation of new variables can help in modeling. Transforming the independent and/or the response variables using mathematical transformation such as log, squared, cubed, square root, and so on, can help to incorporate variables with nonlinear relationships.

The simplest form of a linear regression is one containing a single independent variable, also referred to as *simple linear regression*. In this situation, the model can be drawn as a straight line through the data, plotted on a scatterplot. For example, Fig. 4.17 shows a scatterplot of two variables A and B, and a line drawn through them to represent a linear model. This linear model is represented by the formula for the straight line:

$$A = -0.27 + 1.02 \times B$$

Multiple linear regression analysis involves understanding the relationship between more than one independent variable and a single response variable. The analysis does not imply that one variable causes another variable to change; it only recognizes the presence of a relationship. This relationship is difficult to visualize when dealing with more than one or two independent variables. The relationship between the response variable and the independent variables for the entire population is assumed to be a linear equation of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_n + \varepsilon$$

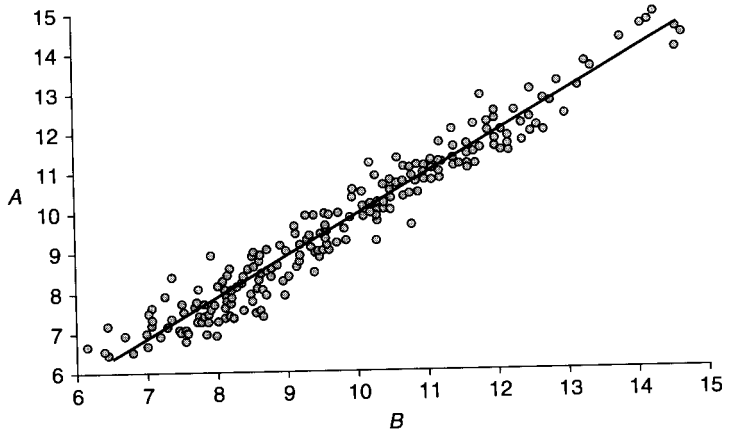

Figure 4.17 Illustration of a simple linear regression model, represented as a straight line

In this equation, y is the response, $\beta_0 - \beta_k$ are constant values referred to as *beta coefficients*, $x_1 - x_n$ are the input independent variables, and ε is a random error. Since the model will be built from a sample of the population, building a multiple linear regression model will estimate values for the beta coefficients, and the equation generated will look like:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_n$$

In this equation, $\hat{y}$ is the predicted response, $x_1 - x_n$ are the independent variables, and $\hat{\beta}_0$ through $\hat{\beta}_k$ are the estimated values of the beta coefficients.

For example, a multiple linear regression equation to predict a credit score, using two variables *LOANS* and *MISSED_PAYMENTS* may look like:

$$CREDIT\_SCORE = 22.5 + 0.5 \times LOANS - 0.8 \times MISSED\_PAYMENTS$$

A number of assumptions must be made when building a multiple linear regression model. These assumptions can be tested once the model has been built, as described later in this chapter. These assumptions are:

- *Linear*: A multiple linear regression will only generate models that describe a linear relationship between the independent variables and the response.
- *Homoscedasticity*: This refers to the assumption that the variation of error terms should be constant with respect to the independent variables; that is, there should be no relationship between the independent variable's variation and the error term.
- *Independence*: The error values should not be a function of any adjacent values, for example, to avoid errors that result from the passage of time.
- *Normally distributed error term*: The frequency distribution of the errors (predicted value minus the actual value) is assumed to follow a normal distribution.

The variables used as independent variables should not be correlated to one another. This situation, known as *multicolinearity*, will cause the models to fail. A scatterplot

can be used to check for multicolinearity. Correlations between continuous variables, as well as categorical variables, should be checked. An example of a multicolinearity situation involving continuous variables is a dataset used to predict a health condition. This data set may contain two or more tests that, in fact, measure the same phenomena, and hence only one should be included. A multicolinearity situation involving categorical variables involves a model used to predict the success of a marketing campaign. In this example, the independent variable *color* can take three values: red, green and blue. To use this variable within the model, the *color* variable is transformed into three dummy variables corresponding to each color, as discussed in Chapter 1. This results in the generation of three new variables; however, an observation where blue is 1 always occurs when red and green are 0. Hence, only two variables are really needed to capture all possible scenarios. The inclusion of all three variables would include correlations between the *color* variables and therefore all three should not be used. After a model is built, multicolinearity may be observed through interpreting the beta coefficients to look for, as an example, unexpected positive or negative numbers that do not reflect how the model is expected to behave.

The following sections discuss how a multiple linear regression model is built, and, once built, how its assumptions are tested through an analysis of the errors. Most statistical software packages generate a series of statistics concerning the models built, such as the *standard error* and the *coefficient of multiple determination*, along with metrics to assess the significance of the models and the parameters. Finally, in building a model, alternative combinations of variables should be used to build multiple models, and alternative data transformations (such as product of variables or variables squared) can be considered to improve the quality of the models built.

4.3.2 Generating Models

A multiple linear regression model is an equation describing the linear relationship between a response variable and a series of independent variables. The equation is a weighted sum of all variables, where $\hat{\beta}_1 - \hat{\beta}_k$ correspond to the weights and $x_1 - x_n$ correspond to the independent variables. $\hat{y}$ is the response variable being predicted and $\hat{\beta}_0$ is a constant added to the equation.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_n$$

The multiple linear regression formula can be rewritten using matrix multiplication:

$$y = X\hat{\beta}$$

The response variable, y, is a column vector of values, where $y_1 - y_n$ are the response values for the n observations:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{pmatrix}$$

The independent variables, X, are represented as a matrix, where n is the number of observations and k is the number of variables to be used as independent variables. The first column is all 1s and relates to the intercept $\hat{\beta}_0$.

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ 1 & x_{3,1} & x_{3,2} & \cdots & x_{3,k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix}$$

The β coefficients are also described as a vector, where $\hat{\beta}_0$–$\hat{\beta}_n$ are the individual coefficients:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \cdots \\ \hat{\beta}_n \end{pmatrix}$$

To generate a multiple linear regression model, estimates for the β coefficients are derived from the training data. The objective of the process is to identify the best fitting model for the data. A procedure referred to as *least squares* attempts to derive a set of coefficients to minimize the model's error (ε). This error is assessed using the *sum of squares of error* (SSE), such that:

$$\text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2$$

The formula calculated is based on the error (ε) squared. The error is squared so that positive and negative errors do not cancel each other out. The error is calculated from the difference between the predicted value ($\hat{y}_i$) and the actual response value (y_i):

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Replacing $\hat{y}_i$ with the equation for the multiple linear regression results in the following:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_k x_k)^2$$

This equation is then solved using calculus, and the details of this calculation are provided in Rencher (2002). The β coefficients can then be calculated using the following matrix formula:

$$\hat{\beta} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y$$

In this formula the superscript T represents a transposed matrix and the superscript -1 represents an inverse matrix. This calculation is always performed with a

TABLE 4.9 Table of Data Relating Rental Prices to Their Square Footage and Number of Baths

SQUARE_FEET	NOS_BATHS	RENTAL_PRICE
789	1	770
878	1	880
939	2	930
1100	2	995
1300	3	1115
1371	3	1300
1481	3	1550
750	1	560
850	1	610
2100	3	1775
1719	3	1450
1900	3	1650
1100	3	900
874	1	673
1024	2	785
1082	2	809

computer because of the complexity of the matrix operations. See Appendix A for more details.

In the following example, a multiple linear regression model is built to predict the rental price of apartments in a specific neighborhood. The response variable is *RENTAL_PRICE* and the independent variables used are *SQUARE_FEET* and *NOS_BATHS*. Table 4.9 is a data table to be used to build the model.

The independent variables, *SQUARE_FEET* and *NOS_BATHS*, are converted to a matrix, with the first column all 1s (corresponding to the intercept), the second columns *SQUARE_FEET* and the third column *NOS_BATHS*:

$$X = \begin{pmatrix} 1 & 789 & 1 \\ 1 & 878 & 1 \\ 1 & 939 & 2 \\ \cdots & \cdots & \cdots \\ 1 & 1082 & 2 \end{pmatrix}$$

The response variable, *RENTAL_PRICE*, is converted to a vector:

$$y = \begin{pmatrix} 770 \\ 880 \\ 930 \\ \cdots \\ 809 \end{pmatrix}$$

The β coefficients of the model are calculated using:

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$$

$$\hat{\beta} = \left[\begin{pmatrix} 1 & 789 & 1 \\ 1 & 878 & 1 \\ 1 & 939 & 2 \\ \dots & \dots & \dots \\ 1 & 1082 & 2 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} 1 & 789 & 1 \\ 1 & 878 & 1 \\ 1 & 939 & 2 \\ \dots & \dots & \dots \\ 1 & 1082 & 2 \end{pmatrix}\right]^{-1} \begin{pmatrix} 1 & 789 & 1 \\ 1 & 878 & 1 \\ 1 & 939 & 2 \\ \dots & \dots & \dots \\ 1 & 1082 & 2 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} 770 \\ 880 \\ 930 \\ \dots \\ 809 \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} -33.3 \\ 0.816 \\ 46.5 \end{pmatrix}$$

The equation for the model relating the square feet and number of baths for apartments to the rental price is therefore:

$$RENTAL\_PRICE = -33.3 + 0.816 \times SQUARE\_FEET + 46.5 \times NOS\_BATHS$$

4.3.3 Prediction

To make a prediction of the rental price for an apartment with 912 square feet and one bathroom, these values are substituted into the equation, thus resulting in a prediction of 757 for the *RENTAL_PRICE*:

$$RENTAL\_PRICE = -33.3 + 0.816 \times SQUARE\_FEET + 46.5 \times NOS\_BATHS$$
$$RENTAL\_PRICE = -33.3 + 0.816 \times 912 + 46.5 \times 1$$
$$RENTAL\_PRICE = 757$$

The coefficients define the rate at which the model's prediction will change as the independent variables change, when all other independent variables are kept the same. The higher the coefficient, the greater the change. For example, increasing the number of baths in this example to 2, while keeping the square feet the same will increase the rental price by 46.5.

It is usual to make a prediction from data within the same range as the data used to build the model.

4.3.4 Analysis of Residuals

Once a model has been built, a prediction can be computed for each observation by using the actual values for the x-variables within the model and calculating a predicted value for the y variables ($\hat{y}$). For example, in the model previously built for the apartment rental example, a prediction has now been calculated using the regression model and the actual x-variables, *SQUARE_FEET* and *NOS_BATHS*. Table 4.10 shows the predicted values along with the residuals.

TABLE 4.10 Calculations of Predictions and Residuals

RENTAL_PRICE, y	SQUARE_FEET, x_1	NOS_BATHS, x_2	PREDICTION, $\hat{y}$	RESIDUAL, ERR
770	789	1	657	113
880	878	1	729	151
930	939	2	825	105
995	1100	2	957	38
1115	1300	3	1166	−51
1300	1371	3	1224	76
1550	1481	3	1314	236
560	750	1	625	−65
610	850	1	706	−96
1775	2100	3	1819	−44
1450	1719	3	1508	−58
1650	1900	3	1656	−6
900	1100	3	1003	−103
673	874	1	726	−53
785	1024	2	895	−110
809	1082	2	942	−133

For each predicted value, the difference between what the model predicts and the actual value is referred to as the *error* or the *residual*. For example, in Table 4.10, the first observation has an actual value for *RENTAL_PRICE* of 770, and a predicted value of 657. The difference between these two values reflects the error or residual.

$$\text{residual} = 770 - 657 = 113$$

A residual has been calculated for each observation in the data set, as shown in Table 4.10. Since the regression model has been calculated to minimize errors, the sum of all residual values should equal zero. Analysis of residuals is helpful in testing the following underlying assumptions, also mentioned above, of multiple linear regression, that is:

- *Linear*: A multiple linear regression will only generate linear models, and hence understanding whether the relationship is in fact linear is important. This can be seen in Fig. 4.18, where the residual is plotted against the y-variable and a clear "U"-shape can be seen, indicating a nonlinear relationship. Plotting the residual against the individual independent variables will help to identify nonlinear relationship that could be rectified with a mathematical transformation, such as a quadratic term.
- *Homoscedasticity*: This refers to the assumption that the variation of error terms should be constant with respect to the independent variables. This can be tested by plotting, for example, the predicted variable against the residual. If there is a trend indicating a nonconstant variance, as shown in Fig. 4.19, then the underlying assumption of homoscedasticity is not valid.
- *Independence*: The error values should not be a function of any adjacent values, and this can easily be tested by plotting the residual values against

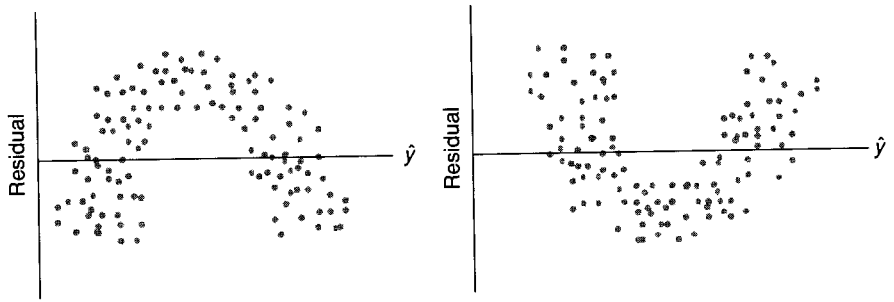

Figure 4.18 Nonlinear relationship shown by plotting the predicted variable against the residual

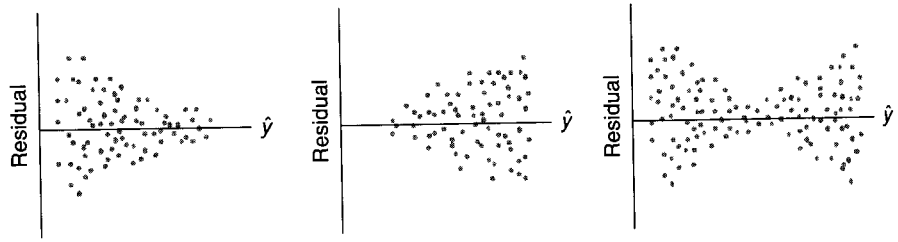

Figure 4.19 A model that violates the homoscedasticity assumption, as seen by plotting the residual against the prediction

the order in which the values were collected. In Fig. 4.20, the residual values have been plotted against the order the observations were taken, and a clear trend is discernable in both graphs, indicating that the assumption of independence is violated. This error may have been introduced as a result of the measurements being taken over time.

- *Normally distributed error term*: Examining the frequency distribution of the residuals, for example, using a frequency histogram or a q–q plot (as discussed in Chapter 2), is helpful in assessing whether the normal distribution assumption is violated. This assumption is usually required to enable computations of confidence intervals, which are not required in many data mining applications.

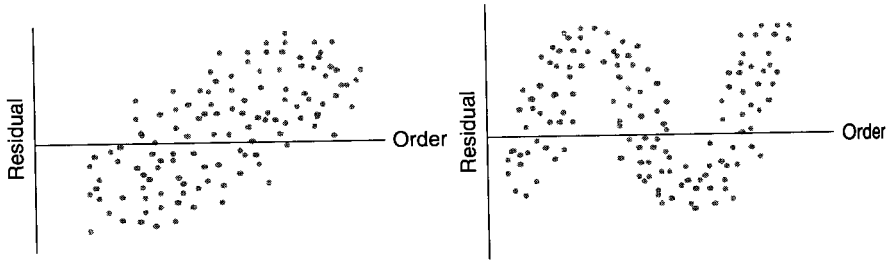

Figure 4.20 Plot of residual against the order the observations were collected, indicating a violation of the assumption of independence

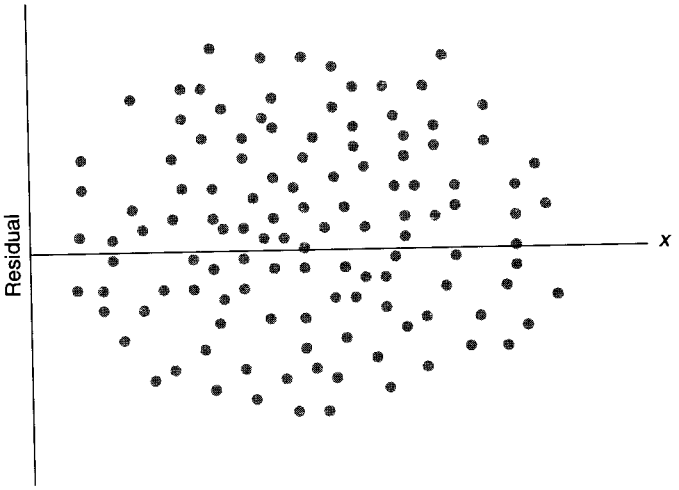

Figure 4.21 Residuals plotted against the x-variable with no discernable trend

There should be no discernable trend in the residual plot of any of the independent variables against the errors, as shown in Fig. 4.21.

The analysis of residuals is helpful in determining whether there are any clear violations in the assumptions. An analysis of the residual plots can also be helpful in identifying observations that do not fit the model, that is, those observations with unusually high positive or negative residual values. These outlier observations may be attributable to errors in the data and should be examined in more detail to determine whether to remove them.

4.3.5 Standard Error

An evaluation of the residuals in the model can help in understanding whether the model is violating any assumptions. An overall assessment of the model error is computed using the SSE. The following formula, as described earlier, is used to calculate the SSE:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Using Table 4.10, the SSE can be calculated as:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{SSE} = (770 - 113)^2 + (880 - 729)^2 + \cdots + (809 - 942)^2$$

$$\text{SSE} = 174{,}096$$

The distribution or spread of the residual values can also be useful in assessing the model. This is achieved by calculating the standard deviation of the residual or the

standard error of the estimate S_e. Assuming that the error terms are normally distributed, approximately 68% of errors will be within one standard deviation, and approximately 95% of errors will be within two standard deviations. The formula for the standard error of the estimate S_e is:

$$S_e = \sqrt{\frac{\text{SSE}}{n-k-1}}$$

where SSE is the sum of the squares of errors, n is the number of observations, and k is the number of independent variables.

In the example model, where SSE is 174,096, the number of observations is 16 (n), and the number of independent variables or k is 2, S_e is:

$$S_e = \sqrt{\frac{\text{SSE}}{n-k-1}}$$

$$S_e = \sqrt{\frac{174{,}096}{16-2-1}}$$

$$S_e = 112$$

This value can help in assessing whether the model is sufficiently accurate. Assuming a normal distribution, approximately 68% of errors should be within one standard deviation or ± 112 and approximately 95% of errors should be within two standard deviations, that is, ± 224.

4.3.6 Coefficient of Multiple Determination

Most statistical software packages that perform a multiple linear regression analysis also calculate the coefficient of multiple determination, or R^2. This coefficient is used to assess how much of the variation in the response is explained by the model. It is determined using the difference between the variance in the data about a naive model, where the mean response is used as the model, against the variance attributable to the fitted model. The value for R^2 varies between 0 and 1, with a high value indicating that a significant portion of the variance in the response is explained by the model. The formula to calculate R^2 is based on the SSE and the *total sum of square* (SST), or error about a naive model. To calculate SST, using $\bar{y}$ as the mean value for y:

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The sum of squares of error has been previously discussed and is calculated using:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The coefficient of determination is therefore calculated using the difference in the variation explained by the fitted model and the naive model (explained variability) as a proportion of the total error:

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}}$$

In the apartment rental example, using a mean *RENTAL_PRICE* value of 1047 ($\bar{y}$), the SST can be calculated as:

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\text{SST} = (770 - 1047)^2 + (880 - 1047)^2 + \cdots + (809 - 1047)^2$$

$$\text{SST} = 2{,}213{,}566$$

Using Table 4.10, SSE can be calculated as:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\text{SSE} = (770 - 113)^2 + (880 - 729)^2 + \cdots + (809 - 942)^2$$

$$\text{SSE} = 174{,}096$$

In the apartment example, the percentage of variation in the response as explained by the model is therefore:

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}}$$

$$R^2 = \frac{2{,}213{,}566 - 174{,}096}{2{,}213{,}566}$$

$$R^2 = 0.92$$

An increasing number of independent variables result in an R^2 value that is overestimated. An adjusted R^2 or R^2_{adj} is usually calculated to more accurately reflect the number of independent variables, as well as the number of observations:

$$R^2_{\text{adj}} = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2)$$

In the apartment rental example, there are 16 observations (n) and two independent variables (k), and the following value of R^2_{adj} is calculated:

$$R^2_{\text{adj}} = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R^2)$$

$$R^2_{\text{adj}} = 1 - \left(\frac{16-1}{16-2-1}\right)(1-0.92)$$

$$R^2_{\text{adj}} = 0.908$$

Usually, R^2_{adj} values are slightly less than R^2 values.

4.3.7 Testing the Model Significance

Assessing the significance of the relationship between the independent variables and the response is an important step. An F-test is most often used, based on the following hypothesis:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_K = 0$

H_a: At least one of the coefficients is not equal to 0

The null hypothesis states that there is no linear relationship between the response and the independent variables. If the null hypothesis is rejected, it is determined that there is a significant relationship. An F-test is performed using the mean square regression (MSR) and the mean square error (MSE). The formula for MSR is:

$$\text{MSR} = \frac{\text{SSR}}{k}$$

The formula for MSE is:

$$\text{MSE} = \frac{\text{SSE}}{n-k-1}$$

The F-test is calculated using the formula:

$$F = \frac{\text{MSR}}{\text{MSE}}$$

The *regression sum of squares* (SSR) is calculated using the following formula:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2$$

For the apartment example, it is calculated as 2,040,244. Using this value, the MSR is:

$$\text{MSR} = \frac{\text{SSR}}{k}$$

$$\text{MSR} = \frac{2{,}040{,}244}{2}$$

$$\text{MSR} = 1{,}020{,}122$$

Using the previously calculated value for the SSE, the value for MSE is:

$$\text{MSE} = \frac{\text{SSE}}{n - k - 1}$$

$$\text{MSE} = \frac{174{,}096}{16 - 2 - 1}$$

$$\text{MSE} = 76$$

An F-value of 76 is calculated and this number is compared to the critical F-value in order to determine whether the null hypothesis is rejected. The critical value is based on the level of significance (α), the degrees of freedom of the regression (k), and the degrees of freedom of the error ($n - k - 1$). Assuming a level of significance of 0.01, the critical value for $F_{0.01,2,14}$ is 6.51, using a standard F-distribution table (see Myatt, 2007). Since the computed F-value is greater than the critical value, the null hypothesis is rejected. A p-value is usually computed in most statistical software packages and can also be used to make this assessment.

In addition to assessing the overall model, each individual coefficient can be assessed. A t-test is usually performed, based on the following hypothesis:

$$H_0: \beta_j = 0$$
$$H_a: \beta_j \neq 0$$

The null hypothesis states that the coefficient is not significant. In the apartment rental example, the independent variable *SQUARE_FEET* has a calculated t-value of 6.53 and a p-value of almost 0, indicating the significance of this variable; however, the *NOS_BATHS* variable has a t-value of 0.803 with a p-value of 0.44, indicating that this variable is less significant within the model for some reason.

4.3.8 Selecting and Transforming Variables

Calculating which variable combinations result in the best model is often determined by evaluating different models built with different combinations of independent variables. Each model may be checked against an indication of the quality of the model, such as R^2_{adj}. An *exhaustive search* of all possible variable combinations is

one approach to identifying the optimal set of independent variables. This approach, however, can be time-consuming and methods such as *forward selection*, *backward selection*, and *stepwise selection* provide faster methods for identifying independent variable combinations. These methods add and remove variables based on different rules, and they will identify solutions more quickly, with the risk of overlooking the best solution. The forward selection method adds independent variables one at a time, building on those additions that result in an increase in the performance of the model. The backwards selection method starts with all independent variables, and sequentially removes variables that do not contribute to the model performance. Finally the stepwise method can proceed in the forward or backward direction and assesses the contribution of the variables at each step. The Further Reading section of this chapter points to additional material on these approaches.

The following example illustrates the process of building multiple models with different sets of independent variables. In this example, a series of models were built to predict *percentage body mass* (the response variable), using up to four independent variables: *weight*, *chest*, *abdomen*, and *hip*. The exhaustive search method is used to generate models using all combinations of the four independent variables, and an adjusted R^2 is calculated for each model generated. It can be seen from Table 4.11 that a model built from two independent variables, *weight* and *abdomen*, yields a model with the highest adjusted R^2 value of 0.716.

When a model would violate one of the underlying assumptions, various mathematical transformations could be applied to either the independent variables or response variables, or both. Transformations such as the natural log, polynomials, reciprocals, and square roots can aid in building multiple linear regression models.

TABLE 4.11 Building Different Models with All Combinations of Independent Variables

Variable 1	Variable 2	Variable 3	Variable 4	R^2_{adj}
Weight				0.371
Chest				0.492
Abdomen				0.659
Hip				0.382
Weight	Chest			0.492
Weight	Abdomen			0.716
Weight	Hip			0.386
Chest	Abdomen			0.668
Chest	Hip			0.494
Abdomen	Hip			0.693
Weight	Chest	Abdomen		0.714
Weight	Chest	Hip		0.515
Weight	Abdomen	Hip		0.715
Chest	Abdomen	Hip		0.694
Weight	Chest	Abdomen	Hip	0.713

4.4 DISCRIMINANT ANALYSIS

4.4.1 Overview

Discriminant analysis is used to make predictions when the response variable is categorical (a classification model). For example, an insurance company may want to predict high or low risk customers, or a marketing department may want to predict whether a current customer will or will not buy a particular product. Discriminant analysis models will classify observations based on a series of independent variables. Figure 4.22 illustrates its use with a data set. Two variables are used to show the distribution of the data. The light circles belong to group A, and the dark circles represent observations in group B. The objective of the modeling exercise is to classify observations into either group A or group B. A straight line has been drawn to illustrate how, on one side of the line, the observations are assigned to group A and on the other they are assigned to group B. Discriminant analysis attempts to find a straight line that separates the two classes, and provides a method for assigning new observations to one of the classes.

The analysis becomes more complex in situations when the data set contains more independent variables, as well as when the response variable has more than two possible outcomes. In these situations, discriminant analysis attempts to identify hyperplanes that separate these multiple groups.

Discriminant analysis is a simple statistical technique that can be used to identify important variables that characterize differences between groups, as well as to build classification models. It is a useful classification method, especially for smaller data sets. The following is a summary of the key assumptions associated with using discriminant analysis:

- *Multivariate normal distribution*: The variables should have a normal distribution within the classes.

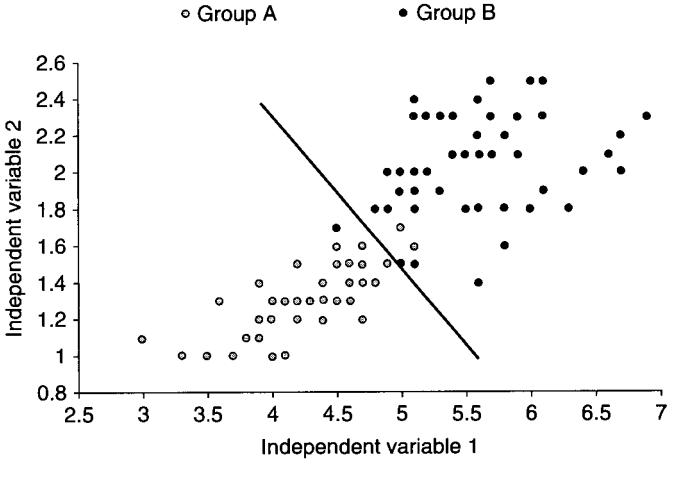

Figure 4.22 A straight line separating two classes (group A and group B)

- *Similar group covariance matrices*: In addition, correlations between and variances of the independent variables in each group used in the model should be similar.

Discriminant analysis is sensitive to outliers and will not operate well in situations where the size of one or more of the groups is small.

4.4.2 Discriminant Function

The method relies on the calculation of a *discriminant function* for each group. There will be k functions generated based on the number of unique categories the response variable can take. If the response variable is *color* with possible values red, green, and blue, then k will be 3. Predictions are made by calculating a score using each group's discriminant function. An observation is predicted to be a member of the group with the highest discriminant function score.

Similar to multiple linear regression, linear discriminant analysis attempts to identify a solution that minimizes the overall error. By making the assumptions described earlier, that is, a normal distribution, with similar group covariance matrices, the following formula can be used to estimate a linear discriminant function:

$$f_k = x^{\mathrm{T}} \hat{S}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^{\mathrm{T}} \hat{S}^{-1} \hat{\mu}_k + \log(\hat{p}_k)$$

In this formula, x is a vector of values for a single observation, $\hat{S}$ is an estimate of the covariance matrix, $\hat{\mu}_k$ is a vector of mean values for the variables corresponding to group k, and $\hat{p}_k$ is an estimate of the prior probability. The superscript -1 represents an inverse matrix and the superscript T represents a transposed matrix. More information on how this formula was derived can be found in Hastie (2003) and Rencher (2002).

One approach to calculating $\hat{S}$, the estimate of the covariance, is to calculate the covariance matrix (described in Section 3.3.5) for each of the groups and then combine the individual matrices into a single pooled covariance matrix.

$\hat{p}_k$ represents the prior probability and can be estimated using N_k, which is the number of observations in category k:

$$\hat{p}_k = \frac{N_k}{N}$$

4.4.3 Discriminant Analysis Example

As an example, a data set of wines that includes a number of their chemical properties can demonstrate these ideas (http://archive.ics.uci.edu/ml/datasets/Wine). Each wine has an entry for the variable *alcohol*, which can take three values: "1," "2," and "3," that relate to the wine's region. A discriminant analysis model will be built to predict this response. The wines are described using a number of independent variables: (1) *malic acid*, (2) *alkalinity of ash*, (3) *nonflavanoids*, and (4) *proline*. There are 198 observations and Table 4.12 presents a number of example observations, which are further summarized in Figs. 4.23–4.27. In Figs.

TABLE 4.12 Data Table Illustrating the Alcohol Classification

Alcohol	Malic acid	Alkalinity of ash	Nonflavanoids	Proline
1	14.23	2.43	3.06	1065
1	13.2	2.14	2.76	1050
1	13.16	2.67	3.24	1185
1	14.37	2.5	3.49	1480
1	13.24	2.87	2.69	735
1	14.2	2.45	3.39	1450
1	14.39	2.45	2.52	1290
1	14.06	2.61	2.51	1295
1	14.83	2.17	2.98	1045

4.23–4.26, the frequency distribution of the four independent variables is presented, with highlighting indicating the three alcohol classes. Similarly, Fig. 4.27 presents a scatterplot matrix of the four independent variables, with the three alcohol classes highlighted.

The three classes are initially summarized as shown in Table 4.13, where a count of the number of observations in each class is presented along with the mean for the four independent variables, corresponding to each class.

One approach to estimating the covariance matrix, detailed in Section 3.3.5, needed for the discriminant function is to calculate a covariance matrix for each group and then to pool the values. This results in the following covariance matrix:

$$\hat{S} = \begin{bmatrix} 0.67 & 0.048 & 0.19 & 166 \\ 0.048 & 0.076 & 0.032 & 19.5 \\ 0.19 & 0.032 & 1.01 & 157.4 \\ 166.2 & 19.5 & 157.4 & 100{,}249 \end{bmatrix}$$

This is inverted, resulting in the following matrix:

$$\hat{S}^{-1} = \begin{bmatrix} 2.62 & -0.56 & 0.23 & -0.0046 \\ -0.56 & 13.96 & -0.070 & -0.0017 \\ 0.23 & -0.070 & 1.33 & -0.0025 \\ -0.0046 & -0.0017 & -0.0025 & 0.000022 \end{bmatrix}$$

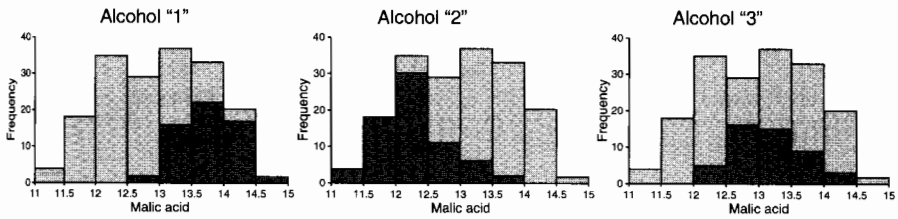

Figure 4.23 Three alcohol groups highlighted on the variable malic acid

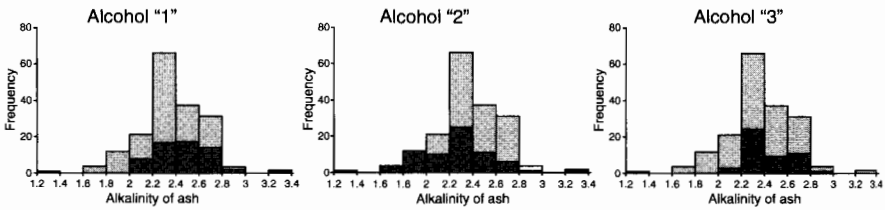

Figure 4.24 Three alcohol groups highlighted on the variable alkalinity of ash

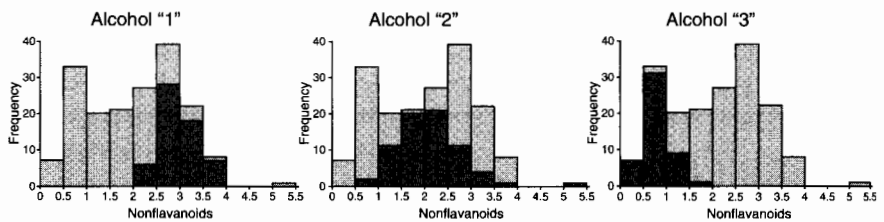

Figure 4.25 Three alcohol groups highlighted on the nonflavanoids variable

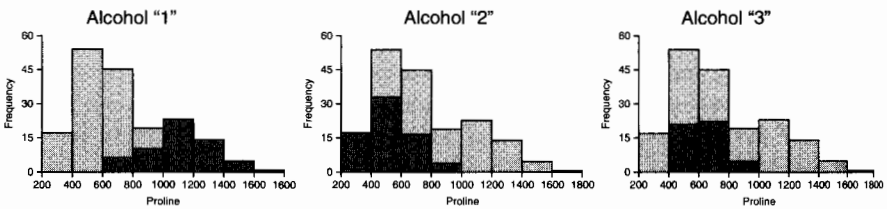

Figure 4.26 Three alcohol groups highlighted on the proline variable

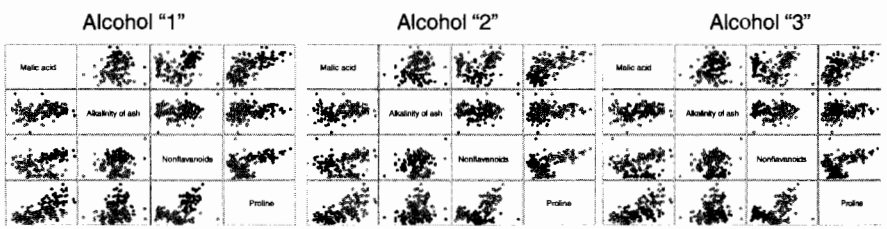

Figure 4.27 Three scatterplot matrices showing the three alcohol classes highlighted over a matrix of the four independent variables

TABLE 4.13 Summary Table of Different Wine Groups

Group name	Count	Mean malic acid	Mean alkalinity of ash	Mean nonflavanoids	Mean proline
"1"	59	13.7	2.46	2.98	1116
"2"	71	12.3	2.25	2.08	520
"3"	48	13.2	2.44	0.78	630

TABLE 4.14 An Observation to be Used in the Discriminant Analysis Model

Malic acid	Alkalinity of ash	Nonflavanoids	Proline
14.23	2.43	3.06	1065

The value p_k is the prior probability and can be estimated using N_k, which is the number of observations in category k:

$$p_k = \frac{N_k}{N}$$

For example, to calculate the prior probability for group 1:

$$p_1 = \frac{59}{198} = 0.33$$

Using this information, a discriminant function score can be calculated for each observation, for each class. In this example, three scores are calculated for each of the response categories, and the values for the independent variables are shown in Table 4.14.

The "1" group function is calculated using the following formula:

$$f_1 = x^{\mathrm{T}}\hat{S}^{-1}\hat{\mu}_1 - \frac{1}{2}\hat{\mu}_1^{\mathrm{T}}\hat{S}^{-1}\hat{\mu}_1 + \log(p_1)$$

To calculate the first part of the equation, $x^{\mathrm{T}}\hat{S}^{-1}\hat{\mu}_1$:

$$x^{\mathrm{T}}\hat{S}^{-1}\hat{\mu}_1 = [14.23 \quad 2.43 \quad 3.06 \quad 1065]$$

$$\begin{bmatrix} 2.62 & -0.56 & 0.23 & -0.0046 \\ -0.56 & 13.96 & -0.070 & -0.0017 \\ 0.23 & -0.070 & 1.33 & -0.0025 \\ -0.0046 & -0.0017 & -0.0025 & 0.000022 \end{bmatrix} \begin{bmatrix} 13.7 \\ 2.46 \\ 2.98 \\ 1116 \end{bmatrix}$$

$$x^{\mathrm{T}}\hat{S}^{-1}\hat{\mu}_1 = 449$$

To calculate the second part of the equation, $\frac{1}{2}\hat{\mu}_1^{\mathrm{T}}\hat{S}^{-1}\hat{\mu}_1$:

$$\frac{1}{2}\hat{\mu}_1^{\mathrm{T}}\hat{S}^{-1}\hat{\mu}_1 = \frac{1}{2}[13.7 \quad 2.46 \quad 2.98 \quad 1116]$$

$$\begin{bmatrix} 2.62 & -0.56 & 0.23 & -0.0046 \\ -0.56 & 13.96 & -0.070 & -0.0017 \\ 0.23 & -0.070 & 1.33 & -0.0025 \\ -0.0046 & -0.0017 & -0.0025 & 0.000022 \end{bmatrix} \begin{bmatrix} 13.7 \\ 2.46 \\ 2.98 \\ 1116 \end{bmatrix}$$

$$\frac{1}{2}\hat{\mu}_1^{\mathrm{T}}\hat{S}^{-1}\hat{\mu}_1 = 432$$

TABLE 4.15 Prediction of Alcohol Class Using the Three Membership Functions

Malic acid	Alkalinity of ash	Nonflavanoids	Proline	Alcohol	f_1	f_2	f_3	Prediction
14.23	2.43	3.06	1065	1	232	230	229	1
13.2	2.14	2.76	1050	1	193	192	190	1
13.16	2.67	3.24	1185	1	200	198	197	1
...	...	...	...	...	...	...	...	...
12.37	1.36	0.57	520	2	166	170	168	2
12.33	2.28	1.09	680	2	182	184	183	2
12.64	2.02	1.41	450	2	198	200	199	2
...	...	...	...	...	...	...	...	...
13.27	2.26	0.69	835	3	200	200	202	3
13.17	2.37	0.68	840	3	200	200	202	3
14.13	2.74	0.76	560	3	252	252	255	3
...	...	...	...	...	...	...	...	...

To calculate the final piece of the equation, $\log(p_1)$:

$$\log(p_1) = \log(0.33) = -1.10$$

The final score for f_1 is:

$$f_1 = 449 - 432 - 1.10 = 232$$

Similarly, scores for f_2 and f_3 are calculated, which are 230 and 229, respectively, and the class corresponding to the largest score is selected as the predictive value ("1").

Table 4.15 illustrates the calculation of the scores for a number of the observations. The highest scoring function is assigned as the prediction. The contingency table in Fig. 4.28 details the cross-validated (using a 5% cross-validation) predictive accuracy of this model.

Discriminant anaysis model summary	
Descriptors	Malic acid, Alkalinity of ash, Nonflavanoids, Proline
Response	Alcohol

Cross validated results	
Accuracy	0.966
Error	0.0337

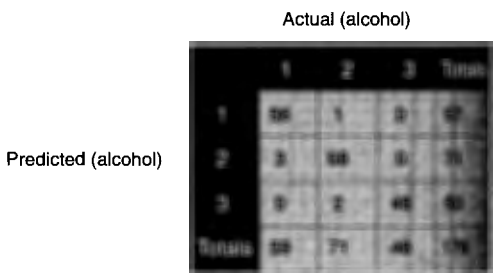

Figure 4.28 Summary of the cross-validated discriminant analysis model

In the same manner as described in Section 4.3.8, different independent variable combinations can be assessed and the most promising approach selected. In addition, interaction terms (such as the product of variables) or other transformed variables, such as quadratic terms, can be incorporated into the model to help in nonlinear situations or if other assumptions are violated. Alternatively, the use of the quadratic discriminant functions would be appropriate (Hastie, 2003).

4.5 LOGISTIC REGRESSION

4.5.1 Overview

Logistic regression is a popular method for building predictive models when the response variable is binary. Many data mining problems fall into this category. For example, a patient contracts or does not contract a disease or a cell phone subscriber does or does not switch services. Logistic regression models are built from one or more independent variables, that can be continuous, discrete, or a mixture of both. In addition to classifying observations into these categories, logistic regression will also calculate a probability that reflects the likelihood of a positive outcome. This is especially useful in prioritizing the results. As an example, a marketing company may use logistic regression to predict whether a customer will or will not buy a specific new product. While the model may predict more customers than the company has resources to pursue, the computed probability can be used to prioritize the most promising candidates.

Unlike discriminant analysis, logistic regression does not assume that the independent variables are normally distributed, or have similar variance in each group. There are, however, a number of limitations that apply to logistic regression including (1) a requirement for a large data set with sufficient examples of both categories, (2) that the independent variables are neither additive nor collinear, and (3) that outliers can be problematic.

4.5.2 Logistic Regression Formula

Logistic regression usually makes predictions for a response variable with two possible outcomes, such as whether a purchase does or does not take place. This response variable can be represented as 0 and 1, with 1 representing the class of interest. For example, 1 would represent the "buy" class and 0 would represent the "does not buy" class. A formula is generated to calculate a prediction from the independent variables. Instead of predicting the response variable, the formula estimates a probability that the response variable is 1, or $P(y = 1)$. A standard linear regression formula would compute values outside of the 0–1 range and is not used for this and other reasons. An alternative function is used in this situation. This function ensures the prediction is in the 0–1 range, by following a sigmoid curve. This curve for a single independent variable is shown in Fig. 4.29. A logistic response function has the following formula:

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$$

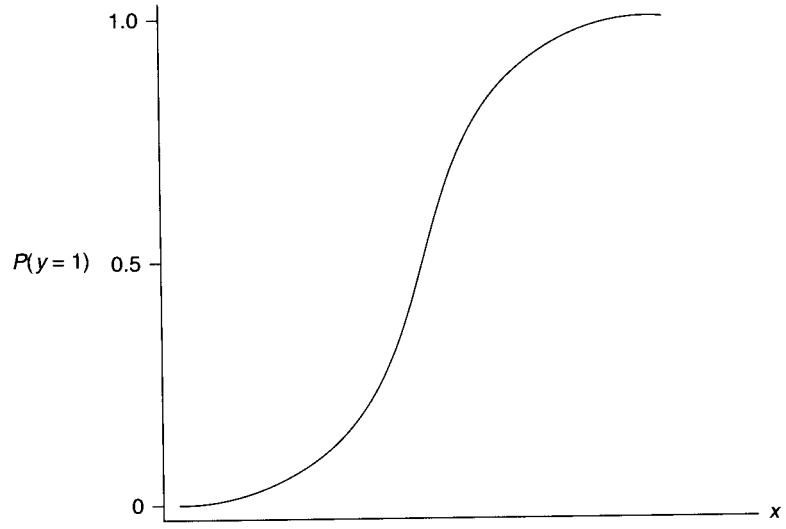

Figure 4.29 A sigmoid logistic response function

where β_0 is a constant, and $\beta_1 - \beta_k$ are coefficients to the k independent variables $(x_1 - x_k)$.

As an example, a logistic regression model is developed to predict whether a cereal would have a high nutritional rating using data from (http://lib.stat.cmu.edu/DASL/Datafiles/Cereals.html). It is based on the following logistic regression formula that uses measurements for variables *calories* (calories per serving), *protein* (grams of protein), and *carbo* (grams of complex carbohydrates):

$$P(y = 1) = \frac{1}{1 + e^{-(11.9 - 0.224 \times calories + 2.69 \times protein + 0.288 \times carbo)}}$$

For cereal A, which has $calories = 90$, $protein = 3$, $carbo = 19$, the predicted probability that this cereal would have a high nutritional rating is:

$$P(y = 1) = \frac{1}{1 + e^{-(11.9 - 0.224 \times 90 + 2.69 \times 3 + 0.288 \times 19)}}$$

$$P(y = 1) = 0.995$$

For cereal B with $calories = 110$, $protein = 2$, $carbo = 12$, the predicted probability is:

$$P(y = 1) = \frac{1}{1 + e^{-(11.9 - 0.224 \times 110 + 2.69 \times 2 + 0.288 \times 12)}}$$

$$P(y = 1) = 0.020$$

A cutoff is usually set such that probabilities above this value are assigned to class 1, and those below the cutoff are assigned to class 0. In this example, by setting a cutoff at 0.5, cereal A is assigned to the category high nutritional value, and cereal B is not.

"Odds" is a commonly used term, particularly in gambling. The odds are often referred to, for example, as "5 to 1" (such as to describe a bet), which translates into a 0.20 probability. The odds ratio considers $P(y = 1)$ vs $P(y = 0)$, using the following formula:

$$\text{odds} = \frac{P(y = 1)}{1 - P(y = 1)}$$

This allows us to rewrite the logistic regression formula as:

$$\text{odds} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i}$$

This equation helps to interpret the coefficients of the equation. For an individual independent variable (x) with corresponding beta coefficient (β), holding all other variables constant, and increasing the value by 1 would result in the odds being increased by e^{β}. For example, in the cereal example, increasing the value of *carbo* by 1 would result in an increase in odds of being a high nutritional value cereal of by $e^{0.288}$ which is 1.33 or 33%.

It is also helpful to consider taking the natural log, which results in the following formula or *logit* function that will return a value between $-\infty$ and $+\infty$:

$$\log(\text{odds}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i$$

4.5.3 Estimating Coefficients

The logistic regression coefficients are computed using a maximum likelihood procedure (Agresti, 2002), where the coefficients are continually refined until an optimal solution is found. The Newton–Raphson method is often used. Since the method is repeated multiple times, the estimated values for the coefficients $\hat{\beta}^{\text{new}}$ are updated using the previous estimates $\hat{\beta}^{\text{old}}$, based on the following formula:

$$\hat{\beta}^{\text{new}} = \hat{\beta}^{\text{old}} + (X^{\text{T}} W X)^{-1} X^{\text{T}} (y - p)$$

where X is the matrix describing the independent variables (with the first column assigned as 1 for calculation of the intercept), p is a vector of fitted probabilities, W is a weight matrix where the diagonal values represent $p(1 - p)$, and y is the response variable.

The method starts by assigning arbitrary values to the β coefficients. The β coefficients are repeatedly calculated using the formula above. Each iteration results in an improved coefficient estimate, and the process finishes when the beta coefficients are not changing significantly between iterations.

In the following example, a data set relating to diabetes is used to illustrate the process of calculating the coefficients (http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes). Table 4.16 illustrates some of the data used to build the logistic regression model and Fig. 4.30 summarizes the data set. The *diabetes* variable is used as the response and the other five variables used as independent variables.

The following illustrates the process of calculating the β coefficients. In the first step, the β coefficients are initialized to an arbitrary value, in this case zero:

$$\hat{\beta}^{\text{old}} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

The X matrix is constructed from the independent variables, again, with the first column containing all 1s for calculation of the intercept value. The first three rows are shown:

$$X = \begin{bmatrix} 1 & 97 & 64 & 18.2 & 0.299 & 21 \\ 1 & 83 & 68 & 18.2 & 0.624 & 27 \\ 1 & 97 & 70 & 18.2 & 0.147 & 21 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

TABLE 4.16 Data Table Concerning Diabetes Information

Plasma glucose	Diastolic blood pressure	Body mass index	Diabetes pedigree function (DPF)	Age	Diabetes
97	64.0	18.2	0.299	21	0
83	68	18.2	0.624	27	0
97	70	18.2	0.147	21	0
104	76	18.4	0.582	27	0
80	55	19.1	0.258	21	0
99	80	19.3	0.284	30	0
103	80	19.4	0.491	22	0
92	62	19.5	0.482	25	0
100	74	19.5	0.149	28	0
95	66	19.6	0.334	25	0
129	90	19.6	0.582	60	0
162	76	49.6	0.364	26	1
122	90	49.7	0.325	31	1
152	88	50.0	0.337	36	1
165	90	52.3	0.427	23	0
115	98	52.9	0.209	28	1
162	76	53.2	0.759	25	1
88	30	55.0	0.496	26	1
123	100	57.3	0.88	22	0
180	78	59.4	2.42	25	1
129	110	67.1	0.319	26	1

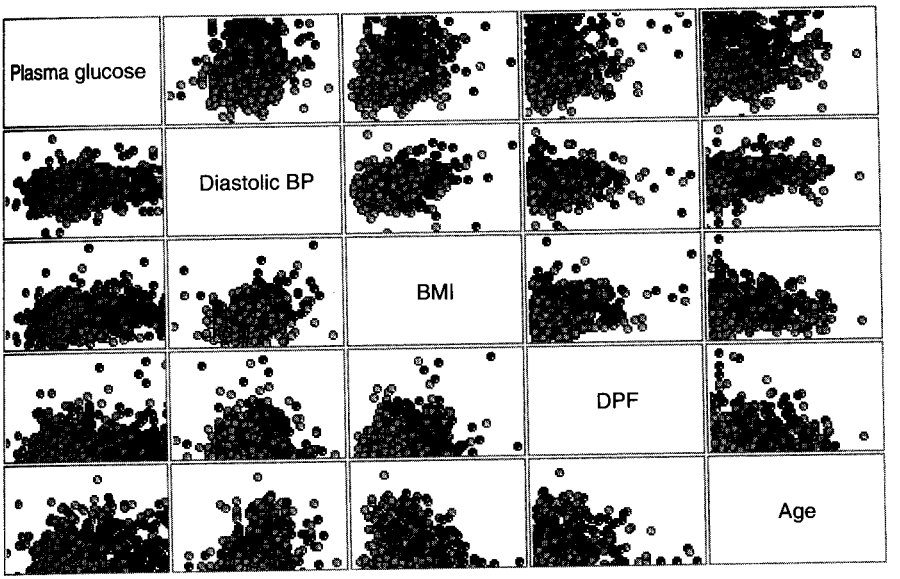

Figure 4.30 Scatterplot matrix with diabetes observations highlighted

The y matrix represents the response data, with the first three rows shown:

$$y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \cdots \end{bmatrix}$$

A p matrix corresponding to the calculated probability for each observation, using the current β coefficients, is calculated. The first three entries are shown here:

$$p = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \\ \cdots \end{bmatrix}$$

A weight matrix, W, where the number of columns and rows both equal the number of observations, is calculated. The diagonal represents $p(1 - p)$, and the first three rows and columns are shown here:

$$W = \begin{bmatrix} 0.25 & 0 & 0 & \cdots \\ 0 & 0.25 & 0 & \cdots \\ 0 & 0 & 0.25 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

These matrices are used to generate an updated value for the beta coefficients:

$$\hat{\beta}^{\text{new}} = \hat{\beta}^{\text{old}} + (X^{\text{T}}WX)^{-1}X^{\text{T}}(y - p)$$

TABLE 4.17 Optimization of the β Coefficients for Logistic Regression

	β_0	β_1	β_2	β_3	β_4	β_5
Step 1	0.0	0.0	0.0	0.0	0.0	0.0
Step 2	−6.162	0.0247	−0.00404	0.0552	0.550	0.0231
Step 3	−8.433	0.0326	−0.00654	0.0814	0.836	0.0320
Step 4	−8.982	0.0344	−0.00733	0.0882	0.918	0.0343
Step 5	−9.009	0.0345	−0.00738	0.0886	0.923	0.0344
Step 6	−9.009	0.0345	−0.00738	0.0886	0.923	0.0344

$$\hat{\beta}^{\text{new}} = \begin{bmatrix} -6.162 \\ 0.0247 \\ -0.00404 \\ 0.0552 \\ 0.0552 \\ 0.0231 \end{bmatrix}$$

In Table 4.17, the first two rows illustrate the β coefficients calculated up to this point. This process is repeated until the coefficients values converge, that is, when the values do not change significantly between steps.

4.5.4 Assessing and Optimizing Results

Once a logistic regression formula has been generated, it can be used for prediction. In Table 4.18, two new columns are added. The logistic regression formula will calculate a probability, and from this value a classification can be assigned. A cutoff value, such as 0.5, can be used with a probability higher than 0.5 assigned to the 1 class and a probability less than or equal to 0.5 assigned to the 0 category. These classifications can be used to assess the model, using a contingency table, such as the one in Fig. 4.31.

The contingency table allows for the calculation of overall accuracy, error rate, specificity, sensitivity, and so on. In addition, since the model also generates a probability, a lift chart and an ROC chart can also be generated. Any models generated can be optimized by varying the independent variables to generate the simplest, most predictive model, as discussed in Section 4.3.8. The cutoff value can also be adjusted to enhance the quality of the model. References to additional methods for assessing the logistic regression model, such as the Wald test, the likelihood ratio test, and Hosmer and Lemeshow χ^2 test of goodness of fit are provided. Like linear regression and discriminant analysis, interaction terms (such as the product of two variables) or higher-order terms (such as a variable squared) can be computed and used with the model to enhance prediction. Chapter 5 illustrates the application of logistic regression to a number of case studies.

TABLE 4.18 Prediction of Diabetes as Well as the Probability of Diabetes is 1

Plasma glucose	Diastolic blood pressure	Body mass index	Diabetes pedigree function (DPF)	Age	Diabetes	Predicted	Probability
97	64.0	18.2	0.299	21	0	0	0.0287
83	68	18.2	0.624	27	0	0	0.0285
97	70	18.2	0.147	21	0	0	0.0240
104	76	18.4	0.582	27	0	0	0.0530
80	55	19.1	0.258	21	0	0	0.0180
99	80	19.3	0.284	30	0	0	0.0425
103	80	19.4	0.491	22	0	0	0.0371
92	62	19.5	0.482	25	0	0	0.0365
100	74	19.5	0.149	28	0	0	0.293
95	66	19.6	0.334	25	0	0	0.0263
129	90	19.6	0.582	60	0	0	0.0416
162	76	49.6	0.364	26	1	1	0.839
122	90	49.7	0.325	31	1	1	0.577
152	88	50.0	0.337	36	1	1	0.828
165	90	52.3	0.427	23	0	1	0.863
115	98	52.9	0.209	28	1	1	0.520
162	76	53.2	0.759	25	1	1	0.909
88	30	55.0	0.496	26	1	1	0.508
123	100	57.3	0.88	22	0	1	0.759
180	78	59.4	2.42	25	1	1	0.993
129	110	67.1	0.319	26	1	1	0.854

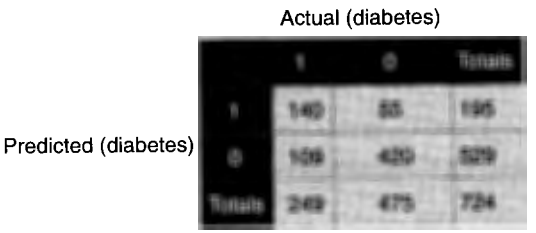

Figure 4.31 Contingency table summarizing the number of correct and incorrect predictions

4.6 NAIVE BAYES CLASSIFIERS

4.6.1 Overview

The *naive Bayes* (also referred to as idiot's Bayes or simple bayesian classifier) is a classification modeling method. It makes use of the *Bayes theorem* to compute probabilities of class membership, given specific *evidence*. In this scenario, the

evidence refers to particular observations in the training set that either support or do not support a particular prediction.

Naive Bayes models have the following restrictions:

- *Only categorical variables*: This method is usually applied in situations in which the independent variables and the response variable are categorical.
- *Large data sets*: This method is versatile, but it is particularly effective in building models from large data sets.

This method provides a simple and efficient approach to building classification prediction models. It also can compute probabilities associated with class membership, which can be used to rank the results.

4.6.2 Bayes Theorem and the Independence Assumption

At the heart of this approach is the Bayes theorem:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

This theorem calculates the probability of a *hypothesis* (H) given some evidence (E), or *posterior probability*. For example, it can calculate the probability that someone would develop diabetes given evidence of a family history of diabetes. The hypothesis corresponds to the response variable in the other methods. The theorem makes use of this posterior probability of the evidence given the hypothesis, or $P(E|H)$. Using the same example, the probability of someone having a family history of diabetes can also be calculated given the evidence that the person has diabetes and would be an example of $P(E|H)$. The formula also makes use of two prior probabilities, the probability of the hypothesis $P(H)$, and the probability of the evidence $P(E)$. These probabilities are not predicated on the presence of any evidence. In this example, the probability of having diabetes would be $P(H)$, and the probability of having a family history of diabetes would be $P(E)$.

4.6.3 Independence Assumption

In situations in which there is only a single independent variable, the formula is straightforward to apply; however, models with a single independent variable would be limited in their usefulness. Unfortunately, the strict use of the Bayes theorem for multiple independent variables each having multiple possible values becomes challenging in practical situations. Using this formula directly would result in a large number of computations. Also, the training data would have to cover all of these situations, which also makes its application impractical. The naive Bayes approach uses a simplification which results in a computationally feasible series of calculations. The method assumes that the independent variables are independent despite the fact that this is rarely the case. Even with this overly optimistic assumption, the method is useful as a classification modeling method in many situations.

TABLE 4.19 Diabetes Data Set to Illustrate the Naive Bayes Classification

Blood pressure	Weight	Family history	Age	Diabetes
Average	Above average	Yes	50+	1
Low	Average	Yes	0–50	0
High	Above average	No	50+	1
Average	Above average	Yes	50+	1
High	Above average	Yes	50+	0
Average	Above average	Yes	0–50	1
Low	Below average	Yes	0–50	0
High	Above average	No	0–50	0
Low	Below average	No	0–50	0
Average	Above average	Yes	0–50	0
High	Average	No	50+	0
Average	Average	Yes	50+	1
High	Above average	No	50+	1
Average	Average	No	0–50	0
Low	Average	No	50+	0
Average	Above average	Yes	0–50	1
High	Average	Yes	50+	1
Average	Above average	No	0–50	0
High	Above average	No	50+	1
High	Average	No	0–50	0

4.6.4 Classification Process

To illustrate the naive Bayes classification process, the training set in Table 4.19 will be used to classify the following observation (X):

$$X : BP = high;\ weight = above;\ FH = yes;\ age = 50+$$

In this example, the observation (X) is an individual whose blood pressure is high ($BP = high$), whose weight is above normal ($weight = above$), who has a family history of diabetes ($FH = yes$), and whose age is above 50 ($Age = 50+$). Using the training data in Table 4.19, the objective is to classify this individual as prone to developing or not prone to developing diabetes given the factors described. In this example, calculating $P(diabetes = 1|X)$ and the $P(diabetes = 0|X)$ is the next step. The individual will be assigned to the class, either has ($diabetes = 1$) or has not ($diabetes = 0$), based on the highest probability value.

$$P(diabetes = 1|X) = \frac{P(X|diabetes = 1)P(diabetes = 1)}{P(X)}$$

$$P(diabetes = 0|X) = \frac{P(X|diabetes = 0)P(diabetes = 0)}{P(X)}$$

Since $P(X)$ is the same in both equations, only $P(X|diabetes = 1)P(diabetes = 1)$ and $P(X|diabetes = 0)P(diabetes = 0)$ are needed.

To calculate $P(diabetes = 1)$, the number of observations is counted in Table 4.19 with $diabetes = 1$, which is 9, divided by the total number of observations, which is 20:

$$P(diabetes = 1) = 9/20 = 0.45$$

Similarly, to calculate $P(diabetes = 0)$, the number of observations in Table 4.19 is counted where $diabetes = 0$, which is 11, divided by the total number of observations, which is 20:

$$P(diabetes = 0) = 11/20 = 0.55$$

Since this approach assumes that the independent variables are independent, the calculation of the $P(X|diabetes = 1)$ is the product of the conditional probability for each of the values of X:

$$\begin{aligned} P(X|diabetes = 1) = {} & P(BP = high|diabetes = 1) \\ & \times P(weight = above|diabetes = 1) \\ & \times P(FH = yes|diabetes = 1) \\ & \times P(age = 50+|diabetes = 1) \end{aligned}$$

The individual probabilities are again derived from counts of Table 4.19. For example, $P(BP = high|diabetes = 1)$ counts all observations with $BP = high$ and $diabetes = 1(4)$, divided by the number of observations where $diabetes = 1(9)$:

$$\begin{aligned} P(BP = high|diabetes = 1) &= 4/9 = 0.44 \\ P(weight = above|diabetes = 1) &= 7/9 = 0.78 \\ P(FH = yes|diabetes = 1) &= 6/9 = 0.67 \\ P(age = 50+|diabetes = 1) &= 7/9 = 0.78 \end{aligned}$$

Using these probabilities, the probability of X given $diabetes = 1$ is calculated:

$$\begin{aligned} P(X|diabetes = 1) = {} & P(BP = high|diabetes = 1) \\ & \times P(weight = above|diabetes = 1) \\ & \times P(FH = yes|diabetes = 1) \\ & \times P(age = 50+|diabetes = 1) \\ P(X|diabetes = 1) = {} & 0.44 \times 0.78 \times 0.67 \times 0.78 \\ P(X|diabetes = 1) = {} & 0.179 \end{aligned}$$

Using the values for $P(X|diabetes = 1)$ and $P(diabetes = 1)$, the product $P(X|diabetes = 1)P(diabetes = 1)$ can be calculated:

$$\begin{aligned} P(X|diabetes = 1)P(diabetes = 1) &= 0.179 \times 0.45 \\ P(X|diabetes = 1)P(diabetes = 1) &= 0.081 \end{aligned}$$

Similarly, the value for $P(X|diabetes = 0)P(diabetes = 0)$ can be calculated:

$$\begin{aligned} P(X|diabetes = 0) = & P(BP = high|diabetes = 0) \\ & \times P(weight = above|diabetes = 0) \\ & \times P(FH = yes|diabetes = 0) \\ & \times P(age = 50+|diabetes = 0) \end{aligned}$$

Using the following probabilities, based on counts from Table 4.19:

$$\begin{aligned} P(BP = high|diabetes = 0) &= 4/11 = 0.36 \\ P(weight = above|diabetes = 0) &= 4/11 = 0.36 \\ P(FH = yes|diabetes = 0) &= 4/11 = 0.36 \\ P(age = 50+|diabetes = 0) &= 3/11 = 0.27 \end{aligned}$$

The $P(X|diabetes = 0)$ can now be calculated:

$$\begin{aligned} P(X|diabetes = 0) &= 0.36 \times 0.36 \times 0.36 \times 0.27 \\ P(X|diabetes = 0) &= 0.0126 \end{aligned}$$

The final assessment of $P(X|diabetes = 0)P(diabetes = 0)$ is computed:

$$P(X|diabetes = 0)P(diabetes = 0) = 0.0126 \times 0.55 = 0.0069$$

Since $P(X|diabetes = 1)P(diabetes = 1)$ is greater than $P(X|diabetes = 0)$ P(diabetes $= 0$), the observations X are assigned to class $diabetes = 1$. A final probability that $diabetes = 1$, given the evidence (X), can be computed as follows:

$$P(diabetes = 1|X) = 0.081/(0.081 + 0.0069) = 0.922$$

The naive Bayes is a simple classification approach that works surprisingly well, particularly with large data sets as well as with larger numbers of independent variables. The calculation of a probability is helpful in prioritizing the results. As with other methods described in the chapter, the predictive accuracy of any naive Bayes model can be assessed using the methods outlined in Section 4.1. Building models with different sets of independent variables can also help.

4.7 SUMMARY

The preceding chapter has discussed two basic types of models:

- *Classification*: model where the response is a categorical variable;
- *Regression*: model where the response is a continuous variable.

TABLE 4.20 Summary of Methods to Assess Regression Models

Mean square error: $\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}$	Mean absolute error: $\frac{\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert}{n}$
Relative square error: $\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	Relative absolute error: $\frac{\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert}{\sum_{i=1}^{n}\lvert y_i - \bar{y}\rvert}$
Correlation coefficient: $\frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{(n-1)s_y s_{\hat{y}}}$	

TABLE 4.21 Summary of Methods to Assess Binary Classification Models

Accuracy:	Error rate:	Sensitivity:	Specificity:
$\frac{TP + TN}{TP + FP + FN + TN}$	$1 - \frac{TP + TN}{TP + FP + FN + TN}$	$\frac{TP}{TP + FN}$	$\frac{TN}{TN + FP}$
False positive rate:	Positive predictive value:	Negative predictive value:	False discovery rate:
$\frac{FP}{FP + TN}$	$\frac{TP}{TP + FP}$	$\frac{TN}{TN + FN}$	$\frac{FP}{FP + TP}$

TABLE 4.22 Summary of Predictive Modeling Methods Discussed in this Chapter

Method	Model type	Independent variables	Comments
Multiple linear regression	Regression	Any numeric	Assumes a linear relationship Easy to explain Quick to build
Discriminant analysis	Classification	Any numeric	Assumes the existence of mutually exclusive groups with common variances
Logistic regression	Classification	Any numeric	Will calculate a probability Easy to explain Limit on number of observations to build models from
Naive Bayes	Classification	Only categorical	Requires a lot of data

Multiple methods have been discussed for assessing regression and classification models, and are summarized in Tables 4.20 and 4.21. Principal component analysis has been discussed as a method for understanding and restricting the number of variables in a data set. Table 4.22 summarizes the methods discussed in this chapter.

4.8 FURTHER READING

For further general discussion on the different methods for assessing models see Han and Kamber (2005) and Witten and Frank (2005). Here additional validation methods, such as bootstrapping, are discussed, along with methods for combining models using techniques such as bagging and boosting. Joliffe (2002), Strang (2006), Jackson (1991), Jobson (1992), and Johnson and Wishern (1998) provide additional detail concerning principal component analysis. Additional information on multiple linear regression can be found in Allison (1998), Draper and Smith (1998), Fox (1997), and Rencher (2002). Discriminant analysis is also covered in more depth in Hastie et al. (2003), McLachlan (2004), Huberty (1994), Lachenbruch (1975), and Rencher (2002). Agresti (2002), Balakrishnan (1992), and Hosmer and Lemeshow (2000) cover logistic regression, and Han and Kamber (2005) as well as Hand and Yu (2001) discuss the use of the naive Bayes approach. Other commonly used methods for building prediction models include neural networks (Hassoun, 1995; Haykin, 1998; and Myatt, 2007), classification and regression trees, rule-based classifiers, support vector machines, and *k*-nearest neighbors, and these are covered in a variety of books, including Han and Kamber (2005), Witten and Frank (2005), Hastie et al. (2003), and Shumueli et al. (2007).