Chapter 6 Potential Applications for a Service Provider

6.1 Context

This chapter focuses on the potential data mining applications for a service provider – in particular, a travel agency. In the last few years, tourism has grown tremendously in the Asia-Pacific region. It is now one of the most important sectors in many Asia-Pacific countries, including Singapore. The substantial growth of tourism is the result of an increase in economic growth, disposable income and leisure time, political stability and aggressive tourism campaigns. Even as early as the late 1990s, the growth of tourism has been predicted to maintain a high rate well into the twenty-first century (Singh, 1997). More recently, China and India have been predicted to be the new powerhouses of the global tourism industry over the next four decades (Yahya, 2003).

Despite impressive growth, the tourism and travel industry has been facing increasing competition. The current environment has been described as the most competitive that the Asia-Pacific tourism industry has ever faced (Hamdi, 2003). Regional destinations have scrambled to make up for the massive losses in tourists brought about by SARS, the Iraqi war and terrorism concerns. Many governments in the region have put aside budgets amounting to many million dollars to help recovery. "Coopetition" – a term coined to suggest a balance between cooperation and competition – is now strongly advocated among different countries and different parties including governments, tourism bodies, airlines, travel agencies and travel professionals (Anonymous, 2003).

On the home front, competition has become more intense too as more travel agencies have emerged to take advantage of the substantial tourism growth. Consequently, to share a bigger slice of the pie, travel agencies have been trying to differentiate themselves in order to capture a larger pool of travellers and attract these travellers to stay as loyal customers with the travel agencies.

At the same time, consumers of tourism products (such as tour packages) are becoming more demanding. Generally, customers of travel agencies are price and quality conscious. To keep their customers (i.e., to reduce churn or customer turnover), travel agencies have to meet the expectations of their customers. In this environment, customer relationship management has become increasingly important. Travel agencies have to know their customers – in particular, their needs and wants. They also have to provide a level of service acceptable to their customers and be innovative and pro-active in their approach to getting and retaining customers.

6.2 Background Information for Data Mining Illustrations

To illustrate potential data mining applications for a service provider, consider Best Travel Agency (abbreviated as Best from this point to facilitate discussion) – a fictitious travel agency that has been operating in Singapore for the past three years. It has five branches in Singapore and 20 employees in total. In view of the increasing competition, Best intends to use data mining to better understand its customers so as to meet their needs and expectations. The agency also hopes to use data mining to improve its operations in terms of the services and tour packages that it provides. Ultimately, based on these, Best wishes to increase its profitability.

To support its data mining applications, Best has built several databases. The first database (Best_des&vis.sav) contains 2000 observations, comprising 400 observations for each of its five branches.

These observations are the result of a survey that the agency has conducted on customers who have travelled to Australia, China, Hong Kong and Thailand, which are the most popular destinations for its tour packages in the last few months. This survey is aimed at obtaining feedback on the agency's products and service standards.

The data collected and hence available for data mining are summarised in Table 6.1 as 17 variables. These variables comprise identification variables, demographic characteristics, tour features and customer satisfaction measures. More information on this database is given in the next section. Other databases will be discussed in subsequent sections.

Table 6.1 List of Variables (Best_des&vis.sav)

Variable	Definition	Label
ID	Identification code of customer	1 to 2000
BrID	Identification code of branch	1 to 5
Gender	Gender of customer	0 = Male 1 = Female
Race	Race of customer	1 = Indian 2 = Malay 3 = Chinese
Age	Age of customer	18 to 69
Educ	Education level of customer	1 = No formal education 2 = PSLE 3 = GCE O Level 4 = GCE A Level 5 = Diploma/degree

Resid	Residence type of customer	1 = 1 to 2 room HDB flat 2 = 3 to 4 room HDB flat 3 = 5 room HDB flat 4 = Condominium 5 = Landed property 6 = Others (HUDC … etc.)
Occup	Occupational status of customer	1 = Working 2 = Unemployed 3 = Retired 4 = Housewife 5 = Student 6 = National service
Country	Country of tour package	1 = Australia 2 = China 3 = Hong Kong 4 = Thailand
Days	Number of days (D) and nights (N) of the tour package	1 = 3D2N 2 = 4D3N 3 = 5D4N
Ptype	Package tour type	0 = Free and easy 1 = Guided tour
Gpsize	Size of tour group	1 = 9 or fewer persons 2 = 10 to 14 persons 3 = 15 to 19 persons 4 = 20 to 24 persons 5 = 25 or more persons
Child	Whether tour group has child/ children	0 = Without child 1 = With child/children

Value	Whether tour package is worth the price paid for	0 = No 1 = Yes
Tour	Whether tour itinerary is enjoyable	0 = No 1 = Yes
Service	Whether tour guide service is satisfactory	0 = No 1 = Yes
Rating	Overall rating of travel agency	1 = Very bad 2 = Bad 3 = Neutral 4 = Good 5 = Very good

6.3 Application 1: Identification of Dissatisfied Customers

In view of the increasing number of customer complaints received in the past few months, Best has decided to identify the branches that are under-performing in order to improve their products and services to meet customers' expectations. To do this, Best has conducted the customer satisfaction survey referred to earlier. The survey asks respondents to give an overall rating of Best on a scale of 1 to 5, where a higher rating represents a more favourable response (see Table 6.1).

In addition, the survey questionnaire includes the following dimensions of customer satisfaction:

1) Is the tour package worth the price paid for?
2) Is the tour itinerary enjoyable?
3) Is the service provided by the tour guide(s) satisfactory?

These questions represent the main areas that Best has some control over and therefore comprise the areas of interest in the data analysis.

At this preliminary stage, Best is interested to explore the survey data using descriptive and visualisation tools. The Histogram node in SPSS Clementine is used to view the overall ratings of the five branches of Best. As shown in Figure 6.1, Branch 4 has the highest number of customers rating the branch "1" and "2" (i.e., "very bad" and "bad"). As resources are limited, Best has decided to focus only on Branch 4 at this time to generically profile the dissatisfied customers. To do this, data from Branch 4 are selected for further analysis.

The SPSS Clementine Distribution node is used to plot the distributions for the variables Value (i.e., whether the tour package is worth the price paid for), Tour (i.e., whether the tour itinerary is enjoyable) and Service (i.e., whether the tour guide service is satisfactory). As shown in Figure 6.1, 62.75%, 77.50% and 70.00% of the customers in Branch 4 are not satisfied with the value that they get from the tour package, the tour itinerary, and the service provided by the tour guides, respectively. To better understand the demographic and tour characteristics of customers who are dissatisfied with Value, Tour and Service, directed web graphs are plotted. The results are presented in Figure 6.2.

As shown in the upper panel of Figure 6.2, customers who perceive that the tour package that they have taken up is not worth the price that they have paid for are mainly customers who have gone on guided tours and to Australia. The tour groups also tend to include children. In addition, these customers are also likely to be working (i.e., currently employed), male customers, aged between 30 to 45 years old and those who live in 3-/4-room HDB flats. With this group of dissatisfied customers identified, Best can attempt to take pre-emptive or remedial actions including contacting them to find out more about their less than satisfactory experience, reviewing the current tour package with a view to improve its value, training counter service staff to spend more time explaining the itinerary to this target group to help them understand what they are paying for and so on.

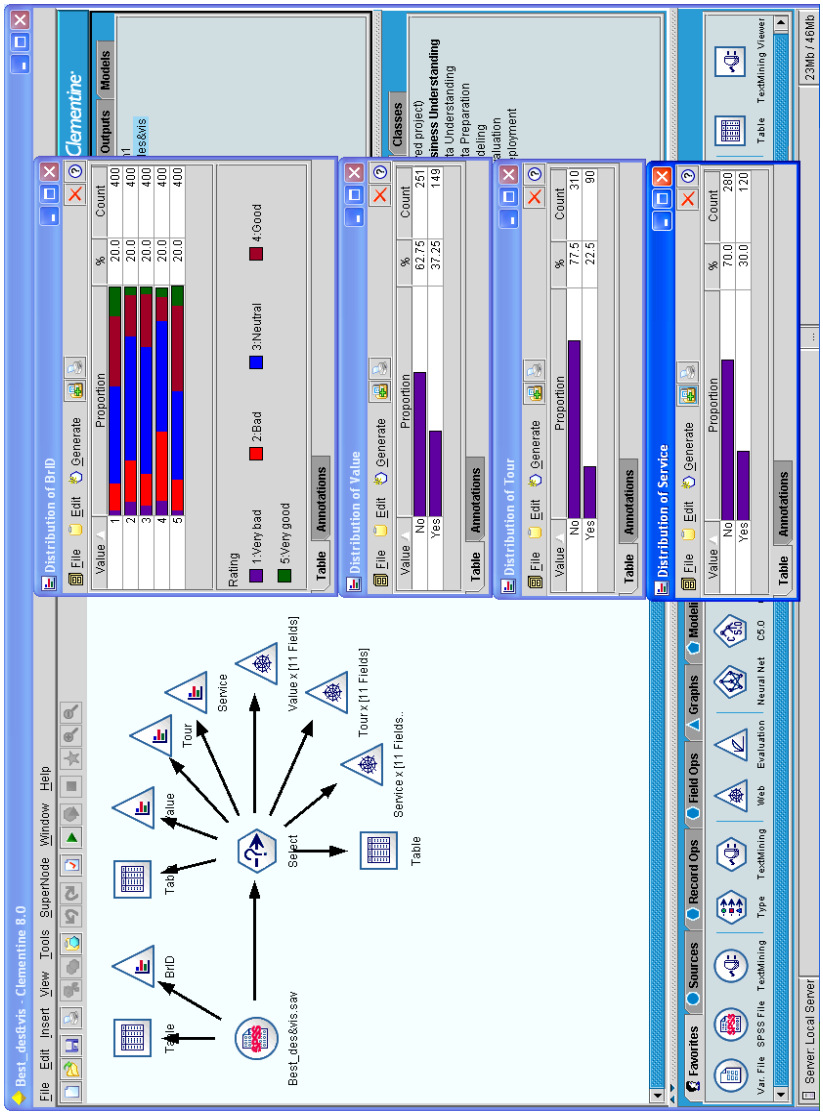


Figure 6.1 Exploration of Dissatisfied Customers

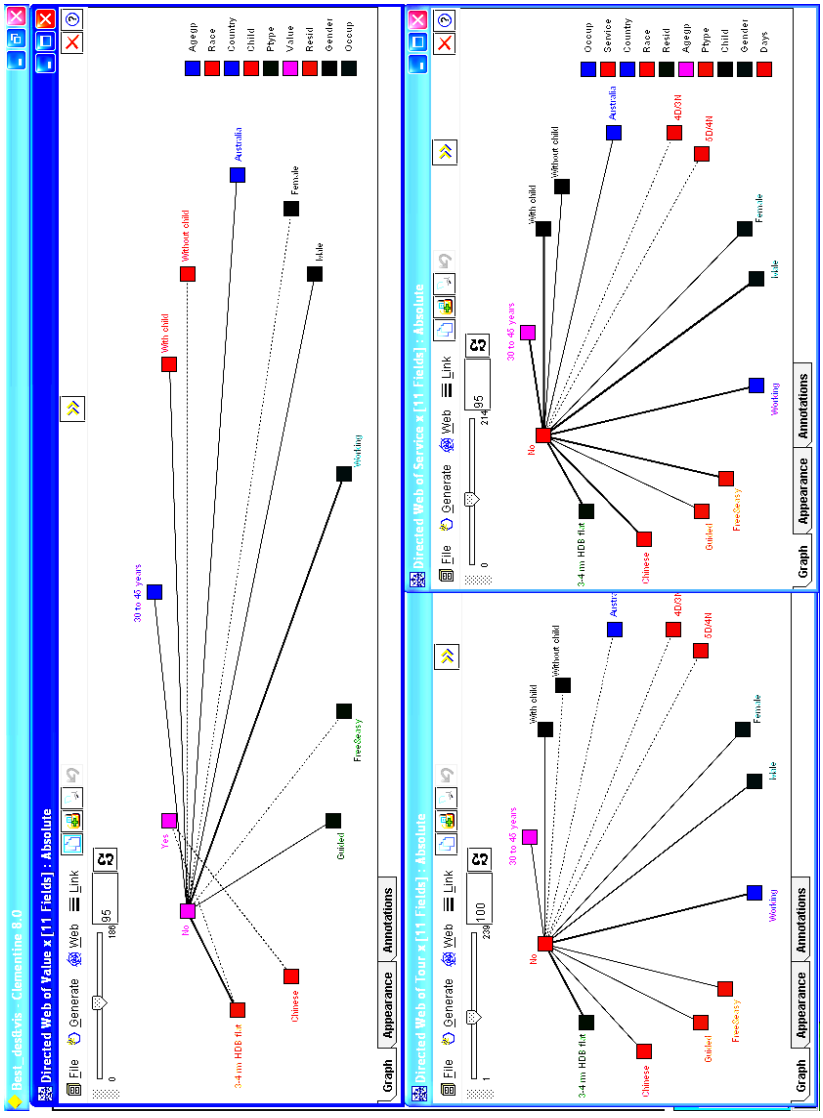


Figure 6.2 Directed Web Graphs of Value, Tour and Service

As for customers who perceive that the tour itinerary is not enjoyable (see the lower left panel of Figure 6.2), these dissatisfied customers are mainly Chinese, working, aged between 30 to 45 years old and those who live in 3-/4-room HDB flats. They are also likely to be in tour groups with children. As in the earlier discussion, Best can attempt to take pre-emptive or remedial actions by contacting the dissatisfied customers to find out more about their less than enjoyable tour and reviewing the current tour itinerary with a view to include what the dissatisfied customers would like to have or exclude what they would not like to have.

Finally, as shown in the lower right panel of Figure 6.2, strong links can be found between customers who perceive that the tour guide service is not satisfactory and those who are working (i.e., currently employed), male, Chinese, aged between 30 to 45 years old and those who live in 3-/4-room HDB flats. The dissatisfied customers are also likely to be on free and easy packages and in tour groups with children. With this group of dissatisfied customers identified, Best can try to reverse the situation by contacting them to find out more about their less than satisfactory experience, assessing the current tour guide service with a view to improve it, and training or hiring tour guides who are able to meet customer needs and expectations.

The results indicate that customers who are dissatisfied with Value, Tour and/or Service overlap. By identifying dissatisfied customers and the areas for improvement, Best can work towards enhancing the quality of its tours and services, the satisfaction and loyalty of its customers and its ability to attract more customers. Over time, these will give Best a significant competitive advantage.

6.4 Application 2: Developing Tour Packages

Best has observed an increasing number of tourists going to Beijing (China) from Singapore and wants to capture a greater share of this tourist segment.

Currently, the itineraries for Beijing tours cover a long list of the most common and popular tourist attractions. Recent feedback from Best's customers has indicated that not all these attractions appeal to all customers. Subgroups of customers seem to be interested in only particular subsets of attractions.

In addition, many of Best's customers who are keen to visit other Chinese cities want to have a brief 1- or 2-day stopover in Beijing. To improve its China tour packages, Best wants to find out which tourist attractions in Beijing can be grouped into subsets that would appeal to different subgroups of customers. Best is considering offering a basic China tour package (outside of Beijing) that incorporates different options to additional attractions in Beijing for different subgroups of customers. For this brief stopover in Beijing, only two or three tourist attractions are feasible.

After conducting several focus group sessions, the following tourist attractions appear to be the ten most popular places of interest for the Beijing options:

1) Forbidden City
2) Great Wall of China
3) Temple of Heaven
4) Summer Palace
5) Ming Tombs
6) Hutong and Courtyard
7) Cultural Village
8) Beihai Park
9) Tiananmen Square
10) Beijing Art Museum

To develop the Beijing options further, Best has engaged a marketing research firm in China to conduct a survey of 2000 tourists who have visited Beijing on tour packages of at least five days. As the survey is conducted at the Beijing International Airport, it is kept very simple by merely

asking the respondents if they have very much enjoyed visiting the ten tourist attractions listed above to the extent of wanting to visit them again and/or recommend them to their relatives and friends. A favourable response for an attraction is taken as an indication of the appropriateness of the attraction for inclusion in the Beijing options. The data collected are then entered into an SPSS file (filename = Best_assoc.sav) for analysis.

Preliminary distribution results (see the topmost panel in Figure 6.3) show that the five most popular attractions (in descending order of popularity) are: (1) Forbidden City, (2) Summer Palace, (3) Cultural Village, (4) Great Wall of China, and (5) Ming Tombs. For this data mining application, association analysis is the primary tool used. This is performed using the SPSS Clementine Apriori and GRI (Generalised Rule Induction) algorithms. The association results are summarised in Figure 6.3.

The Apriori and GRI association analysis results are identical as shown in the middle and lower panels in Figure 6.3. The association rules suggest the following two Beijing options: (1) Summer Place, Cultural Village and Ming Tombs, and (2) Great Wall of China and Forbidden City. These two options are expected to add the greatest value as extensions to the regular China tour packages outside of Beijing.

6.5 Application 3: Profiling of Customers

In this data mining application, Best is interested to profile its most profitable customers with a view to reduce the churn rate of this segment of customers by offering more attractive tour packages and better service to it. This is deemed as critical to the success of Best because of the increasingly competitive environment in the travel industry. For this application, Best has a ready database of 2500 customers. The variables in this database (filename = Best_clus.sav) are listed in Table 6.2.

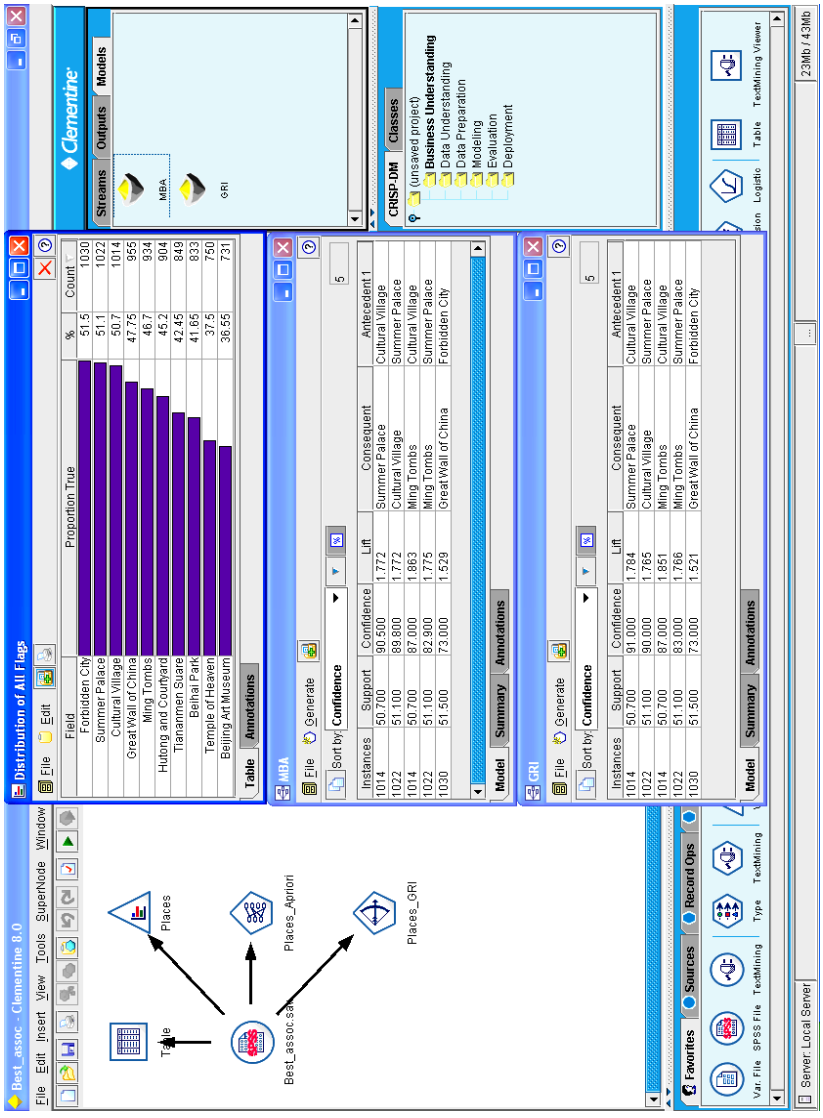


Figure 6.3 Groupings of Attractions in Beijing

Table 6.2 List of Variables (Best_clus.sav)

Variable	Definition	Label
ID	Identification code of customer	1 to 2500
Gender Race Age Educ Resid Occup	As in Table 6.1	As in Table 6.1
Amount	Average amount earned by Best from profit and commission.	S$500 - S$1585 (rounded to the nearest dollar)
Peak	Number of times the customer travelled with Best during peak periods during the last three years	0 to 12
OffPeak	Number of times the customer travelled with Best during off-peak periods during the past three years	0 to 12
Total	Total number of times the customer travelled with Best during the past three years	1 to 12

The first seven variables (i.e., ID, Gender, Race, Age, Educ, Resid and Occup) are identical to the ones listed earlier in Table 6.1. These

variables (except for ID, an identification variable) measure the demographic characteristics of the customers in the database.

The variable "Amount" measures the average revenue earned by Best for each customer in terms of profit (from the sale of tour packages) and commission (from commissions based on purchases made by customers at shops recommended by Best). For each customer, the average revenue is derived from the total amount that Best has earned during a 3-year period divided by the total number of times the customer has travelled with Best during this period. This variable is used as a measure of customer profitability (or value). The remaining three variables (namely, Peak, OffPeak and Total) measure the number of times during the past three years a customer has travelled with Best during peak and off-peak periods and in total, respectively.

To profile the customers, the database is clustered on the basis of variables in Table 6.2 (except ID). The SPSS Clementine TwoStep node is used for this purpose. This clustering algorithm has the advantage of determining statistically the optimal number of clusters among the customers. The clustering results (a 7-clustering solution) are shown in Figure 6.4.

To further analyse the relative profitability of the seven clusters, analysis of variance and multiple comparisons (using the Tukey procedure at an alpha level of 0.05) are performed on the variable Amount with the cluster membership as a factor. The purpose of this analysis is to identify clusters whose profitability is significantly higher or lower than that of other clusters. The results are summarised in Table 6.3. As expected, the results are statistically significant. That is, the clusters do differ significantly with respect to their customer profitability.

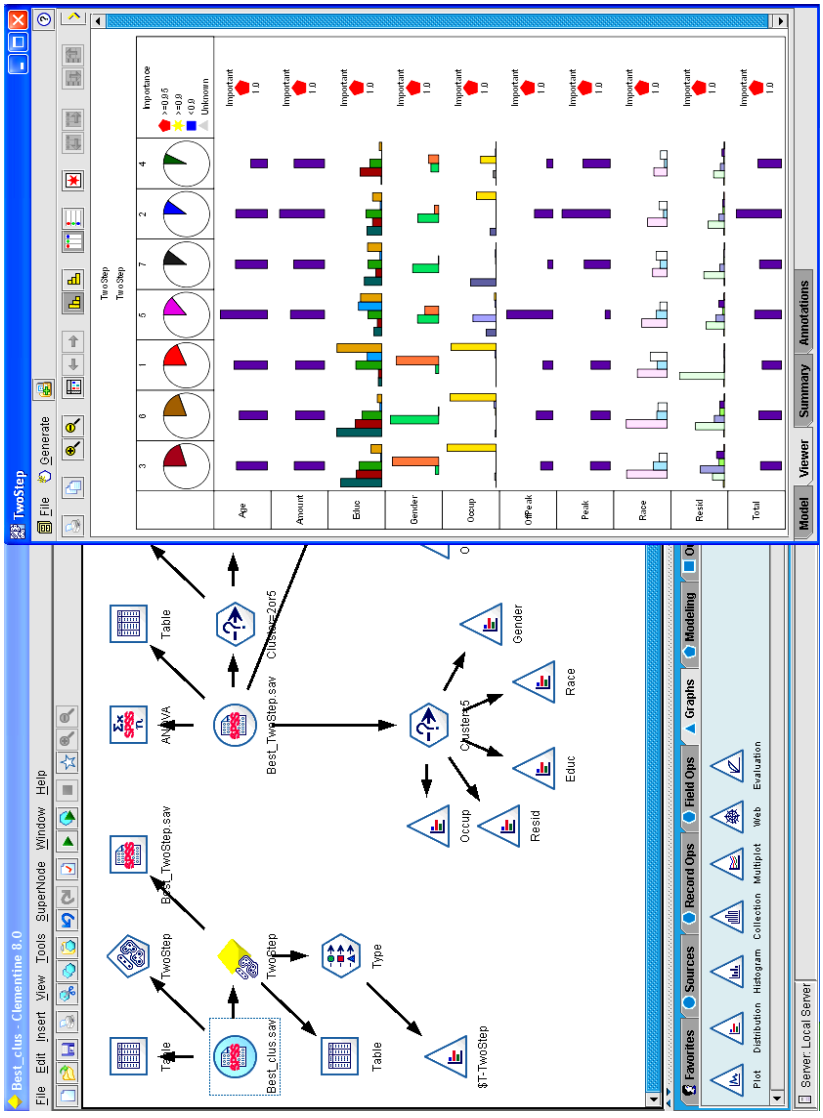


Figure 6.4 Clustering Results of Customers

Table 6.3 ANOVA and Multiple Comparison Results for Cluster Profitability

Cluster Number	Cluster Size	Subset for alpha = 0.05			
		1	2	3	4
2	271	1297.46			
5	321		962.97		
7	296			865.90	
4	196			847.03	
1	463			835.31	835.31
3	455			828.23	828.23
6	498				799.58

Given the objective of this data mining application, Best should focus on Clusters 2 and 5 because Cluster 2 has the highest profitability of $1297.46, and this is significantly higher than the profitability of any other cluster. Similarly, Cluster 5, with a profitability of $962.97, has the second highest profitability. The profitability of Cluster 5 is significantly lower than that of Cluster 2 but is significantly higher than that of the other remaining clusters. With these findings, Best is interested to identify the profile of these two most profitable clusters.

From clustering results in Figure 6.5 (left and middle panels), customers in Cluster 2 can be generally described as having the following profile: (1) an average age of about 38; (2) travelling during peak and off-peak periods about 6 and 2 times in the past three years, respectively; (3) primarily females [85.61%] and Chinese [61.25%]; (4) mainly staying in 3-/4-room HDB flats [63.10%]; and (5) currently working [82.66%]. The education level of Cluster 2 customers is quite diverse.

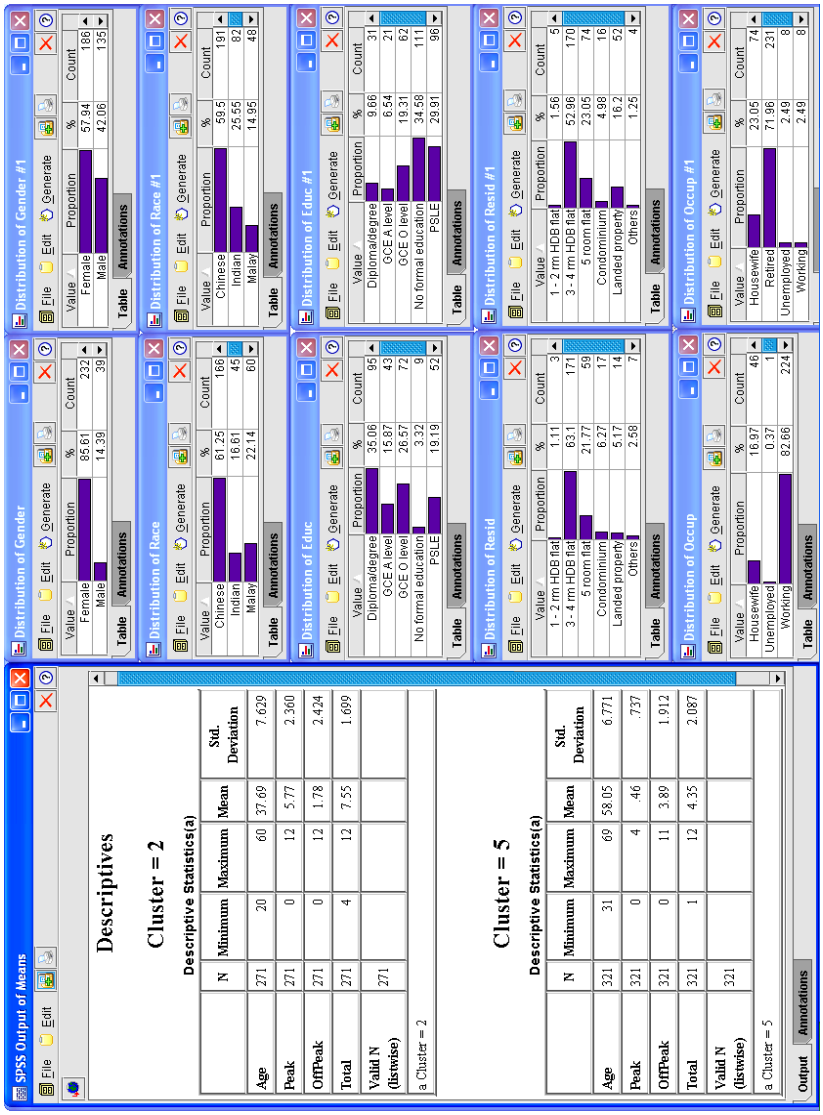


Figure 6.5 Profiles of Cluster 2 and Cluster 5

Similarly, from the left and right panels of Figure 6.5, it can be seen that customers in Cluster 5 can be generally described as having the following profile: (1) an average age of about 58; (2) travelling mainly during the off-peak periods about 4 times in the past three years; (3) primarily females [57.94%] and Chinese [59.50%]; (4) a majority with either no formal education [34.58%] or PSLE qualification [29.91%]; (5) mainly staying in 3-/4-room HDB flats [52.96%]; and (6) currently retired [71.96%].

Given the two profiles of the most profitable customers, Best can gain a better understanding of these customers and hence, design tour packages and provide services that can attract them to continue to travel with Best so as to maintain (if not enhance) its profitability. This will help Best increase its chance of success in the increasingly competitive environment in the travel industry.

6.6 Application 4: Target Mailing Campaign

Best has conducted a mailing campaign a year ago to promote a new tour package to China. The promotion brochure was sent randomly to 2500 customers and the response to the mailing campaign was reasonably good. The data related to this mailing campaign have been entered into a database (with a filename of Best_predict.sav). Best is now reviewing the data and has decided to use data mining to do target mailing for a similar new tour package to China. It aims to increase the response rate and at the same time reduce the cost of the new mailing campaign.

The database Best_predict.sav contains the variables listed in Table 6.4.

Table 6.4 List of Variables (Best_predict.sav)

Variable	Definition	Label
ID	Identification code of customer	1 to 2500
Gender Race Age Educ Resid Occup	As in Table 6.1	As in Table 6.1
Respond	Whether customer respond to the mailing campaign	0 = No; 1 = Yes
China	Whether customer take any tour to China within six months before the mailing campaign	0 = No; 1 = Yes
HK	Whether customer take any tour to Hong Kong within six months before the mailing campaign	0 = No; 1 = Yes
Ltravel	Whether customer take any tour within six months before the mailing campaign to countries other than China and Hong Kong	0 = No; 1 = Yes
Freq	Number of times the customer travelled with the agency during the past three years	1 to 12

In the last mailing campaign, 831 customers responded, giving a response rate of 33.24%. Best believes that this may be improved. Accordingly, Best wants to develop a data mining application to predict responses to a mailing campaign for China tour packages. Best has decided to bin the original age variable (measured on an interval scale) into the following five age group (AgeGp) categories such that the number of customers in each age group is about the same: (1) less than 28 years old; (2) 28 to less than 36 years old; (3) 36 to less than 41 years old; (4) 41 to less than 49 years old; and (5) 49 or more years old. Best believes that age group categories predict response to mailing campaigns better than the actual age does.

In this data mining application, the data set Best_predict.sav (comprising 2500 observations or records) is randomly partitioned into an 80% construction data set (filename = Best_predictc.sav) and a 20% validation data set (filename = Best_predictv.sav). The purpose of the validation data set is to provide an unbiased estimate of the performance of the potential response prediction models when they are applied to data outside the construction data set. In addition, the model which performs the best on the validation data set is selected as the final model for deployment. In the construction and validation data sets, the percentage of responses are 33.70% (or 674 out of 2000 customers) and 31.40% (or 157 out of 500 customers), respectively.

A decision tree model (C5.0), a neural network model and a logistic regression model are constructed to predict response to a mailing campaign for China tour packages. The target variable is Respond and the input variables are Gender, Race, AgeGp, Educ, Resid, Occup, China, HK, Ltravel and Freq (see Table 6.4 for the variables except AgeGp). The models are constructed on the construction data set and validated on the validation data set. The results are summarised in two figures. Figure 6.6 shows the

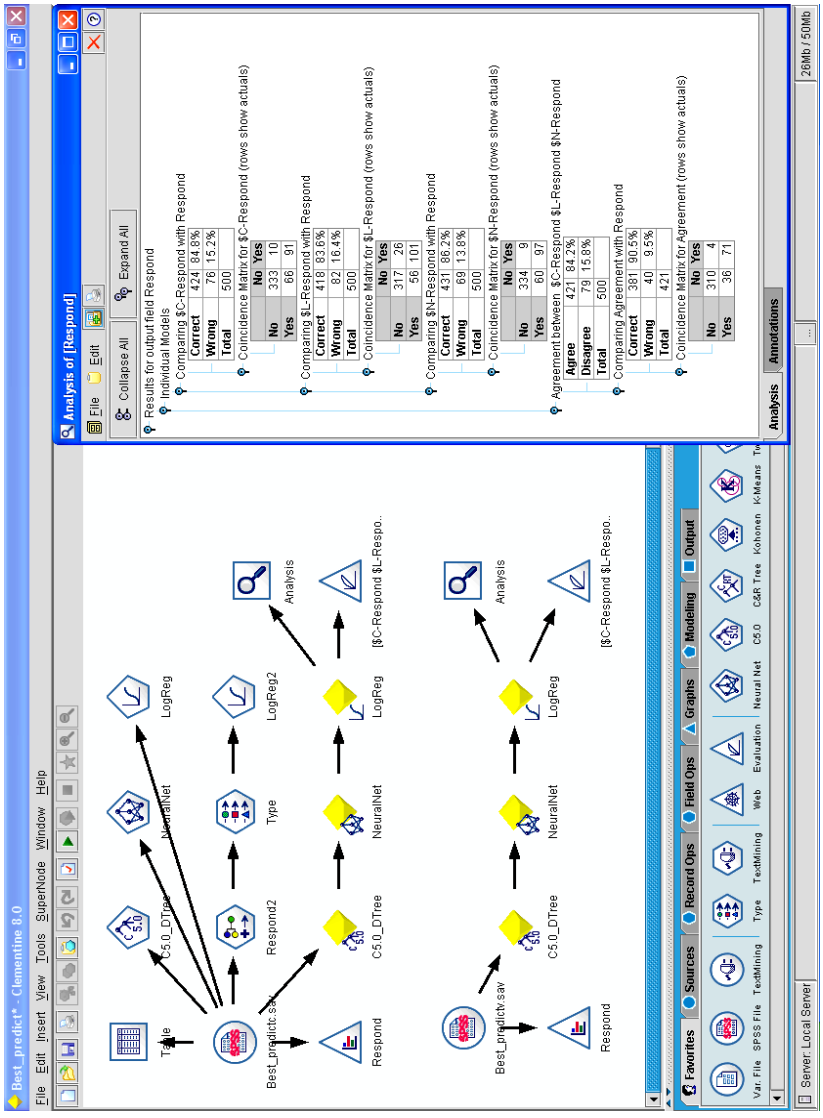
accuracy rates and Figure 6.7 shows the lift charts (both sets of results for the validation data set).

As can be seen from the overall accuracy rates of the three potential response models in Figure 6.6, the performance of the models on the validation set is similar, ranging from a low of 83.60% for the logistic regression model to a high of 86.20% for the neural network model. The performance for the C5.0 decision tree model is 84.80%.

Figure 6.6 also shows that the logistic regression model predicts response most accurately (101 out of 157, or 64.33%) and the neural network model predicts non-response most accurately (334 out of 343, or 97.38%).

The lift charts in Figure 6.7 indicate that the neural network model dominates the logistic regression and decision tree models in the first three deciles. Assuming that with limited resources, Best plans to target at most the top 30% customers with the highest predicted probability of response, the neural network model is the champion or best model to deploy.

As shown in Figure 6.8, the sensitivity analysis of the impact of the input variables on the target variable indicates that the five most important input variables associated with customers' response or non-response to a mailing campaign for China tour packages are (in descending order of importance): (1) race; (2) gender; (3) whether any tour has been taken within six months before the mailing campaign to countries other than China and Hong Kong; (4) whether any tour has been taken to Hong Kong within six months before the mailing campaign; and (5) whether any tour has been taken to China within six months before the mailing campaign. A partial listing of the neural network weights is also given in Figure 6.8. With the weights, Best will be able to predict the probability of response to a mailing campaign for China tour packages for its database of customers.



.

Figure 6.6 Accuracy Rates of Response Models

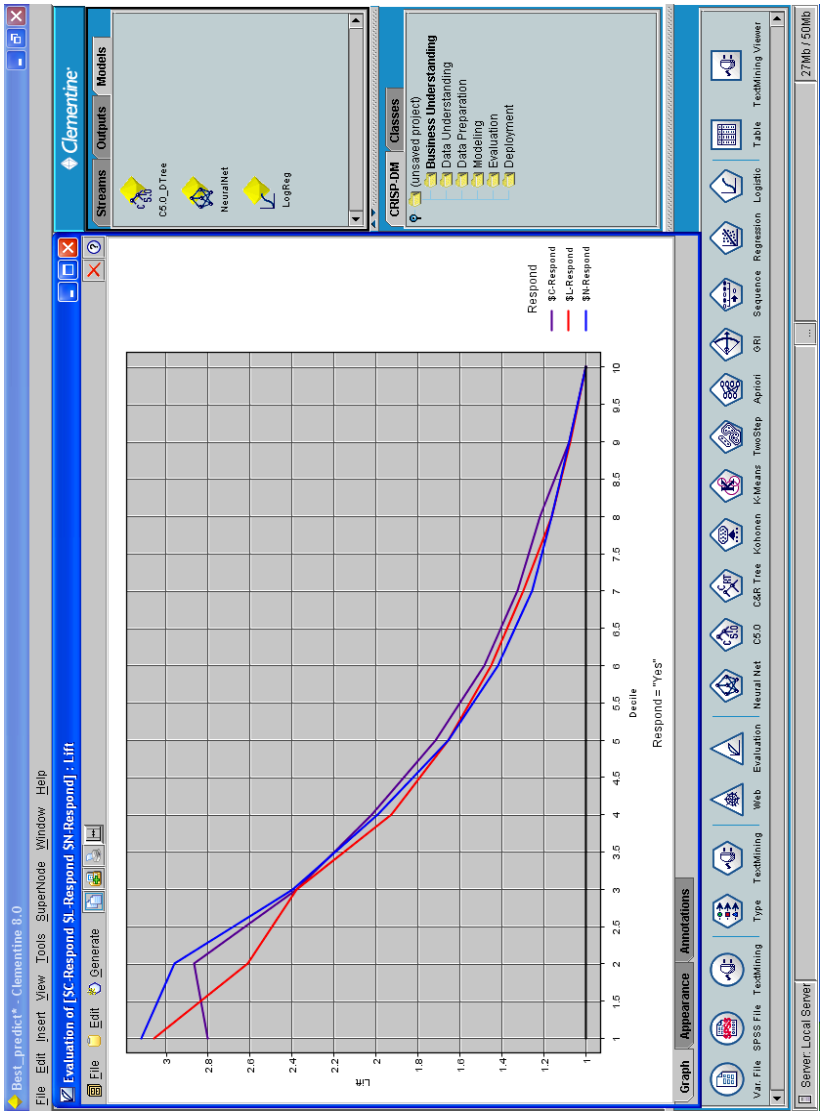
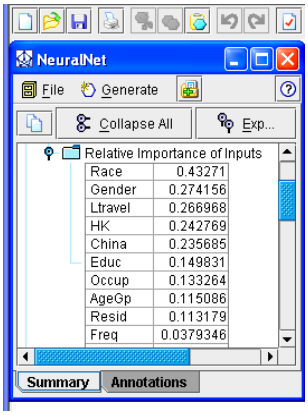


Figure 6.7 Lift Charts of Response Models



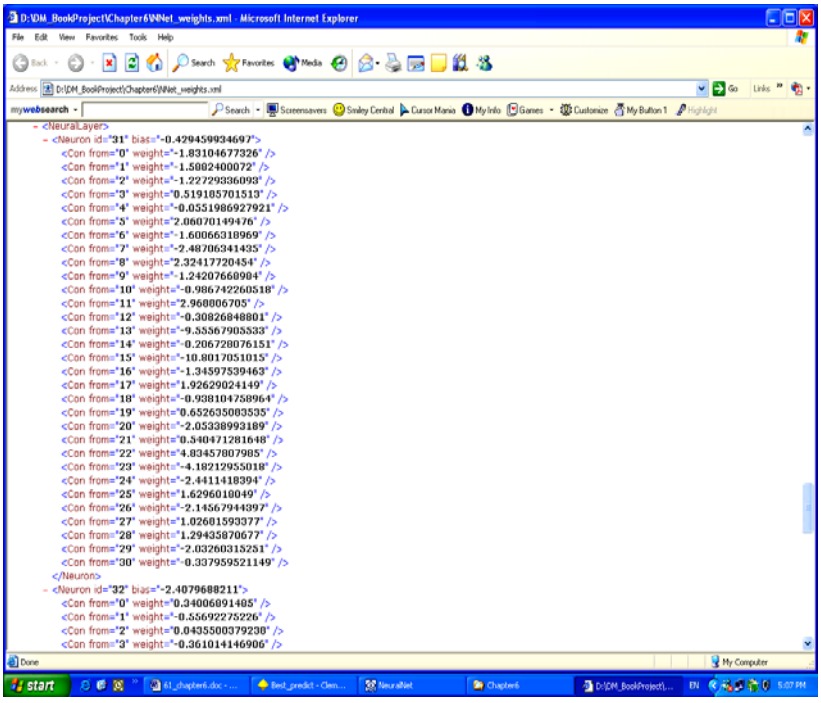


Figure 6.8 Sensitivity Analysis Importance and Model Weights

A disadvantage of the neural network model is that it is a "black box" that does not reveal how the input variables (such as race or gender) are associated with response or non-response. To mitigate this problem, a decision tree model is used to model the neural network predictions (see Section 3.3 of Chapter 3). It should be emphasised that this decision tree models the neural network predictions and not the target variable. Hence, it can reveal the relationships captured by the neural network model (assuming that the performance of the decision tree model can be considered adequate). The resulting decision tree is shown in Figure 6.9.

The results show that customers who respond positively to a mailing campaign for China tour packages are likely to be: (1) customers who are Chinese and female and who have taken a tour within six months before the mailing campaign to countries other than China or Hong Kong [node 10 in Figure 6.9]; (2) Chinese customers who have neither taken a tour within six months before the mailing campaign to countries outside of China or Hong Kong nor to China [node 8]; and (3) customers who are Chinese and female and who have taken a tour within six months before the mailing campaign to China but not to countries outside of China or Hong Kong [node 12].

Similarly, the results show that customers who respond negatively to a mailing campaign for China tour packages are likely to be: (1) customers who are either Indian or Malay [node 2]; (2) Chinese customers who have taken a tour within six months before the mailing campaign to Hong Kong as well as countries outside of China and Hong Kong [node 5]; (3) customers who are Chinese and male and who have taken a tour within six months before the mailing campaign to countries outside of China or Hong Kong but not to Hong Kong [node 9]; and (4) customers who are Chinese and male and who have taken a tour within six months before the mailing campaign to China but not to countries outside of China or Hong Kong [node 11].

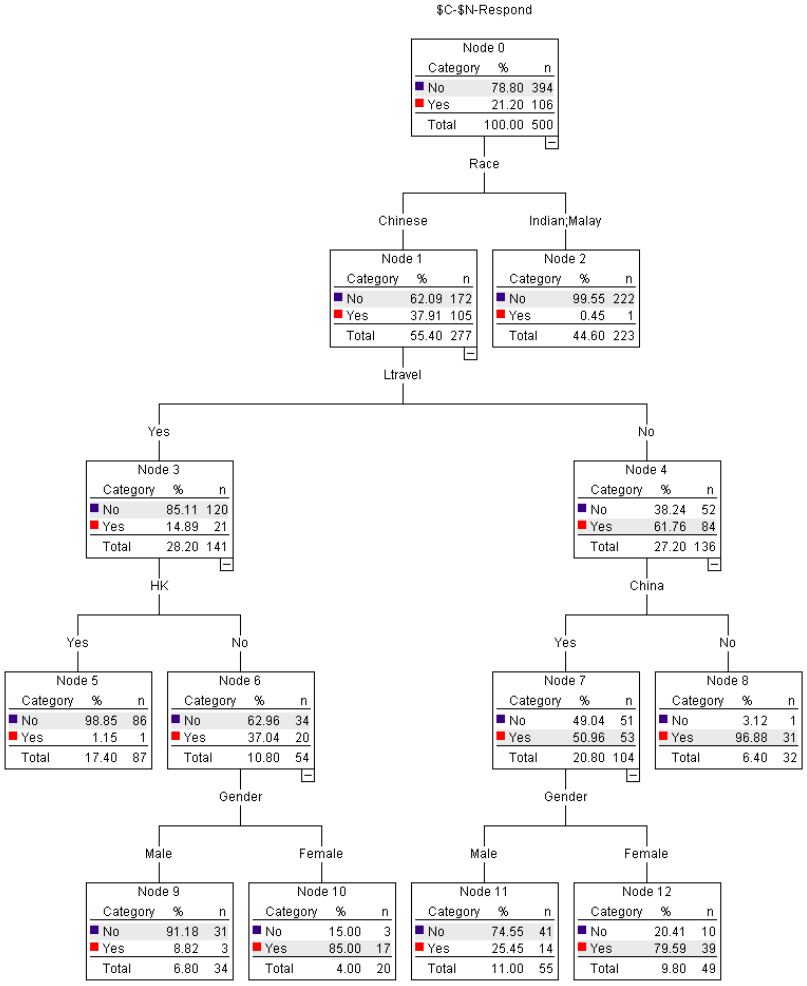


Figure 6.9 Decision Tree Modelling of the Neural Network Model

The relationships listed above appear to be the ones captured by the neural network model. Other than the input variables Race, Ltravel, HK, China and Gender, the other variables (i.e., AgeGp, Educ, Resid, Occup and Freq) do not appear to be associated with response/non-response to a mailing campaign for China tour packages.

In deploying the data mining application, the neural network model will be applied to the target database to compute the predicted probability of response. It is assumed that this database is up-to-date with respect to the variables in the model. Also, it is assumed that Best has knowledge of the tours taken by its customers. The potential customers in the database can then be ranked in descending order of the predicted probability. To increase the response rate and reduce the cost of the mailing campaign, Best can send the campaign brochures only to the top, say 30%, of the potential customers who are the most likely to respond positively.

6.7 Concluding Remarks

This chapter further illustrates the application of data mining in organisations. In particular, it discusses in the context of the service industry (in this case, a travel agency), the use of data mining to identify dissatisfied customers, develop tour packages, profile customers and target potential customers in a mailing campaign. These applications are only illustrative and there can be many other data mining applications in the service industry (see, for example, Berry and Linoff, 2000; Drew et al., 2001; and Bloemer et al., 2002).

Chapter References

Anonymous. (2003), “Gordon to chair PATA board meeting in Singapore”, *BusinessWorld*, October, p. 1.

Berry, M. J. A. and Linoff, G. S. (2000), *Mastering Data Mining: The Art and Science of Customer Relationship Management*. John Wiley & Sons, Inc., New York.

Bloemer, J., Brijs, T., Swinnen, G. and Vanhoof, K. (2002), "Identifying latently dissatisfied customers and measures for dissatisfaction management", *The International Journal of Bank Marketing*, Vol. 20 No. 1, pp. 27-27.

Drew, J. H., Mani, D. R., Betz, A. L. and Datta, P. (2001), "Targeting customers with statistical and data-mining techniques", *Journal of Service Research*, Vol. 3 No. 3, pp. 205-219.

Hamdi, R. (2003), "Region battles to get tourists back", *Media*, August, pp. 22-23.

Singh, A. (1997), "Asia Pacific tourism industry: current trends and future outlook".

Yahya, F. (2003), "Tourism flows between India and Singapore", *International Journal of Tourism Research*, Vol. 5 No. x, pp. 347-367.