Chapter 5 Potential Applications for a Retailer

5.1 Introduction

The first four chapters of the book discuss the fundamentals of data mining. These include: (1) the definition and methodology of data mining; (2) the various data mining tools for description and visualisation, association and clustering, and classification and estimation (i.e., predictive modelling); and (3) important data mining issues such as model validation, optimal cut-off points and evaluation charts.

At this juncture, it is appropriate to present case studies illustrating the potential applications of data mining. This is done in Chapters 5 to 7 for a retailer, a service provider and a manufacturer, respectively. The main objective of these case studies is to help organisations get started in data mining. Throughout these three chapters, SPSS software (in particular SPSS Clementine) is used.

Many of the illustrations presented in this chapter and the next revolve around customer analytics, which in turn, revolves around "getting to know your customers". For example, an organisation can ask the following questions about its customers: (1) who are they? (2) what are their needs? (3) how do they behave? (4) where are they? and (5) why do they churn? Knowing an organisation's customers is critical to its gaining a competitive edge.

To know its customers, an organisation needs to collect data through customer interactions and other sources, analyse the collected data, deploy the analytical results in customer interactions via operational systems

and collect new data. Customer analytics focuses on the analysis and deployment components. Data mining can play a major role here – for instance, in customer relationship management (e.g., customer valuation, cross-/up-selling and churn modelling) and market segmentation (e.g., customer profiling and response-based segmentation). Other more unique applications in customer analytics include counter measures to deter retail criminals (Anonymous, 2004a), customer targeting and retention via e-mail (Anonymous, 2004b) and video mining (Bednarz, 2004).

5.2 Context

This chapter focuses on the potential data mining applications for a retailer – in particular, a supermarket. In an annual survey of retail information technology, Ernst & Young concluded that data mining had grown in usage and effectiveness (Anonymous, 1999). In particular, at the time of the survey, 62 percent of the respondents were using data mining, and 21 percent of the remaining respondents who were not using data mining reported that they planned to use it.

The last decade has seen significant changes in the retail industry. Some of these changes have important implications on data mining. For example, the introduction of bar-code scanners and universal bar-coding has resulted in the accumulation of a wealth of data. Transactional data are now conveniently captured at the point-of-sale. In addition, the use of credit cards and loyalty card programmes has allowed anonymous transactions to be replaced by purchases linked to individual customers. Hence, demographic data and transactional data can now be analysed together to yield richer information on customers and their purchasing patterns.

The advent of the Internet has also opened tremendous marketing opportunities for retailers (e.g., through innovations such as eCommerce and eMarketing). The web has provided a virtual store offering endless

possibilities for tailoring shelf space for each potential customer as opposed to the traditional physical store with predefined shelf placement. In addition, it has generated massive data on what customers buy and what they look at during the online shopping process. Consequently, the Internet has provided another level of detail for understanding customer behaviour.

The retail industry has also seen a shift of focus from products to customers. Meeting customers' needs and keeping customers satisfied are now more important than just pushing products and making sales.

Data mining applications in the retail industry are well documented in the literature. Examples include applications to obtain insights into customer tastes, purchasing patterns, market share, site locations, patronage and targeting (Peterson, 2003), applications to manage inventory, promotions, margin control and negotiation with suppliers (Reid, 2003) and applications to increase returns from customer interactions, up-/cross-/down-selling efforts and multi-channel customer analysis (Fayyad, 2003).

5.3 Background Information for Data Mining Illustrations

To illustrate potential data mining applications for a retailer, consider a fictitious supermarket chain Supreme Supermart. To facilitate discussion, the shortened name "Supreme" will be used from this point onwards to refer to this supermarket chain. Assume that Supreme has been in business for the last 15 years and has six outlets in Singapore, located mainly in housing estates.

Two years ago, Supreme introduced a loyalty card programme to reward its customers who shop frequently and spend substantial amounts in its stores. Each outlet has bar-code scanners at the checkout counters to record each item purchased. With the entry of big foreign supermarkets which can offer even lower prices for groceries and household items, Supreme is facing strong competition from both local and foreign competitors.

The company has decided to make use of its customer and sales data to understand who its customers are and what they buy. It hopes to use the information generated from data mining to raise its profitability.

For the data mining applications illustrated in this chapter, Supreme has three databases. The first database (filename = Supreme_assoc.sav) contains transactional data that are related to baby products as summarised in Table 5.1.

Table 5.1 Database on Baby Products

Variable	Definition	Label
BktID	Identification code of market basket	1 to 10000
SKU1 to SKU5	Milk Powder Brand A to Milk Powder Brand E	T = True; F = False
SKU6 to SKU11	Baby Cereal Brand F to Baby Cereal Brand K	T = True; F = False
SKU12 to SKU17	Baby Diapers Brand L to Baby Diapers Brand Q	T = True; F = False
SKU18 to SKU20	Baby Wipes Brand R to Baby Wipes Brand T	T = True; F = False
SKU21 and SKU22	Baby Detergent U and Baby Detergent V	T = True; F = False
SKU23 to SKU26	Baby Canned Food Brand W to Baby Canned Food Brand Z	T = True; F = False
SKU27 to SKU30	Baby Canned Food Brand AA to Baby Canned Food Brand AD	T = True; F = False
Note: SKU = Stock-keeping unit		

The database captures the purchase of baby products in a single market basket. That is, it captures the baby products that are purchased by a particular customer in a particular transaction at Supreme. The baby products are categorised into 30 SKUs (i.e., stock-keeping units), where purchase is indicated by a "True" value (i.e., "T") and non-purchase by a "False" value (i.e., "F"). To illustrate the interpretation of Table 5.1, SKU12 refers to "Baby Diapers Brand L", SKU28 refers to "Baby Canned Food Brand AB" and so on. The other two databases will be discussed in later sections.

5.4 Application 1: Baby Products Promotion

In this data mining application, Supreme wishes to run an advertisement in the newspapers to promote the purchase of discounted items. The theme is on baby products and the target customers are households with young children. Supreme wants to bundle items together to increase the amount spent by its customers. Specifically, it plans to entice customers to buy the bundled items instead of (fewer) individual items. In the first week, selected pairs of baby products will be offered at a 10% discount. In the second week, a 50% discount will be given to a specified (third) baby product for each purchase of selected pairs of baby products.

To meet Supreme's objective, web graphs and market basket analysis (MBA) are employed to identify baby products that customers tend to purchase together. As web graphs and MBA do not require specific information on the customers, anonymous customer transactional data can be used. The transactional data are put into a format where each record (or observation) represents a "shopping basket" (i.e., it shows all the items purchased by a customer at the counter). Each column represents a baby product sold by Supreme. The value "T" represents an item being purchased and "F" represents the item not being purchased in a particular transaction.

In the database, there are a total of 10,000 records (rows) and 30 baby products (columns).

To gauge the popularity of baby products sold by Supreme, the Distribution node in SPSS Clementine is used to count the frequency and compute the percentage of each of the products. The results are summarised in Figure 5.1.

As shown, the most popular baby product is SKU29 (Baby Canned Food Brand AC), appearing in 45.30% of customer shopping baskets, followed by SKU19 (Baby Wipes Brand S – 37.96%) and SKU20 (Baby Wipes Brand T – 36.88%). The least popular baby product is SKU3 (Milk Powder Brand C), appearing in only 17.96% of customer shopping baskets. From the results, it can be concluded that the baby products sold by Supreme are popular with customers and are in demand.

To select the pairs of baby products to be promoted in the first week, the Web node is used to gauge the pairwise popularity of the baby products (i.e., which two baby products are frequently purchased together). As shown in Figure 5.2, strong links exist for the purchase of: (1) SKU5 [Milk Powder Brand E] and SKU12 [Baby Diapers Brand L]; (2) SKU19 [Baby Wipes Brand S] and SKU29 [Baby Canned Food Brand AC]; and (3) SKU20 [Baby Wipes Brand T] and SKU29 [Baby Canned Food Brand AC]. The number of links is also shown in Figure 5.2 (see right panel). For example, there are 1735 transactions in which SKU5 and SKU12 are purchased together.

Moderate links also exist for the purchase of: (1) SKU14 [Baby Diapers Brand N] and SKU29 [Baby Canned Food Brand AC]; (2) SKU2 [Milk Powder Brand B] and SKU29 [Baby Canned Food Brand AC]; (3) SKU28 [Baby Canned Food Brand AB] and SKU29 [Baby Canned Food Brand AC]; and (4) SKU21 [Baby Detergent U] and SKU29 [Baby Canned Food Brand AC]. The frequent links with SKU29 is not surprising given that Baby Canned Food Brand AC is the most popular baby item sold by Supreme. Figure 5.2 also shows some of the other weaker links.

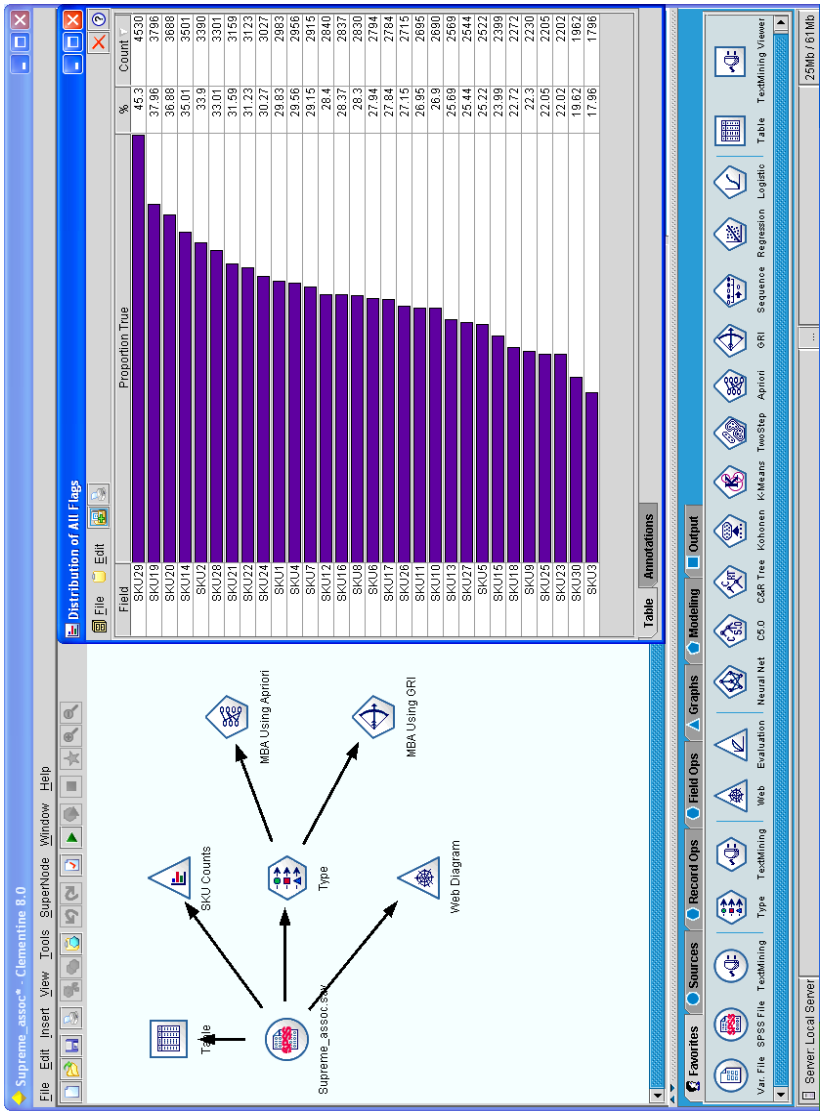


Figure 5.1 Popularity of Baby Products

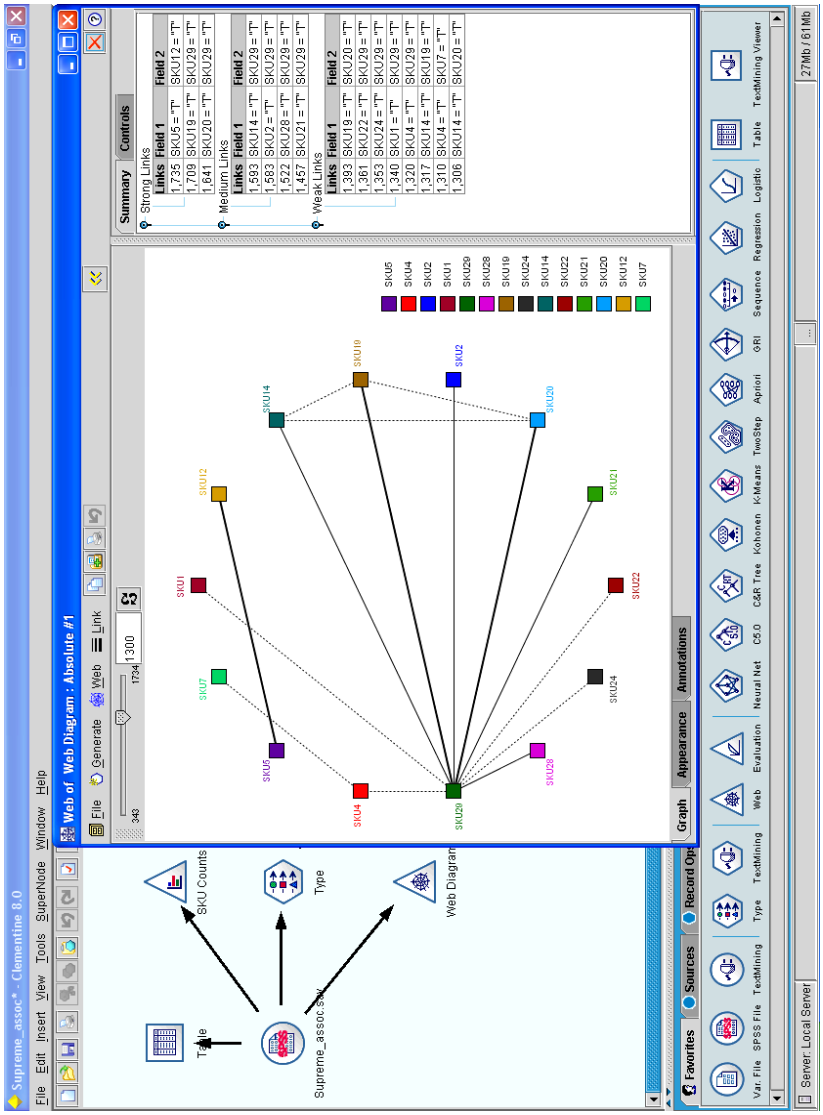


Figure 5.2 Web Graph of Baby Products

Supreme has decided to use all the seven pairs listed above in the first week of the baby products promotion campaign.

For the second week, association analysis (i.e., market basket analysis) is performed to generate association rules for the SKUs. For this purpose, the SPSS Clementine nodes Apriori and GRI (generalised rule induction) are used. These are different algorithms for generating association rules. The association rules are summarised in Figure 5.3.

The rationale for using association analysis in this application is the need to identify the "third" item that is attractive to customers based on their purchase of the first two baby products. As shown in Figure 5.3, the rules generated by Apriori and GRI are consistent. In particular, the three-item association rules can be stated as follows:

1) When SKU8 [Baby Cereal Brand H] and SKU12 [Baby Diapers Brand L] are purchased, there is a 74.6% probability that SKU5 [Milk Powder Brand E] is also purchased;
2) When SKU6 [Baby Cereal Brand F] and SKU4 [Milk Powder Brand D] are purchased, there is a 71.1% probability that SKU7 [Baby Cereal Brand G] is also purchased;
3) When SKU7 [Baby Cereal Brand G] and SKU4 [Milk Powder Brand D] are purchased, there is a 66.0% probability that SKU6 [Baby Cereal Brand F] is also purchased; and
4) When SKU12 [Baby Diapers Brand L] and SKU29 [Baby Canned Food Brand AC] are purchased, there is a 61.3% probability that SKU5 [Milk Powder Brand E] is also purchased.

In generating the rules above, the minimum support level is set at 12.0% (i.e., the antecedents must occur in at least 1200 transactions out of the 10,000 transactions in the database) and the minimum confidence level is set at 60.0% (i.e., the consequent must occur at least 60.0% of the time when the antecedents occur).

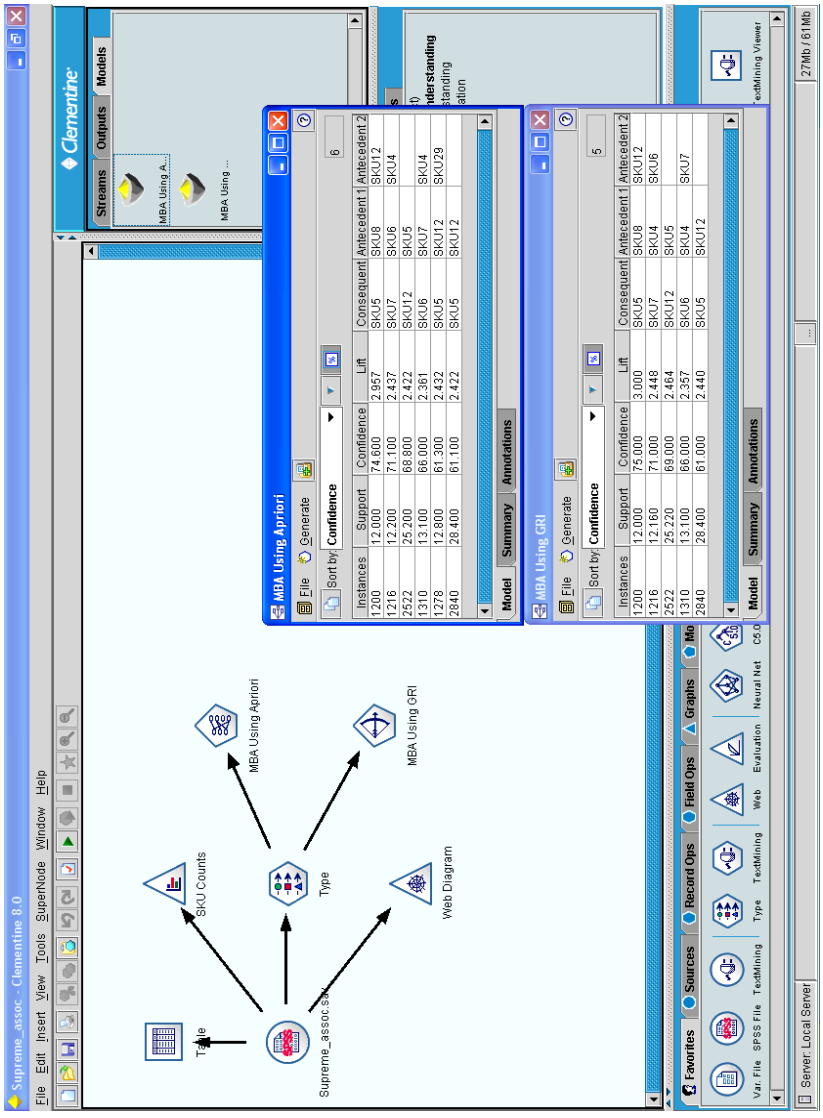


Figure 5.3 Association Rules for Baby Products

For the second week of the baby products promotion campaign, the bundling decisions can be based on the above association rules (i.e., the antecedents are the selected pairs and the consequent the specified third item entitled to a 50% discount). These rules ensure that the baby products that customers frequently purchase together are bundled together in the promotion campaign. This is expected to increase the purchase of baby products in Supreme.

5.5 Application 2: Customer Segmentation

In this data mining application, the second database (with the filename "Supreme_clus.sav") is used. It comprises demographic variables such as a customer's loyalty card number (CustID), gender (Gender), race (Race), age (Age), educational level (Educ) and type of residence (Resid). It also contains the annual amounts spent by the customer on fresh fruits and vegetables (Fruitveg), fresh meat (Fshmeat), dairy products (Dairy), frozen meat (Fznmeat), canned fruits and canned vegetables (Cfrtveg), canned meat (Cmeat), soft drinks (Softdnks) and beer and wine (Beerwine). A summary is shown in Table 5.2.

The objective of this data mining application is to uncover the purchasing patterns of existing customers so that tailored services can be provided to selected customer segments. By serving its most valued customers better, Supreme hopes to induce them to continue to patronise its outlets. To do this, Supreme needs to understand the behaviour and needs of its customer segments.

Although there is much discussion in the CRM (customer relationship management) literature about 1-to-1 marketing, it is more feasible to interact with customers in groups, where customers within the same group have similar behavioural patterns and needs while those in a different group have different patterns and needs.

Table 5.2 Variables in Supreme_clus.sav

Variable	Definition	Label
CustID	Customer loyalty card number	1 to 3000
Gender	Gender of customer	0 = Male; 1 = Female
Race	Race of customer	1 = Indian 2 = Malay 3 = Chinese
Age	Age of customer	
Educ	Educational level of customer	1 = No formal education 2 = PSLE 3 = GCE O Level 4 = GCE A Level 5 = Diploma/degree
Resid	Type of residence 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 (e.g. HUDC)
Annual amounts spent on:		
Fruitveg	Fresh fruits and fresh vegetables	
Fshmeat	Fresh meat	
Dairy	Dairy products	
Fznmeat	Frozen meat	
Cfrtveg	Canned fruits and canned vegetables	
Cmeat	Canned meat	
Softdnks	Soft drinks	
Beerwine	Beer and wine	

Among other things, this group (or cluster) approach can help to facilitate marketing communication (e.g., different advertisements or messages for different groups), product development and customer service (e.g., different products or services for different groups), or the identification of key segments (which can be managed differently to increase their value or reinforce their loyalty).

To meet the above objectives, Supreme has decided to perform customer segmentation using clustering. This analysis is to be applied to the existing customer database to partition the customers into several separate segments based on their purchasing patterns. Eight transaction variables have been measured to capture the customers' purchasing patterns (see variables Fruitveg to Beerwine in Table 5.2). The database contains 3000 customer records (i.e., observations) and 14 variables. Besides the eight transaction variables, five variables describe the demographic characteristics of the customer (see variables Gender to Resid in Table 5.2) and one variable captures the customer loyalty card number (CustID).

Before going into clustering, Supreme has decided to perform some preliminary data exploration. For this, the SPSS Clementine Histogram and Statistics nodes are applied to provide more insight into the customer data. The results are summarised in Figure 5.4.

The Histogram results reveal that the eight transaction variables are highly skewed to the right, with a large number of zero values and the presence of at least one extreme outlier. In addition, the Statistics results indicate that the eight transaction variables have quite different ranges of values. For example, the values for Fshmeat ranges from $0 to $1096.45 with a mean of $204.61 while Cmeat ranges from $0 to $183.06 with a mean of $38.46.

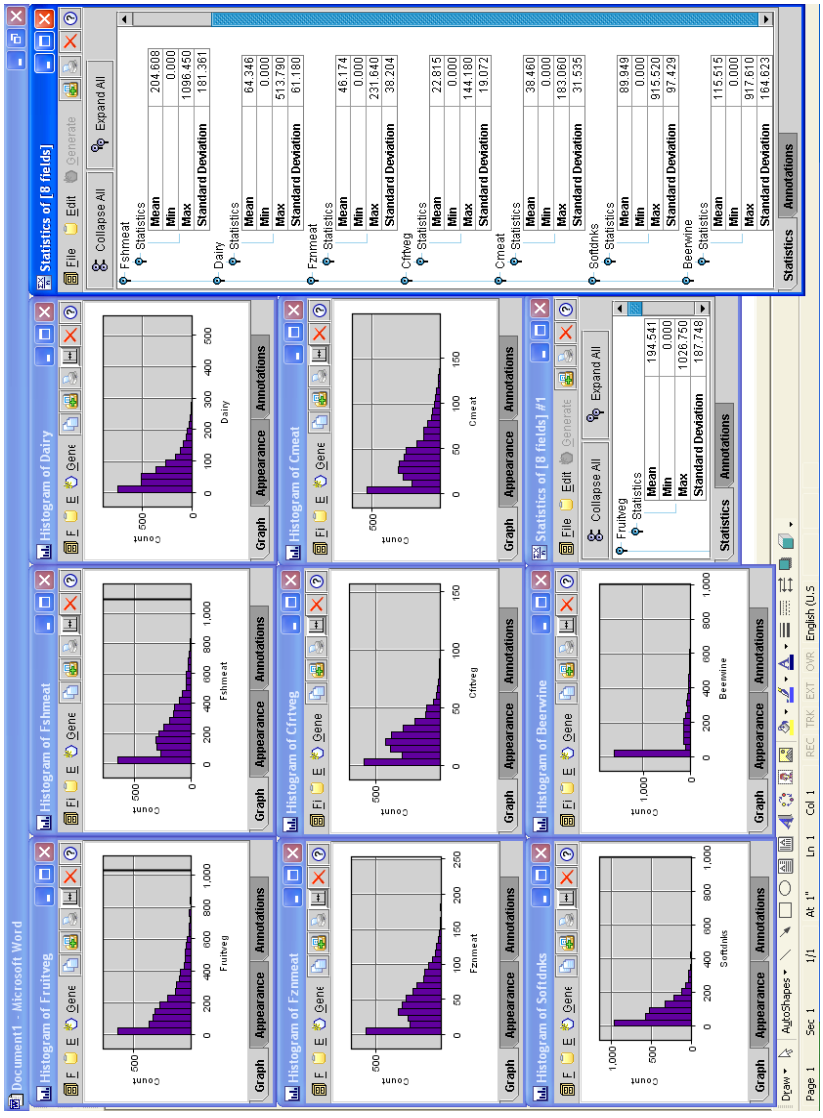


Figure 5.4 Results of Data Exploration

Such observations are common in transactional data and may get in the way of a good clustering solution. Thus, it is necessary to transform the data before clustering is attempted. Some possible transformation methods include: (1) changing the transaction variables to indicator (or dummy) variables that indicate purchase or no purchase, ignoring the actual value of purchases; (2) grouping the values of the variables into several bins (i.e., ranges); and (3) calculating the percentage or proportion of total spending for each variable. The second approach is used in this illustration.

The original transaction variables are binned into deciles (i.e., 10 bins with approximately equal numbers of observations) before performing clustering. For this, Clementine Binning nodes are used. The new binned variables are then used for clustering the customers (on the basis of their purchasing pattern or behaviour). In addition to the creation of binned variables, a new variable "Totalamt" is created to obtain the total amount spent across all the eight transaction variables (i.e., all the categories of food items). This new variable will be used after the final clustering solution is obtained to identify the highest spending customer segment.

To perform clustering, Supreme has decided to use the SPSS Clementine TwoStep clustering node. This clustering algorithm has the advantage of determining the optimal number of clusters. The clustering results are summarised in Figure 5.5.

As shown on the upper left panel of Figure 5.5, the optimal number of clusters is three. The three clusters have sizes of 1543, 1090 and 367 customers, respectively. The graphical view of the cluster profile (see right panel) reveals that Cluster 3 contains customers who are high spenders. In particular, they have large spending (as seen by the relatively higher frequencies in larger deciles) across all the food categories, except for the soft drinks category. Cluster 2, on the other hand, contains customers who are the medium spenders, with moderate amounts spent in all the categories on average.

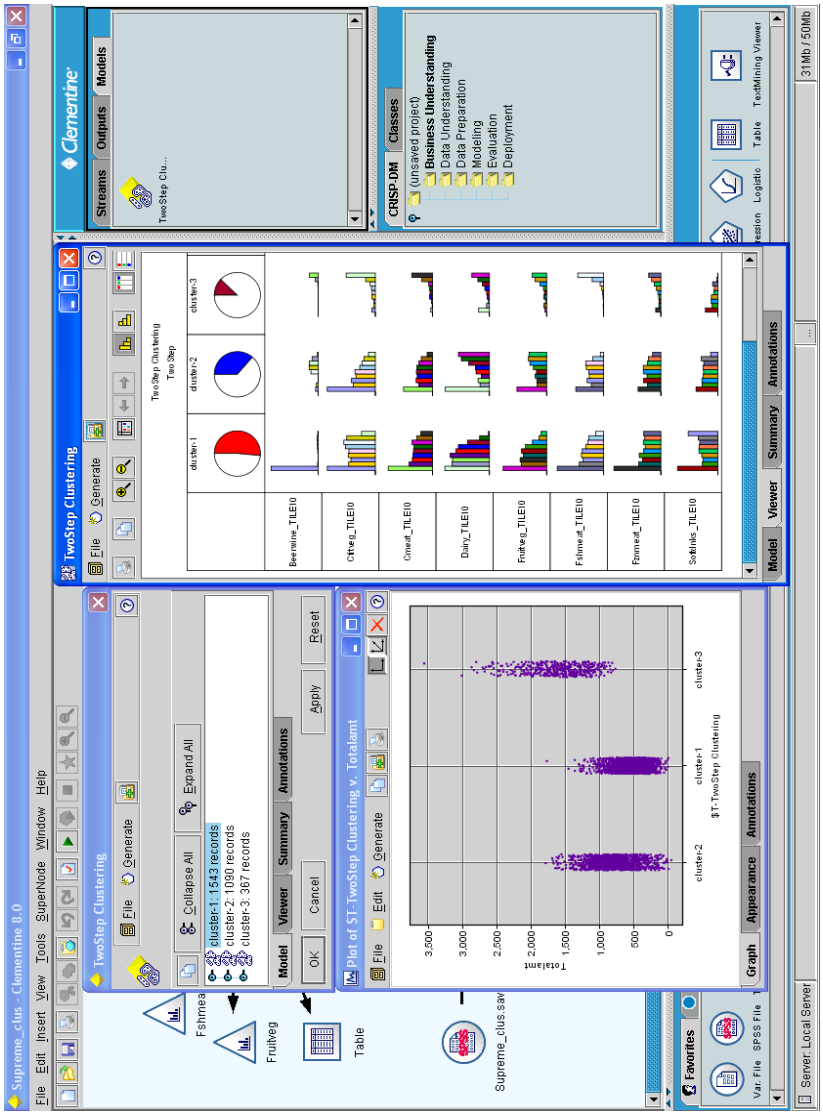


Figure 5.5 Clustering Results

Finally, Cluster 1 contains the low spending customers. However, as compared to the other two clusters, they spend (relatively) more on soft drinks (see distribution of Softdrinks_TILE10). The spending pattern of the three clusters is also confirmed by the plot shown in the lower left panel of Figure 5.5.

Although the optimal number of clusters formed is only three, the clustering solution is able to generate groups of customers who are quite distinctive in their spending behaviour (i.e., high, moderate and low spenders). Further, the number of observations captured in the high spending segment (367 customers) makes it feasible for Supreme to follow up on this (manageable) group of customers.

To understand the cluster profile better, an attempt is made to relate the demographic variables of the customers to their cluster membership. Several description and visualisation nodes in SPSS Clementine are invoked and the results are summarised in Figure 5.6. The main findings are: (1) the high spenders are all female customers; (2) proportionately more Indian customers are high spenders; and (3) there are no discerning relationships between spending patterns and customers' age, education level and type of residence. Similar analyses can also be made for medium and low spenders.

With such information, Supreme can tailor their products and services to the different customer segments. It can also customise marketing communication as well as focus on the high spending segment, among other things.

5.6 Application 3: Churn Modelling

In this final illustrative data mining application for Supreme, the third database (with the filename "Supreme_pred.sav") is used. A summary of the variables is given in Table 5.3.

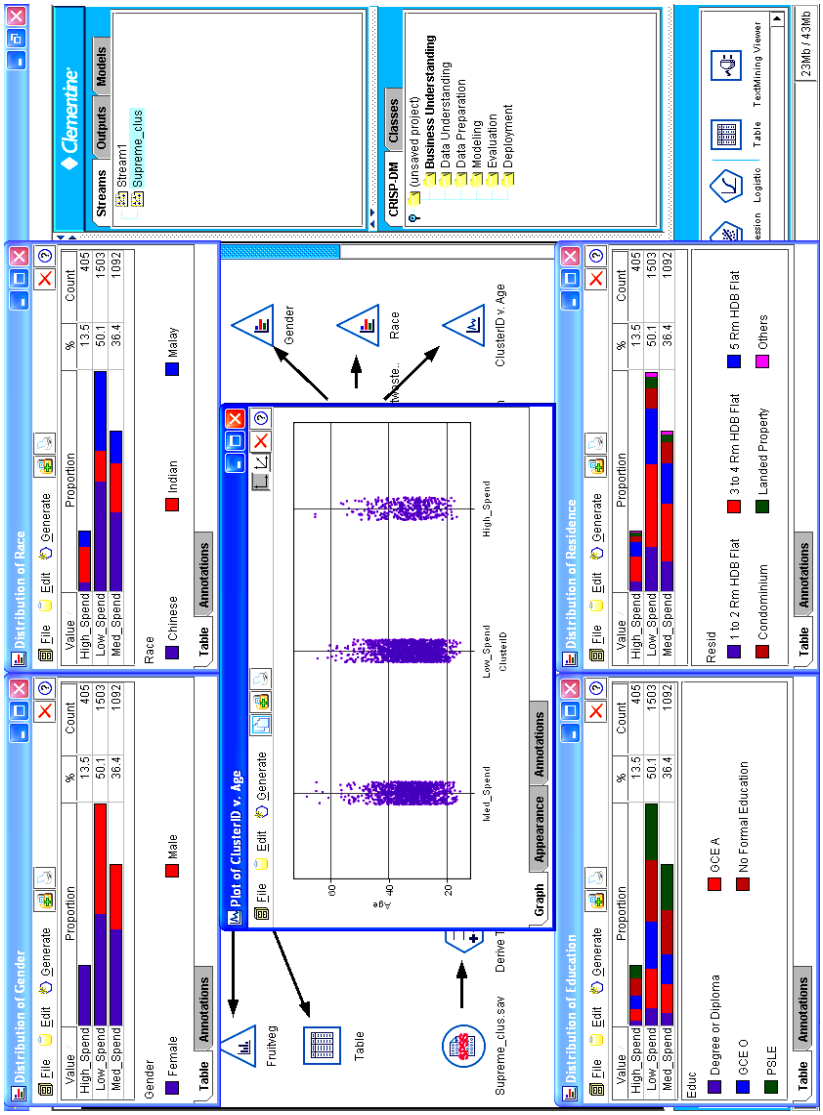


Figure 5.6 Demographic Characteristics and Cluster Membership

Table 5.3 Variables in Supreme_pred.sav

Variable	Definition	Label
CustID	Customer loyalty card number	1 to 10000
Gender Race Age Educ Resid	As in Table 5.2	As in Table 5.2
NeighIn	Neighbourhood indicator – whether customer's address is within 1 km from Supreme's outlet	N = No; Y = Yes
RedeemIn	Redeem indicator – whether customer has redeemed his/her loyalty card points in the last six months	N = No; Y = Yes
DataIn	Data set indicator – to denote the different periods of data	
Amtspt1 to Amtspt7	Monthly amount spent one month to seven months before the month of latency, respectively	
ChurnIn	Churn indicator	0 = No; 1 = Yes

The database comprises demographic variables similar to the database used in the previous section (i.e., CustID, Gender, Race, Age, Educ and Resid). However, there are 10,000 observations or records in the current database (as compared to 3000 observations in the previous database). It also contains the monthly amounts spent one month to seven months before the month of latency (denoted as Amtspt1 to Amtspt7) as well as three indicator variables (i.e., NeighIn, RedeemIn and DataIn). (The

month of latency will be elaborated below). Finally, the churn status is captured in the database as ChurnIn.

In this application, the objective is to predict customers who are likely to churn (i.e., churn modelling is attempted). Supreme intends to apply the data mining results on existing customers to identify those who exhibit the same behaviour as the churners – especially profitable ones – so that actions can be taken to reinforce their loyalty before they are lured away by Supreme's competitors.

Generally, in churn modelling, data from the past (e.g., monthly transactional data) are used to predict future behaviour (i.e., churn). In particular, data from several past months are used to predict churn behaviour in some future month. Also, to be effective, the churn model has to move forward in time. For example, data in and before month x are used to predict churn behaviour in month y, data in and before month (x + 1) are used to predict churn behaviour in month (y + 1) and so on. Since time is needed to collect data as well as take actions to reduce churn, there must be a time gap between months x and y. The month(s) between x and y is (are) referred to as the month(s) of latency. (More discussion on latency can be found in Berry and Linoff, 2000).

To elaborate, in the modelling stage, several months of past transactional data are available and therefore it is possible to use data in and before a particular month to predict churn behaviour in the next month. However, in the deployment stage when the churn model is actually applied, it may be the case that for any particular month when churners are to be identified (i.e., predicted) for the month after, the latest data available are those one month before (e.g., in February, only data for January and before are available). Further, when prediction is done in February, it will be to identify churners in March so that after the prediction (in February), pre-emptive actions can be taken to prevent churn in March. Hence, a realistic churn model will have to be one that uses data one month before to predict

in the current month the potential churners in the next month. As such, there is a time gap of one month between the month for which the latest data are available and the month churn is predicted to occur. The time gap is also necessary to effect pre-emptive actions. In this illustration, the month within the time gap is the month of latency. Data for modelling are counted backwards from the month of latency.

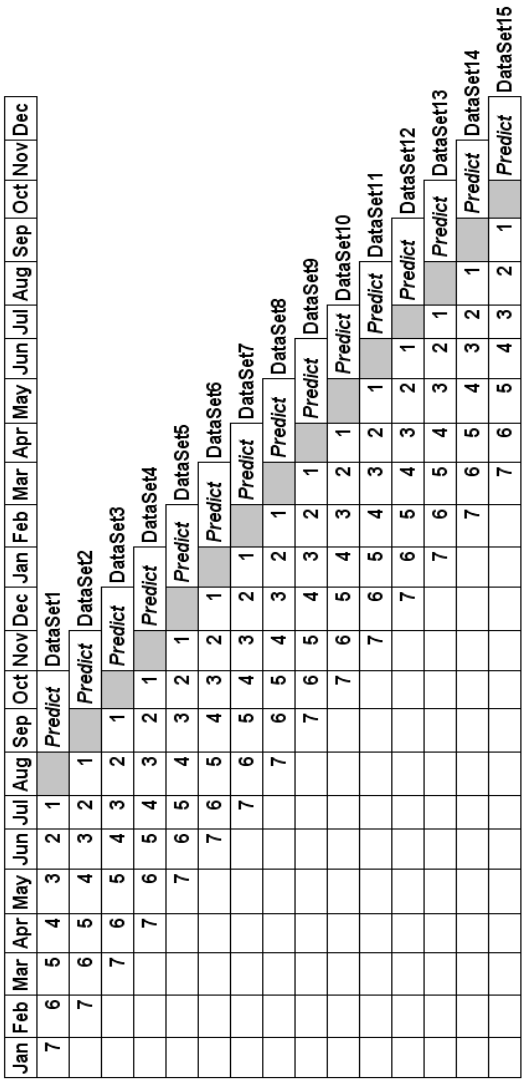
The data in Supreme_pred.sav contain churners and non-churners for January 2003 through December 2004, a total of 24 months. Supreme has decided to use seven months of past transactional data to predict churn two months ahead. That is, there is one month of latency.

Besides the time needed for data collection, the two months' gap also gives Supreme the time to act on the data mining results after the potential churners are identified in the month of latency. This is depicted in Table 5.4 for a two-year period. The numbers in Table 5.4 refer to the number of months counted backwards from the month of latency (shaded).

In this data mining application, the data set Supreme_pred.sav (comprising 10,000 observations or records) is randomly partitioned into an 80% construction data set (filename = Supreme_predc.sav) and a 20% validation data set (filename = Supreme_predv.sav). The purpose of the validation data set is to provide an unbiased estimate of the performance of the potential churn 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 total, construction and validation data sets, the percentage of churners are 23.76%, 23.70% and 24.00%, respectively.

Preliminary data exploration using visualization tools on the total data set (see Figure 5.7) shows that churners (as compared to non-churners) are more likely to be: (1) female customers; (2) customers who do not stay within one kilometre of a Supreme outlet; and (3) customers who have redeemed their loyalty card points in the last six months.

Table 5.4 Month of Latency and Identification of Data Sets



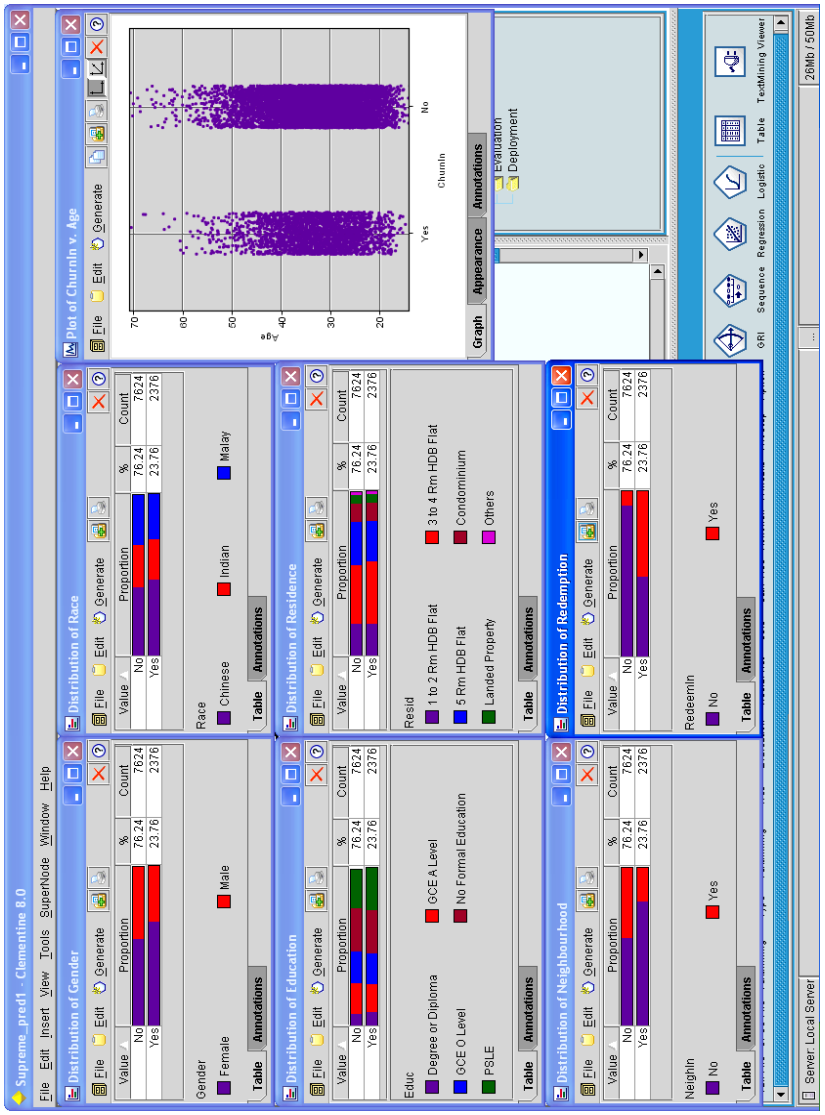


Figure 5.7 Visualisation of Churners and Non-churners

The other demographic variables (i.e., race, age, education level and type of residence) do not seem to show systematic variations between churners and non-churners.

To facilitate churn modelling, new variables (known as derived variables) are created. A derived variable that can help predict churn is the historical churn rate grouped by combinations of levels of selected demographic variables. For this data mining application, the demographic variables selected to compute this historical churn rate are gender, race and education level (e.g., female and Chinese customers who hold diplomas or degrees). The historical churn rate is computed for the 15 data sets shown in Table 5.4. To illustrate, for DataSet1, data from January to July 2003 are used to compute the historical churn rate for each combination of gender, race and education level.

Historical churn rate computations are done using the Aggregate and Derive nodes in SPSS Clementine. The computed rates are then merged back to both the construction and validation data sets.

Another set of derived variables used in the churn model is the set of six new variables created to compare the amount spent in successive months. The six ratios are defined as follows:

1) Ratio1 = (Amtspt1 – Amtspt2)/Amtspt2;
2) Ratio2 = (Amtspt2 – Amtspt3)/Amtspt3;
3) Ratio3 = (Amtspt3 – Amtspt4)/Amtspt4;
4) Ratio4 = (Amtspt4 – Amtspt5)/Amtspt5;
5) Ratio5 = (Amtspt5 – Amtspt6)/Amtspt6; and
6) Ratio6 = (Amtspt6 – Amtspt7)/Amtspt7.

The following predictive modelling tools are used to construct the potential churn models: decision trees (using the C5.0 and CART algorithms), neural networks and logistic regression. The target variable is ChurnIn and the input variables are Gender, Race, Age, Educ, Resid, NeighIn, RedeemIn, DataIn, Amtspt1, Ratio1 to Ratio6 and the historical churn rate. Here, DataIn

can be considered as a trend variable to capture changes over time. The models are constructed on the construction data set and validated on the validation data set. The results are summarised in two figures. Figure 5.8 shows the accuracy rates and Figure 5.9 shows the lift charts (both sets of results for the validation data set).

As can be seen from the overall accuracy rates of the four potential churn models in Figure 5.8, the performance of the models on the validation set is very similar, ranging from a low of 82.15% for the neural network model to a high of 82.95% for the C5.0 decision tree model. Also, the C5.0 decision tree model predicts churners the most accurately (306 out of 480, or 63.75%).

The lift charts in Figure 5.9 indicate that the neural network model is dominated by at least one of the other three models (in terms of lift values and hence, hit rates – see Chapter 4) across the different deciles. Hence, it can be discarded at this assessment stage, especially given its "black box" disadvantage as well (see Chapter 3).

As mentioned earlier, Supreme is interested to identify potential churners so as to take pre-emptive actions to discourage churn and to encourage loyalty. However, with limited resources, Supreme can afford to take pre-emptive actions only on customers with the highest predicted probabilities of churn.

If actions are to be taken only on the top 20% of potential churners, then the logistic regression model is the best model to use given its superior lift value (see Figure 5.9). Anything beyond the top 25% (approximately), the C5.0 decision tree is the best model because its lift value is the highest among the four models after the mid-point of the second and third decile. Given these findings, Supreme has decided to take pre-emptive actions on the top 30% of potential churners. Therefore, the C5.0 decision tree model is the champion model to use.

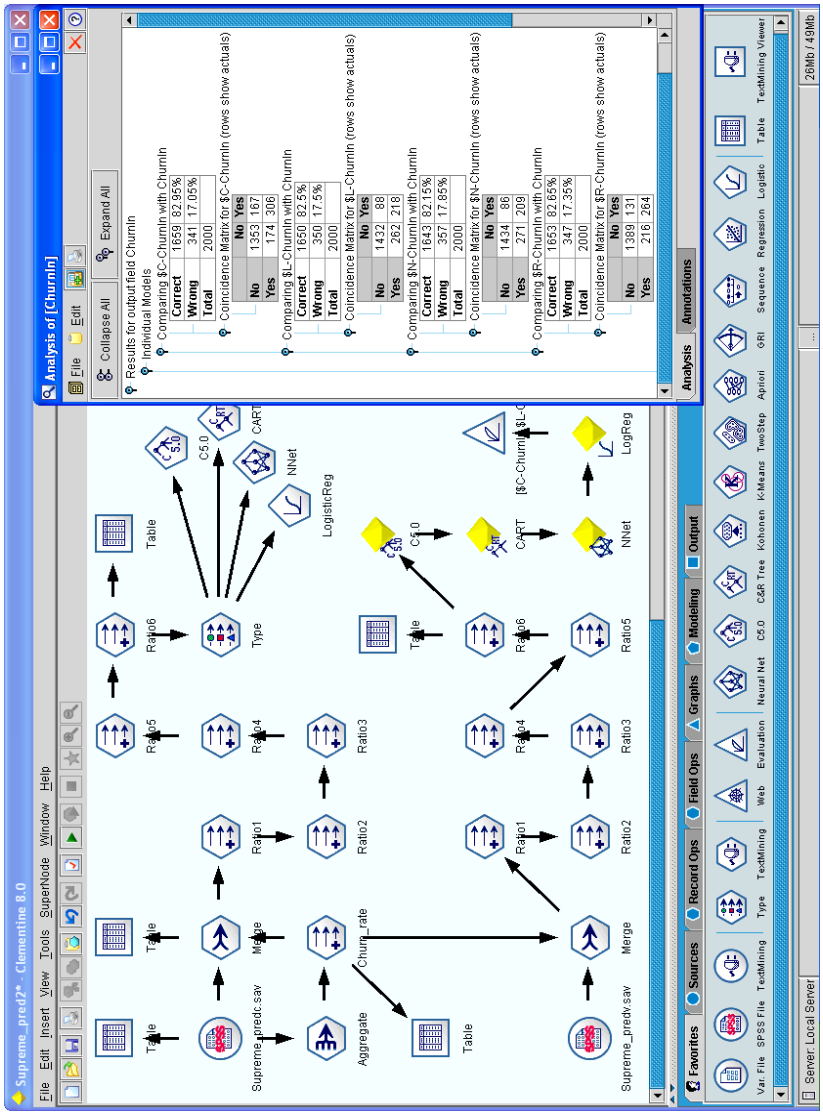


Figure 5.8 Accuracy Rates of Churn Models

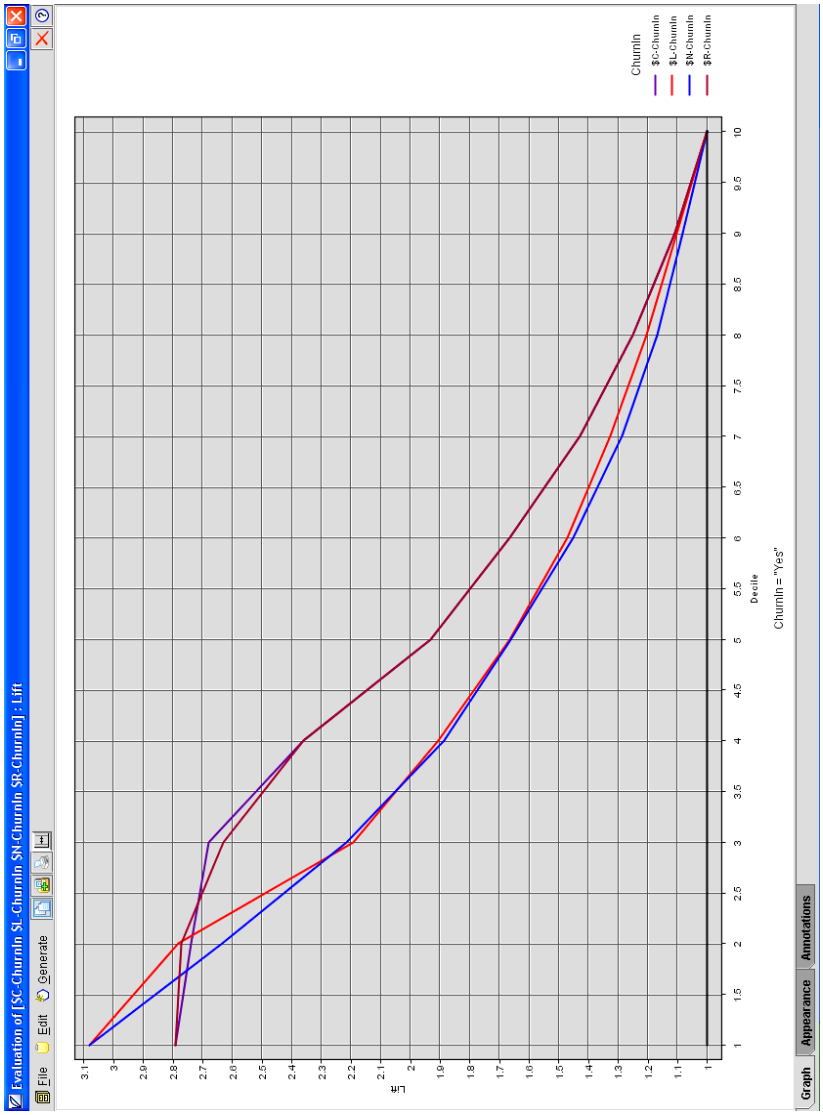


Figure 5.9 Lift Charts of Churn Models

A graphical representation of the decision tree model (using the construction data set) is given in Figure 5.10. It is an excellent way to visualise the predictive modelling results and relationships between the input variables and target variable. Generally, input variables appearing higher up in the decision tree have a stronger association with the target variable and hence in the current application are more important for predicting churn (i.e., in identifying potential churners).

As shown in the decision tree, RedeemIn (indicating whether a customer has redeemed his/her loyalty card points in the last six months) is the most important variable associated with the churn status. Customers who churn are likely to be those who have redeemed their points. Further, for this group of customers, the probability of churn is higher during periods when the historical churn rate is also higher. This pattern is also observed at the last two levels of the decision tree. For customers who have not redeemed their loyalty card points in the last six months (see right side of the decision tree in Figure 5.10), Ratio1 and Ratio2 suggest that churn is associated with progressively decreasing spending at Supreme. Also, at the next level, NeighIn indicates that customers who live within one kilometre of a Supreme outlet are less likely to churn.

With these findings, Supreme is better able to identify customers who are likely to churn in order to take pre-emptive actions to retain them. In particular, Supreme can apply the decision tree model on its customer database and identify those with high predicted probability of churn.

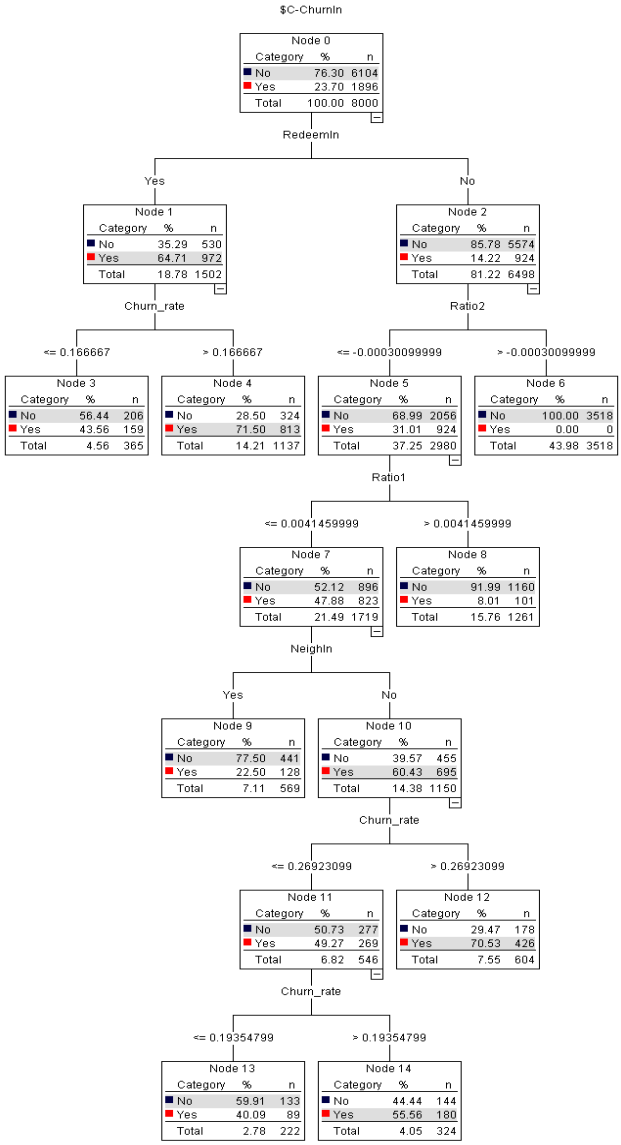


Figure 5.10 Decision Tree Representation of Churn Model

5.7 Concluding Remarks

This chapter does not attempt to illustrate all potential data mining applications in the retail industry. Instead, it aims to link the earlier discussion on data mining to specific applications in particular industries. For example, visualisation and association are used to facilitate baby products promotion, clustering is used to perform customer segmentation, and predictive modelling is used to construct churn models.

The next two chapters extend the discussion in this chapter to include more illustrations and potential applications as well as more sophisticated ones. Chapter 6 focuses on the service industry while Chapter 7 focuses on the manufacturing industry.

There is no doubt that data mining is a very powerful methodology and technology that can be applied in many different commercial and non-commercial contexts. With some imagination and creativity (in addition to the pre-requisites of data mining discussed in Chapter 8), it can go a long way towards enhancing the competitive advantage of organisations.

Chapter References

Anonymous. (1999), "Data mining", *Chain Store Age*, October, p. 42.

Anonymous. (2004a), "Counter measures to retail criminals", *In-Store*, February, p. 29.

Anonymous. (2004b), "Online retail: customer retention", *New Age Media*, March, p. 5.

Bednarz, A. (2004), "It's all about the data for retailers", *Network World*, Vol. 21 No. 20, p. 14.

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.

Fayyad, U. (2004), "Optimizing customer insight", *Intelligent Enterprise*, Vol. 6 No. 8, pp. 22-26, 33.

Fitzgerald, K. (2004), "Grocery cards get an extra scan", *Credit Card Management*, Vol. 16 No. 13, pp. 34-39.

Reid, K. (2003), "Digging into data", *National Petroleum New*, Vol. 95 No. 8, pp. 28-32.

Peterson, K. (2003), "Mining the data at hand", *Chain Store Age*, Vol. 79 No. 6, p. 36.