Chapter 4 Data Mining Issues

4.1 Introduction

Data mining tools can be used for description and visualisation, association and clustering, and predictive modelling. Chapter 2 focuses on description, visualisation, association and clustering while Chapter 3 focuses on predictive modelling and the three most commonly used predictive modelling tools, namely, regression, neural networks and decision trees. In both chapters, theoretical aspects are discussed and practical applications are illustrated. This chapter extends the discussion on predictive modelling by looking at a few important data mining issues.

In particular, Chapter 4 explores the validation of prediction models and highlights the factors affecting the determination of optimal cut-off points. It also discusses measures of prediction effectiveness (i.e., the use of evaluation charts and financial assessments as means to evaluate competing prediction models). Finally, a comprehensive illustration of predictive modelling incorporating tools discussed in Chapter 3 and some of the issues explored in this chapter is presented.

4.2 Model Validation

To facilitate discussion, suppose that a data mining application is developed to predict a non-metric target variable on the basis of several input variables. (The case of a metric target variable will be discussed later). When statistical methods (e.g., logistic regression) are used for predictive modelling, statistical results are available to assess the adequacy of the resulting prediction model. These include test statistics and p-values for the model and input variables and other measures such as R-square. However, statistical results are often not available for artificial intelligence models (e.g., neural networks) and machine learning techniques (e.g., decision trees). Therefore, these tools frequently use non-statistical measures to assess the adequacy of the models. Accuracy rates are the most commonly used non-statistical measure of model adequacy.

The assessment of the adequacy of prediction models are also referred to as the validation (or testing) of models. In particular, the models are validated to ensure that they are adequate (or sufficiently accurate) for use. In model validation, it is important that a validation (or testing) sample is used instead of the construction (or training) sample. The construction sample comprises the data used to construct a prediction model. If model validation is also performed on the construction sample, then the accuracy rates computed will be upward biased. In other words, if the same data are used for model construction as well as model validation, then the computed accuracy rates will be higher than what they would actually be when the model is deployed.

Conceptually, it is best to have a construction data set to construct the prediction model and a validation data set to validate the constructed model. This will lead to the least biased estimates of the accuracy rates of the model when it is used. Having separate construction and validation data sets may not be a major problem in many data mining applications as such applications are likely to involve a lot of data. Hence, the available data can be easily partitioned into two large data sets. The important requirement is for both the data sets to be representative of the population of interest on which the prediction model is to be applied. For example, if a model is to be constructed to predict the acceptance of a particular healthcare service for retired persons, then the construction and validation data sets should reflect the population of retired persons.

If the construction data set does not reflect the population of interest, then the constructed model will not perform well when applied to the population of interest. On the other hand, if the validation data set does not reflect the population of interest (although the construction data set does), then the accuracy rates computed on the validation data set will not reliably estimate the performance of the model when it is eventually applied. Neither will it reliably estimate the accuracy rates of the constructed model.

To apply the above concepts is relatively easy. An organisation needs only to define how the original data set is to be randomly partitioned into a construction and a validation data set. This can be done, for example, by specifying the percentage of data to be used for construction; the remaining will be used for validation. For instance, an organisation can specify that 70% of the available data should be randomly selected for model construction. By default, the remaining (randomly selected) 30% of the data will be used for model validation. It is reasonable to use more data for model construction than for model validation (70%-30% is a good rule-of-thumb). This approach, however, is not data-efficient in that the validation data are used only for validation and hence do not contribute to constructing the model.

To mitigate the above in data mining applications, the n-fold validation method is commonly used. Here, all the data available for predictive modelling are taken to be a single data set. In the first step, this data set is partitioned randomly into equal-size (or approximately equal-size) data sets according to the value of n. For example, if n = 2, then the available data are randomly partitioned into two data sets of equal (or approximately equal) size. In the second step, the prediction model is constructed using (n – 1) of the partitioned data set(s) and then validated on the n^{th} partitioned data set that is not used in the model construction. The validation results (i.e., correct and incorrect predictions or classifications) of the held-out partitioned data set are then noted. This step is repeated until each of the partitioned data sets has had the opportunity to be the held-out data set. In the third and last step, all the validation results are aggregated and the accuracy rates computed for the prediction model.

It is important to differentiate between constructing a prediction model and validating it. In particular, a model can be constructed based on all the available data. However, this model can be validated based using the n-fold validation method where the data are partitioned for validation purposes.

Continuing the discussion with n = 2, the available data will be partitioned into two equal-size (for approximately equal-size) data sets. For simplicity, let these two data sets be called A and B. In the second step, a prediction model will be constructed on A and then validated on B. Next, a prediction model will also be constructed on B and then validated on A. In the third step, the validation results are combined to yield the accuracy rates for the final model that is constructed based on the total of both A and B. The validation accuracy rates estimate the performance of the model when it is used. This method (where n = 2) is also known as the split-half methodology.

At the other extreme is the jack-knife methodology where n = N (where N is the sample size of the total data set). As before, the available data are partitioned into n equal-size data sets. When n = N, this means that each partitioned data set is just one observation. A prediction model will be constructed using all except one (i.e., [n – 1] or [N – 1]) observations. The constructed model is then validated on the held-out observation and the classification result noted. This process is repeated until every observation has had a chance to be the held-out observation. That is, a total of n (= N) number of models will be constructed and validated n (= N) times. As before, the validation results are combined to yield the accuracy rates for the model that is constructed based on the total data set. The validation accuracy rates estimate the performance of the model when it is used.

Between the two extremes of n from 2 to N are a range of possibilities. As a final example of the n-fold validation method, suppose that n = 10. Then, the available data will be randomly partitioned into ten data sets of equal (or approximately equal) size. At each iteration, one of the ten partitioned data sets will be held-out as the validation sample for a model constructed based on the other nine partitioned data sets. Since n = 10, there will be ten such iterations until each partitioned data set has had the opportunity to be the held-out data set. The classification results are observed for each iteration and then aggregated to give the model accuracy rates.

The above discussion focuses on accuracy rates, which are relevant for non-metric target variables. Here, an error is a misclassification. For metric target variables, however, an error is the difference between the actual and predicted value of the target variable. That is, $e = y - \hat{y}$. Several measures of model adequacy (or accuracy) can be constructed based on this. Two common ones are the following:

$$\text{Mean Absolute Deviation (MAD)} = (\sum|e|)/N = (\sum|y - \hat{y}|)/N$$

$$\text{Mean Squared Error (MSE)} = (\sum e^2)/N = (\sum(y - \hat{y})^2)/N$$

where N is the total sample size (i.e., sample size of the total data set) and the summation is made over all the (n) partitioned data sets.

To illustrate, suppose that a prediction model is to be constructed to predict expenditure on healthcare services (a metric target variable) on the basis of several input variables. Suppose further that a 10-fold validation method is to be used to compute MAD for the prediction model that is constructed using all the available data. In this case, the available data will be randomly partitioned into ten data sets of equal (or approximately equal) size. For each iteration, one of the ten partitioned data sets will be held-out as the validation sample for a model constructed based on the other nine partitioned data sets. Since n = 10, there will be ten such iterations until each partitioned data set has had the opportunity to be the held-out data set. The value of ∑|e| is computed for each iteration. Finally, all ten ∑|e|'s are aggregated (i.e., summed up) and the total is divided by N to give the model's accuracy rate (in this case, MAD). Generally, the smaller the MAD or MSE, the more accurate the prediction model is.

Finally, model validation can also be used as a means to reduce over-fitting a model to the data in neural networks and decision trees. Over-fitting occurs when the model "memorises" the unique patterns in the construction data set so that the accuracy rates computed based on the construction data set (known as in-sample accuracy rates) are very high. However, because unique patterns in the construction data set are unlikely to exist in other data sets, the model will perform poorly when it is used. In particular, when over-fitting occurs, the accuracy rates computed based on a validation data set (known as hold-out accuracy rates) will be much lower than the in-sample accuracy rates. Therefore, to reduce the over-fitting problem, it is common during the training of neural networks and construction of decision trees to compare the in-sample and hold-out accuracy rates. For example, in-sample accuracy rates may increase as a decision tree grows. However, the hold-out accuracy rates will decrease if there is over-fitting of the decision tree to the construction data set. Hence, such a situation may activate a stopping rule for the decision tree. The same applies to the training of neural networks. Such stopping rules reduce over-fitting.

4.3 Optimal Cut-off Points

The discussion in this section is applicable only to prediction models with non-metric target variables. In particular, this section examines the factors affecting the optimal cut-off point for such models. To crystallise the concepts, recall the following logistic regression model:

$$\text{Predicted Probability of Event} = 1/[1 + e^{-(b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p)}]$$

where X_1 to X_p are the input variables. As discussed in Chapter 3, the cut-off point for a logistic regression model can be expressed as either a cut-off probability or a cut-off theoretical index. For simplicity, let Z denote the theoretical index without the constant term b_0 and assume that the target variable has only two categories, L (for low) and H (for high).

Since there are two categories of the target variable, there must be two samples (denoted as samples L and H) corresponding to the two categories. Together, these samples form the data set for model construction and validation. The theoretical index (without the constant term) can be expressed as follows:

$$Z = b_1X_1 + b_2X_2 + \dots + b_pX_p$$

Assume that substituting the mean values of the input variables in samples L and H gives Z_L and Z_H, respectively. Then, without additional information, the optimal cut-off point (C) for the prediction model can be computed as:

$$C = (Z_L + Z_H)/2$$

The cut-off point C may be considered optimal in the sense that it minimises the total misclassifications by (roughly speaking) passing through the middle of the two categories L and H.

Assume that $Z_L < Z_H$. Then, if an observation has $Z \geq C$, the observation will be classified as H. Conversely, if an observation has $Z < C$, the observation will be classified as L. This is shown in Figure 4.1.

Generally, a cut-off point C computed in this manner tends to lead to approximately equal accuracy rates for categories L and H. As can be observed from Figure 4.1, if the cut-off point C is moved to the right, the accuracy rate for category L will increase and that for category H will decrease as more observations are now predicted as category L. Conversely, if the cut-off point C is moved to the left, the accuracy rate for category L will decrease and that for category H will increase as more observations are now predicted as category H. Hence, accuracy rates can be affected by adjusting the cut-off point. This also means that for a particular relative accuracy rate desired (e.g., the accuracy rate for category L should be about twice that for category H), it may be possible to adjust C to give this relative accuracy rate (or a relative accuracy rate close to it).

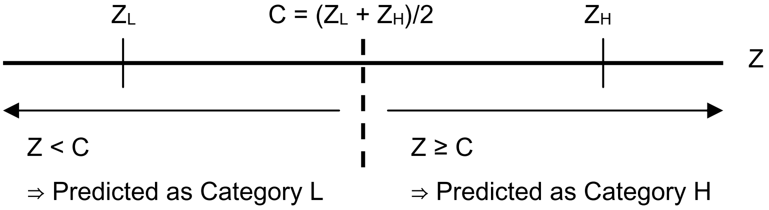


Figure 4.1 Computation of Optimal Cut-off Point

There are two important factors that affect the optimal cut-off point. The first factor is the prior probabilities of the event and non-event occurring. In order to minimise total misclassifications, the category that occurs more frequently (i.e., the category that has the higher prior probability) should be predicted more accurately. Suppose that risk status (i.e., fraud versus non-fraud) is to be predicted for credit card transactions and the prior probability of fraud is 1% (i.e., the prior probability of non-fraud is 99%). (These prior probabilities can be estimated based on a large random sample of credit card transactions). Suppose further that there are two competing prediction models. Model A has accuracy rates of 90% for both fraud and non-fraud, and hence an overall accuracy rate of also 90%. On the other hand, Model B has accuracy rates of 80% for fraud and 91% for non-fraud, giving an (unweighted) overall accuracy rate of 85.5% (i.e., [80 + 91]/2), which is lower than that for Model A. (The weighted overall accuracy rate can be computed by using the relative proportions of fraud and non-fraud transactions in the data set as weights. Alternatively, it can be computed by dividing the total number of correct classifications by the total number of observations in the data set. This is done in the classification tables presented in Chapter 3).

Assume that there are 100,000 credit card transactions on average every month. Given that the prior probabilities of fraudulent and non-fraudulent transactions are 1% and 99%, respectively, then out of the 100,000 credit card transactions, 1,000 can be expected to be fraudulent and 99,000 non-fraudulent. Based on this, the expected performance of the two models can be computed as shown in Table 4.1. As can be seen, Model A has 10,000 misclassifications while Model B has only 9,110 misclassifications. This may seem surprising as Model A predicts fraud better than Model B by a margin of 10% (i.e., 90 – 80) but predicts non-fraud worse than Model B by a margin of only 1% (90 – 91). The reason for the results in Table 4.1 is that non-fraudulent transactions occur a lot more frequently than fraudulent transactions. Therefore, a small decrease in the accuracy rate of predicting non-fraudulent transactions translates into many misclassifications.

Table 4.1 Number of Misclassifications for Models A and B

	Number of Misclassifications		
Prediction Model	Fraud	Non-fraud	**Total**
Model A	100 (0.1 x 1000)	9900 (0.1 x 99000)	10,000
Model B	200 (0.2 x 1000)	8910 (0.09 x 99000)	9,110

From the above, it can be seen that in order to minimise misclassifications, the category of the target variable that occurs more frequently (i.e., a category with a higher prior probability) should be predicted more accurately. This can be done by appropriately adjusting the cut-off point as shown in the following formula:

$$C = (Z_L + Z_H)/2 + \ln(q_L/q_H)$$

where q_i is the prior probability of category i. In the formula above, the categories L and H are defined such that $Z_L < Z_H$.

To illustrate, suppose that for the logistic regression model mentioned earlier, $Z_L = -1.50$ and $Z_H = 1.50$. Also, the prior probabilities for L and H are $q_L = 0.75$ and $q_H = 0.25$, respectively. Then, without considering prior probabilities, the optimal cut-off point C is:

$$C = (Z_L + Z_H)/2 = (-1.50 + 1.50)/2 = 0$$

When prior probabilities are considered, the new optimal cut-off point C* is:

$$\begin{aligned} C^* &= (Z_L + Z_H)/2 + \ln(q_L/q_H) \\ &= (-1.50 + 1.50)/2 + \ln(0.75/0.25) \\ &= 1.10 \text{ (rounded to 2 decimal places)} \end{aligned}$$

That is, the cut-off point is now moved to the right. With reference to Figure 4.1, when the new optimal cut-off point is moved to the right of the original cut-off point, the accuracy rate of predicting category L (the more commonly occurring category) will now be higher at the expense of a decrease in the accuracy rate of predicting category H. This is shown in Figure 4.2.

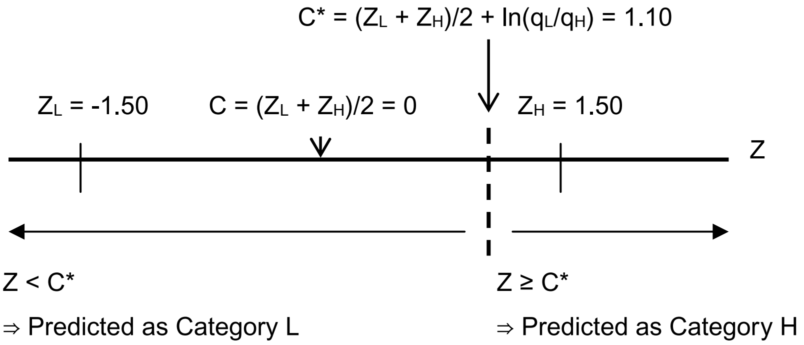

C* = (ZL + ZH)/2 + ln(qL/qH) = 1.10
ZL = -1.50
C = (ZL + ZH)/2 = 0
ZH = 1.50
Z
Z < C*
⇒ Predicted as Category L
Z ≥ C*
⇒ Predicted as Category H

Figure 4.2 Optimal Cut-off Point – Adjusted for Prior Probabilities

The second factor that affects the optimal cut-off point is the relative misclassification costs. This refers to the cost of misclassifying category L as H and the cost of misclassifying H as L. Sometimes, the two misclassification costs may be the same but this is not always the case. As will be seen later, it is the relative misclassification cost that is important in affecting the optimal cut-off point and not the absolute misclassification costs.

Consider, for example, the approval of consumer loans. Suppose that a prediction model that predicts the default risk of payment (say, low-risk versus high-risk) is used in the loan approval process. If a low-risk potential customer is misclassified as a high-risk potential customer, then the loan-granting organisation would not approve the consumer loan. In this case, the misclassification cost (of low-risk incorrectly classified as high-risk) is the interest income forgone (ignoring the effects of ill-will). On the other hand, if a high-risk potential customer is misclassified as a low-risk potential customer, then the loan-granting organisation would approve the consumer loan. In this case, the misclassification cost (of high-risk incorrectly classified as low-risk) is the loan and interest that are not collectible (ignoring the effects of legal expenses and whatever amount that can be recovered). Thus, misclassifying a high-risk potential customer is more costly than misclassifying a low-risk potential customer.

If misclassification costs are equal, then the optimal cut-off point is not affected. Assuming a two-category classification, equal misclassification costs means that the relative misclassification cost is 1:1. However, if misclassification costs are not equal (i.e., when the relative misclassification cost is not 1:1), then the optimal cut-off point should be adjusted so that the category whose misclassification is more costly is more accurately predicted. Only then can the total misclassification cost of using the prediction model be minimised. This total misclassification cost (TMC) can be computed as follows:

$$\text{TMC} = q_L\text{*Prob}(H|L)\text{*Cost}(H|L) + q_H\text{*Prob}(L|H)\text{*Cost}(L|H)$$

where Prob(i|j) is the probability of misclassifying category j as category i and Cost(i|j) is the misclassification cost of incorrectly classifying category j as category i. As defined previously, q_i is the prior probability of category i occurring. Therefore, given the above, q_L*Prob(H|L)*Cost(H|L) and q_H* Prob(L|H)*Cost(L|H) are the misclassification costs of incorrectly classifying L as H and H as L, respectively. Intuitively, q_L*Prob(H|L)*Cost(H|L) applies the misclassification cost of L as H (i.e., Cost[H|L]) on a L occurrence (i.e., q_L) that is incorrectly classified as a H occurrence (i.e., Prob[H|L]). A similar interpretation can be given to q_H*Prob(L|H)*Cost(L|H).

To minimise the total misclassification cost of using the prediction model, the optimal cut-off point can be computed as:

$$C = (Z_L + Z_H)/2 + \ln\{[q_L*\text{Cost}(H|L)]/[q_H*\text{Cost}(L|H)]\}$$

where categories L and H are defined such that $Z_L < Z_H$. The formula shows that it is the relative misclassification cost (i.e., Cost(H|L)/Cost(L|H)) and not the absolute misclassification costs that affect the optimal cut-off point. That is, the two misclassification costs can be increased or decreased by the same percentage and the optimal cut-off point will remain the same. The same comment, however, cannot be made about the prior probabilities (i.e., q_L and q_H) as they have to sum up to 1. That is, increasing or decreasing the prior probabilities by the same percentage will violate the condition that probabilities sum up to 1. The effect of the adjustment is to increase the accuracy rate of the category whose misclassification is relatively more costly.

Continuing with the logistic regression example, suppose that Cost(H|L) = $100 and Cost(L|H)) = $1000. Then, incorporating misclassification costs, the optimal cut-off point (denoted C** in this example) is:

$$C^{**} = (-1.50 + 1.50)/2 + \ln(0.75*100/0.25*1000)$$
$$= 0 + \ln(0.30) = -1.20$$

With reference to the original optimal cut-off point of 0 (which does not consider prior probabilities and misclassification costs), the new optimal cut-off point is now moved to the left. This will increase the accuracy rate of predicting category H (the more costly category to misclassify) at the expense of a decrease in the accuracy rate of predicting category L. This is shown in Figure 4.3.

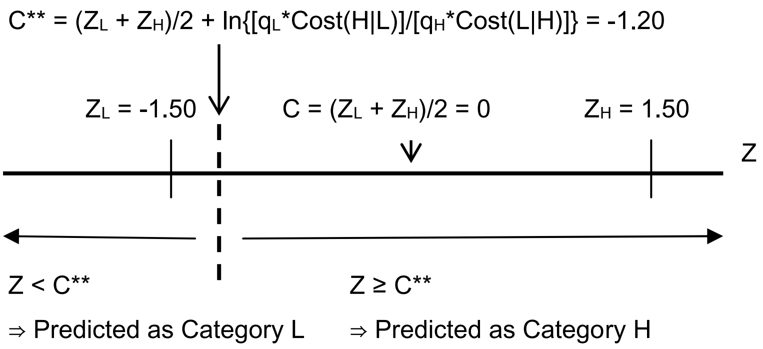

$C^{**} = (Z_L + Z_H)/2 + \ln\{[q_L{*}Cost(H|L)]/[q_H{*}Cost(L|H)]\} = -1.20$
$Z_L = -1.50$
$C = (Z_L + Z_H)/2 = 0$
$Z_H = 1.50$
Z
$Z < C^{**}$
⇒ Predicted as Category L
$Z \geq C^{**}$
⇒ Predicted as Category H

Figure 4.3 Optimal Cut-off Point – Adjusted for Misclassification Costs

The discussion so far focuses on logistic regression. The concepts related to optimal cut-off points, however, can be extended to neural networks and decision trees. Essentially, the idea is to adjust the optimal cut-off points so as to reduce either the number of misclassifications or the total misclassification cost of using the model. The models themselves are not adjusted (e.g., there is no attempt to adjust the logistic regression model coefficients) – only the cut-off point is.

In a data mining application, prior probabilities may be available (say, from previous research). Otherwise, they can usually be estimated from a random sample. Misclassification costs, however, are more difficult to estimate as they may require a fair amount of subjective judgement to be made. As mentioned earlier, it is the relative misclassification costs and not absolute misclassification costs that are important in determining the optimal cut-off point.

More information about prior probabilities and relative misclassification costs can be found be in Koh (1992) and Afifi and Clark (1996).

4.4 Evaluation Charts

As shown in Figure 1.1 in Chapter 1, the modelling stage of the data mining methodology includes the assessment of results and the identification of the final model. Generally, assessment refers to a common framework to compare models and predictions. Factors that may confound the comparison, such as the actual validation data set itself, have to be kept constant.

For any data mining application, there may be more than one data mining tool that can be used. For predictive modelling with a non-metric target variable for example, logistic regression, neural network or decision tree models can be constructed. If two or more these models give acceptable results, then there is a need to compare the performance of these models to select the final model that should be deployed. One common benchmark for comparison is accuracy rates. The most accurate model can be selected as the final model. (The various types of accuracy rates will be discussed later).

Another common benchmark for comparison is evaluation charts. This benchmark is only relevant for the prediction of non-metric target variables. (For metric target variables, evaluation charts can still be plotted if the target variables can be broken down into ranges or categories. One possibility is to dichotomous the values of a metric target variable into values < mean and values ≥ mean). If the target variable has more than two categories, one category of interest can be identified and the remaining categories can be combined in plotting an evaluation chart. For simplicity, this section assumes a non-metric target variable with only two categories: event and non-event, where event is the occurrence of interest.

Suppose that a prediction model has been constructed and is to be validated on a validation data set. In the validation data set, the actual status (i.e., event or non-event) for each observation is known. There are several kinds of evaluation charts that can be plotted. The most popular ones are response charts, gains charts and lift charts. To construct these charts, the model is first applied to the validation data set to generate predictions (i.e., classifications) on the data. These give the predicted status of each observation. The process of generating predictions is commonly referred to as scoring the data. For each observation scored, in addition to the non-metric prediction (i.e., event or non-event), there are usually other related measures that can be computed (e.g., the predicted probability of an event, the computed theoretical index or the confidence of the prediction).

To illustrate, consider a logistic regression model with the predicted probability of the event occurring and the predicted status scored for each observation. In the first step, the predicted probabilities of the event occurring are sorted in descending order. This means that observations that are higher up in the list have greater predicted probabilities of the event occurring. If the prediction model is acceptable, then observations that are higher up in the sorted list are more likely to be actual events. (In this discussion, it is important to differentiate between actual status and the predicted status of event and non-event). The converse is true for observations lower down in the sorted list. That is, observations lower down in the list should have lower predicted probabilities that they are events. They are also less likely to be actual events.

In the next step, the observations as ordered in the list are then grouped into equal-size (or approximately equal-size) segments. While there are no fixed rules on the number of segments that is appropriate, it is common to group the observations into deciles (i.e., ten segments). For a model with high predictive power, actual events are expected to be concentrated in the higher segments (where observations are predicted to have higher probabilities of being events).

To understand the computations involved, let the sample size of the validation data set be N and let the scored and sorted observations be grouped into g segments. For the i^{th} segment, let the sample size be denoted by n_i. If the sorting and grouping of scored observations form segments of equal size, then n_i is the same for every segment. Further, let A be the number of actual events in the validation data set and A_i the number of actual events in segment i (i.e., $A = \sum A_i$, summed over all the i segments). If the actual events are randomly distributed in the validation data set, then for any segment i, the number of actual events in that segment is expected to be $E_i = [(A/N) \times n_i]$. Based on the above, the following three measures can be computed:

1) Response (%) in segment i
= (number of actual events in segment i)/(number of observations in segment i)
= $(A_i/n_i) \times 100\%$

2) Gains (%) in segment i
= (number of actual events in segment i)/(number of actual events in the validation data set)
= $(A_i/A) \times 100\%$

3) Lift value for segment i
= (response in segment i)/(response in the validation data set)
= $(A_i/n_i)/(A/N)$

Since the predicted probabilities of an event are sorted in descending order before they are grouped into segments, the three measures listed above are expected to decrease from the first to the last segment. Also, a good prediction model will show high values for the measures for the first few segments and low values for the last few segments. Generally, the term “response” refers to an event of interest. In many customer relationship management applications, the target variable may actually be some kind of a response (say, customer response to direct marketing campaigns or to cross-selling or up-selling efforts).

The computational formulas listed above are based on the concept of hit rates, which is quite different from the concept of accuracy rates. In predictive modelling, a hit can be defined as the actual occurrence of an event. Consider Table 4.2.

Table 4.2 Illustrative Classification Table

	Predicted Status		
Actual Status	Event	Non-event	**Total**
Event	A	B	(A + B)
Non-event	C	D	(C + D)
Total	(A + C)	(B + D)	(A + B + C + D)

The computation of accuracy rates uses the actual status of the target variable as a base. For example, the accuracy rates for event and non-event are A/(A + B) and D/(C + D), respectively. On the other hand, the computation of hit rates uses the predicted status as a base and focuses on the actual occurrence of the event. Hence, the hit rates for (predicted) event and non-event are A/(A + C) and B/(B + D), respectively. For a good prediction model, the former should be high and the latter low. (The “hit” rate D/(B + D) is also useful as it looks at the extent to which predicted non-events are actually non-events. This rate should also be high for a good prediction model). While accuracy rates answer the question “to what extent are the actual events and non-events predicted correctly?”, hit rates answer the question "given the model predictions of events, to what extent do events actually occur?". Both accuracy rates and hit rates are useful in assessing the performance of prediction models.

Given the above, the following interpretations can be made:

1) Response in segment i indicates the hit rate in segment i. As the predicted probabilities of the event are already sorted in descending order, hit rates are expected to be higher in segments that are higher up in the sorted list.
2) Gains in segment i indicates the percentage of hits captured in segment i.
3) Finally, the lift value measures how much better the hit rate in segment i is compared to the random hit rate (i.e., the hit rate of a segment of randomly selected observations).

In a way, all the measures indicate prediction effectiveness. A higher value is associated with greater effectiveness.

Table 4.3 illustrates the computation of response, gains and lift value. As expected, the measures decrease from the first segment to the last segment (because the segments are already sorted in descending order of the predicted probabilities of the event). There is also a strong similarity between gains and lift value. This will be the case as long as the segments are of equal (or approximately equal) size.

Evaluation charts can be plotted for each of the measures. In particular, response charts have the response in segment i plotted on the y-axis. Similarly, gains and lift charts have gains and lift value in segment i plotted on the y-axis, respectively. For all the evaluation charts, segments are plotted on the x-axis. For the current illustration, segments are grouped into deciles. However, segments can also be grouped into percentiles (i.e., the list of descending predicted probabilities can be grouped into 100 segments), quantiles (five segments) or any other number of segments deemed appropriate for plotting evaluation charts.

Table 4.3 Computation of Response, Gains and Lift Value

Segment		Number	Response (%)	Gains	Lift
Number	Size	of Events		(%)	Value
1	100	80	80.00%	26.68%*	2.68*
2	100	70	70.00%	23.33%	2.33
3	100	60	60.00%	20.00%	2.00
4	100	40	40.00%	13.33%	1.33
5	100	25	25.00%	8.33%	0.83
6	100	15	15.00%	5.00%	0.50
7	100	10	10.00%	3.33%	0.33
8	100	0	0.00%	0.00%	0.00
9	100	0	0.00%	0.00%	0.00
10	100	0	0.00%	0.00%	0.00
Total (Average)	1000	300	(30.00%)**	(10.00%)**	(1.00)**

* Rounded up.

** Related to the baseline model (to be discussed later).

Evaluation charts can also be cumulative or non-cumulative. Cumulative charts are plotted based on cumulative measures. These are measures computed by aggregating the current segment and all earlier/higher segments. Table 4.4 illustrates the cumulative lift value. Both cumulative and non-cumulative lift charts usually show a downward sloping curve (reflecting higher lift values for earlier segments). An erratic curve usually suggests problems with the data or prediction model as it means that segments with lower predicted probabilities of the event can predict the event better than segments with higher predicted probabilities of the event. This is counter-intuitive. This, however, can occur if too many segments are specified in the evaluation chart. Generally, cumulative evaluation charts involve averaging across segments and hence are more gradual compared to non-cumulative charts.

Table 4.4 Computation of Cumulative Lift Value

Segment		Number	Cumulative Number	Cumulative
Number	Size	of Events	of Events	Lift Value
1	100	80	80	2.68*
2	100	70	150	2.50
3	100	60	210	2.33
4	100	40	250	2.08
5	100	25	275	1.83
6	100	15	290	1.61
7	100	10	300	1.43
8	100	0	300	1.25
9	100	0	300	1.11
10	100	0	300	1.00
Total	1000	300	300	1.00
* Rounded up.				

Various benchmarks can be incorporated into an evaluation chart. The most common benchmark is the baseline model. It represents the response, gains or lift value if the observations in each segment are selected randomly. For example, in Table 4.3, there are 300 events out of 1000 observations. Therefore, if observations are selected randomly in each segment, then 30% of the segment can be expected to comprise events (or responses). The baseline model is reflected as a horizontal line anchored at the response, gains or lift value of a “random” segment (see the Total (Average) row in Table 4.3).

An exact model can also be incorporated into an evaluation chart. This model represents a “perfect” model where actual occurrences of the event are associated with the highest predicted probabilities of the event. The exact model in Table 4.3 will have 100 events in each of the first three segments. Graphically, the exact model is usually represented by very high values, followed by sharp drops to very low values.

Based on Table 4.4, a cumulative lift chart with a baseline model is plotted in Figure 4.4. (The SPSS statistics software is used for this purpose).

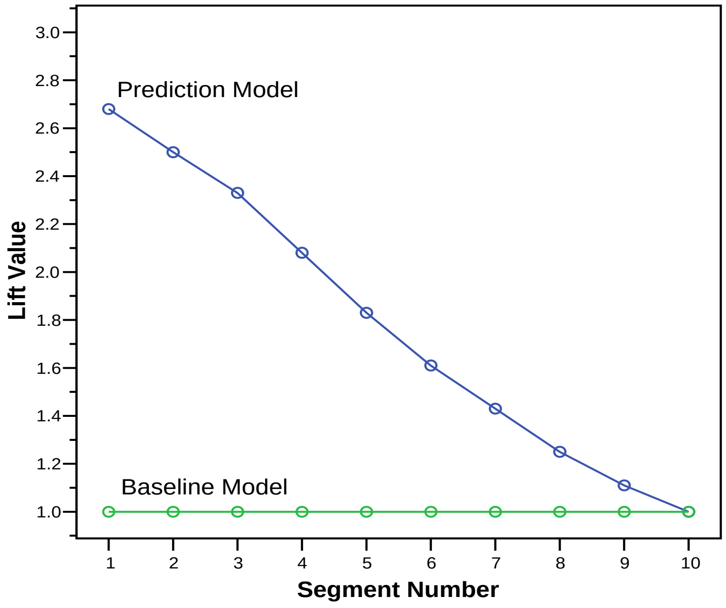


Figure 4.4 Cumulative Lift Chart with Baseline Model

As expected, since the baseline model randomly selects observations, its lift value and cumulative lift value will be 1 (which indicates random distribution of the events among segments). Generally, the higher the lift chart for a prediction model, the better its prediction effectiveness (i.e., hit rates). Ideally, the lift value for the first few segments should be as high as possible. The upper bound (which is the lift value for the first or first few segments of the exact model) is N/A (derived from $(A_i/n_i)/(A/N) = 1/(A/N) = N/A$). This derivation assumes that $A > n_i$ so that all the observations in the first few segments of the exact model are events (implying that $A_i = n_i$ and therefore $A_i/n_i = 1$); otherwise, the upper bound has to be computed using the lift value formula.

Evaluation charts similar to Figure 4.4 can also be plotted with response or gains as the y-axis. Generally, evaluation charts are a good means to compare the performance (i.e., prediction effectiveness) of different prediction models. The best or champion model is the one that dominates the others in terms of having the highest evaluation chart. Usually, only the first few segments (which give the highest predicted probabilities of the event) are of interest. Hence, evaluation charts are usually assessed from the left side of the chart and greater weightage is given to the first few segments (or to a particular portion of the validation data specified in the data mining application [e.g., the first 20% of the data]). This is especially the case when evaluation charts cross. In such a case, no chart is the highest for all the segments. More discussion of evaluation charts will be presented in section 4.6 below.

4.5 Financial Assessment of Prediction Models

Data mining applications are primarily developed and used in the commercial world. Hence, it is not surprising that the assessment of models frequently involves financial aspects in addition to performance measures such as accuracy and hit rates. Financial assessment can also be conducted using evaluation charts and tables.

In addition to the response, gains and lift charts discussed in the previous section, evaluation charts can also take the form of profit charts and ROI (return on investment) charts. As before, these charts are applicable only for non-metric target variable prediction (or metric target variable prediction where the target variable values are broken down into ranges or categories). Both profit and ROI assume a particular business action that leads to cost and revenue. Profit is defined as revenue less cost while ROI is defined as the ratio of profit to cost. Thus, while profit measures absolute return, ROI measures relative return.

Suppose that MailPurchase (the mail order company that has been used in previous illustrations) is planning a mail campaign to promote a particular product. The cost of printing and mailing a brochure to an existing customer is $4. If the customer responds to the mailing campaign and buys the promoted product, the revenue generated is $10. Suppose further that to target the existing customers better, MailPurchase has constructed a prediction model to classify existing customers as potential purchasers or non-purchasers. For this model, the target variable is purchase and non-purchase, where purchase is the response or event of interest. With a cost of $4 and revenue of $10, if a campaign brochure is printed and mailed to an existing customer and this customer does not purchase the promoted product, then MailPurchase would suffer a loss of $4 (which is the cost of printing and mailing the brochure). On the other hand, if the customer purchases the product, then MailPurchase would make a profit of $6 (i.e., $10 - $4). Extending this to the illustration in Table 4.4, a cumulative profit table can be constructed as shown in Table 4.5.

The cumulative profit in Table 4.5 is only the expected cumulative profit and not the actual one; it is based on the validation data set. It is a means to assess the expected performance of the model when it is actually used. With a constant unit cost of $4 per printing and mailing of the campaign brochure and a segment size of 100, the cumulative cost increases at a constant rate of $400 for each additional segment. The cumulative revenue is the product of the cumulative number of events (i.e., responses) and the revenue per unit response of $10. Cumulative profit is just the difference between cumulative revenue and cumulative cost.

Table 4.5 Expected Profit for the Mailing Campaign

Segment		Cumulative Number of Events	Cumulative Cost*	Cumulative Revenue**	Cumulative Profit***
Number	Size				
1	100	80	$400	$800	$400
2	100	150	$800	$1500	$700
3	100	210	$1200	$2100	$900
4	100	250	$1600	$2500	$900
5	100	275	$2000	$2750	$750
6	100	290	$2400	$2900	$500
7	100	300	$2800	$3000	$200
8	100	300	$3200	$3000	– $200
9	100	300	$3600	$3000	– $600
10	100	300	$4000	$3000	– $1000
Total	1000	300	$4000	$3000	– $1000

* Cumulative Cost = $4 x Cumulative Segment Size

** Cumulative Revenue = $10 x Cumulative Number of Events

*** Cumulative Profit = Cumulative Revenue – Cumulative Cost

As shown in Table 4.5, cumulative profit increases steadily from \$400 until it reaches a maximum of \$900 when either 300 or 400 campaign brochures are printed and mailed to the existing customers. This means that to maximise its profit, MailPurchase should send out either 300 or 400 campaign brochures. Both courses of action lead to an expected cumulative profit of \$900. However, MailPurchase may not be indifferent to these two courses of action. Assuming that only these two courses of actions are considered, if MailPurchase wishes to maximise market share of the promoted product or introduce the product to as many consumers as possible (while attempting to maximise profit), then printing and mailing 400 campaign brochures and getting 250 expected responses (i.e., purchases) is the preferred course of action. On the other hand, if MailPurchase faces an immediate budget constraint on its printing and mailing expenditure, then printing and mailing 300 campaign brochures is the preferred course of action.

A more precise table can be constructed with a larger number of segments (and a corresponding smaller segment size). This may, for example, indicate that the optimal course of action to maximise profit is to print and mail 325 campaign brochures, giving a profit of more than \$900.

Table 4.5 also shows that after the fourth segment, cumulative profit decreases and eventually becomes a cumulative loss. A break-even point can be defined as the point of zero profit (i.e., where cumulative profit exactly equals cumulative cost). This occurs between mailing 700 and 800 campaign brochures. As before, a larger number of segments (and a corresponding smaller segment size) can give a more precise break-even point. The break-even point can be a preferred course of action if MailPurchase wishes to have as great a market share as possible without incurring a loss.

Table 4.5 can be plotted as a cumulative profit chart. This is shown in Figure 4.5 using the SPSS statistics software. The baseline model has a loss of $100 ($300 – $400) for each segment.

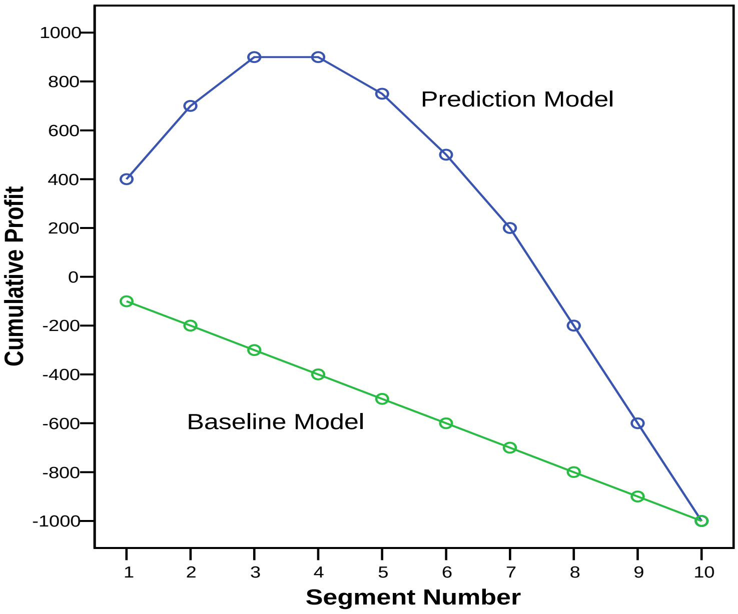


Figure 4.5 Cumulative Profit Chart with Baseline Model

Finally, sensitivity analysis can be performed to assess the impact of selected variables of interest on the outcome by varying the variables of interest (usually one at a time) and observing the resultant changes in the outcome. In Table 4.5, the outcome is cumulative profit and the variable of interest may be, say, the cost of printing the brochure. Assume that different brochures with the same promotional content may be printed, depending on the colour schemes, paper quality, brochure size … etc. This may mean that the per unit cost of printing and mailing each brochure may be $3.50, $4 or $4.50. These possibilities can be used to construct three different cumulative profit tables or charts. Examination of these tables or charts can help MailPurchase assess the effects of different scenarios and decisions and hence make a better decision on its marketing campaign. Sensitivity analysis results can also indicate risk (e.g., the results may show that a small increase in per unit cost may have a very adverse impact on the cumulative profit).

4.6 Illustration of Predictive Modelling – A Revisit

To illustrate some of the data mining issues discussed in this chapter, assume that MailPurchase wishes to extend its predictive modelling efforts in Chapter 3, where the following data are captured:

1) Status: whether the customer has purchased a promoted product in any of the quarterly marketing campaigns last year;
2) Expend: average monthly expenditure on the company's products last year;
3) Numpur: average number of purchases per quarter last year;
4) Age: age of customer as at 1 January last year;
5) Gender: gender of customer;
6) Income: annual income of customer as at 1 January last year (in $'000);
7) Race: race of customer;
8) Marital: marital status of customer as at 1 January last year; and
9) Member: whether the customer is a member of the loyalty card programme last year.

(More details are given in section 2.2.1 of Chapter 2).

To develop the next marketing campaign, MailPurchase is interested to target only existing customers with a high probability of purchase. Hence, it is interested to classify existing customers as likely purchasers or non-purchasers. In this predictive modelling application, the target variable is "status" and the input variables comprise both purchasing patterns (namely, expend and numpur) and demographic characteristics (namely, age, gender, income, race, marital and member).

For this application, MailPurchase has decided to use SPSS Clementine. As shown in Chapter 3, three prediction models can be constructed using the logistic regression, neural network and decision tree nodes. To improve on these analyses, MailPurchase has decided to partition the data into a construction data set (comprising 70% of the database selected randomly) and a validation data set (comprising the remaining 30%). To assess the performance of the three prediction models and to identify a "champion" or best model, MailPurchase has decided to examine their accuracy rates, lift charts and profit charts.

The Clementine data mining stream is shown in Figure 4.6. As can be seen, the available data are now randomly partitioned into a 70% construction data set (with filename MailPurchase_70%.sav) and a 30% validation data set (with filename MailPurchase_30%.sav). Logistic regression, neural network and decision tree models are constructed using the construction data set (see top half of Figure 4.6). The constructed models (represented by the yellow diamond icons) are then bought into the data mining stream to compare their accuracy rates, lift charts and profit charts. A per unit cost of $4 and per unit revenue of $10 (for each response) are assumed in plotting the profit charts. The top half of Figure 4.6 relates to assessment based on the construction data set and the bottom half relates to assessment based on the validation data set.

The assessment results are summarised in Figures 4.7 to 4.9, where the left panels show the in-sample results based on the construction data set and the right panels show the hold-out results based on the validation data set. As mentioned earlier, in-sample results tend to be upward biased because the same data set is used to construct and validate the model.

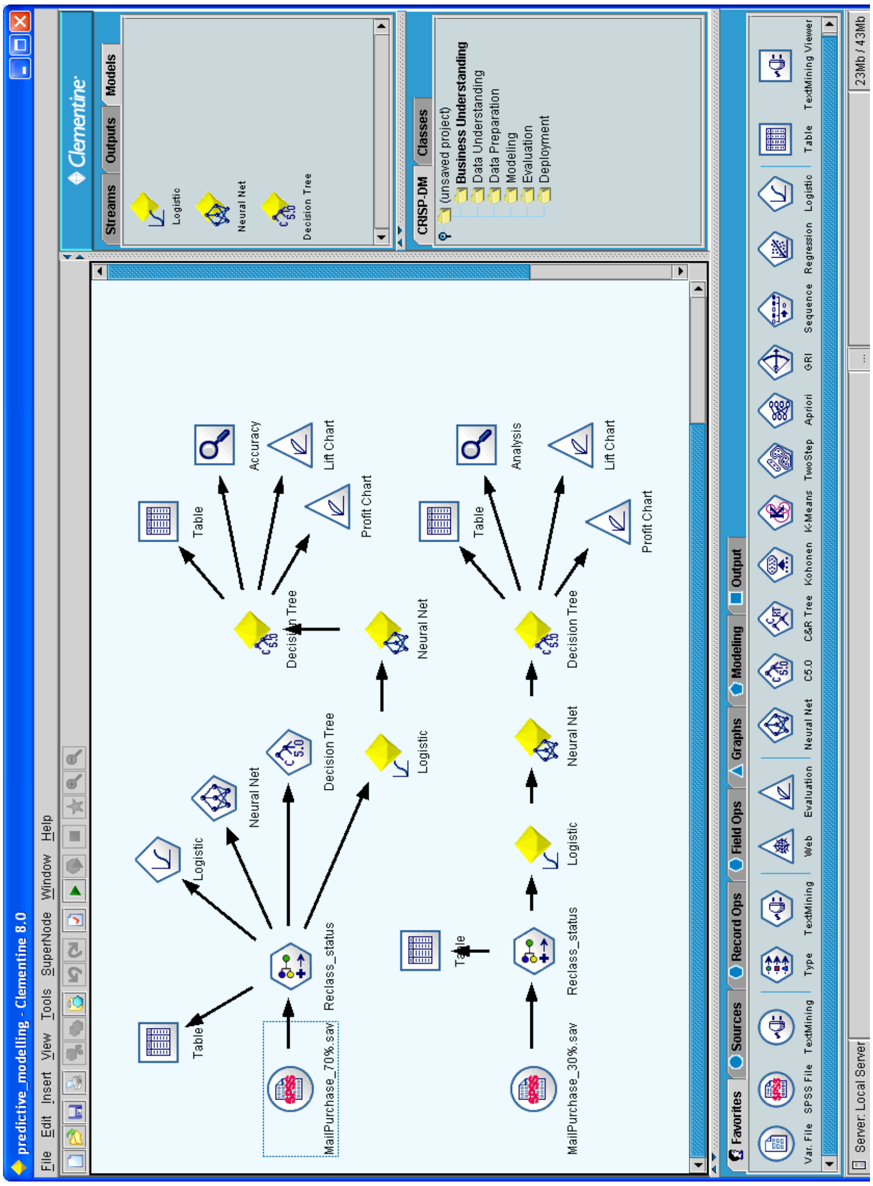


Figure 4.6 Data Mining Stream for Predictive Modelling Illustration

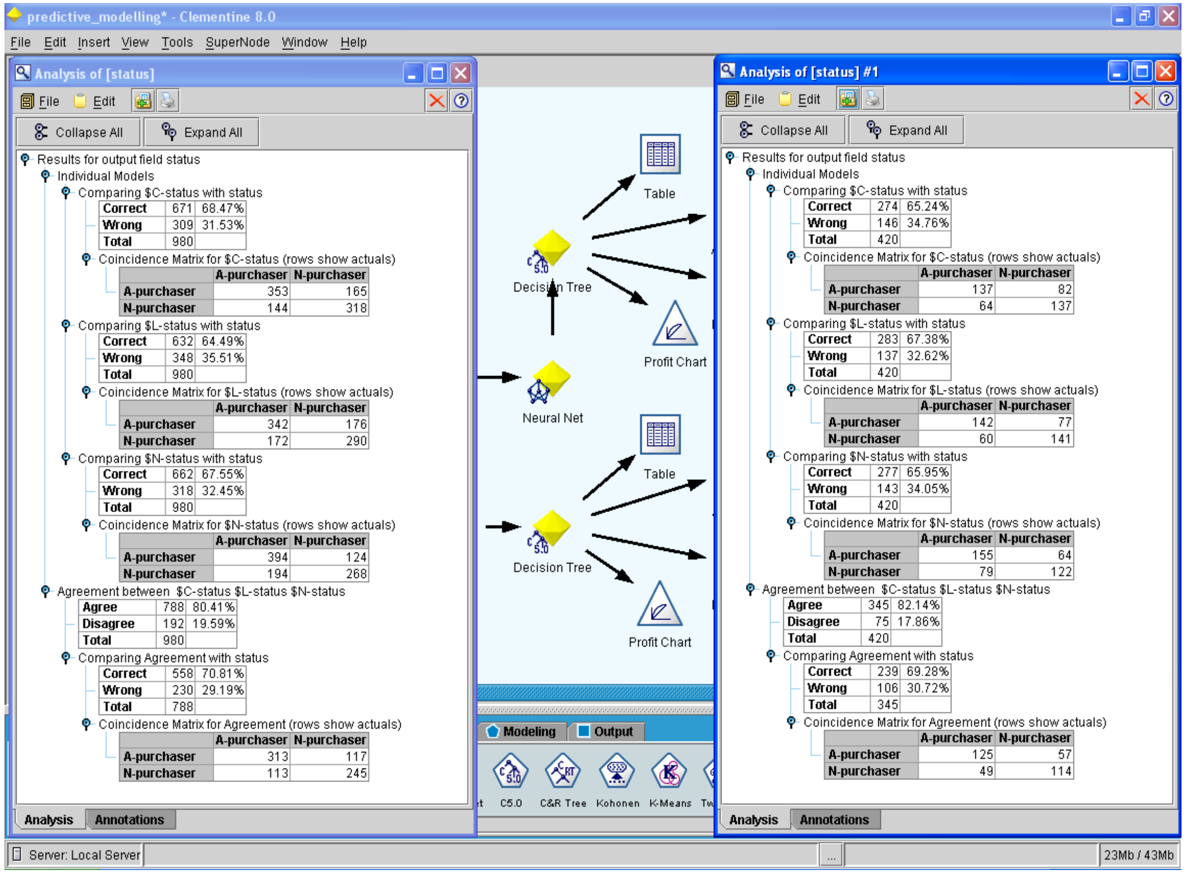
Also, neural networks and decision trees are prone to over-fitting. Hence, it is common for stopping rules of these two models to involve hold-out data sets. In Figures 4.7 to 4.9, the in-sample results are provided for reference only. Interpretation of the assessment results will be based on the hold-out results (i.e., the validation data set).

As shown in Figure 4.7 (right panel), the logistic regression model is the most accurate with an overall accuracy rate of 67.38%, followed by the neural network model (accuracy of 65.95%) and the decision tree model (accuracy of 65.24%). Hence, based on overall accuracy rates alone, the logistic regression model is the best (i.e., champion) model.

The lift charts in Figure 4.8 (right panel) shows a more detailed picture. As highlighted earlier, lift charts are based on hit rates and not accuracy rates. As shown, for the first two segments (or deciles since the validation data set is grouped into ten segments), the neural network model dominates with a lift value of 1.6. However, from the third segment onwards, the logistic regression model dominates. Hence, which model is best depends on how the model will be used.

For this data mining application, this means that which model is best depends on how many customers MailPurchase intends to target. Suppose that MailPurchase wishes to target the top 50% of customers in the database with the highest probabilities (or confidence) of purchase. Then, the logistic regression model will be the best model to use. However, if only the top 20% of customers should be targeted, then the neural network model should be used.

In addition to the above, suppose that MailPurchase wants to consider the financial aspects of using the prediction model (by incorporating the cost and revenue per unit of response to the marketing campaign). In this case, the profit charts in Figure 4.9 (right panel) will be relevant.



Analysis of [status]

```
Results for output field status
  Individual Models
    Comparing $C-status with status
      Correct   671  68.47%
      Wrong     309  31.53%
      Total     980
      Coincidence Matrix for $C-status (rows show actuals)
                    A-purchaser  N-purchaser
        A-purchaser         353          165
        N-purchaser         144          318
    Comparing $L-status with status
      Correct   632  64.49%
      Wrong     348  35.51%
      Total     980
      Coincidence Matrix for $L-status (rows show actuals)
                    A-purchaser  N-purchaser
        A-purchaser         342          176
        N-purchaser         172          290
    Comparing $N-status with status
      Correct   662  67.55%
      Wrong     318  32.45%
      Total     980
      Coincidence Matrix for $N-status (rows show actuals)
                    A-purchaser  N-purchaser
        A-purchaser         394          124
        N-purchaser         194          268
  Agreement between  $C-status $L-status $N-status
    Agree      788  80.41%
    Disagree   192  19.59%
    Total      980
    Comparing Agreement with status
      Correct   558  70.81%
      Wrong     230  29.19%
      Total     788
      Coincidence Matrix for Agreement (rows show actuals)
                    A-purchaser  N-purchaser
        A-purchaser         313          117
        N-purchaser         113          245
```

Analysis of [status] #1

```
Results for output field status
  Individual Models
    Comparing $C-status with status
      Correct   274  65.24%
      Wrong     146  34.76%
      Total     420
      Coincidence Matrix for $C-status (rows show actuals)
                    A-purchaser  N-purchaser
        A-purchaser         137           82
        N-purchaser          64          137
    Comparing $L-status with status
      Correct   283  67.38%
      Wrong     137  32.62%
      Total     420
      Coincidence Matrix for $L-status (rows show actuals)
                    A-purchaser  N-purchaser
        A-purchaser         142           77
        N-purchaser          60          141
    Comparing $N-status with status
      Correct   277  65.95%
      Wrong     143  34.05%
      Total     420
      Coincidence Matrix for $N-status (rows show actuals)
                    A-purchaser  N-purchaser
        A-purchaser         155           64
        N-purchaser          79          122
  Agreement between  $C-status $L-status $N-status
    Agree      345  82.14%
    Disagree    75  17.86%
    Total      420
    Comparing Agreement with status
      Correct   239  69.28%
      Wrong     106  30.72%
      Total     345
      Coincidence Matrix for Agreement (rows show actuals)
                    A-purchaser  N-purchaser
        A-purchaser         125           57
        N-purchaser          49          114
```

Figure 4.7 Comparative Accuracy Rates

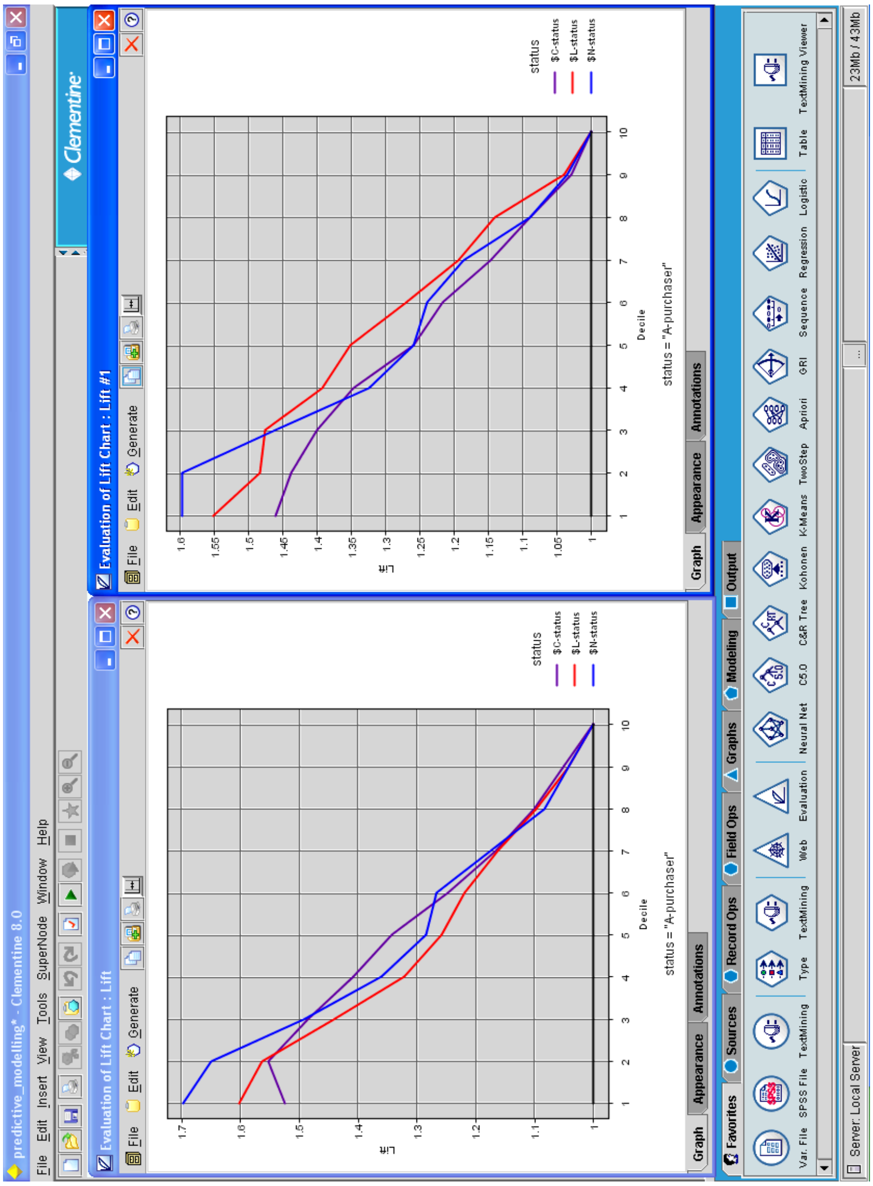


Figure 4.8 Comparative Lift Charts

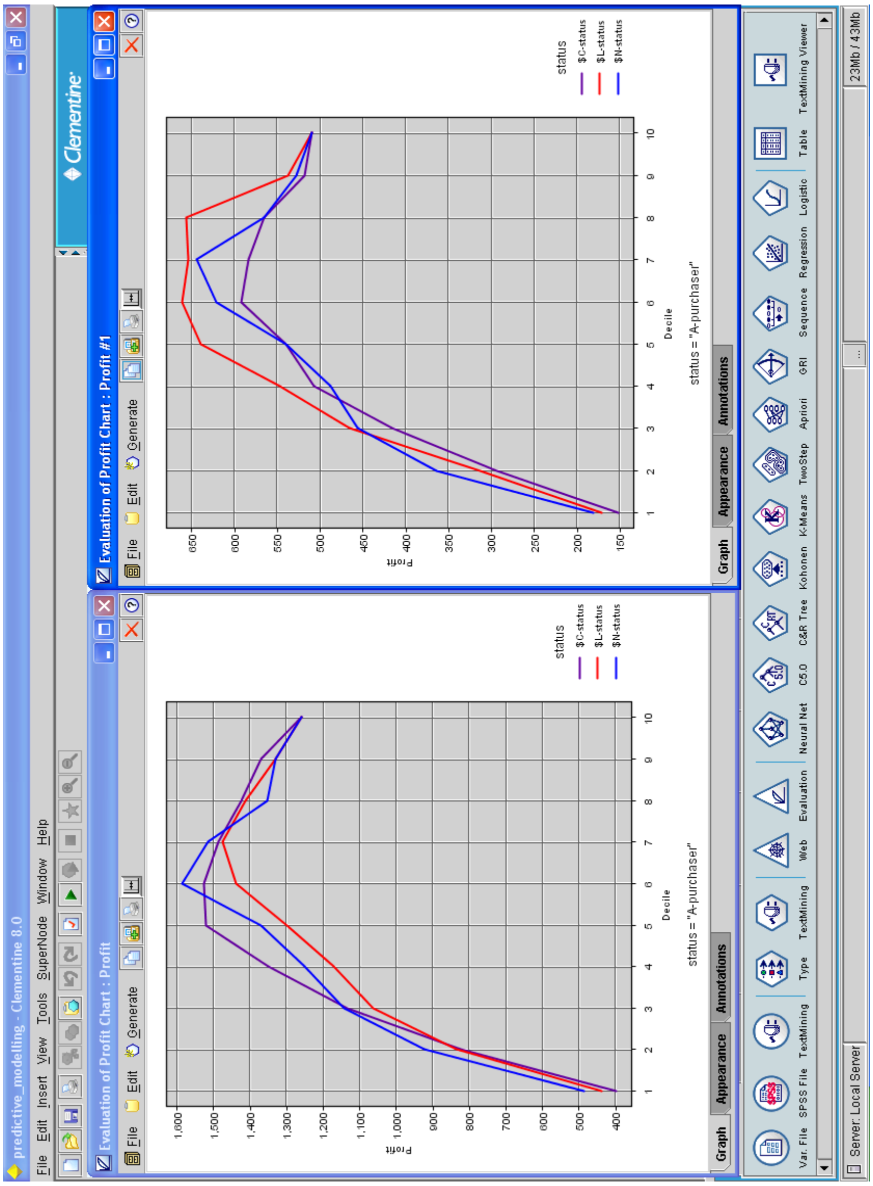


Figure 4.9 Comparative Profit Charts

To maximise profit, MailPurchase should use the logistic regression model to predict the probability of purchase and send the marketing brochure to the top 60% (i.e., first six segments) with the highest predicted probabilities of purchase.

Finally, the comparative results of the logistic regression, neural network and decision tree models discussed above are not generalisable to other data mining applications. Which model is best depends on the data set and the context. Also, the decision tree model does not appear to work well (relative to the other two models) in this illustration. However, it is easy to interpret and use, as shown in Figure 4.10 (based on the construction data set only). In some other data mining applications, decision trees may provide the best performance.

4.7 Summary

This chapter discusses a few important data mining issues that are applicable in the predictive modelling context. In particular, the need for and the common methods of model validation are discussed. Validation (or hold-out sample) results give better assessment of the prediction models.

Next, two factors that affect the optimal cut-off point are highlighted. These are the prior probabilities of event and non-event and the relative misclassification cost. Adjustment to the cut-off point is required to ensure either a minimum number of misclassifications or a minimum expected cost of using the model.

Assessment of models is an important part of the data mining process. In this chapter, evaluation charts based on hit rates (i.e., response, gains and lift charts) are discussed. They can be used to compare competing prediction models and provide an alternative perspective to accuracy rates.

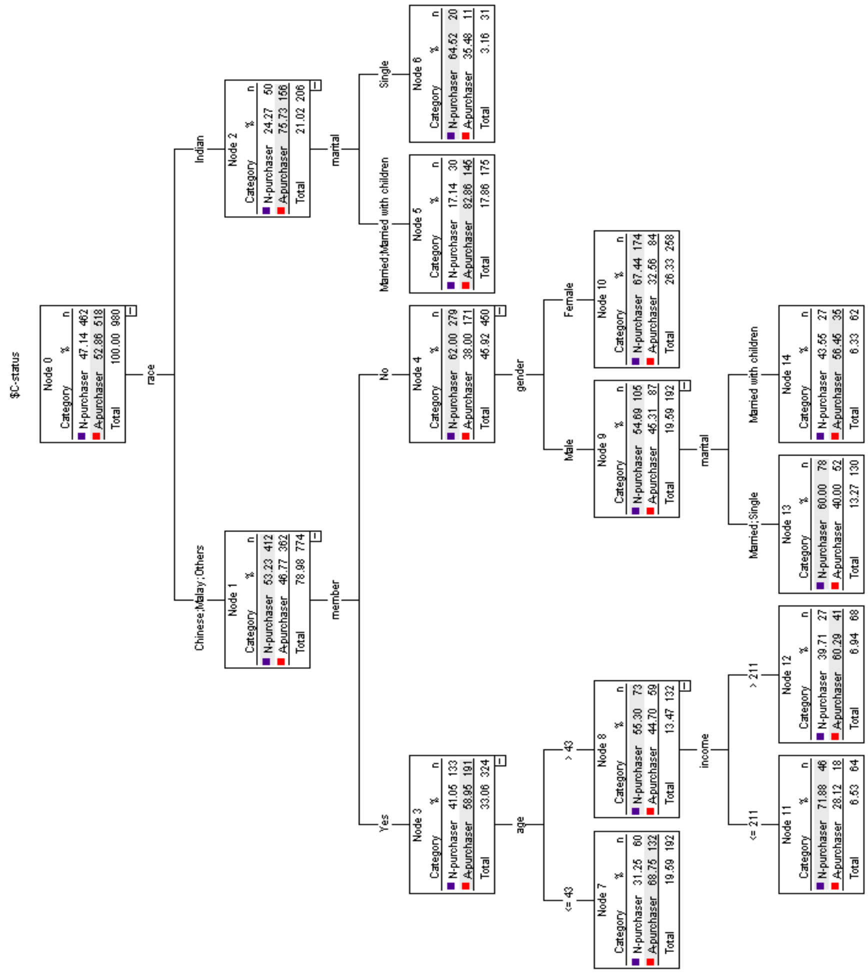


Figure 4.10 Decision Tree Model

Evaluation charts are next extended to incorporate financial aspects. This leads to profit and ROI (return on investment) charts, which relate back to the commercial objectives of data mining applications. Generally, the idea is to maximise either profit or ROI. Evaluation charts can be cumulative or non-cumulative.

Finally, the illustrations in the previous chapter are combined and extended to look at logistic regression, neural networks and decision trees in the same data mining stream. Also, model validation and evaluation charts are employed to assess the prediction models. Some of the data mining issues discussed in this chapter can also be found in Koh and Chan (2002).

Chapter References

Afifi, A. A. and Clark, V. (1996), *Computer-aided Multivariate Analysis*. Chapman & Hall, London.

Koh, H. C. (1992), "The sensitivity of optimal cutoff points to misclassification costs of Type I and Type II errors in the going-concern prediction context", *Journal of Business Finance & Accounting*, Vol. 19 No. 2, pp. 187-197.

Koh, H. C. and Chan, G. K. L. (2002), "Data mining and customer relationship management in the banking industry", *Singapore Management Review/Asia-Pacific Journal of Management Theory and Practice*, Vol. 24 No. 2, pp. 1-27.