
Data mining approaches to modelling insurance risk

Inna Kolyshkina, Richard Brookes

PricewaterhouseCoopers

22 October 2002

Contents

1	Introduction	3
2	What is data mining?	3
3	Why use data mining?	4
4	Data mining methodologies – Decision trees (CART), MARS and hybrid models	5
5	Case study – Claims streaming in Workers Compensation insurance	7
6	Case study – Predicting hospital cost in health insurance	12
7	Appendix - Brief description of CART	16
8	References	18
	Table 1 Misclassification tables for the learning and test data	9
	Figure 1 Gains chart for the CART model	10
	Figure 2 Claim Streaming Evaluation	11
	Figure 3 The gains chart for overall annual hospital cost	14
	Figure 4 The bar chart of averaged actual and predicted values for overall annual hospital cost	15
	Figure 5 Scatter graph	16

1 Introduction

Interest in data mining techniques has been increasing recently among actuaries and statisticians involved in analysing the large data sets common in many areas of insurance. This paper discusses the use of some data mining techniques in insurance and presents two case studies illustrating the application of classification and regression trees (CART), multivariate adaptive regression splines (MARS) and hybrid models. Some of this material has or will be published elsewhere, albeit in a slightly different form (Kolyshkina et al, 2002a, Kolyshkina et al, 2002b).

This paper is not intended to be a treatise on “how to apply data mining to insurance”; this would take several volumes and already be out of date in such a rapidly advancing field. Rather, it is intended to introduce some techniques and sketch briefly some actual examples where they have been used to good effect. For more technical details, readers are referred to the bibliography and in particular Hastie et al, 2001.

2 What is data mining?

Data mining can be defined as a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. Data mining tools are based on highly automated search procedures. The machine does far more of the work than is the norm in classical statistical modeling. This may require millions of times more computation than is needed for conventional statistical analysis but given high speed algorithms and today’s fast processors it usually takes only a few hours at the maximum to run a data mining model.

The best data mining methods can automatically select data to use in pattern recognition, are generally capable of dealing with noisy and incomplete data, include self-testing to assure that findings are genuine and provide clear presentation of results and useful feedback to analysts. Some of the best known data mining methodologies are decision trees, neural networks (for those interested in neural networks some useful references are given in the bibliography, including Ripley (1996) and Bishop (1995)), MARS (Friedman (1991)) and boosting. These methods can be used separately or in combination with each other. Data mining however does not replace traditional statistical techniques. Rather, it is an extension of existing statistical methods and, as we discuss later can be effectively used in combination with them.

3 Why use data mining?

Recently a number of publications have examined the use of data mining methods in insurance and actuarial environment (eg, Francis, 2001, WorkCover NSW News, 2001) and this can be contrasted with the use of “classical” approaches such as generalised linear models (Haberman and Renshaw, 1998; McCullagh and Nelder, 1989; Smyth, 2002). The main reasons for the increasing attractiveness of the data mining approach are as follows:

- It overcomes the shortcomings of traditional methods that operate under the assumption that data are distributed normally (as is the case in linear regression) or according to another distribution in the exponential family, such as binomial, Poisson or Gamma (as is required for a generalised linear model). Classical linear methods are based on such assumptions, which can be incorrect and may be difficult to test.
- It relies more than traditional models on the intense use of computing power. This results in analyses that are less time consuming and more flexible in terms of selection of predictors than those carried out by classical methods. Classical methods applied to large data sets can take longer to develop models and have particular trouble selecting important interactions between predictors.
- It is able to handle categorical variables with a large number of categories (for example, occupation, industry or postcode). Classical methods can have trouble dealing with such variables: as a result, they are either left out of the model, or have to be grouped prior to inclusion.
- Some data mining methods such as CART have methods to handle incomplete or noisy data, which are improvements over those available for traditional linear methods.

Vapnik (1996) and Hastie et al. (2001) give more detail on these points.

4 Data mining methodologies – Decision trees (CART), MARS and hybrid models

There are a wide variety of data mining methodologies and software packages currently available. Examples are Clementine (SPSS) , Intelligent Miner(IBM), Enterprise Miner(SAS), CART and MARS (both Salford Systems). We describe here CART and MARS and a combined methodology in more detail.

A. Classification and regression trees (CART)

The CART methodology is technically known as binary recursive partitioning (Breiman et al, 1984). It is binary because the process of modelling involves dividing data set into exactly two subgroups (or “nodes”) that are more homogeneous with respect to the response variable than is the initial data set. It is recursive because the process is repeated for each of the resulting nodes. The resulting model is called a decision tree or simply tree. To split data into two nodes, CART asks questions that have a yes/no answer. For example, questions might be: “Is age greater than 55?” or “Is the postcode equal to one of 1052, 1530 or 2300?”. CART then compares all possible splits for all values for all variables included in the analysis and conducts an exhaustive search through them all, selecting the split that divides data into two nodes with the highest degree of homogeneity. Once this best split is found, CART repeats the search process for each “child” node, continuing recursively until further splitting is impossible or stopped.

When the data set is sufficiently large, CART builds the model on a randomly selected part (usually two-thirds) of the data (the “learning sample”) and then tests it on the remaining data (the “test sample”). The learning sample is used to grow an overly large tree. The test sample is then used to estimate the rate at which cases are misclassified (possibly adjusted by misclassification costs) and to “prune” the large tree accordingly. The nature of this self-testing process of model building further enhances the predictive power of the model.

The resulting model is usually represented visually as a tree diagram. It divides all data into a set of several non-overlapping subgroups or nodes so that the estimate of the response is “close” to the actual value of the response within each node (Lewis et al, 1993). Important features of CART are being unaffected by outliers and its ability to deal with missing values easily.

An obvious disadvantage of any decision tree including CART is the fact that the predicted value of the response is discontinuous, which means that sometimes a small change in the value of a predictor variable could lead to a large change in the predicted value of the response. Also, the decision tree model is coarse-grained in the sense that a model with n nodes can only predict n different probabilities, which can be an issue, particularly if the tree produces only a small number of nodes (Steinberg & Cardell, 1998a, 1998b).

Another disadvantage is weakness at capturing strong linear structure. A very large tree can be produced in an attempt to represent very simple linear relationships. The algorithm recognizes the structure but cannot represent it effectively and in this case GLM's can perform better.

B. Multivariate adaptive regression splines (MARS)

MARS is a modification of the CART methodology to improve performance where the response is continuous rather than binary or categorical. The MARS algorithm models linear structures more effectively because it foregoes the tree structure and gains the ability to capture additive effects (Hastie et al, 2001).

The MARS procedure builds flexible regression models by fitting separate splines (or basis functions) to distinct intervals of the predictor variables. Both the variables to use and the end points of the intervals for each variable (*knots*) are found via a brute force, exhaustive search procedure. Variables, knots and/or interactions are optimised simultaneously by evaluating a "loss of fit" (LOF) criterion. MARS chooses the LOF that most improves the model at each step. In addition to searching variables one by one, MARS also searches for interactions between variables, allowing any degree of interaction to be considered. The "optimal" MARS model is selected in a two-phase process. In the first phase, a model is grown by adding basis functions (new main effects, knots, or interactions) until an overly large model is found. In the second phase, basis functions are deleted in order of least contribution to the model until an optimal balance of bias and variance is found. By allowing for any arbitrary shape for the response function as well as for interactions, and by using the two-phase model selection method, MARS is capable of reliably tracking very complex data structures that often hide in high-dimensional data (Salford Systems, 2002).

C. Hybrid models

The main disadvantages of CART (and decision trees in general) are discussed above. These include its coarse-grained nature and lack of ability to represent linear structure. Data analysis techniques including linear models, neural networks and MARS provide a continuous smooth response and a unique predicted value for every record but can be sensitive to outliers and missing values.

It is natural to combine these “smooth” modeling techniques with decision trees in such a way that their strengths can be combined effectively. Steinberg and Cardell (1998a, 1998b) describe the one possible methodology. Firstly, a CART tree is built on the data. Then the output of the model in the form of terminal node indicator, or the predicted values or of the complete set of indicator dummies is included among other inputs in the “smooth” model. They point out that the additional effects identified by such hybrid model are likely to be weak as all strong effects already detected by the tree, but nevertheless, a collection of weak effects can be very significant.

We have used this methodology to good effect. Occasionally, we have also used the MARS basis functions produced by such a hybrid model as parameters for logistic regression or a generalised linear model. Whilst not technically optimal, there are some situations where this last approach has yielded good results.

5 Case study – Claims streaming in Workers Compensation insurance

A. Background

In workers’ compensation insurance, serious claims comprise a small proportion of all claims by number but the great majority of the incurred cost. These claims are a natural target for insurers wishing to reduce overall cost. From a practical point of view, the insurer must ensure that the management of claimant injuries is carried out in such a way that the injured person receives the most effective medical treatment at appropriate points in time to prevent his or her injury from becoming chronic and to enable the claimant to return to work in a timely and effective manner.

To do so, the insurer ideally would need to know, at the time of a claim being received, whether the claim is likely to become serious. But in most cases this is not obvious as there are many factors contributing to the result. Therefore, it would be useful to have a model that would account for all such factors and would be able to predict at the outset of a claim the likelihood of this claim becoming serious.

B. Data

The data available for modeling was represented by several years’ worth of information about a large number of claims from the NSW workers’ compensation scheme. This was in the form of a record for each claim for each month in which there was any activity on the claim. The data includes information about:

- The claim; for example, the date the claim was registered, the date when the claim was closed, whether the claim was reopened, whether the claim was litigated, various liability estimates, payments made on the claim, reporting delay.
- The claimant; for example; gender, age, family situation and whether the claimant had dependants, type of employment and work duties such as code for industry and occupation, nature of employment (permanent, casual, part or fulltime), wages.
- The injury or disease; for example time and place of injury, injury type, body location of the injury, cause or mechanism of injury, nature of injury.

Overall there were 83 variables that might have been considered as potential predictors. A number of these were categorical with many categories; for example the variable “occupation code” had 285 categories, “injury location code” had 85 categories.

We defined a serious claim as one where the claimant had received payment for at least three months of time off work, or where the claim was litigated. This results in approximately 14% of reported claims being classified as serious. Those claims make up around 90% of the total claim cost. Claims used for modelling were a random sample of those with a date of injury in the period 24-18 months before the latest data date.

C. Analysis

The two major purposes of the analysis were to:

- Identify the most important predictor variables from a large number of available variables containing information about a claim at the time when the claim is registered
- To build a model based on such predictor variables, which would classify a claim as “likely to become serious” or “not likely to become serious”.

A subsidiary objective of the analysis was to compare the performance of the data mining and traditional statistical modeling approaches in terms of prediction accuracy, computational speed and ability to handle some issues common in insurance data such as presence of missing values and predictors with many categories with the view to select one of the methods for future use with this type of data. This aspect is reported more fully in Kolyshkina et al, 2002.

D. Analysis using CART

CART selected 19 significant predictor variables out of 83 potential predictor variables. It is interesting to note that some variables that turned out to be important predictors were expected to be so on the basis of previous experience and analysis, for example, injury details (nature, location and

mechanism of injury), while some others, such as language skills of the claimant, were unexpected.

E. Evaluation of Model

CART offers two main tools of model evaluation; the gains chart and classification tables. These are calculated for both the learning and test samples. These tools allow us to examine three aspects of the model:

- We can conduct specificity and sensitivity analysis from the classification table.
- We can investigate model stability by comparing classification tables for the test and learning samples (Table 1).
- We can study how well model performs in terms of the ranking of the cases by examining the gains chart (Figure 1).

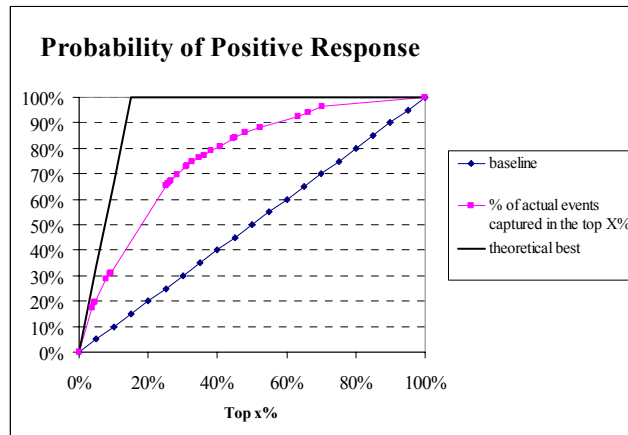
Table 1 Misclassification tables for the learning and test data

<i>Misclassification for learning data</i>				
Class	N Cases	N Misclassified	Percent Error	Cost
Serious	16,922	3,891	22.99	0.23
Non-Serious	105,358	25,744	24.43	0.24

<i>Misclassification for test data</i>				
Class	N Cases	N Misclassified	Percent Error	Cost
Serious	8,558	2,275	26.58	0.27
Non-Serious	52,866	12,923	24.44	0.24

We tried different ways of randomly dividing the data into the learning and test samples, for example 50 percent in each sample, 60 percent in the learning sample and 40 percent in the test sample etc, but the results were very similar each time. We present in Table 1 results based on the default split suggested by CART software which is a randomly selected two thirds of the data as the learning sample and the remaining data as the test sample. Table 1 suggests that the model performs very similarly on the learning and test samples as should be the case when CART is allowed to prune the tree to optimal size. It is possible to select larger trees but these give a much wider performance differential between the learning and test samples (as is to be expected, given that this is essentially overfitting).

Figure 1 Gains chart for the CART model



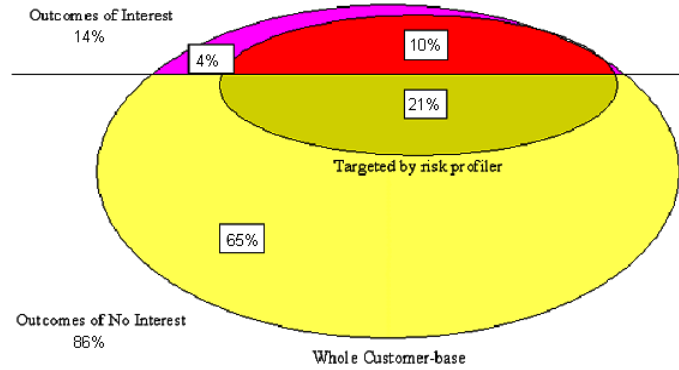
A gains chart is a popular model evaluation tool in data mining. The data are ordered from nodes with the highest proportion of cases in class 1 (in this case “serious” claims) to the lowest. The horizontal axis represents the percentage of the population examined and the vertical axis represents the percentage of “serious” claims identified. The 45° line represents the situation if the tree were giving no useful information about “serious” claims, that is if each node were a random sample of the population. The curved line represents the cumulative percentage of “serious” claims versus the cumulative percentage of the total population. The vertical difference between these two lines gives the gain or lift at each point along the horizontal axis. The ideal gains chart in this case is given by the line shown on the graph which reaches 100% at the actual proportion of serious claims (in this case, around 15%).

Gains charts help the analyst to appreciate how well the model predicts and how well it ranks predicted cases: the higher the curve is above the line, the better the model. For example, consider the gains chart for the CART tree shown in Figure 1. In the first 20% of the population we find 60% of the “serious” claims, a gain or lift of 40%.

Figure 1 is based on a test data sample from the same time period as the learning data. Another valuable validation procedure would be to construct a gains chart based on a test data sample from a later time period. We did not carry out such a validation in this case.

Yet another way of examining model is related to its efficiency in terms of claims streaming and is really just a re-expression of the figures in the misclassification table. The diagram below shows that the model classifies 31% of claims as potentially serious. Of this 31%, 10% turn out to be serious and 21% are not (generally called *false positives*). The model fails to correctly classify 4% of claims which subsequently prove to be serious.

Figure 2 Claim Streaming Evaluation



F. Comparison with logistic regression

Our first step was to attempt building the model using logistic regression, the traditional statistical modeling approach for analysis of data with binary response. Logistic regression is a well-known classical technique and is easily implemented in a number of software packages. In this case, we used SAS.

We identified the following differences between the two approaches:

- Computational speed and time requirements (those interested in the limitations of PROC Logistic in this respect should consult SAS 2002a-c)
- Significant predictor variable selection
- Handling categorical predictors with many categories
- Picking up interactions of predictors
- Handling missing values
- The interpretability of the model
- Individual scoring versus segmenting
- Model evaluation and accuracy.

The detailed report on the differences found can be found in Kolyshkina et al, 2002a. That paper shows that, in terms of ranking, the logistic regression model performs almost as well as the CART model for the first 15% of the population, but significantly worse thereafter. We want to add however that in some situations for instance when the number of predictor variables is relatively small, most of them are numeric rather than categorical, and the

assumptions for logistic regression are clearly valid, logistic regression might be preferred to a tree model.

G. Future direction

There are many ways in which the performance of the resulting claims streamer can be improved. Some of these relate to the model structure and analysis, for instance carrying out more extensive analysis to find a model which responds to the length of time that a claim has been open. This would enable the model to be used for purposes such as ongoing resource management for claims staff. Another area for improvement is adding information not present in the database such as the existence of psychological trauma, or Evidence Based Medicine information sourced from the treating doctor. This latter improvement requires a period of data collection, for instance in the context of a pilot study.

6 Case study – Predicting hospital cost in health insurance

Lifetime customer value is the discounted present value of income less expenses associated with a customer. We have applied data mining methodology to develop a model of total projected customer value for a health insurer. This required a number of sub-models, including:

- Ancillary claim frequency and cost for the next year
- Hospital claim frequency and cost for the next year
- Transitions from one type of product to another
- Births, marriages, deaths and divorce
- Lapses.

These sub-models were combined and projected forward in a multi-state transition structure to give a model of overall projected lifetime customer value.

We will concentrate in this paper on the model that which was used to project hospital claim cost over a single year.

A. Data

De-identified data was available at a member level over a 36 month period. The model structure was to use information available over the first 24 months to fit a model based on outcomes over the last 12 months. There were further exclusions to account for issues such as waiting periods and lapses.

The available variables can be grouped as follows.

- Demographic variables, such as age of the customer, gender, family status
- Geographic and socio-economic variables such as location of the member's residence, socio-economic indices related to location such as indices of education, occupation, relative socio-economic advantage and disadvantage
- Membership and product details such as duration of the membership, details of the product held
- Variables related to claim history and medical diagnosis
- Other variables such as distribution channel, most common transaction channel, payment method.

Overall there were approximately 300 candidate predictor variables.

B. Overall modelling approach

The hospital claim cost consisted of two sub-models; one for the probability of at least one hospital claim over twelve months and the second for the cost given at least one claim. The hospital claims were segregated into three mutually exclusive classes:

- Those lasting only one day
- Those lasting more than one day and with a surgical procedure
- Other claims.

Clinical experts we consulted expected each of these hospital events to have different risk drivers and this was borne out by the model results.

C. Modelling methodology

We first applied CART for the purposes of exploratory analysis. The preliminary tree model suggested segmenting the customer base into four broad groups according to age and previous claims experience.

We then built a separate CART model for each of the segments. The risk drivers for these models were sometimes similar. For example, age and the type of hospital cover were among important predictors for practically all groups but in many cases there were significant differences.

The response (hospital claim cost value) was continuous, and it was important in this case to provide a continuous predicted value and an individual estimate for each record rather than simply segment the data into homogeneous groups in terms of the average predicted hospital cost. Also we expected some of our potential important predictors selected by CART, such as age, to be related to the response in a linear way and, as mentioned above, decision trees can be weak in this situation. For these reasons to further enhance the predictive accuracy of the model, we then built a hybrid

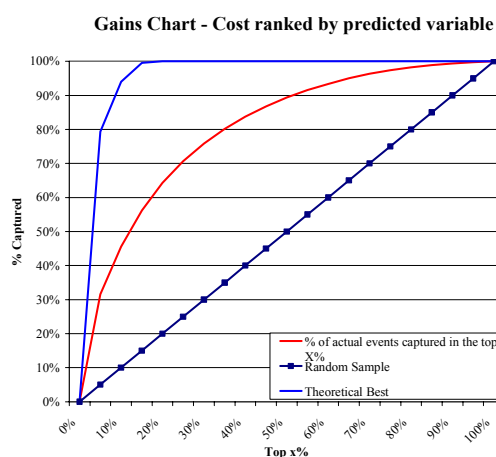
model of CART and MARS using the hybrid modelling methodology (Steinberg & Cardell, 1998a, Steinberg & Cardell, 1998b) outlined above. This was achieved by including the decision tree segment for each record in the form of a categorical variable as one of the input variables into a MARS model. Other inputs into the MARS model were variables selected by CART as important predictors.

MARS, like CART, ranks variables in order of their importance as predictors from the highest to the lowest. It ranked the predictor variable representing the CART decision tree output as the most important. However, as we had expected, other, mostly continuous variables were high in the importance list which showed that the MARS model was finding minor linear effects that were not picked by the decision tree.

D. Evaluation

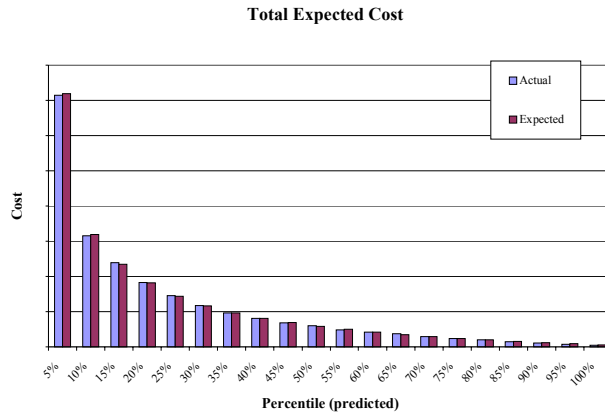
The main tools we used for summarizing the model diagnostics were gains charts and analyses of actual versus predicted values. An example gains chart is given below.

Figure 3 The gains chart for overall annual hospital cost



The gains chart for the overall hospital claims cost model presented in Figure 3 is a continuous analogue of the classification gains chart presented for the previous case study. It shows that we are able to predict the high cost claimants with a good degree of accuracy. As a rough guide, the overall claim frequency is 15%. Taking the 15% of members predicted as having the highest cost by the model, we end up with 56% of the total actual cost. Taking the top 30% of members predicted as having the highest cost by the model, we end up with almost 80% of the total actual cost. Much of this performance is generated simply by the dependence of cost on age but we estimate that the model is improving the gains chart by around 10% of total cost in the high cost part of the population, over a simple model depending on age.

Figure 4 The bar chart of averaged actual and predicted values for overall annual hospital cost



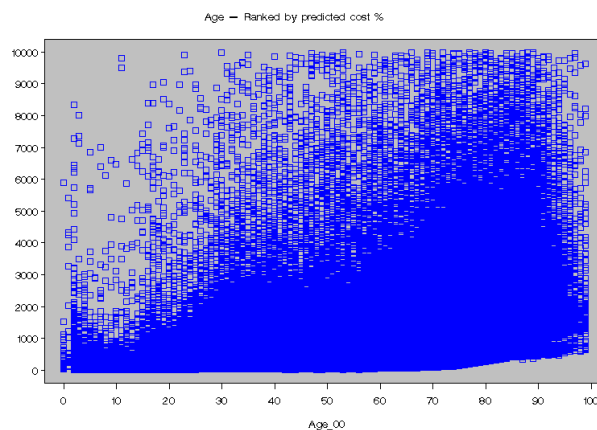
A further diagnostic of model performance is analysis of actual versus expected values of the response variable. A summary chart is shown in Figure 4. The chart shows average predicted and actual values for members ranked from highest to lowest in terms of predicted cost and segmented into 20 equally sized groups.

For a good model such a chart will demonstrate a high degree of differentiation; that is, the highest average predicted value will be much higher than the lowest average predicted value and the actual and predicted average values will match closely. If we consider the chart presented in Figure 4, we will see that the shape of the chart suggests a good degree of differentiation as well as good model fit. The model slightly over-predicts for the lower expected costs but this was of little business importance for the client.

E. Results

Details of the resulting model are commercially sensitive. However, we can state that many of the potential predictors given above were indeed significant to a degree greater than we had expected. One way of showing this is by means of a graph of predicted cost by age which we show below. If age were the only significant variable this would be a single curve. The fact that it is so scattered is evidence that the model is using many other factors to arrive at the predicted cost.

Figure 5 Scatter graph



7 Appendix - Brief description of CART

The description below is based on Salford Systems, 2000, and further details on CART can be obtained from the same source.

The CART methodology is technically known as binary recursive partitioning. It is binary because the process of modelling involves dividing data set into exactly two subgroups (or "nodes") that are more homogeneous with respect to the response variable than the initial data set. It is recursive because the process is repeated for each of the resulting nodes. The resulting model is called a decision tree or simply tree.

To split a node into two child nodes, CART asks questions that have a "yes" or "no" answer. For example, the questions might be: "is age greater than 55?" or "is the location of injury code equal to 1, 15 or 23?"

CART compares groups resulting from all possible splits for all values for all variables included in the analysis. For example, consider a data set with 100 cases and 19 variables. CART considers up to 100 times 19 splits for a total of 1900 possible splits. CART will conduct a brute force search through them

all, selecting the split that divides data into groups with the highest degree of homogeneity. Once this best split is found, CART repeats the search process for each child node, continuing recursively until further splitting is impossible or stopped. (There can be several reasons splitting to be stopped for, including that a node has too few cases.)

The tree-growing methodology is data-intensive, requiring many more cases than classical regression. When the data set is sufficiently large, CART builds the model on a randomly selected part (usually two-thirds) of the data (the “learning sample”) and then tests and refines it on the remaining data (the “test sample”). The learning sample is used to grow an overly large tree. The test sample is then used to estimate the rate at which cases are misclassified (possibly adjusted by misclassification costs). The misclassification error rate is calculated for the largest tree and also for every sub-tree. The best sub-tree is the one with the lowest or near-lowest cost, which may be a relatively small tree.

When data are in short supply and insufficient for a separate test sample, CART employs cross-validation. In such cases, CART grows a maximal tree on the entire learning sample. This is the tree that will be pruned back. CART then proceeds by dividing the learning sample into 10 roughly-equal parts, each containing a similar distribution for the dependent variable. CART takes the first 9 parts of the data, constructs the largest possible tree, and uses the remaining 1/10 of the data to obtain initial estimates of the error rate of selected sub-trees. The same process is then repeated (growing the largest possible tree) on another 9/10 of the data while using a different 1/10 part as the test sample. The process continues until each part of the data has been held in reserve one time as a test sample. The results of the 10 mini-test samples are then combined to form error rates for trees of each possible size; these error rates are applied to the tree based on the entire learning sample.

The upshot of this complex process is a set of fairly reliable estimates of the independent predictive accuracy of the tree. Because the conventional methods of assessing tree accuracy can be wildly optimistic, cross-validation is the method CART normally uses to obtain objective measures for smaller data sets.

The nature of this process of model building further enhances stability of the model.

A CART tree can be built using a number of classification criteria. To give an idea of how such criteria work, we will discuss in more detail Gini and Twoing criteria that CART offers as suitable for a binary response.

The Gini impurity criterion was introduced by Breiman et al (1984) and is based on class heterogeneity. It is given by $i(t)=1-S$, where S is the sum of squared probabilities $p(j | t)$ summed over all levels of the dependent variable. If, for example in our case a node consists exclusively of the serious claims,

then the probability for this class is equal to 1, the probability for non-serious class is equal to 0, and the diversity index $i(t) = 0$. On the other hand, if the node contains an equal number of serious and non-serious claims, it would be characterised by the greatest possible diversity.

The Twoing criterion was introduced by Breiman et al (1984) and is a measure of the difference in the probability that a class appears in the left descendent rather than the right descendent node, and is based on the concept of class separation rather than class heterogeneity like the Gini criterion. The formal criterion to be maximised is $(\frac{1}{4} p_L p_R \sum_j (|p(j|t_L) - p(j|t_R)|)^2)$. The goal is to make the probability that a class j object goes to the left as different as possible from the probability that it goes to the right.

With the first case study presented above we ran a CART model on the full data set using the two methods, Gini and Twoing. Salford Systems (2000) note that sometimes the differences between the models obtained by selecting the splitting criterion will be modest and at other times profound, and mention examples where judicious choice of a splitting rule in CART reduces the error rate by 5 to 10 percent. They add that although there are certain rule-of-thumb recommendations about which splitting rule out of a few offered by CART software for classification-type modelling is best suited to what type of problem, it is good practice to always use a few different splitting rules and compare the results. Experimentation with splitting rules selection should theoretically provide us with different results, and such experimentation also provides us with better understanding of the data-specific issues (Salford Systems, 2000).

We compared the models built using both criteria and it appeared that the models gave very similar results in terms of the decision tree structure, the richness of nodes in serious claims and the selection of the important predictors. This finding further confirmed the stability of the model and the fact that the modeling methodology was correctly identifying the underlying data patterns that characterize this type of data. The Gini criterion-based model however, produced slightly better results in terms of prediction accuracy and ranking the cases than the Twoing criterion-based model.

8 References

Bishop, C., Neural Networks for Pattern Recognition, Oxford Univ. Press, 1995.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). Classification and Regression Trees. Wadsworth, Pacific Grove, CA.

L. Devroye, L. Györfi and G. Lugosi. A Probabilistic Theory of Pattern Recognition., Springer Verlag, New York, 1997.

Fahrmeir L. and G. Tutz (2001) *Multivariate statistical modelling based on generalized linear models* 2nd edition. Springer Verlag, New York, 2001.

Francis, L. (2001). Neural networks demystified. *Casualty Actuarial Society Forum*, Winter 2001, 252–319.

Friedman, J. H. (1991), Multivariate Adaptive Regression Splines (with discussion), *Annals of Statistics*, 19, 1-141(March).

Haberman, S. and Renshaw, A. E. (1998). Actuarial applications of generalized linear models. In Hand, D. J. and Jacka, S. D. (eds). *Statistics in Finance*. Arnold, London.

Han, J., and Camber M. (2001) *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Hastie, T., Tibshirani R. and Friedman, J. (2001). *The elements of statistical learning: Data Mining, Inference and prediction*. Springer-Verlag, New York.

Kolyshkina, I, Petocz P. and Rylander, I. "Modelling Insurance Risk: A Comparison of Data Mining and Logistic Regression Approaches" submitted to Australian and New Zealand Journal of Statistics in October 2002

Kolyshkina, I, Steinberg D. and Cardell, N. S. "Using Data Mining for Modelling Insurance Risk and Comparison of Data Mining and Linear Modeling Approaches" in the book "Intelligent Techniques in the Insurance Industry: Theory and Applications." submitted in October 2002

Lewis, P.A.W. and Stevens, J.G., "Nonlinear Modeling of Time Series using Multivariate Adaptive Regression Splines," *Journal of the American Statistical Association*, **86**, No. 416, 1991, pp. 864-867.

Lewis, P.A.W., Stevens, J., and Ray, B.K., "Modelling Time Series using Multivariate Adaptive Regression Splines (MARS)," in *Time Series Prediction: Forecasting the Future and Understanding the Past*, eds. Weigend, A. and Gershenfeld, N., Santa Fe Institute: Addison-Wesley, 1993, pp. 297-318.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd edition). Chapman and Hall, London.

Ripley, B., *Pattern recognition and Neural Networks*, Cambridge University Press, 1996.

Salford Systems (2000). *CART[®] for Windows User's Guide*. Salford Systems

Salford Systems (2002). MARS[®] (Multivariate Adaptive Regression Splines) [On-line] <http://www.salford-systems.com>, (accessed 08/10/2002).

SAS Institute (2002a). SAS Version 8. [On-line] <http://www.sas.com/>, (accessed 25/09/2002).

SAS Institute (2002b). PROC LOGISTIC in EXACT. [On-line] <http://www.sas.com/service/techsup/intro.html>, (accessed 25/09/2002).

SAS Institute (2002c). *Why is LOGISTIC taking so much CPU or real time?* [On-line] http://www.sas.com/service/techsup/faq/stat_proc/logiproc969.html (accessed 25/09/2002).

Smyth, G. (2002). Generalised linear modelling. [On-line] <http://www.statsci.org/glm/index.html>, (accessed 25/09/2002).

Steinberg, D. and Cardell, N. S. (1998a). Improving data mining with new hybrid methods. Presented at *DCI Database and Client Server World*, Boston, MA.

Steinberg, D. and Cardell, N. S. (1998b). The hybrid CART-Logit model in classification and data mining. *Eighth Annual Advanced Research Techniques Forum*, American Marketing Association, Keystone, CO.

Steinberg, D. and Colla, P. L., (1995). *CART: Tree-Structured Nonparametric Data Analysis*. Salford Systems, San Diego, CA.

Trahair, G. V., (1995) Developments in Personal Lines Rating in the UK, Proceedings of the Tenth General Insurance Seminar, Institute of Actuaries of Australia.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49, 1225–1231.

Vapnik, V. (1996). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

WorkCover NSW News (2001) Technology catches insurance fraud. [On-line] <http://www.workcover.nsw.gov.au/pdf/wca46.pdf> (accessed 08/10/02)