Journal of the Operational Research Society (2000) 51, 532–541
www.stockton-press.co.uk/jors

An analysis of customer retention and insurance claim patterns using data mining: a case study

KA Smith,[1]* RJ Willis[1] and M Brooks[2]

[1]*Monash University and* [2]*Australian Associated Motor Insurers Limited, Australia*

The insurance industry is concerned with many problems of interest to the operational research community. This paper presents a case study involving two such problems and solves them using a variety of techniques within the methodology of data mining. The first of these problems is the understanding of customer retention patterns by classifying policy holders as likely to renew or terminate their policies. The second is better understanding claim patterns, and identifying types of policy holders who are more at risk. Each of these problems impacts on the decisions relating to premium pricing, which directly affects profitability. A data mining methodology is used which views the knowledge discovery process within an holistic framework utilising hypothesis testing, statistics, clustering, decision trees, and neural networks at various stages. The impacts of the case study on the insurance company are discussed.

Keywords: data mining; insurance; neural networks; classification; clustering; case study

Introduction

The insurance industry is a source of a large number of business problems which fall under the concern of the operational researcher: risk assessment, classification of policy holders, portfolio optimisation, decision support, and planning and resource allocation, to name just a few. These problems can each be solved using traditional operational research techniques including regression, goal programming, linear programming, and more recent approaches such as neural networks, expert systems and data mining. Yet the documentation of these approaches to solving problems from the insurance industry is surprisingly scarce. The operational research literature does however provide a few excellent review articles including van Gelder,[1] and Haehling von Lanzenauer and Wright,[2] and we must then conclude that the highly competitive nature of the insurance industry precludes public dissemination of the results of more detailed successful applications of operational research in insurance.

As the insurance industry becomes more excited about emerging technologies like data mining however, the scarcity of related literature creates a challenge for the operational researcher and practitioner. Certainly, there are some reports of successful projects applying data mining and neural networks to problems like fraud detection,[3–5] underwriting,[6–10] insolvency prediction,[11] and customer segmentation,[12] yet few of these studies utilise data mining as a methodology capable of providing solutions to a variety of inter-related problems across the industry.

This paper aims to demonstrate the potential of data mining in the insurance industry through a case study. Insurance is an extremely competitive industry where a combination of market growth and profitability are seen as imperatives to success. Successful marketing campaigns, as well as strategic alliances, mergers and take-overs, can ensure that market growth is attained, but unless there is an understanding of the impact of this growth, profitability is at risk. Emerging techniques like data mining are proving to be of enormous benefit to the business world,[13] in terms of identifying hidden patterns in data, as well as predicting future behaviors of customers. The insurance company in this case study has a large and effective data warehouse which records details of every financial transaction and claim. The company realises that valuable information is hidden in this data, which can help them achieve their objectives of market growth and profitability.

Market growth and profitability are a result of making the right pricing decisions relative to claim costs. Therefore we need to be able to predict average claim costs and frequency of claims, and examine the effect of pricing on profitability. We also need to be aware of the effect of pricing on customer retention patterns, as well as providing the potential for market growth, considering the highly competitive nature of the business.

The purpose of this case study is then to demonstrate the potential of data mining as a means for achieving market growth and profitability. The overall problem of concern is to set a pricing level commensurate with predicted claim costs and at the same time price policies to retain existing

**Correspondence: Dr KA Smith, School of Business Systems, Monash University, Clayton, Victoria 3168, Australia.*
E-mail: kate.smith@infotech.monash.edu.au

customers and acquire new customers. The pricing problem is highly dependent, however, on knowing which customers are likely to renew their policy, as well as their level of risk, and their sensitivity to price increases. Accordingly, the pricing of insurance products cannot be studied without consideration of the many interrelated factors. If we are able to identify the relationship of price to policy acquisition and retention, and if we are able to predict claim costs, we would be able to predict the relationships between price, growth, and profitability.

The approach taken in this case study is to follow a data mining methodology, utilising hypothesis testing and statistics initially, and more sophisticated techniques for the knowledge discovery process. For the highly structured problem of customer retention modeling, regression, decision trees and neural networks are used, while we have used clustering for the less structured analysis of claim patterns. The data mining methodology, together with these techniques and the results of the case study are presented in the following sections.

Data mining methodology

Data mining has emerged over recent years as an extremely powerful approach to extracting meaningful information from large databases and data warehouses.[13] The increased computerisation of business transactions, improvements in storage and processing capacities of computers, as well as significant advances in knowledge discovery algorithms have all contributed to the evolution of the field.[14] Data mining has not been without criticism, however, and it appears that some data mining projects have been unsuccessful for a variety of reasons.[15] Perhaps the most perceptive quote on this topic comes from Small,[15] who observes:

> The new technology cycle typically goes like this: Enthusiasm for an innovation leads to spectacular assertions. Ignorant of the technology's true capabilities, users jump in without adequate preparation and training. Then, sobering reality sets in. Finally, frustrated and unhappy, users complain about the new technology and urge a return to 'business as usual'.

To the operational researcher, data mining may appear to be nothing more than a collection of common techniques: regression, decision trees, neural networks, genetic algorithms and clustering algorithms, etc. This assessment is partly correct in the sense that data mining frequently utilises a variety of these techniques for solving a problem. But the approach of data mining is more grounded in a methodology than this viewpoint implies. It is the adherence to a methodology as well as having an understanding of the individual techniques associated with data mining that can prevent the scenario described in the quote above from occurring. It is for this reason that the operational researcher is likely to find success when applying data mining; the approach is a natural extension of an existing problem solving methodology.

The methodology of data mining used in this case study views the discovery of information from a database as a four step process.[16] The business problem must be identified, then the data must be analysed. Action can be taken based on the results, and the outcomes of the action can be measured. The first and third steps raise mostly business issues, and it is the second step, data analysis, that is most interesting for the operational researcher. Within the process of analysing the data, there are also several steps which should be followed, including data preparation, initial descriptive statistics, hypothesis testing, and knowledge discovery through algorithms such as neural networks, clustering or decision trees. We refer the interested reader to Berry and Linoff[16] for an excellent introduction to data mining methodologies.

The purpose of this holistic approach to data mining is to ensure a full integration of the results with existing business knowledge and procedures. There is little point producing information unless it is directly actionable. The initial descriptive statistics of the data allow us the opportunity to become familiar with the data and to notice and treat any missing values or outliers that might distort the subsequent analysis. It also gives us an opportunity to formulate some initial hypotheses which can be tested using SQL or other database tools. It is at this point, before any actual knowledge discovery has occurred, that the existing business knowledge can be validated or refuted through hypothesis testing.

With the initial analysis completed, knowledge discovery algorithms can be applied to the data. The type of algorithm used depends on the nature of the problem. If the problem can be viewed as a problem of classification or prediction, and a complete set of training data is available, then the problem is well structured. Supervised learning algorithms like multilayered feedforward neural networks with backpropagation,[17,18] regression, or decision trees can be used to learn the relationship between variables and correct decisions. If the problem cannot be viewed as such a well structured task, then a more explorative approach is required. Clustering algorithms like the *k*-means algorithm[19] or self-organising neural network approaches[17,20] allow the structure of the data to be explored in an unsupervised manner. Once formed, the clusters often provide insight into the natural tendencies of the data, allowing new hypotheses to be tested.

Therefore knowledge discovery, and the entire data mining process, may involve initial descriptive statistics, hypothesis testing, supervised and unsupervised learning of the relationships in the data, and integrating this information with existing business knowledge. In this manner, the pitfalls of data mining described above by Small[15] can be avoided.

Customer retention analysis

In the insurance industry, like many competitive industries, the consumer is free to choose their insurer. Their decision is based on a complex combination of price, service, personal preference, convenience, and many other issues. In this environment, it is particularly important to understand which customers are leaving, and why, so that customer retention initiatives can become more focussed. The aim of the analysis presented in this section is to determine reasons for policy termination, and develop a tool for predicting the probability that a policy holder will terminate their policy. This tool can then be used for evaluating the impact of changes to the policy details including premium costs, and can aid in developing more accurate estimates of termination rates for budgetary planning. Furthermore, the tool will allow policy holders who are likely to terminate their policies to be identified prior to their termination, which can provide an opportunity for customer retention initiatives such as direct marketing.[21]

Data collection and preprocessing

Before the analysis could commence, suitable data needed to be obtained which reflected the aims of the analysis. The sample comprised 20914 motor vehicle policy holders whose policies were due for renewal in April 1998. All of these policy holders were contacted by letter in the previous months and notified that their policy was due for renewal and quoted the premium for the next year of cover. 7.1% of the sample did not renew their policy, and are said to have terminated their policy. It is assumed for this analysis that the behavior of policy holders whose policy was due for renewal in this month is representative of all policy holders. (If this assumption is proved invalid then a separate analysis would be required for each month or quarter.) Details on each policy holder included some demographic information (age group, postcode, etc.), policy details (premium, sum insured, etc.), policy holder history (rating, years on rating, claim history), as well as information about the differences in premium and sum insured between the current policy and the renewal policy.

In meetings with the insurance company management team we discussed likely reasons for policy termination. It was agreed that there are three main factors which could affect a policy holder's decision to renew or terminate their policy: pricing, service, and the insured value of their car. A preliminary statistical analysis was performed to investigate the likely impact of these factors. This analysis validated the belief that pricing and sum insured play a large role in a policy holder's decision to terminate or renew their policy, but rejected the impact of service through claims. Examining the impact of other factors such as age, new business, and policy duration, failed to reveal a significant difference between those who renew and those who terminate their policies. It is more likely that many of these factors have an impact when they are combined, rather than individually affecting the decision. In the following section we use data mining as a tool for learning to model the behavior of policy holders based on their demographic and policy information.

Data mining approach

In this analysis we have used the SAS Enterprise Miner software. The Enterprise Miner is a windows based package which allows many common data mining techniques to be applied and compared. Clustering, regression, decision trees, and neural networks are available. A process flow diagram is constructed from predefined nodes of techniques using a drag and drop approach. The process flow diagram constructed for this analysis is shown in Figure 1.

At the left of Figure 1, the data source is specified as a SAS database file. The Insight node has been added to permit statistical exploration of the data. The Variable Selection node is one of the automation features of the Enterprise Miner, and it determines which of the factors or variables the models will consider, based on the impact these have on the decision to terminate or renew. Naturally, considering variables one at a time can obscure the true relationship because of possible variation in the dependent variable due to other factors. As a preprocessing technique though, this approximate approach serves to eliminate unnecessary variables. Once variables have been selected, they can then be transformed through the Transform Variables node, in an attempt to force all interval variables to take on a fairly normalised distribution and to group continuous valued variables to remove the effects of outliers.

Table 1 shows the results of the Variable Selection and Transform Variables nodes. Several variables were rejected from the analysis as a result of their low χ^2 with the dependent variable. ($\chi^2 < 3.92$ is considered low enough for the Enterprise Miner to reject a variable). Some

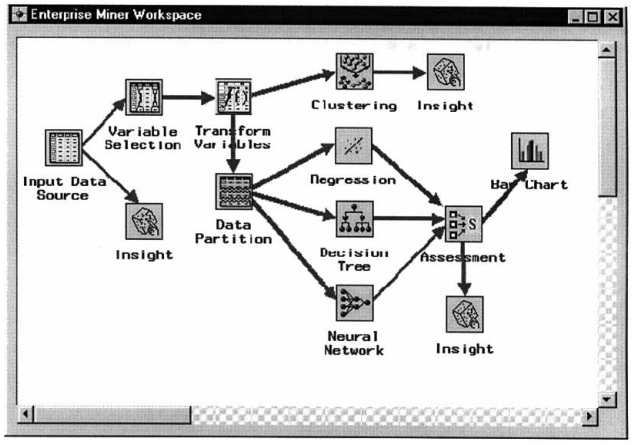

Figure 1 Process flow diagram for customer retention classification.

Table 1 Variables used in the customer retention analysis

Variable	*Data type*	*Status*	*Transformation*
Post code	Categorical	Used	Grouped into 10 bins
New business	Binary	Used	
Vehicle age	Continuous	Used	Grouped into 4 bins
Rating	Categorical	Used	Grouped into 2 bins
Years on rating	Continuous	Used	
Previous company	Categorical	Rejected	
Car category	Categorical	Rejected	
Policy holder age	Continuous	Used	Grouped into 5 bins
Gender	Binary	Used	
Premium	Continuous	Used	Log transformation
Premium diff	Continuous	Used	
Sum insured	Continuous	Used	Log transformation
Sum insured diff	Continuous	Used	
Claim history	Binary	Rejected	
Years on policy	Continuous	Used	
Terminated	Binary	Used	

continuous variables were automatically grouped into bins based on quantiles. Premium and sum insured have been log transformed since their distribution is highly skewed. Categorical data and grouped continuous data are split into binary variables, so for example, the 10 bins produced for the variable PostCode create 10 binary variables. Therefore the total number of variables used for the models is 29 independent variables and one dependent variable (Terminated).

The data is then divided into a training and a test set through the Data partition node. This is important since we need to ensure that our models, whichever they are, do not just memorise the patterns in the data. The test set is used to ensure our findings are valid and can be generalised to enable predictions to be made about new data. Once the data has been partitioned, different modeling techniques can be applied and compared within the Assessment node. In this analysis the data has been modeled by regression analysis, decision trees and neural networks, all using directed data mining to learn to classify policy holders as likely to renew or terminate their policies. The ease of use of commercially available software is an important concern for any organisation assessing the potential of data mining. While we could have experimented with different parameter settings to improve the performance of each modeling technique, we have instead chosen to use only the default values to provide a realistic indication of the kind of results the company is likely to obtain when using such a package as the SAS Enterprise Miner.

For the regression analysis, logistic regression was used with a logit link function. No initial parameter estimates were used, and no interactions between variables were modeled. The decision tree used a splitting criterion based on a χ^2 test with a 0.2 significance level. The minimum number of observations required per leaf was one, with 1% of the observations (209) required for a split search. The maximum number of branches from a node was 2, with a maximum tree depth of 10. Up to 5 splitting rules were saved in each node. The prior probabilities were proportional to the data. The neural network model was the standard three layer feedforward neural network, with 29 inputs (the number of independent variables), 25 hidden neurons, and a single output neuron for the dependent variable. Continuous valued inputs were standardised using the standard normal $N(0, 1)$ distribution. The neurons employed a hyperbolic tangent activation function. Rather than using the common backpropagation learning rule for adapting the weights, the default learning rule utilises a multiple Bernoulli error function and the weights are adapted to minimise the total error using a conjugate gradient technique.

Figure 2 shows a graph representing the comparative performance of each of the three techniques when attempting to classify terminated policies. These results are based on the performance of the test set (or hold-out sample), which was randomly extracted as 20% of the entire data set using stratified sampling. Assuming that the policy holders have been ranked and ordered according to their likelihood to terminate their policy, the graph is a *lift chart* that depicts the percentage of the total terminations that would be discovered if only a certain percentage of policy holders were examined. The horizontal axis represents the percentage of the ranked data set that is under consideration, and the vertical axis gives the corresponding percentage of the total terminations that are found in this group. Ideally, a good algorithm would discover almost all of the terminations without needing to examine the entire data set.

In Figure 2, the neural network is represented by the top (white) line, meaning that it provides the most accurate results. As an example, if we rank all of the policy holders according to the decision of the neural network, with likely terminations first and likely renewers last, and we examine

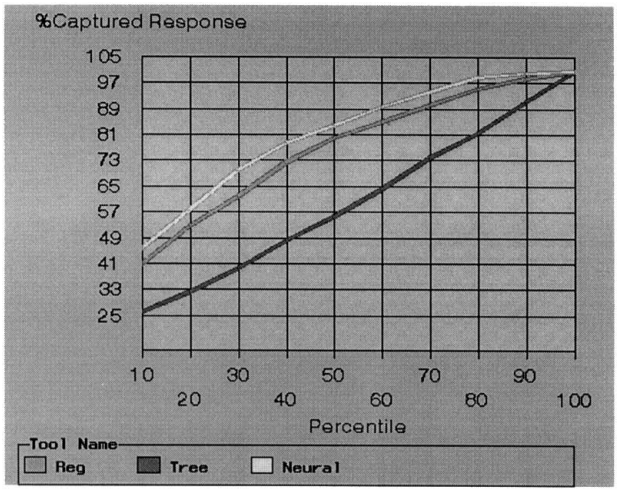

Figure 2 Comparison of neural network, logistic regression and decision tree models for customer retention analysis.

only the top 10% of this list, we will discover nearly 50% of all terminations. The regression and decision tree models would only identify 40% and 28% of the terminations respectively. The further down the ranked list we examine, the more of the terminations we will discover, until we find 100% of the terminations when we examine the entire data set. This concept is particularly useful for direct marketing campaigns, where it is expensive to mail an entire list of customers. It is better to only send mail to those who are likely to respond. The graph is a useful representation to show the comparative predictive powers of the different techniques (using the default parameter settings of the SAS Enterprise Miner).

Results

The previous section showed that the neural network provides the best results for classifying policies as likely to terminate or renew based on the test set or hold-out sample performance and given the automatic method of selecting and transforming variables used by the software. It should be noted that automatic methods of model building seldom find the optimal model because the number of potential models exceeds our ability to comprehensively search the model space. Table 2 shows the accuracy level of the best found neural network model, where a policy is classified as terminated if the likelihood or probability determined by the neural network exceeds 0.5. These results are presented for the entire data set, since similar results were obtained for both the training and test sets, indicating good generalisation of the results. The row and column accuracies of the test set are indicated in brackets.

While these results are quite good, the neural network is reluctant to classify a policy as terminated. Perhaps however, if the neural network outputs a probability of termination around 0.1 we should interpret this as a termination, rather than 0.5. Exploration of the effect of the decision threshold is shown in Table 3 using a probability of 0.1 as the decision threshold. The result is that more of the actual terminations are classified as terminations, but unfortunately the eagerness of the network to classify policies as terminations means a loss in accuracy for the actual renewed policies.

Varying the effect of the decision threshold, which can also be viewed as altering the network's desire to classify a policy as terminated, is useful for exploring how the desired level of accuracy can be achieved. The optimal decision threshold depends upon the purpose of the exercise. In an advertising campaign aimed at enticing likely terminations to renew their policies, we would like to make sure we have as many of the likely terminations as possible (high row accuracy), even if that means contacting some policy holders who are likely to renew. In this case, a lower decision threshold is appropriate to make sure we identify as many of the actual terminations as possible. If, on the other hand, premiums were being adjusted for those policy holders identified as likely to terminate their policies, we need to be very careful that those identified are genuinely likely to terminate (high column accuracy). In this case, a higher decision threshold would ensure that the network only classifies policy holders as terminations if it is quite certain.

Depending on the purpose of the exercise, different cost for misclassification can be assigned, resulting in a profit-loss matrix. As an example, the cost of misclassifying a policy holder as a termination when they were actually a renewer would be the cost of any discount that may have been offered to them to entice them to renew. Likewise the cost of misclassifying a policy holder as a renewer when they actually terminated their policy is the cost of the loss of their premium over the next year (less any claims they might have made over the next year). Clearly, if the misclassification costs can be specified for a given purpose, then the optimal decision threshold can be determined to minimise costs or maximise profits.

Table 2 Classification accuracy of entire data set (and test set) based on neural network model with a 0.5 decision threshold

	Classified as renewed	*Classified as terminated*	*Row accuracy*
Actually renewed	19407 (3879)	20 (6)	99.9% (99.8%)
Actually terminated	1112 (223)	375 (74)	25.2% (24.9%)
Column accuracy	94.6% (94.5%)	95.0% (92.5%)	

Table 3 Classification accuracy of entire data set (and test set) based on neural network model with a 0.1 decision threshold

	Classified as renewed	*Classified as terminated*	*Row accuracy*
Actually renewed	17174 (3440)	2253 (445)	88.4% (88.5%)
Actually terminated	731 (150)	756 (147)	50.8% (49.5%)
Column accuracy	95.9% (95.8%)	25.1% (24.8%)	

The impact of this approach to learning customer retention patterns will be revealed in the following sections. Once the claims and pricing analyses are completed, a policy holder can be evaluated for their risk, and their premium can be set accordingly. This analysis can then be used to determine the likelihood that the policy holder will terminate their policy based upon this premium difference and other factors. Additionally, the neural network tool can be used in isolation to provide more accurate estimates of termination rates for budgetary planning.

Claims analysis

Unlike the customer retention analysis performed in the previous section, the analysis of claims is a highly unstructured problem. Consequently, the data mining approach required is different. The open-ended nature of the analysis requires a more explorative and undirected approach to be taken. The data needs to be examined for hidden trends and patterns of behavior. In this case study, recent growth in the number of policy holders has meant that the insurance company has experienced a decrease in profitability. An understanding of how the growth has affected claim costs is paramount to understanding the business decisions which should be taken in order to combat decreasing profitability.

The insurance industry typically assesses the risk of an individual according to the statistical behavior of others sharing the similar characteristics as the individual. Using a standard statistical risk model, we can assume that claims arrive according to an inhomogeneous Poisson process dependent upon policy holder characteristics and environmental factors. The size of a claim is also assumed to follow a known (usually log–normal) distribution. Maximum likelihood methods can then be used to estimate the model.[1] Therefore, these models depend on assumptions about known distributions (which may not be valid) and tend to be based on pre-defined category rating systems for determining an individual's risk and premium.

For a given category structure (for example, a combination of age, gender, rating, postcode, vehicle worth, and driving record), many researchers have investigated how claims can be better modeled using maximum likelihood and other statistical approaches.[22] The approach taken here however is to use undirected knowledge discovery, namely clustering, to allow these category structures to emerge naturally. The approach is data-driven, rather than predefined, and it is hoped that the impact of any changes to the portfolio of policy holders as a result of growth will be more readily observable using this data-driven approach. Finding the best category structures for a given portfolio of policy holders has been recognised as an alternative approach to the pre-defined categories used by most insurance companies.[22]

This analysis therefore aims to achieve the following:

1. an understanding of where the growth has come from, including a profile of the risks being underwritten;
2. an understanding of the impact this growth has had on profitability, including cost and frequency of claims;
3. a tool for predicting the average claim costs of groups of policy holders;
4. a strategy for increasing profitability based on this analysis.

Key performance indicators

In the insurance industry, profitability is summarily indicated by the *cost ratio*, that is the ratio of claim costs to premiums received. A cost ratio of zero indicates that claim costs were zero and all premiums can be retained. A cost ratio of one indicates that all premiums have been paid out as claim cost, and profitability is zero. Obviously a claim cost greater than one indicates that expenditures are exceeding income. Therefore the cost ratio is defined as:

$$\text{cost ratio} = \frac{\text{sum of claim costs}}{\text{sum of premiums}}$$

The other key performance indicator of note is *claim frequency*. Frequency is measured as the ratio of the number of claims to the number of policy units. If the frequency ratio is one, then there is a claim for every unit on risk. The frequency ratio is defined as:

$$\text{frequency ratio} = \frac{\text{number of claims}}{\text{number of units}}$$

Profitability can be increased by decreasing the average cost ratio across the entire portfolio of policies. This ratio can be decreased in two ways: either the premiums can be increased, or the claim costs can be decreased. While premium increases result in greater profitability (since most policy holders will not claim, and more income has been received), premium increases may result in a higher termination rate as indicated in the previous analysis of customer retention patterns.

Data collection and preprocessing

For this analysis, data was extracted from the data warehouse for all transactions relating to policy holders paying premiums in the first quarter of 1996, 1997 and 1998. Therefore the same group of policy holders are being followed through these years, augmented by new growth and diminished by terminations. The data within each quarter contains information on policy holder characteristics, as well as claim behavior over the preceding 12 month period. Information about each policy holder's contribution to the key performance indicators was added to each row of the data during preprocessing, and was used for statistical

analysis of the clusters only, rather than as an input to the clustering algorithm. This sampling of the portfolio has selected a sub-population of policy holders whose claim behaviors are known. The first quarter was chosen arbitrarily, and we have no reason to believe that policy holders joining the insurance company in quarter one behave differently from the entire portfolio. The temporal analysis performed in this section allows the trends to be observed, and is most suitable for analysing the impact of recent growth. The sample size for each quarter is 114519, 164621, and 212951 respectively for 1996–1998. Note that a different (larger) state has been used for the claims analysis than the customer retention analysis, since this state experienced considerably greater growth.

Analysis and impact of growth

Any attempt to understand the impact of growth on profitability must start with an exploration of where the growth has come from. Descriptive statistics, as well as database queries found that exceptional growth over the last two years has occurred with young people (under 22 years of age), within certain risk areas, and with cars insured over $40 000 (AUD). For each of these categories, the growth rate has been well over 200%, compared to the overall growth rate of 86% over the two year period under consideration. Of course, an increase in the number of young people in the portfolio as well as the number in high risk areas, and with expensive cars, immediately increases the likelihood of claims and puts profitability at risk.

When we query the data sets and examine the key performance indicators for various breakdowns of the portfolio based on the pre-defined category structures used for premium setting, there is not much to be gained. Groups having high costs tend to be the ones anticipated and have correspondingly high premiums. Further analysis of this kind is unlikely to reveal any interesting relationships, particularly when using such an ad hoc query based approach.

Clustering

When using data mining as a predictive tool it is tempting to develop a model for attempting to predict the claim cost of an individual policy holder. This would be pointless, however, since the vast majority of claims cannot be predicted. We are aware of an individual's risk based upon the behavior of the entire group of policy holders like them. We can say that 16–20 year olds are twice as likely to have a claim than 51–71 year olds, but we cannot (and should not expect to be able to) predict which 16–20 year olds are going to claim within the next year.

Therefore a directed data mining approach is inappropriate for this analysis. Instead, we use an undirected clustering of the data. In the insurance industry, premiums are determined using formulae which are based upon a segmentation of the policy holders using demographic information such as driver age, risk area, car age, etc. When we examine the key performance indicators for each segment, the results are fairly uniform. But when we allow the data to tell us the structure it observes, and examine the key performance indicators of these new segments (or clusters), there is quite some variation in the cost and frequency ratios. The inputs used for the clustering are the same as those shown in Table 1, with the dependent variable Terminations omitted. The same transformations were applied to normalise the data and remove the effect of outliers.

When we consider how many clusters to create, we need to find a balance between too few clusters which have little discriminating ability, and too many clusters with few observations in each cluster. After some experimentation, we chose to use 50 clusters using a basic *k*-means algorithm[19] with a least squares criterion. Most of the clusters are not interesting, having a cost ratio of around 0.8 and a frequency ratio of around 0.1 much like the performance of the entire data set. Several clusters, however, have cost and frequency ratios that are significantly different from the average, as shown in the first few columns of Table 4 for quarter 1, 1998. As can be seen, several of these clusters contain policies with cost ratios well above acceptable, while other clusters contain quite low cost ratios. When we apply the same clustering algorithm to the quarter 1, 1996 data (based on the means from quarter 1, 1998) we can see the changing complexion of the clusters, and some of the effect of the growth over the two year period. Some of the more interesting clusters from this perspective include cluster number 26 which has had a large increase in both cost and frequency ratios and a 110% growth. Cluster number 22 has had an increase in cost, though not frequency, while some clusters like numbers 4, 20, 31, 35, and 17 (as well as many of the clusters not shown in Table 4) have been quite stable over the two year period. Due to the confidential nature of the data we cannot disclose the description of policy holders in each cluster, but by exam-

Table 4 Key performance indicators for 10 example clusters, in 1988 and 1996

	Quarter 1, 1998			*Quarter 1, 1996*		
Cluster ID	*Size*	*Cost ratio*	*Frequency*	*Size*	*Cost ratio*	*Frequency*
4	963	0.091	0.051	709	0.113	0.094
20	574	1.154	0.364	365	1.134	0.259
26	874	1.715	0.237	416	1.156	0.208
31	248	0.000	0.000	186	0.07	0.012
36	903	1.117	0.322	900	1.006	0.287
17	1026	1.059	0.288	898	1.231	0.299
22	1485	1.752	0.241	987	1.354	0.235
28	932	1.520	0.215	544	1.100	0.198
34	160	1.974	0.451	150	1.656	0.419
47	963	1.400	0.347	766	1.220	0.308

ining the average values of certain variables in the clusters in 1996 and 1998 we can obtain some insights into how the new growth affects the changing key performance indicators.

The clusters can also be used as a predictive tool. New policy holders can be assessed to determine to which cluster they best belong (based upon maximum similarity), and the predicted claim cost of each new policy holder is then the average claim cost of the cluster. This approach is an alternative to the more traditional viewpoint that calculates predicted claim cost based upon pre-defined structures of the data. Clustering allows the data to tell us about its structures and we then make decisions based upon these evident structures.

The analysis presented in this section has taken a considerably different approach to the previous analysis on customer retention. Predicting which policy holders are going to terminate their policies is a well structured problem. Data mining is successful when solving such problems because there are patterns in the data which can be learned. Many of the variables have a causal relationship with a policy holder's behavior. This analysis into claims is different because it is unlikely that claims can be predicted at the individual policy holder level. Any attempt to force the analysis into a structured format so that modeling techniques like supervised neural networks can be used would be inappropriate and unsuccessful.

Instead, a more explorative and undirected approach has been taken. Clustering has been used to generate segmentations in the data without including information about claims. When the key performance indicators of each cluster are analysed, certain clusters stand out as clearly different from the overall data set. These clusters warrant further investigation and hypothesising, in conjunction with business knowledge. Particularly high cost clusters may provide an alert to fraud, dangerous trends, or inappropriate pricing decisions.

Pricing

In this section we use the previous analyses on customer retention and risks to create a unified pricing methodology. The methodology proposed here for determining appropriate pricing of policies is based upon the interaction of growth, claims and profitability as represented in Figure 3. Information is contained in the data warehouse about the current pricing and profitability, as well as growth patterns and claim patterns. Data mining can be used to analyse this information and to predict future patterns of behavior. The combined information can be used to determine where premiums can be increased to raise profitability without risking policy termination, as well as identifying where premiums could be decreased to retain policies. The overall effect is a methodology for balancing the portfolio of policies in terms of profitability and continued growth and retention.

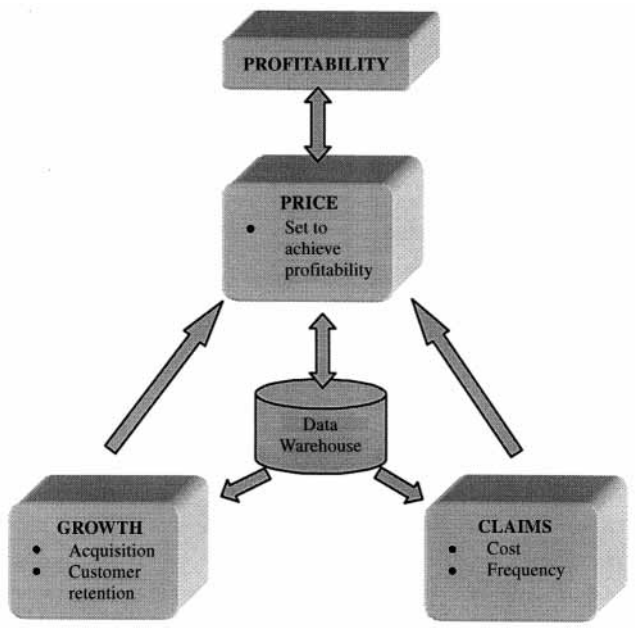

Figure 3 Methodology for price setting to achieve market growth and profitability.

Explicitly, the approach can be divided into four main stages: (1) prediction of claim costs; (2) identification of premiums required to achieve profitability across the portfolio; (3) analysis of likely customer termination patterns; and (4) adjustment where necessary to the offered premium to retain customers while achieving profitability. Each time a group of policies are due for renewal, and before the renewal notices are sent out, the premium being offered for the renewal can be calculated as follows.

Stage 1—prediction of expected claim costs and frequencies

All policies can be passed through the clustering algorithm developed in the claims analysis to arrive at a predicted claim cost for each policy. This is the average claim cost and frequency of the group of policy holders in the determined cluster.

Stage 2—identification of required premium adjustments

Using the premium price of the current policy as the initial default premium, policy holders can be evaluated for their expected cost ratio and frequency ratio. If the cost ratio is acceptable (say between 0.5 and 0.8), then the premiums can remain unaltered. If the cost ratio is either quite low or too high, then premiums can be adjusted to achieve the desired cost ratio and profitability.

Stage 3—classification of customer retention

Of course, this premium adjustment may be considered too much by the policy holder, and termination could result.

The same groups of policies will need to be passed through the neural network model developed in the customer retention analysis. The suggested new premium, together with the premium difference will need to be updated before the data is processed. The neural network will then be able to provide an indication of the likelihood of each policy being terminated as a result of this new premium and other characteristics of the policy. In this case, the decision threshold discussed in the customer retention analysis will need to be set to reflect any current concerns regarding the balance between profitability and customer retention.

Stage 4—price setting

The final premium offered to the policy holder will therefore be the result of several decision processes. If a group of policies under consideration currently have a satisfactory cost ratio, there is probably no need to increase their premiums further. If the cost ratio is too large, a new premium can be determined to achieve the required level of profitability. Whether or not this new premium is the one offered depends upon the likely effect it will have on a customer's decision to renew or terminate their policy. If the cost ratio is currently quite low, premiums could be increased to generate more profit, again within the bounds of tolerance indicated by the customer retention predictive tool. Alternatively, the premiums could be lowered if the price is seen to be affecting customer retention levels.

Clearly, there is a strong relationship between price and retention and growth, which can be used to advantage once it is understood. Likewise, there is a relationship between price and claim costs, known as the adverse selection problem. Unless the price is competitive, low risk customers are likely to terminate their policies in favour of a competitor's cheaper premium, leaving a greater percentage of higher risk policy holders in the portfolio. Therefore it is important to reduce the price where possible to retain low risk customers. This need for diversity (still favouring low risk customers) in the portfolio must be balanced with the goal of profitability. The methodology described in this section is an iterative process designed to find this balance by monitoring key performance indicators over time.

Impacts of this study

Through this case study we have been able to demonstrate the potential of data mining within the insurance industry. We have shown an alternative approach to premium setting and a different way of viewing the data, abandoning the existing pre-defined data structures in favour of a more transparent data-driven approach. Before the methodology can be implemented by the insurance company, there are several key issues that need to be addressed. The insurance company has already restructured its research and development department to focus on data mining, and has engaged in long-term collaborative research with the academic authors of this paper to investigate these issues further.

The first issue relates to the identification of misclassification costs for the customer retention analysis, which will be used to determine the optimal pricing to maximise profit as well as encourage growth and retention. As discussed earlier, these misclassification costs depend on the intended purpose of the knowledge gained from the analysis. Once we know how the knowledge about policy holders' likelihood of terminating their policies is going to be used, we can determine the costs of misclassification, and the optimal pricing level. This issue raises mostly marketing decisions, and reflects the multidisciplinary nature of the insurance industry.

The second issue affecting the implementation of the proposed methodology is the technology transfer that must occur before the insurance company can independently engage in data mining at an operational level. The case study described in this paper was conducted over a 6–8 week period, where much time was spent in the data collection and preparation stages. Staff at the insurance company have since been trained in data mining as an initial step towards transferring the technology of data mining to the company, but there are still issues relating to the real-time operation of the approach to investigate. Pricing decisions need to be made daily as new customers arrive and existing customers' policies become due for renewal. The real-time operation of the system depends on how successfully the methodology described here, and tested for particular groups of the portfolio, generalises so that the models developed can be used across the portfolio. If different models are required for different groups in the portfolio then training, testing and validation of the different models will need to be undertaken before real time implementation is possible.

Finally, many of the other issues related to data mining in the insurance company cannot be tackled without additional information. For instance, early detection of when customers are likely to terminate and why cannot be addressed until a survey is conducted to ascertain the reasons why policy holders terminated their policy. This data is now being collected. The role of the excess in a policy holder's decision to claim must also be better understood before the relationship between price and claim behavior can be established. These are areas of research that we are currently investigating, and the results of which will be used within the data mining process once found.

Conclusion

This paper has discussed a case study from the insurance industry that demonstrates the benefits of data mining to the daily operation. The business problem is the optimal pricing of policies to find a balance between profitability and growth and retention. These often conflicting goals are achieved in

this case study by considering the sub problems of customer retention classification and claim cost modeling.

This case study was divided into three main parts. Firstly, we have examined the potential of data mining for analysing customer retention patterns. Price is found to have a significant impact on a policy holder's decision to renew or terminate their policy. We have also developed a neural network tool for predicting the likelihood of a given policy being renewed or terminated. Secondly, we have used undirected data mining to analyse claim patterns by clustering the data to reveal the natural data structures. Groups with unacceptably high cost ratios (contributing to low profitability) can be identified in this manner. Finally, the results of these two analyses have been combined into a framework for determining the optimal premium price for an individual policy: one that balances the opportunity for profit with the need to retain the customer.

Data mining has proven to be a useful approach for attacking this suite of problems. Each of the sub problems required a different technique and approach, yet the methodology of data mining discussed earlier in the paper provided for a logical fusion of the analysis. Rather than using a single technique for solving a single problem, data mining provides a collection of techniques and approaches that help to fully integrate the solutions with existing business knowledge and implementation.

Data mining is a straightforward application of a methodology and techniques well known to the operational researcher, and as such is likely to become a highly useful and marketable operational research tool in the coming years.

Acknowledgements—The authors gratefully acknowledge the comments and suggestions of the two anonymous referees and the editor. The suggestions have substantially improved the content of the paper, and have provided many ideas for our future research in this area.

References

1 Van Gelder H (1982). Planning and control in insurance. *Eur J Opl Res* **9**: 105–113.
2 Haehling von Lanzenauer C and Wright DD (1991). Operational research and insurance. *Eur J Opl Res* **55**: 1–13.
3 Johnson RA (1997). Digging the dirt. *Best's Rev: Property/Casualty-Insurance-Edition*. **98**: 108–110.
4 Williams GJ and Huang Z (1997). Mining the knowledge mine. The hot spots methodology for mining large real world databases. In: Sattar A (ed). *Advanced Topics in Artificial Intelligence*. Springer-Verlag: Berlin, Germany, pp 340–348.
5 He H, Wang J, Graco W and Hawkins S (1997). Application of neural networks to detection of medical fraud. *Exp Sys with Applic* **13**: 329–336.
6 Gordon RT (1994). An artificial neural network approach to earthquake hazard assessment. In: Dagli CH, Fernandez BR, Ghosh J and Kumara RTS (eds). *Intelligent Engineering System Through Artificial Neural Networks*. ASME Press: New York, **4**: pp 1175–1180.
7 Collins E, Ghosh S and Scofield C (1988). An application of a multiple neural network learning system to emulation of mortgage underwriting judgements. *Proceed IEEE Int Conf on Neural Networks* **2**: 459–466.
8 Hsiung A (1990). Decision making using rule based systems and neural networks. In: Bernold T (ed). *Proceedings 3rd International Symposium. Commercial Expert Systems in Banking and Insurance. Interfaces—Gateway to the Services of the 90s*. SGAICO Technol. Transfer, Lugano, Switzerland; p 117.
9 Nikolopoulos C and Duvendack S (1994). A hybrid machine learning system and its application to insurance underwriting. *Proceedings of the First IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence* **2**: 692–695.
10 Vaughn ML (1996). Extraction of the knowledge base and provision of explanation facilities for a MLP neural network that performs life assurance assessment classification. *IEE Colloquium on Knowledge Discovery and Data Mining* **1**: 1–5.
11 Brockett PL, Cooper WW, Golden LL and Xia X (1997). A case study in applying neural networks to predicting insolvency for property and casualty insurers. *J Opl Res Soc* **48**: 1153–1162.
12 Weber R (1996). Customer segmentation for banks and insurance groups with fuzzy clustering techniques. In: Baldwin JF (ed). *Fuzzy Logic*. Wiley: Chichester, pp 187–196.
13 French M (1998). Mining for dollars: A $6.5 billion market by 2000. *America's Network*. **102**: 24.
14 Bigus J (1996). *Data Mining with Neural Networks*. McGraw-Hill: New York.
15 Small RD (1997). Debunking data mining myths. *Information Week*, January 20: 55–60.
16 Berry M and Linoff G (1997). *Data Mining Techniques*. Wiley: New York.
17 Smith KA (1999). *An Introduction to Neural Networks and Data Mining for Business Applications*. Eruditions Publishing: Melbourne.
18 Rumelhart DE and McClelland JL (eds) (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press: Cambridge, MA.
19 Hartigan JA (1975). *Clustering Algorithms*. John Wiley & Sons: New York.
20 Kohonen T (1988). *Self-Organisation and Associative Memory.* Springer-Verlag: New York.
21 Onn KP and Mercer A (1998). Case study the direct marketing of insurance. *Eur J Opl Res* **109**: 541–549.
22 Samson D (1986). Designing an automobile insurance classification system. *Eur J Opl Res* **27**: 235–241.

Received January 1999;
accepted December 1999 after two revisions