Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006 (pp23-26)

Premium Changes Effects On Insurance Customers Using Neural Networks

ANGELES VARELA[1], HILARIO GÓMEZ MORENO[2], CÉSAR PÉREZ LÓPEZ[3]
[1]Dpto. de Economía, Universidad Carlos III de Madrid, C/ Madrid, 126, 28903 Getafe, Madrid,SPAIN. mavarela@eco.uc3m.es.
[2]Dpto. de Teoría de la Señal yComunicaciones, Universidad de Alcalá, Escuela Politecnica Superior. Campus Universitario, 28805 Alcalá de Henares, Madrid, SPAIN. hilario.gomez@uah.es
[3]Departamento de Estadística e Investigación Operativa III, Escuela Universitaria de Estadística, Universidad Complutense, Avda. Puerta de Hierro, s/n, 28040 Madrid, SPAIN. cesar.perez@ief.minhac.es

Abstract: - This paper examines the use of neural networks for modelling insurance customer retention rates within similar groups. The work is part of a data mining framework for determining optimal premium prices. Clustering is used to arrive at similar groups of policy holders based on insurance company's information. This information is supplemented with premium details, and a neural network is used to model termination rates given premium changes. We have shown that significant improvements in prediction accuracy can be obtained by dividing each cluster to isolate those policy holders with a significant increase in premium. The objective is to determine the optimal premium which reflect the risk of the policy holders, the criteria for grouping has to be similarity in risk. The clusters were grouped according to similarity in claim cost.

Key-Words: - Neural Networks, customer relationship management, customer retention, data mining.

1 Introduction

For many practical applications of classification, the need for a clear explanation of a decision results in a severe limitation when using neural networks. In areas such as insurance companies [1-3], where insurance customers are classified as good or bad before being accepted or rejected for an application, it is unacceptable to use a classification technique that does not provide reasons, to explain the decision. Insurance industry is driven by two main subjects: the need to return a profit to their shareholders, and the need to achieve market growth and retain a certain level of market share. These two goals are imperatives to success, but are often conflicting. Premium prices play an important role in insurance companies to find a balance between these two goals. The challenge is to set a premium prices so expected claims are covered.

Insurance companies have traditionally determined premium prices by assigning policy holders to pre-defined groups and observing the average behaviour of each group. The groups are formed based on insurance companies experience about the perceived risk of different demographic groups of policy holders. With the advent of data warehouses and data mining however comes an opportunity to consider a different approach to premium pricing: one based on data-driven methods. By using data mining techniques, the aim is to determine optimal premiums that more closely reflect the genuine risk of individual policy holders as indicated by behaviours recorded in the data warehouse.

The purpose of this paper is to investigate the second component of the data mining framework: modelling the effect of premium price changes on the customer retention rate.

A case study utilising a database of over 900.000 policy holders is used to evaluate the effectiveness of various techniques within this framework. A strategy for improving the retention rate prediction by dividing the data into more homogeneous groups and using separate neural network models for each group is presented, and the results are compared to a single neural network model. Conclusions are drawn in Section 6.

2 A Data Mining Framework

The framework consists of four main components: identifying risk classifications, predicting claim costs, determining retention rates, and combining this information to arrive at optimal premiums. Firstly, the estimated risk of policy holders must be calculated, and used to determine optimal premium values. The total premiums charged must be sufficient to cover all claims made against the policies, and return desired level of profit. The levels of predicted claims can also be used to forecast profits, when coupled with premium information. However premiums cannot be set at too high a level as customers may terminate their policies, affecting market share. Sales forecasting is determined by marketing information as well as models that predict customer retention or churn rates. When integrated, this information provides a methodology for achieving the two goals of market growth and profitability.

For optimal premiums to be set, the insurance company needs to determine estimated claim costs and the effect of changes in premiums on retention rates. The estimation of claim cost requires an accurate assessment of risk, discussed in the next section.

3 Risk Classification

Insurance companies group policy holders into various risk groups based on factors which are considered predictors of claims. For example additional drivers younger than 25 years old are considered to be a higher risk and so they are charged a higher premium. In designing the risk classification structure, insurance companies attempt to ensure maximum homogeneity within each risk group and maximum heterogeneity between the risk groups. This can be achieved through clustering. In previous work we have shown that the data-driven k-means clustering approach to risk classification can yield better quality predictions of expected claim costs compared to a previously published heuristic approach [4].

The k-means clustering model was used to generate a total of 30 risk categories.

In the insurance industry, risk is measured by frequency or the probability of a claim and the amount of claims. The higher the frequency, the higher the risk. The higher the amount of claims, the higher the risk. The k-means clustering model was able to find clusters which have significantly different claim frequency and claim cost, without being provided with any claim information as input variables. In other words, clustering is able to distinguish between low and high risk groups.

4 Problem Formulation

Neural networks have been used in this paper to learn to distinguish policy holders who are likely to terminate policies from those who are likely to renew. They are an ideal tool for solving this problem due to their proven ability to learn to distinguish between classes, and to generalise their learning to unseen data. Prediction of termination rates or is a significant area of research, particularly in the telecommunications and insurance industries. Several researchers have successfully applied neural networks to churn prediction problems, and to better understand the factors affecting a customer's decision to churn [3-4].

The training set consisted of 270.000 policies with due dates from 1 January to 31 December 1998 while the test set consisted of 140000 policies. The period of overlap was to enable comparison of exposure and retention rates over a one-year period. The training set was used to train the neural networks and while the test set was used to evaluate the results.

In addition to the eleven variables used for risk classification, premium and sum insured information were used as inputs to the neural networks.

A multilayered feedforward neural network was constructed for each of the clusters with 10 inputs, 20 hidden neurons and 1 output neuron. The 10 inputs were :

1. policy holder's age.
2. policy holder's gender.
3. rating of policy holder.
4. numbers of years policy held.
5. category of vehicle.
6. sum insured.
7. vehicle use
8. vehicle age.
9. the vehicle is under finance.
10. change in premium.

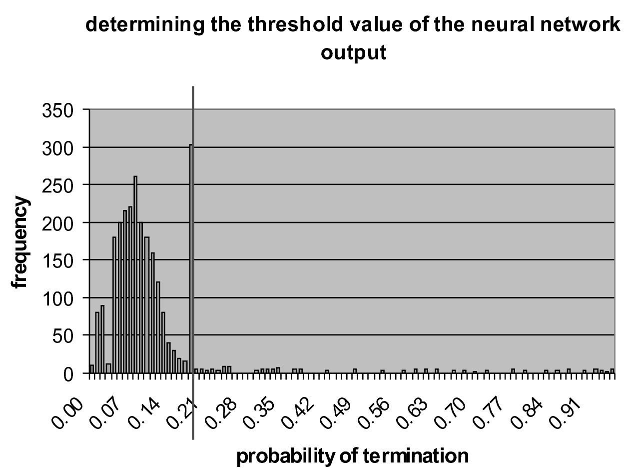

Fig. 1. Determining the threshold value of the neural network output.

The hyperbolic tangent activation function was use. Input variables which were skewed were log transformed. The neural network produces output between zero and one, which is the probability that a policy holder will terminate his policy. Fig. 1 shows the probability of termination of cluster. A threshold value is used to decide how to categorise the output data. For example a threshold of 0.5 means that if the probability of termination is more than 0.5, then the policy will be classified as terminated. In our case, we have set a threshold so that the predicted termination rate is equivalent to the actual termination rate for the whole cluster. For cluster 11, the termination rate, is 14.7% which means a threshold of 0.204 (Fig. 1 marked with a vertical line) is needed to produce the same predicted termination rate.

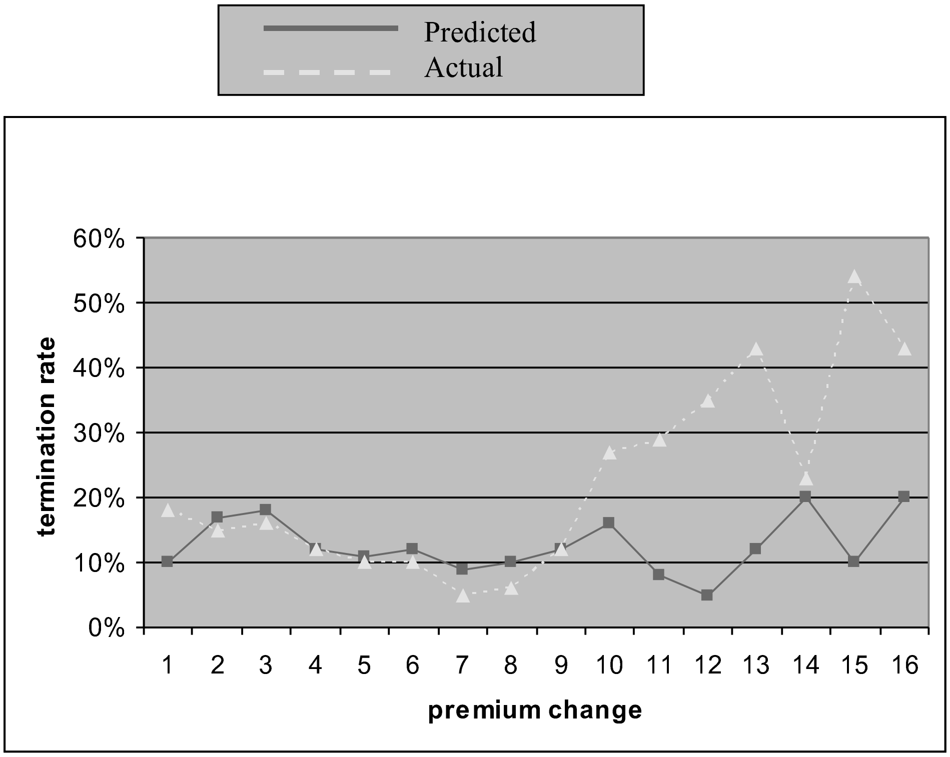

Fig.2. Prediction accuracy for one neural network model

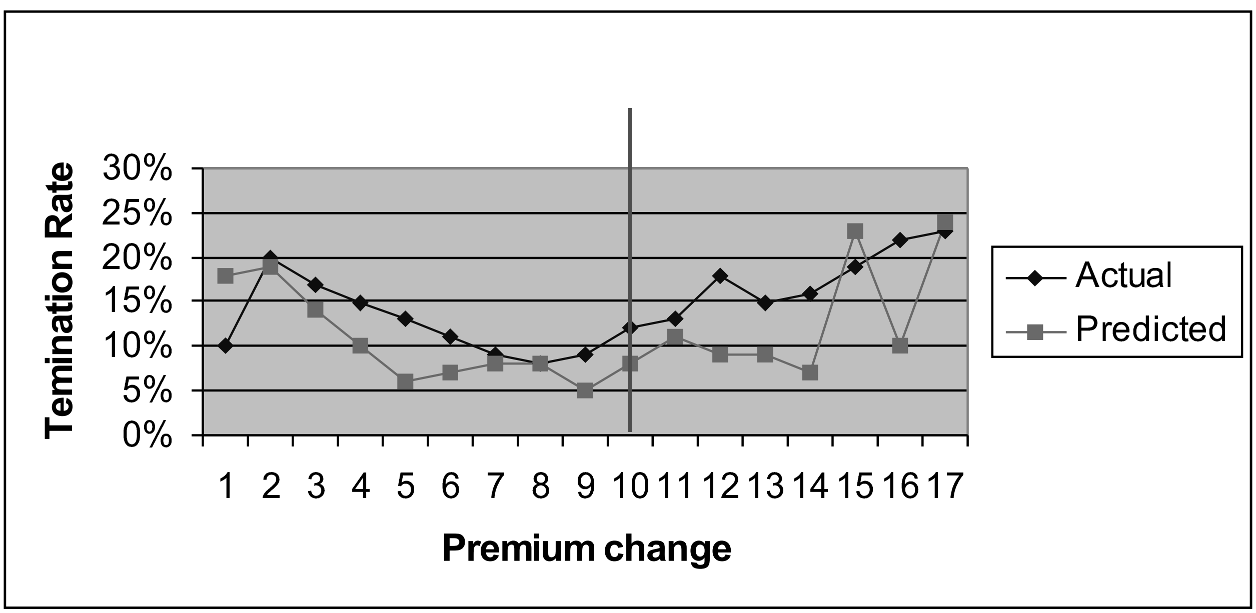

Fig.3. Prediction accuracy for two neural networks model

5 Results

To determine how well the neural networks were able to predict termination rates for varying amounts of premium changes, the clusters were then divides into various bands of premium as follows : decrease in premiums of less than 21%, premium decrease between 14.5% and 21%, premium decrease between 12.5% and 14.5% , etc. In Fig. 2 are shown those results for each band numbered from 1 to 17.

In order to improve the prediction accuracy, the cluster was then split at the point when prediction accuracy starts to deteriorate. Two separate neural networks were trained for each clusters. The prediction accuracy improved significantly with two neural networks.

The average absolute deviation between the actual and predicted termination rates is reduced by employing two neural networks per cluster rather than a single neural network. It appears that a single neural network is unable to simultaneously learn the characteristics of policy holders and their behaviours under different premium changes. In Fig. 3 this result is shown where the first neural net is in the left of the vertical gray line and the other neural net is in the right.

6 Conclusions

This paper examines the use of neural networks for modelling customer retention rates within homogeneous groups. The work is part of a data mining framework for determining optimal premium prices. Clustering is used to arrive at homogeneous groups of policy holders based on insurance company's information. This information is supplemented with premium details, and a neural network is used to model termination rates given premium changes. We have shown that significant improvements in prediction accuracy can be obtained by further dividing each cluster to isolate these policy holders behave differently due to the greater number of these policy holders who have upgraded their vehicles.

In this work we have used neural networks but new classification tools like Support Vector Machines [4] may be used to try new improvements. Besides some techniques to select the best features for classification will be used to reduce time and find better classification rates.

References:

[1] Samson, D. Thomas, H., Linear Models as aids in insurance decision making: the estimation of automobile insurance claims. *Journal of Business Research*, No. 15, pp. 247-256. (1987)

[2] Smith, K.A., Willis, R.J., Brooks, M., An analysis of customer retention and insurance claim patterns using data mining: a case study. *Journal of the Operational Research Society*, No. 51. pp. 532-541. (2000)

[3] Yeo A.C., Smith K.A., Willis R.J., Brooks M., Modelling the Effect of Premium Changes on Motor Insurance Customer Retention Rates Using Neural Networks, *Proceedings of the International Conference on Computational Science-Part II*, p.390-399, May 28-30, 2001

[4] Yeo A.C., Smith K.A., Willis R.J., Brooks M. , Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry, *International Journal of Intelligent Systems in Accounting, Finance & Management*, Vol. 10, N. 1, March, pp. 39-50 (2001).

[5] Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag, New York (1995)