A Comparative Study of Data Mining Algorithms in the Prediction of Auto Insurance Claims

K.P.M.L.P. Weerasinghe[1]and M.C. Wijegunasekara[2]

Department of Statistics & Computer Science, University of Kelaniya

Abstract

Insurance claims are a significant and costly problem for insurance companies. The prediction of auto insurance claims has been a challenging research problem for many auto insurance companies. Identifying the risk factors which are affected for the high number of claims and denying them may lead to increased corporate profitability and keep insurance premiums at a below rate. The key objective of conducting this study is to examine the data mining techniques in developing a predictive model in support of auto insurance claim prediction and a comparative study of them. The research was carried out by using Artificial Neural Network (ANN), Decision Tree (DT) and Multinomial Logistic Regression (MLR) to develop the prediction model. The results indicated that the ANN is the best predictor with 61.71% overall classifier accuracy. Decision tree came out to be the second with 57.05% accuracy and the logistic regression model indicated 52.39% accuracy. Parameters of optimal NN model gives 6 input neurons and 7 minimum hidden neurons with 0.15 learning rate. The comparative study of multiple prediction models provided us with an insight into the relative prediction ability of different data mining methods. The comparison of the results of the decision tree and neural network models showed an interesting pattern. Policies that are misclassified by one model are correctly classified by the other. This might be an indication that the combination of the models could result in a better classification performance.

Keywords: ANN, Auto Insurance, Data mining, Decision Tree, Multinomial Logistic Regression

Introduction

The insurance industry has historically been a growing industry. It plays an important role in insuring the economic wellbeing of one country. Insurance claims are a significant and costly problem for insurance companies in all sectors of the insurance industry. Insurance companies typically employ a claims investigation unit to investigate the factors affected for the claims. Identifying the risk factors which are affected for the high number of claims and denying them may lead to increased corporate profitability and keep insurance premiums at a low rate. 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.

In the insurance industry, there is a huge volume of raw data associated with policy and claims information. These historical data provide the greatest source of information on claim exposure and is the starting place for insurance claim modeling. The incapability of human being to interpret and digest the accumulated data and make use of them for decision-making has created a need for development of new tools and techniques for automated and intelligent huge database analysis. As a result, the discipline of knowledge discovery or data mining in databases, which deals with the study of such tools and techniques, has evolved into an important and active area of research.

Statement of the Problem

In real world applications, classifying particular situation or events as belonging to a certain class, is very important. As an example, predictive modeling in insurance claims can be expressed as a classification problem. We must build an accurate classifier system or model in order to solve such problems. We can use data mining techniques to address such kind of problems. While the traditional data analysis techniques have become inefficient to handle huge data sets, they are also based on the prior assumptions on data. In overall, data mining is more about the search than confirmation of hypotheses. Hence, data mining is not only concerned with algorithmic capabilities, but it also provides tools to accomplish analyzes without strong assumptions or knowledge on the data. Due to the explorative and descriptive nature, intelligible representation and visualization of the found patterns and models are essential for the successful mining process, particularly when the domain expert has limited knowledge of the data mining methodology. This research describes the development of predictive models, which determines the number of claims exposure of motor insurance policies. This research on predicting auto insurance claims is conducted according to the two level knowledge mining (KM) process instead of the traditional KDD model. The KM model provides well-defined interface for domain and method experts. The steps of the process are shown in the Figure 1.

Knowledge Discovery

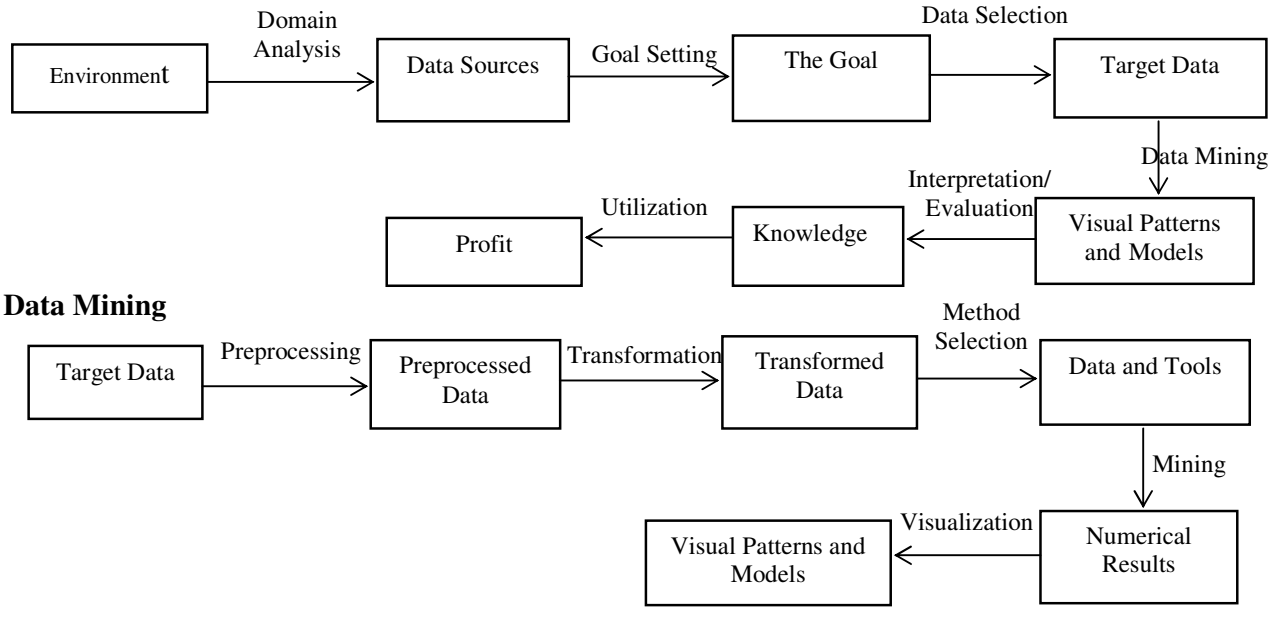

Figure 1: Knowledge Mining Process

Objective of the Study

The key objective of conducting this study is to examine the data mining techniques in developing a predictive model in support of auto insurance claim prediction and a comparative study of them.

Review of Literature

Despite the wide variety of data mining applications, not so many research or development efforts in the context of prediction of insurance claims have been made. In this section a set of research efforts from data mining techniques are reviewed. Chong et al present interesting results for different machine learning techniques (neural networks, decision trees, support vector machines and hybrid decision tree- neural network method) on the automobile accident data set that is collected from the United States. Their results show that hybrid decision tree- neural network approaches outperforms the single classifiers in traffic accident classifier learning.

An attempt has been made by Askale (2001) to assess the application of data mining technology in support of loan disbursement activity at Dashen Bank, one of the private commercial banks in Ethiopia.

The other research, which is undertaken by Gobena (2000), on the possible application of data mining techniques that would help in forecasting flight revenue information in the airline industry (specifically Ethiopian Airlines) is another endeavor, which has been made so far. The results of both previous research works made in these areas have been encouraging.

Andrea Dal Pozzolo (2010) investigates how data mining algorithms can be used to predict Bodily Injury Liability insurance claim payments based on the characteristics of the insured customer's vehicle. The algorithms are tested on real data provided by the organizer of the competition.

Methodology

Preprocessing the input data set for a knowledge discovery goal using a data mining approach usually consumes the biggest portion of the effort devoted in the entire work. The collected data has been preprocessed and transformed in to a form suitable for the particular data mining software used in the study. The data preparation involved discretization, removing outliers and summarization of data.

After preprocessing the data, preparing each variable, and casting aside those variables which will not be helpful in the model, the data set is divided into three approximately equal parts. Two of these, commonly called the "train" and "validation" sets, are for model building, while the "test" is placed aside until the end of the process where it will be used to assess the results.

Here investigated three data mining techniques: Neural Network, Decision Tree and Multinomial Logistic Regression. In this study, a multi-layer network with back propagation (also known as multilayer perceptron) is used. Second technique is the C4.5 decision tree algorithm. C4.5 based on the ID3 algorithm. Using SPSS and Weka software Multinomial stepwise (Backward elimination) logistic regression, decision tree and neural network models were built. Model buildings have been made iteratively using different decision tree and neural network parameter settings until an acceptable result was obtained. Both tools have the facility to partition the dataset randomly into training and testing sets. Number of claims was classified into one of the three possible groups (Low, Fair, and High).

Data Collection and Analysis

Our first step was to attempt building the model using logistic regression, the traditional statistical modeling approach for analysis of data with multinomial response. MLR is a well-known classical technique and is easily implemented in a number of software packages. Stepwise MLR analysis removes the variables which

have no relationship with the claim level. .Multinomial stepwise logistic regression (backward elimination) gives the following model.

Z= 0.940 - 0.865 ([CoverageLevel=1]) - 1.123 ([ClaimType=2])

Z = log (P(Claim Level=2)/P(Claim Level=1))

Z= - 0.348 - 1.498 ([EducationLevel=1]) - 1.141 ([EducationLevel=2]) - 1.087 ([EducationLevel=4]) - 0.819 ([CoverageLevel=3])

Z = log (P(Claim Level=3)/P(Claim Level=1))

[CoverageLevel=1] - (Bodily injury liability)
[ClaimType=2] - (Hit another person/ object)
[EducationLevel=1] - (<9)
[EducationLevel=2] - (Up to O/L)
[EducationLevel=4] - (Diploma)
[CoverageLevel=3] - (Collision cover)

Decision Tree Analysis

The tree was built by changing the parameter 'minimum size of leaves' and using 0.25% confidence level. From each sample 80% records were taken for the training and 20% records were taken for testing. Table 1 indicates the parameters, which are given from optimal decision tree algorithm.

Table 1: Parameters of optimal tree

Min. Size of leaves	Number of leaves	Number of nodes
3	55	42

Neural Network Analysis

In ANN, the output layer contains three neurons; one for each claim level. A one-hidden-layer network is used in this study. Formula used for data standardization is, (x- x_avg) / x_std_dev where x is the selected variable. The learning rate varies in the range of 0.15 to 0.30 in step of 0.05. Table 2 indicates the parameters, which are given from optimal Neural Network algorithm.

Table 2: Parameters of optimal NN model

Input neurons	Minimum hidden neurons
6	7

According to the Table 3, each algorithm gives different risk factors. Thus, the Neural Network shows positive correlation between the variables marital status, vehicle type and age with high claim level by indicating that the married drivers who are less than 38 years old and have vehicle types like bike, bus, have higher claim exposure.

Table 3: Risk factors give from each algorithm

Logistic Regression	Decision Tree	Neural Network
Education_ level	Marital_status	Education_level
Coverage_ level	B_Vage	Vehicle_type
	Vehicle_type	Coverage_level
	Credit_rating	Claim_type
	Claim_type	Marital_status
	Coverage_level	Credit_rating
	Education_level	

Results and Discussion

In this study, the accuracy of three data mining techniques is compared. The goal is to have high accuracy, besides high precision and recall metrics. Although these metrics are used more often in the field of information retrieval, here we have considered them as they are related to the other existing metrics such as specificity and sensitivity. These metrics can be derived from the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics.

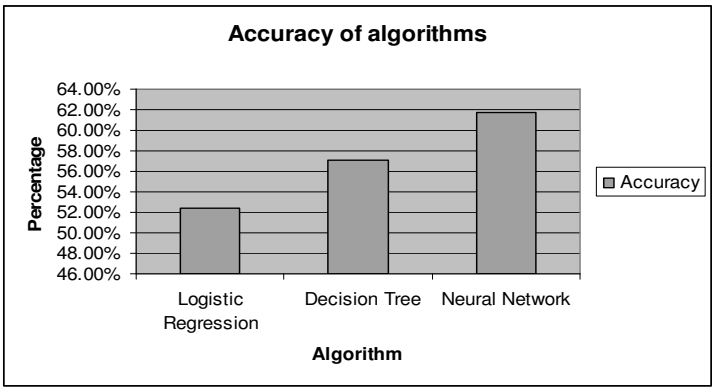

Figure 2: Accuracy of algorithms

In Figure 2, the neural network model outperforms all the other models. The neural network model correctly classifies 61.71% of the total cases. The decision tree model follows, by correctly classifying 57.05% of the total cases and logistic regression achieves an overall performance of 52.39%.

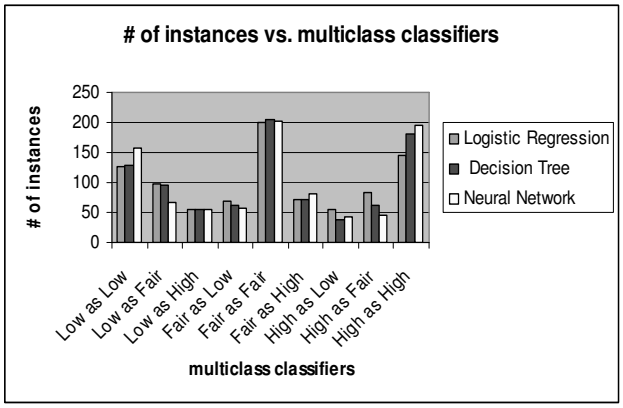

Figure 3: Number of instances vs. multi class classifiers

The number of low level claims and high level claims that has been classified correctly is, however, small than the fair level claims. One possible reason for the misclassification of those claims level might be that the group of low and high level claims is not heterogeneous. Although low and high claims have certain factors, which do not make them fair claim levels. The other possible reason is that, it is not sure whether a claim is fair, and uses the low and high level class as a 'immediate' class. Therefore, a claim is classified, as low or high while it is actually fair level claim. In this case it is not strange that the model shows slightly low performance on low and high claim levels.

Comparison of Models in each Class

In a classification problem the output variable, also called class variable, is a factor taking two or more discrete values. The input variables can take either continuous or discrete values. In this claim prediction problem it has three-classes whereby instances are labeled as Low, Fair and High. A perfect classifier would have zero False Positive and zero False Negative meaning that all the instances are well classified. Table 4 indicates the classifier accuracy of Low, Fair and High claim levels.

Table 4: Classifier accuracy of Low, Fair and High level claims

Low level class classifier diagnostic			
Parameters	**LR**	**DT**	**NN**
Precision	50.06%	55.65%	61.33%
Recall	45.32%	46.04%	56.47%
Specificity	80.25%	83.63%	84.10%
Fair level class classifier diagnostic			
Precision	52.63%	56.20%	64.45%
Recall	59.00%	60.18%	59.88%
Specificity	67.97%	71.70%	80.07%
High level class classifier diagnostic			
Precision	53.68%	59.09%	59.39%
Recall	51.41%	64.08%	69.01%
Specificity	79.58%	9.58%	78.28%

In low level class classifier diagnostic, the neural network model correctly classify low level cases as low level with accuracy of 61.33% and there has a 84.10% of which does not classify low level class claims as fair or high. In fair level class classifier diagnostic, the decision tree model correctly classify fair level cases as fair level with accuracy of 56.20% and there has a 71.70% of which does not classify fair level class claims as low or high. In high level class classifier diagnostic, the neural network model correctly classify high level cases as high level with accuracy of 59.39% and there has a 78.28% of which does not classify high level class claims as fair or low.

Among all the results we can conclude that the neural network model has the best prediction accuracy in high and low levels of claims and the decision tree model has the best prediction accuracy for fair level claims in an auto insurance, which both of the models has the ability of predicting the claims in the drivers claim database.

Decision Tree technique provides insights into the decision-making process, which explains how the results come about. The decision tree is efficient and is thus suitable for large data sets. Decision trees are perhaps

the most successful exploratory method for uncovering deviant data structure. Trees recursively partition the input data space in order to identify segments where the records are homogeneous. Although decision trees can split the data into several homogeneous segments and the rules produced by the tree can be used to detect interaction among variables, it is relatively unstable and it is difficult to detect linear or quadratic relationships between the response variable and the dependent variables.

Conclusion

The ANN model performed the best among three models. This may reflect the fact that neural networks can benefit from more extensive data preparation as well as a greater degree of experimentation with model parameters. The actual claim exposures with which the outcomes of the neural network model have been compared are satisfactory, it can be concluded that the model is quite able to identify most claims with an excessive claim exposure. Though logistic regression model gives less accuracy it remains the clear choice when the primary goal of model development is to look for possible causal relationships between independent and dependent variables, and one wishes to easily understand the effect of predictor variables on the outcome. The comparison of the results of the decision tree and neural network models showed an interesting pattern. Policies that are misclassified by one model are correctly classified by the other. This might be an indication that the combination of the models could result in a better classification performance.

References

[1]. Applying Data Mining Techniques in Property Casualty Insurance, [online], Available: http://www.casact.org/pubs/forum/03wforum/03wf001.pdf.

[2]. Mining road traffic accidents, [online], Available: https://docs.google.com/mining_road_traffic_accidents.pdf.

[3]. Comparison of Data Mining Techniques for Insurance Claim Prediction, [online], Available: http://www.ulb.ac.be/di/map/adalpozz/pdf/Claim_prediction.pdf

[4]. Mining insurance data for fraud detection-final.doc, [online], Available: https://docs.google.com/mining insurance data for fraud detection-final.doc.

[5]. Predicting Breast Cancer Survivability Using Data Mining Techniques, [online], Available: http://www.siam.org/meetings/sdm06/workproceed/Scientific%20Datasets/bellaachia.pdf.

[6]. 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.

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

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

[9]. 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.

[10]. 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.

[11]. 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.