Journal of Data Science **10**(2012), 537-561

A Comparative Analysis of Decision Trees Vis-à-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection

Adrian Gepp[1], J. Holton Wilson[2], Kuldeep Kumar[1] and Sukanto Bhattacharya[3*]
[1]*Bond University,* [2]*Central Michigan University and* [3]*Deakin University*

Abstract: The development and application of computational data mining techniques in financial fraud detection and business failure prediction has become a popular cross-disciplinary research area in recent times involving financial economists, forensic accountants and computational modellers. Some of the computational techniques popularly used in the context of financial fraud detection and business failure prediction can also be effectively applied in the detection of fraudulent insurance claims and therefore, can be of immense practical value to the insurance industry. We provide a comparative analysis of prediction performance of a battery of data mining techniques using real-life automotive insurance fraud data. While the data we have used in our paper is US-based, the computational techniques we have tested can be adapted and generally applied to detect similar insurance frauds in other countries as well where an organized automotive insurance industry exists.

Key words: ANNs, decision trees, fraud detection, logit model, survival analysis.

1. Introduction

The annual cost of settlements from fraudulent insurance claims in Australia was estimated at $1.4 billion dollars in 1997, which added $70 to the annual premium of each insurance policy (Baldock, 1997). These figures are likely to be much higher today as fraud is a growing problem (Morley *et al.*, 2006). While difficult to quantify, the cost is estimated to have increased more than 450% for UK banking card fraud from 1996 to 2007 and also for worldwide telecommunications fraud from 1999 to 2009 (Hand, 2010). The Insurance Council of Australia (ICA) and all major Australian insurance companies[1] are no doubt aware of the cost of

*Corresponding author.
[1]The websites of Allianz, AAMI, SGIC SA, Suncorp and GIO Insurance were visited.

fraud and through their websites are spreading this information to the public to make sure insurance fraud is not viewed as a victimless crime. Online services as well as the ICA's 24-hour telephone hotline are also provided for reporting suspected insurance fraud, almost always with an option for anonymity.

Wilson (2009) has stated that automotive insurance fraud is a global problem. He also outlines that it is costly to automotive insurance businesses, their customers and other general consumers. Insurance businesses have costs from investigating potential fraud and opportunity costs of additional funds that legally must be set aside for all claims. Automotive insurance customers will pay higher premiums and have longer waits for legitimate settlements. This includes more expensive insurance for other businesses, which will in turn be passed on to general consumers through higher prices of goods and services. Thus, the cost to society is much higher than the settlements from fraudulent claims. For example, while fraudulent claim settlements in Australia were estimated at 1.4 billion dollars annually, the total cost to society was estimated to be as high as 9 billion dollars (Baldock, 1997). It should also be noted that insurance companies have further motivations as they can achieve a competitive advantage over rivals by being better at detecting insurance fraud.

Automated statistical techniques for insurance fraud are designed to assist detection of fraudulent claims in a time efficient manner. If successful, this would reduce the costs of fraud outlined in the previous paragraph. However, statistical techniques are inferior to humans at adapting to totally new situations, which do occur in the constantly changing world of insurance fraud, for example, once a specific type of fraud is detected and preventative measures put in place a new type of fraud usually emerges. There is also a risk that those who commit fraud will learn how to conceal their behaviour over time if statistical models are too rigid and predictable. Therefore, statistical techniques should complement, not replace, existing human specialists (Belhadji *et al.*, 2000).

The remainder of this paper is structured as follows. Issues with statistical models for detecting automotive insurance fraud are discussed before a brief look at other research in the field. The main techniques presented in this paper, namely decision trees and survival analysis, are then explained and analysed for automotive insurance fraud detection. This is followed by an explanation of the data and methodology used in a study that empirically assesses these techniques. Following this, we also have included a separate section where we have demonstrated an application of neural networks to a bootstrapped data set based on the same automotive insurance fraud data that has been used in the other four computational techniques; to provide a more exhaustive coverage of our comparative analysis of the prediction performances of alternative computational approaches. After that, concluding remarks on the problem as well as the methods are noted

to round off the paper.

2. Automotive Insurance Fraud Detection

2.1 Some Issues for Statistical Models in the Field

It is important for a fraud detection model to minimise both types of misclassification errors.

- Missing fraudulent claims (Type I error − *Failed Alarm*) is costly in terms of the claim settlement and moreover the success of the fraud might encourage more fraudulent claims. Baldock (1997) estimated the proportion of fraudulent insurance claims to be between 3% and 10% and even higher for automotive insurance, but the number of claims rejected as fraudulent was less than 1%;

- Falsely classifying legitimate claims as fraudulent (Type II error − *False Alarm*) produces wasted costs of investigation plus a potential loss in business reputation resulting from slow and poor handling of legitimate claims.

Baldock (1997) found that that the proportion of insurance claims rejected for fraudulence was between 0.1% and 0.75%. This situation, often referred to as "class imbalance" or "needle in a haystack", presents challenges for statistical models. For example, a simple approach that assumes all claims are legitimate will be more than 99% accurate because of the low rate of fraudulence, but such a model is useless in practice. Bolton and Hand (2002) demonstrates this point with the following example. If 0.1% of claims are fraudulent, then even with a model of 99% accuracy in classifying fraudulent and legitimate claims then only 9 out of 100 claims classified as fraudulent would indeed be so, which is large amount of costly Type II error. Overall, the class imbalance and unequal misclassification costs is important information that must be considered when developing and testing models.

To maintain initial accuracy levels, models implemented in industry will need to be continually updated with new information. This is a challenge as there is a continual flow of new claims. Fan *et al.* (2004) presents a method involving decision trees to handle streaming data that is shown to have good results on a credit card fraud example. In addition to accuracy, automated techniques need to produce classifications in a timely fashion to ensure usefulness.

Another issue relevant to developing fraud detection models is the difficulty in obtaining real-world data for legal and competitive reasons (Wilson, 2009; Phau *et al.*, 2005). And even given data mostly shows deemed rather than true

fraudulence, as that is not known in many cases[2]. This means that models are being trained to repeat mistakes made in the past. There is also a flow-on bias when comparing existing and new methods in that the existing methods have an advantage since they contributed to the actual classification of the real-world test data.

3. Introduction to the Research Field

Phua *et al.* (2005) provide an excellent review of automated fraud detection methods that includes references to other review papers such as Bolton and Hand (2002) who review a subset of fraud detection fields. Phua *et al.* (2005) reveal that the majority of fraud literature is credit card fraud, while automotive insurance fraud came in fourth place. Automotive insurance fraud can be categorised further into different types of fraud schemes as discussed in Wilson (2009) and Phua *et al.* (2005).

There are many research approaches to the area of automotive insurance fraud and they overlap, for example, research into new statistical techniques can reveal new explanatory variables as being important discriminators. The research into automotive insurance fraud more recently includes

- Incorporating other theories into statistical techniques, such as optimal auditing theory into a regression model (Dionne *et al.*, 2009);

- Studying the process taken by companies to optimise existing detection methods, such as Morley *et al.* (2006) who found analysing industry practices can improve the implementation of statistical detection methods;

- Discerning what explanatory variables are important, such as Ganon (2006) who refutes previous suggestions that insurance fraud is more likely to be committed by "average offenders" rather than professionals (Phua *et al.*, 2005) with findings that previous indiscretions such as excessive gambling, license suspension, and tax evasion are significant classifiers in a model for automobile fraud detection;

- Using unsupervised statistical approaches, such as principal component analysis (Brockett *et al.*, 2002);

- Using supervised statistical techniques such as logit analysis (Wilson, 2009) and more complex techniques as presented in this paper to classify claims.

[2]This is more of a problem for supervised, rather than unsupervised, statistical techniques.

Additionally, Viaene *et al.* (2002) studied fraudulence in 1399 personal injury protection claims against automotive insurance policies in Massachusetts and found that See4.5 decision trees to be a poor classifier that was outperformed by other techniques such as logit analysis. Phau *et al.* (2004) attained accuracy improvements by combining See4.5, back-propagation artificial neural networks and a naïve Bayes model when applied to more than 15,000 cases of automotive insurance fraud. Performance assessment was conducted with consideration for the different misclassification costs by using a practical cost minimisation approach, and it is interesting to note that See4.5 was a very important predictor as part of the hybrid model. Very recently, Bhowmick (2011) attempted a comparison of DT-based methods with naive Bayesian classification in detecting auto insurance fraud (which is in essence similar to but albeit narrower in scope to what we have done in this work).

Given that a breakthrough will probably not come from applying a technique to one dataset (Hand, 2010), there is much research still to be done on using decision trees in automotive insurance fraud. While See4.5 performed poorly in one study on one type of automotive insurance, it performed well in a larger study with other techniques. Furthermore, its successor See5 is yet to be used in automotive insurance fraud, which has found success in detecting eBay auction fraud (Chau and Faloutsos, 2005) and is generally preferred over See4.5 in healthcare fraud (Li *et al.*, 2008). Other decision trees such as CART are also yet to be applied to automotive insurance fraud, but have outperformed See5 in other areas such as predicting business failure (Gepp and Kumar, 2008).

Survival analysis techniques that have been used extensively in analysis of medical treatments and shown promise in predicting business failure, are new to automotive insurance fraud and other areas of fraud detection. Insurance fraud studies have been criticised for a lack of time information in data (Phau *et al.*, 2005), such as time-dependent explanatory variables or time-series data, which is interesting as one of the features of survival analysis models is the ability to exploit temporal information.

4. Introduction to, and Analysis of, the Various Techniques

4.1 Survival Analysis

Survival analysis (SA), also known as duration analysis, techniques analyse the time until a certain event. They have been widely and successfully used in biomedical sciences (Kaliaperumal, 2005), but are relatively new to business applications. While other techniques model insurance fraud detection as a classification problem, SA models it as a timeline using functions such as the common survival or hazard function. The survival function $S(t)$ indicates the probability

that an individual survives until time t. When applied to insurance fraud detection, an individual could be a policy owner and survival represents no fraudulent claims being made (or alternatively an individual could be modelled as a policy). Contrastingly, the hazard function $h(t)$ indicates the instantaneous rate of death or fraudulence at a certain time t.

There are many different SA techniques including regression-based models that are well suited for making predictions. These regression-based models define relationships between one of the *descriptor functions* (usually survival or hazard) and a set of explanatory variables. The most prominent is the semi-parametric proportional hazards (PH) model defined by Cox (1972), but there are alternatives such as fully-parametric PH models, accelerated failure time (AFT) models and Aalen's additive model. Cox's PH model is defined as follows:

$$h(t) = h_0(t)e^{\boldsymbol{X}'\boldsymbol{\beta}+c}. \tag{1}$$

- $h_0(t)$ is the non-parametric baseline hazards function that describes the change in the hazard function over time. The flexibility from not having to specify the hazard distribution is one of the key reasons for the model's popularity; and,

- $e^{\boldsymbol{X}'\boldsymbol{\beta}+c}$ describes how the hazard function relates to the explanatory variables ($\boldsymbol{X}$) and is the parametric part of the model, where $\boldsymbol{\beta}$ is a vector of variable coefficients and c a constant estimated by a method very similar to the maximum likelihood method as described by Kalbfleisch and Prentice (1980). Once statistical significance has been established the size of a variable's coefficient does indicate the magnitude of its impact.

The survival function is then computed as follows:

$$S(t) = e^{-H(t)}. \tag{2}$$

Here $H(t)$ is the cumulative hazard function from time 0 to t. The proportional hazards assumption of PH models, such as the Cox model, requires that a unit change in an explanatory variable has a constant multiplicative effect over time. For example, a PH model might show that a policy change doubles the hazard rate of a fraudulent claim, but it could not handle a situation where a policy change doubles the hazard rate initially but has a lesser effect in subsequent years if a fraudulent claim is not made within a year. However, the proportional hazards assumption, which is also not required for AFT or Aalen's models, can be alleviated to a large extent by extending the Cox model to include time-dependent explanatory variables. This can be done with modern statistical packages using a variety of functions to relate explanatory variables and time.

Analysis of SA

SA models can incorporate information from time series (or longitudinal) insurance fraud data, while handling delayed entry and early exit from studies. For example, SA models could consider the number of claims made against a policy each year in addition to the average number of claims per year. This means that SA is different from discriminant analysis (DA) and logit analysis (LA) that assume the process of fraud remains stable over time, which is usually not the case (Hand, 2010).

Unlike cross-sectional models, one SA model could make predictions of fraudulent claims both at the time and before the claims are made. Furthermore, both the easily interpretable survival function and hazard function are available for analysis over time. SA models are also able to model time-dependent explanatory variables. This can allow the coefficients of explanatory variables to change over time, which has been found to happen such as in business failure prediction (Laitinen and Luoma, 1991).

SA techniques, particularly the Cox model, can suffer from multicollinearity problems, but these can be easily avoided by using standard forward and backward variable selection procedures. They can also handle differing misclassification costs in the same way as DA and LA. All three techniques can produce a probability of fraudulence, which is then compared with a cut-off value ranging between 0 and 1 to determine whether the classification is fraudulent or legitimate. Usually this cut-off value is set to 0.5 representing equal misclassification costs, but this value can be changed to represent varying misclassification costs.

Once a policy owner makes a fraudulent claim they are considered to be "dead" by the SA model, which will mean they will have to be re-entered into the model using delayed entry if their policy is not cancelled. This might cause implementation hassles. There are also suggestions that SA models are sensitive to changes in the training dataset, so it is important that they are tested on numerous datasets before drawing any general conclusions.

4.2 Decision Trees

Decision trees (DTs), also known as classification trees, are binary trees[3] that assign data to predefined groups. The tree is built by a recursive process from top to bottom using splitting rules. These rules are usually univariate, but the same variable can be used in zero, one or many splitting rules. When applied to classification problems terminal nodes represent classification groups. Figure 1 shows a simple hypothetical DT for automotive insurance fraud detection that classifies claims as either legitimate or fraudulent.

[3]A binary tree means that each non-terminal node leads to exactly two other nodes.

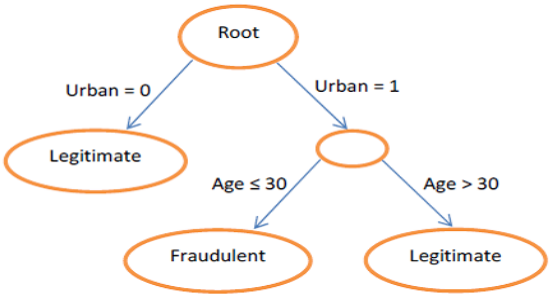

Figure 1: Example DT where "Urban" indicates a rural (0) or urban (1) based policy

Similar to supervised learning with neural networks, DT building algorithms are used to manage the creation of DTs by:

- Choosing the best discriminatory splitting rule at each non-terminal node; and,
- Managing the complexity (number of terminal nodes) of the DT. Most algorithms first create a complex DT and then 'prune' the DT to the desired complexity, which involves replacing multiple node sub-trees with single terminal nodes. Pre-pruning is also possible at the initial creation phase, which creates a simpler tree more efficiently with the risk of reduced accuracy.

Different building algorithms can be used to generate different DTs that often have a large variation in classification and prediction accuracy. Such algorithms include Classification and Regression Trees (CART) (Breiman *et al.*, 1984), Quinlan's Iterative Dichotomiser 3 (ID3) (Quinlan, 1986) and an extension of it called See5, a newer version of See4.5 (Quinlan, 1993).

Analysis of DTs

The major advantages of DTs are that they are non-parametric, can easily model interactions between explanatory variables and are simple to interpret and develop into automated techniques. Unlike parametric DA and LA models, DTs do not need to consider transforming variables as they do not make assumptions about underlying distributions.

The interpretation of DTs is simple with univariate splitting rules and an easy to understand graphical representation. This allows for simple identification of significant variables by comparing their proximity to the root node, where the root node contains the most significant variable[4]. Thus, DTs only identify the

[4]If a variable appears in more than one splitting rule then its significance is measured by the smallest distance to the root node.

relative significance of variables, unlike DA and LA that quantify each variable's significance and impact.

DTs still have the predictive power of a multivariate approach as there are sequences of univariate rules that lead to each classification. Furthermore, these sequences can naturally model interactions between variables without including interaction terms as is required with both DA and LA. Although linear combinations of variables could be used in splitting rules the potentially increased predictive power is not commonly thought to outweigh the increase in complexity and difficulty of interpretation.

DTs can handle missing values and qualitative data (Joos *et al.*, 1998). They can also take different misclassification costs for Type I and Type II Error as inputs, which can then be incorporated into the DT building process at all stages. This is preferable to adjusting the cut-off values after model generation as is done with DA, LA and SA. Arbitrary assignment of cut-off values has been a criticism of DA and LA.

Derrig and Francis (2008) did a fairly exhaustive comparison of DT-based data-mining methods relevant to binomial classification problems. A major disadvantage of DTs is that they do not output the probability of classification as a legitimate or fraudulent claim and consequently no distinction is made between claims in the same classification. DTs building algorithms have also been criticised for not reviewing previous rules when determining future rules (Zopounidis and Dimitras, 1998), but there is no evidence to suggest that this will reduce classification or prediction accuracy ability. Interestingly, DTs also suffer from the same weakness as SA techniques in that their construction is sensitive to small changes in the training dataset (Sudjianto *et al.*, 2010).

4.3 Hybrid Models

Gepp and Kumar (2008) also trialled hybrid DA and LA models that incorporated Cox survival function outputs, but found them to be unsuitable for use on a business failure dataset.

DTs have been used with other comparable techniques, especially domain-specific expert techniques, for forecasting. The primary field of application has been medical diagnostics, for example, Zelič *et al.* (1997) diagnosed sports injures using DTs in conjunction with Bayesian rule-based classification. More recently, Kumar *et al.* (2009) applied a combination of rule-based and ID3 DT case-based reasoning for domain-independent decision support for the intensive care unit of a hospital. Medical cases are often scenario-specific so techniques that combine rule and case based reasoning perform well and interestingly DTs can be used to construct libraries of these cases (Nilsson and Sollenborn, 2004). Chrysler (2005) observed that DTs can also be an efficient method for a knowledge engineer to

use for developing rule-based expert techniques. Bala *et al.* (1995) combined genetic algorithms and ID3 DTs for robust pattern detection and Braaten (1996) showed that cluster analysis can be combined with DTs.

The predictive power of DTs can be further 'boosted' by applying the predictive function iteratively in a series and recombining the output with nodal weighting in order to minimise the forecast errors. Commercial software that can perform DT boosting are available, such as DTREG (http://www.dtreg.com/). Studies using boosted DTs include Chan, Fan and Prodromidis (1999) who used a variant of the AdaBoost DT boosting algorithm for detecting fraudulent credit card transactions and Sudjianto (2010) demonstrating that AdaBoost and LogitBoost can outperform standard DTs in money laundering detection.

Data Analysis and Methodology

This study is designed to empirically assess the suitability of a SA and DT technique new to the area of automotive insurance fraud. Wilson's (2009) study that used LA to detect potential automotive insurance fraud will be extended to include a SA Cox regression and See5 DT, as well as a traditional DA model for comparison purposes.

Dataset

The dataset used for this research is an unaltered copy of the data used by Wilson (2009), which can be referred to for more details. Table 1 shows the main properties of this dataset.

Methodology

DA, LA, Cox and See5 models were developed based on the dataset just described. The in-sample classification ability of all four models was then compared. As a result of their frequency and diverse success, DA and LA serve as excellent benchmarking techniques for the Cox and See5 models. As done by Wilson (2009) the models will not be tested on hold out data because of the small dataset.

Developing complex models reduces their implementation efficiency, interpretation and often its accuracy on new data as the principle of parsimony suggests. Hand (2010) mentions two fraud detection studies that select only a small subset of possible explanatory variables in their chosen model. Although this paper analyses only in-sample classification, it includes only statistically significant variables in final models to assist with future extension of the methodology to include tests on hold-out data. Moreover, the settings used to develop these models are based upon research into business failure prediction that yielded promising empirical results (Gepp, Kumar and Bhattacharya, 2009; Gepp and Kumar, 2008).

PASW Statistics 18 (formerly SPSS) was used to develop the DA, LA and

Table 1: The variables used in the computational techniques to identify the fraudulent claims

Property	**Value**
Source	Initial claim of loss obtained from the Claims Investigation Unit (CIU) of an unnamed US insurance company
Claim type	Stolen and subsequently recovered vehicles
Sample Size	98 total: 49 fraudulent and 49 legitimate
Dependent Variable:	
Fraudulent (1)	Claim denied because it was deemed fraudulent by the CIU
Legitimate (0)	Claims for which there was no involvement of the CIU
Explanatory Variables (6):	
YRS	Number of years the claimant has been a policy owner
CLMS	Total number of claims the claimant has filed with the insurance company
CLMSYEAR	Claims per year calculated as CLMS/YRS where the minimum value of YRS is set to 1 to avoid division by zero
JUA	Boolean variable indicating whether (1) or not (0) the claim is being made on a Joint Underwriting Association policy, which indicates it was placed by the State
NEWBUS	Boolean variable indicating whether (1) or not (0) the claim is being made on a new (less than 1 year old) policy
DATEGAP	Time difference (in months) between insurance claim and the the police report being filed

SA-Cox models with the cut-off values for classification set to 50% indicating equal misclassification costs. Furthermore, all these models were developed using forward stepwise selection methods with the significance level boundaries for variable entry and variable removal set to 5% and 10% respectively.

The Cox model also requires a time (until fraud or legitimate claim) variable that had to be created − all values were set to the same (0) time as the data is cross-sectional. The fact that the data is cross-sectional also means that the PH assumption of the Cox model can't be violated.

The See5 model was developed using Release 2.07[5] with the following settings.

- The default setting of equal misclassification costs;
- The 'minimum cases per leaf node' option was set to 2 to prevent pre-pruning; and,
- The 'pruning CF' option was set to 5%. This controls tree complexity whereby larger values result in less pruning. 'Pruning CF' is expressed as a percentage similar to the significance level for the other models, such that each sub-tree is not pruned if it is significant at the 'pruning CF' level.

[5]Available from http://www.rulequest.com

Results

The resulting LA model as presented by Wilson (2009), which is significant at the 1% level, is:

$$Logit\ Y = -1.135 + 0.671\text{CLMSYEAR} + 1.601\text{NEWBUS}. \tag{3}$$

Here, probability that the ith claim is fraudulent is obtained as follows:

$$P(\text{claim}_i \text{ is fraudulent}) = e^{Logit\ Y_i}/(1 + e^{Logit\ Y_i}). \tag{4}$$

The DA model is also significant at the 1% level with the following equation:

$$\text{Discriminant Score} = -0.717 + 2.341\text{NEWBUS}. \tag{5}$$

Converting the DA score into a probability of fraud is complex compared with LA, but in this case it is analogous to 80.2% chance of fraud if it is a new policy otherwise a 35.1% chance of fraud. Note that if the significance level for variable entry were raised slightly to 5.4% then CLMSYEAR would be included in the model.

The Cox model resulted in the following survival analysis function, which can be interpreted in this case as the probability of a claim being legitimate.

$$S(t = 0) = e^{-0.554e^{0.777\ \text{NEWBUS}}}. \tag{6}$$

So,

$$P(\text{claim}_i \text{ is fraudulent}) = 1 - S_i(t = 0). \tag{7}$$

The model is significant at the 1% level with a significance figure of 0.7%. Interestingly, none of the remaining explanatory variables warrant inclusion in the model even at the 20% significance level.

The See5 model includes only the NEWBUS variable as summarised in Table 2 below, remembering that it does not output probability of group classification. It is also interesting to note that the same tree is generated even if the 'pruning CF' level is raised to 20%.

Table 2: See5 output (the model chose to include only one of the six explanatory variables)

NEWBUS input variable	**See5 DT Classification**
0: not a new policy	0: legitimate
1: new policy	1: fraudulent

In this case, the resulting DA, SA and See5 models are all equivalent with their classifications as represented by the table above. The in-sample classification accuracy of these models is summarised in Table 3 below:

Table 3: In-sample classification results of the DA, SA and See5 techniques compared to logit

	Predicted Classification					
	by DA†, SA and See5			by LA		
Actual	Legitimate	Fraud	Correct	Legitimate	Fraud	Correct
Legitimate	43	6	88%	40	9	82%
Fraud	25	24	49%	20	29	59%
Overall			**68.4%**			**70.4%**

†By increasing the significance level for variable entry to 5.4% and including the CLMS variable in the DA model, the accuracy can be increased to 69.4%.

All the models have produced similar classification accuracy and are superior to a 50% accurate naïve approach of classifying all observations as fraudulent (or legitimate). It is also clear that all the models are better detectors of legitimate, rather than fraudulent, claims. LA is superior in classifying fraudulent claims as well as having slightly superior classification accuracy, but the other models are better at classifying legitimate claims.

Three cases follow that illustrate the use of the four models, which could be undertaken by programming in a spreadsheet or standard computer language. Note that the probabilities always differ between techniques, which indicate that varying the misclassification costs in a larger study might result in significant accuracy differences between the models.

Case 1: a claim for a policy holder who has an average of 1 claim per year and does not represent a new policy. The model outputs are tabulated in Table 4 below:

Table 4: Case 1: classification predictions for each of the four computational techniques

Model	**Probability that claim is fraudulent**	**Predicted Classification**
LA	38.6%	Legitimate
DA	35.1%	Legitimate
SA	42.5%	Legitimate
See5	N/A	Legitimate

Case 2: a claim for a policy holder who has an average of 1 claim per year and does represent a new policy. The model outputs are tabulated in Table 5 below:

Table 5: Case 2: classification predictions for each of the four computational techniques

Model	Probability that claim is fraudulent	Predicted Classification
LA	75.7%	Fraudulent
DA	64.9%	Fraudulent
SA	70.0%	Fraudulent
See5	N/A	Fraudulent

Case 3: a claim for a policy holder who has an average of 2 claims per year and does not represent a new policy. The model outputs are tabulated in Table 6 below:

Table 6: Case 3: classification predictions for each of the four computational techniques

Model	Probability that claim is fraudulent	Predicted Classification
LA	55.2%	Fraudulent
DA	35.1%	Legitimate
SA	42.5%	Legitimate
See5	N/A	Legitimate

Artificial Neural Networks (ANNs)

We have, as a means of providing an even wider coverage of our comparative analysis, also developed, trained and run a back-propagation ANN model having the simplest possible architecture with only one layer of hidden neurons. Of course, more involved architectural variations are possible as is also the prospect of developing an *evolutionarily optimal* network configuration (using e.g., *polyploid* Genetic Algorithm (pGA) optimizer) but we felt that this is best left to a separate research project altogether.

ANNs have sometimes been used in the past in conjunction with other analytical tools – for example; Ohno-machado *et al.* (1995) developed an ANN that estimates survival time more accurately than traditional methods. DTs have also been used in conjunction with ANNs to alleviate some of its black-box nature (Abbass *et al.*, 1999) and extract decision rules without any assumptions about the internal structure (Schmitz *et al.*, 1999). However applications of ANNs as a stand-alone tool in insurance fraud detection haven't been extensively tried before and so the literature is rather thin on this topic. ANNs however have been proposed before as a potent financial fraud detection tool both with respect to financial statement frauds as well as asset misappropriation frauds (Busta and Weinberg, 1998; Bhattacharya, Xu and Kumar, 2011) as they have been observed to fare better in terms of prediction performance on large, complex data sets as compared to linear discriminant analysis and logistic regression models.

A major bottleneck with training and applying ANNs is the size of the data set. ANNs typically work best with large data sets (e.g., data sets having one thousand or more data points) since the data set needs to be partitioned into training, test and blind subsets with the training subset typically consuming at least two-thirds of the entire data set for best learning results (especially in complex data sets). Since our original data set had only 98 data points, we proceeded to bootstrap the data set using a *Monte Carlo* methodology via probability mass functions derived from the actual data distributions of the fraudulent (i.e. "1") and legitimate (i.e. "0") cases. Following the standard Monte Carlo approach, the original data set was bootstrapped to a thousand data points with seven hundred data points in the training set and the remaining in the test set. Effectively this is equivalent to a *random sampling with replacement from the original data set* with the sample size set at 100 and then repeat the process 10 times (7 times to get the training set and 3 times to get the test set).

Our back-propagation ANN architecture is schematically represented in Figure 2 below:

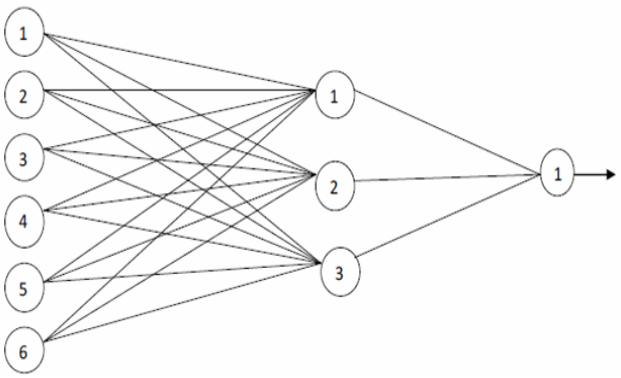

Figure 2: A fully connected, back-propagation, three-layer ANN having a single hidden layer

As stated, our ANN had 3 layers, an input layer consisting of the six input variables, a single hidden layer with three neurons and a single-neuron output layer to record the output as a binary digit (i.e. "1" if the claim is fraudulent and "0" if the claim is legitimate). We used the Neuralyst v1.4^{TM} software (that basically runs as a Microsoft Excel^{TM} add-in) to develop our ANN model as configured above. All network learning parameters were retained as per default *Neuralyst* settings. A hyperbolic transfer function with a floating point calculator was used as it was observed to perform better in terms of the *root-mean-square error* in the training set as compared to the sigmoid, Gaussian and step functions. Results obtained (after 10,000 training epochs) are presented in Tables 7 and 8.

Tables 7 and 8: ANN training & test set predictions for the bootstrapped data (1000 data points)

Network Run Results (Training set)	
0.144803	Root-Mean-Square Error
700	Number of Data Items
619	Number Right
81	Number Wrong
88%	Percent Right
12%	Percent Wrong

Network Run Results (Training set)	
0.575412	Root-Mean-Square Error
300	Number of Data Items
169	Number Right
131	Number Wrong
56%	Percent Right
44%	Percent Wrong

The results show that the ANN was able to detect the underlying pattern in the training data set well enough to get 88% correct predictions (i.e. where a fraudulent claim was correctly categorized "1" and a legitimate claim was correctly categorized "0"). However the test set predictions were not that impressive with the ANN correctly predicting only 169 (i.e. about 56%) of the 300 test data points. Nevertheless even our simple ANN model shows enough promise to be perhaps a good supplementary technique in conjunction with DT or SA. Whether ANNs can be satisfactory "stand alone" techniques remains to be further tested; with other architectures than simply back-propagation.

Re-running the SA and See5 models with a larger (bootstrapped) data set to confirm validity

As the original data set of fraudulent vis-a-vis legitimate claims is of a relatively small size with only ninety eight data points, there can be questions about the validity of the SA and DT prediction results we had obtained earlier. To further confirm the validity of these models, we used a larger dataset (consisting of exactly the same bootstrapped test set of three hundred data points as was used in the ANN model). SA and See5 prediction results with the augmented data set are as follows:

Table 9: DT and SA model predictions with an augmented (bootstrapped) data set (300 data points)

	See5 and SA model predictions		
Actual	Legitimate	Fraud	Correct
Legitimate	134	15	90%
Fraud	76	75	50%
Overall			**69.7%**

The above results show that both the SA and See5 models performed well on new test data without dropping from their training accuracy which corroborates the evidence we got with the original data.

Concluding Discussion

All the LA, DA, SA and See5 models included the NEWBUS variable, which indicates it was the most important in discriminating between fraudulent and legitimate claims. This is consistent with popular incentive strategies designed to encourage policy owners to retain existing policies.

LA had slightly superior classification accuracy, but all four models performed comparably with near 70% overall accuracy. These results support further testing of Cox and See5 models as automotive insurance classifiers, particularly as DT and SA models are known for being sensitive to training datasets. Other techniques can need to be tested, such as the hybrid models mentioned previously along with other SA models and DT models, particularly CART that outperformed See5 in predicting business failure.

Analysis on a much larger dataset is desirable. This would allow for hold-out sample tests as indicated by Wilson (2009). These hold-out sample tests should also have more realistic proportions of fraudulent/legitimate claims rather than the synthetic even split in the data used here. Ideally this larger dataset would also contain a large number of explanatory variables so the nature of fraudulent claims can be better understood. Examples of useful explanatory variables to be included are the number of policies held by a claimant, the regularity of policy changes and details of the claimant's criminal history. It would be advantageous to also include claims referred to the CIU but not subsequently deemed fraudulent as legitimate claims in the training dataset. The reason for this is if models detected these claims before referral then money would be saved by reducing the number of wasteful CIU investigations. Additionally, the use of time-series data might help improve classification accuracy and enable the capabilities of a SA model to be properly tested. The DT and SA approaches introduced in this paper each offer their own advantages. DTs offer an easy to interpret and implement non-parametric model that still has the power of a multivariate approach able to model interactions between variables. Contrastingly, applying SA models in a time-series analysis using both the hazard and survival functions has the potential to reveal more information about automotive insurance fraud, for example, any change in the probability of making a fraudulent claim as the years of policy ownership increase could be useful when designing loyalty programs.

A simple back-propagation ANN model with a 6-3-1 architecture was also trained and run to extend the coverage of our comparative performance analysis of the different techniques. We presented the ANN results separate from the

other four techniques because the ANN model used a bootstrapped data set as ANNs typically perform best with large data sets and our original data set had only 98 data points. However the results show promise in terms of using ANNs as a supplementary method. A larger data set would also make comparisons over varying misclassification costs viable. In the absence of a large, accessible data set of fraudulent automotive insurance claims, we resorted to augmenting the original set via bootstrapping to obtain a bigger data set necessary to effectively train an ANN model. The test data set for the ANN model served as a 'spin-off' to re-test the See5 and SA models and confirm their performance with larger data sets.

To round off, it is unlikely that the field of automotive insurance fraud detection will be advanced by finding that one statistical model that is superior in all situations. However, as more studies are conducted using new data and new techniques, understanding of the process of fraud and the ability to match situations with the most appropriate technique will improve. When implementing statistical models to detect automotive insurance fraud it is important to consider case-specific issues such as resource constraints and to retain human staff in the process to benefit from their superior ability to handle the constant change in the field.

Appendix

The original data set (Tables 10 and 11) used to run our numerical models is provided hereunder as supplementary material for the benefit of future researchers wishing to reproduce/improve on our obtained results.

Table 10: Original data set (identifiers removed) of fraudulent automotive insurance claims

Sorted data (fraud occurred):	yrsmemb	date gap	clms	jua	newbus	clmsyear
1	0	0	1	0	0	0.21
2	0	0	1	0	0	0.25
3	0	0	1	0	0	0.33
4	0	0	1	0	0	0.33
5	0	0	1	0	0	0.35
6	0	0	1	0	0	0.40
7	0	0	1	0	0	0.53
8	0	0	1	0	0	0.59
9	0	0	1	0	0	0.60
10	0	0	1	0	0	0.63
11	0	0	1	0	0	0.67
12	0	0	1	0	0	0.76

Table 10: (continued) Original data set (identifiers removed) of fraudulent automotive insurance claims

Sorted data (fraud occurred):	yrsmemb	date gap	clms	jua	newbus	clmsyear
13	0	0	1	0	0	0.77
14	0	0	2	0	0	0.78
15	0	0	2	0	0	1.00
16	0	0	2	0	0	1.00
17	1	0	2	0	0	1.00
18	1	0	2	0	0	1.00
19	1	0	2	0	0	1.00
20	1	0	2	0	0	1.00
21	1	0	2	0	0	1.00
22	1	0	2	0	0	1.00
23	2	1	3	0	0	1.00
24	2	1	3	0	0	1.00
25	2	1	3	0	0	1.00
26	2	1	3	0	1	1.00
27	3	1	3	0	1	1.00
28	3	1	3	0	1	1.00
29	4	1	3	0	1	1.00
30	4	1	3	0	1	1.00
31	4	1	3	0	1	1.11
32	5	1	4	0	1	1.25
33	5	1	5	0	1	1.40
34	5	2	5	0	1	1.40
35	5	2	5	0	1	1.50
36	5	2	6	0	1	1.61
37	8	2	6	0	1	1.75
38	8	2	7	0	1	2.00
39	9	2	7	0	1	2.00
40	12	3	9	0	1	2.00
41	14	3	9	0	1	2.00
42	17	3	9	1	1	2.00
43	17	4	10	1	1	2.25
44	18	5	10	1	1	3.00
45	26	7	14	1	1	3.00
46	31	7	24	1	1	3.00
47	32	9	25	1	1	3.00
48	33	10	25	1	1	3.00
49	33	52	53	1	1	3.00

Table 11: Original data set (identifiers removed) of valid automotive insurance claims (control set)

Sorted data (no fraud occured):	yrsmemb	dategap	clms	jua	newbus	clmsyear
1	0	0	1	0	0	0.10
2	0	0	1	0	0	0.19
3	1	0	1	0	0	0.20
4	1	0	1	0	0	0.20
5	1	0	1	0	0	0.22
6	1	0	1	0	0	0.23
7	1	0	1	0	0	0.26
8	1	0	1	0	0	0.27
9	1	0	1	0	0	0.27
10	2	0	1	0	0	0.31
11	2	0	1	0	0	0.33
12	2	0	1	0	0	0.33
13	2	0	1	0	0	0.33
14	3	0	1	0	0	0.34
15	3	0	2	0	0	0.40
16	3	0	2	0	0	0.40
17	3	0	2	0	0	0.46
18	4	0	2	0	0	0.48
19	4	0	2	0	0	0.50
20	5	0	2	0	0	0.50
21	5	0	3	0	0	0.56
22	5	0	3	0	0	0.60
23	5	0	4	0	0	0.67
24	5	0	4	0	0	0.73
25	6	1	5	0	0	0.73
26	9	1	5	0	0	0.73
27	9	1	6	0	0	0.75
28	10	1	6	0	0	0.76
29	11	1	6	0	0	0.81
30	11	1	6	0	0	0.85
31	11	1	6	0	0	0.88
32	11	1	6	0	0	0.97
33	13	1	7	0	0	0.97
34	15	1	8	0	0	1.00
35	17	1	8	0	0	1.00
36	17	1	8	0	0	1.00
37	22	1	9	0	0	1.00
38	23	1	9	0	0	1.00
39	23	1	12	0	0	1.00
40	25	1	14	0	0	1.00

Table 11: (continued) Original data set (identifiers removed) of valid automotive insurance claims (control set)

Sorted data (no fraud occured):	yrsmemb	dategap	clms	jua	newbus	clmsyear
41	26	2	14	0	0	1.00
42	29	2	15	0	0	1.17
43	29	2	16	0	0	1.50
44	30	3	19	0	1	1.50
45	31	3	20	0	1	2.00
46	32	4	22	0	1	2.00
47	33	5	26	0	1	2.00
48	35	6	28	0	1	2.33
49	42	7	32	0	1	3.00

Acknowledgements

The authors are deeply grateful to the anonymous reviewer for valuable comments that have helped us greatly in improving the academic quality and general readability of the final version of our paper.

References

Abbass, H. A., Towsey, M. and Finn, G. (2009). C-Net: generating multivariate decision trees from artificial neural networks using C5. University of Queensland, Technical Report: FIT-TR-99-04. http://portal.acm.org/citation.cfm?id=869850

Bala, J., Huang, J., Vafaie, H., DeJong, K. and Wechsler, H. (1995). Hybrid learning using genetic algorithms and decision trees for pattern classification. In *Proceedings of the International Joint Conference on Artificial Intelligence* (*IJCAI*). Montréal, Canada.

Baldock, T. (1997). Insurance fraud. *Trends and Issues in Crime and Criminal Justice*, Paper No. 66. Australian Institute of Criminology, Canberra.

Becker, R. A., Volinsky, C. and Wilks, A. R. (2010). Fraud detection in telecommunications: history and lessons learned. *Technometrics* **52**, 20-33.

Belhadji, E. B., Dionne, G. and Tarkhani, F. (2000). A model for the detection of insurance fraud. *Geneva Papers on Risk and Insurance* **25**, 517-538.

Bhattacharya, S., Xu, D. and Kumar, K. (2011). An ANN-based auditor decision support system using Benford's law. *Decision Support Systems* **50**, 576-584.

Bhowmik, R. (2011). Detecting auto insurance fraud by data mining techniques. *Journal of Emerging Trends in Computing and Information Sciences* **2**, 156-162.

Bolton, R. and Hand, D. (2002). Statistical fraud detection: a review (with discussion). *Statistical Science* **17**, 235-255.

Braaten, Ø. (1996). Artificial intelligence in paediatrics: important clinical signs in newborn syndromes. *Computers and Biomedical Research* **29**, 153-161.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth and Brooks.

Brockett, P., Derrig, R., Golden, L., Levine, A. and Alpert, M. (2002). Fraud classification using principal component analysis of RIDITSs. *Journal of Risk and Insurance* **69**, 341-371.

Busta, B. and Weinberg, R. (1998). Using Benford's law and neural networks as a review procedure. *Managerial Auditing Journal* **13**, 356-366.

Chan, P. K., Fan, W. and Prodromidis, A. L. (1998). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems* **14**, 67-74.

Chau, D. H. and Faloutsos, C. (2005). Fraud detection in electronic auction. In *Proceedings of European Web Mining Forum* (*EWMF 2005*) *at ECML/PKDD.* Porto, Portugal.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

Derrig, R. A. and Francis, L. (2008). Distinguishing the forest from the TREES: a comparison of tree based data mining methods. *Variance* **2**, 184-208.

Dionne, G., Giuliano, F. and Picard, P. (2009). Optimal auditing with scoring: theory and application to insurance fraud. *Management Science* **55**, 58-70.

Fan, W., Huang, Y. and Yu, P. S. (2004). Decision tree evolution using limited number of labeled data items from drifting data streams. In *Proceedings of the Fourth IEEE International Conference on Data Mining.* Brighton, United Kingdom.

Ganon, M. W. and Donegan, J. J. (2004). Self-control and insurance fraud. *Journal of Economic Crime Management* **4**, 1-24.

Gepp, A. and Kumar, K. (2008). The role of survival analysis in financial distress prediction. *International Research Journal of Finance and Economics* **16**, 13-34.

Gepp, A., Kumar, K. and Bhattacharya, S. (2009). Business failure prediction using decision tress. *Journal of Forecasting* **29**, 536-555.

Hand, D. J. (2010). Fraud detection in telecommunications and banking: discussion of Becker, Volinsky and Wilks (2010) and Sudjianto *et al.* (2010). *Technometrics* **52**, 34-38.

Joos, P., Vanhoof, K., Ooghe, H. and Sierens, N. (1998). Credit classification: A comparison of logit models and decision trees. In *Proceedings Notes of the Workshop on Application of Machine Learning and Data Mining in Finance*, 10th European Conference on Machine Learning. Chemnitz, Germany.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, New York.

Kirkos, E., Spathis, C. and Manolopoulos, Y. (2008). Support vector machines, decision trees and neural networks for auditor selection. *Journal of Computational Methods in Sciences and Engineering* **8**, 213-224.

Kumar, K. A., Singh, Y. and Sanyal, S. (2009). Hybrid approach using case-based reasoning and rule-based reasoning for domain independent clinical decision support in ICU. *Expert Systems with Applications* **36**, 65-71.

Luoma, M. and Laitinen, E. K. (1991). Survival analysis as a tool for company failure prediction. *Omega* **19**, 673-678.

Li, J., Huang, K. Y., Jin, J. and Shi, J. (2008). A survey on statistical methods for health care fraud detection. *Health Care Management Science* **11**, 275-287.

Morley, N. J., Ball, L. J. and Ormerod, T. C. (2006). How the detection of insurance fraud succeeds and fails. *Psychology, Crime & Law* **12**, 163-180.

Nilsson, M. and Sollenborn, M. (2004). Advancements and trends in medical case-based reasoning: an overview of systems and system development. In *Proceedings of the 17th International FLAIRS Conference, Special Track on Case-Based Reasoning.* American Association for Artificial Intelligence, Miami, Florida, United States.

Ohno-Machado, L., Walker, M. G. and Musen, M. A. (1995). Hierarchical neural network for survival analysis. In *World Congress on Medical and Health Informatics*, pp. 302-309. Vancouver, Canada.

Phua, C., Alahakoon, D. and Lee, V. (2004). Minority report in fraud detection: classification of skewed data. *SIGKDD Explorations* **6**, 50-59.

Phua, C., Lee, V., Smith, K. and Gaylor, R. (2005). A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review*, 1-14.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning* **1**, 81-106.

Quinlan, J. R. (1993). *C4.5*: *Programs for Machine Learning.* San Francisco, California.

Schmitz, G. P. J., Aldrich, C. and Gouws, F. S. (1999). ANN-DT: an algorithm for extraction of decision trees from artificial neural networks. *IEEE Transactions on Neural Networks* **10**, 1392-1401.

Sudjianto, A., Yuan, M., Kern, D., Nair, S., Zhang, A. and Cela-Díaz, F. (2010). Statistical methods for fighting financial crimes. *Technometrics* **52**, 5-19.

Viaene, S., Derrig, R. A., Baesens, B. and Dedene, G. (2002). A comparison of state-of-the-art classification techniques for expert automotive insurance claim fraud detection. *Journal of Risk and Insurance* **69**, 373-421.

Wilson, J. H. (2009). An analytical approach To detecting insurance fraud using logistic regression. *Journal of Finance and Accountancy* **1**, 1-15.

Zelič, I., Kononenko, I, Lavrač, N. and Vuga, V. (1997). Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries. *Journal of Medical Systems* **21**, 429-444.

Zopounidis, C. and Dimitras, A. I. (1998). *Multicriteria Decision Aid Methods for the Prediction of Business Failure.* Kluwer, Boston, Massachusetts.

Received September 27, 2011; accepted March 5, 2012.

Adrian Gepp
School of Business
Bond University
Gold Coast, Queensland 4229, Australia
adgepp@bond.edu.au

J. Holton Wilson
College of Business Administration
Central Michigan University
Mount Pleasant, Michigan 48859, USA
wilso1jh@cmich.edu

Kuldeep Kumar
School of Business
Bond University
Gold Coast, Queensland 4229, Australia
kkumar@bond.edu.au

Sukanto Bhattacharya
Deakin Graduate School of Business
Deakin University
Burwood, Victoria 3125, Australia
sukanto@deakin.edu.au