Scientific Research Publishing

Theoretical Economics Letters, 2019, 9, 1886-1900
http://www.scirp.org/journal/tel
ISSN Online: 2162-2086
ISSN Print: 2162-2078

A Predictive Modeling for Detecting Fraudulent Automobile Insurance Claims

Hojin Moon*, Yuan Pu, Cesarina Ceglia

Department of Mathematics and Statistics, California State University, Long Beach, CA, USA
Email: *hojin.moon@csulb.edu

How to cite this paper: Moon, H., Pu, Y. and Ceglia, C. (2019) A Predictive Modeling for Detecting Fraudulent Automobile Insurance Claims. *Theoretical Economics Letters*, **9**, 1886-1900.
https://doi.org/10.4236/tel.2019.96120

Received: June 24, 2019
Accepted: August 17, 2019
Published: August 20, 2019

Abstract

Fraudulent automobile insurance claims are not only a loss for insurance companies, but also for their policyholders. The goal of this research is to develop, first, a decision-making algorithm to classify whether a claim is classified as fraudulent or not; and, second, what types of variables should be focused to detect fraudulent claims. To achieve this goal, highly accurate prediction models are built by discovering important sets of features via variable selection algorithms, which can in turn help prevent future loss. In this research, parametric and nonparametric statistical learning algorithms are considered to reduce uncertainty and increase the chances of detecting the appropriate claims. An important set of features for a model is determined by measuring variable importance based on the observed characteristics of a claim via a cross-validation and by testing improvement of the performance at which automobile fraudulent claims are accurately classified using Akaike Information Criterion. We could achieve accuracy above 95% with a set of features selected via a cross-validation. This research would offer some benefit to the insurance industry for their fraud detection research in order to prevent insurance abuse from escalating any further.

Keywords

Classification, Cross-Validation, Prediction Models, Statistical Learning Algorithms, Variable Importance Algorithms

1. Introduction

Fraudulent insurance claims contribute to between 5 and 10 percent of total claims and are costing insurance companies approximately 31 billion dollars annually, with these numbers rising [1]. The current and predicted increase in monetary loss due to automobile insurance fraud is not only a concern for the

insurance companies, but also for the consumer. Boyer [2] quoted a study by the Rand Corporation institute for Civil Justice estimating that in automobile insurance claims, questionable medical claims added between $13 and $18 billion to the nation's total automobile insurance bill in 1993. In order to compensate for the money lost through fraudulent insurance claims, the insurance companies often raise each policyholder's premiums.

Among the many different types of insurance, automobile fraud is found to be the most predominant. Automotive insurance fraud can be brought into many different forms such as staging an accident, the policyholder not involved in the claimed accident, duplicate claims for the same injury, a fake injury, and many other misrepresentations [3]. In order to detect these fraudulent claims, insurance companies and fraud investigators need to know what characteristics lead to a fraudulent claim. Since there are numerous factors and situations that can be attributed to a fraudulent claim, this is a difficult task. It makes more difficult that most insurance companies do not share their claim data with one another, which would collectively enhance the information known about fraudulent claims. Thus, our goal in this paper is to provide a general statistical learning algorithm for building a prediction model for a practical use in an insurance company having heavily lopsided data, not for a particular dataset.

The main goal of identifying fraudulent claims is to find patterns that typically relate to a fraudulent claim. One method in identifying fraudulent claims consists of using cost-related data, such as the cost for vehicle damages and auditing costs [1]. Another common method is to evaluate insurance claim data that are not cost-related, such as the policyholder's demographic and insurance policy information, which is the method that is applied in this paper. Thus, in our study, we are looking for what variables affect the result and the patterns that typically related to a fraud.

Many researchers have used statistical methods for automobile fraud prediction in automobile insurance. A study for fraud detection of an automobile insurance claim was conducted based on a dataset of 1399 personal injury protection (PIP) claims from 1993 accidents collected by the Automobile Insurance Bureau (AIB) [4]. Ciaene, *et al.* [4] used various classification techniques including logistic regression, decision tree, *k*-nearest neighbor, Baysian learning multilayer perceptron neural network, support vector machine, naïve Bayes, and tree-augmented naïve Bayes classification algorithms. For multinomial outcomes, a multinomial logit model was used for fraud detection in data on Spanish automobile insurance claims [5]. On the other hand, in economics application, discrete choice models were used on data claiming for automobile accidents that occurred from 1993 and 1996 in order to detect automobile insurance fraud and misclassified claims [6].

Recently, Wang and Xu [7] proposed a deep learning model for insurance fraud detection that used Latent Dirichlet Allocation (LDA)-based analytics, with the data including both numeric and categorical variables from the Chinese

insurance company claims. Nian *et al.* [8] proposed a new unsupervised spectral ranking method of anomaly (SRA) and illustrated that the spectral optimization in SRA could be viewed as a relaxation of an unsupervised SVM problem. With an auto insurance claim dataset, they provided a solution that the choice of the fraud ranking reference could be made based on whether the cardinality of the smaller class (positive and negative) was sufficiently large, and demonstrated that proposed SRA yielded good performance for a few similarity measures for the auto insurance claim data.

With today's statistical learning algorithms, predictive modeling can more accurately classify a fraudulent claim. There are several predictive modeling methods that could be used in detecting a fraudulent claim in automobile insurance claim data. The most common and fundamental predictive modeling method for classifying fraudulent claims is logistic regression, with an emphasis on variable importance. A goal of this paper is to provide a general statistical learning algorithm best suited especially for highly lopsided data for a practical use in an insurance company. In this paper logistic regression and LASSO (least absolute shrinkage and selection operator) [9] are used for parametric methods. Random Forests [10] is used for a non-parametric ensemble method. Support-vector machines (SVMs, also support-vector networks) [11] are used for kernel-based classification methods. These statistical algorithms are compared based on their performance including classification accuracy and area under ROC curve, balance between sensitivity and specificity, and balance between positive and negative predictive values.

2. Data Description and Preparation

Since the goal of this paper is to propose a general statistical learning algorithm for fraud detection best suited especially for any highly lopsided data from an insurance company for a practical use, we obtained an exemplary dataset from a book entitled *Data Preparation for Data Mining* [12] to illustrate the proposed algorithm. The data set contains 32 predictor variables and a dichotomous response variable for fraud with 15,420 observations. The data were collected over a three-year period from 1994 to 1996. There are 30 categorical variables, one continuous variable, and an identification variable. Since the proposed algorithm is strictly for fraud detection (classification) instead of prediction of insurance premium, none of them was time-sensitive variable. Each categorical variable was translated into dummy variables. The binary response variable describes whether the claim was categorized as fraud or true. There are 923 (6.4%) claims categorized as fraud within the dataset. There is no missing value in the dataset. The data are typically lopsided in insurance fraud detection and it is challenging to build a classification model with such a lopsided dataset.

We note that only 3 years of cases were used in this dataset. The macro-economic changes that impact these variables would not be visible in a small period of 3 years. However, even though only 3 years of data were used in the selected data-

set, our algorithm does not depend on the years of data recruited because our algorithm treats the years as an ordinal variable. The proposed algorithm can be used to recent insurance data recruited for an increased number of years with millions of observations in practice.

The predictor variables include several demographic variables such as age, gender, marital status, etc. Several variables describe the automobile involved in the claim such as type, make, price, age of vehicle, etc. Other variables describe the claim such as time of year, filing of police report, witness present, etc. The rest of the variables describe the type of insurance policy such as deductible, policy type, etc. The variables are summarized in Table 1.

For an initial variable screening of the data, PolicyNumber (the identification variable) was eliminated because it holds no meaning to the analysis. Multicollinearity among the predictor variables was examined by the variance inflation factor (VIF). If the VIF for a variable was greater than 10, then that variable was considered as highly correlated with other predictor variables and was removed from further analysis. The following variables were sequentially removed from consideration based on their VIF: BasePolicy, VehicleCategory, AgeOfPolicyHolder, Month, and AddressChangeClaim. Therefore, there were 26 remaining variables available for further analysis. These variables were defined as the initial 26 variables to be considered. There were no observations eliminated from the dataset.

A learning set and a test set were created from the original dataset. The learning set was used to build all the models in this paper. The test set was used to test and provide the final results of all the models. Since the whole dataset was heavily lopsided with 14,927 non-fraud cases and 923 fraud cases, the learning set was created to balance the data for more accurate results. The learning set was randomly selected for 1000 observations by a stratified random sampling. Five hundred of 1000 observations were randomly chosen from the 14,497 non-fraud cases and the rest 500 observations were randomly selected from the 923 fraud cases. The test set included the rest of 13,997 non-fraud cases and 423 fraud cases, so thus the size of the test set was 14,420.

3. Statistical Methodology

3.1. Logistic Regression

Logistic regression is a popular method to build a prediction model for a binary response variable. A multiple logistic regression model takes multiple predictor variables, $\{x_j\}_{j=1}^{p}$, to predict the binary response variable $\{Y_i\}_{i=1}^{n}$, with Y_i having values of 1 for positive outcome or 0 for negative outcome. A parameter π_i represents the probability that the outcome is positive. The outcome desired (positive) is coded as 1 in the data. The multiple regression model can be written as follows:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \quad (1)$$

Table 1. Summary of automobile claims data.

Variable	Description
Month	From January to December
Week of Month	1, 2, 3, 4, 5
Day of Week	From Sunday to Saturday
Make	Accura, BMW, Chevrolet, Dodge, Ferrari, Ford, Honda, Jaguar, Lexus, Mazda, Mercedes, Mercury, Nisson, Pontiac, Porche, Saab, Saturn, Toyota, VW
Accident Area	Rural, urban
Day of Week Claimed	From Sunday to Saturday
Week of Month Claimed	1, 2, 3, 4, 5
Month Claimed	From January to December
Sex	Male, female
Marital Status	Divorced, married, single, widow
Age (Continuous)	Ages range from 16 to 80
Fault	Policyholder, third party
Policy Type	Sedan—all perils, sedan—collision, sedan—liability, sport—all perils, sport—collision, sport—liability, utility—all perils, utility—collision, utility—liability
Vehicle Category	Sedan, sport, utility
Vehicle Price	(Less than $20,000), ($20,000 - $29,000), ($30,000 - $39,000), ($40,000 - $59,000), ($60,000 - $69,000), (greater than $69,000)
Policy Number	ID variable
Rep Number	1 - 16
Deductible	300, 400, 500, 700
Driver Rating	1, 2, 3, 4
Days Policy Claims	15 - 30, 8 - 15, more than 30, none
Days Policy Accident	1 - 7, 15 - 30, 8 - 15, more than 30, none
Past Number of Claims	1, 2 - 4, more than 4, none
Age of Vehicle	2 years, 3 years, 4 years, 5 years, 6 years, 7 years, more than 7 years, new
Age of Policy Holder	16 - 17, 18 - 20, 21 - 25, 26 - 30, 31 - 35, 36 - 40, 41 - 50, 51 - 65, over 65
Police Report Filed	Yes, No
Witness Present	Yes, No
Agent Type	External, internal
Number of Supplements	1 - 2, 3 - 5, more than 5, none
Address Change Claim	Under 6 months, 1 year, 2 - 3 years, 4 - 8 years, no change
Number of Cars	1, 2, 3 - 4, 5 - 8, more than 8
Year	1994, 1995, 1996
Base Policy	All perils, collision, liability
Fraud Found	0, 1 (response)

The model can also be written as follows:

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} \quad (2)$$

After the correlated predictor variables are removed, a prediction model is built by using stepwise variable selection method. The stepwise selection procedure is similar to the forward selection procedure. The difference is that once a variable is added, the procedure determines whether any of the variables already in the model should be eliminated.

Evaluation and selection of variables via a logistic regression model were based on a variable importance ranking procedure through 20 trials of 10-fold cross-validation (CV). For this method, the learning set was shuffled and partitioned into ten segments. Nine segments were used as a training set to build the logistic regression model. The leftover segment was used to validate the model. This process was repeated ten times with each segment serving as the validation set. The 10-fold CV was performed 20 times to produce 200 logistic regression models. Each model used stepwise selection for its variable selection procedure. For the 200 logistic regression models, each variable that was selected through stepwise selection was counted. For example, a variable that was selected in every model would have a count of 200. From this algorithm, eight variables (Fault, Month Claimed, Policy Type, Age, Agent Type, Deductible, Year and Month) were selected as most important for the logistic regression model by checking AIC (Akaike Information Criterion) improvement using Likelihood Ratio Test (LRT).

The probability response found from the logistic regression model was classified based on a threshold or a cutoff of 0.5. The threshold determines the probability of fraud π_i. This means that a predicted probability greater than or equal to 0.5 will be classified as a fraudulent claim (positive).

3.2. Least Absolute Shrinkage and Selection Operator (LASSO)

The LASSO method [3] [13] is a regression model that penalizes the absolute size of the coefficients, which can cause some regression coefficients to shrink to zero. The penalization, or constraint, allows the LASSO method to estimate a model while simultaneously performing automatic variable selection. Let $\hat{\alpha}$ be the intercept term, and $\hat{\beta}$ be the least squares estimates. Given many predictor variables, $\{x_k\}_{k=1}^{p}$, the LASSO estimate $(\hat{\alpha}, \hat{\beta})$ is defined by

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg\min_{\alpha, \boldsymbol{\beta}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[I_{y_i=1} * \log \pi_i + I_{y_i=0} * \log(1-\pi_i) \right] - \lambda \sum_j |\beta_j| \right\},$$

where π_i is denoted in Equation (2), $\lambda \geq 0$ and $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)^{\mathrm{T}}$. The constraint, λ, regulates the extent of shrinkage that is applied to the coefficient estimates.

The optimal λ is found through 10-fold CV by seeking minimum misclassification error, which can be seen in **Figure 1** in the cross-validation curve. The

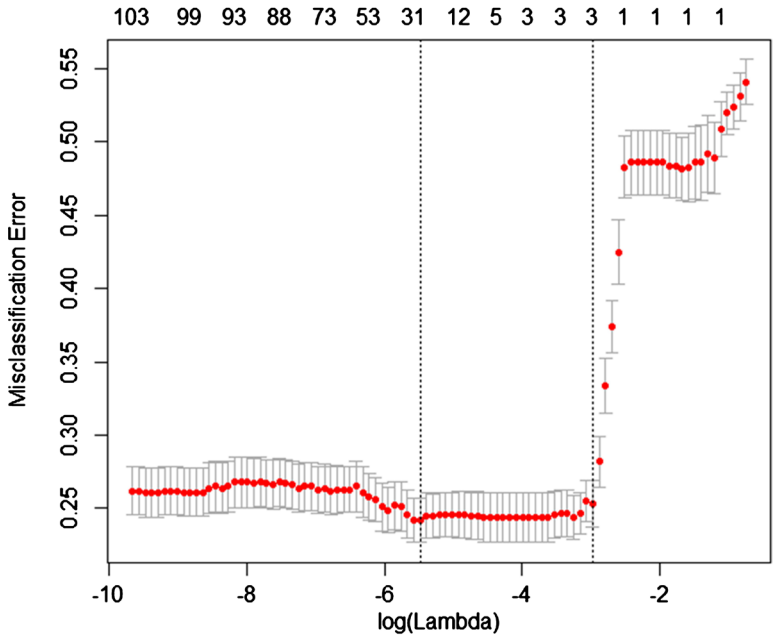

Figure 1. Lambda (λ) values selected through a cross-validation for a LASSO model. The left dotted line defines the minimum lambda and the right dotted line defines the lambda one standard error away from the minimum lambda.

optimal λ produces the lowest misclassification error for the model. Each estimated λ is accompanied by an upper and lower error bound for the estimated misclassification error. The chosen values of λ are designated by the two vertical dotted lines, which represent the minimum λ and the λ_{1SD}, one standard error away from the minimum [14]. We used λ_{1SD}. With smaller values of λ, a LASSO model will produce least squares estimates of a standard regression model. When λ is larger, automatic variable selection occurs by shrinking more coefficients to zero, which removes them from the model. LASSO models are implemented using the R package "glmnet".

The learning set is used to build a LASSO model with the initial 26 variables, and as before in the logistic regression model, the most important variables for the LASSO model were found through 20 trials of 10-fold CV. For the 200 models, each variable was counted if it was significant for the model. The following 8 variables were selected through LASSO variable selection via CV by checking AIC improvement using LRT: Month Claimed, Fault, Sex, Agent Type, Age, Deductible, Year, and Make.

3.3. Random Forests

Random Forests (RF) [10] consists of an ensemble of classification trees, where each classifier is built from different independent and identically distributed bootstrap samples from a training set and each classifier casts a vote for the most popular class. Random Forests algorithm may be summarized in **Algorithm 1** below.

Algorithm 1. Random Forests

1) Create n bootstrap samples from the training set to build n trees (default =

500 trees).

2) For each bootstrap sample, at each node, randomly select m (where $m \leq$ total number of predictor variables; default = $\sqrt{m}$) predictor variables and determine best split among those variables under a feature split criterion (e.g., Gini index).

3) Determine predicted classifications on out-of-bag data (about 1/3 of the sample called the out-of-bag, or OOB, data) that was not in the bootstrap sample by aggregating decisions from the n trees that were grown.

4) Estimate the misclassification rate of OOB data.

5) Aggregate the OOB predictions and calculate misclassification rate (or accuracy) for Random Forests.

Random Forests add more randomness than a single classification tree. The extra randomness is demonstrated through the use of n bootstrap samples and randomly selecting m predictor variables to determine the optimum split at each node for each bootstrap sample. The number of predictor variables m in each node may be extended to the total number of variables in each bootstrap sample in RF.

An important feature of RF is a measure of variable importance. Variable importance is determined by two methods, the Mean Decrease in Impurity (MDI) and the Mean Decrease in Accuracy (MDA). The MDI uses the Gini index as an impurity function and can also be known as the Mean Decrease Gini. The Gini Index $i(t)$ is

$$i(t) = 1 - \sum_{k=0}^{c-1} \left[p(k \mid t) \right]^2,$$

where $p(k \mid t)$ is the fraction of observations belonging to class k at a given node t and c is the number of classes [13]. The MDI measures the importance of a variable X_m using the Gini Index $i(t)$ by taking the sum of the weighted impurity decreases for all nodes and finding the average over all N_T trees in RF. A variable with a higher MDI is deemed as more significant.

The MDA determines the importance of a variable by measuring mean decrease in OOB accuracy for each tree. Each variable's importance is computed by the mean decrease in OOB accuracy before and after a random permutation of each variable [15]. The MDA takes the average difference in accuracies between the OOB data and the permuted OOB data over the N_T trees. A variable with a higher MDA is considered more important.

Twenty trials of 10-fold CV were performed on the RF, a total of 200 RF models, to obtain a feasible set of important variables via variable importance ranking. Random Forests models were built using the R package "randomForest". Each RF implemented with 500 decision trees. A RF model was built with the initial 26 predictor variables in the learning set.

The variable importance plot from this Random Forests model is shown in Figure 2. The plots rank variable importance starting with the most important variable at the top of the plot with the highest MDA and MDI. The variables are

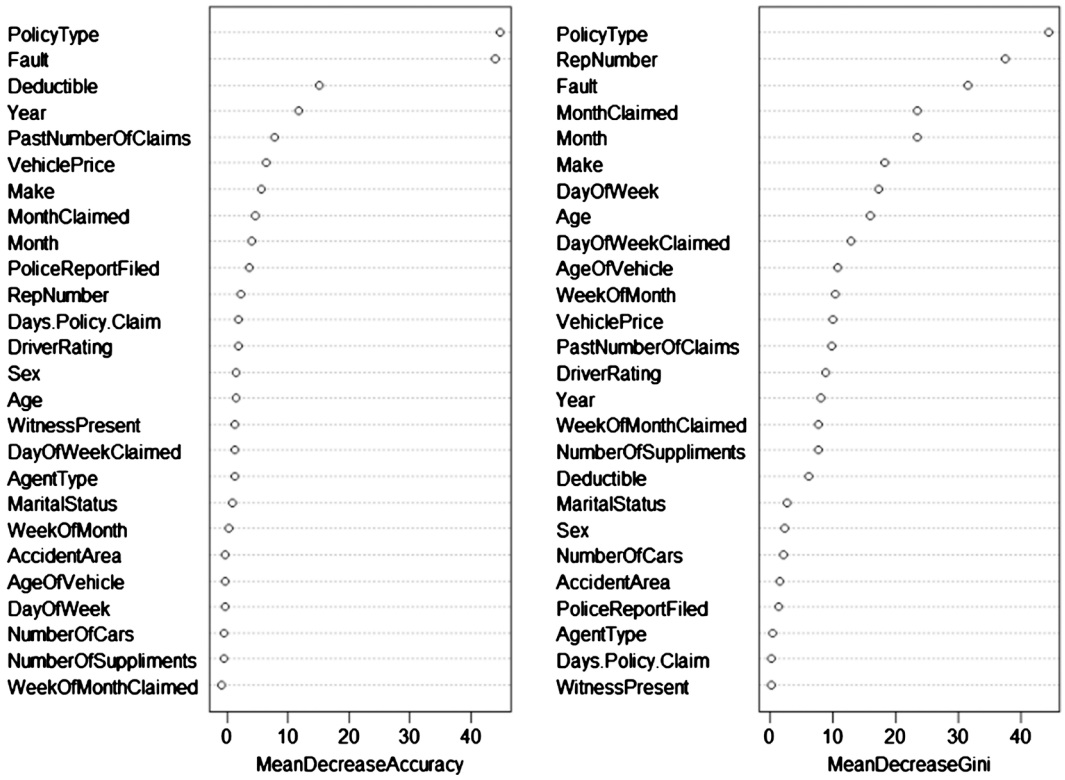

Figure 2. Variable importance plots for a Random Forests model, left for MDA and right for MDI.

then ranked as next most important until the least important variable is reached at the bottom of the plot with the lowest MDA and MDI. The plots display each variable along the y-axis and their importance is shown in the x-axis. RF discovers "Policy Type" as the most important variable. The MDA and the MDI variable importance plots show similar rankings in their most important variables.

For MDI criterion, the following 10 variables were selected as most important variables based off learning set accuracy: Policy Type, Rep Number, Fault, Month Claimed, Month, Make, Day of Week, Age, Day of Week Claimed, Age of Vehicle. On the other hand, using the MDA criterion, 15 variables were chosen based on the learning accuracy. Here are the 15 most important variables: Policy Type, Fault, Deductible, Year, Past Number of Claims, Vehicle Price, Make, Month Claimed, Month, Police Report Filed, Rep Number, Days Policy Claim, Driver Rating, Sex, and Age.

3.4. Support Vector Machine

Support vector machines (SVMs) are kernel-based supervised learning algorithms. Support vector machines consider data points as a p-dimensional vector and separate the data points by (p-1)-dimensional hyperplane for classification. In our study, the two categories (fraud and now-fraud claims) are classified by an optimal hyperplane in a multi-dimensional space, with the largest marginal distance to the nearest training-data point of any class. After a training phase, new observations are mapped into the same space and are classified to a category

based on a side of hyperplane [11].

Let the training data consist of n pairs $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)$ with $\boldsymbol{x}_i \in \mathcal{R}^p$ and y_i are binary response variable. If the data are linearly separable, the SVM finds the closest points in convex hulls and finds a hyperplane (P_0) bisecting the closest points, where P_0 is defined by $\{\boldsymbol{x} : f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta} + \beta_0 = 0\}$, and $\|\boldsymbol{\beta}\| = 1$. Then, the classifier creates a parallel hyperplane (P_1) on a point in class -1 closest to P_0 and a second parallel hyperplane (P_1) on a point in class 1 closest to P_0. The optimal hyperplane that separates the data can be found by maximizing the margin (M) that is a perpendicular distance between two parallel supporting planes P_1 and P_2. A resulting classifier would be $\hat{y} = sign(\boldsymbol{x}'\boldsymbol{\beta} + \beta_0)$.

For datasets that are not linearly separable, SVMs map the data into higher dimensional space where the training set is separable via some transformation $K : \boldsymbol{x} \to \phi(\boldsymbol{x})$. A kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$ computes inner products in some expanded feature space. Some kernel function such as linear $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$ and Gaussian (radial-basis function) $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 / 2\sigma^2\right)$ are widely used [16].

We applied SVM models with four sets of variables that were selected by Logistic Regression, LASSO, Random Forests by MDI, and Random Forests by MDA. We compared learning accuracies between SVM with linear kernel and SVM with Gaussian kernel. Support vector machines with linear kernel had better learning accuracy. We compare the performance in "Results" section for SVM with linear kernel.

4. Results

The results of each model were compared based on accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Receiver Operating Characteristic (ROC) curve analysis (AUC). Accuracy is defined as the percentage of predictions that were correct. Sensitivity is described as the percentage of the amount of positive (fraud) predictions when the actual classification is positive (fraud). Specificity measures the percentage of the amount of negative (non-fraud) predictions when the actual classification is negative (non-fraud). The PPV measures the percentage of accurate predictions when the prediction is positive. The NPV measures the percentage of accurate negative predictions when the prediction is negative. Receiver Operating Characteristic curve analysis is an alternative way to obtain accuracy of the test [17] [18].

For the training set, 500 observations were randomly selected from the 14,497 non-fraud cases and 500 observations were randomly picked from the 923 fraud cases. The remaining 13,997 observations of the non-fraud cases and 423 observations of the fraud cases will be used as a test set. To summarize, we have 1000 observations for the training set and 14,420 observations for the test set.

The results of the logistic regression model with selected features (Fault, Month Claimed, Policy Type, Age, Agent Type, Deductible, Year and Month) found via cross-validation are given in **Table 2**. Since our dataset is highly lopsided (fraud

Table 2. Model Performance (%) (Accuracy, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and Receiver Operating Characteristic curve analysis (AUC)) of Logistic Regression (LR), LASSO, Random Forests via Mean Decrease in Gini Index (RF-MDI) and Random Forests via Mean Decrease in Accuracy (RF-MDA) methods.

	LR	LASSO	RF-MDI	RF-MDA
Accuracy	87.1	97.7	97.0	91.4
Sensitivity	62.4	70.0	70.0	69.3
Specificity	93.1	98.6	98.6	98.2
PPV	56.3	60.7	60.5	54.1
NPV	96.8	97.9	97.9	97.9
AUC	82.4	85.3	83.6	73.8

2.9%; non-fraud 97.1%), our logistic regression with the most important features produces the results showing unbalance between sensitivity (62.4%) and specificity (93.1%). On the other hand, the accuracy and AUC are reasonably high 87.1% and 82.4%, respectively.

The results of LASSO model with the most important variables (Month-Claimed, Fault, Sex, AgentType, Age, Deductible, Year and Make) selected via cross-validation are also given in Table 2. As expected LASSO model has better improved performance compared to the performance from the logistic regression model. It produces the results showing unbalance between sensitivity (70.0%) and specificity (98.6%) due to extremely low fraud rate in the data. The accuracy and AUC are 97.7% and 85.3%, respectively.

The full results of Random Forests models with MDI (RF-MDI) and MDA (RF-MDA) are also shown in Table 2. Results of RF-MDI were based on 10 variables (Policy Type, Rep Number, Fault, Month Claimed, Month, Make, Day Of Week, Age, Day Of Week Claimed, Age Of Vehicle) selected via MDI method. Results of RF-MDA were based on 15 variables (Policy Type, Fault, Deductible, Year, Past Number Of Claims, Vehicle Price, Make, Month Claimed, Month, Police Report Filed, Rep Number, Days Policy Claim, Driver Rating, Sex, and Age) selected via MDA method. Accuracy and AUC for RF-MDI were higher as 97.0% and 83.6% compared to those for RF-MDA of 91.4% and 73.8%, respectively. Sensitivity and specificity from RF-MDI and RF-MDA were almost the same. On the other hand, PPV of RF-MDI was higher as 60.5% compared to PPV of RF-MDA of 54.1%.

The results of SVM with different sets of variables selected by Logistic regression, LASSO, RF-MDI and RF-MDA are shown in Table 3. The performance was similar to the performance of LASSO in Table 2 that was the best model and was also similar to the performance of the other models (LR, RF-MDI and RF-MDA). However, PPV and AUC were substantially lower compared to those of models in Table 2.

In summary, LASSO model with the eight variables selected through LASSO

Table 3. SVM Model Performance (%) (Accuracy, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and Receiver Operating Characteristic curve analysis (AUC) based on four sets of variables selected by Logistic regression (LR), LASSO, RF-MDI and RF-MDA.

	SVM with a Set of Variables Selected by			
	LR	**LASSO**	**RF-MDI**	**RF-MDA**
Accuracy	93.6	95.6	96.7	96.0
Sensitivity	62.9	69.9	72.1	64.1
Specificity	98.3	97.8	98.3	98.3
PPV	52.1	50.0	56.0	53.0
NPV	97.7	97.9	97.9	98.1
AUC	76.9	77.3	80.9	78.0

variable selection method (Month Claimed, Fault, Sex, Agent Type, Age, Deductible, Year and Make) had better performance compared to performance of other models. Ranges of accuracy and AUC among models considered were from 87.1% (LR) to 97.7% (LASSO) and from 73.8% (RF-MDA) to 85.3% (LASSO), respectively. Ranges of sensitivity and specificity were from 62.4% (LR) to 72.1% (RF-MDI) and from 93.1% (LR) to 98.6% (LASSO and RF-MDI), respectively. Ranges of PPV and NPV were from 50.0% (SVM-LASSO) to 60.7% (LASSO) and from 96.8% (LR) to 98.1% (RF-MDA), respectively. The balance between sensitivity and specificity was a bit unbalanced and so does between PPV and NPV because of the lopsided test set. The test set has 2.9% fraud cases and 97.1% non-fraud cases.

5. Conclusions

We applied both parametric and non-parametric supervised classification models with various variable selection methods to identify the fraud from all the insurance claims. It can be seen from the results that the most effective method for classifying fraudulent automobile insurance claims is LASSO method with a set of eight variables selected by LASSO method. The LASSO model consistently has the highest accuracy, AUC, sensitivity and PPV of all the methods. A high percentage for sensitivity is very important since an insurance claim dataset includes significantly more non-fraudulent cases than fraudulent cases, which increases the difficulty in identifying a fraudulent claim. Therefore, the model that produces the highest sensitivity, PPV and accuracy will be the best model to identify fraudulent claims. Sensitivity describes the probability that the model identities the claim as fraud among all fraud claims. The PPV describes the probability that the claim identified as fraud by the model is truly a fraud claim.

In our study, since our whole data are heavily unbalanced with a very low proportion of 6.0% (923 cases out of 14,589 accounts) of all insurance claims being a fraudulent claim, the ability to accurately identify the fraudulent claims is more difficult. Achieving high sensitivities, accuracies and PPVs is more cru-

cial. As shown in Table 2 and Table 3, the statistics show that LASSO model is the most effective model for fraudulent claim detection showing accuracy, AUC, sensitivity, specificity, PPV and NPV of 97.7%, 85.3%, 70.0%, 98.6%, 60.7% and 97.9%, respectively.

When comparing the individual logistic regression and LASSO models, LASSO showed better performance in all six performance measures. On the other hand, LASSO was very competitive to RF-MDI. Their sensitivities, specificities, PPV and NPV were almost the same. Accuracies of LASSO and RF-MDI were 97.7% and 97.0%, respectively. The AUCs of LASSO and RF-MDI were 85.3% and 83.6%, respectively. When LASSO was compared to RF-MDA, LASSO was substantially better in accuracy of RF-MDA (91.4%) and AUC of RF-MDA (73.8%).

Sets of selected variables from LR, LASSO, RF-MDI and RF-MDA were applied to SVM. Among these, SVM with RF-MDI had the best performance with accuracy of 96.7%, sensitivity of 72.1%, specificity of 98.3%, PPV of 56.0%, NPV of 97.9%, and AUC of 80.9%. When SVM via RF-MDI was compared to LASSO model in Table 2, sensitivity was improved, but PPV (56.0%) and AUC (80.9%) were decreased.

Using 10-fold CV to determine variable importance and selection for individual models could result in an improvement in results. We compared several different variable importance and selection methods using a cross-validation method to find the best approach for variable selection for automobile insurance data. Even though the data were collected between the years 1994 to 1996, our methodology can be applied to any similar automobile insurance data.

Our data were highly lopsided with about 6% of the claims in the dataset being fraudulent claims, thus it is a quite challenge to identify those claims. It means that the probability of coming across a fraudulent claim is drastically less than encountering a true claim. However, this research has successfully shown that between 70.0% and 72.1% sensitivities are achieved via LASSO, RF-MDI and SVM with RF-MDI variables.

The goal of this paper is to propose a general statistical learning algorithm for fraud detection best suited especially for any highly lopsided data from an insurance company for a practical use rather than a data-driven algorithm. The first step of the algorithm is to screen the variables by checking VIF's. The second step is to conduct variable importance ranking via a cross-validation. The third step is to build a classification model by conducting a statistical test for model improvement. The last step is to summarize and use the selected variables in the model to classify new cases of claims.

With fraudulent insurance claims on the rise, it is more important than ever to be able to recognize which insurance claims are actually fraudulent. Accurately identifying these fraudulent claims will help prevent the excessive monetary waste within the insurance industry and provide financial relief to both the companies and their policyholders. It can be seen from this research that exploring different classification methods, other than the standard logistic regres-

sion, can improve the rate at which fraudulent claims are detected. Utilizing different classification methods not only increases the chances of correctly identifying fraudulent claims, but also sing amounts of data that is more available and accessible to analyze in practice, these classification techniques are becoming increasingly important to sift out the non-fraudulent cases and hone in on the fraudulent ones. This research should provide some benefit to automobile insurance industry for their fraud detection helps filter out the obvious cases that are not fraudulent claims. With the increasing amount of data that is more available and accessible to analyze in practice, these classification techniques are becoming increasingly important to sift out the non-fraudulent cases and hone in on the fraudulent ones. This research should provide some benefit to automobile insurance industry for their fraud detection research in order to prevent insurance abuse from escalating any further.

Acknowledgements

This work was partially supported by the Research, Scholarly and Creative Activity (RSCA) Award and Office of Research and Sponsored Programs (ORSP) Award from California State University, Long Beach, CA.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Viaene, S., Ayuso, M., Guillen, M., Gheel, D.V. and Dedene, G. (2007) Strategies for Detecting Fraudulent Claims in the Automobile Insurance Industry. *European Journal of Operational Research*, **176**, 565-583. https://doi.org/10.1016/j.ejor.2005.08.005

[2] Boyer, M. (2000) Centralizing Insurance Fraud Investigation. *The Geneva Papers on Risk and Insurance Theory*, **25**, 159-178. https://doi.org/10.1023/A:1008766413327

[3] Derrig, R.A. (2002) Insurance Fraud. *Journal of Risk and Insurance*, **69**, 271-287. https://doi.org/10.1111/1539-6975.00026

[4] Ciaene, S., Derrig, R.A., Baesens, B. and Dedene, G. (2002) A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection. *The Journal of Risk and Insurance*, **69**, 373-421. https://doi.org/10.1111/1539-6975.00023

[5] Caudill, S., Ayuso, M. and Guillen, M. (2005) Fraud Detection Using a Multinomial Logit Model with Missing Information. *The Journal of Risk and Insurance*, **72**, 539-550. https://doi.org/10.1111/j.1539-6975.2005.00137.x

[6] Artis, M., Ayuso, M. and Cuillen, M. (2002) Detection of Automobile Insurance Fraud with Discrete Choice Models and Misclassified Claims. *The Journal of Risk and Insurance*, **69**, 325-340. https://doi.org/10.1111/1539-6975.00022

[7] Wang, Y. and Xu, W. (2018) Leveraging Deep Learning with LDA-Based Text Analytics to Detect Automobile Insurance Fraud. *Decision Support Systems*, **105**, 87-95. https://doi.org/10.1016/j.dss.2017.11.001

[8] Nian, K., Zhang, H., Tayal, A., Coleman, T. and Li, Y. (2016) Auto Insurance Fraud Detection Using Unsupervised Spectral Ranking for Anomaly. *The Journal of Finance and Data Science*, **2**, 57-58. https://doi.org/10.1016/j.jfds.2016.03.001

[9] Tibshirani, R. (2011) Regression Shrinkage and Selection via the Lasso: A Retrospective. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **73**, 273-282. https://doi.org/10.1111/j.1467-9868.2011.00771.x

[10] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[11] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1007/BF00994018

[12] Pyle, D. (1999) Data Preparation for Data Mining. Morgan Kaufmann Publishers, Inc., San Francisco.

[13] Tan, P.-N., Steinbach, M. and Kumar, V. (2005) Introduction to Data Mining. Pearson Addison Wesley, Boston.

[14] Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**, 1-22. https://doi.org/10.18637/jss.v033.i01

[15] Baek, S., Moon, H., Ahn, H., Kodell, R.L., Lin, C.-J. and Chen, J.J. (2008) Identifying High-Dimensional Biomarkers for Personalized Medicine via Variable Importance Ranking. *Journal of Biopharmaceutical Statistics*, **18**, 853-868. https://doi.org/10.1080/10543400802278023

[16] Ahn, H. and Moon, H. (2010) Classification: Supervised Learning with High Dimensional Biological Data. In: Lee, J.K., Ed., *Statistical Bioinformatics*: *A Guide for Life and Bio Medical Science Researchers*, Chapter 6, John Wiley & Sons, Chichester, 129-156. https://doi.org/10.1002/9780470567647.ch6

[17] Metz, C.E. (1978) Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine*, **8**, 283-298. https://doi.org/10.1016/S0001-2998(78)80014-2

[18] Zweig, M.H. and Campbell, G. (1993) Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clinical Chemistry*, **39**, 561-577.