Thailand Statistician
July 2023; 21(3): 552-568
http://statassoc.or.th
Contributed paper

Detecting Fraudulent Claims in Automobile Insurance Policies by Data Mining Techniques

Teerawat Simmachan [a,b], Weerapong Manopa [a], Pailin Neamhom [a], Achiraya Poothong [a] and Wikanda Phaphan*[c,d]

[a] Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani, Thailand

[b] Thammasat University Research Unit in Data Learning, Faculty of Science and Technology, Thammasat University, Pathum Thani, Thailand

[c] Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

[d] Research Group in Statistical Learning and Inference, KMUTNB, Bangkok, Thailand

*Corresponding author; e-mail: wikanda.p@sci.kmutnb.ac.th

Received: 27 November 2022
Revised: 28 February 2023
Accepted: 5 March 2023

Abstract

The insurance industry is a fast-growing industry and handles substantial amounts of data. Fraudulent claims are the main problem in the industry. Auto insurance fraud is one of the most prominent types of insurance fraud. Numerous fraudulent claims affect not only the insurance company but also the sincere policyholders because of the increase in premium amounts. Typically, a fraud report is unbalanced data. Overlooking this generally leads to weak classifiers for predicting the minority class (fraudulent claim). Therefore, the fraud detection is a challenging problem. Traditional approaches are difficult to handle and inefficient. Data mining has recently offered significant contributions to insurance analysis. To overcome this, data mining techniques are used to predict fraudulent claims. The aims of this research are to develop, firstly, what types of features should be used to build the predictive model; and second, a statistical learning strategy to classify whether a fraud report is fraudulent or not. To discover important sets of features, logistic regression (parametric method) and random forest (non-parametric method) are considered as tools of variable selection algorithms. This process is done by cross-validation to reduce uncertainty until two sets of important features are obtained. Four algorithms including logistic regression, random forest, Naïve Bayes, and adaptive boosting are employed as classifiers. A confusion matrix is used to evaluate the algorithm's performance. The results suggest that a set of important features obtained from the non-parametric method provides better performance than the parametric method. The random forest is considered as the best algorithms to identify fraudulent claims with the highest sensitivity (99.19%) and the positive predictive value (93.62%). This work would help in a screening process to investigate claims, thus minimizing human resources and monetary losses in the insurance industry.

Keywords: Naïve Bayes, random forest, adaptive boosting, logistic regression, variable selection

1. Introduction

The insurance industry is one of the fast-growing industries, there are more than a thousand companies worldwide, and more than one trillion dollars in premiums are collected each year (Roy and George, 2017). Fraudulent claims are the main problem in the insurance industry. Fraudulent claims are identified when some person cheats the insurance companies for receiving compensation. There are two types of fraudulent claims: hard insurance fraud and soft insurance fraud (Belhadji et al., 2000) Hard insurance fraud is defined in case a person intentionally fakes an accident. Soft insurance fraud is defined if a person has a valid insurance claim but falsifies part of the claim. One of the important types of insurance fraud is automobile insurance fraud (Roy and George, 2017). Approximately 21%–36% of automobile insurance claims are suspected to be fraudulent claims, but only less than 3% of the suspected fraud is legally preceded (Kowshalya and Nandhini, 2018). When fraudulent claims are undetected, insurance companies increase the premium amount to compensate for the loss. Sincere policyholders are affected by increasing premium amounts. If a company has an effective fraud detection system, then customer satisfaction increases. Accordingly, loss adjustment expenses will be reduced.

There are many manual inspection methods to detect fraudulent claims. The commonly used method is data analysis with its own instruction (Belhadji et al., 2000). Insurance fraud detection relies on auditing and expert inspection. It takes a long time to decide the amount of the claim for applicants. Manual exposure to fraudulent claims leads to higher costs and inefficiency. It deals with the different domains of knowledge. Essentially, claim fraud needs to be detected earlier before the claim payment is done. To overcome this problem, data mining techniques are used to predict automobile fraudulent claims. There are numerous works related to predicting fraudulent automobile claims via data mining techniques. A survey on fraud analytics using predictive models in insurance claims was provided in 2017 (Priya and Pushpa, 2017). A case study on fraud diagnosis using machine learning was proposed in 2002 (Viaene et al. 2002). The most efficient methods were logistic regression, least-squares support vector machine, and Naïve Bayes, respectively. A study focusing on detecting fraudulent claims in automobile insurance using machine learning technique was presented in 2017 (Roy and George, 2017). Decision tree and random forest algorithms were better than the Naïve Bayes algorithm. A case study of auto-insurance fraud detection using deep learning with text analysis was proposed in 2018 (Wang and Xu, 2018). The results showed that machine learning algorithms were more effective than logistic regression. A simulation study on predicting fraudulent claims in automobile insurance using data mining techniques were submitted in 2018 (Kowshalya and Nandhini, 2018). The random forest algorithm performed the best. Predictive modeling for detecting fraudulent automobile insurance claims using parametric and nonparametric statistical learning algorithms together with a cross-validation technique was proposed in 2019 (Moon et al. 2019). The suggested algorithm was the least absolute shrinkage.

This raises the first question: what sets of variables/features should be focused on to detect fraudulent claims. To address this question, a feature selection procedure was necessarily needed. Indeed, there was a large number of features in practice. We may not use all the features in the predictive model. Significant features should be investigated. This procedure can reduce the computational time of model training and also improve the performance of the model. In this research, important features were obtained by binary logistic regression treated as a parametric statistical method and random forest denoted as a non-parametric statistical method. The two sets of important features were obtained via a k-fold cross-validation to reduce uncertainty and increase the chance of detecting the proper claims. Variable importance was measured by the observed statistics of the two methods used for discovering important features.

The second question concerns a decision-making algorithm to classify whether a claim is classified as fraudulent or not. To answer this question, the logistic regression and random forest were also employed as the classifiers. Based on the literature review, the two classifiers were noticed as effective algorithms. Instead of the two algorithms, Naïve Bayes was selected as a control algorithm because it was one of the early methods. Finally, adaptive boosting was chosen as a challenging algorithm

since there were few studies using the boosting algorithm for detecting automobile fraudulent claims. However, there were many studies on fraud detection of other aspects via the boosting technique. For instance, the study of credit card fraud detection in 2018 (Randhawa, 2018) and 2021 (Zou, 2021), and a case study of fraudulent financial operations in 2020 (Belyakov and Karpov, 2020).

The four algorithms were implemented to predict fraudulent claims. Real data set from an anonymous insurance company in the United States in 2015 was used to illustrate the proposed statistical learning strategy. Multicollinearity check was implemented in the initial step. Unbalanced data was an important problem in an insurance fraud detection. In fact, the number of fraudulent claims was typically far fewer than the number of non-fraudulent claims. Ignoring this generally could lead to weak classifiers for predicting the minority class (fraudulent claim). Consequently, a stratified random sampling was employed to overcome this problem. A confusion matrix was used to assess the algorithm's performance. This work would offer some benefit to insurance companies for their fraud detection strategy to minimize human resources and monetary losses.

2. Materials and Methods

2.1. Data description and preparation

The main objective of this research is to offer a general statistical learning strategy for detecting fraudulent claims. Claim details of an insurance company are mostly confidential. Generally, the required dataset is not available. To illustrate the proposed strategy, the auto-insurance dataset is selected from the online source provided by Sharma (2020). The data provided information on claims in automobile insurance policies of the anonymous insurance company in the United States and were collected from January 1, 2015, to March 1, 2015. The dataset consists of 26 features or predictor variables and the dichotomous response with 1,000 records. There is no missing in this dataset. The predictor variables are divided into 15 categorical variables and 11 continuous variables, and their descriptions are presented in Table 1-2, respectively. The response variable is fraud report describes whether the claim was fraudulent or not. It is reported that there are 247 fraudulent claims (24.7%) and 753 non-fraudulent claims (75.3%). This dataset indicates an unbalanced data. Consistent with information in insurance fraud reports, they are often unbalanced in practice, and it is challenging to establish a predictive model with such unbalanced data.

Since logistic regression was a candidate method for variable selection algorithm, multicollinearity check among predictor variables was implemented in the initial step. The required assumption in a general regression model was no multicollinearity problem. If correlations among predictor variables were greater than 0.60, then the corresponding variables were treated as strongly correlated with other predictor variables and were removed from further step. Cramers V was used to measure associations between the categorical variables. Spearman's rank correlation was used to determine relationship between the continuous variables. The following predictor variables were sequentially eliminated based on their correlations: total claim amount, personal damage, and property damage. Accordingly, there were 23 remaining variables for further analysis. There were no observations removed from the dataset.

For data manipulation, the continuous variables were normalized since the range of values of these variables varied widely. This procedure was called feature scaling. Most machine learning (ML) algorithms used the Euclidean distance between two data points; thus, the classifiers may not perform properly without feature scaling (Aksoy and Haralick, 2001). In this study, the Standardization (Z-score Normalization) was selected as a feature scaling method to rescale the continuous features. Another reason why feature scaling was required was that the algorithms converged much faster with feature scaling than without it (Ioffe and Szegedy, 2015).

A training set and a test set were established from the data in the previous step. The training set was used to train all models. The test set was used to evaluate all the models for the final results. As the original dataset was unbalanced with 247 fraudulent claims and 753 non-fraudulent claims, the training set was built to balance the data for better proper results. The training set was randomly selected for 400 observations via a stratified random sampling. A haft of the 400 observations was

randomly picked from the 247 fraud cases and the rest 200 observations were randomly drawn from the 753 non-fraud cases. Subsequently, four hundred observations in the training set were shuffled. The remaining 600 observations consisted of 47 fraud cases and 553 non-fraud cases and were considered to be the test set.

Since there was a small sample size relative to the number of the predictor variables in the training set, high dimension problem may happen in particular when a data cross validation was applied. To create more generalization of the data, 300 observations were randomly selected without replacement from 400 observations in the training set. Ten sets of 300 observations were created independently and were combined. Finally, total 3,000 observations were then shuffled, and this was called the learning set. The learning set was used to create all the models in this work.

Table 1 Categorical variables

Feature Name	Description
policy_month	The month when the insurance policy starts to take effect
policy_year	The year when the insurance policy starts to take effect
policy_state	The city when the insurance policy starts to take effect
insured_sex	Insured's gender
insured_education_level	Education level of the insured
insured_occupation	Occupation of the insured
insured_hobbies	Hobbies of the insured
insured_relationship	The marital status of the insured
incident_type	Type of accident
incident_month	Incident month
incident_severity	The severity of the accident
authorities_contacted	Person to contact after the incident
incident_state	The state where the incident took place
auto_make	Insured car brand
auto_year	The age of the insured car

Table 2 Continuous variables

Feature Name	Description
age	Insured age
policy_deductable	A specified amount of money that the insured must pay before an insurance company pays a claim
policy_annual_premium	Annual premium
umbrella_limit	Coverage limit for umbrella insurance that is made with car insurance
number_of_vehicles_involved	Number of vehicles involved
bodily_injuries	Amount of medical expenses
witnesses	Number of witnesses
total_claim_amount	Total claim amount
injury_claim	Personal damage
property_claim	Property damage
incident_hour_of_the_day	Time of incident

2.2. Data validation tools

2.2.1 K-fold cross validation

This approach involves randomly dividing the set of observations into k groups or folds of about equal size (Gareth et al. 2013). The first fold is treated as a validation set or testing set, and the method or algorithm is fitted on the remaining $k-1$ folds. Accuracy is the proportion of the correct predictions among the total number of cases examined. This procedure is repeated k times, and a different group of observations is treated as a validation set each time. This process results in k estimates of the accuracy criteria such that $accuracy_1, accuracy_2, ..., accuracy_k$. The k-fold cross-validation estimate (CV) is computed by averaging these values.

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} accuracy_i. \tag{1}$$

2.2.2 Confusion matrix

The performance of an algorithm/ method is computed by a confusion matrix shown in Table 3. The positive class indicates the fraud case, and the negative class represents the no fraud case. True positives (TP) indicate the cases in which we predict fraud, and it actually has fraud. Likewise, true negatives (TN) are the cases in which we predict no fraud, and it has no fraud. False positives (FP) specify the cases in which we predict fraud, but actually has no fraud. False negatives (FN) are the cases in which we predict no fraud, but it actually has fraud.

Table 3 Confusion matrix

		Actually	
		Positive	**Negative**
Predicted	**Positive**	True Positives (TPs)	False Positives (FPs)
	Negative	False Negatives (FNs)	True Negatives (TNs)

2.2.3 Assessment criteria

Since insurance companies practically raise more awareness on fraudulent claims rather than non-fraudulent claims, four assessment criteria including sensitivity, positive predictive value (PPV), negative predictive value (NPV), and F1-score were used to evaluate the performance of algorithms. Reminding that in this study, the positive class indicates the fraud case, and the negative class represents the no fraud case. The sensitivity is described as the proportion of the amount of positive (fraud) predictions when the actual classification is positive (fraud). The PPV is defined as the proportion of predicted positives which are actual positives. It reflects the probability a predicted positive is a true positive. The NPV is described as the proportion of accurate negative predictions when the prediction is negative. The F1-score measures the predictive skill of a model by elaborating on its class-wise performance rather than an overall performance as done by accuracy. This criterion is defined as the harmonic mean of the precision and sensitivity. It is typically employed in case data are unbalanced.

The corresponding formulas are given below. The larger the value the higher the performance.

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{2}$$

$$\text{PPV} = \frac{TP}{TP + FP}, \tag{3}$$

$$\text{NPV} = \frac{TN}{TN + FN}, \tag{4}$$

$$\text{F1-score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}. \tag{5}$$

2.3. Algorithms

2.3.1 Binary logistic regression

Logistic regression denoted by LG is parametric method used to create a prediction model representing a relationship between predictors and a categorical response variable. Logistic regression helps us estimate the probability of falling into a certain level of the categorical response given a set of predictors. Binary logistic regression is the case when the response variable (Y) is binary random variable taking on the values 0 and 1. In this study, Y is a fraud report, $y = 1$ if the report is the fraud case, and $y = 0$ if the report is the non-fraud case. Let $x_1, x_2, ..., x_p$ be the p predictors of Y. Let π_i be the probability that $y = 1$ or the probability of the event that there is a fraudulent claim when the values of predictors are given. The multiple binary logistic regression model is given as follows:

$$logit(\pi_i) = ln(\frac{\pi_i}{1 - \pi_i}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \tag{6}$$

where β_0 is a constant and β_j is the coefficient of the j^{th} predictor, $(j = 1, 2, ..., p)$. The above equation can be transformed directly in terms of π_i as

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}. \tag{7}$$

The model parameters $\beta_0, \beta_1, ..., \beta_p$ were estimated by the maximum likelihood estimation.

2.3.2 Random forest

Random forest denoted by RF provides an improvement over bagged trees by way of a small tweak that decorates the trees (Gareth et al., 2013). As in bagging, we build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors. A fresh sample of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$. That is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors. In other words, in building a random forest, at each split in the tree, the algorithm is not even allowed to consider a majority of the available predictors. Suppose that there is one very strong predictor in the data set, along with a number of other moderately strong predictors. Then in the collection of bagged trees, most or all the trees will use this strong predictor in the top split. Consequently, all the bagged trees will look quite like each other. Hence the predictions from the bagged trees will be highly correlated.

2.3.3 Naïve Bayes

Naïve Bayes algorithm denoted by NB is a classification technique using Bayes theorem (Berrar, 2018). It assumes strength is a mathematical concept to get the probability. Predictors are not related to each other and have correlations with each other. All features contribute independently to the

probability of maximizing it. It can work with Naïve Bayes model and does not use Bayesian methods. Naive Bayes learning refers to the construction of a Bayesian probabilistic model that assigns a posterior class probability to an instance: $P(Y = y_j|X = x_i)$. The simple Naïve Bayes classifier uses these probabilities to assign an instance to a class. Applying Bayes theorem and simplifying the notation a little, we obtain

$$P(y_j|x_i) = \frac{P(x_i|y_j)P(y_j)}{P(x_i)}, \tag{8}$$

which we can plug into Eqn. (8) and we obtain

$$P(y_j|x) = \frac{\prod_{k=1}^{n} P(x_k|y_j)P(y_j)}{P(x)}. \tag{9}$$

Note that the denominator, $P(x)$, does not depend on the class − for example, it is the same for class y_j. $P(x)$ acts as a scaling factor (the prior probability of predictor x) and ensures that the posterior probability $P(y_j|x)$, the posterior probability of class (y_j: fraudulent or not) given predictor (x, features), is properly scaled (i.e., a number between 0 and 1). When we are interested in a crisp classification rule, that is, a rule that assigns each instance to exactly one class, then we can simply calculate the value of the numerator for each class and select that class for which this value is maximal. This rule is called the maximum posterior rule in Eqn. (10). The resulting winning class is also known as the maximum a posterior (MAP) class, and it is calculated as $\hat{y}$ for the instance x as follows:

$$\hat{y} = \text{arcmax} \prod_{k=1}^{n} P(x_k|y_j)P(y_j). \tag{10}$$

A model that implements Eqn. (10) is called a (simple) Naïve Bayes classifier. Probabilities are computed differently for nominal and numeric attributes.

For a nominal attribute X with r possible attributes values $x_1, x_2, ..., x_r$, the probability $P(x_k|y_j) = n_{kj}/n$ where n is the total number of training examples for which $Y = y_j$, and n_{kj} is the number of those training examples that also have $X = x_k$. In the simplest case, numeric attributes are assumed to have a normal or Gaussian probability distribution. The probability density function for a normal distribution with mean μ and variance σ^2 is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{11}$$

The mean μ and variance σ^2 are calculated for each class and each numeric attribute from the training set.

2.3.4 Adaptive boosting

The adaptive boosting algorithm denoted by AB solves many of the practical difficulties of the earlier boosting algorithms (Schapire, 2013). The algorithm takes as input a training set (x_1, y_1), ..., (x_m, y_m) where each x_i belongs to some domain or instance space, and each label y_i is in some label set Y. For this paper, we assume $Y = \{-1, +1\}$. Later, we discuss extensions to the case of multiclass. The AB algorithm calls a given weak or base learning algorithm repeatedly in a series of rounds $t = 1, ..., T$. One of the main ideas of the algorithm is to maintain a distribution or set of weights over the training set. The weight of this distribution on the training example on the round is denoted $D_i(i)$. Initially, all weights are set equally, but on each round, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set. The weak learners job is to find a weak hypothesis $h_t : X \to \{-1, +1\}$ appropriate for the distribution D_t. The goodness of a weak hypothesis is measured by its error (ε_t)

$$\varepsilon_t = P_{i\sim D_t}\left[h_t(x_i) \neq y_i\right] = \sum_{i=1}^{m} D_i(i)I(h_t(x_i) \neq y_i). \tag{12}$$

Notice that the error is measured with respect to the distribution D_t on which the weak learner was trained. In practice, the weak learner may be an algorithm that can use the weights D_t on the training examples. Alternatively, when this is not possible, a subset of the training examples can be sampled according to D_t, and these (unweighted) resampled examples can be used to train the weak learner. Relating to the horse-racing example, the instances correspond to descriptions of horse races (such as which horses are running, the odds, the track records of each horse, etc.) The main steps of the AB algorithm can be explained below:

Step 1: Given $(x_1, y_1), ..., (x_m, y_m)$ where $x_i \in X, y_i \in Y = \{-1, +1\}$.

Step 2: Initialize $D_1(i) = 1/m; i = 1, \ldots, m$.

Step 3: For $t = 1, ..., T$, train weak learner using distribution D_t. Get weak hypothesis $h_t : X \to \{-1, +1\}$ with the error ε_t in Eqn. (12).

Step 4: Choose $\alpha_1 = \frac{1}{2} ln(\frac{1-\varepsilon_t}{\varepsilon_t})$.

Step 5: Update D_t

$$D_{i+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha t} \; if \; h_t(x_i) = y_i \\ e^{\alpha t} \; if \; h_t(x_i) \neq y_i \end{cases} \tag{13}$$

$$= \frac{D_1(i) \exp\{-\alpha t y_i h_t(x_i)\}}{Z_t}, \tag{14}$$

where Z_t is a normalization factor chosen so that D_{t+1} will be a distribution.

Step 6: Obtain output, the final hypothesis:

$$H(x_i) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t y_i h_t(x_i)\right). \tag{15}$$

2.4. Feature selection procedure

In this section, the parametric and non-parametric feature selection procedures were described. These procedures were used to obtained sets of significant predictor variables and these variables were implemented for further step.

2.4.1 Feature selection by logistic regression

A binary logistic regression is used to create a predictive model using stepwise variable selection method. The stepwise selection procedure is a combination between the forward selection and backward elimination procedures (Maua, 2012). This procedure determines whether the variables already entered in the model should be removed. In this study, variable inclusion and exclusion from the model are based on the Akaike Information Criteron (AIC) improvement.

Evaluation the significant variables were based upon a variable importance ranking through 20 trials of 5-fold cross-validation (5-fold CV). The 5-fold CV can be described as the following step. The learning set was shuffled and randomly partitioned into five portions with approximately equal size. In this research, the sample size in each portion was around 600. Four portions were used as a training set to create the logistic regression model. The remaining portion was used to validate the model. This procedure was repeated five times with each portion serving as the validation set. The 5-fold CV was conducted 20 times to produce 100 logistic regression models. The stepwise selection was applied for its variable selection procedure in each model. In this study, we applied the 5-fold CV in such a way that the training set (Four portions) was used to build to model, but the remaining portion was ignored. For 100 logistic regression models, each variable that was chosen into the models was counted. As an illustration, a variable that was used in all models was given a count of 100. A variable in this study was deemed to be significant if it appeared at least 70 times out of 100, or 70 percent of the time.

The probability response obtained from the logistic regression model was classified based on a cuttoff or a threshold of 0.5. The probability of fraud π_i was defined by the cottoff. The fraud report

was classified as a fraudulent claim (positive class) if the anticipated probability was greater than or equal to 0.5.

2.4.2 Feature selection by random forest

A significant variable from the RF algorithm is measured by variable importance. The variable importance is determined by two methods, the Mean Decrease in Impurity (MDI) and the Mean Decrease in Accuracy (MDA). A variable with a higher MDI and MDA is considered to be more significant. The Gini index is used as an impurity function in the MDI method, and this method is also known as the Mean Decrease Gini. The Gini index $i(t)$ is defined as

$$i(t) = 1 - \sum_{j=1}^{c-1} [p(j|t)]^2 , \tag{16}$$

where $p(j|t)$ is the function of observations belonging to the class j at a given node t and c is the number of class. The index $i(t)$ is used to assess the importance of a variable by adding the weighted impurity decreases for all nodes and calculating the average over all N_T trees in the RF (Moon et al., 2019).

The variable importance in the MDA method is determined by measuring mean decrease in the out-of-bag (OOB) accuracy for each tree. The importance of each variable is computed by the mean decrease in OOB accuracy before and after a random permutation of each variable (Baek et al., 2008). Over the N_T trees, the MDA takes the average difference in accuracy values between the OOB data and the permuted OOB data.

Like the parametric method, 20 trials of 5-fold CV were performed on the RF algorithm. The resulting 100 RF models were obtained to find the significant variables. Each RF model was implemented with 500 decision trees. The MDI and MDA assessed variable importance ranking. The variable importance of each variable was recorded and averaged over 100 RF models using the MDI and MDA.

2.5. Software used

The results in this paper were obtained using R version 4.1.1 (R Core Team, 2021). The main functions used in this research were listed as follows. The function char_cor() in the package creditmodel proposed by Fan (2022) was used to determined the Cramer'V correlation. The function cor() in R was used to deterimined the Spearman's rank correlation. The packages stats, a part of R and MASS provided by Venables and Ripley (2002) were used in the feature selection step. A logistic regression model, a special case of a generalized linear model, was implemented by the glm() function with the stats package. The stepwise procedure was performed by the stepAIC() function with the MASS package. The AIC improvement determined which variables were included or excluded in the model. The random forest was applied with the randomForest() function in the randomForest package presented by Liaw and Wiener (2002). The Naïve Bayes was executed with the naiveBayes() function by the e1071 package offered by Meyer et al. (2019). The adaptive boosting was accomplished with the ada() function in the ada package contributed by Culp et al. (2006). In the caret package proposed by Kuhn (2008), data was segregated to be training and testing sets with the createFolds() function for the k-fold cross-validation approach, and a confusion matrix was created and the assessment criteria were calculated using confusionMatrix() function.

2.6. Methodology

This section describes the research workflow for modeling and evaluating the fraud report classification as shown in Figure 1. The details of each step were described as follows.

Step 1: Interested data was identified and collected.

Step 2: Data preparation was conducted including data cleansing, multicollinearity check, and feature scaling.

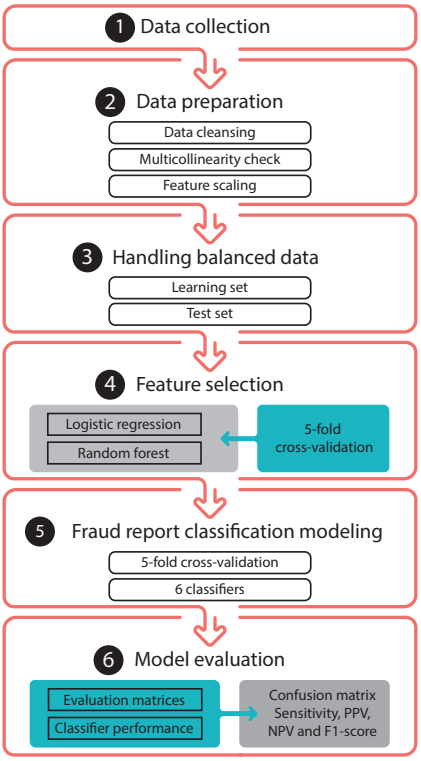

Figure 1 The Research Workflow

Step 3: Unbalanced data was handled via a stratified random sampling. As explained in Section 2.1, the learning set was created for building all the models in this work, and the test set was used for the model assessment.

Step 4: Feature selection was analyzed by the stepwise logistic regression and random forest as described in Section 2.4. Variable importance ranking was identified via 20 trials of 5-fold cross-validation. Sets of important features were obtained and used in the next step.

Step 5: The six algorithms including logistic regression (LG), random forest (RF), Naïve Bayes with features from logistic regression (NB-LG), Naïve Bayes with features from random forest (NB-RF), adaptive boosting with features from logistic regression (AB-LG), and adaptive boosting with features from random forest (AB-RF) were used to create predictive models.

Step 6: The 5-fold cross-validation was performed 20 times to produce 100 models for each algorithms, and the corresponding confusion matrices were calculated to assess the performance of the algorithms. The test set was used as the data validation for each model.

Step 7: Assessment criteria of the algorithms were computed. Finally, conclusion and discussion were addressed.

3. Results

3.1. Feature selection

The variable importance suggested by logistic regression is determined via their importance ranking. Significant predictors are those variables chosen for inclusion in predictive models and appearing at least 70 times out of 100, or 70 percent of all 100 models. Hence, there are 12 significant features recommended by logistic regression: insured_hobbies, incident_type, incident_serverity, authorities_contacted, auto_year, policy_year, policy_month, insured_occupation, incident_state, witnesses, incident_hour_of_the_day, and auto_make.

The variable importance suggested by random forest is presented in Figure 2, left for MDA and right for MDI. The variable importance ranking plots start with the most important variable at the top of the plots. In our case, the insured_hobbies is the most important predictor for both methods. The remaining variables are then ranked as next most important until the least important variable is reached at the bottom of the two plots. The variable importance from MDA and MDI criteria shows similar ranking in the top 15. The two methods provide the same set of predictors and differ only in order of the variable importance. Consequently, the following 15 features were chosen based on the learning accuracy: insured_hobbies, policy_year, auto_year, incident_serverity, auto_make, insured_occupation, policy_month, insured_education_level, insured_relationship, age, authorities_contacted, incident_state, incident_hour_of_the_day, incident_type, and the annual_premium.

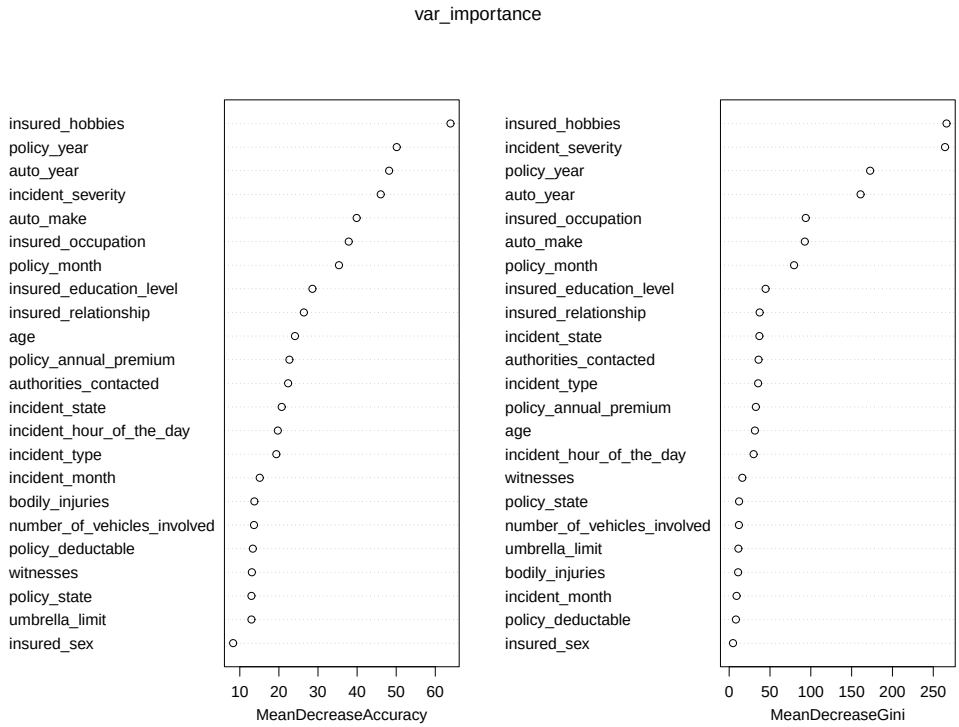

Figure 2 Variable Importance by Random Forest

3.2. Algorithm performance

A performance comparison of the six classification algorithms: LG, RF, NB-LG, NB-RF, AB-LG, and AB-RF is illustrated in Table 4. Averaging their 100 models yields the four assessment criteria for each algorithm. The mean and corresponding standard deviation (SD) presented in the bracket is reported in the table. The algorithm that performed best in each criterion is presented in boldface. For easier consideration, a graphical presentation called a box plot is illustrated in Figures 3-6 for sensitivity, positive predictive value (PPV), negative predictive value (NPV), and F1-score, respectively. The six boxes from each algorithm are illustrated in different colors. The coral box demonstrates the AB-LG, the olive box refers to the AB-RF, the green one presents the LG, the turquoise one indicates the NB-LG, the royal blue box displays, and pink one represents the RF.

According to the results in Table 4 and Figure 3, the RF algorithm has the highest sensitivity (0.9919), followed by the AB algorithms (0.9884 and 0.9862), then the NB algorithms (0.9814 and

Table 4 Model performance comparison

	Sensitivity	NPV	PPV	F1
LG	0.9718	0.7363	0.7489	0.8379
	(0.0018)	(0.0040)	(0.0168)	(0.0027)
RF	**0.9919**	0.6640	**0.9362**	0.7955
	(0.0001)	(0.0090)	(0.0000)	(0.0065)
AB-LG	0.9884	0.7881	0.8915	0.8769
	(0.0017)	(0.0048)	(0.0157)	(0.0031)
AB-RF	0.9862	**0.7908**	0.8702	**0.8777**
	(0.0029)	(0.0070)	(0.0274)	(0.0047)
NB-LG	0.9814	0.6957	0.8447	0.8142
	(0.0029)	(0.0041)	(0.0247)	(0.0028)
NB-RF	0.9756	0.7096	0.7915	0.8216
	(0.0031)	(0.0066)	(0.0262)	(0.0050)

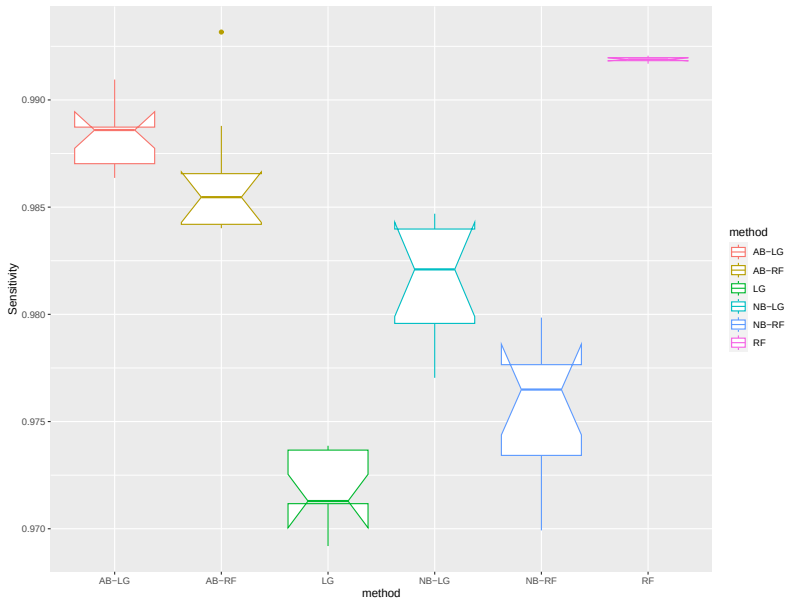

Figure 3 Sensitivity

0.9756), and the LG algorithm has the lowest sensitivity (0.9718). The non-parametric algorithms appear to outperform the parametric one in terms of sensitivity. Further, the difference between the AB algorithms is insignificant.

Based on the findings in Table 4 and Figure 4, the RF algorithm provides the highest PPV (0.9362), followed by the AB-LG and AB-RF algorithms (0.8915 and 0.8702), then the NB-LG and NB-RF algorithms (0.8447 and 0.7915), and the LG algorithm has the lowest PPV (0.7489). In terms of PPV, the non-parametric algorithms seem to perform better than the parametric one. Additionally, the difference between the AB algorithms is negligible.

According to the findings in Table 4 and Figure 5, the AB-RF offers the highest NPV (0.7908), followed by the AB-LG (0.7881), then the LG (0.7363), next the NB-RF (0.7096), then the NB-LG (0.6957) and the RF has the lowest NPV (0.6640). Unfortunately, the RF works poorly for identifying the negative class (non-fraud case). Moreover, the two AB algorithms are comparative in terms of NPV.

The results in Table 4 and Figure 6 follow the same pattern of the NPV criterion. The AB-RF has the highest F1-score (0.8777), followed by the AB-LG (0.8769), then the LG (0.8379), next the NB-RF (0.8216), then the NB-LG (0.8142) and the RF has the lowest F1-score (0.7955). Unfortunately, the RF performs poorly when it comes to identifying the negative class (non-fraud case). Besides, the two AB algorithms are comparative in terms of F1-score.

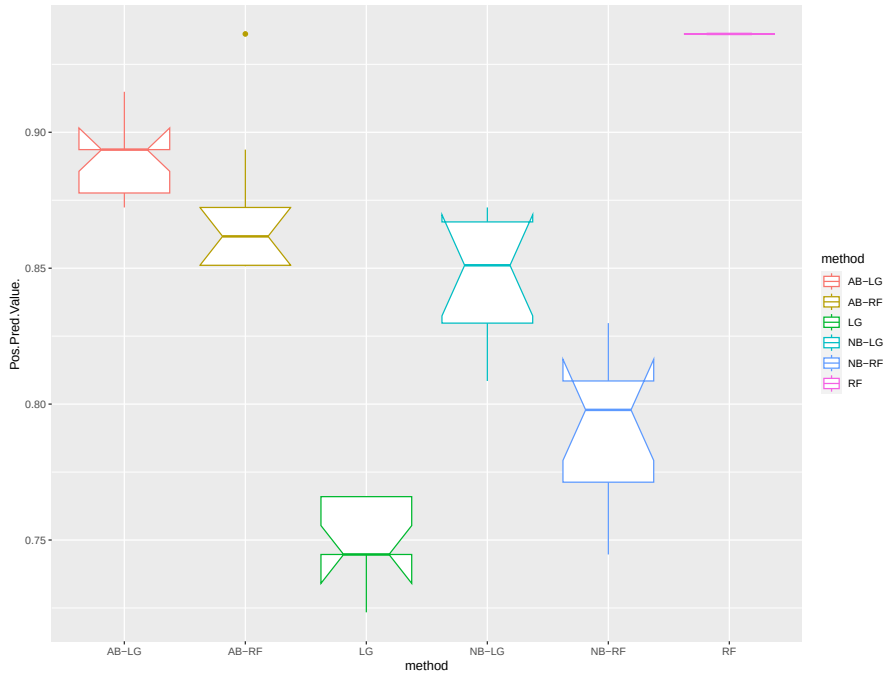

Figure 4 Positive predictive value

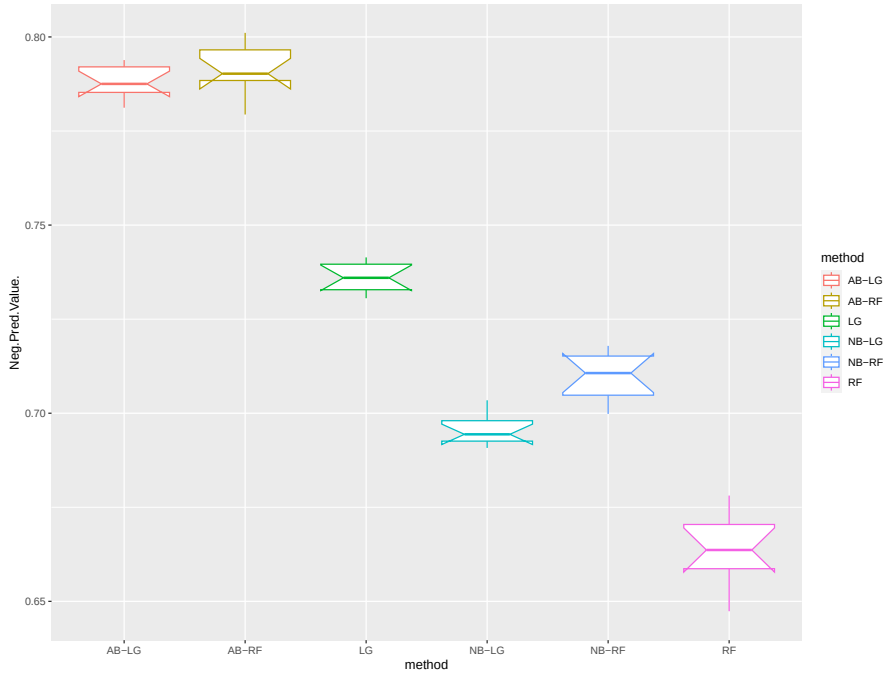

Figure 5 Negative predictive value

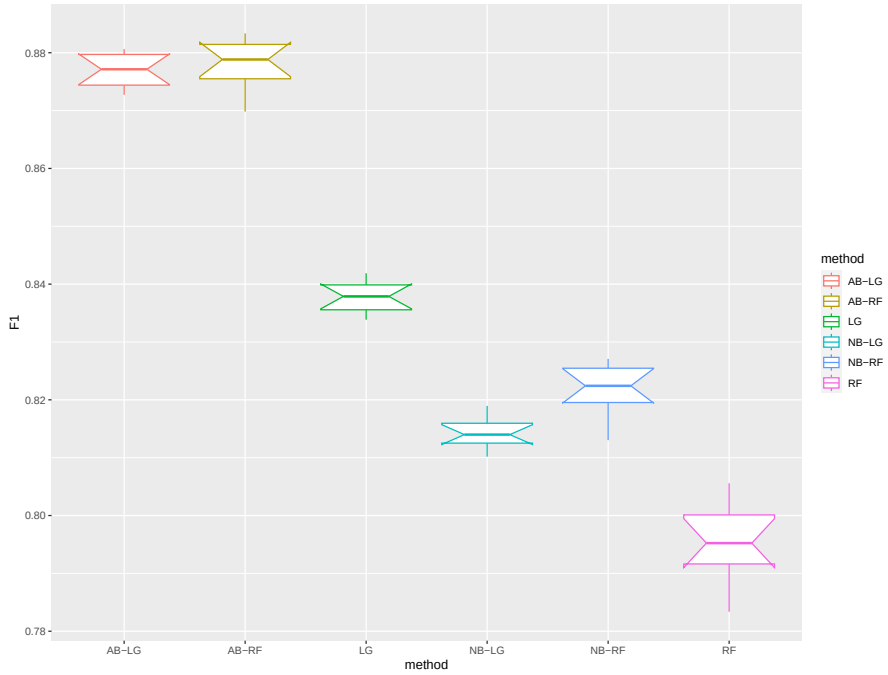

Figure 6 F1-score

4. Conclusion and Discussion

We now return to the questions that motivate this work. The first question concerns what features should be used to detect fraudulent claims. To address this, the stepwise logistic regression is used to determine significant features as the parametric selection method. The results show that there are 12 crucial features as follows; demographic variables such as insured_occupation and insured_hobbies, the time when the policies start such as policy_month and policy_year, the automobile involved in the claim such as auto_make and auto_year, and the details of the claim such as incident_type, incident_serverity, authorities_contacted, incident_hour_of_the_day, incident_state and witnesses. On the other hand, there are 15 significant predictors suggested by random forest, the non-parametric selection method as follows. The are 11 common variables for the two methods: insured_occupation, insured_hobbies, policy_month, policy_year, auto_make, auto_year, incident_type, incident_serverity, authorities_contacted, incident_hour_of_the_day, and incident_state. The additional important variables related to demographic variables such as age, insured_education_level, and insured_relationship. The last variable is policy_annual_premium. These variables obtained correspond to the study of Moon et al. (2019).

The results indicate that the algorithm performance depends on feature selection methods. Overall, the non-parametric method provides more superior performance than the parametric method. The reason for this may be due to most ML algorithms in this work are non-parametric statistical methods and thus they work better together with non-parametric variable selection method than those of the parametric method. However, in some assessment criteria in particular sensitivity, all six algorithms have slightly different sensitivity.

The second question is the decision-making algorithms used to classify whether a claim is fraudulent or not. To answer this, data mining techniques were employed using ML algorithms. The six classification algorithms: LG, RF, NB-LG, NB-RF, AB-LG, and AB-RF were selected. The 5-fold cross-validation was also used to build 100 predictive models for each algorithms. The corresponding confusion matrices were computed to assess the performance of the algorithms. The results reveal that the RF outperforms other algorithms in terms of sensitivity and PPV as expected. The RF algorithm is still an effective classifier. A high value of sensitivity and PPV is very important since an insurance claim dataset includes the number of fraudulent claims was significantly far fewer than the number of non-fraudulent claims. The two criteria measure a predictive capability of classifiers in the positive class or the fraud case. Sensitivity is explained as the probability that the model identifies the claim is fraudulent among all fraudulent claims. The PPV is described as the probability that the claim classified as fraud by the model is actually a fraudulent claim. Therefore, the random forest algorithm is treated as the best model to identify fraudulent claims.

Surprisingly, the AB-RF provides the highest NPV and F1-score. In other words, the AB-RF algorithm can well identify non-fraudulent claims when considering the NPV. An algorithm with a high value of F1-score indicates the well predicted algorithm for both positive (fraud) class and negative (non-fraud) class. Interestingly, the AB-RF and AB-LG provide comparative results for all criteria. Therefore, both AB algorithms offer nice performance, and one can use as an alternative classification algorithm for identifying fraudulent claims. Compared to other non-parametric classifiers, the NB algorithm gives low performance. Although, the NB algorithm is better than the LG algorithm, the parametric classifier, in terms of sensitivity and PPV. Conversely, the LG algorithm is better than the NB algorithm in terms of NPV and F1-score.

When the individual LG and RF algorithms are compared, the RF outperforms the LG in terms of sensitivity and PPV. In terms of F1-score and NPV, however, the LG outperforms the RF. In other words, the RF can accurately classify fraudulent claims while the LG can correctly identify non-fraudulent claims.

The main purpose of this work is to offer a general statistical learning strategy for detecting fraudulent claims in the automobile insurance industry in a practical use rather than a data-driven analysis. In the initial step, variable screening is determined by the Cramer's V and the Spearman's correlations for categorical and continuous variables, respectively. In addition, the Point-biserial

correlation can measure relationship between categorical and continuous variables. However, this method is limited in practice since it works only a dichotomous variable. In the second step, the stratified random sampling is used to handling unbalanced data problem. The training set and test set are obtained. The learning set is created from the training set by a combination of the ten sets of data randomly selected without replacement. In the third step, variable importance ranking from parametric and non-parametric methods is achieved by a k-fold cross-validation. In the fourth step, the predictive models are built by replicating a k-fold cross-validation. In the final step, the model evaluation is conducted using the test set as new case of claims. This research could provide some benefit to auto-insurance companies for their fraud detection strategy to minimize monetary loss.

For future directions of this work, other classification algorithms such as support vector machines should be considered, and other features used for creating predictive models should be investigated by suitably other feature selection methods. The methods used to overcome unbalanced data should be also examined.

5. Acknowledgements

We would like to thank the referees for their useful suggestions on the manuscript. The authors gratefully acknowledge the financial support provided by the Faculty of Science and Technology, Contact No. SciGR 3/2565. This work is also supported by the Thammasat University Research Unit in Data Learning.

References

Aksoy S, Haralick RM. Feature normalization and likelihood-based similarity measures for image retrieval. Pattern Recogn Lett. 2001; 22(5): 563-582.

Baek S, Moon H, Ahn H, Kodell RL, Lin CJ, Chen JJ. Identifying high-dimensional biomarkers for personalized medicine via variable importance ranking. J Biopharm Stat. 2008 Sep 5; 18(5): 853-868.

Belhadji EB, Dionne G, Tarkhani F. A Model for the Detection of Insurance Fraud. Geneva Pap Risk Insur Issues Pract. 2000; 25(4): 517-538.

Belyakov SL, Karpov SM. Identity of Fraudulent Financial Operations using the Machine Learning Algorithm. Vestnik Komp'iuternykh i Informatsionnykh Tekhnologii. 2020; 188: 023-031.

Berrar D. Bayes theorem and naive bayes classifier. In: Ranganathan S, Gribskov M, Nakai K, Schnbach C, editors. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. Oxford: Academic Press; 2019.

Culp M, Johnson K, Michailides G. ada: An R package for stochastic boosting, J Stat Softw. 2006; 17(2): 1-27.

Fan D. creditmodel Toolkit for Credit Modeling, Analysis and Visualization. R Package Version 1.3.1. 2022 [cite 2022 Dec 20]. Available from: https://CRAN.R-project.org/package=creditmodel.

Gareth J, Daniela W, Trevor H, Robert T. An Introduction to Statistical Learning: with Applications in R, ser. Springer texts in statistics. New York: Springer; 2013.

Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. The 32^{nd} International conference on machine learning; 2015, June; pmlr; 2015. p. 448-456.

Kowshalya G, Nandhini M. Predicting Fraudulent Claims in Automobile Insurance. Proceeding IEEE Conference on Inventive Communication and Computational Technologies; 2018 Apr 20-21; India. Coimbatore: IEEE; 2018. pp. 1338-1343.

Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008; 28(1): 1-26.

Liaw A, Wiener M. Classification and regression by randomforest, R news. 2002; 2(3): 18-22.

Maua G, Grbac TG, Bai BD. Multivariate logistic regression prediction of fault-proneness in software modules. Proceeding IEEE of the 35^{th} International Convention on Information and Communication Technology, Electronics and Microelectronics; 2012 May 21-25; Croatia. Opatija: IEEE; 2012. pp. 698-703.

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. Misc Functions of the Department of Statistics (E1071), TU Wien. R J. 2019.

Moon H, Pu Y, Ceglia C. A predictive modeling for detecting fraudulent automobile insurance claims. Theor Econ Lett. 2019; 9(6): 1886-1900.

Priya KU, Pushpa S. A survey on fraud analytics using predictive model in insurance claims. Int J Pure Appl Math. 2017; 114(7): 755-767.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. 2021.

Randhawa K, Loo CK, Seera M, Lim CP, Nandi AK. Credit card fraud detection using AdaBoost and majority voting. IEEE Access. 2018; 6: 1427714284.

Roy R, George KT. Detecting insurance claims fraud using machine learning techniques. Proceeding IEEE Conference on Circuit Power and Computing Technologies; 2017 Apr 20-21; India. Kollam: IEEE; 2017. pp. 1-6.

Schapire RE. Explaining adaboost. Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik. Berlin: Springer; 2013.

Sharma R. Fraud-detection-in-insurance-claims. Kaggle. 2020 [cite 2021 May 20]. Available from: https://www.kaggle.com/roshansharma/fraud-detection-in-insurance-claims/data.

Venables WN, Ripley BD. Modern Applied Statistics with S. New York: Springers; 2002.

Viaene S, Derrig RA, Baesens B, Dedene G. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. J Risk Insur. 2002; 69(3): 373-421.

Wang Y, Xu W. Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decis Support Syst. 2018; 105: 87-95.

Zou H. Analysis of Best Sampling Strategy in Credit Card Fraud Detection Using Machine Learning. Proceeding of the 6^{th} International Conference on Intelligent Information Technology; 2021 Feb 25-28; Vietnam. Ho Chi Minh: Association for Computing Machinery; 2021. pp. 40-44.