Application of Data Mining Techniques in Automobile Insurance Fraud Detection

Kannat Na Bangchang
Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University
Pathum Thani, Thailand
kannat@mathstat.sci.tu.ac.th

Sangdao Wongsai
Teerawat Simmachan*
Department of Mathematics and Statistics, Thammasat University Research Unit in Data Learning, Faculty of Science and Technology, Thammasat University
Pathum Thani, Thailand

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 policy holders because of the increasing in premium amounts. Therefore, detection of insurance fraud is a challenging problem. Traditional approaches are hard 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. This work would help in a screening process to investigate claims, thus minimizing human resources and monetary losses. Three sets of features are obtained by logistic regression models: one with forward selection, one with backward elimination, and one without variable selection. Three algorithms including Naïve Bayes, random forest and adaptive boosting are employed as classifiers. K-fold cross validation is used to evaluate the algorithm performance. The results suggest that the smaller number of features, the better performance. The random forest performs the best with highest accuracy (85.28%), sensitivity (93.85%), and precision (97.91%) whereas the adaptive boosting provides the highest specificity (70.41%) and F-score (89.86%).

CCS CONCEPTS

• **Computing methodologies** → **Boosting**; **Cross-validation**; **Machine learning**; *Classification and regression trees.*

KEYWORDS

Naïve Bayes, Random Forest, Adaptive Boosting, Variable Selection

ACM Reference Format:
Kannat Na Bangchang, Sangdao Wongsai, and Teerawat Simmachan. 2023. Application of Data Mining Techniques in Automobile Insurance Fraud Detection. In *2023 6th International Conference on Mathematics and Statistics (ICoMS 2023), July 14–16, 2023, Leipzig, Germany*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3613347.3613355

ICoMS 2023, July 14–16, 2023, Leipzig, Germany

ACM ISBN 979-8-4007-0018-7/23/07…$15.00
https://doi.org/10.1145/3613347.3613355

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 premiums are collected each year [16]. 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 [2]. 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 [16]. 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 [9]. When fraudulent claims are undetected, insurance companies increase the premium amount to compensate for the loss. Sincere policy holders 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 [2]. 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 automobile fraudulent claims via data mining techniques. A survey on fraud analytics using predictive models in insurance claims was provided in 2017[21]. A case study on fraud diagnosis using machine learning was proposed in 2002[23]. 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[16]. Decision tree and random forest algorithms were better than Naïve Bayes algorithm. A case study of auto-insurance fraud detection using deep learning with text analysis was proposed in 2018[24]. The results showed that machine learning algorithms were more effective than logistic regression. Simulation study on predicting fraudulent claims in automobile insurance using data mining techniques was submitted in 2018[9]. The random forest algorithm performed the best. A predictive modeling for detecting fraudulent automobile insurance

claims using parametric and non-parametric statistical learning algorithms together with a cross-validation technique was proposed in 2019[14]. 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. The second question concerns a decision making algorithm to classify whether a claim is classified as fraudulent or not. To address these questions, the performance of feature selection is compared with 3 ways: no feature selection, using forward selection, and using backward elimination in the preliminary step. Based on the literature review, it suggests that random forest is noticed as an effective algorithm. Naïve Bayes is selected as a control algorithm because it is one of the early methods. Finally, adaptive boosting is chosen as a challenging algorithm since there are few studies using the boosting algorithm for detecting fraudulent claims. However, there are many studies on fraud detection of other aspects via the boosting technique. For instance, the study of credit card fraud detection in 2018[15] and 2021[25], and a case study of fraudulent financial operations in 2020[3]. The three algorithms are implemented to predict fraudulent claims. Real data set from the anonymous insurance company in the United States in 2015 is used. The k-fold cross validation is applied, and a confusion matrix is calculated to evaluate the algorithm performance. This work would offer some benefit to the insurance companies for their fraud detection strategy to minimize human resources and monetary losses.

2 MACHINE LEARNING ALGORITHMS

2.1 Naïve Bays

Naïve Bayes algorithm denoted by NB is classification technique using Bayes theorem [4]. 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

$$P(y_j|x_i) = \frac{P(x_i|y_j)P(y_j)}{P(x_i)} \tag{1}$$

which we can plug into equation (1) and we obtain

$$P(y_j|x) = \frac{\prod_{k=1}^{n} P(x_k|y_j)P(y_j)}{P(x)} \tag{2}$$

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 equation (3). 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:

$$\hat{y} = arcmax \prod_{k=1}^{n} P(x_k|y_j)P(y_j) \tag{3}$$

A model that implements equation (3) is called a (simple) Naïve Bayes classifier. Probabilities are computed differently for nominal and numeric attributes.

2.2 Random Forest

Random forest proposed by [5] and denoted by RF provides an improvement over bagged trees by way of a small tweak that decorates the trees [19]. A number of decision trees is constructed on the basis of bootstrapped samples. However, when creating these decision trees, each time a split in a tree is considered, a random sample of m predictors from the full set of p predictors is chosen as split candidates. 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 $m \approx \sqrt{p}$. In this study, each RF model is implemented with 500 decision trees. For classification tasks, the random forest output is the class chosen by the majority of trees.

2.3 Adaptive Boosting

Adaptive Boosting is a statistical classification algorithm formulated by [7] and is denoted by AB in this study. It solves many of the practical difficulties of the earlier boosting algorithms [17]. It can be combined with a variety of other types of learning algorithms to improve performance. The results of the other learning algorithms or weak learners are merged to create a weighted total that represents the boosted classifier's final results. The AB is typically used for binary classification, but it can be extended to multiple classes or bounded intervals on the real line. The algorithm is adaptive in the sense that subsequent weak learners are biased toward misclassified instances by previous classifiers.

3 VALIDATION TOOLS

3.1 K-Fold Cross-Validation

This procedure involves randomly dividing the set of observations into k portions or folds of approximately equal size [19]. One portion is treated as a validation set or a test set, and the remaining $k-1$ portions are used as a training set to build a predictive model. 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.

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} accuracy_i \tag{4}$$

3.2 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

Table 1: Confusion Matrix

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

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.

3.3 Assessment Criteria

The algorithms' performance is measured using accuracy, sensitivity (also known as recall), specificity, precision, and the F-score, which is the harmonic mean of precision and sensitivity. The associated formulas are listed below. The greater the value, the greater the performance.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$specificity = \frac{TN}{FP + TN} \tag{7}$$

$$precision = \frac{TP}{TP + FP} \tag{8}$$

$$F - score = \frac{2 \cdot precision \cdot sensitivity}{precision + sensitivity} \tag{9}$$

4 RESEARCH METHODOLOGY

4.1 Data Description

Claim details of an insurance company are mostly confidential. To illustrate the proposed strategy, the auto-insurance dataset was chosen from the online source [18]. The data provide information on claims in automobile insurance policies of the anonymous insurance company in the United States, and collected from January 1, 2015, to March 1, 2015. The dataset contains 26 predictor variables and 1,000 records with a dichotomous response. There is no missing in this dataset. Tables 2-3 present descriptions of the predictor variables, which are divided into 15 categorical variables and 11 continuous variables. The response variable is fraud report (fraudulent or not). It is reported in Figure 1 that there are 247 fraudulent claims (24.7%) and 753 non-fraudulent claims (75.3%).

4.2 Data Preparation

4.2.1 Multicollinearity Check. Logistic regression was used for variable selection algorithm. The required assumption of the model

Figure 1: Bar chart of fraud report

Table 2: Categorical Variables

Feature Name	Description
policy_month	The month when the policy starts
policy_year	The year when the policy starts
policy_state	The city when the policy starts
insured_sex	Insured's gender
insured_education	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 in which the incident occurred
auto_make	Insured car brand
auto_year	The age of the insured car

Table 3: Continuous Variables

Feature Name	Description
age	Insured age
deductable	Amount of money that the insured must pay before a company pays a claim
annual_premium	Annual premium
umbrella_limit	Coverage limit for umbrella insurance that is made with car insurance
number_of_vehicles	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	Time of incident

was no multicollinearity problem. Therefore, correlations between predictor variables was considered in the fist step. If the correlations between predictor variables were greater than 0.60, the corresponding variables were considered strongly correlated with other

predictor variables and were eliminated from the next step. The categorical variables' associations were measured using Cramer's V. The relationship between the continuous variables was determined using Spearman's rank correlation. 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.

4.2.2 Feature Scaling. The continuous variables were normalized since their range of values 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 [1]. In this work, the Z-score Normalization was selected as a feature scaling method. The algorithms converged more faster with feature scaling than without it, which was another factor that made feature scaling necessary. [8].

4.3 Feature Selection Method

A binary logistic regression was used for variable selection method. Three sets of important features were obtained as follows: one with forward selection, one with backward elimination, and one without variable selection. The first two procedures determine whether variables should be added to the model and whether variables already in the model should be removed [12]. In this study, variable inclusion and exclusion from the model were based on the Akaike Information Criteron (AIC) improvement. The model without variable selection was called the full model using 23 variables.

4.4 Model Building and Evaluation

In this research, all possible 180 scenarios were generated by the three algorithms, the k-fold cross validation with k = 2, 4, 5 and 10, three feature selection methods, and five assessment criteria. The three algorithms were used to create predictive models of the fraud report. To evaluate the effectiveness of the algorithms, the k-fold cross-validation method was applied, a confusion matrix was built, and the evaluation criteria were determined.

4.5 Software Used

Data analysis in this paper was implemented using R version 4.2.2 [20]. The packages stats and MASS provided by [22] 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, and the forward and backward procedures were performed by the stepAIC() function with the MASS package. In the caret package proposed by [10], 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 by the confusionMatrix() function. The three algorithms were implemented as follows. Firstly, the Naïve Bayes was executed with the naiveBayes() function by the e1071 package offered by [13]. Secondly, the random forest was applied with the randomForest() function in the randomForest package presented by [11]. Finally, the adaptive boosting was accomplished with the ada() function in the ada package contributed by [6].

5 RESULTS

5.1 Feature Selection

There are twelve features suggested by forward selection: the coverage limit for umbrella insurance, number of witnesses, the month when the insurance policy starts, the year when the insurance policy starts, the city when the insurance policy starts, hobbies of the insured, the marital status of the insured, the severity of the accident, persons to contact after the incident, time of incident, insured car brand, and the age of the insured car. However, there are only three features suggested by backward procedure: the coverage limit for umbrella insurance, hobbies of the insured, and the severity of the accident.

5.2 Algorithm Performance

The results in each assessment criterion are presented in Table 4-8. The algorithm that performs best in each criterion is highlighted in boldface in each k-fold option. We compare the algorithm's performance in various k-fold options and feature selection methods, as well as in the overall scene.

The results in Table 4 show that backward elimination provides the highest accuracy followed by forward selection. The model without feature selection produces the lowest accuracy. In general, the performance of the algorithms follows the same pattern for each k value except for k = 2 and the full model. In that case, the AB has the lowest accuracy. For all k cases, the NB works best for forward selection. Furthermore, the RF and AB algorithms produce comparable results, although with slightly lower accuracy than the NB. Moreover, all algorithms are comparable when k = 2. Except for k = 2, the RF performs best for backward elimination. The AB provides the highest accuracy in that case. However, the difference between RF and AB algorithms is negligible. The NB gives the lowest accuracy for all k cases. For the full model, the NB algorithm has the best result for all k cases. However, the difference between NB and AB algorithms is insignificant except the case of k = 2 where the AB has the lowest accuracy. The RF algorithm works poorly with large numbers of features.

According to Table 5, backward elimination has the highest sensitivity, followed by forward selection. The full model has the least sensitivity. Except for k = 2 and the full model, the performance of the algorithms follows the same pattern in each k value. The AB has the lowest sensitivity in this case. For all k cases, the NB and RF algorithms produce the best results for forward selection. The two algorithms differ only marginally. The NB provides the lowest sensitivity. The RF clearly outperforms in terms of backward procedure. Furthermore, the AB outperforms the NB in all k cases. For the full model, the NB gives the best result for all k cases. Besides, the AB outperforms the RF in all k cases except k = 2, where the AB has the lowest sensitivity. With a large number of features, the RF algorithm performs poorly. NB and AB algorithms, on the other hand, work well.

In terms of specificity, Table 6 suggests that forward and backward selection methods produce comparable results. The full model offers the lowest specificity. Except for the full model, the performance of the algorithms in the overall scene follows the same pattern in each k value. The performance of the NB and RF algorithms depends on the k values under the full model.

Table 4: Accuracy Comparison

algorithm	k=2			k=4		
	forward	backward	full model	forward	backward	full model
NB	0.8060	0.8330	0.7890	0.8195	0.8290	0.7890
RF	0.8010	0.8530	0.7490	0.7970	**0.8450**	0.7590
AB	0.8020	**0.8580**	0.5506	0.7979	0.8330	0.7870
algorithm	k=5			k=10		
	forward	backward	full model	forward	backward	full model
NB	0.8230	0.8380	0.7981	0.8209	0.8261	0.7921
RF	0.8080	**0.8510**	0.7580	0.8190	**0.8622**	0.7580
AB	0.8130	0.8450	0.7900	0.8011	0.8531	0.7920

Table 5: Sensitivity Comparison

algorithm	k=2			k=4		
	forward	backward	full model	forward	backward	full model
NB	0.8472	0.8672	0.8604	0.8574	0.8784	0.8746
RF	0.8569	**0.9431**	0.7593	0.8436	**0.9249**	0.7673
AB	0.8150	0.9057	0.5489	0.8156	0.8870	0.8097
algorithm	k=5			k=10		
	forward	backward	full model	forward	backward	full model
NB	0.8633	0.8739	0.8761	0.8638	0.8720	0.8778
RF	0.8613	**0.9394**	0.7633	0.8652	**0.9465**	0.7649
AB	0.8285	0.9007	0.8062	0.8162	0.9000	0.8140

Table 6: Specificity Comparison

algorithm	k=2			k=4		
	forward	backward	full model	forward	backward	full model
NB	0.6380	0.6982	0.5725	0.6804	0.6660	0.5652
RF	0.6079	0.6582	0.4656	0.6166	0.6561	0.5859
AB	0.7071	**0.7129**	0.6642	**0.6847**	0.6671	0.6384
algorithm	k=5			k=10		
	forward	backward	full model	forward	backward	full model
NB	0.6734	0.7111	0.5889	0.6633	0.6684	0.5732
RF	0.6237	0.6611	0.5667	0.6567	0.6764	0.6074
AB	**0.7202**	0.6847	0.6751	0.7044	**0.7099**	0.6448

Table 7: Precision Comparison

algorithm	k=2			k=4		
	forward	backward	full model	forward	backward	full model
NB	0.9057	0.9190	0.8592	0.9098	0.8977	0.8406
RF	0.8845	0.8566	**0.9761**	0.8965	0.8646	**0.9761**
AB	0.9535	0.9057	0.9495	0.9465	0.8924	0.9376
algorithm	k=5			k=10		
	forward	backward	full model	forward	backward	full model
NB	0.9097	0.9177	0.8526	0.9056	0.9017	0.8421
RF	0.8885	0.8593	**0.9841**	0.9016	0.8659	**0.9801**
AB	0.9482	0.8938	0.9495	0.9468	0.9058	0.9390

Clearly, the three algorithms perform the worst in terms of specificity when compared to the other assessment criteria. For forward selection, the AB algorithm outperforms the NB algorithm for all k cases. The RF algorithm provides the lowest specificity. Backward

Table 8: F-score Comparison

algorithm	k=2			k=4		
	forward	backward	full model	forward	backward	full model
NB	0.8754	0.8923	0.8598	0.8826	0.8877	0.8572
RF	0.8701	0.8977	0.8541	0.8691	**0.8936**	0.8591
AB	0.8788	**0.9057**	0.6584	0.8760	0.8894	0.8689
algorithm	k=5			k=10		
	forward	backward	full model	forward	backward	full model
NB	0.8856	0.8949	0.8639	0.8839	0.8864	0.8589
RF	0.8745	0.8966	0.8597	0.8824	**0.9043**	0.8592
AB	0.8842	**0.8967**	0.8719	0.8764	0.9027	0.8718

Table 9: Overall Algorithm's Performance Comparison

algorithm	accuracy			sensitivity		
	forward	backward	full model	forward	backward	full model
NB	0.8173	0.8315	0.7921	0.8579	0.8729	0.8722
RF	0.8062	**0.8528**	0.7560	0.8567	**0.9385**	0.7637
AB	0.8035	0.8473	0.7299	0.8188	0.8984	0.7447
algorithm	specificity			precision		
	forward	backward	full model	forward	backward	full model
NB	0.6638	0.6859	0.5749	0.9077	0.9090	0.8486
RF	0.6262	0.6630	0.5564	0.8928	0.8616	**0.9791**
AB	**0.7041**	0.6937	0.6556	0.9488	0.8994	0.9439

algorithm	F-score		
	forward	backward	full model
NB	0.8819	0.8903	0.8599
RF	0.8740	0.8981	0.8580
AB	0.8789	**0.8986**	0.8178

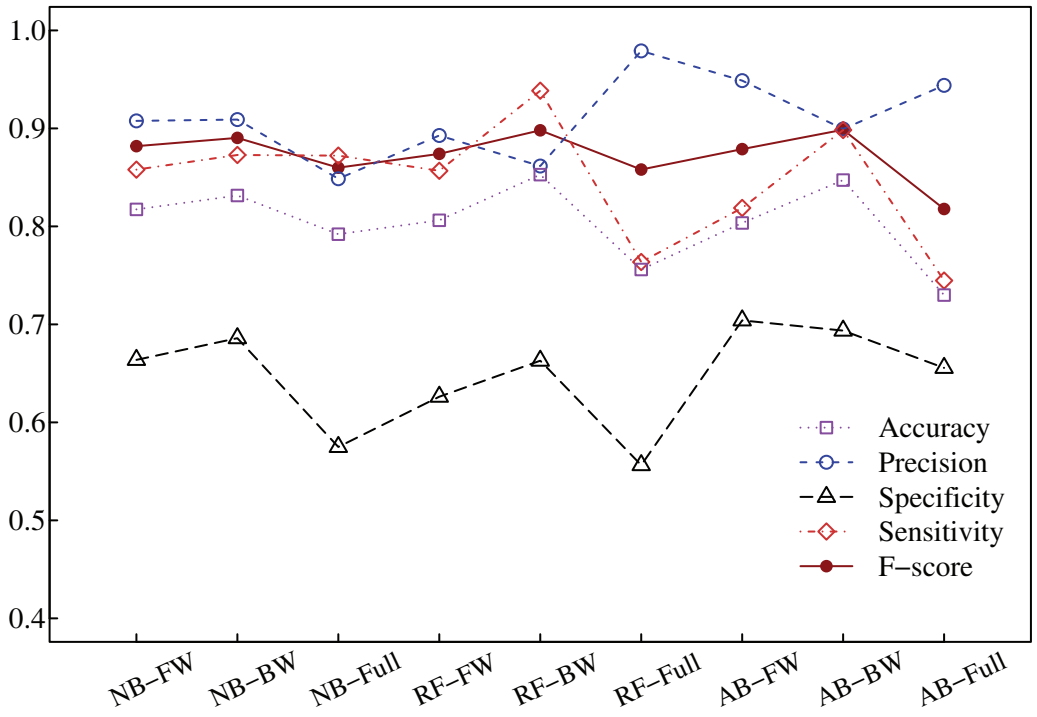

Figure 2: Overall Algorithm's Performance

elimination, like forward selection, the AB yields the best results for all k cases except for k = 5. In that case, the NB has the highest specificity. Furthermore, for k = 4, the AB and NB algorithms are comparable. The RF has the lowest specificity for all k cases. The AB algorithm still produces the best results for all k cases for the full model. Moreover, the RF outperforms the NB for k = 4 and k = 10, while the NB outperforms the RF for k = 2 and k = 5.

According to Table 7, in terms of precision the performance of the three algorithms is obviously highest compared to other evaluation criteria. Furthermore, the performance of each algorithm depends on the feature selection methods. In each feature selection method, their performance shows the same pattern in each k value. For forward selection, the AB algorithm has the best result for all k cases followed by the NB algorithm. The RF algorithm provides the lowest precision. However, the difference between the NB and RF algorithms is insignificant. Unlike forward selection, the NB algorithm has the best result for all k cases followed by the AB algorithm for backward elimination. However, the difference between the NB and AB algorithms is negligible. The RF algorithm still provides the lowest precision. Unlike the two feature selection methods, the RF algorithm gives the best result for all k cases followed by the AB algorithm for the model without feature selection. Besides, the AB gives slightly less precision than the RF. The NB algorithm provides the lowest precision.

Results in Table 8 indicate that backward elimination provides the highest F-score followed by forward selection. The full model produces the lowest F-score. However, the two models with feature selection methods have comparable results. In overall aspect, the the algorithm's performance shows the same pattern in each k value except in case of k = 2 under the full model. In that case, the AB algorithm gives the lowest F-score. In each feature selection method, the algorithm's performance is insignificantly different. For forward selection, the NB algorithm has the best result for almost all cases, while the AB algorithm performs best for the backward elimination and the full model.

The performance of the algorithms and feature selection methods regardless of k-fold cross validation is shown in Table 9. Averaging their assessment criteria from the four k-fold options and combinations of algorithms and feature selection methods yields the corresponding mean values given in the table. The algorithm that performs the best in each criterion is highlighted. Figure 1 shows a graphical representation of the information for simpler consideration. Various types of lines and markers indicate different evaluation criteria. Nine classifiers are represented on the x-axis via a combination of algorithms and variable selection methods. According to Table 9 and Figure 2, the random forest performs the best with highest accuracy (85.28%), sensitivity (93.85%), and precision (97.91%) whereas the adaptive boosting provides the highest specificity (70.41%) and F-score (89.86%). In the overall scene, models with feature selection methods work more efficiently than the model without a feature selection method. Particularly, the predictive model suggested by the backward elimination provides the best results in terms of accuracy, sensitivity and F-score. Additionally, the forward selection and the full model give the best results in terms of specificity and precision, respectively.

6 CONCLUSION AND DISCUSSION

We now return to the two questions that motivate this work. The first concerns what variables should be used to detect fraudulent claims. To address this, a logistic regression is used to gather with three feature selection methods. The results show that there are twelve crucial features suggested by forward procedure. On the other hand, there are only three features suggested by backward method. The details are described in the previous section. The significant variables include demographic variables such as 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 the details of the claim such as incident serverity and authorities contacted. These obtained variables correspond to the study of [14].

The empirical results indicate that the algorithm performance depends on feature selection methods. The backward elimination provides the best results compared to others. A small set of features offers more appropriateness of the predictive model in detecting fraudulent claims. The forward selection presents better results than the full model in several cases. The model without feature selection is unsuitable for detecting fraudulent claims. It also takes longer than other models.

The second is the decision-making algorithm used to determine whether a claim is fraudulent. To overcome this, data mining techniques were employed. The important features were implemented with the algorithms to classify the fraud report. The three algorithms: NB, RF and AB were selected as a classifier. The k-fold cross validation was used as a test option. To evaluate the algorithms' performance, a confusion matrix was created. In the overall scene, the results for the k-fold-option suggest that the level of each assessment criterion for each k follows the same pattern, except when k = 2 under the full model. In practice, one should avoid using the 2-fold option to formulate the predictive model.

According to the empirical results, the RF performs best in terms of accuracy, sensitivity, and precision, while the AB has the highest specificity and F-score. In other words, the RF algorithm can accurately predict fraudulent claims since sensitivity and precision are computed based on the True Positive in the positive class (fraud case). In contrast, the AB algorithm predicts well in both cases. The highest specificity indicates that the AB can predict well in the case of non-fraudulent claims, while the highest F-score represents that it can predict well in the case of fraudulent claims. These results were as expected. The RF algorithm remains a good classifier. The challenging algorithm, the AB algorithm, performs well. In certain cases, the control algorithm, the NB algorithm, produces the best results.

In general, insurance companies maintain their customers and claim details confidentially. However, our intention is to provide insurance companies with a fraud detection strategy in order to reduce both human resources and monetary losses. Other classifiers, such as support vector machines, should be considered in future research, and other features used to create predictive models should be investigated using appropriate other feature selection methods.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their useful suggestions on the manuscript.

REFERENCES

[1] Selim Aksoy and Robert M Haralick. 2001. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern recognition letters* 22, 5 (2001), 563–582. https://doi.org/10.1016/S0167-8655(00)00112-4

[2] El Bachir Belhadji, Georges Dionne, and Faouzi Tarkhani. 2000. A Model for the Detection of Insurance Fraud. *The Geneva Papers on Risk and Insurance. Issues and Practice* 25, 4 (2000), 517–538. http://www.jstor.org/stable/41952549

[3] S. L. Belyakov and S. M. Karpov. 2020. IDENTIFY OF FRAUDULENT FINANCIAL OPERATIONS USING THE MACHINE LEARNING ALGORITHM. *Vestnik Komp'iuternykh i Informatsionnykh Tekhnologii* 188 (2020), 23–31. https://doi.org/10.14489/vkit.2020.02.pp.023-031

[4] Daniel Berrar. 2018. Bayes' theorem and naive Bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* 403 (2018), 412.

[5] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[6] Mark Culp, Kjell Johnson, and George Michailides. 2007. ada: An r package for stochastic boosting. *Journal of statistical software* 17 (2007), 1–27.

[7] Yoav Freund and Robert E. Schapire. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory*, Paul Vitányi (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 23–37.

[8] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 448–456. https://proceedings.mlr.press/v37/ioffe15.html

[9] G. Kowshalya and M. Nandhini. 2018. Predicting Fraudulent Claims in Automobile Insurance. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, Coimbatore, India, 1338–1343. https://doi.org/10.1109/ICICCT.2018.8473034

[10] Max Kuhn. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* 28, 5 (2008), 1–26. https://doi.org/10.18637/jss.v028.i05

[11] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[12] Goran Mauša, Tihana Galinac Grbac, and Bojana Dalbelo Bašić. 2012. Multivariate logistic regression prediction of fault-proneness in software modules. In *2012 Proceedings of the 35th International Convention MIPRO*. IEEE, Opatija, Croatia, 698–703.

[13] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2022. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.* The Comprehensive R Archive Network. https://CRAN.R-project.org/package=e1071 R package version 1.7-12.

[14] Hojin Moon, Yuan Pu, Cesarina Ceglia, et al. 2019. A Predictive Modeling for Detecting Fraudulent Automobile Insurance Claims. *Theoretical Economics Letters* 9, 06 (2019), 1886. https://doi.org/10.4236/tel.2019.96120

[15] Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, and Asoke K. Nandi. 2018. Credit Card Fraud Detection Using AdaBoost and Majority Voting. *IEEE Access* 6 (2018), 14277–14284. https://doi.org/10.1109/ACCESS.2018.2806420

[16] Riya Roy and K. Thomas George. 2017. Detecting insurance claims fraud using machine learning techniques. In *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*. IEEE, Kollam, India, 1–6. https://doi.org/10.1109/ICCPCT.2017.8074258

[17] Robert E. Schapire. 2013. *Explaining AdaBoost.* Springer Berlin Heidelberg, Berlin, Heidelberg, 37–52. https://doi.org/10.1007/978-3-642-41136-6_5

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

[19] Fariha Sohil, Muhammad Umair Sohali, and Javid Shabbir. 2022. An introduction to statistical learning with applications in R. *Statistical Theory and Related Fields* 6, 1 (2022), 87–87. https://doi.org/10.1080/24754269.2021.1980261

[20] R Core Team. 2021. R: A language and environment for statistical computing. Published online 2020. Http://Www.R-Project.Org

[21] I Ulaga Priya and S Pushpa. 2017. A survey on fraud analytics using predictive model in insurance claims. *International Journal of Pure and Applied Mathematics* 114, 7 (2017), 755–767.

[22] WN Venables and Brian D Ripley. 2002. Statistics and computing: modern applied statistics with S. *Springer-Verlag, New York Inc, New York. https://doi. org/10* 1007 (2002), 978–0.

[23] Stijn Viaene, Richard A Derrig, Bart Baesens, and Guido Dedene. 2002. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk and Insurance* 69, 3 (2002), 373–421. https://doi.org/10.1111/1539-6975.00023

[24] Yibo Wang and Wei Xu. 2018. Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems* 105 (2018), 87–95. https://doi.org/10.1016/j.dss.2017.11.001

[25] Hanbing Zou. 2021. Analysis of best sampling strategy in credit card fraud detection using machine learning. In *2021 6th International Conference on Intelligent Information Technology*. Association for Computing Machinery, Ho Chi Minh, Viet Nam, 40–44. https://doi.org/10.1145/3460179.3460186