International Journal of Electrical and Computer Engineering (IJECE)
Vol. 14, No. 1, February 2024, pp. 911~921
ISSN: 2088-8708, DOI: 10.11591/ijece.v14i1.pp911-921

Predicting automobile insurance fraud using classical and machine learning models

Shareh-Zulhelmi Shareh Nordin[1], Yap Bee Wah[1,2], Ng Kok Haur[3], Asmawi Hashim[4], Norimah Rambeli[4], Norasibah Abdul Jalil[4]
[1]School of Computing Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Selangor, Malaysia
[2]UNITAR Graduate School, UNITAR International University, Selangor, Malaysia
[3]Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, Kuala Lumpur, Malaysia
[4]Department of Economics, Universiti Pendidikan Sultan Idris, Perak, Malaysia

Article Info

Article history:

Received Mar 26, 2023
Revised Jul 3, 2023
Accepted Jul 17, 2023

Keywords:

AdaBoost
Data science
Fraud detection
Insurance fraud
Machine learning
Tree augmented naïve Bayes

ABSTRACT

Insurance fraud claims have become a major problem in the insurance industry. Several investigations have been carried out to eliminate negative impacts on the insurance industry as this immoral act has caused the loss of billions of dollars. In this paper, a comparative study was carried out to assess the performance of various classification models, namely logistic regression, neural network (NN), support vector machine (SVM), tree augmented naïve Bayes (NB), decision tree (DT), random forest (RF) and AdaBoost with different model settings for predicting automobile insurance fraud claims. Results reveal that the tree augmented NB outperformed other models based on several performance metrics with accuracy (79.35%), sensitivity (44.70%), misclassification rate (20.65%), area under curve (0.81) and Gini (0.62). In addition, the result shows that the AdaBoost algorithm can improve the classification performance of the decision tree. These findings are useful for insurance professionals to identify potential insurance fraud claim cases.

Corresponding Author:

Yap Bee Wah
UNITAR Graduate School, UNITAR International University
Jalan SS6/3, 47301 Petaling Jaya, Selangor, Malaysia
Email: bee.wah@unitar.my

1. INTRODUCTION

The insurance industry has become a fundamental pillar of our modern world for many years. This industry is relevant to society as it can offer financial security in the event of an accident or illness. However, some irresponsible individuals have taken advantage of making false insurance claims to obtain compensation or benefits which is called insurance fraud. Fraudulent claims are a serious offense as they not only threaten the insurer's or policyholder's profitability but also harmfully affect the insurance industry and the existing social and economic systems.

For the past few decades, the insurance fraud claims problem has been unresolved from the beginning of the insurance industry due to a lack of resources, research, documentation, and technology. However, in the late nineteenth century, the growing systematic data collection allowed the use of pattern recognition techniques such as regression, neural network (NN), and fuzzy clustering [1]. In the late twentieth century, data played a central role in the insurance industry, and today, most insurance companies have obtained more access than before. When the size of databases grows, the traditional approach may overlook a significant portion of fraud as it is difficult to manually review large databases. With emerging technology, this issue can be solved using machine learning models to identify and predict fraud claims [2]–[4].

Several machine learning models have been proposed to accurately predict fraudulent automobile insurance claims. A comparative study by Viaene *et al.* [5] considered various machine learning models which include logistic regression, decision tree (C4.5), naïve Bayes (NB), k-nearest neighbor (KNN), Bayesian neural network (BNN), support vector machine (SVM) and tree augmented naïve Bayes (TAN) for predicting fraudulent claims in automobile insurance in the state of Massachusetts in 1993. The performances were measured with the mean percentage correctly classified and the area under the receiver operating characteristic (ROC). Experimental results showed that the logistic regression and SVM with linear kernel performed excellent predictive capabilities while the naive Bayes also performed well, and decision tree model results are rather disappointing. However, there is no discussion on the distribution of fraudulent data and whether it is balanced in nature. The random forest model was used by Li *et al.* [6] to predict automobile insurance fraud claims using an empirical example. The empirical results show that the random forest model has better accuracy and robustness. In addition, this model is suitable for large data sets and unbalanced data. In [7], several machine learning models (random forest, decision tree (J48), and naïve Bayes) were used to predict fraudulent automobile insurance claims. The random forest model was found to outperform the other models in terms of accuracy, precision, and recall. Similarly, Pranavi [8] compared the accuracy of detecting automobile insurance fraud claims using the random forest and KNN models and found that the random forest model outperforms the KNN model. They also noted the bad performance of the KNN model is due to unevenly distributed data. The findings by Prasasti *et al.* [9] also indicated that random forest outperformed other classifiers namely multilayer perceptron (MLP) and decision tree C4.5, with 98.5% accuracy. Recently, Aslam *et al.* [4] compared the predictive performance of three models (logistic regression, SVM, and naïve Bayes). Their results showed SVM has high accuracy but very low sensitivity. This is due to the problem of imbalanced data. Overall, the logistic regression model performed better than SVM and naïve Bayes, and logistic regression achieved the highest F-measure score.

Recently, some hybrid models have been developed as an extension of the existing machine learning models to improve prediction accuracy. A Bayesian learning neural network was used for automobile insurance fraud detection [10] while Tao *et al.* [11] proposed a dual membership fuzzy SVM (DFSVM) with radial basis function as kernel function for detecting automobile insurance fraud claims to overcome the overfitting problems of SVM models. However, the choice of kernel function used was not discussed. The result based on the cases in Beijing showed that the DFSVM outperforms the regular SVM models evidenced by their F-score, recall, and precision. Sundarkumar and Ravi [12] introduced a novel hybrid undersampling method based on the one-class SVM (OCSVM) and k-reverse nearest neighbor (kRNN) to improve the performance of machine learning models for detecting automobile insurance fraud claims. Their results showed that all models under the hybrid approach improved the overall performance when compared with the same models trained under the original unbalanced distributed data. Decision tree and SVM models performed better than logistic regression and MLP models. In addition, the decision tree model is computationally faster and the "if-then" decision rules are easy to understand. Wang and Xu [2] proposed a latent Dirichlet allocation (LDA)-based text analytics and deep learning model for automobile insurance fraud claims detection. Experimental results show that the combined deep learning NNs and LDA framework outperform those widely used machine learning models, such as random forest and SVM combined with LDA. Other hybrid models include a principal component analysis based random forest with nearest neighbor method [13], an artificial bee colony-based kernel ridge regression [14], back propagation neural network with an improved adaptive genetic algorithm [15], unsupervised deep learning variable importance method [16] and eRFSVM which entails random forest and SVM [17].

Various new complex machine learning algorithms such as AdaBoost [18] and XGBoost [19] have been reported to show potential in better prediction performances, especially for imbalanced and large datasets [20]. Itri *et al.* [21] compared the performance of ten machine learning models including the AdaBoost algorithm and found that the random forest model performs better among all the models compared. Meanwhile, Rukhsar *et al.* [22] conducted a comparative analysis of various machine learning models including the AdaBoost algorithm on insurance fraud prediction. The decision tree was reported to have the highest accuracy of 79% compared to the other models. In addition, the AdaBoost algorithm shows an accuracy of 78% which is comparable to the decision tree model. In the study on the predictive performance of the logistic regression and XGBoost models for the frequency of motor insurance claims, the XGBoost model achieved slightly better than logistic regression. Although the XGBoost model is better in terms of performance, it can easily overfit the data and thus, requires numerous model-tuning procedures to prevent overfitting [20]. Similar findings by Abdelhadi *et al.* [23] also showed that the XGBoost algorithm outperformed NN, J48 and naïve Bayes. However, in [3], the random forest was found to perform better than XGBoost in predicting automobile insurance claims. Useful and informative reviews on machine learning models with applications to automobile insurance fraud claims can be found in [24]–[26].

Literature studies have shown that there is no conclusive evidence that any specific machine learning model outperforms the others for predicting automobile insurance claims. The performance of

machine learning models depends on many factors, such as data quality and model settings. The novelty of this paper is the evaluation of machine learning classification techniques for automobile insurance fraudulent prediction. We also emphasized the importance of the data preparation stage to ensure data quality and statistical tests for identifying the relevant independent variables. The aim of this paper is to analyze the performance of the classical statistical model (logistic regression) with machine learning models (NN, SVM, decision tree, random forest, AdaBoost algorithm) and a semi-naïve Bayesian learning model (TAN). All the models were evaluated using various performance metrics.

The remainder of the paper is organized as follows: section 2 describes the empirical data and data preparation phase. Section 3 discusses the classical statistical model and machine learning models as well as the evaluation criteria to assess the model performance. Discussions of the results are provided in section 4. Section 5 concludes the paper.

2. DATA DESCRIPTION AND DATA PREPARATION PHASE

The car insurance claim data which consists of a dependent variable, fraudulent insurance claim (denoted as *claim_flag*), and 23 independent variables with a total of 10,299 cases were retrieved from Kaggle website [27]. These independent variables can be further categorized into four major groups including drivers' demographic factor: age of the driver (denoted as *age*), gender (*gender*), maximum education level (*education*), job category (*occupation*), income (*income*), marital status (*mstatus*), single parent (*parent*1), number of children (*homekids*), year on the job (*yoj*), home value (*home_val*), number of driving children (*kidsdriv*), distance to work (*travtime*) and home/work area (*urbancity*), vehicle factor: vehicle age (*car_age*), value of the vehicle (*bluebook*), type of car (*car_type*), vehicle use (*car_use*) and red car (*red_car*), insurance factor: number of a claim for the past 5 years (*claim_freq*), the total claim for the past 5 years (*oldclaims*) and time in force (*tif*) and license/charges factor: license revoked for the past 7 years (revoked) and motor vehicle record points (*mvr_points*).

As data preparation is one of the important stages in the data mining process, data cleaning and transformation such as removing illogical data, data reclassification, data encoding, data imputation and binning of continuous variables were carried out via the IBM SPSS Modeler 18. After removing illogical data for age, a total of 10,191 cases are used for the analysis. We observed that the distribution of the dependent variable *claim_flag* with no fraud claim (73.59%) is much higher than with fraud claims (26.41%) which indicates that the *claim_flag* data is imbalanced. Twenty-three independent variables with ten continuous variables and thirteen categorical variables. After performing data cleaning and data transformation, we reclassify three categorical variables (*homekids*, *kidsdriv*, and *clam_freq*), eight variables are encoded (*gender*, *education*, *mstatus*, *parent1*, *urbancity*, *car_type*, *car_use*, and *revoked*), six variables with missing values are imputed (*age*, *occupation*, *income*, *yoj*, *home_val*, and *car_age*), binning was carried out for continuous variables (*income*, *home_val*, *bluebook*, and *oldclaims*). We first carried out chi-squared tests to determine the significance of the relationship between the dependent variable (*claim_flag*) with each of the independent variables. Results show that the value of the vehicle (*bluebook*) and red car (*red_car*) were not significant at the 5% level. Therefore, these two variables were omitted, and the remaining twenty-one independent variables were used for developing the machine learning model. A complete "clean" dataset is available upon request.

3. RESEARCH METHOD

Many machine learning models are used in insurance fraud detection. Among them, the logistic regression model ranked the top with the most common model, followed by NNs, Bayesian belief network, decision trees, and naïve Bayes [24]. As the logistic regression model is easy to implement, this classification model is still widely used in many recent applications [28].

Due to the complexity of the data, machine learning models have shown their potential in classification problems [29]. In this paper, we evaluate the empirical classification performance of seven predictive models, namely: logistic regression, NN, SVM, TAN, decision tree, random forest, and AdaBoost.

a. Logistic regression is a common model for the prediction of a dichotomous dependent variable based on one or more independent variables. Different model selection settings including enter, forward, and backward stepwise are considered.
b. NN is a useful machine learning model for categorical and continuous dependent variables. NN has three or more layers that are interconnected. Each hidden layer consists of a few neurons. These neurons send data to the deeper layers, which in turn will send the final output data to the output layer. Through the backpropagation method, each time the output is labeled as an error during the supervised training phase, the information is sent backward to update the weights until the error is minimized [30].
c. SVM [31] is a classification technique that obtains the optimum separation hyperplane between the target (fraud or not fraud) in a multidimensional space that maximizes the separation between the two classes.

The performance of SVM greatly depends on the choice of the kernel function. Among the kernel functions considered include sigmoid, radial basis function (RBF), linear and polynomial [31], [32].

d. Naïve Bayes model is based on the Bayes theorem with an assumption of independence among independent variables, X. It involves the calculation of the posterior probability ($Y|X$), from $P(Y)$, $P(X|Y)$ and $P(X)$ where Y is the dependent variable. In this model, continuous features are converted to categorical types [32]. Meanwhile, TAN is a semi-naive Bayesian learning model. It relaxes the naïve Bayes attribute independence assumption by employing a tree structure, in which each attribute only depends on the class and one other attribute [33].
e. Decision tree is a non-parametric supervised machine learning algorithm which predicts the target variable by building a tree based on some splitting criteria. If the dependent variable is of continuous type, the model is called a regression tree, while a classification tree involves a categorical dependent variable. The tree structure begins with the parent (or root) node and splits into child nodes and ends with the leaf nodes. Each node of a classification tree is split based on a splitting criterion which is either the Gini index, entropy, or chi-square. In IBM SPSS Modeler, the classification and regression tree (CART) algorithm uses the Gini splitting criteria while C5 and chi-squared automatic interaction detection (CHAID) use the entropy and chi-square, respectively. In the classification tree, each leaf node represents a decision rule to predict the class of the dependent variable [32].
f. Random forest is an extension of the decision tree model. The model built the forest in a random manner and the massive number of trees in the forest provides an ensemble model for prediction. It increases model predictive accuracy by using bootstrap samples of the training data and random feature selection in tree induction. 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 where $m \approx \sqrt{p}$. Prediction is made by aggregating (majority vote or averaging) the predictions of the ensemble [34].
g. The adaptive boosting algorithm, named AdaBoost [18] is a machine learning model for regression and classification problems that produces a predictive model in the form of an ensemble of weak predictive models. It constructs the model in a stage-wise manner as the other methods of boosting. The outputs of weak learner models are combined into a weighted sum that will represent the final output of the weighted classifier. AdaBoost is adapting in the sense that subsequent weak learners are adjusted to support those samples that were misclassified by the previous classifier. The steps involved in the boosting process can be found in [18], [35].

In order to assess the performance of these predictive models, accuracy is the most commonly used performance metric [36]. However, this metric does not really capture the effectiveness of a classifier for imbalanced data. Therefore, many other performance metrics have been proposed in terms of error and fitness [37]. We evaluate the predictive performance using the following seven metrics.

a. $Accuracy = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$ where TP is true positive (fraud predicted as fraud), TN is true negative (non-fraud predicted as non-fraud), FP is false positive (non-fraud predicted as fraud) and FN is false negative (fraud predicted as non-fraud).
b. $Sensitivity/Recall = \frac{\text{TP}}{\text{TP+FN}}$
c. $Specificity = \frac{\text{TN}}{\text{TN+FP}}$
d. $Precision = \frac{\text{TP}}{\text{TP+FP}}$
e. $Misclassification\,rate\ (Error) = \frac{\text{FN+FP}}{\text{TP+TN+FN+FP}}$
f. Area under the ROC curve (AUC) was calculated for the ROC curve. The closer the AUC value is to one, the better the model.
g. Gini coefficient is between 0 and 1. The higher the Gini coefficient, the better the model.

4. RESULTS AND DISCUSSION

A total of 10,191 cases was divided into two groups in which 90% (9,171 cases) of the dataset is used in the training and testing stages for the models and the remaining 10% (1,020 cases) from the dataset for the deployment stage. We then partitioned 9,171 cases into 6,389 cases (70% of data) for training the model. The remaining 2,782 cases (30% of data) were used for testing the performance of the models.

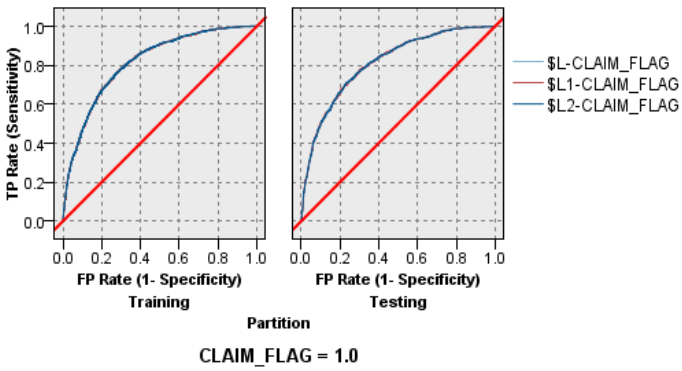
Six different machine learning models, namely logistic regression (enter, forward and backward), NN (MLP and RBF), SVM (sigmoid, RBF, linear, and polynomial), TAN, random forest (number of trees: 50 and 100) and decision tree (CHAID, CART, and C5) and a boosting algorithm (AdaBoost with CHAID, CART and C5) were evaluated using several performance metrics. Each model was evaluated for different settings and the performance results are shown in Table 1.

Table 1. Performance metrics of various models for the training and testing stages

		Performance metric						
Model	Setting	Accuracy	Sensitivity	Specificity	Precision	Error	AUC	Gini
Logistic regression	Enter	78.21	40.38	91.91	63.92	21.79	0.82	0.63
		(79.22)	(41.76)	(92.50)	(66.38)	(20.18)	(0.81)	(0.62)
	Forward	78.34	41.03	91.74	64.11	21.66	0.82	0.63
		(79.19)	(41.62)	(92.50)	(66.30)	(20.81)	(0.81)	(0.62)
	Backward	**78.34**	**41.03**	**91.74**	**64.11**	**21.66**	**0.82**	**0.63**
		(79.19)	**(41.62)**	**(92.50)**	**(66.30)**	**(20.81)**	**(0.81)**	**(0.62)**
NN	**MLP**	**78.38**	**41.39**	**91.68**	**64.13**	**21.62**	**0.80**	**0.61**
		(78.07)	**(41.21)**	**(91.14)**	**(62.24)**	**(21.93)**	**(0.80)**	**(0.59)**
	RBF	74.38	21.49	93.38	53.86	25.62	0.74	0.47
		(75.74)	(22.80)	(94.50)	(59.50)	(24.26)	(0.74)	(0.48)
SVM	Sigmoid	73.57	0.12	100	100	26.43	0.59	0.17
		(73.81)	(0.00)	(99.95)	(0.00)	(26.19)	(0.62)	(0.25)
	RBF	86.41	63.17	94.77	81.26	13.69	0.91	0.82
		(76.89)	(45.74)	(87.93)	(57.31)	(23.11)	(0.77)	(0.55)
	Linear	**78.42**	**39.55**	**92.38**	**65.11**	**21.58**	**0.81**	**0.61**
		(79.26)	**(40.93)**	**(92.84)**	**(66.97)**	**(20.74)**	**(0.81)**	**(0.62)**
	Polynomial	98.69	96.27	99.55	98.72	1.31	0.99	0.99
		(67.90)	(46.70)	(75.41)	(40.24)	(32.10)	(0.66)	(0.31)
TAN		**79.09**	**44.05**	**91.68**	**65.55**	**20.91**	**0.82**	**0.64**
		(79.35)	**(44.70)**	**(91.62)**	**(65.39)**	**(20.65)**	**(0.81)**	**(0.62)**
Random forest	Tree = 50	100	100	100	100	0.00	1.00	1.00
		(78.32)	(35.16)	(93.62)	(66.15)	(21.68)	(0.80)	(0.60)
	Tree = 100	100	100	100	100	0.00	1.00	1.00
		(78.36)	(33.79)	(94.16)	(67.21)	(21.64)	(0.80)	(0.59)
Decision tree	**CHAID**	**86.18**	**35.46**	**90.81**	**58.10**	**23.82**	**0.78**	**0.56**
		(75.84)	**(37.23)**	**(89.53)**	**(55.76)**	**(24.16)**	**(0.76)**	**(0.53)**
	CART	75.88	28.24	93.00	59.18	24.12	0.75	0.50
		(76.74)	(30.36)	(93.18)	(61.22)	(23.26)	(0.77)	(0.53)
	C5	84.38	50.74	96.47	83.77	15.62	0.83	0.66
		(75.77)	(35.99)	(89.87)	(55.74)	(24.23)	(0.74)	(0.49)
AdaBoost	CHAID_boost	79.84	50.56	90.36	65.34	20.16	0.83	0.65
		(75.41)	(43.82)	(86.61)	(53.70)	(24.59)	(0.77)	(0.54)
	CART_boost	**78.21**	**46.18**	**89.72**	**61.76**	**21.79**	**0.78**	**0.56**
		(77.75)	**(44.23)**	**(89.63)**	**(60.19)**	**(22.25)**	**(0.78)**	**(0.56)**
	C5_boost	91.72	73.83	98.15	93.48	8.28	0.94	0.89
		(77.71)	(46.15)	(88.90)	(59.57)	(22.29)	(0.80)	(0.59)

Note: i) The values of accuracy, sensitivity, specificity, precision, and error are given in percentage; ii) Values in parentheses are the results for the testing stage.

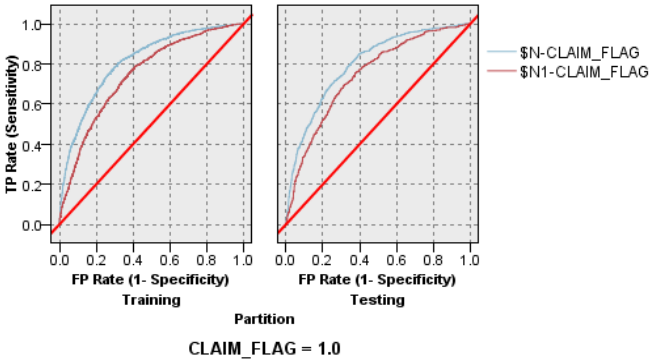
Table 1 shows the seven performance metrics results of various models at the training and testing stages. For the logistic regression model, the result shows that all variables are significant at the 5% level except age, income, parent1, *home_val* and *car_age* which are useful to form the logistic regression model regardless of the selection methods. Thus, the performance metrics also show merely identical results for the training and testing stages regardless of the selection methods. Figure 1 shows that the ROC curves of enter, forward and backward stepwise selection are identical for the training and testing stages. As all enter, forward and backward also provide identical AUC and Gini values in both stages, a logistic regression model with enter setting was selected for comparison with other models.



Note:
$L-CLAIM_FLAG=enter;
$L1-CLAIM_FLAG=forward;
$L2-CLAIM_FLAG=backward

Figure 1. ROC curves of the logistic regression models

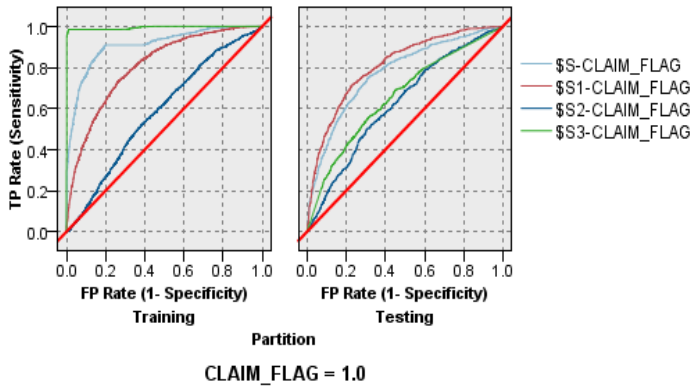
For the neural network model, *urbancity*, *mvr_points*, and *travtime* are ranked the top three important variables based on MLP, while *mvr_points*, *age*, and *travtime* are ranked the top three based on RBF. The NN model based on MLP performed better than RBF in terms of accuracy, sensitivity, precision, misclassification rate, AUC, and Gini for both stages. Although the model based on RBF has a slightly higher specificity rate than the MLP, the overall performance of MLP is still better than that based on RBF. In addition, the ROC curves given in Figure 2 show that the MLP (light blue line) model has a higher curve than RBF (maroon line) model. As a result, the NN model based on MLP was selected to be compared with other models.



Note:
$N-CLAIM_FLAG=MLP;
$N1-CLAIM_FLAG=RBF

Figure 2. ROC curves of the NN models based on MLP and RBF

The performance metrics for the SVM model using four different kernels were then evaluated. All variables are identically important for SVM (sigmoid). The top three important variables are *education*, *occupation*, and *homekids* for SVM (RBF), while *homekids*, *car_type*, and *mstatus* are important variables for SVM (polynomial) and *education*, *mstatus*, and *occupation* are the important variables for SVM (linear). Among these kernels, the linear kernel, in general, outperforms other kernels in terms of their overall metrics for the testing stage. The sigmoid kernel performs the worst in terms of its sensitivity and precision (0.00%). Among these kernels, the linear kernel outperformed the sigmoid, RBF, and polynomial kernels in terms of testing accuracy (79.26%), precision (66.97%), misclassification rate (20.74%), AUC (0.81), and Gini (0.616). The sigmoid kernel has higher specificity (99.95%) but the sensitivity and precision are very low indicating that the SVM (sigmoid) is unable to predict the true positive effectively. The result shows an overfitting issue for SVM (polynomial) as the performance metrics are very much lower for the testing stage. In Figure 3, the ROC curves show that the linear (maroon line) kernel has a consistent curve in both stages compared to RBF (light blue line), sigmoid (blue line), and polynomial (green line) models. Hence, SVM with the linear kernel is selected to be compared with other models.



Note:
$S-CLAIM_FLAG=RBF;
$S1-CLAIM_FLAG=linear;
$S2-CLAIM_FLAG=sigmoid;
$S3-CLAIM_FLAG=polynomial

Figure 3. ROC curves of the SVM models

All variables are equally important for the TAN model. The performance metrics are given by accuracy (79.35%), sensitivity (44.70%), specificity (91.62%), precision (65.39%), misclassification rate (20.65%), AUC

(0.809) and Gini (0.617) during the testing stage. Figure 4 shows the ROC curves have a similar pattern curve in the training and testing stages which indicates that the model has a consistent performance.

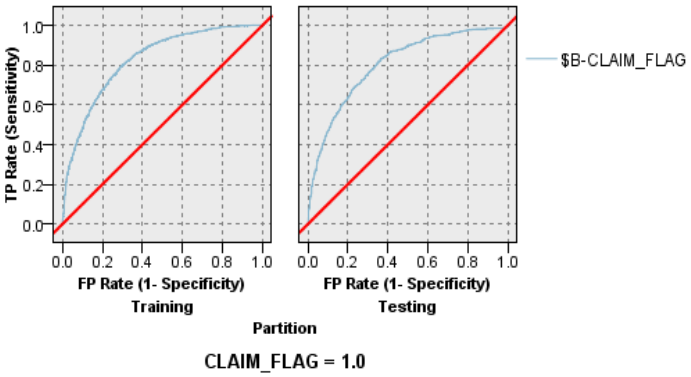


Figure 4. ROC curves of the TAN model

The random forest model with the number of trees 50 and 100 shows that the top three important variables are *age*, *travtime*, and *yoj*. All three models provide perfect accuracy, sensitivity, specificity, and precision at the training stage. However, the accuracy, sensitivity, specificity, and precision decrease at the testing stage regardless of the number of trees used which indicates that the model is overfitting. The ROC curves in Figure 5 show that the curve of each model is not consistent at both stages. Moreover, the AUC values for both stages are inconsistent. As a result, the model is not an ideal model for this insurance fraud data.

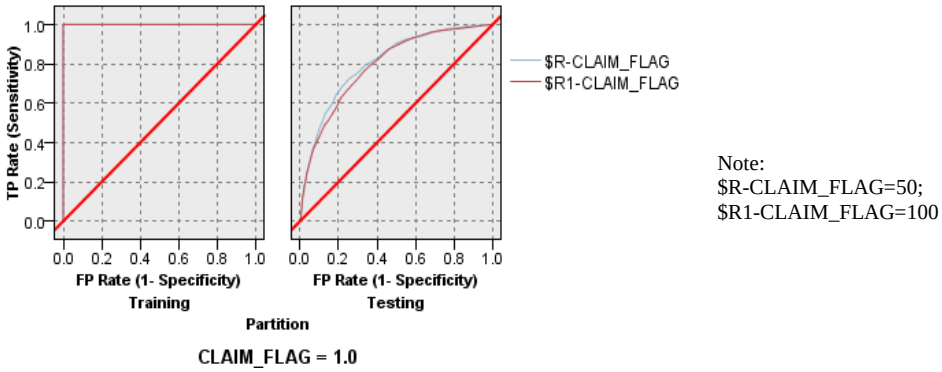


Figure 5. ROC curves of the random forest models

For the decision tree model using CHAID, CART and C5, the splitting criteria based on the Gini index show that C5 produces the highest number (170) of decision rules. The most important variables are *occupation*, *claim_freq*, and *urbancity* for the CHAID, *claim_freq*, *revoked*, and *urbancity* for the CART, while *urbancity*, *gender*, and *mstatus* for the C5. From Table 1, the seven metrics performances of CHAID and CART are consistent for both stages, while the performance of C5 is inconsistent for both stages. In the testing stage, the accuracy (76.74%), specificity (93.18%), precision (61.22%), and misclassification rate (23.26%) for the CART are slightly better than the CHAID. Furthermore, the AUC and Gini values for the CART are slightly better compared to the other model settings in the testing stage. The ROC curves in Figure 6 show that CHAID (maroon line) has slightly better curves than the CART (light blue line) and C5 (blue line) models. As a result, the decision tree with CHAID is selected for comparison with other models.

The AdaBoost algorithm in CHAID, CART, and C5 is denoted as *CHAID_boost*, *CART_boost*, and *C5_boost*. The results show all three model settings gave different important variables. The top three important variables for *CHAID_Boost* are *claim_freq*, *mvr_points*, and *urbancity*. For *CART_boost*, the top three important variables include *claim_freq*, *occupation*, and *mvr_points*, while almost all variables are equally important for *C5_boost*. For the testing stage, the accuracy (77.75%), specificity (89.63%), and precision (60.19%) of *CART_boost* with the highest values. Also, the AUC and Gini values for *CART_boost*

are more consistent for both stages. The ROC curves from Figure 7 show that *CART_boost* (blue line) curve is more consistent than *CHAID_boost* (maroon line) and *C5_boost* (light blue line) curves. Thus, *CART_boost* is selected for comparison with other models. The result shows the AdaBoost algorithm improves the performance of the decision tree.

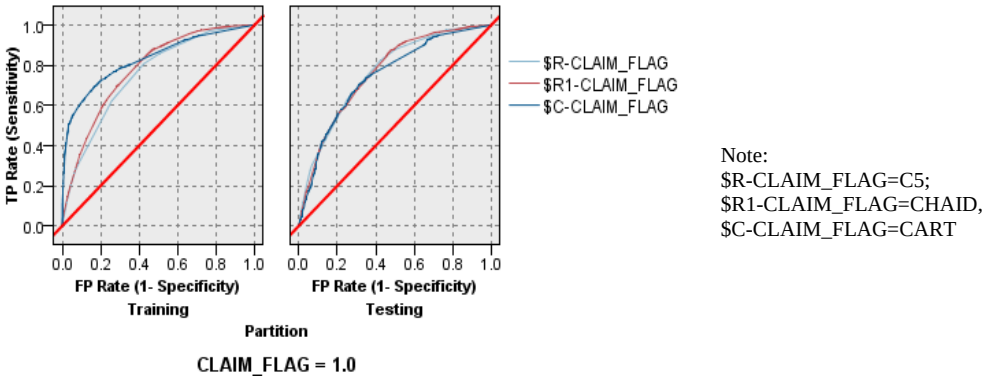


Figure 6. ROC curves for the decision tree models

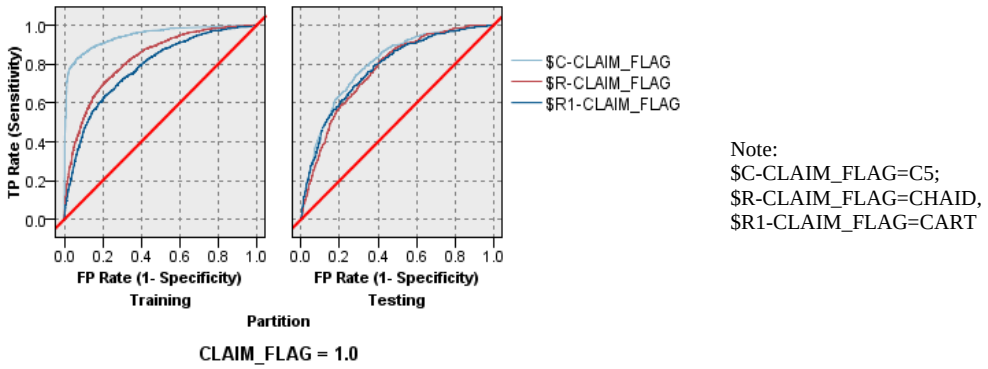


Figure 7. ROC curves for the AdaBoost models

Finally, we compare the performances among all selected models which are the logistic regression (backward), NN (MLP), SVM (linear), TAN, random forest (100), and CART_boost. The logistic regression (backward), SVM (linear), and TAN shared almost similar and higher accuracies which are 79.19%, 79.26%, and 79.35% respectively compared to the CART_boost, NN (MLP), and random forest which are 77.75%, 78.07%, and 78.15% respectively. Overall, the TAN model has the highest testing accuracy (79.35%) and sensitivity (44.70%). Thus, the TAN model was selected for deployment using 10% data (1,020 cases). Results in Table 2 show that the performance of the model at the deployment stage is quite similar to the testing stage.

Table 2. Performance metrics of TAN model for the deployment stage

Model	Performance metric				
	Accuracy	Sensitivity	Specificity	Precision	Error
TAN	78.92	44.16	91.69	66.12	21.98

5. CONCLUSION

Machine learning models are useful tools for classification and prediction problems. The performance of seven machine learning models, namely the logistic regression, decision tree, NN, TAN, SVM, random forest, and AdaBoost for decision tree with different model settings were compared using a publicly available automobile insurance fraudulent claims dataset. This study found that the TAN model has better classification performance than the other models. The result of this study concurs with other studies

that random forests are prone to overfitting issues. In addition, the result shows that the AdaBoost algorithm can improve the classification performance of the decision tree. In this study, the sensitivity of all machine learning models was less than 50%. This could be due to the fact that the dataset is slightly imbalanced with only 26.41% fraudulent cases. Future works should consider sampling techniques such as synthetic minority oversampling techniques to balance the data before applying machine learning models.

ACKNOWLEDGEMENTS

The authors thank the management of UNITAR International University for funding the publication of this paper.

REFERENCES

[1] S. Tennyson and P. Salsas-Forn, "Claims auditing in automobile insurance: fraud detection and deterrence objectives," *Journal of Risk & Insurance*, vol. 69, no. 3, pp. 289–308, Sep. 2002, doi: 10.1111/1539-6975.00024.

[2] Y. Wang and W. Xu, "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud," *Decision Support Systems*, vol. 105, pp. 87–95, Jan. 2018, doi: 10.1016/j.dss.2017.11.001.

[3] M. Hanafy and R. Ming, "Machine learning approaches for auto insurance big data," *Risks*, vol. 9, no. 2, Feb. 2021, doi: 10.3390/risks9020042.

[4] F. Aslam, A. I. Hunjra, Z. Ftiti, W. Louhichi, and T. Shams, "Insurance fraud detection: Evidence from artificial intelligence and machine learning," *Research in International Business and Finance*, vol. 62, Dec. 2022, doi: 10.1016/j.ribaf.2022.101744.

[5] S. Viaene, R. A. Derrig, B. Baesens, and G. Dedene, "A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection," *Journal of Risk & Insurance*, vol. 69, no. 3, pp. 373–421, Sep. 2002, doi: 10.1111/1539-6975.00023.

[6] Y. Li, C. Yan, W. Liu, and M. Li, "Research and application of random forest model in mining automobile insurance fraud," in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Aug. 2016, pp. 1756–1761, doi: 10.1109/FSKD.2016.7603443.

[7] G. Kowshalya and M. Nandhini, "Predicting fraudulent claims in automobile insurance," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Apr. 2018, pp. 1338–1343, doi: 10.1109/ICICCT.2018.8473034.

[8] P. S. Pranavi, "Analysis of vehicle insurance data to detect fraud using machine learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 8, no. 7, pp. 2033–2038, Jul. 2020, doi: 10.22214/ijraset.2020.30734.

[9] I. M. N. Prasasti, A. Dhini, and E. Laoh, "Automobile insurance fraud detection using supervised classifiers," in *2020 International Workshop on Big Data and Information Security (IWBIS)*, Oct. 2020, pp. 47–52, doi: 10.1109/IWBIS50925.2020.9255426.

[10] S. Viaene, G. Dedene, and R. Derrig, "Auto claim fraud detection using Bayesian learning neural networks," *Expert Systems with Applications*, vol. 29, no. 3, pp. 653–666, Oct. 2005, doi: 10.1016/j.eswa.2005.04.030.

[11] H. Tao, L. Zhixin, and S. Xiaodong, "Insurance fraud identification research based on fuzzy support vector machine with dual membership," in *2012 International Conference on Information Management, Innovation Management and Industrial Engineering*, Oct. 2012, pp. 457–460, doi: 10.1109/ICIII.2012.6340016.

[12] G. G. Sundarkumar and V. Ravi, "A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance," *Engineering Applications of Artificial Intelligence*, vol. 37, pp. 368–377, Jan. 2015, doi: 10.1016/j.engappai.2014.09.019.

[13] Y. Li, C. Yan, W. Liu, and M. Li, "A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification," *Applied Soft Computing*, vol. 70, pp. 1000–1009, Sep. 2018, doi: 10.1016/j.asoc.2017.07.027.

[14] C. Yan, Y. Li, W. Liu, M. Li, J. Chen, and L. Wang, "An artificial bee colony-based kernel ridge regression for automobile insurance fraud identification," *Neurocomputing*, vol. 393, pp. 115–125, Jun. 2020, doi: 10.1016/j.neucom.2017.12.072.

[15] C. Yan, M. Li, W. Liu, and M. Qi, "Improved adaptive genetic algorithm for the vehicle insurance fraud identification model based on a BP neural network," *Theoretical Computer Science*, vol. 817, pp. 12–23, May 2020, doi: 10.1016/j.tcs.2019.06.025.

[16] C. Gomes, Z. Jin, and H. Yang, "Insurance fraud detection with unsupervised deep learning," *Journal of Risk and Insurance*, vol. 88, no. 3, pp. 591–624, Sep. 2021, doi: 10.1111/jori.12359.

[17] M. Sathya and B. Balakumar, "Insurance fraud detection using novel machine learning technique," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 3, pp. 374–381, 2022.

[18] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.

[19] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[20] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, "Predicting motor insurance claims using telematics data-XGBoost versus logistic regression," *Risks*, vol. 7, no. 2, Jun. 2019, doi: 10.3390/risks7020070.

[21] B. Itri, Y. Mohamed, Q. Mohammed, and B. Omar, "Performance comparative study of machine learning algorithms for automobile insurance fraud detection," in *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, Oct. 2019, pp. 1–4, doi: 10.1109/ICDS47004.2019.8942277.

[22] L. Rukhsar, W. H. Bangyal, K. Nisar, and S. Nisar, "Prediction of insurance fraud detection using machine learning algorithms," *Mehran University Research Journal of Engineering and Technology*, vol. 41, no. 1, pp. 33–40, Jan. 2022, doi: 10.22581/muet1982.2201.04.

[23] S. Abdelhadi, K. Elbahnasy, and M. Abdelsalam, "A proposed model to predict auto insurance claims using machine learning techniques," *Journal of Theoretical and Applied Information Technology*, vol. 98, no. 22, 2020.

[24] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decision Support Systems*, vol. 50, no. 3, pp. 559–569, Feb. 2011, doi: 10.1016/j.dss.2010.08.006.

[25] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, Jun. 2016, doi: 10.1016/j.jnca.2016.04.007.
[26] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial fraud: a review of anomaly detection techniques and recent advances," *Expert Systems with Applications*, vol. 193, May 2022, doi: 10.1016/j.eswa.2021.116429.
[27] X. Mengsun, "Car insurance claim data," *Kaggle*. https://www.kaggle.com/datasets/xiaomengsun/car-insurance-claim-data (accessed Jan. 12, 2021).
[28] M. K. Severino and Y. Peng, "Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata," *Machine Learning with Applications*, vol. 5, Sep. 2021, doi: 10.1016/j.mlwa.2021.100074.
[29] M.-W. Hsu, S. Lessmann, M.-C. Sung, T. Ma, and J. E. V. Johnson, "Bridging the divide in financial market forecasting: machine learners vs. financial economists," *Expert Systems with Applications*, vol. 61, pp. 215–234, Nov. 2016, doi: 10.1016/j.eswa.2016.05.033.
[30] N. A. M. Salim *et al.*, "Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques," *Scientific Reports*, vol. 11, no. 1, Jan. 2021, doi: 10.1038/s41598-020-79193-2.
[31] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.
[32] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, "Introduction to data mining," Pearson Education India, 2016.
[33] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131–163, 1997.
[34] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, Nov. 2003, doi: 10.1021/ci034160g.
[35] B. W. Yap, K. A. Rani, H. A. A. Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah, "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets," in *Lecture Notes in Electrical Engineering*, Springer Singapore, 2014, pp. 13–22.
[36] B. M. Henrique, V. A. Sobreiro, and H. Kimura, "Literature review: Machine learning techniques applied to financial market prediction," *Expert Systems with Applications*, vol. 124, pp. 226–251, Jun. 2019, doi: 10.1016/j.eswa.2019.01.012.
[37] M. Naser and A. Alavi, "Insights into performance fitness and error metrics for machine learning," *arXiv preprint arXiv:2006.00887*, 2021.

BIOGRAPHIES OF AUTHORS

Shareh-Zulhelmi Shareh Nordin received a B.Sc. degree in science majoring in physics from Universiti Putra Malaysia (UPM), Malaysia, in 2018 and an M.S. degree in data science from Universiti Teknologi MARA (UiTM), Malaysia, in 2021. Currently, he is a data scientist at the Department of Data Science, Petronas Digital Sdn Bhd. His research interests include supervised machine learning, classification, boosting algorithms, and statistical analysis. He can be contacted at sharehzulhelmi@gmail.com.

Yap Bee Wah holds a Ph.D. in statistics from University of Malaya, Malaysia. She has more than 30 years of service at Universiti Teknologi MARA, Malaysia, and recently joined UNITAR International University as Director of the Research and Consultancy Centre. Her research interests are in big data analytics and data science, computational statistics, and multivariate data analysis. She was the conference chair for The International Conference on Soft Computing in Data Science (SCDS) which was held from 2015-2019 and 2021. She was also the conference chair for DaSET2022: International Conference on Data Science and Emerging Technologies. She has served as Guest Editor for Applied Soft Computing and Pertanika Journal of Science and Technology. Her research works are in the healthcare, education, environment, and business domains. She has published more than 100 papers in indexed journals and proceedings. She can be contacted at bee.wah@unitar.my.

Ng Kok Haur received a Bachelor of Science in Mathematics (2000) and a Master of Science in Statistics (2002) from Universiti Putra Malaysia and a Ph.D. in Statistics (2006) from Universiti Malaya. He is currently an associate professor at the Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya. His research interests include volatility modeling and applications, modeling and analysis of high-frequency data arising in financial economics, and statistical process control. He has published research papers in various journals including The North American Journal of Economics and Finance, Studies in Nonlinear Dynamics and Econometrics, International Review of Economics and Finance, International Journal of Industrial Engineering: Theory, Applications, and Practice, Expert Systems with Applications, and Communications in Statistics: Theory and Methods. Apart from that, he also serves as an associate editor and reviewer for journals, guest editor for a special issue of the Malaysian Journal of Science (2019), thesis examiner, and a committee member of local and international conferences. He can be contacted at kokhaur@um.edu.my.

Asmawi Hashim received degrees (1999) and master's (2002) in economics from Universiti Utara Malaysia and a Ph.D. in economics (Finance) from Sultan Idris Education University (2021). He is currently a Senior Lecturer at Sultan Idris Education University, Perak. He has authored or coauthored multiple number of articles published in refereed/indexed journals and conference papers, book chapters, and books. He has published more than 90+ papers in indexed journals and proceedings. His areas of expertise are financial economics, managerial economics, economic development, and macroeconomics. He can be contacted at asmawi@fpe.upsi.edu.my.

Norimah Rambeli received degrees (2002) and master's degrees (2004) in economics from University Kebangsaan Malaysia (UKM) and her Ph.D. in economics from University of Southampton, United Kingdom (2012). She is currently an associate professor at Universiti Pendidikan Sultan Idris (UPSI), Perak under the Economic Department, Faculty of Management and Economics (FPE). She is currently an editor team for Management Research Journal (MRJ), UPSI, and has actively published more than 174 articles in refereed/indexed journals and conference papers, book chapters, and books. Her areas of expertise are environmental economics, financial economics, applied econometrics, and time series data studies. She can be contacted at norimah@gmail.com.

Norasibah Abdul Jalil received degrees in economics from Texas Tech University, USA (1989), Master's (2000), and a Ph.D. in economics from International Islamic University (2010). She is currently an associate professor at Universiti Pendidikan Sultan Idris, Perak. She has authored or coauthored multiple number of articles published in refereed/indexed journals and conference papers, book chapters, and books. Her areas of expertise are macroeconomics, financial economics, and economic development. She can be contacted at norasibah.abduljalil@gmail.com.