Thailand Statistician
January 2025; 23(1): 144-161
http://statassoc.or.th
Contributed paper

Detecting Automobile Insurance Fraud: A Novel Two-Step Strategy Using Effective Ensemble Learning Techniques

Wikanda Phaphan [a,b], Samach Sathitvudh [c], Tikumporn Suntornsuwan [a], Kamon Budsaba [d] and Teerawat Simmachan [d]*

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

[b] Research Group in Statistical Learning and Inference, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand.

[c] Department of Statistics, School of Computer, Data and Information Sciences, University of Wisconsin-Madison, Madison, USA.

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

*Corresponding author; e-mail: teerawat@mathstat.sci.tu.ac.th

Received: 17 June 2024
Revised: 6 August 2024
Accepted: 21 August 2024

Abstract

Like other industries, insurance companies processed large volumes of data during the industrial revolution. The industry's major concern is increasing numbers of fraudulent claims. These claims affect not only financial losses but also the entire industry, honest policyholders, and society. Machine learning (ML) approaches are recently utilized in insurance fraud detection to reduce such losses. To further improve, this article introduces a novel prediction framework for fraudulent claims called the Two-step models. The anonymous US auto insurance dataset was used to demonstrate and evaluate the framework. Under-sampling and synthetic minority over-sampling technique (SMOTE) were used to balance data. Mutual information was employed as a feature selection tool. Five proposed models were built in two steps. Early on, eight basic ML models were implemented. The top three affective models were chosen based on their F-measure scores. Then, their predicted values were used as components to construct the two-step models using ensemble techniques. Statistical tests were utilized to appraise all models. Numerical results indicated that the proposed models yielded significant enhancements. Moreover, the most effective model is a combination of SMOTE and improved multilayer perceptron (IMLP). This research could help insurance firms improve their fraud detection systems to prevent insurance abuse.

Keywords: Corruption, ensembles, fraudulent claims, machine learning, security threats.

1. Introduction

During the industrial revolution, businesses sought digital transformation, with data management being key. Data science, used to find hidden insights in data, has gained popularity as data grows dramatically. It is now used in various industries, including insurance.

Fraudulent claims are the biggest insurance industry risk (Bangchang et al. 2023, Simmachan et al. 2023). Hard and soft insurance fraud depend on accident causation and claim processing (Belhadji et al. 2000). Commercial insurance carriers report increased fraud in all claims (Roy and George 2017, Moon et al. 2019). Auto insurance fraud is serious (Bangchang et al. 2023). Fraudulent claims, comprising 5-10% of all claims, cost insurance businesses $30 trillion annually, with escalating costs (Roy and George 2017, Moon et al. 2019). This is devastating to the industry and customer confidence. An effective fraud detection approach is crucial in the insurance business (Prasasti et al. 2020). A categorization model is needed to determine if a claim is false. ML classification models struggle with uneven class observations. In practice, there are few fraudulent cases and many non-fraudulent ones. ML models ignore the minority class and assign most occurrences to the common class, enhancing accuracy. Kotb and Ming (2021) expect a considerably inferior minority class prediction. To handle imbalanced data, stratified random sampling was used (Moon et al. 2019, Simmachan et al. 2023). ML models were used to compare SMOTE and other resampling strategies (Hanafy and Ming 2021, Kotb and Ming 2021). Subudhi and Panigrahi (2020) used the GA-based Fuzzy C-Means clustering to decrease noisy points through under-sampling. Results from these studies indicate that class-balanced models performed better.

Feature selection is another concern in developing effective classification models. Practical settings have many features to consider, yet predictive models may not include them all. Using feature selection techniques eliminates unneeded features and adds vital ones. However, significant aspects might be chosen differently. Moon et al. (2019) utilized k-fold cross-validation, stepwise logistic regression, variable importance scores in Random Forest, and the Least Absolute Shrinkage and Selection Operator. Forward and backward selection were compared to feature-free models (Bangchang et al. 2023). Shapley additive explanation was used to classify accident severity (Vanishkorn and Supanich 2022). Bayesian variable selection reduced heart disease classification dimensionality (Bangchang 2024). Credit card fraud detection used GA feature selection (Saheed et al. 2020). Association rules were used to select features for detecting tax fraudsters (Matos et al. 2020). Models using feature selections performed well empirically.

Ensemble learning improves insurance fraud detection algorithms. This method optimizes system performance by combining models. Based on the premise that numerous models outperform individual models in prediction. Njoh-Paul (2020) used ensemble learning, Extreme Gradient Boosting and Stacking, and compared to four individual classifiers, including support vector machine, artificial neural network, logistic regression, and linear discriminate analysis, in predicting insurance claims. Abakarim et al. (2023) utilized bagged ensemble learning of CNN models: AlexNet, InceptionV3, and Resnet101. Vosseler (2022) employed outlier ensembles with supervised surrogate models for model explanations and Mutual Information (MI) for creating ensemble models. Numerical results indicate that ensemble models outperform individual models.

ML methods are treated as effective tools and their efficacy relies on the quality of model training. This paper introduces a novel ML prediction framework called Two-step models, based on ensemble learning and other valuable methodologies. The main contributions and procedures are:

- Combining class-balancing techniques and MI as a feature selection indicator to generate five data scenarios: Original, Under, Under+MI, SMOTE, and SMOTE+MI.

□ Building eight common ML models from the previously stated works and additional works (Bhowmik 2011, Kowshalya and Nandhini 2018, Botchey et al. 2020) in the initial step and recording evaluation metrics (accuracy, precision, recall, and F-measure). F-measure is prioritized as it generalizes precision and recall simultaneously and provides the model's overall performance for the fraudulent class.

□ Selecting the best three models based on F-measure scores and using their predicted outcomes as components to create five proposed models using bagging and stacking ensemble techniques.

□ Producing 65 ML models by combining 13 ML models and 5 data scenarios. All models are developed under k-fold cross-validation, with both descriptive and inferential statistics used to identify suitable models and scenarios.

□ Unlike most earlier studies that employ individual models and only use descriptive statistics for comparison, this study selects the three most effective models based on the most important metric for ensemble model construction and employs multiple ensemble techniques.

2. Materials and Methods

Sections 2.1-2.6 outline the materials and methods used in this study. The dataset, data pre-processing steps, and the creation of five data scenarios to handle imbalanced data are described. We detail eight “One-step” ML models and five proposed “Two-step” ensemble models. model validation tools are part of the framework. Statistical tests evaluate model performance across scenarios. The research framework covers data preparation, model construction, and evaluation.

2.1. Data Description

The dataset utilized in this investigation was obtained from GitHub (Phaphan 2024) because insurance claims are usually confidential. The data offers US auto insurance claim data for an anonymous company. Collection began January 1, 2015, and ended March 1, 2015. The 1,000-observation dataset has a binary response variable, fraud report, and 26 predictors or features—14 qualitative and 12 quantitative. No data is missing from this dataset. Tables 1-2 show descriptive statistics for both aspects. There are 247 fraudulent claims (24.7%) and 753 non-fraudulent claims (75.3%). The data presented is imbalanced, which is a common trend in insurance fraud reports.

Table 1 Descriptive statistics of qualitative features

Feature name	Unique value	Top value	Frequency (%)
policy_bind_month	12	December	95 (9.50)
policy_state	3	OH	352 (35.20)
insured_sex	2	FEMALE	537 (53.70)
Insured_education_level	7	JD	161 (16.10)
insured_occupation	14	machine-op-inspct	93 (9.30)
insured_hobbies	20	reading	64 (6.40)
insured_relationship	6	own-child	183 (18.30)
Incident_type	4	Multi-vehicle Collision	419 (41.90)
incident_severity	4	Minor Damage	354 (35.40)
authorities_contacted	5	Police	292 (29.20)
incident_state	7	NY	262 (26.20)
incident_hour_of_the_day	24	17	54 (5.40)
auto_make	14	Dodge	80 (8.00)
auto_year	21	1995	56 (5.60)

Table 2 Descriptive statistics of quantitative features

Feature name	Mean	SD	Min	Max
age	38.95	9.14	19.00	64.00
policy_deductible	1,136.00	611.56	500.00	2,000.00
policy_annual_premium	1,256.41	244.05	433.33	2,047.59
umbrella_limit	1,101,000.00	2,296,257.61	−1,000,000.00	10,000,000.00
capital-gains	25,126.10	27,858.25	0.00	100,500.00
capital-loss	−26,793.7	28,090.04	−111,100.00	0.00
number_of_vehicles_involved	1.84	1.02	1.00	4.00
bodily_injuries	0.99	0.82	0.00	2.00
witnesses	1.49	1.11	0.00	3.00
injury_claim	7,433.42	4,878.51	0.00	21,450.00
property_claim	7,399.57	4,822.31	0.00	23,670.00
vehicle_claim	37,928.95	18,876.81	70.00	79,560.00

2.2. Data pre-processing

In this section, the important procedures used for data pre-processing were described as follows: data cleaning, multicollinearity check, data transformation, feature selection, handling imbalanced data, and data scenario.

2.2.1 Data cleaning

Data preparation is a necessary step before inputting into predictive models. This involves ensuring the data is in a format that can be read by ML models and removing any missing values before analysis.

2.2.2 Multicollinearity check

The absence of multicollinearity assumption in candidate classifiers such as logistic regression and Gaussian naïve Bayes led to the examination of feature correlations. If the correlations were greater than 0.70, they were treated as a serious problem, and the corresponding features were removed from further step. Spearman's rank and Cramer's V were used to measure the relationship between quantitative and qualitative features, respectively. The Spearman's correlations range from -0.05 to 0.81. The total_claim_amount is highly correlated with injury_claim and property_claim. The total_claim_amount was consequently eliminated. On the other hand, there were no serious problems with qualitative features (the range of Cramer's V correlations was 0.03 to 0.45). As a result, 25 features remained for further analysis.

2.2.3 Data transformation

One of the fundamental stages in constructing a predictive model is data transformation. The procedure converts raw data into a readable format, reduces effects of outliers, improves the functionality of ML models (Aksoy and Haralick 2001), and speeds up processing time (Ioffe and Szegedy 2015). To ensure consistency in quantitative features, the range was normalized using the feature scaling method known as Z-score Normalization. For qualitative features, dummy variable encoding was used. To accommodate the restrictions of certain ML models that only accept numerical input, this procedure involves converting categorical features into a binary vector of 0s and 1s.

2.2.4 Feature selection

Choosing relevant and useful features from a bigger set is feature selection. The technique saves computing time, improves model performance, avoids overfitting, and improves data pattern representation, especially in high-dimensional datasets (Kotb and Ming 2021, Vanishkorn and Supanich 2022). This study selected important features using MI. MI measures interdependence of two random variables. As an indicator, it ranks features by response variable influence. Features with higher MI values predict response variables better.

2.2.5 Handling imbalanced data

Imbalanced data refers to a situation in classification where the number of observations for each category (class) of a response variable is not equal. This is a common situation in real world applications, particularly in fraud detection problems. Moreover, the dataset utilized in this research is imbalanced. In this situation, the majority class holds a dominant position over the minority class. Machine learning models tend to favor the majority class, resulting in unreliable performance (Kotb and Ming 2021). This led to the implementation of methods for handling imbalanced data.

2.2.6 Data scenario

To evaluate the proposed strategy, a combination of methods dealing with imbalanced data and feature selection options was implemented. Under-sampling and SMOTE developed by Chawla et al. (2002) were adopted as class-balancing tools. In order to balance the class distribution, under-sampling results in reducing the number of observations in the majority class. SMOTE, on the other hand, interpolates between minority class instances to provide synthetic samples for the minority class. Training ML models with and without feature selection are the two alternatives available for feature selections. Consequently, there are five scenarios as follows:

I. Original data: This is an imbalanced case. We denoted by “Original”.
II. Under-sampling without feature selection: We denoted by “Under”.
III. Under-sampling with feature selection: We denoted by “Under+MI”.
IV. SMOTE without feature selection: We directly denoted by “SMOTE”.
V. SMOTE with feature selection: We denoted by “SMOTE+MI”.

2.3. ML models: One-step models

The eight basic models were chosen and labeled “One-step models”. The following are specific details for each model.

2.3.1 K-nearest neighbor (K-NN)

The K-NN examines similarity under the pre-specified value, k, and classifies new data based on most of those k neighboring components. Similarity is measured using Euclidean distance (Wang et al. 2007, Boonkrong and Simmachan 2016).

2.3.2 Naïve Bayes (NB)

NB is a popular classification method and is particularly suitable for high-dimensional inputs. NB often outperforms more complex ML models despite its simplicity. Because it calculates the probability of each input feature being predicted (Farghaly et al. 2023). The NB uses the Bayes theorem to compute the posterior probability for each class C_j. The resulting classification equation is as follows (Botchey et al. 2020);

$$P(C_j \mid x) = P(x \mid C_i)P(C_i) / P(x),$$

where
$$P(x) = \sum_j P(x \mid C_j)P(C_j).$$

$P(C_j)$ is the probability of class C_j, $P(x)$ is the likelihood density of feature x, $P(x \mid C_i)$ is the class conditional likelihood density of the feature x that belongs to the C_j class, and $P(C_i \mid x)$ is the posterior probability of the C_j class when observing x.

2.3.3 Decision tree (DT)

A decision tree is a non-parametric supervised learning algorithm. Every internal node in the DT denotes a feature test; every branch shows the test's result; and every leaf with a class label is indicated. The pathways from the root to the leaf node express the categorization rules. Furthermore, the DT can be identified as a piecewise constant approximation (Bhowmik 2011).

2.3.4 Random Forest (RF)

Random forest classifies data points based on the multiple decision trees, randomly choosing features and observations from the trees. The final predicted value comes from the most frequent results among those selected decision trees, which can be more accurate than using a single tree (Kowshalya and Nandhini 2018).

2.3.5 Adaptive boosting or AdaBoost (ADA)

A predictive model was constructed based on multiple weak predictive models from decision tree models. AdaBoost repeatedly adjusts the model's weight, depending on the number of incorrect predictions, until it reaches significantly high accuracy (Pedregosa et al. 2011).

2.3.6 Gradient boosting (GB)

An improved method of AdaBoost that alleviates prediction errors while improving accuracy consists of three parts: loss function, weak learner, and additive model. A loss function helps optimize the model. A weak learner generates predictions from observations, while an additive model works with the weak learner to get a minimized loss function (Friedman 2002).

2.3.7 Extreme gradient boosting or XGBoost (XGB)

When compared to several ML models and deep learning, XGBoost works effectively and helps minimize runtime. This method is derived from the original method of constructing non-digestible trees (Dhieb et al. 2019).

2.3.8 Logistic regression (LR)

A statistical model for modeling a qualitative response variable using either quantitative or qualitative independent variables. The data is modeled by a sigmoid curve, representing the probability of the event of interest. We define the dependent variable, $y = 1$ as denoting the occurrence of the event of interest (fraudulent) and $y = 0$ otherwise (non-fraudulent) (Harrell 2015).

2.4. Proposed models: Two-step models

There are five proposed models named "Two-step models". These models were created by combining the best three One-step models based on their F-measure scores under ensemble learning techniques. The first model was built using bagging, and the other ones were formed using stacking.

2.4.1 Majority vote (MV)

This model was developed using bagging, but it was called voting in classification tasks. The predicted values from the top three models were utilized as the three components of the ensemble learning. Then, a majority vote was applied to the three components, and the predicted outcome of the response variable was the result of voting (Pedregosa et al. 2011).

2.4.2 Improved Models

The remaining Two-step models are called "improved models". Different types of models were used to choose the four candidates. The parametric statistical model LR was chosen. GB and XGB are enhancing ensembles. A multilayer perceptron (MLP) in a neural network was used for model construction because it effectively learns non-linear function approximations for classification and regression (Yang et al. 2009). The four improved models were created by stacking the three One-step model components. Layering trains a meta-model to combine base model predictions. Averaging and voting are replaced with base model predictions in the meta-model. It helps the meta-model grasp complex base model prediction linkages, boosting performance. Base model predictions were the three components utilized as supplementary features in the improved logistic regression (ILR), improved gradient boosting (IGB), improved extreme gradient boosting (IXGB), and improved multilayer perceptron (IMLP).

2.5. Model validation tools

The crucial tools utilized for model evaluation were described in this section as follows.

2.5.1 K-fold cross-validation

In k-fold cross-validation, observations are randomly divided into k groups (folds) of approximately equal size. In each iteration, one of these folds serves as the validation set, while the remaining $k-1$ folds are used to train the model or algorithm. The evaluation metric is calculated for each iteration, resulting in k metric estimates ($metric_1, metric_2, ..., metric_k$). The final k-fold cross-validation estimate (CV) is obtained by averaging these metric values. The k-fold CV is one of the most popular cross-validation techniques. In this work, we select a common option, $k=5$.

2.5.2 Confusion matrix

A confusion matrix is often employed as classification metrics (Botchey et al. 2020, Vanishkorn and Supanich 2022). The matrix shows predicted and actual outcome counts. Fraud detection problems use positive and negative classes to represent fraudulent and non-fraudulent claims. Four outcomes are used. TP occurs when the model correctly identifies a fraud scenario. When the model correctly classifies a case as non-fraudulent, TN occurs. FP occurs when the model misclassifies a case as fraud. FN occurs when the model misclassifies a non-fraud case.

2.5.3 Evaluation metrics

The ML models can be commonly evaluated using accuracy, precision, recall, and F-measure (Kotb and Ming 2021, Vanishkorn and Supanich 2022). Model performance rises with value. Table 3 covers four evaluation metrics. Insurance companies warn about fraudulent claims more than non-fraudulent claims (Simmachan et al. 2023). Precision is needed to reduce false warnings and ensure designated circumstances are fraud. To capture the most real fraud cases, a high recall rate is essential. The F-measure evaluates the model's positive class performance by combining precision and recall. Hence, the F-measure was top-prioritized in this work.

Table 3 Evaluation metrics

Metric	Formula	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	This metric indicates the overall correctly prediction.
Precision	$\frac{TP}{TP+FP}$	This metric indicates the ratio of true positive predictions to the total number of positive predictions made by the model.
Recall	$\frac{TP}{TP+FN}$	This metric indicates the correct rate of predicting the fraud class (positive class).
F-measure	$2\times\frac{(\text{Precision}\times\text{Recall})}{(\text{Precision}+\text{Recall})}$	This metric is the harmonic mean of precision and recall.

2.5.4 Statistical tests

Evaluation metrics are used to evaluate model performance; however, they may need further analysis and comparison (Kotb and Ming 2021). Selecting the best model based on capabilities is difficult. The problem can be solved by inferential statistics (Demšar 2006). We used a statistical test to compare many related samples. Four measures will be tested to test the null hypothesis that data scenarios and ML models are the same. Comparing the means of multiple populations using the ANOVA F-test is common. Three conditions of random error must be met to use this test. First, the errors must be independent. Second, the errors must be distributed normally. Finally, error variances must be constant (Simmachan 2019). The Kolmogorov-Smirnov (KS) test (Smirnov 1948) was used for normality assumption. The Friedman test (Friedman 1937, Friedman 1940) will be used if one of the requirements is not met. This test ranks data for each scenario in every model and examines rank values. Likewise, each model was also evaluated in each scenario. The Friedman test returns a mean rank to help choose the best scenario and model.

2.6. Research framework

Before splitting the dataset after data pre-processing, we used stratified random sampling to assure positive and negative classes in the training and testing sets. Next, the dataset was split into 80% training and 20% testing sets. The training set used feature selection and class-balancing to produce five data scenarios successively. After applying a 5-fold CV to the training sets from five scenarios, the testing set was utilized to assess model performance in each iteration instead of the test sets under a basic 5-fold CV. One-step models were created in five scenarios under the 5-fold CV process. Two-step models were built sequentially using the top three F-measure scores of One-step models. Data analysis was done in Python using Google Colab. Finally, four evaluation metrics were generated, descriptive and inferential statistics were compared, and the findings were summarized in conclusion, and discussion sections. The proposed prediction framework is shown in Figure 1.

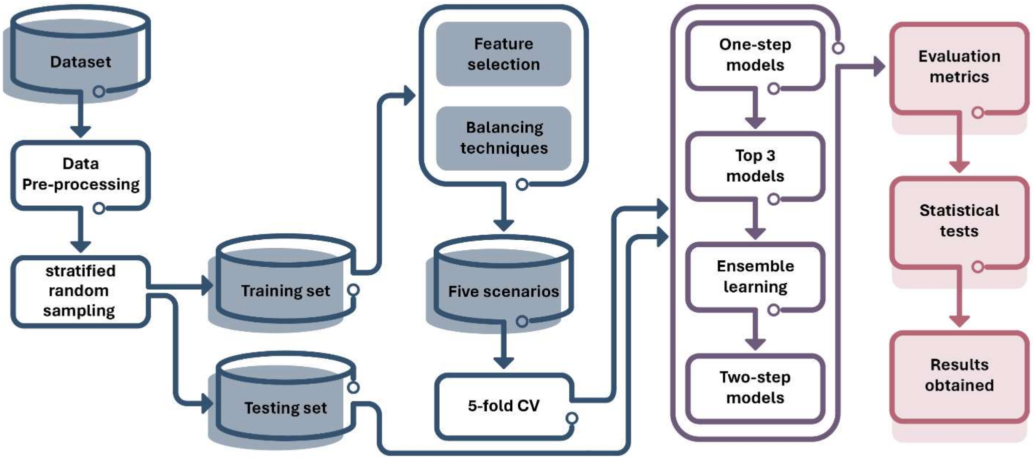

Figure 1 Research framework for predicting insurance fraud report

3. Numerical results

The investigation was conducted to optimize the predictive capability of ML models in predicting fraud reports. More specific and interesting results are illustrated as follows: First, a feature importance score based on MI was reported. Second, comparisons of model performance along with the data scenarios based on descriptive statistics were presented numerically and graphically. Next, the results of the statistical tests were displayed for greater reliability of the comparisons. Finally, the rankings of ML models and data scenarios were given.

3.1. Feature selection using mutual information

We use MI to pick the top 12 influenced features that explain the response variability. Figure 2 shows important features and their MI values in descending order. In this case, the capital-loss is considered the most significant feature.

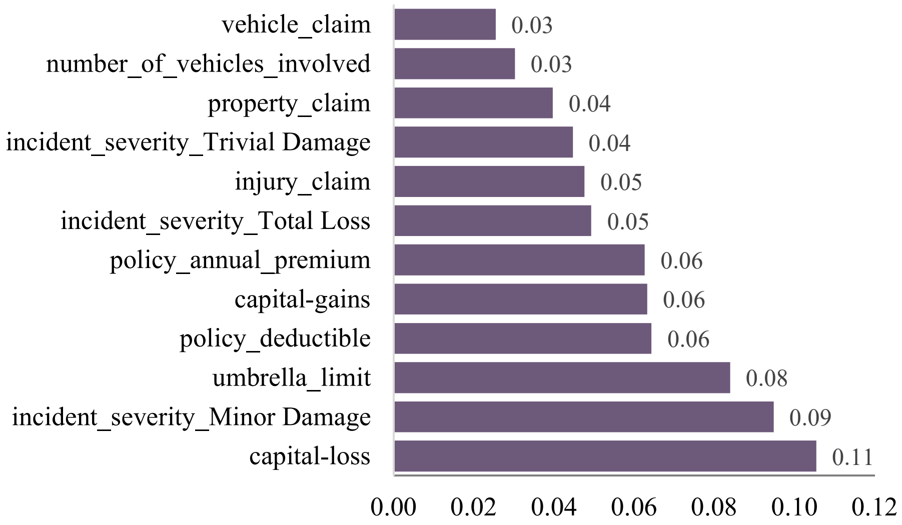

Figure 2 Top 12 feature’s mutual information value

3.2. Comprehensive performance comparisons

Tables 4-5 illustrate the overall performance of One-step and Two-step models, respectively. The combination of 13 ML models, 5 scenarios, and 4 evaluation metrics resulted in these findings. The winners of ML models in each matric are emphasized in bold. The key finding from Tables 4-5 indicates a significant discrepancy between accuracy and recall when using the imbalanced data. We can see that in the column “Original”, all the accuracy results are greater than 70%, while some recall results are less than 20%. The LR model is the reveal case; it has an accuracy of 75.20% but a recall of only 1.27%. The implication of this is that the corresponding models favor the majority class, indicating a bias. Hence, it is crucial to address this problem as it has resulted in misleading results. The class-balancing techniques were employed for this issue, we can observe that in the columns “Under”, “Under+MI”, “SMOTE”, and “SMOTE+MI”. Obviously, the data demonstrates a notable enhancement in the ML models when utilized alongside class-balancing techniques. This is reflected in their strong predictive capabilities towards the minority class. Following the implementation of class-balancing techniques, the recall rate of the LR reached a minimum of 67.30%, and it further improved to 86.43% with the use of SMOTE+MI. The K-NN model is another outstanding advancement. When using imbalanced data, the K-NN had a recall of 5.93%, but when SMOTE was used, the recall raised to 91.22%. Since the F-measure is a combination of precision and recall, it exhibits similar characteristics to recall. As a result, ML models have shown significant improvement in handling imbalanced data. Like recall and F-measure, precision possesses similar qualities, but its advancement is not as remarkable as both metrics.

Among One-step models, the GB with SMOTE model outperforms the others in terms of accuracy, recall, and F-measure, while the NB with SMOTE model is the best in terms of precision. Besides, every top-performing model reaches predictive measurements of at least 88%. Then, the one-step models were compared and ranked according to their F-measure scores in each scenario. The three best-performing models were determined. In addition, the gray-highlighted F-measure scores indicate the top three models. The models performed differently under different scenarios. For original data, the best-performing models are XGB (57.86%), GB (57.56%), and DT (56.33%), respectively. The three best models for the "Under" scenario are GB (80.06%), XGB (76.00%), and RF (75.62%). The LR (72.36%), GB (67.73%), and NB (67.52%) are the most effective models for the "Under+MI" scenario. For the "SMOTE" scenario, the top models are GB (88.83%), XGB (87.31%), and ADA (86.46%), in that order. The best three models for the "SMOTE+MI" scenario are LR (86.00%), ADA (85.71%), and NB (85.50%), respectively. We observed that overall, the SMOTE techniques outperform the others. The under-sampling methods are better than the original data. In addition, the class-balancing methods without variable selection outperform those with variable selection. Nevertheless, there are certain scenarios where variable selection is applicable to the parametric model, LR. As seen in both "Under+MI" and "SMOTE+MI" scenarios, the LR model was the top performer.

Out of the Two-step models, the MV and IMLP with SMOTE achieve the highest accuracy. Similarly, the IMLP with SMOTE is the winer in terms of recall and F-measure. Furthermore, the IXGB with SMOTE demonstrates superior precision compared to other models. Additionally, all the winers yield predictive measurements of at least 88%. To facilitate analysis, bar charts of F-measure scores were generated and displayed in Figures 3-6. Figure 3 compares the LR model to its improved version. Figure 4 compares the GB model to its enhanced variant. A comparison of the XGB model and its enhanced version is shown in Figure 5. A comparison of all five proposed models is shown in Figure 6.

Table 4 Overall performance of one-step models (%)

Models	Metric	Original	Under	Under+MI	SMOTE	SMOTE+MI
K-NN	Accuracy	72.90	47.36	70.23	62.62	85.66
	Precision	28.54	47.27	71.92	58.06	86.25
	Recall	5.93	48.11	66.02	91.22	84.84
	F-measure	9.70	47.62	68.59	70.90	85.49
NB	Accuracy	73.30	60.50	61.11	85.39	85.52
	Precision	38.21	59.98	59.60	**94.38**	85.34
	Recall	10.10	63.72	81.06	75.34	85.76
	F-measure	15.53	61.72	67.52	83.78	85.50
DT	Accuracy	78.60	68.42	60.51	83.33	76.96
	Precision	57.01	69.42	61.25	81.86	76.34
	Recall	55.90	65.86	59.06	85.77	78.05
	F-measure	56.33	67.42	60.00	83.72	77.18

Note: The top 3 models in each data scenario are highlighted by ░. The best ML models for each metric are highlighted in **bold**.

Table 4 (Continued)

Models	Metric	Original	Under	Under+MI	SMOTE	SMOTE+MI
RF	Accuracy	77.50	75.91	67.19	86.06	85.13
	Precision	62.73	76.68	67.98	90.81	86.17
	Recall	21.11	74.85	65.40	80.20	83.64
	F-measure	30.01	75.62	66.47	85.13	84.85
ADA	Accuracy	80.50	73.88	68.01	86.52	85.79
	Precision	64.03	74.65	69.98	86.68	86.10
	Recall	49.24	72.27	63.00	86.37	85.44
	F-measure	55.41	73.20	65.97	86.46	85.71
GB	Accuracy	80.30	79.75	69.62	**88.58**	85.32
	Precision	61.41	78.40	71.31	86.64	84.99
	Recall	54.20	82.01	64.86	**91.25**	85.82
	F-measure	57.56	80.06	67.73	**88.83**	85.38
XGB	Accuracy	80.60	76.31	64.36	87.32	83.60
	Precision	62.15	76.61	65.39	87.07	84.09
	Recall	54.37	75.82	61.75	87.70	82.85
	F-measure	57.86	76.00	63.25	87.31	83.42
LR	Accuracy	75.20	74.70	74.69	85.13	85.99
	Precision	35.00	75.47	78.66	86.54	85.65
	Recall	1.27	73.45	67.30	83.10	86.43
	F-measure	2.42	74.24	72.36	84.76	86.00

Note: The top 3 models in each data scenario are highlighted by ▒. The best ML models for each metric are highlighted in **bold**.

Table 5 Overall performance of two-step models (%)

Models	Metric	Original	Under	Under+MI	SMOTE	SMOTE+MI
MV	Accuracy	77.80	78.53	70.84	**88.12**	85.86
	Precision	68.85	80.04	73.59	88.98	86.39
	Recall	17.33	75.93	64.84	87.02	85.09
	F-measure	27.09	77.81	68.59	87.92	85.69
ILR	Accuracy	80.60	79.35	71.24	87.92	86.12
	Precision	63.06	77.77	73.59	88.33	86.67
	Recall	52.12	82.37	65.71	87.79	85.41
	F-measure	56.92	79.86	69.29	87.83	85.99
IGB	Accuracy	80.80	77.93	71.04	87.98	86.19
	Precision	66.84	77.44	73.86	88.35	86.97
	Recall	47.01	79.03	65.32	87.93	85.13
	F-measure	54.16	77.89	69.04	87.91	86.01
IXGB	Accuracy	81.30	78.54	78.54	87.98	86.19
	Precision	66.20	77.39	77.39	**89.41**	86.97
	Recall	52.56	80.68	80.68	86.58	85.13
	F-measure	57.73	78.78	78.78	87.71	86.01

Table 5 (Continued)

Models	Metric	Original	Under	Under+MI	SMOTE	SMOTE+MI
IMLP	Accuracy	80.50	79.75	79.75	**88.12**	86.32
	Precision	62.43	77.93	77.93	88.30	86.85
	Recall	53.01	83.20	83.20	**88.33**	85.70
	F-measure	57.23	80.37	80.37	**88.08**	86.20

Note: The best ML models for each metric are highlighted in **bold**.

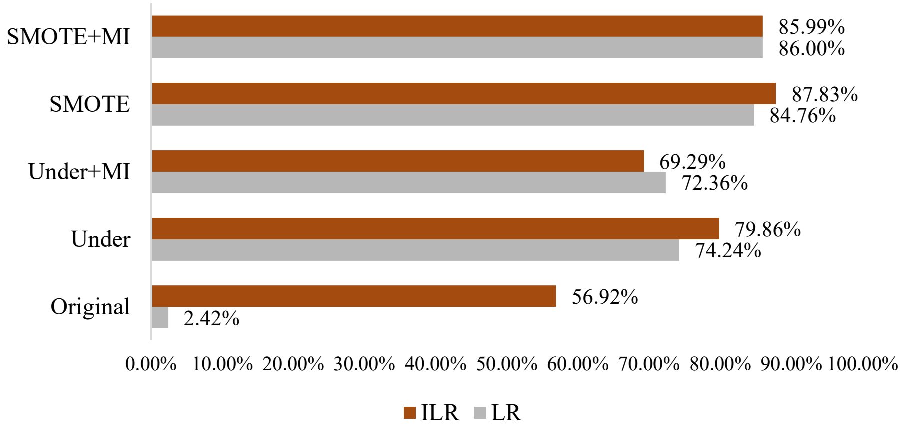

Figure 3 F-measure analysis comparing LR and ILR

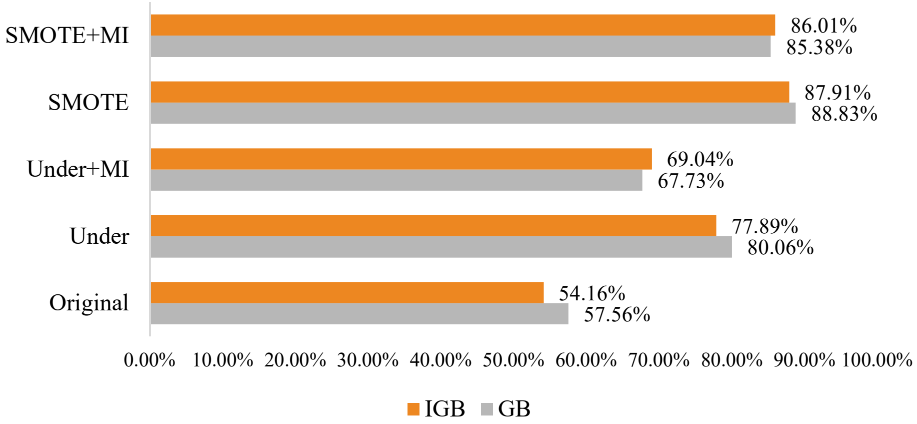

Figure 4 F-measure analysis comparing GB and IGB

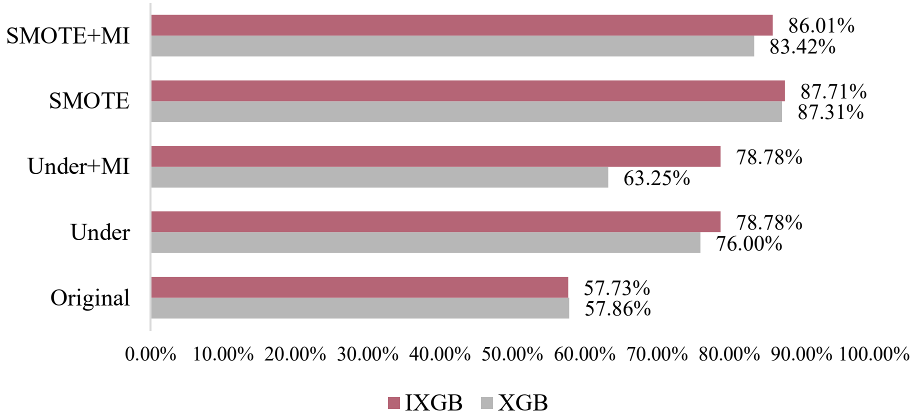

Figure 5 F-measure analysis comparing XGB and IXGB

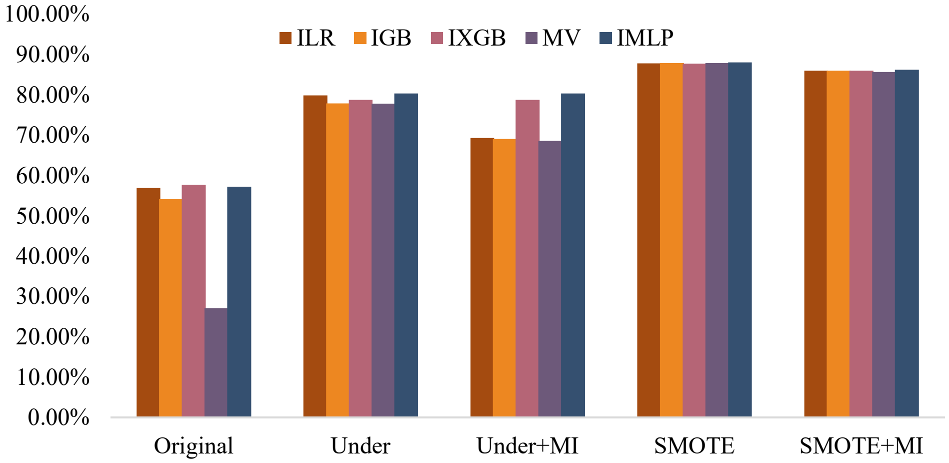

Figure 6 Two-step model performance in terms of F-measure

The ILR model shows a substantial enhancement when compared to the original model in dealing with imbalanced data. For under-sampling schemes, the LR and ILR show similar results. Nevertheless, the upgraded version exhibits slightly superior performance in under-sampling scenario without feature selection. In the context of SMOTE methods, likewise, both models are comparable. In addition, SMOTE approaches outperform under-sampling schemes.

The GB and IGB models provide insignificant results for all scenarios. It can be inferred that the original version of GB is one of permanent ML models. The two SMOTE approaches outperform the others. Without feature selection, under-sampling is more effective than with feature selection, and the imbalanced situation results in the lowest result.

When compared to the original version, the IXGB model presents a considerable improvement in under-sampling and feature selection. In alternative scenarios, IXGB and XGB models show comparable results. However, the SMOTE techniques consistently outperform the other candidates, with the imbalanced data resulting in the lowest outcome.

In terms of F-measure, the overall performance of the proposed models is generally comparable in all scenarios, except for the MV model with imbalanced data which has the lowest result in this situation. Like other models, the proposed models show superior performance when utilizing SMOTE techniques, and they produce unsatisfactory results with imbalanced data. Furthermore, under-sampling yields more consistent results when feature selection is not used.

The performance of ML models varies in different scenarios, making it a challenging task to determine the best model or scenario using descriptive statistics. The ANOVA test is expected to overcome this difficulty. However, this test is valid when all assumptions are satisfied. The normality assumption was initially examined. Statistical tests including normality tests and Friedman tests for model and scenario comparisons, are summarized in Table 6.

Table 6 Results of statistical tests

Metric	Test of Normality		Model Comparison		Scenario Comparison	
	KS Statistic	p-value	Chi-squared	p-value	Chi-squared	p-value
Accuracy	0.127	< 0.000001	39.791	0.000078	42.667	< 0.000001
Precision	0.118	< 0.000001	32.792	0.001043	47.690	< 0.000001
Recall	0.165	< 0.000001	19.217	0.083421	46.698	< 0.000001
F-measure	0.175	< 0.000001	34.957	0.000476	48.682	< 0.000001

According to the KS tests on accuracy, precision, recall, and F-measure in Table 6, the p-values are smaller than 0.05; the normality assumption for all metrics is violated; and the ANOVA test cannot be utilized. Consequently, the non-parametric alternative test, Friedman test, was implemented for model and scenario comparisons. For model comparison, the Friedman test p-values for accuracy, precision, and F-measure are less than 0.05, indicating that at least one ML model performs differently. While the p-value for recall may seem contradictory, we prioritize the previously drawn conclusions which give precedence to F-measure as the main evaluation criteria. All metrics have p-values below 0.05 for scenario comparison, showing that at least one scenario performs significantly.

As seen in Table 7, the mean of ranks and the corresponding ranks for models and scenarios in terms of F-measure are reported. Our findings show that the proposed models are satisfactory. Moreover, all the Two-step models are among the top six ranked. Furthermore, the IMLP model is the most efficient among all other models. Interestingly, the GB model holds the fourth position in the rankings. The SMOTE technique stands out among other scenarios, as reflected in the descriptive statistics. The SMOTE technique with feature selection ranks second, followed by under-sampling techniques without and with feature selection and imbalanced case.

4. Discussion

This study attempts to improve the efficiency of ML models in fraud detection tasks by introducing a novel prediction framework called the Two-step models. Predictive models created with imbalanced data favor the majority class, making them unreliable. Some investigations (Hanafy and Ming 2021, Kotb and Ming 2021) match our findings. Addressing imbalanced concerns improves ML models, notably parametric LR model. Because positive and negative classes have equivalent observation numbers, balanced data supports the probability 0.5 cut-off. SMOTE scenarios hold the top two positions. This is because SMOTE creates synthetic minority class instances to boost dataset representation without duplicating samples (Chawla et al. 2002). Avoiding minority sample duplication by developing synthetic samples decreases overfitting (He and Garcia 2009). It enhances ML models in terms of recall and F1-score on imbalanced datasets (Fernández et al. 2018). Despite employing different feature selection methods than earlier studies, the deductible, incident severity, umbrella insurance coverage limit, and policy type are common relevant features. ML models perform differently in various contexts, making it challenging to identify the best one. The Friedman test selected the optimal model and scenario. All Two-step models rank in the top six, according to our data. IMLP outperforms all models. Four stacking models outperform bagging in two-step models. Effective boosting models include IXGB, GB, and IGB. Similar results were found by Njoh-Paul (2020). Surprisingly, GB retains the fourth spot, and it outperforms One-step models. Furthermore, GB is slightly better than its improved version, IGB. Since GB is inherently an ensemble model, adding extra features to the model may overfit it.

Table 7 Ranking results obtained from Friedman tests in terms of F-measure

Model type	Model	Mean Rank	Rank	Scenario	Mean Rank	Rank
Two-step	**IMLP**	**12.20**	**1**	**SMOTE**	**4.77**	**1**
	IXGB	10.70	2	SMOTE+MI	4.23	2
	ILR	9.60	3	Under	2.77	3
	IGB	9.10	5	Under+MI	2.23	4
	MV	7.50	6	Original	1.00	5
One-step	GB	9.20	4			
	XGB	6.20	7			
	LR	6.20	8			
	ADA	5.60	9			
	RF	4.60	10			
	NB	3.80	11			
	KNN	3.30	12			
	DT	3.00	13			

5. Implementation

ML approaches can detect auto insurance fraud effectively. However, their effectiveness depends on the quality of training data and model architectures. Our prediction framework develops predictive models utilizing a combination of effective tools. We begin with data preprocessing, including cleaning, scaling, and handling missing values and outliers. Stratified random sampling ensures balanced representation of positive and negative classes in training and testing sets, while SMOTE is recommended to addresses imbalanced data in the training set. Feature selection, though inadequate in our current work, is crucial. Extracting informative features from the dataset should be further investigated. One-step models were initially used for model development. Generally, insurance companies emphasize bogus claim alerts. F-measure is preferred among other evaluation metrics due to its generalization of precision and recall, and measures model fraud class performance. The most essential metric of our study should be applied in practice. Based on F-measure scores, we select the top three One-step models to create Two-step models using ensemble methods: voting and stacking. Stacking outperforms voting in our proposed models. We recommend integrating IMLP with SMOTE as the most effective model for fraudulent claim prediction. This work may help insurance companies improve fraud detection systems, but challenges remain. Integrating the model into claims processing, monitoring performance, and retraining are essential. While ML can enhance fraud detection, it should be combined with human expertise and rule-based systems for optimal decision-making in fraud prevention.

6. Conclusions

The introduced prediction framework offers a substantial improvement in developing ML models for fraud detection tasks. The choice of model evaluation metrics is crucial. Within this particular situation, the F-measure is given higher priority. We select the top three productive models based on their F-measure scores. This step leads to the performance of the two-step models. Using effective components to construct ensemble models results in very efficient predictive models. Nevertheless, there exist prospective avenues for additional enhancement in the future. It is advisable to take into account alternative versions of SMOTE, such as the ones proposed by Hanafy and Ming (2021) and Kotb and Ming (2021). The number of components in the Two-step models should be adjusted. Alternative ML models warrant further investigation. Other feature selection techniques such as Sudjai et al. (2023) should be applied especially in case in which high-dimensional data with multicollinearity. Additional ensemble techniques, such as Mathew (2022), Vosseler (2022), Abakarim et al. (2023), and Srisuradetchai et al. (2023), should also be utilized.

Acknowledgements

We acknowledge the referees for their insightful manuscript feedback. This study was supported by Thammasat University Research Fund, Contract No TUFT 3/2567.

References

Abakarim Y, Lahby M, Attioui A. A Bagged ensemble convolutional neural networks approach to recognize insurance claim frauds. Appl Syst Innov. 2023; 6(1): 1-20.

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

Bangchang KN. Application of Bayesian variable selection in logistic regression model. AIMS Math. 2024; 9(5): 13336-13345.

Bangchang KN, Wongsai S, Simmachan T. Application of data mining techniques in automobile insurance fraud detection. ICoMS 2023: Proceedings of the 2023 6th International Conference on Mathematics and Statistics; 2023 July 14-16; Germany. Leipzig: ACM; 2023. pp. 48-55.

Belhadji EB, Dionne G, Tarkhani F. A model for the detection of insurance fraud. Geneva Pap RiskInsur Issues Pract. 2000; 25(4): 517-538.

Bhowmik R. Detecting auto insurance fraud by data mining techniques. J Emerg Trends Comput Inf Sci. 2011; 2(4): 156-162.

Boonkrong P, Simmachan, T. A multigroup SEIR epidemic model with vaccination on heterogeneous network. Chiang Mai J Sci. 2016; 43(4): 896-902.

Botchey FE, Qin Z, Hughes-Lartey K. Mobile money fraud prediction—a cross-case analysis on the efficiency of support vector machines, gradient boosted decision trees, and Naïve Bayes algorithms. Information. 2020; 11(8): 383.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002; 16: 321-357.

Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006; 7: 1-30.

Dhieb N, Ghazzai H, Besbes H, Massoud Y. Extreme gradient boosting machine learning algorithm for safe auto insurance operations. ICVES 2019: Proceeding IEEE of the 2019 IEEE International Conference on Vehicular Electronics and Safety; 2019 September 4-6; Cairo. Egypt: IEEE; 2019. pp. 1-5.

Farghaly HM, Shams MY, El-Hafeez TA. Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt. Knowl Inf Syst. 2023; 65(6): 2595-2617.

Fernández A, Garcia S, Herrera F, Chawla NV. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. J Artif Intell Res. 2018; 61: 863-905.

Friedman JH. Stochastic gradient boosting. Comput Stat Data An. 2002; 38(4): 367-378.

Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc. 1937; 32(200): 675-701.

Friedman M. A comparison of alternative tests of significance for the problem of m rankings. Ann Math Stat. 1940; 11(1): 86-92.

Hanafy M, Ming R. Using machine learning models to compare various resampling methods in predicting insurance fraud. J Theor Appl Inf Technol. 2021; 99(12): 2819-2833.

Harrell FE. Binary logistic regression. In: Frank E, Harrell Jr. regression modeling strategies springer series in statistics. Switzerland: Springer, Cham; 2015.

He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009; 21(9):1263-1284.

Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. ICML'15: Proceedings of the 32nd International conference on machine learning; 2015 July 7-9; Lille. France: JMLR; 2015. pp. 448-456.

Kotb MH, Ming R. Comparing SMOTE family techniques in predicting insurance premium defaulting using machine learning models. Int J Adv Comput Sci Appl. 2021; 12(9): 621-629.

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

Mathew TE. Appositeness of Hoeffding tree models for breast cancer classification. J Curr Sci Technol. 2022; 12(3): 391-407.

Matos T, Macedo JA, Lettich F, Monteiro JM, Renso C, Perego R, Nardini FM. Leveraging feature selection to detect potential tax fraudsters. Expert Syst Appl. 2020; 145, https://doi.org/10.1016/j.eswa.2019.113128.

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

Njoh-Paul IM. A comparative study of ensemble techniques and individual classifiers in predicting insurance claim. MSc [Thesis], Ireland: National College of Ireland; 2020.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. AdaBoost and Voting Classifier. Scikit-learn: Machine learning in Python. JMLR [monograph online] 2011 [cite 2023 May 20]; 12: 2825-2830. Available from: https://scikit-learn.org/stable/modules/ensemble.html.

Phaphan W. Fraud-detection-in-insurance-claims. GitHub. 2024 [cited 2024 Dec 8]. Available from: https://github.com/wikanda-phaphan/Fraud-detection-in-insurance-claims.

Prasasti IMN, Dhini A, Laoh E. Automobile Insurance fraud detection using supervised classifiers. International Workshop on Big Data and Information Security (IWBIS), 2020 October 17-18; Indonesia. Depok: IEEE; 2020. pp. 47-51.

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

Saheed YK, Hambali MA, Arowolo MO, Olasupo YA. Application of GA feature selection on Naive

Bayes, random forest and SVM for credit card fraud detection. The 2020 international conference on decision aid sciences and application (DASA), 2020 November 8-9; Bahrain. Sakheer: IEEE; 2020. pp. 1091-1097.

Simmachan T. Impact of homogeneity of variances violation in single factor components of variance model when sampling from finite population. Sci Eng Health Stud. 2019; 13(1): 29-37.

Simmachan T, Manopa W, Neamhom P, Poothong A, Phaphan W. Detecting fraudulent claims in automobile insurance policies by data mining techniques. Thail Stat. 2023; 21(3): 552-568.

Smirnov N. Table for estimating the goodness of fit of empirical distributions. Ann Math Stat. 1948; 19(2): 279-281.

Srisuradetchai P, Panichkitkosolkul W, Phaphan W. Combining machine learning models with ARIMA for COVID-19 epidemic in Thailand. RI2C: Proceeding of the 2023 Research, Invention, and Innovation Congress: Innovative Electricals and Electronics; 2023 Aug 24-25; Thailand. Bangkok: IEEE; 2023. pp. 155-161.

Subudhi S, Panigrahi S. Use of optimized fuzzy C-means clustering and supervised classifiers for automobile insurance fraud detection. J King Saud Univ Comput Inf Sci. 2020; 32(5): 568-575.

Sudjai N, Duangsaphon M, Chandhanayingyong C. Relaxed adaptive Lasso for classification on high-dimensional sparse data with multicollinearity. Int J Stat Med Res. 2023; 12: 97-108.

Vanishkorn B, Supanich W. Crash severity classification prediction and factors affecting analysis of highway accidents. ICAICTA: Proceeding of the 9th International Conference on Advanced Informatics: Concepts, Theory and Applications. 2022 Sep 28-29; Japan. Tokoname: IEEE; 2022. pp. 1-6.

Vosseler A. Unsupervised insurance fraud prediction based on anomaly detector ensembles. Risks. 2022; 10(7), https://doi.org/10.3390/risks10070132.

Wang J, Neskovic P, Cooper LN. Improving nearest neighbor rule with a simple adaptive distance measure. Pattern Recogn Lett. 2007; 28(2): 207-213.

Yang JB, Shen KQ, Ong CJ, Li XP. Feature selection for mlp neural network: the use of random permutation of probabilistic outputs. IEEE Trans Neural Netw. 2009; 20(12): 1911-1922.