Contents lists available at ScienceDirect

Decision Analytics Journal

journal homepage: www.elsevier.com/locate/dajour

A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions

Jonathan Kwaku Afriyie [a,*], Kassim Tawiah [a,b], Wilhemina Adoma Pels [a], Sandra Addai-Henne [a], Harriet Achiaa Dwamena [a,c], Emmanuel Odame Owiredu [a], Samuel Amening Ayeh [a], John Eshun [a]

[a] *Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana*

[b] *Department of Mathematics and Statistics, University of Energy and Natural Resources, Sunyani, Ghana*

[c] *Biometrics Unit, Council for Scientific and Industrial Research-Crops Research Institute, Kumasi, Ghana*

ARTICLE INFO

Dataset link: https://www.kaggle.com/datasets/kartik2112/fraud-detection

Keywords:
Machine learning
Decision tree
Random forest
Logistic regression
Credit card
Fraud detection and prediction

ABSTRACT

Fraudsters are now more active in their attacks on credit card transactions than ever before. With the advancement in data science and machine learning, various algorithms have been developed to determine whether a transaction is fraudulent. We study the performance of three different machine learning models: logistic regression, random forest, and decision trees to classify, predict, and detect fraudulent credit card transactions. We compare these models' performance and show that random forest produces a maximum accuracy of 96% (with an area under the curve value of 98.9%) in predicting and detecting fraudulent credit card transactions. Thus, we recommend random forest as the most appropriate machine learning algorithm for predicting and detecting fraud in credit card transactions. Credit Card holders above 60 years were found to be mostly victims of these fraudulent transactions, with a greater proportion of fraudulent transactions occurring between the hours of 22:00GMT and 4:00GMT.

1. Introduction

Banks used to provide only in-person services to customers until 1996 when the first internet banking application was introduced in the United States of America by Citibank and Wells Forgo Bank [1]. After the introduction of internet banking, the use of credit cards over the internet was adopted. This has increased rapidly during the past decade and services like e-commerce, online payment systems, working from home, online banking, and social networking have also been introduced and widely used [2]. Due to this, fraudsters have intensified their efforts to target online transactions utilizing various payment systems [3].

In recent times, improvements in digital technologies, particularly for cash transactions, have changed the way people manage money in their daily activities. Many payment systems have transitioned tremendously from physical pay points to digital platforms [4]. To sustain productivity and competitive advantage, the use of technology in digital transactions has been a game-changer and many economics have resorted to it [5]. Hence, internet banking and other online transactions have been a convenient avenue for customers to carry out their financial and other banking transactions from the comfort of their homes or offices, particularly through the use of credit cards.

According to [6], a credit card is designed as a piece of plastic with personal information incorporated and issued by financial service providers to enable customers to purchase goods and services at their convenience worldwide. The unlawful use of another person's credit card to get money or property either physically or digitally is known as credit card fraud [7]. Events involving credit card fraud occur often end in enormous financial losses [8]. It is simpler to commit fraud now than it was in the past because an online transaction environment does not require the actual card and the card's information suffices to complete a payment [9]. [10] postulate that monetary policy as well as business plans and methods used by big and small businesses alike have been imparted by the introduction of credit cards.

The Bank of Ghana (BoG) reported an estimated loss value of GH¢ 1.26 million ($250,000) in 2019 due to credit card fraud which increased to GH¢ 8.20 million ($1.46 million) in 2020, (BoG, 2021). This represented an estimated 548.0% increase in losses in year-to-year terms. All payment channels have experienced persistent increases in fraud in recent years, with digital transactions seeing the largest rise [11]. One such instance is payment fraud, which includes checks, deposits, person-to-person (P2P), wire transfers, automated clearing, house transfers, internet payments, automated bill payments, debit and credit card transactions, and Automated Teller Machine (ATM) transactions [12]. The perpetrators attack victims using Virtual Private Network (VPN) tunnel connections through Anchor- free software or physically rob victims of their valuables from unknown destinations and operate with fictitious identities, because of these their arrest is often a wild goose chase [13]. Following similar patterns,

[*] Corresponding author.
E-mail address: jonathan.afriyie@knust.edu.gh (J.K. Afriyie).

https://doi.org/10.1016/j.dajour.2023.100163
Received 27 November 2022; Received in revised form 15 December 2022; Accepted 10 January 2023
Available online 14 January 2023

compliance and risk management services employed to identify online fraud have shown a lot of interest in AI and machine learning models [12].

Some of these models include Decision Tree, Logistic Regression, Random Forest, Ada Boost, XG Boost, Support Vector Machine (SVM), and Light GBM [14]. This has become necessary because credit card fraud detection is a classification and prediction problem. Supervised machine learning models have been proved as the best models to detect fraud using the above-mentioned algorithms [15]. This study therefore seeks to compare three classification and prediction techniques, namely; Decision Tree, Logistic Regression, and Random Forest in classifying and predicting financial transactions as either fraudulent or not fraudulent.

The remaining portions of the paper are arranged as follows: The review of related literature is presented in Section 2. In addition to describing the dataset that was used and the experimental setup, Section 3 provides a brief overview of the various strategies employed in this research. The results of the analyses are presented in Section 4 whereas Section 5 contains the discussion of the results. The final section, Section 6, highlights the conclusion and recommendations based on the findings.

2. Literature review

Logistic regression is a technique that is used to predict an outcome variable that is binary. This technique does not demand that explanatory variables follow normal distribution, or correlated [16]. The outcome variable in Logistic Regression models is qualitative. Explanatory variables might take the shape of numbers or categories. Numerous scholars have used Logistic Regression to detect financial bankruptcies.

Decision tree is a non-linear classification technique that divides a sample into increasingly smaller subgroups using a collection of explanatory variables. At each branch of the tree, the process iteratively chooses the explanatory variable that, in accordance with a predetermined criterion, has the strongest correlation with the outcome variable [17]. It is nonparametric, therefore there is no requirement to choose unimodal training data and it is simple to add a variety of quantitative or qualitative data structures. However, when applied to the entire data set, decision trees have a tendency to overfit the training data, which might produce bad results. Decision trees can be used to filter spam emails and also to predict the kind of persons who will be vulnerable to a certain virus in the area of medicine.

Random forests, which [18] proposed, are an additional level of randomness for bagging. Random forests alter how the classification or regression trees are built, in addition to employing various bootstrap samples of the data for each tree's construction. In conventional trees, the optimal split among all variables is used to divide each node. Each node in a random forest is split using the best predictor among a subset that was randomly selected at that node. The average of all trees' predictions is then the output for any location. The random forest package in R was used to create bagging and random forest models [19]. Each feature's significance in relation to the training data set can be measured. However, Random forests are biased towards attributes with several levels for data including qualitative variables with differing number of levels. Random forest can be applied as follows; complex biological data analysis in the area of Bioinformatics, segmentation of video and classification of image for pixel analysis.

The categories of credit card fraud recognized by [20] are bankruptcy fraud, counterfeit fraud, application fraud, and behavioural fraud. Depending on the sort of fraud that banks or credit card companies are dealing with, several precautions can be created and implemented. For identifying fraudulent transactions in other jurisdictions, machine learning methods like Logistic Regression, Naive Bayes, Random Forest, K Nearest Neighbour, Gradient Boosting, Support Vector Machines, and neural network algorithms were used by [21]. To choose the top features for the model, they used feature importance approach and reported an accuracy of 95.9% with Gradient Boosting performed better than the other algorithms.

A machine learning-based technique for identifying credit card fraud was developed by [22] for the application of hybrid models with Ada Boost and majority voting strategies. They added noise of around 10% and 30% to their hybrid models to facilitate the approach. A good score of 0.942 was awarded to multiple voting approaches based on data from the sample for 30% more noise. As a result, they settled on the voting system as the most effective technique in the presence of noise. [23] proposed two different types of random forests which were used to teach the behavioural characteristics of typical and abnormal transactions. The study looks at how well these two random forests, in terms of their classifiers, perform in identifying credit card fraud. Data from a Chinese e-commerce company was utilized to analyse the performance of these two different random forest models. According to other study findings, even if the suggested random forests perform well on small datasets, other problems, such as imbalanced data, prevent them from being as effective as other datasets [3].

[24] researched on practical methods for detecting credit card fraud which affects financial institutions. Different machine learning algorithms were employed and were able to determine the best algorithm that predicted fraudulent transactions. Two resampling methods (under sampling and over sampling) were used to train the algorithms. Out of the many algorithms trained, the best models for predicting credit card fraud were found to be Random Forest, Xgboost, and Decision Tree, with AUC values of 1.00%, 0.99%, and 0.99%, respectively.

Machine learning algorithms can help detect fraudulent transactions, classify them, and probably stop the transaction process if required [25]. Credit card fraud detection prognosis consists of modelling past credit card transactions, which have records of transactions that are fraudulent, after which the model will be used on new transactions to detect if it is a genuine or fraudulent transaction [26]; [27].

3. Data and methods

3.1. Data

The data set comprised of simulated transactions of credit cards between January 1, 2020 and December 31, 2020, including both legitimate and fraudulent transactions in the western side of the United States of America available at https://www.kaggle.com/datasets/kartik2112/fraud-detection. Haris's (2020) sparkov data generation was implemented for the simulation. It includes transactions made to a pool of 800 businesses using the credit cards of 1000 customers. The dataset contains every purchase, the customer's name, the merchant, and the type of purchase, as well as information regarding whether or not the transaction was fraudulent. It contains 555 719 rows of observations, which has 23 columns of variables. 12 of these variables are qualitative data.

In the pre-processing stage, the data was cleaned and formatted to eliminate missing values since our analysis is based on complete data. By performing feature scaling, we kept all numeric explanatory variables within the same domain by using range transformation to compute all numeric variables to be in a range of 0 and 1. We also used under sampling on the imbalanced data to prevent biasing of the algorithms towards the majority class [28]. Values less than 5 and greater than 1250 were removed. Because the dataset in this study is significantly skewed, [29] used Synthetic Minority Oversampling Technique (SMOTE) to balance the data, however, we employed under sampling to handle the imbalance in the dataset. Here, in the minority class, this approach decreases the majority cases to equal or slightly equal to the minority class. Fig. 1 shows the undersampled data. Tables 1 and 2 shows the summary statistics of the types of variables used in the study.

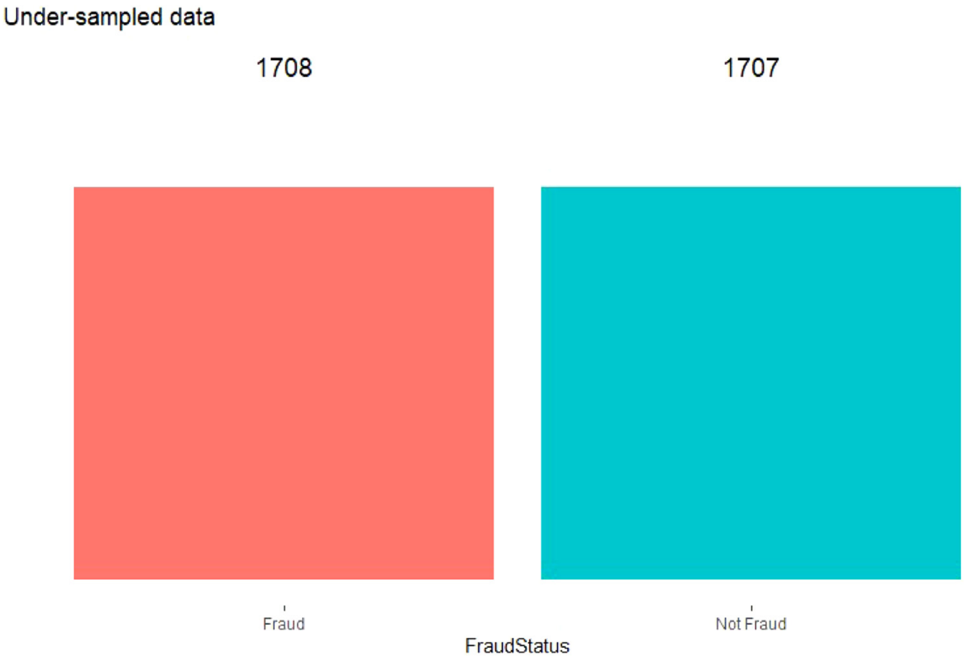


Fig. 1. Under sampled data.

Table 1
Basic Statistics for the character variables.

Name	Count	Unique	Top	Frequency
Transaction date and time	555 719	544 760	2020-12-19 16:02:22	4
Merchant	555 719	693	fraud_Kilback LLC	1859
Category	555 719	14	gas_transport	56 370
First	555 719	341	Christopher	11 443
Last	555 719	471	Smith	12 146
Gender	555 719	2	F	304 886
Street	555 719	924	444 Robert Mews	1474
City	555 719	849	Birmingham	2423
State	555 719	50	TX	40 393
Job	555 719	478	Film/video editor	4119
Date of birth	555 719	910	1977-03-23	2408
Transaction number	555 719	555 719	2da90c7d74bd46a	1

3.2. Methods

In this section, we discuss the supervised machine learning models such as logistic regression, Random Forest, and Decision Tree to classify fraudulent transactions.

3.2.1. Decision tree

Decision trees are non-parametric supervised learning techniques that can be employed for classification [30]. They generate decision rules with a tree-like structure using actual data attributes. Decision Trees evolved from the way humans make decisions [31]. Graphically, they show information in a tree pattern that is easy to understand. The decision tree structure is made up of nodes, edges, and leaf nodes. According to [24], it consists of a set of branches/nodes that are connected by edges. Fig. 2 shows the flow diagram of a decision tree. The decision tree's root node chooses a feature to partition the data into two or more sub nodes to develop decision nodes after the partition into sub nodes and subtrees at the end of the root node [32]. Each sub-tree of the data will once more be partitioned into two sub nodes. Until every training sample is gathered, this process will be repeated. So, at the end of the decision tree, we end up with a leaf node which serves as a representation of the class which aims at classifying.

The decision tree algorithm has the benefit of not needing feature scaling, being robust to outliers, and handling missing values automatically. It is quicker to train and is very good at resolving classification and prediction problems. The decision tree uses the following; the Gini index, information gain, and entropy as a metric for classification into two or more nodes. Fig. 2 shows the decision three algorithm modelling approach.

Entropy is a measure of expected randomness in the features used for splitting the root node by adapting between 0 and 1 [33]. From Fig. 3, values closer to 0 imply a sample that is entirely homogeneous, and near to 1 implies a sample that is evenly divided. The formula for calculating entropy is

$$E\left(X\right) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \tag{1}$$

$$E\left(X\right) = -p(Fraud)\log_2 p(Fraud) - p(Not\ Fraud)\log_2 p(Not\ Fraud) \tag{2}$$

where $p(x_i)$ is the probability of a class 'i' in the feature variable X.

Gini Index is a measure or a metric for splitting root nodes that measures how often a randomly chosen element would be incorrectly identified [34]. As shown in Fig. 3, the Gini index value ranges between the values of 0 and 0.5. This implies that an attribute with a lower Gini index is automatically selected for the splitting. The formula for calculating the Gini Index is

$$E\left(X\right) = 1 - \sum_{i=1}^{n} p(x_i)^2 \tag{3}$$

The information gain (IG) is a statistical characteristic that gauges how effectively a particular variable separates the data into its intended categories [35]. The IG is calculated as the expected reduction of entropy in the form of information gained. The main goal of decision tree construction is to identify the attribute that yields the largest information gain and the lowest entropy. The expressions used for computing IG are

$$G(X,Y) = E(X) - E(X|Y) \tag{4}$$

$$G\left(X,Y\right) = -p\left(Fraud\right)\log_2 p\left(Fraud\right) - p\left(Not\ Fraud\right)\log_2 p\left(Not\ Fraud\right) - \sum \frac{|Sv|}{S} entropy\left(Sv\right) \tag{5}$$

To compute the reduction in uncertainty about Y, given an extra piece of information X about Y, we simply subtract the entropy of X from the entropy of Y. We refer to this as information gain. The amount of information learned about Y from X increases with the reduction in this uncertainty.

Table 2
Basic Statistics for the numeric variables.

Name	Count	Mean	Std	Min	25%	50%	75%	Max
Unique identifier	555 719	277 859	160 422.4	0	138 929.5	277 859	416 788.5	555 718
Credit card number of customers	555 719	4 178 387	1 309 837	6 041 621	1 800 429	3 521 417	4 635 331	4 992 346
Amount	555 719	69.39	156.75	1	9.63	47.29	83.01	22 768.11
Zip	555 719	48 842.63	26 855.28	1257	26 292	48 174	72 011	99 921
Latitude	555 719	38.54	5.061	20.03	34.67	39.37	41.89	65.69
Longitude	555 719	−90.23	13.72	−165.67	−96.8	−87.48	−80.18	−67.95
City population	555 719	88 221.89	300 390.9	23	741	2408	19 685	2 906 700
Time (s)	555 719	1 380 679	5 201 104	1 371 817	1 376 029	1 380 762	1 385 867	1 388 534
Merchant latitude	555 719	38.54	5.1	19.03	34.76	39.38	41.95	66.68
Merchant longitude	555 719	−90.23	13.73	−166.67	−96.91	−87.45	−80.27	−66.95
Fraud status	555 719	0.0039	0.062	0	0	0	0	1

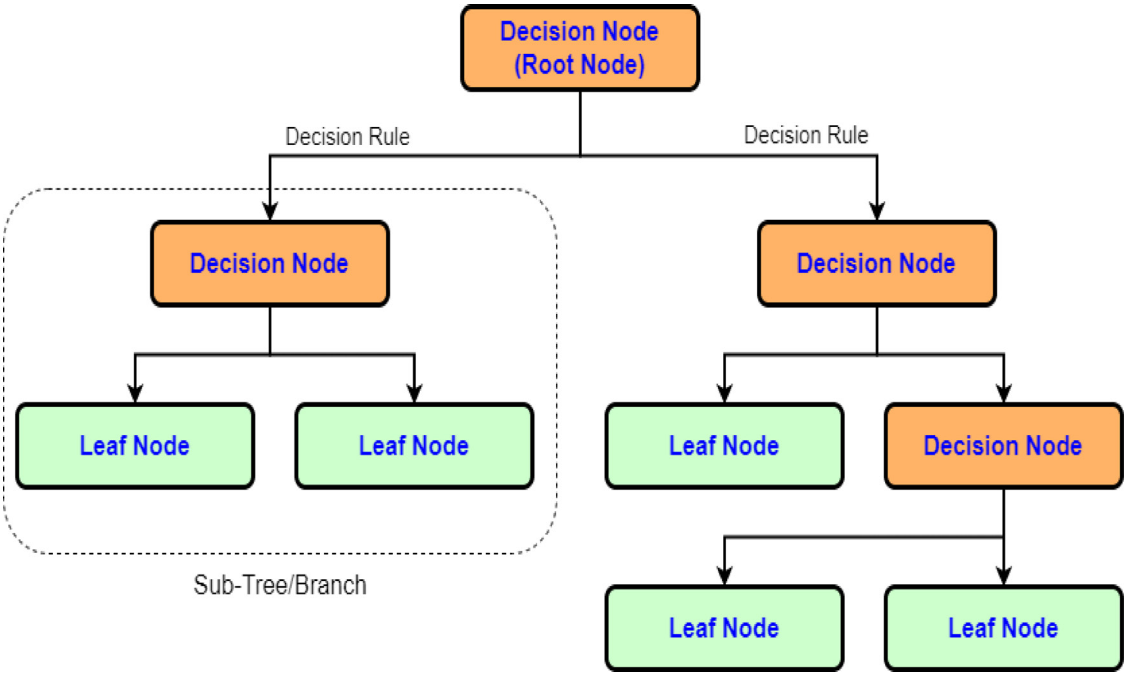


Fig. 2. Decision tree.

3.2.2. Logistic classification

[36] described logistic regression as a type of regression analysis whereby the dependent variable is binary or dichotomous, such as fraud or not fraud. The formula for Logistics Regression is;

$$\ln\left[\frac{p\,(y=1)}{1-p\,(y=1)}\right]=\alpha_0+\alpha_1x_1+\alpha_2x_2+\cdots+\alpha_nx_n \tag{6}$$

where;

α_0 is the intercept of the model

α_i are the model coefficients $,i=1,2,3,\ldots,n$

x_i are the independent variables, $i=1,2,3,\ldots,n$

y is the dichotomous dependent variable

$$p\,(y)=\begin{cases}1, & fraud\\ 0, & no\ fraud\end{cases}$$

The likelihood that an observation belongs to a particular class is usually what we are concerned about when we have a binary classification challenge is given by

$$odds=\left[\frac{p\,(y=1)}{1-p\,(y=1)}\right] \tag{7}$$

The machine learning logistic regression algorithm uses the sigmoid function to describe the relationship that exists between the response variable and the predictor variable (insert reference). It is employed in this work to determine whether or not a transaction is fraudulent. It is calculated as;

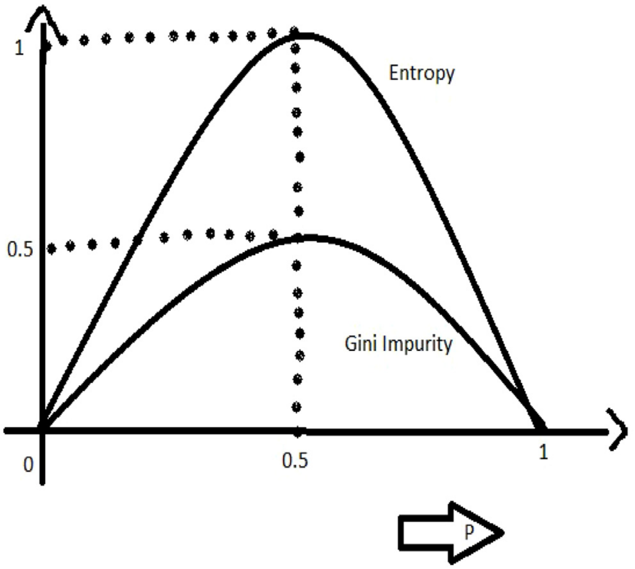
$$p\,(y)=\frac{exp(\alpha+\beta x)}{1+exp(\alpha+\beta x)} \tag{8}$$

Because of the nonlinear nature of the relationship between p(y) and x, the parameters α and β are not as easily interpreted as they are in linear regression. The logistic curve is displayed in Fig. 4. It can be interpreted in terms of probabilities because it is limited to values between 0 and 1.

3.2.3. Random forest

Random Forest is a supervised machine learning algorithm that uses a group of decision tree models for classification and making predictions [37]. Each decision tree is a weak learner because they have a low predictive power. It is based on ensemble learning, which uses many decision tree classifiers to classify a problem and improve the accuracy of the model [38]. As a result, the random forest employs a bagging method to generate a forest of decision trees. Given a dataset (X,Y) with N total observation where X being the predictor variables and Y the outcome variable, the random forest algorithm first creates K_i random variables $(i=1,2,\ldots,N)$ to form a vector and then converts each K_i random vector into a decision tree to obtain the dK_i decision tree $(dK_1(X),dK_2(X),\ldots,dK_N(X))$. The final classification results are as follows:

$$D\,(X)=\arg\max\left\{\sum_i^N dK_i\,(X)=Fraud,\sum_i^N dK_i\,(X)=Not\ fraud,\right\} \tag{9}$$



$$E(S)=\sum_{i=1}^{c}-p_i\log_2 p_i$$

$$Gini(E)=1-\sum_{j=1}^{c}p_j^2$$

Fig. 3. Entropy and Gini Index graph.

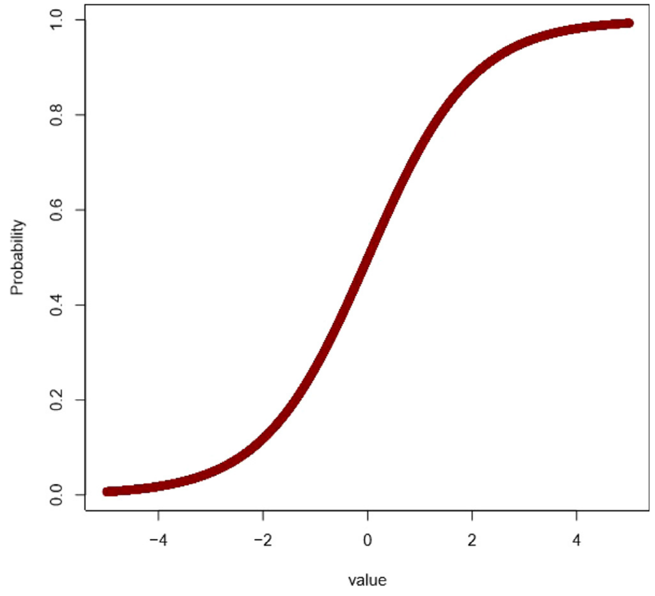


Fig. 4. Logistic curve.

Random forest typically does not require a feature selection procedure [39]. The drawback of this approach is how quickly it may identify data with a wide range of values and variables with numerous values as fraudulent. It is one of the financial sector's most accurate fraud detection algorithms, according to [40]. It is usually more uncertain when the Random Forest method begins to build the tree, so it is crucial to choose the most important feature out of all features for analysis, particularly in node splitting. Fig. 5 illustrates a random forest algorithm technique.

Table 3
Confusion matrix.

	Actual class	
Predicted class	Fraud (1)	Not Fraud (0)
Fraud (1)	True Positive (TP)	False Positive (FP)
Not Fraud (0)	False Negative (FN)	True Negative (TN)

We assessed the model's performance using metrics like accuracy, precision, recall, specificity, and F1-score in order to compare different algorithms. Accuracy is most frequently used to gauge a model's performance [41]. Our dataset is quite unbalanced, thus comparing the model's using accuracy as the only performance metric may not be appropriate in this context. Instead, we must select the best model to identify fraudulent transactions by using other measurements such as area under the curve (AUC) [42] in addition to the accuracy.

The entries in the confusion matrix (Table 3) are defined as the following: False positive (FP) is the total number of incorrect predictions classified as positive; False negative (FN) is the total number of incorrect predictions classified as negative; True positive (TP) is the total number of true predictions classified as positive; and True negative (TN) is the total number of true predictions classified as negative.

Accuracy, as a measurement metric, measures the ratio of the total number of correct predictions of fraud to the total number of predictions (both fraud and not fraud) made by the model [43]. It is calculated as

$$Accuracy=\frac{TN+TP}{TN+TP+FN+FP} \tag{10}$$

Precision metric measures the ratio of correctly classified fraud transactions (TP) to the total transactions predicted to be fraud transactions (TP + FP) [44]. It is calculated as

$$Precision=\frac{TP}{FP+TP} \tag{11}$$

Recall/ Sensitivity, as a metric, measures the ratio of correctly classified fraud transactions (TP) to the total number of fraud transactions [45]. It is calculated as

$$Recall/Sensitivity=\frac{TP}{TP+FN} \tag{12}$$

Specificity measures the ratio of correctly classified not fraud transactions (TP) to the total number of Not Fraud transactions [46]. It is calculated as;

$$Specificity=\frac{TN}{TN+FP} \tag{13}$$

The F1 score metric measures the weighted mean of precision and recall [47]. It ranges between zero and one with a value close to one giving the highest value. It computed using the expression

$$F1\,Score=\frac{2\times precision\times recall}{precision+recall} \tag{14}$$

As illustrated in Fig. 6, the AUC for each threshold value between 0 and 1 is calculated using this metric, which plots the FP rate on the x-axis and the TP rate on the y-axis. A positive real class outcome's likelihood to be predicted as a positive class by the model is shown by the AUC and receiver operating characteristic curve (ROC). The model performs better if it is closer to the top left corner and the higher it moves there; conversely, the model performs worse when it is closer to the curve at 45-degree diagonal of the ROC space. A random classifier is anticipated to provide points that are diagonal by default (FPR = TPR).

4. Results

Table 4 shows the transaction status of the data. We observe that there are 0.4% of fraudulent transactions while the remaining 99.6% were true transactions.

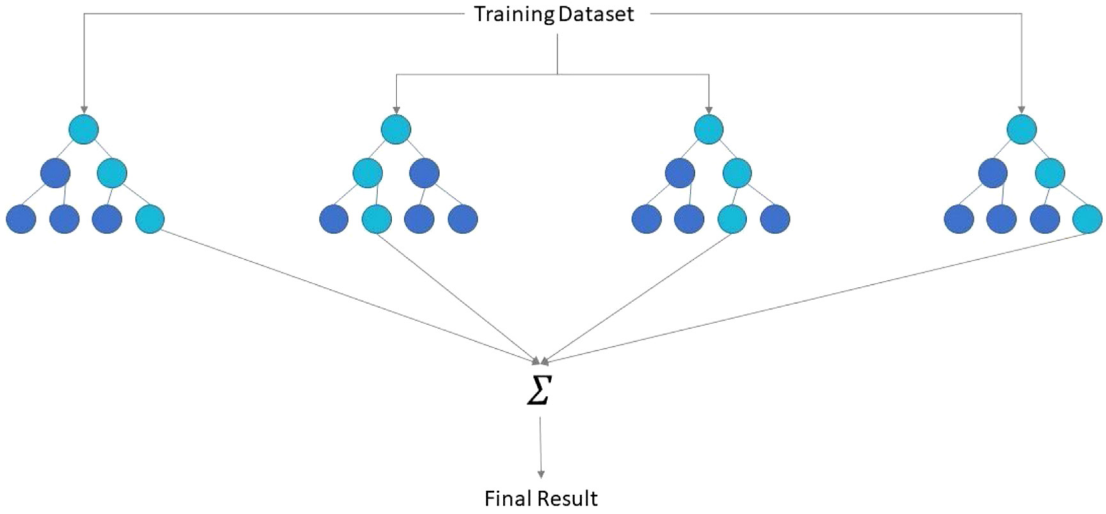


Fig. 5. Random forest.

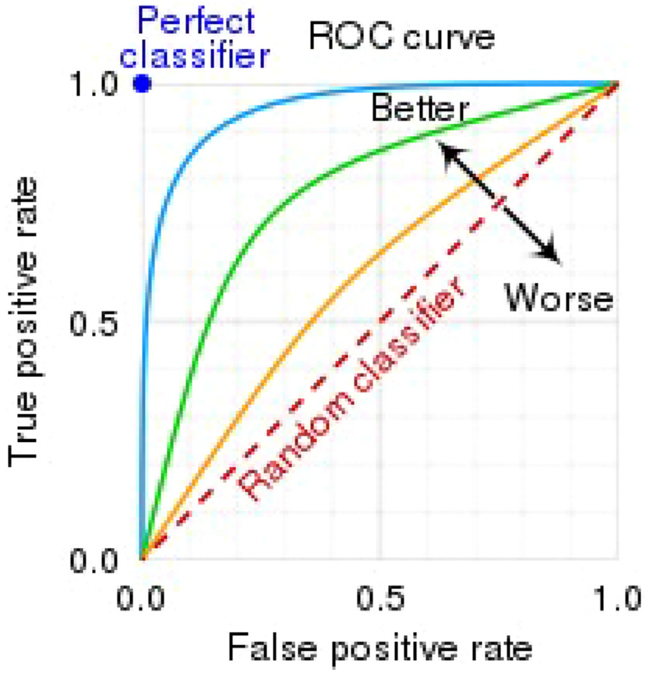


Fig. 6. ROC curve.

Table 4
Transaction description.

Description	Fraud	Non-Fraud
Total	2135	482 672
Percentage (%)	0.4%	99.6

Fig. 7 illustrates the correlation plot of the quantitative variables in our study. The merchant longitude and longitude variables were highly correlated with a value of 1. Moreover, the merchant latitude and latitude variables are highly correlated with a value of 0.99. As a result of the high correlation between the variables, they can affect our model [48]. Hence, we manually removed one before fitting the model.

Fig. 8 illustrates the fraud status and associated transaction amount. Fraudulent transactions had a very high median amount compared to the non-fraudulent transactions. The distribution of the fraudulent amounts is seen to be skewed to the right.

Table 5
Confusion matrix of prediction using decision tree.

	Reference	
Prediction	Fraud	Not Fraud
Fraud	397	8085
Not Fraud	30	88 449

From Fig. 9, 54.9% of the total transactions was made by females, whereas 45.1% were made by male transactions. Thus, females undertook more credit card transactions than males.

As illustrated in Fig. 10, most of the fraudulent transactions occurred in the shopping category (1.19%), followed by grocery (0.73%), miscellaneous (0.36%), transport (028%), and home care (0.16%). It is not surprising that home care transactions recorded fraud since not many transactions occur there.

Most fraudulent credit card transactions affected customers in the cities of Jay and Chatham (Fig. 11). From Fig. 11, the cities of Sprague and Jay had the greatest percentage of fraudulent transactions, with a percentage of 7.56 and 7.37, respectively. The chart for the 15 cities with transactions above 100 and their percentages of fraudulent transactions is shown in Fig. 11.

Fraudulent transactions tend to be higher in the year age group of 31–60 and above 60 years (Fig. 12). Fig. 12 confirms that credit card fraudsters target elderly persons who use credit cards for business transactions.

Fig. 13 shows that the majority of the fraudulent transactions occurred on Sundays, with 372 fraudulent transactions as the largest number of fraud transactions among the days of the week.

According to Fig. 14, fraudulent transactions tend to happen between 22:00 GMT and 4:00 GMT, where the majority of victims are asleep while their credit card information is used to make transactions. Additionally, banks are not operating at that time to monitor credit card transactions, which makes it simpler for fraudsters to take advantage of the available chance. When compared to the daytime, the number of fraudulent transactions is lower.

We outlined the results from the decision tree approach in Fig. 15 and the corresponding confusion matrix associated in Table 5.

Table 5 summarizes the results of the predictions of a confusion matrix when using the Decision Tree model. The model was able to correctly classify 397 fraudulent transactions out of the 427 total fraudulent transactions from the testing data as fraudulent, whereas 30 fraudulent transactions were labelled as not fraudulent. Once more, 8085 Not Fraud transactions were incorrectly classified as Fraud, whereas

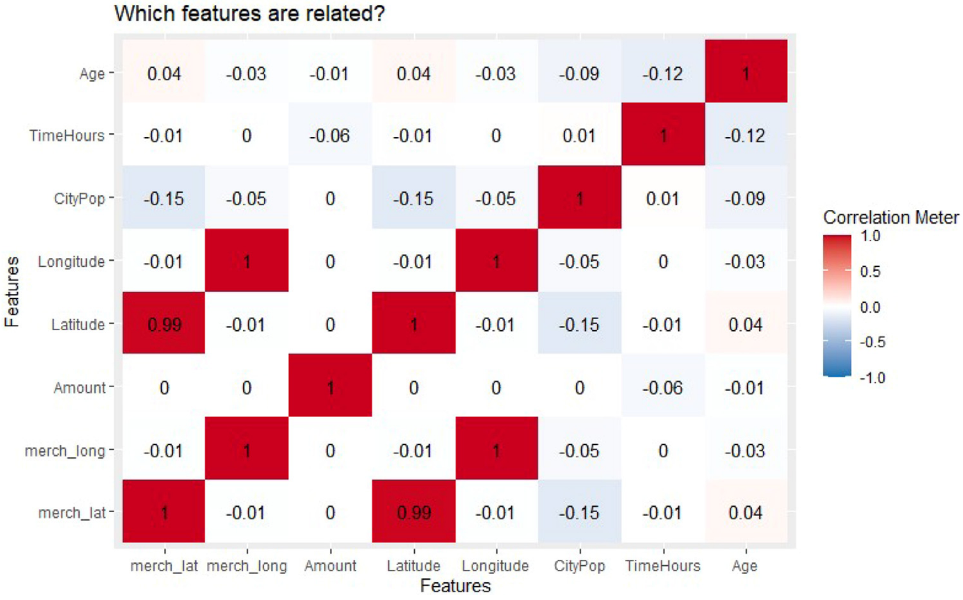


Fig. 7. Correlation plot of quantitative variables.

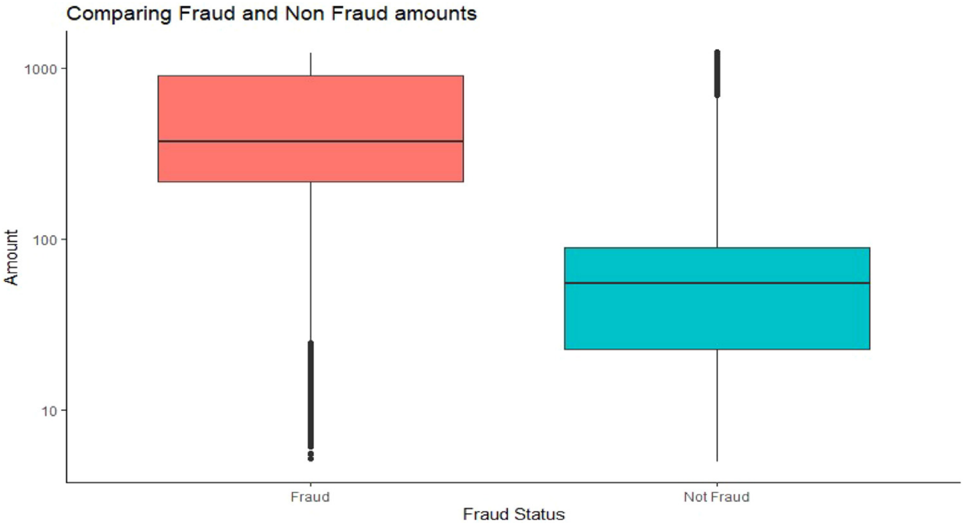


Fig. 8. Number of transactions by fraud status.

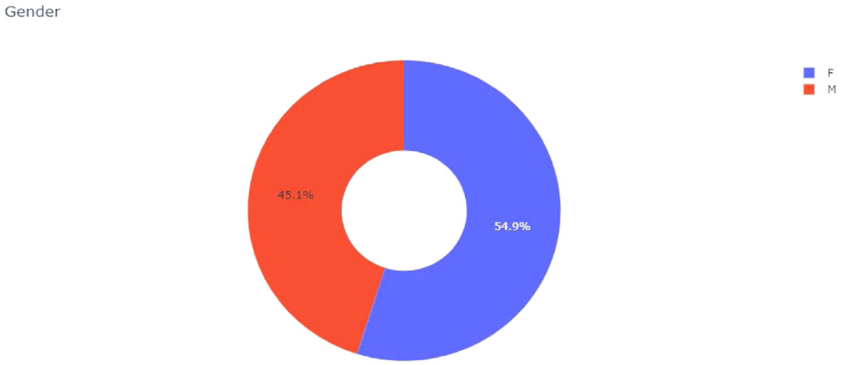


Fig. 9. Fraud status by gender.

88 449 Not Fraud transactions were correctly classified as Not Fraud. Table 6 shows the performance matrix following the Decision Tree model:

Table 7 summarizes the results of the predictions using Random Forest. The model was able to correctly classify 409 fraudulent transactions as fraudulent out of the 427 total transactions from the testing data,

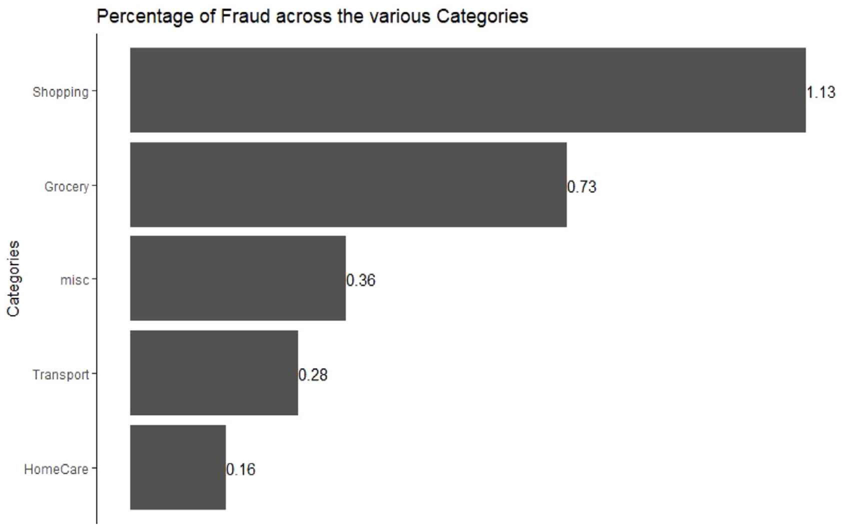


Fig. 10. Fraudulent transaction across merchant categories.

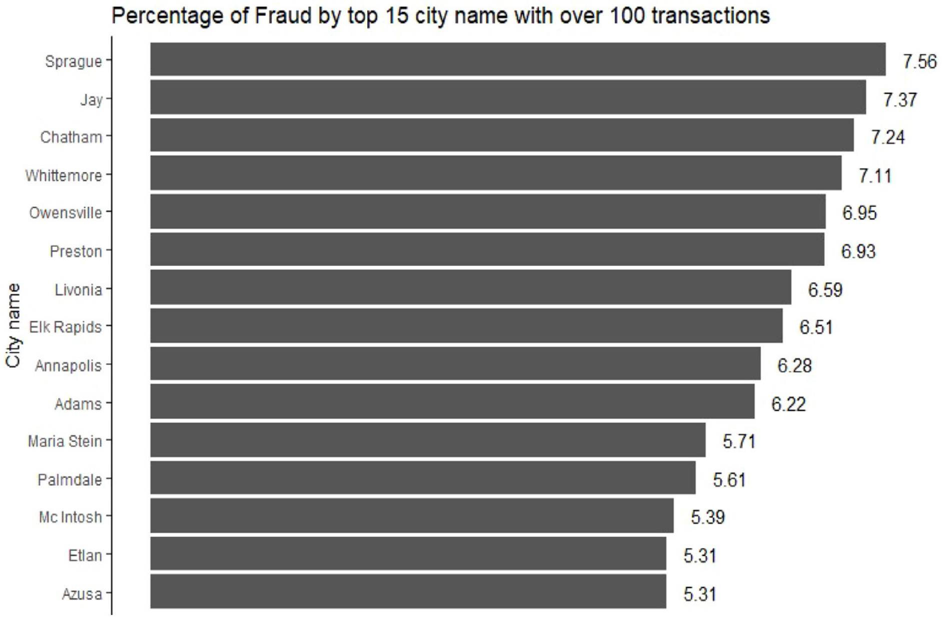


Fig. 11. Percentage of fraudulent transactions among cities with over 100 credit card transactions.

Table 6
Performance of the decision tree algorithm.

Metric measure	Estimate
Accuracy	0.92
Sensitivity	0.93
Specificity	0.92

while 18 fraudulent transactions were classified as not fraudulent. Once more, 4052 not fraud transactions were incorrectly classified as fraud, whereas 92 482 not fraud transactions were appropriately classified as not fraud. Table 8 shows the performance matrix of the Random Forest.

Table 9 summarizes the output of the predictions in a confusion matrix. Out of the 427 total transactions from the Testing Data, the model was able to correctly classify 325 fraud transactions as fraud while 102 fraud transactions were classified as Not Fraud.

Again, 88 803 not fraud transactions were classified correctly as Not Fraud, and 7731 Not Fraud transactions were wrongly classified as Fraud. The model performance matrix is presented in Table 10.

Table 7
Confusion matrix of prediction using random forest.

	Reference	
Prediction	Fraud	Not Fraud
Fraud	409	4052
Not Fraud	18	92 482

Table 8
Performance of the random forest algorithm.

Metric measure	Estimate
Accuracy	0.96
Sensitivity	0.97
Specificity	0.96

The models are compared based on their performance. Table 11 shows the measurement results from these measurement matrices for Accuracy, F1-Score, Recall/Sensitivity, Precision, and Specificity. Among the three models, the Random Forest model as well recorded

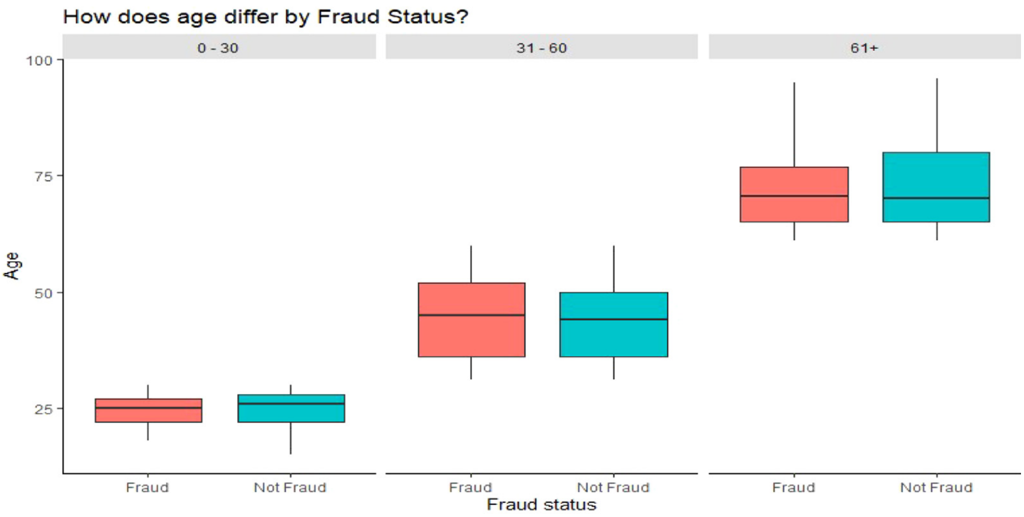


Fig. 12. Distribution of age group and fraud status.

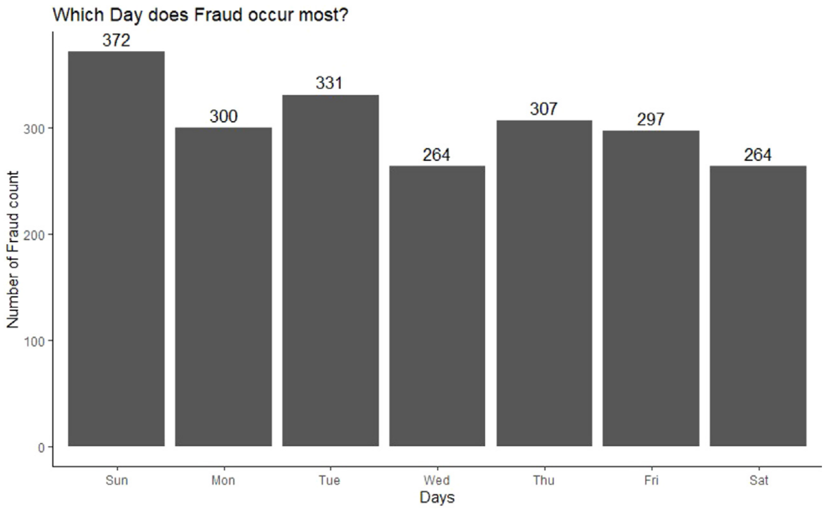


Fig. 13. Fraudulent transactions weekdays.

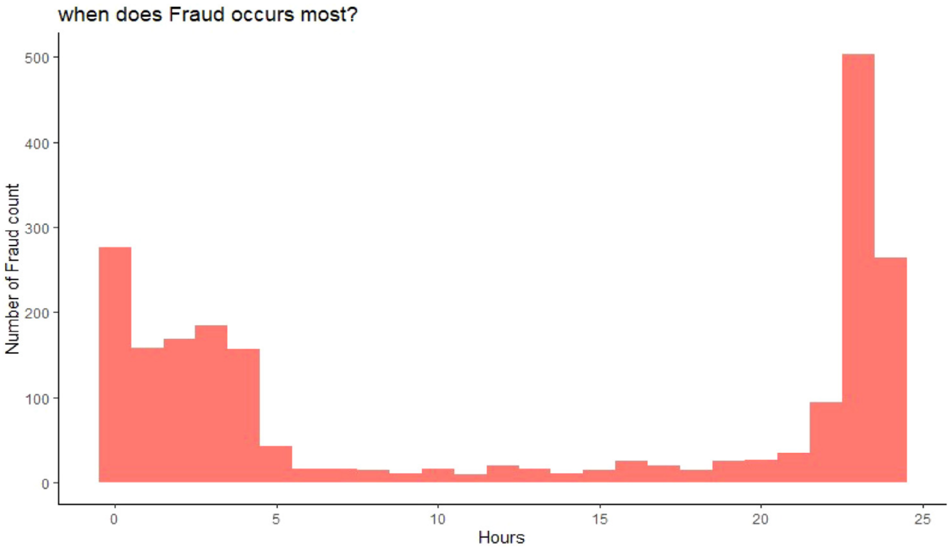


Fig. 14. Fraudulent transactions by time in hours.

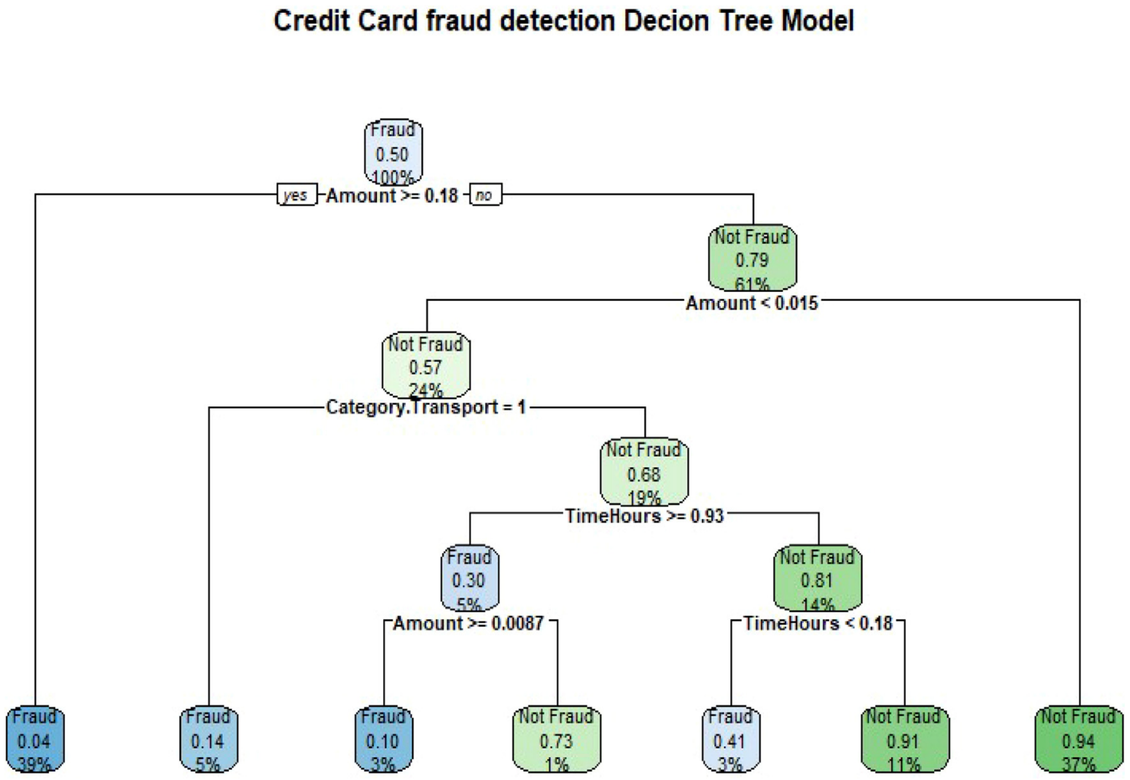


Fig. 15. Decision Tree Model.

Table 9
Confusion matrix of prediction using logistics regression.

	Reference	
Prediction	Fraud	Not Fraud
Fraud	325	7731
Not Fraud	102	88 803

Table 10
Performance of the logistic regression algorithm.

Metric measure	Estimate
Accuracy	0.92
Sensitivity	0.76
Specificity	0.92

Table 11
Comparing the models' performances.

Model name	Accuracy	F1-Score	Recall	Precision	Specificity
Decision tree	0.92	0.09	0.93	0.05	0.92
Random forest	0.96	0.17	0.97	0.09	0.96
Logistics regression	0.92	0.08	0.76	0.04	0.92

the highest values of 96%, 17%, 93%, 9%, and 96%, respectively, in all performance measures mentioned.

Fig. 16 shows the ROC and AUC for all models. The Random Forest model is having the highest AUC with a value of 98.9%, followed by the Decision Tree model with a 94.5% AUC value. The Logistic Regression model shows the AUC value of 87.2%. This information depicts that the Random Forest model is more useful in predicting fraud transactions because the True Positive rate and the False Positive rate were close to 1 (100%).

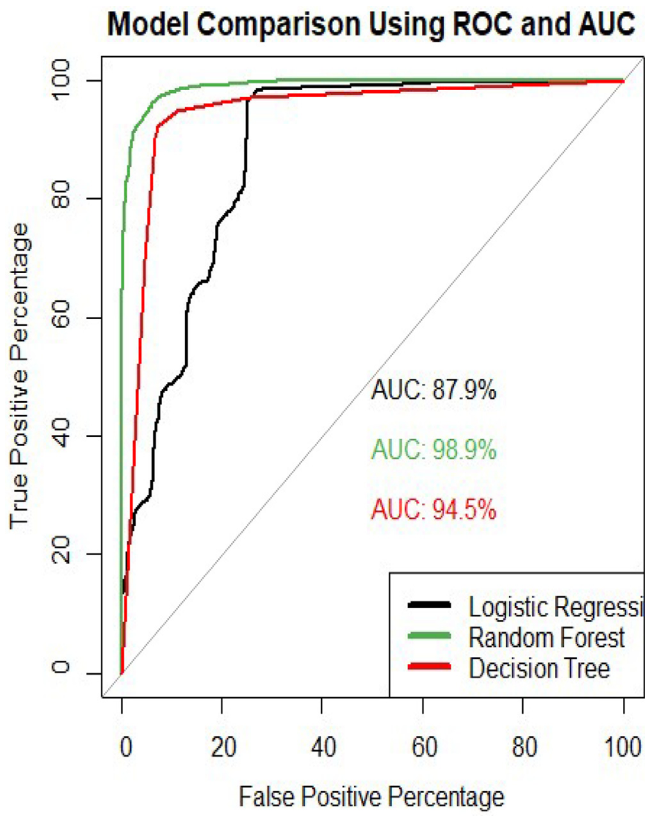


Fig. 16. AUC and ROC for comparison.

5. Discussion

Unlike [24,25,43] where the credit card transaction data set contained between 0.1% and 0.2% fraudulent transactions, our data set contained 0.4% fraudulent transactions. This indicates the richness of our data for effective analyses, detection, and prediction of fraudulent transactions.

The high correlation between merchant longitude and latitude shows that only one of them is a necessary factor in detecting and predicting fraudulent transactions, contradicting the results of [7,44], and [47]. The distribution of fraudulent amount in our study was found to be positively skewed refuting the findings of [7,49], and [50].

We observed that most women are liable to credit card fraud as a result of their frequent usage and complete reliance on transactions using credit cards, confirming the results of [51]. Our results further revealed that most fraudulent credit card transactions occurred in shops, an indication that fraudsters use dubious means to make purchases using credit cards details of others. This was also revealed by [51–55].

Just as in [53,54], and [55] cities like Sprague, Jay, Chatham, and Whittemore witnessed the highest fraudulent transactions. These cities are well known for these acts. We urge authorities to make it a priority to crack down on these fraudsters in their hideouts. Again, the elderly people targeted by these fraudsters as revealed by this study confirms the work of [51] just as the timing of the fraudulent act and the weekdays witnessing most of these uncomfortable and distressing situations.

The performance metrics used for the algorithms in this study are similar to those utilized by [7,49], and [47]. Also, the pattern of the accuracy of our algorithms, F1-score, recall/sensitivity, precision, specificity, AUC, and ROC are similar to those of [7,47], and [49]. These studies, just as in our current study, had the random forest algorithm emerging as the most suitable algorithm for fraudulent credit card transaction detection and prediction.

6. Conclusion

In order to categorize online credit card transactions as either fraud or not, this study built three different classification models, Logistics Regression, Decision Tree, and Random Forest using supervised machine learning. To ensure that the model does not favour solely the majority class and prevent overfitting the model to the data, we balanced the dataset prior to generating the models using the undersampling technique. With an AUC value of 98.9% and an accuracy value of 96.0%, the Random Forest model performed better than the other two models, making it the most suitable model for predicting fraudulent transactions [7,47,49],

Based on the data and analysis, it was determined that the majority of fraud cases occur between the hours of 20 (10 pm) and 5 (5 am). It can be concluded that banks will not be operating to monitor transactions at this time, and victims might be sleeping as well and the possibility of fraudsters to commit fraud is created by this.

The analysis revealed that cardholders over the age of 60 are most frequently the targets of fraudulent transactions. Adults over 60 seem to be more likely to report losses from particular sorts of fraud.

Based on the data and analysis performed, we recommend that the financial institutions should prioritize providing older clients with more in-person services. They must boost their security measures or over online services between the hours of 10 pm and 5 am.

As a matter of urgency, they should develop more robust and fraud-free systems. It is imperative that financial institutions embrace random forest model in predicting and detecting daily credit card fraud. Financial institutions can also implement the strategies outlined by [53], [55,56] for preventing and controlling credit card fraudulent transactions.

Other supervised machine learning algorithms can be considered in future studies with a national or inter regional level data. The present study can also be extended or applied in the health and other sectors for classification purposes.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used is made up simulated transactions of credit cards between January 1, 2020 and December 31, 2020, in the western side of the United States of America available at https://www.kaggle.com/datasets/kartik2112/fraud-detection.

References

[1] K. Yak, D. Tudeal, Internet Banking Development as A Means of Providing Efficient Financial Services in South Sudan. 2 (2011) 139–148.

[2] S. Madan, S. Sofat, D. Bansal, Tools and Techniques for Collection and Analysis of Internet-of-Things malware : A systematic state-of-art review, J. King Saud Univ. - Comput. Inf. Sci. (2021) xxxx, http://dx.doi.org/10.1016/j.jksuci.2021.12.016.

[3] F.C. Yann-a, Streaming active learning strategies for real-life credit card fraud detection : Assessment and visualization, 2018.

[4] V. Nath, ScienceDirect credit card fraud detection using machine learning algorithms credit card fraud detection using machine learning algorithms, Procedia Comput. Sci. 165 (2020) 631–641, http://dx.doi.org/10.1016/j.procs.2020.01.057.

[5] T. Pencarelli, The digital revolution in the travel and tourism industry, Inf. Technol. Tourism (2019) 0123456789, http://dx.doi.org/10.1007/s40558-019-00160-3.

[6] S.B.E. Raj, A.A. Portia, A. Sg, Analysis on Credit Card Fraud Detection Methods. (2011) 152–156.

[7] F. Carcillo, Borgne, Y. Le, O. Caelen, Y. Kessaci, F. Oblé, Combining unsupervised and supervised learning in credit card fraud detection, Inform. Sci. 557 (2021) 317–331, http://dx.doi.org/10.1016/j.ins.2019.05.042.

[8] S. Xuan, S. Wang, Random forest for credit card fraud detection, 2018.

[9] V. Vlasselaer, Van, C. Bravo, O. Caelen, T. Eliassi-rad, L. Akoglu, M. Snoeck, B. Baesens, APATE : A novel approach for automated credit card transaction fraud detection using network-based extensions, Decis. Support Syst. 75 (2015) 38–48, http://dx.doi.org/10.1016/j.dss.2015.04.013.

[10] L.E. Faisal, T. Tayachi, S. Arabia, L.E. Faisal, O. Banking, The role of internet banking in society. 18 (13) (2021) 249–257.

[11] Dorphy, H. Hultquist, 2017 Financial Institution Payments Fraud Mitigation Survey, Federal Reserve Bank of Minneapolis, 2018.

[12] E. Kurshan, H. Shen, H. Yu, Financial crime & fraud detection using graph computing: Application considerations & outlook, in: 2020 Second International Conference on Transdisciplinary AI (TransAI), IEEE, 2020, pp. 125–130.

[13] A.R.K. Alhassan, A. Ridwan, Identity expression—the case of 'sakawa' boys in ghana, Hum. Arenas (2021) 0123456789, http://dx.doi.org/10.1007/s42087-021-00227-w.

[14] B. Lebichot, Y.A.L. Borgne, L. He-Guelton, F. Oblé, G. Bontempi, Deep-learning domain adaptation techniques for credit cards fraud detection, in: INNS Big Data and Deep Learning Conference, Springer, Cham, 2019, pp. 78–88.

[15] B. Lebichot, G.M.P.W. Siblini, L.H.F.O.G. Bontempi, Incremental learning strategies for credit cards fraud detection, Int. J. Data Sci. Anal. 12 (2) (2021) 165–174, http://dx.doi.org/10.1007/s41060-021-00258-0.

[16] B.G. Tabachnick, L.S. Fidell, Using Multivariate Statistics, Harper Collins, New York, 1996.

[17] J.A. Michael, S.L. Gordon, Data Mining Technique for Marketing, Sales and Customer Support, John Wiley & Sons INC, New York, 1997, p. 445.

[18] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[19] A. Liaw, M. Wiener, Classification and regression by randomForest, R News 2 (3) (2002) 18–22.

[20] O. Citation, B. Systems, University of Huddersfield Repository Credit card fraud and detection techniques : a review, 2009.

[21] A. Aditi, A. Dubey, A. Mathur, P. Garg, Credit Card Fraud Detection Using Advanced Machine Learning Techniques. (2022) 56–60. http://dx.doi.org/10.1109/ccict56684.2022.00022.

[22] K. Randhawa, C.H.U.K. Loo, S. Member, Credit card fraud detection using AdaBoost and majority voting, IEEE Access 6 (2018) 14277–14284, http://dx.doi.org/10.1109/ACCESS.2018.2806420.

[23] L. Guanjun, L. Zhenchuan, Z. Lutao, W. Shuo, Random forest for credit card fraud, IEEE Access (2018).

[24] K. Ayorinde, Cornerstone : A Collection of Scholarly and Creative Works for Minnesota State University, Mankato a Methodology for Detecting Credit Card Fraud a METHODOLOGY for DETECTING CREDIT CARD FRAUD Kayode Ayorinde (Thesis Master's), Data Science Minnesota State University Mankato, MN, 2021.

[25] A.D. Pozzolo, G. Boracchi, O. Caelen, C. Alippi, Credit Card Fraud Detection : A Realistic Modeling and a Novel Learning Strategy. 29(8) (2018) 3784–3797.

[26] A. Dal Pozzolo, O. Caelen, Y.A. Le Borgne, S. Waterschoot, G. Bontempi, Learned lessons in credit card fraud detection from a practitioner perspective, Expert Syst. Appl. 41 (10) (2014) 4915–4928.

[27] G. Bontempi, Reproducible machine learning for credit card fraud detection - practical machine learning for credit card fraud detection - practical handbook foreword. May, 2021.

[28] A.D. Pozzolo, O. Caelen, R.A. Johnson, G. Bontempi, Calibrating probability with undersampling for unbalanced classification, 2015.

[29] R. Tyagi, R. Ranjan, S. Priya, Credit card fraud detection using machine learning algorithms. (2021) 334–341.

[30] B.T. Jijo, A.M. Abdulazeez, Classification Based on Decision Tree Algorithm for Machine Learning. 02 (01) (2021) 20–28. http://dx.doi.org/10.38094/jastt20165.

[31] J.F.S. Iii, Evolving Fuzzy Decision Tree Structure that Adapts in Real-Time. (2005) 1737–1744.

[32] P.H. Swain, A.N.D.H. Hauska, The Decision Tree Classifier : Design and Potential. (1977) 142–147.

[33] N. Freitas, Decision Trees, University of British Columbia, 2013.

[34] S. Tangirala, Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm *. 11(2) (2020) 612–619.

[35] Y. Liu, L. Hu, F. Yan, B. Zhang, Information Gain with Weight based Decision Tree for the Employment Forecasting of Undergraduates. (2013) 2–5. http://dx.doi.org/10.1109/GreenCom-iThings-CPSCom.2013.417.

[36] M.D. Begg, An Introduction To Categorical Data Analysis, second ed., Alan Agresti, John Wiley & Sons, Inc., Hoboken, New Jersey, 2009, p. 400, http://dx.doi.org/10.1002/sim.3564, 2007. Price: $100.95. ISBN: 978-0-471-22618-5. In Statistics in Medicine (Vol. 28, Issue 11).

[37] T.R. Prajwala, A comparative study on decision tree and random forest using R tool, Int. J. Adv. Res. Comput. Commun. Eng. 4 (1) (2015) 196–199.

[38] E. Kabir, S. Guikema, B. Kane, Statistical modeling of tree failures during storms, Reliab. Eng. Syst. Saf. 177 (April) (2018) 68–79, http://dx.doi.org/10.1016/j.ress.2018.04.026.

[39] J.L. Speiser, A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data, J. Biomed. Inform. 1172020 (2021) 103763, http://dx.doi.org/10.1016/j.jbi.2021.103763.

[40] N. Donges, A complete guide to the random forest algorithm, 2021, 2019. URL: https://builtin.com/data-science/random-forest-algorithm (*дата звернення*: 25.05. 2021).

[41] J. Zhao, L. Wang, R. Cabral, F. Torre, De, Feature and Region Selection for Visual Learning. 25(3) (2016) 1084–1094.

[42] G. Goy, C. Gezer, V.C. Gungor, Makine Öğrenmesi Yöntemler i ile Kredi KartıSahtecil iği QLQ T espiti. (2019) 350–354.

[43] S. Bagga, A. Goyal, N. Gupta, A. Goyal, ScienceDirect credit card fraud detection ICITETM2020 using pipeling and ensemble learning credit card fraud detection using ensemble a pipeling and goyal c learning, Procedia Comput. Sci. 1732019 (2020) 104–112, http://dx.doi.org/10.1016/j.procs.2020.06.014.

[44] M. Chen, Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches, Comput. Math. Appl. 62 (12) (2011) 4514–4524, http://dx.doi.org/10.1016/j.camwa.2011.10.030.

[45] N. Rtayli, N. Enneya, Journal of Information Security and Applications Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization, J. Inf. Secur. Appl. 55 (1) (2020) 102596, http://dx.doi.org/10.1016/j.jisa.2020.102596.

[46] S. Mittal, Sampling approaches for imbalanced data classification problem in machine learning sampling approaches for imbalanced data classification problem in machine learning. July, 2022, http://dx.doi.org/10.1007/978-3-030-29407-6.

[47] T.C. Tran, B.T. District, H. Chi, M. City, T.K. Dang, H. Chi, M. City, L.T. Ward, T.D. District, H. Chi, M. City, Machine learning for prediction of imbalanced data : Credit fraud detection. Ml, 2021.

[48] M. Gregorich, S. Strohmaier, D. Dunkler, G. Heinze, Regression with highly correlated predictors : Variable omission is not the solution, 2021.

[49] Y. Xie, A. Li, L. Gao, Z. Liu, A heterogeneous ensemble learning model based on data distribution for credit card fraud detection, Wirel. Commun. Mob. Comput. 2021 (2021) 2531210, http://dx.doi.org/10.1155/2021/2531210, 13 pages.

[50] Y. Sahin, E. Duman, [IEEE 2011 international symposium on innovations in intelligent systems and applications (INISTA) - Istanbul, Turkey (2011.06.15-2011.06.18)] 2011 international symposium on innovations in intelligent systems and applications - detecting credit card fraud by ANN and logistic regression, 2011, pp. 315–319, http://dx.doi.org/10.1109/INISTA.2011.5946108.

[51] H. Copes, K.R. Kerley, R. Huff, J. Kane, Differentiating identity theft: An exploratory study of victims using a national victimization survey, J. Crim. Justice 38 (5) (2010) 1045–1052, http://dx.doi.org/10.1016/j.jcrimjus.2010.07.007.

[52] J. Choi, S. Han, R.D. Hicks, Exploring gender disparity in capable guardianship against identity theft: A focus on internet-based behavior, Int. J. Crim. Justice 4 (1) (2022) 25–48, http://dx.doi.org/10.36889/IJCJ.2022.002.

[53] H.Y. Prabowo, Building our defence against credit card fraud: a strategic view, J. Money Laund. Control (2011) http://dx.doi.org/10.1108/13685201111173848.

[54] Y.A. De Montjoye, L. Radaelli, V.K. Singh, A.S. Pentland, Unique in the shopping mall: On the reidentifiability of credit card metadata, Science 3476221 (2015) 536–539, http://dx.doi.org/10.1126/science.1256297.

[55] F. Hayashi, Payment card fraud rates in the United States relative to other countries after migrating to chip cards, Econ. Rev. 104 (4) (2019) 23–40.

[56] K.J. Barker, J. D'Amato, P. Sheridon, Credit card fraud: awareness and prevention, J. Final Crime 15 (4) (2008) 398–410, http://dx.doi.org/10.1108/13590790810907236.