As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 8(06): 89-94, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0806013
Available online at: http://www.ijmtst.com/vol8issue06.html

Revelation of Credit Card Fraud using Machine Learning Algorithm

Vikrant Chole| Ayan Mukherjee| Kshitij Gaikwad | Pradhnya Gawai | Pradhnya Bagde | Rati Mahule | Puja Pawar

Department of Computer Science and Engineering, G.H. Raisoni Institute of Engineering and Technology, Nagpur, India

Corresponding Author Email ID: kshitij62gaikwad@gmail.com

To Cite this Article

Vikrant Chole, Ayan Mukherjee, Kshitij Gaikwad, Pradhnya Gawai, Pradhnya Bagde, Rati Mahule and Puja Pawar. Revelation of Credit Card Fraud using Machine Learning Algorithm. International Journal for Modern Trends in Science and Technology 2022, 8(06), pp. 89-94. https://doi.org/10.46501/IJMTST0806013

Article Info

Received: 28 April 2022; Accepted: 28 May 2022; Published: 30 May 2022.

ABSTRACT

A strategy for 'Revelation of credit card fraud' is created in this study. As the number of scammers grows every day. Credit cards are used for fraudulent transactions, and there are several sorts of fraud. A combination of statistics and decision tree models are used to address this problem. This transaction is assessed individually, and the optimal solution is implemented. The primary purpose by screening the data to detect fraud strategies to achieve a better outcome.

KEYWORDS:Classification, machine learning, fraud detection

1. INTRODUCTION

Scammers use your credit card number and PIN, or a stolen credit card, to make financial transactions from your account without your awareness. When someone unlawfully uses another person's credit card information or spends money they shouldn't. It's referred to as credit card fraud. Identity theft includes credit card frauds, which have grown increasingly widespread in recent years. Scammers steal your card information and use it to conduct unethical transactions on your account without your awareness. Credit card security has become necessary because of such situations.

Because of the large class disparity, this topic is particularly difficult to learn about. The number of valid transactions is well more than the number of fraudulent transactions. In addition, transaction schemes often modify their statistical properties over time.

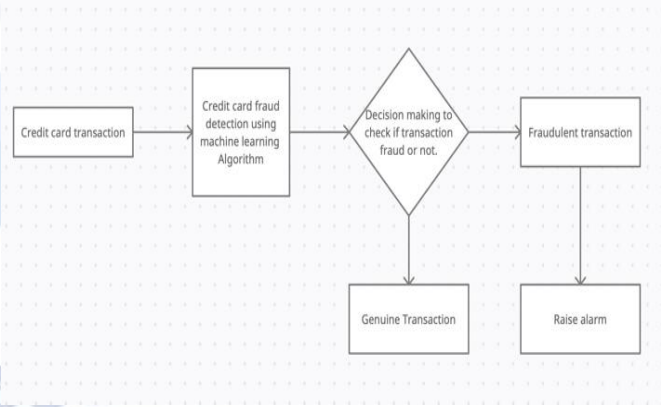


Fig 1. Block diagram of fraud detection

Method used to identify credit card fraud:

1. Logistic model
2. Decision tree model
3. Random decision forest
4. eXtreme Gradient Boosting

2. RELATED WORK

Shi yang Xuan [7]: Two types of randomised forests were used to create the characteristics of regular and abnormal transactions. The researcher examines the performance detection of credit card fraud between these two random forests that are differentiated by their classifier. The data utilised is from a Chinese company and is used to compare the performance of these two types of random forest models. The author has utilised a B2C dataset to identify and detect credit card fraud in this study. Because of this discovery, the researcher decided that while the suggested random forests offer decent results on tiny data sets, there are still issues such as data imbalance that make them less effective than any other data set.

Abhimanyu Roy [9]: To identify fraud in online money transfers, in-depth learning topologies were recommended. This method is based on a time-integrated ANN, long-term memory and short-term memory are types of memory components and a variety of other characteristics. Approximately 80 million online credit card payments have been flagged as potentially fraudulent and legitimate based on their success in identifying fraud. They deployed a distributed high-performance cloud platform. In terms of fraud detection performance, the researchers' work provides an effective reference to the sensitivity analysis of the proposed parameters. The researchers also presented a system for identifying fraud by modifying deep learning topology parameters. By eliminating fraudulent conduct, the financial institution can limit losses.

Changjun Jiang [10]: suggested a four-stage fraud detection approach that is unique. They created clusters of transactions by classifying prior transaction data into categories and then used a sliding window approach to aggregate transactions based on their similar behaviour. This method is used to characterise a cardholder's behavioural model, and then after that, it's utilised to predict their behaviour aggregating, the functionality extraction is finished using the new window. Finally, classification occurs, which divides behaviour patterns and assignments into categories. With an accuracy of 80 percent, the most accurate approaches are Logistic Regression with data (raw), Random Forest with aggregation data (AggRF), andRandom Forest with feedback technique with aggregation data (AggRF +FB).

Kuldeep Randhawa [8]: They used twelve classification techniques algorithms to identify credit card fraud. Researchers keep track of how well reference data and actual data sets perform. AdaBoost and simple majority techniques are also utilised to create hybrid models. The paper describes single and hybrid models as they are connected. They have reported the findings utilising their twelve chosen methods for the two parameters (standard and real-world datasets). As a result, when conventional algorithms were applied with AdaBoost and popular voting techniques below the benchmarks data, the Random Forest approach had the finest accuracy and sensitivity (95 and 91 percent, respectively).

Sai Kiran [11]: A new algorithm for detecting credit card fraud has been presented. The enhanced Nave Bayes K-nearest neighbour approach is what it's called (NBKNN). They use algorithms to uncover the fake transaction in the data they collected. The dataset contains information on European credit issuers who used their cards for 2 days and completed 284,807 transactions, 492 of which were fake. Despite using classification methods, both approaches fared differently on the same dataset. They combined two strategies to improve the algorithm's accuracy and adaptability (Nave Bayes and k-nearest neighbour). As a result, they could achieve accuracy of around 95 percent using Nave Bayes and 90 % using K-nearest neighbour methods.

3. METHODOLOGY

A. Dataset description

First, and importantly, we obtained our data from Kaggle, an analysis of data and dataset publishing site. This dataset has 31 columns, 28 of these are labelled v1-v28 to safeguard the data.

Out of the 284,807 transactions that occurred for two days, this dataset contains 492 scams. Positive transactions represent for 0.172 percent of all transactions, which is severely unbalanced.

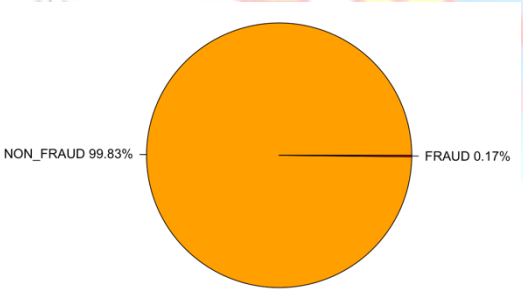
It consists solely of numerical input variables obtained from a PCA transformation. We are unable to provide the original features or other background knowledge about the data due to confidentiality concerns. The principal components generated with PCA include features V1, V2…, V28, except for 'Time' and 'Amount.' The feature 'Time' records the number of seconds that have passed

since the initial transaction in the dataset. The feature 'Amount' represents the transaction Amount, which may be used for cost-sensitive learning based on examples. When there is fraud, the response variable 'Class' is set to 1; otherwise, it is set to 0.

B. Data Pre-processing

The initial phase in machine learning is data exploration, which comprises exploring and visualizing data to uncover early insights or highlight areas or patterns that should really be investigated further. Users can better comprehend the broad picture and get to insights faster with interactive dashboards and point-and-click data exploration. We use a variety of graphs to visually comprehend the dataset and check for errors.

Unbalanced data refers to uneven occurrences of distinct classes in this pie chart. The following graphic depicts the dataset's imbalance of non-fraud transactions and fraud transactions non-fraudulent transactions make up 99.83 percent of total transactions, while likely to be fraudulent make up 0.17 percent. Our information appears to be significantly uneven in terms of the class of interest (Fraud).



This next graph depicts the instances when transactions were completed in less than two days. The least number of transactions happened at night, whereas highest occurred during the day.

Fig 3. Time Variable

This next graph depicts the volume of money transacted. Most number of transactions is small, and only a handful come close to surpassing the maximum amount that may be traded.

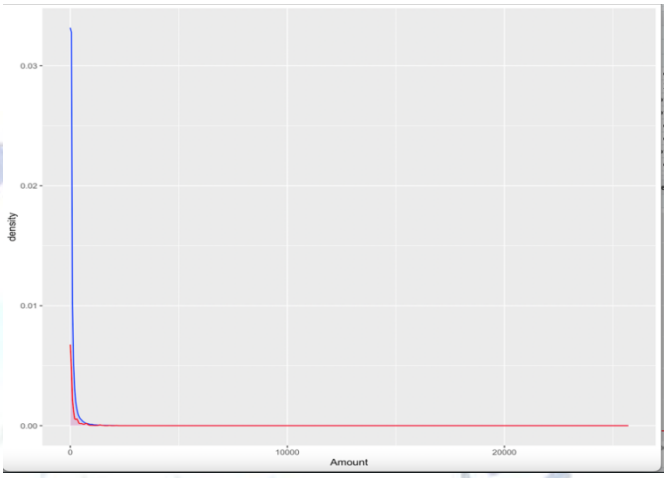


Fig 4. Amount Variable

C. Data Modelling

We will split our dataset into a training and test set with a split ratio of 0.80 once we have standardized our whole dataset. This means that the train data will account for 80% of our information, while the test data will account for 20%.

D. Data Preparation

The preliminary analysis entails choosing an acceptable sample strategy because the dataset is severely imbalanced. Using some technique, such as changing the amount of the original data, sampling methods will convert an imbalanced dataset into a balanced distribution. The difficulty with imbalanced datasets is that most machine learning approaches will disregard the minority class, which is fraud, and hence perform poorly on it.

Undersampling, Oversampling, and ROSE sampling approaches were tested for this objective. The performance of each sampling technique on the dataset is measured using ROC-AUC performance measure. The ROC-AUC gives you a number between 0 and 1, with one being the best and 0 being the worst.

We chose the sampling approach with the maximum area under the curve (AUC) based on the study of the three sample methods since it will deliver the best performance. For our dataset, oversampling produced the greatest AUC.

E. Implementing machine learning algorithms

1. Logistic model

Among the most popular often used algorithms for classification in machine learning is logistic regression (or logit model). The logit model illustrates how continuous, binary, and categorical elements are linked. It's possible tohave binary dependent variables. Based on some predictions, we can estimate whether something will happen or not. We determine the probability of belonging to each category for a given set of predictors.

With just an accuracy of 99.92 per cent and an AUC of 0.821, the logistic model classified 68 of the 106 fraud cases in the sample.

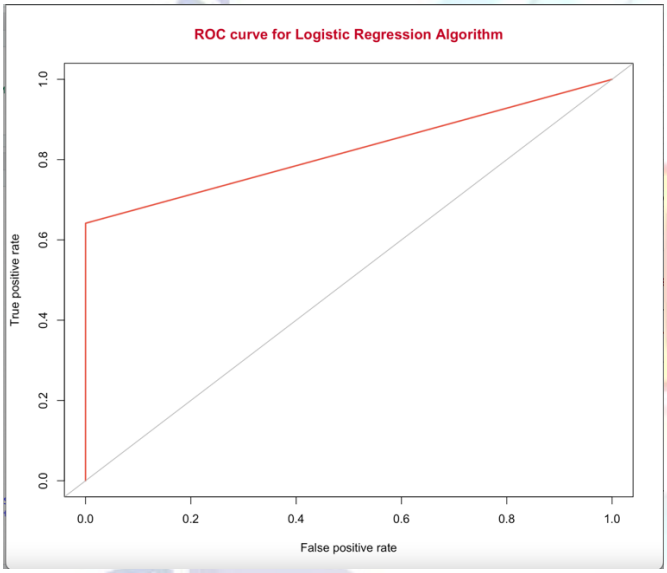


Fig 6. ROC-curve for logistic Regression

2. Decision-tree algorithm

The most basic and extensively used categorization approach is the decision tree algorithm. When constructing a model, this technique assesses all the data's available characteristics and chooses the most relevant.

As a result of this benefit, decision tree algorithms are also employed to determine the relevance of feature metrics. Which was employed in the feature selection process.

The AUC is 0.887, and the accuracy is 99.95 percent, decision tree was able to classify 82 out of 106 frauds. This algorithm is marginally better than Logistic Regression.

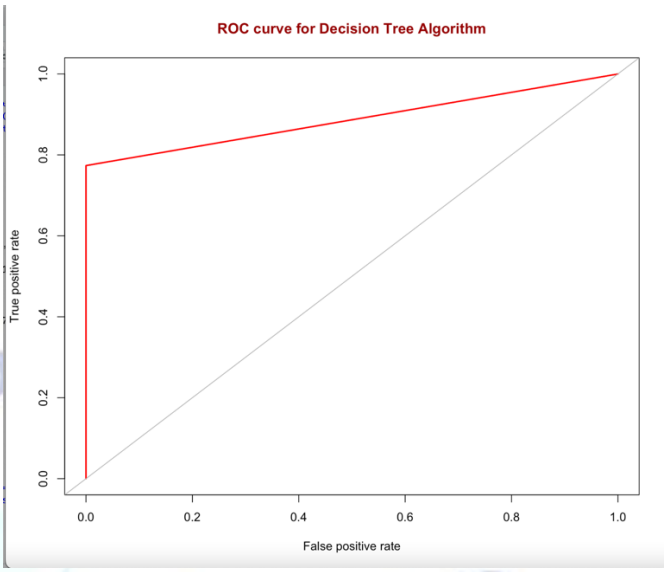


Fig 7. ROC curve for Decision tree algorithm

3. Random Forest

This algorithm selects the outcome based on decision tree classification assumptions. It forecasts by combining the results of numerous trees. The output accuracy improves as the number of trees rises. Using a random forest approach, the shortcomings of a decision tree algorithm are avoided. It improves accuracy while reducing overfitting in datasets. It generates projections without requiring a lengthy list of package parameters.

With an AUC of 0.901, random forest surpasses the other two algorithms, successfully identifying 85 of 106 frauds.

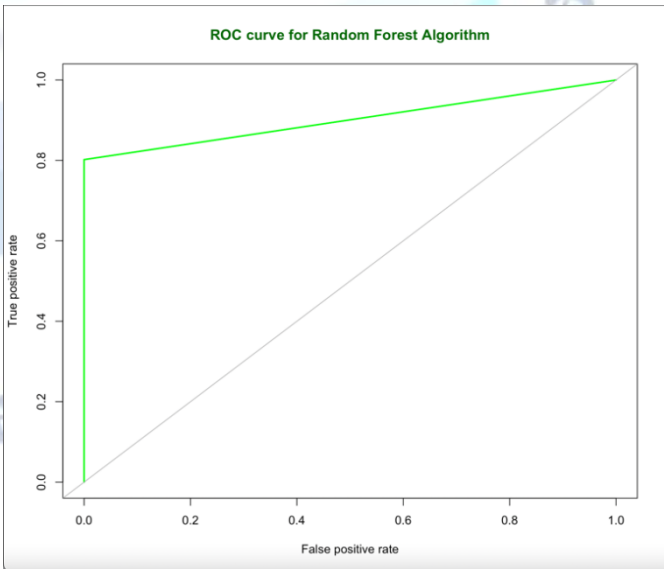


Fig 8. ROC curve for Random Forest

4. XGboost

Extreme Gradient Boosting (XGBoost) is a distributed A scalable GBDT machine learning toolset. It is the ideal machine learning tool for regression, classification, and ranking challenges since it uses parallel tree boosting.

XGboost provides a good result, with an accuracy of 99.96 percent and a classification rate of 87 out of 106 frauds. And a 0.910 AUC.

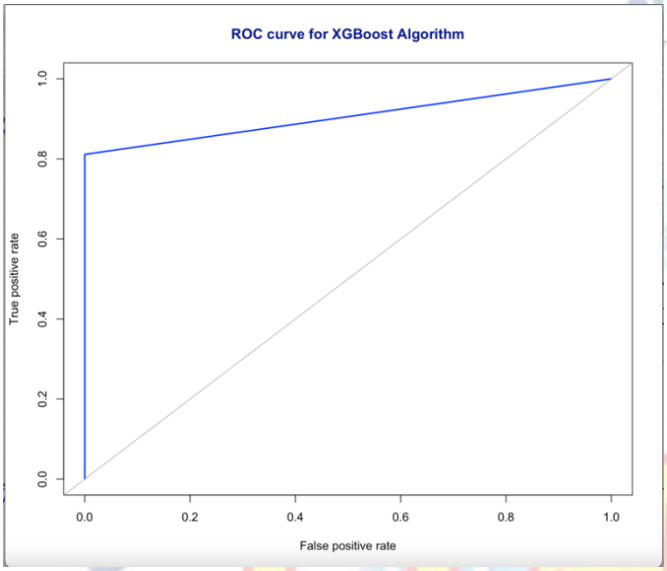


Fig 9. ROC curve for XGboost

4. RESULT

Algorithm	Accuracy	Correct Prediction	Precision	AUC
Logistic Regression	99.92	68	99.93	0.821
Decision tree	99.92	82	99.96	0.887
Random Forest	99.95	85	99.96	0.901
XGboost	99.96	87	99.96	0.910

Based on the data gathered, all the models performed admirably. The accuracy of all the models was more than 90%. The data from the confusion matrix may be used to construct a variety of quality metrics for classification results.

The most essential Precision and accuracy are two metrics, with precision reflecting how near a categorization should be to the genuine (true) values how closely repeated classifications under the same conditions produce the same results. As a result, a classifier might be accurate but not precise, or vice versa.

When compared to other classifiers, eXtreme Gradient Boosting is the best classifier for identifying credit card fraud, with 99.96 % accuracy and precision.

5. CONCLUSION

Because it includes both the customer and the bank, credit card monitoring is a key role for merchant banks. It is necessary to improve credit card monitoring by combining one or more algorithms. The role of machine learning in fraud detection, as well as its methodology, has been thoroughly explored in this work. Analyze and compare the four algorithms, as well as list the numerous applications for each.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] "Machine learning approaches for credit card fraud detection". "S. Venkata Suryanarayana, G. N. Balaji G. Venkateswara Rao". International journal of engineering and technology (IJET) 7(2) PP. 917-920 (2018).

[2] "Detection and prediction of credit card transaction using machine learning". "Kaithekuzhical Leena Kurian and Dr. Ajeet Chikkamannur". International journal of engineering science and research technology (IJESRT): 8(3) ISSN: 277-9655 (2019).

[3] "Credit card fraud detection using machine learning and data science". "Aditya Saini, Swarna Deep Sarkar, Shadab Ahmed, S P Maniraj". International journal of engineering research and technology (IJERT) vol:8 issue 09 ISSN: 2278-0181(2019).

[4] "Machine learning based credit card analysis modelling, detection, and deployment". "Shivkumar Goel and Hitesh Patil" International journal of advanced research (IJAR) ISSN: 2320-5407 (2017).

[5] "Credit card fraud detection using Random Forest" "Devi Meenakshi. B, Janani. B, Gayathri. S, MrsIndira.N" International research journal of engineering and technology (IRJET) VOL. 06 issue: 03 ISSN 2395-0072. (2019)

[6] "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines". Hong Kong, China: The International Multi Conference of Engineers and Computer Scientists. Sahin, Y., &Duman, E. (2011).

[7] Xuan, Shiyang, et al. "Random Forest for Credit Card Fraud Detection." 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), 2018, doi:10.1109/icnsc.2018.8361343.

[8] Kuldeep Randhawa, et al."Credit Card Fraud Detection Using AdaBoost and Majority Voting." IEEE Access, 6 (2018), pp. 14277-14284 doi:10.1109/access.2018.2806420

[9] Roy, Abhimanyu, et al. "Deep Learning Detecting Fraud in Credit Card Transactions." 2018 Systems and Information Engineering Design Symposium (SIEDS), 2018, doi:10.1109/sieds.2018.8374722.

[10] Changjun Jiang, et al. n"Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism." IEEE Internet of Things Journal, 5 (2018), pp. 3637-3647

[11] Sai Kiran, Jyoti Guru, Rishabh Kumar, Naveen Kumar, Deepak Katariya, Maheshwar Sharma, Credit card fraud detection using Naïve Bayes model based and KNN classifier, Published in: International Journal of Advance Research, Ideas and Innovations in Technology, Issue no. 3, vol.no. 4, 2018, pp.44-47.