IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124136

A Review on Credit Card Fraud Detection Using Machine Learning

Dr. Kiran[1], Raju Poovarsha[2], Sanchitha L Anand[3], Soujanya G V[4], Samudyata S[5]

Assistant Professor, Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering, Mysuru, India[1]

Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering, Mysuru, India[2-5]

Abstract: Digitalization enabled all economic opportunities while also perplexing the system with illegal activities. Credit cards are an example of a banking system advancement. The ease of use of credit cards enabled it to attract new users every day. Because of its popularity, the number of fake users, false transactions, and card theft has increased over the years. To puta stop to such illegal acts, fraud detection systems were created.The goal of our proposed paper is to determine whether the completed transaction is true or false. We used ML techniques such as logistic regression and random forest to extract the results. The Random Forest algorithm approach has been shown to provide an accurate estimate of generalization error. The Random Forest algorithm approach was discovered toprovide a good estimate of the generalization error, to be resistant to overfitting, and to be very stable. The obtained results are assessed based on their accuracy, specificity, and precision.

Keywords: credit card, fraud detection, logistic regression, random forest

I. INTRODUCTION

In the twenty-first century, the majority of financial institutions have increased the public's access to business services via internet banking. In today's competitive financial society, electronic payment methods are critical. They have made it very easy to buy goods and services. Customers are frequently given cards by financial institutions that allow them to shop without carrying cash. Credit cards, like debit cards, benefit consumers by protecting them against damaged, lost, or stolen goods. Customers must verify the transaction with the merchant before utilising a credit card. Statistics show that Visa and Mastercard issued 2287 million credit cards globally in 2017. MasterCard issued 1131 million, whereas Visa issued 1131 million.

These statistics demonstrate how card-based transactions became popular among end users. Due to the significant portion of international transactions that fall under this category, fraudsters are laying the groundwork for manipulating this demographic. And if socially engineering people is sometimes easy. Credit cards have a lot of benefits for customers, but they are also connected to problems like fraud and security. This issue is one that banks and other financial organisations are addressing. The issue of credit card fraud is one that banks and other financial organisations are addressing. Through unprotected internet platforms and websites, credit card information is susceptible to theft. They could also be obtained as a result of identity theft. Fraudsters may unlawfully access users' credit and debit card numbers without their knowledge or consent.. One of the main reasons for financial losses in the finance industry, according to "U.K. finance," is due to fraudulent usage of credit and debit cards. It is a major threat that leads to massive financial losses worldwide as a result of technological advancement. As a result, detection is critical in order to reduce financial setbacks. Machine learning is effective at distinguishing between legitimate and fraudulent transactions. The barrier to sharing ideas on fraud detection is one of the most important challenges related with detection methods. A "U.K. finance" research indicates that there are more credit and debit cards in use nowadays.

Credit card fraud detection has increased dramatically in recent years, attracting the attention of most scholars and researchers. This study paper's goal is to analyse and assess many areas of detecting credit and debit fraud. Before suggesting a more effective method for combating credit card fraud, the article looks into various techniques for identifying fraudulent credit card transactions. Researchers are working to overcome some methodological barriers that are limiting the use of ML in real-time applications. The detection of abnormal patterns, biometric identification, diabetes prediction, happiness prediction, water quality prediction, accident prevention at Heathrow, timely diagnosis of bone diseases, and prediction of informational efficiency using deep neural networks are just a few of the studies that have been done in various fields. Researchers are attempting to increase ML's capacity for fraud detection despite these limitations.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124136

II. CREDIT CARD FRAUD DETECTION SYSTEM

Classification of transactions in the dataset which are fraudulent or non-fraudulent by making use of algorithms like random forest algorithm and logical regression is the main objective of this paper. We can determine credit card fraud transactions more accurately by comparing these two algorithms. The diagram with the complete architecture of fraud detection system contains numerous steps starting with data collection to model deployment and the result will be based on the analysis. In this paper we consider Kaggledataset for credit card fraud and pre-processing will be performed on it.

A. Random Forest Algorithm

One of the natural learning algorithm is the Random Forest algorithm. This algorithm is used for regression and classification problem solving. Classification problems are solved primarily using this algorithm. Decision tress are created from Random forest algorithm and each sample data is predicted from it. This algorithm performs single decision trees because it reduces overfitting by averaging the outcome. Hence it is called as the ensemble method.

B. Logistic Regression

Logistic regression performs both regression and classification tasks. Categorical variables are predicted by logistic regression using dependent variables. Sigmoid function or the logistic function employs logistic regression which is one of the most complex cost function. To be linearly related logistic regression does not require variables which are independent and also variance is equal within each group making it less constricted to statistical analysis procedure. As a result the likelihood credit card fraud transactions is employed by this algorithm.

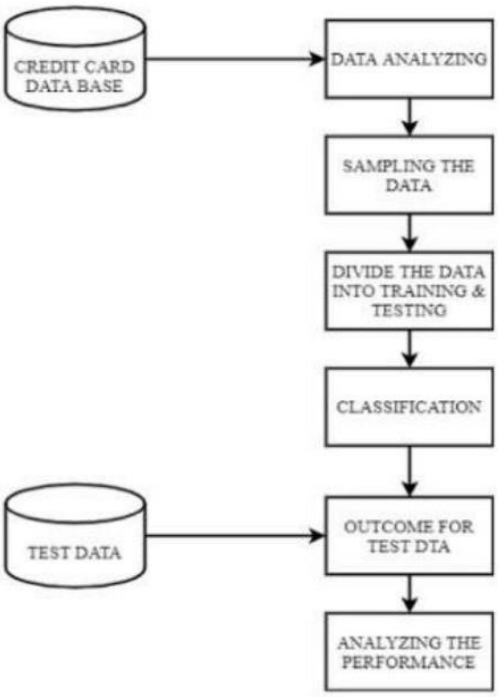

Figure1: General operating flow chart of Credit Card Fraud Detection System

III. LITERATURE REVIEW

Saiju, Sanisa, S. Akshaya Jyothy, Christeena Sebastian, Liss Mathew, and Tintu Sabu [1] In terms of application domain, the supervised algorithm like random forest stands first in the literature. Likelihood fraudulent transactions can be identified easily as soon as the algorithms are integrated into the fraud detection system of a bank.. Larger risks and losses from the banks can be minimized by using various anti- fraud techniques. In contrast to past classification problems, we took a new approach to the study's objective by implementing a variable penalty for misclassification. The suggested system's performance is assessed using precision, f1score, and accuracy. We investigated the information, showing the features and identifying any data imbalances. The suggested system's performance is assessed using precision, f1-score, and accuracy.

DOI: 10.17148/IJARCCE.2023.124136

Arafath, Yeasin, Animesh Chandra Roy, M. Shamim Kaiser, and Mohammad Shamsul Arefin [2] The sequence is put together during the detection phase after the credit card holder's shopping habits are assessed during the training phase using the kmeans clustering method. Using sequence alignment based on real cardholder transaction history and transaction behavioural changes, a successful score is determined in the first stage. The fraudulent transaction signature from the first fraudulent transaction is used to generate the bad score in the second point. If the gap between the good and poor scores is greater than a specific limit, the transaction is unlawful; otherwise, it is allowed.

Awoyemi, John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadare [3] The experiments are presented and discussed in two steps. Eight classification methods are compared in the first stage. Three factors were taken into consideration for the comparison: sensitivity, accuracy, and the area under the precision recall curve (AUPRC). SVM and ANN are two of the top algorithms chosen as a consequence of this comparison. The second phase then compares various imbalance classification methodologies, including Random Oversampling, One Class Classification, and Cost Sensitive, using the chosen algorithms. The SVM's performance is then evaluated against that of the One-Class Classification SVM and the Cost Sensitive SVM when used as a binary classification tool. Additionally, the AutoAssociative Neural Network is used and contrasted with the ANN.

Ileberi, Emmanuel, Yanxia Sun, and Zenghui Wang [4] In this study, a feature selection approach for a machine learning (ML)-based credit card fraud detection engine is proposed. It uses the genetic algorithm (GA). With the help of a dataset created from European cardholders, the effectiveness of the suggested fraud detection engine was assessed. The suggested detection engine uses the ML classifiers Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Artificial Neural Network (ANN), and Naive Bayes after choosing the optimal features (NB).Using the Synthetic Minority Oversampling Technique (SMOTE) oversampling technique, the researcher solved the issue of class imbalance in the dataset. The researcher evaluated the effectiveness of each ML technique using classification accuracy. The imbalance of the dataset was addressed by the authors of this study usinga hybrid sampling method. The MLs LR, NB, and KNN were taken into consideration. The research used a ML framework built on Python. The major performance indicator for evaluating the effectiveness of each ML method was accuracy. According to the experimental findings, the accuracy levels for the NB, LR, and KNN were 97.92%, 54.86%, and 97.69%, respectively.

Sadineni, Praveen Kumar [5] Through the use of Artificial Neural Networks (ANN), Decision Trees, Support Vector Machines (SVM), Logistic Regression, and Random Forest, the current study aimed to identify fraudulent transactions. Accuracy, precision, and false alarm rate are used to assess how well each technique performs. The experiment's data set came from the Kaggle data archive. 150000 trans actions were included in the dataset that was compiled. The data set contained many fields. In order to achieve dimension reduction and only extract the necessary properties, including, among other things, transaction time, amount, and transaction class, they employed principal component analysis to separate useful variables from irrelevant ones. The employment of these tactics has the drawback that not all situations will have the same result. The size and nature of the dataset affect how well the strategies work. The ANN model is reliable, but it requires time consuming and expensive training. SVM delivers outstanding results and performs well with tiny datasets. Decision Tree excels with sampled and preprocessed data, whereas Logistic Regression excels with unprocessed, raw data. With categorical and continuous data,random forest performs well.

Sanobar khan, Sanovar, Suneel Kumar , Mr Hitesh Kumar [6] 28 of the 31 columns in the datasets under examination have the labels v1v28 to protect sensitive data, making a total of 31 columns. Time, Amount, and Class are represented in the remaining columns. The time frame between the first and second transactions in a row is referredto as time. The amount is the sum of money that was transferred. A fraudulent transaction is one that is a valid class0.After these datasets have been analysed, a histogram is shown for each column that was taken into consideration. A graph of the datasets is created as a result. This will guaranteethat no data value is missed. The correlation between the output assumption variables and the class variables is then depicted using a heat map graph of the data.

Sharma, Pratyush, Souradeep Banerjee, Devyanshi Tiwari, and Jagdish Chandra Patni [7] After obtaining the dataset, the data was separated into train, validation, and test sets. The 70/30 guideline was adhered to, with test data making up 15%, validation data being 15, and training data being 70 Because the dataset was rather substantial and did not require any more data points for training, which could have introduced variance and led to classification bias, this ratio was chosen. Several machine learning models were trained using the dataset utilising the logistic regression, support vector machine, and random forest algorithms. The performance of each model is evaluated to produce a comparative analysis after the machine learning models have been trained using all of the machine learning techniques. The macro averages of the F 1 score, recall, and precision are used.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124136

selected point and its neighbours. Random forest algorithm is another method that is employed. From the total of "m" features collected from the multinational dataset, the algorithm chooses "k" features. The created nodes are then sent through a splitting function using the chosen features. The best split function is typically employed when using a decision tree. Recursively dividing the nodes results in the generation of the number of daughter nodes. To limit the amount of nodes generated for each tree, there should be a limit defining how many of these nodes should be created.

Azhan, Mohammed, and Shazli Meraj [10] Study throws light on the use of machine learning and neural networks to spot future fraudsters by examining their past wrongdoings and data on previous fraudsters is explored. Support Vector Machines, Logistic Random Forest Regression, Multinomial Naive Bayes, and a Simple Neural Network are also used. The Machine Learning Group at ULB assembled and made the dataset public (Universite Libre de Bruxelles). It had no missing values, only numerical inputs, and was notably unbalanced that is, the percentage of positive and negative dataseries is very different. The column Class contains the information about whether or not the transaction was fraudulent; a value of 1 indicates fraud and a value of 0 indicates otherwise. A confusion matrix and classification reports generated by the Sklearn software were used to assess the models. The categorization reports have been presented in figures, and the conclusion includes a thorough table of model-by-model comparisons. The classifiers' ROC characteristics score is also displayed on a graph for comparison. Machine learning strategies have shown to be more effective at addressing the issue of class imbalance than a shallow neural network. The distribution of class weights in neural networks barely affects how the class imbalance is managed. There are further techniques that can be used, such as using Cost Sensitive Loss Functions, Over-Sampling, and Under-Sampling.

Meenakshi, B. Devi, B. Janani, S. Gayathri, and N. Indira [8] The credit card dataset is classified by the suggested system using the random forest technique. An approach for classification and regression is called Random Forest. It is essentially a group of decision tree classifiers. Decision trees perform worse than random forests since the former breaks the bad habit of overfitting the training set. Training involves sampling of a subset. Following, a decision tree is built, with each node splitting on a feature selected at random from the entire feature set. Each tree is trained independently of the others, which makes training incredibly quick even for big data sets with numerous characteristics and data instances. The algorithm has been found to be resistant to overfitting and to provide a reliable estimate.

Priya, G. Jaculine, and S. Saradha [9] The suggested device Because there are fewer minority class data in the dataset, SMOTE (Synthetic Minority Oversampling Technique) synthesises minority class elements based on those that already exist. It operates by choosing any unspecified point from the minority class, then calculating its k-nearest neighbours. The synthetic points are added between the

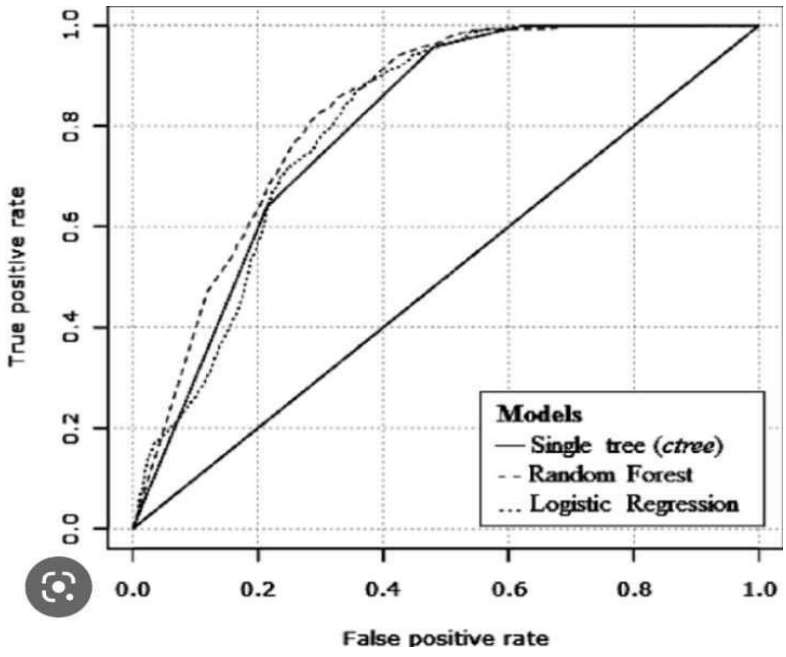

Figure 2: Graphical representation of Logistic regression vs Random forest

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124136

AuthorName	Title		TechniqueUsed	Accuracy
Sadineni, Praveen Kumar	"Detection of fraudulent transactionsin credit card using machine learning algorithms."		Random Forest Algorithm Logistic Regression DecisionTree SVM and ANN	Radom Forest - 99.21% Decision Tree - 98.47%. Logistic Regression - 95.55%, SVM - 95.16% ANN - 99.92%.
Ileberi, Emmanuel, Yanxia Sun, and Zenghui Wang.	"A machine learning based credit card fraud detection using the GA algorithm for feature selection."		G A algorithm Naïve Bayes Artificial Neural Network	GA ANN – 81.82% Naïve Bayes – 99.23% ANN – 88.93%
Saiju, Sanisa, S. Akshaya Jyothy, Christeena Sebastian, Liss Mathew, and Tintu Sabu.	"Credit Card Fraud Detection Using Machine Learning."		Random Forest Isolation Forest Local outlier Factor SVM	Random Forest – 99.92% Isolation Forest – 99.72% Local outlier Factor – 99.65% SVM – 70%
Sharma, Pratyush, Souradeep Banerjee, Devyanshi Tiwari, and Jagdish Chandra Patni	"Machine learning model for credit card fraud detection-a comparative analysis.		ANN SVM Fuzzy Logic DecisionTrees	ANN – 99.71% SVM – 85.45% Fuzzy Logic –77.8% Decision Trees – 97.93%
Awoyemi, John O., Adebayo O. Adetunmbi, and SamuelA. Oluwadare.	"Credit cardfraud detection using machine learning techniques		Naïve Bayes K-Nearest Neighbour	Naïve Bayes – 97.92% K-Nearest Neighbour – 97.69%

Table 1 : Comparison table between methods and their accuracy score

DOI: 10.17148/IJARCCE.2023.124136

IV. CONCLUSION

This review study looks into the many methods used. Conclusion: ML approaches are a great tool to increase the precision of detection methodologies. The model must be trained on huge datasets in order to avoid data imbalance. Real-time datasets can give us access to a wider variety of data, but privacy is still an issue. In order to train the model while protecting privacy, we are taking into account the real- time datasets accessible. The suggested approach can help financial institutions and banks work together to use real-time datasets, which would be beneficial for everyone in terms of creating a system that is effective at detecting fraud. Despite its effectiveness, the proposed method has limits when it comes to real world deployment because it takes a lot of time and engineering resources to integrate, and even then, the outcome is still not ideal because it only uses a small portion of the total data available. Because each bank and finance institution have its own restrictions and relies on internal resources rather than a centralised strategy, adapting the suggested method will be challenging. As a result, even with the constraints still there, more needs to be done to convince banks and other financial organisations to adopt this technology.

REFERENCES

[1] Saiju, Sanisa, S. Akshaya Jyothy, Christeena Sebastian, Liss Mathew, and Tintu Sabu. "Credit Card Fraud Detection Using Machine Learning." *International Journal of Recent Advances in Multidisciplinary Topics* 2, no. 4 (2021): 31-34

[2] Arafath, Yeasin, Animesh Chandra Roy, M. Shamim Kaiser, and Mohammad Shamsul Arefin. "Developing a Framework for Credit Card Fraud Detection." In *Proceedings of the International Conference on Big Data, IoT, and Machine Learning*, pp. 637-651. Springer, Singapore, 2022.

[3] Awoyemi, John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadare. "Credit card fraud detection using machine learning techniques: A comparative analysis."In *2018 international conference on computing networking and informatics (ICCNI)*, pp. 1-9. IEEE, 2018.

[4] Ileberi, Emmanuel, Yanxia Sun, and Zenghui Wang. "A machine learning based credit card fraud detection using the GA algorithm for feature selection." *Journal of Big Data* 9, no. 1 (2022): 1-17.

[5] Sadineni, Praveen Kumar. "Detection of fraudulent transactions in credit card using machine learningalgorithms." In *2020 Fourth International Conference on I- SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 659-660. IEEE, 2020.

[6] Sanobar khan, Sanovar, Suneel Kumar , Mr Hitesh Kumar(2021);Credit Card Fraud Detection Using ML; International Journal of Scientific and Research Publications(IJSRP).

[7] Sharma, Pratyush, Souradeep Banerjee, Devyanshi Tiwari, and Jagdish Chandra Patni. "Machine learning model for credit card fraud detection-a comparative analysis." *Int. Arab J. Inf. Technol.* 18, no. 6 (2021): 789-796.

[8] Meenakshi, B. Devi, B. Janani, S. Gayathri, and N. Indira. "Credit card fraud detection using randomforest." *International Research Journal of Engineering and Technology (IRJET)* 6, no. 03 (2019).

[9] Priya, G. Jaculine, and S. Saradha. "Fraud detection and prevention using machine learning algorithms: a review." In 2021 7th International Conference on Electrical Energy Systems (ICEES), pp. 564-568. IEEE, 2021.

[10] Azhan, Mohammed, and Shazli Meraj. "Credit card fraud detection using machine learning and deep learning techniques." In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 514-518. IEEE, 2020.