International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6141
***Article Received**: 21 December 2022 **Revised**: 18 January 2023 **Accepted**: 24 January 2023*

A Novel Machine Learning Algorithms used to Detect Credit Card Fraud Transactions

M. Sudhakar[1], K. P. Kaliyamurthie[2]
[1]Department of Computer Science and Engineering
Bharath Institute of Higher Education and Research
Chennai, India
sudhakarmtech@gmail.com
[2]Department of Computer Science and Engineering
Bharath Institute of Higher Education and Research
Chennai, India
kpkaliyamurthie@gmail.com

Abstract— During the Covid-19 pandemic, the world was under lockdown, and everyone was inside their home. There are so many restrictions for going out, so many companies introduced online shopping, and this online shopping helped more people; the e-commerce platform also increased their revenue; at the same time, online fraud has also risen worldwide. Everyone adopted online shopping during the pandemic. In 2019 India's 2019 credit/debit card fraud rate was 365, according to the National Crime Record Bureau. The developed countries are the highest rate of credit card fraud in 2020 compared to India; for that reason, we have to implement mechanisms that can detect credit theft. The machine learning algorithm with the R program will play an essential role in credit card fraud detection. The following machine learning algorithm will have used for credit card fraud, Random Forest, Logistic regression, Decision trees, and Gradient Boosting Classifiers. The European bank dataset used in our research and the dataset size is 284808. Here we used two classes, the first one is called the positive class (fraud transactions), and the second one is the negative class (genuine transactions). The final result will show us the outperforms of our proposed system.

Keywords- Credit card, Decision tree, Logistic regression, Gradient boosting classifiers and Random Forest.

I. INTRODUCTION

The bank gives their customers credit cards to buy the products online using their credit card and allows them to withdraw cash in advance within the permitted limit. The fraudulent can buy the product or remove the amount from the credit card without the owner of the card. The Internet's growth keeps increasing, and dishonest people have taken advantage of this development to steal people's information and other details. Everyone started using the Internet. Before the pandemic, e-commerce was not that famous, and credit card usage was also significantly less, but after the pandemic, everyone started using e-commerce using their credit card. So, fraud also increases their activities to attack the transactions that are made using credit cards. Many mechanisms are used to protect credit card transactions but, in most cases, are not fully protected [1]. In 2021 1.7 million people were affected by credit card fraud; compared to 2019 and 2020, it increased. Figure 1 shows the theft rate in the U.S. reached 2019, and the fraud rate increased in 2020 and 2021 due to the vast number of online transactions during the pandemic [2]. Nowadays, criminals are not using physical cards to steal money; instead, they are using credit card details to steal money. In 2019 over 100 million people's credit card information was released by hackers. Hackers are stealing your credit card information in various ways, when we use public WiFi, Phishing attaches, sending malware to your computers, scam phone calls, getting your credit card number after breaching your system and hacking the online store databases. According to the statistical report, 1.1 trillion credit cards were issued to the public from 2012 to 2019, published by the United States.

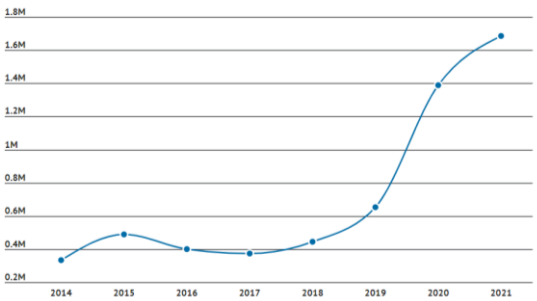

Figure 1. Credit card theft report

The Federal trade commission said that 179 million users' credit card was stolen, and 1579 data breaches happened. Many machine learning algorithms were implemented to detect credit card fraud detection. Still, they provide low accuracy of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6141
***Article Received**: 21 December 2022 **Revised**: 18 January 2023 **Accepted**: 24 January 2023*

credit card fraud detection [3], and our proposed algorithm will perform with high accuracy [11].

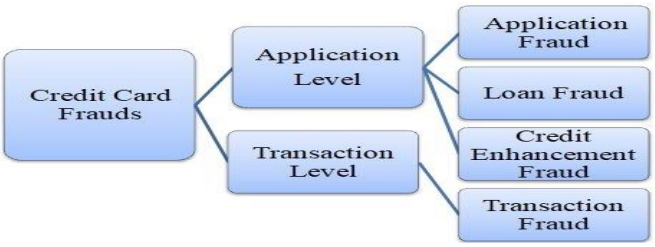

Figure 2: Taxonomy for credit card frauds

The proposed algorithms of Random forest, Decision tree, Logistic Regression and Xgboost have provided better accuracy and performance than the other machine learning used in the credit card fraudulent methods and the AUC score, and the overall performance are also 1%.

II. RELATED RESEARCH WORKS

(Navanshu K et al., 2018) [4] In their research, the authors have implemented four machine learning algorithms (Logistic Regression, Decision Tree, Support Vector Machine and Random Forest) to detect credit card fraud. The author used the year 2013 European cardholder details as a dataset. After performinSg the machine learning approach, the accuracy of the result is 97.70% for logistic regression, 95.50 % accuracy from the decision tree, 97.50% accuracy from the support vector machine and 98.60% from the random forest and the outcomes of the accuracy also is good; and also the author suggested that in future if we implement advanced pre-processing techniques, we can get better accuracy.

(Maniraj et al., 2019) [6] explained the credit card fraud transaction using the modelling of the dataset, and the authors are trying to detect the accuracy for 100%. Still, the result was achieved with 99.6% accuracy, and the precision rate was about 28%. The algorithm used in their research is the Local Factor Isolation Forest Algorithm, and the precision rate increased when they entered all datasets.

(Varmedja D et al., 2019) [5] In their research, the dataset was collected from the Kaggle and used it to detect credit card fraud; this dataset contains only two days' transaction details. This researcher implemented three machine learning algorithms (Random Forest, Naïve Bayes and Multilayer perception), and the result of the credit card detection is as follows, the Random forest will provide 99.96% of accuracy, Naïve Bayes will provide 99.23% of accuracy, and the Multilayer perception will provide 99.93% of accuracy. At the end of the research, the author suggested future work is that more research should be conducted to improve the accuracy of other machine learning algorithms.

(Awoyemi et al., 2017) [16] Explained the two problems of credit card fraud detection, and the first problem is the fraud keeps changing their identity and locations, and the second problem is the credit card fraud datasets are highly skewed. They used 284,807 datasets from European cardholders' transaction details, and three algorithms were used in their research. These are the three algorithms that were used in their study to detect credit card fraud detection Naive Bayes, K-Nearest Neighbor and Logistic Regression). Finally, the accuracy for the Naive Bayes algorithm was 97.92 %, the accuracy for the K-Nearest Neighbor was 97.69%, and the accuracy for the logistic regression was 54.86%.

(Mital et al., 2019) [8] In their research, Artificial Intelligence and Neural networks were used to detect credit card fraud, and the dataset's distribution is highly imbalanced. Therefore, the authors designed and used under-sampling and oversampling techniques to balance the data. Again in their research, data mining techniques were also implemented to achieve more accuracy in the fraud detection system when they used a hybrid model combining pre-existing supervised and unsupervised learning techniques for more accuracy.

(Norton M et al., 2019) [19] In their research work, authors suggested that machine learning, data science and deep learning will help us to credit card fraud detection, and these types of fallacious activities can be done. The advantages of these three combination models will help the banking sectors and financial institutions detect fraud as much as possible before theft. In future, we have to use a combination of supervised and unsupervised learning approaches.

Mohari et al. (2021) [20] the author used Data Science and Machine learning with Deep learning techniques to detect credit card fraud. Their research will help the banks and financial institutions see the fraudulent transaction before it causes damage. The author used different types of unsupervised learning techniques (Logistics Regression, Random Forest, AdaBoost, Artificial Neural Network, Genetic Algorithm, Hidden Markov Model (HMM), KNN Classifier, Decision tree, Isolation Forest, and Local Outlier Factor) and compared them and Local Outlier Factor will provide the highest accuracy among other algorithms.

III. DATASET

The dataset is collected from European credit card cardholder transactions. This transaction happened in 2013, and the total number of transactions is 284807. In this transaction, around 492 credit card frauds happened within two days. The dataset contains only numerical values (V1, V2, V3, V4……V28) [12], and it has thirty features followed by time and amount. There are two types of value used here numerical value 1 represents the fraudulent transactions, and numerical value 0

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6141
***Article Received**: 21 December 2022 **Revised**: 18 January 2023 **Accepted**: 24 January 2023*

represents the typical transaction. Previous research used imbalanced classes to detect credit card fraudulence, but in our research, we will use balanced classes; for that reason, we will use Synthetic Minority over Sampling Techniques [9]. The dataset contains ten rows and 31 columns, 29 contain a feature, and only one contains a class.

Serial Number	Feature	Description
1	Time	Time indicates in seconds. It will specify the elapses between the first transaction and the current transaction.
2	Amount	Details of the transaction amount
3	Class	0 – no fraud transaction 1- Fraud transaction

TABLE I. EUROPEAN ATTRIBUTES

Figure.3 will show the fraudulent and non-fraudulent transactions statistical report. In this output report, the total number of transactions is 1048575.000000; this non-fraudulent amount distribution is 1042569.000000, and the fraud amount distribution is 6006.000000. The report will show us that the minimum non-fraudulent distribution value is 1.00, and the maximum value for the non-fraudulent is 28948.900000. Now let us see that the fraudulent amount distribution ratio for the minimum value is 1.180000, and the maximum value is 1371.810000. Again we will see the mean value for non-fraudulent distribution value is 67.627445 and the fraudulent distribution mean value is 530.573492.

	Row type	Overall amount distribution	Non fraud amount distribution	Non fraud amount distribution
0	Count	1048575.000000	1042569.000000	6006.000000
1	Mean	70.279095	67.627445	530.57492
2	Std	159.951841	153.695606	391.333069
3	Min	1.000000	1.000000	1.180000
4	50%	47.450000	47.220000	391.165000
5	95%	196.260000	189.940000	1085.052500
6	99.9 %	1496.830880	1502.239520	1289.066100
7	Max	28948.900000	28948.900000	1371.810000

TABLE II. DATASET

A. *Data Exploration*

Here we are going to explore the data contained in the credit card data frame by using the head and tail functions.

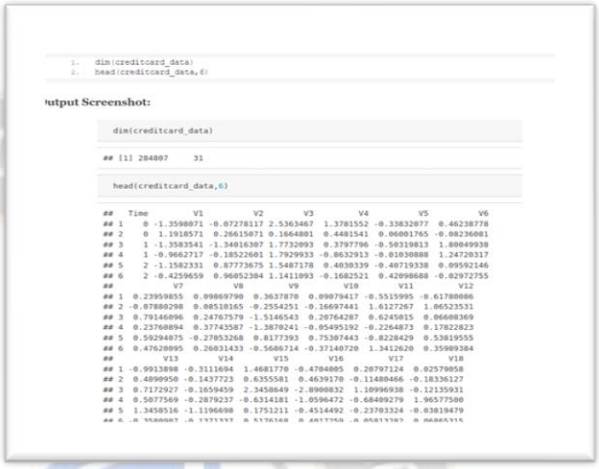

Figure 3 Data Exploration as head function

IV. METHODOLOGY

Figure 4 will show us the workflow of the research. This machine learning model will learn from previous experience and create a new instance from the information given. The dataset will be divided into two sessions one session is used for training, and another session is used to evaluate the model's performance. Two datasets will be used: one data set for training and another for evaluation.

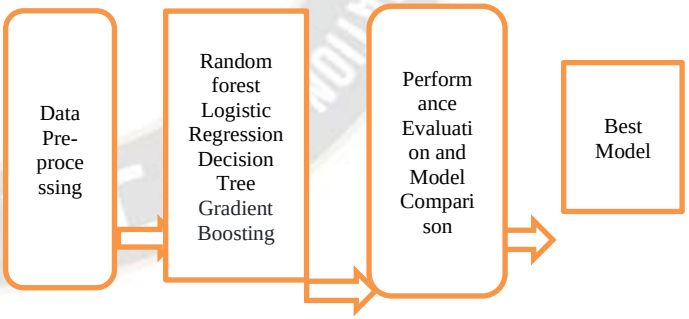

Figure 4: Workflow

A. *Data pre-processing*

Before implementing any machine learning techniques, the data pre-processing mechanism is essential. This step cleans and prepares the data to check missing values and more concise prejudice. The encoding of the data is necessary before using them in the modelling because the data contains both numerical and categorical values. To avoid the data imbalance, we will use the resampling method [10]. The following steps were done data pre-processing, data cleaning,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6141
***Article Received**: 21 December 2022 **Revised**: 18 January 2023 **Accepted**: 24 January 2023*

encoding the categorical data, feature scaling, data resampling, feature correlation and correction, and dataset splitting. Data reassembling is essential because the dataset is imbalanced, so we have to use the under-sampling and over-sampling techniques. In most cases, the dataset belongs to the majority classes, and the dataset will be under-sampling randomly; some instances are not captured for training purposes.

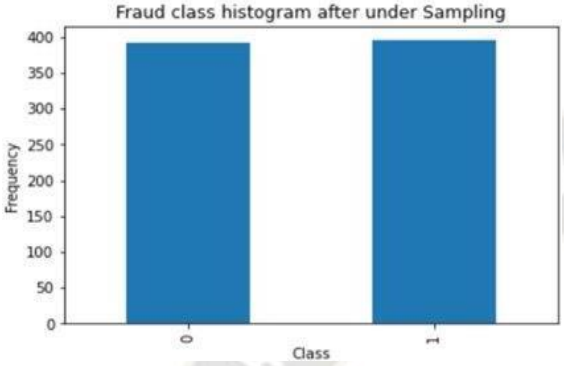

Figure 5 under-sampling after distribution of the classes.

Figure 5 shows that the dataset was under-sampled randomly and reduced the number of classes. If the majority of classes were removed, some critical data instances were not captured for training purposes. Figure 6 shows the better performance because it increases the instances to make the model perform better. Feature selection is an essential and critical part of machine learning methods. Because during the training and testing process, the need for ample feature space may negatively impact the overall performance. The researcher has to select the feature scaling based on their research problem.

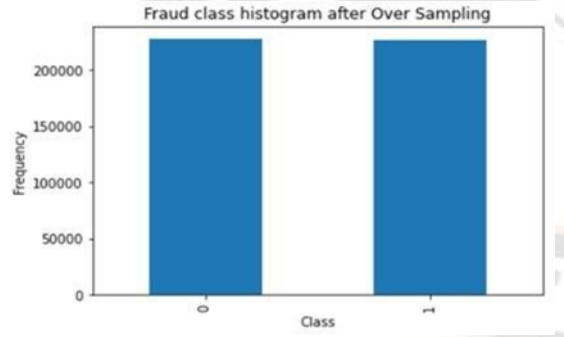

Figure 6. Oversampling after distribution of classes

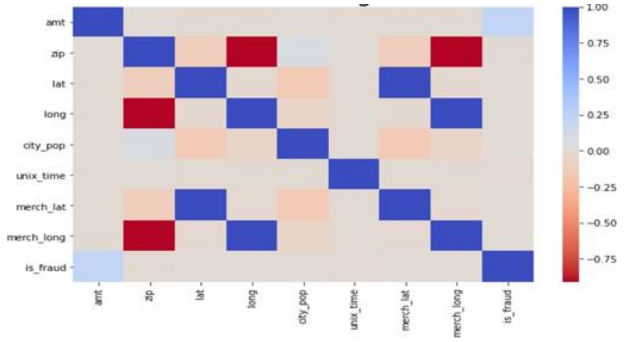

Figure 7. Heat map of the original dataset

Figure 7. Will shows the Original Datasets for the Correlation Heatmap Were Resampled In under sampled and Over-Sampled, Showing That the Information is not revealing In the Heatmap Because Of the More Datasets.

V. RESULT

The experiment will be done in two methods: the first method will classify the process, and the second method will find the efficiency of each algorithm. To process the classification method, we will use feature vector F= {v1, v2, v3, v4}and Random forest, Decision Tree, Logistic regression and Gradient Booster will be trained and tested. In this test, random forest and Xg booster achieved the highest score.

Model	Accuracy	Recall	Precision	F1-Score
Random Forest	99.94	75.22	85.85	80.18
Decision Tree	99.90	76.10	68.80	72.26
Logistic Regression	99.90	53.09	80.800	63.82
Xg Boost	99.95	74.20	87.84	82.35

TABLE III. CLASSIFICATION RESULTS

In this research, we use a machine learning-based binary classification task, and we get the accuracy of each algorithm on the test data, which is the metrics' central performance. Moreover, for each model, we compute the recall, precision and F1 score [17]. Here we used the AUC score to assess the quality of each model. Here the AUC values vary between 0 to 1[18].

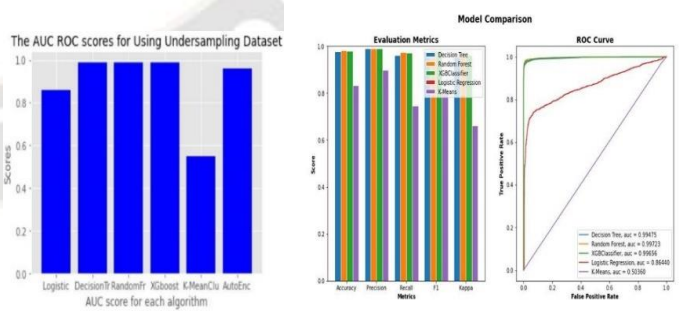

Figure 8 AUC score

Figure 8 will show the under-sampling data for modelling. The performance is still better, and the result is 99%. Without the original dataset, the result still is the same in the decision tree, random forest and Gradient boosting.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6141
***Article Received**: 21 December 2022 **Revised**: 18 January 2023 **Accepted**: 24 January 2023*

	Logistic Regression	Decision Trees	Random Forest	Gradient Boosting
Dataset (Original)	0.74	0.99	0.99	0.99

TABLE IV. UNDER-SAMPLING DATA MODEL SCORE

When we use comparison matrix results for the Random forest, Xg boost, Decision tree and Logistic Regression, we can see that the Decision tree and logistic regression have positive values. Table V. will show the comparison result of each algorithm.

Algorithms	Accuracy	Precision	AUC score
Random Forest	0.9890	0.9521	0.9886
Decision Tree	0.9522	0.9881	0.9982
xgboost	0.9853	0.9822	0.9996
Logistic Regression	0.8179	0.7562	0.8761

TABLE V. COMPARISON RESULTS FOR EACH ALGORITHM

VI. CONCLUSIONS

Today, the growth of technology is making many improvements in society simultaneously, but some are misusing it. Credit card fraud detection is more complex but also a general problem and we have to find a solution for it. In this research, we have discussed credit card and online transaction fraud, which leads to developing machine learning algorithms to detect the online credit card fraud detection mechanism. We used the Random Forest algorithm, Logistic regression algorithm, Decision Tree algorithm and X.G. boost. We compared all the algorithms with different datasets, and all the algorithms gave better results and the XG boost provided better accuracy.

ACKNOWLEDGMENT

I want to thank Dr K.P. Kaliyamurthie and Dr T. Nalini to assist in this reaearch work.

REFERENCES

[1] Iwasokun GB, Omomule TG, Akinyede RO. Encryption and tokenization-based system for credit card information security. Int J Cyber Sec Digital Forensics. 2018;7(3):283–93.

[2] Pumsirirat, A. and Yan, L. (2018). Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. International Journal of Advanced Computer Science and Applications, 9(1).

[3] Thennakoon, Anuruddha, et al. Real-time credit card fraud detection using machine learning. In: 2019 9th international conference on cloud computing, data science & engineering (Confluence). IEEE; 2019.

[4] Navanshu Khare, Saad Yunus Sait, Credit card fraud detection using machine learning models and collating machine learning models. Int J Pure Appl Math. 2018;118(20):825–38.

[5] Varmedja D, Karanovic M, Sladojevic S, Arsenovic M, Anderla A. Credit card fraud detection-machine learning methods. In: 18th international symposium INFOTECH-JAHORINA (INFOTECH); 2019. p. 1-5.

[6] S P, Maniraj & Saini, Aditya & Ahmed, Shadab & Sarkar, Swarna. (2019). CreditCard Fraud Detection using Machine Learning and Data Science. International Journal of Engineering Research and. 08. 10.17577/IJERTV8IS090031.

[7] Khatri S, Arora A, Agrawal AP. Supervised machine learning algorithms for credit card fraud detection: a comparison. In: 10th international conference on cloud computing, data science & engineering (Confluence); 2020. p. 680-683.

[8] S. Mittal and S. Tyagi, "Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection", 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019.

[9] Seera M, Lim CP, Kumar A, Dhamotharan L, Tan KH. An intelligent payment card fraud detection system. Ann OperRes 2021;1–23.

[10] Kasongo SM, Sun Y. A deep long short-term memory based classifier for wireless intrusion detection system. ICT Express. 2020;6(2):98–103.

[11] Abhishek L. Optical character recognition using ensemble of SVM, MLP and extra trees classifier. In: Internationalconference for emerging technology (INCET) IEEE; 2020. p. 1–4.

[12] Kasongo SM. An advanced intrusion detection system for IIoT based on GA and tree based algorithms. IEEE Access.2021;9:113199–212.

[13] Many ID, Sun Y. Improved heart disease prediction using particle swarm optimization based stacked sparseautoencoder. Electronics. 2021;10(19):2347.

[14] Niklas Donges. (2021). A complete guide to the Random Forest algorithm. https://builtin.com/data-science/random-forest-algorithm

[15] Priya, G. & Saradha, S. (2021). Fraud Detection and Prevention Using Machine Learning Algorithms: A Review. 564-568. 10.1109/ICEES51510.2021.9383631.

[16] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare. (2017). "Credit card frauddetection using machine learning techniques: A comparative analysis," International Conference on Computing Networking and Informatics (ICCNI), 2017, pp. 1-9, Doi: 10.1109/ICCNI.2017.8123782.

[17] Raynor de Best. U.S. Credit cards – statistic and fact. (2020). https://www.statista.com/topics/1118/credit-cards-in-the-united-states/

[18] Kasongo SM, Sun Y. A deep long short-term memory based classifier for wireless intrusion detection system. ICTExpress. 2020;6(2):98–103

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2
DOI: https://doi.org/10.17762/ijritcc.v11i2.6141
Article Received: 21 December 2022 Revised: 18 January 2023 Accepted: 24 January 2023

[19] Norton M, Uryasev S. Maximization of auc and buffered auc in binary classification. Math Program.2019;174(1):575–612.

[20] Mohari, Ankit & Dowerah, Joyeeta & Das, Kashyavee & Koucher, Faiyaz & Bora, Dibya & Bora. (2021). A COMPARATIVE STUDY ON CLASSIFICATION ALGORITHMS FOR CREDIT CARD FRAUD DETECTION.