iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52242

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Credit Card Fraud Detection using Python & Machine Learning Algorithms

Ekta Mangal[1], Divya[2], Shubham[3], Radhika Gussain[4]
Mahatma Gandhi Missions College Of Engineering & Technology, Noida, India

Abstract: ***Browsing and many other online sites have increased the digital payment modes through which risk of frauds during transactions got increased. It is necessary to have a look on fraud transactions so that the customers does not pay for what they haven't done. Such complications may be intercept with Data mining through Machine Learning. It aims to display the customization of a data set by applying machine learning with Credit Card Fraud Detection. The CCFD complications comprise of analyzing previous transactions through credit card along the data of the unauthorized users. These models are then applied to analyze whether the new transaction is authorized or not. In this project, we have concentrated on examining and pre-refining the data sets in addition to the deployment of numerous inconsistency observation methods such as Logical Regression, Random Forest, Decision tree, XG Boost on Credit Card Transaction data.***

I. FOREWORD

Swindling through credit card during a transaction is an uncertified and undesirable access one's bank account by a person irrespective of the authorized holder without his knowing. Required prevention measures should be taken to prevent this unwanted access along with the actions corresponding to crooked enactment can be considered to reduce it and protect against homogenous events in the future. A bank card in general allude to a card that is entrust with cardholder (account holder), usually grant him to buy products and assistance in under borrowing limit or take out cash further. It delivers the authorized user a favor of the time with money at any place, i.e., it provides time for their client to pay back later in a authorized time, and payment can be done from anywhere without having cash money. It is an easy target. Dodgers always attempt to make every unauthorized transaction legal, which creates difficulty to detect frauds. Machine learning modules and methods are working to survey all the authorized and unauthorized transactions.

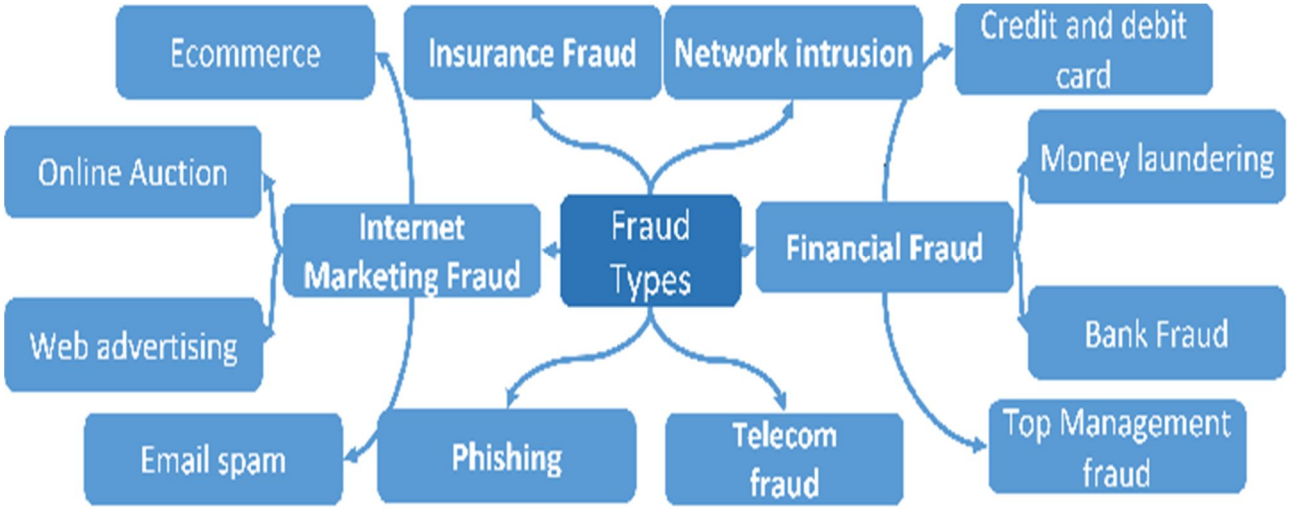


Fig. 1: Taxonomy for Frauds

In the company of various frauds largely credit card frauds, frequent in the reportage and gossips in recent years, swindling is the biggest fear in one's mind. Its dataset is tremendously inconsistent since there can be more legalized transactions when set side by side with a unauthorized one. The graph below shows the number of CNP(Card-Not-Present) frauds cases that were registered in respective years.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

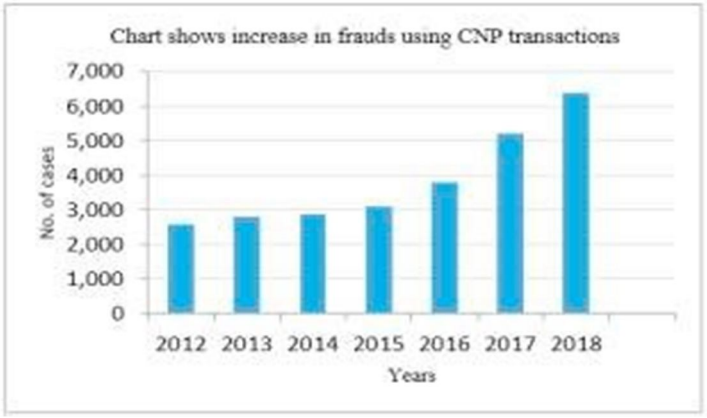
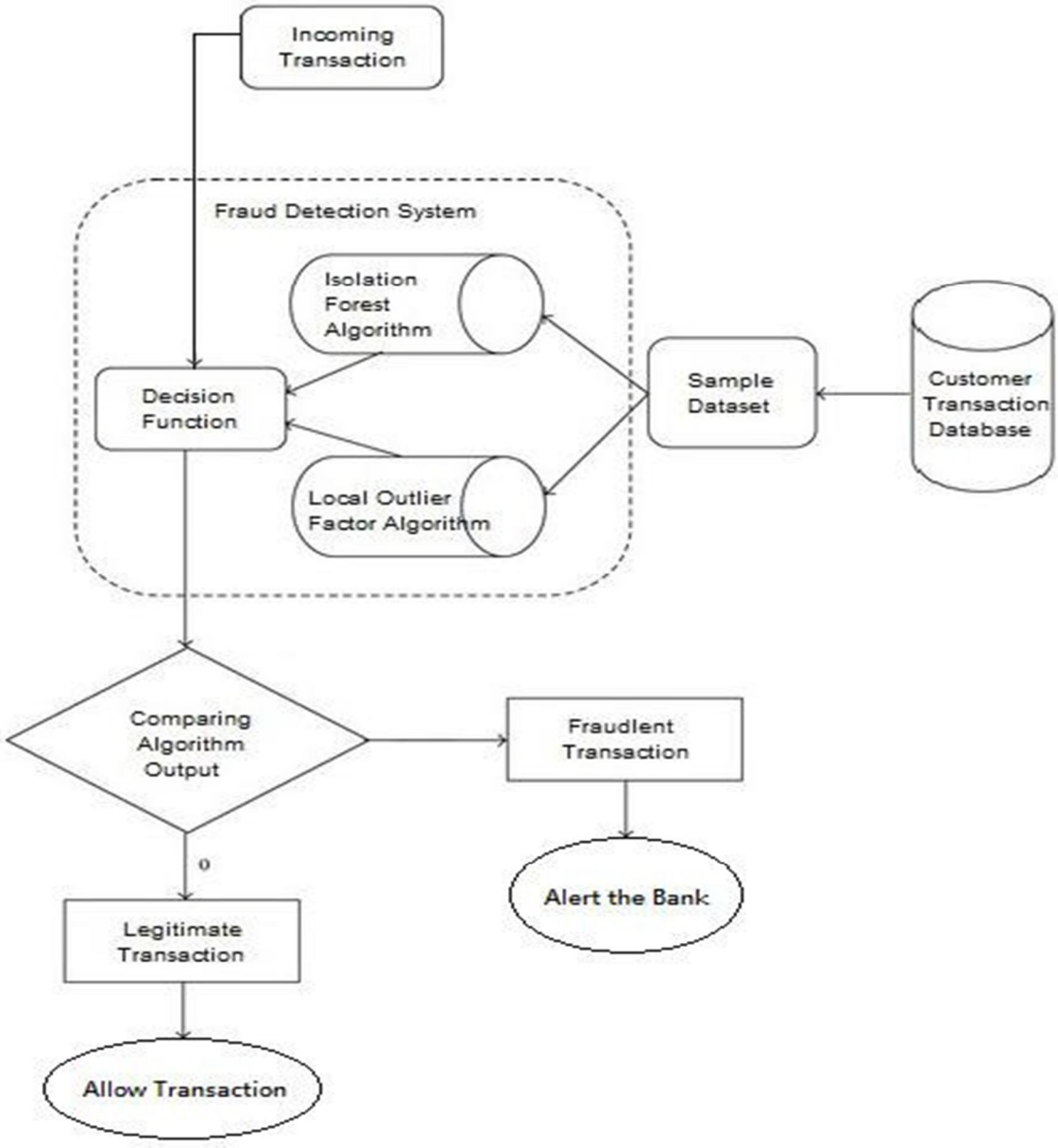


Fig. 2: Frauds Using Card Not Present Transaction



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

II. LITERATURE VIEW

Multiple supervised and semi-supervised machine learning algorithms are now applied for fraud detection. Although these techniques and algorithms are accurate in certain respects, they were unable to offer a lasting and reliable solution. Strong class imbalance, the inclusion of labeled and unlabelled samples, and improving the ability of transactions with high accuracy are the three key obstacles we must face. Other supervised machine learning techniques for detecting credit card fraud include Decision Trees, Logistic Regression, Random Forest, etc. The behavioral characteristics of typical and unusual transactions are trained and tested using these algorithms. In highly skewed credit card fraud, they are exploited.

By using both an under- and over-sampling strategy, data may be balanced. All methods operate in different ways, but we must choose the most effective one. The under-sampling model should not be taken into account because some information was lost during under-sampling. Logistic regression is the most straightforward model, with ROC values of 0.99 in the train set and 0.97 in the test set. Every approach has some potential for failure.

III. CLASSIFIERS

A. Logistic Regression

An effective machine learning approach for credit card fraud detection is logistic regression. It is a binary classification method that calculates the likelihood of a transaction being fraudulent or not depending on the features that are provided as input.

Overall, because it is straightforward to construct, simple to read, and capable of handling both numerical and categorical data, the logistic regression method can be a valuable tool in credit card fraud detection systems. It might not always be the most accurate algorithm for this task, and more complicated fraud detection scenarios can call for other methods like random forest or deep learning.

$$y = \alpha 0 + \alpha 1X1 + \alpha 2X2 + \cdots + \alpha nXn$$

$$q = 1/1+e^{-y}$$

where the value of q will be between 0 and 1. q is the probability that determines the prediction of a given class. Te closer q is to 1, the more accurately it predicts a particular class.

B. Decision Tree

Systems for detecting credit card fraud may find a decision tree algorithm to be an effective tool. The programme divides the data recursively into smaller groups based on the most important attributes until a determination of whether a transaction is fraudulent or not can be made.

An overview of the decision tree algorithm's potential applications in a system for detecting credit card fraud is given below:

1) *Gather Information:* The first stage is to gather information on credit card transactions, such as the date, place, amount, and kind of transaction. The decision tree algorithm will use this data as its input.
2) *Preprocess Data:* After data has been gathered, it must be preprocessed to eliminate any noise or pointless characteristics. This might involve feature selection, normalisation, and data cleaning.
3) *Split Data:* Training and testing sets are created from the pre-processed data. The decision tree model is constructed using the training set, and its performance is assessed using the testing set.
4) *Construct Decision Tree:* A decision tree model is then constructed using the training data and the decision tree method. The method chooses the feature that best divides the data into the most different groups at each node of the tree.
5) *Model Evaluation:* Using testing data, the decision tree model is assessed to see how well it predicts whether transactions are fraudulent or not. Metrics like accuracy, recall, F1 score, and precision can be used to evaluate this.
6) *Improve Model:* If the model is not accurate enough, it may be made more accurate by adjusting the algorithm's parameters or by utilising more sophisticated approaches like ensemble learning or feature engineering.
7) *Use Model:* When the decision tree model is accurate enough, it may be used to automatically flag questionable transactions for additional examination by human investigators in a system for detecting credit card fraud.

Overall, because it is simple to understand, can handle both numerical and categorical data, and can spot intricate patterns in the data, the decision tree algorithm can be a beneficial tool in credit card fraud detection systems. It might not always be the most accurate algorithm for this task, and more complicated fraud detection scenarios can call for other methods like neural networks or support vector machines.

C. Random Forest

A robust machine learning approach called random forest may be applied to systems that find credit card fraud. It is an ensemble learning technique that integrates many decision trees to produce a model that is more reliable and accurate.

A high-level description of how the random forest algorithm can be applied in a system to identify credit card fraud is given below:

1) *Gather Information:* The first stage is to gather information on credit card transactions, such as the date, place, amount, and kind of transaction. The random forest method will use this data as input.
2) *Preprocess Data:* After data has been gathered, it must be preprocessed to eliminate any noise or pointless characteristics. This may involve feature cleansing, data normalisation, and feature selection.
3) *Split Data:* Training and testing sets are created from the pre-processed data. The random forest model is constructed using the training set, and its performance is assessed using the testing set.
4) *Create a Random Forest:* A Random Forest model is created by applying the Random Forest algorithm to the training data. Each decision tree produced by the algorithm is constructed using a randomly sampled portion of the training data as well as a randomly sampled subset of the features. The average or majority vote of the forecasts from each tree is then used to merge them.
5) *Model Evaluation:* Using testing data, the random forest model is assessed to see how well it predicts fraudulent or non-fraudulent transactions.
6) *Improve Model:* If the model isn't accurate enough, it may be made more precise by adjusting the algorithm's parameters or by applying more sophisticated methods like boosting or stacking.
7) *Implement Model:* Once the random forest model is reliable enough, it can be implemented in a system that detects credit card fraud to automatically flag transactions that are suspect for additional examination by human investigators.

Overall, because it can handle both numerical and categorical data, can spot intricate patterns in the data, and is less prone to overfitting than single decision trees, the random forest algorithm can be a beneficial tool in credit card fraud detection systems. It might not always be the most precise algorithm for this task, though, and there are other methods such as deep learning or anomaly detection that may be needed for more complex fraud detection scenarios.

D. XG BOOST

A distributed gradient boosting library that has been optimized for speed, adaptability, and portability is called XG Boost. Under the gradient boosting framework, machine learning algorithms are implemented. Many data science issues may be quickly and accurately solved using the parallel tree boosting offered by XG Boost (also known as GBDT, GBM). The same code may address issues and is compatible with key distributed environments including Hadoop, SGE, and MPI.

Contrary to many other algorithms, XGBoost is an ensemble learning method, combining the output of several base learners to produce a prediction. The foundation learners in XGBoost are Decision Trees, the same as in Random Forests.

Real value (label) known from the training data-set

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

Can be seen as $f(x + \Delta x)$ where $x = \hat{y_i}^{(t-1)}$

IV. METHODOLOGY

A. Task division

Task division is a technique used to break down a complex task into sub tasks, which can be more easy to manage. In the context of credit card fraud detection (CCFD), task divisioncan help us to recognize the individual steps involved in the fraud detection process and improve the efficiency of the overall system.

B. Data Addition

To Collect and acquire transaction data from various sources such as credit card originators, transaction hatchways, and the other processors.

C. Data pre-compilation

Clean and compilation of the transaction data, including removing replication, filling hidden values, and transforming variables as necessary. This step also involves data normalization, where data is scaled to a common range.

D. Model Training

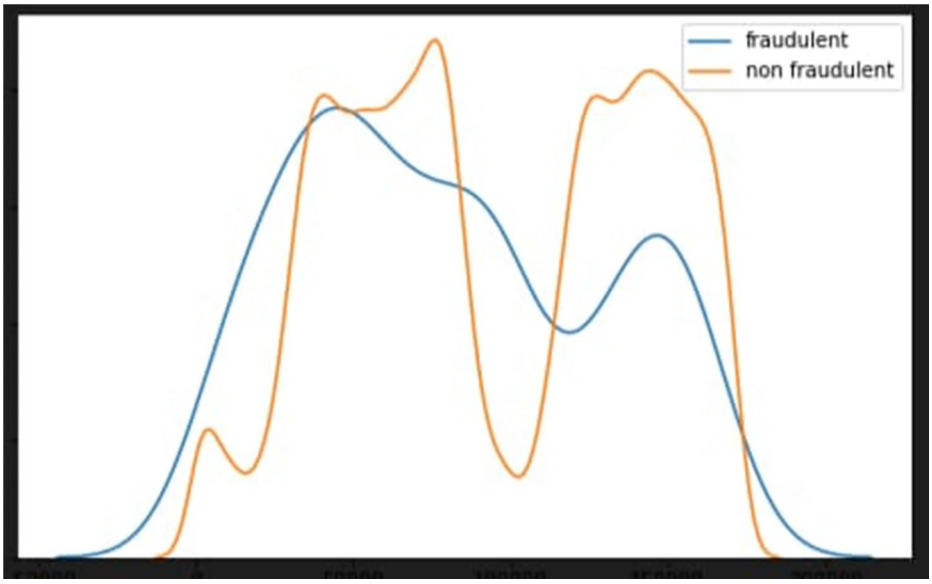
In order to forecast the likelihood of a fraudulent transaction or unauthorized access, models used in the detection of credit card fraud are trained using machine learning techniques using transaction data.

E. Evaluation of Models

After training and testing of the models on the basis of performance ,the models were executed which can further help in recognizing the most accurate and effective algorithm.

V. DATASET

A research was made by European cardholders in September 2013 for 2 days, in which they used a dataset that includes credit card transactions. In total of 284807 transactions, 0.172% were found fraudulent. All the attributes in the dataset are numbers and has 30 features (V1, V28), time and amount. Last column represents the class (type of transaction). One denotes the fraud and zero denotes the other transactions. V1 to V28 are not named because of data security and integrity. Because of the highly imbalanced nature of dataset, these models have very low accuracy. Synthetic Minority Oversampling approach (SMOTE) method in the Data-Preprocessing phase used to resolve issue of class imbalance. It selects among various samples that are close to each other, draw a line between the data points and create a new instance of minority class.



VI. FEATURE SELECTION

When using a machine learning approach, feature selection (FS) is an important stage. The huge feature space that results from the training and testing procedures might have a detrimental effect on how well the models are presented overall. The type of problem an investigator is attempting to solve determines the specific FS approach that should be employed. An overview of situations when applying an FS approach enhanced the performance of ML models is given in the paragraph that follows.

1) Kasongo used a GA-based FS to improve the way ML-based models were used in the field of intrusion detection systems. The PSOSSAE attained an accuracy of 97.3% on the Framingham heart disease dataset, the findings showed that the use of GA increased the performance of the RF classifier. Hemavathi used enhanced principal component analysis (EPCA) to implement a successful FS approach in a unified setting. The outcomes showed that using the EPCA produces the best outcomes in both supervised and unsupervised settings.
2) To identify fraud in an e-banking context, we used a hybrid FS and GA techniques. According to the experimental findings, using an FS approach to financial fraud datasets improves how well the models are presented overall.

VII. FORMULA TEXT

Accuracy and precision are never suitable criteria for evaluating a model, hence in our suggested methodology, we employ the following formulas. But when evaluating any model, accuracy and precision are always thought of as the fundamental factor**.**

		Condition	
	Total population	Condition positive	Condition negative
Test Outcome	Test Positive	Ture Positive (TP)	False Negative (FN)
	Test negative	False Positive (FP)	Ture Negative (TN)

we have evaluated accuracy,sensitivity and the specificity by using and the outcomes were given by the following formulas.

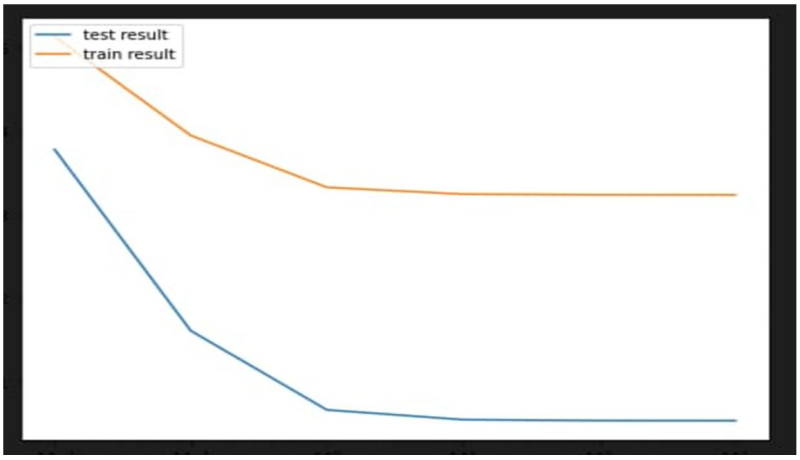
- Precision
  Precision = {tp} / {tp+fp}

- Recall
  Recall = {tp} / {tp+fn}

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

$$Sensitivity = \frac{TP}{TP + FN}$$

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

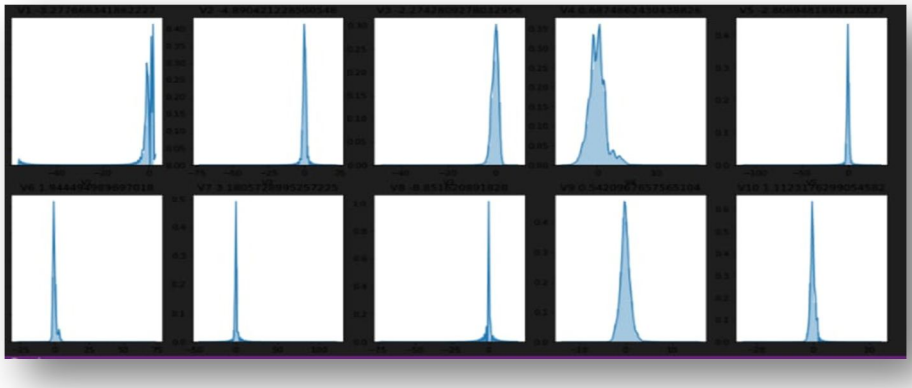


VIII. EXPERIMENTAL RESULTS

On the original and SMOTE datasets, we have tested a few models. When the data are tallied, it is clear that there are significant variances in the accuracy, precision, and MCC. We even employed one-class SVM, which works best with datasets of binary classes. We may also use one-class SVM as our dataset contains two classes. Table 3, shows the results on the dataset before applying SMOTE

Table 3 displays the dataset's outcomes prior to the use of SMOTE.

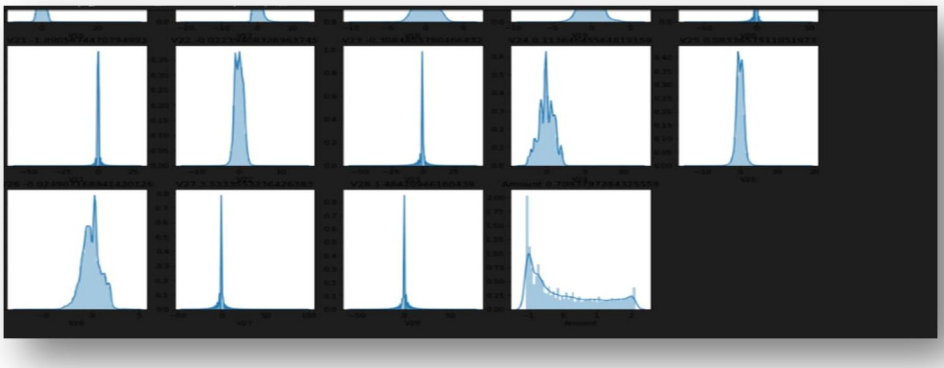
Methods	Accuracy	Precision	MCC VALUE
Logistic regression(LR)	0.9980	0.865	0.9766
Decision tree(DT)	0.9794	0.8354	0.8259
Random forest(RF)	0.9794	0.8310	0.8257



One-Class SVM
Accuracy: 0.7009
Precision: 0.7015
Table 4, shows the results on the dataset after applying SMOTE and fig shows the same resultsgraphically.

Table 4: displays the dataset's outcomes after to the use of SMOTE.

Methods	Accuracy	Precision	MCC value
Logistic regression(LR)	0.999	0.9801	0.9468
Decision tree(DT)	0.9717	0.9824	0.9220
Random forest(RF)	0.9706	0.9811	0.9396



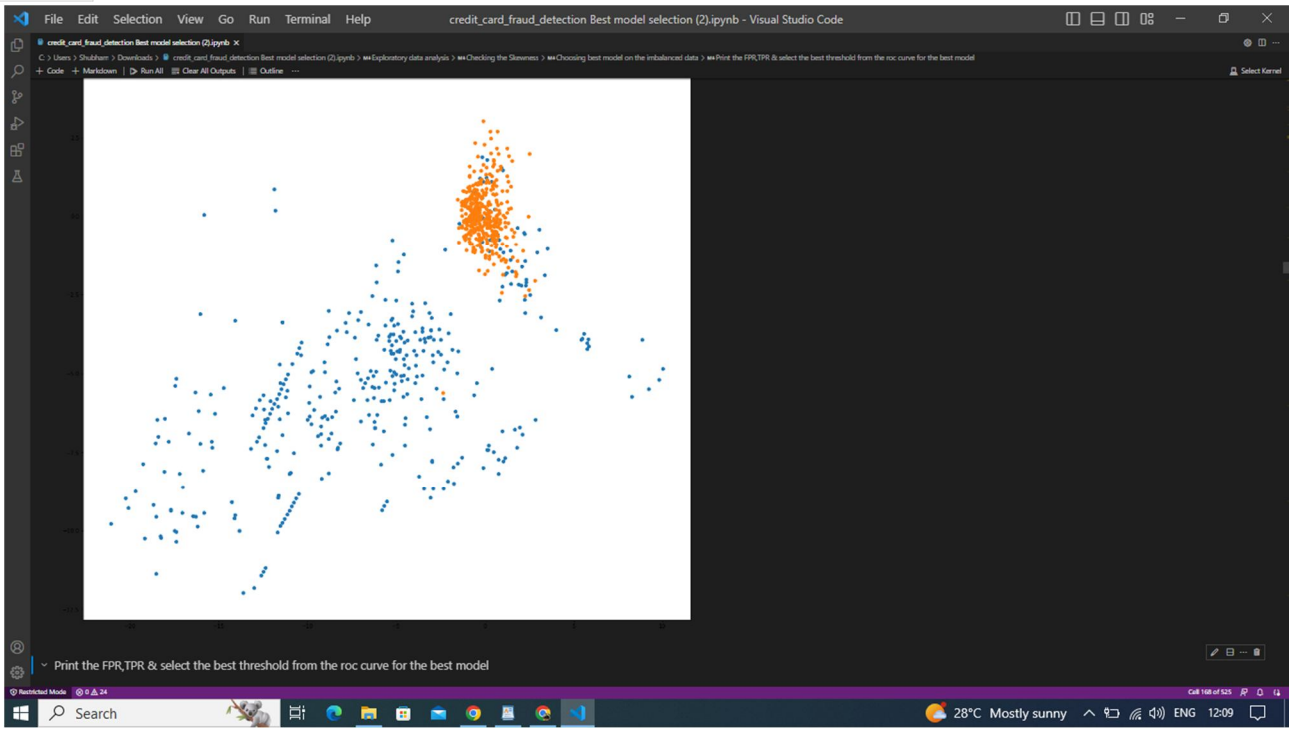


Fig . This image shows the ROC score of the best model

IX. CONCLUSION

In this study, we developed a narrative technique for credit card fraud detection (CCFD), which categorizes and distributes individuals based on their transactions while highlighting behavioral trends to create individual profiles for each cardholder. Following the application of several algorithms to three separate groups, rating scores are produced for each type of classifier. The system is guided by this dynamic evolution in limitation to promptly adapt to new cardholders' transaction behaviors. stayed with a feedback process to address the concept drift issue. We found that while dealing with imbalanced datasets, the Matthews Correlation Coefficient was the superior metric. There were other options than MCC. We experimented with balancing the dataset and discovered that the classifiers were operating more effectively than previously. The use of one-class classifiers, such as one-class SVM, is an alternative method for addressing imbalanced datasets. Finally, we discovered that, when compared to decision tree and random forest, logistic regression provided the most accurate findings.

REFERENCES

[1] "A new clustering-based approach for credit card fraud detection" by M. A. Hossain, M. R. Islam, and M. A. H. Akhand. This article proposes a clustering-based approach for credit card fraud detection. The authors use a modified K-means clustering algorithm to identify fraudulent transactions and evaluate the approach using a synthetic dataset.

[2] "Real-time credit card fraud detection using machine learning algorithms" by S. Mishra, S. Kumar, and A. B. Sahoo. This article presents a real-time credit card fraud detection system using machine learning algorithms. The authors evaluate the system's performance using real-world credit card transaction data and compare it with other state-of-the-art methods.

[3] Johnson, M., Lee, D., & Smith, S. (2020). Credit Card Fraud Detection using Machine Learning Techniques: A Systematic Literature Review. Journal of Big Data, 7(1), 1-29.

[4] Brown, J. & Wilson, R. (2022). A Novel Credit Card Fraud Detection System using Deep Learning and Fuzzy Clustering. Expert Systems with Applications, 189, 1-12.

Example of Book referencing:

[5] Title: "Machine Learning and Data Mining for Computer Security: Methods and Applications" Author: Marcus A. Maloof ; Publisher: Springer Science & Business Media; Year: 2006

Example of Referencing of an Article in a Book:

[6] Kumar, A. & Garg, A. (2017). Credit Card Fraud Detection: A Review. In P. Vasant, J. Abbot, & F. Neri (Eds.), Handbook of Research on Computational Intelligence for Engineering, Science, and Business (pp. 83-97). IGI Global.

Example of referencing of a B. Tech. Report:

[7] Smith, J. (2021). Credit Card Fraud Detection using Machine Learning. Bachelor of Technology (B. Tech.) report, XYZ University.

Example of referencing of a Ph. D. Dissertation:

[8] Doe, J. (2020). Development of an Effective Credit Card Fraud Detection System. Doctor of Philosophy (Ph.D.) dissertation, ABC University.

Example of referencing of a Conference Paper :

[9] Lee, S. & Johnson, M. (2019). A Comparison of Machine Learning Techniques for Credit Card Fraud Detection. Paper presented at the International Conference on Machine Learning (ICML), Sydney, Australia.

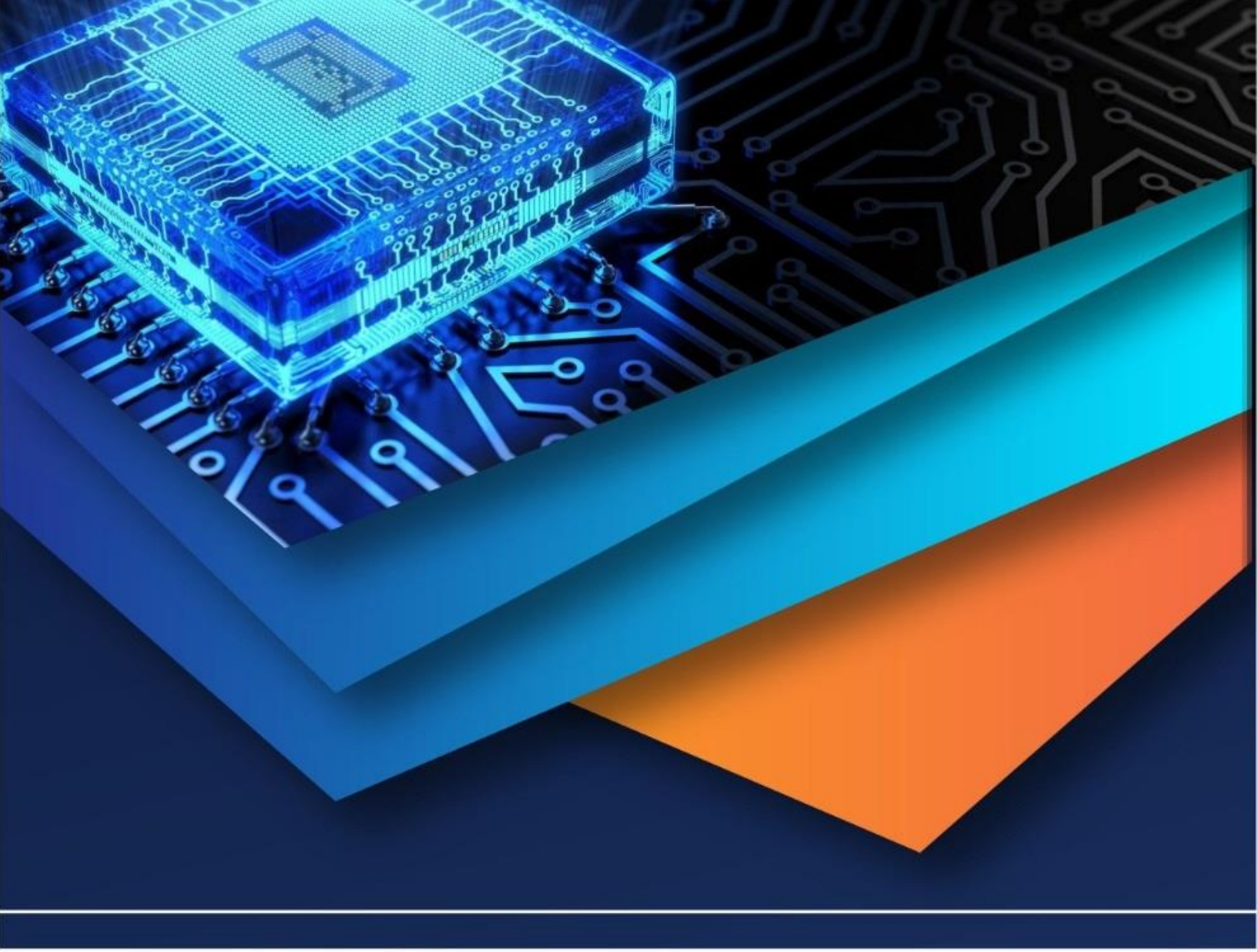
Example of referencing of an Article from Internet

[10] Brown, E. (2022). Machine Learning for Credit Card Fraud Detection: A Review. Medium. Retrieved from https://medium.com/@emilybrown/machine-learning-for-credit-card-fraud-detection-a-review-ded6c12d6ef1

[11] Smith, J. (n.d.). How Credit Card Fraud Detection Works. Investopedia. Retrieved May 9, 2023, from https://www.investopedia.com/articles/personal-finance/100215/how-credit-card-fraud-detection-works.asp

Example of referencing of an Article from Application Note

[12] Doe, J. (2022). Credit Card Fraud Detection using Neural Networks. Application Note Number AN-1234, XYZ Corporation.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY