Credit Card Fraud Detection and Analysis

Rohan Pawar
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
rohan.pawar2020@vitstudent.ac.in

Hardik Kathuria
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
hardik.kathuria2020@vitstudent.ac.in

Praveen Joe I R
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India
praveen.joe@vit.ac.in

Abstract— **In today's digital world, cybersecurity plays a crucial part. Whenever someone talks about security in the digital world, the main task is to recognize abnormal activity. Whenever someone makes any transaction online, a high percentage of people prefer credit cards. The credit limits we set helps the user make purchases even if they can't afford the product at the current moment. These features can be misused by cyber attackers.**

It is necessary for credit card providers to identify fraudulent transactions so that customers are not charged unnecessarily. This paper will provide an approach using machine learning algorithms to detect anomalous transactions.

Keywords—Indispensable, Transaction Fraud Detection, Machine Learning, Ensemble model, Precision, Business Intelligence, Analysis, Logistic Regression, Decision Tree, Random Forest

I. INTRODUCTION

'Fraud' in credit card transactions is unauthorised and unwanted usage of an account by someone other than the owner of that account. Necessary prevention measures can be taken to stop this abuse and the behaviour of such fraudulent practises can be studied to minimise it and protect against similar occurrences in the future.

In other words, Credit Card Fraud can be defined as a case where a person uses someone else's credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used. Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting. Quite simply, the fraud detection system can be depicted in the following manner-

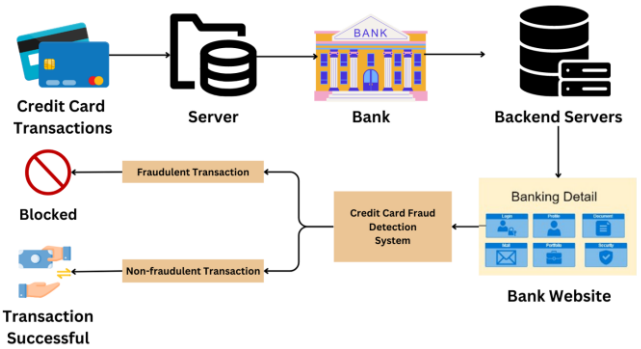

Figure 1 – Credit Card Fraud Detection System Architecture

Credit card fraud costs consumers and the financial company billions of dollars annually, and fraudsters continuously try to find new rules and tactics to commit illegal actions. Thus, fraud detection systems have become essential for banks and financial institution, to minimize their losses.

II. BACKGROUND STUDY

A. Literature Survey

We found various approaches to our problem and with different models. Paper [1] discusses the importance of fraud detection and prevention in the financial industry and how machine learning can help with this task. Isolation forest algorithm and local outlier factor methods from sklearn were used for detection of abnormalities in the dataset which was imported from Kaggle. The authors used ensemble methods to create a model with high accuracy.

The paper also discusses the evaluation metrics used to measure the performance of these models, such as accuracy, precision, recall, and F1-score. In the final result, the proposed model achieved an accuracy of 99.7% with Random Forest Algorithm.

In paper [2], the author talks about the challenges faced in credit card fraud analysis, i.e., data skewness and changing trends in fraudulent behaviours. The model proposed in the paper is a hybrid technique of under-sampling and oversampling carried out on the skewed data. 3 algorithms are used here, namely Naïve Bayes, K-Nearest Neighbors and Logistic regression. Out of the three algorithms, K-Nearest Neighbors performed the best with an accuracy of 98%.

Paper [3] is about using machine learning techniques to prevent credit card fraud. The paper proposes an approach using several machine learning algorithms, including random forest, support vector machines, and k-nearest neighbors. These algorithms take in data from past credit card transactions, including information on the cardholder, the transaction amount, and the merchant, to identify patterns and predict the likelihood of fraud. The model proposed in this paper is an ensemble model combining multiple ML algorithms. Finally, the paper evaluates the performance of these machine learning algorithms on a real-world dataset of credit card transactions, achieving an accuracy of over 99%.

Paper [4] authors use various machine learning techniques such as Decision Trees, Support Vector Machine , Random Forest and Logistic regression to detect fraudulent transactions. They have stated that a fraud detection model should have 3 main properties -

- **Accuracy-** It should perform accurate detection of fraudulent transactions.
- **Real Time Fraud Detection-** It should be able to identify the fraud while it is happening.
- **Low False Positive Rate-** It should not identify real transactions as fraud.

The evaluation matrix used consists of Accuracy, Precision and False Positive Rate. Finally, the results show that the Random Forest Model has the highest accuracy of 99.21%.

In Paper [5], the authors have provided an useful insight about the bagging method using decision trees. Bagging is an ensemble technique that combines predictions of all models from randomly generated testing datasets. Here, the author has called a single decision tree as a "weak learner" whereas the whole model as a "strong learner". Each decision tree in the model gives a vote or basically has a small influence on the final output. The final output prediction is based on the instance that gained the maximum number of votes received from the decision trees.

Finally, the output performance matrix consists of 2 parameters, Fraud Catching Rate and False Alarm Rate. The model with high fraud catching rate and low false alarm rate is considered the best.

B. *About the dataset*

The dataset used to train and test the model contains the real bank transactions made by European cardholders on 1 business day in the year 2013. As a security concern, the actual variables are not being shared but — the data has been transformed versions of PCA. Only 0.17% (492 out of 284,807) transactions are fraudulent. The following graph shows the normal transactions to the fraudulent ones in the form of a pie chart-

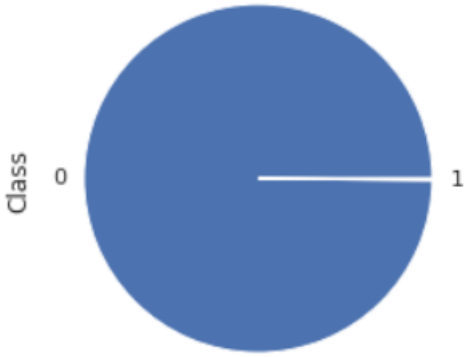

Figure 2 – Piechart representing credit card transactions

III. Proposed Methodology

In this paper, we aim to use various machine learning algorithms to solve the prevalent problems we face in credit card fraud detection. We try to find the general accuracy, F1 score, precision, recall etc by deriving the confusion matrix for algorithms like logistic regression, decision trees, random forest, KNN and Naïve-bayes. Once this is done we try to address the class imbalancement issue that takes place when we split the data. This happens since there is very small amount of fraud transactions when compared to genuine transactions, so when we split the data there is a chance that the data will not be split equally for the training and testing values. So by applying the SMOTE technique we once again train the data for algorithms like logistic regression, random forest and decision trees. The SMOTE technique (Synthetic Minority Oversampling Technique) creates a balance between the target and normal variables in the dataset by using sampling methods where synthetic data is generated based on the already available data by applying few conditions to it.

Once this data is generated, the class imbalancement issue is resolved and there is a neutrality between the number of genuine and fraud transaction data available. This allows us to train our different models for the derived dataset with better insight into the values since the data is now balanced. We find the evaluation metrics for these new algorithms tried with the newly generated dataset and then compare and contrast our results with the scores already derived from previous experiments.

A. *Abbreviations and Acronyms*

SMOTE: SMOTE (Synthetic Minority Oversampling Technique)first selects a minority class instance a at random and finds its k nearest class neighbours. The synthetic instance is then created by choosing a k nearest neighbours b at random and connecting a and b to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances a and b.

PCA: PCA (Principal Component Analysis) is a dimensionality reduction method in machine learning for large size of datasets to make them interpretable whilst preserving maximum information.

B. *Equations*

Accuracy is the percentage of correctly predicted results.

$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN)$$

Recall is the ratio of true positives to all the correctly predicted results.

$$\text{Recall} = TP/(TP+FN)$$

Precision is the ratio of true positives to the total positive observations.

$$\text{Precision} = TP/(TP+FP)$$

Here TP, TN, FP, FN stands for-

TP-True Positive,

TN- True Negative,

FP- False Positive,

FN-False Negative.

F1-score is the harmonic mean of precision, P and recall, R

$$\text{F1 score} = (2*R*P)/(R+P)$$

IV. Results and Analysis

Using the Random Forest Model, we were able to accurately identify the credit card fraud transactions. From the dataset, we identified 5 attributes with the highest correlation to our problem, namely V17, V14, V10, V12, and V11. The preprocessing performed on the dataset are as follows-

1. Split the data using a random, stratified train/test split with a test size of 20%.
2. Box-Cox power transform of the transaction amounts to remove skewness in the data.

3. Mean and variance standardization of all features as part of a machine learning pipeline.

We have used the Matthews correlation coefficient (MCC) to compare the performance of different models. During cross validation, logistic regression achieved a cross-validated MCC score of 0.807, and the random forest model achieved a cross-validated MCC score of 0.856. We therefore chose the random forest as the best model, which obtained an MCC of 0.869 on the testing dataset.

A. Graphs

Number of transactions to Time Bar Graph - Here, we are plotting to number of transaction with respect to time(range 2 days).

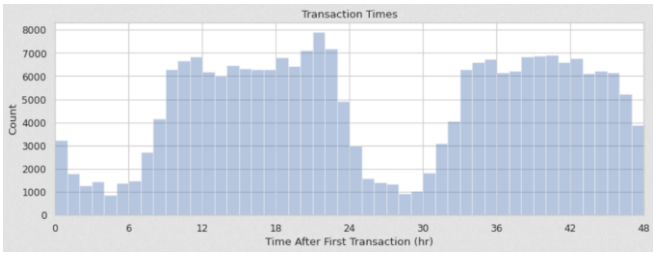

Figure 3 -

Correlation Density Graph - We plotted legal to fraudulent data in a given variable and the highest correlation was found in the variables V10, V11, V12, V14 and V17.

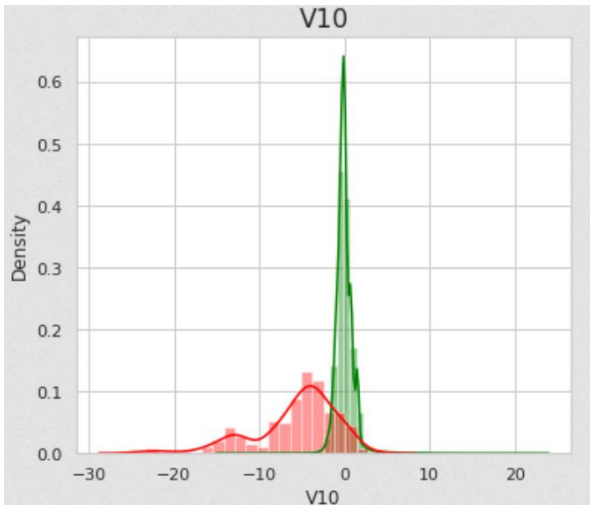

Figure 4 – Correlation Density Graph for V10

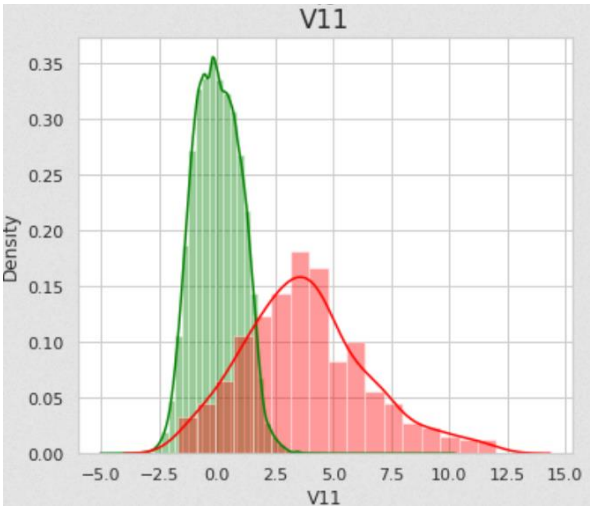

Figure 5 – Correlation Density Graph for V11

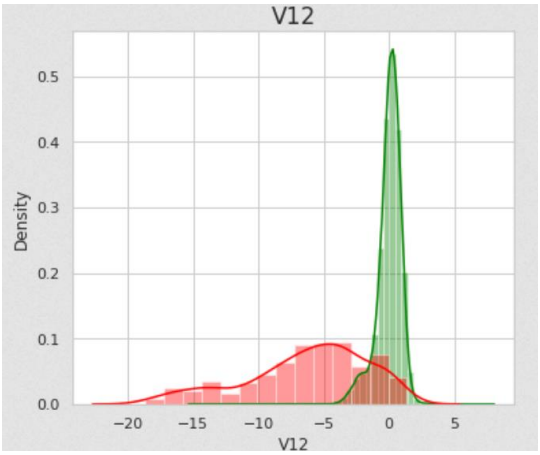

Figure 6 – Correlation Density Graph for V12

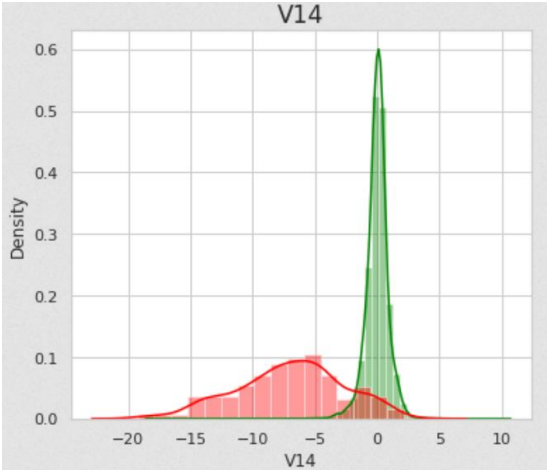

Figure 7 – Correlation Density Graph for V14

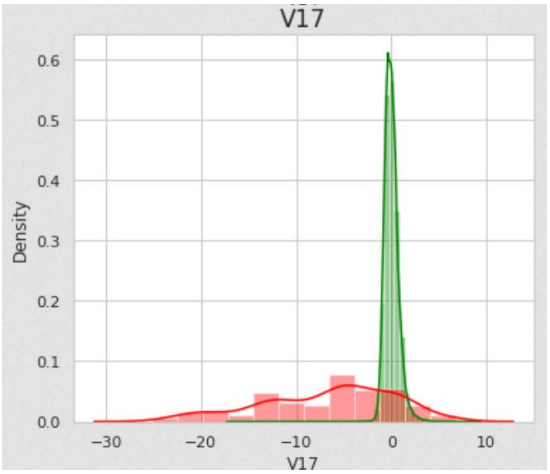

Figure 8 – Correlation Density Graph for V17

Heatmap – A heatmap is a visualisation method which shows us the magnitude of the phenomenon using colour coding. Here, the following heatmap depicts the transactional data variables and the extreme data points shown in deeper colours (any colour other than light blue, which is the normal value), are the abnormalities or the possible fraudulent transactions.

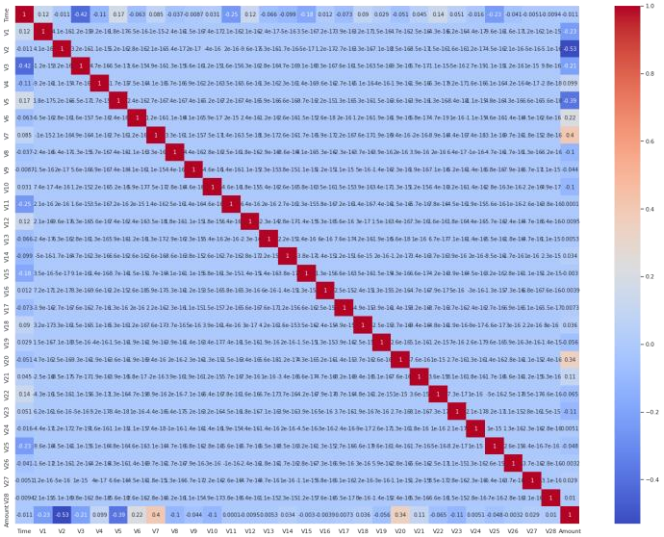

Figure 9 -Heat Map

Boxplot- The boxplot here shows the outliers in transactional data, outliers can sometimes be used as a helpful indicator. For example, in data analytics like credit card fraud detection, outlier analysis becomes important because the exceptions can present useful insight to the analyst.

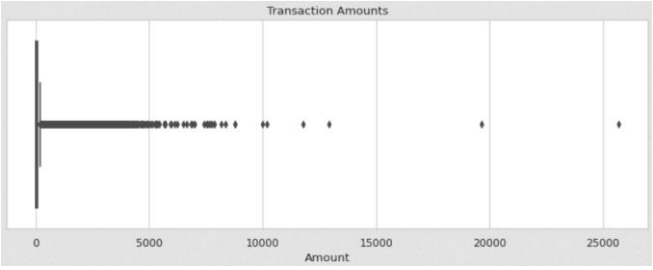

Figure 10 – Boxplot showing outliers

B. Evaluation Matrices

For Imbalanced Dataset: For imbalanced dataset, the following accuracy was obtained-

Classifiers	Accuracy	Precision	Recall	F1-score
Logistic Regression	1.00	0.91	0.71	0.79
Decision Tree	1.00	0.71	0.82	0.76
Random Forest	1.00	0.94	0.86	0.90

Table 1 – Evaluation Matrix for imbalanced dataset

Here, random forest has come out on top with an F1-Score of 0.90, followed by logistic regression with and F1 Score of 0.79 and decision tree with 0.76.

After applying SMOTE method, the following results were obtained-

Classifiers	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.98	0.06	0.92	0.12
Decision Tree	1.00	0.39	0.78	0.52
Random Forest	1.00	0.89	0.84	0.86

Table 2 – Evaluation Metrix after SMOTE

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

C. Conclusion

Fraudulent and true credit card transactions are classified using Naïve Bayes, KNN, Logistic Regression, Decision Tree and Random Forest classifiers. The models are trained and testes on an imbalanced dataset and a balanced dataset using SMOTE method. Logistic Regression shoed the least performance of 0.12 F1-score in balanced dataset, and only performs better on unbalanced dataset with 0.79 F1-score. Finally, the ensemble method Random Forest shows superior performance with the best F1 score in predicting fraudulent transactions of 0.90 on imbalanced dataset and 0.86 on balanced dataset, showing the power of ensemble techniques. With further studies, these machine learning algorithms can be deployed in the real world and prevent fraudulent transactions automatically.

REFERENCES

[1] S P Maniraj, Aditya Saini, Swarna Deep Sarkar Shadab Ahmed, “Credit Card Fraud Detection using Machine Learning and Data Science” ISSN: 2278-0181Vol. 8 Issue 09, September-2019

[2] John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare, “Credit card fraud detection using machine learning techniques: A comparative analysis”, 2017 International Conference on Computing Networking and Informatics (ICCNI), DOI:10.1109/ICCNI.2017.8123782

[3] Shakya, Ronish, "Application of Machine Learning Techniques in Credit Card Fraud Detection" (2018).UNLV Theses, Dissertations, Professional Papers, and Capstones

[4] Sadineni, Praveen Kumar. “Detection of Fraudulent Transactions in Credit Card using Machine Learning Algorithms.” 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) (2020): 659-660.

[5] Masoumeh Zareapoor, Pourya Shamsolmoali, “Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier”, Procedia Computer Science 48 (2015) 679-685

[6] G. Srilakshmi & I. R. Praveen Joe (2023) “A-DQRBRL: attention based deep Q reinforcement battle royale learning model for sports video classification”, The Imaging Science Journal, DOI: 10.1080/13682199.2023.2180022

[7] Praveen Joe I R, E M Malathy, S Aishwarya, R Akila, A Akshaya, “A Hybrid PSO-ACO Algorithm to Facilitate Software Project Scheduling”, International Journal of e-Collaboration (IJeC), 2022, DOI: 10.4018/IJeC.304039

[8] Malathy, E. M., I. R. Praveen Joe, and P. Ajitha. "Miniaturized Dual-Band Metamaterial-Loaded Antenna for Heterogeneous Vehicular Communication Networks." IETE Journal of Research (2021): 1-10.

[9] Joe, IR Praveen, and P. Varalakshmi. "A Two Phase Approach for Efficient Clustering of Web Services." Computational Intelligence, Cyber Security and Computational Models. Springer, Singapore, 2016. 165-170.

[10] I. R. Praveen Joe & P. Varalakshmi (2019) A Multilayered Clustering Framework to build a Service Portfolio using Swarm-based algorithms, Automatika, 60:3, 294-304, DOI: 10.1080/00051144.2019.1590951

[11] Praveen Joe, I.R. and Varalakshmi, P. (2019) "An Analysis on Web-Service-Generated Data to Facilitate Service Retrieval," Applied Mathematics & Information Sciences: Vol. 13: Issue 1

[12] Moorthy Agoramoorthy, Irudayaraj Praveen Joe, “Hybrid cuckoo–red deer algorithm for multiobjective localization strategy in wireless sensor network”, International Journal of Communication Systems, https://doi.org/10.1002/dac.5042, December 2021. pp:47-55

[13] Joe, Praveen. "IR and Varalakshmi, P,,,A Survey on Neural Network Models for Data Analysis‟." ARPN Journal of Engineering and Applied Sciences 10.11 (2015): 4872-4876.

[14] V. Dhanasekar, Y. Preethi, V. S, P. J. I. R and B. P. M, "A Chatbot to promote Students Mental Health through Emotion Recognition," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 1412-1416, doi: 10.1109/ICIRCA51532.2021.9544838.