Credit Card Fraud Detection

Meghna
ECE(AI) Department
Indira Gandhi Delhi Technical University for Women
Delhi, India
meghnas105@gmail.com

Khushi Arora
Mechanical Department
Indira Gandhi Delhi Technical University for Women
Delhi, India
Karora.work10@gmail.com

Nishtha Pabreja
CSE Department
Manipal University
Jaipur, India
nishtha.229301016@muj.manipal.edu

***Abstract*—Credit card fraud poses a substantial and ever-evolving threat to financial institutions and consumers alike. This research paper investigates the application of advanced machine learning techniques for the purpose of credit card fraud detection. Leveraging a comprehensive dataset encompassing legitimate and fraudulent transactions, this study delves into various facets of fraud detection, including data preprocessing, feature engineering, model selection, and ethical considerations. This research includes the exploration of methodologies to address class imbalance within the data, allowing for the development of robust and effective models for fraud detection. An array of machine learning algorithms, including XG Boost, Random Forest and Decision Tree, are evaluated in the context of their ability to discern fraudulent activities. Furthermore, this paper examines the ethical dimensions emphasizing the importance of data privacy and regulatory compliance. In a landscape where financial security is of paramount concern, this research stands as a testament to the potential of machine learning in bolstering the defences against credit card fraud.**

***Index Terms*—credit card fraud, data preprocessing, machine learning, XG Boost, Random Forest, Decision Tree**

I. Introduction

The advent of digital payment systems has revolutionized the way we conduct financial transactions, offering unprecedented convenience to consumers and businesses worldwide. With a mere swipe, click, or tap, we can seamlessly complete purchases, pay bills, and transfer funds, ushering in a new era of financial convenience. However, in this digital transformation, a shadowy adversary looms large—credit card fraud.

Credit card fraud, encompassing a myriad of deceitful tactics such as unauthorized transactions, identity theft, card-not-present fraud, and account takeover, has emerged as a persistent and pernicious threat to financial institutions and cardholders alike. The ramifications of such fraudulent activities extend beyond monetary losses, permeating the very fabric of trust and security within the financial ecosystem. It is against this backdrop that the need for robust and efficient credit card fraud detection systems becomes increasingly pronounced.

This research focuses on the application of the following supervised ML algorithms for credit card fraud detection: Decision Tree (DT), Random Forest (RF) and XG Boost (XGB). Machine Learning(ML) systems are trained and tested using large datasets. In this work, a simulated credit card transaction dataset containing legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020 is utilised. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants. Often sometimes, these datasets may have many attributes that could have a negative impact on the performance of the classifers during the training process. To solve the issue of a high feature dimension space, we implement a feature selection algorithm. Thus, this research paper embarks on a comprehensive exploration of the field of credit card fraud detection, guided by the principles of machine learning. Our primary objective is to investigate, analyze, and innovate in the realm of fraud detection, ultimately contributing to the arsenal of tools available to combat this pervasive issue.

II. Literature Review

The landscape of credit card fraud detection has evolved significantly over the years, driven by the increasing sophistication of fraudulent activities and advancements in machine learning. Traditional rule-based approaches, while effective to some extent, have limitations in adapting to changing fraud patterns. The emergence of machine learning-based techniques, including Logistic Regression, Random Forest and XG Boost has opened new avenues for more accurate and adaptive fraud detection. Also, feature engineering plays a pivotal role in improving model performance. This literature review underscores the dynamic nature of credit card fraud detection and sets the stage for our research, which aims to contribute to the ongoing quest for enhanced financial security and trust in digital transactions.

III. Data Collection

The dataset used for this Research is a simulated credit card transaction dataset containing legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants.

A. Dataset Source and Description

This dataset was generated using the Sparkov Data Generation, GitHub tool created by Brandon Harris. This simulation was run for the duration - 1 Jan 2019 to 31 Dec 2020. The files were combined and converted into a standard format. This simulator has a certain pre-defined list of merchants, customers and transaction categories and then using a Python

library called ”faker”, and with the number of customers and merchants that you mentioned during the simulation, an intermediate list is created.After this, depending on the profile you choose, the transactions are created. So, the transactions were generated across all profiles and then merged together to create a more realistic representation of simulated transactions. The dataset has two csv files fraudTrain and fraudTest on which analysis is performed.

TABLE I
DATASET COLUMN DESCRIPTION

Column Name	Description
index	unique identifier of each row
transdate transtime	Transaction Date Time
cc num	credit card number of customer
merchant	merchant name
category	category of merchant
amt	amount of transaction
first	first name of credit card holder
last	last name of credit card holder
gender	gender of credit card holder
street	street of credit card holder
city	city of credit card holder
state	state of credit card holder
zip	zip of credit card holder
lat	latitude location of credit card holder
long	longitude location of credit card holder
city pop	credit card holder’s city population
job	job of credit card holder
dob	dob of credit card holder
trans num	transaction number
unix time	unix time of transaction
merch lat	latitude location of merchant
is fraud	fraud flag target class
merch long	longitude location of merchant

IV. DATA PPREPROCESSING

Data preprocessing serves as the cornerstone of our research in credit card fraud detection. In this pivotal phase, we meticulously curate and refine our raw dataset to prepare it for the subsequent stages of analysis through a series of strategic steps like data preparation, data imbalance check, data cleaning and feature engineering

A. Data Imbalance Check

To assess the balance within our credit card transaction dataset, we conducted an exploratory data analysis using pie chart and countplot to visualize the distribution of genuine and fraudulent transactions. The pie chart provides a clear depiction of the class distribution in our dataset, revealing the extent of class imbalance between genuine and fraudulent transactions. The countplot visually represents the number of transactions belonging to each class— Genuine and Fraudulent. It is evident from the chart that 99.6 percent of transactions are Genuine, while only 0.43 percent corresponds to Fraudulent transactions. This class imbalance is a critical consideration when building and evaluating credit card fraud detection models, as it can impact the model’s ability to detect fraud effectively.

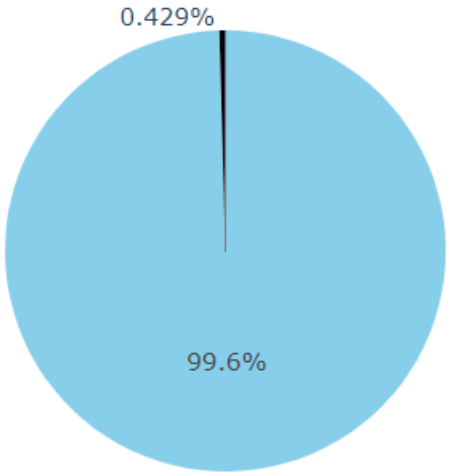

Fig. 1. Fraud vs Genuine Transactions Piechart

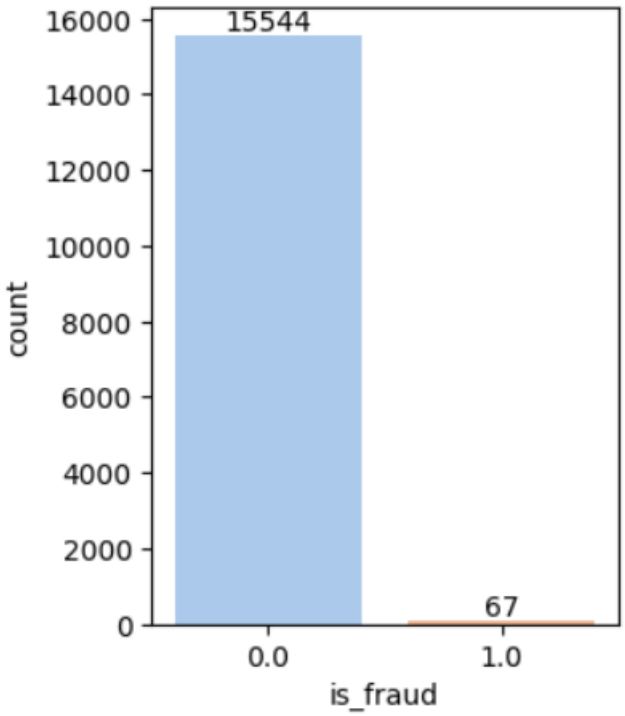

Fig. 2. Fraud vs Genuine Transactions Countplot

B. Data Cleaning

Data cleaning is a crucial step in preparing our dataset for credit card fraud detection analysis. To gain an initial understanding of the dataset, we performed a thorough examination using the info() function. This provided insights into data types, missing values, and column names for both the training and testing datasets. We identified and addressed missing values using the isnull().sum() function. This step is pivotal for ensuring the integrity of our dataset. Missing values, if left unattended, can lead to skewed results and compromised model performance. Certain columns were deemed irrelevant for predicting fraudulent transactions and were subsequently

dropped from the dataset. Columns such as 'Unnamed: 0', 'cc num', 'merchant', 'trans num', 'unix time', 'first', 'last', 'street', and 'zip' were removed for improved model efficiency and relevance. After data cleaning and column removal, we confirmed the dimensions of both the training and testing datasets using shape functions. This ensured that our dataset retained an appropriate size and structure for analysis.

C. Feature Engineering

Feature engineering plays a pivotal role in improving the dataset's richness and effectiveness in modeling credit card fraud. In this section, we elaborate on the feature engineering techniques applied to extract valuable insights from the raw data. We calculated the age of individuals involved in the transactions by subtracting their date of birth ('dob') from the transaction date ('trans date'). The resulting age feature provides a critical demographic dimension to our analysis.

Another critical feature is the calculation of the distance between the merchant and the cardholder's home location. We computed both the latitudinal and longitudinal distances, providing insights into the proximity of transactions and potentially identifying anomalies in geographic patterns. To streamline the dataset and remove noise, we identified and removed columns that were deemed irrelevant for predicting fraudulent transactions. These columns included 'trans date trans time', 'city', 'lat', 'long', 'job', 'dob', 'merch lat', 'merch long', 'trans date', and 'state.'

Then we performed the Gender Encoding in which we transformed the 'gender' column from categorical (e.g., 'M' for male and 'F' for female) into a binary numerical format (1 for male, 0 for female). This simplifies the representation of gender in our dataset. Then we did the One Hot Encoding for the 'category' column, which represents transaction categories, we employed one-hot encoding. This technique converts categorical data into a binary format, creating separate binary columns for each category. This transformation allows our models to understand and utilize category information effectively.

Then we handled the data imbalance recognizing the significant class imbalance between genuine and fraudulent transactions. For this, we employed the Synthetic Minority Oversampling Technique (SMOTE). This technique oversamples the minority class (fraudulent transactions) by generating synthetic examples.

V. MODEL BUILDING

The foundation of our credit card fraud detection system lies in the selection, training, and evaluation of an appropriate machine learning model. In this section, we detail the utilization of the Decision Tree Classifier, along with the steps involved in its application.

A. Decision Tree Classifier

We opted to employ the Decision Tree Classifier for its interpretability, simplicity, and ability to capture non-linear relationships in the data. The Decision Tree is particularly well-suited for fraud detection, as it can help uncover complex decision boundaries that may indicate fraudulent patterns. We initialized and trained the Decision Tree Classifier using the training data (X train and y train). To assess the model's performance, we made predictions on the testing data (X test) and generated a classification report. The classification report provides essential metrics such as precision, recall, F1-score, and support for both genuine and fraudulent transactions. In addition to the classification report, we calculated the accuracy of the Decision Tree model to provide a concise measure of its overall performance.

	precision	recall	f1-score	support
0.0	1.00	0.99	1.00	7775
1.0	0.00	0.00	0.00	22
accuracy			0.99	7797
macro avg	0.50	0.50	0.50	7797
weighted avg	0.99	0.99	0.99	7797

Accuracy: 0.9920482236757727

Fig. 3. Decision Tree Classifier Outcome

B. Random Forest Classifier

To ensure the robustness of our credit card fraud detection system, we explored the application of the Random Forest Classifier, a powerful ensemble learning algorithm known for its versatility and capacity to handle complex patterns in the data. Recognizing the need for a versatile and highly accurate model, we opted for the Random Forest Classifier. This ensemble learning method combines multiple decision trees to provide a powerful and stable prediction model. We configured the Random Forest Classifier with specific hyperparameters to optimize its performance. We trained the Random Forest Classifier on the training data (X train and y train) to harness the collective predictive power of multiple decision trees. To assess the model's efficacy, we made predictions on the test data (X test) and generated a comprehensive classification report. The classification report offers insights into precision, recall, F1-score, and support for both genuine and fraudulent transactions. The accuracy score was calculated to provide a concise measure of the model's overall correctness in classifying transactions. The outcome of the model is given below. The Random Forest Classifier, renowned for its robustness and ability to handle complex data, was effectively applied to the credit card fraud detection task. Evaluation metrics, as presented in the classification report, provide a comprehensive assessment of the model's precision, recall, F1-score, and support for both genuine and fraudulent transactions. The accuracy score serves as a succinct measure of the model's overall performance in detecting fraudulent activities.

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	7775
1.0	0.00	0.00	0.00	22
accuracy			1.00	7797
macro avg	0.50	0.50	0.50	7797
weighted avg	0.99	1.00	1.00	7797

Accuracy: 0.9965371296652559

Fig. 4. Random Forest Classifier Outcome

C. XG Boost

In our pursuit of a robust credit card fraud detection system, we explored the application of the XGBoost classifier, a powerful ensemble learning algorithm known for its exceptional predictive performance and scalability. Recognizing the complexity and non-linearity of fraudulent patterns, we chose the XGBoost classifier due to its ability to handle intricate relationships within the data, making it well-suited for the task of fraud detection. We configured the XGBoost classifier with several hyperparameters to optimize its performance. We trained the XGBoost classifier on the training data (X train and y train) to leverage its ensemble of decision trees for credit card fraud detection. To assess the model's effectiveness, we made predictions on the test data (X test) and generated a comprehensive classification report, offering insights into precision, recall, F1-score, and support for both genuine and fraudulent transactions. The accuracy score was calculated to provide a succinct measure of the model's overall correctness in classifying transactions.

	precision	recall	f1-score	support
0.0	1.00	0.99	0.99	7775
1.0	0.05	0.18	0.08	22
accuracy			0.99	7797
macro avg	0.53	0.59	0.54	7797
weighted avg	1.00	0.99	0.99	7797

Accuracy: 0.9884570988841862

Fig. 5. XG Boost Classifier Outcome

VI. CONCLUSION

In our pursuit of an effective credit card fraud detection model, we explored the performance of three different classifiers: Decision Tree, Random Forest, and XGBoost. These classifiers were evaluated on various metrics to determine their suitability for the task of detecting fraudulent transactions. While the Decision Tree Classifier is known for its interpretability, the XGBoost Classifier, known for its predictive power and the Random Forest Classifier is known for its robustness. Based on the evaluation metrices and key outcomes, it turns out that the **Random Forest Classifier** outperformed the others in terms of accuracy and F1-score, indicating its superior ability to correctly classify both genuine and fraudulent transactions. With an accuracy score of 0.996, the Random Forest Classifier is the preferred model. It is a combination of robust data preprocessing, feature engineering, and an effective machine learning model.

REFERENCES

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

[2] Vaishnavi Nath Dornadula, S Geetha, Credit Card Fraud Detection using Machine Learning Algorithms,Procedia Computer Science, Volume 165, 2019, Pages 631-641, ISSN 1877-0509

[3] Ileberi, E., Sun, Y. and Wang, Z. A machine learning based credit card fraud detection using the GA algorithm for feature selection. J Big Data 9, 24 (2022).

[4] Survey Paper on Credit Card Fraud Detection by Suman" , Research Scholar, GJUST Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering and Technology (IJARCET) Volume 3 Issue 3, March 2014

[5] Mohammed, Emad, and Behrouz Far. "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study." IEEE Annals of the History of Computing, IEEE, 1 July 2018, doi.ieeecomputersociety.org/10.1109/IRI.2018.00025.

[6] 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.

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

[9] Melo-Acosta, German E., et al. "Fraud Detection in Big Data Using Supervised and Semi-Supervised Learning Techniques." 2017 IEEE Colombian Conference on Communications and Computing (COLCOM), 2017, doi:10.1109/colcomcon.2017.8088206.

[10] Guo S, Liu Y, Chen R, Sun X, Wang X. X, Improved SMOTE algorithm to deal with imbalanced activity classes in smart homes. Neural Process Lett. 2019;50(2):1503–26.

[11] https://www.kaggle.com/datasets/kartik2112/fraud-detection?select=fraudTrain.csv

[12] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

[13] https://shorturl.at/irzUX **(dataset)**

[14] https://shorturl.at/twBO7 **(dataset)**