Fraud Transaction Detection System

SUSHANT AGRAWAL
Department of Information Technology

Abstract- ***To prevent customers from being charged for unauthorized purchases, it is crucial for credit card issuers to be able to identify fraudulent transactions. Data science, in conjunction with machine learning, plays a significant role in addressing this issue. This study focuses on utilizing machine learning to model a dataset for credit card fraud detection. The approach involves analyzing past credit card transactions, particularly those that were later identified as fraudulent, in order to assess the legitimacy of new transactions. The objective is to minimize false categorizations of fraud while accurately identifying all instances of fraudulent activity. One prominent example of categorization is the detection of credit card fraud. This approach involves analyzing and preprocessing datasets, as well as employing various anomaly detection techniques on PCA-transformed credit card transaction data.***

I. INTRODUCTION

Credit card fraud refers to the unauthorized and illicit use of someone else's credit card without the knowledge of the cardholder or the card issuer. Preventive measures and the investigation of fraudulent activities are crucial in minimizing and preventing such incidents from recurring. Monitoring user behavior is vital in detecting and preventing fraud, as it allows for the identification of unwelcome behaviors like fraud, intrusion, and defaulting. Machine learning and data analytics play a significant role in addressing this issue by providing automated solutions. However, credit card fraud detection presents challenges, particularly in terms of learning, due to factors such as class imbalance, where there are more legitimate transactions than fraudulent ones, and the dynamic nature of transaction patterns that change over time. Real-world examples rely on automated technologies to swiftly analyze a large volume of payment requests and decide which transactions to approve. Machine learning algorithms are utilized to analyze authorized transactions and identify suspicious ones. Investigators contact cardholders to validate whether a transaction is genuine or fraudulent. The automated system continuously incorporates information from investigators to train and update the algorithm, thereby improving the accuracy of fraud detection over time. Ongoing development of fraud detection techniques is necessary to counter fraudsters who adapt their deceptive strategies. Various types of credit card fraud exist, including card theft, account bankruptcy, device intrusion, application fraud, counterfeit cards, telecommunication fraud, and online and offline credit card fraud. Several techniques, such as Artificial Neural Networks, Fuzzy Logic, Genetic Algorithms, Logistic Regression, Decision Trees, Support Vector Machines, Bayesian Networks, Hidden Markov Models, and K-Nearest Neighbors, are employed to detect and identify fraud in present-day scenarios.

II. LITERATURE REVIEW

Fraud refers to the intentional and unlawful deception aimed at achieving financial or personal gain by violating laws, regulations, or policies. The field of anomaly or fraud detection has seen significant research and the publication of publicly available information. Clifton Phua and colleagues conducted a comprehensive analysis that explored techniques like adversarial detection, automated fraud detection, and data mining applications. Similarly, Suman, a research assistant, discussed techniques such as supervised and unsupervised learning for identifying credit card fraud. While these methods have shown unexpected success in certain cases, they have not provided a reliable and long-term solution for fraud detection.

In a study focused on modelling credit card transaction data from a specific commercial bank, WenFang YU and Na Wang utilized distance sum methods, outlier mining, outlier detection mining, and outlier detection mining to accurately predict fraudulent transactions. Outlier mining, extensively used in the financial and internet industries, aims to identify components that

are disconnected from the main system or transactions that are fraudulent. By considering consumer behavior features and associated values, the gap between the observed value of an attribute and its expected value was calculated.

Unconventional techniques like hybrid data mining/complex network classification algorithms have been successful in detecting illicit activity in medium-sized online transactions. These algorithms leverage the network reconstruction algorithm to create representations of deviations from a reference group, enabling the identification of fraudulent instances within real card transaction datasets. Furthermore, efforts have been made to improve the interaction and feedback process when dealing with fraudulent transactions. In such cases, feedback is sent to the authorized system to deny the ongoing transaction.

One technique that has provided a fresh perspective in fraud detection is the Artificial Genetic Algorithm. This approach effectively detects fraudulent transactions and reduces the frequency of false alerts.

III. METHODOLOGY

The recommended approach in the study utilizes advanced machine learning techniques to identify outliers or unusual behaviors. To begin the process, a dataset was obtained from Kaggle, a platform for data analysis that provides datasets. The dataset consists of 31 columns, with 28 labeled as v1-v28 to protect sensitive information. The remaining columns include Time, Amount, and Class. The Time column indicates the time elapsed between the first and second transactions, while the Amount column represents the total traded amount of money. The Class column categorizes transactions as either Class 0 for legal or Class 1 for fraudulent.

To ensure data quality, histograms were plotted for each column to check for missing values. This graphical representation helps identify any gaps in the dataset, allowing for analysis without the need for missing value imputation. After data preparation and processing, the Class column was removed, and the time and quantity columns were standardized for fair evaluation. The data was then processed using a combination of algorithms from various modules.

The study employed a free and open-source Python library that integrates NumPy, SciPy, and matplotlib modules. This library offers a wide range of tools for data analysis and machine learning, including classification, clustering, and regression algorithms. It is designed to interface with scientific and numerical libraries, providing simple and efficient solutions. The Python application showcasing the recommended method was developed using the Jupyter Notebook platform. Additionally, the Google Colab platform, compatible with Python Notebook files, can be utilized to run the application in the cloud.

IV. IMPLEMENTATION

Implementing this idea presents challenges due to banks being legally obligated and hesitant to collaborate, mainly due to market competition, legal concerns, and the necessity to safeguard consumer data. To gather knowledge, we extensively researched various reference works that employed similar methods. One of these studies applied the proposed method to a comprehensive dataset provided by a German bank in 2006. To preserve banking confidentiality, only a summary of the findings is shared here. The technique resulted in a small number of instances on the level 1 list, but these cases were highly likely to involve fraudulent individuals. All individuals on this list had their cards closed due to their high-risk profiles. The level 2 list posed a greater challenge. However, it still provided sufficient constraints to assess each individual situation. Credit and collections officers believed that at least half of the instances on this list may involve suspicious fraudulent behavior. The last and largest list presented a more complex challenge, with only about one-third of the cases appearing suspect. To improve efficiency and reduce overhead costs, certain queries could be refined by incorporating the first five characters of passwords, email addresses, and phone numbers. These additional questions could be applied to both the level 2 and level 3 lists.

V. RESULT

The code evaluates the number of false positives identified by comparing the predicted values with the actual values. This assessment helps determine the accuracy and precision of the algorithms. The results are presented, with class 0 representing a legitimate transaction and class 1 indicating a transaction classified as fraudulent. The classification report for each method is provided, along with these outcomes. To eliminate any potential false positives, the results are compared with the class values.

Logistic regression

Model summary

- Train set
  - Accuracy = 0.95
  - Sensitivity = 0.92
  - Specificity = 0.98
  - ROC = 0.99
- Test set
  - Accuracy = 0.97
  - Sensitivity = 0.90
  - Specificity = 0.99
  - ROC = 0.97

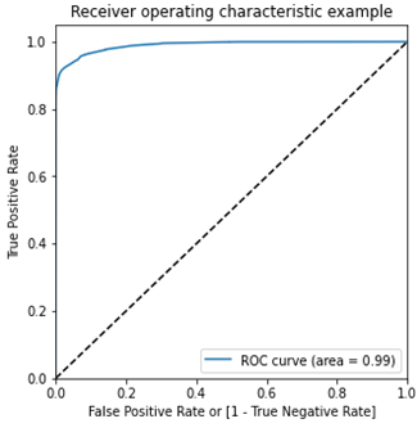

Fig. 1 ROC OF Train dataset using logistic regression

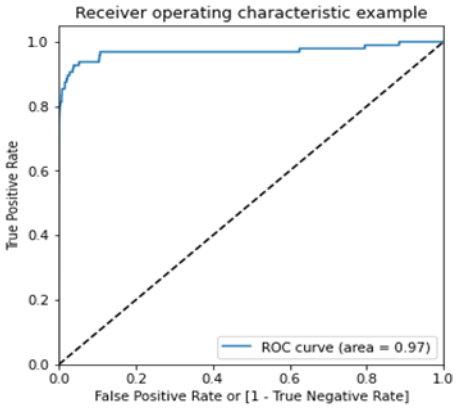

Fig. 2 ROC of Test dataset using logistic regression

Model summary

- Train set
  - Accuracy = 0.99
  - Sensitivity = 1.0
  - Specificity = 0.99
  - ROC-AUC = 1.0
- Test set
  - Accuracy = 0.99
  - Sensitivity = 0.79
  - Specificity = 0.99
  - ROC-AUC = 0.96

Overall, the model is performing well in the test set, what it had learnt from the train set.

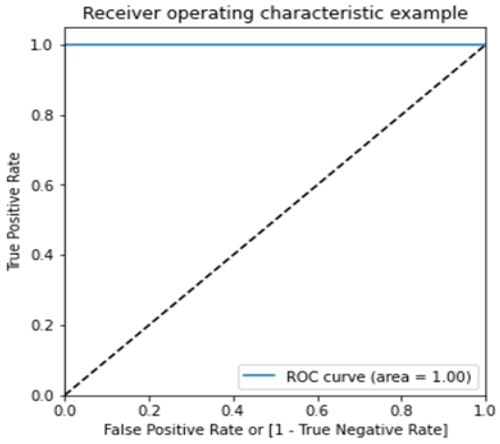

Fig. 3 ROC of train dataset using xgboost algorithm

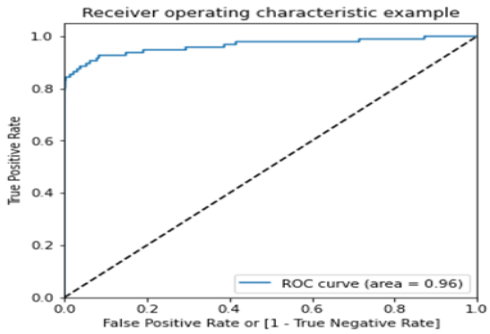

Fig. 4 ROC of Test dataset using xgboost algorithm

Decision Tree

Model summary

- Train set
  - Accuracy = 0.99
  - Sensitivity = 0.99
  - Specificity = 0.98
  - ROC-AUC = 0.99
- Test set
  - Accuracy = 0.98
  - Sensitivity = 0.80
  - Specificity = 0.98
  - ROC-AUC = 0.86

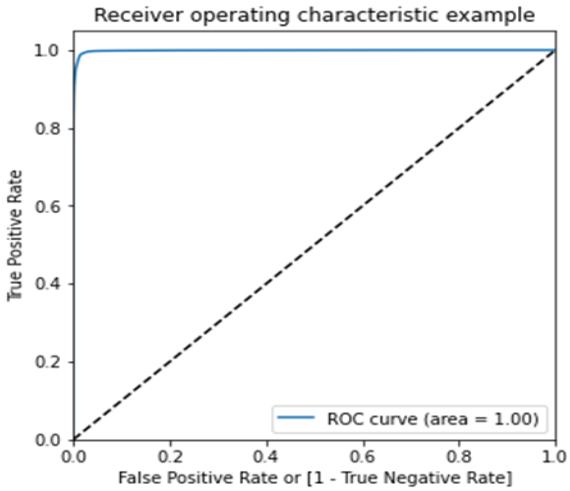

Fig. 5 ROC of train dataset using decision tree algorithm

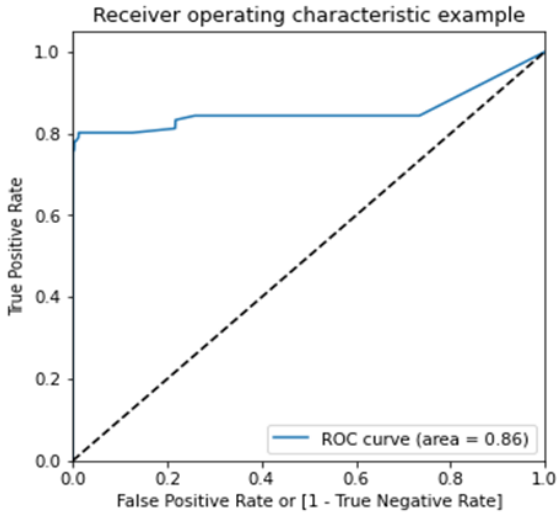

Fig. 6 ROC of test dataset using decision tree algorithm

Table 3 Classification report using decision tree algorithm

CONCLUSION

Undoubtedly, using a credit card fraudulently is a criminal act. This page provides a comprehensive list of the most common fraud schemes and offers guidance on how to identify them. It also discusses recent academic research in the field, including a detailed explanation of how machine learning can be leveraged to enhance fraud detection. The paper includes information such as the technique used, pseudocode, implementation description, and experimental results. However, due to commercial considerations, only a small subset of the dataset, comprising two days' worth of transaction records, can be made public. It is worth noting that the software's efficiency will continue to improve over time as it is based on principles of machine learning. Among the various models, the Logistic model stands out as the best choice due to its ease of interpretation and lower resource requirements compared to heavier models like Random Forest or XGBoost.

FUTURE ENHANCEMENTS

Although our objective of achieving 100% accuracy in fraud detection was not met, we have developed a system that has the potential to approach it with more time and data. Like any similar endeavor, there is

always room for improvement. The project's architecture allows for the integration of multiple algorithms as modules, enabling the combination of their outputs to enhance the accuracy of the final result. Additional algorithms can be incorporated to further enhance the model, as long as their output format aligns with the existing ones. The process of adding these modules is straightforward, as demonstrated in the provided code.

REFERENCES

[1] "Credit Card Fraud Detection Based on Transaction Behavior -by John Richard D. Kho, Larry A. Vea" was included in the proceedings of the 2017 IEEE Region 10 Conference (TENCON), which was held in Malaysia from November 5-8, 2017.

[2] CLIFTON PHUA, VINCENT LEE, KATE SMITH, & ROSS GAYLER are the authors. Published by the School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia, "A Comprehensive Survey of Data Mining-based Fraud Detection Research"

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

[4] Wen-Fang YU and Na Wang's "Research on Credit Card Fraud Detection Model Based on Distance Sum" was published by the 2009 International Joint Conference on Artificial Intelligence.

[5] By Massimiliano Zanin, Miguel Romance, Regino Criado, and Santiago Moral, "Credit Card Fraud Detection using Parenclitic Network Analysis-By, Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages."

[6] AUGUST 2018 IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy"

[7] "Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, and Mridushi" appeared in the January 2016 issue of the International Journal of Advanced Research in Computer and Communication Engineering.

[8] "Plastic Card Fraud Detection Using Peer Group Analysis" Springer, Issue 2008. David J.Wetson, David J.Hand, M. Adams, Whitrow, and Piotr Jusczak.