Grenze International Journal of Engineering and Technology, January Issue

Fraud Detection in Banking Data by Machine Learning Techniques

Narlagiri Vinay[1], G Rakhi[2], Shaik Mahaboob[3] and Mr. R Suvarna Rao[4]
[1-4]Institute of Aeronautical Engineering, Hyderabad, India
Email: vinaynarlagiri4@gmail.com, rakikumar802213h@gmail.com, skmabobcse@gmail.com, r.suvarnararao.iare.ac.in

***Abstract*—With the rise of technology and e-commerce, credit card transactions have surged, leading to an increase in banking fraud and associated costs. This study explores detecting fraudulent transactions using class weight-tuning hyperparameters, optimized through Bayesian methods, to handle un balanced data. We suggest using weight adjustment as a solution preprocessing step and evaluate XGBoost and LightGBM through a majority voting ensemble learning method. Experiments on real-world data show that LightGBM and XGBoost achieve high performance, with ROC-AUC of 0.95; precision 0.79, recall 0.80 and F1 score is 0.79. Incorporating deep learning for hyper parameter tuning further enhances performance, achieving ROC-AUC of 0.94, precision 0.80 and recall 0.82, and F1 score 0.81, surpassing current methods.**

***Index Terms*— XGBoost, LightGBM, Fraud detection, machine learning.**

I. INTRODUCTION

Recently, the number of a financial transactions have dramatically increased due to the expansion of financial institutions and growth of the e-commerce. This increase has led to the growing problem of fraudulent transactions in online banking, making fraud detection increasingly challenging. [1- 2]

As credit cards have evolved, so have patterns of the credit-card fraud. Fraudsters are continually refine their techniques to make fraudulent transactions appears legitimate, learning how to circumvent detection systems and complicating the process of identifying fraud. Consequently, researchers are consistently seeking new methods or improving existing ones to enhance fraud detection effectiveness. [3]

Fraudsters often exploit vulnerabilities in monitoring, control, and security systems within commercial applications to achieve their aims. Timely detection of fraud is crucial to prevent further fraudulent activities. [4][5] To address fraud, there are two main strategies: fraud detection and the fraud prevention. Fraud preventi on is the proactive approach designed to stop the fraud before it happens, while fraud- detection focuses on identifying and managing fraudulent activities once they occur.[7]

Banking fraud detection is a binary classification problem that classifies transactions as either fraudulent or legitimate.

[8] Due to large amount of banking data, manually analyzing and identifying fraudulent patterns are time - consuming and impractical. ML algorithms are essential for fraud detection and prediction, as they can effi ciently handle large datasets. [9] These algorithms, combined with high processing power, enhance the efficiency of fraud detection. Additionally, ML and DL provide quick and effective solutions to the real time fraud detection challenges. [.10]

Grenze ID: 01.GIJET.11.1.353_1

This study presents an effective method for identifying credit card fraud, tested on public datasets with optimized algorithms such as logistic regression, CatBoost, XGBoost, and LightGBM, both individually and through majority voting of combined methods. We also employ deep learning and the hyper-parameter tuning. An ideal fraud - detection systems should identify a high number of fraudulent cases with high precision, ensuring all results are accurate. This accuracy increases customer trust in the bank and prevents it from incurring losses due to incorrect detections. We adopt Bayesian optimisation for fraud detection and propose weight-tuning hyper parameters as a pre-processing step to address data imbalance. To enhance performance, we use XGBoost and CatBoost alongside LightGBM. We chose XGBoost for its fast training speed with big data and effective regularisation, which effectively mitigates overfitting without extensive hyperparameter tuning. We select CatBoost for its ability to deliver good results without requiring hyperparameter adjustments for overfitting control. Our approach utilizes a majority-voting ensemble method that combines LightGBM, XGBoost, and CatBoost to enhance fraud-detection on the real, unbalanced datasets. To handle un_balanced data, we use a recall precision along with the ROC- AUC metric, and we have also measure performance using the F1 - score and MCC. Our findings show that our methods out_perform present approaches.

II. Literature Survey

E-commerce transaction fraud prevention faces the challenge of rapidly evolving and diverse fraud patterns. This paper presents two novel methods: a multi-layer machine learning model and fraud islands. Fraud islands use link analysis to uncover connections between fraudulent entities, helping to reveal intricate and hidden fraud patterns. The multi- layer model tackles the diversity of fraud patterns by integrating various machine learning models, each trained with different fraud labels from banks, manual reviews, fraud alerts, and customer chargebacks. Experiments show that combining these models significantly improves the accuracy of fraud detection. [1] As healthcare schemes grow, fraudulent billing cases are on the rise, making detection difficult due to the complex interactions between doctors, patients, and services. Intelligent fraud detection models are needed to accurately identify fraudulent activities while balancing costs and benefits. This paper proposes a novel approach using sequence mining to detect insurance claim fraud. Unlike previous methods that focus on billing amounts or medication-disease relationships, this approach identifies frequent service sequences within specialties and calculates confidence values. By comparing these sequences with actual patient data, anomalies are detected. The method is validated with five years of transactional data from a local hospital, revealing numerous fraudulent cases. [2] Detecting fraudulent credit card transactions promptly is both crucial and challenging due to the highly skewed nature of the dataset, where fraud cases are rare compared to normal transactions. This work presents an ensemble machine learning approach to address this issue. It observes that RF is more accurate in detecting normal instances, while NN are better at detecting fraud instances. By combining both methods, the proposed ensemble approach leverages the strengths of each, resulting in a highly accurate and confident prediction of transaction labels.[3] This study explores the key drivers, strategic uses, and main challenges of implementing big_data analytics in the banking. A four-rounds Delphi study with 25 experts revealed that "credit risk analysis" and "fraud detection" are the primary applications. The main drivers for these projects are "new product/service development," and "enhanced decision-making" while the major challenge is dealing with "information silos and unintegrated data." These insights enhance the understanding of managing big data in banking and provide valuable guidance for industry leaders and researchers. [4] The surge in e-commerce and online payments has turned credit-card fraud into a significant global concern. Although using the machine learning(ML) algorithms for fraud- detection is gaining traction, challenges such as limited access to public datasets, imbalanced class distributions, and diverse fraudulent behaviors persist. This paper contrasts the performance of three ML algorithms— Logistic Regression (LR), SVM, and Random Forest(RFC), —on real-life credit card transaction data. To address class imbalance, the SMOTE sampling method is used. Incremental learning is employed to handle evolving fraud patterns. The performance of these techniques is evaluated using precision and the recall metrics.

[5] Credit-card fraud is the major issue in financial services, leading to billions in losses each year. This paper applies machine learning algorithms, starting with standard models and advancing to hybrid approaches like AdaBoost and majority voting. Using a publicly available dataset for initial tests and real-world data from a financial institution, the study adds noise to assess algorithm robustness. Results reveal that the majority voting methods excels in detecting fraud with the high accuracy.[6] Healthcare fraud represents a significant financial burden in the United States, leading to increased insurance premiums and potential risks to patients. There is an urgent need for sophisticated digital systems to detect and prevent such fraud. However, the intricate and varied nature of healthcare data systems and models in the US presents substantial challenges to implementing effective

detection solutions. The goal of these systems is to generate leads for investigators to pursue recoupments, recoveries, or the referrals to authorities. This article reviews existing healthcare -
fraud detection methods and summarizes peer- reviewed studies, listing their conclusions, objectives, and data-characteristics. Potential gaps on applying these systems to real-world data are discussed, and several research topics are proposed to address these gaps in future studies. [7]

III. PROPOSED METHODOLOGY

The proposed fraud detection framework, illustrated in Figure, begins with data pre- processing, followed by dividing the data into the training and testing sets. Bayesian's optimization is then applied to training data to find optimal hyperparameters for improved performance. Cross- validation is used to compare performance on an unbalanced dataset, and the algorithms are assessed using metrics like accuracy, precision, recall, MCC, F1_score, and AUC diagrams. The detailed steps are explained further in the text.

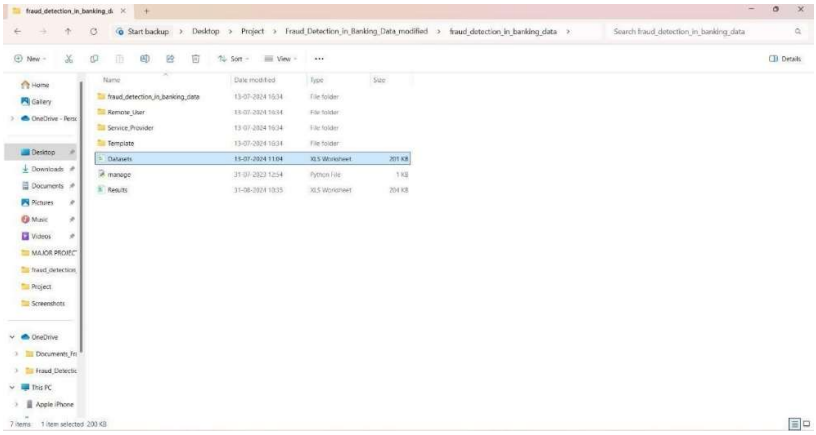

Fig 1 Desktop Site

In the Fig 1 Desktop screen we have seven folders just go inside Datasets folder to view the dataset.

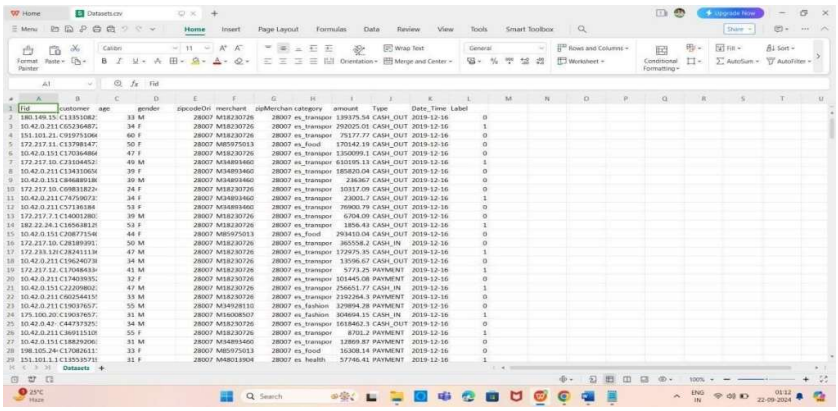

Fig 2. Dataset records

In the fig 2, the dataset records displays transaction data with attributes like Fid,customer, age, gender, zipcodeOri, merchant, zipMerchan category, amount, Type,Date_Time and Labels.
To implement this project, we have designed the following modules:

- Start the Server: To start the server, run the command python manage.py runserver in the command prompt. Once the server starts, copy the provided URL and paste it into any browser to open the user interface.
- Service Provider Login: Navigate to the login page and enter username="Admin" and password="Admin". Upon successful login, you will be directed to the Service Provider dashboard, where you can access various features such as train and test datasets, viewing model accuracy, analyzing predicted results, and downloading predicted datasets.

- Training and Testing Results: Three algorithms are used for fraud detection: Random Forest Classifier, which achieves the highest accuracy of 99.66%, Logistic Regression, and LightGBM. The results are visualized in a bar chart comparing the accuracies of all three algorithms.
- Viewing Fraud Detection Test Data Results: After training the data, the service provider can view and analyze the test data results, which indicate whether fraud was found or not. These results can also be downloaded for further analysis.
- Viewing All Remote Users: The service provider has access to a list of all remote users who have registered. This list includes detailed information such as email, address, mobile number, state, and city.
- User Login Procedure: Users need to register before logging in. Once registered, users can log in and will be directed to their profile page.
- Using the User Interface for Fraud Prediction: In the user interface, users can enter the necessary values, and the system will predict whether fraud is detected based on the input values.

IV. RESULTS ANALYSIS

In the realm of fraud detection in banking data, ensemble methods have proven to be a highly effective approach by combining multiple algorithms to enhance predictive accuracy. Utilizing classifiers such as Random Forest (RF) and Decision Trees (DT), these methods excel at identifying fraudulent transactions while minimizing false positives. This project is designed using the Django framework, where the Python web server is initiated with the command python manage.py runserver, and the resulting URL opens the user interface in any browser. The service provider interface allows administratorsto log in with credentials (e.g., Username: "Admin", Password: "Admin") and access functionalities such as managing datasets, viewing algorithm accuracy, analyzing predicted results, and downloading datasets.

The project employs three algorithms—Random Forest, Logistic Regression, and LightGBM—with Random Forest delivering the highest accuracy at 99.66%. The system also offers graphical visualizations, such as bar charts, to compare algorithm performances. Service providers can view fraud detection results, identifying whether fraud is detected in testing data, and download these results for further analysis. Additionally, they can manage all registered remote users, accessing detailed information like email, address, and contact details.

The user interface allows individuals to register, log in, and access features like profile management and fraud detection predictions by entering transaction data to determine if fraud has occurred.

Ensemble methods are particularly advantageous in banking fraud detection, asthey recognize complex patterns in transactional data and adapt to evolving fraud strategies.

These methods effectively handle imbalanced datasets, where fraudulent cases are rare, and offer insights into feature importance, ensuring system transparency. Furthermore, the real-time processing capabilities of ensemble models allow for quick detection and response to potential fraud, making them a powerful tool in enhancing financial security. By integrating ensemble learning into banking fraud detection, this system provides a robust, adaptive, and transparent solution that significantly improves resilience against fraud in an increasingly digital world.

Fig 3 User Interface

For designing this project, we have used django framework. To start python web server, we have run command python manage.py run server in command prompt. We will get one URL, that need to paste on any browser. Then this user interface will open show in fig 3.

Fig 4 Service Provider Login

By entering username=”Admin” and password=”Admin”, it will open following page as shown in fig 4. This is service provider login page.

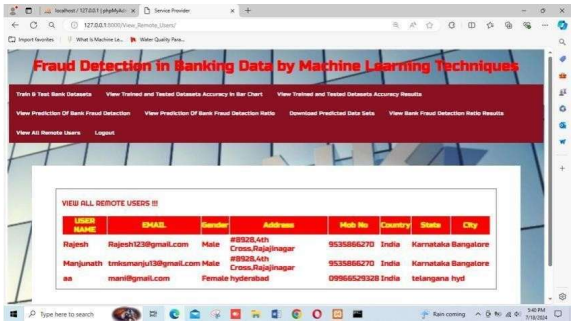

Fig 5 Service Provides Login Access

In the fig 5, service provider has access of train and test dataset, view accuracy, see predicted result, download predicted datasets, etc.

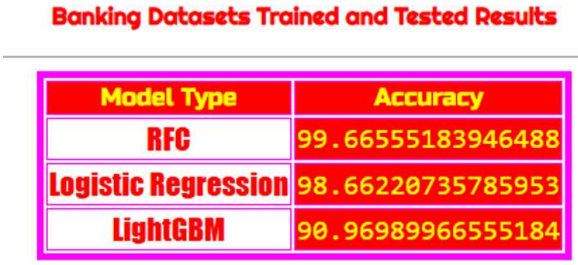

Fig 6 Training and Testing Results

There are three algorithms shown in fig 6, we have used Random Forest Classfier, Logistic regression and Light GBM. Out of which, Random Forest is giving highest accuracy i.e.99.66%.

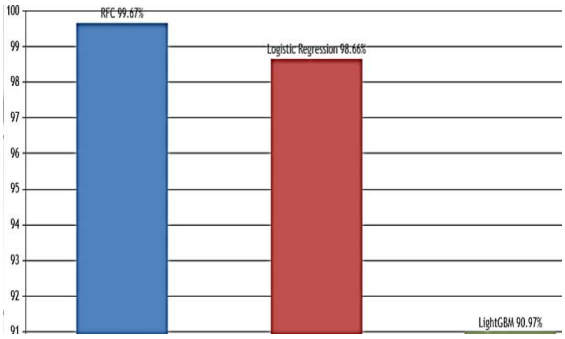

Fig 7 Graphical Visualization

The above fig 7 is the graphical visualization of accuracy in bar chart format of all three algorithms

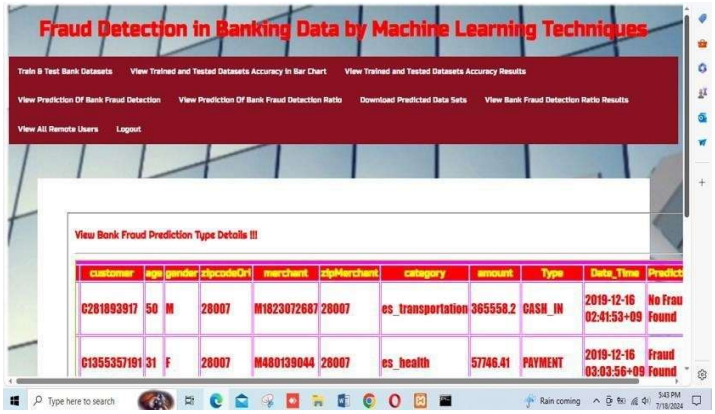

Fig 8 Fraud Detection Test Data Results

After training data, service provider can see and analyse the results of testing data(No fraud found/Fraud found). Here service provider can download this data as shown in fig 8.

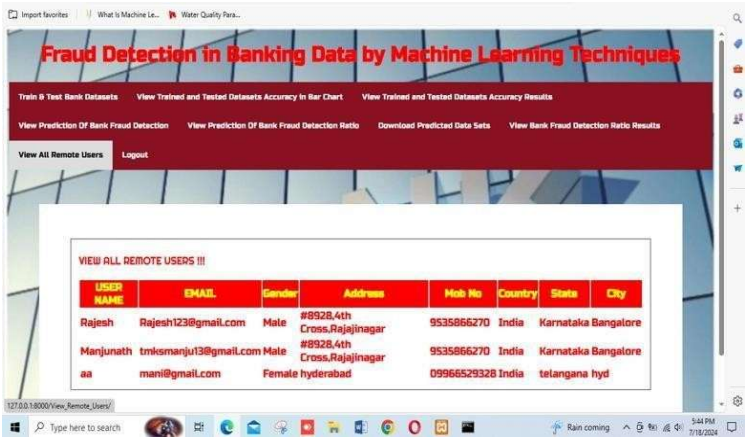

Fig 9 View all Remote Users

The service can view all registered and logged-in users, including their detailed information like email, address, mobile number, state, and city as shown in fig 9.

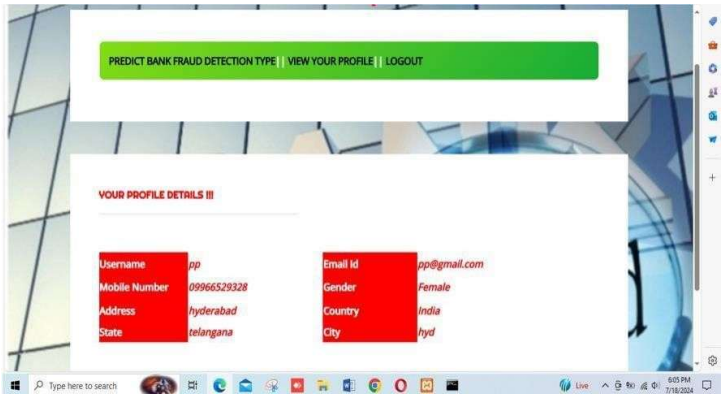

Fig 10 User Interface 2

The above fig 10 is user interface where user can see profile and check fraud detection.

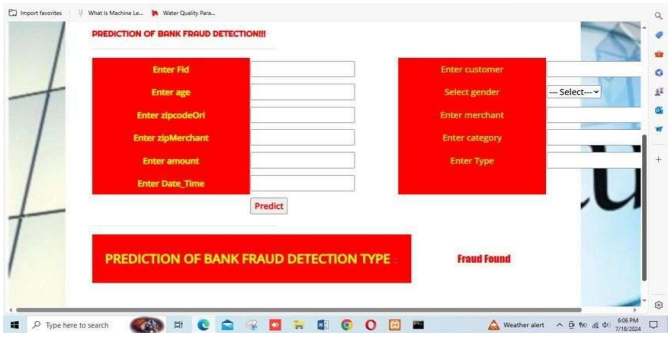

Fig 11 Fraud Prediction

After entering the value for all, it will predict fraud found or no fraud found which is in fig 11.

V. CONCLUSION

In this study, we tackled credit card fraud detection in real, unbalanced datasets by proposing a machine learning approach to enhance detection performance. Using a public dataset, we implemented LightGBM with class weight tuning for optimal hyperparameters, achieving significant improvements in fraud detection and F1-score compared to previous methods. The majority voting algorithm and deep learning methods further enhanced our results. Our findings highlight that addressing data imbalance through hyperparameters, rather than sampling, is more efficient. Future work should explore hybrid models and refine CatBoost hyperparameters, with stronger hardware potentially yielding even better outcomes. There are three algorithms, we have used Random Forest Classifier, Logistic regression and Light GBM. Out of which, Random Forest is giving highest accuracy i.e.99.66%.

FUTURE SCOPE

The future of fraud detection in banking data using machine learning techniques is set to witness significant advancements. The evolution of algorithms, realtime detection systems, adaptive learning, and multi-channel data integration will enhance fraud detection while ensuring compliance with ethical and regulatory standards. Collaboration between institutions, technological innovation, and a focus on customer experience will drive the success of these initiatives.

REFERENCES

[1] J. Nanduri, Y.-W. Liu, K. Yang, and Y. Jia, ‘‘Ecommerce fraud detection through fraud islands and multi-layer machine learning model,’’ in Proc. Future Inf. Commun. Conf., in Advances inInformation and Communication. San Francisco, CA, USA: Springer, 2020, pp. 556–570.
[2] Matloob, S. A. Khan, R. Rukaiya, M. A. K. Khattak, and A. Munir, ‘‘A sequence mining-based novel architecture for detecting fraudulent transactions in healthcare systems,’’ IEEE Access, vol. 10, pp. 48447–48463, 2022.
[3] H. Feng, ‘‘Ensemble learning in credit card fraud detection using boosting methods,’’ in Proc. 2nd Int. Conf. Comput. Data Sci. (CDS), Jan. 2021, pp. 7–11.
[4] M. S. Delgosha, N. Hajiheydari, and S. M. Fahimi, ‘‘Elucidation of big data analytics in banking: A four-stage delphi study,’’ J. Enterprise Inf. Manage., vol. 34, no. 6, pp. 1577–1596, Nov. 2021.
[5] M. Puh and L. Brkić, ‘‘Detecting credit card fraud using selected machine learning algorithms,’’ in Proc. 42nd Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO), May 2019, pp. 1250–1255.
[6] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, ‘‘Credit card fraud detection using AdaBoost and majority voting,’’ IEEE Access, vol. 6, pp. 14277–14284, 2018.
[7] N. Kumaraswamy, M. K. Markey, T. Ekin, J. C. Barner, and K. Rascati, ‘‘Healthcare fraud data mining methods: A look back and look ahead,’’ Perspectives Health Inf. Manag., vol. 19, no. 1, p. 1, 2022.
[8] E. F. Malik, K. W. Khaw, B. Belaton, W. P. Wong, and X. Chew, ‘‘Credit card fraud detection using a new hybrid machine learning architecture,’’ Mathematics, vol. 10, no. 9, p. 1480, Apr. 2022.
[9] K. Gupta, K. Singh, G. V. Singh, M. Hassan, G. Himani, and U. Sharma, ‘‘Machine learning based credit card fraud detection—A review,’’ in Proc. Int. Conf. Appl. Artif. Intell. Comput. (ICAAIC), 2022, pp. 362–368.
[10] R. Almutairi, A. Godavarthi, A. R. Kotha, and E. Ceesay, ‘‘Analyzing credit card fraud detection based on machine learning models,’’ in Proc. IEEE Int. IoT, Electron. Mechatronics Conf. (IEMTRONICS), Jun. 2022, pp. 1–8