

FRAUD DETECTION IN BANKING DATA BY MACHINE LEARNING TECHNIQUES

Mohammed Abdul Naveed¹, Shaik Abdullah², Mohammed Mazeen Ahmed³, Dr. Md Zainlabuddin⁴

^{1,2,3}B. E Student, Department of CSE, ISL College of Engineering, India.

⁴Associate Professor, Department of CSE, ISL College of Engineering, Hyderabad, India.

ABSTRACT:

As technology developed and e-commerce services expanded, credit cards became one of the most popular payment methods, resulting in a rise in the number of banking transactions. In addition, the significant rise in fraud requires high banking transaction costs. As a result, detecting fraudulent activities has become a fascinating topic. In this study, it examines the use of class weight-tuning hyper parameters to control the weight of legitimate and fraudulent transactions. Specifically, it uses Bayesian optimization to optimize the hyper parameters while preserving practical issues such as unbalanced data. It proposes weight-tuning as a pre-process for unbalanced data, as well as Cat Boost and XG Boost to enhance the efficiency of the LightGBM method by taking into account for the voting mechanism. To enhance performance even further, it applies deep learning to fine-tune the hyper parameters, particularly proposed weight-tuning technique. It conducts experiments using real-world data to test the proposed methods. In addition to the standard ROC-AUC, it utilizes recall-precision metrics to better cover unbalanced datasets. Cat Boost, LightGBM, and XGBoost, logistic regression is evaluated individually using a 5-fold cross-validation method. In addition, the majority voting ensemble learning technique is used to evaluate the performance of the combined algorithms. The results show that the proposed methods outperform the cutting-edge methods and achieve a significant improvement in performance.

Introduction

The growth of financial institutions and web-based e-commerce have increased financial transaction volumes in recent years. Fraud detection has always been difficult, but internet financial fraud is rising. Credit card fraud has evolved with credit card development. An ideal fraud detection system detects more fraudulent cases and has high precision, i.e., all results should be correctly detected, which builds customer trust and prevents losses. It addresses internet financial fraud. Fraud detection is difficult and requires appropriate solutions. Deep learning, hyper parameter settings, and machine learning methods like Cat Boost, LightGBM, and XGBoost are suggested to increase fraud detection. For imbalanced data, it recommends pre-processing using the weight-tuning hyperparameter and Bayesian optimization for fraud detection. In addition to LightGBM, it recommends CatBoost and XGBoost for performance. Because of its rapid training in massive data and "regularization," the XGBoost method avoids overfitting by analyzing the difficulty of setting up the tree and hyper parameters quickly. Cat boost is another algorithm. as it gives excellent results without changing hyper settings for over fitting management. not

changing hyper parameters like other algorithmic learning procedures. An ensemble learning system with majority voting is suggested. Combine Cat Boost, XG Boost, and LightGBM to see how they affect fraud detection on imbalanced data. Deep learning is suggested for hyper parameter adjustments, etc. It extensively evaluates the offered methods using real-world data. Recall precision and ROC-AUC are employed to cover imbalanced datasets. It also evaluates performance using F1_score and MCC. Results show that proposed tactics outperform tried-and-true ones. More researchers should disclose source codes and utilize publicly available data sets.

PROBLEM STATEMENT:

The proposed Project for credit card fraud detection including the dataset, pre-processing, feature extraction and feature selection, algorithms, framework, and evaluation metrics, is presented and discusses the evaluation results of the experiments performed, and finally concludes the project with framework predict of credit card fraud.

LITERATURE SURVEY

A sequence mining-based novel architecture for detecting fraudulent transactions in healthcare systems:

Authors: I. Matloob, S. A. Khan, R. Rukaiya, M. A. K. Khattak, and A. Munir

<https://ieeexplore.ieee.org/document/9764751>

Abstract: Built a healthcare fraud detection model with sequence mining. Analyzed service sequences in specialty areas, calculated confidence values, and employed a rule engine to compare them with actual patient data for anomaly detection. Validated using five years of hospital transactional data Built a healthcare fraud detection model with sequence mining. Analyzed service sequences in specialty areas, calculated confidence values, and employed a rule engine to compare them with actual patient data for anomaly detection. Validated using five years of hospital transactional data The implementation of a sequence mining-based healthcare fraud detection system may encounter challenges, including false positives, privacy issues, and difficulties in adapting to emerging fraud tactics. In conclusion, this process-based fraud detection approach using sequence mining shows promise in identifying healthcare insurance claim fraud. It enhances transparency, cost efficiency, and detection accuracy, validated on real hospital data.

Machine learning based credit card fraud detection—A review Authors: K. Gupta, K.

Singh, G. V. Singh, M. Hassan, G. Himani, and U. Sharma

<https://ieeexplore.ieee.org/document/9792653>

Abstract: The methodology entails tackling evolving digital transaction fraud using a sequence of machine learning models. It involves selecting optimal methods through evaluation and enabling real-time fraud detection with predictive analysis using an API module. The approach also emphasizes efficient data handling strategies The proposed system is an advanced real-time fraud detection solution for digital transactions. It utilizes machine learning models and API modules to effectively identify and prevent credit card fraud. Challenges for real-time fraud detection systems include high development costs, false positives, privacy issues, evolving tactics, and machine learning imperfections. In conclusion, the digital era offers advantages but also presents evolving fraud

challenges, especially in online transactions like credit card fraud. This review emphasizes real-time fraud detection through machine learning and efficient data handling strategies.

Analyzing credit card fraud detection based on machine learning models.

Authors: R. Almutairi, A. Godavarthi, A. R. Kotha, and E. Ceesay

Abstract: Utilized machine learning and data science to detect credit card fraud. This involves analyzing historical transaction data, applying diverse modeling techniques, and implementing real-time fraud detection algorithms. The proposed system harnesses machine learning and data science to analyze historical transactions, enhancing real-time monitoring to effectively mitigate credit card fraud risks across diverse payment scenarios. Challenges in adopting a machine learning-based fraud detection system include costs, false positives, slow adaptation, privacy issues, limited effectiveness against evolving fraud, and potential disruption to payment systems. In summary, the widespread use of credit cards, particularly in online transactions, presents a substantial fraud risk. The utilization of machine learning and data science is crucial for effective fraud detection and prevention.

System Architecture

The system starts with raw credit card transaction data, including fraud and authenticity marks. Machine learning requires data preparation, including feature extraction and selection. The dataset has two subsets: trained for model building and tested for performance assessment. Bayesian optimization optimises machine learning hyperparameters. CatBoost, LightGBM, and XGBoost are used to training data using 5-fold cross-validation to guarantee model resilience. The project also considered stacking classifiers. The algorithms' credit card fraud detection and false positive reduction are evaluated using several criteria.

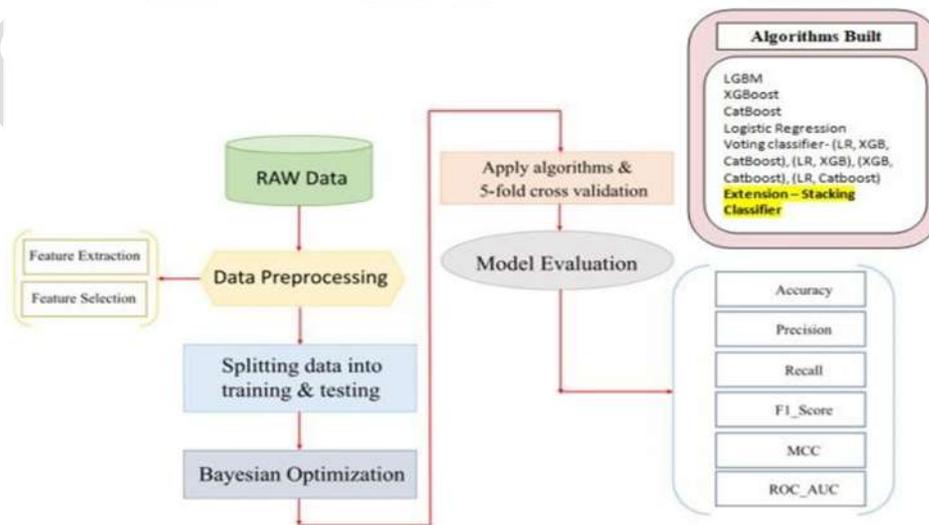


Fig6. System architecture

CLASS DIAGRAM

In software pack age engineering, a class diagram at intervals of the Unified ModellingLanguage(UML)can be the fashion of a static structure diagram that describes the Structure of a system by showing the system`s categories, attributes, operations, and additionally the relationships among objections.

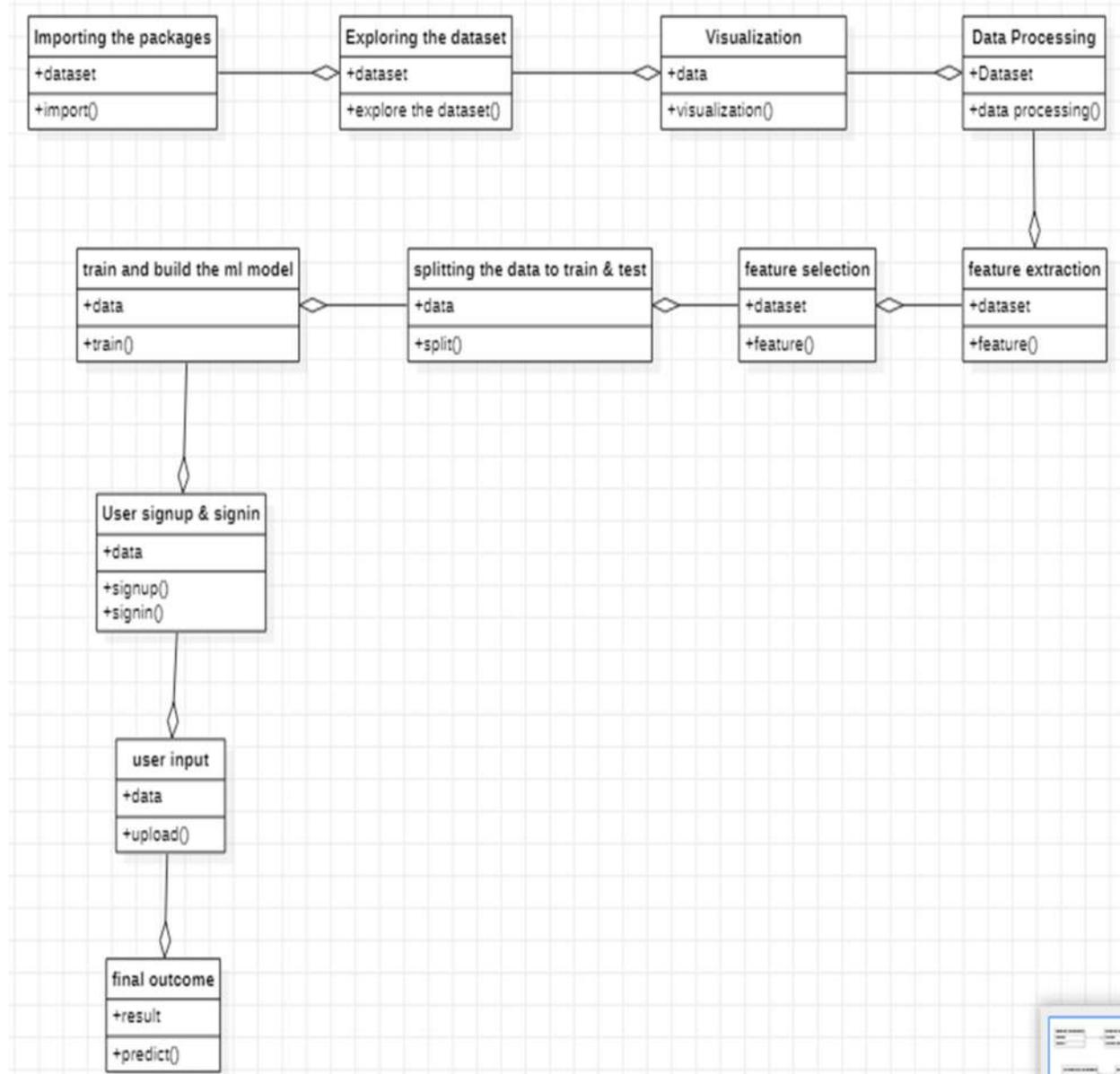


Fig2.ClassDiagram...

SEQUENCE DIAGRAM

A sequence diagram or system sequence diagram (SSD) shows object interactions arranged in timesequence within the field of software package engineering. It depicts the objects involved within the state of affairs and also the sequence of messages changed between the objects required to holdout the practicality of the state of affairs.

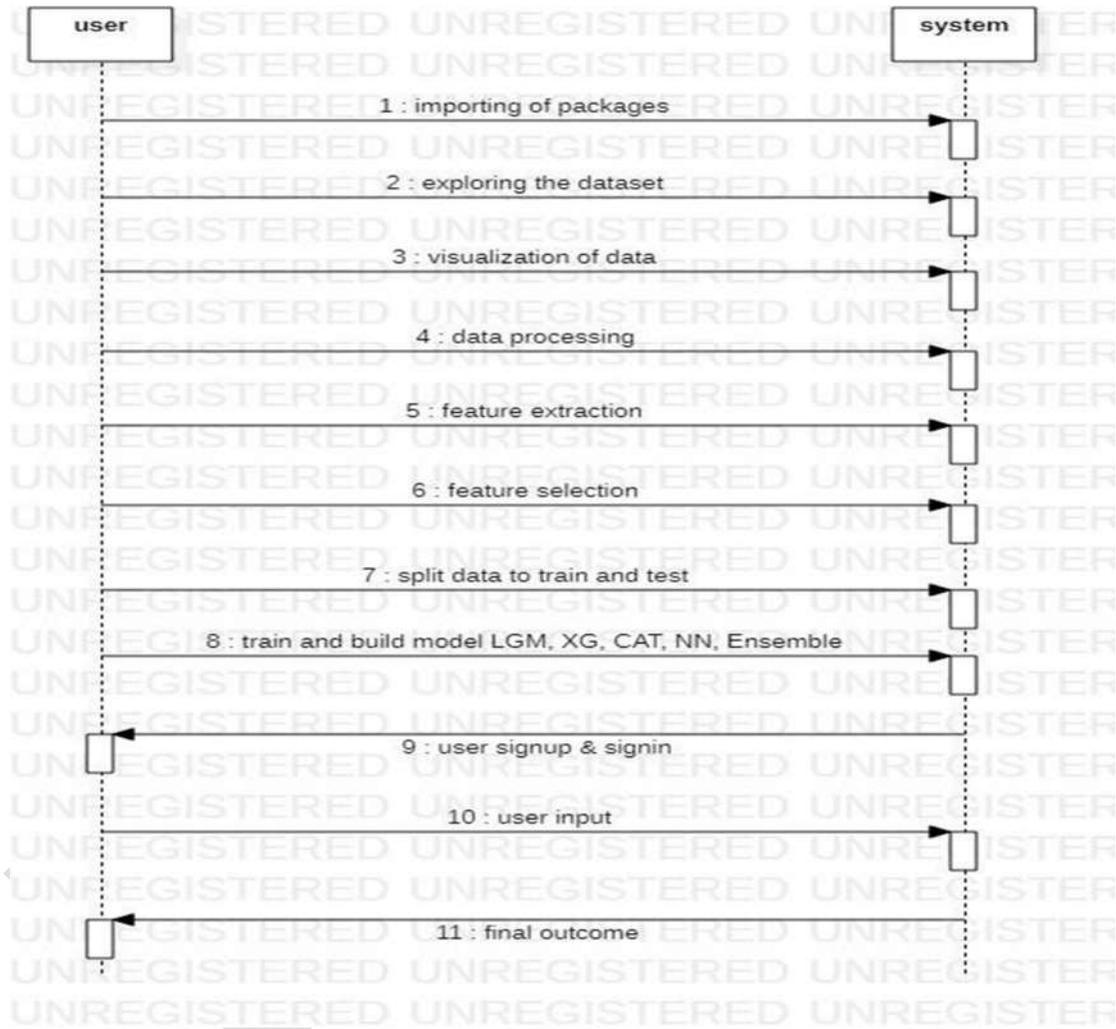


Fig3.SequenceDiagram...

IMPLEMETATION

6.1.1 Algorithms:

Bayesian Optimization: Bayesian Optimization provides a principled technique based on Bayes Theorem to direct a search of a global optimization problem that is efficient and effective. It works by building a probabilistic model

of the objective function, called the surrogate function that is then searched efficiently with an acquisition function before candidate samples are chosen for evaluation on the real objective function.

CV Stratified K-fold: Stratified k-fold cross-validation is the same as just k-fold cross-validation, But Stratified k-fold cross-validation, it does stratified sampling instead of random sampling.

Smote Sampling (over and under sampling): SMOTE is a technique that oversamples the minority class by synthetically generating data points. It uses k nearest neighbours to create new examples that are similar to the existing ones. SMOTE can be combined with random under sampling of the majority class to balance the class distribution and improve the performance of the classifier

Hyper parameters: Hyper parameters that cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.

Light GBM: LightGBM, short for light gradient-boosting machine, is a free and open-source distributed gradient-boosting framework for machine learning, originally developed by Microsoft. It is based on decision tree algorithms and used for ranking, classification and other machine learning tasks. The development focus is on performance and scalability.

XG Boost: XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

Cat Boost: CatBoost is an open-source boosting library developed by Yandex. It is designed for use on problems like regression and classification having a very large number of independent features.

Neural Network: A neural network (NN), in the case of artificial neurons called artificial neural network (ANN) or simulated neural network (SNN), is an interconnected group of natural or artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach to computation.

Ensemble Methods: The ensemble methods in machine learning combine the insights obtained from multiple learning models to facilitate accurate and improved decisions. These methods follow the same principle as the example of buying an air-conditioner cited above. In learning models, noise, variance, and bias are the major sources of error. The ensemble methods in machine learning help minimize these error-causing factors, thereby ensuring the accuracy and stability of machine learning (ML) algorithms.

Stacking Classifier (Gradient Boosting with RF + LightGBM): A stacking classifier is an ensemble learning method that combines multiple classification models to create one “super” model. This can often lead to improved

performance, since the combined model can learn from the strengths of each individual model.

JUSTIFICATION AND RESULTS

7.1 Performance Evaluation Metrics:

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Accuracy: Accuracy is the proportion of correct predictions in a classification task, measuring the overall correctness of a model's predictions.

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

F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balanced measure that considers both false positives and false negatives, making it suitable for imbalanced datasets.

$$F1 \text{ Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$

Performance Evaluation Table:

	Model	accuracy	precision	recall	f1
0	LGBM	0.996908	0.344762	0.413333	0.355455
1	XGB	0.999122	0.883232	0.588889	0.682707
2	CatBoost	0.999262	0.848312	0.713333	0.768263
3	Vot_Lg,Xg,Ca	0.996908	0.344762	0.413333	0.355455
4	Vot_Lg,Xg	0.999122	0.883232	0.588889	0.682707
5	Vot_Xg,Ca	0.999262	0.848312	0.713333	0.768263
6	Vot_Lg,Ca	0.999227	0.830345	0.733333	0.764458

7	Stacking	0.999332	0.85101	0.753333	0.795105
---	----------	----------	---------	----------	----------

7.1.1 Graph

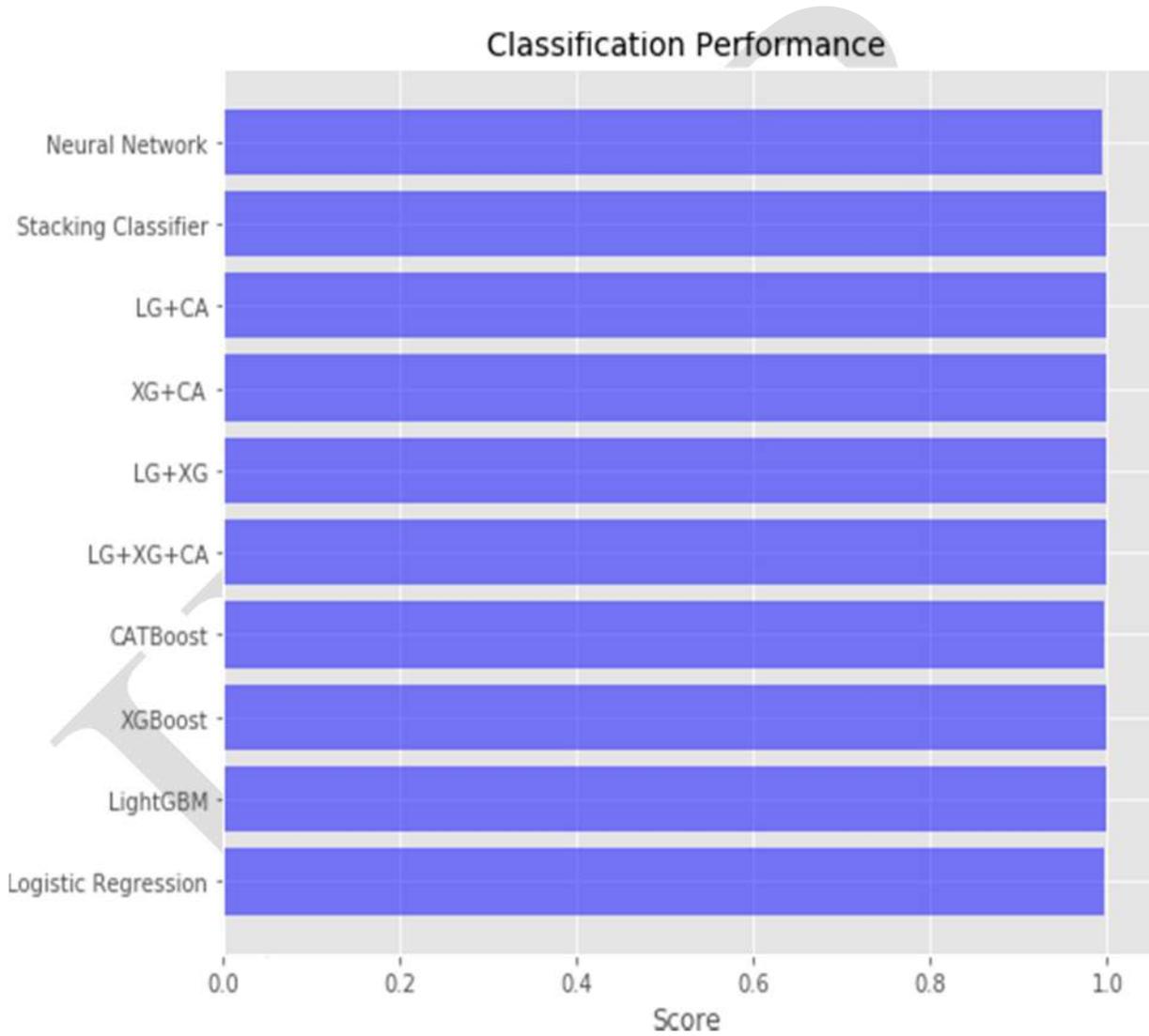


Fig7: Score Graph

7.2 TEST CASES:

Test Case 1:

DESCRIPTION	INPUT	XPECTED OUTPUT	ESTERNAME
Detecting Known Fraudulent Transactions	Dataset containing knownfraudulent transactions	Model correctly identifies transactions as fraudulent	Mohd Abdul Raheem Uddin Arsalan
Detecting Known Legitimate Transactions	Dataset containing knownlegitimate transactions	Model correctly identifies transactions as non-fraudulent	Mohd Abdul Raheem Uddin Arsalan

Test Case 2:

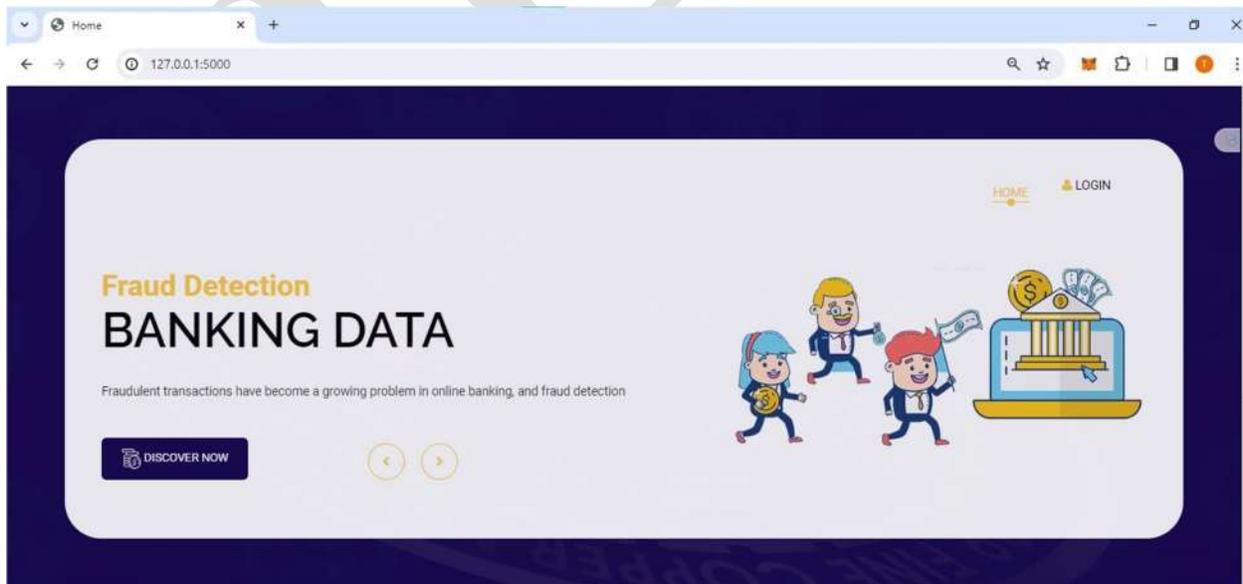
DESCRIPTION	INPUT	XPECTED OUTPUT	ESTERNAME
Large Dataset	Very large dataset	Model processesdata within a reasonable time frame, providing accurate predictions	Syed Shafi Uddin Ahmed
High Frequencyof Transactions	Dataset simulating highfrequency of transactions	Model maintains performance and accurately detects fraud under high load	Syed Shafi Uddin Ahmed

Test Case 3:

DESCRIPTION	INPUT	XPECTED OUTPUT	ESTERNAME
Precision and Recall Balance	Balanced dataset with equal number of frauds/legits	Model achieves a good balance between precision and recall, minimizing false positives/negatives	Mohd Aftab Ahmed Khan
ROC Curve and AUC	Dataset for plotting ROC curve	Model has high AUC value, indicating good overall performance	Mohd Aftab Ahmed Khan

CHAPTER 8: SCREENSHOTS

Fig 8 Home Page





Sign In

[Create an account](#)

Fig 9 Sign in Page

FORM

Fig 10 Upload Input Values



Fraudulent Transaction Happened based on the ML for the Given Input!



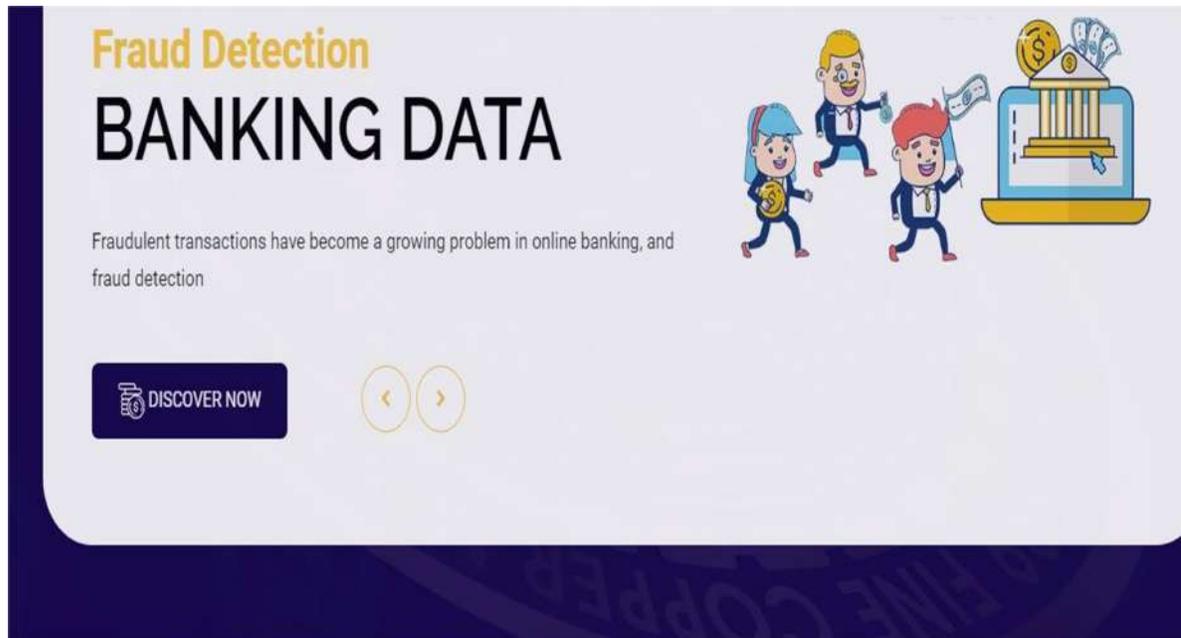
Fig 11 Predict result

FORM

-0.206563459
2.106606021
1.167287618
-0.059554648
0.078147526
0.299318483

Predict

Fig 12 Upload another input values



NON-Fraudlent Transaction Happened based on the ML for the Given Input!

Fig 13 Final Outcome

CONCLUSION

CONCLUSION:

The conclusion is that the proposed machine learning approach using class weight-tuning hyper parameters, Cat Boost, LightGBM, and XG Boost algorithms, and deep learning to fine-tune the hyper parameters significantly improves the performance of fraud detection in real unbalanced datasets. This proposed approach can be used to detect fraudulent activities in banking data and reduce the high banking transaction costs associated with fraud. The result shows that the proposed methods outperform the other cutting-edge methods and achieve a significant improvement in performance. For the Future work, it suggests that using other hybrid models and working specifically in the field of Cat Boost by changing more hyper parameters, especially the number of trees having a chance of increase in the performance of this proposed model. The assurance of the results of MCC for unbalanced data proved that, compared to other criteria of evaluation, it's stronger. By combining the LightGBM and XG Boost methods, it obtained 0.79 and 0.81 for the deep learning method. Using hyper parameters to address data unbalance compared. to sampling methods, in addition to reducing memory and time needed to evaluate algorithms, also has better results. For future studies and work, it proposes using other hybrid models as well as working specifically in the field of Cat Boost by changing more hyper parameters, especially the hyper parameter number of trees. Also, due to hardware limitations in this study, the use of stronger and better hardware may bring better results that can ultimately be compared

with the results of this study.

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 in Information and Communication. San Francisco, CA, USA: Springer, 2020, pp. 556–570.
- [2] I. 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] Ijteba Sultana, Dr. Mohd Abdul Bari ,Dr. Sanjay," *Routing Performance Analysis of Infrastructure-less Wireless Networks with Intermediate Bottleneck Nodes*", International Journal of Intelligent Systems and Applications in Engineering, ISSN no: 2147-6799 IJISAE, Vol 12 issue 3, 2024, Nov 2023
- [7] Md. Zainlabuddin, "*Wearable sensor-based edge computing framework for cardiac arrhythmia detection and acute stroke prediction*", Journal of Sensor, Volume2023.
- [8] Md. Zainlabuddin, "*Security Enhancement in Data Propagation for Wireless Network*", Journal of Sensor, ISSN: 2237-0722 Vol. 11 No. 4 (2021).
- [9] Dr MD Zainlabuddin, "*CLUSTER BASED MOBILITY MANAGEMENT ALGORITHMS FOR WIRELESS MESH NETWORKS*", Journal of Research Administration, ISSN:1539-1590 | E-ISSN:2573-7104 , Vol. 5 No. 2, (2023)
- [10] Vaishnavi Lakadaram, " Content Management of Website Using Full Stack Technologies", Industrial Engineering Journal, ISSN: 0970-2555 Volume 15 Issue 11 October 2022
- [11] Dr. Mohammed Abdul Bari, Arul Raj Natraj Rajgopal, Dr.P. Swetha , " *Analysing AWSDevOps CI/CD Serverless Pipeline Lambda Function's Throughput in Relation to Other Solution*", International Journal of Intelligent Systems and Applications in Engineering , JISAE, ISSN:2147-6799, Nov 2023, 12(4s), 519–526

- [12] Ijteba Sultana, Mohd Abdul Bari and Sanjay," *Impact of Intermediate per Nodes on the QoS Provision in Wireless Infrastructure less Networks*", Journal of Physics: Conference Series, Conf. Ser. 1998 012029 , CONSIPIO Aug 2021
- [13] M.A.Bari, Sunjay Kalkal, Shahanawaj Ahamad," *A Comparative Study and Performance Analysis of Routing Algorithms*", in 3rd International Conference ICCIDM, Springer - 978- 981-10-3874-7_3 Dec (2016)
- [14] Mohammed Rahmat Ali,: *BIOMETRIC: AN e-AUTHENTICATION SYSTEM TRENDS AND FUTURE APLICATION*", International Journal of Scientific Research in Engineering (IJSRE), Volume1, Issue 7, July 2017
- [15] Mohammed Rahmat Ali,: *BYOD.... A systematic approach for analyzing and visualizing the type of data and information breaches with cyber security*", NEUROQUANTOLOGY, Volume20, Issue 15, November 2022
- [16] Mohammed Rahmat Ali, *Computer Forensics -An Introduction of New Face to the Digital World*, International Journal on Recent and Innovation Trends in Computing and Communication, ISSN: 2321-8169-453 – 456, Volume: 5 Issue: 7
- [17] Mohammed Rahmat Ali, *Digital Forensics and Artificial Intelligence ...A Study*, International Journal of Innovative Science and Research Technology, ISSN:2456-2165, Volume: 5 Issue:12.
- [18] Mohammed Rahmat Ali, *Usage of Technology in Small and Medium Scale Business*, International Journal of Advanced Research in Science & Technology (IJARST), ISSN:2581-9429, Volume: 7 Issue:1, July 2020.
- [19] Mohammed Rahmat Ali, *Internet of Things (IOT) Basics - An Introduction to the New Digital World*, International Journal on Recent and Innovation Trends in Computing and Communication, ISSN: 2321-8169-32-36, Volume: 5 Issue: 10
- [20] Mohammed Rahmat Ali, *Internet of things (IOT) and information retrieval: an introduction*, International Journal of Engineering and Innovative Technology (IJEIT), ISSN: 2277-3754, Volume: 7 Issue: 4, October 2017.
- [21] Mohammed Rahmat Ali, *How Internet of Things (IOT) Will Affect the Future - A Study*, International Journal on Future Revolution in Computer Science & Communication Engineering, ISSN: 2454-424874 – 77, Volume: 3 Issue: 10, October 2017.
- [22] Mohammed Rahmat Ali, *ECO Friendly Advancements in computer Science Engineering and Technology*, International Journal on Scientific Research in Engineering(IJSRE), Volume: 1 Issue: 1, January 2017
- [23] Ijteba Sultana, Dr. Mohd Abdul Bari ,Dr. Sanjay, "*Routing Quality of Service for Multipath Manets, International Journal of Intelligent Systems and Applications in Engineering*", JISAE, ISSN:2147-6799, 2024, 12(5s), 08–16;

- [24] Mr. Pathan Ahmed Khan, Dr. M.A Bari,: Impact Of Emergence With Robotics At Educational Institution And Emerging Challenges”, International Journal of Multidisciplinary Engineering in Current Research(IJMEC), ISSN: 2456-4265, Volume 6, Issue 12, December 2021,Page 43-46
- [25] Shahanawaj Ahamad, Mohammed Abdul Bari, Big Data Processing Model for Smart City Design: A Systematic Review “, VOL 2021: ISSUE 08 IS SN : 0011-9342 ;Design Engineering (Toronto) Elsevier SCI Oct : 021
- [26] Syed Shehriyar Ali, Mohammed Sarfaraz Shaikh, Syed Safi Uddin, Dr. Mohammed Abdul Bari, “Saas Product Comparison and Reviews Using Nlp”, Journal of Engineering Science (JES), ISSN NO:0377-9254, Vol 13, Issue 05, MAY/2022
- [27] Mohammed Abdul Bari, Shahanawaj Ahamad, Mohammed Rahmat Ali,” Smartphone Security and Protection Practices”, International Journal of Engineering and Applied Computer Science (IJEACS) ; ISBN: 9798799755577 Volume: 03, Issue: 01, December 2021 (International Journal,U K) Pages 1-6
- [28] .A.Bari& Shahanawaj Ahamad, “Managing Knowledge in Development of Agile Software”, in International Journal of Advanced Computer Science & Applications (IJACSA), ISSN: 2156-5570, Vol: 2, No: 4, pp: 72-76, New York, U.S.A., April 2011
- [29] Imreena Ali (Ph.D), Naila Fathima, Prof. P.V.Sudha ,“Deep Learning for Large-Scale Traffic-Sign Detection and Recognition”, Journal of Chemical Health Risks, ISSN:2251-6727/ JCHR (2023) 13(3), 1238-1253
- [30] Imreena, Mohammed Ahmed Hussain, Mohammed Waseem Akram” An Automatic Advisor for Refactoring Software Clones Based on Machine Learning”, Mathematical Statistician and Engineering Applications Vol. 72 No. 1 (2023)
- [31] Mrs Imreena Ali Rubeena,Qudsiya Fatima Fatimunisa “Pay as You Decrypt Using FEPOD Scheme and Blockchain”, Mathematical Statistician and Engineering Applications: <https://doi.org/10.17762/msea.v72i1.2369> Vol. 72 No. 1 (2023)
- [32] Imreena Ali , Vishnuvardhan, B.Sudhakar,” Proficient Caching Intended For Virtual Machines In Cloud Computing”, International Journal Of Reviews On Recent Electronics And Computer Science , ISSN 2321-5461,IJRRECS/October 2013/Volume-1/Issue-6/1481-1486
- [33] Heena Yasmin, A Systematic Approach for Authentic and Integrity of Dissemination Data in Networks by Using Secure DiDrip, INTERNATIONAL JOURNAL OF PROFESSIONAL ENGINEERING STUDIES, Volume VI /Issue 5 / SEP 2016
- [34] Heena Yasmin, Cyber-Attack Detection in a Network, Mathematical Statistician and Engineering Applications, ISSN:2094-0343, Vol.72 No.1(2023)
- [35] Heena Yasmin, Emerging Continuous Integration Continuous Delivery (CI/CD) For Small Teams, Mathematical Statistician and Engineering Applications, ISSN:2094-0343, Vol.72 No.1(2023)