International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

FRAUD TRANSACTIONS DETECTION USING MACHINE LEARNING

Madhu Bandari, G. Sri Harsha Nayudu

1. Assistant Professor, Department of Data Science and Artificial Intelligence, IcfaiTech School, Telangana, India. madhu.bandari@ifheindia.org

2 Department of Data Science and Artificial Intelligence, IcfaiTech School, Telangana, India. harshanayudu3@gmail.com

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, we examine the use of class weight-tuning hyper parameters to control the weight of legitimate and fraudulent transactions. Specifically, we use Bayesian optimization to optimize the hyper parameters while preserving practical issues such as unbalanced data. We propose 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, we apply deep learning to fine-tune the hyper parameters, particularly our proposed weight-tuning technique. We conduct experiments using real-world data to test the proposed methods. In addition to the standard ROC-AUC, we utilize recall-precision metrics to better cover unbalanced datasets. Cat Boost, LightGBM, and XGBoost , logistic regression are 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.

Index Terms *- Bayesian optimization, data Mining, deep learning, ensemble learning, hyper parameter, unbalanced data, machine learning.*

1. INTRODUCTION

In recent years, there has been a significant increase in the volume of financial transactions due to the expansion of financial institutions and the popularity of web-based e-commerce. Fraudulent transactions have become a growing problem in online banking, and fraud detection has always been challenging. Along with credit card development, the pattern of credit card fraud has always been updated.An ideal fraud detection system should detect more fraudulent cases, and the precision of detecting fraudulent cases should be high, i.e., all results should be correctly detected, which will lead to the trust of customers in the bank, and on the other hand, the bank will not suffer losses due to incorrect detection. So, this paper discusses the problem of fraudulent transactions in online banking. There is lot of challenges for detecting fraudulent activities and the need for effective fraud detection methods. The paper proposes the use of machine learning techniques, such as Cat Boost, LightGBM, and XGBoost, to improve the performance of fraud detection, as well

as deep learning and hyper parameter settings. We employ Bayesian optimization for fraud detection and suggest using the weight-tuning hyperparameter as a pre-process step to address the issue of unbalanced data. In order to enhance performance, we also advise using CatBoost and XGBoost in addition to LightGBM. Using the XGBoost algorithm because of the quick training both in large data and the word "regularization," which overfitting is avoided by evaluating the complexity of it takes little time to set the tree up, and hyper parameters. Cat boost is another algorithm we employ. since it's unnecessary to change the hyper parameters for over fitting management, and it also produces effective outcomes. not modifying the hyper parameters in comparison to other algorithmic learning processes. We suggest an ensemble learning system with majority vote. Combination of Cat Boost, XG Boost, and LightGBM methodology and examine the impact of the combined techniques on the fraud detection performance on actual, unbalanced data. Additionally, we suggest utilizing deep learning for altering adjusting the hyper parameters, etc. • To assess the effectiveness of the suggested techniques, We conduct in-depth tests using data from the real world. In addition to the often used ROC-AUC, we also employ recall precision to better cover the unbalanced datasets. Additionally, we assess performance utilizing F1_score and MCC measures. The findings indicate that the suggested strategies perform better than the tried-and-true techniques. For we make use of publically accessible data sets and also public access to the source codes should be published by more scholars.

The main objective of this project is to predict credit card fraud detection by using different types of Machine learning and Deep learning Models like LightGBM, XG Boost, Cat Boost, Neural Network and Hybrid model like LG + XG+ CAT, LG + XG, LG + CAT, XG + CAT.

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.

Fraud detection in banking is considered a binary classification problem in which data is classified as legitimate or fraudulent [8]. Because banking data is large in volume and with datasets containing a large amount of transaction data, manually reviewing and finding patterns for fraudulent transactions is either impossible or takes a long time. Therefore, machine learning-based algorithms play a pivotal role in fraud detection and prediction [9]. Machine learning algorithms and high processing power increase the capability of handling large datasets and fraud detection in a more efficient manner. Machine learning algorithms and deep learning also provide fast and efficient solutions to real-time problems [10].

2. LITERATURE REVIEW

Ensemble Learning in Credit Card Fraud Detection Using Boosting Methods:

With the ceaseless thriving of the monetary market, MasterCard volume has forever been blasting these years. The extortion organizations are likewise rising quickly. Under this situation, extortion discovery has turned into an increasingly more significant issue. However, the extent of the misrepresentation is totally much lower than the virtuoso exchange, so the

unevenness dataset makes this issue significantly more testing. In this paper we principally advise how to adapt to the Visa misrepresentation identification issue by utilizing supporting strategies and furthermore gave a commitment of the concise examination between these helping techniques.

Ecommerce Fraud Detection through Fraud Islands and Multi-layer Machine Learning Model:

Principal challenge for web-based business exchange extortion counteraction is that misrepresentation designs are fairly unique and various. This paper presents two inventive techniques, extortion islands (interface investigation) and multi-facet AI model, which can actually handle the test of distinguishing assorted misrepresentation designs. Extortion Islands are framed utilizing join examination to explore the connections between various fake elements and to reveal the secret complex misrepresentation designs through the shaped organization. Multi-facet model is utilized to manage the generally assorted nature of misrepresentation designs. As of now, the extortion not entirely set in stone through various channels which are banks' declination choice, manual survey specialists' dismissal choices, banks' misrepresentation alarm and clients' chargeback demands. It tends to be sensibly accepted that different misrepresentation examples could be gotten however unique extortion risk anticipation powers (for example bank, manual audit group and misrepresentation AI model). The analyses showed that by incorporating not many different AI models which were prepared utilizing various sorts of misrepresentation marks, the exactness of extortion choices can be fundamentally moved along.

Detecting Credit Card Fraud Using Selected Machine Learning Algorithms:

Because of the gigantic development of internet business and expanded web based installment prospects, Visa misrepresentation has become profoundly significant worldwide issue. As of late, there has been significant interest for applying AI calculations as information digging method for charge card extortion location. Be that as it may, number of difficulties show up, for example, absence of freely accessible informational collections, exceptionally imbalanced class sizes, variation deceitful way of behaving and so on. In this paper we analyze execution of three AI calculations: Arbitrary Woodland, Backing Vector Machine and Strategic Relapse in recognizing extortion on genuine information containing Visa exchanges. To alleviate imbalanced class sizes, we use Destroyed examining technique. The issue of consistently changing misrepresentation designs is considered with utilizing gradual learning of chosen ML calculations in tests. The presentation of the methods is assessed in view of normally acknowledged measurement: accuracy and review.

Credit card fraud detection using AdaBoost and majority voting:

Visa extortion is a difficult issue in monetary administrations. Billions of dollars are lost because of charge card misrepresentation consistently. There is an absence of exploration concentrates on dissecting genuine Visa information inferable from secrecy issues. In this paper, AI calculations are utilized to identify charge card misrepresentation. Standard models are first utilized. Then, at that point, cross breed strategies which use Ada Boost and greater part casting a ballot techniques are applied. To assess the

model viability, an openly accessible charge card informational index is utilized. Then, at that point, a true charge card informational index from a monetary establishment is broke down. Also, clamor is added to the information tests to additionally survey the power of the calculations. The trial results decidedly demonstrate that the greater part casting a ballot strategy accomplishes great precision rates in recognizing misrepresentation cases in MasterCard.

Feature engineering strategies for credit card fraud detection:

Consistently billions of Euros are lost overall because of Visa misrepresentation. Subsequently, constraining monetary foundations to further develop their extortion recognition frameworks consistently. As of late, a few investigations have proposed the utilization of AI and information mining methods to resolve this issue. Be that as it may, most examinations utilized some kind of misclassification measure to assess the various arrangements, and don't consider the genuine monetary expenses related with the extortion recognition process. Besides, while developing a charge card extortion identification model, it is vital how to separate the right elements from the conditional information. This is typically finished by conglomerating the exchanges to notice the spending personal conduct standards of the clients. In this paper we grow the exchange conglomeration procedure, and propose to make another arrangement of elements in light of examining the occasional way of behaving of the hour of an exchange utilizing the von Mises conveyance. Then, utilizing a genuine Visa extortion dataset given by a huge European card handling organization, we look at cutting edge MasterCard misrepresentation location models, and assess what the various arrangements of elements have a mean for on the outcomes. By including the proposed occasional elements into the strategies, the outcomes show a typical expansion in reserve funds of 13%.

3. METHODOLOGY

In literature they developed a transaction aggregation strategy and created a new set of features based on the periodic behaviour analysis of the transaction time by using the von Mises distribution. In addition, they propose a new cost-based criterion for evaluating credit card fraud detection's models and then, using a real credit card dataset, examine how different feature sets affect results. More precisely, they extend the transaction aggregation strategy to create new offers based on an analysis of the periodic behaviour of transactions. In another study the application of machine learning algorithms to detect fraud in credit cards. They first use Naive Bayes, stochastic forest and decision trees, neural networks, linear regression (LR), and logistic regression, as well as support vector machine standard models, to evaluate the available datasets. Further, they propose a hybrid method by applying AdaBoost and majority voting. In addition, they add noise to the data samples for robustness evaluation. They perform experiments on publicly available datasets and show that majority voting is effective in detecting credit card fraud cases.

Drawbacks:

1. Other hyper parameter tuning methods like gridsearchcv and RandomizedSearchCV that spend more time to reach the highest accuracy model.

2. They used over sampling techniques to address the data imbalance.

3. They considered accuracy rather than precision-recall metrics for evaluating the model performance when the data is imbalance.

4. New set of features are created based on the periodic behavior analysis of the transaction time by using the von mises distribution.

5. Does not use any deep learning to fine-tune the hyperparameters, particularly the proposed weight-tuning one.

The proposed work in the paper is a fraud detection approach using machine learning techniques. We consider the use of class weight-tuning hyperparameters to control the weight of fraudulent and legitimate transactions. We use Bayesian optimization to optimize the hyperparameters while preserving practical issues such as unbalanced data. We propose weight-tuning as a pre-process for unbalanced data, as well as CatBoost and XGBoost to improve the performance of the LightGBM method by accounting for the voting mechanism. Finally, in order to improve performance even further, they use deep learning to fine-tune the hyperparameters, particularly the proposed weight-tuning one.Then examine the algorithms using different evaluation metrics, including accuracy, precision, recall, the Matthews correlation coefficient (MCC), the F1-score, and AUC diagrams

Benefits:

1. The proposed approach uses class weight-tuning hyper parameters to control the weight of fraudulent and legitimate transactions, which helps to address the problem of unbalanced data.

2. We use Bayesian optimization to optimize the hyper parameters, which helps to improve the performance of the model while preserving practical issues such as unbalanced data.

3. The proposed approach combines multiple machine learning techniques, including Cat Boost, LightGBM, XGBoost, and deep learning, which helps to improve the performance of the model compared to cutting-edge methods.

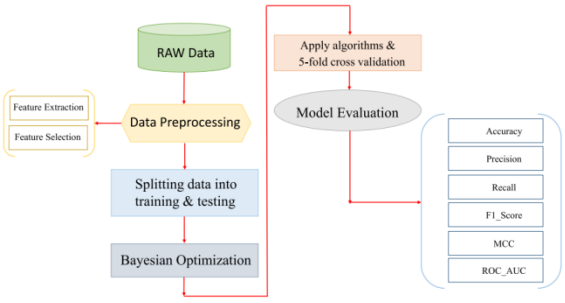

Fig 1 System Architecture

MODULES:

To implement aforementioned project we have designed following modules

- Data exploration: using this module we will load data into system
- Processing: Using the module we will read data for processing
- Splitting data into train & test: using this module data will be divided into train & test
- Model generation: Model building

Bayesian Optimization : LightGBM – XGBoost – CatBoost - Neural Network

CV Stratified Kfold : LightGBM – XGBoost – CatBoost - Neural Network

Smote Sampling (over and under sampling): LightGBM – XGBoost – CatBoost - Neural Network

Hyper parameter Tuning : LightGBM – XGBoost – CatBoost - Ensemble of LG + XG + CAT - Ensemble of LG + XG - Ensemble of XG + CAT - Ensemble of LG + CAT - Neural Network - Stacking Classifier (Gradient Boosting with RF + LightGBM).

- User signup & login: Using this module will get registration and login
- User input: Using this module will give input for prediction
- Prediction: final predicted displayed

Note: As an extension we applied an ensemble method combining the predictions of multiple individual models to produce a more robust and accurate final prediction.

However, we can further enhance the performance by exploring other ensemble techniques such as Stacking Classifier with RF + LightGBM With Gradient Boosting which got 100% accuracy.

4. IMPLEMENTATION

Here in this project we are used the following 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 StratifiedKfold: 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.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

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.

5. EXPERIMENTAL RESULTS

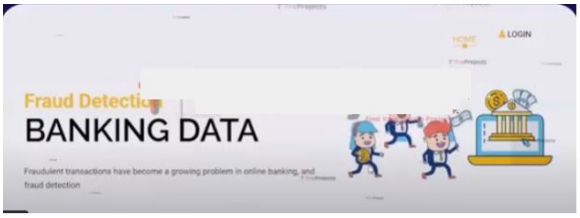

Fig 3 Home page

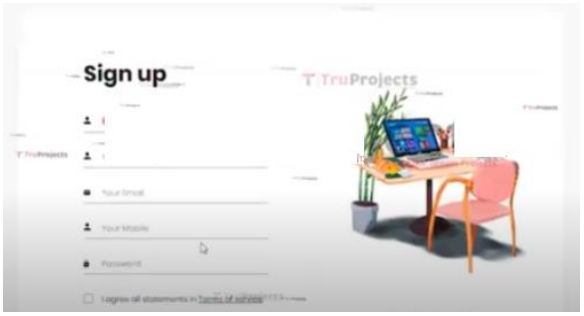

Fig 4 Signup page

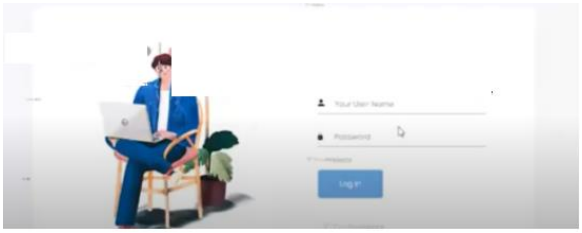

Fig 5 Signin page

Fig 6 User input page

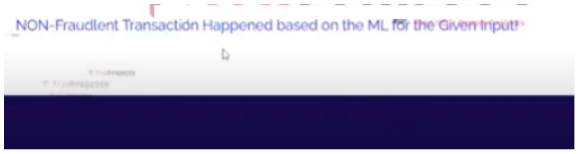

Fig 7 Prediction result

6. CONCLUSION

The conclusion of the paper 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, this paper 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. In this paper, by combining the LightGBM and XG Boost methods, we 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, we propose 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] 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.

[11] N. S. Halvaiee and M. K. Akbari, ‘‘A novel model for credit card fraud detection using artificial immune systems,’’ Appl. Soft Comput., vol. 24, pp. 40–49, Nov. 2014.

[12] A. C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, ‘‘Feature engineering strategies for credit card fraud detection,’’ Expert Syst. Appl., vol. 51, pp. 134–142, Jun. 2016.

[13] U. Porwal and S. Mukund, ‘‘Credit card fraud detection in e-commerce: An outlier detection approach,’’ 2018, arXiv:1811.02196.

[14] H. Wang, P. Zhu, X. Zou, and S. Qin, ‘‘An ensemble learning framework for credit card fraud detection based on training set partitioning and clustering,’’ in Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/ SCI), Oct. 2018, pp. 94–98.

[15] F. Itoo, M. Meenakshi, and S. Singh, ‘‘Comparison and analysis of logistic regression, Naïve Bayes and knn machine learning algorithms for credit card fraud detection,’’ Int. J. Inf. Technol., vol. 13, no. 4, pp. 1503–1511, 2021.

[16] T. A. Olowookere and O. S. Adewale, ‘‘A framework for detecting credit card fraud with cost-sensitive meta-learning ensemble approach,’’ Sci. Afr., vol. 8, Jul. 2020, Art. no. e00464.

[17] A. A. Taha and S. J. Malebary, ‘‘An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine,’’ IEEE Access, vol. 8, pp. 25579–25587, 2020.

[18] X. Kewei, B. Peng, Y. Jiang, and T. Lu, ‘‘A hybrid deep learning model for online fraud detection,’’ in Proc. IEEE Int. Conf. Consum. Electron. Comput. Eng. (ICCECE), Jan. 2021, pp. 431–434.

[19] T. Vairam, S. Sarathambekai, S. Bhavadharani, A. K. Dharshini, N. N. Sri, and T. Sen, ‘‘Evaluation of Naïve Bayes and voting classifier algorithm for credit card fraud detection,’’ in Proc. 8th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), Mar. 2022, pp. 602–608.

[20] P. Verma and P. Tyagi, ‘‘Analysis of supervised machine learning algorithms in the context of fraud

detection,'' ECS Trans., vol. 107, no. 1, p. 7189, 2022.