International Journal of Multidisciplinary Research and Growth Evaluation.

LLM-Enhanced XGBoost-Driven Fraud Detection and Classification Framework

Chao Li [1*], Jiarui Rao [2], Qian Zhang [3]
[1] Georgetown University, DC, USA
[2] Uber Technologies Inc., LA, USA
[3] The Chinese University of Hong Kong, HK

* Corresponding Author: **Chao Li**

Article Info

ISSN (online): 2582-7138
Volume: 06
Issue: 01
January-February 2025
Received: 20-12-2024
Accepted: 22-01-2025
Page No: 1987-1990

Abstract

In this research, the fraud detection model was trained using Matlab R2022a, with the dataset being randomly split into training and testing sets at a 7:3 ratio. Specifically, 70% of the data was allocated for training purposes, while the remaining 30% was used for testing. Upon incorporating the XGBoost model, the confusion matrix analysis indicated that a total of 6,326,133 instances were accurately predicted, with only 36,487 instances misclassified in the test set. This translates to an outstanding model performance, achieving a prediction accuracy of 99%. This demonstrates that the model exhibits robust performance and reliability in detecting fraudulent activities. Through this study, we have not only enhanced our capacity to identify diverse types of fraud but also provided valuable insights for future optimization of fraud detection techniques. The findings hold significant importance for safeguarding personal and organizational financial security and will contribute positively to the development of a more secure and reliable financial and network ecosystem.

DOI: https://doi.org/10.54660/.IJMRGE.2025.6.1.1987-1990

Keywords: XB Goost, fraud detection, confusion matrix

1. Introduction

Fraud detection, as an important area of research, aims to identify and prevent various forms of fraudulent activities, including credit card fraud, Internet fraud, insurance fraud and so on. With the popularity of e-commerce and online payment, fraudulent behaviours have become more hidden and complex, and traditional means are often difficult to deal with. Therefore, the use of machine learning algorithms for fraud detection has become one of the hot spots in current research [1, 2, 3, 4, 5].

Machine learning algorithms are vital in fraud detection for several reasons. Firstly, they can uncover hidden patterns and relationships within large datasets by analyzing vast amounts of information. These algorithms automatically extract relevant features and construct models to differentiate between legitimate transactions and potential fraud. Secondly, they possess the capability to continuously update and optimize model parameters in real-time, enhancing the accuracy and efficiency of fraud detection systems. Additionally, machine learning algorithms can swiftly adapt to emerging types of fraud and adjust their strategies accordingly to improve detection outcomes.

In practical applications, common machine learning algorithms include logistic regression, decision trees, support vector machines, and random forests. These algorithms can be tailored to specific scenarios, selecting the most suitable model for fraud detection and optimizing its performance through training data. Moreover, deep learning techniques, such as neural networks, have also gained significant traction in fraud detection, particularly when dealing with complex issues involving high-dimensional data and non-linear relationships [6, 7, 8, 9, 10, 11, 12, 13].

Machine learning algorithms are essential in fraud detection, and with ongoing optimization and innovation, it is anticipated that increasingly effective methods will be developed and applied in this field. These advancements will contribute significantly to creating a secure and reliable financial and e-commerce environment.

2. Data preprocessing and visualisation

Table 1: Partial text data

Type	Amount	Old balance Org	New Balance Orig	Old balance Dest	New balance Dest	Is Fraud
Payment	9839.64	170136	160296.36	0	0	0
Payment	1864.28	21249	19384.72	0	0	0
Transfer	181	181	0	0	0	1
Cash_out	181	181	0	21182	0	1
Payment	11668.14	41554	29885.86	0	0	0

This experiment uses the open source dataset, which does not contain any missing values, which ensures that our analyses can continue without the complexity and potential bias associated with dealing with missing data. Pie and bar charts were plotted to get a direct feel for the data and the results are shown in Figures 1 and 2.

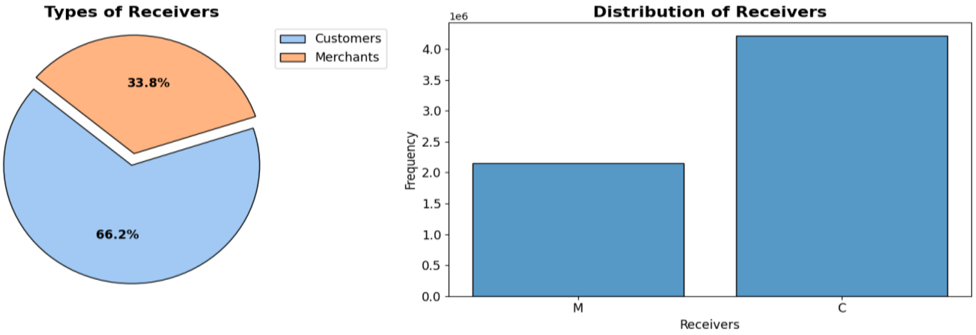

Photo credit: Original

Fig 1: Statistical analysis of data

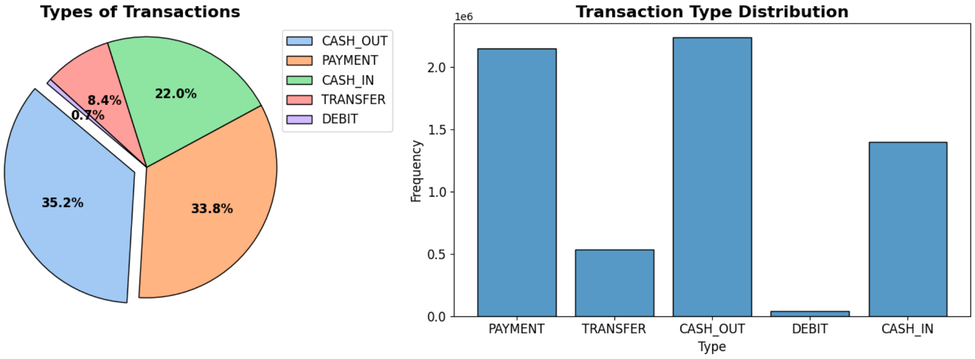

Photo credit: Original

Fig 2: Statistical analysis of data

XGBoost is a highly efficient ensemble machine learning algorithm that has gained widespread adoption across various tasks, including classification, regression, and ranking. It operates on the principles of Gradient Boosting, optimizing model performance by sequentially training multiple decision tree models and integrating regularization terms with loss functions.

To begin with, XGBoost leverages the core concept of Gradient Boosting, which involves iteratively training weak classifiers (decision trees) by focusing on the residuals of the preceding model. In each iteration, XGBoost computes the negative gradient of the loss function relative to the current model and fits a new decision tree to approximate this gradient. This process continues until a predefined number of iterations is reached or the loss function converges [14, 15, 16, 17, 18, 19].

Moreover, XGBoost incorporates regularization techniques to manage model complexity. It employs both L1 regularization (Lasso) and L2 regularization (Ridge) to effectively mitigate overfitting and enhance the model's generalization capabilities. Additionally, during the construction of each decision tree, XGBoost utilizes column and row sampling methods to reduce variance and improve model stability [20, 21, 22, 23, 24].

Another notable feature of XGBoost is its support for custom loss functions. This flexibility allows users to tailor the loss function to suit specific problem requirements, making XGBoost highly adaptable to a variety of machine learning tasks. This adaptability has contributed to its strong performance in numerous competitions.

Overall, XGBoost stands out for its ability to handle large-scale datasets and complex feature spaces. By combining gradient boosting, regularization, and feature sampling techniques, it achieves significant improvements in accuracy and efficiency. Moreover, XGBoost offers excellent interpretability and scalability, making it a highly respected algorithm in the machine learning community and a popular choice for real-world applications [25, 26, 27].

3. Method

In this experiment, Matlab R2022a is used for training, and the training and test sets are randomly divided in the ratio of 7:3, with 70% of the data used for training and 30% of the data used for testing. The upper limit of training times is set to 50 times, the initial learning rate is set to 0.01, the learning rate degradation factor is set to 0.1, and the prediction confusion matrix of the test set is recorded, as shown in Fig. 3. The prediction accuracy of the model test set is recorded and the results are shown in Table 2 [28, 29, 30, 31].

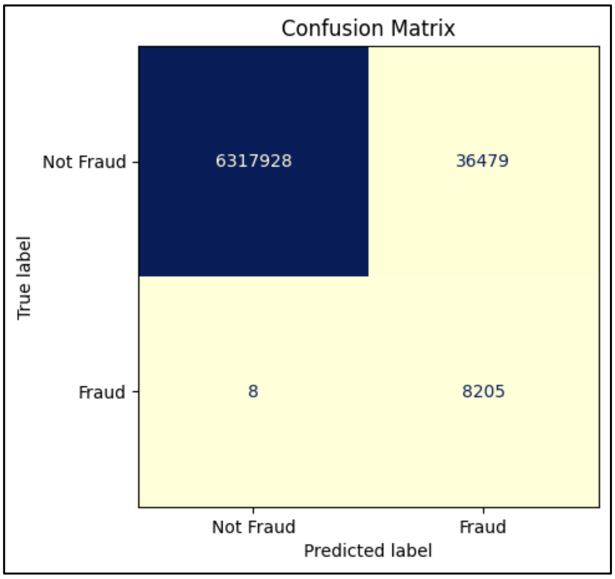

Photo credit: Original

Fig 3: Confusion matrix

Table 1: Partial text data

	Precision	Recall	F1-score	Support
0	1	0.99	1	6354407
1	0.18	1	0.31	8213
Accuracy			0.99	6362620
Macro avg	0.59	1	0.65	6362620
Weighted avg	1	0.99	1	6362620

From the confusion matrix, a total of 6,326,133 instances were predicted correctly and 36,487 instances were predicted incorrectly, with a prediction accuracy of 99%, the model is able to predict fraud detection well.

4. Conclusion

Fraud detection is a critical research domain that impacts social security and economic stability. By leveraging advanced technological tools, particularly machine learning algorithms like the XGBoost model, we can more effectively identify and prevent a wide range of fraudulent activities, such as credit card fraud, online fraud, and insurance fraud. In our experiment, we utilized Matlab R2022a for model training and randomly partitioned the dataset into training and testing sets at a 7:3 ratio, with 70% allocated for training and 30% for testing.

The results of our model training and testing were highly encouraging. The confusion matrix revealed that a total of 6,326,133 instances were accurately predicted, while only 36,487 instances were misclassified, achieving an exceptional prediction accuracy of 99%. This indicates that our model is highly effective in distinguishing fraudulent activities from legitimate transactions, demonstrating significant success in fraud detection.

In essence, by integrating the XGBoost model and utilizing large-scale datasets for training and validation, we have successfully developed an efficient and accurate fraud detection system. This system can assist financial institutions, e-commerce platforms, and other industries in promptly detecting and responding to potential risks. It also helps protect consumer rights and maintain market integrity. Therefore, in future research and practice, we should continue to explore advanced algorithms and methods and strive to optimize model performance to enhance the accuracy and efficiency of fraud detection.

Overall, this study has made a valuable contribution to the field of fraud detection and laid a strong foundation for creating a more equitable, transparent, and secure social environment. We hope that more researchers from related fields will join us in the future to collaboratively build a clearer and more secure digital world.

5. References

1. Fama EF, French KR. Common risk factors in the returns on stocks and bonds. J Financial Econ. 1993;33(1):3–56.
2. Xu Y, Shan X, Guo M, Gao W, Lin YS. Design and application of experience management tools from the perspective of customer perceived value: A study on the electric vehicle market. World Electr Veh J. 2024;15(8):378.
3. Chen M, Chen Y, Zhang Q. A review of energy consumption in the acquisition of bio-feedstock for microalgae biofuel production. Sustainability. 2021;13(16):8873.
4. Chen M, Chen Y, Zhang Q. Assessing global carbon sequestration and bioenergy potential from microalgae cultivation on marginal lands leveraging machine learning. Sci Total Environ. 2024;948:174462.
5. Zheng H, Wang B, Xiao M, Qin H, Wu Z, Tan L. Adaptive friction in deep learning: Enhancing optimizers with sigmoid and tanh function. In: 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS). IEEE; 2024. p. 809–13.
6. Chen M. Annual precipitation forecast of Guangzhou based on genetic algorithm and backpropagation neural network (GA-BP). In: Int Conf Algorithms, High Perform Comput, Artif Intell (AHPCAI 2021). SPIE; 2021 Dec. Vol. 12156, p. 182–6.
7. Zhang X, Soe AN, Dong S, Chen M, Wu M, Htwe T. Urban resilience through green roofing: A literature review on dual environmental benefits. In: E3S Web Conf. EDP Sci; 2024. Vol. 536, p. 01023.
8. Dong S, Xu T, Chen M. Solar radiation characteristics in Shanghai. J Phys Conf Ser. 2022 Oct;2351(1):012016.
9. Wang R, Behandish M. Surrogate modeling for physical systems with preserved properties and adjustable tradeoffs. arXiv preprint arXiv:2202.01139. 2022.
10. Zhang Q, Guan Y, Zhang Z, Dong S, Yuan T, Ruan Z, *et al*. Sustainable microalgae cultivation: A comprehensive review of open and enclosed systems for biofuel and high-value compound production. In: E3S Web Conf. EDP Sci; 2024. Vol. 577, p. 01008.
11. Wang R, Shapiro V. Topological semantics for lumped parameter systems modeling. Adv Eng Inform. 2019;42:100958.
12. Wang R. Consistency analysis between lumped and distributed parameter models [dissertation]. Madison (WI): Univ Wisconsin-Madison; 2021.
13. Yang R. CaseGPT: a case reasoning framework based on language models and retrieval-augmented generation. arXiv preprint arXiv:2407.07913. 2024.
14. Sun Y, Salami Pargoo N, Jin P, Ortiz J. Optimizing autonomous driving for safety: A human-centric approach with LLM-enhanced RLHF. In: Companion of the 2024 ACM Int Joint Conf Pervasive Ubiquitous Comput; 2024 Oct. p. 76–80.

15. Li K, *et al*. Exploring the impact of quantum computing on machine learning performance. 2024.
16. Wang Z, *et al*. Research on autonomous driving decision-making strategies based deep reinforcement learning. arXiv preprint arXiv:2408.03084. 2024.
17. Yan H, *et al*. Research on image generation optimization based deep learning. 2024.
18. Tang X, *et al*. Research on heterogeneous computation resource allocation based on data-driven method. arXiv preprint arXiv:2408.05671. 2024.
19. Su PC, *et al*. A mixed-heuristic quantum-inspired simplified swarm optimization algorithm for scheduling of real-time tasks in the multiprocessor system. Appl Soft Comput. 2022;131:109807.
20. Zhao Y, Hu B, Wang S. Prediction of Brent crude oil price based on LSTM model under the background of low-carbon transition. arXiv preprint arXiv:2409.12376. 2024.
21. Diao S, *et al*. Ventilator pressure prediction using recurrent neural network. arXiv preprint arXiv:2410.06552. 2024.
22. Zhao Q, Hao Y, Li X. Stock price prediction based on hybrid CNN-LSTM model. 2024.
23. Yin Z, Hu B, Chen S. Predicting employee turnover in the financial company: A comparative study of CatBoost and XGBoost models. 2024.
24. Xu Q, Wang T, Cai X. Energy market price forecasting and financial technology risk management based on generative AI. Preprints. 2024. Available from: https://doi.org/10.20944/preprints202410.2161.v1
25. Wu X, Xiao Y, Liu X. Multi-class classification of breast cancer gene expression using PCA and XGBoost. Preprints. 2024. Available from: https://doi.org/10.20944/preprints202410.1775.v2
26. Wang H, Zhang G, Zhao Y, Lai F, Cui W, Xue J, *et al*. RPF-ELD: regional prior fusion using early and late distillation for breast cancer recognition in ultrasound images. Preprints. 2024. Available from: https://doi.org/10.20944/preprints202411.1419.v1
27. Min L, Yu Q, Zhang Y, Zhang K, Hu Y. Financial prediction using DeepFM: Loan repayment with attention and hybrid loss. In: 2024 5th Int Conf Mach Learn Comput Appl (ICMLCA). IEEE; 2024 Oct. p. 440–3.
28. [No title provided] Accurate prediction of temperature indicators in Eastern China using a multi-scale CNN-LSTM-attention model.
29. Rao J, Zhang Q, Liu X. Applications analyzing e-commerce reviews with large language models (LLMs): A methodological exploration and application insight. J Artif Intell Gen Sci (JAIGS). 2024;7(1):207–12.
30. Zhang Q, *et al*. Sea MNF vs. LDA: Unveiling the power of short text mining in financial markets. Int J Eng Manag Res. 2024;14(5):76–82.
31. Rao J, *et al*. Machine learning in action: Topic-centric sentiment analysis and its applications. 2024.
32. Qian C, *et al*. WeatherDG: LLM-assisted procedural weather generation for domain-generalized semantic segmentation. arXiv preprint arXiv:2410.12075. 2024.