Article

Not peer-reviewed version

Application of Machine Learning Model in Fraud Identification: A Comparative Study of CatBoost, XGBoost and LightGBM

Yao Xiao , Li Tan , Jiaran Liu *

Posted Date: 17 March 2025

doi: 10.20944/preprints202503.1199.v1

Keywords: fraud detection; CatBoost; XGBoost; LightGBM; unbalanced data

 doi:10.20944/preprints202503.1199.v1

Article

Application of Machine Learning Model in Fraud Identification: A Comparative Study of CatBoost, XGBoost and LightGBM

Jiaran Liu [1,*], Yao Xiao [2] and Li Tan [3]

[1] School of Architecture and Urban Planning, Beijing University of Technology, Beijing, China
[2] University of Southern California, Los Angeles, CA, USA; yxiao243@marshall.usc.edu
[3] Fuqua School of Business, Duke University, Durham, NC, USA; litan@alumni.duke.edu
* Correspondence: liujiaran1223@163.com

Abstract. In the digital age, credit card fraud seriously affects financial stability and consumer trust, and machine learning technology provides a new way for its detection. In this study, the performance of three machine learning models, CatBoost, XGBoost and LightGBM, in credit card fraud detection is compared and analyzed by using a credit card transaction data set containing more than 1.85 million records. Through hierarchical K-fold cross-validation, with F1 score, accuracy rate and recall rate as evaluation indicators, the results show that the comprehensive performance of CatBoost model is the best, with F1 score of 0.9161, which is excellent in balancing and accurately identifying fraudulent transactions and detecting all fraudulent transactions. At the same time, the top 10 important characteristics that affect the model prediction are determined, such as transaction amount, cardholder age, urban population, etc. These characteristics provide a key basis for constructing and optimizing the fraud detection model. Based on the research results, it is suggested that financial institutions optimize the model with CatBoost as the core, expand the data feature dimension and strengthen real-time monitoring to improve the accuracy and timeliness of fraud detection. This study provides a valuable reference for the field of credit card fraud detection and helps to promote the development of financial security technology.

Keywords: fraud detection; CatBoost; XGBoost; LightGBM; unbalanced data

1. Introduction

In the contemporary digital age, the widespread adoption of credit cards has revolutionized the way people conduct financial transactions. Their convenience and flexibility have made them a preferred payment method for consumers worldwide. However, this surge in usage has also given rise to a significant concern: credit card fraud. Fraudsters are constantly devising sophisticated techniques to exploit vulnerabilities in the system, resulting in substantial financial losses for cardholders, financial institutions, and merchants alike.

The impact of credit card fraud extends far beyond the immediate financial implications. It erodes consumer trust in digital payment systems, undermines the stability of the financial sector, and can even lead to long-term negative consequences for individuals, such as damaged credit scores. As a result, developing effective fraud detection mechanisms has become an urgent necessity.

Machine learning, a subfield of artificial intelligence, has emerged as a powerful tool in the fight against credit card fraud. With its ability to analyze vast amounts of data and identify complex patterns, machine learning can outperform traditional rule-based methods. By leveraging algorithms like CatBoost, XGBoost, and LightGBM, it becomes possible to build highly accurate fraud detection models.

This paper aims to explore the application of these machine learning models in credit card fraud detection. Using a comprehensive Credit Card Transactions Dataset with over 1.85 million rows,

 doi:10.20944/preprints202503.1199.v1

which includes detailed information about transaction times, amounts, and associated personal and merchant details, we will analyze patterns in transaction amounts, locations, and user profiles. Our objectives are two-fold: first, to identify the most effective machine learning model for credit card fraud detection among CatBoost, XGBoost, and LightGBM; second, to determine the key features that significantly influence fraud detection, thereby enhancing the overall performance of fraud detection systems. Through this research, we hope to contribute to the development of more robust and reliable fraud prevention strategies in the credit card industry.

2. Literature Review

In recent years, the rapid expansion of financial transactions and evolving fraud techniques have made financial fraud detection a critical research area. The application of machine learning (ML) has provided new methods to enhance fraud detection accuracy and efficiency. This paper reviews existing ML approaches in financial fraud detection, analyzing their performance, challenges, and future research directions.

Ali et al. highlighted ML's ability to extract fraud patterns from large-scale historical data and enable real-time fraud detection [1]. Their study found that supervised learning methods, such as support vector machines, decision trees, and neural networks, perform best with labeled data, while unsupervised learning methods, including clustering and anomaly detection, are more suitable for unlabeled data. Additionally, ensemble learning (e.g., random forests, AdaBoost) and deep learning models have demonstrated superior performance in big data environments.

Hernandez Aros et al. emphasized deep learning's advantages in handling high-dimensional data and identifying complex fraud patterns [2]. They pointed out that training efficiency and algorithm optimization play crucial roles, particularly when dealing with large-scale financial transaction datasets.

Alsuwailem et al. investigated the effectiveness of various ML algorithms, revealing that traditional models like k-nearest neighbors, naïve Bayes, and logistic regression require extensive preprocessing but train quickly, making them suitable for small datasets [3]. In contrast, deep learning models such as deep neural networks (DNN), long short-term memory networks (LSTM), and convolutional neural networks (CNN) excel in detecting complex fraud patterns, particularly in financial statement and credit card fraud detection.

Obeng et al. explored ML's role in preventing fraud and securing financial transactions [4]. Their research emphasized that fraud detection models must not only improve accuracy but also ensure data privacy and interpretability, which are critical for real-world applications.

Despite these advancements, challenges remain. Ashtiani and Raahemi identified data imbalance as a major issue, where fraudulent transactions are significantly outnumbered by legitimate ones, affecting model performance [5]. Solutions such as oversampling, undersampling, and cost-sensitive learning have been proposed to mitigate this issue.

Furthermore, Whiting et al. pointed out that while ML models perform well in detecting known fraud patterns, they struggle with emerging fraud techniques [6]. Improving adaptability, particularly for real-time fraud detection, remains an important research focus.

Future studies should explore multimodal learning, transfer learning, and adaptive learning. Song et al. suggested that integrating time-series and textual data through multimodal learning can improve fraud detection accuracy [7]. With advancements in deep learning and computational resources, developing more sophisticated models for real-time fraud detection holds great potential.

3. Data Introduction

The Credit Card Transactions Dataset provides detailed records of credit card transactions, including information about transaction times, amounts, and associated personal and merchant details. This dataset has over 1.85M rows.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 March 2025 doi:10.20944/preprints202503.1199.v1

Use machine learning models to identify fraudulent transactions by examining patterns in transaction amounts, locations, and user profiles. Enhancing fraud detection systems becomes feasible by analyzing behavioral patterns.

Table 1 lists variables in the Credit Card Transactions Dataset. It includes transaction-time variables like "trans_date_trans_time", payment-related "amt", cardholder-identifying "cc_num", and demographic details such as "gender", "job". Merchant-related variables like "merchant", "category" are also there. Notably, "is_fraud" is the target variable for fraud detection models, which helps in discerning patterns among other variables to identify fraudulent transactions.

Table 1. Variables and descriptions.

variable	description
trans_date_trans_time	Timestamp of the transaction.
cc_num	Credit card number (hashed or anonymized).
merchant	Merchant or store where the transaction occurred.
category	Type of transaction (e.g., grocery, entertainment).
amt	Amount of the transaction.
gender	Gender of the cardholder.
City/state/zip	Address details of the cardholder.
Lat/long	Geographical coordinates of the transaction.
city_pop	Population of the city where the transaction occurred.
job	Occupation of the cardholder.
dob	Date of birth of the cardholder.
trans_num	Unique transaction number.
unix_time	Unix timestamp of the transaction.
merch_lat/merch_long	Geographical coordinates of the merchant.
is_fraud	Indicator of whether the transaction is fraudulent.
merch_zipcode	Geographical coordinates of the merchant.

Figure 1 shows the distribution of the number of fraudulent and normal transactions in different transaction categories. It can be observed that the number of normal transactions in various categories is generally large, while the number of fraudulent transactions is relatively small, indicating a significant class imbalance. For example, in the "misc_net" category, there are 62,372 normal transactions and only 915 fraudulent transactions; in the "grocery_pos" category, there are 121,895 normal transactions and 1,743 fraudulent transactions. This indicates that different transaction categories face different fraud risks. Although the number of fraudulent transactions in some categories such as "entertainment" and "gas_transport" is small, they still require special attention, providing a data basis for constructing fraud detection models for different transaction categories.

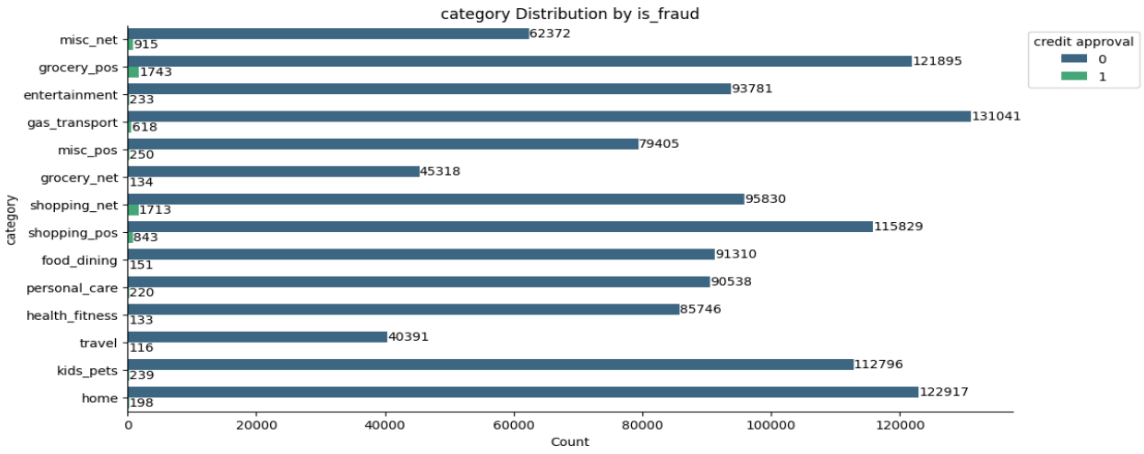

Figure 1. Category Distribution by is_fraud.

 doi:10.20944/preprints202503.1199.v1

Figure 2 presents the change trend of the number of fraudulent transactions from January 2019 to June 2020. It can be seen from the figure that the number of fraudulent transactions fluctuates in different months, without an obvious monotonic increase or decrease trend. For example, the number of fraudulent transactions in April 2019 was 592, then dropped to 527 in May, and rose to 506 in January 2020. This fluctuation may be related to various factors, such as changes in seasonal consumption habits and alterations in transaction patterns caused by marketing activities in different months. By observing the monthly data, it is helpful to discover the patterns of fraudulent transactions in the time series, providing a reference for setting dynamic thresholds in real-time fraud detection systems.

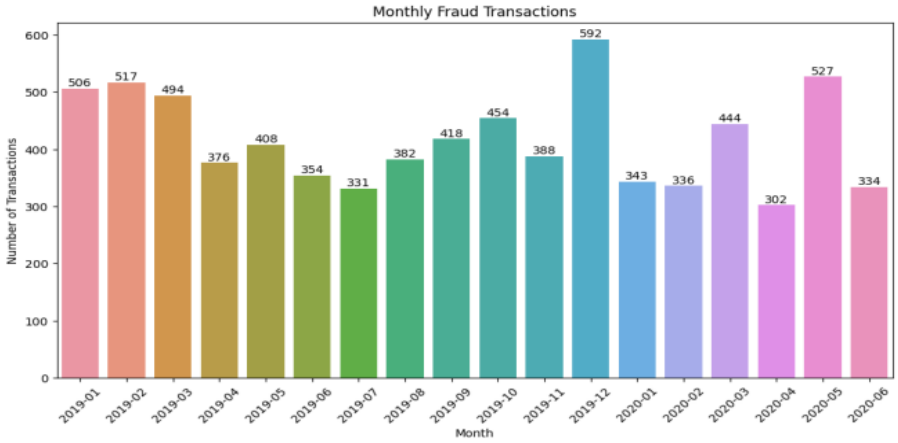

Figure 2. Monthly Fraud Transactions.

Figure 3 shows seasonal and weekly variations in fraudulent transactions from 2019-2020. Fraud peaked in autumn 2019 (1,615 cases) and was lowest in winter 2020 (334), likely due to increased holiday spending. Weekly trends show the highest fraud on Saturdays (175.3) and Sundays (173.7), while Wednesdays (122.7) and Tuesdays (133.6) had the least. Increased weekend fraud may result from higher transactions and reduced oversight, whereas weekdays see stricter monitoring and lower volumes. Understanding these patterns helps optimize fraud detection and prevention strategies.

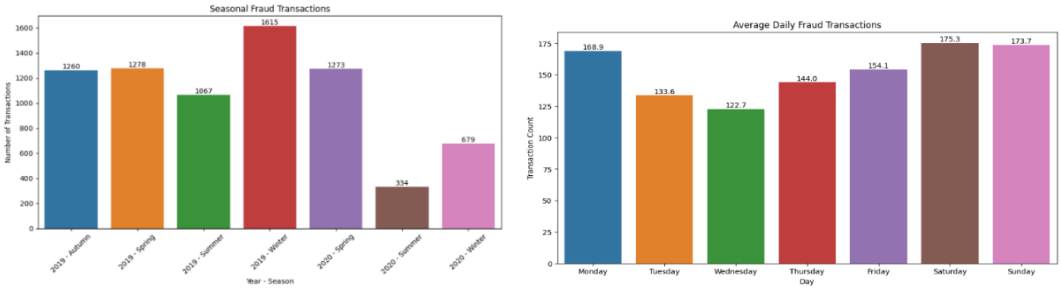

Figure 3. Seasonal Fraud Transactions and Average Daily Fraud Transactions.

The higher figures on weekends may be attributed to increased consumer activities like shopping and entertainment, offering more opportunities for fraudsters. Additionally, potentially reduced regulatory and auditing efforts on weekends might contribute to the higher occurrence of fraud. Conversely, lower fraud-transaction numbers on Tuesdays and Wednesdays could be related to decreased consumption and stricter supervision during weekdays. This information is of great practical significance for credit-card fraud prevention. Financial institutions and regulatory bodies can allocate resources more rationally based on these risk differences, enhancing fraud monitoring and real-time transaction reviews on weekends to more effectively combat fraud and mitigate risks for cardholders and institutions.

 doi:10.20944/preprints202503.1199.v1

4. Model Introduction

This study explores credit card fraud detection using CatBoost [8], XGBoost [10], and LightGBM [11], optimizing hyperparameters with Optuna to enhance performance. Given the dataset's class imbalance, class weights are assigned using the compute_class_weight function, ensuring fair consideration of minority classes in model training.

CatBoost [8], developed by Yandex, excels in handling categorical variables without extensive preprocessing, reducing computational costs and overfitting risks [9]. It effectively captures feature-target relationships and includes early stopping to prevent overfitting. XGBoost [10], known for its parallel tree construction, efficiently processes large datasets with many features, offering high accuracy and tunable hyperparameters. LightGBM [11], from Microsoft, employs a histogram-based algorithm and leaf-wise tree growth [12], improving speed and memory efficiency, making it ideal for large-scale datasets.

To evaluate model performance on imbalanced data, this study uses F1 Score, Precision, and Recall. The F1 Score balances precision (correct fraud predictions) and recall (detected actual fraud cases). Stratified k-fold cross-validation (k=5) ensures a robust evaluation by testing models on different folds and averaging performance metrics. This methodology provides a comprehensive assessment of model effectiveness in fraud detection while mitigating class imbalance issues.

5. Model Results Analysis

Table 2. Comparison of classification results of different models.

Model	Mean F1 Score	Mean Precision	Mean Recall
Catboost	0.9161	0.9338	0.8991
XGBoost	0.8926	0.8925	0.8928
LGBM	0.8812	0.8603	0.9032

This table compares the performance of CatBoost, XGBoost, and LGBM in credit card fraud detection. CatBoost achieves the best results, with an F1 score of 0.9161, precision of 0.9338, and recall of 0.8991, effectively balancing fraud detection and class imbalance. XGBoost follows with an F1 score of 0.8926, showing balanced precision (0.8925) and recall (0.8928) but slightly lower overall performance. LGBM scores 0.8812 in F1, with high recall (0.9032) but lower precision (0.8603), leading to more false positives. These results provide a quantitative basis for selecting the optimal model for fraud detection.

Table 3. Catboost model detailed result index.

category	Precision	Recall	F1-Score	Support
0	0.9995	0.9996	0.9995	257748
1	0.9319	0.9114	0.9215	1501
Accuracy	0.9991	0.9991	0.9991	0.9991
Macro Avg	0.9657	0.9555	0.9605	259249
Weighted Avg	0.9991	0.9991	0.9991	259249

The table evaluates the CatBoost model's performance in credit card fraud detection. For normal transactions (0), precision, recall, and F1 score are all around 0.9995, with 257,748 samples, indicating near-perfect classification. For fraudulent transactions (1), precision is 0.9319, recall 0.9114, and F1 score 0.9215 across 1,501 samples, showing reliable detection despite class imbalance. The overall accuracy reaches 0.9991, with a macro-average F1 score of 0.9605. These results confirm the CatBoost model's strong performance in accurately detecting fraud while balancing imbalanced class data.

 doi:10.20944/preprints202503.1199.v1

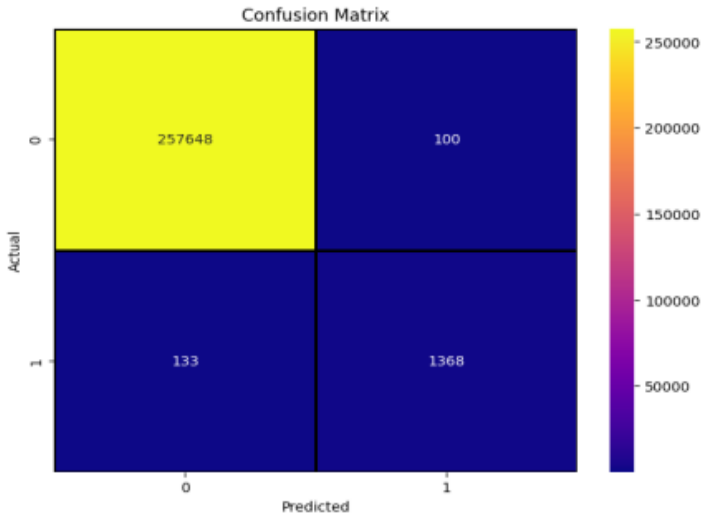

Figure 5. Average Daily Fraud Transactions.

Through the analysis of the tabular data, it can be seen that in credit card fraud detection, the Catboost model has relatively better comprehensive performance. Different models have their own characteristics in terms of precision and recall rate, providing important references for subsequent model improvement and selection, which is helpful for further optimizing the fraud detection system. Calculating and visualising the average feature importances from three machine learning models (CatBoost, XGBoost, and LGBM) to identify the top 10 most important features influencing the model predictions by creating a bar plot for better interpretation of which features are most impactful.

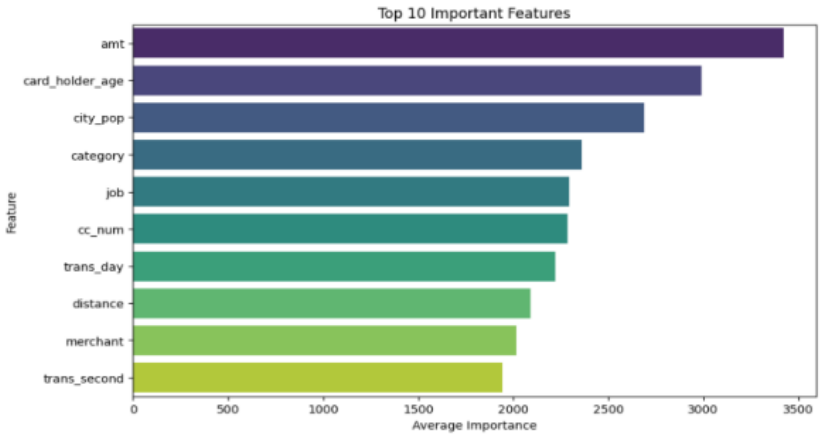

Figure 6. Top 10 Important Features.

The bar chart highlights the top 10 features in credit-card fraud detection. Transaction amount ranks highest, as abnormal amounts often indicate fraud. Card-holder age and city population follow, reflecting fraud risks across demographics. Transaction category and job influence detection, as spending patterns vary. Credit-card number aids fraud pattern recognition. Transaction day and distance impact detection, with unusual locations raising suspicion. Merchant information contributes, as fraud risks differ by business type. Lastly, transaction second suggests timing irregularities. These features play key roles in fraud detection, offering valuable insights for optimizing fraud prevention models.

 doi:10.20944/preprints202503.1199.v1

6. Conclusions

This project uses machine learning technology to detect credit card fraud, adopts three models: CatBoost, XGBoost and LightGBM, and constructs the final model through voting integration method. The final model performed well, with F1 score of 92%, accuracy of 93% and recall rate of 91%. This achievement shows that the machine learning model has significant application value in the field of credit card fraud detection, which can effectively identify fraudulent transactions and provide strong security for financial institutions and cardholders.

At the same time, the study also identified the top 10 important characteristics that affect the prediction of the model, including transaction amount, cardholder age, urban population, transaction category, occupation and so on. These characteristics play a key role in fraud detection. Abnormal transaction amount is often an important signal of fraud. The age of cardholders is different, and the fraud risk is also different. There is a certain relationship between urban population size and fraud mode. Certain transaction types are more prone to fraud, while occupation affects cardholders' consumption and fraud risk. In addition, factors such as credit card number (even if it is hashed or anonymized), transaction date, the distance between trading places and the cardholder's permanent residence, merchant information and transaction time (accurate to the second) are also of great significance to fraud detection. These findings provide a key basis for further constructing and optimizing the fraud detection model.

References

1. Ali A, Abd Razak S, Othman S H, et al. Financial fraud detection based on machine learning: a systematic literature review[J]. Applied Sciences, 2022, 12(19): 9637.
2. Hernandez Aros L, Bustamante Molano L X, Gutierrez-Portela F, et al. Financial fraud detection through the application of machine learning techniques: a literature review[J]. Humanities and Social Sciences Communications, 2024, 11(1): 1-22.
3. Alsuwailem A A S, Salem E, Saudagar A K J. Performance of different machine learning algorithms in detecting financial fraud[J]. Computational Economics, 2023, 62(4): 1631-1667.
4. Obeng S, Iyelolu T V, Akinsulire A A, et al. Utilizing machine learning algorithms to prevent financial fraud and ensure transaction security[J]. World Journal of Advanced Research and Reviews, 2024, 23(1): 1972-1980.
5. Ashtiani M N, Raahemi B. Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review[J]. Ieee Access, 2021, 10: 72504-72525.
6. Whiting D G, Hansen J V, McDonald J B, et al. Machine learning methods for detecting patterns of management fraud[J]. Computational Intelligence, 2012, 28(4): 505-527.
7. Song X P, Hu Z H, Du J G, et al. Application of machine learning methods to risk assessment of financial statement fraud: evidence from China[J]. Journal of Forecasting, 2014, 33(8): 611-626.
8. Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features[J]. Advances in neural information processing systems, 2018, 31.
9. Hancock J T, Khoshgoftaar T M. CatBoost for big data: an interdisciplinary review[J]. Journal of big data, 2020, 7(1): 94.
10. Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.
11. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30.
12. Wang D, Li L, Zhao D. Corporate finance risk prediction based on LightGBM[J]. Information Sciences, 2022, 602: 259-268.