Fraud Detection in Online Transactions: Toward Hybrid Supervised–Unsupervised Learning Pipelines

Shuo Xu , Yuchen Cao , Zhongyan Wang * , Yexin Tian

Posted Date: 14 May 2025

doi: 10.20944/preprints202505.1101.v1

Keywords: Fraud Detection; Supervised Learning and Unsupervised Learning; Class Imbalance; Anomaly Detection; LightGBM; K-Means Clustering Analysis; Online Transactions

Article

Fraud Detection in Online Transactions: Toward Hybrid Supervised–Unsupervised Learning Pipelines

Shuo Xu [1], Yuchen Cao [2], Zhongyan Wang [3,*] and Yexin Tian [4]

[1] Computer Science & Engineering Department, University of California San Diego, La Jolla, USA
[2] Khoury College of Computer Science, Northeastern University, Seattle, USA
[3] Center of Data Science, New York University, New York, USA
[4] College of Computing, Georgia Institute of Technology, Atlanta, USA
* Correspondence: wangzhongyan99@gmail.com

Abstract: Fraud detection in online transactions presents a challenging task due to the rarity of fraudulent events and the evolving nature of fraud strategies. This study presents a comparative analysis of three supervised machine learning models, Logistic Regression, Random Forest, and LightGBM, for detecting fraudulent transactions in an extremely imbalanced dataset. We evaluate each model under both standardized and raw feature preprocessing settings using macro-averaged metrics and AUC. Our findings show that ensemble-based models, particularly LightGBM, significantly outperform linear baselines and exhibit robustness to feature scaling. Additionally, we assess K-Means clustering as an unsupervised baseline, but observe that it fails to meaningfully separate fraud cases, suggesting the need for more informative features or hybrid learning approaches. These results offer practical insights into model selection, preprocessing, and the trade-offs between precision and recall in real-world fraud detection systems.

Keywords: fraud detection; supervised learning and unsupervised learning; class imbalance; anomaly detection; lightgbm; k-means clustering analysis; online transactions

I. Introduction

Fraudulent activity poses a persistent threat across sectors such as finance, insurance, and healthcare, leading to substantial economic and reputational losses. In the financial sector alone, billions of dollars are lost annually due to fraud, prompting increased investment in intelligent detection systems [1,2]. Traditional rule-based approaches, which rely on predefined heuristics and manual reviews, often fail to detect emerging fraud patterns and suffer from high false-positive rates, making them increasingly insufficient in today's dynamic fraud landscape [1].

In response, machine learning (ML) has become a leading paradigm for fraud detection, offering the ability to automatically learn complex behavioral patterns from transaction data. Supervised learning methods, such as Logistic Regression and tree-based models, including bagging (e.g., random forest or RF) and boosting (e.g., gradient boosting machine or GBM, or its light version, lightGBM) techniques, have shown strong performance when sufficient labeled data is available. However, they often struggle with class imbalance, as fraudulent transactions typically represent only a small fraction of the dataset [1]. Furthermore, supervised models can be slow to adapt to novel fraud strategies without ongoing labeling and retraining [3].

On the other hand, unsupervised learning methods, such as clustering algorithms (e.g., K-means) and anomaly detection techniques like Isolation Forest, do not require labeled data and can potentially detect previously unseen fraud patterns. Yet these methods often suffer from limitations in precision, interpretability, and consistency, particularly in high-dimensional or noisy feature spaces.

While both supervised and unsupervised approaches offer unique advantages, each alone is insufficient for building truly adaptive and robust fraud detection systems. This observation

 doi:10.20944/preprints202505.1101.v1

motivates our central hypothesis: a hybrid learning strategy that combines supervised and unsupervised methods may offer improved detection performance, adaptability, and resilience to changing fraud behaviors.

We present a study with both supervised and unsupervised ML models for fraud detection using a publicly available online transaction dataset. Our key contributions are as follows:

- Comparative Evaluation: We benchmark supervised and unsupervised models using precision, recall, F1-score, and AUC-ROC.
- Insight into Trade-offs: We analyze model performance, scalability, and interpretability, identifying when each method is most appropriate.
- Toward Hybrid Methods: We propose a hybrid perspective that leverages the strengths of both paradigms to address real-world constraints such as limited labels and evolving fraud tactics.

The rest of the paper is structured as follows: Section 2 introduces the dataset, preprocessing, and model setup; Section 3 presents performance results and analysis; Section 4 discusses implications, limitations, and the hybrid model direction; and Section 5 concludes with recommendations and future work.

II. Methods

A. Dataset and Preprocessing

We used a publicly available online payment transaction dataset consisting of over 630,000 records, where each entry represents one single transaction. The employed dataset includes both account-level features (e.g., sender and recipient balances pre and post a transaction) and transaction-level attributes (e.g., amount, type of transaction). A binary label indicates whether a transaction was identified as fraudulent.

Fraudulent transactions constitute only about 0.1% of the dataset, reflecting a highly imbalanced distribution typical of real-world fraud detection tasks. To improve signal quality, we focused on transaction types most associated with fraudulent behavior, such as peer-to-peer transfers and cash-outs.

For data preprocessing, categorical transaction types were encoded into binary indicators, and numerical features were either left in their original scale or standardized using z-score normalization. Specifically, we conducted two sets of analyses, one with raw features and another with standardized features, to evaluate the effect of scaling on model performance for this particular dataset. Additionally, we engineered simple derived features based on account balance changes to better capture transactional anomalies.

The dataset was split into 60% training, 20% validation, and 20% testing sets using stratified sampling to maintain class proportions. While no cross-validation was used, the validation set served for model tuning and early evaluation. To address the class imbalance, oversampling of the minority class was applied during training, allowing supervised models to better learn from rare fraudulent instances.

B. Supervised Learning Models

We evaluated three representative supervised learning models to assess their performance in detecting fraudulent transactions: logistic regression, RF, and LightGBM. These three models were selected to capture a range of trade-offs between interpretability (Logistic Regression), robustness to feature interactions (RF), and and scalability with strong performance on imbalanced datasets (LightGBM).

Logistic Regression

Logistic Regression [4] serves as a baseline model due to its simplicity and transparency. It models the probability of fraud as a logistic function of a linear combination of input features. Despite

 doi:10.20944/preprints202505.1101.v1

its limitations in capturing nonlinear patterns, it provides interpretable coefficients that reflect the direction and strength of each feature's association with fraudulent labels [5].

Random Forest

RF is an ensemble (i.e., bagging) method that builds multiple decision trees using bootstrapped samples and aggregates their outputs via majority voting [6]. It captures nonlinear interactions and provides built-in feature importance measures, making it effective for datasets with mixed types and variable interactions.

Light Gradient Boosting Machine

LightGBM is a boosting framework optimized for efficiency and performance [7,8]. It grows trees leaf-wise and employes histogram-based splitting, allowing it to handle large-scale and imbalanced datasets effectively. Its ability to tune class weights and custom loss functions makes it particularly suitable for fraud detection tasks with severe class imbalance.

All supervised models were trained on the labeled training set, and hyperparameters were tuned using the validation set. No resampling techniques were applied; models were evaluated directly on the imbalanced data to reflect real-world conditions.

C. Unsupervised Learning Models

As an unsupervised baseline, we implemented K-Means [9] clustering to explore the structure of transactions without relying on labels. K-Means was selected as a widely used and computationally efficient clustering algorithm, suitable for exploring latent group structures in high-volume transaction data. The optimal number of clusters (i.e., K) was determined using the elbow method [10], which identifies the point at which increasing k produces lower returns in within-cluster variance.

K-Means was applied to both raw and standardized feature sets across experimental conditions. The resulting clusters were later analyzed to assess the distribution of fraudulent and legitimate transactions, enabling evaluation of potential separability in an unsupervised setting.

D. Evaluation Metrics

Given the inherent class imbalance in fraud detection, conventional accuracy can be misleading, as models predicting only the majority class may still appear highly accurate. To provide a more balanced assessment, we evaluated model performance with 4 widely used metrics: precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

For supervised models, precision measures the proportion of predicted fraudulent transactions that are truly fraudulent, while recall indicates the proportion of actual fraud cases correctly identified [11]. The F1-score, as the harmonic mean of precision and recall, offers a balanced view of performance on imbalanced datasets [11]. AUC-ROC was used to assess overall discriminative ability by capturing the trade-off between true and false positive rates across decision thresholds [12,13].

Unsupervised evaluation presents additional challenges due to the absence of direct class labels in training. To address this, we applied standard metrics based on known fraud labels in the test set. We report precision at each k, which measures the proportion of actual fraud cases among the top k transactions ranked by anomaly score [9]. Additionally, ROC curves were constructed using anomaly score rankings to quantify how well the unsupervised model distinguishes fraud from legitimate behavior [12,14].

In summary, we adopted a consistent evaluation framework based on standard classification metrics, allowing direct comparison between supervised and unsupervised approaches under class imbalance conditions.

III. Results

A. Performance of Supervised Learing

 doi:10.20944/preprints202505.1101.v1

To evaluate the effectiveness of different supervised learning models for fraud detection, we compared logistic regression, RF, and LightGBM under two types of feature preprocessing strategies: raw and standardized features. Given the extremely imbalanced class distribution in our dataset, traditional accuracy was insufficient; instead, we focused on macro-averaged metrics to ensure both classes were equally represented in evaluation.

Table 1 reports macro-averaged precision, recall, and F1-score, along with AUC, for each model-feature combination. With only 82 fraudulent transactions out of 63,626 total (~0.13%), macro averages provide a more balanced and representative evaluation across both classes.

Table 1. Supervised Model Performance.

Model	Feature Handling	Macro Average			AUC
		Precision	Recall	F1-Score	
Logistic Regression	Raw	1.00	0.96	0.84	0.76
	Standardized	1.00	0.52	0.55	0.52
Random Forest	Raw	0.97	0.83	0.89	0.83
	Standardized	0.94	0.86	0.89	0.86
LightGBM	Raw	0.94	0.86	0.89	0.86
	Standardized	1.00	0.83	0.90	0.83

Overall, RF and LightGBM (i.e. the tree-based models) consistently outperformed Logistic Regression. While Logistic Regression showed reasonable performance when trained on raw features, its performance dropped substantially after feature standardization—particularly in recall and F1-score. In contrast, both tree-based models (RF and LightGBM) showed strong and stable performance across both feature preprocessing strategies, indicating greater robustness to preprocessing and stronger capability to model non-linear feature interactions.

Among the three models, LightGBM achieved the highest macro F1-score under both preprocessing conditions, highlighting its effectiveness in handling imbalanced and structured transaction data.

We also observed that precision consistently exceeded recall across all models. This is typical in rare-event classification problems like fraud detection, where models tend to be conservative in predicting the minority class. High precision reflects a strong ability to avoid false positives—when fraud is predicted, it's usually correct. However, this often comes at the cost of lower recall, meaning a significant number of actual frauds are missed. Without explicit adjustments to training objectives or threshold tuning, models naturally bias toward the majority class, prioritizing certainty over completeness in identifying fraudulent cases.

B. Unsupervised Clustering Analysis

In practice, unsupervised learning is commonly used in industry for anomaly detection, particularly when labeled data is limited or expensive to obtain. One potential application is the early identification of fraudulent transactions, which are assumed to deviate from normal behavioral patterns.

To explore this direction, we implemented K-Means clustering with $k = 5$, selected using the elbow method. The model was trained on standardized transaction features, and cluster assignments were evaluated against known fraud labels.

As shown in Table 2, the fraud rate across clusters remained nearly uniform, ranging from 0.000997 to 0.001307, with two clusters containing no fraud cases at all. This suggests that K-Means did not uncover feature groupings that effectively separate fraudulent from legitimate transactions.

One possible explanation is that the features available in this dataset do not provide sufficient information to differentiate fraud in an unsupervised setting. Unlike supervised models that can

 doi:10.20944/preprints202505.1101.v1

learn task-specific boundaries from labeled examples, clustering relies solely on the underlying feature distributions—making it less effective when the signal-to-noise ratio is low or fraud patterns are subtle and diffuse.

These findings emphasize the limitations of clustering-based approaches for fraud detection, particularly when the dataset lacks meaningful features that are strongly associated with fraudulent behavior. Without such informative variables to distinguish normal and anomalous patterns, unsupervised methods like K-Means are unlikely to form clusters that correspond to fraud outcomes, highlighting the need for richer, domain-specific features when pursuing unsupervised anomaly detection.

C. Model Interpretation and Feature Importance

All three supervised models evaluated in this study, Logistic Regression, RF, and LightGBM, provide mechanisms for interpreting model behavior through feature importance, albeit with different methodologies. In particular, logistic regression derives importance from model coefficients, allowing not only magnitude but also directional interpretation: positive or negative coefficients increase or decrease the likelihood of a fraud label, respectively. RF uses impurity-based importance, ranking features based on how often and how effectively they reduce uncertainty when used for splits across decision trees. LightGBM reports gain-based importance, quantifying the cumulative improvement in model performance contributed by each feature throughout the boosting process.

Despite these methodological differences, the three models consistently ranked the post- and pre-transaction balances of the sender's account, highlighting that abnormal or inconsistent balance changes are key indicators of fraud. Interestingly, Logistic regression revealed opposite directional effects for these two features: a lower post-transaction balance was associated with higher fraud likelihood, while a higher pre-transaction balance also increased the odds of fraud. Other relevant features included transaction amount, destination balances, and certain transaction types, though their relative importance varied. Categorical and frequency-based variables contributed less across all models.

These findings suggest that sender balance–related features are robust indicators of fraud, and that combining different interpretability metrics provides complementary insights to enhance model transparency and practical decision-making.

IV. Discussion

A. Interpretation of Findings

This study systematically evaluated three supervised models, logistic regression, RF, and LightGBM, on a highly imbalanced fraud detection dataset. The results revealed that ensemble methods (i.e., RF and LightGBM) substantially outperformed Logistic Regression, especially in recall and F1-score, indicating their greater ability to capture non-linear and rare patterns typical of fraudulent activity.

Logistic regression, while interpretable, showed sensitivity to feature preprocessing and failed to generalize well under standardized inputs. On the other hand, RF and LightGBM demonstrated robust performance across both standardized and raw features, suggesting greater flexibility in adapting to varying data characteristics. Among the two, LightGBM achieved the best overall performance, making it particularly well-suited for real-world fraud detection pipelines.

We also observed a consistent pattern where precision exceeded recall, a common outcome in rare-event classification tasks. This indicates a conservative prediction strategy, where models minimize false positives but may overlook some true frauds. Addressing this trade-off may require threshold adjustment, cost-sensitive learning, or alternative loss functions better aligned with recall.

Finally, we explored unsupervised fraud detection using K-Means clustering. The clustering results showed near-identical fraud rates across all clusters, suggesting that the available features

 doi:10.20944/preprints202505.1101.v1

were insufficient for meaningful fraud separation in an unsupervised setting. This highlights the critical importance of feature quality and domain relevance when deploying clustering-based or anomaly detection approaches.

B. Practical Implications

Our findings offer several practical takeaways:

- Logistic Regression still remains useful in regulated environments where model transparency is paramount. However, its performance degrades significantly when fraud patterns are complex or when feature scaling is not carefully managed.
- Tree-based models, particularly LightGBM, offer a strong balance of accuracy, scalability, and interpretability. LightGBM's fast training and robust performance under class imbalance make it a strong candidate for integration into real-time fraud detection systems.

For industry applications without sufficient labeled data, unsupervised methods may still play a role—but only if supported by stronger feature engineering or hybrid learning strategies.

C. Limitations and Future Works

This study has several limitations:

- The dataset used is synthetic and publicly available, which may limit the generalizability to real-world transaction environments.
- Although we included an unsupervised baseline (K-Means), the evaluation was limited to one clustering algorithm. Future work should explore advanced unsupervised or hybrid models, such as Isolation Forest, autoencoders, or semi-supervised anomaly detection.
- Our model evaluation did not include cost-sensitive metrics or business-driven thresholds, which are often crucial in real-world fraud detection and it could be a future direction.
- Lastly, the models evaluated were static. In practice, fraud patterns evolve rapidly. Online learning or continual learning frameworks should be investigated to maintain long-term model effectiveness.

In addition, the emergence of Large Language Models (LLMs) presents new opportunities for fraud detection [15], particularly in scenarios involving text-based transaction metadata, customer communication logs, or behavioral patterns embedded in unstructured data [16,17]. Future research could explore how LLMs can be fine-tuned or integrated with structured fraud detection pipelines to enhance anomaly detection and context-aware classification.

Reference

1. R. K. Rao and V. N. Mandhala, "Unveiling financial fraud: A comprehensive review of machine learning and data mining techniques," Intelligent Systems and Informatics, vol. 29, no. 6, pp. 2309–2334, Dec. 2024.
2. P. Li, M. Abouelenien, R. Mihalcea, Z. Ding, Q. Yang, and Y. Zhou, "Deception detection from linguistic and physiological data streams using bimodal convolutional neural networks," in *Proc. 2024 5th Int. Conf. Inf. Sci., Parallel Distrib. Syst. (ISPDS)*, 2024, pp. 263–267
3. Y. Liu, X. Shen, Y. Zhang, Z. Wang, Y. Tian, J. Dai, and Y. Cao, "A systematic review of machine learning approaches for detecting deceptive activities on social media: Methods, challenges, and biases," arXiv preprint arXiv:2410.20293, 2024.
4. D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed., New York, NY: John Wiley & Sons, Inc., 2000.
5. Y. Zhang, Z. Wang, Z. Ding, Y. Tian, J. Dai, X. Shen, Y. Liu, and Y. Cao, "Tutorial on using machine learning and deep learning models for mental illness detection," *arXiv preprint* arXiv:2502.04342, 2025.
6. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
7. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

 doi:10.20944/preprints202505.1101.v1

8. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS)*, pp. 3149–3157, 2017.
9. J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.
10. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., New York, NY: Springer, 2009.
11. D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
12. J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. on Machine Learning (ICML)*, pp. 233–240, 2006.
13. Z. Ding, Z. Wang, Y. Zhang, Y. Cao, Y. Liu, X. Shen, Y. Tian, and J. Dai, "Efficient or powerful? Trade-offs between machine learning and deep learning for mental illness detection on social media," *arXiv preprint* arXiv:2503.01082, 2025.
14. Y. Cao, J. Dai, Z. Wang, Y. Zhang, X. Shen, Y. Liu, and Y. Tian, "Machine learning approaches for depression detection on social media: A systematic review of biases and methodological challenges," *Journal of Behavioral Data Science*, vol. 5, no. 1, Feb. 2025.
15. Y. Tao, Y. Shen, H. Zhang, Y. Shen, L. Wang, C. Shi, and S. Du, "Robustness of Large Language Models Against Adversarial Attacks," *arXiv preprint* arXiv:2412.17011, 2024.
16. Y. Shen, L. Wang, C. Shi, S. Du, Y. Tao, Y. Shen, and H. Zhang, "Comparative Analysis of Listwise Reranking with Large Language Models in Limited-Resource Language Contexts," *arXiv preprint* arXiv:2412.20061, 2025.
17. Y. Shen, H. Zhang, Y. Shen, L. Wang, C. Shi, S. Du, and Y. Tao, "AltGen: AI-Driven Alt Text Generation for Enhancing EPUB Accessibility," *arXiv preprint* arXiv:2501.00113, 2024.