

Research on Multi-Model Prediction of Financial Transaction Fraud Risk Based on Machine Learning

Guanzhou Shang^{1*}

¹ School of Business, University of Shanghai for Science & Technology, Shanghai, China

*Correspondence: Guanzhou Shang

Received: 12 May 2026; **Accepted:** 3 July 2026; **Published:** 3 August 2026

Abstract:

Amid the deep integration of digital finance and payment services, financial transaction fraud risks show traits of organized crime, intelligent operations, and covert methods. Fraud prevention is crucial for financial institutions' sustainable operations, making precise and efficient fraud prediction models a must. To tackle challenges like scarce fraud data, imbalanced data categories, and feature desensitization, this study uses machine learning algorithms. First, the K - Means algorithm stratifies transaction risks, and the Apriori algorithm finds risk feature combinations. Then, it compares the fraud prediction performance of four models: LR, RF, GBDT, and XGBoost. Using the 6th Shanghai University Open Big Data Analysis Challenge credit card fraud dataset, it evaluates model metrics and suitability for real - time risk management. Results show ensemble tree models are more accurate than traditional linear models, enhancing fraud detection; gradient boosting models reduce false - positive costs, and XGBoost offers balanced detection. The combination of K - Means and Apriori provides business insights. This study supports building intelligent risk control systems and offers a reference for predicting financial transaction fraud risks.

Keywords:

Machine Learning, Financial Transactions, Fraud Risk Prediction, K-Means Clustering, Apriori Association Rules, Financial Risk Control

1. Statement of the Problem

With the deepening market-oriented reforms in digital finance and the transformation of the global payment system, China's non-cash and online payment services have experienced explosive growth. In 2023, the country recorded over 350 billion non-cash payment transactions, while online payment transactions surpassed one trillion for the first time. As transaction volumes expand, payment chains have lengthened with increased participation from multiple entities, elevating financial fraud risks. Statistics show that global financial fraud cases exceeded 200 million in 2023, causing direct economic losses exceeding \$50 billion, while domestic banking systems received nearly 60,000 new daily leads on cybercrime activities. More critically, financial fraud has evolved from simple account impersonation to sophisticated, coordinated operations involving organized criminal groups. Criminals

employ various tactics to evade traditional risk controls, simulating legitimate transactions to test risk thresholds—putting conventional rule-based systems in a dilemma: overly stringent rules may inadvertently harm users and compromise payment experiences and business conversion rates, while overly lenient rules risk overlooking high-risk transactions and triggering losses or compliance violations. Against this backdrop, transitioning from rule-driven to data-driven intelligent risk control has become an urgent priority for financial security, with developing precise, efficient, and interpretable fraud prediction models at the core of financial institutions' digital transformation efforts.

Scholars both domestically and internationally have conducted extensive research on machine learning methods for financial fraud risk identification and credit default prediction. In unsupervised learning and rule mining, studies have demonstrated that the K-Means clustering algorithm effectively identifies high-risk transaction clusters, while the Apriori association rule algorithm extracts characteristic pattern combinations of fraudulent behaviors, providing interpretable hard rules for risk control systems. For supervised learning predictions, Qu Zhongfeng et al. developed a credit risk prediction model using logistic regression (LR), offering quantitative support for credit strategy optimization; Song employed the random forest (RF) algorithm to construct a personal credit assessment model, enabling accurate prediction of default tendencies in online lending; Chi Guotai et al. built a default prediction model for listed companies using the XGBoost algorithm, validating the strong adaptability of gradient boosting algorithms in financial risk forecasting scenarios.

Overall, existing research has covered mainstream machine learning algorithms in financial risk prediction, providing a solid theoretical foundation and technical framework for this study. However, three critical challenges remain unresolved: First, insufficient data scenario adaptability. Current models are primarily built on large-scale, evenly distributed credit datasets, whereas fraudulent transactions typically account for less than 0.2% of real-world financial activities—a classic “low-probability, high-risk” scenario. Moreover, core features are often subjected to PCA desensitization due to privacy compliance requirements, lacking clear business semantics, which leads to issues such as overfitting, weak generalization capabilities, and missed fraud detection. Second, incomplete research workflows. Most studies focus solely on improving model accuracy without establishing a comprehensive analytical cycle encompassing “statistical pattern mining – risk identification – association rule extraction – model optimization – strategy implementation.” This results in models lacking interpretable business rationale, making them inadequate for real-world risk control requirements. Third, lack of business-driven model selection and optimization. Current research often indiscriminately compares multiple algorithms without addressing the core business challenge of financial risk control—balancing both false-negative and false-positive rates—thereby failing to achieve targeted model optimization and validation.

2. Data and Methods

2.1. Data Description and Preprocessing

The experimental data in this study are sourced from the Credit Card Fraud Detection Dataset, publicly available on the 6th Shanghai University Open Big Data Analysis Challenge. This dataset serves as a classic benchmark for addressing

imbalanced classification challenges in financial risk management. All core features have undergone PCA (Principal Component Analysis) desensitization, fully complying with privacy compliance requirements for real-world financial scenarios. The original dataset contains 284,807 transaction records; after quality verification, it exhibits no missing values, all features are numerical, and the data integrity is robust. In terms of category distribution, the dataset contains only 492 fraudulent transactions versus 284,315 legitimate transactions, yielding a fraud rate of merely 0.1727%—a hallmark of the “small sample size and high imbalance” characteristic, closely mirroring the distribution patterns observed in real financial fraud scenarios. The specific field composition of the dataset is presented in Table 1.

Table 1. Dataset Variable Description.

Variable type	Variable	Variable Meaning	Variable Processing and Description
	Time	Transaction Time Interval	Number of seconds elapsed since the first transaction in the dataset
Input Variable	V1~V28	PCA-Anonymized Features	Features anonymized by Principal Component Analysis(PCA)
	Amount	Transaction Amount	Amount Per Transaction
Output Variable	Class	Transaction Fraud Label	Target variable: 0 indicates a normal transaction and 1 indicates a fraudulent transaction

To meet model training requirements and mitigate the adverse effects of class imbalance and feature dimensionality differences, this paper conducts systematic preprocessing and feature engineering on the raw data. The specific steps fully align with the empirical analysis conclusions, as outlined below:

(1) Time feature derivation: Convert the Time field into an Hour feature to capture risk patterns during trading hours. The calculation formula is:

$$Hour = \left\lfloor \frac{Time}{3600} \right\rfloor \bmod 24$$

Based on the statistical analysis findings, a binary feature Is_High_Risk_Hour was constructed for high-risk periods: it takes a value of 1 when trading hours fall between 2:00 and 4:00 AM, and 0 otherwise.

(2) Amount Feature Processing: Apply a logarithmic transformation $Amount_Log = \log(1 + Amount)$ to the transaction amounts to construct a distribution that mitigates the right-skewed long-tail distribution effect. Simultaneously, divide the amounts into nine intervals (0–1, 1–10, 10–50, 50–100, 100–500, 500–1000, 1000–5000, 5000–10000, >10000), thereby discretizing the amount features and establishing a foundation for association rule mining.

(3) Feature Discretization and Interaction Construction: For each of the six core features, calculate the 20th percentile and create a binary variable indicating low values—assigning 1 if the feature value falls below this percentile, otherwise 0. Based on the results of Apriori association rule mining, combine multiple low-value indicators to construct interaction features, such as $Rule_V17_V14_Low = V17_Low \times V14_Low$, to detect strong fraud signals when multiple features occur simultaneously.

In terms of data partitioning, to ensure the fairness of model evaluation and the reliability of results, this study employs a hierarchical sampling method to divide the dataset in a 7:3 ratio into a training set (70%) and a test set (30%). During this process,

the ratio of fraudulent to normal samples remains consistent with that of the original dataset, ensuring uniform data distribution between the training and test sets. The final training set contains 199,364 transactions (including 345 fraudulent transactions), while the test set comprises 85,443 transactions (including 148 fraudulent transactions).

2.2. Core Algorithm Principle

The research framework of this paper encompasses three major modules: unsupervised clustering, association rule mining, and supervised learning prediction. All core algorithms are strictly confined to the research scope, with their specific principles and formulas detailed as follows:

(1) K-Means clustering algorithm

This paper employs the K-Means clustering algorithm to perform unsupervised segmentation of transaction groups [1], identifying high-risk transaction clusters with shared behavioral characteristics. As a classic partitioning-based unsupervised clustering algorithm, its core principle involves dividing samples into K clusters such that intra-cluster similarity is maximized while inter-cluster similarity is minimized, achieved through iterative optimization of the clustering loss function by minimizing cluster centroids.

The objective function of the algorithm (cluster inertia) is:

$$J = \sum_{k=1}^K \sum_{i \in C_k} ||x_i - \mu_k||^2$$

Here, K denotes the preset number of clusters, C_k represents the k-th cluster, X_i is the feature vector of the i-th sample, and μ_k is the centroid of the k-th cluster.

The core execution steps of the algorithm are: (1) Randomly select K samples as initial cluster centers; (2) Calculate the Euclidean distance from each sample to each center and assign samples to the nearest cluster; (3) Recalculate the cluster center (the mean of all samples within the cluster); Repeat steps (2)–(3) until the cluster centers stabilize or the maximum number of iterations is reached, completing the clustering process. Based on business insights and empirical results, this study sets the number of clusters $K = 5$, ultimately achieving risk stratification of the transaction population.

(2) Apriori Association Rule Mining Algorithm

This article employs the Apriori algorithm to mine combinations of strongly correlated features of fraudulent transactions, extracts directly implementable hard rules for risk control, and provides feature enhancement and interpretability support for predictive models[2]. As a classical association rule mining algorithm, it operates on the fundamental principle that “every non-empty subset of a frequent item set must itself constitute a frequent item set.” Through iterative layer-by-layer exploration, the algorithm identifies frequent item sets from the data and subsequently generates robust association rules based on these sets.

This paper focuses on association rules with “fraudulent transactions (Class=1)” as the consequent variable. The calculation formula for the core evaluation metric is as follows:

a. Support: The frequency of an item set appearing in the dataset, measuring the general applicability of the rule. Formula:

$$\text{Support}(X \rightarrow Y) = \frac{\text{count}(X \cap Y)}{N}$$

Here, X denotes the rule antecedent (feature combination), Y represents the rule consequent (Class=1, fraud), count(X∩Y) indicates the number of samples satisfying both antecedent and consequent conditions, and N stands for the total sample size.

b. Confidence: The conditional probability of the consequent occurring given the occurrence of the antecedent, measuring the predictive reliability of the rule. The formula is:

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{count}(X \cap Y)}{\text{count}(X)}$$

c. Lift: The ratio of a rule's confidence level to the posterior baseline probability, measuring the extent to which the rule enhances its predictive power against fraud. The formula is:

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Confidence}(X \rightarrow Y)}{P(Y)}$$

Here, P(Y) represents the baseline probability of fraudulent samples (the fraud rate across the entire dataset). A support value greater than 1 indicates that the rule possesses positive predictive power; the higher the value, the stronger the rule's fraud prediction capability. The algorithm parameters used in this study are set as follows: minimum support 0.0001, minimum confidence 0.05, and maximum rule length 3. This process ultimately yielded strong association rules with fraud as the consequent attribute.

(3) Supervised Learning Prediction Model

This paper systematically compares the fraud prediction performance of four mainstream machine learning models. All models are classic benchmark models in the field of financial risk control, with their core principles and formulas as follows:

a. Logistic Regression (LR)

Logistic regression is a classic linear binary classification model in financial risk control. By combining linear scores with the Sigmoid activation function, it outputs the fraud probability of samples within the (0,1) interval, offering strong interpretability and serving as the benchmark model for risk control rule modeling[3].

The core formula of the model is:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\omega^T X + b)}}$$

Here, X denotes the input feature vector, ω represents the feature weight vector, b is the bias term, and P(Y=1|X) indicates the prediction probability that a sample constitutes a fraudulent transaction.

This paper introduces L2 regularization to construct the model objective function, mitigating the risk of overfitting in small-sample imbalanced scenarios. The objective function is as follows:

$$J(\omega) = -\frac{1}{N} \sum_{i=1}^N [y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)] + \lambda \|\omega\|_2^2$$

Here, N denotes the sample size, y_i represents $\hat{y}_i$ the true label of the sample, p_i is the model's predicted probability, and λ is the regularization coefficient. The objective function is iteratively optimized using gradient descent to determine the optimal model parameters.

b. Random Forest (RF)

Random Forest is a tree model built on the Bagging parallel integration framework. By randomly selecting samples and features, it constructs multiple independent decision trees and ultimately combines their outputs through a majority-voting mechanism, effectively reducing the overfitting risk of a single tree while demonstrating excellent generalization and noise resistance capabilities[4].

The node splitting of a single decision tree is based on the Gini coefficient, which measures the sample impurity of a node, with the formula as follows:

$$Gini = 1 - \sum_{k=1}^2 p_k^2$$

Here, P_k represents the proportion of samples of the k -th category (fraud/normal) within a node. A smaller Gini coefficient indicates higher sample purity in the node and better splitting efficiency.

The final prediction result of the random forest is:

$$\hat{y} = \text{mode}\{T_1(X), T_2(X), \dots, T_M(X)\}$$

Here, M denotes the number of decision trees, $T_m(X)$ represents the prediction result of the m -th decision tree, and $\text{mode}\{\cdot\}$ is the majority voting function. In this study, the number of decision trees M is set to 100, and the core parameters were optimized using 5-fold cross-validation.

c. Gradient Boosting Decision Tree (GBDT)

The Gradient Boosting Tree is a tree model built upon the Boosting serial integration framework, commonly referred to as Gradient Boosting in this paper [5]. Its core concept involves sequentially training a series of shallow decision trees (weak learners), where each new tree aims to fit the negative gradient (residual) of the previous round's model predictions. By progressively reducing these prediction residuals, the model's overall performance is continuously enhanced.

The model's final prediction result is the weighted sum of all base learners, expressed by the formula:

$$F_M(x) = \sum_{m=1}^M \gamma_m T(x; \theta_m)$$

Here, M denotes the number of $T(x; \theta_m)$ decision trees; γ_m represents the m -th decision tree, with its parameters and learning rate denoted as α_m and β_m , respectively.

Using the squared error loss function as an example, the negative gradient (residual) for the m -th iteration is calculated as follows:

$$r_{im} = - \left[\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \right] = y_i - F_{m-1}(x_i)$$

Among them, $L(\cdot)$ is the loss function, F_{m-1} represents the prediction results of the model in the first $m-1$ rounds, and r_{im} denotes the residual to be fitted in the m -th round. By iteratively fitting the residuals, the model's prediction accuracy continues to improve.

d. Extreme Gradient Boosting (XGBoost)

XGBoost represents an engineered enhancement of Gradient Boosting Trees (GBDT), incorporating multiple optimizations over the traditional GBDT framework: it introduces a second-order Taylor expansion in the loss function and utilizes gradients and the Hessian matrix to guide tree growth, thereby improving fitting accuracy; it incorporates L1 and L2 regularization terms to control tree complexity and effectively mitigate overfitting; and it supports parallel computing, significantly boosting training efficiency. XGBoost is currently one of the most widely adopted models in the financial risk management field[6].

The objective function of XGBoost is:

$$Obj(\phi) = \sum_{i=1}^N l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

The first term $l(y_i, \hat{y}_i)$ is a differentiable loss function that measures the deviation between the model's predicted values and the actual values; the second term is a regularization term that controls the complexity of the tree, with the formula as follows:

$$\Omega(f_k) = \gamma T + \frac{\lambda}{2} ||w||^2$$

Here, T denotes the number of leaf nodes, w represents the weight of leaf nodes, γ is the penalty coefficient for leaf node splitting, and λ is the L2 regularization coefficient.

2.3. Model Evaluation Metrics

Given the binary classification nature of financial transaction fraud prediction tasks and the highly imbalanced data characteristics, this paper abandons a single overall accuracy metric in favor of a multi-dimensional evaluation framework. All core metrics align with the critical business requirement in financial risk control—balancing both true-negative and false-positive detection sensitivities.

First, define the core elements of the binary classification confusion matrix: True Positives (TP) are samples that are actually fraudulent and predicted as fraudulent by the model; False Positives (FP) are samples that are actually normal but predicted as fraudulent by the model; False Negatives (FN) are samples that are actually fraudulent but predicted as normal by the model; True Negatives (TN) are samples that are actually normal and predicted as normal by the model. The calculation formulas and explanations for the core evaluation metrics are as follows:

(1) Accuracy: The proportion of correctly predicted samples relative to the total sample size, reflecting the model's overall classification accuracy. The formula is:

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

(1) Precision: The proportion of samples predicted as fraudulent by the model that are actually fraudulent, reflecting the model's false-negative rate for legitimate users. A higher precision indicates a lower false-positive rate and less impact on user experience. Formula:

$$Precision = \frac{TP}{TP + FP}$$

(3) Recall: The proportion of samples actually fraudulent that are correctly identified by the model, reflecting the model's ability to detect fraudulent risks. A higher recall rate indicates a lower false-negative rate and stronger control over financial losses. The formula is:

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

(4) F1 Score: The harmonic mean of precision and recall, which comprehensively reflects the model's balance between these two metrics. A value closer to 1 indicates superior overall performance in risk control scenarios. The formula is:

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

(5) AUC: The area under the ROC curve, plotted with the false positive rate (FPR) on the x-axis and the true positive rate (TPR) on the y-axis. A higher AUC value closer to 1 indicates stronger overall discrimination capability of the model between fraudulent and normal samples, serving as the primary evaluation metric in imbalanced classification scenarios.

3. Empirical Analysis and Results

3.1. Descriptive Statistics and Identification of Fraud Characteristics

This paper first compares the distribution differences between fraudulent transactions and normal transactions across three dimensions—time, amount, and core characteristics—to identify the fundamental patterns of fraudulent behavior, thereby providing a foundational basis for subsequent feature engineering, cluster analysis, and rule mining.

Time Dimension Characteristics: Normal trading strictly follows human daily routines, with trading volume peaking between 8:00 and 20:00 during the day and declining significantly between 0:00 and 06:00 at night. In contrast, fraudulent transactions do not decrease alongside the overall trading volume between 02:00 and 04:00; instead, they exhibit abnormal activity, with the fraud rate exceeding 0.35% during this period—more than twice the daily average fraud rate of 0.1727%. The peak fraud incidence occurs between 02:00 and 03:00, reaching 0.38%. This conclusion indicates that the 02:00–04:00 period represents a high-risk window for fraudulent activities and should be prioritized in risk control strategies.

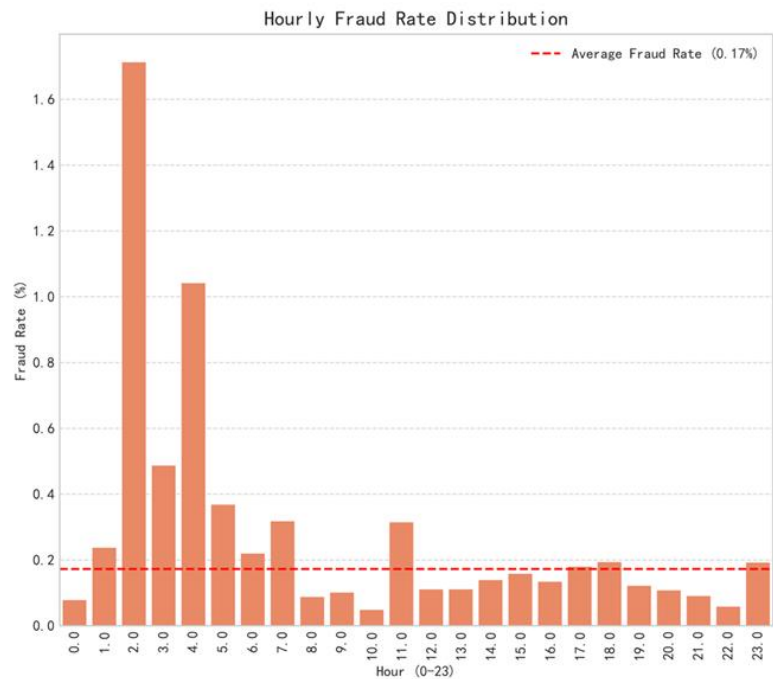
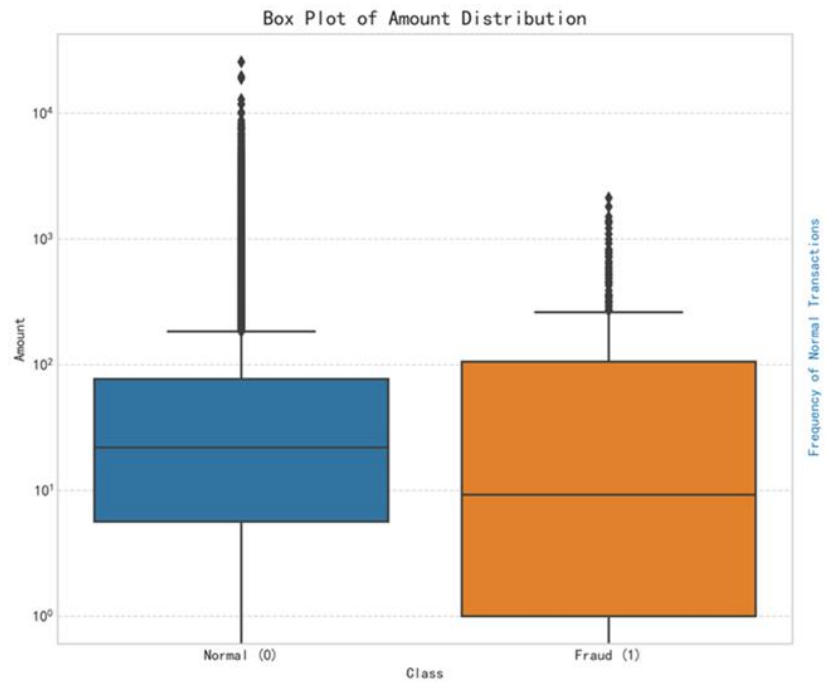


Figure 1. Hourly Distribution of Fraud Rate.

Amount Dimension Characteristics: The average amount of fraudulent transactions was 122.21 yuan, higher than the 88.29 yuan for normal transactions. However, this average was driven by a small number of large-value fraudulent transactions, with over 80% of fraudulent transactions concentrated in the small-value range of 0–100 yuan. The fraud rates for the 0–1 yuan and 1–10 yuan ranges were above 0.4% and 0.2%, respectively, significantly higher than those for the large-value range, while the fraud rate for transactions exceeding 5,000 yuan was extremely low. This characteristic reflects the typical fraudulent strategy of using small-scale probing transactions to circumvent risk control thresholds, indicating that small-value transactions represent a high-risk zone for fraud.



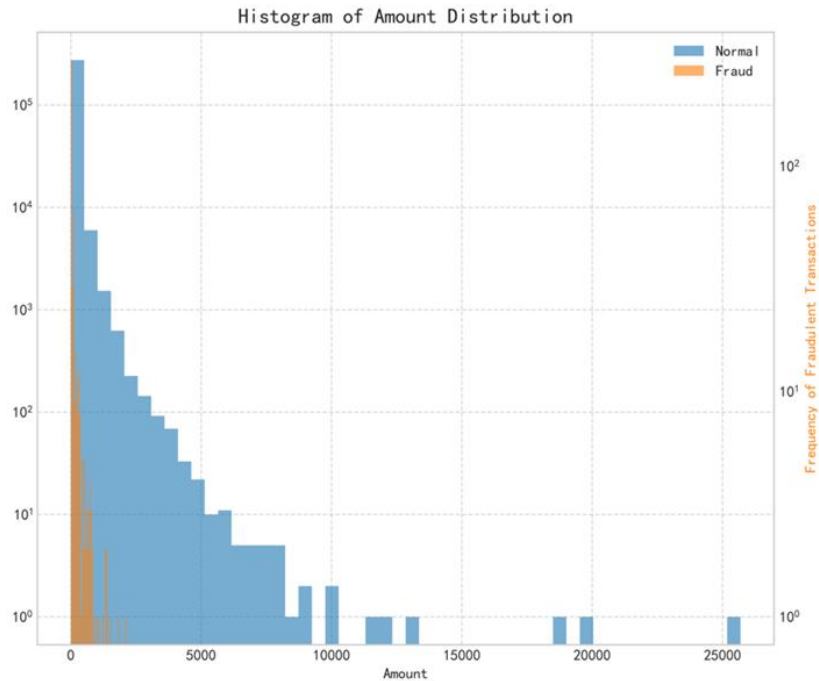


Figure 2. Features of the Amount Dimension.

Core Feature Dimensions: Through Pearson correlation analysis, six features with the highest correlation to fraud detection were identified, ranked by descending absolute correlation coefficients as V17, V14, V12, V10, V16, and V3, all exceeding 0.2—a significantly higher value than other features.

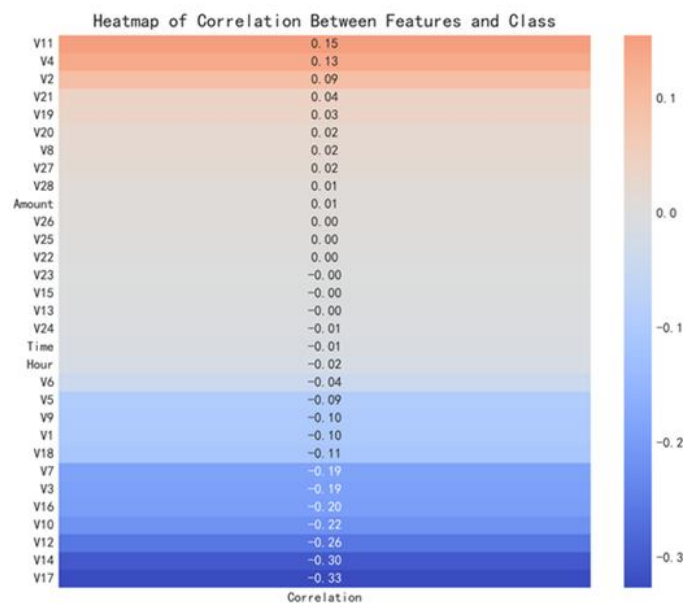


Figure 3. Heatmap showing the correlation between each feature and the Class.

The comparison of nuclear feature distributions reveals that these six features exhibit significant distribution deviations between fraudulent and normal samples, with the median of normal samples approaching zero; V17 shows a bimodal distribution in fraudulent samples versus a unimodal distribution in normal samples. These results demonstrate that the six PCA-processed features contain the core information for distinguishing fraudulent from normal transactions and serve as the primary input variables for subsequent analysis and modeling.

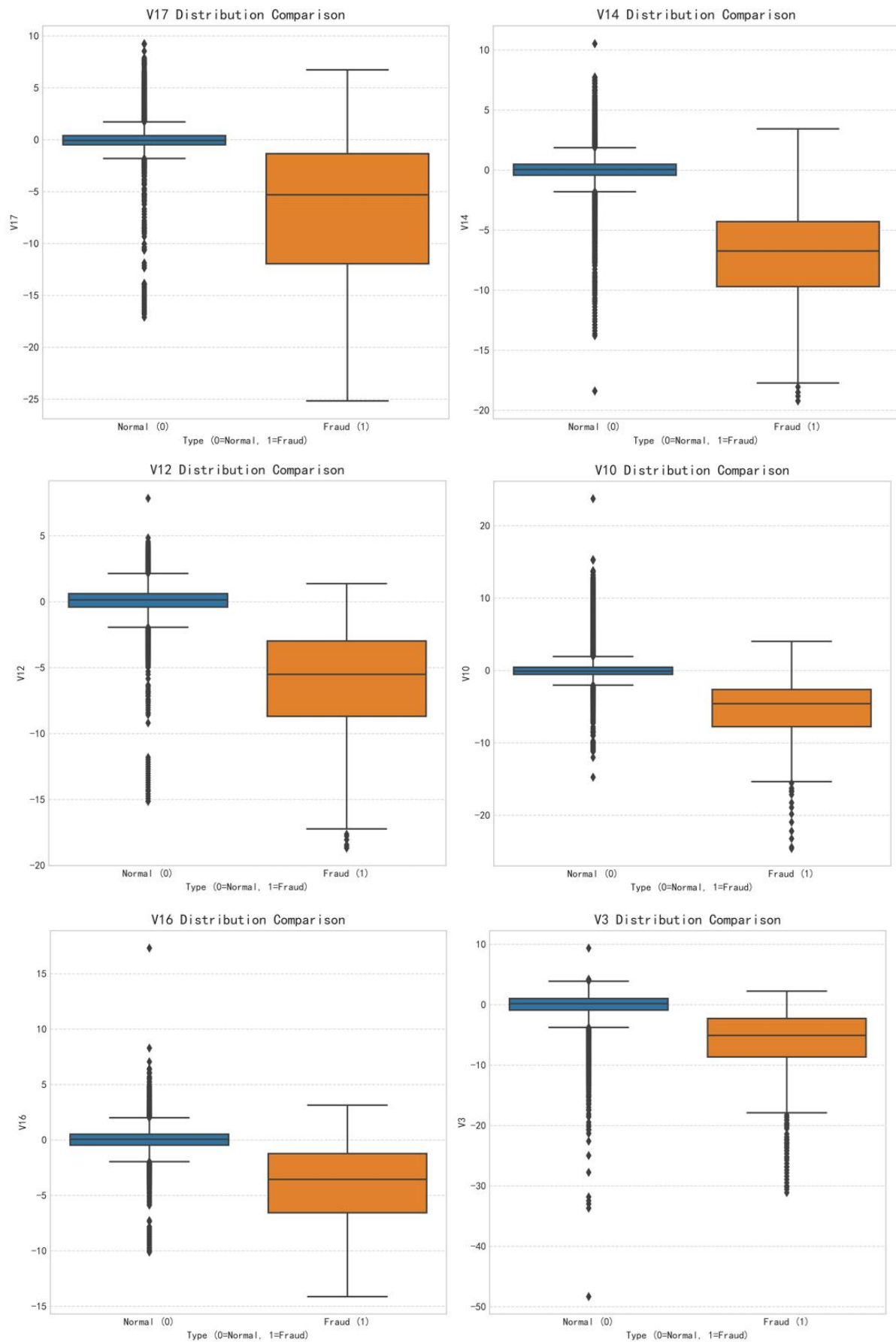


Figure 4. Key Feature Box Plot.

3.2. Results of K-Means Clustering Analysis

Based on the core screening criteria mentioned above, combined with time and amount-derived features, this study employs the K-Means algorithm to categorize all 284,807 transactions into five behavioral pattern groups. The sample distribution and fraud rate statistics for each group are presented in Table 2.

Table 2. K-Means Cluster Statistical Results.

Cluster Number	Sample Size	Sample Proportion	Fraudulent Samples	Fraud Rate
Cluster 0	110316	38.73%	87	0.0789%
Cluster 1	97584	34.26%	68-	0.0697%
Cluster 2	8954	3.14%	24	0.2680%
Cluster 3	67654	23.75%	56	0.0828%
Cluster 4	299	0.10%	257	85.95%

The cluster analysis successfully identified Cluster 4 as an extremely high-risk group, accounting for only 0.10% of the total sample but containing 257 fraudulent transactions—representing 52.2% of all fraudulent samples. The fraud rate within this cluster reached 85.95%, significantly exceeding both other clusters and the overall average.

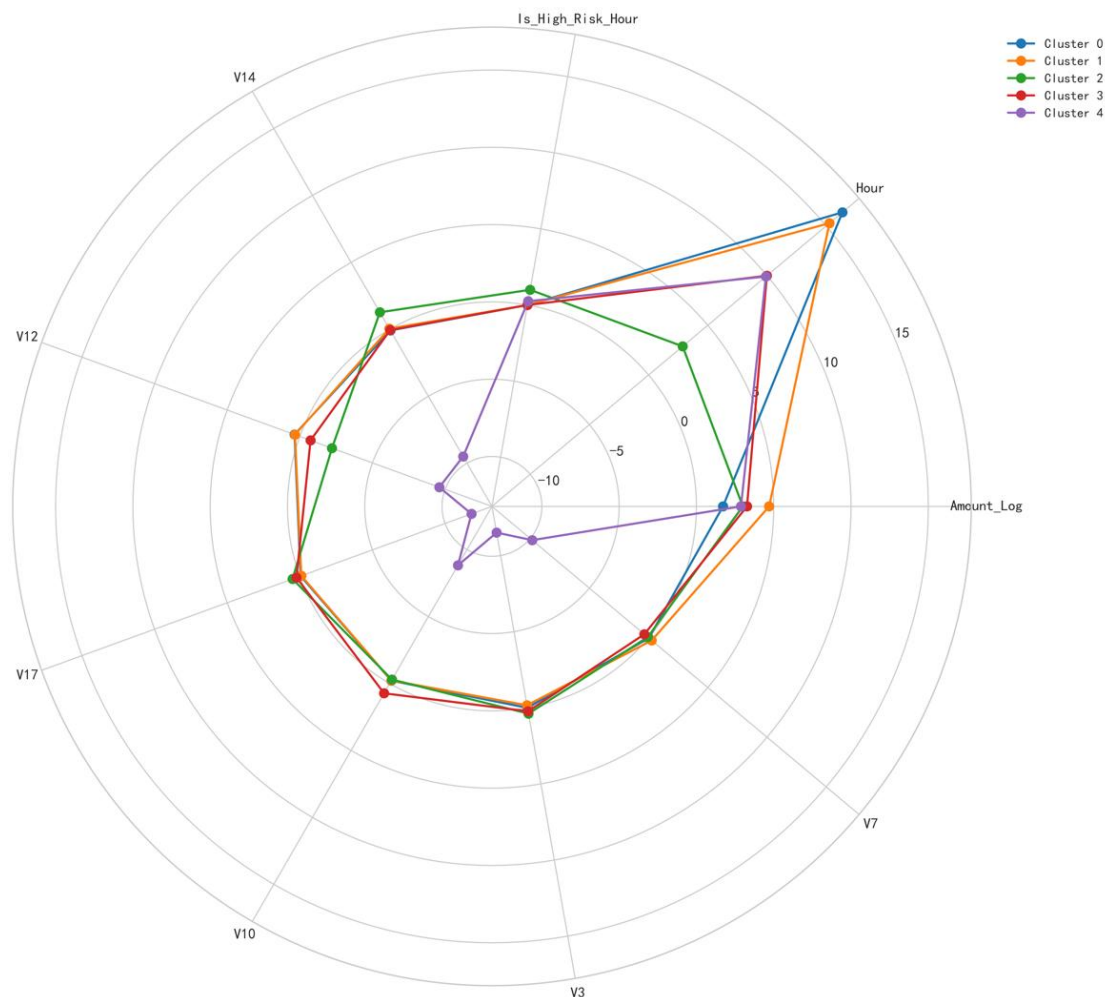


Figure 5. Radar chart of cluster characteristics.

As shown in Figure 5, the typical behavioral characteristics of high-risk groups align closely with the conclusions of the previous statistical analysis: the mean values of core features such as V14, V12, and V17 exhibit significantly negative deviations

from the normal range. These findings demonstrate that K-Means clustering accurately identifies high-risk groups with concentrated fraudulent transactions, providing targeted guidance for precise risk control strategies. Additionally, the clustering labels can serve as feature inputs for predictive models, enhancing their risk identification capabilities.

3.3. Results of Apriori Association Rule Mining

This study employs the Apriori algorithm to perform frequent item set mining on discretized time, monetary amounts, and low-value indicator variables of core features, identifying strong association rules with “fraud (Class=1)” as the consequent variable. Ultimately, over ten strong rules were extracted, each with a confidence level exceeding 20% and a lift value greater than 100. The top ten rules by confidence level are presented in Table 3.

Table 3. Top 10 High-Risk Association Rules.

Serial Number	Association Rules	Confidence	Lift	Covering the number of fraudulent samples
1	V17_Low&V10_Low&V16_Low	58.90%	340.96	364
2	V17_Low & V12_Low&V3_Low	39.78%	230.26	356
3	V17_Low & V12_Low & V10_Low	37.87%	219.24	367
4	V17_Low & V12_Low & V16_Low	37.87%	219.23	370
5	V17_Low & V10_Low&V3_Low	36.95%	213.88	351
6	V17_Low&V14_Low&V10_Low	35.24%	203.98	364
7	V17_Low&V14_Low&V3_Low	32.03%	185.43	353
8	V14_Low&V16_Low&V3_Low	28.10%	162.64	363
9	V17_Low&V16_Low&V3_Low	25.99%	150.47	353
10	V17_Low&V14_Low&V12_Low	25.91%	149.97	371

Fraud risk exhibits a pronounced combination effect: the individual low values of individual features carry limited risk, but when multiple core features such as V17, V14, and V12 simultaneously fall within the low-value range, the fraud probability increases exponentially. For example, the top-ranked rule V17_Low & V10_Low & V16_Low predicts fraud with a confidence level of 58.90% – meaning that when all three features are low, there is a 58.90% probability of the transaction being fraudulent, representing a 340.96-fold increase in predictive power compared to random guessing. High-confidence rules consistently incorporate core features like V17, V14, and V12, aligning closely with findings from correlation and clustering analyses, further confirming these features as pivotal variables in fraud detection.

The Top10 rules each cover over 350 fraud samples, encompassing more than 70% of all fraudulent cases, demonstrating exceptional generalization capabilities. These rules can be directly integrated into the risk control engine as pre-interception criteria, enabling real-time risk interception within milliseconds while providing interpretable decision-making support for black-box prediction models—perfectly meeting the real-time and compliance requirements of financial risk management.

3.4. Comparison of Model Performance and Result Interpretation

This paper trains four classification models: logistic regression (LR), random forest (RF), gradient boosting tree (GBDT), and XGBoost. The comparisons are as follows:

Figure 6 compares the confusion matrices of four models—logistic regression, random forest, gradient boosting, and XGBoost—on the test set. The horizontal axis represents the model's predicted category (0 for legitimate transactions, 1 for

fraudulent transactions), while the vertical axis shows the sample's actual category; the numerical values within the matrix indicate the number of samples in each category. Key performance metrics reveal: First, the logistic regression model exhibited the highest false positives (1,968 cases) and false negatives (19 cases), making it the only model with significant misclassification errors. This highlights its limited ability to capture the nonlinear characteristics of fraudulent transactions, adversely affecting user experience and business conversion rates, rendering it unsuitable for financial risk control scenarios. Second, the gradient boosting model had the fewest false positives (6 cases) and the lowest false-negative rate among legitimate users, demonstrating advantages in consumer payment contexts; however, its false negatives (39 cases) slightly exceed those of XGBoost and random forest, indicating weaker fraud detection capabilities. Third, the XGBoost model demonstrated the lowest false negatives (31 cases) and strongest fraud interception efficacy, effectively minimizing institutional financial losses with only 22 false positives—outperforming gradient boosting and achieving optimal balance between fraud detection and false-negative control, making it the preferred choice for risk management applications. Fourth, the random forest model exhibited balanced performance with 36 false negatives and 24 false positives, showing no significant weaknesses, excellent resistance to overfitting and generalization stability, and superior adaptability to long-tail business scenarios.

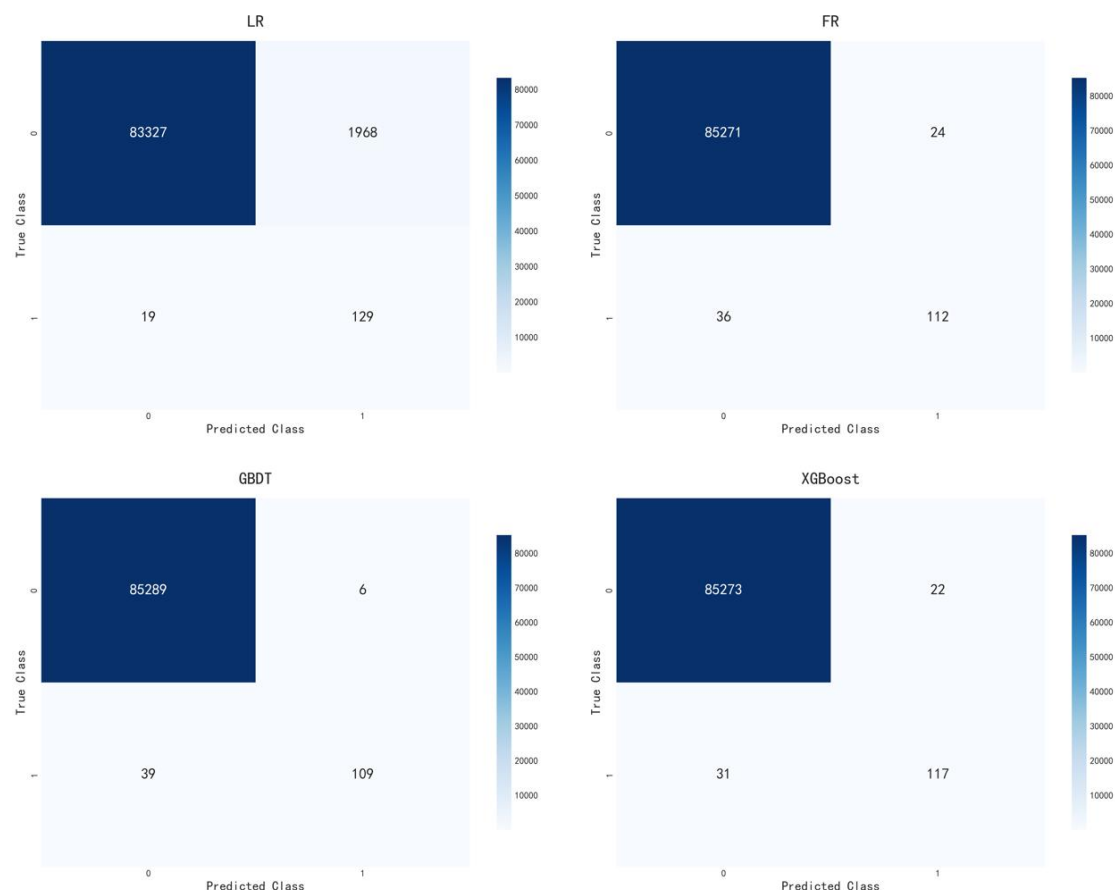


Figure 6. Comparison of Confusion Matrices.

A comparison of the ROC and PR curves, as shown in Figure 7, reveals the ROC curve on the left and the PR curve on the right. The ROC curve indicates that XGBoost achieves the highest AUC, though the differences among models are

minimal. The PR curve more clearly demonstrates XGBoost's ability to maintain high precision even at elevated recall rates, with its PR-AUC significantly outperforming other models, highlighting its superior overall recognition capability on imbalanced datasets. While Gradient Boosting's PR-AUC is relatively lower, it exhibits higher precision and fewer false positives in the low-recall region.

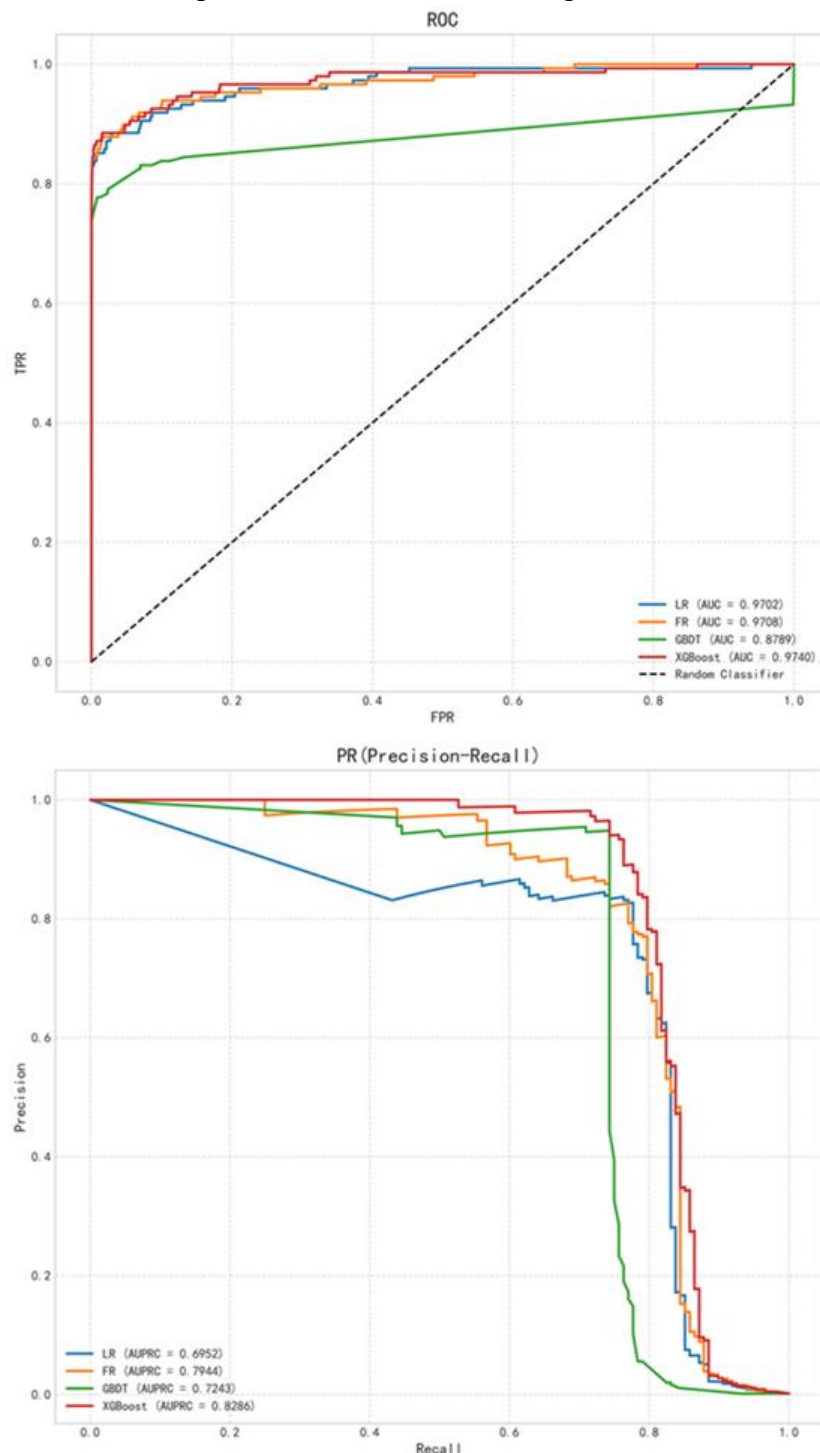


Figure 7. Comparison of ROC-PR Curves.

In summary, the integrated machine learning evaluation metrics presented in Table 4 demonstrate the performance of the four model categories on the test set. These metrics comprehensively cover three key dimensions—overall classification

capability, fraud detection accuracy, and false-positive cost control—fully aligning with the core business requirements of financial risk management scenarios.

As shown in Table 4, XGBoost achieves the best performance in both ROC-AUC and PR-AUC; while Gradient Boosting has the highest F1 score, it requires longer training time and yields a lower ROC-AUC; Random Forest’s PR-AUC is slightly lower than XGBoost’s, but its training speed is satisfactory. Therefore, the Gradient Boosting model is better suited for consumer payment scenarios that demand high user experience and low tolerance for false positives, minimizing the interception of legitimate transactions; the XGBoost model is more appropriate for credit and large-value payment scenarios requiring strict fund security and low tolerance for fraudulent missed detections, enabling higher fraud interception rates; the Random Forest model excels in scenarios with highly volatile data distributions, offering enhanced stability and generalization capabilities.

Table 4. *Evaluation Index of Each Machine Learning Model.*

Serial Number	Model	Accuracy	F1-Score	ROC-AUC	AUPRC
1	LR	0.9767	0.1149	0.9702	0.6952
2	RF	0.9993	0.7887	0.9708	0.7944
3	GBDT	0.9995	0.8289	0.8789	0.7243
4	XGBoost	0.9994	0.8153	0.9740	0.8286

4. Conclusion and Recommendations

4.1. Research Findings

This paper establishes a comprehensive research framework for predicting financial transaction fraud risks, encompassing “statistical pattern mining – group risk identification – association rule extraction – multi-model prediction comparison.” Through multidimensional empirical analysis employing K-Means clustering, Apriori association rules, logistic regression, random forests, gradient boosting, and the XGBoost algorithm, the study yields the following key findings:

First, fraudulent transactions exhibit distinct identifiable behavioral patterns: they show strong clustering between 2:00–4:00 AM in terms of time, and a high incidence of small transactions ranging from 0 to 10 yuan in terms of amount. The identifiers V17, V14, V12, V10, V16, and V3 serve as core features distinguishing fraudulent from legitimate transactions, providing fundamental insights for feature engineering and risk control rule design.

Second, the K-Means clustering algorithm accurately identifies high-risk transaction clusters. Cluster 4, comprising only 0.10% of the total sample, accounts for 52.2% of all fraud samples with an 85.95% fraud rate, making it a key monitoring target for the risk control system. The Apriori association rule algorithm extracts highly reliable fraud feature combinations, with top rules achieving a 58.90% confidence level and covering over 70% of fraud samples. These rules can be directly integrated into the risk control engine for real-time interception, while also providing interpretable business insights and feature enhancements for predictive models.

Third, among the four comparison models, the tree-based ensemble model demonstrates significantly superior overall performance compared to the linear logistic regression model, making it better suited for scenarios involving highly imbalanced financial fraud data and nonlinear features. Specifically, the gradient

boosting model excels in false positive control, minimizing disruption to legitimate users, while the XGBoost model offers more balanced comprehensive detection capabilities, making it the preferred single-model solution for financial fraud detection.

4.2. Application Recommendations

First, financial institutions can establish a dual-track intelligent risk control system combining a “rule engine + machine learning model.” Leveraging the strong association rules extracted by Apriori as the primary defense line enables millisecond-level interception of high-risk transactions, while employing XGBoost or gradient boosting models as the second core defense line to calculate transaction fraud risk probabilities. This creates a complementary and synergistic “hard rules + soft models” risk control framework, enhancing adaptability to complex scenarios such as dynamically evolving fraud patterns and data category imbalances.

Second, based on the core fraud characteristics, high-risk groups, and strongly correlated rules identified in this study, we optimize the end-to-end risk control strategy for credit transactions: implement a “zero-trust” approach for high-risk groups such as Cluster 4, requiring thorough verification for all transactions; increase the risk weight and enhance monitoring for small transactions occurring between 2:00 and 4:00 AM; and directly trigger interception or manual review for transactions matching high-confidence correlation rules, enabling precise and differentiated risk control deployment.

Third, commercial banks, payment institutions, consumer finance companies, and similar entities can leverage the modeling framework presented in this paper, tailored to their specific business data characteristics, to achieve localized adaptation and iterative optimization of their models. During the model iteration process, they should continuously incorporate the latest insights from K-Means clustering and Apriori association rule mining, updating feature engineering and risk control rules to ensure the models and rules remain aligned with the dynamic evolution of fraud patterns and maintain the sustained effectiveness of their risk management systems.

4.3. Research Gaps and Future Prospects

The empirical analysis in this study is conducted using publicly available benchmark datasets. However, external validation of the model's generalization performance across various scenarios and institutions requires further supplementation. The model comparisons primarily focus on empirical outcomes, and the statistical significance testing of performance differences among models needs further refinement. Additionally, the discussion of critical implementation issues—such as incremental learning, real-time deployment engineering, and the construction of gang fraud knowledge graphs—is relatively limited.

To address the aforementioned shortcomings, future research can advance in three key areas: First, incorporate financial transaction datasets from multiple sources and scenarios to conduct cross-sample and cross-institutional model validation, thereby enhancing the generalizability and external validity of research findings; Second, employ statistical significance tests such as the Friedman test and Bonferroni-Dunn test to systematically evaluate performance differences among models at the statistical level, strengthening the robustness of conclusions; Third, integrate online learning, incremental updating, and explainable AI technologies to further optimize the model's real-time performance, stability, and interpretability, while leveraging knowledge

graph technology to uncover correlations among accounts, devices, and IP addresses, identify syndicated and organized fraud patterns, and better align the model with the dynamic risk management requirements of real-world financial operations.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Acknowledgments

As this study nears completion, the authors look back on the fulfilling and challenging experience of the 6th Shanghai University Open Big Data Analysis Challenge, filled with deep emotions and gratitude. The smooth completion of this study is attributed to the strong support and selfless assistance from numerous mentors and friends, as well as the challenging platform provided by the 6th Shanghai University Open Big Data Analysis Challenge. Here, The author would like to acknowledge all parties involved.

References

- [1] Xia, G.W.; He, Q.Q.; Li, Y.S. Research on Financial Customer Risk Based on the K-means Data Analysis Method. *Journal of Insurance Vocational College*, 2022, 36(05), 52-57.
- [2] Du, H.Z.; Yang, C. Application of association rules in bank internal auditing — A case study of A Bank's personal loan customer data. *China Internal Audit*, 2019, 10, 62-65.
- [3] Qu, Z.F.; Wu, H.H.; Li, F.J. Credit risk assessment and credit strategy optimization modeling for small and medium-sized enterprises based on Logistic regression. *Journal of Shandong University (Science Edition)*, 2024, 59(08), 94-100.
- [4] Guo, J.S.; Qian, J.H. Research on Credit Card Default Prediction Based on Random Forest. *Modern Information Technology*, 2020, 4(03), 1-4.
- [5] Zhang, Z.M.; Lü, J.; Liu, Y. Research on a Personal Credit Risk Assessment Model Based on Optimized ACGAN-GBDT. *Financial Theory and Practice*, 2022(5), 84-89.
- [6] Chi, G.T.; Wang, Shanshan. A China listed company default risk prediction model based on XGBoost. *Journal of Systems Management*, 2024, 3, 735-754.
- [7] Yao, D.J.; Fei, B.C.; Zhou, M.B.; Liao, Y.L. Research on Credit Risk Assessment of Unlisted Technology-based SMEs under the Revised BP-KMV Model. *Financial Theory and Practice*, 2024, 11, 1-11.
- [8] Fu, Z.L.; Chen, X.Q.; Ren, W.; Yao, Y. A randomized K-nearest neighbor algorithm with learning process. *Journal of Jilin University (Engineering Edition)*, 2024, 1, 209-220.

- [9] Ma, X.J.; Song, Y.Q.; Chang, B.S.; Yuan, M.; Su, H. Application Research on a P2P Default Prediction Model Based on the CatBoost Algorithm. *Statistical and Information Forum*, 2020, 7, 9-17.
- [10] Qu, Z.F.; Wu, H.H.; Li, F.J. Credit risk assessment and credit strategy optimization modeling for small and medium-sized enterprises based on Logistic regression. *Journal of Shandong University (Science Edition)*, 2024, 8, 94-100.



© 2026 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)