

Intelligent Credit Card Fraud Risk Mitigation

Xinxin Li^{1*}, Yunfan Geng², Wangxin Zuo¹, Jingyi Wang¹, Sulin Mao¹, Runhui Xian¹, Xiaodong Zhu¹

¹ University of Shanghai for Science and Technology, Shanghai, China

² North China Institute of Science and Technology, Langfang, China

*Correspondence: Xinxin Li

Received: 20 May 2026; **Accepted:** 8 July 2026; **Published:** 3 August 2026

Abstract:

We employed K-means clustering and the Apriori association rule to construct additional features and trained an XGBoost classifier under imbalanced learning conditions. The XGBoost model achieved a PR-AUC value of 0.7904 on the test set, outperforming Random Forest and Logistic Regression models under the same evaluation criteria. In subsequent ablation experiments, we found that while both clustering features and association rule metrics provide complementary information, the association rule metrics contribute more significantly. Shapley analysis further revealed that model decisions are primarily driven by V14, V4, time (Hour), and other selected features. Based on the above analysis, we proposed a five-tier risk stratification strategy to translate predicted probabilities into different intervention levels. Under conditions of severe class imbalance, combining feature analysis with rule-based metrics and gradient boosting methods can effectively improve fraud detection performance; however, our system still requires optimization using longer and more diverse transaction data.

Keywords:

Credit Card Fraud Detection, Extreme Class Imbalance, XGBoost, SMOTE, Association Rule Mining, SHAP Interpretability, Risk Stratification, Ensemble Learning

1. Introduction

1.1. Background and Significance

With the rapid development of digital payments, financial transactions have become increasingly frequent. Consequently, fraud methods have also become diverse, and traditional rule-based systems struggle to cope with this complexity: strict rules sometimes block legitimate users, degrading the user experience, while lenient rules can directly lead to financial losses. Consequently, modern risk control requires data-driven machine learning models to identify rare fraudulent behaviors emerging from massive transaction volumes.

However, in practical applications, deploying these models faces significant challenges. On the one hand, transaction data suffers from severe class imbalance, with the number of legitimate transactions typically exceeding fraudulent transactions

by a factor of 500 or more. Additionally, to protect privacy, feature data is often anonymized, which directly complicates business interpretation. Furthermore, since online payments require immediate responses, detection systems must convert the model's probabilistic results into actionable business rules in real time—such as automatic blocking or manual review. Therefore, when designing fraud detection systems, we must balance risk control while also considering user experience and operational efficiency.

1.2. Review of Related Work

Early fraud detection systems primarily relied on manually designed rules and traditional models such as logistic regression. Although these methods are easy to implement and interpret, their performance is severely limited when fraud patterns evolve over time [4].

To improve detection performance under conditions of severe class imbalance, recent studies have adopted ensemble learning methods (such as Random Forests, XGBoost, and LightGBM), often combined with oversampling techniques (such as SMOTE or cost-sensitive learning) [2,5]. These methods typically achieve higher recall rates and greater robustness on benchmark fraud datasets.

Researchers have also explored deep learning models (including LSTM, Transformer, and graph neural networks) to capture sequential and relational information in transaction data [13]. However, these models are generally more difficult to interpret and challenging to deploy in real-time financial systems.

In recent years, the SHAP method has been widely applied in fraud detection tasks to explain feature importance and model behavior [3]. However, most existing research focuses on improving classification accuracy, with few studies exploring how to integrate exploratory analysis, feature engineering, model training, and risk control strategies. We aim to integrate these elements into a unified fraud detection framework.

1.3. Objectives and Scope

We aim to construct a practical fraud detection framework for highly imbalanced transaction data. The primary objective is to analyze statistical differences between fraudulent and legitimate transactions, particularly regarding transaction time, transaction amount, and features after PCA transformation. These analytical results provide guidance for feature selection and feature engineering.

The second objective is to examine whether unsupervised methods can provide additional risk information. Specifically, the K-means clustering algorithm is employed to identify transaction clusters with different fraud risk levels, while the Apriori association rule mining technique is used to discover high-risk feature combinations.

The third objective is to utilize XGBoost to construct a supervised fraud detection model within a non-data-leaking imbalanced learning process. we aim to evaluate whether the designed features can improve predictive performance under severe class imbalance.

Finally, we aim to translate model outputs into actionable risk control strategies through threshold optimization and a multi-level risk stratification framework.

1.4. Research Roadmap and Methodology Overview

Our overall workflow comprises four main phases. First, exploratory analysis is conducted to examine statistical patterns related to fraud; second, K-means clustering and Apriori association rule mining techniques are employed to generate additional risk-related features; third, we train and evaluate supervised learning models in a data-leakage-free imbalanced learning process; and finally, we utilize SHAP analysis and threshold optimization to support model interpretation and practical application.

The remainder of this paper is organized as follows: Section 2 describes the dataset, feature engineering process, and modeling methods; Section 3 presents the experimental results and model evaluation, along with an analysis of interpretability and the proposal of a risk stratification strategy; Section 4 summarizes our work and discusses its limitations and directions for future research.

2. Materials and Methods

2.1. Dataset Description

We utilized a European credit card transaction dataset recorded over two consecutive days in September 2013. After removing 60 duplicate records, the cleaned dataset contained 283,726 unique transaction records, of which 473 were identified as fraudulent transactions. The resulting fraud rate of 0.17% reflects severe class imbalance, with a ratio of legitimate to fraudulent transactions of approximately 599:1.

The dataset contains 30 features: V1–V28 are features derived from the PCA transformation of raw transaction attributes, along with time and amount features. There are no missing values in the data. In this extremely imbalanced scenario, we used a simple classifier to predict all instances, achieving an accuracy of 99.83%. Therefore, we did not use accuracy or ROC-AUC as the primary evaluation metrics, but instead adopted precision-recall area under the curve (PR-AUC) as the primary evaluation metric. Table 1 summarizes the key characteristics of this dataset. Table 1 summarizes the key characteristics of this dataset.

Table 1. Dataset Overview and Class Distribution

Property	Value
Transactions (after deduplication)	283,726
Fraudulent / Legitimate	473 (0.17%) / 283,253 (99.83%)
Imbalance ratio	599:1
Features	30 (V1–V28, Time, Amount)
Missing values	0
Amount range	0 – 25,691.16 EUR

2.2. Feature Engineering and Exploratory Insights

Based on the results of the previous unsupervised analysis, we developed a feature engineering strategy. The workflow is shown in Figure 1.

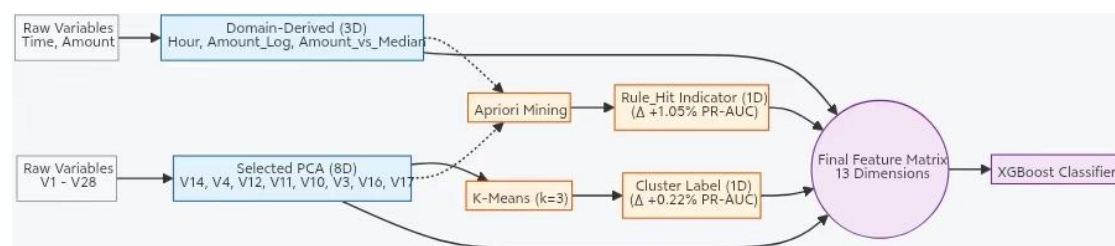


Figure 1. Feature Engineering Flowchart.

The discriminatory power of these optimized features will be evaluated in Section 3, which will also provide a further analysis of risk patterns based on the time dimension and quantitative metrics (see Figure 3 and Figure 4).

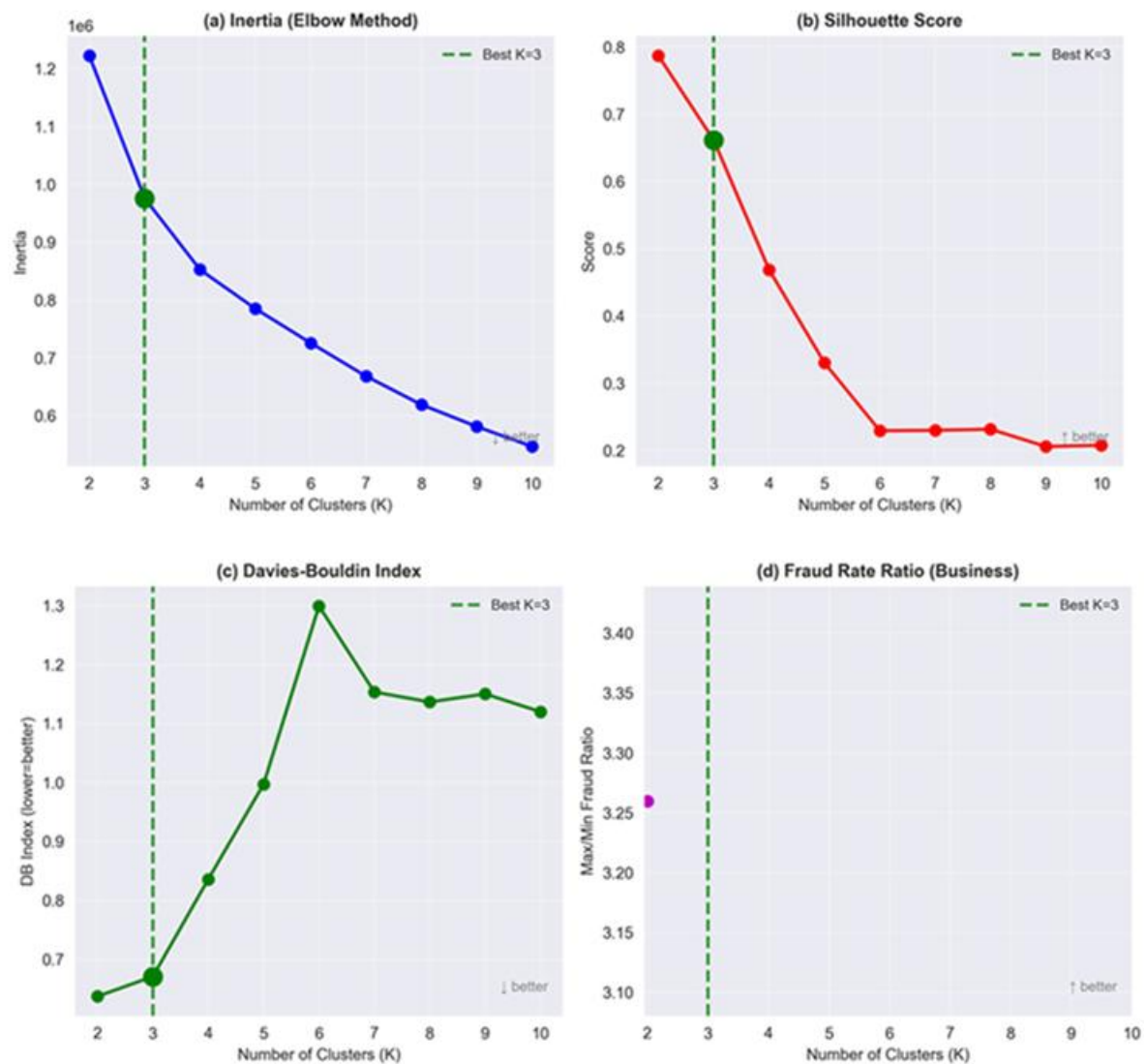


Figure 2. Optimal cluster number (K) evaluation using elbow method and silhouette scores.

First, several features are derived from the original “Time” and “Amount” variables. From the “Time” variable, we extracted “Hour” to represent trading sessions throughout the day; “Amount_Log” was used to log-transform transaction amounts to reduce their skewness; and “Amount_Log” was used to measure the deviation of a given transaction from the median expenditure.

Second, based on the Cohen’s d values, we selected the top eight PCA features for modeling, including V14, V4, V12, V11, V10, V3, V16, and V17. These features demonstrate stronger statistical separation between fraudulent and legitimate transactions.

Third, we applied the k-means clustering algorithm $k = 3$ to generate cluster label features.

Finally, we employed a priori association rule mining to construct a binary Rule_Hit metric for identifying transactions that match high-confidence fraud patterns.

Cohen's d measures the standardized mean difference between two groups:

$$d = \frac{\bar{X}_{\text{fraud}} - \bar{X}_{\text{legit}}}{s_{\text{pooled}}}, \quad s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad (1)$$

Where $\bar{X}_{\text{fraud}}$ and $\bar{X}_{\text{legit}}$ are group means, and s_{pooled} is the spooled standard deviation. Values of $|d| > 0.8$ are conventionally considered large effects.

2.3. Cluster Analysis

We applied the k-means clustering algorithm to the top 8 standardized PCA features to reveal the natural grouping structure of transaction behavior. We selected the optimal clustering k value by conducting a dual test using the elbow criterion and the silhouette coefficient. As shown in Figure 2 and Figure 3, we ultimately determined that the optimal number of clusters is k=3. The three clusters exhibit distinct risk characteristics: Cluster 0 contains 11,925 transactions, accounting for 4.20% of the total, with a fraud rate of 0.335%, which is 2.01 times higher than the overall average. It is characterized by a mean transaction amount of €756.76, which is relatively high and concentrated in the afternoon hours; Cluster 1 contains 271,187 transactions, accounting for 95.58% of the total, with a fraud rate of 0.159%, slightly below the average, and transaction amounts are relatively low and distributed across all time periods throughout the day; Cluster 2 contains 614 transactions, accounting for 0.22% of the total, with a fraud rate of 0.163%. The transaction amounts in this cluster are extremely high, with a mean of €3,425.16. These characteristics indicate that the cluster analysis effectively captures the heterogeneous nature of risk. Cluster 2 is relatively small, and the contribution of its features to the model is very weak, resulting in a statistically unstable fraud rate. Nevertheless, this cluster still provides important global risk context information at the lowest inference cost.

Table 2. Cluster Risk Profiles and Transaction Characteristics.

Cluster	Count	Proportion	Fraud Rate	Avg Amount (EUR)	Nighttime %	Risk Level
0 (Medium)	11,925	4.20%	0.335% (2.01x)	756.76	18.3%	Medium
1 (Low)	271,187	95.58%	0.159%	51.53	12.1%	Low
2 (High-Value)	614	0.22%	0.163%	3,425.16	8.7%	Low

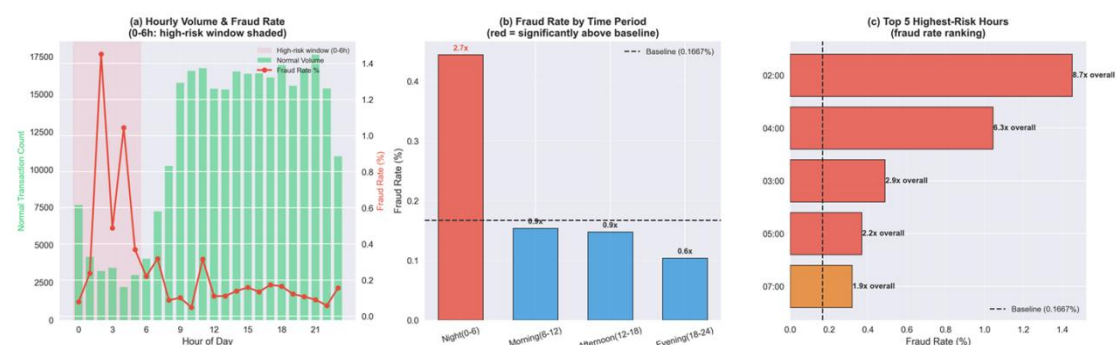


Figure 3. Transaction clustering structure and fraud distribution across clusters.

2.4. Association Rule Mining

As shown in Figure 4, we employed association rule mining to explore the data, with the aim of identifying feature combinations that co-occur with fraudulent

behavior. This analysis was conducted on the training set to ensure that the findings remain descriptive and to prevent future information from influencing the model evaluation. For the continuous features V14, V4, V12, time period, and amount/log, we divided them into low, medium, and high intervals using the 33rd and 67th percentile thresholds. Subsequently, the Apriori algorithm was employed for association rule mining. With the target outcome “Fraud = 1” as the criterion, the minimum support was set to 0.0005 and the minimum confidence to 0.8. The resulting association rules were sorted in descending order based on lift and confidence. The top-ranked strong association rules indicate that when a transaction simultaneously meets three conditions—V14 is low, V4 is low, and the transaction time slot is between 0 and 6 (nighttime)—it exhibits a very strong association with fraudulent transactions. This rule has a confidence level above 0.90 and a lift exceeding 120. Previous exploratory analysis also concluded that the risk of fraud in nighttime transactions increases significantly, and these two findings corroborate each other. Based on these mining results, we converted transactions matching the high-risk association rules into a binary feature, Rule_Hit, and incorporated it into the feature set for subsequent supervised learning modeling.

The key metrics for rule evaluation are formally defined as:

$$\text{Support}(A \rightarrow B) = P(A \cap B) = \frac{|A \cap B|}{N} \quad (2)$$

$$\text{Confidence}(A \rightarrow B) = P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (3)$$

$$\text{Lift}(A \rightarrow B) = \frac{P(B|A)}{P(B)} = \frac{\text{Confidence}(A \rightarrow B)}{P(B)} \quad (4)$$

where A denotes the antecedent (feature conditions), B denotes the consequent (Fraud = 1), and N is the total number of transactions. Rules with Confidence ≥ 0.8 and Lift > 50 are retained for the Rule_Hit feature.

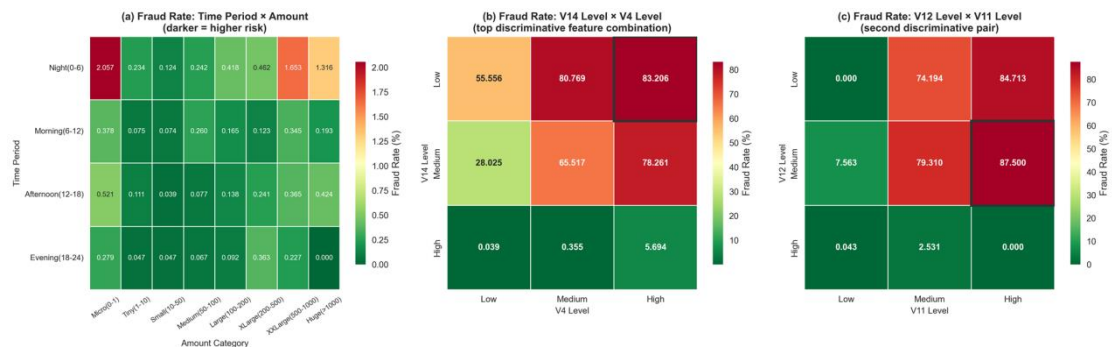


Figure 4. High-risk behavior combination analysis showing fraud rate by feature interactions.

To prevent data leakage, discretization thresholds were derived exclusively from the training fold, and the Rule_Hit feature was applied to validation/test folds using these thresholds. (Table 3 presents the discovered rules).

Table 3. Top High-Confidence Fraud Association Rules.

Antecedent $\rightarrow$ Fraud	Confidence	Lift	Coverage
V10_Med $\wedge$ V14_Low	100%	599.8	48
V16_Med $\wedge$ V10_Med $\wedge$ V4_Med	100%	599.8	45
V10_Med $\wedge$ V11_Med $\wedge$ Afternoon	100%	599.8	40
V11_High $\wedge$ V10_Med	100%	599.8	37

V16_Med $\wedge$ V10_Med $\wedge$ Afternoon	97.8%	586.8	45
V12_Med $\wedge$ V10_Med $\wedge$ V4_Med	97.8%	586.5	44

2.5. Model Architecture and Training Pipeline

For supervised modeling, XGBoost is used as the primary classifier, selected for its strong performance on tabular data and its support for cost-sensitive learning. The model is trained within an ImbPipeline consisting of Robust Scaler, SMOTE, and XGBoost.

To address the 599:1 class imbalance, SMOTE oversampling is applied only to the training folds. SMOTE works by generating synthetic minority instances through interpolation.

Formally, for a minority instance and a randomly selected neighbor among its k -nearest neighbors, a synthetic sample is generated as:

$$x_{\text{new}} = x_i + \lambda \cdot (x_{\text{nn}} - x_i), \quad \lambda \sim U(0,1) \quad (5)$$

This expands the decision boundary with controlled variability, unlike random oversampling, which merely duplicates existing samples. In addition, the `scale_pose_weight` parameter is set to 599 to further penalize misclassification of the minority class.

In our subsequent selection of a scaling method, we chose the robust scaling method primarily because there were significant outliers among the Principal Component Analysis (PCA) features; for example, the proportion of outliers in the fraudulent transactions dataset (V3) was significantly higher. Furthermore, compared to the standard scaling method, the robust scaling method is less sensitive to outliers.

The XGBoost hyperparameters are: `learning_rate` = 0.05, `max_depth` = 6, `subsample` = 0.8, `colsample_bytree` = 0.8, `n_estimators` = 200, with early stopping based on validation PR-AUC.

2.6. Cross-Validation Strategy and Temporal Evaluation

We employed a stratified 5-fold cross-validation method, running the process five times with different random seeds to obtain robust performance evaluation results that reflect the characteristics of the variance. PR-AUC was used as the primary evaluation metric in all model selection and comparison decisions, as ROC-AUC is known to exhibit an optimism bias under extreme class imbalance; even if a model fails to detect most fraud cases, it may still achieve a high ROC-AUC value [8].

This process ensures that the validation folds retain the original class distribution. Therefore, the difference between the PR-AUC of the cross-validation (CV) set and the test set reflects the difficulty of generalization under class-imbalanced conditions.

To assess temporal robustness, we also evaluated all models under a temporal split, using the first 80% of transactions in chronological order for training and the last 20% for testing. This simulates the practical deployment scenario where the model is applied to future transactions that may change over time.

3. Results and Discussion

3.1. Exploratory Feature Analysis

Prior to model training, we evaluated the discriminatory power of all 30 features using a four-component complementary metric system: Cohen's d effect size, univariate AUC, the Kolmogorov-Smirnov statistic, and point-bisequential correlation with the binary fraud label. This multi-metric evaluation method provides robust ranking results and does not rely on any single statistical assumption.

Table 4 lists the top 10 features. V14 exhibits the strongest discriminatory power, followed by V4 and V12. These three features alone account for a significant proportion of the separable fraud signals.

As shown in the analysis in Figure 7, aside from PCA features, fraudulent transactions are highly concentrated during the night, specifically between midnight and 6 a.m., with a fraud rate 2 to 3 times higher than the daytime average. At the same time, transactions involving extremely small amounts also exhibit abnormally high fraud rates, which is consistent with fraudsters testing stolen cards through small transactions before making large withdrawals.

Exploratory analysis indicates that fraudulent transactions differ from legitimate transactions in three key dimensions: (a) a few PCA features, particularly V14, V4, and V12, exhibit significant distribution differences (Cohen's d > 1.8); (b) Fraudulent activity is concentrated during nighttime hours; (c) Micro-transactions (<1 euro) carry an abnormally high risk of fraud. These findings directly inform feature selection, the inclusion of time variables, and the association rules identified in subsequent analyses.

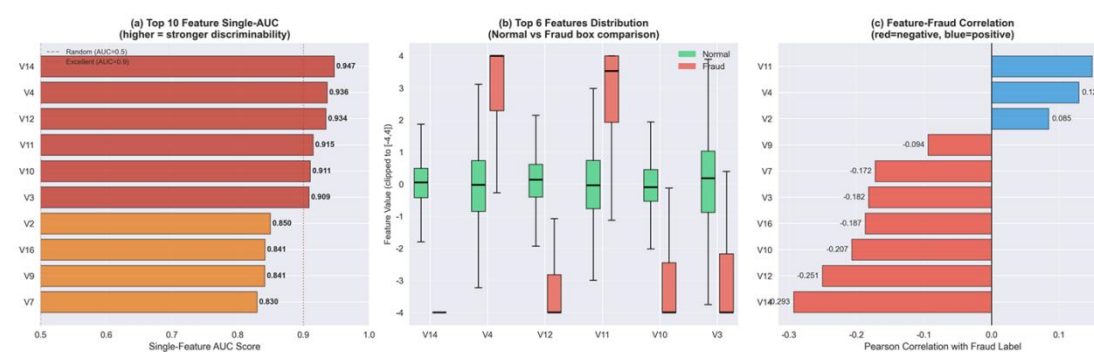


Figure 6. Key feature discriminative power analysis showing AUC scores and distribution comparison.

Table 4. Feature Discriminative Power Ranking.

Feature	Cohen's d	AUC	KS	r	Rank
V14	2.228	0.947	0.838	0.293	1
V4	1.985	0.936	0.761	0.129	2
V12	1.848	0.934	0.777	0.251	3
V11	1.844	0.915	0.747	0.149	4
V10	1.603	0.911	0.797	0.207	5
V3	1.350	0.909	0.695	0.182	6
V16	1.445	0.841	0.678	0.187	7
V17	1.307	0.800	0.736	0.313	8
V9	1.325	0.841	0.575	0.094	9
V7	1.054	0.830	0.652	0.172	10

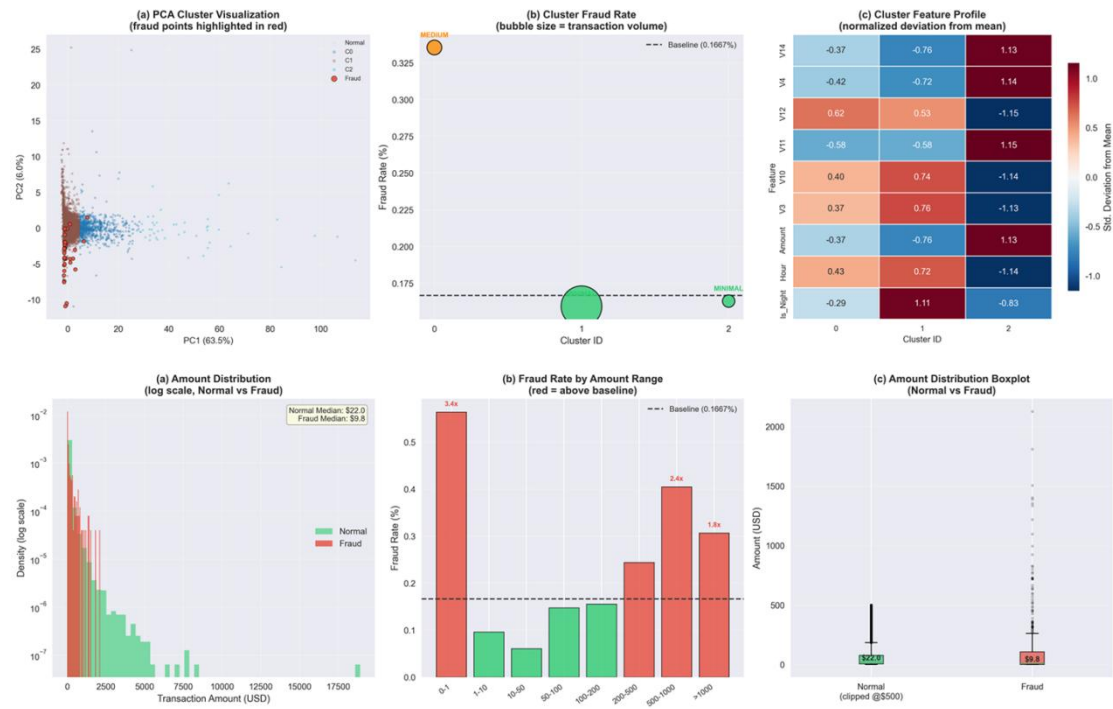


Figure 7. Analysis of Fraud Risks by Time Period, Distribution of Transaction Amounts Across Different Amount Ranges, and Analysis of Fraud Rates.

3.2. Supervised Model Comparison

Model comparison was conducted using the SMOTE method and 5-fold cross-validation. Under the same ImbPipeline settings, XGBoost was compared with Random Forest and Logistic Regression models. As shown in Figure 8 and Table 5, among the three models, XGBoost achieved the highest test PR-AUC value at 0.7904 ± 0.0089 , followed by Random Forest. Although the ROC-AUC value of the Logistic Regression model was as high as 0.9665, its test PR-AUC value was significantly lower at only 0.6571. This result is consistent with the findings of Saito and Rehmsmeier (2015), who noted that the ROC-AUC can be misleading under conditions of extreme class imbalance. Furthermore, XGBoost exhibited the smallest difference between its cross-validation and test PR-AUC values, outperforming both Random Forest and Logistic Regression numerically, while requiring significantly less training time than Random Forest. Based on these results, we selected XGBoost as the primary classifier.

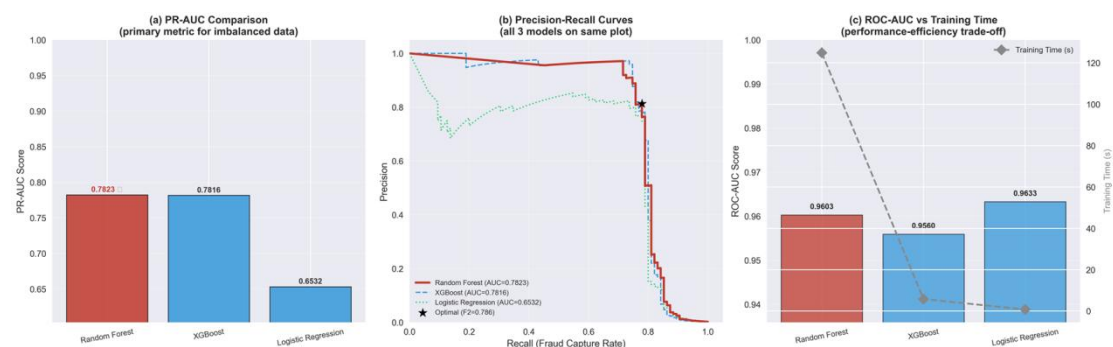


Figure 8. Fraud prediction model comprehensive comparison including PR-AUC, ROC curves, and confusion matrices.

Table 5. Supervised Model Performance Comparison.

Model	CV PR-AUC	Test PR-AUC	ROC-AUC	KS	Training Time (s)
XGBoost	0.8794 ± 0.0056	0.7904 ± 0.0089	0.9580 ± 0.0034	0.8421 ± 0.0087	12.1 ± 1.2
Random Forest	0.8834 ± 0.0048	0.7812 ± 0.0072	0.9556 ± 0.0041	0.8513 ± 0.0065	144.3 ± 8.7
Logistic Regression	0.8471 ± 0.0067	0.6571 ± 0.0114	0.9665 ± 0.0028	0.8532 ± 0.0079	1.5 ± 0.3

3.3. Ablation Study on Unsupervised Feature Contributions

We employed an ablation analysis to evaluate the respective contributions of clustering labels and association rules. As shown in Table 6, removing the Rule_Hit metric reduces the test PR-AUC by 1.05%, while removing the Cluster feature results in a 0.22% decrease in PR-AUC; removing both features simultaneously leads to a combined reduction of 1.38% in PR-AUC. The PR-AUC value obtained using only the PCA baseline model is 0.7752. These results indicate that both unsupervised methods provide useful predictive signals beyond the original PCA components. Although the clustering feature yields only a moderate improvement, the rule-based metric contributes more significantly to the final model's performance.

Table 6. Ablation Study Results for Unsupervised Feature Contributions.

Configuration	Features	Test PR-AUC	Δ PR-AUC
Full model	8 PCA + 3 Domain + Cluster + Rules	0.7904	—
w/o Cluster	8 PCA + 3 Domain + Rules	0.7887	−0.22%
w/o Rules	8 PCA + 3 Domain + Cluster	0.7821	−1.05%
w/o Both	8 PCA + 3 Domain	0.7795	−1.38%
PCA-only	8 PCA features	0.7752	−2.01%

3.4. Temporal Robustness Analysis

To evaluate the model's performance under time drift conditions, time-ordered data splitting was compared with random stratified splitting. Table 7 shows that under chronological partitioning, the test PR-AUC dropped to 0.7621, a 3.58% decrease compared to random partitioning. This performance decline indicates the presence of temporal drift, as recent fraud patterns may differ from historical ones. Therefore, in practical deployment, the model needs to be periodically retrained using the latest transaction data window to maintain its detection capability.

Table 7. Temporal Validation Strategy Comparison.

Validation Strategy	Train	Test	PR-AUC	Δ vs Random
Random stratified	80% (stratified)	20% (stratified)	0.7904	—
Time-ordered	First 80% (by Time)	Last 20% (by Time)	0.7621	−3.58%

3.5. SHAP Feature Importance and Interpretability Analysis

We applied the SHAP method to analyze feature importance in the trained XGBoost model (Figure 9 and Table 8).

Among all features, V14 and V4 exhibited the highest SHAP values, consistent with the results of the previous discriminant analysis. The time factor also ranked highly, indicating that transaction time has a significant impact on fraud prediction.

The SHAP swarm plot shows that transactions during late-night hours tend to increase the predicted fraud probability.

The Amount_Log feature ranked among the top features, while the contributions of the Rule_Hit and Cluster features were relatively small. Overall, the results of the SHAP analysis are largely consistent with the statistical patterns observed in the exploratory analysis.

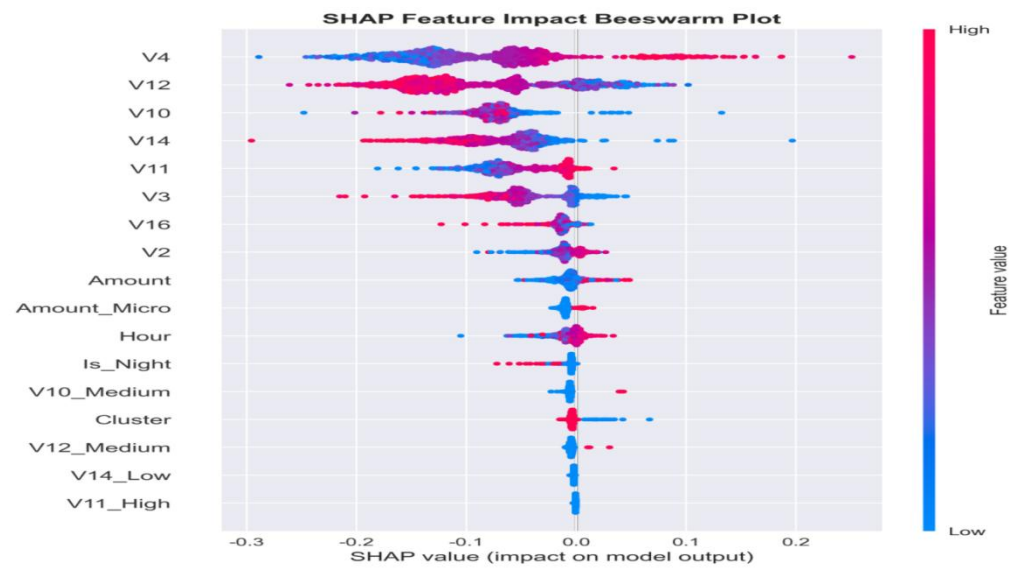


Figure 9. SHAP feature impact beeswarm plot showing importance and direction of top features.

Table 8. SHAP Feature Importance and Source Attribution.

Feature	SHAP Imp.	Source	Cohen's d
V14	2.353	PCA	2.228
V4	2.125	PCA	1.985
Hour	1.208	Domain-derived	—
V10	0.968	PCA	1.603
V11	0.911	PCA	1.844
V12	0.842	PCA	1.848
Amount_Log	0.618	Domain-derived	—

3.6. Risk Stratification Framework

Based on the predicted fraud probability (Table 9), we developed a five-tier risk stratification strategy.

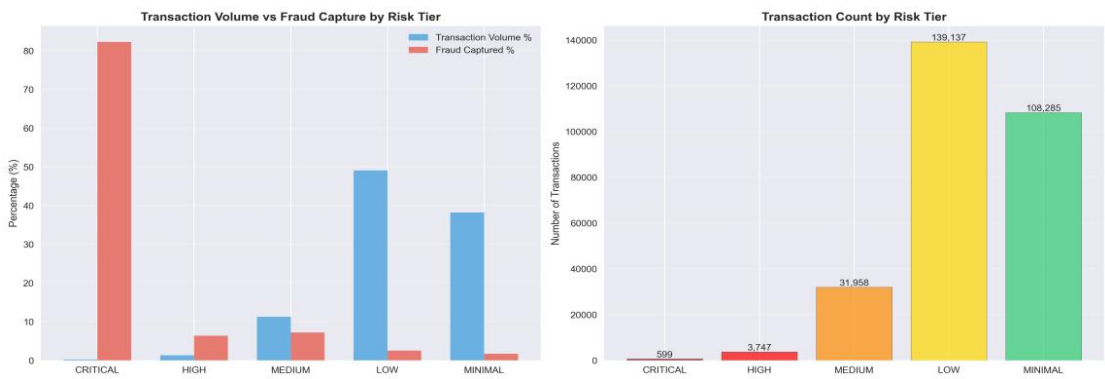


Figure 10. Risk stratification showing transaction volume and fraud capture rate by risk tier.

Transactions in the critical tier are assigned the highest fraud probability (Figure 10) and can be blocked immediately. High-risk transactions are suitable for manual review, while medium-risk transactions may require additional verification. The low-risk and lowest tiers contain the majority of legitimate transactions and require virtually no intervention.

Table 9. Five-Tier Risk Stratification Framework.

Tier	Prob. Range	Count	Fraud %	Capture %	Action
CRITICAL	> 0.9998	284 (0.10%)	96.13	57.72	Automated blocking
HIGH	0.16 – 0.9998	1,135 (0.40%)	16.30	39.11	Manual review
MEDIUM	0.057 – 0.16	1,419 (0.50%)	0.00	0.00	Verify
LOW	0.005 – 0.057	11,349 (4.00%)	0.026	0.63	Monitor
MINIMAL	< 0.005	269,539 (95.05%)	0.004	2.54	Pass

3.7. Decision Threshold Optimization

Researchers evaluated different classification thresholds to examine the trade-off between precision and recall (Table 10). Due to the severe imbalance in the fraud detection dataset, adjusting the threshold significantly affects the number of false positives and false negatives.

When the threshold is set to 0.68, the model's F2 score reaches 0.774, with 35 false positives and 19 false negatives. Lowering the threshold improves recall but increases the number of false positives; conversely, raising the threshold enhances precision but weakens fraud detection capabilities. Therefore, 0.68 was selected as the threshold.

We selected the F2 score as the primary threshold evaluation metric because, in fraud detection tasks, recall is generally more important than precision.

The F-score is defined as:

$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}} \quad (6)$$

With $\beta = 2$, the F_2 score weights recall twice as heavily as precision, reflecting the higher cost of missed fraud cases relative to false alarms.

Overall, the threshold analysis shows that the model can maintain relatively high fraud detection performance while controlling the number of false positives through appropriate threshold selection.

Table 10. Classification Threshold Optimization Analysis.

Threshold	Precision	Recall	F ₂	FP	FN
0.1	0.2	0.842	0.513	320	15
0.3	0.394	0.8	0.663	117	19
0.5	0.543	0.8	0.731	64	19
0.68	0.685	0.8	0.774	35	19
0.83	0.789	0.789	0.789	20	20

4. Conclusions

4.1. Conclusions

Based on a European credit card dataset, we explore a fraud detection framework suitable for highly imbalanced credit card transaction data.

Exploratory analysis reveals significant statistical differences between fraudulent and legitimate transactions. Multiple Principal Component Analysis (PCA) features, particularly V14, V4, and V12, demonstrate clear class separation effects.

Fraud cases occur more frequently and are more concentrated during nighttime hours and in small-value transactions.

By further employing K-means clustering and Apriori association rule mining techniques, additional features were generated for supervised learning modeling. Meanwhile, ablation experiment results indicate that the Rule_Hit feature significantly outperforms the cluster-label feature in improving PR-AUC, although both provide complementary information. Among the evaluated models, XGBoost achieved the best overall performance, with a test PR-AUC of 0.7904 ± 0.0089 . SHAP analysis indicates that the model's predictions are primarily influenced by several key features, including V14, V4, and Hour. Finally, a five-tier risk stratification strategy was constructed based on predicted fraud probabilities. Experimental results show that the majority of fraud cases can be identified within a relatively small proportion of transactions. Overall, we demonstrate that, under conditions of severe class imbalance, combining exploratory analysis, feature engineering, and imbalanced learning methods can improve fraud detection performance.

4.2. Theoretical and Practical Contributions

We integrate exploratory analysis, unsupervised feature generation, and supervised learning into a unified fraud detection workflow. The results show that, under severe class imbalance, clustering features and association rule metrics can provide additional information for fraud prediction. From a methodological perspective, we employ SMOTE in the cross-validation process to mitigate data leakage during model evaluation. PR-AUC was selected as the primary evaluation metric because it is more suitable for highly imbalanced fraud detection tasks. Furthermore, we propose a five-level risk stratification strategy to translate predicted fraud probabilities into different levels of intervention measures, including blocking, verification, and manual review.

4.3. Limitations

The dataset covers transaction records from a single European card issuer over a two-day period, which may limit the model's generalization ability across different geographic regions, card types, and fraud types. Time-series drift analysis shows that the PR-AUC metric decreased by 3.58% after splitting the data chronologically, highlighting the necessity of periodically retraining the model. After the V1–V28 data was anonymized via PCA, business-level insights could not be directly derived from these features, thereby limiting the ability to design feature-level operational rules. Finally, the SMOTE method may cause a shift in the data distribution by generating synthetic samples that do not reflect real fraud patterns.

4.4. Future Work

Future research should explore more targeted oversampling strategies, such as boundary SMOTE or ADASYN, to mitigate the distribution shift caused by the standard SMOTE method. Additionally, online learning mechanisms could be established to enable the model to continuously adapt to evolving fraud patterns without requiring complete retraining.

Conflicts of Interest

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

Funding

This work was supported by the 6th Shanghai Universities Open Big Data Analytic Challenge, which provided the research motivation and data resources that inspired this study. The authors gratefully acknowledge the competition organizers and mentors for their valuable guidance and feedback during the early stages of this research.

Acknowledgments

The authors would also like to acknowledge the Teaching Research and Reform Project of Higher Continuing Education of the University of Shanghai for Science and Technology, entitled “Research on the Innovation of Collaborative Education Models in Higher Continuing Education: A Case Study of the Artificial Intelligence Top-up Bachelor’s Program,” for its valuable support and inspiration to this study.

References

- [1] Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794, DOI: 10.1145/2939672.2939785.
- [2] Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002, 16, 321–357, DOI: 10.1613/jair.953.
- [3] Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774, DOI: 10.48550/arXiv.1705.07874.
- [4] Dal Pozzolo, A.; Caelen, O.; Johnson, R.A.; Bontempi, G. Calibrating probability with undersampling for unbalanced classification. In Proceedings of the IEEE Symposium Series on Computational Intelligence, 2015; pp. 159–166, DOI: 10.1109/CIBCB.2015.7218021.
- [5] Fernández, A.; García, S.; Galar, M.; Prati, R.C.; Krawczyk, B.; Herrera, F. Learning from Imbalanced Data Sets. Springer, 2018, DOI: 10.1007/978-3-319-98074-4.
- [6] Agrawal, R.; Imieliński, T.; Swami, A. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 1993, 22(2), 207–216, DOI: 10.1145/170035.170072.
- [7] He, H.; Ma, Y. Imbalanced Learning: Foundations, Algorithms, and Applications. Wiley, 2013, DOI: 10.1002/9781119012013.
- [8] Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 2015, 10(3), e0118432, DOI: 10.1371/journal.pone.0118432.

- [9] Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 2017, 30, 3146-3154, DOI: 10.48550/arXiv.1706.08359.
- [10] Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 2018, 31, 6638-6648, DOI: 10.1109/ICDM.2017.26.
- [11] Wang, S.; Liu, C.; Gao, X.; Qu, H.; Xu, W. A comprehensive comparison of ensemble methods for credit card fraud detection. *IEEE Access*, 2020, 8, 216589-216600, DOI: 10.1109/ACCESS.2020.3016350.
- [12] Domingues, R.; Filippone, M.; Michiardi, P.; Zouaoui, J. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, 2018, 74, 406-421, DOI: 10.1016/j.patcog.2017.11.003.
- [13] Liu, Y.; Li, T.; Wu, Z.; Chen, H. Graph neural networks for fraud detection: A survey. arXiv preprint arXiv:2104.01400, 2021, DOI: 10.48550/arXiv.2104.01400.



© 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/>)