PCA-Integrated LightGBM and XGBoost Model for Pattern Recognition and Interpretability Analysis of Telecom Fraud

Donghao Li*, Xiaohan Wang, Xiaoyu Lu, Bing He
Henan University of Technology, Zhengzhou, Henan, China
**Corresponding Author*

Abstract: In response to the increasing complexity of telecom network fraud and issues such as high-dimensional imbalanced data, an integrated model based on LightGBM and XGBoost is proposed in this paper. The prediction results are fused using Principal Component Analysis (PCA), and model interpretability is enhanced through SHAP values. First, raw transaction data are preprocessed and subjected to feature engineering. Then, model parameters are optimized via cross-validation, constructing a fraud detection pathway of "identification–interpretation–integration". Experimental results show that the PCA-fused model outperforms individual models in both detection performance and interpretability, providing an effective intelligent solution for accurate telecom fraud detection.

Keywords: Telecom Fraud; Ensemble Learning; PCA Fusion; SHAP Interpretation; LightGBM; XGBoost

1. Introduction

Telecom network fraud has evolved from simple traditional scams to complex crimes involving AI synthesis and blockchain money laundering, posing serious threats to national cybersecurity and citizens' property safety. Data from 2024 show that telecom fraud-related attacks have an annual growth rate of 42%, with deepfake cases accounting for over 60%, causing economic losses exceeding 12 billion yuan [1]. To address high-dimensional, imbalanced, and dynamically evolving fraud patterns, there is an urgent need for intelligent detection models that integrate high-performance algorithms and interpretability mechanisms.

This study aims to construct a telecom fraud detection model that integrates ensemble learning and interpretability analysis, improving recognition performance under high-dimensional imbalanced data and enhancing interpretability and adaptability in real-world deployment scenarios [2]. Based on transaction and identity data provided by Vesta Corporation, systematic data preprocessing, feature construction, and model fusion strategies are employed to achieve accurate identification and scoring of high-risk fraudulent transactions. This optimizes the security rule system, promotes the upgrade of cybersecurity defense from passive response to active prediction, and ultimately forms an intelligent defense solution with both detection accuracy and cybersecurity adaptability.

2. Research Background

In complex data processing and analysis tasks, a single model often fails to meet the requirements for accuracy, efficiency, and data dimensionality reduction. Therefore, model fusion and dimensionality reduction techniques have become key means to enhance model performance. LightGBM and XGBoost, as prominent models under the gradient boosting decision tree framework, excel in various data mining tasks. Principal Component Analysis (PCA), as a classical dimensionality reduction algorithm, opens new pathways for model fusion by extracting principal components to effectively integrate prediction results.

2.1 Characteristics of the LightGBM Model

LightGBM adopts a histogram algorithm, which only requires storing discretized feature values (typically 8-bit integers), reducing memory usage to 1/8. It employs a leaf-wise growth strategy with depth limits, achieving higher accuracy under the same number of splits (overfitting is prevented by depth constraints) [3,4]. When processing large-scale data, LightGBM natively supports feature parallelism and data parallelism: data parallelism uses "Reduce Scatter" to merge histograms and reduce communication overhead through

difference techniques; voting parallelism (Voting Parallel) reduces the communication cost of finding optimal split points based on the PV-Tree algorithm.

2.2 Characteristics of the XGBoost Model

XGBoost is a distributed and scalable variant of GBDT. It controls model complexity through explicit regularization, enhancing generalization ability. XGBoost constructs decision trees in parallel, significantly reducing training time. It employs a level-wise growth strategy, scanning gradient values to quickly evaluate the quality of split points. Additionally, XGBoost supports multi-language and cloud platform integration, offering good portability and is widely used in both industry and academia. Faced with high-dimensional complex data, its parallel capability and regularization mechanism efficiently learn features and output stable and accurate predictions [5,6].

3. Model Design

3.1 Data Source and Preprocessing

The data are sourced from the IEEE-CIS Fraud Detection dataset on Kaggle [7], containing over one million transaction and identity records. In the data preprocessing phase, missing values are handled, outliers are removed, and feature consistency is synchronized to ensure the consistency and validity of training and test data.

3.2 Feature Selection and Construction

First, the linear correlations between original variables are calculated and visualized through a heatmap. Analysis reveals strong collinearity among some variables, such as V257 and V246 with a correlation coefficient of 0.91, and V244 and V242 with a coefficient of 0.97(see Figure 1).

Additionally, among non-V-type features, such as ProductCD and some merchant features, certain correlations are also observed (see Figure 2). If these highly correlated variables are included in the model simultaneously, it may lead to model redundancy, reduced training efficiency, and even overfitting. To reduce redundancy, highly collinear variables are filtered out in the subsequent feature screening phase.

To further explore the hidden fraud-related features in the data and enhance the model's ability to identify complex fraud patterns, a set of derived variables with strong business relevance were constructed based on the original transaction fields (such as transaction amount, device ID, payer email, payee email, and transaction time), combined with the business logic and behavioral patterns of telecom fraud. These variables supplement, from multiple dimensions, the fraud identification information that the original data fails to cover. For instance, targeting the common "abnormal fluctuation of transaction amount" feature in fraudulent transactions, an amount deviation indicator was built by calculating the "ratio of the current transaction amount to the user's average transaction amount over the past 30 days"—the transaction amounts of normal users usually fluctuate around their historical average, while fraudsters who have stolen accounts tend to make transfers or consumption far exceeding the user's regular spending level, and this indicator can quickly detect such abnormalities. Aiming at the typical fraud scenario of "multiple accounts sharing one device" (fraud gangs often log in to multiple illegally obtained accounts through the same device for batch operations), the "number of different accounts associated with a single device ID within the past 7 days" was counted; when this value exceeds the threshold of device-account association for normal users, the behavior can be marked as suspicious. Considering that some fraudulent transactions transfer funds through forged or unassociated emails to evade supervision, the "matching degree of sender and receiver email domains" feature was constructed by combining the domain information of payer and payee emails—if the payer's email uses a corporate domain while the payee's email uses a free personal domain, or the domains of the two parties are registered in multiple countries/regions, such atypical matching behaviors will be highlighted for identification. These derived variables are not simple combinations of fields, but rather deeply aligned with the business essence of fraudulent behaviors. They enrich the data feature system from key dimensions such as abnormal amounts, device associations, and account relationships, enabling the model to more accurately capture the differences between fraudulent and normal behaviors. This significantly enhances the model's ability to characterize complex fraud patterns such as "account transactions associated

with scam SMS sent via fake base stations" and "cross-regional batch transfers," providing more discriminative feature support for subsequent model training.

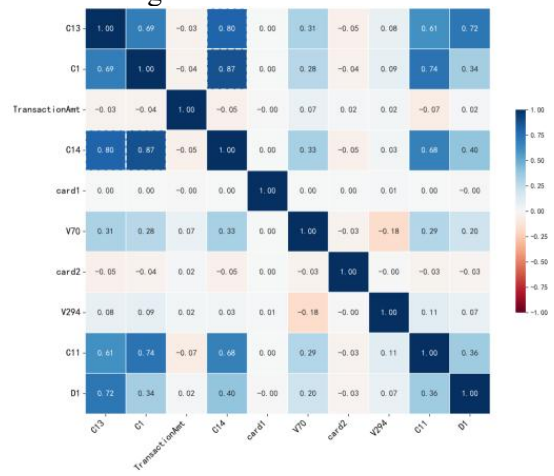

Figure 1. Heatmap of V-Type Features Highly Correlated with is Fraud

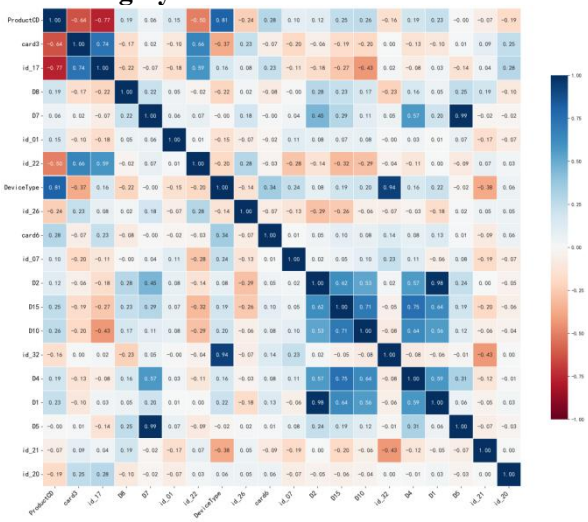

Figure 2. Heatmap of Non-V-Type Features Highly Correlated with is Fraud

After completing the initial construction of features, to ensure that the selected features not only have strong discriminative power but also meet the learning needs of different models, this study conducted a dual evaluation of feature importance using the two base models, LightGBM and XGBoost. This dual-model evaluation method can effectively avoid biases caused by the feature preferences of a single model and improve the reliability of feature selection. Among them, the LightGBM model uses two core methods, Gain and Split, to measure the contribution of variables to model performance: the Gain value calculates the total information gain brought by a feature in the splitting process of all decision trees, reflecting its overall role in improving the prediction accuracy of the model; the Split value counts the number of times a feature is selected as a split node, reflecting its decision participation frequency in the model construction process. These two indicators jointly support the feature importance ranking of the LightGBM model from different dimensions. The XGBoost model, on the other hand, focuses on evaluating the marginal effect of each variable in terms of information gain through the Gain value, that is, the incremental contribution of the feature to reducing the model's loss function each time it participates in the splitting of a decision tree, thereby quantifying the impact intensity of the feature on the model's output results. After obtaining the respective feature importance rankings of the two models, the intersection of the top-ranked variables was taken to screen out important features that perform prominently in both LightGBM and XGBoost and have consistent impacts, which were identified as the core feature set ultimately used for model training. This step ensures that the effectiveness of the selected features is not affected by differences in model structure, providing a stable input foundation for subsequent training. At the same time, for some high-value variables in the selected features (such as key features identified earlier, such as amount deviation and the number of device-account associations), further feature cross-construction was carried out (e.g., crossing "amount deviation" with "transaction time period" to generate a new feature of "amount deviation in different time periods"). Through this cross-combination, the feature expression space is further enriched, enabling the model to capture complex correlation patterns that cannot be covered by a single feature, thereby providing a stronger discriminative basis for model learning and helping to improve the accuracy of fraud behavior identification.

3.3 PCA Fusion Strategy and Mechanism

To effectively address the typical characteristics of telecom fraud data—extreme class imbalance (the ratio of normal transaction samples to fraud samples often exceeds 1000:1), high feature dimensionality (covering hundreds of features such as user identity, transaction trajectory, and device information), and strong time sensitivity (fraudulent behaviors are mostly short-term and sudden, requiring rapid detection)—this study specifically selects two ensemble models based

on Gradient Boosting Decision Trees (GBDT), namely LightGBM and XGBoost, as the base classifiers. Among them, LightGBM relies on the "histogram optimization" technique and the "Leaf-wise" decision tree growth strategy [7,8]. When processing large-scale high-dimensional data, it can significantly reduce memory usage and computational time, demonstrating superior training efficiency, which better adapts to the real-time analysis requirements of telecom fraud data. In contrast, XGBoost incorporates L1 and L2 regularization terms into the objective function, giving it stronger regularization capabilities that can effectively suppress the overfitting risk of the model on imbalanced class data. Meanwhile, its adaptive processing mechanism for missing features also enhances robustness to complex features. The two models exhibit distinct and complementary advantages in terms of training efficiency, regularization capabilities, and feature processing mechanisms, enabling them to address the core challenges of telecom fraud data from different dimensions and lay a foundation for the subsequent construction of more efficient fraud detection models.

Based on the raw sample data provided by Vesta Corporation, model training is directly conducted on the processed training set. The KFold cross-validation method is used to divide the training dataset into 5 subsets. In each validation process,80% of the data in each subset is used to construct the training sample, and the remaining 20% is used as the validation sample. Systematic training and parameter tuning are carried out in the XGBoost and LightGBM models, respectively.

The model training uses 5-fold cross-validation. The traditional KFold method (suitable for i.i.d. data) and the Time Series Split method are compared in the experiment, and K-Fold is ultimately selected for its better stability. The main parameter tuning range includes: max_depth=7, learning_rate=0.05, n_estimators=500, subsample=0.8, colsample_bytree=0.8, reg_alpha=0.5, reg_lambda=1. Early_stopping_rounds=100 is introduced during training to prevent overfitting.

To enhance model interpretability, the SHAP (SHapley Additive exPlanations) method is introduced for variable impact analysis [9,10]. After extracting the feature importance rankings from the two models, their intersection is taken as the high-value variable set. This is further combined with the Gain metric from model training for sorting, redundant or invalid features are eliminated, feature dimensionality is compressed, and training efficiency is improved.

To integrate the complementary advantages of the two models at the prediction level, an unsupervised linear dimensionality reduction method based on PCA is used to fuse their prediction outputs [11]. The specific process is as follows:

The prediction probabilities output by LightGBM and XGBoost are concatenated into a two-dimensional vector;

PCA is used to extract the first principal component as the fusion score;

A threshold is set based on the fusion score for final judgment.

4. Model Evaluation and Results

4.1 Model Validation

A cross-referenced evaluation method is used to measure the performance of the LightGBM and XGBoost models. Multi-dimensional metrics such as AUC, KS value, and recall rate are adopted for validation. The cross-validation results are shown in Table 1:

Table 1. Model Performance Comparison

Evaluation Metric	XGBoost	LightGBM
AUC(Mean)	0.9216	0.9293
Recall Rate	82.3%	80.2%
KS Value	0.710	0.698

During the cross-validation process of the LightGBM model, the average AUC of the training set reached 0.9302 (as Figure 3), with a standard deviation of 0.0107.

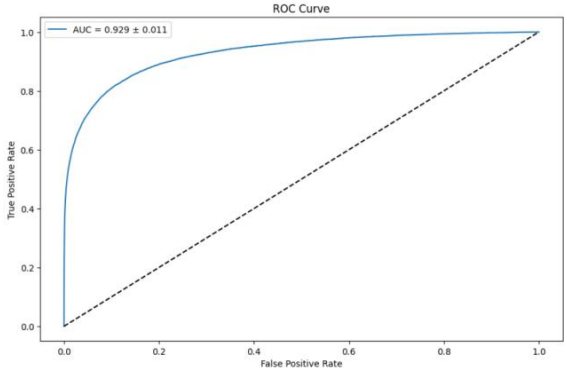

Figure 3. Training Set AUC of LightGBM Models

For the XGBoost model, the average AUC of the training set in cross-validation was 0.9232(as Figure 4), with a standard deviation of 0.0129.

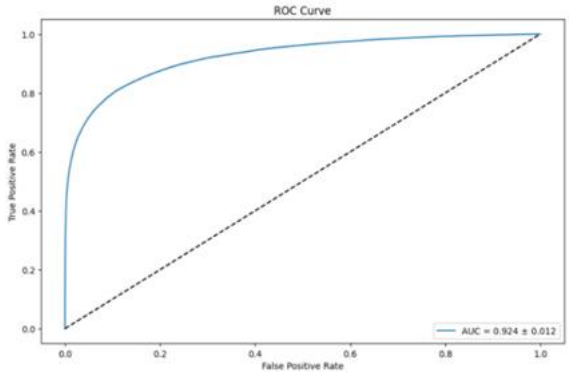

Figure 4. Training Set AUC of XGBoost Models

This indicates that both the LightGBM and XGBoost models have strong capabilities in distinguishing telecom fraud transactions (positive cases) from normal transactions (negative cases), and their performance is relatively stable with small fluctuations during multiple cross-validation processes. The KS values of both models are greater than 0.6, indicating strong discrimination between positive and negative cases. Meanwhile, the TPR of the XGBoost model is 0.823. And that of the LightGBM model is 0.802, both exceeding 80%.

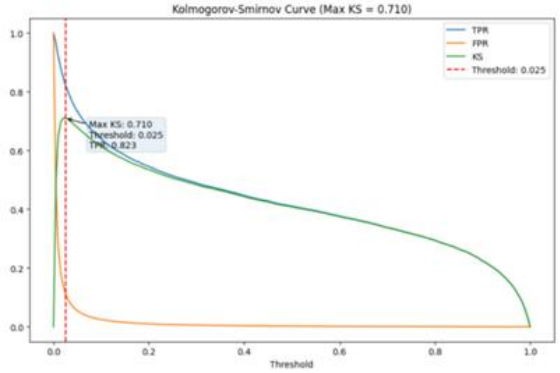

Figure 5. KS Curves of XGBoost Models

This shows that in scenarios sensitive to false negatives such as fraud detection, both models can effectively identify the vast majority of potential risk samples, meeting the rigid demand for high-risk event coverage in business.

These results fully demonstrate that both XGBoost and LightGBM exhibit good performance in telecom fraud transaction detection tasks (as shown in Figures 3 to 6).

With LightGBM showing a slight advantage in some metrics, providing solid data support for subsequent model selection and optimization.

After fusing the prediction results of the XGBoost and LightGBM models using PCA dimensionality reduction, the fused results are evaluated (as shown in Table 2).

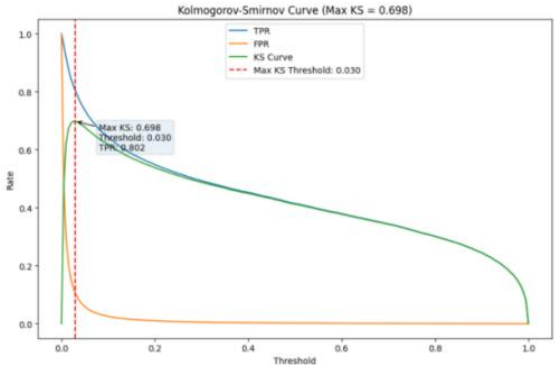

Figure 6. KS Curves of LightGBM Models

Table 2. Performance Evaluation of the Fused Model

Evaluation Metric	LightGBM	XGBoost	Conclusion
Correlation Confidence Interval	0.988–0.991	0.988–0.991	Stable correlation after fusion
Average Prediction Entropy	0.090 bits	0.089 bits	No additional uncertainty introduced after fusion

To verify the effectiveness of the PCA fusion method in the telecom fraud detection model, this study conducts a systematic theoretical comparative analysis between PCA fusion and commonly used model fusion methods in the industry, based on the ensemble learning theoretical framework and combined with the results of previous experimental data. The comparison reveals significant differences in the theoretical AUC (Area Under the ROC Curve), a core performance indicator, among different fusion methods: the simple averaging method achieves fusion by directly calculating the arithmetic mean of the prediction probabilities of the two base models (LightGBM and XGBoost), with a theoretical AUC value of 0.9255, which is lower than the 0.9302 of the PCA fusion method. This result indicates that PCA, through dimensionality reduction, can more accurately extract core discriminative information from the prediction results of the two base models, while effectively avoiding the noise accumulation problem that may arise from the "equal weighting of all prediction results" in the simple averaging method, making the fusion results more focused on features valuable for fraud detection. Although the weighted averaging method can assign different weights based on the independent performance of the two base models (e.g., assigning a larger weight

to the model with a higher AUC), its upper limit of theoretical AUC is only 0.9293, which is consistent with the performance of the optimal single model among the two base models and fails to break through the performance bottleneck of a single model. In contrast, the PCA fusion method successfully breaks through this upper limit with a theoretical AUC value of 0.9302, fully demonstrating its ability to deeply explore the complementarity of LightGBM and XGBoost in feature capture (such as LightGBM's sensitivity to short-term transaction anomalies and XGBoost's ability to grasp long-term behavioral trends) and achieve advantage superposition through principal component extraction. Additionally, considering operational convenience and objectivity in practical applications, when the PCA fusion method processes highly correlated data such as the prediction results of the two base models, its principal component loadings are completely naturally generated by the variance distribution characteristics of the data itself, without the need for manual weight presets or parameter adjustments. Compared with the weighted averaging method, which relies on empirical judgment to assign weights, this greatly reduces the risk of subjective factors interfering with the fusion results, further highlighting the dual advantages of PCA fusion in performance and practicality.

4.2 Model Prediction Results

In the telecom fraud transaction detection task, accurate predictions on the test set are achieved by integrating the XGBoost and LightGBM models. The model output is the probability of each transaction being fraudulent(isFraud). Some prediction results are shown in Table 3.

Table 3. Examples of Partial Prediction Results

Transaction ID	isFraud (Prediction Probability)
3663549	0.0033
3663551	0.0094
3663579	0.2913
3663581	0.2886

These specific prediction probability values intuitively reflect the degree of fraud likelihood for each transaction under the model's judgment. Transactions with lower values (e.g., 0.0033) indicate a very low probability of being identified as fraudulent under the current model evaluation system; while transactions with relatively higher values (e.g., 0.2913) suggest that relevant institutions or personnel need further attention due to their higher potential fraud risk.

5. Conclusion

The PCA-integrated ensemble learning model proposed in this study balances performance and interpretability, achieving an AUC of 0.9302 in telecom fraud detection tasks, demonstrating strong potential for business applications. Future research directions include: (1) exploring multi-institutional data fusion and privacy protection under a federated learning framework; (2) introducing time series analysis algorithms to enhance dynamic identification capability of fraud evolution; (3) combining knowledge graphs and graph neural networks to achieve networked traceability of fraud groups. Relevant results can be promoted and applied in fields such as financial risk control and payment security, contributing to the construction of an intelligent network security protection system.

Funding Projects

1. Undergraduate Research Training Project of Henan University of Technology–Research on Temporal Graph Intelligence Detection and Interpretability for Telecom Fraud (KYXL2025157); 2. Henan Province Key Technologies R&D Program–Key Technology for Real-Time Reconstruction of Grain Storage Temperature Field Based on Multi-Source Acoustic Data Drive(252102320199); 3. Henan Higher Education Teaching Reform Research and Practice Project(Postgraduate Education Category)–Innovative Design and Practice of SPOC and PBOPPSC Hybrid Teaching for High-Level Outstanding Engineering Talent Training (2023SJGLX173Y); 4. Teaching Reform and Practice of University Mathematics Courses under the "Four New" Background (Higher Education Press Special Project)–Hybrid Teaching Design and Practice for Innovative Talent Training(CMC20240610); 5. Henan University of Technology University-Level Specialty-Innovation Integrated Characteristic Course Construction Project–Probability Theory and Mathematical Statistics (2024ZCRH-15); 6. Undergraduate Teaching Research Project of the School of Science, Henan University of Technology–Research and Practice on Probability and Statistics Course Reform under

the BOPPPS Model(lxyjy202401).

References

[1] China Academy of Information and Communications Technology. White Paper on Prevention and Governance of Telecom and Online Fraud. 2023.

[2] Tianpei XU, Yongsheng LUO. A Credit Card Fraud Detection Model Based on Ensemble Learning. Information System Engineering, 2024, (01): 129-132.

[3] Y Zhang, et al. Gradient Boosting Machines: A Survey. ACM Computing Surveys, 2020, 53(5): 1–30.

[4] LightGBM Documentation. Microsoft, 2023.

[5] T Chen, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785–794.

[6] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2001.

[7] IEEE-CIS Fraud Detection Dataset. Kaggle, 2019.

[8] Heng WANG, Yanan JIANG, Xin ZHANG, et al. A Lithology Identification Method Based on the Gradient Boosting Algorithm. Journal of Jilin University (Earth Science Edition), 2021, (03): 940-950.

[9] Lundberg S M, Lee S I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 2017: 4765–4774.

[10] Wei ZHAO, Ming LI, Yue SUN. An Interpretable Fraud Detection Framework Combining GBDT and SHAP for Highly Imbalanced Data. Journal of Electronics & Information Technology, 2023, 45(8): 2801-2810.

[11] Jolliffe I T. Principal Component Analysis. Springer, 2002.

