Online Payment Fraud Detection Model Based On SMOTE-Weighted Stacking Framework

Yuxiang Gao [1, *], Yuxuan Xiao [2]

[1] School of Investigation, People's Public Security University of China, Beijing, China, 102600

[2] Mathematics and Computer, Science College Shantou University, Shantou, China, 515821

* Corresponding Author Email: pla_navy_gyx@163.com

Abstract. Online payment fraud detection plays a key role in protecting public property and curbing economic crimes. To solve the problem of sample imbalance in fraud detection, a classification model framework based on composite minority oversampling and weighted stack ensemble is proposed. SMOTE algorithm is used to synthesize and oversample minority samples, which effectively mitigates the impact of data distribution imbalance on classification performance. At the same time, weighted Stacking ensemble strategy is used to fuse the prediction results of multiple base learners, which improves the prediction accuracy of the model and enhances its robustness. It should be noted that both the base model and the meta-model adopt non-parametric modeling methods in this framework, which avoids the potential impact of model default bias on the integration effect. Experimental results based on real transaction data sets show that the proposed model has significant advantages over traditional ensemble learning methods in precision, recall and F1 - score under different sample imbalance ratios.

Keywords: Fraud Detection, Unbalanced Data, Ensemble Learning, SMOTE, Weighted Stacking Algorithm.

1. Introduction

With the continuous growth of the digital economy's share in China's economic system and the accelerated deep integration of various industries with digitalization, online payment has emerged as the core payment method for social and economic activities due to its unparalleled convenience. However, this rapid advancement has simultaneously given rise to a new form of economic crime—online payment fraud. Criminals threaten public property security and disrupt social and economic order by fabricating e-commerce platform interfaces and impersonating customer service, posing significant challenges to social harmony and stability. Although cross-border police collaboration has effectively cracked down on telecom fraud hubs in northern Myanmar, the covert and dynamic nature of fraudulent behaviors renders traditional manual detection and simple mathematical models inadequate for adapting to the evolving criminal techniques, thus creating an urgent need for innovative detection models to enhance anti-fraud efficiency.

Scholars have made notable progress in addressing the challenges of imbalanced data and model interpretability in fraud detection. Zhou et al. proposed an RF-GBDT intrusion detection model that applies random forest for feature transformation and uses a gradient boosting decision tree model for classification to solve the multi-classification problem of unbalanced data in network intrusion detection [1]. Zhu et al. developed a feature enhancement technology based on neural networks and exponential activation functions, which gradually optimizes reconstructed features during the training process to improve the accuracy of ensemble learning model construction [2]. Shi Hongbo pointed out that SMOTE oversampling is a popular method for improving the classification performance of imbalanced data, which can change the distribution of imbalanced datasets by adding generated minority samples [3]. Shi Jiaqi and Zhang Jianhua proposed a load forecasting method based on a multi-model fusion Stacking ensemble learning model [4]. Shu et al. introduced an XGBoost-based fraudulent transaction detection model to address the shortcomings of other machine learning models in the face of data imbalance [5].

Shi et al. further applied the XGBoost model to identify telecom fraud users by integrating relevant theories of telecom fraud, using the SMOTE algorithm to balance imbalanced sample data and ensure the prediction effect of the data model, and then screening identification variables of telecom fraud users through an embedding method [6]. To overcome the limitations of filter and wrapper feature selection algorithms based on evolutionary learning, Li Zhanshan et al. proposed a new wrapper feature selection algorithm, LGBFS (LightGBM Feature Selection) [7]. Li et al. studied design criteria for cost-sensitive losses and proposed two criteria under the Bayesian optimal classification theory framework [8]. Xu and Chi proposed a series of research strategies to improve the classification accuracy of machine learning algorithms for imbalanced datasets, focusing on data-level adjustments and classification model improvements [9]. Xu Jiwei and Yang Yun emphasized that ensemble learning, as a combinatorial optimization method, can not only derive better composite models by integrating multiple simple models but also enable researchers to design customized combination strategies for specific machine learning problems [10].

However, while machine learning methods enhance recall accuracy through feature engineering and deep learning, the "black-box" characteristic of deep learning models makes it difficult to explain their operational logic when fraud patterns change, requiring frequent fine-tuning with new data and resulting in insufficient agility in practical applications. Although ensemble learning has become a potential solution due to its interpretability advantages, it requires high correlation between data features and classification targets, and existing ensemble models still exhibit weaker performance than deep learning in unbalanced data scenarios. Current research on the integration of unbalanced learning and ensemble learning mostly remains at the level of independent application, lacking systematic framework design—particularly in fully considering the impact of sample distribution on the robustness of base models during construction and integration strategies.

To address these issues, this paper proposes an integrated SMOTE-Weighted Stacking framework, which improves fraud detection performance through a dual mechanism of data-layer preprocessing and model-layer integration. Before training the base models, the SMOTE algorithm is introduced to balance the class distribution of the training set, alleviating the model's bias toward the majority class. A weighted Stacking strategy is adopted to fuse multi-class base models, with a meta-learner dynamically allocating weights to the base models to enhance the capture of complex fraud patterns while maintaining the model's interpretability and robustness. The framework supports any type of base model and meta-learner, reduces the risk of overfitting by integrating heterogeneous models, and uses a weighting mechanism to dynamically adjust the contribution of each base model according to its performance on unbalanced data, thereby improving responsiveness to changes in fraud patterns. Using real transaction datasets, this study verifies the framework's effectiveness under different imbalance ratios. Compared with traditional ensemble methods (such as random forest and Adaboost), the proposed framework demonstrates significant advantages in accuracy, recall rate, and F1 score.

The research results not only provide an accurate and interpretable solution for online payment fraud detection but also explore the deep fusion path of unbalanced learning and ensemble learning, offering a new perspective for classification problems in dynamic and complex scenarios. Future research will further optimize the combination strategy of base models and introduce transfer learning technology to meet cross-domain fraud detection needs.

2. Principle of Stacking Classification Model Based on SMOTE Algorithm

The research results not only provide a solution with both accuracy and interpretability for online payment fraud detection, but also explore the deep fusion path of unbalanced learning and ensemble learning, providing a new perspective for classification problems in dynamic and complex scenarios. Subsequently, the combination strategy of the base model will be further optimized, and the transfer learning technology will be introduced to meet the cross-domain fraud detection needs.

Firstly, SMOTE (synthetic minority oversampling technique) is a classical algorithm to solve the problem of data imbalance. It improves the data distribution by artificially synthesizing minority

samples. Its principle involves the steps of sample distance calculation, neighbor determination, new sample generation and synthetic quantity calculation. Firstly,Euclidean distance is used to measure the similarity between samples, For two n-dimensional eigenvectors $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$ and $X_j = (x_{j1}, x_{j2}, \dots, x_{jn})$,Euclidean distance formula is :

$$d(x_i, y_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{1}$$

k nearest neighbor samples are determined for each minority class sample. When synthesizing new samples, randomly select a neighbor, by equation :

$$X_{new} = x_i + \delta \times (x_{nn} - x_i) \tag{2}$$

Generate a new sample, where the random number in the interval $[0,1]$ controls the position of the new sample on the line connecting the original sample and its neighbors. And when calculating the number of samples to be synthesized,let the number of minority samples be $N_{minority}$、the majority samples be $N_{majority}$, the expected proportion be α,then the number of samples to be synthesized

$$N_{synthetic} = \lceil \alpha \times N_{majority} - N_{minority} \rceil \tag{3}$$

By repeating the above random synthesis process for each minority class sample until it is reached $N_{synthetic}$,a balanced data set is finally formed to optimize the model training effect.

Secondly, Stacking is a powerful ensemble learning method. It combines the advantages of Bagging and Boosting, which can effectively reduce the variance and bias of the model. Through hierarchical structure, it can effectively integrate the prediction results of multiple base models and significantly improve the overall prediction performance. Stacking algorithm is composed of two layers. The first layer generates the probability prediction values of the base model to form the second level features as the input values of the second layer stacking. The specific working principle is shown in Figure 1 below:

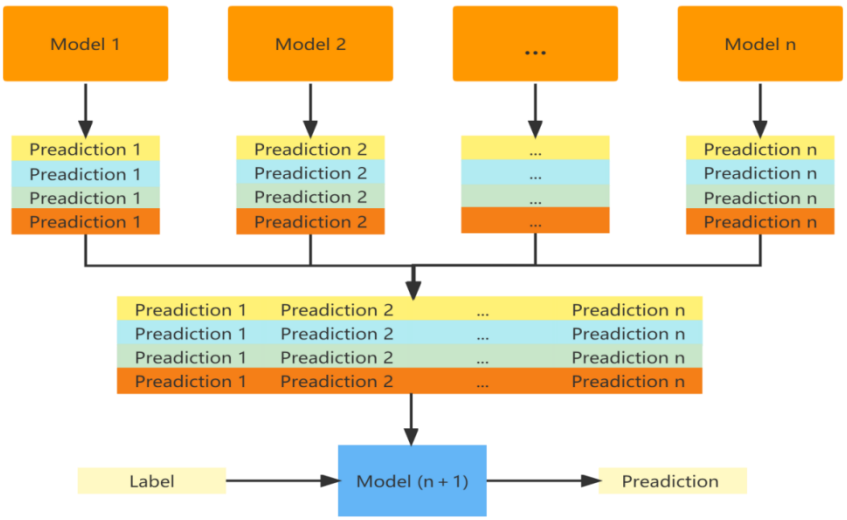

Figure 1 Principle of weighted Stacking model

First, the training data set $D = \{(x_i, y_i)\}_{i=1}^{n}$is divided into a test set and a training set,For each basis model j (j=1.2...m) Train on the training set to learn the relationship between input x and output y. During training on the training set, the training set is divided into a training set and a validation set again by using k-fold cross-validation. The prediction probability value of the base model for each fold is obtained on the validation set, and the prediction probability value of the base model for each fold is spliced to form a secondary feature of the test set under the current base model. For each

sample i(i=1.2...n),collecting prediction probability results under j basis models $\hat{y}_i^{(1)}, \hat{y}_i^{(2)} \ldots\ldots \hat{y}_i^{(m)}$,A new training set sample is formed by combining the actual target values as the input values of the second layer stacking.

$$D_{new} = \left\{\left(\left(\hat{y}_1^{(1)}, \hat{y}_1^{(2)} \ldots . \hat{y}_1^{(m)}\right), y_1\right), \left(\left(\hat{y}_2^{(1)}, \hat{y}_2^{(2)} \ldots . \hat{y}_2^{(m)}\right), y_2\right) \ldots . \left(\left(\hat{y}_n^{(1)}, \hat{y}_n^{(2)} \ldots . \hat{y}_n^{(m)}\right), y_n\right)\right\} \quad (4)$$

In the second layer of stacking, a new training set$D_{new}$is used to train the meta model,which learns how to generate the final prediction output based on the prediction probability results of the base model,enhancing generalization performance and prediction accuracy. The meta model can be any model, and if the meta model is linear regression, the goal is to find a set of weights that minimize the difference between the weighted prediction of the base model and the actual target value y_i.

3. Data analysis

3.1. Data sources and Experimental settings

The data for this article is derived from the Kaggle competition public dataset, which is a labeled online payment fraud detection dataset specifically used to train online payment fraud detection models. This dataset belongs to a typical unbalanced distribution dataset. Each sample contains multiple features and a label of fraud. The data structure is shown in Table 1 below:

Table 1 Description of characteristic variables and response variables

Variable name	data type	explain
step	numeric	The unit of time (1 step = 1 hour) indicates the point in time at which the transaction occurs.
type	character string	The type of online transaction (e.g. transfer, payment, cash deposit/withdrawal, etc.).
amount	numeric	Transaction amount (usually floating point, in monetary units).
nameOrig	character string	Customer ID (may be anonymous account name or user) initiating the transaction.
oldbalanceOrig	numeric	The account balance of the originator before the transaction.
newbalanceOrig	numeric	The account balance of the originator after the transaction.
nameDest	character string	Identification of the recipient of the transaction (which may be an anonymous account name or user).
oldbalanceDest	numeric	The initial balance of the recipient's account prior to the transaction (zero or missing if the transaction type is non-transfer).
newbalanceDest	numeric	The new balance of the recipient's account after the transaction (0 or missing if the transaction type is non-transfer).
isFraud	Boolean type(0/1)	Tag variable that identifies whether a transaction is fraudulent (0 for normal, 1 for fraudulent).

The label variable isFraud often presents a severe class imbalance (a much lower proportion of fraud samples than normal samples), which needs to be mitigated by data preprocessing (such as SMOTE algorithm) or model optimization. Numeric features (such as amount, oldbalanceOrig, newbalanceOrig, etc.) can be used to analyze changes in transaction amounts and account balance anomalies to help identify fraud patterns. Transaction type may be associated with fraud risk (e.g. unusual transfer type may suggest fraud). nameOrig and nameDest are usually anonymous identifiers that cannot be directly associated with real user information, and transaction patterns need to be indirectly mined through other features. For data preprocessing, we did the following:

(1) Coding categorical variables such as type (such as single hot coding or label coding).

(2) Standardize or normalize numerical features to improve model training efficiency.

(3) Handling missing values by statistical methods (such as mean, median) or machine learning algorithms.

In addition, we divide the dataset randomly and construct several sub-models accordingly. SMOTE algorithm is introduced to deal with unbalanced samples before training each sub-model. On the one hand, the proposed method can effectively alleviate the problem of prediction accuracy degradation caused by unbalanced samples. The code logic of cross-validation is to generate sub-models, divide the sub-models into K folds, and then train the models for each fold. Once the result is obtained, a new output matrix is generated, which is used as the input feature of the second layer model (model2), while Y on the training set is still used to train model2. Among them, model1 is the model of the sub-model in the first layer, and model2 is the model of the subsequent second layer. After import, it can be judged whether it is a fraud transaction. The data is then normalized to form a data set. By Monte Carlo simulation, the model will be trained ten times and get ten results randomly. The mean and standard deviation of these ten results will be calculated as the basis for evaluating the effect of this method. In general, higher accuracy indicates stronger predictive performance of the model; smaller standard deviation indicates more robust model. In this paper, we will use a variety of 0-1 ratio data to carry out experimental verification.

In the model validation process, Accuracy, AUC score and F1 score are three core evaluation indicators, which measure model performance from different dimensions. The specific explanations and applicable scenarios are shown in Table 2 below:

Table 2 Application Scenarios of Evaluation Indicators

Index	Focus	Applicable scenarios	Limitations
accuracy rate	global prediction accuracy	categorical equilibrium data	Misleading high in unbalanced data
AUC	positive and negative class discrimination	Imbalanced data, high generalization performance required	Does not reflect performance at specific thresholds
F1 -score	Precision and recall balance	Need to balance missed and misjudged cost tasks	Assessment at a single threshold

3.2. Analysis of prediction and comparison results

Monte Carlo simulation method is used to calculate the final evaluation index in order to evaluate the comprehensive performance of the proposed method comprehensively and accurately. According to the method, a training set and a test set are divided for a plurality of times, and classification evaluation indexes, such as accuracy, F1 score, AUC and the like, are respectively calculated for the data set formed after each division. Then, calculate the mean and standard deviation of these multiple calculations. Among them, the larger the mean, the better the classification performance, which means that the average performance of the model in multiple experiments is outstanding; the smaller the standard deviation, the stronger the robustness of the model, that is, the performance fluctuation of the model under different data set division is small, and the stability is good. Because the label imbalance of samples in the dataset is significant, that is, there is a large difference in the number of fraudulent trading samples and normal trading samples, this imbalance will have an adverse impact on model training and prediction. Therefore, this paper introduces a method based on SMOTE (Synthetic Minority Over-sampling Technique) Multiple ensemble learning models for algorithms, including XGBoost based on SMOTE algorithm (eXtreme Gradient Boosting), Random Forest Based on SMOTE Algorithm (Random Forest), lightGBM (Light Gradient Boosting Machine) based on SMOTE algorithm, and GBDT (Gradient Boosting Decision Tree) based on SMOTE algorithm. SMOTE algorithm is used to oversample minority samples (fraudulent transaction samples) to balance the distribution of sample categories, thus improving the learning ability and prediction accuracy of the model for minority samples. From Table 3 and Figure 2, in terms of accuracy, the Emodel method proposed in this paper has an accuracy mean square error of 0.9956 and a standard deviation of only 0.0003, which is much higher than 0.9167 and 0.0114 of SMOTE+GBDT, and is

also better than SMOTE+XGBoost, SMOTE+RF, SMOTE+lightGBM, and has excellent stability. In F1 score, the mean square error of Emodel is 0.9712, and the standard deviation is 0.0020. Compared with SMOTE+GBDT 0.7118 and other models, it has obvious advantages in balance accuracy and recall. In AUC index, Emodel mean square error is 0.9554, standard deviation is 0.0031, the ability to distinguish fraud from normal trading is strong and stable, while SMOTE+GBDT is only 0.6550. Although the other models have performance, their comprehensive stability is not as good as Emodel. Overall, Emodel outperforms several other methods that combine SMOTE algorithms with different ensemble learning models in predictive performance and stability, enabling more accurate and stable identification of fraudulent transactions in online payment fraud detection.

Table 3 Prediction results of the proposed method and benchmark model

Model name	accuracy rate	standard deviation	F1-score	standard deviation	AUC	standard deviation
Emodel	0.9956	0.0003	0.9712	0.002	0.9554	0.0031
SMOTE+GBDT	0.9167	0.0114	0.7118	0.0198	0.655	0.0148
SMOTE+XGBoost	0.9804	0.001	0.8924	0.0046	0.8324	0.0063
SMOTE+RF	0.9939	0.0006	0.9608	0.0036	0.9938	0.0059
SMOTE+lightGBM	0.9815	0.0011	0.8975	0.0049	0.9396	0.0063

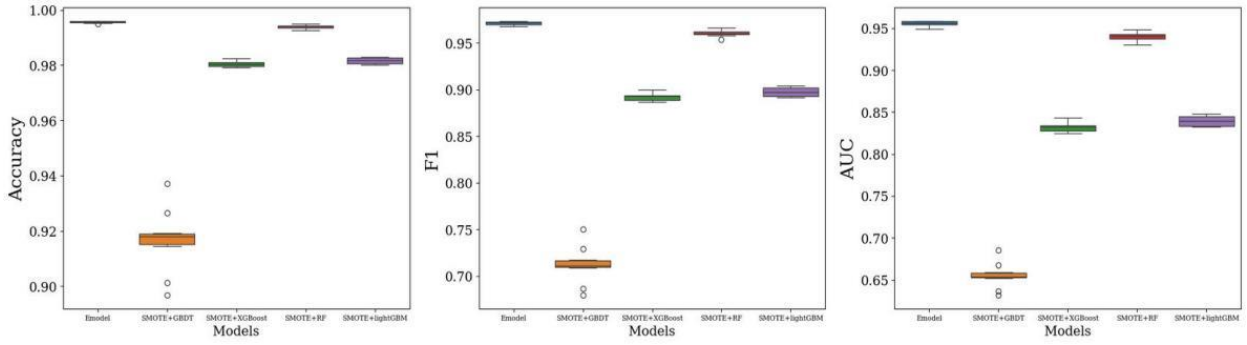

Figure 2 Distribution of evaluation indicators for different prediction methods

After changing the data ratio, from Table 4 and Figure 3, in terms of accuracy, the Emodel method proposed in this paper has an accuracy mean square error of 0.9859 and a standard deviation of 0.0013, which are still much higher than 0.9208 and 0.0141 of SMOTE+GBDT, and are also better than SMOTE+XGBoost, SMOTE+RF, SMOTE+lightGBM, and have excellent stability. In F1 score, the mean square error of Emodel is 0.9373, and the standard deviation is 0.0066. Compared with SMOTE+GBDT 0.7640 and other models, the advantage in balance accuracy and recall is still obvious. In AUC index, Emodel mean square error is 0.9107, standard deviation is 0.0116, and the ability to distinguish fraud from normal trading is strong, while SMOTE+GBDT is only 0.7044. Although the other models have performance, their comprehensive stability is not as good as Emodel. Overall, Emodel still outperforms several other methods combining SMOTE algorithm and different ensemble learning models in prediction performance and stability after changing data matching, and can identify fraudulent transactions more accurately and stably in online payment fraud detection.

After changing the data ratio, from Table 5 and Figure 4, in terms of accuracy, the Emodel method proposed in this paper has an accuracy mean square error of 0.9931 and a standard deviation of 0.0009, which is much higher than 0.9345 and 0.0049 of SMOTE+GBDT, and is also better than SMOTE+XGBoost, SMOTE+RF, SMOTE+lightGBM, and has excellent stability. In F1 score, the mean square error of Emodel is 0.9876, and the standard deviation is 0.0016. Compared with SMOTE+GBDT 0.8944 and other models, the advantage in balance accuracy and recall is still obvious. In AUC index, Emodel has a strong ability to distinguish fraud from normal trading, while SMOTE+GBDT is only 0.8621. Although the other models have performance, their comprehensive stability is not as good as Emodel. Overall, Emodel improved prediction performance and stability after increasing data volume and changing data matching, still surpassing several other methods

combining SMOTE algorithm with different ensemble learning models, proving that it can more accurately and stably identify fraudulent transactions in online payment fraud detection.

Table 4 The prediction results of the unbalanced ratio of 100:6

Model name	accuracy rate	standard deviation	F1-score	standard deviation	AUC	standard deviation
Emodel	0.9859	0.0013	0.9373	0.0066	0.9107	0.0116
SMOTE+GBDT	0.9208	0.0141	0.7640	0.0271	0.7044	0.0232
SMOTE+XGBoost	0.9735	0.0015	0.8931	0.0057	0.8435	0.0074
SMOTE+RF	0.9819	0.0027	0.9205	0.0125	0.8922	0.0196
SMOTE+lightGBM	0.9730	0.0019	0.8918	0.0079	0.8411	0.0102

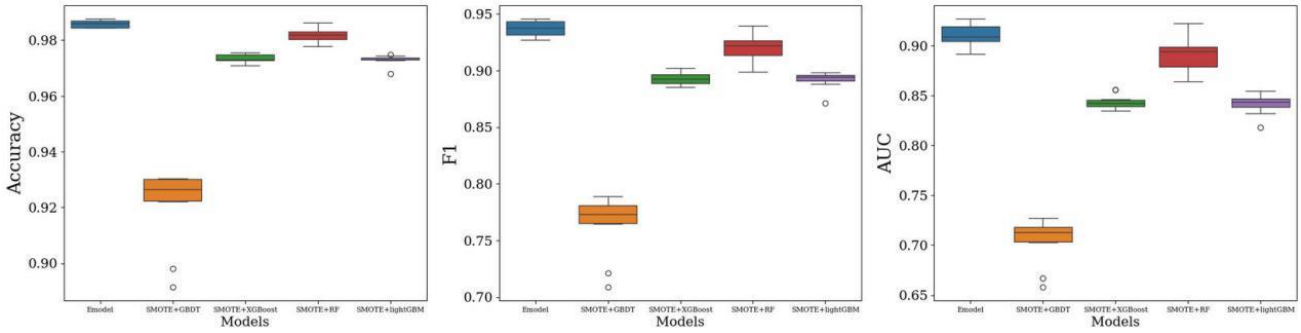

Figure 3 The boxplot of the unbalanced ratio of 100:6

Table 5 The prediction results of the unbalanced ratio of 30:6

Model name	accuracy rate	standard deviation	F1-score	standard deviation	AUC	standard deviation
Emodel	0.9931	0.0009	0.9876	0.0016	0.9835	0.0021
SMOTE+GBDT	0.9345	0.0049	0.8944	0.0073	0.8621	0.0089
SMOTE+XGBoost	0.9844	0.0018	0.9727	0.0032	0.9602	0.0046
SMOTE+RF	0.9915	0.0009	0.9848	0.0016	0.9808	0.0023
SMOTE+lightGBM	0.9851	0.0009	0.9739	0.0015	0.9620	0.0024

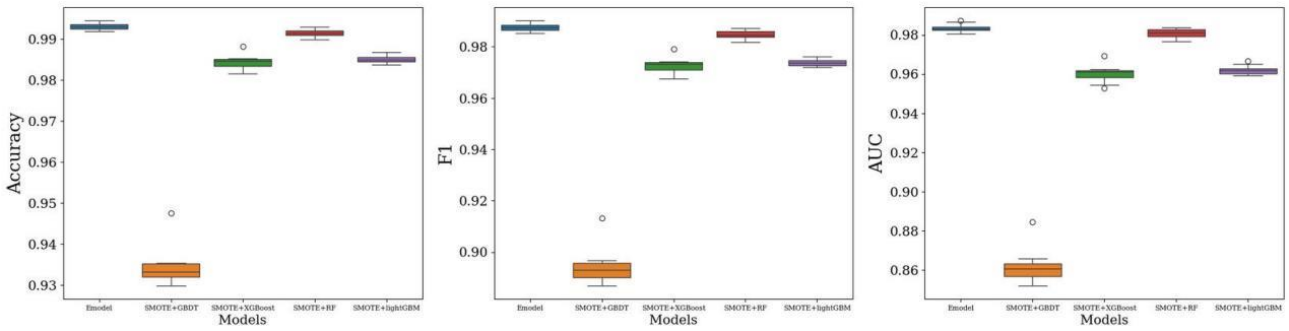

Figure 4 The boxplot of the unbalanced ratio of 30:6

4. Conclusions

Online payment fraud detection model based on SMOTE-weighted Stacking framework focuses on Kaggle competition open unbalanced online payment dataset, which contains multidimensional features such as transaction time, type, amount, account balance change and fraud label (isFraud). The proposed classification model integrates SMOTE data balance technique with weighted Stacking ensemble learning strategy, and provides an effective solution to this class imbalance problem. Specifically, SMOTE algorithm is used to generate synthetic samples to enhance the characterization ability of minority classes (fraudulent transactions), and the boundary sample enhancement strategy is combined to optimize the distribution of synthetic samples to mitigate the overfitting risk of

traditional oversampling. The weighted Stacking framework integrates the diversity output of heterogeneous base learners (such as XGBoost, random forest, etc.), improves the generalization of the model by means of the dynamic weighting mechanism of meta-learners (such as logistic regression), and demonstrates excellent mean and low standard deviation of accuracy rate, F1 score, AUC and other indicators in many experiments of Monte Carlo simulation, which verifies its accurate and robust fraud identification ability in online payment scenarios.

Although the framework has shown good performance in the field of financial risk control, there is still room for improvement in view of the high dimension, real-time and potential noise characteristics of online payment data. Future research will focus on two aspects: one is to optimize SMOTE algorithm, combined with undersampling technology (e.g. Tomek Links) or improved algorithms (such as ADASYN) reduce noise introduction, and design dynamic adaptive sampling strategy to adjust sampling ratio in real time according to transaction characteristic distribution and model feedback, so as to avoid sample redundancy or shortage caused by fixed ratio; The second is to strengthen the theoretical analysis and interpretability of the model, explain the rationality of the weighting mechanism through game theory or information entropy theory, combine SHAP, LIME and other tools to analyze the decision-making logic of the model for high-risk transactions (such as abnormal transfer amount, asymmetric balance change), and improve its credibility in highly sensitive scenarios such as financial supervision and judicial evidence collection.

References

[1] Zhou Jieying, He Pengfei, Qiu Rongfa, et al. Intrusion detection based on fusion of random forest and gradient lifting tree [J]. Journal of Software, 2021, 32 (10):3254-3265.

[2] Zhu Chengyuan. Online payment fraud detection based on ensemble learning [D]. Hainan Normal University, 2024.

[3] SHI Hongbo, CHEN Yuwen, CHEN Xin. Review of SMOTE oversampling and its improved algorithm [J]. Journal of Intelligent Systems, 2019, 14 (06):1073-1083.

[4] Shi Jiaqi, Zhang Jianhua. Load forecasting method based on multi-model fusion Stacking ensemble learning method [J]. Journal of China Electrical Engineering, 2019, 39 (14):4032-4042.

[5] Shu Pengfei. Research on B2C fraud transaction detection model [J]. Fujian Computer, 2019, 35 (12):23-25.

[6] Shi Jianwei. Research on Telecom Fraud User Identification Based on XGBoost Algorithm [D]. Dongbei University of Finance and Economics, 2021.

[7] Li Zhanshan, Yao Xin, Liu Zhaogeng, et al. Feature Selection Algorithm Based on LightGBM [J]. Journal of Northeastern University (Natural Science Edition), 2021, 42 (12):1688-1695.

[8] Li Qiujie, Zhao Yaqin, Gu Zhou. Design of loss function in cost-sensitive learning [J]. Control Theory and Applications, 2015, 32 (05):689-694.

[9] Xu Lingling, Chi Dongxiang. Machine learning classification strategy for unbalanced data sets [J]. Computer Engineering and Applications, 2020, 56 (24):12-27.

[10] Xu Jiwei, Yang Yun. Ensemble Learning Methods: A Review [J]. Journal of Yunnan University (Natural Science Edition), 2018, 40 (06):1082-1092.