Credit Card Fraud Prediction Based on the Improved Data Balancing Technique and the Gradient Boosting Algorithm

Ying Jin[1,*], Yanming Chen[2]

[1]Student, Shantou University, Shantou, China
[2]Student, Shantou University, Shantou, China

*21yjin1@stu.edu.cn,
[2]21ymchen@stu.edu.cn

Abstract. This paper aims to build a credit card transaction fraud classification model by combining improved data balancing techniques and gradient boosting algorithms. After data cleaning and preprocessing, we applied random oversampling, SMOTE oversampling, random undersampling, and Tomek Links undersampling methods to deal with the highly imbalanced dataset. Afterwards, we established classification models using LightGBM, XGBoost and CatBoost algorithms for comparative experiments. Finally, we selected the best performing gradient boosting model under each data balancing method as the first layer models of the Stacking algorithm, and the classification tree model as the second layer model. Its accuracy and F1-score on the testing set reached 0.98.

Keywords: Credit Card Fraud; Imbalanced Data; Gradient Boosting; Stacking; Financial Transaction Security; Classification

1 Introduction

Credit card fraud is an increasingly threatening problem which will not only damage the property of individual users, but also bring negative effects to financial institutions such as banks and even the whole society. Therefore, developing an accurate credit card fraud prediction model has become an important topic in the financial field.

In the pervious studies, many scholars have carried out research on credit card fraud prediction, such as "Fraud Prediction of Credit Card Customers Based on Xgboost Model and Multi-Layer Perception Model" [1], "Credit Card Fraud Prediction and Classification using Deep Neural Network and Ensemble Learning" [2], "Credit Card Fraud Detection Using Artificial Neural Networks and Random Forest Algorithms" [3]. However, conventional research typically employs a single method for addressing data imbalance. This approach may result in model bias and diminish overall performance when confronted with highly imbalanced credit card fraud data. In this study, we employ four methods to handle the highly imbalanced dataset and integrate them with the Gradient Boosting algorithm and Stacking algorithm. This combined approach serves

F. Cao et al. (eds.), *Proceedings of the 2023 5th International Conference on Economic Management and Cultural Industry (ICEMCI 2023)*, Advances in Economics, Business and Management Research 276,
https://doi.org/10.2991/978-94-6463-368-9_74

to enhance the model's robustness and generalization capability, enabling more accurate predictions.

2 Theoretical Foundation

Machine learning [4] is a crucial branch of artificial intelligence algorithms that aims to analyze and process data, construct models, and train them to make predictions or decisions on unknown data. Ensemble learning is a machine learning approach that enhances prediction performance, accuracy, and stability by combining the predictions of multiple individual models.

Oversampling and Undersampling techniques are employed when there is a significant imbalance in the number of samples between different classes in a dataset. These techniques involve adjusting the proportions of samples from different classes to improve the accuracy and generalization ability of the model.

3 Materials and Methods

3.1 Dataset Used in the Study

This paper selects the transaction records of European cardholders in two days from kaggle.com as the data set for the study. There are 31 variables in this dataset, the dependent variable is "Class", where "1" represents fraudulent transactions, "0" represents non-fraudulent transactions. And the remaining 30 are listed as independent variables, "Time" represents the number of seconds between each transaction and the first transaction in the dataset, "Amount" represents the amount of the transaction, and the other V1 to V28 are the results after PCA reduction to protect users' personal information.

The results of descriptive statistics for some variables are shown in Table 1.

Table 1. Descriptive statistical results of some variables

	Mean	Std	Min	Q1	Median	Q3	Max
Class	0.00	0.04	0.00	0.00	0.00	0.00	1.00
Amount	88.35	250.12	0.00	5.60	22.00	77.16	25691
Time	94813	47488	0.00	54201	84692	139320	172792
V5	0.00	1.38	-113.74	-0.69	-0.05	0.61	34.80
V9	0.00	1.10	-13.43	-0.64	-0.05	0.60	15.59
V24	0.00	0.61	-2.84	-0.35	0.04	0.44	4.58

The distribution of these variables is shown in Figure 1.

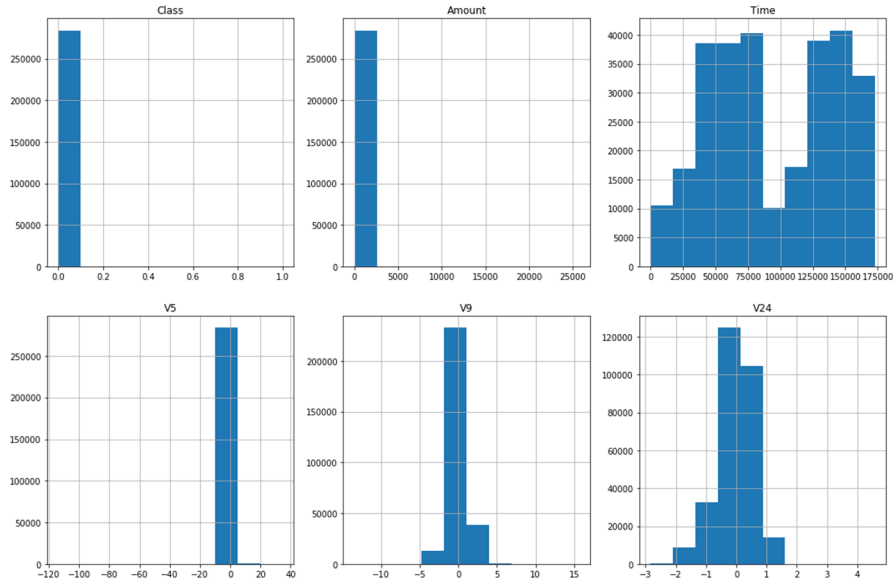

Fig. 1. The distribution of some variables

3.2 Data Standardization

Compared with normalization, standardization can better retain the distribution characteristics of the data, and can eliminate the scale differences between different variables. For large samples, standardization is conducive to improving the robustness of data. Therefore, here we use Standardization for all numerical variables.

First, the original data set is divided into a training set and a test set in a ratio of 7:3, and then formula (1) is used to standardize each data value.

$$x_{new} = \frac{x - \bar{x}}{s} \tag{1}$$

In this formula, $\bar{x}$ is the mean of the sample data, s is the standard deviation of the sample data, x and x_{new} are the data before and after standardizing.

Due to the test set represents the real data of the unknown, if use the mean and standard deviation of the test set to standardize themselves, is likely to interfere with the training process, and too optimistic to model performance evaluation. Therefore, should use the mean and standard deviation of the training set, to better test the model performance.

3.3 Data Correlation Exploration

Point-biserial correlation is used to analyze the correlation between continuous and binary variables. And the P-Value is used to judge whether the correlation is significant,

if the P-Value is less than 0.05, the correlation can be considered significant, otherwise, the correlation is not significant.

According to Point-biserial analysis, most of the independent variables have significant correlation with the "Class", and the top five variables with the highest correlation are illustrated in Table 2.

Table 2. Top five variables with higher correlations

	Correlation Coefficient	P-Value
V17	-0.326	0.00
V14	-0.302	0.00
V12	-0.260	0.00
V10	-0.216	0.00
V16	-0.196	0.00

Only three independent variables, "V22", "V23" and "V25", show an insignificant correlation with "Class", and the results are shown in Table 3.

Table 3. Variables with insignificant correlations

	Correlation Coefficient	P-Value
V25	0.003	0.07
V23	-0.002	0.15
V22	0.001	0.66

Then, we use the Pearson correlation coefficient to calculate the Thermodynamic matrix between the 30 independent variables, as shown in Figure 2.

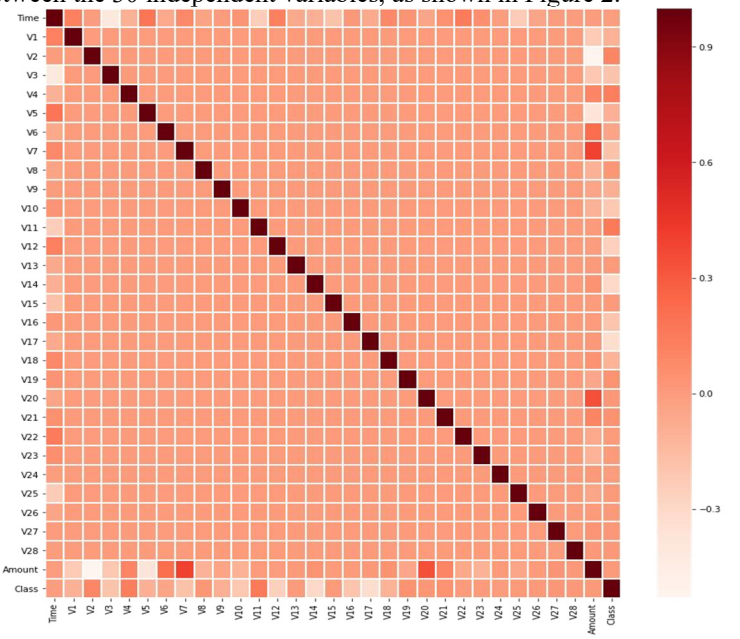

Fig. 2. Thermodynamic matrix between the 30 independent variables

Figure 2 shows that there is no significant correlation from V1 to V28, but some of them such as "V20", "V7", "V2" and "V5" have certain correlation with "Time" or "Amount".

3.4 Imbalanced Data Processing

In the dataset of this study, only 492 of the total 284807 transactions are fraud transactions, and the proportion is only 0.172%, indicating that the dataset is highly unbalanced. Therefore, we need to use Oversampling and Undersampling algorithms in order to balance data. In simple terms, Oversampling increases the number of fraudulent transactions from 492 to 284315, Undersampling reduces the number of non-fraudulent transactions from 284315 to 492.

Then, four algorithms including Random Oversampling, SMOTE Oversampling, Random Undersampling and Tomek Links Undersampling are used to balance the data in this study [5]. Random Oversampling and Radom Undersampling algorithm refer to the random replication or removal of some samples. SMOTE Oversampling is used to balance the dataset by creating new synthetic samples [6]. Tomek Links Undersampling algorithm removes specific pairs of samples based on the distance between samples [7].

3.5 Model Building

Firstly, we use XGBoost, LightGBM and CatBoost algorithms to establish classification models for the datasets after four data balancing techniques. Secondly, uses Bayesian search to adjust parameters. Afterwards, the most suitable model under each method is judged through the calculated accuracy of training set, test set accuracy and test set F1-score.

Finally, we take the best performing model corresponding to these four datasets as the first layer model of the Stacking algorithm [8], and the Decision Tree model with the maximum depth of 4 as the second layer model. The modeling process is shown in Figure 3.

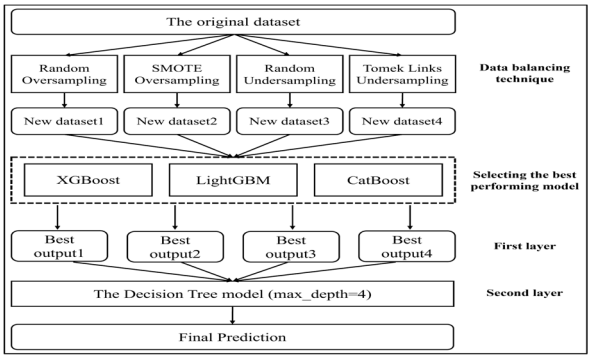

Fig. 3. The main process of improved data balancing based on gradient boosting and Stacking algorithms

3.6 Xgboost Importance Ranking for Feature Analysis

The feature importance of each variable can be generated by XGBoost model, and the final result is shown in Figure 4.

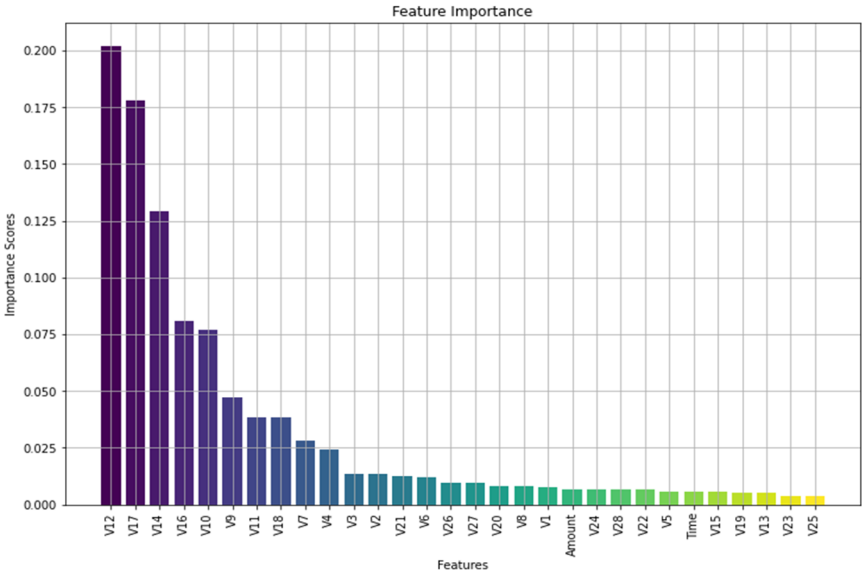

Fig. 4. XGBoost feature importance ranking

4 Experiments & Results

Firstly, we applied the XGBoost, LightGBM and CatBoost models to the processed datasets, and the results are respectively shown in Table 4, Table 5 and Table 6.

Table 4. XGBoost algorithm modeling results

Dataset	Training set accuracy	Testing set accuracy	Testing set F1-score
Random Oversampling	0.98	0.93	0.92
SMOTE Oversampling	0.99	0.92	0.94
Random Undersampling	0.97	0.91	0.90
Tomek Links Undersampling	0.98	0.92	0.92

Table 5. LightGBM algorithm modeling results

Dataset	Training set accuracy	Testing set accuracy	Testing set F1-score
Random Oversampling	0.99	0.93	0.94
SMOTE Oversampling	0.98	0.88	0.91
Random Undersampling	0.96	0.85	0.85
Tomek Links Undersampling	0.95	0.86	0.88

Table 6. CatBoost algorithm modeling results

Dataset	Training set accuracy	Testing set accuracy	Testing set F1-score
Random Oversampling	0.94	0.85	0.86
SMOTE Oversampling	0.95	0.88	0.91
Random Undersampling	0.87	0.83	0.83
Tomek Links Undersampling	0.98	0.92	0.93

The results indicate that the best performing model is LightGBM for dataset using Random Oversampling, XGBoost for datasets using SMOTE Oversampling and Random Undersampling algorithms, and CatBoost for dataset using Tomek Links Undersampling.

Secondly, we employ the Stacking algorithm to build the classification model, and evaluate its performance with accuracy, precision, recall, and F1-score. The results are shown in Table 7.

Table 7. Stacking algorithm modeling results

Metrics	Result
Training set accuracy	1.0
Test set accuracy	0.98
Test set precision	0.99
Test set recall	0.98
Test set F1-score	0.98

Finally, we conduct 10-fold cross validation of the model in order to improve the reliability of the evaluation results. The ROC-AUC curve is shown in Figure 5.

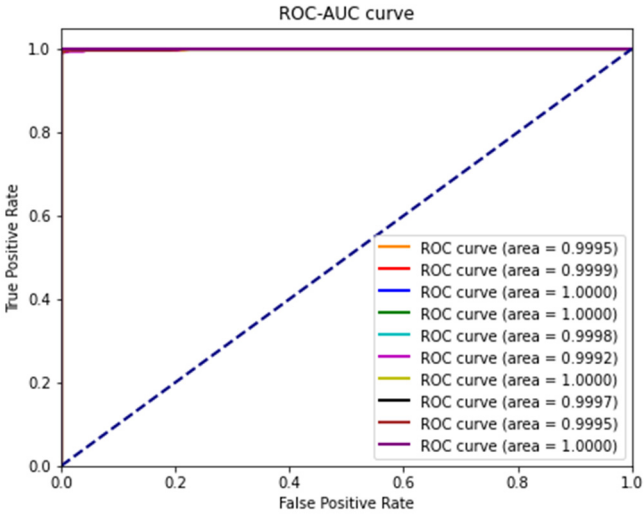

Fig. 5. The ROC-AUC curve

5 Conclusion

In this paper, we construct a Stacking model based on four data balancing techniques and gradient boosting algorithm for predicting credit card fraud transactions. Compared with the traditional machine learning model, the proposed model can better deal with highly imbalanced dataset and make more accurate predictions. However, due to the complexity of the algorithms used, this model is more time-consuming, which requires further feature engineering and attempts to reduce the complexity of the model.

References

1. Zhu, X.P., Li, Q.N., Huang, Y., Huang, L. and Deng, P.Y. (2022) Fraud Prediction of Credit Card Customers Based on Xgboost Model and Multi-Layer Perception Model. In: IEEE International Conference on Advances in Electrical Engineering and Computer Applications. Dalian. pp. 557-561. 10.1109/AEECA55500.2022.9919097.
2. Khan, F. N., Khan, A. H. and Israt, L. (2020) Credit Card Fraud Prediction and Classification using Deep Neural Network and Ensemble Learning. In: IEEE Region 10 Symposium. Dhaka. pp. 114-119. 10.1109/TENSYMP50017.2020.9231001.
3. Pradhan, S.K., Krishna Rao, N.V., Deepika, N.M., Harish P., Kumar M.P. and Kumar, P.S. (2021) Credit Card Fraud Detection Using Artificial Neural Networks and Random Forest Algorithms. In: International Conference on Electronics, Communication and Aerospace Technology. Coimbatore. pp. 1471-1476. 10.1109/ICECA52323.2021.9676142.
4. Gupta, Y. (2022) Using of Machine Learning Techniques to detect Credit Card Frauds. In: OITS International Conference on Information Technology. Bhubaneswar. pp. 124-128. 10.1109/OCIT56763.2022.00033.

5. Baabdullah, T., Alzahrani, A. and Rawat, D.B. (2020) On the Comparative Study of Prediction Accuracy for Credit Card Fraud Detection wWith Imbalanced Classifications. In: Spring Simulation Conference. Fairfax. pp. 1-12. 10.22360/SpringSim.2020.CSE.004.
6. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C. (2004) A study of the behavior of several methods for balancing machine learning training data. J. Acm Sigkdd Explorations Newsletter., 6(1):20-29. 10.1145/1007730.1007735.
7. Cao, S.Q., Wang, S.T., Chen, X.F. (2008) Posterior-probability-based Feature Selection Algorithm for Imbalanced Datasets. J. Computer Engineering., 34(19):1-3. 10. 3901 /JME. 2008.11.304.
8. Abbasi, M. and Shah, M.A. (2022) Credit card fraud detecting using machine learning classifiers in stacking ensemble technique. In: Competitive Advantage in the Digital Economy. pp. 76-81. 10.1049/icp.2022.2044.