ISSN: 2664-9640

Credit Card Fraud Detection: Enhancing Accuracy through Machine Learning Techniques

Chaoran Wu[1,a,*], Xiuliang Yu[1], Chunhai Chen[1], Zejiong Zhou[2]

[1]School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, 233030, China

[2]School of Economics, Anhui University of Finance and Economics, Bengbu, 233030, China

[a]3286211516@qq.com

Abstract

In order to improve the accuracy of credit card fraud detection, this paper uses the experimental data provided by Kaggle website to do the following rigorous processing. Firstly, this paper analyzes the data visually to explore the distribution of the data and the relationship between each other. Considering the diversity of the features in the data, this paper uses the random forest feature importance method to filter the data to select the features that have a greater impact on fraud detection; Aiming at the problem of extreme imbalance between positive and negative samples, this paper uses Smote oversampling processing to achieve the balance of positive and negative samples. On this basis, this paper constructs a credit card transaction fraud behavior prediction model based on five machine learning algorithms, including logistic regression, support vector machine, random forest, extreme gradient boosting tree and Stacking ensemble. The experimental results show that the modeling effect is significantly improved after the importance extraction of random forest features. In addition, the effect of integrated machine learning is better than that of single learning. Among them, Stacking ensemble learning performs the best, with precision, recall and precision scores above 0.99, followed by other models.

Keywords

Credit card fraud detection, Feature importance extraction, Logistic regression, Support vector machine, Extreme gradient boosting tree.

1. Introduction

With the rapid development of e-commerce, credit card payment has been rapidly popularized in China. The explosive growth and highly complex transaction data make it a huge challenge to ensure the stability and security of electronic payment services. According to statistics, the annual loss of China alone has exceeded 10 billion yuan, which has seriously hindered the long-term development of the financial industry and caused immeasurable losses to society. Ensuring the stability and security of electronic payment services has always been an important issue for banks and Internet companies, and in this process, the detection of fraudulent transactions is one of the most critical issues. Therefore, the research on credit card fraud detection is of great significance.

2. Literature Review

In today's digital age, credit card fraud has become a serious problem. With the rise of big data and the improvement of computer computing power in recent years, more and more machine

ISSN: 2664-9640

learning methods have been applied to credit card fraud detection. Research in this field by domestic and foreign scholars is very active.

Domestic scholar Yang Wensi [1] (2020) proposed a credit card fraudDomestic scholar Yang Wensi [1] (2020) proposed a credit card fraud detection method based on federated learning with privacy protection, which achieved an average test AUC of 95.5%, about 10% higher than traditional fraud detection systems. Huang Yongxin [2] (2020) constructed an advanced credit card fraud detection model using the idea of Gaussian mixture distribution and K-nearest neighbors sampling algorithm and deep forest algorithm. Liu Xia [3] (2022) proposed a mixed sampling method based on neighbor geometric space (K-G-Smote for short) and a credit card fraud detection model based on global anomaly detection (G-ADOA for short) to improve the accuracy of credit card fraud detection. Xie Shenghe [4] (2020) analyzed the advantages and disadvantages of logistic regression, decision tree, random forest and other algorithms in credit card fraud information detection.

Foreign scholar K Gayathri Krishna [5] (2023) proposed using big data technology and machine learning algorithms (such as logistic regression, decision tree, random forest, etc.) to predict fraudulent transactions in credit card operations in advance, thereby reducing further risks. Larson, Benjamin James [6] et al. (2020) explored solutions developed to help overcome the challenges of using supervised machine learning on imbalanced data. Padhi [7] (2022) et al. proposed a new feature selection (FS) method based on a meta-heuristic algorithm called Rock Hyrax Swarm Optimization Feature Selection (RHSOFS), and found that the performance of the proposed RHSOFS was excellent. Alamri, Maram [8] et al. (2022) discussed hybrid sampling techniques and their importance in solving data imbalance problems.

3. Data Preprocessing

3.1. Data Description

The experimental data used in this paper comes from the Kaggle website, which contains transaction records of cardholders on a certain day in September 2013 in Europe. There are a total of 284,808 transaction data, including 492 transactions marked as fraudulent. All data are numerical variables, with a total of 31 features. The specific feature information can be referred to in Table1.

Table 1. Credit Card Transaction Data Features

Feature Name	Feature Meaning	Data Type
Time	Transaction Time (sec)	Float
V1-V28	Other trading information	Float
Amount	Transaction amount	Float
Class	Fraudulent Transaction	Int

In the above variables, Time refers to the number of seconds from the transaction time to 0 of the current day; V1-V28 is obtained from the original characteristic information through PCA dimension reduction processing; Amount refers to the amount of the transaction. "Class" is the target variable, where 0 represents normal transactions, i.e., negative samples, and 1 represents fraudulent transactions, i.e., positive samples. Through the analysis of these characteristics, we can try to find the patterns or characteristics of fraudulent transactions, in order to better predict and identify potential fraud.

ISSN: 2664-9640

3.2. Data Analysis

3.2.1. Handling of missing values

The data set contains a large amount of data, so it is necessary to deal with the missing values and outliers of the data before building the model. First, the missingno library is called to quickly visualize the missing values, as shown in the figure.

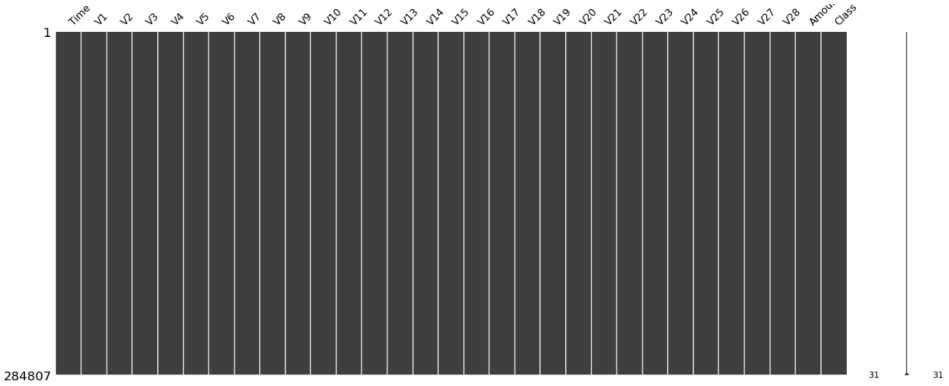

Figure 1. Missing Value Test

3.2.2. Handling of outliers

In a data set, an outlier is a special observation that is clearly different from other observations or deviates from the normal pattern. Outliers may appear as extremely high or low values, away from the concentration of other observations. They may lead to large deviations in the calculation of statistical indicators such as average and variance, and may also affect the accuracy and reliability of the model. Therefore, identifying and handling outliers is one of the important preprocessing steps before data analysis and modeling.

In view of the two situations of normal transaction and fraudulent transaction, this paper draws the data box charts of the two to show the differences between them, as shown in Figure 2.

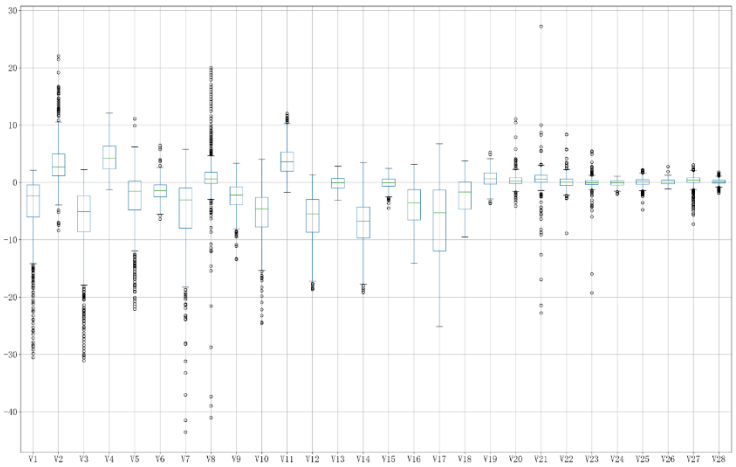

Figure 2. Box Plot of Fraud Transaction Data

ISSN: 2664-9640

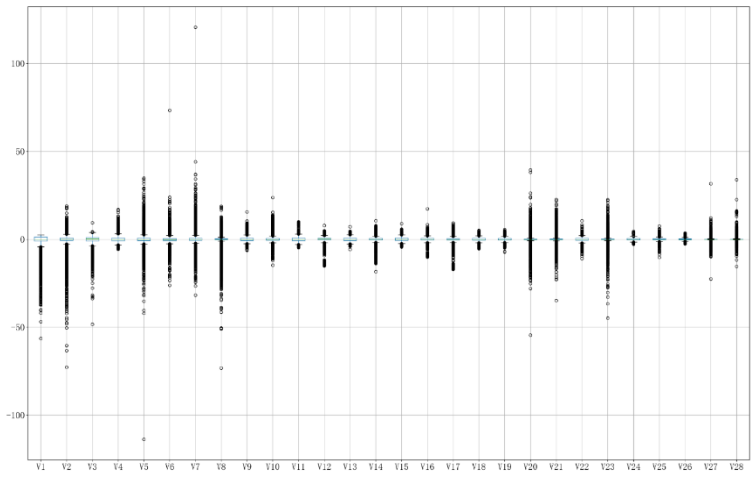

Figure 3. Normal Transaction Data Box Diagram

The boxplot above demonstrates the difference between the indicator data for normal and fraudulent transactions. It can be observed that the index data of fraudulent transactions show obvious left or right deviation, and the degree of data dispersion is high. The normal trading data is relatively uniform. Based on experience and the accuracy of bank transaction data, we can consider the existence of these outliers as a reasonable correction. Therefore, this paper does not deal with outliers.

3.3. Statistical analysis of data

The above outlier analysis simply analyzes the difference between the two data. In order to deeply explore the relationship and characteristics between the indicators, the next step is to visually analyze the two types of data.

3.3.1. Time-transaction frequency analysis

For the convenience of statistical analysis, this paper converts Time from seconds to hours, for example, 7200 seconds = 2 hours, which means that it occurs two hours after 0:00 in the morning. In order to explore the time distribution of different types of transactions, this paper draws the histogram of the quantity-time distribution of two types of transactions.

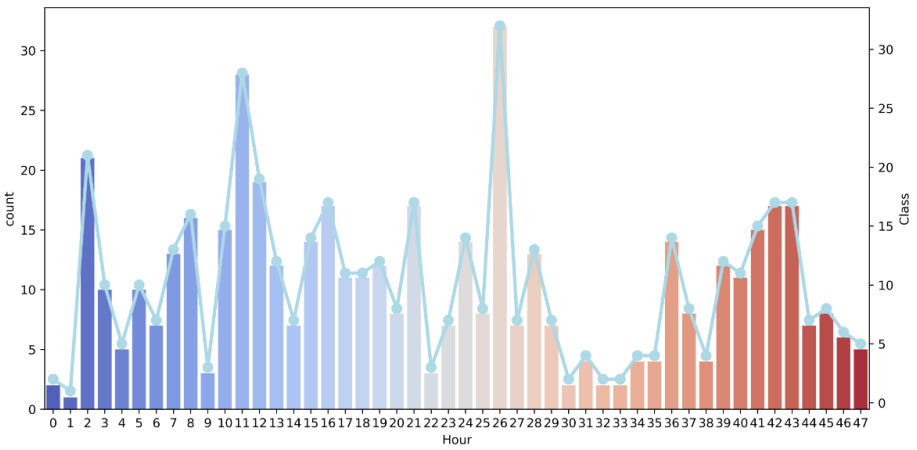

Figure 4. Time distribution of the number of fraudulent transactions

ISSN: 2664-9640

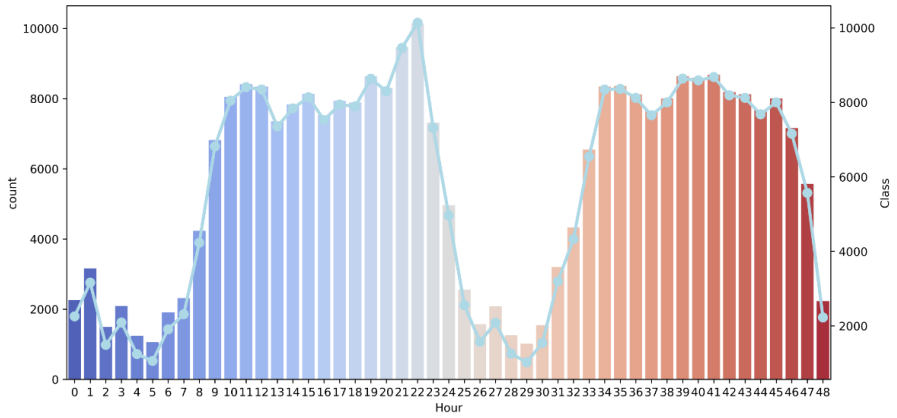

Figure 5. Time distribution of normal transaction times

From the time distribution of the number of transactions, we can find that the distribution of normal transactions is relatively uniform, which is related to the rhythm of people's life and work. In contrast, the distribution of fraudulent transactions is more concentrated, mainly around noon and early morning. This may be because fraudsters are more likely to consume at noon, while people's vigilance is reduced around the early morning, which is easy to be exploited by fraudsters.

3.3.2. Amount Distribution Analysis

Amount for different transaction types is also analyzed here. The quartile table of payment amounts for fraudulent and non-fraudulent transactions is shown in Table 2. You can get the difference from the table. Firstly, the average value of the payment amount of fraudulent transactions is 122.21, which is less than the average value of the payment amount of fraudulent transactions. This indicates that fraudulent transactions have a higher payment amount relative to non-fraudulent transactions in terms of transaction amount. Second, the mean square values of the amounts paid for fraudulent and non-fraudulent transactions are approximately equal. This also shows that the overall amount of fraudulent transactions and non-fraudulent transactions is not very different, which shows that many fraudulent transactions may be hidden in transactions with normal transaction amounts.

Table 2. Amount Quartile Table

Amount	count	mean	std	min	25%	50%	75%	max
1	492.00	122.21	256.68	0.00	1.00	9.25	105.89	2125.87
0	284315.00	88.29	250.11	0.00	5.56	22.00	77.05	25691.16

Figure 6 and Figure 7 show the amount distribution of different transaction types. It can be found that the amount of fraudulent transactions is mainly distributed between 0 and 100, while the amount of normal transactions is mostly distributed below 1000. This trend shows that fraudulent traders are generally more inclined to conduct fraudulent activities at lower amounts, while normal traders are more inclined to conduct normal transactions at higher amounts. This difference may reflect the lower risk appetite and risk tolerance of fraudulent traders.

ISSN: 2664-9640

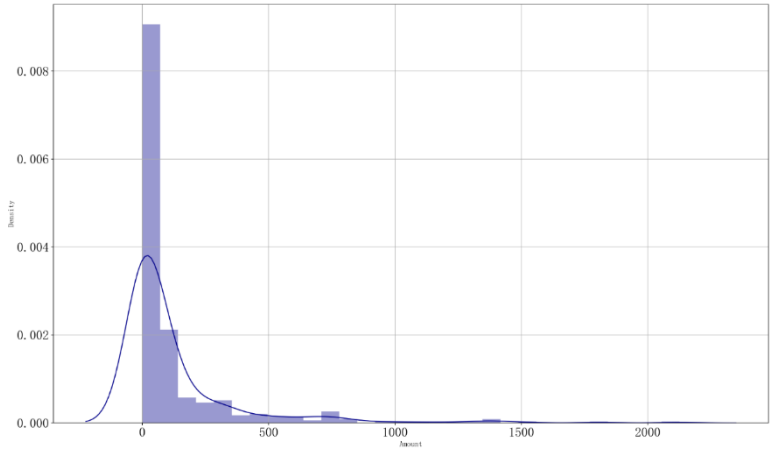

Figure 6. Fraudulent Transaction Amount Distribution

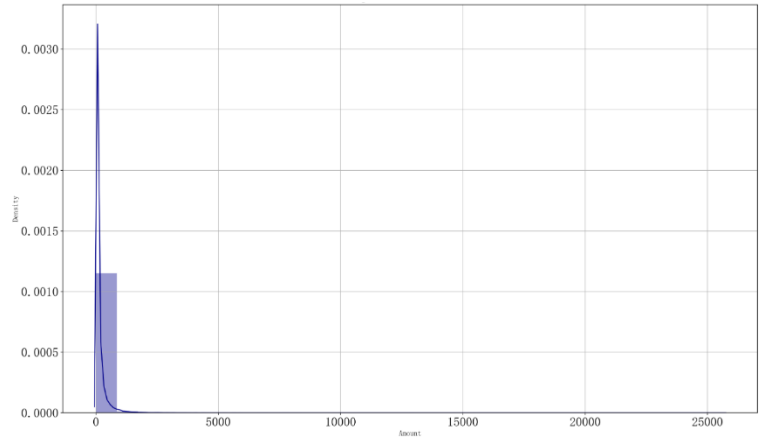

Figure 7. Distribution of normal transaction amount

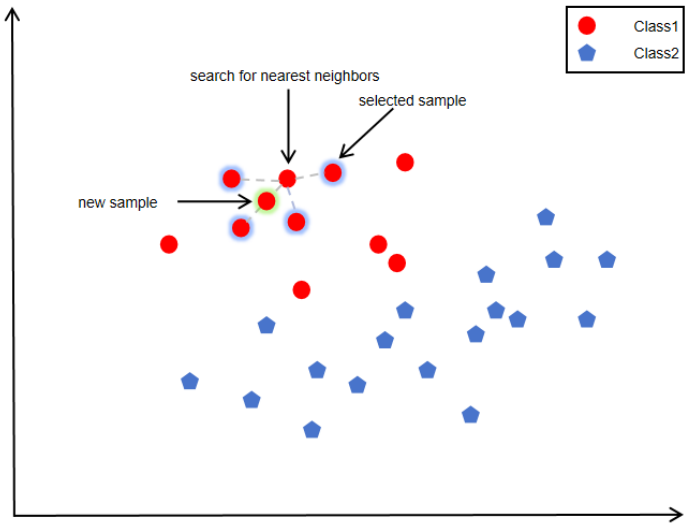

Figure 8. Smote Schematic

4. Smote Processing

In the credit card transaction data, the number of normal transactions is much larger than that of fraudulent transactions, which may lead to overfitting of the model during training, so Smote processing is needed for the data. Smote (Synthetic Minority Over) is a technique for solving class imbalance in classification problems by generating synthetic samples.

ISSN: 2664-9640

Figure 8 shows the flow of the Smote process. Firstly, a normal sample is randomly selected, and then a number of samples nearest to the sample are selected with the sample as the center. Then, a new composite sample is generated according to the characteristics of the neighbor samples. The above steps are repeated until the number of the normal samples and the number of the fraud samples are balanced, thereby obtaining a data set with balanced positive and negative samples.

In this paper, the Smote function of Python is called to balance the data, so that the number of positive and negative samples in the experimental data is the same.

5. Random Forest Feature Importance Extraction

Feature importance is an important part of machine learning models, which can help us understand which features in the data set have a greater contribution to the prediction performance of the model. In credit card fraud detection, feature importance can also help us identify which features are most critical to distinguish between fraudulent and normal transactions.

In the random forest model, the feature importance can be obtained by calculating the number of times each feature is used by the model and the average information gain provided in the decision tree. Specifically, the feature importance of a random forest can be calculated by the following steps:

A random forest model is trained and the decision rules for each decision tree are saved.

The number of times each feature is used and the average information gain provided in the decision tree is calculated for each decision tree.

The importance score of each feature is added up to get the total importance score of the feature.

Divide the importance score of each feature by the total importance score to get the relative importance score of the feature.

Figure 9 demonstrates the feature importance scores for the random forest model. It can be seen that the feature importance score of "V17" is the highest, which indicates that this feature is the most critical to distinguish fraudulent transactions from normal transactions. In addition, features such as "V14", "V10" and "V12" also have high importance scores, which indicates that they have a good contribution to the prediction performance of the model.

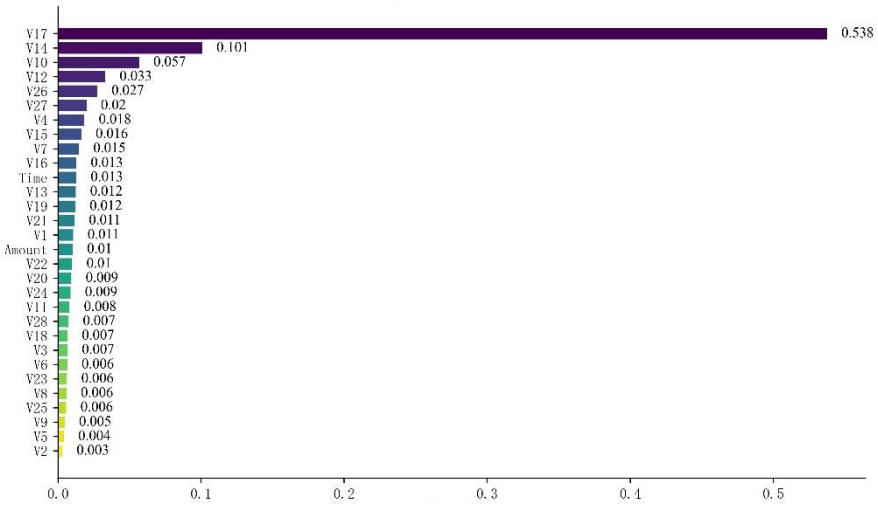

Figure 9. Random Forest Feature Importance Scores

By analyzing the importance of features, we can better understand how the model makes predictions and find the most critical features that distinguish fraudulent transactions from normal transactions. In order to optimize the performance of the model and improve the

ISSN: 2664-9640

prediction accuracy, this paper selects the feature variables whose feature importance scores are 0.01 and above for training.

6. Model Building

6.1. Logistic regression

Logistic regression is a statistical method used to build classification models. It does this by linearly combining the input features with the weights and then converting the result to a probability value using a logistic function, usually a sigmoid function. The key to the model is the use of a sigmoid function, which can map any real number to the sigmoid function, thus completing the conversion from real value to probability, that is, the classification task. The Sigmoid function is shown in the figure:

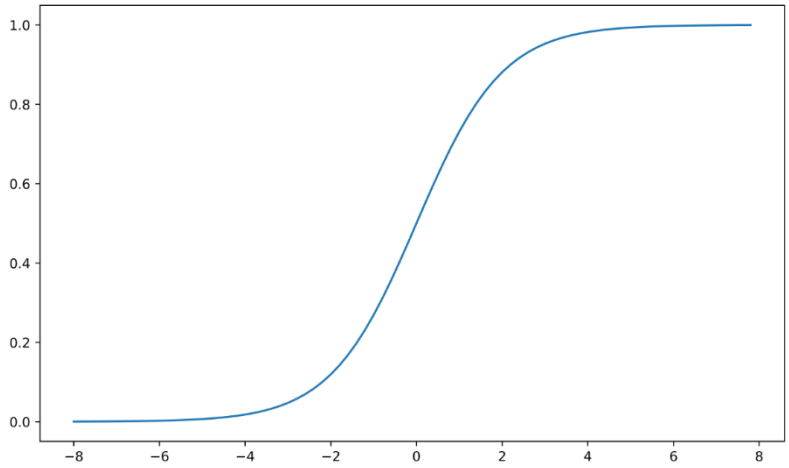

Figure 10. Sigmoid Function

On the image, as the input X increases, the output of the Sigmoid function gradually approaches 1, and as the input X decreases, the output gradually approaches 0. In the extreme case, when X approaches positive infinity, the output of Sigmoid function approaches 1; when X approaches negative infinity, the output approaches 0, and then the classification target is obtained.

In order to obtain a better logistic regression model, a grid search is performed for the regularization term strength C and the regularization term penalty, and cross-validation is used for each parameter in the grid on the training set to obtain the optimal parameter combination of the model as C = 0.76, penalty = 12.

After that, the test set and the training set are predicted to obtain the probability value of each sample output in the data set. Finally, the indexes of the model on the test set are obtained, as shown in the following table:

Table 3. Logistic Regression Performance Metrics

Accuracy_score	Roc_auc_score	Recall_score	Precision_score
0.940333333	0.939646525	0.902853260	0.973626373

Table 3 shows the scores of each indicator of the logistic regression model, in which the recall rate is relatively low and the other indicators are relatively excellent, indicating that the logistic regression model has a certain fraud detection capability.

6.2. Support vector machine

Support vector machine (SVM) is a commonly used machine learning algorithm, which classifies samples by constructing a hyperplane, and uses kernel function to deal with linear or nonlinear classification problems. The credit card transaction data studied in this paper is obviously a nonlinear classification problem. Figure 11 shows the result of constructing a

ISSN: 2664-9640

hyperplane for the binary classification problem. The hyperplane is used to divide the data of different categories, and finally the data is calculated to get the category it belongs to.

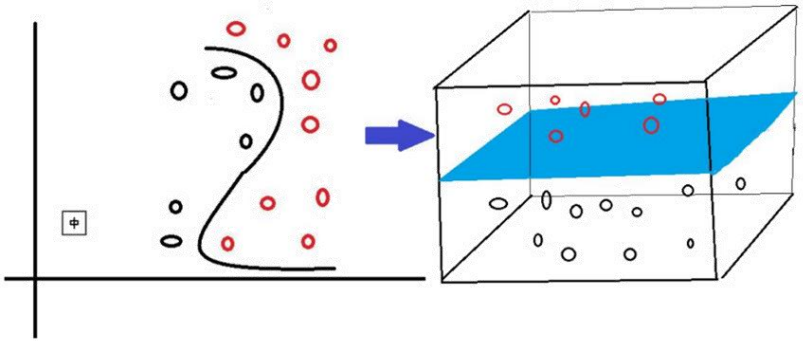

Figure 11. Principle of SVM model

When training the SVM model, we need to choose the appropriate kernel function and penalty parameter C to get a classifier with high accuracy and strong generalization ability. The kernel function can be linear kernel, polynomial kernel or Gaussian kernel, etc. The penalty parameter C controls the penalty degree of classification and the complexity of the model. This paper uses grid search to select the optimal parameter combination of SVM.

After training, the trained SVM model can be used to predict the test set and evaluate the performance of the model. Commonly used evaluation indicators include precision, recall, F1 score and AUC.

Table 4. Performance index of support vector machine

Accuracy_score	Roc_auc_score	Recall_score	Precision_score
0.957333333	0.956783163	0.927309782	0.984848484

Table 4 shows the scores of each index of the SVM model, which are distributed around the 0.95, and the model effect is relatively excellent, indicating that the model has a strong ability to predict fraud.

6.3. Random Forest Classification

Random forest classification is an ensemble learning algorithm, which constructs multiple decision trees and takes the average of their outputs to predict the classification, which consists of a plurality of decision trees, each of which is constructed on one random sampling of the input variables. The model principle is shown in the following figure:

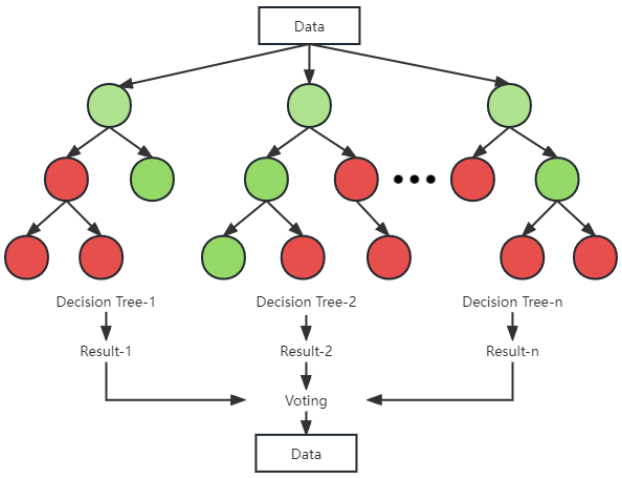

Figure 12. Random Forest Classification Model

ISSN: 2664-9640

In the training process in the figure, each decision tree will be trained according to the training data set and form a classifier. In the process of testing, each decision tree will classify and predict the test data, and take its mode as the final prediction result.

In order to obtain a better random forest classification model, the parameters Max _ depth and n _ estimators of the random forest were searched in a grid, and each parameter in the grid was cross-validated on the training set, and the optimal parameter combination of the model was obtained as Max _ depth = 5 and n _ estimators = 100.

After that, the test set and the training set are predicted to obtain the probability value of each sample output in the data set. Finally, the indexes of the model on the test set are obtained, as shown in the following table:

Table 5. Random Forest Performance Index

Accuracy_score	Roc_auc_score	Recall_score	Precision_score
0.965333333	0.965172616	0.954993306	0.984905385

Table 5 shows the scores of each index of the random forest classification model, and the scores of each index are all above 0.95, so the performance of the model is excellent and the prediction effect is very good.

6.4. Extreme Gradient Lifting Tree

Extreme Gradient Boosting (XGBoost) is an efficient gradient boosting decision tree algorithm, which is more advanced than the random forest classification algorithm. It can effectively deal with various types of data by using weighted residual terms to fit the data set. It also performs well in various machine learning tasks. The training process of the XGBoost algorithm consists of iterations in multiple stages, each of which fits a weighted residual term. In each stage, the algorithm calculates the gradient of each feature and updates the model parameters according to the gradient. In this way, XGBoost can gradually optimize the predictive performance of the model.

When building the XGBoost model, we need to choose some parameters, such as the learning rate, the maximum depth, and the subsample proportion. In this paper, the cross-validation method is used to select the optimal parameter combination. Here, the learning rate and the maximum depth parameter are searched in a grid, and cross-validation is used on the training set to obtain the optimal parameter combination of the model as LR _ rate = 0.1, Max _ depth = 5.

Table 6. Performance index of extreme gradient lifting tree

Accuracy_score	Roc_auc_score	Recall_score	Precision_score
0.986000000	0.985983775	0.981927710	0.989878542

Table 6 shows the scores of each index of the extreme gradient lifting tree model, and the scores of each index are above 0.98, which is obviously better than the performance of the single machine learning model.

6.5. Stacking integration

Stacking is an ensemble learning algorithm, which improves the prediction performance by combining the prediction results of multiple different models. In credit card fraud detection, Stacking can be used to optimize the prediction performance of the model and improve the generalization ability of the model.

The basic idea of Stacking algorithm is to combine the prediction results of multiple different models to get a more accurate prediction result. In the training process, Stacking algorithm needs to train several different models and use these models to predict the training data set. Then, the prediction result of each model is compared with the original label, and the prediction

ISSN: 2664-9640

error and weight of each model are calculated. Finally, these weights are used to combine the prediction results of each model to get the final prediction result.

In credit card fraud detection, this paper uses the previous model for Stacking integration. In the training process, these models will predict the training data set respectively, and get their own prediction results and weights. These weights are then used to combine the prediction results of each model to obtain the final prediction result. The results show that the effect is significantly improved, and the scores of each index of the Stacking integrated model are shown in Table 7:

Table 7. Stacking Performance Metrics

accuracy_score	roc_auc_score	recall_score	precision_score
0.998	0.997999967	0.997991967	0.997991967

Table 7 shows the scores of each index of the integrated model. The scores of each index are all above the 0.99, and the performance is extremely excellent. The prediction effect is significantly higher than that of other models, and the prediction effect is very excellent.

To illustrate the training process of the model, the model ROC curve is plotted here. In Figure 12, the ROC curve for the Stacking ensemble model is presented. The curve shows how the model performs on the training and test sets. As can be seen from the figure, the Stacking ensemble model performs very well on both the training set and the test set, with the ROC curve very close to the upper left corner and the AUC value very high, indicating that the model is able to distinguish between fraudulent and normal transactions very well.

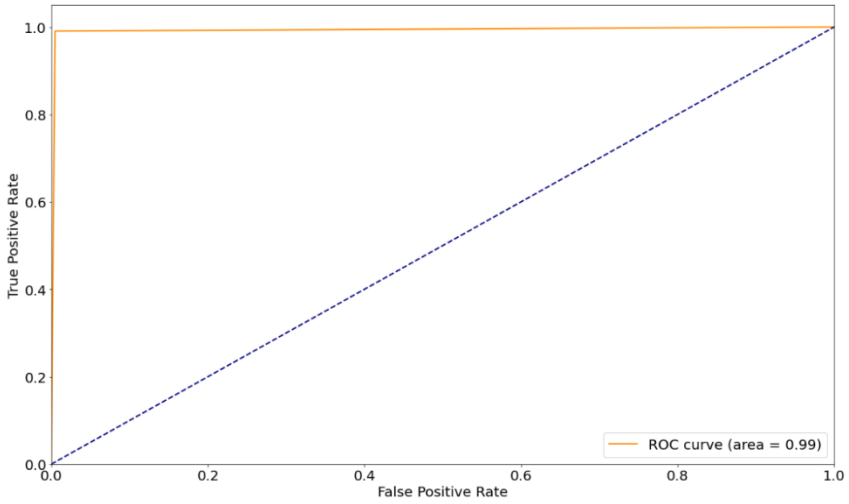

Figure 13. Stacking Integration ROC Curve

7. Conclusion

Through experiments with different machine learning models, using grid search method to optimize different model parameters, and using Stacking algorithm to integrate, this paper obtains a credit card fraud detection model with excellent performance and good prediction effect. The model has excellent prediction performance and generalization ability in credit card fraud detection, can effectively identify fraudulent transactions, and can provide a more accurate and reliable fraud detection tool for financial institutions. At the same time, the model proposed in this paper can also be applied to other similar problems, which provides a useful reference for the application of machine learning in practical problems.

ISSN: 2664-9640

Acknowledgments

This work is supported by Anhui University of Finance and Economics Undergraduate Research and Innovation Fund Project (No.:XSKY23161).

References

[1] Vince Yang. Research on Credit Card Fraud Detection System Based on Federated Learning [D]. University of Chinese Academy of Sciences (Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences), 2020.
[2] Huang Yongxin. Research on Credit Card Fraud Detection Based on Deep Forest [D]. Jinan University, 2020.
[3] Liu Wei. Research on Credit Card Fraud Detection Model Based on Global Anomaly Detection [D]. Guangxi University, 2022.
[4] Xie Shenghe. Comparison of Machine Learning Methods for Credit Card Fraud Detection [D]. Central China Normal University, 2020.
[5] K, G. K., Kulkarni, P., & Natraj, N. (2023). Use of big data technologies for credit card fraud prediction. Piscataway: The Institute of Electrical and Electronics Engineers, Inc. (IEEE).
[6] Larson, B. J. (2020). False positive reduction in credit card fraud prediction: An evaluation of machine learning methodology on imbalanced data (Order No. 28716473).
[7] Padhi, B. K., Chakravarty, S., Naik, B., Pattanayak, R. M., & Das, H. (2022). RHSOFS: Feature selection using the rock hyrax swarm optimization algorithm for credit card fraud detection system. Sensors, 22(23), 9321.
[8] Alamri, M., & Ykhlef, M. (2022). Survey of credit card anomaly and fraud detection using sampling techniques. Electronics, 11(23), 4003.