Enhancing Credit Card Fraud Detection on Imbalanced Datasets

Chao Liu*

College of Science & Engineering, University of Minnesota-Twin Cities, Minneapolis, US

*Corresponding author: liux2693@umn.edu

Abstract. Abstract. The use of credit cards is becoming more and more popular in today's society, especially now with the prevalence of electronic payments. Electronic payments are available through bank apps or payment processors like PayPal, Alipay, etc. Without using cash, people are taking advantage of the convenience of them both online and offline. In addition to bringing people the benefits of efficiency and convenience, credit card fraud has also emerged and caused a great deal of economic losses for cardholders, as well as causing great trouble for banks. The primary goal of this work is to identify fraudulent transactions in an unbalanced dataset. The dataset comprises credit card transactions from just two days in 2013 in Europe. In this study, the original data, the original data with Stratified-KFolds, and the undersampled data will be compared. It is found that undersampling, although it reduces the accuracy by a small amount, can greatly improve the detection of fraudulent transactions. Meanwhile, this study uses different models, one is Logistic Regression, and the others are all Tree-Based method. The study analyzes their confusion matrices, ROC curves, and Precession-Recall curves. The results show that in the undersampling rate dataset, the recall, precision, F1 score and accuracy of Xgboost are optimized to 93.2%, 97%, 95%, and 95% respectively, and the AUC of both the ROC curve and the Presicion-Recall curve are optimized to 99%, so this study concludes that XGboost is the best performer. With excellent algorithms, we can better avoid the leakage of information and loss of money in real life.

Keywords: Imbalanced Data; Credit Card Fraud; Tree-Based Method; XGBoost; undersampling.

1. Introduction

With the popularization of the Internet, e-commerce and online payment have become indispensable in today's human life. Credit card payments and digital payments are fully integrated to develop the currently popular digital credit payment model [1]. However, at the same time, the risk of credit card fraud is increasing. Credit card fraud leads to huge financial losses and personal information leakage. Physical detections are not only expensive and time-consuming, but they also don't produce reliable results [2]. To combat this problem, people are improving the success rate of detecting credit card fraud by combining existing data and technology with advanced artificial intelligence techniques.

In previous studies, to deal with imbalanced data in credit card transaction, pre-application of undersampling to the data usually gives better results than the raw data before modeling. Synthetic minority oversampling technique (SMOTE) technique can effectively augment minority categories and balance the training set [3]. Some studies use an integrated learning framework based on training set partitioning and clustering. This framework integrates the ideas of partitioning and clustering, which not only solves the high imbalance of the credit card dataset, but also ensures the diversity of the training set samples [4]. In the past a lot of studies have also tried to apply different models to deal with credit card fraud. Some researchers have utilized more complex methods to boost detection success rates, such as neural networks and deep learning. To enhance fraud detection performance, three convolutional neural network-based designs are used [5]. Another method that is widely used in credit card fraud is Random Forest. Based on the confusion matrix, some study assesses the Random Forest algorithm's performance and processes the dataset with an accuracy of roughly 90% [6]. To categorize, forecast, and identify fraudulent credit card transactions, some research compares SVM, KNN, random forests, and logistic regression. It has found that random forests outperform other models in predicting the accuracy of fraudulent credit card transactions [7]. However, Random

Forest is commonly learned completely using the Bootstrap approach, although they have limitations in imbalanced classification situations. The Bootstrap technique is rarely, if ever, considered for minority classes. This results in decision trees that are less effective at predicting minority classes [8]. Some study compares KNN, Logistic Regression, Naive Bayes algorithms and came up with the result that logistic regression's predictive model works better [9]. A lot of studies focus on the result of accuracy, but a blind guess will give us a really high accuracy score (by betting on majority class). High accuracy is trivial on an imbalanced classification problem. This means when dealing with an imbalanced dataset, if a model gives a high accuracy does not mean it is a better one. Thus, this study selects several machine learning algorithms, aiming to improve the credit card fraud predictions.

2. Data

A dataset from Kaggle is used in this research. The dataset consists of two days' worth of credit card transactions made by European cardholders in September 2013.

In this section, exploratory data analysis was conducted to understand the data distribution and identify potential outliers. Verifying the distribution of the fraudulent attributive, which is depicted as an imbalance in both the bar chart and the histogram. Out of 284,807 transactions, 492 were fraudulent, with fraud transactions representing only 0.172% of all transactions (See Fig. 1).

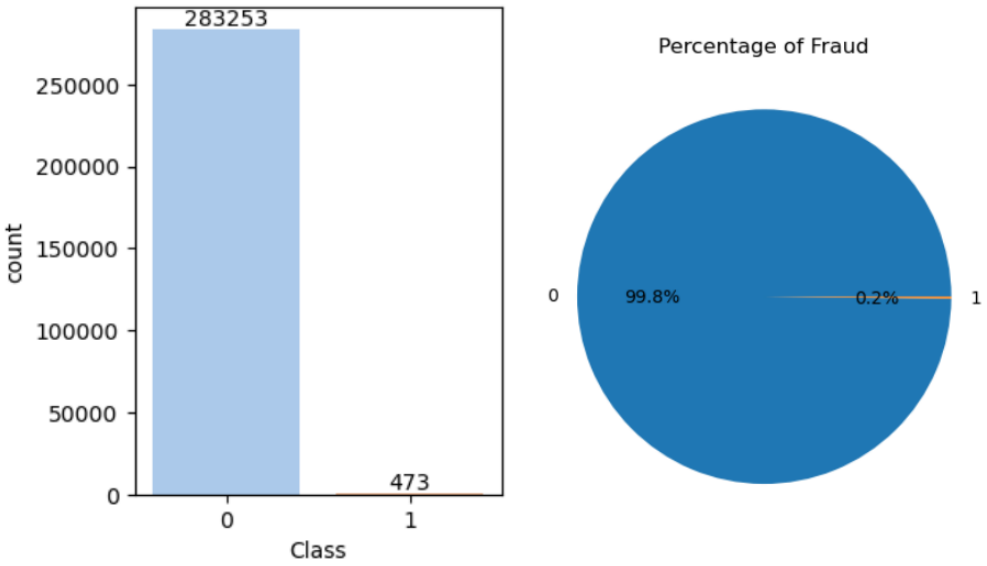


Fig. 1 Bar Chart and Pie Chart

The dataset has a total of 30 feature values. The vast majority of these were obtained through PCA. Because of confidentiality reasons, kaggle has no way to provide specific information other than the numbering (v1-v28). There are two pieces of data for which features, and information are preserved: the "Time" and the "Amount". The "Time" feature refers to the time in seconds between the current transaction and the first transaction in the dataset. The "Amount" feature is literally the amount of the transaction. Finally, "Class" is the result of whether the transaction is fraudulent or not. A "1" means that the transaction is fraudulent and a "0" means that it is a normal transaction.

Simultaneously, Boxplots were plotted for each variable to gain insights into the data. It is obvious that there are lots of outliers (See Fig. 2).

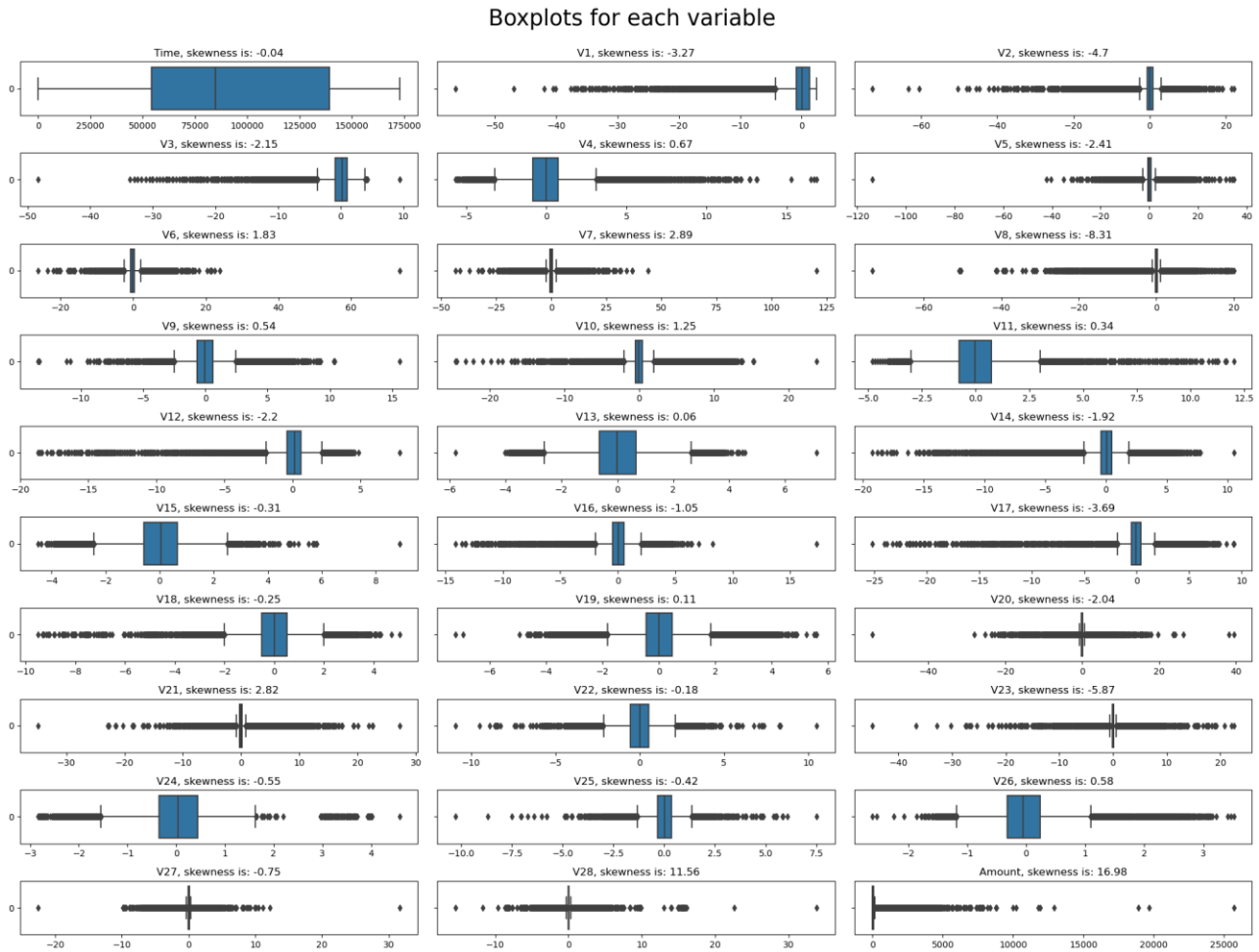


Fig. 2 Boxplot for different features

Normally dropping outliers would be a better choice. However, in the case of credit card fraud, dropping outliers would lose most of the fraudulent transaction data. Fig. 3 shows the result after dropping outliers by applying 1.5 times IQR. Only 26 fraud transactions remain, which means a great loss of important information. Therefore, the research for this article will not drop outliers.

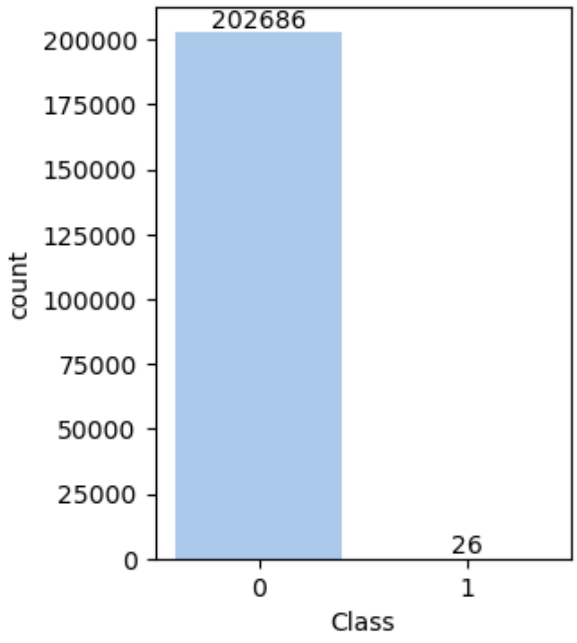


Fig. 3 Transactions after 1.5 times IQR

3. Methology

3.1. Decision Tree

There are a lot of tree-based model. Decision Tree is the most basic of these methods which is used to make classification or regression predictions from input data. It consists of a series of nodes which include root, internal and leaf nodes. Each node represents a feature or attribute and the data is divided based on these attributes. The idea of a decision tree is very plain, and the structure of its branches is an if-else structure. Each internal node is a judgment on an attribute. Each branch downwards represents the output of a judgment. The leaf nodes then represent the final classification result. Fig. 4 shows how a decision tree works.

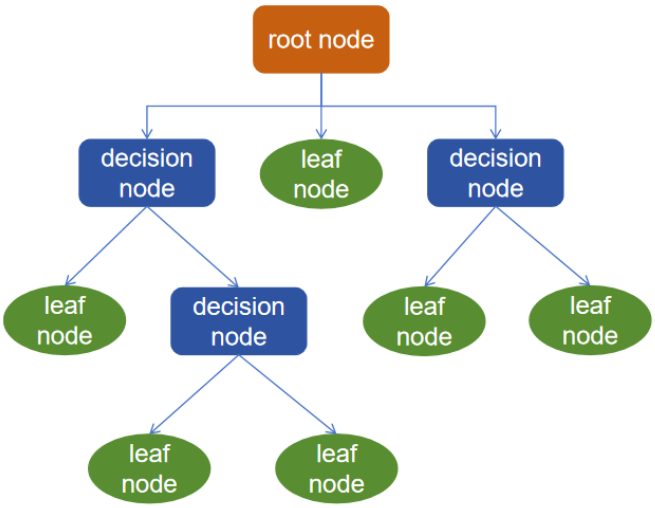


Fig. 4 Decision Tree

There are various methods to affect the classification of decision tree. ID3 method classifies by calculating information gain. C4.5 classifies by information gain rate. This method optimizes the problem of always selecting more values in ID3 method. CART method classifies by calculating Gini coefficients.

3.2. Logistic Regression

Logistic regression is a linear classifier. It is mainly used for binary classification problems [10]. It is computationally inexpensive and easy to understand and implement. There is no need to assume the data distribution in advance, which avoids the problems associated with inaccurate assumption of the distribution. Not only do we predict the categories, but we also get approximate probabilistic predictions. The objective function is derivable of any order. It is prone to underfitting and has poor classification accuracy.

3.3. Random Forest

Random forests can perform both regression and classification in machine learning. Bagging - a general-purpose strategy for lowering the variance of a statistical learning method is the basis of random forest [11]. It can be said that random forest is the combination of bagging and decision tree. Random forest contains many decision trees. It randomly selects an amount of data in the original dataset, after which it randomly selects multiple features and trains them, which constitutes a decision tree. Repeat the above steps to get multiple decision trees. A result is generated on each decision tree and the final result follows the principle of majority rule. Fig. 5 tells how a random forest get a final result.

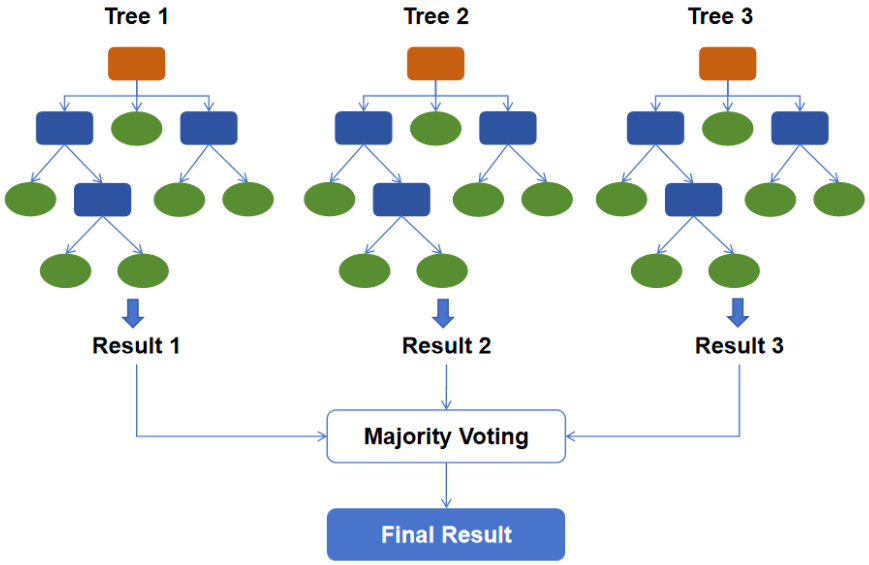


Fig. 5 Random Forest

3.4. XGBoost

XGBoost is an algorithm based on gradient boosting. It builds decision trees sequentially, minimizing residual errors at each step. It handles missing data and features importance analysis efficiently. The gradient boosting algorithm is an iterative integration method that progressively improves the performance of the model by training the weak learner multiple times and adjusting the weights of the training data based on the previous prediction. XGBoost introduces L1 and L2 regularization terms to limit the complexity of the model and reduce the risk of overfitting. Fig. 6 is a general architecture of XGBoost.

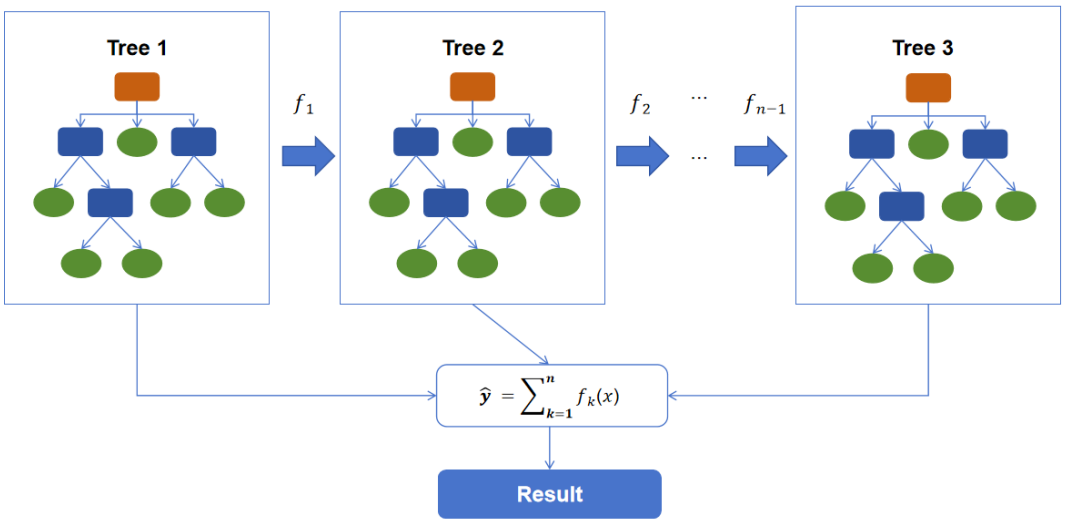


Fig. 6 XGBoost

4. Results

This article employs a variety of criteria to assess the trained model. The criteria include recall, precision, F1-score and accuracy. The criteria aided in determining the models' performance in detecting fraudulent credit card transactions. Additionally, this study analyzes the performance of several models by generating confusion matrix and plotting ROC curves and Precision-Recall curves.

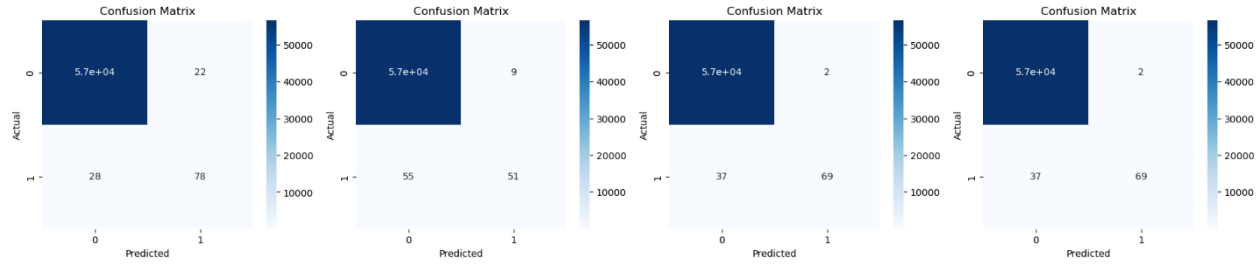
The evaluation of four model classifications was performed for three different datasets: (1) raw data without undersampling, using five-fold cross-validation; (2) raw data without undersampling, using StratifiedKFold; and (3) balanced data after undersampling. The outcomes can be seen in the tables and figures below.

4.1. Raw Data with No Undersampling

This paper construct certain models with raw data and no undersampling, the results are shown in the following Table 1 and Fig. 7.

Table 1. Cross-validation Results for Models on Raw Data with no undersampling

Model Name	Recall	Precision	F1 Score	Accuracy
Decision Tree	0.735849	0.780000	0.757282	0.999119
Logistic Regression	0.556604	0.842857	0.670455	0.998978
Random Forest	0.820755	0.966667	0.887755	0.999612
XGBoost	0.820755	0.966667	0.887755	0.999612



(1)Decision Tree (2)Logistic Regression (3)Random Forest (4)XGBoost

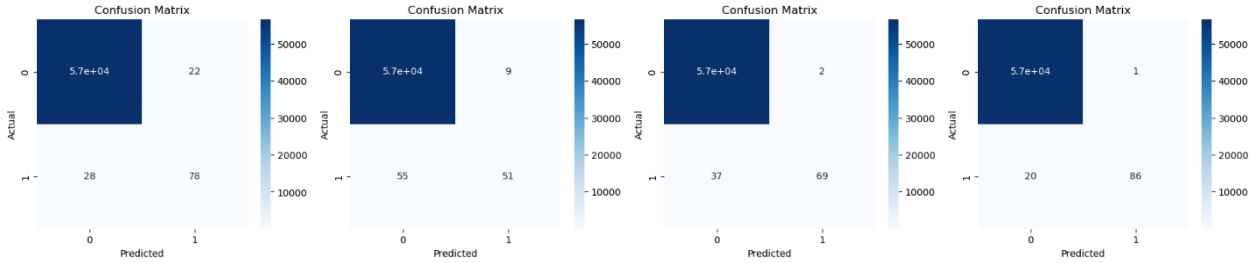
Fig. 7 Confusion Matrix for models on raw data (without undersampling)

4.2. Models on raw data with StratifiedKFold

This paper construct certain models with raw data and no undersampling, the results are shown in the following Table 2 and Fig. 8.

Table 2. Cross-validation Results for Models on Raw Data with StratifiedKFold

Model Name	Recall	Precision	F1 Score	Accuracy
Decision Tree	0.735849	0.780000	0.757282	0.999119
Logistic Regression	0.556604	0.842857	0.670455	0.998978
Random Forest	0.820755	0.966667	0.887755	0.999612
XGBoost	0.820755	0.977528	0.892308	0.999630



(1)Decision Tree (2)Logistic Regression (3)Random Forest (4)XGBoost

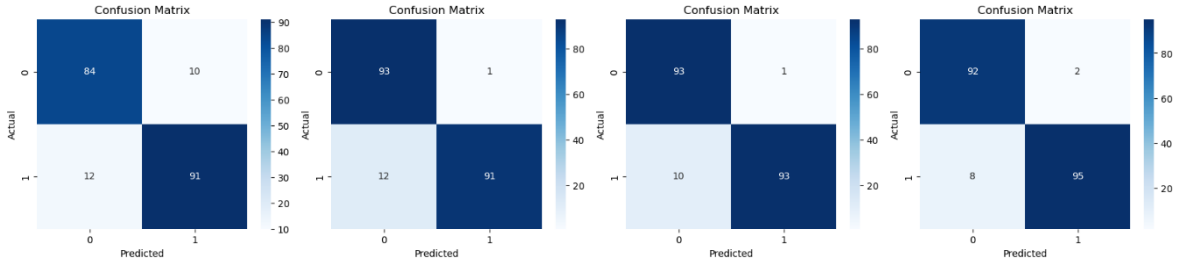
Fig. 8 Confusion Matrix for models on raw data with StratifiedKFold

4.3. Models on Balanced Data after Undersampling

This paper construct certain models with raw data and no undersampling, the results are shown in the following Table 3 and Fig. 9.

Table 3. Balanced Data after Undersampling

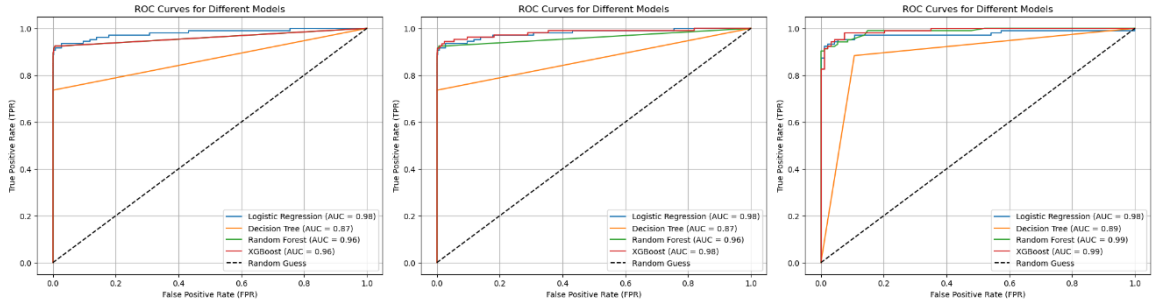
Model Name	Recall	Precision	F1 Score	Accuracy
Decision Tree	0.883495	0.900990	0.892157	0.888325
Logistic Regression	0.932039	0.969697	0.950495	0.949239
Random Forest	0.912621	0.979167	0.944724	0.944162
XGBoost	0.932039	0.969697	0.950495	0.949239



(1)Decision Tree (2)Logistic Regression (3)Random Forest (4)XGBoost

Fig. 9 Confusion Matrix for models on raw data without Undersampling

Furthermore, this paper plots the ROC curve and Precision-Recall curves in Fig. 10.



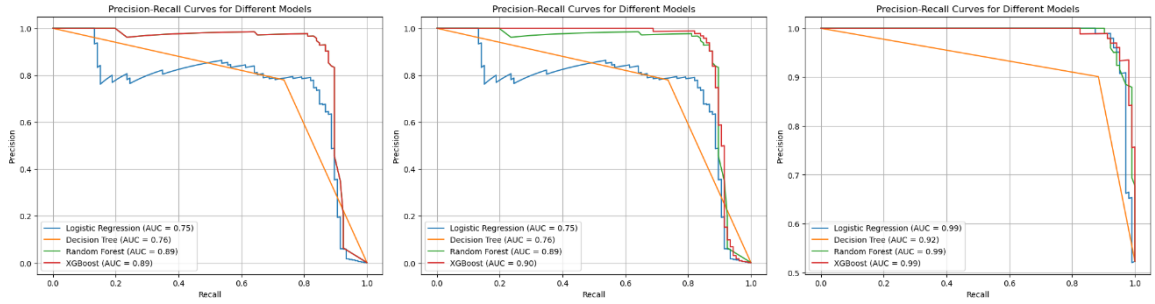
(a)Raw Data (b)Stratified KFold (c)Undersampling

Fig. 10 ROC Curve

ROC is the abbreviation of Receiver Operating Characteristic Curve. The horizontal coordinate is the false positive rate (FPR). It shows the percentage of real negative cases to all negative cases predicted by the classifier in the positive class. The vertical coordinate is the true positive rate (TPR). It shows the percentage of real positive cases to all positive cases predicted by the classifier. AUC is the abbreviation of Area Under Curve. The general efficacy of a model can be evaluated using AUC. In an ROC curve, a bigger value of AUC means a better choice. This assessment method is also applied to the current study to choose a better model for detecting fraud transactions.

For the experimental results, the value of AUC varies under different data processing methods. Specifically, for the original data and the model using the stratifiedKFold method, the AUC values are ≈0.89 and ≈0.95, respectively. On the Undersampled data, the AUC value is closer to 1, reaching ≈0.98. This shows that the model performs best in identifying fraudulent users on the Undersampled data. Another crucial tool for assessing the effectiveness of binary classification algorithms is the Precision-Recall Curve. It focuses primarily on how well the model works with fraudulent users. Precision indicates how many of the samples predicted to be positive are truly positive sample. Recall (recall rate) is the same as TPR mentioned above.

In Precision-Recall Curve, the horizontal coordinate is the Precision, and the vertical coordinate is the Recall (TPR). This curve shows the trade-off relationship between the two different coordinate value. Generally speaking, the Precision is increasing, and the Recall is decreasing when the model's prediction result is more skewed towards positive instances; on the other hand, the Precision is decreasing, and the Recall is increasing when the model's prediction result is more biased towards negative cases (See Fig. 11).



(a)Raw Data (b)Stratified KFold (c)Undersampling

Fig. 11 Precision-Recall Curve

In comparison, the results of the Undersampling model are both higher in Precision and Recall, which can achieve a higher recall rate while maintaining a higher accuracy rate. And XGBoost always shows the best outcome not only in accuracy, precision, recall, f1 score, but also in the AUC of ROC and Precision-Recall Curves. Therefore, in this study, the best results were achieved using XGBoost model after using Undersampling processed data.

5. Conclusion

This study aims to enhance the accuracy of Credit Card Fraud transactions detection by utilizing Decision Tree, Logistic Regression, Random Forests, XGBoost as machine learning algorithms. The results showed significant improvements when applying undersampling to data and using XGBoost algorithm, providing more reliable credit card transaction security in the financial industry.

There are limits for this study. First of all, the use of undersampling results in a dataset that is less than 1% of the size of the original data. Secondly, the more popular and newer neural network models were not utilized due to limited personal time and ability. Future research can explore other ensemble learning methods or deep learning models for addressing imbalanced data problems. Thirdly, it is possible to use neighborhood-based algorithms based on the SMOTE, the most widely used paradigm of imbalanced data resampling. Finally, GridSearchCV, a hyperparameter optimization approach, can be used to improve models.

References

[1] Ghai, V., & Kang, S. S. Role of Machine Learning in Credit Card Fraud Detection. In 2021 3rd International Conference on Advances in Computing, Communication Control and Networking, 2021: 939-943.

[2] Bodepudi, H. Credit card fraud detection using unsupervised machine learning algorithms. Int J Comput Trends Technol, 2021, 69: 1-13.

[3] Dablain, D., Krawczyk, B., & Chawla, N. V. DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[4] Wang, H. Y. Research on Credit Card Fraud Detection Scheme Based on Machine Learning. Thesis for master's degree. Beijing University of Posts and Telecommunications, 2019.

[5] Alarfaj, F. K., Malik, I., Khan, H. U., Almusallam, N., Ramzan, M., & Ahmed, M. Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. IEEE Access, 2022, 10: 39700-39715.

[6] Kumar, M. S., Soundarya, V., Kavitha, S., Keerthika, E. S., & Aswini, E. Credit card fraud detection using random forest algorithm. In 2019 3rd International Conference on Computing and Communications Technologies, 2019: 149-153.

[7] Liu, J., Zhang, Z., Li, X., et al. Machine learning assisted phase and size-controlled synthesis of iron oxide particles. Chemical Engineering Journal, 2023, 473: 145216.

[8] Luo, Z. Q. Improvement and Application of Random Forest Algorithm in Credit Card Fraud Detection. Thesis for master's degree, Beijing Forestry University, 2020.

[9] Itoo, F., Meenakshi, & Singh, S. Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology, 2021, 13: 1503-1511.

[10] LaValley, M. P. Logistic regression. Circulation, 2008, 117(18): 2395-2399.

[11] James, G. G. M. An introduction to statistical learning: with applications in Python.2023.