An Imbalanced Financial Fraud Data Model Based on Improved XGBoost and RUS Boost Fusion Algorithm with Pairwise

Junhao Xian

School of Financial Mathematics & Statistics, Guangdong University of Finance, Guangzhou, China.

201616128@m.gduf.edu.cn

Abstract: As the socio-economic landscape evolves, the investigation into anti-fraud behaviors in shopping gains increasing significance. Although prior studies have utilized machine learning to tackle this issue, they often grapple with two key obstacles. First, an imbalance between positive and negative data samples exists. Second, the presence of redundant features leads to suboptimal model performance. In order to surmount these challenges, we've developed a new machine learning framework. This innovative solution automatically selects features and balances the data set's positive and negative samples. Our framework's outstanding performance on the IEEE-CIS Fraud Detection dataset thoroughly validates the efficacy of our approach.

Keywords: XGBoost, Pairwise, RUSBoost, Ensemble.

1. Introduction

In the ongoing process of socio-economic development, research on shopping anti-fraud behavior has become increasingly crucial. In the current global economic environment, shopping anti-fraud not only pertains to corporate profits and reputation but also directly affects consumers' shopping experience and trust. In recent years, despite some achievements in this field [1][2][3], such as attempting to apply machine learning methods to model [4][5][6] and solve fraud issues, we still face some significant challenges.

The first major challenge we need to confront is the imbalance of positive and negative samples in the data [7] . Generally, the proportion of fraudulent transactions is much lower than that of normal transactions, which presents us with a severe imbalance problem when building models. If we directly feed all the data to the model for learning, the model may lean towards predicting the majority scenario, i.e., normal transactions, and overlook the minority of fraudulent transactions. This leads to a severe decline in the model's predictive power, especially for fraudulent transactions. Faced with such imbalanced data, the research community has proposed many solutions. For example, Chawla, N. V. [8] , introduced Synthetic Minority Over-Sampling Technique (SMOTE). However, the SMOTE algorithm may introduce additional noise, thereby reducing the model's generalization ability. He, H.'s [9] Adaptive Synthetic method (ADASYN) faces similar problems and may lead to a decline in model performance.

Secondly, the issue of feature redundancy is also a key factor impacting model performance. In the era of big data, we can obtain a large number of feature data from various sources. However, not all features are helpful in predicting fraudulent behavior, and some features may even interfere with the model's learning, leading to a decline in model performance. Therefore, selecting features that are helpful for predicting fraudulent behavior from a large number of features is another problem we need to solve. Current solutions mainly involve manual selection and PCA selection. Manual selection is often tedious and inefficient. Simultaneously, PCA selection overlooks the issue of class labels.

To overcome the aforementioned problems, we designed a novel machine learning framework. This framework can automatically perform feature selection to reduce the impact of feature redundancy, and it can also balance the positive and negative samples in the dataset to avoid the problem of sample imbalance. We have validated our method through experiments on the IEEE-CIS Fraud Detection dataset[10], and the results show that our method has excellent performance in shopping anti-fraud problems, demonstrating the effectiveness of our method.

In summary, our contributions are as follows:

1. We designed an innovative machine learning framework for automatic feature selection and sample imbalance resolution. Our framework combines automatic feature selection and sample imbalance solutions, forming a comprehensive solution. Through automatic feature selection, we can select features with Gini index importance, thus reducing the complexity of the feature space. Through the sample imbalance solution, we can balance the distribution of sample categories and enhance the model's ability to recognize fraudulent behavior.

2. To address the problem of imbalance between positive and negative samples, we proposed a new type of loss function. The current approach is not sensitive to the problem of sample imbalance and can lead to a weak recognition ability for the minority category (fraudulent category). To overcome this issue, we introduced a new loss function that gives higher weights to the fraudulent category during the training process to reinforce the model's focus on fraudulent behavior. This novel loss function, based on dynamic adjustments to sample weights, adaptively adjusts weights according to the distribution of sample categories, enabling the model to pay more attention to the samples of the fraudulent category, thereby improving the model's detection accuracy of fraudulent behavior.

3. Our framework achieved good performance on the IEEE-CIS Fraud Detection dataset, proving its effectiveness. Our framework predicted an accuracy of 98.5% and an AUC value of 0.904 on the IEEE-CIS Fraud Detection dataset.

2. Literature Review

2.1 Financial Fraud Detection

The field of financial fraud detection is broad and diverse, encompassing areas such as mobile payments, credit card abuse, and ERP systems, among others. Notably, fraudulent activities in these areas typically account for only a small portion of transactions. Therefore, many financial fraud detection methods are based on anomaly detection. In 2006, Chen, R.C. [11] et al. proposed an effective fraud detection system for unevenly distributed few-sample data, leveraging Support Vector Machine (SVM) and genetic algorithm. In 2018, Yee, O.S.[12] et al. investigated the performance of a variety of Bayesian classifiers on credit card fraud detection tasks. All classifiers demonstrated respectable detection accuracy on datasets processed by Principal Components Analysis (PCA). While these machine learning methods can address the issue of financial fraud to some degree, considering the complexity of credit card transaction behaviors in the real world, how to promptly and accurately extract representative features from limited transaction data remains a topic warranting further exploration.

2.2 Ensemble Tree Methods

In the field of financial fraud detection, XGBoost [13] has proven to be very effective. For instance, F. Wan [14] proposed a supply chain fraud detection model based on XGBoost and random forest hybrid to detect fraud in DataCo intelligent supply chain dataset. C. Meng [15] evaluated the performance of the XGBoost algorithm on the original credit fraud dataset, undersampling, and SMOTE datasets separately; D. Trisanto [16] introduced an enhanced focal loss technique (W-CEL loss unbalanced parameter) for unbalanced XGBoost. This research aimed to refine the focus weight loss, indicating that the method could only yield effective and unbiased machine learning models with mildly unbalanced data.

In financial fraud detection, data imbalance is a prevalent issue as fraudulent transactions are relatively fewer than normal ones. The RUSBoost [17] algorithm addresses this by randomly undersampling the majority class samples, thereby improving the detection rate of fraudulent transactions. B. Liao [18] evaluated the performance of NABAS with the RUSBoost method and compared it with the logistic regression model and the Support Vector Machine (SVM-FK) algorithm with financial kernel; R. Akram [19] proposed a Convolutional Neural Network (CNN) and methods that include RUSBoost Manta Rapaging Optimization (rus-MRFO) and RUSBoost Bird Swarm

Algorithm (rus-BSA) models for analyzing the problem of power theft; S. Mujeeb [20] proposed using Differential Evaluation (DE) to optimize the RUSBoost algorithm classifier heuristically, which was evaluated on the dataset of the State Grid Corporation of China (SGCC).

Despite progress in handling unbalanced data, challenges remain in accurately detecting minority fraudulent transactions and managing data noise. In response, we introduce a novel model fusion strategy combining Pairwise-optimized XGBoost and RUSBoost. This aims to boost performance when dealing with unbalanced financial fraud data.

2.3 Overview of the Algorithmic Process

The algorithm flow is shown in the figure below.

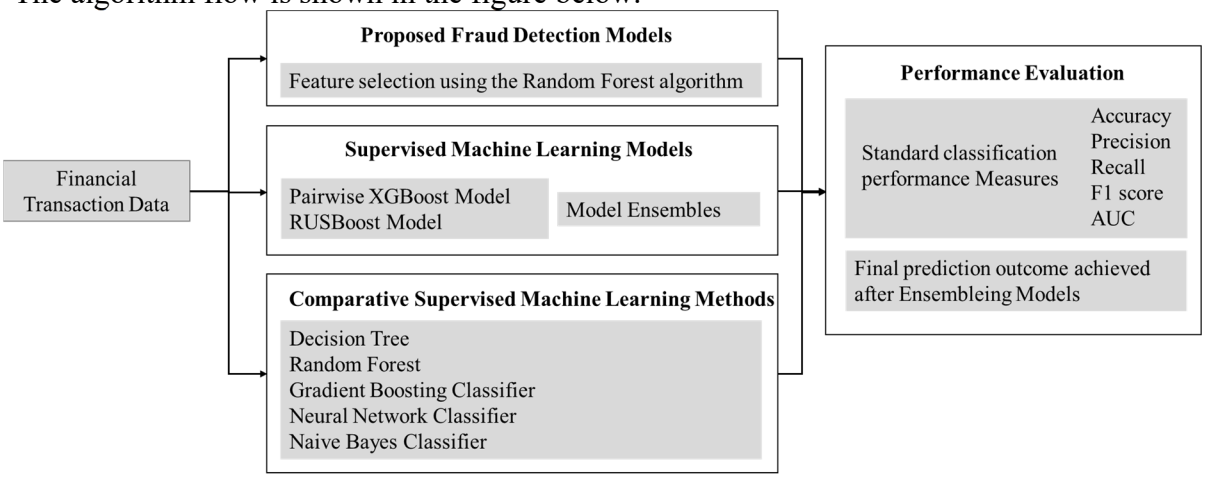

Figure 1 Algorithm flowchart

The algorithmic process is depicted in the following figure (Figure 1). After data pre-processing, we employ Random Forest for automated feature selection and select 20 features of high importance. Then, we input these features into two base models, the XGBoost model, and the RUSBoost model. Both models individually learn and generate their respective predictive results. Ultimately, we utilize a model fusion method to combine the prediction results of the two models, generating the final fused model result. Our model architecture mainly consists of five key steps: data pre-processing, feature selection, XGBoost model training, RUSBoost model training, and model fusion. First, data pre-processing is performed on the raw financial fraud data, which includes data cleaning, feature selection, and handling missing values to ensure data quality and consistency. Second, feature selection is conducted using the Random Forest algorithm, selecting the top 60% of numerical and categorical features of high importance from the many available features for model input. Third, we train the XGBoost model with the selected features as input. We utilize gradient boosting algorithms and Pairwise strategies to enhance model accuracy and generalizability. Fourth, we train the RUSBoost model with the same selected features as input. The Random Under Sampling method is used to address data imbalance problems, thereby improving the model's detection ability of financial fraud. Lastly, model fusion is performed by combining the prediction results of the XGBoost and RUSBoost models. A voting approach is used for the combination to generate the final fused model result. By integrating the prediction results of the two models, we enhance the robustness and performance of the model.

2.4 Data set description

In this section, we utilize the IEEE-CIS fraud detection dataset on Kaggle to validate our proposed model's effectiveness. The dataset comprises 590,540 records, divided into a training set with 472,432 entries and a test set containing 118,108 entries.

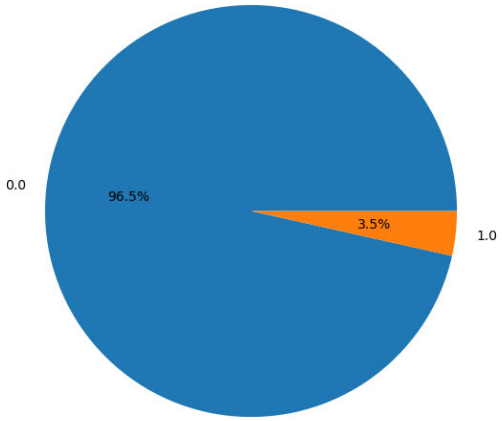

Figure 2 Statistical chart of the number of positive and negative samples in the dataset

Item	Value
Total Count	590540
Training Set	472432
Test Set	118108
Percentage of Minority Class	3.5%
Numerical Features	382
Categorical Features	49

Table 1 Summary of IEEE-CIS Fraud Detection Dataset: In this section, we utilize the IEEE-CIS fraud detection dataset on Kaggle to validate our proposed model's effectiveness.The dataset comprises 590,540 records, divided into a training set with 472,432 entries and a test set containing 118,108 entries.

After merging data tables, not all transaction data requires identification, which also means that there must be a large number of missing values in the data. as the picture shows:

Table 2 Missing Ratio of Training Data Feature Variables

Missing ratio	Numerical variable	Categorical variable
0~25	154	8
25~50	33	5
50~75	2	4
75~100	178	30
Total	367	47

In this dataset, the fraud class, which is the minority, constitutes 3.5% of all samples. The dataset incorporates diverse feature types, including 382 numerical and 49 categorical features. This data, sourced from Vesta's real-world e-commerce transactions and collected in collaboration with security partners, inevitably suffers from data loss. The dataset also includes substantial transaction-sensitive information, closely associated with fraud risk, suggesting the need to preserve the original data as much as possible.

2.5 Feature Selection

In data science and machine learning, feature selection is essential for enhancing models' predictive power and interpretability. Among numerous methods, random forest feature selection is notable. It accounts for feature interactions and exhibits robustness and high computational efficiency. In our study, we employ random forest feature selection, primarily focusing on numerical variables. Our dataset includes 49 categorical variables, the one-hot encoding of which could result in dimensionality issues. As a result, we restrict the application of random forest feature importance to numerical variables, retaining all categorical variables intact.

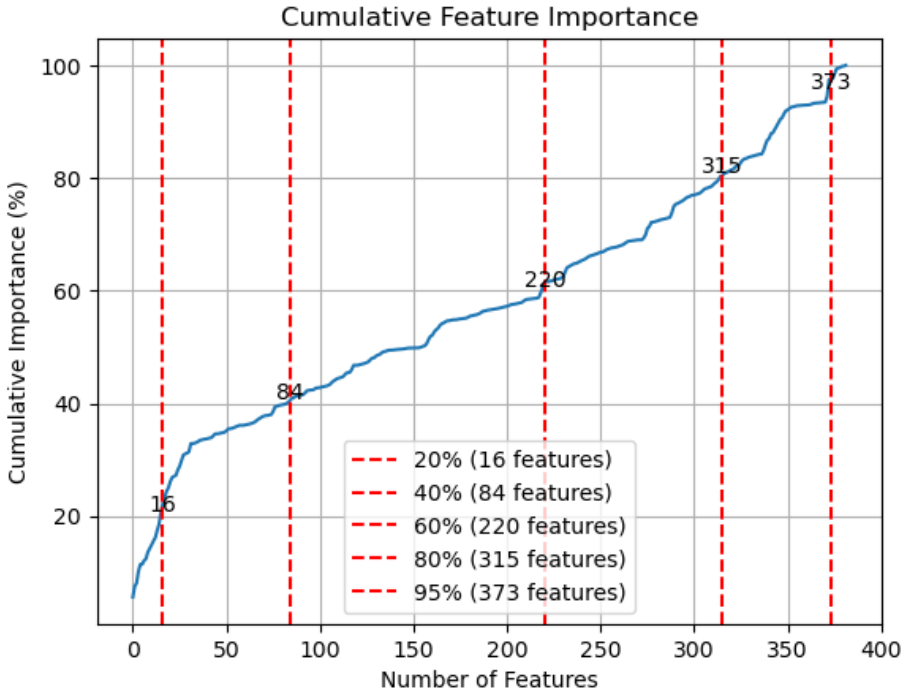

Figure 3 Random forest feature selection

The use of Random Forest for feature selection generated multiple feature subsets, each corresponding to different levels of feature importance. These subsets were formed by selecting the top 20%, 40%, 60%, 80%, and 95% of features. Choosing feature subsets usually requires a balance between model performance, computational expenses, and model interpretability. Consequently, our study incorporates the top 60% of numerical features (220 in total) and all 49 categorical features, leading to a total of 269 input features.

2.6 Contrastive Sorting with XGBoost

In the framework of contrastive-based ranking, our goal is to find a function $f(x)$, where $f(x)=\sum_{k=1}^{K} T_k(x)$, with $T_k(x)$ being the k-th tree and K being the total number of trees. This function should minimize the following objective function as much as possible:

$$obj(\Theta)=L(\Theta)+\Omega(\Theta). \quad (1)$$

Here, L is our loss function, and Ω is the regularization term.

Initially, we define the loss function that is commonly used in Pairwise ranking learning problems, the logistic loss function. For a given pair of samples (i,j), we first set the target difference y_{ij} according to their actual scores. If the actual score of sample i is higher than that of sample j, we set $y_{ij}=1$, otherwise, we set $y_{ij}=-1$. Subsequently, we calculate the predicted score difference $z_{ij}=\hat{y}_i-\hat{y}_j$, where $\hat{y}_i$ and $\hat{y}_j$ are the model's predicted scores for samples i and j. Our objective is to make the predicted score difference z_{ij} approach the target difference y_{ij} as much as possible.

The logistic loss function can be written as $l(y_i,y_j,\hat{y}_i^{(t)},\hat{y}_j^{(t)})=log(1+exp(\hat{y}_j^{(t)}-\hat{y}_i^{(t)}))$. Defining $z_{ij}=\hat{y}_i^{(t)}-\hat{y}_j^{(t)}$, the loss function l can be simplified to

$$l(y_i,y_j,z_{ij})=log(1+exp(-z_{ij}^{(t)})). \quad (2)$$

In this form, the first-order derivative and the second-order derivative of the loss function are:

First-order derivative: $g_{ij}=\frac{\partial l}{\partial z_{ij}}=\frac{-exp(-z_{ij})}{1+exp(-z_{ij})}$

Second-order derivative: $h_{ij} = \frac{\partial^2 l}{\partial z_{ij}^2} = \frac{exp(-z_{ij})}{(1+exp(-z_{ij}))^2}$

We expand the loss function using Taylor's series, resulting in:

$$l(y_i, y_j, z_{ij}^{(t)}) \approx l(y_i, y_j, z_{ij}^{(t-1)}) + g_{ij}(z_{ij}^{(t)} - z_{ij}^{(t-1)}) + \frac{1}{2} h_{ij}(z_{ij}^{(t)} - z_{ij}^{(t-1)})^2 . \quad (3)$$

Here, g_{ij} and h_{ij} are the first-order and second-order derivatives of the loss function at $z_{ij}^{(t-1)}$, respectively..

Substituting the second-order Taylor expansion into the loss term, we get:

$$L(\Theta) = \sum_{i,j} l(y_i, y_j, z_{ij}) \approx \sum_{i,j} \left[l(y_i, y_j, z_{ij}^{(t-1)}) + g_{ij}(z_{ij}^{(t)} - z_{ij}^{(t-1)}) + \frac{1}{2} h_{ij}(z_{ij}^{(t)} - z_{ij}^{(t-1)})^2 \right]. \quad (4)$$

Furthermore, $\Omega(\Theta)$ is the regularization term. For tree models, the regularization term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \| w \|^2 . \quad (5)$$

where T is the number of leaf nodes of the tree, w is the weight corresponding to the leaf nodes, and γ and λ are the regularization coefficients.

Hence, our objective function can be written as:

$$obj(\Theta) = L(\Theta) + \Omega(\Theta) = \sum_{i,j} \left[l(y_i, y_j, z_{ij}^{(t-1)}) + g_{ij}(z_{ij}^{(t)} - z_{ij}^{(t-1)}) + \frac{1}{2} h_{ij}(z_{ij}^{(t)} - z_{ij}^{(t-1)})^2 \right] + \gamma T + \frac{1}{2} \lambda \| w \|^2 . \quad (6)$$

The task is to fine-tune the model parameters, Θ, to minimize the objective function. In the gradient boosting framework, this involves following the negative gradient of the loss function in each iteration. The objective function considers all potential sample pairs during optimization, thereby boosting prediction accuracy. This approach is particularly effective for imbalanced data or multiclass problems. Different computational techniques and optimization algorithms are used in the optimization process to process pairwise samples efficiently and identify optimal model parameters. Ranking XGBoost adeptly merges the complexity of Taylor expansion and tree models, leading to the final optimization objective. We then merge the prediction results of Ranking XGBoost and RUSBoost models. We use a weighted fusion method, based on prediction probabilities, to combine these models, yielding a more robust and accurate model. The fused model outperforms individual models, offering greater accuracy in detecting financial fraud and superior generalization ability.

3. Experiments

3.1 Description of the Dataset

In this study, 60% of the dataset was extracted as the training set, 20% as the validation set, and the remaining 20% as the test set for evaluating the model's performance. For the training set, a 5-fold cross-validation approach was employed. Performance measures such as precision, recall, and accuracy were used to evaluate the model's performance. The experimental results demonstrate the efficacy of the proposed method in accurately identifying fraudulent activities.

3.2 Experimental Setup

The experimental setup had the following configuration:

Table 3 Specific experimental

Name	Configuration environment
CPU	AMD R7-7840H
GPU	RTX Laptop 4060
XGBoost	1.7.5
sklearn	1.0.2

3.3 Controlled experiment

In evaluating financial fraud, we encounter a binary classification problem, categorizing transactions as fraudulent or non-fraudulent. Here, the fraud class is of interest. Spark's confusion matrix, comparing actual and predicted counts, is used to evaluate model performance.

We utilize the AUC (Area Under the ROC Curve) to measure fraud detection efficacy. AUC calculates the area under the Receiver Operating Characteristic (ROC) curve, contrasting false positive and true positive rates. These rates are derived from the confusion matrix and expressed as:

- True Positive (TP): Actual fraudulent transactions correctly identified.
- True Negative (TN): Actual non-fraudulent transactions correctly identified.
- False Positive (FP): Non-fraudulent transactions misclassified as fraudulent.
- False Negative (FN): Fraudulent transactions misclassified as non-fraudulent.

Recall is given by $(\frac{TP}{TP+FN})$ and false positive rate by $(\frac{FP}{FP+TN})$.

The F1-Score provides a more nuanced measure of misclassifications than the Accuracy Metric. As the harmonic mean of Precision and Recall, it's computed as:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}. \quad (7)$$

where Precision is the ratio of True Positives to the sum of True Positives and False Positives, and Recall is the ratio of True Positives to the sum of True Positives and False Negatives.

The AUC is a comprehensive performance metric for classifiers, depicting performance across all thresholds. AUC ranges from 0 to 1, with 1 denoting a perfect classifier, 0.5 signifying random guessing, and values below 0.5 indicating worse than random performance.

Table 4 Machine learning model training results

Model	Accuracy	Precision	Recall	F1	AUC
Dec Tree	0.973	0.659	0.522	0.583	0.756
Grad B CLF	0.972	0.836	0.264	0.401	0.631
Rand FC	0.978	0.798	0.512	0.624	0.753
Neural Classifier	0.973	0.854	0.288	0.431	0.643
Naives Bayes	0.962	0.195	0.02	0.038	0.509

While the discussed machine learning models often demonstrate high accuracy, their AUC values tend to be lower, specifically in the case of the Naive Bayes models. This signifies their limited effectiveness with imbalanced data. Moreover, their inability to compute categorical variables reduces accuracy and robustness. A key strength of our model is its support for numerical variable selection via Random Forest feature selection. It also maintains computation of categorical variables from the original dataset, thus minimizing the effect of missing input values on classification.

For swift model convergence, we utilized the Tree-structured Parzen Estimator (TPE) method for parameter optimization. The model's optimal parameters are:

Table 5 Ranking XGBoost Parameters Table

Parameter	Parameter Description	Value
eta	Learning rate	0.2
lambda	L2 regularization term	0.2
max_depth	Maximum depth of a tree	16
min_child_weight	Minimum sum of instance weight needed in a child	8.6
colsample_bynode	Subsample ratio of columns for each node	0.5
colsample_bytree	Subsample ratio of columns for each tree	0.7

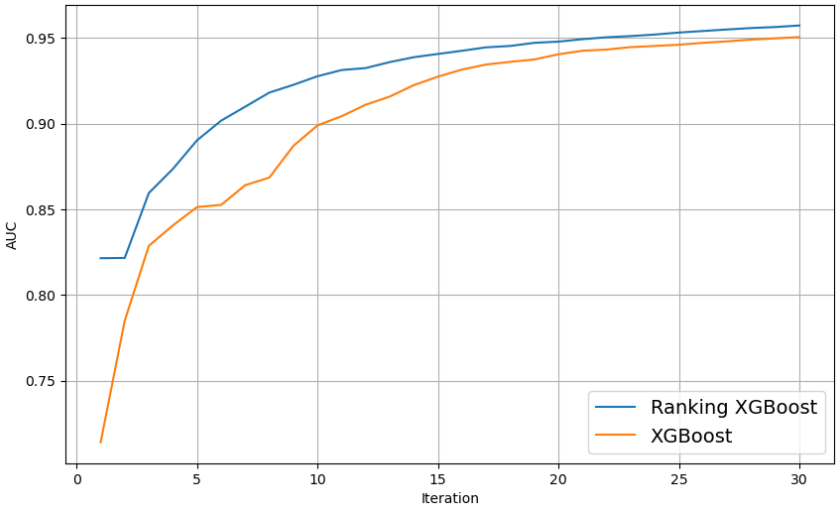

Figure 4 Ranking XGBoost Model and AUC value change table of the first 30 iterations of XGBoost Model

To better analyze model outcomes, we charted the convergence iterations of both the Ranking XGBoost Model and the traditional XGBoost Model for the initial 30 iterations. The graph reveals that our Ranking XGBoost Model, featuring random forest feature importance selection, surpasses the traditional model in terms of AUC values during these early iterations. Moreover, the Ranking XGBoost Model exhibits faster convergence compared to its traditional counterpart.

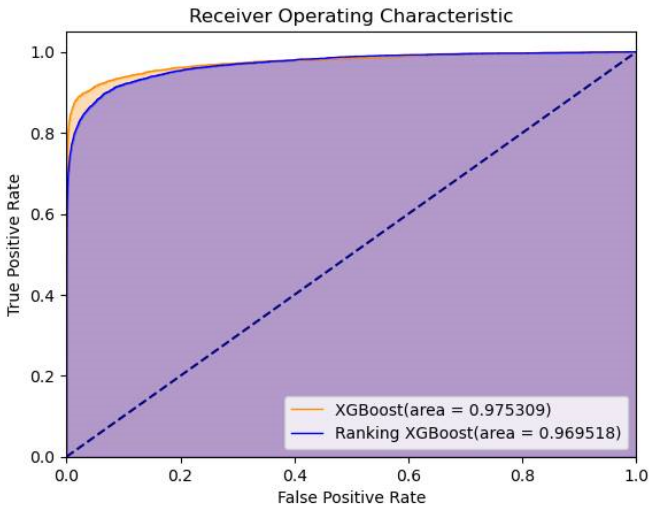

Figure 5 Ranking Last AUC value of the XGBoost model and XGBoost model

In a bid to bolster model robustness, this study incorporates a novel strategy of merging the Ranking XGBoost Model and the RUSBoost Model, employing a prediction probability weighted fusion method. This creates the Proposed model, designed to improve predictive accuracy on new

datasets. Capitalizing on the strengths of both underlying models, the Proposed model delivers enhanced accuracy and performance on unseen data.

Table 6 Comparison of Ensemble model training results

Model	Accuracy	Precision	Recall	F1	AUC	Final iterations
XGBoost	0.989	0.958	0.741	0.836	0.975	1167
Ranking XGBoost	0.980	0.989	0.435	0.604	0.969	507
RUSBoost	0.774	0.114	0.795	0.869	0.8376	30
Proposed Model	0.985	0.780	0.817	0.798	0.904	

The Proposed model, a blend of Ranking XGBoost and RUSBoost, harnesses the strengths of both. RUSBoost excels at handling imbalanced data, and Ranking XGBoost adeptly manages the relative importance of tasks. This synergy allows the Proposed model to outperform individual models in managing imbalanced datasets, achieving a finer balance between precision and recall, and thereby reducing financial fraud likelihood and costs.

Given the high costs associated with each fraud incident, our objective is to minimize fraud as much as possible, even if it means higher false positives. The model might have a slightly lower precision, but its high recall allows it to detect more potential fraudulent activities.

Additionally, the model's iteration count serves as a vital performance metric. Both Ranking XGBoost and RUSBoost show higher training efficiency than the XGBoost model, suggesting faster model outputs in practice. This increases work efficiency, particularly in scenarios demanding real-time feedback.

4. Conclusion

In this study, we proposed and implemented an innovative machine learning framework that addresses the challenges of fraud detection in online shopping behaviors. The framework overcomes issues of data imbalance and feature redundancy commonly encountered in previous research and possesses automatic feature selection and sample balancing capabilities. We conducted experiments on the IEEE-CIS Fraud Detection dataset, achieving a model accuracy of 98.5% and an AUC value of 0.904, demonstrating the effectiveness of the framework. The experimental results show that our machine learning framework exhibits superior performance in dealing with fraud detection in online shopping, validating the practicality and effectiveness of our designed machine learning framework.

5. Summary

Our novel model is based on the powerful learning capabilities of XGBoost and the ability of RUSBoost to handle imbalanced data. By incorporating the Pairwise strategy, our new model improves the recognition rate of the minority class. With this design, our new model not only effectively handles imbalanced data but also enhances the model's ability to identify financial fraud data, demonstrating strong generalization performance. Compared to traditional machine learning algorithms, our new model provides a new and more effective solution with higher recognition accuracy and better generalization performance when dealing with imbalanced financial fraud data.

References

[1] R. Bologa, R. Bologa, and A. Florea, "Big data and specific analysis methods for insurance fraud detection," Database Systems Journal, vol. 4, no. 4, 2013.

[2] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in 2017 International Conference on Computing Networking and Informatics (ICCNI), pp. 1-9, IEEE, 2017.

[3] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit card fraud detection-machine learning methods," in 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), pp. 1-5, IEEE, 2019.

[4] Y. Zhang, J. Tong, Z. Wang, and F. Gao, "Customer transaction fraud detection using xgboost model," in 2020 International Conference on Computer Engineering and Application (ICCEA), pp. 554-558, IEEE, 2020.

[5] C. V. Priscilla and D. P. Prabha, "Influence of optimizing XGBoost to handle class imbalance in credit card fraud detection," in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 1309-1315, IEEE, 2020.

[6] P. Gupta, A. Varshney, M. R. Khan, R. Ahmed, M. Shuaib, and S. Alam, "Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques," Procedia Computer Science, vol. 218, pp. 2575-2584, Elsevier, 2023.

[7] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," Information Sciences, vol. 513, pp. 429-441, Elsevier, 2020.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[9] He, H., & Ma, Y. (2013). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 3247-3252).

[10] A. Howard, B. Bouchon-Meunier, IEEE CIS, inversion, J. Lei, Lynn@Vesta, Marcus2010, Prof. H. Abbass, "IEEE-CIS Fraud Detection," Kaggle, 2019. [Online]. Available: https://kaggle.com/competitions/ieee-fraud-detection.

[11] Chen, R.C.; Chen, T.S.; Lin, C.C. A new binary support vector system for increasing detection rate of credit card fraud. Int. J.Pattern Recognit. Artif. Intell. 2006, 20, 227–239.

[12] Yee, O.S.; Sagadevan, S.; Malim, N.H.A.H. Credit card fraud detection using machine learning as data mining technique. J.Telecommun. Electron. Comput. Eng. 2018, 10, 23–27.

[13] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (pp. 785-794).

[14] F. Wan, "XGBoost based supply chain fraud detection model," in 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), pp. 355-358, IEEE, 2021.

[15] C. Meng, L. Zhou, and B. Liu, "A case study in credit fraud detection with SMOTE and XGboost," in Journal of Physics: Conference Series, vol. 1601, no. 5, article 052016, IOP Publishing, 2020.

[16] D. Trisanto, N. Rismawati, M. M. Femy, and K. F. Indra, "Modified focal loss in imbalanced XGBoost for credit card fraud detection," International Journal of Intelligent Engineering and Systems, vol. 14, no. 4, pp. 350-358, Intelligent Networks and Systems Society (INASS), 2021.

[17] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 40(1), 185-197.

[18] B. Liao, Z. Huang, X. Cao, and J. Li, "Adopting nonlinear activated beetle antennae search algorithm for fraud detection of public trading companies: a computational finance approach," Mathematics, vol. 10, no. 13, article 2160, MDPI, 2022.

[19] R. Akram, N. Ayub, I. Khan, F. R. Albogamy, G. Rukh, S. Khan, M. Shiraz, and K. Rizwan, "Towards big data electricity theft detection based on improved rusboost classifiers in smart grid," Energies, vol. 14, no. 23, article 8029, MDPI, 2021.

[20] S. Mujeeb, N. Javaid, R. Khalid, M. Imran, and N. Naseer, "DE-RUSBoost: an efficient electricity theft detection scheme with additive communication layer," in ICC 2020-2020 IEEE International Conference on Communications (ICC), pp. 1-6, IEEE, 2020.