Research on Credit Card Fraud Risk Identification Based on Integrated Logistic Regression Model

Mo Yu [1,*], Rongpeng Yan [2]

[1] Department of Management Science and Engineering, Harbin Institute of Technology, Harbin, China, 150006

[2] Sports Engineering College, Beijing Sport University, Beijing, China, 100091

* Corresponding Author Email: momo_hit@163.com

Abstract. Effectively identifying credit card fraud using machine learning methods is a significant issue in the financial sector. In this context, machine learning models encounter challenges such as the imbalanced distribution of sample data labels and the high dimensionality of customer feature sets. Addressing these two critical factors, this paper develops an enhanced method for the logistic regression model. This approach not only balances the sample label distribution through resampling but also mitigates the estimation issues arising from the curse of dimensionality. Furthermore, the proposed method addresses the coverage issue of the entire feature set. It solves that resampling can only partially address the curse of dimensionality problem and employs L1 regularization for each logistic regression submodel to further alleviate this issue. Results from simulation experiments and real-world data analysis demonstrate that the proposed method is competitive with logistic regression and several classical classification techniques. This method is not only effective in resolving credit card fraud risks but also has the potential to be extended to other domains.

Keywords: Credit card fraud risk assessment, resample principle, Logistic Regression, Ensemble Learning.

1. Introduction

With the development of e-commerce and continuous innovation in the field of fintech, credit card has gradually become an indispensable financial tool in People's Daily life with its characteristics of convenience, speed and security. However, because of the extensive use of credit cards is also accompanied by the risk of fraudulent transactions, criminals use illegal means to commit credit card fraud to seek improper benefits, resulting in serious economic losses and credibility crisis [1]. Credit card fraud not only causes economic losses to banks and financial institutions, but also may pose a threat to consumer credit records and personal privacy security, thus affecting the stable development of social economy and People's Daily life security. And credit card fraud risk identification is an important means to reduce losses. Therefore, the development of an effective credit card fraud risk identification model is of great significance for maintaining the order of the financial market, protecting the rights and interests of consumers and promoting the healthy development of the economy.

To ensure the security of credit card transactions, numerous scholars have delved into the realm of credit card fraud detection. On one hand, to tackle the issue of high-dimensional data, researchers have historically employed methods such as deep learning. For instance, Yang et al. [2] proposed a neural network system regularized by L1 and L2 norms to address the high-dimensional enterprise credit risk prediction challenge. They incorporated a regularization term into the loss function to achieve feature selection and sparsity. Abhishek et al. [3] suggested leveraging principal component analysis and class imbalance treatment techniques to enhance the classification capabilities in credit card default detection, employing both oversampling and undersampling to correct imbalances in the raw data, thereby accomplishing tasks such as dimensionality reduction and data imbalance processing. Wang et al. [4] advocated for the refinement of credit card inspection data through data reduction, comparing various methods including standalone data sampling, standalone feature selection, data sampling followed by feature selection, and feature selection followed by data

sampling. Their comparative analysis concluded that an integrated approach using Sequential Forward Selection (SFS) and Random Under-Sampling (RUS) for data sampling was particularly effective.On the other hand, to address the problem of data imbalance, Diana et al. [5] proposed techniques grounded in feature selection to identify the most relevant subset of features for fraud prediction by assessing the predictive power of different feature sets, thereby improving the accuracy of machine learning models. Tang et al. [6] introduced a novel approach combining Federated Learning (FL) with Graph Neural Networks (GNN) to deeply explore the relationships between transaction data. Traditionally, researchers have used a one-time fixed sampling method to handle imbalanced samples, which is prone to causing sample bias and non-random results. Furthermore, when dealing with high-dimensional samples, the retained factors may not necessarily be the principal components with significant predictive power for the dependent variable, leading to poor model interpretability and certain disadvantages.

In view of the shortcomings in the above methods, this paper proposes better composite models in both unbalanced data and high-dimensional sample processing. This method alleviates the pathological problem of model estimation while balancing the sample label proportion caused by the dimension disaster. The proposed method has more innovative points considering the characteristics of the whole coverage, through to each logical regression model using L1 regularization technology, processing the resampling method can only partially alleviate the problem of dimension disaster, method more randomness, improve the model of generalization ability and robustness, for credit card fraud risk assessment field provides a new solution.

2. The basic theory and method

2.1. Logistic Regression model

Logistic regression (Logistic Regression) is a statistical model that is widely used in binary classification problems. Despite being name "regression", it is actually a classification algorithm used to predict the probability of an event [7].

The logistic regression model is classified by mapping the output of the linear regression model to the interval [0,1] via a logistic function (Logistic Function). Given the input features $x = (x_1, x_2, \dots, x_n)$, the output p of the logistic regression model is expressed as:

$$p(x) = \frac{1}{1+e^{-(w \cdot x+b)}} \tag{1}$$

Where, $w = (w_1, w_2, \dots, w_n)$is the weight vector of the model, b is a bias term, w · x is the dot product of the weight and eigenvectors and e is the base number of the natural log.

The training objective of the logistic regression model is to maximize the likelihood function, making the model predicts the training data as consistent as possible with the real labels. The likelihood function L is defined as:

$$L(w, b) = \Pi_{i=1}^{m}[p(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i} \tag{2}$$

Where, L(w,b) is an unregularized log-likelihood function for evaluating the likelihood under the model parameters w and b, m is the number of samples, namely, the number of samples in the data set.

The log-likelihood function ℓ is more often used for optimization because it transforms the product to summation, simplifying the calculation:

$$\ell(\mathrm{w}, \mathrm{b}) = \sum_{i=1}^{m}\big[y_i \log\big(p(x_i)\big) + (1 - y_i) \log\big(1 - p(x_i)\big)\big] \tag{3}$$

Model training usually uses a gradient descent algorithm to optimize the ℓ, update the weights and bias:

$$w \leftarrow w - \alpha \frac{\partial \ell}{\partial w} \tag{4}$$

$$b \leftarrow b - \alpha \frac{\partial \ell}{\partial b} \tag{5}$$

Where, α is Learning rate, control weight, and the size of the bias update steps, $\frac{\partial \ell}{\partial w}$ is the log-likelihood function is about the partial derivative of the weight w. $\frac{\partial \ell}{\partial b}$ is the log-likelihood function is about the partial derivative of the bias b.

To avoid overfitting, logistic regression models often use regularization, with L1 (Lasso) and L2 (Ridge) being popular choices. L1 aids in feature selection, and L2 manages model complexity [8]. The log-likelihood function after the regularization is:

$$\ell_{reg}(w, b) = \ell(\mathrm{w}, \mathrm{b}) + \lambda \sum_{j=1}^{n} |w_j| \tag{6}$$

or

$$\ell_{reg}(w, b) = \ell(\mathrm{w}, \mathrm{b}) + \frac{\lambda}{2} \sum_{j=1}^{n} {w_j}^2 \tag{7}$$

Where, λ is regularization parameter, control the strength of the regularization term, used to prevent overfitting, $\boldsymbol{\ell_{reg}}$(w,b): Regularized log-likelihood function, including the original log-likelihood term and the regularization term, $\boldsymbol{\ell_{reg}(w, b) = \ell(\mathrm{w}, \mathrm{b}) + \lambda \sum_{j=1}^{n} |w_j|}$: L1 regularization (Lasso) form,$\boldsymbol{\ell_{reg}(w, b) = \ell(\mathrm{w}, \mathrm{b}) + \frac{\lambda}{2} \sum_{j=1}^{n} {w_j}^2}$: L2 regularization (Ridge) form.

2.2. An integrated logistic regression model based on resampling and rebalancing

The logistic regression model was used to estimate the relationship between the eigenvector X and the dichotomous target variable y. The predicted probability of the model is given by the logical function: formula (1)

The SVM finds the best decision boundary by maximizing the boundary between the two categories. For the case of a linear division, the decision function is:

$$f(X) = sign(w \cdot X + b) \tag{8}$$

Where, w is the normal vector of the decision plane, used to define the maximum interval hyperplane.

Decision tree builds the tree model by recursively splitting the dataset, each segmentation feature based on the optimal threshold. Mathematical descriptions of decision trees often involve conditional probability and information gain.

The KNN algorithm predicts the class of new samples based on the class of k samples of the nearest neighbors. The prediction function is:

$$\hat{y} = mode\left(y_i \middle| for\ i \in nearest(X)\right) \tag{9}$$

Where, $\hat{y}$ is projected category labels, yi is category label of the nearest neighbor sample, nearest (X) can return the index of the k samples closest to the X.

For model evaluation, we created a synthetic dataset with n samples, each featuring d attributes. Among these, k is informative for classification, and rare redundant. The dataset, divided into two categories with distributions controlled by weights w1 and w2, is engineered to mimic imbalanced datasets. The dataset was resampled using the SMOTE algorithm to balance the category distribution:

$$\left(X_{resampled}, y_{resampled}\right) = SMOTE(X, y, k_neighbors = m) \tag{10}$$

Where, $X_{resampled}$, $y_{resampled}$ is features and labels after resampling, m is the number of nearest neighbors used to generate new samples.

For SVM:

$$\hat{w}, \hat{b} = arg \min_{w,b} \{\frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} max(0, 1 - y_i(w \cdot X_i + b))\} \tag{11}$$

Where, C is the regularization constant, which is used to control the penalty intensity of the error term.

The mean AUC value and standard deviation were calculated for each model using the AUC (Area Under the ROC Curve) as the performance indicator:

$$\text{AUC}=\int_0^1 TPR(FPR)dFPR \tag{12}$$

Where, TPR is True Positive Rate, also known as sensitivity or recall. FPR is False Positive Rate.

The application method of the model used in this article is shown in Figure 1.

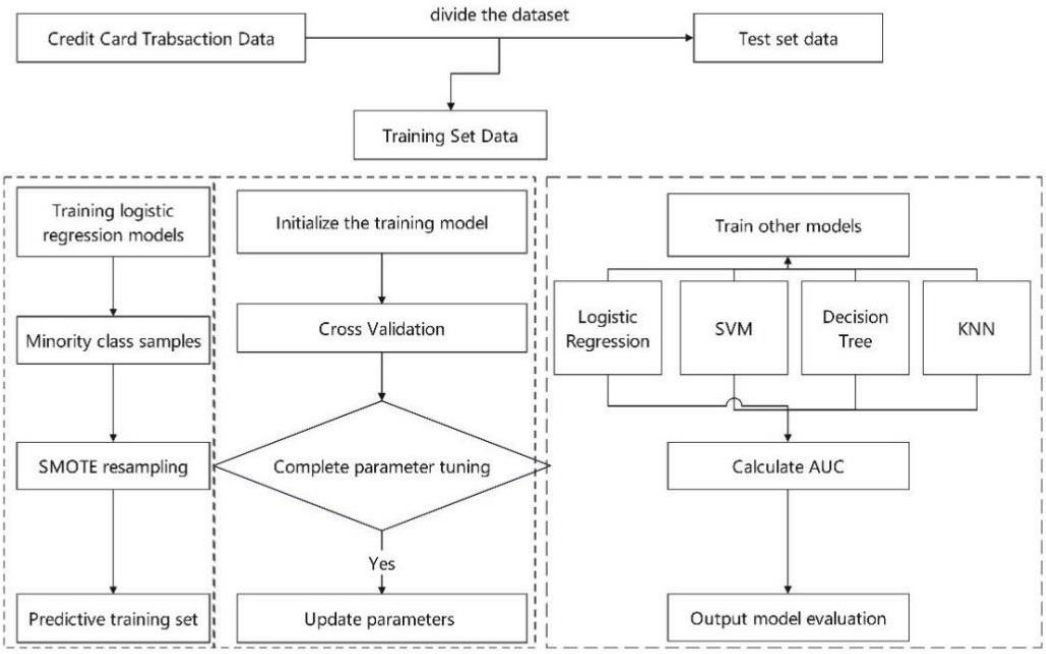

Figure 1. Flow-chart of the fraud detection model

3. Data Analysis

3.1. Data source and experimental setup

The data sources of this paper are divided into two parts: the first part is simulated data from the generation of Sklearn function; the second part is real data. The Credit Card dataset published on the platform can download from: https://www.heywhale.com/mw/dataset/5b56a592fc7e9000103c0442. This data records the transaction information of some credit card holders in Europe in 2013. There are about 280,000 data samples, and only 492 samples are fraudulent transactions. Fraud samples accounted for 0.172%, showing an extremely unbalanced distribution. Because this data set contains a large amount of credit card transaction data, the amount of data is enough to support the establishment and training of effective fraud detection model, and has been widely used in academic research, there are many research results based on this data set, so this paper only uses this data set for experimental evaluation. After desensitization of the original Credit Card dataset, the 28-dimensional features were transformed into V1-V28. Since there are no null values in the fraud detection dataset, there is no need to process missing values, the dataset is first standardized and normalized and then evaluated by cross-validation. The evaluation metrics of the experiments in this paper use AUC values and standard deviations to verify the detection effect of the proposed model [9].

In addition, this paper establishes four comparative methods:

(1) Logistic Regression Model: Logistic regression is a widely used statistical model for binary classification problems. In the context of credit card fraud detection, it can predict whether a

transaction is fraudulent. Its output values naturally fall within the [0,1] range, making it suitable for probabilistic predictions. It is adept at handling linearly separable data and is known for its simplicity, ease of implementation, and interpretability [10].

(2) Support Vector Machine (SVM): SVM is a robust classifier that differentiates categories by identifying the optimal hyperplane in the feature space. It is particularly well-suited for small to medium-sized datasets and can effectively manage high-dimensional data. SVM is also capable of addressing non-linear issues, although it is sensitive to the selection of parameters and kernel functions [11].

(3) Decision Trees: This model forecasts target values by learning straightforward decision rules. It is an intuitive and comprehensible model, ideal for both classification and regression tasks. Decision trees are characterized by their ease of understanding and interpretation, as well as the visual representation of results. They can handle various types of data but are prone to overfitting, necessitating pruning to enhance their generalization capabilities.

(4) K-Nearest Neighbors (KNN): KNN is an instance-based learning method that predicts labels by identifying the closest k training instances to the test sample. It is noted for its straightforward implementation and the absence of a training phase. However, it makes no assumptions about the data distribution and is sensitive to local changes in data, which may result in unstable predictions [12].

In this paper, cross-validation methods to obtain the best parameters corresponding to the constructed detection model on the Credit Card dataset are shown in Table 1.

Table 1. The parameter values on the dataset

Hyperparameter name	Default value	Best hyperparameters
n_estimators	100	100
C	0.1	0.6
feature_sample	3	20
test_ratio	0.4	0.3

In order to facilitate the research, the first 1000 rows of data in the dataset were selected to divide the test set and the training set. Table 2 illustrates the number of samples, their proportion, and categories.

Table 2. Sample size and label of Creditcardfraud Label the sample size ratio

label	Number of samples	proportion
0	284315	99.827%
1	429	0.172%

ACC cannot be used as an evaluation metric because the credit card fraud dataset is an unbalanced sample and using accuracy as an indicator is invalid, so we use AUC as an evaluation metric. In addition, we used simulation experiments and empirical experiments to calculate descriptive indicators such as mean and standard deviation for 10 experiments respectively as the evaluation results of the method.

3.2. Experimental results

3.2.1 Simulation experiments

To validate the efficacy of the S-B-logistic model sampling method, this paper conducts a comparative analysis on a fraud detection dataset using four distinct methods for handling imbalanced data: logistic regression, support vector machine, decision tree models, and the K-nearest neighbors (KNN) algorithm.

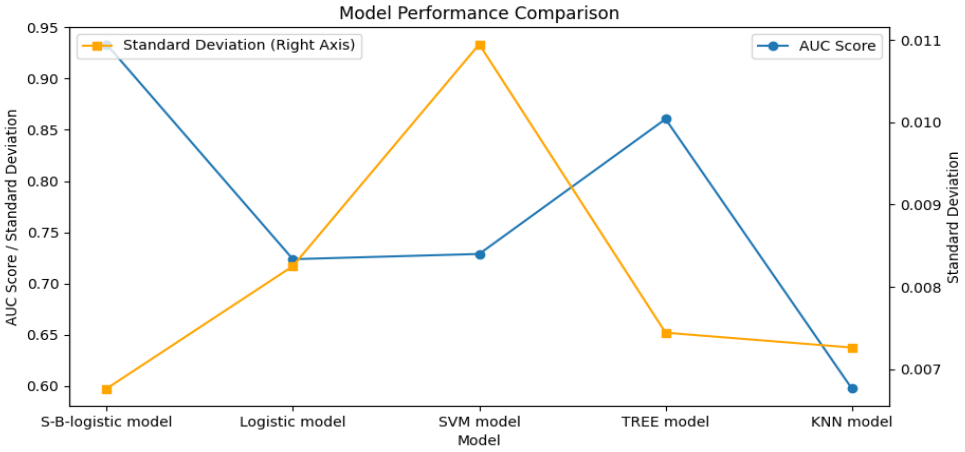

Figure 2. AUC Curves of Different Detection Models

The results synthesized from Table 1 and Figure 3 indicate that the S-B-logistic model sampling algorithm outperforms other algorithms. Specifically, the oversampling algorithm achieved an average AUC of 93% across ten simulation experiments, consistently maintaining a level above 90%, demonstrating high algorithmic efficiency. Moreover, compared to other algorithms, it exhibited a smaller standard deviation, attesting to the model's enhanced stability and reliability. In contrast, other models showed inferior performance in classification tasks when compared to the S-B-logistic model, with greater fluctuations across different tests, indicating poorer stability and, consequently, suboptimal overall predictive performance. Table 3 and Figure 2 show the performance comparison and simulation results and visualization of the five methods.

Table 3. Comparison of sampling methods in terms of performance

Method	AUC	Standard deviation
S-B-logistic model	0.933501	0.006762
Logistic model	0.723737	0.008252
SVM model	0.728827	0.010949
TREE model	0.860525	0.007442
KNN model	0.597357	0.007262

3.2.2 Empirical Experiments

To further assess the effectiveness of the model constructed in this paper, four comparative experiments were set up in accordance with the simulation experiment design, utilizing the Credit Card dataset for empirical experiments. (1) The traditional logistic regression model, which can only validate linear relationships, whereas the credit card fraud dataset contains numerous non-linear relationships. (2) The SVM model [13], which is highly sensitive to parameter selection and has higher costs when processing large-scale datasets [9]. (3) The decision tree model, which is prone to overfitting and has poor generalization capabilities. (4) The KNN model, whose performance is limited when dealing with complex datasets. Ablation experiments were conducted on these five models, and the results demonstrated that the S-B-logistic model showed significant improvements across all metrics compared to traditional models, further validating the effectiveness of the constructed model.

This paper conducts cross-validation on the parameter K of the KNN algorithm, and Table 4 shows the impact of different K values on the model's detection effect.

Table 4. Influence of different K values on the detection effect of the model

K	5	6	7	8	9	10
AUC	0.944025	0.944708	0.945438	0.945715	0.945489	0.944923
Standard deviation	0.006362	0.004020	0.006049	0.005143	0.004012	0.004699

The empirical results of the new model on the Credit Card dataset are shown in Table 5 and Figure 3. Comparison of Model Performance with Different K Values During Oversampling of the Minority Class [10]. As shown in Table 5, as the K value increases, the model's AUC and standard deviation initially rise and then decrease. The model performed best when K was set to 9, with the AUC and standard deviation reaching 94.55% and 0.004012, respectively. Therefore, it is crucial to set an appropriate K value to enhance model performance.

Based on the results of the simulation experiments, the enhanced logistic regression model we proposed showcases significant improvements in handling imbalanced data and high-dimensional features. Compared to traditional logistic regression models, our enhanced model achieved a higher AUC score, indicating superior classification performance.

Table. 5 Experimental results of empirical analysis

Method	AUC	Standard deviation
S-B-logistic model	0.933501	0.006762
Logistic model	0.723737	0.008252
SVM model	0.728827	0.010949
TREE model	0.860525	0.007442
KNN model	0.597357	0.007262

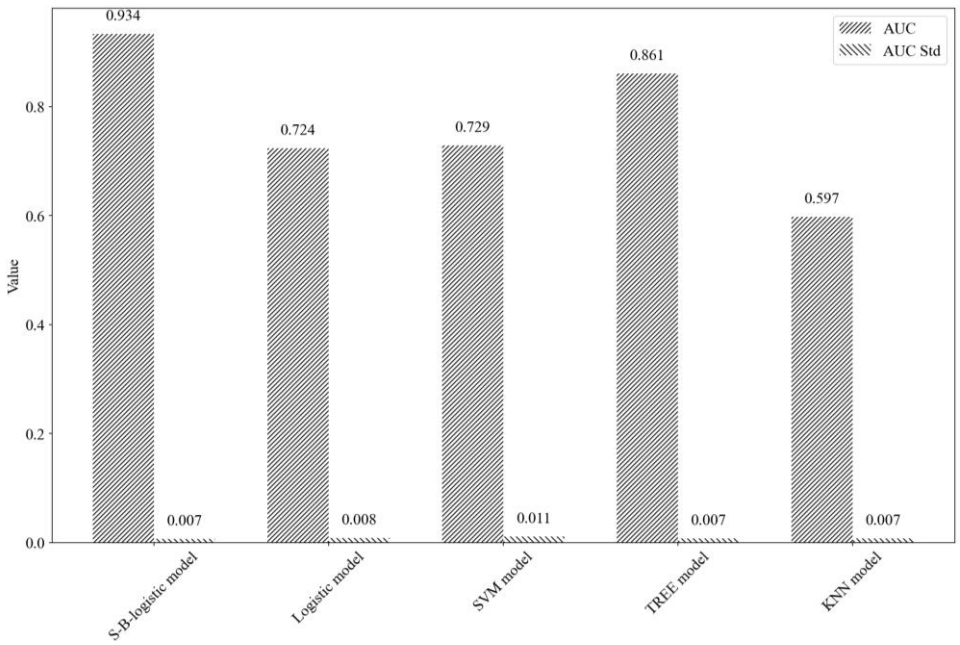

Figure. 3 Bar chart of different model optimization algorithms

In empirical analysis, our enhanced logistic regression model also demonstrated exceptional performance. On the real credit card transaction dataset, the model's AUC reached 0.933, which is significantly higher than that of other comparative models. This result validates the effectiveness and superiority of our enhanced logistic regression model in practical applications.

The enhancements presented in this paper offer several advantages over traditional logistic regression models:

(1) Resampling Technique: By incorporating the SMOTE algorithm, we effectively mitigated the issue of data imbalance, enabling the model to focus more on minority class samples, which is a significant improvement over standard logistic regression techniques that do not inherently address class imbalance.

(2) L1 Regularization: The integration of L1 regularization not only reduced the model's complexity but also enhanced the model's interpretability through feature selection. This is a notable advancement as traditional logistic regression does not inherently perform feature selection.

(3) Ensemble Learning: The use of ensemble learning techniques allowed us to integrate the predictions from multiple models, thereby reducing the model's variance and improving the stability of the predictions. This approach offers a substantial improvement over single logistic regression models that do not leverage the power of ensemble methods.

4. Conclusions

Traditional sampling methods used for imbalanced datasets in credit card fraud detection are characterized by high complexity and sensitivity to outliers, leading to data instability. Additionally, the high dimensionality of the data further degrades the effectiveness of credit card fraud detection. To address these challenges, this paper proposes a credit card fraud detection model based on an ensemble logistic regression model combined with oversampling and the SMOTE algorithm. The proposed algorithm performs oversampling on the credit card fraud detection dataset, synthesizing oversampling for the minority class data to expand the minority class samples, thereby balancing the dataset. The method addresses the issue of feature space coverage by applying L1 regularization to each logistic regression sub-model, solving the problem that resampling methods can only partially mitigate the curse of dimensionality. The integration of an ensemble approach with logistic regression is another key improvement. This method leverages the power of multiple models to reduce variance and improve the stability of predictions, offering a significant advantage over single logistic regression models. Subsequently, the model in this paper was experimentally evaluated on the Credit Card dataset and compared with baseline models. The experimental results show that the proposed model achieved an AUC value of 93.35%, indicating superior detection performance. Further optimization of the model will continue to enhance its stability and detection effectiveness.

The methodology presented in this study not only holds significant application value in the field of credit card fraud detection but also has broad applicability in its core concepts and techniques. In the financial sector, beyond credit card fraud detection, this method can also be applied to loan default prediction, insurance fraud identification, and other risk management scenarios. Moreover, sample imbalance and high-dimensional features are common issues in many fields, such as medical diagnosis, network security, and market analysis, where this method can also be extended to improve the predictive accuracy of models. In future work, we can further explore and optimize resampling and regularization strategies to adapt to a wider range of application scenarios and more complex data distributions. At the same time, we are considering integrating advanced machine learning techniques such as deep learning with this method to further enhance model performance, enabling rapid detection and response to fraudulent activities, and providing more comprehensive decision support for financial institutions.

In summary, the improved logistic regression model proposed in this study provides an effective solution for credit card fraud detection and demonstrates its potential for application in various fields. This method can be further verified and promoted in practical applications, providing a scientific basis for risk management and decision-making in related fields.

References

[1] FOROUGH J,MOMTAZI S.Ensemble of deep sequential models for credit card fraud detection[J].Applied Soft Computing,2021,99:106883.

[2] Mei Y ,K. M L ,Yingchi Q , et al.Deep neural networks with L1 and L2 regularization for high dimensional corporate credit risk prediction[J].Expert Systems With Applications,2023,213(PA):

[3] Agarwal A ,Rana A,Verma N, et al.Enhancement of classification techniques using principal component analysis and class imbalance handling methods in credit card defaulter detection[J].International Journal of Forensic Engineering,2021,5(1):1-18.

[4] Wang H ,Hancock J ,Khoshgoftaar M T .Improving Credit Card Fraud Detection with Data Reduction Approaches[J].International Journal of Reliability, Quality and Safety Engineering,2024,31(04):

[5] Mosa T D ,Sorour E S ,Abohany A A , et al.CCFD: Efficient Credit Card Fraud Detection Using Meta-Heuristic Techniques and Machine Learning Algorithms[J].Mathematics,2024,12(14):2250-2250.

[6] Tang Y ,Liang Y .Credit card fraud detection based on federated graph learning[J].Expert Systems With Applications,2024,256124979-124979.

[7] Zhang Junli, Guo Shuangyan, Ren Cuiping, et al. Study on the personal credit score card model based on logistic regression [J]. Modern Information Technology, 2024,8(05):12-16.DOI:10.19850/j.cnki.2096-4706.2024.05.003.

[8] Du Kang Le. Stochastic optimization algorithm [D] with the L_1 regularization problem. Zhejiang Normal University, 2023.DOI:10.27464/d.cnki.gzsfu. 2023.001931.

[9] Jiang Hongxun, Jiang Junyi, Liang Xun. Review of machine learning-based research on fraud detection of credit card transactions [J]. Computer Engineering and Application, 2023,59 (21): 1-25.

[10] Zhang Junli, Guo Shuangyan, Ren Cuiping, et al. Study on the personal credit score card model based on logistic regression [J]. Modern Information Technology, 2024,8(05):12-16.DOI:10.19850/j.cnki.2096-4706.2024.05.003.

[11] Chen Shou, Yu Xiuyun, Qiu Yongqin, et al. Credit score model based on a semi-supervised SVM [J]. Management Science in China, 2024,32(03):1-8.DOI:10.16381/j.cnki.issn1003-207x.2021.2434.

[12] Ju Chunhua, Chen Guanyu, Bao Fuguang. Consumer finance risk detection model based on kNN-Note-LSTM —— Take credit card fraud detection as an example [J]. Systems Science and Mathematics, 2021,41 (02): 481-498.

[13] DOUZAS G, BACAO F, LAST Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE [J]. Information sciences,2018,465:1-20.