Improved Logistic Regression Model Based on Resampling Techniques

Xinyi Li

College of Mathematical Sciences, Suzhou University of Science and Technology, Suzhou, China, 215009

lixinyi040321@163.com

Abstract. When facing imbalanced samples and high-dimensional features, the performance of traditional logistic regression models may significantly deteriorate, even becoming completely ineffective. Therefore, this paper proposes an improved logistic regression method combined with resampling techniques. Firstly, the proposed method uses the "resampling and encoding" strategy to effectively capture the predictive information that has a significant impact on the classification result while addressing the problems of imbalanced samples and dimension curse. Secondly, the proposed method uses weighted combination of multiple submodels to further use the logistic regression model, thereby improving the model's adaptability and robustness. Furthermore, considering that the number of submodels required to cover all features may be excessive, this paper introduces L1 regularization to penalize each logistic regression submodel, further alleviating the impact of dimension curse on the prediction model. Simulation experiments and real-data analysis show that the proposed method has significantly improved prediction accuracy compared to traditional logistic regression, and has demonstrated certain competitiveness in comparison with some classic classification models.

Keywords: Imbalanced Sample Learning, Dimension Curse, Logistic Regression, Resampling.

1. Introduction

Logistic regression model is a statistical model commonly used for classification problems, especially in binary classification problems [1]. In medical research, logistic regression is often used to analyze risk factors for a certain disease [2-5]; in the financial sector, it is used for credit risk prediction, such as distinguishing "good customers" from "bad customers" or assessing the default probability of customers to help financial institutions conduct risk assessment and decision-making [6-10]. However, in many real-world scenarios, the sample size is often imbalanced and high-dimensional features exist, in which case the traditional logistic regression model may perform poorly or even fail [11-12]. For example, in the personal credit data set, the number of defaulters (positive samples) is usually much smaller than that of non-defaulters (negative samples), and the model may tend to predict negative samples, thereby reducing the prediction accuracy for positive samples. Similarly, in medical diagnosis, the number of patients is obviously smaller than that of non-patients, and directly using logistic regression for training may.

When there is a large difference in sample size, the model may ignore the class with a small sample size in the training process, or assign a lower weight to it, resulting in a decline in the predictive ability of the model. In addition, the dimension curse will also make the model undetermined, resulting in model training errors. In order to solve these problems, regularization technique and resampling method provide effective solutions. Through undersampling or oversampling technology, the majority of class samples can be randomly deleted, or the number of new minority samples can be increased by copying and generating new samples. The weighting method can balance the influence of the uneven number of samples by assigning different weights to different categories of samples. SMOTE algorithm further improves the accuracy and generalization ability of the model for minority class by synthesizing new minority class samples [13-17]. For high-dimensional data problems, regularization methods such as LASSO [18] improve the stability and prediction ability of the model under high-dimensional data by selecting important features. Ridge regression gives up the unbiasability of least square estimation to sacrifice some accuracy for more reliable regression

coefficient. Feature selection in decision tree reduces the number of features by recursively dividing the feature space while retaining key information, thus alleviating the problems caused by high-dimensional features [19].

In order to overcome the limitations of existing models, an improved logistic regression model based on resampling technique is proposed in this paper. To solve the problem of sample imbalance, the method of "resampling and coding" is adopted to resampling unbalanced samples and resampling categorical variables to capture their key prediction information and improve the accuracy of the model. Then, a logistic regression sub-model is constructed for each recoded categorical variable, and the probability of variables is predicted by these sub-models. Another logistic regression model is used to weight the prediction results of each sub-model to obtain the final prediction results, thus enhancing the adaptability and robustness of the model. At the same time, L1 regularization is applied to each submodel to alleviate the dimensional curse problem and improve the explainability of the model. Simulation experiments and actual data analysis show that the proposed method has a great improvement over the original logistic regression in the case of high dimensional data and unbalanced samples.

2. Theory and method

Suppose the prediction vector is $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{p})^\mathrm{T}$, the response variable is $Y = y \in \{0, 1\}$, and given X=x, the conditional probability of Y=1 is as follows:

$$P(y = 1|X) = \frac{1}{1+\exp(-X^T\beta)} \tag{1}$$

Where $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1}, \beta_p)^\mathrm{T}$ is the coefficient vector. The task now is to estimate the coefficient vector based on the sample data $\{X_i, y_i\}_{\mathrm{i}=1}^{\mathrm{N}}$. However, in the case of sample label imbalance and dimension curse, the solution based on the minimization criterion of negative log-likelihood function will fail, on the one hand, because the inverse of Hessian matrix cannot be solved, on the other hand, when gradient update is carried out based on quasi-Newton iterative algorithm, the gradient provided by a few class samples will be covered by the majority class samples. Based on this, this paper proposes a scheme to randomize the sample set to solve these two problems. Specifically, on the one hand, the sample is randomly sampled, assuming that the sampling frequency is K, for k=1,2,... K, there are

$$D_k = \mathcal{B}(\{X_i, y_i\}_{i=1}^N) \tag{2}$$

Where D_k is the subsample data set obtained by the KTH sampling and the number of samples is n = n1+ n2, where n1 represents the number of samples of the majority class and n2 represents the number of samples of the minority class. Then, the features of each subsample data set are randomly sampled, and the number of features is q, q<p. For k=1,2,... K, there are

$$D_k^* = \mathcal{F}(\{D_k : X_{ik}\}) \tag{3}$$

Where D_k is the subsample data set obtained by feature sampling for the KTH time, the number of samples is n = n1+ n2, and the number of features is q. Then it is brought into the KTH logistic regression model for training, whose negative log-likelihood loss function is

$$L(\beta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{4}$$

Where $p_i = P(y_i = 1 \mid X_i)$, the coefficient vector can be solved based on the minimization of the negative log-likelihood loss function. Since we need to cover as many features as possible, the feature dimension of each subsample set cannot be too small.

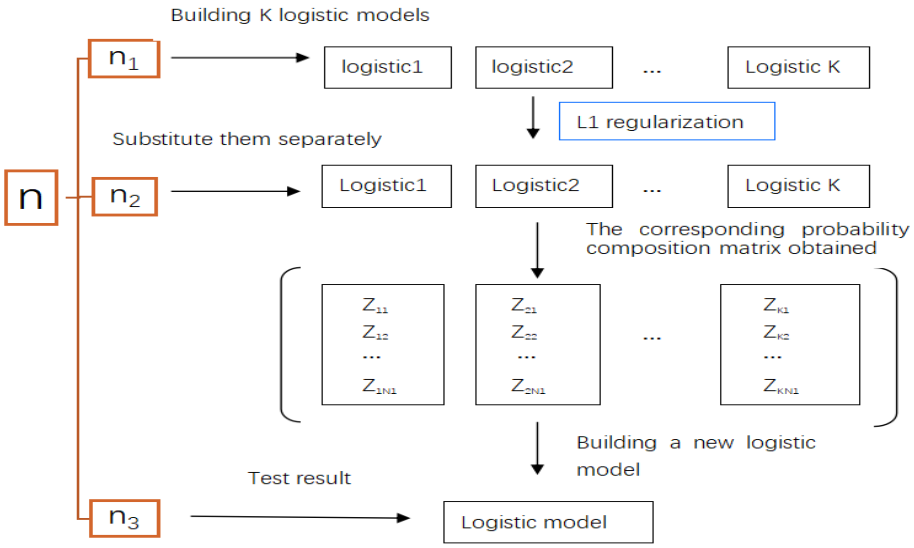

Figure 1. Algorithm flow of the improved logistic regression model based on Bootstrap

Based on this, we use the negative log-likelihood loss function based on L1 penalty, as follows:

$$L(\beta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] + \lambda/2 \parallel \beta \parallel \quad (5)$$

Where λ is the penalty coefficient, $\|\beta\|$ is L1 penalty. Then the coefficient vector is solved based on the quasi-Newton iterative algorithm, and k logistic regression models with randomization property are trained. It is worth noting that the subsample set brought in deals with sample disequilibrium and dimension curse, so the performance of this K logistic regression model will not be affected by label ratio and high-dimensional features. Next, we represented the output probability of K logistic regression models as $\boldsymbol{Z} = (Z_1, \cdots, Z_K)^T$, assumed it to be a K-dimensional predictive random variable, and re-paired it with the response variable to form the corresponding sample data set $\{(\boldsymbol{Z}_i, y_i)\}_{i=1}^{N_1}$, which was retrained in a new logistic regression, and its negative log-likelihood function under the corresponding L1 penalty was shown as follows:

$$L(\beta') = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] + \lambda/2 \parallel \beta' \parallel \quad (6)$$

Where $p_i = P(y_i = 1|Z_i)$, for the optimization of the above formula, we still use the quasi-Newton iterative algorithm. The specific algorithm is shown as follows.

(1) Input:$\{(\mathrm{Zi}, \mathrm{yi})\}_{i=1}^{N_1}$, Initial weight $\beta^{(0)}$, learning rate α, iterations K, penalty coefficient λ

(2) For h=0 to H-1 do

(3) Calculated predicted value:

$$\hat{y}_i^{(k)} = \sigma(\mathrm{Z_i^T}\beta^{(k)}) \quad (7)$$

(4) Calculated gradient:

$$\nabla f(\beta^{(k)}) = \sum_{i=1}^{N_1}(\hat{y}^{(k)} - y_i)Z_i + \lambda * \mathrm{sgn}(\beta^{(k)}) \quad (8)$$

(5) Calculate the Hessian approximation:

$$H^{(k)} = \mathrm{diag}(\hat{y}_i^{(k)}(1-\hat{y}_i^{(k)})) \quad (9)$$

(6) Update weight:

$$w^{(k+1)} = \beta^{(k)} - \alpha H^{(k)} \nabla f(\beta^{(k)}) \quad (10)$$

We call the above method an improved logistic regression model based on Bootstrap, and the calculation process is shown in Figure 1.

3. Data analysis

3.1. Simulation studies

To more fully evaluate the predictive performance of the proposed methods, in this section we use the categorical dataset generation function provided by the sklearn package in Python. We set different experimental conditions, including the number of features, sample size, unbalanced sample proportion and effective feature proportion, etc. The comparison models included Logistic Regression, support vector machine (SVM), Decision Tree and Linear Discriminant Analysis (LDA). In the experiment, we used 70% of the data as the training set and 30% as the test set. Due to the existence of unbalanced samples, traditional evaluation indexes such as Precision may not be able to truly reflect the performance of the classifier. Therefore, we choose AUC (Area Under the ROC Curve) as the main evaluation index, and its calculation formula is as follows:

$$AUC=\sum_{i=1}^{n-1}\frac{(FPR_I+FPR_{i+1})}{2}*(TPR_{i+1}-TPR_i) \tag{11}$$

Where n is the number of points on the ROC curve, FPRi and TPRi are the false positive rate and true rate of the i th point, respectively. To more accurately reflect the overall fitting ability of the model, we ran 10 replicates and measured the model performance using the mean and standard deviation of the AUC score. The higher the AUC mean, the better the prediction performance of the model. A smaller standard deviation means that the model is more stable across experiments.

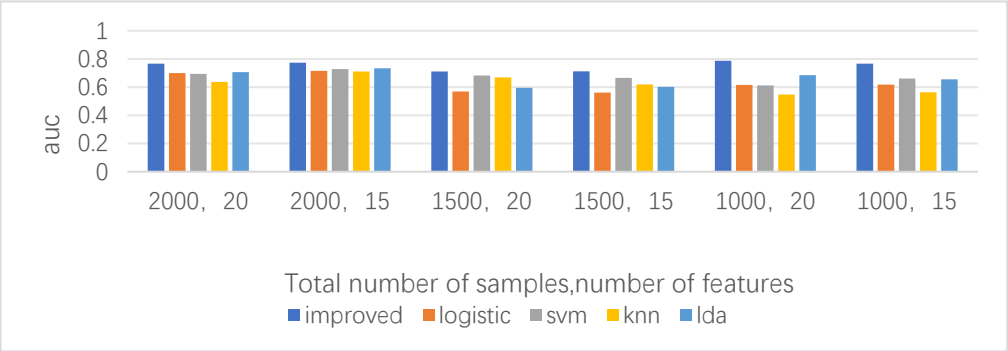

Figure 2. Average prediction performance of each model under different experimental Settings

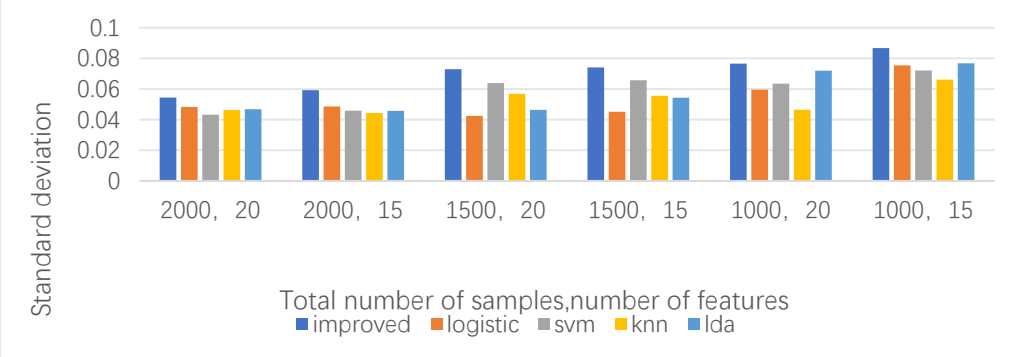

Figure 3. Standard deviation of prediction performance under different experimental Settings

First of all, Table 1 and Figure 2-3 summarize the AUC score performance of each model under the conditions of different sample numbers and feature numbers. Each column represents the predicted AUC value of the model under different experimental Settings, showing the relative performance of each model under different conditions. It is worth noting that the proposed model maintains a good fit under almost all experimental conditions, showing strong prediction ability. It can be seen from the chart analysis that different models show differences under different combinations of the number of features and the number of samples, and this model can stably obtain a higher AUC score in most cases. Although the standard deviation of the proposed method is slightly larger than that of other models, there is little difference on the whole. This standard deviation is

caused by the randomness of bootstrap. Because the standard deviation is small, the proposed method still shows some robustness.

Secondly, we change the unbalanced sample proportion to evaluate the performance of the model. Specifically, we divide the sample into three categories: 0.95: 0.05, 0.9: 0.1, and 0.8: 0.2. Table 2 and Figure 4-5 summarize the AUC score performance of each model under different sample proportions.

Table 1. Summary statistics of fit degree of each model under different experimental Settings

N and P	Improved-log	logistic	svm	knn	lda
2000,20	0.7668 (0.05)	0.6996 (0.05)	0.6940 (0.04)	0.6369 (0.05)	0.7062 (0.05)
2000,15	0.7739 (0.06)	0.7156 (0.05)	0.7277 (0.05)	0.7114 (0.04)	0.7340 (0.05)
1500,20	0.7116 (0.07)	0.5689 (0.04)	0.6832 (0.06)	0.6692 (0.06)	0.5956 (0.05)
1500,15	0.7122 (0.07)	0.5615 (0.05)	0.6654 (0.07)	0.6187 (0.06)	0.6022 (0.05)
1000,20	0.7870 (0.08)	0.6162 (0.06)	0.6126 (0.06)	0.5479 (0.05)	0.6846 (0.07)
1000,15	0.7668 (0.09)	0.6179 (0.08)	0.6606 (0.07)	0.5639 (0.07)	0.6553 (0.08)

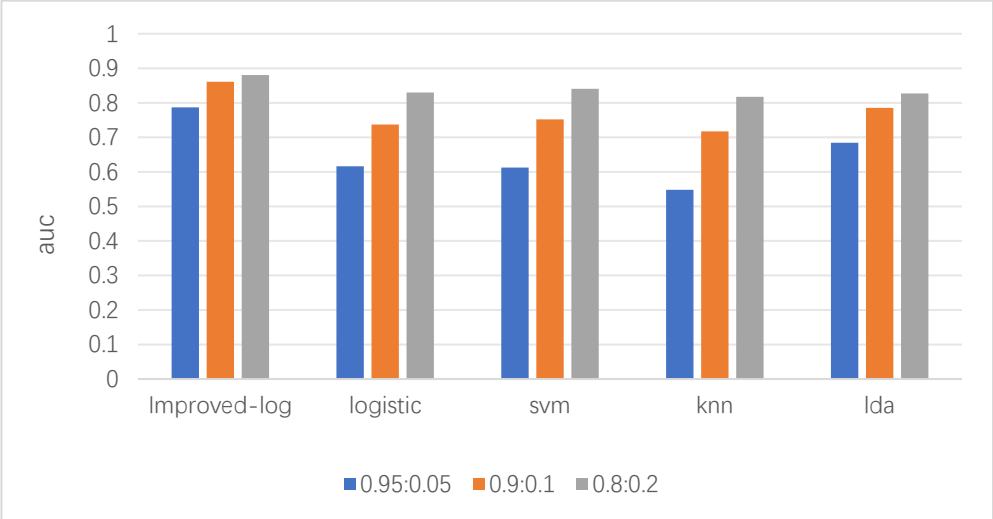

Figure 4. Average prediction performance of each model under different sample proportions

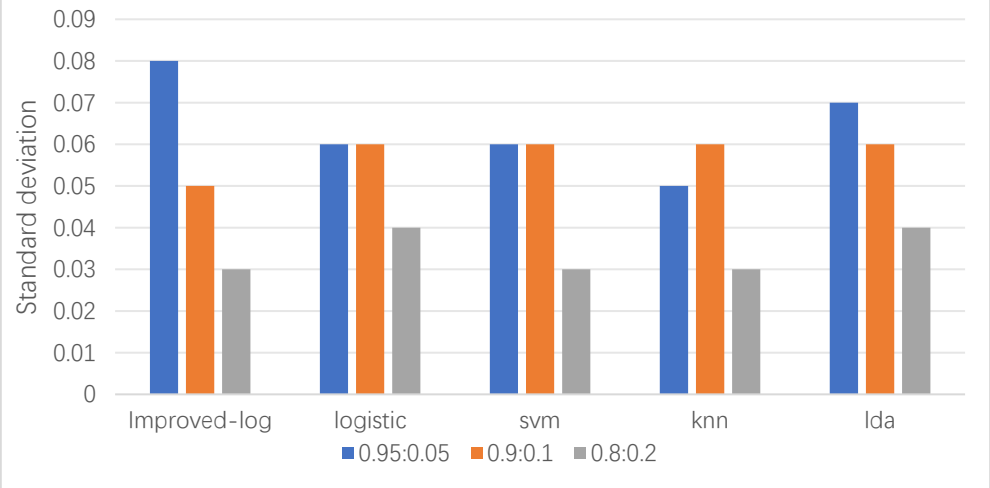

Figure 5. Standard deviation of prediction performance under different sample proportions

Each column represents the predicted AUC value of the model at different scales, showing the relative performance of each model at different scales. It can be seen from the chart analysis that the models show obvious differences when the proportion of samples is different. For the above five models, the smaller the proportion of samples, the higher the AUC value, indicating the better fitting degree of the models. The standard deviation of these models is generally low under different sample proportions. When the sample proportion is 0.8: 0.2, the standard deviation of each model is low, which reflects the robustness of the model.

Table 2. Summary statistics of fit degree of each model under different sample proportions

Radio	Improved-log	logistic	svm	knn	lda
0.95: 0.05	0.7870 (0.08)	0.6162 (0.06)	0.6126 (0.06)	0.5479 (0.05)	0.6846 (0.07)
0.9: 0.1	0.8612 (0.05)	0.7374 (0.06)	0.7522 (0.06)	0.7174 (0.06)	0.7853 (0.06)
0.8: 0.2	0.8806 (0.03)	0.8301 (0.04)	0.8406 (0.03)	0.8173 (0.03)	0.8272 (0.04)

3.2. Real data analysis

In this paper, experiments are mainly carried out on ECG200 data set and financial stock price data set to check the performance of the proposed method in practice. First, in the biomedical field, electrocardiogram (ECG) data is widely used to assess heart health. ECG200 data set from the web site, timeseriesclassification.com, contains 200 patients electrocardiogram (ecg) data, each patient records has 96 ecg curve discrete observations. All subjects were divided into two groups: normal heartbeat and myocardial infarction. The task of this study was to predict whether a patient had a myocardial infarction based on discrete ECG curves. Since this dataset belongs to high-dimensional data, we constructed sample data with unbalanced ratio of positive and negative samples for experiments to verify the performance of the model under unbalanced data. Consistent with the simulation experiment, we used 70% of the total sample as the training set and the remaining 30% as the test set. The mean AUC score and standard deviation of 10 and 100 replicates were used for the evaluation measures. Table 3 and Figure 5-6 show the detailed results.

Table 3. ECG200: Prediction level and robustness of each model after 10 and 100 experiments

method	10-AUC	10-std	100-AUC	100-std
Improved-log	0.7773	0.1132	0.7601	0.1644
logistic	0.7258	0.1055	0.7379	0.1364
svm	0.7083	0.1449	0.6933	0.1872
knn	0.7199	0.1647	0.7028	0.1818
tree	0.6478	0.1302	0.6216	0.1339

Through real data experiments in this field, the improved logistic regression method combined with resampling technique proposed in this paper shows significant advantages in dealing with high-dimensional data and unbalanced samples. Table 3 and Figure 7-8 show the specific results. First, Table 3 summarizes the predictions of different models for a disease based on ECG and visualizes the box plots of AUC scores for 10 and 100 replicates, as shown in Figure 7-8. As can be seen from the chart, the prediction results of the traditional logistic model under different simulation times are inferior to those of the improved model. Compared with other comparison models, the proposed method also shows excellent effect in the prediction of diseases. Second, in the realm of quantitative finance, predicting fluctuations in share prices is a fundamental challenge. This paper utilizes a dataset comprising 2000 days of CSI 300 index data sourced from the Tushare platform as the experimental sample. The focus is not on specific stock exchanges or industry backgrounds; rather, it aims to evaluate the performance of the proposed model. The experiment employs factor data from the previous day to forecast stock price movements for the following day. The factors considered include: percentage change in stock price, current trading price, opening price, intraday high and low

prices, settlement price, trading volume, turnover rate, total turnover, P/E ratio, P/B ratio, market capitalization, circulating market value, and circulating share capital. This data set is also a high-dimensional data set, and we chose a trading period with uneven stock price fluctuations for modeling. Consistent with the simulation experiment, we used 70% of the total sample as the training set and the remaining 30% as the test set.

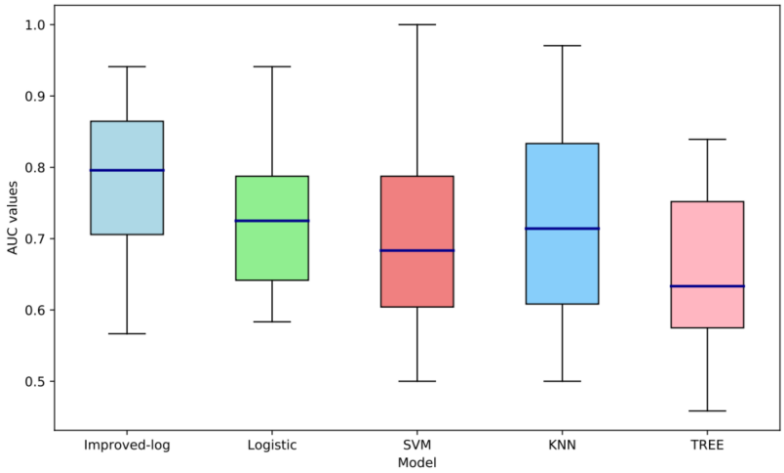

Figure 6. ECG200: Box diagram of AUC scores for 10 experiments

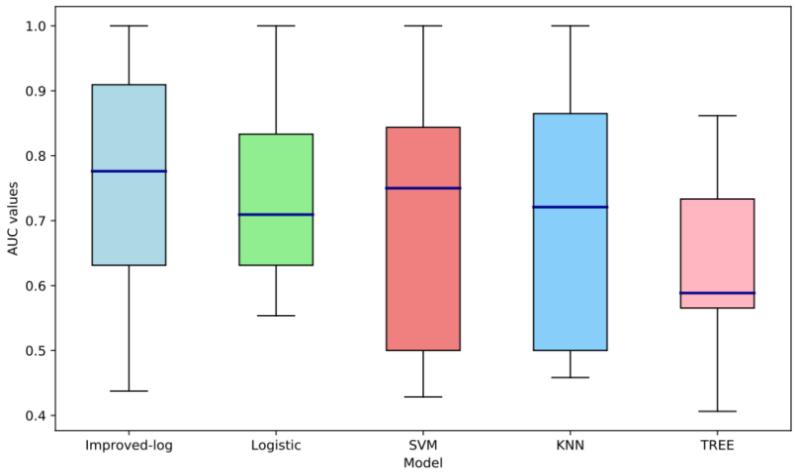

Figure 7. ECG200: Box diagram of AUC scores for 100 experiments

The mean AUC score and standard deviation of 10 and 100 replicates were used for the evaluation measures. Table 4 and Figure 9-10 show the detailed results.

Table 4. Stock prices: Prediction of each model during 10 and 100 experiments

method	10-AUC	10-std	100-AUC	100-std
Improved-log	0.8198	0.1835	0.8429	0.1412
logistic	0.5000	0.0000	0.5000	0.0000
svm	0.4980	0.0128	0.4996	0.0017
knn	0.5757	0.0938	0.5362	0.1250
tree	0.4949	0.0124	0.4877	0.0556

As can be seen from Table 4 and Figure 9-10, the performance of traditional logistic regression completely fails in the stock price prediction task in the face of unbalanced samples and high-dimensional data problems, while the improved logistic regression model still maintains efficient forecasting ability, and despite some instability, it is still optimal compared with other comparison models. This unrobustness can be alleviated by cross-validated hyperparameter optimization method. The reason why the proposed method is better than the original model is that it increases the randomness of the samples through resampling technology, improves the fitting ability of the model by using the idea of divide and conquer, and solves the problem of sample imbalance and alleviates

the curse of dimension. In addition, since the proposed method is also model interpretable, indicators that have an important impact on the rise and fall of stock prices can also be mined in quantitative trading, as shown in Figure 9 below. It can be seen from the figure that the total value of the stock market, the price-earnings ratio and the transaction price have a relatively important impact on the rise and fall of the stock price among all factors.

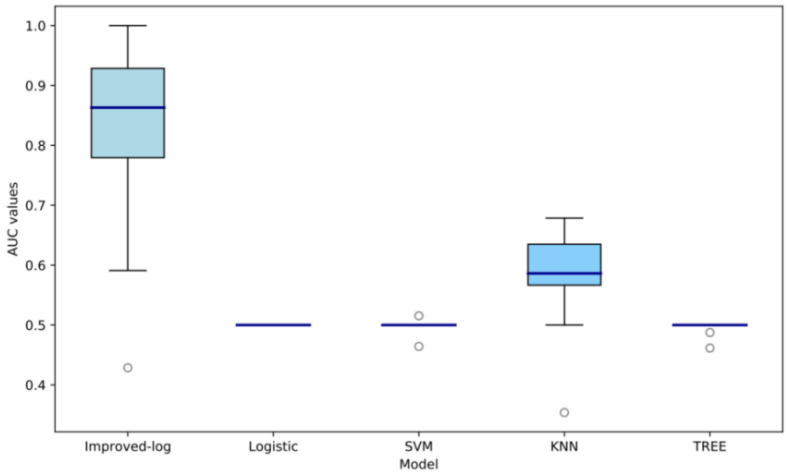

Figure 8. Stock prices: Box plot of AUC scores for experiment 10

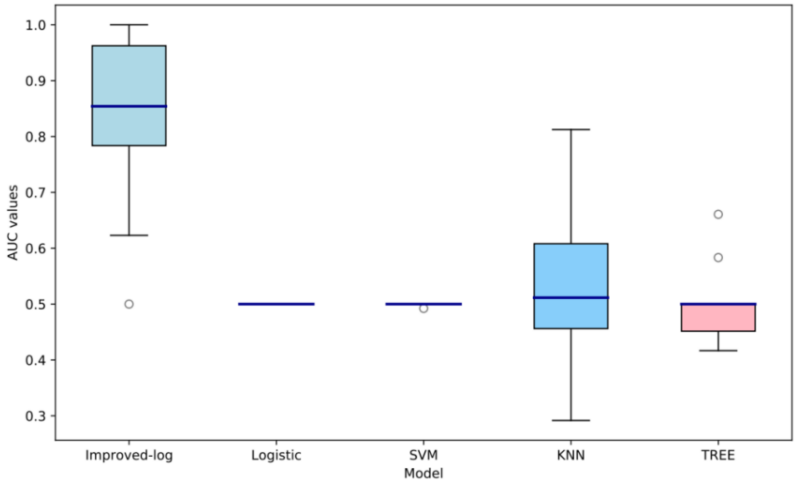

Figure 9. Stock prices: Box chart of AUC scores for 100 experiments

4. Conclusion

According to the above research results, the improved logistic regression method based on resampling technology successfully solves the problem that the performance of traditional logistic regression is significantly reduced under the characteristics of sample imbalance and high dimension. Through the resampling and encoding strategy, this method can effectively capture the feature information that has a key impact on the classification result.

(1) Enhancement of model robustness and adaptability: By weighting multiple submodels and combining them with the quadratic application of logistic regression, this method enhances the model's adaptability and robustness under complex data sets. This divide-and-conquer strategy effectively alleviates the negative impact of dimensional curse on traditional models, and makes the models show higher accuracy in high-dimensional data.

(2) Introduction and performance improvement of L1 regularization: The application of L1 regularization technology to punish each submodel further alleviates the dimensional curse and improves the generalization ability of the model. The experimental results show that this method not only improves the prediction accuracy, but also shows a certain competitiveness in comparison with

other classical classification models, which proves its robustness and applicability under different experimental conditions.

References

[1] Xiao Song, Huang Jiewu. First order approximate knife cut modified ridge estimation in binary logistic regression model [J]. Journal of Hunan University of Arts and Sciences (Natural Science Edition), 2019, 36 (02): 6-13. (in Chinese)

[2] Wang Xuanyu. Application of nearest neighbor classification coupling algorithm based on logistic regression in medical orthopedic classification [J]. Modern Information Technology, 2024, 8 (11): 158-162.

[3] Shi Yulin, Chunyi, Liu Jiayi, et al. Study on lung cancer risk warning model based on tongue image feature logistic regression [J/OL]. Chinese Journal of Traditional Chinese Medicine Information, 2024, 1-8.

[4] Zhang Yiquan. Research on Alzheimer's Disease Classification Algorithm Based on Regularized Logistic Regression [D]. Qufu Normal University, 2021.

[5] BA Mengru, Yin Xiaohong, Li Shaoyuan. A predictive model of cognitive impairment in Parkinson's disease based on multivariate logistic regression [J]. Journal of Intelligent Science and Technology, 2019, 6 (02): 232-243.

[6] Zhang Jian, Liu Lin, Li Qian. Research on optimization of M-Score model based on nonlinear logistic regression [J]. Journal of Business Accounting, 2024, (10): 83-86.

[7] Gu Jiaqing. Research on Credit Risk Assessment Based on Classification Models [D]. Nanchang University, 2024.

[8] Liu Chengxing. Application research of logistic Regression Model in risk user detection of banking financial institutions [J]. Financial Technology Times, 2022, (09): 71-73.

[9] Lai Hongqing. Research on financial data classification method of enterprise secondary entrepreneurship based on logistic regression [J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2021, 38 (05): 114-119.

[10] Xie Rui. Design and Analysis of Personal Credit Scoring Card Based on Logistic Regression [D]. Beijing Jiaotong University, 2022.

[11] Han Meng, Li Chunpeng, Li Ang, et al. Unbalanced data stream classification algorithm based on weighted and Dynamic selection [J/OL]. Computer Engineering and Applications, 2024, 1-18.

[12] Wu Zoling. Oversampling and classification of unbalanced data in diabetes prediction [D]. Xi 'an University of Technology, 2024.

[13] Lin Haitao, Cao Jianming, Li Hansen, et al. An undersampling method for class imbalance problem using Genetic algorithm [J]. Journal of Hanshan Normal University, 2024, 45 (03): 11-23.

[14] Guo Mingjuan, Xu Haning, Xiao Hui, et al. Prediction algorithm of pre-slip stage based on double sampling random forest: a case study of Huangshi No.5 iron ore treatment block in Hubei Province [J]. Science Technology and Engineering, 2019, 24 (14): 5733-5741.

[15] Hou Sai, Cheng Runkun, Liu Da. Transformer Fault Prediction Based on Secondary Sampling and Ensemble Learning Method [J]. Smart Power, 2024, 52 (07): 40-47.

[16] Li Aihua, Liu Wanxin, Chen Sifan, et al. Research on SMOTE-BO-XGBoost Integrated Credit Scoring Model for Unbalanced Data [J/OL]. China Management Science, 1-10 [2022-10-15].

[17] Long Qiangkui, Sun Xuezi, Meng Xiangfu. An unbalanced data over-sampling method based on sample potential and noise evolution [J]. Computer Applications, 2024, 44 (08): 2466-2475.

[18] Li Yi, Zhang Benhui, Guo Yuyan. Lasso-Lssvm prediction model optimized by improved particle swarm optimization [J]. Statistics and Decision, 2021, 37 (13): 45-49.

[19] Li Lili, JIN Shi, ZHOU Kai He. Research on optimal subsampling algorithm of big Data based on Ridge regression model [J]. Systems Science and Mathematics, 2022, 42 (01):50-63. (In Chinese)