Enhanced Logistic Regression Using Stacking Algorithm for Imbalanced and High-Dimensional Data

Yanbo Li

College of Science, Yanbian University, Yanji, China, 133002

Li-Yanbo@outlook.com

Abstract. In binary classification tasks, logistic regression models often perform poorly and can even fail when dealing with issues such as class imbalance and the curse of dimensionality. To address these problems, this paper proposes an improved logistic regression model based on the stacking approach. First, the method constructs multiple logistic regression sub-models by employing a dual randomization strategy on both samples and features. The specific strategy involves retaining a small number of minority class samples and drawing a corresponding number of samples from the majority class in proportion, while also randomly selecting features to form subsets. This strategy not only effectively alleviates the curse of dimensionality but also can handle the issue of sample class imbalance. Second, the method uses the probability outputs of the logistic regression sub-models as new predictive variables, and these encoded representations are used to train a new ensemble model, thereby enhancing the model's ability to balance fitting bias and prediction variance. Moreover, the classifier used to construct the new ensemble classification model is model-agnostic. Through validation with simulation experiments and real data analysis, the results show that this method significantly improves predictive performance compared to the original logistic regression model. Additionally, compared to some classical classification models, the proposed method also demonstrates strong competitiveness.

Keywords: Logistic Regression, Category Imbalances, Dimension Disaster, Stacking, Resampling.

1. Introduction

Given the swift evolution of information technology and the advent of the big data era, the utilization of data analysis and machine learning models has pervaded diverse sectors. Notably, the logistic regression model has emerged as a ubiquitous instrument for addressing numerous classification challenges, owing to its exceptional interpretability, computational simplicity, straightforward result interpretation, and capability to generate probabilistic outputs. Its applications span various domains, including financial credit [1], biomedicine [2], and public health [3]. Specifically, in the financial sector, logistic regression models facilitate personal credit scoring and risk assessment. In the medical realm, they assist in examining disease risk factors and expediting diagnostic processes. Nevertheless, as data volumes and complexities escalate, particularly in binary classification scenarios, the deployment of logistic regression models confronts several formidable obstacles. Notably, the issue of imbalanced samples and high-dimensional features stands out prominently. Primarily, the issue of imbalanced samples poses a considerable challenge in the current application of logistic regression models. In situations where one class of samples vastly outnumbers the other in a classification task, logistic regression models frequently lean towards predicting the majority class. This tendency notably diminishes the predictive capability of the model with respect to the minority class, subsequently decreasing the overall efficacy of the model [4]. This occurrence is particularly pronounced in domains such as financial fraud detection and disease screening, where despite the minority class comprising a small portion of samples, they often hold substantial commercial or medical significance. For example, in the context of credit card fraud detection, the volume of fraudulent transactions is substantially lower than that of legitimate transactions, yet the identification of these few fraudulent transactions is paramount. Furthermore, with the exponential advancement in data acquisition and storage technologies, high-dimensional data has emerged as a prevalent trend. In numerous practical applications, the quantity of features significantly surpasses the number of samples, presenting formidable challenges to model training. The presence of high-

dimensional feature spaces not only intensifies the complexity of the model but also has the potential to induce overfitting issues [5]. This not only undermines the model's ability to generalize but also results in the squandering of substantial computational resources. Consequently, addressing overfitting and bolstering the generalization capability of models in high-dimensional data settings has emerged as a pivotal challenge in the utilization of logistic regression models.

Regarding the aforementioned issues, handling imbalanced samples and high-dimensional features have become two important research directions in the field of machine learning. Many researchers have proposed various effective solutions. For instance, Su Yapeng et al. proposed a novel SA-YOLO adaptive loss object detection algorithm to achieve efficient and accurate target detection under imbalanced samples conditions [6]. Miao Yue et al. used a binary classification balanced cross-entropy loss function based on the improved Focal Loss function as the loss function for the LightGBM model, to improve the situation where model accuracy is reduced due to the imbalance between positive and negative samples [7]. Leng Qiangkui et al. proposed an imbalanced data oversampling method based on sample potential and noise evolution [8]. Wen Xinxin and Kong Xiangbo address the issue of sample imbalance through the Synthetic Minority Over-sampling Technique (SMOTE), extract features using factor analysis, and construct a predictive model (SMOTE-PSO-LSSVM) by optimizing the Least Squares Support Vector Machine (LSSVM) algorithm with Particle Swarm Optimization (PSO) [9]. In terms of handling high-dimensional features, regularization techniques are one of the most widely applied solutions. Regularization reduces model complexity by imposing constraints on model parameters, thereby lowering the risk of overfitting. For instance, Zhou Bu et al. propose a test method based on the L2 norm for the high-dimensional two-way analysis of variance problem where the number of observed samples is less than the number of observed variables [10]. Based on the Gaussian regression model, Li improved the existing regular estimation method by using the coordinate algorithm and KKT condition, and proposed a more effective regularization estimation method of variable selection (or feature extraction) for high-dimensional data [11]. Building a two-stage high-dimensional feature selection ensemble learning method [12] using information entropy, Zhou Gang et al. In recent years, with the continuous development of computer technology, more and more researchers begin to use deep learning technology to process high-dimensional data. For example, Feng Xiaorong and Qu Guoqing proposed a big data feature selection algorithm [13] based on deep learning and random forest in view of the disadvantages of the feature selection algorithm on the dimension reduction effect and poor stability of high-dimensional big data. He Qiang et al. adopts the method of adding penalty constraint to the explanatory variable set, which can effectively deal with the thorny problem with high dimension of Internet big data; the stability of the optimal big data explanatory variable set for predicting quarterly GDP growth is high [14]. Despite the significant advantages of deep learning in high-dimensional data processing, its implicit learning of features greatly reduces the model interpretability. Therefore, in some application scenarios, such as financial credit risk assessment or disease diagnosis, models with interpretability or probability output are still needed. Therefore, how to optimize the logistic regression model on unbalanced samples and high-dimensional features has become an important topic in current research [15].

In summary, this paper proposes an improved logistic regression model [16] based on the stacking idea. Multiple logistic regression submodels are constructed by the double random sampling technique, namely sampling samples and features simultaneously, so as to balance the sample category while reducing the feature dimension. Further apply the model fusion technique, take the probabilistic output of the logistic regression submodel as a new feature, and train a new integrated learning model with stacking ideas to improve the accuracy of prediction. When constructing a new classification model, we follow the model free selection strategy, allowing the selection of the most appropriate classifier according to the specific problem, providing flexibility for the model. To verify the effectiveness of the proposed method, this paper performs simulated data analysis and real data analysis. For the simulation experiment, the comprehensive performance of the model were determined by changing the mean and standard deviation of the scores of the AUC (area under the

curve) under the repeat test with the total number of samples, the number of features and the proportion of unbalanced samples. Finally, the actual data is tested on the two tasks of myocardial infarction prediction and stock price rise and fall. The proposed method effectively improves the original logistic regression, and compared with some other classical classification models, the proposed method also shows strong competitiveness.

2. Theory and method

2.1. Sample resampling strategy

Sample resampling is a technique used to handle the class imbalance problem in machine learning. In many practical applications, there may be significant differences in the number of samples from different categories, and this unbalanced data distribution may lead to bias in the training process of the model, making the model more inclined to predict majority classes and ignore minority classes. Suppose that the original data set is expressed as:

$$\mathrm{D} = \{(\mathrm{x_i}, \mathrm{y_i})\}_{\mathrm{i=1}}^{\mathrm{N}}, \tag{1}$$

Where x_i represents the feature vector of the i th sample, y_i represents the corresponding category label, and N represents the total number of samples. The category label y_i takes a value of 0 or 1 and representing minority and majority classes, respectively. And $P(y_i = 1) = p << 0.5$. To balance the category distribution, based on the resampling strategy:

(1) Retain minority samples: extract all minority samples from the original data set and retain them, denoted as:

$$\mathrm{D_m} = \{(\mathrm{x_i}, \mathrm{y_i}) | \mathrm{y_i} = 1\}, \tag{2}$$

(2) From majority samples: randomly select samples equal to the number of minority samples from majority samples. That is, using the random sampling method, N_m samples are drawn from $D_M = \{(x_i, y_i) | y_i = 0\}$, denoted as: D'_M (where N_m and N_M represent the number of minority and majority class samples in the original data set, respectively). The resampled dataset is expressed as:

$$\mathrm{D_n} = \mathrm{D_m} \cup \mathrm{D_M}, \tag{3}$$

(3) Secondary random sampling: Sampling the characteristics of the uniformly formed subsets of the sample. For the dataset D_n formed from the above steps, perform secondary random sampling on the feature vectors of each sample to obtain a balanced sample set D', to further alleviate the dimensionality curse problem and increase the robustness of the model. Next, we conduct theoretical error analysis of the resampling. Specifically, assume training a model $g(x)$, its expected generalization error can be represented as:

$$\mathrm{E_{gen}(g)} \leq \mathrm{E_{train(g)}} + \sqrt{\frac{1}{2\mathrm{n}}(\mathrm{H(g)} + \log\frac{2}{\delta})}, \tag{4}$$

Where $H(g)$ is the VC dimension of the model, n is the number of samples, and δ is the confidence level. Since the sample distribution in D' is more uniform, $E_{train}(g)$ is more likely to be close to $E_{gen}(g)$, thereby reducing the risk of overfitting. The sample resampling strategy is designed to construct a balanced dataset by retaining a limited quantity of samples from the underrepresented class while drawing an equivalent number of samples from the dominant class. This approach aims to improve the performance of the classification model specifically when predicting samples belonging to the minority class. Additionally, the sample resampling strategy not only equalizes the distribution of classes within the dataset but also strengthens the robustness and predictive accuracy of the model through the utilization of multiple rounds of random sampling.

2.2. Logistic regression submodel construction

Conduct the resampling process k times, each time randomly drawing different samples from D_M to obtain k different balanced sample sets $D_1, D_2, D_3, \ldots, D_k$. Then, train the logistic regression model $f(x;\theta)$ on the resampled subsets, where θ represents the model parameters of logistic regression. Specifically, the logistic regression model is used to estimate the probability of a sample belonging to the minority class (default). Its specific form is as follows:

$$\mathrm{P(y=1|x)} = \frac{1}{1+\mathrm{e}^{-(\beta_0+\sum_{j=1}^{d}\beta_j x_j)}}, \tag{5}$$

Where y is a binary target variable (0 represents normal, 1 represents default), $\mathrm{x} = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_d]$ is the feature vector, β_0 is the intercept term, and β_j is the weight of the j-th feature. The parameters of the logistic regression model are obtained through maximum likelihood estimation (MLE). The likelihood function can be defined as:

$$L_1(\beta) = \prod_{i=1}^{n} [P(y_i = 1 | x_i)]^{y_i} [1 - P(y_i = 1 | x_i)]^{1-y_i}, \tag{6}$$

Where n is the number of samples, and y_i and x_i represent the target value and feature vector of the i-th sample, respectively.

In order to maximize the likelihood function, calculate the log-likelihood, which can be solved using gradient descent or the Newton-Rafson method. The following is the solution process of the Newton-Lafson method:

(1) Define the log-likelihood function:

$$\log L(\beta) = \sum_{i=1}^{n} [y_i \log P(y_i = 1 | x_i) + (1 - y_i) \log[1 - P(y_i = 1 | x_i)], \tag{7}$$

Where, $P(y_i = 1 | x_i)$ are the predicted probabilities of the logistic regression model, which can be calculated by the formula:

$$P(y_i = 1 | x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x_i)}}, \tag{8}$$

(2) Select an initial parameter estimate $\beta^{(0)}$ to calculate the gradient:

$$\nabla \log L(\beta) = \sum_{i=1}^{n} [y_i - P(y_i = 1 | x_i)], \tag{9}$$

(3) Calculate the Hessian matrix:

$$H(\beta) = -\sum_{i=1}^{n} P(y_i = 1 | x_i)[1 - P(y_i = 1 | x_i)] x_i x_i^T, \tag{10}$$

(4) Newton-son iteration, update parameters:

$$\beta^{(t+1)} = \beta^{(t)} - \frac{\nabla \log L(\beta^{(t)})}{H(\beta^{(t)})}, \tag{11}$$

(5) Convergence check, output result: If the parameter update amount is less than a preset threshold or reaches the preset number of iterations, the iteration is stopped. When the iteration stops, $\beta^{(t)}$ is the parameter estimate maximizing the log-likelihood function.

To prevent overfitting, the L2 regularization term needs to be introduced into the optimization problem:

$$\min_{\beta} J(\beta) = -\ell(\beta) + \lambda \sum_{j=1}^{d} \beta_j^2, \tag{12}$$

Where λ represents the regularization parameter, used to control the balance between model complexity and fit ability. Finally, the performance of each logistic regression submodel was evaluated and the best model parameters were selected. Through the above procedure, k logistic regression submodels were constructed. Each submodel can provide probability estimates about the sample belonging to the minority class, which will be used as new features for subsequent model fusion.

2.3. Model fusion framework

After constructing multiple logistic regression submodels and obtaining their probabilistic outputs, we turn to the design of the model fusion framework. Model fusion is an ensemble learning technique that improves the overall performance by combining predictions from multiple models. The probability outputs serve as new features, with each logistic regression sub-model LR_k outputting a probability for each sample x:

$$p_k(x) = P(y = 1 | x; \theta_k), \tag{13}$$

These probabilistic outputs are treated as new features, and the new feature vector $z(x)$ consists of probabilistic outputs from all submodels:

$$z(x) = [p_1(x), p_2(x), ..., p_k(x)], \tag{14}$$

These new feature vectors $z(x)$, when combined with the original feature vectors x, form a richer fused feature space. It can be represented as:

$$Z(x) = [x^T, z(x)^T]^T, \tag{15}$$

In which, x^T represents the transpose of the original feature vector, and $z(x)^T$ represents the transpose of the probability output feature vector.

Next, train the fusion model. On the fused feature space $Z(x)$, train a fusion model F. This model can be any machine learning algorithm suitable for classification tasks, such as logistic regression, support vector machines, random forests, or neural networks. The goal of the fusion model is to use the information extracted from the sub-models to improve classification performance. When choosing a fusion model, consider the complexity of the model, training time, predictive performance, and interpretability. Train the fusion model F on the fused feature space $Z(x)$, select the optimal fusion model parameters φ using techniques such as cross-validation, evaluate the performance of the fusion model on the test set, and output key performance indicators such as AUC, standard deviation, etc. By integrating the forecasts of numerous sub-models, the fusion model is able to capture more intricate decision boundaries, ultimately enhancing prediction precision. Given its foundation on the collective assessment of multiple models, the fusion model exhibits reduced sensitivity to noise and outliers. Furthermore, by amalgamating various models, the fusion model mitigates the potential for overfitting issues that may arise in individual models.

3. Data Analysis

3.1. Simulation studies

In the simulation experiments conducted, the primary objective was to validate the efficacy of the proposed methodology in addressing the issues of sample imbalance and dimensionality disaster. To this end, the study initially generated a binary classification dataset utilizing the sklearn package in Python. This dataset encompassed specific parameters such as the total count of samples, the number

of features present, the quantity of effective features, the count of redundant features, and the proportion of class imbalance.

More specifically, the features predicted within this dataset were primarily categorized into three distinct groups: effective features, redundant features, and random noise features. Effective features held a direct correlation with the target variable, serving as crucial predictors. In contrast, redundant features exhibited a linear relationship with effective features but did not contribute any additional predictive value. Additionally, randomly generated noise features were introduced to ensure their independence from the target variable. The inclusion of both redundant and stochastic noise features was strategic, aimed at simulating the challenging scenario of dimensionality disaster.

To rigorously assess the performance of the enhanced logistic regression model across varying data complexities, the experimental design is outlined as follows: Initially, the dataset is partitioned into three distinct sets: the training set, the validation set, and the test set. The training set is exclusively utilized for the iterative optimization of model parameters, thereby facilitating parameter learning for the model. Subsequently, the validation set serves as the basis for model selection and hyperparameter fine-tuning. Ultimately, once model training is finalized, the test set is employed to quantify the model's performance. To ensure a comprehensive evaluation of the proposed methodology, this study employs the original logistic regression model, along with Support Vector Machine (SVM) [17], K-Nearest Neighbor (KNN) classifier [18], Linear Discriminant Analysis (LDA) [19], and Naive Bayes classification [20] as comparative benchmarks, with a particular emphasis on contrasting the results with the baseline logistic regression. As the primary metric for performance evaluation in this paper, AUC is adopted. AUC effectively gauges the model's capability to discriminate between positive and negative samples across different thresholds, with higher values indicating superior classification performance. This metric is particularly advantageous in mitigating the impact of imbalanced datasets, thereby enabling an objective assessment of the model's classification performance.

The AUC is computed as follows:

$$AUC = \sum_{i=1}^{N-1} \frac{(FPR_i - FPR_{i+1}) \times (TPR_i + TPR_{i+1})}{2}, \tag{16}$$

Among them, N is the threshold quantity, TPR_i and FPR_i are the true positive rate and false positive rate at the i-th threshold, respectively.

Furthermore, to reflect the robustness level of the model, repeated experiments are conducted and the average and standard deviation of all AUC scores are used as the evaluation criteria:

$$\overline{AUC} = \frac{1}{10} \sum_{k=1}^{10} AUC_k, \tag{17}$$

Where AUC_k is the AUC value of the k-th experiment, $k = 1,2,3, \ldots ,10$. The standard deviation is represented as:

$$\sigma_{AUC} = \sqrt{\frac{1}{9} \sum_{k=1}^{10} (AUC_k - \overline{AUC})^2}, \tag{18}$$

The mean AUC value serves as an indicator of the overall efficacy of the model, whereas the standard deviation quantifies the model's performance stability. A diminished standard deviation signifies that the model exhibits uniform performance across varied experimental conditions, underscoring its robust nature. The empirical outcomes are presented in Tables 1-3 and Figures 1-3, which follow.

Table 1. Model comparison results for different total sample settings

Experimental setting	Total number of samples: 6,000		Total number of samples: 8,000	
Indicator output	AUC	standard error	AUC	standard error
Improved_LOG	7.6535E-01	3.3680E-02	7.7521E-01	2.1846E-02
Logistic model	6.9426E-01	3.7220E-02	6.3940E-01	2.4484E-02
SVM model	6.6754E-01	2.8950E-02	6.9815E-01	2.1900E-02
KNN model	5.6016E-01	1.3534E-02	6.5441E-01	2.7003E-02
LDA model	6.9126E-01	3.2758E-02	6.1475E-01	2.3226E-02

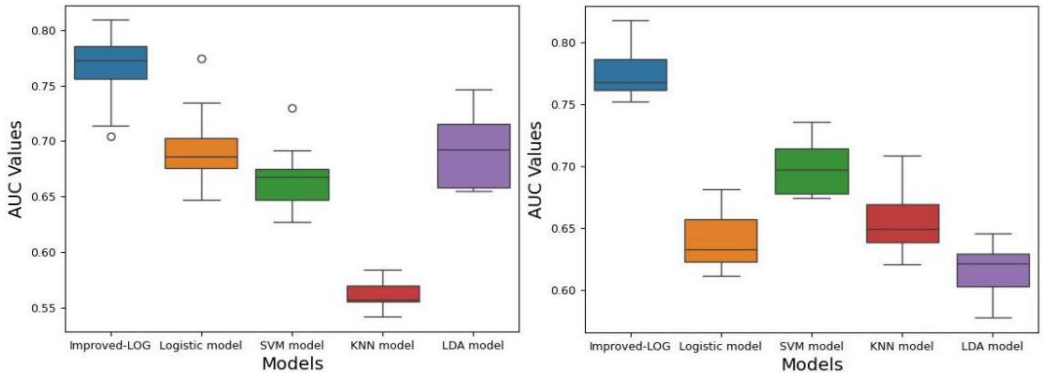

Figure 1. Boxplot of the model contrast under different total number of samples settings

Table 2. Results of model comparison under different number of feature settings

Experimental setting	Number of features: 20		Number of features: 50	
Indicator output	AUC	standard error	AUC	standard error
Improved_LOG	8.1604E-01	2.0068E-02	8.1736E-01	1.0695E-02
Logistic model	7.8604E-01	1.3655E-02	7.9158E-01	1.4978E-02
SVM model	7.7725E-01	1.9304E-02	7.7006E-01	1.2909E-02
KNN model	7.2625E-01	1.5740E-02	6.5070E-01	1.3414E-02
LDA model	7.8304E-01	1.5458E-02	7.9255E-01	1.6521E-02

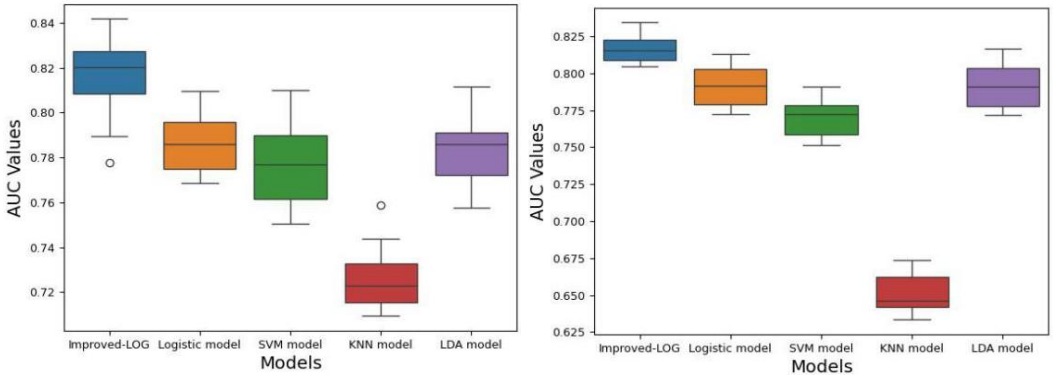

Figure 2. Boxplot of model contrast under different feature number settings

Table 3. Model contrast results under the unbalanced proportions of different categories

Experimental setting	Imbalance ratio: [0.95, 0.05]		Imbalance ratio: [0.85, 0.15]	
Indicator output	AUC	standard error	AUC	standard error
Improved_LOG	7.5670E-01	2.9045E-02	8.2812E-01	8.8660E-03
Logistic model	6.3563E-01	2.1986E-02	7.8069E-01	1.1674E-02
SVM model	6.2120E-01	1.9033E-02	7.6517E-01	1.1704E-02
KNN model	5.8276E-01	2.3218E-02	7.3921E-01	8.9220E-03
LDA model	6.1028E-01	2.4005E-02	7.7172E-01	1.2379E-02

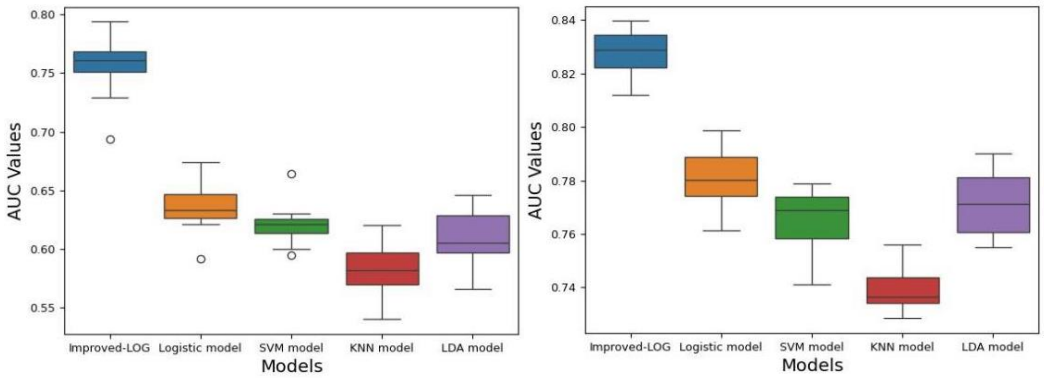

Figure 3. Boxplot of the model contrast under different class of unbalanced scale settings

The improved logistic regression model, utilizing random sampling & model fusion, outperforms original across settings. Mitigates dimensionality, improves class imbalance handling, boosting AUC. Predictive stability enhanced in large datasets. Feature selection & combo boost accuracy, mitigate overfitting. Robust across features, especially for imbalanced classes. Excels in predictive performance with higher median AUC & narrower IQR. Competitive vs. SVMs, KNNs, LDA, achieving optimal in select scenarios.

The improved logistic regression model through random sampling and model fusion, experimental results fully demonstrate the effectiveness of the proposed method in dealing with sample class imbalance and the curse of dimensionality problems, as well as its advantages in enhancing predictive performance and model stability.

3.2. Real data analysis

Two real-world datasets, ECG200 for cardiac anomaly detection and a stock price dataset, were selected for thorough examination. The ECG200 dataset (see Table 4 for experimental output results), comprising 200 samples with 100-dimensional ECG signal features, aims to predict myocardial infarction (MI) amidst imbalanced labels and high dimensionality. Target variables are binary: 0 for normal heart function, 1 for MI, with MI being less common. Accessible via UCR Data, this dataset challenges accurate MI prediction. The stock price dataset tracks historical price fluctuations(see Table 5 for experimental output results), with a binary target (-1 for decrease, 1 for increase) based on technical factors like opening/closing prices, trading volume, turnover, and P/E ratio. The goal is to forecast the CSI 300 index's daily price trend using these factors. Detailed data sourced from Tushare. Both datasets underwent 50 repetitions of training/testing splits, with mean AUC and standard deviations calculated for proposed and benchmark methods, assessing model stability and effectiveness. Plot the boxplots from the experimental results of two real datasets see Figure 4.

Table 4. Output of the experimental results in the ECG200 dataset

	AUC	standard error
Improved_LOG	8.0417E-01	1.5302E-01
Logistic model	5.2500E-01	7.9057E-02
SVM model	6.3036E-01	1.5309E-01
KNN model	5.2857E-01	4.9229E-02
LDA model	6.4682E-01	1.2439E-01

Table 5. Output of the experimental results of the stock price data sets

	AUC	**standard error**
Improved_LOG	7.5022E-01	1.2624E-01
Logistic model	5.0000E-01	0.0000E+00
SVM model	5.0092E-01	6.5270E-03
KNN model	5.2569E-01	7.7897E-02
LDA model	4.8445E-01	5.5387E-02

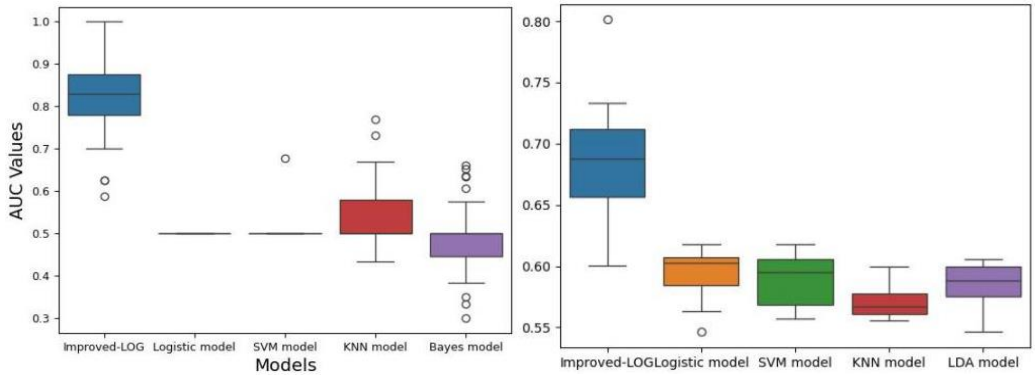

Figure 4. Boxplot of the output results of the ECG200 data set and the stock price data set

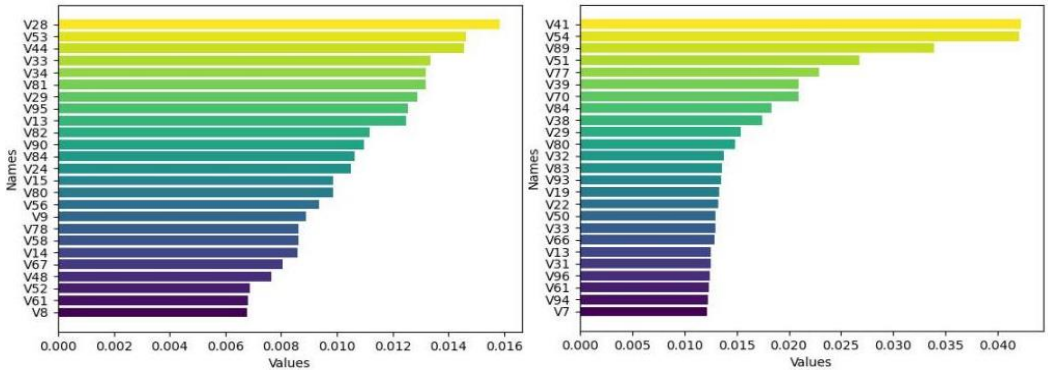

Figure 5. Feature importance output bar plot

In order to visually show the importance of different features in the prediction model, we plot the importance of features for the ECG200 dataset and the stock price dataset respectively (see Figure 5). Each bar in each bar represents a feature whose length represents the importance score of that feature in the model. For the ECG200 dataset, the bars show the importance of various characteristics of ECG signals, such as R peak amplitude, heart rate variability, for predicting potential cardiac problems in patients. In the bar chart of the stock price data set, we can see the importance of characteristics such as historical price, trading volume, and volatility in predicting the future price trend of stocks. Through these bar plots, we can intuitively identify the features that have the most influence on the model prediction results, which helps us to understand the predictive logic of the model.

The results indicate that the proposed improved logistic regression model has high predictive accuracy and low variance on the ECG200 dataset. Through feature importance analysis, it was found that the probability features provided by the logistic regression sub-models have a significant impact on the final prediction, confirming the effectiveness of model fusion. Similar to the ECG200 dataset, we resampled the stock price dataset and trained 10 logistic regression sub-models. The probability outputs of the sub-models were used as input features for the classification model. The results of 50 simulation experiments on the stock price dataset are as follows: Although the predictive performance on the stock price dataset is slightly lower than that on the ECG200 dataset, the proposed method still demonstrates good predictive ability and stability. On the stock price dataset, the GBDT model indicates that the probability features of the logistic regression sub-models are key factors in predicting default behavior. The effectiveness of the improved logistic regression model based on resampling and model fusion was verified through real data analysis. The experimental results on the ECG200 dataset and the stock price dataset show that the method has high predictive accuracy and stability when dealing with class imbalance issues and high-dimensional data. Additionally, feature importance analysis confirmed the contribution of the model fusion strategy in enhancing the predictive power of the model.

4. Conclusion

The improved logistic regression method based on random sampling and model fusion proposed in this study shows significant advantages when dealing with common issues in binary classification tasks, such as class imbalance and high-dimensional data. Specifically, by employing dual random sampling techniques, it effectively alleviates the problem of the curse of dimensionality and balances the sample class proportions. Moreover, the introduction of a model fusion strategy also enhances the model's adaptability to complex data structures. Both simulation experiments and real data analysis results consistently indicate that the classification performance of the proposed method is significantly superior to that of traditional logistic regression models.

Future research should explore more efficient sampling techniques, such as those based on information gain or cluster analysis, to improve model efficiency and accuracy with high-dimensional, class-imbalanced data. Investigating complex model fusion strategies, including combining different learning algorithms or using deep learning frameworks, may enhance the model's predictive power and generalization. Additionally, how to effectively balance class distribution for extremely imbalanced datasets requires further study. The interpretability and computational efficiency of the model are also key areas for future research, as interpretability is vital for fields like financial risk management, and improving computational efficiency is essential for large-scale model application.

Name the method proposed in this paper as improved_LOG.

References

[1] Huang Yiping, Qiu Han. Big technology credit: a new credit risk management framework [J]. Managing World, 2021, 37 (02).

[2] Lu Yuan, Fu Xiaoyan, Wang Chenyu, et al. Study on the gene prediction model of sweet corn endosperm mutation based on logistic regression model [J / OL]. Molecular Plant breeding, 2024, 1-8.

[3] Xu Hao, Huang Huiming, Tang Qiang, et al. Logistic regression analysis of cardiovascular health risk factors among civil servants in Jiangsu Province [J]. Physical Education and Science, 2018, 39 (01): 63-71.

[4] Wang Yongxin, Zhang Dabin, Che Daqing, et al. The LDBSMOTE oversampling method for the classification of unbalanced datasets [J]. Statistics and Decision-making, 2022, 38 (18): 58-63.

[5] Liao Wenxiong, Zeng Bi, Liang Tiankai, et al. Personal credit risk assessment method for high-dimensional data [J]. Computer Engineering and Application, 2020, 56 (04): 219-224.

[6] Su Yapeng, Chen Gaoshu, Zhao Tong. SA-YOLO adaptive loss target detection algorithm under unbalanced samples [J]. Journal of the University of Chinese Academy of Sciences (Chinese and English), 2024, 41 (03): 411-426.

[7] Miao Yue, Wu Chen. Research on the credit risk assessment model based on the RF-FL-LightGBM algorithm [J]. Computer and Digital Engineering, 2024, 52 (03): 808-813.

[8] Leng Qiangkui, Sun Xuezi, Meng Xiangfu. Oversampling method for unbalanced data based on the evolution of sample potential and noise [J]. Computer Applications, 2024, 44 (08): 2466-2475.

[9] Wen Tingxin, Kong Xiangbo. Research on extreme risk early warning in financial market under unbalanced samples [J]. Computer Engineering and Application, 2020, 56 (08): 256-260.

[10] Zhou Bu, Guo Jia, Zhang Jinting. Two-way ANOVA method for high-dimensional data based on the L2 norm [J]. Science in China: Mathematics, 2020, 50 (05): 729-750.

[11] Li Zean. Feature extraction algorithm based on regularization estimation in high-dimensional data mining [J]. Journal of Hefei University of Technology (Natural Science Edition), 2012, 35 (12): 1655-1658.

[12] Zhou Gang, Guo Fuliang. Research on high-dimensional data integration learning methods based on feature selection [J]. Computer Science, 2021, 48 (S1): 250-254.

[13] Feng Xiaorong, Qu Guoqing. Feature selection of high-dimensional data based on deep learning and random forest [J]. Computer Engineering and Design, 2019, 40 (09): 2494-2501.

[14] He Qiang, Dong Zhiyong. Research on the method of predicting quarterly GDP growth rate by using Internet big data [J]. Statistical Study, 2020, 37 (12): 91-104.

[15] Yan Furong, Zhang Wen, Yuan Bao, et al. Research on the application of the improved logistic regression model based on the industry development trend in the risk warning of electricity bill recovery [C] // Power Information Professional Committee of Chinese Society of Electrical Engineering, Information and Communication Branch of State Grid Corporation of China. Proceedings of the 2022 Electric Power Industry Informatization Annual Conference. Beijing CLP PricewaterhouInformation Technology Co., LTD., 2023: 7.

[16] Zhou Xuanlang, Qiu Weigen, Zhang Lichen. A text classification method incorporating text graph convolution and ensemble learning [J]. Computer applications research, 2022, 39 (09): 2621-2625.

[17] Jia Chuanmiao, Xiao Xingyuan, Jiang Tao. Summary of granular SVM studies [J]. Journal of Tianjin University of Technology, 2024, 40 (03): 57-66.

[18] Shi Zhenhua. Research on strongly generalization linear dimension reduction algorithm in classification and regression tasks [D]. And Huazhong University of Science and Technology, 2022.

[19] Li Hua, Liu Jie. Comparative analysis of the LDA classification model with the improved classification model [J]. Journal of Jilin Normal University (Natural Science Edition), 2024, 45 (02): 60-68.

[20] Ou Guiliang, He Yulin, Zhang Manjing, et al. Risk minimization of the weighted Naive Bayes classifier [J/OL]. Computer Science, 1-20 [2024-09-30].