A Credit Scoring Model Based on Integrated Mixed Sampling and Ensemble Feature Selection: RBR_XGB

Xiaobing Lin[1], Zhe Wu[1], Jianfa Chen[1*], Lianfen Huang[2], Zhiyuan Shi[1]

[1] Department of Electronic Science and Engineering, Xiamen University, China
[2] Department of Information and Communication Engineering and Key Laboratory of Digital Fujian on IoT Communication, Architecture and Security Technology (IoTCAS), Xiamen University, China
xiaobinglin@stu.xmu.edu.cn, 23120201150249@stu.xmu.edu.cn, jfchen@xmu.edu.cn, lfhuang@xmu.edu.cn, zyshi@xmu.edu.cn

Abstract

With the rapid development of the economy, financial institutions pay more and more attention to the importance of financial credit risk. The XGBoost algorithm is often used in credit scoring. However, it should be noted that XGBoost has three disadvantages when dealing with small samples of high-dimensional imbalance: (1) the model classification results are more biased towards the majority class when the XGBoost algorithm is used in training imbalanced data, this results in reduced model accuracy. (2) XGBoost algorithm is prone to overfitting in high-dimensional data because the higher the data dimension, the sparser the samples. (3) In small datasets, it is prone to form data fragmentation, resulting in reduced model accuracy. A Credit Scoring Model Based On Integrated Mixed Sampling And Ensemble Feature Selection (RBR_XGB) is proposed on the following issues in this paper. The model first aims at the model failure and overfitting problems of XGBoost in the face of highly imbalanced small samples, and uses the improved hybrid sampling algorithm combining RUS and BSMOTE1 to balance and expand the data set. For feature redundancy problems, the RFECV_XGB algorithm is used to filter features for reducing interference features. Then, considering the strength of the distinguishing ability of different models, the validation set is used to assign weights to different models, and the weighted ensemble is used to further improve the performance of the model. The experimental results show that the classification performance of the RBR_XGB algorithm for high-dimensional imbalanced small data is higher than that of the traditional XGBoost algorithm, and it can be used for commercial use.

Keywords: Credit scoring, XGBoost, Imbalance data, High-dimensional data

1 Introduction

Credit business is the most important asset business and the main profit method of commercial banks and Internet finance companies. As it is used for more and more pre-consumption, the default risk faced by credit businesses also increases. The credit risk of the platform resulted in huge losses such as fraud or financial loss by the operator, overdue repayment by the borrower, and bankruptcy of the P2P platform [1]. Therefore, developing accurate credit scoring models has become a major focus for financial institutions to optimize profits and effectively manage risk exposures [2]. Currently, banks rate users through credit ratings and offer different services to users with different credit levels. This is an important area of research that enables financial institutions to develop lending strategies to optimize profits and to inform users that some actions may jeopardize credit scores [3]. Initially, credit scoring developed a series of rules based on the experience of financial experts and evaluated the user's credit rating according to the rules. This empirical credit scoring model uses mainly the experience of experts and includes a certain degree of subjective will [4].

In the early days, the main traditional credit card scoring mainly used statistical analysis methods such as Logistic Regression [5-7], Linear Discriminant Analysis [8], and Markov model [9]. After the continuous development of machine learning and artificial intelligence technology, many methods of machine learning and artificial intelligence have been applied to credit scoring [10-11]. At present, a single model can no longer meet the needs. For most data, the training results of the ensemble learning method will tend to towards the optimal classifier [12]. For example, Boosting and Bagging models such as XGBoost [13], CatBoost [14], RF [15] are also widely used in credit scoring models. In recent years, technologies such as deep neural networks are also gradually being applied. However, due to its high computational cost and inexplicability, it has not been widely used in the field of credit scoring [16].

However, a well-known problem with credit scoring is the data imbalance [17]. There are far fewer defaulters among users than normal customers. When the classifier is learning on such highly imbalanced data, it is easy for the classifier to favor the majority class. It leads to the failure of many classifiers and many studies have no effective results [18]. The current methods for data imbalance are mainly aimed at three aspects: (1) Balancing the dataset by changing the data distribution [19]. (2) Using evaluation indicators such as AUC (Area Under Receiver Operating Characteristic Curve) to avoid data imbalance which leads to model failure. (3) Improving the model performance by optimizing the classifier [20]. However, traditional undersampling methods only use part of the dataset and lose a lot of information. Only using the traditional oversampling method will generate a large number of similar samples, which is not very helpful for the training

*Corresponding Author: Jianfa Chen; E-mail: jfchen@xmu.edu.cn
DOI: 10.53106/160792642022092305014

of the model, and even the sample aliasing will blur the boundary of the category, making the model more difficult to distinguish.

On the other hand, the performance of the model is closely related to feature selection. Feature screening removes redundant features and saves the cost of model training. It also further improves the performance of the model. However, many studies, especially those of ensemble models [21], ignore the feature screening part of ensemble models.

In this paper, the RBR_XGB algorithm is proposed to address the model failure and overfitting problems of XGBoost in highly imbalanced small samples. The algorithm first balances and expands the dataset through an improved hybrid sampling algorithm combining RUS and BSMOTE1. Then, for the feature redundancy problem in high-dimensional data, the RFECV_XGB algorithm is used to filter features and reduce interference features. Finally, considering the distinction between different models. The algorithm uses the validation set to assign weights to different models, and integrates different XGBoost models to further improve the performance of the model. The main contributions of this paper are listed as follows:

- First, an integrated mixed sampling of RUS and BSMOTE1 is proposed to balance and expand the data sets without losing data information. It can deal with the problems of model failure and data fragmentation when the XGBoost algorithm is used for credit scoring in an imbalanced small dataset.
- Secondly, the RFECV_XGB algorithm is proposed for feature screening of high-dimensional data.It solves the problem which will consume more time and cost due to redundant features and interference features when the XGBoost model is used for credit scoring in high-dimensional data.
- Third, considering the differences between the models, the validation set is used to assign different weights to the XGBoost model to further improve the classification performance of the algorithm.

The remainder of the structure is as follows： Section 1 presents the background theory of credit card default prediction. Section 2 introduces the main methods used in this paper and the solution proposed in this paper. Section 3 mainly contains the experimental design and experimental results of this paper. Section 4 contains a summary of the content of this paper and proposes directions for further research.

2 Method

2.1 XGBoost

XGBoost (Extreme Gradient Boosting) is an extreme gradient boosting algorithm based on Boosting, which integrates weak classifiers to form a powerful classifier. The basic tree model used is the CART regression tree.

Given a dataset:

$$D = \{(x_1, y_2), (x_2, x_2), \cdots, (x_3, y_3)\}, \quad \textbf{(1)}$$

for each point there is a $x_n \in R^m$, and $y_n \in \{0,1\}$， then, the model is represented as:

$$\widetilde{y_p} = \Sigma_{m=1}^{m} f_m(x_n), f_m \in F, \quad \textbf{(2)}$$

where m is the number of decision trees, x_n is the nth sample input, it represents the prediction result of the pth tree, and F is the set of all decision trees.

The objective function and loss function are optimized as:

$$T_t = \Sigma_{n=1}^{k} loss\left(y - \widetilde{y_p}(t-1)_n - f_t(x_n)\right) + \Omega(f_t), \quad \textbf{(3)}$$

where T_t represents the objective function of the t-th iteration, $loss$ represents the loss function, $\widetilde{y_p}(t-1)_n$ is the prediction result of the previous iteration, $f_t(x_n)$ represents the newly added item, $\Omega(f_t)$ represents as regular term, the formula can be simplified to:

$$T_t' = \Sigma_{n=1}^{k}\left[j_n f_n(x_n) + \frac{1}{2} o_n \int_t^2 (x_n) + \Omega(f_t)\right], \quad \textbf{(4)}$$

where j_n is the first derivative, o_n is the second derivative.

The optimal value:

$$Q_j^* = -\frac{p_n}{s_n+\lambda}\ , \quad \textbf{(5)}$$

$$T_t' = -\frac{1}{2}\Sigma_{n=1}^{k}\frac{P_j^2}{s_n+\lambda} + \lambda T, \quad \textbf{(6)}$$

where Q_j^* represents the optimal solution, and T_t' is the value of the objective function.

The XGBoost algorithm has an excellent performance in classification. Therefore, the XGBoost algorithm is widely used in credit scoring.

However, it should be noted that XGBoost has three disadvantages when dealing with small samples of high-dimensional imbalance: (1) the model classification results are more biased towards the majority class when the XGBoost algorithm is used in training imbalanced data, this results in a decrease of model accuracy. (2) XGBoost algorithm is prone to overfitting in high-dimensional data because the samples become more and more sparse as the data dimension increases (3) XGBoost is easy to form data fragmentation when it is used in small data sets because XGBoost uses the divide and conquer idea based on Cart decision tree to divide a problem into many problems. In the case of a few minority class samples, the divided subspace contains little information, and some cross-space features cannot be mined.

2.2 RBR_XGB

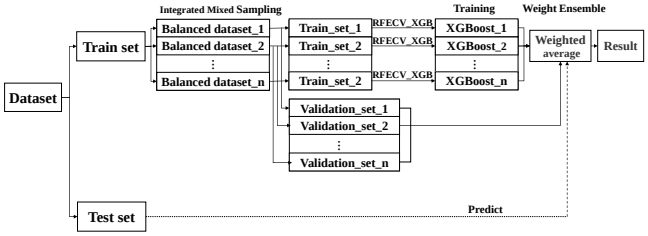

Figure 1. RBR_XGB

RBR_XGB is proposed to solve the problems of imbalanced data distribution, dimensionality disaster, and small datasets. The model proposed for this study is shown in Figure 1: First, an integrated mixed sampling based on

improved RUS and B_SMOTE1 is used to balance and expand the high-dimensional imbalanced dataset to form n expanded balanced datasets. Then, the RFECV algorithm is used to reduce the dimensionality of the balanced datasets respectively. After that, the effective subsets with reduced dimensionality are put into the training of the XGBoost model, and finally, the n XGBoost models are weighted and aggregated by the method of bagging to further improve the classification ability of the algorithm.

2.3 An Integrated Mixed Sampling Based on RUS and B_SMOTE1

It is a practical method to use RUS (RandomUnder-Sampling) to balance the data samples. However, just randomly undersampling the dataset loses a lot of information, resulting in reduced model accuracy. On the other hand, after undersampling, the number of samples in the sub-data set formed by merging the majority class samples and the minority class samples will be too small, which may easily lead to overfitting. In this study, the method in Easyensemble [22] was introduced to improve the RUS algorithm, and the BSMOTE1 (Borderline-SOMTE1) [23] method was used to expand the data set, and an integrated mixed sampling based on improved RUS and BSMOTE1 was proposed to solve the above problems.

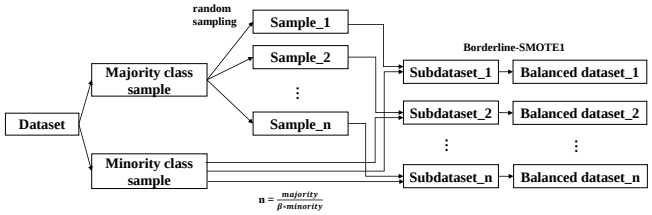

Figure 2. An integrated mixed sampling based on RUS and B_SMOTE1

As is shown in Figure 2, the specific steps based on improving the mixed sampling of RUS and BSMOTE1 are listed as follows:

(1) Calculate the sampling number of majority class and minority class samples, where M annumberse the number of majority class and minority class samples in the sample set D, Q is the desired number of balanced subdataset samples, the number of random sampling without replacement is as follows:

$$P_i = \frac{Q}{2}. \tag{7}$$

The radio of majority class samples and minority class samples after RUS random sampling without replacement is as follows:

$$\beta = \frac{P_i}{N}\ . \tag{8}$$

Adjusting the value of β can adjust the size of the balanced data subset, which can well solve the problem of too small data volume after sampling.

RUS undersampling times can be expressed as:

$$n = \frac{M}{\beta * N}\ . \tag{9}$$

(2) Using RUS to extract n data subsets M_{new_i} from D without replacement, and then form n imbalanced data subsets D_{new_i} with minority class samples N respectively.

(3) Perform BSOMTE1 interpolation on D_{new_i} to generate artificial samples to form balanced dataset D_i.

2.4 REFCV_XGB

In the existing feature selection methods, the RFECV algorithm (Recursive Feature Elimination Cross Validation) filters features according to their feature importance in the model, removes one or more features with the least feature importance in each iteration, and uses cross validation method to compute feature subset scores. After multiple iterations, the feature subset with the highest score is selected, which is a simple and effective wrapping feature selection algorithm. As such, RFECV is widely used in feature screening.

Since the XGBoost algorithm can use hyperparameters to randomly select features when building a decision tree to avoid overfitting, it will eventually lead to too many redundant features being selected, which will reduce the accuracy of a single decision tree. Therefore, feature reduction is performed by combining the RFECV algorithm with XGBoost (RFECV_XGB).

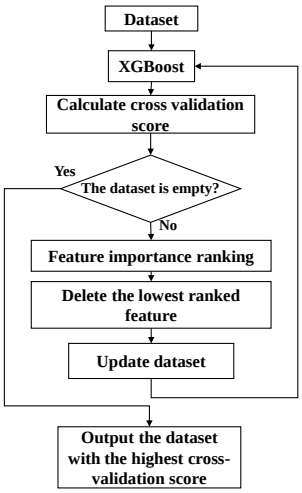

Figure 3. RFECV_XGB

As is shown in Figure 3, the content of RFECV_XGB is as follows:

a) Initialization:

Training sample matrix as:

$$X = [x_1, x_2 \ldots x_n]. \tag{10}$$

Category label as:

$$\mathrm{Y} = [y_1, y_2]. \tag{11}$$

Initial feature set as:

$$S = [f_1, f_2 \ldots f_n]. \tag{12}$$

Feature sorted collection:$R = []$;

b) Looping for feature elimination until $S = [\ ]$;

(a) Extracting the current sample set:

$$X_0 = X(:, S). \tag{13}$$

(b) Training the classifier:

$$\alpha = XGB(X_0, Y, \lambda), \tag{14}$$

where λ is the number of RFECV cross-checks;

(c) Calculating the cross-check score:

$$Score = \frac{1}{\lambda}\sum_i^{\lambda} auc_i. \tag{15}$$

(d) Calculate the feature importance and get the feature or feature set with the smallest ranking coefficient $s(f)$;

(e) Update feature sorted set:

$$R = [s(f), R], \tag{16}$$

where $s(f)$is the feature or feature set to be deleted in this iteration;

(f) Remove $s(f)$ in S:

$$S = S/s(f). \tag{17}$$

(g) Output the subset of features with the largest score $s\_max(f)$.

2.5 Weighted Ensemble XGBoost

The n XGBoost models are obtained from the subset of data filtered by the RFECV_XGB feature after training with the XGBoost classifier. To consider the difference in classification performance between different XGBoost models, the validation set is used to judge the classification accuracy of the decision tree and assign different weights to different XGBoost models. This enhances the weight of models with excellent classification performance and reduces the proportion of models with poor classification performance, which greatly improves the accuracy of ensemble models.

The steps of Weighted Ensemble XGBoost is as follows:

(1) In the balanced data subset, 10% of the data is used as the validation set, and 90% of the data is used as the training set.

(2) The training set is put into the XGBoost classifier for training.

(3) Using the validation set to get the AUC value of the model.

(4) Calculating the weight of different XGBoost, models the calculation formula as:

$$W_i = \frac{auc_i}{\sum_{k=1}^{n} auc_k}\ . \tag{18}$$

(5) The final weighted output as:

$$P = \sum_{i=1}^{n} W_i P_i, \tag{19}$$

where P is the output probability of the xgboost ensemble model, and P_i is the output probability of the xgboost sub-model.

3 Experiments

3.1 Data

The data in this article comes from a well-known domestic financial institution, with a total of 16,592 pieces of data, including 1,393 features including one label. A label of 0 represents a normal user, and a label of 1 represents a default user. The number of positive samples is 16281, and the number of negative samples is 311. The positive and negative sample distribution is as follows Figure 4:

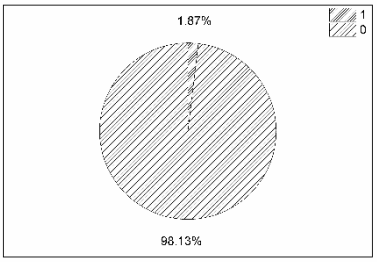

Figure 4. Positive and negative sample distribution

The ratio of positive and negative samples is 52:1, this means the dataset is an obvious high-dimensional imbalanced small data set.

3.2 Data Preprocessing

The dataset has a total of 16592 samples, including 1393 features. Data preprocessing consists of the following 5 parts:

First, features with a high missing rate were screened, and 803 features with a missing rate greater than 60% were removed.

Second, the missing values in data are filled with 0.

Third, mapping the category features that contain quantitative relationships, such as A:1, B:2, C:3, and performing mean encoding on the remaining category features that do not contain obvious quantitative relationships.

Fourth, according to the Pearson correlation coefficient, 389 features with feature correlations greater than 0.975 in the continuous variables are removed, and the remaining 200 features.

Fifth, normalize the remaining features and constrain their range to be between 0 and 1 to prevent some features from exceeding other features and affecting the training of the model.

3.3 Evaluation Indicators

The credit score data is a typical binary classification problem, so evaluation indicators such as Recall, AUC, and KS (Kolmogorov-Smirnov) are used for evaluation in this paper.

3.3.1 Recall

Recall represents the proportion of positive samples in the sample that are correctly predicted by the model. The higher the recall rate, the more positive samples the model can find,

but more negative samples may be misjudged as positive samples. Hence, Recall can be expressed as:

$$Recall = \frac{TP}{TP+FN} \ , \quad \textbf{(20)}$$

where TP (True Positive) and FN (False Negative) are come from Confusion Matrix. Confusion Matrix is an essential tool for visualising the prediction outcome. The Confusion Matrix are shown in Table 1:

Table 1. Confusion matrix

	Predicted Positive	Predicted Negative
Positive	TP	FN
Negative	FP	TN

Where TP is the number of positive samples who are identified as positive samples by model. TN is the number of positive samples who are identified as negative samples by model. FN is the number of negative samples who are identified as positive samples by model and FP is the number of negative samples who are identified as negative samples by model.

3.3.2 AUC

AUC is the area under the ROC (Receiver Operating Characteristic) curve, where the abscissa of the ROC curve is FPR (False Positive Rate), and the ordinate is TPR (True Positive Rate).

$$TPR = \frac{TP}{TP+FN} \ , \quad \textbf{(21)}$$

$$FPR = \frac{FP}{FP+FN} \ . \quad \textbf{(22)}$$

The AUC value ranges from 0 to 1, and the larger the value, the better the classifier performance. AUC can be used to judge the prediction situation of the classifier for imbalanced data, and avoid the situation that the classifier tends to the majority class and the model evaluation index fails.

3.3.3 KS (Kolmogorov-Smirnov)

KS (Kolmogorov-Smirnov) statistic is usually used to measure the performance of the risk control model in the actual financial risk control business. The larger the value of KS, the greater the ability of the model to distinguish between positive and negative samples. The statistic of the KS value is expressed as follows:

$$KS = Max|TPR(th) - FPR(th)|, \quad \textbf{(23)}$$

where th is the threshold used to distinguish positive and negative samples. It can be seen from the above formula that the KS value represents the maximum difference between TPR and FPR when different thresholds are used to distinguish positive and negative samples. It means that the model can find more positive samples with less misjudgment cost when the current threshold is selected as the distinguishing point.

3.4 Experimental Design

In the experiment, β which is the ratio of the majority class to the minority class sample is set to 1.5, and the number of the balanced dataset is set to 34. The parameters of the XGBoost model after Bayesian optimization are shown in the following Table 2.

Table 2. XGBoost hyperparameters

Parameters	Value
Learning rate	0.1
colsample_btree	0.72
gamma	0.92
max_depth	3
min_child_weight	3.65
n_estimators	57
reg_alpha	0.0
reg_lambda	0.87
subsample	0.84

In order to verify the performance of the RBR_XGB algorithm proposed in this paper. RBR_XGB, RB_XGB which combine the integrated mixed sampling based on RUS and B_SMOTE1with XGBoost, RFECV_XGB, and XGBoost are used to classify the data set, and compare their respective recall rates, AUC, and KS values to verify the model classification performance.

3.5 Experimental Results

Since financial institutions are more sensitive to defaulting users, the model pays more attention to finding more potential defaulting users. Therefore, Recall, AUC, and KS values are used as evaluation indicators. The experiments compare the evaluation indicators of the above four algorithms. The final experimental results are shown in Table 3:

Table 3. Evaluation indicators

Algorithm	AUC	Recall	KS
XGB	0.73729	0.701863	0.403726
RFECV_XGB	0.86247	0.767684	0.535367
RB_XGB	0.867496	0.770401	0.540802
RBR_XGB	**0.882975**	**0.805706**	**0.611411**

The line charts drawn by various indicators of the four algorithms and the ROC Curve are shown in Figure 5 and Figure 6 respectively:

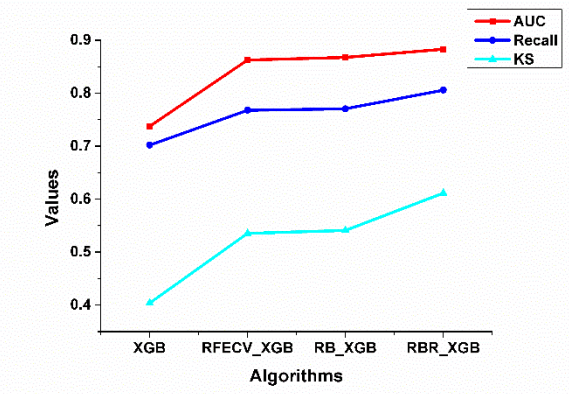

Figure 5. Evaluation metrics

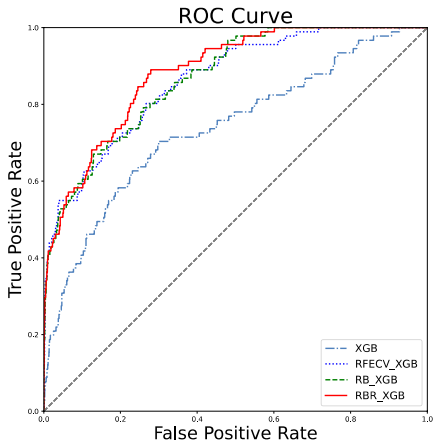

Figure 6. ROC curves

In Figure 5, we find that after the sample is processed by integrated mixed sampling, the prediction of the model has been greatly improved. The AUC value has been increased from 0.73729 to 0.86247, Recall has been increased from 0.701863 to 0.767684, and the KS value has changed from 0.403726 to 0.535367.

As well, the AUC value, Recall, and KS value of the model also has been greatly improved after the sample is processed by RFE feature screening. The AUC value has been increased from 0.73729 to 0.867496, Recall has been increased from 0.701863 to 0.770401, and the KS value has changed from 0.403726 to 0.540802. It can be seen that data balance and expansion of high-dimensional imbalanced data and feature screening can improve model performance.

RBR_XGB is higher than RB_XGB and RFECV_XGB algorithms on all evaluation indicators. The AUC are improved by 1.5% and 2.7% respectively. Recall are improved by 3.8% and 3.5%. Especially, the algorithm has the best effect on improving the KS value. Compared with the other two algorithms, it is increased by 7.1% and 7.6% respectively. The results indicate that the combination of the two methods can further improve the performance of the model compared to only balancing and expanding imbalanced data, or only filtering feature data.

Combining with Figure 6, it can also be seen that most of the ROC curves of the RBR_XGB algorithm are located at the upper left of other curves, and it means that its classification effect is higher than other algorithms.

In order to verify the feature screening effect of the method used in this experiment, we also compare the number of features of four algorithms. The number of features of the four algorithms as shown in Table 4:

Table 4. Characteristics of the number

Algorithm	Numbers
XGB	200
RFECV_XGB	43
RB_XGB	200
RBR_XGB	46 (avg)

Where the number of features of the RBR_XGB algorithm is the average of the RFECV screening results.

As is shown in Table 4, The RB_XGB algorithm can balance and expand the data set to a certain extent, but it does not filter the number of features in the data set and cannot achieve the purpose of feature reduction. The RFECV_XGB and RBR_XGB algorithms can both remove redundant features in the dataset to a certain extent.

In order to analyzes the model feasibility of the RBR_XGB algorithm, the model evaluation table is shown as follows Table 5:

Table 5. Model evaluation

Threshold	Cumsum_ good	Cumsum_ bad	KS
0.00	0.00%	0.00%	0.000
0.05	4.22%	48.35%	0.441
0.10	9.13%	58.24%	0.491
0.15	14.02%	69.23%	0.552
0.20	19.03%	73.63%	0.546
0.25	23.96%	82.42%	0.585
0.29	**27.87%**	**89.01%**	**0.611**
0.30	28.93%	89.01%	0.601
0.35	34.01%	89.01%	0.550
0.40	39.06%	91.21%	0.521
0.45	44.10%	94.51%	0.504
0.50	49.17%	95.60%	0.464
0.55	54.23%	97.80%	0.436
0.60	59.30%	98.90%	0.396
0.65	64.37%	100.00%	0.356
0.70	69.45%	100.00%	0.306
0.75	74.54%	100.00%	0.255
0.80	79.64%	100.00%	0.204
0.85	84.74%	100.00%	0.153
0.90	89.83%	100.00%	0.102
0.95	94.93%	100.00%	0.051
1.00	100.00%	100.00%	0.000

Where $Threshold$ is the threshold for judging default users and normal users, $Cumsum\_good$ is the cumulative normal user ratio, $Cumsum\_bad$ is the cumulative default user ratio, and KS is the KS value when divided by this threshold. The bolded line in the figure is the row with the highest KS.

As can be seen from Table 5, when users are segmented with a threshold of 0.29, the model can capture 89.01% of default users, and the users who are wrongly divided account for only 27.87% of all normal users. The model can find most default users with less misjudgment cost. It can be shown that the RFR_XGB model can be used in actual business.

In summary, the improved algorithm is much higher than the original XGBoost model. Compared with the RFECV_XGB algorithm and the RB_XGB algorithm, the RBR_XGB algorithm improves the overall classification effect in terms of feature reduction and data balance and expansion and can be used in actual business.

4 Conclusion

Based on the characteristics of the finance credit business data, the RBR_XGB algorithm is proposed to solve the problem of low model accuracy when the XGBoost algorithm is used to classify and predict high-dimensional imbalanced small samples in this paper. First, the integrated mixed sampling based on RUS and B_SMOTE1 is used to balance

and expand the data set to solve the model failure and overfitting problems of XGBoost in the face of highly imbalanced small samples; second, the RFECV_XGB algorithm is used to filter features to reduce interference and solve the problem of feature redundancy in high-dimensional data; finally, considering the strength of the distinguishing ability of different models, the validation set is used to give weights to the models, and weighted integration can further improve the performance of the model. Compared with the original xgboost algorithm, RBR_XGB algorithm has improved in AUC, KS and recall, especially in KS, which is improved by 0.2. The model also can capture 89.01% of default users, and the users who are wrongly divided account for only 27.87% of all normal users when users are segmented with a threshold of 0.29. The results show that the prediction performance of the RBR_XGB algorithm in dealing with high-dimensional imbalanced small samples and can be used in business. However, since the RBR_XGB algorithm uses the RFECV algorithm when filtering data, it takes a long time to filter features in high dimensions. In the next step, the filter method can be used to perform feature screening without affecting the feature screening to reduce the time cost of the model.

References

[1] J. Zhao, Internet Finance and its risk prevention and control, *Tax and Economy*, Vol. 23, No. 2336, pp. 52-63, January, 2018.

[2] C. Jiang, Z. Wang, H. Zhao, A prediction-driven mixture cure model and its application in credit scoring, *European Journal of Operational Research*, Vol. 277, No. 1, pp. 20-31, August, 2019.

[3] F. Butaru, Q. Chen, B. Clark, S. Das, A. W. Lo, A. Siddique, Risk and risk management in the credit card industry, *Journal of Banking Finance*, Vol. 72, No. 1, pp. 218-239, November, 2016.

[4] D. Huang, J. Zhou, H. Wang, RFMS method for credit scoring based on bank card transaction data, *Statistica Sinica*, Vol. 28, No. 4, pp. 2903-2919, October, 2018.

[5] S. Kalaycı, M. Kamasak, S. Arslan, Credit risk analysis using machine learning algorithms, *2018 26th Signal Processing and Communications Applications Conference (SIU) IEEE*, Izmir, Turkey, 2018, pp. 1-4.

[6] E. C. Silva, I. Lopes, A. Correia, S. Faria, A logistic regression model for consumer default risk, *Journal of Applied Statistics*, Vol. 47, No. 13-15, pp. 2879-2894, 2020.

[7] E. Dumitrescu, S. Hue, C. Hurlin, S. Tokpavi, Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects, *European Journal of Operational Research*, Vol. 297, No. 3, pp. 1178-1192, March, 2022.

[8] Y. Guo, J. He, L. Xu, W. Liu, A novel multi-objective particle swarm optimization for comprehensible credit scoring, *Soft Computing*, Vol. 23, No. 18, pp. 9009-9023, September, 2019.

[9] D. E. Régis, R. Artes, Using multi-state markov models to identify credit card risk, *Production*, Vol. 26, No. 2, pp. 330-344, June, 2016.

[10] M. Leo, S. Sharma, K. Maddulety, Machine learning in banking risk management: A literature review, *Risks*, Vol. 7, No. 1, Article No. 29, March, 2019.

[11] B. W. Chi, C. C. Hsu, A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model, *Expert Systems With Applications*, Vol. 39, No. 3, pp. 2650-2661, February, 2012.

[12] Y. Li, W. Chen, A comparative performance assessment of ensemble learning for credit scoring, *Mathematics*, Vol. 8, No. 10, Article No. 1756, October, 2020.

[13] L. Munkhdalai, T. Munkhdalai, O. E. Namsrai, J. Y. Lee, K. H. Ryu, An empirical comparison of machine-learning methods on bank client credit assessments, *Sustainability*, Vol. 11, No. 3, Article No. 699, February, 2019.

[14] Y. Xia, L. He, Y. Li, N. Liu, Y. Ding, Predicting loan default in peer-to-peer lending using narrative data, *Journal of Forecasting*, Vol. 39, No. 2, pp. 260-280, March, 2020.

[15] J. Cho, J. Joo, I. Han, The Prediction of Export Credit Guarantee Accident using Machine Learning, *Journal of Intelligence and Information Systems*, Vol. 27, No. 1, pp. 83-102, March, 2021.

[16] B. R. Gunnarsson, S. V. Broucke, B. Baesens, M. Oskarsdottir, W. Lemahieu, Deep learning for credit scoring: Do or don't?, *European Journal of Operational Research*, Vol. 295, No. 1, pp. 292-305, November, 2021.

[17] H. He, W. Zhang, S. Zhang, A novel ensemble method for credit scoring: Adaption of different imbalance ratios, *Expert Systems With Applications*, Vol. 98, pp. 105-117, May, 2018.

[18] O. J. Leong, M. Jayabalan, A comparative study on credit card default risk predictive model, *Journal of Computational and Theoretical Nanoscience*, Vol. 16, No. 8, pp. 3591-3595, August, 2019.

[19] Y. Liu, Y. Cheung, Y. Tang, Self-Adaptive Multiprototype-Based Competitive Learning Approach: A k-Means-Type Algorithm for Imbalanced Data Clustering, *IEEE Transactions on Cybernetics*, Vol. 51, No. 3, pp. 1598-1612, March, 2021.

[20] Y. Chang, K. Chang, H. Chu, L. Tong, Establishing decision tree-based short-term default credit risk assessment models, *Communications in Statistics-Theory and Methods*, Vol. 45, No. 23, pp. 6803-6815, 2016.

[21] T. Zhang, G. Chi, A heterogeneous ensemble credit scoring model based on adaptive classifier selection: An application on imbalanced data, *International Journal of Finance & Economics*, Vol. 26, No. 3, pp. 4372-4385, July, 2021.

[22] Y. Liu, EasyEnsemble and Feature Selection for Imbalance Data Sets, *2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, Shanghai, China, 2009, pp. 517-520.

[23] H. Han, W. Y. Wang, B. H. Mao, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, *International conference on intelligent computing*, Hefei, China, 2005, pp. 878-887.

Biographies

Xiaobing Lin received his B.S. degree from Xiamen University, Xiamen, China in 2020. He is currently pursuing the M.S. degree in circuits and systems with Xiamen University. His research interests include imbalanced credit scoring model and data mining.

Zhe Wu received his B.S. degree from Xiamen University, Xiamen, China in 2020. He is currently pursuing the M.S. degree in electronic engineering with Xiamen University. His research interests include data analysis and imbanlanced credit scoring model.

Jianfa Chen received his M.S. degree in System Engineering in 2007 from Xiamen University. He is an engineer in the School of electronic science and engineering, Xiamen University, Xiamen, Fujian, China. His current research interests include computer networking, artificial intelligence and system simulation.

Lianfen Huang received her B.S. degree in Radio Physics in 1984 and PhD in Communication Engineering in 2008 from Xiamen University. She was a visiting scholar in Tsinghua University in 1997. She is a professor in the Department of Communication Engineering, Xiamen University, Xiamen, Fujian, China. Her current research interests include wireless communication, wireless network and signal process.

Zhiyuan Shi received his B.S. degree in radio physics in 1984, M.S. degree in radio electronics in 1991 from Xiamen University. He is a Professor in the Department of electronic engineering, Xiamen University, Xiamen, Fujian, China. His current research interests include wireless communication, circuit and system.