Loan Prepayment Prediction Based on SVM-RFE and XGBoost Models

1st Qi Mao[1], 2nd Gang Liu[2], 3rd Zhiyu Chen[3], 4th Jianwei Guo[4*], 5th Peng Liu[5*]

[1]2201903016@stu.ccut.edu.cn

[2]lg@ccut.edu.cn

[3]chenzhiyu@stu.ccut.edu.cn

[4*]Corresponding author: guojianwei@ccut.edu.cn

[5*]Corresponding author: 1864460328@wo.com.cn

[1]School of Computer Science and Engineering Changchun University of Technology Changchun, China

[2]School of Computer Science and Engineering Changchun University of Technology Changchun, China

[3]School of Computer Science and Engineering Changchun University of Technology Changchun, China

[4]School of Computer Science and Engineering Changchun University of Technology Changchun, China

[5]Product Department Jilin Heshun Hengtong Technology Co. Changchun, China

Abstract—The problem of large dimensionality of loan data and unbalanced data samples severely affects the classification, and the article proposes a support vector machine for feature recursive elimination and XGBoost for a loan early repayment prediction. Firstly, the combination of Pearson index and SVM-RFE in the data feature layer can reduce the dimension of data, find the best feature subset including more information,and then find more information. Secondly, the weighted cross-entropy loss function is introduced into the XGBoost algorithmto solve the problem of data imbalance. Finally, a comparative experiment is carried out on the LendingClub data set to confirm the effectiveness of the proposed model in predicting and analyzing the personalbehavior of loan prepayment.

Keywords-loan prepayment; feature selection; SVM-RFE; XGBoost; weighted cross-entropy loss function

1 INTRODUCTION

With the rise of online lending, online lending based on borrowers' credit scores is an innovative means to achieve a reasonable asset allocation of funds directly between lenders and borrowers, however, due to borrowers' default and early repayment behaviors can negatively affect the interests of investors as well as the lending platform, which in turn reduces the platform's operating income [1]. In addition, early repayment behavior can also lead to unstable future cash flows of fund products, thus affecting the actual yield and pricing level, and thus the study of early repayment risk is of great significance..

ICIDC 2022, June 17-19, Qingdao, People's Republic of China

DOI 10.4108/eai.17-6-2022.2322765

Early repayment of a loan refers to the behavior of a borrower who pays off the principal in full before the due date of repayment, resulting in lower estimated interest income. Currently, domestic and foreign scholars have studied early repayment behavior. Li Z [2] and others applied multiple logistic regression to predict and analyze the probability of prepayment and default, confirmed that there are different influencing factors between loan repayment and loan default, and paid attention to the need to control the situation of loan prepayment. Liang and Lin [3] developed a two-stage model for early repayment risk. The first stage divides borrowers into several groups using a random forest algorithm, and the second stage constructs a proportional risk model for each group to predict their early repayment time. It was found that the two-stage model possessed a higher accuracy rate than the single-stage model. Chehong Zhu [4] developed and designed a deep learning entity model of mortgage risk, and found that there is a highly optimal control relationship between personal behavior of early repayment of loans and loan characteristics and macroeconomic variables. Mortazavi A [5] removed irrelevant features by searching the optimal feature subset to produce the minimum error on the original dataset. Schapiro et al. [6] proposed the AdaBoost algorithm, which allows the classifier to improve its focus on a small number of classes of samples by focusing on the samples that are misclassified after each round and giving them higher weights. Zhao C et al. [7] used an integrated learning approach to identify social network spam and dynamically adjusted the weights of the base classifier prediction results by setting different misclassification costs for different classifiers, thus effectively improving the classification of unbalanced data sets.

In the above literature, the Filter selection method has fast computation speed and considers the relationship between features, but the selected features do not consider whether the features fit the model, and the SVM-RFE in the Wrapper sign selection method can select a subset of features suitable for this model, but the SVM-RFE selection method does not consider the correlation between features. Among the methods for predicting early loan repayment, the financial model-based prediction methods have strong explanatory power supported by financial theory but are not suitable for large-scale data prediction, while the machine learning-based methods focus only on prediction accuracy and rarely consider the problem that early loan repayment is an unbalanced data set.

Based on this paper for the prepayment of loans ask the following two contributions to this paper, in terms of feature selection, by combining the Pearson correlation coefficient in filtered feature selection with the SVM-RFE in the encapsulated feature selection method to select the optimal feature subset, the selected features have minimal correlation with each other while still having high predictive power. In predictive analysis, the XGBoost loss function is modified by fusing the weighted cross-entropy loss function with XGBoost to make the model focus more on a small number of classes of samples, thus improving the predictive power of the model.

2 LOAN PREPAYMENT PRE*DICTION MODEL*

The proposed SVM-RFE-XGBoost as a loan prepayment model is shown in Fig.1. It is divided into four main modules namely data preprocessing, feature selection, classification prediction module, and model evaluation. The data pre-processing module performs preliminary processing of the data by removing missing values, irrelevant variables, and duplicate data as

well as the conversion of the data format to make the data in a way that initially filters the data characteristics. The feature selection module further selects the input features by combining the Pearson coefficients with the SVM-RFE feature selection method, so that the model accuracy is optimal while the redundancy between the selected features is small. The classification prediction module optimizes XGBoost with a weighted cross-entropy loss function (XGBoost_WCE) so that the model can have better classification results on unbalanced data sets.

3 ALGORITHM DESCRIPTION

3.1 Pearson correlation coefficient

The Pearson correlation coefficient [8] is a linear correlation coefficient that captures the degree of linear correlation between two variables. For any two random variable c $X = x_1, x_2, \dots, x_n, Y = y_1, y_2, \dots, y_n$.The definition is shown as Eq. (1).

$$P(X,Y) = \frac{\sum_i (x_i - \hat{x}_i)(y_i - \hat{y}_i)}{\sqrt{\sum_i (x_i - \hat{x}_i)^2}\sqrt{\sum_i (y_i - \hat{y}_i)^2}} \tag{1}$$

where $\hat{x}, \hat{y}_i$are the mean values of X, Y, respectively $P(X,Y)$ takes values between -1 and 1. When $P(X,Y)$ is 0, it means the two variables are linearly uncorrelated, when $P(X,Y)$ is -1 to 1, it means the two variables are completely correlated.

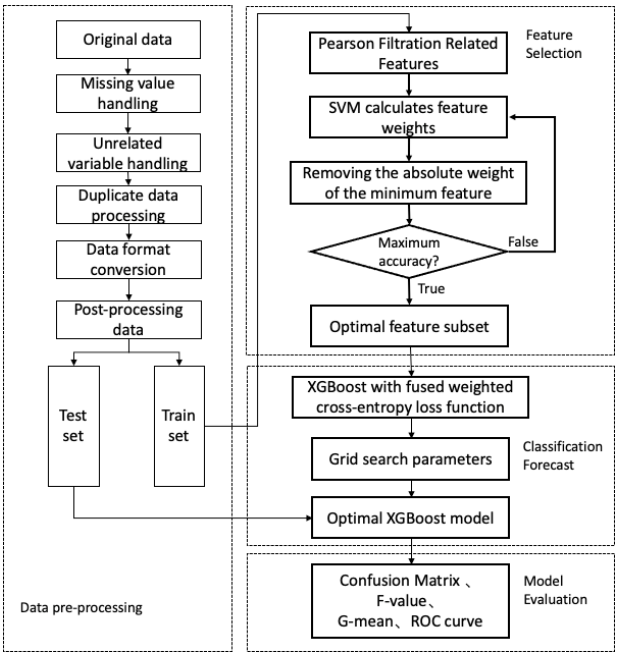

Fig. 1. Loan prepayment prediction model

3.2 SVM-RFE feature selection algorithm

Guyon et al. [9] for a comprehensive examination of SVM, based on which the SVM-RFE feature selection algorithm for ranking the features is proposed. The core idea of SVM is to build a decision surface [10]. The definition is shown as Eq. (2).

$$\omega x + b = 0 \tag{2}$$

to maximize the classification interval. In the nonlinear case, slack variables are introduced and the objective function can be expressed as Eq. (3).

$$\begin{cases} \min J = \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{m}\xi_i \\ s.t\ y_i(\omega x_i + b) \geq 1 - \xi_i, \xi_i\ > 0, i = 1, \dots, m \end{cases} \tag{3}$$

If the first feature is removed, the effect on the objective function according to the Taylor expansion is shown as Eq. (4).

$$\Delta J(i) = \frac{\partial J}{\partial \omega_i}(\Delta\omega_i) + \frac{\partial^2 f}{\partial \omega_i^2}(\omega_i)^2 \tag{4}$$

Then for the optimal solution of the objective function, only the second order can be considered and the expression is shown as Eq. (5).

$$\Delta J(i) = (\omega_i)^2 \tag{5}$$

In SVM-RFE, the importance of a feature is expressed as ω^2, and the feature with the lowest value is removed from the current feature set Feature_current and put into the feature sorting queue Feature_list. The above process is repeated until the current feature set Feature_current is empty, and the queue Feature_list stores the feature sorting. The features in the front and the first to be removed from Feature_current are generally noisy and irrelevant features, the features after them have a strong differentiation ability, and the features at the end of the queue have the strongest differentiation ability for the category. The SVM-RFE algorithm is executed as follows.

Algorithm 1 SVM-RFE

Require: training data X (n samples * m features)
Ensure: feature rank list Feature_list
Feature_list = ∅
Feature_current = all m features

While | Feature_current | > 0 **do**
construct SVM model based on Feature_current
rank the features by ω_i^2 in a descending order
move the bottom ranked feature in Feature_current into Feature_list
end while
return Feature_list

3.3 Weighted cross-entropy loss

Cross-entropy [11] is a concept in Shannon's theory that can be used to measure the similarity of two probability distributions, cross-entropy as a loss function can be used to measure the difference between the true sample and the predicted outcome, the expression of binary cross-entropy is shown as Eq. (6)

$$L_{CE}(y, \hat{y}) = -(ylog(\hat{y}) + (1-y)\log(1-\hat{y})) \tag{6}$$

where y is the true sample label and $\hat{y}$ is the predicted probability.

The premise of the cross-entropy loss function is that the distribution between categories is balanced, and when faced with unbalanced samples, the imbalance between categories leads to a larger proportion of losses occupied by large categories, which causes the model to ignore the learning of small category samples, resulting in the cross-entropy loss function is then no longer applicable. Based on this, a weighted loss function [12] is proposed to make it more applicable to the unbalanced sample set. The mathematical expression of the binary classification weighted loss function is shown as Eq. (7)

$$L_{WCE}(y, \hat{y}) = -(\beta * ylog(\hat{y}) + (1-y)\log(1-\hat{y})) \tag{7}$$

where β is the weighting coefficient of positive samples. The setting of β needs to be decided based on the percentage of samples, and the cost of the loss function of minority class samples is increased by setting the parameter β.

3.4 XGBoost algorithm

XGBoost was explicitly proposed by Chen in 2016 [13] and is one of the boosting algorithms. The idea of boosting the algorithm is to integrate several weak classifiers to produce a strong classifier. XGBoost is a boosting tree model, so it integrates several tree models to produce a very strong classifier. The XGBoost algorithm performs a second-order Taylor expansion of the loss function to take advantage of the first-order derivative and the second-order derivative in the optimization, based on the idea that for a given data set $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)\}, i = 1, 2, \dots,$ n Train T CART trees $F = f_1(x), f_2(x), \dots, f_t(x), \dots, f_T(x)$,where $f_t(x)$ represents the prediction result of sample x_i after the t CART tree, and the prediction result of T CART trees is denoted as $\hat{y}_i = \sum_{t=1}^{T} f_t(x), f_t \in F$ .The optimization objective and loss function of XGBoost are expressed as Eq. (8)

$$L(t) = \sum_{i=1}^{n} I(y, \hat{y}) + \Omega(f_t(x)) \tag{8}$$

where $\Omega(f_t(x) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$, T is the tth tree of the tree, ω_j is the weight of the jth leaf node of the tree, and γ and λ are the adjustment coefficients. The first part $\sum_{i=1}^{n} I(y, \hat{y})$ is the loss function of the model, and the second part $\sum_{i=1}^{K} \Omega(f_t(x)$ is the regularization term of leaf node weights and tree depth added to control the complexity of the model. The smaller the objective function is, the closer the model prediction is to the true value, and it also serves to control the model complexity and enhance the model robustness.

In this paper, $\sum_{i=1}^{n} I(y, \hat{y})$ part is changed to a weighted cross-entropy loss letter to form, XGBoost_WCE algorithm, thus solving the problem of unbalanced data sets. The steps of the XGBoost algorithm (XGBoost_WCE) based on weighted cross-entropy loss are shown below.

4 EXPERIMENTS AND ANALYSIS OF RESULT

4.1 Data sources, data pre-processing and assignment

In this paper, we select the loan information of the 36-month annual personal loan of Lending Club, a US P2P lending platform, in 2015 and the loan status is paid off. After selecting the data, exploratory analysis of data characteristics, data cleaning, removal of missing values, abnormal values, and removal of irrelevant variables were performed, resulting in 165,636 data as the study sample, in which the ratio of normal repayment to early repayment was 1:3.In order to keep the original early payoff ratio and non-early payoff ratio unchanged, the established model is evaluated. In this paper, the study sample is divided into a training sample set and a test sample set by stratified random sampling with a ratio of 3:1, where the training sample set consists of 132508 observation samples and the test sample set consists of 33128 observation samples.

In this paper, the status of having repaid the loan in full before the loan maturity date is considered as early repayment and assigned a value of 0. The duration of early repayment and

on-time repayment is defined as the time between loan disbursement and full repayment. on-time repayment is defined as the time between loan disbursement and full repayment.

Algorithm 2 **XGBoost_WCE**

Require: Training sets D，Test set T
Ensure: Data set prediction categorie Y^0
Input training set D
for i in n **do**
If t == 0 **then**
Initialize sample category labels Y^*
Based on the weighted cross-entropy loss function, the input samples are really class labeled Y and sample class marker predicted values Y^*Calculate the first-order gradient G and second-order gradient H of the sample

Train the tth regression tree according to G, H
The tth regression tree and the previous $t-1$regression trees form a strong learner $F^*(t)$
Use $F^*(t)$ to predict the samples in dataset D and get the dataset category predicted value Y^* and the true value Y
end if
end for
Use the completed training $F^*(t)$ as the final classifier $F(t)$
Input test set T,use $F(t)$to predict and get prediction class label Y^0

4.2 Evaluation Indicators

For the evaluation of the classification performance of the classifier trained by the XGBoost algorithm, a number of discriminant criteria are selected for the comprehensive evaluation, which is generally based on the values in the confusion matrix [14], this is shown in Table 1.

TABLE 1 Confusion Matrix

Type	Predicted Positive Class	Predicted Negative Class
Actual positive class	TP	FN
Actual negativ class	TP	FN

Based on the confusion matrix, the true rate (recall), true negative rate, false-positive rate, false-negative rate, and positive class prediction value (prediction) can be calculated as Eq. (9)

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{9}$$

The accuracy rate indicates the proportion of all correct predictions of the classification model to the total number of observations as shown in Eq. (10)

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$Precision$ rate indicates the proportion of results where the model prediction is a positive case and the model prediction is correct as shown in Eq. (11)

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

Recall indicates the proportion of all outcomes for which the true value is a positive example and the model predicts correctly as shown in Eq. (12)

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

The specificity indicates the proportion of all outcomes for which the true value is a counterexample and the model predicts correctly.

For general classification problems, the most commonly used rating metric is accuracy (accuracy). However, for unbalanced data sets, it is expected that the higher classification accuracy of a few classes is better than the overall good classification performance just because the classification accuracy of most classes is high, and the improvement of the classification accuracy of a few classes is the focus of research, so we need to distinguish the accuracy of the model and construct the discriminant metric accordingly. Common evaluation criteria used for classification of imbalanced data are F-value for positive classes, and G-mean which considers the classification performance of both classes as shown in Eq. (13) and Eq. (14)

$$F - value = \frac{2Precision * Recall}{Precision + Recall} \tag{13}$$

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{14}$$

In order to better reflect the classification effect of the model, ROC curve and AUC value is used to evaluate the generalization ability of the model. The closer the ROC curve is to the (0,1) point, the greater the diagonal error with 450, and the closer the total area under the curve (AUC value) is to 1, the better the actual classification effect of the model.

In summary, the evaluation metrics selected in this paper are Accuracy, Recall, F-value, G-mean, and AUC.

4.3 Analysis of feature selection results

The feature selection part first uses the Pearson correlation coefficient method to evaluate the correlation degree of the features, as shown in Fig.2 is a partial data iso-correlation heat map. The heat map depicts the Pearson coefficient of two features, which indicates the correlation of two features. The correlation coefficient between installment and loan_amnt in the figure is 0.95, which indicates that the two features have strong correlation, so it is enough to choose one feature in installment and loan_amnt.

After filtering out the features with strong correlation, the SVM-RFE feature selection method is used to solve the problem of feature redundancy. Fig.3 shows the change curves of the AUC values corresponding to different number of features based on the SVM method.

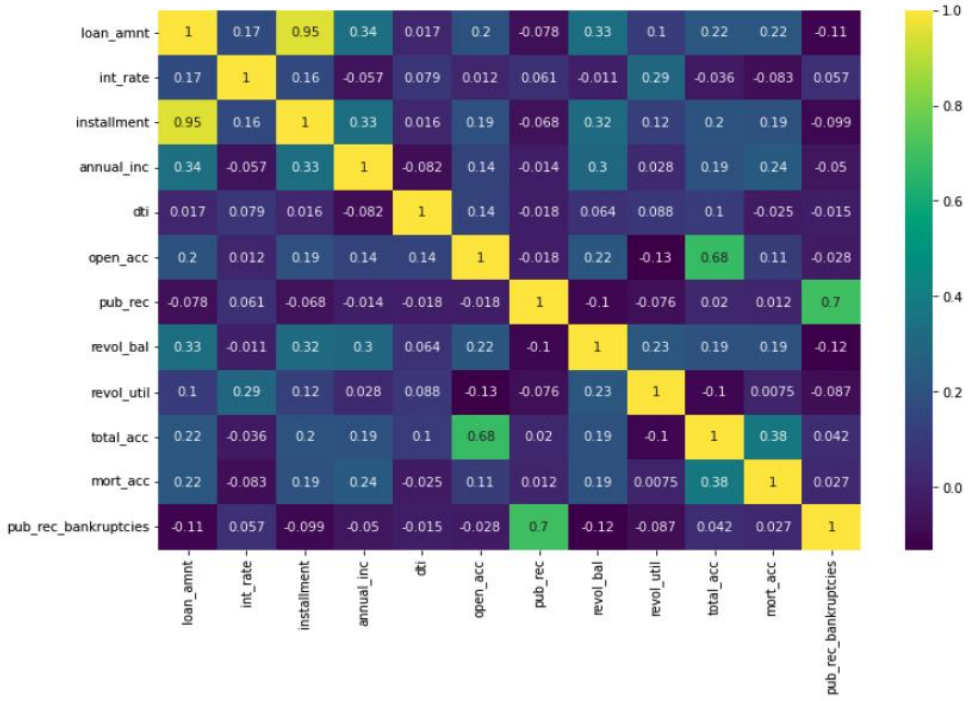

Fig. 2. Heat map of partial feature correlation

TABLE 2 FEATURE SELECTION COMPARISON EXPERIMENT

method	accuracy	recall	F-value	G-mean	Number of characteristics
Original data	0.8562	0.6153	0.6873	0.7774	25
Pearson	0.8692	0.6821	0.7282	0.8080	22
SVM-RFE	0.8794	0.8794	0.7508	0.7619	18
Pearson + SVM-RFE	0.8880	0.8005	0.7860	0.8594	15

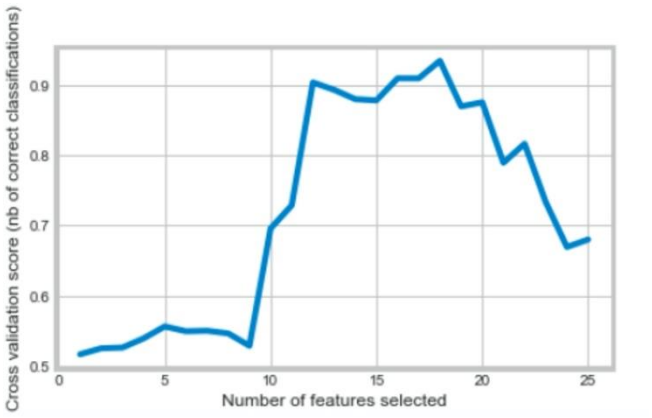

Fig. 3. SVM-RFE feature selection results

In order to further verify the effectiveness of feature selection in this paper, XGBoost is used as the classification model, and the results of the comparison experiments are shown in Table 2. The original feature number is 25, the three features with strong correlation are filtered out by Pearson correlation coefficient to get the feature number of 22, the SVM-RFE feature selection method gets the feature number of 18, and the combination of the Pearson coefficient and SVM-RFE feature selection method finally gets the feature number of 15. By comparing the experimental results, it is found that the classification ability of XGBoost is improved after feature selection, in which the feature selection method with the combination of the Person coefficient and SVM-RFE outperforms other feature selection methods in every index.

4.4 Classification model predictive analysis

Based on feature selection, three integrated learning algorithms, XGBoost, LightGBM and CatBoost, are selected in order to verify the effectiveness of the model after the fusion of XGBoost and weighted cross-entropy loss function (XGBoost_WCE). In terms of the selection of experimental model parameters, the grid search method is used for parameter selection, and the relevant important experimental parameter settings are shown in Table 3, while all other parameter settings are default.

Fig.4, Fig.6, Fig.8 and Fig.10 are the results of the confusion matrix visualization of the prediction results of this algorithm and XGBoost, LightGBM and CatBoost algorithms respectively. Only 631 minority class samples were identified by XGBoost, LightGBM, and CatBoost, and 1050, 2682, and 3046 minority classsamples, which can prove that XGBoost WCE is better than the other three models in identifying minority class samples. XGBoost WCE is effective in identifying minority class samples.Fig.5, Fig.7, Fig.9 and Fig.11 are the ROC curves of the algorithm in this paper with XGBoost, LightGBM and CatBoost algorithms plotted by AUC values, respectively. The AUC values of XGBoost WCE, XGBoost, LightGBM, and CatBoost models have AUC values of 0.9501, 0.9473, 0.9054, and 0.9114, respectively, and it can be seen that XGBoost WCE has the highest AUC value.

The evaluation metric accuracy reflects the overall recognition accuracy of the models XGBoost WCE, XGBoost, LightGBM, and CatBoost models of 0.8975, 0.8880, 0.8299, and 0.8304,

respectively, the results are shown in Table 4, which shows that XGBoost WCE is also better than other models in terms of accuracy. The evaluation metric Recall reflects the accuracy of the model in identifying a few classes of samples, and XGBoost WCE is 0.0237 higher than the XGBoost model. F-value and G- mean, which apply to the evaluation of unbalanced data sets, reflect the equilibrium effect of the model, and the F-value and G-mean of XGBoost_WCE, XGBoost, LightGBM, and CatBoost models have F-value of 0.8224, 0.7860, 0.6793 and 0.6625, respectively, and G-mean values of 0.9061, 0.8594, 0.7889 and 0.7707, respectively, and the F1 values and G mean of XGBoost_WCE are also significantly superior to the other three models. It can prove the classification ability of model XGBoost_ WCE for unbalanced data.

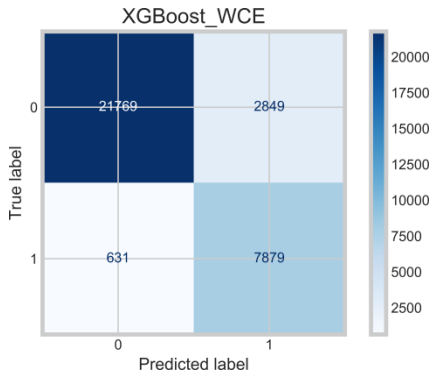

Fig. 4. XGBoost_WCE Confusion Matrix

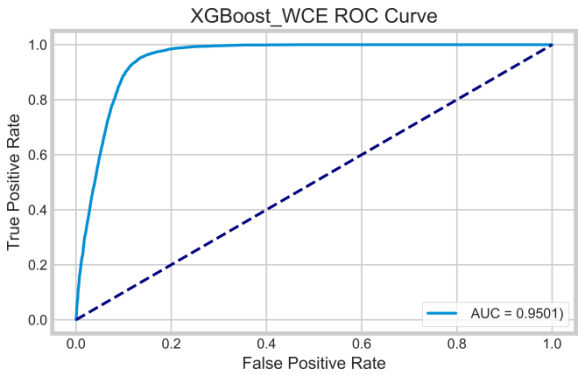

Fig. 5. XGBoost_WCE ROC Curve

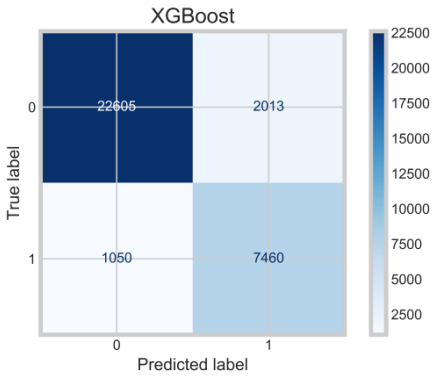

Fig. 6. XGBoost Confusion Matrix

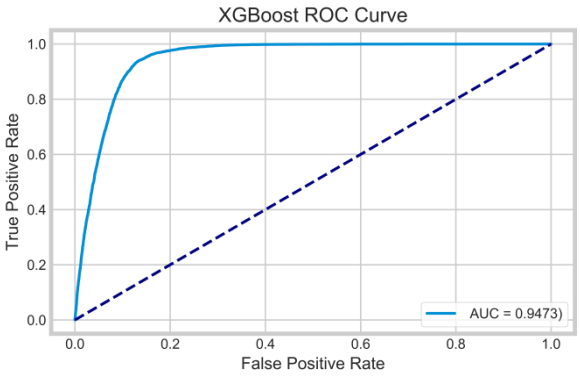

Fig. 7. XGBoost ROC Curve

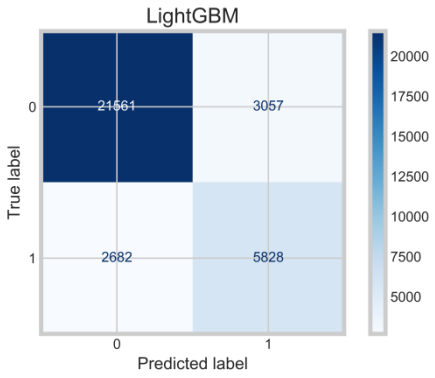

Fig. 8. LightGBM Confusion Matrix

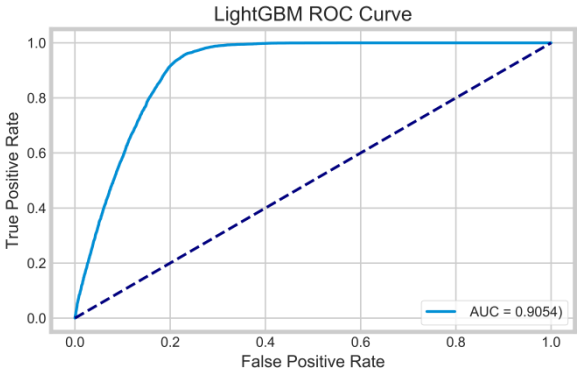

Fig. 9. LightGBM ROC Curve

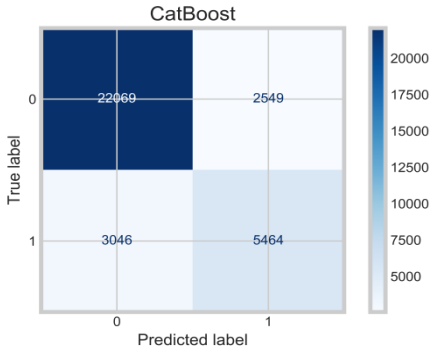

Fig. 10. CatBoost Confusion Matrix

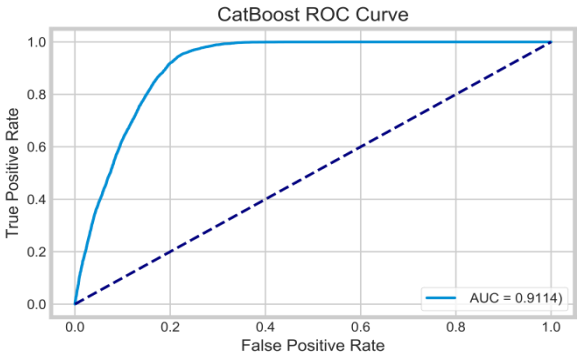

Fig. 11. CatBoost ROC Curve

TABLE 3 Model parameter setting

Model	Main parameter settings
XGBoost	n_estimators=200, learning_rate=0.5, max_depth=6
LightGBM	learning_rate=0.05, num_leaves=300, n_estimators=500
CatBoost	depth=8, itearations=500

TABLE 4 Model prediction results

Algorithm	accuracy	recall	F-value	G-mean
XGBoost_WCE	0.8975	0.9237	0.8224	0.9061
XGBoost	0.8880	0.8005	0.7860	0.8594
LightGBM	0.8299	0.7015	0.6793	0.7879
CatBoost	0.8304	0.6478	0.6625	0.7707

5 CONCLUSION

For the prediction of loan prepayment behavior, a SVM- RFE-XGBoost based loan prepayment prediction model is proposed and validated with the loan dataset of Lending Club. The experimental results show that the feature selection method combining Pearson coefficients and SVM-RFE can select the features with the strongest classification ability while filtering the relevant features. In the classification problem of unbalanced data, introducing a weighted cross-entropy loss function to improve XGBoost can effectively solve the problem that the model is weak in recognizing minority class samples, and then improve the robustness of the model.

In future research, the early repayment problem is further addressed by dividing the forecast period to achieve multi- category forecasting, as well as in the solution of data imbalance problem can be transformed from the level of data to make the proposed model more commercially valuable.

REFERENCES

[1] Li Z, Yao X, Wen Q, et al. Prepayment and Default of Consumer Loans in Online Lending[J]. SSRN Electronic Journal, 2016.

[2] Li Z, Li K, Yao X, et al. Predicting prepayment and default risksof unsecured consumer loans in online lending[J]. Emerging Markets Finance and Trade, 2019, 55(1): 118-132.

[3] Liang T H, Lin J B. A two-stage segment and prediction model for mortgage prepayment prediction and management[J]. International Journal of Forecasting, 2014, 30(2): 328-343.

[4] Sirignano J, Sadhwani A, Giesecke K. Deep learning for mortgage risk[J]. arXiv preprint arXiv:1607.02470, 2016.

[5] Mortazavi A, Moattar M H. Robust feature selection from microarray data based on cooperative game theory and qualitative mutual informa- tion[J]. Advances in bioinformatics, 2016, 2016.

[6] Schapire R E, Singer Y. Improved boosting algorithms using confidence- rated predictions[J]. Machine learning, 1999, 37(3): 297-336.

[7] Zhao C, Xin Y, Li X, et al. A heterogeneous ensemble learning framework for spam detection in

social networks with imbalanced data[J]. Applied Sciences, 2020, 10(3): 936.
[8] Xu J, Tang B, He H, et al. Semisupervised feature selection basedon relevance and redundancy criteria[J]. IEEE transactions on neural networks and learning systems, 2016, 28(9): 1974-1984.
[9] Guyon I, Weston J, Barnhill S, et al. Gene Selection for Cancer Classification using Support Vector Machines[J]. Machine Learning,2002, 46(1-3):389-422
[10] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology,2007, 2(3, article 27).
[11] Ma Y D, Liu Q, Qian Z B. Automated image segmentation using improved PCNN model based on cross-entropy[C]// Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004. IEEE, 2005.
[12] Pihur V, Datta S, Datta S. Weighted rank aggregationof cluster validation measures: a Monte Carlo cross-entropy approach[J]. Bioinformatics, 2007, 23(13):1607.
[13] Chen T, Guestrin C. Xgboost: A scalable tree boosting system [C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785- 794.
[14] He H, Ma Y. Imbalanced learning: foundations, algorithms, and applications[M]. Wiley-IEEE Press, 2013