ISSN: 2664-9640

Personal Credit Risk Forecast Model Based on BO-XGBoost

Gen Zhang [1], Limin Shao [1] and Xiaofang Liu [1,2]

[1] Artificial Intelligence Key Laboratory of Sichuan Province, Automation and Information Engineering, Sichuan University of Science and Engineering Zigong, 643002, China;

[2] School of Computer Science and Engineering, Sichuan University of Science and Engineering, Zigong, 643002, China.

Abstract

With the development of financial business in our country, the consumption mode of personal credit economy is generally accepted in daily life. Personal credit economy brings great convenience, but also brings huge personal credit risk. The accuracy of personal credit risk prediction model is particularly important. The logical regression of the traditional financial risk forecasting model can not solve the nonlinear problem and it is difficult to fit the real distribution of data, resulting in low prediction accuracy. In order to solve these two problems, this paper uses XGBoost (eXtremeGradientBoosting) model, which can solve not only nonlinear problems but also difficult fitting problems, but the prediction accuracy is still not high and there is room for further improvement. In this paper, the super-parameters of XGBoost model are optimized by Bayesian algorithm to form BO-XGBoost (Bayesian Optimization XGBoost) model, and the prediction results are analyzed and compared with traditional logical regression model and unoptimized XGBoost model. The results show that the prediction accuracy of the optimized XGBoost model is higher than that of the traditional logical regression model and the unoptimized BO-XGBoost model.

Keywords

BO-XGBoost, XGBoost, Bayesian Optimization, Personal Credit Risk.

1. Introduction

The balance of non-performing loans and the ratio of non-performing loans of commercial banks and various personal credit platforms across the country have seen a "double rise" since the fourth quarter of 2011, and the asset quality of commercial banks has declined [1]. Effective assessment of personal credit risk is of great significance for banks, regulators and even the financial stability of the whole society. As the mainstream method of credit risk assessment, logical regression is simple and easy to operate, but it can not solve the nonlinear problem and is difficult to fit the real distribution, which leads to the lack of accuracy of credit evaluation, It is difficult to obtain good prediction results [2]. Based on this, this paper uses XGBoost (eXtreme Gradient Boosting) algorithm based on integrated tree model, Proposed by Chen [3], the algorithm has high complexity, can effectively process nonlinear data, and can well fit the real distribution of data. The parameters of xgboost are optimized by Bayesian optimization algorithm to form bo-xgboost (Bayesian Optimization xgboost) model.

The experimental results show that the prediction accuracy of BO-XGBoost model is improved compared with the traditional Logistic Regression Algorithm and the non optimized XGBoost algorithm.

ISSN: 2664-9640

2. XGBoost Model

XGBoost is an extreme gradient boosting algorithm and an integrated tree Model, the objective function of XGBoost is:

$$\begin{cases} obj = \sum_i l(\hat{y}_i + y_i) + \sum_k \Omega(f_k) \\ \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \end{cases} \tag{1}$$

Among them, *obj* is the loss function, used to measure the difference between the predicted value $\hat{y}_i$ and the true value y_i, f_k represents the *k-th* tree model, and the second term $\Omega(f_k)$ is the penalty function, that is, the complexity of the penalty model. In the penalty function, γ is the complexity parameter, *T* is the leaf node tree, and λ is the penalty coefficient of the leaf weight ω [4]. The penalty function $\Omega(f_k)$ helps to smooth the final learned weights to avoid overfitting. The XGBoost model is a stepwise additive model, which is formed by adding *k* tree models. The complete iterative decision tree is:

$$\hat{y}_i^{(k+1)} = \hat{y}_i^{(k)} + \eta f_{k+1}(X_i) \tag{2}$$

f_{k+1} is the *k+1* tree model, η is the step length of the iteration, that is, the learning rate, X_i represents the *i-th* instance. The size of η determines the iteration speed. The smaller the η the slower the convergence speed, but a more accurate optimal value can be found.

Let $\hat{y}_i^{(k)}$ be the predicted value of the *i-th* instance X_i in the *k-th* iteration, use $\hat{y}_i$ in equation (1) as a parameter, and add f_k to minimize the objective function:

$$obj^k \approx \sum_{i=1}^{n}[g_i f_k(X_i) + \frac{1}{2}h_i f_k^2(X_i)] + \sum_{i=1}^{k}\Omega(f_i) \tag{3}$$

Among them, g_i and h_i are the first step statistics and the second step statistics of the loss function.

Define $I_j = \{i \mid q(X_i) = j\}$ as the sample number set of the *j-th* leaf node, and rewrite formula (3) as:

$$obj^{(k)} = \sum_{j=1}^{T}[(\sum_{i\in I_i} g_j)\omega_j + \frac{1}{2}(\sum_{i\in I_i} h_j + \lambda)\omega_j^2] + \gamma T \tag{4}$$

Obtain the first-order partial derivative of the objective function of formula (4) with respect to ω_j, and set it equal to *0* to obtain the optimal weight corresponding to the leaf node *j*:

$$\omega_j^* = -\frac{\sum_{i\in I_j} g_j}{\sum_{i\in I_j} h_j + \lambda} \tag{5}$$

And the optimal value of the objective function is:

$$obj^{\mathrm{k}} = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_j + \lambda} + \gamma T \tag{6}$$

Equation (6) is also a structure score, which is used to judge the quality of the structure tree. It is obtained for different loss functions [5] [6].

ISSN: 2664-9640

3. Optimization Model BO-XGBoost

Commonly used parameter optimization includes manual search method [7], grid search method [8], random search method [9], chaotic particle swarm optimization [10], and Bayesian optimization [11]. Bayesian optimization is an intelligent algorithm based on prior knowledge of hyperparameters. Compared with other optimization algorithms, Bayesian optimization can better select possibly better hyperparameters, reduce unnecessary objective function estimation, and make it faster Find the optimal solution. In this paper, Bayesian algorithm is used to optimize the XGBoost model to form the BO-XGBoost model with AUC as the target.

In this paper, the parameters of Bayesian optimization XGBoost mainly include three categories: Booster parameters, general parameters, and target parameters. The optimized parameters of XGBoost are shown in Table 1.

Table 1: Basic situation of XGBoost optimization parameters

Parameter Name	Adjustment Range	Optimized Value
n_estimator	1000-2500	2190
learning_rate	0.01-0.3	0.04
gamma	0-10	1
max_depth	0-5	4
lambda	0-5	3
alpha	0-5	1

The agent model and acquisition function of Bayesian optimization select TPE and EI to construct Bayesian optimization algorithm. According to the prior probability distribution $p(x|y)$ of the hyperparameter x, the TPE proxy model is used to estimate the corresponding target function risk value distribution $p(y)$, and the next hyperparameter is selected according to the EI acquisition function. Repeat the above process and continuously use the posterior distribution of the surrogate model to select hyperparameters until the optimal solution is obtained.

The TEP probability distribution is defined as follows:

$$p(x|y)=\begin{cases} e(x) & if \quad y<y^* \\ g(x) & if \quad y\geq y^* \end{cases} \tag{7}$$

Where $e(x)$ is the density formed by the observation value x, $g(x)$ is the density formed by the remaining observation values excluding x, x represents the parameter, and y represents the risk value.

The TPE algorithm chooses y^* as a certain quantile γ of the current observed risk value y, and satisfies $p(y<y^*)=\gamma$. Through the $e(x)$ and $g(x)$ of the TPE algorithm, the parameter set is reasonably divided into parts with less risk and greater risk [12]. After further optimization through the maximum expectation improvement, EI is defined as shown in equation (8).

$$EI_y(x)=\int_{-\infty}^{y^*}(y^*-y)p(x\mid y)dy \tag{8}$$

According to Bayes' theorem, formula (8) can be written as:

$$EI_y(x)=\int_{-\infty}^{y^*}(y^*-y)\frac{p(x\mid y)p(y)}{p(x)}dy \tag{9}$$

ISSN: 2664-9640

From $p(y < y^*) = \gamma$ and formula (10)

$$p(x) = \int p(x \mid y)dy = \gamma e(x) + (1-\gamma)g(x) \tag{10}$$

Equation (9) can be rewritten as:

$$\int_{-\infty}^{y^*} (y^* - y)p(x \mid y)dy = \gamma y^* e(x) - e(x)\int_{-\infty}^{y^*} p(y)dy \tag{11}$$

In the end, you can get:

$$EI_y(x) = \frac{\gamma y^* e(x) - e(x)\int_{-\infty}^{y^*} p(y)dy}{\gamma e(x) + (1-\gamma)g(x)} \propto [\gamma + \frac{g(x)}{e(x)}(1-\gamma)^{-1}] \tag{12}$$

It can be seen from equation (12) that when the probability of hyperparameter x in $e(x)$ is as large as possible, when the probability of $g(x)$ is as small as possible, the greatest expected improvement can be obtained. In each iteration, the algorithm will return with the largest EI The value of the hyperparameter.

4. Experiment and Analysis

4.1. Data Preprocessing

The data set used in this article is the desensitized personal credit data set published by Ali Tianchi, and the basic data is shown in Table 2.

Table 2: Basic situation of personal credit data

Data name	Number
Feature number	28
Number of samples	800000
Positive and negative sample ratio	7:1
Missing value	38000

First divide the data set into 8:2, then convert the time feature type of the data to a numeric feature type, and then fill in the missing values of the data by the median. The self-mapping transformation is carried out for the hierarchical category features, and the non-high-dimensional features with more than two categories are one-hot encoded. The data preprocessing situation is shown in Table 3.

Table 3: Basic data preprocessing

Data	Operation method
Time characteristics	Convert to digital features
Missing value	Median padding
Hierarchical categorical characteristics	Self-mapping
Low-dimensional category features	One Hot Encoding
High-dimensional and continuous features	Decision Tree Binning

ISSN: 2664-9640

WOE coding is a supervised coding method, which is widely used in the field of risk control. WOE coding takes the concentration attribute of the prediction category as the value of the coding, and can normalize the value of the feature to the scale of myopia [13].

High-dimensional features and continuity features are binned by a supervised decision tree, and WOE coding is performed after binning. Decision tree binning is to use the threshold value of the internal partition node generated by the decision tree as the boundary of binning. The discrete high-dimensional features and continuous features in the training set are binned in a supervised decision tree, and then the boundary value list is used for the test set binning basis, and WOE coding is performed after binning is completed.

Use the correlation coefficient method for feature selection, and delete the features with a feature correlation value greater than 0.9, such as features n0, n2, n9, subGrade, and ficoRangeLow.

4.2. Experiment Process

The experiment uses k-fold cross-validation [14] to select *k* as 5. First, the data is equally and randomly divided into *5* parts. During the training process, each time *4* folds are used as the training set, and *1* fold is used as the validation set to verify the model. Using the k-fold cross-validation method can make full use of the data, which is more stable and comprehensive than dividing the training set and the test set in a single time.

The personal credit prediction flowchart of the BO-XGBoost model is shown in Figure 1.

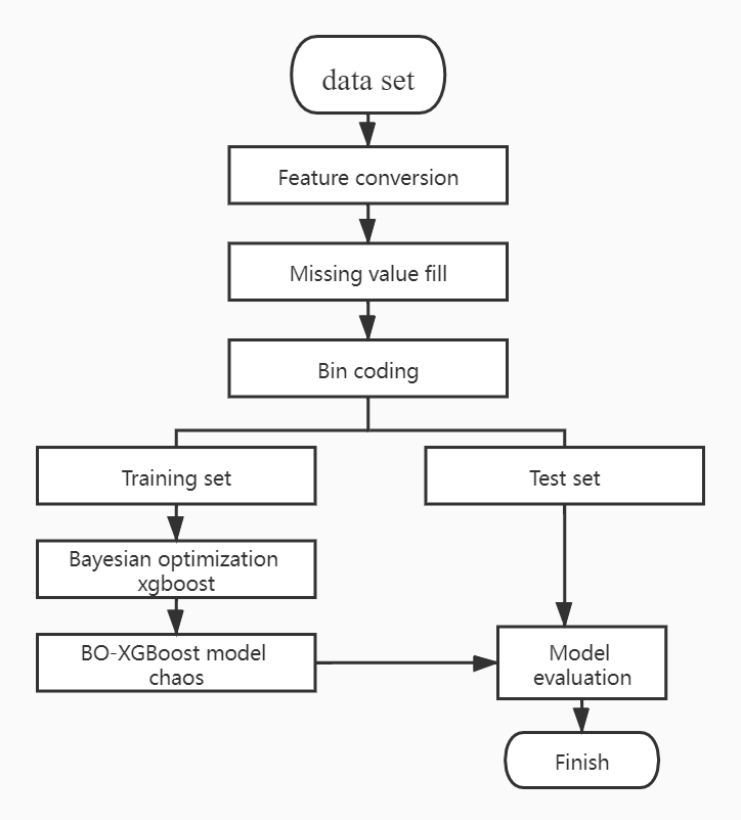

Figure 1: Personal credit prediction flowchart of BO-XGBoost model

4.3. Evaluation Index and Model Comparison Evaluation

This article uses AUC value as the evaluation index. Bring the optimized parameters in Table 1 into BO-XGBoost and compare them with the unoptimized XGBoost model and logistic regression model (LOG). The experimental results are shown in Table 4.

ISSN: 2664-9640

Table 4: Comparison of model experiment results

Model	LOG	XGBoost	BO-XGBoost
1-K	0.63	0.69	0.73
2-K	0.62	0.70	0.72
3-K	0.63	0.70	0.73
4-K	0.63	0.71	0.73
5-K	0.64	0.70	0.74
mean	0.63	0.70	0.73

From Table 4, the following conclusions can be drawn: 1) The unoptimized XGBoost model is better than the traditional logistic regression model. 2) Compared with the unoptimized XGBoost, after Bayesian optimization has increased by 3%, and the AUC of the traditional logistic regression model has increased by 10%.

5. Conclusion

In view of the traditional logistic regression of financial risk prediction models, it cannot solve the non-linear problem and it is difficult to fit the true distribution of the data, resulting in low prediction accuracy. This paper uses the XGBoost algorithm model with higher algorithm complexity, which can better deal with nonlinear problems, and better fit the real data distribution, so that the prediction accuracy is improved. Then use Bayesian algorithm to optimize the super parameters of XGBoost to form the BO-XGBoost model, which further improves the prediction accuracy.

Acknowledgements

This work was supported in part by Sichuan Science and Technology Program under Grant 2017GZ0303, in part by Fund Project of Sichuan Provincial Academician (Expert) Workstation under Grant 2016YSGZZ01, and in part by Special Fund for Training High Level Innovative Talents of Sichuan University of Science and Engineering under Grant B12402005.

References

[1] D.F. Li: *Research on the Optimization Strategy of Personal Credit Risk Assessment of Bank Z's Binjiang Sub-branch* (MS., Zhejiang Gongshang University, China 2019). p.28.

[2] G.B. Fernandes, R. Artes: Spatial Dependence in Creditrisk and its Improvement in Redit Scoring, European Journalof Operational Research, (2016) NO.2, p.517-524.

[3] T.Q. Chen, Guestrin C: XGBoost:a scalable tree boosting system, *The 22 ACM SIGKDD International* (Conference.San Francisco, USA, March 09, 2016), p.785-794.

[4] Y. Xia, C. Liu, Y.Y. Li: A boosted decision tree approachusing Bayesian hyper-parameter optimization for credit scoring, Expert Systems with Applications, Vol. 78 (2017) p.225-241.

[5] M.H. Wang, X.C Liang: Personal credit evaluation based on CPSO-Xgboost, Computer Engineering and Design, Vol. 40 (2019) No.7, p.1891-1895.

[6] L. Zhang, J.Q. Wang, Z.Y. Fei: Bank User Personal Credit Risk Assessment Model Based on RF-SMOTE-Xgboost, Modern Electronic Technology, Vol. 43 (2020) No.16, p.76-81.

[7] W. Gao: *Research on stock robo-advisor strategy based on support vector machine parameter optimization algorithm* (MS., Shanghai Normal University, Chian 2018), p.28.

ISSN: 2664-9640

[8] X.S. Liu, Z.B. Zang: SVM parameter optimization based on improved grid search method, Journal of Jiangxi University of Science and Technology, Vol. 40 (2019) No.1, p.5-9.

[9] J. Bergstra,Y. Bengio: Random search for hyper-parameter optimization, Journal of Machine Learning Research, Vol.13 (2012) No.1, p.281-305.

[10] Z.D. Lu: *Research on Improvement and Application of Hybrid Particle Swarm Algorithm* (MS., Yanshan University, Chian 2020), p.29.

[11] H.T. Shi, Y.J. Shang, X.T. Bai: Research on Early Failure Prediction Method of SWDAE-LSTM Rolling Bearing Based on Bayesian Optimization, Vibration and shock, Vol. 40 (2021) No.18, p.286-297.

[12] J. Song, G.S. Chen, J.F. Chen: Size prediction of injection molded products based on feature selection and Bayesian optimization of LightGBM, Engineering Plastics Application, Vol. 49 (2021) No.08, p.54-60.

[13] Z.F. Li: *Research on loan default risk prediction based on cost-sensitive AdaBoost* (MS., Jiangxi University of Finance and Economics, Chian 2021), p.19.

[14] Y. Bengio, Y. Grandvalet: No unbiased estimator of the variance of K-fold cross-validation, Journal of Machine Learning Research, (2004) No.5, p.1089-1105.