International Journal of Applied Science and Mathematics
Volume 9, Issue 4, ISSN (Online): 2394-2894

The use of the Improved XGBoost Algorithm in Credit Debt Risk Auditing

Xiangrong Shi[1], Qi Zheng[2*], Pengsai Guo[3] and Yifei Ye[4]
[1]Department of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, Hang Zhou, China.
[2,3,4]Department of Accounting, Zhejiang University of Finance and Economics, Hang Zhou, China.
*Corresponding author email id: 2794169863@qq.com

Date of publication (dd/mm/yyyy): 15/07/2022

***Abstract* – In recent years, credit debt default events have occurred frequently, and the current credit debt default has become a common phenomenon in the capital market. Many high-rated companies have also defaulted on their bonds, and credit ratings are no longer sufficient to predict a company's default. This paper utilizes the mainstream Logistic algorithm and XGboost algorithm to construct a credit debt default risk method, the performance of the two is compared and analyzed, and the historical data information is used to improve the original algorithm, and the experimental results testify that the improved model prediction effect is better than the original algorithm, of which the enhanced-XGboost algorithm has the best comprehensive performance in various indicators, which is more suitable for credit debt risk audit.**

***Keywords* – Credit Debt Default, Logistic, XGBoost, Algorithm Boosting, Risk Audits.**

I. INTRODUCTION

In recent years, China's economy has developed rapidly, the demand for companies to raise funds has increased significantly, and the threshold for issuing bonds has been lowered, and the number and amount of bonds issued in the market have continued to rise. This has led to uneven bond quality and more and more credit debt defaults. The default of the bond issuer also brings great difficulties to the auditors, who cannot speculate whether the company will default based on traditional audit methods alone, and may eventually cause the audit to deviate.

Credit ratings were supposed to improve the information asymmetry between investors and issuers, but the recent spate of high-credit-rated corporate credit defaults has led investors to doubt the effectiveness of credit rating systems.

For the research of bond default, Most of the research in China still adopts the method of theoretical and case analysis, and this paper introduces machine learning, adopts XGboost algorithm to build an empirical method, and compares it with the prediction results of Logistic algorithm. By studying the impact of bond issuance data and financial indicators of issuers on bond default risk, to analyze whether corporate bonds will default, the application of the algorithm greatly reduces investors' dependence on credit ratings, and also provides auditors with another indicator for reference, which not only helps to ensure the rationality of bond pricing, but also helps investors weigh the benefits and risks of bonds, avoid blind investment according to a single rating, and then promote the healthy and stable development of the domestic bond market.

II. REVIEW OF RELATED THEORIES AND LITERATURE

A credit bond is a bond issued by a non-government with a fixed principal, interest rate and regular delivery time. It is a bond issued externally with the company's creditworthiness as a guarantee.

2.1. *Risk Theory of Credit Bond Default*

According to the theory of information asymmetry, bond issuers have more information than the outside public, and investors may not know that the company may have had or is about to have financial problems, purchased a credit bond and faced a situation where it could not receive a return when it matured. The risk-reward matching theory testifys that because investors believe that the government will help companies to get the bottom, the investment targets are high-interest rate bonds, thus ignoring the law that the higher the return, the higher the risk. The broken window theory shows that the bond market has entered a vicious circle, and the issuer believes that even if it is unable to repay the bonds, the government will help pay on its behalf, so the issuance of bonds does not need to consider their own solvency. Similarly, investors also feel that there is no need to consider whether the issuer can afford to pay the bond, if it cannot be repaid when it is due, the government will pay it instead, so they all want to buy high-interest bonds.

2.2. *Literature Review*

Some scholars employ a single algorithm to study credit debt defaults. For example, Ping Cao utilized the KMV model to verify the possibility and distance of default between provinces and municipalities directly under the central government, and unified the method for assessing the possibility of default of local government bonds [3]. Chang Yan and Jungang Xu used the Logit algorithm to analyze the financial situation and the company's situation in the current period, calculate the probability of the company's default, and synthesize the indicators to obtain the changes in the company's development [10]. Aoxue Dong constructed a logistic multi-regression bond default risk early warning model [12]. Xuguang Wang constructed a VAR vector autoregressive method to analyze the elements that affect the scale of credit bonds issued by issuers [13].

There are also scholars who adopt combination algorithms to study credit debt defaults. For example, Tianxin Zhang's method of setting up a bond default risk algorithm is different, first using the support vector machine algorithm, then adding kernel functions to the model to solve the nonlinear classification problem, and also creatively adding corporate public opinion indicators [5]. Xuebin Chen, Jing Wu and others have a special method for measuring the probability of default on credit bonds: by using the Bayesian variation Gaussian mixed estimation method, the market index estimation method and the backward estimation method of the default probability change trend, the LSTM is employed to set up a credit bond default risk prediction method [11]. Kesu Wang has pointed out that the Light GBM algorithm has a lower effect on predicting bond defaults than the XGboost algorithm, but much higher than the SVM algorithm and the Logistic algorithm [9].

Di Hu utilized a random forest to build a bond default risk prediction algorithm, and the feature selection adopted the Lasso regression algorithm [6]. Qinghong Zhang constructed a random forest algorithm model and adopted the logistic method as the target, which highlighted the superiority of random forest in predicting the risk of bond default [7]. Yinong Zhu's research confirms that the combination of stochastic forest algorithm and KMV model can effectively assess the risk of bond default [4]. Shijie Zheng employed the KMV method to predict the possibility of credit debt default of unlisted companies, and used an artificial neural network algorithm (ANN) to estimate the asset value and volatility of the enterprise, and then used the default distance obtained by KMV as a variable to put it into the LOGIT model, and obtained the comprehensive method KMV-LOGIT [1]. Peipei Zhang utilized the KMV algorithm, the LOGIT model and the KMV-LOGIT combination method, and the three-party comparison found that the KMV-LOGIT algorithm with default distance worked best [8]. Yaqing Feng adopted the Bayesian optimization algorithm to tune the parameters of the model, and

obtained an improved KMV-XGboost method. Finally, the XGboost algorithm is compared with the traditional logit model and the random forest method, which testifys the superiority of this model [17]. Rongxi Zhou, Hang Peng, and others first employed the random forest algorithm to obtain the characteristics, and then utilized the XGboost algorithm to construct the default risk of bonds, which demonstrates that this algorithm has excellent prediction accuracy[15]. Feng adopted XGboost to construct a method for estimating the default risk of bonds, and compared with the estimation bond default risk model constructed by seven other algorithms, such as random forest and logistic regression, it was found that XGboost and random forest estimation bond default risk algorithm had a higher accuracy rate, but the stability of XGboost was better than that of random forest [14].

Judging from the above literature, the hybrid algorithm is generally more effective than the credit debt default risk prediction method constructed by a single algorithm. Stochastic forest, KMV and XGboost algorithms are widely employed in the construction of credit debt default risk models, and Logit and Logistic algorithms are also frequently used to compare and be compared. Of course, each algorithm algorithm has its own characteristics and scope of application, and it is necessary to adjust the parameters to achieve the optimal solution of the method.

Nowadays, there is not much research on the results of China's bond default risk prediction, mainly because in recent years, bond default events have gradually emerged, and there is a situation of increasing year by year, so the bond default risk prediction method made by the existing literature is still slightly rough and needs to be further improved.

In this paper, the Logistic and XGboost algorithms in machine learning algorithms are applied and optimized for these two algorithms. Because in the previous study, we did not consider combining historical information to make predictions, so we based on the basic assumption that companies that have a history of default records also have a high probability of defaulting in subsequent years. If you utilize information on time series, it is possible to improve the accuracy of predictions, so this paper proposes a lifting model based on historical performance. The creative use of enhanced-logistic and enhanced-XGboost algorithms in the study of credit debt default risk has resulted in different degrees of improvement of the indicators of the two original algorithm algorithms.

III. Model Principle

3.1. *Logistic Model*

Logistic, a probabilistic nonlinear regression, is a multivariate analysis method to study the relationship between binary classification observations and some influencing factors. The basic form of the Logistic algorithm function is as follows:

$$ln\frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad 0 < \pi < 1$$

$$\pi = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)} \quad 0 < \pi < 1$$

$$Y = \begin{cases} 0, & \pi \le A \\ 1, & x > A \end{cases}$$

Meaning of the formula: using matlab multiple linear regression to calculateβ_0,β_1,β_2,...β_k, the resulting values are substituted into the second formula, and the π is obtained, and the result of the ratio of π and A to

determine whether Y is 0 or 1.

Logistic has excellent ability to apply to binary classification, and logistic can be used if it is needed to use binary classification. In addition, logistic methods can also be adopted to predict the probability that something will happen in the future, which plays a role in predicting the default of the issuer's credit debt.

3.2. *XGBoost Algorithm Model*

XGboost's objective function includes two parts, the loss function and the regular term, the loss function represents the degree to which the method fits the data, we usually employ its first derivative to indicate the direction of gradient descent, XGboost also calculates its second derivative, further considering the trend of gradient changes, fitting faster and more accurate. Regular terms are utilized to control the complexity of the algorithm, the more leaf nodes, the larger the model, not only the operation time is long, beyond a certain limit, but also overfitted, resulting in a decline in classification effect. XGboost's regular terms are a penalty mechanism, and the greater the number of leaf nodes, the greater the penalty, thus limiting their number. It also greatly improves the performance of the algorithm through techniques such as pre-sorting and weighted quantiles.

A method whose promotion goal is to minimize structural risk, the general form of function optimization is:

$$obj = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \lambda\Omega(T)$$

Import the addition algorithm of the lift tree to the above equation:

$$obj = \sum_{i=1}^{n} l(y_i, f^{(t-1)}(x_i) + T^t(x_i;\theta)) + \lambda\Omega(T)$$

When the loss function is a square error loss, the above equation becomes:

$$obj = \sum_{i=1}^{n} l(y_i, f^{(t-1)}(x_i) + T^t(x_i;\theta)) + \lambda\Omega(T) = \sum_{i=1}^{n} l(y_i - f^{(t-1)}(x_i) - T^t(x_i;\theta))^2 + \lambda\Omega(T)$$

$$= \sum_{i=1}^{n} (r_i - T^t(x_i;\theta))^2$$

$$= \sum_{i=1}^{n} \left[y_i - f^{(t-1)}(x_i))^2 + (T^t(x_i;\theta))^2 - 2T^t(x^i;\theta)(y_i - f^{(t-1)}(x_i)\right] + \lambda\Omega(T)$$

$$= \sum_{i=1}^{n} \left[l\left(y_i, f^{(t-1)}(x_i)\right) + 2T^t(x_i;\theta)\left(f^{(t-1)}(x_i) - y_i\right) + (T^t(x_i;\theta))^2 + \lambda\Omega(T)\right]$$

If you adopt the general loss function as the loss function, coupled with the second-order Taylor, you have the original form of XGboost:

$$obj = \sum_{i=1}^{n} \left[l(y_i, f^{(t-1)}(x_i) + g_i T^t(x_i;\theta) + \frac{1}{2} h_i (T^t(x_i;\theta))^2\right] + \lambda\Omega(T)$$

There into,

$g_i = \partial_{f_i^{(t-1)}} l(y_i, f^{(t-1)}(x_i))$

$h_i = \partial^2_{f_i^{(t-1)}} l\left(y_i, f^{(t-1)}(x_i)\right)$

Delete the constant$l\left(y_i, f^{(t-1)}(x_i)\right)$,the objective function is converted to:

$$obj = \sum_{i=1}^{n}\left[g_i T^t l(x_i;\theta) + \frac{1}{2}h_i (T^t(x_i;\theta))^2\right] + \lambda\Omega(T)$$

Define the new treeT: $T^{(t)}(x) = \omega_{q(x)}$

The regular term is: $\Omega(T) = \gamma|T| + \frac{1}{2}\lambda\sum_{j=1}^{|T|}\omega_j^2$

Define a new tree along with regular items and import them into the objective function:

$$obj = \sum_{i=1}^{n}\left[g_i T^t l(x_i;\theta) + \frac{1}{2}h_i (T^t(x_i;\theta))^2\right] + \lambda\Omega(T) = \sum_{i=1}^{n}\left[g_i\omega_{q(x)} + \frac{1}{2}h_i\omega^2_{q(x)}\right] + \gamma|T| + \frac{1}{2}\lambda\sum_{j=1}^{|T|}\omega_j^2$$

Define$I_j = \{i|q(x_i) = j\}$ is a collection of leaf nodes, then the upper formula becomes:

$$obj = \sum_{i=1}^{n}\left[g_i\omega_{q(x)} + \frac{1}{2}h_i\omega^2_{q(x)}\right] + \gamma|T| + \frac{1}{2}\lambda\sum_{j=1}^{|T|}\omega_j^2 = \sum_{j=1}^{|T|}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\omega_j^2\right] + \gamma|T|$$

Define$G_j = \sum_{i\in I_j} g_i, H_j = \sum_{i\in I_j} h_i$

Then the objective function becomes:

$$obj = \sum_{j=1}^{|T|}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\omega_j^2\right] + \gamma|T| = \sum_{j=1}^{|T|}\left[G_j\omega_j + \frac{1}{2}(H_j + \lambda)\omega_j^2\right] + \gamma|T|$$

Partially derivative the objective function to ω, make it 0, and get: $\omega_j^* = -\frac{G_j}{H_j+\lambda}$

Import the objective function again and eliminate ω, and get: $obj^* = -\frac{1}{2}\sum_{j=1}^{|T|}\frac{G_j^2}{H_j+\lambda} + \gamma|T|$

The above equation is the final optimization objective function of XGboost.

IV. Case Study

4.1. *Sample Selection and Data Sources*

The sample of this article is taken from the wind database and contains all the companies that have issued credit bonds since 2014-2020, with a total of 4932 data.

The reason why the data was selected from 2014 was because Chao ri Sun was unable to repay the “11 Chao Ri Bonds” due in March 2014, and there was a substantial default, becoming the first public bond in China to default. In 2014, a total of 5 companies defaulted on credit bonds. The initial data selected issuers, issuance years and 30 financial indicators, including total assets, monetary assets, etc., which are also the indicators that

auditors focus on.

4.1.1. *Logistic and XGBoost Data Preprocessing*

Preprocess the sample data:

Step 1: Delete all data rows with #N/A

Step 2: Calculate the proportion of 0 values in each column, and then count the columns with 0 values accounting for more than 30% of the total number of data in the column, delete the statistics of the columns, and reduce the data column from 31 columns to 16 columns;

Step 3: Replace all 0 values with the average of the current column;

Step 4: Delete the first column of issuers and the second column of the issuance period, replace them with serial numbers, and change the 16 columns to 14 columns;

Step 5: Because the same company may have issued multiple credit bonds in a year, duplicate data is generated, and this step is to remove duplicate data and keep only one row of data issued by the same issuer in the same year.

The resulting "credit default" file is the data source file for the operation of XGboost and logistic algorithms.

4.1.2. *Enhanced-Logistic and Enhanced-XGboost Method Data Preprocessing*

In order to further improve the accuracy rate, the Logistic model and the XGboost computing method have been improved, and the enhanced-Logistic algorithm and the enhanced-XGboost algorithm need to be further processed on the preprocessed "credit debt default" file data in the following steps:

Step 1: Sort the data by company name and year;

Step 2: Add a column of "previous year" and put it in the last column, the list means whether the previous year is in default, and because most of the company's credit bonds issued in different years, the actual meaning of this column is whether the company last issued credit bonds in default. Some companies issue credit bonds for the first time, but the last one did not issue, which is also defined as the last issuance without default, marked as "no".

Step 3: Swap the "Previous Year" column and the "Defaulted" column so that the "Defaulted" column is in the last column of the file, which is still the Y value and the "Previous Year" column is the X variable value;

The resulting "credit debt default enhanced" file is used as a data source file for the Logistic optimization algorithm and the XGboost optimization algorithm algorithm.

4.2. *Model Construction*

Step 1: Import all the libraries you need to utilize and read the data source file;

Step 2: Construct four pieces of data, training set arguments and dependent variables, test set arguments and dependent variables;

Step 3: Put the accuracy rate, complete rate, accuracy rate, and auc of the four indicators into the list; Set up 30 loops and place the following method operations in the loop;

Step 4: The training set randomly selects 80% of the data in the sample, and the remaining 20% is employed as the test set, the X value and Y value are constructed, the algorithm is introduced for training, and the test set data is predicted by using the method;

Step 5: Define the confusion matrix formula, type the accuracy rate, check rate, accuracy rate, auc formula;

Step 6: Enter the parameters and draw a confusion matrix graph; Output accuracy, complete rate, accuracy rate, auc 30 times average result.

The construction of Logistic models and XGboost models differs in two ways: different import libraries and different introduction algorithms.

The difference between logistic method and logistic elevation algorithm construction, and the difference between XGboost model and XGboost elevation method construction are in the different data source files read.

Optimization algorithm to add a new column of "whether the previous year is in default", its actual effect is equivalent to the addition of 14 columns of data from the previous year, the 13 columns of this year and the 14 columns of data of the previous year together constitute X variables, and the Y variable "whether it is default" constitutes a model for operation, more accurate data import algorithms, greatly improving the performance of the algorithm method, stability, accuracy, etc. are better than the original algorithm algorithm.

4.3. *Model Results Analysis*

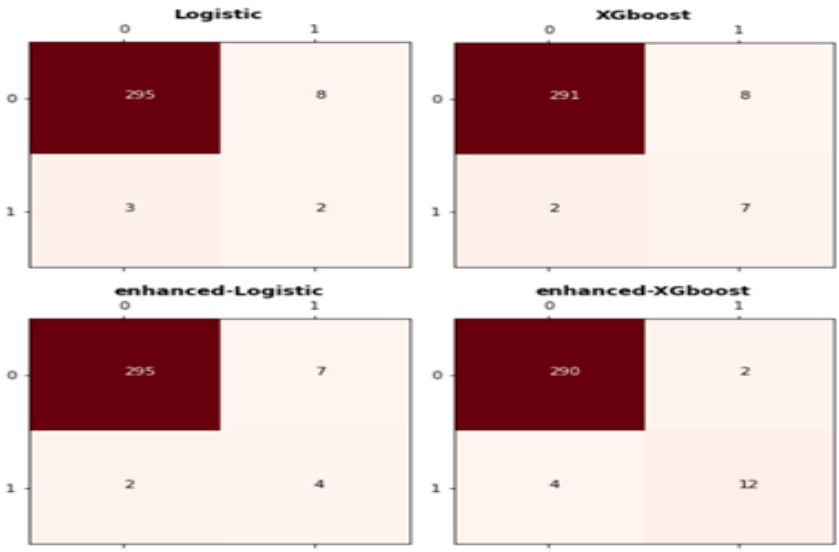

Fig. 1. Model confusion matrix comparison plot.

Combined with the TP, FN, FP, and TN values in Figure 1, the following table is calculated:

Table 1. Comparison table of model confusion matrix values.

Model	TP+TN	Misclassification Rate
Logistic	297	0.02597
XGboost	298	0.02597
enhanced- Logistic	299	0.02273
enhanced- XGboost	302	0.00649

From Table1, it can be found that the correct classification value of TP+TN of the enhanced-XGboost method is much higher than that of the other three algorithms, while the correct classification value of Logistic and XGboost models is the lowest. Logistic and XGboost methods have a misclassification rate of up to 0.02597 for samples, and the number of FPs is also 8; In contrast, enhanced-XGboost greatly improves the misclassification of samples, and also improves the accuracy of samples, with a misclassification rate of only 0.00649 and 2 FPs, which has better classification performance.

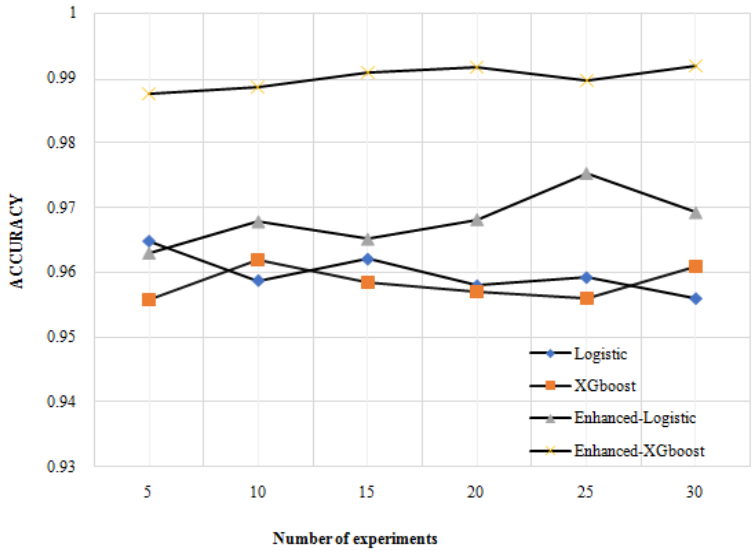

Fig. 2. Model accuracy comparison chart.

As can be seen from Figure 2, the accuracy of the Enhanced-XGboost algorithm is much higher than that of the other three methods, and the accuracy performance is the best. The accuracy of the enhanced-Logistic is higher than that of the other two unoptimized algorithm models. The accuracy rate of the Logistic and XGboost algorithms is almost the same.

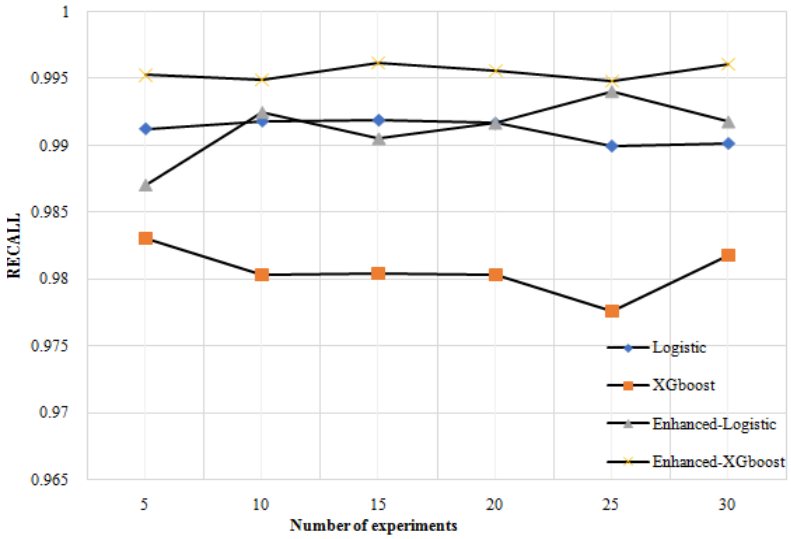

Fig. 3. Model lookup rate comparison.

Looking at the comparison chart of the method 4 algorithm, it can be seen that the checkout rate of the Enhanced-XGboost model is still the highest, the check rate of the Enhanced-Logistic method is slightly higher than that of the Logistic algorithm, and the XGboost model has the worst completeness, far lower than the other

three sets of methods. It can be seen that algorithm optimization also has an improving effect on the completion of the model.

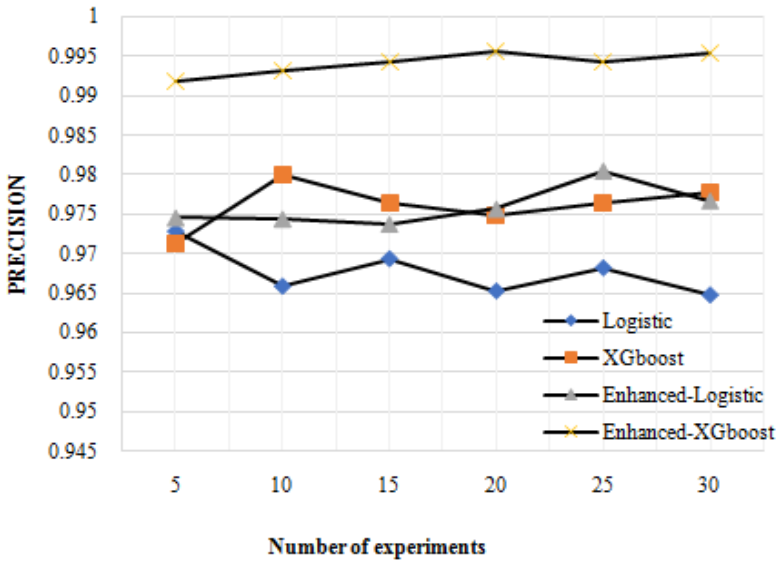

Fig. 4. Comparison of model accuracy rates.

It can be seen from Figure 4 that the accuracy rate of the enhanced-XGboost method is higher than that of other algorithms, and there is a certain gap. The accuracy rate of Enhanced-Logistic is almost the same as that of the XGboost model, but the accuracy of the Logistic optimization method is much higher than that of the Logistic model.

This time the accuracy of the logistic algorithm is at the lowest level. Because some studies have shown that the complete rate is inversely proportional to the accuracy rate, it is very difficult to check the whole and check the accuracy to do a good job. It can be seen from the fact that the lookup rate of the Logistic method is higher than that of the XGboost model, and the accuracy rate of the XGboost model is higher than that of the Logistic method. However, the two optimized algorithms are not bad in terms of both the totality rate and the accuracy rate, and even the XGboost enhancement model maintains optimal performance in terms of the check rate and the accuracy rate, which testifys the feasibility of algorithm method optimization in improving the accuracy rate.

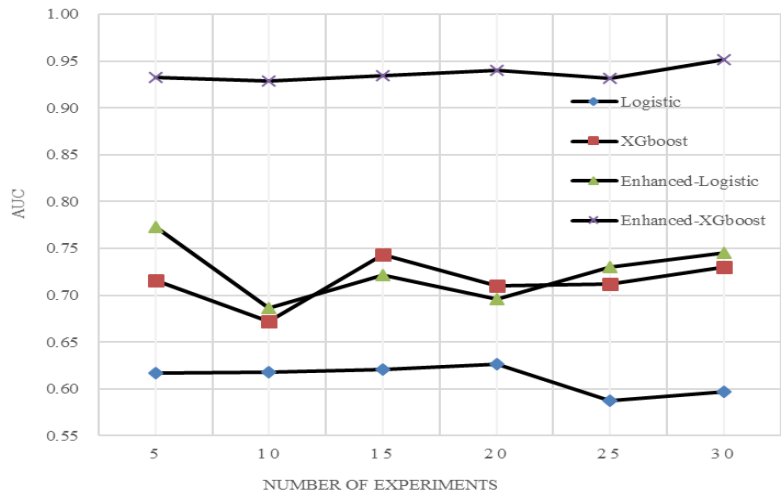

Fig. 5. Model AUC comparison.

As can be seen from Figure 5, the auc value of the Enhanced-XGboost algorithm is much higher than that of the other three methods, which demonstrates that the stability performance of the enhanced-XGboost model is the best. The auc value of the Enhanced-Logistic model is slightly higher than that of XGboost, but it is much higher than the auc value of the Logistic method. It can be seen that algorithm optimization also has an improving effect on the stability of the algorithm.

The following table is the average result after 30 rounds:

Table 2. Comparison table of model evaluation indicators.

Model	Accuracy	Recall	Precision	AUC Area
Logistic	0.9610	0.9885	0.9702	0.6171
XGboost	0.9558	0.9798	0.9742	0.6852
enhanced-Logistic	0.9704	0.9921	0.9775	0.7192
enhanced-XGboost	0.9908	0.9958	0.9946	0.9376

Simulate the same sample data in the same environment. From Table 2, we can see that the accuracy and complete rate of the Logistic model are slightly higher than XGboost, and the accuracy rate and auc value are slightly lower than XGboost. Enhanced-Logistic has improved from Logistic's accuracy rate of 0.9610 to 0.9704. The enhanced-XGboost has increased the accuracy rate from XGboost's 0.9558 to 0.9908, which is better than the improvement effect.

Comprehensive comparison can be found that the accuracy rate, complete rate, accuracy rate and auc value of the XGboost optimization method are the highest, especially the auc value has increased from 0.7192 of the logistic optimization algorithm with the better model to 0.9376, indicating that the XGboost improvement method has been in an extremely stable state, and it is far better than the other three algorithms from the perspective of the auc indicator. Based on the comprehensive comparison of the four models, the Enhanced-XGboost method performed best.

4.4. *Other Results*

By calculating the case sample data in this article, it is also possible to obtain indicators of interest to managers and regulators, as detailed in the following table.

Table 3. Relevant default rates.

Index	Proportion
The probability of default in the previous year and a further default in the following year	91.94%
The probability of default in the previous year and default in the next year	0.41%
The average proportion of defaulting enterprises	4.10%

From the above table, it can be seen that the probability of default in the previous year and the probability of continuing to default in the next year is as high as 91.94%. Regulators should pay attention to those companies that have defaulted in the previous year, and if they issue bonds in the same year, they should strictly review

their operating conditions and financial status to determine whether they are truly capable of repaying the bonds; Investors should be wary of issuers that have defaulted in the previous year, and although their reissued bonds may have a higher interest rate, they should also consider whether the risk is too high. Issuers who have defaulted on the bonds of the previous year need to measure their ability to repay the bonds before the next issuance, so as not to damage the credibility of the company due to the failure to repay.

The probability of defaulting in the previous year is only 0.41%, which can be seen that the companies that did not default in the previous year and the default in the next year accounted for a small proportion, and it is difficult to accurately predict whether the next year will default from many issuers who did not default in the previous year. The average proportion of defaulting enterprises is 4.10%, and regulators need to focus on reducing the ratio of 4.10%, strictly review bond issuance, and check whether the company has the ability to repay the bonds. Investors also need to polish their eyes and try to choose issuers with low default risk, always keeping in mind the theory that the higher the profit, the higher the risk.

V. Conclusions and Implications

Empirical research confirms that Logistic and XGboost algorithms can be used to predict the risk of credit debt default. Adding time series data to the Logistic and XGboost models, the model can continue to improve to the highest 99.08% even if it has reached 95%-96% accuracy, which fully demonstrates the superiority of its algorithm model improvement, especially the algorithm model can eventually achieve 93.76% stability, which shows that the application of the algorithm optimization method proposed in this paper is effective. Among them, the performance of the improved enhanced-XGboost model is better than that of the enhanced-logistic model.

The bond market is still an important part of China’s economic market, in view of the phenomenon of repeated defaults of bonds, the use of the above four algorithm models to analyze whether the bond has a default risk, not only can reasonably guarantee the pricing power of the company's bonds, investors can also compare the income and risk of each bond, to choose the bond that is most in line with the expectations of the heart for investment, no longer rely solely on credit rating. Avoid bonds that are likely to default and recover losses for investors. Auditors should also pay full attention to the issuers of bonds that defaulted in the previous year, be vigilant against their defaults in the next year, and make full use of algorithmic models to predict the risk of credit bond defaults.

Of course, investors and issuers, regulators, and auditors can use the optimal enhanced-XGboost model of the four models to analyze whether bonds will default.

In the future, the amount of data will become more and more large, and the accuracy of model calculations will also increase.

The machine learning algorithm provides a new idea for the future research on credit debt default, and the algorithm model optimization method proposed in this paper optimizes the algorithm model and provides a new method for the improvement of the model.

Acknowledgements

This study is supported by the National Social Science Fund of China. (No. 17BGL047).

REFERENCES

[1] Shijie Zheng. Research on the default risk of credit bond issuers based on ANN-KMV [D]. Southeast University, 2021.
[2] Yiqun Wang. Research on credit debt risk of Chinese listed companies based on KMV-LOGIT mixed model [D]. East China University of Political Science and Law, 2021.
[3] Ping Cao. Analysis of default risk of local government bonds based on KMV model [J]. Securities Market Herald, 2015(08): 39-44.
[4] Yinong Zhu. Credit bond default risk measurement based on KMV-random forest model [D]. Shandong University, 2019.
[5] Tianxin Zhang. Research on early warning model of bond default risk based on machine learning [D]. Nanjing University, 2020.
[6] Die Hu. Analysis of bond default based on Random Forest [J]. Contemporary Economics, 2018(03):28-30.
[7] Qinghong Zhang. Prediction and evaluation of credit bond default in China based on Stochastic Forest Algorithm [D]. Nanjing University, 2020.
[8] Peipei Zhang. Risk analysis of credit bond default of listed companies based on modified KMV-LOGIT Model [D]. Guizhou University of Finance and Economics, 2021.
[9] Kesu Wang. Research on corporate bond credit risk assessment scheme based on combination algorithm [D]. Shanghai Normal University, 2021.
[10] Chang Yan, Jungang Xu. How to dynamically evaluate the default probability of enterprises? - Research on Logit model based on China's credit bond market [J]. Financial Market Research, 2018(01): 124-133.
[11] Xuebin Chen, Wu Jing, et al. Measurement and prevention of individual default risk of credit bonds in my country: Based on LSTM Deep Learning Model [J]. Fudan Journal (Social Science Edition), 2021, 63(03): 159-173.
[12] Aoxue Dong. Current situation of credit debt default in my country and construction of risk early warning model based on Logistic Multiple Regression [D]. Shanghai Jiaotong University, 2019.
[13] Xuguang Wang. Research on the factors affecting the issuance scale of credit bonds-Empirical analysis based on VAR model [J]. Bonds, 2020(02):68-74.
[14] Qiangwei Feng. Research on default risk prediction of credit debt based on XGboost algorithm [D]. University of Electronic Science and Technology of China, 2020.
[15] Rongxi Zhou, Hang Peng, et al. Credit debt default prediction model based on XGboost algorithm [J]. Bonds, 2019(10):61-68.
[16] Yao Qiu, Guowei Yang. User behavior prediction and risk analysis based on XGboost algorithm [J]. Industrial Control Computer, 2018, 31(09): 44-45.
[17] Yaqing Feng. Credit debt risk measurement based on improved KMV-XGboost [D]. Shandong University, 2020.
[18] Chunhua Ju, Guanyu Chen, Fuguang Bao. Consumer finance risk detection model based on kNN-Smote-LSTM-taking credit card fraud detection as an example [J]. System Science and Mathematics, 2021, 41(02): 481-498.

AUTHOR'S PROFILE

First Author
Dr. Xiangrong Shi, received the Ph.D. degree from the department of Control Science and Engineering, Zhejiang University, in 2014. He currently is a Professor with the School of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, China. His research interests include educational management, data mining and artificial intelligence.

Second Author
Qi Zheng, Corresponding author, from Zhejiang, China, zhejiang university of finance and economics graduate first year, auditing major, supervisor is Dr. Xiangrong Shi, currently mainly researching the application of machine learning algorithms in auditing, participated in the whole process of thesis guidance and writing.

Third Author
Pengsai Guo, from Henan, China, is now a first-year graduate student of Zhejiang University of Finance and Economics, majoring in auditing, learning the direction of big data auditing with his mentor, and hoping to become an investor in the future.

Fourth Author
Yifei Ye, from Zhejiang, China, is a second-year graduate student at Zhejiang University of Finance and Economics, majoring in auditing, and follows his mentor to the direction of big data auditing. In the future, he hopes to become a computer teacher.