ISSN 2616-5902 Vol. 4, Issue 7: 88-92, DOI: 10.25236/AJBM.2022.040715

Research on the Credit of High-Yield Corporate Bonds in My country Under Machine Learning

Longyifei Ye*, Haolong Wu, Xiuwen Duan

City College, Zhejiang University, Hangzhou, Zhejiang, 310000, China
**Corresponding author*

Abstract: *As an important part of the bond market, corporate bonds have the characteristics of high yield and high risk. This work aims to design an index system of influencing factors through the literature research on the current situation of corporate bond defaults, credit risk related theories and credit risk model theories, using corporate bond default data, and integrate machine learning models and KMV optimization models to construct the credit risk early warning model of my country's high-yield corporate bonds. Investors, regulators and other relevant parties can use this early warning model to more effectively pay attention to the potential default risk of corporate bonds, and carry out corresponding credit risk control to ensure the stable development of the bond market.*

Keywords: *High-yield bonds, Support vector machine, KMV-Logistic model, Bond default*

1. Introduction

In recent years, my country has vigorously developed direct financing and continuously improved the financing function of the capital market. The bond market is an important part of the financial market. The rapid development of the bond market has also brought about a series of credit default problems. 2014 was a year of vigorous development of China's bond market, but also a year of rising risks and further efforts to regulate the bond market. The current domestic economic environment has been greatly affected by the impact of the new crown epidemic. My country has suffered a huge impact in the fields of production and consumption, resulting in huge pressure on the cash flow of many domestic companies, broken capital chains, and the risk of default on issued corporate bonds. Therefore, the risk of corporate bonds has increasingly attracted the attention and attention of investors and regulators, and has also become an important part of securities market risk research. The market needs a relatively complete risk early warning system to make decisions for investors and regulators.

In the past research on bond defaults, it can be concluded that there are many factors that cause bond defaults. National macroeconomics, corporate governance, and financial conditions will affect the credit risk of the issuer. In terms of the macro factors that lead to the occurrence of default events, macroeconomic variables have a prominent contribution to explaining the default events of the issuer [1]. Financial indicators are a direct reflection of a company’s operating conditions and are closely related to the repayment of debt. The author studies the relationship between bond credit spreads and system and company-level factors, and finds that short-term interest rates have a good explanatory power for bond credit spreads [2]. Changes in the macroeconomic environment also have an impact on the credit status of enterprises. There is a correlation between macro factors such as GDP growth rate and money supply and the credit risk of enterprises [6].

An effective corporate governance mechanism can reduce the agency cost of the company, and at the same time reduce the degree of information asymmetry between the company and the borrower, thereby reducing the risk of bond default [3]. The relationship between the corporate credit of listed companies in my country and the company's shareholding structure and the governance level of the board of directors is studied [4]. Wang Dongjing conducted the theoretical derivation of default probability and found that the higher the company's return volatility and debt ratio, the greater the default risk [5]. Yu Xinyuan selected the influencing factors of bond default through stepwise regression of logistic model [7].

In the context of frequent bond defaults, the bond default early warning model plays an important role. In foreign research, Beaver first proposed the univariate discriminant analysis method, which has a certain prediction effect [8]. Altman first used the multivariate discriminant analysis model to conduct quantitative research, and used the multivariate discriminant analysis method to construct a Z-score

ISSN 2616-5902 Vol. 4, Issue 7: 88-92, DOI: 10.25236/AJBM.2022.040715

model for the US debt distressed enterprises [9]. Ohlson first used the logistic model to predict the company's financial risk, and achieved a high prediction accuracy [10]. Compared with statistical analysis methods, when predicting bond default risk, machine learning has also been applied to the financial field by virtue of its good feature extraction for data, neural network and support vector machine methods. Quantitative analysis based on machine learning models has become one of the mainstream trends in risk research. Empirical research on bond defaulting companies shows that using the SVM algorithm has a high accuracy [11]-[14].

2. Model Establishment

This paper selects 46 listed companies with substantial default in corporate bonds from 2018 to 2021 as default samples. At the same time, according to the ratio of default samples to non-default samples is 1:3, 138 companies that have not defaulted were selected as the empirical samples of healthy companies. The default sample data of the model in this paper are the data in the latest annual or semi-annual report before the actual default of the defaulting company, and the paired non-defaulting companies and defaulting companies use the same period of annual or semi-annual report data.

Based on the previous researches on bond default characteristic indicators, this paper divides the forecast indicators into four parts: financial indicators, non-financial indicators, market rating indicators and macroeconomic indicators. In the financial indicators, in addition to the default distance DD, there are four indicator systems: profitability, solvency, operating capacity, and growth capacity.

2.1. Data preprocessing

Since the dimensional difference between the various features of the sample data will affect the prediction accuracy of the classifier, the data is normalized, and the commonly used Min-Max normalization method is used, which can be expressed as:

$$X_i' = \frac{X_i - \mathrm{Min}(x)}{Max(x) - \mathrm{Min}(x)} \tag{1}$$

2.2. KMV model

In this paper, the debt term T is set to 1 year, and the risk-free interest rate, the book value of the company's debt D and the financial data of the listed company are obtained through the wind database. Since some listed companies in my country still have non-tradable shares, the value of non-tradable shares is calculated according to the relevant correction methods, namely:

Company equity value V_e = value of tradable shares + value of non-tradable shares = annual closing price * number of tradable shares + net assets per share * number of non-tradable shares

Then, the corresponding daily closing price data of the sample companies taken from the wind database are converted into logarithmic returns, and then the GARCH (1, 1) model is used to estimate the stock price volatility δ_D of each company, and then calculate the company's equity value volatility σ_E, which can be expressed as:

$$\sigma_E = \delta_E = \delta_D\sqrt{n} \tag{2}$$

According to the KMV model, the default distance DD of the full sample data is calculated with the help of Matlab software, and the calculation results are shown in table 1.

Table 1: Calculation results

Model results	Experimental group	Control group
Mean	2.9522	1.8974
Sstandard deviation	0.4729	0.8461
Minimum	0.6871	0.0013
Maximum	5.2514	3.2941

Using the default distance (DD) calculated by the KMV model to reflect market expectations, it can determine the size of the issuer's bond default risk to a certain extent. According to the calculation of the distance formula for default, the larger the DD, the smaller the corresponding default risk; otherwise, the default risk will increase. It can be seen from the results that the average default distance of the experimental group is significantly higher than that of the control group, indicating that its default risk is

ISSN 2616-5902 Vol. 4, Issue 7: 88-92, DOI: 10.25236/AJBM.2022.040715

higher than that of the control group, which further proves the effectiveness of the model in measuring corporate credit risk.

2.3. Model construction:

In order to find the most suitable model, the set-out method is used to divide the training set and the test set according to a certain proportion. This paper selects 70% of the samples in the data set as the training set, and the remaining 30% as the test set, and adopts the decision tree model, the logistic model, the BP neural network model and the random forest model for modeling. The specific method is as follows: 70% of the 45 default sample data and 138 non-default sample data are randomly selected, and a total of 129 sample data are obtained to form a training set for training the prediction model; the remaining 30%, that is, 55 sample data, which constitutes the test set for testing the model.

3. Model Efficacy Evaluation

3.1. Evaluation and selection of model performance indicators

When measuring the performance of a two-class machine learning model, in addition to the overall classification accuracy of the sample, it is also necessary to compare the precision rate and the recall rate. In the binary classifier, the above indicators can be obtained by calculating the confusion matrix, as shown in table 2.

Table 2: Indicator results

Actual Results	Forecast Results		Total
	Default state	Normal state	
Default state	TN	FP	TN+FP
Normal state	FN	TP	FN+TP
Total	TN+FN	FP+TP	TN+FP+FN+TP

The formula for calculating the precise is:

$$A = \frac{TP+TN}{TN+FP+FN+TP} \tag{3}$$

The formula for calculating the accuracy is:

$$P = \frac{TP}{FP+TP} \tag{4}$$

The formula for calculating recall is:

$$R = \frac{TP}{TP+FN} \tag{5}$$

Generally speaking, there is some opposition between precision and recall, i.e. when the precision of a machine learning model is higher, the recall is usually lower. For example, if the proportion of samples with positive prediction results is to be higher, the machine learning model may encounter difficulties in the screening process and lose some positive samples, and the model recall rate is low. The F1 metric is the harmonic mean of machine learning model accuracy and recall, which can be expressed as:

$$F1 = \frac{2\times P\times R}{P+R} \tag{6}$$

ROC curves can also be used to measure the predictive power of machine learning models. For each forecast, the FPR and TPR indices can be calculated separately in order to draw the ROC curve, which can be expressed as:

$$FPR = \frac{FP}{TN+FP} \tag{7}$$

$$TPR = \frac{TP}{TP+FN} \tag{8}$$

Both FPR and TPR indicators calculate the corresponding probability from the actual positive and negative samples, so they are not affected by the proportion of positive and negative samples, and have better application to unbalanced samples. In the ROC curve, higher TPR and lower FPR represent the best performance of the machine learning model, that is, the higher the slope of the ROC curve, the better. The area under the curve (AUC) under the ROC curve can be used to measure the performance of a machine learning model. The AUC value measures the classification quality of the sample prediction

ISSN 2616-5902 Vol. 4, Issue 7: 88-92, DOI: 10.25236/AJBM.2022.040715

results. A higher AUC value means a higher classification quality of the sample prediction results. When the AUC value is 1, it means that the model prediction results classify all positive samples before negative samples. Generally speaking, machine learning models have an AUC value between 0.5 and 1, and an AUC value of 0.5 means there is no difference between the model and random judgment. When the AUC value is greater than 0.9, the model performance is better.

3.2. Analysis of model results

Table 3: Model results

Model	Model results			
	A	$F1$	TPR	AUC
Decision tree	78.18%	85.37%	85.37%	0.86
Random forest	87.27%	91.76%	95.12%	0.92
BP neural network	92.27%	95.00%	92.68%	0.97
Logistics returns	85.96%	90.70%	92.86%	0.91

It can be seen from the table 3 that the model accuracy of decision tree is low, while the accuracy of Logistics regression, BP neural network and random forest are all above 85%, the highest accuracy of BP neural network is 92.27%, and the F1 value and AUC The value is also the highest. On the whole, the BP neural network has the best prediction performance and is suitable for building a bond default risk early warning model.

Secondly, the accuracy rate of random forest has also reached 87.27%, and it also has a good reference for prediction. The random forest based on the step-by-step backward evaluation method has a higher contribution to the selection of important indicators. It evaluates and selects the characteristic index system. , extract important feature indicators and use them as the input of the prediction model, which greatly improves the learning efficiency and applicability of the model. In this paper, 25 eigenvalues are used to evaluate the importance of eigenvalues in a backward step-by-step evaluation method through random forest algorithm, and 8 important feature indexes with high influence are obtained, of which 6 indexes belong to financial factors, reflecting the main body of bonds. Its profitability is the main identification ability of its default behavior, the macroeconomic indicators are not important, and the importance of the default distance is also the third, which confirms the validity of the KMV model. The important indicators of random forest screening and their importance scores are shown in table4.

Table 4: Important indicators and importance score table

Feature indicator name	Experimental group
Roe	21.67
Whether it is a state-owned enterprise	17.61
Default distance	15.74
Operating income growth rate	14.32
Shareholding ratio of top ten shareholders	9.12
Total asset turnover	8.17
Assets and liabilities	7.56
Quick ratio	5.81
Total	100.0

4. Conclusion

This paper summarizes the influencing factors of bond default into three levels: the company's external macroeconomics, industry influence, national policy, the company's internal financial and non-financial factors, and the characteristics of the bond itself, and selects the listed companies with substantial corporate bond defaults from 2018 to 2021. As the research object, compared with similar corporate bonds that have not defaulted, the following conclusions are drawn through empirical analysis by establishing a KMV-machine learning model:

Firstly, the corporate bond default prediction model based on BP neural network has good prediction results. According to the sample one year ahead of time, the prediction accuracy of whether the corporate bond will default in the current year is as high as 92.27%, and the probability of error is higher than that of the previous year.

Secondly, the index system after screening by the random forest model is more reasonable, and can

ISSN 2616-5902 Vol. 4, Issue 7: 88-92, DOI: 10.25236/AJBM.2022.040715

filter out some of the eigenvalues that have no significant impact on corporate bond defaults in many original predictors, leaving the eigenvalues with a large enough impact. Make the model converge faster, easy to calculate and easy to generalize. In this paper, 25 eigenvalues are sorted according to their importance through the random forest model, and 8 eigenvalues with significant influence are obtained. A relatively streamlined and efficient index system is constructed, which simplifies the computational workload and makes the model more practical.

Finally, there are many factors that affect whether corporate bonds will default, and the default distance obtained by the KMV model is also an important factor. Among them, there are 6 financial indicators in the top 8 eigenvalues of the random forest model screening results, so the internal financial indicators of the company are the most important factors affecting whether the corporate bond defaults.

Acknowledgements

Supported by the National College Students Innovation and Entrepreneurship Training Program 202113021036

References

[1] Bonfim D. Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics [J]. Journal of banking & finance, 2009, 33(2): 281-299.
[2] Bakshi G, Madan D, Zhang F X. Investigating the role of systematic and firm-specific factors in default risk: Lessons from empirically evaluating credit risk models [J]. The Journal of Business, 2006, 79(4): 1955-1987.
[3] Bhojraj S, Sengupta P. Effect of corporate governance on bond ratings and yields: The role of institutional investors and outside directors [J]. The journal of Business, 2003, 76(3): 455-475.
[4] Li W, Li J. Equity, board governance and corporate credit of Chinese listed companies [J]. Management World, 2003 (9): 103-109.
[5] Wang D, Zhang X, Zhang J. Corporate debt term structure and default risk [J]. Journal of Management Science, 2009, 12(2): 77-87.
[6] Zhou H Yang M, Li Y A review of research on influencing factors of corporate bond credit risk [J]. Economics Dynamics, 2010 (12): 137-140.
[7] Yu X. Research on the influencing factors of corporate bond default based on logistic model [J]. Times Finance, 2017 (18): 194.
[8] Beaver W H. Financial ratios as predictors of failure [J]. Journal of accounting research, 1966: 71-111.
[9] Altman E I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy [J]. The journal of finance, 1968, 23(4): 589-609.
[10] Ohlson J A. Financial ratios and the probabilistic prediction of bankruptcy [J]. Journal of accounting research, 1980: 109-131.
[11] Zhang L, Hu H, Zhang D. A credit risk assessment model based on SVM for small and medium enterprises in supply chain finance [J]. Financial Innovation, 2015, 1(1): 1-21.
[12] Tian J W, Wu K J, Zhuo Z G, et al. Application of support vector machine and Logistic regression model in personal credit prediction [J]. Regional Finance Research, 2018 (11): 25-30.
[13] Shi Y L, Zhang B J, Jiang H. Research on risk identification of P2P online lending platforms [J]. Statistics and Decision, 2018, 16.
[14] Zhao D D, Ding J C. Research on systemic risk early warning in China's banking industry——Modeling analysis based on SVM model [J]. International Business: Journal of University of International Business and Economics, 2019 (4): 100-113.