Forecasting default risk of Chinese real estate credit bonds based on Bagging algorithm

Xuan Liu *

Department of Financial Technology, Shanghai Lixin University of Accounting and Finance

* Corresponding author: 13995004436@163.com

Abstract. In recent years, with Chinese rapid economic upgrading, it has also made the property industry more and more prosperous. However, due to the imbalance between supply and demand in China's real estate industry in recent years has led to an overly rapid rise in property prices, the risk of a real estate market bubble has increased, and many real estate companies have gradually defaulted on their bonds. Against this background, the aim of this paper is to construct a bond default prediction model using the Bagging algorithm in machine learning to predict the default risk of real estate bonds. And the results were also compared with Logistic model and SVM model, eventually it was found that Bagging algorithm can improve the prediction results more effectively. It is possible to identify real estate companies that may default on their bonds in advance, reducing investor losses, strengthening the risk prevention awareness of debt issuers, contributing to the healthy and stable development of real estate.

Keywords: bond default, real estate, Bagging, prediction.

1. Introduction

Chinese rapid economic development in recent decades has largely improved people's quality of life and promoted urbanization, resulting in a booming real estate industry. The real estate industry, as an important pillar of the national economy, has a pivotal role in Chinese economic development, it pulls many industries upstream and downstream to prosper together. Since 2013, Chinese real estate development investment has accumulated 103 trillion yuan, with an average annual growth rate of 9.2%, in 2021, the share of real estate development investment in all investment was 27.1%. However, the risk of a real estate market bubble has increased in recent years due to the imbalance between supply and demand in Chinese real estate sector, which has led to an overly rapid rise in housing prices. The government has introduced several regulatory policies for the real estate market to adhere to the "housing without speculation" and effectively prevent risks in the industry. Under the pressure of a series of policies to deleverage the property business, many illiquid and highly indebted property companies will be exposed to higher financial risks, and the risk of bond defaults by real estate companies has also been gradually exposed after rigid payment was broken, a total of 10 real estate companies defaulted on their bonds in 2022. In this context, an in-depth study of the inducing factors of bond defaults by real estate companies is needed, constructing a bond default prediction model to identify in advance bonds in the real estate sector that are likely to default, minimizing investor losses and enhancing risk awareness among debt issuers, providing decision support for the healthy and stable development of the real estate industry.

In this paper, Bagging, a machine learning algorithm, is introduced into the construction of a bond default prediction model, to predict the default risk of real estate bonds, and compare the prediction results with Logistic model and SVM model. It is found that the application of Bagging algorithm in this paper can effectively improve the prediction effect. Based on this, relevant decision-making suggestions are made to property companies and regulators.

The following section is structured as follows: the second part is the literature review, the third part is the model and data, the fourth part is the empirical results, and finally, the conclusions and recommendations.

2. Literature Review

In the field of bond default prediction, there is a large amount of academic research nowadays. Landschoot [1] (2004) found that the higher the stock market volatility, the higher the risk of bond default. Lee [2] (2010) analyzed the causes of bond defaults and concludes that when corporate governance is poor, it is less able to cope with risk and is vulnerable to the external environment, which in turn leads to bond defaults. Giesecke et al. [3] (2011) show that there is a significant correlation between GDP growth and bond default rates through nearly 150 years of data on US bonds. A study by Kuehn and Schmid [4] (2014) confirms that macroeconomic conditions are negatively associated with bond default risk. Hsu [5] (2015) conducted a study on the volatility and risk profile of firms' cash flows and found that it has a positive relationship with the default risk of the firms themselves. Kiesel [6] (2016) followed the logic of reputation theory to conclude that credit rating agencies will be challenged by the market for stepping on mines and that the impact of their published ratings information on market spreads will diminish. Through the above literature, it can be found that bond characteristic factors, bond issuer characteristic factors, and external macro-environment characteristic factors may affect bond default.

With the development of machine learning algorithms, it has been widely used in risk modelling, and risk assessment and risk prediction have good results. Khandani et al. [7] (2010) applied the RDF (Resource Description Framework) algorithm to predict consumer defaults, which effectively reduced lenders' losses. Barboza et al. [8] (2017) predict corporate defaults in North America and find that machine learning algorithms provide an improvement of about 10 percentage points in AuROC compared to traditional models. Kristina[9] (2023) identifies a non-linear relationship between predictors and crisis risk by identifying economic drivers of machine learning models.

The Bagging model used in this paper has also been applied to forecasting studies in medicine, finance and other fields, showing better results. Wang et al. [10] (2021) used Bagging algorithm and Boosting algorithm to predict mine rock bursts and finally concluded that the best prediction method for rock bursts potential is Bagging. Alam.et al. [11] (2021) used Bagging integrated classifier for birth pattern prediction in Bangladesh and used four traditional machine learning algorithms for comparison (Decision Tree, k-Nearest Neighbor, NB, SVM), which showed that Bagging integrated model outperforms the traditional model in this domain. Wang et al. [12] (2009) designed a financial statement analysis method using decision trees to select 50 financial ratios to predict the direction of change in earnings one year ahead. In order to improve the classification accuracy of the decision tree, Bagging technique is introduced and the results show that the Bagging model is better in stock return prediction compared to the standard decision tree model and the boosted decision tree model. Therefore, this paper innovatively applies the Bagging model to the default prediction of bonds in Chinese real estate industry, which is used to discover the inducing factors affecting the default of real estate bonds and provide decision support for the risk management of bond investment and financing and the formulation of regulatory policies.

3. Models and Data

3.1. Bagging model

Bagging is an algorithm based on integrated learning, where the basic idea is to perform random sampling of the training set with put-back(bootstrap sampling)to obtain multiple distinct training subsets. Then some kind of learning algorithm is used to obtain multiple base classifiers with large variability on different training sets, which are combined after grouping to obtain a high-precision classification model. Specifically speaking, the Bagging algorithm is as follows:

1. Let i=1, 2, 3... ,n. Random sampling with putbacks is performed on the training set to generate a self-help sample set T; Since there is put-back random sampling, some samples are repeated and others are not. This creates variability in the set of n self-help samples.

2. Construct base classifiers h_i,(i=1,2,3…,n) using unstable classification algorithms on self-service sample sets T_i (i=1, 2, 3..,n).

3. Record the prediction categories of the n base classifiers for the test sample x, use balloting to determine the final category of x. H(X) =argmax $(\sum_{i=1}^{T} h_i(x)) = y$

In order to compare the effectiveness of Bagging model, Logistic and SVM are also used as benchmark models in this paper. Logistic regression is a classical forecasting method based on the principle that the results of linear regression are brought into a Sigmoid function, which converts a continuous variable into a probability value in the interval 0-1. When the probability is greater than 0.5, the sample is classified as positive, and when the probability is less than 0.5, the sample is classified as negative. The Sigmoid function is shown in equation (1):

$$F(x)=\frac{1}{1+e^{-x}} \tag{1}$$

The SVM model is a binary classification model, a theory that specializes in the study of the laws of machine learning in the case of small samples. Developed from the optimal classification hyperplane in the linearly divisible case, its basic model is defined as a linear classifier with maximum intervals on the feature space, which uses interval maximization to find the optimal classification hyperplane.

The real estate bond default prediction in this paper is a classification problem, For model training, 80% of the samples were selected as the training set and 20% of the samples were selected as the test set for all types of models.

For the evaluation of model prediction effectiveness, this paper considers the performance of the models in terms of two types of errors. Comparisons were made using the indexs TP (True Positive) and FP (False Positive), Three metrics, precision (precision=$\frac{TP}{TP+FP}$), recall (recall=$\frac{TP}{TP+FN}$), and ROC were also used for model comparison.

3.2. Data

Based on the previous analysis, the factors affecting bond defaults adopted in this paper include three categories: bond characteristics, issuer characteristics and macro-environmental characteristics. Among other things, bond characteristics include: maturity of issue, total amount issued, and credit rating; Characteristics of the issuer include: whether it is a listed company, nature of the company, return on equity (ROE), current ratio, and gearing ratio; Macro-environmental characteristics include: GDP growth rate. The data sample relates to corporate bonds issued in the real estate sector between 2016 and July 2023, and contains a total of 79 samples of defaulted bonds and 892 samples of non-defaulted bonds after excluding samples with missing data. The data source is Choice database. Detailed descriptions of the relevant variables and descriptive statistics of the data are shown in Table 1.

Table 1. Descriptive statistics of variables

Variable name	Description of variables	Sample range	Average	Standard deviation	Maximum	Minimum
Term of Issue	Principal repayment period determined at the time of bond issuance (year)	Full Sample	4.5260	1.7326	15.4849	0.4932
		Sample Defaults	5.1909	2.9260	15.4849	0.5863
		Non-Defaults	4.4425	1.5000	10.0000	0.4932
Aggregate amount Issued	Total amount of funds expected to be raised at the time of bond issuance (billion)	Full Sample	12.3753	9.4179	150.0000	0.5000
		Sample Defaults	15.4165	17.4184	150.0000	1.0000
		Non-Defaults	11.9935	7.7936	72.0000	0.5000
ROE	Return on average net assets	Full Sample	4.1994	78.6104	43.2332	-2386.089
		Sample Defaults	-35.1316	42.0600	14.7813	-139.1131
		Non-Defaults	9.1378	80.7287	43.2332	-2386.089
Current Ratio	Ratio of current assets to current liabilities	Full Sample	4.0860	40.4196	741.7543	0.0000
		Sample Defaults	0.8181	0.5492	2.0181	0.0000
		Non-Defaults	4.4963	42.8668	741.7543	0.3567
Asset-liability Ratio	Ratio of owner's equity to total liabilities of the enterprise	Full Sample	69.7036	17.7066	100.1501	0.0000
		Sample Defaults	59.0388	38.3117	100.1501	0.0000
		Non-Defaults	71.0427	12.4091	91.9026	0.1347
GDP growth Rate	Annual GDP growth rate	Full Sample	7.1539	1.0614	8.6000	1.7000
		Sample Defaults	5.9902	2.9398	8.6000	1.7000
		Non-Defaults	7.3000	0.0000	7.3000	7.3000

Table 1 shows that there are significant differences in certain financial indicators between firms that have defaulted on their bonds and those that have not. For example, real estate firms that have defaulted on their bonds have higher average gearing ratios and have a negative average return on average net assets, whereas non-defaulting firms have a positive average return on average net assets.

4. Empirical Findings

Due to the small proportion of defaulted samples in the full sample, which is typical of unbalanced data, 158 non-defaulted data were randomly selected in a 2:1 ratio with the defaulted samples for model training. And compare the prediction effect of Bagging model with two benchmark models, Logistic and SVM, the comparison of model prediction evaluation indexes is shown in Table 2 and Figure 1. It can be found that the Bagging model has a greater improvement over the Logistic and

SVM models, both in terms of prediction accuracy and in terms of the area under the model ROC curve. In terms of precision, the Bagging model is 0.95, while the SVM model is the lowest at 0.863. The Bagging model is also much higher than the SVM and Logistic models in terms of recall, with a recall of 0.949. For TP values, Bagging model TP value is 0.949, Logistic model TP value is 0.873 and SVM model TP value is 0.865. Thus, it appears that the Bagging model predicts more accurately.

Table 2. Comparison of TP and FP for the three models

	Logistic	SVM	Bagging
TP	0.873	0.865	0.949
FP	0.165	0.188	0.082

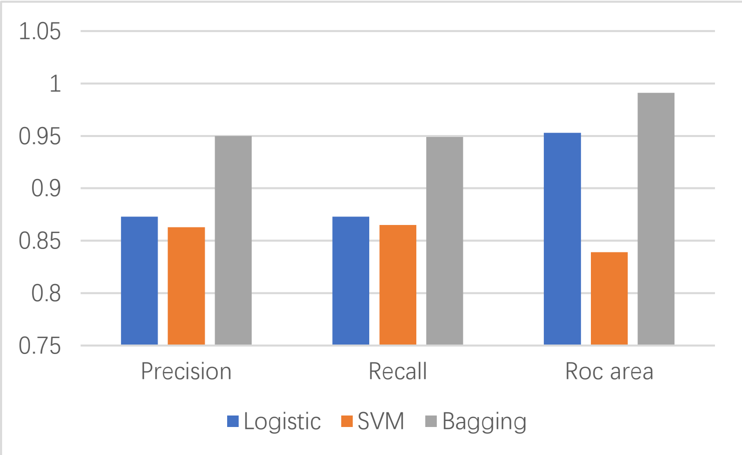

Figure 1. Comparison of precision, recall, and ROC of the three models

In summary, the SVM model, Logistic model and Bagging model are used to predict the default rate of China's real estate bonds respectively, and it can be found that the Bagging model predicts the highest indexes, which is more suitable for the default prediction analysis of real estate bonds compared with other benchmark models.

5. Conclusions and recommendations

This paper takes Chinese real estate enterprises as the research object, constructs the indicator system according to their bond characteristics, issuer characteristics and macro-environmental characteristics, and builds a bond default prediction model through the Bagging algorithm in machine learning. The main conclusions of this paper through data analysis and model predictions are as follows: First, a comparison of data from the default and non-default samples of real estate bonds reveals that there are significant differences in certain financial indicators between the corresponding issuing firms. For example, real estate firms that have defaulted on their bonds have higher average gearing ratios and have a negative average return on average net assets (ROE), whereas non-defaulting firms have a positive average return on average net assets (ROE). Secondly, compared to Logistic and SVM models, the Bagging model selected in this paper can better predict the occurrence of bond defaults by real estate firms, with a model accuracy of 94.9367% and a ROC curve coverage area of 0.995.

Based on the analyses in this paper, the following policy recommendations can be obtained: Firstly, in terms of regulation, regulators can reduce the information asymmetry between investors and bond-issuing entities by using artificial intelligence algorithms, establishing a big data platform, collecting and integrating relevant data, and regularly releasing analysis reports on the financial status of bond-issuing entities to reduce the information asymmetry between investors and bond-issuing entities, and establishing a model with machine learning algorithms for real-time prediction of the probability of default of corporate bonds; Second, the credit ratings of Chinese credit bonds are generally high, but these ratings do not truly reflect the default risk of the bonds, and their usefulness is limited. Therefore, China's credit rating industry needs to rebuild its credibility, while the regulator should strengthen its supervision in order to improve the quality of the ratings, and further regulate the behavior of the rating agencies; Thirdly, as the real estate industry itself has the attribute of high leverage and relies heavily on external financing, real estate enterprises should pay more attention to the management of the company's leverage ratio, appropriately reduce the leverage ratio within a controllable range, and always focus on the company's debt structure, by reducing the scale of interest-bearing liabilities, applying diversified financing methods to reduce the dependence on non-standard financing, and choosing a reasonable development and operation mode suitable for the company, preventing the irrationality of its own debt structure to bring the company debt servicing pressure and refinancing pressure, which will lead to the company's inability to borrow new and repay the old at a later stage, into a financial crisis.

References

[1] Landschoot, V. Determinants of the euro term structure of credit spreads. Social Science Electronic Publishing, 2004, 132 (12): 1303 - 1315.

[2] Lee T S. Corporate Governance and Financial Distress: Evidence from Taiwan. Corporate Governance an International Review, 2010 (3): 23 - 36.

[3] Giesecke, Longstaff, Schaefer. Corporate Bond Default Risk: A150-year perspective. Journal of Financial Economics, 2011, 102 (2): 233 - 250.

[4] Kuehn L, Schmid L. Investment-Based Corporate Bond Pricing. Journal of Finance, 2014, 69 (6): 2741 - 2776.

[5] Lee H, Hsu P, Liur A. Corporate Innovation, Default Risk and Bond Pricing. Journal of Corporate Finance, 2015, 56 (9): 413 - 456.

[6] Kiesel, F. “Do Investors Still Rely on Credit Rating Agencies?Evidence from the Financial Crisis.”The Journal of Fixed Income, 2016, 25 (4), pp. 20 - 31.

[7] Khandani, A. E., A. J. Kim, and A. W. Lo, 2010, Consumer credit-risk models via machine-learning algorithms, Journal of Banking and Finance 34, 2767 – 2787.

[8] Barboza, F., H. Kimura, and E. Altman, 2017, Machine learning models and bankruptcy prediction, Expert Systems with Applications 83, 405 – 417.

[9] Kristina Bluwstein, Marcus Buckmann, Andreas Joseph, Sujit Kapadia, Özgür Şimşek, Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach, Journal of International Economics, 2023, 103773.

[10] Wang, Sm., Zhou, J., Li, Cq. et al. Rockburst prediction in hard rock mines developing bagging and boosting tree-based ensemble techniques. J. Cent. South Univ. 28, 527 – 542 (2021).

[11] M. S. Bin Alam, M. J. A. Patwary and M. Hassan, "Birth Mode Prediction Using Bagging Ensemble Classifier: A Case Study of Bangladesh," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 2021, pp. 95 - 99.

[12] H. Wang, Y. Jiang and H. Wang, "Stock return prediction based on Bagging-decision tree," 2009 IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2009), Nanjing, China, 2009, pp. 1575 - 1580.