A Study on Forecasting the Default Risk of Bond Based on XGboost Algorithm and Over-Sampling Method

Yan Zhang¹, Lin Chen²*

¹Investment Bank Department, Chengdu Agricultural and Commercial Bank, Chengdu, China
²School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, China

Email: *Chenlin2@uestc.edu.cn

Abstract

China’s bond market is an emerging market. The number of bond defaults has been increasing in recent years, but the data set is severely imbalanced. Based on financial data of total 6731 corporate bond issuers which 50 bond issuers had defaulted, this paper uses the XGboost algorithm and an Over-sampling method named SMOTE to predict the default of bond issuers. The results show that the XGboost algorithm has advantages over the traditional algorithm in processing imbalanced data, and SMOTE is one of the effective methods to deal with imbalanced samples. Then, this is an effective way to predict the default risk of bond issuers in an emerging market.

Keywords

XGboost, Default Risk, Bond Issuer, Imbalanced Data, SMOTE

1. Introduction

Corporate credit risk is one of the most key risks for financial institutions and investors. Besides two key theoretical models for evaluating credit risk, one of which is called structural mode (Black and Scholes, 1973; Merton, 1974; Black and Cox, 1976) and the other is called reduced model (Jarrow and Turnbull, 1995; Longstaff and Schwartz, 1995; Jarrow, Lando, and Turnbull, 1997; Duffee, 1999; Duffie and Singleton, 1999; Jarrow and Turnbull, 2000; Duffie and Lando, 2001). A lot of statistical models have been used to predict the corporate credit risk (Altman, 1968; Altman, Haldeman, and Narayanan, 1977; Martin, 1977; Ohlson, 1980; Kim and Expert, 1999; Sung, Chang, and Lee, 1999; Shah and Murtaza, 2000; Nanda and Pendharkar, 2001).
In recent years, a machine learning method named eXtreme Gradient Boosting (XGBoost) has been used to evaluate personal credit risk (Li et al., 2020; Ma et al., 2018; Xia, Liu, and Liu, 2017; Chang, Chang, and Wu, 2018). Since Chen proposed the eXtreme Gradient Boosting (XGBoost) algorithm in 2016 (Chen and Guestrin, 2016), it has been one of the most important, effective, and widely used machine learning methods. The XGBoost has been used in many fields such as disease diagnosis (Liu et al., 2021; Zhang et al. 2020b; Zhang, Deng, and Jia, 2020a; Zhang and Gong, 2020), image recognition (Zhao, Guo et al., 2020; Zhao, Zeng et al., 2020), traffic flow prediction (Zhang and Zhang, 2020), sports analytics (Yigit, Samak, and Kaya, 2020), power system (Xue and Wu, 2020; Raichura, Chothani, and Patel, 2021; Li et al., 2018), e-commerce (Song and Liu, 2020), personal credit (Li et al., 2020; Ma et al., 2018; Xia, Liu, and Liu, 2017; Chang, Chang, and Wu, 2018) forecasting of weather (Jin et al., 2020; Fan et al., 2020; Zheng and Wu, 2019), public security (Feng et al., 2020), fault detection (Zhang et al., 2018; Chen et al., 2019; Lei et al., 2019; Lin et al., 2019) and so on. Those studies show that XGBoost can improve the prediction accuracy better than other statistical methods.

As of December 31, 2020, there are 57,039 issued bonds in the China bond market, and the value of bonds has reached 114.33 trillion yuan. Since 2014, the number of default bonds has increased year by year. There were 145 default bonds in 2020, and the face values of default bonds were 164.7 billion. The default in the bond market caused huge losses to investors, and there are new features of default bonds; for example, many local and central state-owned corporates’ bonds default and high-rating bonds default. So, this paper aims to use XGBoost to forecast the default of bond issuer in China. Moreover, the number of default bonds is still small relative to all samples, the imbalanced data classification is the biggest challenge for modeling. Many Resampling techniques have been developed in the past two decades to cope with imbalanced data classification. Those techniques fall into three groups: Over-sampling methods, Under-sampling methods and Hybrid methods (Chawla et al., 2002; Chawla et al., 2003; Guo et al., 2017). This paper will apply the Over-sampling methods to cope with imbalanced data.

The rest of this paper is organized as follows. Section 2 describes the basic principle of XGBoost and SMOTE. Section 3 describes the data set. Section 4 describes how to use XGBoost to predict defaulted bonds issuer. Section 5 is the conclusion.

2. XGBoost Algorithm and SMOTE Algorithm
The XGBoost algorithm adopts the idea of integration, and it can be used to solve both classification and regression problems. The algorithm mainly uses the integration idea to solve the minimum loss function through the second-order Taylor expansion, determines the split node, and constructs the final model. The basic algorithm as follows (Chen and Guestrin, 2016):
\[ \hat{y}_i = \sum_{k} f_k(x_i) \]  

In Equation (1), \( f_k \) is an additive function and represents a tree, then it is a tree ensemble model and uses \( K \) additive functions to predict the output.

To learn the set of functions used in the model, the model minimizes the following regularized objective.

\[ L(\mathcal{F}) = \sum_{i} l(y_i, \hat{y}_i) + \sum_{f} \Omega(f) \]

\[ \Omega(f) = \gamma T + \frac{1}{2} \| \omega \|^2 \]

In Equation (2), \( l(y_i, \hat{y}_i) \) is a differentiable convex loss function that measures the difference between the prediction \( \hat{y}_i \) and target \( y_i \), and the \( \Omega(f) \) penalizes the model's complexity. The additional regularization term helps to smooth the final learnt weights to avoid over-fitting. \( T \) represents the number of nodes on the tree, and \( \omega \) represents nodes' values on the tree.

Equation (2) includes functions as parameters and cannot be optimized using traditional Euclidean space optimization methods. Chen & Guestrin (2016) put forward a formula for evaluating the split candidates of the tree structure.

In short, Although XGBoost is developed from gradient tree boosting algorithms, XGBoost is still better than traditional gradient tree boosting algorithms for the following reasons: 1) The regular term has been added to the cost function of XGBoost to control the complexity of the model, and make the learned model simpler, and prevent overfitting. 2) The base learning instrument of XGBoost can be both linear classification and CATR classification. The traditional gradient tree boosting algorithms only uses CATR classification. 3) XGBoost learns from random forests and supports column sampling, which reduces overfitting and reduces computation.

SMOTE (Synthetic minority over-sampling technique) is one of Over-sampling methods that a new minority class sample is created in the neighborhood of the minority class sample under consideration (Chawla et al., 2002). Using this technique, a new minority class sample is created in the neighborhood of the minority class sample under consideration. The algorithm of synthesis is to randomly select a minority sample \( A \) and its nearest neighbor \( B \), between \( A \) and \( B \) as a newly synthesized minority sample. For example, the new minority sample \( C \) is synthetized as:

\[ C = A + \text{rand}(0,1) \times |A - B| \]

In Equation (3), \( \text{rand}(0,1) \) denotes a random number between 0 and 1, \( |A - B| \) denotes the Euclidean distance for the continuous features.

Now XGboost algorithm and SMOTE algorithm can be implemented using python packages. SMOTE algorithm is used to deal imbalanced data sets, and the XGBoost algorithm is used to classify the samples.

3. Data Set

We have got financial data of total 6731 bond issuers from the wind database in
2018, which has 50 bond issuers had defaulted in 2019. The data is typical imbalanced data; we will use SMOTE to deal with this imbalanced data. According to existing research, finally this paper selected 16 financial indicators based on analysis of four aspects of bond issuer from profitability, operational capacity, solvency capacity, capital structure (Table 1).

There are some key indicators that have a great effect on the accuracy of prediction: 1) Monetary/total debt is a measure of the solvency of the issuing company. The ratio can reflect the degree of protection of the company’s cash and cash equivalents to the company’s current liabilities. It is generally believed that the larger the ratio, the stronger the solvency of the company. 2) Return on asset and return on equity are indicators to measure the profitability of issuing companies relative to their total assets or equity. Generally speaking, the stronger the profitability of the company, the higher the ability to repay debts, and the smaller the risk of default. 3) The rate of stock turnover indicates how fast a company sells inventory. The ratio reflects the operational capacity of issuing companies. In general, the stronger the operational capacity of the company, the higher the ability to repay debts, and the smaller the risk of default. 4) The asset-liability ratio is a measure of capital structure. It is a leverage ratio that defines the total amount of debt relative to assets owned by a company. The lower the asset-liability ratio, the more repayment ability of the company and the less probability to default.

<table>
<thead>
<tr>
<th>Financial indicators</th>
<th>min</th>
<th>max</th>
<th>average</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>profit rate of core businesses</td>
<td>−25,406.88</td>
<td>15,244.17</td>
<td>11.20</td>
<td>260,017.10</td>
</tr>
<tr>
<td>income growth rate of core businesses</td>
<td>−181.99</td>
<td>346,306.17</td>
<td>172.91</td>
<td>34,596,379.30</td>
</tr>
<tr>
<td>return on asset</td>
<td>−109.20</td>
<td>48.19</td>
<td>3.09</td>
<td>35.04</td>
</tr>
<tr>
<td>return on equity</td>
<td>−832.57</td>
<td>138.85</td>
<td>3.68</td>
<td>463.02</td>
</tr>
<tr>
<td>EBITDA/Total Income</td>
<td>−23,571.09</td>
<td>16,280.29</td>
<td>26.27</td>
<td>205,915.70</td>
</tr>
<tr>
<td>operating cash/EBITDA</td>
<td>−12,286.60</td>
<td>2680.67</td>
<td>−1.80</td>
<td>30,347.40</td>
</tr>
<tr>
<td>liquidity ratio</td>
<td>0.05</td>
<td>5009.00</td>
<td>5.37</td>
<td>7331.34</td>
</tr>
<tr>
<td>quick ratio</td>
<td>0.02</td>
<td>5009.00</td>
<td>3.86</td>
<td>7266.21</td>
</tr>
<tr>
<td>rate of stock turnover</td>
<td>−16.11</td>
<td>48,228.55</td>
<td>37.28</td>
<td>524,720.59</td>
</tr>
<tr>
<td>asset-liability ratio</td>
<td>0.02</td>
<td>800.92</td>
<td>58.81</td>
<td>478.51</td>
</tr>
<tr>
<td>short-term debt/total debt</td>
<td>0.10</td>
<td>100.00</td>
<td>56.71</td>
<td>621.07</td>
</tr>
<tr>
<td>debt with interest/total investment capital</td>
<td>0.00</td>
<td>3308.84</td>
<td>41.66</td>
<td>2415.57</td>
</tr>
<tr>
<td>monetary/total debt</td>
<td>0.00</td>
<td>5009.00</td>
<td>1.99</td>
<td>6575.71</td>
</tr>
<tr>
<td>interest multiples</td>
<td>−2357.01</td>
<td>35,805.41</td>
<td>61.86</td>
<td>857,667.92</td>
</tr>
<tr>
<td>EBITDA/total debt with interest</td>
<td>−12,218.75</td>
<td>378,162.66</td>
<td>172.32</td>
<td>29,637,575.78</td>
</tr>
</tbody>
</table>
4. Effectiveness Analysis of XGboost Algorithm

The classification model performance could be evaluated by some indicators such as Precision, Accuracy, Recall, F-Measure. This paper mainly uses the area under the curve of ROC (Receiving Operating Characteristic) as the performance evaluation of the classification model.

Suppose the issuer of the defaulted bond is a positive sample (positives), and the issuer without defaulted bond is a negative sample (negatives). TP indicates the number of samples that the issuer of a defaulted bond is correctly predicted. FN indicates the number of samples that the issuer of a defaulted bond is not correctly predicted. FP indicates the number of samples that the issuer without defaulted bond is not correctly predicted. TN indicates the number of samples that the issuer without defaulted bond was correctly predicted.

True Positive Rate (TPR) is defined as:

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

False Positive Rate (FPR) is defined as:

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

For every classification model, the ROC curve is formed by a pair of FPR and TPR. ROC curve is a commonly used dichotomous model evaluation standard, which is used to show whether the effect is good or not. Since the ROC curve cannot quantitatively evaluate the classifier, the effect of the model is generally measured by the AUC (Area under the ROC curve), and the AUC value can well describe the overall performance of the model. The larger the FPR, the more real negative classes in the predicted positive class, and the larger the TPR, the more real positive classes in the predicted positive class. The closer the ROC curve is to the upper left corner, the greater the AUC value, and the model is better.

The following algorithm is implemented based on the Python software package. Total samples were classified according to 20% of the test samples and 80% of the training samples. There are six important parameters1 for tree Booster. 1) Learning rate that step size shrinkage used in update to prevents overfitting, this parameter is 0.05; 2) Maximum depth of a tree, this parameter is 5; 3) Minimum sum of instance weight (hessian) needed in a child, this parameter is 3; 4) Minimum loss reduction required to make a further partition on a leaf node of the tree, this parameter is 0.8; 5) Maximum delta step we allow each leaf output to be, this parameter is 0; 6) Subsample ratio of the training instances, this parameter is 1. Those parameters are settled by grid search algorithm.

At first, without the imbalanced treatment of the data, the ROC curves and AUC are obtained in Figure 1. Then before using the XGboost algorithm, we first use SMOTE to rebalance training samples. After that, the results on the test set are shown in Figure 2.

Finally, we compare the XGboost with other classification methods such as the K-Nearest Neighbors (KNN), Logistic regression (LR), Support Vector Machine (SVM), Decision Trees (DT), and Random Forest (RF). The AUC of the six modes list in the Table 2.

According to Table 2, even on very imbalanced data sets, XGboost still gets better results than traditional classification algorithms. It is effective in improving the effect of classification algorithm by data rebalancing.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC of imbalanced data</th>
<th>AUC of balanced data</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.733</td>
<td>0.858</td>
</tr>
<tr>
<td>SVM</td>
<td>0.682</td>
<td>0.694</td>
</tr>
<tr>
<td>KNN</td>
<td>0.722</td>
<td>0.782</td>
</tr>
<tr>
<td>DT</td>
<td>0.494</td>
<td>0.646</td>
</tr>
<tr>
<td>RF</td>
<td>0.774</td>
<td>0.836</td>
</tr>
<tr>
<td>XGboost</td>
<td>0.900</td>
<td>0.914</td>
</tr>
</tbody>
</table>

Figure 1. The AUC without deal with imbalanced data.

Figure 2. The AUC with deal with imbalanced data.
5. Conclusion

China’s bond market is an emerging market. The time of development is short, so the default data show a severe imbalance feature. As the result of Section 4, some traditional statistical classification methods can’t be used to classify the imbalanced data. This paper applies the XGboost algorithms to predict the default risk of corporate bond issuers in China. The conclusions are as follows. At first, the XGboost algorithm has advantages over the traditional algorithm in processing imbalanced data. Secondly, it is necessary to rebalance the imbalanced data by some method before building the prediction model. This paper shows SMOTE is an effective method to deal with imbalanced samples.

Funding

This work is funded by the National Natural Science Foundation of China (Grant No. 71571030).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References


Li, H. et al. (2020). XGBoost Model and Its Application to Personal Credit Evaluation. *IEEE Intelligent Systems, 35*, 52-61. [https://doi.org/10.1109/MIS.2020.2972533](https://doi.org/10.1109/MIS.2020.2972533)


Longstaff, F. A., & Schwartz, E. S. (1995). A Simple Approach to Valuing Risky Fixed and
https://doi.org/10.1111/j.1540-6266.1995.tb04037.x

https://doi.org/10.1016/j.elerap.2018.08.002

https://doi.org/10.1016/0378-4266(77)90022-X

https://doi.org/10.1111/j.1540-6261.1974.tb03058.x

https://doi.org/10.1102/isaf.203

https://doi.org/10.2307/2490395

https://doi.org/10.1049/gtd2.12073


https://doi.org/10.17559/TV-20200808113807

https://doi.org/10.1080/07421222.1999.11518234

https://doi.org/10.1016/j.elerap.2017.06.004

https://doi.org/10.1109/ACCESS.2020.3014644

https://doi.org/10.3233/IFS-189098

https://doi.org/10.1109/ACCESS.2018.2818678

https://doi.org/10.1109/ACCESS.2020.3042848

https://doi.org/10.32604/cmes.2020.011013

Y. Zhang, L. Chen

https://doi.org/10.1080/00268976.2019.1696478