Research on Credit Risk of enterprises based on Logistic Regression and Boost

Ye Yang

School of Economics and Management, Tongji University,200092

15888024556@163.com

Keywords: Credit Risk; Logistic Regression; XGBoost

Abstract: In recent years, the number and amount of defaulted debentures in the bond market keep rising, causing annual economic loss more than 100 billion yuan and this trend is growing. The bond market is only a small part of the huge economic market. Because of the contagion effect, the credit risk of the entire market cannot be ignored. As a financing institution, one of the risks that banks faced is credit risk. The main method currently used by banks to measure credit risk is internal ratings-based approach, which is a method with a large degree of freedom. AS the global trend tighter regulation deepens, because of the drawbacks of excessive freedom of internal ratings based approach, Basel III restricts it. Therefore, looking for a new method to measure enterprises' credit risk has become a consideration for financial institutions. This paper tries to find a new method to measure credit risk by using machine learning method.

This paper collects the enterprises' information of defaulted and undue debentures which is publicly available in the bond market. Obtaining their financial data, combine with the macro economic index indicators and then do data cleaning and sorting. According to the pre-judgment on the impact of various variables on the default, selecte relevant variables and use Logistic Regression and XGBoost to calculate the probability of enterprise default to judge whether a enterprise will has substantial risk.

1. Introduction

The number of defaulted debentures is continually rising in recent years,6 debentures defaulted in 2014, amounting to 1.34 billion yuan. In 2015-2019,the number was 27,56,34,125 and 180, the amounts were 12.18,39.38,31.25,120.96 and 146.6 billion yuan. By the end of March, the accumulative number of defaulted debentures in 2020 exceeded 40 and most of them cannot be get back. They caused economic losses more than 55 billion yuan. With the slowdown of economic growth in last few years, the economic environment is no longer as good as before and the number of default events is increasing. Bond market is just a small part of the economic market, there are more credit risk events happen in other markets .In addition, the challenges in international environment and public health are intensifying from the second half of 2018. Although macro policies provide support to enterprises, credit risk is still a huge challenge for banks.

At present, the major method commercial banks used to measure credit risk is internal ratings based (IRB) approach. As the global trend tighter regulation deepens, Basel III places restrictions on IRB approaches. Seeking for a new measurement is gradually considered by agencies. With the further research of machine learning, using machine learning algorithms to measure bank risk has been studied by many financial institutions. This report uses Logistic Regression and XGBoost to measure credit risk of enterprises.

In a narrow sense credit risk can broadly be viewed as default risk. More generally, by a credit risk we mean the risk associated with any kind of credit-linked events,such as:changes in the credit quality(including downgrades or upgrades in credit ratings),variations of credit spreads, and the default event [1].Credit risk includes expected loss and unexpected loss. This report focuses on expected loss. Expected loss is credit loss that a bank can estimate and reflectes on its balance sheet. Expected loss can be calculated as the product of Probability of Default, Loss Given Default and Exposure At Default: $EL = PD \times LGD \times EAD$ .This report evaluate the expected loss through the

DOI: 10.38007/Proceedings.0001190

ISBN: 978-1-80052-007-3

probability of default calculated by Logistic Regression and XGBoost.

The significance of studying this topic is looking for a method to achieve continuous PD through automatic calculation, implement real time monitoring and avoid manipulation at the same time.

The rest of this report is organized as follows. Section 2 describes data. Section 3 outlines two machine learning models. Section 4 present the results of two models. Finally, section 5 is conclusion and provides the suggestions for further research.

2. Data

The purpose of this report is to study how the macro-econominc index and financial situation of an enterprise affect the probability of default through debentures. This report runs empirical study on the debentures examined and approved by NAFMII,CSRC or filed by exchange, covering 80% of undue debentures with publicly available information in the bond market.

This report collects all defaulted debentures from 2014 to 2019 and obtains financial data of 128 defaulted enterprises. In consideration of sample size, this report uses bootstraping to increase the defaulted samples to 318.According to the nature of defaulted enterprises, data is divided into central state-owned,local state-owned, private,public and foreign-funded enterprises.In term of industry, data is divided into materials, industry, energy, optional consumption, daily consumption, information technology, healthcare, real estate,utilities and finance.Because of the significant stratification of the defaulted samples,report uses stratified random sampling to sample undue bonds and gets 636 undue samples.The ratio of defaulted and undue samples is 1: 2.Table 1 lists the summary statistics of main variables in data.

Table 1. Summary statistics of main variables in data

	Min.	Median	Mean	Max.	SD.
SPREAD	-4.40	1.75	1.63	6.65	1.57
CREDIT LINE USED	0.00	66.53	163.26	7261.43	383.56
CREDIT LINE USED PERCENTAGE	0.00	0.61	0.63	1.00	0.21
GUARANTEE LATEST OUTWARDS	0.00	49950.00	172445.00	7164295.00	447445.00
GROSS PROFIT RATIO	-82.25	18.97	21.25	97.77	19.99
QUICK	0.05	0.87	1.04	49.00	1.80
OPERATING CASHFLOW TO INTEREST DEBTS	-6.40	0.09	0.17	60.44	2.05
CURRENT	0.05	1.20	1.42	49.00	1.86
DEBTS TO ASSETS RATIO	2.04	64.99	70.87	800.93	64.82
YEAR TO YEAR OPERATING REVENUE	-99.71	7.57	31.51	13056.40	462.51
YEAR TO YEAR CASHFLOW	-30598.10	-39.30	691.20	581022.40	19130.60
INVENTORIES TO TOTAL ASSETS	0.00	0.11	0.15	0.77	0.15
SHORT TERM LIABLITIES TO TOTAL LIABILITIES	0.00	0.21	0.21	0.96	0.15
UNDISTRIBUTED PROFIT	-4668186.00	146701.00	408632.00	72718700.00	2638179.00
TOTAL PROFIT	-1100456.00	41698.00	118022.00	11520000.00	638535.00
NET PROFIT	-1132285.00	30091.00	78029.00	8028900.00	492037.00
NET CASHFLOW FROM OPERATING ACTIVITIES	-1857659.00	40530.00	237916.00	35156500.00	1534902.00
GDP GROWTH RATE	6.20	6.20	6.43	7.80	0.37
CPI	594.80	669.80	658.20	669.80	19.40

PPI	353.90	388.20	384.90	389.40	8.80
PMI	49.70	54.00	52.40	54.90	2.00
INFLATION RATE	1.40	2.90	2.57	2.90	0.50
UNEMPLOYMENT RATE	3.60	3.60	3.71	4.10	0.17
CENTRAL PARITY RATE	6.12	6.95	6.88	6.95	0.15

This report fills the samples by financial data and macro-econominc index.For example,report uses the latest financial data to fill the undue samples and fills the defaulted samples with the financial data from previous year to the first three years before the default.Prioritize the data from the year before the default. If it’s difficult to obtain, use the data from the previous two to three years.Macro econominc indicators are also done in this way except exchange rate which is been used by real-time exchange rate. Because the impact of exchange rate changes on enterprises is reflected in real time.After that this report uses median to fill missing data.

The reason for using one-year data instead of three-year weighted average data is considering the enterprise's operating situation, cashflow status and the impact of macro economy on enterprise default within one year instead of observing how the enterpris has gradually deteriorated in three years.

Then this report uses runif method to divide samples into a training set and a validation set with a ratio of 70% and 30%.70% is used for training and 30% for verification.All relevant analyses are calculated in the programming language R.

3. Model

This report uses two models,logistic regression and XGBoost, to calculate probability of default.In terms of variable selection,report puts all variables in logistic regression and then uses stepwise regression to filter variables.When uses XGBoost, report selects variables in advance according to the cause of credit risk.The following is the theoretical basis for selection of variables.

The causes of credit risk can be divided into six aspects,including default risk,recovery risk,exposure risk,migration risk,spread risk and liquidity risk [2].

Default risk is institutions unable to pay off debts because of the increased probability of default.Recovery risk is given the default event happened,the amount recovered by collateral is lower than expected,measured by recover rate.Exposure risk is the amount of debt at default may increase compared with current,measured by EAD.Migration risk is credit quality decline compared with current,such as downgrading.Spread is the difference between interest rate and risk free rate.Spread represents the risk premium and it's the risk compensation required by investors for bearing risks.Spread risk is the expansion of spread and it means investors require more risk premiums because of the higher risks.Liquidity risk includs funding liquidity risk and market liquidity risk.Funding liquidity risk is the current or prospective risk arising from an institution's inability to meet its liabilities and obligations as they come due without incurring unacceptable losses. Market liquidity risk is the risk that the act of buying or selling an asset will result in an adverse price move [3].Table 2 lists the classification of credit risk.

Table 2. Classification of Credit Risk

Credit Risk	Description
Default Risk	The probability of default rises
Recovery Risk	Given the default event,the value of collaterals are lower than expected
Exposure Risk	The exposure of the default debts increase
Migration Risk	Credit quality changes,such as downgrading

Spread Risk	Spread is the premium investors require for high risk compared with treasury bonds .Spread risk means investors require more risk premiums for the rising risk.
Liquidity Risk	Inability to meet obligation or assets are sold lower than expected

Default risk is measured by probability of default(PD) which this report would like to get through logistic regression and XGBoost.Recover rate and Exposure At Default(EAD) are usually given in practice,so Recovery Risk and Exposure Risk are not considered in this report.And downgrading is a lagging indicator.Consider all the reasons above, the variables which are selected major reflect the spread and liquidity risk.In addition, considering the financing capacity and third-party warranty obligation, several additional indicators are added into model.Table 3 lists the variables selected by the above theories.

Table 3. Variables used in XGBOOST

Variable	Description	Notes
SPREAD	Reflect risk premium	It's the difference between interest rate and risk free rate.It’s also the risk premium investors required for bearing the risk.
DEBTS TO ASSETS RATIO	Reflect short-term solvency	It reflects the debt ratio of an enterprise.Moderate debts are conducive to the development of enterprises,but too high will affect the solvency.Different industries have different average levels.It equals debts divided by assets.
SHORT TERM LIABLITIES TO TOTAL LIABLITIES	Reflect short-term solvency	It's proportion of short-term debts to total debts and it reflects the level of borrowing that needs to be repaid within one year.
QUICK	Reflect short-term solvency	It reflects the short-term liquidity of assets.It equals Liquid Capital such as Money Funds,Bills receivable,Accounts Receivable divided by Current Liabilities.
NET CASHFLOW FROM OPERATING ACTIVITIES	Reflect liquidity	It reflects the cashflow obtaining from operating activities.It's one of the indicators that refects whether the cashflow is sufficient.
YEAR TO YEAR CASHFLOW	Reflect liquidity	It reflects the growth rate of enterprise's cashflow per year.It's one of the indicators that refects whether the cashflow is sufficient.
OPERATING CASHFLOW TO INTEREST DEBTS	Reflect liquidity	It's the ratio that debts coveraged by cashflow generated by operating activities.It reflects whether the operating activities can cover the enterprise's debts.
NET PROFIT	Reflect profitability	It's the profitability of business activities.
UNDISTRIBUTED PROFIT	Reflect profitability	It's the benefits delivered the previous year.
YEAR TO YEAR OPERATING REVENUE	Reflect profitability	It reflects the growth of income from business activities per year.

CREDIT LINE USED	Reflect financing ability	It reflects the ability of enterprises to borrow loans from banks.At the same time, it reflects the dependence level of an enterprise on external funds.
CREDIT LINE USED PERCENTAGE	Reflect financing ability	It reflects the ability of enterprises to borrow loans from banks.At the same time, it reflects the dependence level of an enterprise on external funds.
GUARANTEE LATEST OUTWARDS	Reflect Third-Party-Obligation	It's third-party warranty obligation,if the third party defaults,the enterprise needs to bear the payment obligation.

The following is how the variables mentioned above are correlated default.SPREAD is the difference between interest rate of an enterprise and risk free rate.It’s also the risk premium investors required for bearing the risk.The higher risk is ,the more premium they need.So the higher the SPREAD is,the more likely an enterprise will default.DEBTS TO ASSETS RATIO reflects the debt ratio of an enterprise.Moderate debts are conducive to the development of enterprises,but too high will affect the solvency.SHORT TERM LIABLITIES TO TOTAL LIABLITIES is a ratio reflects the proportion of short-term debts to total debts.Short term liabilities needs to repid within a year and long term liabilities will repaid exceed a year,may three to ten year or even more. The higher the ratio is ,the more money funds need to be prepared in a short period of time.So the higher two ratios are,the more money is needed,the higher the probability of insolvency will be.

QUICK is the ratio that an enterprise's ability to liquidate in a short period of time.It equals Liquid Capital such as Money Funds,Bills receivable,Accounts Receivable divided by Current Liabilities.NET CASHFLOW FROM OPERATING ACTIVITIES reflects the cashflow obtaining from operating activities.The higher the two indicators are,the less likely an enterprise will default.

OPERATING CASHFLOW TO INTEREST DEBTS reflects whether the operating activities can cover the enterprise's debts.YEAR TO YEAR CASHFLOW reflects the growth rate of enterprise's cashflow per year.It's one of the indicators that refects whether the cashflow is sufficient.NET PROFIT and UNDISTRIBUTED PROFIT are the profitability of business activities this year and the year before.It reflects the quality of the enterprise's operations.YEAR TO YEAR OPERATING REVENUE is the ratio that reflects the growth of income from business activities per year.So the higher these indicators the better the enterprise is.

CREDIT LINE USED is the financing amount,which is given by banks,already used .And CREDIT LINE USED PERCENTAGE it the proportion of it. The two indicators also reflect the dependence level of an enterprise on external funds.The higher these two indicators are,the easier an enterprise may default.GUARANTEE LATEST OUTWARDS is the amount that an enterprise provides guarantee for other enterprises' external financing.It's third-party warranty obligation,if the third party defaults,the enterprise needs to bear the payment obligation.So the higher the indicator is,the more likely an enterprise will default.

In addition, macro-econominc index indicators such as GDP GROWTH RATE,CPI,PPI etc. are not used into model finally. Because report uses the latest data to fill the undue samples and fills the defaulted samples with the data from the year before the default. However, due to the impact of the economic environment, the latest data are all worse than the previous years.So it is easy to distinguish defaulted and undue samples.Therefore, macro-econominc index indicators are not used for prediction in both Logistic Regression and XGBoost models.

4. Results

4.1 the results of Logistic Regression

This report uses Logistic Regression and XGBoost to compute probability of default.

This report puts all variables in logistic regression and then uses Stepwise Regression to filter variables.The accuracy of Stepwise Regression is 0.8274648 and true positive is 0.670213.Table 4 lists the confusion matrix of Stepwise Regression.

Considering that logistic regression is a linear model, report computes the variance inflation factor(VIF) to see whether there is multicollinearity between variables.Generally, if the the VIF is greater than 5, it means there is multicollinearity.So this report removes highly correlated variables.Table 6 lists the final results after the processing above.

Table 4. Fianl ruselts of Logistic Regression

Fianl ruselts of Logistic Regression	
	default
SPREAD	0.367***
	(0.087)
CREDIT LINE USED PERCENTAGE	3.960***
	(0.757)
GROSS PROFIT RATIO	0.049***
	(0.012)
DEBTS TO ASSETS RATIO	0.019***
	(0.006)
CURRENT	0.540***
	(0.205)
INVENTORIES TO TOTAL ASSETS	-3.263***
	(0.937)
NET PROFIT	-0.00001***
	0
NET CASHFLOW FROM OPERATING ACTIVITIES	-0.00001***
	0
Constant	-5.025***
	-0.854
N	670
Log Likelihood	-230.194
Akaike Inf. Crit.	526.387
Notes:	***Significant at the 1 percent level.

According to fianl ruselts of Logistic Regression, set the probability of enterprise default as p.Dependent variable,y,is binary.The default model is as follows:

y=-5.025+0.367SPREAD+3.960CREDIT LINE USED PERCENTAGE+0.049GROSS PROFIT RATIO+0.019DEBTS TO ASSETS RATIO+0.540CURRENT-3.263-0.00001NET PROFIT-0.00001NET CASHFLOW FROM OPERATING ACTIVITIES

$$\mathrm{p}(y=1)=\frac{\exp(y)}{1+\exp(y)}$$

When y is equal to 1, it means that the default occurs, and the probability of default ,p,can be obtained by using the formula above.

The following is the explain of Logistic Regression coefficient.The final result from Logistic Regression shows that there are 8 variables significantly leading to default at 1% significance level.Five of these variables are positively related to default and the others present the negative related relation.For example,while keeping other variables constant,SPREAD changes by 1 unit, causing y to change by 0.367 units.While keeping other variables constant,INVENTORIES TO TOTAL ASSETS changes by 1 unit,causing y to change by -3.263 units.

4.2 the results of XGBoost

When using XGBoost, this report selects variables in advance and then puts them into model.The accuracy of XGBoost is 0.9119718 and true positive is 0.861702,both indicators are much better than the results of Logistic Regression.It shows that the prediction made by XGBoost is much more accurate.Table 7 lists the confusion matrix of XGBoost.

The above results verify previous thoughts such as the higher the SPREAD is ,the more likely a enterprise will default and the more the NET PROFIT is ,the less likely it will default.So are other variables.

Conclusion

Inferred from the accuracy and true positive above,it is feasible and accurate to calculate PD through two models.XGBoost is significantly better than Logistic Regression both in accuracy and true positive.Compared with Credit Rating Transition Matrix,two models both can get continuous PD and more specific to the individual.Because models are based on automatic calculation,so it can reduce even avoid manipulation and achieves the purpose of real time monitoring.But it has to be recognized that because of the insufficient research ability,unstructured data has not been added into the models.It will addressed in the following research.

References

[1] Bielecki, T. R., & Rutkowski, M. (2013). *Credit risk: modeling, valuation and hedging*. Springer Science & Business Media.

[2] De Laurentis, G., Maino, R., & Molteni, L. (2010). *Developing, Validating and using internal ratings*. United Kingdom: John Wiley & Sons.

[3] Brunnermeier, M. K. (2009). *Deciphering the liquidity and credit crunch 2007-2008*. Journal of Economic perspectives, 23(1), 77-100.

[4] Basel Committee (2017). *Basel Committee on Banking Supervision: High Level Summary of Basel III ReformsAvailable*. www.bis.org.

[5] T. Chen, S. Singh, B. Taskar, and C. Guestrin.(2015). Efficient second-order gradient boosting for conditional random fields. In Proceeding of 18th Artificial Intelligence and Statistics Conference (AISTATS'15), volume 1

[6] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde ,S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar. (2016) .MLlib: Machine learning in apache spark. Journal of Machine Learning Research, 17(34):1–7

[7] T. Zhang and R. Johnson.(2014). Learning nonlinear functions using regularized greedy forest. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(5)