A Research on Corporate Bond Defaults in the Chinese Market

By

Yiwen Chen

B.S. Economics
Shanghai University of Finance and Economics, 2015

SUBMITTED TO THE MIT SLOAN SCHOOL OF MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN MANAGEMENT STUDIES
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MAY 2021

Signature of Author:
____

MIT Sloan School of Management
May 7, 2021

Certified by:
____

Christopher Francis Noe
Senior Lecturer of Accounting
Thesis Supervisor

Accepted by:
____

Jacob Cohen
Senior Associate Dean for Undergraduate & Master's Program
MIT Sloan School of Management

A Research on Corporate Bond Defaults in the Chinese Market

By

Yiwen Chen

Submitted to MIT Sloan School of Management
on May 7, 2021 in Partial Fulfillment of the
requirements for the Degree of Master of Science in
Management Studies.

ABSTRACT

Using data from the Chinese fixed income market, this thesis builds up a logistic regression model mainly consisting of both financial condition variables and financial report quality variables. The analysis suggests the degree of effect for different variables and thus provides a reference for credit risk assessment. Supporting evidence is also provided to show that the model can predict default one year in advance effectively and perform better than the main rating agency companies.

Thesis Supervisor: Christopher Francis Noe
Title: Senior Lecturer of Accounting

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION OF THE STUDY

1.1 Structure of the Study

This paper aims to analyze the default events in the past and predict the bond default events in 2021 in the Chinese bond market. The structure of the study can be mainly divided into chapters.

Chapter 1 is a basic introduction to the current Chinese bond market and the whole study.

Chapter 2 is a summary of existing related literature, including different kinds of static models to assess credit risk and some critiques.

Chapter 3 is the methodology of the research, and this part includes the theoretical foundation and model design. In Chapter 3 I describe how I decide which variables to use in the model.

Chapter 4 is about how I collect data, my method of analysis, and what the model is. At the end of this chapter, I show which bonds, according to the model, would default in the next half year.

Chapter 5 is the conclusion of the study, stating the significance and limitations of the study.

1.2 Background

The Chinese capital market grew rapidly in the past four decades, and is now the world's second-largest capital market. By 2018, China had a capital market that scaled more than 126 trillion RMB[1]. It is also important to note that currently, China is increasingly opening its capital market to the outside world. It is able to do so by gradually removing regulation obstacles for entering the market. Consequently, China is impacting the global market more and more deeply.

The major asset classes that make up the Chinese capital market include stocks, bonds, funds and other investments.

Among all the major asset classes, bonds occupy the largest part of the Chinese market. From 2010 to 2018, the Chinese bond market grew from RMB 16.3 trillion to RMB 86.4 trillion. The average growth rate in this period was 23.18% per year (China People's Bank 2019:38,109; 2011:36[2]).

---

[1] The exchange rate between RMB and US Dollar varied from 6.45-7.10 RMB/USD. (Data resource: Capital IQ https://www-capitaliq-com.libproxy.mit.edu/CIQDotNet/MacroEconomics/EconomicSeries.aspx?economicEntityId=110316468, so the scale in USD is approximately 18 trillion USD.

[2] Data from the Annual Report of China People's Bank, which is a government public document.

The Chinese bond market consists of three main types of assets: the first type is government bonds, including Treasury bonds, issued by the government and three State Policy Banks[3]. The second type is corporate bonds issued by companies, whether state-owned companies or private enterprises. The third type is certificates of deposit issued by commercial banks. At the end of 2019, there were around 60.0 trillion RMB government bonds, 25.0 trillion RMB corporate bonds and 10.5 trillion RMB certificates of deposit in circulation, representing 61%, 26% and 11% of the whole market, respectively[4].

The modern form of the Chinese bond market has a relatively short history compared to the developed countries. It was in the 1980s that the modern bond market began developing in China. Therefore, the Chinese bond market is not fully mature in some respects.

Due to many reasons, including the tight capital control regulation imposed by the Chinese government, the Chinese bond market is not completely bound to the global market, allowing Chinese Treasury bonds a relatively high return rate in the global environment of negative return rates. In 2019, the Sino-U.S. Treasury three-year bond spreads ranged from +0.11% to +1.47% (Wind Information, 2020).

For the corporate bond part, the default rate[5] of corporate bonds in the Chinese bond market is comparably low for the default rate has been below 1% every year. As a reference, the default rate of all the bonds that were rated by Standard & Poor's in 2017 was 1.66% (Standard & Poor's, 2018).

All the special characteristics that the Chinese bond market has make it quite attractive to both Chinese and global investors, whether individuals or institutions. In May 2019, there were 32,329 domestic institutional investors and 378 foreign institutional investors in the inter-bank bond market[6] (National Inter-bank Funding Center, 2020). At the end of the first quarter of 2020, 1.96 trillion of bonds issued in the Chinese bond market were held by foreign investors. What is more, this number has kept increasing for five consecutive quarters since the fourth quarter of 2018 (China People's Bank, 2019:43). There is an expectation that these numbers will certainly keep increasing in the future.

As in the U.S. market, the Chinese market has rating agencies. Five large rating agencies monopolize the market, and almost all the domestic investors use their rating reports as one of the references when making investment decisions. But disappointingly, the rating agencies have rarely made timely down-side rating adjustments before a default happened, which makes the bond ratings largely useless to investors.

[3] The three State Policy Banks include China Development Bank, China Agricultural Development Bank and China Ex-im Bank.

[4] Based on the data from Shanghai Security Exchange, 2020； Shenzhen Security Exchange, 2020；Shanghai Clearing House, 2020; Central Clearing Corporation, 2020

[5] The definition of default includes two types of events: the first is a missed or delayed disbursement of interest and/or principal, including delayed payments made within a grace period; the second is any other interruptions to the timely payment of interest and/or principal.

[6] The inter-bank market holds more than 85% of the bonds in Chinese bond market. It only allows institutional investors to enter.

Moody's, Standard & Poors and Fitch were allowed to enter the Chinese capital market in 2017, but up until now, they have only covered a very limited number of bond issuers.

The first bond default event in the Chinese bond market happened in 2014, and since then more and more default events have occurred in this market.

	2020, up to end of April	2019	2018	2017	2016
Numbers of Default Cases	51	182	130	34	56
Amount of Defaulted bonds (RMB)	64.7 billion	147.6 billion	124.3 billion	31.2 billion	39.4 billion

Figure 1 Defaulted Bond Case Statistics, 2016 to 2020
(Based on data from Wind Information, 2020)

On one hand, a default rate in a reasonable range is necessary for a market to allocate resources efficiently, and thus is characteristic of a mature and healthy market. A reasonable default rate can help maintain the balance between risk and return. On the other hand, frequently occurring default events imply that the bond issuers in this market may have problems with their financing and operations, exposing investors to risk that they badly want to avoid. In such an environment, credit risk measurement and risk control become increasingly important to investors.

Based on a comprehensive study of a sample of default events, I plan to list variables that can reflect debt paying ability in different dimensions. By using the selected variables, a simple quantitative model is estimated. I hypothesize that this model can predict the possibility of default 6-12 months before the default event.

CHAPTER 2: LITERATURE REVIEW

2.1 Overview of Credit Risk Measurement Models

Credit risk measurement research began in commercial banks. Initially the banks conducted qualitative research relying on the experience of experts. Their assessment could have depended highly on subjective analysis and qualitative judgement.

In the later half of the 20th century, Credit Risk Measurement Research entered the quantitative stage with the development of math and computer science. In 1968, Altman established the Zscore bankruptcy prediction model. He selected five factors to conduct multi-factor research. Then Altman, Haldeman and Narayanan (1977) extended the Z-score model to seven factors, creating a model that is more realistic: the ZETA Model. Altman kept revising his model and in 1995 created the Z Double Prime (Z'') Model, which excludes the asset turnover ratio to make the model more applicable across industries.

Merton (1974) presented a model that has been widely used by scholars. This model uses the pricing method of options to calculate the value of a company and argues that when the value of the company assets is below a certain number, its stockholders will choose to default. Due to the theoretical base of the Merton model, it set many assumptions that do not agree with reality. Some scholars, after Merton, removed some of the assumptions. Black and Cox (1976) added subordinated debt and included repayment sequence into the Merton Model. Longstaff and Schwartz (1995) extended the default point from the maturity day to any day before it. Lando (2004) added random interest rate into the Merton Model, and thus broadened the assumption about constant risk-free rate.

As big data analysis techniques developed rapidly, credit risk measurement research stepped into a more advanced stage. Some new quantitative models were invented. In 1997, J.P Morgan developed the Credit Metrics Model. This model simplifies assumptions in the Merton Model and uses credit rating immigration, default rate and recovery rate. Also based on the Merton Model, KMV Company set up the KMV Model to realize constant valuation of debt issuers. The KMV Model claims that the information quality of a company is decided by its specific asset allocation and asset structure. The KMV Model takes stock price information into consideration, but doesn't analyze the relationship with debt default. McKinsey Company built the Credit Portfolio View Model in 1998. In the Credit Portfolio View Model, they take macroeconomic variables, such as economic growth rate and unemployment rate, into consideration.

The above quantitative models brought about widespread discussion. Many scholars gradually developed extended models that have stronger predictability. For example, Norden and Weber (2004) studied 90 companies in Europe, Asia and the US and found that the derivative market anticipated credit rating downgrades around 60-90 days before the announcement day.

Deteriorating financial condition is not the only factor that leads to bond default; financial fraud can also be a powerful impact factor. We cannot omit financial report quality when predicting bond defaults. To assess the financial report quality of bond issuers, there are some existing models to be referred to. The most widely known financial report quality scoring models might be the Mscore and Fscore Models.

Beneish (1999) researched a sample of 74 companies that were punished by the SEC because of financial fraud in 1982-1992 and 2332 other comparable companies. By doing this, he came up with the Mscore Model, which uses accounts receivable index, gross margin index, asset quality index and five other indexes, to judge the possibility of financial fraud.

Dechow et al. (2011) built the Fscore Model on the basis of Mscore Model. They studied 676 fraud companies, and tested the characteristics of financial fraud in five dimensions: accrual quality, financial performance, nonfinancial performance, off-balance-sheet activities, and market-related variables. They finally concluded that the models that only include financial indexes have the best estimation capability, and the accuracy is 69%.

2.2 Credit Risk Research in the Chinese Market

Due to the short history of the Chinese bond market, credit risk research about the Chinese market also started comparatively late. Many domestic scholars did research about credit risk based on the models developed by former scholars. They tested the effectiveness of these models in the Chinese market.

In 2002, Xiang used the Zscore Model to conduct research on 80 listed companies and found that the Zscore Model is efficient in studying management risk of listed companies in the Chinese market. Later Jin (2005) picked the Credit Metrics Model to study the credit risk of a state-owned bank and proved that the Credit Metrics Model has predictability in the Chinese market.

Zheng (2005) tested the KMV Model with data of listed companies in the Chinese market, and he found that the KMV Model cannot predict default possibilities correctly. Zheng thinks the reason is that the stock price in the Chinese market cannot reflect intrinsic value and that the assumption of asset value obeying a normal distribution is not satisfied.

2.3 Summary of Credit Risk Researches Using the Logistic Model

The logit model, a kind of multi-factor regression model, can be used to analyze questions in many different fields. The usage of the logit model in credit risk analysis is quite well-known. There is adequate literature supporting the effectiveness of the logit model in measuring corporate credit risk.

Martin (1977) analyzed the relationship between return on assets, debt ratio and hazard rate using a sample of bankrupted American banks from 1975-1976. Martin's research shows that the logit model has the same accuracy as discriminant estimates. Ohlson (1980) used both logit regression and probit regression to make comparable research of normal enterprises and bankrupted enterprises. From the research, Ohlson discovered that asset scale, capital structure, operational performance and liquidity significantly influence the possibility of bankruptcy, and the logit model has a better discriminant ability. Goss and Ramchandani (1995) used the logit model, neural network model and parametric models to study how to make early warning on financial distress of insurance companies, and the conclusion is that the logit model and neural network model can predict more effectively than parametric models.

For the logit model, the critical factor that decides the effectiveness in predicting bankruptcy and default events is the selection of independent variables in the model. Different scholars used different variables to reveal their different opinions.

Many scholars took only accounting data as the independent variables. For example, Blume, Lim and MacKinlay (1998) used accounting ratios including pretax interest coverage ratio, operating income to sales, long-term debt to assets, and total liability to assets. Some other scholars added non-accounting data, such as the volatility of equity returns, to increase the accuracy. Campbell and Taksler (2002) offered evidence supporting the significant relationship between asset volatility and default possibility. However, later Du (2004) claimed that the relationship between the equity return and credit risk is complicated and cannot be described as positively related or negatively related.

For the Chinese capital market, there are also several papers that use logit regression method to assess credit risk of listed companies. Pang (2006) selected 63 listed companies and built a logit regression credit valuation model, reaching an overall accuracy of 99.06%. But these studies mainly focus on listed companies, so that the sample scale is not large enough to generalize their outcome to all bond issuers. In addition, listed companies are usually more easily connected to finance resources, making the research biased.

CHAPTER 3: METHODOLOGY

3.1 Theoretical Foundation

The Logistic Regression Model is widely used to reflect the relationship between a group of *n* predictor variables $\{X_1, X_2, X_3, \ldots, X_n\}$ and a binary dependent variable.
Let *p* equal the possibility that a default event occurs. If default happens, p=1; if not, p=0.

Then $$Logit(p)=In\left(\frac{p}{1-p}\right)=b_0+\sum_{i=1}^{n} biXi \quad (1)$$

In equation (1), b_i represents the unknown Logistic Regression coefficients, and b_0 is the intercept.

3.2 Design

After choosing the logit model as the basic model type, I then did research to find the variables I needed to build the model and the data used to run the model. Several aspects should be taken into consideration.

3.2.1 Individual Financial Condition Variables

To assess the financial conditions of issuers, I use the Z" Score developed by Altman.

The Z Score Model has been widely used and proven effective in many regions. The Z" Score Model is an extended version of Z Score Model and more suitable for my research: Firstly, since not all the issuers are public companies, it's not appropriate to use the original Z Score Model because it requires market price of equity. TZ" Score Model uses book value of equity instead. Secondly, the sample includes bonds of different industries, and Z" Score Model excludes the ratio of asset turnover, making the model more comparable across sectors and industries.

The Z" Score Model is not the proper option to assess financial companies, such as banks, insurance companies and mutual funds. As mentioned before, the sample data would not include financial companies.

The individual financial variables are:
A). Working Capital to Total Assets
B). Retained Earnings to Total Assets
C). Earnings before Interest and Taxes to Total Assets
D). Book Value of Total Equity to Book Value of Total Liabilities

Variable A, which mainly measures the liquidity of an issuer company, can provide us insight into the ability to liquidate assets, but can sometimes be misleading because the quality of current assets varies among different asset types. So in the next part, a variable that can value asset quality will be added to my model.

Variable B, which measures the size of retained earnings, prefers mature companies to new companies. For bond investors, it is a good indicator because bond investors do not really care about the growth potential of issuers, but care greatly about their accumulated earnings. Variable B implies the leverage usage of a company because a company with high retained earnings would have capability to go forward without incurring more debts.

Variable C, which measures the profitability without impact of interest and taxes, shows us the pure future profit that can contribute to its debt repayment and consistent operations.

Variable D, which measures the ratio of equity to liability, provides a hint about future earning power and leverage. The logic behind variable D is that the equity investors give their capital for future profit, so the book value of equity can take some hidden factors into consideration.

3.2.2 Individual Financial Report Quality and Management Variables

The non-financial variables mainly include financial report quality and management variables.

For financial report quality variables, I use the M-Score Model as a base and combine it with additional variables that coincide with real market conditions. Since it is a long-term process for a company to eventually go bankrupt, the trend can be of more value than the ratios. So I chose to use the ratios of ratios in order to show more than only time-point values.

The individual non-financial variables are as below:

E). DSRI= $(Receivables_t / Sales_t) / (Receivables_{t-1} / Sales_{t-1})$

F). GMI= $Gross\ Margin_{t-1} / Gross\ Margin_t$

G). AQI= $[1-(PPE_t+CA_t)/TA_t] / [1-(PPE_{t-1}+CA_{t-1})/TA_{t-1}]$

H). SGI= $Sales_t / Sales_{t-1}$

I). DEPI= $Depreciation\ Rate_{t-1} / Depreciation\ Rate_t$

J). SGAI= $(SGA_t/Sales_t)/(SGA_{t-1}/Sales_{t-1})$

K). LEVI= $(Debt_t/Asset_t) / (Debt_{t-1}/Asset_{t-1})$

L). TATA= $(Net\ Income_t - Cash\ from\ Operations_t)/Total\ Assets_t$

M). TYPE: Enterprise Type, which means whether the issuer is a state-owned enterprise: 1 means yes and 0 means no.

Variables E, G, I, and L are signals that can result from revenue inflation and cost deflation; thus they are potential evidence of financial manipulation. Variables F, H, J, and K are signals that show the company is facing worse financial conditions, thus predisposing the company to manipulate earnings. Variable M is a variable to measure the possibility of government support. I believe that state-owned companies would get more resources from the government and therefore be less likely to default.

3.3 Data Selection

The next step is to collect data on bonds and issuers. My data comes from Wind, a Chinese database that has a leading position in the field of financial data. Wind provides comprehensive data, news, and analysis of Chinese financial securities and economy, covering stocks, bonds, funds, options, macroeconomics, etc. I also want to mention that I excluded corporate bonds issued by financial institutions and municipal bonds, for they have a different logic of repayment and whether they would default cannot be properly analyzed by issuers' financial conditions.

I selected corporate bonds that mature between 2016/01/01 and 2021/08/08, and I also chose the most updated financial report date on issuers that was at least one year before maturity to be time t.

All the defaulted bonds that belong to the same issuer would be merged into the first default event. The reason is that if an issuer has failed to pay its bond once, it would be obvious that it could not pay the incoming bond obligations. Also, a non-default issuer can have multiple matured bonds because they can issue more than one bond. All the non-default records with the same issuer are merged into one record, the latest one. Thus duplicated records are removed.

I separated the original data set into two sets: one is the raw data intended for model training and selection; the other one is saved for prediction. Basically, the raw data are bonds that matured between 2016/1/1 and 2020/12/31. However, as the default rate is quite low and we have too few records of default bonds, I added some default records that matured in 2021 into the sheet, so that I would have more data to train the models.

3.4 Data Collection and Treatment

Originally, before duplicate removal, I had 4320 records, with 77 default observations and 4243 non-default ones. After removing duplicates, 2305 records remain, with 2243 nondefault and 64 default. It can be seen that non-default bonds are over fifty times more numerous than default ones. Each record of default is so important that I must be very cautious about removing any of them.

I decided to delete non-default observations with missing values. But when it comes to processing missing values in the dataset, I processed them with more caution. There were two defaults that lacked nearly all the variable inputs; I had to remove these two. Then there are three more defaults that only lack one variable, which is variable J, the depreciation rate change ratio. It's understandable that some companies don't disclose their depreciation amount because they don't have a large amount of fixed assets. To address this issue, I replaced these three NAs with medians of the whole dataset. Up to here I have 62 default records, but I would need to compare the performance of the model with a benchmark, which is credit ratings issued by the mainstream credit rating agencies (the definition of credit ratings will be discussed in later chapters). I therefore only kept the records with credit ratings and deleted the others.

After that, there were 58 default observations and 1784 non-default ones. I then mainly used the data mining method introduced by Shmueli (2018) to train, validate, and test a statistical model.

I divided all pre-processed records into a training set, a validation set, and a test set, using 50%, 30%, and 20% of the total sample, respectively. In order to have relatively more default records for training, I reduced the ratio between non-default records and default ones in the training set to 25:1, while maintaining the original ratio (1784 over 58) in the other two sets.

A logit regression model requires that all the variables are numeric; hence I turned all the variables (including dummy variables with only values 0 and 1) into numbers. Then I tested the correlations between all prediction variables and found several pairs of variables had correlations higher than 0.5:

Variable C EBIT to Total Assets & Variable B Retained Earnings to Total Assets;
Variable C EBIT to Total Assets & Variable L TATA;
Variable B Retained Earnings to Total Assets & Variable A Working Capital to Total Assets;
Variable A Working Capital to Total Assets & Variable L TATA;
Variable B Retained Earnings to Total Assets & Variable L TATA;

Variables B and C are both financial performance ratios and since both are related to profitability, it's reasonable that they are highly correlated. Therefore, I decided to drop Variable B.

Variable L is a financial report quality variable, but since it has a correlation higher than 0.5 with three other variables, I chose to drop it. Finally, I used the remaining 12 variables to run logit regression.

	Retained. Earnings/ Asset	Working. Capital/ Asset	EBIT/ Asset	BV. of. Equity/ Liability	DSRI	GMI	AQI	SGI	DEPI	SAI	LEVI	TATA	Enterprise Type
Retained. Earnings/Asset	1.00	0.58	0.58	0.23	0.01	−0.03	0.01	0.00	−0.01	−0.24	−0.34	0.51	−0.07
Working. Capital/Asset	0.58	1.00	0.38	0.27	0.02	0.00	0.06	0.00	−0.02	−0.15	−0.22	0.52	−0.08
EBIT/Asset	0.58	0.38	1.00	0.18	0.00	−0.01	−0.01	−0.01	0.01	−0.28	−0.57	0.82	−0.03
BV. of. Equity/Liability	0.23	0.27	0.18	1.00	−0.01	0.01	−0.01	0.00	−0.02	−0.06	−0.19	0.12	−0.08
DSRI	0.01	0.02	0.00	−0.01	1.00	0.00	0.00	0.00	0.00	0.02	−0.02	0.00	0.01
GMI	−0.03	0.00	−0.01	0.01	0.00	1.00	−0.02	−0.01	0.00	0.00	0.03	0.00	0.01
AQI	0.01	0.06	−0.01	−0.01	0.00	−0.02	1.00	0.00	−0.01	0.06	0.00	0.02	−0.01
SGI	0.00	0.00	−0.01	0.00	0.00	−0.01	0.00	1.00	0.00	−0.06	0.01	0.00	−0.04
DEPI	−0.01	−0.02	0.01	−0.02	0.00	0.00	−0.01	0.00	1.00	0.00	−0.01	0.00	0.01
SAI	−0.24	−0.15	−0.28	−0.06	0.02	0.00	0.06	−0.06	0.00	1.00	0.18	−0.22	−0.04
LEVI	−0.34	−0.22	−0.57	−0.19	−0.02	0.03	0.00	0.01	−0.01	0.18	1.00	−0.47	−0.05
TATA	0.51	0.52	0.82	0.12	0.00	0.00	0.02	0.00	0.00	−0.22	−0.47	1.00	0.01
Enterprise Type	−0.07	−0.08	−0.03	−0.08	0.01	0.01	−0.01	−0.04	0.01	−0.04	−0.05	0.01	1.00

Figure 2 Correlations of Variables

3.5 Statistics of Chinese Bond Market Data

From the processed market data, I first performed some basic statistical analysis. In this way I could have a direct understanding about the default condition in the Chinese market.

There are eleven industries among all the bond issuers. The default record number for each industry varies from zero to seventeen, as shown below. It is worth mentioning that the finance industry does not represent financial institutions, but only companies that focus on industrial investments.

Because the total dataset is going to be divided into a training set, a validation set, and a test set, the default records are too few to realize industry matched sampling, which is the method that pairs one default with a fixed number of non-defaults in the same industry. Therefore I did the sampling process without matching default and non-default records in each industry.

Industry / Bond Category	Material	Telecom	Real Estate	Manufacturing	Public Utility	Finance	Discretionary Consumption	Energy	Daily Consumption	IT	Health Care
Default	8	0	2	17	5	2	11	4	4	5	0
Nondefault	289	4	142	663	119	52	206	113	70	72	54
Total	297	4	144	680	124	54	217	117	74	77	54
Default Ra	3%	0%	1%	3%	4%	4%	1%	9%	5%	5%	9%

Figure 3 Industry Statistic of Sample

I also made a comparison of variable means between default and non-default records. The variable DSRI was largely affected by two outliers which are more than one thousand times larger than the trimmed average, so I excluded these outliers while calculating the average for DSRI. For all variables, I made a two-tailed test on the hypothesis that default mean differs from non-default mean with significance of 80%. The finding is that EBIT/Asset, SAI and LEVI have a mean difference between default and non-default that is statistically significant. Therefore, I consider that these three variables could make a difference in distinguishing defaults.

	Working. Capital /Asset	EBIT/Asset	BV. of. Equity/ Liability	DSRI	GMI	AQI	SGI	DEPI	SAI	LEVI	Enterprise Type
Default	-9.91	-11.44	48.17	1.16	1.31	1.02	0.94	1.35	1.65	1.18	0.14
Non-default	8.16	4.04	77.65	1.39	1.01	1.19	1.46	1.11	1.03	1.00	0.65
Standard Deviation of Non-default	22.71	4.88	70.36	4.93	1.79	2.03	1.40	2.11	0.35	0.14	0.48
Significant Difference in Mean		√							√	√	

Figure 4 Variable Means between Defaults and Non-defaults

CHAPTER 4: RESULTS

4.1 The Model

The logit regression model has been run on the preprocessed data set, and the outcome is shown below.

Coefficients:	Estimate	Std.Erro	z value	Pr(>|z|)
(Intercept)	-2.88	2.11	-1.36	0.17
Working.Capital/Asset	-0.02	0.01	-1.72	0.09
EBIT/Asset	-0.01	0.03	-0.38	0.71
BV.of.Equity/Liability	-0.01	0.01	-1.13	0.26
DSRI	-0.46	0.40	-1.16	0.25
GMI	-0.13	0.14	-0.98	0.33
AQI	-1.20	0.73	-1.65	0.10
SGI	-1.41	0.88	-1.61	0.11
DEPI	0.22	0.09	2.61	0.01
SAI	0.38	0.42	0.89	0.38
LEVI	3.85	1.53	2.51	0.01
Enterprise Type	-2.87	0.61	-4.67	0.00

Figure 5 Summary of Logit Regression Model, with 11 Prediction Variables

The more working capital a company has, compared to its total assets, the stronger ability it has to repay its debt, so Variable A is negatively related with default possibility.

AQI measures the percentage change of the soft asset portion from year t-1 to year t. From the Mscore Model, soft assets are those assets besides current assets and PP&E, and soft assets are usually more difficult to liquidate. Hence, higher AQI is evidence that a company's asset quality is decreasing and its default possibility is becoming greater. From the output of the logit model, however we can see a different conclusion. Variable G is negatively related with default possibility, which does not make economic sense. Future work could try to understand this counterintuitive association better.

SGI is higher simply when sales grow faster from year t-1 to year t. It is almost certain that a company with higher revenue would have higher debt-paying ability. The outcome matches my expectation: Variable H is negatively related with default possibility. This variable also has a very low p value, which supports the certainty of impact direction.

DEPI is higher when a company's depreciation expenses increase slower than revenue from year t-1 to year t. When companies are extending their useful lives, it could be a sign that they cannot afford new assets, a signal of possible default. Thus Variable I is positively related with default possibility.

The higher a company's leverage level, the more radical is its operation style. Debt repayment ability decreases as leverage level increases. LEVI measures the percentage that leverage level increases from year t-1 to year t. Thus Variable K is positively related with default possibility.

Enterprise Type is marked 1 for state-owned holding enterprises and 0 for others. Since the Chinese government has strong financial solvency, state-owned holding enterprises are easier to finance from outside and therefore can more easily survive debt crises. Thus Variable M is negatively related with default possibility. The p value for Enterprise Type is close to zero, which means this variable is very significant and has an unavoidable impact on default possibility.

Except for the variables mentioned above, the others are not significant . Variables C, D, and J have an impact direction that coordinates with my expectation: The more EBIT a company can earn with its fixed total asset scale, the more likely a company can pay its debt -- thus Variable C is negatively related with default possibility. The higher a company's equity is, compared to its liabilities, the more stable is its operation style. Variable D is therefore negatively related with default possibility. Higher SAI means higher increases in SG&A expenses as a portion of sales from year t-1 to year t, leading to worsening management or marketing efficiency. Variable J is thus positively correlated with default possibility. Variables E (DSRI) and F (GMI) are both negatively correlated with default possibility, which does not accord with accounting methods. I assume that they are mostly just noise.

Overall, EnterpriseType, LEVI, AQI, and SGI play the most important roles in the prediction, which indicates that leverage level and enterprise type are two crucial elements in predicting default. These two most important variables show that financing operation style and capability are the keys to debt solvency.

4.2 Findings

4.2.1 Model Accuracy

When investors face bond default, their loss on a single default bond would be large, typically 50% to 100% of total principal value. In contrast, if investors miss a non-default bond, they would at most only miss a profit of 3% to 10% of the principal value per year. The cost of misclassifying a default bond is much greater than that of misclassifying an actually non-default one. The logit regression model produces a predicted possibility of "default" for each record; I then use a cutoff value to make final predictions. The definition of cutoff value is that all records that have a predicted default possibility higher than the cutoff value are classified as "default"; others are classified as "non-default." In most statistical analysis the default value of cutoff is set at 50%, but to better capture the cost structure of bond investment described before, I chose cutoffs lower than usual.

When assessing the accuracy of the model, I used two indexes. One is Missed Default Rate, which is the rate of classifying default=1 as default=0. The other one is False Alarm Rate, which is the rate of classifying default=0 as default=1.

To minimize the cost related to failed detection of default bonds, I adjusted the cutoff value to reduce the missed default rate significantly, while keeping the false alarm rate under 40%.

I trained the model on the training data set and used the validation data set to find an optimal cutoff value. I started with 50% and reduced the cutoff to lower the missed default rate. When the cutoff reached 20%, the missed default rate began decreasing sharply, but the false alarm rate began increasing dramatically. As shown in the table below, though a cutoff of 20% provided a low false alarm rate of 1.86%, its high missed default rate made the prediction unusable. I kept reducing the cutoff and found that 5% is an optimal level where further lowering the cutoff value doesn't pick up many more defaults but ends up misclassifying many more non-defaults. At this optimal cutoff value, the missed default rate was reduced to 12% and the false alarm rate was kept below 20%. This optimal cutoff means that, when all the records with a predicted default possibility higher than 5% are classified as "default," the model detects 88.24% of the defaults one year in advance.

However, there are some individual default records that could hardly be detected by the model. The missed default rate could not be reduced to zero because of their existence. The reason is that not all default companies went bankrupt through a similar path. Some of the default companies just fell for special reasons. For example, the founder of CEFC (China Energy Company Limited) was arrested on bribery charges, causing CEFC to lose financing ability and to fail to repay its debt. The model doesn't have the ability to predict this kind of default.

Cutoff	Missed Default Rate	False Alarm Rate
20.00%	64.71%	1.86%
10.00%	41.18%	8.84%
5.00%	11.76%	19.53%
2.00%	11.76%	30.85%
1.50%	11.76%	34.73%
1.30%	11.76%	36.90%
1.00%	5.88%	41.55%
0.50%	5.88%	66.05%

Figure 6 Model Performance on Validation Data Set

Afterwards I tested the model on the test set, which is composed of data that was never used in training and validating the model. On the test set, the model kept its performance consistent： the missed default rate is 16.67% and the false alarm rate is 24.12%.

Missed Default Rate		16.67%	
False Alarm Rate		24.12%	

Confusion Matrix		Reference	
		0	1
Prediction	0	280	2
	1	89	10

Figure 7 Model Performance Confusion Matrix on Test Data Set

4.2.2 Comparison with Credit Rating

Next I compared the performance of my model with that of rating agencies.

In the Chinese capital market, there are ten mainstream credit rating agencies. Among them, nine are issuer-paying credit rating agencies, and one is an investor-paying credit rating agency. The five largest issuer-paying agencies are: China Chengxin Credit Rating Group; China Lianhe Credit Rating Co., Ltd.; Dagong Global Credit Rating Co., Ltd.; Shanghai Brilliance Credit Rating & Investor Service Co., Ltd.; and Golden Credit Services Co., Ltd. The only investor-paying agency is China Bond Rating Co., Ltd., and it usually issues lower credit ratings than the other agencies for the same bond issuer. The rating system used by the agencies includes nine credit levels, which are AAA, AA, A, BBB, BB, B, CCC, CC and C. Every level can be fine-tuned by using the symbols “+” and “-”.

Credit Level	Interpretation
AAA	Extremely strong capacity to meet financial commitments, almost not affected by negative economic circumstances, extremely low credit risk.
AA	Very strong capacity to meet financial commitments, only slightly affected by negative economic circumstances, very low credit risk.
A	Strong capacity to meet financial commitments, but somewhat susceptible to negative economic circumstances, low credit risk.
BBB	Average capacity to meet financial commitments, strongly affected by negative economic circumstances, average credit risk.
BB	Weak capacity to meet financial commitments, very strongly affected by negative economic circumstances, high credit risk.
B	Very weak capacity to meet financial commitments, strongly affected by negative economic circumstances, very high credit risk.
CCC	Currently vulnerable and dependent on favorable business, financial and economic conditions to meet financial commitments.
CC	Currently highly vulnerable to non-payment, and expected ultimate recovery is low.
C	Payment default on a financial commitment or breach of an imputed promise.

Figure 8 Credit Rating System in the Chinese Capital Market

Currently, most financial institutions use all the credit ratings given by mainstream rating agencies as reference. To obtain more stable data and as many rated records as I can, I chose to use the newest credit rating at the end of year t given by the five largest rating agencies as the prediction.

To compare the performance in an apple-to-apple way, I used the same sample set, which is the sum of the training set, validation set, and test set, to calculate credit rating accuracy.

Based on the definition of credit ratings used by the mainstream agencies, ratings above BBB mean "above average capacity to meet financial commitment." Thus I marked all ratings under A- (which is BBB and below) as 1 and regarded them as predictions of default. I built a confusion matrix with ratings given at the end of year t and actual default conditions. The overall accuracy, which is the total portion of correctly classified records, reached 98.01%, but unfortunately, it was high just because most of the bonds were non-default ones. The missed default rate for the rating agencies is 70.69%%, which means 70.69% of the defaulted bonds were rated higher than A-. The rating agencies failed to detect most of the default bonds, so I can say that the mainstream rating agencies do not provide reliable guidance for bond investment practice.

Missed Default Rate		70.69%	
False Alarm Rate		1.46%	
Boundary: Below A- as default			
Confusion Matrix		Reference	
		0	1
Prediction	0	1758	41
	1	26	17

Figure 9 Confusion Matrix of Credit Rating Given by Credit Agencies with AAA as Boundary

I then re-ran the logistic model over the whole data set. For the same data set, the missed default rate for the logistic model is only 15.52%, which means my model could catch more than 84% of the default bonds at least one year before their maturity. But the model was trained on the same data set, so this accuracy performance might be overstated.

Missed Default Rate		15.52%	
False Alarm Rate		21.41%	
Confusion Matrix		Reference	
		0	1
Prediction	0	1402	9
	1	382	49

Figure 10 Confusion Matrix of Logistic Model

Compared to credit rating agencies, my model did better in detecting default bond issuers in advance. But the credit rating agencies, in most cases, did better in distinguishing non-default bond issuers.

There is a possibility that the credit rating agencies did not work well because the boundary for their recognized default bond is not below A-. Hence, to rule out this possibility, I did similar analyses using AA, AA+, and AAA as boundaries. The results are displayed as follows.

	Missed Default Rate	False Alarm Rate
Logistic Model	16.67%	36.90%
A−	70.69%	1.46%
AA−	65.52%	4.32%
AA	56.90%	14.85%
AA+	18.97%	50.84%
AAA	6.90%	75.06%

Figure 11 Performance Credit Rating Given by Credit Agencies with Different Boundaries

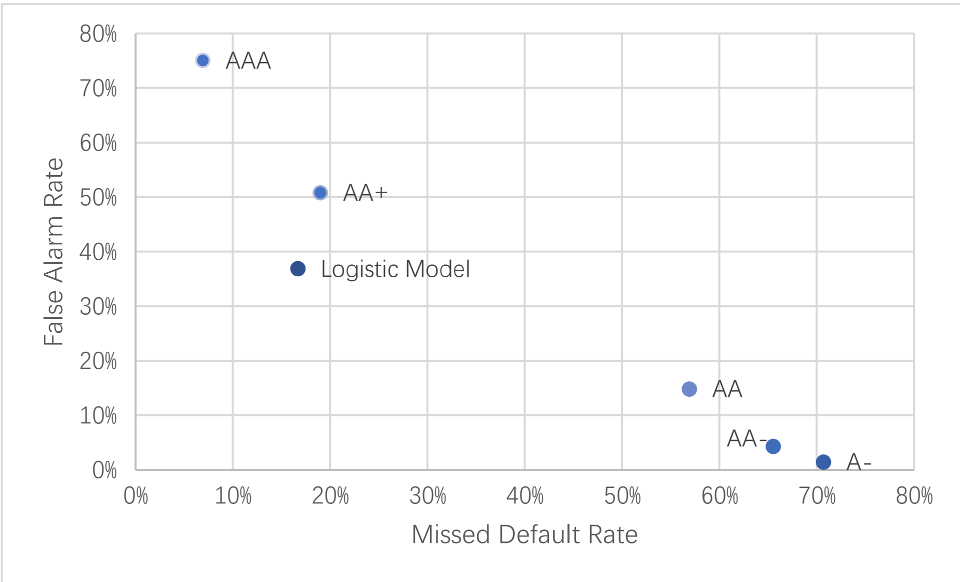

Figure 12 Scatterplot of Logistic Model and Credit Ratings

It's clear that as I lifted the boundary from A- to AA, the missed default rate decreased and the false alarm rate increased. The lower the two rates, the better the performance of classification. The logistic model could generate both a lower missed default rate and a lower false alarm rate compared to credit agencies when using AA+ as a boundary. In this dimension, the logistic model is dominant.

My logistic model analyzed all data as a whole first and then classified the records according to their predicted default possibility. Considering that the cost of a missed default is higher than a false alarm, I mainly focused on lowering the former rate. The cutoff value could be low enough to classify all the records as "default," but that would make the prediction useless. There is a trade-off between the missed default rate and the false alarm rate. I selected an optimal level manually to emphasize reducing the missed default rate but also to balance the two sides. If I had chosen to look at the false alarm rate first and kept it under 2%, then the logistic model would also have missed the majority of the default records.

The rating agencies, in contrast, look at each company separately and assess their comprehensive condition. They are more cautious in rating a bond issuer as a "junk bond" issuer.

In addition, both the logistic model and rating agencies are facing the same problem: good companies could fail in an unexpected way. So it is impossible to accurately forecast all the defaults in advance, at least for now.

4.3 A Potential Reference for Investors: Default Forecast for 2021H1 Existing Bonds

I collected data for all the existing corporate bonds on 2021/03/10. With all corporate bonds issued by financial institutions and municipal bonds removed, there are 1733 bonds with adequate raw financial data and credit ratings given by mainstream credit agencies. Since the annual financial reports for 2020 haven't been released, I used data from 2019 and 2018 annual financial reports. These records make up the prediction data set.

I then ran my model on the preprocessed prediction data set and came up with 403 bonds that are classified as "default" by the model. What is worth thinking about is that the latest credit ratings show that all these bonds that are forecast by the logistic model as "default" are currently rated higher than A-.

A list of these "dangerous" bonds is included in the appendix and might provide a useful reference for bond investors.

I made a comparison of the variable averages between the forecast default and non-default records for this sample. The significance test results showed that contribution scope changed: DEPI and Enterprise Type make the largest differences for whether a record is to be classified as default.

	Working.Capital /Asset	EBIT/Asset	BV.of.Equity/ Liability	DSRI	GMI	AQI	SGI	DEPI	SAI	LEVI	Enterprise Type
Variable Means of Default	11.02	5.22	64.49	1.11	0.71	0.93	0.90	1.82	1.03	1.06	0.06
Variable Means of Non-default	15.49	3.93	85.08	2.02	1.08	1.01	0.97	1.06	1.02	1.00	0.92
Standard Deviation of Non-default	26.87	3.70	84.34	28.00	3.19	0.43	0.42	0.60	0.35	0.11	0.27
Significant Difference in Mean								√			√

Figure 13 Variable Mean between 2021 Forecasted Defaults and Non-defaults

CHAPTER 5: INTEGRATIVE SUMMARY AND CRITIQUE

5.1 Conclusion of the Study

In this paper, I empirically investigated whether a logit regression model using individual financial and non-financial variables is an effective approach for assessing default risk.
From the performance statistics I can conclude that with downward adjustment of the cutoff value, the logit model can generate effective predictions that keep the missed default rate under 20%. In the meantime, the false alarm rate is between 20% and 25%. For this reason, it seems that the logit regression credit risk model could be used (with caution) when assessing a firm's credit quality.

I selected the most widely used credit risk assessment references in the Chinese bond market, the credit rating agencies, as the benchmark, and compared the benchmark with my logit model. I find that the performance of existing credit rating agencies is far from satisfactory—the ratings don't work well, failing to forecast some 70% of the default bonds. Run on the same data set, the logit model can detect more than 85% of the default bonds.

My logistic model and the rating agencies emphasize different aspects: my logistic model is better at detecting defaults, but the rating agencies are better at recognizing non-defaults. From this point of view, I therefore conclude that my model can add some value to the credit research of investors in the Chinese fixed income market.

5.2 Limitation of the Study and Future Expectations

This study leaves room for future research. The prediction is unsatisfying, and the causes can be the following.

In my opinion, one reason could be the lack of default observations, though there are enough non-default observations. It was in 2014 that the first bond default event occurred. If I select only the first default bond for each default issuer from 2014 to 2020, there are no more than 75 default observations (with complete financial data released) that can be collected. To obtain better results for prediction, it may be necessary to include more records in the sample data. An option can be to include some records of defaulted bank loans because bank loans have similar characteristics with bonds, or we can just wait for more default events to appear in the market.

Another reason that the model did not perform very well could be related with the selection of variables. Together with the development of economics, accounting and management, new tricks to manipulate financial reports are constantly showing up. The classic ratios used to detect potential financial fraud and value financial performance may not have enough predictability, making the model out-of-date. Thus it may be helpful to obtain new variables. Three possible choices of additional variables include new tools for financial report quality assessment, average market price and volatility of bonds, as well as data of news and public sentiments about the issuer.

In addition, my model is mainly based on the financial condition of issuers, but financial condition is not the only root cause in bond default. For instance, some industries have significant industrial cycles that can heavily impact their debt paying ability, but I classified all the bonds in the market without separating them into various industries. Analyzing bonds under different industry categories is more likely to capture industry-specific prediction variables.

Similarly, macroeconomic factors can affect the overall default rate of a specific period in the market. Taking time series and macroeconomic factors into consideration could increase the prediction power of the model.

REFRENCENS

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23(4), 589-609.

Altman, E. I. Haldeman R G, Narayanan P. (1977). ZETA analysis: a new model to identify bankruptcy risk of corporations. Journal of Banking and Finance, 1(1): 29-54

Altman, E. I. , and M. A. I. E. (2013). Predicting financial distress of companies: revisiting the Z-Score and ZETA models. Handbook of Research Methods & Applications in Empirical Finance. 428-465.

Beneish, M. D. (1999). The Detection of Earnings Manipulation. Financial Analysts Journal, 55(5), 24–36.

Black F, Cox J C. (1976). Valuing corporate securities: Some effects of bond indenture provisions. The Journal of Finance. 31(2): 351-367.

Blume, M., F. Lim and A.C. MacKinlay, (1998), “The Declining Credit Quality of U.S. Corporate Debt: myth or reality?”, Journal of Finance, 4, 1389-1413.

Campbell, J. and G. Taksler, (2002), Equity Volatility and Corporate Bond Yields. Working paper, Department of Economics, Harvard University

Central Clearing Corporation, (2020). Monthly Bulletin of Statistics [统计月报]. Retrieved 18th May 2020, from https://www.chinabond.com.cn/Channel/19012917

Chen B. and Liu H J. (2012). Which Kind of Companies are on Suspicion of Financial Fraud?

China People’s Bank. (2019). China People’s Bank 2018 Annual Report [中国人民银行 2018 年报]. China People’s Bank.

China People’s Bank. (2011). China People’s Bank 2010 Annual Report [中国人民银行 2010 年报]. China People’s Bank.

DECHOW, P. M., GE, W., LARSON, C. R., & SLOAN, R. G. (2011). Predicting Material Accounting Misstatements*. Contemporary Accounting Research, 28(1), 17–82. doi:10.1111/j.1911-3846.2010.01041.x

Du, Y. (2004). Credit rating, default probability and structural credit risk models (Order No. NQ99984). Available from ProQuest Dissertations & Theses Global; ProQuest Dissertations & Theses Global A&I: The Humanities and Social Sciences Collection. (305087853). Retrieved from https://search.proquest.com/docview/305087853?accountid=10025

Goss, E. P., & Ramchandani, H. (1995). Comparing classification accuracy of neural networks, binary logit regression and discriminant analysis for insolvency prediction of life insurers. Journal of Economics and Finance, 19(3), 1–18. doi:10.1007/bf02920611

Jin L L. (2005). Empirical Study of Domestic Commercial Bank Credit Risks Quantitative Management based on CreditMetrics. [基于 Credit Metrics 的我国商业银行信用风险量化管理实证研究]. Master's thesis. Northeastern University. Shenyang. Liaoning Province. China

Lando, D. (2004). Credit risk modeling. Master's Thesis. Princeton University. Princeton, New Jersey.

Longstaff F A, Schwartz E S. (1995). A simple approach to valuing risky fixed and floating rate debt. The Journal of Finance. 50(3): 7789-819

Martin, D. (1977). Early warning of bank failure. Journal of Banking & Finance, 1(3), 249–276. doi:10.1016/0378-4266(77)90022-x

Merton R C. (1974). On the pricing of corporate securities: The risk structure of interest rates. The journal of Finance. 29(2): 449-470.

National Inter-bank Funding Center. (2020). Local Currency Market Member Statistics [本币市场成员统计]. Retrieved 18th May 2020, from http://www.chinamoney.com.cn/chinese/mtmemrmb/.

Norden, L., & Weber, M. (2004). Informational efficiency of credit default swap and stock markets: The impact of credit rating announcements. Journal of Banking & Finance, 28(11), 2813–2843. doi:10.1016/j.jbankfin.2004.06.011

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, 18(1), 109. doi:10.2307/2490395

Pang S L. (2006). The Application of Logistic Regression Model in Credit Risk Analysis [Logistic 回归模型在信用风险分析中的应用]. Mathematics in Practice and Theory. 2006(9): 129-137

Shanghai Security Exchange. (2020). Bond Data Statistic [债券数据统计]. Retrieved 18th May 2020, from http://www.sse.com.cn/market/bonddata/statistic/.

Shenzhen Security Exchange, (2020). Market Data Statistic [市场数据统计]. Retrieved 18th May 2020, from http://www.szse.cn/market/overview/index.html.

Shanghai Clearing House, (2020). Monthly Bulletin of Statistics [统计月报]. Retrieved 18th

May 2020, from https://www.shclearing.com/sjtj/tjyb/.

Shmueli, G. (2018). In Data mining for business analytics: concepts, techniques, and applications in R. essay, John Wiley & Sons.

Xiang D W. (2002). Empirical Study of Using Zscore Model to Assess Operation Risk of Listed Companies. [运用”Z 计分法”评价上市公司经营风险的实证研究]. Accounting Research. 2002 (11): 53-57.

Wind Information, (2020). Economic Data Base (EDB) [经济数据库]. Retrieved 31th May 2020, from EDB: Factor ID: G0000888; S0059746.

Zheng C. (2005). Research on Credit Retreat of Commercial Bank from the Perspective of Portfolio Management——Based on the Analysis of the Application of Creditrisk+ Model. [组合管理视角下银行信贷的退出——基于 Credit risk+模型的应用分析]. Finance Forum. 14(02): 34-41

APPENDIX A：R Code for Research

```
#Pre-process
data <- read.csv('data.csv')
for (k in 2:14) {data[,k]<-as.numeric(data[,k])}
data$Default<-as.numeric(data$Default)
default <- data[data$Default==1,]
non-default <- data[data$Default!=1,]
non-default <- na.omit(non-default)
dim(non-default)
summary(is.na(default))
default[is.na(default$DEPI),'DEPI'] <- median(data$DEPI,na.rm = T)
default <- na.omit(default)
dim(default)
summary(default$Industry)
summary(non-default$Industry)
#######################################################################
#partition
set.seed(1)
trainrow.d <- sample(row.names(default),29)
trainrow.nd <- sample(row.names(non-default),29*25)
train.d <- default[trainrow.d,]
train.nd <- non-default[trainrow.nd,]
train <- rbind(train.d,train.nd)
validrow.d <- sample(setdiff(row.names(default),trainrow.d),17)
validrow.nd <- sample(setdiff(row.names(non-default),trainrow.nd),21*1784/58)
valid.d <- default[validrow.d,]
valid.nd <- non-default[validrow.nd,]
valid <- rbind(valid.d,valid.nd)
testrow.d <- setdiff(row.names(default),union(trainrow.d,validrow.d))
testrow.nd <- sample(setdiff(row.names(non-default),union(trainrow.nd,validrow.nd)),12*1784/58)
test.d <- default[testrow.d,]
test.nd <- non-default[testrow.nd,]
test <- rbind(test.d,test.nd)
##############################################################
data.sample <- rbind(default,non-default)
summary(data.new[,2:14])
cor(data.new[,2:14])>0.5
##############################################################
#run logit regression
default.logit <- glm(Default ~., data = train, family = "binomial")
options(scipen=999)
summary(default.logit)
```

```
pred.prob.logit<-predict(default.logit,valid,type = "response")
library(caret)
cut<-c(0.2,0.1,0.05,0.02,0.015,0.013,0.01,0.005)
accuracy.logit<-cbind.data.frame(cut,c(1:8),c(1:8))
for (i in 1:8) {
  cutoff=cut[i]
  pred.logit<-ifelse(pred.prob.logit>cutoff,1,0)
    cmlogit<-confusionMatrix(as.factor(pred.logit),factor(valid$Default))
  accuracy.logit[i,2]<- cmlogit$table[1,2]/(cmlogit$table[1,2]+cmlogit$table[2,2])
  accuracy.logit[i,3]<-cmlogit$table[2,1]/(cmlogit$table[2,1]+cmlogit$table[1,1])}
colnames(accuracy.logit)<-c("Cutoff","Missed Default Rate","False Alarm Rate")
# get the confusion matrix for test set
pred.logit.test<-ifelse(pred.prob.logit2>0.05,1,0)
cmlogit.test<-confusionMatrix(as.factor(pred.logit.test),factor(test$Default))
############################################################
#compare with the rating agencies
pred.prob.r<-predict(default.logit,data.sample,type = "response")
pred.r<-ifelse(pred.prob.r>0.05,1,0)
cmr<-confusionMatrix(as.factor(pred.r),factor(data.sample$Default))
high.level<-levels(data.sample$CreditRating)[1:7]
data.sample$ratingaj<-ifelse(data.sample$CreditRating %in% high.level,0,1)
cmraj<-confusionMatrix(as.factor(data.sample$ratingaj),factor(data.sample$Default))
############################################################
#statistic analysis of market data
market.data<-data.frame(data.sample,pred.r)
for (k in 5:16) {market.data[,k]<-as.numeric(market.data[,k])}
default.m <- market.data[market.data$pred.r ==1,]
non-default.m<- market.data[market.data$pred.r !=1,]
sapply(default.m[5:16],FUN=mean)
sapply(non-default.m[5:16],FUN=mean)
############################################################
#predict
virgin <- read.csv('vset.csv')
virgin<-na.omit(virgin)
pred.p.vrigin<-predict(default.logit,virgin,type = "response")
pred.virgin<-ifelse(pred.p.vrigin>0.013,1,0)
conclusion<-data.frame(virgin, pred.virgin)
write.csv(conclusion,'conclusion.csv')
conclusion$Industry<-as.factor(conclusion$Industry)
default.v <- conclusion[conclusion$pred.virgin ==1,]
non-default.v <- conclusion[conclusion$pred.virgin!=1,]
sapply(default.v[8:21],FUN=mean)
sapply(non-default.v[8:21],FUN=mean)
```

APPENDIX B：List of Forecasted Default Bonds in H1 2021

ID	NAME	MaturityDate	ID	NAME	MaturityDate	ID	NAME	MaturityDate
012100433. IB	21振石SCP001	2021-7-26	012001669. IB	20迪安诊断(疫情防控债)SCP001	2021-3-30	101800322. IB	18均胜电子MTN001	2021-3-29
012001974. IB	20皖供销SCP001	2021-8-8	012100493. IB	21中联重科SCP001	2021-4-23	101800324. IB	18美凯龙MTN001	2023-3-30
012002350. IB	20泰豪科技SCP002	2021-3-27	012100590. IB	21长电科技SCP001	2021-5-8	101800365. IB	18致达MTN001	2021-4-11
012002449. IB	20天安数码SCP001	2021-4-9	012100715. IB	21复星医药SCP001	2021-5-27	101800373. IB	18南山集MTN001	2021-4-12
012002465. IB	20红狮SCP004	2021-3-25	012100820. IB	21正泰SCP001	2021-8-31	101800401. IB	18万达MTN001	2021-4-16
012002484. IB	20均瑶SCP006	2021-4-12	012100940. IB	21九龙园SCP002	2021-12-7	101800419. IB	18联合光伏MTN001	2021-4-18
012002572. IB	20铁塔股份SCP007	2021-4-16	031759010. IB	17新城控股PPN001	2022-5-17	101800510. IB	18金光纸业MTN001	2021-4-26
012002629. IB	20腾越建筑SCP002	2021-4-24	031782010. IB	17荣盛地产PPN001	2022-11-21	101800580. IB	18南玻MTN001	2021-5-4
012002645. IB	20东阳光SCP003	2021-4-26	031800302. IB	18中信国安PPN002	2021-4-27	101800588. IB	18天士力医MTN001	2021-5-3
012002759. IB	20浙手工SCP002	2021-5-7	031800354. IB	18康达环保PPN001	2021-6-13	101800683. IB	18福星科技MTN001	2021-6-15
012002816. IB	20华邦健康SCP003	2021-5-11	031800634. IB	18七匹狼PPN001	2021-11-2	101800740. IB	18万科MTN001	2021-7-12
012002823. IB	20花园SCP002	2021-5-11	031800654. IB	18汉当科PPN001	2023-11-14	101800893. IB	18牧原MTN001	2021-8-17
012002830. IB	20广汇汽车SCP001	2021-5-11	031800706. IB	18天士力PPN001	2021-11-29	101800904. IB	18吉利MTN002	2021-8-17
012002912. IB	20新希望SCP004	2021-5-14	031900235. IB	19阳光城PPN001	2022-3-22	101800914. IB	18中国国贸MTN001	2023-8-22
012002916. IB	20鲁钢铁SCP006	2021-5-18	031900619. IB	19广州城建PPN001	2024-8-15	101800979. IB	18广汇实业MTN001	2021-8-29
012002960. IB	20华立SCP002	2021-5-21	031900832. IB	19南翔贸易PPN001	2022-11-14	101801019. IB	18杉杉MTN002	2021-9-5
012002991. IB	20中南建设SCP001	2021-5-22	032000068. IB	20沪世茂PPN001	2023-1-19	101801061. IB	18三星MTN001	2021-9-12
012003008. IB	20利港SCP004	2021-5-23	032000323. IB	20华联投资(疫情防控债)PPN001	2021-4-10	101801080. IB	18顺丰泰森MTN001	2021-9-19
012003150. IB	20桐昆控股SCP002	2021-5-31	032000374. IB	20富力地产PPN001	2024-4-23	101801157. IB	18海运集装MTN001	2021-10-17
012003189. IB	20宝龙SCP002	2021-6-7	032000384. IB	20远洋控股PPN001	2023-4-24	101801300. IB	18岚桥MTN001	2021-11-6
012003205. IB	20复星高科SCP003	2021-6-8	032000614. IB	20城阳开投PPN001	2023-7-6	101801347. IB	18正才MTN001	2021-11-21
012003381. IB	20平安不动SCP007	2021-3-24	032100299. IB	21富阳交通PPN002	2026-3-12	101801408. IB	18TCL集MTN001	2021-12-3
012003533. IB	20科伦SCP003	2021-4-14	042000108. IB	20碧水源(疫情防控债)CP001	2021-3-19	101801416. IB	18伊泰MTN001	2021-12-5
012003898. IB	20闽东百SCP003	2021-5-10	042000213. IB	20富通CP001	2021-4-27	101801458. IB	18湛江港MTN001	2021-12-10
012003957. IB	20上海大众SCP004	2021-5-15	042000227. IB	20五江轻化CP001	2021-5-6	101801506. IB	18冠城大通MTN001	2021-12-19
012004129. IB	20桐昆SCP004	2021-5-25	042000314. IB	20阳谷祥光CP001	2021-7-16	101801515. IB	18紫江MTN001	2021-12-18
012004135. IB	20亨通SCP004	2021-6-3	042000327. IB	20蓝光CP001	2021-7-29	101801531. IB	18奥德MTN001	2021-12-24
012004147. IB	20中粮SCP006	2021-8-27	042000442. IB	20紫江CP001	2021-4-19	101801550. IB	18金风科技MTN001	2021-12-26
012004155. IB	20红豆SCP002	2021-5-31	042000516. IB	20滨江房产CP004	2021-12-10	101900001. IB	19宝安集MTN001	2022-1-4
012004169. IB	20东航股SCP033	2021-6-1	068032. IB	06沪水务债	2021-6-29	101900053. IB	19南京钢铁MTN001	2022-1-18
012004180. IB	20银建SCP002	2021-6-2	081800117. IB	18保利置业ABN001优先A	2038-9-10	101900198. IB	19迪马实业MTN001	2022-2-21
012004246. IB	20威高SCP003	2021-9-7	081900208. IB	19美凯龙ABN001优先	2037-5-7	101900257. IB	19步步高MTN001	2022-3-1
012004322. IB	20中天建设SCP003	2021-9-14	081900310. IB	19比亚迪ABN001优先	2022-8-26	101900289. IB	19格力MTN001	2022-3-8
012004395. IB	20金东纸业SCP002	2021-9-24	082000300. IB	20华凌工贸ABN001优先A2	2021-7-15	101900344. IB	19深航空MTN001	2022-3-18
012004398. IB	20宁波建工SCP001	2021-6-23	082001080. IB	20山河智能ABN001优先A1	2021-10-20	101900385. IB	19南通三建MTN001	2022-3-22
012004406. IB	20华强SCP002	2021-9-23	082100007. IB	21九州通ABN001优先A	2022-11-8	101900461. IB	19圆通蛟龙MTN001	2022-4-3
012004413. IB	20福建阳光SCP003	2021-9-24	101652016. IB	16大连万达MTN001	2021-3-29	101900557. IB	19华南工业MTN001	2022-4-18
012100089. IB	21金科地产SCP001	2021-9-25	101652033. IB	16金地MTN003	2021-8-18	101900631. IB	19居然之家MTN002	2021-4-29
012100093. IB	21亚厦SCP001	2021-9-30	101654068. IB	16苏沙钢MTN001	2021-8-9	101900701. IB	19尧柏水泥MTN001	2022-5-8
012100096. IB	21百业源SCP001	2021-8-9	101654096. IB	16鲁宏桥MTN001	2021-10-26	101900827. IB	19中建新疆MTN001	2022-6-24
012100142. IB	21珠江投管SCP001	2021-10-11	101663007. IB	16阿波罗MTN001	2021-3-23	102000019. IB	20佳源创盛MTN001	2023-1-9
012100153. IB	21立讯精工SCP001	2021-10-12	102000118. IB	20福耀(疫情防控债)MTN001	2023-2-17	101674002. IB	16隆鑫MTN001	2021-5-6
012100232. IB	21格盟SCP001	2021-10-16	101758032. IB	17兴湘投资MTN001	2022-8-28	102000209. IB	20永达MTN001	2023-3-18
012100254. IB	21南航集SCP001	2021-3-19	102000218. IB	20特变(疫情防控债)MTN001	2023-3-4	101780003. IB	17幸福基业MTN001	2022-5-23
012100271. IB	21东航SCP002	2021-3-23	102000264. IB	20希望六和(疫情防控债)MTN001	2023-3-9	101800045. IB	18心连心MTN001	2023-1-25
012100304. IB	21南通二建SCP001	2021-10-18	101800232. IB	18奥克斯MTN001	2021-3-20	102000322. IB	20苏宁易购MTN001	2023-3-13
012100309. IB	21豫园商城SCP001	2021-5-21	101800252. IB	18中瑞实业MTN001	2021-3-21	102000348. IB	20中天金融MTN001	2023-3-12
012100337. IB	21中燃投资SCP001	2021-3-23	101800311. IB	18德力西MTN001	2021-3-28	102000354. IB	20歌尔MTN001	2023-3-13
012100369. IB	21南航股SCP003	2021-3-26	102000366. IB	20特变股份(疫情防控债)MTN001	2023-3-17	101800313. IB	18顾家MTN001	2021-4-4
102000423. IB	20沪打捞MTN001	2023-3-20	136301. SH	16龙盛03	2021-3-17	116364. SZ	彩1号优5	2021-8-30
102000501. IB	20锦州港MTN001	2023-3-25	136307. SH	16协信03	2021-3-17	116599. SZ	17华联04	2021-7-5
102000596. IB	20传化MTN001	2023-4-9	136317. SH	15智慧01	2021-4-5	117118. SZ	18大族EB	2021-8-24
102000638. IB	20乐普MTN001	2023-4-13	136322. SH	16宇通02	2023-3-22	117168. SZ	20三花EB	2023-7-24
136347. SH	16永利债	2021-3-28	102000685. IB	20人福(疫情防控债)MTN001	2022-4-17	117179. SZ	20中希E1	2023-12-25
136374. SH	16建业01	2021-4-12	102001038. IB	20康缘集(疫情防控债)MTN001	2022-5-22	1180183. IB	11中国泛海债02	2021-12-13
102001140. IB	20中兴新MTN001	2025-6-10	136405. SH	14亿利02	2021-4-26	118602. SZ	16万通03	2021-4-7
102001144. IB	20海亮MTN001	2022-6-10	136421. SH	16春秋01	2021-6-2	118647. SZ	16穗建03	2023-4-28
102001191. IB	20东菱凯琴MTN001	2023-6-18	136433. SH	16晟晏债	2021-5-19	118704. SZ	16隆地02	2021-6-8
102001216. IB	20通威MTN001	2023-6-19	136457. SH	16希望01	2021-5-30	122393. SH	15恒大03	2022-7-8
102001358. IB	20龙湖拓展MTN001A	2023-7-15	136475. SH	16华宇01	2021-6-8	122496. SH	15世茂02	2022-10-16
102001495. IB	20温氏食品MTN001	2023-8-11	136521. SH	16鸿坤01	2021-7-8	123002. SZ	国祯转债	2023-11-24
102001749. IB	20中化岩土MTN001	2025-9-7	136567. SH	16凯华01	2021-7-22	123004. SZ	铁汉转债	2023-12-18
102002144. IB	20兆泰MTN001	2025-11-16	136592. SH	16鄂稻01	2021-8-4	123007. SZ	道氏转债	2023-12-28
110045. SH	海澜转债	2024-7-13	136632. SH	16亚洲浆	2021-8-24	123010. SZ	博世转债	2024-7-5
110047. SH	山鹰转债	2024-11-21	136647. SH	16华新01	2021-8-22	123011. SZ	德尔转债	2024-7-18
110056. SH	亨通转债	2025-3-19	136662. SH	16友阿01	2021-8-24	123012. SZ	万顺转债	2024-7-20
110068. SH	龙净转债	2026-3-24	136690. SH	16恒安01	2021-9-8	123013. SZ	横河转债	2024-7-26

ID	NAME	MaturityDate	ID	NAME	MaturityDate	ID	NAME	MaturityDate
110076.SH	华海转债	2026-11-2	136744.SH	16祥源债	2021-9-29	123014.SZ	凯发转债	2023-7-27
111069.SZ	16格林G1	2023-10-31	136749.SH	G16博天	2021-10-12	123015.SZ	蓝盾转债	2024-8-13
111072.SZ	18美尚01	2025-9-11	136792.SH	16中筑01	2021-11-7	123018.SZ	溢利转债	2024-12-20
111082.SZ	19木森G1	2024-6-3	136842.SH	16银鹰01	2021-11-23	123019.SH	PR阳纸业	2021-7-21
112229.SZ	14白药01	2021-10-16	136850.SH	16宝丰02	2021-11-23	123023.SZ	迪森转债	2025-3-20
112359.SZ	16魏桥03	2021-3-22	136877.SH	16合盛01	2021-12-14	123025.SZ	精测转债	2025-3-29
112367.SZ	16华西01	2021-3-28	137074.SZ	君玺1优	2022-9-5	123027.SZ	蓝晓转债	2025-6-11
112370.SZ	16新纶债	2021-3-30	137080.SH	19美04EB	2022-5-10	123028.SZ	清水转债	2025-6-19
112373.SZ	16奥瑞金	2021-4-11	137104.SH	20卧龙EB	2023-5-8	123029.SZ	英科转债	2025-8-16
112378.SZ	16盛润债	2021-4-12	137288.SZ	20雅居1A	2023-10-28	123049.SZ	维尔转债	2026-4-13
112383.SZ	16当代债	2021-4-20	137563.SZ	21路劲A	2032-11-4	123058.SZ	欣旺转债	2026-7-14
112391.SZ	16龙基02	2021-5-19	138247.SZ	粤湾01优	2021-12-15	123061.SZ	航新转债	2026-7-22
112396.SZ	16华美02	2021-5-26	138385.SZ	20奥创优	2022-1-20	123092.SZ	天壕转债	2026-12-24
112421.SZ	16信利01	2021-7-28	138945.SZ	20新尾A	2021-8-3	123094.SZ	星源转2	2027-1-20
112426.SZ	16东林02	2021-8-10	143037.SH	17中科01	2022-3-28	123095.SZ	日升转债	
112434.SZ	16正商03	2021-8-26	143139.SH	17长园债	2022-7-13	124018.SZ	易成定转	2026-12-16
112435.SZ	16绿景01	2021-8-26	143143.SH	17鹏博债	2022-6-16	124130.SH	13陕东岭	2023-1-15
112451.SZ	16利德01	2021-9-23	143149.SH	17广汇01	2022-6-22	124815.SH	14天瑞02	2024-6-25
112460.SZ	16安控债	2021-10-24	143363.SH	18广汇G1	2021-8-8	127003.SZ	海印转债	2022-6-8
112461.SZ	16龙控02	2021-10-21	143389.SH	17永钢01	2022-11-13	127004.SZ	模塑转债	2023-6-2
112473.SZ	16海普瑞	2021-11-8	143418.SH	17亚通01	2022-12-19	127013.SZ	创维转债	2025-4-15
112474.SZ	16合泰01	2021-11-9	143494.SH	18香江01	2022-3-9	127144.SH	15黄河债	2021-3-27
112483.SZ	16宝新01	2021-11-28	143514.SH	PR璞泰02	2021-3-19	127325.SH	15凯嘉债	2021-11-30
112489.SZ	16嘉美债	2021-12-8	143539.SH	18凤祥01	2023-3-29	127839.SH	G18树业	2022-8-14
112511.SZ	17正通01	2022-3-24	143595.SH	18君华01	2021-5-4	128015.SZ	久其转债	2023-6-8
112522.SZ	17阳普S1	2022-4-28	143617.SH	18迈科01	2021-4-26	128018.SZ	时达转债	2023-11-6
112532.SZ	17凯撒03	2022-6-16	143644.SH	18复地01	2021-8-27	128022.SZ	众信转债	2023-12-1
112548.SZ	17兴森01	2022-7-19	143664.SH	18保集01	2021-6-7	128028.SZ	赣锋转债	2023-12-21
112576.SZ	17齐科01	2022-8-24	143673.SH	18伊泰01	2021-6-8	128029.SZ	太阳转债	2022-12-22
112612.SZ	17正邦01	2022-11-17	143708.SH	18中庚G1	2023-6-29	128032.SZ	双环转债	2023-12-25
112620.SZ	17制药02	2022-11-23	145418.SH	17融禾01	2022-3-21	128035.SZ	大族转债	2024-2-6
112623.SZ	17丽鹏G1	2022-12-1	145707.SH	17九通01	2021-8-17	128036.SZ	金农转债	2024-3-9
112624.SZ	17鑫能G1	2022-12-7	149027.SZ	20荣安01	2025-1-13	128039.SZ	三力转债	2024-6-8
112631.SZ	18英唐01	2022-1-3	149060.SZ	20清新G1	2025-3-13	128040.SZ	华通转债	2024-6-14
112658.SZ	18朗姿01	2023-3-19	149118.SZ	20金桂01	2025-4-30	128041.SZ	盛路转债	2024-7-17
112661.SZ	18传化01	2021-3-21	149338.SZ	20海能01	2023-12-29	128042.SZ	凯中转债	2024-7-30
112680.SZ	18物美01	2021-4-17	149898.SH	PR18红1B	2021-8-30	128044.SZ	岭南转债	2024-8-14
112684.SZ	18联创债	2023-4-20	150369.SH	18东投01	2021-5-11	128046.SZ	利尔转债	2024-10-17
112827.SZ	18卓越06	2021-12-18	150390.SH	18俊发01	2022-10-22	128049.SZ	华源转债	2024-11-27
112892.SZ	19宝德01	2022-11-20	150423.SH	18海伟01	2021-5-24	128050.SZ	钧达转债	2024-12-10
112903.SZ	19电科01	2022-5-14	150600.SH	18明诚03	2021-8-15	128051.SZ	光华转债	2024-12-14
112908.SZ	19沃尔01	2022-5-29	150880.SH	18建资01	2023-11-30	128056.SZ	今飞转债	2025-2-28
112942.SZ	19中骏01	2023-8-1	151060.SH	18鑫业02	2022-1-4	128071.SZ	合兴转债	2025-8-16
113012.SH	骆驼转债	2023-3-24	151234.SH	19翔宇01	2022-3-8	128079.SZ	英联转债	2025-10-21
113016.SH	小康转债	2023-11-6	151287.SH	G19高能1	2022-3-14	128087.SZ	孚日转债	2025-12-17
113029.SH	明阳转债	2025-12-16	151629.SH	19天域01	2022-6-12	128090.SZ	汽模转2	2025-12-27
113038.SH	隆20转债	2026-7-31	155092.SH	18花样年	2021-12-17	128130.SZ	景兴转债	2026-8-31
113505.SH	杭电转债	2024-3-6	155375.SH	19中天01	2022-4-24	128131.SZ	崇达转2	2026-9-7
113508.SH	新凤转债	2024-4-26	155574.SH	19江河01	2021-7-30	131800003.IB	18东华能源GN001	2021-6-21
113519.SH	长久转债	2024-11-7	155735.SH	19国贸01	2024-9-26	131900010.IB	19新疆能源GN001	2022-6-17
113524.SH	奇精转债	2024-12-14	159650.SH	19宝龙A	2037-9-6	132023.SH	20幸福EB	
113526.SH	联泰转债	2025-1-23	159924.SH	如皋优	2021-10-11	135842.SH	16旭辉02	2021-9-23
113527.SH	维格转债	2025-1-24	1624005.IB	16随供水项目NPB	2023-3-28	135862.SH	16花园02	2021-9-29
113530.SH	大丰转债	2025-3-27	1624043.IB	16东北袜业项目NPB01	2021-12-19	136101.SH	15合景01	2021-12-17
113534.SH	鼎胜转债	2025-4-9	162461.SH	19平金01	2024-11-7	136106.SH	15三友02	2022-12-17
113578.SH	全筑转债	2026-4-20	162797.SH	19深业01	2024-12-25	136135.SH	16联泰01	2022-1-6
113588.SH	润达转债	2026-6-17	163796.SH	20茂业01	2022-8-4	136139.SH	16国美01	2022-1-7
113595.SH	花王转债	2026-7-21	166064.SH	20叠石02	2022-3-2	136170.SH	16景瑞01	2021-3-17
113602.SH	景20转债	2026-8-24	167265.SH	20福华01	2023-7-17	136249.SH	16海怡01	2021-3-30
113605.SH	大参转债	2026-10-22	168040.SH	20德清A1	2021-3-19	136296.SH	16珠投04	2021-3-16
113607.SH	伟20转债	2026-11-2	168233.SH	迪马优A	2038-4-21	114550.SZ	19融投01	2023-8-26
114010.SZ	16华源01	2021-10-24	169125.SH	世欧优1	2038-7-14	114645.SZ	19新国都	2022-12-27
114365.SZ	18融创01	2021-8-28	169432.SH	20龙元01	2023-3-27	114829.SZ	20西建01	2023-9-29
114389.SZ	18碧桂01	2021-11-16	175481.SH	20名城债	2023-12-1	116288.SZ	融侨A10	2021-5-30
114427.SZ	19唐德01	2022-1-18	175505.SH	20方圆01	2024-12-3	1880001.IB	18国轩绿色债01	2023-4-13
114469.SZ	19洪涛01	2022-4-24	175873.SH	21星发01	2026-3-17	114520.SZ	19科陆01	2022-7-11
1780324.IB	17能兴专项债	2022-10-16						