Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 221 (2023) 300–307

Procedia Computer Science

www.elsevier.com/locate/procedia

10th International Conference on Information Technology and Quantitative Management

Research on Default Prediction Model of Corporate Credit Risk Based on Big Data Analysis Algorithm

Qingyan Xianyu, Mo Hai*

School of Information, Central University of Finance and Economics, Beijing, 100081,China

Abstract

In recent years, with the rise of many technologies such as big data and artificial intelligence, the digitalization and information transformation of enterprises have gradually penetrated into the financial industry. Based on different big data analysis algorithms, we aim to establish a default prediction model for corporate credit risk, further optimize different models, compare the model performance, and analyze the robustness of the optimal model. We collect 21 items of financial and non-financial index data from more than 1,000 listed companies, standardize, balance and normalize the data, use correlation coefficient to screen the index, and establish two in-depth learning models, convolutional neural network model and recurrent neural network model, based on Pytorch framework of Spark platform, to complete the model optimization, and compare them with two traditional machine learning models: random forest model and logistic regression model. Finally, the comparison experiment shows that the recurrent neural network is the optimal model with an accuracy rate of 0.93, a recall rate of 0.96 and a F1 value of 0.93. For the optimal recurrent neural network model, the robustness of the model is analyzed by modifying the number of indicators, changing the number of samples and eliminating non-financial factors. The results show that the evaluation indicators of the model do not change much under the three conditions, and the model has a good robustness.

Keywords: Default Prediction Model; corporate credit risk; deep learning; big data; Spark

1. Background introduction

1.1. Research and application background

In order to ensure the business safety of commercial banks and reduce the occurrence probability of default events, it is necessary to evaluate the credit of lending enterprises. The traditional credit evaluation method needs to collect a lot of information, which not only consumes a lot of manpower, material resources and time, but also brings huge cost, and also has the problems of inaccurate and untimely evaluation.

On this basis, it is necessary to incorporate the financial technology of big data analysis into the risk management mechanism, build an intelligent management mechanism for preventing and controlling bank risks, and explore new ways to manage and evaluate bank risks. Based on big data analysis, data mining, machine learning and other methods, this study uses business information, registered legal persons and core teams, litigation information, intellectual property rights, tax assessment, corporate magnitude, financial data and other information to intelligently assess the credit default risk of lending enterprises, and establishes a default probability prediction model, which is compared and optimized with the traditional enterprise credit risk prediction model.

* Corresponding author:Hai Mo.
E-mail address: haimo@cufe.edu.cn.

1877-0509
10.1016/j.procs.2023.07.041

1.2. Summary of related work

1. Random forests

Random forest is an integrated learning algorithm, which classifies samples by constructing multiple decision tree models. The term was first proposed by Tin Kam Ho of bell laboratory in 1995[1].

In 2010, Fang Kuangnan et al[2] studied the principle of random forest (RF) method, which shows that the RF method has been applied to the field of individual credit risk assessment under the condition of asymmetric information. In 2017, Li Xiangpei[3] simulated the parametric and non-parametric measurement of corporate credit risk based on the research on the change of market value of corporate assets over time. In 2019, Zhu Yinong[4] analyzed the credit default phenomenon in China's bond market and studied it through the background of phased development.

2. Logical regression

Logistic regression is a generalized linear regression analysis model. The earliest concept of "regression" was put forward by Gao Erdun in 1886 to study the relationship between father and son's height[5].

In 2013, Tan Yanni[6] used data mining method to identify outliers, proposed the method of constructing credit risk assessment model of listed companies in China, established decision tree and logistic regression model, and verified the model. In 2020, Li Siqi[7] used simple and classic logistic regression analysis, support vector machine, BP neural network, Light GBM and XGBOOST to predict, established a prediction model based on three analysis methods, and compared the chaotic matrix generated by its final prediction results with the ROC curve. In 2020, Yuan Jiang[8] used logistic regression, support vector machine, BP neural network, XGBOOST and Light GBM to study the prediction performance of each model in bond credit default.

3. Convolution neural network

Convolution neural network is a kind of feed-forward neural network with depth structure. As early as 1995, the concept of convolution neural network was proposed. Time delay network and LeNet-5 are the earliest convolution neural network[9].

In 2017, Wang Chongren[10] proposed a customer default risk prediction method based on convolution neural network for credit risk assessment of online financial industry. In 2019, Tang Yiwei[11] tried to use the convolution neural network model, which is rarely applied in the field of green credit assessment, to study and discuss the green credit assessment of commercial banks.

4. Recurrent neural network

The development of recurrent neural network algorithm originated in the 1980s and 1990s and gradually prevailed in the 21st century. Recurrent neural networks commonly used today include bidirectional recurrent neural networks and long-term and short-term memory networks[12].

In 2020, He Xintian[13] proposed an in-depth learning comprehensive credit scoring model. The recurrent neural network (RNN) and its extended form of bidirectional recurrent neural network (BRNN) are introduced into the credit scoring domain for the first time to avoid the limitations of the model. In 2021, Xia Jiajia et al[14] established an RNN regression neural network model with memory ability to assess financial risks, based on a detailed analysis of the current financial risk situation in Anhui Province and taking into full consideration of the multidimensional and non-linear financial risks and the factors affecting the time series of sample data.

5.Spark platform

From 2003 to 2004, Google Labs published three academic papers to introduce the big data processing platforms MapReduce[15], GFS[16] and BigTable[17] it designed and used internally, which laid the foundation for the vigorous development of big data framework and later many big data platforms.

Since then, with the further development of Spark as a big data analysis and processing platform, Tang Zhenkun[18] designed and implemented the classical algorithms based on Spark platform in 2014.At present, Spark platform as a new big data computing platform has been applied in all aspects of life. Rao Jiyong et al[19] applied the Spark platform to the field of medical equipment operation and maintenance, and Xu Tao et al[20] applied the Spark platform to the field of logistics decision-making.

In the existing research work, machine learning algorithm is widely used in the field of credit risk prediction, and neural network algorithm is also gradually applied in recent years; As a new big data platform, Spark platform has been combined with traditional machine learning algorithm and applied to medical treatment, decision-making and

other fields. At present, the Spark platform has not been basically applied to the field of corporate credit risk prediction, and its application combined with in-depth learning neural network algorithm is also less.

In the selection of indicators, most of the models only use the financial indicators that the enterprise fails to take into account, and basically do not include all non-financial indicators.

The innovation of our work are as follows:

(1) Combination of financial and non-financial indicators

(2) Build the model by combining Spark platform with depth learning algorithm

(3) Include a distinction between different levels of risk

2. Research process

2.1. Data collection

According to the industry classification of the CSRC, this study selected the listed companies in three major industries (manufacturing, wholesale and retail, and transportation, warehousing and postal services), and collected all kinds of relevant financial and non-financial data through CSMAR Database, Choice Data, Dongfang Fortune, Sina Financial, Tonghua Shun and the annual reports disclosed by the companies.

The credit risk of listed companies generally arises from the borrower's failure to repay the debts in time and in full or from the default of bank loans due to various reasons. Generally speaking, this kind of risk has a very strong correlation with the profitability, solvency and cash flow of the enterprise, so the collected sample data is divided into two groups based on the performance of the listed company's operation, i.e. whether the company has been marked as ST (special treatment) or *ST (delisting warning) by the exchange. Among them, the company data marked as ST or *ST by the exchange constitute the high risk group of corporate credit default, and the companies not marked as ST by the exchange constitute the low risk group of corporate credit default.

Considering the time span of sample selection, the earliest time for collecting various financial and non-financial indicators of the company is set as 2012 based on the earliest year for which most of the company data can be collected. On the other hand, considering that the operation of each company suffered a great impact from the epidemic from 2020 to 2022, this paper only studies the credit risk of companies under normal operation, so this paper selects the data of 1704 listed companies from 2012 to 2019 as the research object.

The above factors that may affect the credit default risk of the company are divided into two categories: financial indicators and non-financial indicators, under which there are further sub-divided primary indicators and further sub-divided secondary indicators. Based on the data collection, 21 sub-indicators as shown in Table 1 were collected as the data set indicators used in this paper. In addition, year dates are included as indicators.

Table 1. Indicators of dataset collection.

	First-level indicators	Second-level indicators
Financial indicators	Debt paying ability	Liquidity ratio
		Quick ratio
		Cash ratio
		Asset-liability ratio
		Property ratio
		Interest coverage ratio
	Operating capacity	Receivable turnover ratio
		Current assets turnover
		Turnover of total assets
	Profitability	Operating margin
		Net operating profit rate
		Net profit rate of total assets

		Rate of return on common stockholders' equity
	Ability to grow	Operating income growth rate
		Net profit growth rate
		Total assets growth rate
		Price/earning ratio
		PB ratio
Non-financial indicators	Internal control	Proportion of independent directors on the board of directors
		Whether the specialized committees under the board of directors are perfect
		Records of non-compliance by corporate entities

2.2. Experimental environment

Windows S10 host, VMware-workstation15.5.0 virtual machine, Ubuntu16.04 virtual operating system, Python3.6.8, jdk-8u201-linux-x64, pseudo-distributed installation of Hadoop3.1.3, Spark2.4.0, Scala2.12.8, jupyter-notebook6.0.0.

2.3. Data processing

1. Standardized processing

After reading the data, firstly, the erroneous data and blank data in the data set are eliminated. Because there are still some data that can't be trained, in order to solve this problem, consider the year and convert it into float data.

2. Factor standardization

Through the analysis of the selected indicators that may affect the credit default risk of the company, it can be seen that there is the problem of dimensional inconsistency among each indicator. Standardization of indicators can eliminate the impact of different dimensions. In this study, the Z-Score method is selected to standardize the data. Based on the original mean and standard deviation of each index, each index can be converted into a variable that approximately obeys the standard normal distribution.

3. Select the correlation index for real application based on the correlation

From the preliminary correlation analysis, it can be seen that not all the collected indicators have strong correlation with the corporate credit risk that needs to be predicted. If all indicators are taken into account when establishing the model, the prediction result of the model may be inaccurate and the effect may not be ideal. In this paper, the Pearson coefficient and Kendall coefficient of each index are further screened to further streamline the relevant indicators.

Finally, combining the results of the two correlation coefficients, taking 1/3 of the average coefficient as the standard, the indicator with higher correlation coefficient is deleted, and the remaining 13 indicators are stock code, year date, default record, net profit margin on total assets, return on net assets, operating gross profit margin, operating net interest rate, total assets turnover rate, net profit growth rate, current ratio, quick ratio, property right ratio and market-to-book ratio.

4. Balancing treatment

The data set selected in this paper is unbalanced data because the amount of company data marked with ST and *ST in listed companies is much smaller than the amount of company data not marked with ST. In the two-class classification problem, when the data is unbalanced, the accuracy of the model is often high, and the model has no reference value. Therefore, it is necessary to adopt methods to balance the data from a data perspective.

Based on the advantages and disadvantages of various data imbalance processing methods and the characteristics of the data selected in this paper, the over-sampling method is finally selected to balance the data, i.e. random sampling is performed from a small number of samples so as to add new samples. After the balancing process, the original 4,293 pieces of data were expanded to 8,100 pieces.

2.4. Model building

2.4.1. Random forests

Compared with the traditional random forest algorithm, the random forest algorithm based on Spark platform is optimized to a certain extent, thus improving the operation efficiency. In the implementation of Spark, because the data is stored in a distributed manner, frequent access to the data will reduce the efficiency of the operation. Therefore, the Spark platform uses a breadth-of-view approach, which builds one level of all trees at a time after each data read.

The data set is randomly divided according to the proportion of 70% training set and 30% test set. The trainClassifier method in RandomForest Forest module in Spark MLlib is selected to train and model the divided training set, and the input parameters are adjusted by grid search method. Here, the random seed (random_state), proportion parameter (proportion) and number of decision trees (numTrees) are adjusted respectively, and the change of precision of the model is shown in fig. 1. We can find that: the model has the best prediction effect when the optimal parameter combination is random seed 7, the proportional parameter is 0.8 and the number of decision trees is 21. The prediction accuracy rate of the model is 0.85, the recall rate is 0.85, and the F1 value is 0.85.

2.4.2. Logical regression

Similarly, the data sets are randomly divided into 70% training sets and 30% test sets. Spark Mllib encapsulates LogisticRegression classifier, loads training model, and also uses grid search method to adjust parameters. Here, the split parameters of random seed (random_state) and training set and test set are adjusted respectively, and the change of precision of the model is shown in fig. 2. We can find that: the model has the best prediction effect when the random seed parameter is 9 and the proportion is 40% training set and 60% test set. The prediction accuracy rate of the model is 0.61, the recall rate is 0.61, and the F1 value is 0.60.

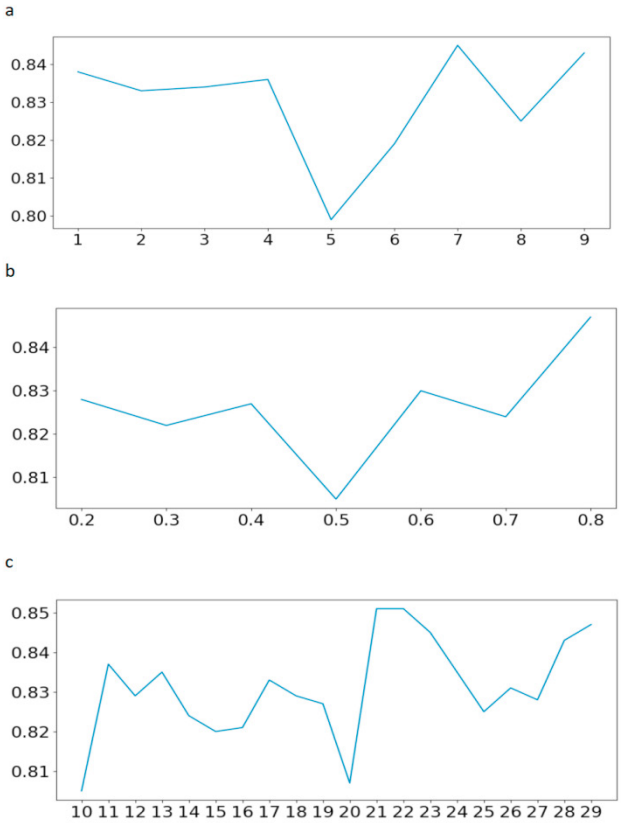

Fig. 1. (a) The Relationship between Random Seed Parameters and RF Model Accuracy Rate; (b) The Relationship between Proportional Parameters and RF Model Accuracy Rate;(c) The relationship between the number of decision trees and the accuracy rate of RF model.

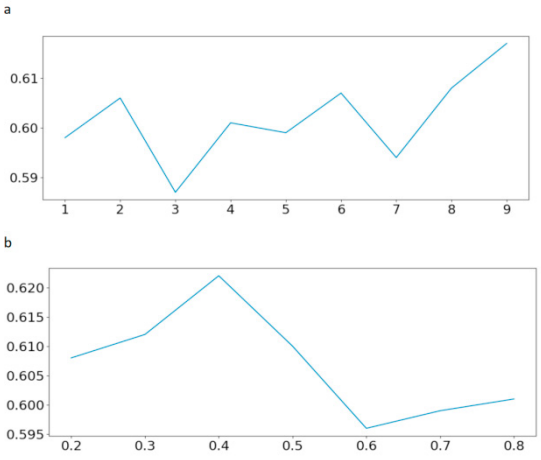

Fig. 2. (a) The Relationship between Random Seed Parameters and the Change of Accuracy Rate of Logistic Regression Model; (b) The Relationship between Partition Proportion and the Change of Accuracy Rate of Logistic Regression Model.

2.4.3. Convolution neural network

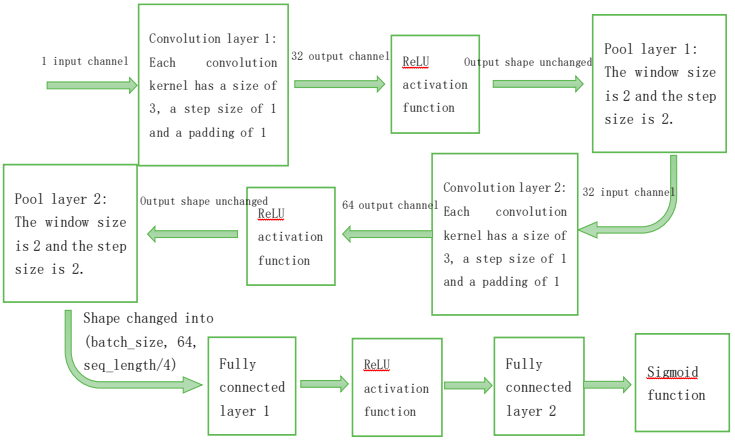

Fig. 3. Hierarchical diagram of convolutional neural network model

As shown in fig. 3, a multi-layer convolution neural network model is established. We can find that: when the learning_rate is 0.00001 and the training times (epochs) is 10000, the model prediction accuracy rate is 0.67, the recall rate is 0.41 and the F1 value is 0.54.

2.4.4. Recurrent neural network

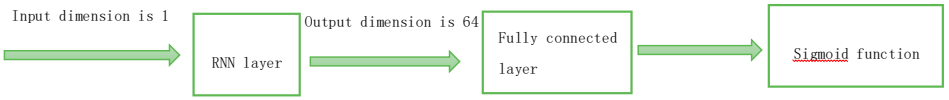

Fig. 4. Hierarchical diagram of recurrent neural network model

As shown in fig. 4, a three-layer recurrent neural network model is established. We can find that: when the learning_rate is 0.0001 and the training times (epochs) is 10,000, the model prediction accuracy rate is 0.93, the recall rate is 0.96 and the F1 value is 0.93.

3. Model comparison and result analysis

3.1. Model evaluation and comparison

The empirical results are shown in Table 2 and Figure 5, which show that the prediction accuracy, recall rate and F1 value of the recurrent neural network model built based on the pytorch framework of Spark platform in this study

are higher than those of the other three models. This model is effective in predicting corporate credit default risk and provides a new idea for how to identify corporate default risk. Each model has the best evaluation index under the parameter combination.

Table 2. Comparison of model results.

	ACC	REC	F1-Score
RF	0.85	0.85	0.85
logistic	0.61	0.61	0.60
CNN	0.67	0.41	0.54
RNN	0.93	0.96	0.93

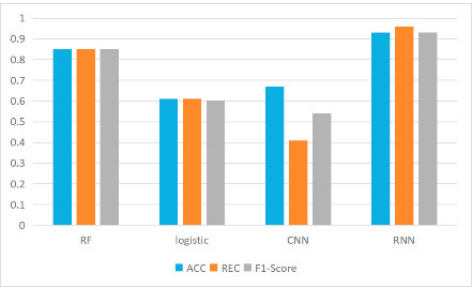

Fig. 5. Hierarchical diagram of recurrent neural network model

3.2. Robustness analysis

For the optimal model, we performed robustness analysis on the circular neural network model based on pytorch framework of Spark platform. In this study, the data set used in the model is changed by modifying the number of indicators, changing the number of samples and eliminating non-financial factors, so as to analyze and evaluate the robustness of the model. Among them, for the way to modify the number of indicators, we randomly deleted an indicator from the data set to re-train and evaluate the model; For the way of changing the number of samples, this study randomly deleted 50 pieces of data in the data set to re-train and evaluate the model; For the method of excluding non-financial factors, this study deleted the "corporate non-compliance record" indicator in the data set to re-train and evaluate the model. The results of evaluation model indicators obtained by these three adjustment methods are shown in Table 3 and Figure 6.

Table 3. Robustness analysis of RNN model.

Model	ACC	REC	F1-Score
Original model	0.93	0.96	0.93
Modify the number of indicators	0.97	0.99	0.97
Change the number of samples	0.95	0.99	0.95
Eliminate non-financial factors	0.94	0.98	0.94

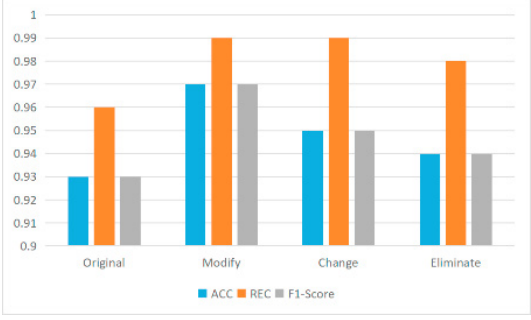

Fig. 6. Robustness analysis of RNN model

4. Summary

In this paper, the big data analysis algorithm is combined to predict the company's credit risk default, and the emerging data technology is integrated into the traditional enterprise risk management and control mechanism. We try to introduce non-financial indicators, use the Spark big data platform based on pytorch framework and deep learning algorithm, and use it in the field of corporate credit default, and compare it with other traditional machine learning algorithm prediction models, and finally get better prediction results. Our proposed model is expected to serve many levels in the future, providing reference for the enterprise itself and many parties with trading relations with the enterprise.

Acknowledgements

This research is supported by the NSFC(61100112,61309030), Key projects of National Natural Science Foundation of China NSFC(71932008), Program for Innovation Research in CUFE, Education and Teaching Reform Fund of Central University of Finance and Economics in 2022(2022ZXJG36), Emerging Interdisciplinary Project of CUFE, Discipline Construction Foundation of Central University of Finance and Economics, CUFE Young Teacher Foundation(QJJ1706).

References

[1]Ho, Tin Kam. “Random Decision Forests.” In Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1, 278. ICDAR ’95. Washington, DC, USA: IEEE Computer Society.
[2] Fang Kuangnan, Wu Jianbin, Zhu Jianping. Research on credit risk of credit card under asymmetric credit information [J]. Economic Research, 2010,45(S1):97-107.
[3] Li Xiangpei. Research on corporate credit risk based on the time series changes of asset market value [D]. Shanghai Jiaotong University, 2017.
[4] Zhu Yinong. Credit bond default risk measurement based on KMV- stochastic forest model [D]. Shandong University, 2019.
[5]Francis Galton,1886.Regression towards mediocrity in hereditary stature [J].Nature,1889,15:246-263.
Tan Yanni. Application of data mining in credit risk assessment [D]. Changsha University of Science and Technology, 2013.
[7] Li Siqi. Study on the default prediction model of China's credit bonds [D]. Southwestern University of Finance and Economics, 2020. DOI: 10.27412/D.CNKI.GXNCU.20000.000000000607
Yuan Jiang. Research on the credit risk prediction of corporate bonds in China based on data mining [D]. Southwestern University of Finance and Economics, 2020. DOI: 10.27412/d.cnki.gxncu.20000.000000000606
[9]LeCun, Y. and Bengio, Y., 1995. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), 1995.
[10] Wang zhongren, Han Dongmei. research on credit risk prediction of internet finance based on convolutional neural network [J]. microcomputer and its application, 2017,36 (24): 44-46+50. doi: 10.19358/j.issn.1674-7720.2017.20017181826
[11] Tang Yiwei. Research on Green Credit Evaluation Based on Convolutional Neural Network [D]. Shanghai Normal University, 2019.
[12]Ng, A., Kian, K. and Younes, B., Sequence Models, Deep learning.Coursera and deeplearning.ai.2018
[13] He Xin-hui. Research on integrated model of credit score combined with deep learning optimization algorithm [D]. Northwestern University, 2020. DOI: 10.27405/d.cnki.gxbdu.20000.00000000000005
[14] Xia Jiajia, Jiang Tao. Research on financial risk early warning based on RNN recurrent neural network-taking Anhui Province as an example [J]. Journal of hubei university of arts and science, 2021,42(11):26-32.
[15]DEAN J,GHEMAWAT S.MapReduce:Simplified data processing on large clusters[J].Communications of the ACM,ACM,2008,51(1):107-113
[16]GHEMAWAT S,GOBIOFF H,LEUNG S-T.The Goole file system[C].ACM SIGOPS Operating systems review,2003,37(5):29-43
[17]CHANG F,DEAN J,GHEMAWAT S,et al.Bigtable:A distributed storage system for structured data[J].ACM Transactions on Computer Systems(TOCS),ACM,2008,26(2):4
[18] Tang Zhenkun. Design and implementation of machine learning platform based on Spark [D]. Xiamen University, 2014.
[19] Rao Jiyong, Li Cong, Qian Xuezhong. Research on information mining of medical equipment operation and maintenance based on Spark [J]. Computer and Digital Engineering, 2021,49(03):525-529.
[20]Xu Tao, Sun Jian, Liu Chenwei. Adaptive ant colony algorithm based on Spark to solve CVRP problem [J/OL]. ZTE Technology: 1-7[2023-01-04].