Prediction of Financial Crisis Based on Machine Learning

Huang Junyu
Tsinghua University, China
thuhjy@163.COM

ABSTRACT

The financial crisis is an inevitable catastrophic event in the operation of the entire capital market. And it may cause significant losses to the entire market. So as individuals, if they can predict and respond in advance, it will reduce a lot of losses and make the company's life cycle longer. Here we used the data on credit defaults with a total sample of 1,000 samples containing Germany's credit default records and some basic personal information. Logistic regression, random forest and Xgboost were applied to discover useful information behind these data. The results showed that the machine learning method fitted the data relatively well, and the accuracy of Xgboost has reached about 80%. Existing checking account and foreign worker were two most important indicators to help predict financial crisis. In this way, both companies and the country could reduce their losses, so that they can spend the time of the financial crisis more smoothly and promote social prosperity.

CCS CONCEPTS

• Social and professional topics; **• Professional topics**; **• Computing and business**; **• Economic impact**;

KEYWORDS

Financial Crisis, Logistic regression, Random forest, Xgboost

ACM Reference Format:
Huang Junyu. 2020. Prediction of Financial Crisis Based on Machine Learning. In *2020 The 4th International Conference on Business and Information Management (ICBIM 2020), August 03–05, 2020, Rome, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3418653.3418674

1 INTRODUCTION

The financial crisis refers to a significant deterioration in indicators such as interest rates, exchange rates, asset prices, and solvency. During the process of the catastrophic crisis, a large number of companies go bankrupt and stock prices fall sharply, which is a huge blow and loss to every participant in the financial market [1]. Taking the example of the US subprime mortgage crisis, we can see that the sudden burst of the real estate bubble caused a chain reaction, the Hong Kong Hang Seng Index plunged, a large number of funds closed, and many banks had a large number of shortfalls in the crisis, which caused an extremely large impact. Therefore, how to judge in advance before major crisis events needs to be considered and studied by each risk control department. There is a panic index in the market today that can be used for trading. It is an effective signal of the crisis, but far from enough, because there are massive and redundant data that interferes with our judgment. Machine learning is an excellent solution to one of the problems. Its powerful inductive classification ability can help us find the rules and help us judge the coming of the financial crisis in advance.

It was found that in the work of predecessors, many researchers used machine learning methods to study the time point of the financial crisis, and got many very good results. Wang et al. used text mining and BP neural network technology based on particle swarm optimization to mine investors' attention including emotions and social interactions, and constructed a stock market crisis early warning model. This model had an early warning accuracy rate of up to 80% in the 7th crisis month, a 100% warning accuracy rate in the non-crisis month, and a 96.3% warning accuracy rate for the overall sample [2]. The popular and most widely-used application of machine learning algorithms were Bayes, support vector machines and neural networks. News media reports allows us to understand the emotional changes of individual stocks and society as a whole. Through monitoring investors' future view, not only could institutions benefit, but regulators also prevent excessive ups and downs [3]. An important harbinger of the financial crisis was the deterioration of the credit situation, so there were also literatures to study the feasibility of using improved support vector machine to evaluate credit risk. The literature showed that if a reasonable kernel function was selected, the improved support vector machine method was closer to empirical judgment [4]. There were also articles that used artificial neural networks and genetic algorithms to analyze the economic data of South Korea and propose a better financial risk warning system than the previous model [5]. A team from Portugal used the improved support vector machine method to improve the model for predicting default risk. Their results showed that this method was better than the prediction results of the general and multi-task learning support vector machine algorithm [6]. Limitation existed while using the machine learning to deal with the data. For example, a large amount of data were required in the training model. But just few financial crisis occurred after the creation of globalized world market. Moreover, there is no theoretical explanation for the results of machine learning. It is only a phenomenon rather than a cause and effect relationship [7].

Logistic regression, random forest (RF) and Xgboost are the main methods used in this paper to analyze the financial crisis data and find the most influential factors. In the following section, the source and analysis of the data will be introduced.

2 DATA RESEARCH

In this experiment, the data comes from 1,000 samples collected by Professor Dr. Hans Hofmann of the University of Hamburg. The

ICBIM 2020, August 03–05, 2020, Rome, Italy

ACM ISBN 978-1-4503-8797-2/20/08…$15.00
https://doi.org/10.1145/3418653.3418674

Table 1: Statistics of the data set

	mean	std	min	25%	50%	75%	max
existing checking account	1.58	1.26	0	0	1	3	3
Duration in month	20.92	12.06	4	12	18	24	72
Credit history	2.54	1.08	0	2	2	4	4
Purpose	3.28	2.74	0	1	3	4	9
Credit amount	3273.36	2823.37	250	1368.5	2320	3972.5	18424
Savings account/bonds	1.10	1.58	0	0	0	2	4
Present employment since	2.38	1.21	0	2	2	4	4
Installment rate	2.97	1.12	1	2	3	4	4
Personal status and sex	1.68	0.71	0	1	2	2	3
Other debtors	0.15	0.48	0	0	0	0	2
Present residence since	2.84	1.10	1	2	3	4	4
Property	1.36	1.05	0	0	1	2	3
Age in years	35.51	11.34	19	27	33	42	75
Other installment plans	1.67	0.71	0	2	2	2	2
Housing	0.93	0.53	0	1	1	1	2
existing credits	1.41	0.58	1	1	1	2	4
Job	1.90	0.65	0	2	2	2	3
Maintenance Providers	1.16	0.36	1	1	1	1	2
Telephone	0.40	0.49	0	0	0	1	1
foreign worker	0.04	0.19	0	0	0	0	1
Repay or not	1.30	0.46	1	1	1	2	2

data set contains Germany's credit default records and some basic personal information. The professor collected the data of the interviewee with 20 dimensions. All the characteristics of the companies or individuals can be categorized into two parts: one is continuous factors while the other is discrete. Status of the account, the duration, credit history, personal status and age are all discrete factors. The other factors like the number of workers, the amount of cash in the bank can be categorized as continuous factors. These factors show diverse dimensions of a company like financial condition, credit history and resistance of the risk. This information is so complete that a concrete description of the company can be given through the information. The focus of the research is the relationship between these characteristics and the occurrence of a credit crisis, discovering various situations that will occur before the crisis, and predicting the subsequent crisis.

In the data, 30% of the interviewee come out to be dishonest. The correlation, mean, quartile, and variance were calculated for the features, and the calculation results were shown in the Figure 1 and Table 1. In Table 1, several values with relatively large absolute values were found: Credit amount and Duration in month. The occurrence of the relationship is high because the more money you borrow, the longer it will take for the borrower to repay. If the amount of the borrowed money is in a large quantity, the difficulty to pay back in a short time is obviously rather big, so it is normal for the two to have a clear positive correlation. The correlation coefficient will be closer to 1 than the other factor pairs. There is also a group called Credit History and existing credits. This relationship may be because the better the credit history, the better the bank is willing to lend. As the usual process of bank conducted before loaning, the bank will evaluate the user before borrowing.

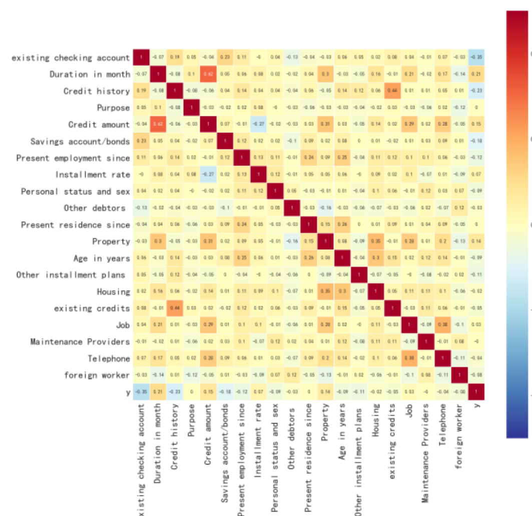

Figure 1: Correlation between different factors

The previous credit history will not borrow if the history is not good, and the good credit history in the past will have a beneficial effect on borrowing. From the Figure 1, the most important relationship is the existing checking account. The situation of dishonesty is obviously negatively related to the balance in the account. The more the balance, the smaller probability of dishonesty. As a result, a negative correlation exists here, which is logically reasonable.

3 METHODOLOGY

Machine learning has this very important application in processing large sample data. Machine learning algorithms can be divided into classification and regression. In the experiment, methods of logistic regression, random forest and Xgboost are applied, some introductions are given below.

3.1 Logistic Regression

This method is powerful and robust in classification, making it become the widely-used method in data mining and automatic diagnosis. When the problems we care about can be finally classified as a binomial distribution, logistic is often the first choice to classify the data. The pros of logistic regression is as follows: First, it has a good extensiveness. There is no high requirement for data. Any data can be classified by Logistic regression; on the other hand, it is normalized. The data obtained when doing logistic regression is between 0-1, which can be used as a probability and has a more intuitive meaning. Logistic regression is a good 0-1 classification method.

The given data becomes a binomial distribution after processing. Two parts of data consist of the whole set, one part is assigned a value of 0, and the other value 1. Then from the model, we assume that the distribution function that all data obey is:

$$P(\mathrm{x}) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Then the maximum likelihood method is used to estimate the value of the two unknown parameters. When there are more parameters, you can use the same method to estimate. Since there will be overfitting, there is a certain penalty for overfitting. Adding some additions about coefficients to the loss function, so that the overfitting situation is reduced. The formula for L2 regularization is as follow:

$$J(\beta) = -\sum_{i=1}^{N} (y_i \log(P(x)) + (1 - y_i) \log(1 - P(x))) + \lambda \sum_{i=1}^{N} \omega_i^2$$

The formula for L1 regularization is as follow:

$$J(\beta) = -\sum_{i=1}^{N} (y_i \log(P(x)) + (1 - y_i) \log(1 - P(x))) + \lambda \sum_{i=1}^{N} |\omega_i|$$

Then steepest descent method, Newton iteration or genetic algorithm can be used to calculate the minimum value of the above loss function, so that the values of the various parameters obtained are more appropriate [8] [9]

3.2 Random Forest

In our ordinary life, we will use some simple decision tree ideas. For example, naturalists will classify the creatures they see. They are first divided into very large categories such as animals, plants, and microorganisms, and then there are different genera and in the phylum of animals. Species have a more detailed classification and can grasp their characteristics. The idea of decision trees is similar. Some characteristics have been designed in the survey and shown in the data, but how to classify them? The method we generally follow is the method of maximum entropy separation. The greater the entropy, the greater the impact of this feature on the result. Information Entropy was proposed by Shannon in 1948

$$i(p) = -\sum_{j} P(\omega_j) \log_2 P(\omega_j)$$

When separating features, find the feature that maximizes the entropy and through this process, we have a decision tree,which is shown in Figure 2. Random forest classification is a combined classification model composed of many decision tree classification models. We can use the bootstrap method to sample from a large amount of data with replacement, classify and model the data extracted from each group, and then combine the prediction results of multiple decision trees to obtain by voting. Random forests are widely used in many ways, and some people have made relatively good results in economic management, credit risk assessment, bio informatics, and automatic monitoring [10].

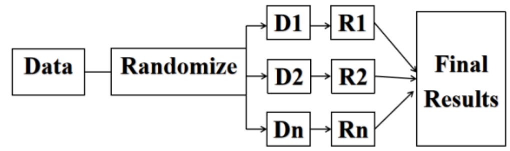

Figure 2: The steps in random forest

Since the method of bootstrap is used, the random forest method can better restore and use the data that it has, in which way better mine the information in the data. Proved by a large number of experiments, the RF algorithm has a higher prediction accuracy, resistance to outliers and noise is also very strong, which means the robustness is very excellent. Additionally the random forest over-fitting is also less than other algorithms. Overall, due to the statistical superiority, random forest is a good algorithm [10] [11]. The specific steps of the model is shown in the figure.

Our concern is to find the best parameters. Using Boostrap method, we first randomly extract data from the data set, so that the parameter sequence we get can be regarded as independent and identically distributed random variables. And then we model the decision tree. For calculation, each decision tree has a result. Then the result of each decision tree is voted, and the result with the most votes can be used as the optimal classification result [10-11].

3.3 Xgboost

Xgboost holds a difference from the previous random forest algorithm. The final decision of the random forest is the voting method. Under each decision tree, you will get a value to indicate the accuracy of the classification given by the decision tree, and then sum it to get the final score as a prediction. The formula is:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

Table 2: Evaluation of Logistic regression, random forest and Xgboost

model	accuracy	precision	recall	F1 score	roc
L1 Logistic regression	0.78	0.85	0.84	0.85	0.72
L2 Logistic regression	0.77	0.86	0.83	0.84	0.71
random forest	0.75	0.90	0.79	0.84	0.69
Xgboost	0.78	0.84	0.85	0.84	0.72

The loss function can be freely defined

$$L(\varphi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

The first term is a loss function defined by the researchers, which represents the gap between the predicted value and the actual value. The second term is the penalty term, which represents the complexity of the decision tree. After minimization, we can get the result. The process can be divided into two steps: greedy algorithm and Taylor expansion. Each time we split, we only focus on the current optimal features for splitting (greedy algorithm), and then the loss function is generally second-order derivable. After quadratic Taylor expansion, we get a quadratic function. The optimization is very simple. In this way, the optimal coefficients and splitting characteristics are obtained step by step. There are still many details in the specific implementation process, which will not be repeated here.

Xgboost mainly has the following advantages: Xgboost's loss function has regular terms, so overfitting rarely occurs. Unlike other algorithms, the Taylor expansion of Xgboost expands to the second term. It makes the result more accurate. The loss function can be freely defined, enhancing operability. At the same time, Xgboost supports column sampling and considers the sparse values that often occur in large calculations, which not only improves the accuracy, but also greatly reduces the complexity of the calculation [12].

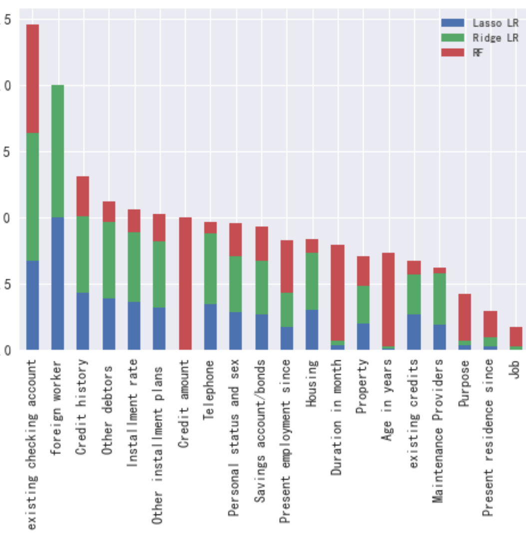

Figure 3: Different significance in different models

4 RESULTS

Many factors accounted for different proportions in different models. In two types of Logistic regression methods, 0.1 was used as the penalty coefficient. In Xgboost, the penalty parameter was also 0.1. In the random forest, 200 decision trees were established.

It can be seen from the Figure 3 that different factors have different effects on the final prediction under different methods. Existing checking account and foreign worker were two most important indicators to help predict financial crisis. The three methods have the same proportion in the factor of existing checking account, which also takes a relatively large proportion in the full model, but there is a big difference in the other. Only the random forest has a proportion to the credit amount factor, and as for foreign worker factor, only linear models have a proportion. In the four factors: Age in years, duration in month, occupation and present employment, random forest has a large proportion. Since the methods of Xgboost and random forest are essentially the same, Xgboost assigns weights to the factors that can be considered consistent with random forest. Among other factors, linear regression weighs a large proportion on them. For factors that are more important to generalized linear regression, the proportion in random forest is very low. Through the above analysis, it can be considered that there is a big gap between the model obtained by the decision tree and the model obtained by the linear regression method.

However, as the numbers in Table 2 demonstrated, we can find that the value of the F1 score describing the accuracy of the overall model is almost the same, all around 0.84. The performance of L1 Logistic regression is slightly better. As for accuracy, the performance of L1 Logistic regression and Xgboost will be better, 0.78. In precision, random forest algorithm performed better among the used methods, 0.9. In recall, the performance of Xgboost method is better. Due to the huge destructive nature of defaults, we should control the probability of the second type of error. Rather than lending, we cannot accept too many bad debts. Here is a problem, because the model gap exists, but the prediction results are similar, indicating that there is a problem of collinearity in the data.

5 CONCLUSION

This paper used the machine learning model to analyze the data of the financial crisis. A relatively new algorithm Xgboost was adopted to fit. The conclusion showed that the fitting of these methods was relatively good, and the accuracy was very high, which provided a new method of financial crisis prediction. Several improvements should made in many directions. There are not enough features in the data. Moreover, the financial crisis is a macro concept, so it is not reasonable to limit the analysis to the individual scale. Text information processing including public opinion analysis can be added to get people's panic or excessive optimism. There is also a very important part of the national macroeconomic data. When bubbles exist in large numbers, financial crisis may hit the market with a much higher possibility. With the assistance of these data under individual data, the accuracy of the model prediction can be polished.

REFERENCES

[1] Rudd K. The global financial crisis[J]. Monthly, The, 2009 (Feb 2009): 20.
[2] Elagamy M N, Stanier C, Sharp B. Stock market random forest-text mining system mining critical indicators of stock market movements[C]//2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP). IEEE, 2018: 1-8.
[3] Schumaker R P, Zhang Y, Huang C N, *et al.* Evaluating sentiment in financial news articles[J]. Decision Support Systems, 2012, 53(3): 458-464.
[4] Wang Y, Wang S, Lai K K. A new fuzzy support vector machine to evaluate credit risk[J]. IEEE Transactions on Fuzzy Systems, 2005, 13(6): 820-831.
[5] Oh K J, Kim T Y, Kim C. An early warning system for detection of financial crisis using financial market volatility[J]. Expert Systems, 2006, 23(2): 83-98.
[6] Ribeiro B, Silva C, Chen N, *et al.* Enhanced default risk models with SVM+[J]. Expert Systems with Applications, 2012, 39(11): 10140-10152.
[7] Mohseni S, Zarei N, Ragan E D. A survey of evaluation methods and measures for interpretable machine learning[J]. arXiv preprint arXiv:1811.11839, 2018.
[8] Yuan Z, Wang Z, Xu H. Credit Risk Assessment of Peer-to-Peer Lending Borrower Utilizing BP Neural Network[C]//International Conference on Emerging Internetworking, Data & Web Technologies. Springer, Cham, 2017: 22-33.
[9] Trofimov I, Kornetova A, Topinskiy V. Using boosted trees for click-through rate prediction for sponsored search[C]//Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy. 2012: 1-6.
[10] Fang K N, Wu J B, Zhu J P, *et al.* A review of technologies on random forests[C]//Statistics & Information Forum. 2011, 26(3): 32-38.
[11] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.
[12] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.