Financial Credit Default Forecast Based on Big Data Analysis

Huihui Jin[1, *], Longyin Luo[1], Xinyi Wang[2], Xiaoqi Zhu[3], Lian Qian[4], Zhice Zhang[4]

[1]Wenzhou-Kean University, Wenzhou, Zhejiang Province, China
[2]Sichuan University, Chengdu, Sichuan, China
[3]Australian National University, ACT, Australia
[4]The Affiliated High School to Hangzhou Normal University, Hangzhou, Zhejiang, China
**Corresponding author: 1098784@wku.edu.cn*
These authors contributed equally to this work

Abstract: *How to effectively evaluate and identify the potential default risk of borrowers and calculate the default probability of borrowers before issuing loans is the basis and important link of the credit risk management of modern financial institutions. This paper mainly studies the statistical analysis of historical loan data of banks and other financial institutions with the help of the idea of non-balanced data classification, and uses machine learning algorithms (not statistical algorithms) such as random forest, logical regression and decision tree to establish loan default prediction model. The experimental results show that neural network and random forest algorithm outperform decision tree and logistic regression classification algorithm in prediction performance. In addition, by using the random forest algorithm to rank the importance of features, the features that have a greater impact on the final default can be obtained, so as to make a more effective judgment on the loan risk in the financial field.*

Keywords: *Random Forest, Bank Credit, Loan Default Prediction, Data Mining*

1. Introduction

With the vigorous development of the world economy and the gradual deepening of China’s reform and opening up, whether it is the development needs of enterprises or the change of people’s consumption concept, loans have become an important form for enterprises and individuals to solve economic problems. With the introduction of various bank loan businesses and the increasing demand of people, the probability of non-performing loans, that is, loan default, also increases sharply. In order to avoid loan default, banks and other financial institutions will evaluate or score the credit risk of borrowers when granting loans, predict the probability of loan default, and make a judgment on whether to grant loans or not based on the analysis results. How to effectively evaluate and identify the potential default risk of borrowers before issuing loans is the basis and important link of credit risk management of financial institutions. Using a set of scientific machine learning algorithm model and systematicness to determine the risk of loan default can minimize the risk and maximize the profit.

This paper mainly studies how to analyze the historical loan data of banks and other financial institutions with the idea of non-balanced data classification, and predict the possibility of loan default based on random forest classification model and logistic regression classification model. The first section of this paper mainly introduces the non-balanced data classification and random forest algorithm. The second section mainly carries out data preprocessing and data analysis. The third section mainly constructs the random forest classification model to predict loan default, and obtains the AUC value of the evaluation result of the model. By comparing the random forest algorithm with the decision tree model and the logistic regression algorithm model, it comes to the conclusion that the random forest algorithm is better. Finally, through the evaluation of the importance of each feature, the conclusion about which features have a greater impact on the final result of default is drawn. The fourth section summarizes the full text.

2. Random Forest Algorithm

2.1 Unbalanced data classification

Unbalanced data, which means that the data of the majority class is far more than the data of the minority class, is prevalent in many fields, such as network instruction detection, financial fraud transaction, and text classification and so forth. And in most cases, we are only more interested in the classification of the data of the minority class. The classification problem of unbalanced data can be solved by the penalty weight of positive and negative samples. The idea is that different weights are given to the categories of different sample numbers in the classification, and then calculation and modeling are carried out in the process of algorithm implementation. Generally, the category weight of small sample size is high, while the category of large sample size is low.

2.2 Introduction of Random Forest

Random forest algorithm is a combination learning algorithm based on decision tree, which builds a forest in a random way. The basic idea of random forest algorithm is that in the process of constructing a single tree, some variables or features are randomly selected to participate in the node division of the tree, which is repeated many times to ensure the independence of the established trees. After the random forest is obtained, every decision tree in the forest will make a judgment on the sample when a new input sample enters, and get the result of which category the sample belongs to. Finally, the sample will be predicted according to which category in the whole forest gets the highest votes.

2.3 Principle and characteristics of random forest algorithm

Random forest algorithm includes classification and regression problems, and its algorithm steps are as follows:

It can be seen from the above algorithm process that the randomness of random forest is mainly reflected in two aspects: the randomness of data space is realized by Bagging (Bootstrap Aggregating), and the randomness of feature space is realized by random subsample. For the classification problem, each decision tree in the random forest makes classification prediction for new samples, and then aggregates the decision results of these trees in some way to give the final classification results of samples. Therefore, there are following characteristics of random forest:

a) The introduction of the randomness of rows (data recording) and columns (variate) in the training data set effectively reduces the probability of the random forest falling into overfitting.

b) Random forest has a good ability to resist noise.

c) When there are a large number of missing values in the data set, random forest can effectively estimate and process the missing values.

d) Random forest has strong adaptability to data: it can process both discrete and continuous data, and data sets do not need to be normalized.

e) Random forest can calculate the importance of the variable order, which makes it easy to interpret the variable.

There are two ways to calculate the importance of variables in a random forest. One method is based on the average drop accuracy of OOB (Out of Bag): In the process of generating the decision tree, OOB samples are used for testing, and misclassified samples are recorded. Then, values of a column of variables in the Bootstrap samples are randomly shuffled, and the decision tree is used to predict them again, and the number of misclassified samples is recorded again. The sum of the prediction errors divided by the total number of OOB samples is the change of the error rate of the decision tree, and the average sum of the change of the error rate of all the trees in the random forest is the average decline accuracy. The other method is based on the drop of Gini at the time of splitting. During the generation of the decision tree, the random forest splits nodes according to the decrease of Gini impurity, and all the nodes in the forest that choose a variable as the split variable are summarized to get the Gini drop.

2.4 Random forest method for non-equilibrium data classification

The random forest algorithm defaults to a weight of 1 for each class, which assumes that the misclassification costs of all classes are equal. In scikit-learn, the random forest algorithm provides the argument of class_weight, which can be a list or dict value, and the weights of different categories need to be specified manually. If the parameter is balanced, then the random forest algorithm automatically adjusts the weight using the value of y so that the weights of each category are inversely proportional to the frequency of the categories in the input data.

The calculation formula is as follows: n_samples / (n_classes * np. bincount(y))

The model of "balanced subsample" is similar to the model of "balanced" in that they both use the number of samples in a sample with a fallback expression for calculation, rather than the total number of samples. Therefore, we can solve the problem of unbalanced data classification by this method.

3. Data exploration and statistic

3.1 Data set

The data samples originated from banks' credit department which provides individual unsecured loan. The purpose of bank is to mitigate credit risk occurred in the individual loan approval according to related features. The data set includes 250000 personal loans, of which 150000 is regarded as training set and 100000 is regarded as test set respectively.

The training data set has 150000 samples of individual creditors, of which 10026 creditors default in payment, accounting for 6.684%. Those individual creditors who repay the loan timely are 139974(93.316%). The extreme difference between individual creditors shows that the training data set is skewed. The training data set contains 11 variables including the dependent variable 'SeriousDlqin2yrs', categorical variable, and 10 predictive features which are related to individual creditors. The detailed variables are displayed in table 1.

Table 1: Description of variables of data set.

Variable	Attribute	Type
SeriousDlqin2yrs	Default	Categorical
RevolvingUtilizationofUnsecuredLines	The percentage of total credit amount (exclude mortgage, car loans) in credit line	Numerical
Age	Creditors' age	Numerical
NumberofTime30-59DaysPastDueNot Worse	Number of Time 30-59 Days Past Due in past 2 years	Numerical
DebtRatio		Numerical
MonthlyIncome	Creditors' monthly income	Numerical
NumberofOpenCreditLinesAndLoans	Number of Open Credit Lines and Loans	Numerical
NumberOfTimes90DaysLate	Number of times 90 days late in past 2 years	Numerical
NumberRealEstateLoansOrLines	Number of real estate loans or lines	Numerical
NumberOfTime60-89DaysPastDueNotWorse	Number of time 60-89 days past due in past 2 years	Numerical
NumberOfDependents	Number of dependents (spouse and children)	Numerical

3.2 Data statistic

The training data set is implemented in Anaconda3 and Python3 environment. The experiment is aiming at analyzing the distribution of default rate on each independent variable. The frequency distributions of each independent variable are displayed in the tables as follows.

Table 2: Frequency distribution of 'RevolvingUtilizationOfUnsecuredLines'

RevolvingUtilizationOfUnsecuredLines	Number	Percent	Default Amount	Default Rate
Below 0.25	87657	58.44%	1873	2.14%
(0.25, 0.5]	21055	14.04%	1114	5.29%
(0.5, 0.75]	13764	9.18%	1394	10.13%
(0.75, 1.0]	24203	16.14%	4408	18.21%
(1.0, 2.0]	2950	1.97%	1183	40.10%
Above 2.0	371	0.25%	54	14.56%

The default rate (40.10%) is the highest for creditors with relatively large utilization of unsecured lines (1.0, 2.0] while the default rate (2.14%) is the lowest for creditors with less than 0.25 utilization of unsecured lines (Table 2). The default rates increase as the percentage of utilization of unsecured lines increase. However, only 14.56% of creditors with above 2.0 utilization of unsecured lines choose to default (Table 2).

Table 3: Frequency distribution of 'Age'

Age	Number	Percent	Default Amount	Default Rate
(0, 25]	3028	2.02%	338	11.16%
(26, 35]	18458	12.3%	2053	11.12%
(36, 45]	29819	19.9%	2628	8.81%
(46, 55]	36690	24.5%	2786	7.59%
(56, 65]	33406	22.3%	1531	4.58%
(65, 100)	28599	19.1%	690	2.41%

According to Table 3, creditors aged less 25 have the highest default rate (11.16%), subsequently followed by creditors aged between 26 and 35 (11.12%). The default creditors accounts for 2.4% of the total creditors who are over 65 years old (Table 3). Although the number of default creditors is largest in the age interval (46, 55], the default rate is relatively small. There is a negative relationship between age and default rates.

Table 4: Frequency distribution of 'NumberOfTime30-59DaysPastDueNotWorse'

NumberOfTime30-59DaysPastDueNotWorse	Number	Percent	Default Amount	Default Rate
0	126018	84.01%	5041	4.00%
1	16033	10.69%	2409	15.03%
2	4598	3.07%	1219	26.51%
3	1754	1.17%	618	35.23%
4	747	0.50%	318	42.57%
5	342	0.23%	154	45.03%
6	140	0.09%	74	52.86%
7 and above	104	0.07%	50	48.08%

The default rates of creditors who have more than 4 times of overdue are over 40%. Only 4% of creditors with no record of overdue would default.

Table 5: Frequency distribution of 'DebtRatio'

DebtRatio	Number	Percent	Default Number	Default Rate
Below 0.25	52361	34.91%	3126	5.97%
(0.25, 0.5]	41347	27.56%	2529	6.12%
(0.5, 0.75]	15728	10.49%	1484	9.44%
(0.75, 1.0]	5427	3.62%	596	10.98%
(1.0, 2.0]	4092	2.73%	539	13.17%
Above 2.0	31045	20.70%	1752	5.64%

Creditors with high debit ratio between 1.0 and 2.0 have the highest default rate (13.17%). The default rates of creditors with debt ratio below 0.25 and above 2.0 are 5.97% and 5.64% respectively.

Table 6: Frequency distribution of 'MonthlyIncome'

MonthlyIncome	Number	Percent	Default Number	Default Rate
Below 5000	55859	37.24%	4813	8.62%
(5000, 10000]	46091	30.73%	2752	5.97%
(10000, 15000]	13035	8.69%	547	4.20%
Above 15000	5284	3.52%	245	4.64%

The default rates for creditors whose monthly income is below 5000 and above 15000 are 8.62% and 4.64% respectively. The monthly income is negatively associated with default rates.

Table 7: Frequency distribution of 'NumberOfOpenCreditLinesAndLoans'

NumberOfOpenCreditLinesAndLoans	Number	Percent	Default Number	Default Rate
Below 5	46590	31.06%	3922	8.42%
(6, 10]	60400	40.27%	3345	5.54%
(11, 15]	29184	19.46%	1804	6.18%
(16, 20]	9846	6.56%	676	6.87%
(21, 25]	2841	1.89%	191	6.72%
(26, 30]	785	0.52%	62	7.90%
Above 30	354	0.24%	26	7.34%

In the Table 7, there is no distinct difference on default rate for creditors with different open credit lines and loans. Default rates are between 6% and 8%.

Table 8: Frequency distribution of 'NumberRealEstateLoansOrLines'

NumberRealEstateLoansOrLines	Number	Percent	Default Number	Default Rate
Below 5	149207	99.47%	9884	6.6%
(6, 10]	699	0.47%	121	17.3%
(11, 15]	70	0.05%	16	22.8%
(16, 20]	14	0.009%	3	21.4%
Above 20	10	0.007%	2	20%

According to Table 3, 99.47% creditors have less than 5 loans and lines about real estate with the lowest default rate (6.6%). Creditors who have 11 to 15 loans about real estate have the largest default rate (22.8%). Only 2 creditors who have over 20 loans default, accounting for 20% of total creditors in this internal. Creditors with more real estate loans are easier to default.

Table 9: Frequency distribution of 'NumberOfDependents'

NumberOfDependents	Number	Percent	Default Amount	Default Rate
0	86902	57.93%	5095	5.86%
1	26316	17.54%	1935	7.35%
2	19522	13.01%	1584	8.11%
3	9483	6.32%	837	8.83%
4	2862	1.91%	297	10.38%
5 and more	990	0.66%	99	10.00%

Creditors with no dependents have the lowest default rate (5.86%) and creditors with 4 dependents have the highest default rate (10.38%). Creditors with more dependents are under larger financial pressure.

For above independent variables, we get tables of frequency distributions on each variable through statistical analysis. All independent variables are related to default rates except for variable 'NumberOfOpenCreditLinesAndLoans'.

3.3 Data preprocessing

The preliminary data exploration demonstrates that there are missing values in 'MonthlyIncome' and 'NumberOfDependents' (29731 and 3924 respectively). The smallest value in variable 'Age' is 0, which is an abnormal value. Besides, there are several particular values (96 and 98) in three variables including 'NumberofTime30-59DaysPastDueNot Worse', 'NumberOfTime60-89DaysPastDueNotWorse' and 'NumberOfTimes90DaysLate'. Those values might be abnormal values or behavioral codes with special meanings.

When we use pandas to read data, we reset parameter 'na_values' in the function pd.read_scv()by adopting self-defined list, which use 'NaN' to replace '0', '96' and '98' in corresponding variables. Then we use imputer function in library sklearn to replace all 'NaN' with means of corresponding columns.

4. Conclusion

As the introduction of various bank loan business and the increasing demand of people, the probability of non-performing loan default also increases sharply, effective evaluation and identification of the borrower's potential default risk and calculation of the borrower's default probability become the basis and important link of credit risk management of modern financial institutions. This paper mainly focuses on the common loan default problems in the financial field, using the random forest method of unbalanced data classification to establish a loan default prediction model, to find a set of scientific machine learning algorithm model and systematicness to determine the risk of loan default can minimize the risk and maximize the profit. The basic idea of random forest is that in the process of constructing a single tree, some variables or features are randomly selected to participate in the tree node partition, and repeated many times to ensure the independence of these trees. Aiming at the unbalanced data, the random forest method can automatically adjust the weight according to the distribution of target-Y variable by adjusting the parameters, which could effectively solve the classification problem of unbalanced data.

The experimental results show that the classification performance of random forest algorithm is better than that of decision tree and logistic regression model, which has important reference significance for loan default prediction in financial field. In addition, we find that the age of the borrower, debt ratio and the number of real estate and mortgage loans have a great impact on the final default during this experiment by measuring the importance of each feature. The method of measuring the importance of features has important reference significance for other feature selection problems in data mining, and benefit financial institutions effectively judge the loan risk in the financial field.

References

[1] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446-3453.

[2] Gao Jiawei, Liang Jiye. (2008). Research progress on classification of unbalanced data sets (Doctoral dissertation).

[3] Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. IEEE access, 5, 16568-16575.

[4] Lu Hongyan, & Feng Qian. (2019). Review of random forest algorithm. Journal of Hebei Academy of Sciences, 3

[5] Khashman, A. (2011), Credit Risk Evaluation Using Neural Networks: Emotional versus Conventional Models; Applied Soft Computing, 11, pp.5477-5484.

[6] Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012, July). How many trees in a random forest?. In International workshop on machine learning and data mining in pattern recognition (pp. 154-168). Springer, Berlin, Heidelberg.

[7] Python Software Foundation (2018).Python Language Reference, version3.5. http://www.python.org

[8] Shen Chu (2019). The method of non-equilibrium data classification based on the generation model and Its Application Research (master's thesis, Hebei University)

[9] Wei Zhengtao, Yang Youlong, & Bai Jing. (2018). Improvement of random forest classification algorithm based on unbalanced data. Journal of Chongqing University, 41 (4), 54-62

[10] Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. Procedia Computer Science, 162, 503-513.