ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0066

Research on Credit Risk Evaluation and Forecast Method Based on Machine Learning Model

Liuliu Zhang[1*], Luling Bu[2,a], Chen Ding[3,b], Yulin Wang[4,c], Yifu Wang[5,d], Haozhen Xie[6,e]

[1]Macau University of Science and Technology, Macau, China,

[2]Jiangnan University, Wuxi, Jiangsu, China

[3]University of Toronto, Toronto, Canada

[4]Wenzhou-Kean University, Wenzhou, Zhejiang, China

[5]National Taipei University of Technology, Taipei, Taiwan, China, 1528647407@qq.com

[6]Zhejiang University City College, Hangzhou, Zhejiang, China, alexresent@163.com

*Corresponding author: 1847795319@qq.com

[a]1678633954@qq.com, [b]chen.ding@mail.utoronto.ca, [c]1246113142@qq.com, [d]1528647407@qq.com, [e]alexresent@163.com

These authors have contributed equally to this work

Abstract

How to evaluate and identify the potential default risk of the borrower before issuing the loan and calculate the default probability of the borrower 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 using the idea of non-equilibrium data classification, and uses a random forest algorithm to establish a loan default prediction model. The experimental results show that the random forest algorithm surpasses the decision tree and the logistic regression classification algorithm in the prediction performance. In addition, by using the random forest algorithm to rank the importance of features, it is possible to obtain features that have a greater impact on the eventual default, so that it can more effectively determine the risk of lending in the financial sector.

Keywords

Random forest, loan default forecast, data mining.

1. Introduction

With the continuous innovation of Internet technology and mobile communication technology and the rapid integration with the financial industry, Internet finance has developed rapidly. Artificial intelligence is already the third wave in history, called "Industry 4.0". There are breakthrough achievements, but there are still some mysteries. To actually create a cognitive "life" -- that's a lot harder. The explosive generation of data in modern society is brought about by the rapid development of computers. There are tens of thousands of data generated every moment. How to use such huge data has become a hot topic, and machine learning is undoubtedly one of the best methods.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0066

Previous risk assessment models are mainly linear and rely heavily on individual subjective initiative and empirical knowledge. With the integration of the world economy and increasingly diversified cultures, risks are no longer limited to one country or one province. Risks that break through time and space are hidden in all walks of life, especially credit risks in the financial sector. Traditional risk assessment models can no longer meet the needs of risk management. JinYunyun (2020) proposed that digitalized and scenari-based credit demand requires high transaction frequency and high timeliness, which requires Banks to quickly make risk judgment and decisions through fast and effective data collection, quickly handle credit business, and improve the efficiency of financial services. In recent years, machine learning technology has made rapid progress and is constantly approaching the goal of artificial intelligence in our mind. Most of the artificial intelligence technologies have been applied in our actual life.

A simple overview of machine learning is to learn through a large amount of data to mine useful data from the data, that is, the computer USES the existing data instead of human rational thinking to process data, and accurately calculate the data pattern process. Rather than programming to identify data, machine learning focuses on finding patterns in the data and using those patterns to make budgets. The active learning process is to select the data first, build the model data second, verify the data second, test the data again, use the data again, and finally tune the data. So use machine learning to enhance analysis to reduce risk.

Machine learning is the statistical analysis of a large amount of Internet data. The greater the amount of data learned, the higher the accuracy of prediction. In fact, many machine learning algorithms give estimates that are themselves highly probabilistic. This high probability assessment is just right for decision making and for assessing risk. The significance of risk lies in the possible deviation of the expected outcome in a given space and time, and the size of such deviation. Machine learning algorithms can predict the probability of adverse events and ensure that risk assessment models can be applied, so the application of machine learning technology in the field of risk assessment is becoming more and more popular.

In this paper, through the idea of subsection statistics, the Max value of each column section is counted from the sample data for data analysis. The first section of this paper mainly introduces the algorithm design and describes the stochastic forest, logistic regression and decision tree algorithms, and applies these algorithms to the default risk prediction scenarios. Section 2 describes the python code process and shows the output in Section 3. The fourth section introduces the relevant work content.

2. Data Analysis

2.1 Understand the data

The data set is divided into training set and test set. The purpose is to develop an application scorecard model, predict the probability of the borrower's default within a period of time in the future, and evaluate and score the customer's credit. The training set sample data consists of 150,000 pieces, with 10 independent variables and 1 dependent variable (SeriousDlqin2yrs). The details are as follows:

Variable Name	Description	Type
SeriousDlqin2yrs	**Person experienced 90 days past due delinquency or worse**	**Y/N**
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
DebtRatio	Monthly debt payments, alimony,living costs divided by monthy gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

Figure 1 Data Dictionary

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0066

According to the variables in the dataset, these variables can be categorized to help deepen the understanding of variables. It can be roughly divided into the following dimensions:

Demographic information: borrower's age, number of family members

Credit history: 35 to 59 days overdue within 2 years, 60 to 89 days overdue within 2 years, 90 days overdue or more within 2 years

Solvency: debt ratio, monthly income, amount of loans, unsafe quota recycling, real estate loan or quota

2.2 Data reading and parsing

Import data and make descriptive statistics on the original data

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	150000.00000	75000.50000	43301.41453	1.00000	37500.75000	75000.50000	112500.25000	150000.00000
SeriousDlqin2yrs	150000.00000	0.06684	0.24975	0.00000	0.00000	0.00000	0.00000	1.00000
RevolvingUtilizationOfUnsecuredLines	150000.00000	6.04844	249.75537	0.00000	0.02987	0.15418	0.55905	50708.00000
age	150000.00000	52.29521	14.77187	0.00000	41.00000	52.00000	63.00000	109.00000
NumberOfTime30-59DaysPastDueNotWorse	150000.00000	0.42103	4.19278	0.00000	0.00000	0.00000	0.00000	98.00000
DebtRatio	150000.00000	353.00508	2037.81852	0.00000	0.17507	0.36651	0.86825	329664.00000
MonthlyIncome	120269.00000	6670.22124	14384.67422	0.00000	3400.00000	5400.00000	8249.00000	3008750.00000
NumberOfOpenCreditLinesAndLoans	150000.00000	8.45276	5.14595	0.00000	5.00000	8.00000	11.00000	58.00000
NumberOfTimes90DaysLate	150000.00000	0.26597	4.16930	0.00000	0.00000	0.00000	0.00000	98.00000
NumberRealEstateLoansOrLines	150000.00000	1.01824	1.12977	0.00000	0.00000	1.00000	2.00000	54.00000
NumberOfTime60-89DaysPastDueNotWorse	150000.00000	0.24039	4.15518	0.00000	0.00000	0.00000	0.00000	98.00000
NumberOfDependents	146076.00000	0.75722	1.11509	0.00000	0.00000	0.00000	1.00000	20.00000

Figure 2 describe table

2.3 Data cleaning

The original data information is usually miscellaneous, we need to clean out the unnecessary information, in order to mine useful information from the data

1.Clean the missing data first

First look at the missing data. There are too many missing values for Monthly Income and NumberOfDependents. We use the mean instead. Because the other items are complete, we do not need to populate them.

2.Modify the data display format

The Dataframe default float type data in Pandas is displayed using scientific counting, which makes it inconvenient to observe and uses decimal instead.

3.Selective Data Analysis

Remove variables that are not important to the analysis and save the cleaned data for subsequent analysis.

4.Using cut function to convert continuous variables into categorical variables

By classifying the data variables, you can better analyze the different types of loan risk assessment.

5.Generate frequency distribution table for each variable

Converts abstract data into a frequency distribution table for easy understanding and analysis.

Each independent variable is then categorized in order, for example:

Age	Number Count	Percent	Defaulters Count	Defaulters Per
Below25	3028	2.01867%	338	11.16248%
26-35	18458	12.30533%	2053	11.12255%
36-45	29819	19.87933%	2628	8.81317%
46-55	36690	24.46000%	2786	7.59335%
56-65	33406	22.27067%	1531	4.58301%
Above65	28599	19.06600%	690	2.41267%

Figure 3 age Frequency distribution table

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0066

It can be seen that the default rate of the people under 25 and the people aged 26-35 exceeds 10%.Default rates decrease with age.

Num of Loans	Number Counts	Percen	Defaulters Co	Defaulters
Below5	149207	99.47%	9884	6.62435%
6-10	699	0.466%	121	17.31044%
11-15	70	0.04667%	16	22.85714%
16-20	14	0.00933%	3	21.42857%
Above 20	10	0.00667%	2	20%

Figure 4 Frequency distribution table of NumberRealEstateLoansOrLines value

It can be seen that 99.47% of the borrowers have less than 5 real estate and mortgage loans, but the default rate of borrowers with more than 5 loans has increased significantly, and the default rate of borrowers with more than 10 loans is above 20%.

NumberOfTime30-59DaysPastDueNotWorse	Number Count	percent	Defaulters Counts	Defaulters Perc
0	126018	84.0120%	5041	4.0002%
1	16033	10.6887%	2409	15.0253%
2	4598	3.0653%	1219	26.5115%
3	1754	1.1693%	618	35.2338%
4	747	0.4980%	318	42.5703%
5	342	0.2280%	154	45.0292%
6	140	0.0933%	74	52.8571%
7 and above	104	0.0693%	50	48.0769%

Figure 5 Frequency distribution table of NumberOfTime30-59DaysPastDueNotWorse value

It can be seen that the default rate of borrowers who have not overdue for 30-59 days is only about 4%, but with the increase of overdue times, the default rate has increased significantly.For the other 2 variables, the frequency distribution table of overdue times of 60-89 days for borrowers and overdue times of 90 or more for borrowers also shows the same trend as figure 2.5.5.Therefore, it can be concluded that the more overdue the borrower occurs, the higher the default rate.

DebtRatio	Number Coun	Percent	Defaulters Counts	Defaulters Perc
Below0.25	52361	34.90733%	3126	5.97%
0.25-0.5	41347	27.56467%	2529	6.11653%
0.5-0.75	15728	10.48533%	1484	9.43540%
0.75-1.0	5427	3.61800%	596	10.98213%
1.0-2.0	4092	2.72800%	539	13.17204%
Above2	31045	20.69667%	1752	5.64342%

Figure 6 Frequency distributionof DebtRatio

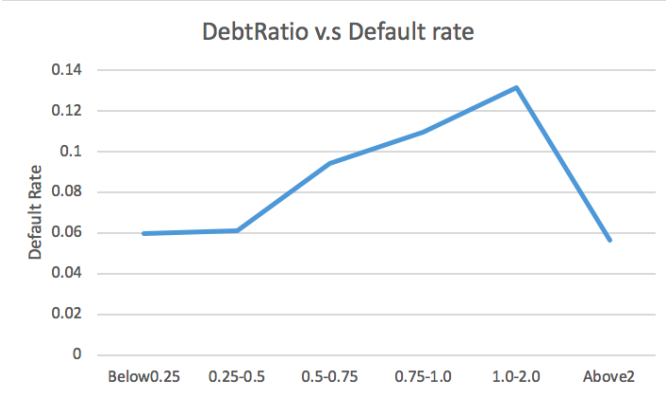

Figure 7 DebtRatio

TheDebtRatio in Figure.6 is the ratio of monthly debt payment, dependents' relief and living cost over monthly income, i.e., expense over income. As can be seen from figure 7, default rate increases as DebtRatio moving up until the ratio reaches 2, where default rate will start decrease.

NumOfDependents	Number Coun	Percent	Defaulters Counts	Defaulters Perc
0	86902	57.93467%	5095	5.86293%
1	26316	17.54400%	1935	7.35294%
2	19522	13.01467%	1584	8.11392%
3	9483	6.322%	837	8.82632%
4	2862	1.908%	297	10.37736%
5 and more	990	0.66%	99	10.00000%

Figure 8 Frequency distributionof NumberOfDependents

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0066

Figure 8 illustrates that the default rate increases with the number of dependents, reaching the rate 10% eventually. Hence, we can conclude that families tend to have a tight budget as the number of dependents increases, which then caused high default rate.

Monthly Income	Number Counts	Percen	Defaulters Co	Defaulters
Below 5000	55859	37.239%	4813	8.61634%
5000-10000	46091	30.72733%	2752	5.9708%
10000-15000	13035	8.69%	547	4.19639%
Above 15000	5284	3.52267%	245	4.63664%

Figure 9 Frequency distributionof MonthlyIncome

Figure 9 shows that the group of people whose monthly income below 5000 has the highest default rate(8.61%) compared to other groups. The default rate goes down as monthly income rise.

Column1	Number Counts	Percen	Defaulters Co	Defaulters
Below5	46590	31.06%	3922	8.41812%
6-10	60400	40.26667%	3345	5.53808%
11-15	29184	19.456%	1804	6.18147%
16-20	9846	6.564%	676	6.86573%
21-25	2841	1.894%	191	6.72298%
26-30	785	0.52333%	62	7.89809%
Above 30	354	0.236%	26	7.34463%

Figure 10 Frequency distributionof NumberOfOpenCreditLineAndLoans

From figure 10, there's no significant pattern showing that the number of open credit lines and loans has a relation with the default rate.

Therefore, among the 10 variables have been researched, 9 of them are related to default rate.

3. Related work

A close study to ours is Risk assessment in social lending via random forests (Malekipirbazari & Aksakalli, 2015). The authors researched and evaluated 4 different methods (random forest, support vector machines, logistic regression and k-NN) to predict social lending risks and concluded that random forest has the highest accuracy among the methods they tested by comparing accuracy, AUC, RMSE, TP Rate and FP Rate. Besides, an interesting finding is that they claimed random(RF) forest outperformed FICO scores and LC grades: RF classifier has a lower defaulting rate(3.1%) compared with FICO scores(8.2%). (FICO, a publicly traded corporation, produces scoring models that are most commonly used and distributed by TransUnion, Equifax, and Experian.) The dataset author considered contained are both numerical and categorical data such as loan status and loan purpose, which helped better predicting the default risk. In terms of database size, the authors researched about 68k data while we conducted 250k in our study, hence the accuracy and AUC value is higher than they did.

A slight difference for the RF model, when compared with ours, is that the authors choosea forest size of 80 with 5 split points and tree reaches a depth of 25, while our research has 100 forests with the number of splits no less than 2. Besides, the authors computed Information Gain and correlation to determine the importance of the features they researched with respect to the loan status, while our research use a different method to determine it, which we will discuss later in the article.

Ghatasheh (2014) used the Random Forest model, choosing more categorical than numerical data (13:7) to predict credit risks.Confusion matrix and AOC value are studied to evaluate the random forest model. The highlight of this article is that the author focuses on researching different RF model by tuning 3 parameters: r (the ratio of the individual tree samples to be constructed), m (the number of variables used in growing the trees) and nT (number of the model's trees). Throughout several rounds of tests the author conducted, he concluded that Bagging of Random Forest Trees using 100 Trees and 16 variables outperformed in most of the measures, and has 80% area under the ROC curve.

S.Y.Zhou（2020） also adopted the Random Forest model.In the first stage, XGBoost algorithm was mainly used to select and output the score of feature importance in the data samples, which could also

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0066

improve the interpretation of the model. In the second stage, Random Forest (RF) was used to classify the selected indicators in the first stage. In the choice of the data, they used the German credit and included a total of 1000 customers basic information, containing 700 no default record of good customers and 300 customers who have default record . Each customer's information contains 24 attribute index. First ， they used the XGBoost in python software programming calculation, characteristics of importance of all index scores, including the credit usage, credit period, the guarantor, savings accounts, credit history, which had a greater influence on the current account for the borrower default risk; Foreign workers, age and other payment methods have little influence and are basically consistent with actual experience. Then, they used RF to conduct classification tests on 1000 data samples by using the obtained 14 attribute indicators. And the ROC curve was adopted to conclude that the improved XGBoost-RF model made the model perform better than before by optimizing the data indicators, with higher classification accuracy.

4. Modeling and experimental results

4.1 random forest model

This experiment uses sklearn. ensemble. Random Forest Classifier in Python to build a random forest model.

Some parameters are set as follows:

N_ Estimators: Set the number of decision trees to 100.

oob_ Score: Whether to use out of pocket data, set to true.

min_ samples_ Split: When dividing nodes according to attributes, the minimum number of samples for each division is set to 2.

min_ samples_ Leaf: The minimum sample number of leaf nodes, set to 50.

N_ Jobs: Number of parallels, set to -1. How many jobs are started as many cores as the computer CPU has.

class_ Weight: Set to 'balanced_subsample', use the y value to automatically adjust the weight, each weight is inversely proportional to the category frequency in the input data.

Bootstrap: Whether to use bootstrap sample sampling, set to true.

4.2 Model evaluation

The model evaluation index used in this experiment is the AUC (Area under the ROC curve) value. AUC is defined as the area under the ROC (Receiver Operating Characteristic) curve. Obviously, the value of this area will not be greater than 1. The horizontal axis of the ROC curve is False Positive Rate (FPR), and the vertical axis is True Positive Rate (TPR). Since the ROC curve is generally above the line y=x, the range of AUC value is between 0.5 and 1. The AUC value is used as the evaluation criterion because in many cases the ROC curve does not clearly indicate which classifier performs better, and as a value, a classifier with a larger AUC performs better. We compared the random forest model with the logistic regression classification model and the decision tree classification model. The results are as follows.

Algorithm	AUC value
Random forest	0.86
Decision tree	0.80
logistic regression	0.80

Figure 11 AUC function values of different models

As can be seen from Figure 11, the AUC value of the random forest algorithm is higher than that of the decision tree and logistic regression algorithm, so the prediction performance of the random forest algorithm is also better than the other two algorithms.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0066

4.3 Feature importance measurement

This experiment uses thefeature importances_ of sklearn. ensemble. Random Forest Classifier. The importance of each feature is shown in the following table.

variable	feature_importance
RevolvingUtilizationOfUnsecuredLines	0.3411
NumberOfTime30-59DaysPastDueNotWorse	0.1694
NumberOfTimes90DaysLate	0.1594
NumberOfTime60-89DaysPastDueNotWorse	0.0727
age	0.0677
DebtRatio	0.0625
MonthlyIncome	0.0488
NumberOfOpenCreditLineAndLoans	0.0442
NumberRealEstateLoansOrLines	0.0223
NumberOfDependents	0.0117

Figure 12 Feature Importance Table

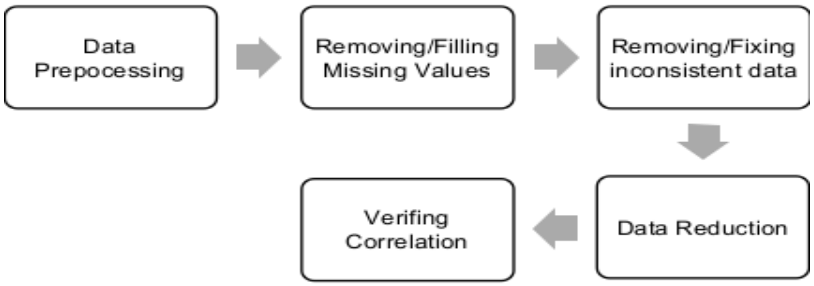

Figure 13 Raw data

As can be seen from Figure 12, the ratio of the total loan amount of the borrower to the total credit line, the number of 30-59 days overdue in the past two years and the number of over 90 days overdue in the past two years are among the top three in importance. It has a greater impact on whether the contract is ultimately defaulted, so when processing loan applications, you can pay special attention to these characteristics of the borrower.

5. Conclusion

This paper mainly studies the common problems of loan default in the financial field, and uses the random forest method of non balanced data classification to establish the model of loan default prediction. The basic idea of random forest is to randomly select some variables or characteristics to participate in the tree node division in the process of constructing a single tree, repeat for many times and ensure the independence of the established treesThrough parameter adjustment, the random forest method can automatically adjust the weight according to the y value, so as to effectively solve the problem of non balanced data classification.

The experimental results show that the classification performance of the random forest algorithm is better than that of the decision tree and the logistic regression model.In addition, by measuring the importance of each feature, in this experiment, we can obtain the 3 features of the borrower's debt ratio, the number of historical overdue times and the ratio of total loans to total credit, which have a great impact on the ultimate default. The method to measure the importance of the feature also has a more important reference significance for the feature selection problem in other data mining.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0066

References

[1] Lu Minfeng. Epidemic crisis and commercial bank digital countermeasures research observation and thinking [A]. Observation and thinking, 2020 (5): 36-43.

[2] Milad Malekipirbazari, VuralAksakalli(2015)Risk assessment in social lending via random forests.

[3] Nazeeh Ghatasheh(2014) Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study.