ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0014

Big Data Analysis Method based on Statistical Machine Learning: A Case Study of Financial Data Modeling

Yushi Miao[1,*], Tianran Fang[2,a], Da Huo[3,b], Tong Pan[4,c], Yue Yu[5,d], Yupeng Li[6,e]

[1] Feng Chia University, Taichung, Taiwan, China

[2] Beijing No.80 High School, Beijing, China

[3] Duke Kunshan University, Kunshan, Jiangsu, China

[4] Zhejiang University of Finance & Economics, Hangzhou, Zhejiang, China

[5] Sunnybrook secondary school, Wuxi, Jiangsu, China

[6] The Woodlands Secondary School, Mississauga, Canada

[a]F18510511833@163.com, [b]332872362@qq.com, [c]tp2199@foxmail.com, [d]yuyue2022@yeah.net, [e]yupeng04.yl@gmail.com

[*]Corresponding author: 937178600@qq.com

These authors contributed equally to this work

Abstract

Statistics and machine learning algorithms are combined for large numbers. According to analysis modeling is an integrated analysis method, widely used in Internet, finance. Data analysis scenarios are the focus of current analysis modeling methodology. This paper presents an analysis that fuses statistical and machine learning models, and the big data statistical analysis modeling method is applied to financial data. Analysis. We analyzed real financial lending data, and the experimental results show that in King. Application of stochastic forest algorithm in statistical machine learning in default analysis of loans and loans. If the random forest algorithm is used to rank the importance of features, it can be obtained. To the characteristics that have a greater impact on the final default, so that it can be carried out more effectively. Lending risk judgment in the financial field.

Keywords

Big Data Analysis; Statistical Machine Learning; Financial Data Modeling.

1. Introduction

With the introduction of various bank loan services and people's increasing demand, non-performing loans, that is, the probability of loan default, have also increased sharply. To avoid loan defaults, banks and other financial institutions will evaluate or score the borrower's credit risk when issuing loans to predict the probability of loan default and make decisions based on the results. How to effectively evaluate and identify the potential default risk of borrowers before issuing loans is the foundation and important link of credit risk management for financial institutions. By using a set of scientific models and systems to determine the risk of loan default, risks and maximize profits can be minimized.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0014

We analyze the historical loan data of banks and other financial institutions based on big data and statistical machine learning and predict the possibility of loan default based on the random forest classification model.

The unbalanced data in big data means that the data of one type (majority type) far exceeds the data of another type (minority type). Unbalanced data is common in many fields such as network intrusion detection, financial fraud transaction detection and text classification. In many cases, we are only interested in the classification of minority type. The classification problem of dealing with unbalanced data can be solved by the penalty weight of positive and negative samples. The idea is to assign different weights to categories with different sample sizes in the algorithm implementation process. Generally, the weight of the small sample size category is high, and the weight of the large sample size category is low, and then calculation and modeling are performed.

2. Method Design

At present, financial data is structured data. Statistical analysis and machine learning are commonly used for big data analysis and modeling. Random Forest algorithm is a type of machine learning algorithm based on statistics. It builds a forest in a random manner and is a combined learning algorithm based on decision trees. The basic idea of Random Forest is to randomly select some variables or features to participate in tree node division in the process of constructing a single tree, repeat it many times and ensure the independence between these trees. After the Random Forest is obtained, when a new input sample enters, each decision tree in the forest will judge the sample, get the result of which class the sample belongs to, and finally see which one in the entire forest belongs to has the highest votes, predict which class the sample belongs to.

Random Forest algorithm, including classification and regression problems. The algorithm steps are as follows:

Random Forest algorithm

Input:

T=training dataset;

N_{tree}=number of decision trees in the forest;

M=number of predictors in each sample;

M_{try}=number of variables participating in the division in each tree node;

$S_{samplsize}$=Bootstrap's sample size

Algorithm process:

for(i_{tree}=0;1<i_{tree}≤N_{tree};i_{tree}++)

{

1. Using T, generate a Bootstrap data sample with a size of $S_{samplsize}$.
2. Use the Bootstrap data generated above to construct an unpruned tree i_{tree}. In the process of spanning tree i_{tree}, m_{try} variables are randomly selected from M and the best variable is selected according to a certain standard (Gini value) to branch.

}

Output:

Regression problem: Take the average of all individual numbers returned as the prediction result

Classification problem: Use the classification results of most decision trees as the prediction results

Figure 1. Random forest algorithm

From the above algorithm process, the randomness of Random Forest is mainly reflected in two aspects: The randomness of the data space is determined by Bagging (Bootstrap Aggregating), the randomness of the feature space is made up of random sub-samples (Random Subspace) way to achieve. For classification problems, each decision tree in the random forest performs classification

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0014

prediction on a new sample, and then gathers the decision results of these trees in a certain way to give the final classification result of the sample.

1) The introduction of the two randomness of the row (data record) and the column (variable) in the data makes the Random Forest not easy to fall into overfitting.

2) Random Forest has good anti-noise ability.

3) When there are a large number of missing values in the data set, Random Forest can effectively estimate and process the missing values.

4) Strong adaptability to data sets: it can handle both discrete data and continuous data, and data sets do not need to be standardized.

5) It is possible to sort the importance of the variables to facilitate the interpretation of the variables.

Table1. Variable name and data type

Variable	Variable name	Variable description	Type
Y	SeriousDlqin2yrs	Breach of contract	Y/N
X1	RevolvingUtilizationOfUnsecuredLines	The total amount of credit cards and personal credit loans (excluding mortgages, instalments like car loans, etc.) divided by the sum of credit lines	Percentage
X2	age	Borrower's age	Integer
X3	NumberOfTime30-59DaysPastDueNotWorse	Number of borrowers 30-59 days overdue in the past two years	Integer
X4	DebtRatio	The number of monthly debt payments, alimony, cost of living, etc. divided by the total monthly income	Percentage
X5	MonthlyIncome	monthly income	Real number
X6	NumberOfOpenCreditLinesAndLoans	Number of open loans (Open loans, installment payments such as car loans and mortgages) and lines of credit (such as credit cards)	Integer
X7	NumberOfTimes90DaysLate	The number of times the borrower was 90 days or more overdue in the past two years	Integer
X8	NumberRealEstateLoansOrLines	Mortgage and real estate loans include the number of mortgage loans	Integer
X9	NumberOfTime60-89DaysPastDueNotWorse	The number of times the borrower is 60-89 days overdue in the past two years	Integer
X10	NumberOfDependents	The number of people (spouse, children, etc.) who need to be supported in the family, excluding myself	Integer

The model is referred to as: Y (SeriousDlqin2yrs)=Fmodel=RF(x1, x2,……, x9, x10).

There are two methods for calculating the importance of variables in random forests: One method is based on the average drop accuracy of OOB (Out of Bag). That is, in the process of growing the decision tree, first use the OOB sample to test and record the number of wrong samples, then randomly scramble the value order of a column of variables in the Bootstrap sample, re-use the decision tree to predict it, and record the number of wrong samples again. The number of two

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0014

prediction errors divided by the total number of OOB samples is the error rate change of this decision tree. The error rate changes of all trees in the random forest are aggregated and averaged to get the average decreasing accuracy rate. The other is based on the GINI drop method during splitting. The growth decision tree of the random forest is split according to the decline in GINI impurity. All the nodes in the forest that select a variable as the split variable are summarized to get GINI drop amount.

This paper designs a big data modeling method for the financial lending behavior data set. The loan default data set contains a total of 250,000 samples, of which 150,000 samples are used as the training set and 100,000 samples are used as the test set. The training set has a total of 150,000 historical data of borrowers, of which 10026 default samples, accounting for 6.684% of the total sample, the loan default rate is 6.684%, and 139,974 non-default samples, accounting for 93.316% of the total sample. This data set is a typical highly unbalanced data. The data set includes the age, income, family, etc. and loan status of the borrower. There are a total of 11 variables, among which SeriousDlqin2yrs is the label, and the other 10 variables are the predictive features.

The following table lists the variable names and data types.

3. Experimental Results

The experiment is based on Anaconda 3 and Python3. We first operated preliminary analysis on the data and then mainly focused on the distribution of the default rate on each independent variable, generating a frequency distribution table as shown in Table 2. (all decimals are rounded).

Table 2. Frequency distribution table for variable age

age	number of people	proportion	number of people defaulted	proportion of people defaulted
below 25	3028	2.02%	338	11.16%
26-35	18458	12.31%	2053	11.12%
36-45	29819	19.88%	2628	8.81%
46-55	36690	24.46%	2786	7.59%
56-65	33406	22.27%	1531	4.58%
above 65	28599	19.07%	690	2.41%

Considering the relationships of the age of the borrower and default rate (in Table 2), we could see that the default rate exceeds 10% for both those younger than 25 and those aged between 26 and 35. As age increase, the default rate decreases.

Table 3. Frequency distribution table for variable NumberRealEstateLoansOrLines

NumberRealEstateLoansOrLines	number of people	proportion	number of people defaulted	proportion of people defaulted
below 5	149207	99.471%	9884	6.62%
6-10	699	0.466%	121	17.31%
11-15	70	0.047%	16	22.86%
16-20	14	0.009%	3	21.43%
above 20	10	0.007%	2	20.00%

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0014

The findings in Table 3 suggest that 99.47% of borrowers had less than 5 real estate and mortgage loans, but the default rate increased significantly for those with more than 5 loans. Specifically, the results show that all borrowers with more than 10 loans had a default rate of 20% or more.

Table 4. Frequency distribution table for variable NumberOfTime30-59DaysPastDueNotWorse

NumberOfTime30-59DaysPastDueNotWorse	number of people	proportion	number of people defaulted	proportion of people defaulted
0	126018	84.16%	5041	4.00%
1	16032	10.71%	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%
above 7	104	0.07%	50	48.08%

As in Table 4, the default rate for borrowers with no past due for 30-59 days is only about 4%. With successive increases in the number of overdue loans, the default rate rose significantly. Two other variables, the number of borrowers with 60-89 days and with 90 days or more past due, also indicated the same trend. Therefore, it can be concluded that a higher number of overdue cases leads to a higher default rate for one borrower.

Our dataset for the experiment contains ten variables. After statistically analyzing each variable, we have obtained the frequency distribution table above. Except for Number of Open Credit Lines and Loans, the variable representing the number of open loans and lines of credit was not significantly correlated with the default rate; the rest was related to whether borrowers default or not. Our preliminary exploration of the data revealed missing values in the variables Monthly Income and Number of Dependents, with their numbers of 29731 and 3924, respectively.

Outliers were discovered in the variable age, with 0 as its minimum value. Moreover, the number 96 and 98 have been found for few times in three variable of numbers of days past due: NumberOfTime30-59DaysPastDueNotWorse, NumberOfTime30- 59DaysPastDueNotWorse, NumberOfTimes90DaysLate. The value may be exceptions or a code for certain behaviour.

Several techniques have been developed to process and analyze the data. When reading the data with pandas library of Python, we set the na_values parameter in the pd.read_csv() function to our list. The zeros in the age variable and all numbers of 96 and 98 in the three overdue variables were processed as NaN values. We then use sklearn. preprocessing. Imputer to replace all NaNs in the dataset with the average of the corresponding columns.

3.1 Integration of Big Data Modeling, Statistical Analysis, and Machine Learning

Recent advances in big data and statistical machine learning have facilitated an experimental approach to integrating statistical analysis modeling and machine learning, building a random forest model with sklearn. ensemble. RandomForest Classifier in Python.

Below are the settings of some parameters:

n_estimators: The tree numbers were set to 100;

oob_score: Whether to use out-of-bag samples to estimate the generalization score, True;

min_samples_split: The minimum number of samples required to split an internal node was 2;

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0014

min_samples_leaf: The minimum number of samples required to be at a leaf node. It was set to 50;

n_jobs: The number of jobs to run in parallel. It was set to -1, meaning to use all processors;

class_weight: was set to 'balanced_subsample', in which the y values automatically adjust weights inversely proportional to class frequencies in the input data.

bootstrap: Whether bootstrap samples are used when building trees. It was set to true.

3.2 Diagnostic Method of the Model after Training

The experiment used AUC to assess the outcome. AUC is defined as the area under the ROC (Receiver Operating Characteristic) curve. The value should not be greater than 1. The horizontal axis of the ROC curve is the False Positive Rate (FPR), and the vertical axis is the True Positive Rate (TPR). Since the ROC curve is generally above the y=x line, the value of the AUC ranges between 0.5 and 1. The AUC value is used as an evaluation criterion in cases that the ROC curve fails to indicate which classifier is more effective. A more significant value of AUC often represents better effects for the classifier.

We compared the random forest model with the logistic regression and decision tree classification models, and the results are presented in the following table.

Table 5. Comparison between Random Forest and other methods

Algorithm	AUC Value
Random Forest	0.86
Decision Tree	0.80
Logistic Regression	0.80

Table 5 shows that the AUC value of the random forest algorithm is higher than that of the decision tree and logistic regression. The predicted performance of the random forest method should outweigh the others.

Table 6. Importance of the variables

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
NumberOfOpenCreditLinesAndLoans	0.0442
NumberRealEstateLoansOrLines	0.0223
NumberOfDependents	0.0117

We carried out the importance of all characteristics with the attribute feature_importances_ in sklearn. ensemble. RandomForestClassifier. As in Table 6, the three characteristics, namely the ratio of the total loan amount to total credit amount, the number of loans past due for 30-59 days in the past two

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0014

years, and the number of loans overdue for more than 90 days in the past two years are considered the top three that have a more significant impact on whether a borrower eventually defaults on a loan. It would be wise to focus on these characteristsics of the borrower when processing loan applications.

4. Related Work

Bayesian methods have also developed rapidly in the past 20 years and become a very important machine learning method, at last, the paper also gives a brief introduction and prospect to the problem of large-scale Bayesian learning and summarizes its development trend. By the way, When the distribution of credit fraud data is extremely imbalanced, the noise errors caused by information distortion, periodic statistical error and reporting bias will disturb the training model and easily produce over-fitting phenomenon. In view of this, a deep belief neural network integration algorithm is proposed to solve the problem of extremely imbalanced credit fraud. Another related work is that the latest research progress of big data processing technology is discussed and Statistical relationship learning is a new research focus in the field of artificial intelligence. It combines relationship representation, likelihood theory and machine learning to better solve complex relational data problems in the real world. At last, proposed a new and efficient feature selection and rule extraction method for data classification.

5. Conclusion

This paper mainly studies the common problem of loan default in the financial field, and makes. The random forest method with unbalanced data classification is used to establish the prediction of loan default. The model, the basic idea of a random forest, is that in the process of constructing a single tree, random. Select some variables or features to participate in tree node division, repeat several times and ensure the establishment. The independence between these trees, for unbalanced data, through parameter adjustment. So that the random forest method can automatically adjust the weight according to the y value, so that the effective solution. Classification of non-equilibrium data. Experiments show that the random forest algorithm is better than. Decision trees and logical regression models have better classification performance for loans in the financial field. The problem of default prediction has important reference significance. In addition, through the weight of each feature. In this experiment, the age and debt ratio of the borrower can be obtained. And the number of real estate and mortgage loans affects the eventual default. Larger, this measure of feature importance also applies to other features in data mining. The selection problem has important reference significance.

References

[1] Xiaodong Wang, Qing Wang, and Ye Tao. A User Profile Analysis Framework Driven by Distributed Machine Learning for Big Data. In Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science (AICS 2019). Association for Computing Machinery, New York, NY, USA, 358-363.

[2] Jian Liu, Ke Ji, Runyuan Sun, Kun Ma, Zhenxiang Chen, and Lin Wang. 2019. Abnormal Phone Analysis Based on Learning to Rank and Ensemble Learning in Environment of Telecom Big Data. In Proceedings of the 2019 11th International Conference on Machine Learning and Computing (ICMLC '19). Association for Computing Machinery, New York, NY, USA, 301–305.

[3] Zhihan Lv, Ranran Lou, Hailin Feng, Dongliang Chen, and Haibin Lv. 2021. Novel Machine Learning for Big Data Analytics in Intelligent Support Information Management Systems. ACM Trans. Manage. Inf. Syst. 13, 1, Article 7 (March 2022), 21 pages.

[4] Toshiro Minami and Yoko Ohura. 2021. Small Data Analysis for Bigger Data Analysis. In 2021 Workshop on Algorithm and Big Data (WABD 2021). Association for Computing Machinery, New York, NY, USA, 1-8.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0014

[5] Yi-Chuan Chiu, Hsing-Hung Lin, and Yung-Tsan Jou. 2019. A Model Selection Method for Machine Learning by Differential Evolution. In Proceedings of the 2019 4th International Conference on Big Data and Computing (ICBDC 2019). Association for Computing Machinery, New York, NY, USA, 135–139.

[6] Hamdi Kavak, Jose J. Padilla, Christopher J. Lynch, and Saikou Y. Diallo. 2018. Big data, agents, and machine learning: towards a data-driven agent-based modeling approach. In Proceedings of the Annual Simulation Symposium (ANSS '18). Society for Computer Simulation International, San Diego, CA, USA, Article 12, 1–12.

[7] Zouiten Mohammed. 2017. Machine learning algorithms for oncology big data treatment. In Proceedings of the 2nd International Conference on Computing and Wireless Communication Systems (ICCWCS'17). Association for Computing Machinery, New York, NY, USA, Article 76, 1–6.

[8] Ping Wang, Yan Li, and Chandan K. Reddy. 2019. Machine Learning for Survival Analysis: A Survey. ACM Comput. Surv. 51, 6, Article 110 (February 2019), 36 pages.

[9] Yihao Li and Jin Wang. 2019. Online Updating Algorithms of Statistical Methods for Big Data. In Proceedings of the 2nd International Conference on Computing and Big Data (ICCBD 2019). Association for Computing Machinery, New York, NY, USA, 81–85.

[10] Venkamaraju Chakravaram, Vidya Sagar Rao G., Jangirala Srinivas, and Sunitha Ratnakaram. 2019. The Role of Big Data, Data Science and Data Analytics in Financial Engineering. In Proceedings of the 2019 International Conference on Big Data Engineering (BDE 2019). Association for Computing Machinery, New York, NY, USA, 44-50.