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

Jiaqi Pang
susipang2022@gmail.com

Miami College of Henan University, Zhengzhou, Henan, China

Abstract: The combination of statistics and machine learning algorithms for big data analysis modeling is an integrated analysis method, which is widely used in data analysis scenarios such as the Internet, finance, etc. It is a hotspot in the current analysis and modeling methodology research. We propose an analytical modeling framework that can integrate statistical models and machine learning models, and applies statistical analysis modeling methods of big data to financial data analysis. We analyzed real financial loan data. The experimental results show that in the application of financial loan default analysis, the random forest algorithm in statistical machine learning works very well. We propose an improved random forest algorithm to accurately and efficiently identify key variables to better judge loan default or not, thus allowing financial institutions to more accurately assess loan risk.

Keywords: Machine Learning, Financial Data Modeling, data analysis modeling

1 INTRODUCTION

How to accurately predict potential loan defaults is a common pain point and difficulty faced by banks and other financial institutions. The unbalanced data in big data means that the data of one type (majority category) in the data far exceeds the data of another type (minority category). It is used in a variety of areas, including intrusion detection, financial fraud detection and data classification. Furthermore, we are frequently only concerned with the classification of a few categories. The weight of the penalty for positive and negative samples can be used to solve the classification problem for unbalanced data. During the implementation of the algorithm, categories with different sample sizes will be assigned different weights. In general, small sample size categories are given high weights and large sample size categories are given low weights, which are then calculated and modeled.

2 METHOD DESIGN

The current financial data is structured data, and statistical analysis and machine learning are commonly used for big data analysis and modeling. The random forest is one of the most used machine learning techniques. It builds a forest in a randomized manner and is a combinatorial learning algorithm based on decision trees. Random forest refers to a classifier that uses multiple trees to train and predict samples. In machine learning domain, a random forest is a classifier that consists of multiple decision trees, and the classes it outputs are determined by the mode of the classes output by the individual trees. When a fresh input sample enters the

ICICA 2022, December 02-04, Chongqing, People's Republic of China
Copyright © 2023 EAI
DOI 10.4108/eai.2-12-2022.2327928

random forest after it has been created, each decision tree in the forest judges the sample to determine which class it belongs to. Finally, the random forest method may forecast which class the sample belongs to based on whatever class in the entire forest receives the most votes.

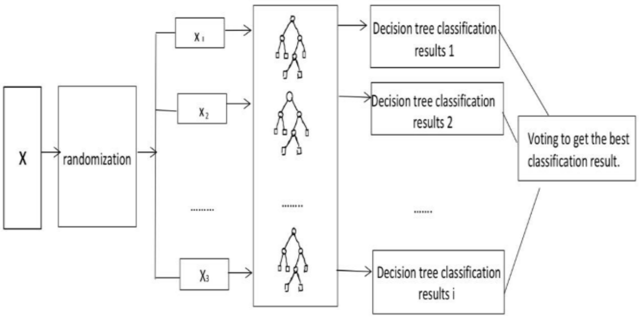
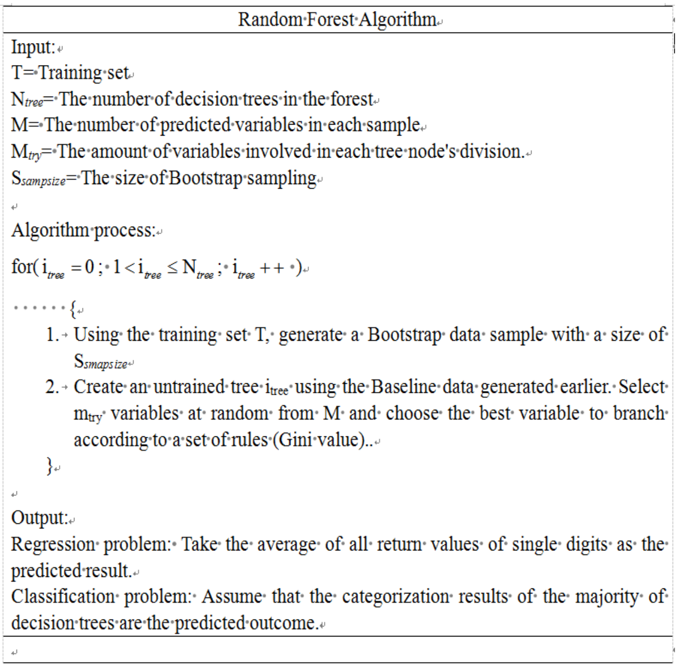

X
randomization
X_1
X_2
....
X_3
....
Decision tree classification results 1
Decision tree classification results 2
....
Decision tree classification results i
Voting to get the best classification result.

Figure 1. An illustration of the random forest algorithm

Regression and classification problems are also addressed by the random forest technique. The implementation steps are as follows:



```
Random Forest Algorithm
Input:
T= Training set
$N_{tree}$= The number of decision trees in the forest
M= The number of predicted variables in each sample
$M_{try}$= The amount of variables involved in each tree node's division.
$S_{sampsize}$= The size of Bootstrap sampling

Algorithm process:
for( $i_{tree} = 0$ ; $1 < i_{tree} \le N_{tree}$ ; $i_{tree}$ ++ )
······{
    1. Using the training set T, generate a Bootstrap data sample with a size of $S_{smapsize}$
    2. Create an untrained tree $i_{tree}$ using the Baseline data generated earlier. Select $m_{try}$ variables at random from M and choose the best variable to branch according to a set of rules (Gini value)..
    }

Output:
Regression problem: Take the average of all return values of single digits as the predicted result.
Classification problem: Assume that the categorization results of the majority of decision trees are the predicted outcome.
```

Random forest randomly selects sample data to build a model. First, use the bootstrap method to generate Ntree training sets. Bootstrap is a method of generating many available sub-data sets of the same sample size from the sample data set T by sampling with replacement, wherein the data in each sub-sample data set is repeated. Second, for each training set, construct a decision tree. All data is entered into decision trees, and each decision tree will get a result. Finally, the new data of the test set is input into the random forest composed of sub-decision trees, and the final output of the random forest is obtained by voting through the classification results of each sub-decision tree.

1.The random forest is difficult to overfit due to the insertion of random data records and variables columns.

2.The anti-noise capacity of random forest is excellent.

3. The random forest algorithm can successfully predict and handle missing values in datasets and filling correct values.

4. The random forest has good fitting ability. When building a forest, it can internally produce an unbiased estimate of the generalized error.

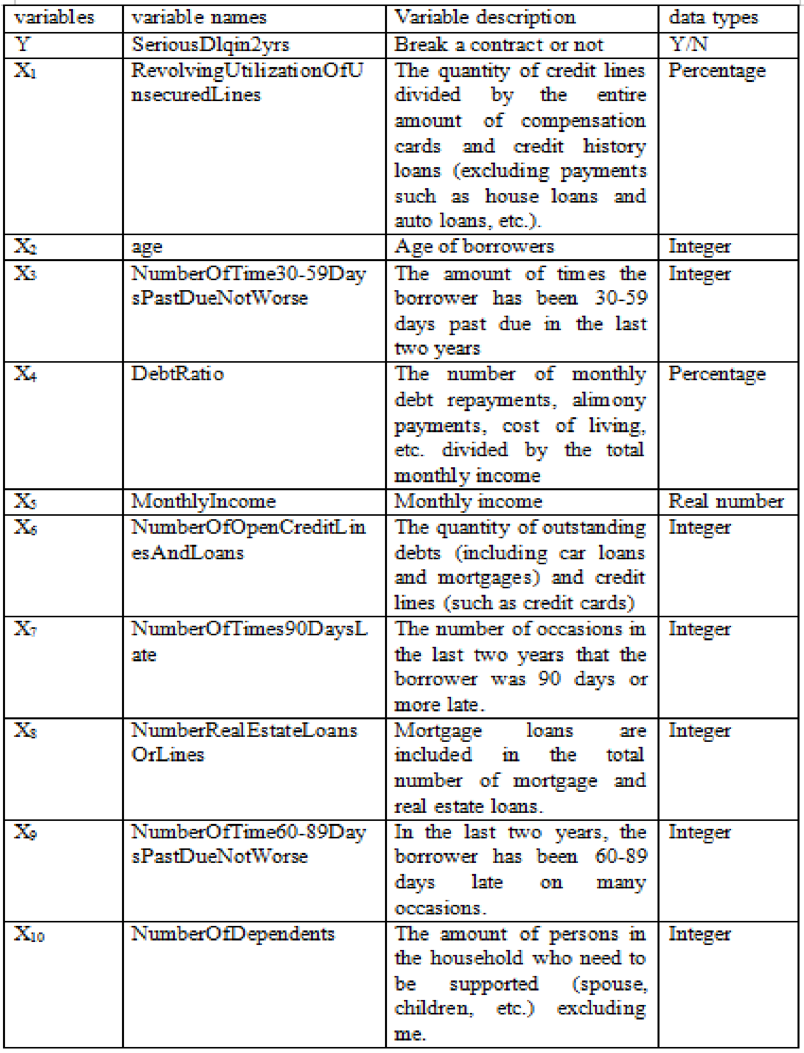
5. The random forest can compute the importance of variables to make the interpretation of variables easier. In random forests, there are two techniques to calculate the correlation of variables: The first technique uses the mean descent accuracy of Out of Bag (OOB).

The randomness of random forest is not only reflected in the random sampling of sample data, but also in the random sampling of part of the characteristics of each data. Selecting features from all features in a certain order is significant in the process of building a decision tree. The candidate feature is the set of features that have not been selected before the step. When building a decision tree, instead of splitting nodes through all the features, a part of the features is randomly extracted from the features, and the optimal solution is found among the obtained features. In this way, the samples and features are sampled at the same time, which improves the diversity of the results in the random forest system, avoids overfitting, and improves the classification performance.

In this paper, a big data modeling approach is designed for the financial lending behavior dataset. There are 250,000 samples in the loan default dataset, of which 150,000 are used as training samples and 100,000 as testing samples. Borrowers' age, income, household and loan status are included in the dataset. There are 11 variables in total, of which Y is the label, and Y indicates whether the loan defaults. {X1, X2, X3, …, X10} are feature variables. Since feature selection may have explosive combinations, feature selection by Importance may be able to replace feature combination exhaustion to a certain extent. Importance selection refers to scoring each feature and selecting features based on the score.

The names and data types of the differences are listed in the TABLE1:

Table 1. Variable names and data types



variables	variable names	Variable description	data types
Y	SeriousDlqin2yrs	Break a contract or not	Y/N
X_1	RevolvingUtilizationOfUnsecuredLines	The quantity of credit lines divided by the entire amount of compensation cards and credit history loans (excluding payments such as house loans and auto loans, etc.).	Percentage
X_2	age	Age of borrowers	Integer
X_3	NumberOfTime30-59DaysPastDueNotWorse	The amount of times the borrower has been 30-59 days past due in the last two years	Integer
X_4	DebtRatio	The number of monthly debt repayments, alimony payments, cost of living, etc. divided by the total monthly income	Percentage
X_5	MonthlyIncome	Monthly income	Real number
X_6	NumberOfOpenCreditLinesAndLoans	The quantity of outstanding debts (including car loans and mortgages) and credit lines (such as credit cards)	Integer
X_7	NumberOfTimes90DaysLate	The number of occasions in the last two years that the borrower was 90 days or more late.	Integer
X_8	NumberRealEstateLoansOrLines	Mortgage loans are included in the total number of mortgage and real estate loans.	Integer
X_9	NumberOfTime60-89DaysPastDueNotWorse	In the last two years, the borrower has been 60-89 days late on many occasions.	Integer
X_{10}	NumberOfDependents	The amount of persons in the household who need to be supported (spouse, children, etc.) excluding me.	Integer

The model is referred to as:

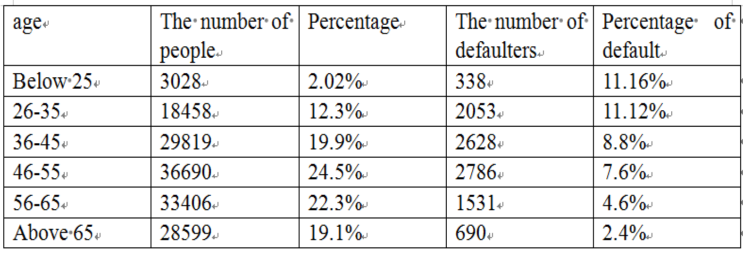
Y(SeriousDlqin2yrs)=FRFmodel{X1, X2, X3, …, X10}

3 EXPERIMENTAL RESULTS

Anaconda3+Python3 was the experimental environment used in this article. The data is first preliminarily examined. This experiment focused on analyzing the default rate distribution of various autonomous variables and generated a table of the frequency distribution, which is shown in Table 2. (These decimals are rounded).

As seen in Table 2, the default rate exceeds 10% for both those under 25 and those between 26 and 35 years of age. As age increases, the default rate decreases.

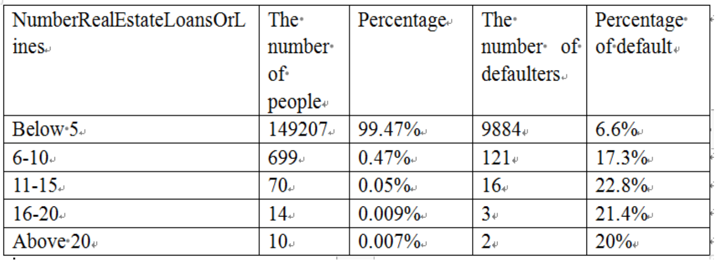
Table 2. Table of variable age frequency distribution



age	The number of people	Percentage	The number of defaulters	Percentage of default
Below 25	3028	2.02%	338	11.16%
26-35	18458	12.3%	2053	11.12%
36-45	29819	19.9%	2628	8.8%
46-55	36690	24.5%	2786	7.6%
56-65	33406	22.3%	1531	4.6%
Above 65	28599	19.1%	690	2.4%

Table 3 shows that 99.47 percent of borrowers have fewer than five real estate and mortgage loans, but the default rate for borrowers with more than five loans has risen dramatically, and default rates for borrowers with more than ten loans are all above 20%.

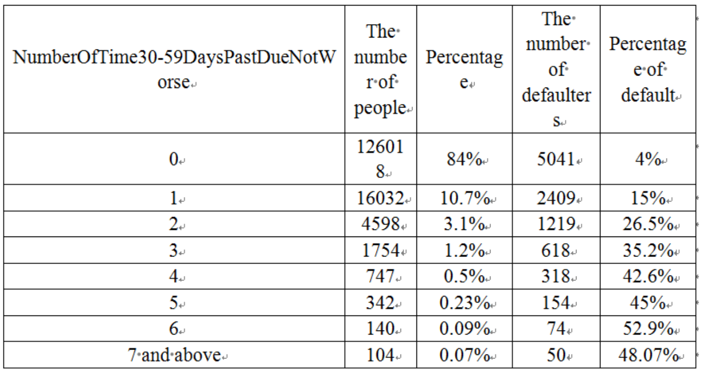
Table 3. Variable frequency distribution NumberRealEstateLoansOrLines



NumberRealEstateLoansOrLines	The number of people	Percentage	The number of defaulters	Percentage of default
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%

As shown in Table 4, the default rate for debtors that were 30-59 days past due and not overdue was only about 4%, but rose sharply as the number of days of delay increased. The frequency distributions of the other 2 factors, the number of borrowers with 60-89 days of delinquent payments and the number of borrowers with 90 days or more of overdue payments, show the same pattern as Table 4. As a result, the higher the borrower's default rate, the more times he or she has been late.

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



NumberOfTime30-59DaysPastDueNotWorse	The number of people	Percentage	The number of defaulters	Percentage of default
0	126018	84%	5041	4%
1	16032	10.7%	2409	15%
2	4598	3.1%	1219	26.5%
3	1754	1.2%	618	35.2%
4	747	0.5%	318	42.6%
5	342	0.23%	154	45%
6	140	0.09%	74	52.9%
7 and above	104	0.07%	50	48.07%

The experimental results has ten variables. Each variable is statistically analyzed by using the frequency distribution of samples. All independent variables in the experimental sample data set are related to the dependent variable Y.

The two variables X5 and X9, these two variables include missing values, and we discard variables with more than 80% missing values, that is, do not participate in the model training process. X5 and X9 have with 29730 and 3925 missing values respectively.

Outlier situation: The value of the Age field cannot be less than 0.

There are three overdue days, i.e., X3, X7, X9, which can be indicated as an outlier value.

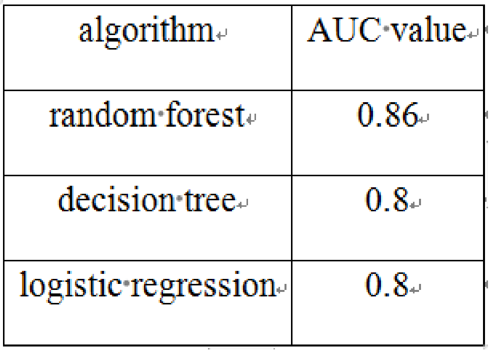
We use Python pandas library to load data, and use sklearn.preprocessing library to process NULL Values. All missing data are set to as the average value of the associated column using the Python Imputer library.

The experimental method of integrating big data statistical analysis modeling and machine learning is: Use sklearn.ensemble. To create a random forest model, use Python's Random Forest Classifier. The random forest parameters are set as follows. We set the maximum number of decision trees to 100, and the out of bag score is set to TRUE. The control the height of random forest tree, the minimum number of samples for each division of splitting nodes is set to 5, and sample number of tree leaf node is set to 50. Start as many jobs as there are computing cores of server. Use the y value to automatically update the weights when the mode is set to balanced subsample. The objective class frequency is set to balanced weight. We set bootstrap sample sampling option to True.

Model evaluation method after training: The AUC value was employed as the model assessment index. The area under the subject operating characteristic (ROC) curve is referred to as the AUC. Obviously, this area's value will be no more than one. AUC (Area Under Curve) is defined as the area enclosed by the ROC curve and the coordinate axis. Obviously, the value of this area will not be greater than 1. And because the ROC curve is generally above the line y=x, the value range of AUC is between 0.5 and 1.

In this paper, we compared the accuracy of different algorithms such as random forest, random regression, decision tree, etc. and found that random forest has the highest accuracy. The comparison result is described as Table 5.

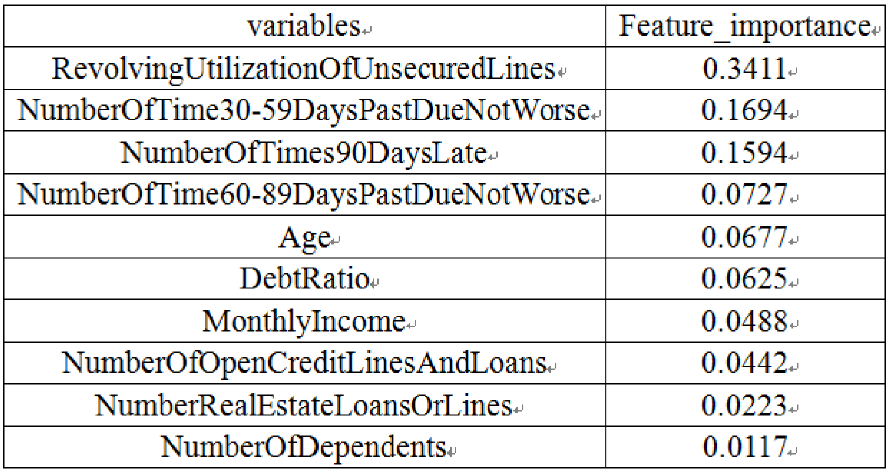
Table 5. Random forest and other algorithms are compared



algorithm	AUC value
random forest	0.86
decision tree	0.8
logistic regression	0.8

As shown in Table 6, the entire loan is associated with the borrower's overall credit limit, 30-59 days past due in the first two years, and more than 90 days past due in the first two years as the three most important factors. These three variables are the top three important characteristics in determining whether a loan defaults

Table 6. Importance varies



variables	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

4 RELATED WORK

Machine learning is a good method for financial loan prediction, which is not only efficient, adaptable, but also accurate. Sohony used a comprehensive random forest and neural network ensemble algorithm to deal with the problem of unbalanced data. Kazemi et al. proposed to use an autoencoder to extract features from transaction data samples and use a softmax network to identify sample categories. Roy et al. proved that the neural network algorithm and topology have a strong ability to deal with credit fraud problems through cloud computing. Luo used deep belief neural network algorithm to build a credit evaluation model and compared it with traditional machine learning models, and concluded that DBN has better classification capabilities. Kim et al. experimentally show that the complex neurons of the feedforward neural network algorithm are more suitable for processing high-dimensional and complex credit fraud data than the bagging ensemble algorithm.

The research on the problem of unbalanced data currently focuses on two aspects: resampling and ensemble algorithms. For resampling, SMOTE has been developed for a long time. Nguyen et al. used SMOTE random sampling in the SVM separator to significantly improve the classification effect. However, the SMOTE method also has the problem of amplifying invalid samples and reducing the classification accuracy. For the ensemble algorithm, Liu et al. proposed the easy ensemble algorithm, which uses undersampling and AdaBoost algorithm to solve the problem of most types of sample loss. Xia et al. proposed an iterative resampling model, in which iteratively optimizes the model's tilt for the majority and minority samples.

5 CONCLUSION

This paper proposes a machine learning-based financial loan default prediction method. The method shows that the effect of random forest is the best, which brings us an idea, which is to dynamically calculate whether a customer's lending behavior is based on machine learning. cause financial risk. We use a large sample data set to train the random forest model and its comparative machine learning model. The practical effect proves that the larger the sample data volume, the higher the accuracy of the model. We also found that there are 5 main factors that affect whether customers' lending behavior will have default risk, which provides a good identification tool model for loan officers in financial fields such as banks and trusts. The loan default prediction model of machine learning can quickly calculate the default probability of customers, decide whether to issue loans based on the order, and calculate the overdue rate and bad debt rate of repayment in real time.

REFERENCES

[1] Wang Qing. Automatic Classification of Library Documents Based on Random Forest. Automation Technology and Application (07),51-53+72. doi:10.20033/j.1003-7241. (2022)07-0051-04(2022).Davis, A. R., Bush, C., Harvey, J. C. and Foley, M. F., "Fresnel lenses in rear projection displays," SID Int. Symp. Digest Tech. Papers 32(1), 934-937 (2001).

[2] Gong Yangguang. Design and implementation of financial data analysis system based on data mining technology (MA Thesis, Hubei University of Technology) (2018). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201901&filename=1019806975.nh

[3] Liang Wei. Research on Anomaly Detection Methods for Financial Data (MA Thesis, Nanchang University). (2019). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202001&filename=1019210542.nh

[4] Xiao Xiangmei, Yu Jian, and Lin Zhixing. A network abnormal traffic detection method using random forest. Journal of Sanming University 39.03:84-91. (2022).doi:10.14098/j.cn35-1288/z.2022.03.0010.

[5] Zhao Zhihong & Du Ying. Research on credit fraud risk prediction model based on machine learning. Modern Electronic Technology(24),177-181.(2021). doi:10.16652/j.issn.1004-373x.2021.24.038.

[6] Ling Min. Financial data analysis method based on data mining. Small and medium-sized enterprise management and technology (mid-term issue) (11),70-72.(2021).

[7] Liao Jianxi, Lu Yong, Wang Zhenyu & Zhao Kai. Research on EEG Emotion Recognition Based on Random Forest Algorithm. Computer and Information Technology (03),1-4. (2022). doi:10.19414/j.cnki.1005-1228.2022.03.007.

[8] Wang Bin, He Binghui, Lin Na, Wang Wei & Li Tianyang. (2022). Remote sensing extraction of tea gardens based on random forest feature selection. Journal of Jilin University (Engineering Edition) (07), 1719-1732. doi:10.13229/j.cnki.jdxbgxb20210138.

[9] Chen Yang. (2022). Research on quantitative analysis method of financial data based on machine learning (MA Thesis, University of Electronic Science and Technology of China). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFDTEMP&filename=1022545837.nh

[10] Zhao Ningjie & Li Xuefei.(2020).Research on Decision Tree Classification Algorithm Based on Bagging Thought. Journal of Beijing Institute of Fashion Technology (Natural Science Edition) (03),43-48. doi:10.16454/j.cnki.issn.1001-0564.2020.03.007.

[11] Du Keke, Zhang Yue & Zhao Kai.(2022). Personal credit evaluation model combining similarity measure and random forest. Journal of Chongqing Technology and Business University (Natural Science Edition) (03),54-60. doi:10.16055/j.issn.1672-058X.2022.0003.008.