International Core Journal of Engineering Volume 8 Issue 2, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0035

Predicating Loan Risk with Machine Learning Algorithm

Tengzhou Jiang[1, *], Jing Bao[2, a], Boxuan Hou[3, b], Yihan Li[4, c]

[1] Beijing Forestry University, Beijing, China

[2] Maple Leaf International School- Shanghai, Shanghai, China

[3] ShanDong Career Development College, Jining, Shandong, China

[4] Beijing University of Technology, Beijing, China

[a]1078749273@qq.com, [b]3474070029@qq.com, [c]alice4734@sina.com

[*]Corresponding author: 2417695949@qq.com

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 credit risk management of modern financial institutions. This paper mainly studies the statistical analysis of the historical loan data of banks and other financial institutions with the help of the idea of unbalanced data classification, and uses the random forest algorithm to establish a loan default prediction model. Experimental results phenotype, neural network and random forest algorithm outperform decision tree and logistic regression classification algorithm in prediction performance. In addition, by using random forest algorithm to rank the importance of features, features that have a greater impact on the final default can be obtained, so as to make a more effective judgment of lending risks in the financial field.

Keywords

Random Forest; Bank Credit Investigation; 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, loans have become an important way for enterprises and individuals to solve economic problems, whether it is the development of enterprises or the change of people's consumption concept. With the launch of various kinds of bank loans and people's increasing demand, the probability of non-performing loans, that is, loan default, is also soaring Increase. In order to avoid loan default, banks and other financial institutions will evaluate or score the credit risk of borrowers when issuing loans, predict the probability of loan default and make a judgment on whether to issue loans according to the result. How to effectively evaluate and identify potential default risks of borrowers before issuing loans is the basis and important link of credit risk management of financial institutions. Using a set of scientific models and systems to determine the risk of loan default can minimize risks and maximize profits.

This paper mainly studies how to use the idea of unbalanced data classification to analyze the historical loan data of banks and other financial institutions and predict the possibility of loan default based on random forest classification model. The first section of this paper mainly introduces unbalanced data classification and random forest algorithm. The second section mainly carries on the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0035

data preprocessing and the 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 calculating the random forest compared with decision tree and logistic regression model, random forest algorithm is better. Finally, through the evaluation of the importance of each feature, which features have a greater impact on the final result of default. The fourth section summarizes the full text.

2. Random Forest Algorithm

2.1 Classification on Imbalanced Data

Imbalanced data refers to data from one class (majority) far exceeds data from another (minority),and is common in many areas such as network-based intrusion detection, financial transaction fraud detection, text classification. The classification problem of dealing with imbalanced data can be calculated and modeled by assigning different weights to the categories of different sample sizes in the classification.

2.2 Introduction to Random Forest

Random Forest Algorithm refers to an ensemble learning algorithm based on decision trees and is to build a forest in a random way. The basic idea of random forest is to randomly select some variables or characteristics 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 getting the random forest, when a new input sample enters, each decision tree in the forest will judge the sample, getting the result of which class the sample belongs to. Then, predict the sample should belong to the class which has the highest number of votes in the entire forest. The process is shown in Figure 1.

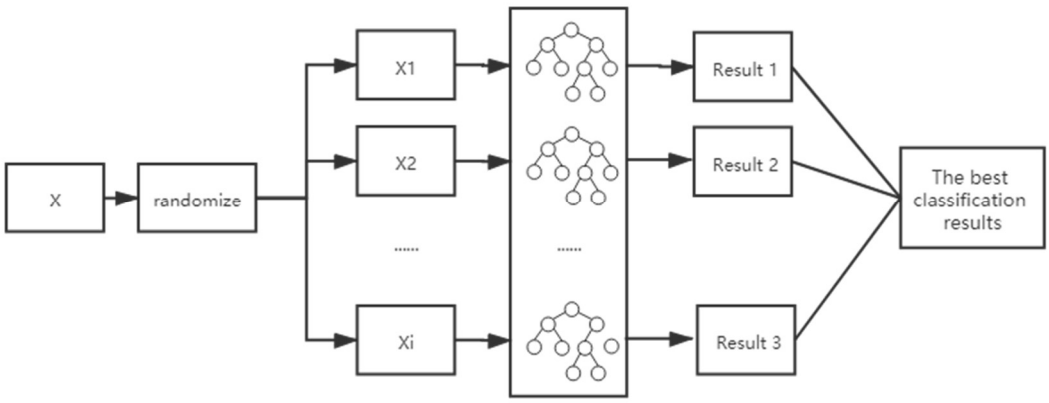

Figure 1. diagram of random forest

2.3 Principles and Characteristics of Random Forest Algorithm

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

Random Forest Algorithm.

Import:

T=Training Set.

Ntree=Number of decision trees in the forest.

M=The number of predictors in each sample.

Mtry=The number of variables participating in the division in each tree node.

Ssampsize=Bootstrap sample size.

Algorithm process:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0035

for(itree=0; 1<itree≤Ntree; itree++)

{

1) Use the training set T to generate a Bootstrap data sample with a size of Ssampsize.

2) Use the Bootstrap data generated above to construct an unpruned tree. In the process of constructing the tree, randomly select Mtry variables from M and select the best variable to branch according to a certain standard(Gini value).

}

Output:

Regression problem: take the average value of all individual tree return values as the prediction result.

Classification problem: use the classification result of most decision trees as the prediction result.

Random forest has the following characteristics: As can be seen from the above algorithm process, the randomness of random forest is mainly reflected in two aspects: the randomness of the data space is realized by Bagging (Bootstrap Aggregating), and the randomness of the feature space is realized by random sub-samples (Random Subspace) way to achieve. For classification problems, each decision tree in the random forest classifies and predicts new samples, and then aggregates the decision results of these trees in a certain way to give the final classification results of the samples.

The random forest algorithm has the following advantages:

1) The introduction of two randomness in the data row (data record) and column (variable) 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. There are two methods for calculating the importance of variables in random forests: one 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 wrong sample, then randomly scramble the value of a column of variables in the Bootstrap sample, re-use the decision tree to predict it, and record again The number of wrong samples. The number of two 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 of GINI impurity. All the nodes in the forest that select a variable as the split variable are summarized to get GINI drop amount.

2.4 Random Forest Method for Classification of Imbalanced Data

The random forest algorithm defaults to the weight of each class as 1, which means that the misclassification cost of all classes is the same. In scikit-learn, the random forest algorithm provides the class_weight parameter, whose value can be a list or dict value to manually specify the weights of different categories. If the parameter is "balanced", then the random forest algorithm uses the y value to automatically adjust the weights, and various weights are inversely proportional to the category frequency in the input data.

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

"balanced_subsample" is similar to the "balanced" mode. The calculation uses the number of samples in the sampling with replacement instead of the total number of samples. Therefore, the problem of unbalanced data classification can be solved through this method.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0035

3. Data Pre-processing and Analysis

3.1 Data Set

The credit investigation is used in this thesis: there are 250000 samples for loan default data set in total, of which 150000 samples are used as training set and 100000 samples are used as test set.

There are 150000 historical data of borrowers in the training set, including 10026 default samples which account for 6.684% of the total sample with 6.684% loan default rate, while 139974 non-default samples account for 93.316% of the total sample. From these, it can be seen that this data set is a typically highly-unbalanced data, which contains borrower's age, income, family, loan and other conditions. In these total 11variables, SeriousDlqin2yrs is the label, and others are predictive characteristics. The following table lists the variable names and data types:

Table 1. Variables

Variable Names	Description	Types
SeriousDlqin2yrs	Default or not	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 monthly 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

3.2 Data Analysis

The experimental environment used in the experiment is Anaconda3+Python3. First of all, we conducted preliminary analysis on data to mainly show the distribution of default rate on every independent variables and generate the frequency distribution table in Table 2 (decimals are all rounded).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0035

Table 2. the Frequency Distribution Table of Variable 'Age'

Age	The total number of people	Proportion	The number of defaulters	Default proportion of this age range
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%
More than 65	28599	19.1%	690	2.4%

From Table 2, it can be obviously seen that default rate of age ranges of below 25 and 26-35 both exceed 10%. In the meantime, as age increases, default rate falls.

Table 3. the Frequency Distribution Table of Variable 'NumberRealEstateLoansOrLines'

NumberRealEstateLoansOrLines	The total number of people	Proportion	The number of defaulters	Default proportion of this range
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%
More than 20	10	0.007%	2	20%

From Table 3, it can be obviously seen that the number of real estate and mortgage loans for 99.47% borrowers are less than 5, while the default rates for whom borrow more than 5 times increase dramatically, of which those more than 10 are all over 20%.

Table 4. the Frequency Distribution Table of Variable 'NumberOfTime30-59DaysPastDueNotWorse'

NumberOfTime30-59DaysPastDueNotWorse	The total number of people	Proportion	The number of defaulters	Default proportion of this range
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%

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0035

From Table 4, it can be obviously seen that borrowers who has never been 30-59 days past due in the last 2 years. However, with the increasing of the number of times that are past due, the default rate significantly rises. For other two variables, the frequency that number of times borrower has been 60-89 days past due but no worse in the last 2 years and that over 90 days also have the same trend as this in Table 4. Therefore, it can be concluded that the more the number of times that borrowers has been past due is, the higher default rate is.

There are 10 variables that are used in the data set of this experiment. We statistically analyze each variable and obtain frequency distribution tables shown above. Excluding that the variable Number Of Open Credit Lines And Loans (the number of open loans and credit loans) has no obvious correlation with default rate, other variables are all correlated to whether borrowers default or not.

3.3 Data Pre-processing

Firstly, in accordance to the preliminary exploration of data, it can be found that missing values are existed in these two variables, Monthly Income and Number Of Dependents, the amounts of which are 29731 and 3924 respectively.

When it comes to abnormal value, the minimum value is 0 in the variable age, which is abnormal. In these three variables of the number of days that has been past due, NumberOfTime30-59DaysPastDueNotWorse, NumberOfTime30-59DaysPastDueNotWorse and NumberOfTimes90DaysLate, there are the minority of values 96 and 98, which may be abnormal values or may be certain behavior codes.

Regardless of the method for data pre-processing, when we use pandas base in Python to read the data, we should define the list by setting parameter na_values in functioning pd.read_csv(), followed by treating 0 in variable age and 96, 98 in three past due variables as NaN value. And then all NaN values in the data set are replaced by the average value of the corresponding columns by using sklearn.preprocessing.Imputer base.

4. Modeling and Experimental Results

4.1 Random Forest Model

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

Random forest is a very important algorithm in the experiment, so we used Sklear.ensemble. RandomForestClassifier in Python to build the model of random forest in this experiment. Some parameters are required to build the model for calculation.

The following lists some parameter Settings:

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 partition is set to 2,

Min_samples_leaf: minimum sample number of leaf nodes, set to 50,

N_jobs: indicates the number of parallel jobs. Set this parameter to -1. Start as many jobs as the CPU has cores.

Class_weight: set it to 'balanced_subsample' and use the y value to automatically adjust the weight.

The frequency of categories in the data is inversely proportional,

Bootstrap: Specifies whether to use bootstrap sample sampling. Set this parameter to True.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0035

The most important setting is n_estimators, n_estimators are theoretically better, but the calculation time increases accordingly. Therefore, it is not true that bigger is better, and the best prediction results will occur in a reasonable number of trees.

Random forest is to get the best bifurcation attribute in a random subset, while ET is completely random to get the bifurcation value, so as to realize the bifurcation of decision tree.

When training the random forest, it is recommended to use the cross_VALIDATED (cross-validation) method. One piece of data n is used as the validation set, and the rest data is used to train the random forest and predict the test set. You end up with n results, and you average the final results

4.2 Modeling Evaluation

Each model has an evaluation, and the model evaluation index used in this experiment is AUC(Area under the ROC Curve) value. What is an AUC? AUC is defined as the area under the Receiver Operating Characteristic (ROC) curve. Obviously, the value of this area is not greater than 1. The horizontal axis of ROC curve was False Positive Rate (FPR), and the vertical axis was True Positive Rate (True).

Positive rate, TPR), and the ROC curve is generally above the straight line y=x, so the value norm of AUC.

It's between 0.5 and 1.The closer AUC is to 1.0, the higher the authenticity of detection method is. When the value is 0.5, the authenticity is the lowest and has no application value. AUC values are used as evaluation criteria because many times the ROC curve is not clear.

Indicate which classifier performs better, and as a numerical value, the classifier with a larger AUC performs better. Therefore, we compared the random forest model with the logistic regression classification model and the decision tree classification model, and the results are shown in the following table.

Table 5. Comparison between random forest and other algorithms

arithmetic	AUC
random forest	0.86
decision-making tree	0.80
logistic regression	0.80

As can be seen from the table, the AUC value of random forest algorithm is higher than that of decision tree and logistic regression algorithm, so random forest. The prediction performance of the algorithm is better than the other two algorithms. The AUC value of random forest is closer to 1, which indicates that its authenticity is higher than other values that are not so close to 1.

4.3 Measure of Feature Importance

There are indeed several ways to get feature “importances”. As often, there is no strict consensus about what this word means.

In scikit-learn, we implement the importance as described in [1] (often cited, but unfortunately rarely read…). It is sometimes called “gini importance” or “mean decrease impurity” and is defined as the total decrease in node impurity (weighted by the probability of reaching that node (which is approximated by the proportion of samples reaching that node)) averaged over all trees of the ensemble.

In the literature or in some other packages, you can also find feature importances implemented as the “mean decrease accuracy”. Basically, the idea is to measure the decrease in accuracy on OOB data when you randomly permute the values for that feature. If the decrease is low, then the feature is not important, and vice-versa.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0035

To the degree of importance of each feature is shown in the table below.

Table 6. To the degree of importance of each feature is shown

variable	feature_importance
RevolvingUtilizationofUnsecuredLines	0.3411
NumberOfTime30-59DaysPastDueNotWorse	0.1694
NumberOfTimes90Days	0.1594
NumberOfTime60-89DaysPastDueNotWorse	0.0727
age	0.0677
DebtRatio	0.0625
MonthlyIncome	0.0448
NumberOfOpenCreditLinesAndLoans	0.0442
NumberRealEstateLoansOrLines	0.0223

As see, The values of each variable are very different, and they are very different, and then we can analyze and focus on those values to find the points that need the most attention.

The feature Importance algorithm of Python calculates the top five important variables is "RevolvingUtilizationofUnsecuredLines","NumberOfTime30-59DaysPastDueNotWorse","NumberOfTimes90Days","NumberOfTime60-89DaysPastDueNotWorse" and "age"

Therefore, when processing loan applications, special attention can be paid to these characteristics of the borrower.

5. Conclusion

This paper mainly studied the loan defaults of common problems in the financial sector, and using the random forest of unbalanced data classification method to predict default model is established, the basic idea of random forest is in the process of a single tree structure, some random variables or characteristics involved in tree node, repeated, and ensure the independence between the trees, According to the unbalanced data, the random forest method can automatically adjust the weight according to the Y value through parameter adjustment, so as to effectively solve the classification problem of unbalanced data. Experiments show that the stochastic forest algorithm has better classification performance than decision tree and logistic regression model, and it has important reference significance for loan default prediction in the financial field. In addition, based on the importance of the characteristics of the measurement, in this experiment can be lending a person's age, debt ratio and number of real estate and mortgage of the three characteristics of the final is greatly influenced by default, the feature importance measure method is the other feature selection problem in data mining to have the important reference significance.

References

[1] Georgiev, Sven Giesselbach, et al. Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems, IEEE Transactions on Knowledge and Data Engineering, 2021.

[2] Ning Li, Li Zhao, A new heuristic of the decision tree induction, 2009 International Conference on Machine Learning and Cybernetics.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0035

[3] Kunal Pahwa, Neha Agarwal. Stock Market Analysis using Supervised Machine Learning, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon).

[4] Chenn-Jung Huang, Ming-Chou Liu, Application of machine learning techniques to Web-based intelligent learning diagnosis system, Fourth International Conference on Hybrid Intelligent Systems (HIS'04).

[5] Yu-Fang Shi, Ping-Ping Song. Improvement Research on the Project Loan Evaluation of Commercial Bank Based on the Risk Analysis, 2017 10th International Symposium on Computational Intelligence and Design (ISCID).

[6] Hong Qiao, Xue-Chen Dong, Research on the risk evaluation in loan projects of commercial bank in financial crisis, 2009 International Conference on Machine Learning and Cybernetics.

[7] Xiu-hua Wang, Ling Liang, The Study on the Loan Risk Pricing of the Bank in Internal Rating-Based Approach, 2008 International Conference on Risk Management & Engineering Management.

[8] Yu Guo-an, Xu Hong-bing, Wu Chao, Design and implementation of an agent-oriented expert system of loan risk evaluation, IEMC '03 Proceedings. Managing Technologically Driven Organizations: The Human Side of Innovation and Change (IEEE Cat. No.03CH37502).

[9] Zhang Xinmei, Yuan Quan, Li Bin, Study on the credit risk mitigation techniques of small loan companies, 2011 International Conference on Product Innovation Management (ICPIM 2011).

[10] R. Gerritsen, Assessing loan risks: a data mining case study, IEEE IT Professional, vol.3(6), 1999.