A Novel Financial Risk Identification Algorithm for Online Fintech Platform

Yingjie Wei[1,a]*, JiaXing[2,b], Pingye Zhou[3,c], Tianye Tu[4,d], Shangzhe Wei[5,e]

[1]Senior High School BCOS, Jiaxing, Zhejiang, China
[2]University of Sheffield, Sheffield, UK
[3]Huawei Foreign Language School, Shaoxing, Zhejiang, China
[4]Henan Experimental high school, Shang Qiu, Henan, China
[5]Kang Chiao International School, Suzhou, jiangsu, China
[a]3036200899@qq.com, [b]zhoupingye0116@163.com, [c]1328693691@qq.com, [d]2083759633@qq.com, [e]1497637462qq. com
**Corresponding author: 3036200899@qq.com*
These authors contributed equally to this work

Abstract: *The issues we study are risk identification, credit granting, pre-loan and post-loan lending risks in consumer lending and trust loans. How to effectively evaluate and identify the potential default risk of borrowers and calculate the probability of default of borrowers before granting loans is a fundamental part of credit risk management in modern financial institutions. This paper focuses on the statistical analysis of historical loan data from banks and other financial institutions with the help of the idea of non-equilibrium data classification, and the use of the random forest algorithm to establish a loan default prediction model. Then a fixed threshold screening in the form of a black and white list is used. The limitation of the current method is that it cannot meet the lending needs of large-scale transaction classes and cannot be judged quickly. When the number of decision trees in a random forest is large, the space and time required for training are large. In addition, it is insensitive to data and has low accuracy. The experiments found that the machine learning-based financial lending risk prediction method can solve the problem better, and the experimental results phenotyped that the neural network and random forest algorithm outperformed the 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 whether the final default can be obtained, so that the lending risk judgment in the financial sector can be more effective.*

Keywords: *component; lending risk identification, fintech, machine learning, algorithmic models, risk prediction*

1. Introduction

With the development of large data and artificial intelligence, more and more financial institutions are turning traditional strategies to quantitative means such as relying on machine learning models. Many scenarios, such as risk identification, credit and loan risk after pre-credit and trust loans, are also suitable for the application of machine learning models. Machine learning methods can be used to perceive risks, analyze risk rules and risk behaviors, assess risk impact, and quantify risks. Machine learning technology provides a new method to improve the risk management level of information technology. There is a certain probability of risk occurrence, which can be predicted by machine learning algorithm.

Most of the existing loan default data are unbalanced. This essay mainly studies how to use the idea of unbalanced data classification to analyze the historical loan data of banks and other financial institutions to predict the possibility of loan default. When there is a large difference in the number of samples from different categories in a data set for a classification task, we usually refer to the data set as "class imbalance". For example, there are 99 positive samples and 1 negative sample in the training set. In many cases, regardless of sample imbalances, the learning algorithm will cause the classifier to discard negative case predictions, because dividing all samples into positive ones will result in up to 99% training classification accuracy. There are three types of processing for data sets: Under-sampling (a technique that reduces the number of samples to improve class balance, such as keeping all samples in negative samples and randomly selecting samples in positive samples to approximate positive and negative samples); Oversampling (techniques that add additional data points to improve class balance, such as sampling negative samples with sampling with replacement until the number of positive and negative samples is balanced); Class uniform sampling method (with some processing on each type of data, so that the probability of each type of data being collected by the final sampling is the same) ; Threshold shift (adjusting the threshold for categorization based on reality, usually specifying a threshold of 0.5, but moving the threshold based on reality increases the weight of a class, A way to alleviate class imbalances. In this paper, we can use the random forest algorithm to build a loan default prediction model, and then use a black-and-white list to filter for a fixed threshold. Loan rights will not be granted if the probability of violation reaches 0.5.

Although the random forest algorithm is fast enough, when the number of decision trees in a random forest is large, the space and time required for training will be large, which will result in slower models. When classifying or regressing problems, random forests are unable to make predictions beyond the range of training set data, which may result in over-fitting of some noise-specific data when modeling. Ultimately, it results in insensitivity to data and low accuracy.

In this essay, we will introduce three machine algorithms: Logical Regression, Random Forest and Neural Network. Logistic regression, also known as logistic regression analysis, is a generalized linear regression analysis model, which is often used in data mining, automatic disease diagnosis, economic prediction and other fields.Logical regression is a bi-classification problem in nature. Simply put, it predicts the probability of an event by fitting a logical function.So it predicts a probability value and the output value should be between 0 and 1.

Random forests can be used for almost any prediction problem, such as classification of discrete values, regression of continuous values, unsupervised learning clustering, and outlier detection. It is a relatively new machine learning strategy and belongs to the large category of integrated learning in machine learning. Random forests are made up of many decision trees, and there is no correlation between different decision trees. When we perform the classification task, new input samples are entered and each decision tree in the forest is judged and classified separately. Each decision tree will get its own classification result. Random forests will take the one with the most classification results in the decision tree as the final result.

Neural network algorithms are inspired by the principles of human neurons (axons, dendrites, nuclei). Neural network models are built on many neurons, each neuron can be considered as a learning unit.These neurons take certain characteristics as input and get outputs based on their own models. The most famous algorithm is back propagation in 1980, which minimizes errors on the basis of more than two parameters, improves the method of using gradient descent, and constantly modifies the values of the two parameters to minimize the final errors. Its basic principle is that the partial derivative of the error is calculated by propagating the final output forward, and then weighted by this partial derivative and the hidden layer ahead, so that the layers are passed back until the input layer (without calculating the input layer) is reached, and the partial derivative from each node is used to update the weight.

2. Related work

Any direct and indirect lending relationship in the banking, P2P, or consumer finance scenarios will produce risks. The generalized risk includes not only the default of the lender but also the operational risk of the lending institution or participant, as well as various uncontrollable factors such as policy and environment. After controlling for policy, environment, institutions, and other factors, the risk of chivalrousness is simply that the lender does not pay back the money. There are a variety of reasons why people don't pay back loans, whether they can't help it or they just don't want to. To ensure that the borrower can repay, the necessary risk control means should be adopted, such as screening before lending to ensure the repayment ability of the lender. And credit risk is greatly affected by the economic cycle: in the period of economic expansion, credit risk decreases, because the strong per capital income capacity reduces the overall default rate; During periods of economic contraction, credit risk increases as the overall per capital income situation deteriorates, increasing the likelihood that borrowers will not be able to repay in full and on time for various reasons. In recent years, China's economic growth rate has slowed down, but it is still in the rising period. The overall default rate is low. With the change of the economic cycle, if there is an obvious economic crisis and large-scale unemployment, the default rate will rise rapidly.

Random Forest has lots of advantages because of the integrated algorithm, its accuracy is better than the most single algorithm, so the accuracy is high. It can deal with data of high dimensions without feature selection, and has strong adaptability to data sets: it can deal with both discrete data and continuous data without normalization of data sets. The training is fast and can be applied to large data sets. Default values can be processed without additional processing. Due to out-of-pocket data (OOB), unbiased estimates of true errors can be obtained during model generation without the loss of training data. Because of its simple implementation, high precision, and strong over-fitting resistance, it is suitable to be used as a reference model when facing nonlinear data.

3. Design or algorithm model

3.1. Random Forest

Random forest is a kind of cluster classification model. Random forest is a random way to build a forest, which is composed of many decision trees, and there is no correlation between each decision tree. This is mainly applied to regression and classification. Bootstrap sampling is conducted for the random forest, and each node variable is generated only among a few variables randomly selected when each tree is generated. Through “the bootstrap” resampling technique, k samples (k is generally the same as N) were randomly and repeatedly selected from the original training sample set N to generate a new training sample set. Then N classification trees were generated according to the bootstrap sample set to form a random forest. Its essence is an improvement to the decision tree algorithm, which merges several decision trees, and the establishment of each tree depends on an independent sample set. OOB statistics of random forest model: when the random forest model uses the Bootstrap sampling method to extract samples, N represents the number of samples in the training set. When N is large enough, the probability of not being selected in each sample in the training set is deduced according to the limit to converge to 36.8%. In other words, nearly 37% of the samples in the original data will not appear in the Bootstrap samples. These data become out-of-pocket data and can be used to estimate the model generalization error.

First of all, put-back sampling is taken from the original data set to construct a sub-data set. The data amount of the sub-data set is the same as that of the original data set. The elements in different sub-datasets can be repeated, and the elements in the same sub-data set can also be repeated. Second, sub-decision trees are constructed by using sub-datasets. The data is put into sub-decision trees, and each sub-decision tree outputs a result. At last, if there are new data that need to be classified by random forest, the output results of random forest can be obtained by voting on the judgment results of sub-decision trees.

3.2. Logistic regression

Logistic regression is one of the classification and prediction algorithms. Predicting the probability of future results through the performance of historical data is mainly used for two classification problems: the first is used for prediction, and the second is to find the influencing factors of dependent variables (logistic regression is used for classification to know which influencing factors are dominant, so an event can be predicted). For example, we can set the probability of purchase as a dependent variable, set the user's characteristic attributes, such as gender, age, registration time, etc., as independent variables, and predict the probability of purchase according to the characteristic attributes. The process of logistic regression can be summarized as: facing a regression or classification problem, establish the cost function, then iteratively solve the optimal model parameters through the optimization method, and then test and verify the quality of our model. In the regression model, y is a qualitative variable, such as y = 0 or 1. The logical regression method is mainly used to study the probability of occurrence at some time.Logistic function (or sigmoid function), in the form of:

$$g(z) = \frac{1}{1+e^{-z}}$$

The main process of logistic regression is divided into:(1) Build a prediction function. Before construction, it is necessary to determine whether the function model is linear or nonlinear according to the data.(2) Build the cost function. This function represents the deviation between the predicted value (x) and the actual value (y), which can be the difference between them (X-Y) or other forms. Calculate the sum or average "loss" of all data, and record it as J(θ) Function, which represents the deviation between the predicted value of all data and the actual value.(3) Gradient descent algorithm, minJ(θ).When the amount of data is large, the gradient descent algorithm is not efficient, and the random gradient descent algorithm can make up for its defects.

3.3. Neural network

Neural network is an algorithm model that simulates the human brain. The principle of neural network algorithm is the concept of information transformation replaced by symbols, and then inferences are derived from symbolic operations, and can be converted into instructions for computers to execute. The principle is to store information in a distributed manner and to process it in parallel. The network composed of a large number of units can realize the calculation of complex data and the complex nonlinear learning system.

The algorithm model is divided into input layer, hidden layer and output layer. The hidden layer can contain a lot of hidden information. For example, a typical three-tier network. Each circle represents this neuron. There are also more common formats like the output layer has two nodes, or multiple hidden layers.

Each node in the input layer must perform a point-to-point calculation with each node in the hidden layer. The calculation method is weighted sum + activation. Then use the hidden layer to calculate each value, and finally use the same method and output layer to calculate. Initially, the input layer values are respectively propagated to the hidden layer through network calculations, and then propagated to the output layer in the same way. Finally, the output value is compared with the sample value and the error is calculated.

But then the error back propagation algorithm, the BP algorithm, became a more effective calculation method and enabled the research of neural network calculation to progress.

The neural network of BP algorithm can be used for classification, clustering, prediction and so on. But a certain amount of historical data is needed. Through the training of historical data, the network can learn the hidden knowledge in the data. In the kind of problems encountered, we must first find the characteristics of the problem and the corresponding evaluation data, and use these to realize the training of the network nerve. This is an effective method for calculating partial derivatives. But the BP algorithm still has some shortcomings: the calculation is complicated, the speed is slow, and it is easy to fall into the local optimal solution.

4. Experiment result

In this experiment, we used the credit data set, which contains 250,000 samples about the loan default data in total, 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 borrowers' historical data, of which 10,026 are defaulted samples, accounting for 6.684% of the total sample, with a loan default rate of 6.684%, and 139,974 non-defaulted samples, accounting for 93.316% of the total sample. It can be seen that the data set is a typical highly unbalanced one. We use the formula n_samples / (n_classes * np.bincount(y)). The calculation of uses the number of samples in a sample with a put-back formula, rather than the total number of samples. Thus, we can solve the problem of unbalanced data classification by this method. The data includes the borrower's age, income, household, etc. and loan status for a total of 11 variables, where SeriousDlqin2yrs is the label and the other 10 variables are predictive characteristics. The table I shows the variable names and data types.

Regarding the second part of the experiment, the correlation between the variables. There are many examples that can be cited, such as food production and fertilization; people's weight and age increase, etc., among which there are some similar functions. Among the correlations, for example, the more fertilizer is applied, the greater the food output will be, and the weight of people within a certain range of age will increase (young children).

The multiple correlation coefficient is the relationship between a dependent variable and multiple independent variables. Take a very simple example: the seasonal demand for goods in supermarkets and the salary of employees, the price of goods, etc.

Table 1 Data set variables situation

Variable Name	Description	Type
SeriousDlqin2yrs	Person experienced 90 days past due delinquency or worse	Y/N
Revolving Utilization Of Unsecured Lines	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
Number Of Time 30-59 Days Past Due Not Worse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	Integer
Debt Ratio	Monthly debt payments, alimony,living costs divided by monthly gross income	Percentage
Monthly Income	Monthly income	Real
Number Of Open Credit Lines And Loans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	Integer
Number Of Times 90 Days Late	Number of times borrower has been 90 days or more past due.	Integer
Number Real Estate Loans Or Lines	Number of mortgage and real estate loans including home equity lines of credit	Integer
Number Of Time 60-89Days Past Due Not Worse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	Integer
Number Of Dependents	Number of dependents in family excluding themselves (spouse, children etc.)	Integer

The correlation coefficient is a statistical indicator designed by statistician Carl Pearson to describe the amount of linear correlation between variables. Generally, it is represented by the letter r. Due to different research objects, the correlation coefficient has different definitions. , The most commonly used is called Pearson correlation coefficient. The Pearson correlation coefficient is not the only correlation coefficient, but the most common correlation coefficient.

Regarding how to build the algorithm model, first we need to pre-process the data and observe if there are any missing values in the data. We can see that there are 22,729 missing values in the test sample and 33,655 missing values in the training sample. The next step is to look at the underlying descriptive information for each variable and get some guesses from the observations. In the process of tuning the parameters of the random forest regression model, we need to find the best value and then tune max_depth. next, the parameters are tuned. First, the optimal value of n_estimators is explored using gridsearchcv. Secondly, the optimal value is found for the maximum depth of the decision tree max_depth and the minimum number of samples needed to subdivide the internal nodes. The minimum number of samples needed for internal node repartitioning, min_samples_split, cannot be determined together for the moment, as it is also related to other parameters of the decision tree. Third, find the optimal parameters for min_samples_split and min_samples_leaf for the minimum number of samples needed for internal node repartitioning and for the minimum number of samples for leaf nodes. Fourth, find the optimal parameters for the maximum number of features max_features. Finally, we aggregate the best parameters we searched for and train them. Tuning the parameters can optimise the results.

Decision Tree AUC is 0.80. The essence of Decision Tree learning is to generalize a set of classification rules from the training data set. The loss function of Decision Tree learning is usually the regularized maximum likelihood function, and the learning strategy is to estimate the conditional probability model from the training data set. After the construction of the Decision Tree is completed, the evaluation function (objective function) is used to judge the quality of the current Decision Tree.

Random Forest AUC is 0.86. Random Forest is an algorithm that integrates multiple trees through the idea of Ensemble Learning. Its basic unit is the Decision Tree, and its essence belongs to a large branch of machine Learning 一 Ensemble Learning. Random Forest has a wide range of applications, from marketing to health care insurance. It can not only be used to model marketing simulation, calculate customer sources, retention and loss, but also be used to predict the risk of disease and the susceptibility of patients.

Back Propagation AUC is 0.80. Logical thinking refers to the process of reasoning according to logical rules. It first conceptualizes information and expresses it with symbols. Then, it makes logical reasoning according to symbolic operation according to serial mode. This process can be written as serial instructions for the computer to execute. Intuitive thinking, however, is the synthesis of distributed storage of information, resulting in a sudden idea or solution to a problem.

5. Conclusion

This paper focuses on the commonly seen loan default problem in the financial field uses a random forest approach for unbalanced data classification to build a model for predicting loan defaults. The basic idea of random forest is that in constructing a single tree, some variables or features are randomly selected to participate in the tree node division, repeated several times, and the independence between these established trees is guaranteed. For non-equilibrium data, the parameter adjustment makes the random forest method can adjust the weights automatically according to the y-values. Thus, the data classification problem of unbalanced data can be solved effectively.

It is shown through experiments that the random forest algorithm has better classification performance than decision tree and logistic regression models. It is an essential reference for the loan default prediction problem in the financial field. In addition, by measuring the importance of each feature, in this experiment, we can get three features of lender's age, debt ratio, and the number of real property and mortgage loans can be obtained in this experiment. And also whether they ultimately have a more impact on default. This measure of feature the importance of features also has a more important reference for other feature selection problems in data mining.

References

[1] John W. Tukey. The Future of Data Analysis. July,1 1961: 2-10.
[2] Xindong Wu & Xingquan Zhu. Data Mining with Big Data. Jan 2014: 1-4.
[3] Placebo 2019. Unbalanced data classification 1 - undersampling / oversampling.
[4] DataVisor 2018. Why is unsupervised machine learning anti fraud mainstream.
[5] Daniel T. Larose. Data Mining Methods and Models. 2006: 2-20.