ISSN 2616-5902 Vol. 3, Issue 8: 39-45, DOI: 10.25236/AJBM.2021.030808

Identification and Prediction Methods of Financial Anti-fraud

Kejia Zhu[1, *], Pengzhou Fang[2], Haocheng Li[3], Ruihan Shi[4], Yutian Shi[5] , Yanyuzi Chen[6]

[1]University of Nottingham Ningbo China, Ningbo, Zhejiang, China
[2]University of Toronto, Toronto, Ontorio, Canada
[3]Beijing World Youth Academy, Beijing, China
[4]Southwest Jiaotong University, Chengdu, Sichuan, China
[5]Shanghai University of Finance and Economics, Shanghai, China
[6]The High School Attached to Northwest Normal University, Lanzhou, Gansu, China
**Corresponding author: biykz4@nottingham.edu.cn*
These authors contributed equally to this work

Abstract: *Nowadays, evaluating and identifying the potential fraud risk of borrowers effectively and calculating the fraud probability of them are the basis and significant steps of credit risk management in modern financial institutions before issuing loans. This paper mainly studies the statical analysis of the historical loan data of financial institutions based on the idea of unbalanced data classification and establishes the prediction model of loan fraud through random forest, decision tree and regression algorithm. The prediction performance of random forest algorithm is better than the other two mentioned methods. Additionally, it may obtain the feature that have a remarkable impact on the final fraud by ranking the importance of those features, which leads to a more effective judgment on the credit risk in the financial field.*

Keywords: *Random forest, Bank reference, Prediction of loan fraud, Data mining.*

1. Introduction

China's economy has been developing rapidly with deepening of the reform and opening-up since 21st century. Under the context of flourishing world economy, people's consumption concept and companies' development patterns have undergone great changes and loans are gradually becoming important ways to solve personal and companies' economic problems. With people's needs increasingly expand, banks' loan business changes as well and non-performing loans, whose another name for loan fraud is soaring. After entering the digital era, the financial anti-fraud measures have gradually developed from the spread of financial knowledge at the beginning to artificial intelligence and machine learning (Su, 2018). At present, the most common financial anti-fraud methods are credit reference system, blacklist system, user behavior risk identification engine, etc., which are usually used in the Internet finance industry (Su, 2018). In order to avoid loan fraud, financial institutions such as banks are taking strict comprehensive appraisals on credit risks of borrowers to predict their fraud probabilities when they apply for loans. These institutions are taking appraisal results as standards to grant loans. Using a scientific model and appraisal standard to predict the probabilities of loan fraud can lower risks and improve profit margins to the greatest extent.

However, the current financial anti-fraud management is still not perfect, and the analysis and identification technology of financial risk also needs to be further enhanced. Specifically, some risk information is usually separated and fragmented, which will affect the accuracy of the results in the process of machine learning, and ultimately may lead to risk misjudgment and serious losses (Su, 2018). Therefore, establishing an effective appraisal system to identify potential fraud risks of the borrowers is a foundation and key link of credit risk management in financial intuitions.

This article discusses how to use unbalanced data classification for historical loan data in financial institutions and make analyses on history data based on classification models of Random Forest, Logistic Regression and Decision Tree so as to have prediction models of probabilities of loan fraud.

ISSN 2616-5902 Vol. 3, Issue 8: 39-45, DOI: 10.25236/AJBM.2021.030808

2. Random Forest

2.1 Introduction of Random Forest

The algorithm of Random Forest is an ensemble learning method based on units of decision tree, which means to establish a random forest by using decision trees with random selection. Every none-leaf node in the decision tree is a random test of feature property, later, multiple branches are created by tests of every feature property and select output branches according to their values and use stored categories of leaf nodes as decision results (Zhou, 2014). Constituting random forests is to constitute multiple independent decision trees in repeat ways. From the original training sample set, repeatedly randomly select k samples with replacement to generate a new training sample set, and then generate k classification trees to form a Random Forest based on the new training sample set. After entering a new sample, each decision tree in the forest will classify and judge the sample, and the classification result with the highest vote in the random forest is the predicted category of the sample. The schematic diagram of the process is as follow:

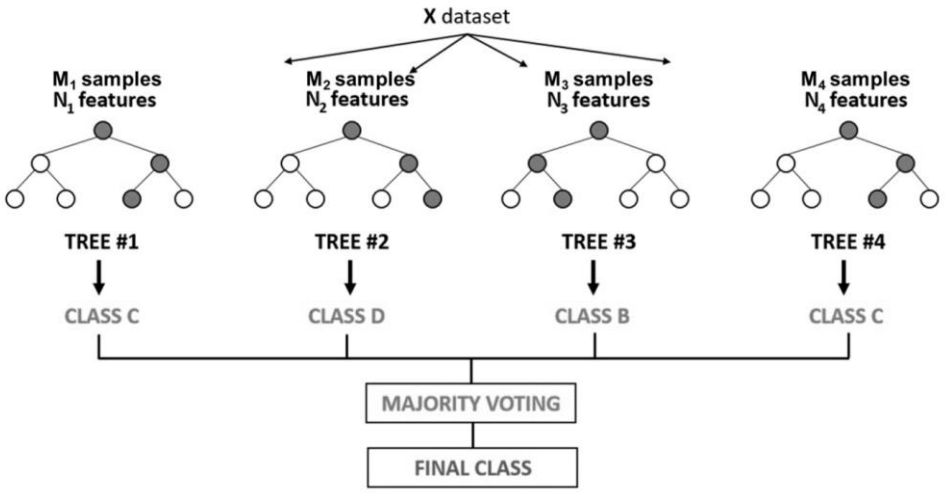

Figure 1: The schematic diagram of the process

2.2 Principles and Characteristics of Random Forest Algorithm

Random Forest algorithm includes two aspects: classification and regression. The algorithm procedures are as follow:

Random Forest Algorithm

Import:

T=training set

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

M=the number of predictor variable per sample

M_{try}=the number of variables that participate in the partition in each node

$S_{sampsize}$=the sample size of Bootstrap

Procedure:

For($i_{tree} = 0; 1 < i_{tree} \leq N_{tree}; i_{tree}++$)

{

1.Use the training set T to generate a data sample of Bootstrap which's size is $S_{sampsize}$.

2.Use the Bootstrap data generated above to construct an untrimmed tree: t_{tree}. During the procedure of constructing t_{tree}, select M_{try} variables randomly from M and choose the best variable to branch according to some criteria (GINI).

}

Export:

Classification: take the average of all the returned values as the prediction result.

Regression: take the classification results of most decision trees as the prediction result.

ISSN 2616-5902 Vol. 3, Issue 8: 39-45, DOI: 10.25236/AJBM.2021.030808

Randomness of Random Forests mainly embodies in two aspects from the previous algorithm procedures: randomness of the data space is achieved by Bagging (Bootstrap Aggregating) and randomness of feature space is achieved by Random Subspace. This can make samples and creation of node variables achieve randomness.

Because the final classification result of each new sample is obtained by the ensemble learning of each decision tree in the Random Forest, and at the same time has good randomness, the Random Forest has the following characteristics:

First, randomly introduced rows (data records) and columns (variables) in the data. Random Forest is not easy to fall into overfitting;

Second, Random Forests have good anti-noise ability;

Third, for data sets with many missing values, Random Forests can also effectively estimate and process the missing values and get better results;

Fourth, the ability to adapt to the data set is strong. There is no need to standardize the data, and can deal with both discrete data and continuous data;

Fifth, it can evaluate the importance of each feature in the classification problem and sort it. Random Forest calculates the importance of variables in two ways. One method is based on the average drop accuracy of OOB (Out of Bag). For each decision tree, first use OOB samples to test and record the number of misclassified samples, then randomly shuffle the order of a column of attribute values in the Bootstrap sample, re-use the decision tree to predict, and record the number of misclassified samples again. The ratio of two misclassified samples to the number of OOB samples is the change in the error rate of the decision tree, and the average error rate of all decision trees is the error rate of the Random Forest, which is the average decreasing accuracy rate (Liu, 2014). The other method is based on the GINI drop during the split. The node splitting of Random Forest refers to the decline of GINI impurity. The amount of GINI decline is the sum of all nodes in the forest that select a variable as the split variable.

3. Unbalanced Data

3.1 Classification of Unbalanced Data

Unbalanced data means that the number of one type of data far exceeds the other type. In this article, the samples without loan fraud far exceed the samples with loan fraud. Unbalanced data is common in many fields, such as network intrusion detection, financial fraud transaction detection, text classification (Guo, 2013). The classification problem of dealing with unbalanced data can be based on the penalty weight of positive and negative samples, that is, when building the model, different weights are assigned to categories of different sample sizes. Generally, the weight of small sample size is higher, and the weight of large sample size is lower.

3.2 Random Forests Methods of Unbalanced Data Classification

The Random Forest algorithm assumes that the misclassification costs of all classes are equal, that is, the weight of each class is 1 by default. In python scikit-learn, the Random Forest algorithm provides the class_weight parameter, whose value can be a list or dict value, and you can manually specify the weights of different categories. If the parameter is "balanced", the Random Forest algorithm uses the y value to automatically adjust the weight, and the weight of each category is inversely proportional to the category frequency in the input data.

The "balanced_subsample" is similar to the "balanced" mode. The calculation uses the number of samples with replacement instead of the total number of samples (Xiao, 2013).

4. Experiment Method

4.1 Data Preprocessing and Data Analysis

4.1.1 The Data Set

In the data set, we have totally 250000 samples, where 150000 samples are used for traing and the rest are used for testing. The dataset includes 11 variables that have impact on the financial fraud decision,

ISSN 2616-5902 Vol. 3, Issue 8: 39-45, DOI: 10.25236/AJBM.2021.030808

including age, income and family background. Moreover, there are 10026 financial fraud samples and 139974 clean samples in the dataset. The following table lists variable names with description and data types.

Table 1: Data Set Variables

Variable Name	Description	Type
SeriousDlqin2yrs	Indicator of whether there is a financial fraud	Y/N, 1/0
RevolvingUtilizationOfUnsecuredLines	The percentage of personal credit debt to the total personal credit	Percentage
age	The age of the borrower	int
NumberOfTime30-59DaysPastDueNotWorse	How many times the borrower pay back late in 30-59 days after the due date in the past 2 years	int
DebtRatio	The propotional of monthly debt payment and living cost to the monthly income	Percentage
MonthlyIncome	Monthly income of the borrower	Real number
NumberOfOpenCreditLiesAndLoans	The sum of quantity of open lonas and lines of credit	int
NumberOfTimes90DaysLate	How many times the borrower pay back the debt after 90 days past the due date in the past 2 years	int
NumberRealEstateLoansOrLines	The number of real estate loansand credit lines	int
NumberOfTime60-89DaysPastDueNotWorse	How many times the borrower pays back the debt after 60-89 days past the due date in the past 2 years	int
NumberOfDependents	Exclude the borrower, how many people in the family needs to be raised by the borrower	int

The dataset itself does have potential problems. First and foremost, there are missing values in MonthlyIncome variabe and Number of Dependents variable. Besides, we found that in age variable, there are values below zero, which is an error value, since the minimum of the age should be 0. Also, in Number of Time 30-59 Days Past DueNotWorse, NumberOfTime60-89DaysPastDueNotWorse and Number of Times 90DaysLate variables, there exist a small amount of values such as 96, 98, which may be caused by error or codes. To deal with such issues, when we implement pandas in Python to read the date, we set na_values in pd.reast_csv() as self-defined list and turn all missing values and error values into NaN, and later turn all NaN value into their corresponding average of that column, by sklearn. preprocessing.Imputer.

4.1.2 Data Analysis

The experimental environment used in this paper is Anaconda3+Python3. At first, we proceeded a preliminary analysis and this experiment mainly analyzed the distribution of financial fraud rate on each independent variable and generated the frequency distribution table as shown below(all decimals are rounded).

Table 2: Frequency Distribution Categorized by Ages

Age	Number of people	Proportion	Number of fraud people	fraud rate
Under 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%

It is clear from Table 2 that the proportion of people who are in default aged both under 25 and 26-25 years old is just above 10% and the fraud rate declines with age going up. The over 65 years-old group shows the lowest fraud rate.

Table 3: Frequency Distribution of Variable 'NumberRealEstateLoansOrLines'

NumberRealEstateLoansOrLines	Number of people	Proportion	Number of fraud people	fraud rate
Under 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%

It can be seen from the table 3 that 99.47% of borrowers had less than 5 real estate and mortgage loans whose fraud rate is only 6.6%. However, the fraud rate of borrowers with more than 5 loans

ISSN 2616-5902 Vol. 3, Issue 8: 39-45, DOI: 10.25236/AJBM.2021.030808

increased significantly and the fraud rate of borrowers with more than 10 loans was more than 20%.

Table 4: Frequency Distribution of Variable 'NumberOfTime30-59DaysPastDueNotWorse'

NumberOfTime30-59DaysPastDueNotWorse	Number of people	Proportion	Number of fraud people	fraud rate
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%
Above 7	104	0.07%	50	48.07%

It is clear from Table 4 that the fraud rate of borrowers who have not been overdue for 30-59 days is only about 4%, but with the increase of overdue times, the fraud rate increases significantly. Although the other two variables 'NumberOfTime60-89DaysPastDueNotWorse' and 'NumberOfTimes90Days Late' do not show detailed tables here, both fraud rate of them illustrate the similar trend like Table 4. Therefore, it could be concluded that the more times the borrower overdue, the higher the fraud rate.

There are 10 variables in the data set of this experiment. We got the frequency distribution tables as shown above by statistical analysis of each variable. Moreover, all variables are related to whether the borrower cheating or not, except the variable 'NumberOfOpenCreditLiesAndLoans'.

4.1.3 Data Preprocessing

Firstly, a preliminary exploration of the data reveals that there are missing values in the variables of 'MonthlyIncome' and 'NumberOfdependents' and the number of missing values is 29731 and 3924. Additionally, an outlier condition is observed, which is that the minimum value in the age variable is 0. Also, in those three overdue days variables 'NumberOfTime30-59DaysPastDueNotWorse' 'NumberOfTime60-89DaysPastDueNotWorse' 'NumberOfTimes90DaysLate', there are a few values of 96 and 98, which could be outliers or some behavior code.

During the data preprocessing, we set the parameters na_values in the function pd.read_csv() to our own defined list and treated the minimum value 0 in the age variable and values of 96, 98 in the three overdue days variables as NaN. After that, the 'sklearn.preprocessing.Imputer' library was used to replace all NaN in the data set with mean value of the corresponding columns.

4.2 Modeling and Experimental Results

4.2.1 Modeling of Random Forest

In this paper, we implemented random forest model by sklearn.ensemble.RandomForestClassifier in Python. The following table shows the initial parameters in this model and their description.

Table 5: Model Parameters

Parameters	Description
n_estimators	Set the number of decision trees to 100
obb_score	Bool of whether using outside data, initially to be true
min_sample_split	The minimum number of samples to each spot, set to be 2
Min_sample_leaf	The minimum number of samples to each leaf point, set to be 50
N_jobs	The number of jobs can be working together, according to the cores of the CPU
Class_weight	"balanced_subsample", automatically adjust weight, according to y
bootstrap	Whether use bootstrap, set to be True

For analysing the model, we implement AUC (Area under the ROC curve) index, where AUC is between 0.5 and 1. And the larger AUC, the higher efficiency and effectiveness of the model.

4.2.2 Model Evaluation

The AUC (Area under the ROC curve) value was used to evaluate the performance of the model. The x-axis of the ROC (Receiver Operating Characteristic) curve is the FPR (False Positive Rate), while the y-axis is the TPR (True Positive Rate). The area under the ROC curve at all times is less than or equal to 1. In addition, as the ROC curve in general is above y=x, the value of AUC is between 0.5 to 1. The closer the value is to 1, the higher the performance of the model. The AUC value was used because the

ISSN 2616-5902 Vol. 3, Issue 8: 39-45, DOI: 10.25236/AJBM.2021.030808

ROC curve cannot legibly state which model is best one.

The model used for this experiment, Random Forest, was compared with Logistic Regression and Decision Tree. The results are shown in the table below.

Table 6: Random Forest Compared with Other Models

Model	AUC value
Random Forest	0.86
Logistic Regression	0.80
Decision Tree	0.80

As shown in the table, the AUC value of Random Forest is higher than Logistic Regression and Decision Tree. Hence, it can be concluded that the performance of Random Forest is higher than that of the other two models.

4.2.3 Measurement of Feature Importance

The results are derived from using the method of feature_importance in sklearn. Ensemble. RandomForest Classifier. The importance of each feature is shown in the table below.

Table 7: Feature Importance of Variables

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	0.223
NumberOfDependents	0.0117

As shown in the table, the Revolving Utilization of Unsecured Lines, Number of 30-59 days overdue in the past two years, number of more than 90 days overdue in the past two years have the highest feature importance and contribute to ultimately determining loan fraud. Thus, when processing a loan application, the focus should be on these features.

5. Conclusion

By establishing the forecasting model of loan fraud based on the Random Forest method for non-equilibrium data classification, this paper provides some effective suggestions for predicting loan fraud in the financial field. Our experiment shows that, compared with Decision Tree model and Logistic Regression model, Random Forest model has a higher AUC value, which indicates that this model has a better classification performance in predicting loan fraud and higher application value. The basic idea of this method is to form a complete single tree by self-selecting random vectors to generate different nodes, and then repeatedly generate a large number of mutually independent trees to form a forest, and use this random forest for classification and regression (Zhang et al., 2014). In this paper, aiming at the non-balance of data, we adjust some parameters so that the Random Forest method can automatically adjust the weight according to the Y value, so that the optimized random forest method can reflect the classification situation better.

The above methods are used to analyze the experimental data we collected and it can be concluded that while the age of the borrower, the debt ratio and the number of real estate, loans or lines are the three features that have the greatest impact on the fraud results, the number of open credit lines and loans is the only irrelevant variable among all ten predictor variables. This method of measuring the importance of features also has important reference significance for feature selection in other data mining.

References

[1] Guo, W. (2013) Classification of Imbalanced Datasets Research Based on Ensemble Learning.

ISSN 2616-5902 Vol. 3, Issue 8: 39-45, DOI: 10.25236/AJBM.2021.030808

Available at: http://www.doc88.com/p-0753734954688.html (Accessed: 27 February 2021).
[2] Liu, J. (2014) 'Research on Classifying Unbalanced Data Based on Penalty-based SVM and Ensemble Learning', Computer Applications and Software, 31(1), pp. 186-190. Available at: https://www.ixueshu.com/document/214c9856c46110072903558263d62dcc318947a18e7f9386.html (Accessed: 27 February 2021).
[3] Su, B. (2018) 'Strengthen Financial Anti-Fraud Capability', HINA Finance, pp. 72-74. Available at: https://www.ixueshu.com/document/22cde1e97521595bc1b0dd42e57f8555318947a18e7f9386.html (Accessed: 14 March 2021).
[4] Xiao, J. (2013) Research on Imbalanced Data Classification Method Based on Random Forest Algorithm. Available at: https://www.doc88.com/p-6562591316684.html (Accessed: 27 February 2021).
[5] Zhang, L. et al. (2014) 'the basic principle of random forest and its applications in ecology: a case study of Pinus yunnanensis', Acta Ecological Sinica, 34(3), pp. 650-659. Available at: http://www.ecologica.cn/stxb/ch/html/2014/3/stxb201306031292.html (Accessed: 24 February 2021).
[6] Zhou, B. (2014) Classification and Application of Ensemble Learning in Unbalanced Data. Available at: https://www.ixueshu.com/document/a46919a580be9f73b5095400f4449332318947a18e7f9386.html (Accessed: 27 February 2021).