The 2022 International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/5/20220081

Loan Prediction Using Machine Learning Methods

Simiao Wang[1, a, *, †], Shengqi You[2, †], and Shenwei Zhou[3, †]

[1]West Vancouver Secondary School, West Vancouver, BC, Canada
[2]College of Information Engineering, Zhejiang University of technology, Wenzhou, China
[3]College of Information Engineering, Shenzhen University, Shenzhen, China
a. carlw249@edu.sd45.bc.ca
**corresponding author*
[†]These authors are equally contributed

Abstract: Credit risk has always been the most important risk faced by commercial banks. Credit risk management has important practical significance for preventing credit risk. With the emerging of machine learning algorithms, numerous frameworks, including linear regression, support vector machine, random forest and decision tree are proposed with satisfying performance and robust accuracy. This paper will focus on predicting credit outcomes and calculating forecast accuracy from a given dataset. This paper adopts three algorithms, decision tree, random forest and logistic regression, to calculate the data set from the Bank of Portugal separately and obtain relevant conclusions. Finally, the authors evaluate the advantages and disadvantages of the three methods according to the accuracy of the prediction results, and the conclusion is described as follow, First, all three methods have great potential on handling loan prediction task. Second, the logistic regression algorithm is the most accurate, which obtains 86.4% accuracy.

Keywords: Loan prediction, Machine learning, Logistic regression

1. Introduction

Among commercial banks, credit business is the largest proportion of banks and the most important income and efficiency assets, but one-sided pursuit of asset scale expansion is huge risk. The long-term consumer loans in the credit business are important categories of personal loans. Because of its long-term income period, low adverse rate, and significant comprehensive efficiency, it has long-term attention from commercial banks. The so-called prediction and analysis of loans is to take corresponding scientific methods to predict, calculate, analyze and judge a series of future uncertain factors such as the development status, trend and results of the production and circulation process of bank loans, and put forward scientific data and demonstration. It provides services for the optimal loan issuance, ensures the rational use of loans, and improves the economic benefits of society and banks themselves. Credit risk is the most important risk faced by commercial banks. With the slowdown of national economic growth and the deepening of interest rate liberalization, the deposit and loan spreads shrink, and the profit space and profitability of banks continue to shrink. At the same time, due to the downward trend of the national economy and the deterioration of the financial conditions of bank customers, the credit risks faced by banks are constantly increasing. Therefore, in the current environment, it is particularly important to do a good job in credit risk management, improve the level of operation and management, enhance the ability of credit risk prevention and control and resolution, and promote the rapid and effective development of various credit businesses. Research on loan default risk of enterprises not only has important practical significance for financial institutions to solve the problem of "reluctance to lend" and prevent credit risks, but also can put forward targeted suggestions and measures for enterprises to standardize their own operations and improve their financial conditions.

At present, many researches have been made on bank credit. Some scholars look for the law of bank credit data according to its changing characteristics and establish a forecasting model. Li et al.[1] first analyzed the general principle basis of the bank's loan reserves from the perspective of loan risk premium. On the basis of this, the theoretical basis for bank prediction loan preparations under the incomplete information conditions of the credit market. And the current status and existing problems of the banking industry's banking industry, and then the Markov chain predicting theory has built the theoretical model of forecast loan reserve method, and analyzed its rationality. Finally, the loan reserve method can be used as a reference to the reference loan reserve policy "conclusion. Jiang et al.[2] established the time series ARIMA model of bank loan scale based on the total monthly loan scale of Chinese banks from 2007 to 2010. They believed that the model fitted the predicted value and the actual value to a high degree, and had important reference value for financial institutions to make decisions. Chen et al.[3], according to an organization's corporate loan default data study of loan risk of default, the first missing value of original data processing, feature selection and unbalanced data processing, and then using logistic regression, random forests, XGBoost and LightGBM four machine learning methods for data modeling and analysis model and compare advantages and disadvantages, the most Finally, the GBDT model is used to calculate the feature importance. The results show that: (1) the prediction effect of the three integrated models is significantly better than that of the single model; (2) LightGBM model shows the best prediction performance among the integrated models; (3) The tax payment and credit granted by the enterprise can be used as an important reference to judge whether the enterprise will have overdue loan phenomenon. Li et al.[4] used the combination model of cointegration regression and ARMA to make short-term forecasts of residents' medium- and long-term consumer loans through the housing sales price index. First, the data of 37 periods from January 2007 to January 2010 are used to Granger causality test, and then the combination of cointegration regression and ARMA is used to establish a prediction model. The model forecasts the medium - and long-term consumer loans of residents in 5 periods from February to June 2010. Compared with the actual data, the relative error of the prediction is less than 1.5% Policy recommendations for Guan.

In addition, some scholars studied the important factors affecting the change of bank loans and analyzed the trend of bank loans through these factors. Li et al.[5] analyzed a series of real estate bubble crises in the United States, Japan and East Asia since the 1980s, and pointed out that the rise of real estate prices, the emergence and collapse of real estate bubbles, and the occurrence of economic crises were closely related to the expansion of bank credit. Goodhart et al.[6] based on the data of 12 countries, through VAR impulse response function analysis, showed that the housing price in most countries has a significant impact on bank credit, but there is no obvious evidence to confirm the significant impact in the opposite direction or the interaction between the two.

According to the related works and the importance of loan prediction[7-10], this article first uses the three methods of decision Tree, Random Forest, and Logistic Regression to conduct separate algorithms and obtain relevant conclusions, and then integrate and data sorting the experiments listed different algorithms listed. The dataset gets 100% accuracy and proves that the Decision Tree algorithm can achieve the best performance. In order to generate accurate algorithms, there is a need for sufficient variables and details of an accurate big data set. After searching the database online, the utilized data set comes from Kaggle, which accurately records information about the

1,162 bank customers and 17 variables of a Portuguese bank, including their age, occupation type, etc. For numerical variables, the authors analyze their statistical values (specifically: their average, standard deviation, maximum value, Q1, Q2, Q3, minimum value). Through the Decision Tree method, it can be seen from the output result of the experiment that the training accuracy is 100 %, but the test accuracy is reduced. Therefore, the model is overfit, so this article has tried some high parameters adjustment.

The dataset is taken from the Bank of Portugal in the 20th century and has an enormous data entries with 11162 clients. 7 numerical variables and 10 categorical variables of the clients are recorded. These include their age, job, marriage status, education status etc. In order to analysis categorical variables, they are first converted into numerical values. For example, under the categorical value “marriage status”, the value “single” is recorded as “0”, the value “married” is recorded as “1”, and the value “divorced” is recorded as “2”.

2. Methods

2.1. Decision Tree

Decision Tree in general is a tool to make decisions. Values of an entry are compared with parameters before it split and eventually output an outcome. In Machine Learning, the tree can be formed by two entities: decision nodes and leaves. The leaves are the decisions or the final outcomes, and the decision nodes play a role of splitting the entry. After preprocessing the data set, the model is trained and tested with following result (figure 1).

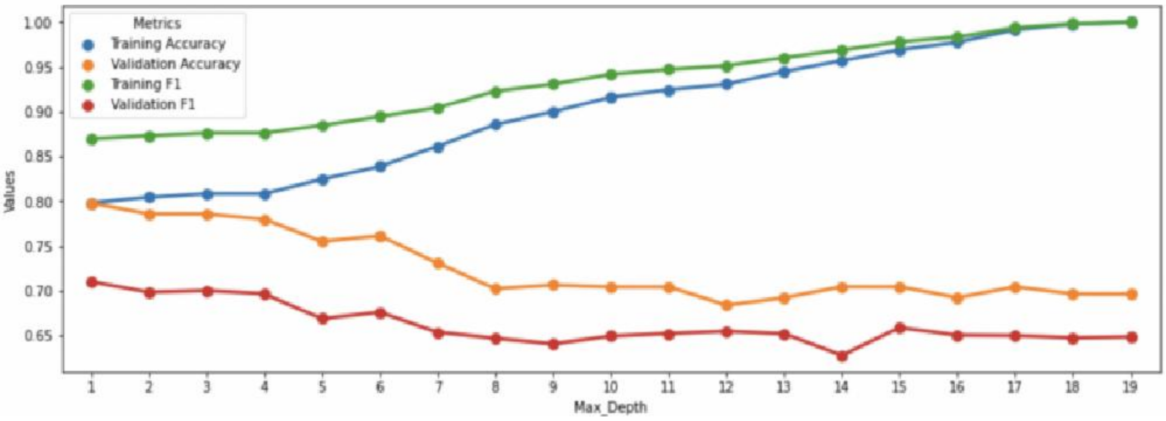

Figure 1: Training and evaluation statistics of decision tree.

From the image, the “Max_Depth” variable is tuned as a hyperparameter, the training accuracy increases as the tree goes deeper, but the validation accuracy decrease as a result of overfitting. Thus, the hyperparameter is tuned between a value between 1 and 20 and Max_depth =11 is the best output with a 79.2% validation mean accuracy with a high training accuracy (the result are shown in figure 2).

2.2. Random Forest

Random Forest is a way of utilizing Decision Tree. Instead of one big tree, it contains multiple Decision Trees and therefore the category of its output is determined by the mode of the category output by individual trees. Suppose there are M observations and P features in training data set. Instead of trying all features for each decision node, only a subset ---a random sample of M not bigger than P for each tree, at each split is tried. Note that if M = P, the method is called bagging, M ≈\sqrtP is for classification, and M ≈ P/3 is for regression.

Different from decision tree algorithm, random forest algorithm has many advantages over decision tree.

(1) Reducing overfitting. The decision tree is using all the features and samples, so overfitting is more likely to occur. On the other hand, Random Forest is multiple decision trees constructed by using the partial features of some samples (have taken back the sampling), so that features and data are reduced on a single decision tree, therefore preventing overfitting.

(2) Reducing the Impact of Outliers. Since only a selected part of data is used to build a tree, even if some trees are inaccurate due to outliers, the prediction results are still obtained by the majority of the trees, so the impact of outliers are reduced effectively.

The training data was used for generating a random forest algorithm. Three hyperparameters were found, and they are tuned by testing rough values. These three hyperparameters are “N_estimators”, “max_depth”, and “min_sample_leaf”. “N_estimators” indicates the number of trees in the forest. “max_depth” indicates the depth of each of the trees, it also determines the number of randomly selected features for each decision tree. And “min_sample_leaf” indicates the number of the branches per layer. The values tested for each hyperparameters are N_estimators=500,200,100,50,25,20, max_depth=2,3,4,5,6,7,8,9,10, min_samples_leaf =3,5,7,10,20,30

Final best values of each hyperparameters best number of estimator=500, best max depth=10, best min samples of leaves=3

2.3. Logistic Regression

Regression is an algorithm aiming to find a mapping function f from training data and to find a continuous output value y. y is a continuous quantity; therefore it can be utilized in projects like pricing optimization, sales forecasting, and rating forecasting. Regression predictions can be evaluated using the mean squared error. In some cases, a classification problem can be converted to a regression problem. The conversion can be done by calculating the probability for each category. Logistic Regression have the following advantages over Decision Tree and Random Forest models. The training speed is fast. When classifying, the computation amount is only related to the number of features. It is simple and easy to understand, and the model is very interpretable. The influence of different features on the final result can be noticed from the weight of features. It is suitable for binary classification problem as it does not need to scale the input features. Only the eigenvalues of each dimension need to be stored, So the memory footprint is small. There's a couple of modules that make it easier to import, fit, and predict. Finally, the authors start to calculate the accuracy at different thresholds and plot the picture. As the data set is big having 10,000 pieces, the problem of overfitting is in consideration. Eventually, data cleaning is utilized and less important features are selected like “education” and “contracts” and removed out of the model.

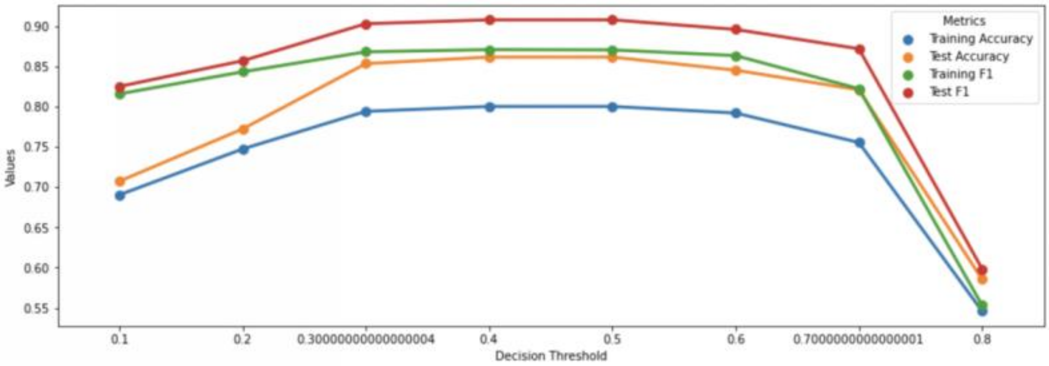

Figure 2: Training and evaluation statistics of logistic regression.

According to graph, when the value of the hyperparameter threshold is 0.4, the accuracy is the highest. Thus, the value is used and the final accuracy is determined. 86.4% test accuracy is achieved, and it is the highest among three example models.

3. Results

Table 1: Overall results of three methods on the test subset.

method	Accuracy	F1-score
Decision tree	79.2%	0.869
Random forest	84.5%	0.842
Logistic regression	86.42	0.910

According to Table 1, the authors can clearly see that logistic regression has the best results according to the test accuracy and F1-score. The algorithm of decision tree needs to use all the features and samples. In addition, with the increase of the tree, the training depth increases, which is very easy to lead to overfitting, resulting in the reduction of verification accuracy. Although its hyperparameters are tuned between 1 and 20, its training accuracy is still inferior to random Forest and logistic regression.

Random forest, however, is an advanced version of decision tree, which contains multiple decision trees and only uses part of the features of the sample, so that the features and data are reduced in a decision tree, so as to prevent over-fitting to a certain extent.

Logistic regression is also good at excluding insignificant features from the model to prevent overfitting, and the test accuracy is about the same as random forest. In terms of performance, the training speed of logistic regression is fast, and the amount of computation is related to the number of specialties. You can clearly see that logistic regression has the best results.

The algorithm of decision tree needs to use all the features and samples. In addition, with the increase of the tree, the training depth increases, which is very easy to lead to overfitting, resulting in the reduction of verification accuracy. Although its hyperparameters are tuned between 1 and 20, its training accuracy is still inferior to random Forest and logistic regression.

Random forest, however, is an advanced version of decision tree, which contains multiple decision trees and only uses part of the features of the sample, so that the features and data are reduced in a decision tree, so as to prevent over-fitting to a certain extent.

Logistic regression is also good at excluding insignificant features from the model to prevent overfitting, and the test accuracy is about the same as random forest. In terms of performance, the training speed of logistic regression is fast, and the amount of computation is related to the number of specialties. It is suitable for dichotomous classification problems because it does not need to scale the input features and only needs to store the eigenvalues of each dimension, so the memory footprint is small. Finally, because logistic regression is the most straightforward model among the three models, the model is highly interpretable. The influence of different features on the final result can be seen from the weight of features.

To sum up, logistic regression has the most powerful performance in loan forecasting analysis. It is suitable for dichotomous classification problems because it does not need to scale the input features and only needs to store the eigenvalues of each dimension, so the memory footprint is small. Finally, because logistic regression is the most straightforward model among the three models, the model is highly interpretable. The influence of different features on the final result can be seen from the weight of features.

4. Conclusion

Among the three methods ----Decision Tree, Random Forest Classifier, and Logistic Regression, Logistic Regression turns out to have the most accurate testing accuracy of 86.2%. However, it doesn’t mean that the model always performs best in all loan prediction situations. The result might be different among banks in different countries or time periods as the data for these banks have different IQRs or Standard Deviations which affect the best model. In our method, hyperparameters play a big role of a model. Hyperparameters that better fit the model help generate more accurate result. In Decision Tree, the maximum depth of the tree is treated as a hyperparameter and is tuned by expirments. In Random Forest Classifier, the number of trees, the maximum depth of the trees, and the maximum number of branches of a node are tured also by expirments. As for Logistic Regression, the value of threshold is tuned by a trend graph. In the future, bank loan always holds a strong place of the world’s economics and money cycle. To update the prediction model in the future to make it more accurate, more kinds of model such as KNN or Neural Network can be applied and categories like gender, race, and religion can be added as a value in the dataset and education, housing, marrige can be specified and converted into numerical values. Unnecessary values like day, p days can be ignored to generate a more accurate model that can contribute to the world economy in a positive way.

References

[1] Li Yujia, & Lu Jun. (2007). Risk Premium, Expected loss and forecast loan loss Provision. Contemporary Finance and Economics (12), 7.

[2] Jiang Zuobin, Xie Shuangqin, & Zhang Huan. (2010). Application of Arima model in the prediction of bank loan scale. Finance and Economics (7), 3.

[3] Chen Xulan, Han Suwan, & Pang Jianhua. (2021). Enterprise loan default risk prediction based on machine learning method. Modeling and Simulation, 010(003), P.890-897.

[4] Li Yunmeng, & Qian Xin. (2011). Long-term consumer loan forecasting based on co-integration and arma model. Statistics and Decision (11), 3.

[5] Xiang Weixing, Li Hongjin, & Bai Dafan. (2007). Bank Credit Expansion and Real Estate Bubble: Lessons from the United States, Japan, and East Asian Countries and Regions. International Studies of Finance (3), 7.

[6] Goodhart, C. , & Hofmann, B. . (2004). Monetary transmission in simple backward-looking models: the is puzzle.

[7] Goyal, A., & Kaur, R. (2016). A survey on ensemble model for loan prediction. International Journal of Engineering Trends and Applications (IJETA), 3(1), 32-37.

[8] Arun, K., Ishan, G., & Sanmeet, K. (2016). Loan approval prediction based on machine learning approach. IOSR J. Comput. Eng, 18(3), 18-21.

[9] Gupta, A., Pant, V., Kumar, S., & Bansal, P. K. (2020, December). Bank Loan Prediction System using Machine Learning. In 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 423-426). IEEE.

[10] Ratadiya, P., Asawa, K., & Nikhal, O. (2020). A decentralized aggregation mechanism for training deep learning models using smart contract system for bank loan prediction. arXiv preprint arXiv:2011.10981.