# Research on loan default prediction based on logistic regression, randomforest, xgboost and adaboost

*Jinchen* Lin
Guangdong University of Technology, Guangzhou, 510006, China

**Abstract.** Lenders often experience loan defaults, resulting in huge losses to lenders. Lenders are required to conduct a credit assessment of borrowers before making loans. Machine learning plays an essential role in loan credit analysis. This study analyzes the application of machine learning in loan credit analysis through a dataset of borrowers from Kaggle and looks for an excellent algorithm.This study use Logistic Regression, randomforest, XGBoost and AdaBoost to fit the dateset and compare their accuracy in prediction.In terms of results, XGBoost performed well while logistic regression performed poorly. For banks or lending institutions, using Gradient Boosting Decision Tree like XGBoost to predict loan default can increase profit.

## 1 Introduction

In the field of finance, credit risks are a common phenomenon in relation to mortgages, credit cards, and other types of loans. There is always the possibility that the borrower won't repay the loan in full. Risk assessment in connection with a loan application is the main issue for lending institutions' survival in a cutthroat market and for profitability. Lending institutions receive numerous loan applications from consumers each day, but not all of them are granted. These institutions use a variety of techniques to assess the applicant's information in order to make the best choice. Despite this, many people fail to make yearly loan payments. Lenders must therefore deal with this enormous loss [1].

Artificial intelligence technology can achieve in-depth mining and analysis of big data to address financial risks and challenges brought by financial technology. Compared to the shortcomings of traditional expert ratings, machine learning models for predicting bank credit loan defaults perform better [2]. Artificial intelligence can utilize big data and machine learning technologies to analyse borrowers' personal information, credit history, and other related data, helping banks and other financial institutions assess credit risks and make more accurate loan decisions. Artificial intelligence, as an emerging technology, will undoubtedly become a great driving force for the development of the financial industry. It will reduce the customer default rate of bank loans and make the bank's capital flow normal

In the past, banks and other financial institutions often used artificial analysis to determine the credit of customers. Based on previous data, artificial credit analysis is an inefficient and time-consuming method. It cannot process large amounts of data like machine learning. At the same time, the accuracy of Artificial credit analysis is much lower than that of machine learning. Machine learning can develop more automated and reliable risk assessment procedures to save lenders more costs [3]. Machine learning in credit analysis enables lenders to predict a borrower's credit more accurately. By analyzing large datasets and identifying borrowers' characteristics, machine learning algorithms can generate more precise credit scores and reduce the risk of loan defaults.

Many scholars use machine learning to predict loan defaults. Kumar et al. examined the precision of customers' loan eligibility with decision tree, Random Forest, support vector machine, k-nearest neighbour, and decision tree using adaboost technology[4]. On the basis of the data set, Singh et al. presented three machine learning (ML) algorithms: XGBoost, Random Forest, and Decision Tree. ML approaches are applied to the loans data for a borrower from the public bank sector. Seven performance indicators are employed to assess performance, and the outcomes are contrasted. The highest accuracy is 91.7% for Random Forest. Additionally, it has high rates of sensitivity and specificity. It performs better than classifiers using a single base [5].

There is still great room for improvement in machine learning for predicting loan defaults. In this paper, Logistic Regression, RandomForest, XGBoost, AdaBoost will be used to predict default or not according to the borrowers' data from Kaggle. The model of XGBoost maybe has a highest accuracy. At the same time, the selection of data characteristics also affects the accuracy of the prediction model. Without choosing characteristics, it would be challenging to examine the loan data because it has many characteristics [6]. For different data characteristics, the same algorithm may have different performance. This study is a supplement to loan default prediction, hoping to find a better method. The research on machine learning in loan default will drive the development of the loan industry. This not only reduces the losses of

Corresponding author: 3121004707@mail2.gdut.edu.cn

loan institutions, but also promotes more stable economic development.

## 2 Data and Method

### 2.1 Data

This study uses the dataset about Loan default from Kaggle. After checking, the validity of this dataset is high. It serves as the base for this experiment and can be utilized to train the loan default model.

#### *2.1.1 Sample and Features*

The dataset contains 32,581 records. It includes the personal information of the borrower and the purpose and amount of the loan. The dataset consists of the features as listed in table 1. Amont these features, loan_Status is the actual result. It indicates whether the borrower has defaulted, 1 means the borrower has not repaid the loan, and 0 means the opposite. The accuracy of the model can be obtained by combining this data with the results of training data of the classification model.

**Table 1** Features and Description

| Features | Type |
|---|---|
| Age | int |
| person_income Annual Income | int |
| Home ownership | object |
| Employment length (in years) | float |
| Loan intent | object |
| Loan grade | object |
| Loan amount | int |
| Interest rate | float |
| Loan status (0 is non default 1 is default) | int |
| Percent income | float |
| Historical default | object |
| Credit history leng | int |

#### *2.1.2 Data Cleansing*

To improve the accuracy and availability of data, the dataset needs to undergo data cleaning. Missing data and outlier will lead to inaccurate prediction of the model. The existence of outlier in the dataset will greatly affect the statistical characteristics of the data, such as mean, variance, correlation, etc. This may cause the predicted results of the machine learning model to deviate from the actual situation.

In this dataset,there are a total of 3116 missing values. For these data, measures were taken to delete them in this study. At the same time, outlier of numerical data are also deleted. For example, the data showing that the working age of borrowers is greater than their age is clearly abnormal. Or borrowers over 60 years old are clearly extremely unreasonable.

#### *2.1.3 Data Distributions*

Figure 1 shows that 20.5% of borrowers are in default. The ratio of default to non default is about 1:4. Through comparing the differences in features between default preson and non default person, it's clear that there is a certain regularity in the characteristics of these two groups, which is the basis of using the algorithm to predict the result of default. For example, the Figure 2 shows that the Loan proportion income in the default data group is significantly lower than that in the other group of data. It reflects that the Loan proportion income to income can serve as a basis for determining whether borrowers will default.

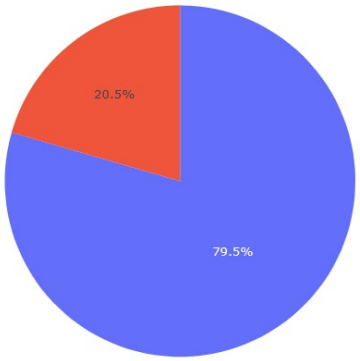


**Fig. 1.** Proportion of Loan_status

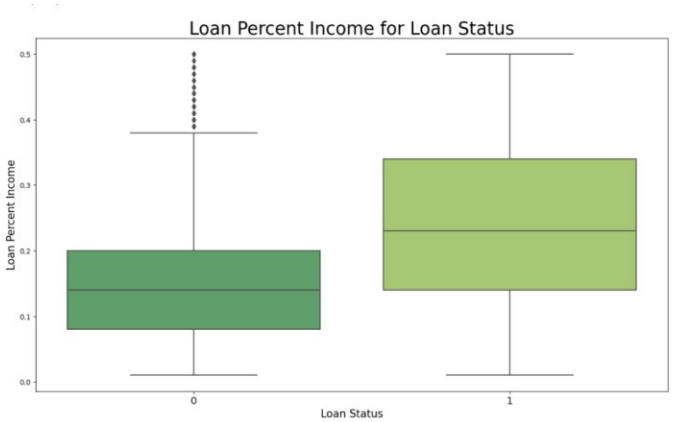


**Fig. 2.** Loan Percent Income

### 2.2 Method

In this paper, Logistic Regression, randomforest, XGBoost and AdaBoost are used to predict the loan default.These algorithms are common methods for binary classification problems. Logistic regression belongs to supervised learning in machine learning. It's the classic model. RandomForest, XGBoost and AdaBoost are also frequently used in financial risk analysis.

#### *2.2.1 Logistic Regression*

The statistical technique of logistic regression is used in data analysis and machine learning to forecast the likelihood of an outcome based on one or more input features. It is nonlinear classification, which is also suitable for the problem of qualitative indicators in dependent variables. Moreover, the Logistic discriminant function establishment method — — maximum likelihood estimation method has good statistical characteristics [7]. Finding the ideal weights that maximize the likelihood of predicting the right result for each observation in the dataset is the core objective of logistic regression. This is achieved through an optimization algorithm, such as gradient descent. Once the optimal weights are found, they can be applied to forecast the likelihood of a result for fresh data points.

#### *2.2.2 RandomForest*

A supervised learning algorithm called Random Forest is applied for both classification and regression tasks [8]. A randomly selected subset of the training data and features is used to create multiple decision trees, one for each tree. The output of each individual tree is combined to produce the result. RandomForest is well known for its capacity to handle high-dimensional datasets with numerous features and to avoid overfitting. They are widely used in many industries, including finance, marketing, and healthcare, for feature selection and predictive analysis.

#### *2.2.3 XGBoost*

Extreme Gradient Boosting abbreviated as XGBoost, is a gradient boosting framework that is effectively implemented as an open-source software library that is scalable, portable, and effective. Tianqi Chen invented this algorithm in 2016, and it has been demonstrated in the literature that its model has a low computational complexity, a quick processing time, and a high degree of accuracy [9]. It utilizes parallel and distributed computing, cache-aware algorithms, and out-of-core computing for large datasets. It is also very flexible, supporting various objective functions and customizable evaluation metrics. Moreover, XGBoost provides support for missing values, interpretable models, and built-in cross-validation.

#### *2.2.4 AdaBoost*

AdaBoost is short for adaptive boosting. A number of "weak" classifiers are combined in this machine learning algorithm to produce a strong classifier. When attempting to predict a binary outcome in classification problems, it is frequently used. While most ensemble learning algorithms construct more and more complex classifiers to improve prediction accuracy,Ada Boost seeks to combine the simplest weak classifiers that are slightly better than random guesses to get strong classifiers [10]. AdaBoost is a powerful and widely used algorithm that can achieve high accuracy in many different types of classification problems.

## 3 Results and discussion

### 3.1 Data Split

Aimed to accurately evaluate and validate the model, the data segmentation is mainly divided into the training set and the testing set. In this study, the ratio of the training set to the test set was 7:3.

A training set is a dataset used to train a model. Through continuous iterative training, the parameters of the model are continuously adjusted, resulting in a relatively good model.A test set is a set of datasets used to test the final effect of a model, and the data in the test set cannot overlap with previous training and validation data.

### 3.2 Train Result

As listed in table 2, XGBoost and AdaBoost have the best performance. RandomForest is slightly inferior to these two algorithms. Logistic regression performed much worse than other algorithms. XXX It can be seen from the values of f1 and recall that when the actual result is default, the performance of logistic regression model is poor and the accuracy rate is far less than 25%.

**Table 2.** Accurancy of Train Data

| Model | accurracy | precision_0 | precision_1 | recall_0 | recall_1 | f1_0 | f1_1 |
|---|---|---|---|---|---|---|---|
| Logistical Regression | 81.06% | 0.815 | 0.705 | 0.984 | 0.143 | 0.891 | 0.238 |
| Randomforest | 93.66% | 0.927 | 0.988 | 0.997 | 0.701 | 0.961 | 0.820 |
| XGBoost | 96.07% | 0.954 | 0.989 | 0.997 | 0.818 | 0.975 | 0.895 |
| AdaBoost | 95.11% | 0.942 | 0.996 | 0.999 | 0.766 | 0.970 | 0.866 |

### 3.3 Test Result

As listed in table 3, the average accuracy of the test results decreased comparing with the training results. But XGBoost's behave is still the best. When the actual result is non default, the accuracy of these three algorithms is similar. When the actual result is default, the prediction accuracy of XGBoost is significantly higher than that of RandomForest and AdaBoost.

**Table 3.** Accuracy of Test Data

| Model | accurracy | precision_0 | precision_1 | recall_0 | recall_1 | f1_0 | f1_1 |
|---|---|---|---|---|---|---|---|
| Logistical Regression | 81.20% | 0.817 | 0.678 | 0.983 | 0.135 | 0.893 | 0.225 |
| Randomforest | 92.72% | 0.922 | 0.957 | 0.992 | 0.669 | 0.956 | 0.788 |
| XGBoost | 93.26% | 0.932 | 0.933 | 0.987 | 0.717 | 0.958 | 0.811 |
| AdaBoost | 92.37% | 0.927 | 0.901 | 0.980 | 0.699 | 0.953 | 0.787 |

### 3.4 ROC Curve

The ROC curve can be used to choose the best classification threshold, balance false-positive and false-negative rates, and visualize the trade-off between sensitivity and specificity in classification problems. It is also helpful for comparing the performance of various models to find the one that is most appropriate for the current issue.

Figure 3,4,5,6 is the ROC curve of Linear Regression, RandomForest, XGBoost, AdaBoost.

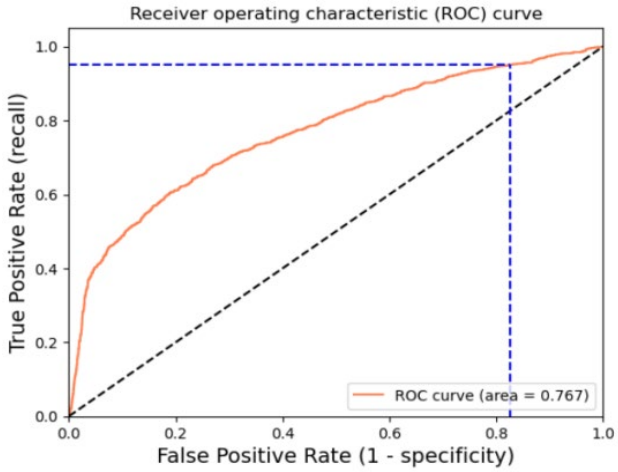


**Fig. 3.** Linear Regression

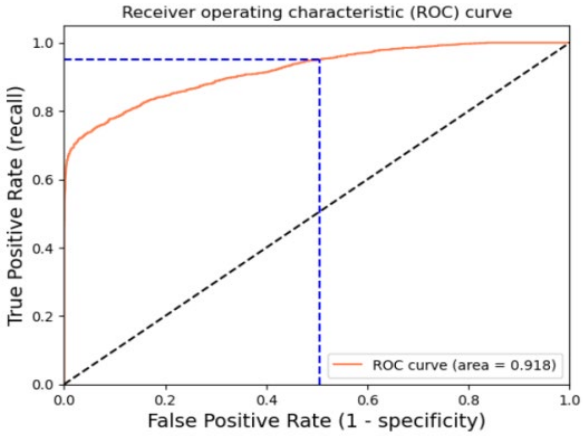


**Fig. 4.** Random Forest

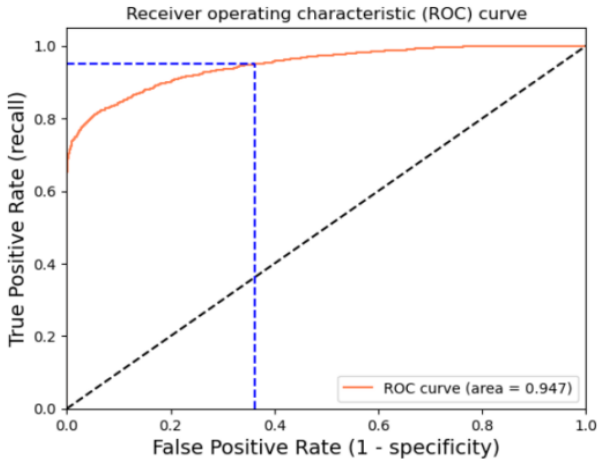


**Fig. 5.** XGBoost

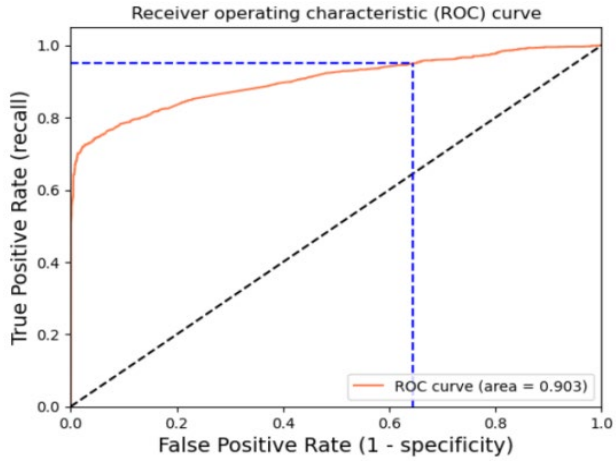


**Fig. 6** AdaBoost

### 3.5 Discusion

From the perspective of prediction accuracy(Table 2-3), the accuracy of test results of XGBoost, AdaBoost and RandomForest exceeded 92%. Among them, XGBoost has the highest prediction accuracy among the four algorithms. But the accuracy of Logistic Regression is only 81.20%. Its accuracy is far lower than the other three algorithms.

Comparing the f1 scores of the four models (Table 2-3), XGBoost has the highest score. This indicates that the XGBoost model has good accuracy and recall in classification. And the accuracy of Logistic Regression is the lowest.

It can be seen from ROC curve(Figure 3-6) that the order of misjudgment rate of the four algorithms is: XGBoost<RandomForest<AdaBoost<Logistic Regression. XGBoost has excellent sample recognition ability.

The data characteristics of this study are different from those of other similar studies. Although similar to the results obtained by other scholars, the algorithm accuracy of this study is higher [4, 5, 11,12]. It can be seen that the data characteristics in this research are beneficial to prediction. The model of Logistic Regression cannot provide more accurate results. It is not suitable for predicting loan credit default. On the contrary, the model of XGBoost can predict borrowers' default status well, providing lenders with more accurate predictions.

## 4 Conclusion

In this study, four machine learning algorithms are used, including Logistic Regression, Random Forest, XGBoost, and AdaBoost. After training and evaluation, The results of the study showed that XGBoost has higher prediction accuracy. XGBoost is a new algorithm in recent years. It is an improvement on Gradient Boosting. Because it uses feature-based parallel processing, it can process massive amounts of data with high accuracy. And because the sample is too complex, the logistic regression model has overfitting phenomenon when processing the training data. Logistic regression performed the worst.

This study verifies the advantage of Gradient Boosting framework for loan default prediction through XGBoost model. Therefore, according to the characteristics of loan default prediction, Gradient Boosting algorithm can be improved to get a higher accuracy rate.

However, this study has several limitations. To maximize loan profit is the primary goal of loan default prediction. But the final payoff predicted by different algorithms was not calculated in this study. The balance between accuracy and benefit should be realized.

In future research. more data or other characteristic can be used to optimize the model. The accuracy of the machine learning model will be impacted by the choice of various features. Supplement the loan interest rate to calculate the final payoff on the loan. Alternatively, other machine learning algorithms can be explored to continuously optimize the predictive performance of the model.

## References

1. Ashish Pandit, "Data Mining On Loan Approved Dataset For Predicting Defaulters", A Project Report Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science Computer Science (2016)
2. ZHANG Liying, YANG Ruoji, "Research on the application of personal loan default prediction model based on machine learning[J]",Financial Supervision Research,**6**,126 (2022)
3. Mingrui Chen, Yann Dautais, LiGuo Huang, and Jidong Ge. “Data driven credit risk management process: a machine learning approach”,In Proceedings of the 2017 International Conference on Software and System Process, Association for Computing Machinery, New York, NY, USA (2017)
4. C. Naveen Kumar, D. Keerthana, M. Kavitha and M. Kalyani, "Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore,1007-1012 (2022)
5. Vishal Singh, Ayushman Yadav and Rajat Awasthi, "Prediction of Modernized Loan Approval System Based on Machine Learning Approach", 2021 International Conference on Intelligent Technologies (CONIT) (2021)
6. Pratheeksha Hegde N, Deepa, Chinmai Shetty, Rashmi N, Dhananjaya B, Prathvakshini, "Predictive Analysis of Loan Data using Machine Learning", 2022 International Conference on Artificial Intelligence and Data Engineering (AIDE), 272-276 (2022)
7. Liu Yiliang, Yin Kunlong, Liu Bin. “Application of logistic regression and Artificial neural network model to spatial prediction of landslide disaster”. Hydrogeology engineering geology, **5** , 37 (2010)
8. ]Lin Zhu, Dafeng Qiu, Daji Ergu, Cai Ying, Kuiyi Liu,A study on predicting loan default based on the random forest algorithm,Procedia Computer Science,**162** (2019)
9. Chen T, Guestrin C.XGBoost:a scalable tree boosting system[C]//ACM SIGKDD International Parallel&Distributed Processing Symposium (2018)
10. Ying Cao,Qiguang Miao,Jiachen Liu,Lin Gao.Research progress and prospect of AdaBoost algorithm [J]. Acta Automatica Sinica, 6, 39 (2013)
11. S.K.Shaheen and E.ElFakharany, "Predictive analytics for loan default in banking sector using machine learning techniques," 2018 28th International Conference on Computer Theory and Applications (ICCTA), Alexandria, Egypt, 66-71 (2018)
12. Lifang Zhang, Jianzhou Wang, Zhenkun Liu,“What should lenders be more concerned about? Developing a profit-driven loan default prediction model”,Expert Systems with Applications, **213**, Part B (2023)