ISSN 2616-5775 Vol. 6, Issue 9: 32-37, DOI: 10.25236/AJCIS.2023.060905

Analysis and Comparison of Loan Default Prediction Models Based on XGBoost and LightGBM Algorithm

Xia Dong[1,a], Wei Xue[1,b], Jing Chen[1,c,*]

[1]*Shengbao Institute of Financial Technology, Geely University, Chengdu, Sichuan, 610000, China*
[a]*dongxia1114@foxmail.com*, [b]*xuewei@bgu.edu.cn*, [c]*jingc0278@gmail.com*
[*]*Corresponding author*

Abstract: *Based on the default loans caused by information asymmetry and uncontrollable factors at this stage, this paper will use two algorithm models, XGBoost and LightGBM, to extract and screen the relevant information of the applicant and build a loan default prediction model to predict the default situation of the loan. And the two different models were compared and evaluated to provide data reference for financial institutions to select and build the loan default prediction model to reduce their risks and bank losses to a certain extent.*

Keywords: *XGBoost algorithm; LightGBM algorithm; Loan default prediction*

1. Introduction

With the continuous development of society, people generally accept loan consumption as a new lifestyle, and the loan business profits by recovering the principal and interest through lending and deducting costs. Because the loan is out of the bank's control, there is a greater risk that the principal and interest cannot be recovered on time. According to the RMB credit balance data of financial institutions released by the People's Bank of China, it can be concluded that the amount of various loans in China has been increasing year by year recently. In addition, the latest data shows that the total amount of RMB loans of financial institutions in January 2023 was 219,745,514 million yuan. This data is enough to illustrate the importance of the loan business in commercial banks and affirms the importance of loan default prediction.

Today's research on the Internet financial credit industry is mainly about studying loan default risk control, including exploring the default behavior of borrowers and building a loan default prediction model [1]. Domestically, Fang Kuangnan et al. and others proposed a credit scoring model based on semi-supervised generalized additive Logistic regression to evaluate the default risk of personal credit loans. With algorithm technology's continuous update and development, gradient-boosting tree frontier algorithms such as XGBoost and LightGBM have great advantages in the training set's utilization, efficiency, and accuracy. Zhou Rongxi et al. built a credit default prediction model based on the XGBoost algorithm. Ma Xiaojun et al. believed that the results of the LightGBM algorithm have an excellent fitting effect on the actual situation through empirical research. At the same time, the LightGBM algorithm effectively alleviates the disadvantages of traditional machine learning algorithms, such as overfitting, long learning time, and strong subjectivity of parameter settings, having high accuracy of default prediction [2].

According to the above reviews, there are some experiences for reference. Still, there are also some problems: many dependent variables affect loan prediction, including their attributes and unpredictable and uncontrollable factors. For example, the calculation speed of the XGBoost algorithm is too slow in the actual operation, especially in the Bayesian parameter tuning process and the model training process. Then, the feature redundancy in the financial data set is high, and there are many missing values, the model's classification accuracy is weak, and the model generalization ability is insufficient [3].

Therefore, the main work of this paper is to experiment with the existing data. After the data is processed, two algorithm models, XGBoost and LightGBM, are constructed to predict the loan default situation. In addition, the model scores are compared through the AUC value to provide commercial banks with an algorithm to get the factors that significantly impact default to reduce the risk of bank loan default.

Sichuan Xinwang Bank provides the data in this paper, and all data are masked.

ISSN 2616-5775 Vol. 6, Issue 9: 32-37, DOI: 10.25236/AJCIS.2023.060905

2. Research Design

2.1 Sample and Variable Selection

The data set in this paper has 2.15 million data, with a total of 204 columns of variable information.

Explanatory variables: jieju_dubil_amt (IOU) IOU amount, jieju_mbank_contri_amt (IOU) capital contribution amount of the bank, jieju_obank_contri_amt (IOU) capital contribution amount of cooperative bank, jieju_dubil_bal (IOU) balance of IOU, jieju_mbank_prin (IOU) principal balance of the bank, jieju_dubil_bal (IOU) Balance, shouxin_aval_limit (credit) available amount, kehu_age (customer) age, kehu_yr_incom_amt (customer) annual income amount, etc.

Explained variable: whether the loan is in default is regarded as the explained variable.

2.2 Model Algorithm Introduction

2.2.1 XGBoost Algorithm

In 2016, Chen Tianqi formally proposed it in the paper "XGBoost: A Scalable Tree Boosting System."

XGBoost (eXtreme GradientBoosting) is a gradient-boosting algorithm. It is an efficient and flexible machine-learning algorithm that can be used for classification and regression. XGBoost creates a more robust prediction model by combining many weak prediction models. It optimizes weighted residuals and introduces new decision trees in each iteration. In addition, XGBoost also has regularization techniques to prevent overfitting and improve generalization. Due to its high precision, high efficiency, and scalability, XGBoost has become one of the most popular machine learning algorithms in numerous data science competitions and practical applications. Its prediction model is:

$$\hat{y}_i = \sum_{k=1}^{K} f_{k(x_i)} \tag{1}$$

The loss function is:

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

Of which, K is the total number of trees, f_k represents the Kth tree, $\hat{y}_i$ represents the prediction result of the sample x_i, and $l(y_i, \hat{y}_i)$ represents the training error of the sample x_i, $\Omega(f_k)$ represents the regular term of the kth tree. Obj, also known as the structure score, is a function like the Gini coefficient to evaluate the tree structure. The smaller the score, the better the tree structure [4].

2.2.2 LightGBM Algorithm

LightGBM (Light Gradient Boosting Machine) is a decision tree framework based on a gradient boosting algorithm. Unlike traditional gradient boosting algorithms, LightGBM uses a technique called "histogram-based decision tree learning," which divides the dataset into discrete bins and builds a decision tree within each bin. This reduces memory usage and speeds up training. In addition, LightGBM also introduces some new features, such as a Leaf-wise growth strategy, Gradient-based One-SideSampling (GOSS), etc., to improve the accuracy and efficiency of the model. LightGBM has become one of the most popular algorithms in machine learning and has performed well in various competitions and applications.

2.2.3 AUC Value Test and ROC Curve

AUC is the area under the ROC curve, which can be directly calculated. The area is the sum of the small trapezoidal areas (curves), and the calculation accuracy is related to the threshold accuracy.

The ROC curve is composed of two variables, TPR and FPR. This combination uses FPR to TPR, that is, costs to benefits.

The x-axis is the false positive rate (FPR): the proportion of incorrect classifier prediction among all negative samples.

ISSN 2616-5775 Vol. 6, Issue 9: 32-37, DOI: 10.25236/AJCIS.2023.060905

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

The y-axis is the true positive rate (TPR): the proportion of correct classifier prediction (equal to Recall) among all positive samples.

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

3. Empirical Results and Analysis

3.1 Data Preprocessing

The training set and test set are merged to increase the sample size. The data set in this paper is 2.15 million, with a total of 204 columns of variable information. The features are manually selected based on the correlation between the independent and dependent variables. The following feature values are selected: jieju_subj_num (IOU) principal account number, jieju_dubil_amt (IOU) amount, jieju_mbank_contri_amt (IOU) capital contribution amount of the bank, jieju_obank_contri_amt (IOU) capital contribution amount of the cooperative bank, jieju_dubil_bal (IOU) balance, jieju_mbank_prin (IOU) principal balance of the bank, jieju_co_bank_prin (IOU) principal balance of the cooperative bank, jieju_assetal_flow_trans_bal (IOU) balance of asset circulation, etc.

3.1.1 Missing Value Processing

After selecting the feature values, the missing value processing is performed on the data, visualizing the missing variables, as shown in Figure 1. Ultimately, this paper chooses to use missing values as variables to maximize the retention of all original data.

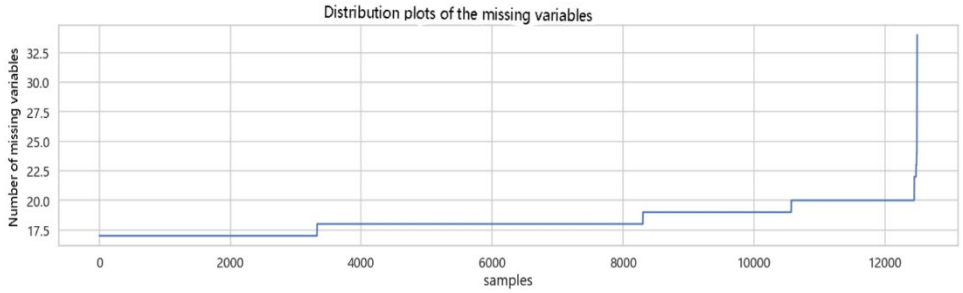

Figure 1: Data visualizing

First, the feature values with missing values accounting for more than 80% are deleted. Then the missing values of continuous variables are filled with the mean value, and the rows where the missing values of non-continuous variables are deleted.

3.1.2 Outlier processing

Due to this paper's large amount of data, the 3sigma criterion is applied. The probability of the numerical distribution in (μ—3σ, μ+3σ) is 0.9974, where σ represents the standard deviation in the normal distribution, μ represents the mean value, and x=μ is the image symmetry axis. First, it is assumed that a test data set contains only random errors. Then, the data were calculated and processed to obtain the standard deviation, and an interval was determined according to a certain probability. It is believed that any error exceeding this interval is not a random error but a gross error. The data containing gross errors should be removed [5]. As shown in Figure 2, it is subsequently determined whether to filter outliers based on the model training effect.

ISSN 2616-5775 Vol. 6, Issue 9: 32-37, DOI: 10.25236/AJCIS.2023.060905

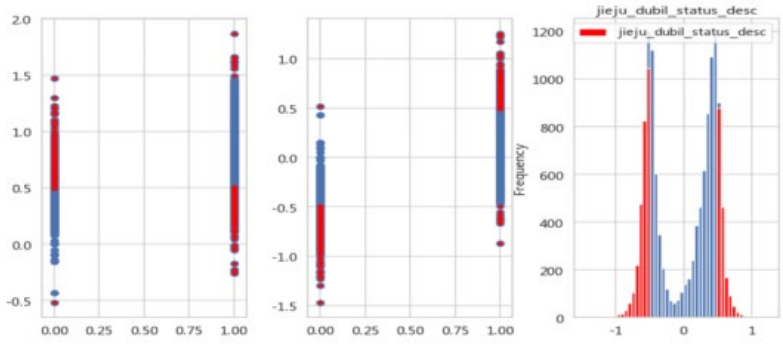

Figure 2: Data calculation and processing

3.2 Analysis of Feature Importance

XGBoost and LightGBM have their methods for calculating feature importance. The importance of features is related to the number of times the feature is used and the gain when using the feature. The more times the feature is used for splitting, the greater the gain brought by the splitting, and the more important the feature is. Based on this, this paper draws the important features that significantly impact loan defaults, as shown in Figure 3.

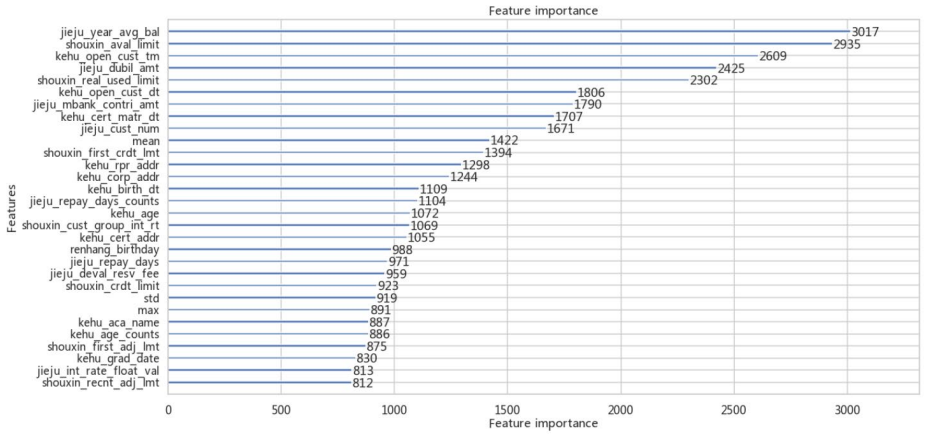

Figure 3: Feature importance

It can be concluded that the features that have a more significant impact on loan defaults are: (IOU) annual average daily balance, (IOU) customer opening time, (credit) available amount, (IOU) amount, (credit) used amount, etc.

Among them, (IOU) annual average daily balance: For individuals, the annual average daily balance means the level of the person's financial ability; the higher the balance, the better the individual's financial ability and the stronger the repayment ability, and vice versa.

(Credit) Available amount: The available amount will decrease with each consumption and will be restored correspondingly with each repayment of the cardholder. The amount level directly reflects the lender's credit amount and repayment ability.

(IOU) Amount of IOU: The amount reflects the size of the loan business, and the repayment risk is more significant for businesses with a larger amount.

(Credit Grant) Actual used amount: Banks can predict the property status and credit of the lender based on the used amount, which can avoid the risk of default to a greater extent.

According to the analysis, there are 82 features whose importance are 0 and 155 features whose importance is not 0. Then, after continued optimization and selection, citytype (city type), jieju_cust_num ((IOU) customer number), jieju_subj_num ((IOU) principal Subject number), and other 104 features are determined as the final number of features.

ISSN 2616-5775 Vol. 6, Issue 9: 32-37, DOI: 10.25236/AJCIS.2023.060905

3.3 Model Construction

3.3.1 Building XGBoost Model

When selecting XGBoost parameters, the improved grid search method is used to optimize the parameters of the number of sub-models, the maximum depth of the tree, and other critical parameters in XGBoost, as shown in Table 1.

Table 1: Parameter Setting

Parameter	Parameter Description	Parameter value
seeds	Random number seed	12
eta	By reducing the weight of each step, the robustness of the model can be improved	0.01
gamma	The value of the parameter is closely related to the loss function	0.1
min_child_weight	The sum of the minimum sample weights	1.1
max_depth	The maximum depth of the tree	5
lambda	Used to control the regularization part of XGBoost	10
subsample	Controls the proportion of random sampling for each tree	0.7
colsample_bytree	Used to control the proportion of the number of columns randomly sampled per tree (each column is a feature)	0.7
colsample_bylevel	Used to control each split of each level of the tree, the proportion of sampling the number of columns	0.7

The AUC value of the model evaluation is 0.9283, and the precision rate is 0.964, as shown in Figure 4.

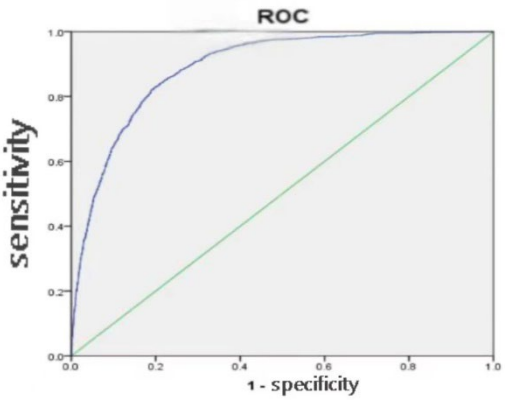

Figure 4: AUC of the XGBoost model

3.3.2 Building the LightGBM Model

First, the model's optimal parameters are calculated using the ten-fold cross-validation and grid search method. The results are shown in Table 2 so that the model can obtain the optimal classification accuracy parameters, and then the model is trained based on the LightGBM algorithm.

Table 2: Optimal parameters

Parameter	Parameter Description	Parameter value
max_depth	Limit the maximum depth of the tree model	8
min_child_weight	The minimum hessian sum on a leaf	0.1
num_leaves	The number of leaves on a tree	255
min_data_in_leaf	The minimum amount of data on a leaf	30
learning_rate	Control the learning progress of the model	0.005
feature_fraction	The proportion of randomly selected features in each iteration	0.8
bagging_fraction	Randomly select part of the data without resampling	0.8
bagging_freq	A non-zero value means performing k bagging	5

The AUC value of the model evaluation is 0.8174, and the precision rate is 0.856, as shown in Figure

ISSN 2616-5775 Vol. 6, Issue 9: 32-37, DOI: 10.25236/AJCIS.2023.060905

5.

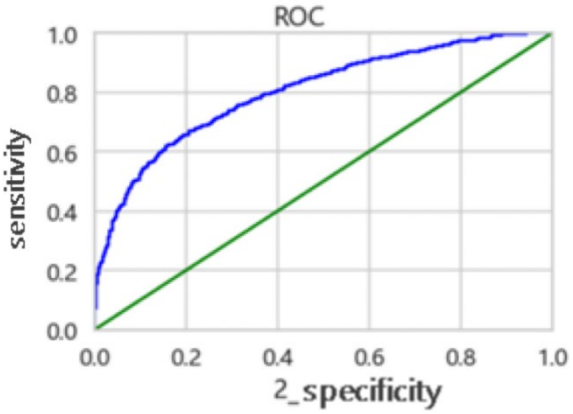

Figure 5: AUC of the LightGBM model

4. Conclusion

The practice has proved that both models have good learning and prediction ability. By comparing the two models, XGBoost and LightGBM, this paper finds that the XGBoost model has a higher AUC value and a higher precision rate regarding loan default prediction.

The selection of important feature values, such as jieju_year_avg_bal (IOU) annual average daily balance, shouxin_aval_limit (credit) available amount, kehu_age (customer) age, kehu_yr_incom_amt (customer) annual income amount, etc., provides the direction of credit evaluation for major financial institutions so that they can make a correct evaluation at the beginning, and minimize loan defaults.

References

[1] Tan Z, Zhang J, He Y, et al. Short-term load forecasting based on the integration of SVR and stacking [J]. IEEE Access, 2020, 8: 227719-227728.
[2] Xia Y, He L, Li Y, et al. Predicting loan default in peer-to-peer lending using narrative data [J]. Journal of Forecasting, 2020, 39(2): 260-280.
[3] Gao B, Balyan V. Construction of a financial default risk prediction model based on the LightGBM algorithm [J]. Journal of Intelligent Systems, 2022, 31(1): 767-779.
[4] Liu Y, Yang M, Wang Y, et al. Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China [J]. International Review of Financial Analysis, 2022, 79: 101971.
[5] Jin Y, Xiao Q, Jia H, et al. A novel detection and localization approach of open-circuit switch fault for the grid-connected modular multilevel converter [J]. IEEE Transactions on Industrial Electronics, 2022, 70(1): 112-124.