Credit Card Default Prediction based on Machine Learning Techniques

Zixuan Zhang*

Faculty of Economics and Businicess, The University of Hong Kong, Hong Kong, China

*Corresponding author: u3589687@connect.hku.hk

Abstract. In recent years, with the development of society and economy, credit cards have been popularized due to their low interest rate and easy payment. However, with the advent of the epidemic era, the unemployment rate has increased, making the probability of credit card defaults rising. The prediction of credit card default helps banks and financial institutions balance the risk and economic interests, contributes to the stable and healthy development of the financial industry, and plays an important role in bank credit control. Therefore, this paper addresses the credit card default prediction problem by using Random forest, Decision tree, LightGBM, XGBoost, Logistic regression, and Adaboost models to make predictions and compare the results. The outcomes demonstrate that LightGBM algorithm has the most outstanding prediction score, and its AUC value can reach 0.78 and recall rate reaches 0.95.

Keywords: Credit card default; SMOTE; machine learning models.

1. Introduction

Credit cards appeared in the United States in 1951 and are a common means of payment in the world today. They are bank-based proof of credit for individuals, and it offers customers a spending method that enable them to spend first and pay back later. Banks are able to make significant profits by conducting credit card business. Credit cards are usually based on different credit limits for different customers depending on their ability to repay. As the customer continues to overdraw the credit card amount, the amount available to the user will gradually become smaller, but the credit card limit will be restored when the customer repays his overdraft amount within a specified period of time. If the cardholder is unable to complete the repayment of the minimum amount within the time limit set by the bank, the status of the into account is considered to be expected and the credit card customer will pay a late payment fee (i.e., interest for the month due to untimely repayment).

Over the past few decades, the credit card industry has flourished in the United States, Europe and other countries, and the creation of credit cards has boosted people's desire to spend money and thus the economy. Credit card delinquency is a common and unavoidable problem for banks, which usually manage credit card delinquency collections for different types of customers with different statuses. As a customer's overdue time increases, the bank's collection efforts increase. Generally speaking, if a customer is overdue for less than 90 days, the bank will collect through emails and SMS, while once the customer has not repaid for more than 90 days, the bank will take the means of door-to-door collection or legal action. However, with the gradual slowdown of the economic trend in recent years, especially after the impact of the epidemic and the rise of unemployment, the credit default risk borne by banks has been increasing and the probability of large-scale credit card defaults is gradually increasing, which will bring serious losses to banks and financial institutions. Once this happens, the regular collection methods used by banks will no longer be effective Therefore, it is important for banks and financial institutions to evaluate the basic information of credit card holders and analyze it together with their spending behavior to predict potential defaulters in advance and control their credit limits. This can minimize the loss of the bank.

In recent years, many experts have proposed solutions to this problem and have identified the main factors influencing credit card customer defaults by examining credit card customer data. Steenackers et al. first used logistic regression to find that the significant factors affecting credit card customer default were age, place of residence, occupation, annual salary, and length of employment [1]. Edwaretal et al. found that low-income groups were more likely to default by regressing credit card

customers' income on factors such as debt [2]. Bhattacharyya et al. analyzed and predicted the credit card customer data of an international bank by using three models: LR, SVM and RF, and the results showed that the logistic regression model has better prediction results [3]. Butaru conducted a study by collecting credit card data from six different kinds of banks and using decision tree, Random Forest and Logistic Stipulation three machine learning models for customer default study, and the results showed that the best prediction models corresponding to them are different for different kinds of banks [4]. Yang Lei et al. adopted Transformer as an encoder in deep learning to self-code customer credit card data from a regional bank in Taiwan and predicted user default by a classification prediction model, and found that the prediction results of adopting a combination of Transformer and machine learning models outperformed the prediction results of a single machine learning model [5]. Qiao Cuiyi et al. conducted a default prediction study by using a logistic regression model and a decision tree model on credit card data of Hubei Postal Savings Bank customers, and discovered that the decision tree model's prediction impact was superior than the logistic regression model using model credibility assessment and performance evaluation as evaluation measures. [6]. Fan weiqiang et al. constructed a BP neural network for default prediction of credit card customers of a bank study, and found that abnormal data had a large impact on the algorithm of this model [7]. Wang Jiao et al. used SMOTE algorithm and ADASYN algorithm to oversample the original data set and built various credit card prediction models for the problem of unbalanced sample data, respectively, and the final results showed that logisticstic regression has better fitting ability [8].Mei Ruiting et al. conducted default by building Lasso-Logistic and random forest models for prediction and using F-score and prediction accuracy as evaluation metrics, the results showed that random forest outperformed Lasso-Logistic model [9]. Soui et al. proposed that the credit card risk assessment problem can be viewed as a genetic programming algorithm (GP) classification problem that maximizes the accuracy of the model [10].

2. Method

2.1 Logistic Regression

Logistic regression is a generalized linear regression, which is usually used in 01 variable classification problems, and the output is discrete values. Its probability distribution function is a sigmoid function.

The classification probability is found by the above formula.

$$p(x_i) = \frac{1}{1+e^{-w^T x+b}} \tag{1}$$

2.2 Random Forest

Random forest is a tree model-based algorithm. The method is based on bagging and tree model. Each tree is predicted to obtain a predicted value. By sampling the data with put-back and selecting only a random subset of the features into each tree, the variance of the random forest estimate is reduced, which in practice reduces the variance more significantly and reduces the possibility of over-fitting. The final classification results are obtained by averaging the predicted values for multiple trees.

2.3 Decision tree

A fundamental classification regression technique is the decision tree. This approach generates legible rules from the preprocessed data using induction techniques before making choices.

2.4 LightGBM

LightGBM is an integrated modeling framework whose principles mainly rely on gradient boosting decision trees. The gradient boosting decision tree belongs to one of the integrated learning methods of boosting, and its strategy of combining base learners becomes gradient boosting.

2.5 Adaboost

Adaboosting is defined as 'Adaptive Boosting'. It was put forward by Schapire in 1990, training a set of different classifiers (weak classifiers) and then gather them into a strong final classifier (strong classifier), which can achieve promising results.

2.6 XGBoost

For the binary classification problem, XGBoost can use probability loss function and category loss function, and the binary classification probability loss function of this method is:

$$L = \sum_1^m y^{(i)} logh\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) log\left(1 - h\left(x^{(i)}\right)\right) \quad (2)$$

Among them, since the focus of this paper is on the binary classification problem, only its binary loss function is introduced.

2.7 SMOTE

To address the problem of unbalanced dataset, Chawla et al. proposed the SMOTE algorithm to oversample the unbalanced dataset. This algorithm reduces the prediction errors due to classification imbalance by increasing the number of minority class samples to make all types of data in the dataset balanced. The method consists of three main steps, first calculating the Euclidean distance between two of the minority class samples to obtain the k-nearest neighbors, then determining the sampling multiplicity P according to the imbalance ratio, randomly selecting samples from the k-nearest neighbors, and finally obtaining the new data set by linear interpolation on the original data set.

3. Results and Discussion

The experimental data in this paper comes from the public platform Kaggle. The data contains a total of 30,000 credit card default data of a bank in Taiwan from April to September 2005, and the data independent variables include user repayment information, credit card limit, pre-payment record, personal characteristics information, gender, education level, age, marital status, etc. Each data is divided into two categories: defaulted and non-defaulted, of which 6,646 data are included in defaulted and 23,364 data are included in non-defaulted (Figure 1).

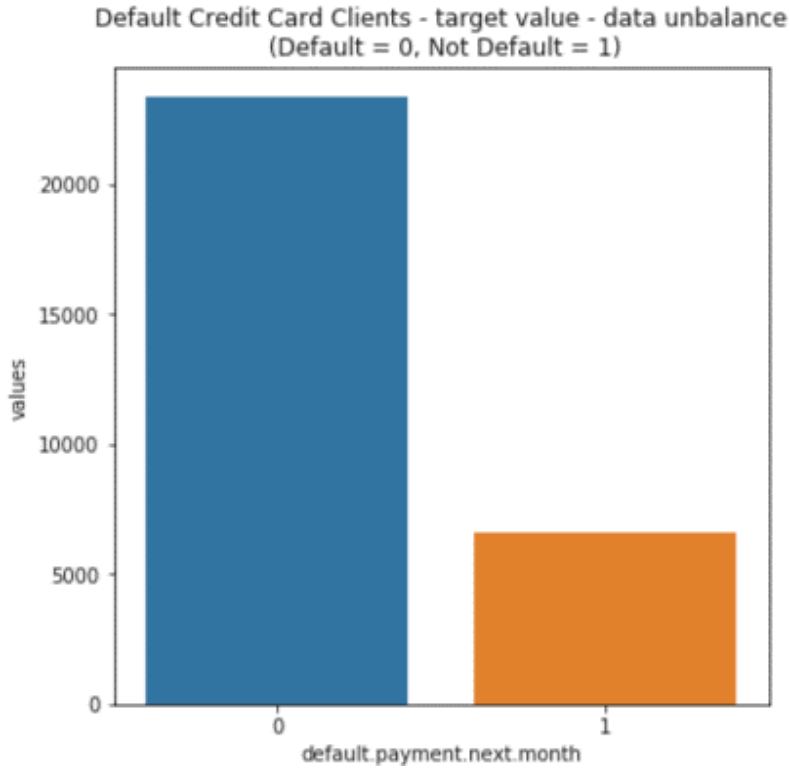

Figure 1. Default Ratio

The dataset was then classified using logistic regression, decision tree, random forest, XGBoost, LightGBM, SVM, and Adaboost algorithms and compared each result in terms of MSE, AUC, and R2 score. The table 1 shows the variable from the dataset.

Table 1. Variables

Attribute Name	Description	Type
LIMIT_BAL	Amount of the given credit(NT dollars)	numerical
SEX	Gender(1=male, 2=female)	categorical
EDUCATION	Education	categorical
MARRIAGE	Marital status	categorical
AGE	Age(year)	categorical
P_1 TO P_6	The payment status from April to September, 2005	numerical
BA_1 TO BA_6	Amount of bill statement from April to September, 2005	numerical
PA_1 TO PA_6	Amount paid from April to September,2005	numerical
Default or not	Whether it will default in the next month	categorical

The programming language of this paper is python3, the operating system is Windows, and the machine learning framework is sklearn.

3.1 Data Pre-processing and Partitioning

Before building the model, the original data needs to be checked because the data has a significant correlation with the data quality for the correct prediction rate of the model, and the process of data collection, storage and extraction all may have manual data errors, data attribute errors, missing data and many other problems, which will directly lead to unreliable prediction results. The results of descriptive statistics are shown in table 2.

Table 2. Descriptive statistics results

	mean	std	min	25%	50%	75%	max
LIMIT_BAL	167484.3	129747.7	10000	50000	140000	240000	1000000
SEX	1.603	0.489	1	1	2	2	2
EDUCATION	1.842	0.744	1	1	2	2	4
MARRIAGE	1.557	0.521	1	1	2	2	3
AGE	35.485	9.217	21	28	34	41	79
P_1	-0.016	1.123	-2	-1	0	0	8
P_2	-0.133	1.197	-2	-1	0	0	8
P_3	-0.166	1.196	-2	-1	0	0	8
P_4	-0.220	1.169	-2	-1	0	0	8
P_5	-0.266	1.133	-2	-1	0	0	8
P_6	-0.291	1.149	-2	-1	0	0	8
BA_1	51223.33	73635.86	-165580	3558.75	22381.5	67091	964511
BA_2	49179.08	71173.77	-69777	2984.75	21200	64006.25	983931
BA_3	47013.15	69349.39	-157264	2666.25	20088.5	60164.75	1664089
BA_4	43262.95	64332.86	-170000	2326.75	19052	54506	891586
BA_5	40311.4	60797.16	-81334	1763	18104.5	50190.5	927171
BA_6	38871.76	59554.11	-339603	1256	17071	49198.25	961664
PA_1	5663.581	16563.28	0	1000	2100	5006	873552
PA_2	5921.164	23040.87	0	833	2009	5000	1684259
PA_3	5225.682	17606.96	0	390	1800	4505	896040
PA_4	4826.077	15666.16	0	296	1500	4013.25	621000
PA_5	4799.388	15278.31	0	252.5	1500	4031.5	426529
PA_6	5215.503	17777.47	0	117.75	1500	4000	528666

In this paper, the data pre-processing mainly contains two parts: data cleaning and non-equilibrium data processing. After analysis, there are no redundant attributes in this dataset and there are no missing values. For the marriage variables, they were combined into one category and recorded as others because they lacked practical meaning for either marriage=0 or 3. For the education variables, the variables for education=5, 6, and 0 were uniformly classified as 0 and recorded as the others category. The following statistical analysis of this dataset for the variable attributes are shown in Table 2.

The information described in Table 2 contains the number count, minimum Min, maximum Max, standard deviation std, and 25%, 50%, and 75% quartiles of each variable. The next step is to perform the data transformation. Since the actual meaning of each quantity in the dataset varies leading to inconsistencies in the absolute magnitude of each variable, which can lead to anomalies in the results of data mining and analysis, the dataset is normalized before performing exactly the analysis. According to the attributes of the data, the data in this dataset contains discrete data and continuous data. The discrete variables include nine variables such as education, marital status and repayment status, which are one-hot coded. For the remaining continuous variables, such as age, credit card limit, and repayment amount of different periods, they were normalized with the following equation:

$$x_{scale} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \tag{3}$$

Where x is the data to be normalized, x_{min},x_{max} are the minimum and maximum values of the preprocessed data respectively, and x_{scale} is the normalized data.

3.2 Dealing with the Imbalanced Data

A categorical dataset generally includes balanced and unbalanced datasets, and a dataset is considered unbalanced when the majority of categorical labels exceed a certain percentage. For unbalanced data, people tend to pay more attention to the few label items that appear in the data set,

which can be interpreted as anomalous data in the data set. In the field of credit card default prediction research, financial institutions tend to focus more on the types of customers who are more likely to default and the common characteristics they have, but default is a small probability time, so it is important to increase the proportion of default information. To address the problems associated with imbalanced data, Chawla et al. in 2002 proposed SMOTE, or (Synthetic Minority Oversampling Technique) algorithm to scale imbalanced data by adding a small number of sample data as a way to reduce prediction errors. The SMOTE function in the imblearn data package provided by Python is able to oversample the data, and the results of the defaulted and non-defaulted data after processing are shown in Figure 2 below:

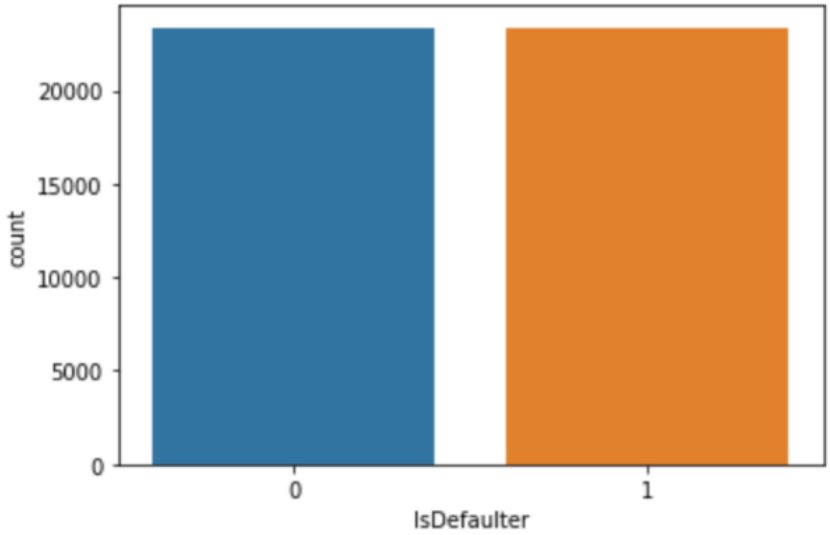

Figure 2. Distribution after SMOTE

After the balancing process, the default data increased from 6,646 to 23,364, which are equal to the non-default data.

3.3 Evaluation Metrics

Evaluation indicators are important for judging the results. The main purpose of credit card default prediction is to identify potential users who may default. The higher the recall rate, the higher the probability that the actual default and is predicted, indicating that the model is more effective, so this paper first uses the recall rate as the result evaluation index. It is calculated as follows:

$$Recall = \frac{TP}{(TP+FN)} \tag{4}$$

Where TP stands for true positive rate (i.e., actual and predicted all positive), and FN stands for false negative rate (i.e., actual positive but predicted negative). Besides, this paper also incorporates the classical index AUC for machine learning model evaluation and the evaluation indexes MSE and R2 in traditional measurement to evaluate the model effectiveness comprehensively.

3.4 Experimental results

Based on the above six machine learning predictions, we evaluate the effectiveness of the predictive classification model by comparing the known true values of the test set with the predicted values given by the model, and the evaluation metrics for the predictive effectiveness of the model are shown in Table 2.

Table 3. Comparison results

	Decision.T	XGBoost	Lightgbm	Adaboost	Logistic.R	Random.F
Recall_rate	0.725	0.821	0.953	0.823	0.827	0.822
AUC	0.614	0.661	0.785	0.648	0.659	0.659
R2_score	-0.624	-0.059	-0.094	-0.043	-0. 021	-0.052
MSE	0.274	0.179	1.572	0.176	0.172	0.177

As shown in Table 3, except for the decision tree model, all other models achieved good results in terms of accuracy and their AUC metrics were generally above 0.65. The best model was LightGBM, whose AUC value was as high as 0.785996 and its accuracy was much higher than other classification models. The recall rate of this model is 0.953, which is the highest among all models.

4. Conclusion

Combining the recall, AUC, R2_score and MSE of the above six models with the model's SMOTE method to construct a balanced data set in the case of an unbalanced ratio in the original data set, it can be concluded that the LightGBM model has higher accuracy in prediction, significantly better than other models, and is suitable to be adopted by banks to predict potential default customers, and to avoid potential risks.

This paper also analyzes the three most important variables under this algorithm, which are BA_1, Amount of bill statement in September, 2005(latest month); LIMIT BAL, amount of given credit in NT dollars; P_1, payment status in September 2005(latest month).

There are still deficient parts of this article that can be improved in the future. For example, the highest AUC value reached by the LightGBM model in this paper is 0.78, there is still room for prediction accuracy optimization, and more customer behavior information can be added in the future. Besides, the traditional onehot encoder method is adopted to encode the qualitative variables in this paper, and the latest NLP transformer method can also be tried to model the user behavior data.

References

[1] Steenackers A, et al. A credit scoring model for personal loans. Insurance: Mathematics and Economics, 1989, 8 (1): 25 - 34.

[2] Edwaretal, et al. The Important and Subtlety of Credit Rating Migration. Journal of Banking & Finance, 2009, 35 (8): 1234 - 1247.

[3] Bhattacharyya S, et al. Data mining for credit card fraud: A comparative study. Decision Support Systems, 2011, 50 (3): 602 - 613.

[4] Butaru F, et al. Risk and Risk Management in the Credit Card Industry. Journal of Banking & Finance, 2016, 23 (8): 218 - 239.

[5] Chen Ying, et al. Research on credit catd default prediction based on K-means SMOTE and BP neural network. Complex, 2021.

[6] Qiao Cuiyi et al. Application of data mining in bank credit risk management. University of Electronic Science and Technology, 2007.

[7] Fan Weiqiang, et al. Credit card default risk prediction based on BP neural network. Computer Knowledge and Technology, 2011, 7 (10): 2348 - 2349.

[8] Wang Jiao. Analysis of credit card user default prediction based on oversampling method. Northeast Normal University, 2019.

[9] Mei Ruiting, et al. Study on Analysis and Influence Factors of Credit Card Default Prediction Model, 2016.

[10] Makram Soui, et al. Credit card default prediction as a classification problem. Statistical Research, 2018.