Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/96/20241285

Prediction of Credit Default based on the XGBoost Model

Yi Chen[1], Youzhong Dong[2,4,*], Wen Liu[3]

[1]Guangdong Country Garden School, Foshan, Guangdong, 528000, China
[2]Malvern college Chengdu, Chengdu, Sichuan, 610000, China
[3]Guangzhou No.7 Middle School, Guangzhou, Guangdong, 510000, China

[4]Eric, Dongyouzhong@malvernchengdu.cn
*corresponding author

Abstract. Loans are an important source of revenue for most banks and lending institutions. Improving the accuracy of personal repayment capacity predictions is particularly crucial. It is of great significance for reducing credit risk, optimizing the credit assessment system, and ensuring market stability. This paper uses various models for the predictions. In terms of evaluation, this paper uses the ROC curve as a criterion to assess the practicality of the models. At the same time, to ensure that the most practical model does not suffer from overfitting or underfitting, this paper also uses learning curves to ensure the usability of the entire model. Experimental results have demonstrated that the XGBoost model outperformed other models in predicting credit defaults, achieving a high ROC score of 0.71. Although it cannot predict very well in actual situations, it can also demonstrate through various assessments that the XGBoost model is highly usable and has great potential. The XGBoost model can become a trend in future prediction models by continuously improving the dataset and conducting more tests. This can help governments or banks to roughly understand which customers will default in the next month and take corresponding actions accordingly.

Keywords: Credit risk forecasting, XGBoost model, Feature extraction.

1. Introduction

In recent years, with the rapid development of the Global economy, consumer credit has already permeated into our daily lives. Furthermore, loans are one of the important sources of income for the bank and lending institutions. Over the past several years, as credit granting decisions have been created, credit scoring models have been developed for calculating whether the customer can take out a loan. Data shows that from 2011 to 2018, the annual growth rate of personal consumer credit in China exceeded 25%, demonstrating the strong growth potential and broad prospects of the market. As the policies related to the national credit economy are perfected, the number of institutions engaged in credit business in China continues to increase, the types of credit products become increasingly diverse, and the credit balance has grown rapidly. Behind the rapid growth of the credit economy is the increasing demand for intelligent credit risk control. Credit risk control generally includes pre-loan review, loan risk management, and post-loan collection, among which the most critical stage is the pre-loan review. A good pre-loan risk prediction model can minimize future risks. Therefore, this paper primarily focuses on the pre-loan approval of customer credit and attempts to establish an efficient and accurate customer credit default prediction model. However, despite the rapid development of the market, the existing credit system still has many shortcomings, especially since the credit investigation capabilities of credit agencies have not been effectively enhanced. This often leads to inaccurate assessments of an individual's repayment ability, which can lead to high-risk lending and an increase in default rates. This phenomenon not only threatens the stability of financial institutions, but it may also affect the overall health of the financial system. Therefore, improving the accuracy of individual repayment ability prediction is particularly crucial as it is of great significance for reducing credit risk, optimizing the credit assessment system, and ensuring market stability.

In the field of credit default prediction, commonly used models currently include empirical analysis methods, linear regression, decision tree models, and random forest regression models. The empirical analysis method has strong subjectivity and can only provide a range or grade of default loss rates, which is not sufficient to meet the demands of large-scale credit risk management. Linear regression models perform poorly when dealing with (0,1) data and fail to fully capture the complexity of the data. Decision tree models, although capable of non-linear fitting, have high requirements for data integrity and are prone to overfitting. Random forest regression models perform well in many aspects, but they lack interpretability and make it difficult to reveal the characteristics of the predicted results. To solve these problems, researchers have proposed a variety of improvement methods. Wang Zhaohui proposed an improved Q-learning algorithm based on reinforcement learning, which optimizes the hyperparameters of the XGBoost and improves the iteration process of the Q-value table in the Q-learning algorithm, effectively solving the problem of reward value transmission in hyperparameter optimization in traditional algorithms [1]. This innovation provides a new approach to improving the accuracy of default prediction. On Github, Roberto Franceschi predicts whether customers will default next month based on machine learning. In addition, Roberto uses a lot of familiar models in order to find out which model is the best for predicting credit default. To measure the practicality of the models, Roberto also utilizes several metrics like precision, recall, F1 score, AUC, etc., to express the practicality of each model [2].

The research objective of this paper is to use the XGBoost model to predict consumer credit defaults, thereby improving the accuracy of predictions. To achieve this goal, this paper will first explore and compare several commonly used machine learning models, including Logistic Regression (LR), Support Vector Machine Classification (SVC), Random Forest (RF), and XGBoost [3,4]. Based on the comparison of these models, this paper will analyze their performance in default prediction and conduct parameter tuning for them. Secondly, this paper will employ feature extraction techniques, using the Random Forest model to evaluate the importance of features, in order to select the features that are most informative for default prediction. This process not only helps to improve model performance but also reduces computational complexity. Finally, this paper will construct and validate an XGBoost-based default prediction model, assessing its effectiveness and reference value in actual credit decisions through empirical research [5,6].

By following these steps, this paper is going to establish an accurate and relatively efficient default prediction model to provide scientific decision support for financial institutions. This not only helps to optimize credit risk management strategies but also enhances the health and stability of the entire consumer credit market. Through comprehensive research and comparison of various models, this paper hopes to provide new perspectives and solutions for future credit risk prediction, offering valuable reference for research and practical applications in related fields.

2. Data presentation

2.1. Data

The dataset that is used in this paper is "Default of Credit Card Clients Dataset." The dataset was obtained through the Kaggle platform and can be used as a practice dataset for machine learning and data analysis to study how to effectively use customer information for risk assessment and prediction. This dataset provides information about whether credit card clients will default. The data was gathered by a bank in Taiwan and includes information on default payments, demographic factors, credit data, payment history, and billing statements of credit card clients from April 2005 to September 2005.

2.2. Method

Random forests are highly effective at determining the significance of individual features by evaluating their contribution to the model's predictive performance. There are two primary methods for measuring feature importance in random forests: Mean Decrease in Impurity (MDI)entails calculating the impact of each feature on reducing node impurity (e.g., Gini index or information gain) and averaging these effects across all the trees in the model. Mean Decrease in Accuracy (MDA) evaluates the changes in model accuracy by perturbing feature values [7,8].

Random forests perform exceptionally well with high-dimensional data, capable of extracting the most informative features from a large set, thus reducing computational complexity. Their ensemble nature provides strong robustness and resistance to overfitting. Additionally, random forests do not rely on specific data distribution assumptions, allowing them to process raw data directly, which minimizes the need for data preprocessing. They have been widely utilised in fields like image processing, text analysis, and bioinformatics, where assessing feature importance effectively enhances model performance and predictive accuracy [9,10]. As a feature extraction technique, random forests are superior in terms of efficiency and intelligence. By analyzing feature correlations, they can reveal underlying patterns and structures in the data, selecting the most informative features. Highly correlated features may introduce redundant information, and by removing or combining these features, the model can be simplified and its performance improved. By extracting personal credit and default cases, calculating the linear correlation between features using the Pearson correlation coefficient, and visualizing the results with a heatmap, one can analyze the features with the highest correlation to defaults as shown in Figure 1.

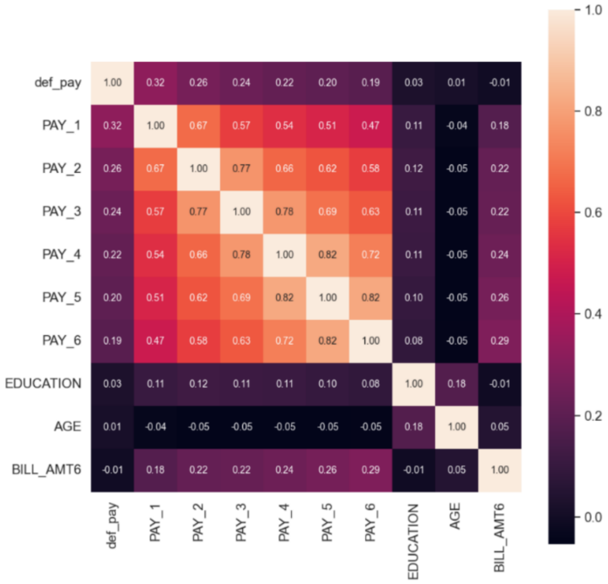

Figure 1. Feature-related thermodynamics

Logistic Regression (LR) is indeed a statistical model used primarily for binary classification problems, although its name suggests otherwise. Here's a brief explanation of logistic regression in English: Despite the term "regression" in its name, it is a classification algorithm. The core idea of this model is to use a linear model to predict the category to which the input data belongs and then map the linear output to a probability value ranging from 0 to 1 using a logistic function (Sigmoid function) [11]. The model's output represents the likelihood that the input data falls into a specific category. Generally, if this probability is 0.5 or higher, the model classifies the sample as belonging to the positive class (e.g., class 1); if it is lower, the sample is classified as part of the negative class (e.g., class 0). {"code":0, "data": "The model's output represents the likelihood that the input data falls into a specific category. Generally, if this probability is 0.5 or higher, the model classifies the sample as belonging to the positive class (e.g., class 1); if it is lower, the sample is classified as part of the negative class (e.g., class 0). The main advantages of this model are that it is easy to understand, computationally efficient, and suitable for dealing with linearly separable binary classification problems. It is widely used in fields such as spam classification, credit scoring, medical diagnosis, etc. In summary, logistic regression is a basic algorithm that is easy to understand and performs well in many practical classification problems.

The K-Nearest Neighbors (KNN) algorithm is a straightforward and efficient supervised learning method, commonly applied to both classification and regression problems. It has performed exceptionally well in numerous data science competitions and is It is widely used in a variety of tasks like classification and regression tasks XGBoost improves the model performance by integrating multiple weak learners, typically decision trees, where the latest new tree can be used to correct the previous model's mispredictions. The algorithm also incorporates regularization terms into the model, assisting in controlling the model's complexity to prevent overfitting. KNN algorithm is a simple yet effective supervised learning algorithm widely used for classification and regression tasks. It operates by calculating the distance between a new sample and all samples in the training set, then selecting the nearest K neighbours to vote or calculate the average value to predict the category or numeric value of the new sample. KNN does not require an explicit training process but instead stores all the training data and performs calculations when there is a new sample to predict, hence it is called a "lazy learning algorithm." Due to the need to calculate distances between all samples, KNN can have high computational costs when dealing with large datasets.

Support Vector Classification (SVC) is a specific application of Support Vector Machines (SVM) tailored for classification tasks. SVM is a robust supervised learning algorithm capable of addressing both linear and non-linear classification challenges. SVC‘s operational principle is finding a hyperplane that maximizes the margin between different classes, which effectively enhances the model's generalization capability. SVC can employ kernel functions (like linear, radial basis function (RBF), polynomial, etc.) to transform the input data into a higher-dimensional feature space. This allows for the identification of a linearly separable hyperplane in that space, thus addressing non-linear classification issues.

3. Experimental results and analysis

All experiments were conducted on a personal computer equipped with a 13th Gen Intel(R) Core(TM) i9-13980HX 2.20 GHz, 32.0 GB RAM, and the Microsoft Windows 11 operating system. The machine learning programs in this study were written and executed in Python 3.8 based on the Anaconda development environment. The random forest feature selection method significantly improved the model's performance. This indicates that the random forest can effectively identify the most important features for predicting defaults and reduce the dimensionality of the feature space, thereby making the model more accurate and stable. In comparison, although Logistic Regression (LR) also showed some improvement, its performance gain is not as significant as that of the random forest.

Table 1. AUC score table of each model

Model’s name	AUC	Chart sequence number
RF	0.70	1)
LR	0.66	2)
XGBoost	0.71	3)
SVC	0.54	4)
KNeighbors	0.61	5)

From Table 1, it is clear that there is not much difference in scores between the Random Forest and XGBoost, and their feature selection methods significantly improved the model's performance. This suggests that Random Forest can effectively pinpoint the most critical features for predicting defaults and reduce the dimensionality of the feature space, which enhances the model's accuracy and stability. In comparison, LR also showed some improvement, but its performance gain is not as significant as that of the Random Forest. However, overall, XGBoost performs better. Furthermore, through the confusion matrices in Figure 1, Figure 2, Figure 3, Figure 4, and Figure 5, it can also be seen that XGBoost has better overall performance. Therefore, this article ultimately chooses XGBoost for the final experiment.

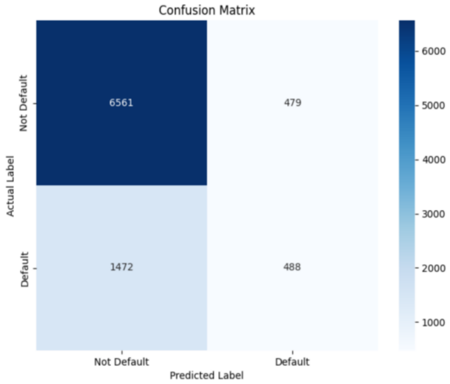

Figure 1. Confusion matrix of a random forest

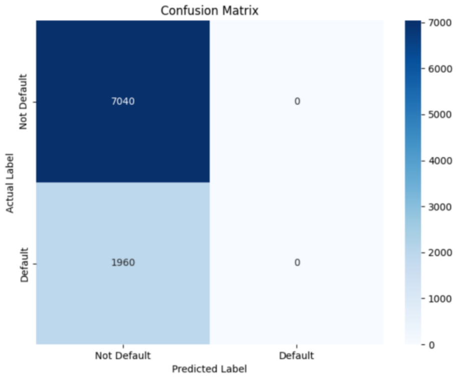

Figure 2. Confusion matrix for logistic regression

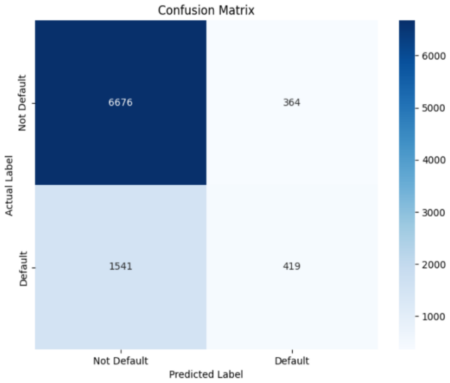

Figure 3. Confusion matrix of XGBoost

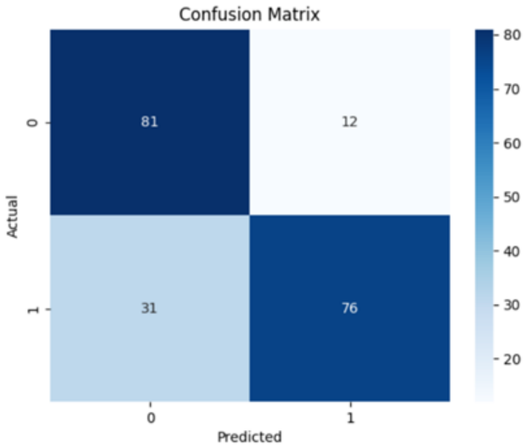

Figure 4. Confusion matrix supporting vector classification

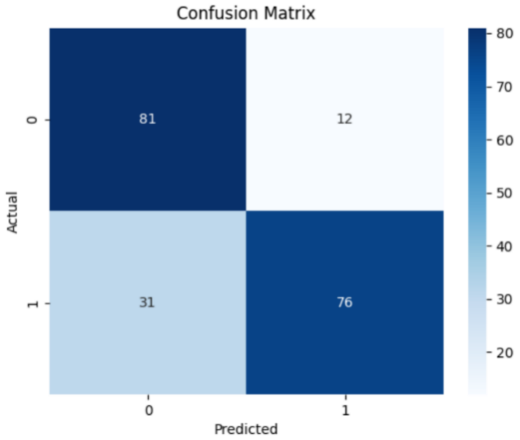

Figure 5. Confusion matrix of the k-proximity algorithm

4. Experimental process

The objective function of XGBoost consists of a loss function and a regularization term. The loss function is used to assess the error between the model's predictions and the true values, reflecting the accuracy of the model. The regularization term, on the other hand, is used to constrain the complexity of the model to reduce the risk of overfitting, thus improving the model's ability to generalize on new data.

During the tree construction process, XGBoost can automatically handle missing values in the data by finding the optimal split direction for missing values during training. The use of feature parallelism and data parallelism techniques significantly improves computational efficiency during the tree construction process.

During the training process, it disposes of the missing values by finding the best-split direction for the missing values automatically. By iterative training multiple weak learners (tree models), each new tree model learns from the residual of the previous round (prediction error), attempting to correct the mistakes of the previous prediction and therefore improving the prediction accuracy.

After our group compared the models that we have been mentioned, the experimental results indicate that the XGBoost model performs the best in credit default prediction. Its AUC score reached 0.71, significantly outperforming the other models. Figure 6 shows that it has the optimal predictive performance in dealing with credit default data. Due to inadequate algorithm optimization and the limited amount of data, there is still room for improvement in the prediction results.

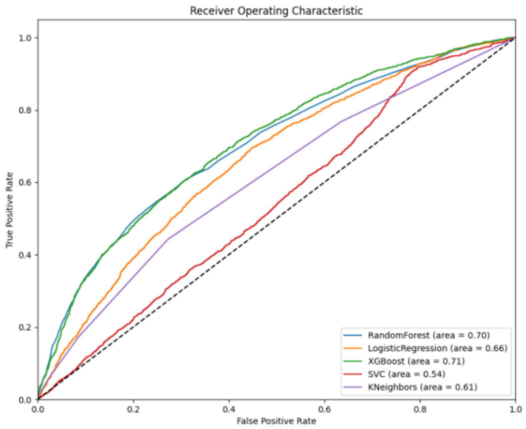

Figure 6. ROC curve

5. Conclusion

Credit default prediction holds significant practical importance in the current financial sector and has profound prospects for the future. Through this research, which is based on deep learning neural network models for predicting credit defaults, we have delved deeply into how to utilize advanced machine learning technologies to improve the accuracy and reliability of predictions. Not only does this assist financial institutions in assessing customer credit risk more accurately, but it can also effectively reduce credit losses and improve the quality and stability of overall loan portfolios. In the credit default risk prediction dataset, the superior performance of the XGBoost model in metrics such as AUC scores and accuracy highlights its effectiveness in capturing complicated patterns and interactions in the data. This superior capability is attributed to its ability to handle more dimensional information, better handling of missing values, and stronger algorithmic principles.

In the future, with the continuous advancement of data science and artificial intelligence technologies, credit default prediction is expected to see more innovation and expansion. Firstly, models can be further optimized by integrating more data sources and more complex feature engineering to improve the robustness and adaptability of prediction models. Secondly, with the development of blockchain technology and big data analytics, it is possible to explore the establishment of safer and more efficient credit assessment systems, thereby improving customer experience and the efficiency of financial services. In addition, the advancement of cross-industry cooperation and cross-international data sharing will also bring new possibilities and opportunities for credit risk management on a global scale.

In conclusion, credit default prediction is not only a vital component of the development of financial technology but also one of the key tools for achieving inclusive finance and financial stability. Through ongoing research and innovation, we are confident that we will be able to advance this field in someday, contributing to the sustainable progression of the financial industry.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Mao X. Model study for the estimation of microloan default loss rate. Southwestern University of Finance and Economics, 2020.

[2] Mei R, Xu Y, Wang G. Analysis of credit card default prediction model and exploration of influencing factors. 2016.

[3] Wang Z. Credit risk prediction study based on reinforcement learning and XGBoost. Taiyuan University of Technology, 2022.

[4] Franceschi R. Machine learning algorithms for predicting default of credit card clients. Available from: https://github.com/robertofranceschi/default-credit-card-prediction#:~:text=The%20aim%20of%20this%20study%20is%20to%20exploit%20some%20supervised

[5] Yotsawat W, Wattuya P, Srivihok A. Improved credit scoring model using XGBoost with Bayesian hyper-parameter optimization. International Journal of Electrical and Computer Engineering. 2021;11(6):5477-5487.

[6] Zhang L, Wu Z. Comparison of credit scorecard based on XGBoost machine learning model and logistic regression model. Journal of South-South University for Nationalities (Natural Science Edition). 2023;42(6):846-852.

[7] Xu M. Online credit default prediction based on ensemble learning. Southwest University, 2023.

[8] Xie L. Research on customer credit default prediction based on deep learning. East China Normal University, 2022.

[9] Shan H. A study based on machine learning. Data Mining. 2019;9(4):8.

[10] Li J. A loan default forecast study based on the improved XGBoost algorithm. Northern University for Nationalities, 2024.

[11] Fei H, Huang H. Research on internet credit and credit risk prediction based on model fusion. Statistics and Application. 2019;8(5):12.