International Journal of Data Warehousing and Mining
Volume 19 • Issue 1

Loan Default Prediction Based on Convolutional Neural Network and LightGBM

Qiliang Zhu, North China University of Water Resources and Electric Power, China*
https://orcid.org/0000-0001-7592-0184
Wenhao Ding, North China University of Water Resources and Electric Power, China
Mingsen Xiang, North China University of Water Resources and Electric Power, China
Mengzhen Hu, North China University of Water Resources and Electric Power, China
Ning Zhang, North China University of Water Resources and Electric Power, China

ABSTRACT

With the change of people's consumption mode, credit consumption has gradually become a new consumption trend. Frequent loan defaults give default prediction more and more attention. This paper proposes a new comprehensive prediction method of loan default. This method combines convolutional neural network and LightGBM algorithm to establish a prediction model. Firstly, the excellent feature extraction ability of convolutional neural network is used to extract features from the original loan data and generate a new feature matrix. Secondly, the new feature matrix is used as input data, and the parameters of LightGBM algorithm are adjusted through grid search so as to build the LightGBM model. Finally, the LightGBM model is trained based on the new feature matrix, and the CNN-LightGBM loan default prediction model is obtained. To verify the effectiveness and superiority of our model, a series of experiments were conducted to compare the proposed prediction model with four classical models. The results show that CNN-LightGBM model is superior to other models in all evaluation indexes.

KEYWORDS

AUC, Boxplot, CNN-LightGBM, Confusion Matrix, GBDT, Heat Map, Histogram, Logistic Regression, Normalize, XGBoost

INTRODUCTION

As credit consumption has become the lifestyle of more and more people, credit consumption has become an important part of the national economy and has played a great role in promoting the actual economy. From 2014 to 2019, China's Internet consumer credit scale expanded rapidly, from 18.7 billion yuan to about 16.3 trillion yuan. Credit consumption has become a new driving force for economic growth. However, loans without collateral are bound to be accompanied by bad behavior such as fraud and default. Frequent loan defaults have also become an important factor that restricts the development of the credit industry and even hinders social and economic growth (Boateng & Oduro, 2018). Financial institutions will not be able to deal with bad debts in time due to a large

DOI: 10.4018/IJDWM.315823 *Corresponding Author

number of loan defaults, resulting in huge losses and even the risk of bankruptcy. In order to control the loan default ratio within a safe range, the risk prediction of loan default has become one of the most important tasks of financial institutions.

In recent years, machine learning technology has been widely used in the financial industry. The improvement of efficiency and reliability brought by machine learning algorithms makes it indispensable in this field. With the gradual rise of neural networks, data mining, and other technologies, more and more scholars apply these technologies to the prediction of loan default risk (Teply & Polena, 2020). Compared with the traditional logistic regression prediction model, the two-stage credit evaluation model and the ensemble model of two or more algorithms used at this stage greatly improve the prediction ability of loan default. However, limited by the feature extraction ability of standard feature engineering data, their prediction accuracy has not made a qualitative breakthrough. In order to reduce the negative impact of complex feature engineering on loan default modeling, this paper introduces convolutional neural network and LightGBM algorithm into the field of loan default prediction uses convolutional neural network instead of feature engineering to extract data set features and establishes a hybrid algorithm model to improve prediction accuracy and prediction efficiency. From the perspective of deep learning, a convolutional neural network has excellent performance in obtaining information. Compared with feature engineering technology, a convolutional neural network can obtain the actual features of datasets more quickly and comprehensively. The fully connected layer used for classification in a convolutional neural network is to further realize the mapping of feature space to target space on the basis of convolutional layer feature extraction, and because the fully connected layer has a large number of parameters, it is easy to produce overfitting phenomenon when the training data is not enough. Therefore, this paper proposes that after the neural network training is completed, only the output of the convolutional layer is extracted as a newly derived variable, and the classification results are obtained by further learning by the machine learning algorithm. For the choice of machine learning algorithms, we think LightGBM is the most ideal. LightGBM algorithm is an improvement on the GBDT algorithm, supports high-efficiency parallel training, and has advantages such as faster training speed, lower memory consumption, better accuracy, support for distributed can quickly process massive data. It is one of the best machine learning models at present. Compared with the current mainstream model, the combined model of convolutional neural network and LightGBM algorithm has better performance in the accuracy of loan default risk prediction.

RELATED WORK

With the rapid development of the Internet financial industry at home and abroad, the shortage and existing problems of online credit have become increasingly prominent, and the default of loan users has become increasingly common. What kind of algorithm model can be used to predict user loan risk more effectively has now become a research hotspot of many scholars.

In recent years, various scholars have tried to apply machine learning algorithms to loan default prediction and made good progress. Zhang et al. established a random forest model for loan default prediction by sorting the importance of features and calculating the important features that affect the default. The results show that the prediction performance of the random forest algorithm is better than the decision tree and logistic regression classification algorithm (H. Zhang et al., 2020). By measuring the importance of each feature, Zhang et al. obtained the borrower's debt ratio, the number of historically overdue times, and the ratio of total loans to total credit, which had a great impact on the ultimate default (L. Zhang et al., 2021).

Chotwani et al. (2019) took the fraudulent loan data set as the research object, studied the data mining algorithm, looked for the data mining algorithms with better performance than the algorithm, and used it to predict the fraudulent loan. Cerchiello and Scaramozzino (2020) applied text analysis to augment the traditional set of account default drivers with new text-based variables, classifying bank account users into different customer profiles by using ad hock dictionaries and distance