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

metrics and verify that they can be effective predictors of default through a supervised classification model. Dendramis et al. (2020) proposed the use of an asymmetric binary link function to extend a proportional hazards model for forecasting loan default, demonstrating that ignoring the actual level of asymmetry leads to a severely biased estimate of the slope coefficient. Niu et al. (2020) proposed a further resampling ensemble model based on data distribution, which effectively improves the comprehensive classification performance of P2P lending imbalance credit risk assessment. Although the above methods have made some achievements, there are still some shortcomings in prediction accuracy and universality.

Some researchers have established hybrid algorithm models, or established multiple algorithm models, and selected appropriate models through comparison. Luong and Scheule (2022) found that co-borrowers, loan contracts, and external characteristics are significant in explaining forward credit risk and developed a hybrid model to predict default probability. Cai and Zhang used the loan data set provided by the LendingClub platform to evaluate the actual data by using a decision tree and binomial Logistic Regression algorithm and demonstrated that the decision tree algorithm can improve the accuracy of preliminary screening and more accurately predict the default of a borrower's probability (Cai & Zhang, 2020). Using logistic regression, random forest, gradient boosting, and CatBoost classifiers, Patel et al. (2020) obtained results for predicting defaulters, respectively, and experiments show that the gradient boosting process provides better or equivalent results compared to logistic regression. Jin et al. used the idea of unbalanced data classification to study the statistical analysis of historical loan data of financial institutions such as banks and used machine learning algorithms such as random forest, logistic regression, and decision tree to establish a loan default prediction model. The experimental results demonstrate that the neural network and random forest algorithm outperform the decision tree and logistic regression classification algorithm in prediction performance (Jin et al., 2021).

Neural networks and LightGBM have gradually become new research focuses. Ma et al. used "multi-observation" and "multi-dimensional" data cleaning methods, applied LightGBM and XGBoost algorithms to model P2P transaction data, and observed that the LightGBM algorithm based on the classification and prediction results of multiple observation data sets is the best thing (Ma et al., 2018). Zhang combined random undersampling and SMOTE oversampling to effectively avoid the problem of algorithm failure when the data are extremely unbalanced and verified that the performance of the integrated algorithm is preferable to that of a single algorithm, and the performance of LightGBM is better (X. Zhang, 2022). Qian et al. utilized the data of public loan transactions as experimental data and successfully predicted the default risk of loan users by using feedforward neural network and least squares support vector machine (Qian & Hu, 2019).

PREDICTION MODEL

LightGBM Algorithm

LightGBM is an iterative boosting tree system provided by Microsoft, an improved variant of gradient boosting decision tree (GBDT; Ke et al., 2017). The classic GBDT generally only uses the first-order negative gradient of the loss function (Yang et al., 2020). Still, it also uses the first-order and second-order negative gradients of the loss function to calculate the residual of the current tree and uses this result to fit the new tree in the next round.

To minimize the specified loss function $L\left(y_1, f\left(x\right)\right)$, the LightGBM algorithm finds the approximate value $\hat{f}\left(x\right)$ of $f\left(x\right)$ through training, where $\hat{f}\left(x\right)$ is also called the optimization function, which can be expressed as:

$$\hat{f}\left(x\right) = \arg\min E_{y,x} L\left(y_1, f\left(x\right)\right) \tag{1}$$

Integrated K regression trees in the LightGBM model to fit the final model. This process can be expressed as:

$$f_k\left(x\right)=\sum_{i=1}^{k}f_1\left(x\right) \tag{2}$$

The regression tree in the model is represented by $M_{q\left(x\right)}, q\in\left\{1,2,...,J\right\}$, where M is the sample weight vector of the leaf node and J is the number of leaves in the regression tree. In particular, the information of the previous $\left(t-1\right)$ trees will be used when the *t*th tree is generated, so the objective function generated after *t* iterations will be as follows:

$$\eta_t=\sum_{i=1}^{n}L\left(y_1,F_{t-1}\left(x_i\right)\right)g_if_i\left(x_i\right)+\sum_{j=1}^{m}\Omega\left(f_j\left(x\right)\right) \tag{3}$$

In the above formula, $\Omega\left(f_m\left(x\right)\right)$ is the regularization term, and the purpose is to let the model to avoid overfitting when training the data, carrying out the second-order Taylor expansion of the objective functions. The expanded objective function can be expressed as:

$$\eta_t=\sum_{i=1}^{n}\left(L\left(y_1,F_{t-1}\left(x_i\right)\right)g_if_i\left(x_i\right)+\frac{1}{2}f_i^2\left(x_i\right)\right)+\sum_{j=1}^{m}\Omega\left(f_j\left(x\right)\right) \tag{4}$$

After determining the tree structure as $q\left(x\right),$ the corresponding objective function is:

$$\eta_t^*=-\frac{1}{2}\sum_{j=1}^{J}\frac{\left(\sum_{i\int I_j}g_i\right)^2}{\sum_{i\int I_j}h_i+\lambda}+\gamma J \tag{5}$$

In Equation 5:

$$-\frac{\left(\sum_{i\in I_j}g_i\right)^2}{\sum_{i\in I_j}h_i+\lambda}$$

is the optimal weight score of each leaf node, the optimization problem that the model needs to implement. It refers to minimizing the objective function, maximizing the splitting income by calculating the splitting income of the leaf nodes of the regression tree, selecting the splitting feature with the most considerable income, and continuing to iterate this process until the conditions are met. The split payoff can be expressed by:

$$F = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I_K} g_i\right)^2}{\sum_{i\in I_K} h_i + \lambda}\right] \tag{6}$$

In addition, it also adopts the decision tree algorithm based on a histogram, which can save half the time of GBDT algorithm execution (Peng et al., 2019), and its leaf growth strategy with maximum depth limit makes the overall implementation of the algorithm more efficient.

Deep Convolutional Neural Network Model

A *convolutional neural network* (CNN) is a kind of feedforward neural network with convolution calculation and deep structure, and it is one of the representative algorithms of deep learning (Gu et al., 2018; Janiesch et al., 2021). Convolutional neural networks can learn representations and perform translation-invariant classification of input information according to its hierarchical structure (Myburgh, 2021). A deep convolutional neural network consists of three parts: a convolutional layer, a pooling layer, and a fully connected layer. The core of the convolutional neural network is the convolutional layer, which mainly contains various latent features. After several layers of convolution and pooling operations, the obtained feature maps are expanded row by row, connected into the vectors, and input into the fully connected network to classify the finally got high-level features. Equation 7 calculates the classification probability for the softmax formula:

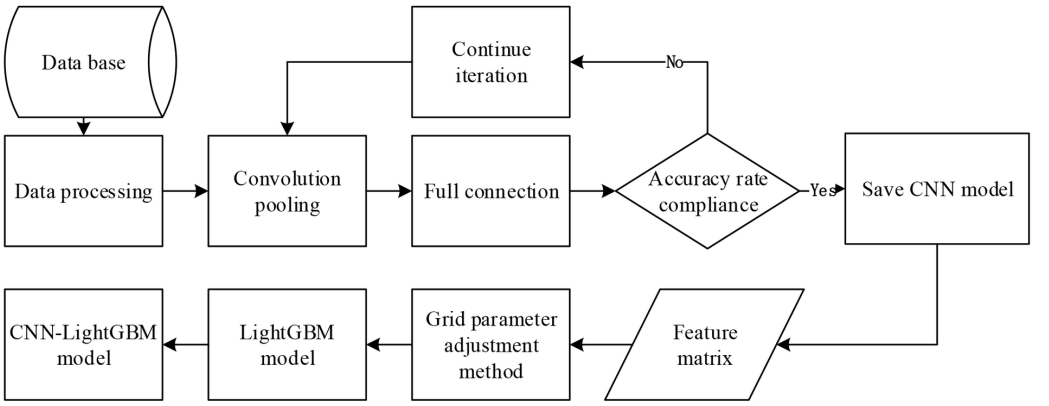
$$y_i = \frac{e^{zi}}{\sum_{j=1}^{n} e^{zj}} \tag{7}$$

Among them, z is the high-level feature; n is the number of categories; y is the probability that the sample belongs to a certain category.

CNN-LightGBM Model

The feature matrix obtained by the fully connected layer of the convolutional neural network is invoked as the input data of the LightGBM algorithm. In general, the feature matrix extracted by the neural network is sparse, and the LightGBM algorithm can automatically learn its classification direction, saving a lot of training time. The construction architecture of the model is shown in Figure 1. First, the neural network extracts the original data feature information. If the correct rate calculated by softmax fluctuates within a predetermined range, increase the number of iterations. Otherwise, output the feature matrix, save the model, and use the feature matrix. After adjusting the parameters of the LightGBM model through a combination of grid parameter tuning and cross-validation, the LightGBM model is trained again using the feature matrix. Finally, a CNN-LightGBM model that predicts whether the loan will default under different conditions is formed.

Figure 1. CNN-LightGBM model



Parameter Adjustment

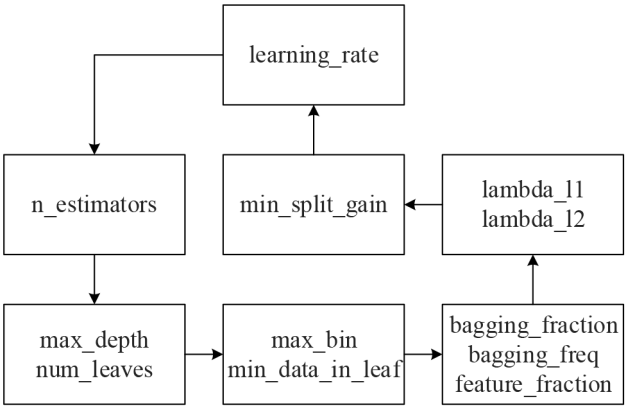
This paper uses the LightGBM module to analyze the data. The LightGBM algorithm is a multi-parameter supervised model, and adjusting parameters determines the model's performance to a large extent. Table 1 gives the main parameters that affect the quality of the model, where the learning rate and the number of trees are closely related. In general, a lower learning rate means more weak learners.

Table 1. Main parameter

Main parameter	Meaning
learning_rate	Learning rate
n_estimators	Number of weak learners
num_leaves	Number of leaf nodes in the decision tree
max_bin	Feature maximum split
min_data_in_leaf	Minimal data for one leaf
max_depth	Depth of the tree
bagging_fraction	Scale of data used at each iteration
lambda_l1	Regular coefficient of $L_1(\gamma)$
lambda_l2	Regular coefficient of $L_2(\lambda)$

According to the relationship between the parameters, the parameter adjustment steps based on the grid parameter adjustment method in this paper are shown in Figure 2. First, set the value of the learning rate, other parameters default, adjust the number of weak learners, and then set the number of optimal trees (n_estimators) and the preset learning rate are fixed, adjust the following two parameters, and so on, and finally fine-tune learning_rate to get a set of optimized parameters.

Figure 2. Parameter optimization



DATASET FEATURE EXTRACTION

Data Preprocessing

The data in this article come from the data set of the financial risk control competition on the Tianchi platform (https://tianchi.aliyun.com/competition/entrance/531830/information). There are 1 million data sets. We randomly select 200,000 of them as the data set for this paper. The data set contains 47 columns of variable information, 15 of which are unknown variables. Among them, $isDefault$ is the target variable indicating whether the loan default and the variable data such as $employmentTitle$, $purpose$, $postCode$ and $title$ have been desensitized. The variables can be roughly divided into four categories: loan information, borrower information, borrower credit information, and *n*-series of unknown variables. The introduction of some variables is shown in Table 2.

Table 2. Introduction to some variables

	Variable	Variable description	Variable type
Loan information	loanAmnt	loan amount	Numerical
	term	loan term (year)	Numerical
	installment	Installment amount	Numerical
	subGrade	child of loan class	type value
Borrower information	employmentLength	Years of employment (years)	Date variable
	annualIncome	Annual income	Numerical
	dti	debt-to-income ratio	Numerical
Borrower credit information	openAcc	The number of outstanding lines of credit in the borrower's credit file	Numerical
	pubRecBankruptcies	Number of public record removals	Numerical
	revolUtil	Cycle quota utilization	Numerical
n-series anonymous features	n0-n14	Lender behavior counting features	Numerical

The dataset preprocessing in this paper mainly includes type data for numerical processing, missing value data processing, noise data processing, etc. Among them, date variables such as $employmentLength$ need to be substituted into the model. The $employmentLength$ format is "1 year, 2 years, 8 years, 10+ years". After numerical encoding, the $employmentLength$ becomes a range of 0 to 10 values. The data were selected and deleted for each data variable with more than two missing values. The rest of the data with missing values were filled with the mean for numerical variables and the mode for categorical variables. Due to the significant variance of some data in loan data, such as loan amount, this paper uses the box plot method to filter abnormal data, which is a fast and effective unnatural data processing method. In the boxplot program, the criteria for identifying outliers are set as the upper limit U+1.5IQR and the lower limit L-1.5IQR, where U is the upper quartile of the dataset, L is the lower quartile, and the interquartile range IQR = U-L.

Feature Engineering

The amount of loan default data in the dataset is far from the average repayment data, and the resulting sample imbalance will affect the prediction accuracy. In the research, the oversampling method and the undersampling method are used to reconstruct the dataset. The oversampling method achieves data balance by randomly repeating the data with a small proportion to expand its ratio, and the undersampling rule achieves data balance by deleting the data with a large proportion. The data set in this paper uses a combination of the above two methods to achieve sample balance while keeping the total number of data unchanged.

Numerical gaps between variables are large. As different dimensions and dimensional units may affect the model convergence rate and prediction results. Therefore, we need to normalize the variables before feature screening using the most value interval scaling method (Durga & Jeyaprakash, 2019). The calculation formula of this method is:

$$x' = \frac{x - \min\left(x\right)}{\max\left(x\right) - \min\left(x\right)}$$

In the formula, x is the value of the sample data, $\min\left(x\right)$ is the minimum value in the sample data, $\max\left(x\right)$ is the maximum value in the sample data, and x' is the sample data scaled to the [0,1] interval of the corresponding value above.

In the process of machine learning modeling, the number of features is not "the more, the better." It is necessary to exclude irrelevant or insignificant features to improve the generalization ability of the model and reduce the operational cost of the model. Constructing a correlation coefficient heat map shows the correlation between the predictor variable and the target variable and the multicollinearity between the variables. The correlation between the predictor variable $isDefault$ and other features can be found through the heat map. Some feature correlations are shown in Table 3. It can be seen from Table 3 that some features are negatively correlated with predictors, and only 32 features with positive correlations are retained in this paper as Dataset A for subsequent experiments.

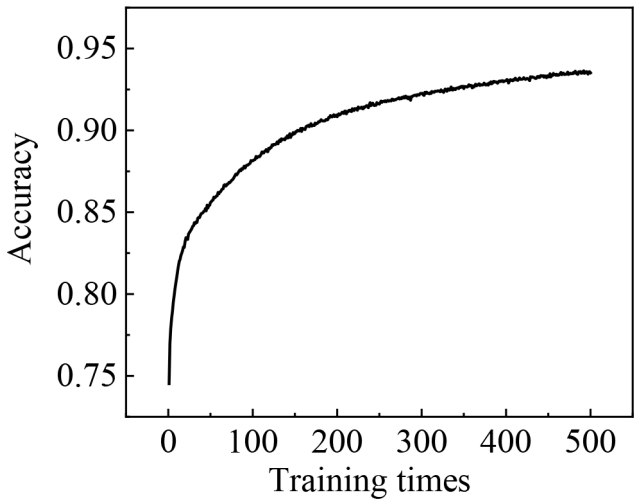
Table 3. Correlation of isDefault to other features

Features	Correlation	Features	Correlation
loanAmnt	0.087944	employmentTitle	-0.021616
term	0.153772	employmentLength	-0.039561
interestRate	0.335937	homeOwnership	-0.093670
installment	0.070230	annualIncome	-0.054608
grade	0.236875	verificationStatus	-0.046666
subGrade	0.324914	issueDate	0.039430

CNN Feature Extraction

In this paper, for the problem of loan default prediction, latent feature extraction is performed on the original dataset through the network characteristics of CNN. This paper uses the Keras neural network framework to build a feature extraction network. The number of neurons in each layer is 512, 256, 128, 64, 32, 26, and 1, and the parameters of the dropout layer are 0.2 and 0.1, respectively. The optimizer is adam, the loss function is binary_crossentropy, and the performance evaluation indicator uses accuracy. Use the original data set as the training matrix, and divide the training matrix into two parts: One part trains the CNN network, and the other part tests the CNN network and outputs the correct rate. When the correct test rate does not change within the predetermined range of the number of iterations or is 100%. As shown in Figure 3, the output of the fully connected layer is extracted, and 32 features are extracted as Dataset B for the subsequent experiments.

Figure 3. CNN training times versus accuracy



EXPERIMENTS

Comparison Method

To verify the prediction performance of the CNN-LightGBM integrated model, this paper uses the loan default data set provided by the Tianchi big data platform to conduct multiple sets of simulation experiments. An average of 10 rounds of prediction results was taken in the experiment, and the prediction effect of the CNN-LightGBM fusion model was compared with the single algorithm of logistic regression, XGBoost, and LightGBM.

Comparative Indicators

To evaluate the detection performance of the model, the confusion matrix, accuracy, precision, and recall are used as model evaluation metrics. The confusion matrix is a matrix that describes the classification results in detail. The two-class problem is a 2×2 matrix, and the multi-class problem is an $n \times n$ matrix. The confusion matrix for the two-class problem is shown in Table 4. The sum of the first row is the number of samples whose actual class is positive, and the sum of the row is the number of pieces whose exact type is negative.

Table 4. Confusion matrix

Forecast results	Actual result	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Among them, TP represents the number of samples that predict the positive class as a positive class. FN represents the number of samples that predict a positive class as a negative class, and FP represents the number of samples that indicate a negative type as a positive class. TN means the negative course is the number of instances of the negative category. The model in this paper is used to predict loan default. It is aimed at the binary classification problem, and the data have been balanced. Therefore, the four indicators of ACC, Precision, recall, and AUC are used to evaluate the model. ACC is the correct rate, reflecting the proportion of correctly classified samples in the total, calculated as:

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{8}$$

Precision reflecting the proportion of positive samples in all samples, calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall represents the proportion of all positive samples that are found. These indicators are calculated from the classification results of positive and negative samples, calculated as:

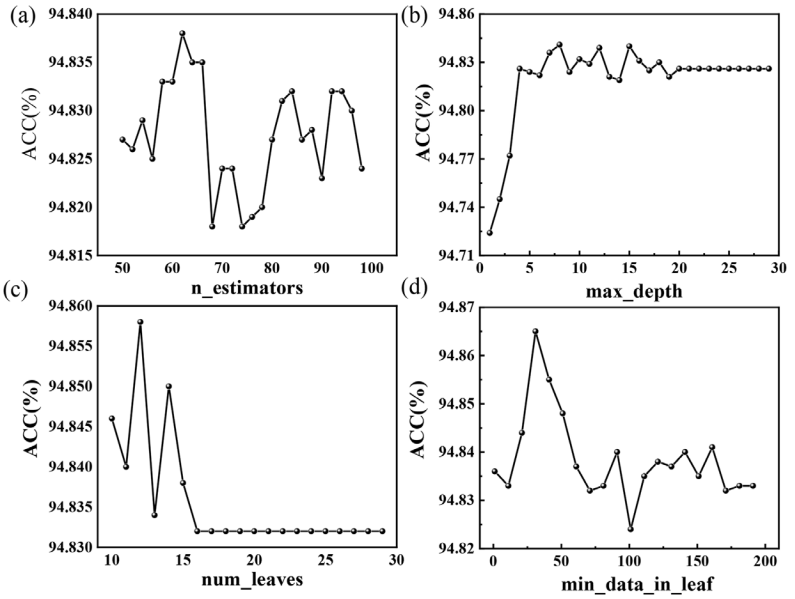
$$Recall = \frac{TP}{TP + FN} \tag{10}$$

AUC is the area under the ROC curve and is generally used to evaluate the performance of a classifier.

Parameter Tuning

The data set B is used as the input data of the CNN-LightGBM model. The adjustment process of the model parameters is shown in Figure 2. The default parameters are used, the learning_rate of the model is set to 0.1, and the five-fold cross-validation is used to determine the number of trees, and n_estimators set in the range of 50 to 100, use grid search algorithm, perform parameter search every two units, and draw a line graph of n_estimators and prediction accuracy. As can be seen from Figure 4(a), when n_estimators is set to 62, ACC reaches the maximum value for a test error of 0.0516.

Figure 4. CNN-LightGBM model parameters and ACC variation curve



Max_depth, the maximum depth parameter of the tree, is an important parameter of the LightGBM model. According to experience, it is set in the range of 1 to 30, and every other unit performs a parameter search. As shown in Figure 4(b), the error is the smallest when max_depth is adjusted to 8. num_leaves is closely related to max_depth and is also one of the most important parameters to control the complexity of the model. Since the value of max_depth is 8, num_leaves is set between 10 and 30, and a parameter search is performed for every unit. As shown in Figure 4(c), when num_leaves is set to 12, the ACC reaches the maximum. Set min_data_in_leaf between 0 and 100, as shown in Figure 4(d). When min_data_in_leaf is set to 31, the ACC is the largest. Next, set max_bin to 85, and set feature_fraction, bagging_fraction, and bagging_freq to 0.6, 0, and 0.7, respectively. Finally, adjust lambda_l1 and lambda_l2 to 1e-05 and 1e-03, and set min_split_gain to 0. 0463. After cyclic adjustment, the parameter values in Table 5 are finally obtained, and the test error is reduced to 0.0463.

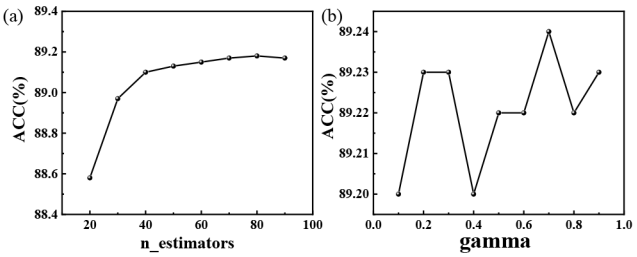
Table 5. Optimal parameter

Parameter	Value
learning_rate	0.1
n_estimators	62
max_depth	8
num_leaves	12
max_bin	215
min_data_in_leaf	31
bagging_fraction	0.9
bagging_freq	40
feature_fraction	0.7
lambda_l1	1e-05
lambda_l2	1e-03
min_split_gain	0.8

The logistic regression model is widely used in loan default prediction (Fox & Hammond, 2019). Dataset A is divided into 75% of the training set and 25% of the test set and imported into the logistic regression algorithm module. The algorithm parameters are adjusted. That is, the penalty is L2, and the regularization coefficient λ is the reciprocal *C* is adjusted to 10000 to obtain the optimal parameters.

The XGBoost model, also known as the extreme gradient boosting tree model, was improved by Chen and Guestrin (2016) based on GBDT. It is an ensemble learning algorithm formed by combining base functions and weight through the idea of boosting. Dataset A is used as the input data of the XGBoost model for training, and the parameters are adjusted. Figure 5 shows the relationship between some parameters of the XGBoost algorithm and the accuracy. It can be seen from Figure 5 that when n_estimators is in the range of 0–100, the accuracy rate increases and tends to be stable. When n_estimators is set to 80 and gamma is 0.7, the model accuracy rate is the highest.

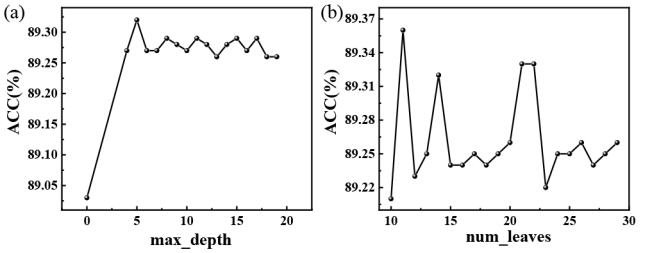
Figure 5. XGBoost model parameters and ACC variation curve



As the framework of GBDT algorithm, LightGBM supports efficient parallel training, and has the advantages of faster training speed, lower memory consumption, better accuracy, and support for distributed and fast processing of massive data. Dataset A is used as the input data of the LightGBM model to train and adjust the parameters. Figure 6 shows the relationship between some parameters of the LightGBM algorithm and the accuracy. It can be seen from Figure 6 that under the condition

that n_estimators is set to 86, the accuracy rate is the highest when max_depth is equal to 5. As the depth increases, the accuracy rate fluctuates slightly and tends to be stable. The change of num_leaves has a great influence on the accuracy. When num_leaves is set to 11 in the interval of 10 to 30, the accuracy reaches the maximum value.

Figure 6. LightGBM model parameters and ACC variation curve



ANALYSIS OF RESULTS

In the construction of the logistic regression model, XGBoost, LightGBM, and CNN-LightGBM model, 32 features are used as input variables. Table 6 shows the comparison results of the prediction effects of each model. It can be seen from Table 6 that XGBoost and LightGBM, as the most popular algorithm models in the field of machine learning, still have good performance in this data set and have a great improvement in all aspects of evaluation indicators compared to logistic regression. It is also the boosting algorithm of GBDT. Since LightGBM adopts the decision tree algorithm based on the histogram, it realizes faster training speed. The leaf-wise splitting method generates more complex trees than the level-wise splitting method, achieving higher accuracy. In addition, we can prevent overfitting by setting the max-depth parameter to limit the maximum depth of the tree resulting from splitting. Experiments show that LightGBM had better performance than XGBoost in this dataset. Therefore, this paper uses LightGBM in combination with convolutional neural networks.

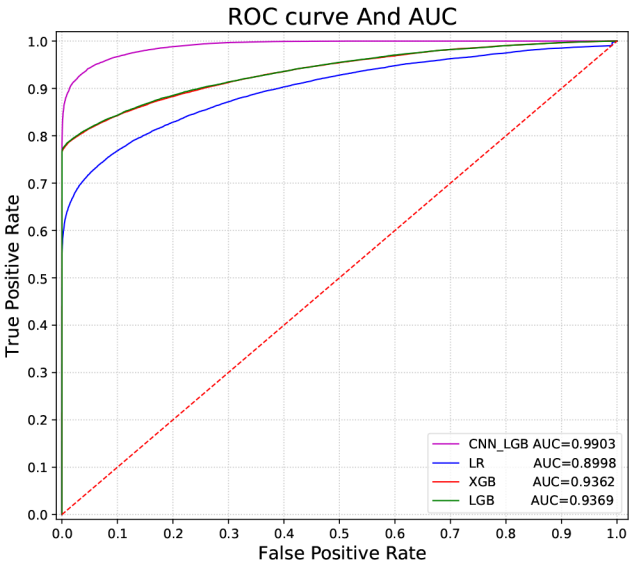
Table 6. Comparison of classification results of different algorithms

Algorithm	Precision	Recall	ACC
Logistic regression	0.85	0.83	0.83
XGBoost	0.90	0.88	0.88
LightGBM	0.91	0.88	0.88
CNN-LightGBM	0.95	0.95	0.95

In this paper, the grid search method is combined with five-fold cross-validation to train the model parameters, and the CNN-LightGBM model with optimal parameters is obtained through multiple experimental verifications. Compared with XGBoost and LightGBM, the precision, recall, and ACC results of the CNN-LightGBM model have increased by 5% to 7%, reaching 95%, which is difficult for traditional machine learning algorithms to achieve. The AUC value of the same CNN-LightGBM model is also greatly improved. As shown in Figure 7, the AUC value of the CNN-LightGBM model

is enhanced by more than 0.05 compared with the XGBoost and LightGBM models and is improved by 0.09 compared with the logistic regression model.

Figure 7. AUC values of different models



CONCLUSION

In order to improve the accuracy of loan default prediction, this paper constructs a learning method for the LightGBM model based on convolutional neural network feature extraction. The performance of this method is verified by comparing it with a single model. Experimental results show that compared with logistic regression, XGBoost, and LightGBM algorithms, the CNN-LightGBM model has higher accuracy and classification performance in loan default prediction. The prediction results are higher than 90%, and the AUC is higher than 95%, which verifies the feasibility of this method. In the future work, we will further verify the feasibility of CNN-LightGBM model for predicting and studying other financial problems and explore how this method can widely identify and deal with other regression prediction and classification problems.

CONFLICT OF INTEREST

The authors of this publication declare there is no conflict of interest.

FUNDING AGENCY

The open-access processing charge for this article was covered in full by the authors of this article.

REFERENCES

Boateng, E. Y., & Oduro, F. T. (2018). Predicting microfinance credit default: A study of Nsoatreman rural bank, Ghana. *Journal of Advances in Mathematics Computer Science*, *26*(1), 1–9. doi:10.9734/JAMCS/2018/33569

Cai, S., & Zhang, J. (2020). Exploration of credit risk of P2P platform based on data mining technology. *Journal of Computational and Applied Mathematics*, *372*, 112718. doi:10.1016/j.cam.2020.112718

Cerchiello, P., & Scaramozzino, R. (2020). On the improvement of default forecast through textual analysis. *Frontiers in Artificial Intelligence*, *3*, 16. doi:10.3389/frai.2020.00016 PMID:33733135

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. doi:10.1145/2939672.2939785

Chotwani, P., Tiwari, A., & Hooda, M. (2019). Fraudulent loan prediction using machine learning algorithms. *Indian Journal of Public Health Research & Development*, *10*(5), 845–850. doi:10.5958/0976-5506.2019.01187.2

Dendramis, Y., Tzavalis, E., Varthalitis, P., & Athanasiou, E. (2020). Predicting default risk under asymmetric binary link functions. *International Journal of Forecasting*, *36*(3), 1039–1056. doi:10.1016/j.ijforecast.2019.11.003

Durga, V. S., & Jeyaprakash, T. (2019). An effective data normalization strategy for academic datasets using log values. *Proceedings of the 2019 International Conference on Communication and Electronics Systems*, 610–612.

Fox, W. P., & Hammond, J. (2019). Advanced regression models: Least squares, nonlinear, Poisson and binary logistics regression using R. Data Science and Digital Business, 221–262.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377. doi:10.1016/j.patcog.2017.10.013

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685–695. doi:10.1007/s12525-021-00475-2

Jin, H., Luo, L., Wang, X., Zhu, X., Qian, L., & Zhang, Z. (2021). Financial credit default forecast based on big data analysis. *Academic Journal of Business Management, 3*(8).

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

Luong, T. M., & Scheule, H. (2022). Benchmarking forecast approaches for mortgage credit risk for forward periods. *European Journal of Operational Research*, *299*(2), 750–767. doi:10.1016/j.ejor.2021.09.026

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGBoost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, *31*, 24–39. doi:10.1016/j.elerap.2018.08.002

Myburgh, J. C. (2021). *Parametric studies of translation invariance and distortion robustness in convolutional neural networks*. North-West University.

Niu, K., Zhang, Z., Liu, Y., & Li, R. (2020). Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. *Information Sciences*, *536*, 120–134. doi:10.1016/j.ins.2020.05.040

Patel, B., Patil, H., Hembram, J., & Jaswal, S. (2020). Loan default forecasting using data mining. *Proceedings of the 2020 International Conference for Emerging Technology*, 1–4.

Peng, B., Chen, L., Li, J., Jiang, M., Akkas, S., Smirnov, E., Israfilov, R., Khekhnev, S., Nikolaev, A., & Qiu, J. (2019). Harpgbdt: Optimizing gradient boosting decision tree for parallel efficiency. *Proceedings of the 2019 IEEE International Conference on Cluster Computing*, 1–11. doi:10.1109/CLUSTER.2019.8890990

Qian, M., & Hu, F. (2019). An empirical study on prediction of the default risk on P2P lending platform. *Proceedings of the IOP Conference Series: Materials Science and Engineering*, *490*(6), 062048.

Teply, P., & Polena, M. (2020). Best classification algorithms in peer-to-peer lending. *The North American Journal of Economics and Finance*, *51*, 100904. doi:10.1016/j.najef.2019.01.001

Yang, J., Zhao, C., Yu, H., & Chen, H. (2020). Use GBDT to predict the stock market. *Procedia Computer Science*, *174*, 161–171. doi:10.1016/j.procs.2020.06.071

Zhang, H., Bi, Y., Jiang, W., Luo, C., Cao, S., Guo, P., & Zhang, J. (2020). Application of random forest classifier in loan default forecast. *Proceedings of the International Conference on Artificial Intelligence and Security*, 410–420. doi:10.1007/978-981-15-8101-4_37

Zhang, L., Bu, L., Ding, C., Wang, Y., Wang, Y., & Xie, H. (2021). Research on credit risk evaluation and forecast method based on machine learning model. *International Core Journal of Engineering*, *7*(11), 441–448.

Zhang, X. (2022). Research on credit risk forecast model based on data mining technology. *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, 131–136.

Qiliang Zhu is currently a lecturer at the School of Information Engineering, North China University of Water Resource and Electric Power. He received his Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunication in 2018. His research interests include services computing, data mining, and recommender systems.

Wenhao Ding is a graduate student in the School of Information Engineering of North China University of Water Resources and hydropower. He is currently studying for a master's degree. His main research areas include data mining, machine learning and deep neural networks.

Mingsen Xiang is an associate professor of School of Information Engineering, North China University of Water Resource and Electric Power. His principal research interests include information security and network security.

Mengzhen Hu is a graduate student of the School of Electronic Engineering, North China University of Water Resource and Electric Power. She is currently studying for a master's degree. Her main research areas include data mining, text data analysis.

Ning Zhang is a graduate student in the School of Information Engineering, North China University of Water Resource and Electric Power. She is currently studying for a master's degree. Her main research fields includes machine learning and computer vision.