Proceedings of the 2nd International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/87/20240979

Credit Risk Classification Prediction Based on Optimised Adaboost Algorithm with Long Short-Term Memory Neural Network (LSTM)

Qiong Zhang[1,a,*], Chang Zhang[2], Xin Zhao[3]

[1]Cornell University, 616 Thurston Ave, Ithaca, NY, 14853, USA
[2]Northeastern University, 360 Huntington Ave, Boston, MA, 02115, USA
[3]Texas A&M University, 210 Olsen Blvd, College Station, Texas, 77845, USA
a. qz222@cornell.edu
[*]Corresponding author

Abstract: In this paper, the Adaboost algorithm is optimised to classify and predict the user's credit risk by combining the long and short term memory neural network LSTM. The dataset was firstly divided, transposed, normalised, tiled and format converted and then the model was trained and tested. During the training process, it is observed that the loss on the training set gradually decreases and the model gradually optimally fits the data and gradually converges to the optimal solution. The confusion matrix shows that the credit risk of 2914 customers is correctly predicted in the training set with an accuracy of 83.3%. The model performs well on the training set and is able to predict the credit risk of the customers accurately. On the test set, 1211 customers' credit risks were correctly predicted with 80.7% accuracy. Compared to the training set, the prediction on the test set has slightly decreased, but it still copes well with the demand of predicting customers' credit risk. This indicates that the model has some generalisation ability and can achieve better performance on unknown data. Overall, the Adaboost algorithm based on LSTM optimisation proposed in this paper shows high accuracy and reliability in the credit risk classification prediction task. By combining neural networks and traditional machine learning methods, it improves the model's ability to accurately predict the credit risk of customers, providing an effective solution in the financial field.

Keywords: Credit Risk, LSTM, Adaboost

1. Introduction

Credit risk prediction is a crucial work in the financial field, which aims to assess the probability of borrower default, help financial institutions to reasonably formulate lending policies and reduce risks, so as to ensure the stable operation of the financial market [1,2]. In the past, traditional credit risk assessment mainly relied on manual experience and simple statistical analysis methods, which had problems such as difficulty in obtaining information, low efficiency, and susceptibility to subjective factors. With the development of big data and artificial intelligence technology, machine learning algorithms have gradually shown a powerful role in credit risk prediction [3].

DOI: 10.54254/2754-1169/87/20240979

Machine learning algorithms have the following advantages in predicting the credit risk of user credit: Firstly, machine learning algorithms can deal with massive and complex data and mine the laws and features hidden behind the data, improving the accuracy and stability of the prediction model. Secondly, machine learning algorithms can achieve automated modelling and continuous optimization, constantly learning and adapting to new data, and updating models in a timely manner, making credit risk prediction more real-time and flexible [4]. In addition, machine learning algorithms can effectively identify nonlinear relationships, handle high-dimensional data, and have a certain degree of robustness to outliers and noise, and perform well in the complex and changing financial market environment.

Machine learning algorithms commonly applied in credit risk prediction include, but are not limited to, logistic regression [5], decision trees [6], random forests [7], support vector machines [8], and neural networks [9]. These algorithms are trained to identify the association between borrower characteristics and defaults by training models and perform predictive analyses based on historical data. For example, in the field of credit scoring, using these algorithms it is possible to personalise a customer's score and accordingly determine whether to grant a loan and the size of the loan [10].

Machine learning algorithms play an increasingly important role in credit risk prediction, providing financial institutions with more accurate, efficient and reliable decision support. In this paper, the Adaboost algorithm optimised based on the long and short-term memory neural network LSTM is used to classify and predict the user's credit risk, which provides a certain research basis for subsequent studies.

2. Data set sources and data analysis

The dataset used in this paper is selected from the open-source dataset, which records information about the age, income, family, years of employment, intention, amount, interest rate status, percentage of income, years of credit, and default of each user, with a total of 28638 pieces of data, and each piece of data is the borrowing and lending information of one user, which records the user's personal situation as well as the eventual default situation (defaulted or not), and some of the data are shown in Table 1 shows.

Table 1: Selected data sets.

Age	Income	Intent	Amount	Rate	Status	Percent income	Cred length	Default
22	59000	1	35000	16.02	1	0.59	3	1
21	9600	2	1000	11.14	0	0.1	2	2
25	9600	3	5500	12.87	1	0.57	3	2
23	65500	3	35000	15.23	1	0.53	2	2
24	54400	3	35000	14.27	1	0.55	4	1
21	9900	4	2500	7.14	1	0.25	2	2
26	77100	2	35000	12.42	1	0.45	3	2
24	78956	3	35000	11.11	1	0.44	4	2
24	83000	1	35000	8.9	1	0.42	2	2

The data were statistically summarised, including maximum, minimum, mean and standard deviation, and the results are shown in Table 2.

DOI: 10.54254/2754-1169/87/20240979

Table 2: Data statistics.

Variable name	Sample size	Max	Min	Average	Standard deviation
Age	4999	144	20	23.577	3.175
Income	4999	500000	9600	50665.21	35549.295
Emp length	4999	123	0	3.723	3.676
Intent	4999	6	1	3.296	1.676
Amount	4999	35000	500	8381.506	6895.22
Rate	4999	21.74	5.42	11.215	3.196
Status	4999	1	0	0.292	0.455
Percent income	4999	0.83	0.01	0.18	0.122
Cred length	4999	4	2	2.991	0.821
Default	4999	2	1	1.813	0.39

3. Method

3.1. Short- and long-term memory networks

Long Short-Term Memory Network (LSTM) is a special kind of Recurrent Neural Network (RNN) designed to solve the problem of gradient vanishing or gradient explosion that exists in ordinary RNNs.LSTM controls the flow of information by introducing three gating structures: forgetting gates, input gates, and output gates, so as to better capture long term dependencies. Specifically, the forgetting gate is used to control forgetting the memory of the previous moment; the input gate is used to selectively update the memory; and the output gate is used to determine the output of the current moment.

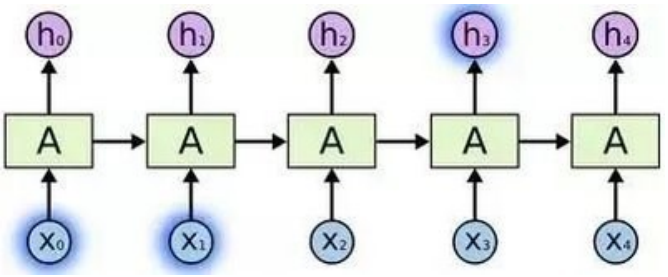

Figure 1: LSTM structure diagram.
(Photo credit: Original)

In LSTM, each time step has a cell state that is responsible for transferring information and is regulated by each gating unit. The forgetting gate will decide which information to discard based on the input data and the output of the previous time step; the input gate will update the cell state based on the current input and the output of the previous time step; and finally, the output gate will produce the final output based on the current input and the output of the previous time step. This design enables LSTM to effectively process long sequential data and achieve excellent performance in a variety of tasks.LSTM is able to effectively capture long-term dependencies through well-designed memory cells and gating structures, and has become one of the important and widely used models in the field of deep learning.

3.2. Adaboost

Adaboost is an integrated learning method that builds a strong classifier by combining multiple weak classifiers. The principle is based on a "boosting" strategy, i.e., iteratively training a series of

DOI: 10.54254/2754-1169/87/20240979

classifiers, each of which tries to correct the errors of the previous one, and eventually combining them to form a stronger model.

The basic principle of Adaboost is that given a training set of N samples, each sample is given a weight, and initially each sample is given an equal weight. Then, in each iteration, Adaboost trains a weak classifier (e.g., a decision stump) and adjusts the weights of the samples based on their performance on the training set. Specifically, for each sample, when computing the next round of the training set, if that sample is correctly classified by the current classifier, its weight is decreased; conversely, its weight is increased. This makes the next classifier pay more attention to the samples misclassified by the previous classifier, thus continuously improving the model performance.

In the final prediction stage, Adaboost combines the results of all the weak classifiers in a weighted combination, where the weight of each weak classifier depends on its performance during training. Typically, the better performing weak classifiers are given higher weights and thus play a greater role in the overall prediction.

One of the advantages of Adaboost is that it is able to handle high-dimensional data and complex feature spaces and is not prone to overfitting. In addition, Adaboost also performs well when dealing with unbalanced datasets due to its strategy of adaptively adjusting sample weights. However, Adaboost also has some drawbacks, such as being sensitive to noisy data and outliers, which may lead to a decrease in model performance when faced with noisy or many outliers.

3.3. Adaboost based on LSTM optimisation

The Adaboost classification algorithm based on the optimisation of Long Short-Term Memory (LSTM) neural network is a method that combines deep learning and integrated learning. By applying the LSTM network to the Adaboost algorithm, the powerful sequence modelling capability of the LSTM network is used to extract the temporal features in the data and combined with the integrated learning advantage of the Adaboost algorithm to achieve effective classification of the sequence data.The LSTM is able to capture the long term dependencies in the data, which improves the classification accuracy and generalization ability; at the same time, the Adaboost algorithm can further improve the overall classification performance by iteratively training multiple weak classifiers and weightedly combining their results. This approach combines the advantages of LSTM network in sequence modelling and Adaboost algorithm in integrated learning, providing an effective and efficient solution for dealing with classification problems with time series features.

4. Experiments and Results

4.1. Data processing

The data processing part includes the process of dividing the dataset, data transposition, data normalisation, data tiling and data format conversion. Firstly, the original dataset is divided into training set and test set according to different categories in the ratio of 7:3 and processed accordingly. Then transpose operation is performed on the training set and test set to get the number of samples. Then the input data is normalised to ensure that the data range is between 0 and 1. The data is flattened into a 1-dimensional array to match the model input requirements. Finally, the processed training and test set data formats are converted into a form suitable for model processing.

DOI: 10.54254/2754-1169/87/20240979

4.2. Parameter setting

Table 3: Parameterisation.

Parametric	Setting
Number of weak regressors	10
Number of hidden layer nodes	6
Gradient descent algorithm	Adam
Maximum number of training sessions	1000
Initial learning rate	0.01
Learning rate decline factor	0.1

In terms of experimental setup, this paper uses matlab R2022a to conduct experiments, the number of weak regressions is set to 10, the number of hidden layer nodes is set to 6, the gradient descent algorithm is set to Adam, the maximum number of training times is set to 1,000, Initial Learning Rate is set to 0. 01, and Learning Rate Decrease Factor is set to 0.1.

4.3. Result

Firstly output the change of loss value during training process of training set, the change of loss value reflects the good or bad result of model training. The credit credit risk dataset is trained using the training set and tested using the test set, and the classification confusion matrices of the training and test sets are outputted respectively, and the results are shown in Fig. 2 and Fig. 3. Finally the accuracy of classification of training and test set is output.

As the training process proceeds, the value of loss gradually becomes smaller, the loss curve gradually decreases, the model gradually optimally fits the data during the training process, and the model gradually converges to the optimal solution.

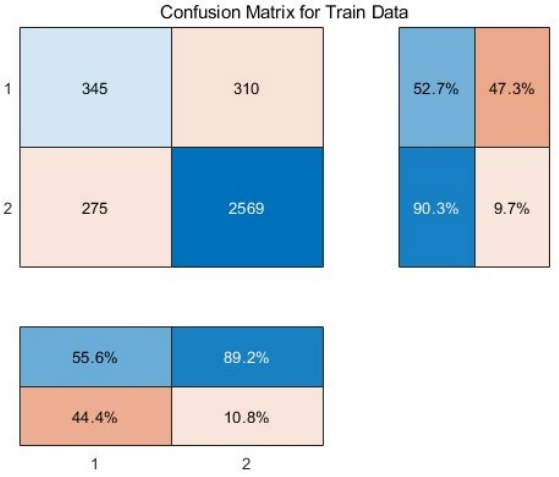

Figure 2: Training set confusion matrix.
(Photo credit : Original)

From the confusion matrix of the training set, 2914 customers' credit risk profiles are predicted correctly and 585 customers' credit risk profiles are predicted incorrectly, with an accuracy of 83.3% in the training set. The prediction of the model in the training set is good, and it can predict the credit risk of the customers based on their individual situation relatively accurately.

DOI: 10.54254/2754-1169/87/20240979

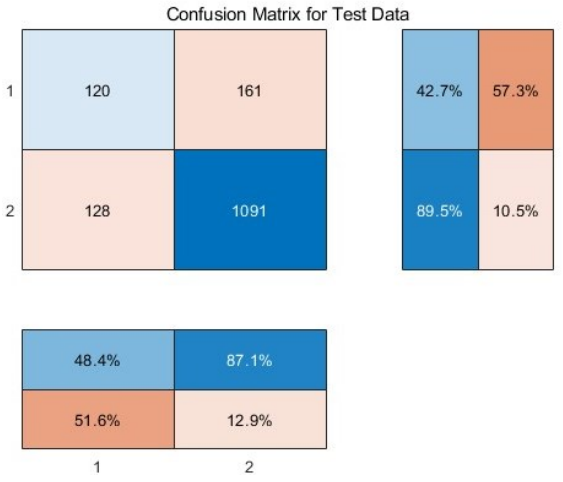

Figure 3: Test Set Confusion Matrix.
(Photo credit : Original)

From the confusion matrix of the test set, there are 1211 customers whose credit risk profile is predicted correctly and 289 customers whose credit risk profile is predicted incorrectly, and the accuracy of the test set is 80.7%. The prediction effect of the model in the test set is 2.6% lower than the training set, the model accuracy does not decrease much and it can predict the credit risk of the customers well.

5. Conclusion

In this study, Adaboost algorithm based on Long Short Term Memory Neural Network (LSTM) optimisation is used to classify and predict the credit risk of users. The dataset was first divided, transposed, normalised, tiled and format converted and then the model was loaded for training and testing. During the training process, it was observed that the loss value of the training set gradually decreases and the model gradually optimally fits the data and converges to the optimal solution. According to the confusion matrix of the training set, 2914 customers' credit risks were correctly predicted with an accuracy of 83.3%. The model performs well on the training set and is able to accurately predict the credit risk of the customers.

A further look at the confusion matrix on the test set shows that 1,211 customers' credit risks were correctly predicted with an accuracy of 80.7%. Although the accuracy on the test set is slightly lower than that on the training set, it still remains at a high level and has only decreased by 2.6% compared to the training set. This indicates that the model has a good generalisation ability and can effectively predict the credit risk profile of customers on unseen data. Therefore, the Adaboost algorithm based on LSTM optimisation proposed in this study performs well in credit risk classification prediction, provides an effective and reliable prediction tool in the financial field, and has positive significance in improving the efficiency of risk management and reducing financial risks.

References

[1] *Md, Abdul Quadir, et al. "Novel optimization approach for stock price forecasting using multi-layered sequential LSTM." Applied Soft Computing 134 (2023): 109830.*

[2] *Yun, Kyung Keun, Sang Won Yoon, and Daehan Won. "Interpretable stock price forecasting model using genetic algorithm-machine learning regressions and best feature subset selection." Expert Systems with Applications 213 (2023): 118803.*

[3] *Mahmoodi, Armin, et al. "A developed stock price forecasting model using support vector machine combined with metaheuristic algorithms." Opsearch 60.1 (2023): 59-86.*

DOI: 10.54254/2754-1169/87/20240979

[4] *Jorgenson, Dale W., et al. "Can neural networks predict stock market?(LON: MSMN Stock Forecast)." AC Investment Research Journal 220.44 (2023).*

[5] *Kożuch, Anna, Dominika Cywicka, and Krzysztof Adamowicz. "A comparison of artificial neural network and time series models for timber price forecasting." Forests 14.2 (2023): 177.*

[6] *Zhang, Wen, et al. "An ensemble dynamic self-learning model for multiscale carbon price forecasting." Energy 263 (2023): 125820.*

[7] *Sonkavde, Gaurang, et al. "Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications." International Journal of Financial Studies 11.3 (2023): 94.*

[8] *Shrotriya, Lalit, et al. "Cryptocurrency algorithmic trading with price forecasting analysis using PowerBI." International Journal of Engineering, Science and Technology 15.4 (2023): 1-8.*

[9] *Karakuş S ,Kaya M,Tuncer A S.Real-Time Detection and Identification of Suspects in Forensic Imagery Using Advanced YOLOv8 Object Recognition Models[J].Traitement du Signal,2023,40(5):*

[10] *Gao, Jie. "Research on stock price forecast based on Arima-GARCH model." MSIEID 2022: Proceedings of the 4th Management Science Informatization and Economic Innovation Development Conference, MSIEID 2022, December 9-11, 2022, Chongqing, China. European Alliance for Innovation, 2023.*