Proceedings of the 2nd International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/88/20241007

Credit Risk Classification and Prediction Based on Deep Neural Network Algorithm

Xiqing Liu[1,a,*]

[1]*Columbia University, 2101 Mission St, San Francisco, CA 94110 USA*
a. liuxiqing1995@gmail.com
**corresponding author*

Abstract: This study predicts whether a user has defaulted based on correlation analysis and deep neural network algorithms. The results of the study show that the occurrence of default by a user is positively correlated with age, family, years of employment, and credit length, and negatively correlated with income, amount, rate, status, and percentage of income. After model training and testing, the prediction accuracy was 81.68% on the training set and 81.68% on the test set. Specifically, there were 2858 correct predictions and 641 incorrect predictions in the training set, of which 469 incorrectly predicted that no default had occurred as having occurred and 172 incorrectly predicted that a default had occurred as not having occurred, and there were also 2858 correct predictions and 641 incorrect predictions in the test set. The results of this study show that the established model has high reliability and accuracy in accurately predicting whether a user has defaulted or not, which provides an important reference for risk assessment and decision-making.

Keywords: Credit Risk, Deep Neural Network, Spearman correlation analysis

1. Introduction

Bank CREDIT RISK FORECASTING is a very important task in the field of finance, which refers to the assessment and prediction of the likelihood of a borrower's default by a bank in the process of lending, so that the bank can better manage the risk, protect the assets, and ensure profitability. Accurate credit risk prediction can help banks avoid non-performing loans, improve the efficiency of loan approval, reduce asset losses, optimise asset allocation, and ensure the stability and healthy development of the financial system.

Traditional credit risk prediction methods are mainly based on statistical models, such as logistic regression and decision trees. However, with the development and application of deep learning technology, deep learning algorithms have also shown strong advantages in credit risk prediction [1].

Firstly, deep learning algorithms can handle large-scale and high-dimensional data. In credit risk prediction, it is usually necessary to consider a variety of characteristic factors, such as personal information, financial status, and historical credit history [2]. Traditional models may not be able to deal with these complex data effectively, while deep learning algorithms are able to automatically learn the complex laws and relationships between features hidden in the data through the neural network structure. Secondly, deep learning algorithms have strong generalisation ability [3]. Traditional models may have limitations when dealing with nonlinear relationships, while deep learning algorithms can better fit complex data distributions through a multi-level neural network

DOI: 10.54254/2754-1169/88/20241007

structure, and can better adapt to new data sets for accurate prediction. In addition, deep learning algorithms can automatically extract features and perform feature combinations, which reduces the need for manual feature engineering to a certain extent, and can further improve model performance by continuously adjusting the network structure and parameters.

In conclusion, deep learning algorithms have great potential and broad application prospects in credit risk prediction. In the future, with the continuous development and improvement of technology, it is believed that deep learning algorithms will play a more and more important role in the financial field and bring more innovation and change to the banking business.

2. Data sources and statistics

The data selected for this paper comes from an open source dataset that contains borrowing and lending data for more than 30,000 customers, and each entry records the customer's age, income, family, years of employment, amount, rate, status, percentage of income, length of credit, and whether or not a default has occurred. Some of the data are shown in Table 1. The data were statistically analysed by calculating the maximum, minimum, mean, standard deviation and median and the results are shown in Table 2.

Table 1: Partial data.

Age	Income	Emp length	Amount	Rate	Status	Percent income	Cred length	Default
24	10980	0	1500	7.29	0	0.14	3	2
22	80000	3	33950	14.54	1	0.42	4	1
24	67746	8	33000	12.68	1	0.49	3	2
21	11000	3	4575	17.74	1	0.42	3	1
23	11000	0	1400	9.32	0	0.13	3	2
24	65000	6	32500	9.99	1	0.5	3	2
21	11389	5	4000	12.84	1	0.35	2	1
21	11520	5	2000	11.12	1	0.17	3	2
25	120000	2	32000	6.62	0	0.27	2	2
26	95000	7	31050	14.17	1	0.33	3	1

Table 2: Data statistics.

Variable Name	Max	Min	Mean	Standard	Median
Age	144	20	23.577	3.175	23
Income	500000	9600	50665.21	35549.295	40872
Home	4	1	1.584	0.876	1
Emp length	123	0	3.723	3.676	3
Amount	35000	500	8381.506	6895.22	5000
Rate	21.74	5.42	11.215	3.196	11.14
Status	1	0	0.292	0.455	0
Percent income	0.83	0.01	0.18	0.122	0.15
Cred length	4	2	2.991	0.821	3
Default	2	1	1.813	0.39	2

DOI: 10.54254/2754-1169/88/20241007

3. Relevance analysis

In order to explore the correlation between a customer's age, income, family, years of employment, amount, rate, status, percentage of income, and length of credit and whether or not a default occurs, this paper performs a Spearman correlation analysis of each variable, which is a non-parametric statistical method for measuring the degree of correlation between two variables. Unlike the Pearson correlation coefficient, Spearman correlation analysis does not require a linear relationship between the variables, but rather compares the variables based on their rank order. The Spearman rank correlation coefficient is calculated by comparing the ranks of each pair of data points [4]. The value of this coefficient ranges from -1 to 1, with 0 indicating no correlation, 1 indicating a perfect positive correlation, and -1 indicating a perfect negative correlation. The results of the correlation heat map are shown in Figure 1.

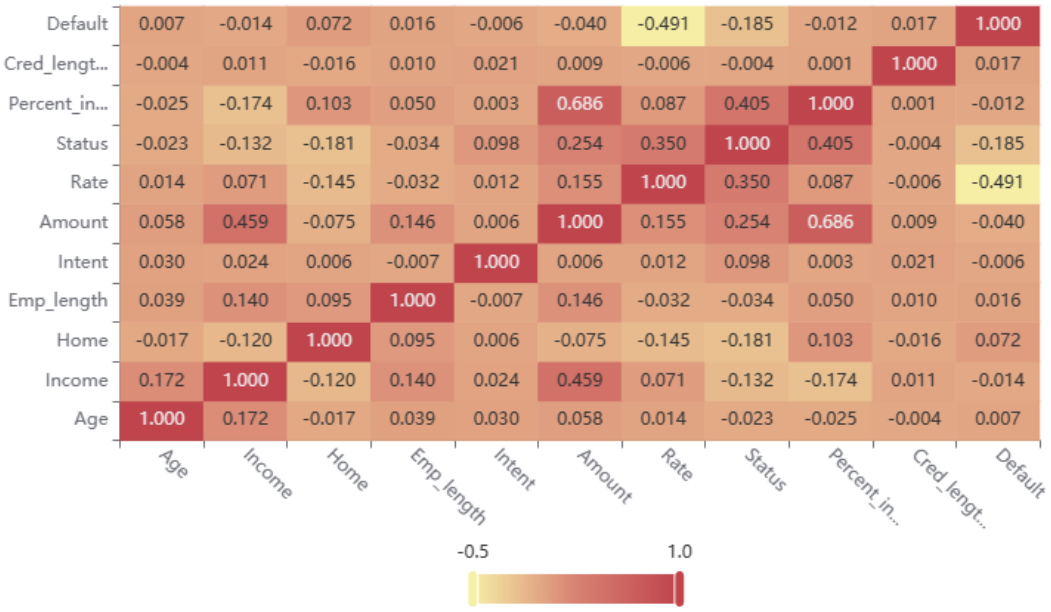

Figure 1: Correlation heat map.
(Photo credit: Original)

Deep Neural Network (DNN) is a machine learning model based on artificial neural networks with a structure consisting of multiple layers of neurons.The principle of DNN is based on the connection and information transfer between neurons, and extracts the abstract features in the data by means of multiple hidden layers, so as to achieve the learning and recognition of complex data patterns [5,6]. The model structure of DNN is shown in Fig. 2.

DOI: 10.54254/2754-1169/88/20241007

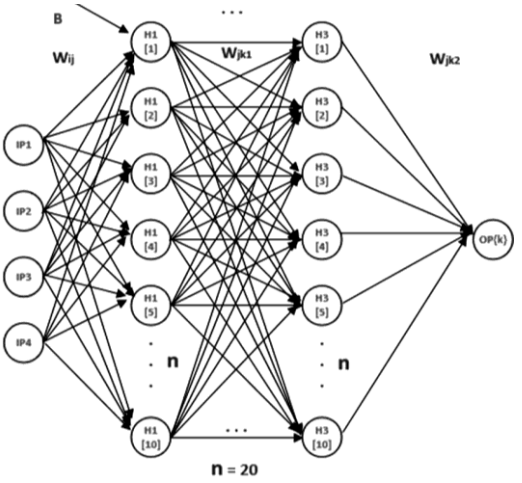

Figure 2: The model structure of DNN.
(Photo credit: Original)

DNN consists of an input layer, a hidden layer and an output layer. The input layer receives raw data as input, the hidden layer is responsible for feature extraction and transformation of the data, and the output layer generates the final prediction [7]. Each neuron has weights and bias parameters that are used to adjust the influence of the input signal as well as to introduce nonlinear transformations.

In DNN, information transfer is achieved through two processes: forward propagation and back propagation. Forward propagation refers to calculating the output value of each neuron layer by layer from the input layer to the output layer and using it as input for the next layer. The hidden layer introduces a nonlinear transformation through the activation function, enabling the network to learn complex data patterns [8,9].

Backpropagation is a crucial step in the DNN training process by calculating the gradient of the loss function with respect to the weights and bias parameters and using optimisation algorithms such as gradient descent to update the parameters to minimise the loss function. In this way, the continuous iterative tuning of parameters on the training set allows the network to gradually optimise to learn better feature representations and improve prediction accuracy [10].

The ability of DNNs to effectively handle complex data patterns is mainly due to the feature representation capability brought about by their deep structure. By stacking multiple hidden layers together, the network can represent data features in a level-by-level abstraction, gradually extracting more abstract and complex feature information from low to high levels.

Deep Neural Networks (DNNs) are artificial neural networks that mimic the structure of the human brain through multiple neural network layers to learn and understand complex data patterns. In terms of predicting whether a user will default or not, a deep neural network predicts whether a user will default or not by analysing the correlation analysis of each variable as well as deep neural network algorithms. In specific application scenarios, each variable is analysed, such as age, family situation, years of employment, length of credit, income, and expenditure ratio, to understand its relationship with default.

4. Experimental setup

Firstly, on the dataset division, the dataset is divided into training set and test set according to the ratio of 7:3, the training set is used for model training and the test set is used for testing after the model training is finished, and the accuracy is used to evaluate the goodness of the model prediction effect.

DOI: 10.54254/2754-1169/88/20241007

The deep neural network (DNN) model created in this paper consists of an input layer, three hidden layers and an output layer. Among them, the hidden layers are three layers each containing 10 nodes.

In each layer of the model, different types of neural network layers are used, sequenceInputLayer is used to define the format of the input data, fullyConnectedLayer is the fully connected layer, reluLayer is the activation function ReLU layer, softmaxLayer is the Softmax normalised output layer, the classificationLayer is the classification output layer.

The Adam optimisation algorithm is used for training with a maximum number of 1000 training sessions, 128 samples are used for gradient estimation in each iteration, the initial learning rate is 0.01, the L2 regularisation parameter is 1e-4, the dataset is disrupted in each round of training, and the option of displaying a chart of the training progress is made available during the training process.

5. Results

At the end of training, the test accuracy of the test set of the model is output and the predicted confusion matrices of the training and test sets are output and the results are shown in Figures 3 and 4.

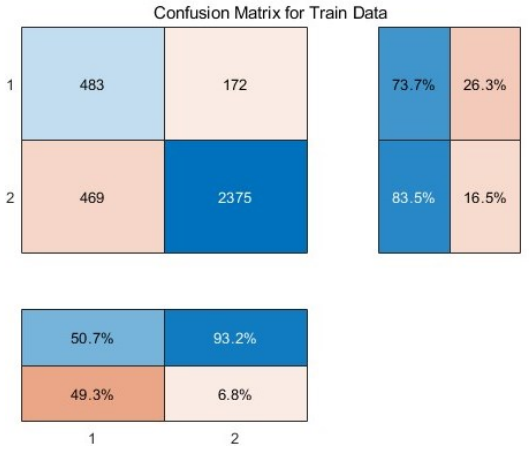

Figure 3: Confusion matrix.
(Photo credit: Original)

From the prediction confusion matrix of the training set, a total of 2858 predictions were correct and 641 predictions were incorrect in the prediction of credit status (default or not), of which 469 should have been predicted as not defaulting but were predicted as defaulting, and 172 should have been predicted as defaulting but were predicted as not defaulting, giving a prediction accuracy of 81.68%.

DOI: 10.54254/2754-1169/88/20241007

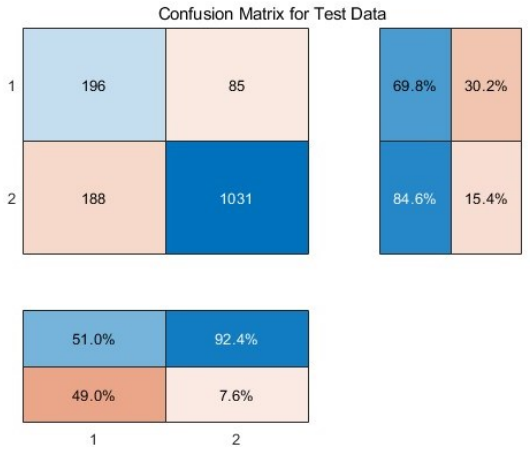

Figure 4: Confusion matrix.
(Photo credit: Original)

From the prediction confusion matrix for the test set, it can be seen that a total of 1,227 predictions were correct and 273 predictions were incorrect on the prediction of credit status (whether or not to default), of which 188 should have been predicted as not defaulting but were predicted as defaulting, and 85 should have been predicted as defaulting but were predicted not to be defaulting, giving a prediction accuracy rate of 81.8 per cent.

6. Conclusion

In this paper, based on correlation analysis and deep neural network algorithms to predict whether a user has defaulted or not, correlation analysis of each variable shows that the occurrence of default by a user is positively correlated with age, family, years of employment and credit length, and negatively correlated with income, amount, rate, status, and percentage of income. After the introduction of model training and testing, the results showed that there were a total of 2858 correct predictions and 641 incorrect predictions on the training set, of which 469 should have been predicted as no default but were predicted to occur, and 172 should have been predicted to occur but were predicted not to occur, resulting in a prediction accuracy of 81.68%. There were a total of 2858 correct predictions and 641 incorrect predictions on the test set, of which 469 should have been predicted as not having defaulted but were predicted to have defaulted, and 172 should have been predicted to have defaulted but were predicted not to have defaulted, giving a prediction accuracy of 81.68%. This study successfully predicts whether a user will default or not through a deep neural network algorithm, and also reveals the correlation between each variable and default. Our model demonstrated high accuracy on both the training and test sets, providing important reference information for financial institutions and helping to develop more effective risk management strategies and credit decisions.

References

[1] *Bhatt, T. K., Ahmed, N., Iqbal, M. B., & Ullah, M. (2023). Examining the determinants of credit risk management and their relationship with the performance of commercial banks in Nepal. Journal of risk and financial management, 16(4), 235.*

[2] *Rao, C., Liu, Y., & Goh, M. (2023). Credit risk assessment mechanism of personal auto loan based on PSO-XGBoost Model. Complex & Intelligent Systems, 9(2), 1391-1414.*

[3] *Li, Z., Liang, S., Pan, X., & Pang, M. (2024). Credit risk prediction based on loan profit: Evidence from Chinese SMEs. Research in International Business and Finance, 67, 102155.*

DOI: 10.54254/2754-1169/88/20241007

[4] *Khan, N., Ramzan, M., Kousar, T., & Shafiq, M. A. (2023). Impact of bank specific factors on credit risk: Evidence from Islamic and conventional banks of Pakistan. Pakistan Journal of Humanities and Social Sciences, 11(1), 580-592.*

[5] *Gilchrist, S., Wei, B., Yue, V. Z., & Zakrajšek, E. (2024). The Fed takes on corporate credit risk: An analysis of the efficacy of the SMCCF. Journal of Monetary Economics, 103573.*

[6] *Al-Qudah, A. A., Hamdan, A., Al-Okaily, M., & Alhaddad, L. (2023). The impact of green lending on credit risk: Evidence from UAE's banks. Environmental Science and Pollution Research, 30(22), 61381-61393.*

[7] *Halim, M. A., Moudud-Ul-Huq, S., Sobhani, F. A., Karim, Z., & Nesa, Z. (2023). The Nexus of Banks' Competition, Ownership Structure, and Economic Growth on Credit Risk and Financial Stability. Economies, 11(8), 203.*

[8] *Yfanti, S., Karanasos, M., Zopounidis, C., & Christopoulos, A. (2023). Corporate credit risk counter-cyclical interdependence: A systematic analysis of cross-border and cross-sector correlation dynamics. European Journal of Operational Research, 304(2), 813-831.*

[9] *Duho, K. C. T., Duho, D. M., & Forson, J. A. (2023). Impact of income diversification strategy on credit risk and market risk among microfinance institutions. Journal of Economic and Administrative Sciences, 39(2), 523-546.*

[10] *Wang F ,Lin Y ,Xu J , et al.Risk of papillary thyroid carcinoma and nodular goiter associated with exposure to semi-volatile organic compounds: A multi-pollutant assessment based on machine learning algorithms[J].Science of the Total Environment,2024,915169962-.*