Banking Credit Risk Analysis using Artificial Neural Network

Charles Maruma, Chunling Tu, and Claude Nawej

Abstract Banking credit risk analysis is a form of evaluation conducted by financial institutions to determine applicants' ability to repay their debt obligation. Financial institutions, such as banks, set objectives to offer credit to creditworthy customers, after spending time trying to evaluate their repaying capacity. In this paper, we propose a credit risk analysis system based on an artificial neural network (ANN) to identify customers who will default. A feedforward propagation algorithm is used to train the model consisting of three layers. Data pre-processing is performed to clean the datasets and check for missing variables. The datasets were normalized using min–max normalization to get the correlation among the variables. The datasets are applied to the proposed model and logistic regression models, and the comparison shows the proposed model which has a better performance.

Keywords Credit risk analysis · Artificial neural network · Logistic regression · Default · Credit · And algorithm

1 Introduction

The banking industry focuses mainly on offering credit/loans [1]. Although the banking industry does offer other services such as policies and insurances. Since banks offer credit as their primary service offerings, they might suffer a great financial loss if customers fail to pay their credit back [2]. For banks to avoid financial loss as a result of customers defaulting on credit, banks must perform a credit risk analysis to predict and minimize the risk by discriminating between customers with a high risk of default and customers with a low risk of default. Credit risk analysis

C. Maruma (✉) · C. Tu · C. Nawej
Tshwane University of Technology, Pretoria, South Africa
e-mail: Charlesmaruma1@gmail.com

C. Tu
e-mail: Duc@tut.ac.za

C. Nawej
e-mail: NawejMC@tut.ac.za

X.-S. Yang et al. (eds.), *Proceedings of Seventh International Congress on Information and Communication Technology*, Lecture Notes in Networks and Systems 447,
https://doi.org/10.1007/978-981-19-1607-6_76

is a process used to predict whether or not a customer will default on a loan or credit [3]. Credit applications for consumer or commercial loans are being processed and evaluated using credit risk analysis before approving or denying them. Credit risk analysis is usually performed based on historical data such as past transactions, credit history, and other relevant information. The decision to approve or deny credit is based on customer's risk of default; if the customer has a high risk of default, then the credit will be denied, however, if a customer has a low risk of default, the credit will be approved [4]. Credit scoring is a method used to predict and analyze credit risk. Credit scoring is categorized into two different scoring types, such as behavioral scoring and application scoring. Behavioral scoring is usually performed after the credit has been approved, it monitors how the customer repays the credit [5]. Application scoring is performed before approving or denying credit, based on the customer's credit history. When a customer applies for a loan, his/ her credit information is obtained from the credit bureau. The credit information includes features such as employment status, past bank transactions, salary, and credit history [1].

Credit scoring is a technique used to analyze and predict credit risk. When a customer applies for credit, a credit risk analysis is performed before approving or denying the credit based on the customer's credit history and other relevant information [6]. This process is performed to minimize and manage the risk of customers faulting on loans. Bank manages the flow of money between investors and borrowers. Investors put money in the bank, then the bank borrows that money to customers/borrowers in the form of credit/loans, then the bank manages the risk on how the borrower pays back the loan with interest [7]. When the borrower pays back the loan, the money is put back into the investor's account with a small portion of the interest, and the other portion of the interest goes to the bank for managing the risk. The study of credit risk analysis is the most researched area in the banking industry. The credit risk analysis technique plays an important part in the banking industry, particularly for big banks with large data that is hard to work with and process. Credit risk analysis is a form of evaluation process conducted by financial analysts to determine applicants' ability to repay their debt obligation. After a person or a company applies for credit at the bank, the bank processes and evaluates the costs and profit related to the credit [8]. The credit risk analysis model is utilized to predict the costs and risks related to the credit. The objective of this study is to address credit risk analysis issues by applying an ANN to credit risk models [9]. ANN algorithm is applied to credit risk/loan application datasets of public data to predict loan risk. The credit risk model produces the output of "1" or "0" to predict whether or not the customer will default.

2 Literature Review

Credit risk analysis using historical data predicts how different customer characteristics determine whether or not the customer will be able to repay the loan. This method uses a score to categorize customers according to their risk of default on credit. The

customer with a high score has a low risk of defaulting on the loan and he/she will be able to repay the loan, while the customer with a low score has a high risk of defaulting on the loan and will not be able to repay the loan. Normally, logistic regression and discriminant analysis were the most used machine learning algorithms in credit risk analysis models. The first machine learning algorithm applied to the credit scoring model was discriminant analysis. The use or application of linear discriminant analysis has regularly been disapproved because of how it deals with categorical data of datasets, and the classification of creditworthy and not creditworthy classes is not likely to be accurate. Another machine learning algorithm used in credit risk analysis is logistic regression. Logistic regression is used as a machine learning algorithm of choice in predicting customers who will or will not default on their credit.

Artificial neural network (ANN) algorithm in artificial intelligence is a data modeling technique that is similar to the brain of the human and nervous system, and it works similarly to how the human brain works [10]. ANN has a network of inter-connected neurons to find the functionality of the model. A few tests were performed on different machine learning algorithms to measure the accuracy of the individual models. During the performance testing process, it was discovered that artificial neural networks performed better and produced more accurate results when compared to logistic regression [11]. Normally for the artificial neural network to perform results prediction, it entails being trained on the input and target variables of the given datasets. There are many successful real-world applications of artificial neural networks such as edited file detectors (checks if a file has been modified), unusual banking transaction detectors, and other predicting technologies. The application of machine learning algorithms to credit risk has improved the performance of credit risk analysis [12].

Discriminant analysis technique was considered the foremost common technique for developing customer credit risk analysis models [10]. Even though the discriminant analysis technique has been criticized by the analysts because of the way, it processes and handles datasets of categorical variables to predict customers with a high risk of default and customers with a low risk of default. In machine learning, the neural network algorithm is a nonlinear technique that provides a new alternative to linear methods, especially in situations where the dataset has more composite relationships between the independence of the nonlinear variables [13]. An artificial neural network (ANN) is a machine learning algorithm that develops a relationship between the independent variables and dependent variables, in consideration that there is a correlation among the variables. ANNs are artificial intelligence algorithms that mimic the structure of the human brain and work similarly to the nervous system [14]. An artificial neural network is consists of a network of neurons organized in a matrix form. Neurons are associated by joins with related weights which decide how data are being processed [15].

The credit risk analysis model uses a feedforward neural network [16]. Feedforward neural network is an ANN technique. The inter-connected links between nodes in the feedforward neural network do not a form cycle [17]. In feedforward neural networks, input variables enter the model via the input layer, and the variables are multiplied by the weights. The values of each variable are added to get the total sum

of the input variables. We get an output of 1 if the sum of the values is over a given threshold, however, if the sum of the values is below a specific threshold, then less than 1 is a product at the output.

3 Methodology

In this paper, ANN is used to build a credit risk analysis system. Publicly accessible datasets obtained from the Internet are applied to train the model. The dataset consists of about 1000 customer records and 10 categorical and numerical variables.

The independent variables are

- Age: Age of the borrower
- Sex: Gender of the borrower (male or female).
- Job: Employment status of the borrower (employed or unemployed).
- Housing: Checks if the borrower owns or rents a property.
- Saving accounts: Banking history of savings account.
- Checking account: Banking history of a cheque account.
- Credit amount: Loan amount given to the borrower.
- Duration: The term in months the borrower will settle the loan amount.
- Purpose: Reason for taking a loan.

The dependent variable is the risk represented with 1 or 0. The risk predicts whether the customer will be able to repay the full loan amount on time. If the customer's risk prediction is "1", it means that the customer will default on the loan and will not be able to pay back the loan otherwise, it is a creditworthy customer.

The first step is variable-processing. Usually, variables in the dataset do not come in a way that can be directly used. Pre-processing is therefore needed. Categorical variables have labels instead of numbers. For example, the gender has "female" or "male" labels. Numerical variables have numerical values. In datasets, categorical variables could have a lot of null values. For this situation, it is a significant loss if the concerned data are discarded. So we can replace those null values with random labels. The other important pre-processing to improve the performance of the model is data normalization. Normalization is carried out to get the correlation of the data. In this paper, min–max normalization techniques are used to normalize input/independent variables between 0 and 1 as shown in Eq. (1).

$$Z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where $x = (x_1, x_2, \ldots, x_n)$ and Z_i are the i-th normalized variable data.

Target/independent variable is normalized between 0 and 1, as shown in Eq. (2).

$$y2 = \log(1 + y) \tag{2}$$

The ANN-based credit evaluation system architecture is set up as follows: the feedforward propagation network is used as shown in Fig. 1. There are three layers in the network: 1 input layer with 9 independent variables, 1 hidden layer with 10 neurons, and 1 output layer with 1 dependent variable representing if the customer is creditworthy or not. The random weight/bias rule is used as the training function to train the neural network. The feedforward propagation algorithm makes the neurons perform better by reducing the error between the actual and the desired results to the least possible amount.

The root mean squared shown in Eq. (3) is used as the training error of the ANN.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2} \tag{3}$$

where n is the number of observations, O_i is the observations, and S_i is the predicted values.

To reduce the error of neural networks, weights must be adjusted by a small amount. Choosing the correct parameters is crucial, especially for the learning coefficient and the number of hidden neurons. Based on the proposed neuron network structure as depicted in Fig. 1, neurons are represented by nodes in each layer and the lines between them weight. The RSME is reduced by retraining the model or it

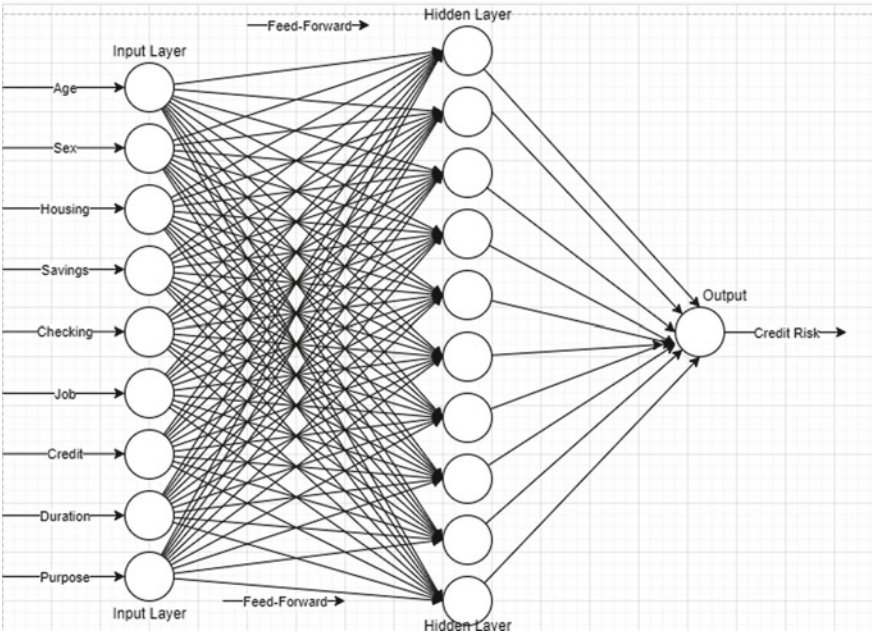

Fig. 1 Feedforward ANN

can be reduced by adjusting the settings on ANN such as reducing the number of neurons.

In this paper, artificial neuron network has been applied to the datasets. Feedforward networks consist of 3 layers, namely input, hidden, and output layer. The dataset is divided into 3 categories; training, test, and validation sets. During the training stage, all variables are tested 1 by 1, to check if they improved the performance of the model. If variables do not improve the performance of the model, they are discarded. The most important variables remain in the dataset. The model was trained until it produced better results.

4 Results

The dataset with 1000 records was fed into ANN with the following parameters and properties:

- Network type: feedforward back prop as shown in Fig. 1
- Datasets division method: random
- Training function: random weight/bias rule
- Adoption learning function: lean GDM
- Performance function: root mean squared error (RMSE)
- Number of hidden layer neurons: 10
- Transfer function: hyperbolic tangent sigmoid

The output error of the model is 0.02 for the trained ANN. The accuracy of the neural network model is compared to the logistic regression model. As per the comparison, neural network performed better than logistic regression, as the error is shown in Table 1. The output values for logistic regression range between 0 and 1.

Table 1 shows that the RMSE for the proposed model is smaller than logistic regression, which means the proposed system is more accurate to predict the risk levels.

The proposed model has a nonlinear activation layer which makes input variables have a nonlinear impact on the credit risk status as the weights have a generalized weight of more than 1. The 9 input variables have been normalized to get the correlation between them before they were added to the neural network. The output of the neural network model is classified as 0 or 1. Data cleaning and pre-processing were performed on the datasets. The datasets were pre-processed to ensure that there were no missing values.

Table 1 Error comparison for proposed system and logistic regression

Proposed model	Logistic regression
0.021005	0.032532

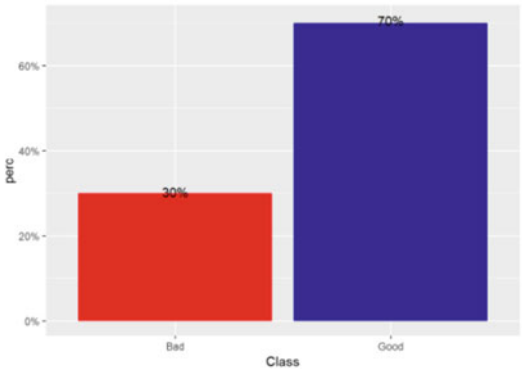

Fig. 2 Total customer credit risk status

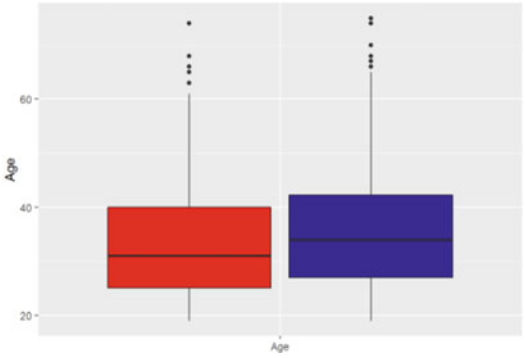

Fig. 3 Age groups by credit risk status

For the total customers in the dataset used in this paper, there are approximately 70% of customers are creditworthy and have a low risk of default, while approximately 30% are at high risk, as shown in Fig. 2, where blue color represents low risk while red color represents a high risk.

In Fig. 3, the age groups show different risk levels, red color shows that customers below 40 years old have a higher risk of default, while the blue color shows the group above 40 years old age has a lower risk of default.

5 Conclusion

In this paper, we studied banking credit risk analysis using ANN and logistic regression algorithms to detect whether customers are likely to default on their credit, based on the given customer information on the dataset. The data were firstly cleaned by a pre-processing stage, to fill in missing values and handle exceptions. The correlations among the independent/input variables are detected. The independent/input variables have been normalized using min–max normalization.

The proposed ANN-based system was compared with the logistic regression model. The experiment results show that the proposed method performed better than logistic regression.

References

1. Lavrushin O, Sokolinskaya N (2020) Confidence level and credit risk analysis in Russian banks. Banks and Bank Syst 15(2):38–46
2. Shan Y (2017) Systemic risk and credit risk in bank loan portfolios. SSRN Electronic J
3. Livshits I (2015) Recent developments in consumer credit and default literature. J Econ Surv 29(4):594–613
4. Singh M, Dixit G (2018) Modeling customers credit worthiness using enhanced ensemble model. Int J Comp Sci Eng 6(7):1466–1470
5. Giannopoulos V (2018) The effectiveness of artificial credit scoring models in predicting NPLs using micro accounting data. J Accounting & Marketing 7(04)
6. Miroshnychenko I, Ivliieva K (2019) Assessing credit risk using machine learning methods. Efektyvna Ekonomika (12)
7. Aslam M, Kumar S, Sorooshian S (2019) Predicting likelihood for loan default among bank borrowers. Int J Financial Res 11(1):318
8. Amat O, Manini R, AntónRenart M (2017) Credit concession through credit scoring: Analysis and application proposal Intangible Cap 13(1):51
9. Demma C (2017) Credit scoring and the quality of business credit during the crisis. Econ Notes 46(2):269–306
10. Ettensperger F (2019) Comparing supervised learning algorithms and artificial neural networks for conflict prediction: performance and applicability of deep learning in the field. Qual Quant 54(2):567–601
11. Maler L (2020) Neural networks: how a multi-layer network learns to disentangle exogenous from self-generated signals. Curr Biol 30(5):R224–R226
12. Abkowitz M, Camp J (2017) Structuring an enterprise risk assessment protocol: traditional practice and new methods. Risk Manag Insurance Rev 20(1):79–97
13. Shah S, Ng J (2020) Hands-on artificial intelligence for banking. Packt Publishing, Limited, Birmingham
14. Kang E, Baek S (2019) Humanistic brain that artificial intelligence can't mimic and artificial intelligence challenging human ambivalence (creativity and limitation). J Contemp Psychoanalysis 21(2):143–154
15. Bondarenko A, Borisov A, Alekseeva L (2015) Neurons vs weights pruning in artificial neural networks. Environment. Technology. Resources. Proceedings of the International Scientific and Practical Conference, 3, p.22.
16. Amardeep R (2017) Training feed forward neural network with backpropogation algorithm. Int J Eng Comp Sci
17. Fuangkhon P (2021) Normalized data barrier amplifier for feed-forward neural network. Neural Net World 31(2):125–157