Credit Risk of Bank Customers can be Predicted from Customer's Attribute using Neural Network

Subrata Saha
Department of Information and Communication Technology
Mawlana Bhashani Science and Technology University, Santosh, Tangail, Bangladesh

Sajjad Waheed
Department of Information and Communication Technology
Mawlana Bhashani Science and Technology University, Santosh, Tangail, Bangladesh

ABSTRACT
The aim of this paper is to present a model based on Multi-layer perceptron neural networks to recognize bad or good credit customers. Credit risk is one of the major problems in banking sector. Banks are faced with credit Risk while doing their tasks. Credit risk is the probability of non-repayment of bank loan granted to lenders. Decreasing Credit Risk, banks may perform better duties and responsibilities successfully for the economic growth of the country. This study will help for a banker to select a right borrower for investing bank fund and hereby may reduce non-performing loan. Artificial neural network is used for loan applicants' credit risk measurement and the calculations have been done by using SPSS and WEKA software. Number of samples was 101 and 12 variables were used to identify good customers from bad customers. The results showed that, History of borrower (Defaulter or non-defaulter), amount of loan, type of collateral security (physical assets or financial assets) and Value of collateral security had most important effect in identifying classification criteria of good and bad borrowers. The main contribution of this paper is specifying for credit rating of bank customers in Bangladesh's banking sector.

Keywords
Credit risk, neural network, multilayer perceptron, Bank credit Customers

1. INTRODUCTION
The main job of commercial banks is to collect deposits of real and legal bank customers and also allocate them to borrowers. Banks are responsible for relationship between lenders and borrowers. Loan applications are evaluated by subjective judgmental assessment of the credit officer or through the use of various statistical (classification) techniques. That is inefficient and non-uniform. Therefore, a neural network or data mining tool is needed to assist in decision making regarding the loan application. Financial system efficiency plays an important role in economic development of Countries.

Banks face problems such as "credit risk". Credit Risk is the probability of non-repayment of received loans at the due date. Credit risk is widely studied topic in bank lending decisions and profitability [1]. Borrowers usually have better information about the projects to be financed, but lenders usually don't have sufficient information about those projects [2].

Before giving a credit loan to borrowers, bank decides who is bad or good Customer. The prediction of customer status i.e. in future borrower will be defaulter or non-defaulter is a challenging task for bank.

If a bank gets some good customers, definitely Bank will have more power in lending loans and increasing in profit will occur. But if bank gets bad customers who will not repay loans in due date, then decreasing in profit will occur.

Risk analysis in today's financial markets is one of important factor that could be applied with neural networks [3]. Artificial Neural Networks (ANN) plays an important role in financial applications for such tasks as pattern recognition, classification, and time series forecasting. Factors affecting the presence of credit risk are divided into two groups: within the organization and outside the organization.

i. Factors outside the organization cannot be controlled by bank management teams and are considered exogenous for bank. Political changes, earthquakes, war and etc. are included in this category.

ii. Factors within the organization are those which bank management teams could take the matter under their control. These are called endogenous factors.

In this research we are trying to figure out the endogenous factors affecting credit risk within an organization. The main purpose of this paper is to identify and rank the factors that have effects on credit risk in one of the commercial banks in Bangladesh. This research can estimate the credit risk of each customer and helps to make the right decision toward granting of loan to customers.

ANN model is used in this research to achieve the above-mentioned. The ANN models use the same input and output parameters as in the linear models. These models have three primary components: the input data layer, the hidden layer and the output layer. Each of these layers contains nodes, and these nodes are connected to nodes at adjacent layer(s).

The hidden layer(s) contain two processes: the weighted summation functions and the transformation function. Both of these functions relate the values from the input data to the output measures. The weighted summation function is typically used in a feed forward/back propagation neural network model.

2. LITERATURE REVIEW
Credit evaluation is one of the most crucial activities in lending [1]. It is the process by which lenders identify loan applicants who have potential to default. Credit evaluation can be undertaken primitively through the subjective judgmental assessment of the credit officer or through the use of various statistical (classification) techniques also known as credit scoring models [4]. Credit rating is one of technical factor in credit risk evaluation. The aim of credit rating is to categorize the applicants into two groups; applicants with good credit and applicants with bad credit. ANN models have a high

predictive power. This means that the networks are capable of adapting to arbitrary and unknown functional forms, with an arbitrarily specified degree of precision. Neural network and logistic regression in forecasting customer credit risk have same efficiency [5]. The important roles of ANN in financial application are pattern recognition, classification and time series forecasting [6].

2.1 ANN Model

An ANN is a mathematical model inspired by biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In this context, a neuron is the basic computation unit. In these units a series of mathematical operations are developed. Afterwards they decide the next step in the computation pathway depending on the results obtained, which is called the activation function.

In this study we have used a Multilayer Perception (MLP) network [7, 8, 9] which as a special sort of ANN comprises architecture of several layers of neurons.

In our case, which is very common, we chose three layers (where each one is fully connected to the next one): an input layer that receives external inputs, one hidden layer, and an output layer which, normally, and it also happens in our case, it generates the classification results [see Figure 1].

After the first level of neurons (in the input layer) the rest of neurons form computational elements with a nonlinear activation function. In order to summarize the MLP functioning, the strategy of the network is that when data are presented in the input layer, the remainder neurons run calculations in the consecutive layers until an output value is achieved for the output neurons which will specify the correct class for the input data.

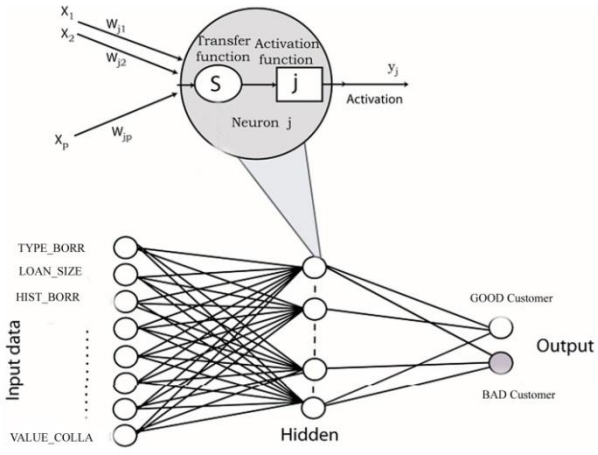

Figure 1: Architecture of the MLP network (input layer, hidden layer and output layer). For a single neuron in the hidden layer it is shown the computation topology (x_i are inputs value; w_{ji} is the connection weight between the input $x_i$and the neuron j; θ_j is the threshold; f_j is the activation function; y_j is the output).

Figure 1 indicates how the hidden layers neurons compute weighted sums of their inputs and add a threshold. The resulting sums are used to calculate the activity of the neurons by applying a sigmoid activation function.

This Process is defined as follows:

$$v_j = \sum_{i=1}^{p} w_{ji} x_i + \theta_j , y_j = f_j(v_j)$$

Wherev_j is the linear combination of inputs $x_1 x_2 x_3 \ldots . x_p$ and the threshold v_j, w_{ji} is the connection weight between the input $x_i$and the neuron j, andf_j is the activation function of the jth neuron, and y_j is the output. The sigmoid function is a common choice of activation function. It is defined as:

$$f(t) = \frac{1}{1 + e^{-t}}$$

A single neuron in the MLP is able to linearly separate its input space into two subspaces by a hyperplane defined by the weights and the threshold. The weights define the direction of this hyperplane whereas the threshold term θj offsets it from origo.

This is essentially what it does MLP very suitable for the classification problems. The MLP network uses the back propagation algorithm [10], which is the most suitable algorithm for similar works [11, 12].

The different layers of the architecture MLP is assembled as follows:

a) Layer 1: It is built automatically from the input vector.

b) Layer 2: To decide the number of hidden neurons for this layer is the hardest issue in the network's architecture. This number represents the equilibrium between a good accuracy and the possibility of over fitting. In the matter of fact, the precise number of neurons in a hidden layer will improve the capacity of the network of generalization from new data notably [13].

c) Layer 3 is the output (Classification) layer. The output corresponds to the different problem classes. In our example there are two outputs, good or bad Customer.

The weights and the threshold of the MLP are calculated during an adaptation process. In the results section the number of hidden neurons of the MLP will be established.

All the computations of the neural network was developed using the WEKA software [14]. WEKA includes most of the machine learning methods and it is Open Source software developed by the University of Waikato in New Zealand.

3. METHODOLOGY

3.1 Data and Variables

Data has been collected over the period 2015-16, based on documents and records of applicants for a Bangladeshi commercial bank. Sample 101 are selected as the number of samples, which are derived from individual customers' profiles.

The variables are as follows:

i. Dependent variable: good and bad customers; in this study we select and realize good or bad customer. Here, good customer is a person who repays his/her loan at the due date and bad customer is a person who don't repay his or her Loan at the due date. We differentiate between good and bad customers in our neural network model calculations by assigning 0 to identify good customers and 1 to identify bad customers.

ii. Independent variables: In this study, 12 variables are defined as independent variables:

1. Type of Borrower (TYPE_BORR): data samples are divided into two groups, Such as Individual or group/private Company.

2. Relationship with Borrower (RELAT_BORR): Borrower's relationship with the bank in years.

3. History of customer (HIST_BORR): Borrower default history, HIST_BORR=1 if customer was defaulter or 0 if non-defaulter.

4. Gender (GENDER_BORR): data samples are divided to female and male according to their gender.

5. Amount of loan (LOAN_SIZE): amount of money that is given to the customer.

6. Time period of loan (LOAN_DURA): Time period of loan in Months.

7. Loan Supervision (SUPERV_LOAN): Loan supervises and monitor by Lender.

8. Type of loan (TYPE_LOAN): data samples are divided into three groups Cash Credit, Agriculture Loan and SOD-Secured Over Draft.

9. Loan recovery Status (RECOV_STATUS): Loan Recovery Status is divided into five: UC-Unclassified, SMA-Special Mention Account, SS-Sub Standard, DF- Doubtful Loan and BL-Bad Loan.

10. Interest rates (INTT_RATE): it expressed as percentage and it determine amount of bank's profit.

11. Type of collateral (TYPE _COLLA): data samples are divided in two categories: physical assets like home and property; and financial assets like equity and long term deposit.

12. Value of collateral (VALUE _COLLA): Value of Collateral security.

Table 1: Variable definitions and measurements

Variable	Definition	Measurement
Dependent Variable		
CUST_STATUS	good and bad customers	Dummy(1, if customers is Good, 0 if Bad)
Borrower characteristics		
TYPE_BORR	Type of borrower	Dummy (1, if Individual borrower, 0 if corporate borrower/Company)
RELAT_BORR	Years of relationship with the bank	Scale (year)
HIST_BORR	Borrower default history	Dummy (1, if borrower has history of default, 0 if otherwise)
GENDER_BORR	Data samples are divided to female and male according to their gender.	Dummy (1, if borrower is Male, 0 if female)
Loan characteristics		
LOAN_SIZE	Loan size	Scale (Lac)
LOAN_DURA	Loan duration	Scale (month)
SUPERV_LOAN	Loan supervision and monitoring by lender	Dummy (1, if bank supervised loan, 0 if otherwise)
TYPE_LOAN	Data samples are divided into Three groups Cash Credit, Agriculture Loan and SOD-Secured Over Draft.	Dummy(1, if loan is Cash Credit, 0 if loan is SOD, -1 if loan is Agriculture Loan)
RECOV_STATUS	Loan recovery status (UC-Unclassified, SMA-Special Mention Account, SS-Sub Standard, DF- Doubtful Loan and BL-Bad Loan.)	Dummy(1, if loan is Unclassified, 0.5 if loan is SMA, 0 if loan is Sub Standard, -0.5 if loan is Doubtful, -1 if loan is Bad Loan)
INTT_RATE	Growth in Interest Rate	Scale(%)
Collateral characteristics		
TYPE _COLLA	data samples are divided in two categories: physical assets like home and property; and financial assets like equity and long term deposit.	Dummy(1, if Collateral is physical assets, 0 if financial assets)
VALUE _COLLA	Value of collateralized property	Scale (Lac)

3.2 Model Estimation

We apply data into neural network model to estimate the probability that the customers are good or bad. Table 2 shows network information.

Table 2: Network information

Input Layer	Number of units (excluding the bias unit)	12
	Rescaling method for covariates	Standardized
Hidden Layer	Number of hidden layers	1
	Number of units in hidden layer 1 (excluding the bias unit)	6
	Activation function	Hyperbolic tangent
Output Layer	Dependent variable	CUST_STATUS
	Number of units	2
	Activation function	Softmax
	Error function	Cross-entropy

4. DATA ANALYSIS AND RESULTS

To analyze this dataset we use WEKA Open Source tool for Data mining [15] these information is summarized on figure 2 shown the Input variables and the Target Variables.

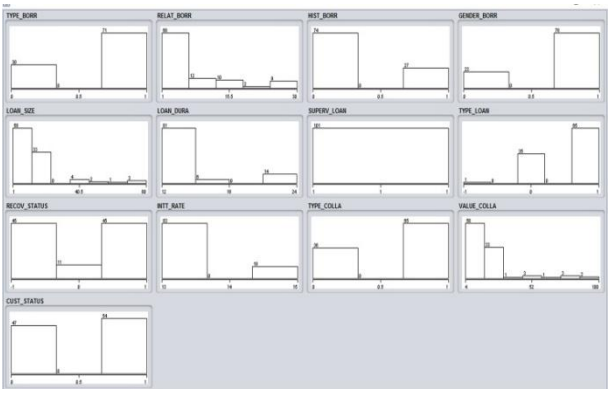

Figure 2: the Input variables and the Target Variables

Table-3 shows Minimum, Maximum, Mean and standard deviation (σ) of all variable.

Table 3: Descriptive statistics

Variable	N	Minimum	Maximum	Range	Mean	Std. Dev. (σ)
CUST_STATUS	101	0	1	1	0.535	0.501
TYPE_BORR	101	0	1	1	0.703	0.459
RELAT_BORR	101	1	30	29	7.911	7.199
HIST_BORR	101	0	1	1	0.267	0.445
GENDER_BORR	101	0	1	1	0.772	0.421
LOAN_SIZE	101	1	80	79	15.564	15.667
LOAN_DURA	101	12	24	12	14.02	4.266
SUPERV_LOAN	101	1	1	0	1	0
TYPE_LOAN	101	-1	1	2	0.634	0.504
RECOV_STATUS	101	-1	1	2	0.054	0.768
INTT_RATE	101	13	15	2	13.356	0.769
TYPE_COLLA	101	0	1	1	0.644	0.481
VALUE_COLLA	101	4	100	96	21.891	18.98

WEKA Process: Step-1. Input all data set in WEKA; Step-2. Select Classifier: Multilayer Perceptron and get result; Step-3. Supplied test data and get result – bad or good customer but Relative absolute error= 48.832%.

The data description is shown in Table 1. The dependent variable (Y) is an ordinal variable. Some of the independent variables are also ordinal and some of them are scale. SPSS (version 24) software [16] is used for modeling (neural network model). For building the provided model, only one hidden layer with hyper tangent activation function is used. The numbers of nodes in the inner layer will be selected automatically. The output of software analysis is shown in Table 4.

Table 4: Classification

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	32	1	97.0%
	1	2	36	94.7%
	Overall Percent	47.9%	52.1%	95.8%
Testing	0	13	1	92.9%
	1	1	15	93.8%
	Overall Percent	46.7%	53.3%	93.3%

Table 4 is divided to two parts: Training and Testing. In the first part of table (Training), 97.0 percent of customers who classified to good customers and 94.7 percent of customers who classified to bad customers were estimated correctly. In second part of table (Testing), the correct predicted percent are 92.9% and 93.8%. Also according to Table 5, overall percent error data is 10.52%. It is clear that in provided model, all the variables don't have the same effect on estimation and some are more effective in this model.

Table 5: Model Summary

Training	Cross Entropy Error	k10.520
	Percent Incorrect Predictions	4.2%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.03
Testing	Cross Entropy Error	8.239
	Percent Incorrect Predictions	6.7%

In Figure 3, importance of variables in the model is presented as normalized. Based on this figure, History of Borrower (Defaulter), Amount of loan, Type of collateral and Value of collateral Security have greatest effect on customer's good or bad estimation.

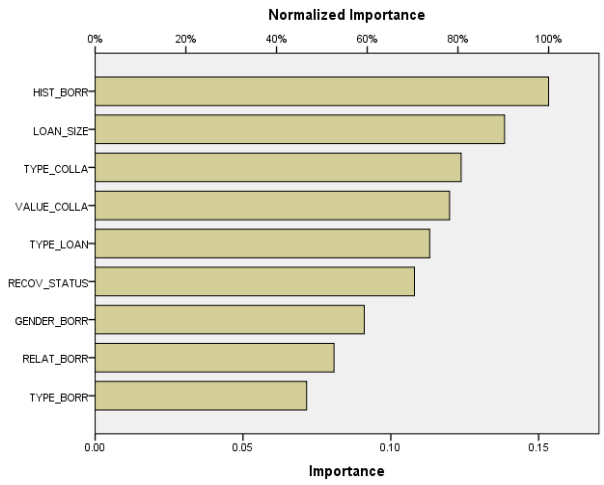

Figure 3: Importance of variables in the model

Figure 3 shows the importance analysis of the independent variables. It evaluates the degree of influence that the independent variables have on the classification power of the ANN model. Relative importance ranges from 0% representing 'no effect' to 100% representing the variable that dominates the prediction.

Customer will be good if History of Borrower (Defaulter) = No, Value of collateral Security > Amount of loan and also Type of collateral= financial asset is better than physical Asset. So those attribute is more important for selection of good or bad bank customer.

5. CONCLUSION

This paper represents an application of artificial neural network to select good or bad borrowers for giving new Loan and aim of this paper is to detect classification principles for good and bad borrowers in Bangladeshi banks. The architecture of artificial neural network models and the importance of customers' credit risk measurement are discussed here. After defining necessary variables, the collected data were entered into the model. Results of this study show that History of borrower (Defaulter or non-defaulter), amount of loan, type of collateral security (physical assets or financial assets) and Value of collateral security have most important effect in identifying classification criteria of good customers and bad customers. It means that bank managers and policy makers should focus on History of borrower (defaulter or non-defaulter), amount of loan, type and value of collateral Security. Hoping that, this strategy reduces credit risk and increases the bank's profit.

In the future, the objective is to integrate multilayer perceptron neural network model to core banking software for that a bank credit officer can easily identify good or bad borrower for giving new loan. Moreover, the extra data sets have to be classified to get more accuracy and compare with other neural network algorithm.

6. REFERENCES

[1] Angelini, E., di Tollo, G., &Roli, A. (2008). A neural network approach for credit risk evaluation. The Quarterly Review of Economics and Finance, 48(4), 733-755.

[2] Matoussi, H., &Abdelmoula, A.k (2009). Using a Neural Network-Based Methodology for Credit–Risk Evaluation of a Tunisian Bank. Middle Eastern Finance and Economics, 4, 117-140.

[3] Vasconcelos, G. C., Adeodato, P. J. L., &Monteiro, D. S. M. P. (1999, July 20-22). A NeuralNetwork Based Solution for the Credit Risk Assessment Problem. Paper presented at the IV Brazilian Conference on Neural Networks, São José dos Campos.

[4] Abdou, H. A., &Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. Intelligent Systems in Accounting, Finance and Management, 18(2-3), 59-88.

[5] Salehi, M., &Mansoury, A. (2011). An evaluation of Iranian banking system credit risk: Neural network and logistic regression approach. International Journal of the Physical Sciences, 6(25), 6082-6090.

[6] Eletter, S. F., &Yaseen, S. G. (2010). Applying Neural Networks for Loan Decisions in the Jordanian Commercial Banking System. International Journal of Computer Science and Network Security, 10(1), 209-214.

[7] Haykin SS. Neural networks: a comprehensive foundation. London: Prentice-Hall; 1999: 842.

[8] Ripley BD. Pattern recognition and neural networks. Cambridge: Cambridge university press; 1996: 403.

[9] Bishop CM. Neural networks for pattern recognition. Oxford: Clarendon press; 1995: 482.

[10] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986; 323:533-536.

[11] Gil D, Johnsson M, Garcia Chamizo JM, Paya AS, Fernandez DR. Application of artificial neural networks in the diagnosis of urological dysfunctions. Expert SystAppl 2009; 36:5754- 5760.

[12] Gil D, Girela JL, De Juan J, Gomez-Torres MJ, Johnsson M. Predicting seminal quality with artificial intelligence methods. Expert SystAppl 2012; 39:12564-12573.

[13] Pal M, University of Nottingham - GB. Factors Influencing the Accuracy of Remote Sensing Classification: A Comparative Study. University of Nottingham; 2002:

[14] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 2009; 11:10-18.

[15] WEKA Open Sources tools for Data Mining; http://www.cs.waikato.ac.nz/ml/weka/

[16] IBM - Statistical analysis software package - SPSS Statistics; http://www.ibm.com/software/products/en/spss-statistics