Journal of Mathematics and Statistics

Original Research Paper

Comparison of Multiple Linear Regression and Neural Network Models in Bank Performance Prediction in Botswana

Hassan Kablay and Victor Gumbo

Department of Mathematics, University of Botswana, Private Bag UB00704, Gaborone, Botswana

Article history
Received: 08-08-2021
Revised: 09-10-2021
Accepted: 12-10-2021

Corresponding Author:
Hassan Kablay
Department of Mathematics,
University of Botswana, Private
Bag UB00704, Gaborone,
Botswana
Email: hassankablay@gmail.com

Abstract: Bank performance is critical to the banking sector and the economy as a whole. In this study, Multiple Linear Regression (MLR) technique and feed forward Neural Network (NN) are used to predict the performance of 11 banks in Botswana. Return on Assets (RoA) is used as the dependent variable, while management quality, credit risk, liquidity, financial leverage and capital adequacy are used as the independent variables. The data is sourced from the financial reports for the year range 2015-2019. When using MLR, the cost-to-income (C_I) ratio (management quality measure) and the loan loss provision to total loans (LLP_TL) ratio (credit risk measure) are found to be the two most significant drivers of bank performance. The NN has an R^2 value of 84.37% which is significantly higher than the R^2 value of 70.00% for the MLR. The cost-to income ratio is found to be the most important driver of the NN. The performance of the two methods (MLR and NN) is then assessed using the Mean Absolute Error (MAE) and Mean Square Error (MSE) as the performance indicators. When using the validation sample, it was found out that the MLR has a MAE of 0.00611 while the NN has a MAE of 0.00472. The MLR has a MSE of 0.00008 in comparison to the NN with a lower MSE of 0.00004. It was then concluded that the NN has better predictive abilities than the MLR.

Keywords: Multiple Linear Regression, Neural Networks, Bank Performance

Introduction

The stability of any country depends heavily on the performance of its banks. The prediction and/or monitoring of bank performance using statistical techniques has evolved over the years with multiple linear regression being the most commonly used. Bank performance prediction is of utmost importance as it promotes better managerial decisions both in banks and other financial sectors. Multiple linear regression is easy to interpret due to the regression equation derived as compared to neural networks which carry out all their calculations in a “black box” and provides the predicted result (the user gets no information on the estimation of parameters). However, neural networks have been found in most literature to outperform the multiple linear regression in terms of predictive ability.

In this study, ROA (dependent variable) was used as the bank performance measure while 5 other financial ratios that have been used in past literature were selected as the independent variables. The study utilizes MLR and feed-forward NN for analysis. The performance of the two methods is compared using the MAE and the MSE.

Objectives

- To predict the performance of Botswana banks using 5 financial ratios under MLR and ANN.
- To compare the performance of the two methods (MLR and ANN) using MAE and MSE.

Literature Review

The financial performance of banks has been studied globally using different methods such as Artificial Neural Networks (ANN), data envelopment analysis and discriminant analysis, as well as the commonly used statistical method, multiple linear regression.

Bakar and Tahir (2009) carried out a study on predicting bank performance using MLR and ANN. In their study, both techniques were assessed to determine which one had better predictive abilities. They found out that C_I and LLP_TL ratios were the most influential determinants of bank performance when using MLR with an R^2 value of about 60.9%. The ANN testing results established an R^2 value of 66.9% which is greater than that of the MLR. The ANN was proved to be the more powerful tool in predicting bank performance with a Mean Square Prediction (MSPR) value of 0.0061 against the MLR MSPR value of 0.6190.

Sarokolaei *et al*. (2012) studied the performance of 10 Iranian banks using MLR and ANN with ROA as the dependent variable. One of their major findings were a positive relationship between cost to income ratio and inflation rate when using MLR. When using ANNs, 7 different inputs were used and a neural network with 9 neurons was obtained. Sarokolaei *et al*. (2012) used the MSPR to measure the performances of the two methods and they found out that the regression method performed better than the neural networks.

Kamande (2016) conducted a study on 11 Kenyan commercial banks for the years 2011 to 2015. ROA was used as the dependent variable, while capital adequacy, assets quality, management efficiency, earnings ability and liquidity were used as independent variables. Some of the findings of the research were that asset quality of the bank has the highest influence on ROA. In another study in Kenya, Ongore and Kusa (2013) investigated the moderating effect of ownership structure on bank performance using MLR and generalised least squares on panel data to estimate the parameters. In their study, it was concluded that board and management decisions are the main drivers of financial performance in Kenyan banks, while macroeconomic factors have insignificant contribution.

Shah and Jan (2014) used regression analysis and correlation technique to study the performance of 10 commercial banks in Palestine. Some of their findings were that bank size and operational efficiency were negatively related to ROA. In a study by Karim and Alam (2013), the performance of 5 banks in Bangladesh was evaluated using financial ratios for the period 2008 to 2012. Their study employed MLR analysis and they found out that the strongest model in measuring bank performance as seen by the Adjusted R-Square was ROA, followed by the Economic Value Added and finally the Tobin's Q model.

Fakhri *et al*. (2019) investigated the factors that affect the performance of Sharia and Conventional banking using ANN. They found out that inflation was the most influential variable that affected the Sharia banking performance, as well as for Conventional banking although it was not too significant. In another study by Sapuan *et al*. (2017), the performance of Islamic banks in Malaysia was assessed using MLP neural networks and pooled regression. They found out that total assets (representing size) were the most influential drivers.

In most recent research in Botswana, Kablay and Gumbo (2021a) investigated the determinants of financial distress of Botswana banks using logistic regression analysis. Their study was conducted over a range of 5 years on 11 banks. The findings revealed that Return on Equity (ROE) and Non-Performing Loans (NPL) ratios are the most influential drivers of financial distress for Botswana banks. In another research by Kablay and Gumbo (2021b), C_I, ROA and ROE were used as dependent variables to investigate financial performance of 11 banks over a 5 year period using multiple linear regression. Interest income on loans over average total assets was found to be the most influential driver of ROA and ROE, while interest expense over assets proved to be the most influential driver of C_I ratio.

Methodology

The study was carried out on 11 Botswana banks over a 5 year period. The banks' financial reports were used as the data sources. The dependent variable used was ROA and the predictor variables used were 5 financial ratios mostly utilized in past banking literature, as shown in Table 1.

Multiple Linear Regression

This technique models the relationship between the response variable and multiple explanatory variables. This model is the generalisation of the simple linear regression model and is widely used in the banking industry to predict bank performance, among others.

The general linear regression model is:

$$Y_i = \beta_0 + \sum_i \beta_i * X_i + \mu_i \tag{1}$$

where,

- i ranges from 1 to 5
- β_0 is the constant term (intercept)
- β_i are the slope coefficients for each explanatory variables
- X_i are the explanatory variables
- μ_i is a random error

The 55 observations in this study were used for developing the multiple linear regression model and two-thirds (2/3) of the total sample was used to validate the model as this study was carried out on a small dataset.

Artificial Neural Network

Neural Networks is a technique that has Artificial Intelligence at its very core in decision process (Anderson, 2007). Hastie *et al*. (2016) states that neural networks are an effective learning method that is widely used in various fields of study. Much like the human brain, neural networks get knowledge from their environment via a learning process and store it in interneuron connection weights (Haykin, 1998).

Multi-Layer Perceptron (MLP)

MLP is a neural network with an input and output layer and one or more intermediate layers. Figure 1 shows an

MLP network which is of feed-forward type and mostly uses a back propagation algorithm.

Each input neuron receives input signals x_i that have connection weights w_{ki}. A weighted summation ν of the inputs is processed and then a suitable activation function f(ν) transforms the weighted summation into the output y_k.

$$y_k = f(v) = f\left(\sum_{i=1}^{n} x_i w_{ki}\right) \tag{2}$$

where,

- x_i is the network's input
- y_k is the network's output
- w_{ki} is the synaptic weight between output of neuron k and input of neuron i
- $v = x_i w_{ki}$
- $f(v)$ is the activation function

For this study, the sigmoid function is used as the activation function and is shown in Eq. 3:

$$f(v) = \frac{1}{1+e^{-v}} \tag{3}$$

where,

- $v = x_i w_{ki}$

Comparison of Performance

To compare the predictive abilities of the two models under study, namely MLR and ANN, the MAE and the MSE are used as loss functions on the validation sample. These functions measure the predictive abilities of the model.

Mean Absolute Error (MAE)

The first type of loss function used to compare the bank performance prediction ability of the two models is the MAE and it is defined below:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{4}$$

where,

- y_i is the target value,
- $\hat{y}_l$ is the predicted value

Mean Square Error (MSE)

This is the most commonly used regression function and it will be used as the second loss function, which is defined below:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{5}$$

where,

- y_i is the target value
- $\hat{y}_l$ is the predicted value

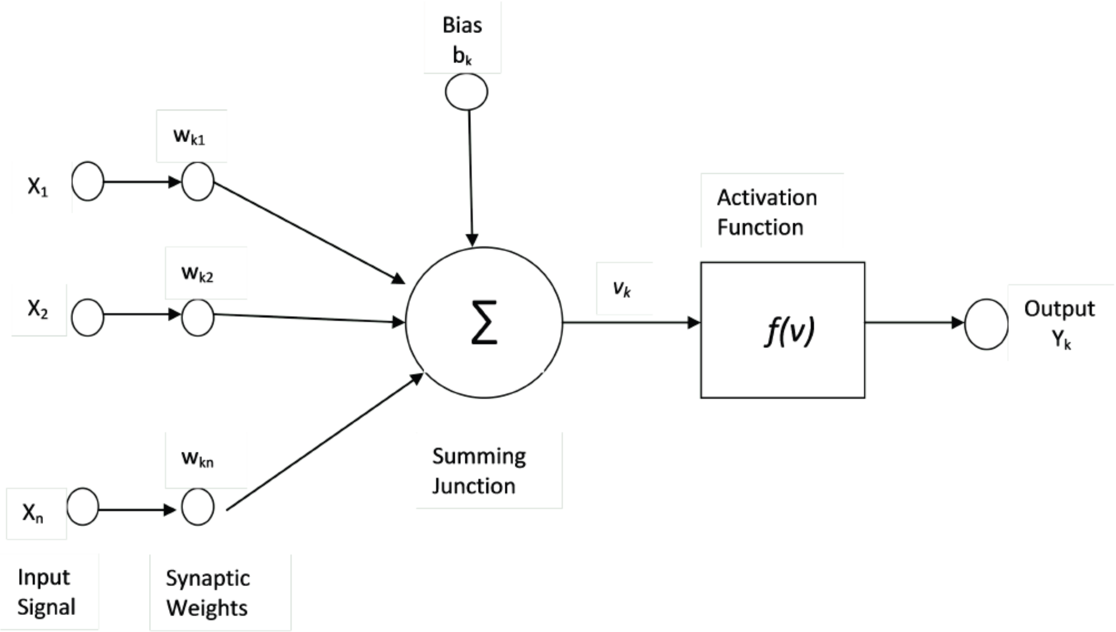

Fig. 1: Multilayer Perceptron network

Table 1: Variables used to predict bank performance

Variable	Ratio	Notation
Dependent variable		
Performance	Net Profit after taxes/Assets	ROA
Independent variables		
Credit Risk	Loan Loss Provisions/Total Loans	LLP_TL
Liquidity risk	Loans/Deposits	LDR
Management quality	Operating Expenses/Operating Income	C_I
Financial leverage	Assets/Equity	A_E
Capital adequacy ratio	(Tier 1 Capital + Tier 2 Capital)/Risk Weighted Assets	CAR

Table 2: Descriptive statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error
ROA	55	-0.056	0.031	0.010	0.015	0.000	-2.132	0.322
A E	55	3.344	19.493	9.678	3.686	13.587	0.055	0.322
C I	55	0.318	2.179	0.697	0.284	0.081	3.192	0.322
CAR	55	0.152	1.063	0.246	0.159	0.025	3.345	0.322
LDR	55	0.000	2.933	0.956	0.528	0.279	2.542	0.322
LLP_TL	55	-0.003	0.057	0.010	0.010	0.000	2.093	0.322

Table 3: Multicollinearity test results

	ROA	A_E	C_I	CAR	LDR	LLP_TL
ROA	1					
A E	-0.0639	1				
C I	-0.7652	-0.0431	1			
CAR	-0.2789	-0.6437	0.5150	1		
LDR	0.2102	-0.4167	-0.1715	0.4298	1	
LLP_TL	-0.0069	0.0835	-0.3969	-0.2095	-0.1374	1

Data Analysis

The five independent variables, A E, C_I, CAR, LDR and LLP_TL were selected to study their influence on the ROA of banks in Botswana. The study used data for 11 banks and ranged from 2015 to 2019, which is a total sample of 55 observations.

Descriptive Statistics and Multicollinearity Test

The analysis begins with the preliminary analysis of the 6 variables providing the means, standard deviations and skewness for the continous variables as shown in Table 2.

The multicollinearity test shown in Table 3 was carried out to evaluate the relationships between the constructed variables. It was found out that no multicollinearity existed between the variables as this can be seen by all correlation coefficients of less than 80%. The highest positive significant correlation was identified among the CAR and C_I ratio (r = 0.515) while the most negative significant relationship was found to be between C I and ROA (r = -0.765).

Results and Discussion

Multiple Linear Regression Model

The MLR model was utilized on all 5 predictor variables and stepwise regression was used.

The general ROA model is:

$$\begin{aligned} ROA &= \beta_0 + \beta_1 * A\ E + \beta_2 * C\ I + \beta_3 \\ &*CAR + \beta_4 * LDR + \beta_5 * LLP\ TL + \mu_i \end{aligned} \tag{6}$$

where,

- β_0 is the constant term (intercept)
- βi *are* the slope coefficients for each explanatory variables
- μ_i is a random error

In Table 4, the R^2 value of 70.00% and adjusted R^2 value of 68.80%, indicate the good explanatory power of the regression model. The R^2 value shows that 70.00% of the variation in the dependent variable (ROE) is explained by the explanatory variables. The validation sample for the MLR was found to have an R^2 value of 69.23%, a MAE of 0.00611 and a MSE of 0.00008.

As shown in Table 5, the significance value which is less than 0.001 is less than the selected level of significance of 5%, hence the model is significant. This implies that, for the 11 banks under study, the C_I and LLP_TL ratios have a significant impact on ROA.

The ROA model is:

$$ROA = 0.048 - 0.047 * CI - 0.522 * LLPTL \tag{7}$$

The regression analysis established that C_I and LLP_TL are significant in predicting the performance of banks, as shown in Table 6. This is in agreement with a study by Bakar *et al*. (2009), who found out that the same variables were significant in determining the performance of banks in Malaysia over 6 years. In Table 6, the C_I ratio has a coefficient of -0.047 that is significant and negatively correlated to ROA, therefore for every 1-unit increase in C_I there is a 0.047 decrease in ROA when all other variables remain constant. Moreover, the LLP_TL ratio has a coefficient of -0.522 that is significant and negatively correlated to ROA, therefore for every 1-unit increase in LLP_TL, there is a 0.522 decrease in ROA when all other variables remain constant.

Artificial Neural Network

The neural network was developed using the data collected for the 11 banks (100% sample). The sample consisted of only 55 observations and since this is a small data set, two-thirds (2/3) of the total sample was used for validating the model.

A 3-Layer Multilayer Perceptron (MLP) neural network with 1 input layer, 1 output layer and 1 hidden layer was developed with the selected variables namely, A_E, C_I, CAR, LDR and LLP_TL. The input layer comprised of 5 neurons while the hidden layer comprised of 3 neurons and the output layer comprised of 1 neuron as seen in Table 7. The sigmoid function and the identity function are used as the activation functions in the hidden layer and output layer, respectively.

Supervised learning was used by providing the network with the actual output for each input and as a result the re-adjustment of the weights was performed in the hidden neuron in an effort to minimise the error between the predicted output and the actual output.

The feed forward NN obtained is shown in Fig. 2. The development sample for the network has an R^2 value of 84.37% with a MAE of 0.00392 and a MSE of 0.00003. On the other hand, the validation sample for the network has an R^2 value of 83.00% with a MAE of 0.00472 and a MSE of 0.00004.

The importance of the 5 independent variables in the ANN model is shown in Table 8. As indicated on the table, C_I has the greatest effect on bank performance as it has an importance of 0.392 followed by CAR with 0.244, LLP_TL with 0.179, LDR with 0.120 and lastly, A_E with the least importance of 0.065.

Figure 3 below shows both the importance and the normalized importance graphically

Comparison of the MLR and ANN Model

The MAE and MSE were used to compare the performance of the MLR and ANN model in predicting the performance of the banks. When using the development sample, the MLR was found to have an R^2 value of 70.00% in comparison to the ANN with an R^2 value of 84.37%.

The performance measures were then evaluated on the validation sample and the results in Table 9 were obtained. The NN model was found to have better predictive abilities as observed by both a lower MAE and MSE when compared to the MLR model.

Table 4: Model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	0.765	0.585	0.578	0.009	
2	0.837	0.700	0.688	0.008	1.460

a. Predictors: (Constant), C I
b. Predictors: (Constant), C I, LLP TL
c. Dependent Variable: ROA

Table 5: ANOVA

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	0.007	1	0.007	74.859	<0.001
Residual	0.005	53	0.000		
Total	0.011	54			
2 Regression	0.008	2	0.004	60.668	<0.001
Residual	0.003	52	0.000		
Total	0.011	54			

a. Dependent Variable: ROA
b. Predictors: (Constant), C I
c. Predictors: (Constant), C I, LLP TL

Table 6: Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1 (Constant)	0.038	0.003		11.044	0.000	0.031	0.045		
C_I	-0.039	0.005	-0.765	-8.652	0.000	-0.048	-0.030	1.000	1.000
2 (Constant)	0.048	0.004		12.780	0.000	0.041	0.056		
C_I	-0.047	0.004	-0.911	-11.015	0.000	-0.055	-0.038	0.842	1.187
LLP_TL	-0.522	0.117	-0.369	-4.455	0.000	-0.756	-0.287	0.842	1.187

a. Dependent Variable: ROA

Table 7: Network Information

		1	A_E	
		2	C_I	
Input Layer	Covariates	3	CAR	
		4	LDR	
		5	LLP_TL	
	Number of Units			5
	Rescaling Method for Covariates		Standardized	
	Number of Hidden Layers			1
Hidden Layer (s)	Number of Units in Hidden Layer 1			3
	Activation Function		Sigmoid	
	Dependent Variables	1	ROA	
	Number of Units			1
Output Layer	Rescaling Method for Scale Dependents		Standardized	
	Activation Function		Identity	
	Error Function		Sum of Squares	

a. Excluding the bias unit

Table 8: Independent variable importance

	Importance	Normalized Importance
A_E	0.065	16.6%
C_I	0.392	100.0%
CAR	0.244	62.3%
LDR	0.120	30.8%
LLP_TL	0.179	45.6%

Table 9: Comparison of performance

	MAE	MSE
MLR	0.00611	0.00008
NN	0.00472	0.00004

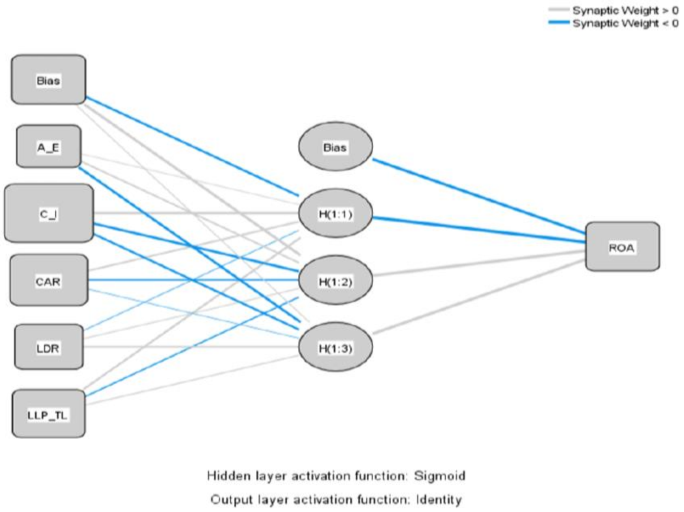

Fig. 2: Neural network

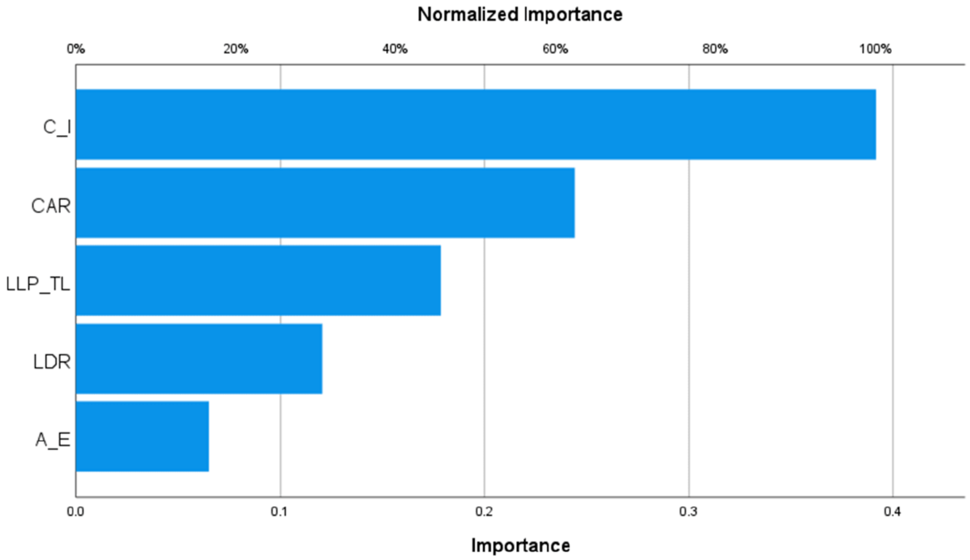

Fig. 3: Normalized importance

Conclusion

In comparison to the MLR model, the neural network was found to have better predictive abilities as seen by both a lower MAE and MSE. Although the neural network was found to be the better method when predicting bank performance in Botswana, the MLR was found to be simpler to interpret/use. This is because a regression equation is obtained when using multiple linear regression while the neural network consists of a "black box" where all calculations are performed and the user gets no information on the estimation of the parameters. The MLR established 2 significant drivers of bank performance which are C_I and LLP_TL, whereas when using the neural network the C_I, CAR and LLP_TL carried more weight in the prediction of bank performance in Botswana. The ANN does not require any distributional assumptions, hence there is no violation of assumptions as compared to when using the MLR model on real data. The neural network performed better than the multiple linear regression in predicting bank performance and this is consistent with Bakar and Tahir (2009).

Further studies can be carried out using more data mainly because the neural networks are good at training on a large dataset. This will allow the sample to be split into 3 parts, which are training, testing and validation. Moreover, the COVID-19 Pandemic has considerably affected the banking industry, therefore more recent data (2020) may be used in future research in order to accommodate such changes.

Acknowledgement

Special thanks to my supervisor Dr Victor Gumbo for the continuous support of my research, for his motivation and patience.

Ethics

The author confirms that this article is original and contains unpublished material and the author has read and approved the manuscript and no ethical issues involved.

Author's Contributions

Hassan Kablay: Collected the data, designed the study, performed the statistical analysis and wrote the draft of the manuscript, gave final approval of the version to be submitted.

Victor Gumbo: Participated in data collection, managed the analysis of the study, reviewed the article for significant intellectual content, gave final approval for the version to be submitted.

References

Al Karim, R., & Alam, T. (2013). An evaluation of financial performance of private commercial banks in Bangladesh: Ratio analysis. Journal of Business Studies Quarterly, 5(2), 65. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.651.5655&rep=rep1&type=pdf

Anderson, R. (2007). The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation. Oxford University Press.

Bakar, N. M. A., & Tahir, I. M. (2009). Applying multiple linear regression and neural network to predict bank performance. International Business Research, 2(4), 176-183. doi.org/10.5539/ibr.v2n4p176

Fakhri, U. N., Anwar, S., Ismal, R., & Ascarya, A. (2019). Comparison and predicting financial performance of islamic and conventional banks in indonesia to achieve growth sustainability. al-Uqud: Journal of Islamic Economics, 3(2), 174-187. doi.org/10.26740/al-uqud.v3n2.p174-187

Hastie, T., Tibshirani, R., & Friedman, J. (2016). The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, 2nd Edition.

Haykin, S. (1998.) Neural Networks: A Comprehensive Foundation. Pearson, 2nd Edition.

Kablay, H., & Gumbo, V. (2021a). Bank Distress Prediction Model for Botswana. Asian Research Journal of Mathematics, 47-59. doi.org/10.9734/arjom/2021/v17i230273

Kablay, H., & Gumbo, V. (2021b). Financial Performance of Banks in Botswana. Journal of Mathematical Finance, 11(3), 386-397. doi.org/10.4236/jmf.2021.113022

Kamande, E. G. (2017). The effect of bank specific factors on financial performance of commercial banks in Kenya (Doctoral dissertation). http://41.89.55.71:8080/xmlui/handle/123456789/3057

Ongore, V. O., & Kusa, G. B. (2013). Determinants of financial performance of commercial banks in Kenya. International journal of economics and financial issues, 3(1), 237. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.827.1383&rep=rep1&type=pdf

Sapuan, N. M., Bakar, S., & Ramlan, H. (2017). Predicting the performance and survival of islamic banks in malaysia to achieve growth sustainability. In SHS Web of Conferences (Vol. 36, p. 00016). EDP Sciences. doi.org/10.1051/shsconf/20173600016

Sarokolaei, M. A., Alinezhad, P., & Khosroshahi, M. A. (2012). A Comparative Study of Iranian Banks' Efficiency by Using Artificial Neural Networks and Multi-Linear Regression. In 2 nd International Conference on Management and Artificial Intelligence IPEDR. http://ipedr.com/vol35/016-ICMAI2012-E10027.pdf

Shah, S. Q., & Jan, R. (2014). Analysis of financial performance of private banks in Pakistan. Procedia-Social and Behavioral Sciences, 109, 1021-1025. doi.org/10.1016/j.sbspro.2013.12.583