Journal of Statistical and Econometric Methods, vol.5, no.1, 2016, 29-47
ISSN: 1792-6602 (print), 1792-6939 (online)
Scienpress Ltd, 2016

A Study of Discriminant Analysis and Artificial Neural Network in Prediction of Stock Market in Nigeria

R.A. Kareem[1] and O.A. Adeoti[2]

Abstract

This paper analyses financial and macroeconomic data responsible for predicting stock market in Nigeria using four company-specific variables and five macroeconomic variables. The variables are inflation, investment, consumer price index, unemployment, lending interest rate, net revenue, net income and net asset. Discriminant analysis and artificial neural network were employed to determine the variables responsible for good and bad investment choices. The result of study has shown that earnings per share, lending interest rate, inflation and net income are important variables that contributed towards good and poor investment choices and the ANN model trained with scaled conjugate gradient algorithm using five hidden nodes perform better in discriminating between good and poor investment

[1] Department of Mathematics and Statistics, Lagos State Polytechnic, Ikorodu, Nigeria.
Email: raskareem4@yahoo.com
[2] Department of Mathematics and Statistics, Bowen University, Iwo, Nigeria.
E-mail: tim_deot@yahoo.com

Article Info: *Received* : October 17, 2015. *Revised* : November 23, 2015.
Published online : March 3, 2016.

choices and has higher percentages of classifying groups.

Mathematics Subject Classification: 62H30; 68T10; 62P20

Keywords: Stock market; Investment choice; Artificial neural network; Discriminant analysis

1 Introduction

The stock market is a financial game of winners and losers and most often unpredictable since a seemingly uncountable number of variables influence the stock market and company's performance. Global crashes in stock market operation do not occur all of a sudden but are as a result of local and regional crashes in emerging economies. Also the interdependence among macroeconomic indicators and financial ratios affect the probabilities of the different types of stock market crashes. It is important for potential investors and shareholders to be aware of relevant financial and macroeconomic indicators that may determine good and bad investment choices. Predicting the stock performance as to what constitutes good or bad investment choices decision is certainly difficult and complicated as no accurate and comprehensive model has been developed for predicting stock market in Nigeria.

The performance of stock has been analyzed, to some extent, on the financial indicators presented in the company's annual report. The report contains vast amount of information that can be transformed into financial ratios for assessing the future performance of the company. Most analyst, investors and researchers use financial ratios in fundamental analysis to predict the future performance of a company stock. So financial ratios help to form the basis of judging whether a company stock value will rise or fall in a given year. The level of importance given to financial ratio differs from country to country and sector to sector.

However, it is observed that in some situations that the performance of stock is not only affected by financial ratio, sometimes it is affected by some macroeconomic variables which are at variance with the financial information of the different companies listed on the Nigerian stock exchange So, there exist numerous variables both financial and macroeconomic influencing the performance of quoted companies on the stock market. Thus, selecting appropriate financial ratio and macroeconomic indicators is very crucial in the effort to classify stock as good or bad investment choices.

In the literature, a number of different methods have been applied in order to predict Stock market and assess their performances. Min and Jeong [8] proposed a binary classification method for bankruptcy prediction and showed that the prediction accuracy can serve as a good alternative when compared with other methods such as logistic regression, decision tree and multi-discriminant analysis. Li et al [7] use the logistic regression as a comparative method to build a model for the prediction of stock returns. Dutta et al. [1] predicted the out performing stock in Indian market using logistic regression and eight financial ratios as the independent variables to determine the financial indicators that significantly affect the performance of the stock and concluded that the model can enhance the stock price forecasting ability of an investor. However, macroeconomic variables which also can influence the stock price were not considered in their study. The use of data-mining techniques in predicting stock return was studied by Ogut and Aktas [9]. They proposed a binary classification method for predicting corporate failure based on genetic algorithm and show that data-mining techniques are better than multivariate techniques. Pan et al. [10] presented a computational approach for predicting the Australian Stock market index using the neural network from time series data exploiting dynamical swings and inter-market influences. Kara et al.[4] studied the prediction direction of stock price index using ANN and SVM for the Istanbul stock exchange. They attempted to develop models and compared their performances in predicting the direction of movement in the Istanbul stock

Exchange. Li and Sun [6] use homogenous multiple classifiers in predicting return on investment and show that classifiers using majority voting perform best when predicting stock returns. Karlen and Poulsen [5] investigates the performance of stocks in America shunned (sin) industries in an attempt to increase the awareness of how the stock of companies perceived as engaging in sinful business activities has performed.

This paper explores the use of discriminant analysis and artificial neural network to develop a model for classifying the company stocks into two categories of good and poor investment choices for nine financial and macroeconomic variables. Discriminant analysis offers a tool for separating a large dataset into two or more mutually exclusive groups using the variables of the dataset and ANN is a tool that can be applied to complex but unknown relationship between variables. They can learn from existing data set and do not require a apriori model. These if properly implemented can accurately classify financial and macroeconomic data set, that they have not been trained on in the Nigeria stock market.

2 Methodology

2.1 Data

The dataset used in this study consist of data of 25 companies' annual reports quoted on the Nigerian Stock Exchange (NSE) randomly selected using the simple random sampling technique from different sectors as well as some macroeconomic variables. The financial and macroeconomic data were sourced from the Nigeria stock exchange and the world bank data respectively. It consists of four company –specific and five macroeconomic variables that we feel might account for a particular company at a specific time being a good or bad investment choice for the period 2010- 2013. The nine variables are: Profit before taxation

(Net Revenue), Profit after taxation (Net Income), Net asset, earnings per share, inflation rate, unemployment rate, lending interest rate, annual investment and consumer price index. The variables are tested individually for univariate normality using the Shapiro-Wilk test and those that are not normal are transformed to achieve normality. Four variables are transformed into normality using the log function. The data was randomly split into two sets. The first, containing 80% of the data, was the training and test set, from which samples were randomly drawn to form training and test data set. The second comprised the validation set, which was used to perform an independent evaluation of the prediction accuracy for each target stocks.

2.2 Discriminant Analysis

The Discriminant analysis readily provides a rule for classifying observations from a multivariate data set into two or more populations.

In the case of two populations defined by $\Pi_1 = N_p(\mu_1, \Sigma_1)$ and $\Pi_2 = N_p(\mu_2, \Sigma_2)$, we can derive a classification rule that can be used to classify an element **x** into one of the populations. Each **x** is assumed to be p-variate normal. When the population parameters are unknown, as is often the case, one must obtain training samples for estimation of the mean and covariance of each population, as well as derive the classification rule.

The estimates of μ_1, μ_2, Σ_1, Σ_2 are $\bar{x}_1$, $\bar{x}_2$, S_1 and S_2, respectively. If the covariance matrices for the populations, Σ_1 and Σ_2 are equal, then the common covariance matrix, Σ is replaced by the pooled estimate

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

The estimates $\bar{x}_1$, $\bar{x}_2$, S_1, S_2 and S_p are unbiased estimators.

Next, look for the classification of **x** into $\Pi_1$and Π_2.We use an optimal linear discriminant function, $\hat{a}'x$ where $\hat{a} = S_p^{-1}(x_1 - x_2)$, to assign **x** into a population based on the decision rule:

Classify **x** into Π_1 if $\hat{a}'x > \frac{1}{2}(x_1 - x_2)'S_p^{-1}(x_1 - x_2)$ otherwise classify **x** into Π_2 if $\hat{a}'x \leq \frac{1}{2}(x_1 - x_2)'S_p^{-1}(x_1 - x_2)$.

2.3 Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is a computational modelling tool that have found extensive acceptance in many disciplines for modelling complex real-world problems. Schalkoff [12] defined ANN as structures comprised of densely interconnected adaptive simple processing elements (called artificial neurons) that are capable of performing massively parallel computations for data processing and knowledge representation. Haykin [3] defined an artificial neural network as "massively parallel distributed processor that has a natural propensity for storing experiential knowledge, making it available for use and resembles the brain in two respects: *It acquire knowledge through a learning process and interneuron connection strength known as synaptic weights which are used to store the knowledge"*. ANN learns to perform a function (an input/output map) from data and has the ability to learn, recall and generalize knowledge. It is built with a systematic step-by-step procedure to optimize a performance criterion or to follow some implicit internal constraint, which is commonly referred to as the learning rule. During the learning phase, known data set are used as a training signal in input and output layers.

2.3.1 ANN Model, Training and Evaluation

A three layer multilayer perceptron (MLP) neural network model which were found to perform better than traditional statistical classification methods (Guh and

Hsieh, [2]; Rumelhart et al., [11]) was used in this study. The learning algorithms used are the gradient descent and scaled conjugate gradient algorithms. The input vectors and target vectors of the dataset are normalized such that all the features are of zero mean and unit variance. The supervised learning rule was used and the activation function of input variable is hyperbolic tangent function $f(z)=\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$ and that of output is sigmoid function $f(z)=\frac{1}{1+e^{-z}}$. The number of neurons in the hidden layer of the MLP model is 5 and 10 respectively. The number of epoch is 500 and the learning rate is 0.8. The training and test dataset was randomly divided into 60% training and 20% testing respectively. The neural network will minimize the error during training process. The prediction accuracy of the MLP over the training dataset was then evaluated to determine how well the network had learned and over the testing dataset how well the network generalized. The accuracy was measures in terms of sum of square errors of the training dataset and percentages of correct classification of the model using the confusion matrix. The contribution of each variable to the output of network was also determined. After the completion of the trial, the MLP model with the highest percentages of correct classification (that best generalize the network) and least sum of squares error in the training sample was selected as the best model which is then compared with that of discriminant analysis model.

3 Results and Discussion

3.1 Discriminant Analysis

Before performing discriminant analysis, we first need a method of classifying a company as a good or poor investment, for a given year. While there is no definite method for defining a stock market investment as “good” or “poor”, we decided that if the value of a company’s stock which was obtained as the market price on the first trading day of a given year rose compared to the closing price on the last day of the preceding year, it is classified as a good investment, otherwise it is classified as a poor investment. A value of zero is poor investment and a value of one is good investment choices. Analysis was based on data set of 25 companies for all years from 2010-2013. This made a sample size of 98 distinct company-year observations as information for year 2010 was unavailable for two companies.

Table 1: Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Inflation	.919	8.325	1	95	.005
Unemployment	.885	12.382	1	95	.001
CPI	.877	13.286	1	95	.000
Investment	.922	8.058	1	95	.006
Lending interest	.983	1.649	1	95	.202
Log_NetRev	.989	1.058	1	95	.306
Log_NetIncom	.985	1.447	1	95	.232
Log_NetAsset	.994	.542	1	95	.463
Log_Earnings	.954	4.562	1	95	.035

In Table 1, significant group differences were observed for five predictors on the dependent variables. The log determinant was quite similar in Table 2 but Box M indicated that the assumption of equality of covariance matrix was violated in Table 3. In this case Box M is 60.564 with F=1.993 which is significant at $p < 0.001$.

Table 2: Log Determinants of Group

grpdata	Rank	Log Determinant
poor	7	-13.862
good	7	-14.012
Pooled within-groups	7	-13.302

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Table 3: Box M Test Results

Box's M		60.564
F	Approx.	1.993
	df1	28
	df2	31191.482
	Sig.	.001

Tests null hypothesis of equal population covariance matrices.

It was observed in Table 4 that three significant prediction variables namely consumer price index, unemployment and inflation have strong effect on allocation to groups. However, this is not a good model for any meaningful prediction because the discriminant function model account for only 29.6% of the variation in the grouping variable while 70.4% cannot be explained in Table 5 and 6. Also, we observed that investment and lending interest rate were unused in the analysis because they failed the tolerance test.

Table 4: Structure Matrix

	Function
	1
CPI	-.577
Unemployment	.557
Inflation	.456
Investment[a]	.449
Log_Earnings	-.338
Lending interest[a]	-.203
Log_Netincome	.190
Log_NetRev	.163
Log_NetAsset	.116

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions Variables ordered by absolute size of correlation within function.

a. This variable not used in the analysis.

Table 5: Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.421[a]	100.0	100.0	.544

a. First 1 canonical discriminant functions were used in the analysis.

Table 6: Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.704	32.130	7	.000

The discriminant function (equation) is given as

D= 2.937*inflation + 171.176*unemployment + 0.885 *CPI - 0.715*NetRev + 1.309*Netincome – 0.130*asset – 0.879*earnings- 1440.810

Table 7: Canonical Discriminant Function Coefficients

	Function
	1
Inflation	2.937
Unemployment	172.176
CPI	.885
Log_NetRev	-.715
Log_Netincome	1.309
Log_NetAsset	-.130
Log_Earnings	-.879
(Constant)	-1440.810

Unstandardized coefficients

3.1.1. Classification in Discriminant Analysis

In Table 8, the discriminant analysis model was able to classify 40 observations as “poor choices” out of 47 observations. Thus 85.1% classification accuracy of poor choices for poor group. On the other hand, the model is able to classify 31 (62.0%) observations as good choices out of 50 observations. Thus the model is able to generate 73.2% classification accuracy for both groups. However, because the model can only explained 29.6% of the total variation of the data set, the discriminant analysis is not a good model for discriminating between good and poor investment choices in this study, though, three variables which are CPI, unemployment and inflation in order of importance are seen to be responsible for the classification.

Table 8: Classification Results[a,c] of good and poor choices using Discriminant Analysis

		grpdata	Predicted Group Membership		Total
			poor	good	
Original	Count	Poor	40	7	47
		Good	19	31	50
	%	Poor	85.1	14.9	100.0
		Good	38.0	62.0	100.0
Cross-validated[b]	Count	Poor	38	9	47
		Good	21	29	50
	%	Poor	80.9	19.1	100.0
		Good	42.0	58.0	100.0

a. 73.2% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 69.1% of cross-validated grouped cases correctly classified.

3.2 Artificial Neural Network

A predictive multi-layer perceptron (MLP) neural network model was obtained using SPSS 23 neural network package for the classification of variables that is responsible for good and poor investment choice. Also, the MLP neural network is used to predict the importance of individual variables towards the type of investment decisions of good and poor choices. The result derived from the model provides valuable source of information for future decisions on good and poor investment choices by investors.

3.2.1 Predictive model summary of Artificial Neural Network

Table 9- 12 is the predictive model summary of dataset that shows the results of MLP neural network when trained with scaled conjugate gradient and gradient descent algorithms for five and ten nodes in the hidden layer.

Table 9: Predictive Model Summary of Dataset for Scaled Conjugate Gradient (5 hidden nodes)

Training	Sum of Squares Error	9.598
	Percent Incorrect Predictions	22.6%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.04
Testing	Sum of Squares Error	3.920
	Percent Incorrect Predictions	27.8%
Holdout	Percent Incorrect Predictions	11.8%

Dependent Variable: grpdata

a. Error computations are based on the testing sample.

We observed that the percentages of incorrect prediction of the respective neural network is equal to 22.6% and 27.1% in the training samples for scaled conjugate gradient and gradient descent algorithms respectively when trained with five hidden nodes (see Table 9 and 10) while the percentages of incorrect prediction using the same algorithms when trained with ten hidden nodes is 18.0% and 23.8% in the training samples (see Table 11 and 12).

Table 10: Predictive Model Summary of Dataset for Gradient Descent (5 hidden nodes)

Training	Sum of Squares Error	11.364
	Percent Incorrect Predictions	27.1%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.03
Testing	Sum of Squares Error	4.689
	Percent Incorrect Predictions	30.4%
Holdout	Percent Incorrect Predictions	20.0%

Dependent Variable: grpdata

a. Error computations are based on the testing sample.

So the percentages of correct prediction are 77.4%, 62.9%, 82.0% and 76.2% respectively. We also observed that the predictive model summary of dataset using scaled conjugate gradient with 5 and 10 hidden nodes gives 88.2% and 76.2% correct predictions in the validation (holdout) sample to data that have not been previously used in both the training and testing dataset while the neural network model trained with gradient descent algorithm gives 69.2% and 61.5% correct predictions using the validation dataset. Similarly, the model trained with scaled conjugate gradient algorithm gives lower sum of squares error in the training and testing dataset when trained with 5 and 10 nodes in the hidden layer. So the model trained with the scaled conjugate gradient is considered to be a good model. Particularly, the model with 5 hidden nodes is a good model for predicting the performance of stock market since it has a good prediction rate for new dataset that has not used previously in the analysis. The predictive model using the validation dataset is summarized in Table 13.

Table 11: Predictive Model Summary of Dataset for Scaled conjugate (10 hidden nodes)

Training	Sum of Squares Error	6.854
	Percent Incorrect Predictions	18.0%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.05
Testing	Sum of Squares Error	2.308
	Percent Incorrect Predictions	11.5%
Holdout	Percent Incorrect Predictions	23.8%

Dependent Variable: grpdata

Error computations are based on the testing sample.

Table 12: Predictive Model Summary of Dataset for Gradient Descent (10 hidden nodes)

Training	Sum of Squares Error	10.287
	Percent Incorrect Predictions	23.8%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.04
Testing	Sum of Squares Error	4.561
	Percent Incorrect Predictions	33.3%
Holdout	Percent Incorrect Predictions	38.5%

Dependent Variable: grpdata

a. Error computations are based on the testing sample.

Table 13: Predictive Model Summary of ANN for Scaled Conjugate Gradient and Gradient Descent Algorithms

Model		No of hidden nodes	Original Poor choices	Original Good choices	Predicted Poor choices	Predicted Good choices	% success rate	% failure rate
Artificial Neural Network	Scaled conjugate	5	8	0	2	7	88.2	11.8
		10	11	1	4	5	76.2	23.8
	Gradient descent	5	8	3	0	4	78.2	20.8
		10	4	1	4	4	61.5	38.5

Table 14: Independent Variable Importance using ANN

	Importance	Normalized Importance	Rank
Inflation	.156	66.4%	3
Unemployment	.043	18.5%	7
CPI	.096	40.9%	5
Investment	.042	17.9%	8
Lending interest	.176	75.0%	2
Log_NetRev	.091	39.1%	6
Log_Netincome	.138	59.1%	4
Log_NetAsset	.024	10.2%	9
Log_Earnings	.234	100.0%	1

The normalized importance of individual variables which influence the output variable (good or poor investment choices) using ANN is given Table 14. The importance of the independent variables is a measure of how well the neural network model predicts value changes for different independent variable. The dominant variables are found to be earnings (100%), lending interest rate (75%), inflation (66.4%) and net income (59.1%). This result indicate that two macroeconomic variables and two financial variable are key factors responsible for the good or poor investment decision in the stock market using artificial neural network.

3.2.2 Classification in Artificial Neural Network

Table 15 shows the classification of company-specific and macroeconomic variables as good or poor investment when the neural network model is trained with scaled conjugate gradient algorithm with five hidden nodes. We observed that the model was able to generalize the network to data not yet seen with about 88.2% which can be regarded as a good score. A value of zero is poor investment and a value of one is good investment choices.

Table 15: Classification Result of good and poor choices using ANN

Sample	Observed	Predicted		
		poor	good	Percent Correct
Training	poor	22	4	84.6%
	good	10	26	72.2%
	Overall Percent	51.6%	48.4%	77.4%
Testing	poor	10	3	76.9%
	good	2	3	60.0%
	Overall Percent	66.7%	33.3%	72.2%
Holdout	poor	8	0	100.0%
	good	2	7	77.8%
	Overall Percent	58.8%	41.2%	88.2%

Dependent Variable: grpdata

4 Conclusion

This study was aimed at finding the best model for the prediction of quoted companies in the Nigerian stock market into good and poor investment decision. Discriminant analysis and ANN models was applied and the results were compared using the percentages of correct classification for the two methods. The study shows that the classification model based on ANN was more relevant than the one based on Discriminant analysis because the ANN was able to deal with errors during the training phase and has higher percentages in generalizing the network to data not yet seen by the model. On the other hand, the discriminant analysis model was unable to explain about 70.4% of variation in the dataset. Among the ANN models, the MLP model trained with scaled conjugate gradient algorithm for five nodes in the hidden layer was found to be more appropriate. Also, the model shows that earnings per share, lending interest rate, inflation and net income are found to contribute significantly to the performance of stock market in Nigeria.

References

[1] A. Dutta, G. Bandopadhyay and S. Senupta, Prediction of Stock Performance in Indian Market using Logistic Regression, *International Journal of Journal of Business and Information,* **7**(1), (2012), 105- 136.

[2] R.S. Guh and Y.C. Hsieh, A Neural Network Based Model for Abnormal Pattern Recognition of Control Charts, *Computers and Industrial Engineering*, **36**, (1999), 97-108.

[3] S. Haykin, *Neural Networks: A Comprehensive Foundation,* Second edition, Prentice-Hall, New Jersey, 1999.

[4] Y. Kara, M.A. Boyacioglu and Ö. K. Baykan, Predicting Direction of Stock Price Index Movement using Artificial Neural Networks and Support Vector

Machines: The Sample of the Istanbul Stock Exchange, *Expert Systems with Applications*, **38**, (2011), 5311–5319.

[5] A. Karlén, and S. Poulsen, Can It be Good to be Bad? Evidence on the Performance of US sin Stocks. *MSc dissertation*, Umea School of Business and Economics, (2013).

[6] H. Li, and J. Sun, Empirical Research of Hybridizing Principal Component Analysis with Multivariate Discriminant Analysis and Logistic Regression for Business Failure Prediction, *Expert Systems with Application*, **38**(5), (2011), 6244-6253.

[7] H. Li, J. Sun and J. Wu, Predicting Business Failure using Classification and Regression Tree: An Empirical Comparison with Popular Classical Statistical Methods and Top Classification Mining Methods, *Expert Systems with Applications*, **37**(8), (2010), 5895-5904.

[8] J. Min and C. Jeong, A Binary Classification Method for Bankruptcy Prediction, *Expert Systems with Applications*, **36**(3), (2009), 5256-5263.

[9] H. Ogut, M. Mete Doganay and R. Aktas, Detecting Stock-price Manipulation in an Emerging Market: The Case of Turkey, *Expert Systems with Applications***, 36**(9), (2009), 11944-11949.

[10] H. Pan, C. Tilakaratne and J. Yearwood J, Predicting Australian Stock Market Index Using Neural Networks Exploiting Dynamical Swings and Intermarket Influences, *Proceedings of 16th Australian conference on Artificial Intelligence*, Perth, T.D. Gedeon T.D and L.C.Che Funf, (editors), **Dec3-5**, (2003).

[11] D.E. Rumelnhart, D.E. Hinton and R.J. Williams, *Learning Internal Representations by Error Propagation in Parallel Distributed Process*, MIT Press, Cambridge, MA, 1986.

[12] R.J. Schalkoff, *Artificial Neural Networks*. McGraw-Hill, New York, 1997.