International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

# A Generic Prediction of Nigeria Stock Exchange Market, Implementation Using Naive Bayes Model

Abubakar S. Magaji, Department of Mathematical Sciences, Faculty of Science, Kaduna State University, Kaduna-Nigeria;
Victor Onomza Waziri, Department of Cyber Security, School of ICT, Federal University of Technology Minna-Nigeria;
Audu Isah, Department of Mathematics & Statistics, SSSE, Federal University of Technology Minna Nigeria;
Adeboye K.R., Department of Mathematics & Statistics, SSSE, Federal University of Technology Minna-Nigeria.

## Abstract

*This paper researches on Nigerian Stock Exchange (NSE), which serves as a platform where industries and commercial ventures converge to raise the public capital that paves way for them to expand, and in the process they create new jobs, products, services and opportunities. The uncertainty on the market dynamics is undesirable. We present the Naive Bayes algorithm as a tool for predicting the Nigerian Stock Exchange Market, a transformation of the NSE data and subsequent implementation of the algorithm on the WEKA platform, and from the results obtained we concluded that the Naive Bayes provides an avenue for predicting the Nigerian Stock Exchange.*

**Keywords:** *Nigerian Stock Market, Prediction, Data Mining, Machine Learning, Naive Bayes*

## 1. Introduction

It is a common tradition to note that a vast amount of capital is traded through the Stock Markets in each country and around the World. The performance of each country at the stock exchange Market gives an insight to the economic growth of that country; generally known as the GDP.

Nigerian Stock Exchange (NSE) is definitely an avenue for trading shares and bonds, but it also has a great influence on the Nigerian Economy. It serves as a platform were industries and commercial ventures converges to raise the public capital that paves way for them to expand, and in the process they create new jobs, products, services and opportunities.

As observed during the recent Economic melt-down, one of the specific characteristics that all Stock Markets have in common is the uncertainty, which is related with their short and long-term future state. This feature is undesirable for the investor but it is also unavoidable whenever the Stock Market is selected as the investment tool. The best that one can do is to try in reducing this uncertainty, or if possible wipe it out completely. Stock Market Prediction (or Forecasting) is one of the instruments in the process of achieving this dream.

There is no doubt that the majority of the people related to stock markets are trying to achieve profit. Profit comes by investing in stocks that have a good future (short or long term future) [1]. Thus what they are trying to accomplish one way or the other is to predict the future of the market. But what determines this future?

Predictions based on different models include some of the followings: Autoregressive Moving Average (ARMA), Random Walk (RW), Neural Network (NN), are being exploited now; all in an effort to improve the predictions, and thus, make reasonable indicial pronouncements that can guides the economic growth of a country. Other models that are useful in the forecasting of the economic indices are Naive Bayes (NB) and Support Vector Machines (SVM) which are useful in classifications or prediction of GDP.

The focus of our paper is time series forecasting, we intend to carry out prediction of the Nigerian Stock Market using Naive Bayes. Our study of the Stock Exchange Market is limited to Nigerian reference frame of Nigerian Stock Exchange Market. In view of this all our training data set shall be acquired within the Nigerian Stock Exchange reference context. Our studies border models shall also be limited to the Naive Bayes model.

## 2. Related Works

In the literature the data that are related to the stock markets are divided in three major categories [2]:

(1) *Technical data*: are all the data that are referred to stocks only. Technical data include:

- The price at the end of the day.
- The highest and the lowest price of a trading day.
- The volume of shares traded per day.

(2) *Fundamental data*: are data related to the intrinsic value of a company or category of companies as well as data related to the general economy. Fundamental data include: Inflation, Interest Rates, Trade Balance, Indexes of industries (e.g. heavy industry), Prices of related commodities

(e.g. oil, metals, and currencies), Net profit margin of a firm, and Prognoses of future profits of a firm.

(3) *Derived data*: this type of data can be produced by transforming and combining technical and/or fundamental data.

Many experts in the stock markets have employed the technical analysis for better prediction for a long time. Generally speaking, the technical analysis derives the stock movement from the stock's own historical value. The historical data can be used directly to form the support level and the resistance or they can be plugged into many technical indicators for further investigation. Conventional research addressing this research problem have generally employed the time series analysis techniques (i.e. mixed Auto Regression Moving Average (ARMA)) [3] as well as multiple regression models. Considerable evidence exists and shows that stock market price is to some extent predictable [4].

Within the context of Machine Learning, researchers have dwelt into a number of areas for prediction purposes; these includes Applied Mathematics (Yue et al.,2008, Berwald et al., 2011, Khudabukhsh et al.,2012), Business and Finance (Pompe et al., 1997, Holmes et al., 1998, Huang et al., 2005, Shah,2007, Kinlay et al.,2008, Fletcher et al,2008) , Computer Science (Boetticher,1994, Murray et al.,2005, Singh et al., 2006, Gammerman et al.,2007, Andrzejak et al., 2008, Alonso et al., 2010, Kumar et al., 2012, Haffey, 2012, Malhotra et al., 2012), Life Science (Muggleton et al., 1992, Calder et al., 1996, Demser et al., 2005, Cruz et al., 2006, Garzon et al., 2006, Cheng et al., 2008, Nugent, 2010, Kruppa et al., 2012, Qatawneh et al., 2012, Miller,2012), Energy (Gross et al., 2005, Arnold et al., 2006, Sharma et al., 2011), Sports (Lyle,2005, Joseph et al.,2006, Warner, 2010, Davis et al,2012), etc.

A closer look at the previous works done in predicting stock markets reveals Shah,2007 [5] whose paper discusses the application of Support Vector Machines, Linear Regression, Prediction using Decision Stumps, Expert Weighting and Online Learning in detail along with the benefits and pitfalls of each method. The main goal of the project was to study and apply as many Machine Learning Algorithms as possible on a dataset involving a particular domain, namely the Stock Market, as opposed to coming up with a newer (and/or better) algorithm that is more efficient in predicting the price of a stock.

Kinlay et al., 2008 [6] they applied SVM techniques to forecast market direction in the S&P 500 index, and also used a competitive model framework provided by the 11Ants modeling system to select the best performing combinations of non-linear models employing a variety of non-linear classification techniques. Fletcher et al,2008 [7] used the following algorithms, Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Relevance Vector Machines (RVM) to predict daily returns for an FX carry basket.

As for the most recent works that featured to predict Nigerian Stock Markets, three out of five conducted researches (2009 -2011) used the ANN models. Akinnwale et al (2009) [8] used error back propagation algorithm and regression analysis to analyze and predict untraslated and translated Nigerian Stock Market Price (NSMP). Based on the findings of the study, translated NSMP prediction approach was more accurate than untranslated NSMP using either regression analysis or error back propagation algorithm.

Olabode et al [9] presented the use of a neural network simulation tool for stock market price, where various neural models like Multi-Layered Perceptron (MLP), Radial Basis Function (RBF), Generalized Regression Neural Networks (GRNN), Generalized Feed Forward Neural Networks (GFFNN) and Time Lagged Recurrent Networks (TLRN) were tested. The TLRN network architecture with one hidden layer and five processing elements was able to model the problem, as it came out to be the best model with good generalization capability.

Bello et al (2011) utilized ANNs model to predict closing price of AshakaCem Security in Nigerian Stock Market price index. They employed Feed Forward Artificial Neural Network (FFANN) Architecture and obtained results, which were evaluated on four performance indicators [Mean Square Error (MSE), Correlation Coefficient (r), Normalize Mean Square Error (NMSE) and Mean Absolute Error (MAE)] [10].

Whereas Agwuegbo et al (2010) urged that "The daily behaviour of the market prices revealed that the future stock prices cannot be predicted based on past movements" [11]. Though the result from the study provided evidence that the Nigerian stock exchange is not efficient even in weak form and that NSE follow the random walk model; thus concluded that Martingale defines the fairness or unfairness of the investment and no investor can alter the stock price as defined by expectation. The other work found in the literature, that did not make use of the ANN is the one presented by Emenike K.O (2010) [12], the Autoregressive Integrated Moving Average *(p,d,q)* model [ARIMA] was used to models and forecasts stock prices of the Nigerian Stock Exchange. The predictions failed to match market performance between certain periods of time, thus the adequacy of ARIMA (1.1.1) model to forecast the NSE index was questioned. The researcher concluded that the deviations found between forecast and actual values indicate that the global economies crisis destroyed the correlation relationship existing between the NSE index and its past.

## 3. The Naive Bayes Model

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classifier is based on Bayes' theorem. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computation involved and, in this sense, is considered “naive”.

## 3.1 Bayes Theorem

Let $X = \{x_1, x_2, \dots, x_n\}$ be a sample, whose components represent values made on a set of *n*-attributes. In Bayesian terms, $X$ is considered “evidence”. Let H be some hypothesis, such as that the data X belongs to a specific class C. For classification problems, our goal is to determine P(H|X), the probability that the hypothesis H holds given the ”evidence”, (i.e. the observed data sample X). In other words, we are looking for the probability that sample X belongs to class C, given that we know the attribute description of X. P(H|X) is the a posteriori probability of H conditioned on X.

## 3.2 Naïve Bayesian Classifier

The naive Bayesian classifier works as follows:

Let T be a training set of samples, each with their class labels. There are $k$ classes, $C_1, C_2, \dots, C_k$. Each sample is represented by an n-dimensional vector, $X = \{x_1, x_2, \dots, x_n\}$, depicting $n$ measured values of the $n$ attributes, $A_1, A_2, \dots, A_n$, respectively.

Given a sample X, the classifier will predict that X belongs to the class having the highest a posteriori probability, conditioned on X. That is X is predicted to belong to the class $C_i$ if and only if $P(C_i|X) > P(C_j|X)$ for $1 \le j \le m, j \neq i$.

Thus we find the class that maximizes $P(C_i|X)$. The class $C_i$ for which $P(C_i|X)$ is maximized is called the maximum posteriori hypothesis. By Bayes’ theorem

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (3.1)$$

As P(X) is the same for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class a priori probabilities, $P(C_i)$, are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \cdots = P(C_k)$, and we would therefore maximize $P(X|C_i)$. Otherwise we maximize$P(X|C_i)P(C_i)$. Note that the class a priori probabilities may be estimated by $P(C_i) = freq(C_i, T)/|T|$.

Given data sets with many attributes, it would be computationally expensive to compute $P(X|C_i)$. In order to reduce computation in evaluating$P(X|C_i)P(C_i)$, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample. Mathematically this means that

$$P(X|C_i) \approx \prod_{k=1}^{n} P(x_k|C_i) \quad (3.2)$$

The probabilities $P(x_1|C_i), P(x_2|C_i), \dots, P(x_n|C_i)$ can easily be estimated from the training set. Recall that here $x_k$ refers to the value of attribute $A_k$ for sample X.

(a) If $A_k$is categorical, then $P(x_k|C_i)$ is the number of samples of class $C_i$ in T having the value $x_k$for attribute $A_k$, divided by freq($C_i$, T), the number of sample of class $C_i$ in T.

(b) If $A_k$ is continuous-valued, then we typically assume that the values have a Gaussian distribution with a mean $\mu$ and standard deviation $\sigma$ defined by

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} exp - \frac{(x-\mu)^2}{2\sigma^2} \qquad (3.3)$$

so that

$$P(x_k|C_i) = g(x_k, \mu_{ci}, \sigma_{ci}) \qquad (3.4)$$

We need to compute $\mu C_i$ and $\sigma C_i$, which are the mean and standard deviation of values of attribute $A_k$ for training samples of class $C_i$.

In order to predict the class label of X, $P(X|C_i)P(C_i)$ is evaluated for each class $C_i$. The classifier predicts that the class label of X is $C_i$if and only if it is the class that maximizes$P(X|C_i)P(C_i)$.

The Laplacian correction (or Laplace estimator) is a way of dealing with zero probability values.

Recall that we use the estimation $P(X|C_i) \approx \prod_{k=1}^{n} P(x_k|C_i)$ based on the class independence assumption. What if there is a class,$C_i$, and X has an attribute value, $x_k$, such that none of the samples in $C_i$ has that attribute value? In that case $P(x_k|C_i) = 0$, which results in $P(X|C_i) = 0$ even though $P(x_k|C_i)$ for all the other attributes in X may be large. We can assume that our training set is so large that adding one to each count that we need would only make a negligible difference in the estimated probabilities, yet would avoid the case of zero probability values.

## 3.3 Methodology

The objectives of this work is to illustrate that Naive Bayes can effectively be used to predict the Nigerian Stock Exchange Market (NSE) index values using previous day’s index values, and previous day’s NGN/USD exchange rate.

In this study the following input variables would be considered to ultimately affect the stock exchange market index value.

- NSE All Share index (according to closing price) (NSE_ASI)
- NGN/USD exchange rate (NGN_USD)
- NSE Market Capitalization (according to closing price) (NSE_MCAP)
- Volume (VOL_CLOSING)
- Value (VAL_CLOSING)

Considering the input variables, the following system model was considered for the prediction stock exchange market index value:

$$VAL_{cls}=f(NSE_{ASI},NGN_{\$},NSE_{Mcap},VOL_{cls})(3.5)$$

Experimental data were downloaded from the websites of three prominent/registered Nigerian stock brokers these are Cowry, CashCraft and BGL. The data collected is for a period of 570 days starting from January 4, 2010 to April 30, 2012 excluding weekends and public holidays.

While pre-processing our data the mean of each the five attributes ($NSE_{ASI}$,$NGN_{\$}$,$NSE_{Mcap}$,$VOL_{cls}$ and $VAL_{cls}$) were used to further transformed data into nominal values of "small" and "large" for the first four attributes, while nominal values of "low" and "high" were used for the fifth attribute. We implemented the Naive Bayes algorithm using the WEKA software and results were obtained as presented in section 4 below.

## 1. Experiment and Results

4.1 Sample of the raw data showing a noisy data

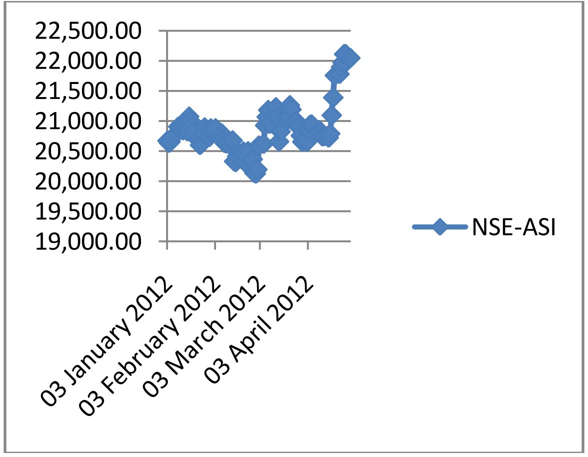

Figure 1: Chart showing a sample of NSE-ASI [Jan-April 2012]

4.2 Results of the Naive Bayes Algorithm

Table1:Statistical Inferences of the 570 instances trained with a 10-fold cross-validation Naïve Bayes Algorithm

Table1A:

| Attribute | Mean | |
|---|---|---|
| | Bad | Good |
| NSE-ASI | 23,048.03 | 24,650.77 |
| NGN_USD | 151.46 | 149.93 |
| NSE_MCAP | 6,862.00 | 6,907.30 |
| VOL_CLS | 272.47 | 443.79 |
| VAL_CLS | 1,876.37 | 3,782.70 |

Table1B:

| Attribute | Standard Deviation | |
|---|---|---|
| | Bad | Good |
| NSE-ASI | 2,170.12 | 2,128.19 |
| NGN_USD | 3.04 | 2.49 |
| NSE_MCAP | 903.57 | 1,000.52 |
| VOL_CLS | 115.76 | 197.72 |
| VAL_CLS | 434.41 | 1,177.39 |

Table1C:

| Attribute | Weight Sum | | Precision | |
|---|---|---|---|---|
| | Bad | Good | Bad | Good |
| NSE-ASI | 348 | 222 | 14.7641 | 14.7641 |
| NGN_USD | 348 | 222 | 0.0662 | 0.0662 |
| NSE_MCAP | 348 | 222 | 6.9139 | 6.9139 |
| VOL_CLS | 348 | 222 | 4.2691 | 4.2691 |
| VAL_CLS | 348 | 222 | 19.0245 | 19.0245 |

Table2: Results of Stratified Cross validation for the 570 instances

| | |
|---|---|
| Correctly Classified Instances | 521 (91.4035%) |
| Incorrectly Classified Instances | 49 (8.5965%) |
| Kappa statistic | 0.8167 |
| Mean absolute error | 0.1225 |
| Root mean squared error | 0.2435 |
| Relative absolute error | 25.7430% |
| Root relative squared error | 49.9441% |

=== Confusion Matrix ===

a b <-- classified as

$$\begin{pmatrix}332 & 16\\ 33 & 189\end{pmatrix}\begin{matrix}|a=Bad\\ |b=Good\end{matrix}$$

## 5 Discussion

The Chart in figure 1 clearly shows a noisy data, thus before subjecting our data to the Naïve Bayes implementation and analysis, we employed robust statistical techniques to alleviate the problem of noise sensitivity [13]; this process had greatly enhanced the quality of our data.

The Naive Bayes algorithm learned all the 570 instances as categorized under the five attributes (Table 1), with an unbiased estimates of the instances as seen from the weight sum column; also the precision results seems to have a positive influence on the learning process [14].

The results of the stratified cross validation of the instances (Table 2) came up with medium values of relative absolute error (25.7430%) and root relative squared error (49.9441%), however other indices most especially Kappa

Statistics, mean absolute error and root mean squared error depicts a positive results.

## 6 Conclusion/Further Works

Based on our findings, it is clear that through the processes of data mining (semi-transformation of the data before analyzing it) the Naive Bayes algorithm can effectively predict NSE.
Further works will involve research work on using other Machine learning algorithms to predict the NSE.

## Acknowledgments

The authors are thankful to IJACT Journal for the support to develop this document.

## References

[1] Marwala L.R., ‘Forecasting the Stock Market Index Using Artificial Intelligence Techniques’2007

[2] Hellstrom T. And Holmstrom K., Predicting the Stock Market, Technical Report Series, Malardalen Univeristy Sweden, 1998.

[3] S.M. Kendall and K.ord, *Time Series*, 3rd ed. (Oxford University Press, New York, 1990)

[4] A.W. Lo and A.C. Mackinlay, Stock market prices do not follow random walks: Evidence from a simple specification test, *Review of Financial Studies 1 (1988) 41-66*

[5] Shah H.V., 2007, Machine Learning Techniques for Stock Prediction, **Foundations of Machine Learning |** *Spring 2007*

[6] Kinlay J andRico D., Can Machine Learning Techniques be used to predict Market Directions?, Journal of Finance, 2008

[8] Akinwale A.T, Arongundade O.T and Adekoya A.F. (2009), *Translated Nigeria Stock Market Prices using Artificial Neural Network for Effective Prediction*, ***Journal of Theoretical and Applied Information Technology, 36-43.***

[9] Olabode O., Adeyemo A.B, and Coker C.O., *Artificial Neural Network for Stock Market Forecasting*, *www.sciencenigeria.org/index2.php?option=com_docman...doc*.

[10] Bello M.Y. and Chiroma H. (2011), *Utilizing Artificial Neural Network for Prediction in the Nigerian Stock Market Price Index*, ***Computer Science and Telecommunications 2011/No.1(30)***

[11] Agwuegbo S.O.N., Adewole A.P., and Maduegbuna A.N. (2010), *A Random Walk Model for Stock Market Prices*, ***Journal of Mathematics and Statistics 6(3): 342-346***

[12] Emenike K. O. (2010), *Forecasting Nigerian Stock Exchange Returns: Evidence from Autoregressive Integrated Moving Average (ARIMA) Model*, ***Social Science Research Network.***

[13] Chu F., Wang Y., and Zaniolo C., (2004), Mining Noisy Data Streams via Discriminative Model, Discover Science-Lecture Notes in Computer Science, Vol.3245,2004, pp 47-59

[14] Mendes R.M.S. and Godinho M.A.B., Knowledge of Results Precision and Learning: A Review, Sousa 2003, pp 23.