International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

A PROGNOSIS APROACH FOR STOCK MARKET

MS. V. Anitha[1], K.Charitha[2], K.Harshitha[3]

[1]Assistant Professor, CSE Department, SCSVMV University, Kanchipiram, Chennai, India
[2,3]CSE Department, SCSVMV University, Kanchipuram, Chennai, India

---

Abstract: *The Stock market process is full of uncertainty and is affected by many factors. Hence the Stock market prediction is one of the important exertions in finance and business.*

In this system technical analysis are considered. Technical analysis is done using historical data of stock prices by applying machine learning. The learned model can then be used to make future predictions about stock values. The system is trained by using machine learning algorithm.

Then the correlation between the stock values is analyzed. The learned model can then be used to make future predictions about stock values. It can be shown that this method is able to predict the stock performance.

In this system we applied prediction techniques approach in order to predict stock prices for a sample companies.

In stock predictions, a set of pure technical data, fundamental data, and derived data are used in prediction of future values of stocks. The pure technical data is based on previous stock data while the fundamental data represents the companies' activity and the situation of market.

1. INTRODUCTION

Future being a mystery is always a challenging task to predict. From the ages, human nature has always been more curious about the future. Forecasting refers to an approach of predicting what is likely to occur in the future by observing what has happened earlier in the past and what is occurring at present. In other words, it is just similar to driving a car in forward direction by keeping an eye on the rear-view mirror of a car. Forecasting is an important problem but with vital importance in all areas of real world like business and industry, medicine, social science, politics, finance, government, economics, environmental sciences and others. In recent years, with the rise of social media and other promising applications, stock market forecasting has attracted huge interest from people in general and business in particular. Advances in financial sectors are responsible for growth and stability of overall economy [1]. In business domain, forecasting is considered as one of the difficult tasks owing to the various complexities of the market [2] [3] [4]. But it is important since it helps to plan for future by providing a solid idea about how to allocate the resources and plan for foreseen costs in the forthcoming period of time. Investors always try to monitor the risks in real time so that the return on investments could be higher. Forecasting helps in safeguarding the trade of securities among the buyers and the sellers as well as elimination of the risks involved. This paper discusses an ARIMA (Auto Regressive Integrated Moving Average) model for prediction of stock market movement. An ARIMA model is a vibrant uni-variate forecasting method to project the future values of a time series. The remaining of the paper is arranged as follows. Section II describes the forecasting process. Section III discusses the forecasting techniques, while section IV discusses the financial forecasting. Section V presents the Time Series Analysis. In section VI, we try to explain in detail, the various statistical models for forecasting. Data collection and methodology are discussed in section VII, while the final section of this paper provides a brief conclusion

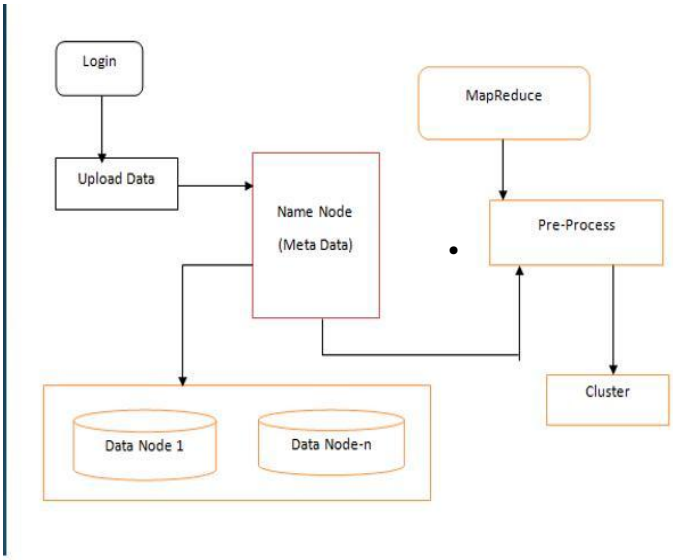
2. *ALGORITHM*

SUPPORT-VECTOR MACHINES:

- *In machine learning, support-vector machines SVMs, also support-vector networks are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.*
- *Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).*

- *An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.*
- *In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.*
- *When data are unlabeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups.*
- *The support-vector clustering algorithm, applies the statistics of support vectors, developed in the support vector machines algorithm, to categorize unlabeled data, and is one of the most widely used clustering algorithms in industrial applications*

3. ARCHITECTURE DIAGRAM



4. WORKING MODEL

- *User authentication*
- *Data uploading*
- *Preprocessing*
- *Data clustering*
- *Svm classification*
- *Report Prediction*

Description of MODULES:

- ***User Authentication*:**

Every last client login the page at that point makes the • exchange and utilize this application. Validness is confirmation • that a message, exchange, or other trade of data is from the source it cases to be from. Validness includes verification of character. We can check validness through confirmation. Enroll and login choice in landing page. Every single client needs to enlist as the new client for login. Client need to Fill the all prerequisite for security reason just, so fill the all subtle elements unique points of interest. Every one of the subtle elements spared in various ways. Make new table for every client and spare points of interest in like manner table.

DATA UPLOADING:

The readied informational index will store the Hadoop document framework. HDFS occurrences are partitioned into two segments: the name node, which keeps up metadata to track the arrangement of physical information over the Hadoop case and data nodes, which really store the information.

The information stacking to the hdfs utilizing hdfs url way. The transferred information will be kept up by the name hub and information hubs. The informational index traits are kept up in the name hub like record name, measure, get to consent, and so forth.

The crude information kept up by the information hubs. The information hub controlled by the name node. The transferred information can't change any qualities in light of the fact that hdfs have compose once perused many time property.

PREPROCESSING:

Information preprocessing is an information mining method that includes changing crude information into a reasonable arrangement. True information is frequently inadequate, conflicting, as well as ailing in specific practices or drifts, and is probably going to contain numerous blunders. Information preprocessing is a demonstrated strategy for settling such issues.

The transferred information recover from the hdfs. The recovered information going to the Map Reduce calculation and information will be composed into organized configuration. In this procedure have expelling the unusable qualities from the informational indexes. The information diminishment process is lessened portrayal of the information in an information distribution center.

DATA CLUSTERING:

To gather those information into those bunches whose stock information class has been as of now characterized. In this way it develops a procedure to anticipate the promoting of the up and coming days. This one procedure gathering the information and ought to be made out of focuses isolated by little separations, in respect to the separations between groups. The information will gathering in light of the value, open, high, low, shut and time. In this bunching will apply the map reduce with svm approach. It give more productive in high volume information bunching process.

- *For SVM, it's the one that maximizes the margins from both categories. In other words: the hyper plane (remember it's a line in this case) whose distance to the nearest element of each category is the largest.*
- *CandidateSV = {closest pair from opposite classes} while there are violating points do*
- *Find a violator*
- *CandidateSV = CandidateSV S violator*
- *if any αp < 0 due to addition of c to S then*
- *CandidateSV = CandidateSV \ p*
- *repeat till all such points are pruned*
- *end if*
- *end while*
- *SVM is generally used for text categorization. It can achieve good performance in high-dimensional feature space. An SVM algorithm represents the examples as points in space, mapped so that the examples of the different categories are separated by a clear margin as wide as possible.*
- *The basic idea is to find the hyper plane which is represented as the vector w which separates document vector in one class from the vectors in other class.*
- ***REPORT PREDICTION :***
- *The cluster value will be different ranges. Those values are gathered and compared to each other's. Finally we will get the low and high result based on the calculation. The predicted values will be given the graphical representation graph.*

5. FORECASTING TECHNIQUE

Forecasts are being employed in a wide range of situations. In spite of all the numerous real life circumstances that involve forecasts, there exist only two types of forecasting techniques to be implemented[9]: a) Qualitative Forecasting models. b) Quantitative Forecasting models. The Qualitative forecasting models are generally subjective in nature and are mostly grounded on the opinions and judgments of experts. Such types of methods are generally used when there is little or no past data available that can be used to base the forecast. An example of the qualitative forecasting model is the Delphi method which engages a board of experts who are supposed to be well-informed about the problem. Hence, the outcome of the forecast is based upon the knowledge of the experts regarding the problem. On the other hand, the Quantitative forecasting models make use of the data available to make predictions into future. The model basically sums up the interesting patterns in the data and presents a statistical association between the past and current values of the variable. Likewise, we can say, that quantitative forecasting models are used to extrapolate the past and present behaviour into future. Some examples of the Quantitative models include the regression analysis models, smoothing models and the time series models.

6. FINANCIAL FORECASTING

A Financial forecast can be elaborated as a forecast regarding the future business circumstances that are expected to affect a company, organization, or a country. A financial forecast visualizes the movements in relevant historical data and then projects these movements in order to help the decision-makers by providing information regarding the forthcoming financial status of the company. Simply we can say that, a financial forecast is a business plan or budget for a business. It is basically considered as an estimate of two vital forthcoming financial outcomes of a business – the projected revenue and the costs. Prediction of the financial state of a business is never an easy task; with most of the forecasts go wrong. But still it is a better idea to have an educated guess about the future than to not forecast at all, since "Best" educated guesses about future are more valuable for purpose of planning and budgeting. There are many advantages of an efficient financial forecast as below: a) Controls the financial practicability of a new business project. Thus aides in designing of models for how the business would perform economically, if certain approaches, procedures and tactics are undertaken. b) Helps in comparing the real financial operation with the forecasted financial plan and make modifications where needed. c) Drives the business in an accurate direction and controls the flow of cash. d) Specifies a point of reference against which the future performance can be supervised. e) Identifies the probable threats and the cash deficits in order to keep the business away from the financial catastrophe.

f) Helps in knowing the future cash needs and whether any additional borrowing is needed.

A. STOCK MARKET PREDICTION

A "share market or equity market or a stock market" is a public market that exists for issuing, buying and selling of stocks or shares [10][43]. A Stock denotes a partial ownership in a company or an industry, with rights to share in its profits. A person, who invests in a stock or buys a stock of a company, is termed as stockholder of that company. A stock market is a dynamic ingredient of a free- market economy where organizations get access to capital in exchange by providing the sponsors a share in proprietorship of the organization. A stock market plays a significant role for organizations to raise revenue along with the debit markets. The stock market allows the business units to be openly traded and raise surplus financial capital for growth and development by selling shares of the proprietorship of a company in a public market. History has proved that stock prices and prices of other assets have a dynamic impact on the economic activity and is also an indicator of social mood state. An economy is considered as a rising economy if its stock market is on rise. India, being one amongst the fast developing economies of the world provides investors, both domestic as well as foreign investors, an opportunity to make right investments in Indian stock market. India has two main stock exchanges as: □ "National Stock

Exchange (NSE)" □ "Bombay Stock Exchange (BSE)"

The BSE is the oldest stock exchange in India that came into existence in 1875 while the NSE came much later and started working in 1994. All of the major companies or business firms in India are listed under these two stock exchanges. BSE has around 5000 firms listed to its name while the rival NSE has listed about 2000 firms to its name [11]. In spite of having a lesser number of firms listed to its name, NSE still enjoys the dominant share in share trading with about 70% of market share[11] [12]. The performance of overall stock market is calculated by Index. It is an indicator of the overall stock market movement. Indian stock market has two main indicators or indexes which are as: □ Nifty.

□ Sensex. There are many other indexes that reflect the

performance of a particular sector example bank index, IT index, automobile index and many others, but the above two are the prime ones. The Sensex (or sensitive Index) comprising of 30 (thirty) stocks, reflects the whole market sentiments of major

firms listed under Bombay Stock Exchange (BSE). The UP or DOWN movement of Sensex indicates that the majority of stocks prices under BSE have gone up or down respectively. Nifty, on the other hand comprises of 50 (Fifty) stocks, is an indicator of the firms listed under National Stock Exchange (NSE) [13]. Stock market prediction can be concluded as an approach to estimate the upcoming worth of a company‟s stock traded on a stock exchange. A constructive prediction of a company‟s stock price can bring fortunes to the company.

Calculation of the Index There are two main components that are required while calculating the value of index for the next day. These components include "the index value" and "the total market capitalization of the previous day". The index is calculated as follows: ***Index Value = (Todays Market Capitalization / Yesterdays Market Capitalization) x Yesterdays Index point***

7. FORECASTING MODEL

The selection of prediction model is of remarkable significance as it reveals the fundamental structure of the time series. Time series models can be "linear or nonlinear" based on whether the present value of the series is a "linear or non-linear" function of earlier observations. Univariate time series models try to interpret a number of economical phenomena through the historical behaviour of a dependent variable. There are mainly two widely used linear models [24][25][26][27][45][47], "Auto-Regressive models (AR)" and "Moving Average (MA) models". A grouping of these two models result in the formation of another model referred as Auto-regressive Moving Average (ARMA). We have one similar type of model as "AutoRegressive Integrated Moving Average model" (ARIMA) [28][29][30][33][46]. All these models have been elaborated below:

A. AR (p) MODEL An "Auto regressive (AR) model" is used to calculate the future behavior of a variable under consideration, using a linear combination of historical values of the variable [34][35]. The word "auto-regression" designates that there is a regression of the variable against itself. It could be considered as a function of past values. i.e. "yt = f (yt-1, yt-2, yt-3, yt-4, …….. ϵt)" Auto regressive model that

is influenced by „p‟ of its earlier values, referred as AR(p) is represented as:

"yt = c + ϕ1yt-1 + ϕ2 yt-2 + + ϕpyt-p + ϵt "

yt = c + i=1Σp ϕ1yt-I + ϵt (1) where ϵt is the error term.

In AR(p) model, P is the parameter, which can change its values as: When p = 0, AR(0) becomes yt = c When p = 1, AR(1) becomes yt = c + ϕ1yt-1 and so on for the values of p as 2,3… Normally we restrict autoregressive models to stationary data, but When p≥3, the restrictions are much more complicated. We need to find the best value of p for forecasting.

B.

MA (q) MODEL Instead of using the earlier values of a variable for forecasting as in case of AR(p) model, the "Moving Average (MA) model" uses earlier errors terms for prediction. MA model can be considered as a function of error

terms as: yt = f (ϵt, ϵt-1, ϵt-2, ϵt-3 ,. ……. ϵt-q)

In regression, we get an error term when we regress a series with its past values as : Regression of yt over yt-1: yt = μ + θ1 ϵ t-1 +

ϵ1 where μ= constant, and ϵ1 = error

Similarly we get other error terms as ϵ2, ϵ3 and so on when we regress the series with different values. So instead of

using the past values, we use the error terms in this model. As such the "moving average model" can be put forward as: "yt = μ + θ1 ϵ t-1 + θ2

ϵt-2 + ……. + θq ϵt-q + ϵt" yt = μ +

j=1Σq θj ϵt-j + ϵt (2) The error terms ϵt

are supposed to be white noise processes, i.e having "zero mean" and "variance constant" σ2. μ represents the "mean of the series", and "θj (j = 1,2,3,4..q)" represent the parameters of the model and "q is the order of the model". MA

model is more difficult than AR model to fit to a time series as the random error terms are not foreseeable [37].

C. ARMA (p,q) MODEL "Autoregressive" (AR) along with "moving average" (MA) models are used collectively to get a new unit of time series models referred as "ARMA models" (i.e. AR + MA = ARMA model)[35][36] [37] [44]. The general notation for the "ARMA (p,q) model" is as:

"yt = (c + ϕ1yt-1 + ϕ2 yt-2 +… + ϕpyt-p + ϵt) + (μ + θ1 ϵ t-1 + θ2 ϵt-2 + .. + θq ϵt-q + ϵt)" yt = c + ϵt + i=1Σp ϕ1yt-1 + j=1Σq θj ϵt-j

(3) In the above given equation, the parameters„ p" and „q" correspondingly refer to the "auto-regressive" and "moving average model".

D. ARIMA (p,d,q) MODEL "ARIMA (AR+I+MA)" stands for "Auto-regressive Integrated Moving Average". This model is also often called by the famous "Box-Jenkins model". The "ARMA model" is best suited for the stationary time series data but the thing is that most of the time series data from real world shows non-stationary behavior. This model claims that a non-stationary series could be changed to stationary by means of differencing it[37][47] . The common form of an "ARIMA model" for yt is as:

" yt = c + ϕ1yt-1 + ϕ2 yt-2 +… + ϕpyt-p + θ1 ϵ

t-1 + θ2 ϵt-2 + .. + θq ϵt-q + ϵt " (4)

Where yt is "differenced time series" that might have been differenced once or more. This model is called as "ARIMA (p,d,q)" model, in which the parameters p, d, q represent the following: „p" is the "order of auto-regressive part", „d" is the "degree of differencing", „q" is the "order of the moving average part".

The "ARIMA model" combines 3 basic methods:

Auto-regressive (AR)

This specifies that the values of the given time series is to be regressed with its own lagged value. This is specified through the "p value of an ARIMA model".

Differencing part

(I for integrated) Integration is implied as the inverse of differencing. It is the degree on differencing that needs to be done on data. In order to transform a "non-stationary time series into a stationary one", the series needs to be differenced. Differencing can be implied as: D(i)

= Data (i) – Data(i-1). This differencing of the time series is represented by the „d" value of the "ARIMA model". There are some situations that may arise with the value of „d", likewise below: When „d" = 0, it signifies that the series under consideration is stationary, so we don"t require to take the difference of it. If „d" =1, it signifies that the current series is not stationary, we need to take the first difference of the series. If „d" = 2, it signifies that the series under consideration has been differenced two times.

Moving Average (MA)

The "moving average" component of an ARIMA model is denoted by „q". This simply refers to the total number of lagged values of the error term. For instance, when „q" = 1 it means that there exists an error term and there is auto-correlation with one lag.

8. DATA COLLECTION

The work in this research paper is focused on the data regarding the stock market. Data considered as raw oil, is being generated with every passing second. [31]. This pragmatic study began with the analysis of Indian stock market data related to Sensex and Nifty. The publically available Stock market data sets contain historical data about all the stocks [32] has been collected from yahoo. The dataset specifies the "opening price, lowest price, closing price, highest price, adjusted closing price and volume" against each date. The historical data of the Indian stock market collected through a span of five years beginning from "January 2012 to December

2016" has been taken into consideration for this work. The data has been divided into two parts – "the training part and the testing part". The "training part" from the time series data is used for formulation of the model while the "testing part" is used for the validation of the proposed model.

METHODOLOGY

Auto-regressive processes have a certain degree of unpredictability or randomness built in, that occasionally makes it capable to predict future trends pretty well. However, this needs to be assumed that they are never 100% accurate [38][39]. After analyzing the time series data collected regarding the stock market, the 1st thing to do, is to ensure that the series is stationary or not. If the series is non-stationary, then the series has to be differenced so as to make it stationary. As such, we need to find the auto-co-relation and partial auto co-relation of the series. ARIMA model relies on the stationarity of the series [39]. So we will start with a fleeting portion about the stationary time series.

STATIONARITY OF A TIME SERIES

A time series needs to be lacking trend and seasonality, in order to be stationary. Such type of time series are characterized by having a constant variance and constant mean over a given period of time. The "trend and seasonality" component may affect a time series at different instants [40]. As ARIMA model takes into account, the earlier values of the series to model its prediction, so modeling a steady series with regular properties involves little insecurity. In order to design a model that is efficient in predicting future values of series, the primary time series has to be Stationary one. There are certain tests that assist in checking whether the series is stationary or not [33]. Some of these include "W-D test", "Auto-correlation function (ACF)", "Partial auto-correlation function (PACF)", "L-jung-Box test", "t-statistic test", the "Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test" and "Augmented Dickey-Fuller unit root test (ADF test)". Example- The "augmented Dickey-Fuller (ADF)" test is a standard statistical test for stationarity. In the ADF test, if the value of „p" is less than 0.05 or 5% level for a time series, then the series is supposed to be stationary. However, there are cases when the series is non- stationary, at such times the "p value is greater than 0.05 or 5% level". A„non-stationary time-series" needs to be corrected by means of differencing [41]. An easy way to alter a nonstationary time series into stationary one needs to compute the differences between consecutive observations. This is referred as differencing i.e.(yt –yt-1).

This differencing is the „I" (integration) part of ARIMA model and denoted by „d". If we put on the differencing method twice or more, this gives birth to 2nd order difference and so on. These differenced values are then noted and thus it gives rise to a new dataset of the time series nature that can be used to test and discover new remarkable statistical properties and correlations.

All the stock market investors aim to maximise the returns over their investments and minimise the risks associated. Stock markets being highly sensitive and susceptible to quick changes, the main aim of stock trend prediction are to develop new innovative approaches to foresee the stocks that result in high profits. This research tries to analyse the time series data of Indian stock market and build a statistical model that could efficiently predict the future stocks.

9. CONCLUSIONS

- *The stock advertising information is expanding day by day with the information will be produced in various associations.*
- *The information will gathered and stacked into the HDFS utilizing the Hadoop system.*
- *The put away information broke down utilizing map reduce calculation used to characterize and grouping process.*
- *The mining method to foresee the stock promoting status in view of utilizing the authentic information like value , low ,high ,open and close everything utilizing the verifiable information.*

10. REFERENCES

- *Khodabakhsh, A., Arí, I., Bakír, M. and Ercan, A.O., 2018. Multivariate Sensor Data Analysis for Oil Refineries and Multi-mode Identification of System Behavior in Real-time. IEEE Access, 6, pp.64389-64405*
- *Raicharoen, T., Lursinsap, C., & Sanguanbhokai, P. (2003, May). Application of critical support vector machine to time series prediction. In Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on (Vol. 5, pp. V-V). IEEE.*

- *Zhang, M. and Pi, D., 2017. A New Time Series Representation Model and Corresponding Similarity Measure for Fast and Accurate Similarity Detection. IEEE Access, 5, pp.24503-24519*
- *Pati, J., Kumar, B., Manjhi, D. and Shukla, K.K., 2017. A Comparison Among ARIMA, BP-NN, and MOGA-NN for Software Clone Evolution Prediction. IEEE Access, 5, pp.11841-11851*
- *Zhang, Q., Li, F., Long, F. and Ling, Q., 2018. Vehicle Emission Forecasting Based on Wavelet Transform and Long Short-Term Memory Network. IEEE Access, 6, pp.56984-56994.*
- *Conejo, Antonio J., Miguel A. Plazas, Rosa Espinola, and Ana B. Molina. "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models." IEEE transactions on power systems 20, no. 2 (2005): 1035-1042.*
- *Anaghi, M.F. and Norouzi, Y., 2012, December. A model for stock price forecasting based on ARMA systems. In Advances in Computational Tools for Engineering Applications (ACTEA), 2012 2nd International Conference on (pp. 265-268). IEEE.*
- *A. Akbar, G. Kousiouris G, H. Pervaiz, J. Sancho J,F. Carrez, and K. Moessner, "Real-Time Probabilistic Data Fusion for Large-Scale IoT Applications," IEEE Access, 6:10015-27, 2018.*
- *E. Olmezogullari and I. Ari, "Online association rule mining over fast data," In Big Data (BigData Congress), IEEE International Congress on, pp. 110– 117, 2013, IEEE.*
- *E. Lughofer, M. Pratama, and I. Skrjanc, "Incremental Rule Splitting in Generalized Evolving Fuzzy Systems for Autonomous Drift Compensation,"IEEE Transactions on Fuzzy Systems, 2017.*