Forecasting on Stock Market Time Series Data Using Data Mining Techniques

Ananya Preeti Padma [1], Anil Kumar Mishra [2]

[1]Asst. Professor, Einstein Academy of Technology & Management, Bhubaneswar
[2]Professor, Einstein Academy of Technology & Management, Bhubaneswar

Abstract: *A stock exchange market depicts savings and investments that are advantageous to increase the effectiveness of the national economy. There are many factors that affect share prices. However there is no specific cause for the prices to rise or fall. This makes investment subject to various risks. The future stock returns have some predictive relationships with the publicly available information of present and historical stock market indices. ARIMA (Autoregressive integrated moving average) is a statistical model which is known to be efficient for time series forecasting especially for short-term prediction. In this paper, we propose a model for forecasting the stock market trends based on the technical analysis using historical stock market data and ARIMA model. This model process future stock price indices and provides assistance for financial specialists to purchasing and/or selling of stocks at the right time. The forecast results are visualized using R programming language. Results of ARIMA model have a strong potential for short-term prediction of stock market trends.*
Keywords: *Stock Market, Data Mining, ARIMA, Prediction, Time Series Data.*

I. Introduction

The stock market system consists of the primary market and the secondary market.Primary market deals with the new issues of securities and securities are bought by way of the public issue directly from thecompany. An official prospectus is published under the Corporations Law and contains all the information that is reasonably required to allow you to make an informed investment decision about the company. Secondary market: it is where existing securities are bought and sold. Secondary market deals with outstanding securities. In the secondary market shares are traded among investors. This market is made of organized exchanges and may have a trading floor, where orders are transmitted for execution. Stock market forecasting contains uncovering the market trends, planning investment tactics, identifying the right time to purchase the stocks and which stocks to purchase. A stock exchange or equity business sector is a non-direct, non-parametric framework that is difficult to model with any sensible exactness. The basic assumption made while forecasting stock data is that future market trends are influenced by the stock information available publicly in the past. This means, the historical stock data provides an insight into its future behavior.According to the Random Walk theory for stock markets, "stock market prices evolve according to a random walk and thus cannot be predicted". The theory is further divided into 2 separate parts. The first hypothesis states that successive price changes in an individual security are independent. The second hypothesis states the prices conform to a certain probability distribution. However, it is the probability distribution of data or the form of distribution that allows academicians and investors to forecast stock data. Recent studies have shown that Time Series data analysis techniques provide verifiable information for forecasting stock prices. Time series data is sequence of data collected over specified period of time. Time series data for stock market prediction can be collected on a daily, weekly, monthly or yearly basis. The analysis of the time series data extracts useful statistical information to understand characteristics of data. Time series forecasting techniques involve using models to predict future values based on past information.

R is an open source programming language and software environment for statistical computing and graphics. It has numerous applications in the field of data analysis and widely used by statisticians and data miners. Along with a command line interface, it has several graphic front-ends. R is extensible through functions, extensions and packages, contributed by the global R community.

As of 2016, 7801 additional packages are available for installation. This user created packages like forecast, stats, ggplot2 allows the user to perform specialized statistical and graphical procedures.

RStudio is an open source integrated development environment (IDE) for R. The software is written in C++ programming and uses Qt framework for graphical user interface. It supports direct code execution as well as tools for statistical analysis, debugging and workspace management. There are 2 editions of RStudio, RStudio Desktop and RStudio Server. RStudio Desktop runs the program as a regular desktop application. Using the

RStudio Server, RStudio running on a Linux server can be remotely accessed via a web browser. RStudio allows users to manage multiple working directories using projects. It also has extensive package development tools.

The remaining part of the paper covers the following topics. Literature review discusses about pre published research papers related to forecasting stock market trends. Next discusses about System analysis covers the problem statement and the papers approach to forecast the market trends. Implementation describes in detail the methodology used to predict stock prices. Model simulation describes the R code used to develop the model. The section also consists of visualized outputs.

II. Literature Review

To predict stock returns, scholars and researchers rely upon fundamental analysis and technical analysis. The author [Suresh A.S] [1] describes fundamental analysis as the examination of underlying forces that affect the wellbeing of the economy. Fundamental analysis combines economic, industry and company analysis to derive a stocks fair value known as intrinsic value. According to fundamental analysis if the fair value is not equal to the current stock price, then the stock is either undervalued or overvalued. Fundamental analysis takes into account macroeconomic factors and individual specific factors. Fundamental analysis is believed to be effective predicting long-term trends. The same paper describes technical analysis as a supplement for to fundamental analysis but more focused on predicating the price of a security. Technical analysis takes into account the change in demand and supply of securities as a function of time. Therefore it is preferred over fundamental analysis for short-term and medium term forecasting. Technical analysis is defined as the art and science of forecasting future prices based on the examination of past price movements by the author [C.Boobalan][2]. In addition to past stock prices, technical analysis also considers company fundamentals, broader economic factors, market psychology and prices them into the stock.

The authors [Ayodele A] [Adebiyi] [3] have used the ARIMA model to develop an extensive process of building stock price predictive model by obtaining data from NYSE and NSE. Artificial Neural Networks (ANNs) model is very popular due to its ability to learn patterns from data and infer solution from unknown data. Hybrid approaches also engaged to improve stock price predictive models by exploiting the unique strength of each of them. The results obtained from real-life data demonstrated the potential strength of ARIMA models to provide investors short-term prediction that could aid investment decision making process.

The most efficient way to forecast the future is to understand the present scenarios. The author [Banerjee D] [4] tried to develop an appropriate model that helps to forecast the unseen values of the Indian stock market, based on the information collected on the monthly closing stock indices. Based on the ARIMA model they predict the future stock indices which have the strong performance of the Indian economy. It is very important to understand the present status of the market because for many economists, investors and researchers the Indian stock market is the center of interest. It has been predicted that the performance of the Indian stock market presents a suitable time series ARIMA(1,0,1) model which helps to create the appropriate values of the future indices.

Author [Linhao Zhang][5] describes the effect of public sentiment on stock prices by analyzing Twitter messages. The author examines the effects of tweets on stock prices and also determines which words in the tweet correlate to stock price change. The author uses machine learning techniques like Naïve Bayes classification, Maximum Entropy classification and Support Vector Machines to determine sentiment. Data is fetched from the Twitter's Search API and classified into 2 different datasets for training. However, it was found that the classifiers were not much effective for negation statements. The author also correlates tweet sentiments and stock market prices on an intra-day scale. Yahoo's finance API was used to gather data on 10 exchange traded funds or ETF's. The paper uses the Pearson correlation coefficient to account for any time difference. The experimental results show that for the data to be truly effective a larger time granularity is required. To find deeper correlations, the system needs to be used over a longer period of time with more aggregate data. The author also describes the numerous challenges that analyzing Twitter sentiment poses. The very first challenge is searching for the right tweets without getting too arbitrary. Searching for keywords and interpreting slang jargon is another inherent challenge. For effective results, the system needs to be trained on much more data over a larger period of time. A relatively new method, Approximation and Prediction of Stock Time series data (APST) has been proposed by authors [Vishwanath R.H.], et al. [6]. The system generates a sequence of approximated values using multi-scale segment mean approach after preprocessing historical stock time series data. To identify the similar set of objects, the authors use Euclidian distance approach and find the nearest neighbors. Experimental results show that the average Mean Error Relative and average Mean Absolute Error for APST are 5.90% and 0.37%. This implies that the system shows a high level of accuracy.

To forecast stock price trend the authors [Tao Xing] [Yuan Sun] [7] have introduced a method based on Hidden Markov Model. Hidden Markov Model first proposed by Baum and Egon, which is a kind of Markov

Chain and is used for the pattern recognition technique. This paper finds the hidden relationship existing between the Hidden Markov Model and stock prices. The experimental results show that, this method can get attractive accurate result, particularly efficient in short period prediction.To forecast time series data analysis for stock prediction the aurthors[MahantesgAngadi][Amogh Kulkarni][8] have ARIMA model automate the process of direction of future Stock price indices and provide assistance for financial specialists to choose the better timing for purchasing and/or selling of stocks.

III. System Analysis

A. Problem Statement

The financial market or stock market is complex and evolutionary. It functions as a non-linear dynamic system. According to academic investigations movements in market prices are not random and depend upon numerous factors that correlate it with present and historical stock data. It is not possible for every investor to comprehend the various factors that cause the prices to change. Hence every investor desires a system to predict the future stock prices to help them take appropriate decisions.

B. Existing Systems

Numerous qualitative and qualitative analysis methods have been developed to estimate stock trends. There are various statistical models for forecasting stocks and decide the right time to sell or hold a stock. Depending upon the format of the data, a particular forecasting model can be used by the investor to predict trends.

C. Proposed Study

The paper proposes a model for predicting time series stock market data. The model based on technical analysis using ARIMA aims to automate the process of change of stock price indices. With the help of Data Mining techniques a prediction model is developed. R programing language in RStudio IDE is used for visualizing the experimental results.

IV. Proposed Method

Data mining is used to discover patterns in large data sets and has wide application s in the field of statistics. Data mining techniques are devised to address forecasting problems by providing a reliable model with data mining features. We use the auto-regressive integrated moving average (ARIMA) model to predict the market trends. The complete architecture of the system is shown below.

Fig.1. Proposed Method

System architecture contains the information regarding the constituent elements of a system. It also describes the relationship between these elements. It is a model that provides information about the behavior of a system by breaking it into subordinate systems that perform the same functions. The ARIMA system includes seven major steps to implement the system and each step is explained below.

A. Understanding the Objective

The objective describes the essential requirements of the system. It helps in better understanding of the problem statement as well as the expected results. The objective this paper is to develop a system that can be used by investors to find the direction of the market trends and make right investment decisions. The

experimental results are provided in a graphical format for better interpretation

B. Data Collection

Understanding the objective also aids in collating the right datasets. Data collection involves gathering information relevant to the required variables and measuring them to evaluate outcomes. The paper uses R script to collect data from Google using the function getSymbols() available in the QuantMod package.

QuantMod

Quantmod refers to Quantitative Financial Modelling and Trading Framework for R. It is quantitative tool that helps traders in developing and testing trade based statistical models. The quantmod package makes modelling easier and faster by excluding repeated workflow. The package consists of comprehensive tools for data management and visualization. To extract and load the data from multiple sources we use a method called getSymbols (). As a source for obtaining the stock market data, most of the stock investors use Google finance or Yahoo finance.In our project the OHLC data is not directly downloaded from the Google finance (finance.google.com),or Yahoo finance(finance.yahoo.com) instead a call to getSymbols() is used to fetch data. We didn't specify the source here so the data is downloaded from default reference ie:-www.finance.yahoo.com.

Table 1 intra- day aaplsample data from Google finance

	AAPL.Open	AAPL.High	AAPL.Low	AAPL.Close	AAPL.Volume	AAPL.Adjusted
2007-01-03	12.32714	12.36857	11.70000	11.97143	309579900	10.73159
2007-01-04	12.00714	12.27857	11.97429	12.23714	211815100	10.96978
2007-01-05	12.25286	12.31428	12.05714	12.15000	208685400	10.89166
2007-01-08	12.28000	12.36143	12.18286	12.21000	199276700	10.94545
2007-01-09	12.35000	13.28286	12.16429	13.22429	837324600	11.85469
2007-01-10	13.53571	13.97143	13.35000	13.85714	738220000	12.42201
2007-01-11	13.70571	13.82571	13.58571	13.68572	360063200	12.26833
2007-01-12	13.51286	13.58000	13.31857	13.51714	328172600	12.11722
2007-01-16	13.66857	13.89286	13.63571	13.87143	311019100	12.43481
2007-01-17	13.93714	13.94286	13.54571	13.56428	411565000	12.15918
2007-01-18	13.15714	13.15857	12.72143	12.72429	591151400	11.40648
2007-01-19	12.66143	12.80714	12.58857	12.64286	341118400	11.33348
2007-01-22	12.73428	12.73714	12.23571	12.39857	363506500	11.11449
2007-01-23	12.24714	12.50143	12.21571	12.24286	301856100	10.97491
2007-01-24	12.38286	12.45000	12.29714	12.38571	231953400	11.10297
2007-01-25	12.44429	12.64286	12.29000	12.32143	226493400	11.04534
2007-01-26	12.44429	12.48143	12.14143	12.19714	246718500	10.93393
2007-01-29	12.32857	12.37857	12.21857	12.27714	225416100	11.00564
2007-01-30	12.34714	12.35571	12.17857	12.22143	144492600	10.95570
2007-01-31	12.12286	12.28571	12.05000	12.24714	214017300	10.97875
2007-02-01	12.31857	12.32429	12.10571	12.10571	166085500	10.85197

C. Data Pre-processing

Data collection is loosely controlled and more than often garbage values get added to the dataset. A high concentration of redundant information (noise) makes the data irrelevant and useless for further processing. Hence pre-processing of data is necessary to prepare the final dataset from given raw information. The method described in this paper converts the input data into a differentiated vector list. The function c{base} is used to address the combined vector list.

Data Frames

A data.frame() object in R has same dimensional properties as a matrix. But unlike matrices, data frames may contain both categorical and numeric data. It can be said that data frame is a list of variables with components as columns of a table. A list of variables with same number of rows and distinct row names of a class is defined as a data frame. The row names decide the number of rows, if no variables are involved.. The behavior of the data.frame() object can be changed by writing methods according to its class.

D. Data Processing: Training Data

The first step in data processing is to train the data. The ARIMA(p, d ,q) model is used to process data. Investors and analysts two methods to predict stocks namely auto regression and moving average. R provides

auto.arima () method to forecast the time series data according to ARIMA (p, d, q). The ARIMA model is a tool for technical analysis. It focuses on repeated parameter estimation and forecasting to find the right approximation model.

Auto Regression (AR)

A model that uses the dependent relationship between an observation and some number of lagged observations.Auto regression technique estimates the future values based on the previous values. The function of an autoregressive model is denoted by AR(p), where p represents the order of the model. AR(0), the simplest process, involves no dependence between terms, preceding or current. For a first order autoregressive model AR(1), the preceding term and a percentage of error contribute to the output. AR(2) model takes into account 2 preceding values and noise to predict the output.

Integrated(I)

The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.

Moving Average (MA)

A moving average is a technique to model datasets that vary according to single factor. It finds the future trends based on the previous values that do not follow a definitive pattern. The two commonly used moving average techniques are exponential moving average (EMA) and the simple moving average (SMA).

Order of ARIMA

The order of an ARIMA model is a class of statistical models for Analyzing for forecasting time series data.
Generally, represented asARIMA(p,d,q),
Where,
p = order of the autoregressive part.
d = degree of first differencing involved.
q = order of the moving average part.
Here if d=0, then the model becomes ARMA which is linear stationary model. The same stationary and in-variability conditions that are used for autoregressive and moving average models apply to this ARIMA (p,d,q) model. Selecting the appropriate values for p, d and q can be challenging. The auto.arima ( ) function in R will do it automatically.

Model Estimation for ARIMA

Model estimation for ARIMA can be achieved based on the pre-processed historical data.

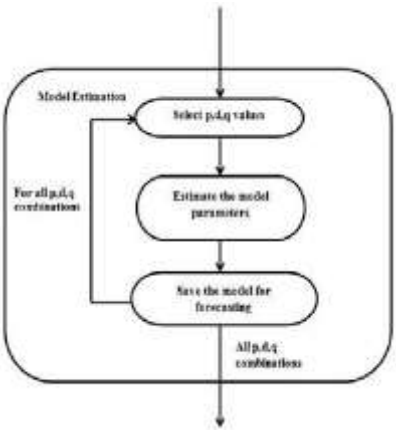

Figure2. pre-processed data

In ARIMA model, the identification is to be accomplished using auto co-relation function and partial auto co-relation function in order to identify p, d and q standards. For any realistic time sequence generally p, d and q values vary between 0 and 2, but model estimation is executed for all probable combinations of p, d and q values. The pictorial representation of these steps shown in figure2.

ARIMA() Function in R

The auto.arima() is a very useful function in R, but anything automated can be a little dangerous. This function examines over conceivable models within the edict limitations provided and returns the best ARIMA model. The value of d also has an effect on the prediction intervals i.e., the more complex the value of d, the more rapidly forecasting intervals surge in size. For d=0, the long-term prediction average deviance will go to the typical deviance of the historic data. It is usually not possible to tell merely from a time plot, what values of p and q are suitable for the specific kind of data. Sometimes it is conceivable to use the ACF plot and closely related PACF plot to govern the appropriate values.

Special Cases of ARIMA Model

Table 2. Special Cases of ARIMA Model

Special Cases	ARIMA Values
White noise	ARIMA(0,0,0)
Random walk	ARIMA(0,1,0) with no constant
Random walk with drift	ARIMA(0,1,0) with a constant
Auto regression	ARIMA(p,0,0)
Moving average	ARIMA(0,0,q)

E. Forecasting Results

Forecasting allows us to predict future values based upon the knowledge of current and historical stock data. The model specified here uses the forecast package for R for predicting future stock values. The forecast package contains tools for analyzingunivariate time series data using state space models and ARIMA modelling. The Arima () and auto. arima () functions used to model future stock prices are a part of the forecast package.

F. Plot Visualisation

Plot visualization involves representing the numerical data in graphical format. In the given methodology, line charts and histograms are used to represent the stock data. This is done using the plot () function provided in R. The addBBands () function adds two additional lines that make data interpretation easier. The x-axis represents the represents time period in terms of year/months and days while the y axis shows stock price values.

G. View and Analyze Results

Once after plotting the results in-terms of visualizations we can find out the correlations to get the short-term predictions. In the next section we provide some of the figures by which the investor can analyze and predict the future stock trends of a particular company at a specific time period. So the investors in the stock market can use this as assistance to sell/buy/hold a share.Figure3 shows the candlestcks graph of Apple price data ,figure4 shows chart series graph and in figure5 shows Chart series graph for AAPL data with bollinger bands.

Figure 3 candlesticks graph for APPLE data

Figure 4 chart series graph for APPLE company data

Figure 5 Chart series graph for AAPL data withbollinger bands

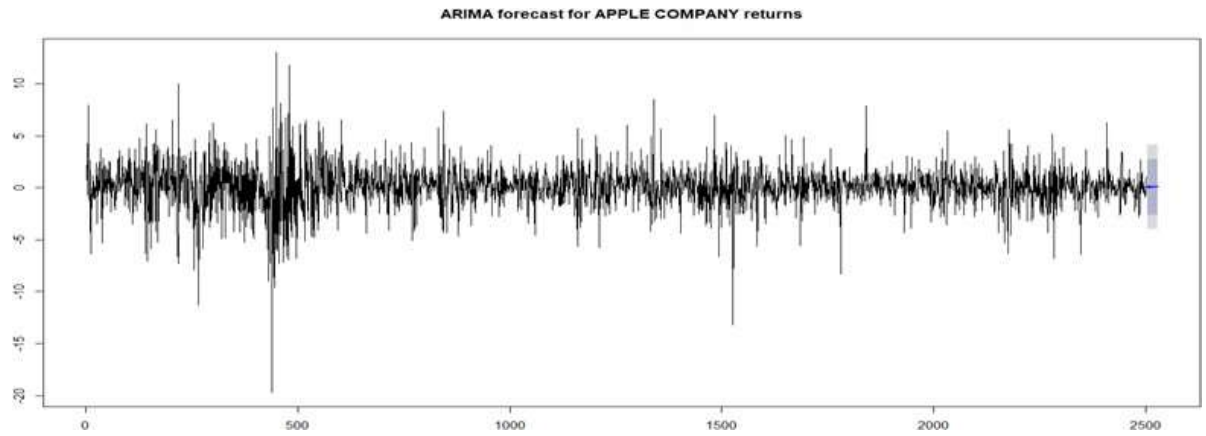

Figure6: ARIMA forecast for aapl returns

Above figure6 are the results that we obtain with a simple ARIMA(2,0,2) model.The deeply shaded region provide us the 99% confidence level and the lightly shaded region provide the 95% confidence level for the forecast.An intrinsic shortcoming of the ARIMA models, which is evident from the plot above, is the assumption of the mean reversal of the series. What this means is that after some time in future the forecasts would tend to the mean of the time series's historical values thus making it a poor model for long term predictions. Calculating the accuracy of the model.The lower the value of RMSE (Root Mean Square Error) better is the accuracy of the model. Here we find the RMSE value to be 0.7841242 which is quite low hence the model is found to be pretty accurate.

V. Conclusion

In this paper an attempt was made to forecast the stock market prices of the APPLE stock by developing a prediction model based on technical analysis of historical time series data and data mining techniques.This paper succesfully predicted the stock price indices for short-term period using an ARIMA model.The potential of the ARIMA model in finding future stock price indices which will enable stock brokers/investors to make pofitable investment is huge.The only drawback of this model as compared to its competitors is the tendency tio compute the mean of the historical data as forecast when it comes to long term prediction.Thus it is not advisable to use this model for long-term forecasting of stock price indices.

VI. Future Scope

The possibility of integrating this model with fundamental analysis can lead to better decision making when it comes to making decisions like buy/hold/sell a stock.Through a sentiment analysis performed by collecting public opinions from social media data and combining it with the ARIMA forecast better profitable investment decisions could be made.In this way we can improve better results for invest in stock market.

References

[1]. AS,Suresh.,"A Study on Fundamental and Technical Analysis".International Journal of Marketing Services and Management Research.Vol.2,No.5,May 2013.http:// indianr esearchjournals.com/pdf/IJMFSMR/2013/May/6.pdf
[2]. C.Boobalan "Technical Analysis in Select Stocks of Indian Comapanies." International Journal of Business and Administration Research Review, Vol.2 Issue 4, Jan-March 2014. http://ijbarr.com/downloads/2014/vol2-issue4/4.pdf
[3]. Ayodele A. Adebiyi, Aderemi O. Adewumi, Charles K. Ayo, "Stock price prediction using the ARIMA model", 16th IEEE International Conference on Computer Modelling and Simulation (UKSim), March 2014, pp. 106 -112.
[4]. Banerjee, D., "Forecasting of Indian stock market using time-series ARIMA model", 2nd IEEE International Conference on Business and Information Management (ICBIM), January 2014, pp. 131-135.
[5]. LinhaoZhang,"Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation" Department of Computer Science, the University of Texas at Austin. April 16,2013.http://apps.cs.utexa s.edu/tech_reports/reports/tr/TR-2124.pdf.
[6]. VishwanathR.H,LeenaS.V,SrikantaiahK.C,Shreekrishna Kumar K,DeepaShenoyP,VenugopalK.R,Patnaik L.M, "Approximation and Prediction of Stock Time-Series Data using Pattern Sequence" ELSEVIER. http://searchdl. org/ public/book_series/elsevierst/2/ICDMW18.pdf
[7]. Tao Xing, Yuan Sun, Qian Wang, Guo Yu, "The analysis and prediction of stock prices", IEEE International Conference on Granular Computing (GrC), December 2013, pp. 368–373.
[8]. MahanteshC.Angadi,AmonghP.Kulkarni "Time Series Data Analysis for Stock Market Prediction using Data Mining Techniques with R",International journal of Advanced Research in Computer Science,vol.6,No.6,july-august 2015.
[9]. Han, J., Kamber, M., Jian P., "Data mining concepts and techniques". San Francisco, CA: Morgan Kaufmann Publishers, 2011.

