A Fusion System for Forecasting Stock Prices

R.M.C.D.K. Rajasinghe¹, W.D.N.M. Weerapperuma², W.U.N.N. Wijesinghe³, K.K.K.P. Rathnayake⁴, L. Seneviratne⁵, B.H. Kasthuriarachchi⁶

Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

Abstract— Stock market is a network of buyers and sellers where large organizations sells stocks to investors. A stock represents a part of the company and the investor becomes a shareholder of the company once a stock is purchased. Shareholders are entitled to an agreed amount of dividend and the investor should be wise enough to choose profitable stocks. A price forecasting mechanism can help the investors in selecting profitable stocks. The Random Walk Hypothesis governs the Stock Prices as it changes constantly due to various factors and the Hybrid System Mechanism address these factors and analyses the stock prices to generate accurate price predictions.

Keywords – Stock Market, Random Walk Hypothesis, Forecasting Algorithms, Feature Extraction, Sentimental Analysis

I. INTRODUCTION

Large organizations issue stocks to the general public in order to fund their financial activities. Shareholders are entitled dividends for the invested amount, once the company generate a profit. Investing profitably is a tricky process where it requires an investor to be well aware of the company’s economical background as well as the performance of the company in order to determine how the company will perform in future. If a company is performing well the stock prices tends to go up. Investor should buy the stock before the price go up and also be wise enough to sell the stock in the ideal time before the price fall back down.

Fundamental factors such as the financial stability of the company and also the sentimental factors such as the public opinion of the company plays a huge role in keeping the market efficient and non-stationary. So it is very hard to determine how companies will perform and how the stock prices will vary in the future.

Stock market investments can be both direct and indirect. Direct investments are carried out either through exchanges or over-the-counter-markets. In stock exchanges companies can issue stocks as well as the investors and shareholders can trade their stocks. In trading stocks, shareholders will have to analyze the market and decide which stocks to keep and which stocks to sell in order to generate profits. Shareholders have to undertake the risk of losing the investment in determining the profitable stock correctly. Return on the investment will only be profitable if the performance of the company in the time period is profitable. So, evidently, investors shoulder a huge financial risk in becoming shareholders, where the credibility of companies are very unpredictable and uncertain due to the dynamically changing nature of the world economies. The trader should be wise and experienced enough to gain profits by trading the stocks at the ideal time.

Indirect Investments is a fairly new trend in investing where the investor is not directly involved in physically purchasing the stock. The investor or the bidder in Indirect Investing context places bets regarding the price variations in the future. This is also known as spread betting [1]. A spread betting company quotes two prices, the bid and the spread, and investors bet whether the price of the underlying stock will be lower than the bid or higher than the spread. The random nature of the prices can affect spread betters as well although they don’t own the underlying stock [2].

Taking all the above factors in to consideration one could arrive at a conclusion that investing in the share market either direct or indirect involves a lot of risk. The investor has to assume and accumulate the worth of undertaking the risk in order to earn profits. Currently there are many tools available that help the investors in investing more effectively and profitably but the main problem is most of the available tools focus only on a single factor that affects the stock prices. The Hybrid System is implemented to generate rich predictions that takes multiple factors that affect the stock prices in to consideration in generating the prediction.

II. RESEARCH METHODOLOGY

The research is carried out to develop a system that will forecast stock price variations with a good accuracy. The forecasting process is designed in a way that it addresses and analyses stock market in multiple standpoints that affects it. Fig. 1 depicts the basic solution outline suggest by the research. Generation of predictions are implemented under four main methodologies which include;

1. Data Mining
2. Statistical Analysis
3. Graphical Analysis
4. Sentimental Analysis

All the four methodologies will generate predictions that will be used in generating the final hybrid prediction from the Dynamic Weight Distributor which will allocate weights for each prediction based on their accuracies. The main inputs for the process are historical stock quotes and twitter feeds regarding financial activities of companies in the stock market. The analysis process will be done dynamically once the user defines the necessary data sources and time periods.
Stock Markets deal with multitude of data in a day and it is impossible to keep track of these data and identify patterns manually. Data mining can be used to extract the knowledge embedded within this data to provide a predictive analysis which will provide an insight to the investor that will invoke his judgment and experience into the equation.

Since stock market data is of time series data type it contains three main components apart from actual data, the trend, seasonal pattern and the irregular or remainder component. Loess data smoothing methodologies are used on data to further clarify the underlying patterns that are distorted due to seasonal and irregular elements. Once the data is decomposed using the seasonal-trend decomposition a seasonally adjusted set of data can be obtained using an additive model. Fig. 2 depicts the three components in the time series data decomposed using seasonal-trend decomposition. The seasonally adjusted data is used to generate predictions.

Naïve methods consider the most recent fluctuations in the data to determine price variation in the future using the seasonally adjusted data. The patterns in the recent data are projected to continue in the future. Fig. 3 depicts the output from the naïve forecast.

The time series data is further analyzed in the Statistical Analysis process using time series analysis methodologies. Mainly two methods will be used:

- ARIMA (Autoregressive Integration Moving Average)
- (G) ARCH (Autoregressive Conditional Heteroscedastic)

ARIMA aims to describe the autocorrelations in a dataset. A stationary time series is one whose properties do not depend on the time at which the series is observed. In ARIMA (p,d,q), p is the number of autoregressive terms, d is the number of non-seasonal differences, and q is the number of lagged forecast errors in the prediction equation. ARIMA models are defined for stationary time series and if the data is not stationary, differencing process is done until stationary time series is obtained. If the differencing process is done d times, then in ARIMA (p,d,q) model, d will become the order if differencing used.

In a multiple regression mode, the variable of interest for the forecasting is combination of predictors. In an auto-regressive model, forecast the variable of interest using a linear combination of past values of the variable. Term auto-regression indicates that it is a regression of the variable against itself. In ARIMA, p can be written as,

$$y_t = c + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \varepsilon_t,$$

(1)

Here c is a constant and \(\varepsilon_t\) is white noise. This is similar to multiple regression but with lagged values of \(y_t\) as predictors. This is referred to as an AR(p) model. Rather than using past values of the forecast variable in a regression method, a moving average model uses past forecast errors in a regression-like model. In ARIMA, q can be written as,

$$y_t = c + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

(2)

Where \(\varepsilon_t\) is white noise. This is referred to as an MA(q) model. So the ARIMA model would be, [3]

$$y'_t = c + \varphi_1 y'_{t-1} + \cdots + \varphi_p y'_{t-p} + \varphi_1 \varepsilon_{t-1} + \cdots + \varphi_q \varepsilon_{t-q} + \varepsilon_t,$$

(3)
The auto.arima() methodology is used in the implementation and it is proved to be very useful, but anything automated can be a little risky, and worth understanding the behavior of it. The auto regressive model will choose a model for the forecasting. In above equation, constant C has an important effect on the long term forecasts,

- If c=0 and d=0, forecasts will go to zero.
- If c=0 and d=1, forecasts will go to a non-zero constant.
- If c=0 and d=2, forecasts will follow a straight line.
- If c≠0 and d=0, forecasts will go to the mean of the data.
- If c≠0 and d=1, forecasts will follow a straight line.
- If c≠0 and d=2, forecasts will follow a quadratic trend.

The value of d also has an effect on the prediction intervals, the higher the value of d, the more rapidly the prediction intervals increase in size. The value of p is important if the data show cycles.

The GARCH model is specialized in characterizing a dataset. They are used whenever there is reason to believe that, at any point in a series, the error terms will have a characteristic size, or variance. Term p is the lag length of a GARCH(p,q) and it can be found by using three steps [4],

1. Estimate the best fitting AR(q) model Autoregressive (AR) model is a representation of a type of random process. Example economics,
   \[ y_t = a_0 + a_1 y_{t-1} + \cdots + a_q y_{t-q} + \epsilon_t = a_0 + \sum_{i=1}^{q} a_i y_{t-i} + \epsilon_t \]  
   \( (4) \)

2. Compute and plot the autocorrelations of \( \epsilon^2 \) by Autocorrelation, also known as serial correlation, is the cross-correlation of a signal with itself
   \[ \rho = \frac{\sum_{t=t+1}^{T} (\epsilon_t^2 - \bar{\epsilon}^2)(\epsilon_{t-1}^2 - \bar{\epsilon}^2)}{\sum_{t=1}^{T}(\epsilon_t^2 - \bar{\epsilon}^2)} \]  
   \( (5) \)

3. The asymptotic, that is for large samples, standard deviation of \( \rho(i) \) is \( 1/\sqrt{T} \). Individual values that are larger than this indicate GARCH errors. To estimate the total number of lags, Ljung-Box test is used until the value of these are less than 10% significant. The Ljung-Box Q-statistic follows distribution with n degrees of freedom if the squared residuals are uncorrelated. It is recommended to consider up to T/4 values of n.

The Graphical Analysis is a part of Technical Analysis and it can be used to identify stock patterns in past datasets. Feature selection is a preprocessing step that aims to select the most relevant subset of attribute by eliminating unrepresentative attributes from the dataset.

Stock data patterns are the best way to represents trends in a stock. Even someone with less knowledge in stock price variations can analyze chart patterns without much troubles. These patterns are further classified in to two categories. Those are reversal and continuation patterns. Each pattern which occurred in past may have a certain similarity to present. The goal behind this analysis is finding an appropriate relationship between the past and current patterns. This analysis is done by using three main sub-systems. Those are,

- Candlestick pattern analysis
- Algorithmic analysis for stock pattern recognition
- Artificial Neural Network based pattern recognition

Candlesticks is a graphing methodology which represents data in a more descriptive manner. A sample candlestick chart for Stocks of Apple Organization is depicted in Fig 5. A candlestick is based on four variables. They are open, high, low and closing prices of a stock. It’s further categorized into increasing or decreasing. If the closing price of a stock is higher than the opening one, it implies increasing candlestick. The decreasing is the vice versa of increasing. There are several types of standard patterns found in candlestick charts such as doji, gravestone doji, dragonfly doji, hammer patterns and long shadows. Based on these patterns candlestick pattern analysis system is working as supporting system, which is responsible for determining whether a stock trend will reverse after a special pattern such as doji occurs. This enables the decision making process of the dynamic wright distributor much simpler. Below are the steps of the process,

- Analysing past stock data and retrieve dataset containing all the doji occurrences
- Consider all the doji data points one by one and check data points of days following the pattern and before the pattern and determine whether the underneath trend reversed
- Get same calculations for all patterns and develop a constant value which is dedicated for a stock
- Whenever the relevant stock is used for the prediction, relevant calculated constant can be used for validate the generated prediction

The stock data considered as noisy and rapidly varying. In order to identify patterns using the algorithmic graphical
analysis, the first step is to filter out the noise. Sliding window mechanism is used for filtering. This technique is based on turning points that change the trend of a stock dataset. It’s done in two levels. First row stock data will be considered and fragmentation of the dataset into smaller parts is done. Then consider a single part and identify turning points by using local maximums and local minimums. In level one, for each segment only local minimums and maximums are taken and points which do not affect for the under laying trend are ignored [5].

After selecting the points in separate segments, merging process takes over, combine neighbouring segments one another, and generate a new dataset. In level two, the dataset from level one is used. This extract points which are more contribute for the under laying trend. This is done by using an algorithm. Below are basic step for eliminating intermediate points. P1, P2, P3 and P4 are consecutive data points in the stock dataset.

After accruing filtered dataset, pattern recognition algorithm works on it. The algorithms is able to extract head and shoulder patterns, inverse head and shoulder patterns, broadening patterns like flat-top broadening, flat-bottom broadening and semantic broadening and triangle patterns like, flat-top triangles, flat-bottom triangles and semantic triangles. In each of above mentioned patterns, formed using seven consecutive data points. Each seven data points for each pattern has its own characteristics. The algorithms uses those characteristics for the identification of a pattern [6].

Fig. 5 is an illustration of how the algorithm evaluates the data and identify a symmetrical triangle pattern. The algorithm will check consecutive seven filtered segments and check whether they form a symmetrical triangle. After the identification of all the patterns in a considered stock dataset, immediate next data points after the occurrence a certain pattern is analysed and price variation percentage is calculated. This value is generalized only for a certain stock.

Next method is the neural network and it is capable of finding stock price patterns as the algorithmic solution but in a more efficient manner. This solution is generated using a three-layered neural network. First layer is the input layer, which has seven neurons[7]. Each of these neurons accepts a data segment as the algorithmic solution. Middle layer contains two hundred neurons and output layer has three neurons, each dedicated for a specific pattern. The neural net follows feed forward design and it uses Sigmoid as the learning algorithm [8]. Dataset for learning is generated and taken from a data generation algorithm, which is capable of generating random data points that has the characteristics of each pattern. Fig. 6 illustrates a high-level arrangement of the neural network.

The steps taken for identifying a pattern from a dataset,

1. Filtering the data
2. Segmentation: Segmentation is important since the neural net will consider only seven data points for an instance
3. Inserting each segment to the neural net to identify patterns

Because the neural net is pre trended, no dynamic training involved. So it will reduce the over training problem.

Pseudo code:
Get all min points \{p1,p3,p5,p7\}
Get all max points \{p2,p4,p6\}
If \(h1 < \text{TRESHOLD\_HEIGHT}\) and \(h2 < \text{TRESHOLD\_HEIGHT}\) and \(p2 > p4\) and \(p4 > p6\) and \(p1 < p3\) and \(p3 < p5\) and \(p5 < p7\) then
The segment is valid pattern

Fig. 5: Point identification in a head and symmetrical triangle

Fig. 6: High-level design of neural network

Fig. 7: Patterns Identified By NN
Fig. 7 is a sample pattern identified by the neural net for apple stock, which shaped from 18th of Oct 2000 to 20th of Nov 2000. Prediction methodology is same as the one in algorithmic method [9].

Human Emotions play a huge role in stock market as the investors are often manipulated with new information and trends in the market. Researches has proved that the social opinion of a company is an indicator of the stability of the company and the confidence people have on the company. The impression the company reflects on its investors and the financial world can encourage or discourage the investors to invest. In Sentimental Analysis approach the reaction of the society toward an action of a company is inspected to determine the psychological effect it has on the investors to invest more or invest less in a company.

Extracting the public opinion is a tricky process as emotions and opinion is subjected to change from one person to another so for implementing this practically, Twitter website is used as it is the largest microblogging website in the world. Twitter data is obtained from twitter API based on a certain scenario or a company. The data is then processed and mined to extract the polarity and the emotion that is embedded within it. This is done using a data dictionary that classifies positive and negative words and world combinations. The obtained data will be mapped with the dictionary and the percentage of polarity is calculated to represent public acceptance or denial to the scenario. Fig. 8 depicts the outcome of the sentimental approach regarding the release of the new Apple IPhone.

The final step of the Hybrid Forecaster in to merge all the results and generate a single forecast for a given stock. This is the responsibility of the weight distributor to allocate appropriate weights to each prediction and generate a single prediction which the investor can rely on easily. The weight distributor is implemented in a way that it takes the accuracy of the predictions in to consideration when allocating the weights. Accuracies of different approaches may change from company to company as well as from season to season. These factors are taken in to consideration in designing the weight distributor.

III. RESULTS AND EVIDENCE

Naïve based method used in predictions used in Seasonally adjusted data can deliver predictions up to 2 years by projecting the last obtained actual value. The main problem with this methodology is that as the prediction length goes beyond 3 months, the accuracy of the prediction tend to decrease drastically. TABLE I define how the predictions of the Naïve method vary within the next 3 months.

<table>
<thead>
<tr>
<th>Month</th>
<th>Low Range</th>
<th>Mean Prediction</th>
<th>Higher Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>September</td>
<td>1909.13</td>
<td>1974.39</td>
<td>2039.64</td>
</tr>
<tr>
<td>October</td>
<td>1892.03</td>
<td>1987.31</td>
<td>2076.6</td>
</tr>
<tr>
<td>November</td>
<td>1874.62</td>
<td>1987.64</td>
<td>2100.66</td>
</tr>
</tbody>
</table>

Once the results are considered with the actual observations an idea about the accuracy of the results could be gained. The predictions and actual result variation is shown in Fig. 9.

Fig. 10 depicts how the stock prices of S&P 500 has varied with the ARIMA model. The predicted daily data for 10 days are compared here.
Fig. 11 depicts how the S&P 500 index has behaved with regard to the predictions from the GARCH model. The predicted daily data for 10 days are compared here.

![GARCH model prediction variance](image)

**Fig. 11 GARCH model prediction variance**

Chart pattern analysis only produce results when a certain pattern occurs in a dataset. Fig. 12 represents the results acquired from algorithmic pattern analysis on Apple dataset from 2000-Jan to 2014-Aug. The x-axis depicts number of patterns and y-axis depicts the prices.

![Pattern recognition](image)

**Fig. 12: Actual and predicted values from pattern recognition**

Since all these methods are inputted to a weight distributor methodology including the sentiment analysis, the final generated prediction will be enriched with various standpoints that would affect the stock prices in different proportions. The scalability of the application will not be a problem and the market for this kind of application is clearly evident.

**IV. CONCLUSION AND FUTURE WORK**

The research was based on the Random Walk Hypothesis which suggests the unpredictable nature of prices in financial market. The hypothesis was tested using four main modules in the research which included both purely quantitative and qualitative aspects that affect the stock price variations. The results obtained in the process suggests the possibility of the predicting the stock prices even though it is completely governed by the Random Walk Hypothesis. The results proves the prices can be predicted to some extent.

A mechanism can be very effective which can be used to map other irregular factors that can affect the stock market other than public opinion itself. The same kind of mechanism can be used to find out how stock market has reacted to political, geographical and natural phenomena’s. The Hurricane Katrina that happened in the mid of 2005 is a perfect example to this as it directly affected the world economies in the long run. The solution can be further developed to incorporate the irregular indirect factors along with the current features which mainly focus on the direct influences on the stock market.

**REFERENCES**


