INTRODUCTION

Prediction of the Stock Market has been a topic of interest since the inception of NYSE in 1792. The reason, which is evident by the fact that it is an open market (not exactly a physical entity) where buyers and sellers come to trade stocks, which represent ownership claims on businesses. These may include securities listed on a public stock exchange, as well as stock that is only traded privately. Stock exchanges list shares of common equity as well as other security types, e.g. corporate bonds and convertible bonds.

The prices of stocks move on supply and demand. For a stock that is higher in demand and low in price and abundant in quantity would be soon sized up and then later sold for higher prices. Similarly, a higher priced stock may face lower demand for trade than is usual and thus incur a loss. These motivations – the demand and supply – depend not only on the performance and a promising future of a company, but also on the market sentiment. Even though the company might be gaining huge profits, it might run into bad publicity and lose the value of its stocks.

The variables in this world are many. Having said so, many people attempt to predict the future of a business by buying its stock, hoping for greater returns in the future. Successfully predicting the market sentiment along with financial success is not a job for an ordinary man, yet the stock market remains a popular platform for investing one's money, as the profits can be extraordinary. This however, usually, requires careful analysis of the market. Therefore, prediction of the stock market has been stirring the minds of investors since day one.

There are two prices that are critical for any investor to know: the current price of the investment they own, or plans to own, and its future selling price. Investors are constantly reviewing past price history and using it to influence their investments decisions. A common way to invest is to look at the price history of a stock and look at its "momentum". A stock that has been increasing for some time will continually grow before its downfall. Similarly, a stock that is low and continues to fall, will not rise. Another common approach- one taken by experienced investors, is more psychological and abides by the view that the market evens out over time. Therefore, the stocks that are high for a long time will lose value over the years and the low stock will gain value over a period in such a way that the loss and gain would balance out. Another possible method is to take to the view that past returns do not matter. According to this Random Walk theory, the valuation of an investment depends only on the current price and estimated volatility of the stock.

A popular method, one taken by value investors, look to buy stock cheaply-one that has been underpriced by an inefficient market- and expect to be rewarded later.

The approach this study takes, assumes that stock prices are dependent on past prices and on the current price and sentiment of the market. The reason for doing so is simple. Often, investments are made by either analyzing past trends or by analyzing the sentiment. Both methods (The Random Walk and Temporal Analysis) limit themselves by not considering the value of the other, as both have been seen to work. By combining both these methods, this approach takes leverage of both sides of the coin to form an enhanced predicting strategy.

The rest of the paper is organized as follows, Literature review – where analysis of current research done in the field is done. Then in Methodology – where the approach and theory behind the models and tools is explained. Then finally in Implementation – the practical application of the approach is described, followed by the results and finally conclusion of this study.

LITERATURE REVIEW

For this study, analysis of the current literature has been done. Most of them produced satisfactory results. However, they had their own assumptions and limitations.

Firstly, to ascertain the validity of using data mining approaches, the paper, 'Does the use of technical & fundamental analysis improve stock choice? A data mining approach applied to the Australian stock market' by Hargreaves et al., concludes its hypothesis as true by producing higher returns with various models (decision trees, neural networks) than the Australian Ordinary Index (AOI) during a 20-day trading period. The only limitations were a limited small data set where missing data was imputed using the median value and their selection criteria of the stock, i.e. the sector (Industrial) to which their strategies were limited.

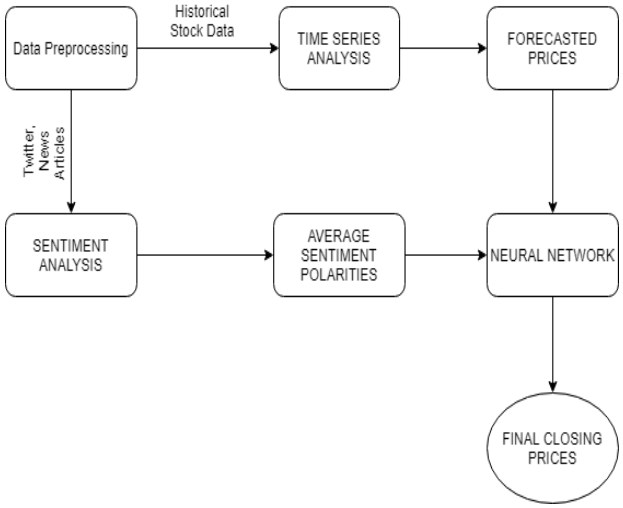
According to 'Price Trend Prediction of Stock Market using Outlier Data Mining Algorithm' by Zhao et al., stock prices have no underlying pattern. The trading volume therefore is also random, and hence if there are anomalies (due to insider trading and market manipulation) in the distribution of trading volume, market becomes inefficient, thus making long term prediction possible. By using tick-by-tick data instead of time series data, the researchers have calculated these anomalies and plotted them to produce an upward trend. However, any cyclical, seasonal patterns or even the general trend that is clearly evident in a time series plot of the stock price. has not been taken into account. Another paper, 'Short-term prediction for opening price of stock market based on self-adapting variant PSO-Elman neural network', applies a recurring neural network that feeds output from the hidden layer back to the input of the hidden layer via delay and storage. This makes the network sensitive to historical data. When iterating, the algorithm narrows the searching space of species, while the variant operation can jump out of the optimal position it had searched before, and search in a bigger space, increasing the chances that the algorithm finds out an even better value for optimization of the neural network. However, the drawbacks of this are that this network takes only a few recent values before it. There is also only one hidden layer in this network, when recent developments in deep learning prescribe larger neural networks for better performance owing to the amount of large datasets available nowadays. Furthermore, it works under the assumption that there are no major fluctuations in the time series, which is highly unlikely as stock prices tend to have major fluctuations. In 'Stock market prediction with multiple classifiers' by Rasheed et al., using a consistent voting ensemble method classifier combining neural networks, k-nearest neighbor and decision trees, to achieve the best average error of 34.64% versus the best average error of only ANN (out of all the three) of 39.91%. The ANN used however, again has only one hidden layer, i.e. it is shallow. Another approach has been to use Sentiment Analysis of Social Media to forecast stock prices. In 'Using social media mining technology to assist in price prediction of stock market' by Wang et al., the researchers have analyzed various Chinese websites and social media platforms for the forecasting algorithm combined with a Support Vector Machine (SVM). The results, however, of SVM and the combined model are roughly similar. This indicates that a better learning algorithm than SVM could be used to combine with Sentiment Analysis.

The authors propose an approach where, both temporal analysis and sentiment analysis are combined in a novel manner, by raising or sinking the forecasted prices obtained from temporal analysis with the aid of sentiment polarities, obtained from sentiment analysis.

METHODOLOGY

The stock data used in this study are historical daily stock prices obtained from Yahoo Finance. The raw dataset comprises of 7 elements: date, open price, high, low, close price, adjusted close, and volume. Of these only date and close price are of importance. The prediction of stock price in this research is represented by the closing price of a particular day. Closing price is chosen because it reflects the overall gain or loss in a trading day, and also helps in identifying the trend of the stock, i.e. is it trending upwards and is therefore profitable to invest in? Therefore, the time series thus formed is the closing price of the stock in a daily interval.

For this study, temporal analysis is done using autoregressive integrated moving average model (ARIMA). The autoregressive (AR) component of the ARIMA indicates that the evolving variable of interest is regressed on its own lagged values (i.e. prior values). The moving average (MA) part indicates that the regression error is a linear combination of its previous errors. The I (Integrated) part signifies that the data values have been replaced with the difference between their values and their previous values (This differencing process may be done several times). The purpose of all these features is to fit the model as precisely as possible.



ARIMA models can be seasonal and non-seasonal.

For temporal analysis, the only features required from the dataset is the Date and Close price. All other features are therefore removed. Then, the dataset finally comprises of historical stock prices indexed with their respective dates.

But before the prediction, it is of paramount importance to learn which timeframe for training the model is the most effective. It is generally believed the model should be trained within a time frame that is closest and shortest to that of the forecast period. However, we used K-fold cross validation on the dataset, as is applicable in a time series. In this method, with 10 folds, the time series is fragmented into 10 sets of miniature time series, each of which approximately spans the same duration of time. Therefore, there is a collection of test sets, each consisting of a single observation. The corresponding training set consists of observations that occurred before the observation of the test set. This way, no future observations can be used for the forecast. The results of the accuracy of these forecasts of each test set, depicts the best timeframe (1 week/ 1 month/ 3 months/ 6 months and so on) to train the model on. It is also possible that two or more timeframes result in a similar accuracy, in which case either can be used.

After establishing best training period, the ARIMA model is trained on it, analyzing the various seasonal and cyclic trends and forecasts the closing prices corresponding to a given timeframe (1 week, or 2 weeks, or 1 month and so on).

Along with these forecasted prices, we need the corresponding average sentiment polarity. The average sentiment polarity is calculated by averaging all the sentiment polarities of all the news articles, tweets, etc., of the selected company, of the day before, as an indicator of public sentiment about the company.

Sentiment Analysis is achieved by segmenting a sentence into tokens. Then from these tokens, common words like 'to', 'is', 'are' that are only used to join sentences are removed. The remaining words (like 'Profit', 'Achievement', 'Failure', 'Bankrupt') are assessed based on their preloaded weights. Certain words have higher weights than others. Compounding all these weights results in the sentiment of the sentence. This process is repeated for a paragraph, or an entire text consisting of several paragraphs.

This is done for every article for a day, and then the mean is taken to result in final average sentiment for that date. Then the average sentiment polarity and the forecasted price are taken in as input features to trained deep neural network, which predicts the final forecast prices.

IMPLEMENTATION

The tools used for implementation is the Pyramid Arima (Python equivalent of R's AutoArima), to analyze the various parameters of an ARIMA model (p, d, q, P, D, Q) given a certain frequency, and return with the best possible set of parameters for a given time series dataset. The Pyramid Arima selects the best model according to Akaike information criterion (AIC) value. The AIC is a relative estimator of the quality of statistical models for a given set of data. In other words, given several models for the same dataset, the AIC estimates the quality of each model relative to the each of the other models. The set of parameters that produce the least AIC value is regarded as the best. However, some parameters must be manually defined, like the frequency (m). Frequency is the number of times observations are taken in a seasonal cycle. Typically, m corresponds to some recurrent periodicity such as: 7- daily, 12-montly, 52-weekly. Usually setting frequency requires analysis of the time series under study and requires technical analysis into its seasonal patterns. Since our dataset corresponds to a daily time period distribution, we set frequency as 7, which is to be interpreted as a weekly observation in each seasonal cycle.

For feature extraction, the stock data obtained must be reduced to a time series. After it has been reduced to a time series, with the index of the data frame as the date, and its close price as the variable that is changing with time, the time series is divided into two sets, i.e., Train, Test. The Train set is for training the ARIMA model, and the Test set is for finally testing the model's performance.

It is important to note that since each time series is different, k-fold cross validation is a vital step to help determine the size of training and test sets. The training set then is passed into the auto-arima function to prescribe the best model that fits the dataset.

The dataset comprises of stock prices from 2017-09-11 to 2018-09-07. The total day span is of 362 days. For 10-fold cross validation, the training sets are selected with an interval of 36 days. The last set comprises of 38 days. In total, we have 9 training sets. Correspondingly, there are 9 test sets, each of which exceed their corresponding test sets by 1 observation period.

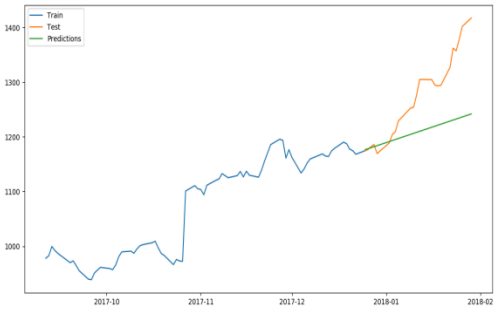
Results are calculated using the root mean squared error (RMSE) values of predicted values against the actual values. The RMSE is used because the simple difference between the forecast and actual values maybe either positive or negative.

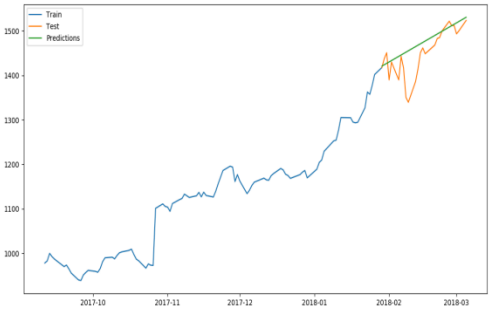
For each of the test sets, the RMSE is as follows:
[64.65638174324005, 24.1280649595207,
86.16796603788796, 42.642049185654336,
127.10919841296234, 100.02248714334554,
26.856386568675617, 45.21713092778753,
27.54472147346794]

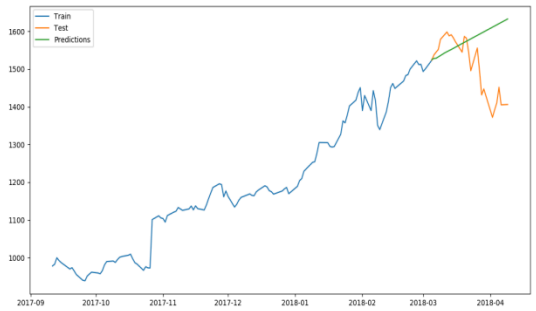
Finally, the mean error of all these errors is:

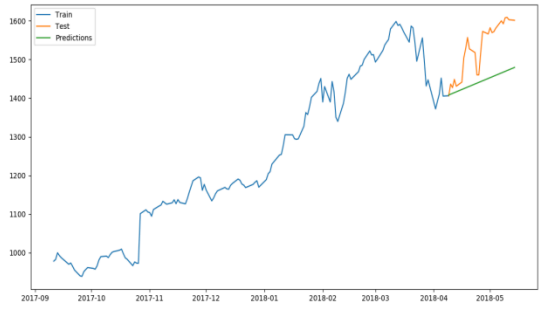
60.482709605838004

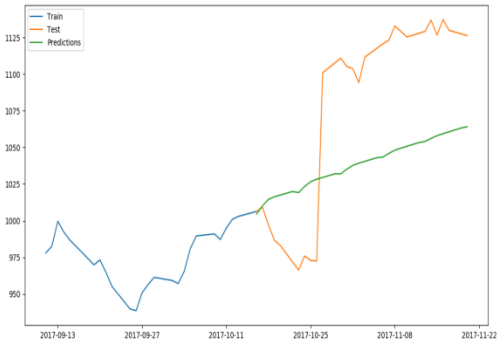
It implies that the 2nd test set, the 7th and the 9th have the best predictions. For future predictions, each can be checked to test the best performance.

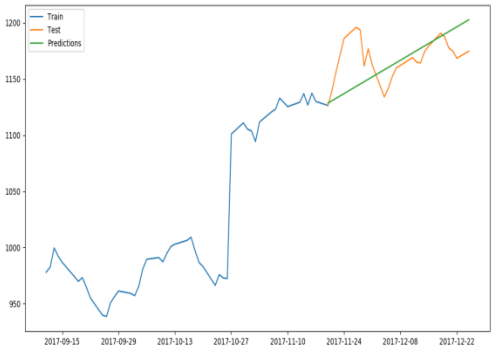


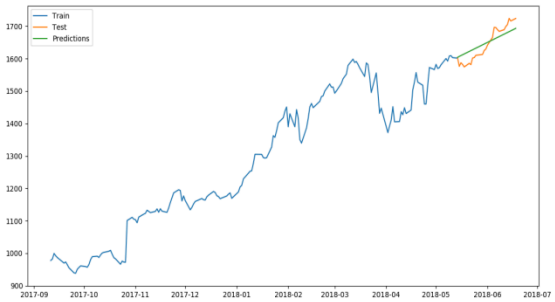


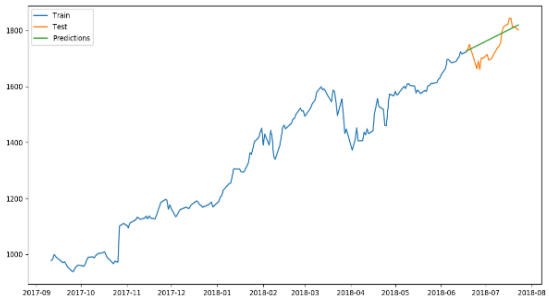


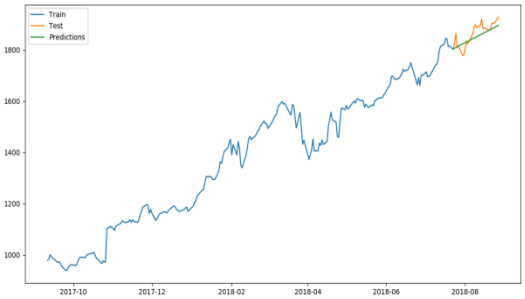








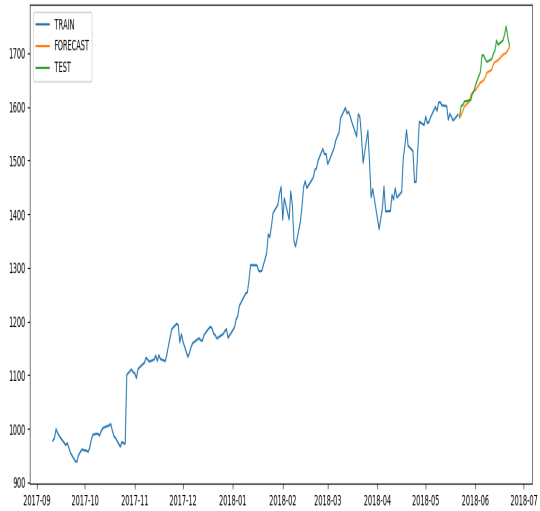




As can be seen from the graphs, the 2nd, 7th and 9th test set have the best prediction values, and obviously the lowest errors.

In this experiment, the chosen set sizes are 252-day ratio to 30-day ratio as training data versus testing data. According to Pyramid's AutoArima function, the following model fits best.

(ARIMA: order= (2, 0, 0) seasonal order= (2, 1, 2, 7); AIC=2167.196, BIC=2195.239, Fit time=6.229 seconds)



RMSE = 24.126771509548018

The prediction is nearly accurate, save for a couple of peaks and just by simple time series analysis. However, to make it account for the peaks as well, the predicted time series thus obtained is combined with the sentiment analyzer to output the final prediction values. This concludes the time series analysis.

According to the time series analysis, the best time frames for predicting were then used individually, to predict the time frame from '2019-01-01' to '2019-01-07'. The time frame, between 2018-10-20 to 2018-12-31, was selected for time series analysis and produced an RMSE of 58.08. This produced the forecast values for 2019-01-01 to 2019-01-07.

Data for sentiment analysis is extracted from various financial news websites like Nasdaq for the period 2019-01-01 to 2019-01-07. For sentiment analysis, the TextBlob library is used to extract sentiment from the gathered news articles, tweets, etc. for each day in the respective forecasting time frame. Then the sentiment for each day is calculated by averaging the sentiment values.

The average sentiment polarity along with the ARIMA predicted prices are fed into a neural network of 11 hidden layers and 2 outer layers, pretrained on the Amazon stock data, along with sentiment, trained. Each hidden layer and the first layer consist of 128 nodes, while the last layer contains a single node. For every layer, the activation function used is Rectified Linear Unit (ReLU), except for the last layer, where the activation is a linear function. This is done using the Keras and TensorFlow packages.

The NN predicts then final closing price.

Average Sentiment Polarity:

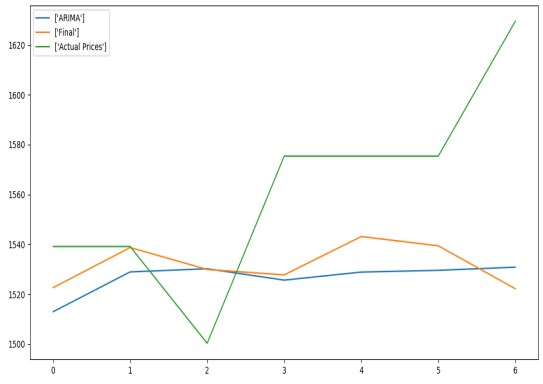
Date	Sentiment
01/01/2019	0.16953
02/01/2019	0.172247
03/01/2019	0.122231
04/01/2019	0.058794
05/01/2019	0.2
06/01/2019	0.17233
07/01/2019	0.026804

RESULTS AND DISCUSSION

Time series Analysis Results:

Training Period for ARIMA	RMSE	AIC of ARIMA MODEL
2018-10-20 – 2018-12-31	58.08	751.73
2018-05-29 – 2018-12-31	81.26	2102.83
2018-01-05 – 2018-12-31	75.78	3416.04

The 2018-10-20 – 2018-12-31 period result with RMSE 58.08 was selected as an input feature, along with the sentiment results of the testing time frame.



The final MAE was reduced to 38.54409086. from 58.08.

In order to validate the hypothesis, the time lag between the ten fold cross validation sets, and the actual test set is taken to be approximately 3 months. The tenfold cross validation is also considered for the prediction of 36 days and not 10 days as has been done. Another limitation is the failure of sentiment analysis. The sentiment results are poor. The reasons are a limited variety of financial news articles about the particular stock being analyzed and inability of existing natural language processing libraries (like TextBlob) to process financial sentiment accurately.

If the time series analysis is done on a recent time series relative to the time frame that is being predicted, and the corresponding sentiment is calculated accurately, the results can even be extraordinary.

Overall, given the limitations, the combined model of time series analysis using ARIMA model and sentiment analysis has performed satisfactorily.