An Comparative Study of Machine Learning Models for Stock Market Rate Prediction

G.V.Soni Meera[1], Assistant Professor, ECE, Marthandam College of Engineering & Technology, E-mail: soni.rosh@gmail.com

M Joselin Kavitha [2], Assistant Professor, ECE, Marthandam College of Engineering & Technology, E-mail: drjoselinkavitha@gmail.com

Dr.R.Isaac Sajan[3], Professor, ECE, Ponjesly College of Engineering, Nagercoil, E-mail: isaacsajan@gmail.com

Abstract---Predicting the direction of movement of the stock market index is important for the development of effective market trading strategies. It usually affects a financial trader's decision to buy or sell a stock. Closing price is one of the important factors in effective stock trading. Successful prediction of closing stock prices may promise attractive benefits for investors. Machine learning techniques have potential capability to process the historical stock trends and predict near accurate closing prices. This study compares three diverse machine learning models - ARIMA time series forecasting model, Support Vector Regression and LSTM Neural Network in terms of complexity of analysis, predictive accuracy for closing prices and customization.

Keywords---Machine Learning, Stock, Prediction, ARIMA, Support Vector Regression, LSTM Neural Network.

1. Introduction

Stock market prediction is undoubtedly one of the most challenging issues for traders and researchers. The opportunities of high profit offered by the stock market have always attracted a large number of investors. Researchers consider stock market prediction as a challenging task due to the difficulty in capturing the nonlinear and non-stationary variation in data [1]. The application of machine learning techniques for stock market prediction is a well established area.

Short-term trading refers to the trading strategies in the stock market or futures market in which the time duration between entry and exit is within a range of few days to a few weeks. Motivation of this study is that Short-term trading can be risky and unpredictable due to the volatile nature of the stock market. Within the time frame of a day and a week, many factors can have a major effect on a stock's price. Company news, reports, and consumer's attitudes can all have a positive or negative effect on the stock going up or down. Most investors invest in stocks that yield in long term because of the risk involved in short-term trading. Although the short-term trading could be risky, with relevant historical data and powerful machine learning techniques, it is possible to predict the variation in stock market. The closing price for the short term (10 to 15 days ahead) could be predicted to help the users choose a better stock to trade using three different machine learning models and analyzing when a particular model would be appropriate. Three diverse machine learning models have been compared: Autoregression Integrated Moving Average (ARIMA) [1], Support Vector Regression (SVR) [2] and Long Short Term memory (LSTM) Neural Network [3].

ARIMA, a statistical machine learning model is majorly used in time series forecasting. Methods for analyzing changing patterns of stock prices have always been based on fixed time series. Considering that these methods have ignored some crucial factors in stock prices, ARIMA model has been used to predict stock prices. Support Vector Regression (SVR) is considered to be one of the most suitable regression based machine learning algorithms available for time series prediction.

Finally, using LSTM neural network which is the most popular deep learning model for sequential analysis and analysis for time related data, prediction of the closing price is done and the comparison of these three models is conducted.

The rest of the paper is organized as follows. Section II focuses on the related work. Section III presents the methodology adopted and experimental setup. Section IV presents results and discussion. Section V depicts the conclusion and future work.

2. Related Work

Stock market is a highly volatile and changing environment and hence difficult to predict its future trends and movements with high accuracy. In the past, stock market forecasting was a question of intuition and feeling of traders. As the market and trading volumes grew, more sophisticated tools and models are used to perform more accurate and robust predictions.

Prediction of stock price using time series analysis has been done in [1],[5]. Results obtained revealed that the ARIMA model has a strong potential for short-term prediction and can compete favorably with existing techniques. Market performance was analyzed for Karachi Stock exchange by M Usmani et al. [6] where prediction model uses different attributes as an input and predicts market as Positive or Negative. This is done using ARIMA, Radial basis function and Multi layer Perceptron. Using Holt Winters, neural Network and ARIMA, opening price of Nifty50 was predicted along with sentiment analysis by S. Tiwari et al. [7]. It was found that feed forward network is the most efficient. Nearest Neighbor and Multilayer perceptions are compared for the financial analysis of time series data R.K. Dase et al. [8]. A. Greaves et al. [9] analyzed the Bitcoin to predict the price of Bitcoin using support vector machines and artificial neural networks (ANN) reporting price direction accuracy of 55% with a regular ANN.

R.P. Schumaker et al. [10] studied the correlation between financial news and stock prices. Long Short-Term Memory (LSTM) is a variant of RNN and proves to be demonstrating good performances in time series learning as LSTM can maintain contextual information as well as temporal behaviors of events. A Abraham et al. [11] have researched on application of hybridized soft computing techniques for automated stock market forecasting and trend analysis. They make use of a neural network for one day ahead stock forecasting and a neuro-fuzzy system for analyzing the trend of the predicted stock values. Implementation of Recurrent Neural Network and LSTM is done on GPU and CPU for analyzing the direction of Bitcoin price by S. McNally et al. [12] and was found that LSTM in combination with GPU gives the best classification accuracy.

3. Methodology

This section elaborates the Machine Learning models which are being used to forward the study from data towards inferences.

Theoretical Framework

Autoregressive Integrated Moving Average

Time series modelling has long been used to make forecast in different industries with a variety of statistical models currently available. The general model includes autoregressive as well as moving average parameters, and explicitly includes differencing in the formulation of the model. Specifically, the three types of parameters in the model are: the autoregressive parameters (p), the number of differencing passes (d), and moving average parameters
(q) [1]. The ARIMA model is provided with time series data and the model is fitted according to this data to understand the nature of data or to predict the future values (forecasting).

ARIMA models are the most common class of models for forecasting a time series which is first made stationaryby differencing. The prediction equation is a linear equation that refers to past values of original time series and past values of the errors. The trend can be correctly predicted by the model when both the model and independent variables are correctly selected [1],[2],[5],[6].

ARIMA is also known as Box-Jenkins approach. Box and Jenkins claimed that non-stationary data can be made stationary by differencing the series, Yt. The general model for Yt is written as,

$$Y_t - \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \theta_q \epsilon_{t-q}$$

Where Yt is the differenced time series value ϕ and θ are unknown parameters and ϵ are independent identically distributed error terms with zero mean. Yt is expressed in terms of its past values and the current and past values of error terms.

The ARIMA model combines three basic methods:

1. Auto Regression (AR)-In auto-regression, the values of a given time series data are regressed on their own lagged value, which is indicated by the p value in the ARIMA model.
2. Differencing (I-for Integrated)-Here differencing is done to convert non stationary graph into stationary graph by removing the trend. This is indicated by the d value in the ARIMA model. If d = 1, it looks at the difference between two-time series entries, if d = 2 it looks at the differences of the differences obtained at d =1, and so forth.
3. Moving Average (MA)-q value is used to represent the number of lagged values in the error part q value represents the moving average nature of the ARIMA model.

Support Vector Regression

SVR is considered to be one of the most suitable algorithms available for time series prediction. This supervised algorithm can be used in both, regression and classification. The SVR involves plotting of data as point in the space of n- dimensions. These dimensions are features that are plotted on particular coordinates. SVR algorithm draws a boundary over the data set called the hyper-plane [6],[7],[9].

Long Short term Memory

LSTM is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Such networks have a short term memory capability and the hypothesis to explore here is that this feature can present gains in terms of results when compared to other traditional approaches in Machine Learning field.

Unlike standard feed forward neural networks, LSTM has feedback connections. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over random time intervals and the gates in the cell regulate the flow of information in the cell. LSTM is chosen over the other Neural networks such as Recurrent Neural Network (RNN) as LSTM is best suited for the time series analysis [8],[11],[13].

Experimental Setup

Data Pre-processing and Visualization

The dataset considered for stock price prediction is taken from New York Stock Exchange (NYSE) [19]. It represents various fortune 500 companies in NSE. It is the performance indicator or benchmark of all listed companies of NSE. From the website of NYSE the stock prices for various firms are obtained from Jan 2013 to June 2018. The first step is the data pre-processing. Machine learning needs two things to work with -data and models. Acquiring the data from various sources with right features is the major task.

In Data pre-processing raw data is transformed into clean data sets. Null values, inaccurate, noisy and inconsistent data is removed. Descriptive Statistics - maximum, minimum, standard deviation, mean is computed on features for removal of outliers.

Data Visualization with Seaborn plotting is used for identifying correlated features to further aid in the selection of effective features for predictions.

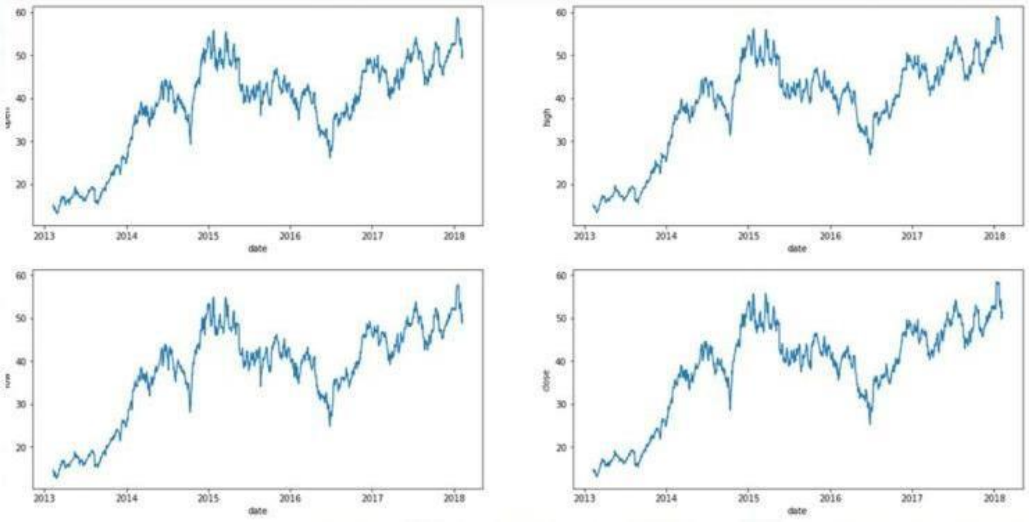


Figure 1: Variation in Closing Price Prediction with different Features

Figure 1 shows the variation of closing price with time when different combinations of features are used.

Model Building and Training ARIMA

The classical approach for fitting an ARIMA model is to follow the Box-Jenkins Methodology. Before running the ARIMA model, the time series is to be made stationary. Firstly the rolling mean for a window = 12 for a period of 12 months is calculated and a cumulative sum of the rolling mean is calculated. The rolling mean and rolling standard deviation is compared against the original data by plotting. Dicky Fuller Test [1] is conducted to check if the Time Series is Stationary. Differencing is done until stationarity is obtained. The value of p, the lag value and critical values of 1% 5% and 10% is calculated.

Log of Differentiated shift in mean is plotted to check for trends and seasonality. Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) plotting is done. p" and q" values as explained in section 3.1 are assigned according to the ACF and PACF plots obtained. When the user gives a forecast period for estimation, a graph comparison of predicted and original closing price is shown along with the confidence interval [1].

Algorithm:

```
1. Plot (dataset)
   while (graph is non-stationary = True) (smoothen graph to make it stationary)
2. ACF/PACF(stationary Graph)
3. Estimate all possible model parameters
4. Calculate Akaike Information Criterion values of all model parameters
5. Plot (with model parameters)
   if( no lag)
      (Forecast Dataset with these parameters)
   else
      choose other estimated model parameters and repeat step 3
```

SVR

The proper output from SVR model mostly depends on selecting appropriate kernel function and its parameters setting. Here, Radial Basis Kernel function is used because itis faster and also able to produce good output results.

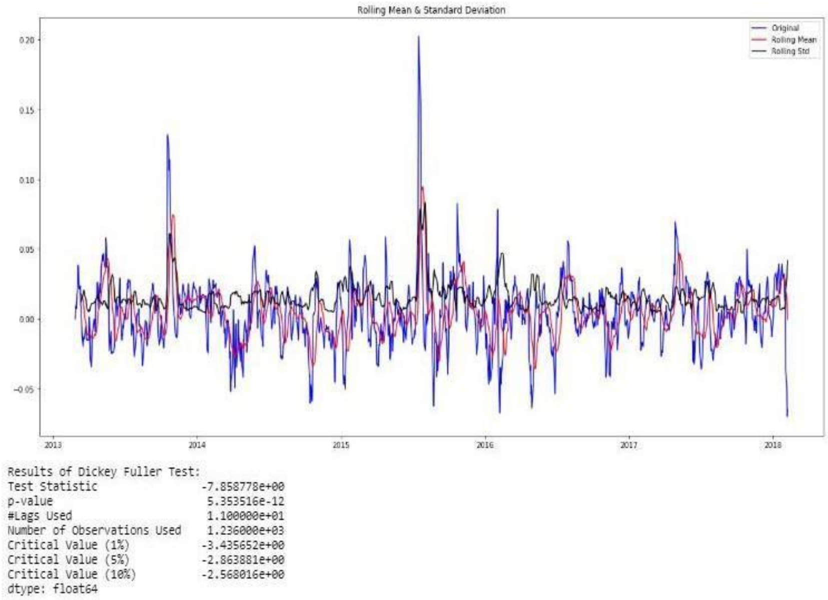
LSTM

LSTMs were developed to deal with the exploding and vanishing gradient problems that can be encountered when training traditional RNNs[4]. Relative insensitivity to gap length is an advantage of

LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications. In this study, 3 hidden layers have been used and the activation function used for the LSTM network is the rectified Linear unit which is best suited and mitigates the effect of vanishing/exploding gradient issue.

4. Results and Discussion

Data was partitioned for training and testing with the same partitions and similar inputs used for all the three machine learning models used for this study. Closing prices of NSE data which included daily stocks of companies such as GOOGLE, UBER etc was predicted.



```
Results of Dickey Fuller Test:
Test Statistic                -7.858778e+00
p-value                        5.353516e-12
#Lags Used                     1.100000e+01
Number of Observations Used    1.236000e+03
Critical Value (1%)           -3.435652e+00
Critical Value (5%)           -2.863881e+00
Critical Value (10%)          -2.568016e+00
dtype: float64
```

Figure 2: Dicky Fuller Test in ARIMA Model

Figure 2 depicts the Dicky Fuller test performed during model building and training in ARIMA Model. In Figure 2, the black line represents rolling Standard Deviation plot, blue line represents original data plot and Red line represents rolling mean plot. The plotting is done to compute the parameters required to input the ARIMA model and the critical values computed.

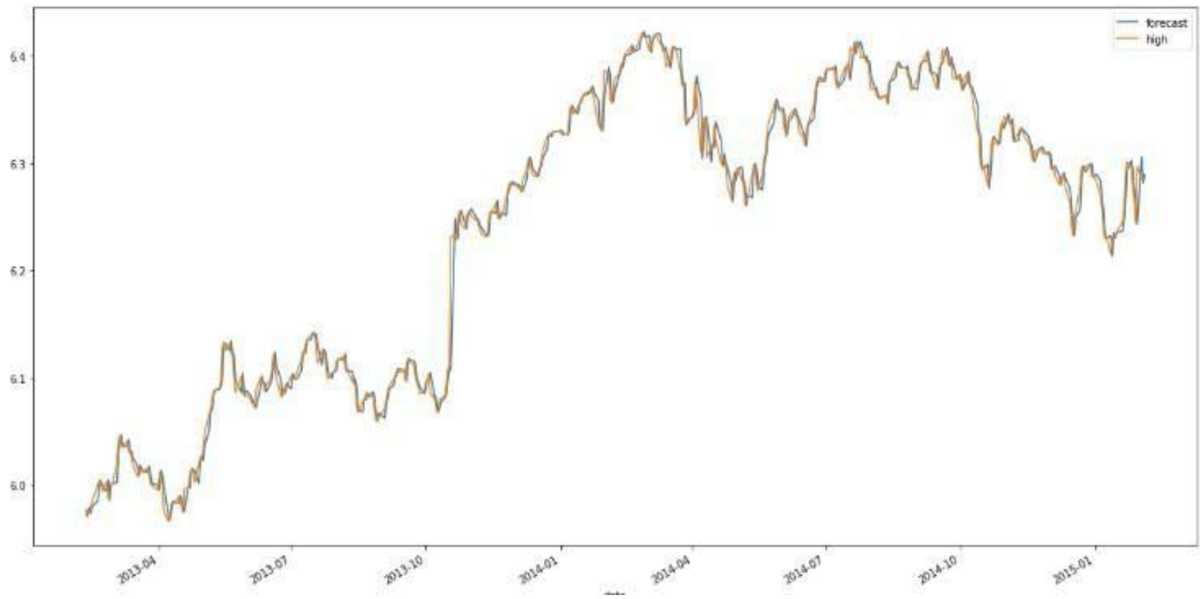
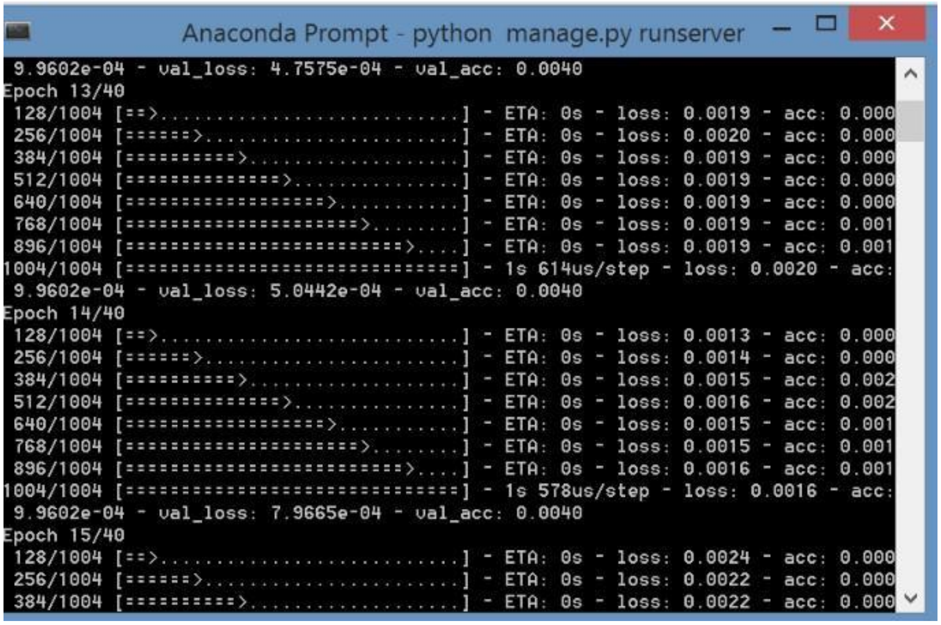


Figure 3: Predicted Change in the Closing Price with Time

Figure 3 is the result obtained from the ARIMA model which is the plotting of predicted closing price to the original closing price against the time. In Figure 3, the blue line represents forecasted closing price and the orange line represents original data plot.



```
 9.9602e-04 - val_loss: 4.7575e-04 - val_acc: 0.0040
Epoch 13/40
 128/1004 [==>...........................] - ETA: 0s - loss: 0.0019 - acc: 0.000
 256/1004 [======>.......................] - ETA: 0s - loss: 0.0020 - acc: 0.000
 384/1004 [==========>...................] - ETA: 0s - loss: 0.0019 - acc: 0.000
 512/1004 [==============>...............] - ETA: 0s - loss: 0.0019 - acc: 0.000
 640/1004 [==================>...........] - ETA: 0s - loss: 0.0019 - acc: 0.000
 768/1004 [=====================>........] - ETA: 0s - loss: 0.0019 - acc: 0.001
 896/1004 [=========================>....] - ETA: 0s - loss: 0.0019 - acc: 0.001
1004/1004 [==============================] - 1s 614us/step - loss: 0.0020 - acc:
 9.9602e-04 - val_loss: 5.0442e-04 - val_acc: 0.0040
Epoch 14/40
 128/1004 [==>...........................] - ETA: 0s - loss: 0.0013 - acc: 0.000
 256/1004 [======>.......................] - ETA: 0s - loss: 0.0014 - acc: 0.000
 384/1004 [==========>...................] - ETA: 0s - loss: 0.0015 - acc: 0.002
 512/1004 [==============>...............] - ETA: 0s - loss: 0.0016 - acc: 0.002
 640/1004 [==================>...........] - ETA: 0s - loss: 0.0015 - acc: 0.001
 768/1004 [=====================>........] - ETA: 0s - loss: 0.0015 - acc: 0.001
 896/1004 [=========================>....] - ETA: 0s - loss: 0.0016 - acc: 0.001
1004/1004 [==============================] - 1s 578us/step - loss: 0.0016 - acc:
 9.9602e-04 - val_loss: 7.9665e-04 - val_acc: 0.0040
Epoch 15/40
 128/1004 [==>...........................] - ETA: 0s - loss: 0.0024 - acc: 0.000
 256/1004 [======>.......................] - ETA: 0s - loss: 0.0022 - acc: 0.000
 384/1004 [==========>...................] - ETA: 0s - loss: 0.0022 - acc: 0.000
```

Figure 4: LSTM Training

F Figure 4 shows the training of the LSTM network which is undergoing 14th and 15th iteration and shows the variation in accuracy.

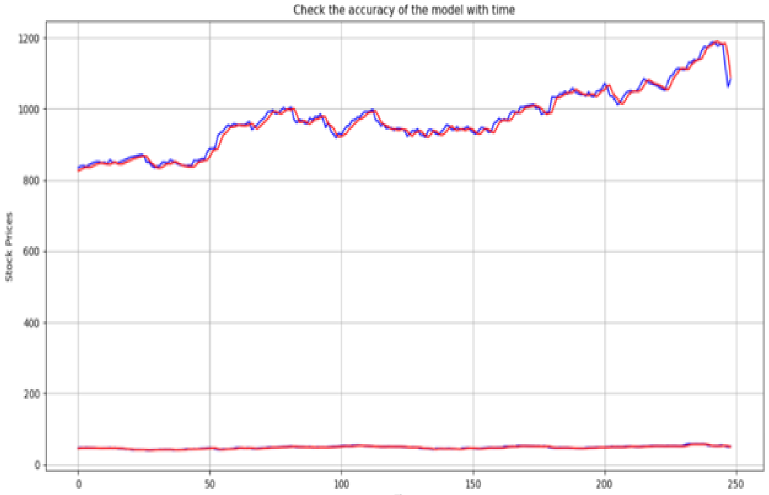


Figure 5: Original v/s Predicted Closing Price from LSTM

Table 1 depicts the comparative analysis of the three machine learning models chosen for predictions based on complexity, accuracy and customization. ARIMA being one of the most popularly used statistical approach is easy to build and train as time complexity and the computational space required is much lesser than the other models compared. It was observed that ARIMA gives higher forecast window which could be adjusted as needed whereas for SVR the forecast period is limited in order to not compromise with accuracy. Support vector regression provides the best result when used with radial basis function instead of the linear or polynomial functions. Although SVR gives better accuracy it is not as robust compared to deep learning approach. LSTM network gives the best accuracy even when the window period for closing rate prediction is increased, but the training of the neural network takes significantly higher amount of time with higher computation requirement. An advantage with LSTM network is that when there is a need for filtering or cleaning of data or additional information required on the input data, this could easily be achieved by adding successive layers to the network as LSTM is a sequential network.

Table 1: Comparative Analysis for ARIMA, SVR and LSTM

Approach used	**Training Time**	**Accuracy**	**Classification of stock based on companies**
ARIMA	85.2%	95% confidence interval	Not Applicable
SVR-RBF	96%	95% confidence interval	Not Applicable
LSTM	96.2%		Closing prices could be personalized based on the companies by adding a filter.

5. Conclusion

Stock price forecasting is a popular and important domain in financial and academic studies. To understand and identify trends in the stock market has been the goal of numerous market analysts, but the patterns of stock trend is always difficult to predict. In this study, three diverse machine learning models have been implemented and analysed for comparing the predictive accuracy of closing price which is one of the important aspects of stock trend analysis and investments for the stock traders. The models are analysed for complexity in model building, training, analysis and customization. The advantages and limitations of the models are discussed with respect to the mentioned parameters to enable users to choose the model that best fits their usage. For Future enhancement sentiment analysis could be carried out determining the popularity of a company and its influence on the stock price along with these model discussed.

References

1. A.A. Adebiyi, A.O. Adewumi, C.K. Ayo.Stock Price Prediction Using the ARIMA Model. AMSS 16th International Conference on Computer Modelling and Simulation, 2014.
2. Tian Ye. Stock Forecasting Method Based on Wavelet Analysis and ARIMA-SVR Model 3rd International Conference on Information Management , 2017.
3. C.C. Aggarwal.Neural Networks and Deep learning Springer Publication India, 2018.
4. Tom M. Mitchell.Machine Learning, McGraw Hill Education, India, 2017.
5. P. Li C. Jing T. Liang M. Liu Z. Chen L. Guo.Autoregressive Moving Average Modeling in the Financial Sector,2O15.
6. M. Usmani, S.H. Adil, Kamranraza , S.S. Ali.Stock Market Predictions Using Machine Learning Techniques.
7. 3rd International Conference on Computer and Information Sciences (ICCOINS), 2016.
8. Tiwari, S., Bharadwaj, A. and Gupta, S., Stock price prediction using data analytics. In International Conference on Advances in Computing, Communication and Control (ICAC3) (pp. 1-5). IEEE, 2017.
9. R.K. Dase, D.D. Pawar, D.S. Daspute Methodologies for Prediction of Stock Market: An Artificial Neural Network, International Journal of Statistika and Mathematika Vol 1 Issue 1 pp O8-15, 2011.
10. Greaves B. AuUsing the bitcoin transaction graph to predict the price of bitcoin No Data 2O15.
11. R.P. Schumaker. H. Chen.Textual Analysis of Stock Market Prediction Using Financial News, Americas Conference on Information Systems, 2006.
12. A Abraham, B Nath,P.K Mahanti, Hybrid intelligent systems for stock market analysis in: Computational Science-ICCS, Springer, pp. 337-345, 2007.
13. S. McNally, J. Roche S. Caton. Predicting the Price of Bitcoin Using Machine Learning,26th Euromicro, International Conference on Parallel, Distributed, and Network-Based Processing , 2018.

14. P.C. Chang C.H Liu J.L Lin, A neural network with a case based dynamic window for stock trading prediction. Expert Systems with Applications Vol 36. pp.6889-6898, 2009.
15. Q. Cao K.B., Leggio, M.J. Schniederjans. A comparison between Fama and French"s model and artificial neural networks in predicting the Chinese stock market, Computers Operations Research, 32, 2499-2512, 2005.
16. B.G. Malkiel, A random walk down Wall Street: including a life- cycle guide to personal investing, Completely, 1999.
17. Y. Hu, K. Liu, K. Zhang, L. Su, M. Liu, Application of evolutionary computation for rule discovery in stock, Algorithmic trading: A literature review. Applied Soft Computing Vol 36, pp. 534-551, 2016.
18. S.B. Achelis. Technical Analysis from A to Z. McGraw Hill New York, 2OO1.
19. Hussain, Sadiq, C. Akif, Josan D. Tamayo, and Aleeza Safdar. Big data and learning analytics model. International Journal of Computer Sciences and Engineering Vol 6, Issue 7 pp: 654-663,2018.
20. Fernandes, Marie. Data Mining: A Comparative Study of its Various Techniques and its Process. International Journal of Scientific Research in Computer Science and Engineering Vol 5, Issue 1, 19-23, 2017.