Stock Prediction using Python and Machine Learning

Samayak Malhotra
Prachish Pandey

Bits Pilani Dubai
Campus, Academic City
Faculty :Dr. Siddhaling Urolagin

I. Abstract

Stock trading has become very common in these todays world. Predicting the future cost of a share is the key to making revenue inside the stock marketplace. Stock brokers in particular use fundamental analysis and time series for the stock price prediction but these techniques are very slow and unpredictable. The programming language that we utilized in predicting the stock price is Python. In this research paper we propose a Machine Learning method referred to as Support Vector Machine so one can be taught from the given shares data, benefit intelligence after which use the obtained expertise for an accurate prediction of the shares within the inventory marketplace. A lot of experiments have been performed and the highest accuracy we achieved with the SVM model was 89.71% for Delta Airlines (DAL) inventory.

Keywords: *Machine Learning, Linear Regression, Stock Market, Predictions, Support Vector Machine, LSTM.*

II. Introduction

Predicting how the stock market will perform is one of the most difficult things to do these days. The stock market could be very volatile, as we saw during the spread of COVID-19 there has been a crash in the market, the U.S. market dropped low by 2000 points. There are such a lot of elements to look at while making the prediction – physical factors vs. Physiological, rational and irrational behavior, etc. All those elements integrate to make share prices terrific risky and really difficult to predict with a high level of accuracy. Using capabilities like the modern announcements about an enterprise, their quarterly revenue outcomes, etc come into play when estimating the inventory rate. News performs a prime function in inventory price movement. Machine studying strategies have the capability to deliver out the patterns and insights we haven't seen earlier and those can be used to make impeccable predictions.
In this article, we will work with historical facts and latest data available to use from different sources. We will enforce a mix of machine learning algorithms and LSTM to predict the future inventory rate of this employer. The main idea behind this text is to show how these algorithms are carried out.

III. Literature Survey

Stock Market prediction has usually had a positive enchantment for researchers and has won a variety of significance over the last few years. While numerous clinical tries were made, no method has been observed to accurately expect stock fee movement.[1]

Since monetary inventory markets generate sizable quantities of records at any given time a exquisite quantity of records needs to go through evaluation earlier than a prediction may be made.

Hence, we see that there's a awesome need to broaden a dependable stock prediction technique. Even with a loss of steady prediction techniques, there have been some moderate successes in this area. Stock marketplace research sums up 2 elemental buying and selling ideologies; Fundamental and Technical procedures [2].

In Fundamental evaluation, Stock fee movements are believed to derive from a security's relative statistics. Fundamentalists use profits, ratios, and management effectiveness to decide destiny forecasts. In Technical evaluation, market timing is the important thing. Charts and modeling techniques are used to discover traits in price and volume.[3] One vicinity of confined success in Stock Market prediction comes from textual records. Information from quarterly reviews or breaking news testimonies can dramatically have an effect on the stock fee. These techniques assign weights to key phrases in part with movement of a percentage price. These sorts of analyses have shown a specific but feeble capacity to forecast the route of inventory prices.[4]

We also collected statistics approximately Linear regression that is one of the strategies utilized in stock charge prediction. Each of techniques given beneath Linear regression has its personal benefits and downsides over its opposite numbers. This technique works with the aid of fitting the least rectangular technique and lowering a disabled variation of the least rectangular loss characteristic. Even if we use the identical dataset, the output could be extraordinary for every method. Conversely, least squares technique may be applied to in shape nonlinear fashions.[5]

The studies paper tries to predict the inventory price the usage of the preceding economic information. There are many thing affect the price of stock market fees such as public sentiment, traders opinions, information and different events including wars, elections, pandemic occasions, and so on. [6] Changes within the organisation's control, adjustments in the services presented by way of the employer can exceedingly have an effect on the company's stock value. The prediction can be made more accurate if use an accurate dataset and a model that takes the above factors into consideration [7].

In [8] the editor talks about the volatility in forecasting the stock marketplace and the numerous variables which can have an effect on it. The clinical studies is conducted via studying the feelings of humans and the enterprise. Social media has a big effect and can be beneficial in predicting the inventory marketplace rate. Machine learning algorithms are used and technical analysis is carried out that includes collecting records from social media accounts, extracting sentiments from news platforms. Stock fees from the beyond years are collected [9]. Time collection is mainly based on the Support Vector Machine(SVM) and the effects acquired via this approach had been then established using binary type . The SVM parameters were without difficulty treated the usage of Cuckoo test and Swarm Intelligence technique used for optimization.[10]

IV. Proposed System

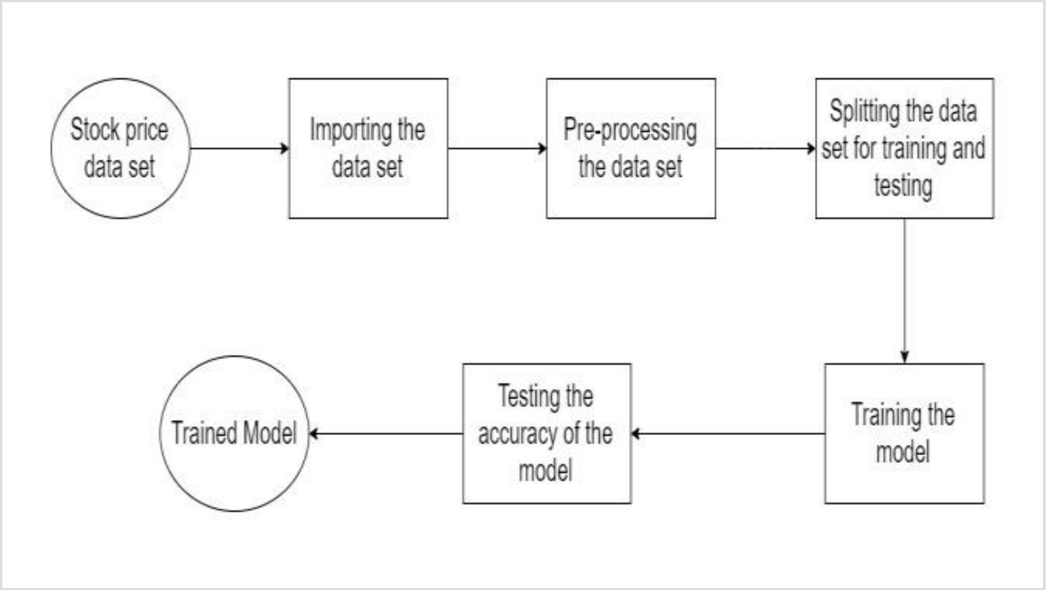

Stock price data set
Importing the data set
Pre-processing the data set
Splitting the data set for training and testing
Training the model
Testing the accuracy of the model
Trained Model

Figure 4.1 - Block Diagram for working of Model

We made use of several modules already available in Python .The "SciKit-learn" library contains algo for classification ,regression and clustering.
Numpy library had also been used to manage multi-dimensional arrays and huge set of matrices. We further used Matplotlib and Pandas for plotting the data and operating on the time series .

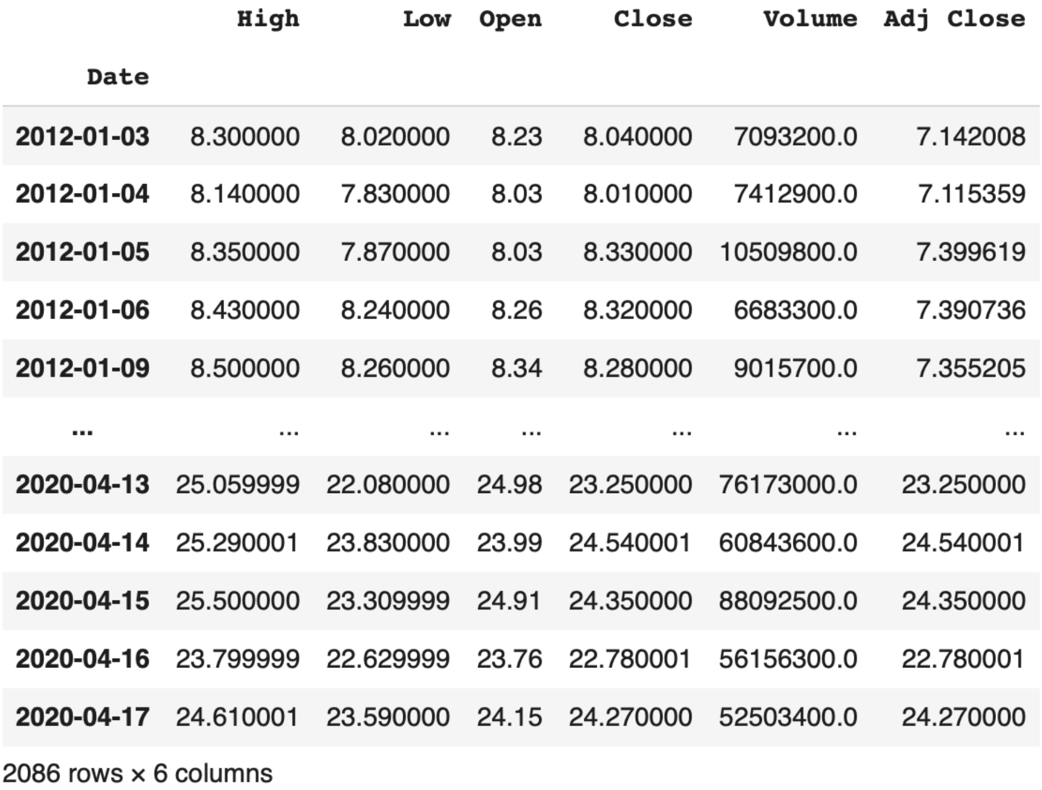
a)Data Collection

8 years of dataset and news data were collected from Yahoo finance platform from the year 2012 to 2020 for DEL (Delta airlines) shares. The data collected from the source contained the following records :- open bid, close bid, high bid, and low bid, volume and adj-close.

b)Importing and Preprocessing of Data

Stock price data is taken from yahoo finance and as seen in Figure 4.2 our dataset contains 2066 records that contains DEL stock data from 2012 to 2020.

We import this dataset and look for the shape of the dataset. This gives us an idea about the size of the data. We then move on to data cleaning by removing the irregularity and any outlier data points that exist. We use inbuilt functions to fill the respective vacant spaces in case of missing data . The dataset is then divided into 80:20 ratio .The 80 % of the dataset is used for the training and the rest 20% is used for checking the accuracy of the model.



	High	Low	Open	Close	Volume	Adj Close
Date						
2012-01-03	8.300000	8.020000	8.23	8.040000	7093200.0	7.142008
2012-01-04	8.140000	7.830000	8.03	8.010000	7412900.0	7.115359
2012-01-05	8.350000	7.870000	8.03	8.330000	10509800.0	7.399619
2012-01-06	8.430000	8.240000	8.26	8.320000	6683300.0	7.390736
2012-01-09	8.500000	8.260000	8.34	8.280000	9015700.0	7.355205
...	...	...	...	...	...	...
2020-04-13	25.059999	22.080000	24.98	23.250000	76173000.0	23.250000
2020-04-14	25.290001	23.830000	23.99	24.540001	60843600.0	24.540001
2020-04-15	25.500000	23.309999	24.91	24.350000	88092500.0	24.350000
2020-04-16	23.799999	22.629999	23.76	22.780001	56156300.0	22.780001
2020-04-17	24.610001	23.590000	24.15	24.270000	52503400.0	24.270000

2086 rows × 6 columns

Figure 4.2- Dataset for DEL from 2012-2020

c) Training the model

The training sets are used to tune and fit the models. The test sets are untouched, as a model should not be judged based on unseen data. The training of the model includes cross-validation where we get a well-grounded approximate performance of the model using the training data. The idea behind the training of the model is that we some initial values with the dataset and then optimize the parameters which we want to in the model. This is kept on repetition until we get the optimal values. Thus, we take the predictions from the trained model on the inputs from the test dataset. Hence, it is divided in the ratio of 80:20 where 80% is for the training set and the rest 20% for a testing set of the data. After the training is done the model is then used to predict the future stock price values. We pass the rest 20 % of input variables to the model and command it to predict the target variable . The target variable that’s predicted by the model is the output generated by the trained model. Finally to check the models working efficiency we need to compare the predicted values against the real-world values.

d) Support Vector Machine

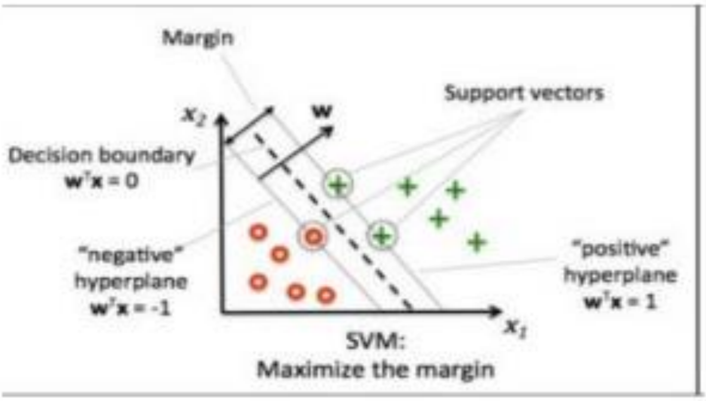


Figure 4.3 -SVM Decision making boundary

The Support Vector Machine algorithm draws a boundary over the data set called as the hyper-plane that separates the data into two classes as shown in the Figure 4.3 above.

Considering the same figure, if μ is some unknown data point and w is vector which is perpendicular to the hyper-plane, then the SVM decision rule will be

$\overline{w}\overline{\mu} + b \geq 0$ ——(1)

The width w of the hyper-plane is to be maximized

w = [2/ || w ||] .... (2)

w = (max [2/.|| w ||]) ……..…………………. (3)

Applying lagrange's multiplier as

L = 0.5|| w $||^2$ →-$\sum \alpha_i [y_i (\omega_i x_i + b) - 1]$................ (4)

L = $\sum \alpha_i - 0.5 \Sigma_i \Sigma_j \alpha_i \alpha_j y_i y_j x_i x_j$................................. (5)

The updated decision rule will be

($\sum \alpha_i y_i x_i$)μ+b ≥ 0

e)Time Series Analysis

Any metric that is measured over regular time intervals forms a time series. Analysis of time series is commercially importance because of industrial need and relevance especially with respect the demand, sales, supply etc).Each data point (*Yt*) at time *t* in a Time Series can be expressed as either a sum or a product of 3 components, namely, Seasonality (St), Trend (Tt) and Error (et).

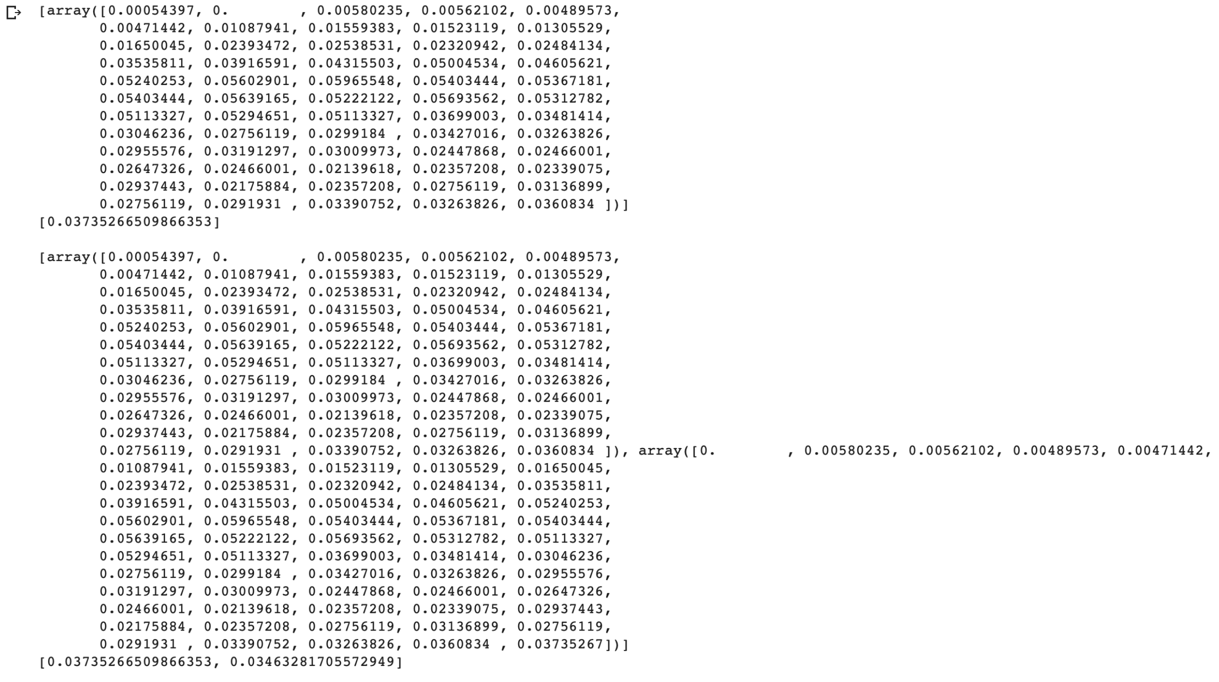
For Additive Time Series,

$Y_t = S_t + T_t + \epsilon_t$

For Multiplicative Time Series,

$Y_t = S_t \times T_t \times \epsilon_t$

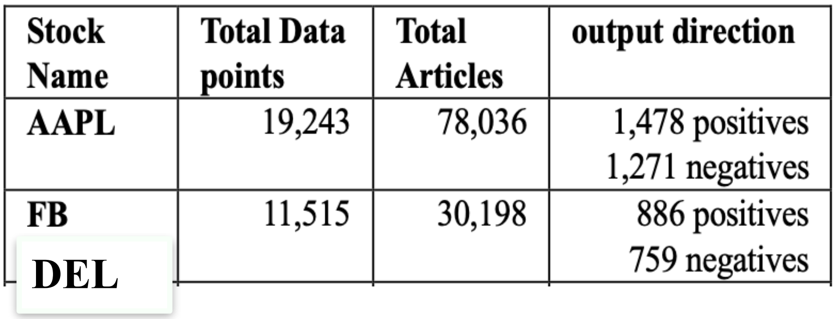
Figure 4.4- Numpy arrays



```
[array([0.00054397, 0.        , 0.00580235, 0.00562102, 0.00489573,
       0.00471442, 0.01087941, 0.01559383, 0.01523119, 0.01305529,
       0.01650045, 0.02393472, 0.02538531, 0.02320942, 0.02484134,
       0.03535811, 0.03916591, 0.04315503, 0.05004534, 0.04605621,
       0.05240253, 0.05602901, 0.05965548, 0.05403444, 0.05367181,
       0.05403444, 0.05639165, 0.05222122, 0.05693562, 0.05312782,
       0.05113327, 0.05294651, 0.05113327, 0.03699003, 0.03481414,
       0.03046236, 0.02756119, 0.0299184 , 0.03427016, 0.03263826,
       0.02955576, 0.03191297, 0.03009973, 0.02447868, 0.02466001,
       0.02647326, 0.02466001, 0.02139618, 0.02357208, 0.02339075,
       0.02937443, 0.02175884, 0.02357208, 0.02756119, 0.03136899,
       0.02756119, 0.0291931 , 0.03390752, 0.03263826, 0.0360834 ])]
[0.03735266509866353]

[array([0.00054397, 0.        , 0.00580235, 0.00562102, 0.00489573,
       0.00471442, 0.01087941, 0.01559383, 0.01523119, 0.01305529,
       0.01650045, 0.02393472, 0.02538531, 0.02320942, 0.02484134,
       0.03535811, 0.03916591, 0.04315503, 0.05004534, 0.04605621,
       0.05240253, 0.05602901, 0.05965548, 0.05403444, 0.05367181,
       0.05403444, 0.05639165, 0.05222122, 0.05693562, 0.05312782,
       0.05113327, 0.05294651, 0.05113327, 0.03699003, 0.03481414,
       0.03046236, 0.02756119, 0.0299184 , 0.03427016, 0.03263826,
       0.02955576, 0.03191297, 0.03009973, 0.02447868, 0.02466001,
       0.02647326, 0.02466001, 0.02139618, 0.02357208, 0.02339075,
       0.02937443, 0.02175884, 0.02357208, 0.02756119, 0.03136899,
       0.02756119, 0.0291931 , 0.03390752, 0.03263826, 0.0360834 ]), array([0.        , 0.00580235, 0.00562102, 0.00489573, 0.00471442,
       0.01087941, 0.01559383, 0.01523119, 0.01305529, 0.01650045,
       0.02393472, 0.02538531, 0.02320942, 0.02484134, 0.03535811,
       0.03916591, 0.04315503, 0.05004534, 0.04605621, 0.05240253,
       0.05602901, 0.05965548, 0.05403444, 0.05367181, 0.05403444,
       0.05639165, 0.05222122, 0.05693562, 0.05312782, 0.05113327,
       0.05294651, 0.05113327, 0.03699003, 0.03481414, 0.03046236,
       0.02756119, 0.0299184 , 0.03427016, 0.03263826, 0.02955576,
       0.03191297, 0.03009973, 0.02447868, 0.02466001, 0.02647326,
       0.02466001, 0.02139618, 0.02357208, 0.02339075, 0.02937443,
       0.02175884, 0.02357208, 0.02756119, 0.03136899, 0.02756119,
       0.0291931 , 0.03390752, 0.03263826, 0.0360834 , 0.03735267])]
[0.03735266509866353, 0.03463281705572949]
```

V. Experimental Results



Stock Name	Total Data points	Total Articles	output direction
AAPL	19,243	78,036	1,478 positives 1,271 negatives
FB DEL	11,515	30,198	886 positives 759 negatives

In the above Table 5.1 we see the stock data details of Apple and Delta Airlines stocks. The evaluation metrics are directional accuracy, precision and recall .Based on accuracy model values we see that SVM outperforms other models like Recurrent NN, Support Vector Regression and Deep Neural Networks. All the accuracies achieved by the model were above 80% and in the case of DEL the accuracy was calculated to 89.71%. All our models achieved better results than other models we saw in the literature.

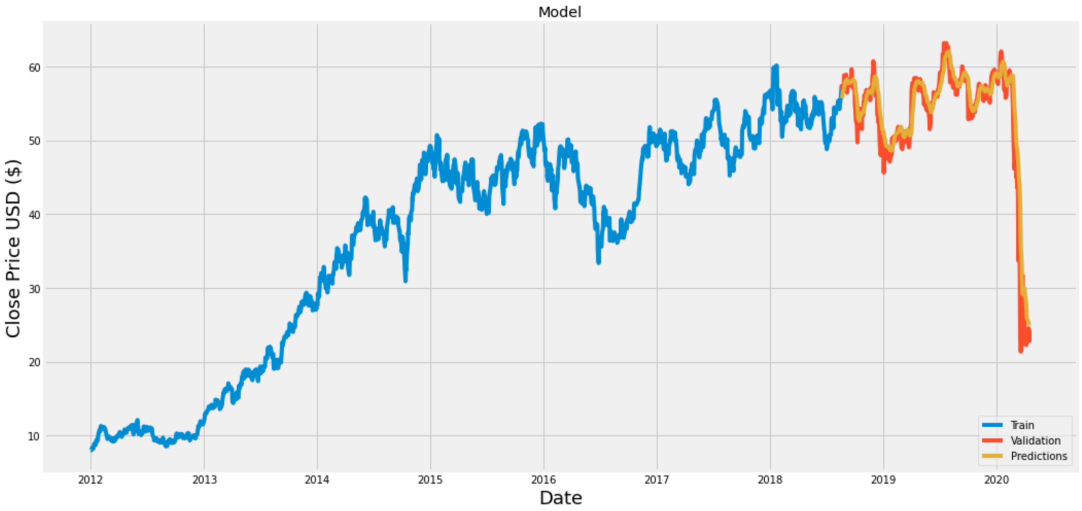
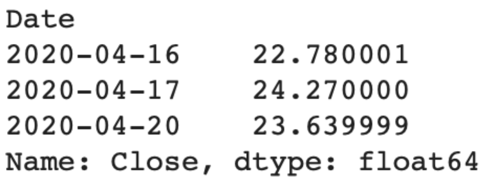


Figure 5.1.1- Data plot after training of Data

The values in the Figure 5.1.2 are the closing prediction done by our stock prediction model on the DEL stock on the given dates.



```
Date
2020-04-16    22.780001
2020-04-17    24.270000
2020-04-20    23.639999
Name: Close, dtype: float64
```

Figure 5.1.2- Closing predictions DEL

VI. Conclusion

A large body of work was presented in this report. Two of the most widely used methods, Fundamental Analysis and Technical Analysis showed little promise in the experiments carried out. Technical Analysis specifically shows little to no potential of ever producing any statistically significant result when the correct methodology is applied. Machine learning techniques were then tested on a wide variety of data sets. The result of few of these models looked hopeful, but failed when put up through the realistic trading systems. Stock market is different in practice from theory we already know and this report shows that stock market prediction is an extremely difficult.

VII. Reference Papers

1) https://ieeexplore.ieee.org/abstract/document/4424794/

2) https://ieeexplore.ieee.org/abstract/document/6889969

2) https://ieeexplore.ieee.org/abstract/document/5726498

3) https://ieeexplore.ieee.org/abstract/document/501826

5)https://link.springer.com/chapter/10.1007/978-3-319-11310-4_76

6)https://www.researchgate.net/publication/259240183_A_Machine_Learning_Model_for_Stock_Market_Prediction

7)http://www.academia.edu/Documents/in/Stock_Market_Prediction

8) https://www.researchgate.net/publication/331829113_Stock_Market_Trend_Prediction_using_Machine_Learning

9) https://www.researchgate.net/publication/330677469_Stock_Market_Data_Prediction_Using_Machine_Learning_Techniques_Proceedings_of_ICITS_2019

10) https://jfin-swufe.springeropen.com/articles/10.1186/s40854-019-0131-7