# Performance based comparison of different Machine Learning and Deep Learning models for Stock Prediction

Adhaar Sharma Nikhil Gupta Nishant Braiwal
Electronics and Communication Engineering
Bharati Vidyapeeth College of Engineering, Paschim Vihar, Delhi

***Abstract:*** *The uncertainty among the stock prices is area of interest among many investors during these unpredictable times of pandemic. It becomes increasingly useful to be able to predict how stocks behave in such anomalous cases. Recurrent Neural Network (RNN) and Long Short-Term Memory prove to be extremely competent in terms of performance and accuracy in predicting the graph of a particular stock in the near future as well as over a long period of time..*

***Keywords:*** *Logistic Regression, Decision Trees, Recurrent Neural Network (RNN), ARIMA, SARIMAX, Long Short-Term Memory (LSTM)*

## I. INTRODUCTION

There are many variables to consider while considering the fluctuation in price of a stock. These variables often tend to be very subtle. How a stock rises or falls often depends on the mindset of the investor. During Stock Market Price of 1929, Recession during 2008 and amidst Pandemic 2020, the buyers often tend to panic and the price of the stock tend to fluctuate more than usual. Prediction of stock in times like these becomes more and more challenging and therefore, more advanced machine learning and deep learning models have to be employed to be able to predict the graph accurately. This would help a smart investor to make a lot of money even during financial crisis while managing risk.

When we need to employ a neural network to process input data in which sequence is of utmost importance and the output data varies with respect to time, then time series analysis models like ARIMA and various Recurrent Neural Networks (RNN) prove to be extremely powerful for sequential data processing. When we read a fiction book, we understand the context of the storyline on the basis of what we have already read, and we don't need to think from scratch everytime we read a word. What we have while reading a book is a sequential understanding of the narrative which we try to replicate in Recurrent Neural Network model while understand a stock dataset varying with time.

Traditional Neural Networks are incapable of classifying the event happening in the plot of the book at a particular time. The working of Traditional Neural Network while predicting future based on past events is unclear. This is the shortcoming which RNNs aim to solve. The basic machine learning models like logistic regression prove to be extremely primitive as they can only classify the whether the prospect of investing in the stock is good or bad. They fail to examine the extent to which the stock fluctuate and the accuracy of the model tends to be low as the dataset becomes more and more complex. Employing such model during times of high uncertainty is not a good idea from the point of view of a smart investor.

The Logistic Regression model despite its simplicity is extremely useful for predicting the trajectory of stock which are considered to be comparatively safe i.e. stocks with low fluctuations. Therefore, prediction of government bonds and index funds using this model is still feasible for investing. The model would also serve those who tend to invest from a long term point of view. Logistic regression is used to find whether the company is going to suffer financially and if it is bad to invest in it. If we use multiple independent variables to predict the stock trajectory, a decision tree algorithm comes extremely handy. It is one of the earlier models used for the purpose of stock prediction. A decision tree technique runs a loop to give a result and when the output is graphed, it appearance is similar to that of a tree and thus, it is so called. Using this model is a major improvement over logistic regression because realistically, we tend to have many relationships over the multiple variables.

While computing day to day real time data, we tend to employ time series analysis models of deep learning like ARIMA and SARIMAX. Time series analysis comprises of taking a set of data at a particular time in an equally spaced intervals and previously observed values are used to predict future values. In time series, Time is the mandatory variable. The major components of Time Series Analysis are Trend, Seasonality, Irregularity and Cyclic. medicine (EEG analysis), finance (Stock Prices) and electronics (Sensor Data Analysis) are the areas where Time Series is prevalent technique for prediction. Idea of connecting previous information to present task is one of the major appeals of RNNs. Sometimes, only recent information is required to perform the present task i.e. gap between the relevant information and the place that it is needed is small and RNNs are great at dealing with such predictions. Sometimes the context of the data which is needed to be understood by the neural network may be more subtle and harder to identify. It is also possible that the gap between relevant information and the point where it is needed is extremely large. This is the problem of 'long term dependency' which is solved by employing Long Short-Term Memory (LSTM) model. LSTMs are very much capable of remembering information for longer periods of time as it is practically their default behaviour.

## II. RELATED WORKS

Over the past century, a lot of data on price of various stock has been accumulated and therefore, we are capable of inferring a lot of information about the future through various models which have been studied over the past few decades.

In the field of prediction stock market values, the most primitive approach was the use of OLS technique on the data. Statisticians and Machine Learning practitioners collaborate to come up with models based on time series. Some of such popular models are ARIMA, SARIMA and kNN-TSPI. Other models like Granger causality test with other regression variations are also prevalent.

Significant work has been done in the field. A testament to which is work of Chuanrui Fu [1] used empirical analysis with nonparametric quantile regression to find a correlation using absolute return and trading volume as the parameters. They used China stock market dataset for the same. Robert P. Schumaker [2] and Hsinchun Chen employed language processing techniques with other trends based on finance and statistics to predict discrete stock prices. Avijan Dutta [3], Gautam Bandopadhyay and Suchismita Sengupta employed ratios used for stock analysis and applied Logistic Regression model on it to predict stock prices.

Chih-Fong Tsai [4] and Yu-Chieh Hsiao improved upon feature engineering, data mining and existing pre-processing techniques. They found ways to eliminate redundant data and made their model more performance efficient for stock prediction. C.N.W. Tan [5] and G.E. Wittig used deep learning techniques based on artificial neural networks (ANN) with improved backpropagation optimizers and weight updating methods to improve upon the model's capability to predict the stock prices.

Anshul Mittal [6] and Arpit Goel used human psychology and language processing techniques to predict the mindset of the investors. They collaborated sentiment analysis with machine learning techniques to predict whether the investor will invest or sell a particular stock. The uncertainty of the graph and corresponding fluctuation in the graph was also anticipated. Ayodele Ariyo Adebiyi [7], Aderemi Oluyinka Adewumi, and Charles Korede Ayo used time series analysis and various models based on the statistics behind time series to forecast the stock graph. ARIMA and artificial neural networks model were employed and the dataset used was stock data obtained from New York Stock Exchange.

## III. METHODOLOGY

Various machine learning and deep learning models have been so chosen such that they are suitable for stock price prediction. The models we have chosen for the same are given below.

Machine Learning Models

1. Logistic Regression
2. Decision Trees

Deep Learning Models

1. ARIMA
2. LSTM

We have divided our methodological process into various stages given below:

1. Selection of Dataset: While choosing our dataset, we should care for its validity as well as consider its age. We should pick the latest dataset for training our model from appropriate source. We have chosen our dataset containing stock price of Apple stock over past few years from Kaggle.
2. Data Preprocessing
3. The clean dataset is now ready to be divided into training and testing sets. Training values consist of 90% of the dataset whereas the remaining dataset is used as testing the trained model.
4. The required features to be fed to the model for training are extracted. In our dataset, features that we have selected are date, open, high, low, close, and volume.
5. Data is now fed to the model so that it could be trained. Hyperparameter tuning is done using cross-validation techniques. This model would then be able to predict output.
6. The predicted output is compared to the actual output in the test set of the dataset.

## IV. MODELS

### A. Linear Regression

One of the most basic models to find a linear relationship to predict an output is Linear Regression model. The major advantage which it has to offer is its simplicity. It takes certain data points and tries to fit a line in such a way that it is average of all the data points of the dataset and therefore, we arrive at a correlation which helps us in predictions. For example, we may use variables affecting salary of a person to predict it if the variables are already given. These variables may be skillset, experience, position etc.

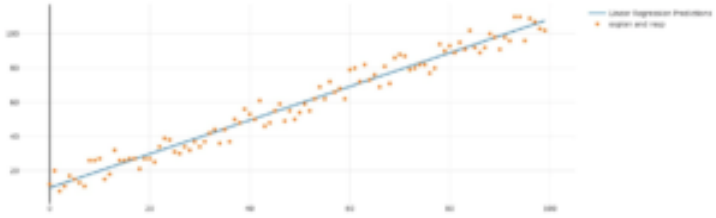

*Figure 1: Example linear regression equation: 4x+7y-10z+15=Y*

### B. Logistic Regression

Logistic Regression is used for tradition statistics and mainly used for classification problems. For example, if we want to find if a person is obese or not on the basis of the height and weight of the person then Logistic Regression would prove to be useful in such cases. In such classification cases, a linear regression model may try to overfit the data giving inaccurate results. Both Linear Regression and Logistic Regression models are capable of working with continuous data as well as discrete data. In other words, its result is either one thing or another. Logistic Regression estimates the link between the dependent variable and the additional independent variables, by approximating likelihood of using its fundamental logistic function.

A yes or no output is required by these probabilities to make an appropriate prediction. Sigmoid function is one of the most commonly used function for this algorithm. The Sigmoid function appears to be S-shaped and the range of this function is (0,1). A threshold classifier is used to give an output which is either 0 or 1.

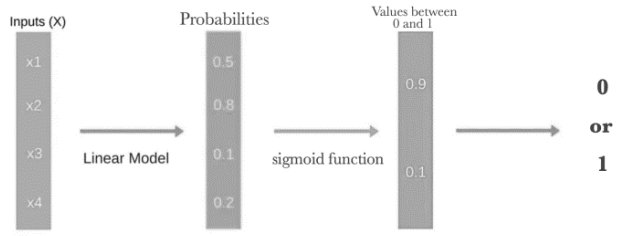


*Figure 2: The depiction above demonstrates the procedure of logistic regression..*

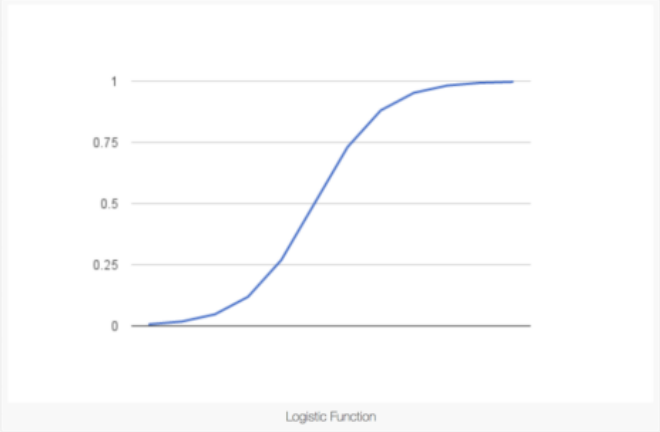


*Figure 3: Representation of sigmoid function*

To get the best likelihood, we can take an arbitrary data and use for categorisation process. This is a common methodology for approximating probabilistic model parameters. Various optimization techniques or inbuilt optimizers are used to reach a better accuracy. We can use cross-validation process to find better models for the same. Newton's Method is one of the popular optimization technique. Gradient Descent can also be used for the same.

The effectiveness of this model is proved again and again, it does not need numerous high end gpu for execution and therefore, it is highly interpretable by low end machines as well. Input scaling is not required, regularization is easy and its results are well-calibrated and are statistically accurate.

Logistic Regression is as efficient as linear regression and works really well when elimination of various attributes which are not useful to the output parameter and other attributes are related to each other in some way or the other. Feature Engineering is one of the very important techniques used to determine how well the algorithm performs and either Logistic Regression, Linear Regression or any other model is selected accordingly. Logistic Regression has a very easy implementation and the training process is also extremely efficient. Logistic Regression model is often used as a reference comparison and we can select algorithms with higher complexity accordingly.

A shortcoming of Logistic Regression is that non-linearity affects the performance negatively and we have to use dataset with linear relationships to have higher accuracy and therefore, it is not feasible to implement it everywhere because of its lower flexibility to adapt to different relationships among data points. Looking at the example below will make things clearer.

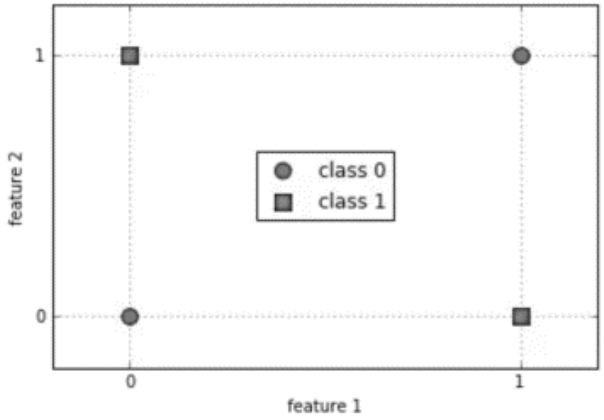


*Figure 4: Two binary features from two examples.*

Linearly separable line cannot be drawn for the below given classes without compensating for the huge inaccuracy. Implementing decision tree is more apt choice for such cases.

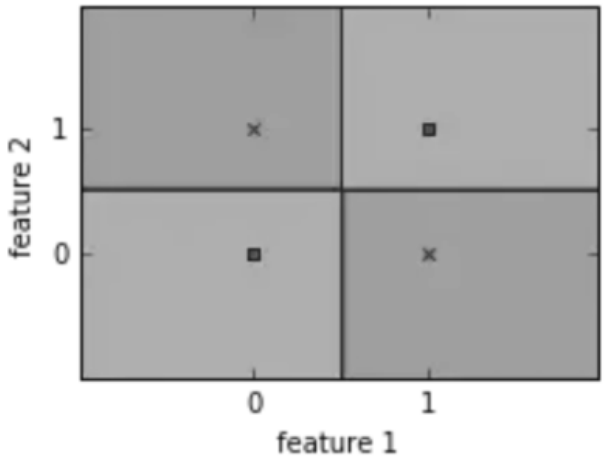


*Figure 5: Linearly Inseparable Data points.*

C. Decision Tree

The entropy of a system is measure of its disorder and of the unavailability of energy to do work.

Mathematical expression of Entropy is denoted as -

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

We need to decide which feature to use while splitting the nodes of decision tree and this is done by calculating the entropy beforehand. We may also use ID3 approach for the same. If we select high entropy feature then the decision tree will branch a lot which will make the algorithm less performance efficient. We have considered 1000 points here in the dataset which we are working on. 200 and 800 data points are given to positive and negative class respectively. P+ is equal to 0.2 and P- is equal to 0.8

Calculating the entropy using above mentioned entropy formula, we get.

$$-\left(\frac{2}{10}\right) * \log_2\left(\frac{2}{10}\right) - \left(\frac{8}{10}\right) * \log_2\left(\frac{8}{10}\right) = 0.217$$

We have calculated the entropy in this case to be 0.217 which is considered pretty low and we can use it as node given if it is lower than other features. This is part of feature engineering which is much explored area in field of machine learning. The range of Entropy is between 0 and 1. There may be cases where entropy may come to be larger than 1 which may mean that it has very high level of disorder and therefore unsuitable feature to select first.

Decision Trees are based around the idea of asking questions in a sequence such that each question lead us closer to the expected answer. The goal is to guess what expected value is going to be. By asking these sequential questions that can only be answered by "Yes" or "No," if we are playing well, then each answer helps you ask a more specific question until you get the right answer. Questions are selected in such a way that the decision tree that guides you further to more specific questions and ultimately to the answer. Questions are based around the independent variable.

D. ARIMA

Time series are usually hard to extrapolate to give an accurate result. Data scientist although can predict the stock patterns to some extent but it is not a good idea to do so as it is hard to account for all the independent variables which affect the stock prices. Investing based on their models, having forecast with precision and accuracy is easier said than done. In reality, the market is not outperformed by the hedge funds and accuracy in the prediction is extremely short-lived for a given dataset and new data needs to be constantly updated. There are issues with a lot of fluctuations and unbeknownst factors may surface leading to errors in prediction. This has to be accounted for and new techniques have to be used to prevent this. Models tend to have naive dependencies which behave sporadically the interaction between linear and nonlinear parameters of the dataset may confuse the model to derail its prediction.

ARIMA model is one of the most proficient model for brief prediction but they fail to predict the future graph for a longer time period. If the predictions need to be performed for only a short period, then the ARIMA model has an edge over all the other models for the purpose of stock prediction. ARIMA stands for Autoregressive Integrated Moving Average. ARIMA is better known as the Box-Jenkins approach.

The mathematical depiction for the same is denoted by:

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

Accuracy and reliability of ARIMA output graphs is way more than traditional regression models. There are very few uni-variate models which are used and one of the most popular ones is ARIMA model. Uni-variate model is a model which has only one working variable. It is harder to manipulate the necessary variables to improve the accuracy. The importance of these variables may be high in certain case and in those case it would be better to go for multi-variate models as they can compensate for more than one variables. Dynamic Regression will be better in such cases.

Seasonality, stationarity are some of the important factors which need to be considered while implementing ARIMA models. ARIMA models work well with non-seasonal dataset and if dataset has any sort of seasonality then it has to be accounted for with data pre-processing techniques. p represents autoregressive terms, d represents non-seasonal differences needed for stationarity, and q represents the lagged prediction inaccuracies in the equation of the model.

Seasonal ARIMA consists of six major parameters which are p,d,q,P,D,Q, P denotes seasonal autoregressive terms, D denotes seasonal differences and Q denotes seasonal moving-average terms.

Inferring from the succeeding conclusions given below, interaction with seasonal ARIMA to get a feel for the intricacies involves graph manipulation using various techniques.

This model is more constrained. Given fundamental system can be too multifaceted and graph fitting will prove to be a harder task. For a simpler fundamental models, this model can be easily implemented without any issue and it is much more effective than other deep learning approaches.

E. Long Short Term Memory (LSTM)

Long Short Term Memory networks abbreviated as "LSTMs" are unique type of RNNs, efficient in learning dependencies with large gap between them. They were improved upon and pioneered by various individuals working in the field. Their usefulness was proved on a different kinds of problems, and is now prevalent in various applications like language processing, face detection and context understanding processes.

The unequivocally aim of LSTM is to take into account those cases where we may find dependency with huge gaps problem. The special attribute of these models is to retain useful data over a significantly large data and to discard any information which doesn't fit into the expected result from the dataset during training process. Other models on the other hand struggle with cases of long term dependencies.

Sequential chain of repeating units are used in Recurrent Neural Networks. These RNN are very versatile and we can use vector to sequence model, sequence to sequence model, vector to vector model or sequence to vector model for the purpose of mapping one kind of variable to different kind of variable or different form of variable. We use LSTM cells in the LSTM models which consists of Input gate, Output gate, Forget gate and Cell state. The flow of information through LSTM is explained through the diagram given below.

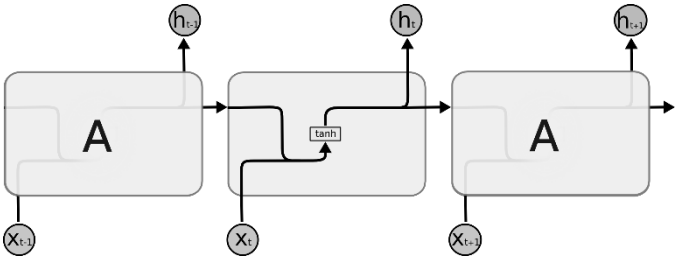


*Figure 6: The repeating module in a standard RNN contains a single layer.*

Each layer in the LSTM has a purpose. The forget gate discards the information not needed to get to predicted result. All the information is updated in the cell state which acts as a conveyer belt of information. Input gate and Output gate respectively is used to feed and receive information from each cell unit of LSTM.

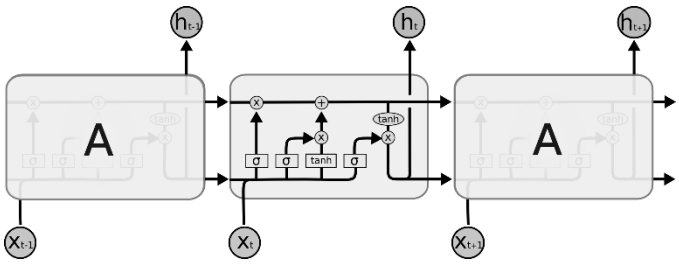


*Figure 7: The repeating module in an LSTM contains four interacting layers.*

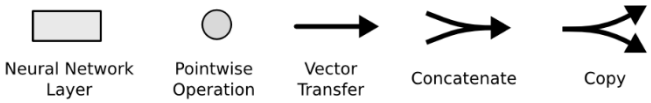


*Figure 8: Notations*

## V. EXPERIMENTAL ANALYSIS

Results obtained using Linear Regression:

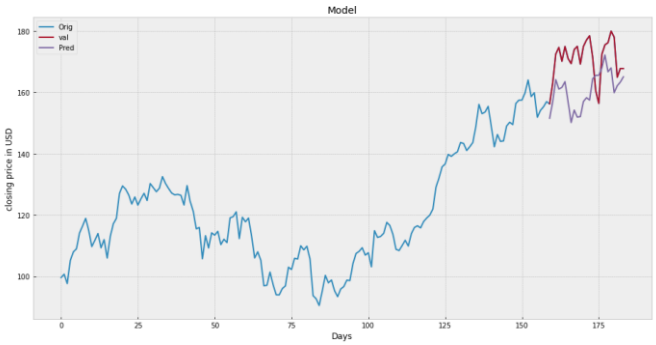


*Figure 9: Forecast using Linear Regression.*

```
regressor = LinearRegression()
regressor.fit(x_train,y_train)
accuracy = regressor.score(x_test,y_test)
print(accuracy*100,'%')
```

```
34.658382388342865 %
```

Results obtained using Decision Tree:

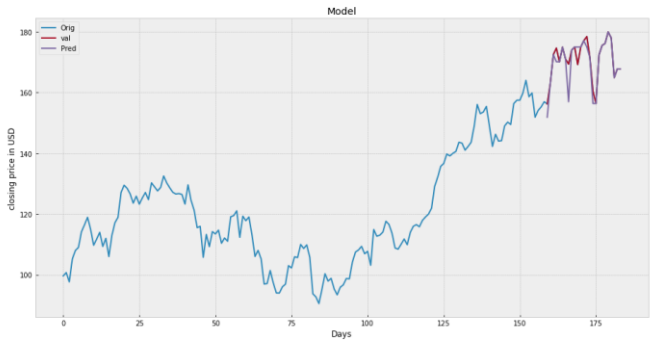


*Figure 10: Forecast using Decision Tree.*

```
regressor = DecisionTreeRegressor()
regressor.fit(x_train,y_train)
accuracy = regressor.score(x_test,y_test)
print(accuracy*100,'%')
```

```
4.751131752726279 %
```

Results obtained using ARIMA

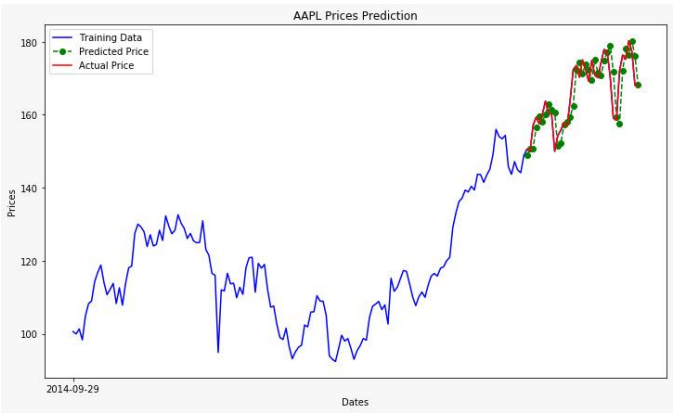


*Figure 11: Apple Stock Forecast using ARIMA.*

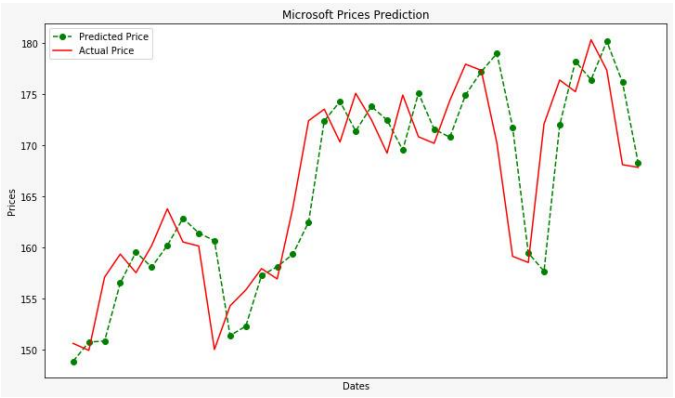


*Figure 12: Microsoft Stock Forecast using ARIMA.*

Result using LSTM:

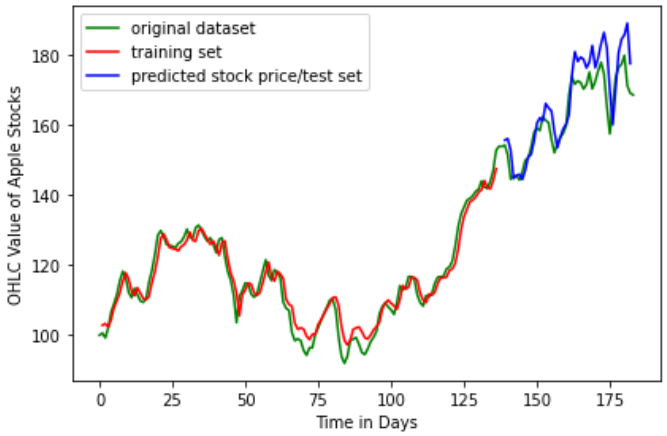


*Figure 13: LSTM Model using Stock Prediction*

## VI. CONCLUSION

Machine learning and deep learning models which can efficiently predict the stock market price values over time have been designed and used successfully to predict stock prices over a year. (2018).

While implementing various machine learning models, we have found that their accuracy is good for simpler datasets. However, for complex dataset, deep learning are better suited.

LSTM will perform better if it is provided with larger datasets for training, but with our data sets also it have produced quite good results.

As different techniques of machine and deep learning have produced different performances, as a future scope of work we plan to explore the possibility of fine tuning these models and aim for higher prediction accuracy.

## VII. REFERENCES


C. Fu, "Nonparametric quantile regression analysis on the price-volume relationship in China stock market," *2010 2nd IEEE International Conference on Information Management and Engineering*, Chengdu, 2010, pp. 86-91, doi: 10.1109/ICIME.2010.5477637.

Schumaker, Robert P., and Hsinchun Chen. “A Quantitative Stock Prediction System Based on Financial News.” *Information Processing & Management*, Pergamon, 29 May 2009.

Dutta, Avijan, and Gautam Bandopadhyay. “prediction of stock performance in Indian Stock Market using logistic regression.” International Journal of Business and Information.

Tsai, Chih-Fong, and Yu-Chieh Hsiao. “Combining Multiple Feature Selection Methods for Stock Prediction: Union, Intersection, and Multi-Intersection Approaches.” Decision Support Systems, North-Holland, 21 Aug. 2010.

Torres, Edgar & Alvarez, Myriam & Torres, Edgar & Yoo, Sang Guun. (2019). Stock Market Data Prediction Using Machine Learning Techniques: Proceedings of ICITS 2019. 10.1007/978-3-030-11890-7_52.

Nabipour, Mojtaba & Nayyeri, Pooyan & Jabani, Hamed & Mosavi, Amir. (2020). Deep learning for Stock Market Prediction.

Masum, Shamsul & Liu, Ying & Chiverton, John. (2020). ARIMA-VS-LSTM

D. M. Q. Nelson, A. C. M. Pereira and R. A. de Oliveira, "Stock market's price movement prediction with LSTM neural networks," *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, 2017, pp. 1419-1426, doi: 10.1109/IJCNN.2017.7966019.

S. Liu, G. Liao and Y. Ding, "Stock transaction prediction modeling and analysis based on LSTM," *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, Wuhan, 2018, pp. 2787-2790, doi: 10.1109/ICIEA.2018.8398183.

X. Tang, C. Yang and J. Zhou, "Stock Price Forecasting by Combining News Mining and Time Series Analysis," *2009*

A. A. Ariyo, A. O. Adewumi and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," *2014 UKSim-AMSS*

J. Gong and S. Sun, "A New Approach of Stock Price Prediction Based on Logistic Regression Model," *2009 International Conference on New Trends in Information and Service Science*, Beijing, 2009, pp. 1366-1371, doi: 10.1109/NISS.2009.267.