Stock Prediction Using LSTM, Decision Trees and Linear Regression

Varun Ghantasala
gvarun@vt.edu

Abstract

In today's world of finance understanding stock prices is of utmost importance. A lot of stockbrokers are dive deep into various analysis for stock price prediction to place their bets. This paper explains an approach to understand the volatility of stock prices using Machine Learning Techniques. This approach is ready to be used for any stock to conduct a detailed analysis and predict stock prices of the future. In this paper Machine Learning strategies called Linear Regression, LSTM and Decision Trees are used to predict stock prices. Their accuracy and reliability shall be explored in the paper. A comparision between the methods shall also be made.

1 Introduction

The volatility of stock prices makes it very hard and can result in inconsistent analysis if done by humans. Quantitative stock brokers try to find patterns in stock to invest in them. In today's world, Machine learning techniques are greatly used in understanding stock since a computer can compute prices and take into various variables to predict stocks. Some of these Machine Learning methods have proven to be useful. With methods that are depicted in the paper, a step towards understanding pattern and prediction of data can be made. In this paper, Linear Regression, Decision Trees and LSTM models have been used to predict stock prices.

1.1 LSTM

Long short term memory is a type of neural network under the subset of Recurrent Neural Networks. Unlike other neural networks which have forward feed, LSTM uses feedback connections. An LSTM unit has a cell, input gate, output gate and forget fate. LSTM can keep track of arbitrary long term correlations in input sequences. LSTM can also handle the vanishing gradient problem faced by RNN due to forget gate. The forget gates chooses whether to discard previous inputs. The input gate tries to learn new information. Whereas the output gate passes the new information from current timestamp to the next.LSTM can be used for classifying, forecasting and processing. Major uses of LSTM include handwriting and speech recognition, time series analysis and video games.

1.2 Linear Regression

Linear Regression is one of the most well-known algorithms in machine learning. It was initially developed in the field of statistics. It understands relationship between input and output variables in the form of a linear relationship by reducing the cost function error. In other words it attempts to find the relationship between two variables by fitting a linear equation to the observed. The simplest form of the regression equation with one dependent and one independent variable is y=a + b*x where y is expected value, a is the constant, b is called the regression coefficient and x is the true value. Major uses of Linear Regression include forecasting an event, trend forecasting, understanding the correlation between two values.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

1.3 Decision Trees

As the name suggests, Decision tree has a tree like model of decisions. Decision Tree algorithm is a sub-category of supervised algorithms. Simply put, Decision Trees tries to create a model from training data by understanding inherent decision rules. This algorithm uses multiple algorithms to create and split a node into sub-nodes. As the tree grows it decides what feature to choose and what functions to use for splitting. It can solve both regression and classification problems. It could be used for both categorical and continuous variables. It could be used in ecommerce, diagnosis of diseases and to detect frauds.

2 Methodology

2.1 LSTM

The input values are Adjusted Close stock data prices of AAPL. The data for X label had an array of 30 values with a label of 31st value.The data fed was first split in to train and test data with 4:1 ratio respectively. The model had 5 LSTM layers including dropout layer. After data preprocessing, the data was fed into LSTM model. Additionally, the loss parameter was mean squared error and optimizer used was Adam Optimizer. The layers were initially taken from research papers as standards and tested while changer the layer parameters to arrive at the best solution. After this, the number of epochs was also decided based on the epochs vs loss graph. Furthermore, batch size was notably of the order 2 which was also experimented to arrive at the best batch size model. The optimal epochs and parameters were used to create and fit the model. After which, the training data was used to predict values. This was compared using the RMSE performance metrics.

2.2 Linear Regression

The input values are the same as considered in LSTM. Whereas in this case the input X and Y values for fitting the model was training data and its respective label after a fixed number of days. In my case the fixed number of days was set to 30. In other words, the first data points label shall be 30th data point. In this was the model was resourced from scikit learn to fit X and Y values. After which training data was used to predict the expected values which was then compared with test values.

2.3 Decision Trees

The input values are the same as considered in LSTM and Linear Regression i.e. Adjusted Close stock prices. Whereas in this case the input X and Y values for fitting the model was training data and its respective label after a fixed number of days. In my case the fixed number of days was set to 30. In other words, the first data points label shall be 30th data point. In this was the model was resourced from scikit learn to fit X and Y values. After which training data was used to predict the expected values which was then compared with test values.

3 Data Collection and Preprocessing

Date Collection is fundamental module needed for any type of analysis. The data of stock prices under analysis in this paper is that of Apple. Inc with stock ticker abbreviation 'AAPL'. The data is extracted from Yahoo Finance under the historical prices section. The data collected 1st Jan 2019 to 31st December 2019. This data is limited to 2019 since COVID related events has caused an inconsistency in data patterns from 2020 and has been shown in figure 1.

3.1 LSTM

LSTM For data preprocessing for LSTM requires an input of three-dimensional arrays. The first step was to normalize or scale the data to range (0,1) since LSTM is very sensitive to the scale of data. This was followed by reshaping into 3-dimensional arrays as (samples, time steps, features).

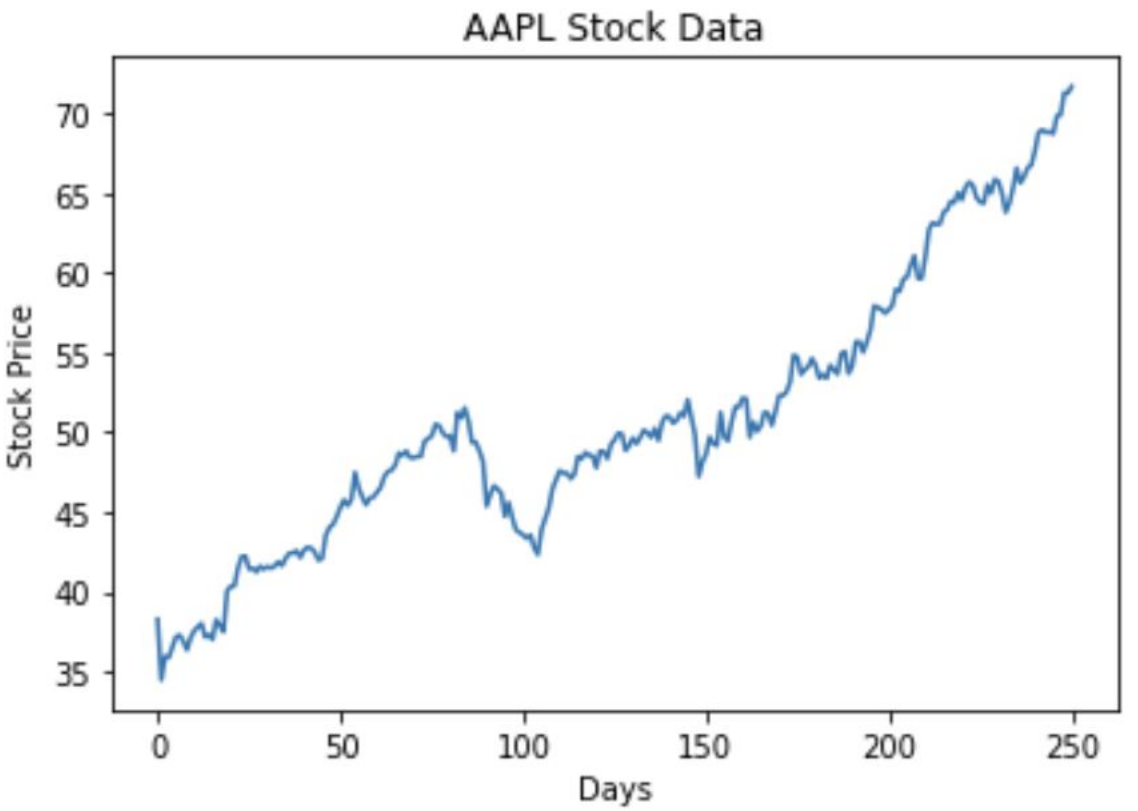

Figure 1: Orignial APPL Stock Price Data

3.1.1 Linear Regression and Decision Trees

Not much preprocessing of data was done since the data was already available in data frames. The target data was the 30th data point of the first data point which is the X variable in this case. The data was split in training and test in 4:1 ratio. The X and Y data points were finally fed into the Linear Regression model and Decision Trees models respectively.

4 Results

4.1 LSTM

The LSTM predictions with respect to original stock data is shown below in Fig 2. The variations do not seem to be much from the visual aspect which can also be proven through RMSE values. The RMSE value for LSTM test predictions is 1.23 and train predictions is 0.983. The training loss vs epochs for the LSTM model is shown below the epoch chosen for my case is 500.

The epoch values were selected based on Fig.3 to arrive at an optimum epoch value

4.2 Linear Regression

The results on APPL stock prices predictions are shown below. The variations from the graph seem to be way higher than the other two models. The RMSE value is also high and comes out to be 7.035

4.3 Decision Trees

The results of prediction are given below in the graph. The variations in the graph are not as much compared to Linear Regression. The prediction error is 4.037

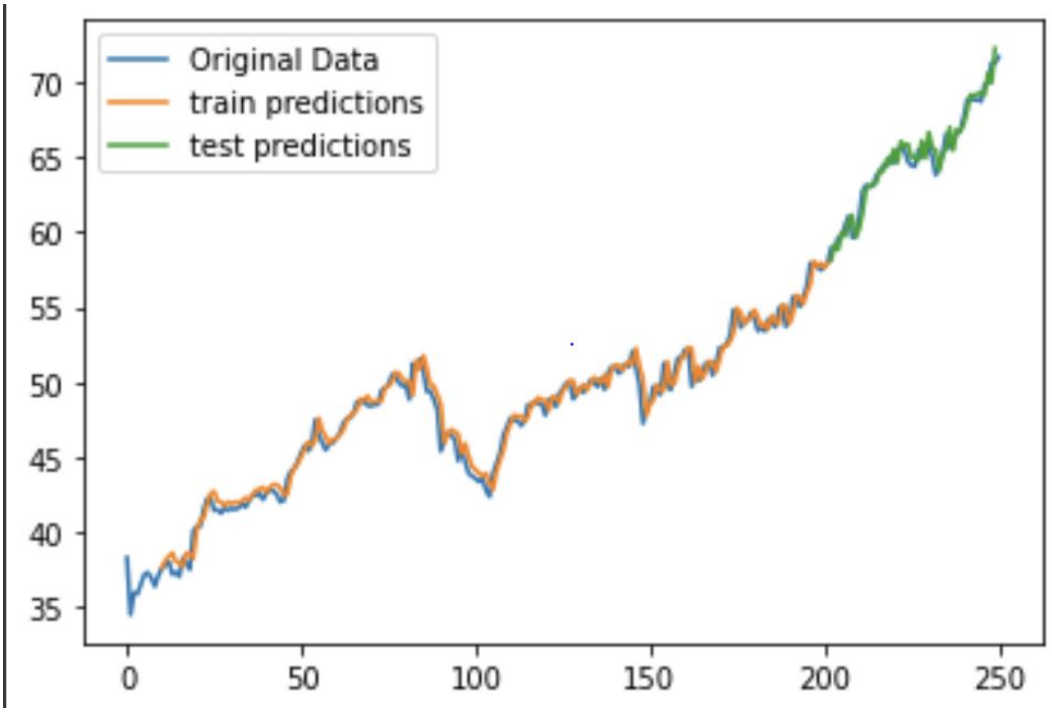

Figure 2: LSTM Prediction Results

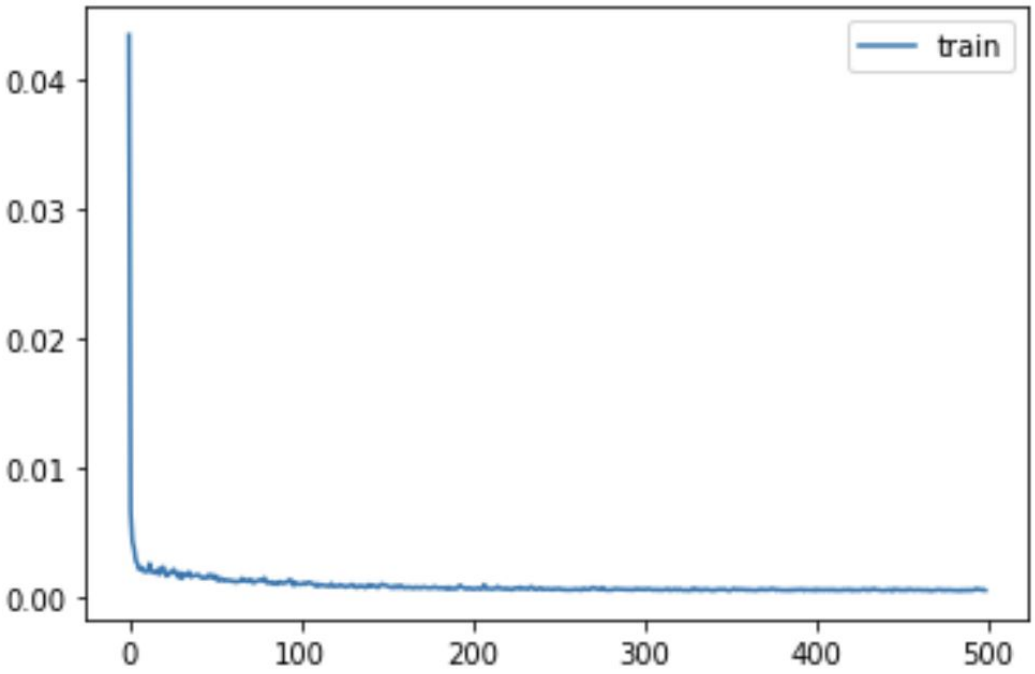

Figure 3: training loss

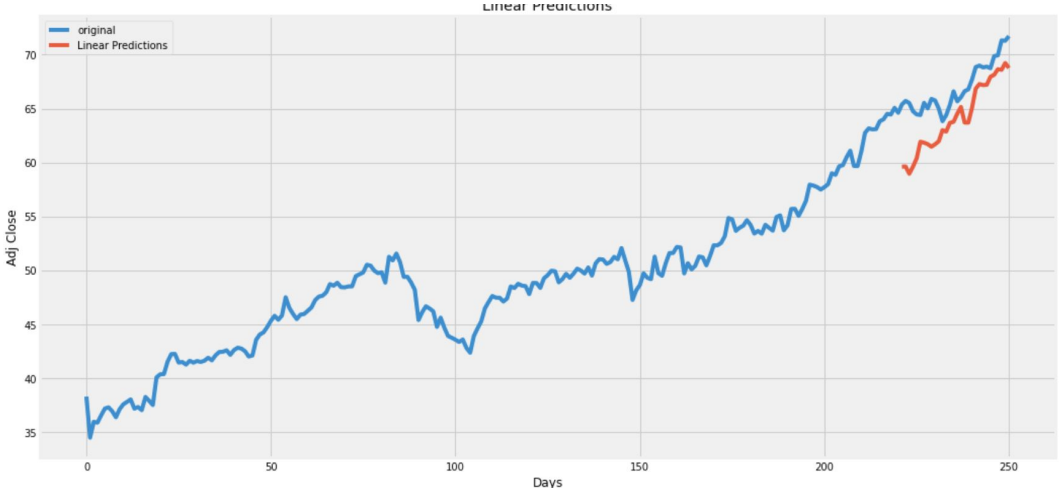

Figure 4: Linear Regression Prediction Results

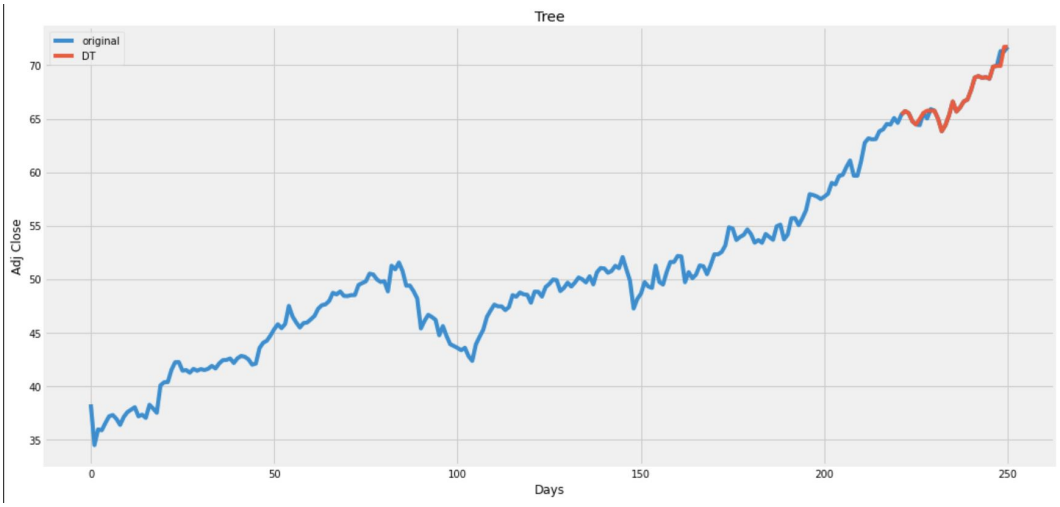

Figure 5: Decision Tree Prediction Results

5 Future Scope

Future scope of this is to try on more data and also different stock price data to understand the efficiency of the models in use such as LSTM, Decision Trees and Linear Regression. Additionally, further fine tuning of the models also needs to be experimented. Another effective approach is to include more data features such as volume, open price, and close stock price to get more efficient analysis.

6 Conclusions

The graph plots explain that LSTM performed better than Decision Trees followed by Linear Regression. The RMSE value also explains the same comparison which is that LSTM is more efficient in predicting values than Decision Trees followed followed by Linear Regression. A more detailed analysis with more features and optimum parameters will give more efficient results which can be used to predict future data making the job of stockbrokers a lot easier. Though these models need to be only used as a second perspective since the market events have lot of factors that could affect the stock prices of the future days. I truly believe that the true value of stock prices cannot be predicted. Since that would only mean predicting future events which is not possible. In other words,

if you cant predict future events then you cant predict stock prices since the stock prices are proven to be sensitive to the events around us.

7 Appendix

Fig1: Originial APPL Stock Price Data - Pg. 3
Fig2: LSTM Prediction Results - Pg. 4
Fig3: Training loss - Pg. 4
Fig4: Linear Regression Results - Pg. 5
Fig5: Decision Tree Results - Pg. 6

8 References

Singh, S.,Rehan S., Kumar v. (2022). Stock Price Prediction Using Linear Regression, LSTM and Decision Tree. *Easy Print*. Preprint no.7805. https://easychair.org/publications/preprint/rD6d

https://www.tensorflow.org/apidocs/python/tf/keras/layers/LSTM

Karim, Rezaul Alam, Md Hossain, Md. (2021). Stock Market Analysis Using Linear Regression and Decision Tree Regression. 1-6. 10.1109/eSmarTA52612.2021.9515762.