**STOCK MARKET PREDICTION USING MACHINE LEARNING ALGORITHMS**

**PYTHON PROJECT**

*S*ubmitted by

**MONICA JAWAHAR - 2017115619**

**APARNA S ANAND - 2017115535**

Submitted to the faculty of

**Department of Computer Science and Engineering**



**DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY**

COLLEGE OF ENGINEERING, GUINDY

CHENNAI - 600025

APRIL 2020

**ABSTRACT**

In Stock Market Prediction, the aim is to predict the future values of the financial stocks of a company. The recent trend in stock market prediction technologies is the use of machine learning which makes predictions based on the values of the current stock market indices by training on their previous values. Machine learning itself employs different models to make prediction easier and authentic. This project focuses on six methods based on Machine learning to predict stock values. Factors considered are open, close, low, high and turnover. The successful prediction of stock will be the great asset for the stock market institutions and will provide real-life solutions to the problems that stock investors face.

**INTRODUCTION**

The Stock market is basically an aggregation of various buyers and sellers of stock. The prediction is expected to be robust, accurate and efficient. The system must work according to the real life scenarios and should be suited to real world settings. Machine learning involves artificial intelligence which empowers the system to learn and improve from past experiences without being programmed time and again. The stock price movement over a long period of time usually develops a linear curve. People tend to buy those stocks whose prices are expected to rise in the near future. The uncertainty in the stock market refrain people from investing in stocks. Thus, there is a need to accurately predict the stock market which can be used in a real-life scenario. The methods used to predict the stock market includes a time series forecasting along with technical analysis, machine learning, modelling and predicting the variable stock market. The datasets of the stock market prediction model include details like the closing price opening price, the data and various other variables that are needed to predict the object variable which is the price in a given day.

**PROBLEM DEFINITION**

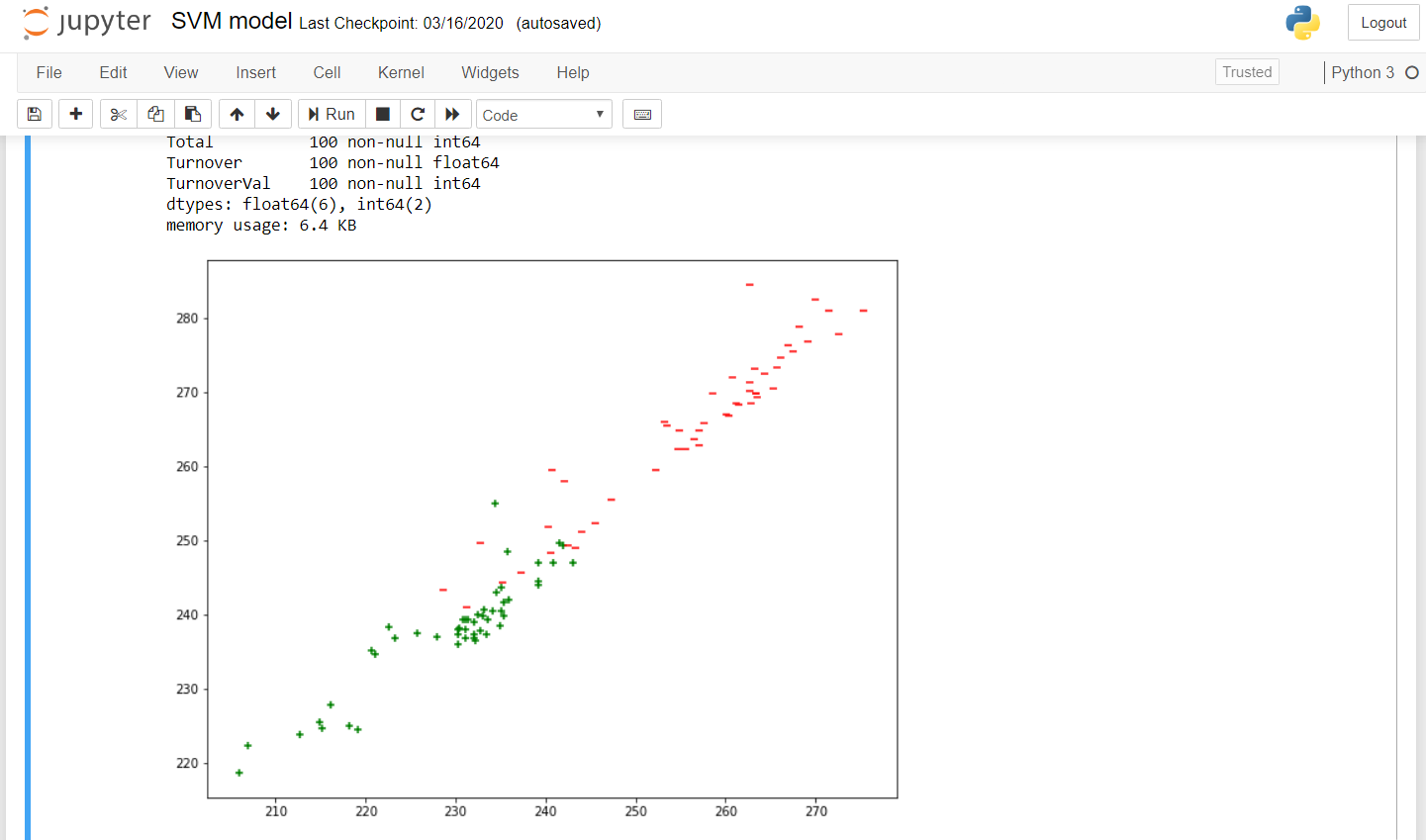
Stock market prediction is basically defined as trying to determine the stock value and offer a robust idea for the people to know and predict the market and the stock prices. It is generally presented using the dataset. Hence, we are contemplating towards the study of machine learning with datasets integration to predict the market and the stock trends. The problem with estimating the stock price will remain a problem if a better stock market prediction algorithm is not proposed. Predicting how the stock market will perform is quite difficult. Once, the right data is collected, it then can be used to train a machine and to generate a predictive result.

**METHODOLOGIES**

Regression analysis is used to determine the magnitude of relationships between variables as well as to model relationships between variables and for predictions based on the models. Regression performs operations on a dataset where the target values have been defined already. And the result can be extended by adding new information. The relations which regression establishes between predictor and target values can make a pattern. This pattern can be used on other datasets which their target values are not known. Therefore, the data needed for regression are 2 part, first section for defining model and the other for testing model. In this section we choose linear regression for our analysis. First, we divide the data into two parts of training and testing. Then we use the training section for starting analysis and defining the model.

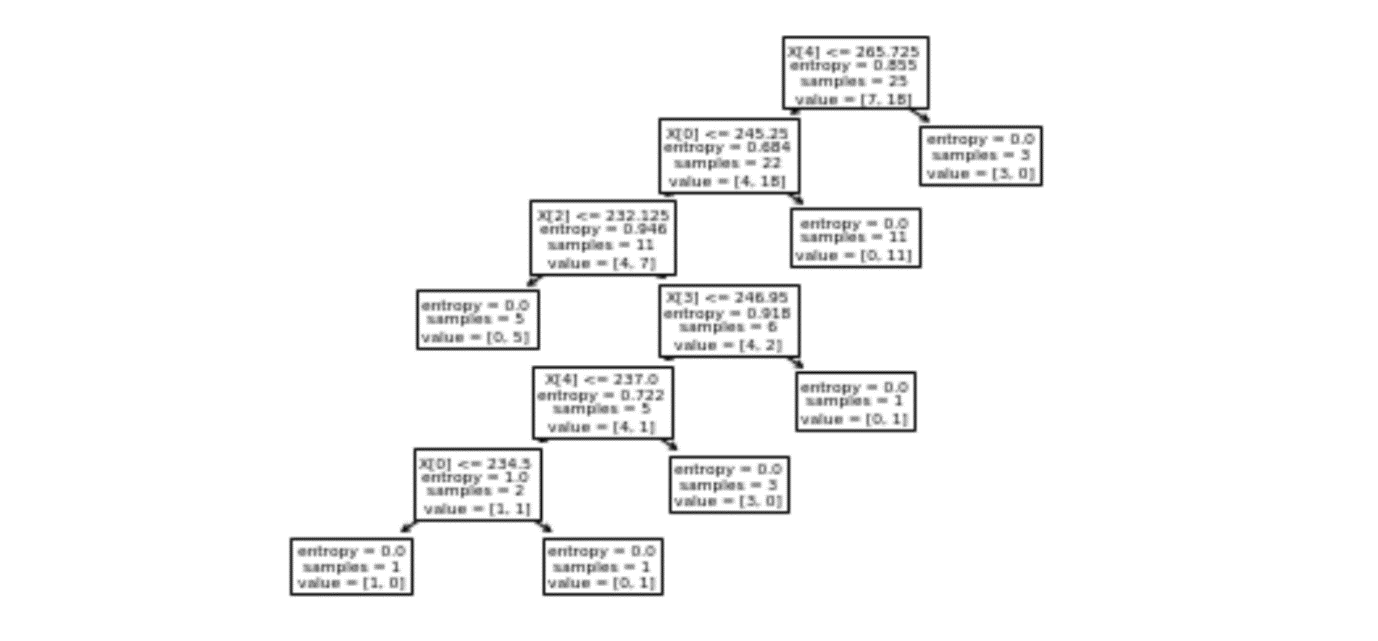
1. **SUPPORT VECTOR MACHINE(SVM)**

* The main task of the support machine algorithm is to identify an N-dimensional space that distinguishably categorizes the data points.
* Here, N stands for a number of features. Between two classes of data points, there can be multiple possible hyperplanes that can be chosen.
* The objective of this algorithm is to find a plane that has maximum margin. Maximizing margin refers to the distance between data points of both classes.
* The benefit associated with maximizing the margin is that it provides is that it provides some reinforcement so that future data points can be more easily classified.
* Decision boundaries that help classify data points are called hyperplanes.
* Based on the position of the data points relative to the hyperplane they are attributed to different classes.
* The dimension of the hyperplane relies on the number of attributes, if the number of attributes is two then the hyperplane is just a line, if the number of attributes is three then the hyperplane is two dimensional.
* Here by using features like Date, Open, Close, High, Low and Turnover SVM is concluded with input and output values.

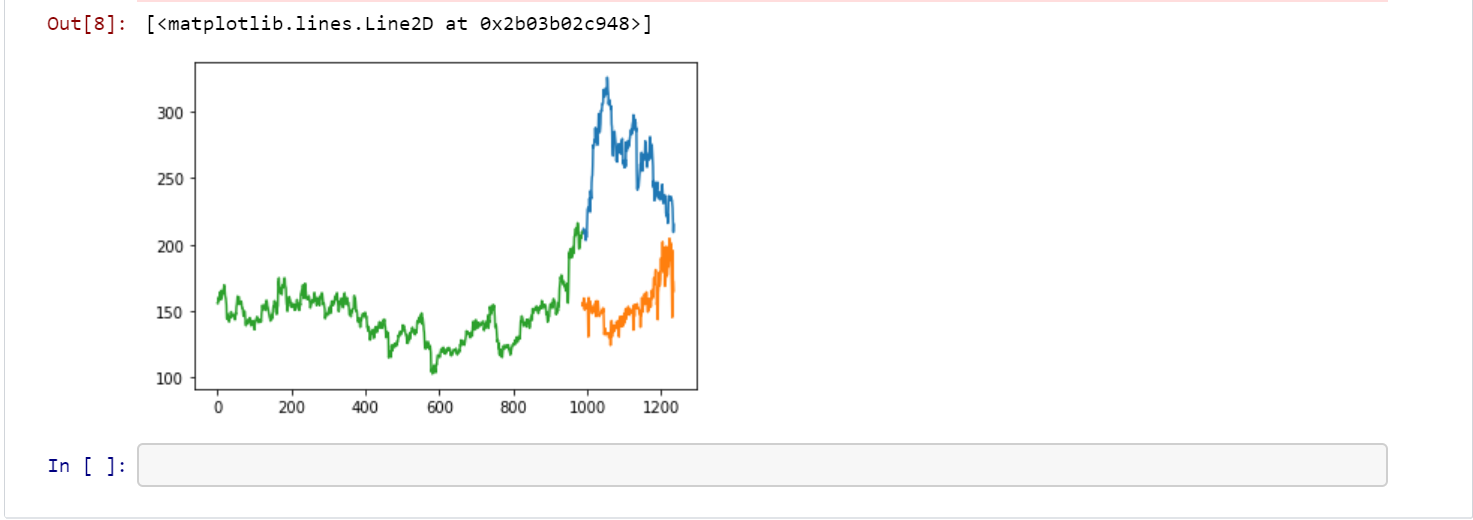
****

1. **DECISION TREE**

* Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
* The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested.
* Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.
* The core algorithm for building decision trees called ID3 by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. The ID3 algorithm can be used to construct a decision tree for regression by replacing Information Gain with Standard Deviation Reduction.

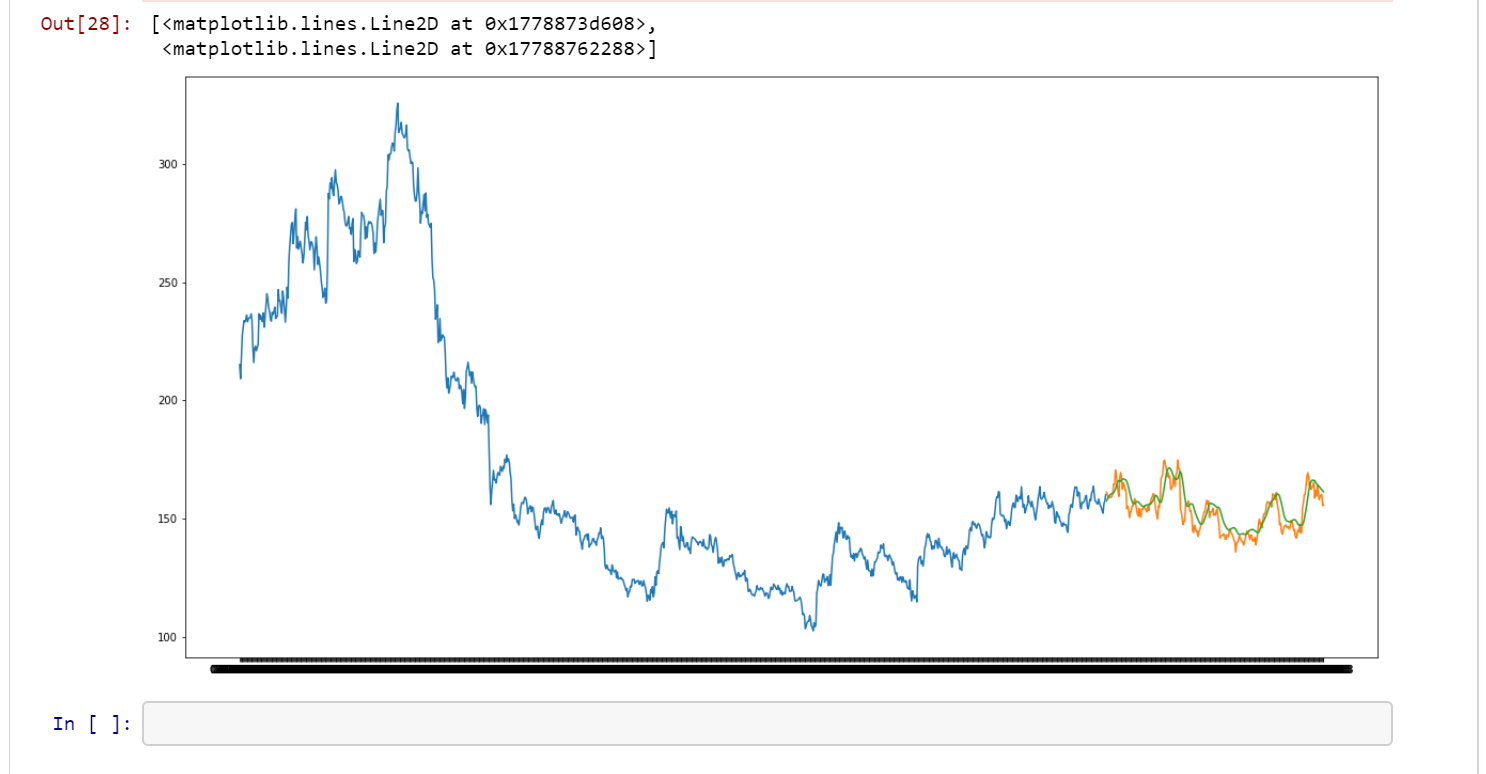
****

1. **K NEAREST NEIGHBOURS**

* The input consists of the k closest training examples in the [feature space](https://en.wikipedia.org/wiki/Feature_space). The output depends on whether k-NN is used for classification or regression:
* In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive [integer](https://en.wikipedia.org/wiki/Integer), typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
* In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.
* k-NN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood— calculating the distance between points on a graph.
* ****A simple implementation of k-NN regression is to calculate the average of the numerical target of the K nearest neighbours.  Another approach uses an inverse distance weighted average of the K nearest neighbours. k-NN regression uses the same distance functions as k-NN classification.

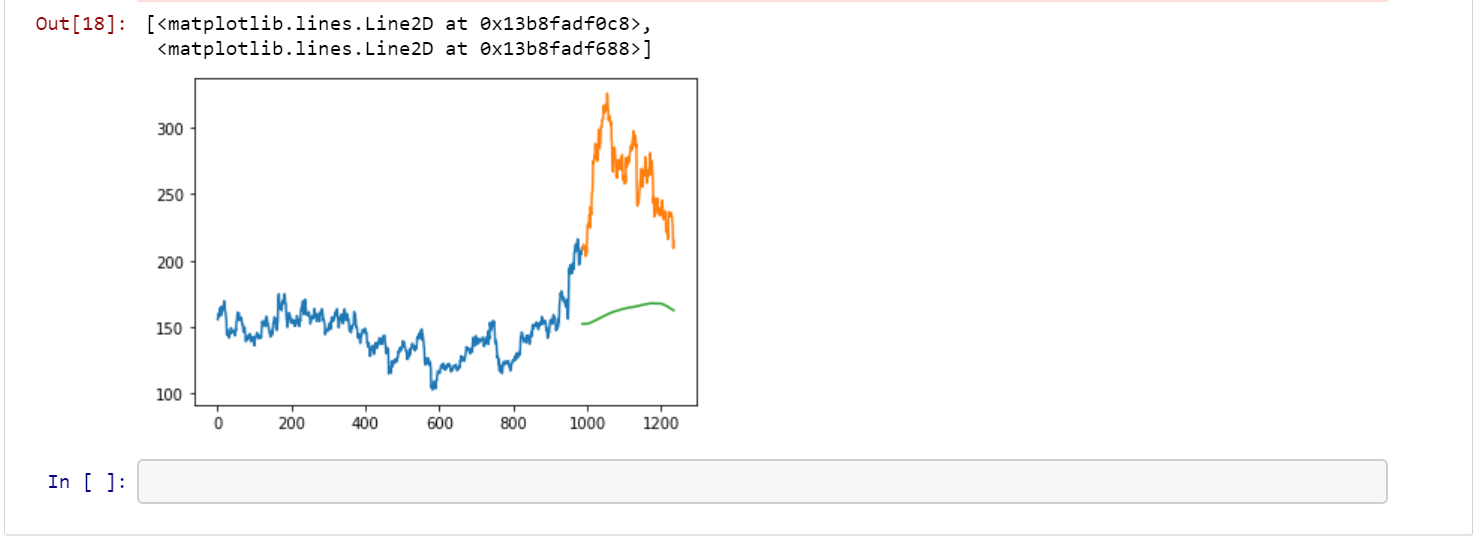
1. **LONG SHORT TERM MEMORY**

* LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series.
* LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs.
* The core concept of LSTM’s are the cell state, and its various gates. The cell state act as a transport highway that transfers relative information all the way down the sequence chain.
* The cell state, in theory, can carry relevant information throughout the processing of the sequence. So even information from the earlier time steps can make its way to later time steps, reducing the effects of short-term memory.
* As the cell state goes on its journey, information gets added or removed to the cell state via gates. The gates are different neural networks that decide which information is allowed on the cell state. The gates can learn what information is relevant to keep or forget during training.

****

1. **MOVING AVERAGE**

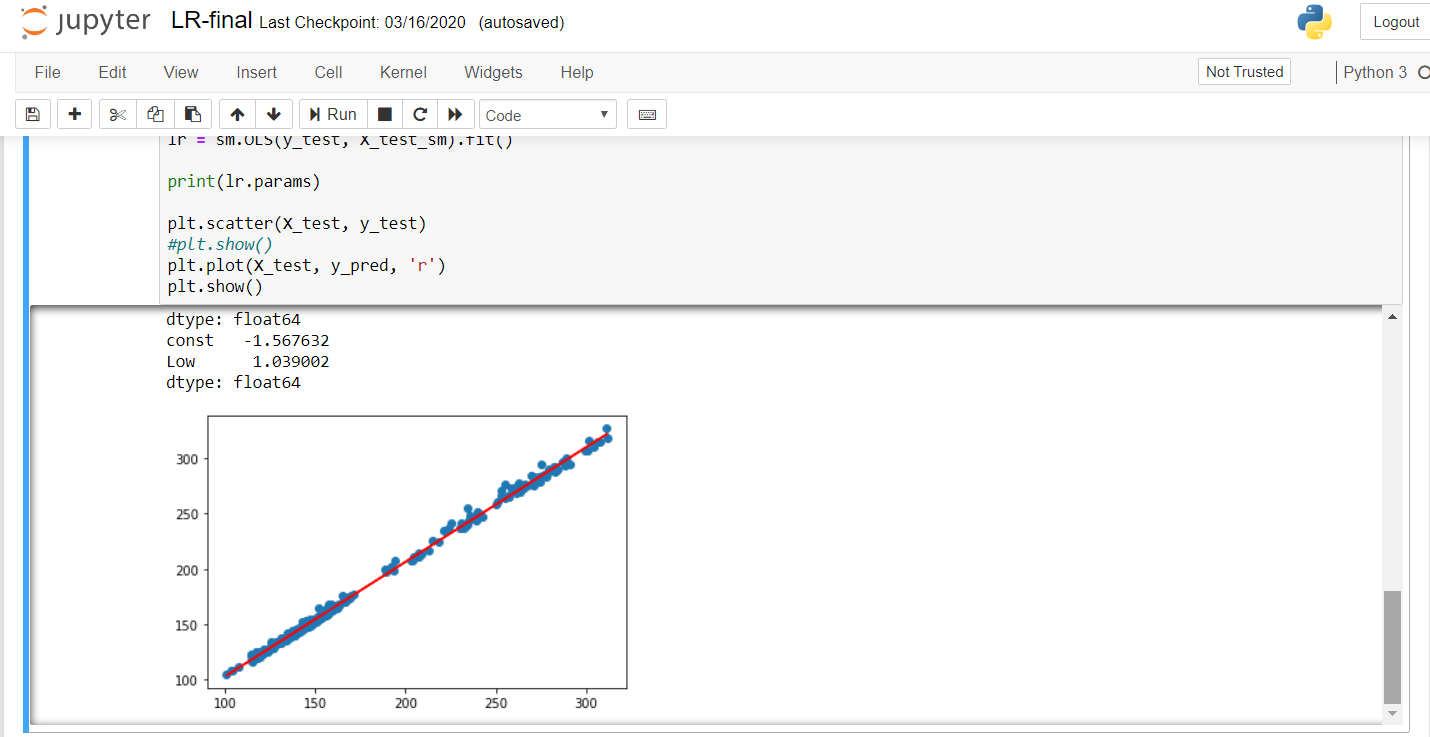
* Moving average smoothing is a naive and effective technique in time series forecasting. It can be used for data preparation, feature engineering, and even directly for making predictions.
* Calculating a moving average involves creating a new series where the values are comprised of the average of raw observations in the original time series.
* In the moving average method, the predicted value will be the mean of the previous N values.
* In our context, this means we set the current adjusted closing price as the mean of the adjusted closing price of the previous N days. The hyper parameter N needs to be tuned.

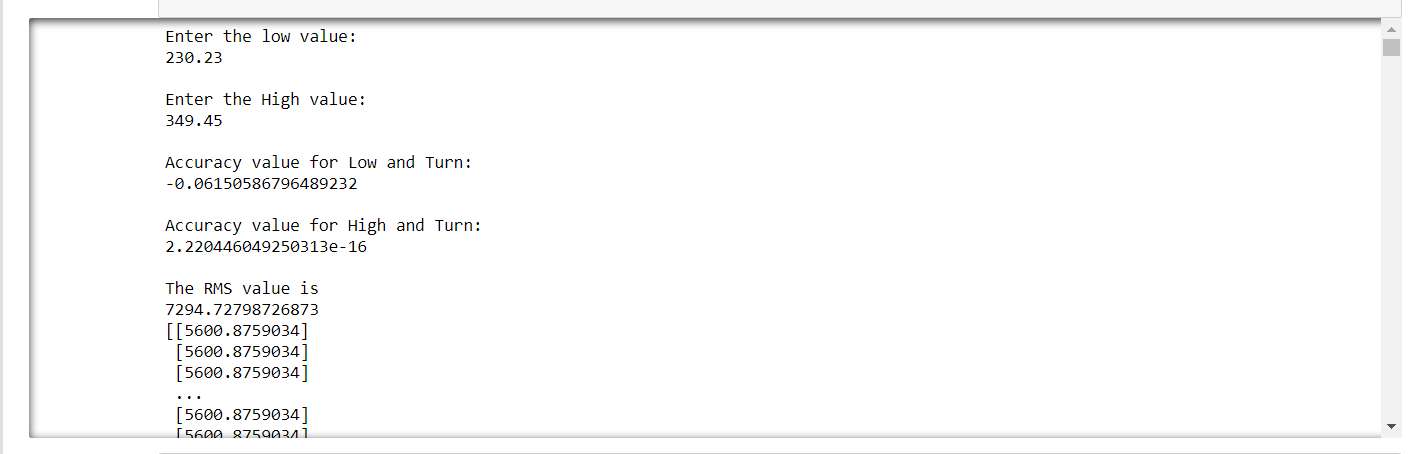


1. **LINEAR REGRESSION**

* **Linear regression** is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).
* Simple linear regression or multiple linear is applicable when this relationship is assumed to be linear.
* However, a number of non-linear techniques could be used to obtain a more accurate regression if the relationship between variables is not linear in parameters.
* If the goal is prediction, linear regression can be used to fit a predictive model to an observed data set of y and X values.
* After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted.

The accuracy for this method is the highest among the 6 methods we’ve tried. Hence, we prefer Linear Regression for prediction.





**SYSTEM ARCHITECTURE**

**PREDICTED RESULT**

**TRAINED DATA RESULT**

**TESTING DATA**

**TRAINING DATA**

**FEATURE EXTRACTION**

**RAW DATA**

**CONCLUSION**

By measuring the accuracy of the different algorithms, we found that the most suitable algorithm for predicting the market price of a stock based on various data points from the historical data is the linear regression algorithm.

The algorithm will be a great asset for brokers and investors for investing money in the stock market since it is trained on a huge collection of historical data and has been chosen after being tested on a sample data.

The project demonstrates the machine learning model to predict the stock value with more accuracy as compared to previously implemented machine learning models.