**Stock Market Prediction**

**Team Members**

Lina Khalid Alismail

Maryam Arif Aloudah Alfadhli

**12th of November**

Table of Contents

[List of Tables iii](#_Toc119186955)

[List of Figures iv](#_Toc119186956)

[1 Introduction 1](#_Toc119186957)

[1.1 Problem Statement 1](#_Toc119186958)

[2 Methodology 1](#_Toc119186959)

[2.1 Dataset Description 2](#_Toc119186960)

[2.2 Models 4](#_Toc119186961)

[*2.2.1* Linear Regression Model 4](#_Toc119186962)

[*2.2.2* LSTM Model 6](#_Toc119186963)

[2.3 Split Test and Train Data 7](#_Toc119186964)

[2.4 Training The Model 7](#_Toc119186965)

[2.5 Predict Tomorrow Price 8](#_Toc119186966)

[2.6 Performance Evaluation 8](#_Toc119186967)

[3 Results and Discussion 9](#_Toc119186968)

[3.1 Linear Regression Result 9](#_Toc119186969)

[3.2 LSTM Model Result 11](#_Toc119186970)

[3.3 Discussion 12](#_Toc119186971)

[References 13](#_Toc119186972)

# List of Tables

[Table 1 Dataset Description 2](#_Toc119187247)

[Table 2 Correlation table 3](#_Toc119187248)

[Table 3 LR performance metrics table 10](#_Toc119187249)

# List of Figures

[Figure 1 Google Stock Dataset Sorted 2](#_Toc119187268)

[Figure 2 Stocks info 3](#_Toc119187269)

[Figure 3 Correlation matrix 3](#_Toc119187270)

[Figure 4 Stock price (open, close) over the years. 4](#_Toc119187271)

[Figure 5 Regression line 5](#_Toc119187272)

[Figure 6 our model regression line 5](#_Toc119187273)

[Figure 7 Correlation Between Columns 5](#_Toc119187274)

[Figure 8 LSTM Model 6](#_Toc119187275)

[Figure 9 Splitting the dataset 7](#_Toc119187276)

[Figure 10 LSTM model fitting 8](#_Toc119187277)

[Figure 11 Root mean square error 8](#_Toc119187278)

[Figure 12 R2 score 9](#_Toc119187279)

[Figure 13 Mean squared error 9](#_Toc119187280)

[Figure 14 Mean Absolute Error 9](#_Toc119187281)

[Figure 15 Close vs prediction LR table 9](#_Toc119187282)

[Figure 16 Close vs prediction LR graph 10](#_Toc119187283)

[Figure 17 R2 score of lr 10](#_Toc119187284)

[Figure 18 Preformance Metrics of LR 10](#_Toc119187285)

[Figure 19 Close vs prediction LSTM table 11](#_Toc119187286)

[Figure 21 Closing prediction LSTM graph 11](#_Toc119187287)

[Figure 20 Preformance Metrics of LSTM 11](#_Toc119187288)

# Introduction

Machine learning is widely used in tasks that predict the future with high speed and accurate results. Before machine learning, scientists were manually calculating predictions using mathematical and observation methods of the factors and trends affect the prediction result. Machine learning algorithms played a significant role on changing the way of predicating the future data. It has been approved with many applications that machine learning algorithms can achieve high accurate predictions with great efficiency.

## Problem Statement

The stock market is notorious for its volatility, dynamism, and non-linearity. Accurate stock market price prediction is extremely challenging due to the verity of factors that relevant to the stock market price such as politics, global economic condition, unexpected events, and company’s financial performance.

The stock market prediction is time series forecasting problem. Those kind of problems understands the historical data in order to identify patterns then it forecast the future. The dataset of stock market contains large number of values to search through in order to find patterns, which could be extremally hard to do manually. For that reason, machine learning algorithms is important and can be used to process those large number of values and dealing with unstructured data.

One of the greatest advantages of predicting the future prices of stock market, it could minimize the risk of losing money by trading in the market and lead to huge profits if it predicts the correct price.

In this project, model to predict the closing prices a of stock market were built by applying supervised machine learning algorithms on the stock data. This paper explains the methodology used to make the prediction of the stock market closing price including the chosen dataset, analysis of the dataset and the models built to train and predict the prices. It also discusses the result of the models and the accuracy of its prediction.

# Methodology

Stock market is a complicated problem because the prices changing according to many factors. When plotting stock market prices, the figure does not have a certain patten and it is stochastic. It seems there is no relation between the data yet by using the machine learning algorithms, it is possible that the model could be trained to learn the relationship from the historical data to predict the next day prices. This project focuses on applying two of the most promising models to predict close prices of stock market, Linear Regression model and LSTM model.

## Dataset Description

On this project we worked on Google stocks dataset, It has a collection of data from 2004-09-01 to 2022-11-01. [1]

On the website attached, the system can be tested over detest taken from Yahoo Finance. the user can search for any valid stock. The prices can be predict using the algorithms described below.

The dataset has the following attributes:

|  |  |
| --- | --- |
| Open | Opening price of the stock for each day, it is the same price when trading starts |
| Close | Close price of the stock for each day, it is the same price when trading is closed/finished for the day |
| Low | The lowest value of the stock for a given day |
| High | The highest value of the stock for a given day |
| Adj Close | The closing price after paying dividends to the investors |
| Volume | The total number of shares traded/bought on a day |

Table Dataset Description

The main goal of our project is to predict stock market prices to make safer choices of selling and buying. The dataset has 6 columns or features which are open, close, low, high, adj close and volume. We must understand the relationships among those columns in order to make better predictions.

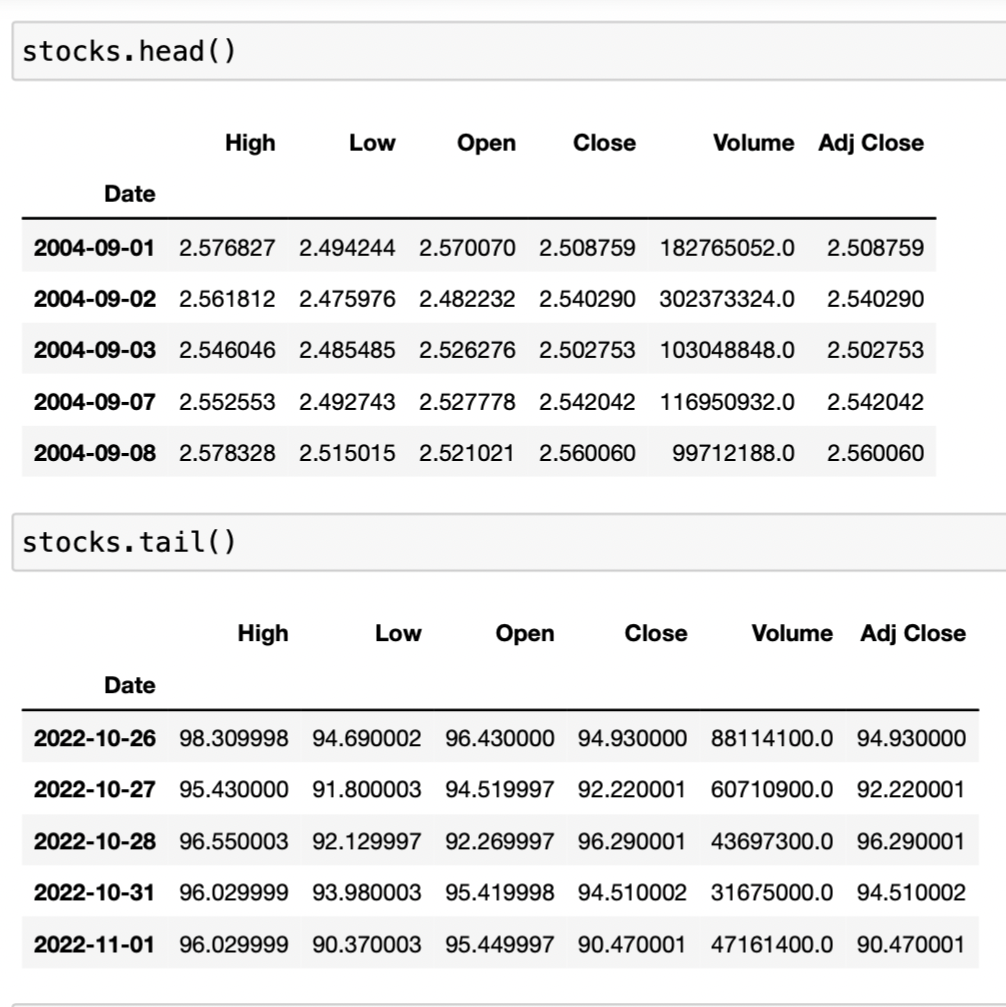


Figure Google Stock Dataset Sorted

The dataset includes data of over 18 years, starting from 2004-09-01 to 2022-11-01.

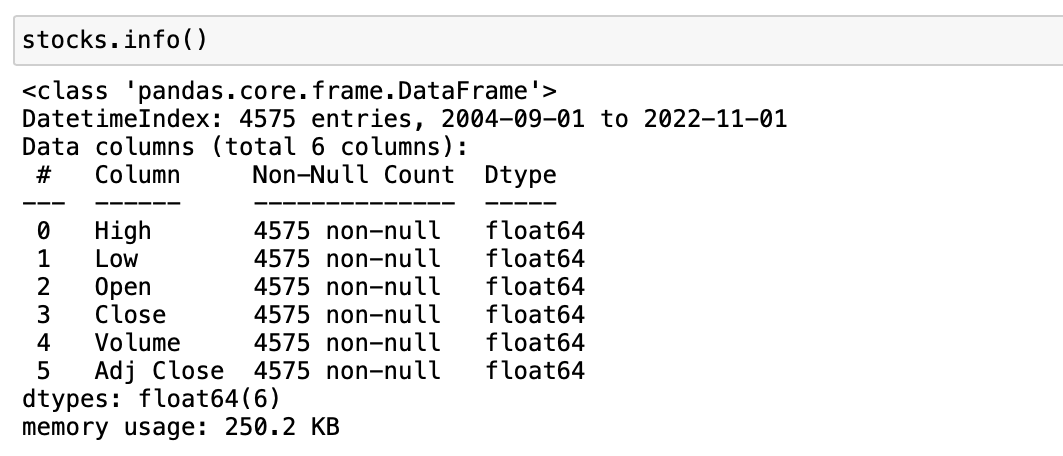
****

Figure Stocks info

The dataset set used has 4575 with no NaN / Null values, and it attributes has the type Float64.

We’ve re-arranged the columns to draw the correlation matrix in a cleaner way.

The correlating matrix will help us find the relationship among the attributes.

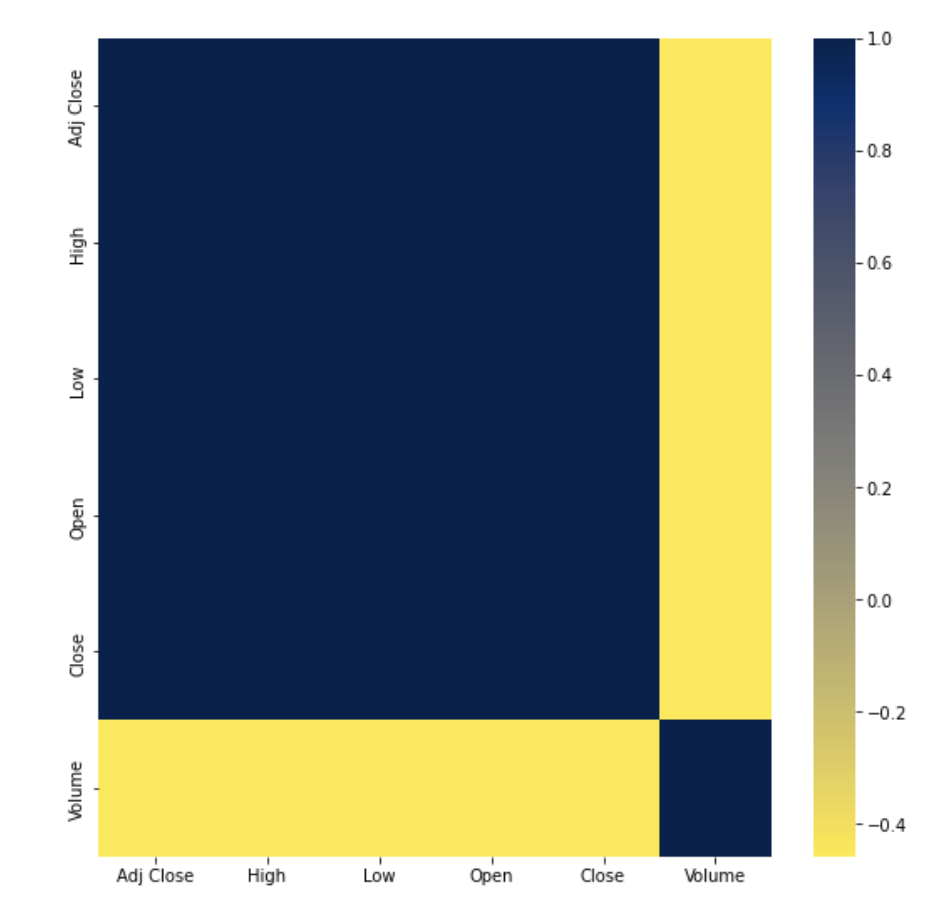


Figure Correlation matrix

As seen by the matrix, the correlation among the columns can be described as the following table:

|  |  |  |
| --- | --- | --- |
| Columns | High correlation | Low correlation |
| Open | Close, Low, High, Adj Close | Volume |
| Close | Open, Low, High, Adj Close | Volume |
| Low | Close, High, Adj Close, Open | Volume |
| High | Close, Low, Adj Close, Open | Volume |
| Adj Close | Close, High, Low, Open | Volume |
| Volume |  | Close, High, Low, Open , Adj Close |

Table Correlation table



Figure Stock price (open, close) over the years.

As seen in the figure above, **Google** stock prices of close and open columns are very similar. The graph shows the huge jump in prices over the last 4 years.

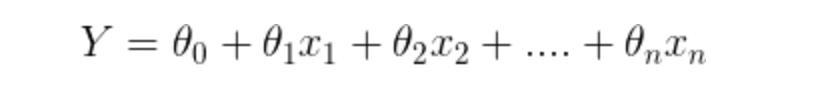
## Models

### Linear Regression Model

It is a model that describes the relationship between a 1 or more dependent(y) and independent variables, X. a linear model assumes a linear relationship between the input x and the single output y variables. By using linear combination of the input variables x, we can calculate y.

We can refer to the dependent variable as a response variable. We can also refer to independent variables as explanatory variables.

**We can represent** Linear Regression by the following equation:



* Y : predicted value
* θ₀ : the bias term
* θ₁,…,θₙ : the parameters of the model
* x₁, x₂,…,xₙ : the feature values.

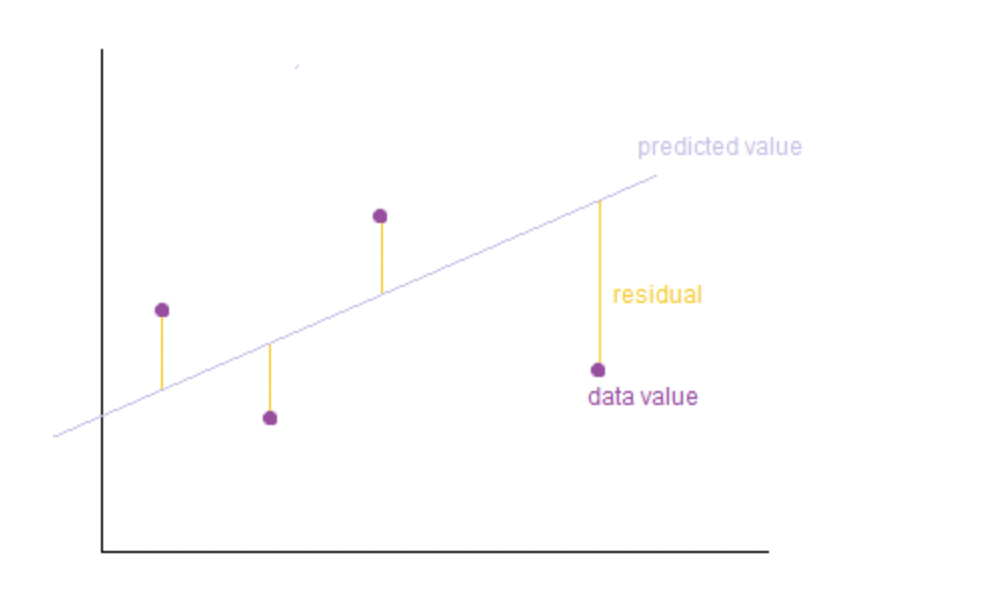


Figure Regression line

Regression line is the best fit line where the errors between the predicted values and the observed values are minimum. Errors can be called residuals, which can be visualized by lines which represent the distance between data value to the regression line. [2]

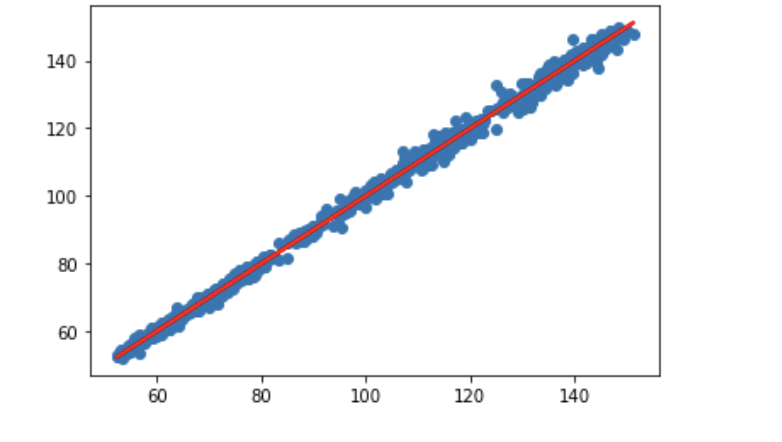


Figure our model regression line

The image shows plotting of the regression line over the test data which are Open and Close columns, Data values are not far from the regression line which minimizes the errors or residuals.

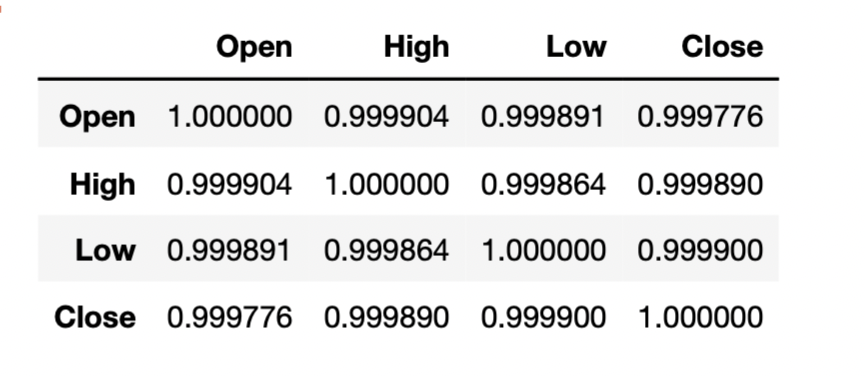


Figure Correlation Between Columns

We’ve used linear regression model because it’s great for predicting dependent variable using independent variable. In this model our variables are ‘Open’ and ‘Close’.

### LSTM Model

Long Short Term Memory known as LSTM model. It is type of Recurrent Neural Networks (RNN). LSTM mechanism involve a long-term dependency that can predict value by finding the relationship between the old and the current information. The LSTM model improve the efficiency in the problems that depend on storing data for a long period of time. It is widely use in natural language processing and time series forcasting.

The model was chosen in stock market prediction because the prices depend on many factors that are influenced by long term of historical stock data. By using this model, it can preserve old data to make more accurate prediction

The LSTM structure contain a cell that has input gate, output gate and a forget gate, the cell works as a memory to remember the data for long term and each gate responsible to regulate those data.

This project uses sequential model which has multiple layer above each other. In each LSTM cell the time steps will be 60 days. The model also uses a dropout with 20% precent of the total number of neurons in the network in order to prevent overfitting problem and decrease the training process time. The last layers are dense layer where all the neurons connect to produce the final output.

Diagram

Description automatically generated

Figure LSTM Model

The chosen evaluation metric for the model is mean squared error. It monitors the model's performance and calculate the accuracy of the model.

Adaptive Moment Estimation known as Adam is an optimizer used to calculate the learning rate for the parameters. it optimizes noisy problems that involve many parameters and handles stochastic sparse gradients. In the model is used to optimize weights in each layer to give better results and minimize the error.

## Split Test and Train Data

In splitting the data is important to note, the data should be arranged according to the date time because the date effect the prediction result.

To train the LSTM model, the model will concentrate only in one feature in the stock dataset, the closing prices. The main goal is to predict the closing price of next day based on the historical values of the closing price.

Splitting the date depends on the time steps chosen. After trying multiple time steps, 60 days give the best result. The first 60 days work as a base for the model.

The figure shows clearly how the data was splitting in a way that make is a supervised learning problem that has features and labels

A picture containing diagram

Description automatically generated

Figure Splitting the dataset

On the other hand, to train the linear regression model, the model will use opening prices data as a feature to predict the closing price

## Training The Model

Important note before building the LSTM model the training data should be stored in a three-dimensional array. In this problem, it has only one feature and each 60 data are gathered in a one batch. The shape of the data will be

**(number of rows in the stock data , 60 , 1)**

After building the model, it trained using the training sets and as shown in the figure below, in each time the model fits the training set, it calculates the mean squared error and it is relatively got smaller in each iteration.

**Table

Description automatically generated**

Figure LSTM model fitting

Note that it is recommended to use smaller epochs numbers in order to avoid overfitting problem.

## Predict Tomorrow Price

In the project we tried to predict the value of closing price for tomorrow. To achieve this goal, the model used the previous 60 days and added one day on the testing data.

I believe in order to predict the next month closing price, the model can use the same method to predict day by day and in addition to that it stores all the predicted values to use it to predict the future closing price. It could give an accurate result based on the evaluation of the model’s accuracy prediction.

## Performance Evaluation

To evaluate the performance of the models, we’ve used common performance measures that helped us assess the predictions accuracy. We have used Root mean square error (RMSE), Coefficient of determination (R2 score) and the mean squared error (MSE). They helped us identify weather the predictions were close to the original values or not.

Root mean square error (RMSE) is the square root of the average of sum of the squares of residuals, The below equation demonstrates the RMSE.

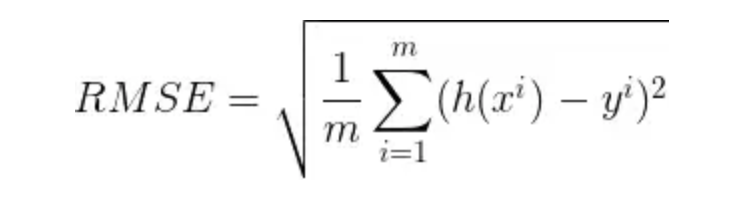


Figure Root mean square error

Coefficient of determination (R2 score) explains how much least square regression can reduce the total variance of the response/dependent variable

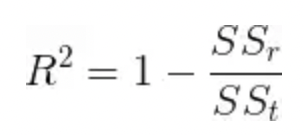


Figure R2 score

The mean squared error (MSE) The Average the square of the difference between original and predicted values.

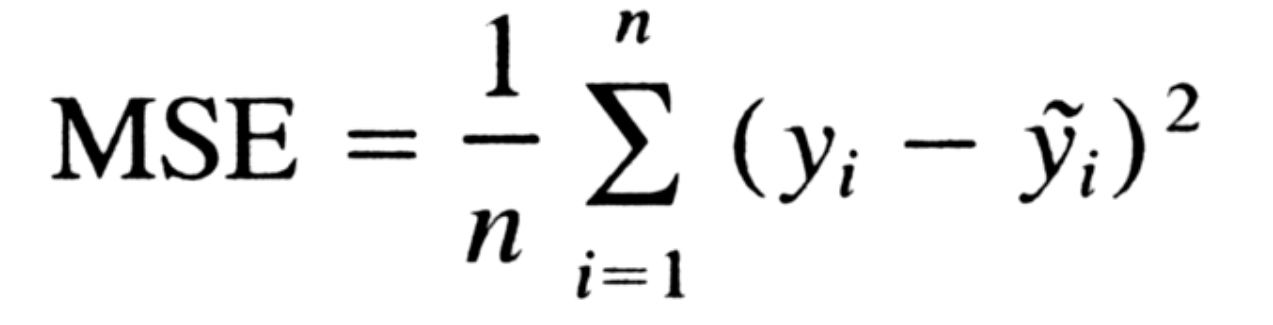


Figure Mean squared error

Mean Absolute Error (MEA) Finds the average of all absolute errors

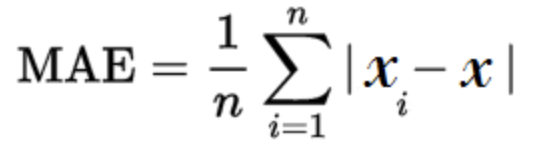


Figure Mean Absolute Error

# Results and Discussion

## Linear Regression Result

A linear regression model was fit to the graph as seen above in Figure 6. The values of RMSE, MSE, MAE, and Coefficient of determination were used to evaluate the performance of the model.

1. The prediction table:

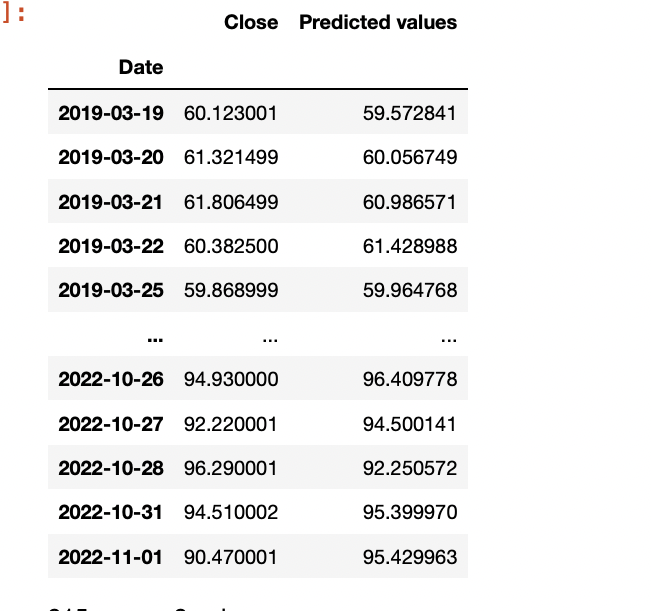


Figure Close vs prediction LR table

1. The prediction graph:



Figure Close vs prediction LR graph

1. The performance evaluation metrics:

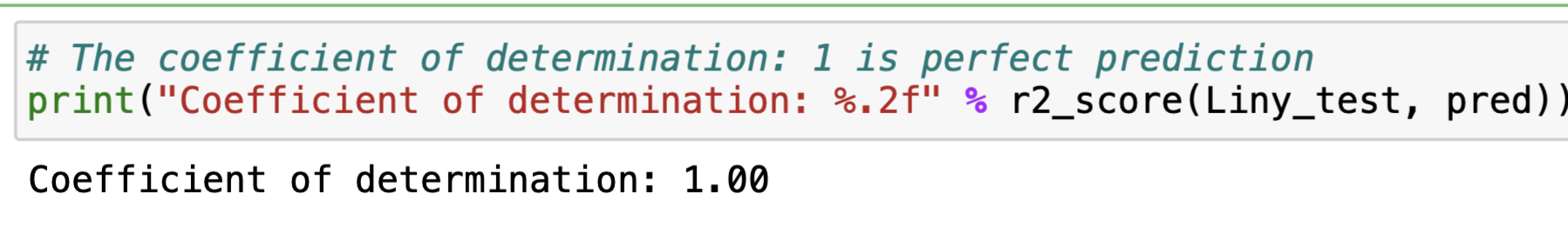
****

Figure R2 score of lr

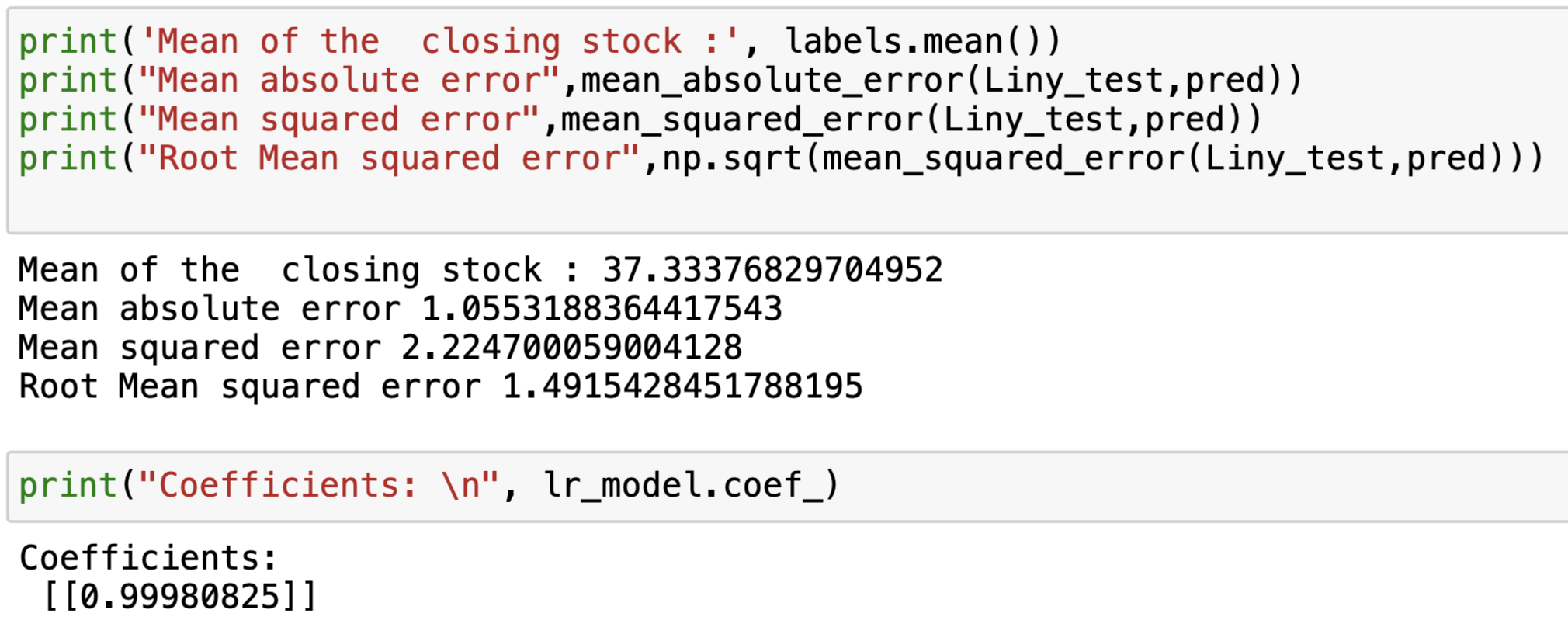


Figure Preformance Metrics of LR

Linear regression has the following performance metrics:

|  |  |
| --- | --- |
| **Mean of the Closing stock** | 37.33376829704952 |
| **Mean absolute error** | 1.0553188364417543 |
| **Mean squared error** | 2.224700059004128 |
| **Root Mean squared error** | 1.4915428451788195 |

Table LR performance metrics table

## LSTM Model Result

1. The prediction table:

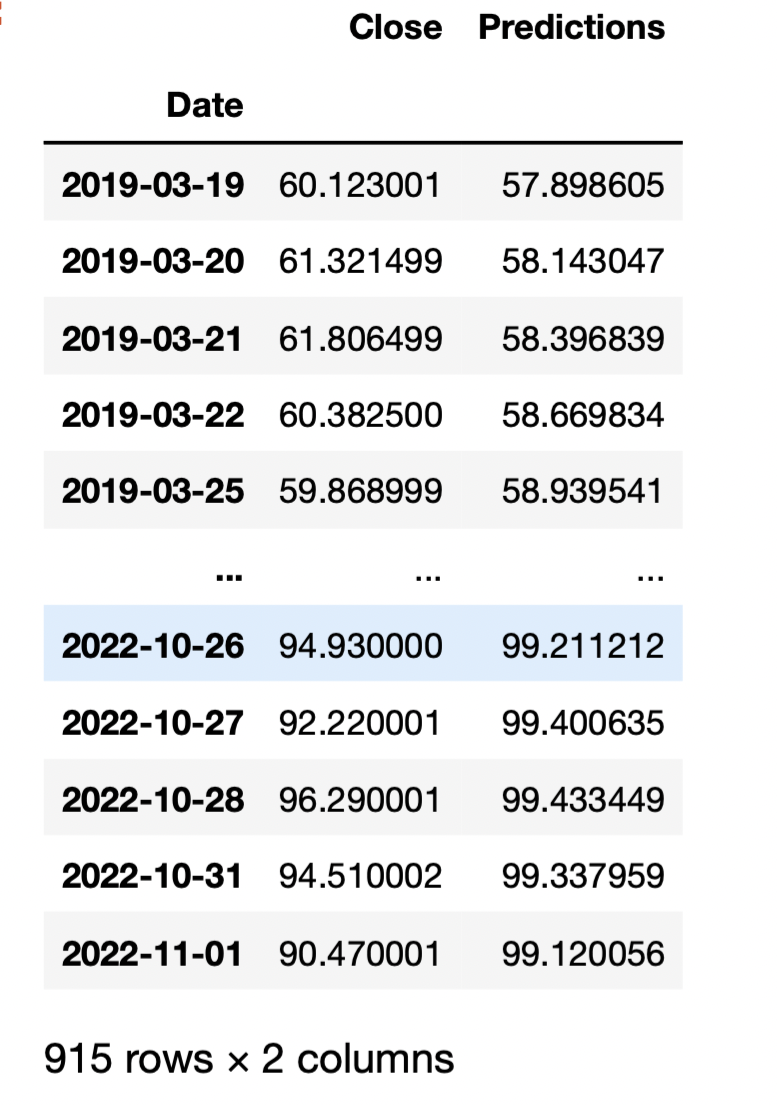


Figure Close vs prediction LSTM table

1. The prediction graph:



Figure Closing prediction LSTM graph

1. The performance evaluation metrics:

Graphical user interface, text, application

Description automatically generated

Figure Preformance Metrics of LSTM

## Discussion

Stock market is a challenging area to for price prediction due to non-linearity and volatility. It has many factors that can affect the price of the stock and the change happens almost every second. Many factors can be taken to predict the next price but this project we’ve used numerical values from the given dataset only.

As seen by the results for this attempt, linear regression model gave more accurate result than LSTM. Since Linear regression captures linear relations between the response and predictors it gave good results, The linear regression model didn’t include any date dependent forecasting attributes (Not dependent on date column) which helped with the result. But otherwise, the LSTM would perform best, since it is better with pattern recognition, which serves best in stock market prediction where the order of input or date is the main factor. For further enhancement in the LSTM model, it is advised to change the layers and try multiple different arrangement of LSTM layers.

In our project we have used both LSTM and linear regression to predict the closing price. The models of those techniques include previous values provided by the models that helps the accuracy of the results. The stock market prices are dependent on many other factors in life other than what was provided in the dataset which could affect the result significantly such as natural disasters or economic crises. It is recommended not to completely depend upon simple ML prediction model to buy and sell stocks but to use them to support buying or selling decisions.

# References

|  |  |
| --- | --- |
| [1] | Y. Finance, "Yahoo Finance," [Online]. Available: https://finance.yahoo.com. |
| [2] | A. Agarwal, "Linear Regression using Python," 2018. [Online]. Available: https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2. |
| [3] | I. Parmar, "Stock Market Prediction Using Machine Learning," National Institute Of Technology, Hamirpur, INDIA. |
| [4] | M. U. Ghan, "Stock Market Prediction Using Machine Learning(ML)Algorithms," Advances in Distributed Computing and Artificial Intelligence Journal, Faisalabad. |
| [5] | T. D. Smedt, "Evaluating and understanding text-based stock price prediction models," Belgium. |