Closing Stock Price Prediction for Walt Disney using Regression

A Project Report

Presented To

Professor Gayathri Namasivayam

Department of Computer Science

San José State University

Class CS 256

By

Gayatri Hungund

Riti Gupta

May 2019

**Abstract**

The investment and financial sector have changed the way finances are dealt daily due to which, large number of people invest in stock market and contribute to the profit or loss of a given organization. Stock market has gained importance over the past few decades and is one of the most important factors contributing to the growth of an economy. Millions of dollars are invested in stock trading daily which can lead to sharp rise or decline in financial status of an organization. After a company attains financial stability, it generally invests in the stock market to increase the revenue by considering various factors including demand and supply. Understanding the trend of financial data becomes an important task and can also help the investors to follow the trend to predict rise and fall of a stock. Statisticians use many techniques to predict the closing price of a stock for a given time frame and help the investors make better decision before investing money. The project uses four regression models to compute the closing price of one of the popular stocks which is the Walt Disney stock at a given instant of time. Data analysis and predictions obtained using machine learning models thus prove to be a smarter and quicker way to analyze financial data.

**Keywords – Stock Market, Finances, Closing Price, Machine Learning, Regression, Data Analysis**

**I. INTRODUCTION**

As the technology grows day by day, all the sectors have started adopting the use of growing smart technologies to solve mighty challenges in their fields in a faster and more efficient way. Machine learning algorithms help to train a machine to simulate human behavior and handle data more accurately as per the given training module. Machine learning techniques help in accelerating complex mathematical calculations and simplify them with minimal human interaction. Every day, millions of transactions are processed, and huge amount of data is being generated that needs to be analyzed to find interesting patterns and help in making data analysis smarter. Artificial intelligence automates the process of data analysis and real time result generation for a smarter tomorrow.

Financial data is generated by New York Stock Exchange every second and market analysis needs to be carried out to help the investors make profitable investment. Trading transactions that involve buying and selling of shares can decide the rise and fall of an economy of a country [1]. Stock market investment is a new way for the companies to invest when the company grows. The stock investment can help in increasing the annual revenue generated by the company. Also, it is easy to carry out trading transactions using the stock market and the shares brought at any instant of time, being a liquid asset can be sold whenever the investor wants to sell them to liquidate it to cash [2].

Buying a share or stock from a company makes an investor a shareholder of percentage defined for profit made by the organization [2]. Price of a stock is decided based on various factors in which revenue of a company is very important factor affecting the price of stock. Other factors affecting the rate of stock is business, company clients, market value, profit margin and so on. The trading starts early morning at NYSE based on which the prices of stock fluctuate every minute. Investment in the stock market can be risky, as the investor can lose their money if stock price falls. Thus, if the price of stock goes high and the stock is sold to another investor, the investment yields profit [3].

This project this aims to predict the closing price of stock (target variable) using artificial intelligence and help the investors to get a probable idea about the closing price of stock at

given instance of time. This will help the investors to decide if buying a particular stock will be profitable or not. The numerical features used will include opening price, high price, low price, volume, and adjusted prices of stock.

**II. UNDERSTANDING THE STOCK MARKET DATASET**

In order to get meaningful patterns, and study the data distribution, exploratory data analysis is important. In context of stock market data, the exploratory data analysis will help to understand the data well by knowing the statistical distribution, and correlation between the various variables that introduce dependency to alter the prices of stock. Familiarity with basic concepts related to stock market is needed to understand the dataset properly. The dataset selected for the project contains the stock market data from for Walt Disney stock from January 1962 to May 2019. The dataset consists of the 14441 stock records and 10 columns each of which is interpreted as a feature. The dataset consists of end-of-day(daily) stock prices for Walt Disney and is available for download from “quandl” for registered users. The dataset has following columns:

1. **Open:** The opening price of stock is the price at which the trading starts for given stock. At the beginning of day in stock market, when first trading transaction is done, the price of stock is recorded and is considered as the open price.
2. **Close:** The closing price of a stock is the price of the stock at the end of the day when last trading transaction was done for the given stock before the stock market was closed for the day.
3. **Low:** On a given day, the low price of the stock is the lowest price at which the stock was traded at the stock market.
4. **High:** On a given day, the high price of the stock is the highest price at which the stock was traded at the stock market.
5. **Volume:** The volume is an indicator of number of shares that were brought or sold in a particular time frame. For example, if a person buys 10 shares and sells the same 10 shares, then volume is considered to be 10.
6. **Dividend:** Dividend is the extra number of shares that is given to the shareholders when the revenue of company increases. For example, if a shareholder gets 10 percent more shares and currently holds 100 company shares, the total number of shares is 110.
7. **Split:** The stock split parameter is generally decided by the higher management of the company. When a company decides to give out more stocks, the split parameter is decided to increase the number of shares.
8. **Adjusted Prices and volume:** The dataset consists of adjusted open, close, high, and low prices. The adjusted open, high, low and close prices are the values that show the exact amount in cash price for a stock after taking into consideration, the factors such as split, dividend and other transactions that occur after the stock market is closed. These prices are more accurate than the non-adjusted prices. The adjusted volume is counted by taking into consideration, the trading transactions done after stock market closure.

Following figure shows a snippet of dataset used for prediction closing stock price for Walt Disney stock:

A screenshot of a cell phone

Description automatically generated

Fig. 1: Dataset snippet

As per the above snapshot, the data entries for the dates 2019-05-11 and 2019-05-12 are missing. The reason for the entries being missing is that the stock market is closed on weekends and all public holidays. Thus, the trading session is discontinued on weekend and data values cannot be added to the dataset.

**Note:** The use of API key to fetch the data was tried but the use was avoided as Quandl website constantly keeps adding and deleting the data records due to which there was an inconsistency in number of data samples.

**III. EXPLORATORY DATA ANALYSIS**

**1. Statistical Analysis:**

Before studying the univariate and bivariate distribution for predicting the closing stock price, a detailed statistical analysis is done by using describe function to check the minimum, maximum price so that the stock price range can be determined. Following is the output snippet of statistical analysis of the stock market dataset used:

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Fig. 2: Statistical Analysis of stock data

As per the statistical analysis of the dataset, for each feature the range of can be determined. As the dataset used contains historic data records ranging from 1962 to 2019, the prediction accuracy is increased. This distribution helped to study the spread of data. As per the records, the minimum price of Walt Disney stock was 13.8 dollars and the maximum price raised to 244 dollars till date. The dataset was queried to get the date of lowest stock price and following is the snippet of the result:

A screenshot of a cell phone

Description automatically generated

Fig. 3: Walt Disney lowest stock price

As per fig. 3, we can see the lowest stock price for the Disney stock was 13.48. As per the records from CNN Money [4] Disney stock prices slashed down by 11 percent due to low demand in the stock market on 8th August 2002. After querying the dataset to check the number of days the high stock price was greater than mean high stock price of 62.738. About 5367 days from 1962 to 2019 the stock price was in range of 62.738 to 135 dollars for Walt Disney. On 15th May 2019, the stock price raised to 135.21 dollars. As per the records, the price of Walt Disney stock raised by 1.7 percent which is an indicator of positive growth [5].

Since stock market prices are sensitive to various features, the price range definition will help predict the closing price of the stock more accurately. Also, the presented statistical analysis will serve as a proof to support the risky investment decisions [6]. Thus, statistical data analysis will help in understanding the trend of Walt Disney prices getting affected over the period of time and also various factors affecting the price changes and fluctuations in the stock market for Walt Disney.

**2. Data Visualization using histograms:**

Plotting the distribution of stock market data will help in understanding the possible trend followed by the stock values, help in understanding the spread of data and will be useful in further price predictions based on historical data. Following graphs demonstrate the difference between the adjusted prices and actual cash prices of Walt Disney stock, plotted using the dataset downloaded from Quandl:

A screenshot of a cell phone

Description automatically generated

Fig. 4: Open Price vs. Count

The graph demonstrated above shows the distribution of open stock prices and the count for the respective price. The X-axis represents the range of open prices for the Walt Disney stock for the years 1962 to 2019 and Y-axis represents the frequency of the price on represented on X-axis (number of occurrences). As seen from the graph there is a considerable difference between the adjusted prices and the actual open cash prices of the Walt Disney stock price. That happens because the “Open” price is the actual cash price of the particular stock at start of the day. While, the adjusted open price is calculated by taking the actual open price as a starting point and calculated a new price by taking features such as dividend, split into consideration. As it can be clearly seen, the highest open price (denoted in blue) is around 240 while the highest adjusted open price is approximately 150 dollars. The open price of Walt Disney was in range 35 to 50 dollars for majority of time period while the adjusted open price was in the range 15 to 20 dollars approximately.

Similarly, the financial data is studied to see the variation in between the adjusted high price and high price which is visualized using histogram as follows:

A screenshot of a cell phone

Description automatically generated

Fig. 5: High Price vs. Count

The prices displayed on X-axis vary from 1 dollar to 247 dollars, the highest price attained by Walt Disney stock till date lies in the range 245 to 250 dollars approximately. The “high” price is the highest price at which the stock was bought or sold on that day. From the year 1986 to 2019 the variation of prices is as shown in the histogram above which shows that there is an increase in the stock prices with time, but, there are abrupt variations and other factors such as turnover of Walt Disney, that resulted in sharp decline in price of stock leading to low prices such as 13 dollars offered per stock.

Following graph shows the histogram plotted for showing the variations in the “low” prices of Walt Disney stock from the year 1962 to year 2019:

A screenshot of a cell phone

Description automatically generated

Fig. 6: Low Price vs. Count

The low-price range for Walt Disney stock varies from 3.08 dollars to 158 dollars approximately. The most probable lowest price after visualizing the low-price data lies in between the range 35 dollars to 40 dollars as per the above histogram.

A screenshot of a social media post

Description automatically generated

Fig. 7: Volume vs. Count

As shown in the histogram above, the number of shares Walt Disney shares that were traded per day investing in Walt Disney stock is in range of 0.01\*10 raised to 8 which is mentioned as 1e8 using matplotlib. This denotes that, on an average number of people greater than 10,00,000 invest in the Walt Disney stock making it one of the popular stocks for trading transactions and profit benefits. After taking other influential factors such as turn-over of company, revenue of the company, split of stock and dividend, the adjusted volume of shares decreases.

The dividend remains in the range 1 to 1.5 and split remains in the range of 0 to 0.2 as shown in the following histograms:

A screenshot of a social media post

Description automatically generated

Fig. 8: Variation in split value of Walt Disney stock

A screenshot of a social media post

Description automatically generated

Fig. 9: Variation in dividend value of Walt Disney stock

**3. Data Visualization using Pairplot:**

The bivariate distribution of data can also be plotted to understand the probabilities of a particular event to happen given all the combinations of the features used in the dataset. Probabilistic analysis is very important for stock market data analysis To study the rise and fall in the closing price of stock, the bivariate distribution was plotted. By plotting the bivariate distribution, we checked how the increase or decrease in value of a certain feature impacts the increase or decrease in the closing price of stock. In. order to understand the relation between the pair of variables, pairplot was used in this project. The pairplot makes it easy to plot the dependency of other features on determining the value of closing price for Walt Disney dataset. Following is the bivariate distribution plotted for the dataset used:

A picture containing window

Description automatically generatedA picture containing text

Description automatically generated

Fig. 9: Walt Disney stock data visualization using pairplot

The features used to plot the pairplot are as listed below:

Open price, High price, Low price, Close price, Volume, Dividend, Split, Adjusted Open price, Adjusted High price, Adjusted Low price, Adjusted Close price, and Adjusted Volume.

The diagonal histograms show the marginal distribution while the upper and lower triangular portions represent the joint distribution. Pairplot is used to determine the relation between the features across multiple dimensions. As we have multiple features in the dataset, we need to visualise the dataset in a different way. As per the pairplot when taking features such as open, , high, and low prices the relation established is almost linear while the points appear to be scattered for volume in combination with other features.

**3. The Correlation Matrix:**

The co-relation matrix also gives an idea about how the features are co-related by assigning them a score. Using a correlation matrix, we can decide how a change in a particular feature, for example, the Open price of a stock, will impact our prediction of closing price of the stock.

Following is the snippet of co-relation matrix calculated for the Walt Disney closing stock price prediction:

A screenshot of a cell phone

Description automatically generated

Fig. 10: Correlation Matrix

The diagonal of the matrix is 1 which indicates that each variable has a perfect correlation with itself [7]. The features having negative correlation tell that as one quantity increases, the other decreases. For example, the Open price and Volume have correlation coefficient of -0.24 which means that as the Open price increases the Volume of trades done decreases which is logical as, when the stock becomes costly people buying the stock will decrease. Similarly, it can be inferenced that, the closing price of the Walt Disney stock decreases with increase in volume, adjusted volume and split as they have negative correlation coefficient value. Also, the closing price will increase with increase in Open price, adjusted Open price, High price, adjusted High price, Low price, Adjusted Low price, and Dividend as they have positive correlation coefficient value.

**IV. DATA PREPROCESSING**

Data Preprocessing is an important step in any machine learning technique. There can be many steps in data preprocessing like removing features that does not add any value (zero variance), removing corrupt data, getting rid of noise from the data. **In our project, we have used Open Price, High price, Low price, Volume Dividend, split, adj\_open, adj\_high, adj\_low and adj\_volume to predict closing price of the stock.** We have used data cleaning and feature scaling to preprocess our dataset.

**1. Data Cleaning:**

Data cleaning was used to remove the null values from the dataset. We also tried to get rid of features which have zero variance from our dataset. Fortunately, none of the features had zero variance. Following is the output of code executed to find NULL values in the dataset:

Feature     Number of null values

Open          0

High                   0

Low             0

Close           0

Volume          0

Dividend        0

Split           0

Adj\_Open        0

Adj\_High        0

Adj\_Low        0

Adj\_Close                0

Adj\_Volume        0

**2. Feature Scaling:**

Feature scaling is an important step to preprocess the data as it helps the model to converge faster. It basically brings the data between a particular range. Feature scaling was done on all the features of our dataset. We used MinMaxScaler to scale our data. It scaled the data such that all the values of a feature lie in between 0 and 1 in our case, as none of the values were negative. We used min-max scaler in our dataset as values of all the features were varying greatly.

Following is the snapshot of dataset before and after scaling:

Before scaling:

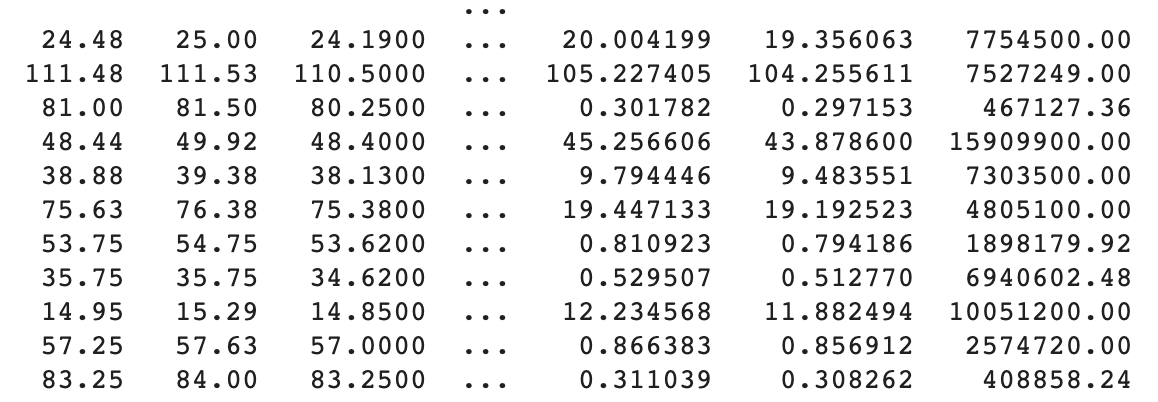


Fig. 11: Dataset before scaling

Dataset After scaling:

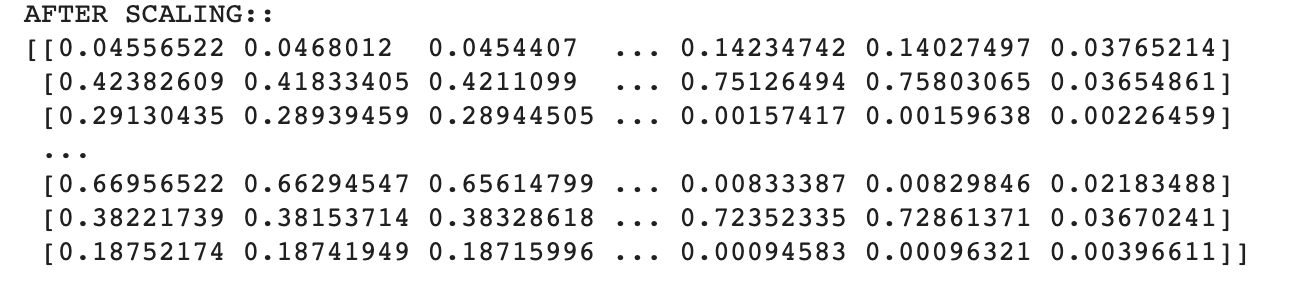


Fig. 12: Dataset after scaling

The values before scaling vary largely as the ranges in which the data is represented is different. The prices are mentioned to be in dollars but the number of people trading stock is a numerical value greater than that of price data. The values after scaling lie in the range between 0 and 1 and now can be used for regression.

**V. MACHINE LEARNING MODELS USED and EVALUATION:**

The goal of our project was to predict the closing price of the stock. Since this is a regression problem as values of the predictions are continuous, we decided to use regression techniques like linear regression, Random Forest Regressor, Decision Tree Regressor and ADA Boost Regressor in our project.

As we were working on a regression problem, we decided to use Mean Squared Error and R2\_score as our evaluation metrics. Mean Squared Error is the average of squares of errors. Higher the value of Mean Squared error, higher is the error. **R2\_score** is the score used to determine how well the data fits the regression line. If the value of r2\_score is higher, the model is more accurate than a model with a lower score. This score basically compares the ratio of variance explained by the model with the actual variance.

Following Regression models were used to predict the closing price of Walt Disney stock:

**1. Linear Regression:**

Linear Regression is the most commonly and intuitive Machine Learning model for kind problems where data involved is in the form of time series. It establishes relationship between dependent and independent variables by finding the coefficient in a linear equation. The independent variables are the features whereas dependent variable is the value to be predicted. Following were the results obtained for linear regression.

**Mean squared error = 5.801599091194399**

**R2\_score = 0.9955086946393326**

The graph of predicted and actual values for closing price of the stock are as shown in the scatterplot below:

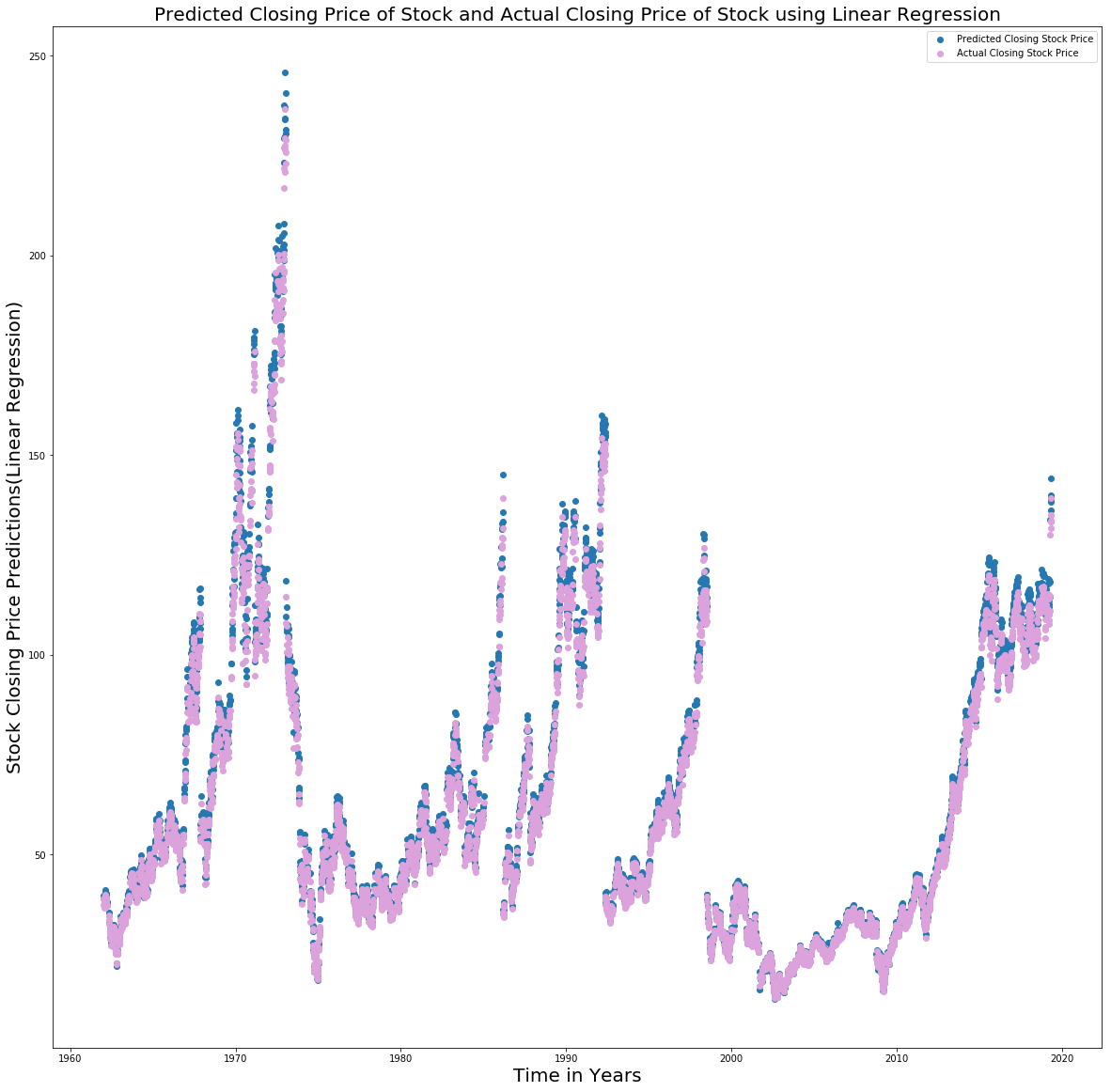


Fig. 13: Actual and predicted values of closing stock prices using linear regression

The pink dots show the actual value whereas the blue dots in the graph depict predicted values. We can see from the graph that the values of actual and predicted values are very close to each other.

**2. Decision Tree Regressor:**

Decision Tree Regressor was the next model that we worked to experiment with predicting closing price of Walt Disney Stock. Decision Trees work well with both continuous and categorical data. The nodes in the decision tree are features and the edges are the values of the features. After trying out various combinations of parameters passed decision tree regressor function, we used following parameters in our training model:

*criterion='mse', splitter='best',random\_state=1,presort=True*

‘Mse’ was the criterion used to evaluate the quality of split. We chose ‘best’ as the splitting strategy. The data was presorted for speeding up of finding best splitting of data.

We obtained following results with our decision tree model:

**Mean squared error = 3.405197555602182**

**R2\_score = 0.9973638678241624**

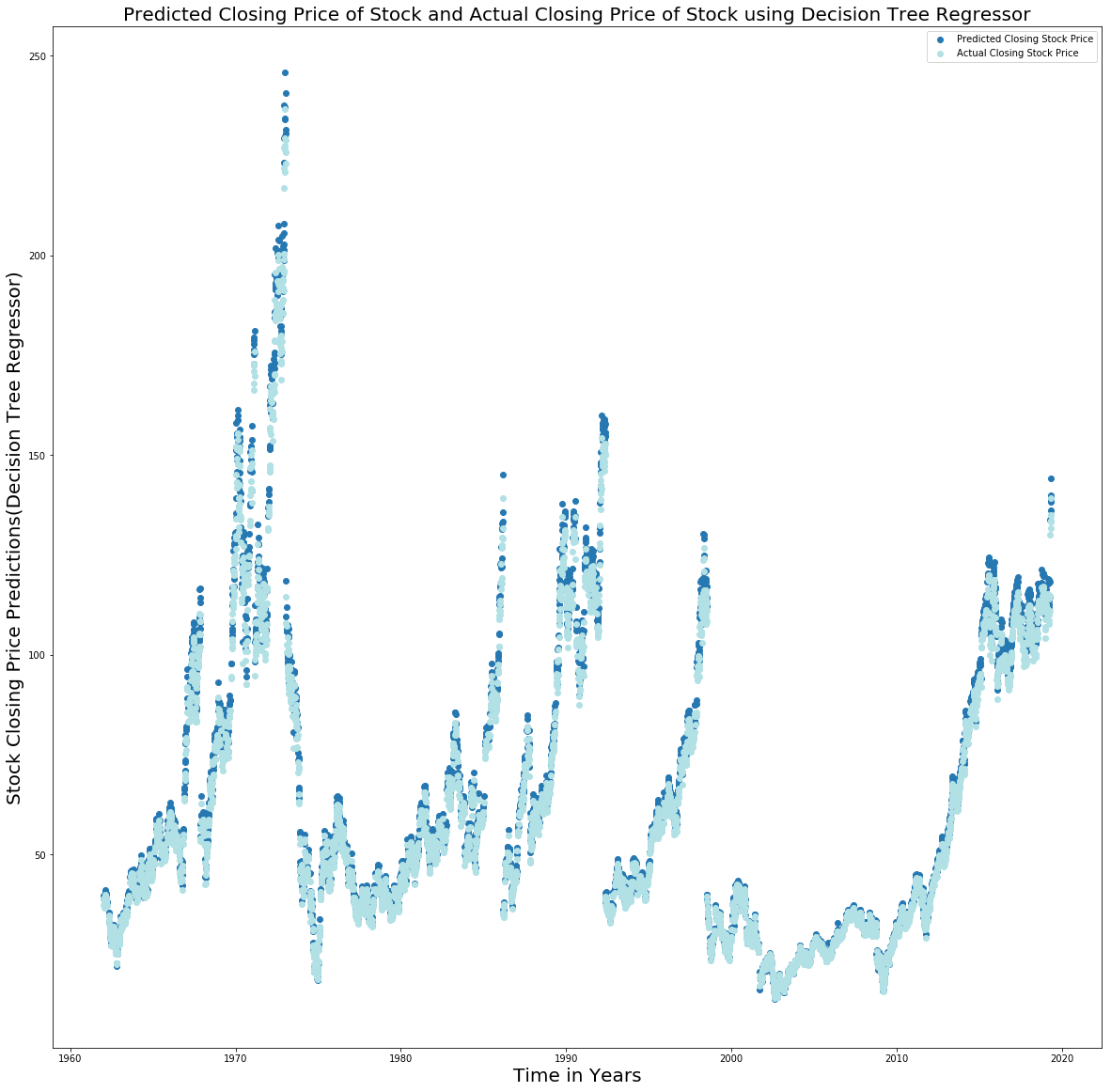


Fig. 14: Actual and predicted values of closing stock prices using decision tree regressor

**3. Random Forest Regressor:**

Random Forest Regressor was the next model we experimented with. Since the decision trees might result in overfitting, random forest builds various trees based on random samples and random features. After trying out various combinations of parameters passed decision tree regressor function, we used following parameters in our training model:

*random\_state=1,n\_estimators=59,max\_depth=50,bootstrap=True,n\_jobs=-1*

Following were the results we obtained with our dataset:

**MSE      = 2.6612175127825544**

**r2\_score = 0.9979398196440006**

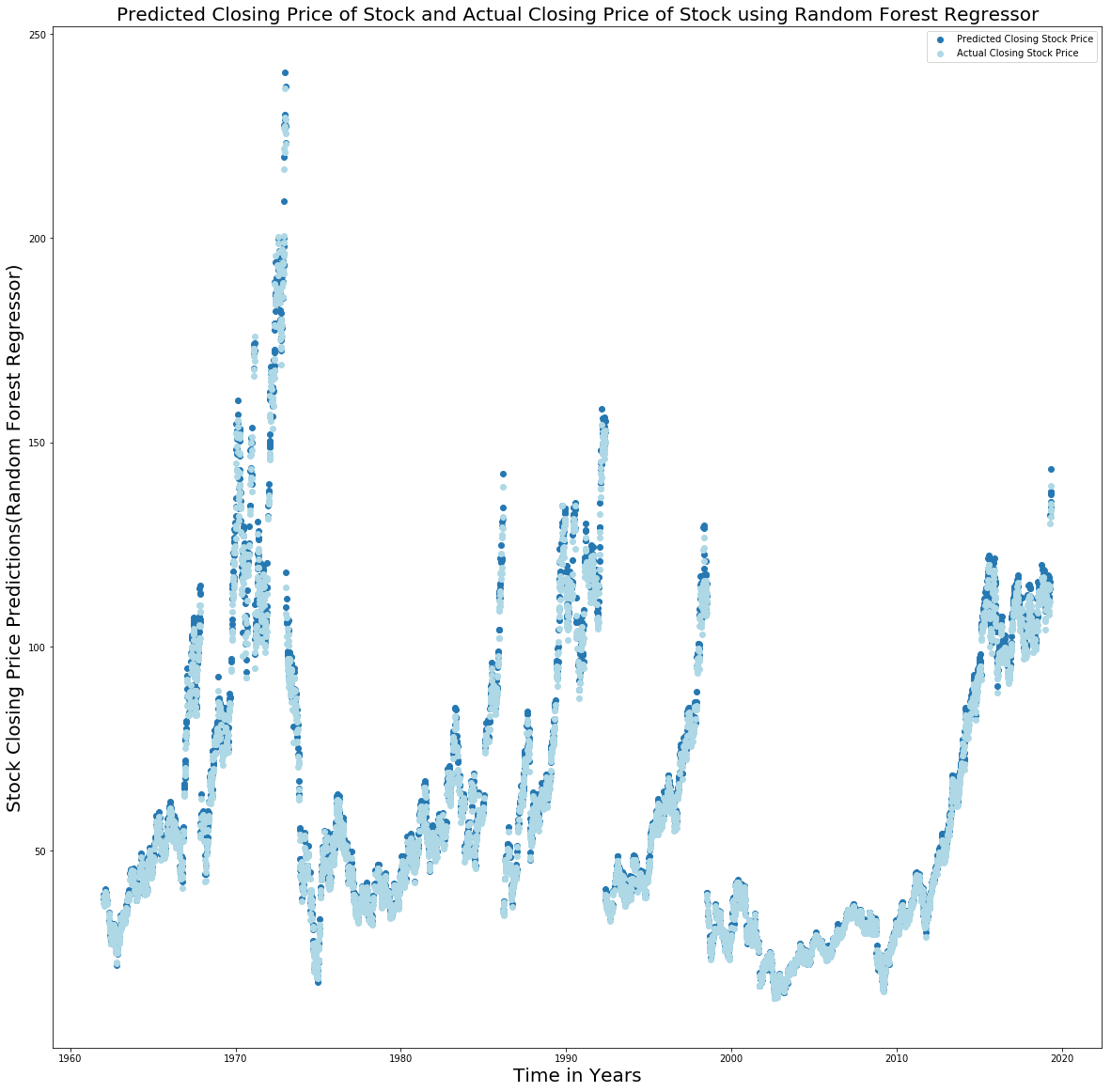


Fig. 15: Actual and predicted values of closing stock prices using random forest regressor

**4. ADA Boost Regressor:**

AdaBoost Regressorwas the next model we experimented with to predict the closing price of stock. It takes into consideration, the various weak regressors and tries to learn with each subsequent regressor in comparison to random forest for the evaluation is done parallelly on all the models and the most voted value is selected. It has higher weights for the data that was wrongly predicted in the previous model for it to be predicted correctly. Following are the parameters that were used in our model:

*random\_forest,n\_estimators=30,random\_state=1,learning\_rate=0.000165*

The base regressor that we used in our model was random forest regressor defined and used previously.

Following were the results we obtained with our dataset:

**MSE      = 2.5793235204733764**

**r2\_score = 0.9980032178417877**

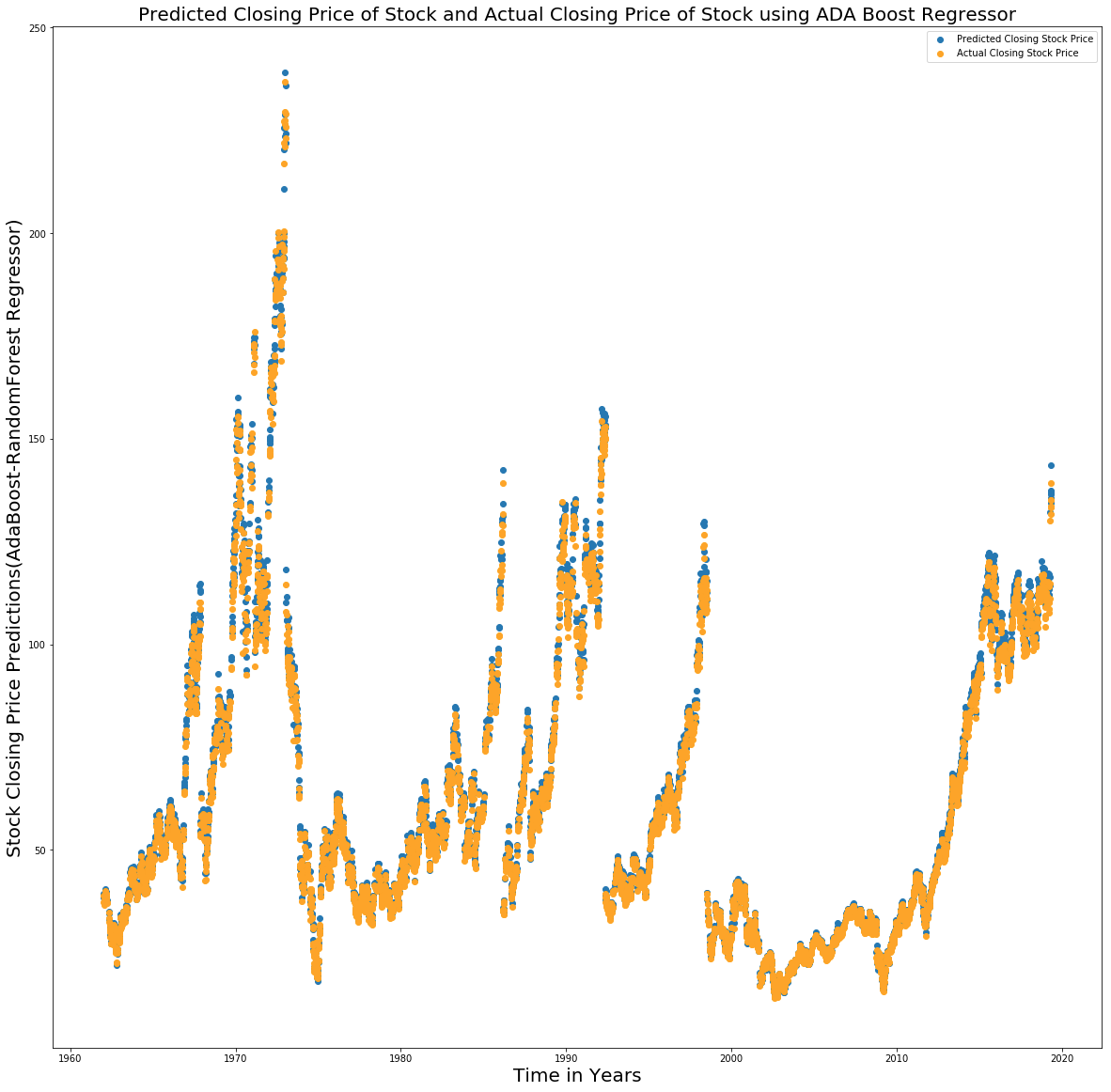


Fig. 15: Actual and predicted values of closing stock prices using ADA Boost regressor

**VI. RESULTS**

Following is the comparison of various techniques:

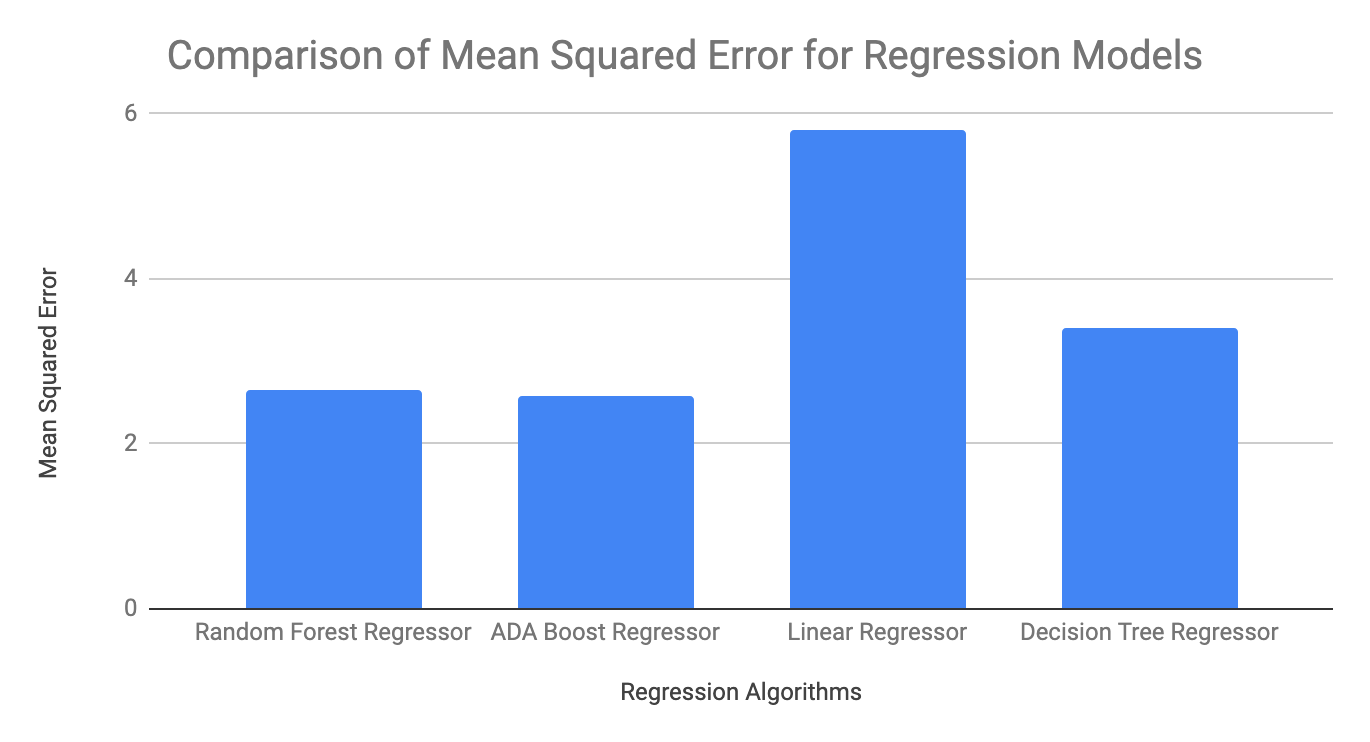


Fig. 16: Comparison of MSE

We obtained Minimum Squared Error with ADA Boost regressor. The results have improved in comparison to Decision Tree which might have been overfitting the data. ADA Boost regressor would have kept improving with each classifier being included.

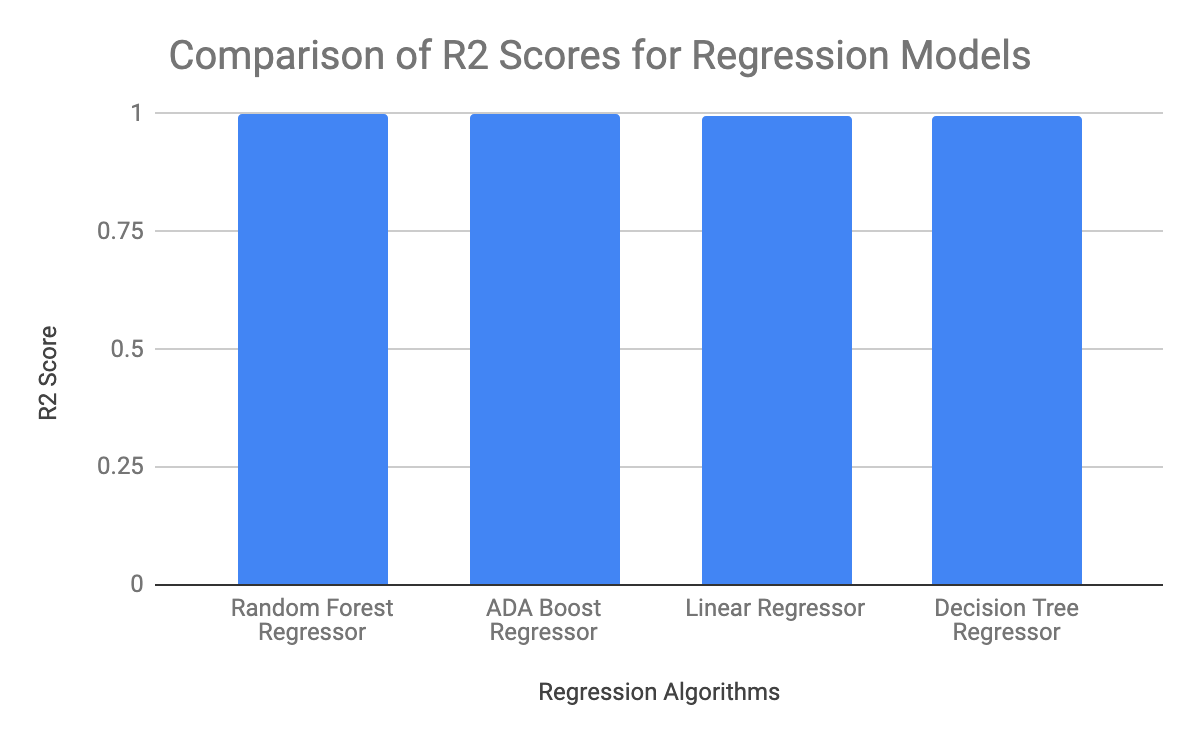


Fig. 17: Comparison of R2 Score

The R2\_scores for all the techniques are around 0.99. But, with ADA Boost regressor, the result is slightly higher.

**VII. CONCLUSION**

From this project, we learnt various techniques like data visualization, preprocessing, building the models and comparison of various results altogether to make predictions for closing prices of a stock and make an investors life easy. This project aims to help the investor in investing the hard-earned money at right place and time. All the steps carried out for achieving the ultimate goal are very important to build a good model for predicting the unseen data in an efficient manner. For our dataset, we obtained best results with ADA Boost regressor with Mean Squared Error of 2.57932352 and R2\_score of 0.9980032178.

**References**

[1]  A. O’ Shea, K. Voigt, “Stock Market Basics: What Beginner Investors should Know”, 2019, Accessed on: May 15, 2019. [Online]. Available: https://www.nerdwallet.com/blog/investing/stock-market-basics-everything-beginner-investors-know/

[2] K. Amadeo, “Benefits of Investing in Stocks Versus Disadvantages”, 2019, Accessed on: May 15, 2019. [Online]. Available: <https://www.thebalance.com/stock-investing-for-the-individual-investor-3306182>

[3] K. Voigt, A. O’Shea, “What is Sotck?”, 2019, Accessed on: May 15, 2019. [Online]. Available: https://www.nerdwallet.com/blog/investing/what-is-a-stock/

[4] CNN Money, “Disney Shares tumble 11%”, Accessed on: May 15, 2019. [Online]. Available:

https://money.cnn.com/2002/08/02/news/companies/disney/index.htm

[5] D. Sparks, “Will 2019 Be the Year Disney Stocks Rises Higher?”, 2019, Accessed on: May 16, 2019. [Online]. Available: https://www.fool.com/investing/2019/04/10/will-2019-finally-be-the-year-disney-stock-rises-h.aspx

[6] J. Williams, “The Importance of Statistics in Management Decision Making”, 2019, Accessed on: May 16, 2019. [Online]. Available: https://smallbusiness.chron.com/importance-statistics-management-decision-making-4589.html

[7] T. Bock, “What is a Correlation Matrix?”, Accessed on: May 17, 2019. [Online]. Available: https://smallbusiness.chron.com/importance-statistics-management-decision-making-4589.html

[8] GeekforGeeks, “Linear Regression”, Accessed on: May 19, 2019. [Online]. Available: https://www.geeksforgeeks.org/linear-regression-python-implementation/

[9] “sklearn.tree.DecisionTreeRegressor”, Accessed on: April 30, 2019. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

[10] “sklearn.ensemble.RandomForestRegressor”, Accessed on: April 30, 2019. [Online]. Available:

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[11] “sklearn.linear\_model.LinearRegression”, Accessed on: April 30, 2019. [Online]. Available:

https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html

[12] “sklearn.ensemble.ADABoostRegressor”, Accessed on: April 30, 2019. [Online]. Available:

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html