ISM 6136 Data Mining

**Stock Price Prediction Model**

By:

Divyang Mehta (U 5159 2551)

Mukund Kathait (U 2768 3820)

Rashesh Kothari (U 9836 8729)

Shaolin Pu (U 1181 0151)

Table of Contents

[Tables of Figures 3](#_Toc24818642)

[Objective 4](#_Toc24818643)

[Data Overview 5](#_Toc24818644)

[Ticker Extraction: 5](#_Toc24818645)

[Attributes Extraction: 5](#_Toc24818646)

[Extracted Data: 6](#_Toc24818647)

[Data Models 9](#_Toc24818648)

[Recommender Systems: Buy, Sell and Hold 11](#_Toc24818649)

[Summary 12](#_Toc24818650)

[Future Scope 13](#_Toc24818651)

[References 14](#_Toc24818652)

# Tables of Figures

[Figure 1: Tickers from S&P 500 5](#_Toc24818658)

[Figure 2: Attributes Description 6](#_Toc24818659)

[Figure 3 Extracted Data Sample 6](#_Toc24818660)

[Figure 4: Apple Adjusted Close 7](#_Toc24818661)

[Figure 5: Microsoft - Adjusted Close 7](#_Toc24818662)

[Figure 6: Facebook - Adjusted Close 8](#_Toc24818663)

[Figure 7: Predictive Model 9](#_Toc24818664)

[Figure 8: Model Observations 10](#_Toc24818665)

[Figure 9: Recommender System – Buy, Sell or Hold 11](#_Toc24818666)

# Objective

Investments and trading have become a smarter way to grow and manage funds for an individual as well as an organization. Stock markets have always been of an interest to those who know where to invest and grow their money. In this paper, the objective is to try and predict the stock prices for certain companies based on the historical data of various attributes from 2010 to 2018 and suggest if a person should buy, sell or hold the stock.

# Data Overview

## Ticker Extraction:

To extract the data of various companies, we have taken a list of top 500 Standard and Poor’s from [Wikipedia](https://en.wikipedia.org/wiki/List_of_S&P_500_companies). Using python, we request a webpage from Wikipedia. Information from the webpage is then converted into text used to identify the list of table class. After identifying the class, the ticker names of companies are saved into a text file (figure 01).

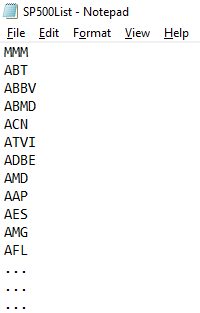


Figure 1: Tickers from S&P 500

## Attributes Extraction:

Using the ticker names of 500 S&P companies, we extract various attributes and store them into individual data file for each company. Each file consists of thirteen attributes and approximately 100 records.

Using the ticker data, the following attributes have been extracted from [www.quandl.com](http://www.quandl.com)

|  |  |
| --- | --- |
| * Date | * Adjusted Open |
| * Open | * Adjusted High |
| * High | * Adjusted Low |
| * Low | * Adjusted Close |
| * Close | * Adjusted Volume |
| * Volume |  |

Extracted data contains above mentioned 11 attributes of last day of each month from 2010 to 2018. Since the data is in monthly format, we have taken our target attribute to be “Adjusted Close”. Adjusted close is an interval type of attribute that can range corresponding to the historical data of the stock profile.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Attributes** | **Description** | **Attribute Type** |
| 1 | Date | Last day of each month from 2010 to 2018 | Independent |
| 2 | Open | Open-High-Low-Close shows the high and low price a stock attained for a period as well as the opening and closing prices of the stock for the same period |
| 3 | High |
| 4 | Low |
| 5 | Close |
| 6 | Volume | Number of stocks traded in a day |
| 7 | Ex-Dividend | Describes a stock that is trading without the value of the next dividend payment. | Excluded |
| 8 | Split Ratio | Ratio of corporate actions in which a company divides its existing shares into multiple shares to boost the liquidity of the shares |
| 9 | Adjusted Open | Adjusted Open-High-Low-Close shows the adjusted high and adjusted low price a stock attained for a period as well as the adjusted opening and adjusted closing prices of the stock for the same period | Independent |
| 10 | Adjusted High |
| 11 | Adjusted Low |
| 12 | Adjusted Close |
| 13 | Adjusted Volume | Target |

Figure 2: Attributes Description

## Extracted Data:

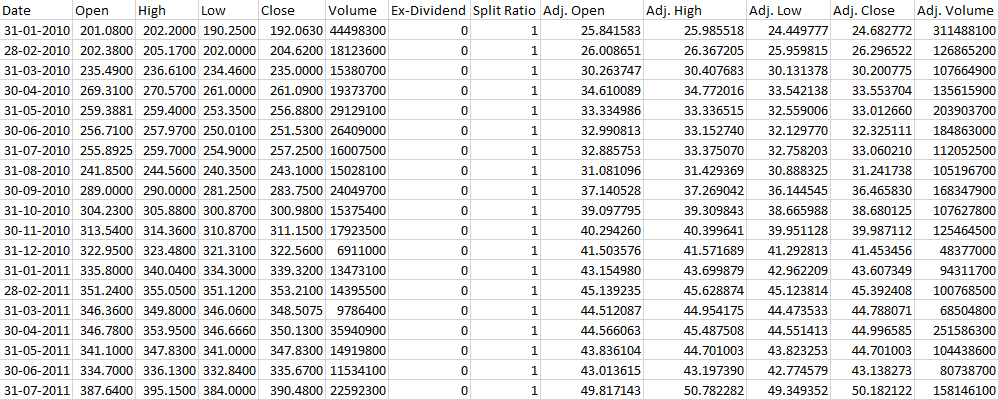


Figure 3 Extracted Data Sample

Since the extracted data consists of data from 2010 to 2018, it has been cleaned for each company. Any attribute that has a cell value equal to null or if it is missing, then it has been replaced by zero. This helps in increasing the accuracy and if a company was started in 2013 then, the financial data of that company would begin from 2013 instead of 2010.

Figure 4: Apple Adjusted Close

Figure 5: Microsoft - Adjusted Close

Figure 6: Facebook - Adjusted Close

# Data Models

We have implemented 3 different regression models for these companies in order to predict the monthly values of Adjusted Close.

* Linear Regression
* Neural Network Regression
* Boosted Decision Tree Regression

The model below is built on Azure Machine Learning Studio where the dataset is of Apple, it is cleaned and split into training and testing sub datasets before training them on various regression techniques.

We observed these models and found a good amount of correlation among 4 attributes (Adjusted Open, Adjusted High, Adjusted Low, Adjusted Close).

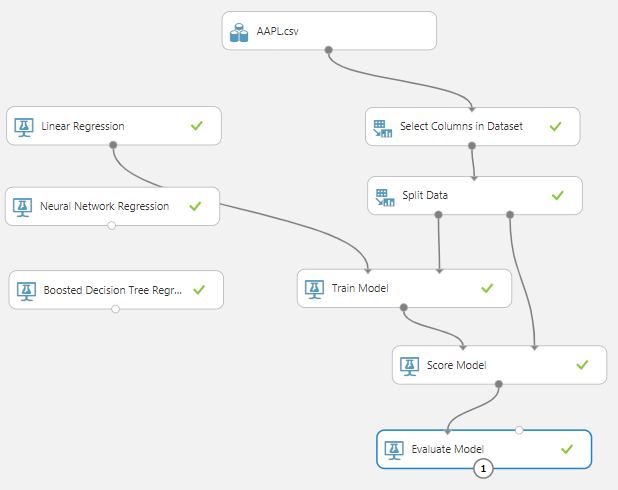


Figure 7: Predictive Model

Upon evaluating the above-mentioned models for three companies, the Root Mean Squared Error and Coefficient of Determination have been considered as a measure of accuracy of a model for that dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Facebook** | | **Apple** | | **Microsoft** | |
|  | **Root Mean Squared Error** | **Coefficient of Determination** | **Root Mean Squared Error** | **Coefficient of Determination** | **Root Mean Squared Error** | **Coefficient of Determination** |
| Linear Regression | 0.8823 | 0.9997 | 0.5200 | 0.9997 | 0.2508 | 0.9998 |
| Boosted Decision Tree Regression | 10.7658 | 0.9631 | 5.4036 | 0.9776 | 4.2706 | 0.9381 |
| Neural Network Regression | 1.7848 | 0.9990 | 1.8370 | 0.9974 | 1.0311 | 0.9964 |

Figure 8: Model Observations

# Recommender Systems: Buy, Sell and Hold

We have used python to categorize the stocks as either Buy, Sell or Hold. This prediction has been done based on the historical data. It suggests whether the investor should buy the stock, sell the stock before the price drops or hold the stock based on the historical data. This model can predict the bucket category for the stocks for up to three months in the future with decent accuracy. It also allows a user to specify the profit margin. For example, if profit margin for Apple is set to 5%, the model would predict if a stock of Apple would be able to gain 5% three months in the future and categorize it either as Buy, Sell or Hold.

The table below indicates if a stock should be bought, sold or held in order to achieve the desired profit. For example, from the historical data of Apple, the model recommends if a user should buy the stock 56 times out of 99. Similarly, it predicts 6 times to hold the stock and recommends selling the stock 35 times out of total of 99 predictions.

|  |  |  |  |
| --- | --- | --- | --- |
| **Profit %** | **Buy** | **Hold** | **Sell** |
| **5%** | 56 | 6 | 35 |
| **10%** | 41 | 31 | 27 |
| **15%** | 24 | 57 | 18 |
| **20%** | 15 | 73 | 11 |

Figure 9: Recommender System – Buy, Sell or Hold

# Summary

* Knowing that there are many different techniques for data mining, it cannot be said with certainty that the above-mentioned models are the best for this kind of dataset as there are number of combinations of data cleaning techniques, data partitioning and creating a predictive model and it is probable that a different model might give better results in terms of Adjusted Close.
* It has been concluded that the comparison done is among the three models built for this report.
* Considering just three companies from the list of 500, we have built nine models in total.
* Three models have been built for each company and all the observations indicate that the accuracy is highest for Linear Regression Models for all the companies.
* In terms of accuracy, the Boosted Decision Tree Regression Model has lowest accuracy for all the companies.
* Recommend if a user should buy, sell or hold a stock based on historical data

# Future Scope

Stock Price is dependent on many other factors other than the data used in this project, few of them are listed below:

#### THE MARKET PLACE

The marketplace is one of the influencers of stock price. There is no algorithm or mathematical formula that lets the investors know exactly how the prices will abate or blow up. However, there are several factors that can alter the stock price.

#### DEMAND AND SUPPLY

Demand and supply in the market also plays an important role in manipulating the prices of shares. When the demand for shares exceeds the supply, i.e. the buyers are more than the sellers, the prices inflates and when the demand is less than the supply, i.e. that the buyers are less than the sellers, the prices subside.

#### ECONOMY

Fluctuations in the economy feature what are commonly referred to as “booms and depressions”. Under favorable conditions the share prices are at their peak and at their lowest point during depressions (the most recent and devastating one being in 2008). Share prices will gradually rise during recovery and fall during recessions.

#### INVESTORS

Market players are the investors have an impact on the stock prices. With more bulls than bears, the prices flourish and with more bears than bulls, share prices tend to dwindle.

# References

* <https://medium.com/@tomyuz/a-sentiment-analysis-approach-to-predicting-stock-returns-d5ca8b75a42>
* <https://github.com/silburt/Machine_Learning/blob/master/Fundamental_Analysis/Fundamental_Analysis.ipynb>
* <https://www.quandl.com>
* <https://www.quandl.com/tools/python>
* <https://www.goodreads.com/book/show/8640735-the-psychology-of-finance>
* <https://www.youtube.com/watch?v=2BrpKpWwT2A&list=PLQVvvaa0QuDcOdF96TBtRtuQksErCEBYZ>