Bayesian Theory and Data Analysis

Final Project Report

TOPIC: Stock Market Prediction

By: Aastha Hurkat

Meet Valia

Priyanshi Gupta

Vijeet Sharma

Index

Stock Market Prediction

I. Aim:

We aim to implement a successful model for stock price prediction to gain insights about market trends over time. We want to investigate the reliability of the Bayesian linear regression model by comparing the predicted valued to actual value of the dataset. We also aim to build a Monte Carlo simulation which could provide us with future stock prices prediction for a particular time duration. Stock Market Prediction can be a motivating factor for investors for making prudent investment which can lead to economic growth and capitalize on the growing prices of the stock market.

II. Data Set:

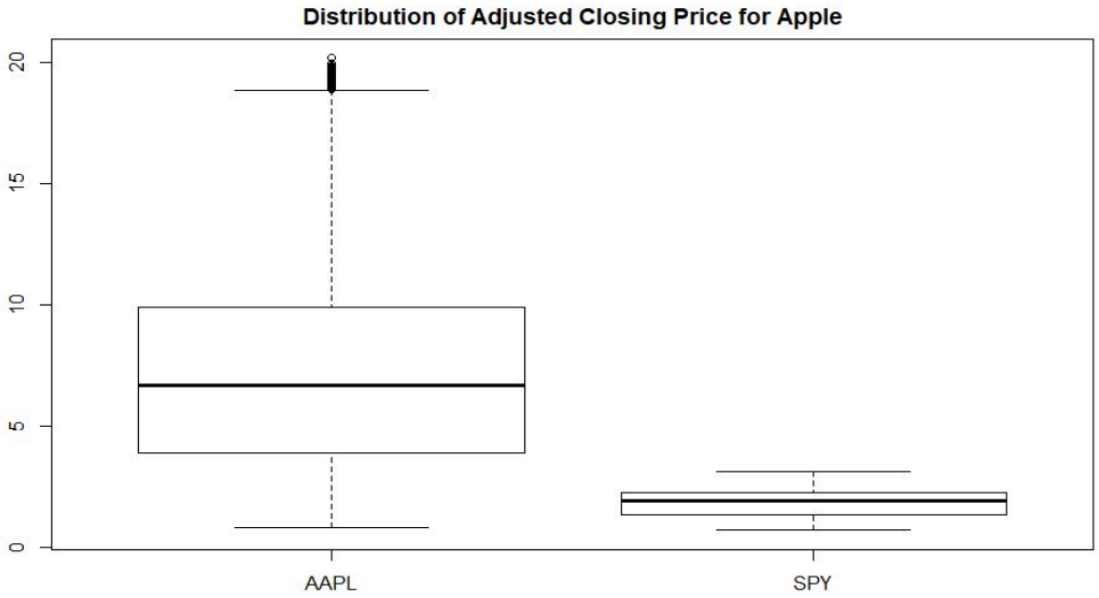
We have taken historical stock prices of Apple from 2009 to 2018 along with the GSPC index found on Yahoo finance. GSPC Index (S&P Index) is a capitalization-weighted index of 500 stocks. It is a popular index and is used to measure the performance of the large cap U.S. stock market. The 500 stocks of this index are selected to be a representative sample of leading companies. Money managers often index their portfolios to match or beat the S&P 500. GSPC is the ticker symbol [1]. Following are the fields of the dataset:

- **Date** - In format: YY-MM-DD
- **Open Price** - price of the stock at market open (this is NYSE data so all in USD)
- **High Price** - Highest price reached in the day
- **Low Price** - Lowest price reached in the day
- **Close Price** – Closing price of the day
- **Adjusted Close Price** – Closing price adjusted for dividends and stock splits
- **SPY-** It is the price values of the S&P 500 index. Each stock is reactive to some extent to the market.

The major fields we are using in this project are “Date” and “Adjusted Close Price” of a particular company in question.

III. Statistical Analysis of Dataset:

The above graph shows price vs time graph for the stocks of AAPL and GSPC index over a year. The GSPC index stock price is increasing at a steady rate whereas there is high volatility for the Apple stock.



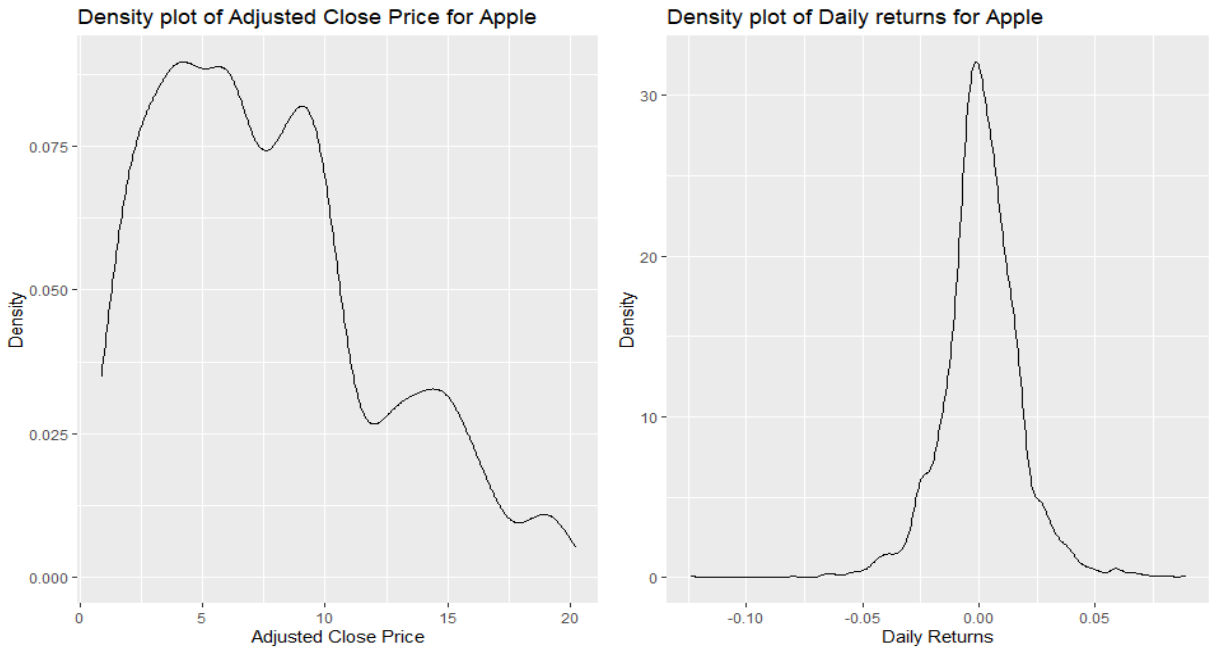
The above graph clearly has been created by normalizing the adjusted closed prices for Apple and GSPC index. We can see that the median of AAPL adj. prices is much higher than the GSPC index adj. prices. It is also evident that the inter quartile range for AAPL is more than twice the range of the box for GSPC index. Even the overall spread (distance between adjacent values) is comparably different. We can say that the dispersion in Apple stock prices is higher than GSPC index adjusted prices.

By looking at the boxplots, we can also say that there is no skewness in the GSPC index data whereas Apple data is slight skewed to the right. Overall, the data looks symmetric for GSPC index but the potential outliers in AAPL data would influence any calculations of skewness.

Investment decisions are made based on 2 important factors:

- How much the stock would grow over the next period
- Best identify the noise which is usually unpredictable. Also considered as risk factor.

Let us compare the density plots of the stock price and the daily returns to identify the further methodologies that would be required to make these decisions.



We can infer from the plot above that Daily Returns are normally distributed whereas we would need a more complex model like Multiple Linear Regression to predict the stock prices. Apple Inc. stock prices seems to have a right skewed multimodal distribution compared to the standard normal distribution of the Daily Returns for the same.

We can identify the maximum risk factor from daily returns which would to some extent which include the noise introduced in the daily stock prices by taking random walks using the Monte Carlo Simulation.

IV. Bayesian Linear Regression Method:

In this phase of the project we aim to implement Bayesian linear regression through R. Through this method we aim to use various features of stock market to predict the adjusted price in future in our case next day's adjusted price. Following is the methodology we followed:

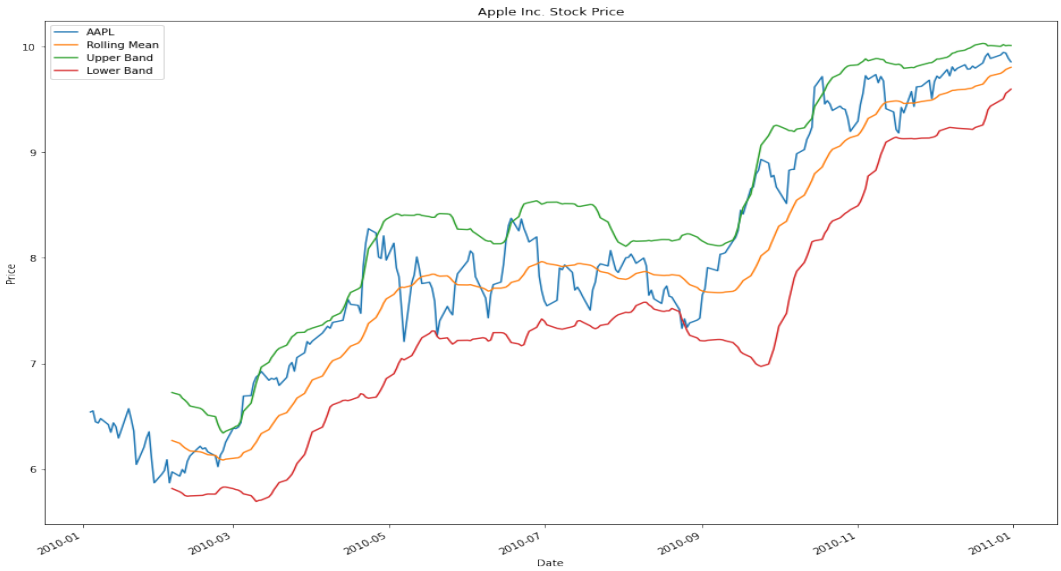
➢ Feature Augmentation:

We create some of the Technical Analysis parameters regularly used by Quantitative Traders to predict the stock prices over a period. The Blue line represents the actual stock price for Apple Inc. We use Adj.Close and High / Low from the data that we downloaded from Yahoo Finance. The parameters generated are:

(i) **Simple Moving Average**: The orange line in below graph represents the moving average of a 24-day window period. Since the first window is created on the 24th day, we lose the data for the first 23 days. Starting from the 24th day, each data point represents a mean of the current as well as the last 23 days. This helps smoothen out the curve of prices a bit and helps eliminate the noise.

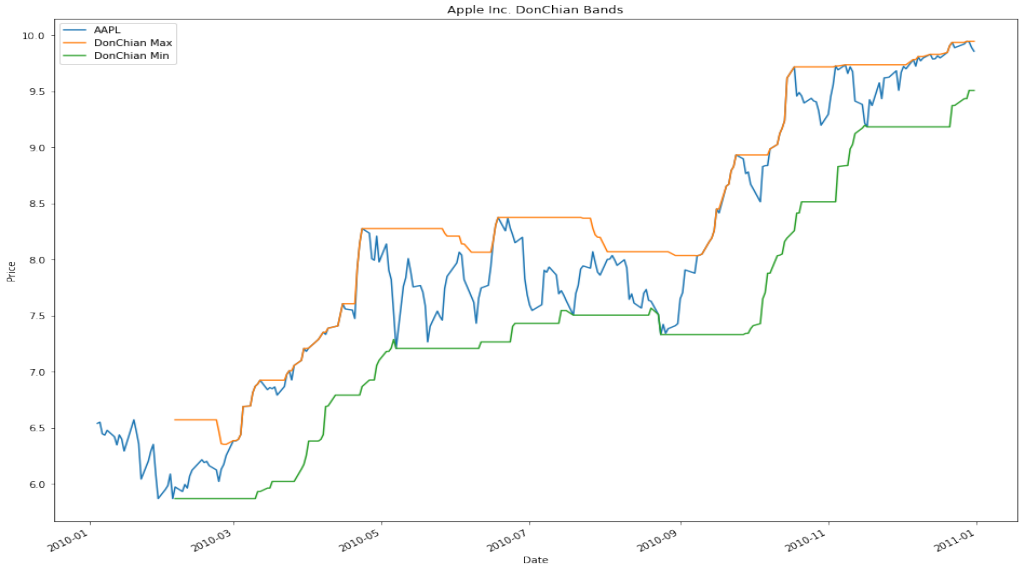
(ii) **Bollinger Bands**: These are a combination of two price bands which run parallel to the Simple Moving Average. One band, the green one, runs 2 standard deviation above the SMA and other, red one, runs 2 standard deviation below the SMA. The standard deviations are also calculated for a period of 24 day rolling window period. These bands are used to identify the points when the actual stock price goes beyond an accepted standard. For example, when the actual price crosses the green band, it represents a sell signal, whereas when it crosses the red band, it is a buy signal.

Below is the graph depicting these Bollinger Bands and Simple Moving Average:



(iii) **Donchian channels:** The rolling maximum of high price (Orange line) for the last 24 days and rolling minimum of low prices (Green line) of the last 24 days.

Below is the graph depicting these Donchian Channel:



➢ Date Preparation

For predicting the stock adjusted closed price of a particular day, we have taken the features of previous days as predictors.
Therefore, we will take the features of day N and use it to predict the adjusted price of day N+1, we have taken features and shifted the adjusted price upward by 1 day to create a lag and thus actual value of the model. For fitting the data in the regression model, we split the data into training and testing set according to the 80/20 rule where training dataset is 80% of the whole data and testing is 20%.

➢ Linear Regression Model:

The regression model we will be working on is as follows:

$$Y_i \approx Normal(B_0 + B_1 SPY + B_2 SMA24_i + B_3 UpperBand_i + B_4 LowerBand_i + B_5 Max24_i + B_6 Min24_i, \sigma^2)$$

Where B_0 is the intercept and B_1---B_5 are the coefficient of the features of the dataset.

The notation is simpler to be defined as:

$$X = \begin{pmatrix} 1 & SPY_1 & SMA24_1 & \dots & Max\ 24_1 & Min24_1 \\ 1 & SPY_2 & SMA24_2 & \dots & Max\ 24_2 & Min24_2 \\ . & . & . & & . & . \\ . & . & . & & . & . \\ . & . & . & & . & . \\ 1 & SPY_n & SMA24_n & \dots & Max\ 24_n & Min24_n \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_6 \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_6 \end{pmatrix}.$$

The regression model is defined as follow:

$$E[y_i|x] = B_0 + B_1 SPY + B_2 SMA24_i + B_3 UpperBand_i + B_4 LowerBand_i + B_5 Max24_i + B_6 Min24_i$$

$$V[y_i|x] = \sigma^2$$

$$Here$$
$$x_i = (1, SPY, SMA24_i, UpperBand_i, LowerBand_i, Max24_i, Min24_i)$$
$$and$$
$$B = (B_0, B_1, B_2, B_3, B_4, B_5, B_6)$$

➢ Zellner g-prior:

The g-prior is an objective prior for the regression coefficients of a multiple regression:

$$\beta|Y, \sigma^2 \sim \text{Normal}\left(\frac{g}{g+1}\hat{\beta}, \sigma^2 \frac{g}{g+1}(X^T X)^{-1}\right)$$
$$\sigma^2|Y \sim \text{IG}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + SSR_g}{2}\right)$$
$$SSR_g = Y^T\left(I - \frac{g}{g+1}(X^T X)^{-1} X^T\right).$$

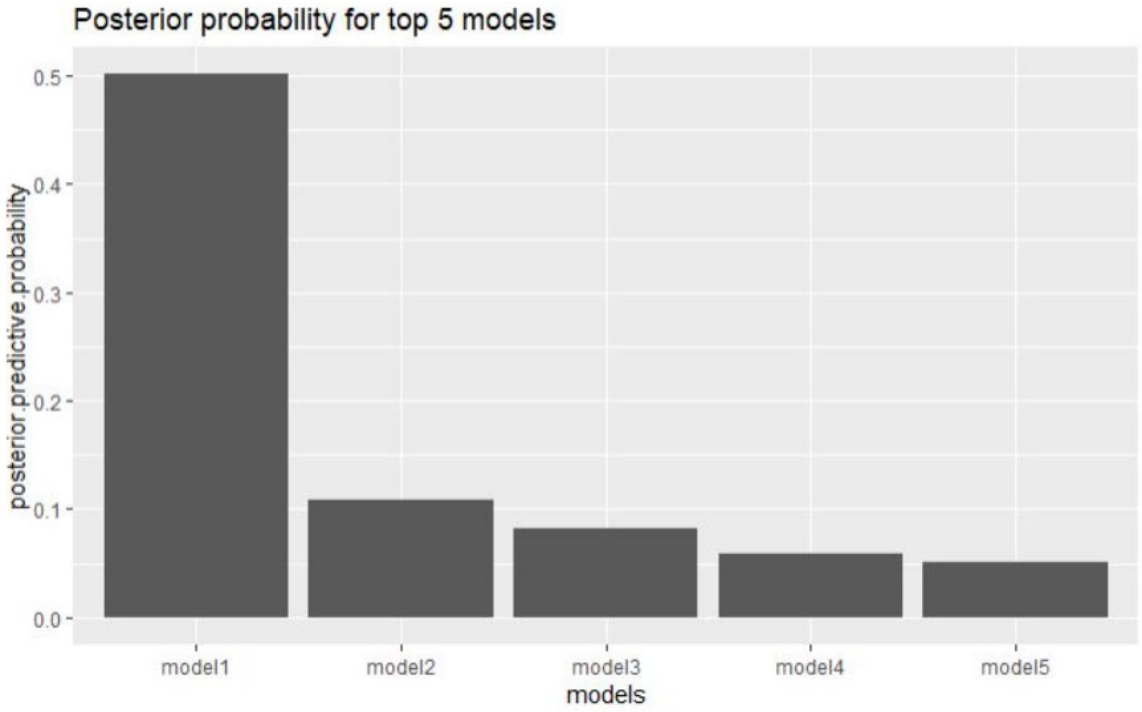
Here we have taken g to be equal to the number of rows for diffuse prior.

➢ Model Selection:

Since we will have a finite set of models at our disposal {z}, that we can use to explain our data, and let us write Bi for the parameters of model z. We can get the posterior probability over the models via Bayesian theorem for the training dataset.

$$p(z|y, x) = \frac{p(z)p(y|x, z)}{\sum_{z=1}^{m} p(z)p(y|x, z)}.$$

Here Pr(z) is a prior distribution over models that we have selected. We have taken uniform prior over all the models. Following is the posterior probability of dfferent models that we attained after running the linear regression model on training dataset:



As we can see from above, model 1 has the highest posterior probability from the first 5 models.

Coeffecients Included in Model 1:

- SPY
- SMA24
- MAX24
- MIN24

Therefore the final regression model we will be working with is as follows:

$$E[y_i|x] = B_0 + B_1 SPY + B_2 SMA24_i + B_5 Max24_i + B_6 Min24_i$$

➢ Model Prediction:

In this phase of the project, we used the model which has been selected in the model selection process to fit our testing dataset in order to get prediction of adjusted close price. Following graph shows us the actual adjusted close price (Stock Price) vs the predicted adjusted close price (Predicted Stock Price) trend.

As we can see from the above graph, both actual price and predicted price are quite close to one another, however there is some percetage of error between them.
This error is due to the fact that Bayesian Linear regression always predicts the expected value, it cannot capture the noise due to randomness in stock prices.

- **<u>Prediction of Adjusted Stock Prices by Linear Regression:</u>**

We have also predicted the Adjusted close Price using Frequentist Linear Regression approach as given below:

$$\hat{Y} = a + b * X$$

$$b = r * (S_y / S_x)$$

$$a = y - b * x$$

In this case, linear regression assumes that there exists a linear relationship between the response variable and the explanatory variables.

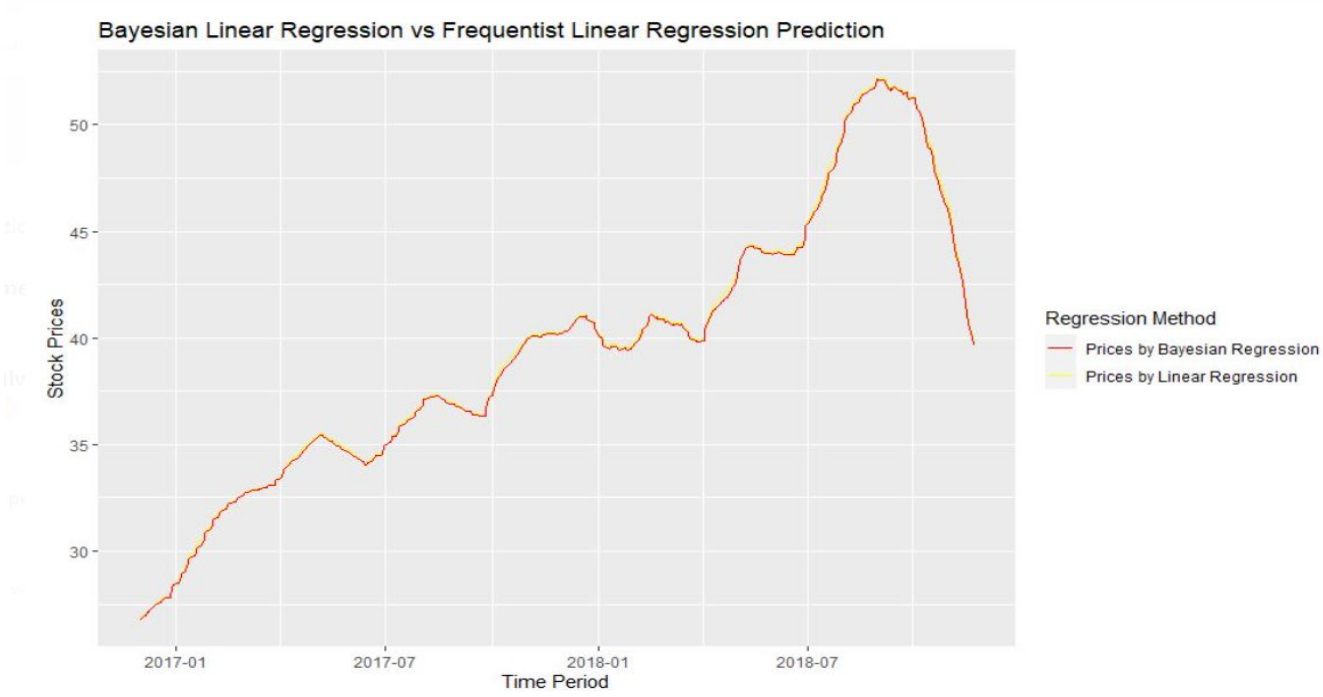
We then compare the prediction of Stock price using Frequentist Linear Regression to the Bayesian Linear Regression.

➢ Evaluating Bayesian Linear Regression and Frequentist Linear Regression:

For evaluation the result of Bayesian model we have used Mean Squared Error. The purpose of choosing this metric was to compare the results with frequentist linear reression results.

Mean Squared Error measures the average squared difference between the estimated values and the actual value.

Regression Method	MSE Value
Bayesian Linear Regression	5.102
Frequentist Linear Regression	4.884



We can see that both results are comparable to one another. Above graph shows us how results from bayesian linear regression are comparable to frequentist linear regression results.

V. Risk Prediction By Monte Carlo Simulation:

Monte Carlo Simulations are one of the most widely used approach to estimate the risk of investing in stocks. We use Monte Carlo simulations that attempts to predict the worst likely loss for a given stock given a confidence interval over a period.

We use the Apple Inc. stock to test our analysis over a period of 15 days. We apply the Monte Carlo Simulation approach to Geometric Brownian Motion which is one of the most basic financial models in the world of finance. The stock price follows a random walk and is consistent with the weak form of Efficient Market Hypothesis, which is basically states that all information is reflected in the share price.

The formula to predict the next price is as follows:

$$PriceToday = PriceYesterday * e^r$$

Where r is the Brownian Multiplier

We apply the random walk Monte Carlo Simulation on the log returns of the daily percent change in the stock prices. To calculate that we use the following formula:

$$LogReturns = log(PercentChange + 1)$$

where $PercentChange = (Today'sPrice/Yesterday'sPrice) - 1$

The components to calculate e^r are as follows:

- **Drift:** The direction that the rate of returns has followed in the past. It is basically the expected value of the returns. We will multiply it by 0.5 since historical values are eroded over time.

$$Drift = \mu - (1/2)\sigma^2$$

Mu & Sigma derived from previous 2500-day prices

- **Volatility:** The historical volatility multiplied by a random standard normal variable.

$$Volatility = \sigma * Z$$

- This random standard normal variable represents our prior belief that the daily changes of stock prices follows a random function which is standard normal.

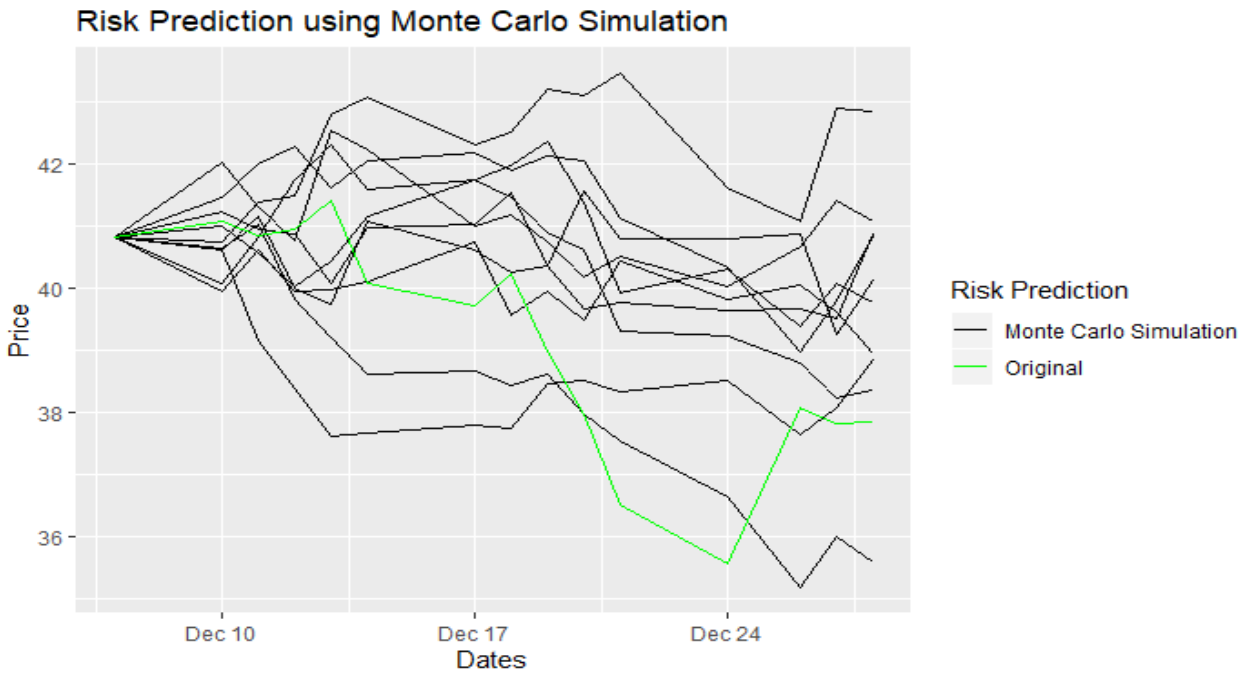
$$Z \approx Normal(0, 1)$$

- **Brownian Multiplier**:

$$r = Drift + Volatility$$

We can use this Brownian multiplier to predict the prices of next day in the Monte Carlo Simulation.

We use 1000 random walks to predict all the possible outcomes of the Apple Stock over a period of 15 days.



In the above plot, the black lines represent the output of 10 random walks out of the 1000 that we have created of the Monte Carlo Simulation while the green line represents the actual value of the Apple Stock Price.

As we can observe, the mean of the random walk prices would predict a slight decrease in the price of the stock. But a few MC simulations does predict a sharp drop in the stock price. We can consider it as a maximum risk factor and make our decisions for investment accordingly.

Further, we would also be using a regression model to predict the actual stock price using various parameters technical analysis parameters that are used by Quantitative Traders for making investment decisions.

VI. Subject Matter Implication of Study

The Monte Carlo Simulations and Regression Model combined together can be used to make investment decisions. The Monte Carlo simulation should be used to identify the maximum possible risk involved in investing in the stock. This would give the investors a pessimist opinion of the amount of loss they are willing to incur in case their other hypothesis fails. On the other hand, the regression model gives an estimate and the direction in which the stock price would follow if the hypothesis stands true. They can either make decisions to buy, sell or hold stocks by predicting the future prices of the stock.

VII. Executive Summary

The stock market is an essential component of the nation's economy, where most of the capital is exchanged around the world. Therefore, the stock market's performance has a significant influence on the national economy. It plays a crucial role in attracting and directing the distributed liquidity and savings into optimal paths. So, predicting the stock prices becomes crucial for the market. The index price is hard to forecast due to its uncertain noise.

The prediction of a stock market direction may serve as an early recommendation system for short-term investors and as an early financial distress warning system for long-term shareholders. So, Investment firms, hedge funds and even individuals have been using financial models to better understand market behavior and make profitable investments and trades. The stock index price prediction has always been one of the most challenging tasks for people. To predict the stock price, we have considered the company history, industry trends and other lots of data points. This project seeks to implement Bayesian Linear Regression for the prediction of the Apple's Adjusted Closing Stock Prices, and we have also compared that with the Frequentist Linear Regression. The MSE for Bayesian Linear Regression and Frequentist Linear Regression are 5.104 and 4.884, respectively. So, they have captured a significant trend to predict the stock prices and both the models are not much different from each other.

Another issue for Investors to is to calculate the risk for investing in a particular stock- how high or low will the price of that stock will go. They are familiar with the saying, "buy low, sell high" but this does not provide enough context to make proper investment decisions. Before an investor invests in any stock, he needs to be aware how the stock market behaves. Investing in a good stock but at a bad time can have disastrous results, while investment in a mediocre stock at the right time can bear profits. Financial investors of today are facing this problem of trading as they do not properly understand as to which stocks to buy or which stocks to sell to get optimum profits. So, we also predicted how high or low will the stock price of the Apple will go in the future 15 days. We took Drift and Volatility into account for the prediction of future stock prices and the predictions are quite like the original price of the Apple. Looking at these statistics, an investor can decide if he wants to invest in the Apple and if interested, when to invest the money.

VIII. Future Work:

Stock market prediction is the act of trying to determine the future value of company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield a significant profit. The efficient-market hypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable.;
We can implement the following approaches for future in the field of finance:

- We can take factors like volume burst risks, Political and social upheavals, Current affairs, News, and other factors into consideration for predicting the stock prices.
- We can apply dimensionality reduction techniques to consider the most important features and we can also assign more weights to the more important features.
- Finance is highly nonlinear and sometimes stock price data can even seem completely random. So, we can apply Time Series prediction or Neural Networks for better prediction of stock prices.
- We will consider the correlation analysis between the stock market valuation and the economic situation of business entities growth.

Reference:

[1] https://www.yourdictionary.com/s-amp-p-500-index
[2] https://www.cse.wustl.edu/~garnett/cse515t/fall_2019/files/lecture_notes/7.pdf
[3] https://www.investopedia.com/articles/07/montecarlo.asp