# STOCK MARKET PREDICTION

Group17

IST718

Rishabh Agarwal, Wesley Stevens, Haoyang Shang, Xin Bai

# Table of Contents



# ABSTRACT

The stock market is a place where fortune is gained, traded, and lost. Many people and companies buy and sell stocks at the market to make a profit, but many fail to do so because the stock market is ever changing and difficult to predict. We perform an exploratory data analysis on stock market data over a period of 5 years. In this analysis, we determine which stocks perform the best and worst overall. We also determine which stocks have the highest and lowest daily returns and volume of shares.

We evaluate 3 statistical models on predicting the closing price of a stock using a running error metric. We show that the closing price is linearly correlated with the other attributes in the dataset. We found that the volume of traded shares has a very low impact on the closing price whereas the opening price has a much higher influence. Finally, we explore the idea of moving averages and show it's benefit to traders as a signal for when to buy, hold, and sell a stock.

For the inference part, we were able to find that surprisingly the volume attribute wasn't a significant predictor in predicting the closing price as its feature importance score is quite less. AAPL,MSFT,INTC,CSCO and PFE had the highest daily average volume of shares traded whereas UNH, UTX, GS, MMM and TRV had the lowest. 3M (MMM), UnitedHealth Group Inc (UNH) , Boeing Co (BA), Microsoft (MSFT) and Intel Corp (INTC) had the highest average daily return whereas DWDP, Merck (MRK), United Technologies (UTX), Pfizer (PFE) and Disney (DIS) had the lowest. AAPL had recorded the lowest daily return (-10.4%) whereas

XOM had the highest (8.34%) daily return on a particular trading day. DowDuPont Inc. (DWDP) was inducted into the ^DJI universe on 1st September 2017.

During this project, we encountered several problems such as: took some time to figure out how to visualize data using groupby objects and also cause there are too many companies (30), the visualization became too messy. The most confusing part of the ML is that after hyperparameter tuning, we didn't get much improve on the RMSE scores.

## DATASET COLLECTION/CLEANING/EXPLORATION

The dataset we used contains stock market data for 30 companies over a period of 5 years ("DOW30_5yr"). These 30 companies are a part of Dow Jones Industrial Average (DJI) ETF, which is a benchmark stock index that tracks 30 economy-leading blue-chip industrial and financial companies in the U.S. These stocks are on the Nasdaq and NYSE and are subjectively picked by the editors of The Wall Street Journal.The DJI is used in the media as a barometer of the broader stock market and the economy as a whole.

The DJIA is calculated by adding up all the stock prices of its 30 components and dividing the sum by the Dow divisor. However, the divisor is continuously adjusted for corporate actions, such as dividend payments and stock splits.

There are 36620 data points and 7 attributes in the dataset. The *Date* ranges from 2013/2/8 to 2018/2/7. The *Open*/*Close* indicates the opening/closing price for a stock, *High* indicates the highest price a stock attained, *Low* indicates the lowest price a stock attained, and *Volume* indicates the number of shares traded on that day. *Ticker* is an abbreviation for the company name. The dataset seems to be in accordance with good data science practices, and required no cleaning aside from turning the Date attribute into a DateTime object.

| | Date | Open | High | Low | Close | Volume | Ticker |
|---|---|---|---|---|---|---|---|
| 0 | 2/8/2013 | 67.7142 | 68.4014 | 66.8928 | 67.8542 | 158168416 | AAPL |
| 1 | 2/11/2013 | 68.0714 | 69.2771 | 67.6071 | 68.5614 | 129029425 | AAPL |
| 2 | 2/12/2013 | 68.5014 | 68.9114 | 66.8205 | 66.8428 | 151829363 | AAPL |
| 3 | 2/13/2013 | 66.7442 | 67.6628 | 66.1742 | 66.7156 | 118721995 | AAPL |
| 4 | 2/14/2013 | 66.3599 | 67.3771 | 66.2885 | 66.6556 | 88809154 | AAPL |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 36615 | 2/1/2018 | 87.5 | 89.25 | 87.35 | 89.07 | 17971012 | XOM |
| 36616 | 2/2/2018 | 85.13 | 86.01 | 82.9978 | 84.53 | 29822144 | XOM |
| 36617 | 2/5/2018 | 83.28 | 83.99 | 78.13 | 79.72 | 30452693 | XOM |
| 36618 | 2/6/2018 | 78.51 | 80.35 | 76.9 | 78.35 | 36262761 | XOM |
| 36619 | 2/7/2018 | 78.44 | 79.41 | 76.92 | 76.94 | 21994450 | XOM |

36620 rows × 7 columns

*Fig 2.1*

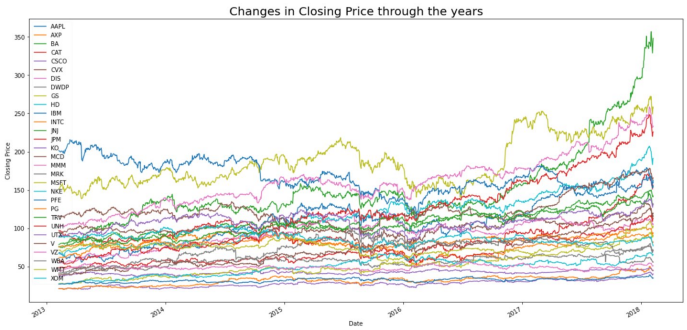



*Fig 2.2*


*Fig 2.3*


*Fig 2.4*

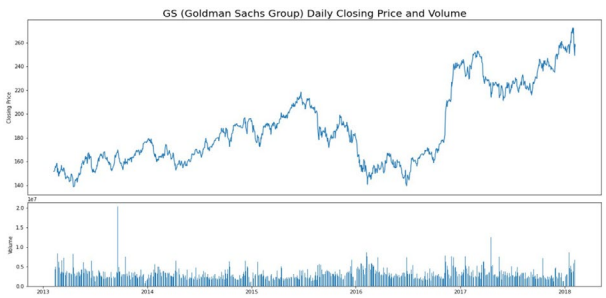



*Fig 2.5*

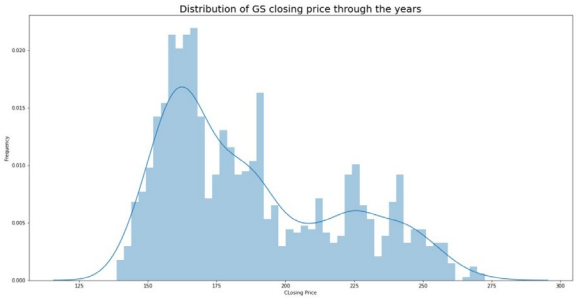



*Fig 2.6*

Fig 2.1 shows a brief view of the dataset.

We will predict the closing price for each company stock using the other attributes. The Fig 2.2 shows the changes in closing price of different companies through the years As seen below, it is difficult to follow each stock, and is difficult to compare how each one does compared to the others. This is because each stock has its own opening price. So for better understanding of the stock market, we divide each closing price with its first closing price in the

period as seen in the figure below. We see in the figure that JPM seems to be the best performing in general, until the end when BA overtakes it.

Fig 2.3 is easier to understand than the one before, but it is still difficult to read; Thus, we created the Fig 2.4. Fig 2.4 is a word cloud based on the average closing price of each company through the years. This shows that companies like IBM, BA, MMM, and GS have the highest closing prices and companies like MRK, VZ, and KO have the lowest average stock closing prices.

From the two visualizations above, we can see that GS (Goldman Sachs Group) is outstanding. So we choose GS as an example to visualize the daily closing price and volume. As seen in the Fig 2.5, stock market prices seem stochastic and there does not seem to be a simple correlation between volume and closing price.

Fig 2.6 shows the closing price of GS through the years. Interestingly enough, the closing price seems to follow a gamma distribution. This makes sense since gamma probability distributions most often occur in processes where there are waiting times between events. In the stock market, one must wait 24 hours between closing prices to get the next datapoint.

## METHODOLOGY

We are performing regression on the closing price of a stock. Each model is trained on 4 years of stock data and tested on the last one for each stock. Since each attribute is already consumable by each model, so there is no data wrangling involved. We also perform a gridsearch on each model to tune the parameters.

## MODELS

For our analysis, we chose to use 3 different regression models: random forest (RF), gradient-boosted tree (GBT), and linear regression (LR). Each model is trained on 4 years of stock data and is evaluated on the final year using a 7-fold cross validated RMSE score averaged over all of the 30 stocks.

## RANDOM FOREST

We chose to use a RF model since it works well in high-dimensional, high variance situations, with highly correlated features. This seems well suited for stock market data since there seems to be a lot of variance and the opening costs seems to be correlated with the closing costs.

In Fig 4.1.1, we show how close a fit RF is on one of the top performing stocks: AAPL. As seen in the figure, it stays close to the truth, but seems to only give a general approximation. Toward the end, the RF seems to perform extremely poorly.

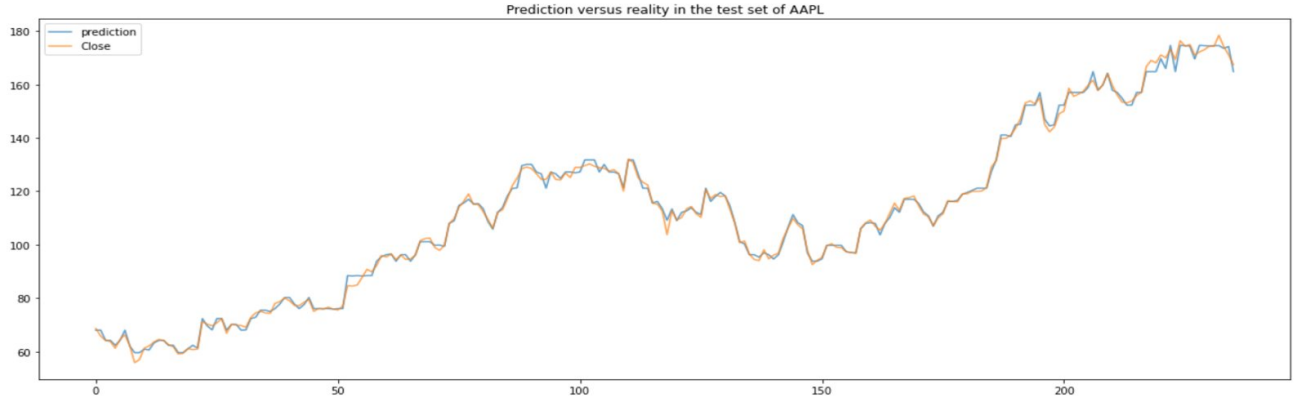



*Fig 4.1.1*

## Gradient Boosted Tree

We chose GBT because it is often considered a "better" version of random forests, but is better for data that is highly biased or imbalanced. As seen in the first figure, our data has a wide range of bias, so we believe this may do well at predicting closing price.

Fig 4.2.1 shows how the GBT performs on another top performing stock: MSFT. As seen in the figure, the GBT model seems to follow the truth better, but still performs poorly at the end.

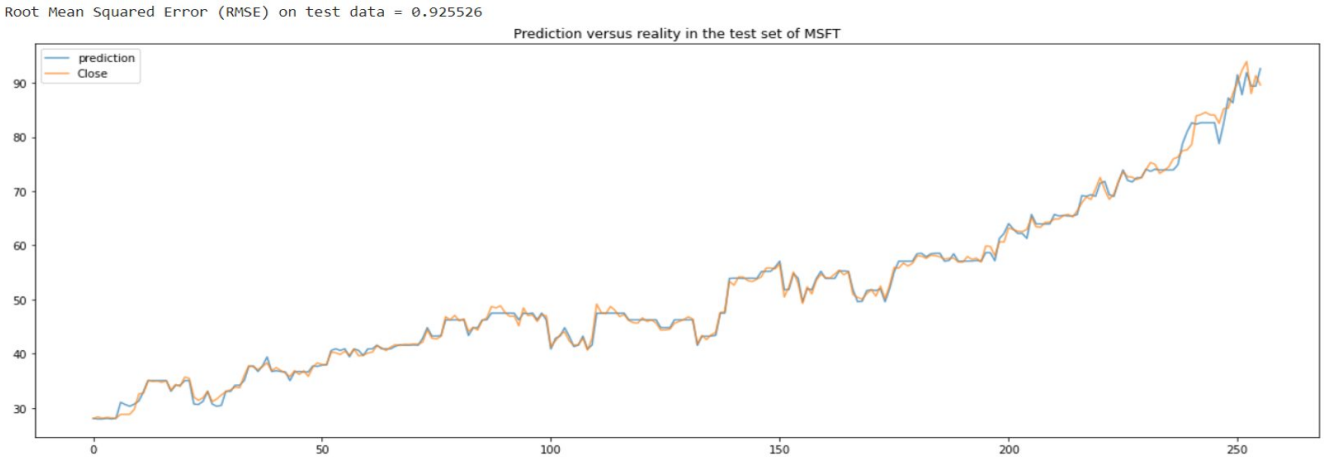
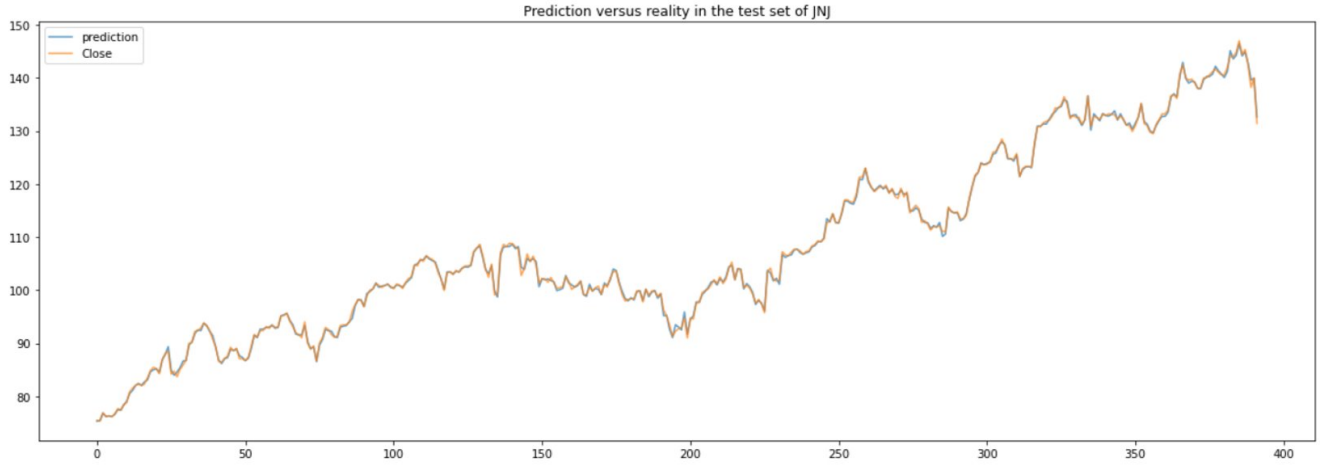



*Fig 4.2.1*

## Linear Regression

Finally, we chose to use LR because it works well on linear data, and the overall trends of closing prices tend to be linear. We suspect the LR model will also perform well since the data changes in a linear manner (straight lines between data points).

The LR model prediction is shown in the following figure on another top performing stock: JNJ. We see that LR does a great job at predicting the closing price of a stock, and even captures the sudden fall toward the end of the time sample.



We note that the sharp fall in the closing price at the end of the analysis period might be due to micro or macroeconomic factors affecting the company.

### MOVING AVERAGES

We will also explore the idea of the moving average (MA). MA is a technique often used in Time Series analysis that produces a generally reliable, "smooth" forecast. Stock traders often use this as a flag to buy or sell. It works by averaging the inputs over a time window. We also note that the idea of MA can also be applied to a moving STD, which can represent the volatility of a stock.

## CONCLUSION

Each of the prediction models performed well. As seen in the table, LR outperformed the others by a large margin and GBT outperformed RF. This seems to indicate that the closing price is far more linearly dependent on the other attributes than initially suspected. The fact that GBT performed better than RF also seems to indicate that the data suffers more from high bias than high variance.

| | RF | GBT | LR |
|---|---|---|---|
| RMSE | 1.44697 | 0.9195 | 0.3823 |

The RMSE scores for each stock with each model is visualized below. We note that while LR is the best overall, it is not the best model for each stock. Namely, the VZ and the DWDP stocks are better modeled with a GBT model than a LR model. It is also interesting to note that BA and HD are far better modeled by LR than any of the others.

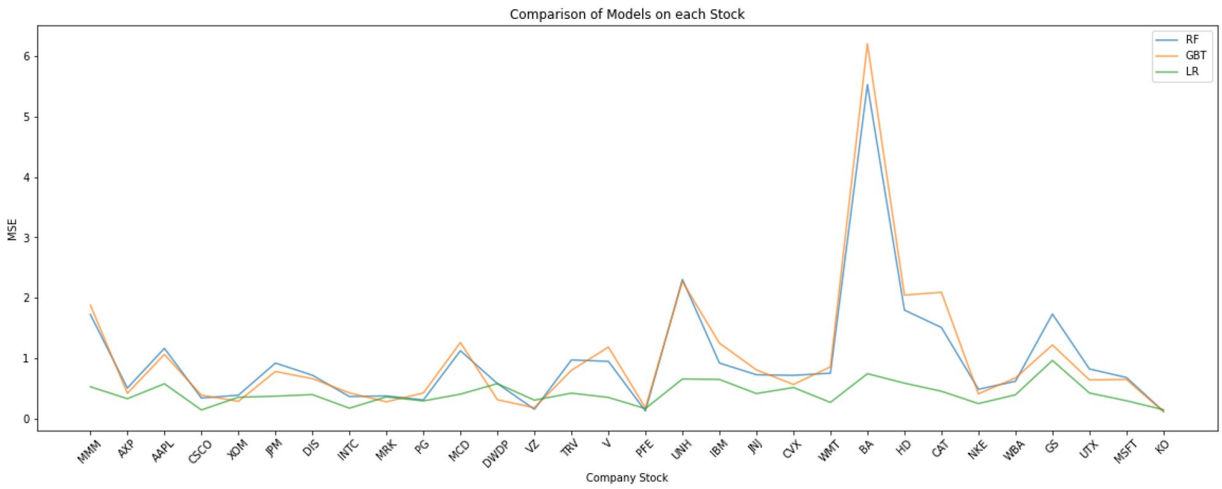



*Fig 5.1*

In Fig 5.2, we chose 5 top performing companies and plotted their MA forecast together. As compared with the raw stock data, it is much easier to read and to notice when a stock is performing well and when it is not. It is also easier to see when a stock overtakes another, such as AAPL and JNJ in the figure below. We see many flags to buy when the MA of a stock has a positive slope, and to sell when its slope is negative.

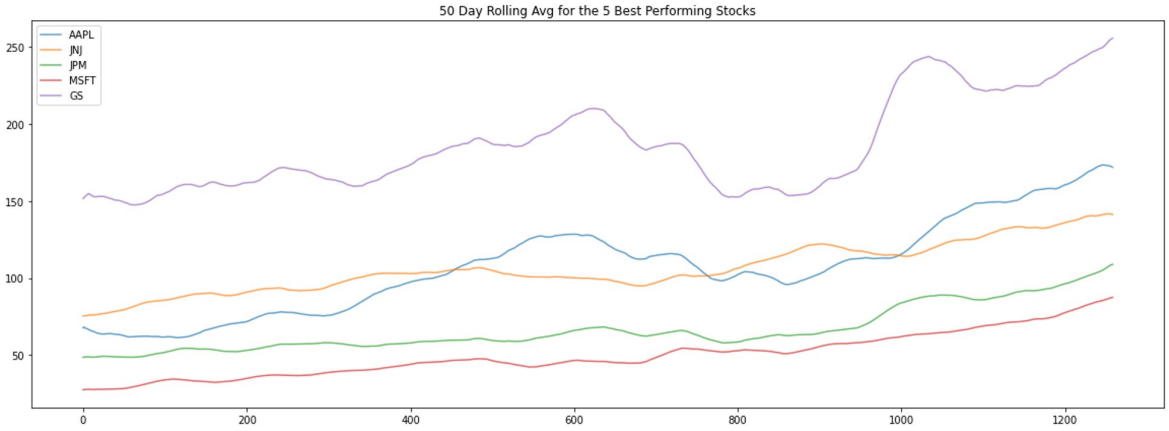



*Fig 5.2*

For the same 5 companies, Fig 5.3 shows the moving STD, or volatility. As seen in the figure, AAPL seems to be the most volatile and JNJ seems to be the least. We see a direct correlation with the GS stock around the 1000 time step where it is considered highly volatile in the figure below, but we see a huge growth in the figure above.

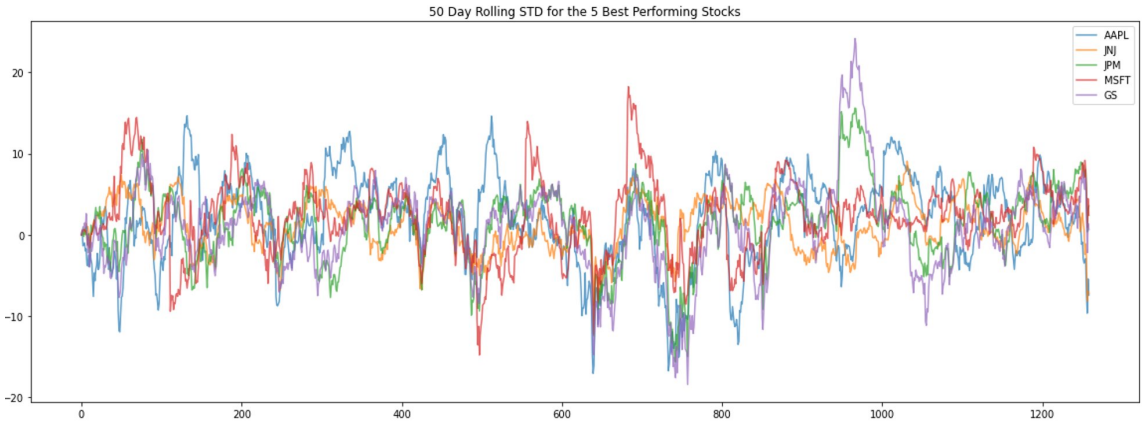



*Fig 5.3*

Finally, we can see with Fig 5.4 that using the MA signal, we could have predicted the price trend of JNJ. When short-term crosses above long-term we get a 'buy' signal. When short-term passes below the longer-term we get a 'sell' signal. This is a visual testimony of how simple, but effective this basic algorithm is.


Fig 5.4

The stock market is widely popular as a source of long-term and short-term investments. But due to its stochastic nature, it is difficult to predict when a stock will rise or fall in value. Through our analysis, we found that the volume of traded stock shares is not highly correlated with closing price. We also found that among the 30 DJI companies, AAPL, MSFT, INTC, CSCO, and PFE had the highest daily average volume of shares traded whereas UNH, UTX, GS, MMM, and TRV had the lowest. Additionally, 3M (MMM), UnitedHealth Group Inc (UNH) , Boeing Co (BA), Microsoft (MSFT) and Intel Corp (INTC) had the highest average daily return whereas DWDP, Merck (MRK), United Technologies (UTX), Pfizer (PFE) and Disney (DIS)

had the lowest. Finally, the AAPL was very volatile and had recorded the lowest daily return (-10.4%) whereas XOM had the highest (8.34%) daily return on a particular trading day.

We were able to build prediction models using random forests, gradient-boosted trees, and linear regression. We found that linear regression far out-performed the other models in predicting the closing price of a stock. This means that closing price is linearly correlated with the other stock attributes. We found that GBT performed better than RF, which indicates that the stocks suffered more from high bias than high variance.

# APPENDIX

## Bibliography

"DOW30_5yr." *DOW30_5yr*, https://drive.google.com/file/d/12AlXVGyI289zs7k-1Vuy6HEecTesFYSa/view?usp=sharing. Accessed 1 11 2020.