Machine Learning Algorithms for Stock Market Prediction

Chae Won Lee[#1], Kyle Davis[#2], William Jones[#3AI]

[#]*College of IST, The Pennsylvania State University*
University Park, PA 16802
[1]cvl5801@psu.edu,[2]ked5465@psu.edu,[3]wdj5029@psu.edu

Abstract

Predicting stock market trends is a very difficult task, and many people have attempted to do so and failed. There are many external factors causing the difficulty of this task, such as recessions, pandemics, and other unforeseen external factors that can influence a stock. On top of this, it seems as though the psychology of investors significantly influences their decisions, and therefore, the stock market. Investors have looked to machine learning models to try and help remedy the issue of poor predictions [1]. Time and time again people have tried and failed in using historical stock data and machine learning algorithms to predict future prices and trends. A newer trend has seemingly emerged, trying to use public sentiment to predict stock market trends. In this area however, there have also been many shortcomings. In this study we will explore the performance of many popular machine learning algorithms in predicting the changing price of stocks. The machine learning algorithms we will use are Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost) and Linear Regression (LR). We will use data obtained from Yahoo Finance, covering from January 2017 to January 2022 of stock data for Apple, AMD, Facebook, Ford and Snapchat. We will then try to use subreddit data to create a feature based on the frequency of certain keywords that are related to the stocks we have chosen.

Keywords— stock market, machine learning, sentiment analysis

I. Introduction

Stock markets worldwide all affect and influence each other, and this affects each country's individual economy as well as the world economy as a whole. The stock market has a huge effect on individuals' lives as well as world government entities. When something major happens economically in any major country, it has an effect on the entire world. For example, the housing crash in 2008 had rippling effects felt worldwide. It led to stock market crashes globally, people losing homes, and fortunes. Understanding stock market trends and being able to predict the future trends can help governments, individuals and companies alike not only make a lot of money, but also give stability.

There are many factors that influence the stock market. Many of these factors are outside the companies control, making it very hard to predict. These could be local, domestic or foreign factors that affect stocks of companies, especially global ones. These factors could be anything from political uncertainty, change in interest rates, natural disasters and so much more, the list is quite literally endless. On top of these many internal factors go into it, such as earning reports, change in management, mergers, and again this list is quite long as well. This makes it near impossible to combine all the factors that can affect the stock market.

There are also more factors that have to do with the investors. The psychology of investors has a major effect on the stock market. If investors think the stock market will boom, they will invest, driving prices up. If the investors think the stock market will crash, they will sell, again having a cause and effect relationship. Something as simple as a CEO making a politically charged tweet could drive a stock down. All of these factors make the stock market highly complex and difficult to predict.

To try and include some of the outside forces affecting the stock market we will also be including an extra feature built from posts on Reddit. We will be looking for mentions of stocks on some financial based subreddits. From these mentions we will get a frequency by date to match up with our benchmark suite models. Hopefully this will improve our models performance, but if not we will discuss why.

II. Literature Review

In the financial industry, being able to predict future stock trends is very important. To be able to get accurate predictions can allow someone to make lots of money. The stock market is affected by so many people and so many different things. Strictly speaking it has buyers and sellers of the stocks, and companies of which the stock is for. The influence on a stock goes far outside these basic parameters. To be able to get a precise prediction, one needs to try and take into account different key elements that influence a stock.

Many people have tried in the past to predict future stock prices and trends. Everytime some sort of new technological way of modeling comes out, it seems to be tested against the stock market. The main types of stock market prediction fall into four categories: "(i) *Fundamental Analysis*, (ii) *Technical Analysis* (iii) *Prediction of Time Series with Traditional Models* and (iv) *Machine Learning Methods*" [1].

Fundamental analysis uses statistics, projections and fundamentals of the economy to predict stock prices. This method measures security's intrinsic value by examining different factors that affect the value of the security [2]. These factors can be macro or microeconomic, that affect the economy of the industry or the specific company's issues.

Technical analysis studies historical market data, such as the high, low, volume, open and close. In technical analysis they typically use chart patterns and technical indicators [3]. The assumption of technical analysis is that the historical stock information tells you everything you need to know. Using past trends to predict future trends does not work terribly well on its own when it comes to the stock market. Many believe that the stock market reacts almost at random.

Prediction of time series with traditional models is pretty much how it sounds. They use linear models to track patterns in historical data [1]. This is very similar to technical analysis, just using time series. Time series is a collection of data from observations made over time. Usually it is repeated measurements over time, which essentially create points on a graph that are based on the X axis of time [4]. The Y axis can represent any variable that can be measured through time, in this case the price of the stock. These models are broken into "univariate models and multivariate regression models" [1].

The fourth and final method is using machine learning. This again is self-explanatory, using time series data, and machine learning models, one can try and predict future stock trends. Machine learning models are meant to find patterns, and use them to predict future ones. In this case it is no different, using historical stock data a machine learning model can be applied to see if future trends can be found.

Within the financial and science industries, there are two schools of thought about stock prediction. The first group believes it is possible to predict market behavior, the other believes in the efficient market theory, which says "the price of stock follows a random (unpredictable) walk" [1]. This essentially says that prediction of stocks using patterns (historical data) is not possible. This is because they believe that there is no way for one to incorporate all the data/information that the market does.

People have begun to try and incorporate outside information (not stock data) to try and predict stocks. The main technique used here is sentiment analysis. This technique attempts to incorporate public sentiment into models by using twitter or other public social media outlets. For example, if one tried to predict the price of Microsoft's stock, they could get all the tweets that mention Microsoft. They then assign the tweet a score, saying that the tweet is a positive tweet or a negative tweet. Using this information on all the tweets they have, they model it to see if there is a comparison to be made with the stock price. They can use machine learning models to see how accurately they can predict stock market shifts based off of public sentiment.

III. Data Collection

In this study, five years of historical data of the top five most traded stocks on NASDAQ were obtained from Yahoo Finance from January 2017 to January 2022. The top five stocks include Apple, AMD, Facebook, Ford, and Snapchat. The data includes 7 features: Date, Open, High, Low, Close, Adj Close, and Volume. Out of these 7 features, we mainly focused on Date and Adj Close price. We have tried to calculate the net price movement using this formula: (Close price - Open price) to set the direction of the price movement. It would show whether the closing price of the stock went up, down, or no change from the previous trading day. This price movement feature would create a numerical feature where 1 denoted up, -1 denoted down, and 0 denoted no change. However, we later realized that this price movement feature was not needed in our study. The Adj Close feature factors in several actions which include stock splits and dividends. It is useful when studying a detailed analysis of historical data. [6] Therefore, we mainly focused on Adj Close and Date features to predict the future stock prices.

We then obtained the data of r/wallstreetbets posts using the Pushshit API. The time frame is the same as the stock data we got for the stocks excluding weekends and holidays. We filtered out the posts using subreddit of "wallstreetbets", "stocks","investing","stockmarket","finance","security analysis","financial independence" to only include posts that are relevant to our study. With these posts, we converted the body to lower case, removed URLS and HTML tags, and removed punctuations for better word embeddings. After filtering out the posts, we were left with around 600 posts. We then counted the frequency of each stock that was mentioned in the posts. After preprocessing the reddit posts and calculating the frequency, we combined them with the stock data that we obtained from yahoo.

IV. Methods

For our methods, we started by creating a benchmark suite with the data we collected from the API. Our machine learning models we chose to use Logistic Regression, Support Vector Machine (SVM), and XGBoost.

A. Logistic Regression

Logistic Regression is a statistical analytic approach that uses past observations of a data set to predict a binary result, such as yes or no. A logistic regression model predicts a dependent data variable by understanding the correlation between one or more pre-existing independent variables. A logistic regression, for example, might be used to forecast whether a political candidate will win or lose an election, or if a high school student would be admitted or not to a specific institution. These binary outcomes allow for simple choices between two choices.

B. SVM

Support Vector Machine (SVM) is a supervised learning model for pattern recognition and data analysis. It is primarily used for classification and regression analysis. SVM can be applied to both linear and nonlinear classification problems. Given a set of data from either category, the SVM algorithm creates a non-stochastic binary linear classification model that determines which category the new data belongs to based on the given data set. The developed classification model is represented as a boundary in the space where the data is mapped, and the SVM algorithm is a search algorithm that finds the boundary with the greatest width.

C. XGBoost

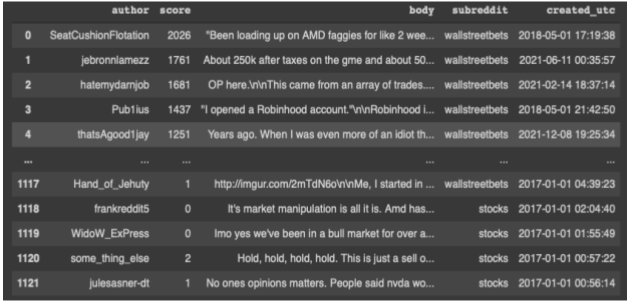
XGBoost is a machine learning method for tree boosting that uses a tree-based (GBDT) to solve many data science problems quickly and accurately. It is a distributed gradient boosting library optimized for efficiency, flexibility, and portability, and it implements machine learning algorithms within the Gradient Boosting framework. One advantage would be a Flexible Learning System, which allows for the creation of an optimal model by adjusting various parameters. When compared to neural networks, visualization is simpler and more intuitive.

We decided to use 65% for training data, 15% for testing data, and 20% for the final validation set, so we don't overfit our data. Finally, we calculate the Root Mean Square Estimation (RSME) for each of the stocks, based on the models we used. This will serve as our baseline score for how well models performed solely off of API data we obtained,

D. Subreddit Feature

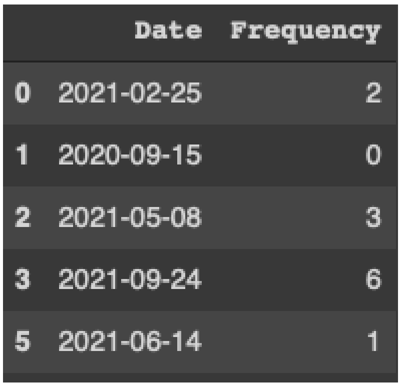
After creating our benchmark, we move to creating our subreddit feature. The subreddits will be used to determine how much each of the stocks have been talked about during the five year period. From the frequency of how much the stocks are being talked about, we hope to see any significant changes to the model when implementing the feature back into our models.

Using the PushiftAPI and the different subreddits we chose to use (wallstreetbets, stocks, financialindependents, stockmarket, etc.), we get a query list of the different comments from each of the subreddits, as seen below.



	author	score	body	subreddit	created_utc
0	SeatCushionFlotation	2026	"Been loading up on AMD faggies for like 2 wee...	wallstreetbets	2018-05-01 17:19:38
1	jebronnlamezz	1761	About 250k after taxes on the gme and about 50...	wallstreetbets	2021-06-11 00:35:57
2	hatemydarnjob	1681	OP here.\n\nThis came from an array of trades....	wallstreetbets	2021-02-14 18:37:14
3	Pub1ius	1437	"I opened a Robinhood account."\n\nRobinhood i...	wallstreetbets	2018-05-01 21:42:50
4	thatsAgood1jay	1251	Years ago. When I was even more of an idiot th...	wallstreetbets	2021-12-08 19:25:34
...	...	...	...	...	...
1117	Hand_of_Jehuty	1	http://imgur.com/2mTdN6o\n\nMe, I started in ...	wallstreetbets	2017-01-01 04:39:23
1118	frankreddit5	0	It's market manipulation is all it is. Amd has...	stocks	2017-01-01 02:04:40
1119	WidoW_ExPress	0	Imo yes we've been in a bull market for over a...	stocks	2017-01-01 01:55:49
1120	some_thing_else	2	Hold, hold, hold, hold. This is just a sell o...	stocks	2017-01-01 00:57:22
1121	julesasner-dt	1	No ones opinions matters. People said nvda wo...	stocks	2017-01-01 00:56:14

We then go about cleaning the data, including making the comments all lowercase and removing any unreadable characters to make it easy to find the keywords of each of the stocks. Next, we ran a search of the comments of each different stocks we have based on the keywords of the stock, i.e. Snapchat usually goes by “Snap” and Facebook can go by “FB”. From this, we gain the frequency of each comment of stocks and we later join the frequency of the stocks based on the date of each subreddit comment was created.



	Date	Frequency
0	2021-02-25	2
1	2020-09-15	0
2	2021-05-08	3
3	2021-09-24	6
5	2021-06-14	1

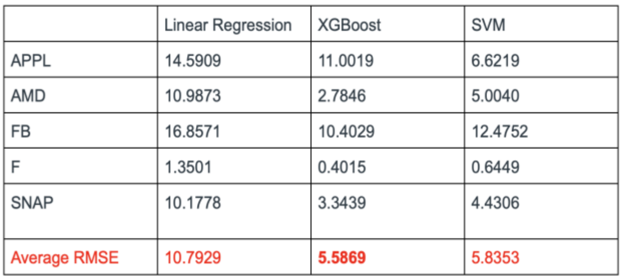
Now that we have our frequency via the subreddits, we then move the data back over to our original model to implement the models using the subreddit feature, using the frequency of the comments and the Adj Close values of the stocks all based on the corresponding date. Finally, after running the same models over, now including the subreddit feature, we hope to get a different RMSE score than the first time to see if the feature does affect the score.

V. Results

The RMSE score was used as an evaluating metric for the models. RMSE can be calculated by the following formula. Y-hat is the predicted value, y is the true value, and n is the sample size.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

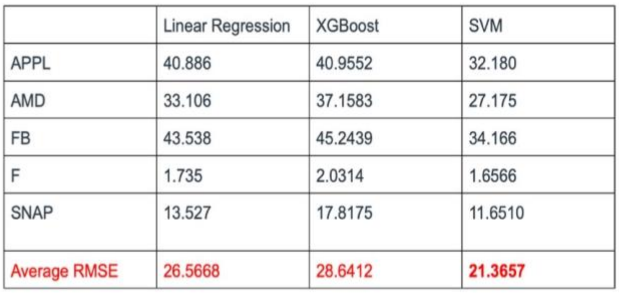
After the individual RMSE scores were calculated for each of the three algorithms for each of the 5 stocks, the final conclusions were based on the Average RMSE score. The results can be found in Table 1. The result shows that XGBoost performed the best RMSE score of 5.5869.



	Linear Regression	XGBoost	SVM
APPL	14.5909	11.0019	6.6219
AMD	10.9873	2.7846	5.0040
FB	16.8571	10.4029	12.4752
F	1.3501	0.4015	0.6449
SNAP	10.1778	3.3439	4.4306
Average RMSE	10.7929	**5.5869**	5.8353

Table 1

After implementing the subreddit feature to our model, the results can be found in Table 2. The result shows that SVM performed the best RMSE score of 21.3657.



	Linear Regression	XGBoost	SVM
APPL	40.886	40.9552	32.180
AMD	33.106	37.1583	27.175
FB	43.538	45.2439	34.166
F	1.735	2.0314	1.6566
SNAP	13.527	17.8175	11.6510
Average RMSE	26.5668	28.6412	**21.3657**

TABLE 2

VI. CONCLUSION

As stated in the beginning, the stock market is one of the most complex systems on the planet. Its constant fluidity and changes that can happen almost unexpectedly makes it next to impossible to consistently predict the best possible outcome for anyone. We understand that this task is much deeper than we can imagine. Our study is more of an idea and question of "Can machine learning models help predict stock prices from more than just a quantitative means?". This is where the subreddit feature comes into play. People talking and sharing different information about the stock market can have an impact on how the stocks are traded. It's no different to a CEO of a Fortune 500 company sending out a controversial tweet on Twitter and the stocks of the company suffer from it. Overall, based on our findings and our best efforts, we feel that the stock market trading can be affected by subreddit data and comments.

VII. FUTURE WORK

There are definitely some improvements that can be made in the future for our project. Firstly, we would want to explore exactly how the different models were affected by the subreddit feature. XGBoost did the best with the sole historical data we provided, however adding the subreddit feature, it shows XGboost performed the worst of the three. Another aspect of the project we hoped to tackle later is using non-numerical features in our models to see how that would affect the RMSE results. Mainly, using positive/negative sentiment analysis through Twitter posts and implementing it into our models.

With more time and resources allotted for our team, these would be the steps we take to better understand our models in relation to predicting stock prices.

REFERENCES

[1] Fagner A. de Oliveira, Cristiane N. Nobre, Luis E. Zárate, "Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index – Case study of PETR4, Petrobras, Brazil,Expert Systems with Applications" Volume 40, Issue 18, 2013, Pages 7596-7606,

[2] Segal, T. (2022, February 8). *What is fundamental analysis?* Investopedia. Retrieved April 14, 2022, https://www.investopedia.com/terms/f/fundamentalanalysis.asp

[3] Chen, J. (2022, February 8). *Technical analysis of stocks and Trends*. Investopedia. Retrieved April 14, 2022, https://www.investopedia.com/terms/t/technical-analysis-of-stocks-and-trends.asp

[4] *What is time series data?: Definition, examples, types & uses*. InfluxData. (2022, March 25). Retrieved April 14, 2022, from https://www.influxdata.com/what-is-time-series-data/

[5] Content. (2022, February 15). *All the factors that affect stock market*. ABC of Money. Retrieved April 14, 2022, from https://www.adityabirlacapital.com/abc-of-money/factors-affecting-stock-market

[6] Ganti, Akhilesh. "Adjusted Closing Price Definition." *Investopedia*, Investopedia, 19 May 2021, https://www.investopedia.com/terms/a/adjusted_closing_price.asp.