Final Project Report for Stats 170B, Spring 2021

Project Title: Stock Price Prediction

Student Names
Yu Miyauchi, 23233603, ymiyauch@uci.edu

Sean Imai, 62603849, saimai@uci.edu

Hien Huynh, 17782540, huynhhp@uci.edu

Github: https://github.com/Yu821/stock_price_prediction

1. Introduction and Problem Statement (1 or 2 paragraphs)

Nowadays, investors often spend a large amount of money trading stocks, so the stock price is always one of the top concerns. If price can be predicted correctly, it will bring great success. And vice versa, it will cause loss of money, sometimes even lead to bankruptcy. Therefore, stock analysis has become a large business and tends to grow - at the increasing demand of investors - because the market is developing more diversified and more complex. Our project objective is to predict price and make an alert system to notify traders to buy, sell, or hold stocks in the technology industry in the short term (less than a year ahead). We specifically chose the technology field because of the continuous development and the important role of it in the modern world.

Basically, our technical method is similar to professional investors and companies. We predict the stock prices by looking at the trading data of a stock in the past. Technical investment experts always believe that all information about a company is represented by its stock price. Therefore, we will use the stock news, market news, the financial statements, the historical daily stock price and the historical daily stock volume for analyzing and predicting in this project. All the datasets we use can be extracted from IEX Cloud API (https://iexcloud.io/docs/api/).

2. Related Work: (1 or 2 paragraphs)

[1] Strader, Troy J.; Rozycki, John J.; ROOT, THOMAS H.; and Huang, Yu-Hsiang (John) (2020) "Machine Learning Stock Market Prediction Studies: Review and Research Directions," Journal of International Technology and Information Management: Vol. 28 : Iss. 4 , Article 3.

We were able to utilize the article to create the baseline for what kind of machine learning models we could use for predicting stock price and what others have tried in the past. The authors in the article conduct a peer review of four relevant articles in which they try and “identify directions for future machine learning stock market prediction.” Since the problem we are trying to predict is extremely complex, we felt it would be best to learn from others than to try and reinvent the wheel.

This study introduces algorithms that are relevant to our project which are neural networks (NN), support vector machines, and genetic algorithms. Additionally, the study does delve into the subject matter of the importance of financial investment theory being a “strong driver underlying ML systems’ inputs, algorithms, and performance measures.” Being that the system needs to stand the test of time, and tested in multiple different scenarios such as different risk and volatility environments.

[2] Indexing, Ijar. “Stock Market Prediction Using Neural Networks and Sentiment Analysis of News Articles.” *Www.academia.edu*, www.academia.edu/32423873/Stock_market_prediction_using_Neural_Networks_and_sentiment_analysis_of_News_Articles. Accessed 20 Apr. 2021.

This article was able to be used in order to understand what factors are generally used in order to analyze and predict a stock prices movement. We were able to use this information in order to formulate what data we would need to collect. From the article, the prediction of a stock's price is formulated using fundamental factors, technical factors, and market sentiment.

3. Data Sets [at least 1 page]

There are five datasets we used: historical prices, balance sheet, income statement, market news and stock news. The data for the first three datasets were extracted utilizing the IEX Cloud API (https://iexcloud.io/docs/api). The news for the last two datasets were extracted using the Datanews API (https://datanews.io). Data was collected for ten technology stocks which include: Apple, Facebook, Adobe, Intel, Paypal, Alibaba, Qualcomm, Shopify, AMD, and Microsoft. We chose to specifically choose stocks pertaining to the same market sector, since the price movement of stocks in the same sector tend to behave similarly.

Historical Prices

Historical prices dataset describes the prices of our ten technology stocks from the past 5 years (from 04-04-2016 to 04-04-2021). This dataset has 25 columns and 12588 rows. Some important features which are useful for our prediction are the symbol of the companies’ tickers, the day of trading, closing price (the price at the end of the day), the open price (the price starts at the beginning of the day), the highest/lowest price of that stock within that day, and the volume of trading on that day.

label	change	fClose	fOpen	fVolume	symbol	change	fHigh	fLow
2021-04-01	0.00	123.00	123.66	75089134	AAPL	0.00	124.18	122.49
2021-03-31	-0.85	122.15	121.65	118323826	AAPL	-0.85	123.52	121.15

Table 3.1: Sample of historical prices dataset .

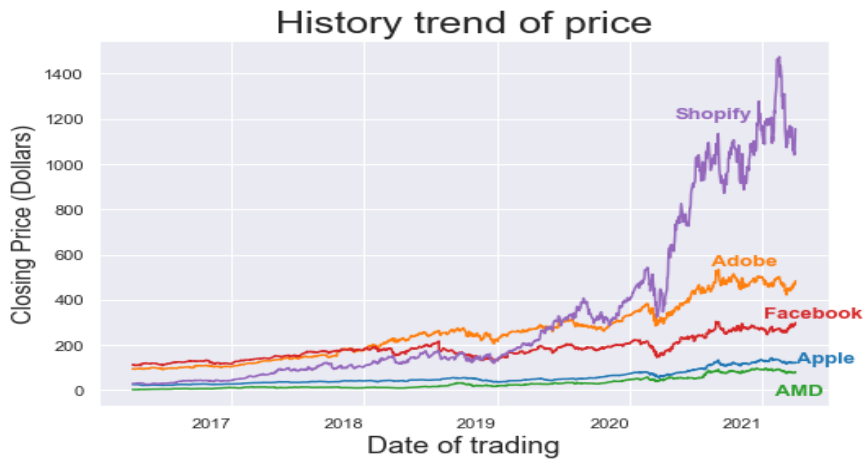

Fig 3.1: The time series within 5 years of 5 stocks: AAPL, SHOP, FB, AMD, ADBE

The fig 3.1 shows the trend of stock prices of 5 stocks from the past 5 years. Most of the 10 stocks we analyze present the increase in price. This makes sense because the technology field is very important in the modern world. There is a light decrease at the beginning of 2020, it can be caused by the starting of the pandemic. But after that, we observe the strong increase in price of stocks, especially the graph of Shopify. It can be explained by the demand of people in technology while they have to work at home and shop online.

Balance sheet

Balance sheet dataset describes the assets, loans, and funds of companies from 2016 to 2021. It has 38 columns with 235 rows. Some important variables are total assets, long term debts, long term investments, capital surplus, current assets, and fixed assets. Capital surplus is all the money collected from investors that companies haven't used. Current assets are the assets that can be converted to cash within one year such as deposits while fixed assets are the assets that are hard to convert into cash within one year such as companies' buildings.

	key	fiscalYear	fiscalQuarter	totalAssets	longTermDebt	longTermInvestments	intangibleAssets	otherAssets	totalLiabilities
0	AAPL	2021	0.0	3.540540e+11	9.928100e+10	1.999480e+11	0.0	4.327000e+10	2.878300e+11
1	AAPL	2020	7.5	3.238880e+11	9.866700e+10	1.801750e+11	0.0	4.252200e+10	2.585490e+11
2	AAPL	2020	5.0	3.173440e+11	9.404800e+10	1.772790e+11	0.0	4.100000e+10	2.450620e+11

Table 3.2 Examples of balance sheet data

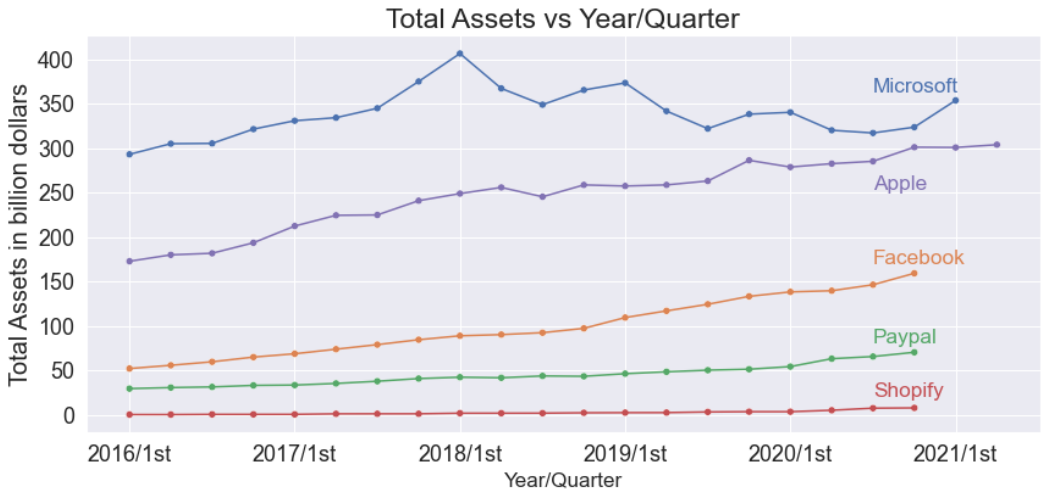

Fig 3.2. The time series data of total assets in five different technology companies

From figure 3.2, we are able to see that all five technology companies have an increasing trend. Also, Apple is the one which has the largest slope. Interestingly, Shopify had the most increasing trend in stock price dataset (fig 3.1), but its asset hasn't increased yet.

Income Statement

Income statement dataset mainly describes the revenues and expenditures of companies from 2016 to 2021. It has 27 columns with 235 rows. Some important variables are total revenue, gross profit, net income, operating income, and operating expenses. Total revenue is the total amount of sales while total income is the total amount of profits. Operating income is the amount of income only from sales while operating expenses are expenses produced by sales such as labor costs.

	key	fiscalYear	fiscalQuarter	netIncome	netIncome	operatingIncome	operatingExpense	grossProfit	incomeTax
0	AAPL	2021	0.0	2.875500e+10	2.875500e+10	3.353400e+10	7.790500e+10	4.432800e+10	4.824000e+09
1	AAPL	2020	7.5	1.267300e+10	1.267300e+10	1.477500e+10	4.992300e+10	2.468900e+10	2.228000e+09
2	AAPL	2020	5.0	1.125300e+10	1.125300e+10	1.309100e+10	4.659400e+10	2.268000e+10	1.884000e+09

Table 3.3 Examples of income statement data

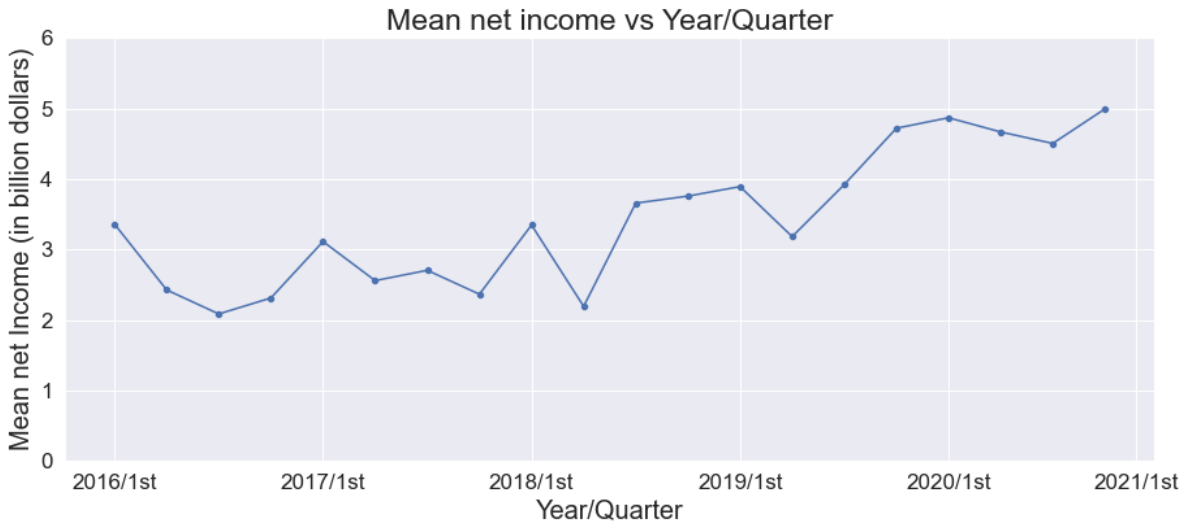

Fig 3.3. The time series data of mean total income

From figure 3.2, we are able to see that mean total assets has an increasing trend. Also, we notice that its value increased rapidly after the 1st quarter of 2019 and continue to have an increasing trend toward 2021. It seems that the impact by COVID-19 is limited.

Market/Stock News

	ticker	headline	date	score
0	AAPL	Dow Futures Soar as New York Begins Winning th...	2020-04-06	-0.1280
1	AAPL	Apple Snaps Up AI Startup Inductiv As Analysts...	2020-05-28	0.4019
2	AAPL	Gates Foundation Buys Up Amazon Apple Twitter ...	2020-05-25	0.1779

Table 3.4. Example of Market/Stock News dataset

The market news dataset includes the headline, publish date, and sentiment score of multiple news articles. We collected roughly one hundred articles per month for the past year, April 1 of 2020 to April 1 of 2021. The sentiment score for each article was calculated using the VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon and rule-based sentiment analysis tool that is specifically designed to extract sentiments expressed in social media. The sentiment score ranges from -1 to 1, 1 being very positive, 0 being neutral, and -1 being very negative.

The stock news dataset contains all of the same details of that in the market news dataset, but a ticker label is attached to each article, for which company the article relates to. Additionally, one hundred articles were collected for each stock, per month in the same time frame as the market news.

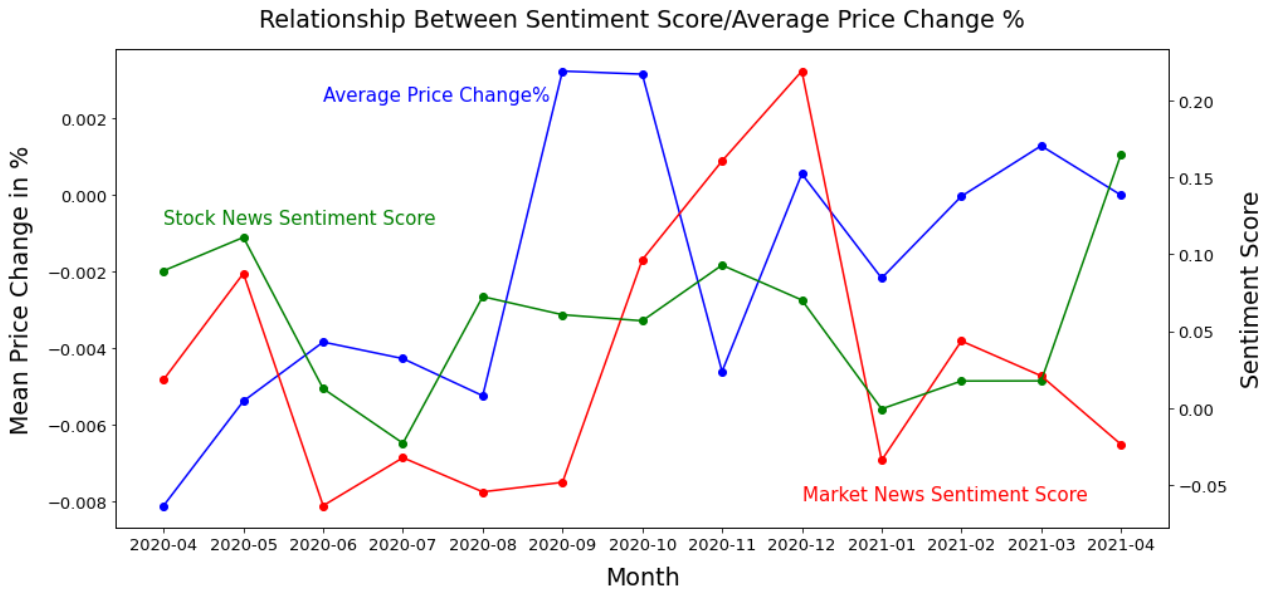

Fig 3.4. Time series data of change in market price and sentiment score

This plot shows the relationship between the sentiment of the stock news and market news with the mean change in price as a percentage of all 10 stocks month by month. Now if we look at the plot some key insights we gathered were how the change in price is sometimes affected by the market news or stock news depending on which sentiment has changed the most. Additionally, there is sometimes a delay in the change in price, from when a major spike in sentiment occurs. This chart is no surprise since the nature of the movement in a stock price is often never based purely on good or bad news, but on a number of other reasons. Oftentimes you can even see the price of a stock drop given good news.

Aggregated data

We aggregated stock price, income statement, and balance sheet by first aggregating income statement and balance sheet on its fiscal year and quarter. Then, we aggregated it with the stock price on the dates after those statements were released. About the news dataset, we aggregated them based on the date that news was released. In more detail, we put each sentiment score on one week of price dataset after the news was released.

Here is the correlation map of income statement and balance sheet.

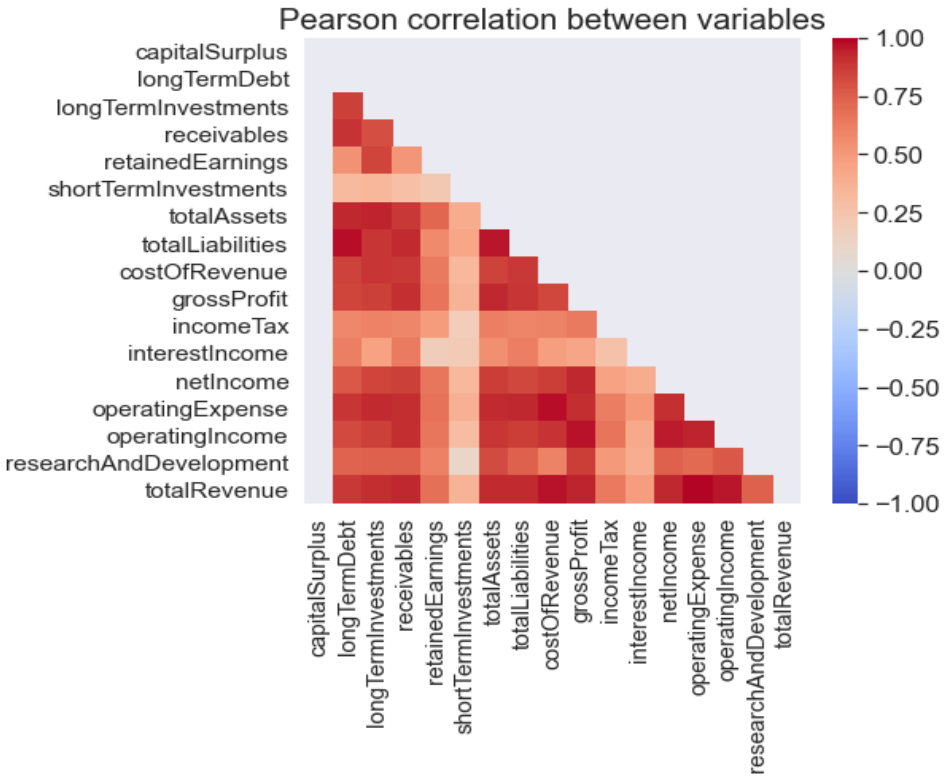

Fig 3.5. Correlation map of income statement and balance sheet.

From this figure, we can see that most variables have positive correlations on each other. Interesting point is that net income and long term debt and net income have positive correlation. This may be because companies invest money on technologies to earn more incomes.

4. Overall Technical Approach [at least 1 or 2 pages]

Data Management

We took two approaches to store the dataset. The first one is AWS Shared Postgres Database. There are two reasons. The first reason is that it is safer to store the dataset in the cloud as it backups the database. The second reason is that, since the dataset is shared with teammates, we are able to see changes made to our dataset and having a shared database reduces work to upload the dataset in each one of our computers. The second way to store data is local files. This is because it may be more handy to have a dataset in the file format.