STOCK PRICE PREDICTION USING MACHINE LEARNING

by:
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 to the firm or individual. Vice versa, if implemented wrong could cause loss of profit or even lead to bankruptcy. Therefore, stock price analysis has become a large business and tends to grow - at the increasing demand of investors and technological resources to do so - because the market is developing more diversified and more complex. Our project's objective is to predict the change in stock price of ten technology companies one day and five days ahead. We specifically chose all of the stocks to be related to the technology field since stocks in the same industry tend to have similar price movements.

Our technical method is similar to professional investors and firms. We predict the stock prices by looking at the trading data of a stock's past. Technical investment experts always believe that all information about a company is represented by its stock price, sentiment, and market sentiment. Therefore, we will use a stock's news, market news, financial statements, historical stock price for analyzing and predicting a stock's price movement in this project.

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.

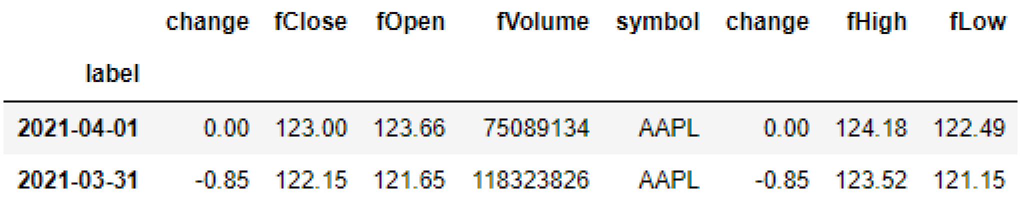
This article was able to be used in order to understand what factors are generally used in order to analyze and predict a stock prices movement. We were able to use this information in order to formulate what data we would need to collect. From the article, the prediction of a stock's price is formulated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

3. Data Sets

There are five datasets we used: historical stock 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 tends to behave similarly.

3.1 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 are 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.

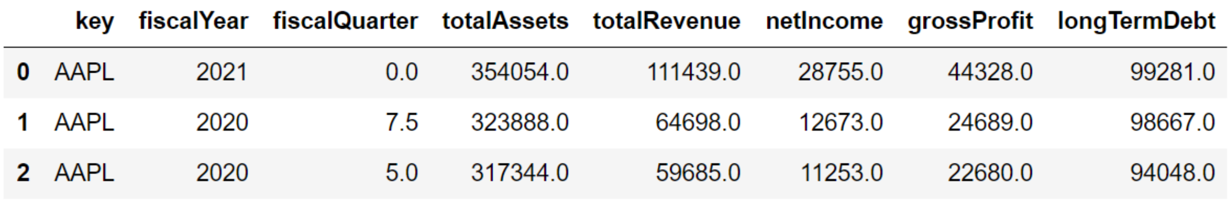


	change	fClose	fOpen	fVolume	symbol	change	fHigh	fLow
label								
2021-04-01	0.00	123.00	123.66	75089134	AAPL	0.00	124.18	122.49
2021-03-31	-0.85	122.15	121.65	118323826	AAPL	-0.85	123.52	121.15

Table 3.1: Sample of historical prices dataset .

3.2 Balance sheet / Income Statement

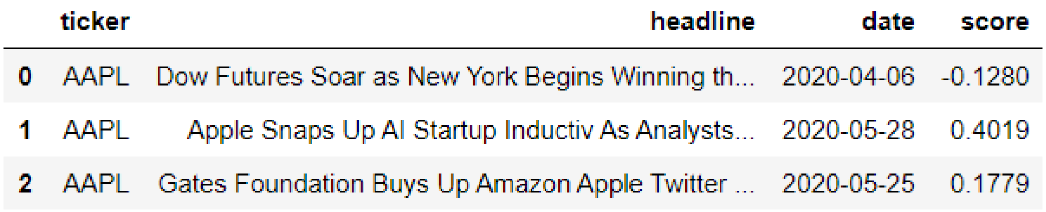
Balance sheet and Income statement dataset describes the assets, loans, revenues, and expenditures of companies from 2016 to 2021. It has 65 columns with 235 rows. Some important variables are total assets, total revenue, total income, long term debt, and capital surplus. Capital surplus is all the money collected from investors that companies haven't used. 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	totalAssets	totalRevenue	netIncome	grossProfit	longTermDebt
0	AAPL	2021	0.0	354054.0	111439.0	28755.0	44328.0	99281.0
1	AAPL	2020	7.5	323888.0	64698.0	12673.0	24689.0	98667.0
2	AAPL	2020	5.0	317344.0	59685.0	11253.0	22680.0	94048.0

Table 3.2 Examples of balance sheet data (values in million dollars)

3.3 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.

3.4 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.

4. Overall Technical Approach

4.1 Data Management and preprocessing

We took two approaches to store our datasets. The first one is AWS Shared Postgres Database. This is because storing the dataset in the cloud is safer as it backups the database. 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.

About preprocessing data, we discarded columns that had all null values and filled in any null values using the median of all the rows. Also, we converted categorical variables into one hot encoded or numerical variable to feed them into the models.

4.2 Exploratory Data Analysis

In order to get a deeper understanding of our data we conducted our data exploratory analysis using visualizations libraries Seaborn and Matplotlib in Python.

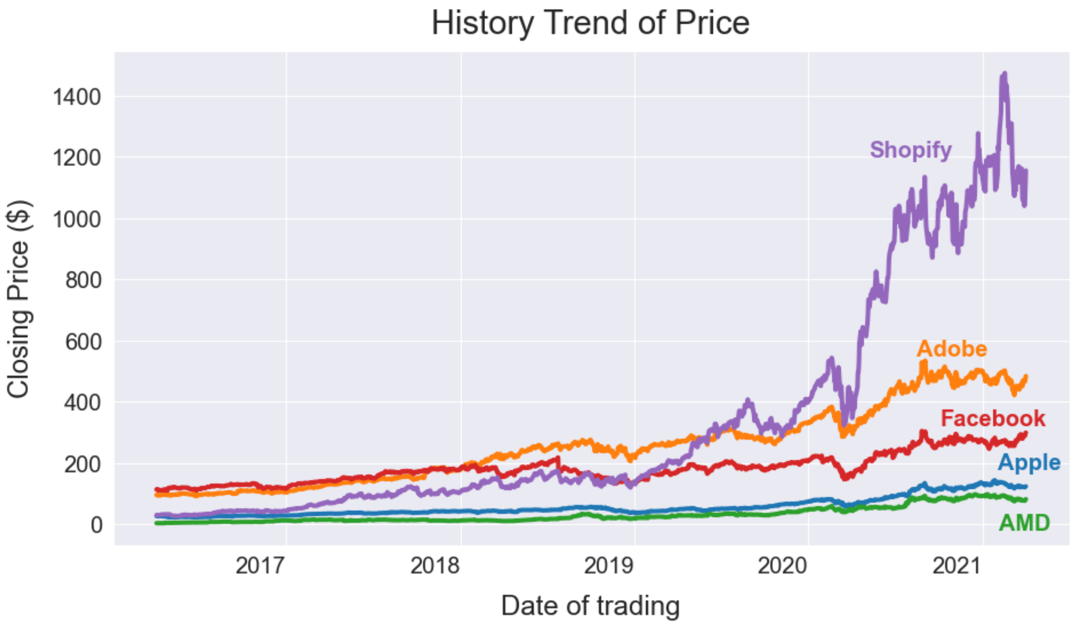


Fig 3.1: The time series within 5 years of 5 stocks: AAPL, SHOP, FB, AMD, ADBE

We first wanted to try and uncover any common trends of the price action between all of our stocks. Fig 3.1 shows the trend of stock prices of 5 stocks from the past 5 years. We found most of the 10 stocks we analyzed present a major spike in price during the past year. There is a light decrease at the beginning of 2020, which could be caused by initial shock of the pandemic. But after that, we observe a strong increase in price, especially Shopify. It can be explained by the demand of technology services while people are forced to work at home and shop online.

We then wanted to see if there were any common trends between a stocks sentiment score and its average price change, along with the market's sentiment. To do this we plotted the monthly aggregated stock and market sentient scores along with the mean monthly price change of all ten companies together which can be seen in Fig 3.4.

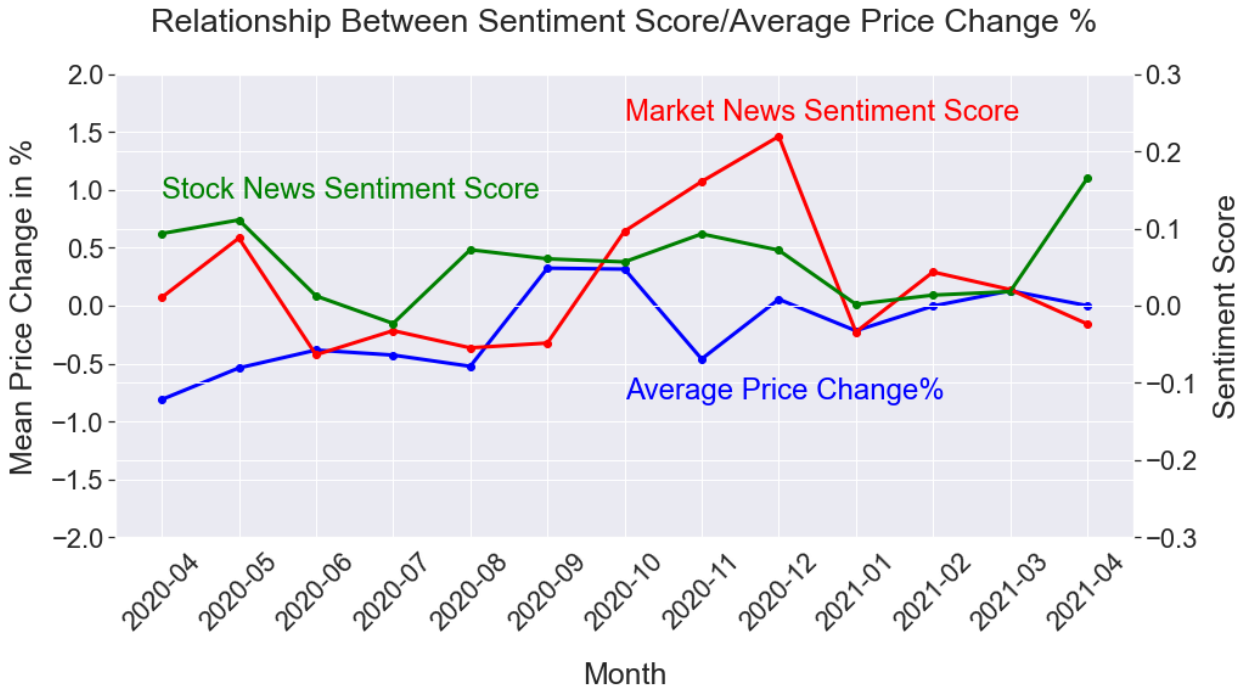


Fig 3.4. Time series data of change in market price and sentiment score

From the plot some key insights we gathered was how the change in price was 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.

Lastly, we tried to find any correlation between some of our joined attributes in the full dataset. Here is the correlation map highlighting a few attributes from the income statements and balance sheet data.

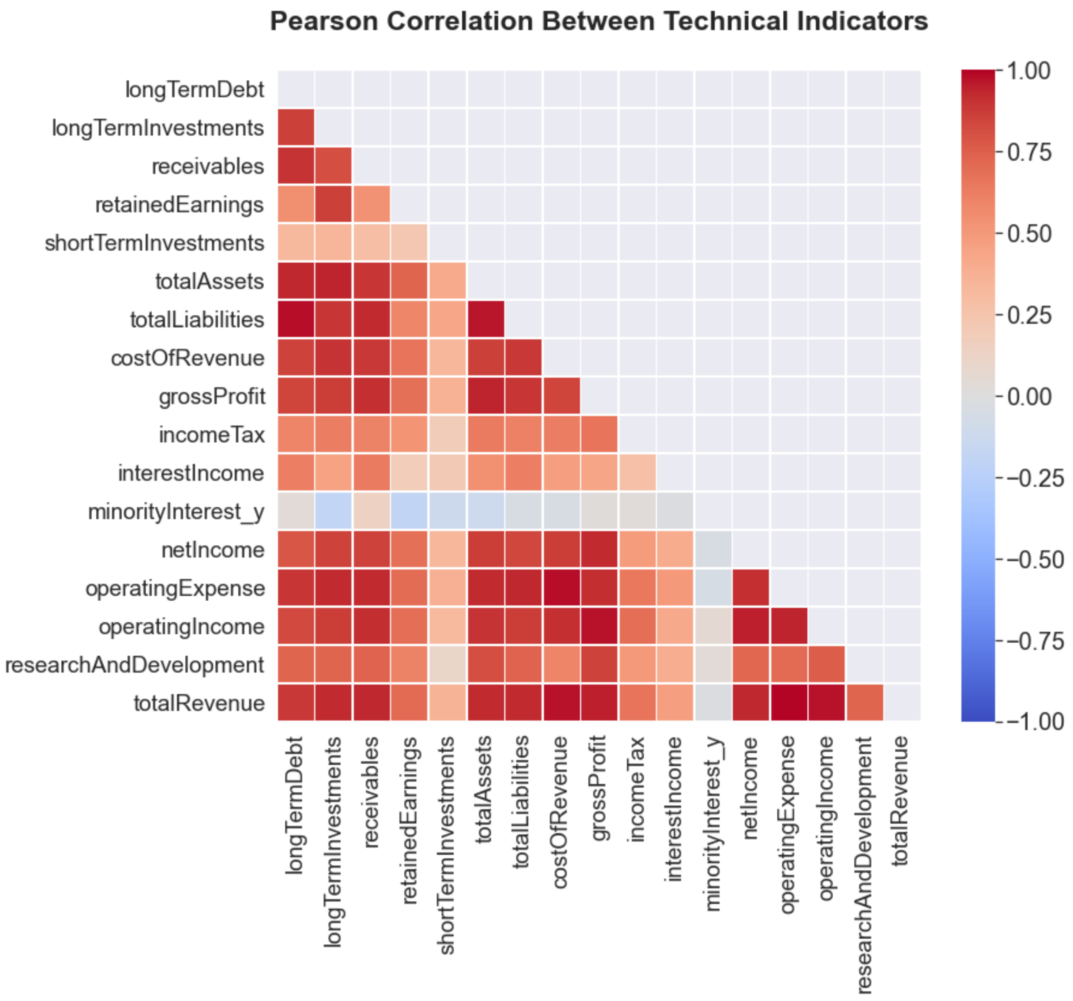
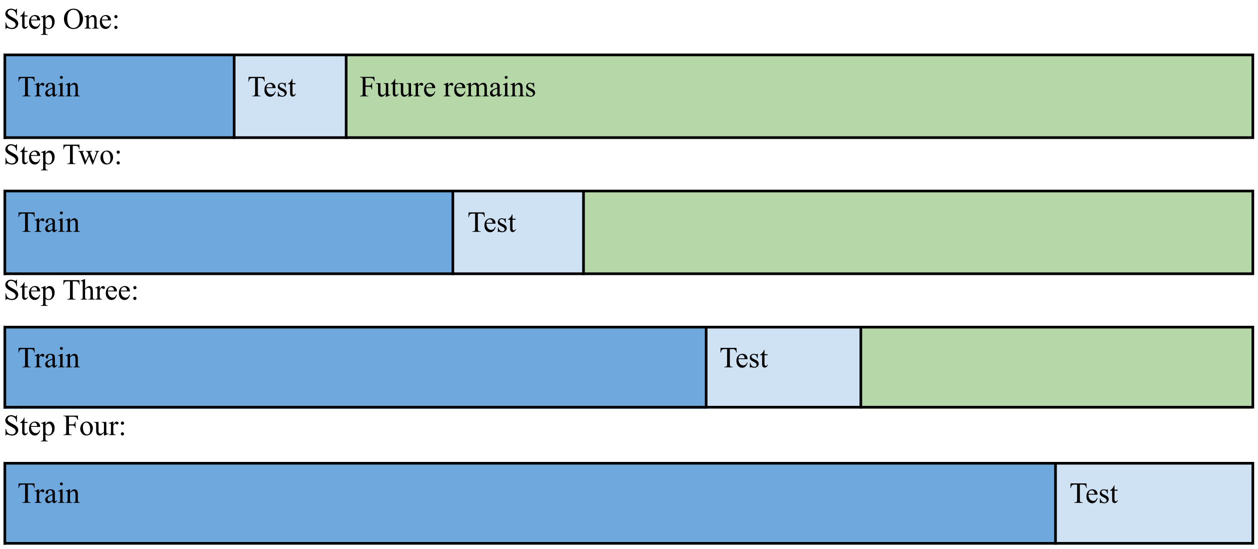


Fig 3.5. Correlation map of income statement and balance sheet. (CapitalSurplus was removed because it only included NaN values)

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 income.

4.3 Data Modeling

Each model was trained and tested on four different time frames. Each time frame was split 80/20 between the training and test set. This was so we could evaluate the strength of our models overtime rather than in one instance. In order to evaluate the accuracy of our model we decided to see how far away our predictions were from the actual open price of the stock in increments of 1, 2, and 3 percent and MAE. We use mean absolute error (MAE) as the evaluation metric because it is more robust to outliers. This should be the good reason for stock price predictions because it often has outliers generated by unexpected events such as covid 19. (In reality, we will add new data into our model each time we obtain it.)



Step One:

Step Two:

Step Three:

Step Four:

Figure 4.1

Model 1: Gradient Boosted Decision Tree (GBDT)

LightGBM is GBDT which adds different decision trees to improve errors from the gradient of loss function. We use this model because it can express more sophisticated and generalized predictions. One assumption to use this model is that features should be either one hot encoded or numerically encoded. About feature engineering, we created more than 180 features which include volatility and trend features.

Model 2: Long Short-Term Memory (LSTM)

LSTM is a type of recurrent neural network (RNN) which uses memory cells in order to retain information overtime. Each cell tries to "remember" all past knowledge that the network has seen so far and "forgets" information it deems as irrelevant. This process is done by introducing different activation function layers called "gates" which each gate has a different purpose. The algorithm has been found in research to be one of the most efficient algorithms for predicting stock prices since it is able to predict the price of a stock using what it knows about the past trading days. It is also great at self optimizing feature importance and uncovering patterns in the fluctuation in price. We used the past five days of data in order to predict the next day and the past twelve days of data to predict the next five. The data needed to be normalized to a scale of 0-1 to ensure that attributes with large domains did not slow down the learning and convergence of our network.

5. Software

The self-written software developed for this project is as follows, listed in order of use:

Name	**Description**
Collect_data folder	Folder which includes all files to collect data.
visualizations.ipynb	Script to conduct exploratory data analysis.
merge_data.ipynb	This file combines all datasets into one merged dataset.
tree_models.ipynb	File contains code for Gradient Boosted Decision Tree models.
LSTM_Models.ipynb	File contains code for LSTM models

Third-party softwares or libraries we utilized are as follows:

Library/Software	**Description**
ta	Technical Analysis library used to create technical features such as volatility, trend, and change in values.
IEX API	API used to extract price, balance sheet, and income statement data.
Mediastack API	API used to extract related stock and market news articles
PostgreSQL Database	Used to house all of our data
scikit-learn	Python library used to build our LSTM models

6. Experiments and Evaluation

Model 1: Gradient Boosted Decision Tree (GBDT)

Here is the result of feature importance for lightGBM which is the GBDT model.

*Mean target values = Setting all predictions to be the mean of the target.

Stock (Ticker)	1% Away	2% Away	3% Away
All tickers (in average) 1-day	43.39%	59.04%	66.81%
All tickers (in average) 5-day	12.62%	24.05%	34.01%

1-day	Predictions	Mean values
MAE	20.6 ($)	94.33
MSE	7983.283	23271.649

5-day	Predictions	Mean values
MAE	24.85 ($)	94.5
MSE	32228.09	23530.63

Table 6.2 Result of LightGBM

$MAE = (\Sigma_{i=0}^{n} |y_i - pred_y|)/n$ $MSE = (\Sigma_{i=0}^{n} |y_i - pred_y|)^2/n$

To see if we have satisfactory results, we compared our MAE with one when we set all predictions to be the mean of the target (In the future, the criteria for the good model will be the amount of money we could actually earn). (Response to "Not clear what would be a satisfactory performance for your task") From the table, we can see that our model is predicting the target well because the error for model predictions are much lower than those of "mean of the target"; however, MSE is not so good. This may be because the models are able to predict general change in price well; however, it may not predict outliers caused by some unexpected events such as COVID-19.

Compared to LSTM models, we have higher MAE and MSE. This may be because while the LSTM model is created for each company, LightGBM is created for all companies. It is true that the tree model can still split the result by the feature (ticker=company); however, it might have suddenly affected the result regardless. Since we decided to make a target to be the actual price, but not change in price to compare with LSTM models, this lost the stability of the target because each company has a different price range. If we make the target to be "change in price", all the companies should have almost the same price range which should increase model stability. (Replying to "Model comparison")

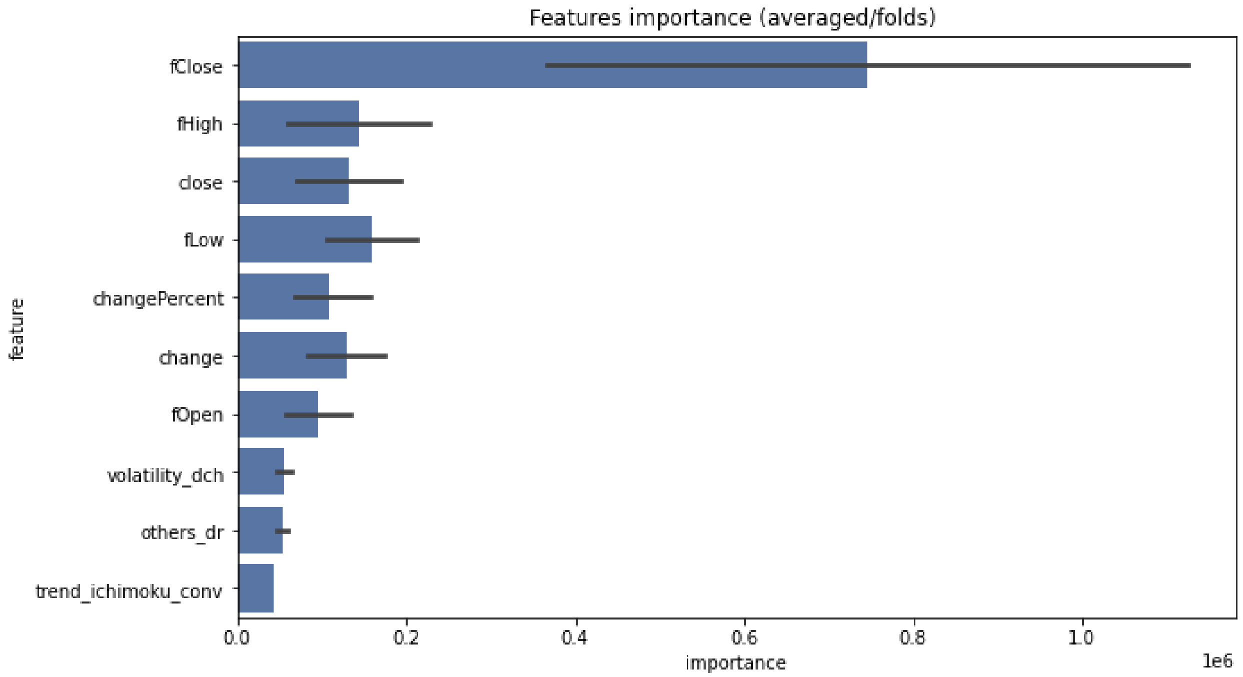


Figure 6.1 Feature importance of LightGBM

We noticed that important features were closing price, highest price, and lowest price. This may be obvious because we are just predicting the one-day ahead closing price. If we know the closing price, we pretty much know the closing price of tomorrow.

Model 2: Long Short-Term Memory (LSTM)

Through experimentation, we found that training and validating only using the last few years of data for each stock produced the best results in terms of price prediction. We found this could be due to the fact that the trend in price action for the past few years was largely altered due to economic changes from COVID-19.

The LSTM model aims to predict one day ahead using two days of past data. The model was fitted using only technical indicator data and sentiment data since our financial/income statement data proved to be insignificant when predicting the daily price change of a stock. All features were fed into the model since LSTM takes care of dictating feature importance.

Because we are using LSTM to predict the daily price change of stock we chose to determine the accuracy of our model based on how far away the prediction price was by one, two, and three percent. This is since a stable technology stock will only move one to five percent on average daily. Additionally, other means of measuring the accuracy of our predictions were found to be unuseful.

Stock (Ticker)	1% Away	2% Away	3% Away	MAE
All tickers (in average) One Day	25.60%	46.94%	60.72%	16.84 ($)
All tickers (in average) Five Days	20.06%	39.05%	52.75%	20.99 ($)

1-day	Predictions
MAE	5.579 ($)
MSE	321.91

5-day	Predictions
MAE	5.516 ($)
MSE	205.52

Table 6.2 Result of LSTM

Compared with LightGBM, we can see that the LSTM model is more stable regardless of targets (either 1-day or 5-days ahead). We also notice that MSE is much lower because LSTM has a low error around COVID-19. These were possibly because LSTM required data normalization and LSTM models were made per company. (Replying to “Model comparison”)

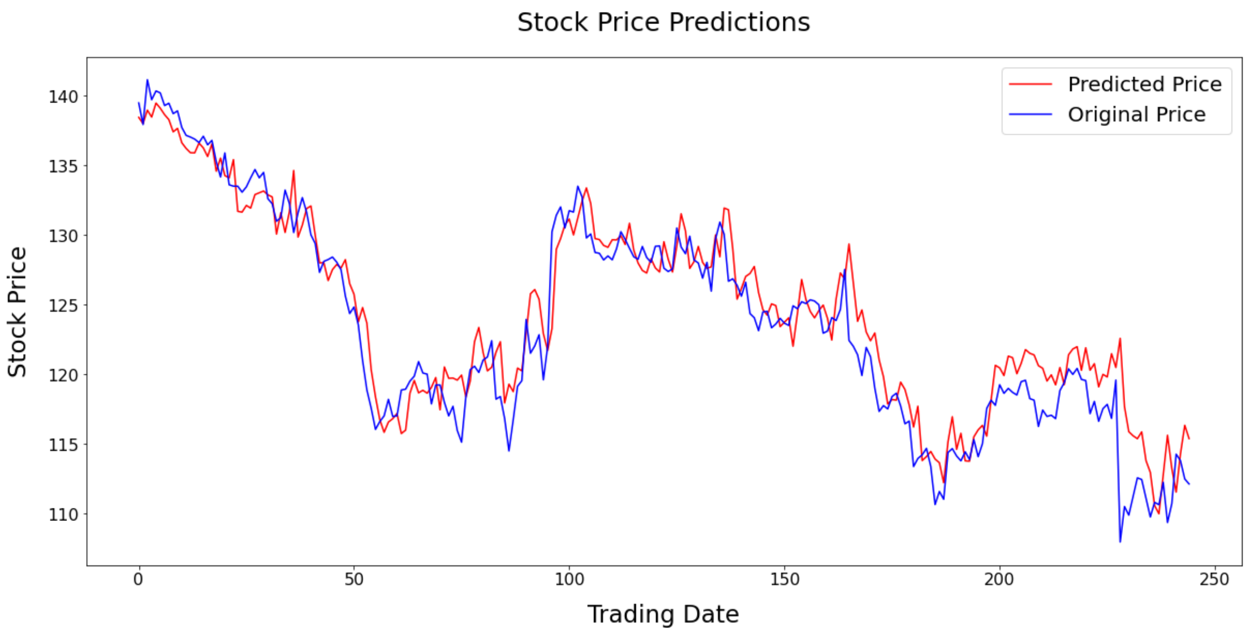


Figure 6.2 LSTM prediction line

We can see that the predicted price pretty much follows the original price although it doesn't follow at the start of COVID-19. This implies that the model is generalized, but can't capture the sudden spikes caused by some events such as COVID-19.

7. Notebook Description

We included the modelling part in our notebook. This is because our objective was to make actual models to predict stock price and analyze results based on models output.

8. Members Participation

Each team member contributed to the project by the kind of dataset. You worked on the balance and income sheet. Hien worked on price and Sean worked on news dataset.

Data collection	
News data	Sean
Price data	Hien
Income and Balance sheet	Yu

Data Management	
AWS (Data base)	Yu
Data Integration	Yu
Data Cleaning	Everybody 33.33%

Modelling	
LSTM	Sean
ARIMA	Hien
LightGBM	Yu

Data Visualization	
News data	Sean
Price data	Hien
Balance sheet/ Income	Yu

Feature Engineering	
Technical indicators	Sean, Hien: 40% Yu: 20%
Balance sheet/ Income	Yu

Integrating files	
Visualization	Yu
Models	Yu

9. Discussion and Conclusion

- What we learned about our models including strength and limitations.

We noticed that LSTM and GBDT models could capture the significance of our features and learn decently well over a relatively short period of time. One of the limitations we found with both models was that they could generalize well but couldn't catch drastic changes in predictions.

We also noticed that tree models make it able to interpret more about features from the plot such as feature importance. About the ARIMA model, we noticed that prediction with univariate variables is not useful, but multivariate models may be worth trying in the future.

- Difficult and surprising things we found in our project.

The first process of collecting our data was a bit difficult due to the limitations of our API's and the overlap of some of the news content generated by the Datanews API. Additionally, early into the project, it became apparent that the same LSTM model did not perform the same among all of our stocks so we had to implement a way in which we could train a specific model for each individual stock.

- Other lessons we learned unexpectedly

Comparing model results was harder than we expected. This was because there were limitations for each model. For example, the target for the time series model was suited to the price but not change in price and they were only suited to use for one time series meaning that they could only use one company stock data.

- Ideas and directions in the future

One idea is to research a method to calculate the sentiment of our stock and market news more accurately based on the article rather than just the headline. This could be beneficial to capture significant events that may cause a dramatic change in price. Another idea is to get all companies' data in the stock market and select only top companies to invest in. This will increase stability and reduce risk to invest money. (Response to the comment "Can you relate your results to trading strategies? ") In addition, in the future, we need to change our project objective to whether we can actually earn money to see if we have satisfactory results. (Response to "Not clear what would be a satisfactory performance for your task")

7. APPENDIX

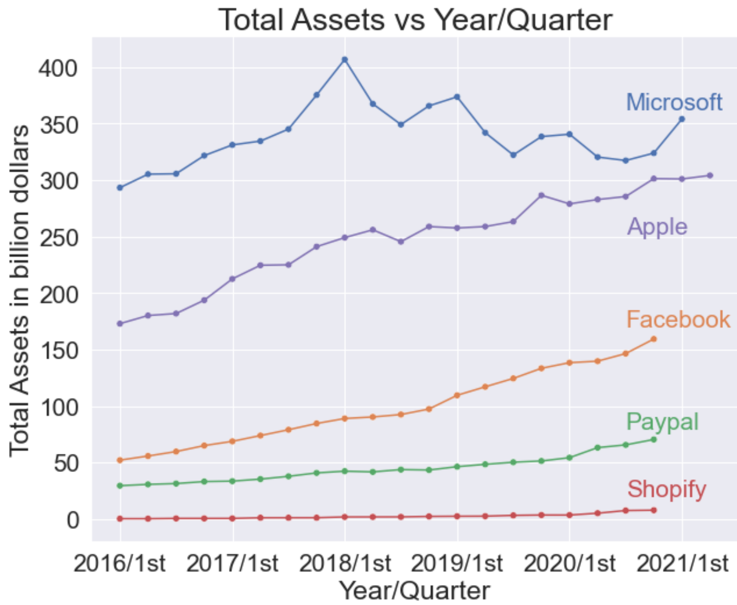


Fig 3.2. The time series data of total assets in five different technology companies

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 the stock price dataset (fig 3.1), but its assets haven't increased yet.