Predicting whether an Apple stock will increase or not using news articles as features pre and post COVID period

Author: Young Cho

Problem Statement:

Retail Investors (aka personal/individual investors) represented just 10% of the US Stock market in 2019. However, the proportion rapidly climbed to as much as 25%[1] of the stock market's activity in 2020, a year that can be characterized by COVID-19 pandemic. One of the major contributing factors to the enormous growth to the trend can be explained by the rise of commission-free brokers starting from Robinhood, which has forced even the traditional brokers to drastically reduce commission fees or go completely commission-free to compete winning over a growing representation of retail investors.

Retail investors are primarily characterized by "responding quickly to overnight returns and short-term news" and "pursing both momentum and contrarian strategies"[2]. This naturally leads to the primary hypothesis of this analysis. Since retail investor group has grown so much bigger from pre COVID-19 period to post COVID-19 period, and this group is known to react much more sensitively to news media and having commission-free brokerage apps like Robinhood at their fingertips, the influence/predictive strength of news media over explaining short-term stock movement must have grown as well. It should be clear that the purpose is not to build a high performing stock prediction model leveraging multiple categories of features.

In summary, this analysis will be in 3 part.

1. Build multiple stock prediction models with varying ML algorithms that only uses news media data as dependent variables and identify the best one.
2. Using the best model, slice time in chunks to train model in each time partition to evaluate consistency in performance.
3. Run step 1 & 2 for pre COVID-19 period and post COVID-19 period and compare them.

Methodology:

Pre COVID-19 period in the analysis indicates a period between 2019-04-08 and 2019-12-09. Post COVID-19 period is a period between 2020-04-06 and 2020-12-04. Please note that these periods were selected with the following intentions:

1. Removing seasonality effect
2. To avoid the COVID-19 stock collapse (S&P losing 33% from Feb. 2020 to March 2020) period where there was extreme amount of volatility in the US stock market.

[1] https://markets.businessinsider.com/news/stocks/retail-investors-quarter-of-stock-market-coronavirus-volatility-trading-citadel-2020-7-1029382035

[2] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3703815

Apple stock was picked for the main entity of the analysis because Apple is the 2nd largest stock[3] in S&P500 index (market cap weighted) and has a strong branding on consumer electronics with a disproportionate distribution of profits from a single product (iPhone) that is upgraded every year, making its short-term stock price movements to sensitive to what was said on the news.

Stock prediction model can be built in both regression and classification forms. In this analysis, we will build binary classification model that predicts whether the price of Apple stock will "increase (1)" the next day or "not increase (0)", using news data from 1 day before. In the attempt to make the model practical and more robust, observations with increase less than 0.5% from the previous day's price have been classified as not increase (0)".

Next, create categorical features with the results obtained from running a text vectorizer on the body of text with varying the following parameters:

- Ngram range: The min and max length of the sets of words to be tokenized
- Minimum Count Limit: The min of repetitions to qualify for tokenization (as one-time occurrence of a specific word is not useful)

In total, 4898 binary categorical features (tokenized word or phrase less than 2 words) were used in training classification models (below) for each of ~240 observation (8 months).

- Vanilla linear model (OLS)
- Linear model with L1 penalty (Lasso)
- Linear model with L2 penalty (Ridge)
- Support Vector Regressor (SV Regressor)
- AdaBoost with Decision Tree (ADABOOST)
- Random Forest (RF)

Next step is to pick the best model and evaluate its consistency of performance using overall accuracy by slicing time in 12 chunks (~3 weeks of trading data for each time partition) to train model in each time partition. And finally compare the performance of the best model between pre COVID-19 period and post COVID-19 period.

[3] https://www.investopedia.com/articles/investing/053116/10-largest-holdings-sp-500-aaplamznfb.asp

Results:

Pre COVID-19 Period Model Performance

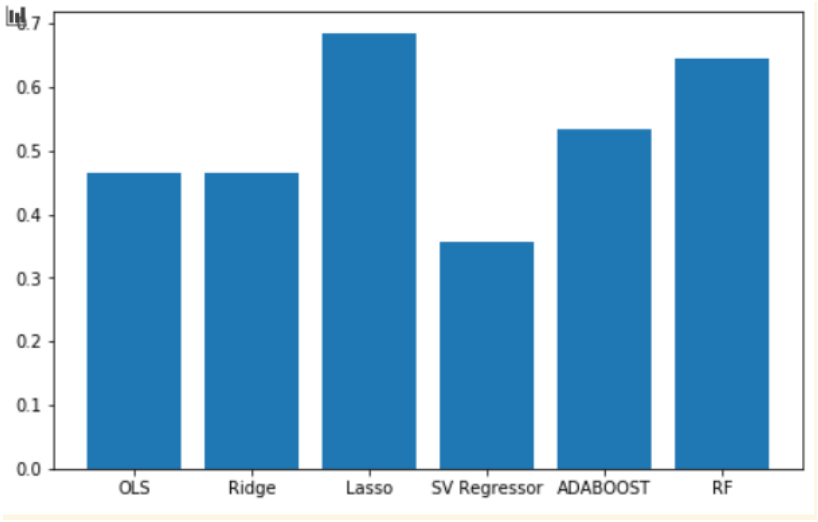

Chart1: Pre COVID-19 Period Classifier Performance Comparison

	accuracy	conf_mat	roc	time_chunk
0	0.571429	[[8, 7], [5, 8]]	0.553846	0.00
1	0.687500	[[13, 6], [4, 9]]	0.589069	0.05
2	0.542857	[[9, 11], [5, 10]]	0.591667	0.10
3	0.641026	[[16, 4], [10, 9]]	0.713158	0.15
4	0.547619	[[10, 12], [7, 13]]	0.579545	0.20
5	0.608696	[[18, 5], [13, 10]]	0.612476	0.25
6	0.551020	[[15, 10], [12, 12]]	0.550000	0.30
7	0.622642	[[19, 7], [13, 14]]	0.644587	0.35
8	0.589286	[[20, 8], [15, 13]]	0.620536	0.40
9	0.550000	[[19, 11], [16, 14]]	0.647778	0.45
10	0.531250	[[24, 8], [22, 10]]	0.527344	0.50
11	0.507463	[[30, 3], [30, 4]]	0.549465	0.55

Chart2: Pre COVID-19 Period Classifier Performance Consistency over each time chunk

```
print(results_df.accuracy.mean())
```

```
0.5792321977962018
```

Chart3: Pre COVID-19 Period Mean of Best Classifier Performance across 12 time chunks

Post COVID-19 Period Model Performance

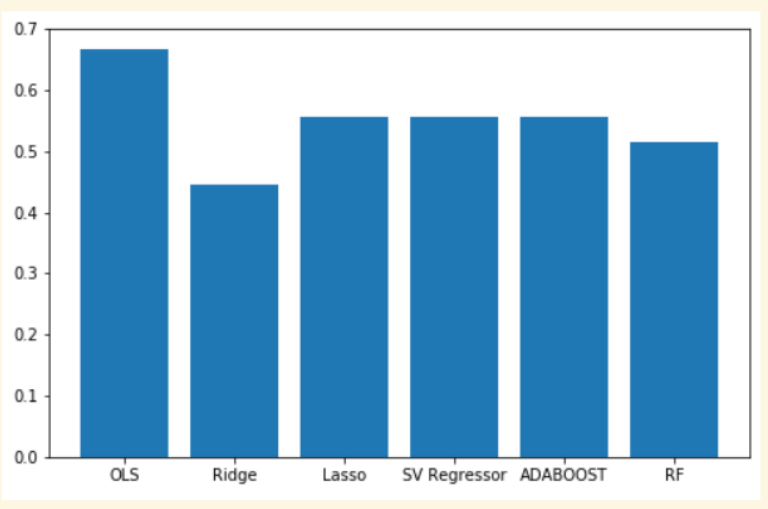

Chart4: Post COVID-19 Period Classifier Performance Comparison

	accuracy	conf_mat	roc	time_chunk
0	0.464286	[[7, 1], [14, 6]]	0.606250	0.00
1	0.516129	[[3, 6], [9, 13]]	0.540404	0.05
2	0.514286	[[6, 6], [11, 12]]	0.586957	0.10
3	0.631579	[[8, 6], [8, 16]]	0.629464	0.15
4	0.714286	[[6, 9], [3, 24]]	0.701235	0.20
5	0.555556	[[5, 13], [7, 20]]	0.521605	0.25
6	0.653061	[[4, 16], [1, 28]]	0.531897	0.30
7	0.653846	[[4, 16], [2, 30]]	0.588281	0.35
8	0.571429	[[4, 18], [6, 28]]	0.554813	0.40
9	0.644068	[[6, 16], [5, 32]]	0.622850	0.45
10	0.571429	[[6, 18], [9, 30]]	0.564637	0.50
11	0.621212	[[8, 18], [7, 33]]	0.564904	0.55

Chart5: Post COVID-19 Period Classifier Performance Consistency over each time chunk

```
print(results_df.accuracy.mean())
```

```
0.5925970930878806
```

Chart6: Post COVID-19 Period Mean of Best Classifier Performance across 12 time chunks

Discussion and Conclusion:

The final overall performance of the best classifier for pre COVID-19 period is 57.9%, while for post COVID-19 period it is 59.3% (difference of 1.34%). It appears that our original hypothesis “increased influence of news media over certain stock’s short-term returns grew from pre COVID-19 period to post COVID-19 period, with rising retail investors representation with commission-free broker apps” is true.

Whether or not the above conclusion can also be made to other stocks or other types of assets should be examined further, but the hypothesis will be that “if there was a growth of retail/individual investors and a rise of commission free brokerage, then the influence of news media over the asset’s short-term return would’ve also grown.