# Stock Prediction Using COVID News

—

Dana Chernysheva * Ying Guan † Ziyue Su ‡

December 13, 2020

## Abstract

The COVID-19 pandemic has had a significant influence on the economy and the stock market, from imposing lock downs to restricting travel. We were curious about exploring the relationship between COVID news/factors and stock prices. In this project, we focused on four industries: healthcare, technology, airlines, and retail. For each industry, we chose 10 stocks to analyze (a detailed lists of these stocks can be found under the same directory). Choosing 10 stocks per industry enabled us to assess how COVID affects sectors differently, since some industries have been affected more than others by COVID. For example, before beginning our analysis, we believed that the airline industry has been negatively impacted by COVID since people are less likely to go on vacation/travel; this suggests a negative exposure to COVID news/factors.

In this project, we aim to use the intensity scores of news headlines (how positive or negative the headline is), in addition to other COVID-related factors, like keywords in headlines, positive cases, and deaths, to identify any relationships with stock prices. Ultimately, our goal is to predict stock price moves based on the COVID-related factors. We understand that correlation does not imply causation and that COVID is not the only factor affecting stock prices, but with the pandemic affecting our lives, we thought it would be interesting to analyze its impacts on the stock market.

*Cornell University, M.Eng. in Financial Engineering, *djc*464@*cornell.edu*
†Cornell University, M.Eng. in Financial Engineering, *yg*532@*cornell.edu*
‡Cornell University, M.Eng. in Financial Engineering, *zs*366@*cornell.edu*

# 1 Introduction

## 1.1 Data Description

### 1.1.1 Data Wrangling

Our data set consists of the adjusted daily closing prices of 40 stocks (10 stocks for each industry) from 02/11/2020 to 10/30/2020, obtained via Yahoo Finance. The COVID news headlines were scraped from the World News subreddit with the flair "COVID-19." This was done by utilizing the PushShift API. These headlines began accumulating on 02/11/2020. Thus, we chose this date as our starting point. Extracting all of the relevant headlines required running our program for about eight hours. Additionally, we are using cumulative deaths and positive COVID cases (obtained from the COVID Tracking Project) as potential features.

We cleaned the headlines data by removing all N/A values and duplicates. Frequently, the PushShift API could not handle the continuous number of calls we were making to collect our data, so we employed exception handling rules to ensure that if the API did not handle the call properly, it would restart at the previous headline it accessed. Moreover, we also deleted any headlines consisting of non-English words.

### 1.1.2 Missing Values Imputation

Since there are daily values for the COVID-related features, we decided to impute the missing stock data (weekends/holidays). Based on a Stanford research paper (see references), we decided to impute the missing stock prices using the following method:

$$\text{price} = \frac{\text{next available price} + \text{previous available price}}{2}.$$

The paper justifies this imputation because stock data tends to follow a concave function, except at anomaly points of sharp rises/falls.

### 1.1.3 Feature Selection and Engineering

The COVID-related features that we considered and found most useful include:

- *Sentiment of headline.* We used the VADER sentiment package to give each COVID-related headline a sentiment score. We chose VADER because it is considered to be a powerful tool to analyze sentiments expressed in news. It gives the attitude/intensity for how positive or negative each headline is, and does not require training data. The sentiment score of each headline is represented by its compound score, which is weighted by its positive score, neutral score, and negative score. While COVID-related news tends to be negative, "positive" news headlines, such as "New Zealand COVID Cases Hit Zero" exist. From the headlines on each specific day, we then computed the total daily sentiment scores.

- *Representative keywords.* We extracted some representative keywords from the headlines as features, which include the number of times that words related to death (i.e. death, die, fatal, etc.), vaccine, spread (i.e. spread, increase, transmission, expose, etc.), and shut (i.e. shut, close, lock) occurred per day. We felt that such COVID-related keywords largely determine the overall sentiment of each headline and would be critical features to keep track of.

- *Numbers of U.S. cases and deaths.* We believe that it was essential to include the number of positive cases and deaths involving COVID in the U.S. as two useful features in addition to our news-related features since they are important trackers of COVID.

- *Moving average of stock prices.* We found that it would be effective to use the moving average for the past $n$-days as a feature; we will discuss

our reasons for this further in the Model Selection section.

- *Stock splits.* For tech stocks, we included a feature regarding stock splits (no other industries had any in this time frame). We included this because stock splits cause the price per share to decrease dramatically. For example, Tesla split its stock on a 5-for-1 basis on 08/31/2020; this sudden price drop was not influenced by factors related to COVID, so we wanted to account for that.

- *Daily number of U.S. passengers.* Airline stock prices tend to display seasonality effects, with travel increasing during summer months and holidays. Therefore, to consider the natural fluctuations within this industry, we included a feature for the average daily number of U.S. passengers travelling domestically and internationally from the past three years (data obtained from the Bureau of Transportation).

## 1.2 Data Visualization

### 1.2.1 COVID Headlines

Figure 1 shows the distribution of sentiment scores for each day's headlines. Based on the chart, most scores are negative (207 days out of 263), as expected. Note that the distribution is right skewed due to a very large, positive outlier.

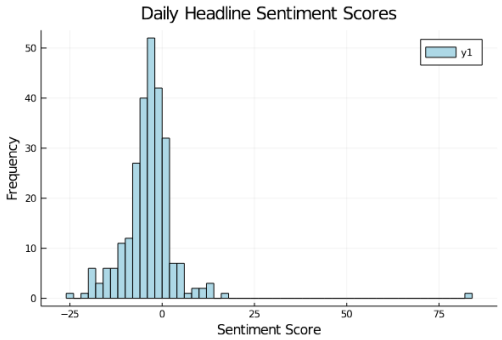


**Figure 1:** Daily Headline Sentiment Scores

Figure 2 illustrates the total frequency of the keywords in the headlines; words related to death occur most often. However, all keywords show up over 500 times, so we believe there are a sufficient number of occurrences of each of the keywords to use them as features.

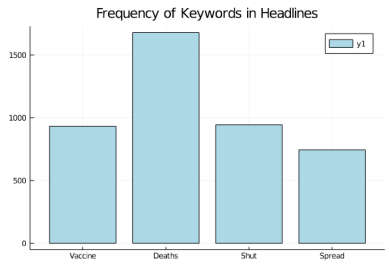


**Figure 2:** Frequency of Keywords in Headlines

In Figure 3, we plotted the keyword frequencies on 10 random dates. The figure shows that almost every day, each of the keywords arise in the headlines; as we saw in Figure 2, Figure 3 also shows that on these 10 dates, the keywords related to "death" tend to occur at higher frequencies.

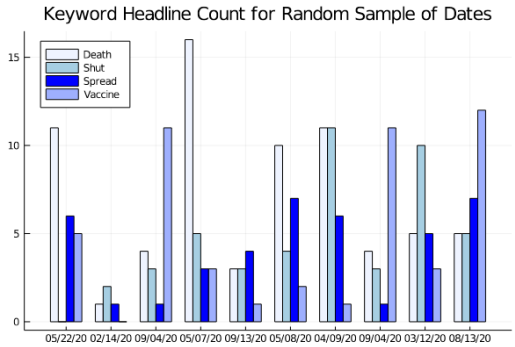


**Figure 3:** Keyword Headline Count for Random Sample of Dates

Figure 4 is a correlation heat map of our potential features, using Delta stock as an example of the stock price feature. Since this is an airline stock, we also have the feature relating to average daily travel. Here, we see the total daily sentiment has a decent correlation with stock price (0.21). Because of this, we felt that using the total daily sentiment in our model could

potentially be useful in predicting stock prices. Additionally, the figure illustrates that our keyword features (vaccine, shut, spread, and death) and COVID positive cases/deaths are correlated to stock price. For example, the correlation between death and stock price is -0.49.

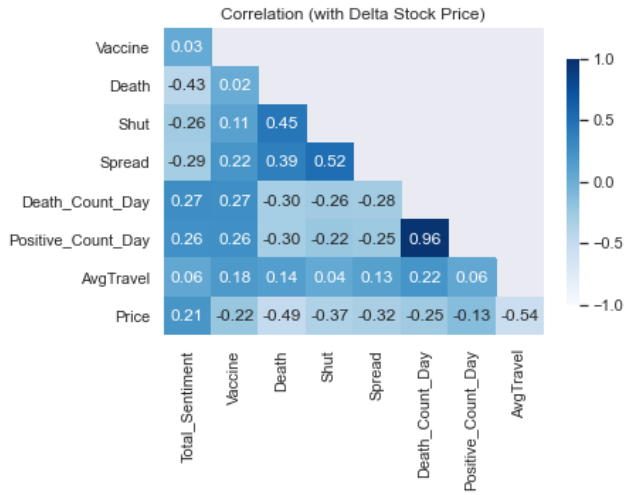


**Figure 4:** Correlation Heatmap for Delta

# 2 Initial Modeling

## 2.1 Rolling K-Fold Cross Validation

To avoid overfitting and underfitting, we fit our data using different models, choosing the best model using cross validation. However, k-fold cross validation cannot be applied directly since we are exploring time series data; it would be meaningless to fit the model using future data and predict on past dates. Instead, we utilized cross validation on a rolling basis. Specifically, we divided our overall time interval into six sub-intervals, each with 43-44 consecutive days in it. To begin, we used the first subset as our training set and the second subset as our test set. Then, we used the first and second subsets as our training set and the third subset as our test set, and so on. For each training set, we used an $80-20$ split for training and validation. The general idea of is displayed in Figure 5.

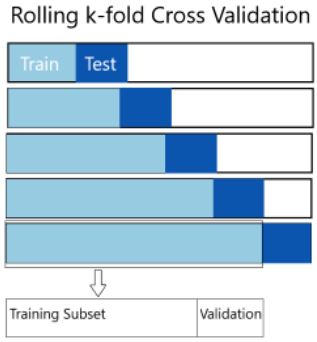


**Figure 5:** Rolling k-fold Cross Validation

## 2.2 Model Selection

We first investigated some time series models such as Autoregressive models with different lags, where AR(p) is defined as follows:

$$y_i = \phi_0 + \phi_1 y_{i-1} + ... + \phi_p y_{i-p} + \epsilon_i.$$

Here, $\epsilon$ represents white noise since there is randomness in the model. AR models predict future data using past data, going back $p$ days (seen in the $y_i$ values). However, these models, while producing low validation and test errors, created predictions that seemed to shift the predicted stock prices horizontally. This indicated that the models were placing too much emphasis on the previous day's stock price. As an example, Figure 6 shows that for United, the AR(1) model produced the least error while displaying this horizontal shift.

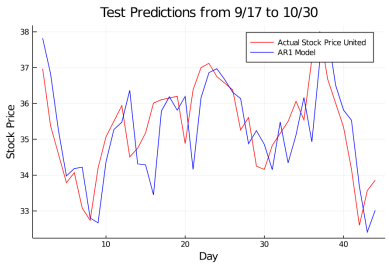


**Figure 6:** United Horizontal Shift

For this reason, we decided to standardize our data and utilize different loss/regularizers. Some of these loss functions include: quadratic loss, L1 loss, Huber

loss, and Quantile loss. In terms of regularizers, we considered One regularization and Quadratic regularization. We also decided to include moving averages as a feature. Moving averages smooth out stock price trends by looking at an n-day average of the prices. Since our time frame is short-term, we decided to use 5-day moving averages. Our goal was to use the moving average so that we did not place excessive weight on the previous day's stock price (to remove the horizontal shift we were seeing before). Figure 7 highlights how the moving average removes some of the noise from the stock movements, creating a smoother curve.

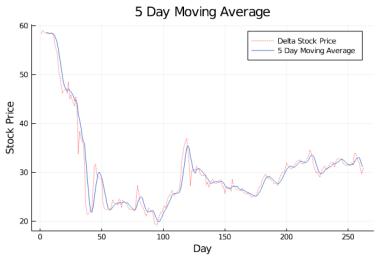


**Figure 7:** Moving Average Example

### 2.3 Error in the Model Fit

By utilizing rolling k-fold cross validation, we fit different models and selected the best model for each stock as the model that had the smallest mean absolute error, defined as

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}.$$

We experimented with several types of errors, including MSE and MAE. We chose MAE because it makes more sense in the context of looking at how different our stock price predictions were compared to the actual stock price, as opposed to mean squared error. After finding the model that had the lowest mean absolute error for the training data, we used this model on our test set.

## 3 Analysis

To choose the optimal $\lambda$, we first fit the models using $\lambda$ ranging from (0,1) with step-size 0.1 and computed the mean absolute errors (MAE) of our predictions using the fitted models; the results are displayed in the figures below. For conciseness, we only included the graphs for two industries: healthcare and tech. As we can see from the healthcare graphs, the MAE of predictions does not change much as $\lambda$ changes, and for the majority of stocks that we are analyzing, the MAE increases as $\lambda$ increases. Therefore, in order to keep the prediction as precise as possible, yet still being able to set some weight constraints, we decided to use $\lambda = 0.1$.

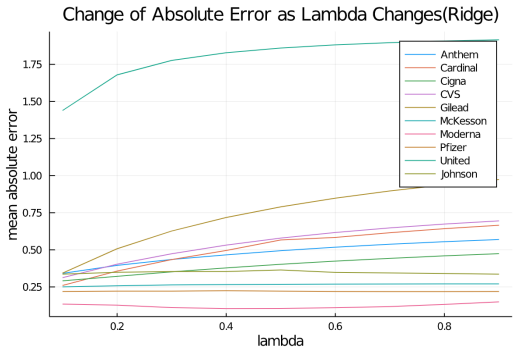


**Figure 8:** Mean Absolute Error of Prediction Using Ridge Regression as $\lambda$ Changes (Healthcare)

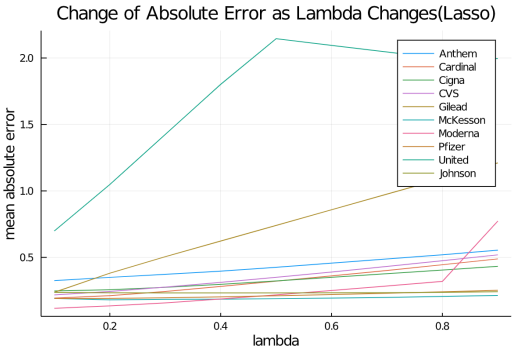


**Figure 9:** Mean Absolute Error of Prediction Using Lasso Regression as $\lambda$ Changes (Healthcare)

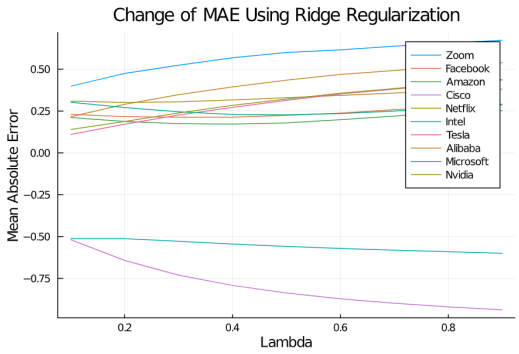


**Figure 10:** Mean Absolute Error of Prediction Using Ridge Regression as $\lambda$ Changes (Tech)

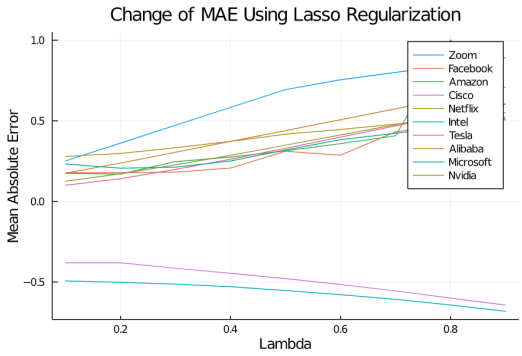


**Figure 11:** Mean Absolute Error of Prediction Using Lasso Regression as $\lambda$ Changes (Tech)

We applied similar ideas to the rest of the industries. For example, for airlines, using a similar approach, we used $\lambda = 0.1$ and a quantile of 0.6.

Below is a table summarizing the models chosen for each stock in the healthcare industry using rolling k-fold cross validation, as well as the respective average validation error and test set error. The stock models in the healthcare industry do not seem to overfit or underfit, as the validation and test errors are relatively low and are close to each other in magnitude.

| Healthcare Industry | | | |
|---|---|---|---|
| Company Name | Loss/Reg | Val Error | Test Error |
| Anthem | Quad/Quad | 0.5303 | 0.3426 |
| Cigna | Quad/One | 0.3895 | 0.3147 |
| Cardinal Health | Quad/One | 0.3511 | 0.3215 |
| CVS Health | Quad/One | 0.3232 | 0.2986 |
| Gilead Sciences | Quad/One | 0.3359 | 0.2958 |
| McKesson | Quad/One | 0.3545 | 0.2859 |
| Moderna | Quad/One | 0.1449 | 0.1565 |
| Pfizer | Quad/One | 0.2620 | 0.3524 |
| UnitedHealth | Quad/Quad | 0.2473 | 0.4947 |
| Johnson&Johnson | Quantile/Quad | 0.3994 | 0.3617 |

Below is a similar table for the tech industry. For the tech industry, the validation and test errors are not hugely different, and are relatively small, so it is reasonable to believe that there is not much evidence of overfitting or underfitting.

| Tech Industry | | | |
|---|---|---|---|
| Company Name | Loss/Reg | Val Error | Test Error |
| Zoom | Quad/One | 0.1031 | 0.2894 |
| Facebook | Quad/One | 0.2294 | 0.6500 |
| Amazon | Quad/One | 0.0162 | 0.2066 |
| Cisco | Quad/One | 0.5099 | 1.0904 |
| Netflix | Quad/One | 0.0839 | 0.1637 |
| Intel | Quad/One | 0.5711 | 0.5086 |
| Tesla | Quad/One | 0.0309 | 0.4946 |
| Alibaba | Quad/One | 0.0200 | 0.5105 |
| Microsoft | Quad/Quad | 0.1716 | 0.1707 |
| Nvidia | Quad/One | 0.1167 | 0.2625 |

The table below pertains to the airline industry. From this table, we observe that the models do not seem to overfit or underfit. In the airline industry, most of the COVID-related features had a negative impact on the stock price, which makes sense consid-

ering people are traveling less during the pandemic. Overall, for this industry, quadratic loss and quadratic regularization seemed to be the best model for most of the stocks.

| Airline Industry | | | |
|---|---|---|---|
| Company Name | Loss/Reg | Val Error | Test Error |
| Delta | Quad/Quad | 0.3292 | 0.2068 |
| United | Quad/Quad | 0.2767 | 0.2055 |
| Southwest | Quad/Quad | 0.2195 | 0.2525 |
| Jet Blue | Quad/Quad | 0.2700 | 0.2224 |
| American | Quad/Quad | 0.2284 | 0.2983 |
| Hawaiian | Quad/Quad | 0.3032 | 0.2428 |
| Spirit | Huber/Quad | 0.2765 | 0.2224 |
| Alaskan Air | Quad/Quad | 0.2312 | 0.2733 |
| Skywest | Quant/Quad | 0.3499 | 0.2626 |
| Mesa Air | Quant/Quad | 0.3158 | 0.2502 |

Using the COVID-related features and moving averages to predict airline stock prices seemed to capture the general trends of the stocks. For example, in Figure 12, we observe that the model chosen via rolling k-fold cross validation tends to follow the up/down stock price movements of the actual Southwest stock between 09/17 and 10/30. The same trend is seen with United in Figure 13. Since most of our features are COVID-related, and do not take into account other factors that are used in general stock price prediction, we were not surprised that our models differ slightly from the actual price movements. However, we believe that using our model, we were able to capture some of the effects of COVID on the stocks, seen as the stock fluctuations are followed relatively well.

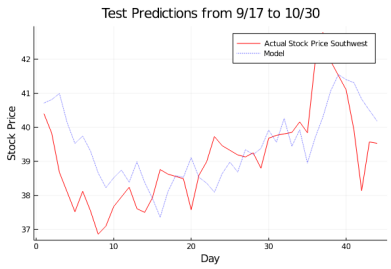


**Figure 12:** Southwest Model vs. Actual

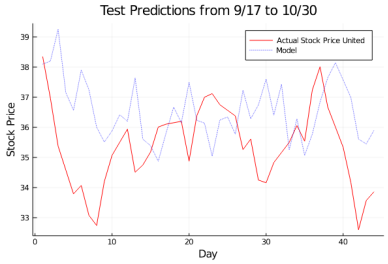


**Figure 13:** United Model vs. Actual

Below is a similar table for the retail industry. The retail industry has more varied validation and test errors; for example, the Costco model seems to have much bigger validation error compared to test error. As we discuss in our section for potential improvements, we would ideally like to implement more complex models to try to lower these errors.

| Retail Industry | | | |
|---|---|---|---|
| Company Name | Loss/Reg | Val Error | Test Error |
| Walmart | Quad/One | 0.0518 | 1.4305 |
| Target | Quad/One | 0.4421 | 0.1227 |
| Macy's | Quad/Quad | 0.2090 | 1.3112 |
| Lululemon | Quad/One | 0.0754 | 1.0006 |
| Costco | Quad/Quad | 4.6038 | 0.3899 |
| Nordstrom | Quad/Quad | 4.6813 | 1.5131 |
| Home Depot | Quad/One | 0.2781 | 0.2533 |
| Kohl's | Quad/Quad | 0.6562 | 1.6593 |
| JCPenney | Quad/Quad | 2.0840 | 0.6088 |
| Ulta | Quad/Quad | 0.0748 | 0.4859 |

# 4 Conclusion

For each industry, we averaged the model coefficients related to COVID features. Below is a chart showing the rankings of the features from highest to lowest importance/weight (1 is highest). For healthcare stocks, the keyword vaccine was the most important feature, followed by positive COVID cases. For tech stocks, positive cases and total sentiment were the most important features when making predictions. For airline stocks, positive cases and the keyword spread were the most important features, and for retail stocks, positive cases and the keyword vaccine were the most important. From this, it seems that positive COVID cases are an important feature across all industries.

| Feature Importance Rankings | | | | |
|---|---|---|---|---|
| Features | Health-care | Tech | Airline | Retail |
| Sentiment | 6 | 2 | 6 | 6 |
| "Death" | 3 | 4 | 5 | 5 |
| "Vaccine" | 1 | 4 | 8 | 2 |
| "Shut" | 7 | 3 | 4 | 4 |
| "Spread" | 5 | 7 | 2 | 3 |
| DeathNum | 3 | 4 | 7 | 5 |
| PositiveCases | 2 | 1 | 1 | 1 |
| StockSplits | N/A | 8 | N/A | N/A |
| TravelNum | N/A | N/A | 3 | N/A |

In addition, for each industry, we averaged all of the model coefficients related to COVID features in order to assess the industry's exposure to these features. Figure 14 demonstrates this. It appears that the airline industry has the most negative exposure to COVID, followed by the healthcare industry. Retail has a slightly negative exposure. As expected, the tech industry has positive exposure, a trend that we have seen in the markets. These results are not surprising given what we have observed in reality; airlines seem to have been hugely negatively affected by COVID and the tech industry has generally remained strong during the pandemic.

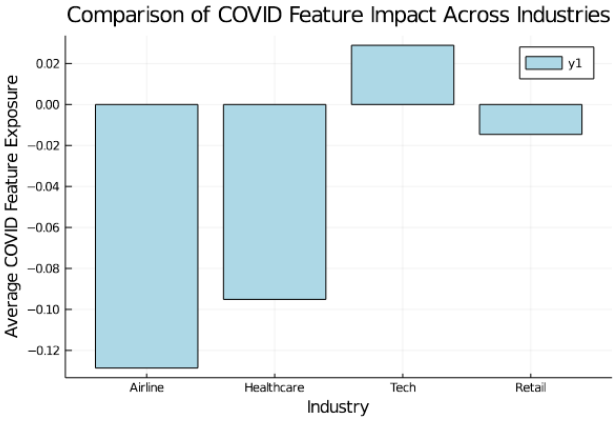


**Figure 14:** COVID Exposure for Each Industry

Furthermore, by analyzing the fitted weights for individual stocks, we note that for the healthcare industry, companies with higher stock prices, namely Anthem, Cigna, and Johnson&Johnson, have corresponding models that tend to be more sensitive to COVID headlines compared to those with lower stock prices (Moderna, Pfizer, Cardinal Health etc.)

As the graphs corresponding to the airline industry indicate (Figure 12 and Figure 13), we have mostly captured the up/down stock price movements, so it is logical to conclude that our COVID features do have a relationship with stock prices. Since most of our features are COVID-related and our models do not take into account other factors that are used in general stock price prediction, it is reasonable that our models differ slightly from the actual stock price movements. However, we believe that our models were able to capture partial effects of COVID on the stocks, especially since they do capture most of the up/down movements.

In general, we are highly confident with our predictions as most of the test errors are below 0.5. As mentioned previously, most of our predictions were able to demonstrate similar price fluctuations to the actual

stock price movements, meaning that these models are explaining the COVID exposure of different industries well. By taking into account that our models are solely revealing the impact of COVID-related news headlines on stock prices, and that in reality, many other factors also influence the movement of stock prices, we can utilize our models as one indicator of our companies' next moves, together with other measurable indicators.

# 5 Potential Improvements

There are several ways to improve our current project. Firstly, we could include headlines from more sources, in order to gain a broader pool of daily headlines. Secondly, we believe it is also worth considering modeling our results as a classification problem (stock price moves up or down). It would be interesting to use classification models such as the Support Vector Machine to compare results to our current models and check if the results align. With more time, we also would have explored dimensionality reduction and using low rank models to see what results could arise. In addition, some of the errors in the retail industry were somewhat high, so another way we could improve our models would be to fit the retail industry using more complex models, while ensuring there is no overfitting or underfitting. Although this is beyond the scope of this course, we could explore the ARCH model, GARCH model, Hidden Markov Model, and Kalman Filter as well.

# 6 Weapons of Math Destruction and Fairness

The idea of fairness as defined in lecture notes does not fully apply in our project because we are not looking at individuals and potential biases that can arise (e.g. gender or race biases). However, biases could have arisen in the news headlines collected. To limit the bias of headlines, we used headlines from the world news subreddit, which included a large set of daily headlines from a myriad of news sources; we believe that the level of bias was not very high. However, one way to improve this further is to include headlines from more sources so that our pool of headlines is even bigger.

We do not believe that our model is a weapon of math destruction. The outcome (movement in stock price) is easily measurable in the stock market, and so is our test error. Additionally, we do not believe that the predictions can have very negative consequences; the purpose of the project was not to use COVID features to create trading strategies, but was to explore the relationship between these factors and stock prices. For this reason, informed investors realize that there are many other factors (e.g. momentum, macroeconomic factors, etc.) that also influence stock prices. Therefore, as we do not believe our model has the ability to perfectly predict market movements, our predictions should not and were not intended to be used for trading purposes. Because of this, we do not believe our project creates a self-fulfilling feedback loop.

# 7 References


1. http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf
2. https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9
3. https://stats.stackexchange.com/questions/14099/using-k-fold-cross-validation-for-time-series-model-selection
4. https://towardsdatascience.com/heatmap-basics-with-pythons-seaborn-fb92ea280a6c
5. https://www.investopedia.com/articles/active-trading/052014/how-use-moving-average-buy-stocks.asp