Project Midterm Report

Project Overview

We altered our project to look at how COVID-19 news headlines can impact stock prices for certain sectors of the stock market. We are focusing on tech, healthcare, airlines, and retail, and choosing 10 stocks in each sector to analyze (a detailed list of stocks can be found under the same directory). Once we have analyzed the individual stocks to see their average exposure to COVID news, we want to look at the sectors overall and what kind of exposure they have to COVID. For example, based on our knowledge, we believe that the airline industry has been negatively impacted by COVID, so it will have negative exposure to COVID. "Positive" COVID related news headlines exist, which include titles such as "New Zealand COVID Cases Hit Zero" (see histogram below for average daily headline sentiment values). We are not overly concerned with bias in these headlines because there are a large variety of headlines per day from a myriad of sources. Also, even if there was bias, the future bias would be the same as the past bias, so it will not be too impactful in our analyses. We know that correlation does not imply causation and that COVID is not the only factor affecting stock prices, but with the COVID-19 pandemic affecting our lives, we thought it would be interesting to explore the exposure of stocks to COVID.

Description of the Dataset

Our dataset consists of the adjusted daily closing prices from 02/11/2020 (when the first news headlines regarding COVID began accumulating in the subreddit) to 10/30/2020 of 40 different stocks for the four industries listed above (prices downloaded from Yahoo Finance). We also collected news headlines from 02/11/2020 to 10/30/2020 from the Reddit World News subreddit with the flair "COVID-19" (by utilizing the PushShift API). Duplicated headlines were removed, and each headline received a sentiment score using the VADER package. We chose to use VADER because it is considered to be a good tool to analyze sentiments expressed in news, it gives how positive or negative the headlines are (intensity), and does not require any 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. Since there are more than one headline per day, we computed the average daily sentiment intensity by averaging the sentiment scores of all headlines published on each day. There are 263 days between 2/11/20 and 10/30/20; to account for missing stock data (no weekend/holiday trading), we decided to impute and use the adjusted closing prices for the previous traded day as the price for weekends and holidays. We have 263 days' worth of average daily sentiments for headlines, and 263 days of stock data for each stock we are considering. Thus, we have 2 features currently, which are the daily average sentiment and the previous day's stock price; however, this is subject to change if we think that we need to use more lagged periods (more than just the previous day's price). For each stock, we have 263 examples. Then, since we have 10 stocks in each industry, we will have 2630 examples for each sector.

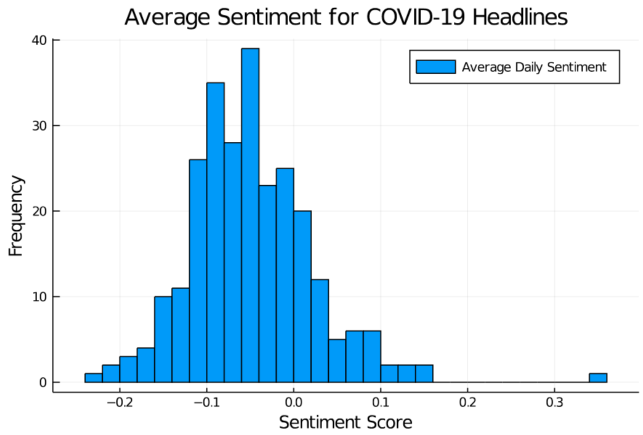
Avoiding Overfitting and Underfitting

To avoid overfitting and underfitting, we plan to fit our dataset using different models and choose the best model using cross validation. However, k-fold cross validation cannot be applied directly since we are exploring time series, and it would be meaningless to fit the model using future data and predict on previous dates. Instead, we will be using a cross validation method that is commonly used for time series, namely cross validation on a rolling basis. Specifically, we will divide the time interval evenly into 6 subintervals with 43-44 consecutive days each and divide the dataset into subsets accordingly. To begin with, we will be using the first subset as our training set and the second subset as our test set. We will then be using the first and second subset as our training set and the third subset as our test set. We will be fitting the model five times, with the last one using the first to fifth subset as our training set and the sixth subset as our test set. For each training set we have, we will use 80% of the data to fit the model and use the rest 20% for validation. We will then test the fitted models using validation sets and choose the model with least average test error. We are also going to monitor our train/test errors for each model we build to ensure that these values are relatively similar. If our features have low significance, we might consider looking at more or less lagged models (using more than just the previous day's stock price or headlines).

Testing the Effectiveness of Our Model

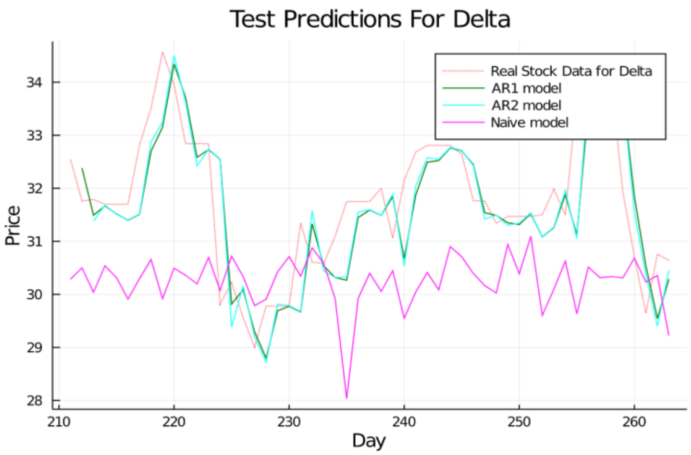
To test the effectiveness of our model, we will be testing the model we have selected on the test sets described above. We will use the average test error as an estimate of how effective our model is. Since the test sets are not used while training, we expect the test error to be a fair estimate of the effectiveness of our model.

Exploratory Data Analysis/Preliminary Analyses

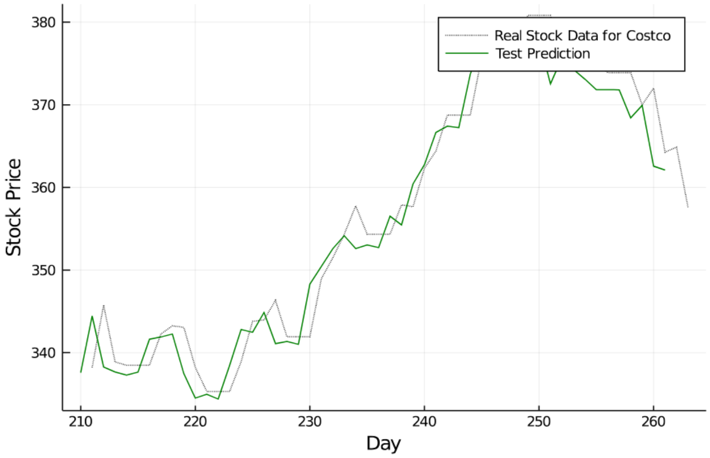


We have made an initial histogram for the average daily sentiment scores based on the headlines we accumulated. There are more headlines with negative scores, which we expected since there have been spikes in COVID cases, but there are also a decent number of positive/neutral headlines in our sample. Also, we will be considering the intensity of the scores as well (how negative or positive). The max sentiment score is 0.3504, the mean is -0.0634 and the min is -0.2313.

We decided to run a couple of simple models for exploratory purposes. First, we considered Delta Airlines stock, and ran a few models on it. For simplicity, we split up the data into an 80% training set and a 20% test set (no cross-validation was implemented yet). We first fit a simple time series model (naive model) on the daily average sentiment score with an offset term. From the prediction we generated from the test set, some of the ups and downs from our prediction follow the fluctuation of the actual stock price, but the model fails to reflect the long term trend of the stock's rise/fall. We then ran a simple autoregressive model (AR1) using the daily average sentiment score and previous day's adjusted closing price (and offset term) and an AR2 model with the adjusted closing price for the day before the previous day added. These two models yielded very similar predictions and have small training and test error (For the AR1 model, the Train Absolute Mean Error = 0.9439 and Test Absolute Mean Error = 0.2083). We graphed our test set predictions vs. the actual Delta stock prices for the same period of time. Absolute error was considered in this simplified model because absolute error in the stock price makes more sense to us than MSE does; however for future models we will also consider other error functions that we believe make most sense in the context of this problem.



However, using the same AR1 model for Costco yielded a much worse result. The model for Costco seems to overfit to the training set since the absolute mean error for the training and test sets are 3.1939 and 324.0240. We plan to address this issue by the ways described in the above sections (cross validation). Just from these preliminary results, we believe that it is possible that stocks in different sectors should be modeled differently; it is also possible that some companies are exposed more to COVID than others, which could be the reason that Delta's AR1 model performed decently well compared to Costco's. Also, using more complex models is another way to improve (see Further Steps).



Further Steps

Our preliminary analyses have given us a sense of where we want to go with this project and the types of models we want to consider. The rest of the semester will focus on building models for other stocks within the four sectors we are exploring. Once we have built some individual stock models, our goal is to generalize the effect of COVID headlines to each industry to see which of these sectors have the most exposure to COVID. We will use cross validation on a rolling basis to avoid over/underfitting and will consider other types of time series models besides the simple autoregressive model we used for exploratory purposes, such as autoregressive models (AR) of different orders, combinations of AR and linear models, smoothed models, and autoregressive integrated moving average models (ARIMA).