Data Collection and Format

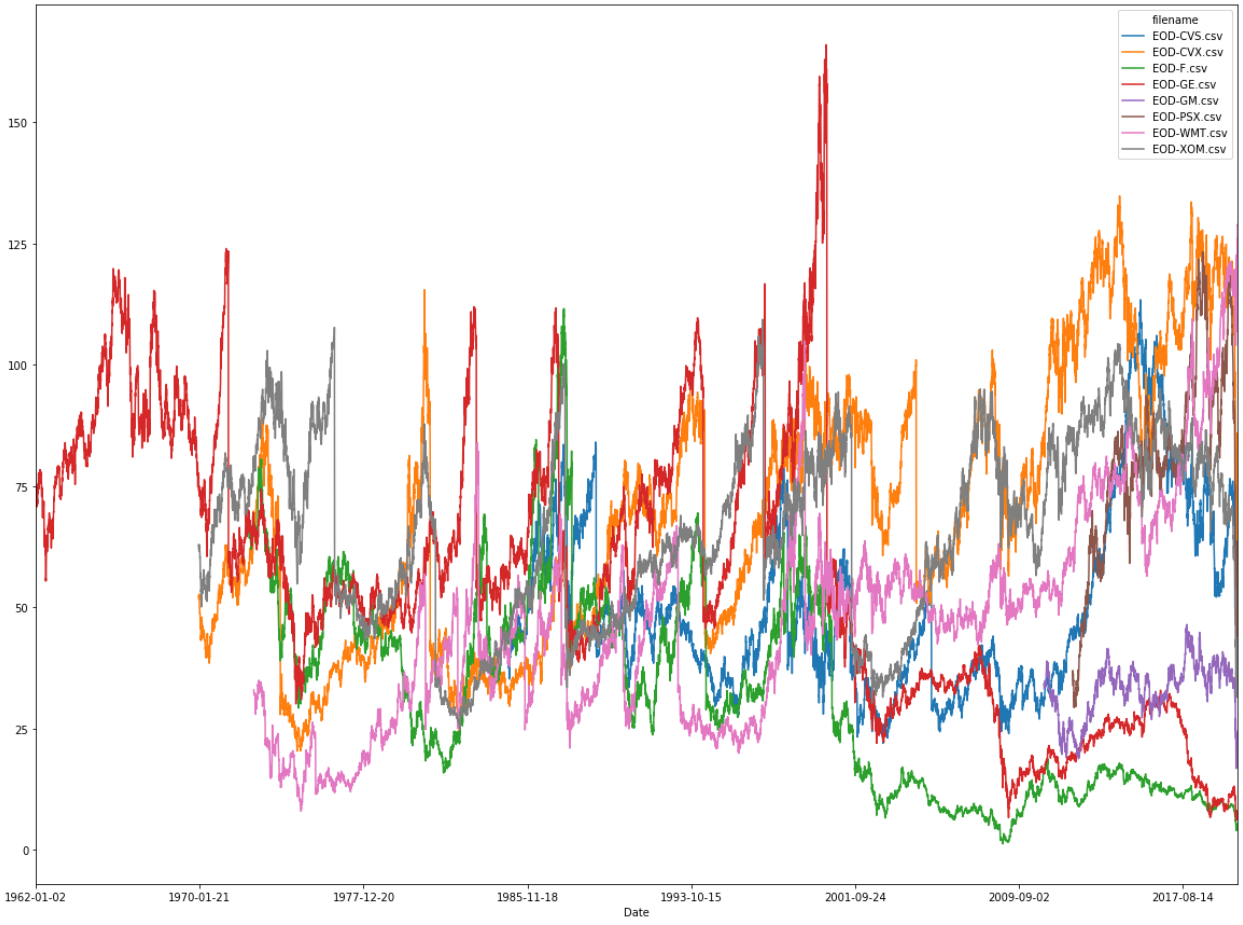
The data collected is a compilation of various stock market prices. This includes the opening and closing prices, along with the high and low per trading day. There are also adjusted prices for each and well as the volume. This was collected by searching through a database of datasets on Quandl. Since the Google and Yahoo finance APIs are depreciated, this was the next best option. They have also provided a python package to easily retrieve data from their site.

The data set right now is individually divided by companies. Each CSV file contains the stock history as far back as can go and up until the present.

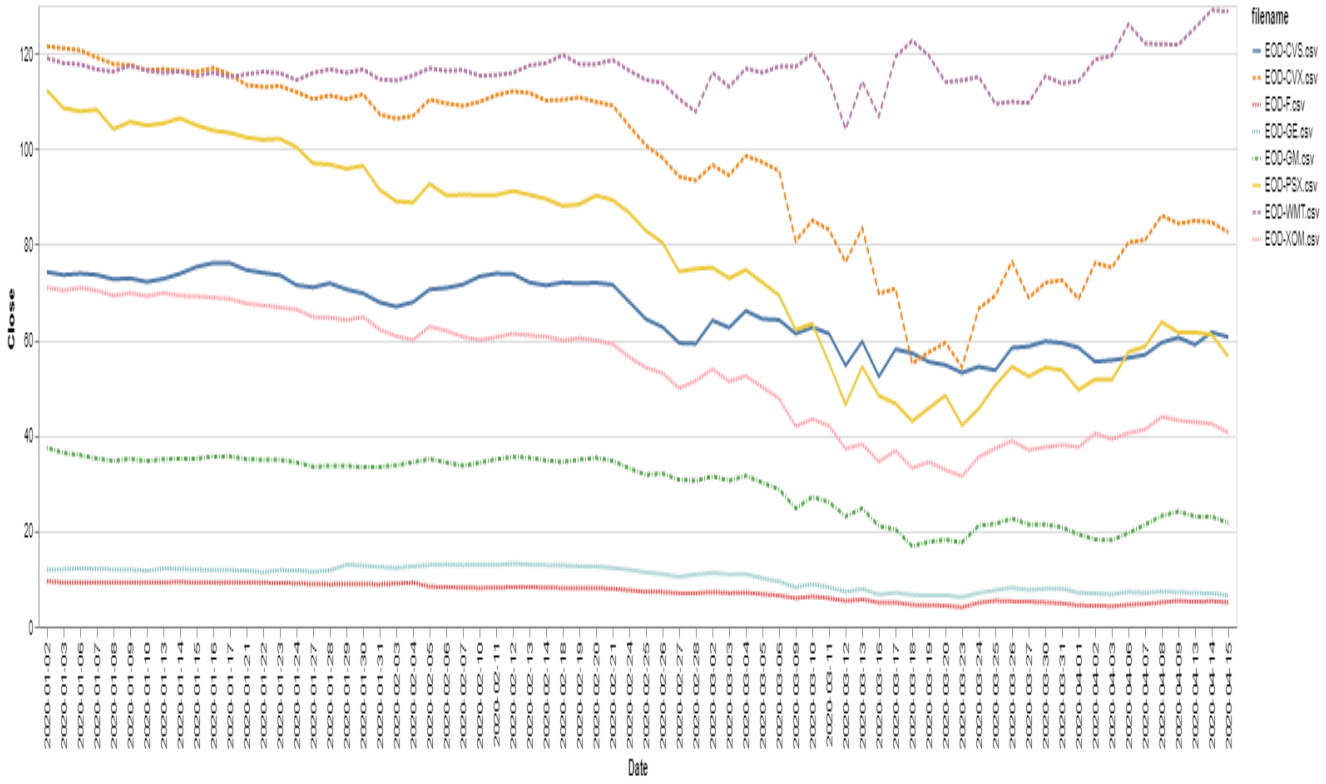
Descriptive Statistics

Generally, the standard deviation for the opening/closing day prices of the stocks were relatively the same across the time series. However, of course when the stock market showed high volatility, it was also reflected in the standard deviation. These areas are what my project will mostly focus on.

Data Analysis, Visualization, and Insights



Here is a sample chart of the stock performance of the top 10 companies in the Fortune 500 2015 edition. The most notable drops in performance were seen at around the 2000 mark, when the dot-com bubble happened and then again after the 9/11 attacks. Nearing the end of the chart we can see another instance of economic downturn.



Here is an Altair visualization that takes a closer look at the 2019-2020 period. Although the first instance of COVID-19 was reported at around the beginning of the year, the impacts of this isn't shown until around late February. This would be when the outbreak started to hit the United States.

Future Plans

As I mentioned before, the goal of this project is to focus on the periods of economic downturn. Many companies have risk projects of scenarios such as Brexit, 9/11, the housing market crash, and now, COVID-19. Similar to that, I plan to use historic stock data and create a machine learning model that can analyze these times. To move on to the next step, I would use various statistical models to determine which variables are the most important and then start to create the machine learning model (probably with Python's sci-kit packages).

Machine Learning Model

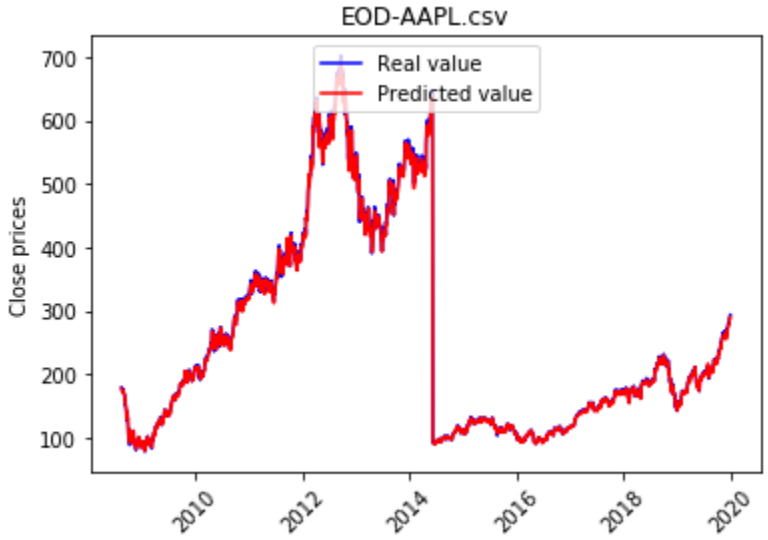
The model implemented was with Python's sci kit learn. I utilized a linear regression to analyze each individual companies historical stock price performance. The model was then used to forecast future data, particularly in early 2020 and on to the present.

Related Work

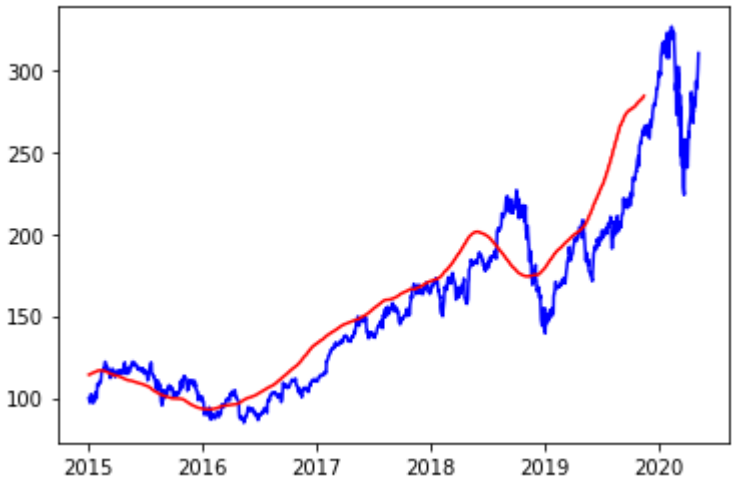
Although many stock market predictions have been done already, the purpose of this project was to explore specifically the ramification of economic downturn and using those portions to predict the worst case scenario given knowledge of an upcoming event.

Results and Conclusions

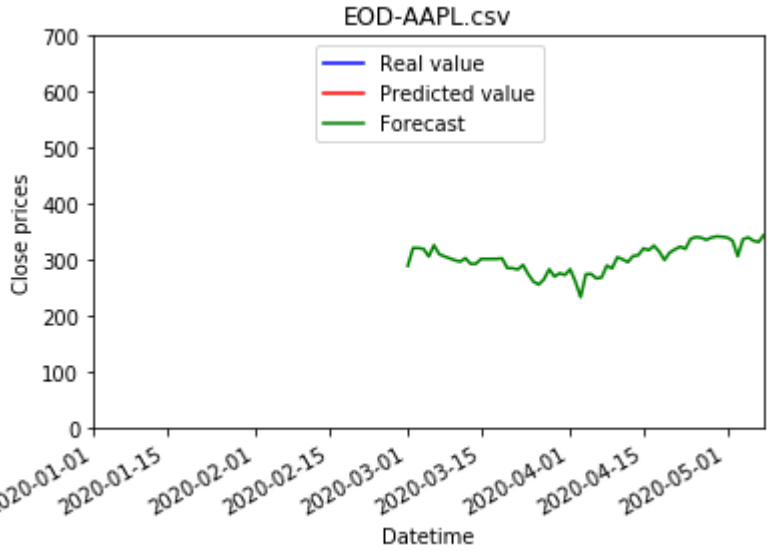
The goal was to ultimately see if the implemented model could predict the effects of the COVID 19 outbreak on the stock market. The linear regression model trained on labels such as volume, dividend, splits, and highs in order to predict the end of day closing prices. Pictured



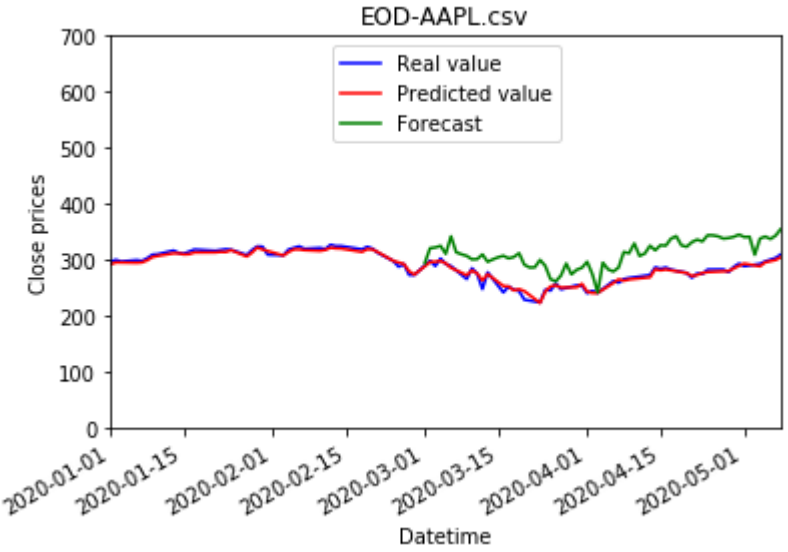
above is a sample graph taken from the model of Apple. Given these labels, the model had a score of 99%. This would be due to the labels having a high bias towards the "high" historical prices. This would cause a lot of overfitting for this time period, and the scores of the forecasted data later on reflect that. The mean squared error was around 12. The next step was to forecast the data past 2020. I initially used a moving average approach to better see how the forecast would go.



The red line indicates the smoothed average contrasted with the blue real value of the stock prices. Next, I trained once again with the previous linear regression model and forecasted the data past 2020. The graph of the results are below. This time, the model had a linear regression



score hovering around 80. Given this data, I then compared it to the actual present prices. Pictured below is the comparison. As you can see, the prices were generally higher than



the actual prices. This is due to the fact that the model could not account for the effects of COVID 19 on the economy. However, this does not mean that the info isn't useful. Given the range of error between these two means that we can apply the discrepancy/difference when we know another economic disaster event is going to happen. We can apply that to the predicted prices and find the worst case scenario change of prices. Some future adjustments to this would be to find more "random" labels for the initial linear regression model. Introducing these labels

would help reduce the amount of overfitting that occurred. Also, the model only included three previous economic recessions to help predict the next one (Gulf War recession, 9/11 and dotcom recession, housing market crisis recession). I would try to add more of these and find more datasets that include events that occurred before these to better analyze the next drop.

Acknowledgements

- Quandl.com
- Oftomorrow.github.io
- Towardsdatascience.com