Midterm Report - Covid US Stock Markets

Adrienne Zheng [yz839], Melissa (Xiaoyuan) Mao [xm76], Tainon Chen [tc457]

Dataset Description

We obtained historical prices for 191 S&P industry and sub-industry indexes, which span a large portion of the stock market (henceforth stock prices). We ran into an issue with locating a free data source, so we ended up using MarketWatch, writing a Python script to bypass the one-year data download limit.
Our stock data is structured the same way as conventional stock data, consisting of daily opening, closing, high, and low. Following convention, we used daily closing prices to represent stock movement.
Regarding the COVID-19 data, our current dataset consists of features including daily COVID cases, deaths, vaccination rates, and stringency indexes (stringency index is a measure of severity of governmental response) in the US, obtained from a variety of government sources. On top of that, we added first derivatives (velocity) and second derivatives (acceleration) for each feature.

Dataset Cleaning, Missing + Corrupted Data Handling

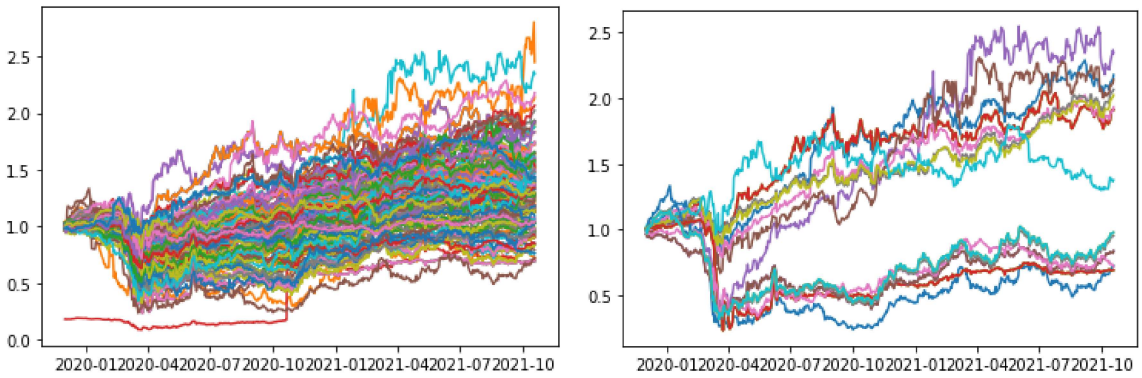
Some COVID-related features, especially the vaccination rate, don't have daily data, instead relying on weekly updates. Additionally, stock prices omitted data on weekends and holidays. For consistency, we decided to make the indices for both datasets daily. For missing vaccination data, we assume the same number of vaccinations on each missing day. For stock prices, whenever we encounter weekends or holidays, we forward fill the missing values with the latest trading day's data.
Additionally, we used news sources to determine the day that vaccination first started, and assumed that full vaccinations started 14 days following.

Overfitting and Underfitting

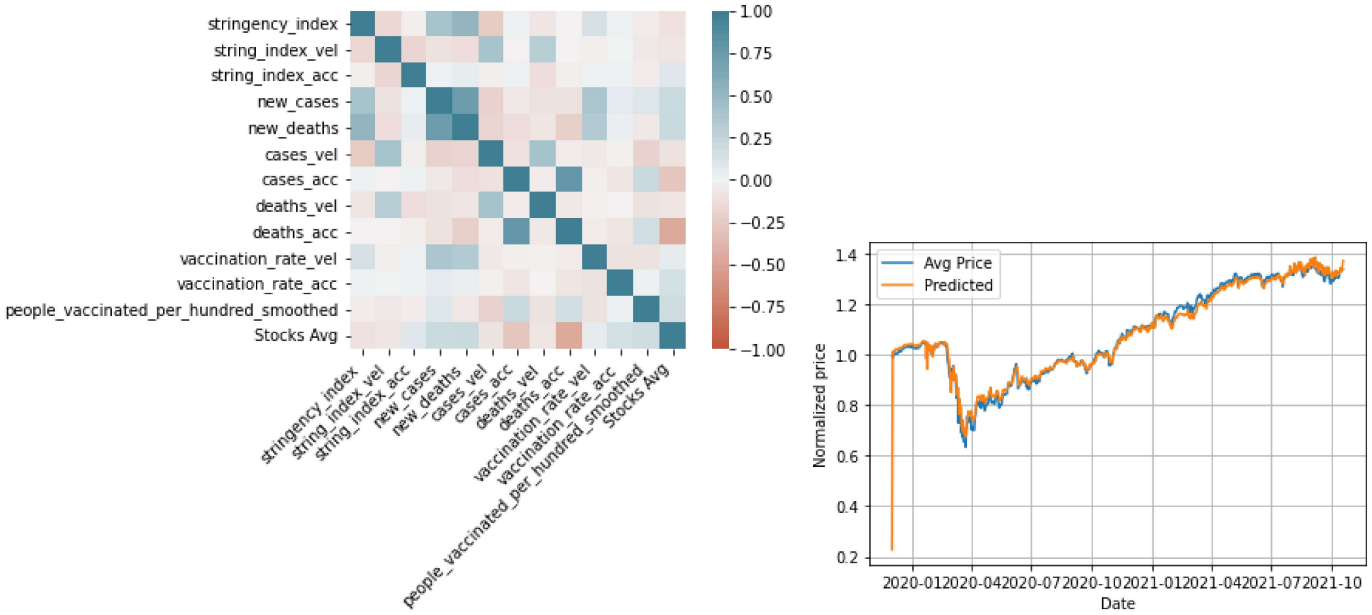
To avoid underfitting: We tested out combinations of parameters to determine which model works best. We decided which parameters to use mainly based on their correlation to stock prices, but tried a few combinations which seemed intuitive, as well. Also, we include the previous day's stock price in our dataset, as we would not otherwise have enough data to make an educated prediction about stock price.
To avoid overfitting: We try not to use too many different parameters, or ones which may be out of the scope of our project. Additionally, in the future, we may utilize ensemble methods such as random forest.

Descriptive Plots

Normalized stock prices plot: stock prices normalized to pre-COVID price. The first plot has information on all stock indexes. The second only includes a few of the highest or lowest-changing stocks. The index with most positive change was S&P 1500 Technology Hardware, Storage & Peripherals Sub-Industry Index, while the one with most negative change was S&P 1500 Oil & Gas Drilling Sub-Industry Index.

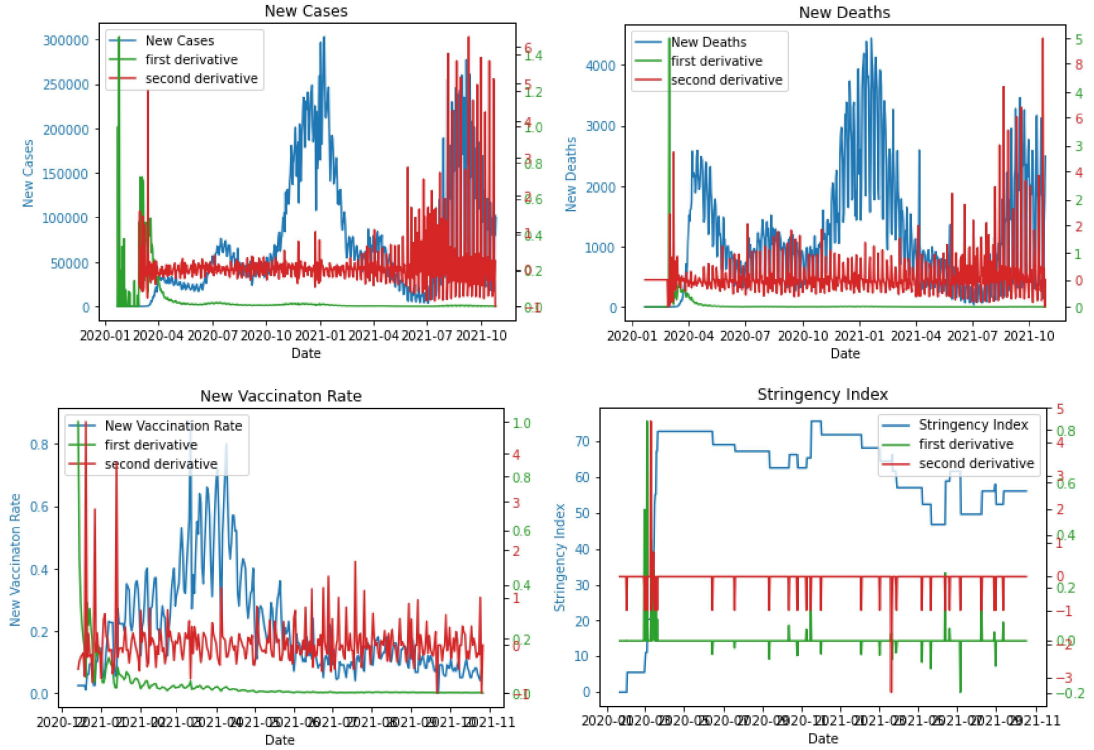


Correlation matrix: We computed the correlation matrix for our features and index prices. The corresponding heatmap is shown below. We also identified the features with the highest correlation to average index price movement, which turned out to be: vaccination rates, stringency index, new cases, cases velocity, and deaths velocity.



Preliminary regression plot: We fitted the aforementioned combination of features to the data, along with previous day's stock price. Here, we use the resultant weights to predict the stock prices given the feature data it was derived from.

Cases, deaths, vaccinations; and their first and second derivatives. Stringency index as well.



How Many Features and Examples

There are 13 total features: The 12 in our dataset, plus previous day's stock price.

There are 191 stock indexes, but we currently treat them separately. Each stock index has approximately 2 years' worth of daily data.

Preliminary Analysis

We try fitting least-squares regression models on different sets of features. The most basic one consisted of the base data (without derivatives), and previous day's stock data. The addition of previous day's stock data makes

sense because although covid-19 has some significant impacts on the markets, the stock price cannot be wholly determined by the covid-related features alone. Instead, covid drives the price movement, i.e. increase or decrease compared to the previous trading day. We then try adding more features and refitting the regression model, using these different combinations of features.
We then use the linear regression weights to predict the stock price, for the given stock data. The results are shown below. Note that the "None" case includes only information about previous day's stock prices.
We can see that adding more features decreases the MSE, thus improving the prediction. Using the highest-variance features seemed to lower the MSE the most, with least number of features. Also, all cases were an improvement over the "None" case.

	Cases	Cases'	Cases''	Deaths	Deaths'	Deaths''	Vacc	Vacc'	Vacc''	Strin	Strin'	Strin''	MSE
None													0.0015306
Base Data	X			X			X			X			0.0013186
Base Data + Vel	X	X		X	X		X	X		X	X		0.0012592
Base Data + Vel + Acc	X	X	X	X	X	X	X	X	X	X	X	X	0.0012532
3 Highest Correlation	X	X					X						0.0013528
5 Highest Correlation	X	X			X		X			X			0.0012665
7 Highest Correlation	X	X	X		X	X	X			X			0.0012615

Future Work
The current model is still rudimentary - we need more data and need to experiment with more different features. There are five changes we can implement for our current data.

1. Try out new features/new ways of representing features, such as higher-order derivatives, normalized new covid cases and categorical data such as different sectors (different industries tend to react differently to covid). And there may be some other ways to represent the stringency index as although the stringency index captures national policy pretty well, different stocks would likely respond to different specific policies in different ways, such as government lockdown policies, different social distancing measures implemented by different industries, etc.
2. Use better/new data: use S&P 500 data, which has more comprehensive industry information. This should also resolve some of the missing data issues - the S&P 500 may be better-kept than S&P 1500. However, this has the drawback of being less representative than the entire S&P 1500 index.
3. Try a more advanced way to deal with missing values. We currently assume the same number of vaccinations on each missing day. We could instead try to treat it as a time series and use a rolling average.
4. Try creating a generalized model for all industries, rather than a different model for each industry/stock index. Currently, we are trying to predict stock market performance from COVID data separately for each index/industry
5. Moreover, we are considering adding media sentiment as a feature. However, unless a pre-existing dataset exists, this could be extremely time-consuming to work on.

We can also try out more advanced methods:

1. Regularization. However, we may need more parameters and features in order to consider this.
2. Random Forest or Control Burn to better identify feature importance
3. Try different objectives, such as predicting daily price changes of the stocks instead of the daily stock prices, or predicting returns rather than actual prices.