Final Report - Covid US Stock Markets

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

Problem Description

COVID-19, which was first identified in late 2019, has had a severe impact around the world, impacting both individuals and corporations. It has also had similar effects on the US stock market, where a buildup of investor concerns resulted in a stock market crash starting February 2020 [1]. Since the end of the crash, however, the stock market has since been in recovery, reaching pre-crash levels around November 2020. Despite this apparent recovery, the ongoing pandemic and the stock market remain closely intertwined [2].

Different industries' stocks have also been affected differently by the pandemic. Some, like the hotel industry, were negatively affected, while the pandemic arguably might have a positive impact on tech industries. Recent publications have explored the varied effects of the pandemic on different industries. For example, S&P Global recently published an article using probabilities of default to conclude that airlines, leisure facilities, and oil drilling were negatively affected the most, while insurance and REITs (real estate investment trusts) were affected the least [3]. However, the relationships between specific aspects of the pandemic on different industries has yet to be properly explored.

To that end, the problem we are trying to solve in this report is twofold:

1. Use industry index performance to analyze the impacts of each COVID factor
2. Develop a model which uses COVID data to predict stock movement for different industries

The corresponding questions we are trying to answer are:

1. [Feature Importance] Which COVID factors have the greatest impact on each industry?
2. [Prediction] Which method predicts the stock price changes most effectively?

The inputs of our model consisted of aggregated COVID data, including stringency index, cases, deaths, and vaccinations, as well as their rates of change (first derivative) and convexity (second derivative). The responses for prediction were the day-to-day changes in industry index prices, for each of the 11 main S&P industries. Henceforth, the industry index prices will be referred to as stock prices.

Dataset Description

There were two main types of data used in our project: COVID data and market data.

For COVID data, we obtained data on infections, deaths, and vaccinations, due to their effectiveness in summarizing the progression of the pandemic. Vaccination data included number of total vaccinations, number of fully vaccinated people, and number of booster shots. We also used stringency index, a composite measure of government response severity, with the reasoning that performance of some industries may be affected by government actions, such as lockdown.

We obtained national COVID data on infections, deaths, and vaccinations from the CDC website [4][5]. We obtained national stringency index data from the University of Oxford website, per their Covid-19 Government Response Tracker [6].

For market data, we obtained data on S&P sector and industry index performance. In the S&P hierarchy, each firm falls into a specific industry. Similar industries are then grouped to form sectors [7]. For example, Ford and GM both fall into the automobile industry. The automobile industry, along with other industries such as hotels, restaurants, apparel, and luxury goods, make up the consumer discretionary sector, which covers non-essential goods. The eleven S&P sectors and their corresponding numbering in this project are: Communication services (0), Consumer discretionary (1), Consumer staples (2), Energy (3), Financials (4), Health care (5), Industrials (6), Information technology (7), Materials (8), Real estate (9), and Utilities (10). The stock data we used had price information for all the industries within each sector.

We obtained market data for the different industries from the MarketWatch website [8]. We opted to use the S&P 1500 industry indices, rather than the more widely-known S&P 500 indices. The S&P 1500 includes all of large, mid-sized, and small companies, which provides a more comprehensive picture of the various sectors and industries.

Dataset Manipulation

Dataset cleaning and consistency

Dates were the main indices for all our data, whether COVID-based or stock-based. Here are the three main issues relating to index consistency, and how we solved them:

- Missing COVID data: Some COVID datasets did not contain data for each date
  - For missing data having to do with infections, deaths, and vaccinations, we used a smoothing process which combined a seven-day moving window with the assumption that local data was mostly linear, the process used by the dataset providers. This allowed us to impute missing values so that they would resemble the actual values.
- Consistency between COVID and stock data
  - Stock data was not recorded on the weekends or holidays, as opposed to COVID data, which was recorded daily. We opted to prioritize the indices corresponding to stock data, by removing the dates during which there was no stock data.

Feature generation

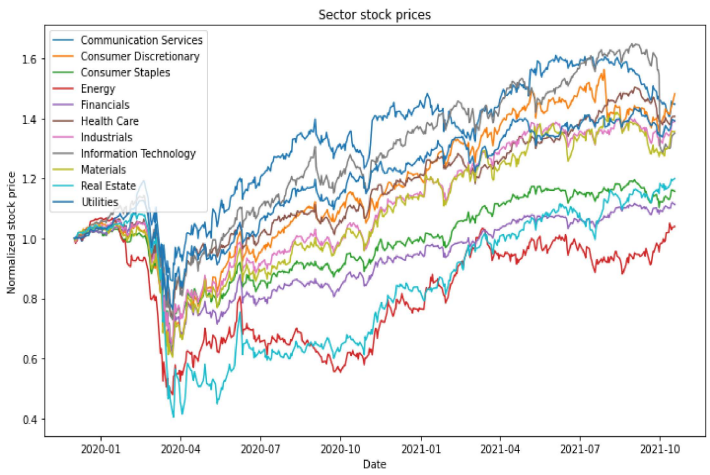
The source COVID dataset included only base statistics on total and new cases, total and new deaths, total and new vaccinations, and stringency index. We added higher-level derivative terms, with the expectation that changes in these variables would be more important in predicting changes in stock price, than the base data. We took the difference between consecutive dates for new cases, new deaths, new vaccinations, and stringency index, to obtain an approximation of the first derivative. Such data can also be interpreted as the "velocity" of the corresponding data. We then took the difference of these differences to approximate the second derivative, which can be interpreted as the "acceleration." The aggregate of base data, velocities, and accelerations served as the source of our features (our "X").

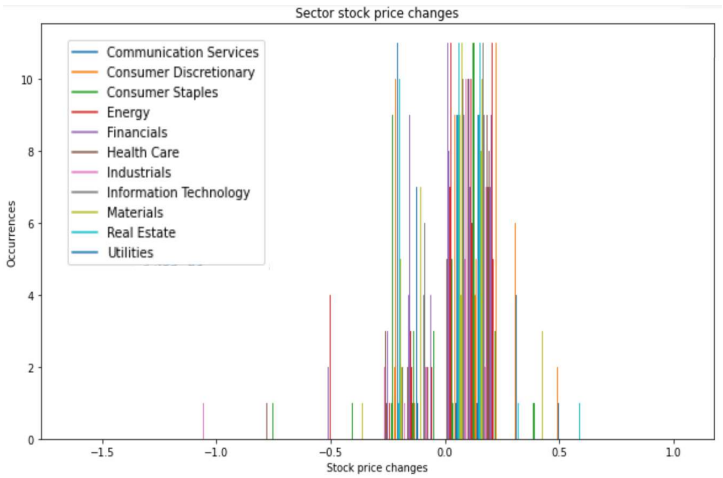
The only new aspect of stock data we added was the day-to-day price change. This was done to account for the fact that although COVID may have influenced actual stock prices, it was never the sole cause. However, we could still expect various aspects of COVID to have an effect on the day-to-day price change. Using the daily differences had the additional benefit of allowing us to treat the data as individual points rather than a time series. This enables us to shuffle the points around when generating our train and test sets, creating a model that utilizes data from all points of time without overlap between training and testing. This was our primary prediction response (our "y").

Note: Predictions on actual stock prices requires some sort of reference to past data, either the previous day's data or a historical moving average. However, trying to fit a model to predict the actual stock price always resulted in overemphasis on this historical data, with very little of the prediction attributed to COVID data.

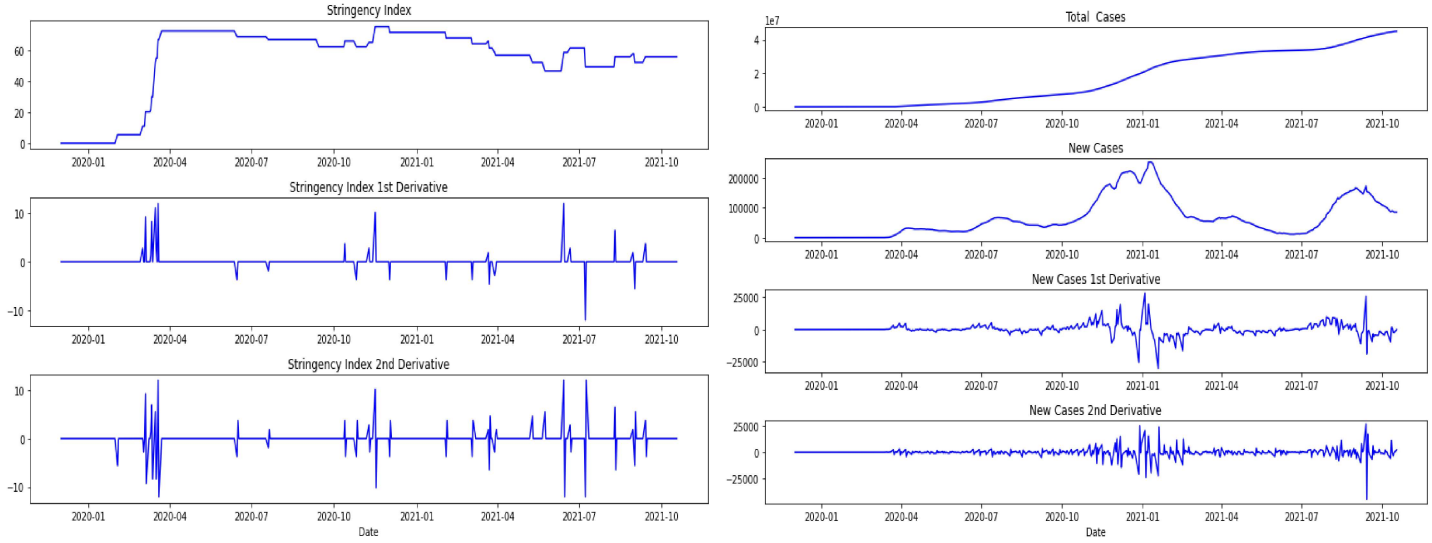
Preliminary Analysis

We first plotted the data to ensure validity, and get a general idea of what the different features and reponses looked like.



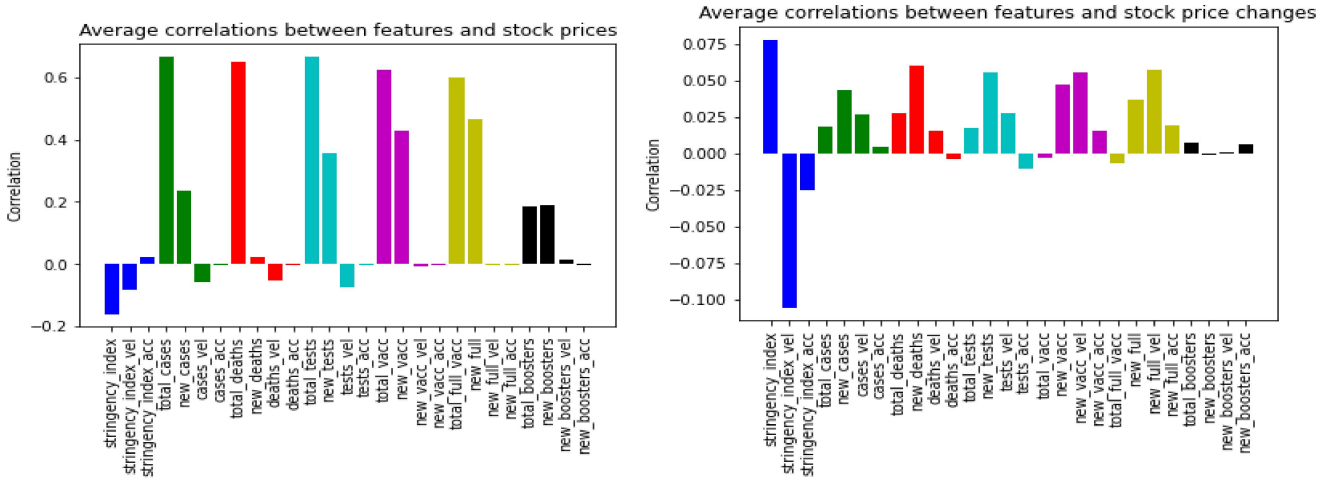


We can see that the prices for the sectors all experienced a sharp drop followed by general upward growth, representing the market crash and subsequent recovery. Although the degrees to which the price of each stock dropped initially and rose afterwards are different, the general shape is always the same. The stock price difference histogram offers an alternate view of the same data. We can also observe the extreme negatives, and large prevalence of the positives.



Stringency index sees a sharp initial rise, then a slow gradual decrease, while total cases, deaths, and vaccinations all monotonically increase (as expected: cases aren't removed from the total after recovery, and deaths and vaccinations are irreversible). New cases, deaths, and vaccinations typically experience one or two peaks where the value is significantly higher for a span of time. Their first and second derivatives reflect this.

Next, we calculated the correlations between each feature and the response for each sector, obtained by averaging the correlations of its industries. This was done with the intention of identifying which features were more and less important, without using any complex methods.



On the left graph, we see pretty strong correlations between total cases, deaths, and vaccinations, and to a smaller extent, new cases, deaths, and vaccinations, with the stock prices. The total cases, deaths, and vaccinations likely aren't very useful, however. Keep in mind that the totals all monotonically increase, similar to how the stock price (barring the initial drop) also increases pretty consistently.
On the right graph, we see that the correlations between features and price changes are significantly weaker, to the point that nearly all are insignificant. The one with largest magnitude is stringency index first derivative, which has a very weak correlation of around -0.1. This is essentially the change in stringency index. The relationship can be interpreted as follows: A decreasing stringency index (restrictions being lifted) is correlated to higher stock price changes, which seems pretty reasonable.

In the individual graphs for the sectors, we also observe that the correlations between features and response were all very low, usually with magnitude below 0.1, with maximums around 0.15. The correlations also didn't always make sense in terms of explaining the stock price changes. For example, stringency index correlation was usually positive, meaning that high stringency index (more restrictions) was correlated with higher stock price changes.

How we Tried to Solve the Problem
Our end goal is to identify important features so that we can better study which COVID-related factors could potentially drive the stock price changes, and eventually construct a model to predict the future price changes with the selected features by minimizing the prediction error (MSE).

We decided to apply a few methods that were discussed in class to our data to better interpret our data. There were three main ones, which will be discussed in subsequent sections:

- Unregularized linear models: Linear regression
- Regularized linear models: Ridge regression, Lasso regression and Huber regression
- Tree-based methods: Random forest, Control burn

All these methods were used with the intention of predicting stock price changes from the given features. However, a subset of these methods were also used to analyze feature importance: Lasso regression, Random forest, and Control burn.

Linear Regression without Regularization
Linear regression is used to model the linear relationship between input features and desired response [9].

We decided to use unregularized linear regression as an initial test for how well linear models can predict our data, and later as a benchmark against which to analyze our regularized linear regression methods. For ease of comparison, we included the data and plots of linear regression along with the regularized linear models in the section below.

Linear Models with Regularization
Regularized linear regression imposes additional constraints on the coefficients to prevent overfitting [10].

We fitted a few regularized regression models and compared the output with unregularized linear regression. We trained three different regularized linear models: Ridge, Lasso and Huber Regression, and used mean squared test error to measure prediction performance. To tune the regularization parameters (referred to as alpha), we further split the full dataset into a validation and a test set and used the SciPy optimizer (scipy.optimize.minimize) to find the parameter with the minimum mean squared validation error. We then fitted the optimal Ridge, Lasso, and Huber Regressions to each industry dataset to identify which one performed the best.

We first examine the most important features of each industry, which makes use of LASSO Regression's property of providing sparse solutions. We looked at the three features with the biggest absolute coefficient generated by LASSO Regression, the results of which are summarized in the table below.

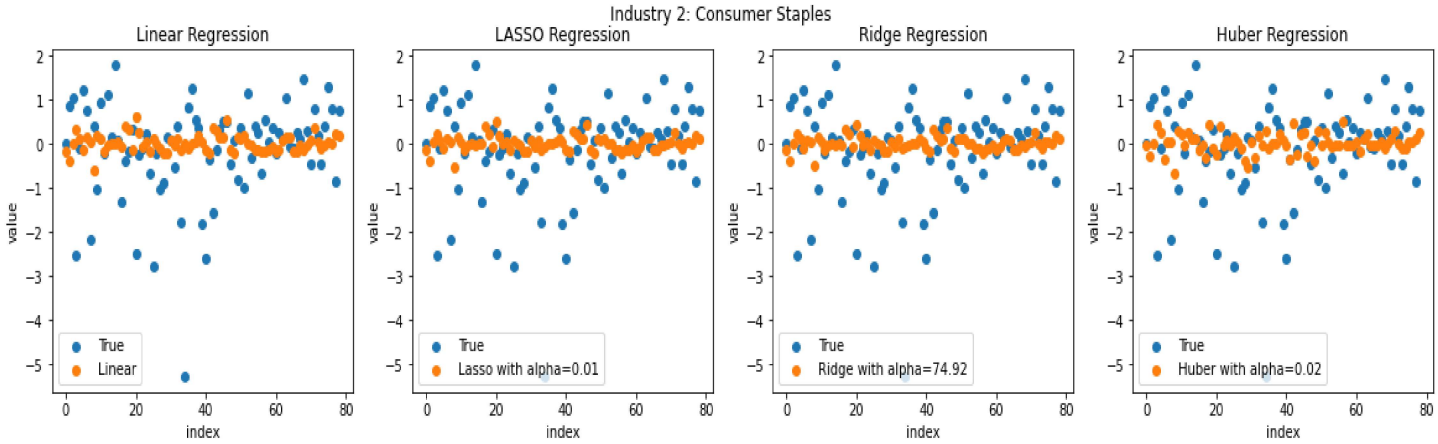
Industry	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities
Top Feature 1	new_tests	stringency_index	stringency_index	tests_vel	stringency_index_acc	new_tests	new_vacc_vel	new_boosters	new_cases	new_tests	stringency_index_acc
Top Feature 2	stringency_index	new_full_vel	stringency_index_acc	new_deaths	tests_vel	stringency_index	stringency_index	new_boosters_acc	new_deaths	new_cases	new_full_vel
Top Feature 3	new_vacc_vel	new_full_acc	new_vacc_vel	tests_acc	new_vacc_vel	new_vacc_vel	tests_vel	new_boosters_vel	tests_acc	stringency_index_vel	cases_acc

We can see that features related to stringency index and vaccination rate are the most common top features among all the industries. This makes sense, given that for most industries, government policy is directly related to their operation and therefore profit. Additionally, the distribution of vaccines and boosters influences society's confidence in economic recovery, which also impacts companies' operation and profit. However, for Information Technology, the top three features are all related to boosters, which is a bit counterintuitive since boosters didn't appear until pretty recently, relative to stock data. This may be attributed to the fact that linear models may not capture the trend well for this particular industry.

We then examine the results from predicting stock price changes, using each of the four linear methods. The below table shows the mean square test error for all linear models in each industry and the best model for each industry.

Industry	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities	Mean MSE
Linear	1.4	1.32	1.38	1.32	1.32	1.33	1.34	1.35	1.31	1.21	1.41	1.33
Ridge	1.39	1.27	1.33	1.28	1.29	1.29	1.34	1.28	1.28	1.21	1.34	1.3
LASSO	1.4	1.26	1.34	1.29	1.29	1.32	1.32	1.3	1.3	1.21	1.3	1.3
Huber	1.3	1.29	1.4	1.27	1.34	1.28	1.42	1.31	1.31	1.25	1.3	1.31
Best model	Huber	LASSO	Ridge	Huber	LASSO	Huber	LASSO	Ridge	Ridge	Linear	LASSO	Ridge

It is clear that unregularized linear regression typically performed the worst, relative to the regularized models. This could be because we have fewer stocks in each industry relative to the number of features and some features are correlated with each other. In addition, due to the volatile nature of the stock market, the robust models, LASSO and Huber, tend to perform slightly better in general as they are not affected by outliers.



The above scatter plots show true values and predicted values for the example industry Consumer Staples. We can see that the plots for unregularized regression and Huber regression deviate more from the true values, which makes sense given their higher MSEs. Ridge and LASSO regression yielded comparably similar results.

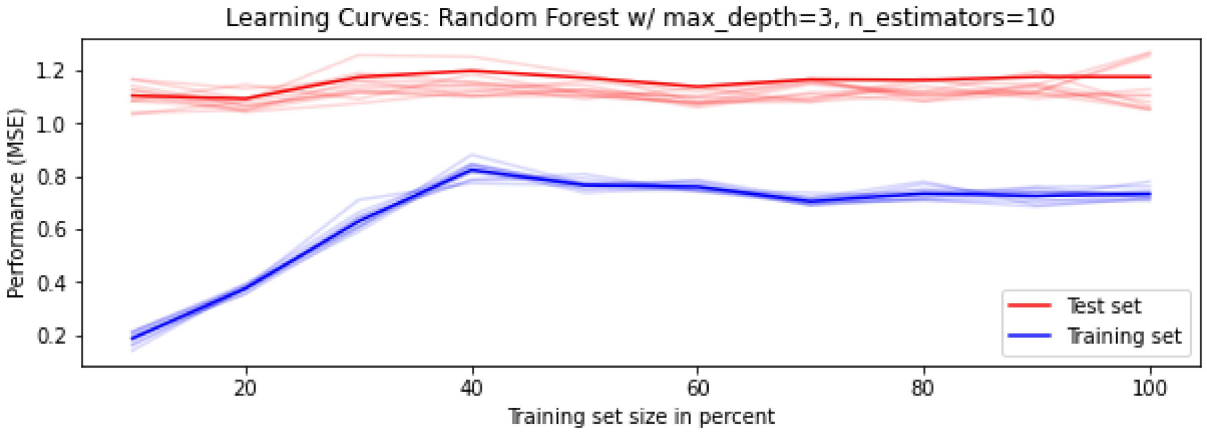
To gain a better understanding of the relationship between COVID and stock price changes by better analyzing feature importance, as well as to improve our prediction results, we turned to more complex nonlinear models.

Random Forest and Control Burn

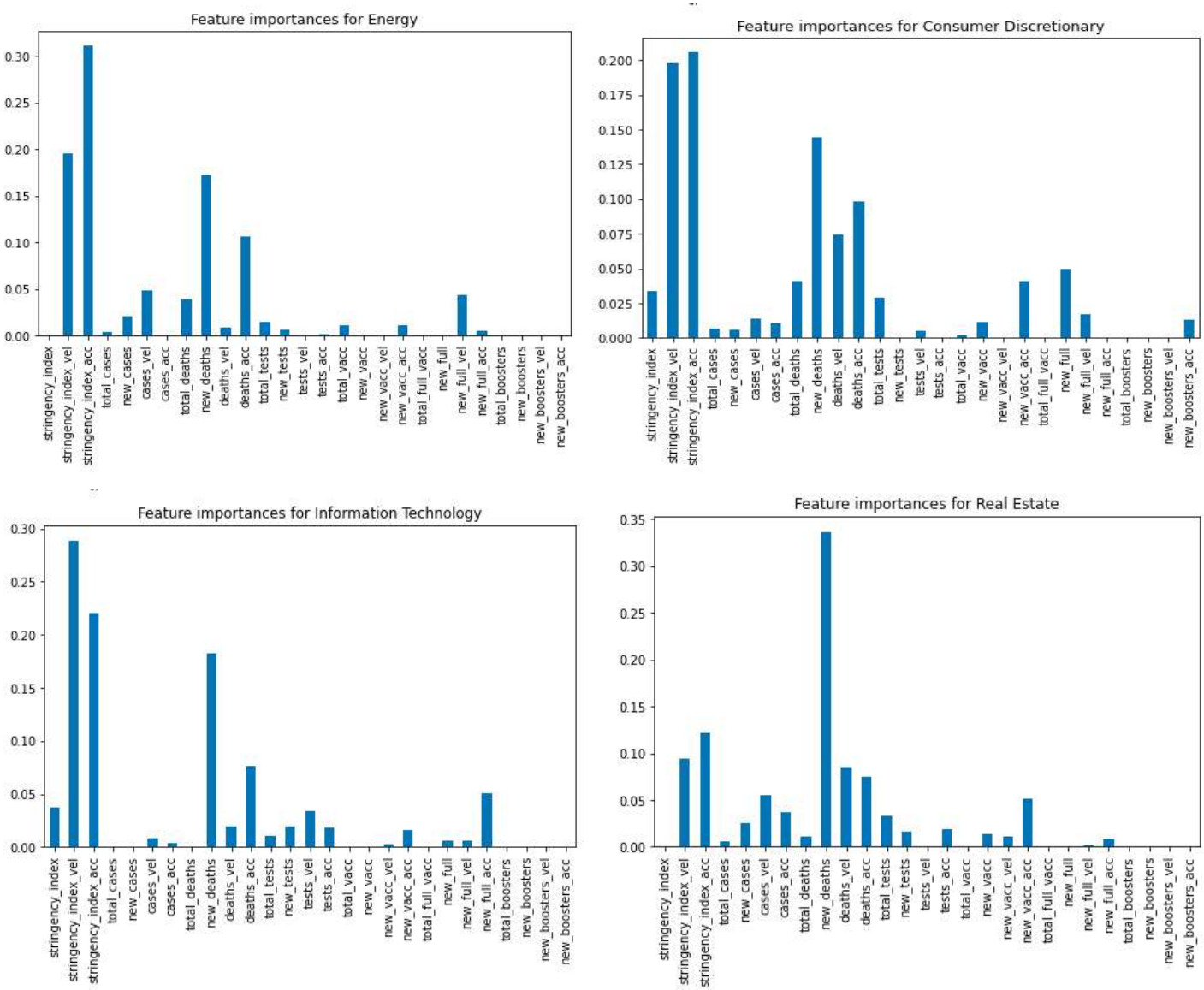
Random forest and control burn are both based on decision tree-based ensemble methods, which aggregate the decisions made by multiple decision trees to obtain a better model.

To better understand what features contribute to the stock price changes, we decided to use Random Forest and Control Burn regressors to extract feature importance. Although multiple factors should be considered in order to analyze the price movement, in order to avoid overfitting, it is important to identify a subset of features that can best explain the price pattern. As we have discussed in class, Control Burn can prevent overfitting and select a sparse subset of important features from a tree ensemble. By limiting the tree depth and the number of trees, Random Forest will be also able to select a subset of features that minimizes the difference between the training error and the testing error, and thus avoid overfitting.

After parameter tuning, we chose a maximum depth of 3 with 10 trees, and performed the regression on all 11 different industries. The resulting learning curve with this choice of parameters turns out to be the smoothest with the training errors and testing errors being the closest.



Plots which illustrate feature importance for various industry sectors, via random forest, are shown below.



By observing these plots for all 11 industries, we observe that stringency index velocity (rate of change), stringency index acceleration (convexity), and new deaths tend to be the dominating factors across the industries that affect price change.

The feature selection results from control burn are mostly similar, with the only exception being new deaths velocity replacing new deaths.

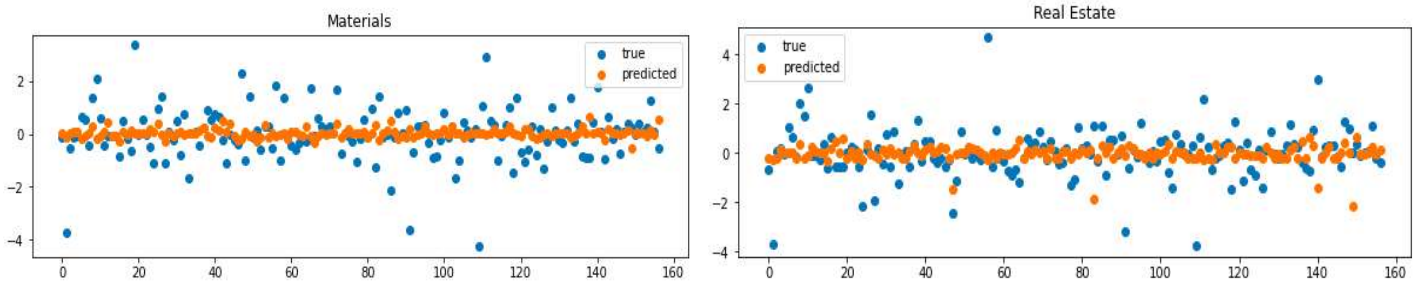
The result makes sense, given that business performance during COVID is largely dependent on government intervention and how cautious consumers are regarding the pandemic. The stringency index velocity and acceleration are the most important factors affecting all industries, but with magnified results in the consumer discretionary, health care, and utilities sectors. Stricter lockdowns and increasingly limited operational capacity tend to restrict people's mobility and businesses' ability to operate, thus negatively impacting those providing non-essential services and goods, and the overall utility consumptions across the country. It has also been studied by multiple researchers that restricted mobility and quarantine could cause health issues, thus it could be reasonable to assume that health care providers are facing a rising amount of claims. For the information technology industry, stringency index velocity also has been listed as the most important feature. Given the blooming of tech stocks during the pandemic, we may assume that in this case, the stringency index could have a positive impact on the industry as restricted mobility could directly result in people spending more time online. In fact, from the results generated by Control Burn, stringency index velocity is also considered the most contributing factor for almost all industries, followed by death velocity and stringency index acceleration.

On the other hand, the number of deaths can also serve as a measure of severity for the pandemic, as people are more likely to take precautions against warnings and events that could be life threatening. The industry affected the most by new deaths is real estate. Rising number of deaths could cause panic and stop people from traveling, or simply spending time outside. During the severe stage of the pandemic, hundreds of shops and office buildings were left unused, which makes it reasonable that the real estate industry would be significantly affected by the new deaths. Other industries including health care and utilities are also impacted by new deaths.

Below is a collection of MSE from the 11 industries from predicting stock price changes using random forest and control burn regressors. The average random forest MSE was 1.1568, and the average control burn MSE was 1.3506. It can be observed that random forest achieved better results across all industries compared to both the linear models and Control Burn. For reference, unregularized and regularized linear regression had average MSEs ranging from 1.3 to 1.35.

Industry	Communication Services	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Industrials	Information Technology	Materials	Real Estate	Utilities
RF_MSE	1.089	1.199	1.116	1.160	1.192	1.177	1.067	1.194	1.193	1.110	1.228
CB_MSE	1.382	1.380	1.359	1.347	1.334	1.335	1.353	1.359	1.385	1.285	1.369

The below scatter plots illustrate true versus predicted price changes for random forest, for the same four industries. Blue dots represent the true values, and the orange dots represent the predicted values.



We can see that the results similarly resemble those from linear regression, but with improvements. The improvements are more noticeable in some industries, such as materials and real estate.

Questions and Answers

The data analysis questions we were trying to answer are:

1. [Feature Importance] Which COVID factors have the greatest impact on each industry?

    - We analyzed the most important features for each industry in the previous sections. To summarize:
    - For linear models (regularized and unregularized), stringency index and vaccinations-related features were the ones with greatest impact
    - For random forest and control burn, stringency index velocity, stringency index acceleration, and new deaths were the factors with the greatest impact
    - It makes sense that stringency index is present as a key feature in both methods. Stringency index can be viewed as a summary of government policy, which probably most directly affects company operation and profitability.
2. [Prediction] Which method predicts the stock price changes most effectively?
    - We tested out unregularized linear regression, regularized linear regression, random forest, and control burn. We chose these methods because of their added functionality in helping to determine feature importance.
    - We found that random forest predicted the stock price changes most effectively, with an average MSE of approximately 1.15. This was a pretty significant improvement over the other methods, which had MSEs within the range of approximately 1.3 to 1.4.

Model Deployability

We would recommend deploying this model alongside other ones which predict stock profitability, or incorporate this model's findings into another more comprehensive model. This model is not intended to be deployed independently. However, it may provide valuable insights about how COVID-19 impacts the markets, and how investors could cope with the pandemic in the future when combined with other stock-predicting models.

Although our model seems slightly capable of predicting stock price changes, particularly random forest, it should be noted that the results by themselves aren't reliable enough to deploy on its own, for the sake of generating profit in the stock market. However, this shouldn't be entirely surprising, given that stock price changes are dependent on many many other variables, not solely COVID data, which we analyzed in this project.

Fairness and Weapon of Math Destruction

Our model does not make use of any insider or illegal information - it only employed publicly available COVID and stock market data. In a sense, this makes it fully legal, and thus "fair" and "unbiased". However, fairness in a social context is not very applicable to our project. In terms of Weapons of Math Destruction, the outcomes of our analysis and model are numerical-focused, and thus objectively measurable. The result we achieve does not take any stand on any particular social issues and our sources of data mostly depend on no customized survey nor personal response, and thus has no negative consequence and won't cause any feedback loop.

Further Improvements

We can expand the scope of this project by looking for patterns in how the stock market responds to epidemics/pandemics, as well as the subsequent recovery. This can be difficult, however, as pandemics do not happen very often, and epidemics vary drastically in scale and severity.

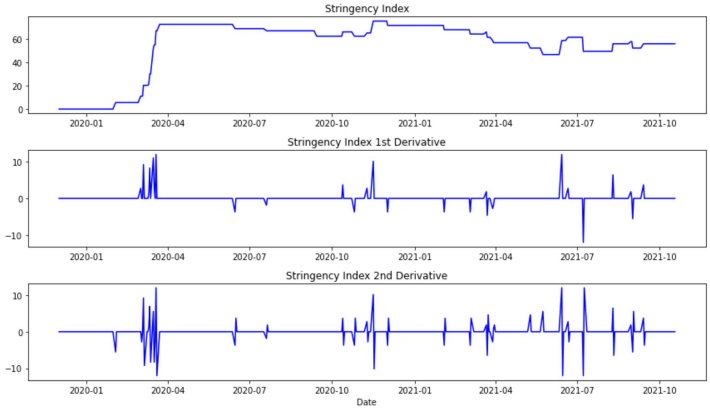
Using a neural network may be effective in fully capturing the complex relationships between our features and outcomes. We omitted using neural networks since their results are less interpretable than those of linear regression or tree-based methods.

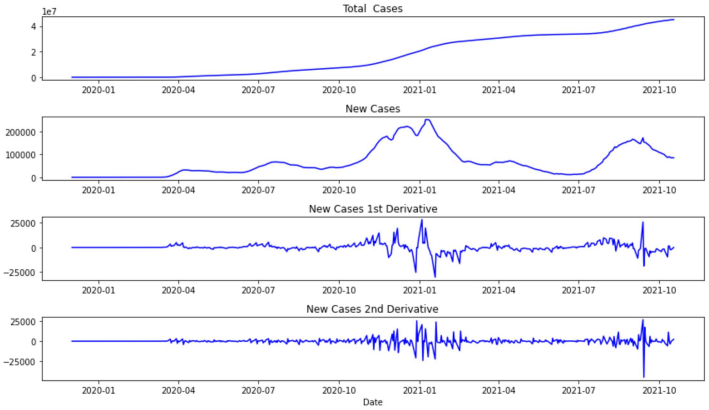
Linear regression effectiveness may have been limited by the relatively small number of industry stock prices relative to the high number of features we used. We could potentially obtain better linear regression results by using the raw data for multiple companies inside the industry to obtain more stock price points. The reason we decided on using industry indices was due to simplicity.

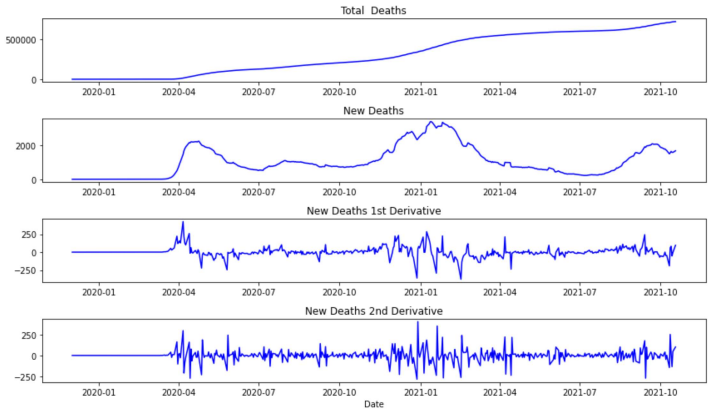
Bibliography

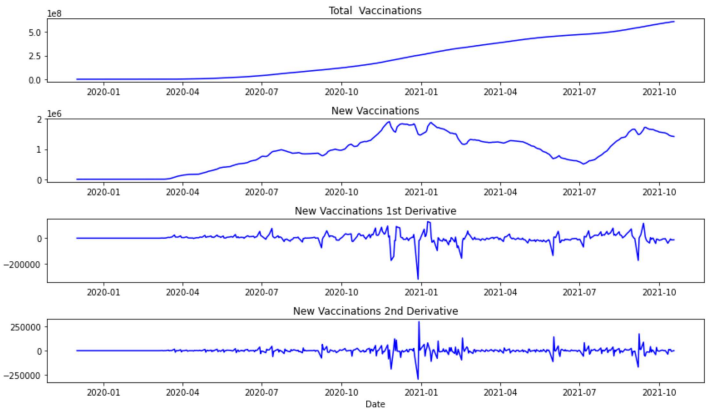
[1]: Making Sense of Stocks' Rude Awakening to the Virus Scare; Feb 2020; Wall Street Journal; https://www.wsj.com/articles/making-sense-of-stocks-rude-awakening-to-virus-scare-11582637181
[2]: COVID-19 and the march 2020 stock market crash. Evidence from S&P1500; Jan 2021; Finance Research Letters; https://www.sciencedirect.com/science/article/pii/S1544612320306668
[3]: Industries Most and Least Impacted by COVID-19 from a Probability of Default Perspective – September 2020 Update; Sep 2020; S&P Global Market Intelligence; https://www.spglobal.com/marketintelligence/en/news-insights/blog/industries-most-and-least-impacted-by-covid19-from-a-probability-of-default-perspective-september-2020-update
[4]: United States COVID-19 Cases and Deaths by State over Time; Dec 2021; CDC Data; https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36
[5]: COVID-19 Vaccination Trends in the United States,National and Jurisdictional; Dec 2021; CDC Data; https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3yt2
[6]: Oxford COVID-19 Government Response Tracker; Dec 2021; University of Oxford; https://covidtracker.bsg.ox.ac.uk/
[7]: Sectors & Industries Overview; Dec 2021; Fidelity Investments; https://eresearch.fidelity.com/eresearch/markets_sectors/sectors/sectors_in_market.jhtml
[8]: List of Indexes; Dec 2021; MarketWatch; https://www.marketwatch.com/tools/markets/indexes/a-z/S
[9]: Linear Models and Linear Least Squares; Oct 2021; ORIE 5741 Lectures; https://people.orie.cornell.edu/mru8/orie4741/lectures/linear.pdf
[10]: Regularization; Oct 2021; ORIE 5741 Lectures; https://people.orie.cornell.edu/mru8/orie4741/lectures/regularization.pdf
[11]: Trees; Oct 2021; ORIE 5741 Lectures; https://people.orie.cornell.edu/mru8/orie4741/lectures/trees.pdf
[12]: ControlBurn; Oct 2021; ORIE 5741 Lectures; https://people.orie.cornell.edu/mru8/orie4741/lectures/ControlBurnSlides.pdf

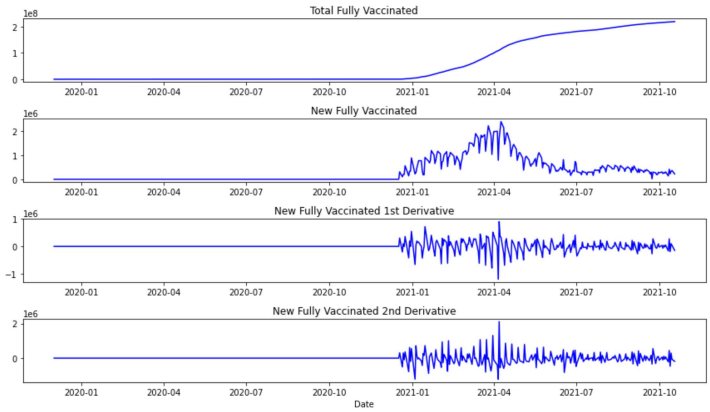
COVID data visualizations

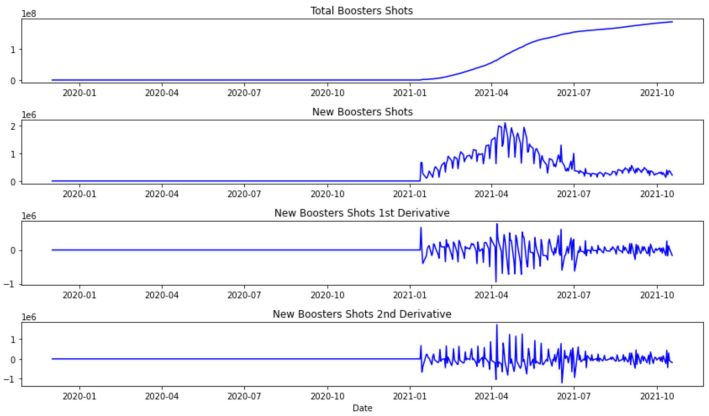




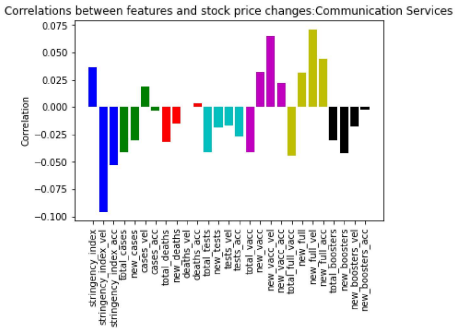


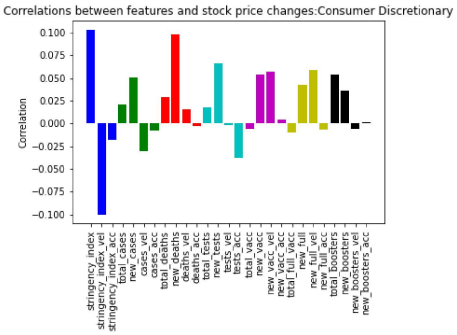


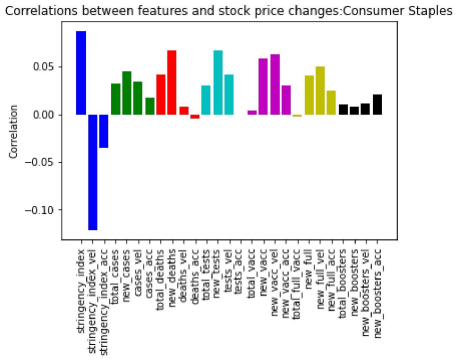


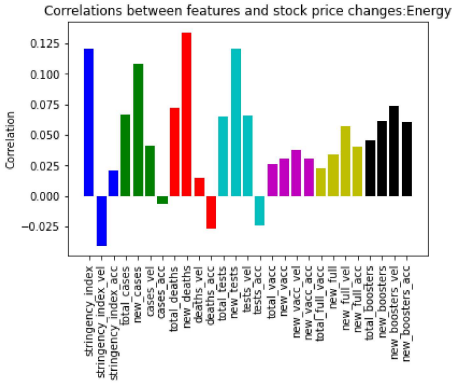


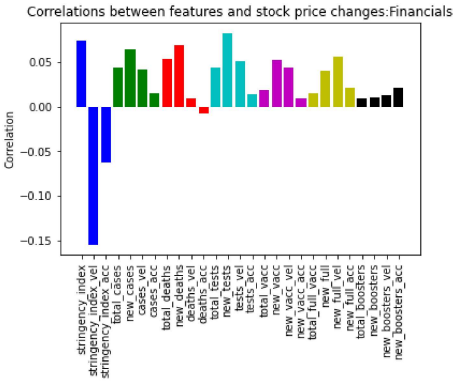
Correlation plots for all sectors

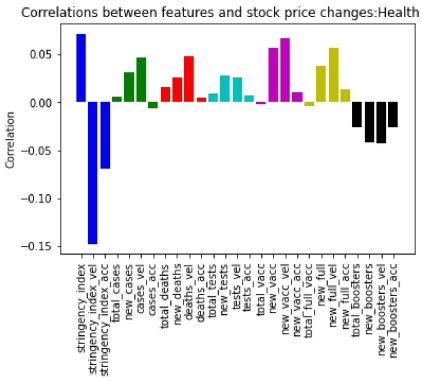


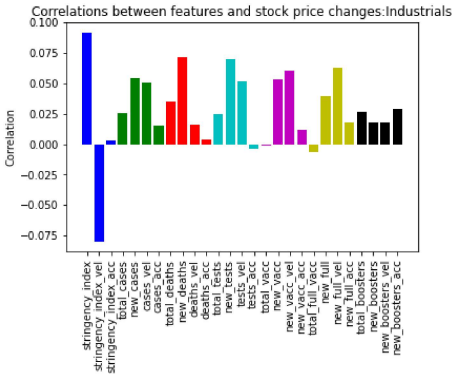


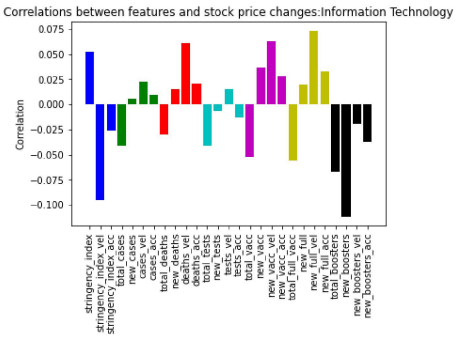


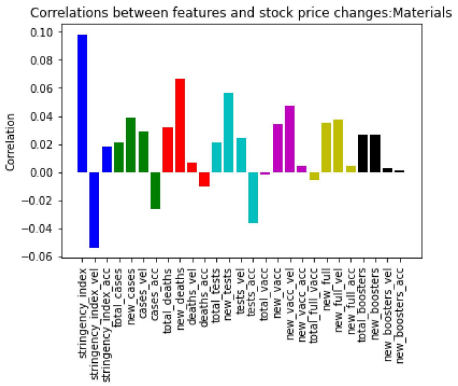


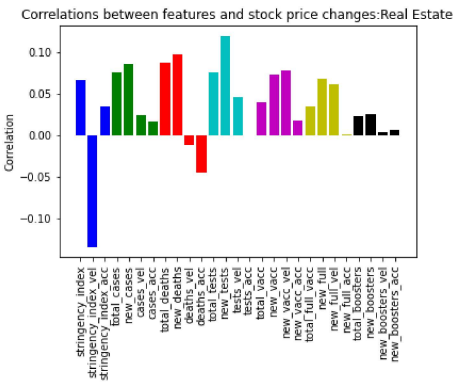


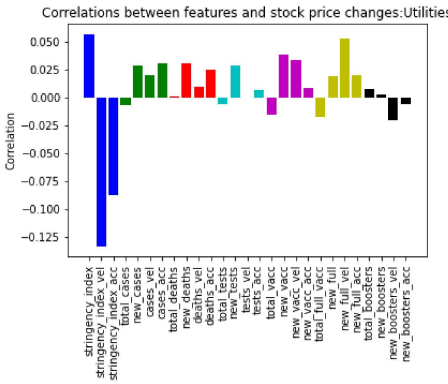






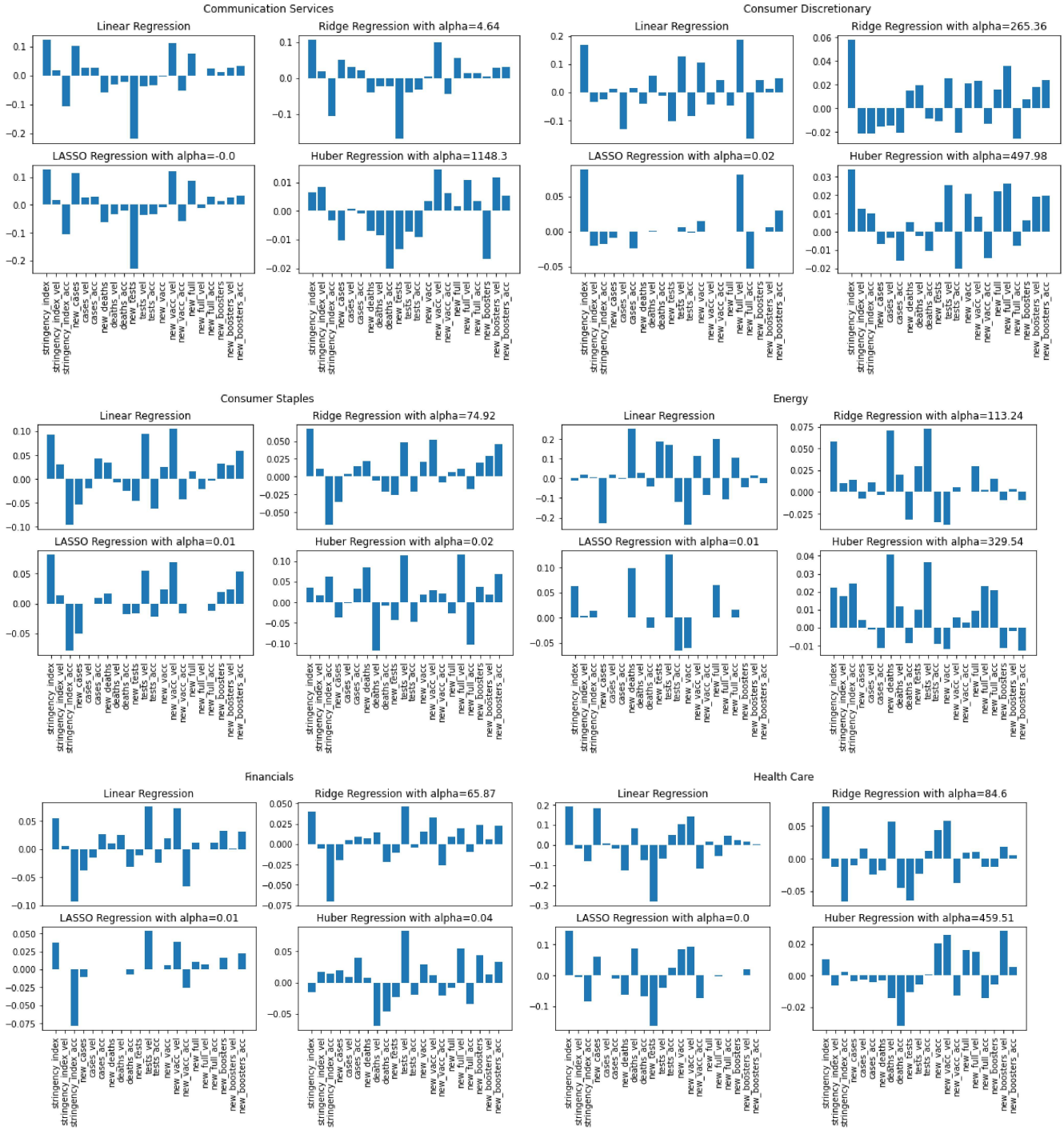




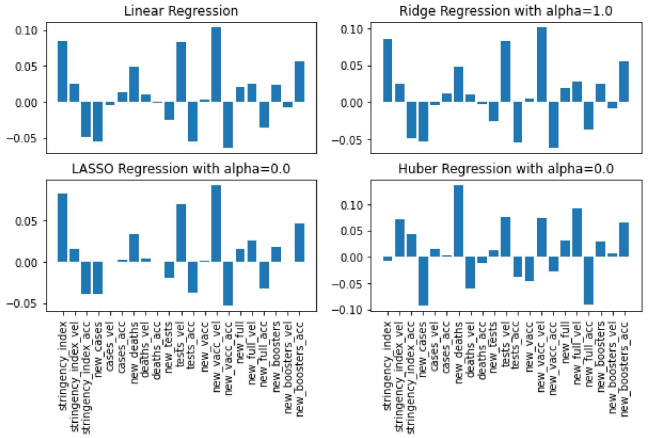


Plots for linear models

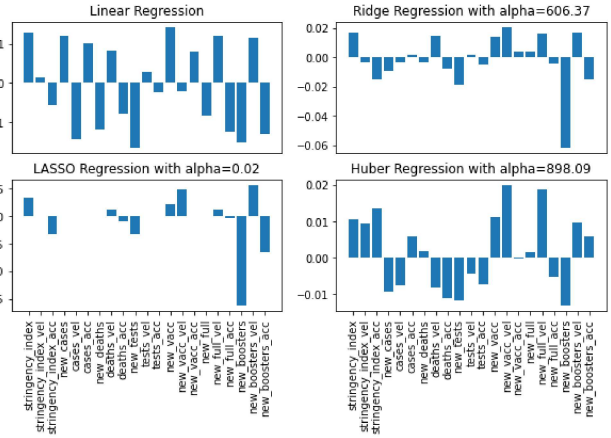
Below are plots of coefficients for Linear, Ridge, LASSO, and Huber Regression.



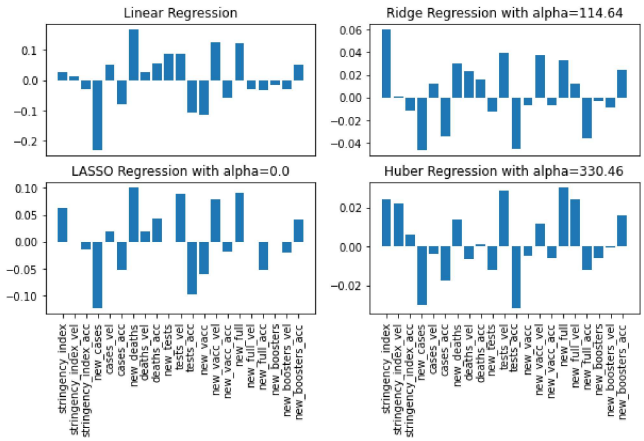
Industrials



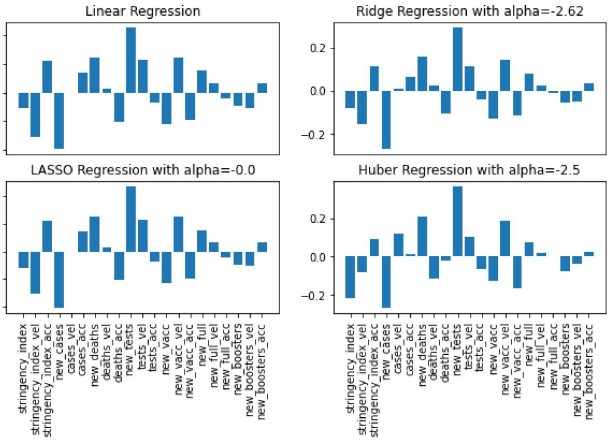
Information Technology



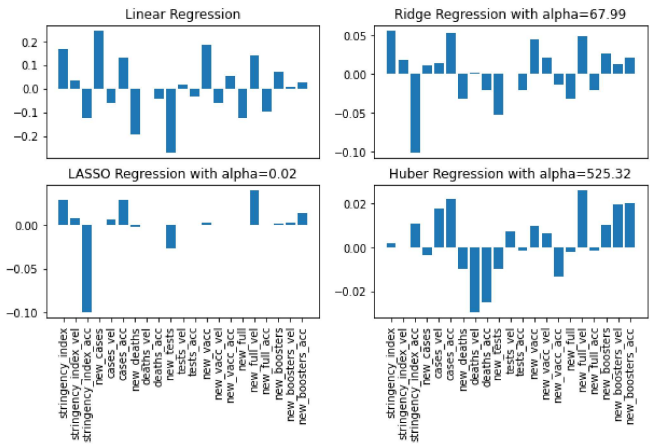
Materials



Real Estate

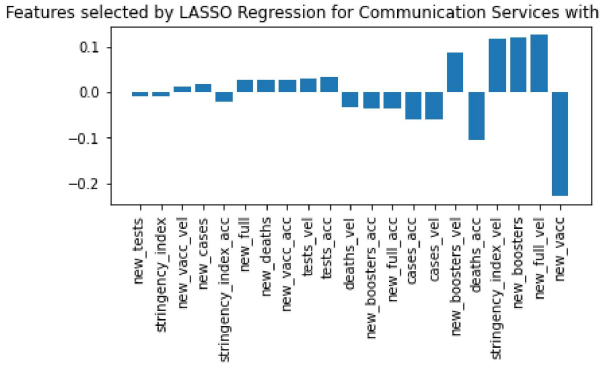


Utilities

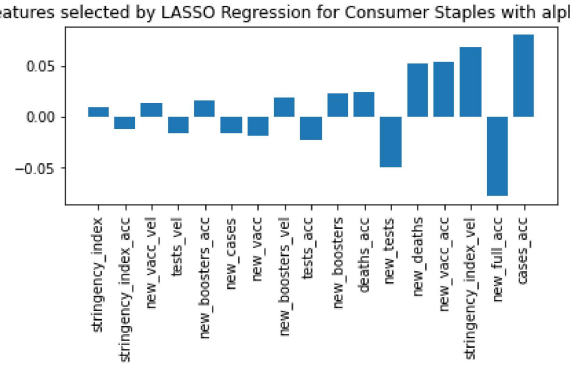


Plots of coefficients for LASSO:

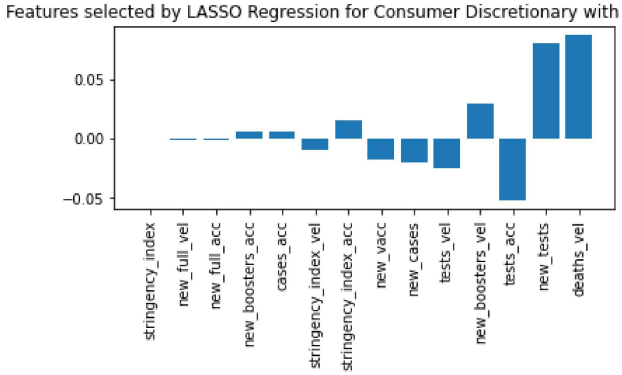
Features selected by LASSO Regression for Communication Services with alpha=-0.0



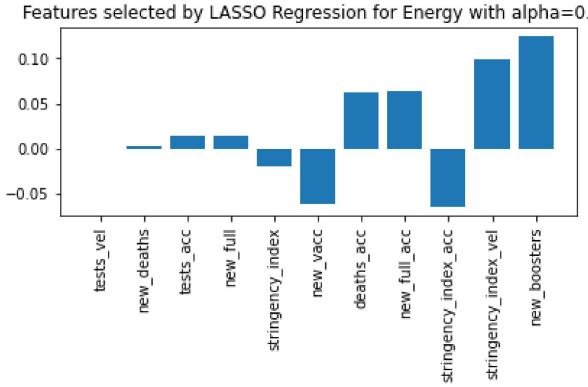
Features selected by LASSO Regression for Consumer Staples with alpha=0.01



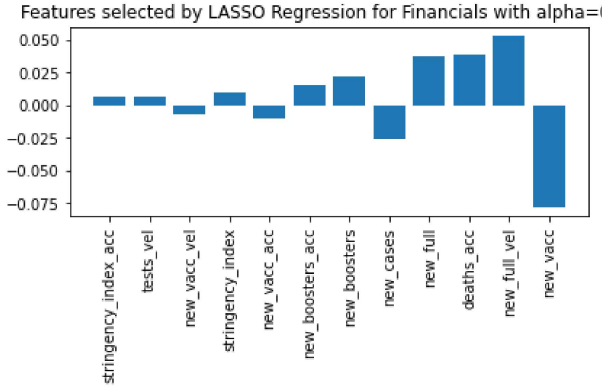
Features selected by LASSO Regression for Consumer Discretionary with alpha=0.02



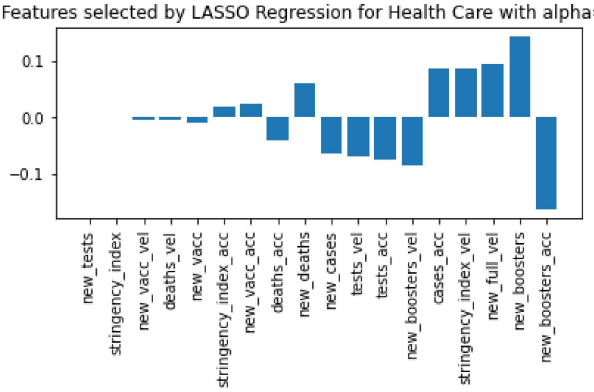
Features selected by LASSO Regression for Energy with alpha=0.01



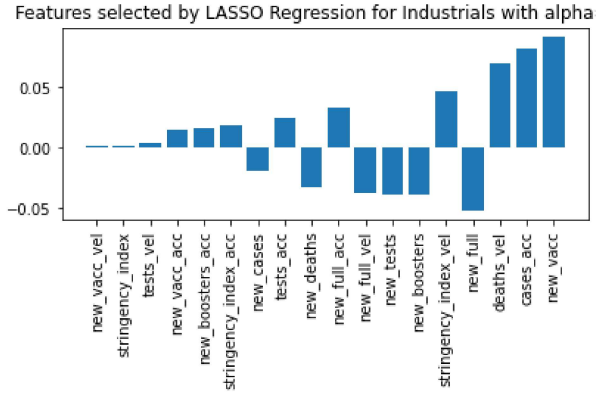
Features selected by LASSO Regression for Financials with alpha=0.01



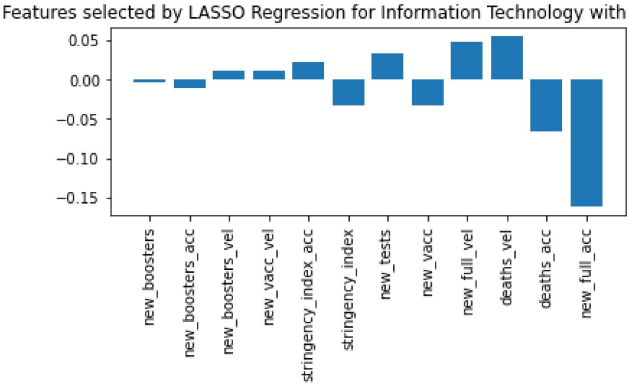
Features selected by LASSO Regression for Health Care with alpha=0.0

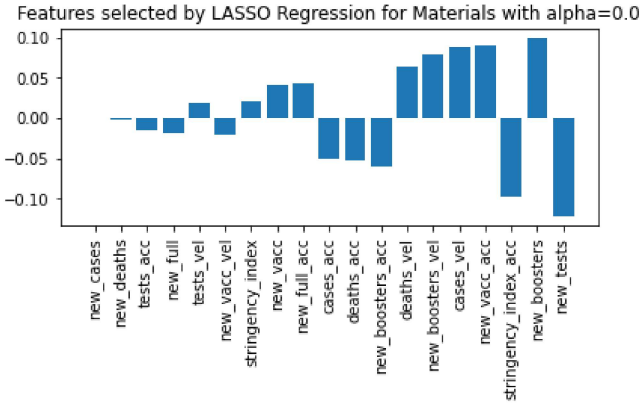


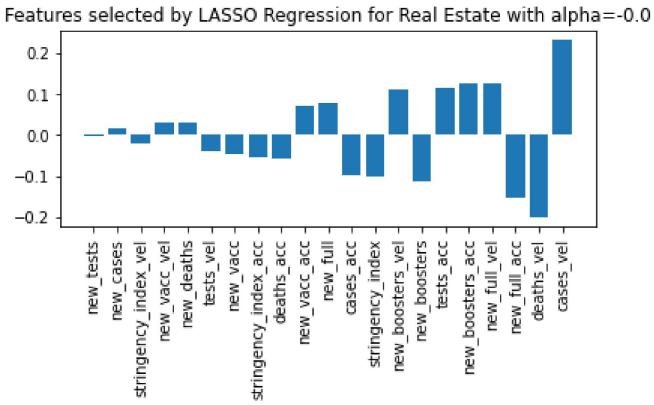
Features selected by LASSO Regression for Industrials with alpha=0.0

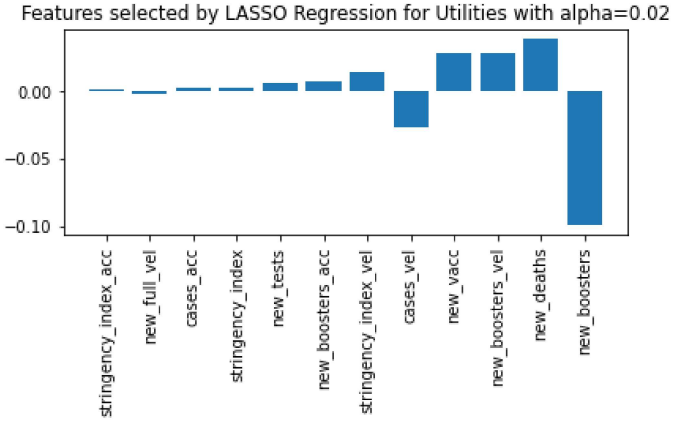


Features selected by LASSO Regression for Information Technology with alpha=0.02

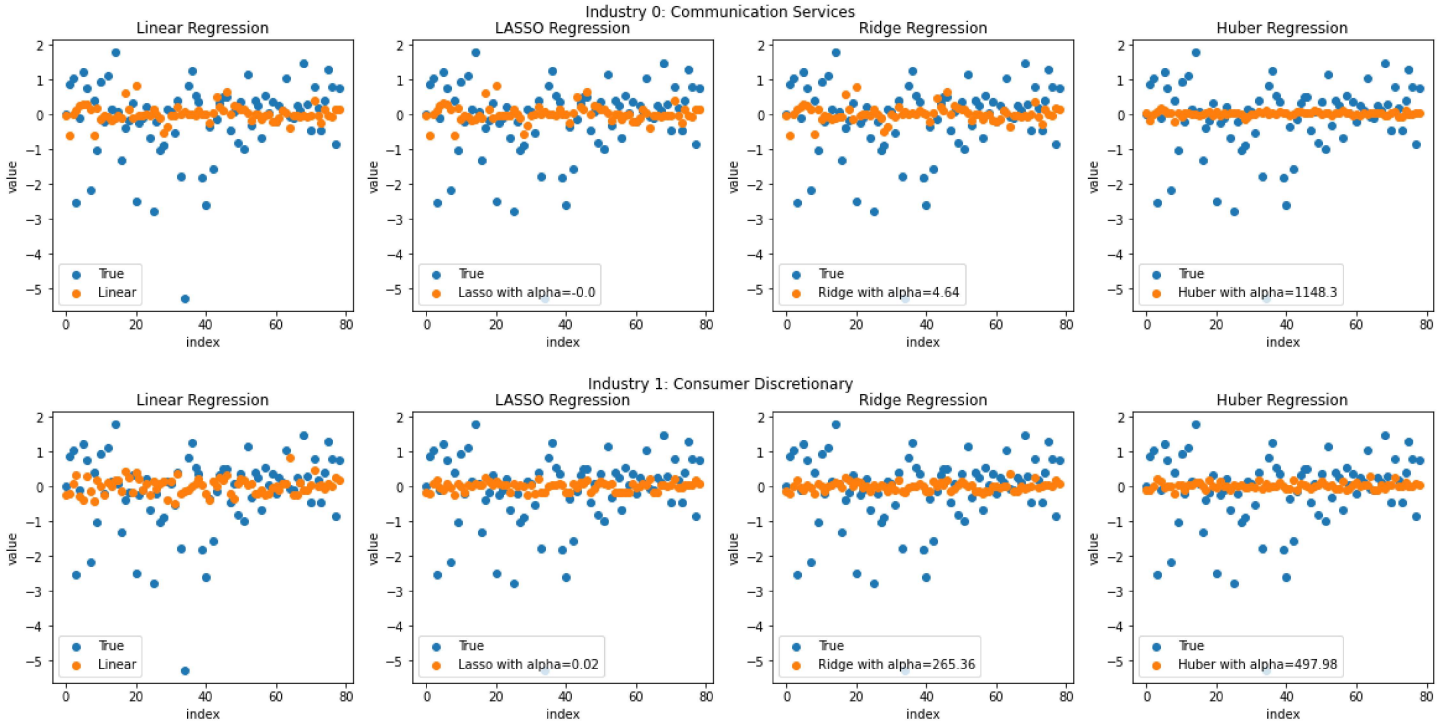


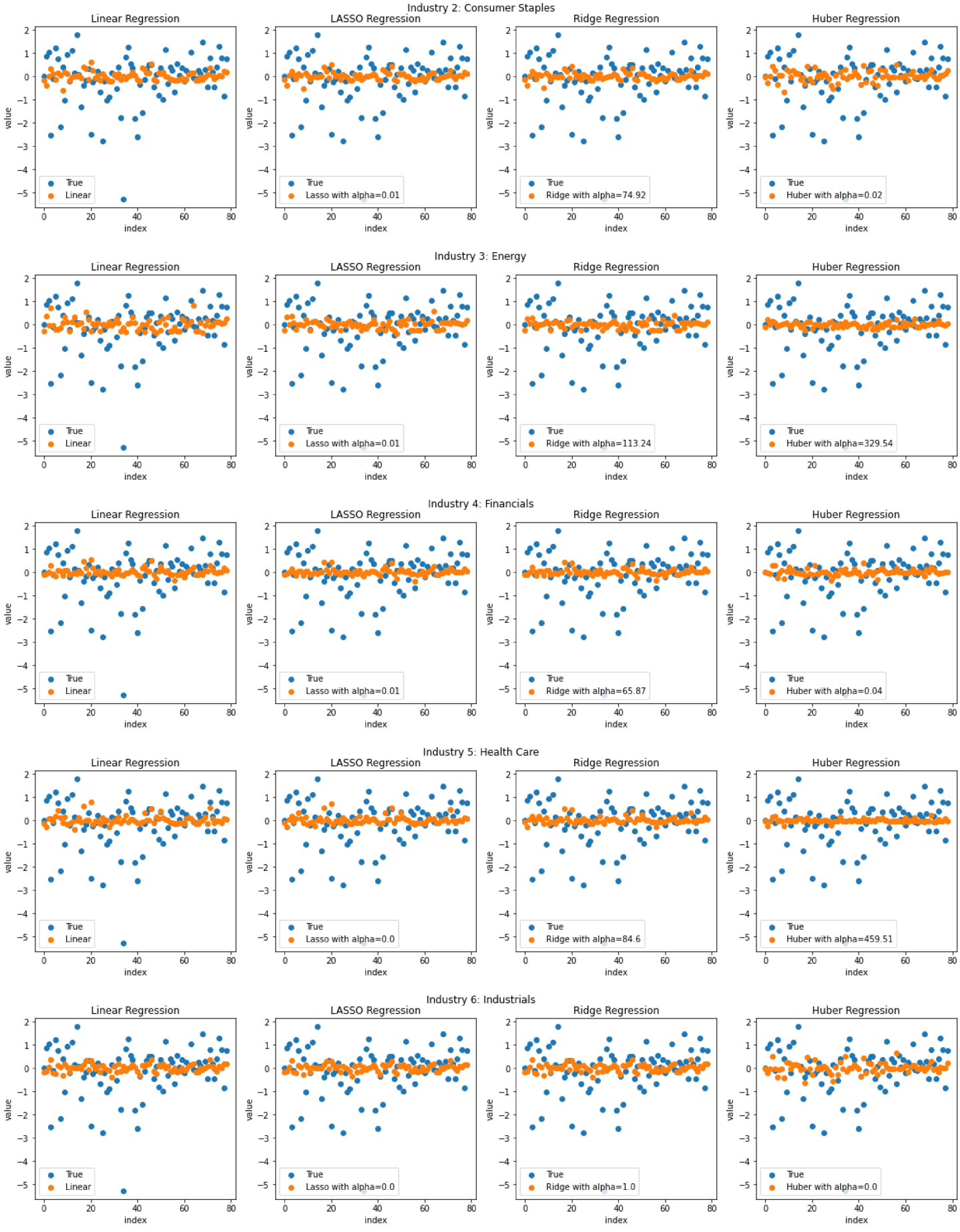




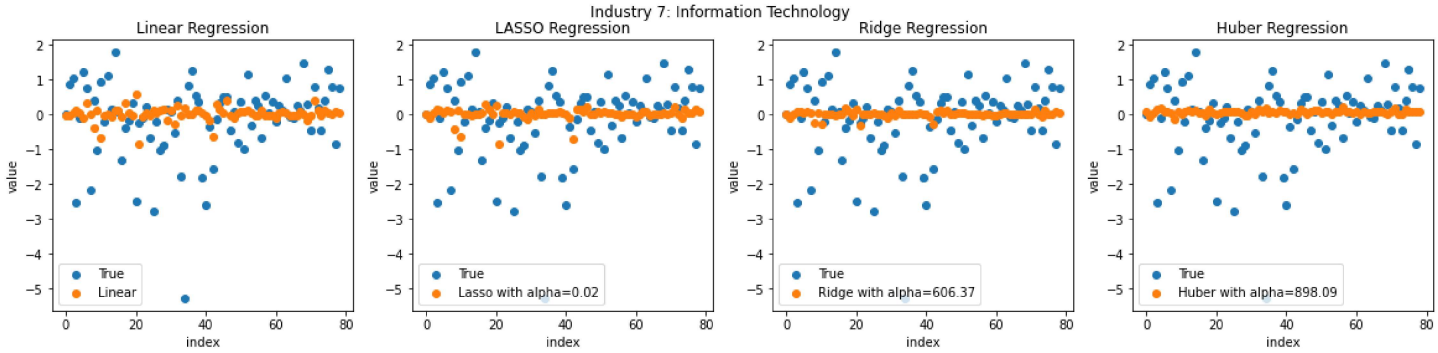


Scatter plot of pred and true

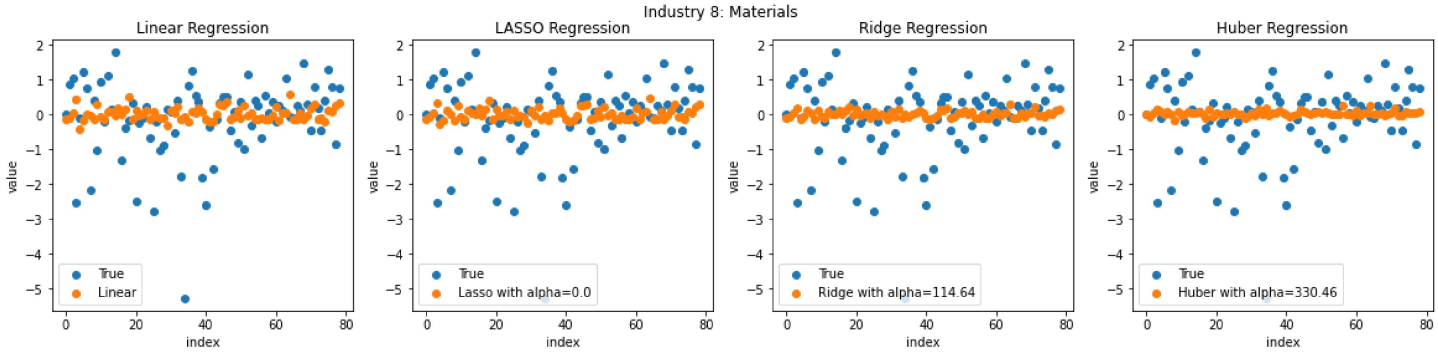



Industry 2: Consumer Staples
Linear Regression
LASSO Regression
Ridge Regression
Huber Regression
True
Linear
Lasso with alpha=0.01
Ridge with alpha=74.92
Huber with alpha=0.02
Industry 3: Energy
Linear Regression
LASSO Regression
Ridge Regression
Huber Regression
True
Linear
Lasso with alpha=0.01
Ridge with alpha=113.24
Huber with alpha=329.54
Industry 4: Financials
Linear Regression
LASSO Regression
Ridge Regression
Huber Regression
True
Linear
Lasso with alpha=0.01
Ridge with alpha=65.87
Huber with alpha=0.04
Industry 5: Health Care
Linear Regression
LASSO Regression
Ridge Regression
Huber Regression
True
Linear
Lasso with alpha=0.0
Ridge with alpha=84.6
Huber with alpha=459.51
Industry 6: Industrials
Linear Regression
LASSO Regression
Ridge Regression
Huber Regression
True
Linear
Lasso with alpha=0.0
Ridge with alpha=1.0
Huber with alpha=0.0
value
index

Industry 7: Information Technology

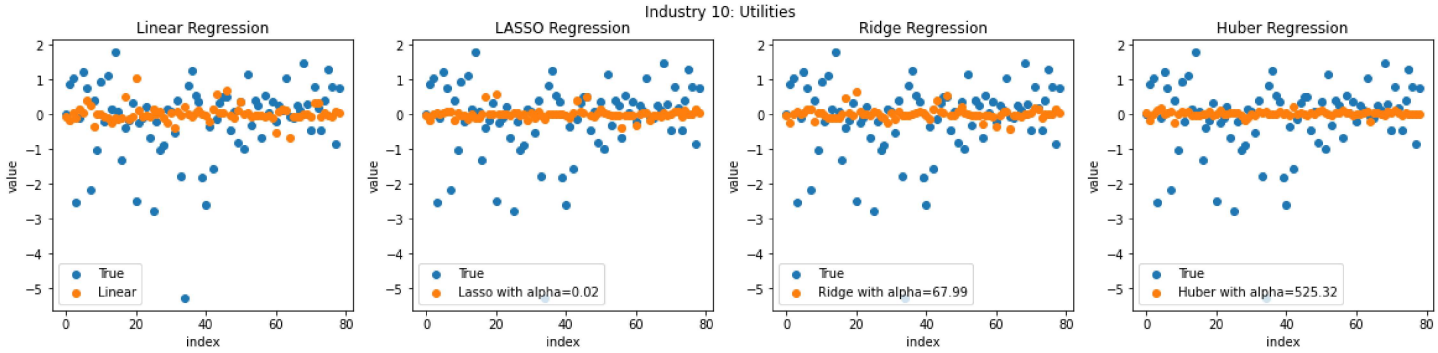


Industry 8: Materials



Industry 9: Real Estate

Industry 10: Utilities



Scatter plots of predicted vs true for random forest