Cornell University
Operations Research and
Information Engineering

Can we use Covid-19 Data to Predict Stock Market Performances?

Yujie Cao (yc2628), Jinyi Jiang (jj756), Qianqian Wu (qw273)

Abstract

In 2020, the sudden worldwide outbreak of Coronavirus has caused a tremendous impact on our economic, social, and spiritual life. Currently, we are facing enormous uncertainty, which is the last thing investors like to see. In this project, we are interested in identifying the relationships between Covid-19 data and stock indices in China and the US. The impact of COVID-19 on the economy, and on stock markets in particular remains unclear. With our project insights, investors can better allocate their investment portfolios by investigating the relationship between confirmed and death cases and stock price variations in the US and China.
Specifically, we are interested in:
How does COVID-19 affect the stock market?
In this report, we will explore the following questions with a correlation table and linear regression model with a quadratic loss function.

- What are the impacts of new confirmed cases and new deaths on the stock indices in the US and China and across sectors?
- How would the COVID-19 cases in China impact the US stock market or vice versa?

Which factors contribute to high and low stock indices, respectively?
We will explore the question using quantile regression.
With COVID-19 information, which model can predict the stock market indices more accurately?
We will perform various models such as linear regressions, support vector regression, and neural network, and compare the mean squared error (MSE) and mean absolute error (MAE) of these models and decide which model has the most predictive power.

Data Gathering

As people were not practically aware of the existence of Covid-19 until early January in 2020, we only used data since January 2020 in our model to ensure validity. Our dataset consists of three main parts and we collected them from various sources.
Coronavirus data:
We obtained the Covid related data from the World Health Organization, and the variables we chose to include in our model are the daily new identified cases and death in China and the US respectively. Each Covid related data has 343 observations.
China market indices:
As for the Chinese stock indices, we chose to use the CSI300, the Chinese consumer products index, and the Chinese healthcare index to represent the general market and specific sector performances. The indices were collected from the Wind Terminal, which is a popular Chinese financial platform that is comparable to Bloomberg. The China market indices each have 228 observations.
US market indices:
For stock indices in the US, we gathered S&P500, Dow Jones U.S. Retail Index, and Dow Jones U.S. Pharmaceuticals & Biotechnology Index from Investing.com. Additionally, market performance is subject to the unexpected, and such information is hard to be quantified and captured. Therefore, we used COMEX Gold and US 10-Year Treasury Bill as proxies for the black swans, and the data were collected from the Wind Terminal. The US market indices each have 237 observations.

Data Cleanup

We first constructed the dataset by combining the data downloaded from the abovementioned sources. When having a closer look at our data, we noticed that the scale of the US Covid data is extremely large in comparison with the Chinese data, and the same rationale applies to the market data as well. To cope with this situation and given that our project purpose is to explore the market sensitivity in response to the Covid impact, we performed a log transformation to all of our variables. In addition to making the US and China data more comparable, the benefit of having all of the variables in a log scale is that when we construct the linear regression model, the coefficients can be interpreted as the percentage impact on the dependent variable as the independent variables change by one percent.

We also believed that the stock markets would be affected by its previous day's performance. Taking this lingering effect into consideration, we constructed a new independent variable for each market index that is simply just the market index lagged by one day. In addition, as Covid data will not be released until the end of the day, we also lagged the Covid related variables by one day.

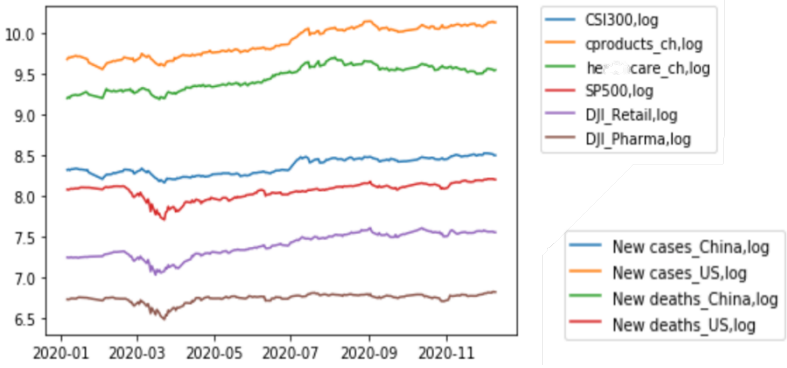
Last but not least, since market openings are subject to holidays and weekends, there are 115 missing values for each Chinese index and 106 missing values for each US index. As Covid related data, there are no missing values. Since market indices are both dependent and independent variables in our model, it does not make sense to fill up the missing values by either using the sample mean or via imputation. Hence, we decided to drop the observations with missing market values, and our dataset now consists of 207 non-null observations in total.

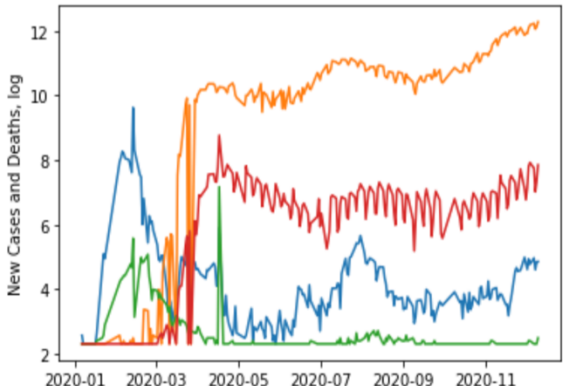
We used data up to November 05, 2020, as our training set, and the data afterward to December 10, 2020, was used as our testing set. This date was chosen arbitrarily as it marks the submission of the project midterm report. The training set has 184 observations and the testing set has 23 observations.

Data Description

- **Dependent Variables**

Country	Variable Name	Mean	Standard Deviation	Variable Type	Industry indicators
China	**CSI300**	8.3707	0.1014	Numeric, Continuous	Stock market
	cproducts_ch	9.8984	0.1783	Numeric, Continuous	Consumer products
	healthcare_ch	9.4602	0.1475	Numeric, Continuous	Healthcare
United States	**SP500**	8.0615	0.1028	Numeric, Continuous	Stock market
	DJI_Retail	7.4021	0.1488	Numeric, Continuous	Consumer products
	DJI_Pharma	6.7464	0.0589	Numeric, Continuous	Healthcare

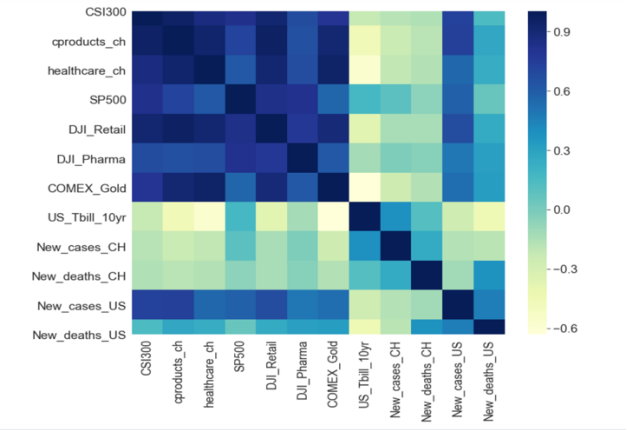




- **Independent Variables**

Measurement Type	Variable Name	Mean	Standard Deviation	Variable Type	Note
Covid-19	**lNew_cases_CH**	4.1084	1.3325	Numeric, Continuous	When assessing the Chinese stock market, the US Covid-19 variables serve as intercountry influencers, and vice versa.
	lNew_deaths_CH	2.6482	0.7318	Numeric, Continuous	
	lNew_cases_US	8.9542	3.3122	Numeric, Continuous	
	lNew_deaths_US	5.7625	1.8856	Numeric, Continuous	
Black Swan	**lCOMEX_Gold**	7.4792	0.0812	Numeric, Continuous	These variables serve as haven assets.
	lUS_Tbill_10yr	-0.1978	0.3277	Numeric, Continuous	
Historical Influencers	**lCSI300**	8.3699	0.1010	Numeric, Continuous	When assessing the Chinese stock market, the US stock indices serve as the intermarket influencers, and vice versa.
	lcproducts_ch	9.8963	0.1781	Numeric, Continuous	
	lhealthcare_ch	9.4586	0.1483	Numeric, Continuous	
	lSP500	8.0609	0.1023	Numeric, Continuous	
	lDJI_Retail	7.4007	0.1488	Numeric, Continuous	
	lDJI_Pharma	6.7459	0.0587	Numeric, Continuous	

Correlation Matrix



After feature engineering, we ran a correlation analysis on our dataset. We found that all of the dependent variables are highly correlated with each other. In addition, the number of new cases and death in the US are also strongly and positively correlated with the market indices, whereas the 10 Year US T-Bill rate is negatively correlated with the market indices.

Naïve linear regression

Based on the preliminary data exploration, the first model we were interested in running on our data was the naïve linear regression. We used the GLM module in Julia to perform the linear regression on different dependent variables and compared the significance of each feature across the regressions. The two tables below show the P-values for models on China's and the US's stock market.

Here are the observations common to both stock markets:

- Every stock index is close dependent on its yesterday price.
- Lagged new cases in each country significantly contribute to its own stock market.
- Lagged new deaths are more related to the healthcare industry.
- The US's Covid situation (new cases and new deaths) doesn't have a significant impact on China's stock market, and vice versa.

P-Values for China Stock Price Indices

China	CSI300		cproducts_ch		healthcare_ch	
Intercept	0.0078	**	0.0037	***	0.0189	*
lNew_cases_CH	0.0044	***	0.0097	**	0.0039	***
lNew_deaths_CH	0.8680		0.7680		0.1322	
lNew_cases_US	0.2265		0.2801		0.2182	
lNew_deaths_US	0.5103		0.7561		0.9094	
lCOMEX_Gold	0.9149		0.5642		0.9741	
lUS_Tbill_10yr	0.0599		0.0470	*	0.0590	
lCSI300	0.0000	***	0.0803		0.0237	*
lcproducts_ch	0.3500		0.0000	***	0.0154	*
lhealthcare_ch	0.0424	*	0.0604		0.0000	***
lSP500	0.8169		0.1405		0.3407	
lDJI_Retail	0.0021	***	0.0000	***	0.0002	***
lDJI_Pharma	0.3171		0.1489		0.4570	

*p<0.05, **p<0.01, ***p<0.005

Here are the observations unique to China's stock markets:

- The US 10-year treasury bill works better as a safe haven asset for China's market than for the US.
- The lagged Dow Jones Retail Index is highly significant to all China's market indexes.

P-Values for the US Stock Price Indices

US	SP500		DJI_Retail		DJI_Pharma	
Intercept	0.0305	*	0.2915		0.0032	***
lNew_cases_CH	0.0786		0.0752		0.1497	
lNew_deaths_CH	0.0479	*	0.0672		0.0894	
lNew_cases_US	0.0034	***	0.0847		0.0005	***
lNew_deaths_US	0.0330	*	0.1242		0.0046	***
lCOMEX_Gold	0.9940		0.9276		0.5083	
lUS_Tbill_10yr	0.7589		0.4597		0.2696	
lCSI300	0.3631		0.3077		0.4494	
lcproducts_ch	0.9700		0.5784		0.8225	
lhealthcare_ch	0.4491		0.3818		0.2784	
lSP500	0.0000	***	0.4215		0.9531	
lDJI_Retail	0.0722		0.0000	***	0.1864	
lDJI_Pharma	0.3100		0.8319		0.0000	***

*p<0.05, **p<0.01, ***p<0.005

Here are the observations unique to the US stock markets:

- Both the lagged new cases and deaths are insignificant to the Dow Jones Retail Index.
- Gold and the 10-year T-Bill are not significant influencers for the US stock market as a whole.

Linear regression with Regularizations

Due to market volatility, we assume that our dataset contains extreme outliers. Additionally, as our sample size is not very large, we want to penalize outliers by a lot via using a quadratic loss function. Our code is mainly based on class demos and homework. We also combined the quadratic loss function with different regularizations to compare across models. We provided tables of the mean squared error and the mean absolute error for each combination.

MSEs and MAEs for different regularization methods (train dataset)

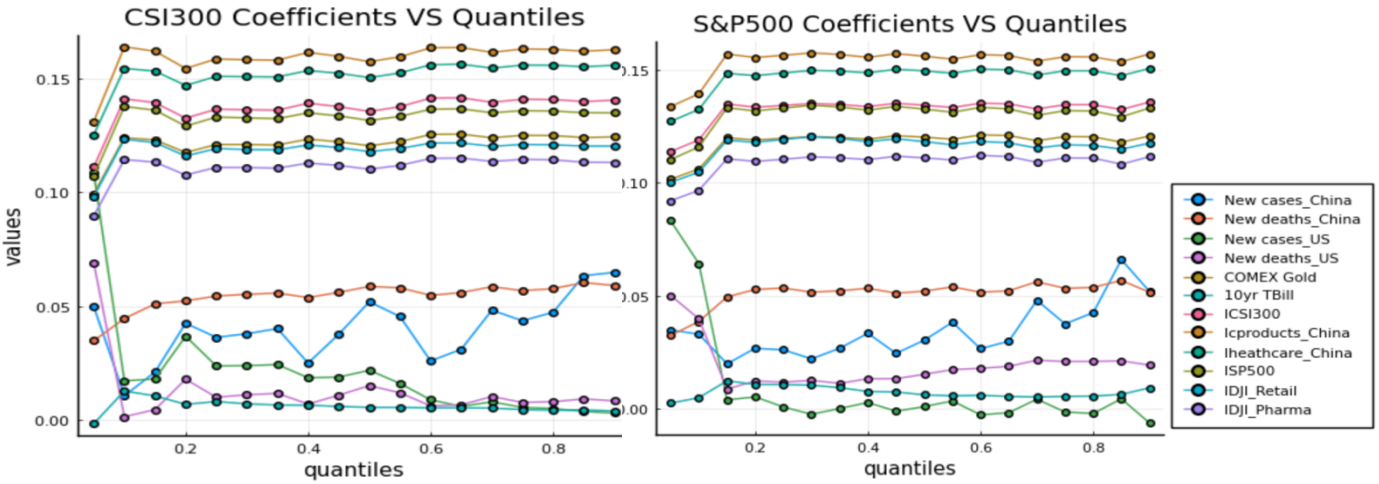
	No Regularization		L1 Regularization		L2 Regularization	
	MSE	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
CSI300	0.0011	0.0244	0.0056	0.0654	0.0068	0.0532
cproducts_ch	0.0019	0.0334	0.0042	0.0589	0.0094	0.0655
healthcare_ch	0.0015	0.0275	0.0046	0.0604	0.0076	0.0567
SP500	0.0020	0.0330	0.0083	0.0654	0.0072	0.0625
DJI_Retail	0.0027	0.0388	0.0062	0.0702	0.0064	0.0616
DJI_Pharma	0.0022	0.0421	0.0076	0.0699	0.0063	0.0647

MSEs and MAEs for different regularization methods (test dataset)

	No Regularization		L1 Regularization		L2 Regularization	
	MSE	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
CSI300	0.0008	0.0250	0.0066	0.0795	0.0015	0.0306
cproducts_ch	0.0006	0.0216	0.0074	0.0854	0.0021	0.0380
healthcare_ch	0.0129	0.0110	0.0023	0.0442	0.0273	0.1578
SP500	0.0025	0.0482	0.0069	0.0818	0.0007	0.0208
DJI_Retail	0.0004	0.0175	0.0058	0.0747	0.0016	0.0323
DJI_Pharma	0.0012	0.0322	0.0003	0.0142	0.0063	0.0774

Quantile Regressions

We were also interested to see how the variables would contribute to the indices differently when the market is performing well and poorly. As can be seen on the following plots, except for the new cases in China and the US, all other independent variables' contributions to the stock indices are relatively consistent regardless of the stock performance. For both the US S&P500 and China CSI300, the number of new cases in China contributes more to the market indices when the market is performing well, whereas new cases number in the US contributes more when both markets are underperforming.

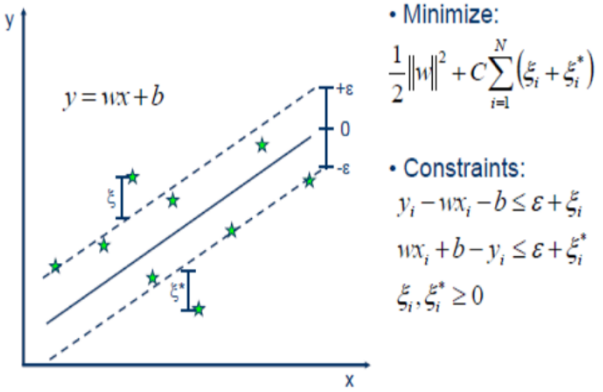


Support vector regression

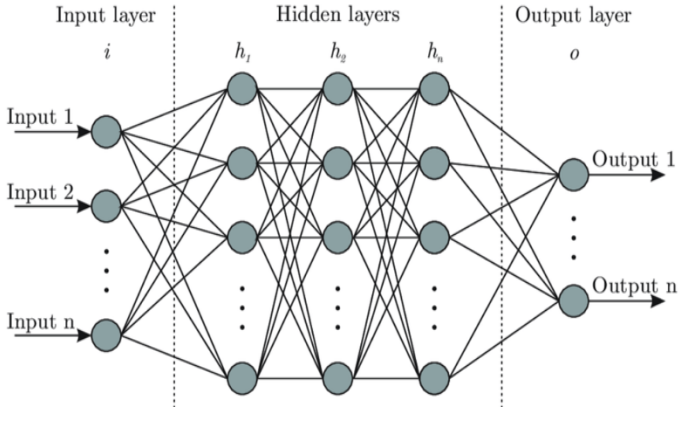
Support vector regression (SVR) introduces a non-linear part to the model. SVR creates an interval band on both sides of the function. No loss will be calculated for all samples that fall into the interval band, and the optimized model is obtained by minimizing the total loss.

Therefore, we used SVR for the six dependent variables and built six SVR models respectively. Among the various kernels[1], we decided to choose the linear kernel. We tried in python that if we use RBF or Polynomial kernels for our training data and testing data, overfitting problem will occur. All of the independent variables and dependent variables have been standardized.

The MSEs and MAEs for testing data are small and the R-squared scores are large, indicating that when we add a non-linear part, the models fit data well. The result makes sense because we cannot use the simplest naïve linear model to explain and predict the complex movement in stock indices.



SVR structure (source: Data Mining Map)[2]

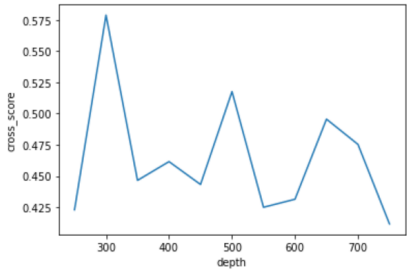


Neural network structure (source: KDnuggets)[3]

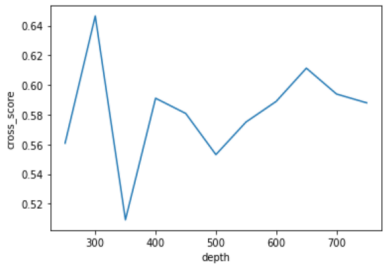
[1] https://scikit-learn.org/stable/auto_examples/svm/plot_svm_regression.html
[2] https://www.saedsayad.com/support_vector_machine_reg.html
[3] https://www.kdnuggets.com/2019/11/designing-neural-networks.html

Neural Network

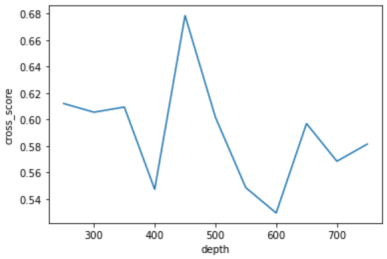
A Neural network works like a human brain in the sense that it has an input layer to input the data, hidden layers for improving prediction using backpropagation, and an output layer that produces the outputs. In our model, we use two hidden layers. Our rationale is that two or even fewer layers are sufficient for simple datasets, and problems that need more than two hidden layers are rare before deep learning.[4] When choosing the number of hidden neurons, we used cross-validation to find a good fit. Additionally, we used the same hidden neurons for all hidden layers.[5]



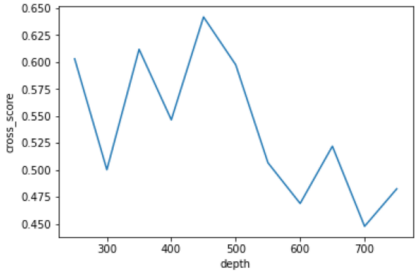
Cross-validation of US SP500



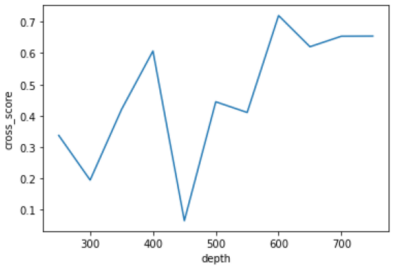
Cross-validation of US Products



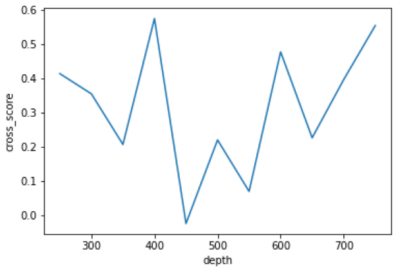
Cross-validation of US Healthcare



Cross-validation of China CSI300



Cross-validation of China Products



Cross-validation of China Healthcare

As we have standardized all of the variables for Neural Network, the MSEs and MAEs for testing data are small, but larger than those of training data. It indicates that our neural network models may have an overfitting problem. Due to the complex structure, our neural network models may not predict the stock indices very well.

Model Selection

We experimented with four different models (Support vector regression, Neural network, Linear regression with quadratic loss function and l1 regularization, Linear regression with quadratic loss and l2 regularization).

Advantages[6]

Linear regression: The interpretation of a linear regression model is very intuitive, and various regularization methods can be applied to avoid overfitting.

Neural network: The classification accuracy is high, the learning ability is extremely strong, and it can approximate any nonlinear relationship.

SVR: SVR is easy to obtain a nonlinear relationship between data and features when working with small to medium size samples. It can avoid problems associated with neural network structure selection and local minimum problems, and it is also easy to interpret.

Disadvantages[7]

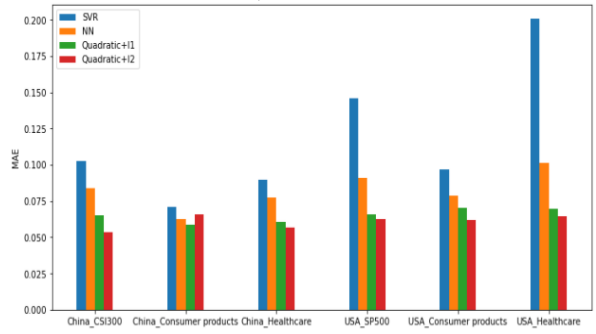
Linear regression: Linear regression is not the best at dealing with non-linear relationships, and it is not flexible enough to recognize complex patterns.

Neural network: The learning process cannot be observed, and the output results are difficult to interpret. Therefore, results generated by neural network models would have potential credibility issues.

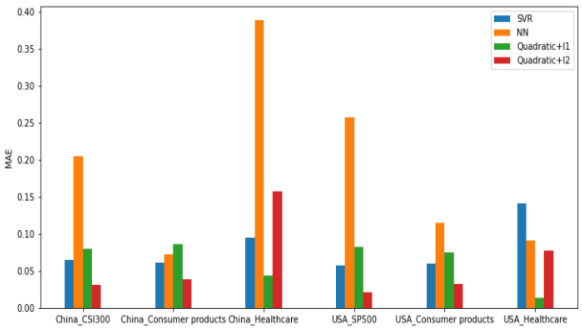
[4] https://www.heatonresearch.com/2017/06/01/hidden-layers.html
[5] https://www.kdnuggets.com/2019/11/designing-neural-networks.html
[6] https://www.cnblogs.com/tianqizhi/p/9714634.html
[7] https://blog.csdn.net/qq_27825451/article/details/84132680

SVR: It is relatively difficult to find a suitable kernel function for SVR, and the result is highly sensitive to the parameters chosen.

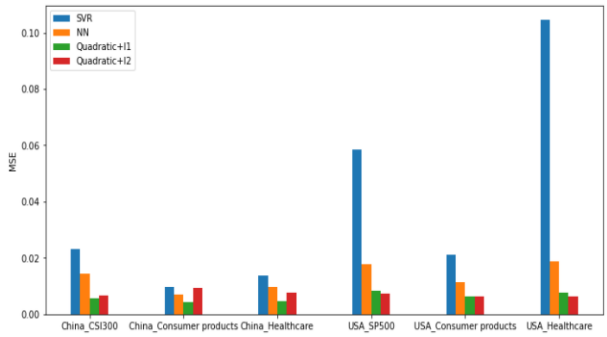
For each model, we calculated the MSE and MAE for both the training and the testing data.



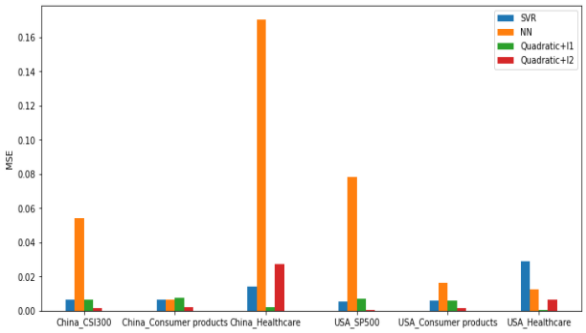
MAE for training data



MAE for testing data



MSE for training data



MSE for testing data

As shown by the graphs above, we could see that the SVR model with linear kernel predicts well for the stock index in China. As for the stock indices in the US, however, the SVR has an underfitting problem in comparison with other models.

The neural network model has an overfitting problem for the CSI300 index, China healthcare stock index, and S&P 500 index. As for other stock indices, the performance is lacking compared with other models. The fact that the Neural network structure is too complex for our data could be a potential cause.

As for linear regression with quadratic loss function and l1 regularization, the MSEs and MAEs are small for training data and testing data, indicating that the model performs fairly well on our dataset. The model predicts the China and US healthcare stock index best among these four models, as it has the smallest MSE and MAE in testing data.

Last but not least, the linear regression model with quadratic loss function and l2 regularization yields small MSE and MAE for both training and testing data. It produced the smallest MSE and MAE for the China and US overall stock index and consumer product stock index with the testing data.

In summary, we would use the linear regression with quadratic loss function and l1/l2 regularization to predict China CSI300 index, China consumer product stock index, the US S&P500 index, and the US consumer product stock index.

Future Improvement

Although we have done a relatively comprehensive analysis, we believe that there are still areas for future improvements that would complement our work.

- **Test our models with other countries' stock and Covid-19 data.**

  In our report, we only discussed the US and China. We have shown that Covid-19 indeed has an impact on stock performance in these two countries, therefore we suspect that it could potentially be an international phenomenon. Hence, it would be interesting to test our methods and models with data from other countries and markets.

- **Explore models such as deep learning and compare them with previous models.**

  We only compared four models in our report to predict the future movement of the stock indices. As more models are being developed in the field of Machine Learning, we could test more models on our dataset.

- **Analyze news reports and use NLP (Natural Language Processing) to simulate black swans.**
  Most black swan events are disclosed in news reports. Therefore, we could collect relevant news by using data mining and utilize NLP to convert the news articles into usable features. With this addition, we could better capture the effect and the extent of the black swan incidents.
- **Add a vaccine feature to the model when Covid vaccines are officially launched in the market**
  As Covid vaccines have entered Phase 3 Efficacy Trials in both China and the US, the official launch date seems to be approaching. Once the vaccines are launched, investors' confidence towards the market might be boosted, thus leading to an increase in the stock index. Therefore, we hypothesize that data regarding vaccine usage will have a strong impact on stock market performances.

Discussion

Weapon of Math Destruction

This model is not a Weapon of Math Destruction. Because we won't provide a negative feedback loop in the market. We assume that everyone believes in this model and will act rationally. Therefore, investors would purchase stocks whenever the model predicts an increase in the indices, driving the indices to grow even more. While the government tries to stimulate the economy during the public health emergency, our model helps to strengthen the market confidence, thus subsiding the recovery of the slowed-down economy.

Fairness

Our models do not have fairness issues. This is because our variables do not contain Booleans, thus it would not have the fairness harm of false positives and false negatives. Moreover, there are no potential discrimination problems, which shows that there is no disparate treatment or disparate impact in our algorithm.

Conclusion

- From the linear regression, we learned that every stock index is closed dependent on its previous day performance. Additionally, lagged new cases in each country significantly contribute to its own stock market. Finally, the US Covid situation does not have a significant impact on China's stock market, and vice versa.
- With quantile regression, we have found that new Covid cases in China would contribute more to the market indices when the market is performing relatively well, whereas new Covid cases in the US are more impactful when the market is underperforming.
- For the model to predict stock index, we choose linear regression with quadratic loss function and l2 regularization for China CSI300 index, China consumer product stock index, the US S&P500 index and the US consumer product stock index, and linear regression with quadratic loss function and l1 regularization for China and the US healthcare stock index.