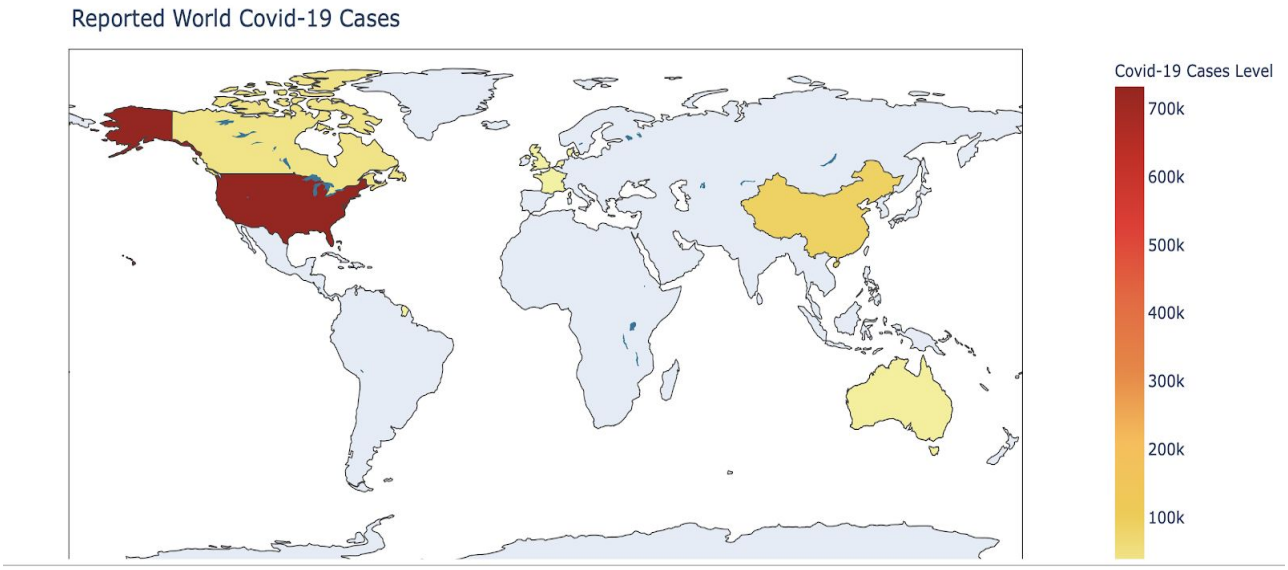
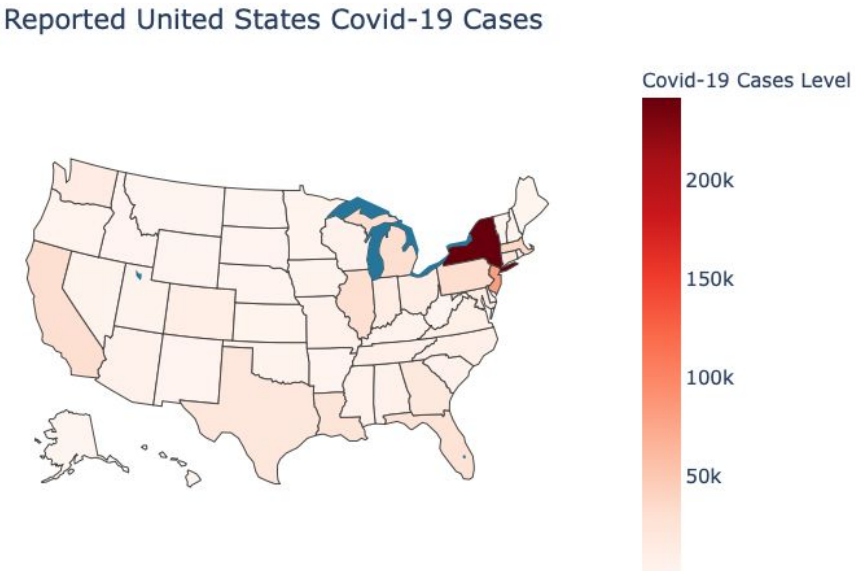
United Stated COVID-19 Data Analysis

BY: Bolun Du, Tianyu Yao, Qiaoan Yang

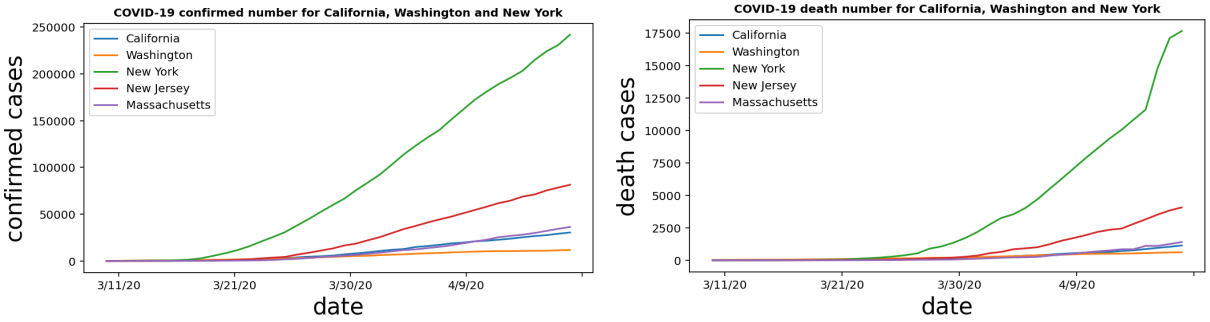


Introduction:

Coronavirus (Covid-19) has become the most buzzed topic these days. Its outbreak has taken the world by storm. The data we have for 2020-4-18 only contains 8 countries: Australia, Canada, China, Denmark, France, Netherlands, US, United Kingdom. There are 859,536 people in those eight countries confirmed to have the COVID-19 and 44,835 people dead. The above figure shows that the United States is the country with the most cases of COVID-19. In this project, we only focus on the cases in the United States.



This is the Choropleth map of the United States about the COVID-19. New York has over 200k cases on 4.18 became the most severe area. All the other states are about 50k cases.



The figures above show the confirmed and death cases in five states: California, Washington, New York, New Jersey and Massachusetts from 3.1 to 4.18. The left figure shows confirmed cases in those five states, and we could see that the green line which represents New York is extremely higher than other states. The figure on the right shows the deaths cases, and we could see that confirmed cases are proportional to the death cases. New York has the most death cases as well. When we did our model, we also found out that New York has a big effect on the linear model, and other models as well.

Question Frame

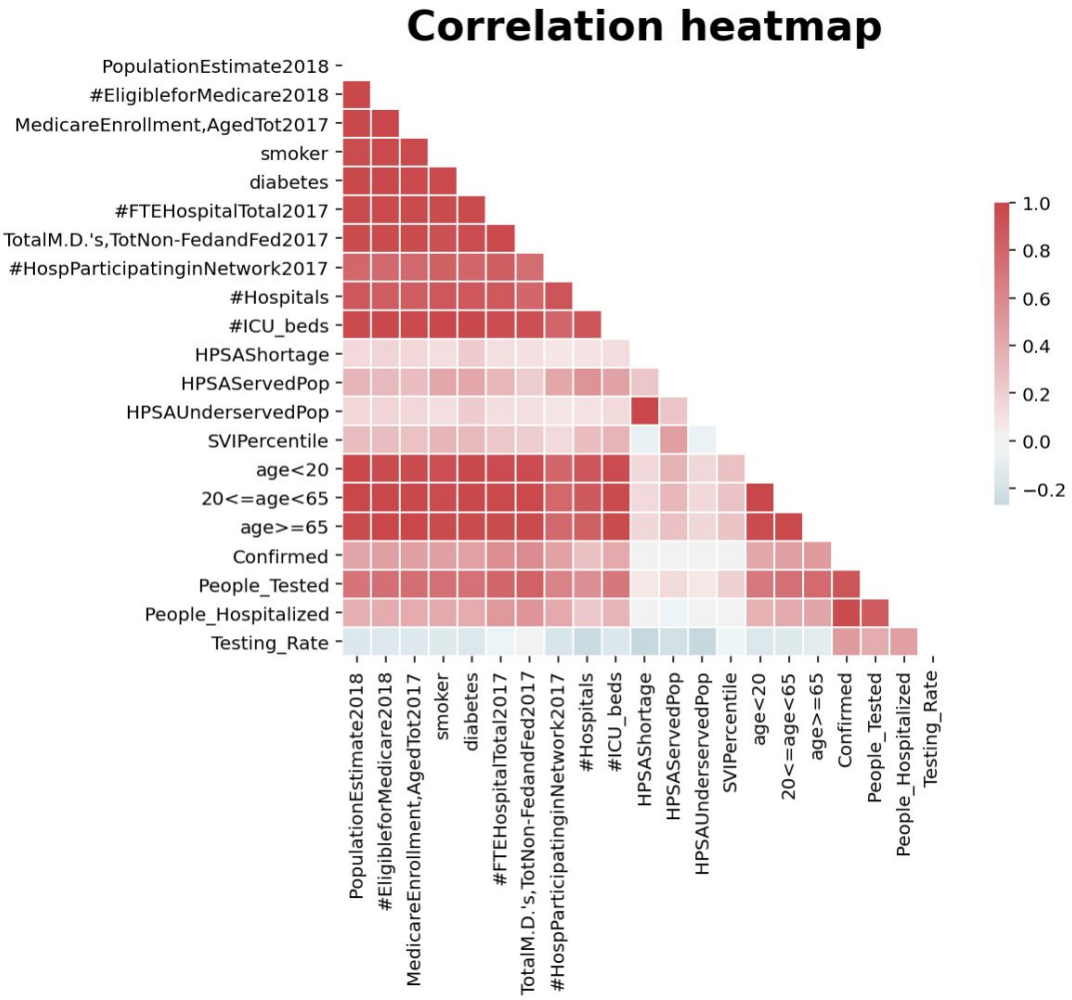
From the COVID-19 datasets, we have confirmed cases data and death cases data from 1.22 to 4.18, detailed information of 4.18 and counties information. We could see that the resources and population in different counties are relatively large. There are several counties even without hospitals. It might affect the number of confirmed COVID-19. The mortality of some other diseases, such as diabetes, smoke, and heart disease also has large gaps, since people with some other diseases will have lower immunity, so the possibility of infection will be greater, which might also affect the number of confirmed COVID-19. And also, from the 4.18 data, we could find there is much useful information such as testing rate, hospitalization rate, and the number of people tested, etc. Therefore, we are curious about what kind of features would affect the number of people confirmed in different states.

In this project, we use some different models and try to find out which features have more impact on affecting the number of people confirmed COVID-19.

EDA (data cleaning, analysis):

Data cleaning:

1. First of all, all the missing values are treated as 0, the columns that have missing string values are dropped since there is an index column that sufficiently contains info we need.
2. For the two time-series data, we calculate the confirmed and death number for each day in different states and remove other property info such as location and identification code.
3. For 4.18 states data, we select all the features we need and set all missing values as 0.
4. For abridged_couties, we calculate the number of people in different age ranges(0-20, 20-65, 65+), the number of smokers and diabetes for each state. We removed columns that only for identification except for state FIPS code which will be used to join tables and we removed data from two Princesses Cruises and inhabited territories of the US since they don't have enough information.
5. We inner join abridged_couties and 4.18 states data together by FIPS.
6. Data were normalized when training the model.



Then we inspect the correlation of each variable in our dataset. We plot a correlation heatmap and visualize the correlation between each feature.

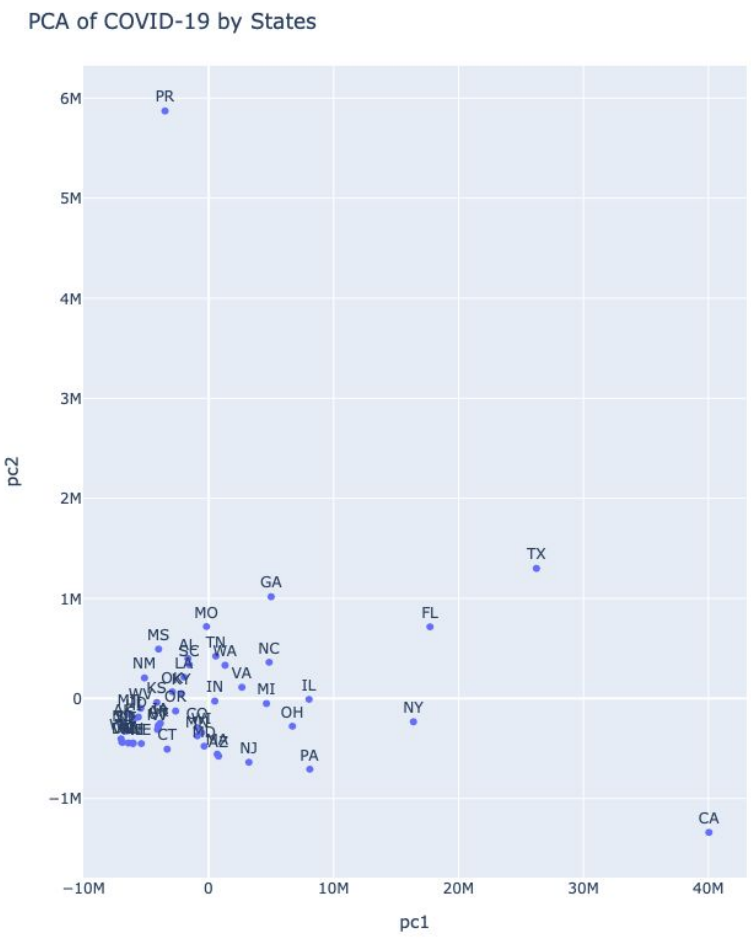
Correlation map shows what features are highly correlated with confirmed cases. We found out that the last four features are either nearly 0 correlated or negatively correlated with confirmed cases, which are 'SVIPercentile','HPSAShortage','HPSAUnderservedPop', and 'HPSAServedPop'. Therefore, we dropped those four features before we started modeling.

```
# the correlation of confirmed number with all features
corr["Confirmed"].sort_values(ascending=False)
```

```
Confirmed                              1.000000
People_Hospitalized                    0.973668
People_Tested                          0.905835
TotalM.D.'s,TotNon-FedandFed2017       0.581842
#FTEHospitalTotal2017                  0.562816
age>=65                                0.492052
```

We also notice among those features, they do have collinearity problems, thus we preserve all of the related features and perform backward selection in our final model

Model:

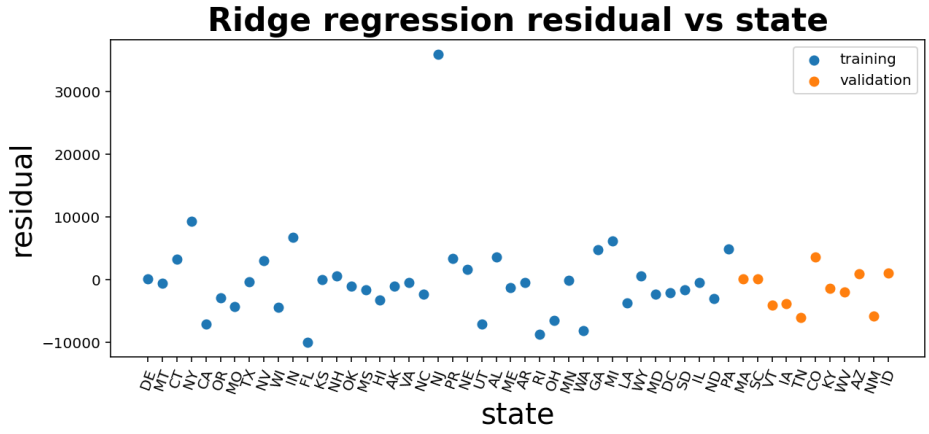


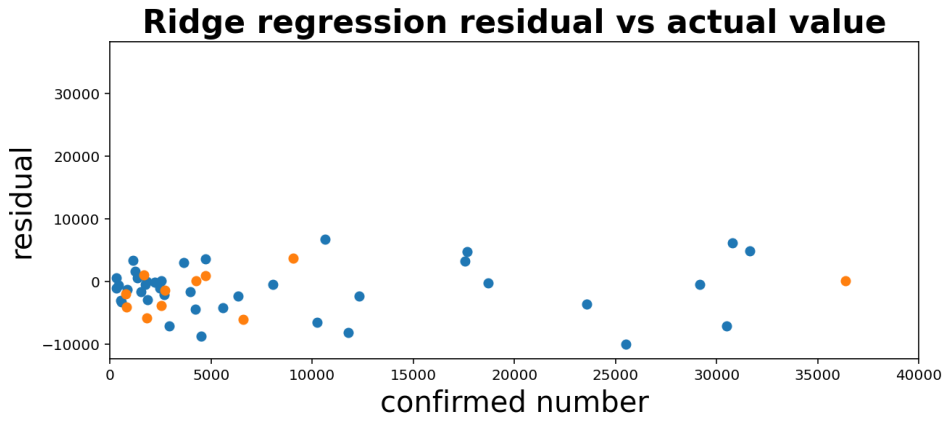
PCA

We want to investigate what variables are related to confirmed cases. Our first approach is using PCA to model confirmed cases given state level statistics. We take all columns in training data, center each column, form a design matrix and calculate the first 2 principal columns. The plot shows states' data in transformed coordinate . From the plot we cannot easily infer what each axis is represented for. We realize PCA is not a good model to figure out confirmed cases by state, since it is hard to interpret principal components.

Ridge regression with full model

Since our samples are relatively small and the heatmap suggests multicollinearity exists in our data, we perform a ridge regression to figure out what variables are related to confirmed cases. We train a ridge model by cross validation, calculate mean square error on both training and test sets. This ridge regression takes all our features in. Model has training residual mean square error (rmse) 7074.7 and we run prediction on validation set, scoring rmse 3343.5



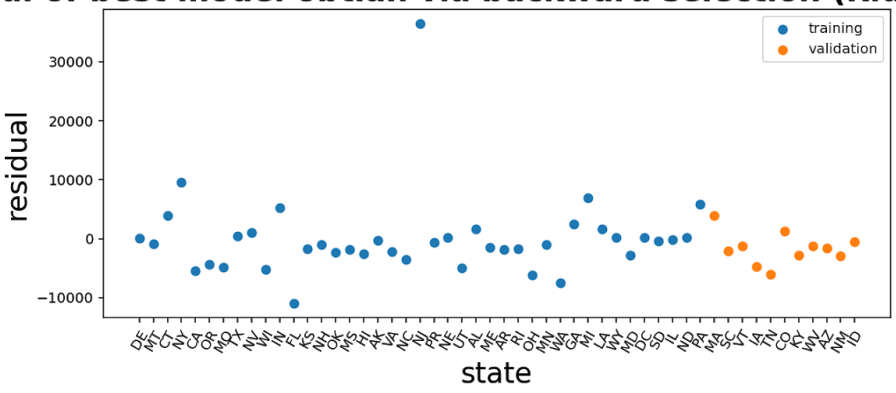


The above plot is a residual versus confirmed number. This graph did not include two states with confirmed infection over 80000 since they are too big to plot here.

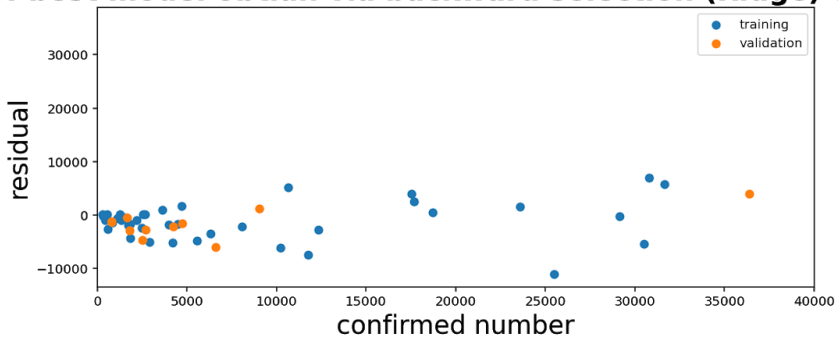
Ridge regression with backward selection

Our full model is still not an ideal model since we are already aware that some features are not necessary. Thus we run a backward selection on the full model, kicking out redundant features by measuring mean-square error each time. The result suggest that we should drop these 4 features: testing rate','Testing_Rate', "TotalM.D.'s,TotNon-FedandFed2017", '#FTEHospitalTotal2017', 'age>=65'. Our model training rmse and validation rmse are decreased. The residual plots below approximately align to a line around zero. Compare with previous model, we could observe the residuals slightly decrease. Although we still have outliers, backward selection yields a better model. Model has training residual mean square error (rmse) 6902.33 and we run prediction on validation set, scoring rmse 3047.36

Residual of best model obtian via backward selection (Ridge) vs state



Residual of best model obtian via backward selection (Ridge) vs actual value



```
Ridgemodel.coef_
```

```
array([-1.08713449e-04, -8.23805248e-04, -9.27395546e-04,  9.02898446e-04,
        1.55331059e-03,  1.30533379e+02, -6.15002237e+01, -9.42627112e-01,
        1.26708545e-04,  1.84646342e-04,  1.55651882e-01,  2.58804072e+00])
```

```
x_train_backward.columns
```

```
Index(['PopulationEstimate2018', '#EligibleforMedicare2018',
       'MedicareEnrollment,AgedTot2017', 'smoker', 'diabetes',
       '#HospParticipatinginNetwork2017', '#Hospitals', '#ICU_beds', 'age<20',
       '20<=age<65', 'People_Tested', 'People_Hospitalized'],
      dtype='object')
```

The plot shows the ridge model we finalized contains 12 features and corresponding non-zero coefficients. The alpha is 0.1 and intercept 301.06.

We found that the coefficient of the number of hospitals participating in Network (HospParticipatinginNetwork2017) has the largest absolute value, which implies an increase in one unit of hospital join the network corresponding to an increase of 130 confirmed cases. This model shows that the normalized unit number of hospitals participating in networks will have a great impact on the number of confirmed cases of that state.

Analysis:

We find the number of smokers and the number of diabetes are interesting features. These two features have a correlation close to 1, this might reflect people's health awareness of a certain region.

One particular feature that draws our attention is the number of full time employees at Hospitals (#FTEHospitalTotal2017). It would be useful since employees would affect the detection capabilities. We think that if there are more full time employees, then there might be more detection each day. If there are more detections, then it might lead to more confirmed cases. So the number of full time employees would affect more confirmed cases. But this feature has been dropped by backward selection, which means that confirmed cases are not relied on this variable.

We find data processing and model selection are two challenging parts of this project. We spend a lot of time understanding and selecting columns in different

datasets. We collect county level data, group them to state level and then join them with another dataset to obtain the number of confirmed cases by state. The groupby part is tricky since there are lots of missing or invalid entries in county name columns and we are also responsible to apply different aggregate functions to county level variables. The second hardest part is designing a proper model to figure out what role each variable played in determining confirmed cases. After group counties data to states, we find we have relatively small data sets and therefore need to handle overfitting issues. We have to explore multiple models to find which one is the most decent.

At the first time, when we see the topic we want to do a model for predicting the future. But the data that proved has pretty limited information, then it is hard to do a prediction model based on much fixed information.

We face a dilemma about whether to drop an outlier without explicit evidence why it becomes one. When we build our model, we find New Jersey is an outlier state (has high residual), and our model will have lower MSE if we drop it. We are tempted to drop it. However, we do not find a specific reason why NJ is an outlier state and hence we keep it in our training data even though our model will perform better if we throw it away.

We want population density and economic data of each state, because we suspect the number of confirmed cases also depends on social distance and local economy. The population density could serve as an indicator of how sparse people are living. Low population density states might slow down the speed of infection, and therefore has a low number of confirmed cases. Economic data might also help explain confirmed cases since a state with a strong economy will invest more on the local health system and attract rich people, who are willing to spend more on personal protection in a pandemic. This might solve the outlier problem in our data.

We need more personal health data but this part is hard to access since it is about privacy. There is always a trade off between analyzing data to serve people better and giving away personal data. We believe that data are important for both individuals and data scientists. Scientists and researchers need more data to push the boundary of fields but do need to obtain data with the consent of individuals.

Conclusion:

We first try PCA but find it is overfitted and hard to interpret what our principal components are after plotting data in 2 principal components' coordinates. We are bothered by this problem a while and eventually decide to build a ridge model to prevent overfit, and run backward selection to drop features that are useless since we are aware that some of our features are uncorrelated with confirmed cases. Our model shows what features affect the number of confirmed cases.

The limitation of our model is it cannot predict what will happen with time. Because we collect only 4.18's data and design a model for detecting which features are associated with confirmed cases, it is hard to use current data to predict what will happen in future.

Surprisingly, the coefficient of our final model indicates that numbers of hospitals in the network and the number of hospitals have the most impact on the confirmed number in positive and negative direction. We didn't figure out the reason why those two features are kind of similar, but impact the confirmed cases in different directions. One potential explanation might be people are more willing to be tested if there are more hospitals in the network such that their treatment fee can be reduced if they were tested positive unfortunately.

In conclusion, our model selected features that made the most impact to the number of confirmation and most of those features can be interpreted reasonably. The variables selected by our model can give people a general idea when a pandemic comes, what factors will lead to a higher infection number. The governments and individuals could be aware of their situations and then adjust their strategies to better protect people around them.