Explore Covid-19 Data Set

Author: Yanling Liu

Introduction

In this final project, we will be using some Covid-19 data sets and linear regression to explore two main questions:

1. What happens if we try to use known features to predict the number of confirmed cases of each *state* on a given day? Is it a conductible way for the future to predict the number of confirmed cases of a state on a given day given the available features?
2. What are the best features to use in order to predict the number of confirmed cases of each *county* on a given day? What different combinations of features are there to give an ideal result?

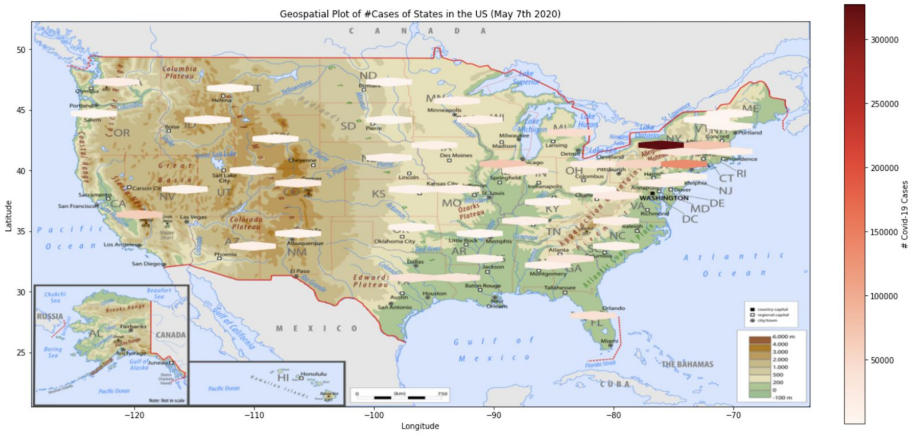
We want to build linear regression models that could help predict the daily number of confirmed covid-19 cases either for a county or a state. In order to make the process easier, we built sklearn pipelines that help select desired columns, normalize the data, and fit selected features to models.

There are 2 main parts of the project. The first part focuses on predicting daily numbers of confirmed covid-19 cases for each state, and the second part focuses on predicting daily numbers of confirmed covid-19 cases for each county. After experiments of different combinations of features and models, we found that predicting for each state may be difficult given the information we have, and predicting for each county is more efficient and accurate with county features that we obtain.

Visualisations and Findings

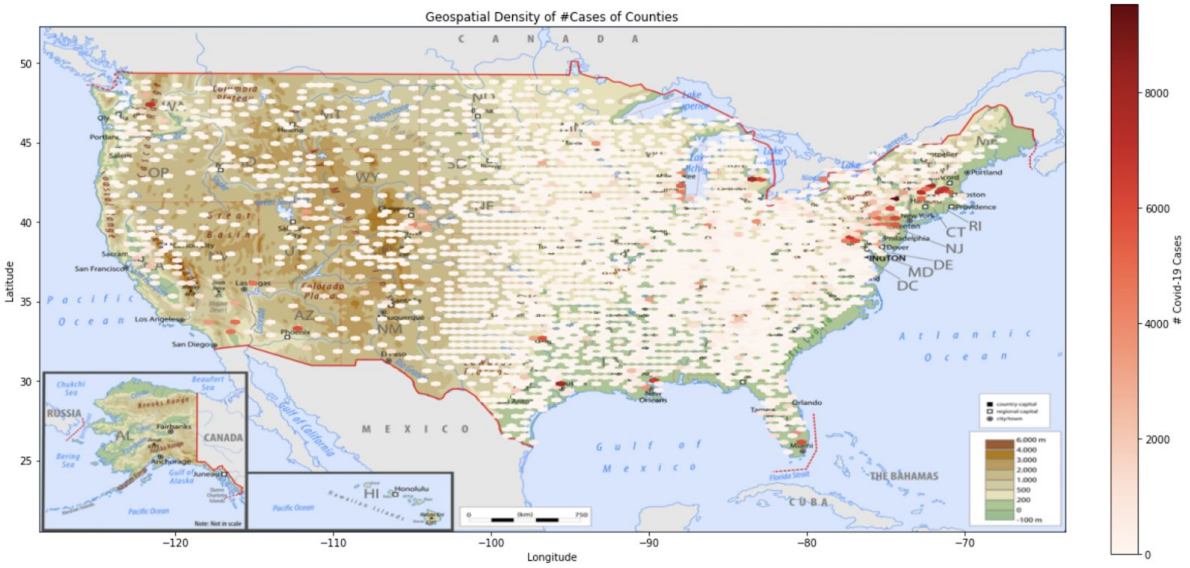
1. Geospatial Plots

Geospatial plot of Number of Confirmed Cases of each State on May 7th 2020



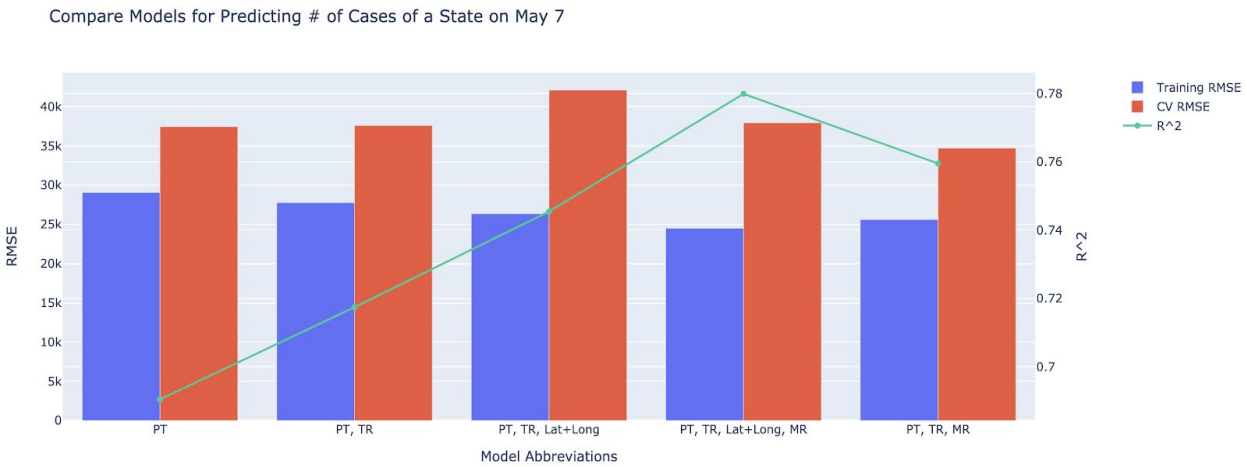
In this geospatial visualization of the number of confirmed Covid-19 cases for states in the United States on May 7th, we can see that the Northeast of the US has the most confirmed cases, especially areas near NY. We can see that color code gets lighter as we go in towards the middle of the United States, indicating more cases in the outskirt states and less in the middle of the US. Therefore we can see that geographics somewhat show patterns in the number of confirmed cases and geographical indicators should maybe be considered in our future model.

Geospatial Plot of Number of Confirmed Cases of each County on May 10th 2020



The outcome of this graph is similar to the state's plot as more counties with large numbers of confirmed cases are centered around New York, the east coast. With this graph, we are able to see some counties that have more cases than the surrounding counties in other areas of the US.

2. Compare performances of State Models



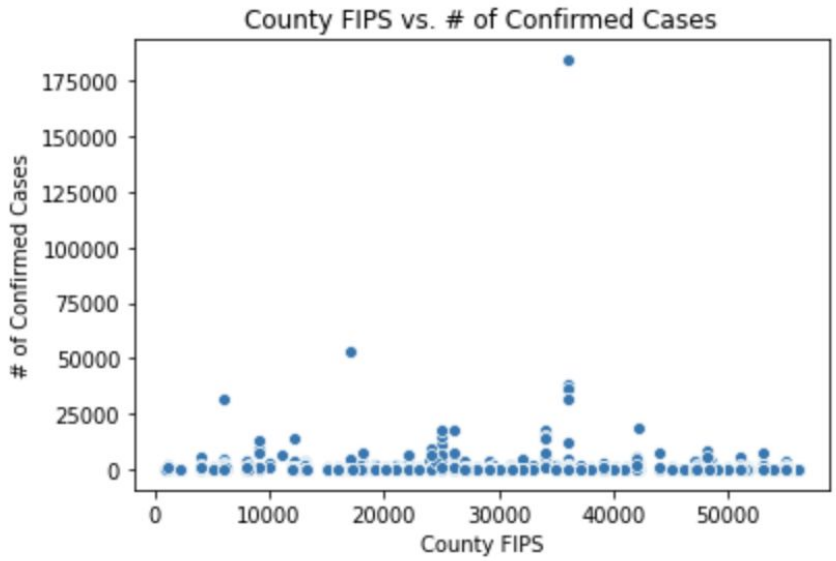
By fitting all 5 feature columns in the data set *People_Tested, Testing_Rate, Lat, Long_, Mortality_Rate*, we are able to maximize R^2, which means all 5 features explain the most variances of our data.

The RMSE and CV remain huge as more features are used. But the errors do reduce with more features used. However, the value of errors are around 30k to 40k, which indicates the models are not as ideal. Recall the geospatial plot for the number of cases in each state on May 7th, the value of confirmed cases in the United States has a wide range from around 5000 to around 30k. Therefore, we suspect that there are outliers that result in huge prediction errors.

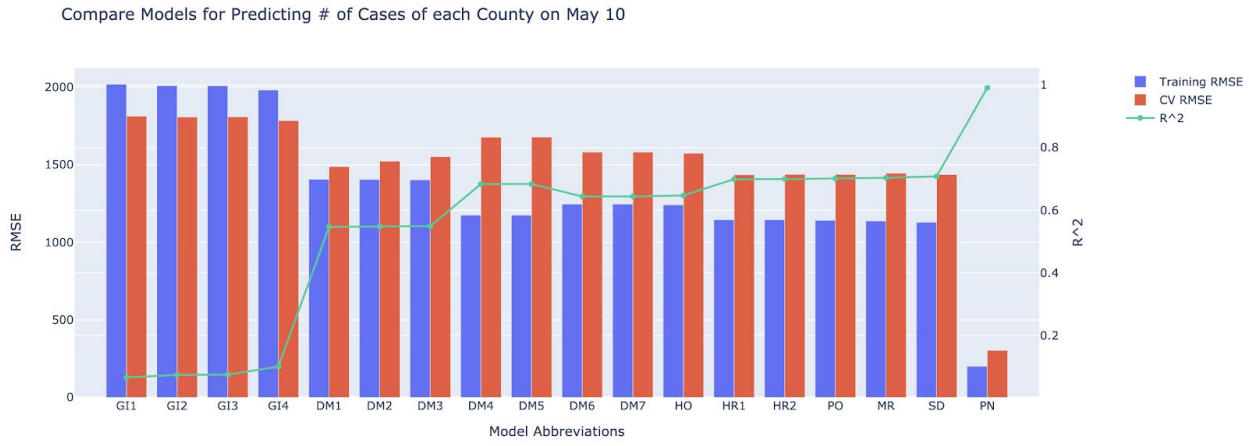
With the limited amount of features of each state, predicting the number of cases of each state on a given day seems harder than expected. As counties have more available features, we wonder what will happen if we try to predict the number of cases in a county based on available information? Would the results be better than predicting based on states? What are the best features to use to predict the number of cases of a county on a given day?

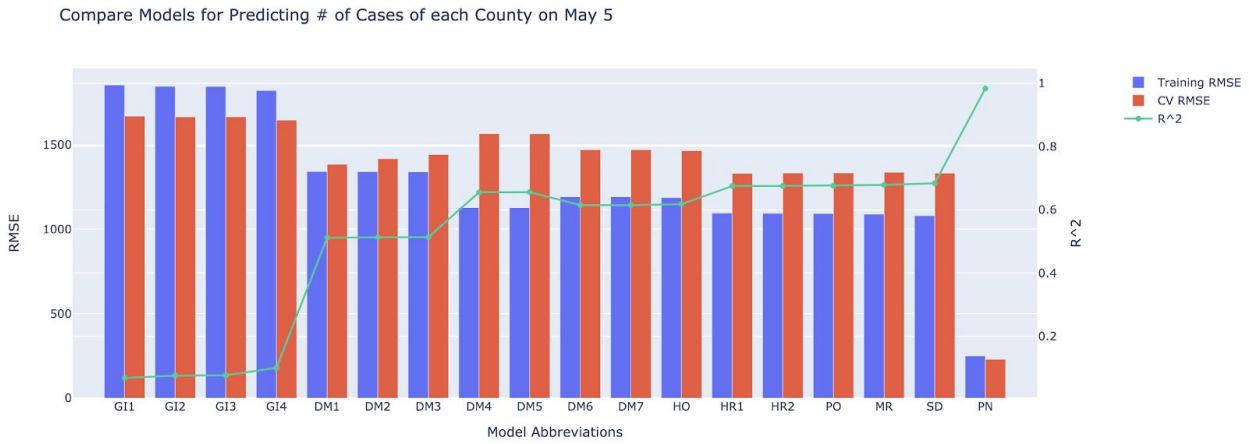
3. Visualizations on County Data Set

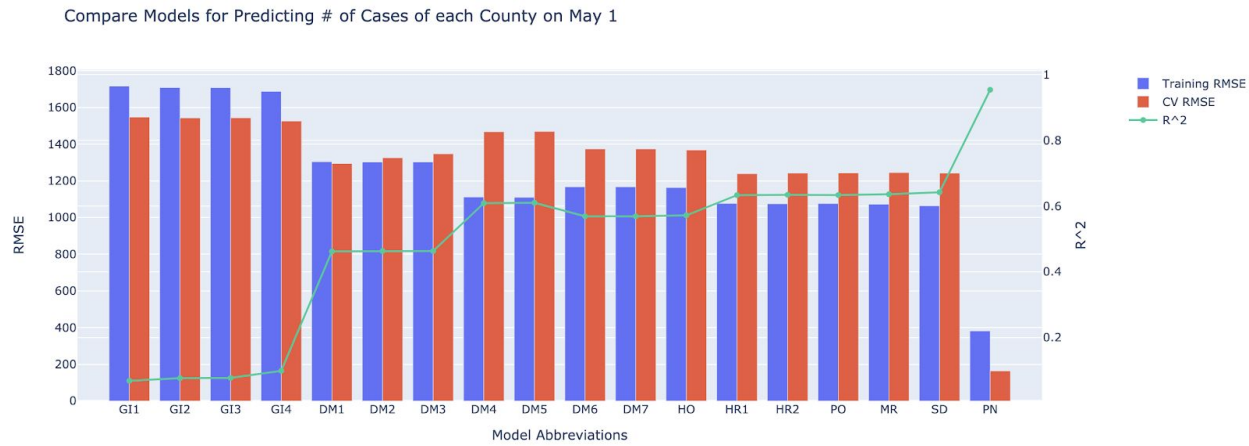
After splitting county data set into training and test set, we visualized our training data set and found out an outlier that had # confirmed cases > 175k. We discovered the outlier to be New York. New York has had the most # of confirmed cases in the US and the difference between NY and the rest of counties is very significant. We will remove New York from the training set because no other counties in the training set have similar situations.



4. Compare Performance of County Models







As indicated by the 3 graphs above, the linear regression models perform similarly predicting the three selected dates in May. In terms of features, we notice that adding all the available features other than known #cases of previous dates would reduce rmse to about 1200. Adding data of # confirmed cases before our predicting date significantly reduces our errors and increases R^2 to almost 1. We would rate the best model for predicting # covid-19 confirmed cases of a county on a given day to be the 'PN' model, which contains all indicators and some

information of # confirmed cases in April. However, if we do not have # confirmed case information, the best model to use will be among 'HR1/2', 'PO', 'MR', 'PN', which are models containing the majority of the available features.

Interesting Features that came across

There are many interesting features that came across when performing analysis on the county data sets. For example, the census population indicators in 2010, which is consisted of how many people are in each age group and sex, improved the model more than we expected. By fitting all census population indicators in 2010, except for the total census population in 2010, R^2 of our model is increased by about 0.15, meaning census population indicators helped to explain more variances in our data. However, it also caused our cross validation to increase, which may indicate overfitting. Using only male indicators in 2010 does not change the result.

In addition, the "dem_to_rep_ratio" feature, which indicates the ratio of popularity of democratic to republican candidates, is also interesting. Originally, we did not expect political popularity in a county to affect covid-19 outcome, but this indicator turned out to increase R^2 by about 0.01, and reduced our errors by a small amount. Therefore, political popularity in a county may actually be related with the situation of covid-19 in a county.

Ineffective Feature

One feature we thought would be useful was mortality rates of different age groups. Mortality rates could be indicators of how well medical resources in a county is. If a county has relatively low mortality rates, we may expect the county to be better at controlling the spread of coronavirus and have less confirmed cases. We have mortality rates of 5 age groups in all counties, and the abundance of the data made us expect more from its improvement in our models.

However, putting all mortality rates we have into the model only has a little improvement on the model, which is slightly bigger than the effect of political popularity, but so much less than other health resource availability indicators.

Challenges

We were most challenged when dealing with trying to use linear regression to predict the number of confirmed cases of each state based on available information. The features are limited, but we do not know that this is the biggest problem. The values of the number of confirmed cases of each county are very extreme and differ from each other by a large amount. The errors of state models were around 30k, which are really big, even though the R^2 of the models is not as bad. Looking back, we think we should categorize the states into different severity levels of covid-19 based on the number of confirmed cases on an earlier date, and design models that correspond to different categories of severity of states.

Limitations of the Analysis, Assumptions that could be incorrect

When dealing with features of counties, there were many missing values in mortality rates of different age groups of each county. Some mortality rates have about 1000 missing values, which is almost ⅓ of the data set. We decided to replace most mortality rates with mean values because we expected mortality rates to be closely related to the number of confirmed cases. However, the assumption that many state's mortality rates can be replaced with mean values of other states, can be incorrect. After fitting 5 age groups mortality rates to our model, the indicators did not bring much improvement to the models. Therefore, the missing values were limitations to our analysis, and obtaining additional information on mortality rates may bring differences to the performances of our models.

Ethical Dilemmas

When we added all census population indicators in 2010 to our country models, the cross validations errors of the model started to increase. Therefore we think that fitting all census population indicators may be overfitting the model. In order to address overfits, we decided to use only male census population indicators in 2010, and exclude those of females. The result turned out to be no different from fitting all indicators. However, if we actually decide to use indicators of one gender and not the other, there might be ethical problems, for instance, maybe we should not assume the composition of male population and that of females have similar correlations with confirmed covid-19 cases.

Additional Data that will Strengthen the Analysis

As mentioned in section "Limitations of the Analysis", if we have more data on mortality rates, we would be interested to know how much actual mortality rates of each age group would improve our models in predicting daily covid-19 cases for each county.

Ethical Concerns and How to Address

As addressed in "Ethical Dilemmas", the choices of data may raise ethical concerns in terms of what age group, gender, race is dominant in the data you are using to fit the model. A general case would be, for example, by fitting more data of male over female, your model may be biased towards male and be less accurate predicting on female. These ethical concerns may arise in any model training and machine learning. In order to address such problems, we should be careful and fair in choosing data, be mindful of different types of people, and try to pick the most efficient features considering a bigger picture of the project.

Conclusions

Performing linear regression analysis on state and county datasets, we found that predicting for each state may be difficult given the information we have, and predicting for each county is more efficient and accurate with county features that we obtain. The best features to predict daily number of confirmed covid-19 cases of each county includes but not limits to population density indicators of a county, when the county starts to restrict social gathers, hopital resource indicators and population composition indicators. We believe that the models can be improved further. Other attempts could be done trying other models and predicting methods like

principle component analysis. In addition, because the value of the number of cases in a county differ greatly, we could categorize counties into severity groups based on the previous number of confirmed cases and build models on each severity group. By fitting different models to different categories, the accuracy of the models may be improved significantly.