Covid-19 Project Report

Jerry Shan, Joseph Pang, Michael Go

Abstract

In this project, we explored the historical data for Covid-19 and came up with a linear regression model which predicts the number of confirmed cases for a new day for a state based on its data for the last 14 days as well as certain features about the state, such as its population density. Alternative numbers of days were considered, but 14 appeared to be an optimal balance between data availability and accuracy. The model attained an average root mean square error of ≈ 189 cases in its prediction for the test set. It was also found that the day right before the target day has the strongest correlation with it. Other useful indicators were "stay at home" policies and restrictions on > 500 gatherings.

Introduction

As Covid-19 spreads across the world, monitoring the statistics regarding the number of confirmed and new cases at the state level becomes important for many people. As students of Data 100, we were interested in exploring the patterns underlying the trajectories of confirmed cases, and to use technology to analyze what factors are most responsible for the growth of new cases. Of the many forecasting techniques, we decided to use linear regression because we thought that it's not unreasonable to model the number of new cases as a linear combination of previous data, so we decided to build a model that could forecast the number of confirmed cases on the next day based on past data.

Forecasting is also a great opportunity for us to learn about factors that drive the trajectory of confirmed cases. One of the available datasets is a variety of statistics about each county, such as the dates that it started certain Covid-19 precautionary policies (e.g. stay at home policies), gender proportions, age distribution of the population, percentage of smokers, etc. This is the main source of information that we used to assist in our forecasting model besides historical confirmed cases counts.

Description of Data

The data that we are working with is from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The data includes states and counties from around the world, along with the populations, confirmed cases count, death count, and other information in regards to their respective states/counties. Additionally, other data that could be helpful in regards to COVID-19 such as >50/>500 gathering and hospital count is also included.

Finally, we also used 2 other datasets from outside sources. First is the 2019 US Census, where we used the estimated population of each state. Second is the state area data where we got state area in sq miles.

Data Cleaning + EDA

We started off having a lot of data to clean. Of the 4 provided datasets for Covid-19, we decided to use three of them (counties data, states total statistics as of 5/6/2020, and confirmed cases by date by county), as well as two additional data sets related to the population and area of the US by state.

The counties data set was a particular useful one, because it contains a lot of statistics that are potentially relevant to Covid-19. Of course, not all features are readily convertible to numerical data useful for training, and others are obviously uncorrelated with Covid-19 confirmed cases (e.g. latitude and longitude, area codes) that we decided to discard right away, so we extracted a selection of features from the counties table to start exploring. The purpose of EDA is then to eliminate the weakly correlated features from our selection to reduce the amount of overfitting.

EDA is further divided into two stages. In the first stage, we were interested in learning about the current situation for Covid-19 and didn't target any specific feature. We generated a couple of diagrams for this purpose. One of them displays the prevalence of confirmed cases and deaths among all states.

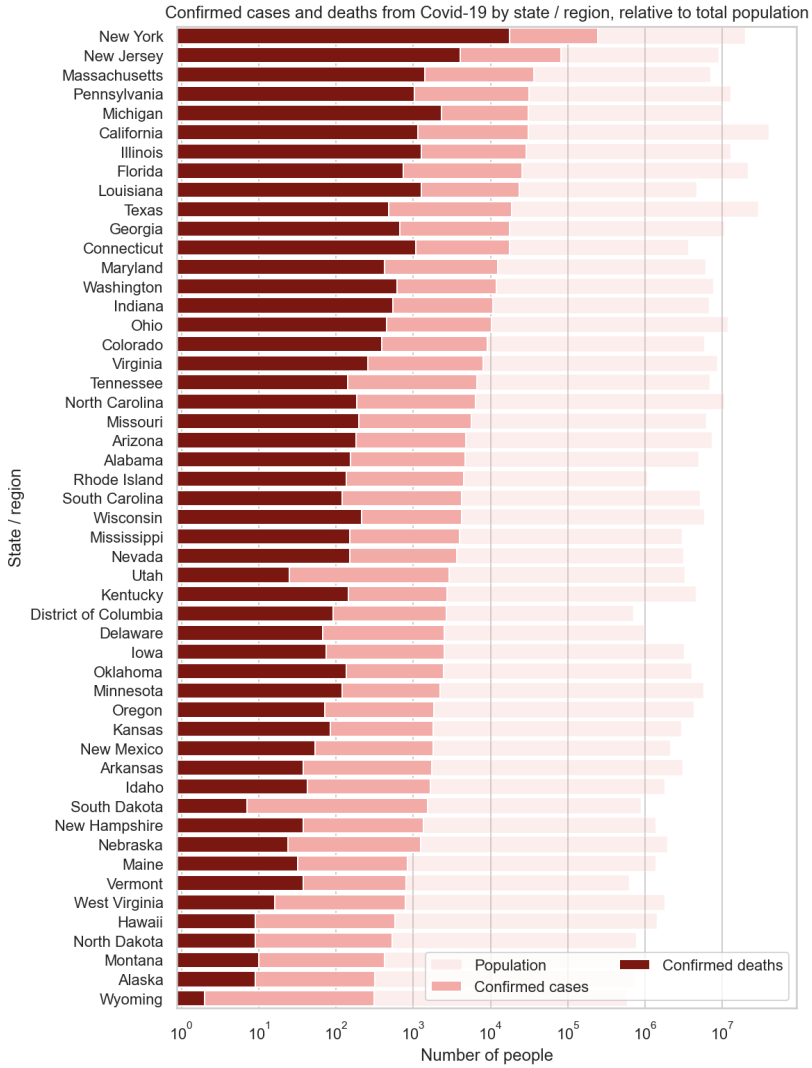

Figure 1: Confirmed and death prevalence for Covid-19 in the US.

From those, we learned that except for New York, an outlier, most other states are evenly distributed in terms of confirmed cases. Next, we decided to visualize the growth trend of confirmed cases for a subset of the states to get a feel for whether forecasting makes sense in the first place, and the graph reassured us that there must be patterns. In particular, we saw that the confirmed cases are increasing at a rate slower than exponential (since the curves are concave upward).

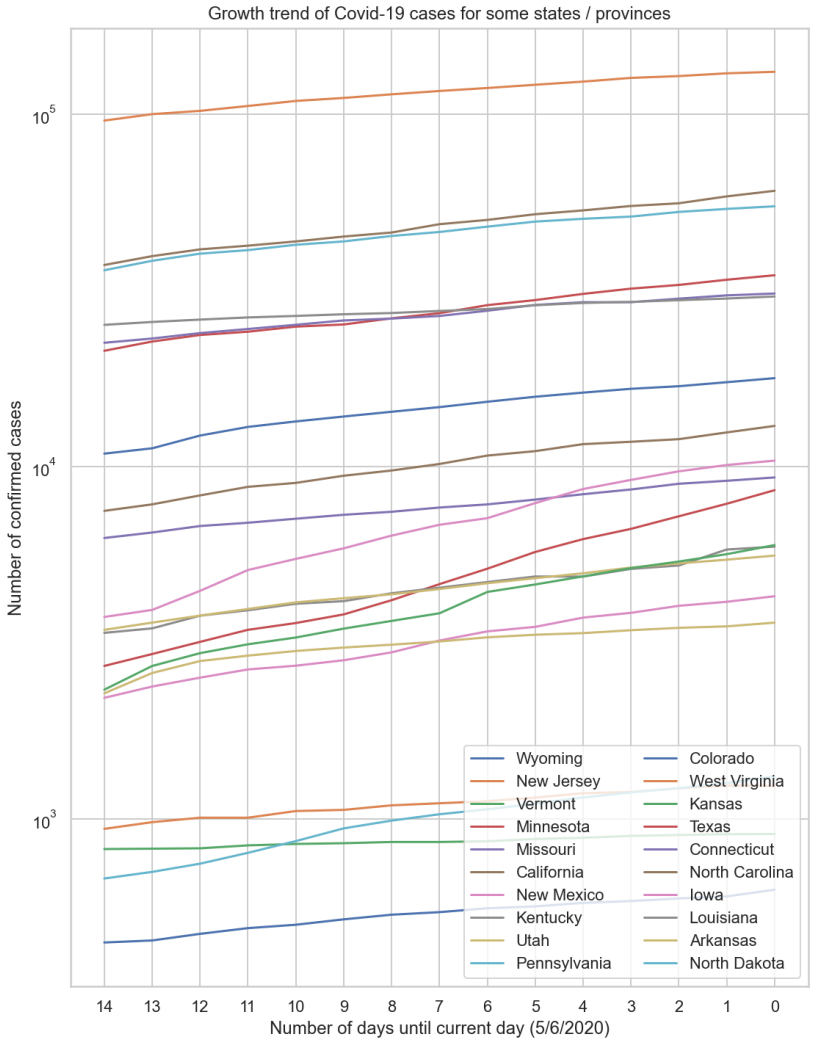

Figure 2: Confirmed cases trends as of 5/6/2020, for selected states.

The next part of EDA consists of examining features that could be useful in making predictions about the number of confirmed cases. To prepare the data, we merged the cases data with everything else from the datasets which we thought might help with forecasting, and started drawing scatter plots to rule out the useless features. For instance, gender proportion was found to be not as useful, whereas the number of hospitals and population density were found to have positive correlations with the number of confirmed cases. Also, we found that two features had a correlation of 0.983, so we dropped one. Eventually, we removed weakly-correlated features from the training data. The feature with the strongest correlation ($r = 0.426$) turned out to be the population size of the state / province.

Data Transformations

The two datasets that were rather difficult to transform were the counties statistics and the confirmed cases statistics. The former was difficult because we would like to know the statistics about states, not counties, so we must group the counties by their state and aggregate a few selected values. Aggregation was performed twice because some columns should be aggregated by mean while others should be summed. The resulting table was called `state_info` and contains a range of candidate features for each state

which we thought might be relevant to Covid-19. The timestamps were later encoded appropriately to boolean variables indicating whether the policies were in effect since that day. However, for clarity, the table below did not show the transformations.

	MedianAge2010	Smokers_Percentage	FracMale2017	stay at home	>50 gatherings	>500 gatherings	entertainment/gym	restaurant dine-in	#Hospitals	#EligibleforMedicare2018	#ICU_beds
State											
Alabama	39.356716	19.989231	0.486846	2020-04-03	2020-03-20	2020-03-13	2020-03-28	2020-03-19	86.0	1080141.0	1533.0
Arizona	38.653333	16.483911	0.503349	2020-03-31	2020-03-17	2020-03-17	2020-03-19	2020-03-19	76.0	1346727.0	1559.0
Arkansas	40.316000	20.388849	0.494874	NaN	2020-03-26	2020-03-26	2020-03-19	2020-03-19	74.0	670352.0	732.0
California	38.503448	12.091600	0.504910	2020-03-19	2020-03-19	2020-03-19	2020-03-15	2020-03-15	329.0	6466995.0	7338.0
Colorado	41.265625	14.297498	0.519244	2020-03-25	2020-03-24	2020-03-13	2020-03-17	2020-03-20	80.0	965047.0	1095.0

Figure 3: Other features about each state, taken from `states_info` table (only showing 5 states).

For the confirmed cases table, we needed to do a lot of transformation because the dates were originally given as columns, whereas we would like to represent them as another row value. Moreover, for each day, we would like to associate its confirmed cases with its last 14 days. This was achieved by melting the columns down and then group the entries by their state and sum up their confirmed cases. Following that, we iteratively joined the table with the confirmed cases records to obtain the data for the last 14 days. Eventually, we were able to transform the confirmed cases data into the following dataframe structure:

	Province_State	Current_Date	Current_Confirmed	1 day(s) ago	2 day(s) ago	3 day(s) ago	4 day(s) ago	5 day(s) ago	6 day(s) ago	7 day(s) ago	8 day(s) ago	9 day(s) ago	10 day(s) ago	11 day(s) ago	12 day(s) ago	13 day(s) ago	14 day(s) ago
0	Alabama	2020-02-05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Alabama	2020-02-06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Alabama	2020-02-07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Alabama	2020-02-08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Alabama	2020-02-09	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5239	Wyoming	2020-05-02	579	566	559	545	536	520	502	491	473	453	447	443	317	313	309
5240	Wyoming	2020-05-03	586	579	566	559	545	536	520	502	491	473	453	447	443	317	313
5241	Wyoming	2020-05-04	596	586	579	566	559	545	536	520	502	491	473	453	447	443	317
5242	Wyoming	2020-05-05	604	596	586	579	566	559	545	536	520	502	491	473	453	447	443
5243	Wyoming	2020-05-06	631	604	596	586	579	566	559	545	536	520	502	491	473	453	447

Figure 4: Confirmed cases by date and state, after transformation, taken from `confirmed` table.

This table was then joined with the other features about the state (from the `state_info` table) to form the training data.

Methodology

Our model was a linear regression instance (regularization was considered but not used because there aren't so many features), and we prepared the training data by having the number of confirmed cases for each state on each day as the response variable, and the past 14 days data and other information about the state as covariates.

Then, we split the data into a training set and test set, containing 90% and 10% of the total data respec-

tively. The training data was then used to train the model. Using cross validation with a split count of 5, we determined that the root mean square loss of the model was around 251.84. After we finalized model, it was found that the root mean square loss of the model on the test set was ≈ 189. The 21-day model had a loss of 201.67 on the test set, and the 7-day model had a loss of 230.34 on the test set.

Some assumptions that we made include the policies implemented each state. Due to the lack of data at our disposal, we are unable to account for governmental policies implemented, which are a heavy factor to the spread of COVID-19. Thus, as of now our model is making the assumption that policies make negligible difference in confirmed cases.

Another big assumption is that the growth rate of confirmed cases is a linear growth. However, if the actual growth rate was a polynomial growth, then our linear regression model would not be a good model for prediction. This assumption is partly due to our current capabilities and knowledge of models.

Results

Testing

In order to test the accuracy of our model without directly testing on the testing data, we used cross-validation on the training data (with a fold of 5). We created 3 iterations of our model, with cv rmse scores of about 251, 278, and 259 cases respectively. Finally, we tested our best model, the first one, on the testing data and got an rmse score of 189 cases.

Interesting + Useful Features

Here were the coefficients for our final model regarding the features we included:

Feature name	**Coefficient**
`Current_Confirmed`	1.518
`1 day(s) ago`	-0.367
`2 day(s) ago`	-0.084
`3 day(s) ago`	-0.033
`4 day(s) ago`	0.145
`5 day(s) ago`	-0.028
`6 day(s) ago`	0.015
`7 day(s) ago`	-0.244
`8 day(s) ago`	0.139
`9 day(s) ago`	-0.09
`10 day(s) ago`	-0.136
`11 day(s) ago`	0.098
`12 day(s) ago`	0.045
`13 day(s) ago`	0.022
`14 day(s) ago`	-9.98
`stay at home`	18.303
`>50 gatherings`	0.178

>500 gatherings	12.516
entertainment/gym	10.111
restaurant dine-in	-0.022
#Hospitals	0.002
#ICU_beds	0.0
Population	0.047

Table 1: Coefficient values for our model.

The coefficients are actually messier than what we expected. In particular, we didn't expect that the coefficients for the days are so different from one another, since during EDA, we found that the trajectories were quite smooth. This led us to conclude that perhaps we overfitted.

We could also tell that 'stay at home' and restrictions to '>500 gatherings' were popular policies when confirmed cases go up, whereas '>50 gatherings' restrictions weren't so popular. Moreover, restrictions to 'restaurant dine-in' were even less popular.

It was also surprising to find that the number of hospital beds and the number of hospitals had little to do with the number of confirmed cases. During EDA, we used the scatter plots to see that there are visible correlations.

Evaluation

One of the challenges we encountered was that the coefficients for our final model did not directly inform us that what were the most influential factors in the spread of Covid-19, because after all, the model's goal is to make forecasts, not to analyze the weight of the features considered. For instance, we found it surprising that there is a positive correlation between restrictions on large gatherings and the number of confirmed cases, as seen by the coefficient +18.3 for the boolean indicator variable '>50 gatherings'. However, we cannot simply infer that restrictions on large gatherings had a *causal relationship* with the spread of Covid-19. We could only tell that when there are higher numbers of confirmed cases, state governments were more willing to put up restrictions on public activities.

Another limitation in our model is that it's inherently linear in the features that we used, as mentioned in our assumptions. However, the actual growth trajectories of the disease may be super-linear in nature, so perhaps including the squares or exponential powers of historical data in our features could potentially improve the model. Unfortunately, we didn't have enough time to carry out this approach.

Finally, we must admit that the datasets themselves certainly didn't capture all of the trends that underly the spread of Covid-19. The spread of a disease is not easy to model due to so many factors, some of which are not quantifiable and/or retrievable, such as political factors and the influence of the social media. However, we've shown that there are certainly patterns in how a disease spreads, and data science is a powerful discipline that could find these patterns.

Ethical Dilemmas and Concerns

While working on this project, we realized that if we could get more data on the individuals who were confirmed to have the virus, would could potentially get a more accurate model. Data such as listing out specific areas with higher confirmed cases or proportion of ethnicity groups that are confirmed could drastically improve the model. However, this creates a potential ethical dilemma. Although more data could increase the accuracy of the model, to collect and use this data would require us to infringe on the privacy of many people. Especially with ethnicity data, this could bias our results and cause us to come to conclusions that paint certain groups in a false negative light.

Furthermore, we also have ethical concerns with how a model like ours could be used. Hypothetically speaking, if we're able to improve our model to have a high accuracy, our predictions can be used by state governments to decide what policies to implement (such as shelter in place). Although our predictions can be helpful in deciding what policies to implement, many don't realize that our model doesn't explain everything. Government officials could potentially take a model such as our and base their policies solely based on the predictions generated. As a result, they would be overlooking a lot of information that we don't address in our model and as a result, their policies could be hurting people who being impacted heavily by Covid-19. Furthermore, if the general public were to have access to our predictions, it could create unwarranted panic for many who don't understand the limitations of our model.

To address this, transparency and reflection is necessary. If one wishes to use a model such as ours, they must first understand how the model was created, what is addresses, what it overlooks, and potential problems with it.

Additional Data

We believe the following features could strengthen our model:

- Shelter-in-Place Implementation: Information on when or whether or not the state or county implemented a shelter-in-place and for how long
- Policies Implemented: What policies have the government implemented to fight the spread of Covid-19
- Incoming Flights: How many flights that dropped people off in the respective state or how many people entered the state on that day
- Government Aid: Whether or not the government has provided aid for the state citizens and if so what kind of aid as well
- The availability of precautionary items such as masks, gloves, sanitizers, etc. These items are believed to help reduce the spread of Covid-19 and was in shortage for extended periods of time.

Future Work

In the future, we hope to find new features + test out new models in order to increase the accuracy of our model. One way change we could implement is using a non-linear regression model. Due to the scope

of this class, we are only comfortable with linear and logistic regression models for prediction. Thus, to improve our model, deviation from a linear regression model might be necessary. Additionally, more research on COVID-19, and collection of potential features data will be crucial as well. With the global pandemic happening right now, we hope to help contribute to a solution in any we can, regardless through data science or not.

Another potential area of research is to build a similar model to predict the number of deaths. Of course, this probably requires another EDA to determine which factors are strongly correlated with death tolls, and adjust the training data accordingly. For instance, although median age appeared to be weekly correlated with the number of confirmed cases, it is not impossible that it turned out to be more strongly correlated with death cases, since it is believed that Covid-19 is more threatening to elderly people in general. Building a death forecasting model would be beneficial to learn about measures that might reduce death counts.