Summary

This ML project leverages the dataset from Kaggle about the recently developing coronavirus. Given the open-ended nature of this exercise, I have first conducted an exploratory data analysis to understand the data and derive any actionable insights. Then, based on these insights, I defined a useful question for which my project will be based. Finally, I implemented several models for predictive analytics, and evaluated them in relation to relevant metrics. Please keep in mind all relevant coding was done in R.

Given that there was both individual and aggregate level data, it was essential to understand the underlying data. I approached this problem by first trying to figure out interesting questions that could be answered by looking at the death rates. It turns out by no surprise that there were very limited death rate data in the individual level data. It all points to very concentrated places in China like Wuhan and most among those of old age. There was no particularly interesting prediction or model building tasks here that could be extrapolated.

With that in mind I thought one interesting direction to take would be try to predict trends of confirmed cases in the time series data. One useful application of this is that I could use such model to extrapolate how the coronavirus might develop in places that are just starting to experience exposure such as New York or Boston. However, this extrapolation is extremely dangerous as New York and Boston is not similar to China in any ways. Even if it was currently the virus is more “alerted” than it was in the past and thus the development of the virus won’t have parallel trends.

Nevertheless, it was still an interesting direction to take as at least the time series model I built could be applied locally to that current country/province and see how the confirmed cases might grow or stabilize in the future. In terms of aggregate data, I only analyzed the countries with more than 1000 confirmed cases, which leaves me to the following 5 places: Guangdong, Hunan, Zhejiang, Hubei, and Henan. From a very simple EDA the trends for all five places are very similar. Therefore, just for the sake of time I have only analyzed Guangdong in depth. However, the analysis and code could be generalized to include the analysis of any 5 of this.

The first thing I did was to build an ARMA model by looking at the differenced plot, the acf, and the pacf to determine reasonable values of my (p,d,q) parameters. Seasonality didn’t seem to be an issue here so I stuck with this. From the EDA I determined reasonable values of p, d, q and tested through AIC and cross validation what the best model was. I then plotted by trying to predict the next 20 days of what might the confirmed cases in Guangdong look like including the confidence intervals.

I could have stopped here but one possible modelling approach while looking at the first graph in the EDA would be to try to parametrically model it with a sigmoid function. A sigmoid function could be modelling in many ways but one way would be to use a smoothing spline with varying degrees of freedom. A reason why I wanted to do this is because one problem with the ARMA is that it will constantly project the future as having more and more cases. This obviously does not match real life. A sigmoid function will approach a limit asymptotically for eventually have a fixed number of confirmed cases. This is what one should expect for a virus evolving because at a certain point people stop getting it and the rate definitely decays exponentially. ARMA does capture this well but a sigmoid function could do better.

As a result, I also fit a smoothed spline with degrees of freedom cross validated the same way. I plot the next 20-day prediction with the red points indicating the predicted points. We see that the predicted smoothing spline definitely has a much better “leveling-off” characteristic and we can see that at the very end when we see the point by point comparison as the last point prediction for the spline is around 1349 while the ARMA is close to 1370, which is a much faster unrealistic growth rate.

In conclusion one could theoretically take these models and project other countries/regions of what the coronavirus confirmed cases might look like over time. However, as explained above this extrapolation is not justified, as there are certain restrictions of the dataset. On one hand, there is limited availability of data extending from only 01/22/20 to 02/20/20. On the other hand, countries or cities located outside of China have few reported cases that might not be substantive enough to test the model. Nevertheless, this is a well-developed model in the end. At the very least we could use these models that I found to have good asymptotic properties and well cross-validated to locally predict what might happen to each of the respective places like Guangdong over the next few days. I thoroughly enjoyed this case study, and thank you for reviewing my work.