CMPT 733 Project Report:

Predicting COVID-19 and Analyzing Government Policy Impact

Team: Kacy Wu, Fangyu Gu, Steven Wu, Yizhou Sun, Srijeev Sarkar

1. Motivation and Background

On the 11th of March, the World Health Organization (WHO) officially declared ‘COVID-19’ (the Novel Coronavirus) a pandemic. With more than 2 million cases spread over 140 countries, and over 160,000 lives have been lost. In North American, the United States of America is now an epidemic centre with over 700,000 cases, while Canada is, unfortunately, catching up with 30,000 cases in a short period.

COVID-19 has negatively impacted our society and economy. The International Monetary Fund reports that the global economy has entered a recession since the outbreak and the current situation is worse than the 2008 global financial crisis. Millions of people have lost their jobs and unemployment rates are at an all-time high. Canada has seen a steep jump in its unemployment rate from 2.2% to 7.8%, the largest increase in a one month window since 1976. As the world imposes rigid government policies and invests large sums of money to better healthcare facilities, we as Data Scientists would like to contribute by working together to estimate future progressions of the disease.

2. Problem Statement

The objectives and contents of this project are divided into two fundamental parts:

Part A: Develop, test and compare multiple predictive time-series machine learning models to estimate and understand the spread of COVID-19.

Part B: Understand the impact of government policy and develop a model to provide insights on how effective the implemented policies are in terms of “flattening the curve”.

3. Data Science Workflow

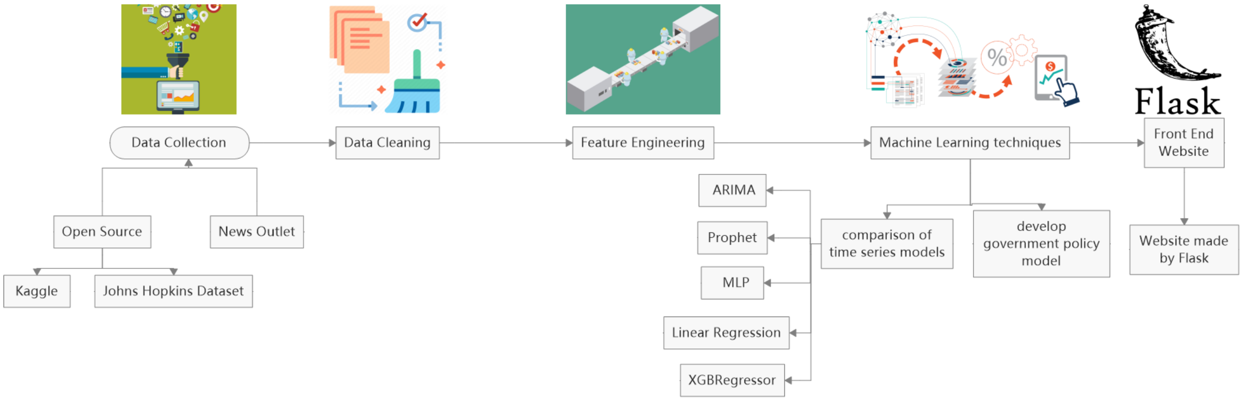


Figure 1: Data Workflow Pipeline

4. Datasets

- 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE.[1]
- Kaggle Novel Coronavirus 2019 Dataset.[2]
- COVID-19 containment and mitigation measures.[3]
- Canadian government policy dataset (Collected ourselves).

5. Exploratory Data Analysis

At the start of March, the pandemic started spreading at a rapid rate. On the 3rd of March 2020, China had almost 100,000 people infected, followed by South Korea with 5,545 cases and Italy with 2,706. In North America, the US only had 45 people infected, and Canada with 17 cases at that time.

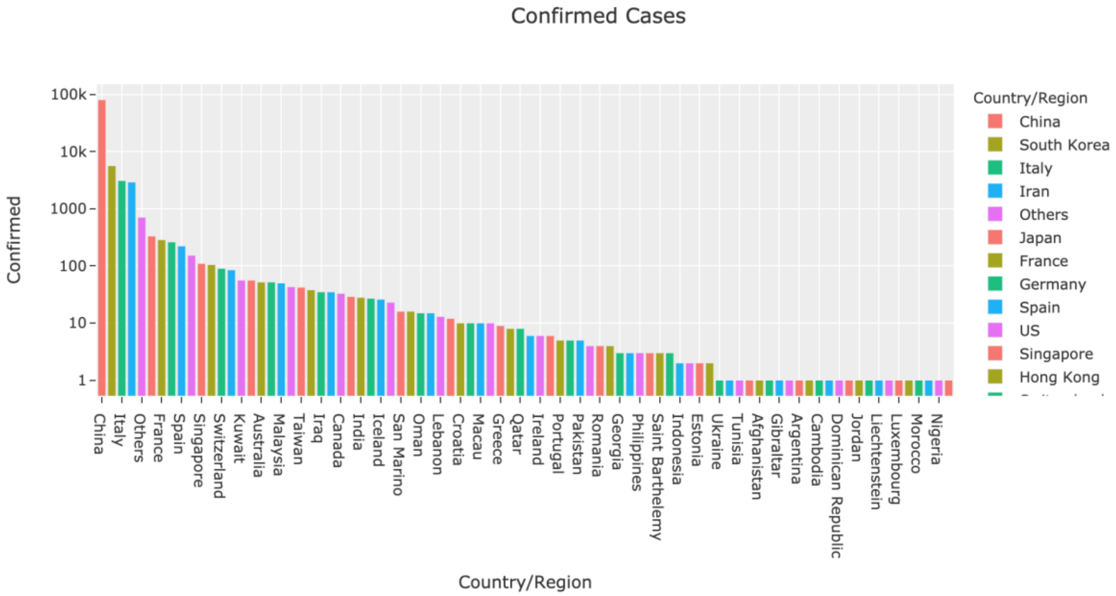


Figure 2: Global situation at the beginning of March 2020

Towards the middle of March, we were curious to see how the pandemic was developing across the globe. By using COVID-19 location data on folium maps, we were able to develop a clearer understanding on global confirmed cases, recoveries, and deaths.

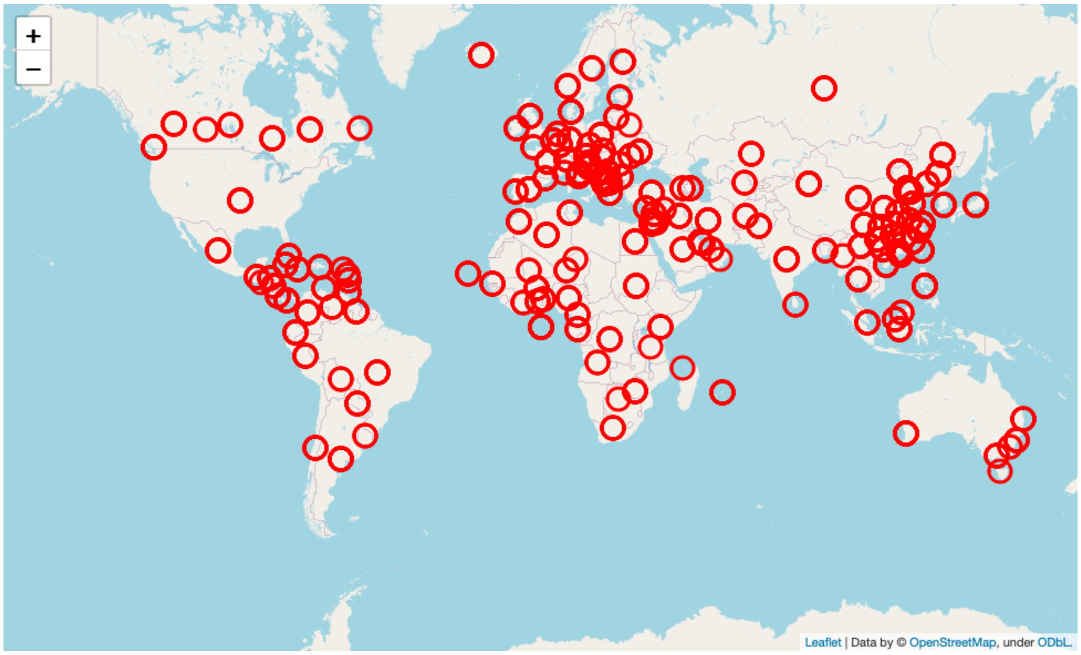


Figure 3: Pandemic deaths locations at the middle of March 2020

By the end of March, most countries started imposing travel restrictions, some format of lockdowns and encouraged people to stay at home. After constructing plots to understand the trend, it was very clear that countries such as Italy, Spain, Germany, United States of America saw a steep rise in the number of cases.

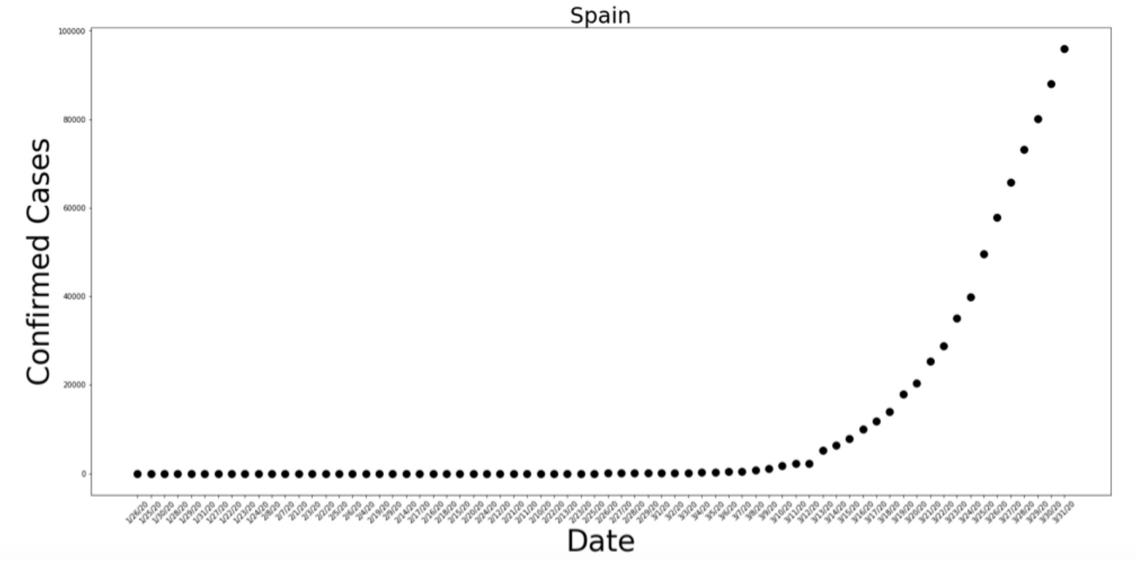
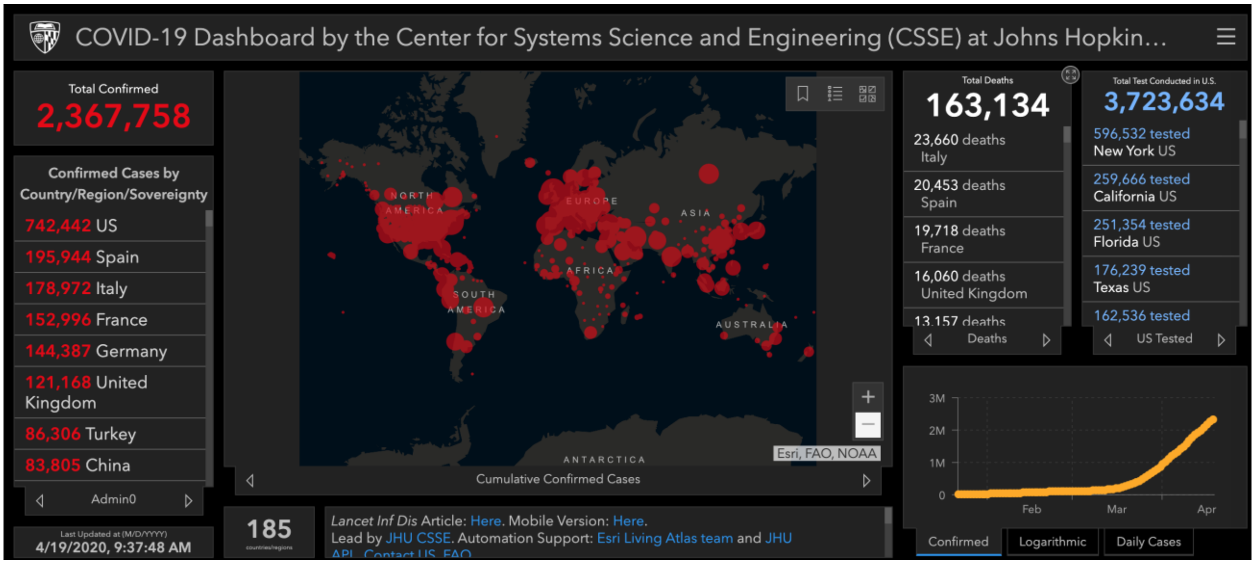


Figure 4: Pandemic Spread in Spain (Confirmed Cases) by the end of March 2020

Moving into the current global situation today, COVID-19 has infected more than 2 million people. Just after a month, the US became the new epidemic center with 764,177 confirmed cases, while Canada also took a huge jump to 35,036. Out of all the closed cases, the death rate is at 21%[4] and it is said that the virus is far more deadly to smokers and the elderly.



COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkin...

Total Confirmed
2,367,758

Confirmed Cases by Country/Region/Sovereignty

742,442	US
195,944	Spain
178,972	Italy
152,996	France
144,387	Germany
121,168	United Kingdom
86,306	Turkey
83,805	China

Admin0

Last Updated at (M/D/YYYY)
4/19/2020, 9:37:48 AM

185
countries/regions

Cumulative Confirmed Cases

Lancet Inf Dis Article: Here. Mobile Version: Here.
Lead by JHU CSSE. Automation Support: Esri Living Atlas team and JHU
API. Contact US. FAQ.

Total Deaths
163,134

23,660 deaths	Italy
20,453 deaths	Spain
19,718 deaths	France
16,060 deaths	United Kingdom
13,157 deaths	

Deaths

Total Test Conducted in U.S.
3,723,634

596,532 tested	New York US
259,666 tested	California US
251,354 tested	Florida US
176,239 tested	Texas US
162,536 tested	

US Tested

Figure 5: Global situation on the 19th of April, 2020 (Source: CSSE JHU)

6. Part A - Comparison of Time Series Models

6.1 Data Preparation

The data preparation for most time-series models primarily involved extracting the confirmed cases, death cases, and date-time data. The missing values were filled up, group by operations were conducted on the countries and day-count features were created for certain tasks.

For the Canadian provincial model, the provincial data for Canada was filtered and the most impacted provinces were sorted in ascending order for further analysis.

6.2 Methodology

6.2.1 ARIMA, Prophet and MLP Models

ARIMA

ARIMA is a generalization of the ARMA (an autoregressive moving average) model. The model is usually fit with time-series data for the purpose of forecasting future results.

Since COVID-19 is an ongoing event, and one of our assumptions is that the datasets are not seasonal.*

For non-seasonal ARIMA models, there are three-parameters: p,d, and q:

- p is the autoregressive model time lags number.
- q is the number of the moving-average model.
- d is the difference that time series needed to get a stationary series[5].

The COVID ARIMA model can be estimated using an approach known as the 'Box-Jenkins' approach[6]. The goal here is to check the data 'seasonality' and 'stationary' nature , and we checked these two features using sequence plots.The results indicate that there is no seasonality. The plots also show a rising up tendency, making the time series 'not stationary'.

We then run the autocorrelation plot and partial autocorrelation plot to generate a confidence interval using the ARIMA library. Finally, we trained the model using the p, d, and q value. Our model was first trained using data starting from January. However, the model did not fit very well. Because most countries have the outbreak around March, we changed our training dataset timeline starting from March. We used confirmed cases as input and made predictions for the next 30 days.

Prophet

In comparison to ARIMA, Prophet is an automatic time-series forecasting library. Ironically, there isn't a lot of 'time' spent in estimating the model's parameters.

First of all, we learned that the Prophet model automatically handles the non-continuous time series and fits non-linear trends, and It functions best with time-series data that have strong seasonal effects[7]. In this case, we can directly train our model in a similar approach.The prediction and actual confirmed cases were evaluated by MSE and log loss.

Multilayer Perceptron Model

We built the MLPRegressor with three hidden layers (32,32,10), and the model was trained using data starting from March as well.

6.2.2 Country Regressor Model

Regression being the gold standard of predictive analytics was the key philosophy behind this model. The two models used were the Linear Regression model from the python library Sklearn and XGBRegressor from the C++ (developed for python now) library XGBoost.

To perform training, The Johns Hopkins dataset is encoded, featurized and a newly created feature called day count is added.The day count is trained against the confirmed cases and fatality cases for the month of March.

The Linear Regression and XGBRegressor models can predict such cases for any day in April. The choice of country and date is based on user input and the resultant visualizations are available on the web front-end

6.2.3 Candian Province Model

The filtered data for Canada is first analyzed to understand which provinces have been impacted the most by COVID-19, and multiple Linear Regression models are used to train for each province.

The Linear Regression models were used to train the most impacted provinces in Canada, to make predictions for the month of April.

The training data is provincial-based confirmed cases and death cases data for March. All the predictions are pulled together and a master dataset is created. From our dashboard, the user is given the ability to select a date in April to see the provincial comparison of death cases/confirmed cases for Alberta, British Columbia, Nova Scotia, Ontario, and Quebec.

6.3 Results

6.3.1 ARIMA, Prophet & MLP

For these three models, we focused on analyzing the situation in these five countries:Canada, US, Germany, Italy and Spain.

ARIMA

From the following figures we can see that the confirmed cases in the US are growing the fastest, while Canada is the slowest. From the prediction result for Canada we can see that ARIMA is very sensitive to fluctuations in the training set. This is a possible reason why ARIMA performs better using the US data compared to the other four countries.

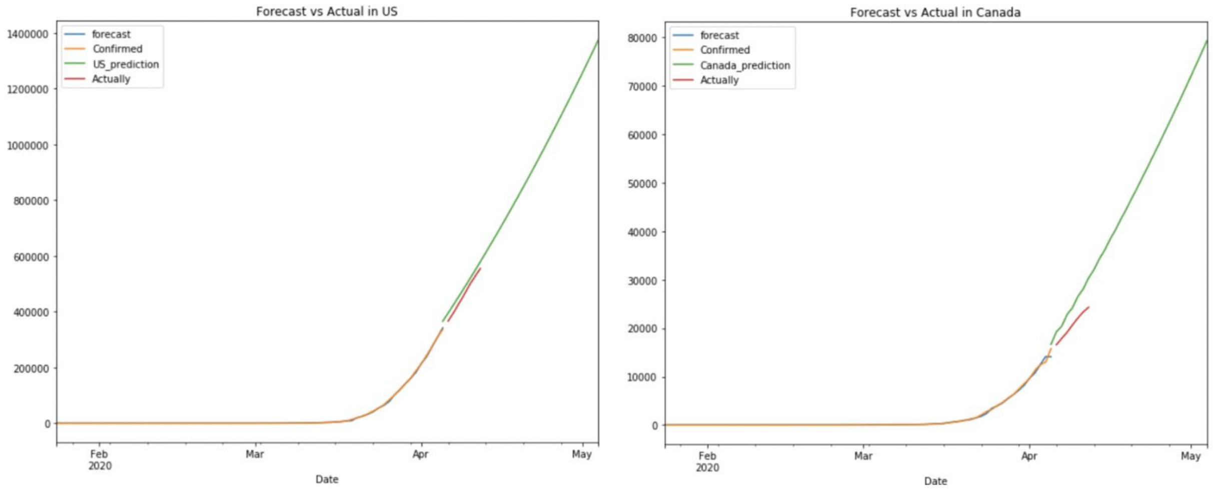


Figure 6,7: ARIMApredictions for the US (left) and Canada (right)

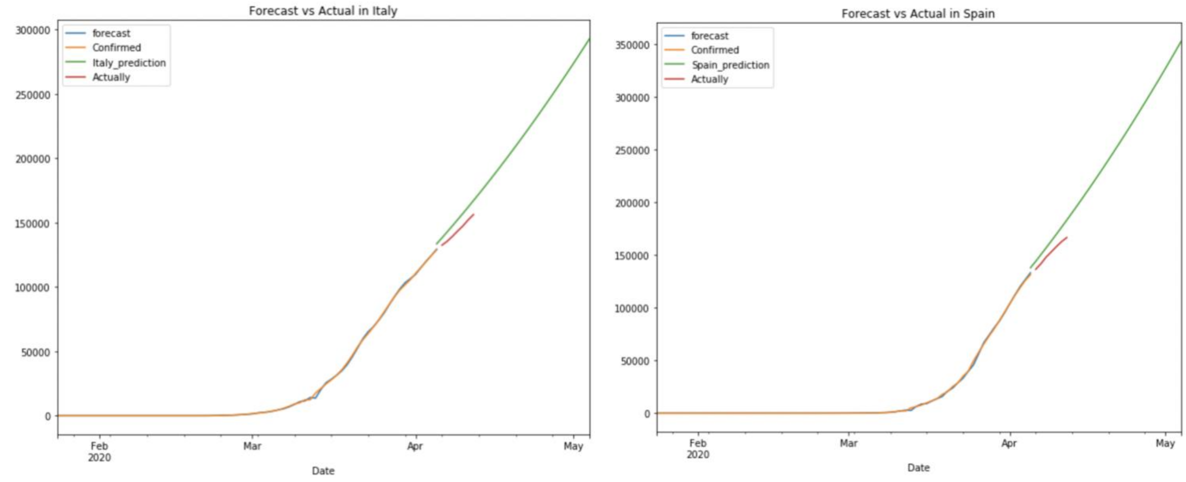


Figure 8,9: ARIMA predictions for Italy (left) and Spain (right)

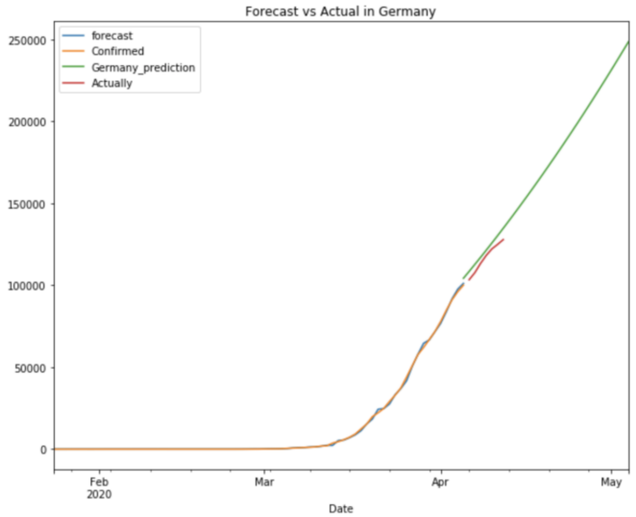


Figure 10: ARIMA prediction for Germany

	US	Canada	Italy	Spain	Germany
MSE	621,837,985.1	16,892,101.5	77,564,712.8	131,374,312.7	23,538,176.4

Figure 11: ARIMA MSE Comparison

Prophet

The MSE error is the largest for the US dataset, and the prediction is a bit off. One possible reason is that when dealing with non-stationary datasets, the Prophet model chooses to fit the trend. On the other hand, the US has adopted an aggressive testing strategy from the beginning of April, that might also explain why the prediction error is large here.

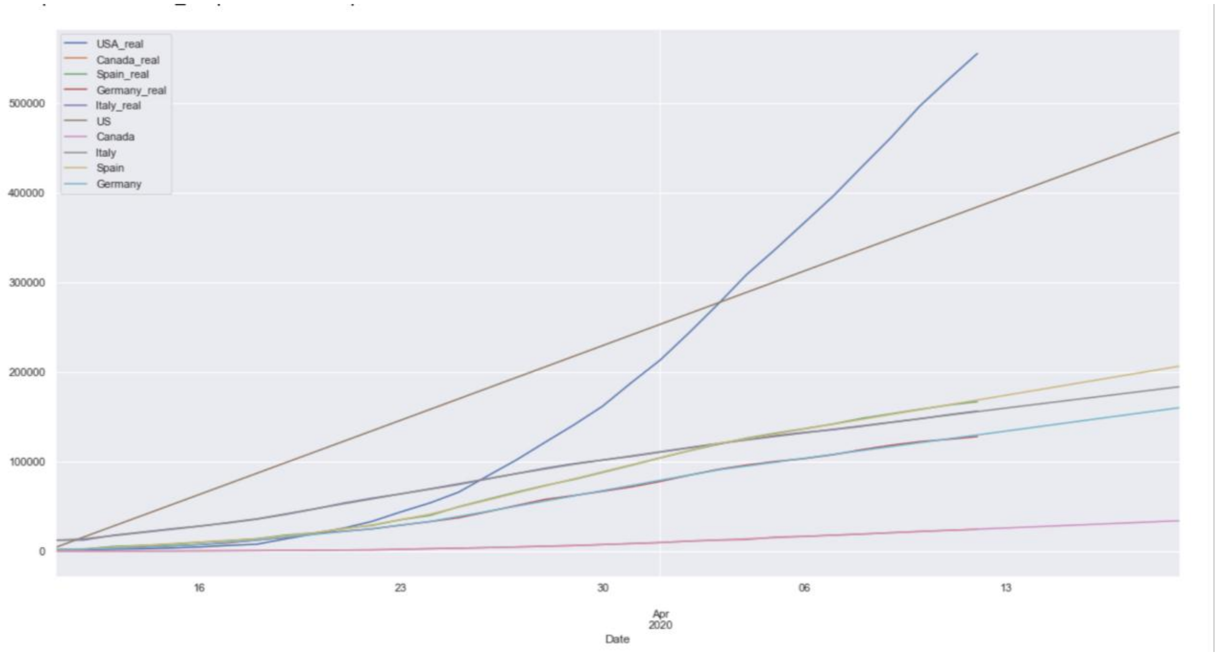


Figure 12: Prophet predictions for the five countries

	US	Canada	Italy	Spain	Germany
MSE	5,214,521,318.	29,853.4	241,674.2	318,319.5	518,466.6

Figure 13: Prophet MSE Comparison

Multilayer Perceptron

As the results shown below, It is easier to tell that MLP c performs better than ARIMA and Prophet. It is not very sensitive to fluctuation and avoids fitting the trend in a simple straight line. One interesting fact here is that both MLP and ARIMA models show that the confirmed cases in the US will increase to one million and Canada will increase to 40,000 before May 1st.

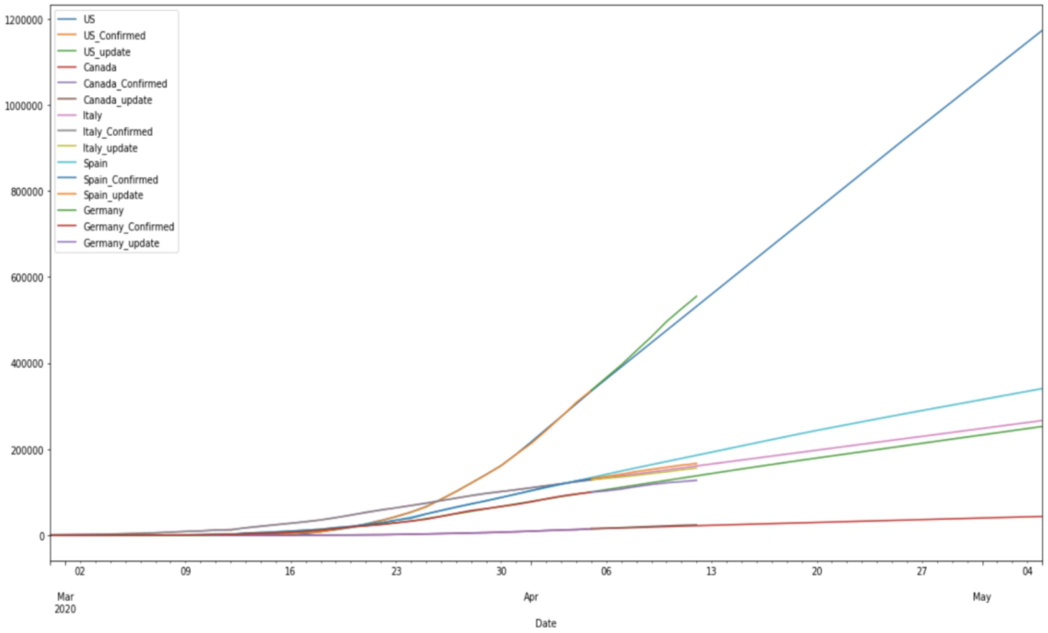


Figure 14: MLP predictions for the five countries

	US	Canada	Italy	Spain	Germany
MSE	2,698,011.7	36,312.6	148,188.2	596,618.7	477,422.8

Figure 15:MLP MSE Comparison

6.3.2 Country Regressor

The model consists of a Linear Regressor and an XGBRegressor[8]. It can predict the cases for any country besides China, the USA, and Canada. Here we run the model by selecting Spain, and the first 15 days in April.

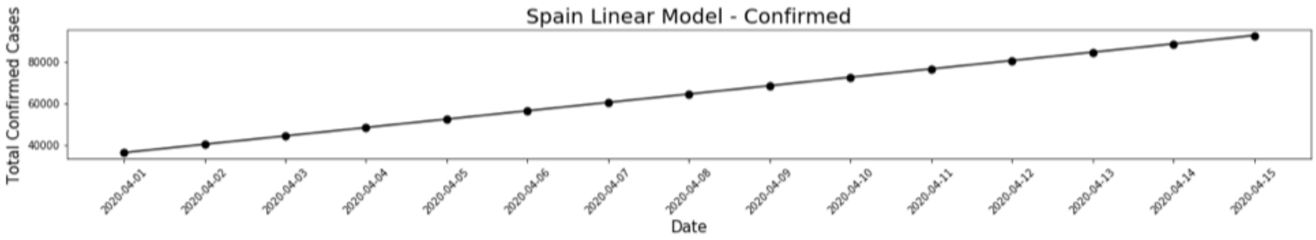


Figure 16: Linear Regressor (confirmed cases prediction) - Spain

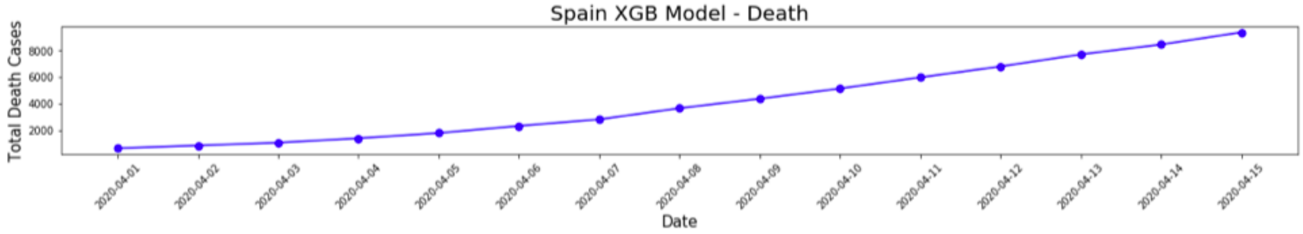
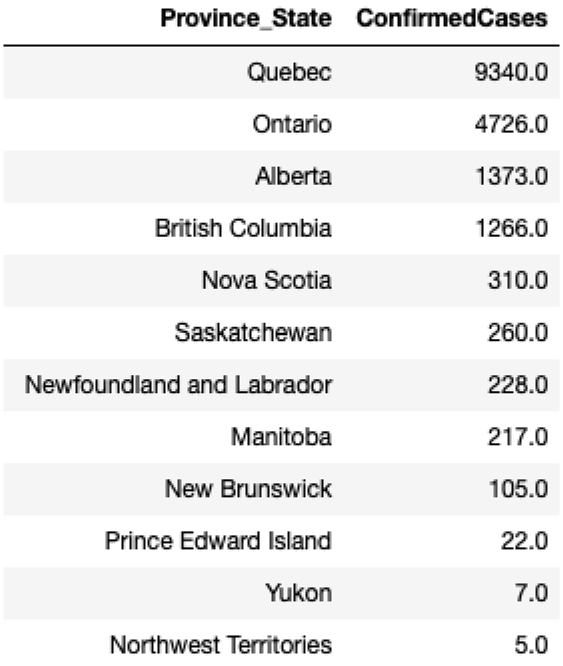


Figure 17: XGBRegressor (death prediction) - Spain

6.3.3 Canadian Province Model

The number of total confirmed cases in Canada by province helps to determine the most impacted province.



Province_State	ConfirmedCases
Quebec	9340.0
Ontario	4726.0
Alberta	1373.0
British Columbia	1266.0
Nova Scotia	310.0
Saskatchewan	260.0
Newfoundland and Labrador	228.0
Manitoba	217.0
New Brunswick	105.0
Prince Edward Island	22.0
Yukon	7.0
Northwest Territories	5.0

Figure 18: Confirmed cases by province (till March 30th)

The user input ‘day’ creates a forecast comparison of bar charts to compare predicted confirmed cases and deaths amongst Canadian provinces.

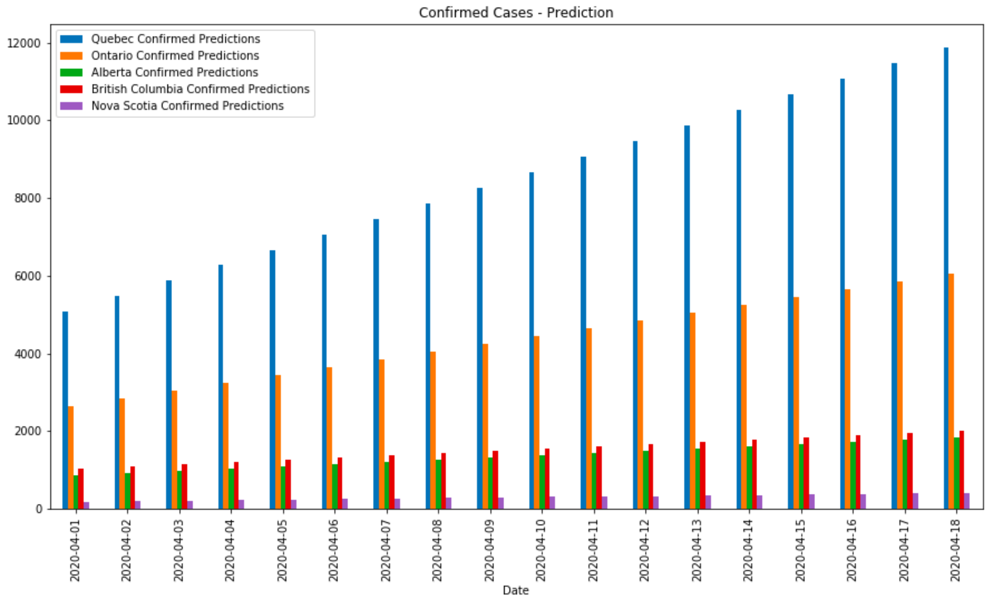


Figure 19: Predicted confirmed cases by province (till April 18th)

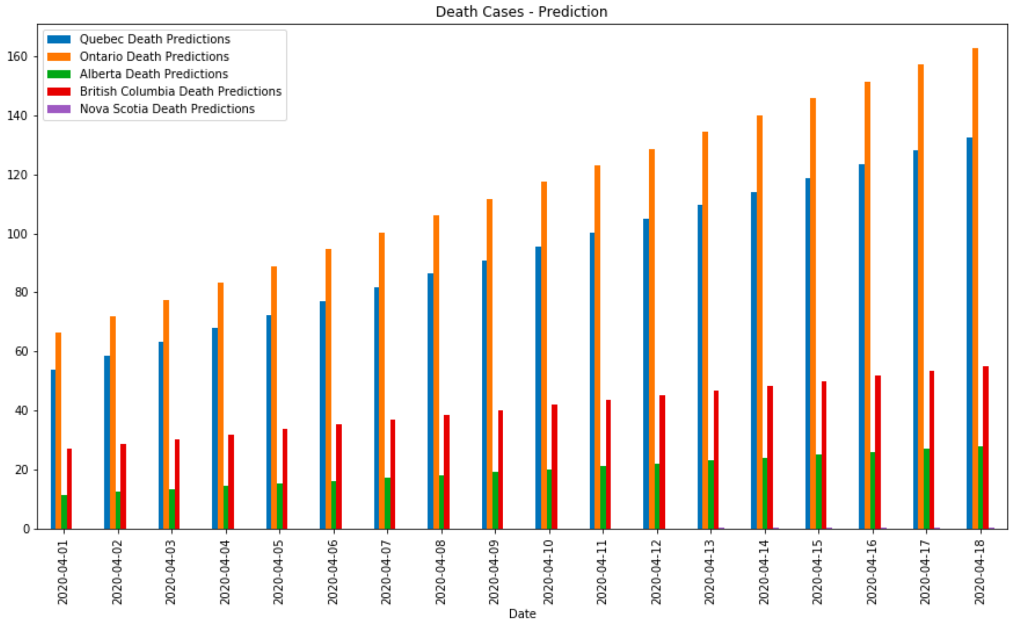


Figure 20: Predicted death cases by province (till April 18th)

6.4 Evaluation

The ARIMA, Prophet, MLP model used the same training data to make predictions. In terms of accuracy, after applying cross-validation, mean square error techniques. The MLP model has the most accurate numbers.

The Country Regressor model consisted of Linear Regression and XGBRegressor models, of which, the XGBRegressor model performs better when compared to the Worldometer statistics data for COVID-19.

One interesting finding is that the Canadian Province model made accurate predictions for the confirmed cases as per the local news. However, the death toll prediction amongst provinces was slightly inaccurate. The reason for that might be due to Quebec having the outbreak much later but the situation becoming severe very quickly.

6.5 Model Limitations

ARIMA - It is unable to make a prediction for more than thirty days and is too acute to fluctuation.

Prophet - The performance is hindered as the data isn't very stationary. It is unable to curve the predicted plot if the data changes at a rapid rate, because the model tends to fit the trend when facing a non-stationary dataset.

Multilayer Perceptron - Performs poorly on rolling predictions. In rolling prediction, we try to apply the dataset with 7 columns containing the confirmed cases from Monday to Sunday, and predict for the next day,and then we put the prediction into the training set and keep on moving forward. The prediction for the next 1-2 days is very accurate.However, the prediction starting from the third day has very large MSE compared with the actual confirmed cases. One possible explanation is that rolling predictions for MSE need more features rather than only use confirmed cases, since there will be other reasons that affect the result.

Country Regressor - The model doesn't perform for countries such as Canada, China, and the USA. These countries also include provincial data with different date-time stamps. In other words, the model functions on countries without provincial data on the dataset.

Canadian Province Model - The predictions are limited to comparison amongst the top five most-affected provinces in March,but can be applied to more provinces if needed.

7. Part B - Government Policy Model

7.1 Data Preparation

The first dataset used is called “containment and mitigation measures”, which contains the description of each country’s government policies for COVID-19 listed by dates. The second dataset is the Canadian government policy dataset, which contains Canadian government policy description, date, and policy levels. For Canadian Policy Dataset, we manually collected ourselves, with information from news outlets.

7.2 Policy Labelling

One issue with the “containment and mitigation measures” dataset is that the government policies descriptions are categorical. We decided to create a procedure that takes categorical government policy as input and outputs a numerical value indicating policy strength.

The policies were categorized into 4 levels manually following the best practices published by CovidActNow, which is built by a group of public health experts[9]:

- Level 0/Limited Action: refers to if no interventions are put in place and the disease is able to spread at its natural rate.
- Level 1/Social Distancing: refers to voluntarily social distancing and advocacy of enhancing public health. For example, restricted travel, ban on events over 50 people, possible school closures and passive monitoring.
- Level 2/Stay Home: refers to voluntarily staying at home quarantine and policies in helping achieve this goal. For example, non-essential businesses shutdown, school close, public relief bill and restricted travel.
- Level 3/Mandatory Quarantine: refers to mandatory community-wised lockdown to force people to stay at home. For example, the full shutdown of businesses, borders close, population-wide mandatory testing.

7.3 Policy duration

To extract more features relevant to this model, we took into consideration ‘Policy Duration’ based on the assumption that policies can only take effect after a certain amount of time.

To integrate this piece of information into our dataset, we added a “duration” column into the “containment and mitigation measures” dataset, representing how many days a country/territory or province/state has been staying in the current level of policy strength.

7.4 Time series to Supervised Learning

To fit our model for training, the data was formatted in a features-target relationship with each row consisting of 21 features and 1 target. 21 features are the number of confirmed cases for the past 7 days, the policy levels for the past 7 days and the duration days for the past 7 days. The target is the number of confirmed cases for today.

7.5 Methodology

For the policy model, a time series approach was applied and then we transferred the model to supervised learning. We first used China’s policy data as the training dataset, since China has the longest outbreak and the government has controlled the pandemic in three months. After that, we use the China model to predict Italy’s situation, and then we combined China and Italy to make predictions for the UK, Canada, and the US.

After the steps from our data pipeline explained above, we have our data in a nice 21 features and 1 target format each row. We decided to use a linear regression model to train the data. This is because a linear regression model is the most simple and also one effective approach to try. The curve of an epidemic is usually exponential first and then flattened towards the end. Linear regression with basis functions can capture the shape of an epidemic curve. Since this pandemic is ongoing, we might consider changing our approach for training the model. For example, if there are more policy related data available in the future, we can consider using a recurrent neural network for training such as LSTM or GRU.

7.6 Results

We did show that government policies will have an impact on the confirmed cases in the future. Supposedly, if we put in policy level 3 for future predictions, the increasing rate for the number of confirmed cases will tend to be flattened. However, this happened when we put in lower policy levels. When we put in higher policy levels, the increasing rate for the number of confirmed cases will not tend to be flattened.

We inspected the coefficients of our linear model, and found that the ones corresponding to policy level are positive while they should be negative (Figure 21 & 22) We think this is due to lack of training data because up to now China is the only country which shows a flattened curve, while other countries show exponential growth curves. And thus the model will tend to lighten the importance of policy levels.

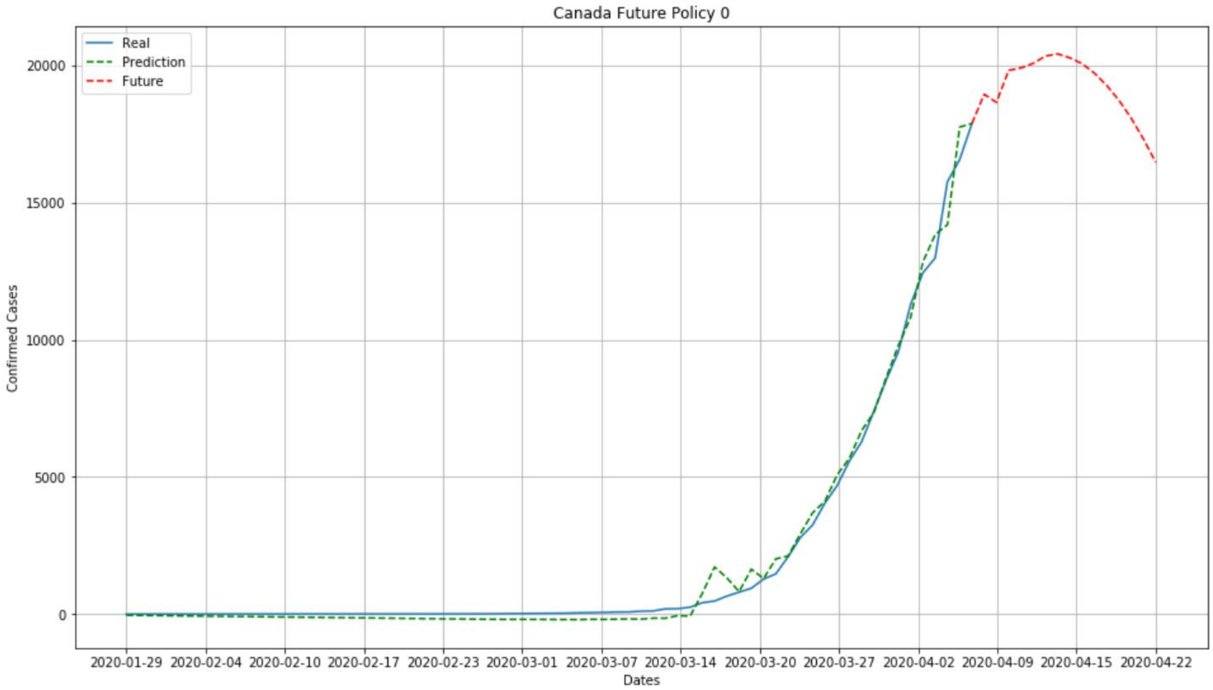


Figure 21: Prediction of the confirmed cases in Canada on policy 0

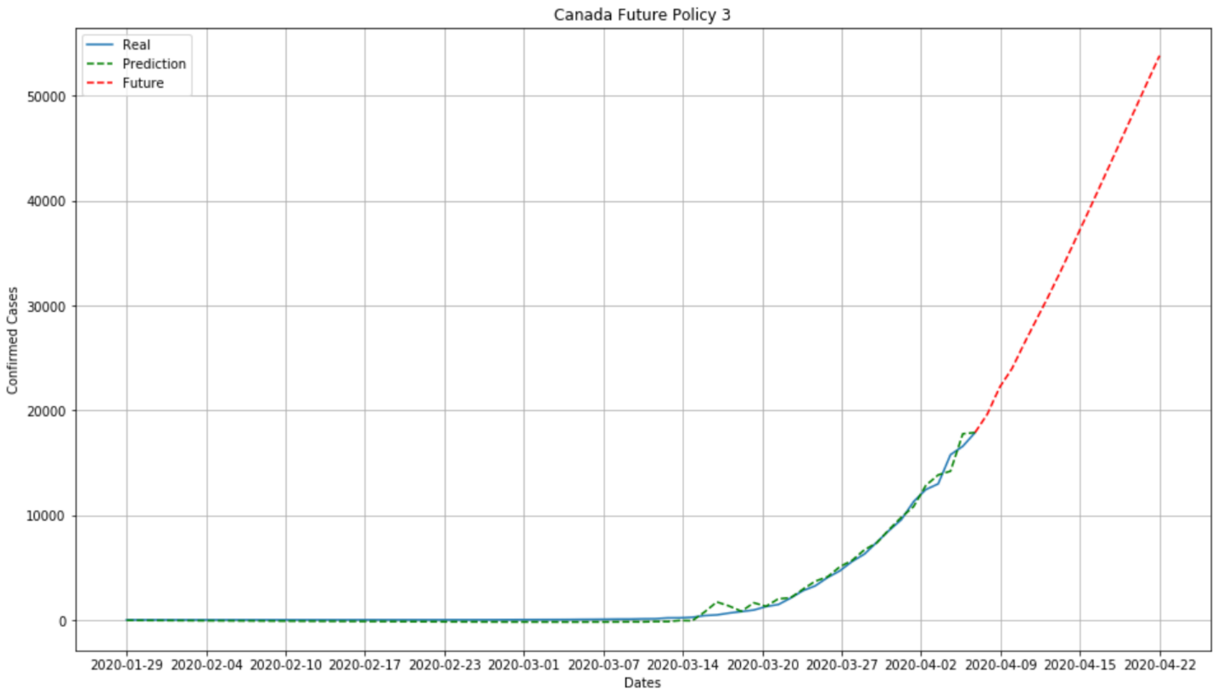


Figure 22: Prediction of the confirmed cases in Canada on policy 3

7.7 Policy Model Evaluation

Since our model is linear regression, the choice of metrics for evaluating the model is mean squared error. We evaluate our model by comparing predictions with actual scenarios. For example, we want to have predictions for the next 7 days starting from today. Today’s data will be fed into the model to get tomorrow’s number, and then tomorrow’s number will be added into the data. These steps are done recursively so that future 7 days’ data is acquired. After this, we just wait 7 days until real world data appears, which is used to compute MSE between predictions and actual values.

We tried to apply our model to make predictions on different provinces in Canada, and we found out our model doesn’t apply to provinces. The reason for that might be because first of all, we trained our model using all country level data and policies. Secondly, countries and provinces’ numbers vary in magnitude, which may also affect the accuracy for predictions.

7.8 Limitations

COVID-19 is a current ongoing crisis that evolves by the day. It is difficult to tell what type of policy has the biggest impact on slowing down the spread of the pandemic. The current model is a preliminary one and is based on the best information we currently have. We do have a couple of thoughts on how this project will move forward in the future and we discussed it in the Lessons Learned Section.

8. Data Product

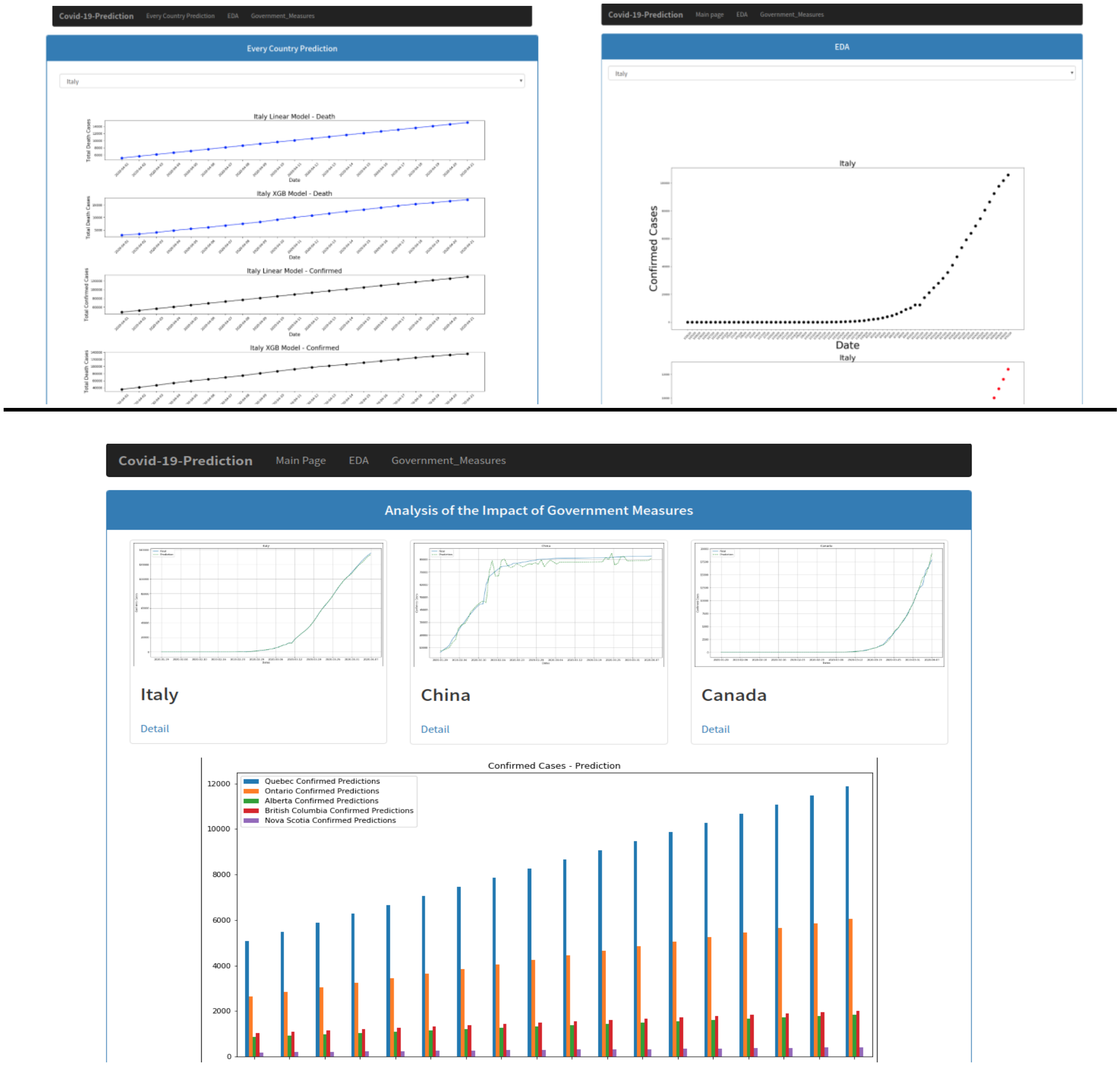


Figure 23: Frontend Display

We made a website by using Python Flask to show our work. It displays the result map for every countries' regressor model map, EDA map, and other analysis results.

9. Lessons Learnt

Predicting an ongoing pandemic is very challenging, as the situation evolves every day.

For comparison of time series models, we learned that, first of all, using less training data in certain situations is a better choice. When developing a correlation between variables such as date vs confirmed cases or date vs fatalities for coronavirus, using data from March rather than using data from January, gave better predictions. Due to the fact that the spread is more exponential in March than in January and February.

For government policy impact prediction, we tried to update our data as best as we can, but we might still lack some information and our model needs to be updated if things change.

In the future, we plan to keep working on our model, for example, One idea we have is to combine the policy model with SIR data such as hospital beds, population, fatality ratio. In this way, we will be able to provide a better picture on how policy level will influence the spread of the pandemic, specifically on the healthcare system.

10. Summary

The mission of our project is to develop a deep understanding of the COVID-19 pandemic spread and forecast its impact in the future.

To accomplish the goal, we broke our tasks into multiple sections. In terms of exploratory data analysis, we created multiple visualizations on maps, plots etc. to understand how the virus is spreading across different countries and causing deaths.

Post EDA work, we focused on these two parts: a comparison of current time series prediction models on COVID-19 and government policy impact on this pandemic. Most outputs were made available in a front-end for easy access as well.

Firstly, we tested multiple time-series machine learning models to forecast the pandemic and compared how each model performs and how our predictions ranked up against real world data. Our models include ARIMA, MLP, Prophet, Linear+XGBRegressor, and a Canadian Provincial model.

We also developed a Government policy model which used a dataset that was built by ourselves. We collected Canada policy data from news outlets and manually labelled them into different levels using domain knowledge. We built a linear regression model which shows that government policies have an impact on the epidemic in terms of “flattening the curve”, however, more data would be required to improve model accuracy.

References

1. Dataset: 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by Johns Hopkins CSSE. https://github.com/CSSEGISandData/COVID-19
2. Dataset: Kaggle Novel Corona Virus 2019 Dataset. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset
3. Dataset: Kaggle COVID-19 containment and mitigation measures. https://www.kaggle.com/paultimothymooney/covid19-containment-and-mitigation-measures
4. Worldometer COVID-19 information. https://www.worldometers.info/coronavirus/
5. Autoregressive integrated moving average. (2020, April 13). Retrieved from https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average
6. ARMA Models and the Box Jenkins Methodology. Journal of Forecasting, 147–163. Author: Spyros, M. S., & Hibon, M. (1997).
7. Prophet is a forecasting procedure implemented in R and Python. It is fast and provides completely automated forecasts that can be tuned by hand by data scientists and analysts. (n.d.). Retrieved from https://facebook.github.io/prophet/
8. XGBoost information. https://towardsdatascience.com/a-beginners-guide-to-xgboost-87f5d4c30ed7
9. CovidActNow: https://covidactnow.org/