BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
K. K. BIRLA GOA CAMPUS

BITS F464 - MACHINE LEARNING

MAJOR PROJECT

Predicting and analyzing the spread of Covid-19 in Canada

By

Shahsank Kumar	2018A4PS0069G
Aditya Bhandari	2018A7PS0805G
Shobhit Mehta	2018A8PS0417G
Nilesh Raghuvanshi	2018A8PS0415G

CONTENTS

ABSTRACT

The highly infectious coronavirus disease (COVID-19) was first detected in Wuhan, China in December 2019 and subsequently spread to 212 countries, infecting millions of people. Numerous studies have been conducted to forecast the spread of Coronavirus. We aim to complement the available work by examining the Canadian COVID-19 data for the period March 1, 2020, to April 21, 2021. Use a model to forecast the spread of Covid-19.

INTRODUCTION

Canada reported its first covid case on January 25, 2020. The total number of confirmed infections in Canada today has crossed 1.2 million and is growing steadily. This pandemic continues to challenge medical systems worldwide in many aspects, including sharp increases in demands for hospital beds and critical shortages in medical equipment. Many healthcare workers have also been infected. Thus, the capacity for immediate clinical decisions and effective usage of healthcare resources is crucial. Our forecast model should therefore help officials in understanding how the present dynamics could affect future cases. Accordingly, they can adopt necessary measures to mitigate the possible ramifications. The model also attempts to highlight how vaccination can help curb Covid's spread.

1.DATA COLLECTION

For our machine learning model, we scraped/downloaded three types of data. In the mode, we have used data for the period: 1st Feb 2020-21st April 2021. Data was collected individually for each Canadian province.

1.1 Covid-19 data

To build a machine learning model that forecasts covid-19 data, we first collected all the available statistics related to the disease. Besides case and death counts, we aggregated data corresponding to active cases, recoveries, vaccination counts, number of tests carried, number of hospitalizations, and number of patients admitted to ICU. All data fields were available on a daily and cumulative basis.

1.2 Mobility data

Mobility data quantify public engagement in various activities such as driving, walking, recreation, transit, work, etc. It was obtained from two sources. Google community reports provided mobility data only for the year 2020, while apple mobility data were available up to 21st April 2021. Both Google and Apple consider public engagement on a certain date as the baseline. Accordingly, numbers were assigned to future dates comparing the engagement on that day to the baseline.

We also scraped data on international travelers and vehicles from the United States entering the four provinces. However, this data was available monthly. So, we assumed the daily data followed a normal distribution with a mean equal to the month's total divided by the number of days and standard deviation equal to 10% of the mean.

1.3 Google trends

Using the Pytrends library in python, we collected the daily search volume on keywords like Coronavirus, Vaccine, and Lockdown. We also included comparable queries like Coronavirus Quebec, Vaccine Ontario, etc.

1.4 Temperature

Following initial claims that temperature could affect covid spread, we decided to include daily temperature to check for the argued association. International Airports, which are weather stations, were selected to collect the temperatures for each province. It didn't show sufficient correlation with covid spread.

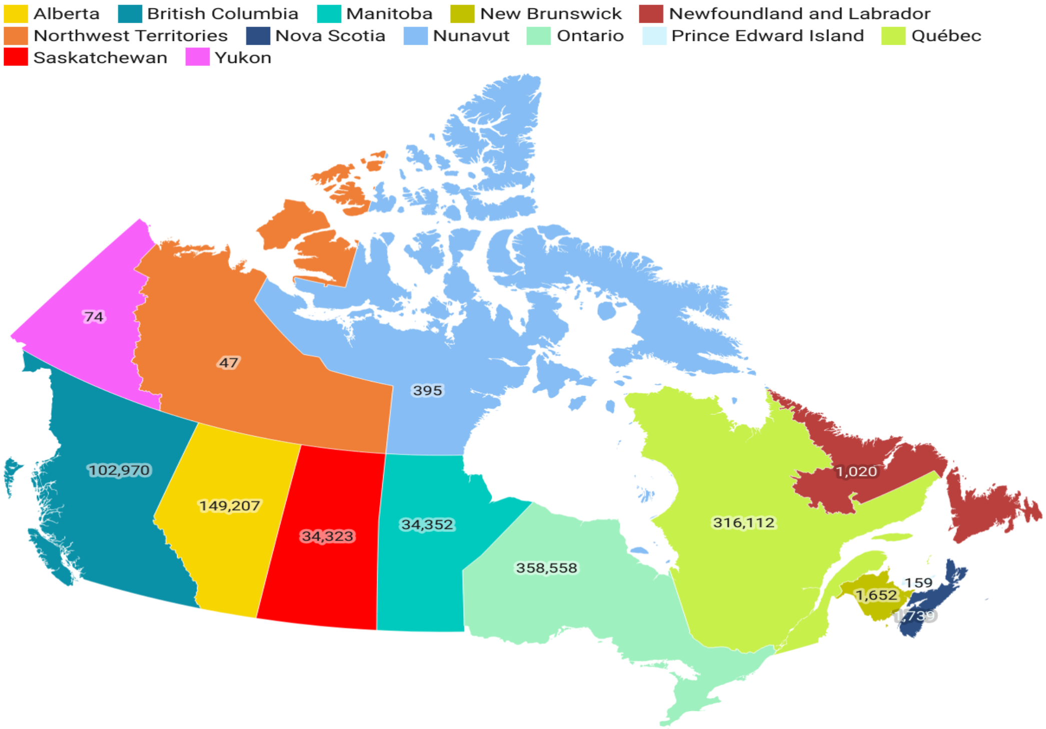
2.PRE-PROCESSING

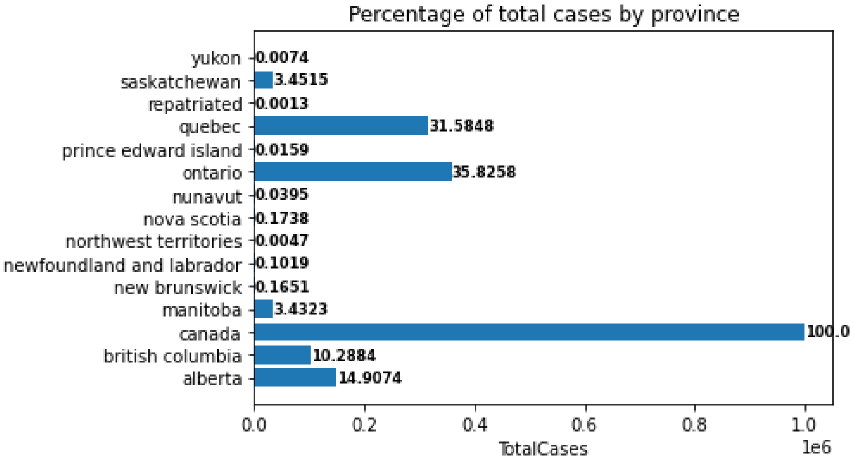
2.1 DATA CLEANING

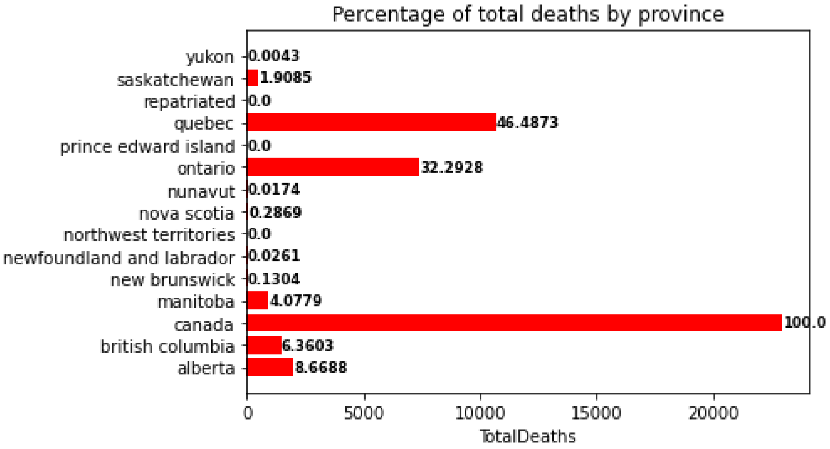
Similar to any other machine learning project, we first started by cleaning and organizing the collected data. This included string to float conversions, unifying the date format across collected data, and treating null values.

2.2 INITIAL EXPLORATION

We searched the collected data for some useful information. Around 93% of the total cases and deaths were reported from four Canadian provinces: Ontario, Quebec, Alberta and British Columbia.







Consequently, we decided to focus on only these four provinces. Also, the vaccination drives began in all provinces on/after 9th January 2021. Hence, we chose to make forecasts for two periods:

1)Pre vaccination: 1st March 2020- 31st December 2020

2)Post vaccination: 1st January 2021-21st April 2021 (Vaccinations began from 9th January, but we included the previous 8 days to simplify our task)

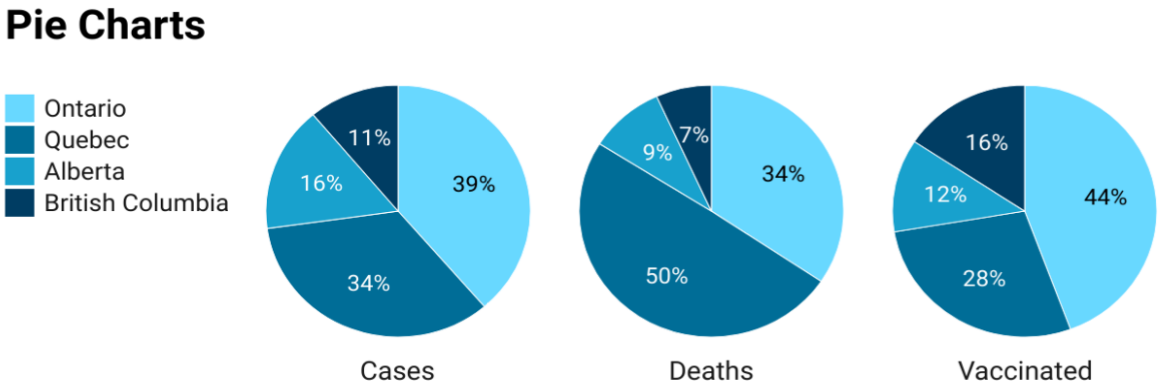
Next, we segregated data from each period into training and testing sets.

1)Pre vaccination

a. Training: 1st March 2020-31st October 2020
b. Testing: 1st November 2020-31st December 2020

2)Post vaccination

a. Training: 1st January 2021-31st March 2021
b. Testing: 1st April 2021- 21st April 2021



Note, the aim was to build a model that predicts covid cases on future dates. So, we did not use randomised splits for creating test and train datasets. To avoid extrapolation errors, we modified the dependent variable.

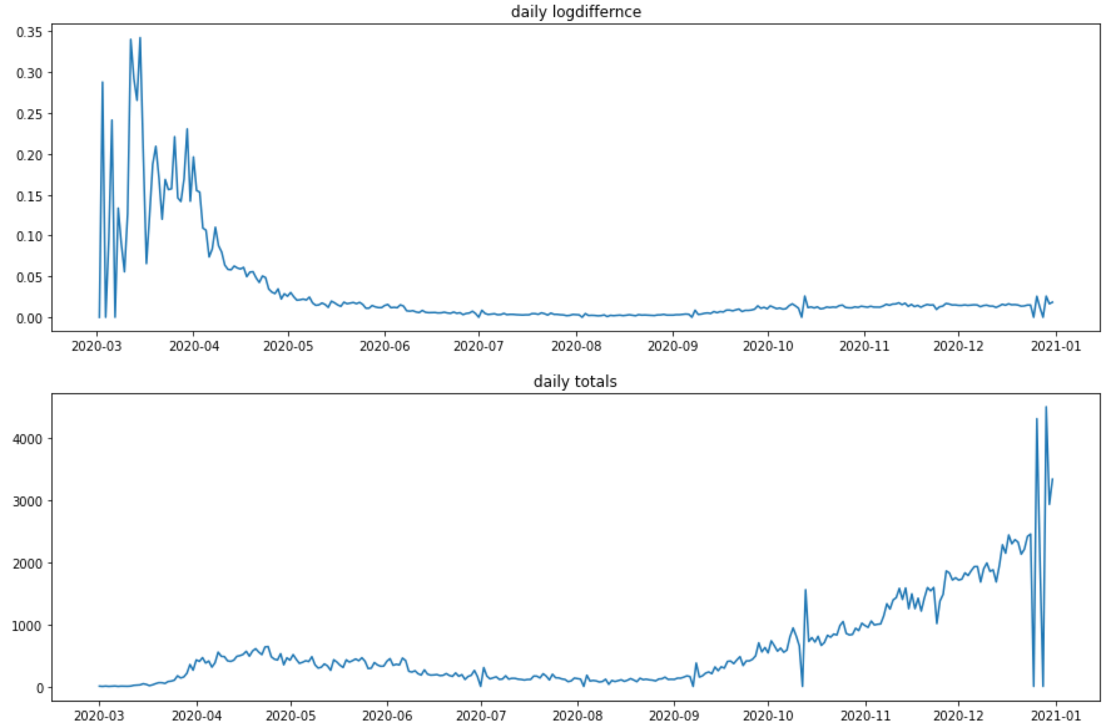
2.3 DATA ORGANIZATION

After narrowing down to four provinces, our next step was to determine the variables for forecasting models. We did it separately for each province.

2.3.1 Dependent variable

In our model, we have forecasted the number of covid cases. The obvious choices for the dependent variable were cumulative total case count(CTC) and daily case count(DCC). But there were a few issues. Unlike CTC, most other data columns had a fluctuating nature. So, the former showed weak correlations with most variables. Contrarily, it was strongly correlated, Pearson coefficient> 0.95, with other cumulative data variables like total recovery. But that was because both have an increasing nature. Thus, we believed predicting CTC using other cumulative data could result in overfitting.

We overlooked DCC because of the selection criteria for training and testing data. The test data, November-December 2020, had daily case counts that were beyond the range of training data. The same was also true for the 2nd period where we used data till 31st March for training purposes. Hence to avoid extrapolation issues, we decided to use the daily log difference of cumulative total cases as the dependent variable. So, if X and Y are the cumulative case totals on days d-1 and d, the log difference for d is: log(Y) - log(X)



(Ontario data)

2.3.2Candidates for explanatory variables

As discussed previously, we had four types of data. We chose appropriate features from them by applying time lag.

2.3.2.1 Time lag

We established the belief that covid symptoms generally appear 1-2 weeks after contracting the disease, as a prior. Using that, we ran tests to determine which day's data had the most impact on the dependent variable. For example, consider the number of daily recoveries (R) as an explanatory variable and the log difference (X) on a day d. We aimed to determine which previous day's R had the most impact on X.

$$R_d = (A)X_{d-t}$$

Here most impact refers to the highest Pearson correlation. A is the coefficient for a linear model(used arbitrarily to explain the above example), and t is the time lag, difference in number of days. The adequate time lags for mobility,google search and covid data(besides case and death counts) were calculated independently. Furthermore, the lags were found separately for each province and each period. No time lag was applied to temperature.

2.3.2.2 Google trends keywords

For each province, we tested how the search volume for certain keywords was associated with the daily log difference of cases. For instance, for Ontario, the keywords were: Coronavirus, Lockdown, and Coronavirus Ontario. We used the first two words for each province and modified the province name in the third keyword.
Next, for the post-vaccination period, we changed the keywords to Coronavirus, Vaccine, and Vaccine + ‘province name’. We found time lags for these keywords as well. Finally, a single keyword was chosen, considering Pearson correlation, for each province.

2.3.2.3 Final datasets

After finding the time lags for all types of variables, we finalised eight datasets, a pre and post vaccination dataset for each province. All these datasets included the following:

1. Mobility columns: transit,walking,driving,recreation,essentials,parks,station_transit workplace,residential,international travel(named same as the province) and vehicles entering from the United States (named same as the province +’-us’ ex. ontario-us)
2. Covid columns: Total hospitalized, Daily tested,Total ICU,Total recovered, Total active
3. Google search column: This included a separate single keyword/search query for each province
4. Daily mean temperature

Again, the time lags were different for the four types of variables(0 for temperature), and were also different across the eight datasets. We have provided the time lags for each province in the results section.

2.4 FINAL VARIABLE SELECTION

Across the eight datasets, we had 18 different candidates for forecast model features. Explanatory variables were selected separately for each dataset.

2.4.1 Scatter plots

We made scatter plots for all 18 variables against the dependent variable. From these plots, we ascertained more about the relationship between candidate features and daily log difference. Some variables presented a hyperbolic or log relationship and were accordingly modified. This was done to improve the Pearson correlation.

2.4.2 Correlation heatmaps

Next, we used correlation heatmaps to select the best features. We made sure not to include features that were highly correlated amongst themselves. So, if two features had a Pearson coefficient greater than 0.6, we included only one of them. The decision generally depended on which of the two was more correlated to log difference. But in some cases, we altered.

3.FORECASTING

We applied three algorithms, Linear regression and Random forest regressor on each of the eight datasets. Their performance was judged by the ability to predict daily log difference using the passed explanatory variables. We considered two metrics: mean absolute error(MAE)mean square error(MSE). As discussed, forecasts were carried independently for each province and for the two periods. While the model choice remained same and the dependent variable was same, we modified the explanatory variables in each case.

3.1 Linear regression

We have used multivariate regression to create a baseline model and compared it to some advanced algorithms.

3.2 Random forest regressor :- Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble . Each individual tree in the random forest spits out a class prediction and the class with the most votes is returned as the model’s prediction. For our project, we used a base random forest model, and a hyper parameterized random forest model using random search which performs 45 iterations to get the best parameters.

3.3 Vaccine influence

With the beginning of vaccination, explanatory variables also change. That is, mobility increases, google searches change, and most importantly covid related data also alters. We had 111 days of data (second period) describing the trends during the vaccination drive. To quantify the effect of vaccination, we first needed to forecast the circumstance in its absence. So, we assumed that had there been no vaccine, the resulting situation would be similar to the 111 days before the start of vaccination. Essentially, the explanatory variables describing log difference 111 days before vaccination would have remained constant. With that assumption, we predicted the cumulative cases for the next 111 days from 31st December 2020 using the two versions of random forest regressor. Note, only the model built for 1st period, pre-vaccination, was used for this forecast. Finally, we compared the predictions with the actual data available for the vaccination period.

Besides forecasting, we also visualized the data from the vaccination period to draw some qualitative inferences.

1) Compared the death to case ratio: the ratio of total deaths to total cases for the two periods
2) Examined the effect of daily vaccinations on daily cases using plots

4.RESULTS

4.1ONTARIO

4.1.1Pre vaccination

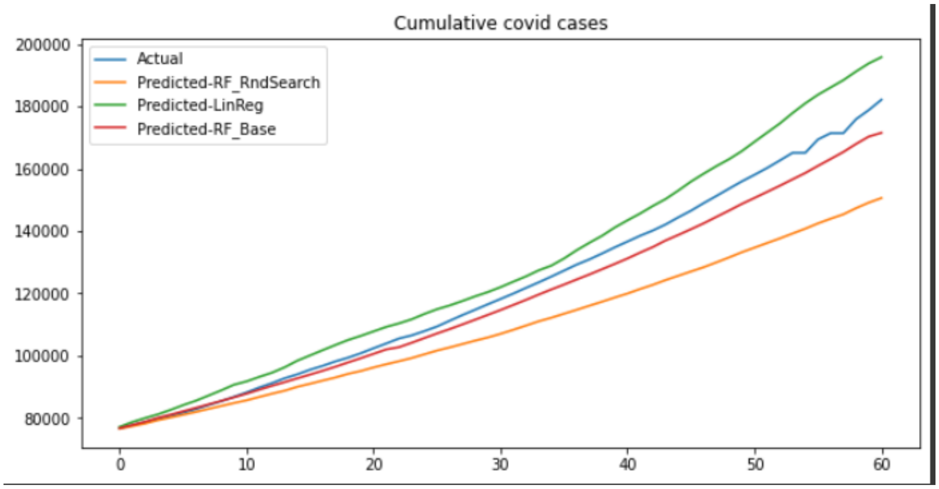
Time lags:

1. Mobility data: 9 days
2. Covid data: 10 days
3. Google search: 10 days

Features used: Total Recovered, Coronavirus(google trends)

Models Performance

Metric	Performance	Model
MAE	0.003680728636316352	Linear Regression
MAE	0.002490701224836064	Random Forest Base
MAE	0.003885157488274561	RF Random Search
MSE	2.347388421111153e-05	Linear Regression
MSE	1.6026484381947406e-05	Random Forest Base
MSE	2.342727560064374e-05	RF Random Search



(Forecast for 1st Nov 2020-31st Dec 2020)

4.1.2Post vaccination

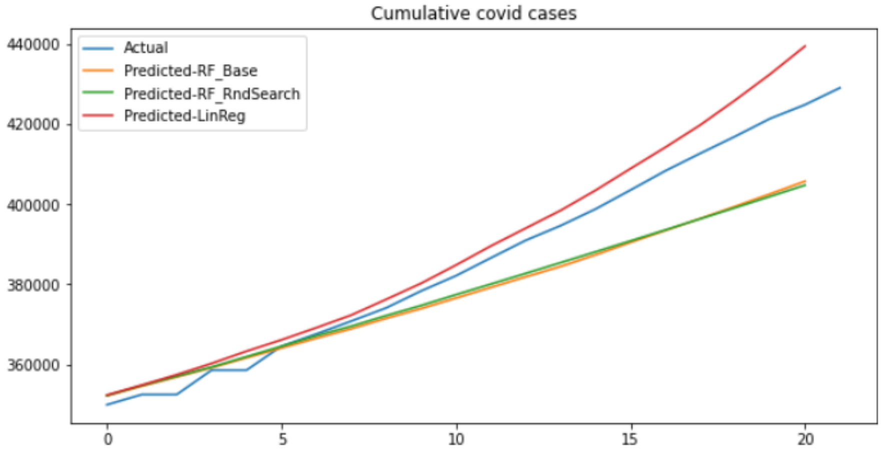
Time lags:

4. Mobility data: 8 days
5. Covid data: 6 days
6. Google search: 3 days

Features used: Total Recovered, Total Active, International travellers, Ontario Vaccine(google trends)

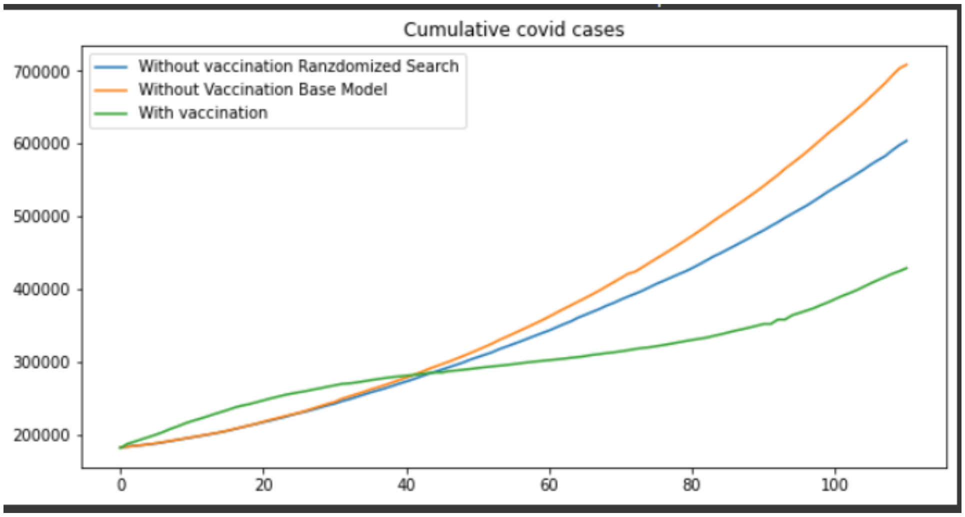
Models Performance

Metric	Performance	Model
MAE	0.0030923440678320337	Linear Regression
MAE	0.003914621600651795	Random Forest Base
MAE	0.004078395444535839	RF Random Search
MSE	0.0030923440678320337	Linear Regression
MSE	2.17930667384124e-05	Random Forest Base
MSE	2.288925992414512e-05	RF Random Search

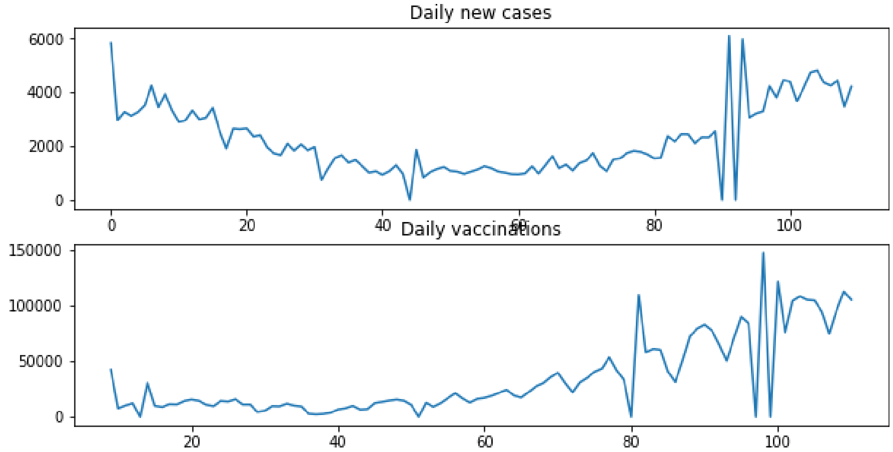


(*Forecast for 1st April 2021-21st April 2021*)

4.1.3Influence of vaccines



(*Forecast of covid spread in absence of vaccine*)



Pearson correlation between daily new cases and daily vaccinations(From data)=0.57187926
Difference in cumulative cases on 21st April 2021=279340 (Base model)
Difference in cumulative cases on 21st April 2021=174878 (Random search model)
Death/Case before vaccination=0.024870432185523542
Death/Case after vaccination= 0.013196255324662703
Percentage reduction in Death/Case= 46.93998388840257

4.2QUEBEC

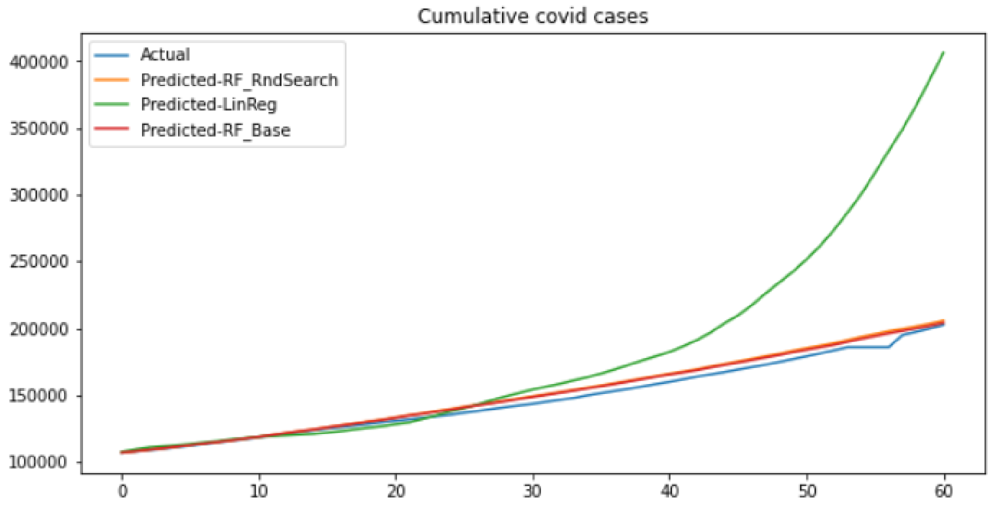
4.2.1Pre vaccination

Time lags:

7. Mobility data: 9 days
8. Covid data: 6 days
9. Google search: 4 days

Features used: Total Recovered, Coronavirus(google trends)
Models Performance

Metric	Performance	Model
MAE	0.012912527032349894	Linear Regression
MAE	0.0024763091240612802	Random Forest Base
MAE	0.0025754590770779486	RF Random Search
MSE	0.00034519452856731085	Linear Regression
MSE	3.2590923458284435e-05	Random Forest Base
MSE	3.3661650679534553e-05	RF Random Search



(*Forecasting for 1st Nov 2020-31st Dec 2020*)

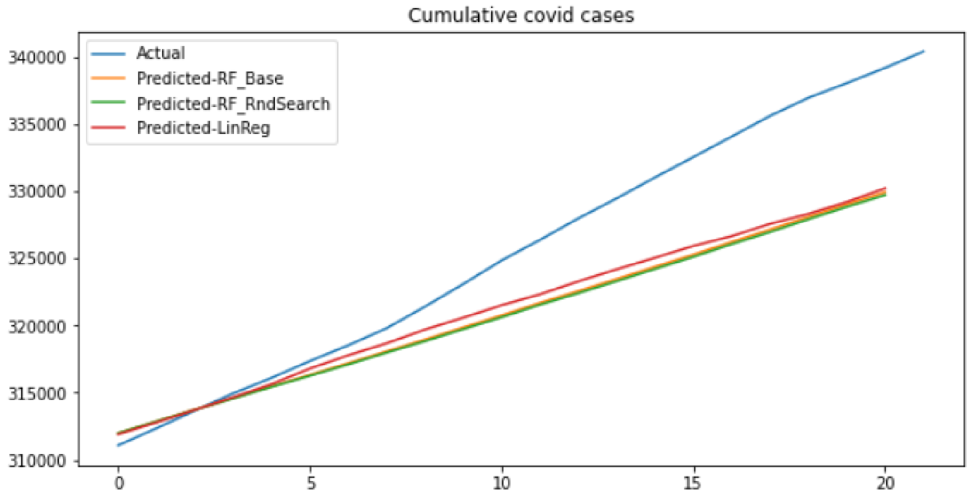
4.2.2Post vaccination

Time lags:

10. Mobility data: 9 days
11. Covid data: 6 days
12. Google search: 3 days

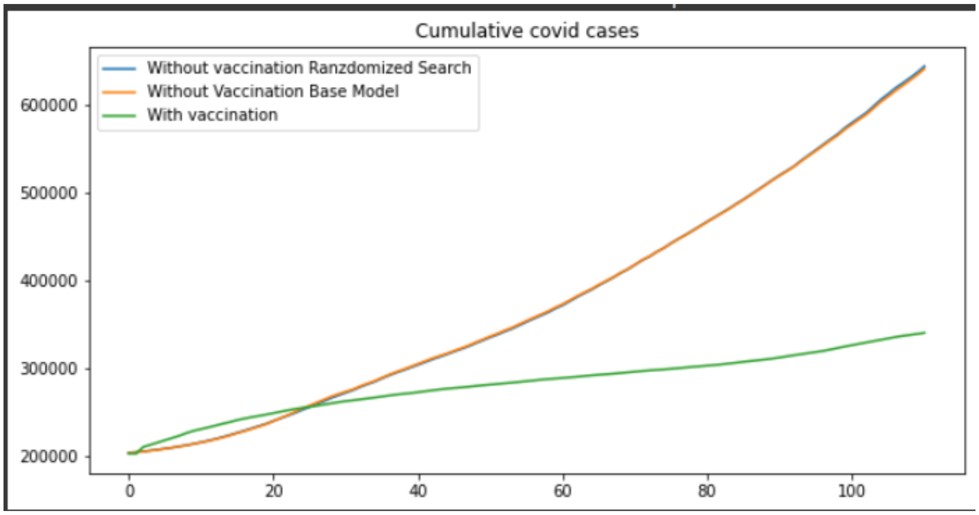
Features used: Total vaccinated, walking(mobility), international traveller
Models Performance

Metric	Performance	Model
MAE	0.001444354999090745	Linear Regression
MAE	0.0014851022432347896	Random Forest
MAE	0.0015159688082583005	RF Random Search
MSE	2.5662029390500557e-06	Linear Regression
MSE	2.5546383546179453e-06	Random Forest
MSE	2.6452822202306706e-06	RF Random Search

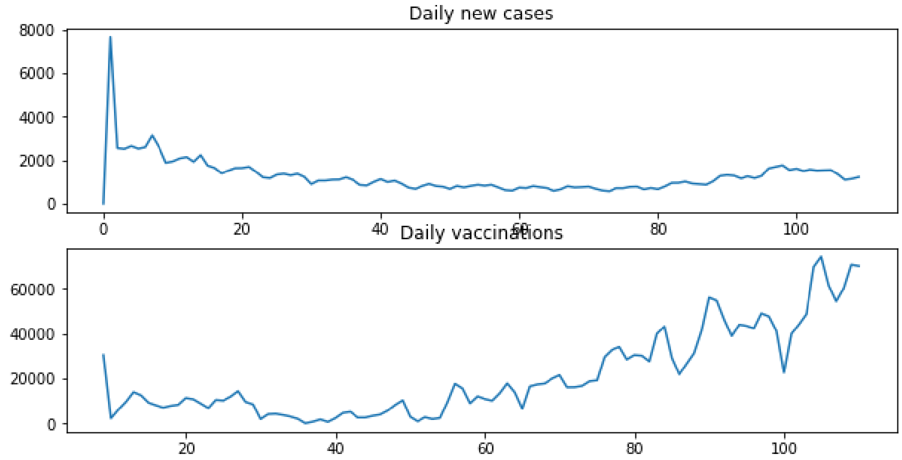


(Forecast for 1st April 2021-21st April 2021)

4.2.3 Influence of vaccination



(*Forecast of covid spread in absence of vaccine*)



Pearson correlation between daily new cases and daily vaccinations(From data)=0.18980577
Difference in cumulative cases on 21st April 2021=301058 (Base model)
Difference in cumulative cases on 21st April 2021=303980 (Random search model)
Death/Case before vaccination=0.04059415712593762
Death/Case after vaccination= 0.018961061587154097
Percentage reduction in Death/Case= 53.29115584705925

4.3 ALBERTA

4.3.1 Pre vaccination

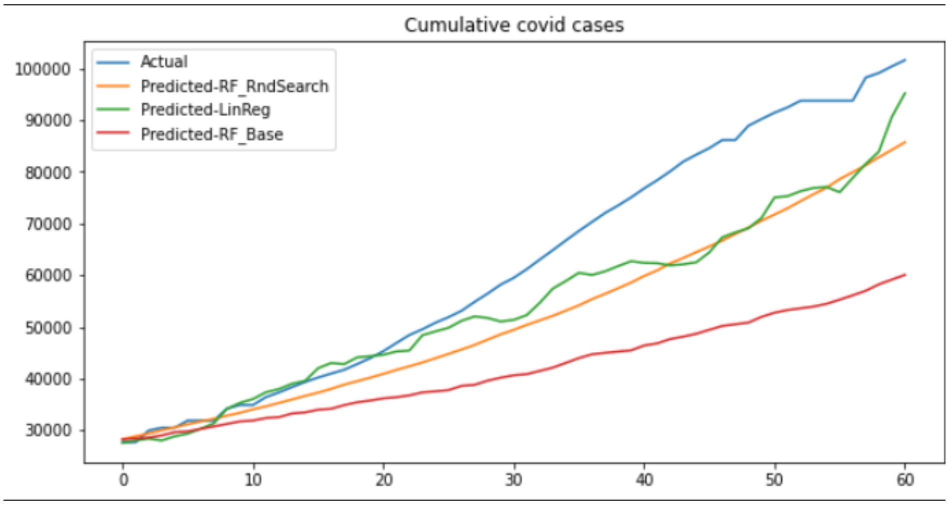
Time lags:

13. Mobility data: 9 days
14. Covid data: 12 days
15. Google search: 5 days

Features used: Total Recovered, Coronavirus(google trends)

Models Performance

Metric	Performance	Model
MAE	0.01956789412526039	Linear Regression
MAE	0.014090245103278687	Random Forest Base
MAE	0.011067778909327938	RF Random Search
MSE	0.0005823120878165411	Linear Regression
MSE	0.0003524772325261248	Random Forest Base
MSE	0.00024163076442523817	RF Random Search



(*Forecast for 1st Nov 2020-31st Dec 2020)*

4.2.2Post vaccination

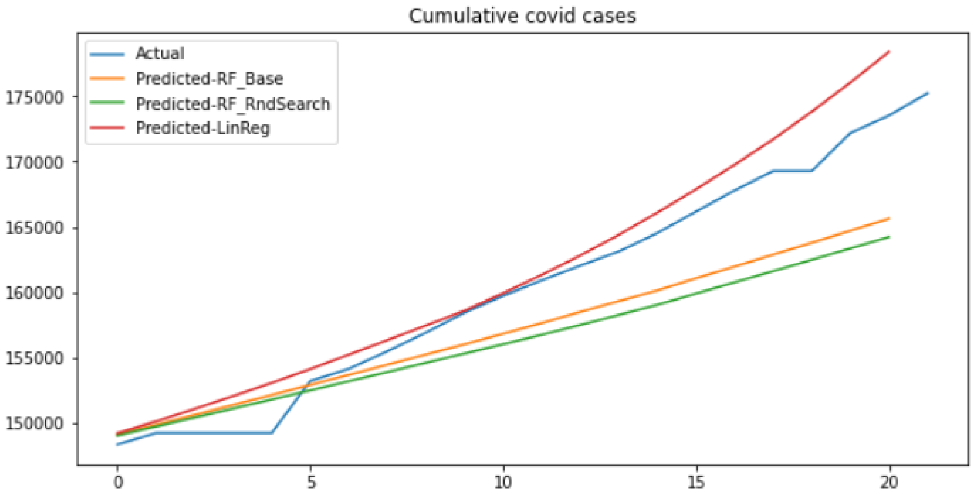
Time lags:

16. Mobility data: 13 days
17. Covid data: 6 days
18. Google search: 7 days

Features used: Total vaccinated, Total active

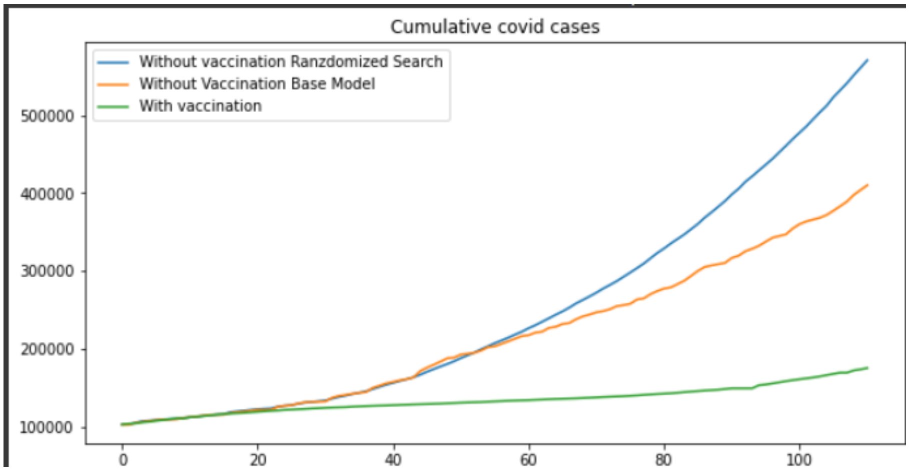
Models Performance

Metric	Performance	Model
MAE	0.0038699178127267917	Linear Regression
MAE	0.004656282828702952	Random Forest Base
MAE	0.004907033028919095	RF Random Search
MSE	3.481986485184669e-05	Linear Regression
MSE	4.0383430404397306e-05	Random Forest Base
MSE	4.259267169361823e-05	RF Random Search

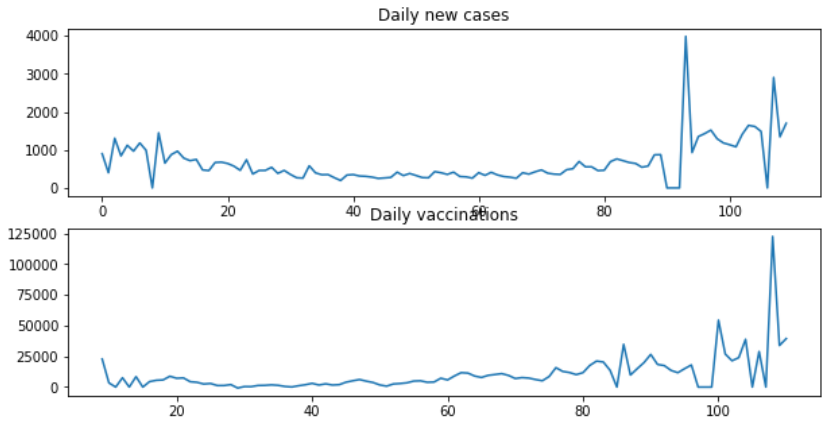


(Forecast for 1st April 2021-21st April 2021)

4.2.3Influence of vaccination



(*Forecast of covid spread in absence of vaccine*)



Pearson correlation between daily new cases and daily vaccinations(From data)=0.51233228
Difference in cumulative cases on 21st April 2021=234738
Difference in cumulative cases on 21st April 2021=395193
Death/Case before vaccination= 0.010292439091588931
Death/Case after vaccination=0.013858537799784238
Percentage reduction in Death/Case= -34.64775138780809

4.4 BRITISH COLUMBIA

4.4.1 Pre vaccination

Time lags:

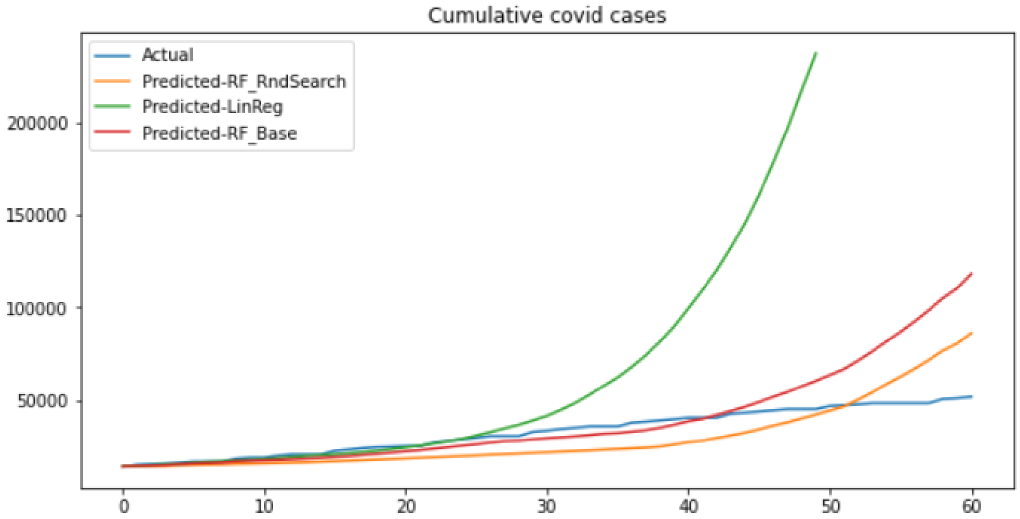
19. Mobility data: 7 days
20. Covid data: 9 days
21. Google search: 14 days

Features used: Transit(mobility), Total Hospitalized

Models Performance

Metric	Performance	Model
MAE	0.05234602360827308	Linear Regression
MAE	0.02670392608978142	Random Forest Base
MAE	0.028338689771077848	RF Random Search
MSE	0.0038737953599077613	Linear Regression

MSE	0.0011148439096847643	Random Forest Base
MSE	0.0012924649124571525	RF Random Search



(Forecast for 1st Nov 2020-31st Dec 2020)

4.4.2Post vaccination

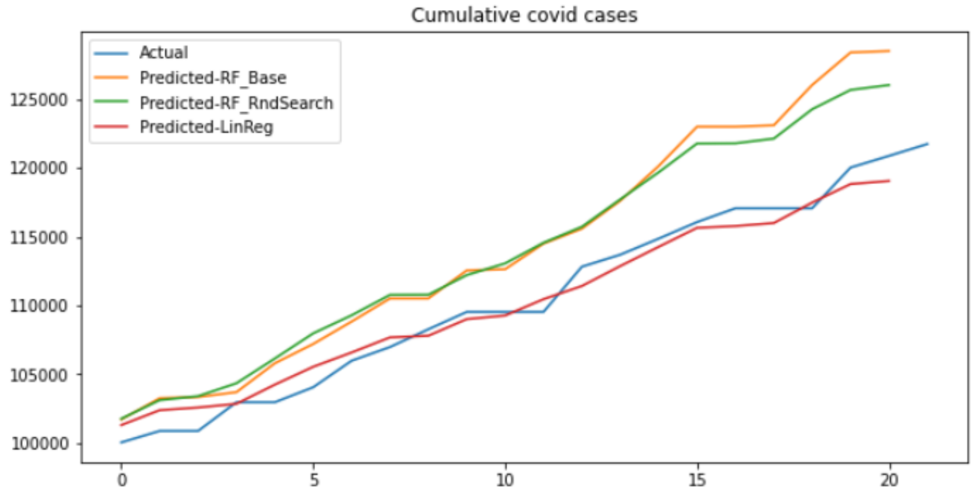
Time lags:

22. Mobility data: 8 days
23. Covid data: 7 days
24. Google search: 6 days

Features used: Daily vaccinated, Coronavirus(google trends)

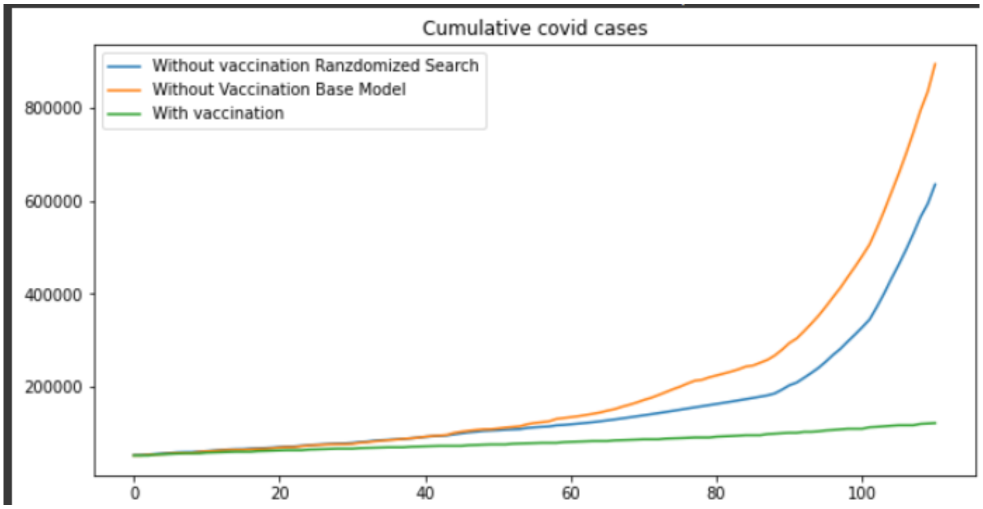
Models Performance

Metric	Performance	Model
MAE	0.005891104912035994	Linear Regression
MAE	0.008039318064896824	Random Forest Base
MAE	0.007246043684432591	RF Random Search
MSE	6.411231646161023e-05	Linear Regression
MSE	9.855947906712274e-05	Random Forest Base
MSE	7.576485512651097e-05	RF Random Search

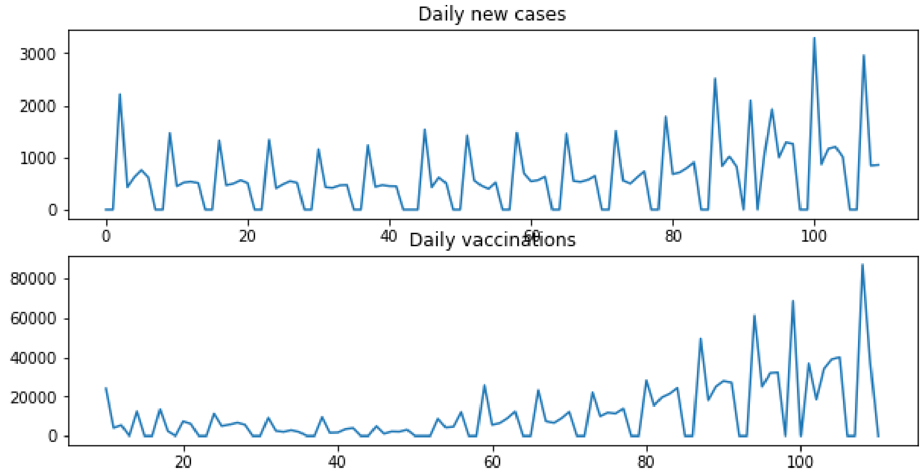


(Forecast for 1st April 2021-21st April 2021)

4.4.3 Influence of vaccination



(*Forecast of covid spread in absence of vaccine*)



Pearson correlation between daily new cases and daily vaccinations(From data)= 0.61517419
Difference in cumulative cases on 21st April 2021=772585
Difference in cumulative cases on 21st April 2021=513550
Death/Case before vaccination =0.017332922934862067
Death/Case after vaccination 0.009245853700491679
Percentage reduction in Death/Case: 46.65727335638641

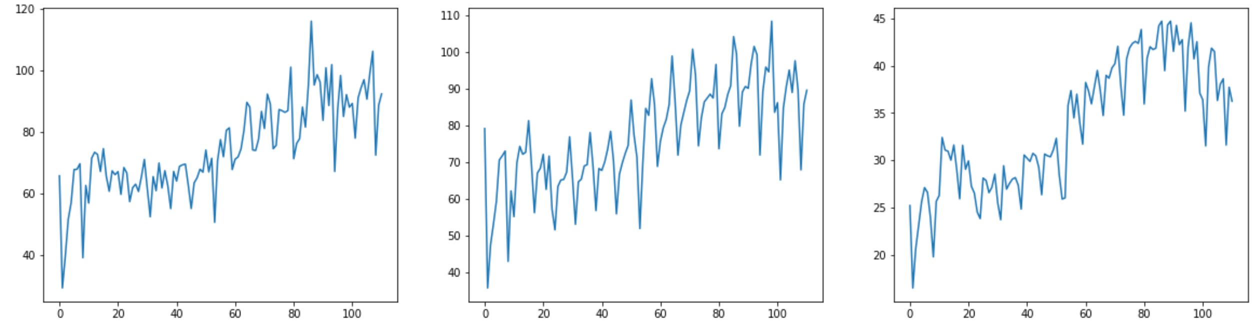
4.5 Evaluation of models

MAE and MSE scores for all models were close to zero. However, this was also because the dependent variable, daily log difference, is small in magnitude. Overall, the performance of models varied across all the datasets. For the pre-vaccination period, random forest regressors proved better than a linear regression model. However, for the post-vaccination period, the performance of linear regression was comparable/better than that of random forest regressors. We attribute this to the fact that less data is available for the vaccination period. The forecast for covid spread in absence of vaccination seemed reliable for all provinces barring British Columbia where the prediction was likely excessively high. Our forecasts may also be inaccurate due to the earlier discussed assumption. It's unlikely that explanatory variables would have followed a similar pattern in absence of vaccination. Moreover, these variables are also linked to covid cases. That is, depending on how the number of confirmed cases and deaths change, explanatory variables also alter. Hence, we believe a model that considers the dynamics between all variables would work better.

4.6 Qualitative evaluation of vaccine drive

According to the our base version of random forest regressor the cumulative covid cases on 21st April 2021 would have been 13,35,183 more than the actual in absence of vaccination. Note, this forecast is for just four Canadian provinces. Moreover, the deaths/case ratio for all provinces apart from Alberta reduced by almost 50% during the vaccination period. Given that senior citizens were prioritized for vaccination, the reduction aligns with the belief that Coronavirus is fatal amongst old people.

We also observed something interesting while visualizing the vaccination data. From the plots of daily new cases vs daily vaccinations provided for each province, it’s discernible that the daily new cases don’t necessarily decrease with rise in vaccination. All provinces barring Quebec experienced an increase in daily cases towards March end. So, albeit vaccination significantly curbed the number of cases when compared to the pre-vaccination period, it’s influence in the period 1st Jan-21st April 2021 seemed a bit inexplicable. A reason for this could be that with the onset of vaccination, citizens lost their guard and started engaging in public activities which resulted in a minor growth in daily new cases. This is also evident from the below plots for mobility from Ontario during the vaccination period.



5.CONCLUSION

Due to the factors discussed in 4.6, predicting the spread of Covid-19 is a cumbersome task. We also used PCA for feature extraction, but the results didn’t vary much. PCA, we believe, would have boosted model performance with more data. By more data, we refer to advanced and more detailed data metrics like reproduction number of the virus, age/sex/medical history details of individuals succmbing to covid-19,etc. Moreover, we considered each province as a single region. A better model would consider the cities,towns and villages independently in each province. Scraping all this data was not possible because government websites don’t provide it publicly. Even if they do, the data is not available for all provinces. Nevertheless, we believe that our model provides an idea of how factors like mobility and google trends can be used to predict the spread of Covid-19. We also reckon that the model provides a rudimentary image of how vaccination curbs covid spread.

REFERENCES AND DATA SOURCES

1.https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19/epidemiological-economic-research-data.html
2. https://resources-covid19canada.hub.arcgis.com/datasets/provincial-daily-totals/data
3. https://climate.weather.gc.ca/historical_data/search_historic_data_e.html
4.https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74
5.https://covid19.apple.com/mobility
6.https://www.google.com/covid19/mobility/
7.https://trends.google.com/trends/?geo=US