The COVID-19 Trend Analysis and Prediction in Canada

Chang Liu 40056360 Kevin Hoang-Nam Trinh 40057470

Abstract

Abstract—In this project, we investigated the accuracy of different models when predicting COVID-19 new cases in Canada. We compare between Linear Regression, Prophet, ARIMA, and Auto ARIMA. We also investigate the correlation between predicted values, the average temperature and the wind speed on our results.

1. Introduction

Time Series and Regression are popular tools used to predict or forecast dependent variables. It attempts to estimate the relationship between dependent and independent variables by extrapolating and interpolating the data. Although regression is most often used to find the relationship among variables, it can also be used in for time series, e.g. auto-regression. This report explores the accuracy of Linear Regression [4], Prophet, ARIMA [1], and Auto ARIMA [7] models on COVID-19 related data.

The goal of this project is to make predictions for current and future trends of COVID-19, visualize the data, analyze and find out the reasons for the rising case numbers in order to help researchers to better understand the impact of COVID-19, and provide data samples for medical diagnostic and treatment purposes.

To achieve this goal, an accuracy of 70% and above will be considered as a success. The data is taken from COVID-19 Canada Open Data Working Group. ”Data are entered in a spreadsheet with each line representing a unique case, including age, sex, health region location, and history of travel where available.”[5] (The example of sample data and output is shown in Figure 1 and Figure 2)

2. Methodology & Experimental Results

To achieve our goal, we first had to study and acknowledge the related information of COVID-19 in Canada. Meaning a good data-set needed to be found as it would be one of the most important aspect of the project. Starting with a good data-set would facilitate the tasks at hand and provide more insightful information.

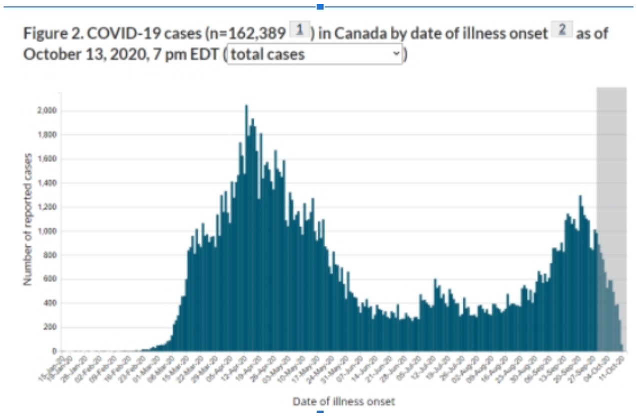
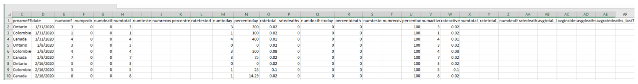


Figure 2. COVID-19 cases (n=162,389 1) in Canada by date of illness onset 2 as of October 13, 2020, 7 pm EDT (total cases ⌄)

Figure 1. Sample output aimed to achieve.



	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1	prnameFR	date	numconf	numprob	numdeath	numtotal	numteste	numrecov	percentre	ratetested	numtoday	percentoday	ratetotal	ratedeaths	numdeathstoday	percentdeath	numteste	numrecov	percentac	numactive	rateactive	numtotal_	ratetotal_	numdeath	ratedeath	avgtotal_l	avgincide	avgdeaths	avgratedeaths_last7	
2	Ontario	1/31/2020	3	0	0	3					3	300	0.02	0	0	0			100	3	0.02									
3	Colombie	1/31/2020	1	0	0	1					1	100	0.02	0	0	0			100	1	0.02									
4	Canada	1/31/2020	4	0	0	4					4	400	0.01	0	0	0			100	4	0.01									
5	Ontario	2/8/2020	3	0	0	3					0	0	0.02	0	0	0			100	3	0.02									
6	Colombie	2/8/2020	4	0	0	4					3	300	0.08	0	0	0			100	4	0.08									
7	Canada	2/8/2020	7	0	0	7					3	75	0.02	0	0	0			100	7	0.02									
8	Ontario	2/16/2020	3	0	0	3					0	0	0.02	0	0	0			100	3	0.02									
9	Colombie	2/16/2020	5	0	0	5					1	25	0.1	0	0	0			100	5	0.1									
10	Canada	2/16/2020	8	0	0	8					1	14.29	0.02	0	0	0			100	8	0.02									

Figure 2. Sample Data used for our various models.

As mentioned earlier, we found our data from a public GitHub which neatly organizes data from open data Canada [5]. The first step we took was to validate the data and we quickly realized that no further preprocessing was needed as the data was mostly numerical. No null values were present and the data was already presented under different columns displaying the current number of COVID-19 cases, deaths, tests, and their respective rate.

With our data-set, the logical way to split it into training and testing data was by time periods. The team did think about splitting by provinces, however that had too many independent factors which we did not have data for. Those unknown features could invalidate our findings, since for example, the health and safety guidelines for Quebec and Ontario have been different throughout the year. Having decided to use time periods as the main factor to obtain our training and testing data, we opted for the 85/15 rule. This meant that our training data would be taken from the month of March to September and the training data would consists of only the month of October. With these ratios and the overall size of each data-set we estimated that models could be trained on our computers and their would be no risk of scalability or engineering difficulties.

With the data now separated, we decomposed it further down into three distinct components, namely the trend, seasonality, and noise. We can now make our predictions with different models and see which one is the most accurate compared to the actual data. Accuracy in our case was calculated as $Accuracy = 1 - [(actualcases - predictedcases)/actualcases]$

Linear regression is a linear model, a model that assumes a linear relationship between the input variables (x) and the single output variable (y). In our linear Regression Model, we first define numbers of the week as variable X_2 and data of cases as y_2. Then we predict the future trend in the linear model with the relationship $y_2 = -1128.3 + 88.346(X_2)$. Next, we define the weeks for prediction as X_3 and data of forecasting cases as y_3. We calculated the relationship between forecasting weeks and cases as $y_3 = -952.4 + 45.08(X_3)$. Through the observation of the graphs, we found that although our method and predictions can predict the general trend, there still exist some errors in the accuracy which could not provide the ideal results.

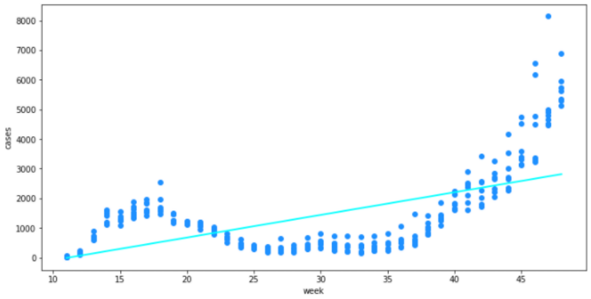


Figure 3. Linear Regression model COVID-19 Daily Cases prediction.

Prophet model is used for forecasting time series based on either an additive or multiplicative manner where non-linear trends are fitted with seasonality [3]. In our case, the decomposition was made using an additive model, with daily and weekly seasonality as data changes everyday say Friday values differ from Monday ones. Also, we tried to take into account how the general situation had hospitals and official count number sometimes being backlogged during the weekends.

The same procedure as linear regression was used to predict the future cases. As expected, the model performed poorly compared to our goal. This is due to it being based on linear regression as well. An additive Prophet model can be represented as: $y(t) = g(t) + s(t) + h(t) + e_t$ Where t represents the time, $g(t)$ the trend, $s(t)$ the seasonality, $h(t)$ the holidays, and e_t the error. We obtained an accuracy of approximately 40.38% with this model.

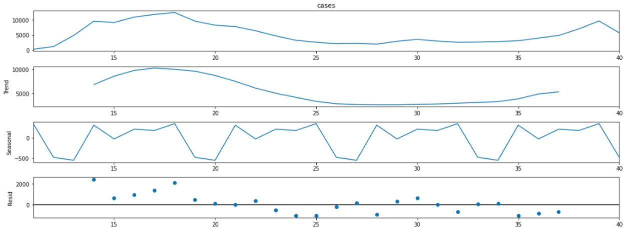


Figure 4. Prophet decomposition values.

Linear Regression and Prophet are trained on squared error which gave not ideal condition/results due to only looking at 'peaks'. Age, sex, travel history are good features but we did not have enough data surrounding it to use them properly and predict trends. Instead of training based on different features, we were more concerned on standardizing the data with seasonality and periodicity. We also tried using 7-day moving averages instead of seasonality/periodicity when it was not applicable.

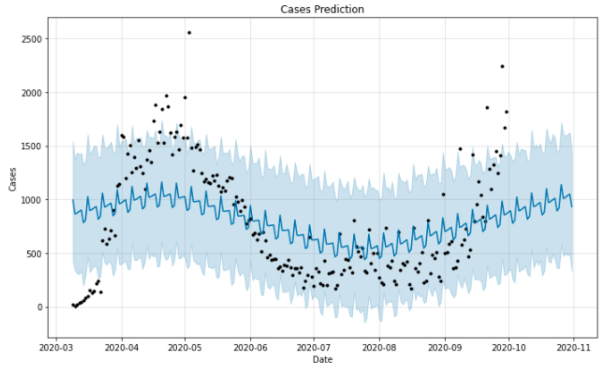
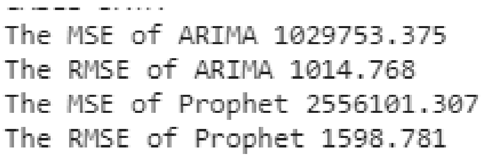


Figure 5. Prophet model COVID-19 Daily Cases prediction.

The team also investigated using an ARIMA, or autoregressive integrated moving average model. Which ”is a generalization of an autoregressive moving average (ARMA) and is fitted to time-series data in an effort to forecast future points. ARIMA models can be especially efficacious in cases where data shows evidence of non-stationarity.” [8]

Using the same data separation as the previous models, we were able to create an ARIMA model. The model was first created with various order and we found that the best one was $(4, 0, 0)$. This represents an auto-regressive model of order 4, with 0 degree of differencing, and with order 0 for the moving-average as our data would already take into account the moving-average. This was done to avoid looking only at peaks but rather investigate trends. The model was then fitted with the mle method since it maximizes the exact likelihood of events and with $transparams$ argument being set to $True$ to ensure the data was stationary.

With this model, we could now calculate the Mean Squared Error and Root Mean Squared Error and compared it to the previous models. This would determine which model is better over the others. As observed in the below figure, we could clearly see that the ARIMA model was a better choice over the Prophet and Linear Regression one.



```
The MSE of ARIMA 1029753.375
The RMSE of ARIMA 1014.768
The MSE of Prophet 2556101.307
The RMSE of Prophet 1598.781
```

Figure 6. MSE and RMSE for Prophet and ARIMA model.

We could now simply predict the future values for the entire month of October using the model's default predict function with our start and end dates as parameters. After the model had finished predicting its values, we were able to plot them against the actual cases we were expecting and thus also determine the accuracy.

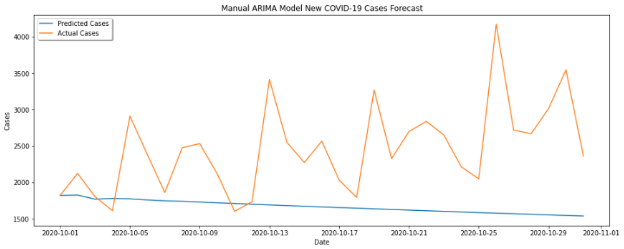


Figure 7. ARIMA model COVID-19 Daily Cases prediction.

Even though seasonality and the order of differentiation were not ideal and non evolving, we obtained an accuracy of 70.96%. The Mean Absolute Error was also calculated to be 806.704 which is relatively high with the numbers we were dealing with. This represented around 30% to 50% variance/uncertainty on our predicted values.

The last model investigated was the AUTO ARIMA. This model uses the same principles as the ARIMA one. However, for each data-point fed into the model, it differentiates with 4 different tests (i.e., Kwiatkowski–Phillips–Schmidt–Shin, Augmented Dickey-Fuller or Phillips–Perron) to determine the order of differencing. Since we had the parameter $seasonal$ set to true, the model also "seeks to identify the optimal P and Q hyper- parameters after conducting the Canova-Hansen to determine the optimal order of seasonal differencing, D" [8].

The same procedure as the regular ARIMA was applied to obtain the predictions. Furthermore, we trained the model with two different data sets to investigate how the period of data would affect it. One data set had values for every single day from 2020-03-08 to 2020-09-30 while the other had values for every week. The data for every week was obtained by summing each day's data. Both models using AUTO ARIMA were then used to obtain 31 days of predicted new cases. With the new month of October forecasts, we plotted them against the same actual cases that occurred in Canada.

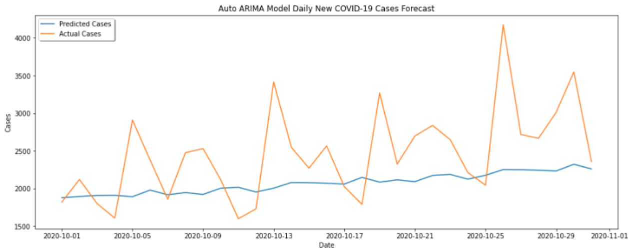


Figure 8. Auto ARIMA model COVID-19 Daily Cases prediction.

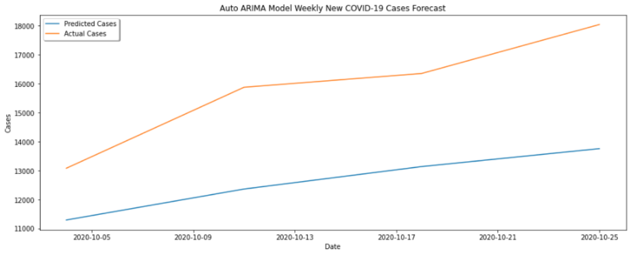


Figure 9. Auto ARIMA model COVID-19 Weekly Cases prediction.

We were able to notice a slight increase in performance and accuracy with the daily data. The team obtained an accuracy of 83.26% compared to 80.21% with weekly data. This slight increase may not seem to be a lot, but on the order of things, we know that the higher the accuracy we try to achieve, the harder it is to get. There are diminishing return but not in our scenario. The only downside is that the model took significantly more time to train.

Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were also used to determine the models' accuracy. MAE was not used often as the results may not represent the true nature of our model.

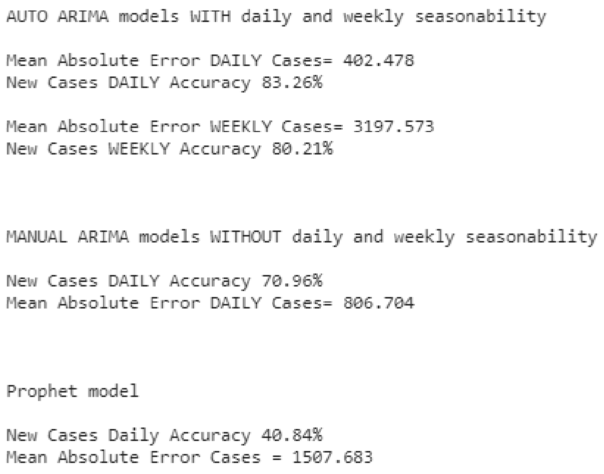
$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. It calculates the difference between the actual values and the prediction, by averaged the absolute difference over the data set. [2] The range of MAE is $[0, +\infty)$,When the predicted value is completely consistent with the actual value (equals 0), the model can be considered as a perfect model. Therefore, the greater the error, the greater the value, and the model is less accurate.

$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. It calculates the difference between the actual values and the prediction, by squared the absolute difference over the data set. [2] In fact, the $(y_i - \hat{y}_i)^2$ is the loss function of the linear regression. The purpose of linear regression is to minimize the loss function. However, this judgment method will amplify the error since it squares the error values. In Figure 6, the MSE of prophet model is nearly 2.5x larger than the MSE of ARIMA. It is because that the number of mean errors exist in the prophet model is larger than the one in the ARIMA model, which means ARIMA model make better prediction than the prophet model.

$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ "RMSE increases with the variance of the frequency distribution of error magnitudes." [6]. RMSE is the squared root of MSE. Since MSE gives us very large numbers to analyse, RMSE can make the data more intuitive. In Figure 6, the result values of RMSE is much smaller than the MSE, and we can clearly conclude that the ARIMA model has higher accuracy

We found that Auto ARIMA came to the best results because it uses multiple order of differentiation instead of just one like regular ARIMA. Predictions from the Linear Regression did not take advantage because it used simple regression/gradient descent but our data is non-linear. In addition, Linear Regression is trained on the R squared which gave not ideal condition/results due to only looking at 'peaks'.

To maximize performance we simple used multiple different models of periodicity and seasonability. By ignoring prediction quality in the smaller scale regime, we opted to look at trends and averages when training our model to not over-fit them. We tried to correlate age, sex, etc. to our models, but there was no direct correlation we could observe and thus decided to go straight to the time-series data instead. As a last test, we tried to relate the average temperature and wind speed with the predicted number of cases and actual cases. This is discussed more in detail in the "Appendix: Extra Results" section as it was coming from another dataset and we had to manually clean and match the dates. Our models were able to predict accurately without being trained with multiple features as their algorithm are based more on trends, seasonality, and periodicity. Even though Canada had many public health restriction, the models predicted a gradual increase for the month of October which matches the 2^{nd} COVID-19 wave in Canada.



```
AUTO ARIMA models WITH daily and weekly seasonability

Mean Absolute Error DAILY Cases= 402.478
New Cases DAILY Accuracy 83.26%

Mean Absolute Error WEEKLY Cases= 3197.573
New Cases WEEKLY Accuracy 80.21%



MANUAL ARIMA models WITHOUT daily and weekly seasonability

New Cases DAILY Accuracy 70.96%
Mean Absolute Error DAILY Cases= 806.704



Prophet model

New Cases Daily Accuracy 40.84%
Mean Absolute Error Cases = 1507.683
```

Figure 10. Accuracy for different models and datasets periods.

3. Conclusions

Throughout the project, we investigate and compare the prediction's performance of Prophet, ARIMA, Auto ARIMA, and Linear Regression Models with the current trends data of COVID-19 in Canada. The result shows that Auto ARIMA models take more advantage in prediction accuracy than Prophet and Linear Regression Models, and the test results accurately match the 2^{nd} wave in Quebec. Hence we can measure this project as a "success"(It is important to note that these results come from our COVID-related data and may not be identical to other types of data sets.)

Furthermore, for future research, we highly recommend implementing and comparing more specific information of the patient's gender, age, region, and related-symptoms. Using these updated data, these predictions can provide researchers with more comprehensive information and data for analysis, and provide medical scientists with data samples to prepare for the second wave in advance. However, from a realistic point of view, COVID-19 may be very unpredictable due to human behaviors and the surrounding environment. Therefore, large spikes negatively degrade these properties.

4. Statement of Contribution

Chang and Kevin both worked on every aspect of the code and compared their findings. The most optimal or adequate results were kept and built upon. As for the the paper, we all contributed equally. Both team members worked in tandem to learn and reinforce their knowledge equally as they have similar experiences.

References

[1] Mustafa Abusalah. Brent oil prices prediction using prophet & arima. 1

[2] Anonymous. Regression model accuracy (mae, mse, rmse, r-squared) check in r. 4

[3] Ryan Ben et al. Prophet: Automatic forecasting procedure. 2

[4] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006. 1

[5] Jean-Paul R. Isha et al. Epidemiological data from the covid-19 outbreak in canada. 1

[6] JJ. Mae and rmse — which metric is better?Mean Absolute Error versus Root Mean Squared Error. 4

[7] Jose Marcial Portilla. Using python and auto arima to forecast seasonal time series. 1

[8] Taylor G Smith. pmdarima: Arima estimators for python. 2, 3

Appendix: Extra Results

Besides, there are many rumors that people are more likely to catch the flu in a cooler environment. Influenza damages the human immune system and makes people more susceptible to coronavirus. Some scientists believe that a warm and humid environment can reduce the spread of coronavirus.

Hence, our team considered putting average temperature and wind speed as features. First, we visualized the temperature and wind data in October 2020. Then, We combined the graphs from our forecasting models with the bar charts of average temperature and wind speed. By comparing with the AUTO ARIMA (highest accuracy) and the bar charts. We found that the trend of the COVID-19 might not affected by wind speed changes and the temperature drop might become a factor in the increase in the number of infected people. Although the certainty of the evidence generated is low, there is homogeneity among the results of the included research reports.

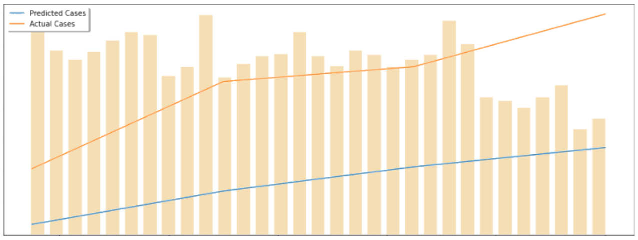


Figure 11. Auto ARIMA Prediction & Temperature Bar Chart.

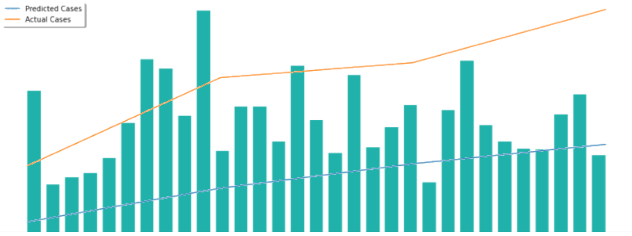
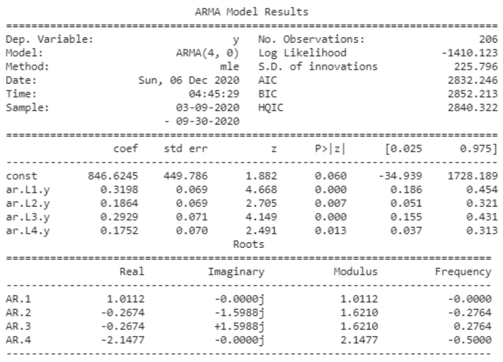


Figure 12. Auto ARIMA Prediction & Wind Speed Bar Chart.



```
                              ARMA Model Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  206
Model:                     ARMA(4, 0)   Log Likelihood               -1410.123
Method:                           mle   S.D. of innovations            225.796
Date:                Sun, 06 Dec 2020   AIC                           2832.246
Time:                        04:45:29   BIC                           2852.213
Sample:                    03-09-2020   HQIC                          2840.322
                         - 09-30-2020
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        846.6245    449.786      1.882      0.060     -34.939    1728.189
ar.L1.y        0.3198      0.069      4.668      0.000       0.186       0.454
ar.L2.y        0.1864      0.069      2.705      0.007       0.051       0.321
ar.L3.y        0.2929      0.071      4.149      0.000       0.155       0.431
ar.L4.y        0.1752      0.070      2.491      0.013       0.037       0.313
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.0112           -0.0000j            1.0112           -0.0000
AR.2           -0.2674           -1.5988j            1.6210           -0.2764
AR.3           -0.2674           +1.5988j            1.6210            0.2764
AR.4           -2.1477           -0.0000j            2.1477           -0.5000
-----------------------------------------------------------------------------
```

Figure 13. ARIMA model results.