Prediction and Visualization of The Transmission of COVID-19 Using Machine Learning Models

DAI Jie

Tongji University

Abstract

Background: The whole world is significantly influenced by the pandemic of COVID-19. Though vaccination has been studied and promoted, the number of confirmed cases is growing without an attempt to stop. This project aims to predict the development of the pandemic on time and space scales using machine learning models.
Methods: Four different machine learning models, namely Linear Regression, Support Vector Regression, SARIMAX and Facebook's Prophet have been employed to forecast the epidemic trend in the top 5 affected countries.
Results: Facebook's Prophet turned out to be the best model in performance. Implementation of Facebook's Prophet has been performed on worldwide level and individual country level with inevitable system errors and externalities. The obtained database requires further development for unsolved issues. The possibility of cross validation between data from different countries has been discussed.

Key Words: *COVID-19, Machine Learning, Prediction, Time series analysis, Linear Regression, SVR, SARIMAX, Prophet*

1. Introduction

The world has been continually suffering from the Novel Coronavirus (COVID-19) pandemic since the first reported outbreak in December 2019 in Wuhan, China. The virus is highly contagious and has been affecting people all over the world. Until September 24, 2021, the total infected cases has been over 230 million and the total deaths has been over 4.7 million. Despite governments of countries in the world have been declaiming countermeasures including social distancing, quarantine and vaccination, the transmission of COVID-19 still seems undefendable for most countries. Hence understanding how the virus spreads on both time and space scales is undoubtfully pivotal for governments and international organizations such as WHO to take truly effective measures in terms of fighting the epidemic.

In this project, the author visualized the worldwide transmission of COVID-19 with plotly express package and analyzed the COVID-19 data from multiple countries using four machine learning models, namely the Linear Regression model, the Seasonal Auto Regressive Integrated Moving Average with eXogenous factors model (SARIMAX), the Support Vector Regression model (SVR) and Facebook's Prophet prediction model.

The main objects of this project are: (a) to visualize the worldwide transmission of COVID-19 on time and space scales, (b) to find the best machine learning model in terms of prediction of the pandemic by comparing performance of the above-mentioned 4 models in forecasting the pandemic of top 5 affected countries, and (c) to predict the future trend of the pandemic of the world and individual countries using the selected model in (b).

2. Literature Review

Ahmad Bani Younes et.al[1] presented a dynamic model based on the Lotka–Volterra model coupled with an extended Kalman Filter so as to evaluate the transmission of COVID-19. The model showed that the spread grows exponentially if none of control measures were imposed, and social distancing was the most effective way to reduce the spread.

Manav R. Bhatnagar et.al[2] developed a simple mathematical model of prediction of COVID-19 based on transmission theory. The model was implemented to data of the USA, Italy, France and India and proven effective and accurate.

Gopi Battineni et.al[3] conducted 60-day forecasting of total infections of COVID-19 in four high hitting countries -- USA, Brazil, India and Russia -- using Facebook's prophet machine learning model. The model performed accurately with the R2 value of 0.995 and gave both the trend and seasonality properties of the pandemic. The authors also found some underestimation and overestimation of daily cases.

Alok Kumar Sahai et.al[4] adopted Auto Regressive Integrated Moving Average (ARIMA) model to predict the incidence and spread of the COVID-19 in the five most badly hit countries -- India, Russia, Brazil, Spain and the US. The ARIMA model showed a good performance in terms of mean absolute deviation (MAD) and mean absolute percentage error (MAPE).

Amit Kumar Gupta et.al[5] predicted the active rate, the death rate, and the cured rate of COVID-19 in India using Linear Regression, Support Vector Machine (SVM) and Prophet Forecasting Model. Facebook's Prophet model was found the most accurate predictive model compared to the other two, in terms of MAE, MSE and RMSE.

3. Methodology

This project has mapped the time lapse of the transmission of COVID-19 onto a geographical density map and employed four machine learning modes -- linear regression, SVR, SARIMAX and Facebook's prophet model – to forecast the trend. The materials and methods are described as the following sections.

3.1 Data Source

The dataset was extracted from the "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University", and the url: https://github.com/CSSEGISandData/COVID-19.

In terms of space and time, this dataset has daily-level and province/state-level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. The dataset contains useful information including observation dates Provinces/States, Countries/Regions, corresponding latitudes and longitudes, and number of Confirmed cases, Deaths and Recovered cases, respectively. It is a time series data and the number of cases on any given day is the cumulative number. The

data for this project is available from 22 Jan, 2020 to 29 May, 2021.

3.2 Data Preprocessing

In order to present the time-lapse spread of COVID-19 onto a world map, the observation dates, province/state-level administrative regions, corresponding geographical coordinates, and the cumulative number of confirmed cases were integrated.

In order to see the cumulative trend of worldwide confirmed, active, deaths and recovered cases, all data was integrated into daily level information. Also, active rates, death rates and cured rates were calculated as below for presentation.

$$\text{Active Cases} = \text{Confirmed Cases} - \text{Deaths} - \text{Recovered Cases}$$

$$\text{Active Rate} = \frac{\text{Active Cases}}{\text{Confirmed Cases}} \times 100\%$$

$$\text{Death Rate} = \frac{\text{Deaths}}{\text{Confirmed Cases}} \times 100\%$$

$$\text{Cured Rate} = \frac{\text{Recovered Cases}}{\text{Confirmed Cases}} \times 100\%$$

In order to fit the machine learning models, the dataset was integrated into daily-and-country/region-level information. Also, the integrated data was split into a 70% training set and a 30% test set, with a divide date of 1 Jan, 2021.

3.3 Linear Regression Model

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.[6]

Mathematically, a simple LR model can be expressed in matrix notation as:

$$y = X\beta + \varepsilon \tag{1}$$

Where y is a vector of observed values $y_i\ (i = 1, \dots n)$ of the variable called the dependent variable, X is a matrix of a series of vectors x_i called the explanatory variable, β is a parameter vector and ε is the error term. [7]

In this project, y is the cumulative numbers of confirmed cases and X is observational dates.

3.4 Support Vector Regression (SVR)8

The Support Vector Machine (SVM) is a supervised learning model constructing a hyperplane in a high dimensional space, which can be used for classification and regression. In this project, its extension Support Vector Regression will be applied.

In general, SVR model does not consider training data that is too similar to the predicted values. The object of SVR is to minimize the coefficients with constraint where the identified margin is greater than or equal to the absolute error.

Mathematically, training the SVR models means:

$$MIN \ \frac{1}{2}\left|\left|\omega\right|\right|^2 \tag{2}$$

$$Subject\ to \qquad \epsilon \geq |y_i - x_i\omega_i|$$

Where x_i is the training sample with target value y_i, ω_i is the set of coefficients and ϵ is the identified margin known as maximum error.

3.5 Seasonal Auto Regressive Integrated Moving Average with eXogenous factors model (SARIMAX)[9]

SARIMAX is an updated version of the ARIMA model. ARIMA includes an autoregressive integrated moving average, while SARIMAX includes seasonal effects and eXogenous factors with the autoregressive and moving average component in the model. Mathematically, SARIMAX can be presented as:

$$\phi_p(L)\tilde{\phi}_P(L^s)\Delta^d\Delta_s^D y_t = A(t) + \theta_q(L)\tilde{\theta}_Q(L^s)\epsilon_t \tag{3}$$

Where:

- $\phi_p(L)$ is the non-seasonal autoregressive lag polynomial
- $\tilde{\phi}_P(L^s)$ is the seasonal autoregressive lag polynomial
- $\Delta^d\Delta_s^D y_t$ is the time series, differenced d times, and seasonally differenced D times.
- $A(t)$ is the trend polynomial (including the intercept)
- $\theta_q(L)$ is the non-seasonal moving average lag polynomial
- $\tilde{\theta}_Q(L^s)$ is the seasonal moving average lag polynomial

3.6 Facebook's Prophet[10]

Prophet model is a decomposable time series model designed by Facebook for forecasting non-linear trends. It has three main model components: trend, seasonality and holidays, combined in a simple linear equation as:

$$y(t) = g(t) + s(t) + h(t) + \epsilon(t) \tag{4}$$

Where $g(t)$ is the trend function which models non-periodic changes in the value of the time series, $s(t)$ is periodic changes (e.g., weekly and yearly seasonality), and h(t) represents the effects of holidays which occur on potentially irregular schedules over one or more days. The error term $\epsilon(t)$ represents any idiosyncratic changes which are not accommodated by the model.

4. Results

This section presents the visualization of worldwide spread of COVID-19, comparative analysis of forecasting the pandemic in the top 5 affected countries with the four machine learning models based on train/test performance, prediction of world-level and country-level of the pandemic using the most accurate model—Facebook's Prophet and cross validation between different countries with this model.

4.1 Visualization of The Spread of COVID-19

We plotted the data to see the cumulative trend of worldwide confirmed, active, deaths and recovered cases, as Figure 1 shows below. In general, all these four numbers are steadily growing, which means the pandemic is still spreading widely and fast. Interestingly, the number of recovered cases is higher than that of active cases, and the number of deaths remains relatively low, suggesting that defeating the pandemic is quite promising.

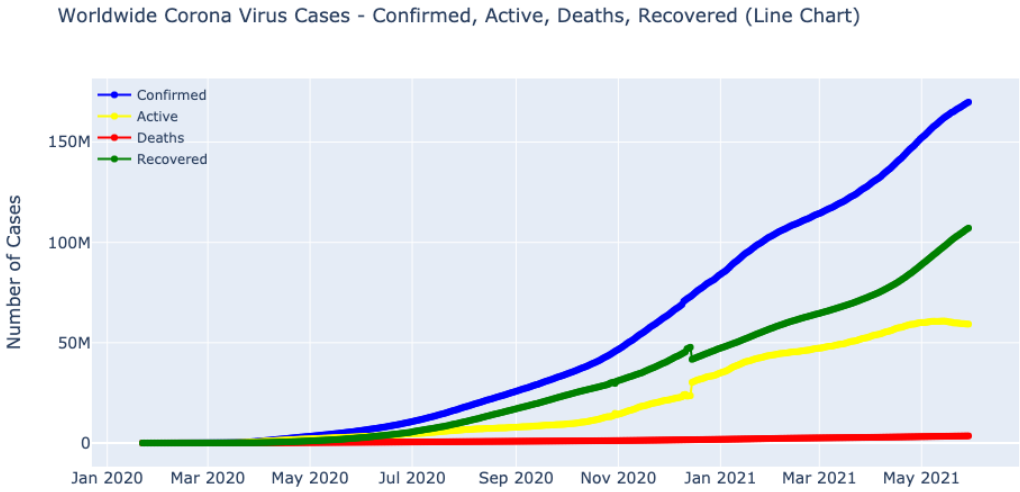

Figure 1 Worldwide Confirmed, Active, Deaths and Recovered Cases

We also plotted data of active rate, death rate and cured rate, as Figure 2 shows. The active rate is unstably changing, suggesting that the situation of the transmission is more complicated than we consider, probably because of the increasing multiple variants of COVID-19 virus. The cured rate is promisingly going up and the death rate stays low, which again suggests that the prevention of the epidemic is effective in general. The abrupt discontinuities occurring in Figure 1 and Figure 2 may be on account of the change of the statistical standard of recovered cases in some countries or simply large-scale false negative results of testing.

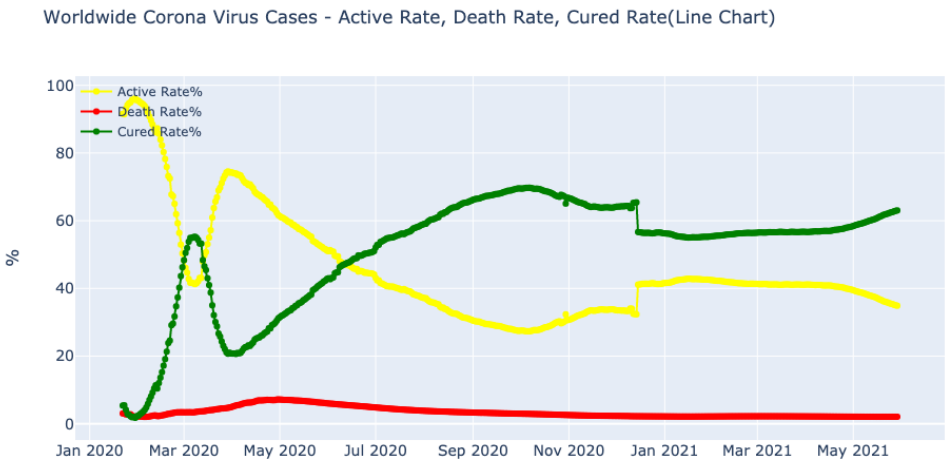

Figure 2 Worldwide Active Rate, Death Rate and Cured Rate

To intuitively percept the worldwide spread of COVID-19, we made a dynamic time-lapse density map as Figure 3 shows. Unfortunately it can't be vividly demonstrated due to the inherent statical restriction of papers.

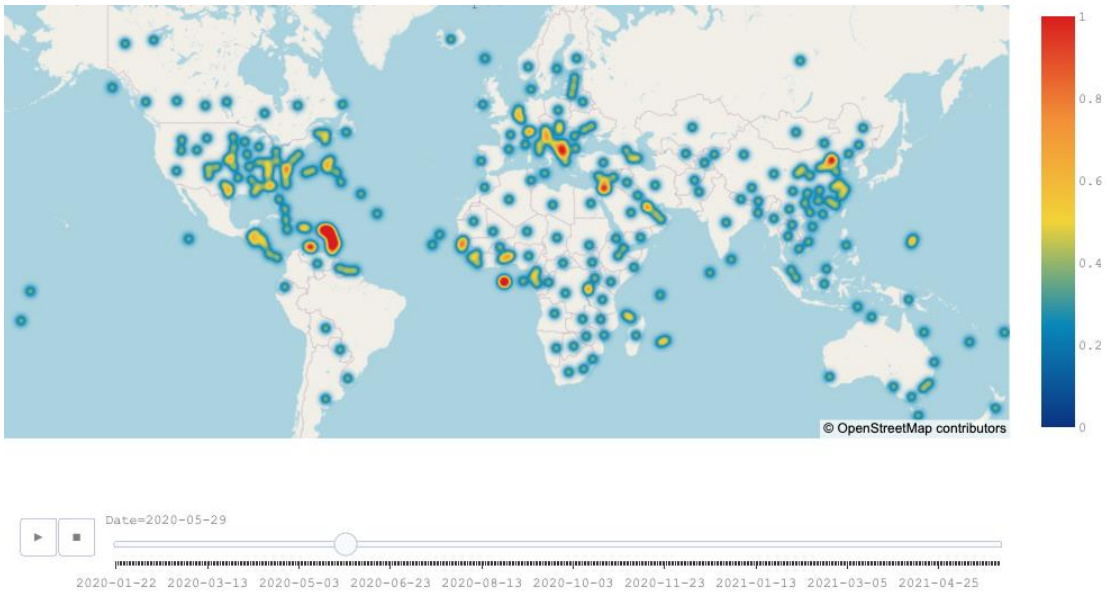

Figure 3 Worldwide Confirmed Cases Time Lapse (framed at date 2020-05-29)

Also, we plotted the cumulative confirmed cases on the last observation date of the dataset onto natural earth projection, as Figure 4 shows. Top 5 affected countries – US, India, Brazil, UK and Russia can be easily recognized in Figure 4.

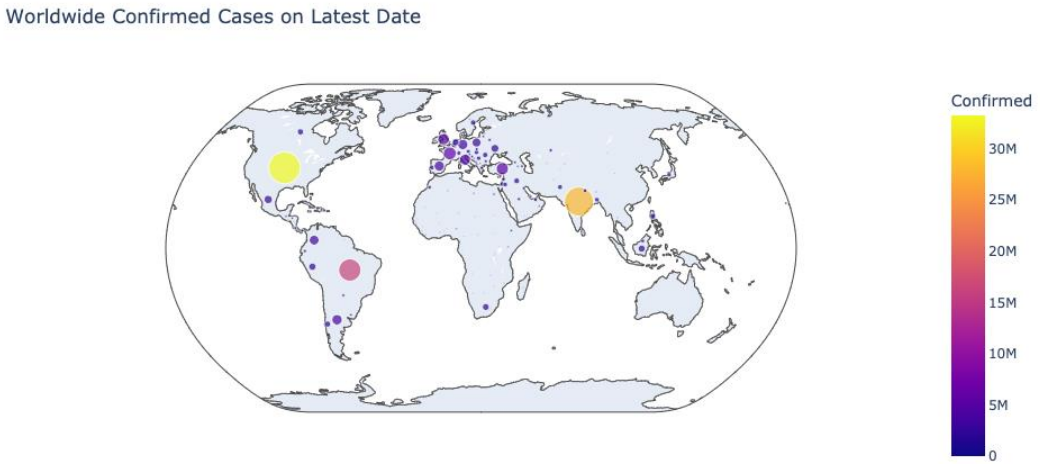

Figure 4 Worldwide Confirmed Cases on Latest Date

4.2 Comparative Analysis of Forecasting The Pandemic with Four Machine Learning Models – Linear Regression, SVR, SARIMAX and Facebook's Prophet

This project applied LR, SVR, SARIMAX and Prophet for forecasting the cumulative number of confirmed cases of COVID-19 in the top 5 affected countries mentioned above. The dataset was split into separated 70% training set and 30% test set, in order to measure the performances of different models. It turned out that Facebook's Prophet is the best model in terms of accuracy in forecasting the COVID-19 time series data. The seasonal module uses the Fourier transformation series for analysis of data based on yearly, weekly, or daily basis. Prophet model automatically identify the change points in data which affects the trends . The SARIMAX gave the second best accuracy considering also seasonal effects but not change points. As the simplest model, the LR method does not use any of the seasonality or other component which may affect the data on the parameter of time, and hence LR returned a monotonous prediction in this project. The SVR performed the worst in this project. It may shows more accuracy in the area of classification rather than time series, considering its relatively complicated high dimensional features.

Table 1 shows performance in terms of R square (R2), Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) for the four models. Figure 5 to 8 show the graphs for predicted and actual datasets of the top 5 affected countries using SVR, LR, SARIMAX and Prophet models.

Table 1 Performance Metrics of Different Models

Country	Model	R2	MAE	RMSE	MAPE
US	Prophet	-1.138	3848918	4978018	0.1246
	SARIMAX	-6.496	7198718	9321495	0.2289
	LR	-11.773	12074354	12167970	0.4162
	SVR	-56.869	25674500	25899283	0.8838
India	Prophet	0.374	2727961	4215304	0.1524
	SARIMAX	-1.255	4895704	7997562	0.2507
	LR	0.17	2692189	4850842	0.1314
	SVR	-6.308	13376351	14397630	0.9228
Brazil	Prophet	0.417	1658460	2003050	0.1256
	SARIMAX	0.669	1194996	1508445	0.0885
	LR	-0.331	2657868	3026403	0.2079
	SVR	-12.514	9277973	9641616	0.773
UK	Prophet	-0.624	533595	596843	0.1335
	SARIMAX	-37.208	2261917	2894862	0.5169
	LR	-23.607	2303906	2323188	0.5623
	SVR	-65.388	3787016	3815893	0.9245
Russia	Prophet	-1.756	627466	826931	0.1352
	SARIMAX	-4.117	893966	1126791	0.1937
	LR	-7.25	1420971	1430805	0.3328
	SVR	-49.83	3516329	3551472	0.8206

Figure 5 Forecasting Confirmed Cases with SVR

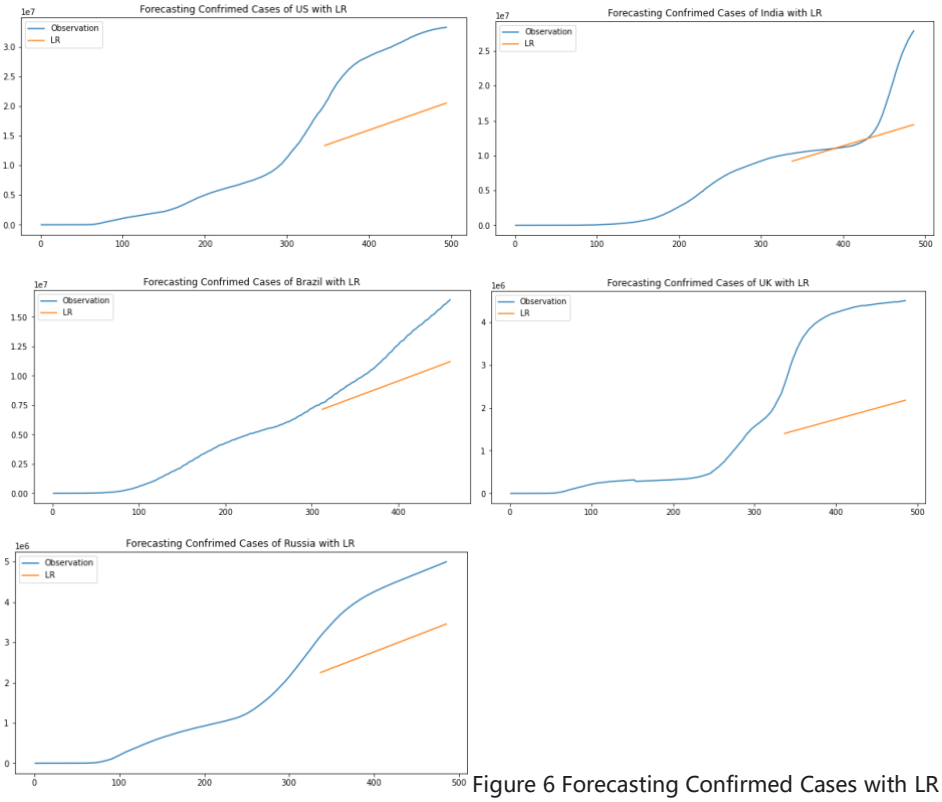

Figure 6 Forecasting Confirmed Cases with LR

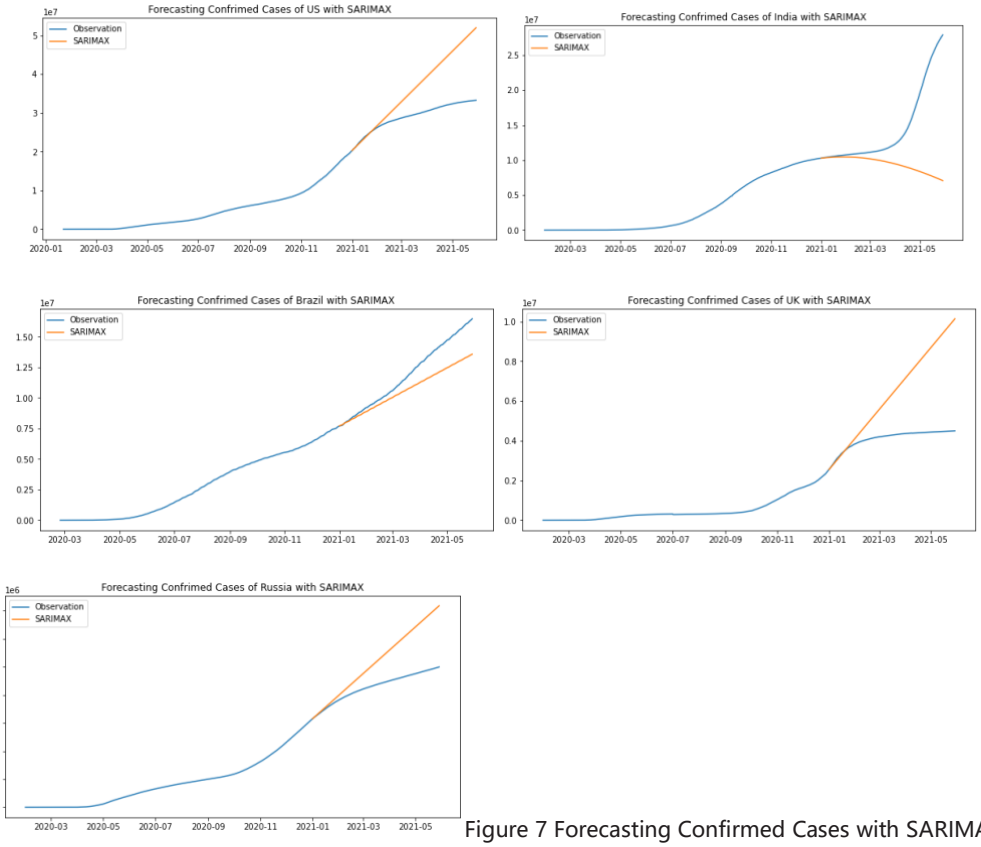

Figure 7 Forecasting Confirmed Cases with SARIMAX

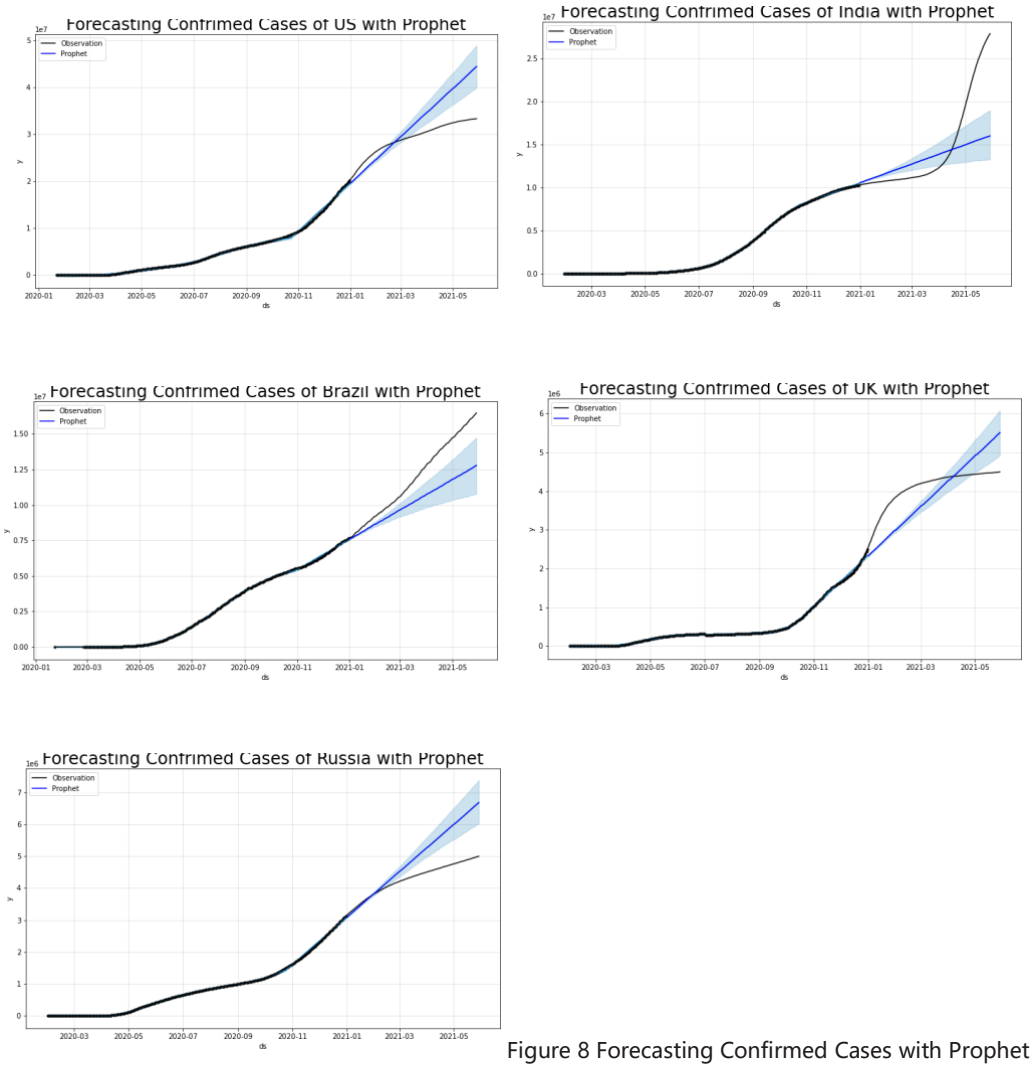

Figure 8 Forecasting Confirmed Cases with Prophet

4.3 Implementation of Facebook's Prophet for Forecasting Worldwide and Country Level Confirmed Cases of COVID-19

To finalize the selected best model – Facebook' Prophet for predicting worldwide and country level pandemic trend, we fit the model with the whole dataset instead of only training set and made a 150-day prediction as Figure 9 shows. As the prediction suggests, by the end of September 2021, the total number of worldwide confirmed cases of COVID-19 will be over 240 million (It's nearly 231 million on 25 September) and by the end of October, the number will be over 260 million. Figure 10 shows the seasonality features including monthly and weekly trends of worldwide confirmed cases. Not surprisingly, the number is highest at the weekends in the weekly trend since weekends are usually when most people travel and gather to party. Interestingly, in terms of monthly trend, it is highest around the middle of the month. However, considering that testing results may take longer time to be available in some areas and some virus carriers do not have symptoms, the delays and unseen positives as system errors cannot be detected and reflected by the model.

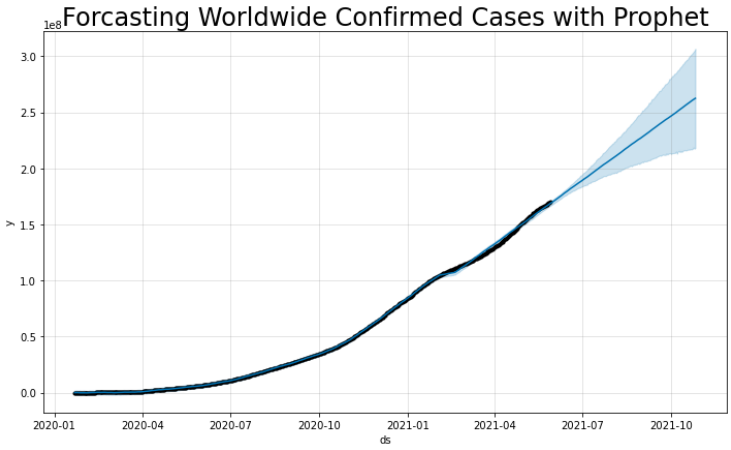

Figure 9 Forecasting Worldwide Confirmed Cases with Prophet

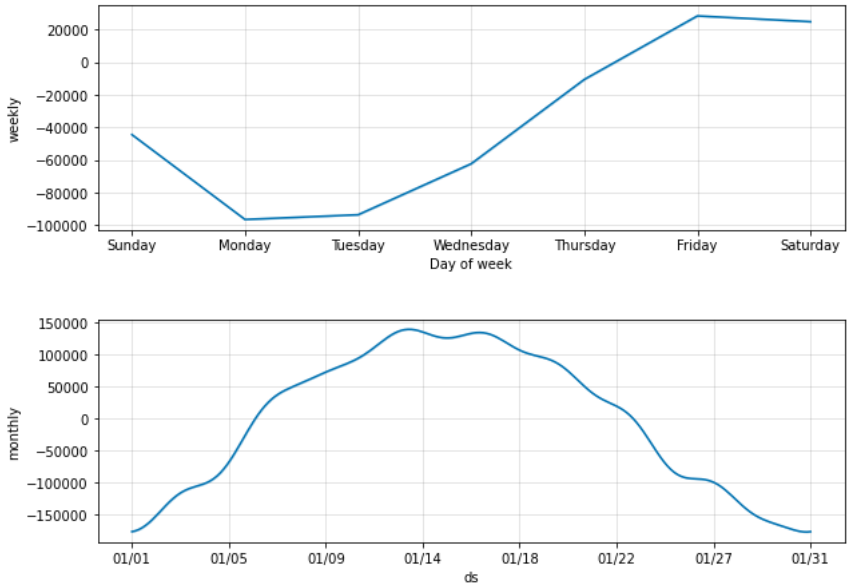

Figure 10 Seasonality of Worldwide Confirmed Cases

The accuracy standing the test, this project also employ Facebook's Prophet for prediction of pandemic ongoing trend of every individual country in the world, as long as its data is available. In the long run, Prophet model tend to produce underestimates results in this project. For example, as the model predicts, by the end of October, the total confirmed number of the US will be over 42 million, which has already been reached until this day (25 September). Also, that number of UK by the end of October could be around 5 million as the model suggests. However, until this day this number has been over 7.6 million! It could be related to the lack of the adoption of necessary countermeasures or new variants of the virus.

Apparently more study is required to unravel the mystery of the underestimate characters of Prophet model. The database of individual-country-level trend and seasonality awaits to be further developed to study the issues and factors mentioned above. Given the more accurate scenarios, it can serve to provide decision makers with a clear and intuitive perspective so as to find a solution to halt the pandemic at some point.

4.4 Exploring Cross Validation between Different Countries*

Some countries have similar scale in terms of total confirmed cases, e.g. Sweden and Portugal. This section will explore the possibility of applying a model trained on dataset of Sweden to predict the trend of Portugal, or the other way around. As amazingly as Figure 11 and Table 2 show, this idea unexpectedly works! Needless to say, one reason is the strong robustness of Facebook's Prophet. However, another explanation behind this phenomenon could be that Sweden and Portugal share not only similar scale in the number of confirmed cases, but also similar patterns of countermeasures (passive herd immunity), geographical location (European countries) and population size (10.23 million and 10.28 million in 2019, respectively). Therefore, the success of this cross validation is a result of both the endogenous robustness of the model and exogenous circumstances regarding to politics, geography and demography. More examples of such implementation can be furthered studied.

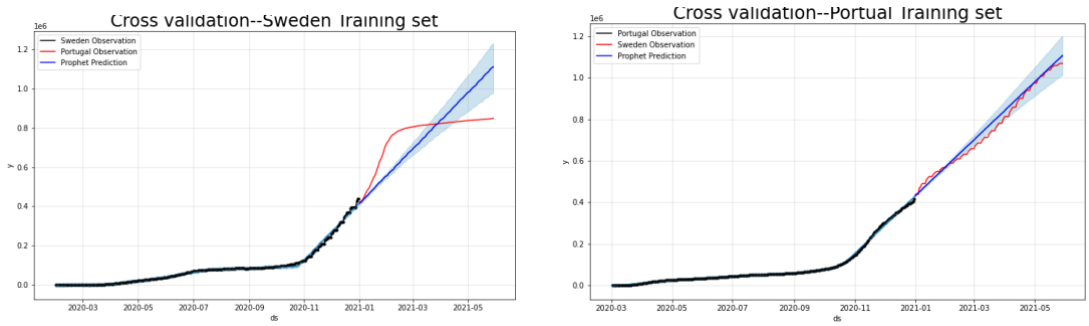

Figure 11 Cross Validation between Sweden and Portugal

Table 2 Performance Metrics of Cross Validation

Parameter	Sweden as training set	Portugal as training set
R2	-0.18	0.981
MAE	108447	21699
RMSE	128769	25819
MAPE	0.1383	0.0307

5. Discussion and Future Scope

In this project, the worldwide spread of COVID-19 has been visualized in varies types of plots. Also, the SVM, LR, SARIMAX and Prophet Model has been experimentally tested for the time series analysis of forecasting the COVID-19 pandemic dataset of top 5 affected countries. This result suggested that Facebook's Prophet gives the best performance compared to other models. Using Prophet, the prediction of future transmission of COVID-19 both on worldwide level and individual country-level has been made. This database requires to be furthered enhanced and improved for study the pattern of COVID-19 epidemic.

This project also has mainly the following limitation -- current machine learning models do not consider system errors such as delayed effect of the data and undetected positive cases, and externalities such as change of policies and emergence of new variants of the virus. These factors could lead to overestimation or more usually underestimation of predictions. It's difficult to take these factors into consideration since the uncertainty is huge.

Despite of the limitation, the author believes that this work will help not only predict the pandemic in any country with high accuracy, but also search for a solution to prevent its transmission. In order to achieve such vision, future work including studying the effect of different countermeasures and the influence of different variants of virus should be put on the agenda.

6. Acknowledgement

The author thank Professor Ramin Ramezani and the teaching assistant Junjie Wang for helpful advice and supportive guidance throughout the whole project.

7. Reference

1. Bani Younes A, Hasan Z. COVID-19: Modeling, Prediction, and Control. *Applied Sciences*. 2020;10(11):3666. doi:10.3390/app10113666
2. Bhatnagar, Manav R. 2020. "COVID-19: Mathematical Modeling and Predictions," 7.
3. Battineni, Gopi, Nalini Chintalapudi, and Francesco Amenta. 2020. "Forecasting of COVID-19 Epidemic Size in Four High Hitting Nations (USA, Brazil, India and Russia) by Fb-Prophet Machine Learning Model." *Applied Computing and Informatics* ahead-of-print (ahead-of-print). https://doi.org/10.1108/ACI-09-2020-0059.
4. Sahai, Alok Kumar, Namita Rath, Vishal Sood, and Manvendra Pratap Singh. 2020. "ARIMA Modelling & Forecasting of COVID-19 in Top Five Affected Countries." *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14 (5): 1419–27. https://doi.org/10.1016/j.dsx.2020.07.042.
5. Gupta, Amit Kumar, Vijander Singh, Priya Mathur, and Carlos M. Travieso-Gonzalez. 2021. "Prediction of COVID-19 Pandemic Measuring Criteria Using Support Vector Machine, Prophet and Linear Regression Models in Indian Scenario." *Journal of Interdisciplinary Mathematics* 24 (1): 89–108. https://doi.org/10.1080/09720502.2020.1833458.
6. http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm
7. https://en.wikipedia.org/wiki/Linear_regression
8. https://en.wikipedia.org/wiki/Support-vector_machine
9. https://analyticsindiamag.com/complete-guide-to-sarimax-in-python-for-time-series-modeling/#:~:text=SARIMAX(Seasonal%20Auto%2DRegressive%20Integrated,average%20component%20in%20the%20model.)
10. Taylor, Sean J, and Benjamin Letham. 2017. "Forecasting at Scale." Preprint. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.3190v2.