Predicting COVID19 with Prophet

Ashmitha Shetty
Kavya Todi
Vidya Venkatesan
Rutgers University
Piscataway, NJ, USA

I. Project Description

The outbreak of Covid-19 is developing into a major international crisis, and it's starting to influence important aspects of daily life. A strong model that predicts how the virus could spread across different countries and regions may be able to help mitigation efforts. The goal of this project is to implement a model to predict the future trends of the COVID19 data set at country and state/province level using the Prophet algorithm.

The project has four stages: Gathering, Design, Infrastructure Implementation, and User Interface.

A. Stage1 - The Requirement Gathering Stage.

COVID19 is a novel respiratory virus that was first identified in China. Since then this virus has spread throughout the globe causing millions of death worldwide. Typically for diseases like influenze, researchers have lot of historical data to create a prediction model. Since we have no previous data from this disease outbreak we are not clear how easily or sustainably this virus is spreading between people. There are many unanswered questions about how this virus gets transmitted and what are the other factors that impact its spread. Incorporating all this information will help us make a reliable prediction model of COVID19 infections. With an effective model we can prepare the government and people to take or implement the control measures proactively to reduce the impact of coronavirus disease.

In this project, we focus on implementing a model to predict the future trends of the novel COVID19.It adopts machine learning technology and applies prophet algorithm on the prediction system.We used the dataset collected by John Hopkins University for this purpose.

- The general system description: Predicting COVID19 with Prophet.
- The types of users: It is a general system that works works for any user.
- The user's interaction modes: Inputting the daily case reports for all the countries and outputting the forecast for per country and on a global level basis.
- The real world scenarios:
  - Scenario 1
    - Description: The user wants to know how many people recovered from the virus in their country/World-wide
    - System Data Input: None.
    - Input Data Types: String.
    - System Data Output: The result will be displayed.
    - Output Data Types: Performance metrics and Graphical visualization.
  - Scenario 2:
    - Description: The user wants to know how many people died from the virus in their country/world-wide
    - System Data Input: None.
    - Input Data Types: String.
    - System Data Output: The result will be displayed.
    - Output Data Types: Performance metrics and Graphical visualization.
- Project Time line:
  - April 15 to April 18- Planning the ideas for the project
  - April 19- Project proposal
  - April 20- April 27- Gathering the required data and deciding data types
  - April 28- May 2- Designing the layout and implementing the algorithm
  - May 3- May 11- Testing and deployment

B. Stage2 - The Design Stage.

In this section we describe the data transformation process, the overall approach used for the prediction with high level pseudo code for the system operation, as well as the system time and space complexity.

A brief textual description of the overall functional operation of the system

- Data selection and pre-processing.
- Split the data for training and testing
- Using Prophet algorithm we train the model on the training dataset.
- Make the predictions for the test data and visually represent the forecast and other trends as a line plot.
- Evaluate the performance of the model by comparing the results against the number of actual cases in test set using measures like RMSE and MAE.

Data processing

To start with we used daily cases report collected by John Hopkins University. This dataset includes time series

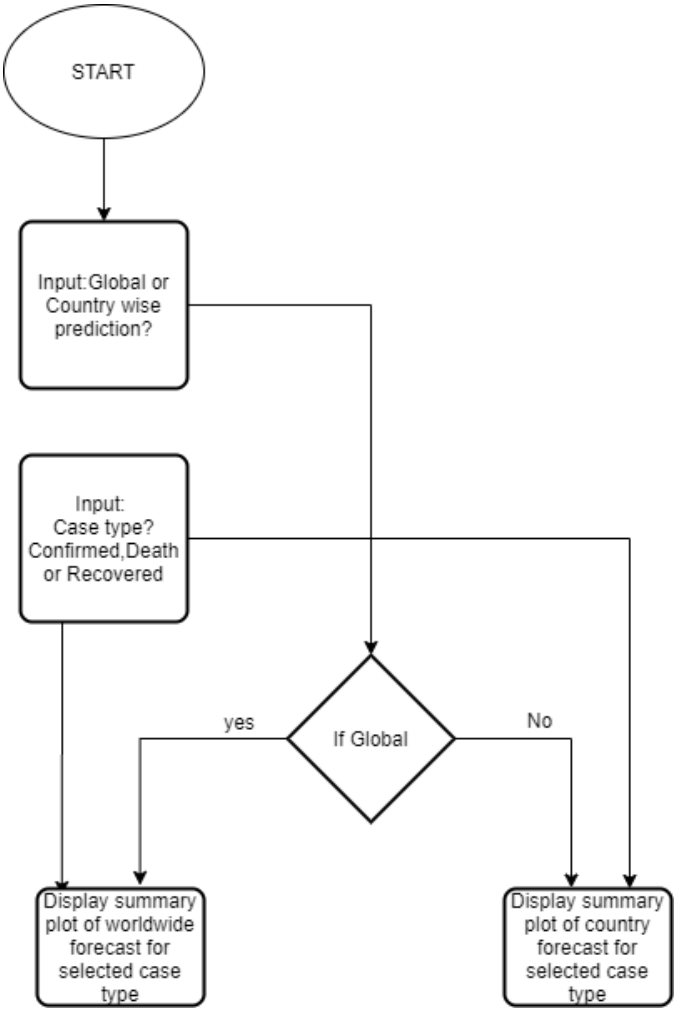

Fig. 1. User Interface of the system

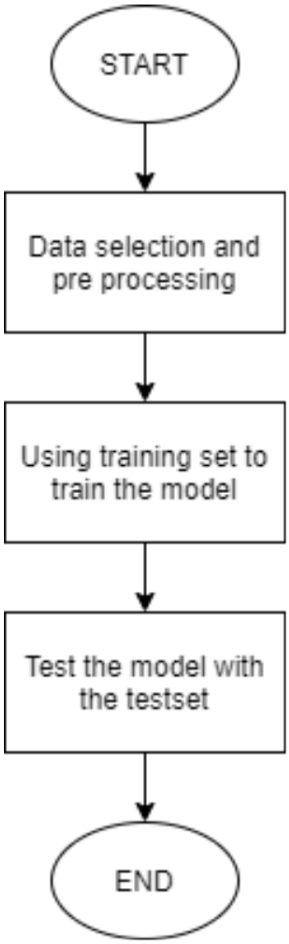

Fig. 2. High level flow diagram of the system

data tracking the number of people affected by COVID-19 worldwide, including

- confirmed tested cases of Coronavirus infection
- the number of people who have reportedly died while sick with Coronavirus.
- the number of people who have reportedly recovered from it.

In order to generate accurate predictions, we then combined this with the average temperature and the population density of each country and merged to a single dataset for our use.
The entire dataset is then divided into training and testing set. About 80 percent of the cases are put in the training data set and the remaining 20 percent of the cases are put in the testing dataset.

Prediction

Here we describe the algorithm that we used to make the forecast. Predictions for this model is based on Prophet algorithm. It is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. At itâs core, Prophet is an additive model with the following components: y(t) = g(t) + s(t) + h(t) + e

- g(t) models trend, which describes whether data increases or decreases in the future.
- s(t) models seasonality with Fourier series, which describes how data is affected by seasonal factors such as the time of the year.
- h(t) models the effects of holidays or large events that impact business time series.
- e is the error term.

We first used the times series data Prophet model with date and total number of daily cases already seen to get the base prediction. Then we added additional external regressors to fine tune the model. From our literature survey we identified three parameters that are believed to impact the spread this virus. We incorporated these factors to improvise our model to make more accurate predictions.
Following are the regressors used

	Province/State	Country/Region	Lat	Long	Date	Confirmed	Death	Recovered
29255	NaN	Western Sahara	24.215500	-12.885800	5/10/20	6	0	NaN
29256	NaN	Sao Tome and Principe	0.186360	6.613081	5/10/20	208	5	NaN
29257	NaN	Yemen	15.552727	48.516388	5/10/20	51	8	NaN
29258	NaN	Comoros	-11.645500	43.333300	5/10/20	11	1	NaN
29259	NaN	Tajikistan	38.861034	71.276093	5/10/20	612	20	NaN

Fig. 3. Sample data snippet

- The time when shutdown was enforced. Many countries have adopted social distancing and quarantine measures to mitigate the impact of the pandemic. We modeled this information to see its effect in the predictions.
- The average temperature of the country. Studies have shown that the virus thrives on cooler weathers. So we took average temperature of each country and divided them in two zones: Hot and cold. We added a positive weight if the country belongs to cold category and zero if it is a hot country.
- The population density. We also wanted to see if the density of countries is associated with the rate of spread of the disease. So we added a positive weight for high dense countries and zero for the low dense countries.

After integrating all these values we made the final prediction. To visually inspect what the model is capturing from the data we plot the forecast and each of the individual components in the trends.

Evaluation

To evaluate the performance of the model we calculated the error between the actual and predicted values. We compute some useful statistics of the prediction performance like mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percent error (MAPE), and coverage of the lower and upper boundaries of the estimates.

C. Stage3 - The Implementation Stage.

We used Python programming language for this project, and specially fbProphet package for its power in scientific computing.

- Sample data snippet: Fig 3 is a snippet of the data used after prepocessing it.
- Sample output: Fig 4 represents the forecast of the confirmed cases worldwide. Breaking this down a bit further Fig 5 represents the components of forecast.
- Sample findings:
  From our experiments we observed that the model with additional regressors gives better results than the model without regressors. The table shows all the evaluation metrics computed with and without regressors.
  For a dataset of size 1KB the rmse value for without regressor evaluates to 30,3961. But by combining social distancing, climate and population density of the countries the rmse substantially reduced to 78.5.
  Some anecdotes for the predictions of the model
  - Predicted number of total cases on Feb 25th are 80192 and actual cases are 80406.
  - Predicted number of total cases on Feb 24th are 79776 and actual cases are 79561.
- Data size: Because it is a novel disease we only have few months of data available to us. With good amounts of data the predictions can get stronger.
- As we can see from the trend of the global confirmed cases in Fig 5 , the number of new cases have been increasing linearly.
  The trend for the global recovery cases in Fig 6 shows

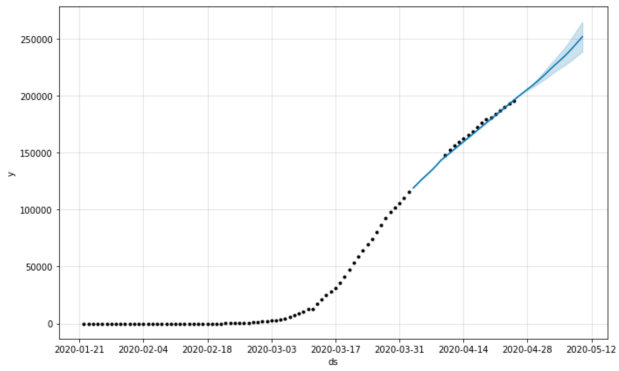

Fig. 4. Forecast of the confirmed cases in Italy

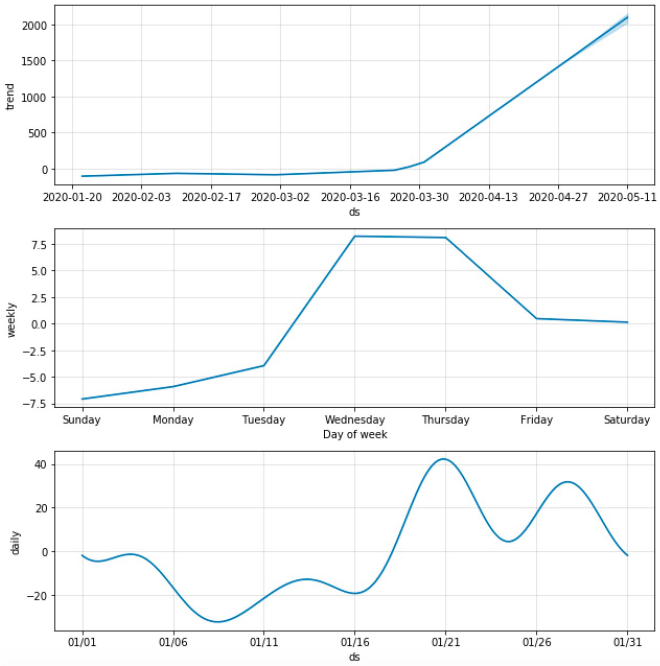

Fig. 5. Forecast components for global confirmed cases

horizon	mse	rmse	mae	mape	mdape	coverage
1 days 12:00:00	9.239277e+10	303961.793247	172062.458902	0.216104	0.142423	0.125000
2 days 00:00:00	5.871824e+10	242318.462530	130724.483819	0.178863	0.053175	0.125000
2 days 12:00:00	1.197342e+11	346026.315672	217386.421936	0.245473	0.179520	0.041667
3 days 00:00:00	1.089686e+11	330103.848943	217240.995583	0.273275	0.223917	0.000000
3 days 12:00:00	1.489523e+11	385943.361146	257066.863600	0.303413	0.424692	0.000000

TABLE I
PERFORMANCE METRICS WITHOUT REGRESSORS

horizon	mse	rmse	mae	mape	coverage	coverage
2 days 00:00:00.000000000	6164.983	78.5174	58.19829	0.560145	0	0.125000
2 days 12:00:00.000000000	5440.993	73.76309	52.57491	0.635867	0	0.125000
3 days 00:00:00.000000000	10764.61	103.7526	74.85328	0.632503	0	0.041667
3 days 12:00:00.000000000	9477.031	97.35004	68.65318	0.690608	0	0.000000
4 days 00:00:00.000000000	17180.12	131.0729	98.95908	0.673324	0	0.000000

TABLE II
PERFORMANCE METRICS WITH REGRESSORS

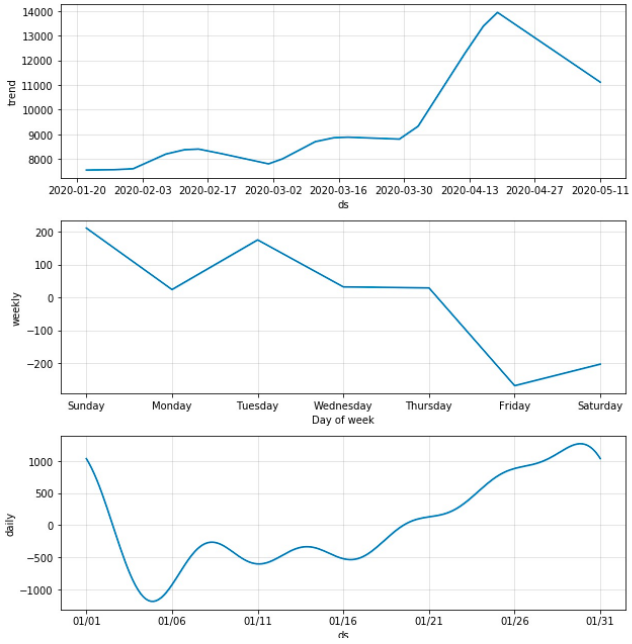

Fig. 6. Forecast components for global recovery cases

that it slowly increases till peak and decreases later. This could be because the rate at which the confirmed cases is increasing is very high compared to the recovery

- The outbreak of Covid-19 is developing into a major international crisis, and it's starting to influence important aspects of daily life.A strong model that predicts how the virus could spread across different countries and regions may be able to help mitigation efforts. The goal of this project is to build a model that predicts the progression of the virus throughout future.

D. Stage4 - User Interface.

The modes of user interaction with the data:
The user is expected to input the type of case like confirmed, death or recovered and also if he/she likes to see the predictions on a global scale or specific for a country. If the user is interested in a country level view he/she should input the name of the country.

- If the user tries to given an invalid input an error message pops up along with the explanation as to why the violation took place.
- If all the input given by the user is correct we then display the forecast plot, weekly and daily components of the requested case type.

II. PROJECT HIGHLIGHTS.

Challenges:

The main challenge we faced when implementing the model was that the existing data-set was small and for only the past few months as this is a new disease. On the other hand, the population is vastly compartmentalized and it is increasingly difficult to accurately predict how and when the disease will spread.

Conclusion:

We implemented an algorithm to calculate and predict the intensity of the outbreak and the number of people affected, recovered or dead from it. The user can decide whether they want to view the country-wide result or a global one.

- The trends show that the severity of the virus began increasing around Mid-March and it has been growing ever since.

- The number of people affected by covid-19 usually peaks around the weekend and goes down mid-week.

Future Extensions:

The project can further be updated with an UI to increase readability. The UI could have features like zooming and panning to integrate results in one page, it can also be made to have interactive graphs where the user can click on different parts of globe to close-in on the location of their preference instead of typing it in as an input. As mentioned earlier, the data can be compartmentalized based on gender, age group, race and current location. We can then predict the spread of virus based on these compartments and analyse the set of people affected most,.

Acknowledgements:

We would like to thank our professor Antonio Miranda for this great opportunity to have hands-on work on the subject and also for the valuable support and feedback for this project.

References:

Daily cases report collected by John Hopkins University.