COVID-19 Predictions in India

Anuj Tanwar

Abstract

COVID-19 outbreaks don’t only impact lives of people but disrupts economy and healthcare infrastructure of the county. So, it is important to study the pandemic and predict potential future infectious disease outbreaks. A comprehensive understanding of the spread and past count of cases will help in predicting recent future and enable administration by giving a heads-up to stay alert and can provide sufficient time to facilitate support interventions.

Keywords: Covid-19, Predictions

COVID-19 Predictions in India

India has seen an increase in COVID-19 case again during the start of year 2022. A simple google search can tell us that on April 4th, there were only 795 new cases, on Apr 14th there were 949 and in last week count of new cases have been on constant rise and going above 2000 per day now. This trend shows the cases have been rising. As per Ministry of Health and Family Welfare Government of India (https://www.mohfw.gov.in/) there had already been more than half a million documented deaths due to COVID and active cases are on rise. Government of India has also issued new guidelines and restrictions. Referring to the below google graph, we can see that COVID spread had increased in India during the summer months in last 2 years. 2022 summer season has already started in India, so the concerning questions that we have are:

1. Are we going to have another wave of COVID?
2. If so, then how severe it can be?
3. When will we see the peak?

These are some of the questions that this project will target to answer. There will be more questions that will be answered during the project to study the past trends such as :

1. How long did the previous waves last?
2. What was the trend in death toll every day?
3. What was the cured trend

I will use data mining techniques learnt in this course to study the data and make prediction on it.

Why it is important to solve this problem?

Below are some of the reasons why it is very important and useful to predict COVID-19 trends:

1. It saves lives by keeping the numbers low
2. Reduces impact on country's economics. International Monetary Fund (IMF) estimated that median global GDP dropped by 3.9% from 2019 to 2020, making it the worst economical downturn since great depression.
3. Advance predictions will allow healthcare sector to be prepared for drastic rise in cases.
4. Predictions can help in determining type (partial or complete) and timeline of lockdowns, if required.
5. Help supply chain in managing and distribution by providing estimates on demand of products such as PPE, ventilators, sanitizer, etc

Data Source

COVID-19 Data is collected from below sources:

1. **Datameet** website. It is community collected, cleaned and organized COVID-19 datasets about India, sourced from different government websites which are freely available. https://projects.datameet.org/covid19/ https://github.com/datameet/covid19/tree/master/data
2. Ministry of Health and Family Welfare Government of India: https://www.mohfw.gov.in/

EDA Summary

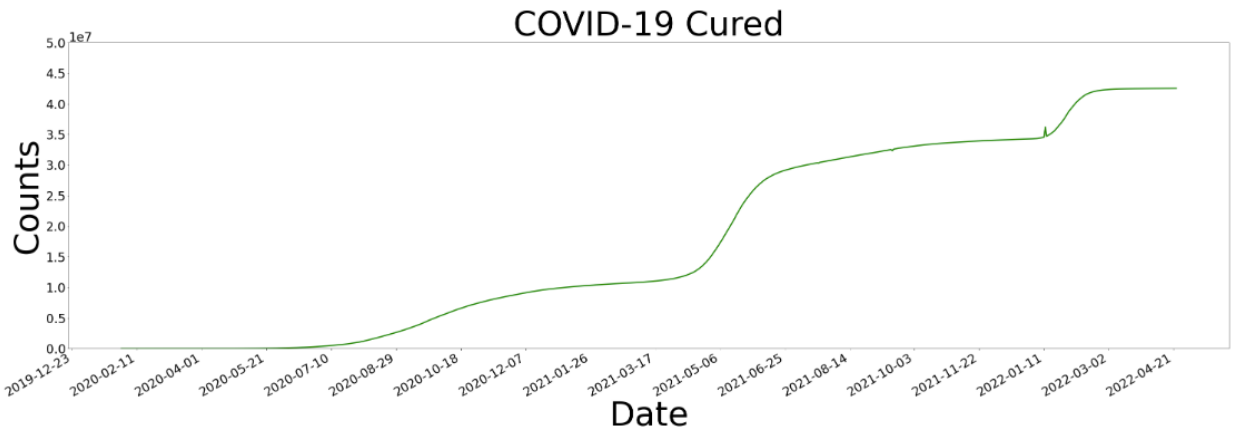
Data Collection: Data was collected from all_totals.json file from datameet git location https://github.com/datameet/covid19/tree/master/data.

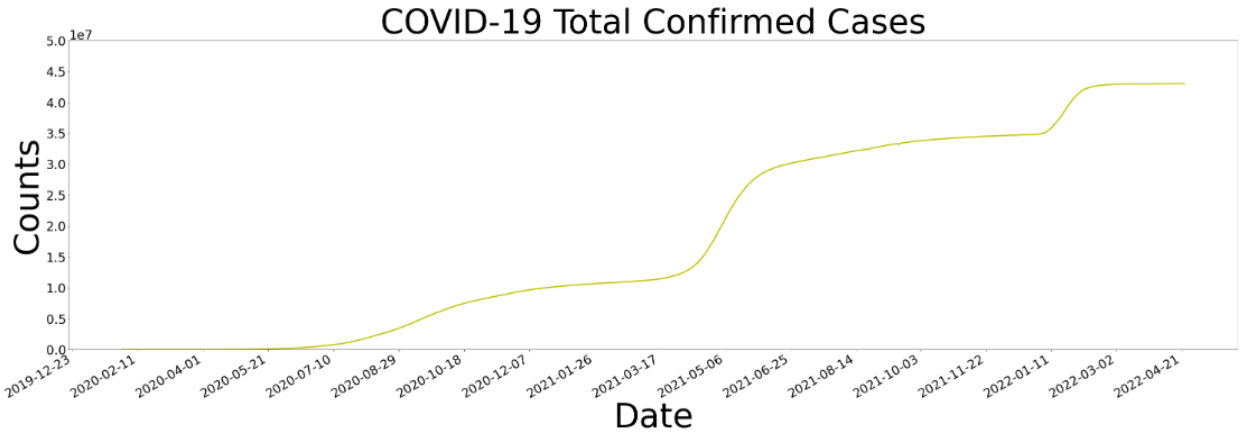
Data Cleaning: Keys and values were read separately from JSON dataset and then combined into a dataframe. At this point, dataframe has multiple lines for each date, one row for each type(active cases, deaths, cured, total confirmed cases). This need dataframe to be pivoted on date field to convert types to individual columns as shown below.

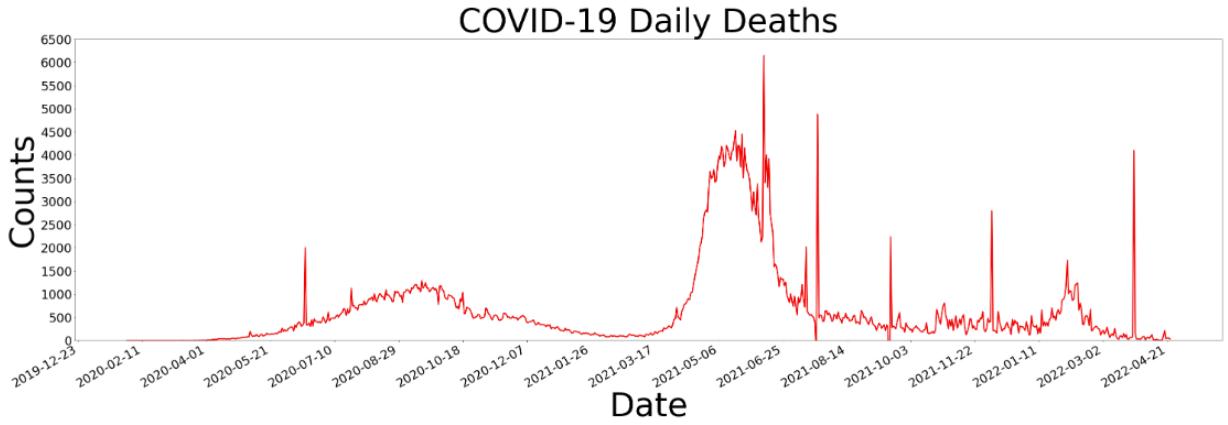
	dt	type	value
0	2020-01-30	active_cases	1
1	2020-01-30	cured	0
2	2020-01-30	death	0
3	2020-01-30	total_confirmed_cases	1
4	2020-02-02	active_cases	2
...	...	...	...
3347	2022-04-22	total_confirmed_cases	43052425
3348	2022-04-23	active_cases	15079
3349	2022-04-23	cured	42517724
3350	2022-04-23	death	522149
3351	2022-04-23	total_confirmed_cases	43054952

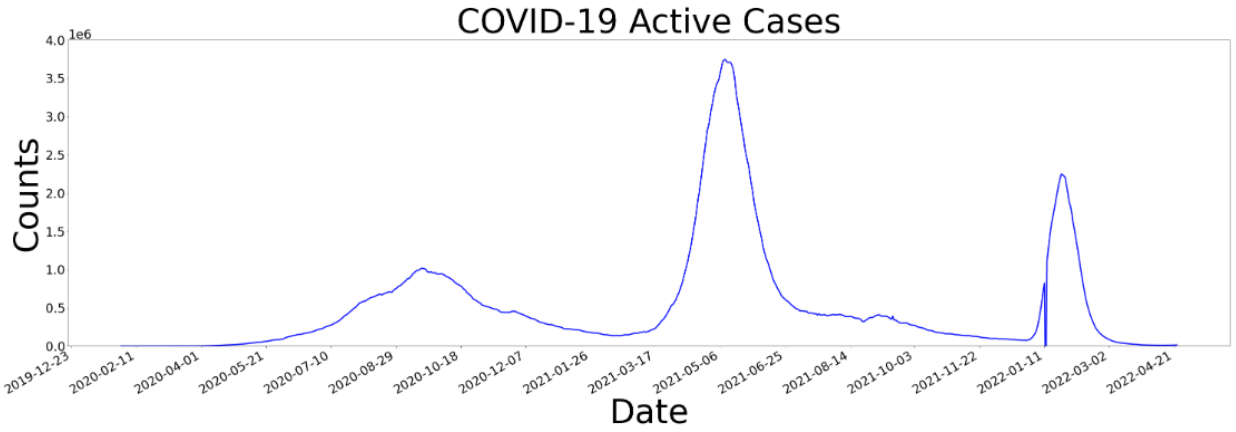
dt	active_cases	cured	death	total_confirmed_cases
2020-01-30	1.0	0.0	0.0	1.0
2020-02-02	2.0	0.0	0.0	2.0
2020-02-03	3.0	0.0	0.0	3.0
2020-03-02	5.0	0.0	0.0	5.0
2020-03-03	6.0	0.0	0.0	6.0
...	...	...	...	...
2022-04-19	11860.0	42511701.0	521966.0	43045527.0
2022-04-20	12340.0	42513248.0	522006.0	43047594.0
2022-04-21	13433.0	42514479.0	522062.0	43049974.0
2022-04-22	14241.0	42516068.0	522116.0	43052425.0
2022-04-23	15079.0	42517724.0	522149.0	43054952.0

Univariate Time Series: Each variable was plotted in time-series to study the trend.




COVID-19 Total Confirmed Cases
1e7
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Counts
2019-12-23
2020-02-11
2020-04-01
2020-05-21
2020-07-10
2020-08-29
2020-10-18
2020-12-07
2021-01-26
2021-03-17
2021-05-06
2021-06-25
2021-08-14
2021-10-03
2021-11-22
2022-01-11
2022-03-02
2022-04-21
Date


COVID-19 Daily Deaths
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Counts
2019-12-23
2020-02-11
2020-04-01
2020-05-21
2020-07-10
2020-08-29
2020-10-18
2020-12-07
2021-01-26
2021-03-17
2021-05-06
2021-06-25
2021-08-14
2021-10-03
2021-11-22
2022-01-11
2022-03-02
2022-04-21
Date


COVID-19 Active Cases
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Counts
2019-12-23
2020-02-11
2020-04-01
2020-05-21
2020-07-10
2020-08-29
2020-10-18
2020-12-07
2021-01-26
2021-03-17
2021-05-06
2021-06-25
2021-08-14
2021-10-03
2021-11-22
2022-01-11
2022-03-02
2022-04-21
Date

Data Preparation

1. Joined case dataset with Vaccine data(mohfw_vaccination_status.json) to get additional information.
2. Dropped features not useful for analysis.
3. Transformed features such as report_date and used it as key to join case dataset with vaccine dataset.
4. Dropped records with invalid values such as counts cannot be negative.
5. Used imputer to fill in NANs with mean.
6. Used MinMaxScaler() to rescale each column.
7. Filled missing data with mode of respective column.

Model building and evaluation

- Model Used: Long short-term memory (LSTM). It is an artificial neural network used in the fields of artificial intelligence and deep learning.
- I have used sequential model from tensorflow.keras package.
- Added a long short-term memory layer with 100 memory units
- Used rectified linear activation function (RELU) which will output the input directly if it is positive, otherwise, it will output zero.
- Used 20% dropout.
- Compiled the model with adam optimizer

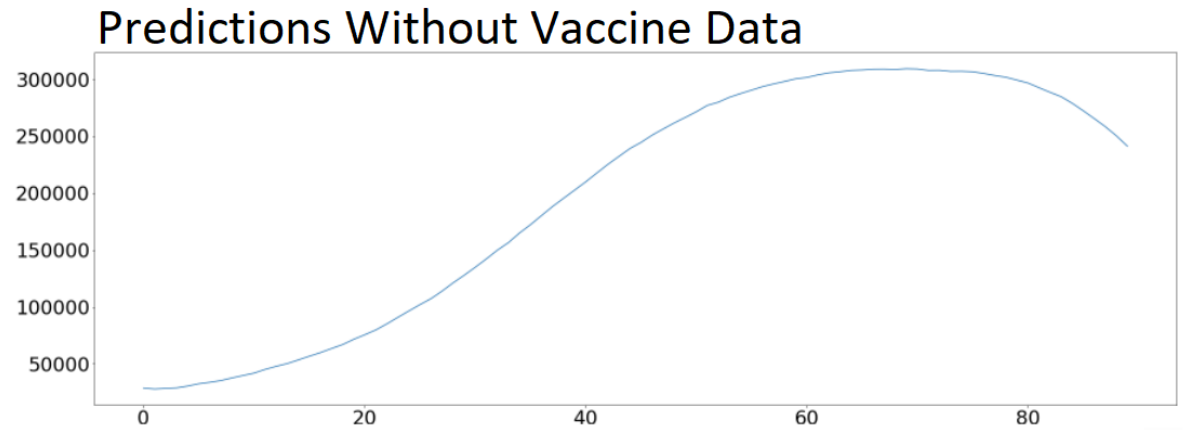
```
Model: "sequential_31"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_29 (LSTM)              (None, 100)               76400

 dropout_83 (Dropout)        (None, 100)               0

 dense_82 (Dense)            (None, 100)               10100

 dropout_84 (Dropout)        (None, 100)               0

 dense_83 (Dense)            (None, 90)                9090

 dropout_85 (Dropout)        (None, 90)                0

 dense_84 (Dense)            (None, 1)                 91

=================================================================
Total params: 95,681
Trainable params: 95,681
Non-trainable params: 0
_________________________________________________________________
```

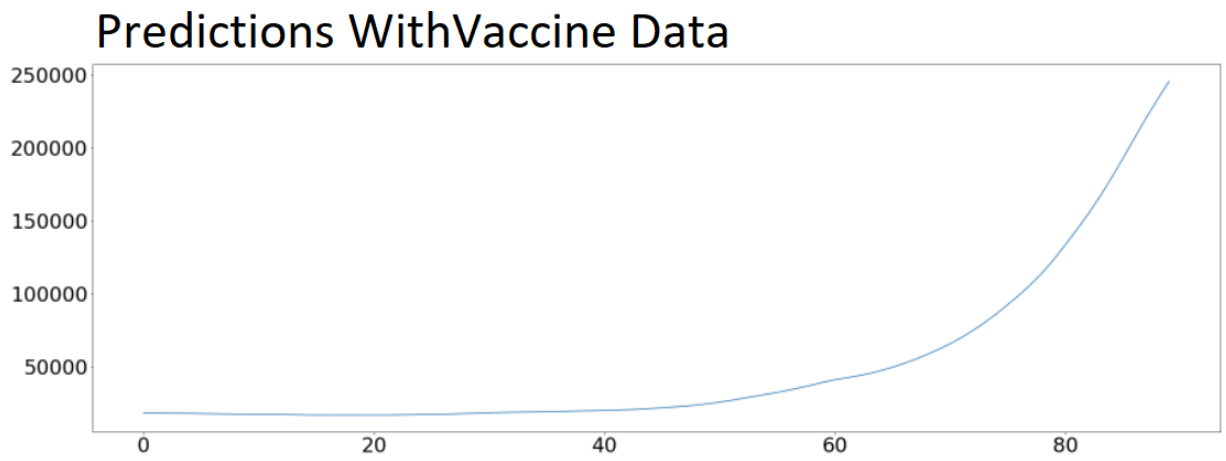
- Data Fitting used EarlyStopping class from keras callbacks. This enables model to stop training when a metric has stopped improving.

Conclusion

1. What does the Model tell us?

Model clearly suggest that vaccine reduces covid counts. Without first dose the 90 days total case prediction is 17.7 million but with just first dose it comes down to 4.5 million.





2. Is the model ready to be deployed?

Yes, the model is ready to be deployed.

3. What are the recommendations?

Case estimations are purely on past data, Future numbers depends on lot of factors such as restrictions (mask mandate, social distancing, etc) and vaccine dosage and boosters. Those factors can affect the numbers. Model considered only first vaccine dose and based on that we can recommend encouraging public to take vaccine and booster shots whichever is applicable.

4. Challenges/New Opportunities that need to be explored?

Model can be enhanced to consider below factors in future predictions:

- Second and subsequent vaccine dosage
- Enforced Restrictions and guideline such as mask mandates, lock downs, number of people allows in public transportation such as buses and trains
- Weather conditions such as summer or winter counts

References

https://www.kff.org/global-health-policy/issue-brief/economic-impact-of-covid-19-on-pepfar-countries/