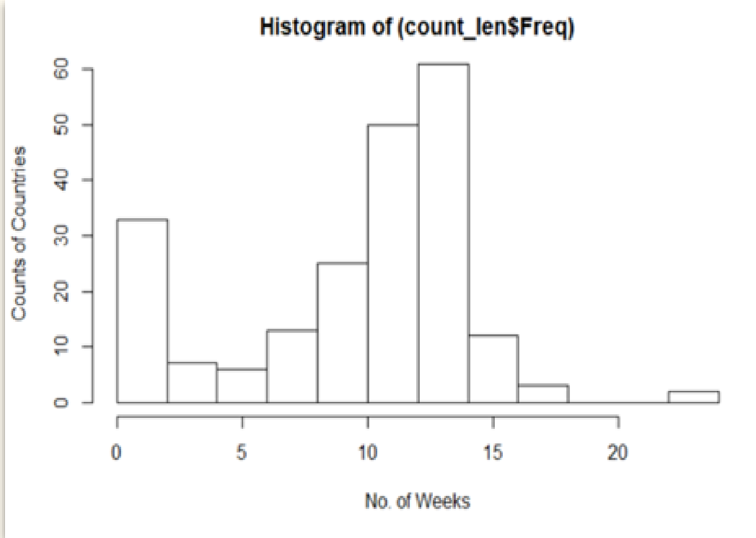
Pure ML Approaches for COVID-19 Peak Prediction

Motivation and Problem Statement

- Covid-19 pandemic has led to the the entire world be in a state of turmoil and has significantly affected all walks of life.
- Predicting the peak of cases is significant for any country to effectively fight against the disease and ensure minimal damage.
- In this work, we try to determine the effectiveness of Machine Learning models to predict the peak for Covid-19 cases for India.

Details of the Dataset

- We use the ‘Our World In Data’ dataset, which is a standardized collection of data from the ECDC.
- The dataset provides daily Covid-19 data observed in countries across the world and other relevant details like population density, GDP per capita, Median age of the population.
- Source: https://ourworldindata.org/coronavirus-source-data

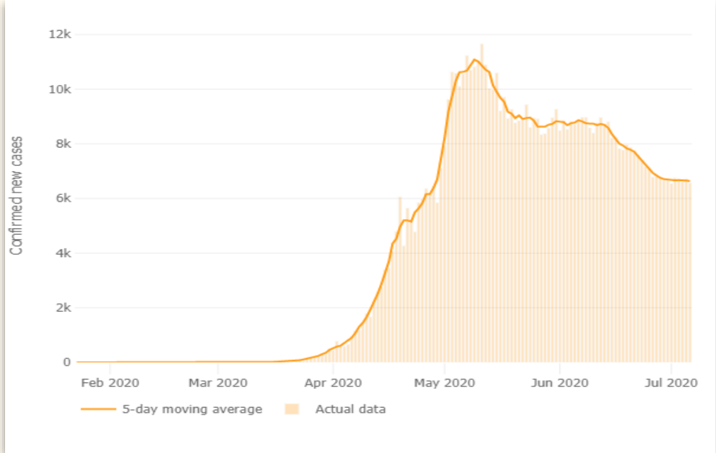


Features in the Data Used

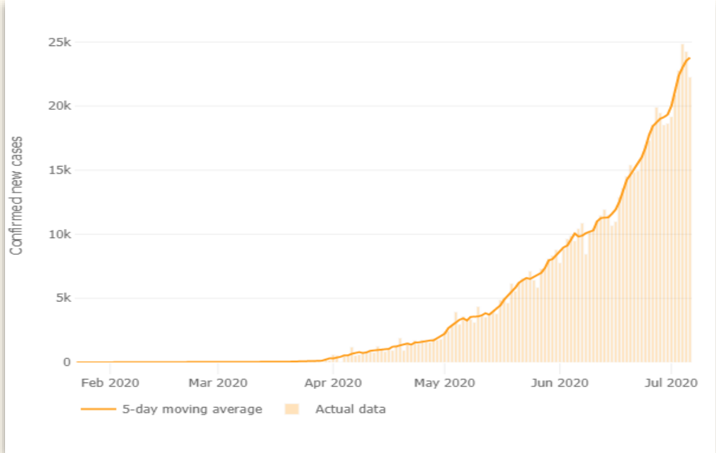
- Sequential features in the data used - Total cases, New Cases, Total Cases per Million, New Cases per Million, Computed 7-Day Moving Averages of New Cases
- Non-sequential features in the data used (Covariates)- Population, Population Density, GDP per Capita

Defining a Peak

The week with the maximum number of new cases in a country.



New Cases Graph for Russia



New Cases Graph for India

Models Used

Linear Regression on week number in which the country peaks

Linear Regression to predict the curve for the rates of change of new cases

LSTM for regression on daily rates of change to predict peak of cases

LSTM for regression on day number on which the country peaks

LSTM based sequence classifier with new cases

LSTM based sequence classifier with rate of change

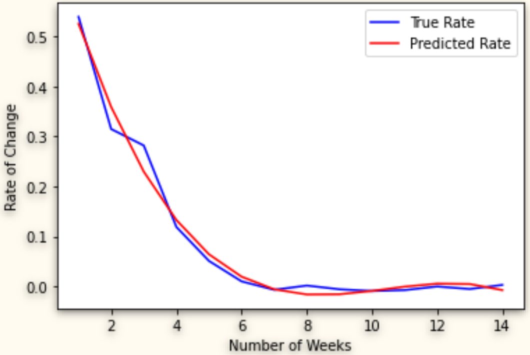
Linear regression on week number in which the country peaks

- One model trained on the data of all the countries
- X: 10 Features corresponding to the first 10 weekly values of new cases
- Y: The week number from the start of the data when the country peaks
- The results were not satisfactory as expected. Training RMSE turned out to be 3 weeks and Validation RMSE was more than 300.

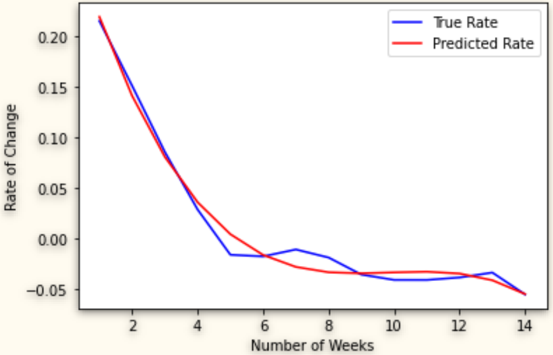
Linear regression on rate of change to predict the week of peak

- We try to predict the curve of the rates of change of new cases
- X: Week number and its powers
- Y: Rate of change of new cases for the corresponding week
- We observe that the rates of change follow a logarithmic decrease
- We obtain a good performance for most of the known countries but fail to perform well for India

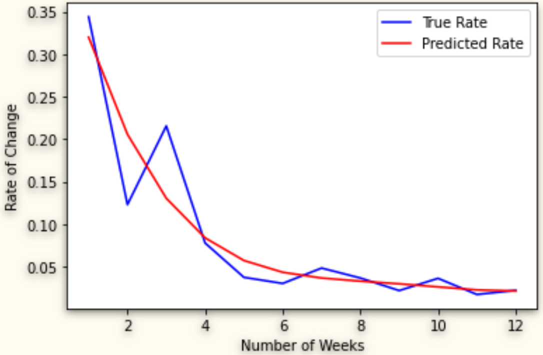
Linear regression on rate of change to predict the week of peak



U.S.A.

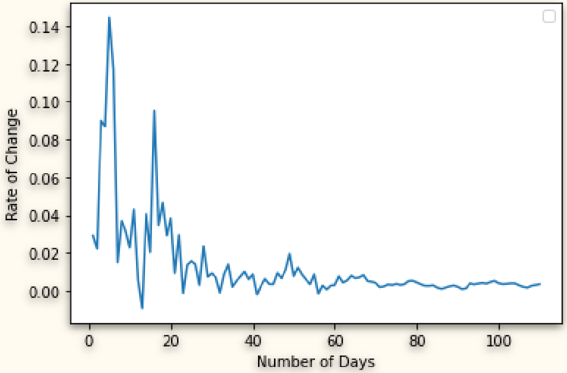


Italy

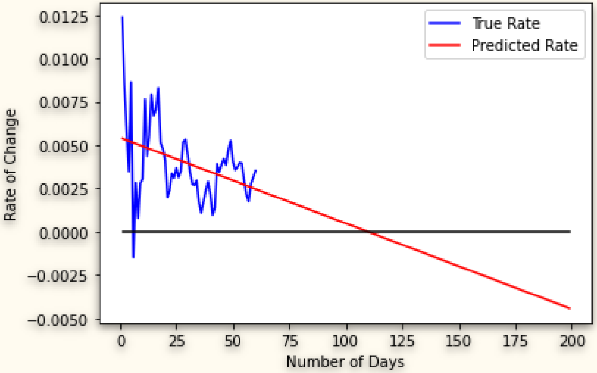


India

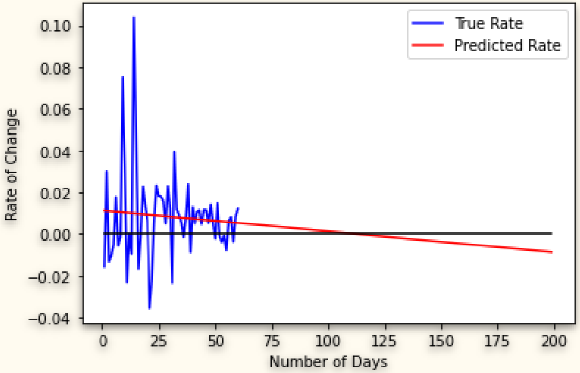
Linear regression on rate of change to predict the day of peak



India's Rate of Change Curve



Prediction for India



Prediction for a country similar to India

LSTM for regression on future rates of change

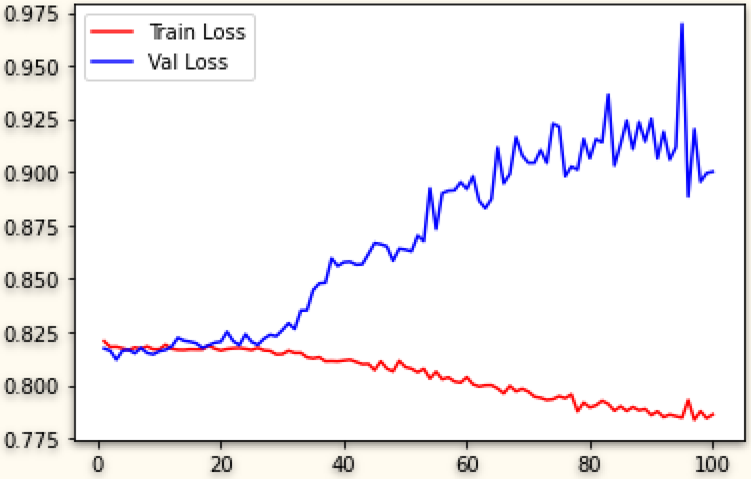
- The premise was to see if the past rates of change of new cases could be used to predict the future rates
- X: Last 20 days' rates of change of new cases
- Y: Next 7 days' rates of change of new cases
- Covariates like population density and GDP per capita are concatenated as to the output of LSTM layer.
- We fail to train the model correctly and believe that the recent past does not have enough information to make the predictions for the future.

LSTM for regression on day of peaking

- We try to use the past rate of changes to predict how long will it take for the country to peak
- X: Last 20 days’ rates of change of new cases
- Y: Number of days from the current timestamp when the country peaks
- Note- The split between training and validation set is made keeping in mind the cultural and economic factors for India.

LSTM for regression on day of peaking

We observe that though training loss goes down, the validation loss doesn't. A possible explanation is that the training and validation set differ too much from each other



LSTM Based Sequence Classifiers

- Models to predict if the week after n weeks from now is the week where the covid new cases will be max for a country, i.e. it’ll peak

Further, we find the peak of each country and take the last (x+1)*7 to 7 days previous to peak as sequential features for the LSTM model with the following six features:

- **Total_cases,New_cases,Total_cases_per_million,New_cases_per_million**
- **moving_average(new_cases),moving_average(new_cases_per_million)**

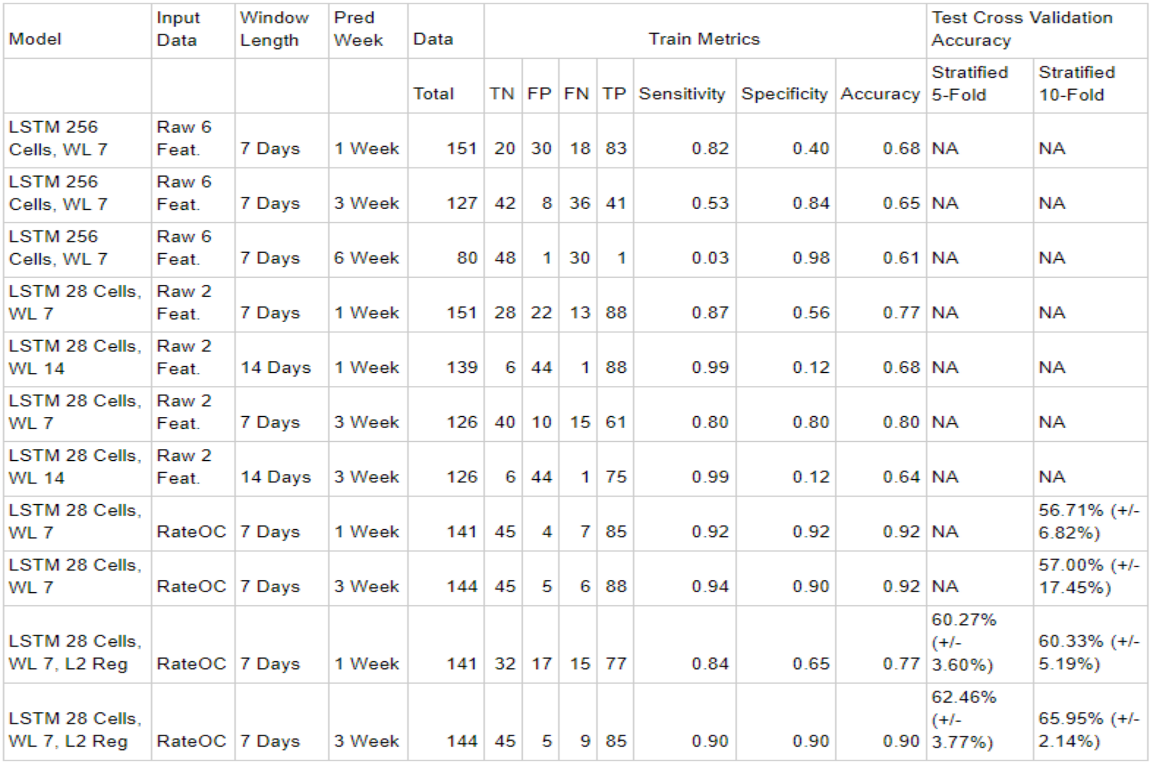
So as to form,

X :n_countries*(x+1)7*6 features for training

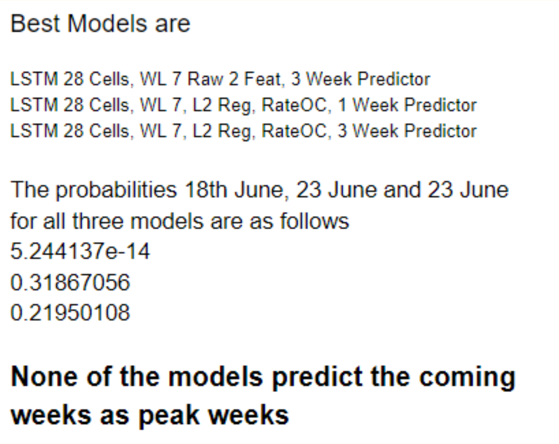
Y : 0, 1 classification

On some models we only take rate of change of **moving_average(new_cases) & moving_average(new_cases_per_million)**

Results (LSTM Based Sequence Classifiers)



Model	Input Data	Window Length	Pred Week	Data	Train Metrics							Test Cross Validation Accuracy	
				Total	TN	FP	FN	TP	Sensitivity	Specificity	Accuracy	Stratified 5-Fold	Stratified 10-Fold
LSTM 256 Cells, WL 7	Raw 6 Feat.	7 Days	1 Week	151	20	30	18	83	0.82	0.40	0.68	NA	NA
LSTM 256 Cells, WL 7	Raw 6 Feat.	7 Days	3 Week	127	42	8	36	41	0.53	0.84	0.65	NA	NA
LSTM 256 Cells, WL 7	Raw 6 Feat.	7 Days	6 Week	80	48	1	30	1	0.03	0.98	0.61	NA	NA
LSTM 28 Cells, WL 7	Raw 2 Feat.	7 Days	1 Week	151	28	22	13	88	0.87	0.56	0.77	NA	NA
LSTM 28 Cells, WL 14	Raw 2 Feat.	14 Days	1 Week	139	6	44	1	88	0.99	0.12	0.68	NA	NA
LSTM 28 Cells, WL 7	Raw 2 Feat.	7 Days	3 Week	126	40	10	15	61	0.80	0.80	0.80	NA	NA
LSTM 28 Cells, WL 14	Raw 2 Feat.	14 Days	3 Week	126	6	44	1	75	0.99	0.12	0.64	NA	NA
LSTM 28 Cells, WL 7	RateOC	7 Days	1 Week	141	45	4	7	85	0.92	0.92	0.92	NA	56.71% (+/- 6.82%)
LSTM 28 Cells, WL 7	RateOC	7 Days	3 Week	144	45	5	6	88	0.94	0.90	0.92	NA	57.00% (+/- 17.45%)
LSTM 28 Cells, WL 7, L2 Reg	RateOC	7 Days	1 Week	141	32	17	15	77	0.84	0.65	0.77	60.27% (+/- 3.60%)	60.33% (+/- 5.19%)
LSTM 28 Cells, WL 7, L2 Reg	RateOC	7 Days	3 Week	144	45	5	9	85	0.90	0.90	0.90	62.46% (+/- 3.77%)	65.95% (+/- 2.14%)



Best Models are

LSTM 28 Cells, WL 7 Raw 2 Feat, 3 Week Predictor
LSTM 28 Cells, WL 7, L2 Reg, RateOC, 1 Week Predictor
LSTM 28 Cells, WL 7, L2 Reg, RateOC, 3 Week Predictor

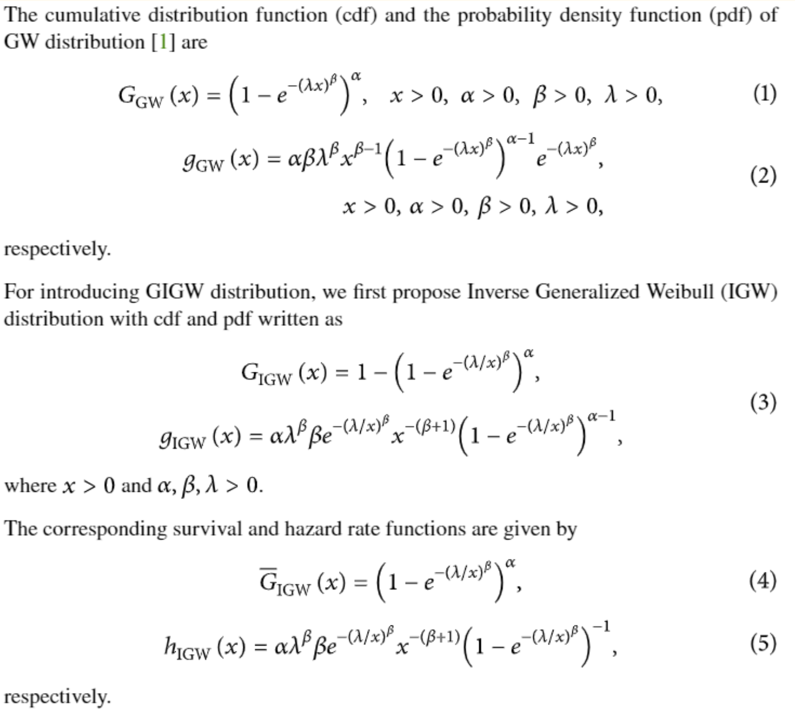
The probabilities 18th June, 23 June and 23 June for all three models are as follows
5.244137e-14
0.31867056
0.21950108

None of the models predict the coming weeks as peak weeks

Weibull Distribution Curve (Simple Non-ML Ap.)

While machine learning and linear fit like linear regression may not give a good fit for coronavirus cases, we could make mathematical inferences/assumptions from observed data.

The one thing observed among countries that have peaked is the best mathematical distribution fitting for coronavirus is Generalized Weibull Distribution, which is given by as follows:-



The cumulative distribution function (cdf) and the probability density function (pdf) of GW distribution [1] are

$$G_{\mathrm{GW}}(x) = \left(1 - e^{-(\lambda x)^{\beta}}\right)^{\alpha}, \quad x > 0,\ \alpha > 0,\ \beta > 0,\ \lambda > 0, \tag{1}$$

$$\begin{gathered} g_{\mathrm{GW}}(x) = \alpha\beta\lambda^{\beta}x^{\beta-1}\left(1 - e^{-(\lambda x)^{\beta}}\right)^{\alpha-1}e^{-(\lambda x)^{\beta}}, \\ x > 0,\ \alpha > 0,\ \beta > 0,\ \lambda > 0, \end{gathered} \tag{2}$$

respectively.

For introducing GIGW distribution, we first propose Inverse Generalized Weibull (IGW) distribution with cdf and pdf written as

$$\begin{gathered} G_{\mathrm{IGW}}(x) = 1 - \left(1 - e^{-(\lambda/x)^{\beta}}\right)^{\alpha}, \\ g_{\mathrm{IGW}}(x) = \alpha\lambda^{\beta}\beta e^{-(\lambda/x)^{\beta}}x^{-(\beta+1)}\left(1 - e^{-(\lambda/x)^{\beta}}\right)^{\alpha-1}, \end{gathered} \tag{3}$$

where $x > 0$ and $\alpha, \beta, \lambda > 0$.

The corresponding survival and hazard rate functions are given by

$$\overline{G}_{\mathrm{IGW}}(x) = \left(1 - e^{-(\lambda/x)^{\beta}}\right)^{\alpha}, \tag{4}$$

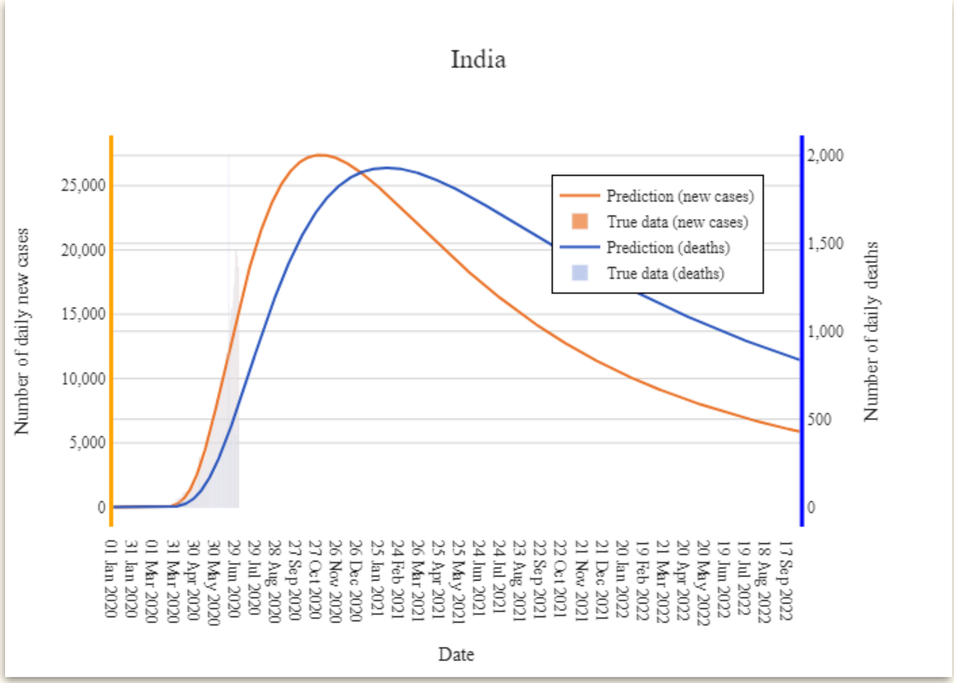
$$h_{\mathrm{IGW}}(x) = \alpha\lambda^{\beta}\beta e^{-(\lambda/x)^{\beta}}x^{-(\beta+1)}\left(1 - e^{-(\lambda/x)^{\beta}}\right)^{-1}, \tag{5}$$

respectively.

Results (Weibull Distribution Curve)

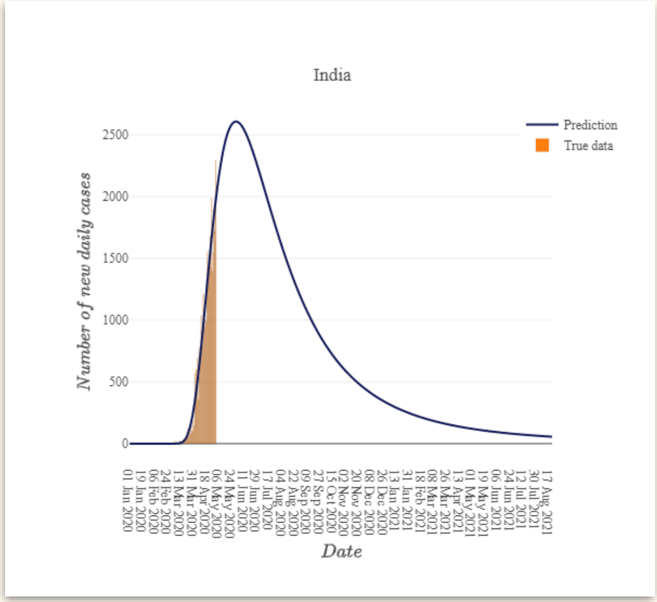
This is however prone would best work assuming that lockdown had limited and minimal effects on the curve which is not the case.

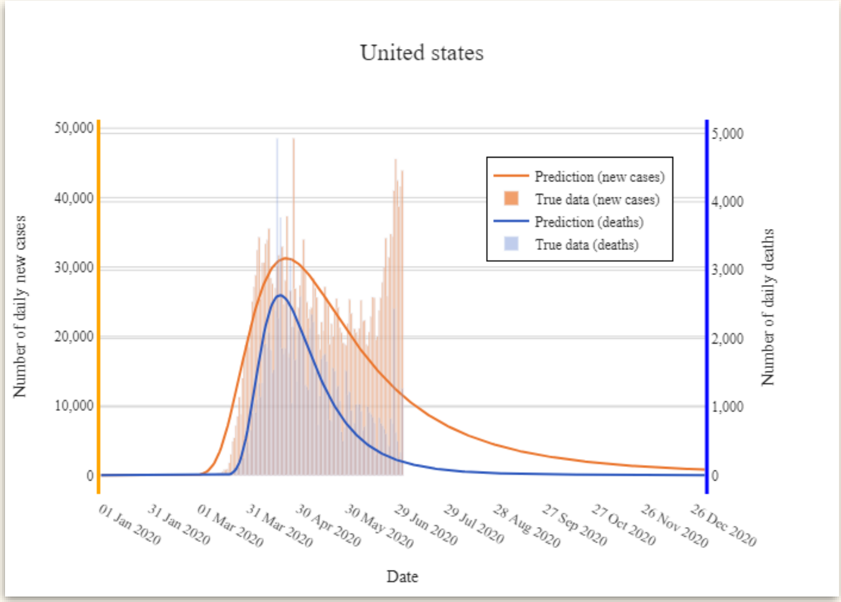
But assuming the same the peak would occur in the first week of November



The same model with data till May 2 (vs the on above I ran on 3 July predicting November Week 1) had predicted the peak to happen around June 2.

Which might even have happened is a continuous lockdown was maintained. It also does not work if we assume an uneven distribution/multiple waves.





Further possibilities / Questions

- Compartmental ML Models
- Statistical inference based models, a Weibull curve based LSTM regression.
- Other possible ways to improve test accuracy in LSTM classifiers
- Models like ARMA, ARIMA and ARIMAX to see if higher orders of difference gives us stationarity

Questions We Have

- Does the Linear Regression approach for the prediction of the curve of the rates seem viable? The input to the model is simple, however, the predictions seem to be somewhat correct.
- Do you have any suggestions to improve the training of the LSTM approach of predicting the day of peaking?
- Could we incorporate the covariates more effectively to lay emphasis on features that are of significance for the problem?

Fin