Covid-19 UW Datathon

Question

How will the number of cases and deaths progress over the next month (30 days) in the US?

Hypothesis

Null Hypothesis: The number of deaths and cases from covid-19 will remain the same over the next 30 days.

Our Alternate Hypothesis: We think that the number of cases and deaths due to covid-19 will keep increasing in the next 30 days.

Assumptions

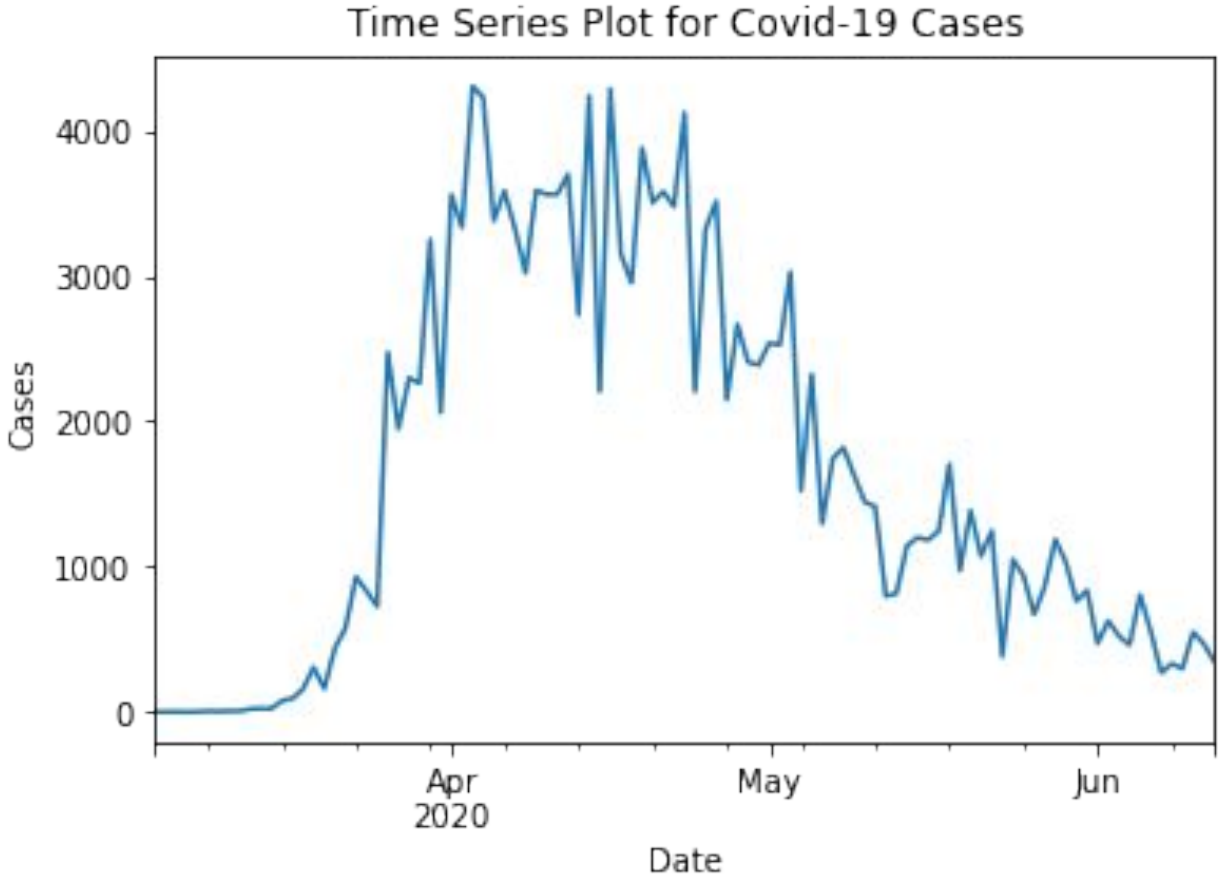
- We realized that the covid dataset is fairly new and it might be too early to observe any seasonality. So, we steered clear of time series models incorporating seasonality like sARIMA and other variants.
  - ASSUMPTION: No seasonality
- Since the covid time series is evolving so fast, the mean, standard deviation and maximum values might change with incoming data. So, we decided against scaling/ normalizing data.
  - ASSUMPTION: No need to standardize
- We observed that the time series for covid ‘cases’ and ‘deaths’ progressed relatively differently. So, we performed separate time series analysis for the two.
  - ASSUMPTION: Separate time series and forecasts for covid ‘cases’ and ‘deaths’

Data Profile

- We took covid-19 data for each of the US states
- Daily time series
- Our focus was on daily covid ‘cases’ and ‘deaths’
- Dates range from 21st January 2020 to 12th June 2020

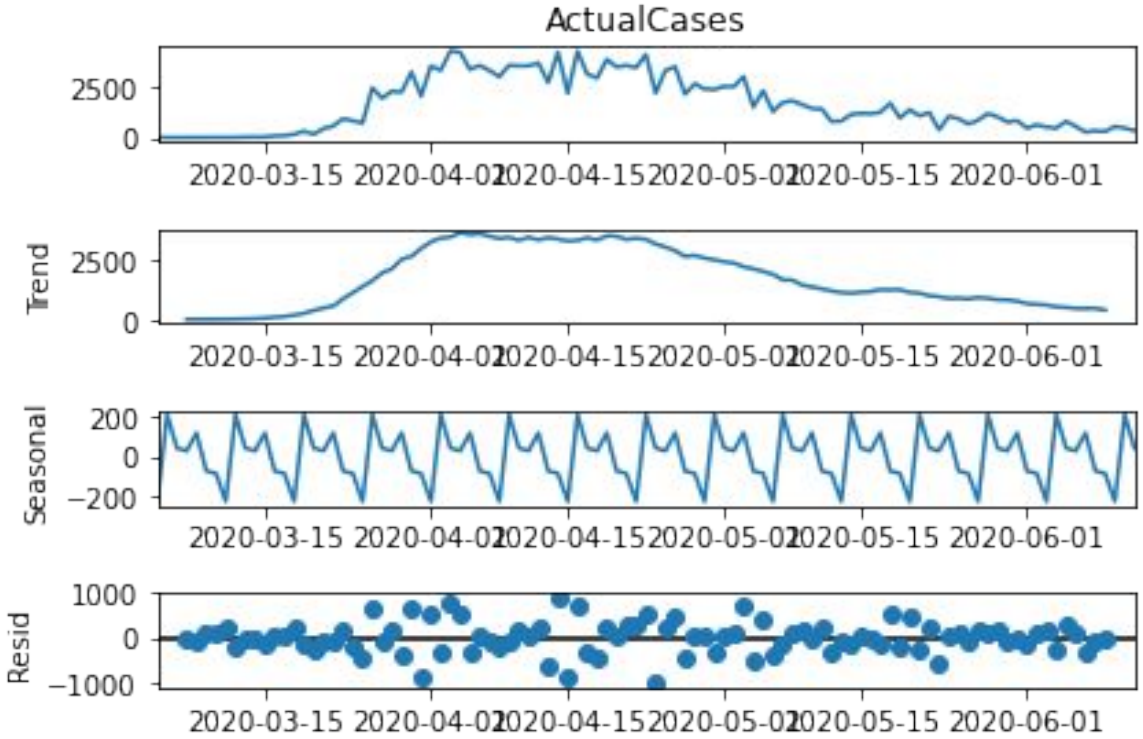
Exploratory Data Analysis

COVID “cases”

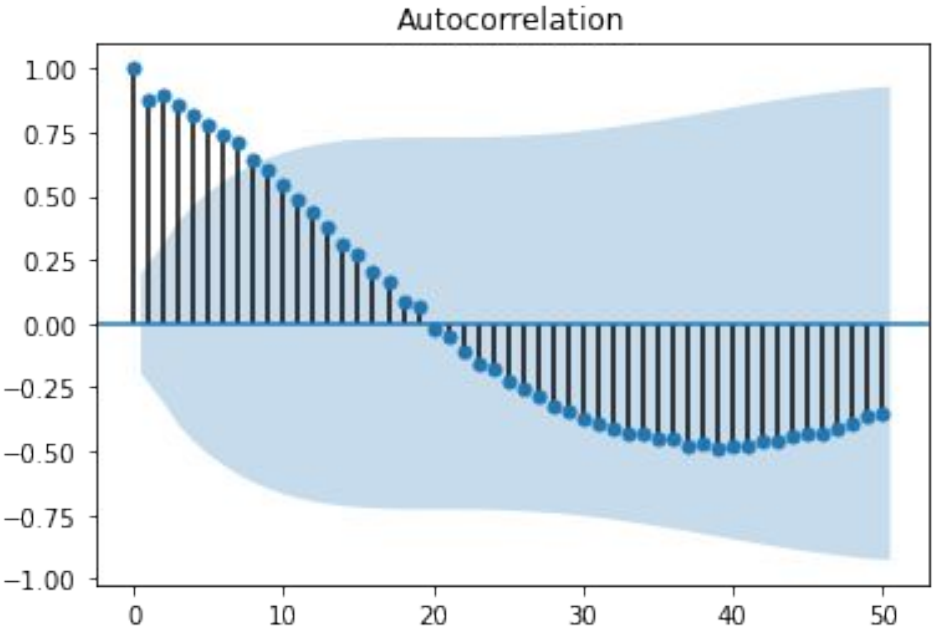

Time Series Plot for Covid-19 Cases
4000
3000
2000
1000
0
Cases
Apr
2020
May
Jun
Date

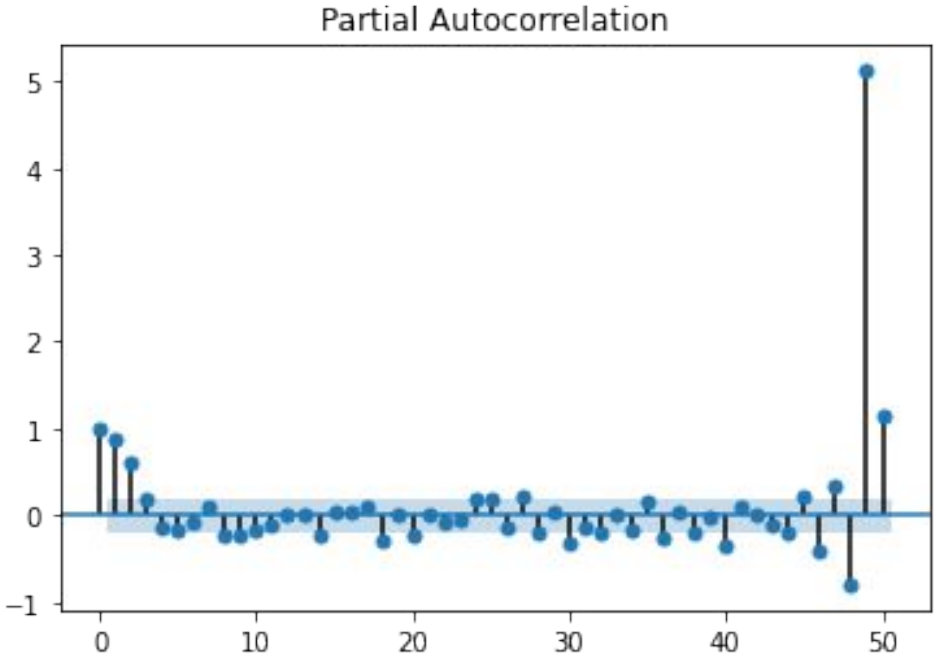
STL decomposition

Here, we can observe trend, seasonality, and residual progression.

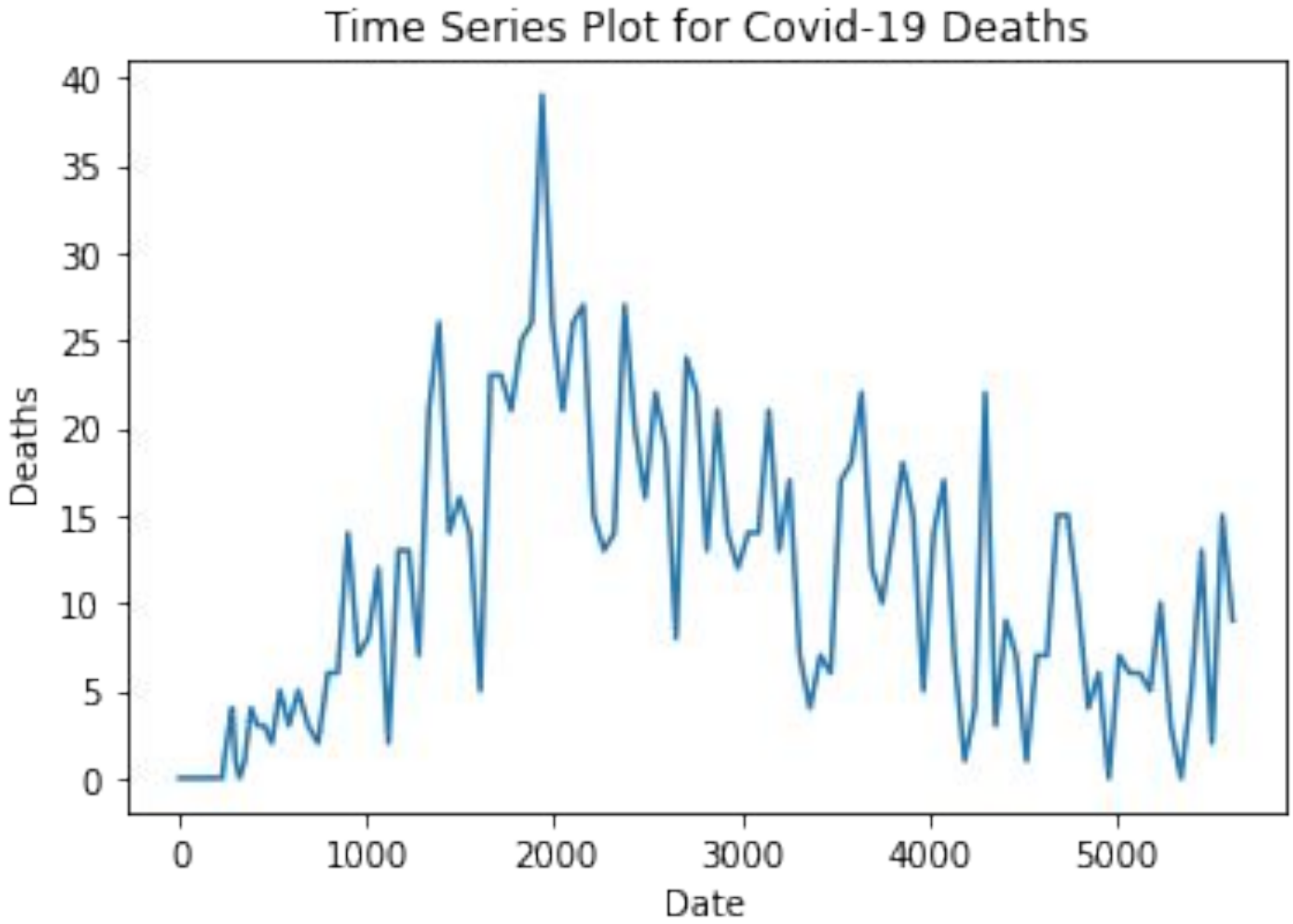


Autocorrelation and Partial Autocorrelation Functions



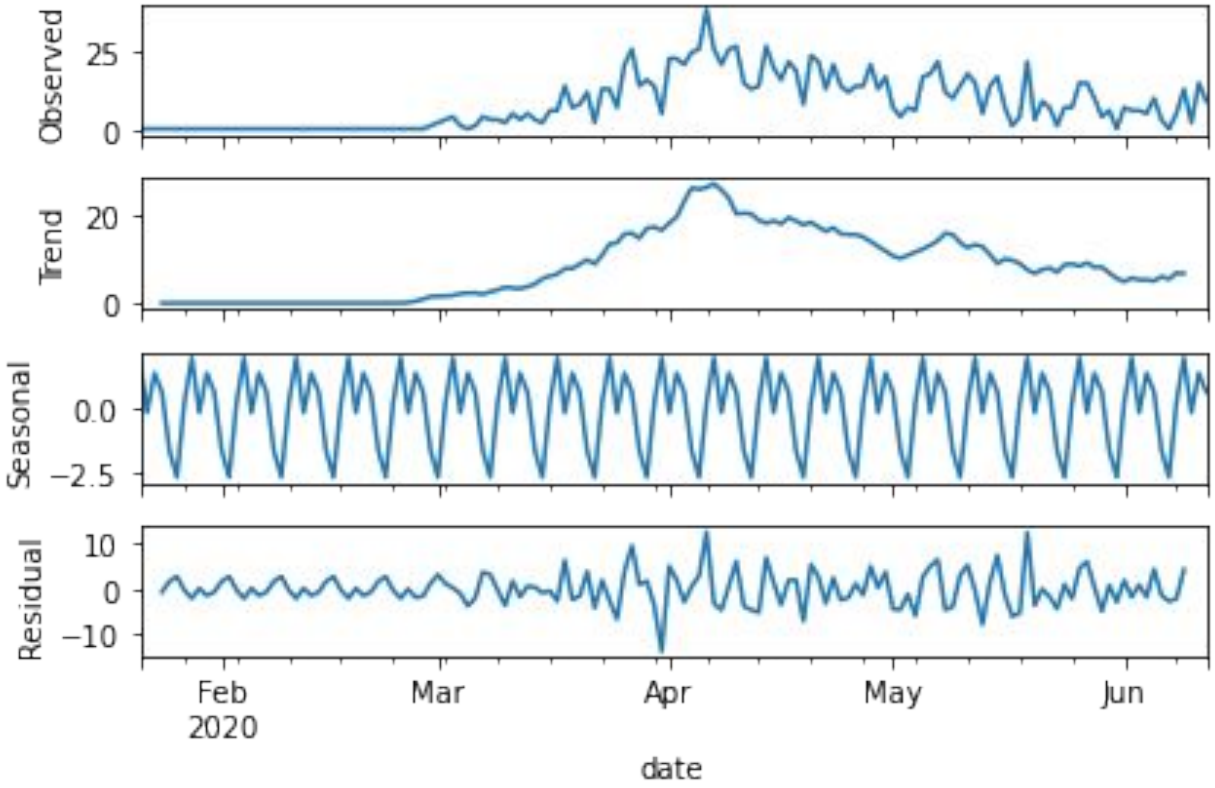


COVID “deaths”

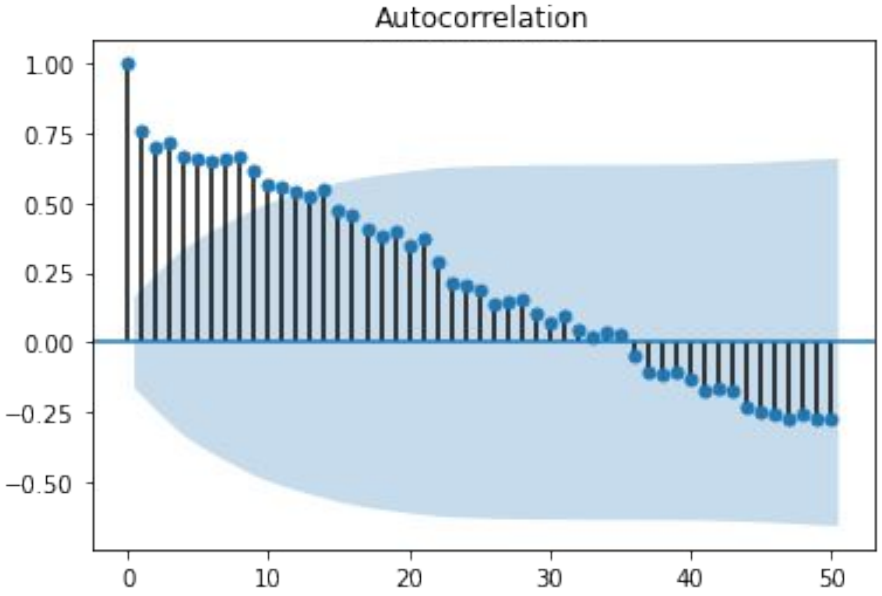

Time Series Plot for Covid-19 Deaths
Deaths
Date
0
5
10
15
20
25
30
35
40
0
1000
2000
3000
4000
5000

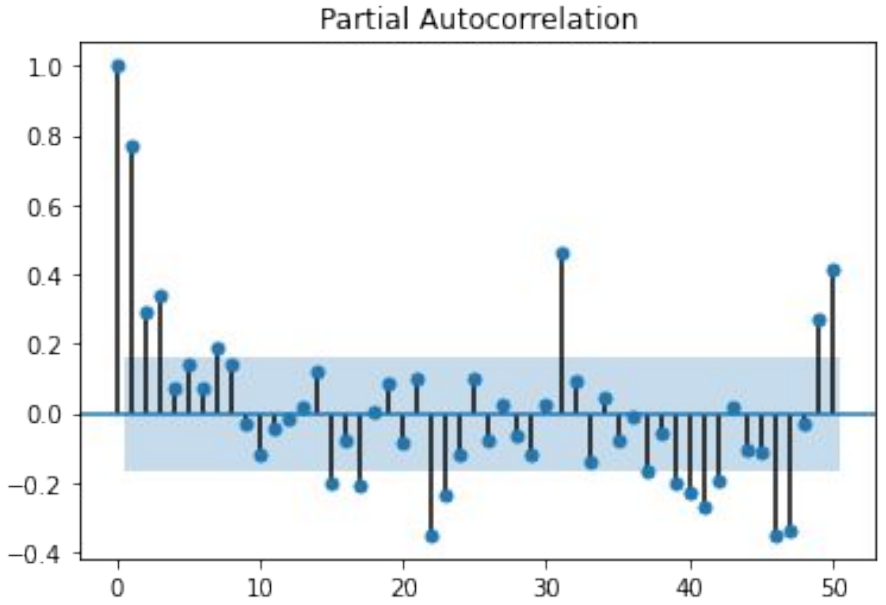
STL decomposition

Here, we can observe trend, seasonality, and residual progression.



Autocorrelation and Partial Autocorrelation Functions





Feature Quality

- **Feature Selection:** We chose to operate on number of ‘cases’ and ‘deaths’.
- **Feature Engineering:** We realized the ‘cases’ and ‘deaths’ for each day were represented as a cumulative sum of cases upto that day. So, we found the distinct number of cases for each day to get a better idea of daily progression.
- **Data Quality:** The data was pretty clean. There were no missing values. Although time series for some states started earlier, since we treat each state separately, this was managed easily.

Problem Solving

- We chose to implement a Long Short Term Memory network or **LSTM.** As the name implies, LSTMs are able to learn long-term dependencies, especially in time series, due to their use of ‘feedback’ structures.
- We also implemented an Auto Regressive Integrated Moving Average or **ARIMA** model. This kind of time series model uses past values of the time series, to forecast future values, making it apt for covid-19 data where seasonality is hardly present to forecast upon.
- Additionally, we implemented a 3rd ensemble of the LSTM and ARIMA models to check whether it performed better.
- **Everything described ahead was done for both covid ‘cases’ and ‘deaths’.**

LSTM Network (model specifications-Covid cases)

- ReLu activation function
- Look-back period of 7 days (1 week) or 15 days (2 weeks)
- Adam optimizer
- 3-layer LSTM with different levels of Dropout for **regularization**
- Select model with minimum validation error for the two look back periods

LSTM Network (model specifications-Covid deaths)

- ReLu activation function
- Look-back period of 1 day or 7 days (1 week)
- Adam optimizer
- 3-layer LSTM with different levels of Dropout for **regularization**
- Select model with minimum validation error for the two look back periods

Hyperparameter Tuning

- **Look-back period:** While choosing an appropriate look-back period, we decided to choose either 1 day, 1 week, 2 weeks or 1 month as our look-back values. After running the model for each of these values, we realized, the higher the look back, the more memory the LSTM has and hence, the higher the accuracy. So, to find a sweet spot where the LSTM did not overfit the data and still remained fairly accurate, we chose a look-back period of 7 days (1 week) and 15 days (2 weeks) for cases and 1 day and 7 days (1 week) for deaths.
- **Choice of number of layers, dropout percentage, number of nodes**: We ran a hyperparameter search with several values for each and came up with the combination with lowest mean square error (MSE).

ARIMA analysis

- We implemented a stepwise ARIMA model to find the best values of p,d,q for our analysis.
- **Criterion to find best ARIMA model: Akaike Information Criterion (AIC)**
- We chose AIC because it assesses both the model's **goodness of fit** as well as its **simplicity**

Choice of Ensemble Model

- We had two options to ensemble: 1. pick the model with minimum error rate for each state **OR** 2. average the predictions from both models
- Since both the models were comparable in accuracy, we created an ensemble by finding the simple average of the **LSTM** and **ARIMA** models.
- We wanted to maintain a uniform ensemble model, so, we decided against choosing the minimum RMSE model for each state.
- Our resulting model was more interpretable.

Model Evaluation

LSTM:

- We used MSE as a loss function to tune our hyperparameters.
- We evaluated the model using both MSE and Mean Absolute Percent Error (MAP).

ARIMA:

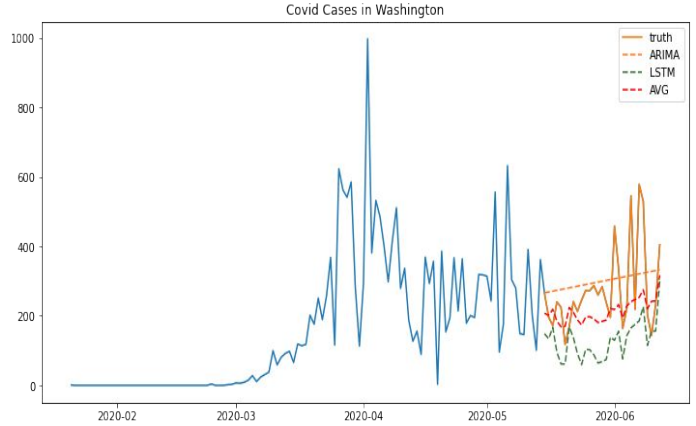
- We used AIC as our metric to find the best hyperparameters (values of p,d,q)
- We evaluated the model using MSE and MAP.

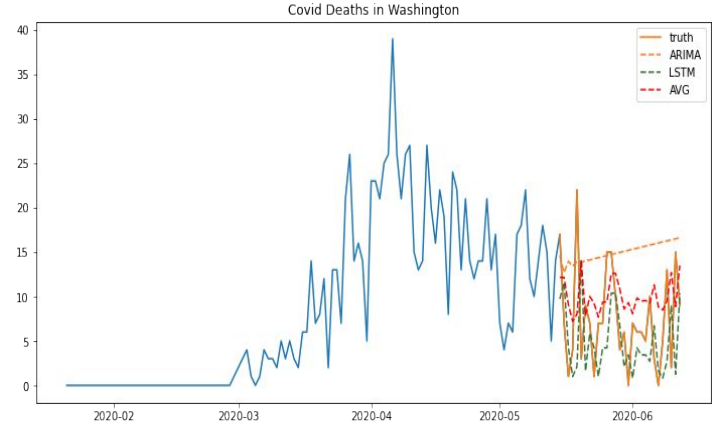
Results

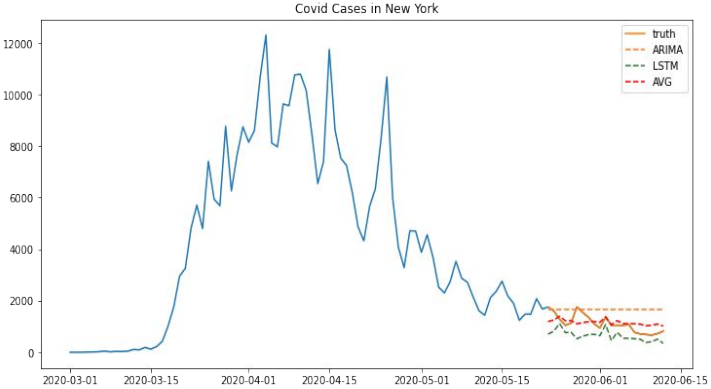
State	Type	Evaluation Metrics	ARIMA	LSTM	Ensemble
New York	Cases	RMSE	651.25	551.11	289.81
New York	Cases	MAP	23.96	22.02	15.82
New York	Deaths	RMSE	75.91	21.85	41.45
New York	Deaths	MAP	8.41	4.03	6.22
Washington	Cases	RMSE	113.24	177.09	119.49
Washington	Cases	MAP	9.54	12.03	9.21
Washington	Deaths	RMSE	9.32	6.94	6.27
Washington	Deaths	MAP	2.88	2.34	2.29

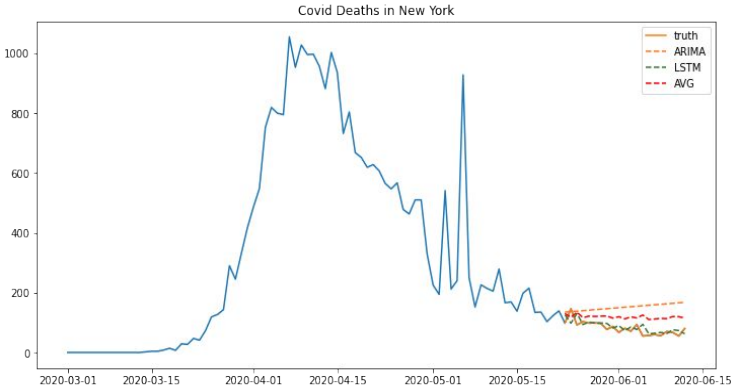
Comparing Forecasts for ‘cases’ and ‘deaths’

Covid cases and death forecasts (Washington and NY)





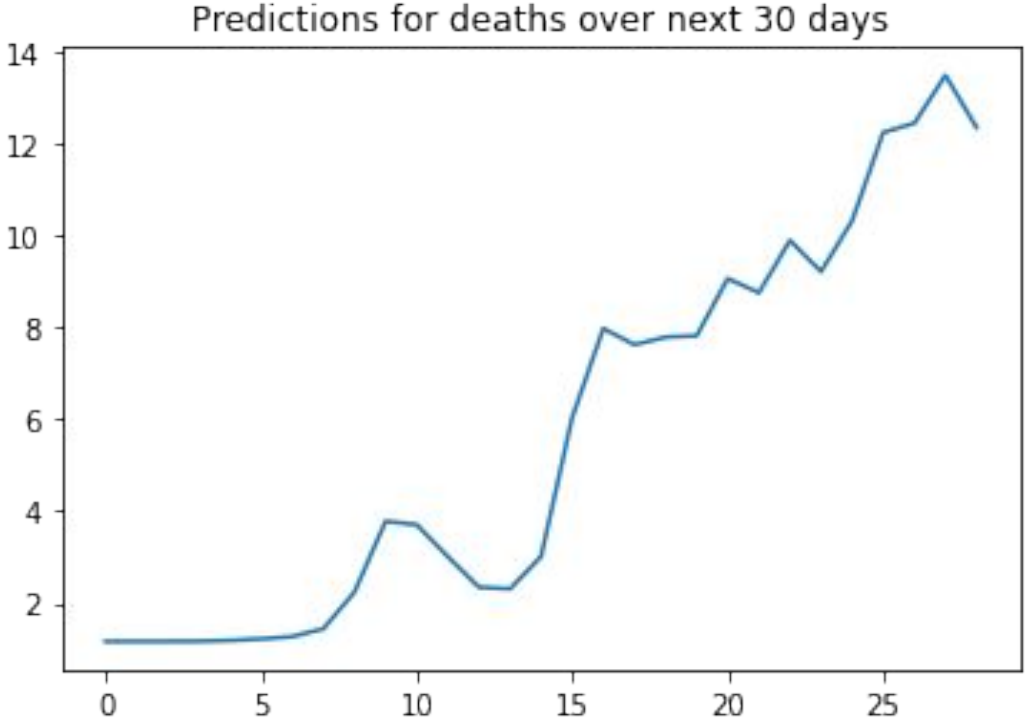




Comprehensive Take-Aways

- LSTM models are highly variable and can produce varied predictions in each iteration.
- The ARIMA model, on the other hand, cannot handle the changes in the time series and produces a straight line forecast.
- Combining both the models, gives us a much more stable and accurate forecast. Due to LSTM’s variability, we are able to predict the changes in the time series, and the ARIMA model stabilizes the prediction by reducing LSTM’s underestimation or overestimation.

Death Predictions for next 30 days



Observations and Hypothesis Validation

We can tell from the above graph that there is an increasing trend to the predictions for deaths in the next 30 days. So, our null hypothesis can be rejected and we can conclude that the **deaths seem to increase** for the next 30 days.

Reproducibility Checks

- A random seed of 5 was picked in order to exactly replicate our algorithm for models like the LSTM network which incorporate randomness in terms of Dropout noise
- Code is provided on the following Github link to recreate our process end-to-end:
- https://github.com/hariniramp/UW-Covid-19-Datathon