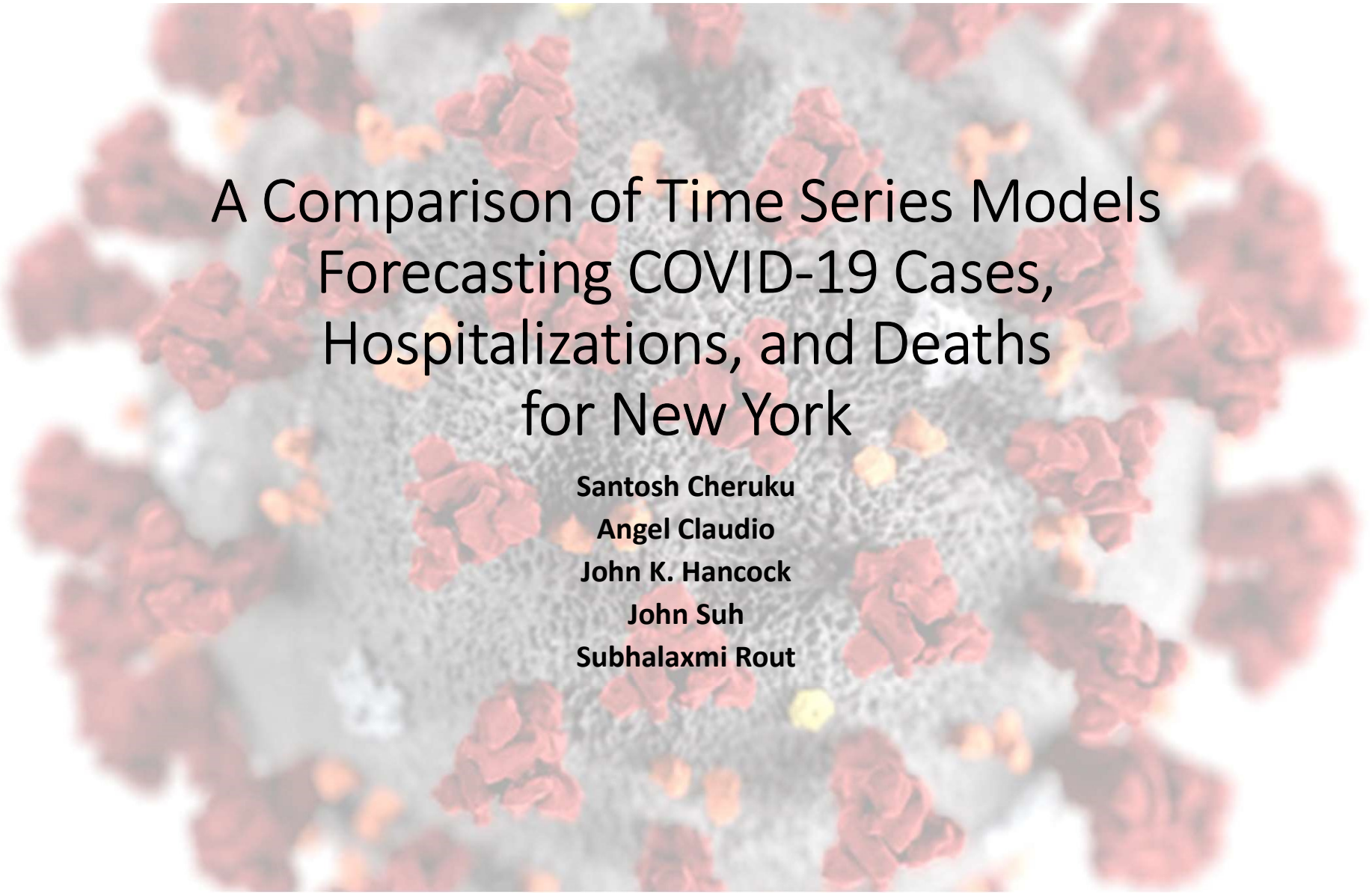


A Comparison of Time Series Models Forecasting COVID-19 Cases, Hospitalizations, and Deaths for New York

Santosh Cheruku
Angel Claudio
John K. Hancock
John Suh
Subhalaxmi Rout

Introduction

- December, 2019: The World Health Organization (“WHO”) reported the earliest onset of symptoms of a new highly contagious disease in Wuhan, China

- The virus, Severe Acute Respiratory Syndrome Coronavirus 2, would become known globally as COVID-19 and is spread through aerosol transmission of human respiratory droplets

- The World Health Organization (“WHO”) declared the virus a pandemic in March 2020

CUNY MS Data 698, Spring 2021: Santosh Cheruku, Angel Claudio, John K. Hancock, John Suh, and Subhalaxmi Rout

Introduction

- First case of COVID-19 in the US was reported in January 2020
- First death of COVID-19 in the US was reported in February 2020
- By April 2020, New York became the epicenter of the pandemic with over 12,000 cases and over 29,000 deaths by May 2020
- Public health and government officials were not prepared for the pandemic

Research Questions

- Could a Time Series model mitigate the overwhelming effects of the pandemic by preparing public health officials?

- Could time series data taken from March through May of 2020 predict cases, hospitalizations, and deaths for New York for June 2020?

Objective

The objective of this project is to compare nine time series models, from basic to the more advanced in order to determine which model can accurately predict cases, hospitalizations, and deaths.

The best model(s) could be used by public health officials to prepare for future pandemics.

Literature Review

- Reviewed over 25 articles on Time Series predictions and COVID-19
- Researchers pursued building time series models to predict the spread of the virus
- Researchers chose certain classes of models
- We did not find one article that compare multiple classes of models

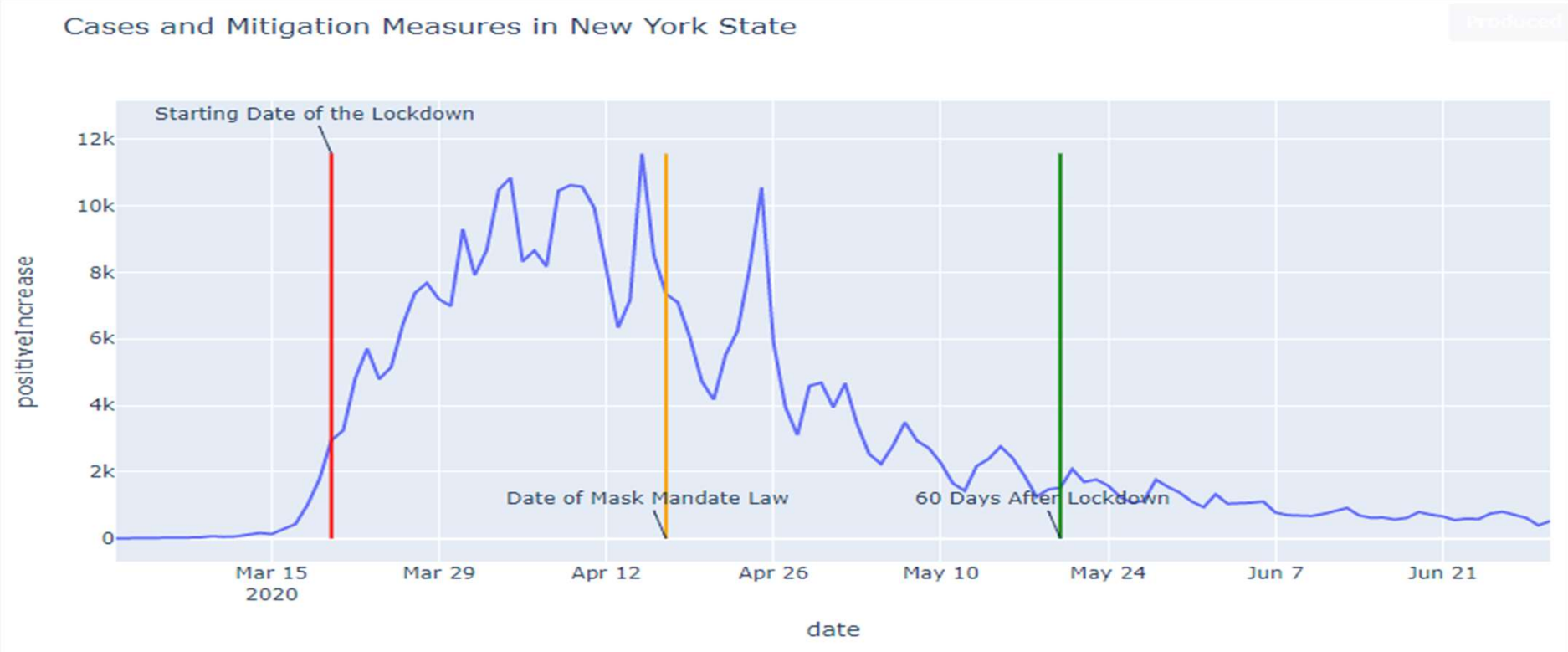
Data and Methods - Overview

- Data Sources
- Explore the Data
- Discuss the 9 models that we evaluated and compared
- Discuss the best model for our analysis

Data and Methods - Data Sources

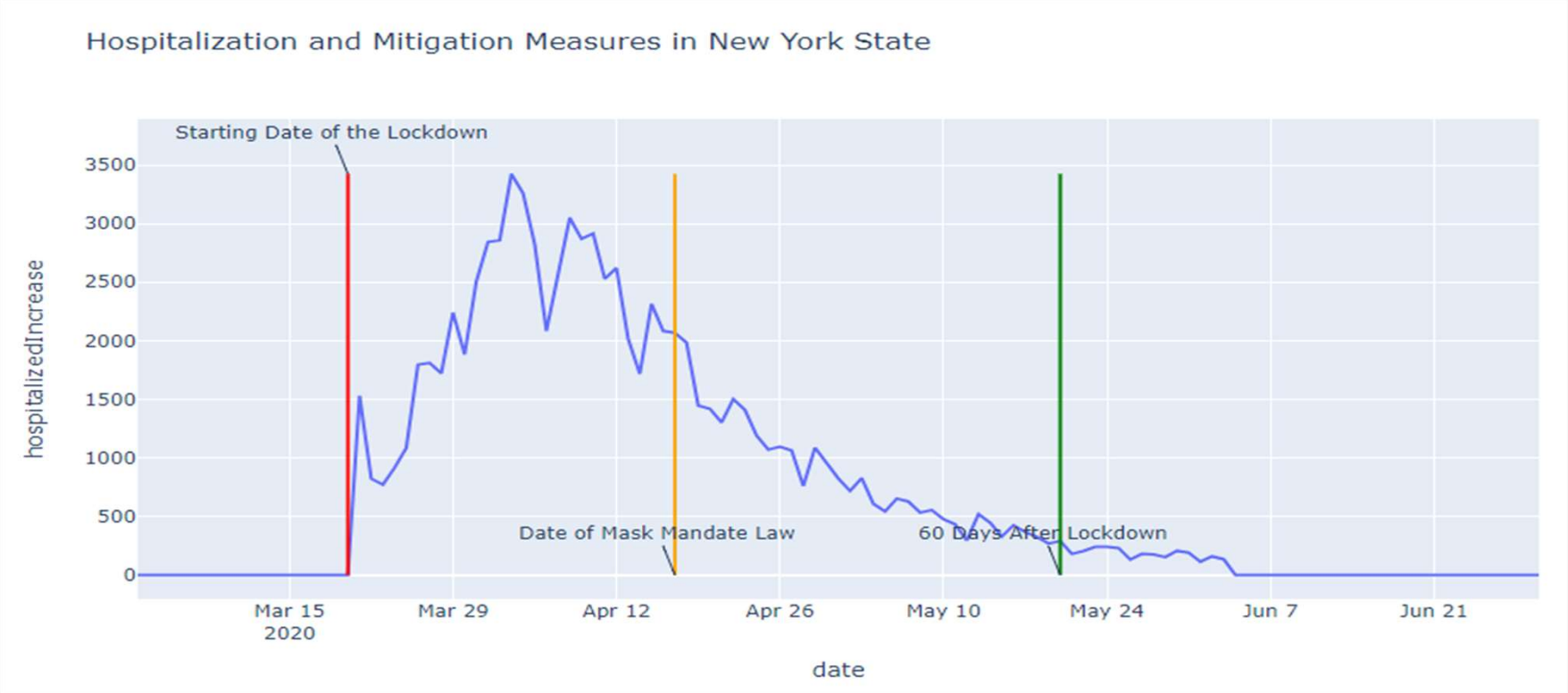
- Reviewed data collected by Johns Hopkins and the CDC
- Both datasets lacked the detailed information that we needed for analysis
- The best source that we found was The Atlantic's COVID-19 Tracking project, https://covidtracking.com[1]

Data and Methods - Exploratory Data Analysis (Cases)

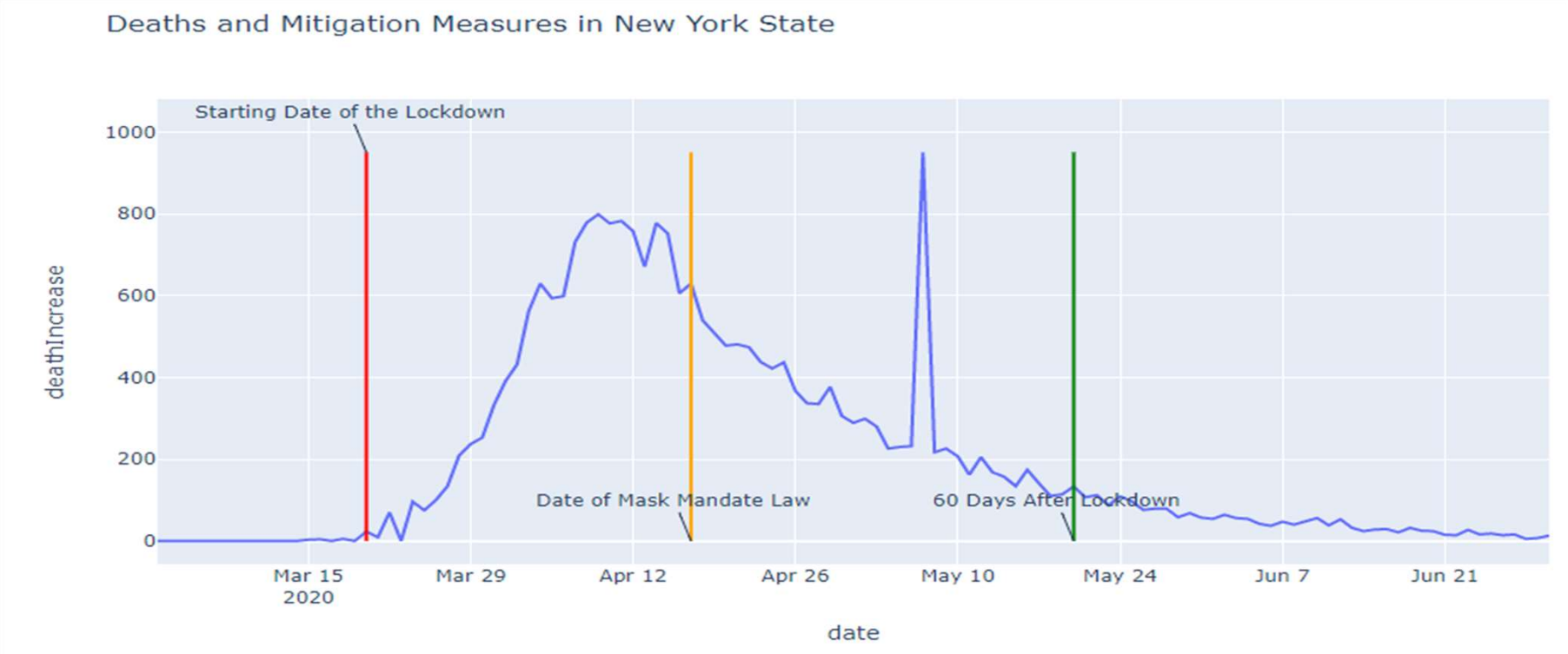


CUNY MS Data 698, Spring 2021: Santosh Cheruku, Angel Claudio, John K. Hancock, John Suh, and Subhalaxmi Rout

Data and Methods - Exploratory Data Analysis (Hospitalizations)



Data and Methods - Exploratory Data Analysis (Deaths)



CUNY MS Data 698, Spring 2021: Santosh Cheruku, Angel Claudio, John K. Hancock, John Suh, and Subhalaxmi Rout

Data and Methods - Data Models

Three Classes of Models

I. Basic Models

a. Seven-day Rolling Average
b. Simple Exponential Smoothing
c. Holt Winters

II. Autoregressive and Moving Average Time Series Models

a. ARIMA
b. SARIMA

III. Advanced Time Series Models

a. Facebook's FBProphet Model
b. Neural Network ("NN")
c. XGBoost
d. Long Short Term Memory ("LSTM")

Data and Methods - Basic Models

Seven-day Rolling Average

An average of the previous seven day observations:

$$MA_t = \frac{x_t + x_{t-1} + x_{t-2} + x_{t-3} + x_{M-(t-1)}}{M}$$

Simple Exponential Smoothing

Uses an alpha parameter as weights for past observations:

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots,$$

Holt Winters

A forecasting model that captures trend and seasonality

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(k+1)}$$

Data and Methods - Advanced Time Series Models

Facebook's FBProphet Model

An open-source library for univariate time series data forecasting. additive time series forecasting model, and the implementation supports trends, seasonality, and holidays.

Neural Network ("NN")

A system of inter-connected layers of nodes that takes input data and passes signals from one node to another and then from one layer to another through activation function.

XGBoost

A class of ensemble machine learning algorithms that can be used for classification or regression predictive modeling problems. Ensembles are constructed from decision tree models.

Long Short Term Memory ("LSTM")

A Recurrent Neural Network ("RNN") has the same structure as a NN with the major difference that an RNN has loops in them that allows for information to persist.

Autoregressive integrated moving average model (ARIMA)

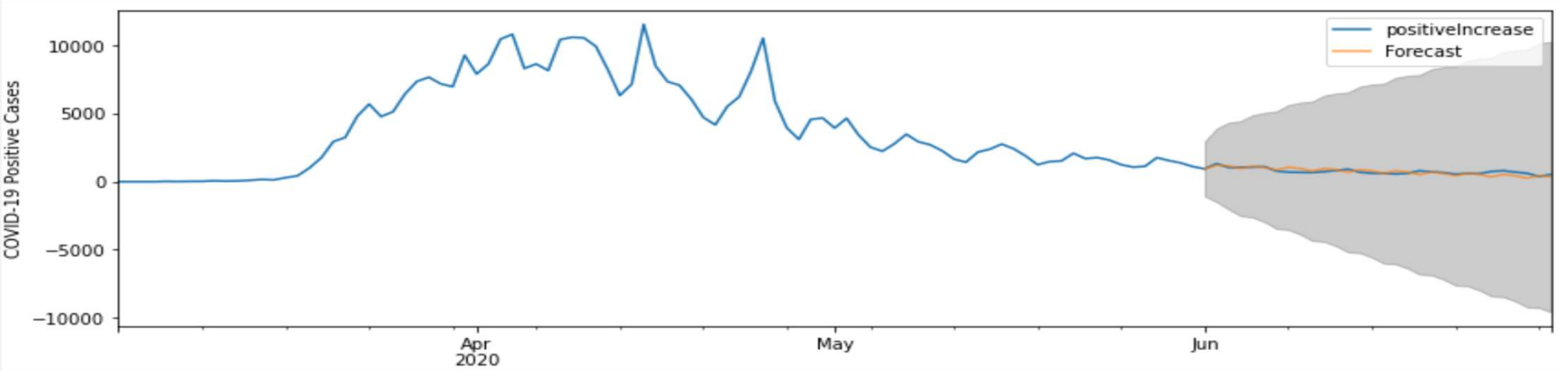
- ARIMA stands for autoregressive integrated moving average. The model is used to understand past data or predict future data in a series.
- ARIMA model is classified as ARIMA(p,d,q)
  a) p is the number of AR terms
  b) d is the number of nonseasonal differences needed for stationarity,
  c) q is the number of lagged forecast errors in the prediction equation
- ACF and PACF plot
- ARIMA popular because of its generality

Forecasting process:

- Plot the data, difference until series is stationary
- Examine differenced series and pick p and q from PACF and ACF
- Fit ARIMA(p, d, q) model to original data
- Check model diagnostics and predict forecast

Seasonal Auto Regressive Integrated Moving Average model(SARIMA)

- SARIMA model is extension of the ARIMA model to account for the seasonal nonstationary behavior of some series.
- SARIMA model is denoted SARIMA(p,d,q)x(P,D,Q,s)
  a) s = season
  b) P,D,Q represents the same p,d,q but they applied across a season



Comparative Results - Cases

Table 10: Positive Cases

Model	RMSE	MSE	MAE	MAPE
7-day Rolling Avg.	600.771	360,926.03	86.70	566.47
Simple Expo Smoothing	408.32	166,464	369.71	0.58
Holt Winters	**185.48**	**34,402.83**	**151.95**	**0.2136**
ARIMA	765.70	586,296.49	740.96	0.98
SARIMA	**188.46**	**35,517.17**	**155.43**	**0.223**
Facebook Prophet Model	2939.3	8,639,484.49	2,934.72	4.20
Neural Network	272.10	74,042.14	202.34	26.33
XGBoost	385.43	148,555.63	366.36	54.68
LSTM	507.27	257,331.84	455.52	64.96

Comparative Results - Hospitalizations

Table 11: Hospitalizations

Model	RMSE	MSE	MAE	MAPE
7-day Rolling Avg.	172.286	29,682.463	167.36	Undefined
Simple Expo Smoothing	182.356	33,253.689	177.70	Undefined
Holt Winters	62.95	3962.70	53.73	Undefined
ARIMA	**42.37**	**1795.22**	**15.18**	**Undefined**
SARIMA	**32.30**	**1043.29**	**13.94**	**Undefined**
Facebook Prophet Model	823.5	678,152.25	752.64	3.244
Neural Network	431.11	185,555.55	429.40	Undefined
XGBoost	183.22	354,816.4	566.70	Undefined
LSTM	144.70	20,935.26	132.9	Undefined

Comparative Results - Deaths

Table 12: Deaths

Model	RMSE	MSE	MAE	MAPE
7-day Rolling Avg.	44.852	2011.68	41.76	243.06
Simple Expo Smoothing	**31.82**	**1012.51**	**27.64**	**1.90**
Holt Winters	66.86	4470.26	60.49	3.5490
ARIMA	**33.74**	**1138.39**	**29.53**	**0.9**
SARIMA	**12.77**	**163.07**	**3.36**	**0.392**
Facebook Prophet Model	129.54	16,780.61	100.75	7.00
Neural Network	142.59	23,429.21	152	761.48
XGBoost	43.82	1920.4	204.8	207.47
LSTM	39.34	1547.65	37.68	174.46

Summary Findings

- The most consistent and highest accuracy for predicting cases, hospitalizations, and deaths was the SARIMA model
- Holt Winters model also performed well for forecasting cases
- ARIMA did well for hospitalizations, and the SES model performed well for Deaths
- The advanced models, Facebook Prophet Model and Neural Network, underperformed all other models
- The LSTM model was the best advanced model on Hospitalizations and Deaths
- XGBoost was the best advanced model for predicting cases

Conclusions

- Time series models are effective at forecasting a future time period for the COVID-19 pandemic
- Accuracy of any model is dependent on the accuracy of the collected data
- There was a wide variation in the performance of each model
- Most of the models that we saw were not built to model the spread of a virus specifically

Recommendations

- Strict protocols for data collection and reporting
- Develop a time series model specifically for predicting highly infectious diseases
- Forecasts from such a model need to be published so as to assist public health officials

References

1. The Atlantic, “The Covid Tracking Project”, 2021. [Online]. Available: https://covidtracking.com