Time Series Analysis and Forecast for COVID-19
Date: 11/22/2020
Raj Patel, Samarth Patel, Mansi Chaubey

Problem: Using the COVID-19 dataset to analyze and forecast the number of positive cases for the next 10 days by Time Series Analysis and building multiple models.

Overview of Data:

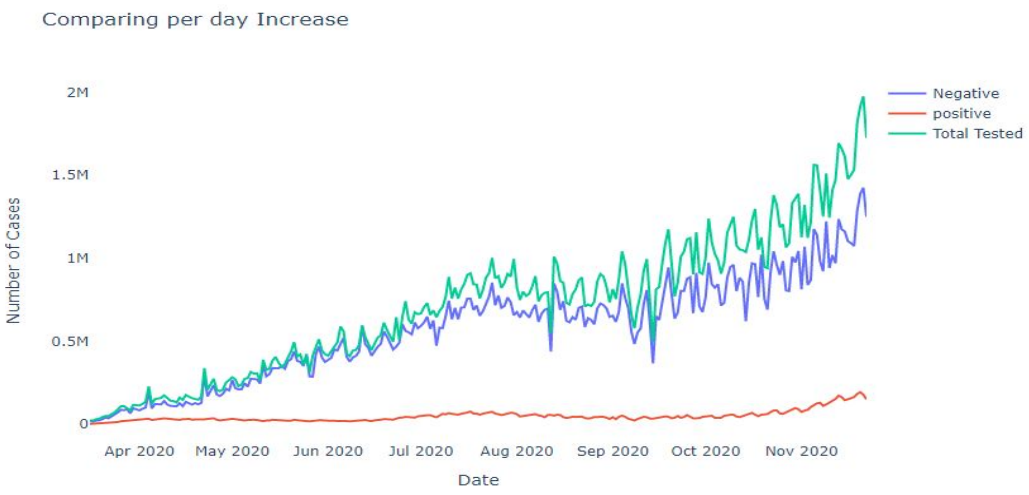
- **Data specification:**

  The dataset consists of 306 Rows and 18 Columns, where each row represents the data of each day across all US states and terretories. The data provided was in the CSV files. The data contains 18 features listed below.
- **Date:** Each day starting from March 16,2020 to Today
- **States:** Number of states US Including Territories.
- **Positive:** Total number of positive cases.
- **PositiveIncrease:** Number of Positive Case increased in each day
- **Negative:** Total number of negative cases.
- **NegativeIncrease:** Number of Negative Case increased in each day
- **Death:** Number of Fatalities in care facilities defined by the state.
- **Recovered:** Number of people Recovered from covid-19
- **DeathIncrease:** Number of death increased in each day
- **InICUCummulative:** Total number of individuals hospitalized in the ICU
- **HospitalizedCumulative:** Total number of individual hospitalized
- **HospitalizationIncrease:** Total number of individual hospitalized increase
- **OnVentilatorCumulative:** Total number of individuals hospitalized ventilation.
- **TotalTestResults:** Total number of completed antibody tests.
- **TotalTestResultsIncrease:** Total number of completed antibody tests per day

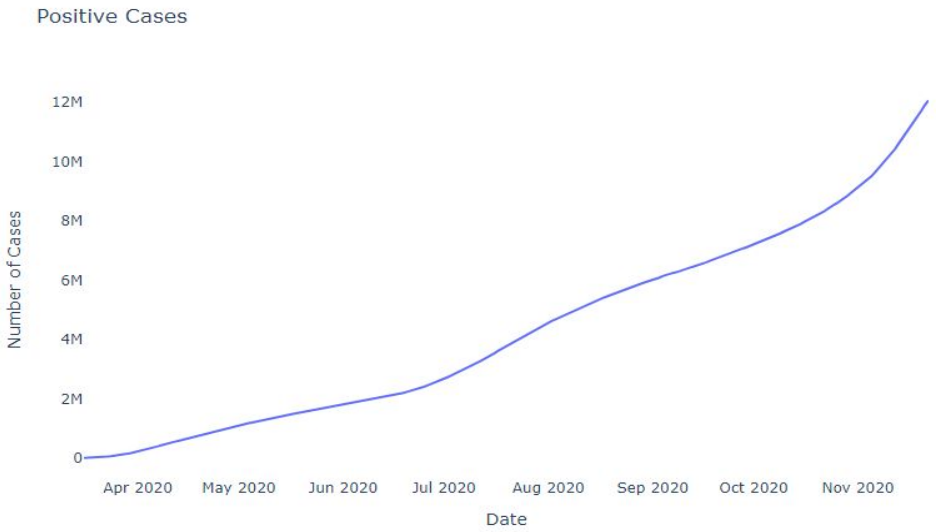
Preprocessing:

- The Dataset initially did not have data for all 56 states and territories, so the data if used as it is would have not given the complete picture for the whole US. so the samples where the data was not available for all states were removed.
- After reducing the dataset for all 56 states and territories, Null values which were present for some features were replaced with 0 as initial data was missing.
- Some features like **InICUCummulative, OnVentilatorCumulative, HospitalizationIncrease** were removed as not all states report those values.
- Feature extraction was done by finding correlation between features and removing those features whose value was more than 0.9.
- Final features which were used for forecasting were 'positive', 'positiveIncrease', 'deathIncrease','hospitalizedIncrease'.

Visualization and insights:



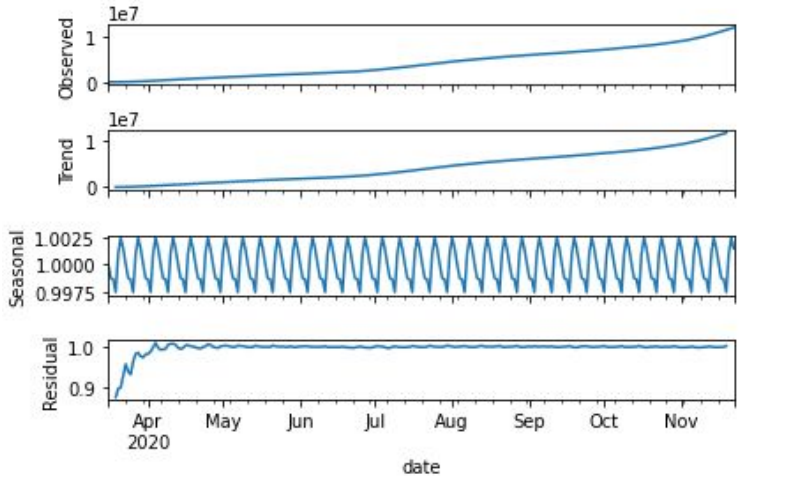
- The above graph shows the number of tests conducted and the outcome of tests as positive and negative.
- The number of tests conducted has increased over the period and so does the negativity and positivity which shows a positive correlation and upward trend.



- The above plot is a number of positive cases with respect to time.
- Plot shows that there is an upward trend which initially starts with linearly turns into exponential growth towards the end.
- The plot does not show seasonality and it can be seen that it is non stationary as the mean is increasing as time increases.

Time Series Analysis:

- Time Series data for positive cases can be decomposed into 3 factors, Trend, Seasonal and Residual.



- Time Series data for positive cases can be decomposed into 3 factors:
  - Trend
  - Seasonal and
  - Residual
- Looking at the decomposition above, an upward trend is observed with a constant seasonal and residual pattern.
- The above decomposition confirms that the trend is increasing, therefore the series is not stationary.
- Looking at the decomposition above, it can be seen that the trend is exponential towards the end which suggests that multiplicative trend. Seasonality has a constant pattern suggesting additive seasonality and residual is constant after initial increase suggesting no residual.

Models: Decomposition showed that the data is non-stationary so models like Exponential smoothing, ARIMA,and VAR were used for forecasting.

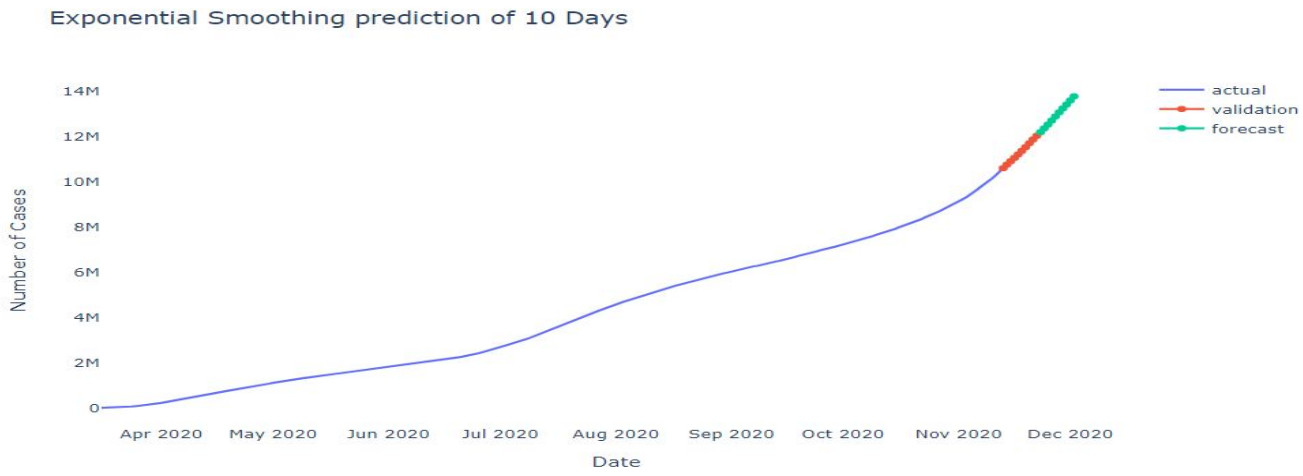
- Models were built using the stats model library, where the last 10 observations as validation set and rest as training set.

Exponential Smoothing:

- Exponential Smoothing is a technique that calculates the rolling mean with weights in order to forecast. It uses alpha as a parameter which is the smoothing factor.
- Model performance:
  - **R-squared value** : 0.9914
  - **MAE** : 37368.85
  - **RMSE**: 48769.29
- 10-day forecast:

2020-11-23	12163096.81
2020-11-24	12329842.43
2020-11-25	12498650.40
2020-11-26	12674341.43
2020-11-27	12855835.97
2020-11-28	13034714.60
2020-11-29	13207773.61
2020-11-30	13379864.99
2020-12-01	13559344.70
2020-12-02	13741020.04

- Plot with validation and forecast:

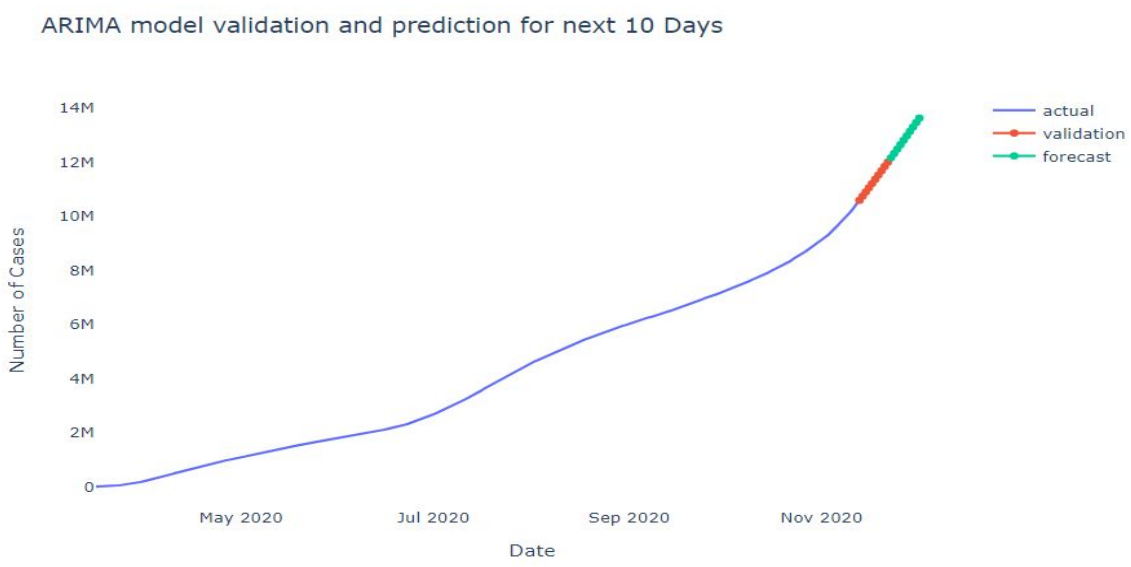


ARIMA: Auto Regressive Integrated Moving Average

- For this model the parameters (p=1,d=2,q=0) were used, where **p** is the number of autoregressive terms, **d** is is the number of nonseasonal differences needed for stationarity and **q** is the number of lagged forecast errors in the prediction equation.
- $\hat{Y}_t = \mu + Y_{t-1} + \phi_1 (Y_t - 2Y_{t-1} + Y_{t-2})$, this is the predictive equation corresponding to the values of p,d,q. Value of p=1 indicates a first order auto regressive model.
- Model performance:
  - **R-squared value** : 0.9876
  - **MAE** : 41614.69
  - **RMSE**: 53109.42
- 10-day forecast:

2020-11-23	12148467.85
2020-11-24	12309027.83
2020-11-25	12470211.79
2020-11-26	12632019.75
2020-11-27	12794451.69
2020-11-28	12957507.62
2020-11-29	13121187.54
2020-11-30	13285491.45
2020-12-01	13450419.34
2020-12-02	13615971.22

- Plot with validation and forecast:

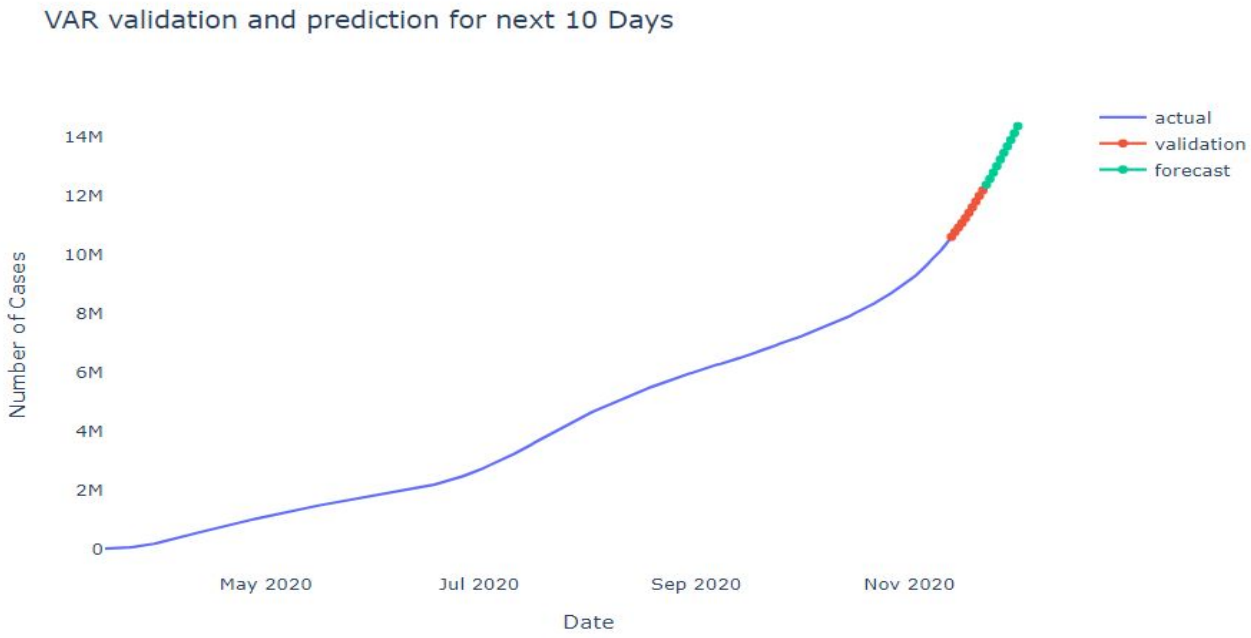


VAR: Vectorized Auto Regressive Model

- VAR is a multivariate Regression model i.e it depends on the other features other than the time and the y variable. For this model Positive cases and positive increase variables were used to forecast total positive cases for next 10 days.
- This model uses a first order regression equation just like ARIMA but compared to ARIMA which is univariate and VAR uses multiple variables.
- Model performance:
    - **R-squared value** : 0.9936
    - **MAE** : 27924.57
    - **RMSE**: 38041.25
- 10-day forecast:

2020-11-23	12356662.59
2020-11-24	12555071.62
2020-11-25	12767397.51
2020-11-26	12989129.47
2020-11-27	13214515.62
2020-11-28	13438219.14
2020-11-29	13657278.49
2020-11-30	13876025.95
2020-12-01	14103597.88
2020-12-02	14344364.62

- Plot with validation and forecast:



Model Insights and Evaluation:

	MAE	R2	RMSE
Exponential Model	41562.11	0.99	48769.29
VAR	27924.57	0.99	38041.25
ARIMA	41614.70	0.99	53109.42

- The above table shows how each model performed. Looking at the mean absolute error it is conclusive that the VAR model performs better relative to the other models.
- Mean Absolute Error calculated the mean error for all the points in the validation set, which shows that the prediction for a given point have on mean error about MAE value for that model.
- The R-Squared value for all the models were almost the same.

Conclusion:

- Predicting the Cases for COVID-19 without considering the external factors , and using only previous data provides an overview of how rapid the cases are increasing and will keep on increasing unless major steps are taken.

- The VAR model which is best among the others, predicts 144344364 cases for December 2 and it is a realistic number if we take into account the recent festive season.
- The VAR model which is multivariate performs better than the univariate models for the given TIme Series data.

Commented Video:

NAME	**NETID**	**Question Asked for (Group #)**	**Question Answered from (name)**
Raj Patel	rpate375@uic.edu	Group 1	Mohammad Khan
Samarth Patel	spate504@uic.edu	Group 1	Indu singh
Mansi Chaubey	mchaub2@uic.edu	Group 1	Jakub Krezeptowski-Mucha

REFERENCES:

https://www.kaggle.com/nitishabharathi/the-story-of-covid-19-in-india-eda-and-prediction
https://otexts.com/fpp2/arima.html
https://otexts.com/fpp2/expsmooth.html
https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm
https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775
https://www.aptech.com/blog/introduction-to-the-fundamentals-of-time-series-data-and-analysis/