PROJECT SYNOPSIS

***Project Title*: COVID-19 Infections and Death Projections for US**

Team

Project Guide: Dr Narayana D

Team Members: Amol R, Devendra B, Kushal T, Sanku L, Uday K

Abstract

The United States is the country with the most reported cases of 2019 Novel Coronavirus (COVID-19). Several research organizations, regulatory agencies and Universities have been tracking and forecasting COVID infections and fatalities at country (US) level especially since it peaked in Spring 2020 in the US. Our project aims to provide the following,

a) Forecast daily infections and deaths for the US for the period Nov 2020 – Apr 2021
b) Identify which State is going to lead the infections
c) Evaluate if demographic distribution is directly responsible for infections and fatalities.

With this the government and medical authorities will be able to prepare a mitigation plan to curb new infections.

Introduction

COVID 19 has been one of the most researched topics in recent times not only in the medical fraternity but also in the arena of AI and Machine Learning. It is not just the magnitude of this pandemic but also the unavailable cure due to which Scientists and Healthcare organizations around the world are working tirelessly to find a solution. Till the time a cure is in horizon Governments across the world are taking precautionary measures on how to contain further spread of the virus. This requires a two-fold approach, tactical measures at ground like containment, lockdowns etc. and long-term planning to minimize impact on the economy and population at large. A major input required for long term planning is the information on the rate of spread of virus in future, which demographics, or geographies it is going to spread more to and other factors like incubation period etc. Many universities, research organizations and government bodies have already been working to provide these forecasts based on how COVID has impacted countries and populations since its origin. John Hopkins has been the front-runner in sharing both the comprehensive COVID data to the public in general as well as defining the problem statements to determine statistics.

Our project utilizes the data provided by John Hopkins University for the period Jan 2020 – Oct 2020 as a reference to define the problem statement of forecasting as well as deriving a solution for the same. Further, the project attempts to arrive at a solution on three fronts, forecasting of total infections/deaths in future, State which is

going to lead these infections and answer whether demographic distribution is directly responsible for the spread.

We would be considering the broad framework of Machine Learning techniques like Supervised Learning (Regression), Multiclass Classification, Unsupervised Learning and Time-Series Forecasting Analysis using Univariate methods for non-stationary data to arrive at the three-pronged solutions.

In the first step, we would be using the unsupervised learning techniques like K-means Clustering and Hierarchical Clustering Dendrogram to identify if any particular regions or counties are exhibiting any hidden trend in terms of spike in infections or deaths. For this we would utilize different contributing factors e.g. age, population distribution, median income, weather (humidity / temperature / snow window), proximity to sea / water bodies etc. for each of these counties. One reason for this approach is because we observed inconsistencies in overall COVID infections development from one region in the existing projections available in the public domain. This has led us to believe that there are underlying factors which could be playing much enhanced roles rather than just a handful of factors like existing comorbidities or age or access to medical treatment, though these factors are equally critical. Once these clusters are identified we could apply forecasting and other stochastic techniques to determine how these clusters are performing in terms of rate of infections and death in future.

We would be using the widely used Machine Learning technique, Time Series Forecasting, to derive the projections based on the number of infections and deaths for the past months. Time Series Forecasting pertains to the sequence of observations collected in constant time intervals be it daily, monthly, quarterly, or yearly. Time Series Analysis involves developing models used to describe the observed time series and understand the "why" behind its dataset. This involves creating assumptions and interpretation about a given data. **Time Series Forecasting** makes use of the best fitting model essential to predicting the future observation based on complex processing current and previous data. For our problem we would be using the data containing infections and death of all US Counties and States for the last 10 months along with the demographics of each County.

A step-by-step framework will be incorporated to implement Forecasting and its associated model. As a first step we would be analyzing the three main components, Trend, Seasonality and Noise of the past data available through Kaggle. Plotting a time series of daily deaths and infections of counties will provide us a visual aid and identify if there is any upward or downward trend, repeating pattern or cycles and noise or irregularity in the data.

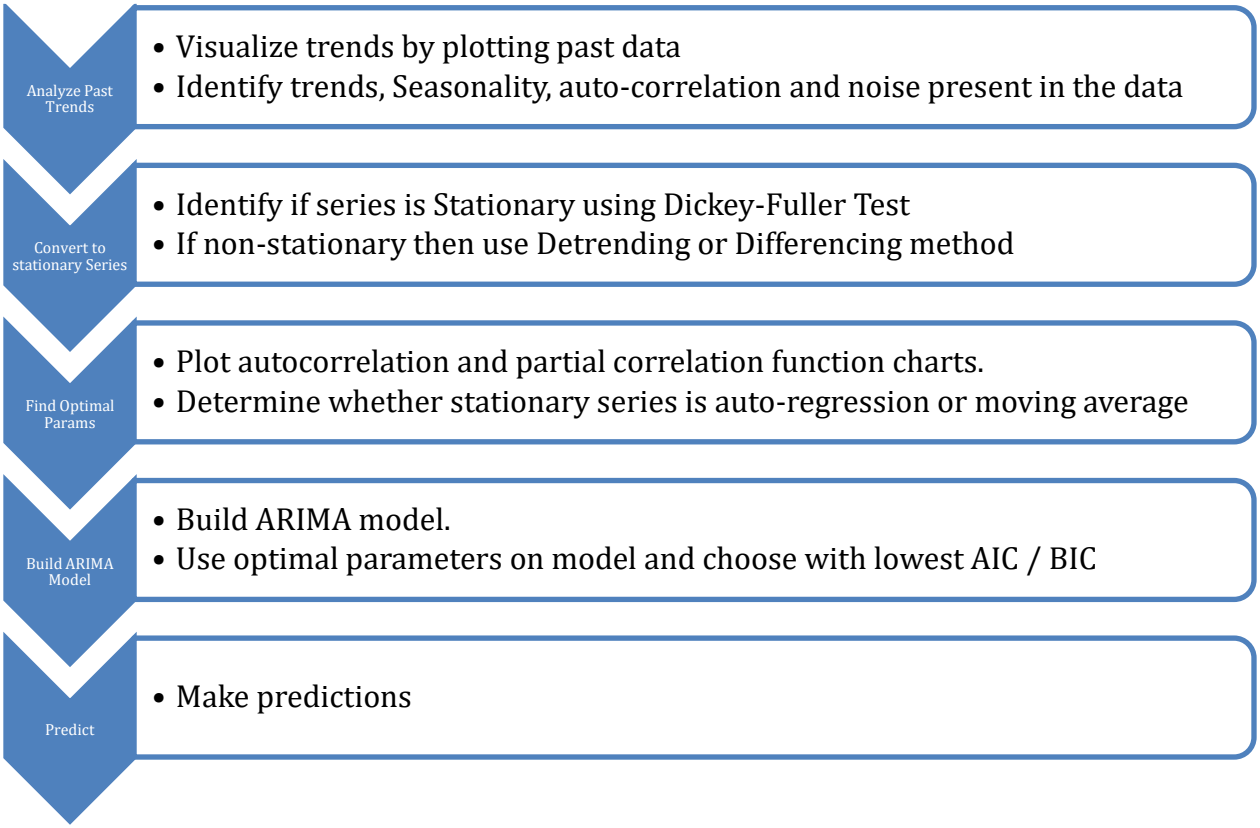

Figure 1: Workflow - Determining Projections

The hypothetical mathematical equation for time series could be represented as below:

$$y_t = a^{n*}y_{t-n} + \Sigma\varepsilon_{t-i}*a^i$$

where, y_t is the value at the time instant t, ε_t is the error term, i is the number of observations

Next, we check for whether this time series is stationary or non-stationary. This is primarily because if a series is not stationary, then that means that either the mean is not constant or the variance-autocovariances are dependent on time. For a series to be stationary it should exhibit three properties, no upward or downward trend i.e. trend should not change, constant variance even if the trend is not changing and equal spread across the time. If a time series does not exhibit these features, it becomes much more difficult to develop theory and models for time series. We would be conducting the stationary check primarily using unit root statistical technique, ADF (Augmented Dickey Fuller) popularly known as Dickey Fuller Test and validated through Kwiatkowski-Phillips-Schmidt-Shin or KPSS Test. This test will be used to determine the presence of unit root in the series, and hence help us understand if the series is stationary or not.

If the value of a is 1 (unit) in the above equation, then the predictions will be equal to the y_{t-n} and sum of all errors from t-n to t, which means that the variance will increase with time or a non-stationary output. A null hypothesis will be applied to determine

whether the series has a unit root(a=1) and thereby non-stationary. In case the series is non-stationary we would be applying Differencing (Integrating) technique to make the series Stationary. Once the series is determined as stationary, we would determine the optimal parameters (p, q, d) using ACF and PACF plots for use in ARIMA model where,

- **p** is the order of Autoregressive term (lagged predictions)
- **q** is the order of Moving Average (lagged errors)
- **d** is the difference of consecutive terms in the series

Finally, the ARIMA model will be built using the p, q and d values. The ARIMA model could be represented as:
where,

- α **r**epresents some value of white noise
- ϵ_{t-1} and ϵ_t refer to the past and current period respectively
- β and Φ represent respective coefficient for previous time value and error

An optimal set of these values could be derived using different combinations with values having minimum AIC and BIC chosen for building the final ARIMA model. We chose AIC and BIC methods over other approaches like Training/Validation/Test and Resampling model selection approaches as this method combines the complexity of the model with the performance of the model into a score. Moreover, this method also does not require to segregate data into subsets to first build the model upon and then test.

Finally, the predictions for infections and deaths for future time for all the clusters containing counties (from either a single state or multiple states) will be made using the ARIMA model. We can also visualize the trends to cross validate if the model works fine.

To address another part of the problem to identify which factors are most responsible for infections/deaths we will be applying the classification supervised learning algorithm on the identified clusters. We will first classify the clusters which will act as labels for our dataset. Next, we divide the entire set into training, validation, and test set. Once this is done, we can apply any advanced classification technique (e.g. Logistic/Random Forest) to determine the predictions for cluster labels. The accuracy of the chosen model can be evaluated using RMSE or MAE. This could also be supplemented with Precision/Recall and F1 scores to validate the accuracy scores.

The end goal of this project is to imbibe the learning and techniques of machine learning to attempt at finding factors responsible for COVID 19 infections as well as providing a futuristic projection. Though we do not dive in the clinical pathophysiology of COVID and its interaction with humans we have taken into consideration environmental and economic factors which could be major participants in spreading infections. Above all, we attempt to derive the solutions from a mathematical perspective rather than from a domain angle.

References

i. https://www.kaggle.com/headsortails/covid19-us-county-jhu-data-demographics#us_county.csv
ii. https://www.iunera.com/kraken/big-data-science-intelligence/time-series-and-analytics/top-5-common-time-series-forecasting-algorithms/
iii. https://www.analyticsvidhya.com/blog/2018/09/non-stationary-time-series-python/
iv. https://codeit.us/blog/machine-learning-time-series-forecasting
v. https://towardsdatascience.com/cluster-then-predict-for-classification-tasks-142fdfdc87d6
vi. https://medium.com/analytics-steps/introduction-to-time-series-analysis-time-series-forecasting-machine-learning-methods-models-ecaa76a7b0e3
vii. https://machinelearningmastery.com/probabilistic-model-selection-measures/
viii. https://www.analyticsvidhya.com/blog/2015/03/introduction-auto-regression-moving-average-time-series/
ix. https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/
x. https://link.springer.com/article/10.1007/s41403-020-00165-z
xi. https://towardsdatascience.com/simple-exponential-smoothing-749fc5631bed
xii. https://towardsdatascience.com/using-machine-learning-to-model-the-growth-of-covid-19-2f3b0af304bb
xiii. https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/
xiv. https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas
xv. https://www.kaggle.com/johnjdavisiv/us-counties-covid19-weather-sociohealth-data

Glossary

I. ACF – Auto Correlation Function
II. AIC - Akaike Information Criterion
III. AR – Auto Regression
IV. ARIMA - Auto Regressive Integrated Moving Average
V. BIC - Bayesian Information Criterion
VI. MA – Moving Average
VII. MAE – Mean Average Error
VIII. PACF – Partial Correlation Function
IX. RMSE – Root Mean Square Error
X. FIPS - Federal Information Processing Standards