COVID-19 Time Series Analysis

Names: Priyanka Sahoo

Supervisor: Shagufta Henna, PhD

Degree: MSc. Big Data Analytics,

Module: Big Data Analytics

Abstract—

”Novel Corona virus, Pandemic, Covid-19, Lockdown” are the only buzzwords in the past year across the planet. Within a few months this contagious corona virus spread from country to country universally and disintegrated almost every industry considered as pillar of world’s economy e.g. health-care, transportation, aerospace, agriculture, entertainment, education, manufacturing, hospitality and so on leaving a trail of crumbled humanity behind. This research paper will mainly focus on dynamics of COVID-19 pandemic worldwide with time series analysis and its future predictions. There is a discussion on the trend analysis of the disease spread on the world map in addition to the visualizations of new cases, confirmed, recovered and death in the most affected countries with various graphs. The most important features of this exploration are established with segregating the world population into confirmed, death, recovered and new cases corresponding to different countries. Various machine learning regression models are developed and trained with the World Health Organization (WHO) and ArcGIS (Geographic Information System) steadfast data. The study will demonstrate a life cycle of data science project prototype using Big Data Analytics and architecture. Notable results are observed with truthful cumulative predictions country wise. Moreover, there are also discussions on comparison of the root mean squared values from different models in the conclusion part with regards to validation of machine learning model’s credibility and the future statistics of COVID-19 could be helping hand for government, NGOs, hospitals etc. for the battle against present pandemic.

Introduction—

COVID-19 is the foremost question in the year 2020. World health organization declared this as global pandemic during the first quarter last year, impacted more than 90% of the continent’s population. It is spreading exponentially daily and become a never-ending process. Few medical studies even revealed that this corona virus will reside inside human body eternally, only antibodies within us can conquer it gradually with time. As per today’s date, human race is trying to balance off between the corona virus vaccines and the new covid-19 strains emerging from different realms. In this research paper, there will be analysis on the pattern of spread using the features like new and confirmed case, mortality rate and recovered records in most affected countries with respect to time.

Time series analysis can be defined as the study done on a data set where each data point is observed under particular time instance [1]. As part of this venture, various analytics process will be able to appease further down queries.

1. Stats of total deaths and recovered daily/weekly/monthly across all over the world.
2. COVID-19 variations and trends over time.
3. COVID-19 cases spread Visualization across countries in 2019-2020.

4. Predictions for the new confirmed cases, death and recovered numbers through machine learning regression models.
5. Forecasting the future of pandemic through deep learning.

In view of, variation of COVID-19 impact with time, forecasting for count of new confirmed cases and deaths for near future done using deep learning. Apart from COVID-19 trends visualization, the study presents comparison between 7 types of regression models and their integrity. The research paper is organized in different units coined as dataset overview, system architecture, exploratory data analysis, ML model development, forecasting, results, and conclusion. Additionally, the whole analytics process is done in Databricks community edition that provides clusters that run on top of AWS and adds a convenience of already hosted notebook system. It supports variety of spark APIs as well and has its own storage like DBFS, S3, Hive etc.

Code files—

L00151175_COVID-19 Project_Code_DataBricks.ipynb

Data File— WHO-COVID-19-global-data-1.csv

Datasets sourced from ArcGIS and WHO websites,
https://services1.arcgis.com/FeatureServer/1/query
https://covid19.who.int/WHO-COVID-19-globaldata

Methods section—

- ### How you cleaned, prepared the dataset with samples of intermediate data

The json data file is pulled directly from ArcGIS url to get the latest information for visualization. Main aim is to get the interactive graph plots displaying the results for the same day as the code executed without any lag in latest statistics. As discussing in the above section dataset overview, first column 'features' is sliced out for further examination. Excavating deep through the facts, the figure 7 is a snapshot of the raw, normalized data from json file. Column 'states' corresponding to 'Countries' were filled with null values initially, which was replaced with spaces. The column 'Last Update' holds date values like [1.614418e+12] which is formatted to actual timestamp value e.g. [2021-02-27 00:23:52] as part of data pre-processing.

For time series analysis and Machine learning model implementation up-to-the-minute WHO COVID data stored in databricks file system read as spark data frames as well as pulled directly from API for including today's data also. At the outset, the comma separated files are warehoused to the databricks distributed file system. Then the stored data files are read and constructed spark data frames for further data wrangling. The data frame contains features as 'New cases', 'Cumulative cases', 'New deaths', and 'Cumulative deaths' that is aggregated by 'Date reported'. And then this aggregated data set is used further.

- ### Tools you used for analysing the dataset and the justification (tools, models, etc.)

The whole analytics process is done in Databricks community edition that provides clusters that run on top of AWS and adds a convenience of already hosted notebook system. It supports variety of spark APIs as well and has its own storage like DBFS, S3, Hive etc.

Databricks is used due to its optimized usage cost, readily available notebook system and compatibility with python plus spark. The procedure is conducted on cluster of 16 GB and 2 cores processor having configuration of DBR 7.5, Spark 3.0.1, Scala 2.12. Moreover, databricks files system (DBFS) is used to store the data. DBFS is distributed file storage system, installed on Databricks work space. The study is done, by exploiting Python Scikit learn and MLlib Spark libraries.

- How did you model the dataset, what techniques did you use and why?

First the data set is split into training and test and various machine learning models were used for predicting the new cases and deaths due to COVID-19. Here, OLS, LASSO, Ridge, Elasticnet, Bayesian, OMR Regression models imported from python Scikit learn package and Linear regression, K-Means clustering, Gaussian mixture and Random forest models imported from MLLib PySpark algorithms used for predicting the counts new cases and deaths due to corona virus effect.

- Did you have a hypothesis that you were trying to prove?

The main goal of this research paper are the prediction and visualization of below,

1. Stats of total deaths and recovered daily/weekly/monthly across all over the world.
2. COVID-19 variations and trends over time.
3. COVID-19 cases spread Visualization across countries in 2019-2020.
4. Predictions for the new confirmed cases, death and recovered numbers through machine learning regression models.
5. Forecasting the future of pandemic through deep learning.

- Did you just visualize the dataset, and if so, why?

As part of this research, first the maximum number of corona virus confirmed cases are aggregated by equivalent countries, organized in descending order. Same done for active, recovered and death cases. Later, forecasting for new cases and deaths are plotted in graphs using Facebook Prophet open source libraries. Prophet is used for time series prediction. It decomposes the time series into trends, seasonality, and holidays. It has inbuilt easily tuned, instinctual hyper parameters.

Results/Conclusions section—

- What did you find and learn?
- How did you validate your results?

Following the findings, in this research paper, an observation could be made for the pattern in COVID-19 spread globally. United States, India, Brazil, Russia, UK, France, Spain, Italy, Turkey, and Germany are the topmost affected countries in the world. The month of December 2020 displays the maximum spread of virus in USA, which could be assumed that this may happened due to Lockdown restriction liberty during Christmas. India has a spike in confirmed cases in the month of September, after which Indian Government declared Level 5 restriction and here it shows quite a good decrease by October. The time series analysis between the cumulative new cases and deaths generates a polynomial curve of nth degree for various countries.

Discussing over the effective variance and root mean squared error experienced from OLS, Ridge, Lasso, Baeyesian, ElasticNet, Lasso Lars, OMR and Tweedie regressor the values are approximately same as 1.0. This clearly shows a linear increase in Cumulative death cases with respect to new cases. And the same is forecasted as well using Fbprophet time series predictions.

The comparison between the least R- squared values calculated using linear regression and n-degree polynomial regression model proved the later one is good approach. However, higher order of polynomial could result in overfitting where the curve is going out of its way to accommodate outliers. A high R-squared simply means the curve fits training data well, but it may not be a good predictor.

Additionally, The Kmeans clustering model shows a defined clusters with silhoutte score of 0.86. The inter cluster distance called as centroids is also calculated to check the cluster centers. However, kmeans only considers the mean to update the centroid while Gaussian Mixture model takes into account the mean as well as the variance of the data [25]. The values observed in GMM are more distinct as compared to Kmean clustering.

Random forest regressor algorithm is also used to do predictions for new cases with respect to dates. There is a calculation of R-square on both train data and test data have values as 2.15716 and 3.53081 respectively. Larger R-square values signifies better regression model that fits to the observations. Here it shows test data performs better that trained data in predicting the new cases with respect to date. The trained dataset is fitted in the model and got an average prediction of 11.7.

Future work—

On the closing note, would only say that accurate results mostly depend on various features used. Sometimes, there may be unforeseen scenarios which may not fit with a specific technique for the first time. As said earlier, big data analytics with machine learning will get trained and become more mature with the experience. A future work could be trying Apache Spark implementation is to on a machine with higher GPU, which may increase the performance. And also trying the research with real time data could be helpful for better insights. More comprehensive studies to describe the context of a broader type of time series analysis with covid data are currently being considered by larger research communities.