John Marshall

Udacity Deep Learning Nanodegree Capstone Proposal

Domain Background:

In late 2019, an infectious disease called Coronavirus (COVID) was first diagnosed in Wuhan, China prior to spreading globally before being labeled a pandemic on March 11th, 2020 by the World Health Organization (Adhanom *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*). Given the grave consequences concerning the proliferation of this pandemic, I thought it would be prudent to try and tackle some forecasting on the expected spread of this disease. While an infectious disease can naively be modeled with either an exponential growth/decay or some sort of Markov chain, these tools lack the capability of capturing more rich details. As of this instant, there have been 5,991,102 confirmed cases and 366,875 deaths spread across 188 countries/regions (Hopkins *Covid Dashboard*).

Problem Statement:

Given an initial condition (total confirmed cases and total deaths as of date X), how many new cases/ new deaths can we expect over a specified interval. If we can properly predict the evolution of this pandemic, it could allow more effective resource allocation for all the independent entities that are trying to contain this pandemic. My aim is to predict the pandemic one week out at a time. I am guessing perturbations inherent in the data (given the uncertainty about the unconfirmed cases) will cause the projections to grow increasingly unstable for each additional timestep. If I find success at the weekly interval, I will try extending the timeframe to see what type of accuracy I can achieve.

Dataset and Inputs:

I am going to use these datasets I found from Kaggle:

1. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_deaths.csv (Srk *Novel Corona Virus 2019 Dataset*)
2. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_confirmed.csv (Srk *Novel Corona Virus 2019 Dataset*)

I also intend to do some more exploring given new datasets are being developed as we organize more data about this pandemic. If I do end up including more data, I will make sure to update my proposal/include it in the final analysis.

Solution Statement:

I intend to replicate the timeseries notebook from the previous module and use SageMaker's DeepAR forecasting algorithm.

Benchmark Model:

I will compare this deepAR model to a more elementary exponential smoothing algorithm and/or an ARIMA model as my baseline.

Evaluation Metrics:

I intend to use root mean squared error as my basis for comparison across the three different models.

Project Design:

My intention is to condense the province/state and region labels into one categorical variable and then convert the dates into a proper format. Once I clean the data, I will investigate at multiple levels whether there are any discernible patterns. One thing I might want to include, is grabbing and normalizing the countries 2018 GDP as an instrumental variable to control for countries/regions with less equipped institutions to ensure social distancing is followed. After that I will start building my models and comparing the results not only across the models, but also across the time series timeline.

Works Cited

Adhanom, Tedros. “WHO Director-General's Opening Remarks at the Media Briefing on COVID-19 - 11 March 2020.” *World Health Organization*, World Health Organization, 11 Mar. 2020, www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

Hopkins, Johns. “Covid Dashboard.” *ArcGIS Dashboards*, 30 May 2020, www.arcgis.com/apps/opsdashboard/index.html.

Srk, SRK. “Novel Corona Virus 2019 Dataset.” *Kaggle*, 29 May 2020, www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_confirmed.csv.

Srk, SRK. “Novel Corona Virus 2019 Dataset.” *Kaggle*, 29 May 2020, www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_deaths.csv.