Machine Learning Engineer Nanodegree

Capstone Proposal

Ashraf Hussain June 19th, 2020

Proposal

Forecasting COVID-19 Cases – A DeepAR Model

Domain Background

On December 31, 2019, the World Health Organization (WHO) was informed of an outbreak of “pneumonia of unknown cause” detected in Wuhan City, Hubei Province, China. Identified as coronavirus disease 2019, it quickly came to be known as COVID-19 and has resulted in an ongoing global pandemic. As of 20 June 2020, more than 8.74 million cases have been reported across 188 countries and territories, resulting in more than 462,000 deaths. More than 4.31 million people have recovered.[^1]

In response to this ongoing public health emergency, Johns Hopkins University (JHU), a private research university in Maryland, USA, developed an interactive web-based dashboard hosted by their Center for Systems Science and Engineering (CSSE). The dashboard visualizes and tracks reported cases in real-time, illustrating the location and number of confirmed COVID-19 cases, deaths and recoveries for all affected countries. It is used by researchers, public health authorities, news agencies and the general public. All the data collected and displayed is made freely available in a GitHub repository.

Problem Statement

This project seeks to forecast number of people infected and number of deaths caused by COVID-19 for a time duration of 14-days based on historical data from JHU. I will be using Amazon SageMaker DeepAR forecasting algorithm, a supervised learning algorithm for forecasting scalar (one-dimensional) time series using recurrent neural networks (RNN) to produce both point and probabilistic forecasts[^2]. DeepAR is an underutilized approach in this area.[^3] The dataset contains hundreds of related time series, and DeepAR outperforms classical forecasting methods including but not limited to autoregressive integrated moving average (ARIMA), exponential smoothing (ETS), Time Series Forecasting with Linear Learner for this type of applications.

Datasets and Inputs

The datasets are accessed from files provided by the JHU GitHub repository time_series_covid19_confirmed_US.csv

The file have the same columns:

- UID - UID = 840 (country code3) + 000XX (state FIPS code). Ranging from 8400001 to 84000056.
- iso2- Officially assigned country code identifiers 2 Chr (US, CA, ...)
- iso3 - Officially assigned country code identifiers 3 Chr.(USA, CAN, ...)
- code3- country code USA = 840
- FIPS -Federal Information Processing Standards code that uniquely identifies counties within the USA.
- admin2 - County name. US only.
- Province_State - The name of the State within the USA.
- Country_Region - The name of the Country (US).
- Combined_Key - Province_State + Country_Region
- Population - Population
- Number of cases are is columns where each column is a day

Solution Statement

The solution will offer predictions of next 14-day cases in an easy-to- use, intuitive forecasting model. The 14-day time period was selected on the basis of the incubation period of the novel coronavirus. Since the data sets are relevantly clean, I expect to spend 50% of the time on data cleaning and DeepAR processing and 50% of the time on training models and tweaking parameters. To minimize the `cost` the following `Hyperparameters` will be used:

`early_stopping_patience` : enabled as when this parameter is set, training stops when no progress is made within the specified number of epochs. The model that has the lowest loss is returned as the final model.

`likelihood` : will set to *negative-binomial*: Use for count data (non- negative integers). As the model generates a probabilistic forecast, and can provide quantiles of the distribution and return samples. Depending on your data, select an appropriate likelihood (noise model) that is used for uncertainty estimates.

`time_freq` : will be set to 14D as will be forecasting the next 14-days only other Hyperparameters will be set to default.

Benchmark Model

For this problem, the benchmark model will be ‘Time Series Forecasting with Linear Learner’ to prove the hypothesis theoretically that the DeepAR model is more suited for this type of forecasting.

Evaluation Metrics

Prediction results are ±80% confidence intervals between predicted values and the ground truth.

Project Design

The project will be executed based on the following template:

- Load and Explore the Data
- Pre-Process the Data
    - where i need to convert the data form a wide-table format to tall-table.
- For DeepAR
    - Create Time Series
    - Splitting in Time Series into Training and Test
    - Convert Time Series to JSON
    - Uploading Data to S3
    - Training a DeepAR Estimator
    - Setting Hyperparameters
    - Creating Training Job
    - Deploy and Create a Predictor
    - Generating Predictions
    - Display the Results
- For Time Series Forecasting with Linear Learner
    - Creating Training Job
    - Deploy and Create a Predictor
    - Generating Predictions
    - Display the Results
- Clean up
    - Delete the Endpoints
    - Delete S3 bucket
    - Delete Notebook

Endnotes

[^1]:COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)". ArcGIS. Johns Hopkins University. Retrieved 20 June 2020.

[^2]:DeepAR Forecasting Algorithm. Amazon Web Services. Retrieved 20 June, 2020

[^3]:Time series prediction. Telesens. Retrieved 20 June, 2020.