Udacity Machine Learning Nanodegree

Capstone Proposal

Prediction of positive cases of COVID-19 in Peru using Time-Series Forecasting

Tania Incio

February 2021

Domain Background

The coronavirus COVID-19 has been classified by the World Health Organization as a public health emergency of international importance. The pandemic began in December 2019 in Wuhan, China and its spread was rapid and global, causing the deaths of thousands of people around the world.

In this scenario, we can say that the current COVID-19 pandemic is devastating, despite the wide implementation of control measures.

In an analysis of the panel of COVID-19 cases in Peru until February 26, 2021, the country's Ministry of Health had confirmed 1,308,722 cases, in addition to 45,903 deaths. The cases are distributed throughout the national territory, with a greater concentration in the capital Lima.

The data for Peru are alarming. In this sense, this project will aim to analyze and predict positive cases of COVID-19 in Peru using Machine Learning techniques such as Time-Series Forecasting.

There are some of real-world applications for this research, such as:

- Estimated population infected by Covid-19 using generalized logistic regression and optimization heuristics
- Interpretation of forecasts of new cases of covid-19
- Prediction of COVID-19 cases in Peru
- Prediction of the number of infections and deaths from COVID-19 in Mexico

The links of each project will be added at the end of this document as references.

My personal motivation for working on prediction of positive cases of COVID-19 is my current job, I work with an epidemiologist doctor I am in charge to analyze the COVID-19 data from Peru frequently so I am familiar with the data then It would like to apply my machine learning knowledge using time-series forecasting to this domain.

Problem Statement

Taking into account the situation that Peru is going through, this project will help epidemiologists to have one more tool to estimate the risk of this pandemic by predicting positive cases of covid-19 in Peru and some control measures can be taken.

To achieve the objective I will analyze the data of the Ministry of Health of Peru and use Machine Learning techniques such as Time-Series Forecast for the predictions of new cases.

Datasets and Inputs

For this project we will use the open data of the Peruvian Ministry of Health. These published data correspond to the total of reported cases that tested positive for COVID-19, by department, province and district. In addition, within the data set there are data that allow identifying the main characteristics of the patient such as age, sex and date of obtaining the positive result. At the level of update frequency of the report, the Open Data Portal indicates that the Ministry of Health carries out a daily update of the information.

Link of open data of Ministry of Health:
https://www.datosabiertos.gob.pe/dataset/casos-positivos-por-covid-19-ministerio-de-salud-minsa

I will work with data obtained from March 6 to February 26, 2021, there can be many records in a single day, every record means a single positive case of COVID-19, the dataset contains 1,093,938 records.

I will predict the number of infections of COVID-19 per day so my labels will be the number of new cases of COVID-19 per day.

The dataset has the following inputs:

- UUID: Universally unique identifier, identifier of each record
- DEPARTAMENTO: State where the infected person lives
- PROVINCIA: Province where the infected person lives
- DISTRITO: District where the infected person lives
- METODODX: Method applied such as rapid test, molecular test and antigen test
- EDAD: Age of the infected person
- SEXO: Sex of the infected person
- FECHA_RESULTADO: Date where the result of the test of the infected person was known

I will count the number of records per day to obtain the number of infected people per date after that I will split the data in chronological order.

Solution Statement

The proposed solution to this problem is to apply Machine Learning techniques such as Time-Series Forecasting.

First I will extract the data from the Open Data Portal of the Peruvian Ministry of Health .

The next step will be to clean and analyze the data of COVID-19.

After that I will create training and test sets of time series and I will apply the algorithm taught in classes, DeepAR is a supervised learning algorithm for forecasting time series that uses recurrent neural networks.

Then I will make predictions and I will use the evaluation metrics such as coeffcient of determination to compare the performance of the training set against the test set.

Benchmark Model

For the benchmark model, we will use the algorithms outlined in the paper "Prediction of the number of infections and deaths from COVID-19 in Mexico" (CONACyT-CONABIO, Inder Tecuapetla-Gómez, 2020) [1]. The paper considers the semi parametric regression model:

$$y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{K} u_j (x_i - \tau_j)^3 + \varepsilon_i, \quad 1 \leq i \leq n$$

This model is equivalent to a mixed linear model where yi represents the number of new cases per day

Metrics	R^2
For new cases	0.9999211
For new deaths	0.9996476

Evaluation Metrics

Since this is a regression problem the evaluation metric will be R^2 or Coefficient of Determination.

Project Design

Loading and exploring the data

I will get the data from the open data portal of the Ministry of Health using the *requests* and *pandas* libraries to read from the web. I will review the type of data in each column, I will check how many nulls there are for each column and if they are necessary to do data imputation I won't need to scale or normalize the data. I will make graphs for better understanding.

Data Preprocessing

Since I am familiar with the data I will make the following steps for preprocessing:

- I will remove the columns UUID since this column doesn't add any information to my problem
- I will remove records with null values in the columns such as 'DEPARTAMENTO', 'PROVINCIA' and 'DISTRITO' those columns indicate the place where the infected people live.
- I will remove records with null values in the column 'FECHA_RESULTADO', since this column is important because it indicates the date of the infection.
- I will group the number of rows per day, so in this way I will get the total number of infected cases by date.

Data Splitting

Since this is a time-series problem, I will split the data in chronological order in order to get the training set and validation sets.

Model training

I will instantiate and train using models for time-series forecasting such as DeepAR, ARIMA or FB Prophet, I will test each model to see how it is better for my problem. After I instantiate the estimator which will launch my training job, I will check the R^2 metric and if necessary I will tune the hyperparameters as epochs, time_freq, prediction_length, learning_rate, num_layers, early_stopping_patience.

Evaluation

The predictor will be evaluated using R^2 or Coefficient of Determination.

References

https://espanol.cdc.gov/coronavirus/2019-ncov/cases-updates/forecasts-cases.html

http://www.coeeci.org.pe/wp-content/uploads/2020/03/Prediccion-Covid19-Peru-mar_24.pdf

https://www.biodiversidad.gob.mx/media/1/atlas/files/predictCOVID.pdf[1]

https://arxiv.org/pdf/2004.01207.pdf