Using DeepAR to predict COVID-19 in Spain

Project background and context

Following the epidemic we are facing these days and in order to give clarity to the bunch of data we are exposed daily the idea is to:

- Create **interactive dashboard** to see the status of the pandemic daily by community in Spain.
- Relevant **indicators** to see the historical and other information.
- Using this historical data by community **use DeepAR to predict the historical series** of the pandemic.
- Show this prediction also in the web dashboard.
- **Deploy the application** to be able to access publicly.
- Open source all in a **GitHub repository**.
- Document the process in **blog articles** in my personal website (in development).

The problem we are trying to solve is **to have better visualization of COVID data** in Spain and trying **to apply deep learning to predict the evolution** of the pandemic.

Talking about the background of the problem, I have been 3 weeks exploring all the different dashboards all over the world to showcase the data about the moment we are living; also reading a lot about coronavirus (papers, articles, forums…).

In order to get read in a coronavirus context we can follow:

- COVID-19 Open Research Dataset (CORD-19)
- LitCovid
- Computer Scientists Are Building Algorithms to Tackle COVID-19
- Fastai Covid 19 Forum
- Covid19 Dashboards
- Kaggle Covid 19 Forum

All this literature suggest that another interesting machine learning project would be to use computer vision to detect COVID-19 in x-ray images of patients; the decision of taking the DeepAR and dashboard project is because it requires less medical knowledge and experience. Also this Nanodegree is not about computer vision and deep learning specifically.

Roadmap of the project

The idea and the steps to success (the roadmap of the project) can follow this schema:

1. **Explore historical Covid data in Spain:** using Jupyter Notebook with standard EDA libraries (pandas, matplotlib, plotly, seaborn and some statistical libraries).
2. **Develop a web dashboard:** we will use Dash by plotly to create a web dashboard to showcase data in a cool UI.
3. **Feature engineering:** in order to improve our model performance we will add additional features, based in scientific information but also in "knowledge" as spanish inhabitant.
4. **Model** definition, training, hyperparameter tuning, testing and deploy.
5. Add **model results** to the web dashboard.
6. *(Optional)* Automate data updating.
7. *(Optional)* Test other models like Facebook Prophet or other RNN types.

Data and technology

For this project, we will be using the following data:

- COVID-19 Spain datadista data (updated at the same time that official data). (Public and free to use with source citation)
- Topojson of Spain Communities; to use the geojson information for a map visualization.
- Own data based on news and official releases.

Technology to be used:

- Python 3.6
- Jupyter Notebook.
- Pandas, matplotlib, plotly, seaborn, numpy, scipy.
- Dash.
- Amazon AWS for model development, training, testing, endpoint and batch transformation.
- Heroku or similar service to deploy the app.

Metrics and benchmarking

Based on the fact that we will be working on time-series based project and following the premises that in machine learning normally “less is more” and “simple is better than complex” we will go from simple models like Random Forest Regression or Gradient boosting traditional solutions to more complete things like simple ARIMA models, Facebook Prophet, DNN or DeepAR models.

As we are facing a regression problem we will based our predictor’s quality on RMSE (Root Mean Squared Error) and we will also do some visualization to compare our predictor vs real values.

All this model’s will be available in a Jupyter Notebook publicly and the best performing one will be also exposed in the dashboard website.