Machine Learning Engineer Nanodegree

Stefano Signor
Capstone Project Proposal

Bicing Mobility Forecasting

March 21, 2021

Overview

As part of the Machine Learning Engineer Nanodegree, a closing Capstone Project is required to complete the nanodegree. The present document sketches the choiced subject and the machine learning algorithms and techniques that will be used to solve the selected problem.

The coming section will outline:

- domain background — the field of application
- A problem statement — a problem being investigated for which a solution will be defined;
- The datasets and inputs — data or inputs being used for the problem;
- A solution statement — the solution proposed for the problem given;
- A benchmark model — some simple or historical model or result to compare the defined solution to;
- A set of evaluation metrics — functional representations for how the solution can be measured;
- An outline of the project design — how the solution will be developed and results obtained.

Domain Background

The city council of barcelona provides a bicycle sharing system commonly known as Bicing. The utility cycling is based on a short term hire scheme and its subscription service provides approx. 7000 bikes available to all citizens of Barcelona willing to pay a small year fee. The bike's fleet includes electrical and mechanical bikes, docked on stations (517) spread over the whole city area. The user can unlock the bike through a phone application and return it to any other dock station of the city.

Distilling information from the data generated by the service can help the service provided to improve the user experience, increase efficiency of the operations and help the council to take actions enhancing the policies related to mobility and the environmental impact in the city.

The prediction of bikes usage and occupancy of the station provide a good context to apply machine learning techniques, particularly time-series forecasting. Multivariate Time series can be considered for this kind of application because of the relevant number of bike stations involved whose occupation is correlated by the non random usage.

Problem Statement

To successfully fulfill the goal of bicing, a part of the regular maintenance activities, the provider of the service faces another challenge: redistribute the bikes to ensure the maximum availability to the final users. This is done by moving around bikes with customized vans from dock stations that are full to empty ones.

The activity of redistributing the bikes is done in reactive rather than in active fashion, which lead to:

- Inefficiencies: bikes may be moved unnecessarily as the users itself can contribute to the redistribution in certain situations.
- insatisfaction of the users: the redistribution of the bikes may not respond in time to the demand of bikes. Empty stations when and where there is high demand of bikes and missing spaces for docking when and where needed by the users.

Datasets and Inputs

The dataset used in this project is publically available on the website of the city hall of Barcelona, within the OpenData BCN initiative.

The dataset includes .csv files with monthly informations about:

- The dock stations characteristics, such as the geolocalization, elevation, capacity, status of the station, etc.
- The logs of the number of bikes docked for all the stations on an asynchronous base in any certain moment in time. The dataset also gives this insign on docked electric bikes of the fleet.

The data recorded in the dataset does not include information about the users or the use of the bikes by itself (probably for privacy concern) but the available information is enough to estimate, with some level of accuracy, the usage of the bikes through time and the level of occupancy of the stations.

Solution Statement

The predictive estimation of the bike usage across the bike stations can be highly effective to reduce the cost of the operations by better planning the redistribution of the bikes. A machine learning tool based on time-series forecasting can provide such predictive estimations and anticipate by weeks the required resources as well as giving insightful information for their optimization.

Benchmark Model

No benchmark public model is available to compare the results of the project work. Thanks to the good amount of data the validation of the model will be done through the tests series.

Evaluation Metrics

Besides the visual inspection of the results with comparison to the test data, the rmse (root mean square error) will be used to quantify the accuracy of the predictions of the model.

Project Design

The project will consist of code splitted in two Jupyter Notebook: a first notebook for preprocessing of the data and a second notebook making use of the sagemaker python libraries for the training of a ML model , creation of a predictor and testing of a predictor in Sagemaker.

1. Data Processing

Given the dataset available, only the change in number of docked bikes is available for each station and not actual use of bikes.

The first step in the project will require the processing of the data to extract the information of the used bikes for each station (derivative of the docked bikes). Within this process, the bikes removed and added by the users to a station shall be screened from the bikes moved by the service provider with the purpose of redistributing them or putting them in maintenance. This is achieved by assuming that the second activity usually leads to a remotion or addition of at least 5 to 10 bikes in a short period of time for the station. This data processing is applied to all the bike stations.

The sets of bike usages for each station are properly converted to be fed into the Sagemaker for the selected ML algorithm for training purpose as well as splitted in testing and training series.

The testing and training data are saved in .csv files to be then transferred into a S3 instance.

2. Training and Testing the ML algorithm

After the data processing of the training and testing data, those are uploaded into a S3 instance and the estimator is trained with data (Deep AR Estimator will be the first choice).

The resulting trained model will be used to create a predictor and generate a series of predictions to be compared with the testing set.

If a good agreement with the testing set is achieved, the predictor can then be deployed and accessed to a web-based instance that the service provider can use to plannify the operations.

Considering the important number of stations (more than 500), as first iteration, the Deep RRN algorithm will be limited to a single station at the time. So 517 independent univariate predictors can be deployed.

Further developments and margins of improvement

Few areas of development can then be investigated to improve the accuracy of the predictions. Those may or may not be explored depending on the computational resources available.

Here few proposed improvements:

- a single multivariate time-series forecasting can be trained including the data of all the stations,
- global correlation between the usage of the bikes and the weather conditions can be leveraged in the prediction model, as it is likely that bad weather will inhibit the use of the bikes.