Machine Learning Engineer Nanodegree

Stefano Signor
Capstone Project Report

Bicing Mobility Forecasting

May 20, 2021

Overview

As part of the Machine Learning Engineer Nanodegree, a closing Capstone Project is required to complete the nanodegree. The present document describes the five phases of the project development:

- Problem definition — problem to be solved and approach to the resolution
- Analysis of the problem and data — analysis of the data to better understand the problem;
- Implementation — preprocessing of the data, selection of the algorithm and metrics.
- Results — exposure of the results obtained in the implementation
- Conclusion — analysis of the results and further steps.

Problem definition

For a detailed introduction about the problem, please refer to the Project Proposal.

The project intends to analyse data publicly available of the bicing service of Barcelona and assess if a forecasting tool, using Machine Learning techniques, can provide useful informations to better plan the redistribution of the bicycles throughout the available stations.

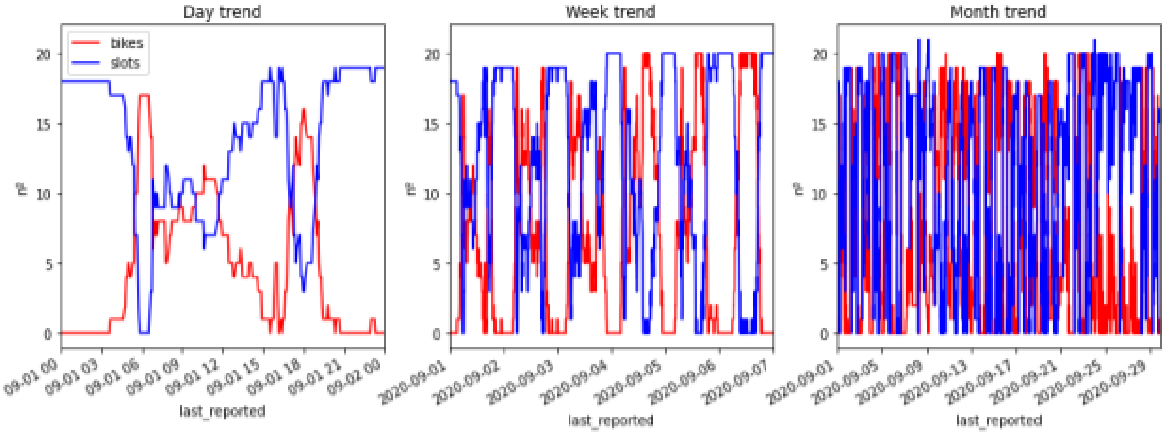
Knowing in advance the status of the station occupation can reduce the cost of redistribution of the bicycle and improve the availability of the bikes for the customers.

A prediction horizon of 1 day is considered reasonable for this task but a time period of 7 days was considered due to the seasonality involved.

Analysis

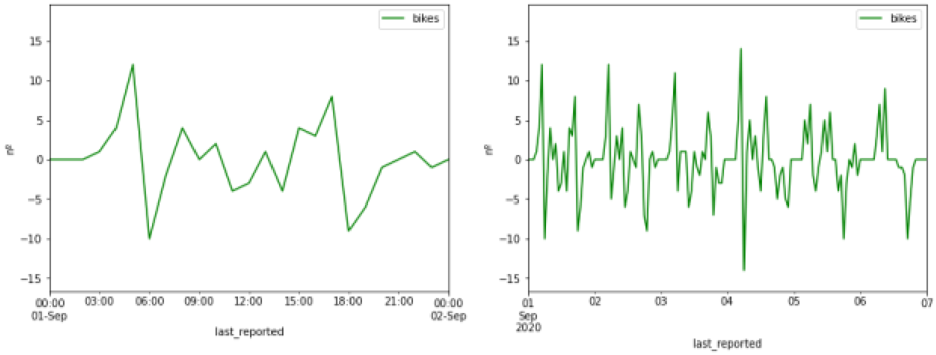
The available data includes informations about:

- 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.



Raw data of slots and bikes available for one station

Those data are not directly useful to calculate the usage of the bike service. For this reason the data were processed to obtain the actual bikes in use related to each station where they were docked or undocked. The figure above shows the bike and the slots available for a station during one day, one week and one month.

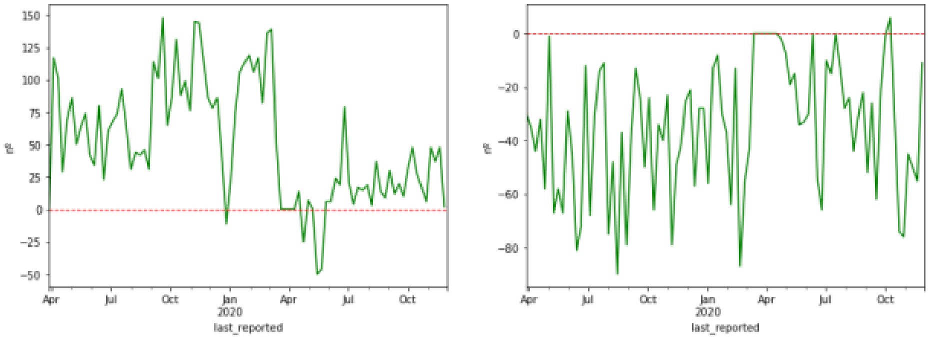


Calculated bicycle usage for one station

From the variation of the number of slots and docked bikes it is possible to calculate the number of bicycles that were docked and undocked to a specific station, like in the figure above. This is the basic information that we can use to predict the bicing usage.

Few difficulties shall be overcome:

- change in the number of bicing can be also due to the replacement activity operated by the service. The bikes are redistributed changing the number of bikes and slots in the stations. Usually this operations involves a conspicuous number of bike docked or undocked in a short period of time, so spurious effect on the data can be mitigated by cutting off rate changes in the number of bikes and slot above a specific threshold.



Cumulative bicycle usage for two stations removing redistribution

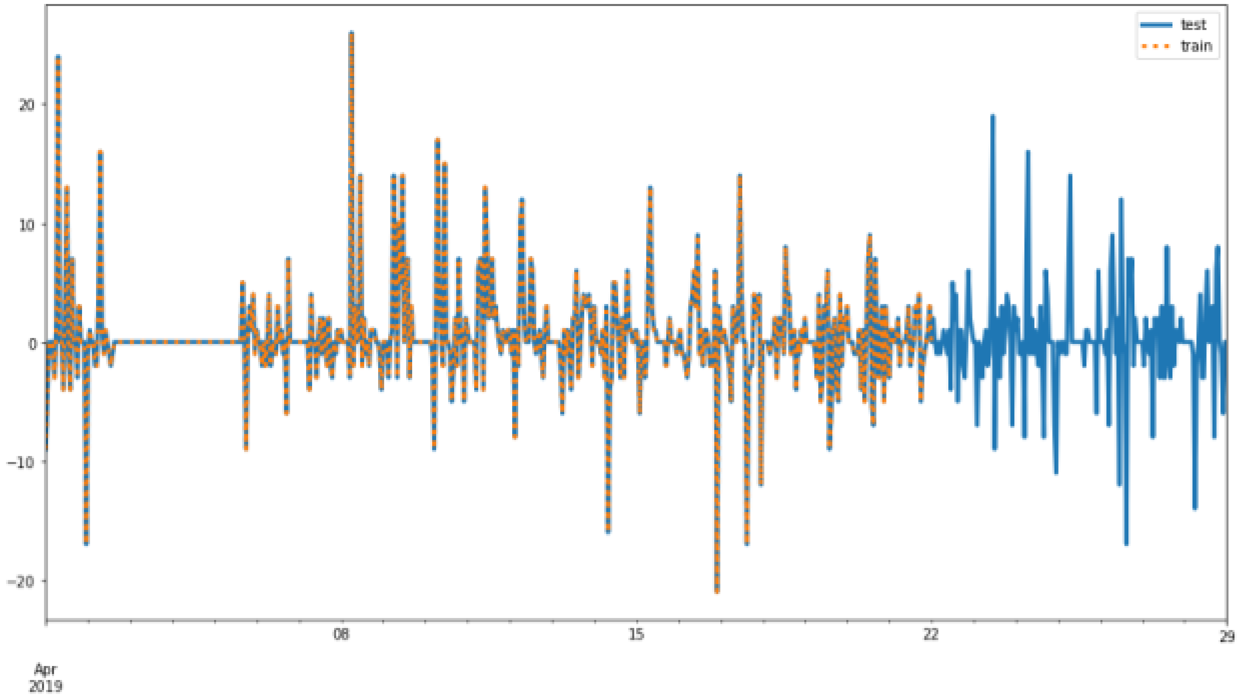
The figure shows how, removing the redistribution carried by the service (with the strategy described above), the bike transfer due to customer usage produce in the large term imbalance of docked bikes between stations.

- The station may saturate or empty while customers would actually benefit from docking or undocking a bike. This "potential" usage cannot be easily predicted from those datasets, but an optimized redistribution activity driven by the prediction can mitigate the problem providing the right amount of bike and slot at every time.

Implementation

The model used was the Sagemaker Deep AR. As per implementation DeepAR employs LSTM-based recurrent neural network architecture to the probabilistic forecasting.

The train/test sets are split in intervals of months with prediction length covering the last week of each month. The picture below shows test and train data for April 2019. The dataset is provided for several months.



One month test and train dataset for model training

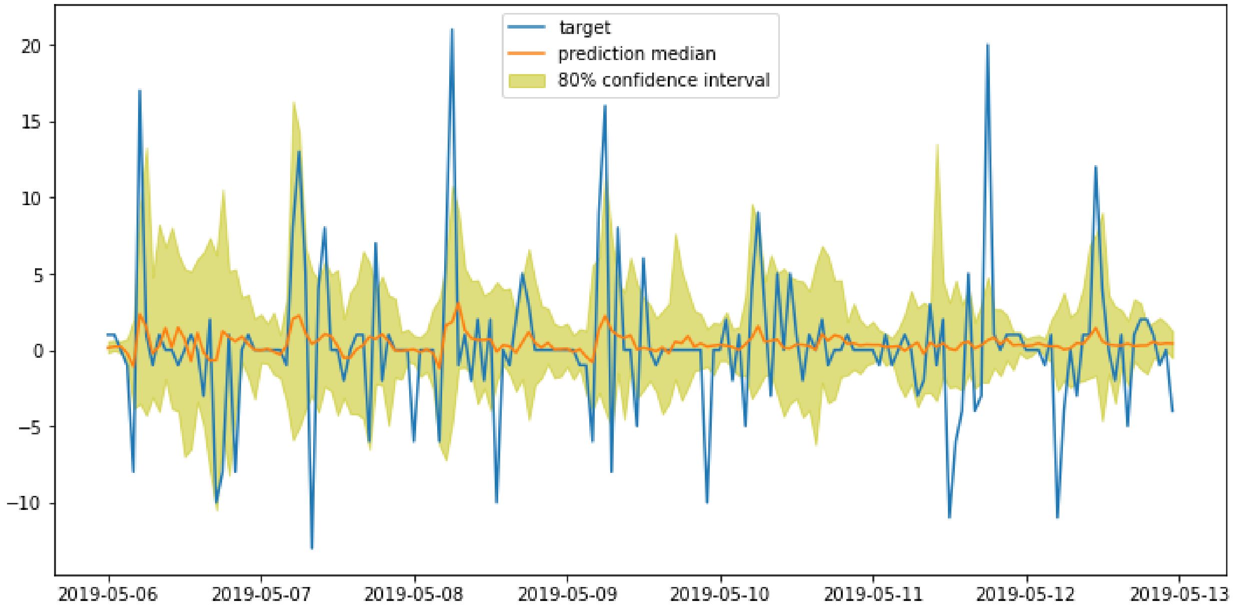
The hyperparameter setup used to train the model is the following:

```
hyperparameters = {
    "epochs": "50",
    "time_freq": "H”,
    "prediction_length": "168",
    "context_length":: "168”,
    "num_cells": "50",
    "num_layers": "2",
    "mini_batch_size": "128",
    "learning_rate": "0.001",
    "early_stopping_patience": "10"
}
```

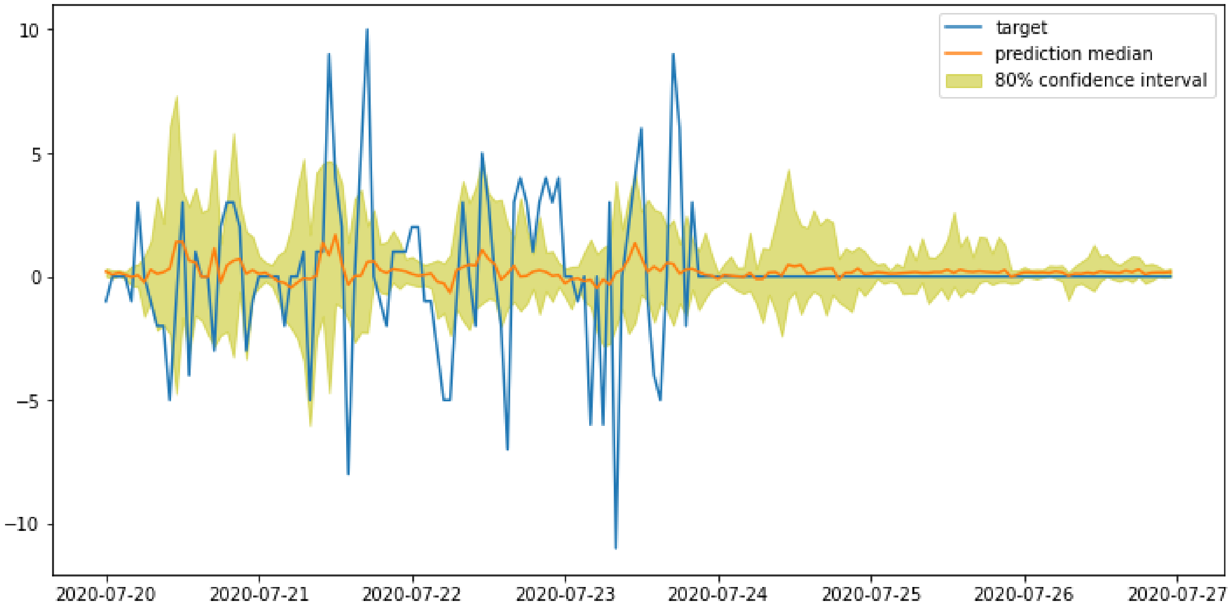
A total number of two hidden layers with 50 cells each is used in the neural network. The input data was averaged on an hourly base covering the 7 days of the forecast horizon.

Results

The predictive estimation aims to estimate the probability distribution of a time series’ future given its past, results exposed here are based on a 80% confidence margin and the prediction median. The prediction is limited to the week horizon and, as it can be observed in the figures, the seasonality, as well as peaks of usage are detected by the model and in accordance with the data.



When the station is out of service the usage prediction tends to converge to zero, even though the status of the station is an available data that can be used to override the prediction based on the time series.



Conclusions

The estimation produced by the prediction model shows clearly the daily seasonality of the bicycle usage for the station. The 80% confidence margins still underestimate in average the real usage of bicing.

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. It is clear that the un-docking of a bike from a station is reflected into the docking in another station within a limited amount of time. The total amount of bicycle is also limited

- the global correlation between the usage of the bikes and the weather conditions can be leveraged to improve the accuracy of the predictions, as it is likely that bad weather or low/high temperatures will inhibit the use of the bikes.
- Improvement in the amount and quality of data: one of the main constraints is the type of data available. No direct data of the bicycle usage is available, either if the bicycle is in transit or maintenance. This information could help identify patterns in the movements of the bicycles through the city which can be valuable to improve the accuracy and better plannify the redistribution.