BiciMad Bike-Sharing System Dataset creation and Preliminary Hourly Demand Prediction using Machine and Deep Learning Approaches

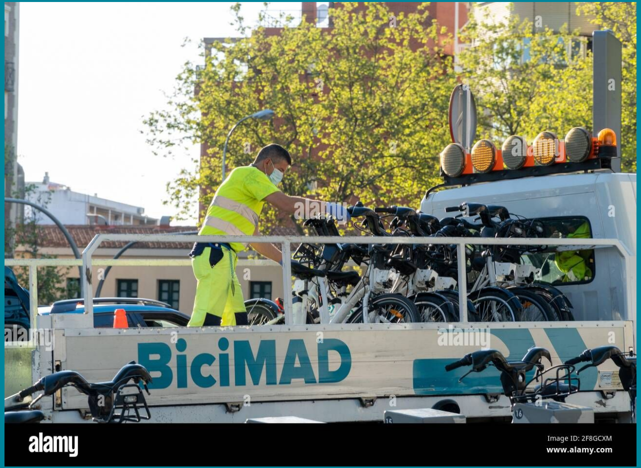
Erika Gutierrez
Javier Roset

Barcelona School of Economics

INTRODUCTION

Barcelona School of Economics

Bikesharing Systems



1) Bikes station

2) Bikes restocking

Objective

Creation of a dataset of BiciMAD bike-sharing system data **oriented to bike demand prediction** that covers from 2019 until the end of 2022, as well as a **preliminary prediction approach** using state-of-the-art techniques such as **XGBOOST and GCNN**

Scope

- Creation of a coherent BiciMAD dataset
- Preliminary prediction study using state-of-the-art techniques (XGBOOST and GCNN)
  - Prediction of plugs and unplugs in each station at each hour
- Hyperparameter optimization of an XGBOOST
- Analysis of the results obtained

- data analysis to determine the quality of the final dataset
- Complete feature engineering to exploit all the potential of the data in terms of predicting power
- Research to find the best model to predict demand in BiciMAD bike-sharing system

Literature

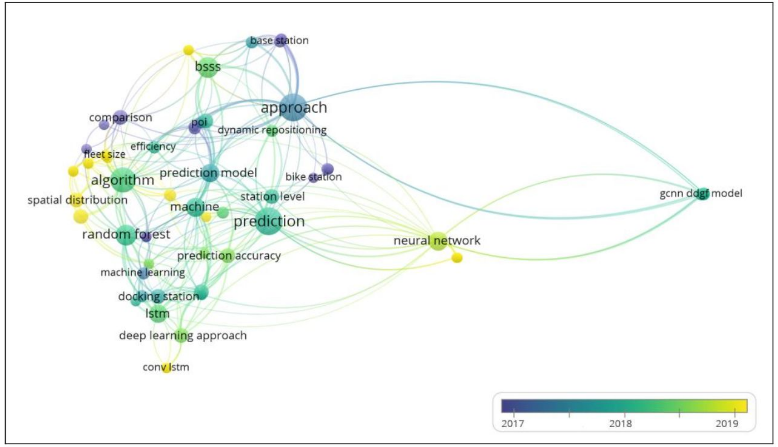


Figure 10. Title and abstract text occurrence by year overlay visualisation.

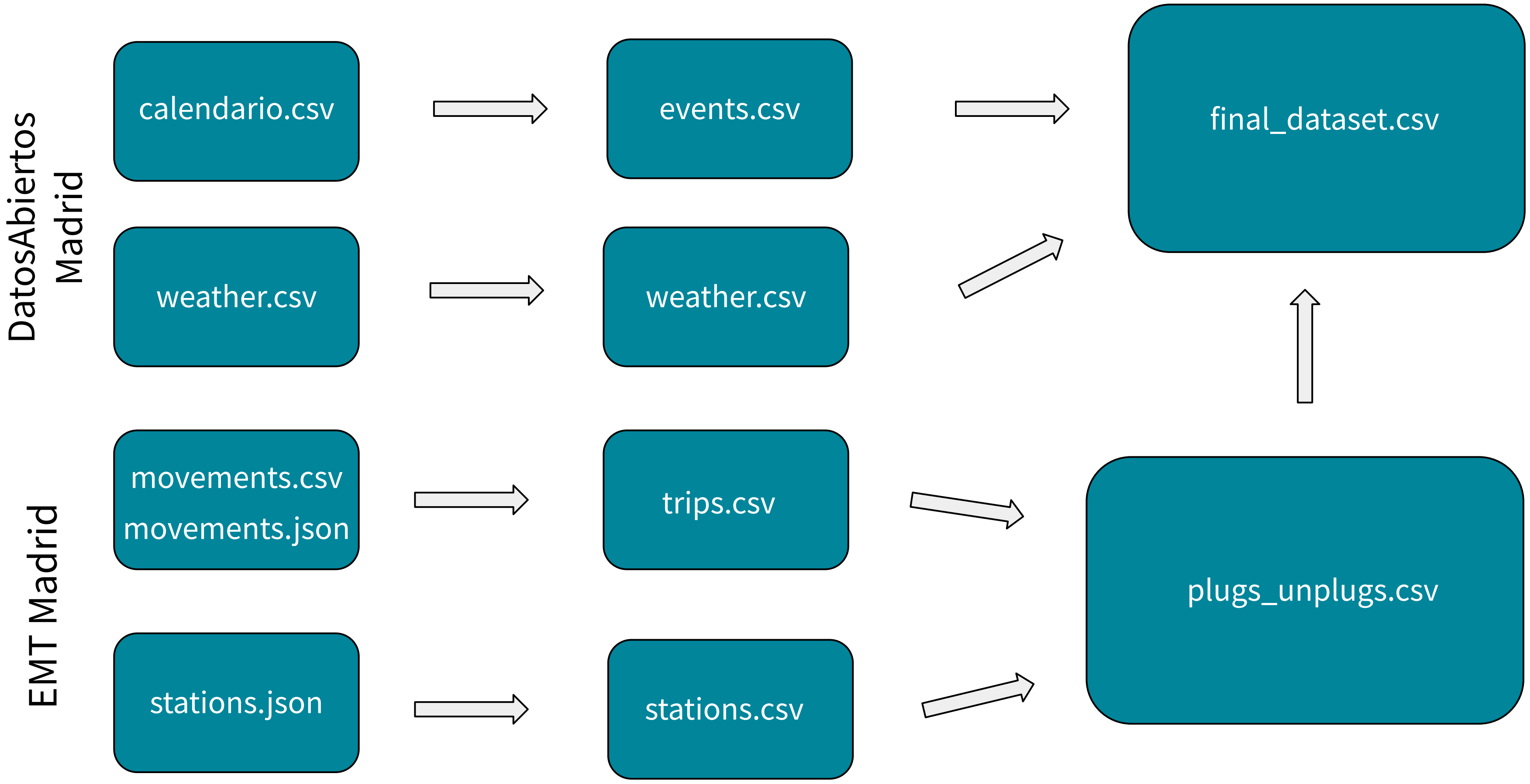
Papers used to guide our project:

- TS Kim et al. (2019) “Graph convolutional network approach applied to predict hourly bike-sharing demands considering spatial, temporal, and global effects.”
- R. Guo et al. (2019)"BikeNet: Accurate Bike Demand Prediction Using Graph Neural Networks for Station Rebalancing,"
- Lei Lin et al. (2019) “Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach”

DATA

Barcelona School of Economics

Data Pipeline



Final Dataset Specifications

- Hourly level observations from January 1st 2019 to December 31st 2022 for 266 bike stations
- Dimensions:
  - 8, 185, 250 rows
  - 33 columns
- Transformed temporal data to sine and cosine expressions to capture cyclic behavior
- 2 target variables: bike **plugs** and **unplugs**
- Inclusion of weekly lag of target variables to account for weekly seasonality
- Standard scaling of all the features

<u>Example weather variables</u>

Relative humidity
Temperature
Precipitation

<u>Example time variables</u>

Week of year
Hour
Month

<u>Example bike variables</u>

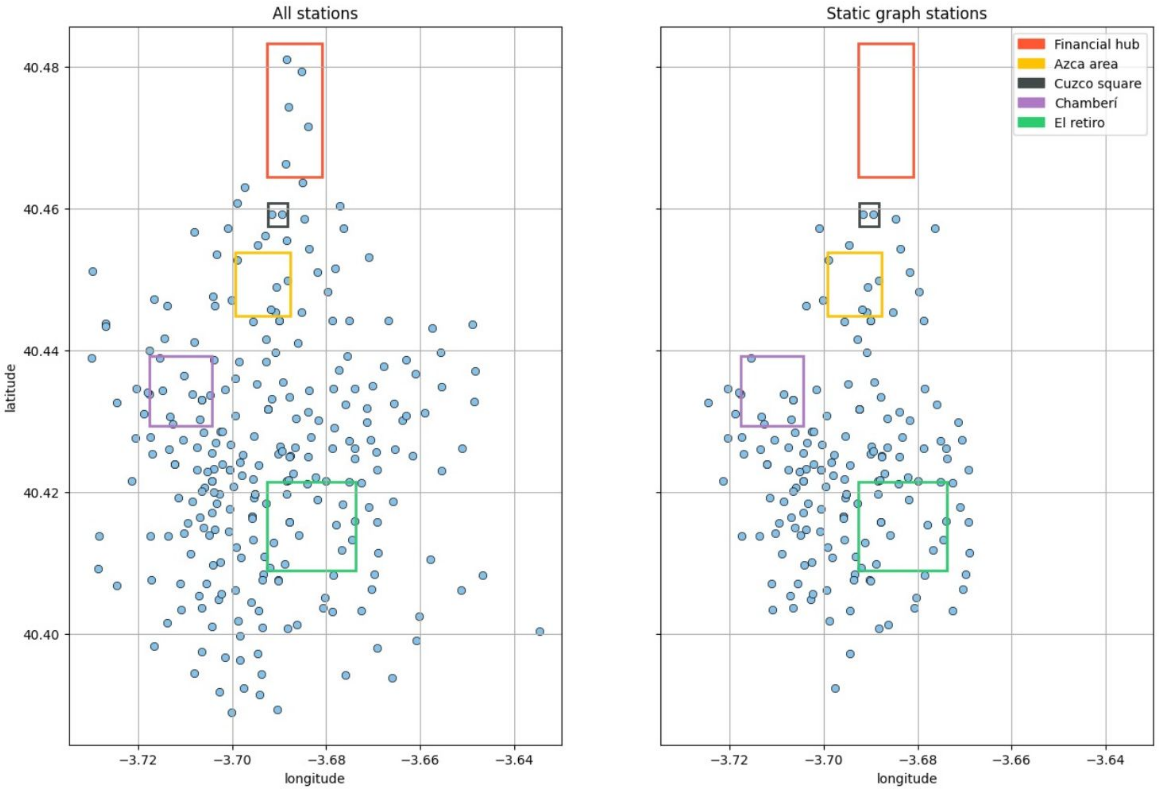
Number of reserved bikes
Station availability
Level of occupation

<u>Example events variables</u>

Workday indicator
COVID-19 indicator

Examples selected based on the level influence in XGBOOST model apart from lag variables

Spatial Component



METHODOLOGY

Barcelona School of Economics

Models

Baseline and XGBOOST

Baseline

- It predicts the value of the previous week at that time (day and hour)
- Will be used to compare our models with a trivial solution

XGBOOST

- Provided best results of Innova
- Can handle seasonality (by adding lagged variables)
- Allow to obtain feature importance of the model

Models

Graphic Convolutional Neural Network (GCNN)

GCNN

- Very popular in the literature
- Well suited for capturing spatial relationships and leveraging with graph structured data
- Can handle dynamic graphs (graphs with changing number of nodes over time)

Considerations

- We will use static graphs

Training process

Training sets creation and hyperparameter search

Splits analyzed

Innova split:

- Training: from 2019 to June 19th of 2021
- Test: from june 19th to july 3rd of 2021

2022 split:

- Training: from 2019 to January 2nd of 2022
- Test: From January 2nd of 2022 to December 31st of 2022

Hyperparameter search optimization

OPTUN

- Smart hyperparameter search, always looks to improve results
- Pruning stops useless trainings

Training process

Specificities of our approach

XGBOOST

2 approaches for training:

- Use all data to train in one go
- Use a rolling window

Considerations

- All stations where used
- Rolling window approach did not give very good results
- 1 separated model per target

GCNN

1 approach for training:

- Use a rolling window

Considerations

- Subsample of the data due to static graph approach
- Iteration lagging weather variables from next week to simulate weather forecast
- 1 separated model per target

RESULTS

Results

Metrics

Test set Innova

MAE comparison for the Innova-TSN test set

Model	Plugs	Unplugs
Innova-TSN XGBOOST	1.106	1.106
Our XGBOOST	2.362	2.371
GCNN without lagged weather	2.844	2.84

Test set 2022

Metrics of the models for plugs on the 2022 test set

Model	RMSE	MAE	R^2
Baseline	2.831	1.573	.077
XGBOOST	2.053	1.395	.331
GCNN with lagged weather	2.326	1.541	.2
GCNN without lagged weather	2.364	1.584	.168

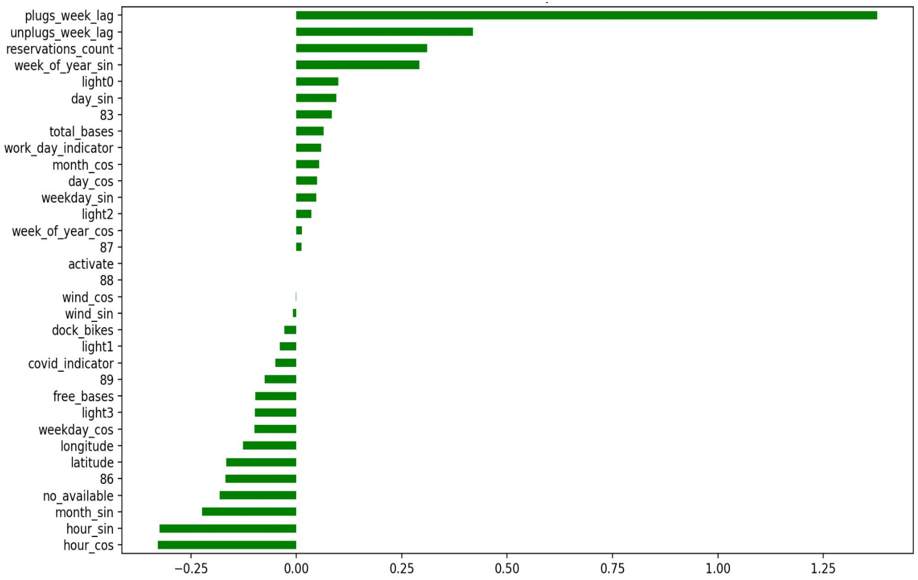
Metrics of the models for unplugs on the 2022 test set

Model	RMSE	MAE	R^2
Basic Baseline	2.831	1.573	.077
XGBOOST	2.219	1.513	.273
GCNN with lagged weather	2.464	1.635	.151
GCNN without lagged weather	2.46	1.647	.147

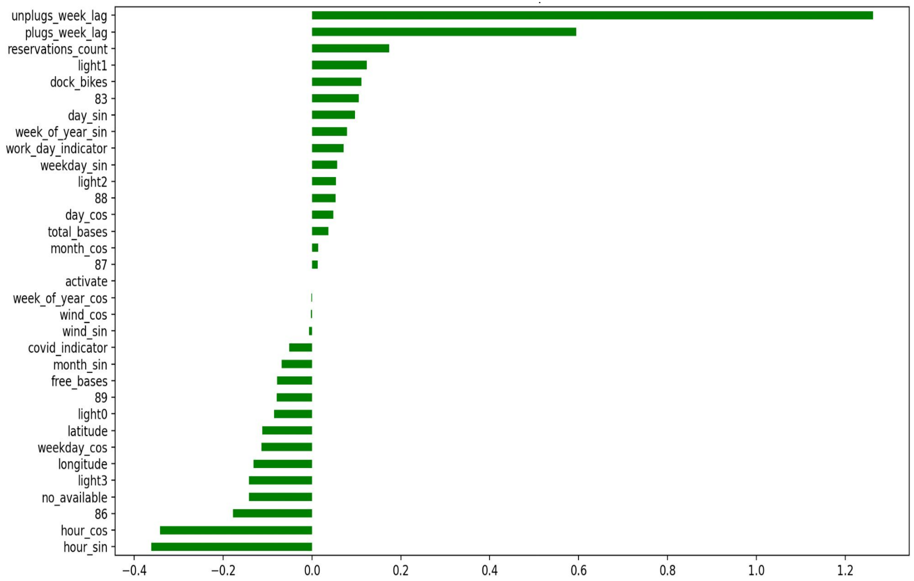
Results

Feature importances of XGBOOST

Plugs



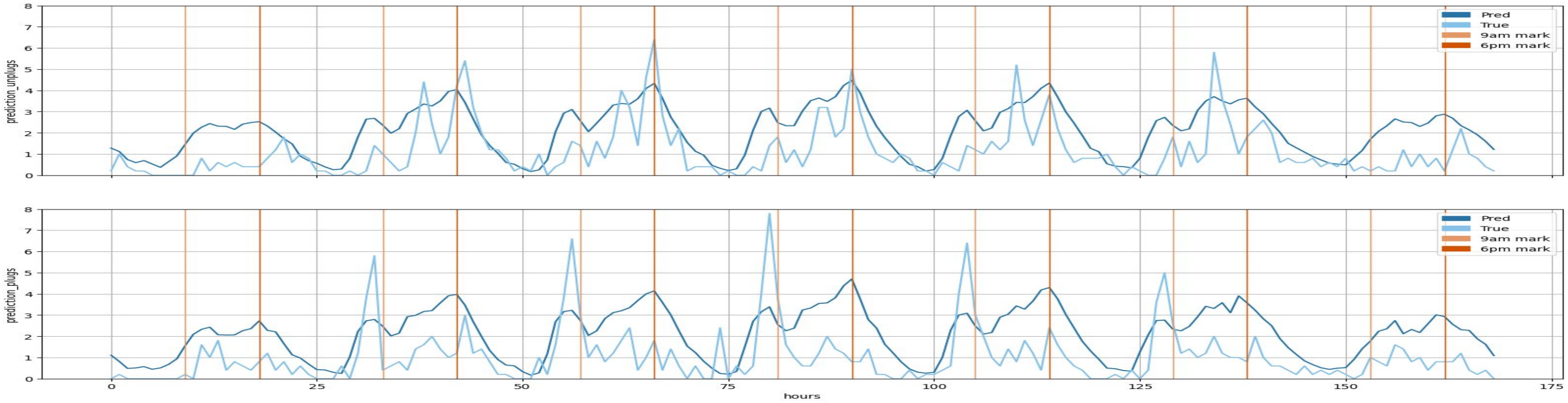
Unplugs



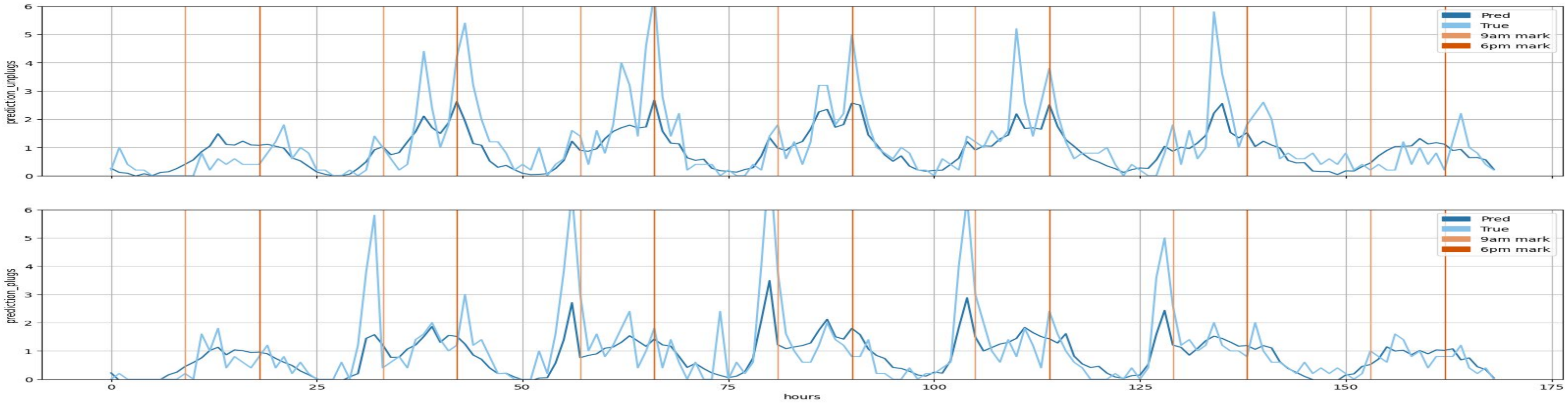
Results

Comparisons over Azca area (office district)

GCNN



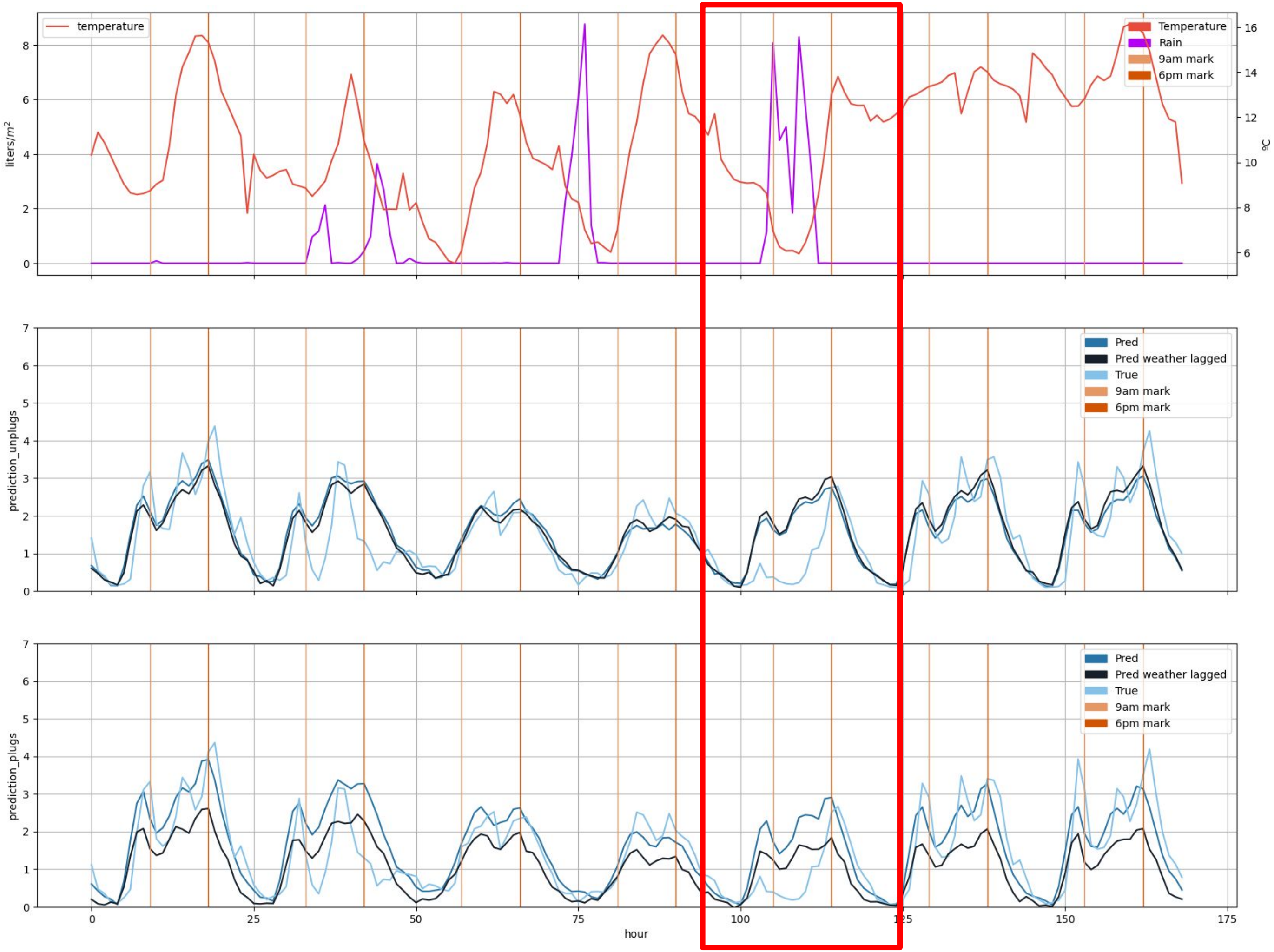
XGB



Results

Comparisons over rain

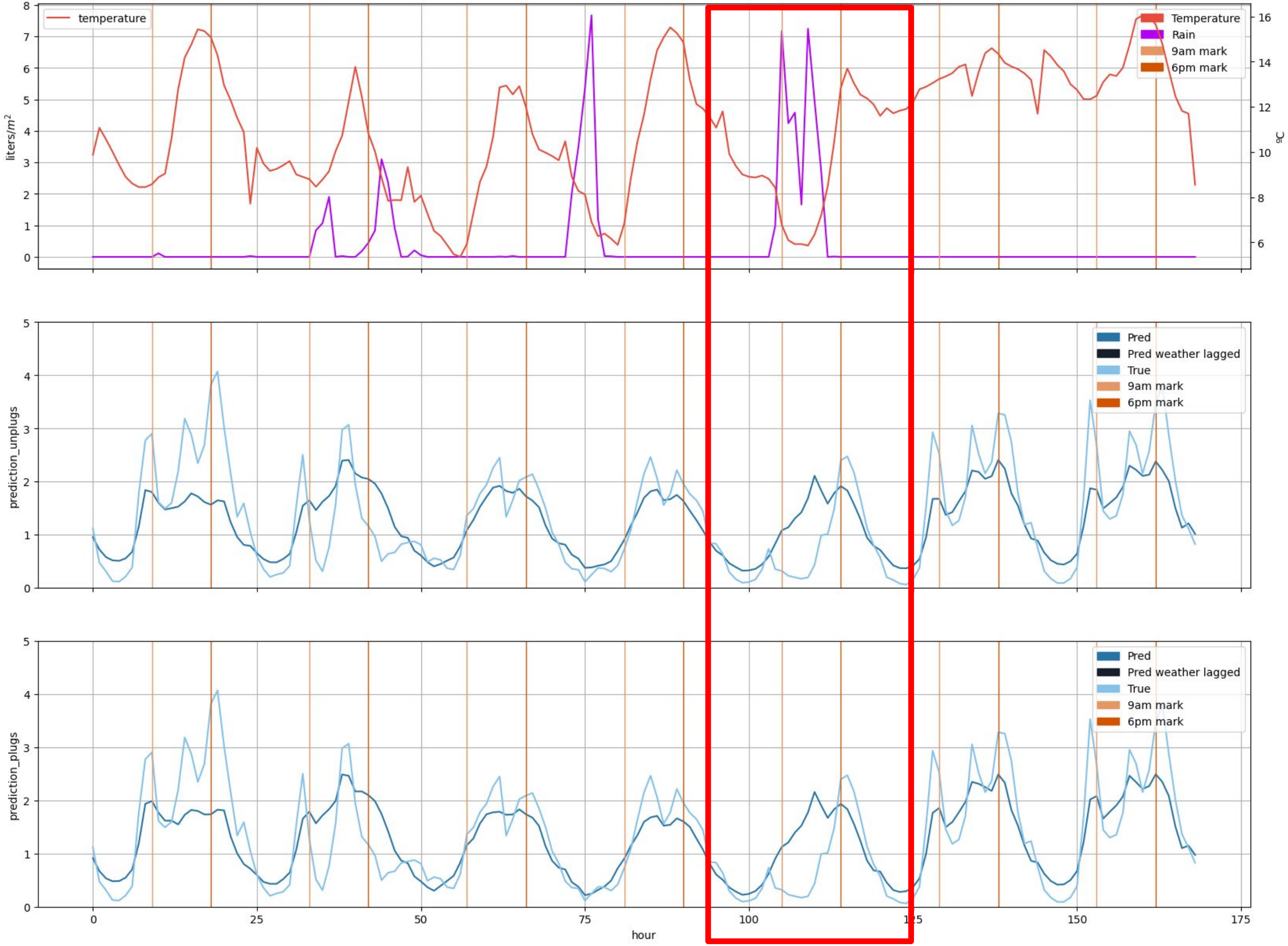
GCNN



Results

Comparisons over rain

XGB

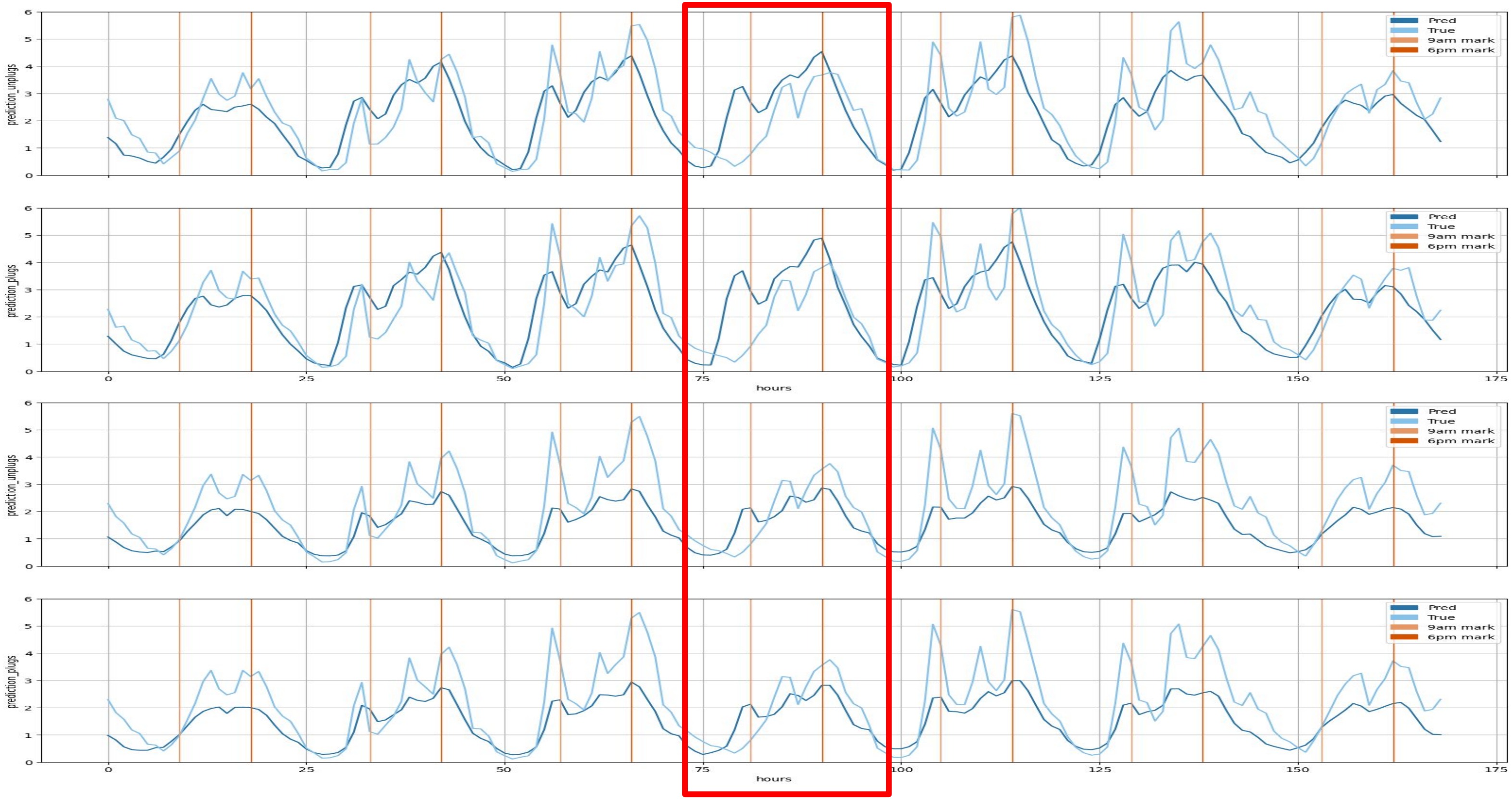


Results

Comparisons over “Día de la hispanidad”

GCNN

XGB



CONCLUSIONS AND FUTURE LINES OF WORK

Conclusions

- We have created a coherent BiciMAD bikesharing system dataset oriented to plugs and unplugs demand prediction with hourly weather data per each station
- We have done a preliminary prediction study using SOA techniques (XGBOOST and GCNN)
- We have observed how the models capture the behaviors present in the data

Future lines of work

- deep quality data assessment
- Enrich the dataframe with more variables (such as interactions with other public transports)
- Improve feature engineering to exploit the potential of the data
- Implement GCNN with dynamic graphs
- Work on AI explainability for GCNN

THANK YOU FOR YOUR ATTENTION

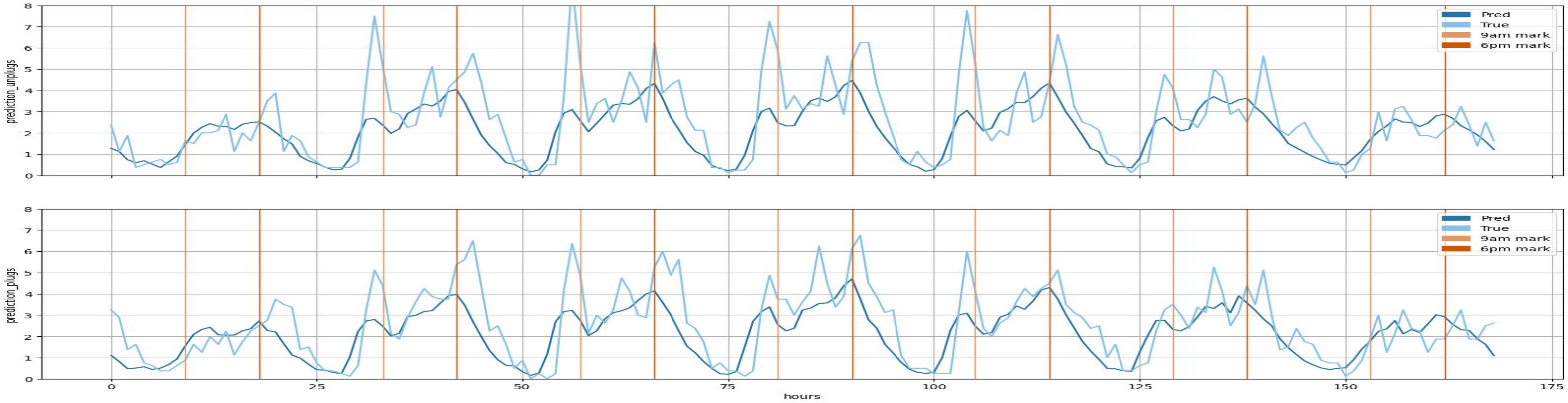
Barcelona School of Economics

APPENDIX

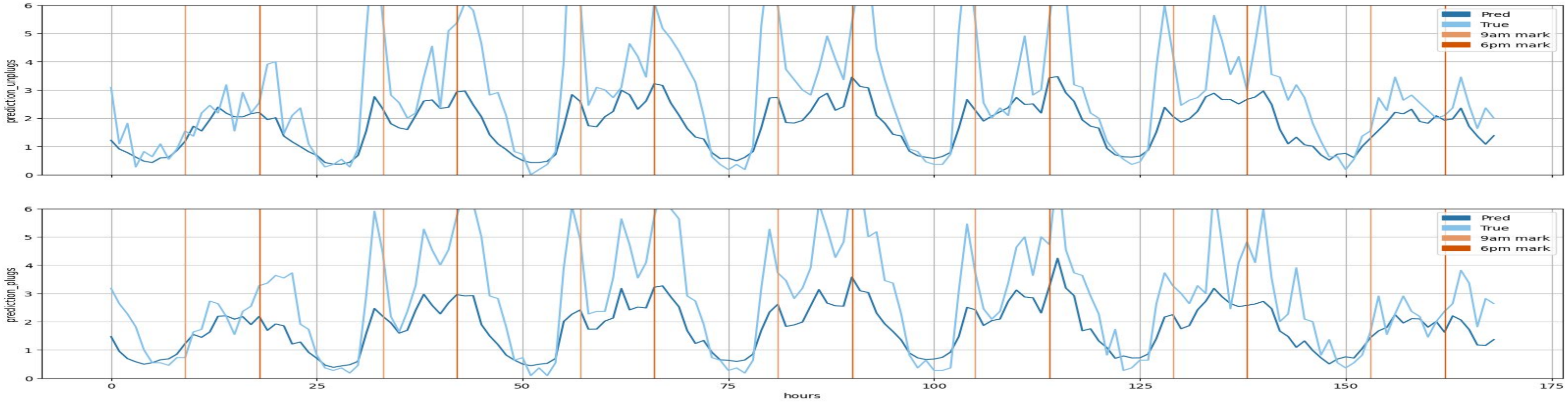
Results

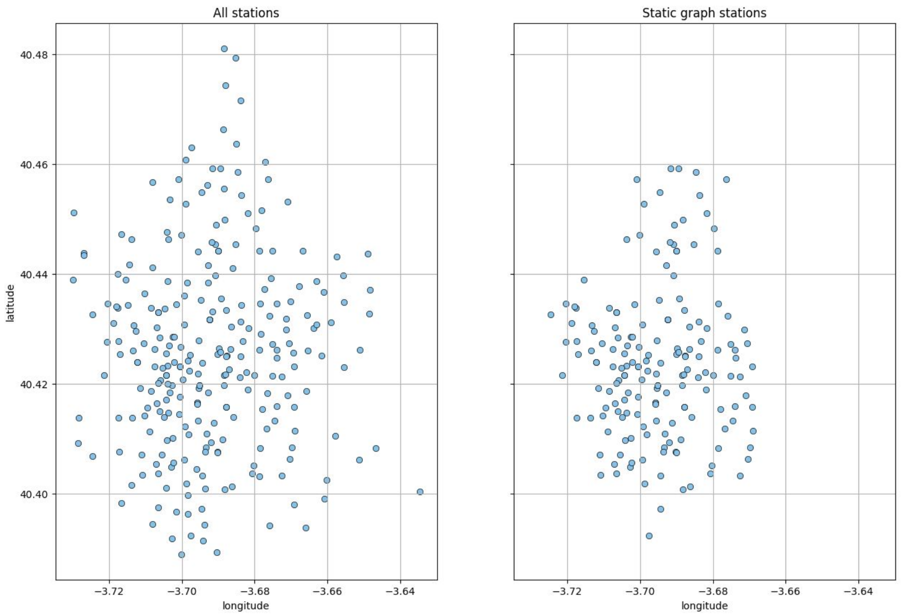
Comparisons over Chamberí area (residential)

GCNN



XGB




All stations
Static graph stations
latitude
longitude
longitude
40.48
40.46
40.44
40.42
40.40
−3.72
−3.70
−3.68
−3.66
−3.64