Predicting Bike Sharing Demand

Machine Learning Project, Hertie School

Mona Borth, Linus Hagemann, Luis Windpassinger

Bike sharing

- Key role in sustainable urban mobility, reducing car dependence and emissions in urban environments

- Prediction of station-level departures
  - Operations: Bike redistribution and fleet management
  - City planning: Identify infrastructure needs

- Weather data: Apart from cyclical patterns, most prominent factor influencing day-to-day cycling behavior

Data Sources

Trips

- All trips in 2023
- 4.467.334 rows
- Last year with station-based data
- No ID for bikes ☹️

capital bikeshare

Stations

- All stations currently in the network
- Not versioned on year ☹️
- No stable ID for stations ☹️

capital bikeshare

Weather

- Historical weather for each station in every hour of 2023

OpenMeteo

Final Dataframe

df.shape:

6.464.880 rows
94 columns

738 stations over 1 year

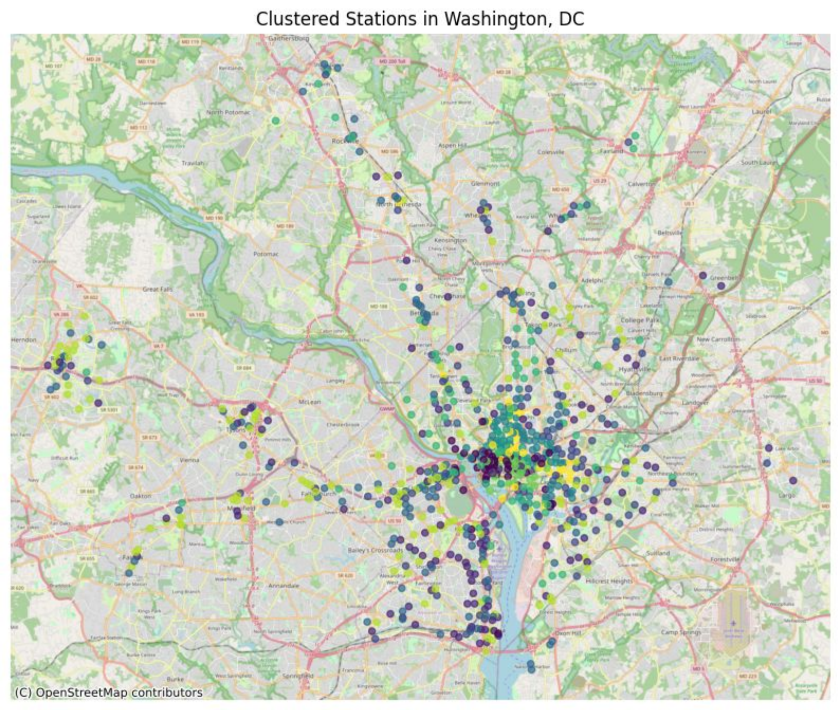
Each row represents:

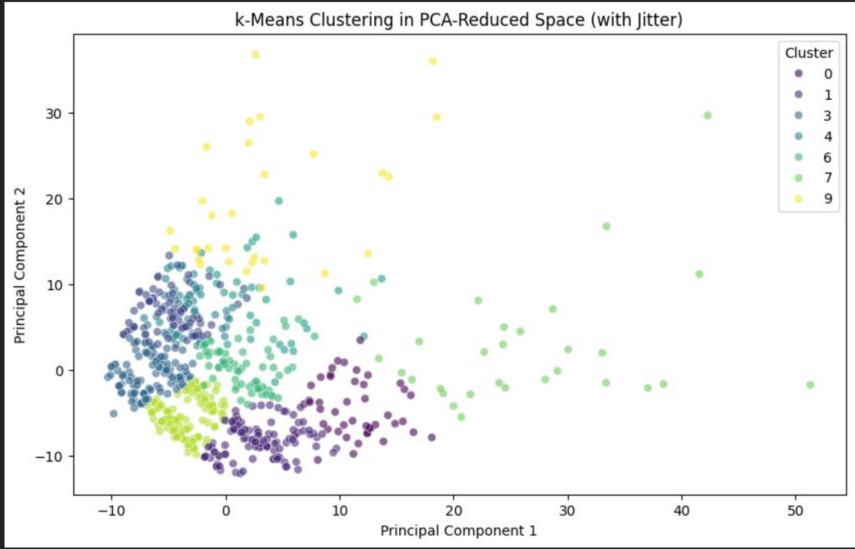
- ➢ 1 hour
- ➢ 1 station

Features include:

- ➢ Departures/Arrivals per station
- ➢ Weather at station
- ➢ Station characteristics (stable)
- ➢ Time info (work hours, holiday, …)

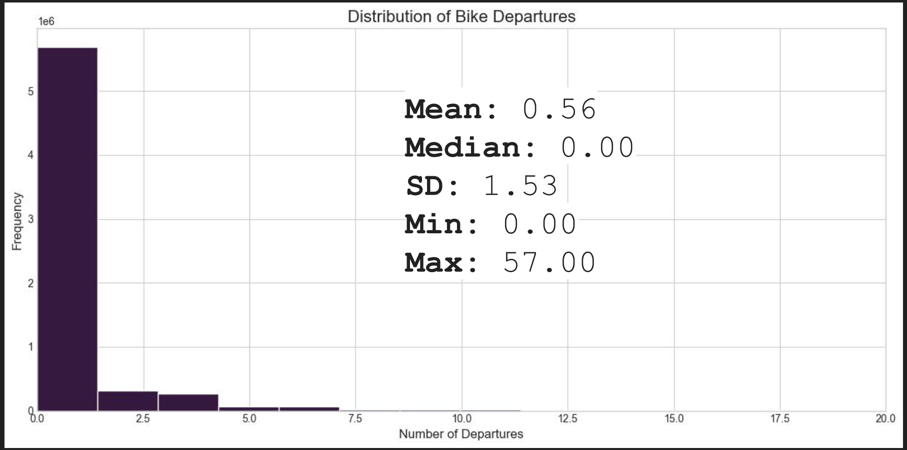
Station Clustering

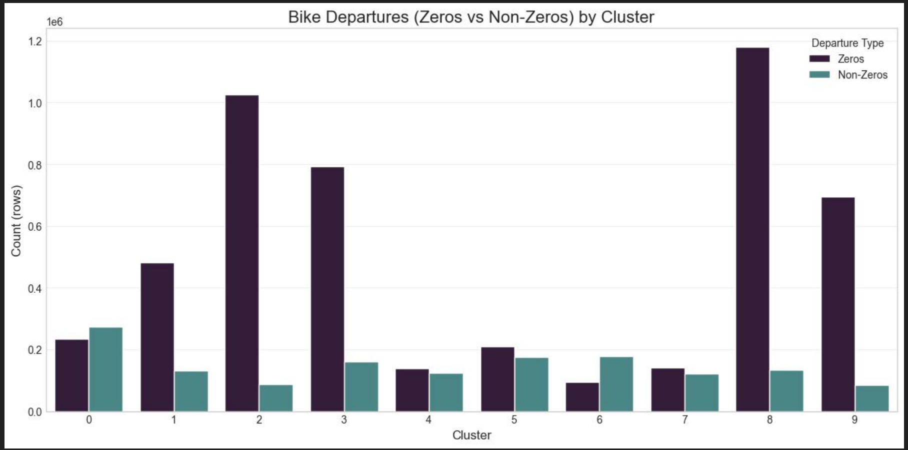




Constrained k-Means

Descriptives: Target I





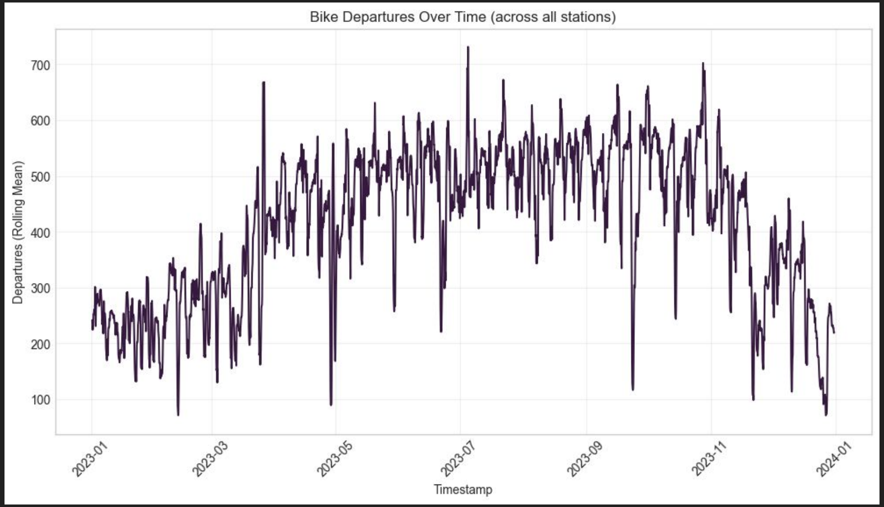
Target: Departures

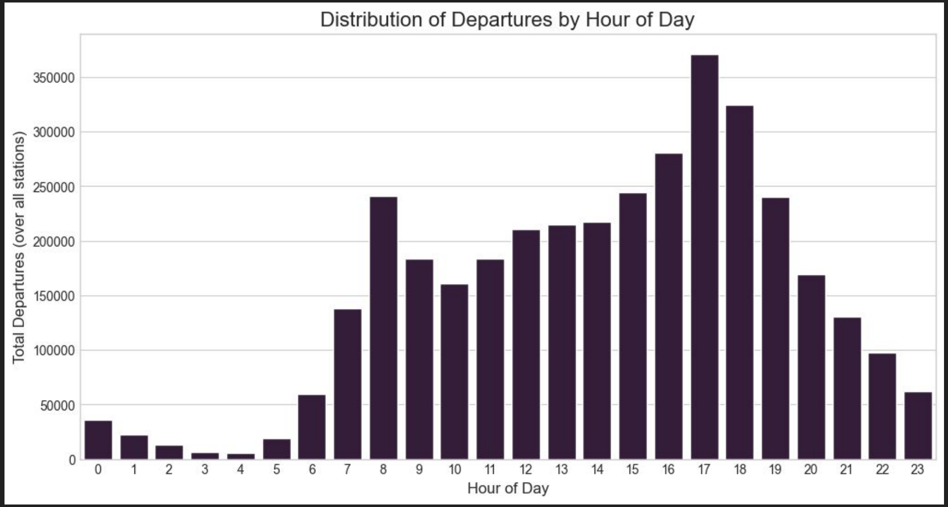
- ➢ Strong right skewed distribution
- ➢ Share of Zero: 77%

Departures per Cluster

- ➢ Imbalanced distribution across clusters (zero vs. non-zero)

Descriptives: Target II





Yearly patterns

- ❄️ Less activity
- ☀️ More activity

Daily patterns

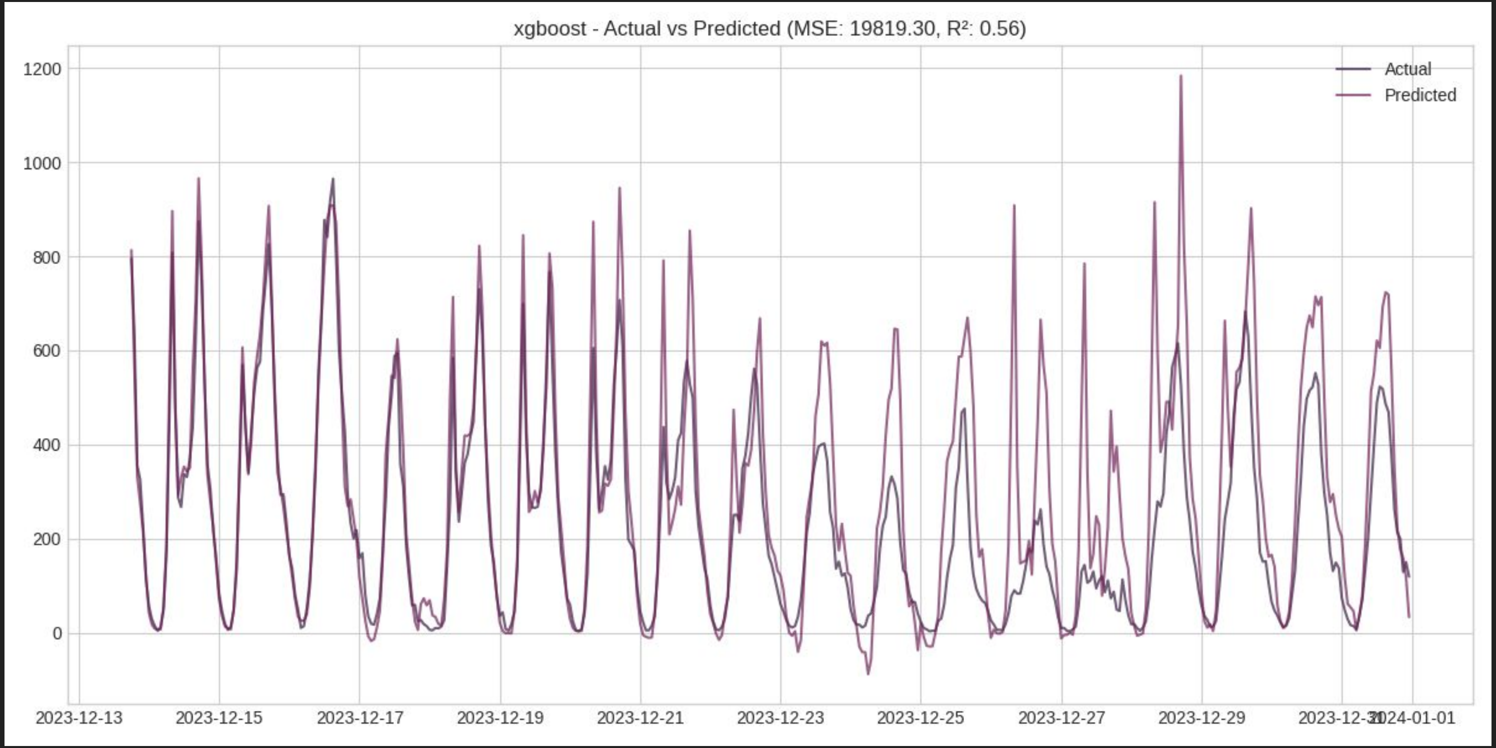
- Night hours (22-6 hr) significantly fewer departures

What Did We Do?

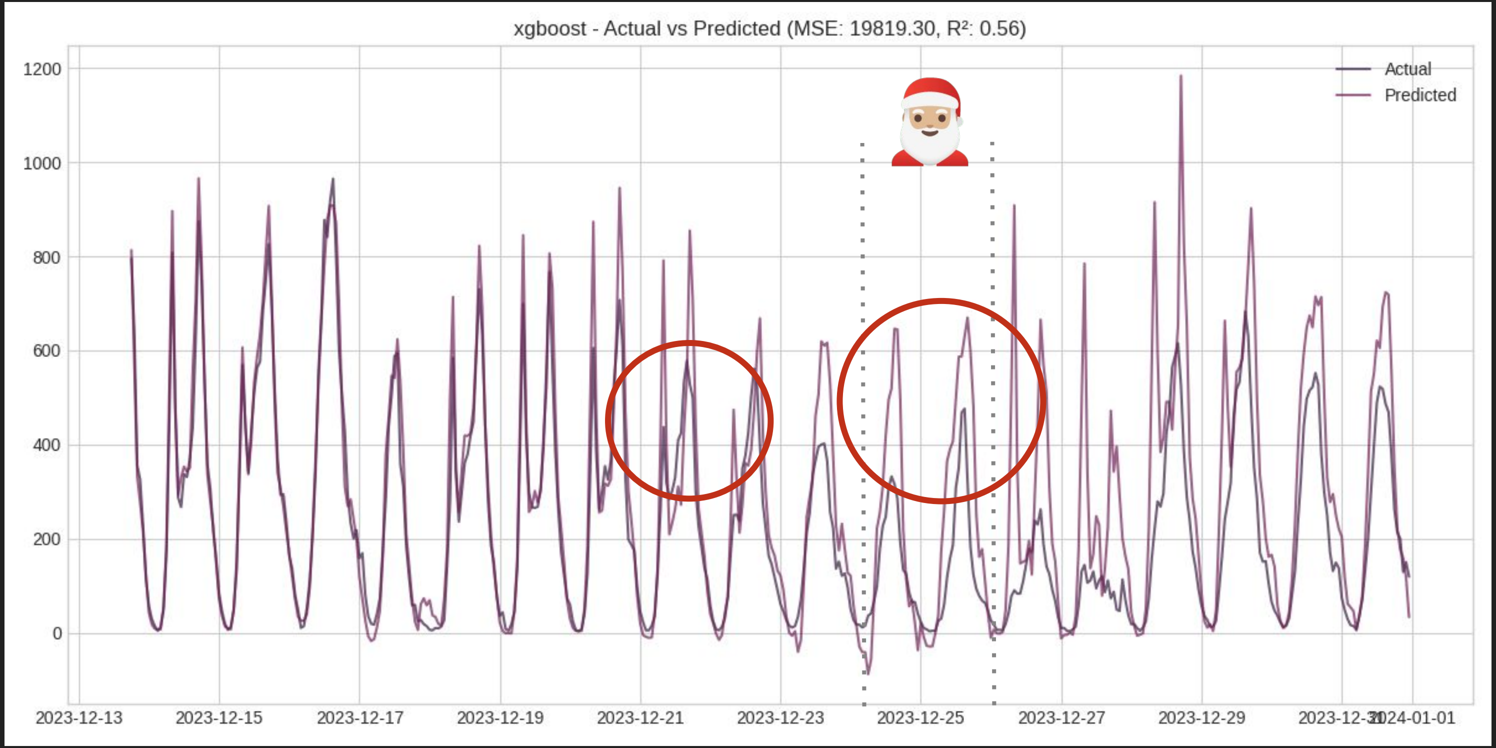
- ➢ Citywide Model for Aggregate Departures
- ➢ Unpooled Models (per Cluster)
- ➢ Fully Pooled Model
- ➢ Semi-Pooled Model
  - ○ Varying Slope - Varying Intercept - Negative Binomial Regression

- ➢ Linear Regression
  - ○ Lasso & Ridge
- ➢ Polynomial Regression
- ➢ Decision Tree
- ➢ XGBoost
- ➢ Random Forest

Results - Citywide Predictions

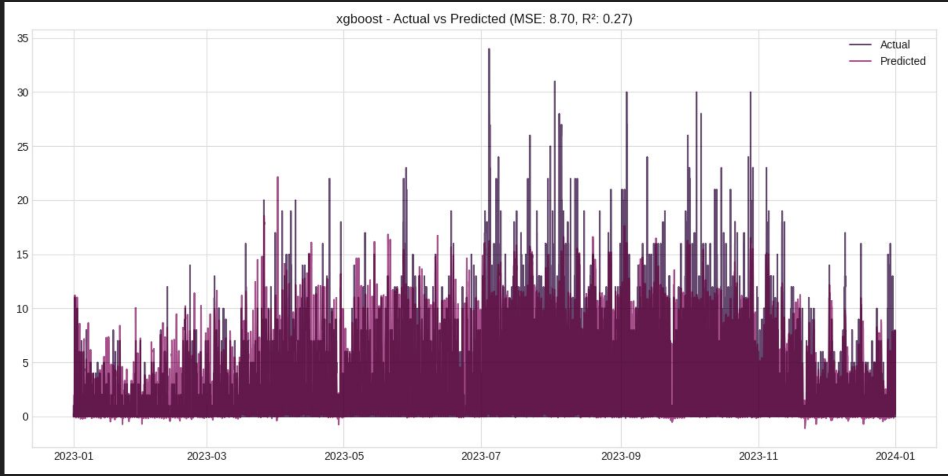


Results - Citywide Predictions

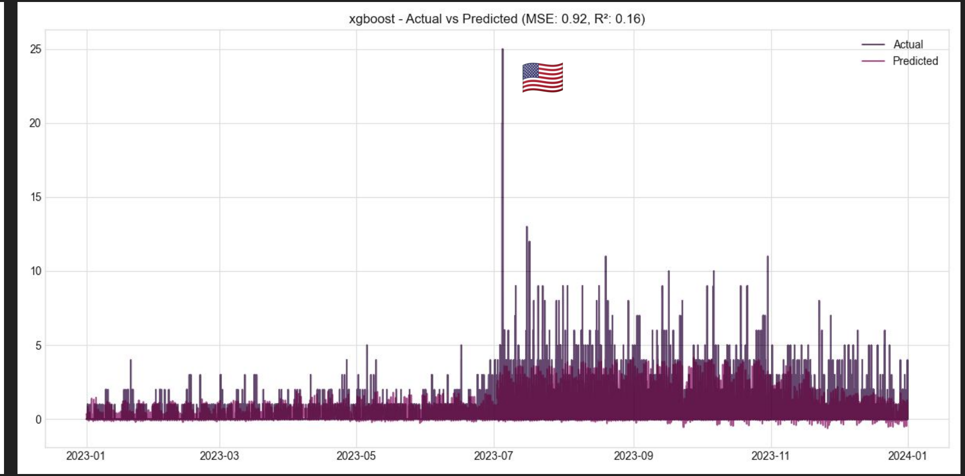


Results - Cluster Predictions

Cluster 4

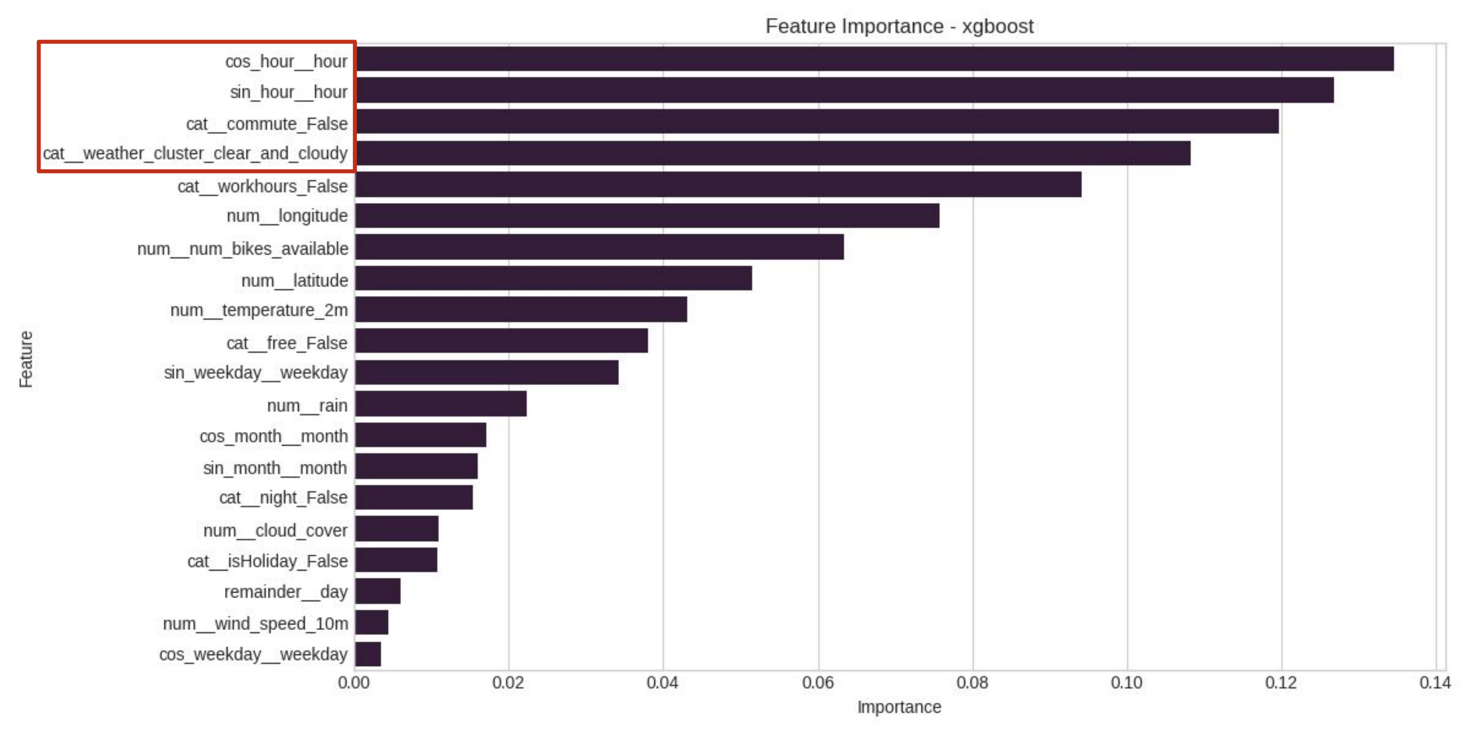


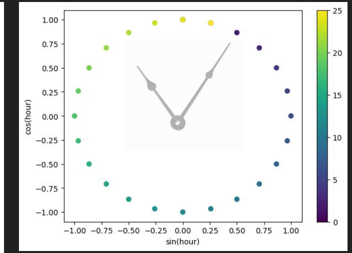
Cluster 7



Predictions on Test Set (5% of Year), with potential Time-Information Leakage

Results - Cluster 4 - Feature Importance





Challenges, Limitations & Conclusion

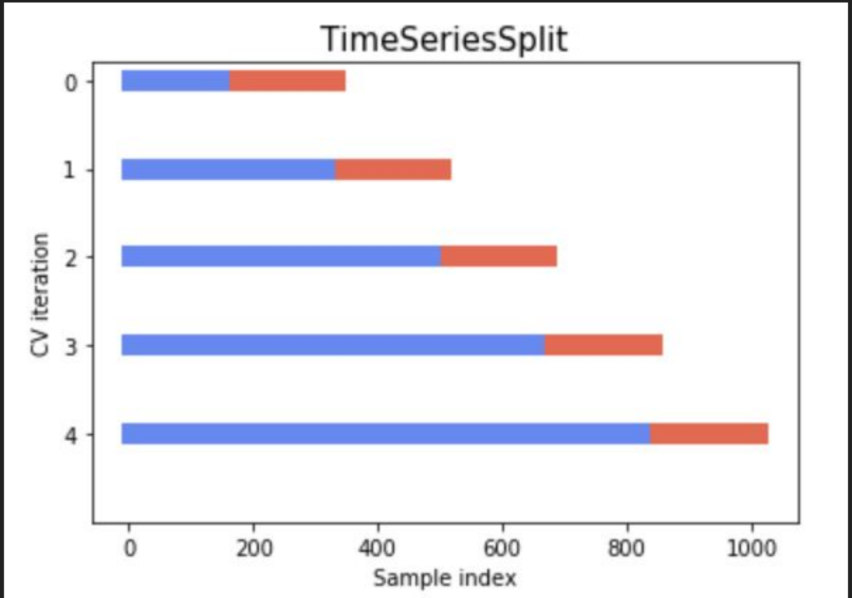
- Mixed results for the cluster approach
  - Unbalanced clusters could warrant something like Negative Binomial Regression
  - Distinct patterns got captured in the more balanced clusters (zero vs. non-zero)
- Citywide Model shows strong predictability based on time and weather

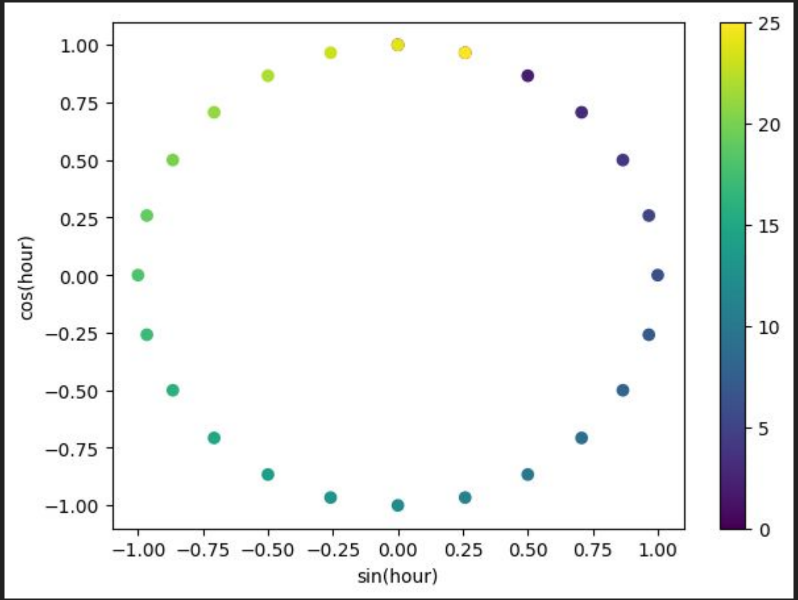
---

- Processing power (ideally: 1 model per station)
- Extend dataset
  - Individual station characteristics
  - Population density
  - Events
  - …
- We don’t know if people would rent when no bikes are available
- Open Data
  - Incompatibility issue with available datasets
  - Station capacity over time not known
  - Overall number of bikes in the system not known

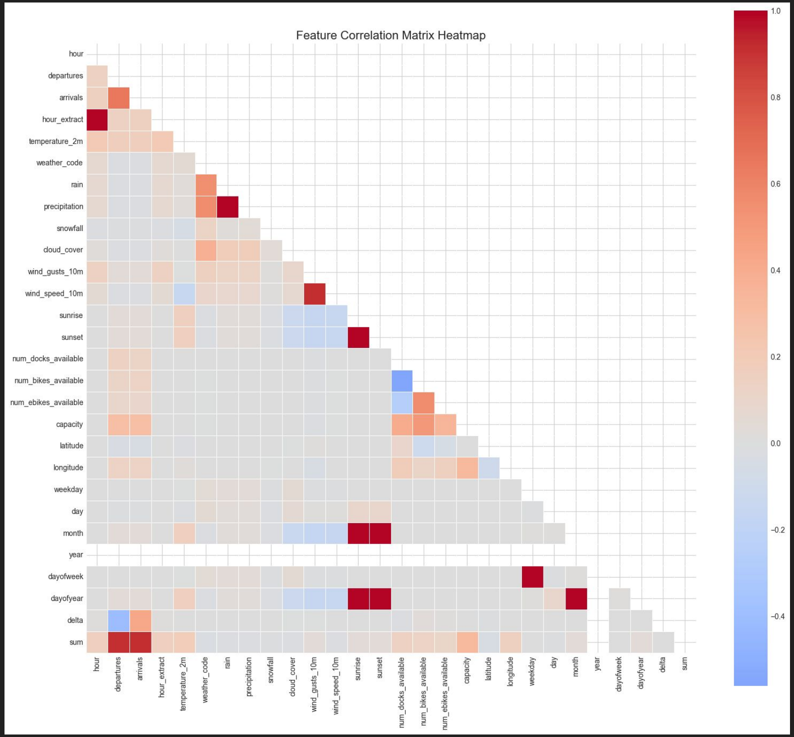

I ride my bike
I reduce CO2 emissions
I run all the models!
I run all the models!
imgflip.com

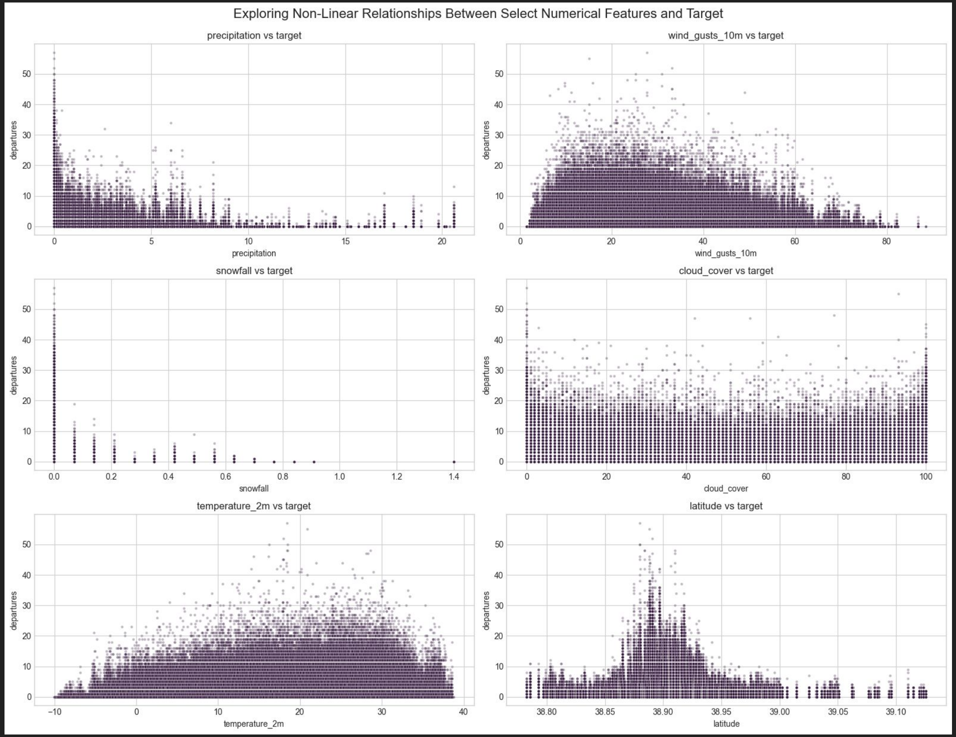
Time data specifics





Descriptives: Features





Preprocessing