PREDICTING CYCLIST TRAFFIC IN PARIS

Solal Zana & Adam Berdah

SOMMAIRE

I. PROJECT PRESENTATION

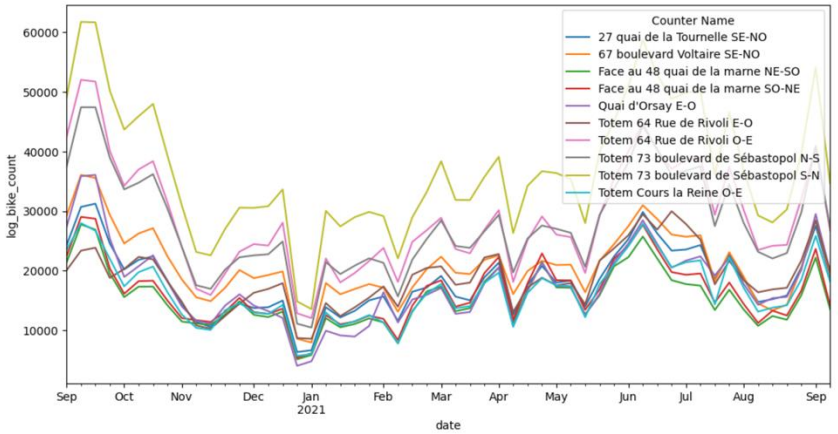
- **GOAL**: Predict bike count at a given hour across 56 counters in Paris.
- **METRIC**: Root mean square error (RMSE)
- **TIMEFRAME**:
  - Training data: September 1st, 2020 – September 9th, 2021
  - Public Test data: September 10th, 2021 – October 18th, 2021
- **DATA**:
  - training dataset partly fursnihed
  - Testing set: 50% public, 50% private.

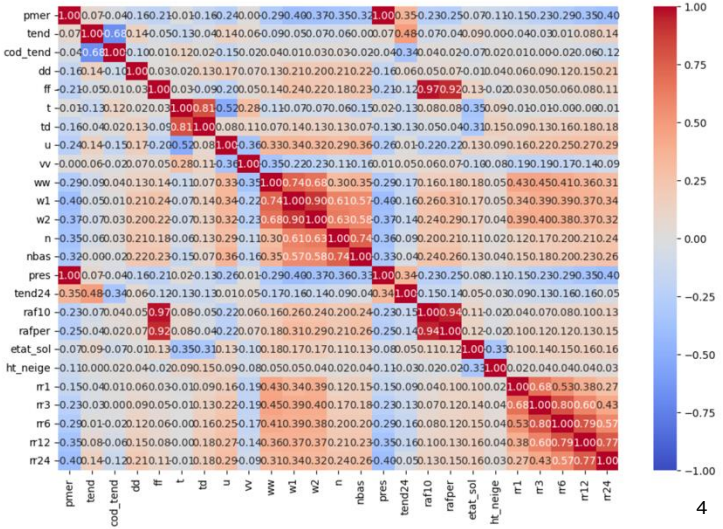
II. EDA

ADDING EXTERNAL DATA

Exploration:

- Checked proportion of missing and unique values
- Plotted relationships between features and target
- Checked the correlation table





II. EDA

ADDING EXTERNAL DATA

Cleaning:

- Missing values
  - Deleted columns containing > 10% missing values
  - Replaced the rest with the median
- Deleted columns with a single value
- Deleted some highly correlated columns

Added features:

- French school holidays (zone C)
- Bank holidays
- Quarantines
- Seasons
- Distance from city center

ISSUE: merging both datasets gave a huge error (RMSE ≈ 2.15)

II. EDA

ENCODING

Standard Scaler	One-Hot	Cyclical encoding
All numerical columns	Counter name Years Weekday	• Hour • Day • Weekday • Month

IV. MODEL EXPERIMENTATION & SELECTION

- Models tested:
  - Ridge Regression
  - Random Forest
  - XGBoost
  - Catboost
  - LightGBM

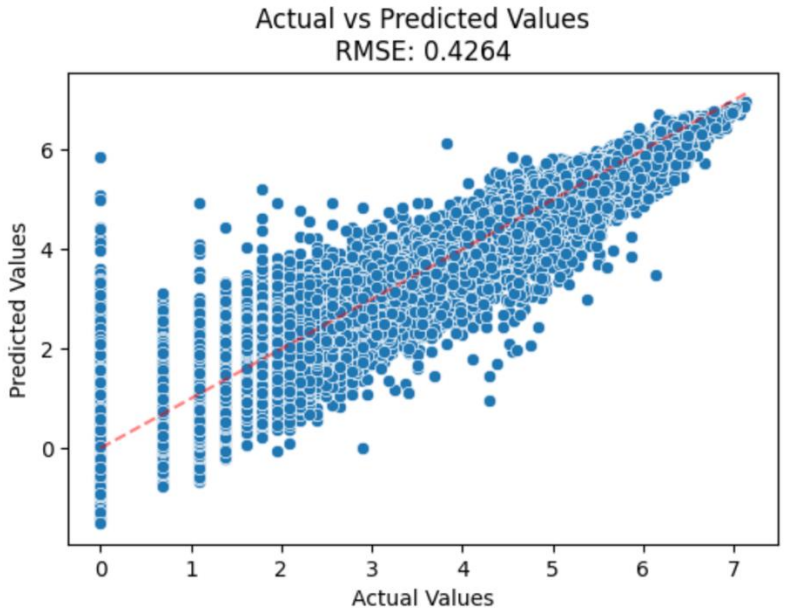
- Evaluated model performance on XGBoost using cross-validation and RMSE as metrics.

- Grid Search on the most probant model

V. EVALUATION OF FINAL MODEL

OUR FINAL BEST RESULT:

Submission and Description	Private Score	Public Score
submission_xgboost.csv Complete (after deadline) · Solal Zana · now	**0.7247**	**0.7366**



VI. FUTURE RESEARCH

- Analysis of construction work iin Paris from additional datasets ?
- Test other types of encoders ?
  - Gap encoder or Target encoder
- Integrating additional data like geographical traffic density or inhabitants density
- Exploring neural networks

THANK YOU FOR YOUR ATTENTION