

Predicting Cyclist Traffic in Paris

56

We had to **predict bike count** at a given hour given 56 counters in Paris

RMSE

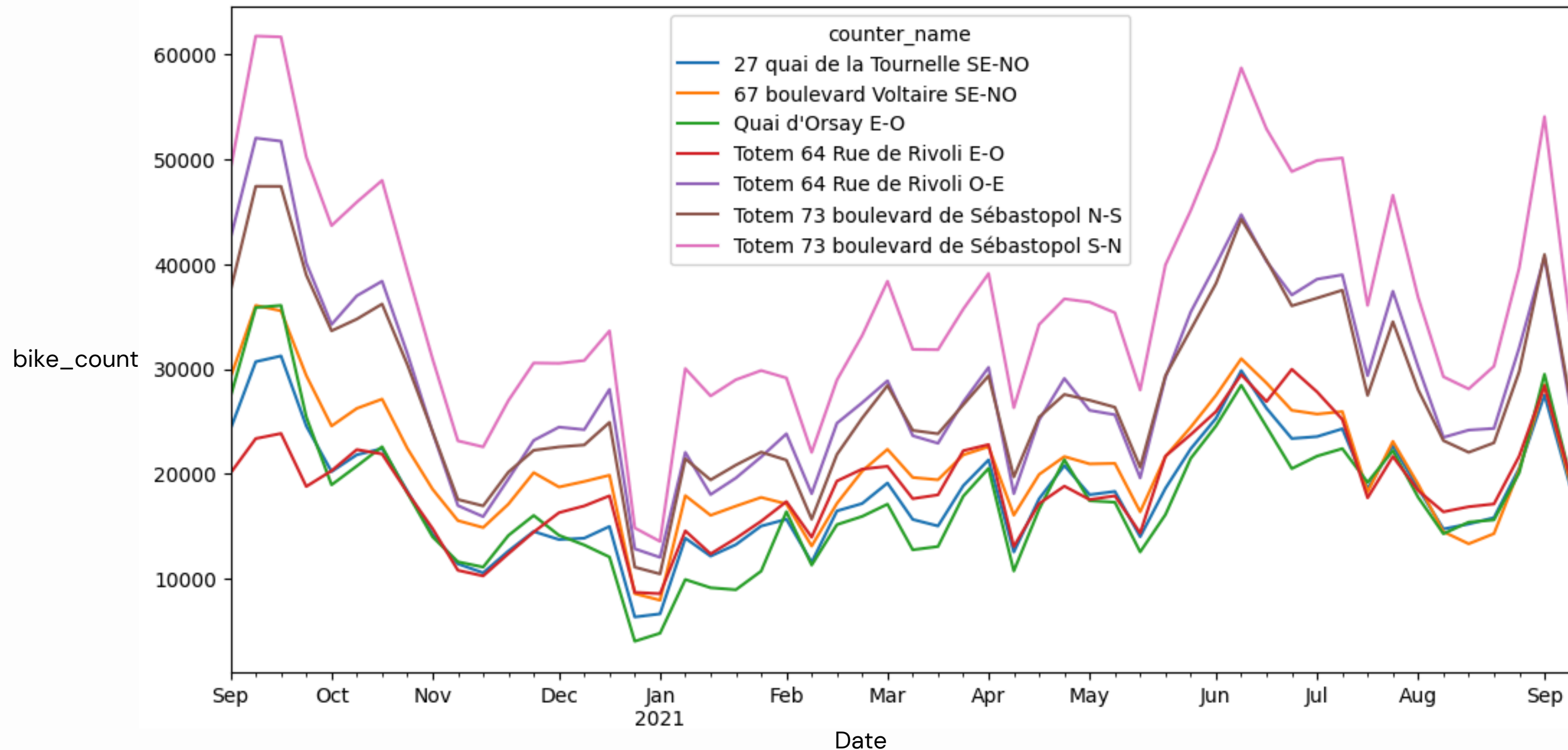
We used **RMSE** for our prediction: The Root Mean Squared Error

Time



The **timeframe** for our training data was: 2020-09-01 to 2021-09-09

Exploratory Data Analysis

bike_count per counter



Issues faced - merging

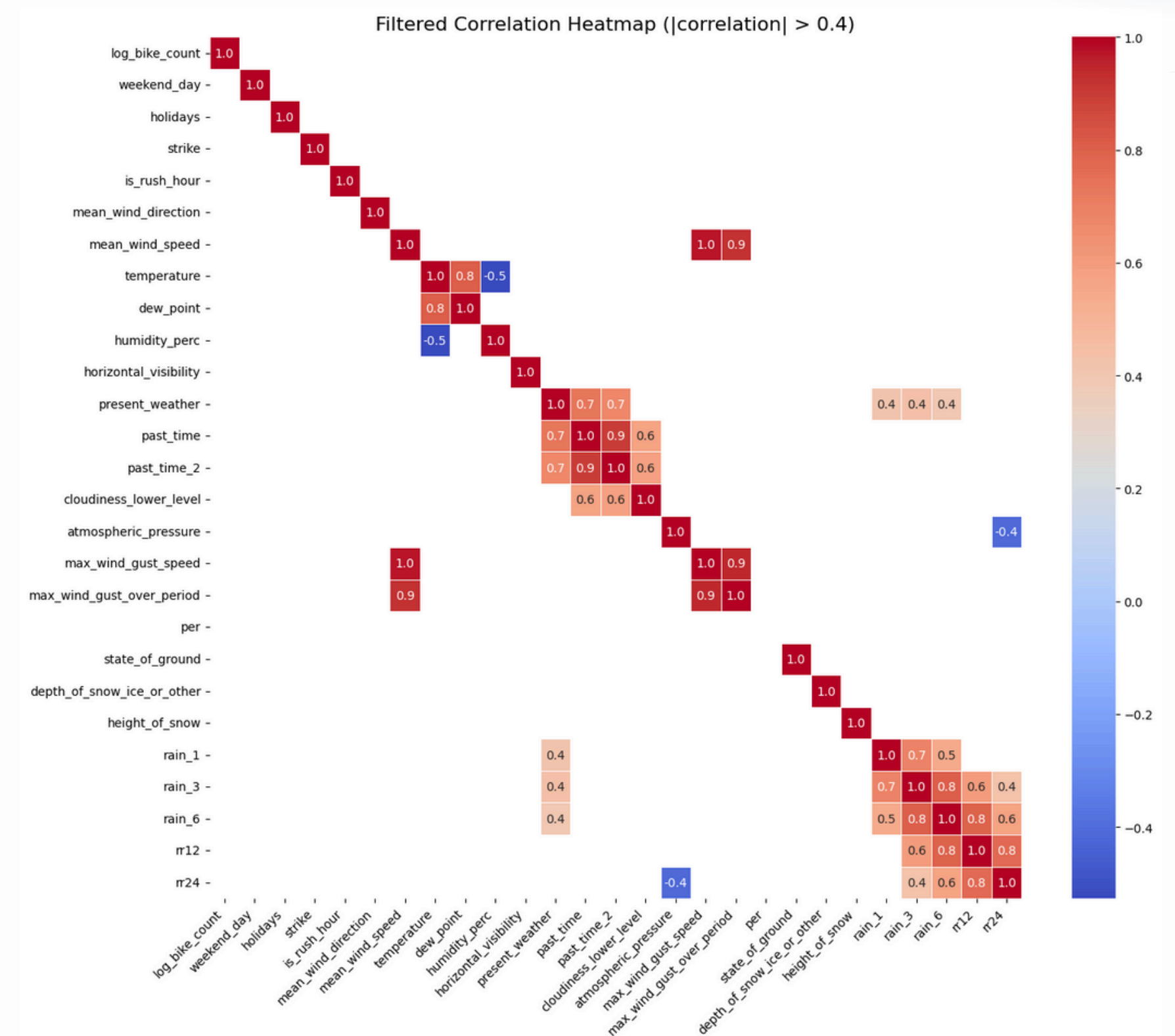
Submission and Description		Private Score ⓘ	Public Score ⓘ	Selected
	submission.csv Complete (after deadline) · Philip Bunford · 5d ago	0.6386	0.6505	<input type="checkbox"/>
	submission.csv Complete · Philip Bunford · 10d ago · .	2.1442	2.1463	<input type="checkbox"/>

Without the external Data

With the external Data

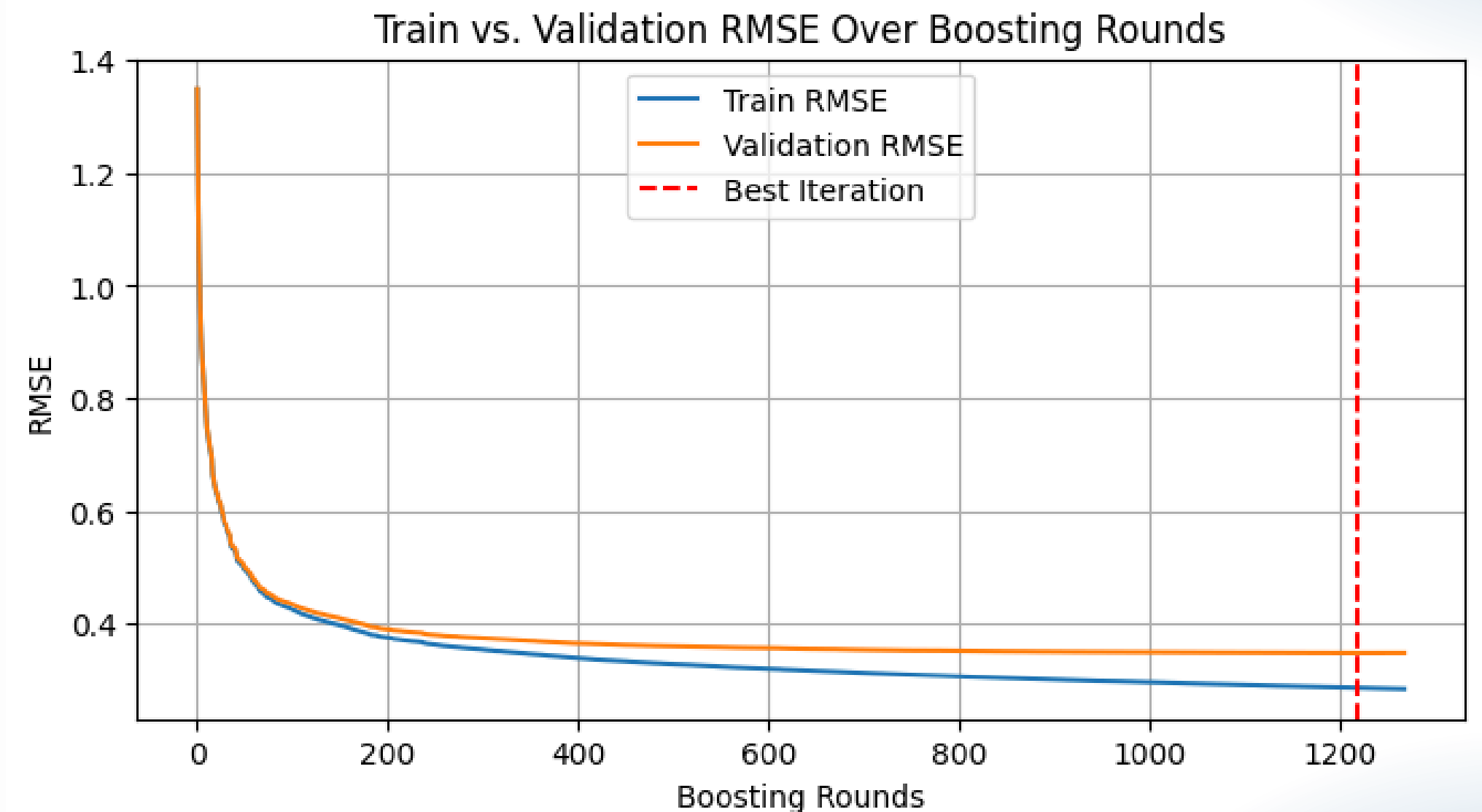
Correlation heatmap (merged data)

- Pairs of **numerical features** with **correlations above** the chosen **threshold** (e.g., >0.4) were identified.
- One feature was removed from each **highly correlated pair** to reduce redundancy.
- This process was repeated until **no pairs exceeded the threshold**.
- The final feature set became **more concise** and **informative** for further analysis.



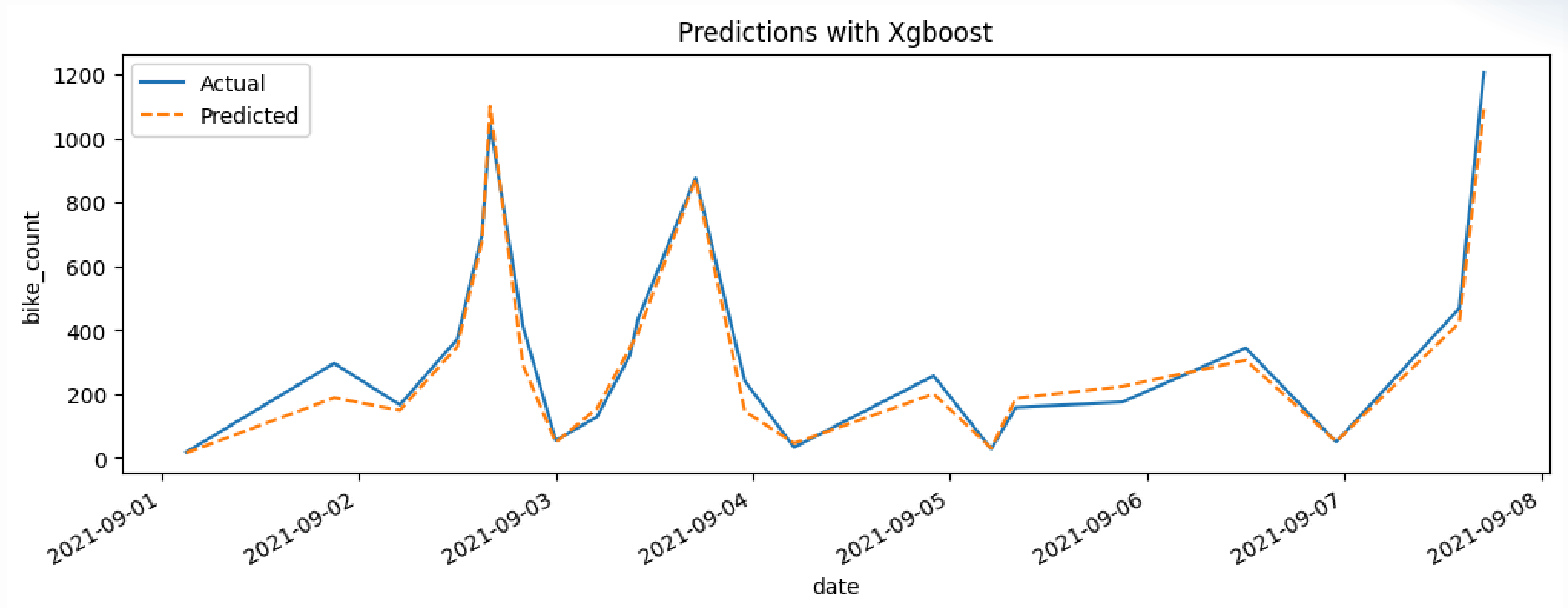
Model selection – early stopping

- Early stopping monitors the **validation** metric each round.
- **If it stops improving** after a set number of rounds, **training halts**.
- The red line at round **1217** marks the **best iteration** before overfitting.
- This ensures the model is saved at its **most generalizable** state.

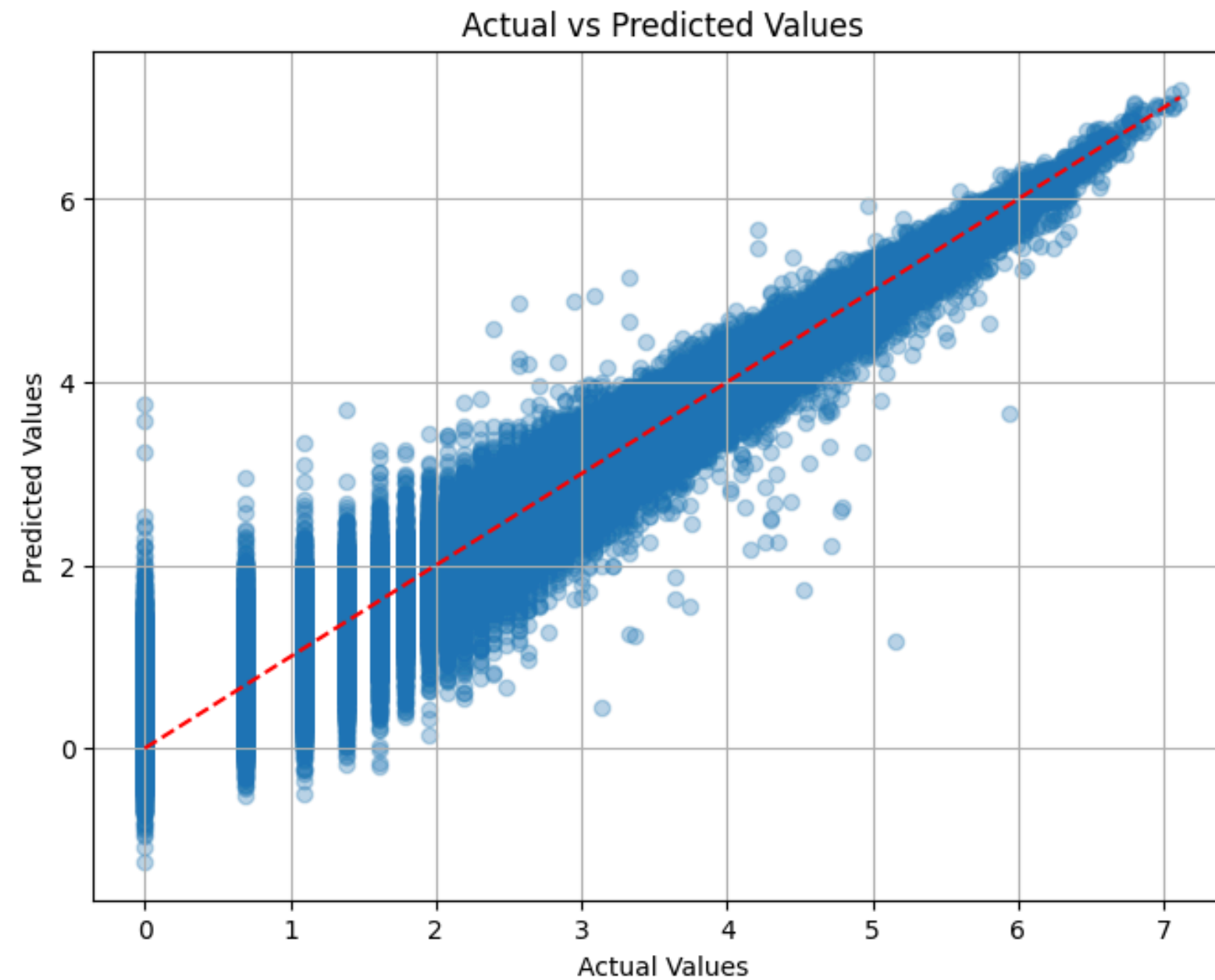


Model selection – predictions

- Both lines follow a **similar pattern**, indicating the model's ability to **approximate the observed trends**.
- Peaks and troughs generally align, suggesting that **the model captures major variations in the data**.



Evaluation of the final model



20/12/2024

thank you!

Philip Bunford | Alessandro Giuliani