Kaggle Project Oral Presentation

Predicting Cyclist Traffic in Paris

Raphael Guenoun & Roy Bensimon

SOMMAIRE

I. PROJECT PRESENTATION — p. 3

II. EDA AND EXPLORATORY DATA ANALYSIS — p. 4-8

III. MODEL SELECTION — p. 9

IV. EVALUATION OF THE FINAL MODEL — p. 10

V. FUTURE RESEARCH — p. 11

VI. CONCLUSION — p. 12

PROJECT PRESENTATION

- GOAL: Accurately predict hourly cyclist traffic using temporal and external data across 56 counters in Paris.
- Performance Metric: Root Mean Square Error (RMSE), chosen for its ability to measure prediction accuracy by penalizing larger errors more heavily.
- Data Preparation: Handled 496,827 entries with no missing values
- Timeframe: - Training data: September 1, 2020 - September 9, 2021

  - Public Test data: September 10, 2021 - October 18, 2021

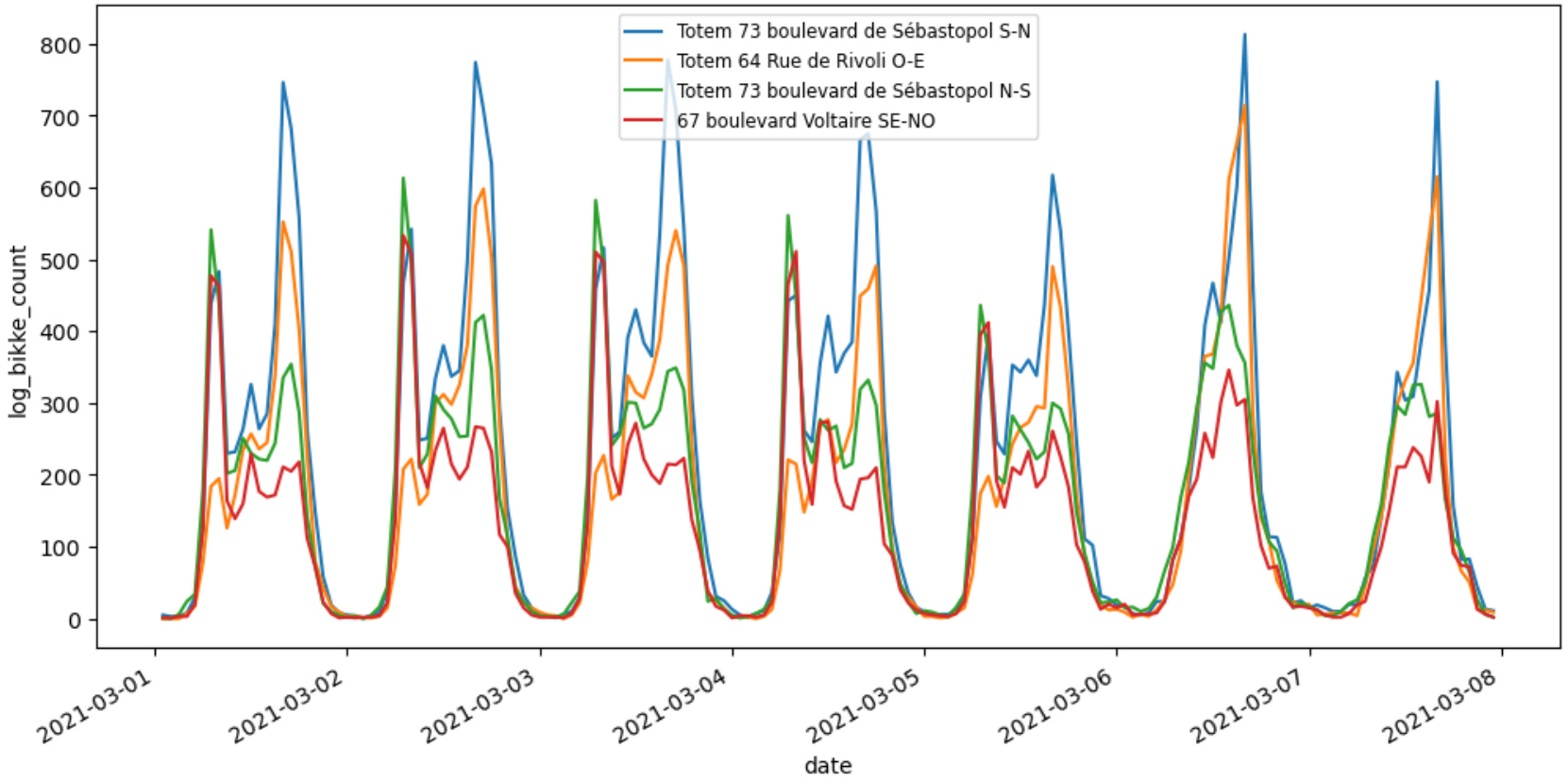
ENCODING OF THE DATA

One-Hot	Cyclical Encoding	Standard Scaler	Target Encoder	Interaction Column
Counter Name Years Weekday	Hours Days Months	All Numerical Columns	Counter Name	Hour x Month

- **One-Hot Encoding**: Maintains the unique identity of categorical variables like `counter_name` without implying any order.
- **Cyclical Encoding**: Captures the periodic nature of temporal variables such as `hour` and `month` for better modeling.
- **Standard Scaling**: Ensures all numerical features have a consistent scale, improving model convergence.
- **Target Encoding**: Highlights the average impact of each counter on `log_bike_count`, leveraging its predictive power.

ENCODING OF THE DATA

The **cyclical** character of days and hours



External Dataset

Handling Numerical Features

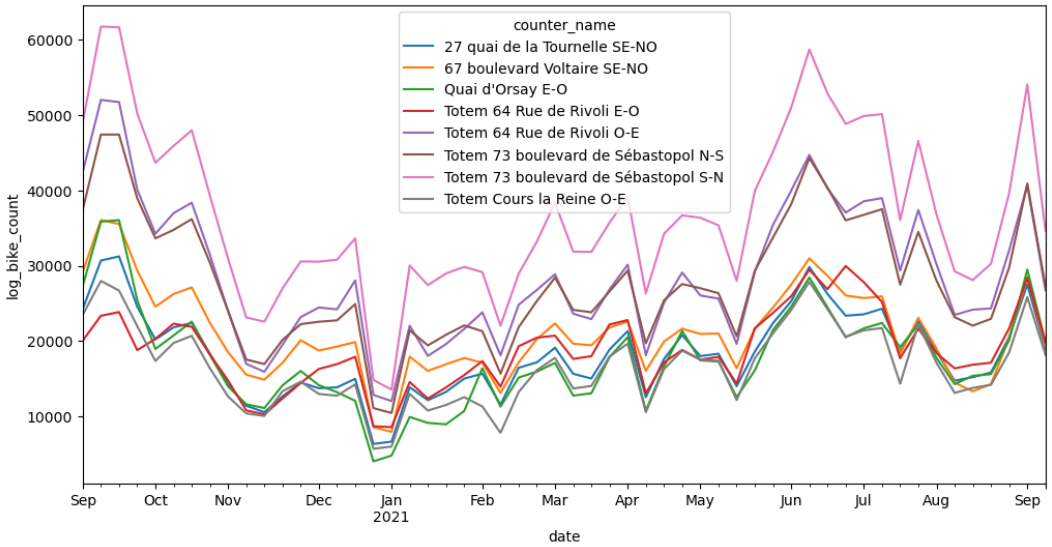
Addressed missing or irrelevant values by:

- Eliminating features with over **10% missing data**.
- Filling remaining gaps using the **median** to minimize outlier impact.
- Removing columns containing **single constant values**.

Incorporating Boolean Indicators

Added key event-related features to capture patterns:

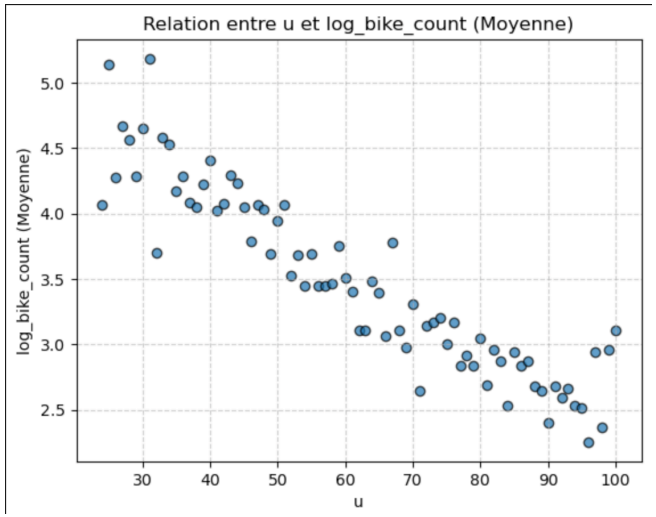
- *Holiday*
- *Bank Holiday*
- *Quarantine periods*
- *Christmas Holidays*



Bike count trends showing a dip during the Christmas period.

Exploratory Data Analysis

Plot showing the relationship between features and bike count



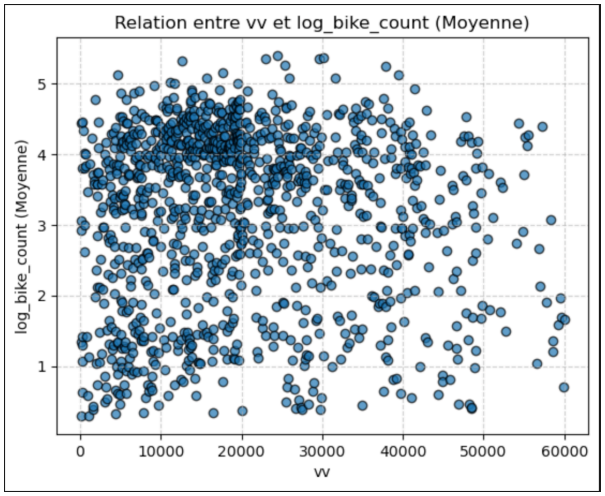
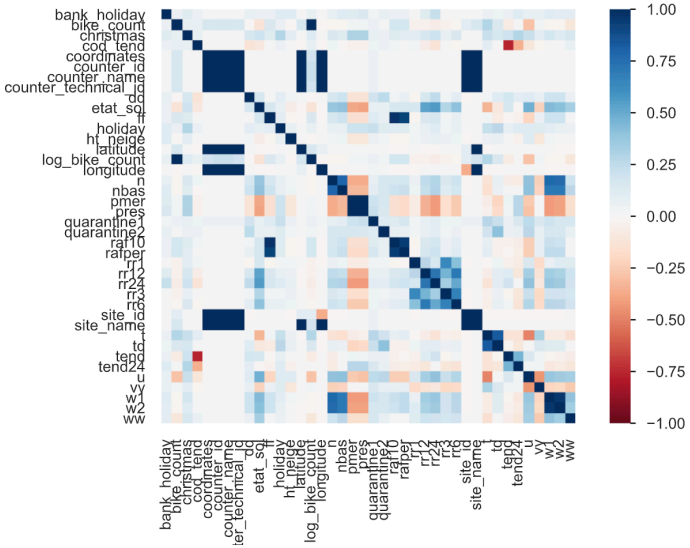


Chart displaying the correlation among variables,



This analysis establishes a baseline understanding of the relationships between variables, enabling the identification of key predictors to be included in the final model.

MODEL SELECTION

Models Tested:
- Linear Regression
- Random ForesT
- XGBoost (final choice)

Evaluation Metric:
Root Mean Square Error (RMSE)

XGBoost Hyperparameter Tuning:
- `learning_rate`
- `max_depth`
- `n_estimators`
- `subsample` using GridSearchCV.

Final Model:
XGBoost: Selected for its superior performance and ability to capture non-linear relationships effectively.

MODELS	RMSE BEFORE TUNING	RMSE BEFORE TUNING
LINEAR REGRESSION	1.25	-
RAMDOM FOREST	0.85	0.75
XGBOOST	0.68	0.66

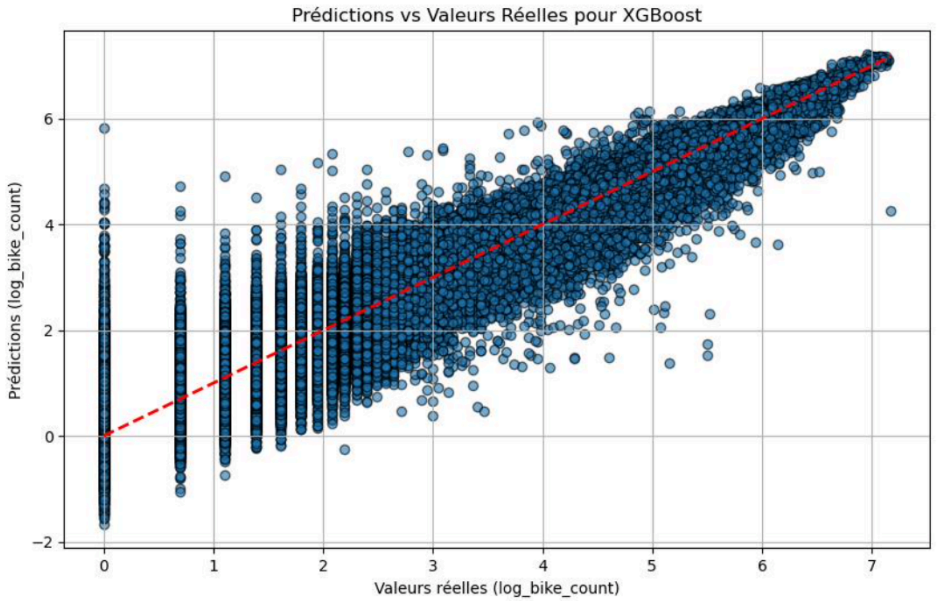
Model Selection

Submission and Description	Private Score	Public Score
submission_Ridge.csv Complete · Raphael Guenoun · 8d ago · Ridge_before_tunning	1.0368	1.0426
submission_RandomForest.csv Complete · Raphael Guenoun · 8d ago · random_forest_before_tunning	0.7676	0.7841
submission_XGBoost.csv Complete · Raphael Guenoun · 8d ago · XG_boost_before_tuning	0.6783	0.6942
submission2_XGBoost_Grid.csv Complete · Raphael Guenoun · 8d ago	0.6611	0.6745

EVALUATION OF THE FINAL MODEL

Our final best result on Kaggle:

✓	**submission2_XGBoost_Grid.csv** Complete · Raphael Guenoun · 8d ago	**0.6611**	**0.6745**	☐



Scatter Plot of Actual vs Predicted Log Bike Counts

FUTURE RESEARCH

Investigating outliers with unusually low or high bike counts
Examine causes like road closures, weather anomalies, or data collection errors.

Integrating additional external data sources
Include metrics such as traffic density, air quality, or public transportation usage to enhance model accuracy.

Adopting advanced time-series models
Explore LSTMs and GRUs to better capture temporal dependencies in bike usage trends.

Implementing counter-specific models
Tailor predictions by training individual models for each counter to address location-specific patterns.

Thank you for your attention