Predicting Trip Duration for London Rental Bikes

Team Members : Arnav Sachdeva, Atishay Jain,
Nilay Jaini, Shrinidhi Bhide

Colab link

Why It Matters:

It optimizes travel for tourists and visitors in London, enhances resource allocation and maintenance, informs infrastructure improvements, and enables location-based marketing and promotions.

Stakeholders

A

Bike-sharing companies

Improve resource allocation and maintenance scheduling.

B

Sustainable Transportation

Roles in public transit, biking infrastructure, and urban planning.

C

City planners

Enhance urban infrastructure planning.

D

Businesses

Target riders with location-based marketing.

E

Riders

Ensure safer, more efficient commutes.

Data Description

Data Source & Scope:

- Includes records from 2015 to 2022.
- Excludes incomplete data from 2023.
- Total records analyzed: 83 million.

Sampling Method:

- Calculated average trip durations for routes with the same year, month, day of the week, and start hour.
- Data filtered for the top ten start and end station names based on usage.

Dataset Division for Analysis:

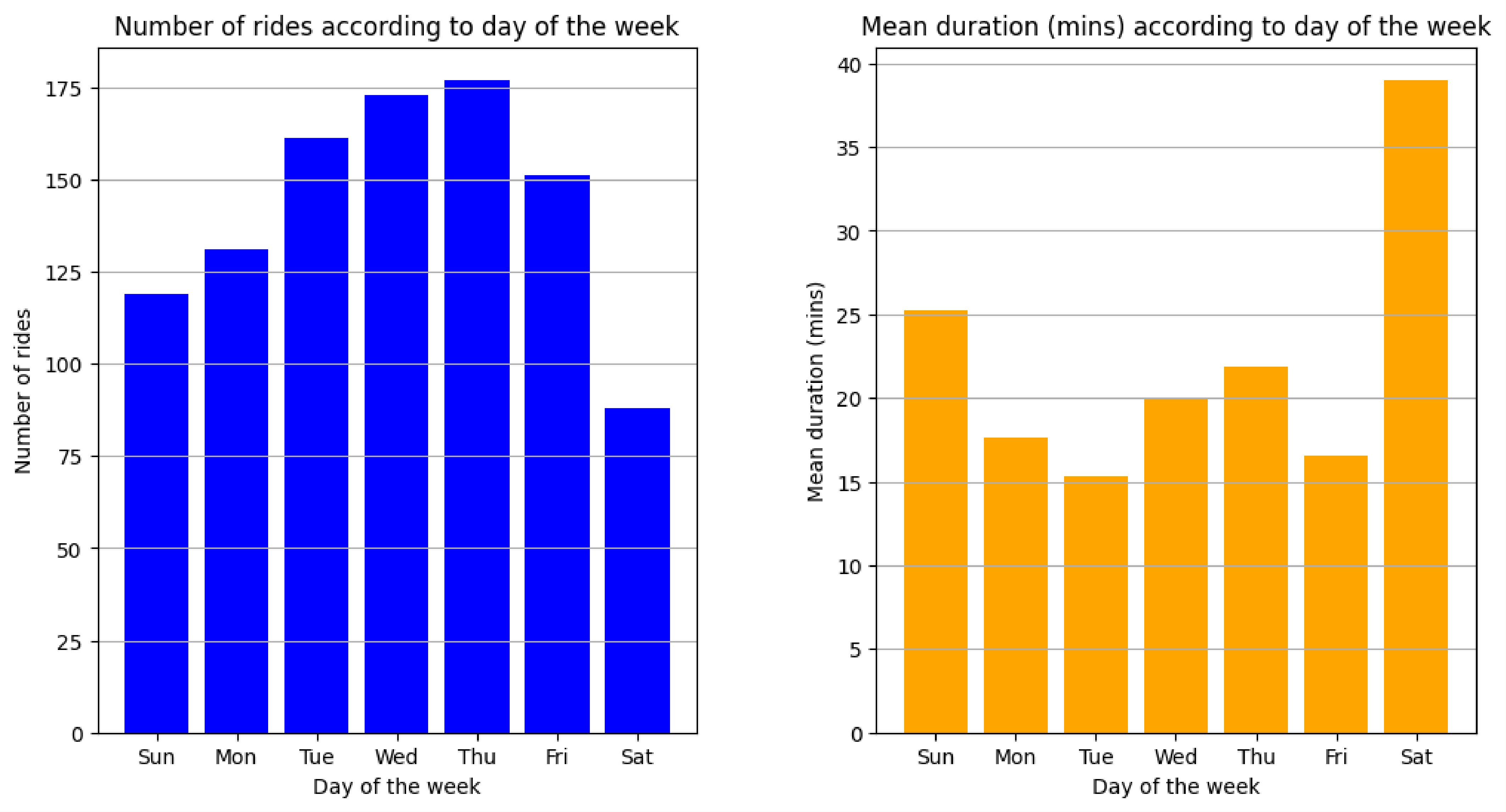
- Training Data: Years 2015 and 2019.
- Testing Data: Year 2022.

cycle_hire QUERY SHARE COPY

SCHEMA DETAILS PREVIEW TABLE EXPLORER PREVIEW

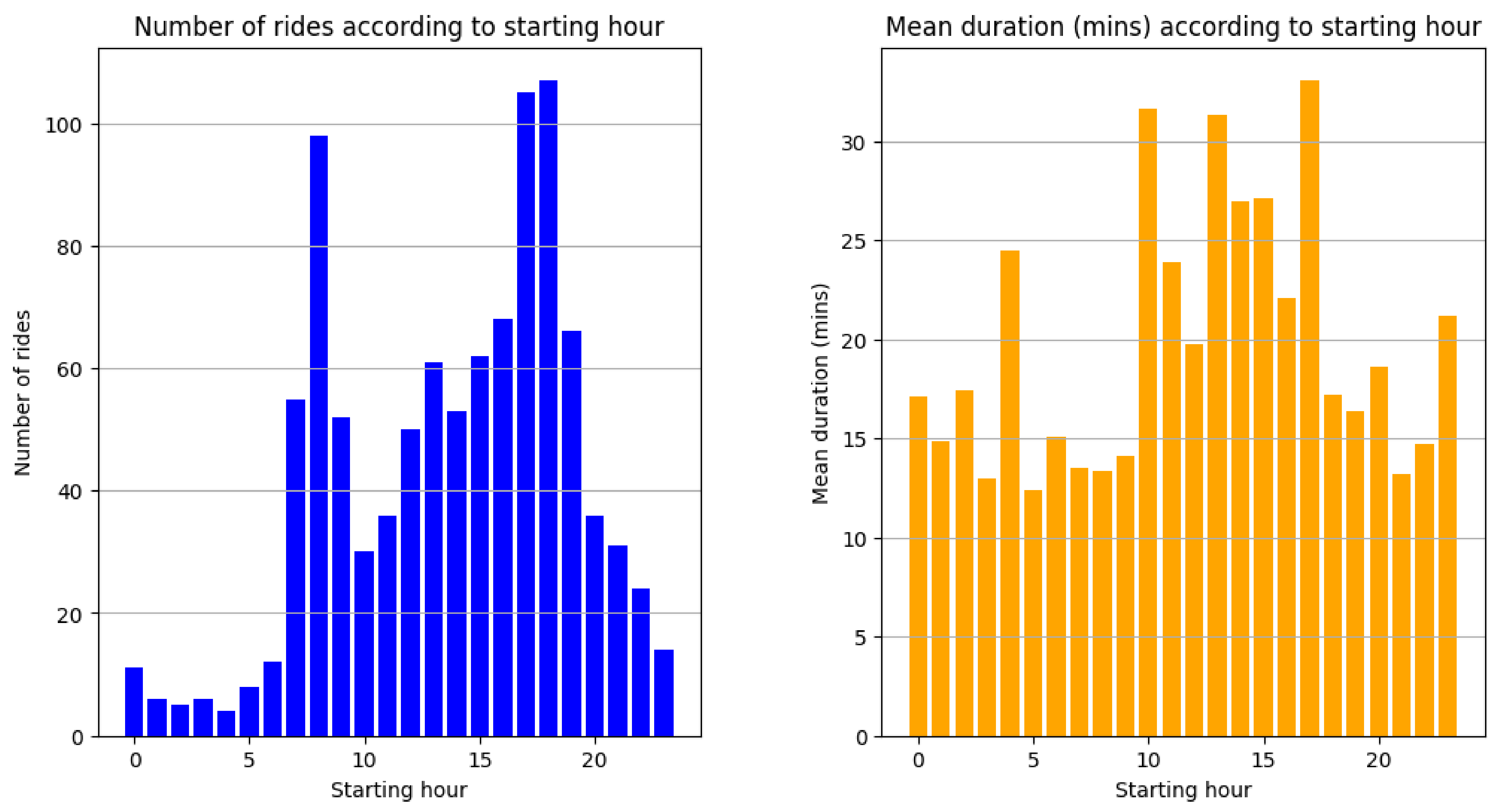
rental_id	INTEGER	REQUIRED
duration	INTEGER	NULLABLE
duration_ms	INTEGER	NULLABLE
bike_id	INTEGER	NULLABLE
bike_model	STRING	NULLABLE
end_date	TIMESTAMP	NULLABLE
end_station_id	INTEGER	NULLABLE
end_station_name	STRING	NULLABLE
start_date	TIMESTAMP	NULLABLE
start_station_id	INTEGER	NULLABLE
start_station_name	STRING	NULLABLE
end_station_logical_terminal	INTEGER	NULLABLE
start_station_logical_terminal	INTEGER	NULLABLE
end_station_priority_id	INTEGER	NULLABLE

Row	year	month	day_of_week	start_hour	start_station_name	end_station_name	avg_duration
1	2015	1	1	0	Waterloo Station 3, Waterloo	Waterloo Station 3, Waterloo	0.0
2	2015	1	1	0	Craven Street, Strand	Craven Street, Strand	1500.0
3	2015	1	1	1	Duke Street Hill, London Bridge	Duke Street Hill, London Bridge	0.0
4	2015	1	1	2	Bethnal Green Road, Shoreditch	Brushfield Street, Liverpool Stre...	180.0
5	2015	1	1	2	Bethnal Green Road, Shoreditch	Bethnal Green Road, Shoreditch	1720.0



Exploratory Data Analysis (EDA)

- Weekday Dominance: Higher number of rides, reflecting commuting trends.
- Weekend Patterns: Longer ride durations, likely due to leisure activities.
- Wednesday Thursday: above average number of trips, likely due to errand running



- Morning and Evening Peaks: Ride demand peaks align with typical office commute times.
- Early Morning Trends: Around 4 AM, ride durations are higher, averaging around 20 minutes.
- Morning Rush Hour: During the morning commute, ride durations decrease.
- Evening Trends: Durations increase again around 5 PM, coinciding with the end of the workday.
- Late Night Surge: There is an unusual rise in ride durations at around 11 PM.

OUR APPROACH

01 Data Preprocessing

- Converted duration from seconds to minutes.
- One-hot encoding for categorical variables.
- Standard Scaling for numeric variables

02 Train/Test Split

- To ensure robust model evaluation we took training data from the years 2015 and 2019.
- The test predictions were evaluated based on data from 2022.

03 Machine Learning Models Explored

- XG Boost (Fastest Best Performer)
- Random Forest (Slow Best Performer)
- SVR, KNN, Linear, Polynomial, Ridge (Underperforming)

Hyperparameter Tuning

Techniques Used: -

- Grid Search: Explores all parameter combinations in a grid
- Random Search: Samples random combinations of parameters.
- Halving Search: Iteratively reduces parameter combinations for efficiency.

Evaluation Metrics:-

- Lowest Root Mean Squared Error (RMSE) on Test dataset
- Negative Root Mean Squared Error inside the Searching CVs

ML Results and performance comparison

y_test

	avg_duration
count	10006.000000
mean	12.663612
std	3.347391
min	7.200000
25%	10.000000
50%	12.500000
75%	15.300000
max	19.000000

dtype: float64

y_pred

	0
count	10006.000000
mean	12.974766
std	2.511582
min	8.704916
25%	10.935959
50%	12.792801
75%	15.326238
max	18.513562

Lowest Test RMSE: 2.47 mins

Achieved by:

Random Forest (Random Search CV)
{'regressor__max_depth': 13, 'regressor__min_samples_leaf': 12, 'regressor__n_estimators': 249}

Random Forest (Halving Search CV)
{'regressor__max_depth': 12, 'regressor__min_samples_leaf': 20, 'regressor__n_estimators': 200}

XGBoost (Random Search CV)
Same result as above but 5 times faster

Challenges & Opportunities

A.

CHALLENGES

1. Dataset complexity (e.g., too many rows).

2. Hyperparameter tuning efficiency.

B.

SOLUTIONS

1. Feature important analysis with Random Forest.

2. Halving Search was used to reduce computation time

FUTURE STEPS

RECOMMENDATIONS

- Computing resource allocation optimization.
- Improved urban planning.
- Personalized customer experiences.

- Incorporate weather data and bike type data (regular or e-bike) for more robust predictions
- Partnership with local businesses for better optimization

CONCLUSION

• Random Forest with Random Search CV achieved best Test RMSE of 2.4694 minutes, marginally better than XGBoost's 2.4780 minutes

• Three models compared:

1. Random Forest with Random Search (7 minutes runtime)
2. XGBoost (1 minute runtime)
3. Random Forest with Halving Search (3 minutes runtime)

• All three models achieved similar RMSE of approximately 2.47 minutes

• Practical interpretation: For a predicted 12-minute trip, actual duration could range between 9.5 to 14.5 minutes

• XGBoost is recommended due to:

1. Fastest computation time
2. Similar performance to other models
3. Better scalability for larger datasets