An Analysis of Cyclist Bike Share Dataset

Term Paper for Business Intelligence II

Group 2 (Option 2)

Redwan Hossin, Fatih Karahan, Fidan Tahirova

02.10.2025

Table of Content

Motivation

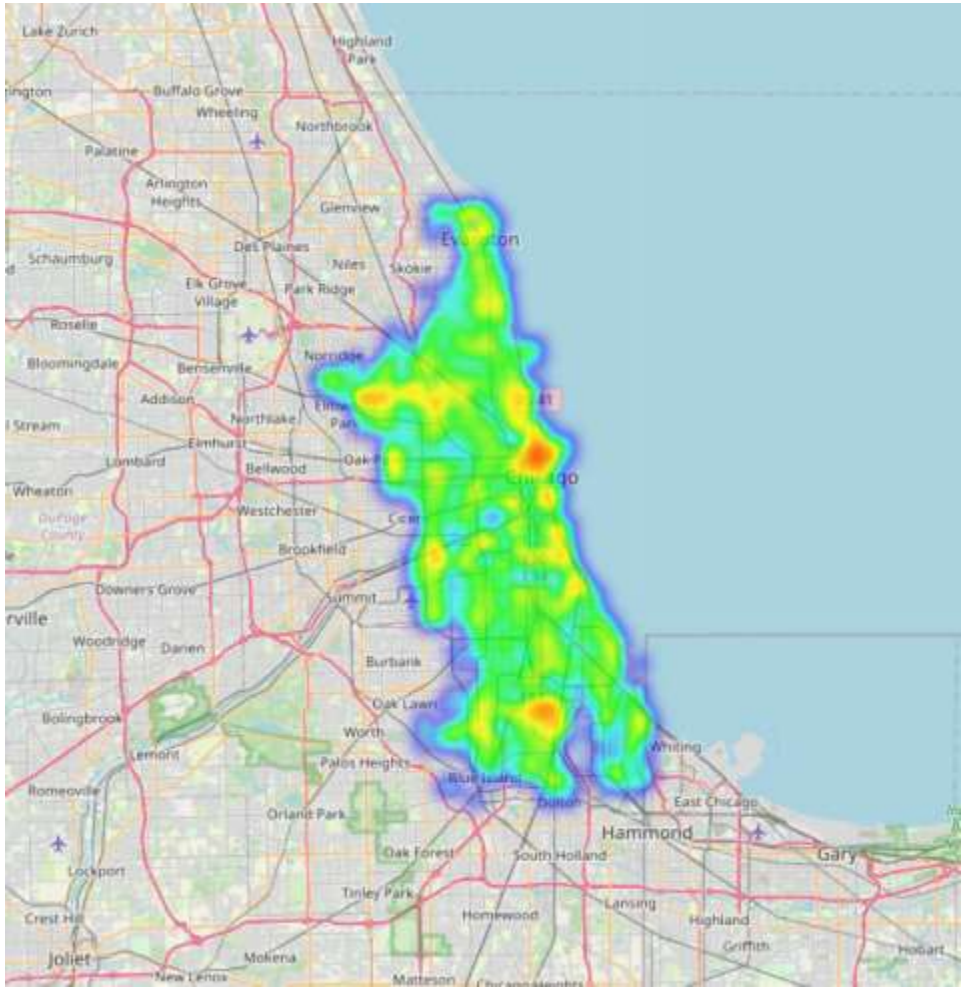
Bike Rider Types

Casual Riders

Member riders

Introduction

Dataset Description



12 months of public trip data from Divvy (Chicago's bike-share system)

5.7 Million individual rides analyzed

Key feature points:

— Ride start/end times and locations

— Bike type (Classic, Electric, Docked)

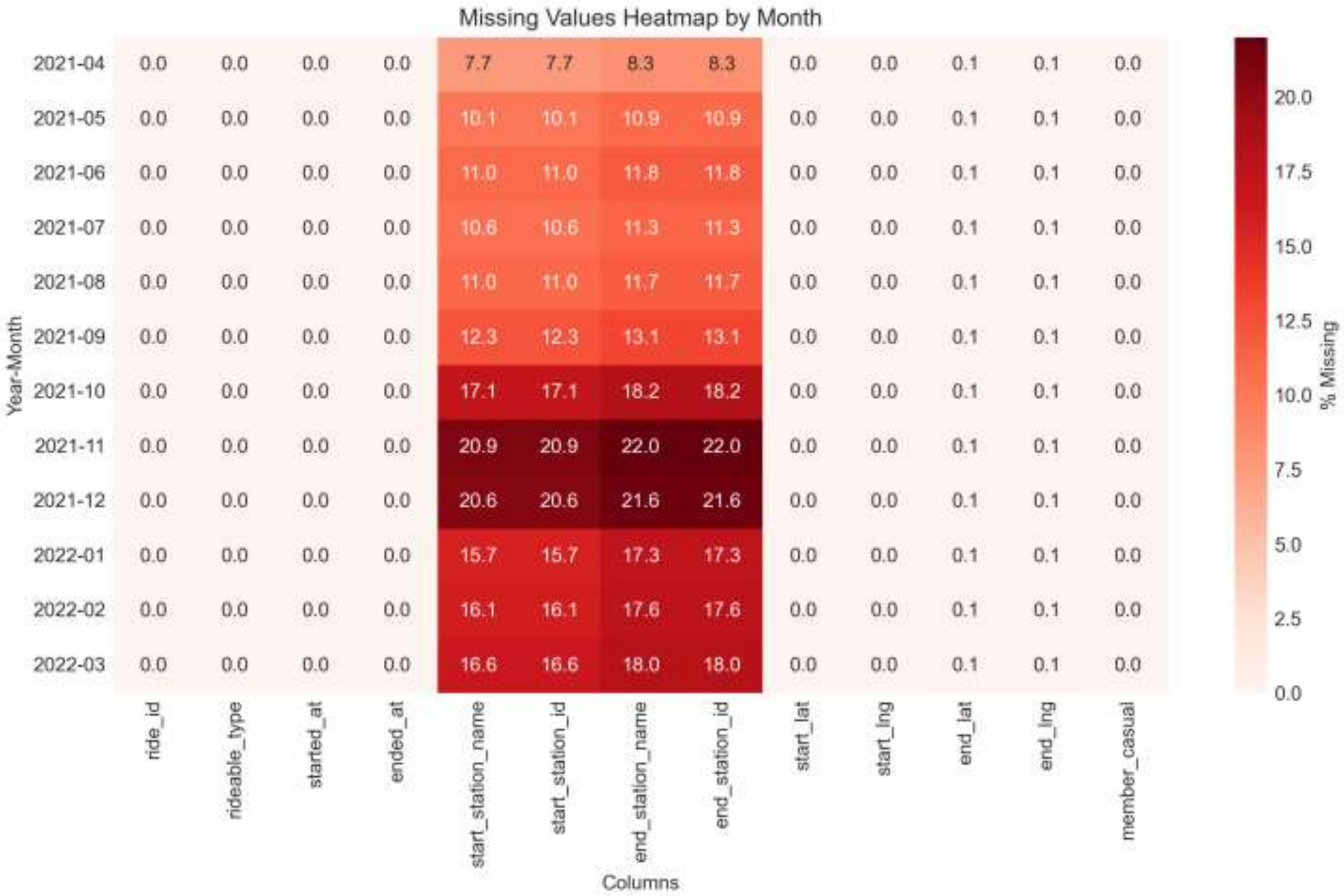
— Rider type (Casual vs. Member)

Data Preprocessing - Feature Engineering

1. Unification and Cleaning
   a) Combined *12 monthly files into one master dataset.*
   b) Handled missing station data by creating an „**Unknown**" category, presenving over *800,000 rides from dockless bikes.*
2. Feature Engineering
   a) Created new metrics like **ride duration, distance,** and **speed**
   b) Added temporal context: **hour, day of week**, and **rush hour flags**

Preprocessing - Missing Values

Summary of Missing Values



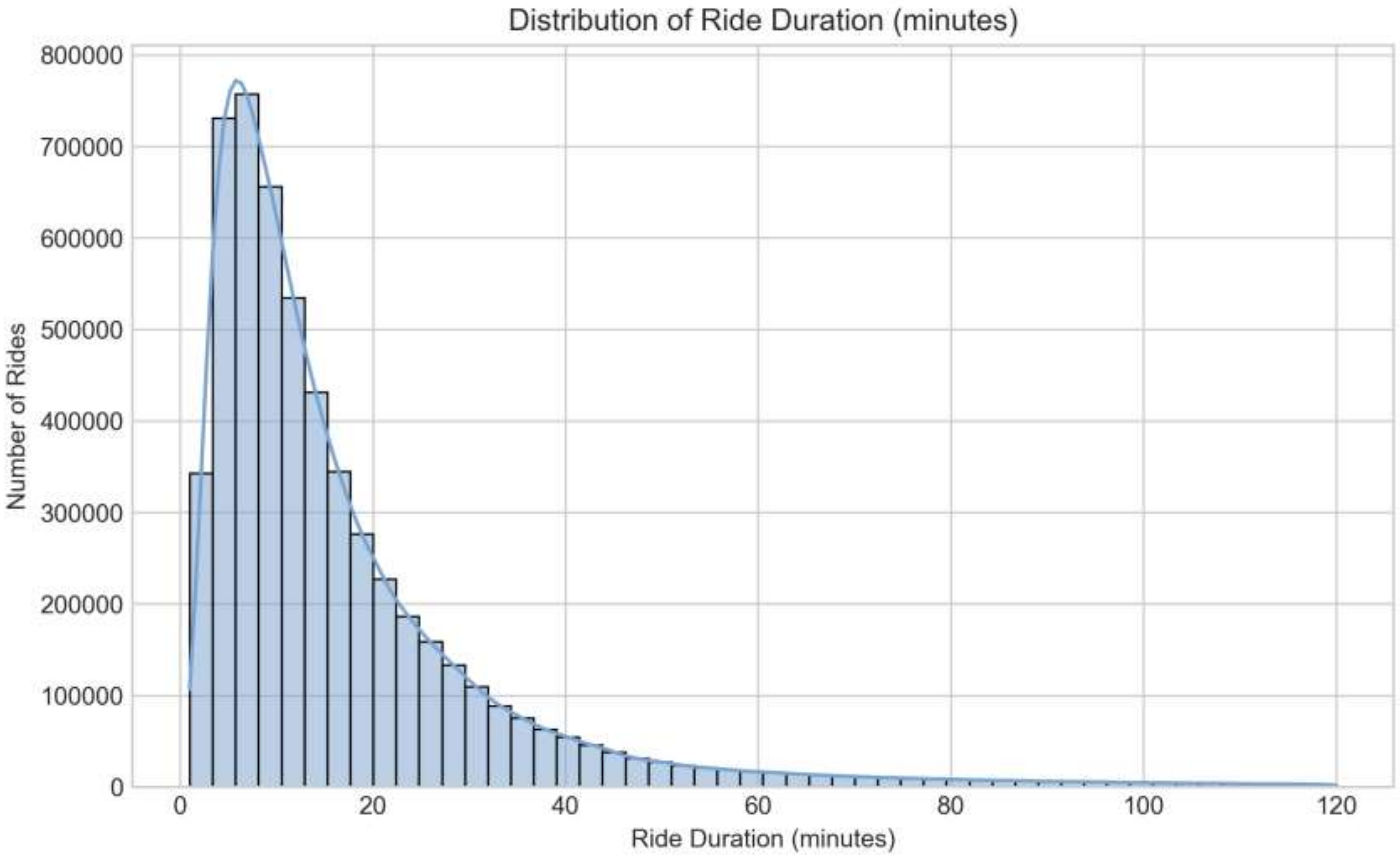
Classic Bikes – must be returned to a docking station.

Docked Bikes – picked up and returned at specific docking stations.

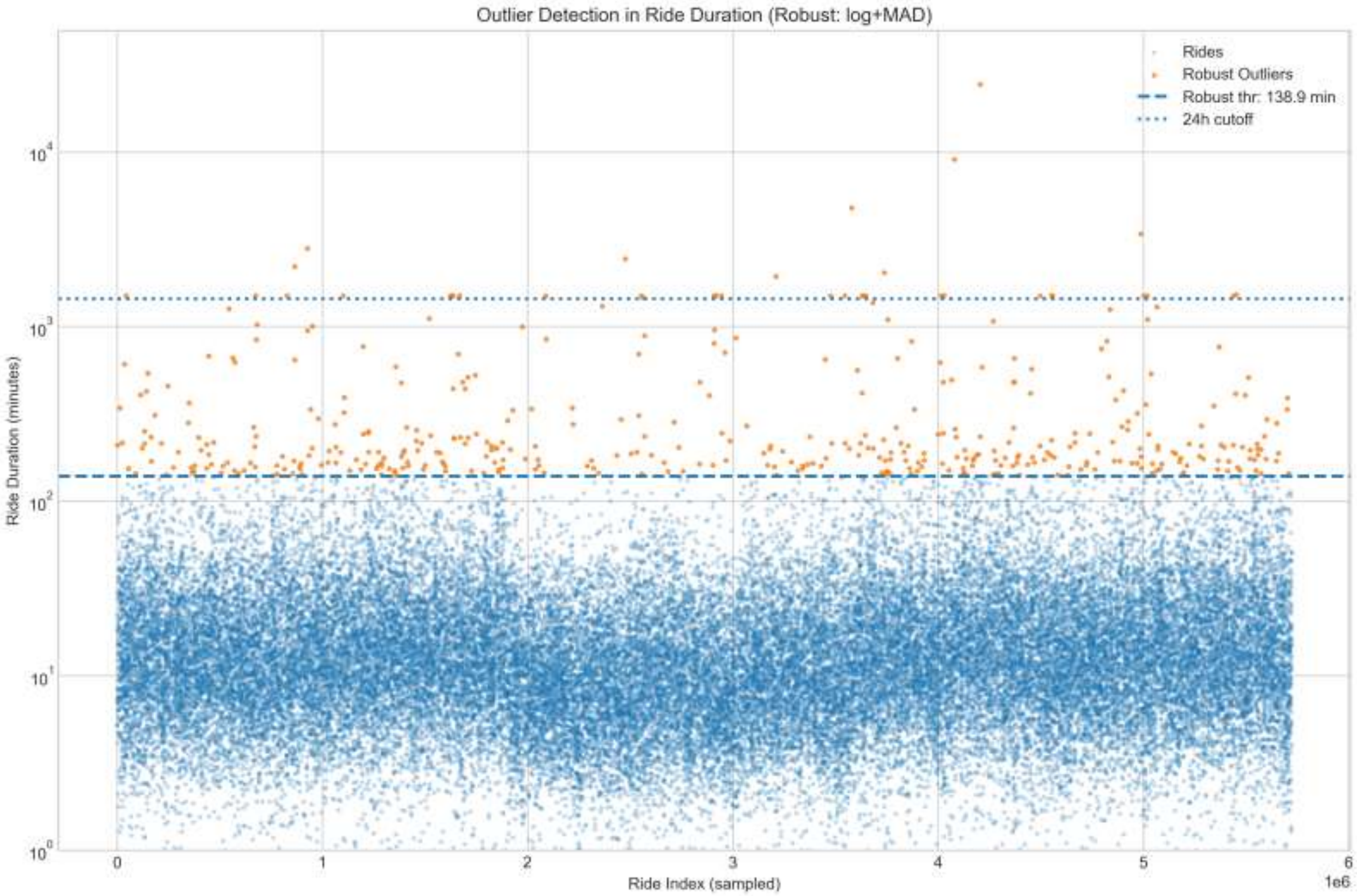
Electric Bikes – can be parked anywhere within the service area.

Preprocessing - Outlier Detection

Data Spread



Preprocessing - Outlier Detection



Preprocessing - Outlier Identification

Noise in the Data

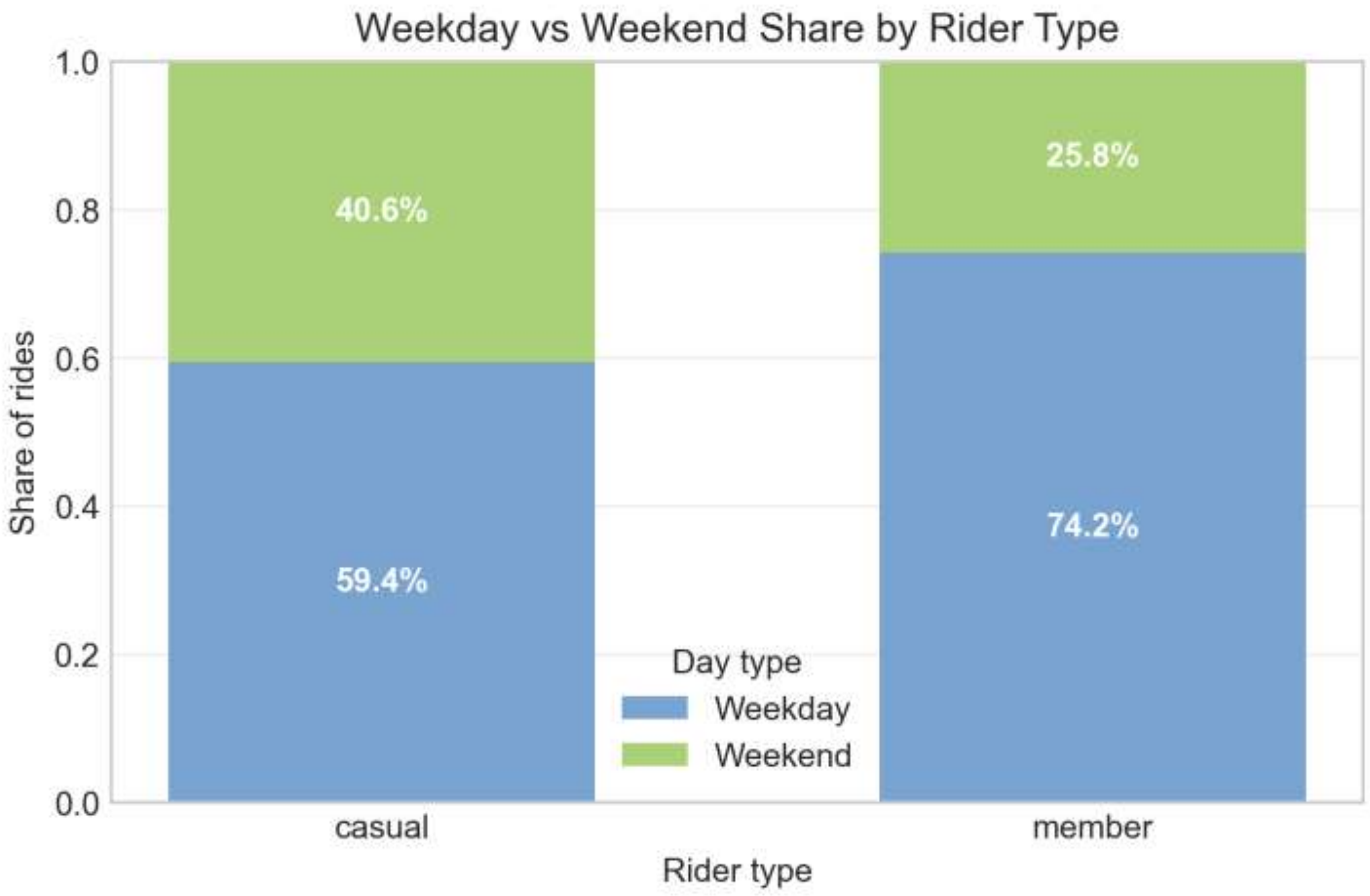
- 145 negative duration rides
- 514 zero-minute rides
- 4,138 rides longer than 24 hours

Impossible data entries

System errors

Unclosed trips

EDA – Weekday vs. Weekend usage

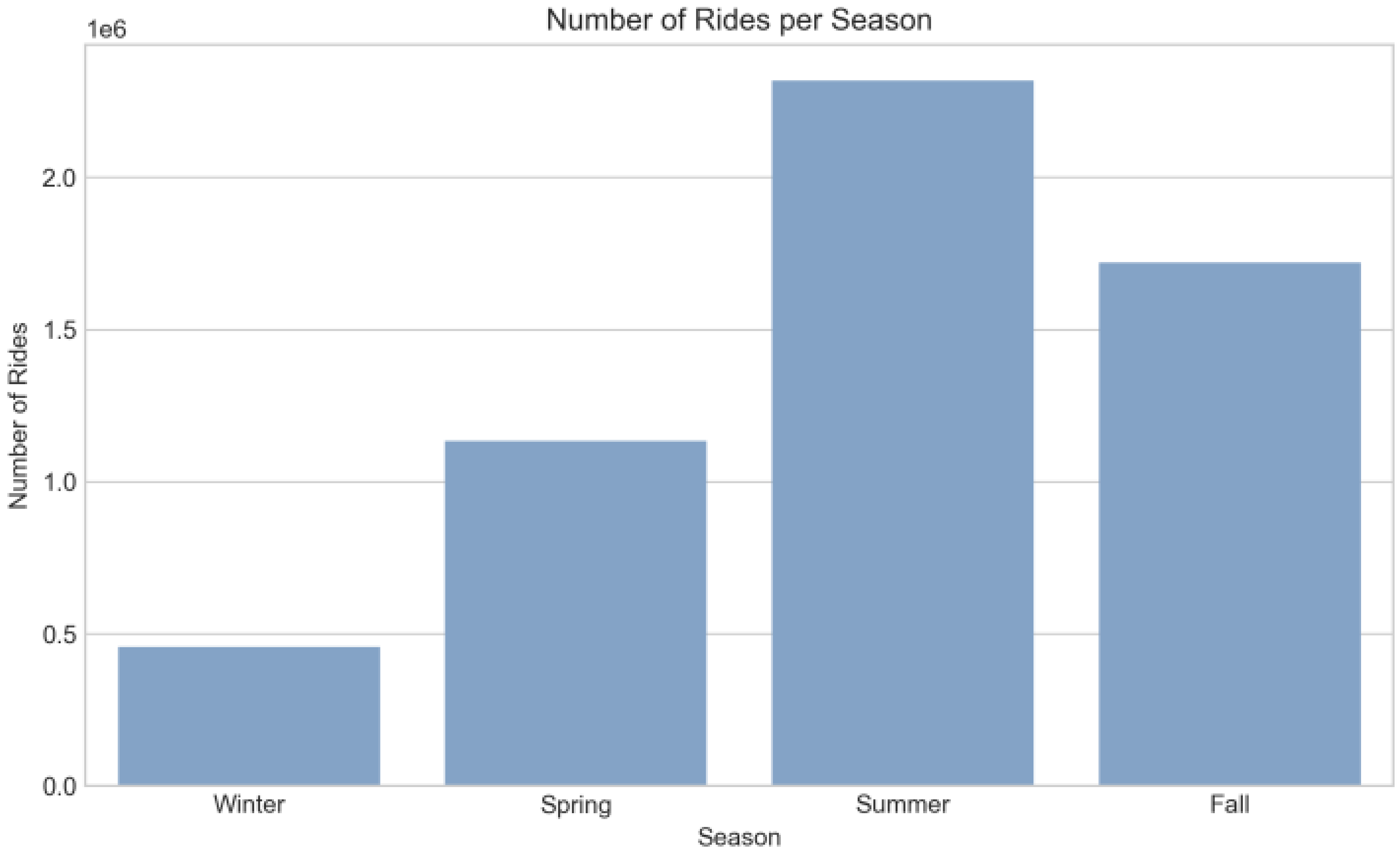


— Casual riders are weekend-focused

— Members are weekday commuters

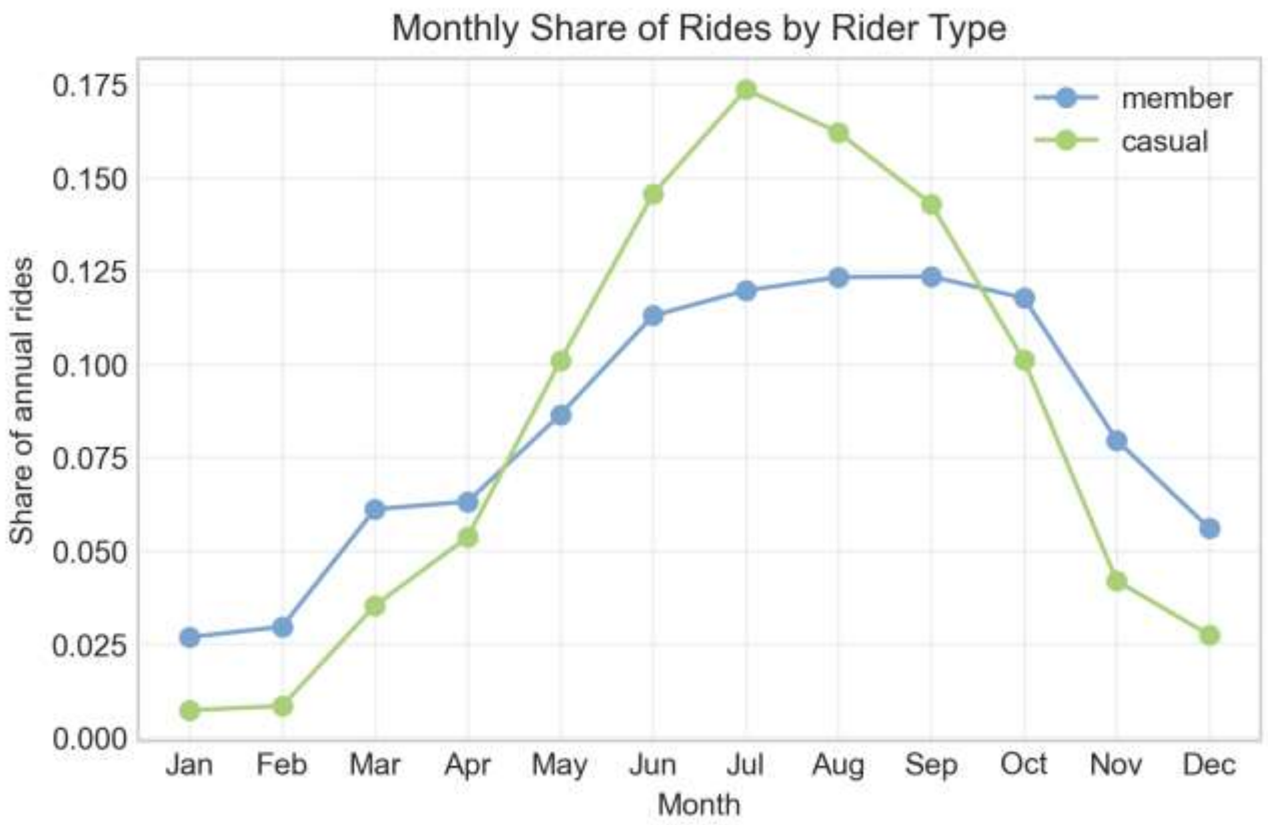
EDA

Impact of Seasonality



- Strong Seasonality

EDA – Impact of Seasonality



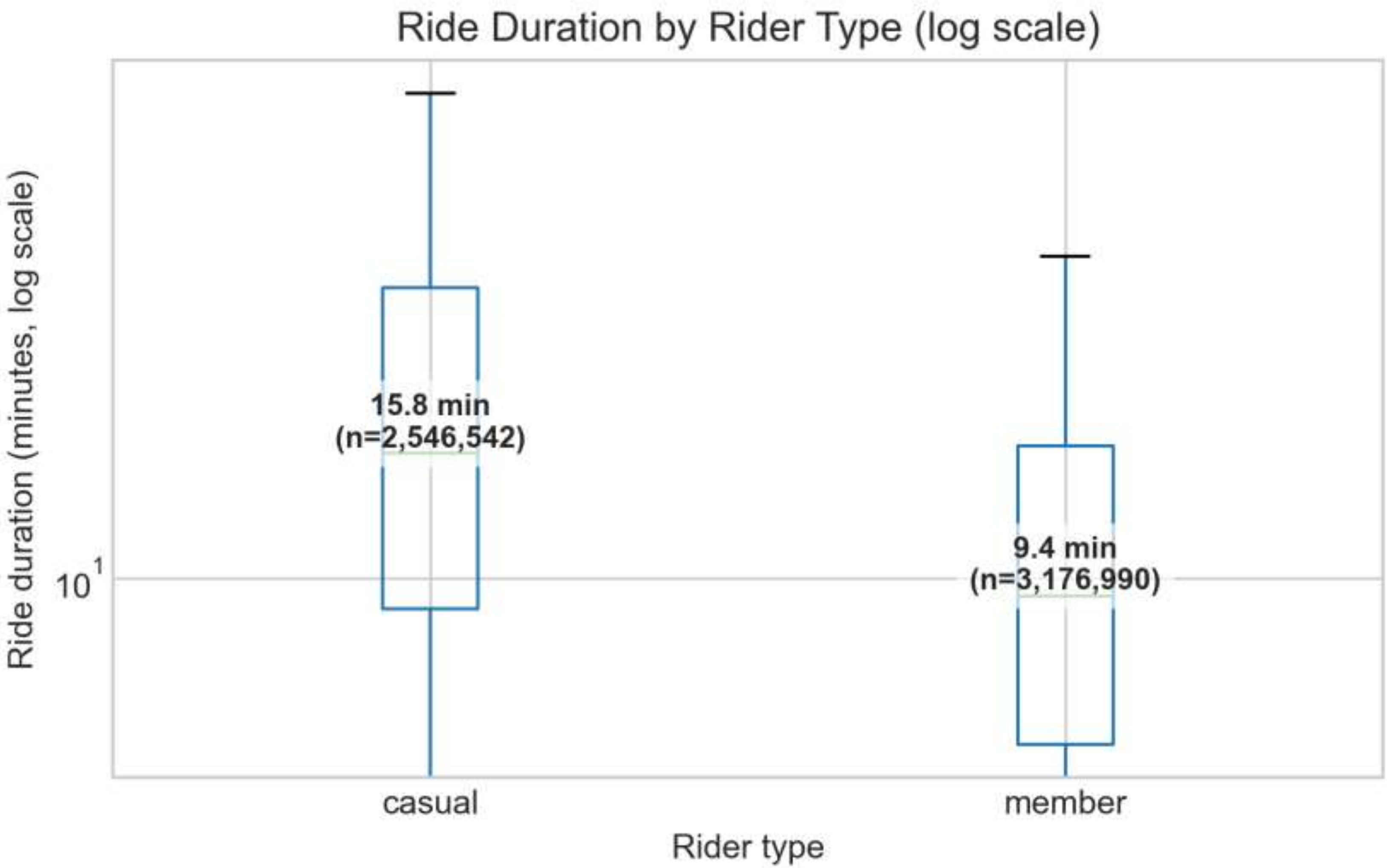
Casual Riders

- Strong summer peak: June–August
- Sharp winter decline: Jan–Feb
- Highly sensitive to weather/season

Annual Members

- Flatter seasonal trend
- Stable usage even in winter
- Less influenced by season

EDA – Casuals Take Longer Rides



External Data Integration

Temperature Data | **Population Density** | **Public Holidays & Events** | **COVID-19 Regulation Data**

Machine Learning - Introduction

Algorithms

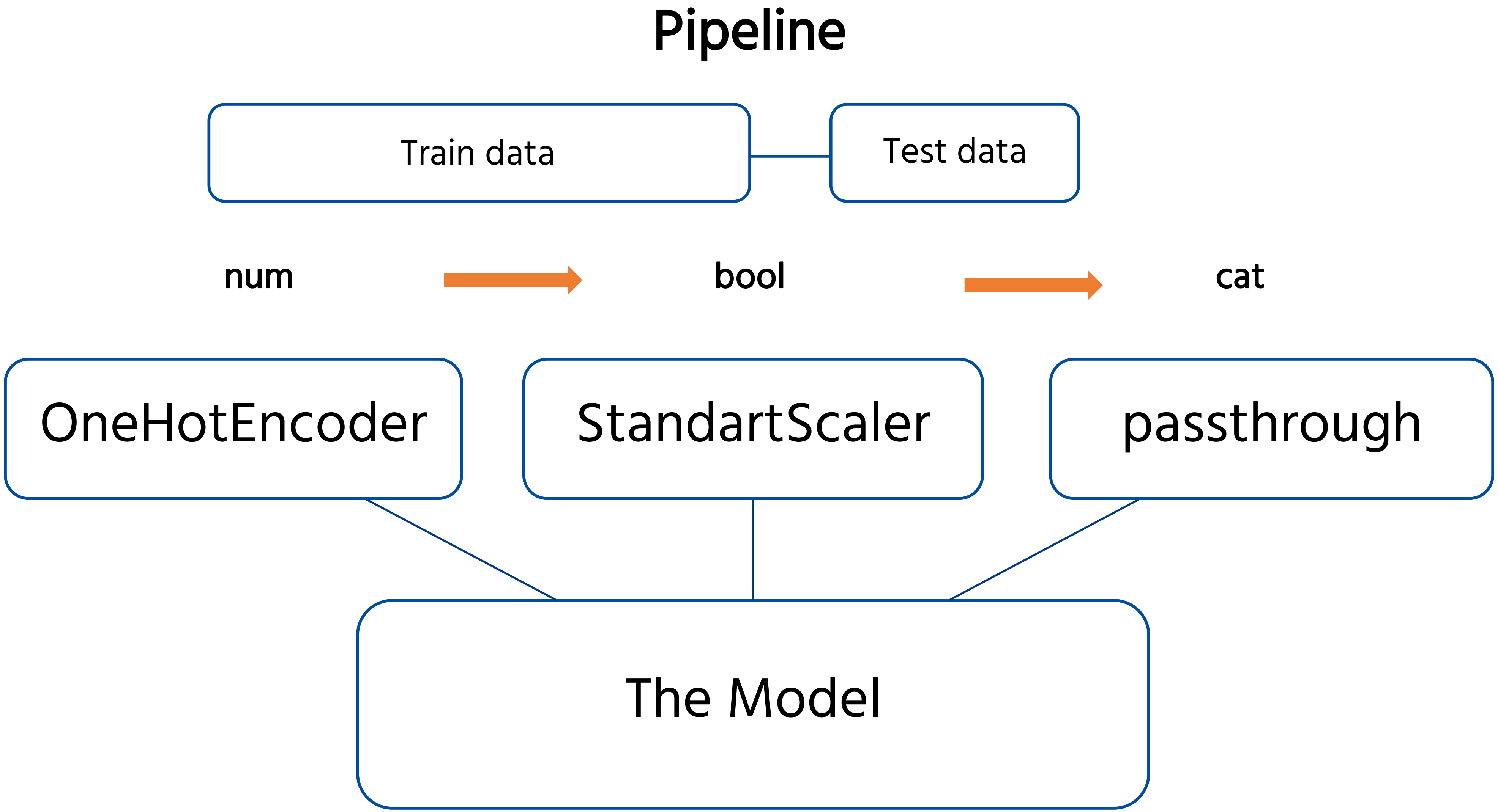
- Logistic Regression – Simple Benchmark
- Random Forest – Strong and Interpretable
- XGBoost – Efficient and High Accuracy

Features Used in Data Modeling

Type	Features
Numeric (scaled)	ride_distance_km, ride_duration_min, speed_kmh, start_hour, start_month, temp_c, precip_mm, wind_kmh, rh_pct, cloud_pct
Boolean (binary)	is_weekend, is_rush_hour
Categorical (one-hot encoded)	rideable_type, season, duration_category, temp_bin, precip_bin, wind_bin

Machine Learning

Pipeline Structure



Machine Learning

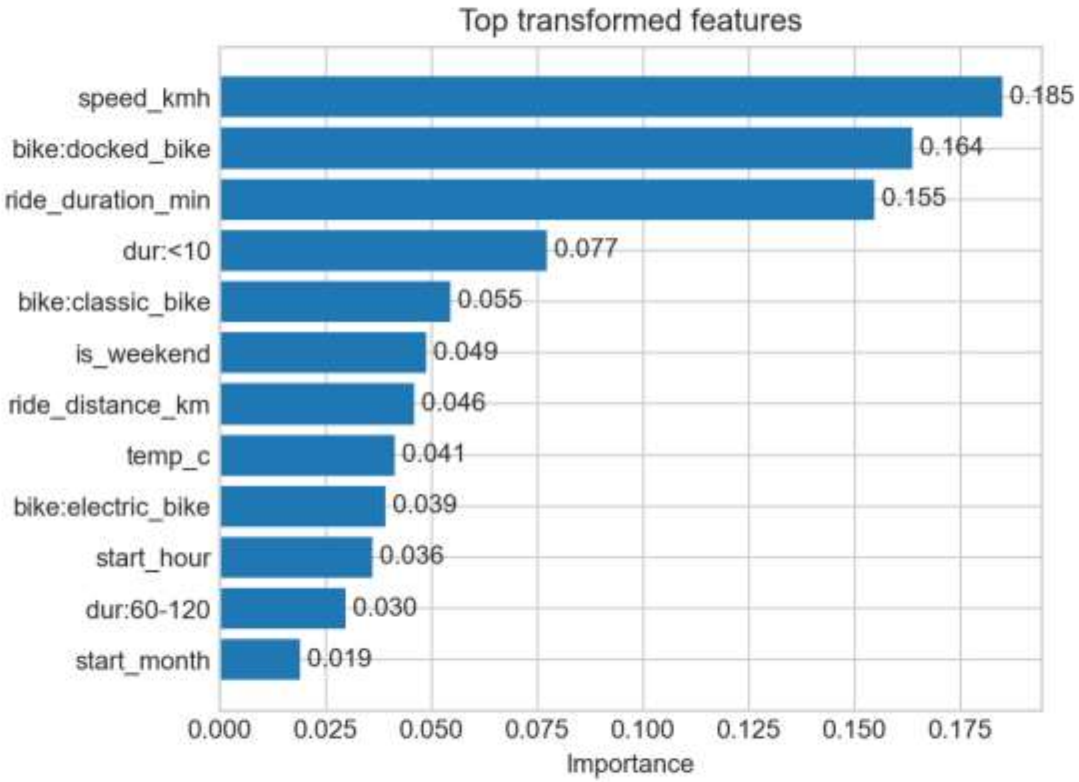
Random Forest Classification Report

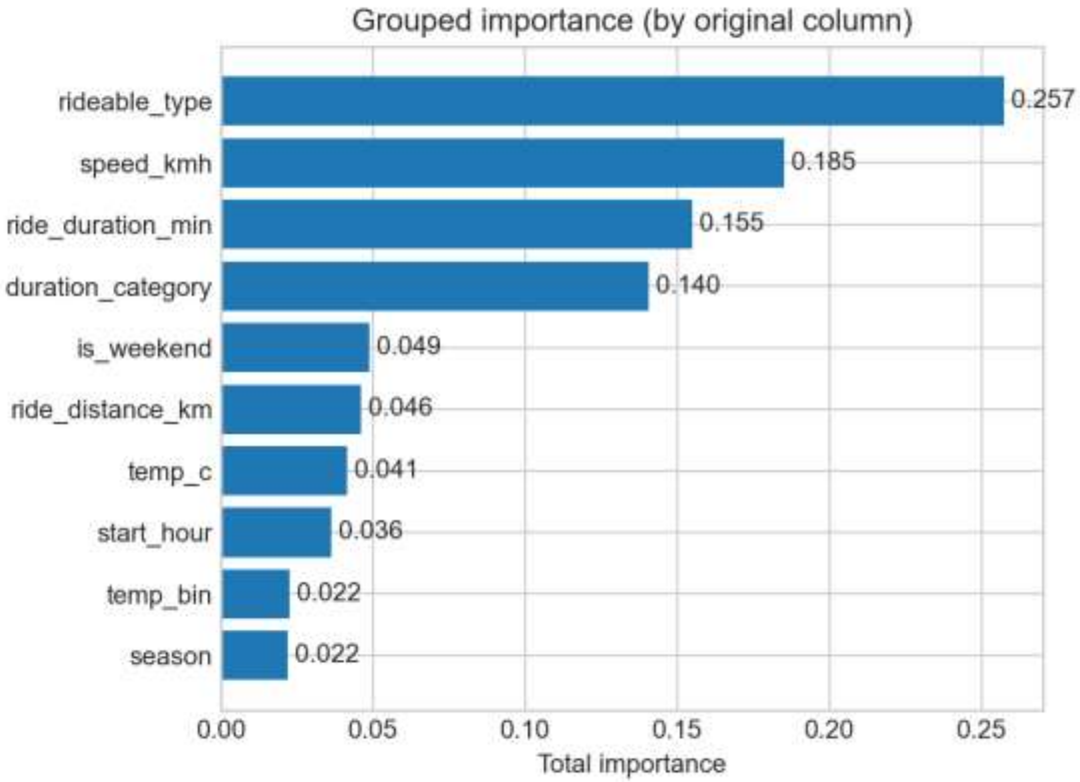
	Predicted Casual	Predicted Member
Actual Casual	37%	63%
Actual Member	7%	93%

	Precision	Recall	F1-score	Num(#)
Casual	0.67	0.37	0.48	109k
Member	0.79	0.93	0.86	282k
Macro avg.	0.73	0.65	0.67	391k
Weighted avg.	0.76	0.77	0.75	391k

Machine Learning – Random Forest

Key Drivers





Model Performance & Feature Importance

Confusion Matrix

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.76	0.71	0.64	0.65
Random Forest	0.77	0.73	0.65	0.67
XGBoost	0.77	0.77	0.72	0.74

Outlook and Future Steps

- Weekdays vs. Weekend ride patterns, action point, positioning of the bikes
- Seasonality
  - Summer - Casual rider campaigns
  - Winter - Member rider campaings
- Docked Bike Stations

References

Source of Data and Algorithms

1. Karahan, F., Hossin, R., & Tahirova, F. (2025). *An Analysis of Cyclist Bike Share Dataset Code*. GitHub. https://github.com/Fatih0234/bike-share-bi2-project

2. Gower, E. (2023). Cyclistic bike-share dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/evangower/cyclistic-bike-share

3. Cox, D. R. (1958). *The regression analysis of binary sequences*. Journal of the Royal Statistical Society: Series B, 20(2), 215–242.

4. Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5–32.

5. Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

6. Open-Meteo. (2025). *Free weather forecast API for non-commercial use*. Retrieved from https://open-meteo.com/

Thank you for your attention!

Are there any questions?