PREDICTING NEXT-DAY DEMAND FOR CAPITAL BIKESHARE STATIONS USING MACHINE LEARNING

Matt Larsen

DCP 4300

Introduction

Capital Bikeshare is an important part of the transportation network in the Washington DC and Arlington Virginia region. It provides a low cost and sustainable travel option that helps people reach transit stations workplaces and local destinations without relying on cars. Because bikeshare stations can become either empty or full very quickly operators must constantly rebalance bikes across the system to keep it functioning smoothly. Predicting where and when demand will occur is therefore critical for both riders and system managers. This project uses historical Capital Bikeshare trip data to build a machine learning model that predicts next day trip demand at the station level. By focusing on short term forecasting using only past usage and calendar information this work explores how data driven methods can improve bikeshare operations and support more efficient transportation planning.

Dataset

This project uses the official Capital Bikeshare trip history dataset for October 2025, which contains 624,869 individual bikeshare trips. Each trip record includes start and end times, station names and IDs, bike type, geographic coordinates, and whether the rider was a member or a casual user. After cleaning the data and removing invalid trip durations, the trips were aggregated into a station day dataset where each row represents the total number of trips originating from a single station on a single date. The final modeling dataset contains 21,892 station day observations across 819 unique stations, covering the period from September 30 to October 31, 2025. Additional calendar and historical usage features were engineered to support next day demand prediction.

=== Summary statistics (numeric) ===

	start_station_id	end_station_id	start_lat	start_lng	end_lat	end_lng
count	543240.000000	538711.000000	624869.000000	624869.000000	624527.000000	624527.000000
mean	31407.163850	31408.513496	38.906640	-77.033362	38.906036	-77.033247
std	334.726827	336.933423	0.027511	0.036127	0.027550	0.036070
min	30200.000000	30200.000000	38.766844	-77.420000	38.710000	-77.450000
25%	31210.000000	31212.000000	38.894851	-77.045640	38.894841	-77.045916
50%	31297.000000	31296.000000	38.905697	-77.031737	38.905578	-77.031686
75%	31615.000000	31616.000000	38.920000	-77.012289	38.918155	-77.012289
max	33200.000000	33200.000000	39.130000	-76.810000	39.130000	-76.810000

Dataset Summary

Input/Output

```
predict_trips_for_station_and_date(
    station_id=31000,
    target_date_str="2025-10-31",
    model_df=model_df,
    feature_cols=feature_cols,
    model=gbr,
    use_baseline=True
```

Input: Station ID & Date

```
Station ID: 31000
Target date: 2025-10-31
ML model prediction (Gradient Boosted Trees): 21.49 trips
Baseline (Station + Weekday Mean) prediction: 22.50 trips
You can compare ML vs baseline for this station/date.
```

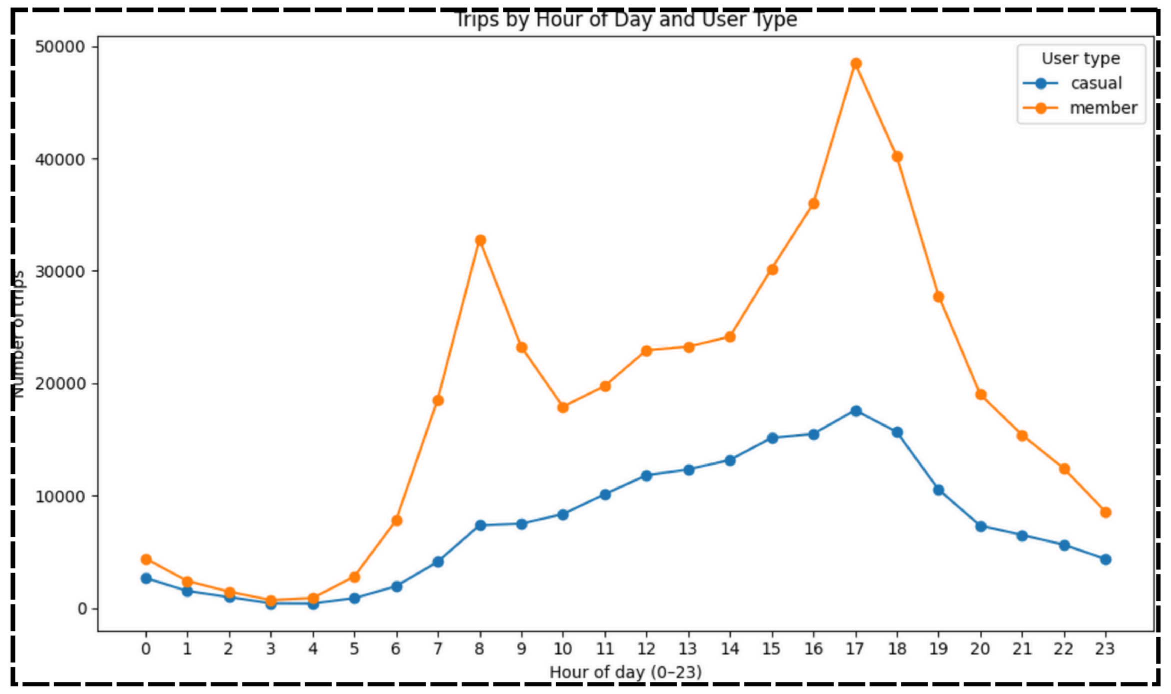
Output: Rideshare prediction

Research Statement

“This project investigates whether short term station level bikeshare demand can be accurately predicted using only historical usage and basic calendar features. Using Capital Bikeshare trip data from October 2025, I develop a supervised machine learning model to forecast next day trip counts for individual stations. The goal is to evaluate how well lightweight data driven models without weather or demographic inputs can support operational rebalancing decisions and improve system reliability in dense urban transportation networks.”

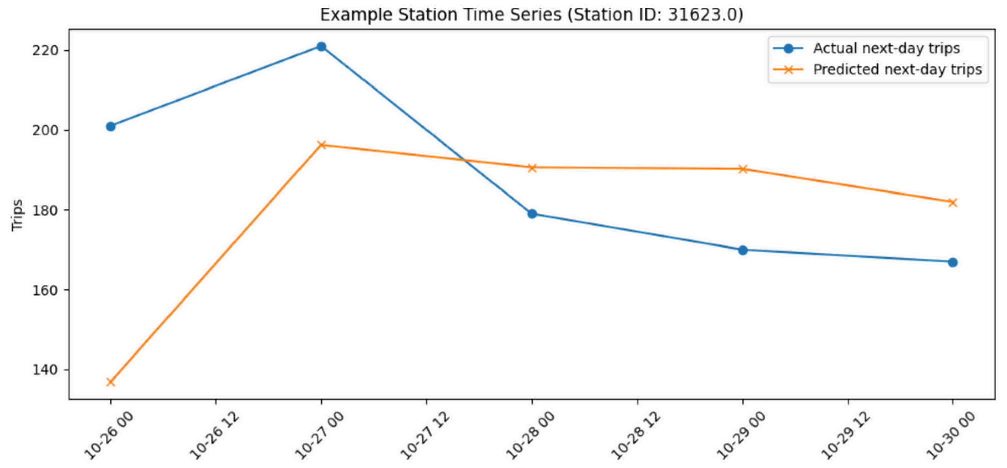
Main Approach

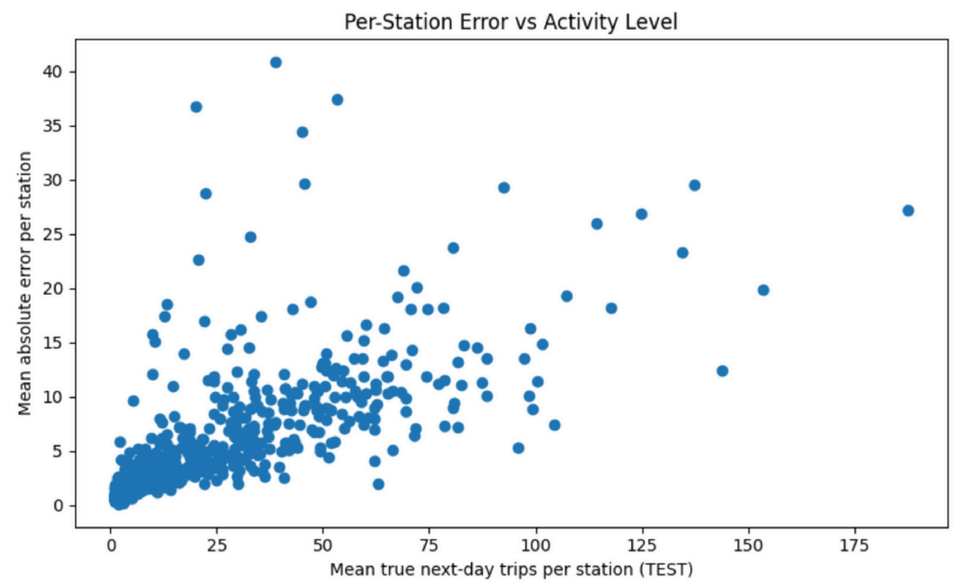
This project frames bikeshare demand prediction as a supervised learning problem at the station day level. Each model input represents a single station on a single day, and the goal is to predict how many trips will originate from that station on the following day. The model uses features derived entirely from historical demand and the calendar, including the previous day's trip count, the same weekday's trip count from the prior week, a seven day rolling average of trips, the station's long term average demand, the numerical day of the week, the month, and a weekend indicator. Station IDs are label encoded to allow the model to learn station specific demand behavior. The primary predictive model is a Gradient Boosted Decision Tree, which was selected because it can capture nonlinear relationships and interactions between recent demand trends and weekly patterns. The model is trained using a time based split to ensure that predictions are always made on future data rather than shuffled observations. Predictions are generated through a custom forecasting function that allows users to request next day demand for any station and date, with an automatic baseline comparison for reliability.



Error Analysis

Model errors were analyzed using residual distributions, station level error comparisons, and predicted versus actual trip plots. Results showed that the model performs best for medium to high demand stations, while low activity stations tend to have higher relative error due to limited historical patterns. Some large errors occurred on days with unusual ridership behavior, likely caused by special events or disruptions that were not included as features. Additionally, early in development, convergence warnings from the Poisson regression model and data indexing warnings in pandas highlighted limitations in model stability and data handling. Overall, most prediction errors were driven by unobserved real world factors and short term demand spikes not captured in the dataset.





Order of Operations

01 **Raw Trip Data**
Capital Bikeshare trip records are collected with timestamps locations bike types and user categories.

02 **Station Day Aggregation**
Individual trips are grouped by station and date to compute total daily trip counts.

03 **Feature Engineering**
Calendar features and historical demand lags are created to capture weekly patterns and recent trends.

04 **Gradient Boosted Model**
A gradient boosted decision tree model is trained to learn nonlinear demand patterns across stations.

05 **Next Day Trip Prediction**
The trained model generates a forecast of how many trips will occur at each station the following day.

Metrics

Model performance was evaluated using three quantitative metrics. Mean Absolute Error measures the average number of trips the model is off by per station per day and is the most interpretable for operations. Root Mean Squared Error penalizes larger prediction errors more heavily, making it useful for identifying models that occasionally produce extreme mistakes. Weighted Absolute Percentage Error normalizes the total error by total demand, allowing comparison across stations with different ridership levels.

Baseline Comparison

Two baseline models were used to evaluate the performance of the proposed machine learning approach. The first baseline predicts demand using each station's historical average for the same day of the week, while the second baseline uses the trip count from the same weekday one week earlier (lag 7). On the test set, the station weekday mean baseline achieved a Mean Absolute Error (MAE) of 5.89 trips, while the lag 7 baseline performed worse with an MAE of 6.74 trips. The Gradient Boosted Trees model outperformed both baselines with a lower MAE of 5.40 trips, demonstrating improved accuracy in capturing short term demand patterns.

=== Results Summary ===

	Model / Split	MAE	RMSE	WAPE
0	Baseline 1: Station+Weekday Mean (VAL)	5.409112	8.910593	0.214260
1	Baseline 1: Station+Weekday Mean (TEST)	5.889018	9.303757	0.252238
2	Baseline 2: Lag-7 (VAL)	6.882556	11.880837	0.272625
3	Baseline 2: Lag-7 (TEST)	6.740986	11.253552	0.288730
4	Linear Regression (VAL)	5.561608	8.769029	0.220301
5	Linear Regression (TEST)	5.821461	8.698834	0.249345
6	Gradient Boosted Trees (VAL)	5.018524	8.566889	0.198789
7	Gradient Boosted Trees (TEST)	5.402173	8.973139	0.231386

Results Summary

References:

Gebhart, K., & Noland, R. B. (2014). The impact of weather conditions on bikeshare trips in Washington, DC. Transportation, 41(6), 1205–1225. https://doi.org/10.1007/s11116-014-9540-7

Noland, R. B., Smart, M. J., & Guo, Z. (2016). Bikeshare trip generation at stations in New York City. Transportation Research Part A: Policy and Practice, 94, 164–181. https://doi.org/10.1016/j.tra.2016.09.020

Rixey, R. A. (2013). Station-level forecasting of bikesharing ridership. Transportation Research Record, 2387, 46–55.