A Time-Based Look at Bike Share Station Demand

DS 340H Capstone Project

Jennifer Ruffin

Introduction

This study analyzes bike share demand at the 5 most frequently used Blue Bike stations across different months and times of day during the time frame of May 2024-August 2024.

Given these time periods and stations, the project aims to predict future departure counts as well as peak usage times during the busy summer months.

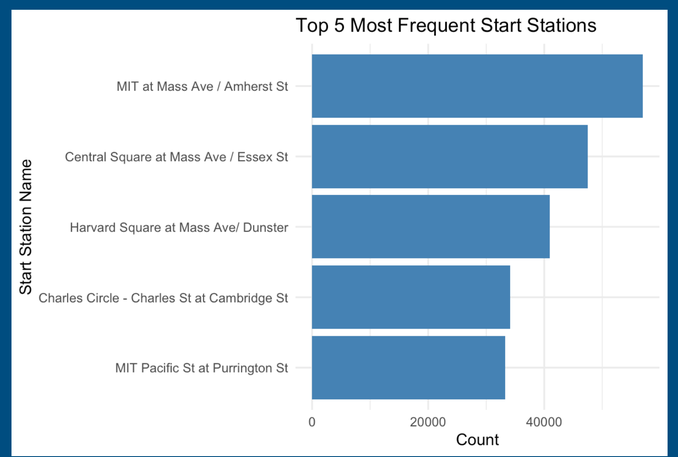

Figure 1: Bar plot showing top 5 most frequently used origination stations

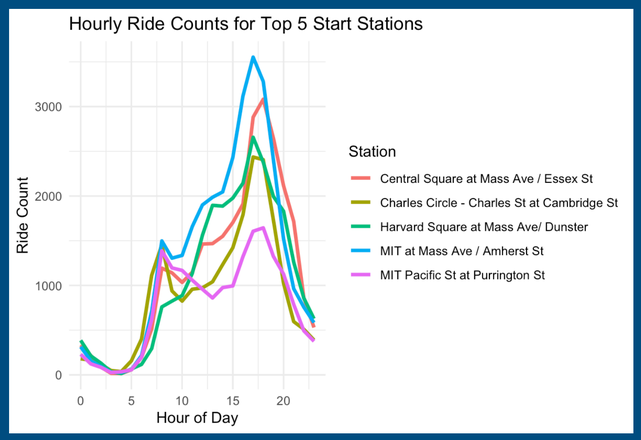

Figure 2: Line plot of the top 5 stations ride count by hour of the day

Research Questions

Q1: How does station-level demand (in terms of trip origin) vary across different months and times of day?
Q2: Can predictions be made for peak usage periods for the most popular origination stations?

Data

Source: Blue Bike Ride Data
Population: April 2023-August 2023
Manipulation: Month and start station variables encoded numerically; variables related to arrival stations removed

Methods

Negative Binomial Regression utilized for count of bike trips and number of departures at a station within a specific time interval
Random Forests used for predicting peak usage times

Q1 Visualizations

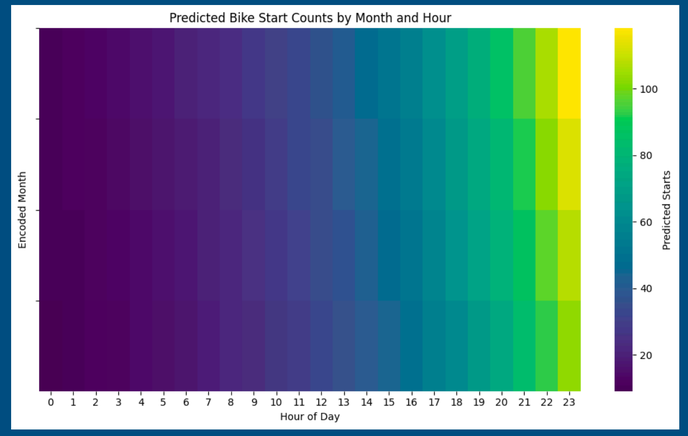

Figure 3: Heat map for predicted bike start counts by encoded hour of the day during encoded month

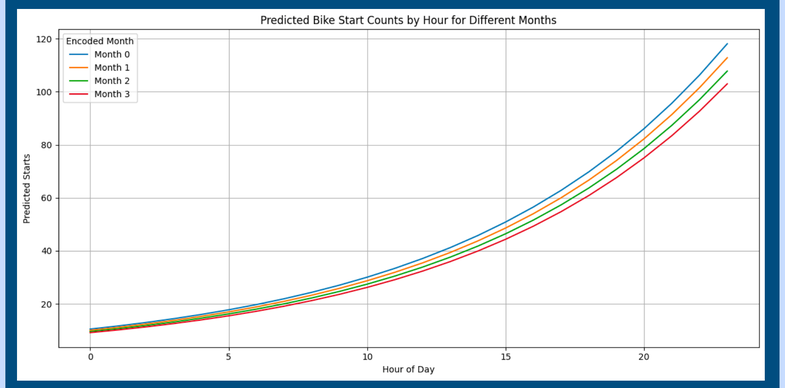

Figure 4: Line graph for predicted bike departures during encoded months by encoded hour of the day

Results

Part A: Negative Binomial

- **Month:** For each unit increase in the encoded month, the rate of bike trip starts decreased by approximately 4.5% ($p < 0.05$).
- **Hour of Day:** For each one-hour increase in the day, the rate of bike trip starts increased by approximately 11.1% ($p < 0.001$).
- **start_station_name_encoded** and **day_of_week_encoded** were not statistically significant predictors of bike start counts in this model ($p > 0.05$)

Part B: Random Forests

Overall accuracy was 93.5% in predicting peak usage.

- Precision was 95%, indicating that when the model predicted non-peak usage, it was correct 95% of the time.
- Recall was 98%, indicating that the model correctly identified 98% of the actual non-peak usage instances.
- Precision was 77%, indicating that when the model predicted peak usage, it was correct 77% of the time.
- Recall was 61%, indicating that the model correctly identified 61% of the actual peak usage instances.
- The F1-score was 0.68.

Conclusion

- The Negative Binomial model showed limitations in accurately predicting start counts across all hours.
- Counterintuitive high predictions during off-peak hours suggest the model didn't fully capture the drop in nighttime demand.
- This may be due to insufficient temporal features or model complexity.
- Future work should explore better time representations (cyclical, interactions) or alternative models.
- Improving prediction accuracy, especially during low demand, is a key consideration.