An Emerging Mode of Transportation: Predictors of Electric Bike Usage in Boston

Nile Rankin-Watson, Wellesley College ‘25, Data Science Major Capstone

Background

As the world becomes more environmentally conscious, citizens have begun to swap individual cars for more eco-friendly options like public transportation, ride-sharing, and, increasingly, bicycling. In 2017, Boston and its surrounding municipalities adopted the regional bike sharing system: Blue Bikes. In 2023, Blue Bikes introduced e-bikes as an alternative to the classical mechanical bike. Since then users have increasingly opted into this option, with 2520 electric trips and 7480 classic bike trips in 2024.

Research Question

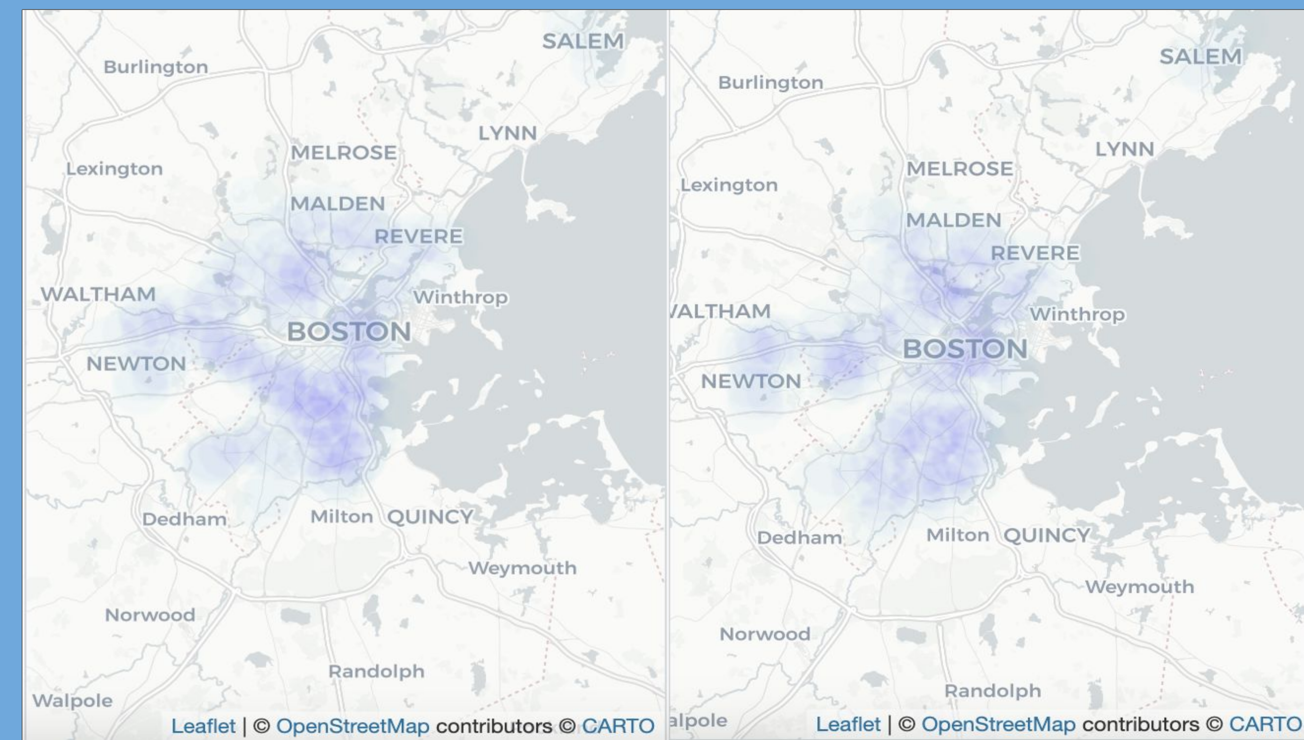
What are the predictors of a Blue Bike rider’s usage of an e-bike rather than a classic bike in 2024?

Data

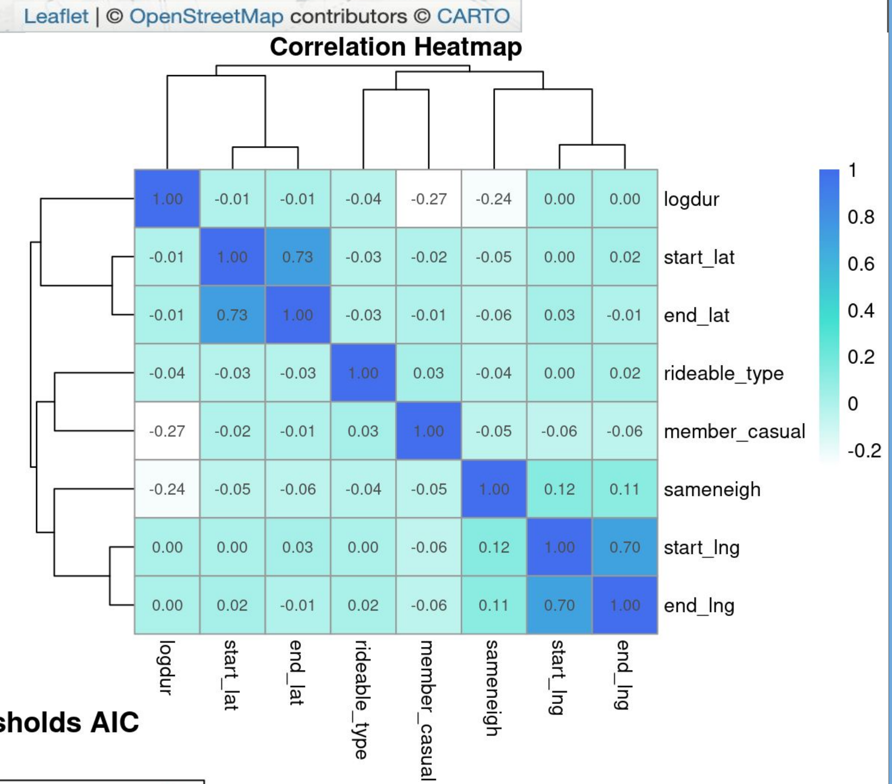
Source: “Blue Bikes Comprehensive Trip Histories”, a comprehensive dataset outlining individual trips made by patrons including the time of the trip, the start and ending stations, the bike’s ID, the bike type, and the patron’s user type.

Manipulation:

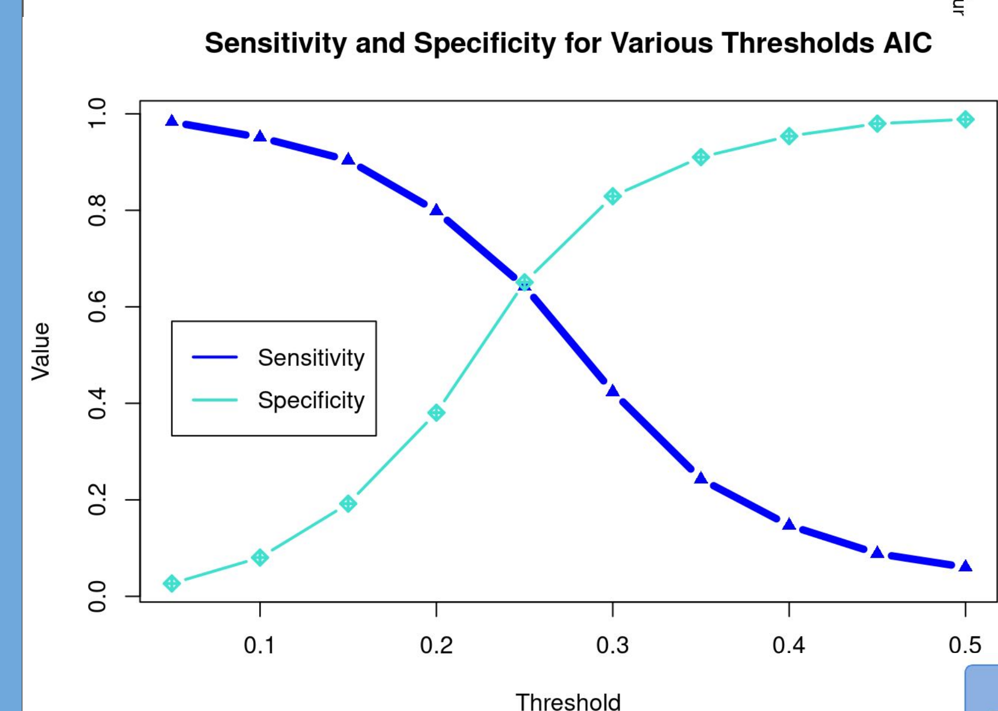
- Combined all 2024 trip history data into one dataset, sortied the data by date.
- Created a separate date, time started, and time ended variables for each trip.
- Deleted rows with NA values.
- Created variables representing the neighborhoods in which users started and ended their rides, as well as one signifying whether or not users started and ended in the same neighborhood.
- Created a subset of 10,000 random rows for analysis, marking 8000 for training models and the other 2000 for testing.
- Ran a variance inflation factor test in order to assess multicollinearity. No variables had severe multicollinearity so all variables remained in the dataset for modeling.



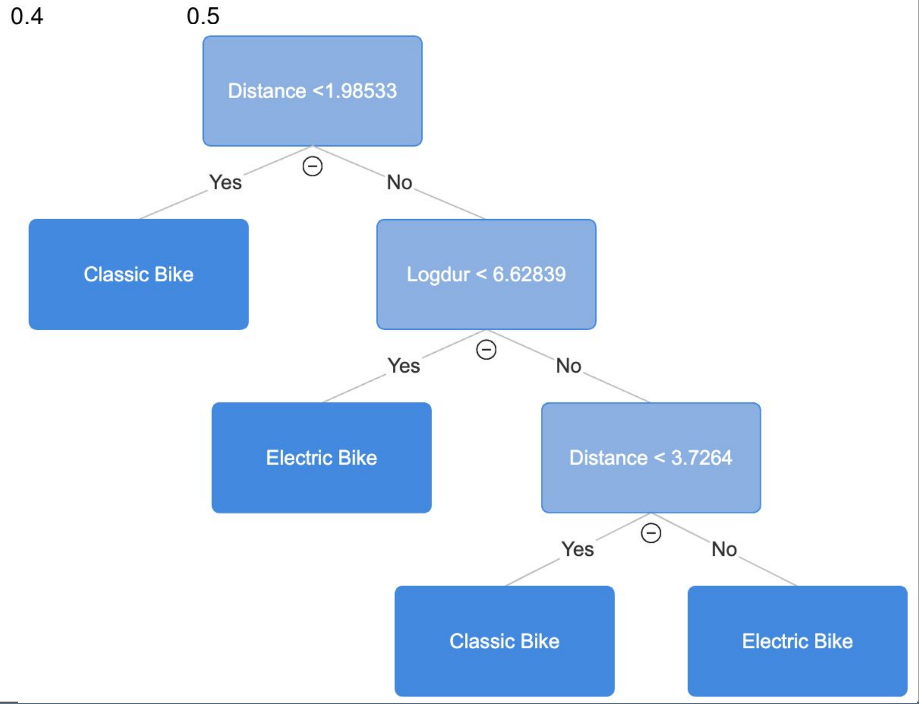
Maps of electric bike usage at starting (left) and ending (right) points. More people end their trips in the heart of Boston than start them.



Correlation heat map matrices. Many of the variables listed have small correlations to rideable_type (response variable). Latitude, trip duration, and starting and ending in the same neighborhood all have negative correlations, while longitude and user type (member_casual) have positive correlations



Overlaid line chart depicting sensitivity and specificity at a range of thresholds. Sensitivity is highest when threshold is low while the opposite is true for specificity.



Classification tree model. The tree created had four terminal nodes and used only two variables: Distance and logged duration. Based on 10-fold cross validation, this is the optimal pruned tree model.

Models & Analysis

I created three models based on my response variable (rideable_type). I coded a pruned tree model and two binary classification logistic regression models using stepwise regression: Akaike Information Criterion (AIC), and a Bayesian Information Criterion (BIC). After comparing all three models using various measurements (listed below), the AIC model proved to be the most optimal.

	Accuracy	Sensitivity	Specificity	AUC
AIC	**0.742**	**0.060**	0.060	**0.674**
BIC	0.740	0.045	**0.991**	0.673
Tree	0.735	0.180	0.935	0.598

Next, I assessed different thresholds to find the optimal threshold for sensitivity and specificity. I prioritized sensitivity because I wanted to predict the usage of electric bikes specifically. I thus decided to use the optimal threshold for sensitivity: 0.015.

Results & Discussion

Result: The final model uses the variables of logged duration, distance travelled, the neighborhood and city in which they started, whether or not their trips started and ended in the same neighborhood, and starting and ending latitudes.

Key Findings:

- People are less likely to rent e-bikes if their trip starts and ends in the same neighborhood.
- E-bike rentals are more likely in neighborhoods like Dorchester and East Boston and cities like Brookline and Revere, but less likely in neighborhoods like Cambridge and Riverside.
  - Neighborhood data may be skewed due to limited sample size & response imbalance
- Trips ending further east are more likely to involve e-bikes, while trips starting eastward are less likely. Users may use e-bikes to commute into cities like Boston but use other means for departure.
- Trip duration and distance have a smaller effect, but faster and farther trips are more likely to involve e-bikes, as they offer a time-saving alternative for longer journeys.
- This information can guide future decisions on e-bike distribution and station allocation.