From Station to Station

Predicting Blue Bikes Trip Duration Using Boston Neighborhood Characteristics

Hannah Chiou '25 | Wellesley College Data Science Capstone

WELLESLEY

Research Question

What aspects of Boston neighborhoods are the most significant predictors of Blue Bike trip length?

Background

→ Blue Bikes are used across Boston and other municipalities as a means of transportation for both residents and tourists

→ How are people using the bikes? Can we infer this by looking at trends across length and geography?

Data

Primary data sources

→ Blue Bikes ride data from January 2023 - filtered only for rides that began and ended in Boston

Secondary data sources

→ Blue Bikes station data - filtered only for Boston neighborhoods

→ Boston zip code information - used to determine neighborhoods and whether or not trip began or ended in same neighborhood

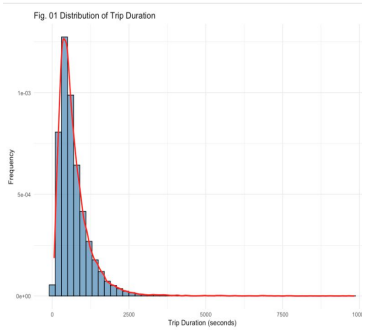
Manipulation

→ Station data was reverse geocoded using latitude and longitude data to get zipcode, then combined with zipcode data to get neighborhood

→ Ride and station data were merged to get information about start and end neighborhoods of rides

→ Neighborhood information was aggregated (count of total number of stations and docks in each neighborhood) and also added to the final dataset

→ Crucially, response variable was logged (see Figure 01) due to highly skewed distribution



Methodology

→ Run three different methods for variable selection and compare resulting models using k-fold cross validation and resulting RMSE

Stepwise regression analysis

→ *Selecting for best AIC:* log_tripduration ~ end_neighborhood_stations + distance_km + same_neighborhood

→ *Selecting for best BIC:* log_tripduration ~ distance_km + same_neighborhood

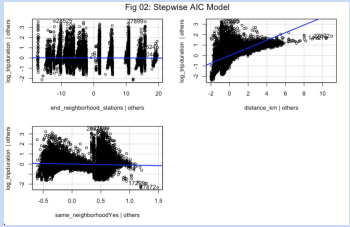
Best subset selection

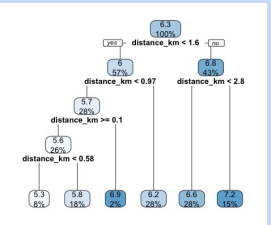
→ *Selecting for best AIC: n/a*

→ *Selecting for best BIC:* log_tripduration ~ distance_km + same_neighborhood

Regression tree

→ *Model:* log_tripduration ~ distance_km





	Model	CV_RMSE
1	Stepwise AIC	0.5215855
2	Stepwise BIC	0.5216887
3	BIC Subset	0.5217552
4	Tree Model	0.5235816

AIC model is best → best predictors are # of end neighborhood stations, distance, and start/end

Discussion and Conclusion

→ Must be careful with interpretation: *percentage change* in the trip duration for each unit change in the predictor variables

→ Though we conclude that the best set of predictors are the number of stations in the ending neighborhood, the distance of the trip, and whether or not the trip began and ended in the same neighborhood, our results are not strong overall (RMSE values not very low)

→ Our partial regression plots reveal that *distance* is the strongest predictor for logged trip length

→ Limitations and future work: rider-specific and neighborhood characteristics could consider more complex non-linear models for better predictive performance