Predicting Membership Status of Blue Bike Trips

Caroline Jung ’25, Data Science Major Capstone

Introduction

Motivation
- Bike share is a form of public transportation that promotes sustainability
- Collect data on bike usage and its membership marketing tactics for the city of Cambridge

Research Question
- How can we best predict the membership status for Blue Bike trips taken within Cambridge in 2024 based on rider & trip attributes?

Data Cleaning

- Dataset: Blue Bikes 2024 trip history and station info, only trips taken within Cambridge (962052 observations)
- New variables:
  - Month (categorical)
  - Round-trip (binary categorical)
  - Trip length (in logged mins) – difference in start & end time
  - Time of day (categorical) – start & end times based on hour
- Dropped rows with suspected recording errors (start times > end times), which was only 0.00003% of observations
- Created smaller data subset due to limited processing power
- Dimensions: 12000 observations x (8 predictors + 1 response)

Visualizations

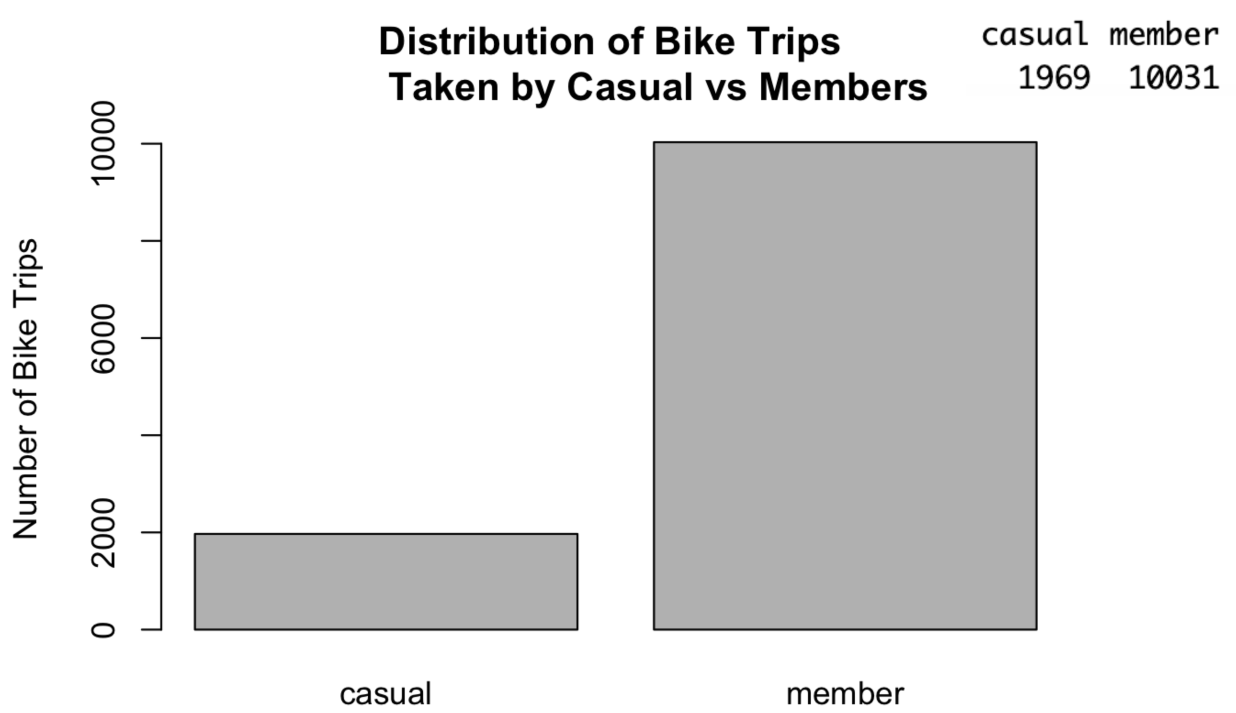


Fig 1. Bar plot of the class imbalance for binary response variable (member vs casual).

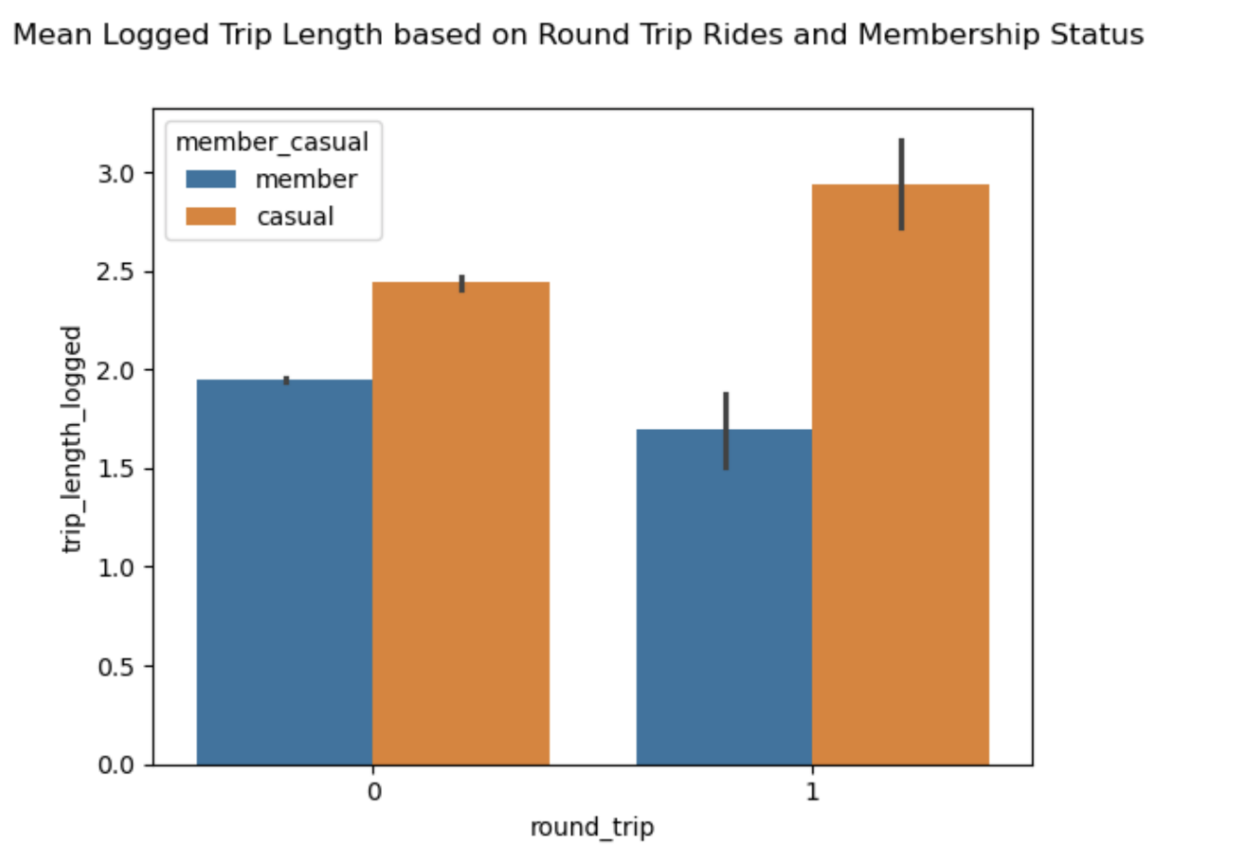


Fig 2. Side-by-side bar plot of the relationship between round trips, trip lengths (logged), and membership status.

Methodology

STEP 1: Screen for multicollinearity ⇒ none

STEP 2: Identify best first-order model & tune threshold via 10-fold CV

	F-measure				Difference (|TPR-TNR|)			
	Logit		Probit		Logit		Probit	
Threshold	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
0.8	0.838	0.841	0.835	0.838	**0.263**	0.314	**0.242**	0.301
0.7	0.900	0.906	0.899	0.906	0.651	0.697	0.653	0.703
0.6	0.913	0.913	0.913	0.913	0.821	0.834	0.827	0.841
0.5	0.915	0.915	0.915	0.915	0.873	0.895	0.879	0.901
0.4	0.915	0.915	0.915	0.915	0.925	0.931	0.930	0.936

Table 1. Performance metrics for first-order models at different thresholds via 10-fold CV.

- F-measure: precision & sensitivity tradeoff regarding correct predictions for members → optimize for high value
- Difference in TPR and TNR: equilibrium considers tradeoff between correct predictions for members vs non-members → optimize for low value
- Best first-order model: logit AIC with threshold 0.8

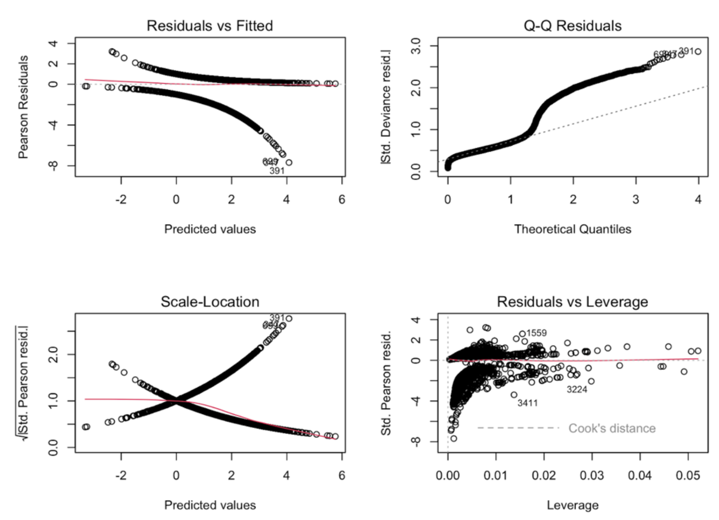
STEP 3: Consider higher order models & interaction terms in regression

	First-order logit AIC	Interaction logit BIC	Tree (bagging)	Tree (random forest)	Support Vector Machine
F-measure	**0.838**	0.838	0.789	0.826	0.914
|TPR-TNR|	**0.262**	0.296	**0.166**	0.292	0.828

Table 2. Performance metrics for all potential (higher-order) models at threshold 0.8, validated via 10-fold CV.

Two proposals of best models:
1. First-order logistic regression, selected by AIC criterion: allows for quantitative analysis, prediction, and accessible interpretation
2. Classification tree with ensemble method (bagging): allows for easy interpretation and list of most important predictors

STEP 4: Check regression diagnostics, outliers, and influential observations
- Outliers do not seem to be highly influential ⇒ kept full dataset



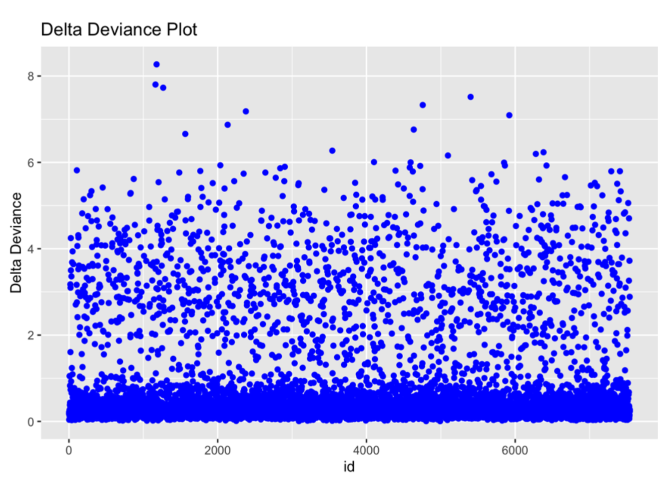


Fig 3 (left). Regression diagnostic plots for first-order logistic AIC model to identify outliers.
Fig 4 (right). Delta deviance plot for first-order logistic AIC model to identify influential observations.

ID	Member ship Status	Bike Type	Month	Round Trip	Start station #docks	End station #docks	Started time of day	Ended time of day	Trip length logged
699	casual	electric	**1**	**0**	19	19	morning	**morning**	**0.624**
391	casual	electric	**1**	**0**	27	53	morning	**morning**	1.540
347	casual	**classic**	**1**	**0**	27	18	**evening**	evening	**0.587**
3411	casual	**classic**	4	**0**	15	19	**night**	morning	2.287
3224	casual	**classic**	4	**0**	19	19	**night**	morning	3.580
1559	member	classic	1	0	19	19	**morning**	morning	2.401

Table 3. Observed data values for predictor & response variable for identified outliers. Bolded values indicate departures from the direction of correlation based on membership status in first-order logistic AIC model.

Results

Classification Tree (Bagging)

```
                        MeanDecreaseGini
rideable_type                   77.19494
month                          466.46088
round_trip                      42.71039
start_station_total_docks      370.24665
end_station_total_docks        332.92416
started_time_of_day            142.12780
ended_time_of_day              136.97625
trip_length_log               1707.39568
```

Top 3 most important variables:
1. **Trip length (logged)**
2. **Month**
3. **Total # docks at stations**

Fig 5. Importance plot of classification tree determined by bagging ensemble method.

- MeanDecreaseGini: avg decrease in node heterogeneity from splitting on variable
- Higher scores → more homogeneous nodes → indicate important variables

First-Order Logistic AIC Model

	Predictor	Coefficient	Interpretation: expect odds of being a member to…	
	Bike type: Electric	-0.044	- 4.30%, compared to trips taken with classic bike	
	February (month 2)	-0.097	- 9.24%	compared to trips taken in January (month 1)
	March (month 3)	-0.262	- 23.05%	
***	April (month 4)	-0.613	- 45.83%	
***	May (month 5)	-0.666	- 48.62%	
***	June (month 6)	-0.675	- 49.08%	
***	July (month 7)	-0.977	- 62.36%	
***	August (month 8)	-0.824	- 56.13%	
***	September (month 9)	-0.755	- 53.00%	
***	October (month 10)	-0.550	- 42.31%	
***	November (month 11)	-0.513	- 40.13%	
	December (month 12)	-0.288	- 25.02%	
***	Round trip: yes	-0.946	- 61.17%, compared to non-round trips	
**	Start station total docks	0.013	+ 1.31%	
***	End station total docks	0.030	+ 3.05%	
	Started time of day: afternoon	0.276	+ 31.79%	compared to trips starting in the morning
**	Started time of day: evening	0.708	+ 102.99%	
	Started time of day: night	0.346	+ 41.34%	
***	Ended time of day: afternoon	-0.777	- 54.02%	compared to trips ending in the morning
***	Ended time of day: evening	-0.931	- 60.58%	
**	Ended time of day: night	-0.900	- 59.34%	
***	Trip length (logged)	-0.887	- 58.81%	

Table 4. Regression coefficients & interpretation on odds scale for first-order logistic AIC model. Positively (blue) and negatively (green) correlated coefficients indicated.

Discussion & Limitations

Discussion & Conclusion
- Both models indicate trip length, month, and number of docks at stations to be among the most significant predictors – **trips are more likely to be taken by members if trips were shorter, taken in January, and started/ended at stations with more docks**
- I propose two models to the city of Cambridge:
  - The classification tree may be preferred due to easy interpretation of important trip attributes and no inherent assumptions
  - The first-order logistic AIC model may be preferred for a more comprehensive quantitative analysis of significant attributes, while maintaining accessible interpretation (no higher order terms)

Limitations & Future Work
- All models do not account for time series analysis (month was treated as a categorical variable with independent levels), concentrated to one year
- Comparative analysis with other Massachusetts cities (spatial analysis)

References: https://bluebikes.com/system-data