Predicting Membership Status of Blue Bike Trips

Caroline Jung ’25, Data Science Major Capstone

Introduction

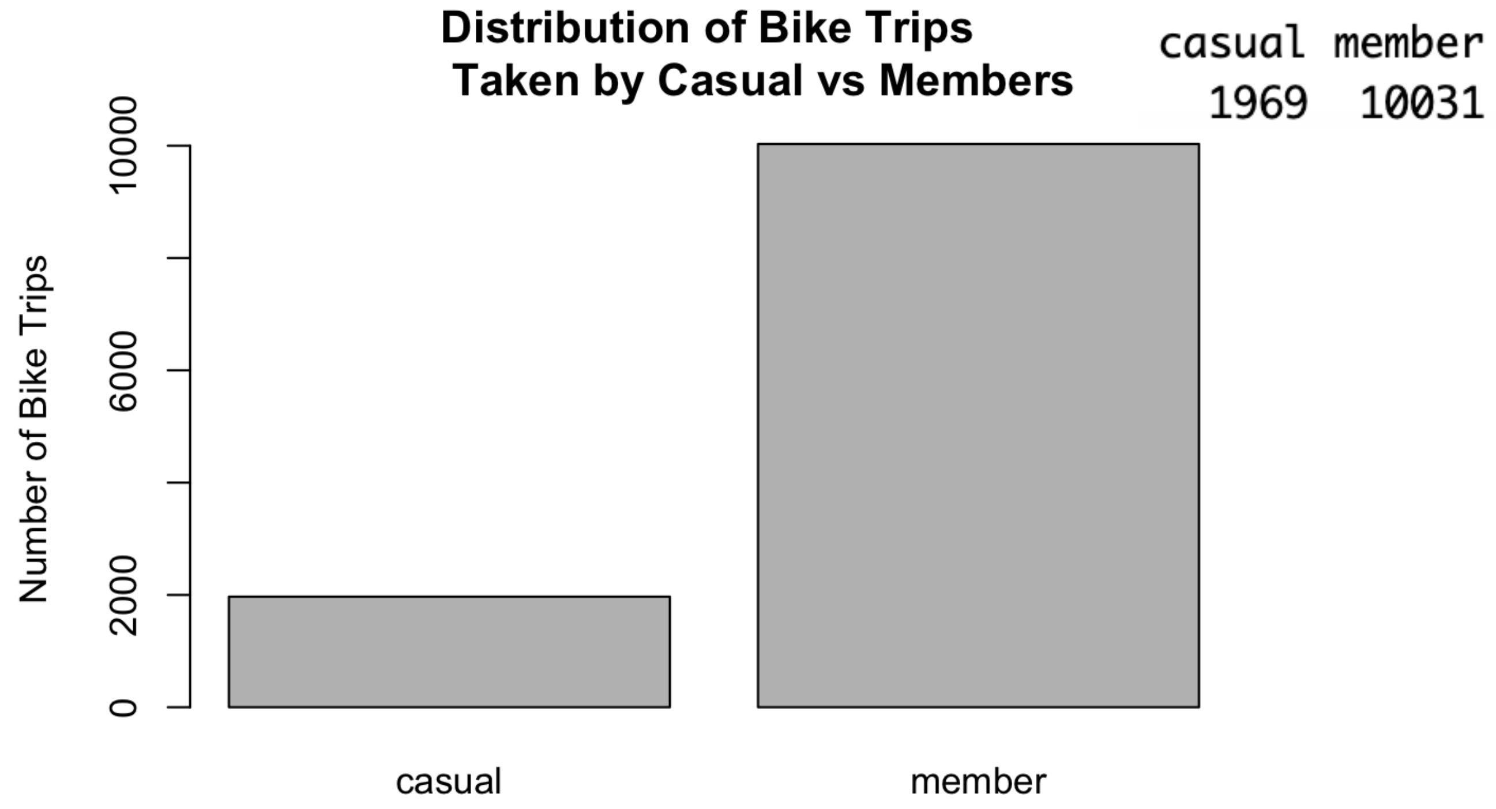
Research Question

- 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)
  - Time of day (categorical)
  - Dropped rows with suspected recording errors
- Used smaller data subset due to limited processing power
- Dimensions: 12000 obs. x (8 predictors + 1 response)

Visualizations



```
casual member
  1969  10031
```

Fig 1. Bar plot of the class imbalance for binary response variable.

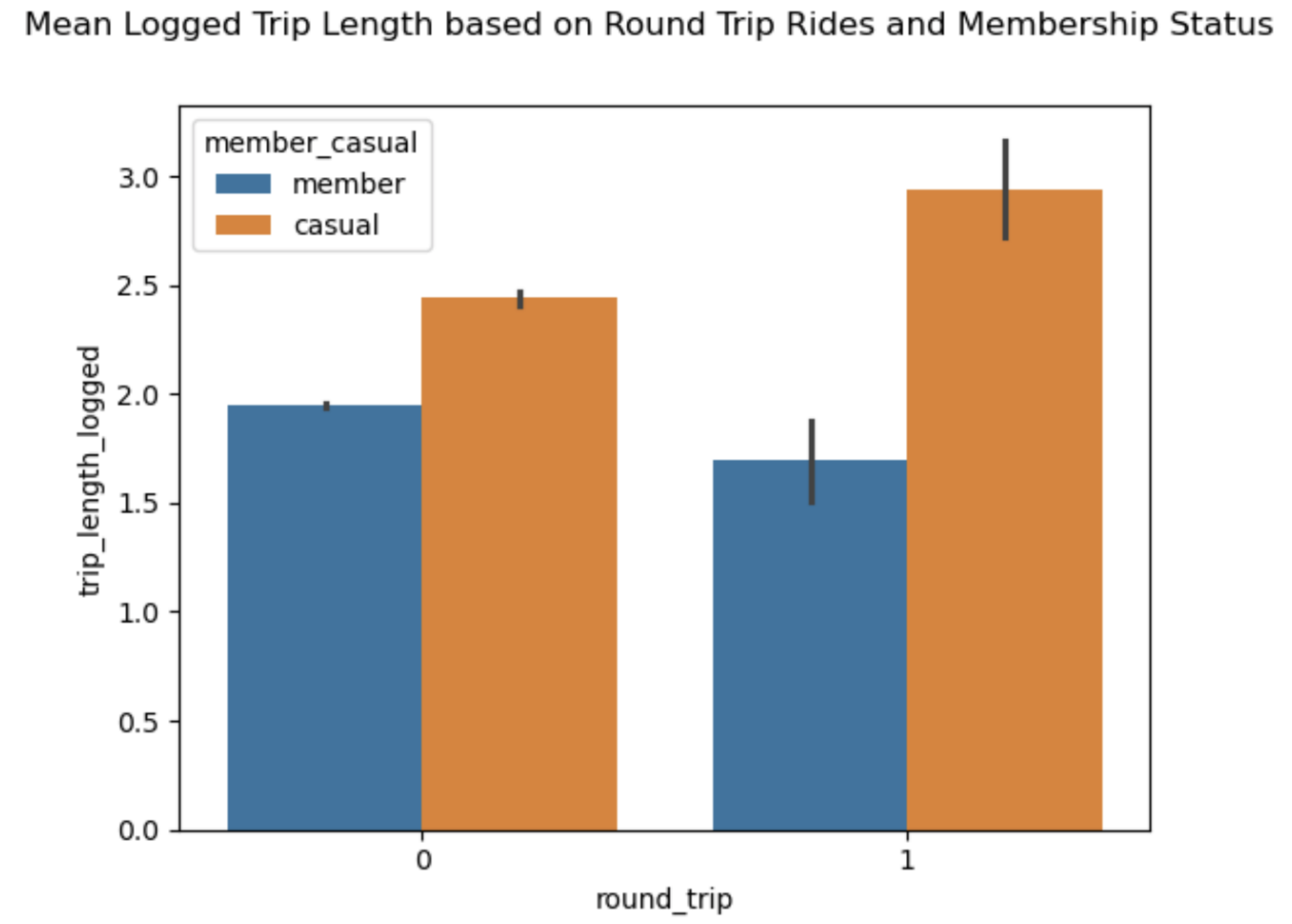


Fig 2. Side-by-side bar plot of the relationship between round trips, trip lengths (logged), and membership status.

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|)				TNR (specificity)			
	Logit		Probit		Logit		Probit		Logit		Probit	
Threshold	AIC	BIC	AIC	BIC	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	**0.524**	0.483	**0.537**	0.490
0.7	0.900	0.906	0.899	0.906	0.651	0.697	0.653	0.703	0.280	0.249	0.278	0.245
0.6	0.913	0.913	0.913	0.913	0.821	0.834	0.827	0.841	0.155	0.143	0.150	0.138
0.5	0.915	0.915	0.915	0.915	0.873	0.895	0.879	0.901	0.113	0.096	0.109	0.089
0.4	0.915	0.915	0.915	0.915	0.925	0.931	0.930	0.936	0.069	0.065	0.065	0.060

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

- F-measure: precision & sensitivity tradeoff → optimize high value
- Difference in TPR and TNR: tradeoff correct predictions for members vs non-members → optimize low value
- TNR (specificity): TPR for non-members → optimize high 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
TNR	**0.525**	0.495	**0.542**	0.482	0.150

Table 2. *Performance metrics for higher-order models at threshold 0.8, validated via 10-fold CV.*

Two proposed best models:

1. First-order logistic regression, selected by AIC criterion: allows for quantitative analysis, prediction, accessible interpretation
2. Classification tree from bagging: allows for easy interpretation, list of most important predictors

STEP 4: Check regression diagnostics, outliers, and influential observations

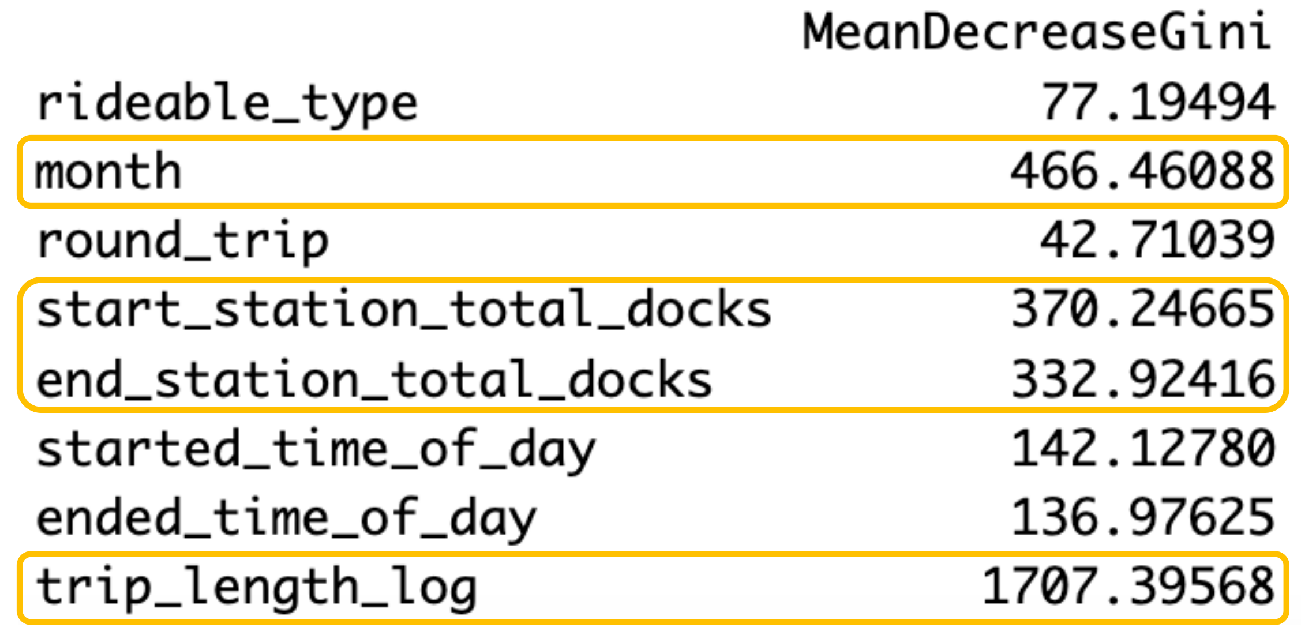
- Outliers do not seem to be highly influential ⇒ kept full dataset

Results

Classification Tree (Bagging)

Top 3 most important variables:

1. Trip length (logged)
2. Month
3. Total # docks at stations



```
                       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
```

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

First-Order Logistic AIC Model

- Positive correlation: start station total docks, end station total docks, started time of day
- Negative correlation: bike type, month, round trip, ended time of day, logged trip length

Conclusion

Discussion

- 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