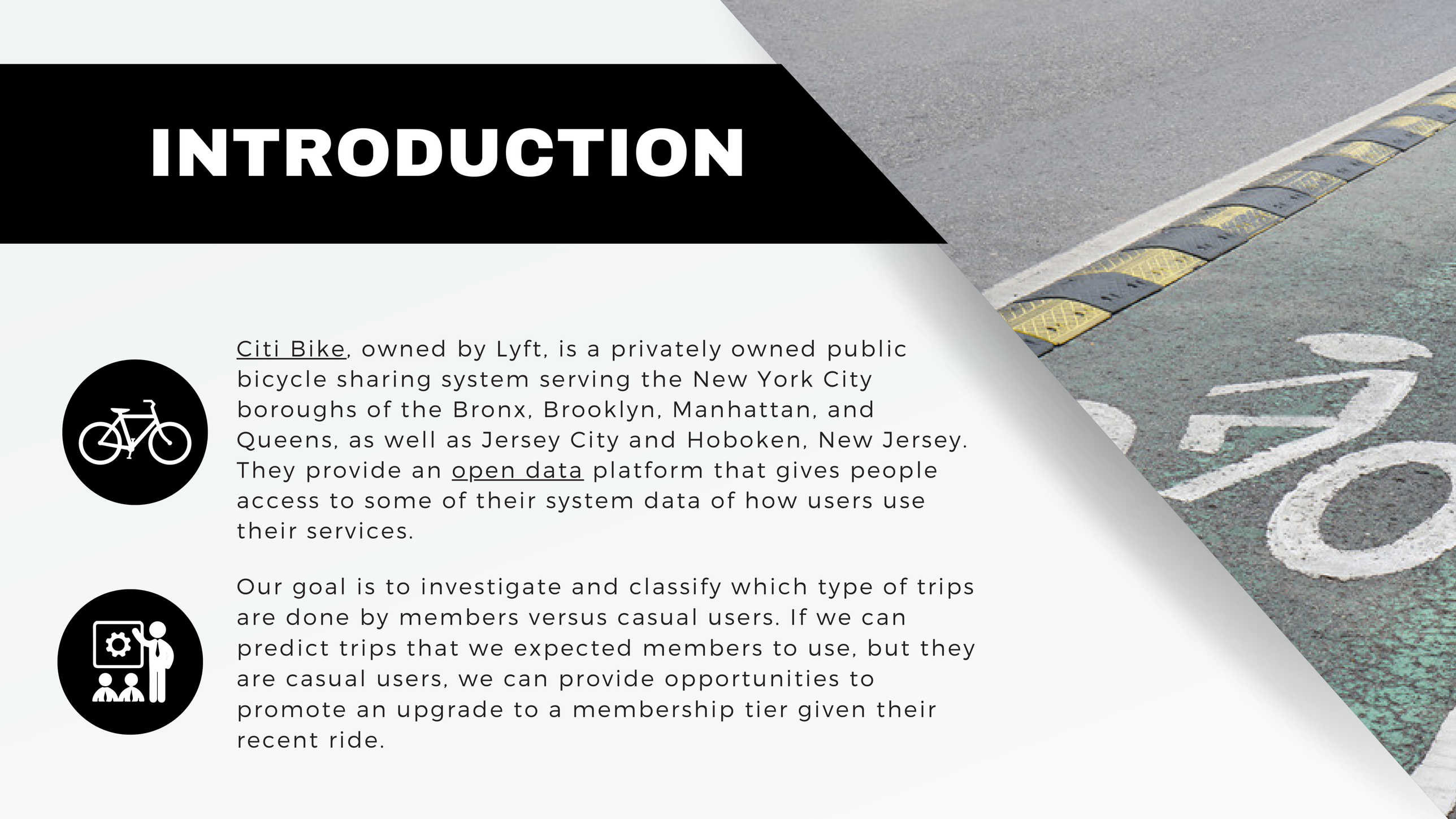
DATA 621 - FINAL PROJECT

CITI BIKES

DANIEL CRAIG, JOHN CRUZ, SHAYA ENGELMAN,
NOORI SELINA, GAVRIEL STEINMETZ-SILBER

CONTENT

INTRODUCTION



Citi Bike, owned by Lyft, is a privately owned public bicycle sharing system serving the New York City boroughs of the Bronx, Brooklyn, Manhattan, and Queens, as well as Jersey City and Hoboken, New Jersey. They provide an open data platform that gives people access to some of their system data of how users use their services.

Our goal is to investigate and classify which type of trips are done by members versus casual users. If we can predict trips that we expected members to use, but they are casual users, we can provide opportunities to promote an upgrade to a membership tier given their recent ride.

LITERATURE REVIEW

Literary Findings and References

- Fischman et al. (2015) identified several predictors of bikeshare membership, including reactions to mandatory helmet legislation, riding activity over the previous month, and the degree to which convenience motivated private bike riding.
- Kaviti et al. (2019) highlighted differences in trip purposes, demographic profiles, and pricing preferences between bikesharing members and casual users.
- Reilly et al. (2020) found that low-income people are more likely to become members of bikesharing programs due to having a smaller rate of car ownership.
- Chen et al. (2024) emphasized affordability as a primary barrier for lower-income earners in bikesharing usage.

LITERATURE REVIEW

Strengths and Weaknesses

- Methodological Rigor: Each study employs robust methodologies such as logistic regression and custom surveys, ensuring comprehensive analysis and reliable insights.
- Diverse Perspectives: The studies explore various aspects including user behaviors, demographic profiles, and socioeconomic disparities, enriching the understanding of bikesharing dynamics.
- Policy Implications: The findings offer actionable insights for bikesharing operators and policymakers to enhance accessibility and address disparities in bikesharing usage.

- Limited Generalizability: The findings may be specific to the studied locations (Melbourne, Brisbane) and may not fully represent bikesharing dynamics in other cities or regions.
- Potential Bias: The surveys and data collection methods used in the studies may introduce biases, affecting the accuracy and validity of the findings.
- **Incomplete Understanding: While the studies provide valuable insights, gaps remain in understanding how immutable ride characteristics can predict trip types, indicating avenues for further research.**

DATA EXPLORATION

01 Given computational restrictions, we only used March 2024 data as the file size is over 1 GB and contains over 2.7 million rows of data

ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casual
0FC89A53DF9D7E90	electric_bike	2024-03-07 19:49:43	2024-03-07 20:20:33	48 St & Skillman Ave	6283.05	Kingston Ave & Park Pl	4016.03	40.74615	-73.91619	40.67308	-73.94191	member
0FF38F5D1277746B	electric_bike	2024-03-15 17:45:30	2024-03-15 17:55:39	Liberty St & Broadway	5105.01	Mercer St & Spring St	5532.01	40.70886	-74.01023	40.72363	-73.99950	member
DE040AD144FB0BFA	electric_bike	2024-03-19 18:00:52	2024-03-19 18:07:26	W 56 St & 6 Ave	6809.07	E 43 St & Madison Ave	6551.11	40.76341	-73.97722	40.75355	-73.97897	member
5C7DFD80B04BBA5A	electric_bike	2024-03-05 17:25:30	2024-03-05 17:30:17	W 56 St & 6 Ave	6809.07	E 43 St & Madison Ave	6551.11	40.76306	-73.97767	40.75355	-73.97897	member
5C0A03B95B0D0A0F	electric_bike	2024-03-22 13:18:37	2024-03-22 13:23:24	5 Ave & W 126 St	7701.21	Frederick Douglass Blvd & W 115 St	7658.13	40.66095	-73.98304	40.80387	-73.95593	member
6F87E826DC6091B5	classic_bike	2024-03-13 09:34:15	2024-03-13 10:04:35	6 Ave & W 45 St	6593.15	Adam Clayton Powell Blvd & W 118 St	7670.09	40.75695	-73.98263	40.80437	-73.95148	casual

TRIPS DATASET

02 Trips Predictor Variables

rideable_type	type of bike rented (electric or classic)
started_at	datetime rental was taken from the station
ended_at	datetime rental was returned to a station
start_station_name	bike taken from station
start_station_id	ID of start_station_name
end_station_name	bike returned to station
end_station_id	ID of end_station_name
start_lat	starting station latitude
start_lng	starting station latitude
end_lat	ending station latitude
end_lng	ending station latitude

Trips Response Variable

member_casual	whether the rental was used by a member or casual (one-time rental) user. We will encode 1 as member and 0 as casual.

DISTANCE TRAVELED

03 We have latitude and longitude values, where if we had the monetary allowance, we could create a query using gmapsdistance, that would calculate the distance traveled based on the mode of transportation. It would provide the best accurate results to determine traveled distance on these bikes. However, given these restraints we could estimate the values based on the following data

- Classic bikes travel around 8.3 miles per hour
- Electric bikes travel up to 20 miles per hour**

Developing Predictor Variables

usage_time	how long the ride was for (in seconds)
est_distance	estimated traveled distance based on bike usage (in miles)

** To limit the maximum speeds on the electric bikes given the crowded nature of New York City, we will arbitrarily limit the speeds up to 15 miles per hours in the calculation.

WEATHER DATASET

04 Weather Predictor Variables

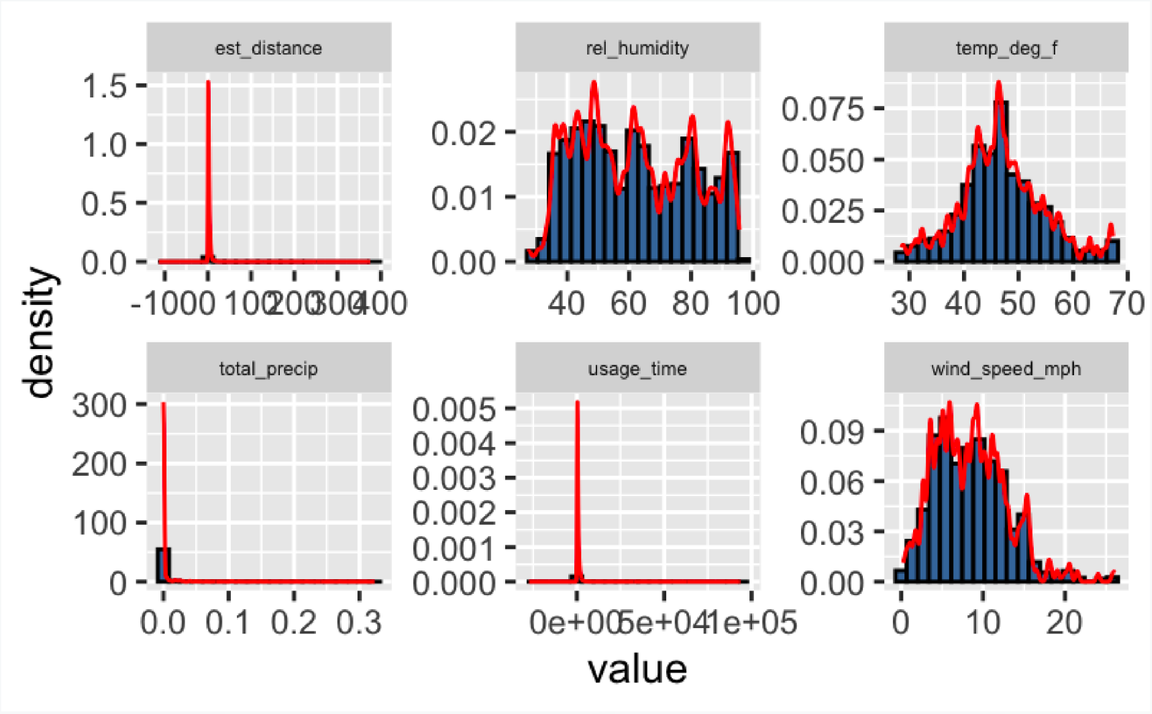
We also obtained weather data for the month of March to use as predictors in our models

temp_deg_f	temperature (Fahrenheit)
rel_humidity	relative humidity
total_precip	total precipitation (inches)
wind_speed	wind speed (miles per hour)
day_of_week	day of the week (Monday, Tuesday...)

DENSITY PLOTS

05

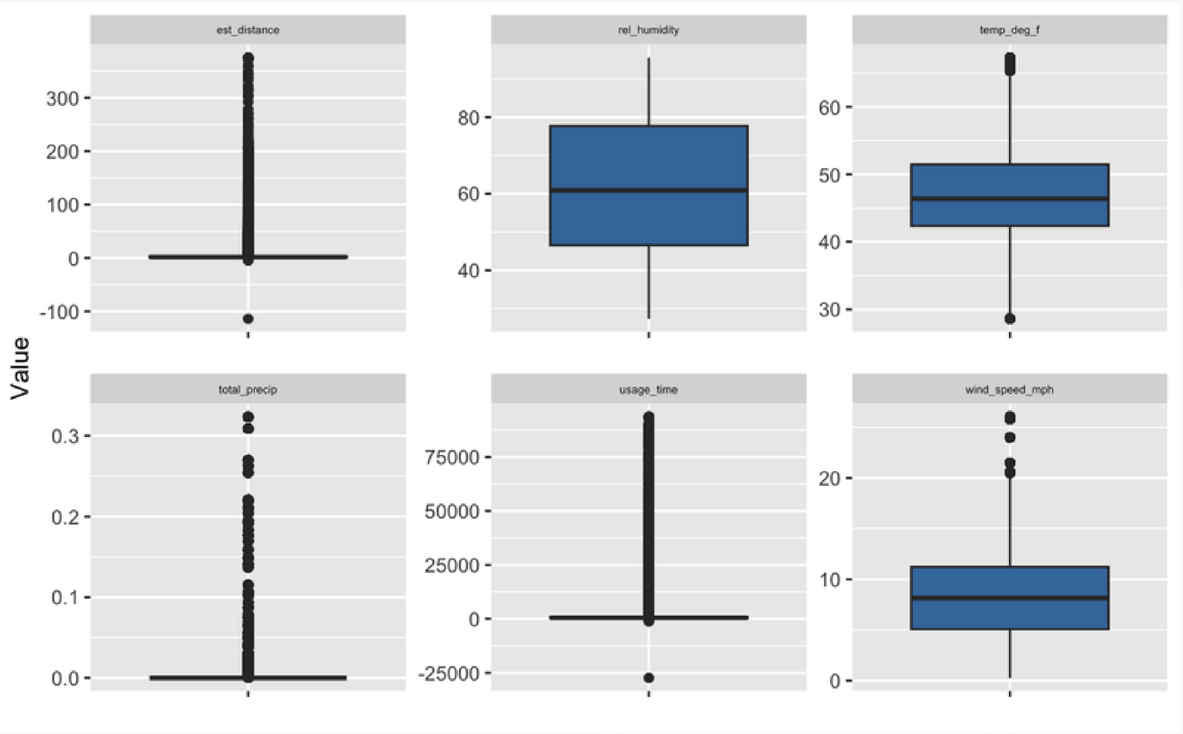
- The plots show significant right skew in **wind_speed_mph**, **usage_time** and **est_distance**
- We have a Poisson distribution in **total_precip**. These skewed variables might be candidates for transformation.
- The plot also shows **rel_humidity** is multi-modal and hovers around every 20% increase in humidity.
- We also see issues with outliers that we need to investigate.



BOX PLOTS

05

- These boxplots further confirm the skewness mentioned earlier.
- They also reveal that variables **est_distance**, **total_precip** and **usage_time** all have a large amount of outliers.



CORRELATION MATRIX

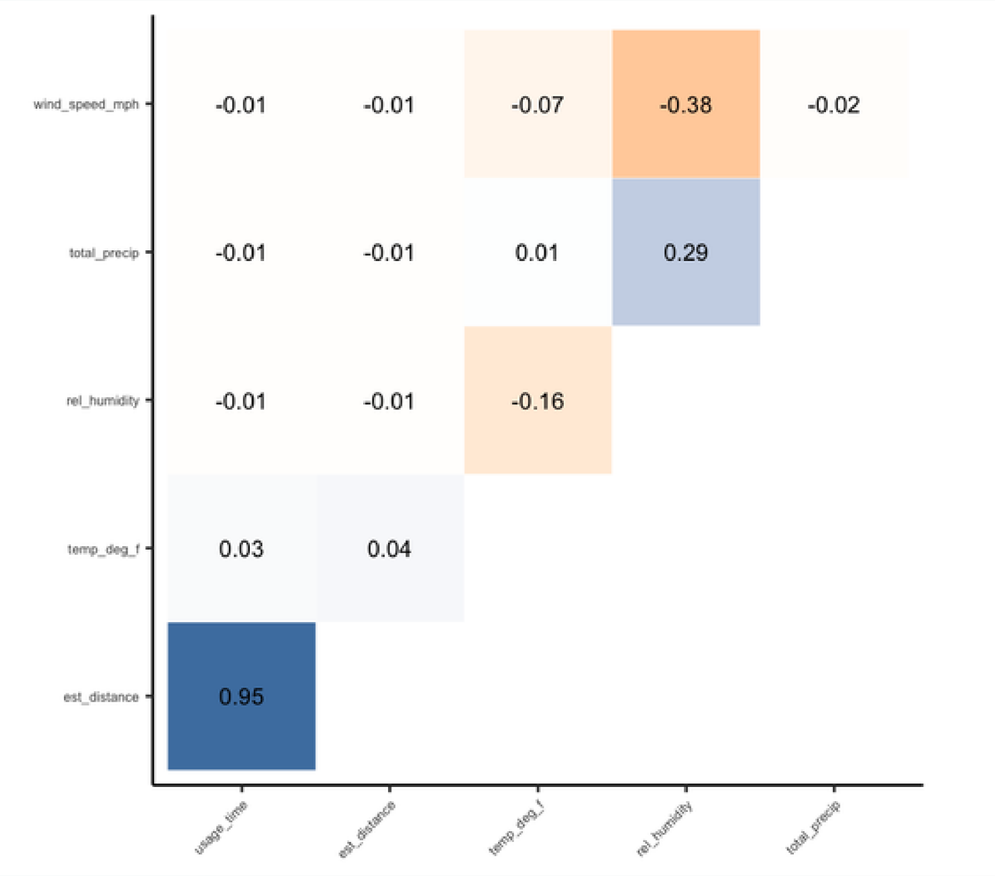
06

Negative Correlations:

- Predictors **wind_speed_mph** and **rel_humidity** exhibit negative correlations with each other, indicating that as the relative humidity increases, the likelihood of the wind speed being above the median decreases. It is interesting, as we could have assumed more humidity brings a higher chance of rain and also wind speeds.

Positive Correlations:

- Conversely, predictors such as **usage_time** and **est_distance** exhibit strong positive correlations with each other. This makes sense as we derived the distance traveled based on how long they rode the bike for. We also see some positive relationship between **rel_humidity** and **total_precip** which intuitively makes sense.

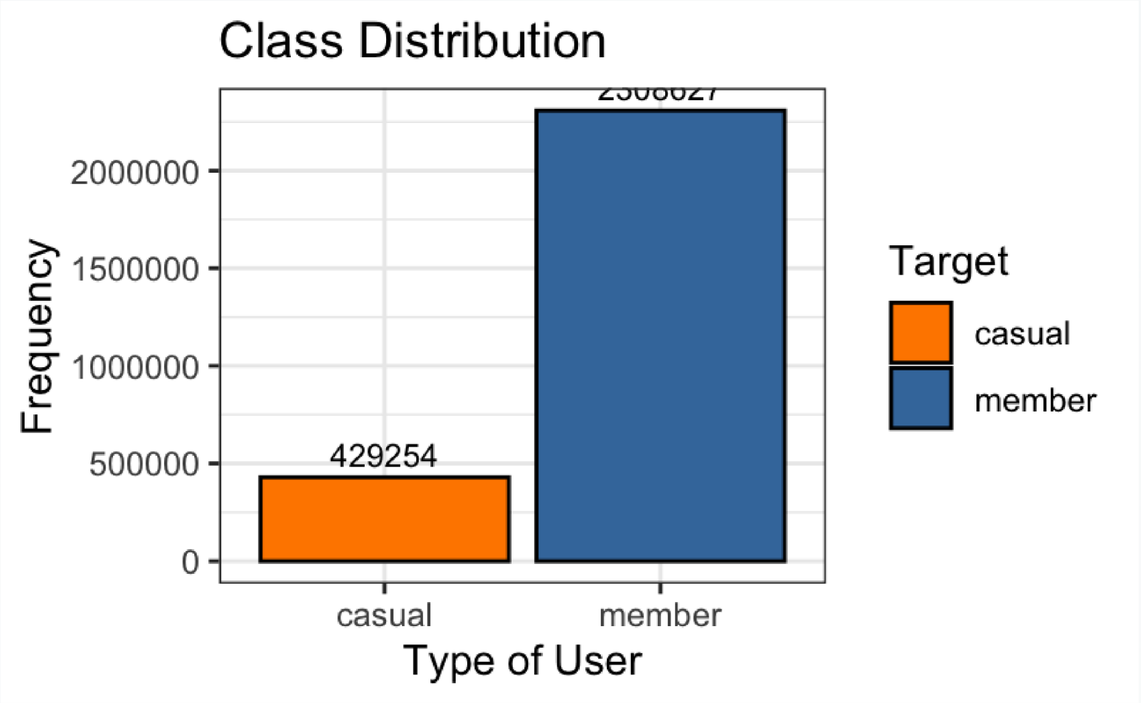


TYPES OF USERS

07

Class Imbalance:

- We definitely see most users are members as opposed to casual non-member users. We need to keep in mind which metrics we will use to evaluate our models because of this.

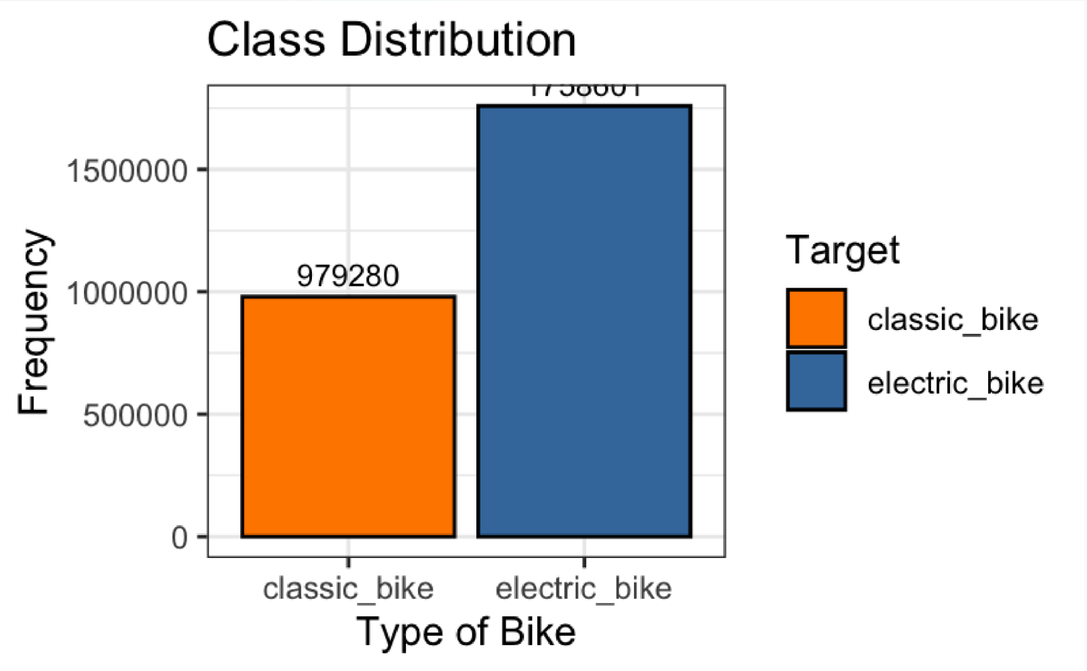


TYPES OF BIKES

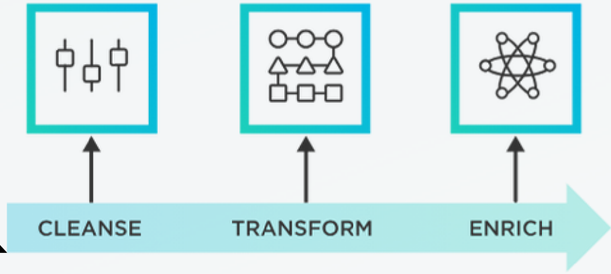
08

Class Imbalance:

- We also see how electric bikes are about twice as likely to be used compared to the classic bike. This may be due to inventory or preferences on the bike users are willing to take trips with.



DATA PREPARATION



1. Data Cleaning and Splitting

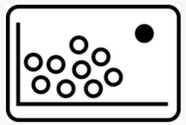
- We dropped missing values from our dataset to ensure data integrity,
- We then split the dataset to training and testing sets using a 70/30 ratio to prevent data leakage.

2. Best Normalization:

- We used the **‘bestNormalize’** function to identify optimal transformations for skewed variables ensuring they meet normal distribution assumptions.

3. Transformations:

- Following the identification of the best transformations, we applied the recommended methods, including logarithmic and square root transformation.
- This normalized the distributions, crucial for improving the accuracy and reliability of our models.



4. Outliers:

- We decided to keep outliers in the dataset, believing they could provide valuable insights and help us capture the full score of data variability. This decision will be re-evaluated in the modeling phase.

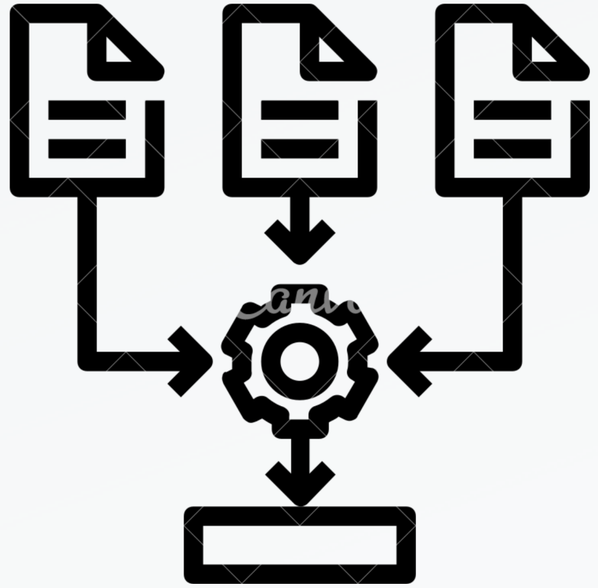
MODELING

01 There were two goals in modeling:

- Establish what features distinguish casual users from members
- Consider when casual users are more likely to exhibit “member-like behavior”

02 To predict casual users vs. members, we used binary logistic regression

03 To predict when casual users are likely to act like members (we will clarify the meaning of this shortly), we used binary logistic regression with splines as well as linear discriminant analysis (LDA)



MODELING (PART 1)

01 Undersampling was used to addreess class imbalance. After variable selection, including dropping variables due to multicollinearity, the variables included *rideable_type* (electric vs. regular), weather variables, *day_of_week*, *usage_time*, and *time_of_day*

02 Takeaways include: using an e-bike increases the likelihood of being a member, so too does uncomfortable weather. Longer usage times are correlated with members, and early morning rides are correlated with casual users

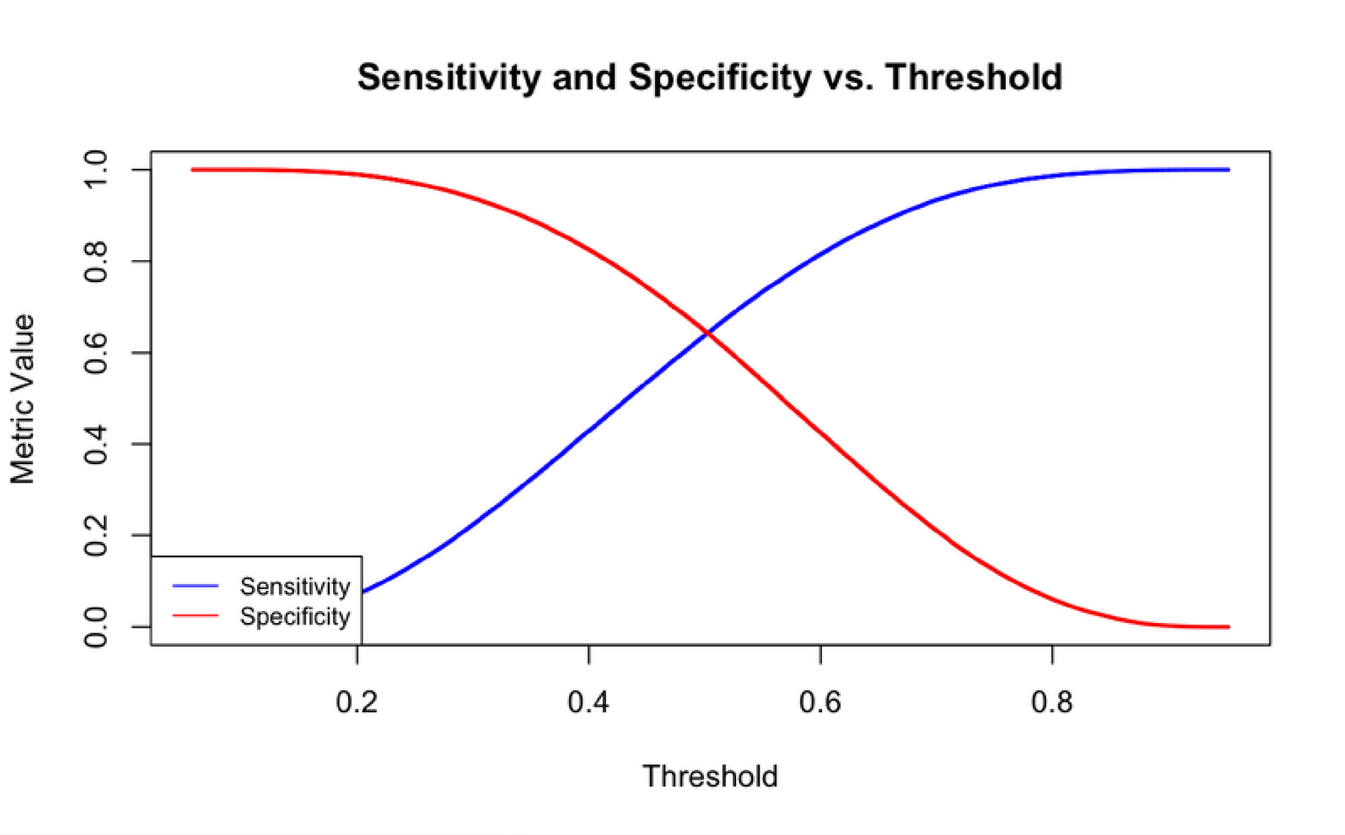
```
Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.437517   0.015663 -27.932  < 2e-16 ***
rideable_typeelectric_bike  0.419141   0.009052  46.304  < 2e-16 ***
temp_deg_f_transformed      0.114022   0.005064  22.514  < 2e-16 ***
rel_humidity_transformed   -0.055519   0.005687  -9.762  < 2e-16 ***
total_precip_transformed   -0.039670   0.004916  -8.070 7.04e-16 ***
wind_speed_transformed     -0.047863   0.004926  -9.717  < 2e-16 ***
day_of_weekTuesday         -0.003779   0.017054  -0.222   0.8246
day_of_weekWednesday       -0.046528   0.018151  -2.563   0.0104 *
day_of_weekThursday        -0.007831   0.017517  -0.447   0.6548
day_of_weekFriday           0.243263   0.015697  15.497  < 2e-16 ***
day_of_weekSaturday         0.557794   0.017187  32.454  < 2e-16 ***
day_of_weekSunday           0.575379   0.015910  36.166  < 2e-16 ***
usage_time_transformed      0.584232   0.004444 131.468  < 2e-16 ***
`time_of_dayEarly Morning` -0.495642   0.019844 -24.977  < 2e-16 ***
time_of_dayEvening         -0.274401   0.012453 -22.035  < 2e-16 ***
time_of_dayMorning         -0.303975   0.012901 -23.562  < 2e-16 ***
time_of_dayNight           -0.161684   0.014559 -11.106  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 354714  on 255871  degrees of freedom
Residual deviance: 322326  on 255855  degrees of freedom
AIC: 322360

Number of Fisher Scoring iterations: 4
```

MODELING (PART 1)

01 Even though the model was adjusted repeatedly to improve performance, it remains imperfect; the ROC score is 0.698

02 Still, we can use this model to determine “member behaviors.” Two such behaviors are biking in poor weather and using e-bikes.

03 A second model with an attempt at using start and end station names was used for comparison



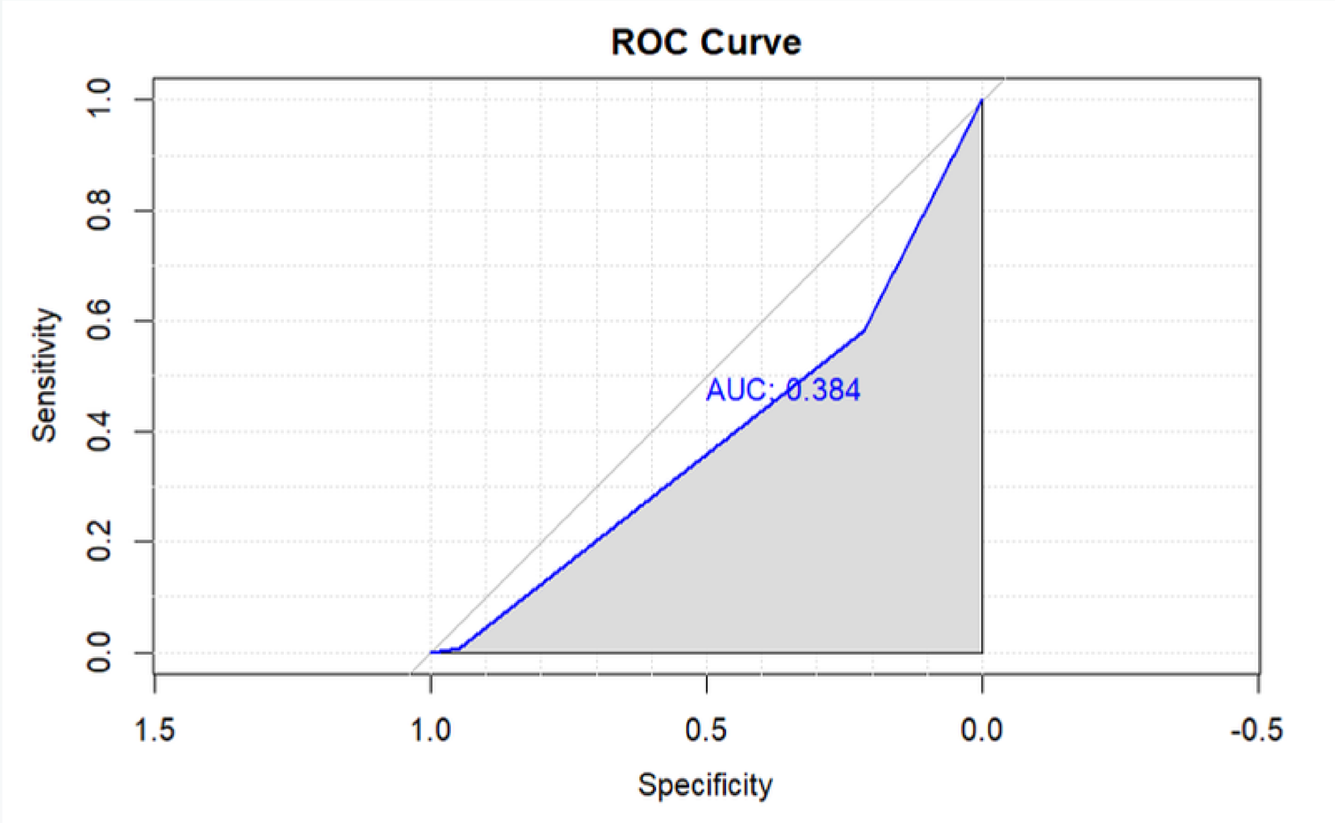
MODELING (PART 1)

01 A second model was attempted using a **boosting logistic regression** model with the start and end stations. The number of different stations counting over 2000 creating many, potentially, weak predictors

02 Class weights were used on a ratio of 6 to 1, to try a different method of dealing witht he imbalanced dataset

03 AUC: 0.384
Sensitivity: .053

04 Since this model performed so weakly, the first model was used to help determine when casual riders act like members

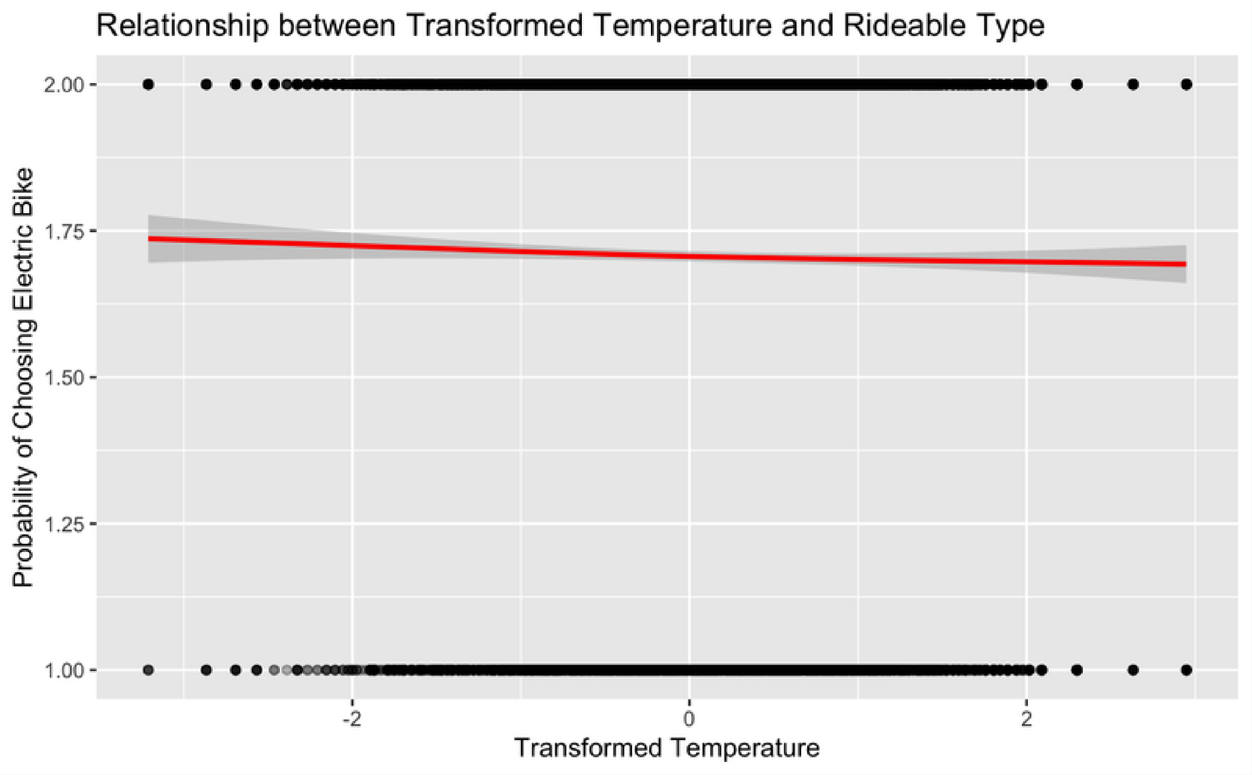


MODELING (PART 2)

01 We thus filter the data to include only casual members and attempt to answer: what factors make casual members more likely to use e-bikes?

02 Subtle as it is, the impact of temperature does not appear to be linear, and we therefore make use of splines to address this non-linearity

03 In the next model, then, we can see when casual users are likely to act like members--that's when CitiBike should attempt to convert them to members



MODELING (PART 2)

01 Surprisingly, almost all features were ***not*** associated with e-bike usage, with two notable exceptions. Rides by casual users on weekends increase the log odds of the rides being on e-bikes

02 This suggests that CitiBike should attempt to market to casual users more frequently before weekends and emphasize members' e-bike savings

03 However, caution must be exercised; the model explains only 5.33% of the deviance--the key metric for binary logistic models

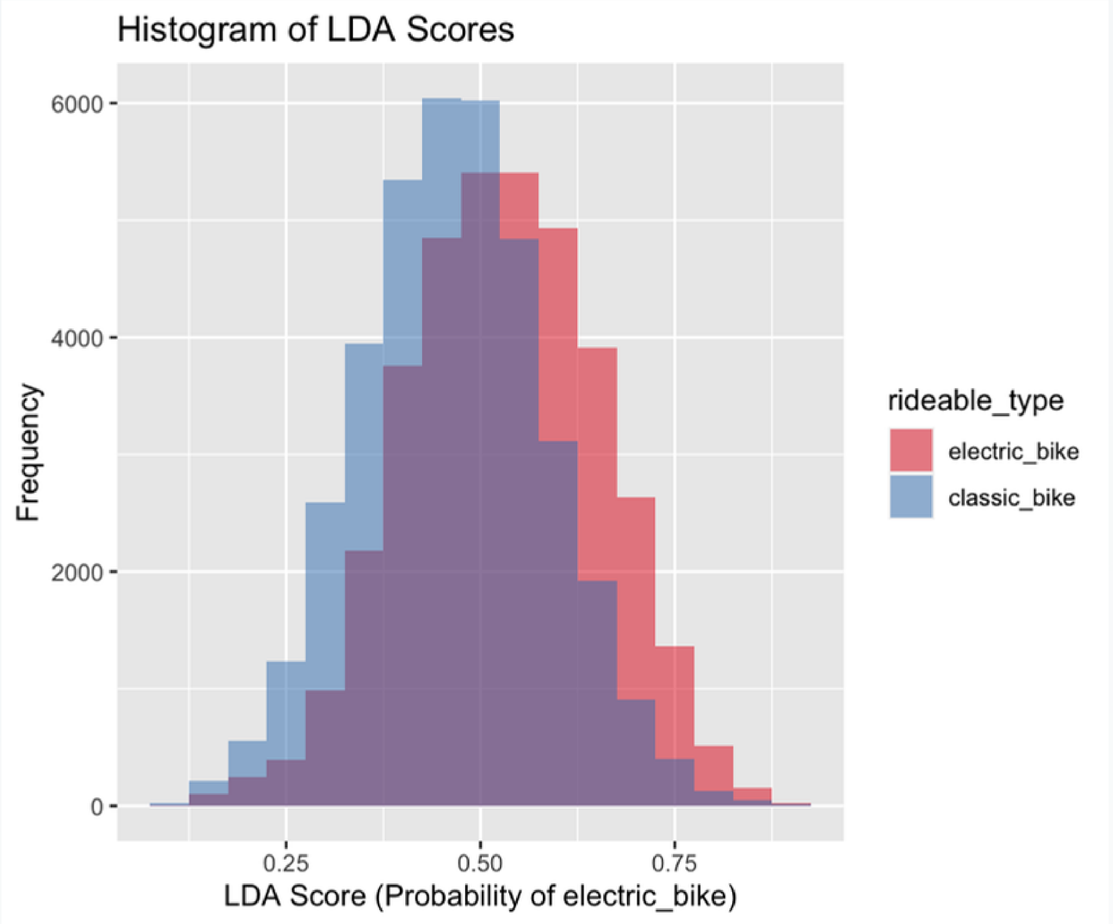
```
Formula:
rideable_type ~ s(temp_deg_f_transformed, bs = "cs") + rel_humidity_transformed +
    total_precip_transformed + wind_speed_transformed + day_of_week +
    est_distance_transformed + time_of_day

Parametric coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 0.373550   0.028952  12.902  < 2e-16 ***
rel_humidity_transformed   -0.029385   0.012479  -2.355  0.01854 *
total_precip_transformed   -0.197384   0.038791  -5.088 3.61e-07 ***
wind_speed_transformed     -0.023515   0.009832  -2.392  0.01677 *
day_of_weekTuesday         -0.006238   0.034662  -0.180  0.85718
day_of_weekWednesday       -0.064190   0.036910  -1.739  0.08202 .
day_of_weekThursday        -0.095246   0.036718  -2.594  0.00949 **
day_of_weekFriday           0.072730   0.031810   2.286  0.02223 *
day_of_weekSaturday         0.283551   0.032269   8.787  < 2e-16 ***
day_of_weekSunday           0.370257   0.030426  12.169  < 2e-16 ***
est_distance_transformed   -0.508418   0.008229 -61.780  < 2e-16 ***
time_of_dayEarly Morning   -0.995717   0.044452 -22.400  < 2e-16 ***
time_of_dayEvening         -0.320166   0.022337 -14.333  < 2e-16 ***
time_of_dayMorning         -0.076546   0.024105  -3.176  0.00150 **
time_of_dayNight           -0.736610   0.026899 -27.384  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                             edf Ref.df Chi.sq p-value
s(temp_deg_f_transformed) 7.294      9  178.8  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0721   Deviance explained = 5.33%
```

MODELING (PART 2)

01 Given these limitations we also used LDA. The derived function is:

$$
\begin{aligned}
LD1 = \; & -0.9239170 \times \textit{est\_distance\_transformed} \\
& -2.0514485 \times \textit{time\_of\_dayEarlyMorning} \\
& -0.6915460 \times \textit{time\_of\_dayEvening} \\
& -0.1769963 \times \textit{time\_of\_dayMorning} \\
& -1.4816669 \times \textit{time\_of\_dayNight} \\
& +0.2550056 \times \textit{temp\_deg\_f\_transformed} \\
& -0.4662773 \times \textit{total\_precip\_transformed}
\end{aligned}
$$

02 But the visualization shows clear limitations; the model does an inadequate job separating the classes



MODELING (CONCLUSION)

01 Thus, we began by building a model classifying users as casual or members. We then built further models predicting when casual users exhibit “member-like” behavior

02 But while the logic is solid, even after many moves to improve the models, they were unable to separate classes effectively (for both types of models)

03 This suggests that the data is incomplete; we likely need data about the users (e.g. income, etc.) to improve performance. Still, following our methodology with more complete data can yield clear insights that CitiBike can act on to convert casual users into members