Predicting Ford GoBike Rider Gender

Dan Weiss

5/16/18

Outline

1. Introduction/ Background
2. Data
3. Exploratory Data Analysis
4. Model/ Results
5. Next Steps

Introduction

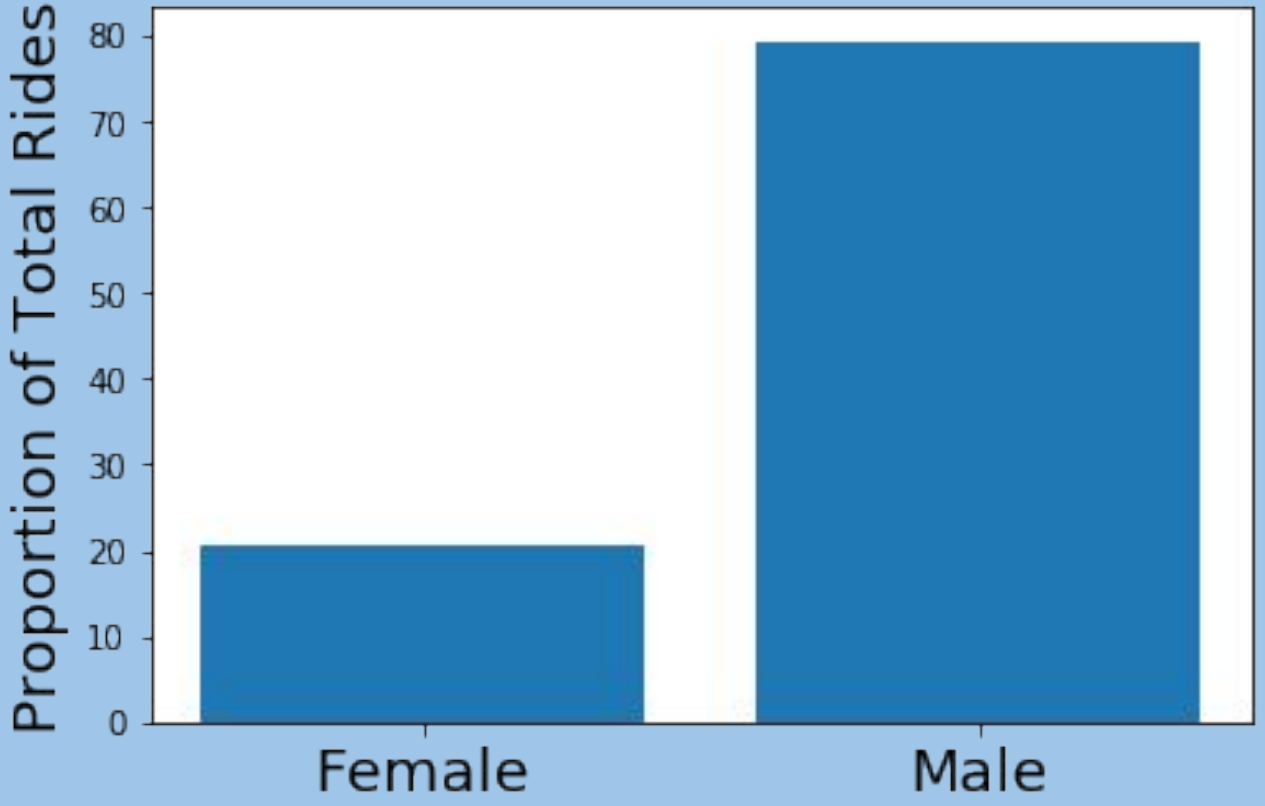
- Large gender imbalance in bike share programs nationwide
  - NYC- 76% of riders are male
  - Chicago - 79% of riders are male
  - Boston- 75% of riders are male
- SF Ford GoBike 79% of riders are male
- Do men and women who use the bike share system interact with it in different ways?
  - Can machine learning techniques be used to classify a ride as one taken by a male or female?

Source: https://www.buzzfeed.com/jsvine/these-maps-show-a-massive-gender-gap-in-bicycle-riding?utm_term=.sfpqABX5ob#.ghrBWp9kPM

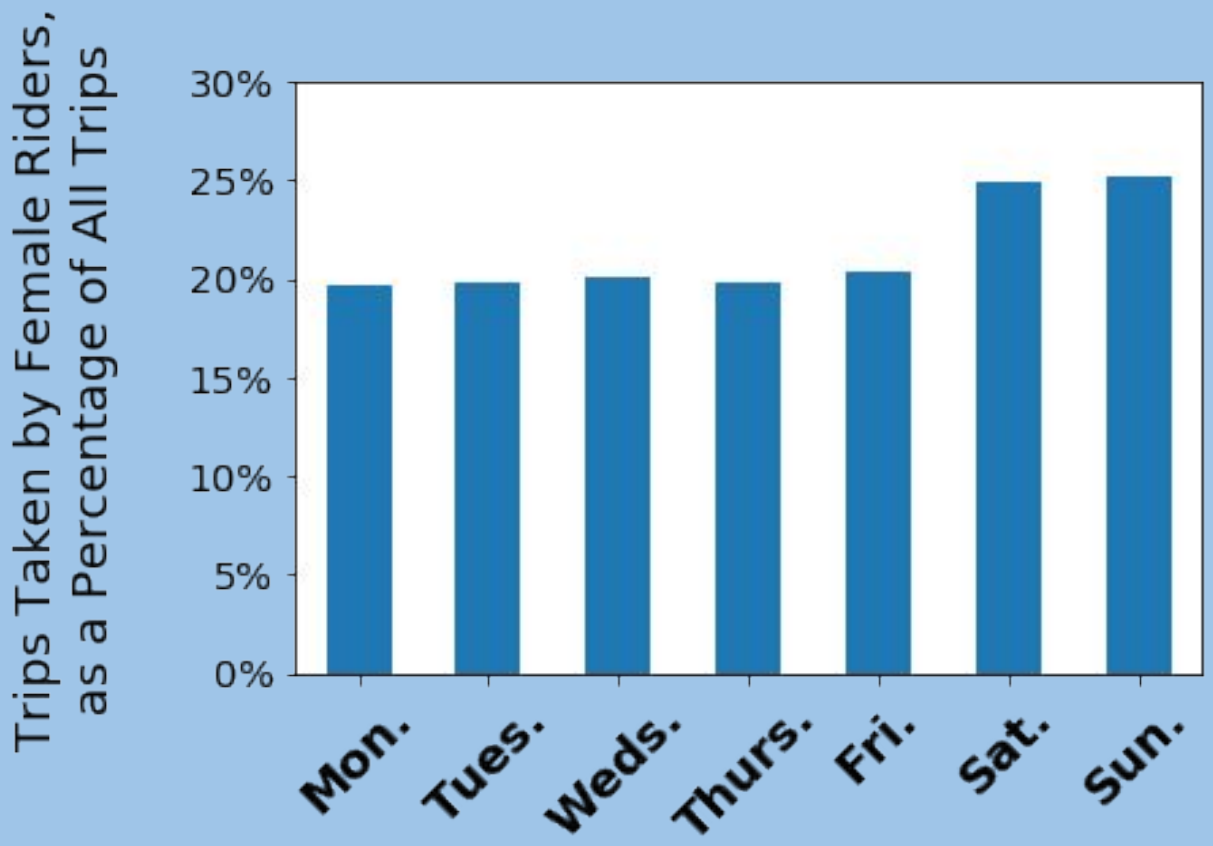
Data

- Publicly available data from Ford GoBike June - December 2017
  - ~350,000 rides
  - Starting station, ending station, start time, end time
  - Member year of birth
  - Member gender
- NOAA Weather data
  - Hourly temperature
  - Sunrise and Sunset
  - Hourly wind speed
  - Hourly precipitation
- SF Neighborhood Shapefile

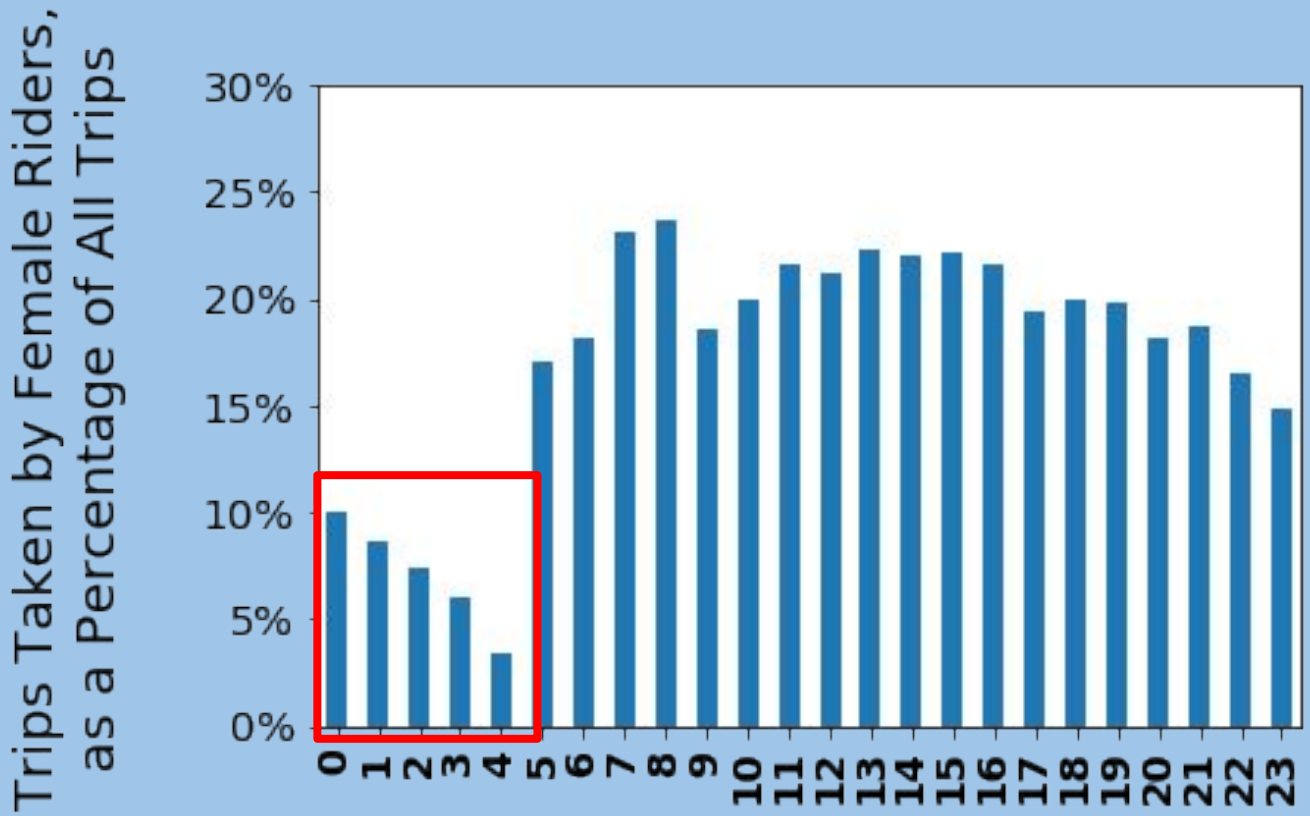
Ridership on Ford GoBike



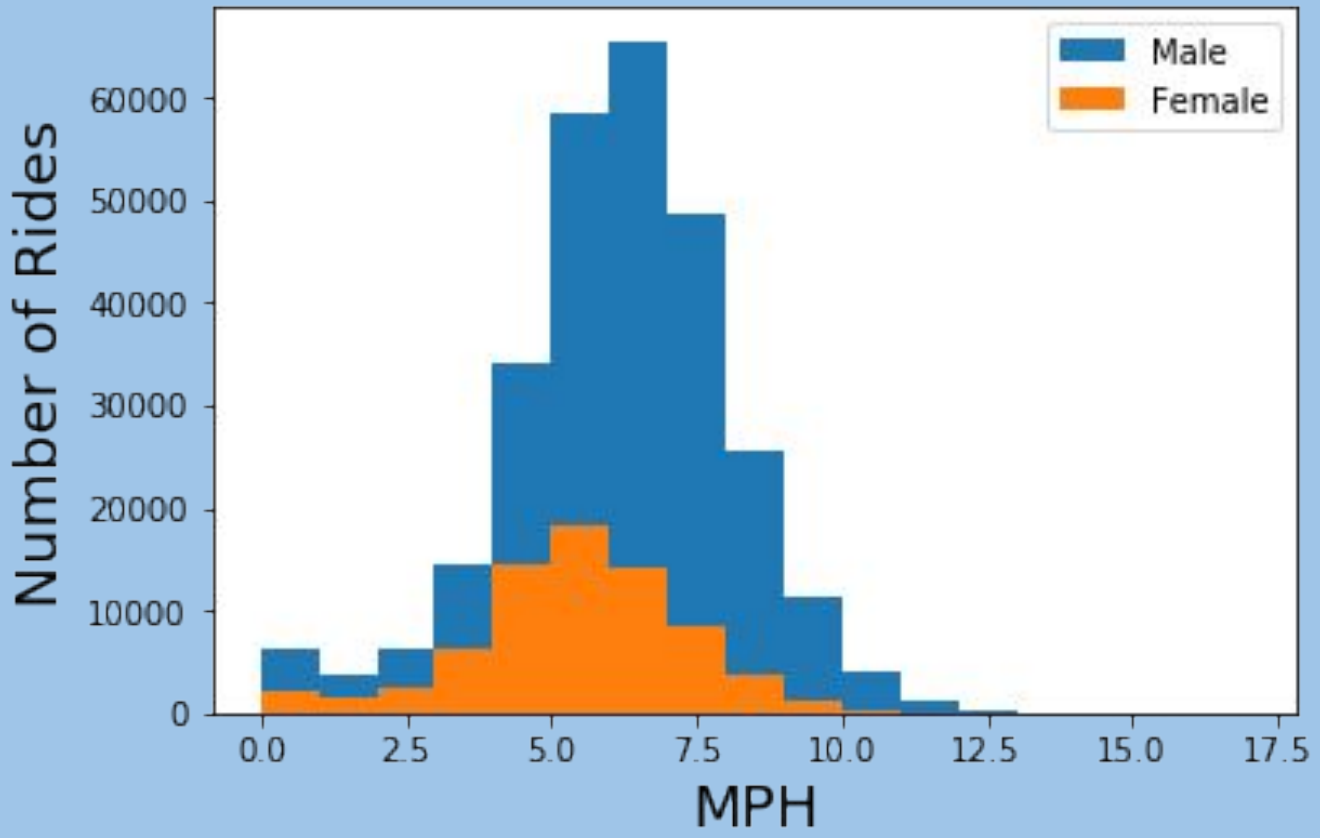
Female Ridership on Ford GoBike by Day of Week



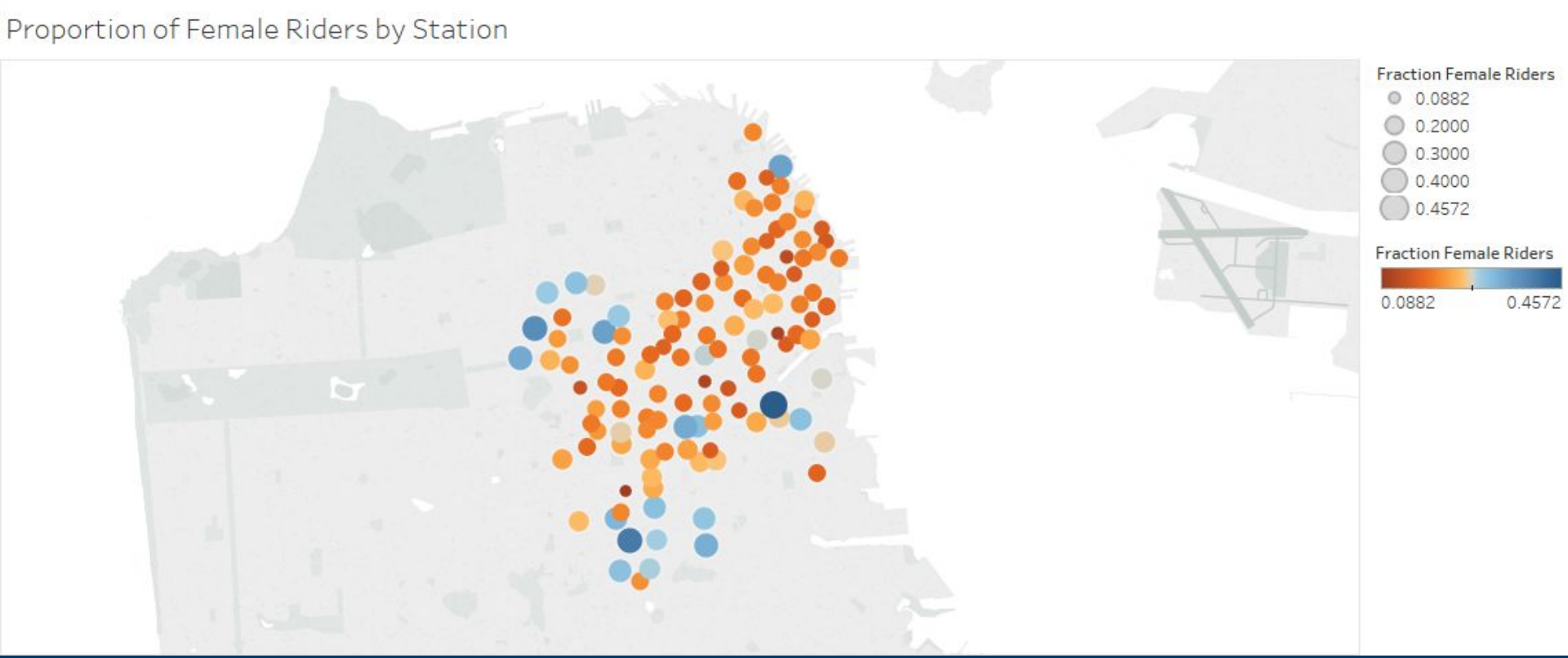
Female Ridership on Ford GoBike by Hour of Day



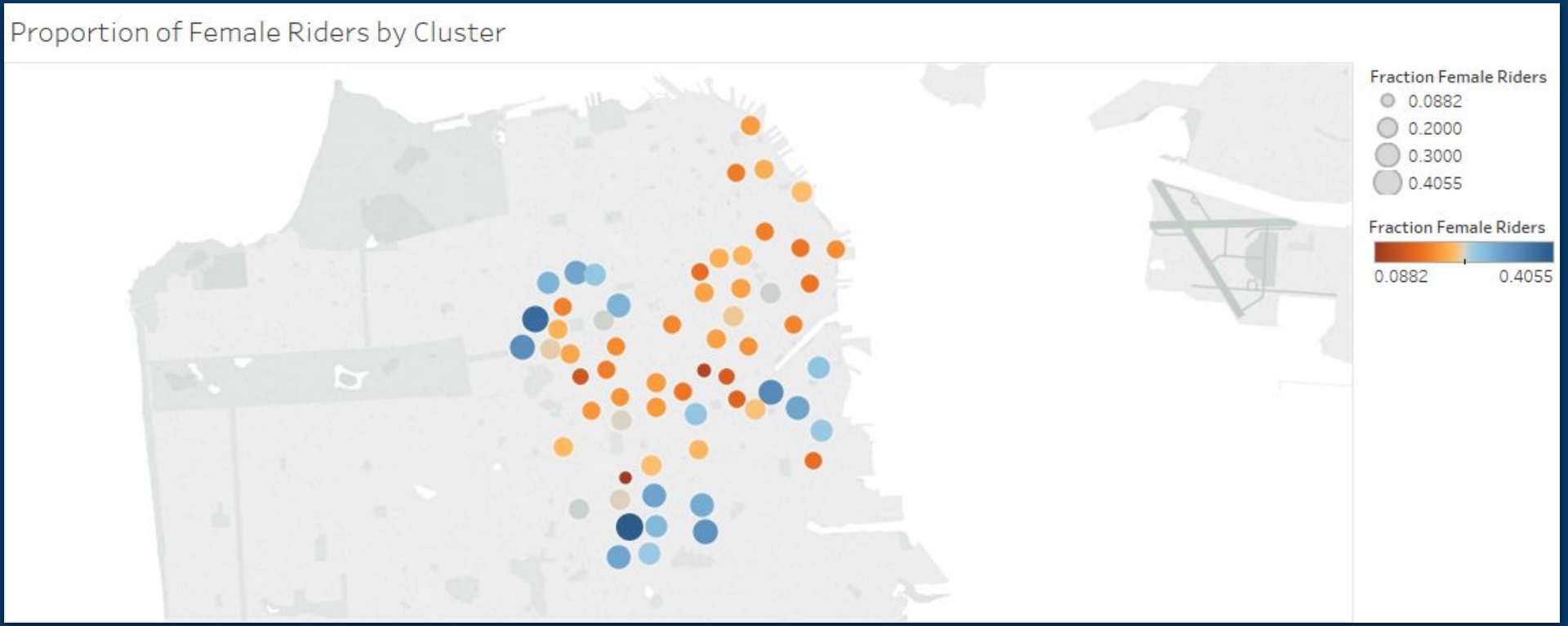
Speed of Male and Female Rides



Proportion of Female Riders by Station



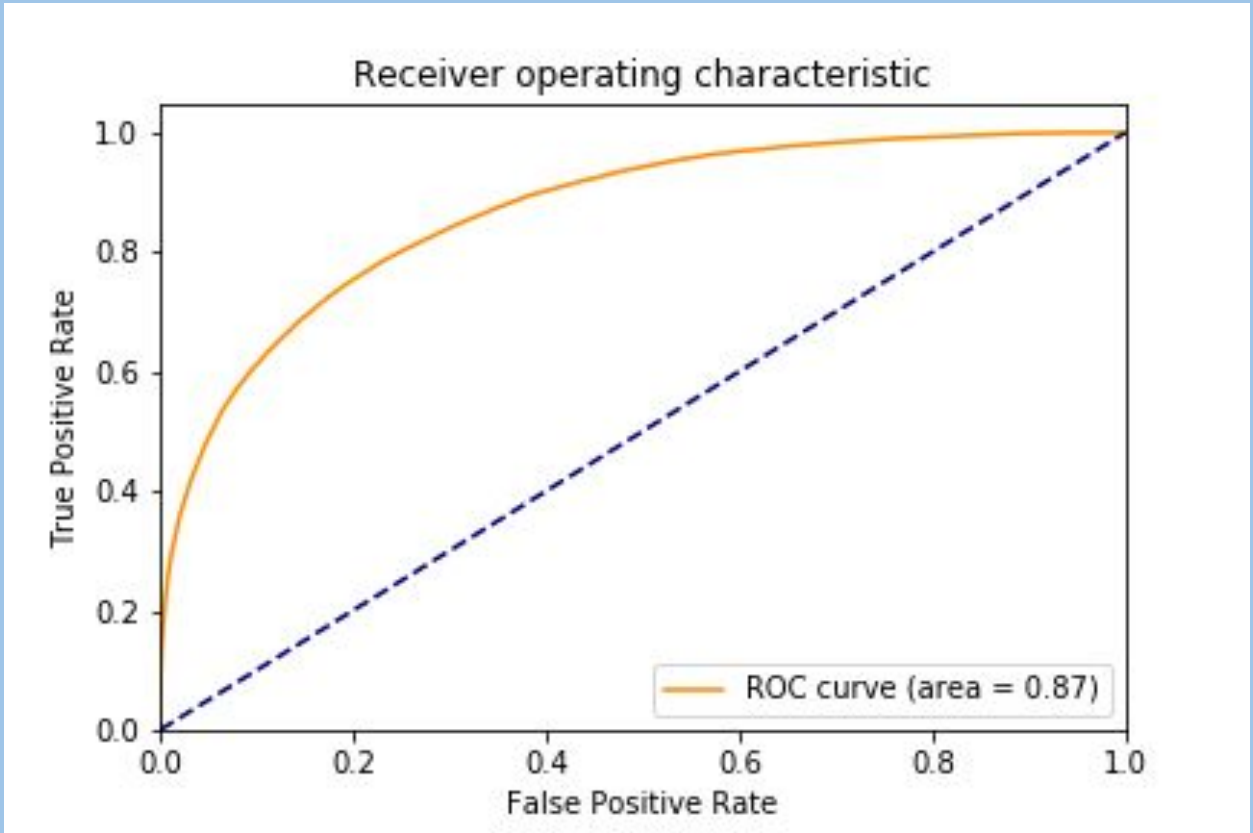
Proportion of Female Riders by Station Clusters



Models

- Attempted several different models (e.g. Logistic, KNN, SVM, Naive Bayes)
- Only Random Forest Model Outperforms a simple ‘Always Male’ Prediction
- Final Model is Random Forest with 250 Trees
- Percentage of riders who are male ~79.3%
- Accuracy of the Random Forest model ~84%

Model Results



Next Steps

- Additional Models: Train Neural Network
- Other Data Cleaning: alternate clustering techniques, feature reduction techniques
- Outside Data:
  - Leverage data from other bike share networks
  - Look into MUNI Delay Data

Thanks!

Dan Weiss

Slide Link:

https://github.com/DanWeiss1/Ford-Go-Bike-Gender-Prediction/blob/master/Slides.pdf