Capital Bike Share Prediction Analysis

DANIEL NASON

PERSPECTIVES IN DATA SCIENCE/PROFESSIONAL SKILLS FOR STATISTICIANS

Agenda

- Item
  - Executive Summary
  - Introduction/Background
  - Data/Processing
  - Exploration
  - Model Development
  - Performance Evaluation
  - Maintenance Plan
  - Discussion
  - Appendix
- Page Number

Executive Summary

- Purpose: Capitol Bikeshare wants to develop a maintainable model that predicts bike availability at their stations for a given time
- Data: 2018 and 2019 bikeshare data was collected, processed, and combined with other variables (weather, station information) to generate a dataset to predict arrivals/departures
- Methods: Trained multiple machine learning algorithms and found that XGBoost performs best on out-of-sample data
- Results: On average, the model's prediction is off by approximately 2.4 bikes for arrivals and departures
- Discussion: Reasonable extensions to improve model performance and data limitations require additional work to deploy model to production

Introduction

- Project: Develop a model for Capital Bikeshare to predict bike availability for their stations
  - Deploy model to inform staff on when/how many bikes to relocate to stations
- Deliverables:
  - Predictive model
  - Presentation of findings and utility
  - Strategy for model maintenance

Background

- Bike-share: affordable short-trip transportation alternative
- Client: 3rd-generation bike-share company with scaled services in limited major metropolitan markets
- Goal:
  - Lower operational costs to generate profit for client
  - Improve customer experience by using predictive model to tell client when and where to relocate bikes to meet daily demand

Data Overview

Station Information

- Source: Bike Station Dataset
- Variables:
  - Station
  - Geocoordinates
  - Capacity

Bikeshare Data

- Source: Capitol Bikeshare Dataset
- Variables:
  - Station
  - Trip ID/Bike ID
  - Timestamp

Weather Data

- Source: NCEI (NOAA) Dataset
- Variables:
  - Temperature
  - Precipitation
  - Timestamp

Assumptions:

Contractors 'reshuffle' bikes (bike changes station without being ridden)
Station capacity identical between our data
Bikes are not stolen/broken

Data Limitations

- Time discrepancy: Data not during similar timeframe
  - Bike data at hourly level, weather data at daily level
  - Bike data from 2018-2019, Station data from 2021
- Missing data: Bikes are randomly relocated between station*
- Formatting: Data reporting fundamentally changes between March 2020 and May 2020

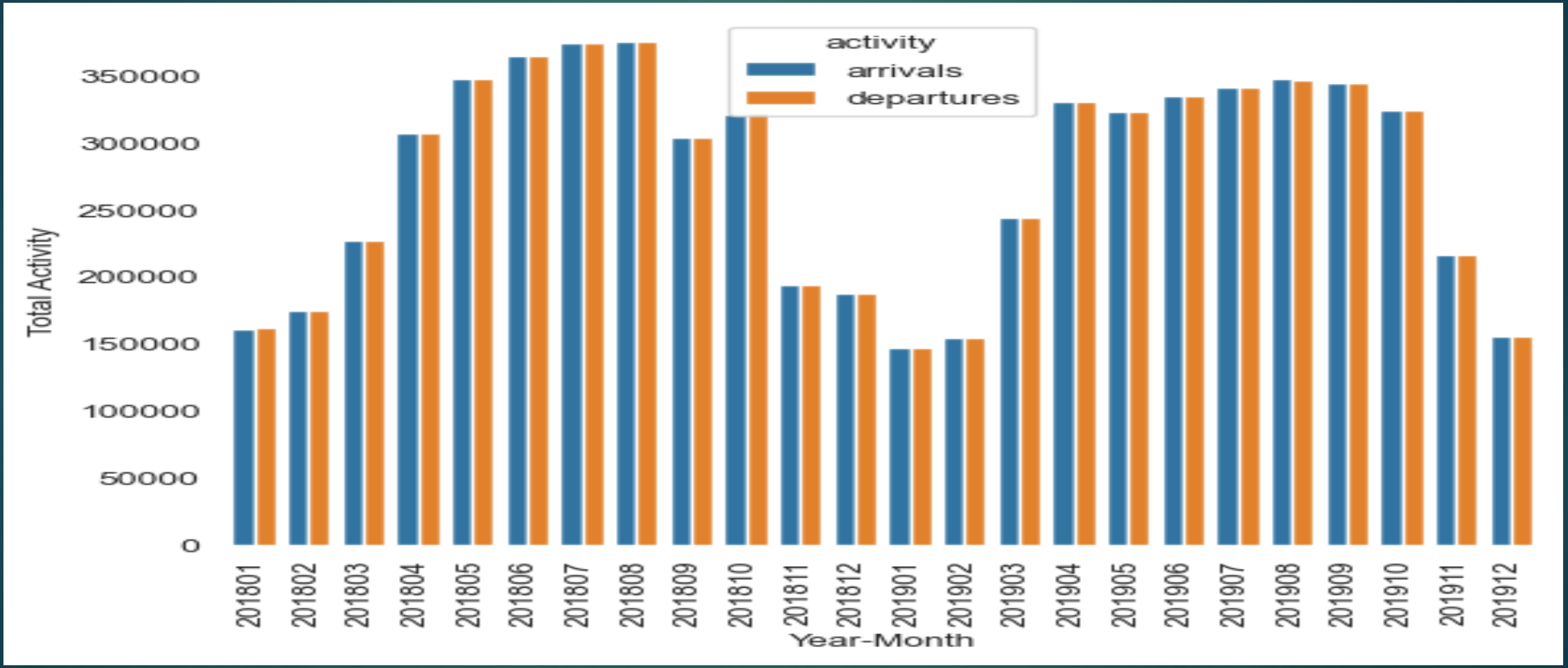
*See Appendix for details

Outcome of Interest: Arrivals and Departures

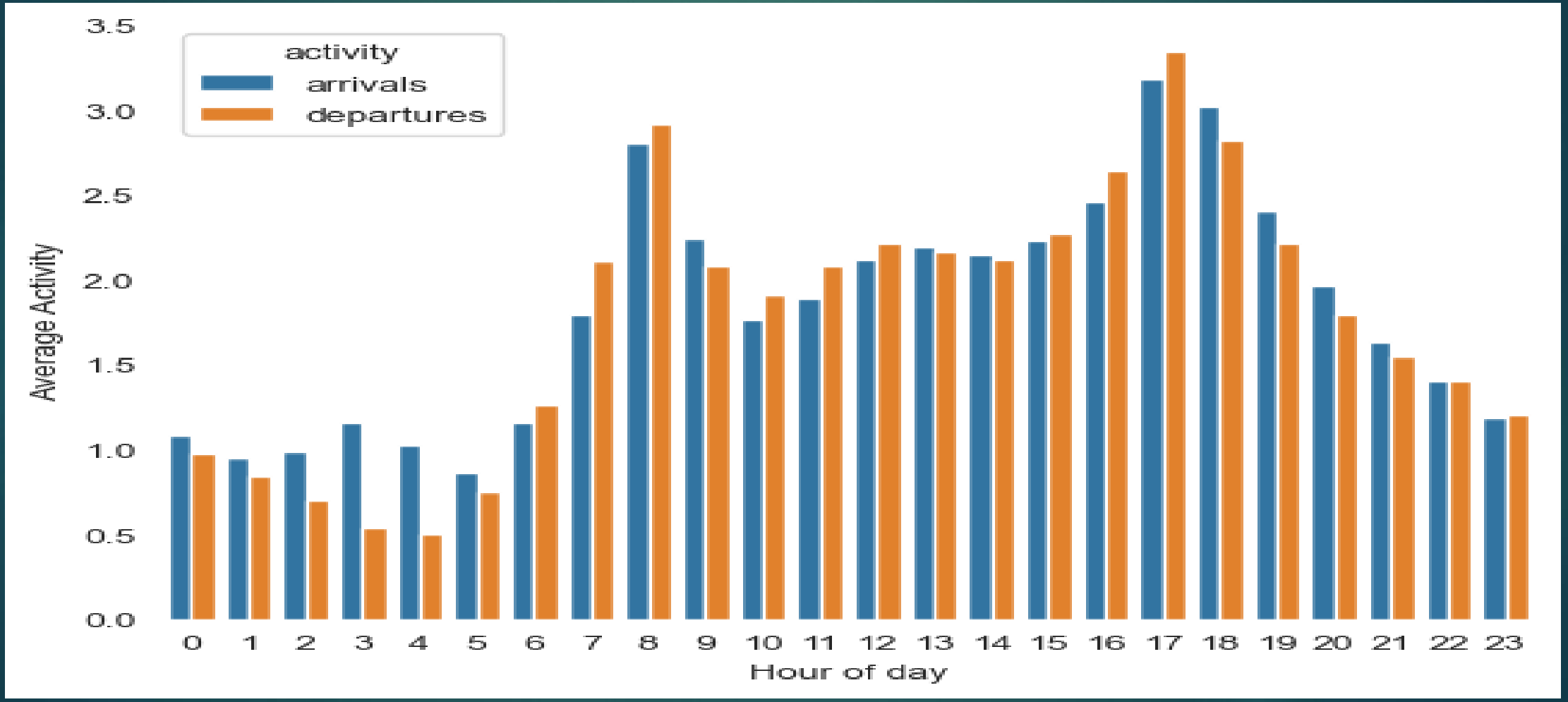
$$Availability = \frac{Currently\ Available + Net\ Departures\ and\ Arrivals + Net\ Reshuffling}{Station\ Capacity}$$

- Targets: Departures and Arrivals at each station at each hour

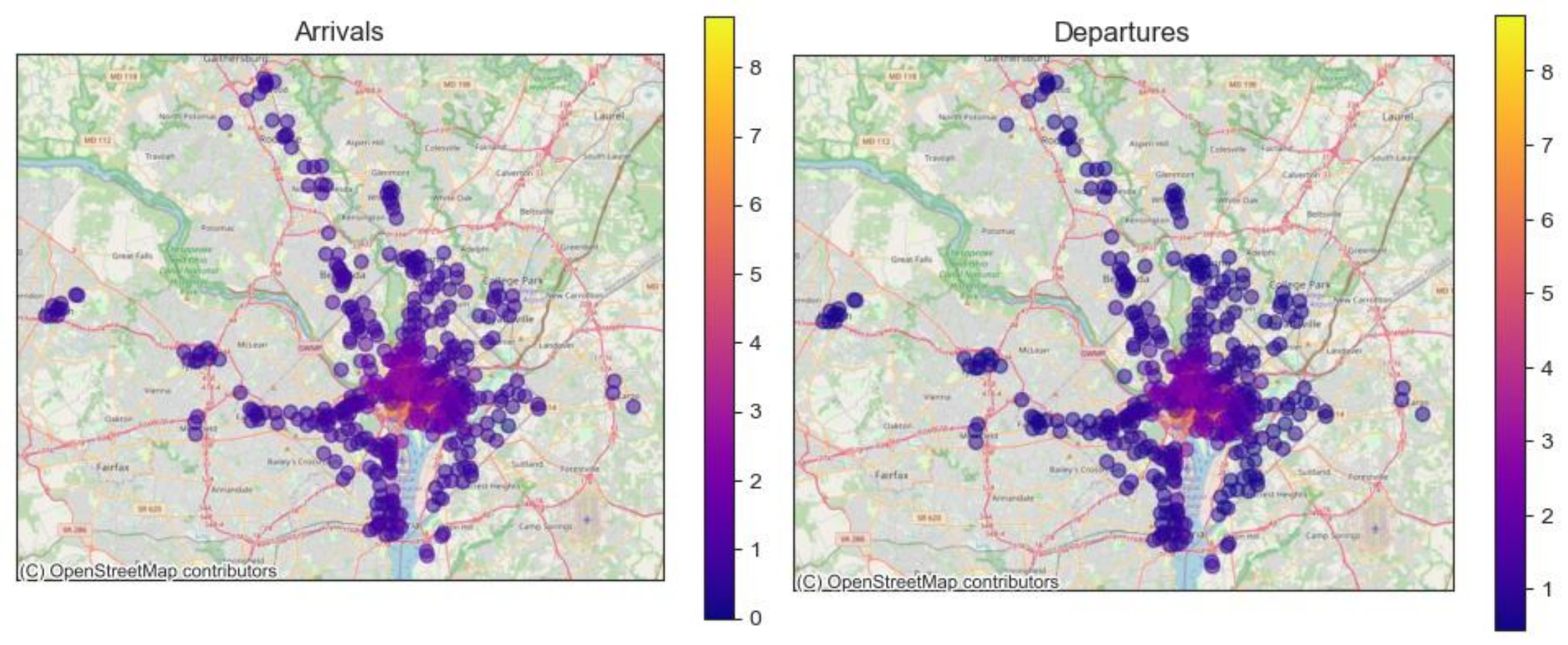
Bikeshare activity peaks during the warmer seasons of the year.



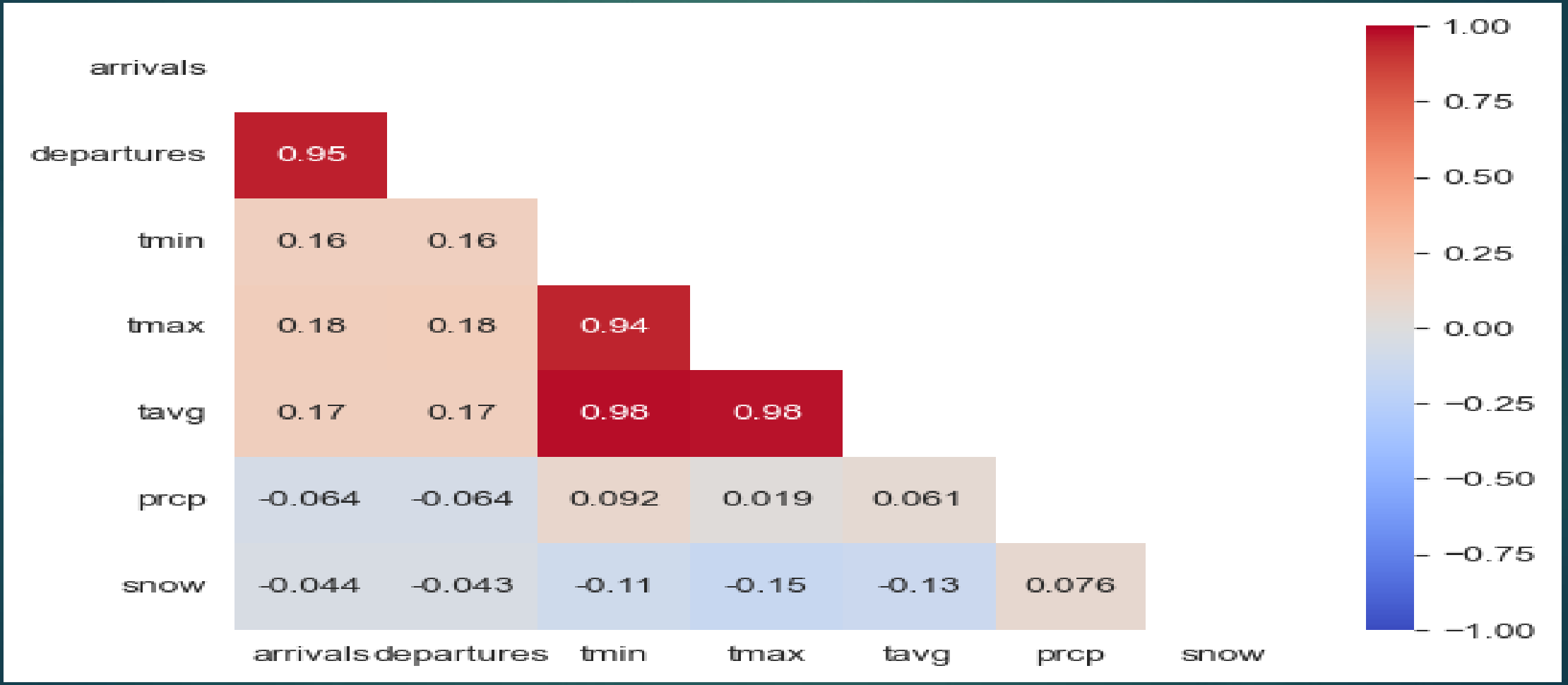
Daily bikeshare activity peaks during commuting rush hours.



Highest bikeshare activity is concentrated in downtown D.C.



Outcome variables only strongly correlated with each other.



Modeling Approach Requirements

- Non-linear relationships exist in the data
- Prediction is most important, but interpretability would still be useful
- Scalable, requires model that can handle moderately sized data
- Focus only on stations with sufficient sample size for each month across the entire dataset

Model Development

- Feature engineering: date/time features, location, weather data
- Performance evaluation:
  - Select features and tune hyperparameters on 2018 data, using predictive modeling best practices (hyperparameter tuning, cross-validation)
  - Evaluate model performance on 2019 data unseen by the model, using root mean squared error to evaluate performance*
    - Larger penalty on bad predictions

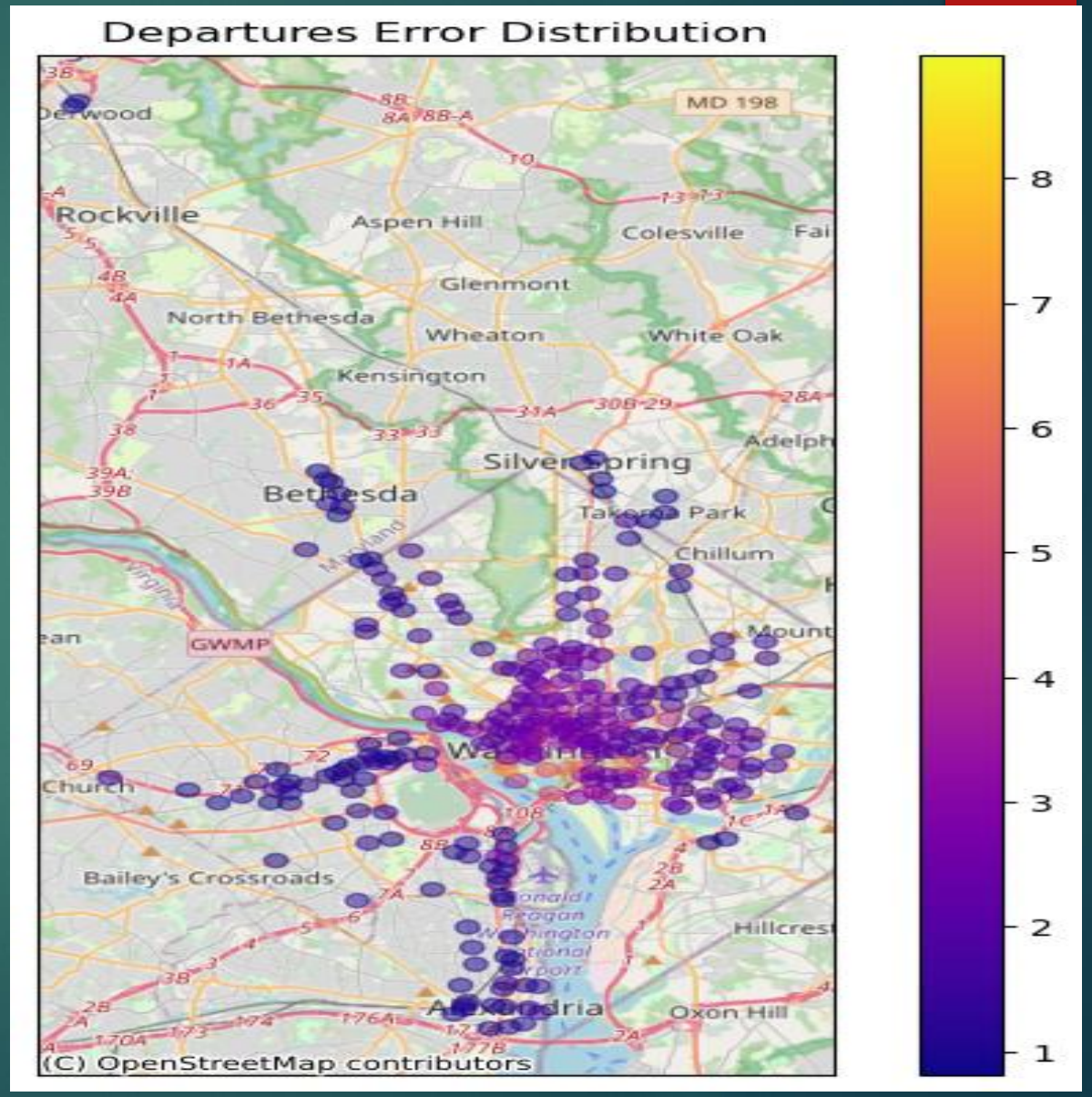
*See Appendix for details

XGBoost Performs Best for Arrivals and Departures

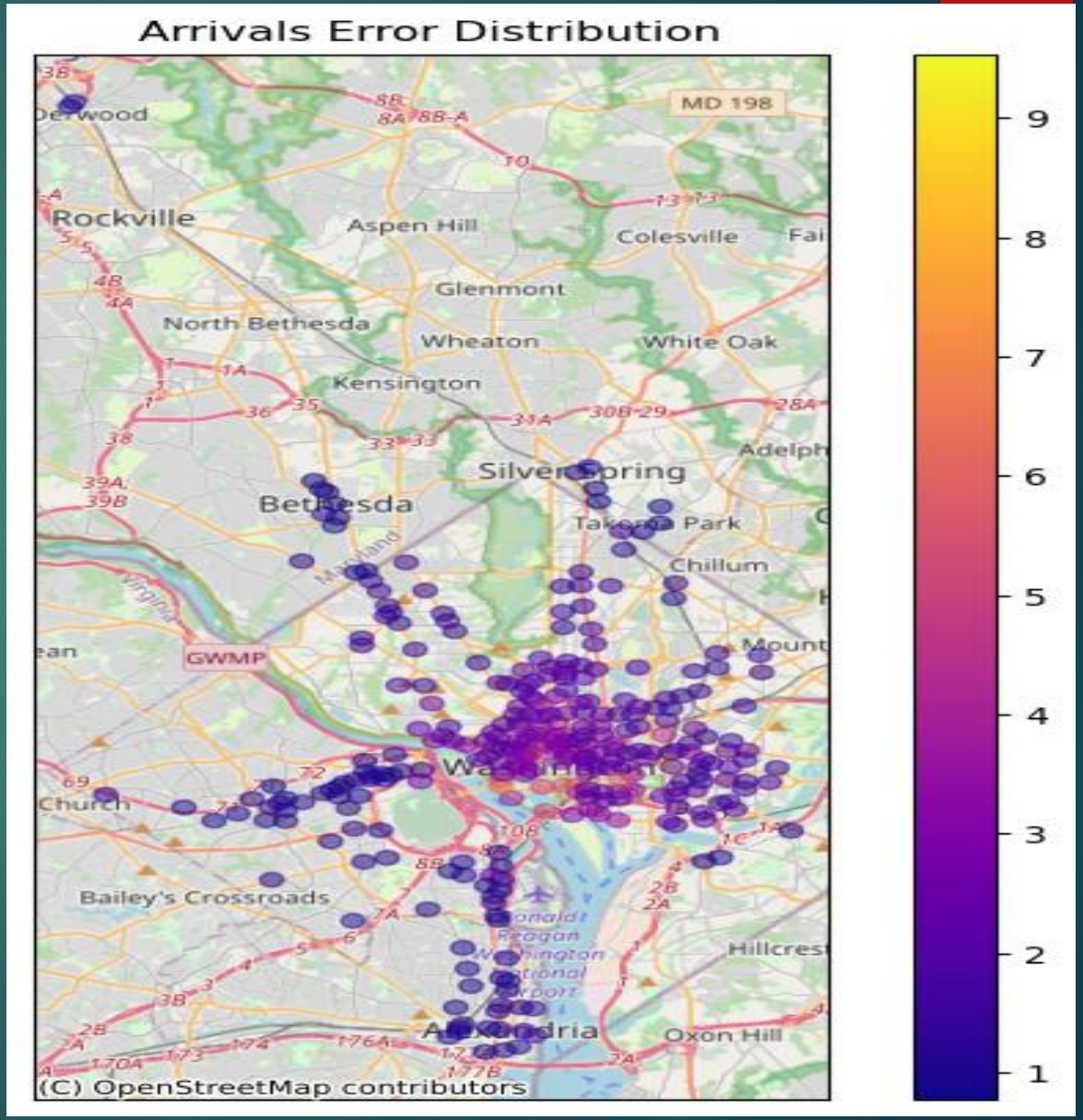
- Models evaluated on identical, unseen training data to evaluate performance
- XGBoost minimizes error for both arrivals and departures, with an RMSE of 2.4
- Interpretation: The model's prediction in each case misses the expected arrivals and departures by 2.4 bikes

Model	Arrivals 2019 RMSE	Departures 2019 RMSE
Poisson	2.8	2.8
Random Forest	2.8	3.1
Gradient Boosting	3.0	3.0
XGBoost	2.4	2.4

XGBoost departures model struggles the most with downtown Washington DC.



XGBoost arrivals model also has larger relative errors in Washington DC area.



Analysis Limitations

- Extensive data cleaning/processing and simplifying assumptions to model the data
- Operational Protocols needed to be accounted for in the data
- Pre-COVID-19 data may not be valid today
  - Changes in consumer behavior could lead to few trips
  - Fundamental reporting shifts by Capitol Bikeshare would require re-engineering of data pipeline to accommodate current data structure

Next Steps

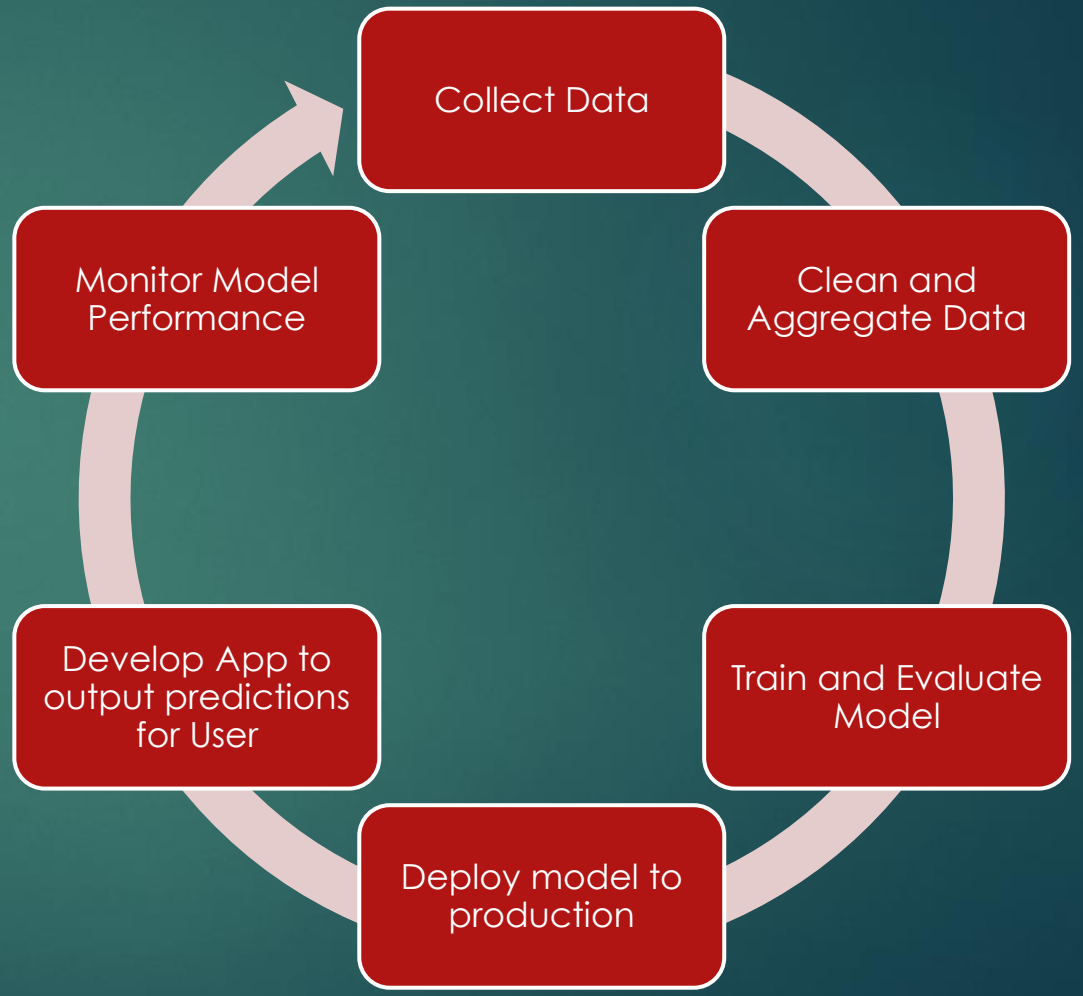
- Extensive data cleaning/processing and simplifying assumptions to model the data
- Operational Protocols needed to be accounted for in the data
- Pre-COVID-19 data may not be valid today
  - Changes in consumer behavior could lead to few trips
  - Fundamental reporting shifts by Capitol Bikeshare would require re-engineering of data pipeline to accommodate current data structure

Extending the Project

- Data:
  - Update pipeline to account for new data format
  - Metro and bus locations, university semester schedules
  - Other indicators: sports, political activities, etc.
- Modeling: include alternate approaches to evaluate performance of other modeling types (classification, time-series, etc.)
- Develop algorithm to recommend reshuffling by contractors and deploy app to production

App Deployment and Model Maintenance

- Model predictions improve operational efficiency by deploying contractors to reshuffle bikes
- Mobile app allows seamless interaction between workers and data
- Can redeploy bikes to stations with most demand
- Opportunity: extend app to users to let them choose a station with a bike nearest to them



Thank you! Questions?

Appendix

Reshuffling: Before Processing

Trip Start	Trip End	Start Station	End Station	Bike ID
1/30/2019 11:49	1/30/2019 12:00	24th & N St NW	11th & M St NW	W23345
1/30/2019 12:10	1/30/2019 12:18	11th & M St NW	7th & F St NW / National Portrait Gallery	W23345
1/31/2019 6:51	1/31/2019 6:58	New Jersey Ave & N St NW / Dunbar HS	8th & H St NW	W23345

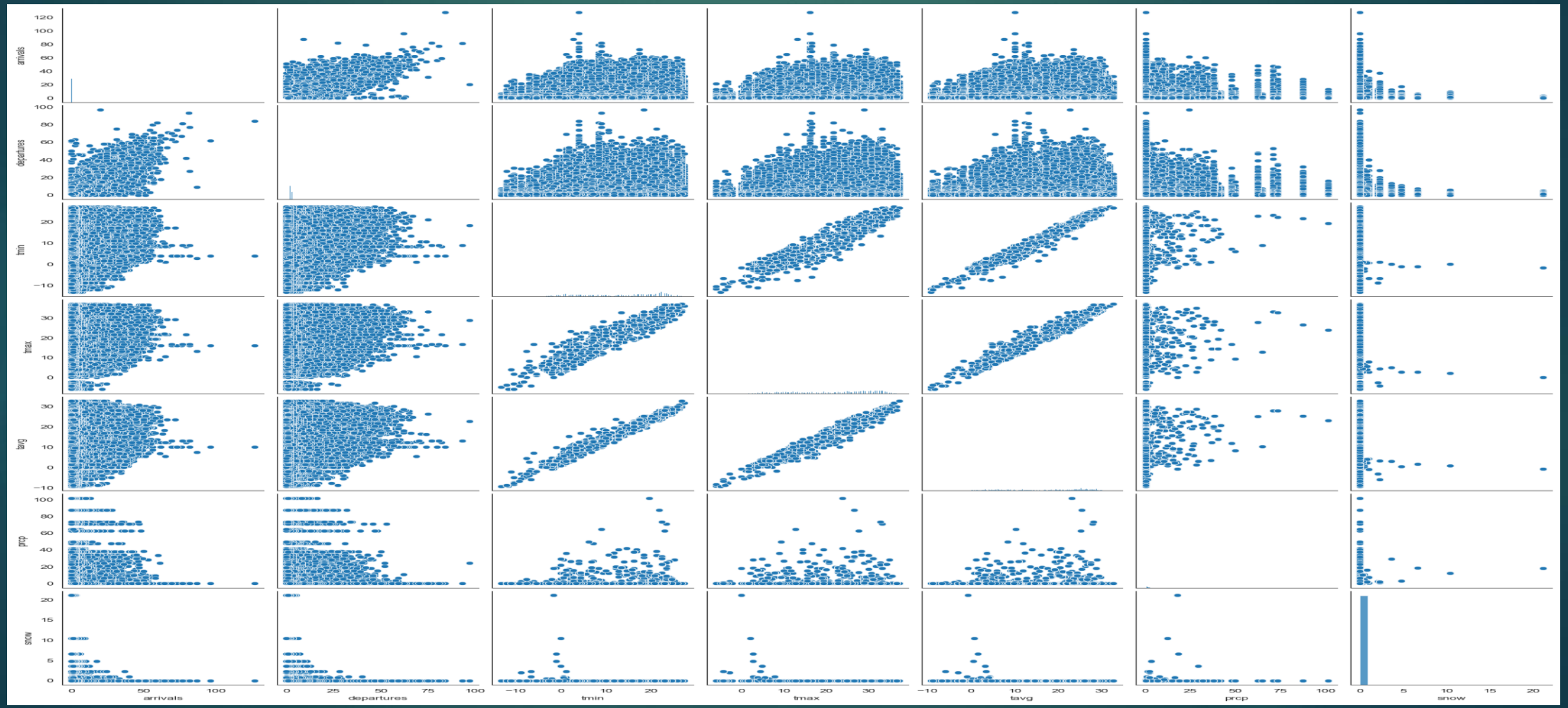
- Bikes are arbitrarily 'reshuffled' in the data, resulting in missing trip information
- Need to account for this behavior to get an accurate count of bikes at each station over time

Reshuffling: After Processing

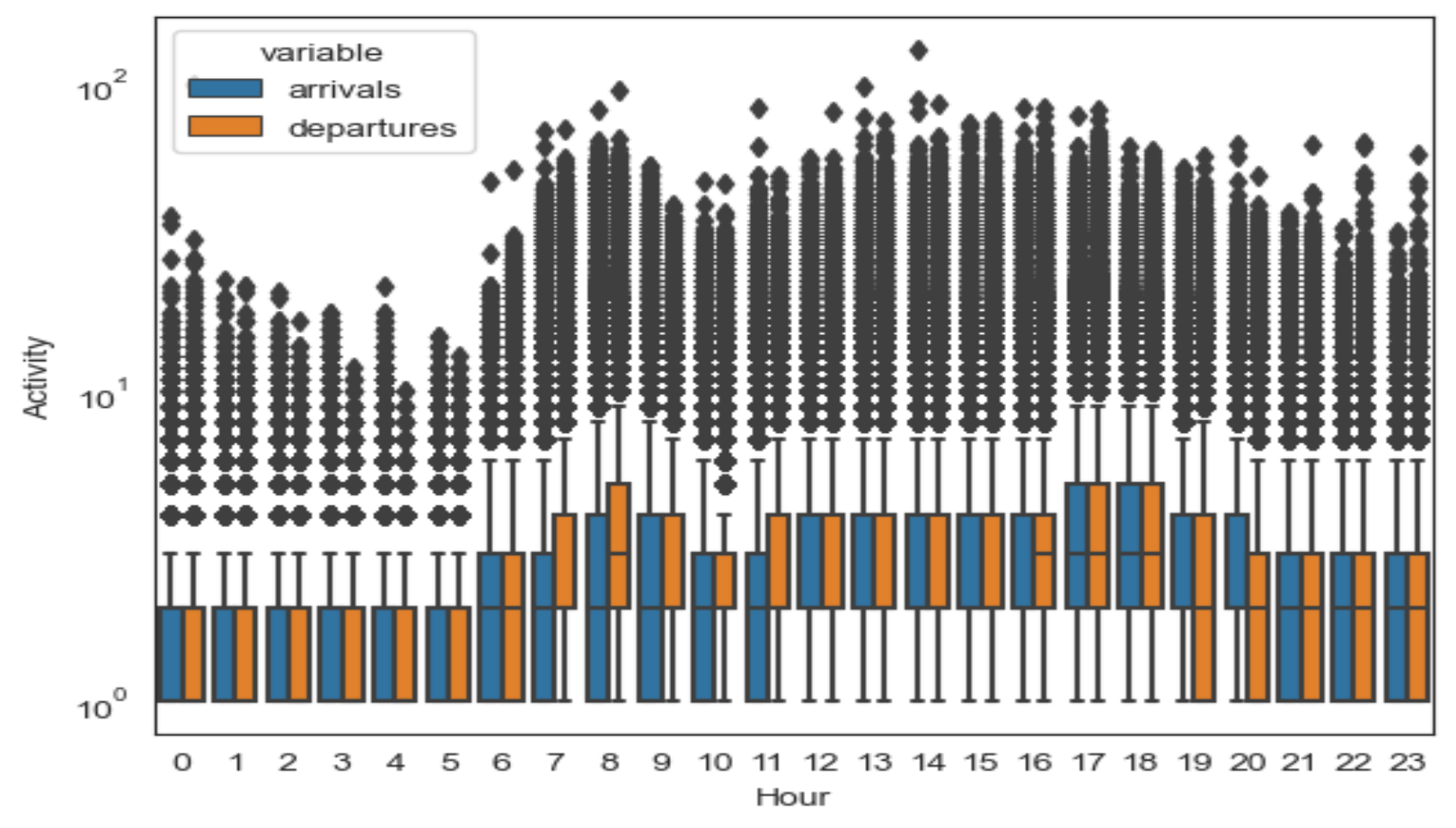
Bike ID	Time	Station	Event	Reshuffled
W23345	1/30/2019 11:49	24th & N St NW	Departure	No
W23345	1/30/2019 12:00	11th & M St NW	Arrival	No
W23345	1/30/2019 12:10	11th & M St NW	Departure	No
W23345	1/30/2019 12:18	7th & F St NW / National Portrait Gallery	Arrival	No
W23345	1/30/2019 6:18	7th & F St NW / National Portrait Gallery	Departure	Yes
W23345	1/31/2019 0:18	New Jersey Ave & N St NW / Dunbar HS	Arrival	Yes
W23345	1/31/2019 6:51	New Jersey Ave & N St NW / Dunbar HS	Departure	No

- Processing changes data format from 'wide' to 'long' for ease of analysis
- More easily account for missing information by adding rows

Outcome variables not strongly linearly correlated with features.

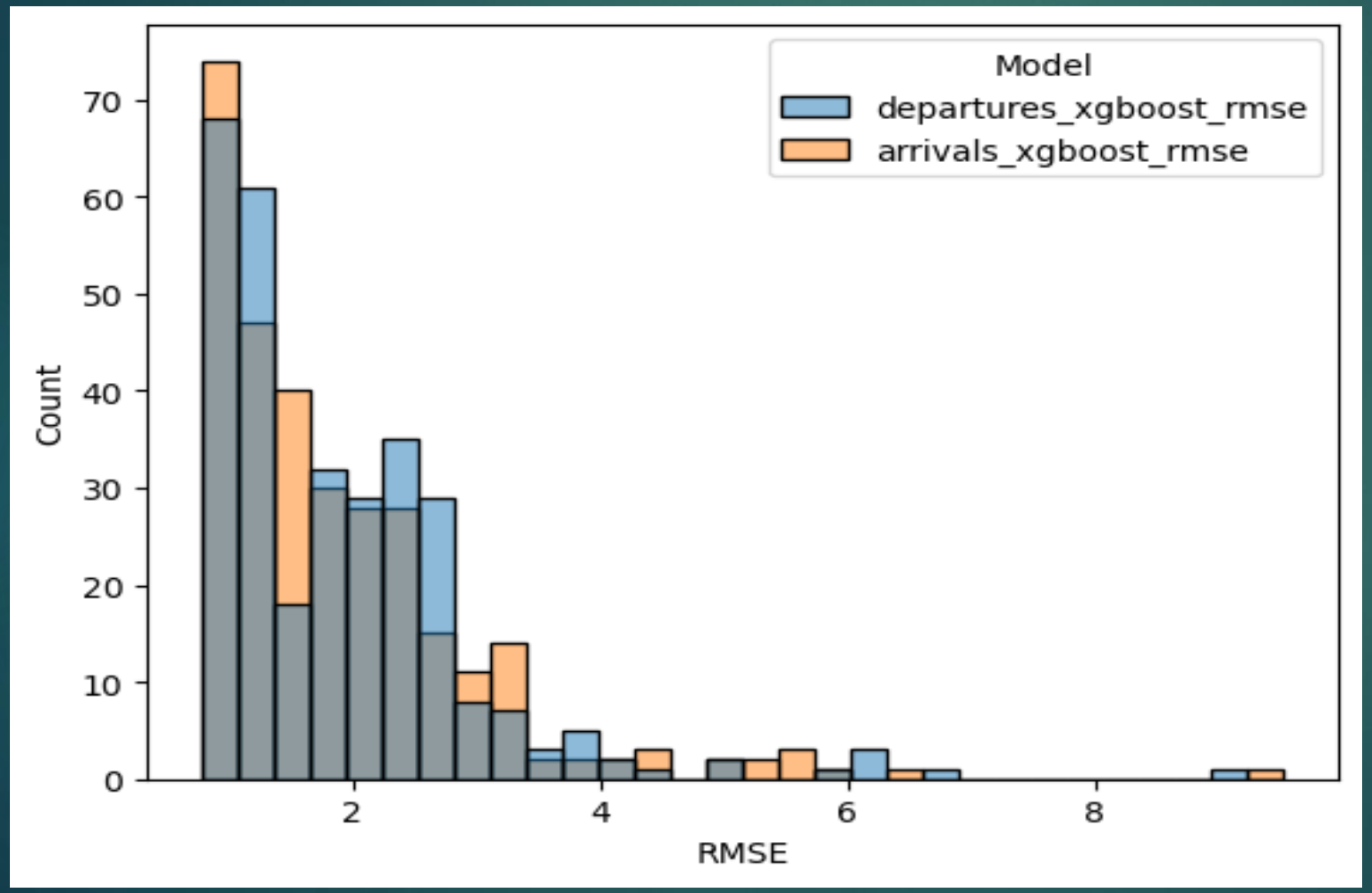


Hourly activity is similarly distributed for arrivals and departures by hour.



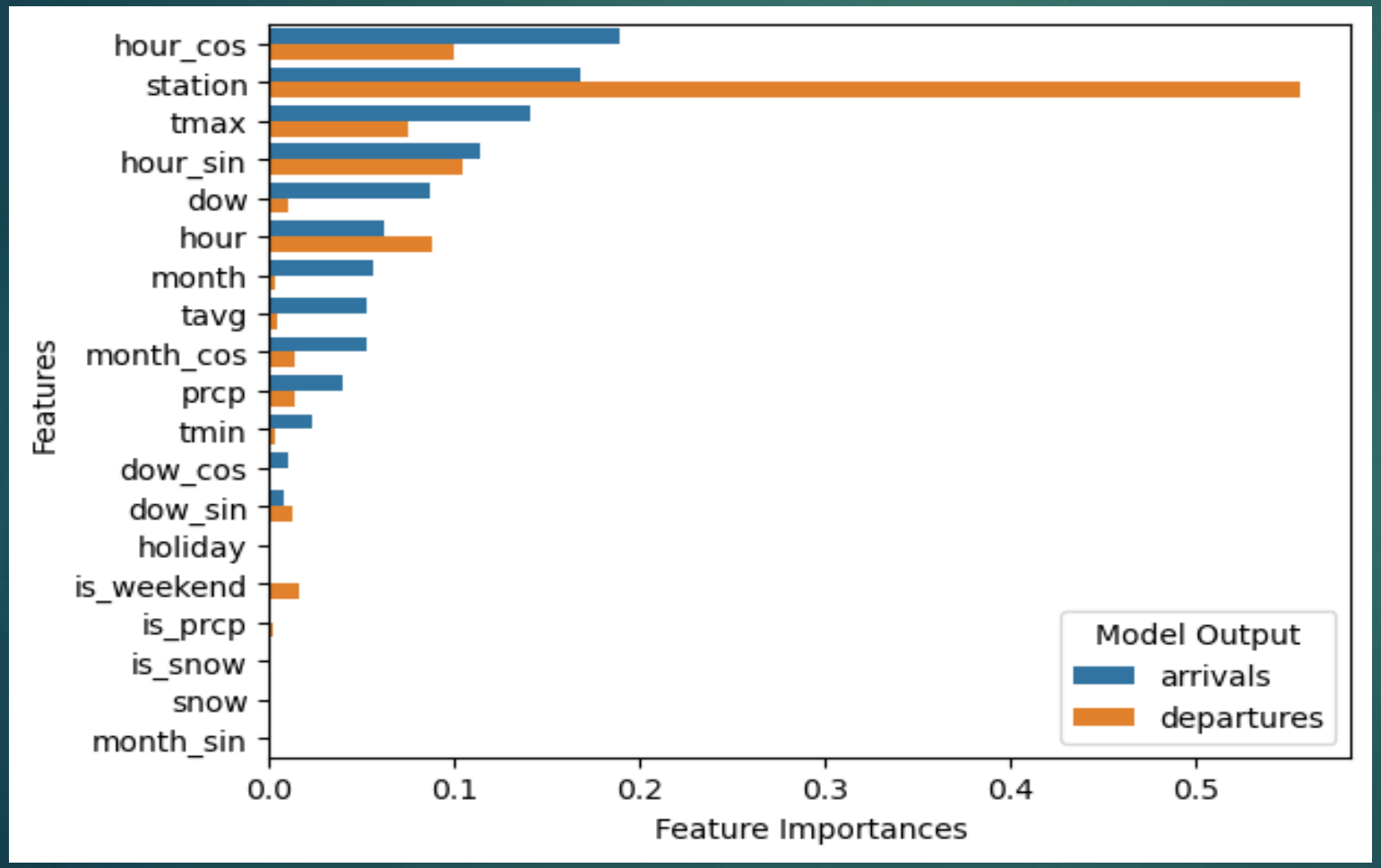
- Majority of hourly activity by hour is under 10 bikes
- Outliers likely largely driven by larger stations with greater bike capacity

Errors roughly identically distributed across stations for both models.



- Majority of RMSEs below 2 suggest that the models are fairly accurate for many stations
- Some larger RMSEs at or above 6 require further investigation to see whether outliers are driving these larger errors

Station, weather, and time of day features most important for predictions.



- Arrivals model more balanced in assigning importance to features, with time and station being roughly equivalent
- Departures places far greater emphasis on the station than any other variable
- Interestingly, holidays and snow do not seem to be useful predictors for either model