BLUEbikes SM

Station Optimization

Analysis by Stefanie Porcaro

The Goal:

To evaluate how well each station is stocked with bikes and predict if there will be available bikes for riders at each station at the beginning of each day

Data Information

Data acquired for the period: January 2018 - November 2020

Areas included: Boston, Cambridge, Somerville, Brookline, Everett, Arlington, Watertown, Newton, and Revere

Number of Records: 6,290,023

Files format: csv

Data Cleaning + Aggregation

Created 3 aggregated dataframes to merge:

1. Station information - station name, station distance to center of the city, total docks, district
2. Percentages of trip features for each day of the week at each station
3. Net bike percentages for each day at each station

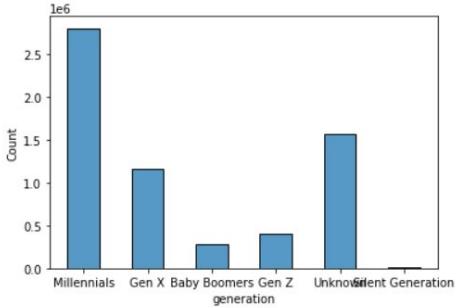
Features:

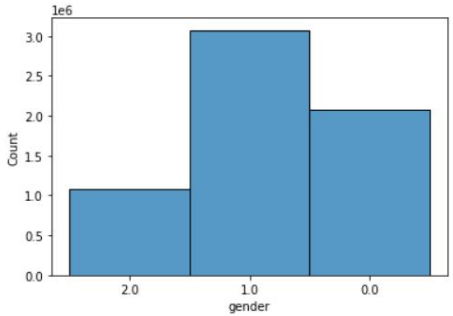
Day of Week
Month
Station Name
Station Distance to Center of City
Station District
Customer/Subscriber Percentage
Gender Percentage
Birth Year → Generation Percentage

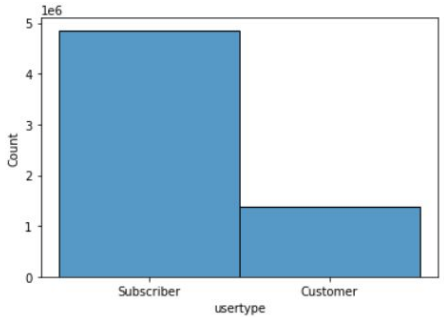
Response Variable:

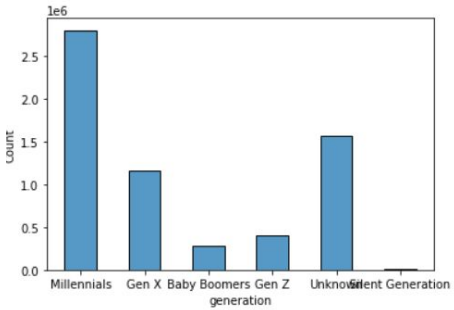
1. Net Bike Percentage at End of Day
2. Binary Variable indicating if Bikes are available at EOD or not

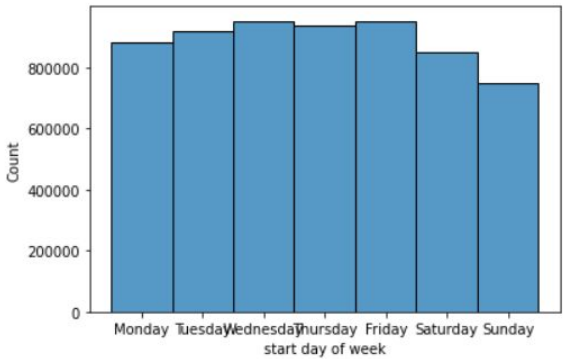
Exploratory Data Analysis

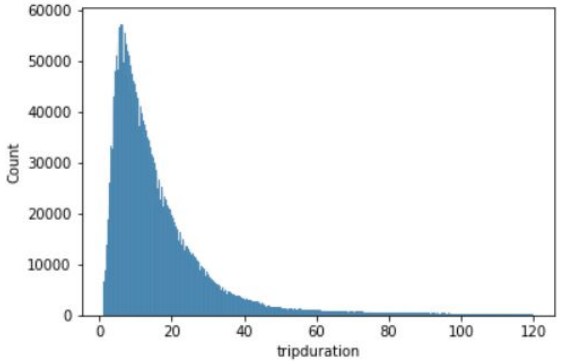


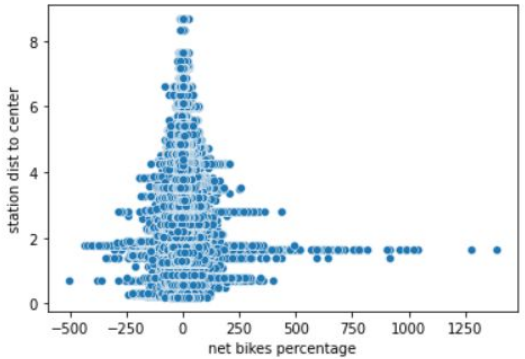












Preprocessing:

One-hot Encoding

Standard Scaler

Principal Component Analysis

Clustermap

Train/Test Data Split - 80%/20%

Findings:

Upon examining the covariance of the data's features, some relationship between gender and generation existed. Specifically, unknown gender varied with the millennial variable, and unknown generation varied with the male variable.

Through PCA, 348 columns were cut to 195 and still preserve 95% of the variance caused by each of these features.

To prepare the data for modeling, 80% of the data was used to train the model and the 20% leftover was used to test it.

Supervised Regression Models

Response variable: Percentage of net bikes at end of each weekday per each station

	Accuracy	MAE	RMSE
Linear Regression	8.29%	20..67	33.19
Lasso Regression	8.29%	20.67	33.19

Supervised Binary Classification Models

Response variable: Binary, 0 or 1, indicating that there are or are not available bikes at each station at the end of the day

Decision Tree	Random Forest	Gradient Boosting	AdaBoost	knn
.567	.574	.603	.600	.565

Best Model = Random Forest

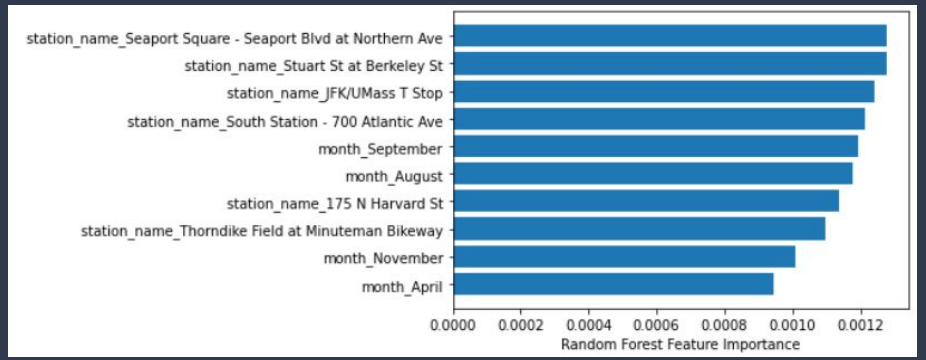
Accuracy Score: .602

	Precision	Recall	f1
Not enough bikes EOD	.63	.20	.30
Enough bikes EOD	.60	.91	.72

Hyperparameters:
n_estimators = 100
max_depth = 9
min_samples_split = 5
min_samples_leaf = 1

Feature Importance: →

1. Station: Seaport Square
2. Station: Stuart St at Berkeley St
3. Station: JFK/Umass T stop



Future Applications

- This model can be used to see if BlueBikes is evenly distributing bikes at each dock. It will let us know if bikes are available or not at each station at the end of the day.
- The model can easily be adjusted to determine bike availability at a particular hour throughout the day. This will determine where new stations may need to be built.
- The model would be best adjusted to a time series plot to measure bike-availability trends over seasons.
- This project would benefit from access to specific user IDs as well. The data could be used to dive deeper into customer segmentation and better predict rider patterns. The model could be used to predict how each area of the city is expanding.