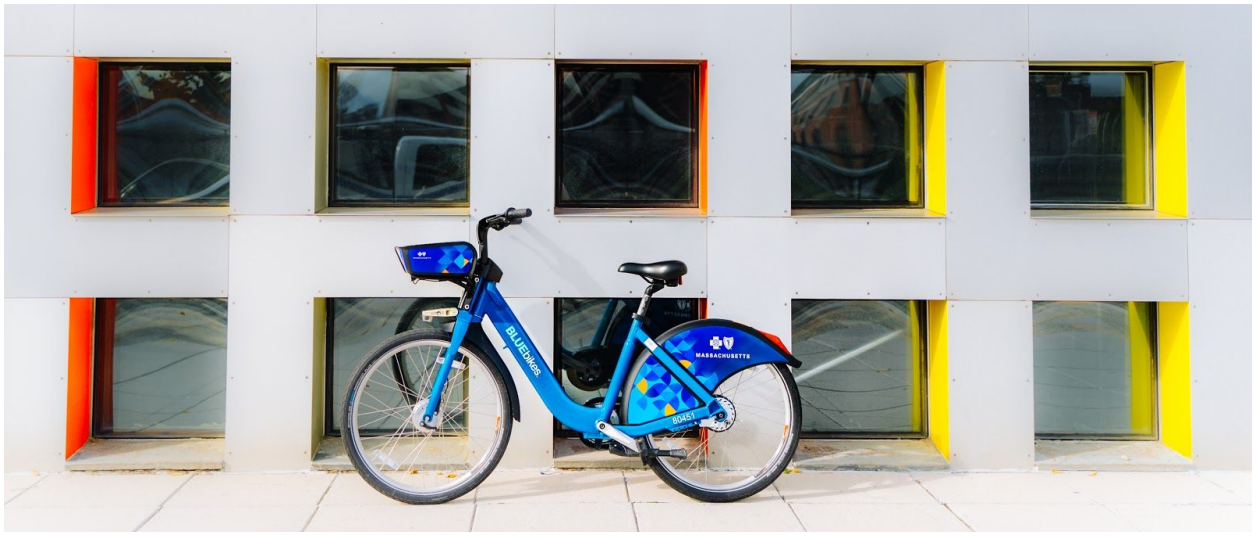
BlueBikes Station Optimization



BlueBikes: MetroBoston's BikeShare Program

BlueBikes is Boston's bicycle sharing system. Since its launch in 2011, BlueBikes has grown to include more than 3,750 bikes at 365 stations across Boston, Cambridge, Somerville, Brookline, Everett and now Newton, Arlington, Revere and Watertown. Over the last three years, in terms of the amount of trips taken, BlueBikes has grown at an average rate of 21%.

In this project, the goal is to evaluate how well each station is stocked with bikes and predict if there will be available bike(s) for riders at each station at the beginning of each day.

Data

The data for this project was gathered from the BlueBikes official website at https://www.bluebikes.com/system-data. Trip data was used from January 2018 through November 2020, gathered at https://s3.amazonaws.com/hubway-data/index.html along with specific station information from https://s3.amazonaws.com/hubway-data/index.html.

The trip data includes:

- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station Name & ID

- End Station Name & ID
- Bike ID
- User Type (Customer = Single Trip or Day Pass user; Subscriber = Annual or Monthly Member)
- Birth Year
- Gender, self-reported by member (0=unknown; 1=male; 2=female)

The station information data includes:

- Station Number ID
- Station Name
- Latitude
- Longitude
- District
- Total Docks at each station

Data Caveats:
Over the course of the three years of trips used in this project, stations were added and station dock numbers may have varied as well. The total frequency of trips at each station is not representative of the current popularity of each station. However, for this project, we just examined how busy a station is during its time of existence.

Data Cleaning and EDA

For gender data, unknowns were changed to 0 and unknowns for total docks number, 0's were deleted. For birth year, I deleted a couple of entries that were logged as before 1920 and I created a new categorical column binning ages by generation and creating an 'Unknown' label for nulls. For trip duration, the measure units were changed from seconds to minutes for better comprehension. A new column for day of the week was created to use for aggregations later.

For each of these features, I aggregated by count on station name and day of the week and then got percentages of each of these features.

I used latitude and longitude to create a column for the distance of each station to the city center which I placed in the financial district of downtown Boston.

To get a target variable of net bikes percentage per day, I subtracted a count of trips at each station on each day from the total docks, then divided by total docks. Later, I made this response variable binary, by calling rows 1 when the net bikes percentage was positive and 0 when it was negative. This would signify if there were or weren't bikes available at each dock on each day.

Preprocessing

For preprocessing, I used one hot encoding to handle categorical data such as generation and station district. I scaled the data using StandardScaler so the data would be ready for all types of models. I also employed Principal Component Analysis keeping 95% variance of all of these features to reduce the dimensionality of the independent variables. This reduced the number of columns from 348 to 195.

Furthermore, I saw from a clustermap of the features that most of the features were not highly correlated. However, there did exist some relationship between gender and generation. Specifically, unknown gender varied with the millennial variable, and unknown generation varied with the male variable.

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.

Modeling

With net bikes percentage as the continuous dependent variable, I tried using a Linear Regression to model the data. The accuracy score was very low at 4% and the root mean squared error was very high 45.68. After trying a Lasso Regression which minimized uninfluential features, the accuracy score did not improve significantly. These results show that this particular data cannot be modelled using linear regression.

Therefore, I turned the problem into a binary supervised classification prediction. I aimed to create a model that would tell us if bikes simply were or were not available at the end of each day. After testing Decision Tree, Random Forest, Gradient Boosting, AdaBoosting, and Support Vector Machine methods on a base level and considering runtime, Random Forest was found to be the best. I then tuned the hyperparameters to yield the best results from the model.

This particular data did not yield very high accuracy scores in the models that we used. After tuning, the best Random Forest model accuracy score was only 60.2%. With that mentioned, we can still interpret the precision and recall of the model.

If we take a positive *prediction* of “negative bikes end of day”, the probability that there were indeed a negative number of bikes at the end of the day is 63%. If we take a random true positive *example* of “negative bikes end of day”, the probability that it was predicted as such is 20%.

If we take a positive *prediction* of “positive bikes end of day”, the probability that there were indeed a positive number of bikes at the end of the day is 60%. If we take a random true positive *example* of “positive bikes end of day”, the probability that it was predicted as such is 91%.

Upon examining feature importances, it was evident that specific stations were the key features that influenced availability, as opposed to the station distance to the center or

rider attributes such as generation or gender. The top three stations that influenced the model were Seaport Square, Stuart Street at Berkeley Street, JFK/Umass T stop, and South Station. Other features that showed significance were months, specifically September, August and November.

Model use and Future Improvements

As new rider data comes in, this model can solve the business problem of if bikes are being allocated most efficiently based on the type of riders and their locations. It could help BlueBikes use its resources most efficiently and determine how many bikes should be in each dock at existing or new stations.

Something to note is that this model may best be used as a time series to account for the seasonality patterns of ridership. The data could be smoothed out over groups of three months to correspond to the seasons of Boston's weather.

This model can also be modified to be used to determine if bikes are available at each station in the greater Boston area at a particular hour of the day. For example, at high commuting hours. It also could be modified to determine bike availability during particular months of the year.

Additionally, if there were access to customer IDs, this data could be very useful in customer segmentation and predicting rider patterns and habits to see specifically how each area of the city is expanding.