Predicting Shared Bike Usage in San Francisco

Proposal for Springboard Capstone Project
J Slaga

Bike Share: Predicting Usage

SF Bay Area Bike Shares provides data on all trips taken on shared bikes as well as data on station' bike availability and capacity, and weather data for each day. We would analyse data to create a model predicting bike usage throughout the city.

Client:

Bike share companies can use this data in their planning for expansion throughout the city. The data would also be useful in their day to day operations for resetting bikes at various stations. Knowing where most bikes end up as well as when they are most needed, trucks that transport bikes back to stations could be more efficiently dispatched. Analysis of this data could also be useful to city planners in understanding bike routes and usage for expanding bike lanes across the city.

Data:

Data was found on Kaggle: https://www.kaggle.com/benhamner/sf-bay-area-bike-share. Original data is released regularly on http://www.bayareabikeshare.com/open-data. There are four files available: station.csv, status.csv, trip.csv, and weather.csv. We will be building a predictive model based on weather, trips, and station data.

Deliverables:

Code, report, and slidedeck.

Predicting Shared Bike Usage in San Francisco

Bike Share: Data Wrangling

Data Wrangling

The three datasets for this project are fairly clean. The weather dataset had the most missing data, primarily in the 'events' column (~85% missing). This was expected as this column recorded significant weather events such as storms. Missing data was filled with NAN. Max gust speeds also had a significant amount of missing data (~25%). As this column records gusts and regular wind patterns were recorded elsewhere, missing data was filled with NAN.

The trip dataset is very complete. The data was subsetted to just trips in San Francisco (a majority of the data). The main cleaning in this set was in removing outliers, specifically trips which seemed to be errors. Plotting the trip durations revealed many implausible trip times. Trips with durations over ten hours were removed. There were many trips with durations of less than five minutes. Further exploration revealed that many of their start and end stations were the same and presumably due to user error in checking out the bikes. Trips with the same start and end stations were removed.