ORIE 4741: Project Proposal

Divya Talesra (dt377), Michelle Chao (mc2244),
Ansh Godha (ag759), Pooja Gokhale (pg334)

October 2020

1 Project Proposal

Bike-Sharing Rides Datasets:
NYC Bike Sharing: https://www.kaggle.com/akkithetechie/new-york-city-bike-share-dataset

SF Bay Area Bike Share: https://www.kaggle.com/benhamner/sf-bay-area-bike-share

To begin with, we chose two similar data sets, which both relate to bike-sharing rides in major US cities. One dataset contains information about specific bike-sharing rides in NYC in 2015-2017. Each row has information about the ride duration, start/end station, and gender/age of the rider. The second dataset contains information about the Bay Area bike share for 2013-2015. There are tables with information pertaining to the stations, trips, and weather for several days.

In our project, we plan on analyzing both data sets and building many models in order to make accurate predictions. Ultimately, we also hope to see how models trained on each of the separate datasets will perform on the other dataset. The problem statement we are trying to answer is: what factors are closely related to the time it takes a biker to go from their start station to their end station? These factors can be gender, age, weather, etc.... For instance, specifically with respect to the Bay Area dataset, it may take an elderly person in icy weather longer than expected to complete the trip. We will also analyze both data sets (together as well as individually) in order to gain a better understanding of the demographics of the riders. This problem is important, because it can be useful to know additional factors that relate to a trip's duration such as rider demographic information besides just the start and end location of the trip. This can then help in more accurately predicting how much time it may take a person to go from their starting location to ending location.

From the data analysis perspective, we will analyze both sets of data and see how often and how closely the duration from point A to point B closely resembled Google Maps' expected duration to go from point A to point B.

We think that we have a very sufficient amount of data to guarantee us accurate results.