Data 1030 - Project Proposal
Michael Harder
9/30/2019

Citi Bike Investigation

In an effort to not only reduce traffic, carbon emissions, and roadwear, but also improve public health in New York City; public officials launched a public bike share transportation alternative known as Citi Bike. Citi Bike became officially operational in 2013 and has been expanding ever since. Servicing Manhattan, Brooklyn, Queens, and Jersey City provides its share of logistical challenges. Applying machine learning techniques to publicly published Citi Bike data on their website, we can better understand their riders' habits.

NYC Open Data has published datasets containing Citi Bike station data as well as Citi Bike trip data; the latter representing 30 million Citi Bike trips from 2013 to present. These datasets have been made public via various avenues. A classification algorithm will allow us to identify how Citi Bike annual subscription riders differ from 24-hour pass or 3-day pass users. By successfully predicting whether or not a trip was conducted by a subscription rider or an everyday customer we can potentially find interesting insights into how their behaviors differ by investigating the model's parameters. For the purposes of this project I will be looking into Citi Bike trip data over the last two years, from August, 2017 to August, 2019. The features included in the data set are: trip duration (in seconds), start time and date, start station name, end station name, station ID, station latitude and longitude, bike ID, user type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member), gender (0 = unknown, 1 = male, 2 = female), and birth year. Because we want to classify if s trip was made by a customer or a subscriber, the target variable will be user type.

In order to preprocess this data I will apply the following steps:

- One Hot Encoder will be applied to the following features:
    - start station name, end station name, and birth year
    - One Hot Encoder is used here because they are categorical without any obvious rank
    - Please note: One Hot Encoder would have been applied to gender as it is a non rankable categorical feature. However, it's values are already supplied as a 0,1, or 2 (0 = unknown, 1 = male, 2 = female)
- Standard Scaler was applied to the following feature:
    - trip duration
    - Standard Scaler is used here because outliers have potential to not ensure reasonable upper and lower bounds
- Label Encoder was applied to the following feature:
    - User type
    - Label Encoder is used here because it is a categorical target variable
- Min Max Encoder was not applied to any features

- Ordinal Encoder was not applied to any features. The only categorical with an obvious rank was user type based on the subscription of the user. However, user type was preprocessed with the Label Encoder because it is our target variable.
- Left alone: longitude and latitude of start and stop station
- Dropped: start station ID, end station ID, bike ID
    - These identification numbers will not provide any added value to our modeling