Capstone Three Proposal

Taxi Trip Prediction

Problem + Context

Taxi trip durations in NYC can vary wildly because of several factors from weather and traffic to simply geography of the pickup and drop-off locations. For instance, New York is abounding with one-ways, side streets, and known for the ever-present number of pedestrians. With everyone engaging in a mad rush to get from point A to point B, you'll invariably find yourself late for whatever you need to be on time for. The goal of this analysis is to determine which factors predict the taxi trip duration.

Stakeholders + Audience

The main beneficiary will be the Taxi & Limousine Commission of NYC. They'll better be able to control costs and maintain a level of accuracy relating to 'appropriate' travel times for taxi drivers. Another group that will benefit are passengers that will be better able to know how long trips will take, ahead of time.

Data

The main data stems from the 2016 NYC Yellow Cab trip record data from Big Query on the Google Cloud Platform.

Two main datasets will be used for this project:

- Taxi_Train
- Taxi_Test

Taxi_Train
Each row corresponds to a specific taxi trip with features detailing pickup & drop off times, passenger count and other variables. It consists of 1,458,644 observations and 11 variables, provided via Kaggle. The target variable is named **trip_duration** and it indicates the length of a taxi trip. Data is 81 MB compressed and available as a single zip file.

Taxi_Test
Similar to the Driver_Train dataset, each row is associated with a trip. This dataset consists of 625,134 observations and 10 variables. Data is 21 MB compressed and available as a single zip file from Kaggle. We will be predicting the 'trip_duration' variable for this dataset.

Data Description:

- In the train and test data, features that belong to similar groupings are tagged as such in the feature names (e.g., ind, reg, car, calc).
- Feature names include the suffix 'bin' to indicate binary features and 'cat' to indicate categorical features.
    - Features without these designations are either continuous or ordinal.
- Values of -1 indicate that the feature was missing from the observation.

Constraints + Scope

- Not all features in our data set will be useful.
    - Feature selection is an iterative process where we run an initial model with either few or many features and then step-by-step add or remove some features based on the results of the previous run.
- The features are anonymized, but there’s still a chance that it may end up being tracked back to the passenger(s).
    - The anonymity of the features will make it difficult to interpret the characteristics of the taxi trip and any relations to other features.

Approach

Given the built-in time & geographic nature of the features, the first order of business is to determine which features, or groups of features, have the most impact (correlation) with the trip duration (target). In order to take this on, a significant number of visualizations/plots will be employed during the Exploratory Data Analysis portion of this capstone. In particular, an alluvial plot to determine multi-feature interactions.

Once there’s been sufficient exploration of the connections within the current features, I’ll use these insights to build new features.

I will be using XGBoost to be able to predict how long the trip will last when taking a specific route.

Deliverables

A project jupyter notebook will be developed to detail the python code used and, in particular, exhibiting my adherence to the Data Science Method. A slide deck & report summarizing findings of note, analysis process & recommendations will be developed alongside associated figures/plots.