SPRINGBOARD PROJECT PROPOSAL

New York City Taxi Trip Duration

New York City taxi rides paint a vibrant picture of life in the city. The millions of rides taken each month can provide insight into traffic patterns, road blockage, or large-scale events that attract many New Yorkers. With ridesharing apps gaining popularity, it is increasingly important for taxi companies to provide visibility to their estimated ride duration, since the competing apps provide these metrics upfront. Predicting duration of a ride can help passengers decide when is the optimal time to start their commute. This problem was posted NYC Taxi and Limousine Commission as competition in Kaggle.com challenging us to build a model that predicts the total ride duration of taxi trips in New York City.

The primary goal of this project is to predict trip duration of NYC Taxis based on features like trip coordinates, duration date and time.

Primary data for this analysis was released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables. Training dataset has close to 1.5 Million and 630k records in test dataset. Each row contains one taxi trip.

As a beginner take on this problem I will be applying Data Wrangling and EDA techniques learned in this course, to understand the underlaying data and to filter the data for my analysis. In order to understand various Machine Learning algorithms, I'm planning to employ Random Forest and XgBoost algorithms to build a model on subset of data with counts close 500K records.

We will deliver R codes of EDA, Modelling and predictions and also R Markdown detail report.