ORIE 4741 - Project Proposal

Audrey Yap, Sung Mo Kim
amy34, sk2346

Abstract and Background

New York is infected by taxis. The yellow cabs are synonymous with the Big Apple, and they are part of both the cultural identity as well as the historical identity of the city. When taxis were first introduced and used by city dwellers, people did not have a resource to compare fares other than word of mouth. Whatever fare a taxi driver - or perhaps carriage driver, to make a more historical comparison - chose to charge, the customer would not have an opportunity to identify if they were paying more or less than other customers. Along with the price, wait times for taxis were ambiguous. People waiting in lines for taxis would not know how long it would be for their turn since the taxis are going everywhere in the city and occasionally going back to the lines. People not in the lines would just have to try their luck at flagging a vacant one down. However, now customers live in an information-rich world where it is extremely easy to make these comparisons.

This information revolution has brought with it its own problems. Namely, now that customers are able to compare their fares with others, they have found some apparent inconsistencies and differences between their fares and another competitor, or customer's, charged fare. Customers have also found inconsistencies with estimated travel times from navigation apps and tend to become more impatient when the actual travel time starts to exceed the estimated.

The Dataset

In order to make accurate predictions regarding the time of travel and the cost of the travel, we decided to analyze two types of datasets. First is a CSV file that contains taxi trip details on yellow taxis in a given month. There are multiple datasets we can view for each month, so we look to randomly select sufficient data points from each month/year and create a new CSV file that will be more comprehensive in terms of the timeline. The second dataset contains similar information, but it includes the pickup location and drop-off location expressed in terms of longitude and latitude. These figures will allow an accurate calculation of the distance travelled, which will be an important feature in making our proposed predictions.

The data fields that will be mainly used include figures such as pickup time, drop-off time, pickup location id/longitude, latitude, drop-off location id/longitude, latitude, travel time, and fare amount which will all be necessary features in making accurate predictions.