Final Project Proposal

Team Member: Jiahui Luo, Qin Yin, Zhe Feng

Based on the 2016 NYC Yellow Cab trip record data originally published by the NYC Taxi and Limousine Commission (TLC), we are going to build a model that predicts the ride duration of taxi trips in New York City, since it can help passengers decide the optimal time to start their commute. In the meanwhile, given a dataset, we could use this prediction model for other cities as well.

Data Information:

- Data link: https://www.kaggle.com/c/nyc-taxi-trip-duration/data
- Data type: csv file (continuous, binary, datetime)
- Training size: 1458644
- Testing size: 625134
- Data Description:

1) id: a unique identifier for each trip
2) vendor_id: a code indicating the provider associated with the trip record
3) pickup_datetime: date and time when the meter was engaged
4) dropoff_datetime: date and time when the meter was disengaged
5) passenger_count: the number of passengers in the vehicle (driver entered value)
6) pickup_longitude: the longitude where the meter was engaged
7) pickup_latitude: the latitude where the meter was engaged
8) dropoff_longitude: the longitude where the meter was disengaged
9) dropoff_latitude: the latitude where the meter was disengaged
10) store_and_fwd_flag: This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
11) trip_duration: duration of the trip in seconds

Tasks:

1. data processing

Even though the data was sampled and cleaned for the purposes of Kaggle competition, we might still need to do some reprocessing for the purposes of our project, e.g. drop some unnecessary features.

2. data visualization

- plots: distance vs duration; average duration vs time(day) and time(week); average duration/distance vs location clusters
- heat map: pick-up and drop-off location, animation of taxi demand

3. prediction model

we will try to use different regression methods (Linear regression, ridge regression, random forest, gradient boosting) and ensemble methods to build prediction model and make comparisons.