New York City Taxi Fare Prediction

Team members: Huang Zhibin, Qiu Zhizhou, Tang Tianyi

1. Introduction:

   It is a supervised regression machine learning tasked with predicting the fare amount (inclusive of tolls) for a taxi ride in New York City given the pickup and drop-off locations, travel time and number of passengers.

2. DataSet:

   It is a Kaggle related project.

   The dataset basically contains 55M rows with 6 features. The results of "fare amount" is our goal, and models are trained to predict the continuous values.

fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
4.500	2009-06-15 17:26:21	-73.844	40.721	-73.842	40.712	1
16.900	2010-01-05 16:52:16	-74.016	40.711	-73.979	40.782	1
5.700	2011-08-18 00:35:00	-73.983	40.761	-73.991	40.751	2
7.700	2012-04-21 04:30:42	-73.987	40.733	-73.992	40.758	1
5.300	2010-03-09 07:51:00	-73.968	40.768	-73.957	40.784	1

Data Features:

- id - a unique identifier for each trip
- vendor_id - a code indicating the provider associated with the trip record
- pickup_datetime - date and time when the meter was engaged
- dropoff_datetime - date and time when the meter was disengaged
- passenger_count - the number of passengers in the vehicle (driver entered value)
- pickup_longitude - the longitude where the meter was engaged
- pickup_latitude - the latitude where the meter was engaged
- dropoff_longitude - the longitude where the meter was disengaged
- dropoff_latitude - the latitude where the meter was disengaged

- 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
- trip_duration - duration of the trip in seconds

3. Plans:
   - Understands task mission and dataset:
   - Explore and extract valid data:
     - try to find out abnormal or incorrect data and make corrections.
     - figure out the relations with each feature.
   - Evaluate and select module:
     - linear regression
     - random forest
   - Improve the module:
     - feed more data
     - build more features
   - Interpretation and make predictions

   Reference:
   https://www.kaggle.com/c/new-york-city-taxi-fare-prediction