**New York City Taxi Fare Prediction**

Team members: Huang Zhibin, Qiu Zhizhou, Tang Tianyi

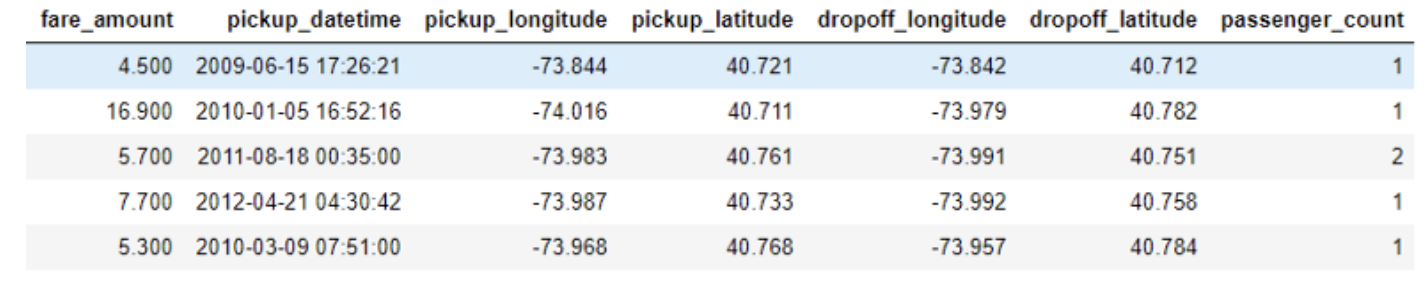
1. Introduction:

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.

1. 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.



**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

1. 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