Project-Proposal

Hukai Luo

12 October 2018

Introduction

The goal of this project is to predict the *duration of taxi rides in NYC* based on features like trip coordinates or pickup date and time. The data comes in the shape of 1.5 million training observations and 630k test observation. Each row contains one taxi trip.

Since there are lots of parameters when we are estimating the trip duration, we will first study and visualise the original data, engineer new features, and examine potential outliers. Then we add two external data sets, one is the NYC weather and the other is the fatest routes. We will visualise and analyse the new features within these data sets and their impact on the target trip duration values. Finally, we will choose some important parameters to curve fit the data using the least square method.

We use the *multiplot* function, courtesy of R Cookbooks to create multi-panel plots. We use *data.table's* fread function to speed up reading in the data. We can have an overview of the data sets using the *summary* and *glimpse* tools.

We find:

- *vendor_id* only takes the values 1 or 2, presumably to differentiate two taxi companies
- *pickup_datetime* and (in the training set) *dropoff_datetime* are combinations of date and time that we will have to re-format into a more useful shape
- *passenger_count* takes a median of 1 and a maximum of 9 in both data sets
- The *pickup/dropoff_longitute/latitute* describes the geographical coordinates where the meter was activate/deactivated.
- *store_and_fwd_flag* is a flag that indicates whether the trip data was sent immediately to the vendor ("N") or held in the memory of the taxi because there was no connection to the server ("Y"). Maybe there could be a correlation with certain geographical areas with bad reception?
- *trip_duration:* our target feature in the training data is measured in seconds.

That means we need to deal with these unusual data.