Prediction of Trip Duration Time in Divvy Bike Data

Sunyi Chi, Bangyao Zhao, Xuemei Ding

Introduction

The Divvy Bikes is a bike sharing program in Chicago, Illinois, which offers a convenient and affordable way to get around the city. 6,131 bicycles can be rented from and return to 620 stations. Given that the number of bicycles at each station is relatively limited, it is meaningful to predict the renting time of each bicycle so as to reasonably arrange the resting bicycles.

Preliminary Analysis

1. Data background:
   The Divvy data set contains Divvy Bikes' using information in 2018. 3,603,082 samples (trips) were recorded. For each trip, there is information of start and end time and station, as well as information of the user. Our goal is to predict the duration time of each trip.
2. Data cleaning process:
   There are missing values in gender and age. We deleted those trips with missing data, where 3,040,517 samples (trips) remain.
3. Data description:
   a. trip_id: trip ID
   b. The outcome of interest:
      tripduration: end_time - start_time, in minutes.
   c. Potential Covariates:
      Continuous Variables:
         Age: age of user
      Categorical Variables:
         1. from_station_name: name of the station where the trip starts. There are 620 different stations.
         2. usertype: user type. Can be either subscriber (Annual Members) or customer (users who bought single ride or one-day pass)
         3. Gender: gender of the user
         4. Season: in which season the trip took place.
         5. Start_time_cat: in which time slot of a day the trip took place
         6. weekends: indicator of whether the trip happens on weekends

Research Goal and Challenge

We are interested in predicting the trip duration time using both information of the trip (start time and start place) and information of the user (type, age, gender). As the number of trips and the number of stations (as categorical variables) are huge, we are facing a large n, large p challenge. If we want to fit a linear model, the R function *lm* breaks with a memory issue. Further, if we fit the linear model on a subset of the data, we also see a non-random trend in the residual plot. To solve the memory issue and improve prediction accuracy, we are going to use nonparametric models with parallel computing and cluster computing to predict duration of time using Divvy bikes in Chicago.