New York Citi Bike Trip Duration Prediction

CSE 258 Assignment 2

Zhuo Cheng
A xxxx
UC San Diego
San Diego, CA
zhc262@eng.ucsd.edu

Jiamin He
A53243479
UC San Diego
San Diego, CA
jih426@eng.ucsd.edu

Tianran Zhang
Axxxx
UC San Diego
San Diego, CA
tiz217@eng.ucsd.edu

ABSTRACT

In this project, a model that predicts the total ride duration of taxi trips in New York City is built. The patterns related to the geographical locations and pick-up and pick-off time with passenger businesses are observed and evaluated. We utilize the dataset released by the NYC Taxi and Limousine Commision, which provides the necessary information to accomplish our goal We manipulated and mined the dataset by multiple models we learned from CSE 258 course taught by Prof.Julian McAuley and the further reading materials. [MODELS MODELS MODELS] [RESULTS RESULTS RESULTS] Predicting fare[??] and duration of a ride can help passengers decide when to start a commute and help drivers decide an optimal ride. Our model can be used to give an estimated fare and duration prediction.

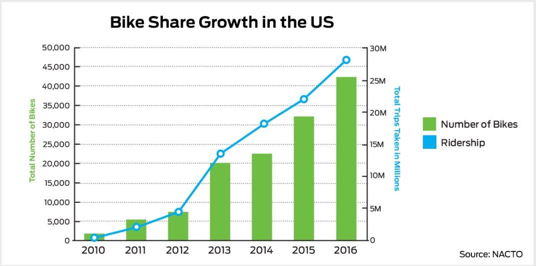
KEYWORDS

Give me FIVE!!!!!!

1 INTRODUCTION

Bike shares are riding a wave of popularity in the intermodal transit planning community. Through bike sharing systems in a city, people are able to rent a bike from a one location and return it to a different place on an as-needed basis. The number of bike share systems, defined as publicly-available systems with at least 10 stations and 100 bikes, has steadily increased year-over-year, from four systems in 2010 to 55 systems in 2016 across U.S, with over 42,000 bikes available in cities of all sizes[1]. In addition, 80% of systems that have been in operation for more than a year have expanded since they launched. The number of bikes in the nation also increased substantially, up 30%, as existing large systems have continued to grow. Figure 1 shows the bike share growth in the US from 2010 to 2016[1].

The growth of bike share shows no signs of stopping. A number of U.S. cities, such as Detroit, New Haven, and New Orleans, have either selected vendors or are planning to launch systems, and many existing systems are also rolling out major expansions: New York's Citi Bike Program [2] is adding another 2,000 bikes, for a total of 12,000; Houston is more than tripling in size to over 100 stations; and the San Francisco Bay Area is expanding from a 700 to a 7,000 bike system[3]. Here in our assignment, we conduct a case study based on the bike sharing data on the Citi Bike Share System in New York City in order to predict the duration for a user in one bike trip at a certain time. We try to forecast the duration of a share-bike trip to help improve and expand the bike sharing system in NYC. Predicting duration of a bike ride can also help users to make wise decisions on hailing a cab or riding a bike, for example.



Source: NACTO

Figure 1: Bike Share Growth in the US

In this project, we obtained the primary dataset provided by Citibike System Data. The data generated by these bike share systems are consisted of duration of travel, start time and date, stop time and date, start station name, stop station name, station ID, station latitude and lonitude, bike ID, user type(Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member), user gender and user's year of birth. Given adequate data from Jul 2013 to Sep 2017, we used data in 2016 year round. In our analysis, we attempt to evaluate the importance of different features and extract the non-trivial ones to create a prediction model for this problem. We then compare performances of different models and discuss their effectiveness and shortcomings. In this assignment, We mainly use the method of multiple linear regression analysis, Random Forest and XGBoost to forecast the bike share trip duration in New York City. An ensemble of Random Forest and XGBoost is also applied in this report. We evaluate these models based on the Fraction of Variance Unexplained (FVU).

2 DATA

We obtain the data from the data system of Citi Bike, which is the national largest bike share program, with 10,000 bikes and 600 stations across Manhattan, Brooklyn, Queens and Jersey City[2]. It consists of data from the program's launch date, from 07/2013 to 09/2017. The dataset provided information is as shown in Table 1.

[1] Figure taken from https://nacto.org/bike-share-statistics-2016/, December 02, 2017

Variate	Format
Trip Duration	in seconds format
Start Time and Date	Timestamp
Stop Time and Date	Timestamp
Start Station Name	String
End Station Name	String
Station ID	Number
Station Lat/Long	Number
Bike ID	Number
User Type	Customer or Subscriber [2]
Gender	Number [3]
Year of Birth	Number

Table 1: Dataset Variable and Type

Basically, there were [XXXX] bike stations and total number of rides in dataset is [XXXXX]. Total number of customers' rides is [XXXXX] and subscribers' rides is [XXXX].

Before choosing useful features, we do some cleaning work since there are some deficient data in the dataset: some rides miss parts of the usersâĂŹ information such as the age and sex; also by looking at the start time and stop time fields, we find some rides take over 1 million seconds, roughly around one month, which is unreasonable. Besides, we calculated the speed of each ride and find that some speeds are too low or too high. To avoid the effect of these outliers, we restricted the speed of each ride to be greater than 1 m/s and less than 8.5 m/s, which falls in the range between the speed of walking and driving.

After the cleaning process, we randomly choose 30,000 data from each month of the 2016 to form our dataset with size 360,000 in total. And we split it into three sets: training, validation and test, each with 120,000 rides.

3 PREDICTIVE TASK

4 MODELS AND METHODOLOGY

5 LITERATURE

6 RESULTS AND CONCLUSIONS

REFERENCES

[1] [n. d.]. *Bike Share in the US: 2010-2016.* https://nacto.org/bike-share-statistics-2016/.

[2] [n. d.]. *Citi Bike Share system in New York City.* https://www.citibikenyc.com.

[3] [n. d.]. *How bike share is changing American cities.* https://www.curbed.com/2017/3/21/15006248/bike-share-ridership-transit-safety.