# 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.

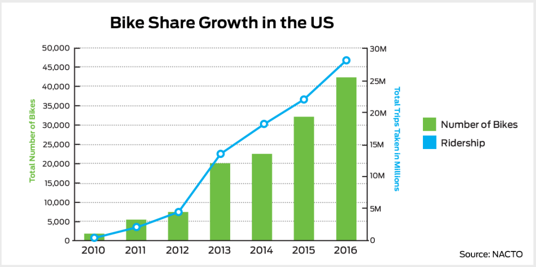

**Figure 1: Bike Share Growth in the US**

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

### 2.1 Dataset Variables

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

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

consists of data from the program's launch date, from 07/2013 to 09/2017. The dataset provided information[2][3] is as shown in Table 1.

**Table 1: Dataset Variable and Type**

| 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 |
| Gender | Number |
| Year of Birth | Number |

## 2.2 Data Cleaning

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].

From a preliminary analysis on these data, we need to detect and deal with some abnormal data. For some cases too short duration or even zero or negative is deleted; cases that takes too long duration in a trip (> 1 million seconds, roughly around a month) are also removed; for some cases the users' information are missing, age and sex, for example; for some the ratio of distance by duration is unreasonably high (30 mile/hour) or low (1 mile/hour). A lower bound (2.2369 mile/hour) and an upper bound (19.0137 mile/hour) is used in this analysis to protect our models from being skewed by extremities and make them immune to outliers, which is a reasonable range from walking to driving[10].

After all these data cleaning, we randomly choose 30k trips from each month in 2016, which creates a dataset with a size of 360k in total. And we split it into three sets: training, validation and test, each with 120,000 rides.

## 2.3 Exploratory Analysis

To understand [Here we need a picture to show the overall distribution of bike duration. –> each year][show the time period that the dataset cover...] [Different Pictures...] [EDA]

*2.3.1 prevelance.* Fig.X illustrates an increase in the number of trips starting from 2014, when the citi bike program just launched to 2017. This shows that more and more people start to [PIC!!!!!!]

# 3 PREDICTIVE TASK

Based on the given dataset, what we want to do is to predict a single trip duration in the Citi Bike Share System in New York City.

## 3.1 Evaluation Method

We evaluate the performance of different models on the basis of the fraction of variance unexplained (FVU). FVU is calculated as:

$$FVU(f) = 1 - R^2 = \frac{MSE(f)}{Var(y)}$$

FVU is suitable for our problem because FVU in statistics, is the fraction of variance of the regressand (dependent variable) Y which cannot be explained, i.e., which is not correctly predicted, by the explanatory variables X. When trying to predict Y, the most naive regression function that we can think of is the constant function predicting the mean of Y, i.e., $f(x_i) = \bar{y}$. It follows that the MSE of this function equals the variance of Y and the FVU then reaches its maximum value of 1. So if the explanatory variables X tell us nothing about Y, $FVU(f) = 1$. But as prediction gets better, MSE can then be reduced and ideally the best modle will have $FVU(f) = 0$.

## 3.2 Data Preprocessing

In this dataset, 30k observations is randomly extracted from data in each month in 2016. And this aggregated dataset now contains 360k observations and is well distributed in all year round. It is then shuffled and divided into three splits, individually 1/3 for training, validation and test. Each training sample consists of 15 entries – the first 11 contains the basic information for a certain bike trip. In addition, there are also 4 entries that represent the rider's information. To make a better use of these training samples, we preprocess and convert them into a more intuitive form as described below.

(1) Day of month:
From the given timestamp of the start date and end date we can easily compute the day of the month, and we use it as 31 features to represent different days in a month.
(2) Day of week:
Using the timestamp provided in our dataset, by using one-hot encoding, we construct 7 features to represent different days in a week in order to provide more explantion on weekdays and weekends.
(3) Year:
As we focused mainly on data in 2016, so the year feature here is set all 2016, and in later discussion be left behind.
(4) Month:
With the same technique, we use 12 features to represent the months from January to December.
(5) Hours:
Given the specific start time and stop time, by one-hot encoding we can use 24 features to represent the 24 hours in a day.
(6) Minutes:
From the timestamp value present in the data set, we divide 0~60 minutes into four features, each of them covers a range in 15 minutes.
(7) Distance:
Given latitude and longitude of the start location and stop location, we can calculate In order to represent the latitude and longitude information of the start and end station in some useful forms, we consider several different ways, such

[2] Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member

[3] 0=unknown; 1=male; 2=female

as mapping them into zip code and use one-hot encoding to represent them into categorical features. Due to the largeness of our data set, it takes too long to convert each pair of latitude and longitude into a zip code representation. We choose to use the distance between the start station and end station as the feature. The conversion is simply by using the vincenty distance of the Geopy library in Python[4].

(8) Gender:
In the orginal dataset, 1 for males and 2 for females. After processing, we convert it into a boolean feature, 1 for females and 0 for males.

(9) Age:
In order to represent the age in a useful form, we set 30 year-old as a watershed and use two categories to represent it. If the given age is smaller than or equal to 30, we set the negative value of given age to the first category and 0 to the second category. If the given age is greater than 30, we set 30 to the first category and the difference between the given age and 30 to the second category. For example, if the given age is 15, then we set -15 to the first category and 0 as the second. If the given age is 40, then we set 30 to the first category and 10 to the second. The intuition to represent the feature age in this way is that we consider the age 30 and below is negatively related to the duration of the ride, which means that the closer to the 30, the shorter duration it takes. And we consider the age above 30 is positively related to the duration of the ride, which means that the further to the 30, the longer duration it takes.

## 3.3 Features Selection

The previous exploratory analysis on different features illustrated a few imporatant features that we should include. We firstly incorporated the latitude and longitude of the start and end station, month, week, hour, minutes, gender and age as our feature vectors. To analyze the importance of each feature, we remove each of these features from our feature set to see how the performance changes. Fig.X shows the importance of each chosen features. [PIC!!!!!]

[Time this part... ]

When we consider if selecting the day or week as one of the features and if adding time into the feature vector, there are different opinions in our group, so we draw a bar graph (Fig. 2) to compare four cases: week, week and time, day, day and time. We add one of these combinations into our feature set each time and run our linear regression model to see which one works best. As shown in the graph, by adding the features week and time, the performance is the best.

# 4 MODELS AND METHODOLOGY

# 5 LITERATURE

## 5.1 Relevant Dataset

Our dataset is provided by New York Citi Bike Share System and most of the bike share system provide opening dataset in similar formula to public. Similar dataset can also be visited and gained through Hubway Bike Share System in Boston, MA[4]; DecoBike in Miami, FL[5]; Relay Bike Share System in Atlanta, GA[6]; Reddy Bike Share System in Buffalo, NY[7]; Divvy Bike Share System in Chicago, IL[8]; Capital Bike Share System in Washington, DC[9],etc. These systems generate a great deal of data relating to various ride details, including trip duration, start and end location, etc.

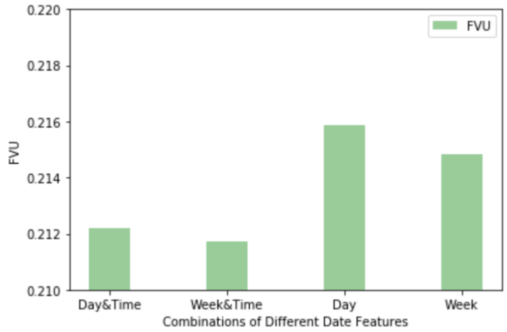

**Figure 2: Date and Time Feature Selection**

The dataset we used here is of great importance for us to study public transporation and city mobility. Leveraging the historical data provided can benefit us to providing more accurate prediction for bike trip duration at a time.

## 5.2 Related Work

Many previous research and investigation have been conducted in bike trip prediction. But most of them are forcasting the share bike demand at a certain time or at a certain place. Also, there has been a share bike demand prediction Competition running on Kaggle three years ago. But few of relevant projects are delved into bike trip duration prediction as what we did in this project.

For those who are predicting the share bike demand, some techniques and key features that they have used and worth mentioning are as follows:

### 5.2.1 *Geographic and Distance Information.*

Javier etc.[7] develops a rapid response ridership forecast model, based on the combined use of Geographic Information Systems (GIS), distance-decay functions and multiple regression models. Analyses carried out show that weighting the variables according to the distance-decay functions provides systematically better results. The choice of distance threshold also significantly improves outcomes. Osvaldo etc. [5] develops models based on GIS and proves its considerable advantages over the traditional four-step model, including simplicity of use, easy interpretation of results, immediate response and low cost. This study also uses geographically weighted regression (GWR) to deal with the staion prediction problem. Patrick etc.[9] refrains from building a spatial model to assess the bike share repositioning services.

[4] Hubway Bike Share: http://www.cambridgema.gov/CDD/Transportation/gettingaroundcambridge/bikeshare
[5] Deco Bike, Citi Bike: http://citibikemiami.com
[6] Relay Bike Share: http://relaybikeshare.com/system-data/
[7] Relay Bike Share: https://reddybikeshare.socialbicycles.com
[8] Divvy Bike Share: https://www.divvybikes.com
[9] Relay Bike Share: https://www.capitalbikeshare.com/system-data

*5.2.2 Bike Docking Station Locations.* Yexin Li etc.[8] proposed a bipartite prediction algorithm based on hierarchical clustering to cluster bike stations into groups and therefore get different levels of hierarchical bike stations.

*5.2.3 User Subscription Plans and User Habits.* Elliot etc.[6] investigates several significant predictors of membership includingreactions to mandatory helmet legislation, riding activity over the previous month, and the degree to which convenience motivated private bike riding. These results provide insight as to the relative influence of various factors impacting on bike share membership in Australia. The findings may assist bike share operators to maximize membership potential and help achieve the primary goal of bike share âĂŞ to increase the sustainability of the transport system.

*5.2.4 Time Series Data.* To predict the demand of bikes, a multi-similarity-based interference model based on gradient boosting regression tree is proposed to predict the rent proportion across clusters and the inter-cluster transition, based on which the number of bikes rent from/returned to each cluster can be easily inferred[8].

### 5.3 Analysis on this task

From the aboved analysis, we can see most of the prediction task are trying to solve the re-allocation problem or the arrangement of the bike docking station problem. So they pay more attention on the specific bike docking stations and try to deeply understand users' habits. Also, a majority of these investigations are made to predict the bike demand for an entire area and that is why they care more about geographic features. However, our predictive task is different from above mentioned work because we are trying to predict the duration of each bike trip. So we convert the specific location information such as longtitude and latitude to a more concrete distance feature during a trip. Also, the user's habit and subscription would be helpful in our prediction and we also included sexuality, age and other user information in our predcition models.

## 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.

[4] [n. d.]. *Python Geopy Toolbox.* https://pypi.python.org/pypi/geopy.

[5] Osvaldo Daniel Cardozo, Juan Carlos García-Palomares, and Javier Gutiérrez. 2012. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography* 34 (2012), 548–558.

[6] Elliot Fishman, Simon Washington, Narelle Haworth, and Angela Watson. 2015. Factors influencing bike share membership: An analysis of Melbourne and Brisbane. *Transportation research part A: policy and practice* 71 (2015), 17–30.

[7] Javier Gutiérrez, Osvaldo Daniel Cardozo, and Juan Carlos García-Palomares. 2011. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *Journal of Transport Geography* 19, 6 (2011), 1081–1092.

[8] Yexin Li, Yu Zheng, Huichu Zhang, and Lei Chen. 2015. Traffic Prediction in a Bike-sharing System. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '15).* ACM, New York, NY, USA, Article 33, 10 pages. https://doi.org/10.1145/2820783.2820837

[9] Patrick Vogel and Dirk Christian Mattfeld. 2010. Modeling of repositioning activities in bike-sharing systems. In *World conference on transport research (WCTR).*

[10] Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. 2008. Learning Transportation Mode from Raw Gps Data for Geographic Applications on the Web. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08).* ACM, New York, NY, USA, 247–256. https://doi.org/10.1145/1367497.1367532