*

Pittsburgh Bike Share Data: Progress Report

Ayush Mishra (amm428@pitt.edu) Junwei Zhuge (juz25@pit.edu)

ABSTRACT

The aim of this project is to predict the bike share counts for different stations across the city of Pittsburgh. The raw data is transformed to obtain actual results. In this Progress report, we made use of K-NN, LOGISTIC REGRESSION(LR) and NAÏVE BAYESIAN(NB) methods to train our model and predict the counts. A further aim is also to compare these counts against the rack quantities at each of the bike stations. The predictions based on our models can help determine the optimum time matrix to "re-balance" the bike accumulation and restore them to their optimum levels across the city.

1. INTRODUCTION

In this report, we first combine the Q3 and Q4 bike share data into a single entity to ease the processing. This gives us a consolidated view of the data and we can work on this to give us a better view of the response variable that we need to predict. At this point however, we do not have the actual response variable. We further process the dataset to get rid of "missing values". After this, we split some consolidated columns such as "StartTime" and "StopTime into more edifying and workable columns.

Our goal is to predict the counts that are *rented out* and are *deposited in* to each of the stations each hour. Thus, we move one step forward by splitting the above mentioned consolidated columns into more processible data such as "StartMonth", "StartDay", "StartHour", "StopMinute", etc. This brings us closer to determining the response variable.

We aggregate the dataset by grouping together predictor variables and determining the counts. This count data is our response variable. The detailed methods of our predictions are shown in the following sections.

2. RELATED WORK

First, we plot some density plots based on those potential useful attributes, including "Hour", "Day", "Month", and "UserType". Plots are showed in Figure 1 to Figure 4.

According to Figure 1, the "Pitt Bike Trips by Hour of Day" plot, the fastigium of Pitt bike trips is between 8 - 20. According to Figure 2, the "Pitt Bike Trips by Day of a Month" plot, we will not take account of "Day" element, instead, we will calculate weekdays and weekends from Month and Day, and take weekdays and weekends as variables. According to Figure 3, the "Pitt Bike Trips by Month" plot, July and August have the most bike trips, and the bike trip amount declines month by month till December. In Figure 4, 1 represent Member (pay as-you-go customer), 2 Subscriber (deluxe and standard monthly member customer), and 3 Daily (24-hour pass customer).

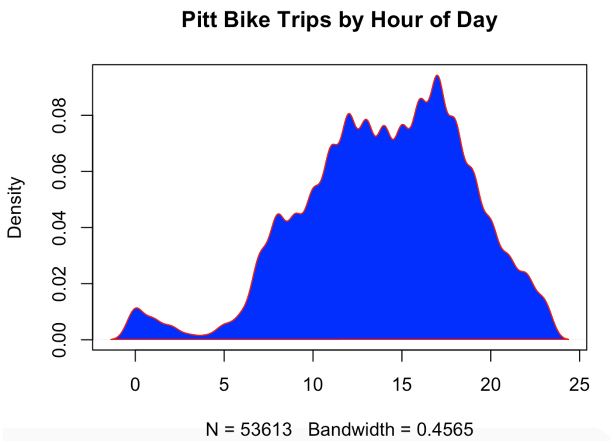

Figure 1- Pitt Bike Trips by Hour of Day

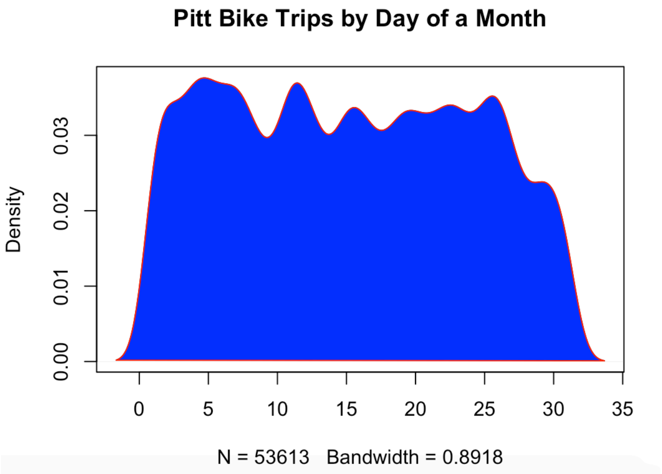

Figure 2- Pitt Bike Trips by Day of Month

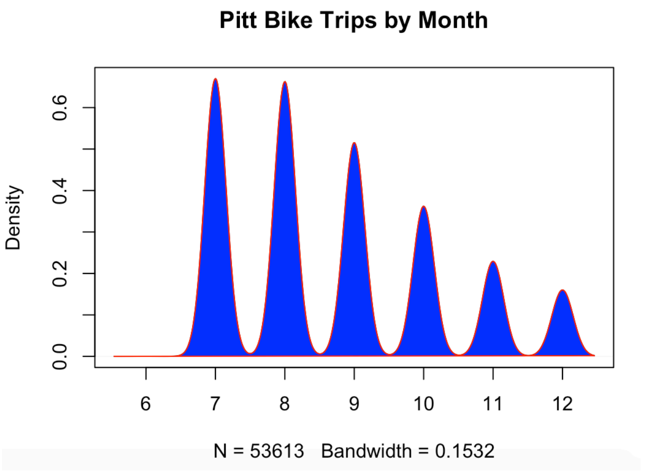

Figure 3- Pitt Bike Trips by Month

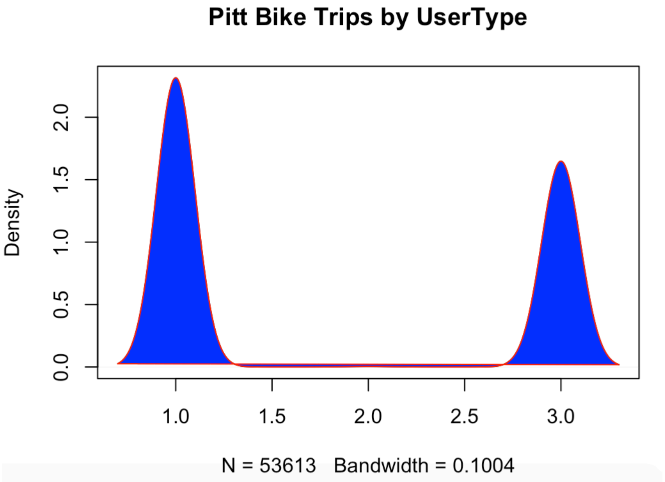

Figure 4- Pitt Bike Trips by UserType

Besides, we deleted some attributes that are not important for the results including "StartMinutes" and "TripId", etc., and add some important attributes such as "Weekdays" and "Qnty". Split the data set into two data sets, "Start" data set and "Stop" data set. Calculated the "StartCount" and "StopCount" separately from these two data sets based on "UserType", "StationId", "Month", and "Day". Finally, we calculated the "Final Count" from 'StartCount" and "StopCount". If the Count is larger than 0, it means bikes

In our final table, the variables will be "StationID", "UserType", "Month", "Weekday", "RackQnty", and the "y" will be the bike trip count of a specific station at a day, values with "-" means out, "+" means in. So we can compute the final bike amount of a station at one day, so we can decide the rebalancing problem based on the amount and the "density plot based on Hour".

What's more, according to the Figure 1, "Hour of Day" seems an important variable. However, the "StartHour" and "StopHour" of every trip between training and testing set cannot be exactly the same, so I delete these two column, but I will use "Hour" element as an important factor to decide what time (a time period in a day) is the best period to retransfer the bikes, just as "density plot based on StartHour" showed.

3. DATA SET

The data set used for building our models is from preprocessing after multiple iteration, including deleting many attributes and adding useful attributes, and splitting variables to get more meaningful variables. We calculated "Weekdays" from year, month and day. Although there are only totally 4 weekdays in the dataset, Tuesday, Thursday, Friday, and Sunday, they still could be very important and meaningful attributes to our models. Finally, we separate the dataset into training set and testing set according to the day in a month, before 20th and after 20th in a month.

Data Column	**Description**
y	Count = StartCount - StopCount
StationId	All Pittsburgh station ID
Usertype	User Type (1 - Member (pay as-you-go customer), 2 - Subscriber (deluxe and standard monthly member customer), 3 - Daily (24-hour pass customer))
Month	In which Month
RackQnty	The rack quantity of one station
Weekdays	Days in a week (Tuesday – 5, Thurday – 4, Friday – 1, Sunday – 3)

Table 1- Dataset Variables and Descriptions

4. METHODS

For the initial processing we choose to apply K-NN (1,5,10), LR and NB methods to show the contrasts in the predicted data. We separate the counts to the ones that "Start" at a station and the ones that "Stop" at the stations. We merge these datasets to create a total counts dataset. We then create the training and test sets for predictors. These methods give us a peek at the prediction of the response counts as described in the following:

4.1 LOGISTIC REGRESSION

We extract the count data for the number of bikes *rented out* of a station named as the start station count ("StartCount") and the number of bikes *deposited in* to a station named as station stop count ("StopCount"). The actual response variable for the model predictors are determined by the difference in start counts and stop counts each hour. This gives us a clear picture of the residual number of bikes at a given station ("Count"). This *Count* variable

5. FUTURE SCOPE

The analysis shown above is just a preliminary analysis of the data. We can use these models to predict actual counts on an hourly basis and can compare these data to the individual rack quantities at each of the stations and thus can determine the most appropriate times to “re-balance” the bikes for efficient bike sharing across the city.

In order to make this efficient, we shall explore further number of methods such as SVM, DECISION TREES and ADA.

For example, the K-NN method can be used to produce the actual count predictions as follows: