Bike sharing system-report

Dor Dveer-305315467

04/07/2020

Abstract

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return have become automatic. Through these systems, user can easily rent a bike from a particular position and return at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousand bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Introduction

Apart from interesting real-world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data. The main goal of this work was to predict the number of rental bike users directly or throw registered or casual users.

Data

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors. The core data set is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available in http://capitalbikeshare.com/system-data.

The data set include 17,380 records.

Variables:

After quick research and data exploration, i decided to create some variables, and change others. The result is the following:

season : season (springer, summer, fall, winter).

holiday : if day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule).

weekday : day of the week.

workingday : if day is neither weekend nor holiday is 1, otherwise is 0.

weather :

Good: Clear, Few clouds, Partly cloudy, Partly cloudy.

Normal: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist.

Bad: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds.

Very Bad: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow.

temp : Temperature in Celsius.

atemp: Feeling temperature in Celsius.

hum: Humidity.

windspeed: Wind speed.

casual: Count of casual users.

registered: Count of registered users.

cnt: Count of total rental bikes including both casual and registered.

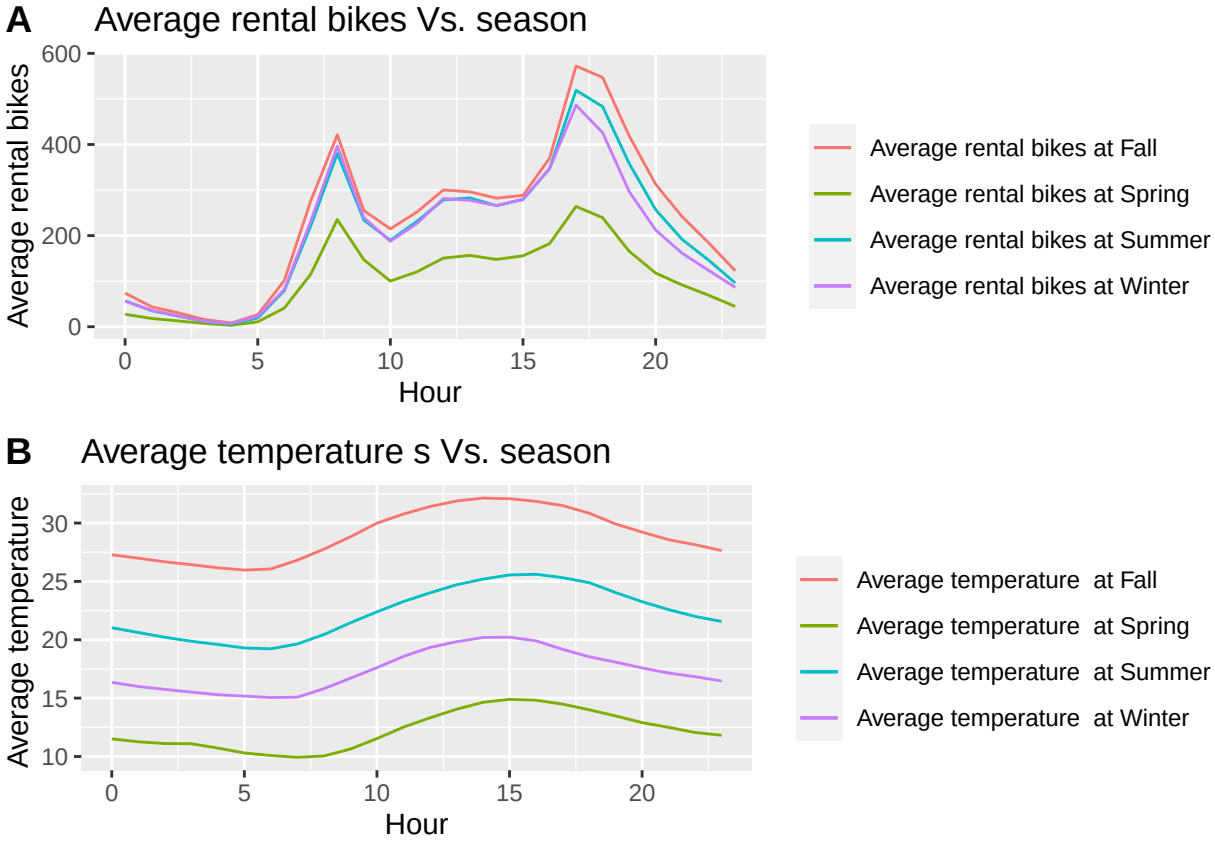
hour : Hour (0 to 23).

month : Month (1 to 12).

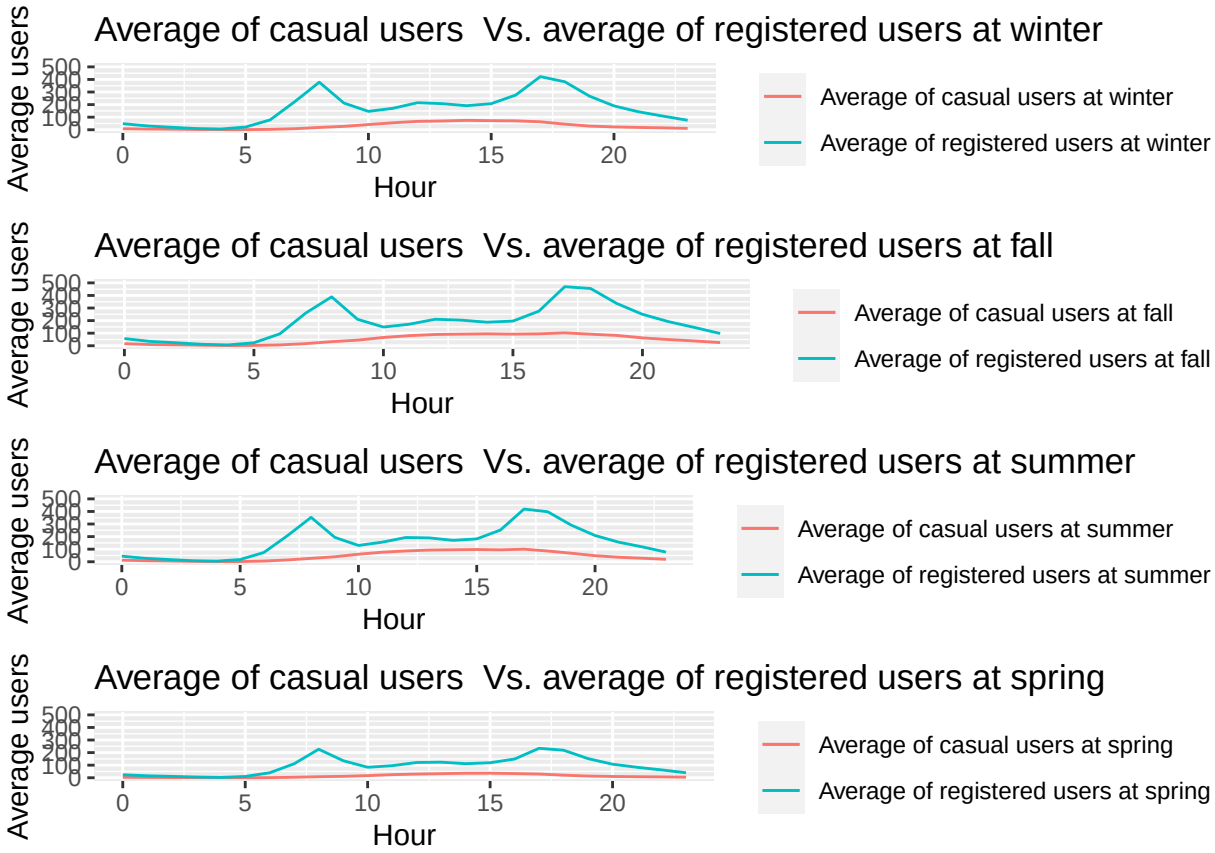
year : Year (2011, 2012).

day : Day (1 to 31).

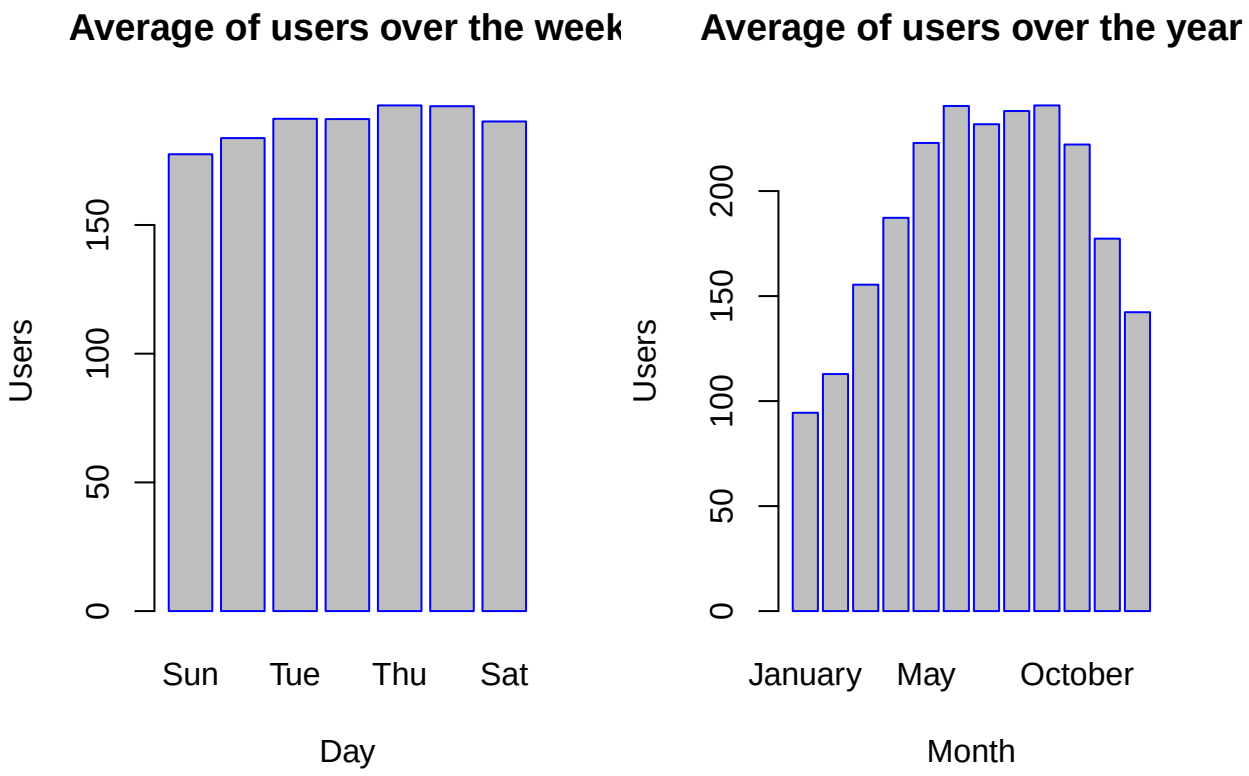
Exploring the data:



We can see that at fall the demand for rental bikes is the highest. It's also noticeable that on 7-8 at the morning and 17-18 in evening all year long, the demand for bike in very high although that the temperature is not the most convenient.



It is seemed that most of the registered users using a bike in the rush hours which can indicate that most of the users in the rush hour are regular users regardless the weather or season.



Surprisingly, for each day the number of rental bikes is the same. The lowest temperature is in the spring, probably that is the reason that the demand is the lowest in the spring.

Splitting to test and train

Since we are talking about time series the splitting to test and train will be as following: the training set is comprised of the first 21 (3 weeks) days of each month, while the test set is the 22th (last week) to the end of the month.

Models

At the following section I tried to predict the number of rental bikes users, the number of registered users and casual users. In each section I tried: *First* I used a quick Random forest model to select the important feature for using at all models.

Second, tuning each model:

- Decision tree-a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.
- Random Forest- The random forest is an algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.
- Boosting-e build a family of models that are aggregated to obtain a strong learner that performs better. However, unlike bagging that mainly aims at reducing variance, boosting is a technique that consists in fitting sequentially multiple weak learners in a very adaptative way: each model in the sequence is fitted giving more importance to observations in the dataset that were badly handled by the previous models in the sequence. Intuitively, each new model focus its efforts on the most difficult observations

to fit up to now, so that we obtain, at the end of the process, a strong learner with lower bias (even if we can notice that boosting can also have the effect of reducing variance). Boosting, like bagging, can be used for regression as well as for classification problems.

and for the first prediction

- NN- set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input.

Third, fitting each model in all train data and predict the test set.

Last, comparing the results according to RMSE:

$$RMSE = \sqrt{\sum_i \frac{(y_i - \hat{y}_i)^2}{n}}$$

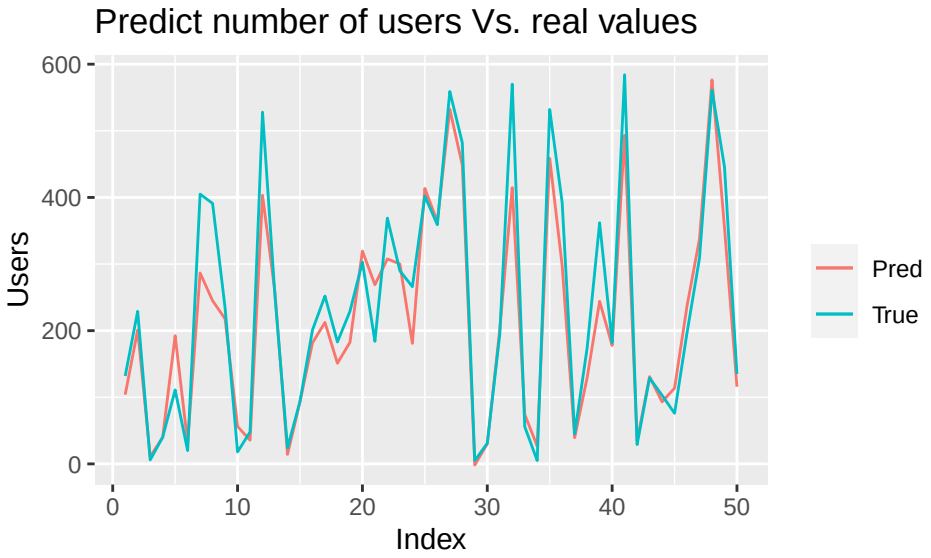
Predication of bike rental count hourly or daily based on the environmental and seasonal settings

BY using Random forest feature importance I selected the features for the future models: year, hour, workingday, month, atemp and day, assuming that the value of registered and casual users is unknown at the model time.

Summary of the results

Model name	Model error
Random Forest	62.154
Boosting	45.065
Boosting with scaled temp	45.119
Decision tree	68.538
NN	183.807

Plotting some predication from the winning model- boosting- in order to "sense" the error.



Most of the time the predictions are good. There are some points where the error is very larger, and this is probably the reason for the large RMSE. It is seemed the Boosting did a good job, but still it was far from the research question expectation.

Predication of registered users on daily based on the environmental and seasonal settings

The most important features are year, hour, workingday and month. At the following models I will use the variables that were selected according to random forest feature importance function.

Summary and results

Model name	Model error
Random Forest	97.77
Boosting	61.275
Decision tree	77.501

The error is still bigger than expected.

Predication of casual users on daily based on the environmental and seasonal settings

The most important features are: hour, day, temp, hum, atemp, workingday and month.

Summary and results

Model name	Model error
Random Forest	152.635
Boosting	149.514
Decision tree	149.92

The error for the last section was the biggest. Trying scaling or other method did not improve it.

Conclusion

- Time variable and temperature are the most important and have the biggest influence about the number of rental bikes.
- Registered users do not consider weather as important and they tent to use the bike anyway most of the year.
- The spring was with the lowest demand, due to low number of both registered and casual users.
- The fall was with the highest demand, mainly due to high number of casual users.
- The model succeeds to predict the total users with relatively small error. But the trail to prediction of registered and casual users along was fail.