Jisun Lee (U37416487)
Dr. Eugene Pinsky
CS 677 A1
Project Report

Bike Share in Washington D.C.

Bike sharing has grown in popularity over the last few decades. People are increasingly gravitating toward healthy, livable communities where activities such as bike sharing are widely accessible. It has evolved into a more environmentally friendly mode of transportation. In Washington D.C., the bicycle sharing system has gradually increased, expanding the market for more stations and bikes. Riders could choose between two types of passes: casual and registered. Casual has some options that are a single trip and 24-hour pass. Registered also has two options: Monthly or Annual Membership. This dataset contains hourly and daily rental bike counts as well as climate and seasonal data from 2011 to 2012. Moreover, the training set consists of the first 19 days of each month, and the test set consists of 20 days of each month until the end of the month. The dictionary of dataset is following:

datetime	hourly date + timestamp
season	1: Spring 2: Summer 3: Fall 4: Winter
holiday	whether the day is considered a holiday
workingday	whether the day is neither a weekend nor holiday
weather	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp	temperature in Celsius
atemp	"feels like" temperature in Celsius
humidity	relative humidity
windspeed	wind speed
casual	number of non-registered user rentals initiated
registered	number of registered user rentals initiated
count	number of total rentals

To do basic exploratory data analysis, I set the index in datatime and set the frequencies in hours. Then, I would like to see the trends of the data with the count variable. The number of using bikes is getting increase when time goes past. Analyzing data by weekdays and weekends, there is high demand during weekdays rush hours like morning and evening in registered. And, on weekends, the number of casual is higher than the number of registered. Moreover, casual people usually rent the bike afternoon. Analyzing weekday and weekends, it implies that registered people are mostly working. People who do not have membership usually rent the bike during afternoon weekends for going somewhere, hanging out, or exercising. After analyzing data on weather, it shows that people like to ride bikes when the weather is good (1.0); however, when the weather is bad (4.0), mostly registered people ride the bike. When analyzing the data with temperature, casual and registered people like to ride their bikes when the temperature is moderately warm. Even though the number of count is high in moderately warm between 20 to 30, the increment of the graph is not extremely high. When the wind speed is more than 22, the number of registered is gradually decrease and the number of casual is getting decrease when the wind speed is around 32. When analyzing data of season, the trends of casual and registered show similar shapes. However, the result of the data is quite impressive because both casual and registered people use more bikes in winter than the number of usage bicycles in spring. In the result of the correlation function, count, and registered are highly correlated because the number of registered people has a high portion in the count variable. The relationship between atemp and temp shows the highest relationship, but they are not good for feature because it is obvious that the temperature is affective to feeling temperature.

To compare with test, data and to predict data, season, weather, temp, atemp, humidity, year, hour, dayofweek, holiday, and workingday is the best feature because test data does not have count, casual, and registered features. For calculating error, I choose Root Mean Squared Logarithmic Error (RMSLE) because the larger the determination value, the larger the error value, which prevents the overall error value from increasing due to some large error values. I predict data with the linear regression and random forest. After calculating with using two methods, the RMSLE value for linear regression is 0.9804, and for random forest is 0.1067. Furthermore, I create the graphs for each method, and it seems similar. However, the error with using random forest is much less than the error with using linear regression because it ensembles the predictions of several decision tree regressions.