**BIKE SHARING DEMAND**

**Introduction**

In this project, my goal is to show when are the rider’s willing to pick the bikes from a rental station. The goal of this project is to leverage the historical data provided in order to forecast future bike rental demand within in the city. I have applied linear and logistic regression methods to increase the machine learning performance. Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world. The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city.

I have considered the weather data and compared it with other categorical variable to check the demand for the rental bikes.

**Training and Test Data:**

I have taken the Train and Test data from kaggle. Below is the data available from the train and test datasets.

atetime – 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:

Clear, Few clouds, Partly cloudy, Partly cloudy

Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

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

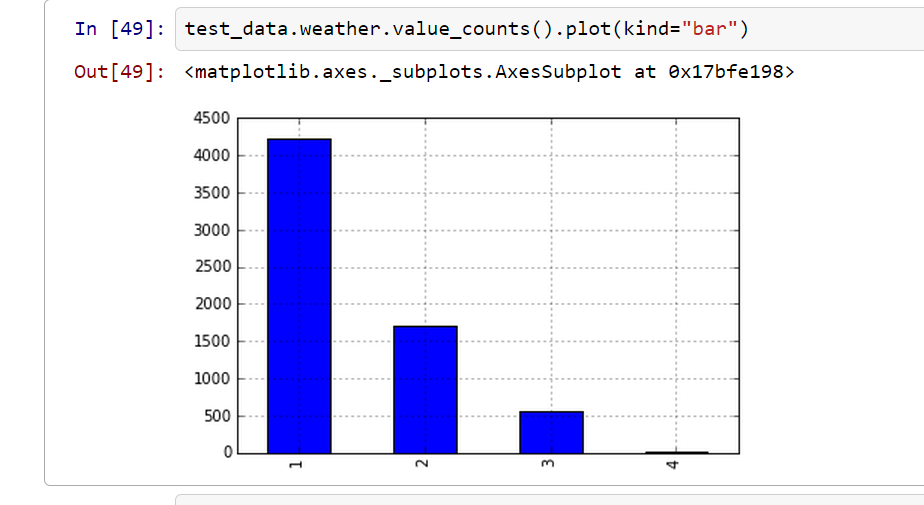
count – number of total rentals

This data is within the span of 2 years. The training dataset consists of the data for the first 19days and test dataset consists of the remaining days in a month. Hoping that year, month, day of week and hour will be of great importance, I have developed a model.

Looking ahead, we anticipate that the year, month, day of week, and hour will serve as important features for characterizing the bike demand at any given moment. These features are easily extracted from the date, time formatted-values loaded above. In the following lines, we add these features to our Data frames.

**Estimated data Analysis:**

I have checked for missing values for all the variables in my training and test datasets and found none. I have produced the bar graphs for each and every attribute in the dataset and found the mean, standard deviation, min and max values for few of them. A sample of the bar chart is below:



**Post Data Analysis:**

I have performed the post data analysis for the model. I have taken the output log which is obtained after running the machine model. The attributes are datetime and count. I have checked if there are any missing values in the output log for both the attributes and found none. Mean, Standard deviation, min and max values are not necessary since the values are continuous. I produced the bar chart for datetime attribute, which is not so clear. The reason being the column might have string as well as float values.

**Machine Learning:**

The machine learning model I have used is random forest regression to check the accuracy of the model. Using the cross validation method, I have predicted the accuracy for the casual bikers and registered bikers and found the accuracy to be 63$ and 69% respectively. Finally I have generated the output data frame with datetime and count attributes. The accuracy can still be improved but considering the current time I could only get near to 70% which I feel is decent.

**Explanation about the Bike Sharing Demand:**

If we consider the weather conditions, more bikes can be rented if we have good weather. Season also plays a major role in determining the demand for the bikes, it may be casual bikers or registered bikers. If the season is spring, due to snow fall, the bikers will not be willing to go for a ride what so ever. So, we see very less or no demand for bikes during spring. When it comes to fall, bike lovers will mostly freak out and try to get some refreshment in hot sun. So, obviously more demand for the bikes and they can be rented heavily. Temperature is based on the region where the bikes are being rented. So people who live in less humid conditions get more bikes and people who live in more humid conditions mostly might not prefer. So, over all in my view season and weather are the categorical variable that gives us the exact results of the model. The other variable also plays important roles but not as much as these two.

**Advantages of the model:**

I have checked the accuracy for both casual and registered users. Comparing both, I found this is more accurate for the registered users which is around 70%.

**Disadvantages:**

I agree that the accuracy is only 70% and can be improved. This is the best model that I could build right now.

**Feedback:**

Finally I would like to say, this is one of the best courses that I have opted for this semester and no regrets. I am not at all aware of the subject at the beginning and now I can say I knew some part of it and there is a chance for me to excel in this field if I wish to choose after my graduation ☺ Thanks a lot to you Mike for all the valuable material that you have shared and being more friendly to everyone of us. Kudos to you.