University at Buffalo
The State University of New York

University at Buffalo
School of Engineering and Applied Sciences

Exploratory data analysis and Modelling of Bike rental forecast

Project Guide:

Professor Mohammad Zia

Submitted by:

Group 16

Vikas Deep Shukla

ACKNOWLEDGEMENT

We would like to thank to the course instructor Professor Mohammad Zia for letting us undertake the Project, reviewing our work throughout the process of conducting the project and for providing knowledge and material necessary for the project.

Sincerely,
Vikas Deep Shukla

Abstract

Dataset is a Washington City Bike rental Data of 2011 and 2012. Bike rental data set is a breakdown of every feature which effects bike rental system. In dataset, each record represents a trip in Washington by Date, Season, Year, month, hour, holiday, weekday, working day, weather, temp, humidity, windspeed, casual, registered. The objective of the project is to predict total rentals per hour for the 2011 and 2012. Determine and plot features which have high and low correlation with response variables. Plots graphs to understand how data is spread across features and with respect to response variable. These reports/visualizations can be used by the public to know the bike rental peak timing. And these reports can be used by the bike companies to improve inventory to meet the demand.

Objective

Predict total rentals per hour for the last quarter of 2012. Determine and plot features which have high and low correlation with response variables. Plots graphs to understand how data is spread across features and with respect to response variable.

Process

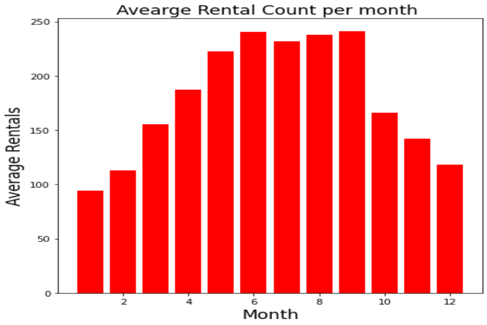
CSV to SQLdb

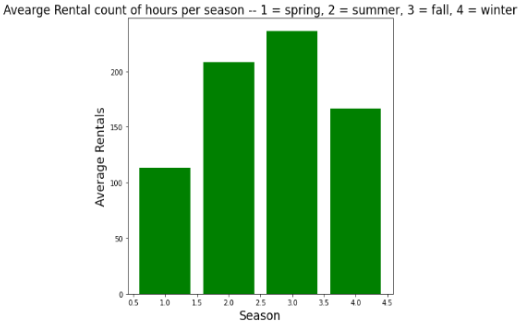
We have three tables – days, weather, and rentals. And we have joined on the basis of Instant column.

- ## Exploratory data analysis

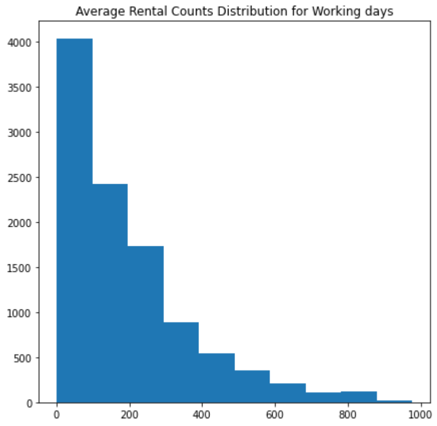
This is an Exploratory Data Analysis for the Bike rental dataset. The data comes in the shape of 15211 training observations. Sanity of the data - there seems to be no NA values in this dataset.

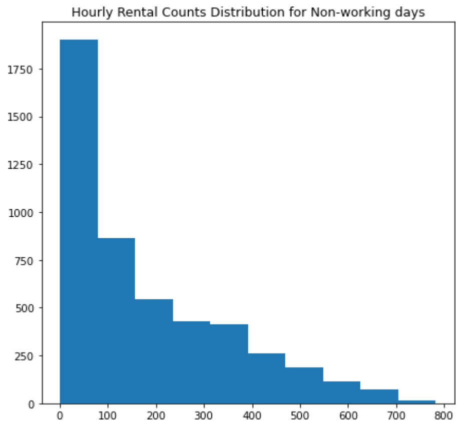
Inference:



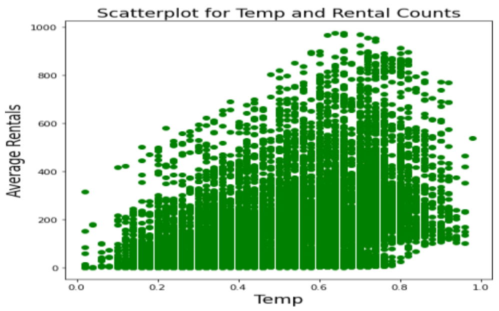


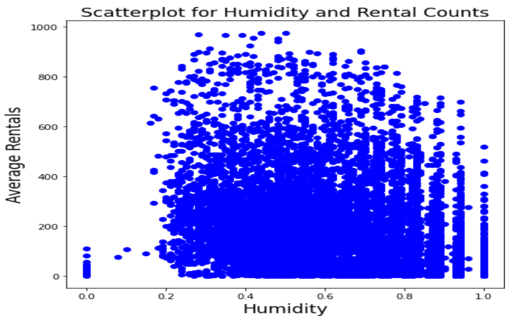
From the barplot, we can observe a spike in rentals in month of July to September. Spring and summer weather have maximum rentals whereas lowest in spring. Moreover, we have noticed a sudden spike in rentals between 4 PM to 7 PM.





For temperature vs count we observe an increasing trend and it is maximum in between 0.7-0.8. Further, windspeed and count has decreasing trend, but not much towards very high windspeed and we can't comment due to less data. For weathersit (clear, few clouds, partly cloudy) – plot is bit left skewed in weathersit 2 and 3 which is due to less count and for weathersit 4 we can't comment due to less observations.





Modelling (Classifying problem)

- Season is basically work as a dummy variable. And for categorical variable, there is no proper relationship exists, so the integer encoding is not enough. Therefore, one hot encoding can be applied to the integer representation.

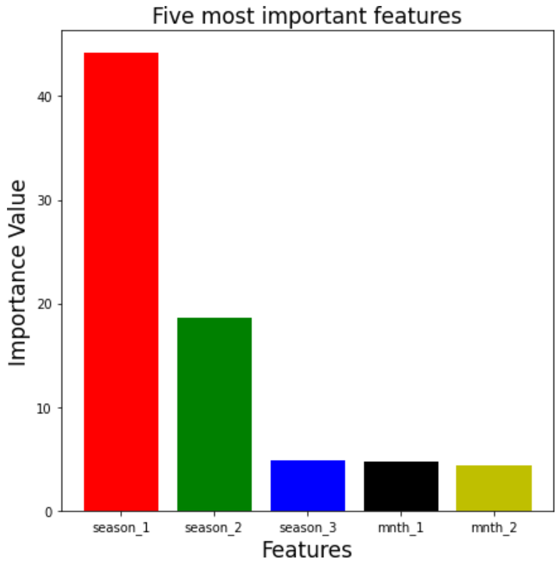
- After fetching data :

In test data we don’t have count value , so we can’t compare on this basis. So, we have selected the test data from train data by shuffling. And on the basis of train data, we have classified test data as well.

- Applied Gradient boosting classifier

  Train Accuracy - 0.91
  Test Accuracy - 0.88

- Five most important feature:



Conclusion:

- Bike rental counts are higher during summer and fall.
- Bike rental are higher during office opening and closing hours.
- With increase in temperature bike rentals increase and decrease with increase in humidity.
- As weather becomes harsh bike rentals decrease.
- Percent of casual rides are more during nonworking days as compared to working day