**Bike Renting**

**By Sachin Sharma**

**– Oct’2019**

1. **[Introduction](#_bookmark0)** 
   1. [Problem Statement](#_bookmark1)
   2. [Data](#_bookmark2)
2. [**Methodology**](#_bookmark3)
   1. [EDA](#_bookmark4) (Exploratory Data Analysis)
      1. Visualization
      2. Outlier Analysis
      3. Feature Selection
3. Correlation
4. Anova
   * 1. Feature Scaling
   1. Modelling
      1. Model Selection
5. Random Forest
6. Linear Regression
   * 1. Visualizing models
7. Prediction Plots
8. **Conclusion** 
   1. Model Evaluation
      1. Mean Absolute Error(MAE)
      2. Mean Squared Error(MSE)
   2. Model Selection
9. ***Introduction*** 
   1. ***Problem Statement***: The objective of this Case is to Predication of bike rental count on daily based on the environmental and seasonal settings.
   2. ***Data:*** As the dataset given has dependent and independent values, it will come under supervise Machine learning. Our task is to build Regression models which will help us predicting the count of bikes which will get rented depending on the factors provided. Given below is a sample of the data set that we are using for our prediction.

|  |  |
| --- | --- |
| **Variable** | **Explanation** |
| instant | Daily customer index |
| dteday | Date index for both the years |
| season | Season (1:springer, 2:summer, 3:fall, 4:winter) |
| yr | Year (0: 2011, 1:2012) |
| mnth | Month (1 to 12) |
| holiday | weather day is holiday or not (extracted fromHoliday Schedule) |
| weekday | Day of the week |
| workingday | If day is neither weekend nor holiday is 1, otherwise is 0. |
| weathersit | (extracted fromFreemeteo) 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 | Normalized temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale) |
| atemp | Normalized feeling temperature in Celsius. The values are derived via (t-t\_min)/(t\_maxt\_min), t\_min=-16, t\_max=+50 (only in hourly scale) |
| hum (Humidity) | Normalized humidity. The values are divided to 100 (max) |
| windspeed | Normalized wind speed. The values are divided to 67 (max) |
| casual | count of casual users |
| registered | The number of registered users at a given day |
| cnt (Count) | Total Rentals with both casual and registered users |

1. ***Methodology***
   1. ***Exploratory Data Analysis***: Before we proceed to create our model on top of the provided data. It is necessary to do Exploratory Data Analysis to understand the data which will help to create model accordingly. Weather Variables are useful or not, which variables has the impact on model creation and so on.
      1. **Variable Identification:** In Order to understand the data, we need to first, Identifying Predictor (Input) and Target (output) variables. Then, Identifying the data type and category of the variables.
   2. ***Hypothesis***:- On Getting the data and problem statement,

People Hire Bike

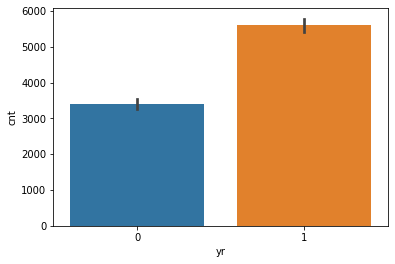
– For Official purpose (Going to Office or Going Back to Home)

-- For Personal Use (Roaming Around) if weather is good (Enjoyable), Season is favourable

On the basis of this Hypothesis we will start Exploring the data and will dive deep into the data to analyse about the Bike Hiring.

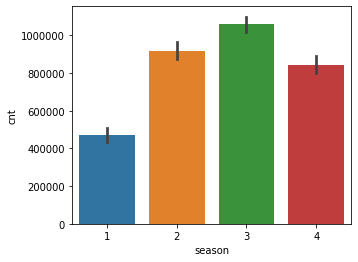
* + 1. ***Visualisations***

1. First of all we will see that the bike hiring trend w.r.t year



Bike hiring has increased over the year. People are hiring bike more than the last year.

1. Now in Which season people tend to hire bike more (1:springer, 2:summer, 3:fall, 4:winter)

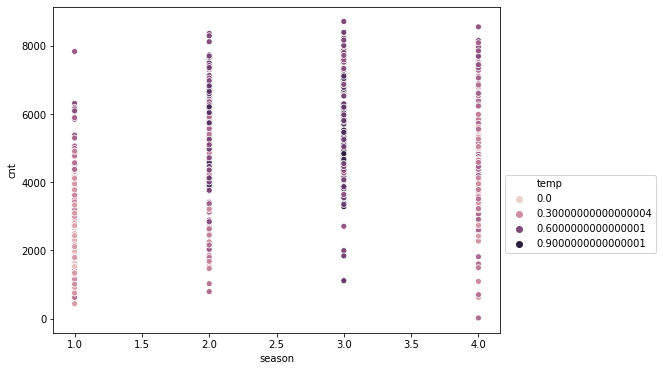


**During Fall Season (Neither Hot nor Cold) – Pleasant Season. Bike Hiring is more.**

season

1-471348 2-918589 3-1061129 4-841613

1. **Now let’s see what is the effect of temperature on bike hiking w.r.t the seasons**

****

**When temperature is of moderate range bike hiring is high.**

1. **Checking the count working days Vs non-working day**

****

**This Graph shows that in working days people hire bike more, to commute to office-home**

**But if we check the count of weekend and weekdays and check the proportion**

**🡺data[(data['holiday']==0) & (data['workingday']==0)]['cnt'].sum() # WEEKEND**

921834

**🡺** **data[(data['holiday']==0) & (data['workingday']==1)]['cnt'].sum() # WEEKDAYS**

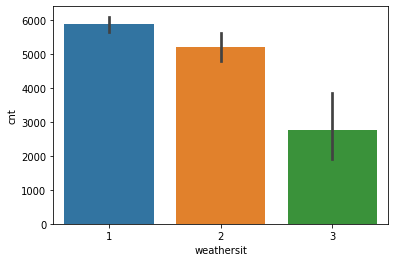
2292410

**Now lets check the proportion**

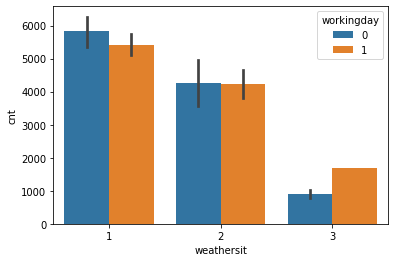
### --- On comparing Both the counts it seems like proportaional because 2292410/921834 = 2.5 approx

### and 5days/2days = 2.5

1. **Checking the impact of weather**

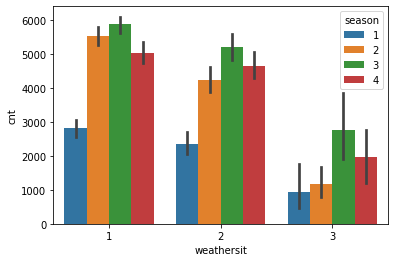
****weather is Clear, Few clouds, Partly cloudy Bike hiring is High

With weather lets check the working day effect

****

**Above Graph is showing that during pleasant weather on non-working day bike hiring is high**

1. **Lets check the season wise effect of weather on bike hiring**

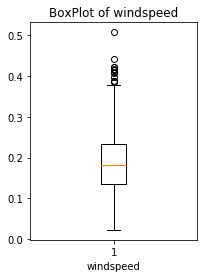
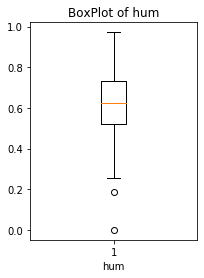
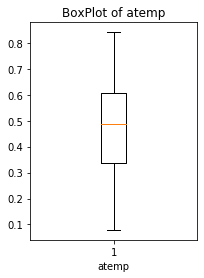
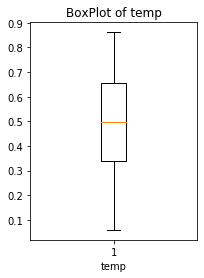
****

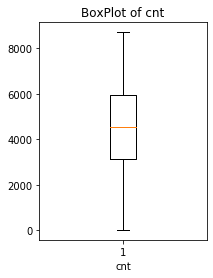
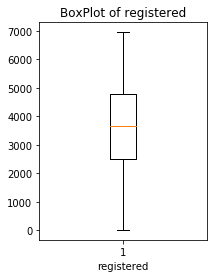
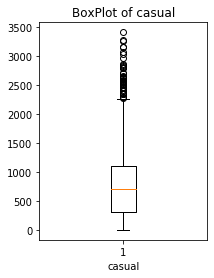
**In fall season when weather is clear bike hiring is high.**

**Hence on visual analysis of the data we can easily conclude that , People hired the bike more in Pleasant Season (Fall ) with weather condition is clear. That means People hired more bike for enjoying the ride with great enjoyable weather and season conditions.**

**2.1.2 Outlier Analysis**

the presence of extreme values which can effect the analysis of the data are the Outliers. The best method to check the presence of outliers is visualization using BOX-Plot

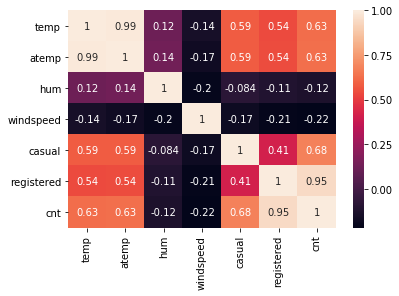


****

**2.1.3 Feature Selection**

There can be few features we can carry same information, those features will be Redundant in the model. So before developing model we will be selecting the features which are useful

Correlation Analysis: Using Heat map we will be analysing about the co-relation of Continous-variables among each-other.



**Anova Test: Analysis of Variance**

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. As our target variable is numerical we will use ANOVA for feature selection technique to see whether any categorical variable is related to target variable. The higher the variance between the variables, the less likely that they are related (or correlated). The result of anova is as follows:

season

df sum\_sq mean\_sq F PR(>F)

data[i] 3.0 9.505959e+08 3.168653e+08 128.769622 6.720391e-67

Residual 727.0 1.788940e+09 2.460715e+06 NaN NaN

------Table ends here------

------Table Starts here------

yr

df sum\_sq mean\_sq F PR(>F)

data[i] 1.0 8.798289e+08 8.798289e+08 344.890586 2.483540e-63

Residual 729.0 1.859706e+09 2.551038e+06 NaN NaN

------Table ends here------

------Table Starts here------

mnth

df sum\_sq mean\_sq F PR(>F)

data[i] 11.0 1.070192e+09 9.729021e+07 41.903703 4.251077e-70

Residual 719.0 1.669343e+09 2.321757e+06 NaN NaN

------Table ends here------

------Table Starts here------

holiday

df sum\_sq mean\_sq F PR(>F)

data[i] 1.0 1.279749e+07 1.279749e+07 3.421441 0.064759

Residual 729.0 2.726738e+09 3.740381e+06 NaN NaN

------Table ends here------

------Table Starts here------

weekday

df sum\_sq mean\_sq F PR(>F)

data[i] 6.0 1.765902e+07 2.943170e+06 0.782862 0.583494

Residual 724.0 2.721876e+09 3.759498e+06 NaN NaN

------Table ends here------

------Table Starts here------

workingday

df sum\_sq mean\_sq F PR(>F)

data[i] 1.0 1.024604e+07 1.024604e+07 2.736742 0.098495

Residual 729.0 2.729289e+09 3.743881e+06 NaN NaN

------Table ends here------

------Table Starts here------

weathersit

df sum\_sq mean\_sq F PR(>F)

data[i] 2.0 2.716446e+08 1.358223e+08 40.066045 3.106317e-17

Residual 728.0 2.467891e+09 3.389960e+06 NaN NaN

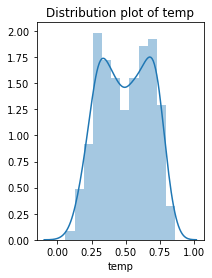
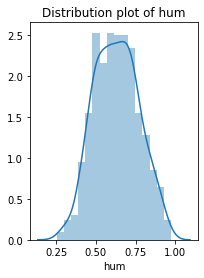
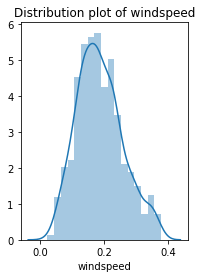
------Table ends here------

**Using the result of feature selection technique we will be dropping the**

**Holiday, workingday, registered, atemp, casual variables as they contain redundant information, which will not be using to create the model.**

**2.1.4 Feature Scaling**

Scaling of features is done to bring the data in a standard scale. to check scaling is required or not in the data we can analysis the data variance

**Data already seems to be distributed hence no need for Normalisation .**

**3 Modeling**

**Random Forest** Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree. It means to build each decision tree on random forest we are not going to use the same data.

MAPE: 12.619964720683937

R-Sq: 0.9144359104973715

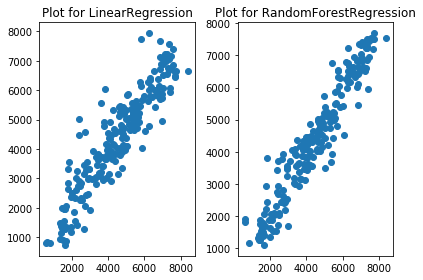
**Liner Regression**

Linear Regression is one of the statistical methods of prediction. It is applicable only on continuous data. To build any model we have some assumptions to put on data and model. Here are the assumptions to the linear regression model.

MAPE: 15.377919361757925

R-Sq: 0.8385211149370677

* 1. **Model Selection**

****

**Random forest model is best fitted among the two.**

**-R sq value is better**

**-Plot shows uniformity**

**Conclusions:** On Exploring the data we can conclude that the bike hiring is more in favorable conditions of enjoyable Season and nice weather conditions.