**Prediction of Bike Rental Count**

**Pravin Chaudhary**

**7-10-2019**

1. **Introduction**
   1. **Problem Statement**
   2. **Data**
2. **Methodology**
   1. **Pre – Processing**
   2. **Outlier detection & removal**
   3. **Relationship of features against Bike count**
   4. **Feature selection**
3. **Modelling**
   1. **Model Selection**
   2. **Random Forest**
   3. **Linear Regression**
4. **Conclusion** 
   1. **Model Evaluation – MAPE**
   2. **Model Selection**

**Chapter 1. Introduction**

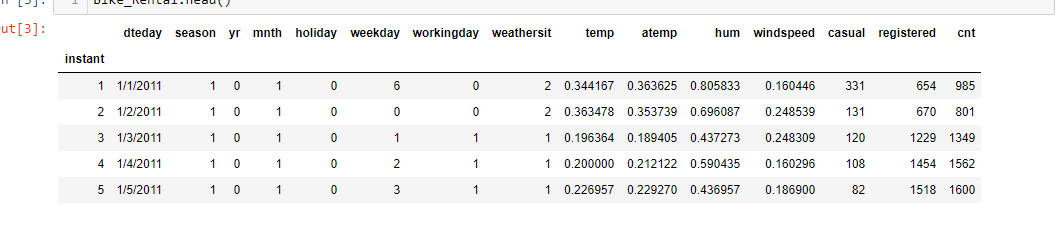
**1.1 Problem Statement :**

Predicting the count of the bike rental based on the seasonal and environmental setting.This will inturn help with managing the no of bikes required on daily basis.

**1.2 Data :**

Our aim is to build a regression model which will predict the count of bike required based on environmental and seasonal setting.

Below is the sample of the data which we are using to predict the number of bikes.



As you can see in the table below we have the following 12 variables, using which we have to correctly predict the count of bikes:

|  |  |
| --- | --- |
| Sno | feature |
| 1 | Dteday |
| 2 | Season |
| 3 | Yr |
| 4 | Mnth |
| 5 | Holiday |
| 6 | Weekday |
| 7 | Workingday |
| 8 | weathersit |
| 9 | Temp |
| 10 | Atemp |
| 11 | Hum |
| 12 | windspeed |

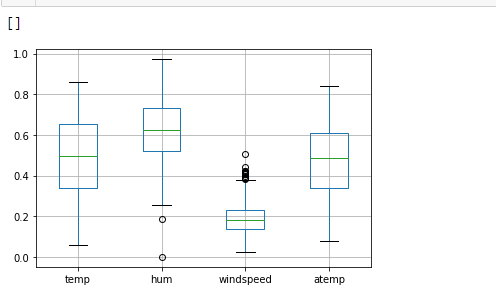
**Chapter 2: Methodology**

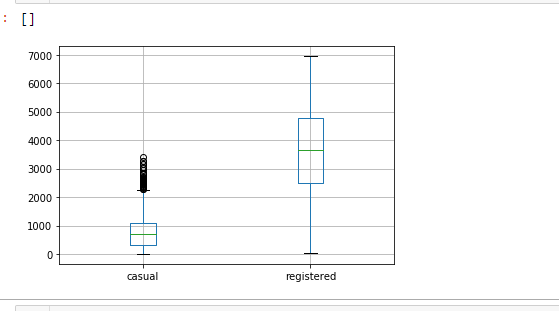
2.1 Pre – Processing

Before we feed the data to the model. We need to do some pre-processing on the data. It includes cleansing the data, finding the relationship of input features against the output label.

Also visualizing the data to find meaningful features and their distrubutions.

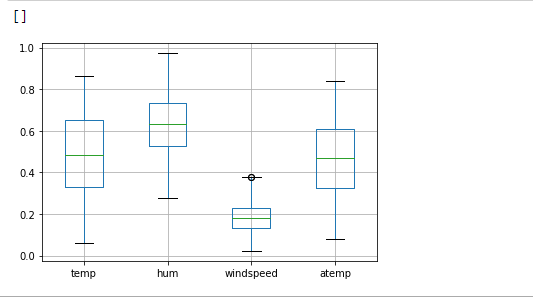
**2.2 Outlier detection and removal**

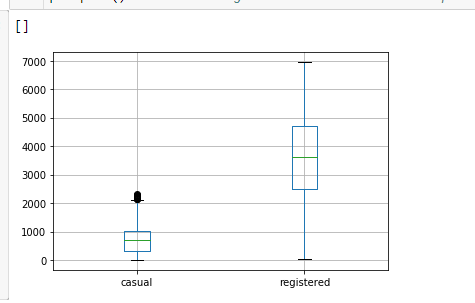




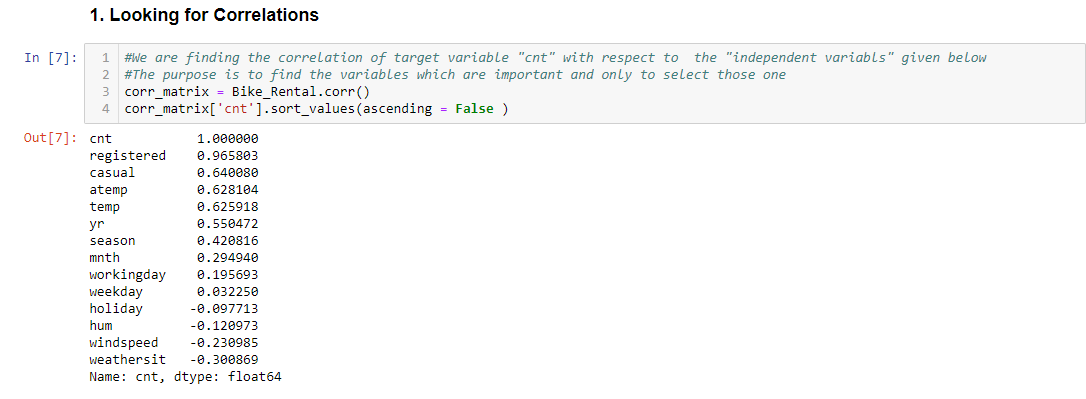
Outliers can be removed by calculating Inter Quartile Range (IQR), minimum, maximum values are calculated for the variables.Any values ranging outside the minimum and maximum value are discarded.

**Box plot of above features after outlier removal :**

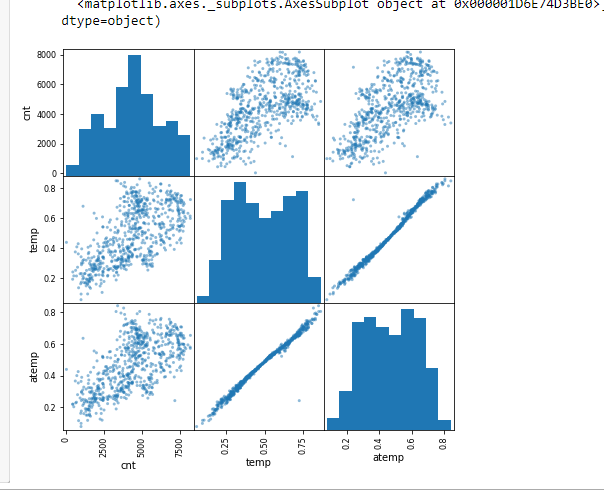




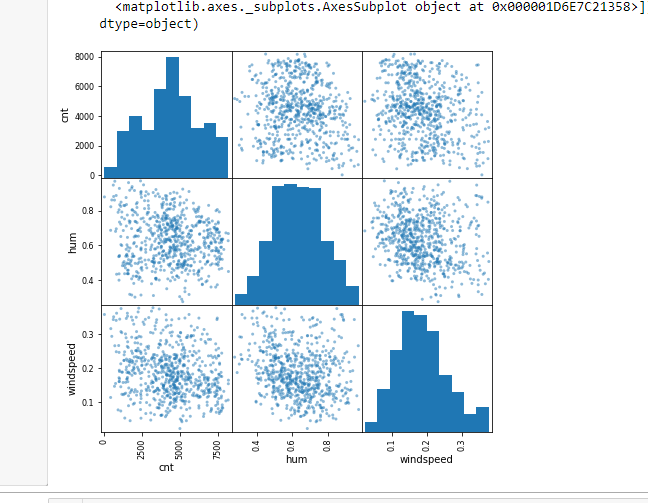
* 1. **Relationship of features against Bike count :**



As per the above fig. temp, atemp, yr, season, mnth, workingday are having positive correlation with the cnt variable whereas hum, windspeed and weathersit are having negative correlation with cnt variable.



**Fig 1.**



**Fig 2.**

The above fig 1. Is having correlation plot for temp, atemp and cnt variable whereas the fig2. Is having correlation plot for hum, windspeed and cnt variable.

**4.4 Feature Selection :**

Feature selection plays an important role in building our model.There might be lot of features in our dataset but not all features are important. We need to only select those features which plays an important role in predicting our output variable.

There are different methods to find the correlation of input features with the output variable. One of them is pearson correlation.

**It is really difficult to select the some specific features out of all bcz :**

**1. The Names of the features are not meaningful which inturn does not helps us in understanding their meaning.**

**2. Also the correlation does not show if a variable is strongly related with the output variable.**

**Chapter 3 . Modelling**

* 1. **Model Selection:**

Our output variable is bike count which is a countinous variable. So we will choose regression models for this problem.The regression model we choose are Random Forest and Linear Regression.

And The error metrices we have choose is MAPE ( Mean absolute percentage error )

* 1. **Random Forest**

Random Forest consist of one or more decision tree. In simple terms it try all possible combinations and gives us the best outcome.The more the tree in our random forest the more robust our model and accuracy.There are two type of random Forest model one is for classification and other is for regression problem. We will go with random forest regressor model since our output variable is continous.

* 1. **Linear Regression :**

It is one of the most commonly used model for continuous dependent variable and one more independent variable.If we have one depedent and one independent variable then it is called Simple linear regression whereas if we have more than one independent variable then it is called multiple linear regression.

So here we are choosing multiple linear regression model since we have more than one independent features.

**Chapter 4. Conclusion**

* 1. **Model Evaluation – MAPE :**

Now when it comes to model evaluation we need some error metrices which will measure the error of our models and help us in selecting the best model for our dataset depending upon the error. So here we are using MAPE (mean absolute percentage error).

The model which gives the least error among other model we choose that.

MAPE for random Forest : 11.90

MAPE for linear regression : 16.97

* 1. **Model Selection :**

Now based on MAPE, we find the accuracy of model based on below formulae :

Accuracy : 100 – Error ( MAPE )

So,

Accuracy for Random Forest : 88.09

Accuracy for linear regression : 83.02

Since the Accuracy for random forest is better than linear regression model. We choose random Forest as our model to predict the bike count based on environmental and seasonal setting..