Bike Renting

The objective of this project is to build a Machine learning model to predict Bike Rental count. We have receive the data set of 731 rows with 16 variables, which are

instant: Record index
dteday: Date
season: Season (1:springer, 2:summer, 3:fall, 4:winter)
yr: Year (0: 2011, 1:2012)
mnth: Month (1 to 12)
hr: Hour (0 to 23)
holiday: weather day is holiday or not (extracted from Holiday 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: 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: count of registered users
cnt: count of total rental bikes including both casual and registered

Explotary data analysis

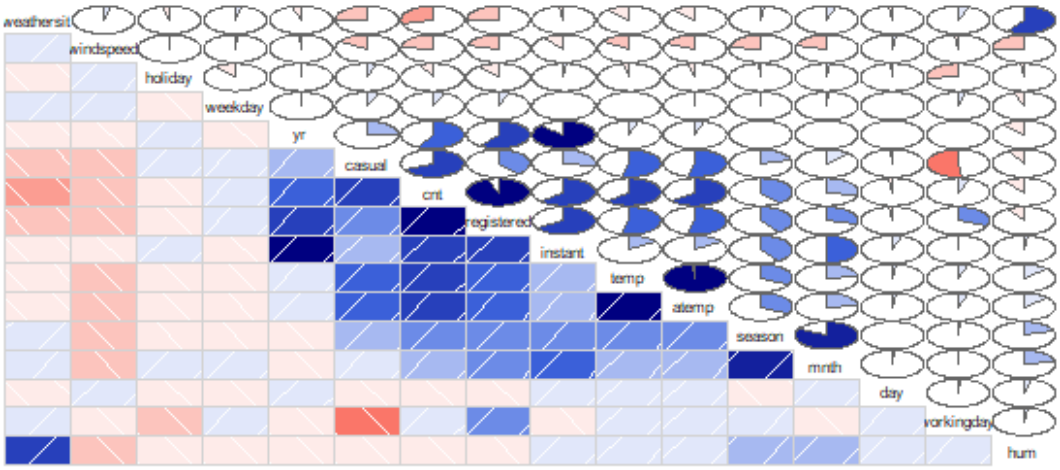
Missing Value Analysis

After running code to find out the no. of missing values, we found out that there are no missing or NA's value in the data set, so we don't have to worry about imputing any data point.

Feature Selection

We have to plot a correlation graph by which we can select features which we can use for model building & can reduce the complexity of the model.

Correlation Plot



-atemp & temp are highly correlated with each other,so we can drop atemp to avoid multicollinearity.

-registered and casual are the subset of the count and they are correlated.

-humidity is highly negatively correlated with count, which shows humidity can be a good dependent variable.

Outlier Analysis

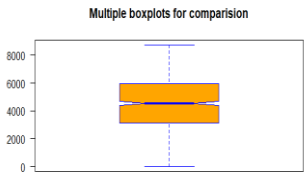
Outliers are the data which fall away from dataset.

Causes of outliers

-Poor Data quality/contamination

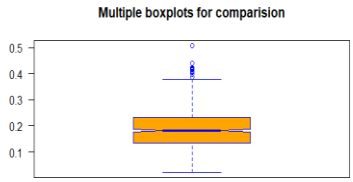
-Lower quality measurement,malfunctioning equipatient etc.

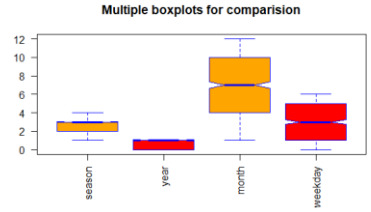
Box Plot

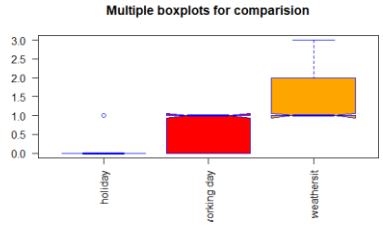


Count

Box plot for windspeed.







As we can see count,season,year,month,weekday,holiday,working day ,weathersit,temp,atemp,doesn't have outliers but windspeed has outliers & in the above correlation plot we found out that windspeed is not that much important in predicting target variable.After converting windspeed range from 1 to 34 which which doesn't looks like malfunction or huge deviation or purely wrong data & in model buiding we are using Random Forest which are less sensitive to outliers.

Feature Engineering

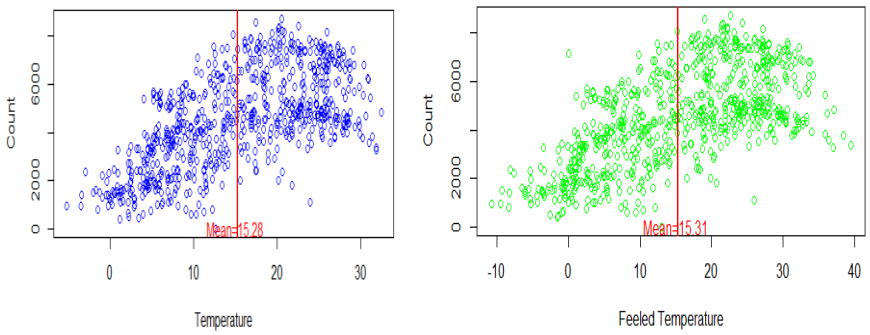
Feature Engineering is a technique where we can extract new feature from existing feature.
We have extracted date from datetime column & converted temp,atemp,humidity & windspeed from normalized data to do analysis.

So,we have 5 new variable i.e day,con_temp,con_atemp,con_windspd,con_humid.

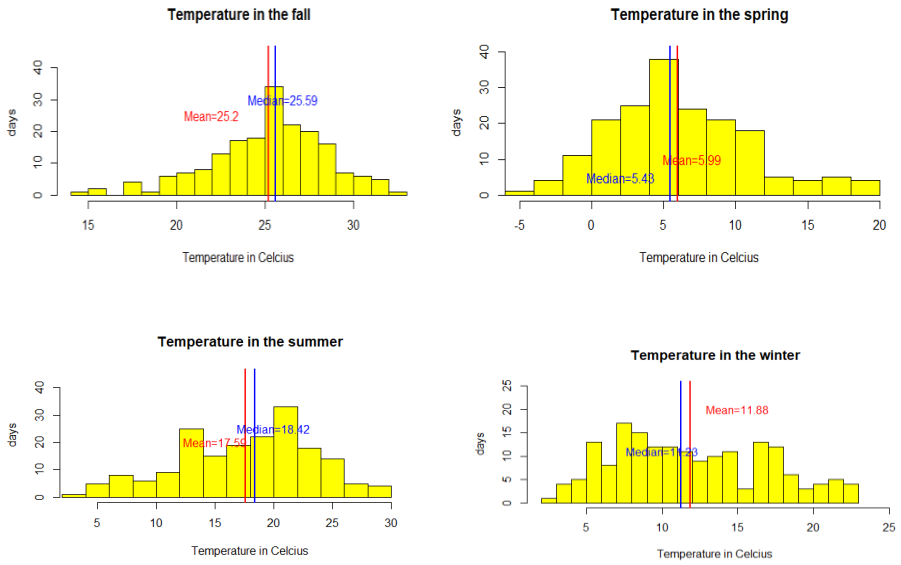
ANALYSIS

-**Relation between temp & atemp.**

We can see in the below scatter plot that the mean of the temperature & feeled temperature doesnot have much difference,so passing two variable to the model will make the model heavy.



-Analysis of temperature with respect to season



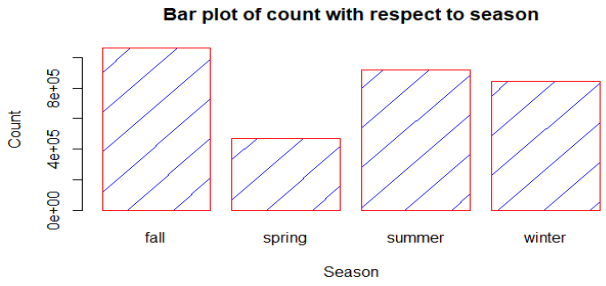
Mean temperature in fall is 25.19
Median temperature in fall is 25.58
SD temperature in fall is 3.32

Mean temperature in spring is 5.99
Median temperature in spring is 5.43
SD temperature in spring is 4.82

Mean temperature in summer is 17.58
Median temperature in summer is 18.41
SD temperature in summer is 5.76

Mean temperature in winter is 11.87
Median temperature in winter is 11.23
SD temperature in winter is 5.06

-How the number of bike rentals is affected by change of season.



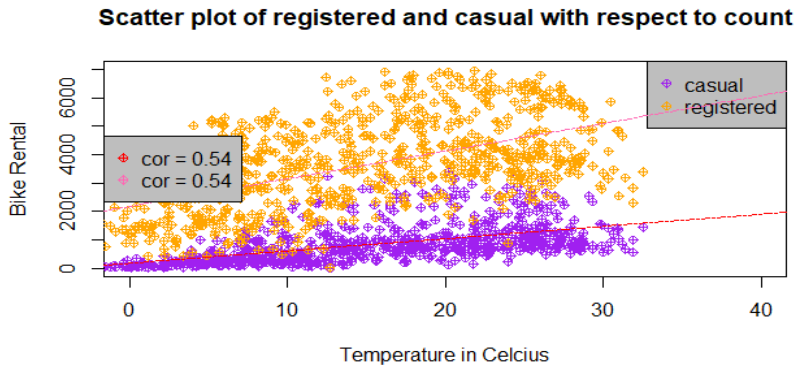
Mean count in summer is 4992.3
Median count in summer is 4941.5
SD count in summer is 1695.9

Mean count in winter is 4728.16
Median count in winter is 4634.5
SD count in winter is 1699.61

Mean count in spring is 2604.13
Median count in spring is 2209
SD count in spring is 1399.94

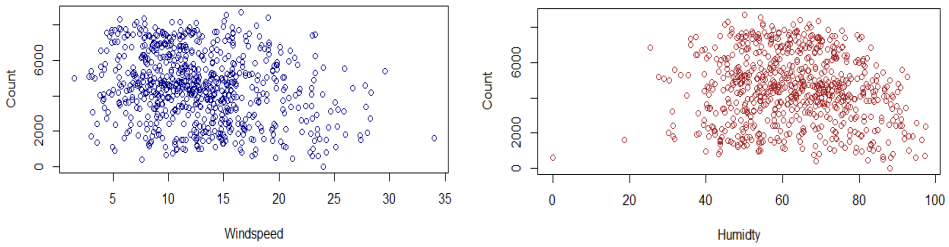
Mean count in fall is 5644.30
Median count in fall is 5353.5
SD count in fall is 1459.80

-scattered plot of register & casual bike with respect to temperature



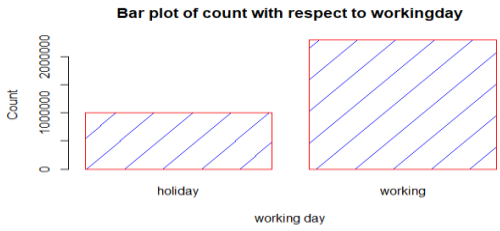
We can visualize that there are more bike rented by registered customer than casual which is a good sign as they are renting on regular basis & most bike are rented when the temperature is between 12 to 28.

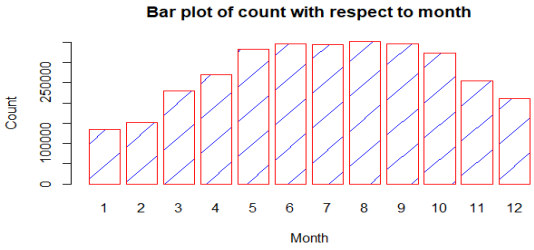
-scattered plot of windspeed & humidity with respect to count

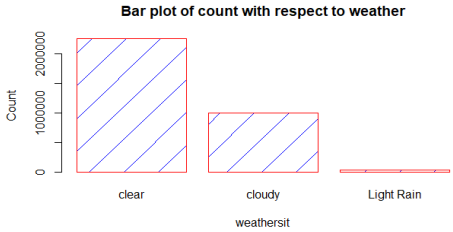


More bike is rented when the windspeed is between 7 to 15 & humidity is between 50 to 75.

-Bar Graph of month & working day







-Conclusion

- We can see that there are more bike rented in working day ,which shows that people are mostly likely using rental service to go to the office or work so there must be more rental point in offices area. But we can't say that there are less bike rented in holidays as there are less holidays compare to working days in a year.

-Highest number of bike rentals are between the month of July & September, whereas the lowest is in January & February.

-Highest number of bike is rented in fall followed by summer with average count of 5645 & 4992 respectively & lowest is spring with average of 2604.

-In the above correlation scatter plot it is visible that the temperature are equally correlated with casual & registered.

-There are maximum no. of bike rental in clear weather than cloudy & zero when it is heavily raining but very less in light rain.

-So we can conclude that the organization should take care between the month of November & February & take some step to increase the count at least in light rain like discount or providing rain cover.

MODEL BUILDING

SAMPLING

I have divided the training data into train data (80%) to build the model & test data(20%) to test the data before deploying to actual test data.

LINEAR REGRESSION

It is a statistical Model,while in machine learning algorithm we save the pattern from historical data,where as in statistic model,it will save numbers in form of coefficients

Linear regression is measured by two parameters

- R squared(It is the proportion of variance in the dependent variable which can be explained by the independent variable.

  Adjusted R squared(Adjustment of the RSquared that penalizes the addition of extraneous predictors to the model.

```
Call:
lm(formula = cnt ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-4022.2  -421.3    16.6   515.1  2952.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1790.088    253.146   7.071 4.49e-12 ***
season        425.007     62.368   6.815 2.40e-11 ***
yr           2058.591     71.918  28.624  < 2e-16 ***
mnth           -4.067     19.872  -0.205  0.83793
weekday        64.282     17.753   3.621  0.00032 ***
workingday    103.118     77.080   1.338  0.18149
weathersit   -618.759     85.305  -7.253 1.32e-12 ***
temp         5010.686    215.821  23.217  < 2e-16 ***
hum          -921.987    343.858  -2.681  0.00754 **
windspeed   -2811.522    487.694  -5.765 1.34e-08 ***
day            -5.560      4.056  -1.371  0.17102
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 855.1 on 573 degrees of freedom
Multiple R-squared:  0.8006,   Adjusted R-squared:  0.7971
F-statistic: 230.1 on 10 and 573 DF,  p-value: < 2.2e-16
```

-Min -4022.2 & Max 2952.6 means that there is no much deviation in error.
- Pr(>|z|) says the amount of significance of each variable in deciding the target variable, the star mark after the value tells that how strong that variable is in deciding the target.
-Null Deviance will tell how well the target variable is predicted by the model with the help of intercept.
-Residual Deviance tell how well the target variable is predicated using Null Deviance & other independent variable.
-AIC(Akaike information criterion) will help to choose better model,less AIC give more accuracy.
-R Squared 0.8006 means 80% variation in the dependent variable is explained by independent variable.
-Adjusted R Squared shows the perfect variation that is 79%.

Why RMSLE as Error Metrics over RMSE?

RMSLE- $$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(x_i+1) - \log(y_i+1))^2}$$

RMSLE is the difference of log function between X predicted value & Y actual value. RMSE value increases in magnitude if the scale of error increase for ex. RMSE for LR in this case in R is 978.67 where as RMSLE is 0.32. RMSLE is also used in this case because overestimation of target variable can be good for this buiseness than underestimation.

RMSLE in Linear Regression.

	R	Python
RMSLE	0.32	0.26

DECISION TREE

-A predictive model based on a branching series of boolean test
-can be used for classification & regression

2 popular decision tree algorithm

C5.0,CART

RMSLE in Decision TREE

	R	Python
RMSLE	0.34	0.32

MODEL TUNING

Model Tuning is the process of maximizing a model's performance without overfitting or creating too high of variance. We have to change the parameter setting which is more favourable(tuned) or less favorable(Untuned). For e.g in KNN we will take no. of neighbors as 3,5,7,9 to see which give better performance. Similarly in Random Forest we will change the no. of trees from 100 to 300,500,700 to get a better model.

RANDOM FOREST

Random Forest is ensemble which consist many number of decision trees, if the data is big we will use Random forest, single decision tree will not able to handle it.

Random forest use bagging method where once an error occurred in a decision tree,that error is fed to next decision tree to improve accuracy .So, this improve the correct prediction & decrease the error rate.

Some Rules

```
[1] "cnt<=3 & season<=0.5 & yr<=1.5 & weathersit<=0.432373 & weathersit<=0.24
875 & temp<=0.4639585"
[2] "cnt<=3 & season>0.5 & yr<=1.5 & weathersit<=0.432373 & weathersit<=0.248
75 & temp<=0.4639585"
[3] "cnt<=3 & yr<=1.5 & workingday<=2.5 & weathersit<=0.432373 & weathersit<=
0.24875 & temp>0.4639585"
[4] "cnt<=3 & yr<=1.5 & workingday>2.5 & weathersit<=0.432373 & weathersit<=0
.24875 & temp>0.4639585"
```

	R Code	Python
RMSLE(n=100)	0.287	0.224
RMSLE(n=300)	0.285	0.220
RMSLE(n=500)	0.286	0.218
RMSLE(n=700)	0.284	0.218

KNN

K-nearest Neighbor predicts the value by checking the distance with from other feature with respect to the Kth value. Applying the K-nearest Neighbors on our data set with n_neighbors=3,5,7,9. Checking which kth value fit best for the respective data set.

	R Code	Python
RMSLE(k=3)	0.56	0.404
RMSLE(k=5)	0.53	0.432
RMSLE(k=7)	0.62	0.451
RMSLE(k=9)	0.69	0.453

Conclusion: From the above error metrics we can conclude that Random Forest will give the best predictions for test data as the RMSLE is lowest.