Capstone Project

Bike Sharing Demand Prediction

Apoorva KR
apoorvargowda1@gmail.com

Problem statement

The contents of the data came from a city called Seoul. A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis for a price or free.The data had variables such as date, hour, temperature, humidity, wind-speed, visibility, dew point temperature, solar radiation, rainfall, snowfall, seasons, holiday, functioning day and rented bike count.The problem statement was to build a machine learning model that could predict the rented bikes count required for an hour, given other variables

Points to discuss

- Data description and summary
- Analysis of categorical variable
- Analysis of numerical variable
- Handling outliers
- Regression plot
- Machine learning algorithms
- conclusion

Data description

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

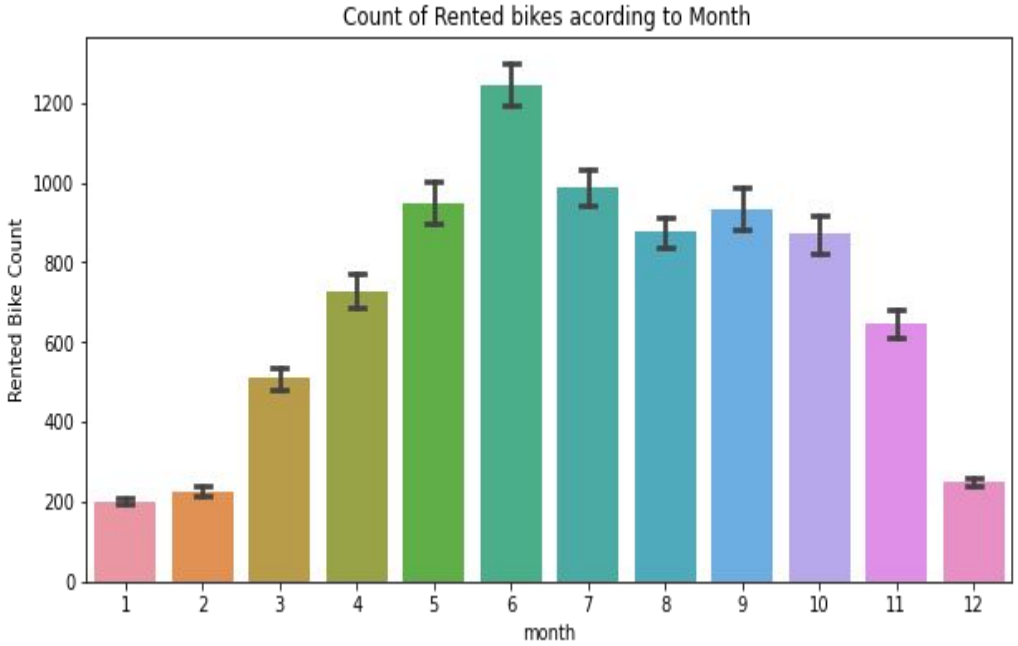
- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm

Data description(cont,.)

- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

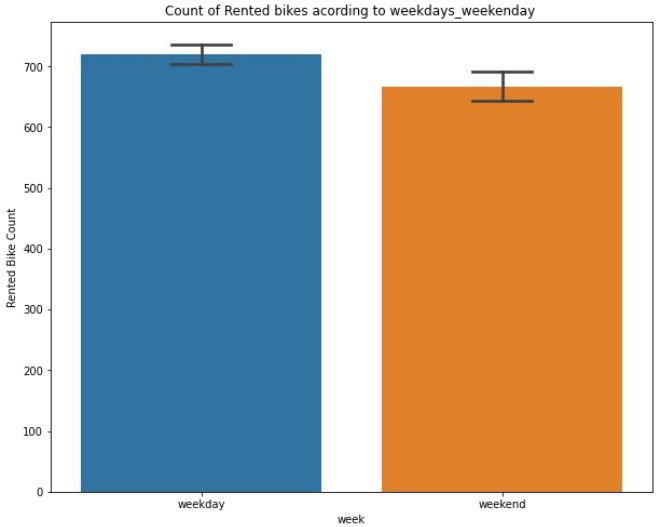
1. This dataset contains 8760 lines and 14 columns
2. Numerical variables - temperature, humidity,wind,visibility,dew point temp, solar radiation,rainfall,snowfall
3. Categorical variables - seasons,holiday and functioning day
4. Rented bike column - which we need to predict for new observations

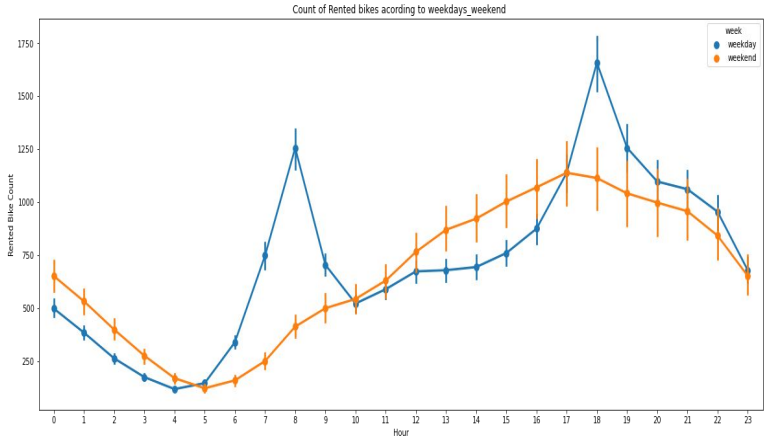
Month



The demand of the rented bike is high from the month 5 to 10

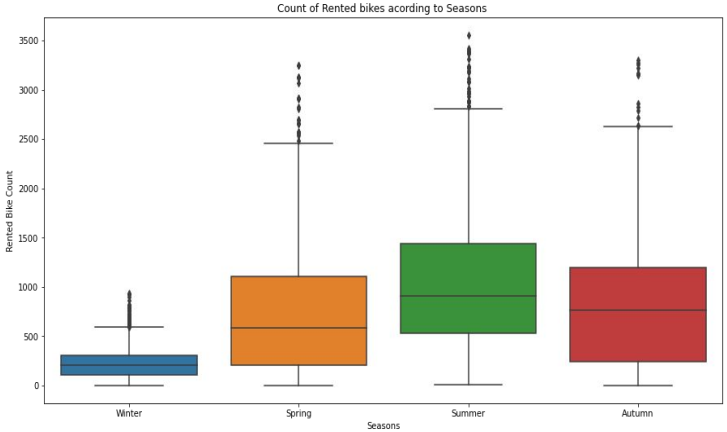
week

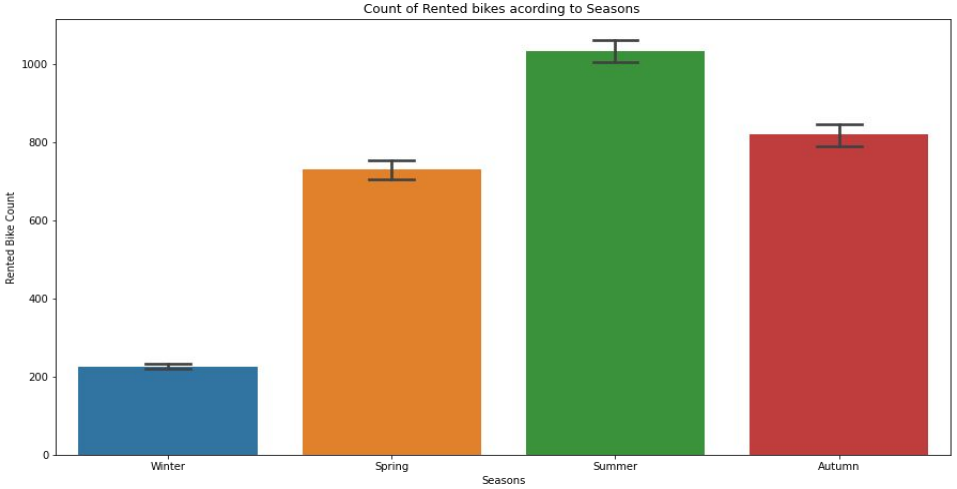




From the above point plot and bar plot we can say that in the week days which represent in blue colur show that the demand of the bike higher because of the office. Peak Time are 7 am to 9 am and 5 pm to 7 pm The orange colour represent the weekend days, and it show that the demand of rented bikes are very low specially in the morning hour but when the evening start from 4 pm to 8 pm the demand slightly increases

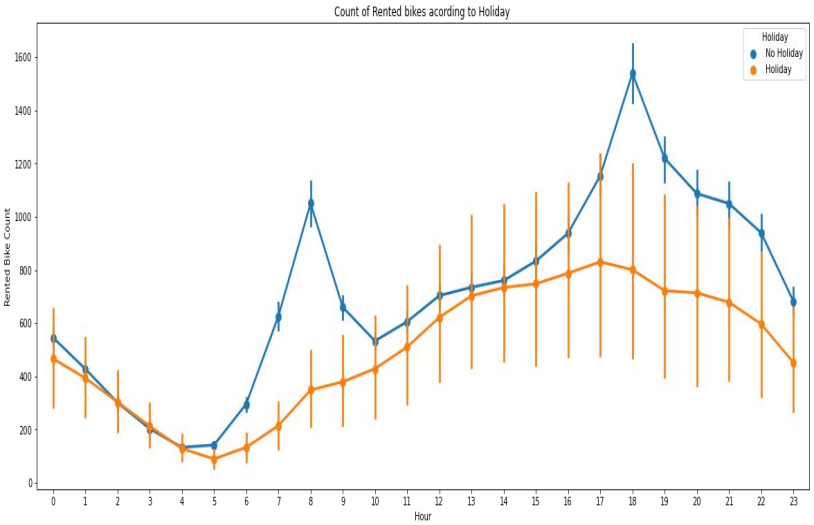
seasons





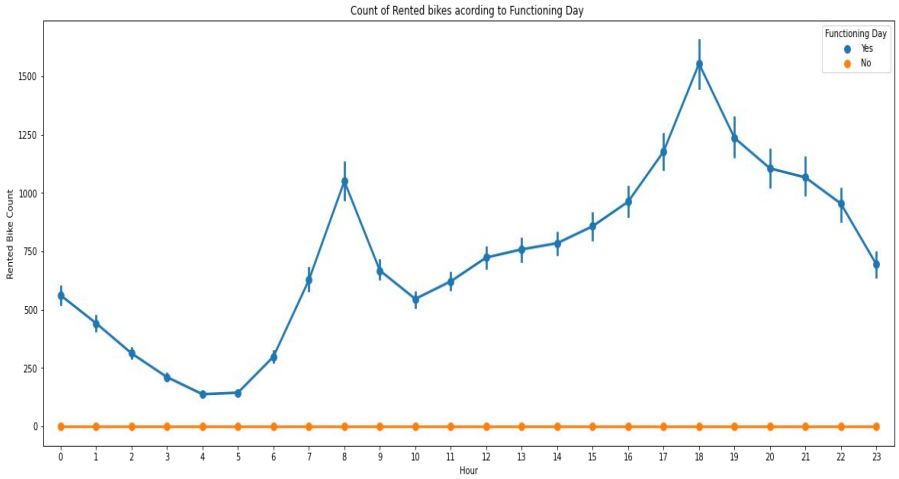
In the above box plot and bar plot which shows the use of rented bike in in four different seasons, and it clearly shows that, In summer season the use of rented bike is high In winter season the use of rented bike is very low because of snowfall.

Holiday



In the above point plot which shows the use of rented bike in a holiday, and it clearly shows that, plot shows that in holiday people uses the rented bike from 2pm-8pm

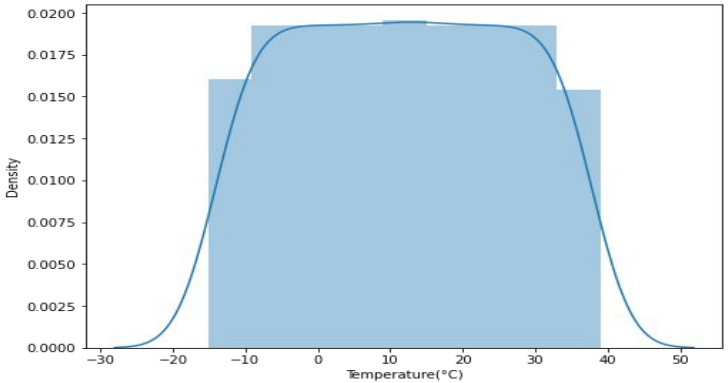
Functioning day



In the above point plot which shows the use of rented bike in functioning day or not, and it clearly shows that, Peoples dont use reneted bikes in no functioning day

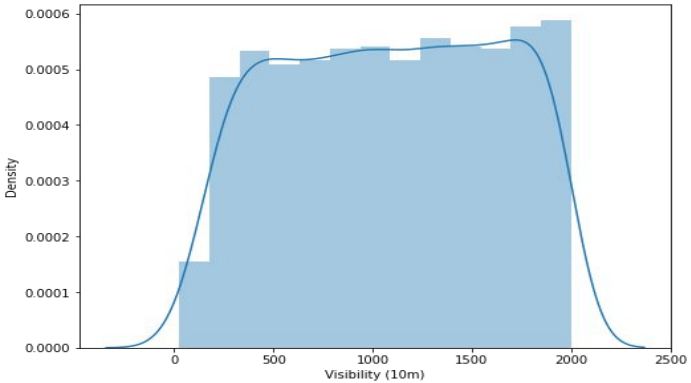
Numerical variables

Temperature



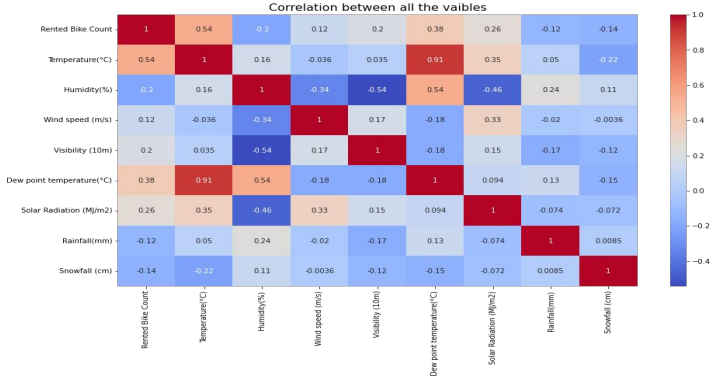
Above plot shows that people tend to rent bikes when the temperature is between -5 to 25 degrees.

visibility

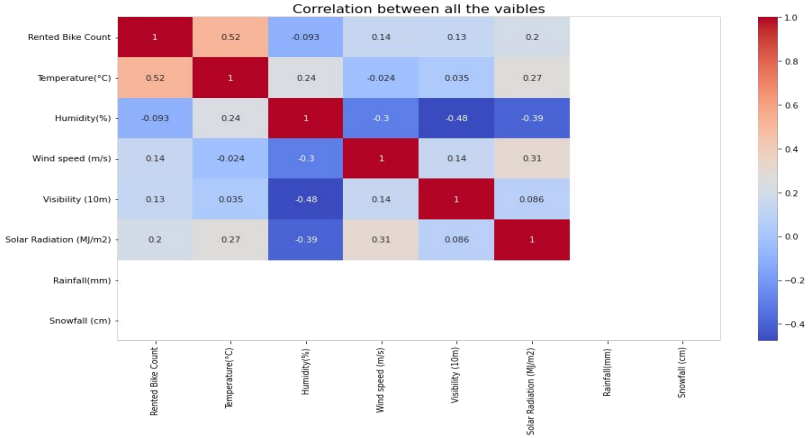


Above plot shows that people tend to rent bikes when the visibility is between 300 to 1700

Heat map

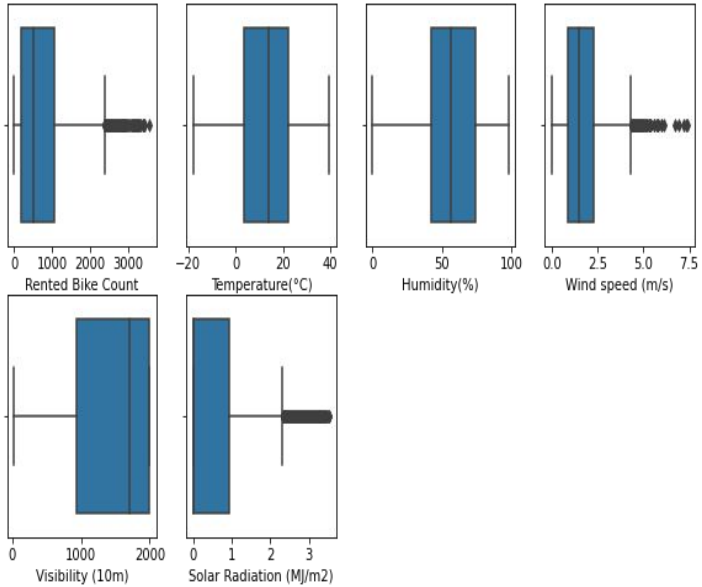


From the above heat map i can conclude that Temperature and Dew point temperature(°C) has the high correlation . we drop this column then it dont affects the outcome of our analysis



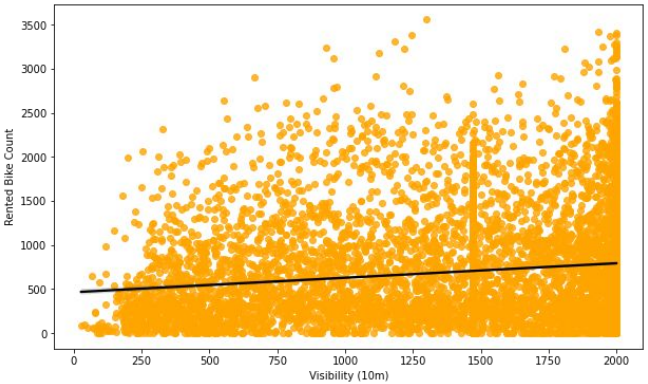
After removing the Dew point temperature

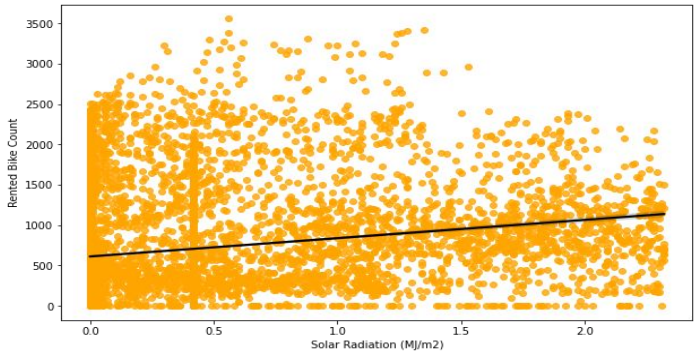
Handling outliers

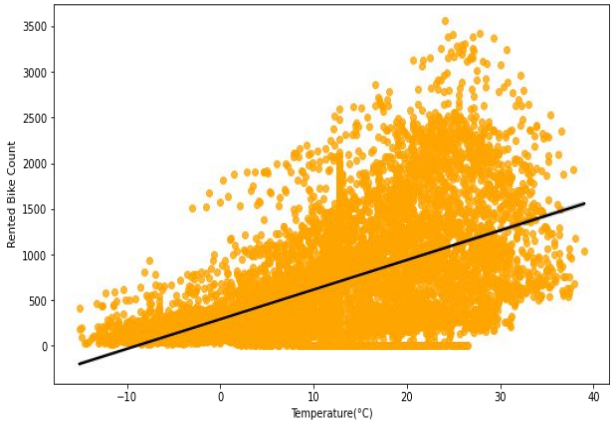


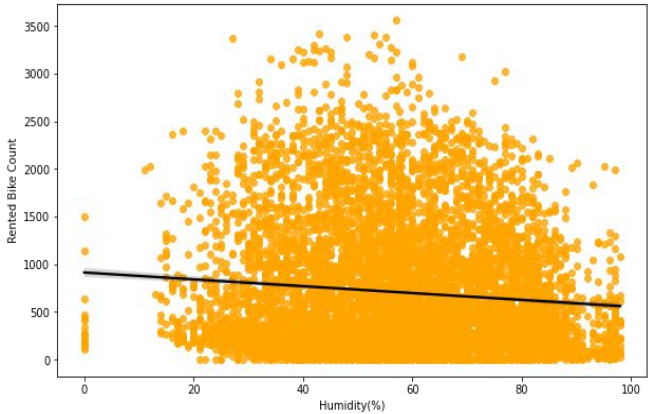
An Outlier is a data-item/object that deviates significantly from the rest of the (so-called normal)objects,The interquartile range (IQR) is the difference between the 75th and 25th percentile of the data. It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outlier

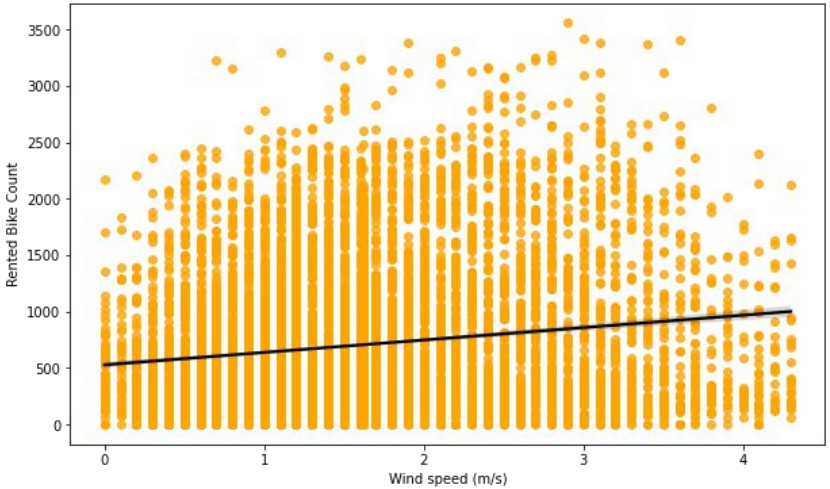
Regression plot











- Temperature,solar radiation,wind speed,visibility are positively related to target variable ,the rented bike count increases with increase of these features

ML algorithms

1. Linear regression
2. Ridge regression
3. Elastic net
4. Decision tree
5. Random Forest regressor
6. SVR
7. Gradient boosting

- Linear regression

```
MSE : 61.15584375724336
RMSE : 7.820220185982192
MAE : 5.923260387972038
R2 : 0.6028600936479797
Adjusted R2 : 0.5979004370931673
```

- Ridge regularization

```
MSE : 60.83384712272495
RMSE : 7.799605574817546
R2 : 0.6148937632652991
Adjusted R2 :  0.6100843884309619
```

- linear regression with elastic net

```
MSE : 61.36957118125049
RMSE : 7.833873319198523
MAE : 5.93382066577593
R2 : 0.6014721692250582
Adjusted R2 : 0.5964951796640391
```

Decision tree

- The r2 score of decision tree is 0.7619882672759375
- the r2 score of decision tree with hyper perameteres tunning is 0.8013257980814106

Random Forest Regressor with GridSearchCV

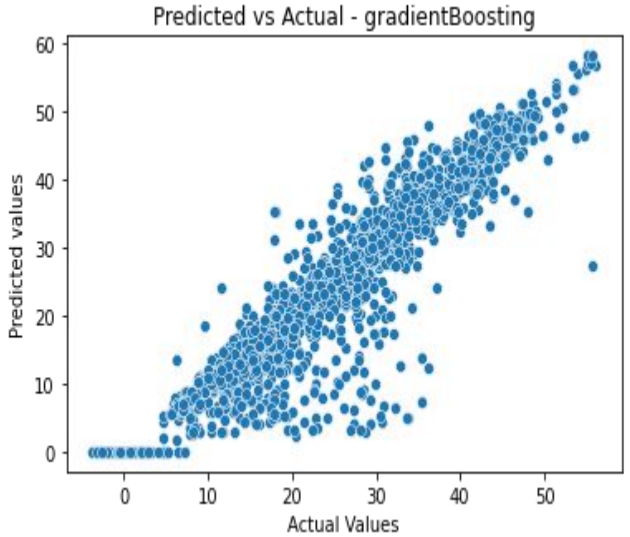
MSE= 7.548469131956553
RMSE= 2.747447748721812
R2_Score_train=
0.9509809996888273
MSE= 19.900384576287564
RMSE= 4.460984709264039
R2_Score_test=
0.8740214111679199

SVR using grid search cv

The MAE of training set = 4.312974309168605
The MSE of training set = 44.525880818380806
The R2_score of training set = 0.7108534025195494
The MAE of test set = 4.674553587050071
The MSE of test set = 46.854036819585005
The R2_score of test set = 0.703392393398722

Gradient Boosting Regressor with GridSearchCV

R2 score of training data: 0.91%
R2 score of test data: 0.869495



conclusion

- Hour of the day holds most importance among all the features for prediction of dataset
- It is observed that highest number bike rentals counts in Autumn/fall Summer Seasons and the lowest in Spring season.
- We observed that the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day
- the top 5 important features of our dataset are: Season_winter, Temperature, Hour, Season_autumn, Humidity
- Peoples dont use rented bikes in no functioning day
- people tend to rent bikes when the temperature is between -5 to 25 degrees
- people tend to rent bikes when the visibility is between 300 to 1700
- for all the above experiments we can conclude that gradient boosting and random forest regressor with using hyperparameters we got the best results