CAPSTONE PROJECT - 2

Bike SharingDemand Prediction
(Supervised Machine Learning regression)
BY

Team Members

Priyvrat Sharma
Richa Pandya

Bike Sharing Demand Prediction

Abstract: Our data set contains Bike sharing prediction information came from city called seoul. It includes information such as date, hour, temperature, humidity, wind speed, visibility, dew point temperature, solar radiation, rainfall, snowfall, seasons, holiday, functioning day and rented bike count. ***In upcoming slide, we will analysis and get insight from the data.***

POINTS FOR DISCUSSION

- Descriptive summary
- Bike sharing analysis by visualization
- Analyze the numerical columns
- Analyze the relationship between Rented_Bike_Count & temperature
- Prepare data for modelling
- Model selection and evaluation
- Hyperparameter tunning
- Conclusion

PYTHON LIBRARIES USED

- NumPy
- Pandas
- Seaborn
- Matplotlib
- Plotly lib

REGRESSION MODEL

- Linear Regression
- Lasso Regression
- Ridge Regression
- Decision tree
- Gradient boosting
- Hyper parameter tuning

PREPARING OUR DATA FOR DEEP ANALYSIS

- Step 1: Overview the data.
- Step 2 : Exploratory Data Analysis
- Step 3 : Analysis of data by visualization
- Step 4 : Numerical columns
- Step 5 : Preparation of model building
- Step 6 : Linear, Lasso, Ridge regression
- Step 7 : Decision tree, Gradient boosting, Hyper parameter tuning
- Step 8 : Conclusion

Data Collection and Understanding

- We had a Seoul Bike Data for our analysis and model building
- The dataset contains weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.
- In this we had total 8760 observations and 14 features including target variable.

Data Description:

Date : year-month-day.
Hour - Hour of he day.
Temperature - Temperature in Celsius.
Humidity - %.
Wind speed - m/s.
Visibility - m.
Dew point temperature - Celsius.
Solar radiation - MJ/m2.
Rainfall - mm.
Snowfall - cm.
Seasons - Winter, Spring, Summer, Autumn.
Holiday - Holiday/No holiday.
Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours).
Rented Bike count - Count of bikes rented at each hour (Target Variable i.e Y variable).

Data Wrangling and Feature Engineering

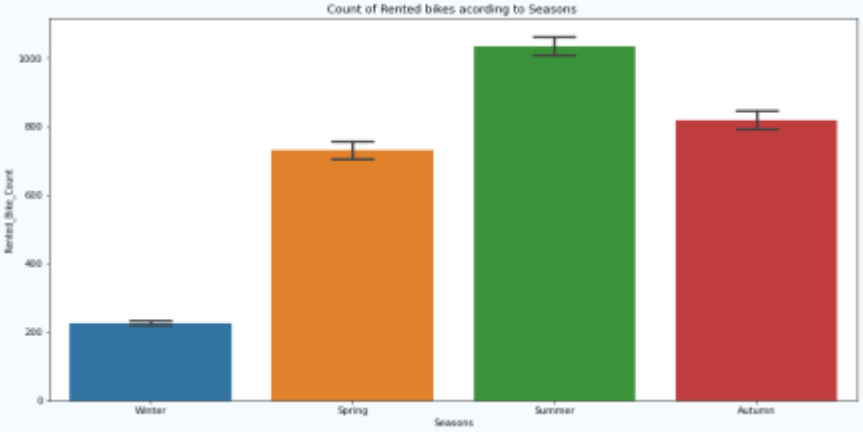
As we know we had 8760 observations and 14 features.

- **Categorical Features:** Seasons, Holiday and Functioning day.
- **Numerical Columns:** Date, Hour, Temperature, Humidity, Wind speed, Visibility, Dew point temperature, Solar radiation, Rainfall, Snowfall, Rented Bike count .
- **Rename Columns:** We renamed columns because they had units mentioned in brackets and was difficult to copy the feature name while working.

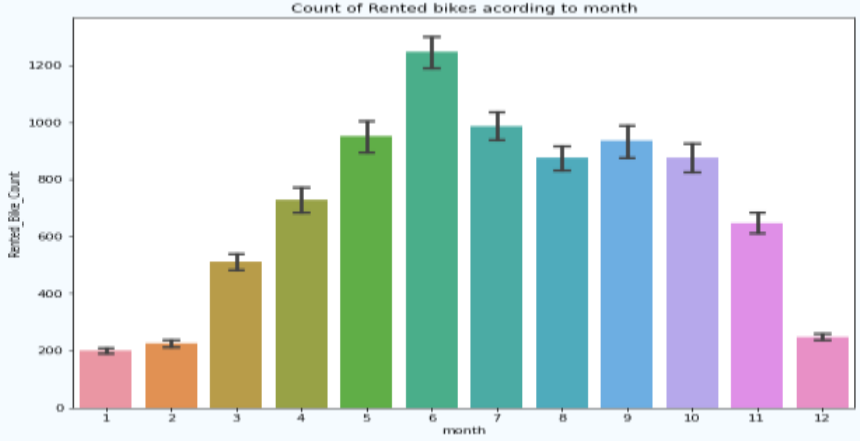
Data Wrangling and Feature Engineering

- We had zero null values in our dataset.
- Zero Duplicate entries found.
- We changed the data type of Date column from ‘object’ to ‘datetime64[ns]’. This was done for feature engineering.
- We Created two new columns with the help of Date column ‘Month’ and ‘Day’. Which were further used for EDA. And later we dropped Date column.

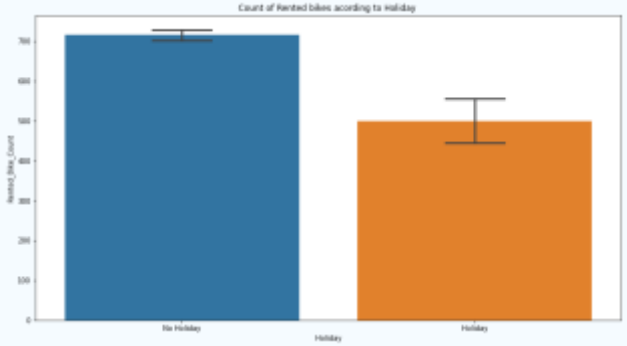
EDA (Exploratory Data Analysis)



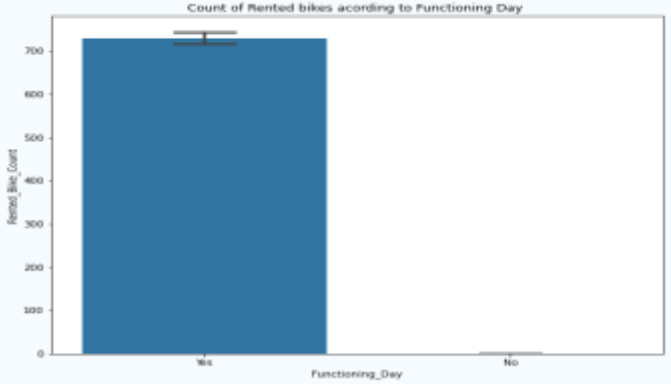
Summer season had the highest Bike Rent Count. People are more likely to take rented bikes in summer. Bike rentals in winter is very less compared to other seasons.



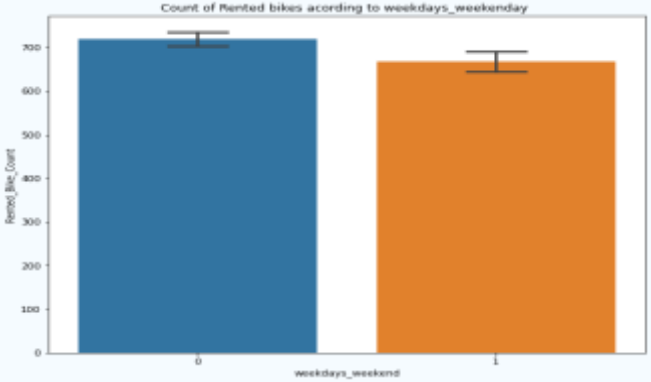
From March Bike Rent Count started increasing and it was highest in June.



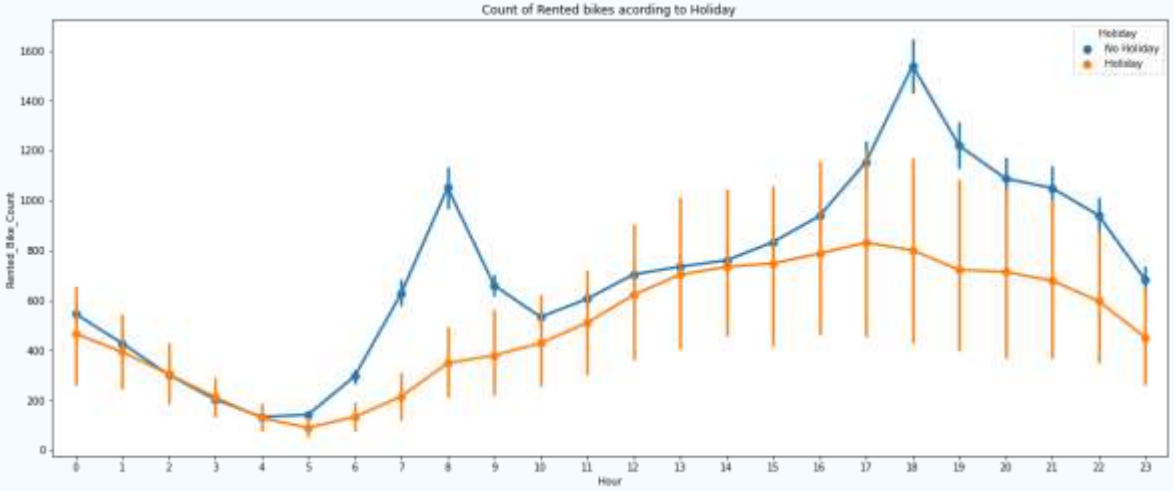
Demand high **on no holiday** which is almost 700 bikes.



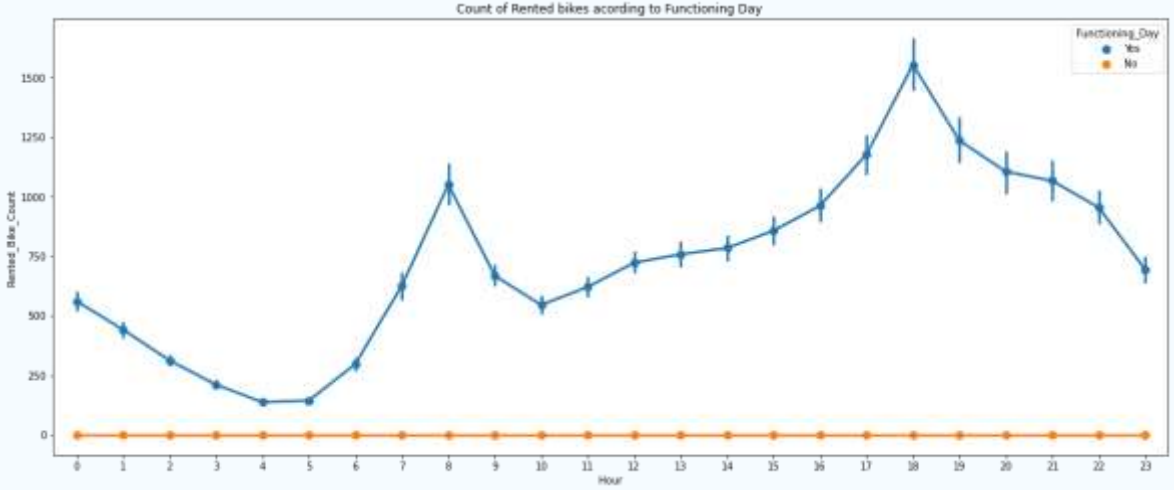
On no functioning day there is no demand & on functioning day more than 700 bikes were rented.



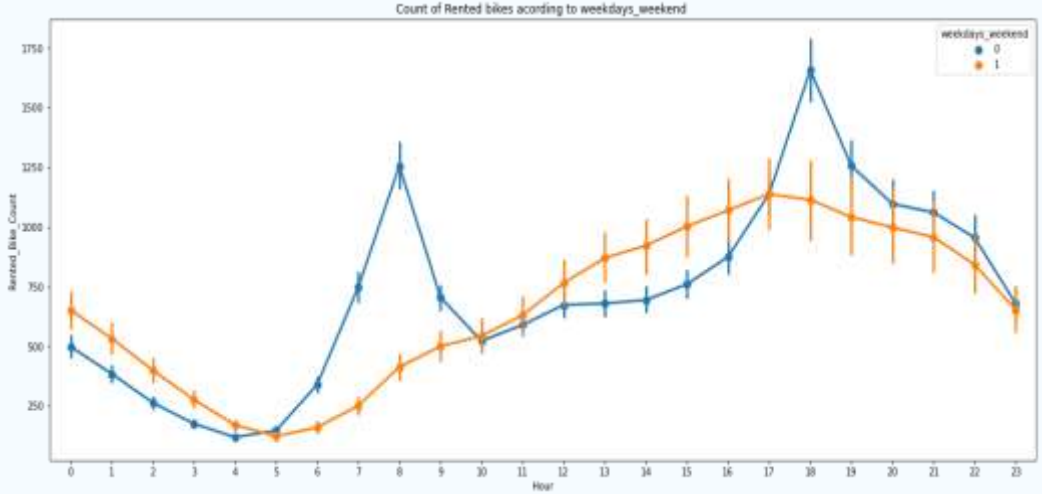
More than 700 bikes were rented **on weekdays**. On weekend, almost 650 bikes were rented.



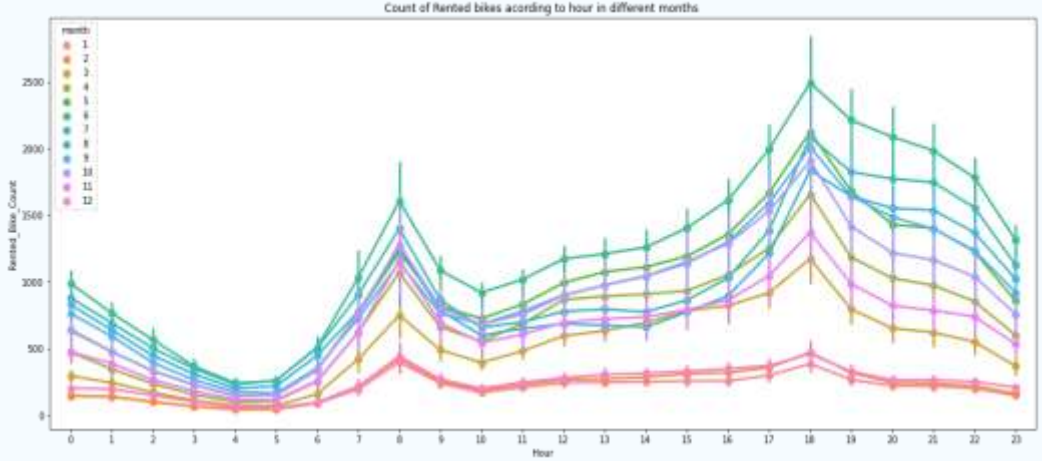
Observations: In this situation, we found that the hourly trend for bike rentals is basically consistent across all possibilities.



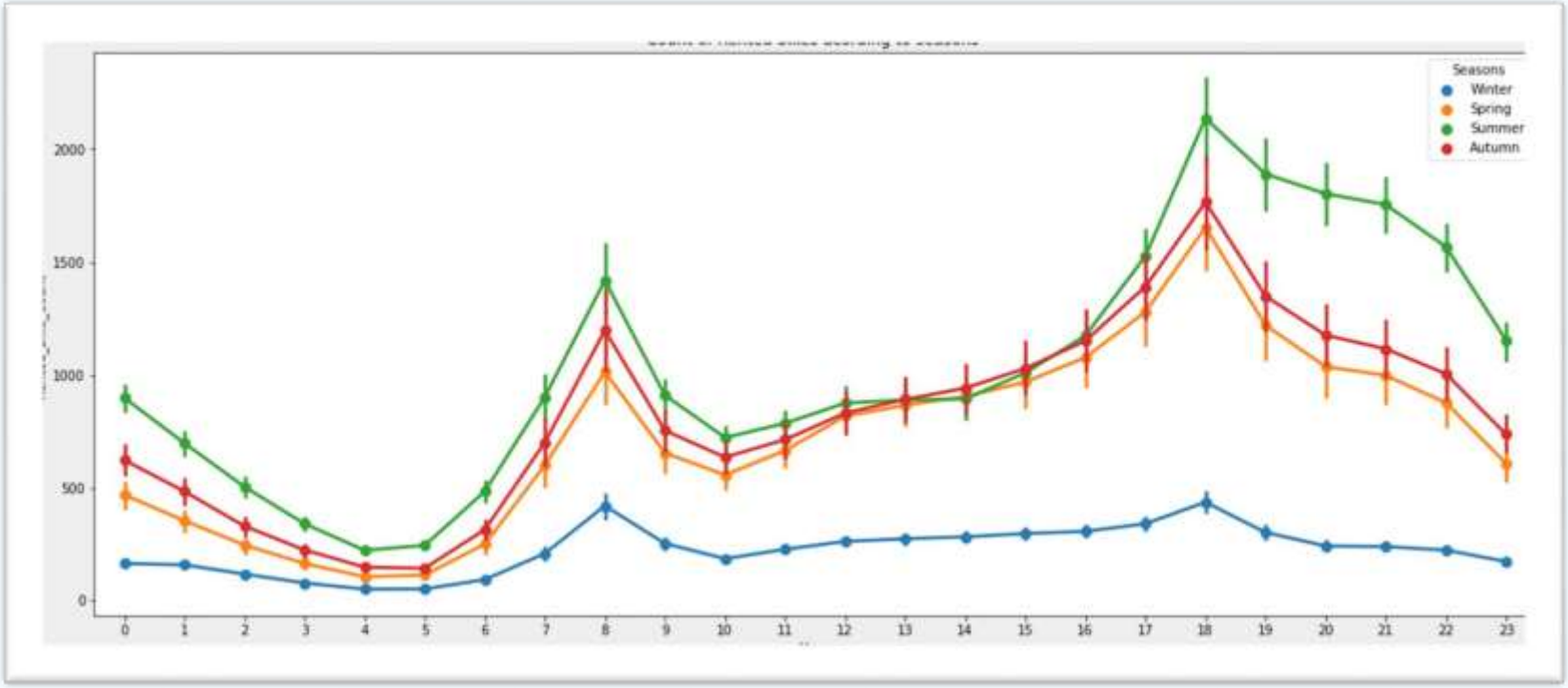
Between 6/7 and 10 AM, there is a sudden rise. Possible reasons for this highest activity on NO Holiday including office/college going hours. However, at holidays, there were considerably fewer bike rentals.



Once again, the peak hours are from 4 to 7 PM. Maybe the mentioned people are leaving for work. (NO Holiday).

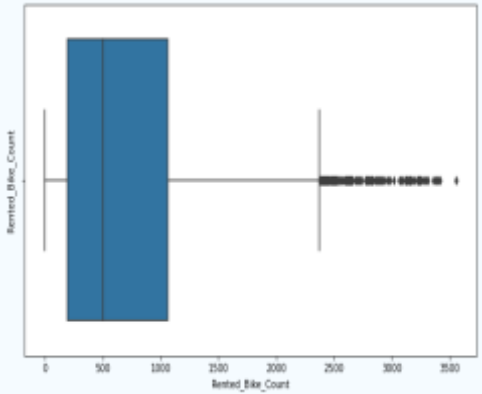


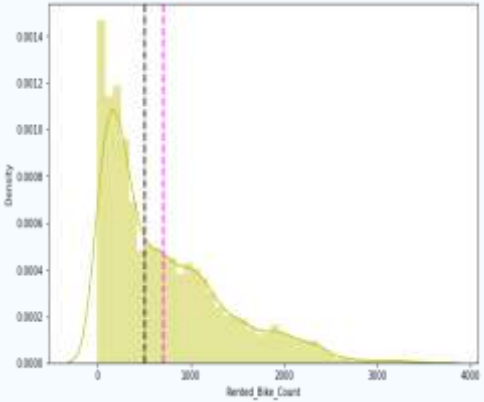
Here the trend for functioning day is same as of No holiday. Only the difference is on No functioning day there were zero bike rentals.

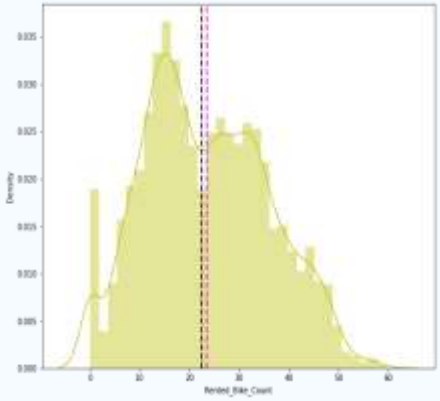


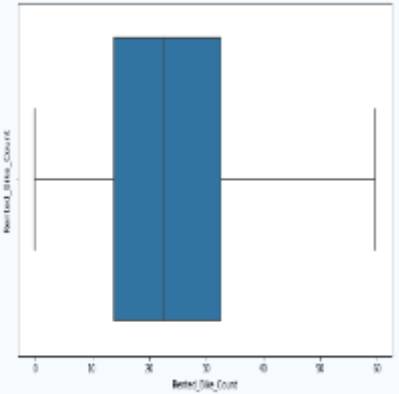
Summer season had the highest Bike Rent Count. People are more likely to take rented bikes in summer. Bike rentals in winter is very less compared to other seasons.

OUTLIER









There are some outliers present and the distribution is slightly skewed.

We applied the square root method to normalise the distribution. The outcomes of normalisation show no outliers.

Model Selection and Evaluation

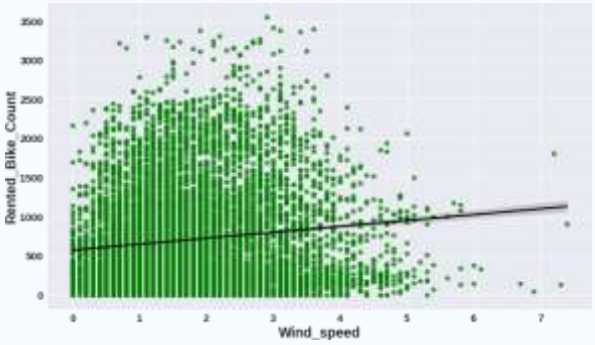
We are trying to determine continuous value because this is a regression problem. We applied the following regression for this models.Linear Regression

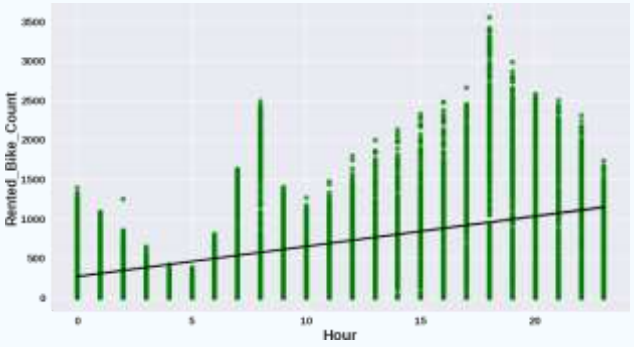
- Lasso regression
- Ridge Regression
- Decision Tree regression.
- Random forest regression
- Gradient Boosting regression.

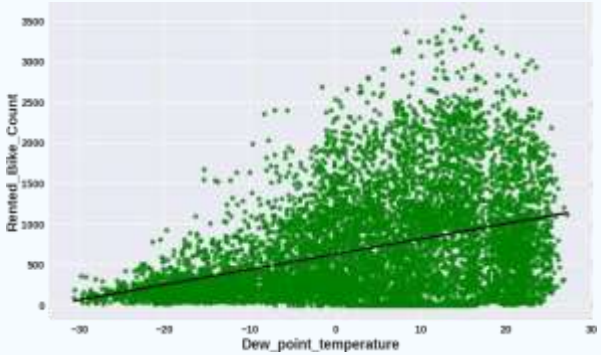
Assumptions of regression line:

1. It should be nearly linear how the dependent and independent variables relate to one another.
2. The average residual value should be zero or as near to zero as possible. To determine if our line truly is the line of "best match," this is done.
3. A regression model must have homoscedasticity or equal variance. According to this presumption, all values of the predictor variable have the same variance around the regression line (X).
4. Regression models shouldn't have many collinearities. When two or more independent variables have substantial correlations with one another, multi collinearity typically results.
5. By looking at the distribution of residuals, a scatter plot of the actual and predicted values, and reducing multi-colinearity among independent variables, we were able to verify our regression assumptions before and after using these models.

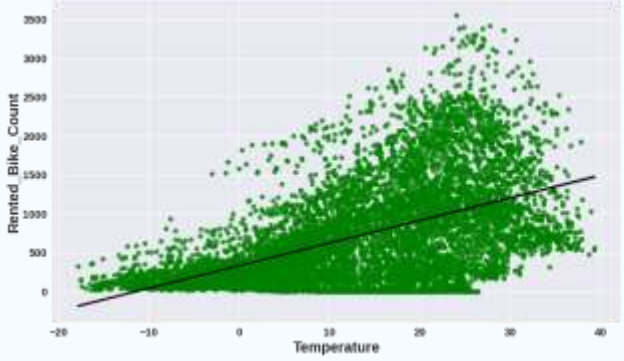
Model Selection and Evaluation

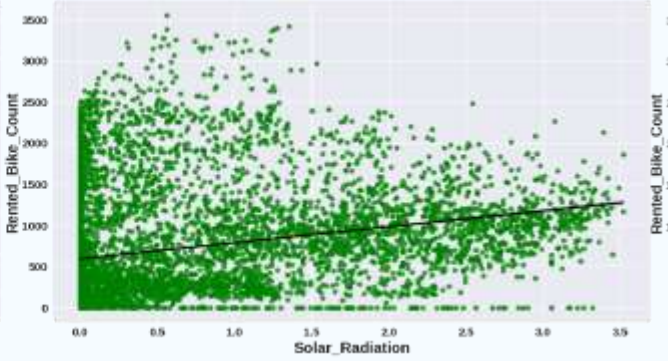


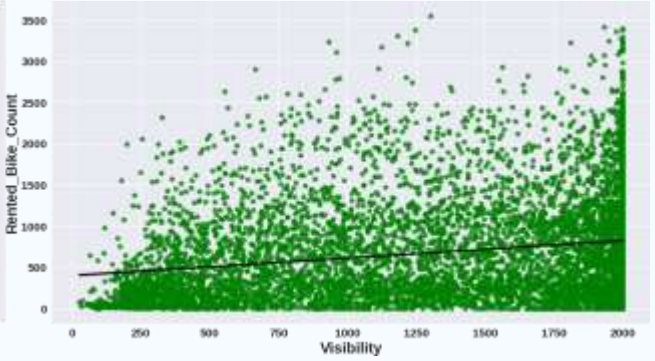




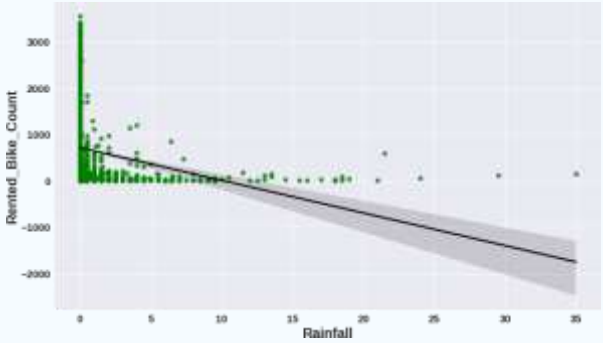
The columns "Temperature," "Wind speed," "Visibility," "Dew point temperature," and "Solar Radiation" are positively related to the target variable, which means the number of rented bikes rises with an increase in these features, as can be seen from the above regression plot of all numerical features.

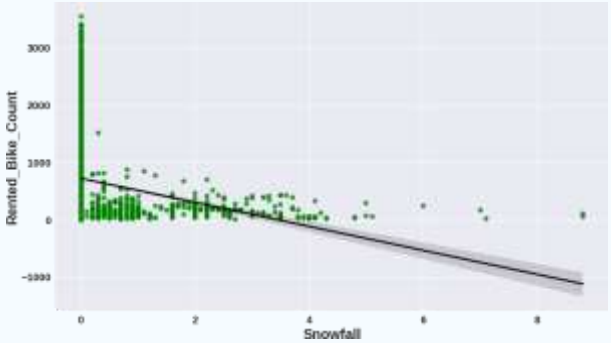


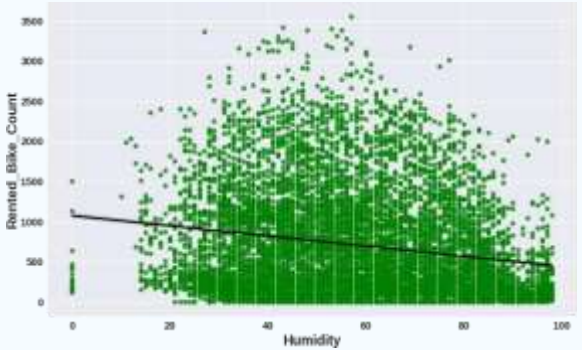




Model Selection and Evaluation







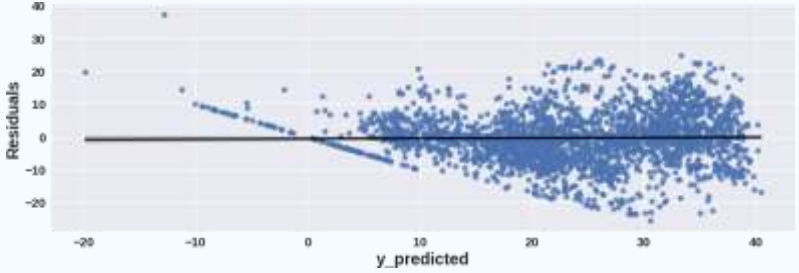
Rainfall, Snowfall, and Humidity are factors that are negatively correlated with the objective variable, meaning that when these features get worse, fewer people rent bikes.

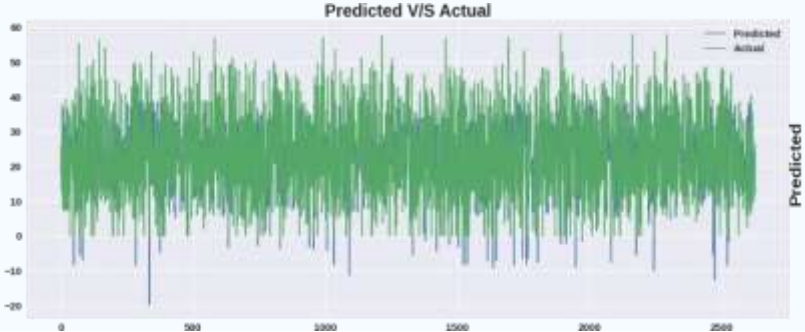
Model Selection and Evaluation

➢ **Linear Regression**

Scores on Trainset

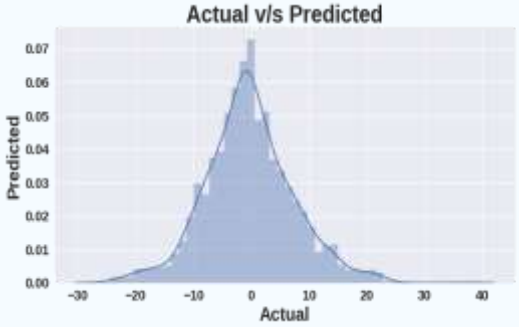
The Mean Absolute Error (MAE) is 4.702038874450511.
The Mean Squared Error(MSE) is 40.28177356393986.
The Root Mean Squared Error(RMSE) is 6.346792383869182.
The R2 Score is 0.7410406683214645.
Adjusted R2 : 0.7365274344231167

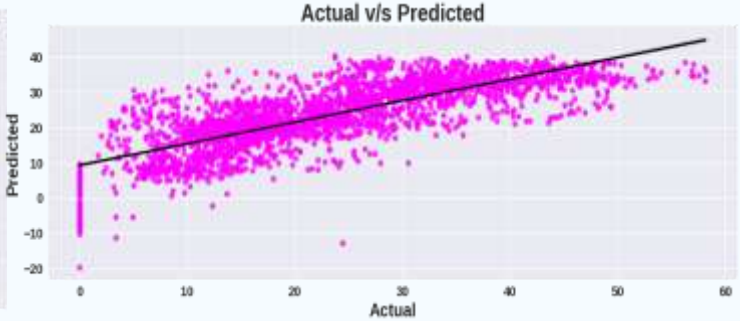




Scores on Testset

The Mean Absolute Error (MAE) is 4.615067638078279.
The Mean Squred Error(MSE) is 37.079711816860836.
The Root Mean Squared Error(RMSE) is 6.0893112760689805.
The R2 Score is 0.758592617752073.
Adjusted R2 : 0.7543852853736235





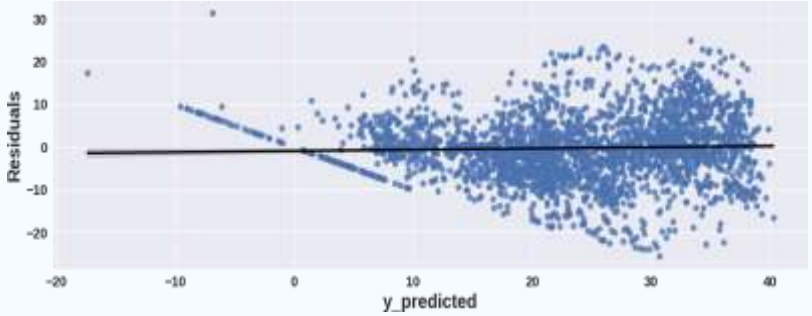
Mean of residuals should be zero or close to 0 as much as possible. It is done to check whether our line is actually the line of "best fit"

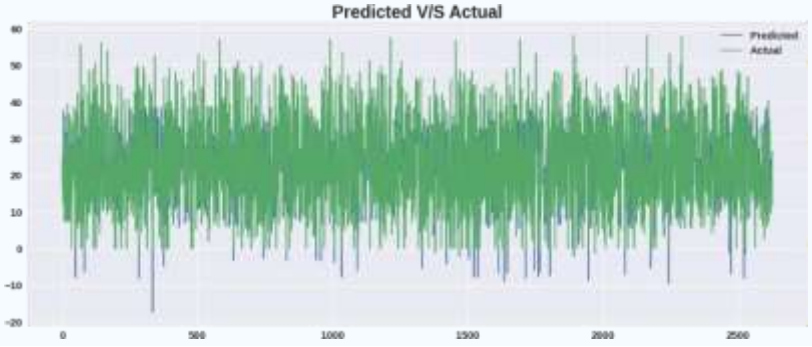
Model Selection and Evaluation

➢ **Lasso Regression (Hyper-parameter tuned- alpha=0.01)**

Scores on Train set

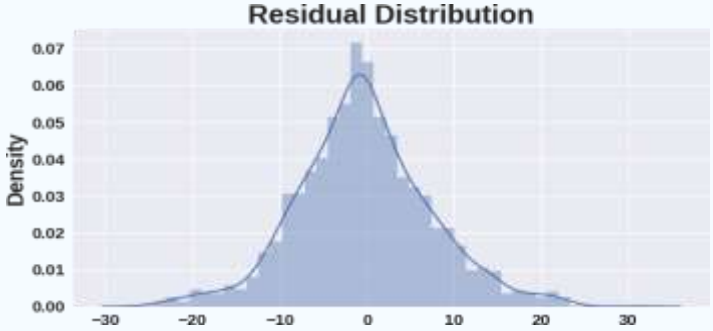
The Mean Absolute Error (MAE) is 4.706302197698486.
The Mean Squred Error(MSE) is 40.53831856326989.
The Root Mean Squared Error(RMSE) is 6.366970909566801.
The R2 Score is 0.7393914181595638.
Adjusted R2 : 0.7348494405519652

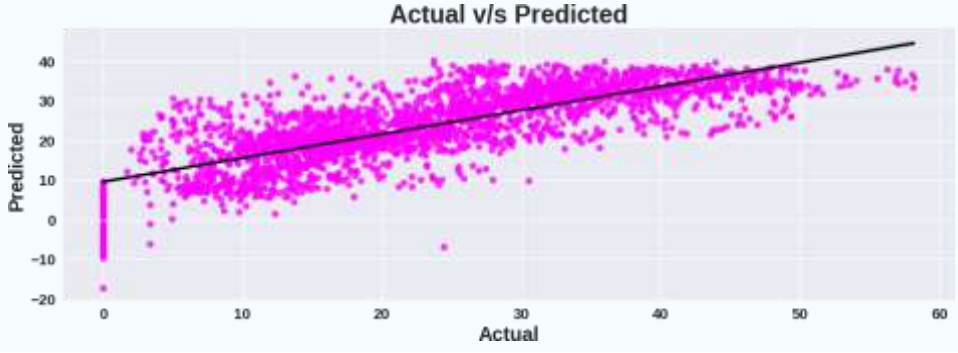




Scores on Testset

The Mean Absolute Error (MAE) is 4.617535387661918.
The Mean Squred Error(MSE) is 37.43193945811274.
The Root Mean Squared Error(RMSE) is 6.118164713221828.
The R2 Score is 0.756299440468227.
Adjusted R2 : 0.752052141793196





Model Selection and Evaluation

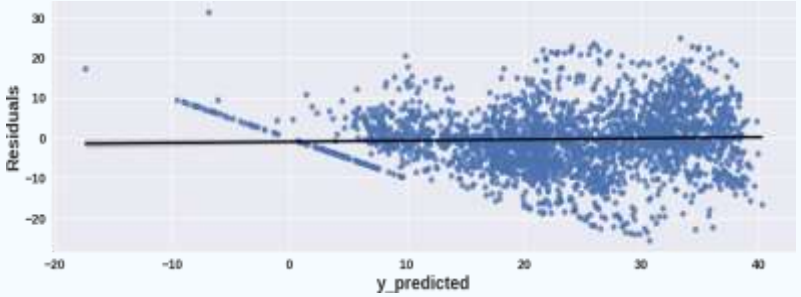
➢ **Ridge Regression(Hyper-parameter tuned- alpha=0.1)**

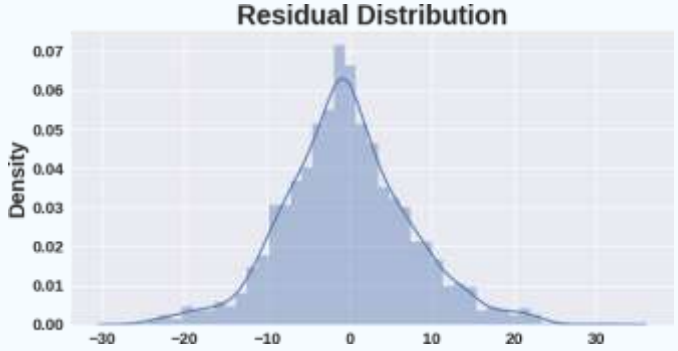
Scores on Train set

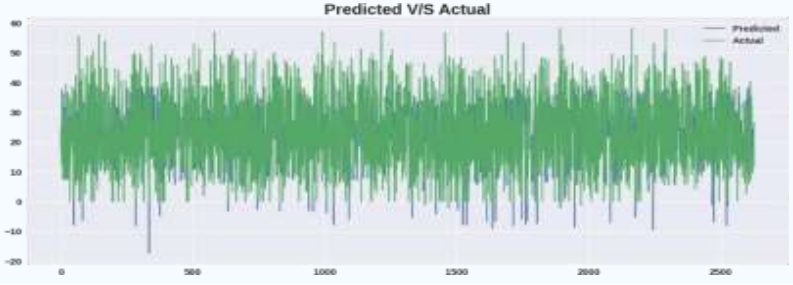
The Mean Absolute Error (MAE) is 4.706302197698486.
The Mean Squred Error(MSE) is 40.53831856326989.
The Root Mean Squared Error(RMSE) is 6.366970909566801.
The R2 Score is 0.7393914181595638.
Adjusted R2 : 0.7348494405519652

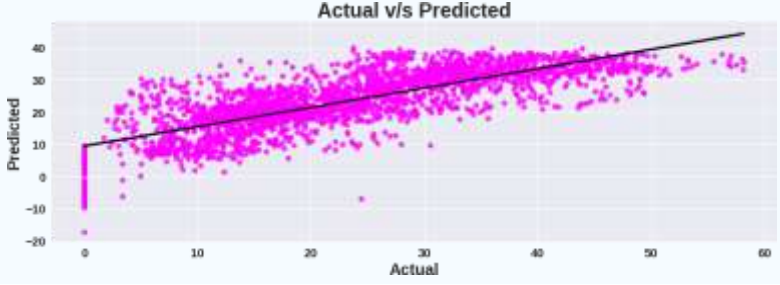
Scores on Testset

The Mean Absolute Error (MAE) is 4.617535387661918.
The Mean Squred Error(MSE) is 37.43193945811274.
The Root Mean Squared Error(RMSE) is 6.118164713221828.
The R2 Score is 0.756299440468227.
Adjusted R2 : 0.752052141793196









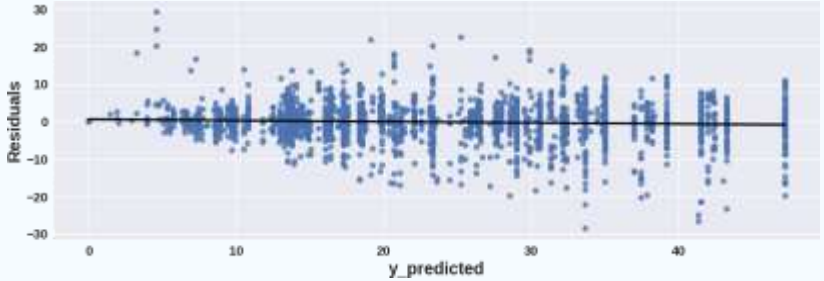
Model Selection and Evaluation

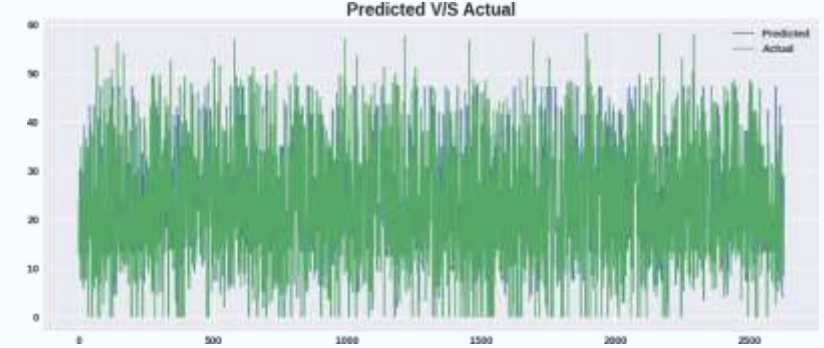
➢ **Decision Tree regression(Hyper-parameter tuned- max_depth=9,max_features='auto')**

The number of features to consider when looking for the best split

Scores on Trainset

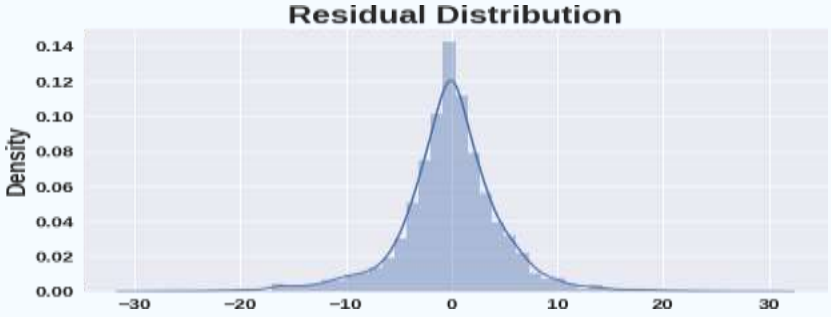
Model Score: 0.7034940289496471
MSE : 46.122247492641385
RMSE : 6.791336208187707
MAE : 4.854903777528903
R2 : 0.7034940289496471
Adjusted R2 : 0.698326419074641

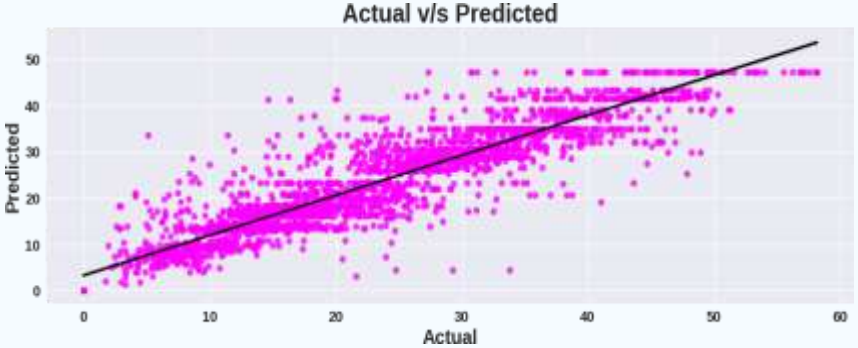




Scores on Testset

MSE : 46.835175022157074
RMSE : 6.843622945644878
MAE : 4.9537747782895964
R2 : 0.6950796960055869
Adjusted R2 : 0.6897654381900375



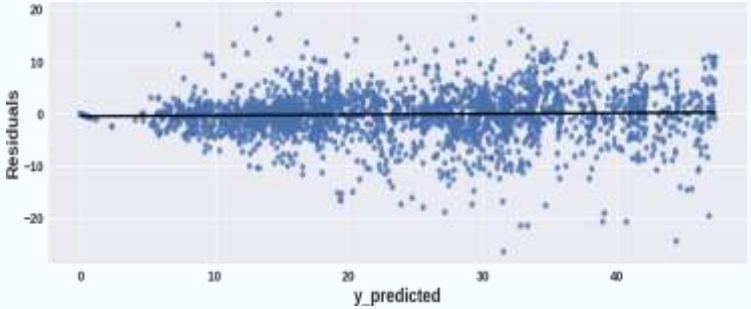


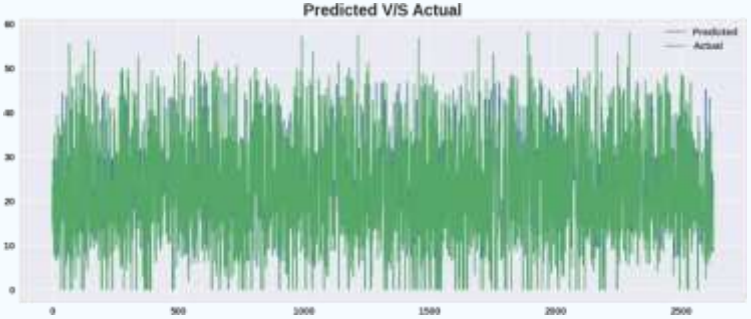
Model Selection and Evaluation

➢ Random Forest Regression

Scores on Trainset

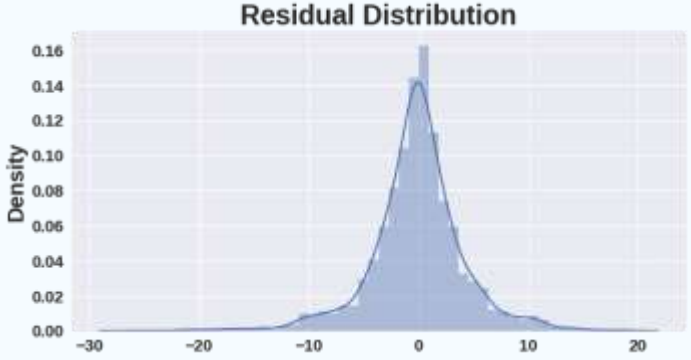
The Mean Absolute Error (MAE) is 0.8530259996132592.
The Mean Squred Error(MSE) is 1.9030432256298093.
The Root Mean Squared Error(RMSE) is 1.3795083274956368.
The R2 Score is 0.9877659110246918.
Adjusted R2 : 0.9875526910386775

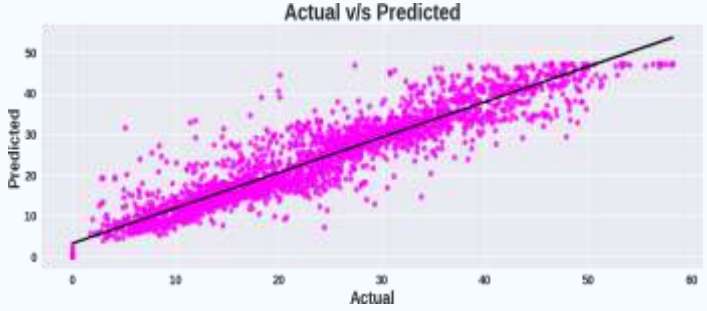




Scores on Testset

The Mean Absolute Error (MAE) is 2.2615080527193454.
The Mean Squred Error(MSE) is 12.842458270297694.
The Root Mean Squared Error(RMSE) is 3.583637575187772.
The R2 Score is 0.9163892036709126.
Adjusted R2 : 0.9149320054389959



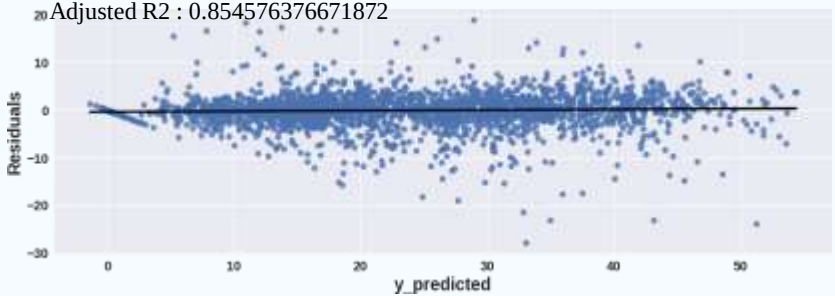


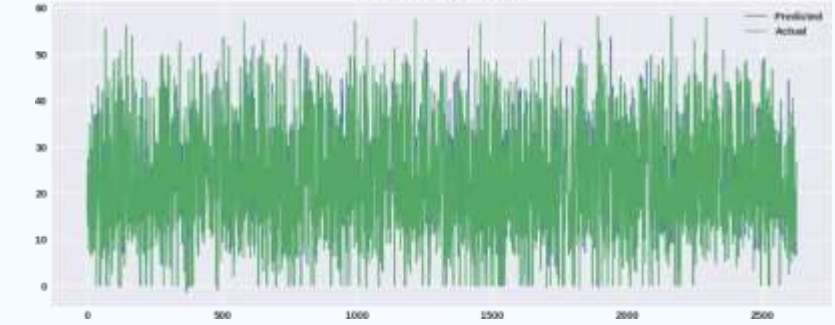
Model Selection and Evaluation

➢**Gradient boosting regression(Hyper-parameter tuned- 'learning_rate': 0.04, 'max_depth': 8, 'n_estimators': 150, 'subsample': 0.9)**

Scores on Train set

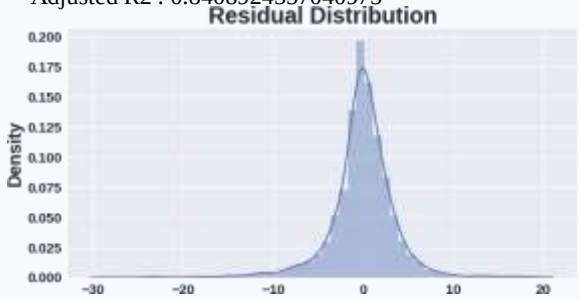
Model Score: 0.8570674551072606
MSE : 22.233515861225204
RMSE : 4.715242927063801
MAE : 3.4816393449441096
R2 : 0.8570674551072606
Adjusted R2 : 0.854576376671872

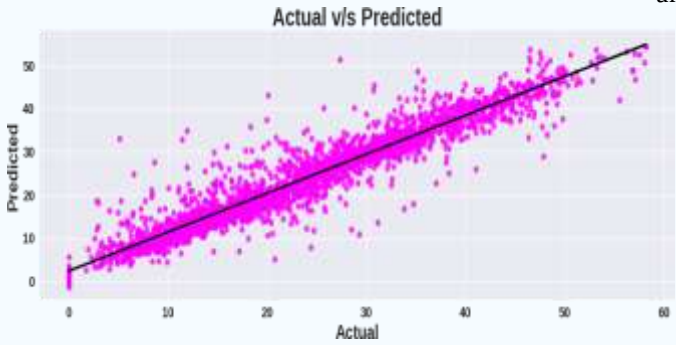




Scores on Testset

MSE : 24.01998552633507
RMSE : 4.901018825339795
MAE : 3.621289671891889
R2 : 0.8436179370490978
Adjusted R2 : 0.8408924557040975

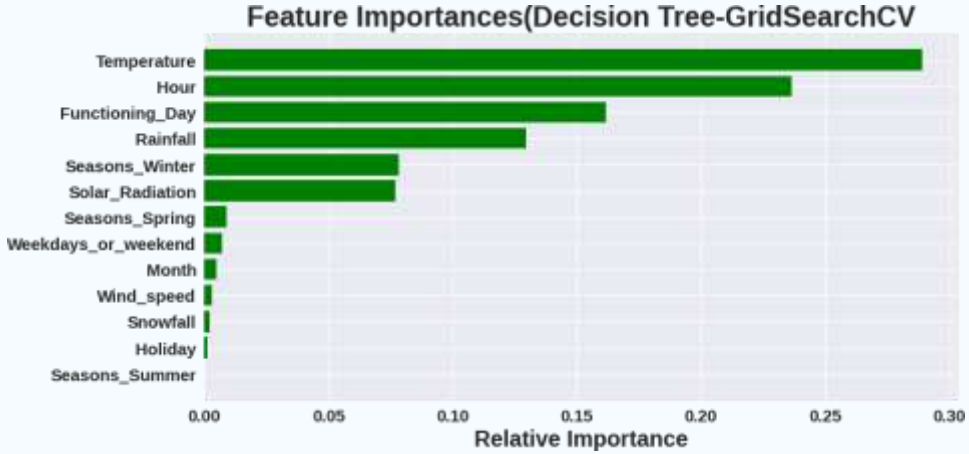


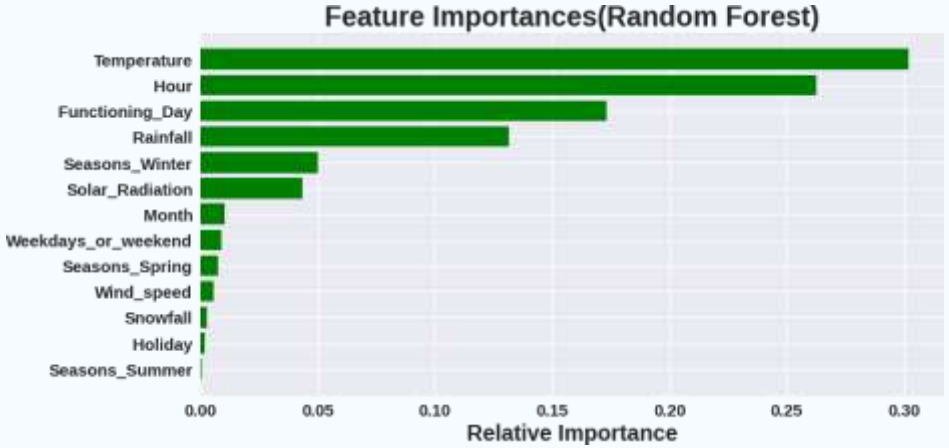


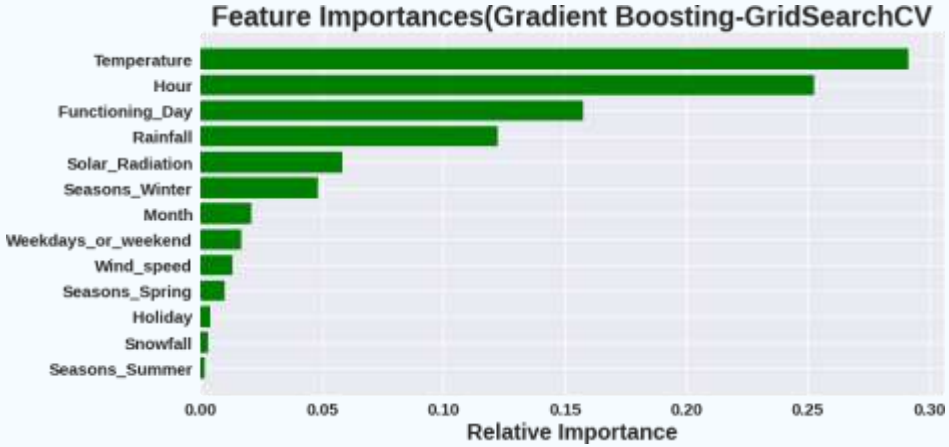
➢Learning rate shrinks the contribution of each tree by learning_rate. There is a trade-off between learning_rate and n_estimators.

➢Choosing subsample < 1.0 leads to a reduction of variance and an increase in bias.

Feature importance's







We can determine from all three models that the top three significant features are temperature, hour, and functional day.

Conclusion

Train data result

	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
0	Linear Regression	4.7020	40.2818	6.3468	0.7410	0.74
1	Lasso	4.7063	40.5383	6.3670	0.7394	0.73
2	Ridge GridSearchCV	4.7063	40.5383	6.3670	0.7394	0.73
3	Dicision tree regression	6.0820	72.1970	8.4970	0.5360	0.53
4	Gradient boosting regression	3.4820	22.2340	4.7150	0.8570	0.85
5	Random Forest	0.8607	1.9751	1.4054	0.9873	0.99

Test data result

	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
0	Linear Regression	4.6151	37.0797	6.0893	0.7586	0.75
1	Lasso	4.6175	37.4319	6.1182	0.7563	0.75
2	Ridge(GridsearchCv Tunned)	4.6175	37.4319	6.1182	0.7563	0.75
3	Dicision tree regression	6.3970	80.7760	8.9880	0.4740	0.46
4	Gradient boosting regression	3.6220	24.0230	4.9010	0.8440	0.84
5	Radom forest	2.2414	12.7087	3.5649	0.9173	0.92

As we have calculated MAE,MSE,RMSE and R2 score for each model. Based on r2 score will decide our model performance.
Our assumption: if the difference of R2 score between Train data and Test is more than 5 % we will consider it as over fitting.

Linear, Lasso and Ridge:
Linear, Lasso, Ridge and Elastic regression models have almost similar R2 scores(74%) on training and (75%) on test data.

Decision Tree Regression:
On Decision tree regressor model, without hyper-parameter tuning, we got r2 score as 70% on training data and on test data it was very less. Thus our model memorized the data. So it was a over fitted model.
After hyper-parameter tuning we got r2 score as 53% on training data and 47% on test data which is quite good for us.

Random Forest:
On Random Forest regressor model, without hyper-parameter tuning we got r2 score as 90% on training data and 91% on test data. Thus our model memorized the data. So it was a over fitted model, as per our assumption
After hyper-parameter tuning we got r2 score as 98% on training data and 91% on test data which is very good for us.

Gradient Boosting Regression(Gradient Boosting Machine):.
Our model performed well without hyper-parameter tuning. After hyper-parameter tuning we got r2 score as 85% on training data and 84% on test data, thus we improved the model performance by hyper-parameter tuning

Conclusion

Thus Gradient Boosting Regression(GridSearchCV) and Random forest(GridSearchCv) gives good r2 scores. We can deploy this models.

THANKYOU