Capstone Project

Seoul bike sharing demand prediction

By-Deepak kumar Dubey

Problem statement

- A **bicycle-sharing system**, **bike share program**, **public bicycle scheme** or **public bike share** (**PBS**) **scheme** is a shared transport service in which bicycles are made available for shared use to individuals on a short-term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" and return it at another dock belonging to the same system.

- It is important to make the rental bikes available and accessible to the public at the right time as it lessens the waiting time.

- We have around 8760 hours of observations for the whole year that covers all seasons and weather conditions to predict the demand of rental bikes at any particular hour using supervised machine learning algorithm.

- This is a regression problem that comes under supervised machine learning.

Steps used

This problem has been solved following these steps-

- Data understanding
- Data exploration
- Data preprocessing
- Feature engineering
- Model training
- Model validation
- Error analysis-Mean square error,mean absolute error,r2 score for test and train score,adjusted r2 score were calculated and compared for various models.

Data understanding

- Date: year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Wind Speed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - (Non-Functional Day), Fun (Functional Day)

Dependent and independent variables

1. Dependent variable -Rented bike counts per hour
2. Independent variable-

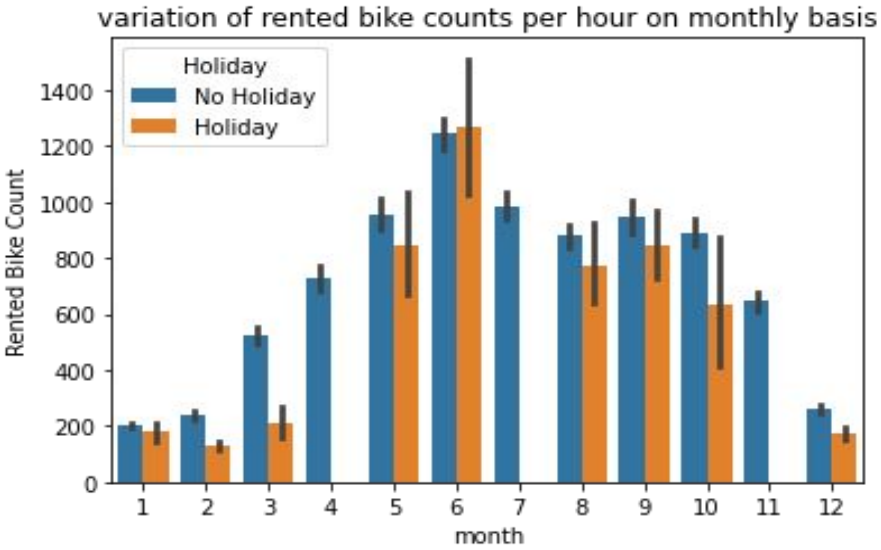
Numerical variables-

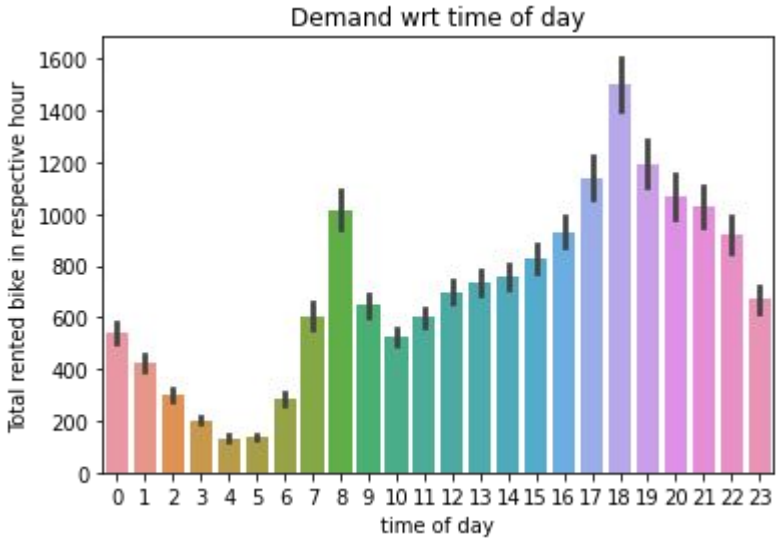
a. Temperature
b. Humidity
c. wind speed
d. Humidity
e. Rainfall
f. solar radiation
g. Wind speed
h. visibility

Categorical variables-

i. Hour
j. Holiday
k. Functioning day
l. Seasons

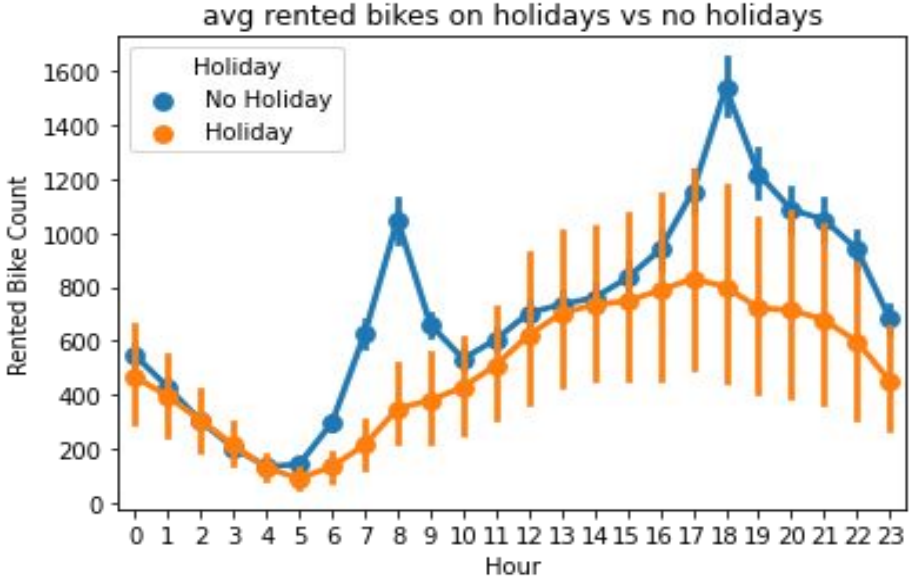
Exploratory data analysis

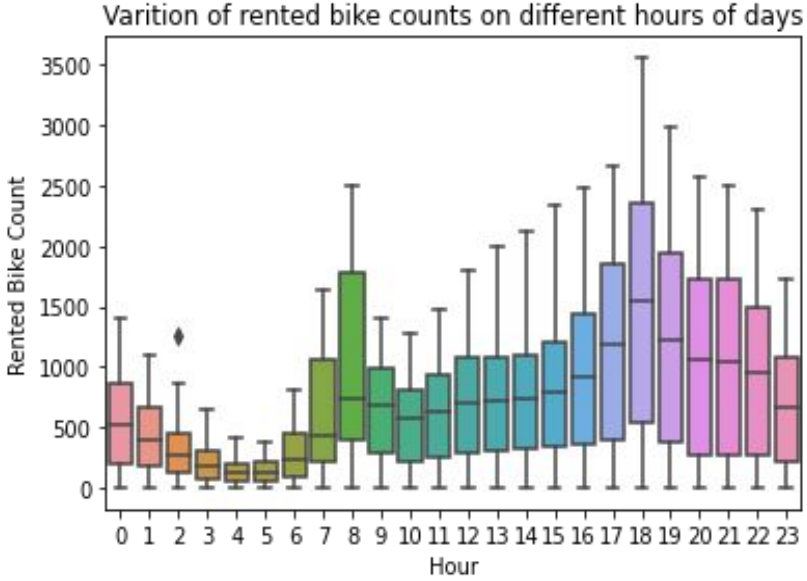




6th month of the year i.e june has highest rented bike counts while november, december ,jan ,feb have least count per hour.

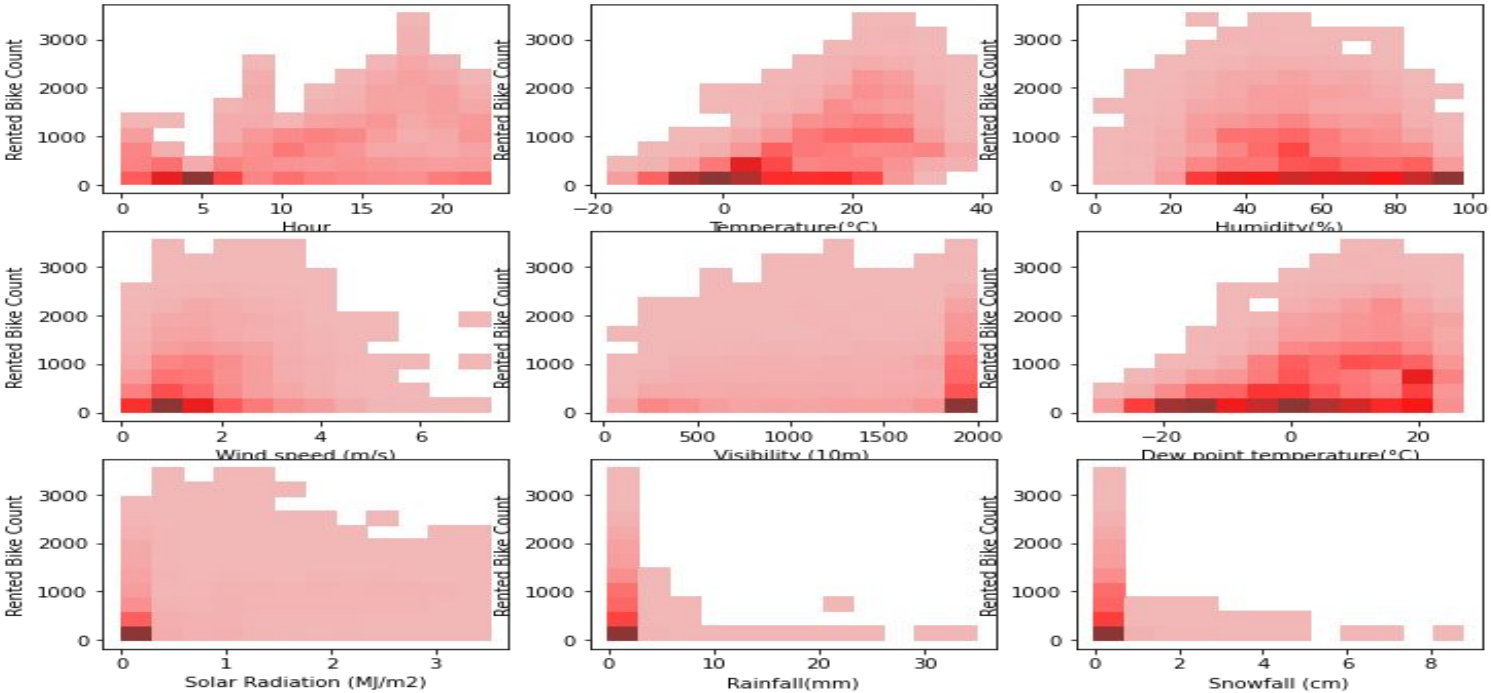
We can clearly conclude from the above chart that 6:00 pm in evening has highest count and rented bike count from 4 to 9 in evening remains higher as compared to other hours.





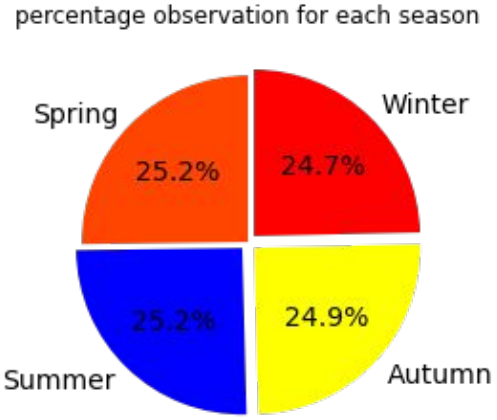
For holidays and no holidays, we can see avg rented bike counts distribution on hourly basis is different.It remains almost uniform from 1:00pm to 9:00 pm on holidays while varies significantly on no holidays.

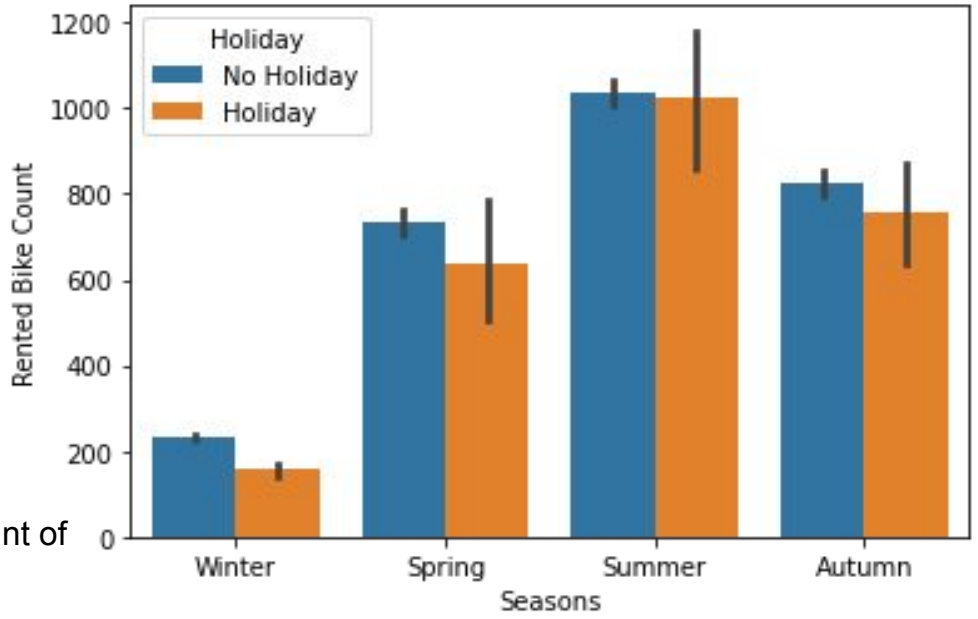
We can observe the variation for all hours of a day through a boxplot as well.



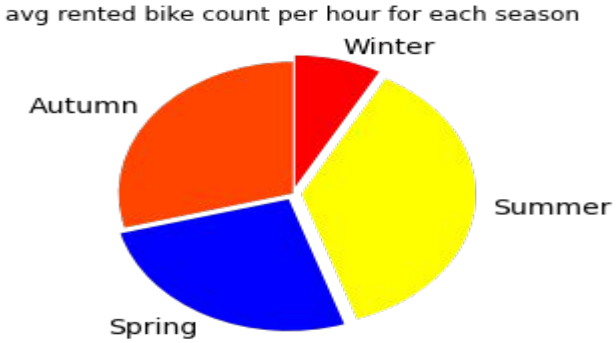
From above histograms we can note down following things-

1. When temperature is low rented bike counts per hour is low while when temperature value is between 20-30 degree count is high.
2. When wind speed is greater than 4 m/s count is low.
3. When there is rainfall or snowfall,number of bike count is quite low as compared to when there is no rainfall or snowfall.



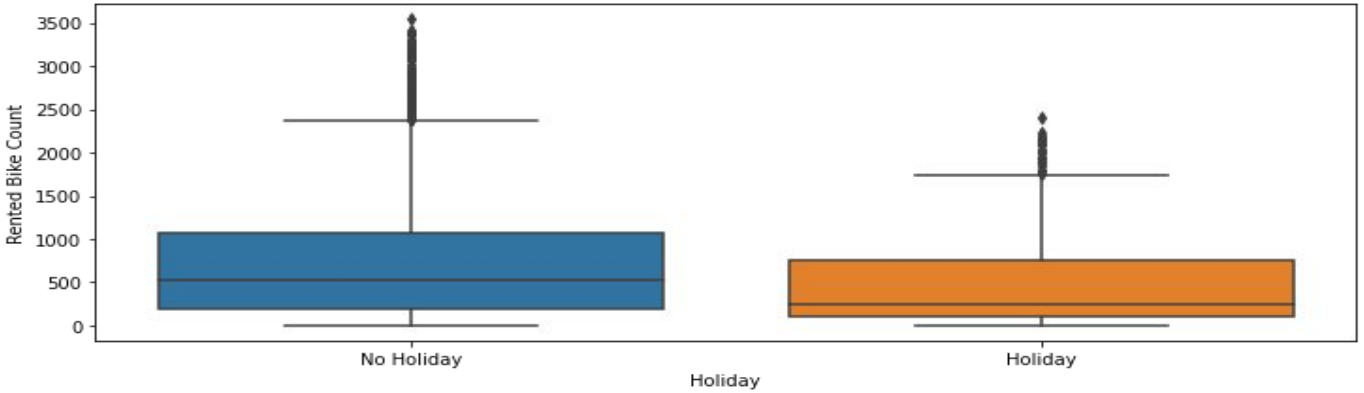


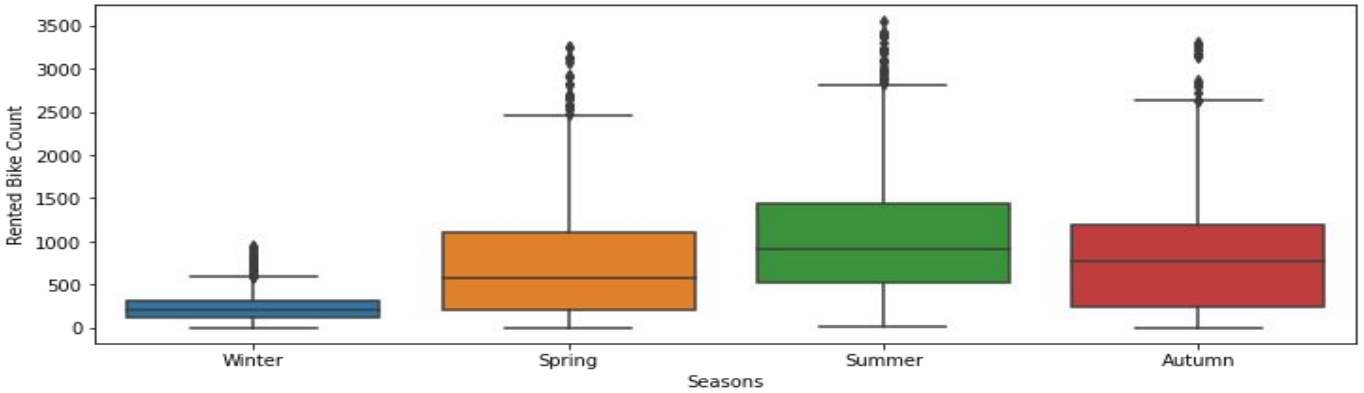
This barchart shows we have equal amount of data for all seasons in ourdata.



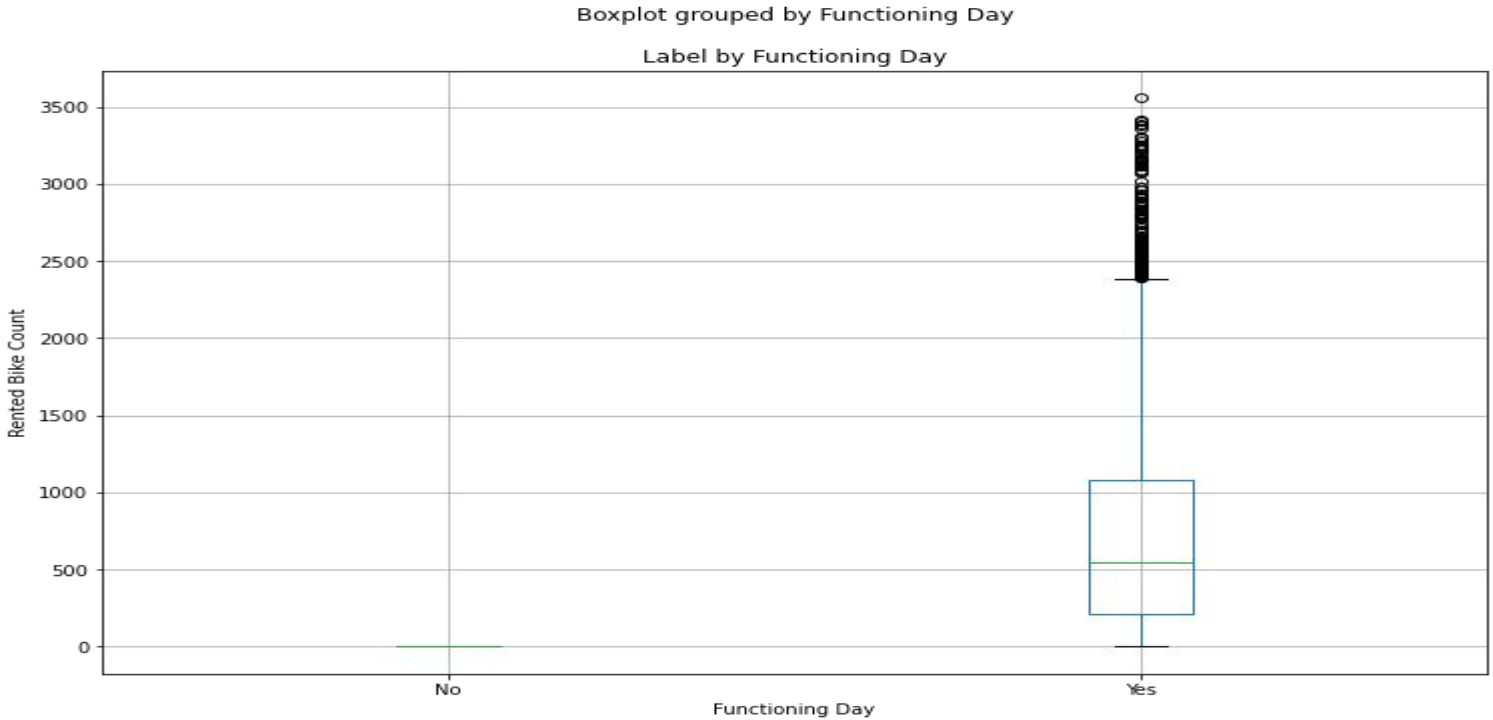
1)We can clearly see from the below pie chart that avg rented count per hour is higher when there is no holiday as compared to when there is holiday.

2)Avg rented bike counts for summer season is highest while winter is lowest

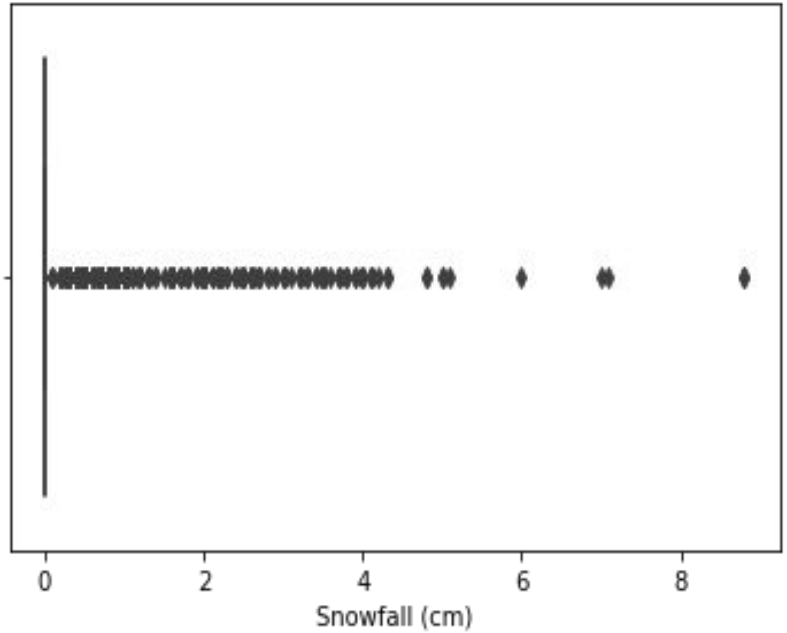


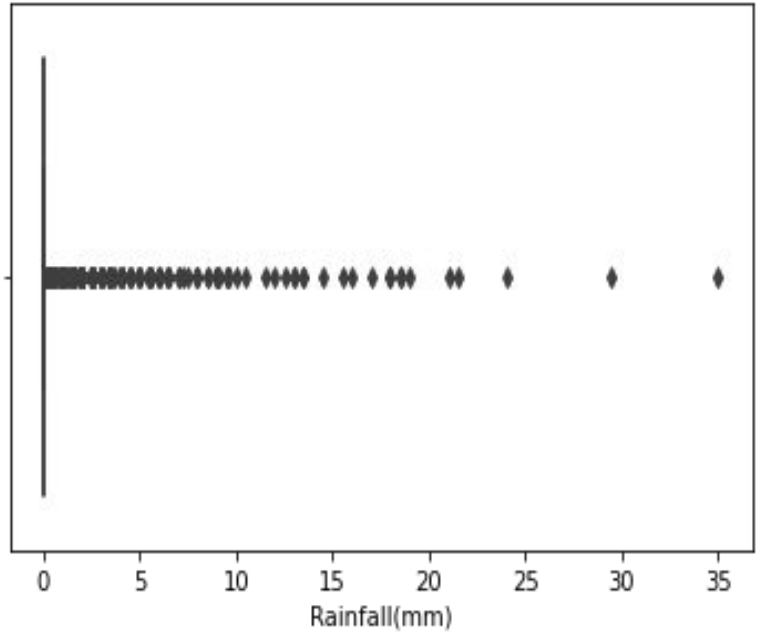


158 hours out of 8760 hours of observation have unusually high count of rented bikes.



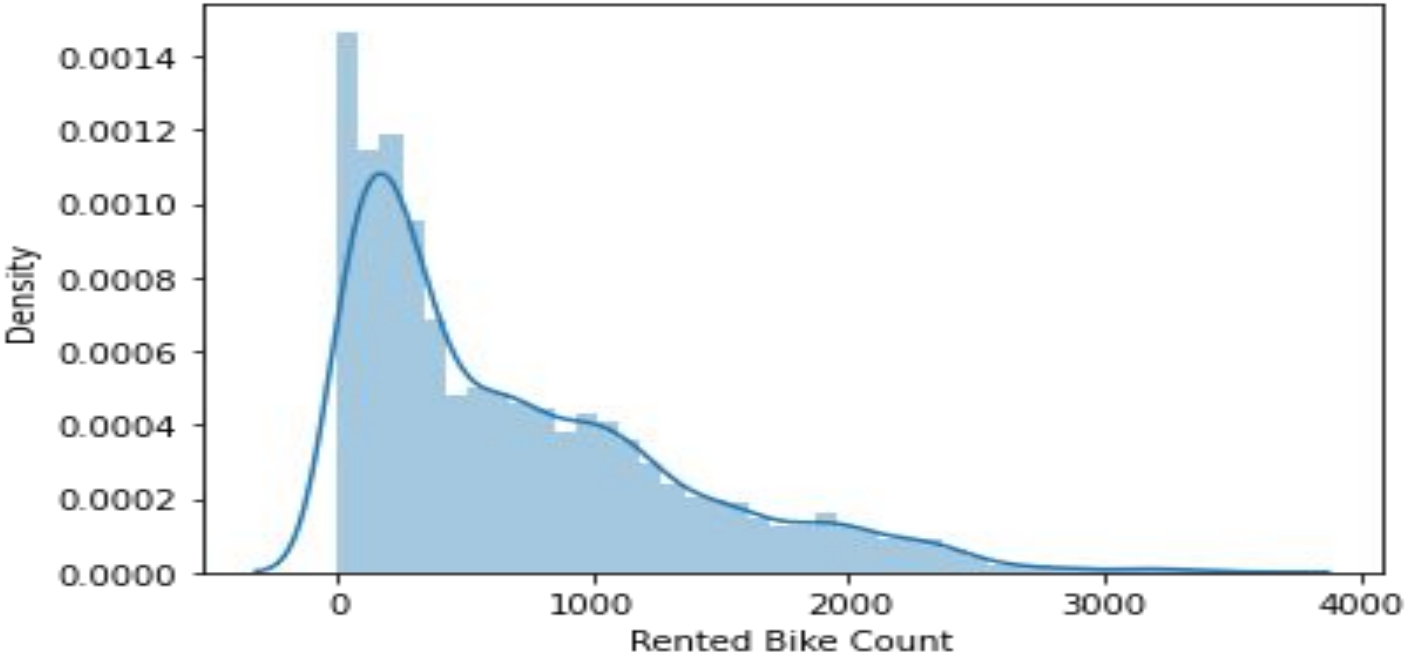
When there was not a functioning day,rented bike count in those hours was 0.



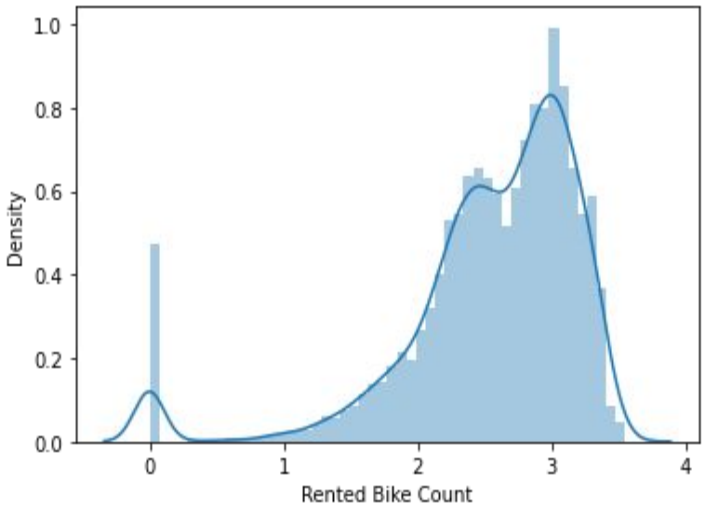


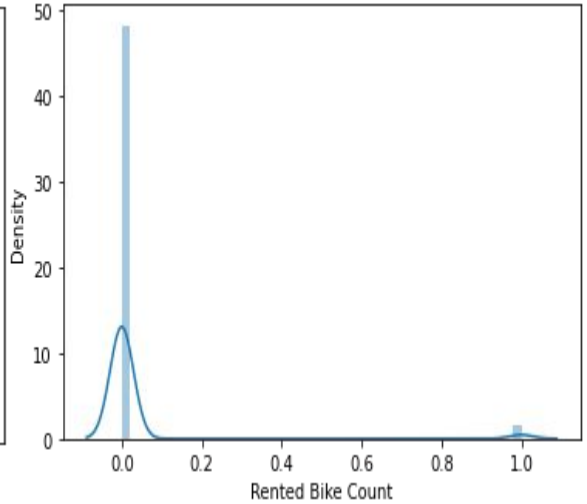
From above boxplots,we can conclude that there are very few hours out of 8760 hours when there is rain or snowfall but in those hours rented bike count falls quite low.

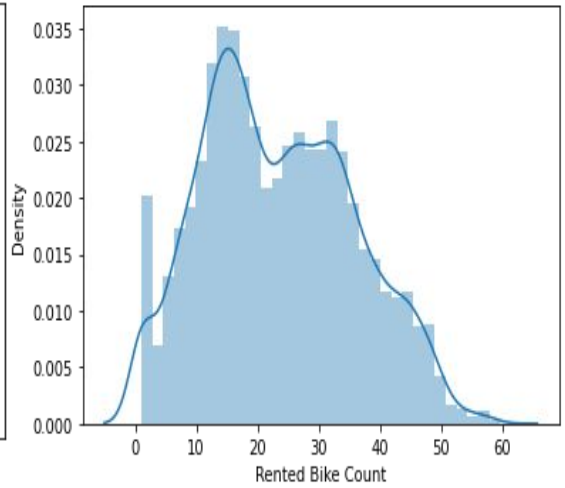
Dependent variable distribution



This is a right skewed distribution .Transformation has to be used to get a normal distribution.







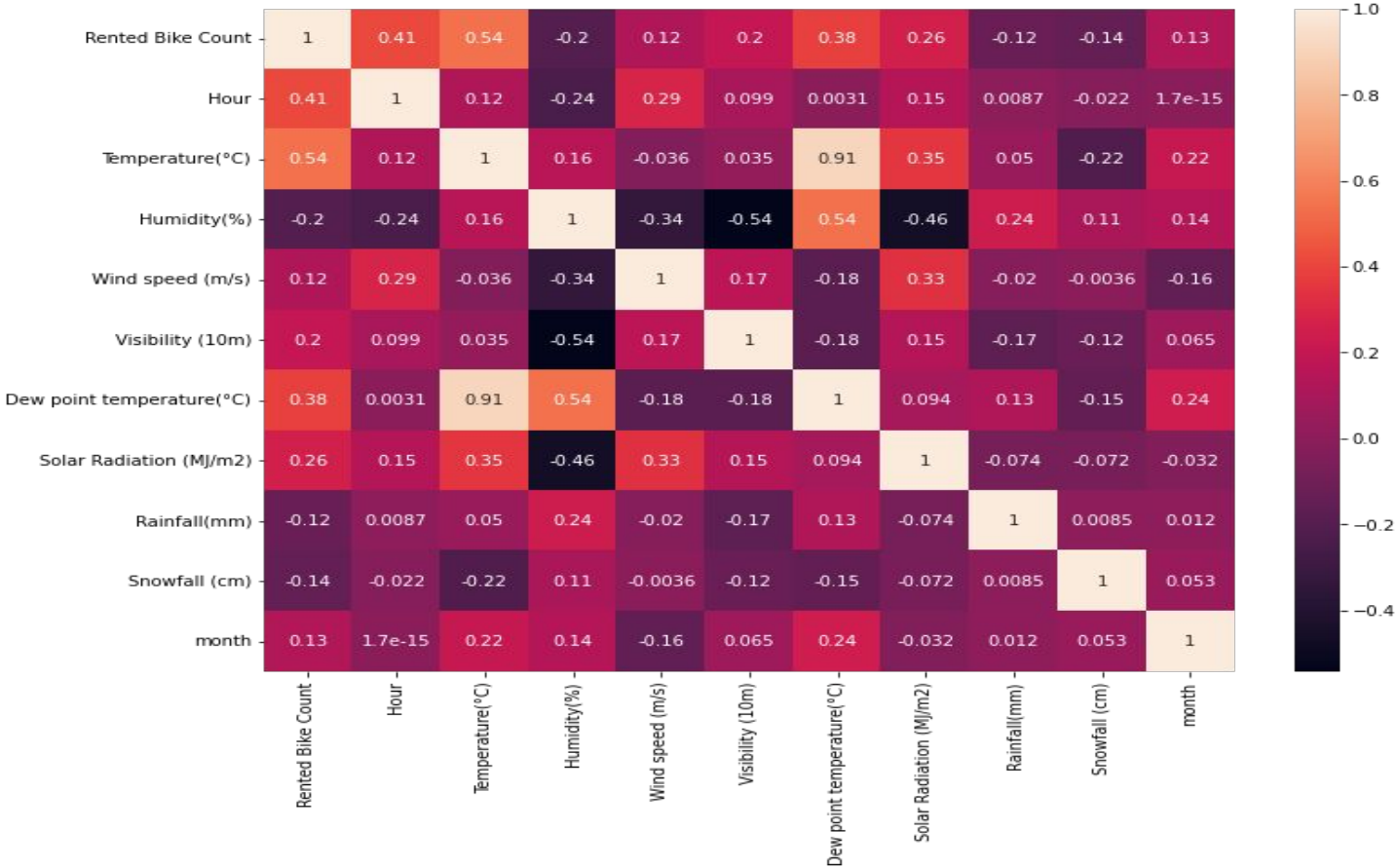
Log transformation

inverse transformation

sqrt transformation

We can see that square root transformation gives a distribution that looks somewhat like normal dist.

MULTICOLLINEARITY



We can see correlation coefficient between dew point temperature and temperature is 0.91 which represents almost a perfect correlation.So we will drop dew point temperature.

Feature engineering

- One hot encoding was used for categorical features.
- Z score pruning method was used to detect outliers.
- Square root transformation is used to convert dependent variable distribution to normal distribution.

Models used

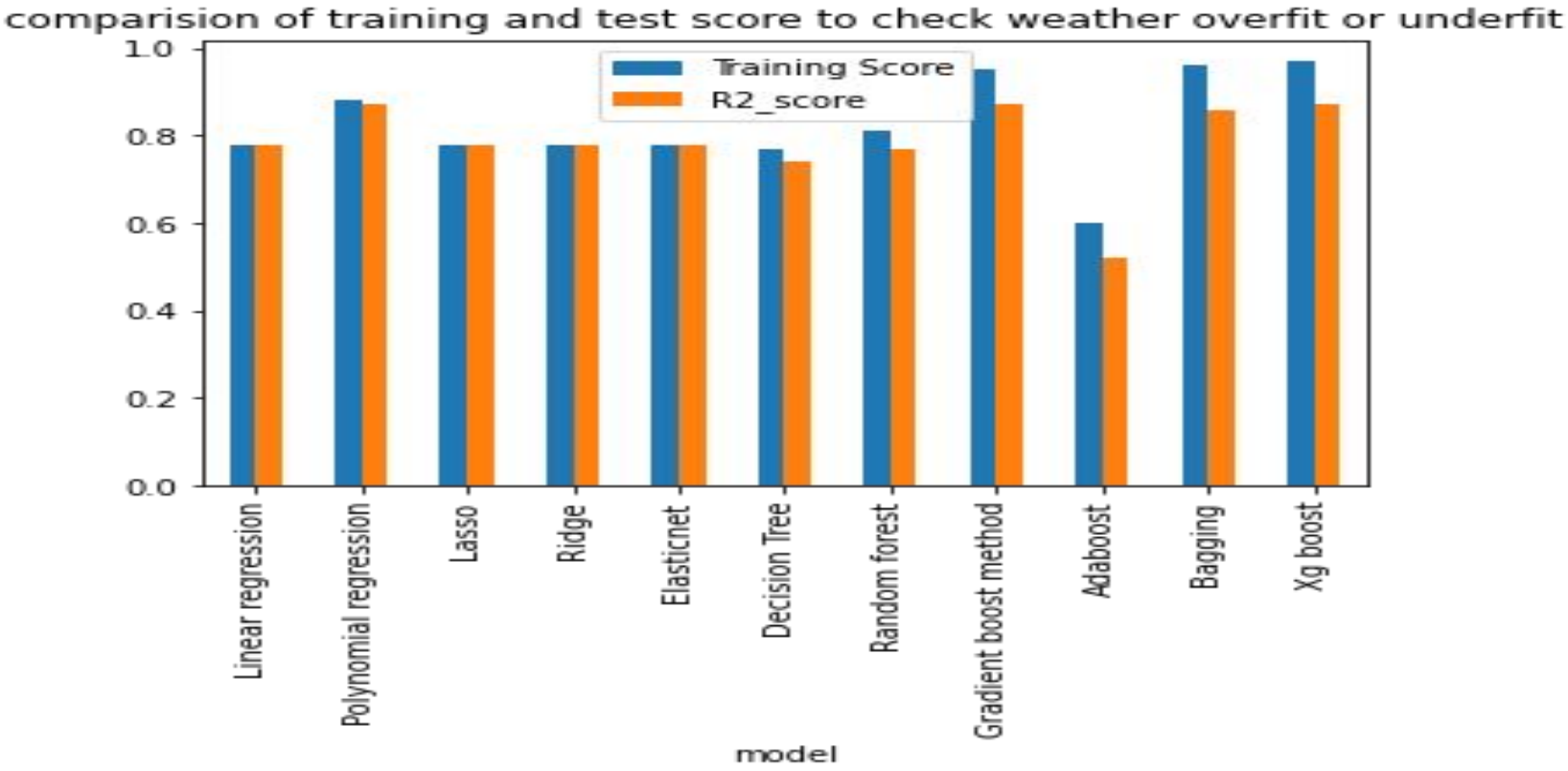
- Linear Regression
- Polynomial Regression
- Lasso Regression
- Ridge Regression
- Elastic net regression
- Decision tree regressor
- Random forest regressor
- Gradient boosting method
- XGboost regression model
- adaboost regression model
- Bagging regression model

Error Analysis

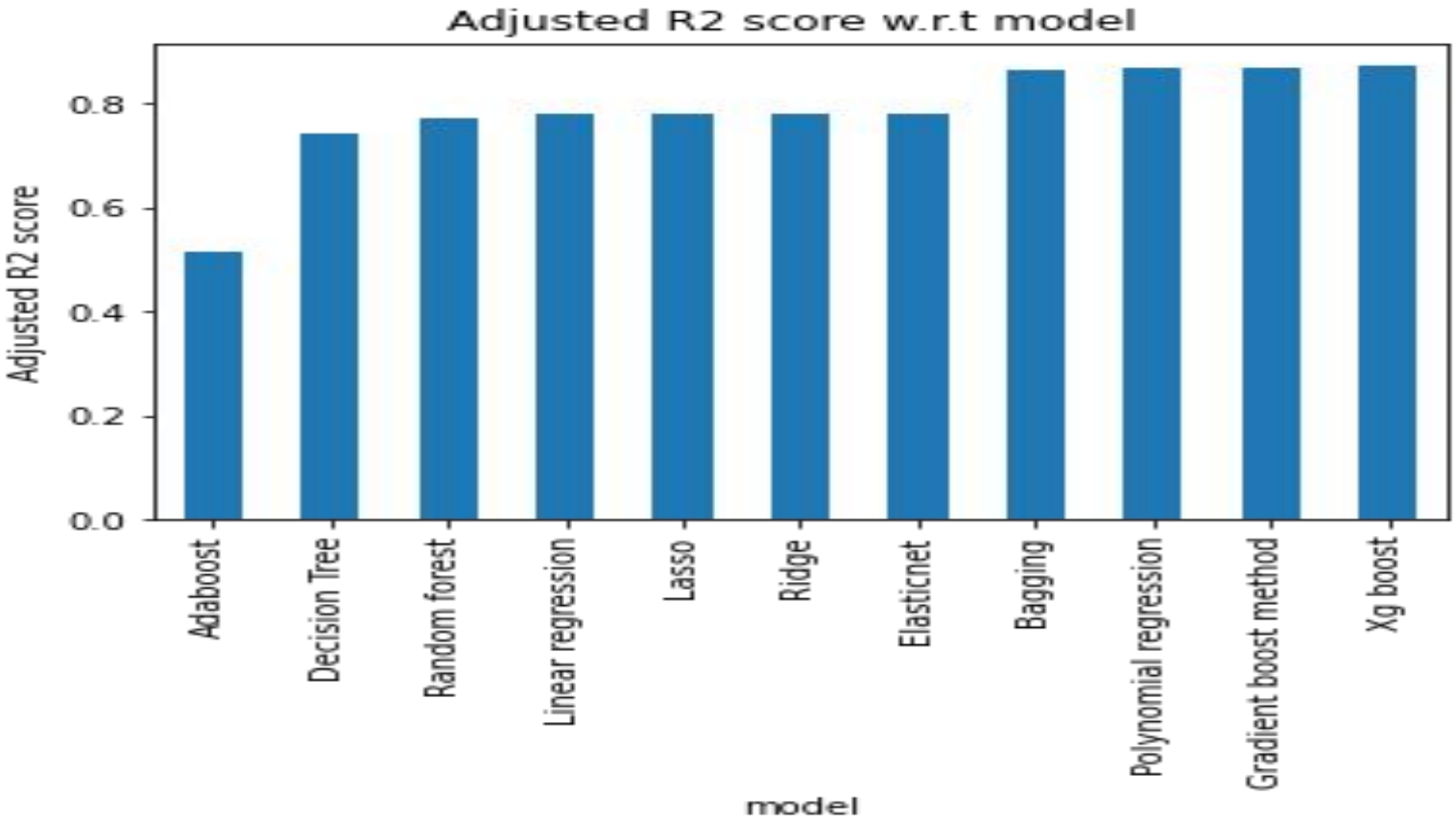
model	MSE	RMSE	MAE	Training Score	R2_score	Adjusted_R2
Linear regression	89591.33	299.41	202.52	0.78	0.78	0.78
Polynomial regression	53336.51	230.95	144.84	0.88	0.87	0.869
Lasso regression	89588.76	299.31	202.51	0.78	0.78	0.78
Ridge regression	89608.41	299.34	202.56	0.78	0.78	0.78
Elastic net regression	89614.45	299.35	202.56	0.78	0.78	0.78
Decision Tree regression	106589.86	326.48	210.79	0.77	0.74	0.74
Random forest regression	93720.39	306.14	198.11	0.81	0.77	0.77
Gradient boost method	52276.59	228.64	137.67	0.95	0.87	0.87
Adaboost	197572.58	444.49	311.62	0.6	0.52	0.516
Bagging	55946.63	236.53	140.28	0.96	0.86	0.862
Xgboost	52096.23	228.24	138.47	0.97	0.87	0.872

Simple models as well as models with cross validation and hyperparameter tunings have been used. Best results have been put in this table.

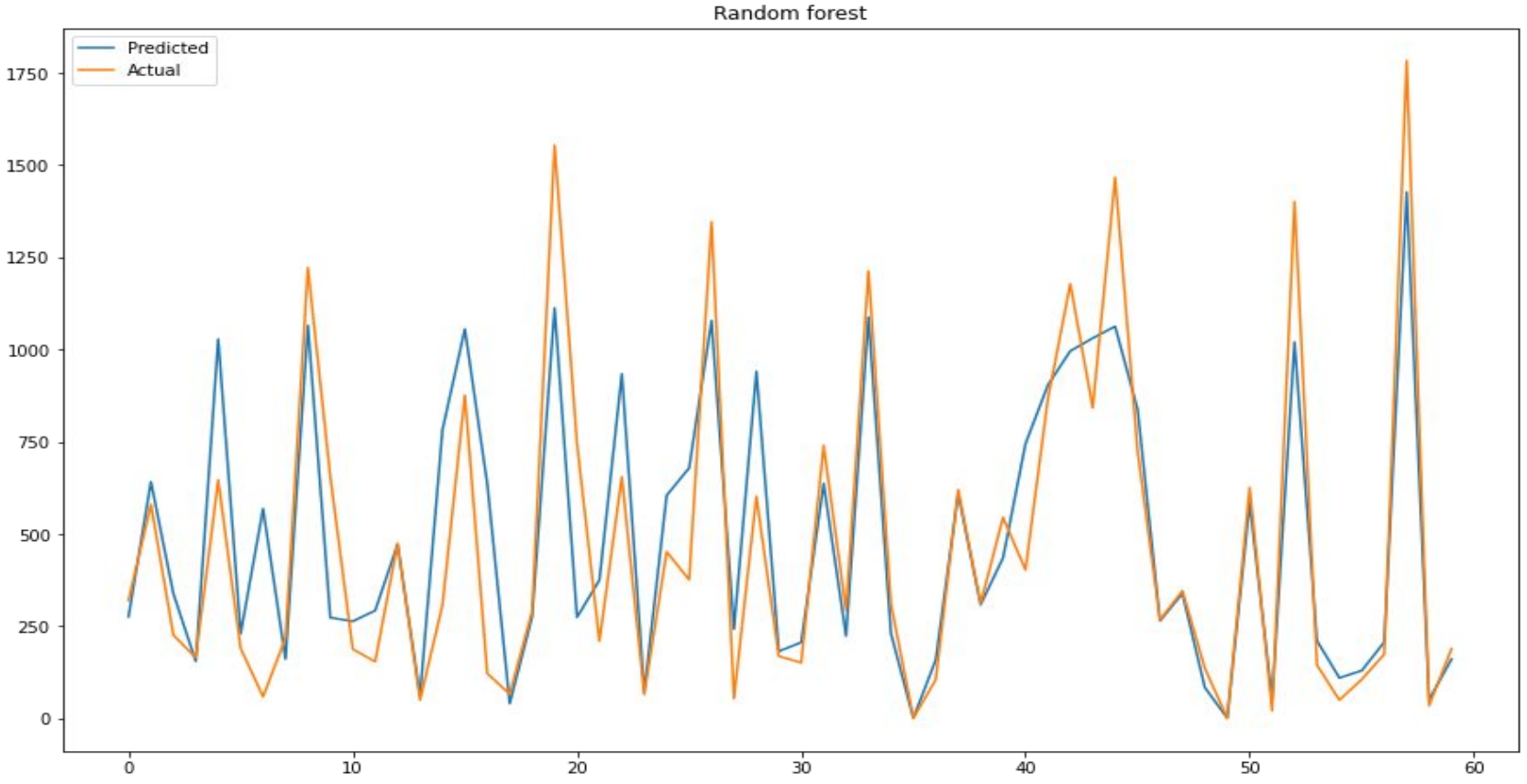
Train score vs test score



Comparison of Adjusted R2 value variation for different models



Predicted value vs actual value of dependent variable for random forest model



Conclusions from model fitting and validation

- Linear models such as linear regression ,lasso,ridge are performing good on both train and test data.R2 score is around 0.8 for both train and test set.
- Polynomial regression is giving better results than linear models and giving a good r2 score on both test and train set.r2 score is around 0.85 for both train and test sets.
- Ensemble models are giving quite good results on train data i.e R2 score more than 0.9 but are a little bit overfit i.e r2 score of 0.86-0.88 on test data.

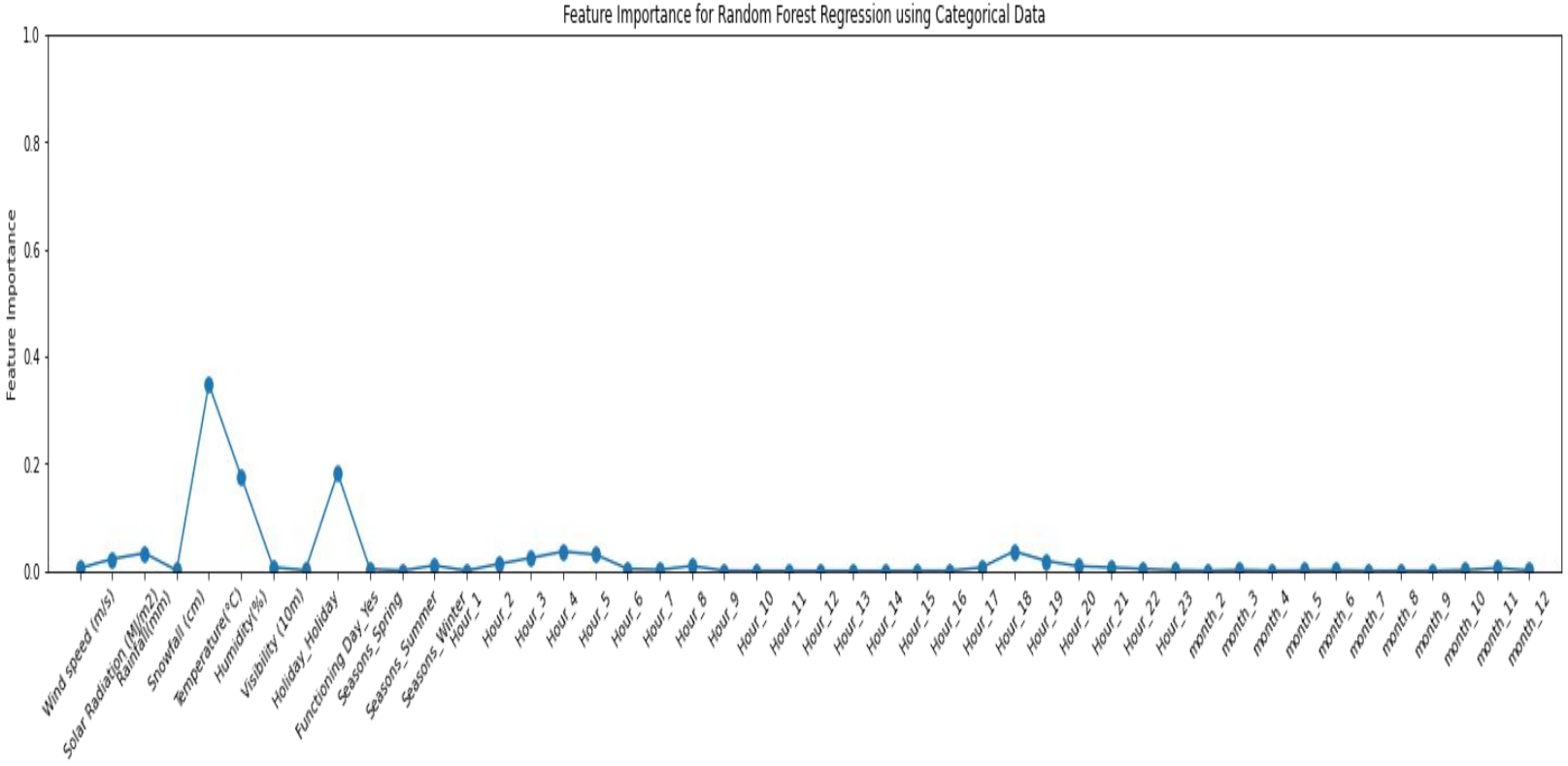
Important features based on feature importance plot

- Temperature,humidity,hour_18,season_winter,Functioning day_yes

Feature importance table for Random forest with hyperparameter tuning

index	Feature	Feature Importance
0	Temperature(°C)	0.35
1	Functioning Day_Yes	0.18
2	Humidity(%)	0.18
3	Hour_4	0.04
4	Hour_18	0.04
5	Rainfall(mm)	0.03
6	Hour_5	0.03
7	Hour_3	0.02
8	Hour_19	0.02
9	Solar Radiation (MJ/m2)	0.02

Feature importance plot



Conclusion

- We observed that bike rental count is high during no holidays than holidays.
- The rental bike counts was at its peak at 8 AM in the morning and 6pm in the evening, an increasing trend can be observed from 5am to 8am, the graph touches the peak at 8am in the morning then dips a bit. Later we can see a gradual increase in the demand until 6pm, the demand is highest at 6 pm, and reduces until midnight.
- People prefered to rent bikes when temperature was between 20 degrees to 35 degrees celsius temperature and even when it is little windy.
- Highest bike rental count was found in Autumn and summer seasons and the lowest is in winter season.
- Bike rentals were highest during the clear days and lowest on snowy and rainy days.
- When there was not a functioning day no bikes were booked.