Capstone Project
Bike Sharing Demand Prediction

Team Members

Gaurav Bisht

Deepak Karki

BIKE SHARING SYSTEM

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place an there need.

Bike rental service 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.

Predicting bike-sharing demand can help bike-sharing companies to allocate bikes better and ensure sufficient circulation of bikes for customers.

COMPLICATION STATEMENT

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

STEPS INVOLVED

1. Analyzing the dataset : Exploring the features & target variable, checking for duplicates & null values, plotting the distribution of target variables etc.

2. Exploratory data analysis : Working on numerical features & catagorical features seperately, VIF analysis, outlier detection etc.

3. Initialization of data : Train – test split, Tranformation of data, Scaling etc.

4. Creating Models : Creating different models & evaluating them using different metrics.

OBJECTIVES

1. PRIMARY OBJECTIVE:

> Prediction of bike count required at each hour for the regular supply of rental bikes in bike sharing system.

> To build a superior statistical model to predict the no. of bicycles that can be rented with availability of data.

2. SECONDARY OBJECTIVE:

> To learn how real time data is represented in datasets.

> To understand how to pre process such data.

> To study the comparision of results achieved by various machine learning algorithms such as Linear regression, Random forest, XG Boost & Gradient boosting machine.

DATA SUMMARY

The following dataset consists of 8760 rows & 14 columns.

DATA

Features / Independent variable

Numeric:
1. Hour
2. Temperature
3. Humidity
4. Wind Speed
5. Visibility
6. Dew point temperature
7. Solar radiation
8. Rainfall
9. Snow fall

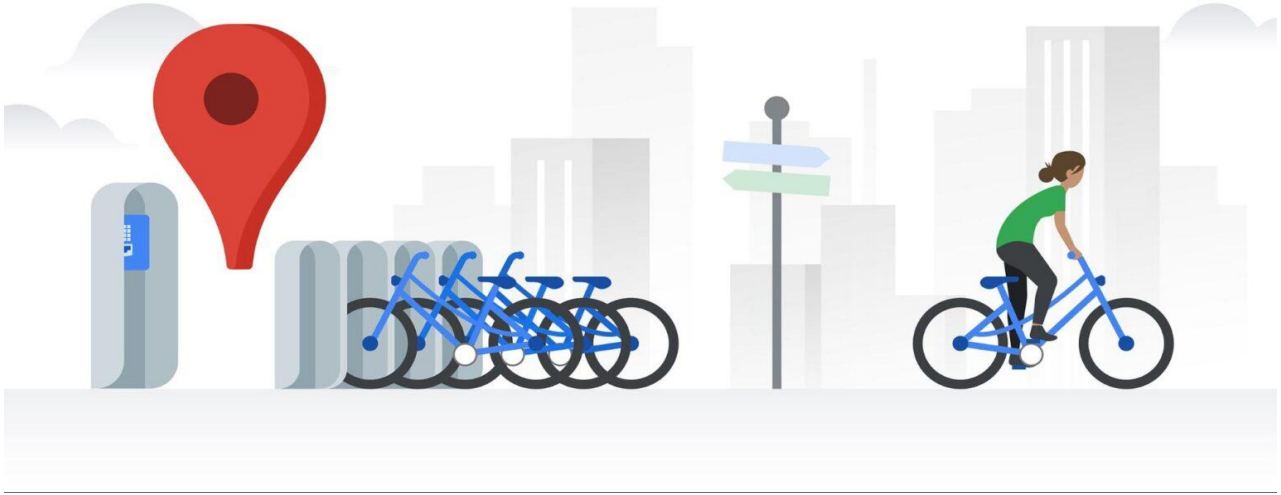
Categorical:
1. Season
2. Functioning Day
3. Holiday

Target variable

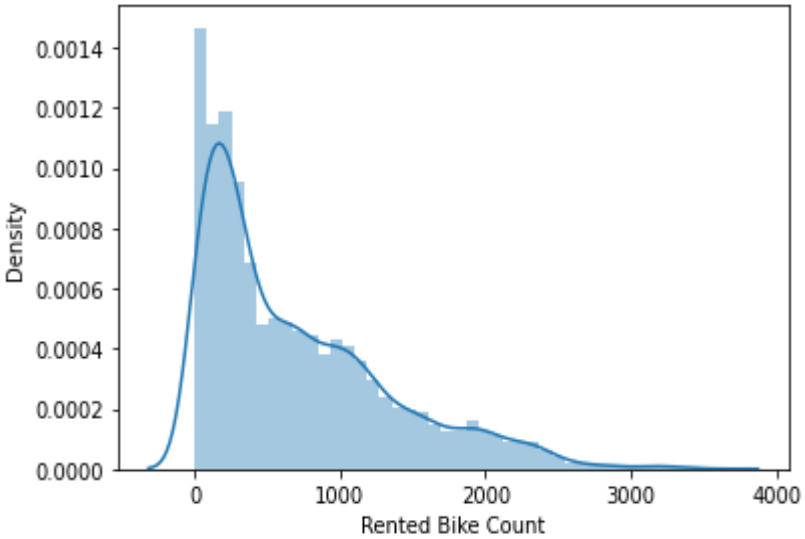
Rented Bike Count

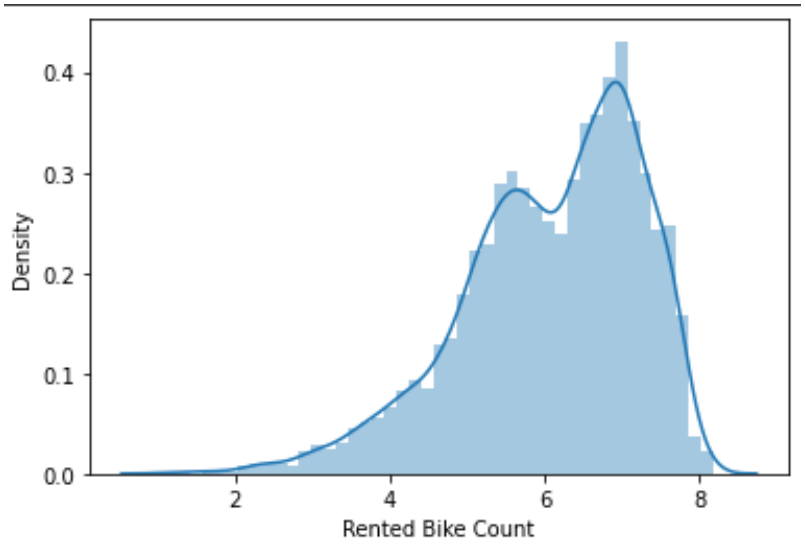
DEPENDENT VARIABLE

In the given dataset the dependent variable or basicially the target variable is " Rented Bike Count", The goal of this project is to combine the historical bike usage patterns with the weather data in order to forecast bike rental demand & to reduce the waiting time.



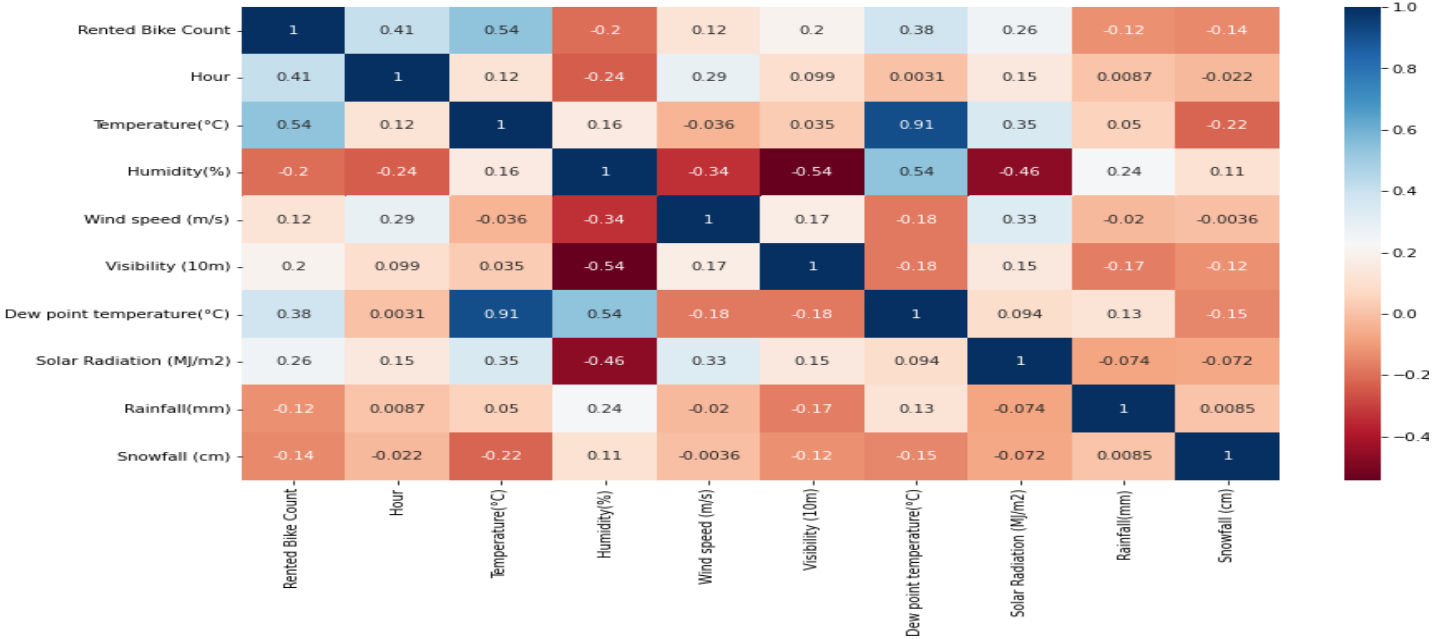
Plotting distribution of dependent variable





In the above observation we can clearly see that a distortion or asymmetry is deviating from the symmetrical bell curve plot by the given data, hence the distribution of the dependent variable is skewed. So we will apply numpy.log1p() for transforming the data.

Checking collinearity between variables via heatmap plot



Insights from the above Heatmap:

> The heatmap clearly shows which all variable are **multicollinear in nature**, and which variable have **high collinearity** with the **target variable**.

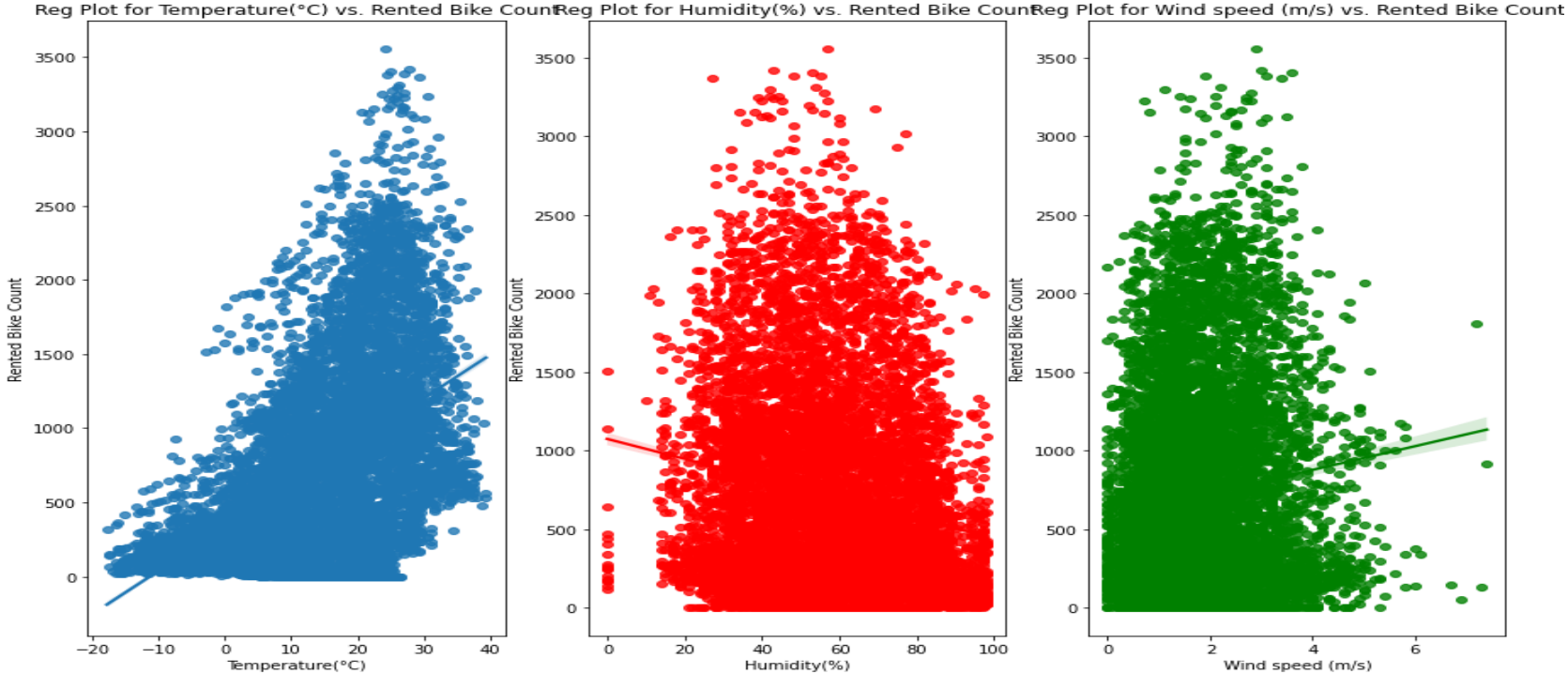
> We can observe that the numeric variables Temperature and Dew Point Temperature exhibit a high correlation, Also humidity & visibiity(10m) have a good negative correlation.

> We will refer this map back-and-forth while building the linear model so as to validate different correlated values along with **VIF & p-value**, for identifying the correct variable to select/eliminate from the model.

Correlation Analysis

Regression Plots vs. Temperature, Humidity and Windspeed

Using seaborn to get regression plots with respect to Temperature, Humidity and Windspeed.



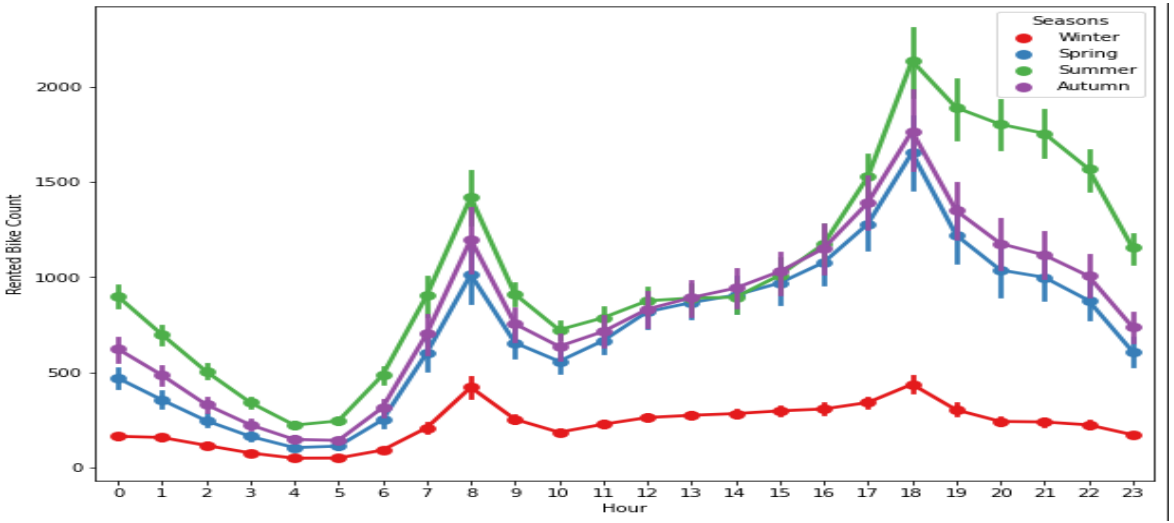
VIF Analysis

	variables	VIF
0	Hour	3.921832
1	Temperature(°C)	3.228318
2	Humidity(%)	4.868221
3	Wind speed (m/s)	4.608625
4	Visibility (10m)	4.710170
5	Solar Radiation (MJ/m2)	2.246791
6	Rainfall(mm)	1.079158
7	Snowfall (cm)	1.120579

$$VIF_i = \frac{1}{1 - R_i^2} = \frac{1}{\text{Tolerance}}$$

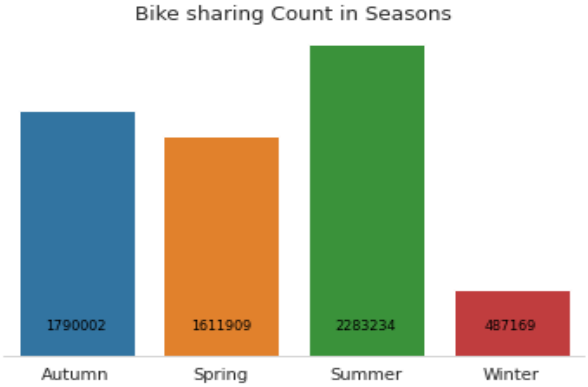
VIF value is under 5. Therefore we assume that the multicollinearity between the independent variables is negligible. So now basicially it is good to build the linear regression mode.

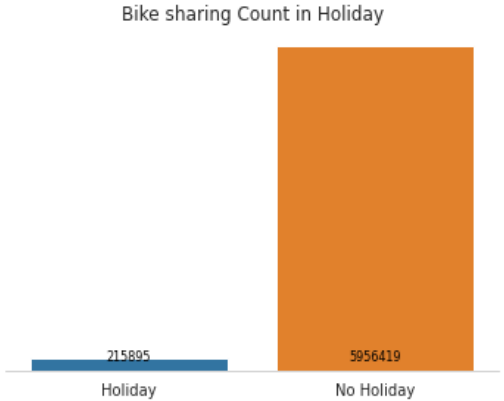
Demand for rental bikes during different hours of the day.

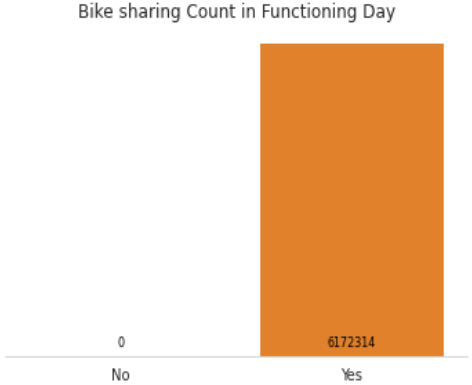


> People prefer to hire bikes during morning & evening hours, as seen by the sharp increase in rentals from 6:00 am to 9:00 am and 5 pm to 7 pm is the peak time.
> We can claim that there is a lot of high demand at the opening and closing hours of offices because it is apparent that demand increases steadily at 10 a.m.
> After 10:00 AM and through 6:00 PM, there is a steady increase in the demand for rental bikes.
> 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.
> Early hours(1:00 AM to 6:00 AM) is when there is least demand for bicycles. Regardless,of the seasons, this has been the general trend noticed.

Bike demand based on SEASONS, HOLIDAY & FUNCTIONING DAY.



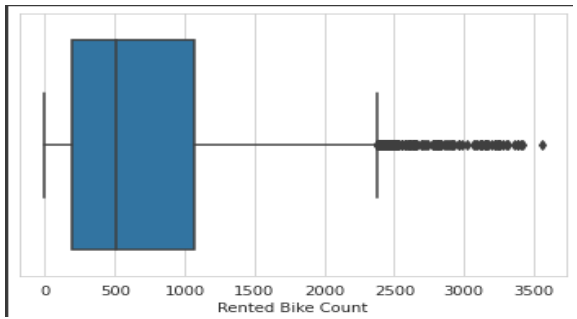


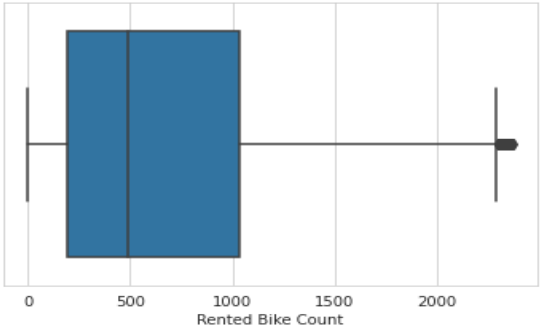


> Highest no. of bike were rented in the summer, the total no. of bikes rented in summer was 2.28 million.
> Second highest Bikes were rented in Autumn around 1.79 million followed by Spring in which 1.6 million bikes are rented.
> Winter appears to be the least popular season for bike rentals. In the winter, just 487K bikes were rented.
> People prefer to use the bike on Non-holiday more compared to Holidays.
> 5.9 million bikes are rented on Non-holidays, only about 216K (approx) bikes were rented on holidays.
> It's reasonable to conclude by looking at the data that the majority of clients in the bike rental sector are from Seoul's working class.
> All the bikes rented were on the functioning days.

OUTLIER DETECTION : BOX PLOT

Checking for outliers in the dataset by plotting a boxplot for the target variable “ Rented Bike Count” as it depicts presence of a high range of outliers.



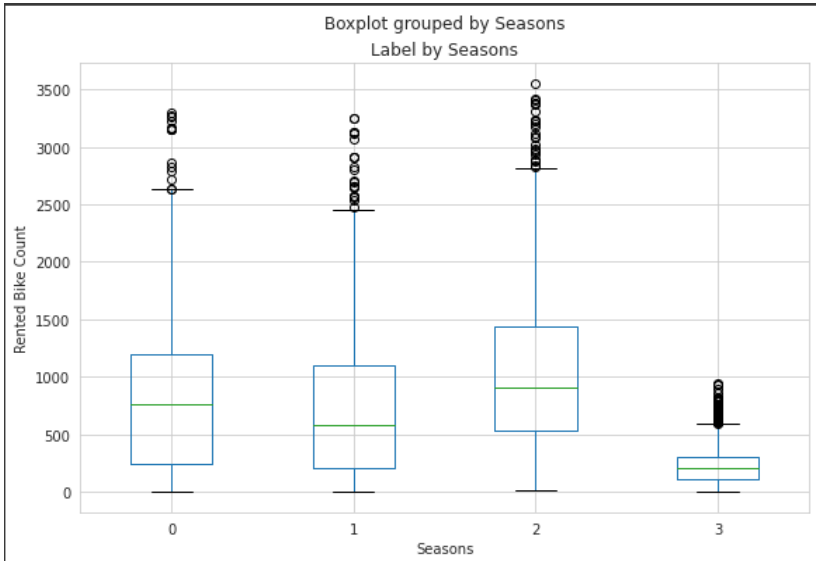


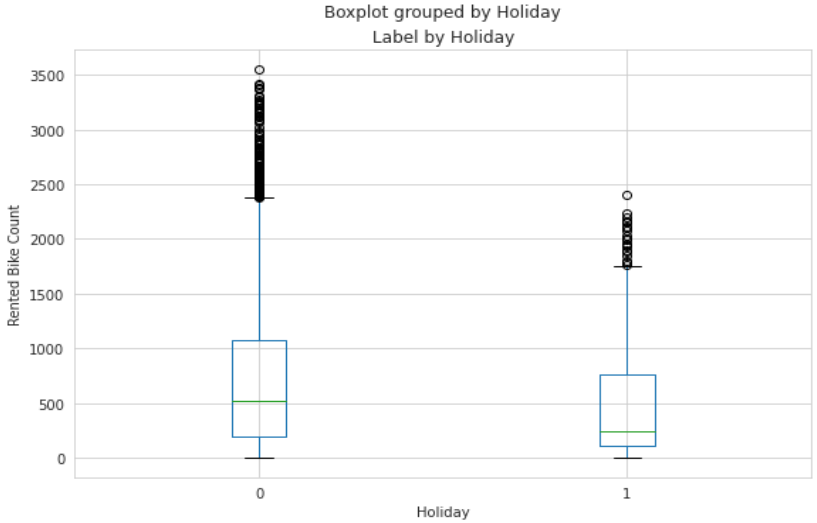
Using IQR process to remove outliers :

Q1 = df.quantile (0.25)
Q3 = df.quantile (0.75)
IQR (Inter quartile range) = Q3 – Q1

Lower Range = Q1 – 1.5*IQR
Higher Range = Q3 – 1.5*IQR
As we see there are many outliers in the dataset after using IQR process we dropped them for the accurate model predictions as seen in the box plot.

Outlier detection for categorical variable like seasons and holiday





By obtaining boxplot we can depict that they contain outliers values for both seasons (different seasons) & holiday.

Prepairing dataset for modelling

Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes
100	5	-6.4	37	1.5	2000	-18.7	0.0	0.0	0.0	Winter	No Holiday	Yes

Training data : (7008, 11)

Testing data : (1752,11)

For preparation of dataset for modelling we have splitted data into train & test and then defined functions to get evaluation metrics score and feature importance.

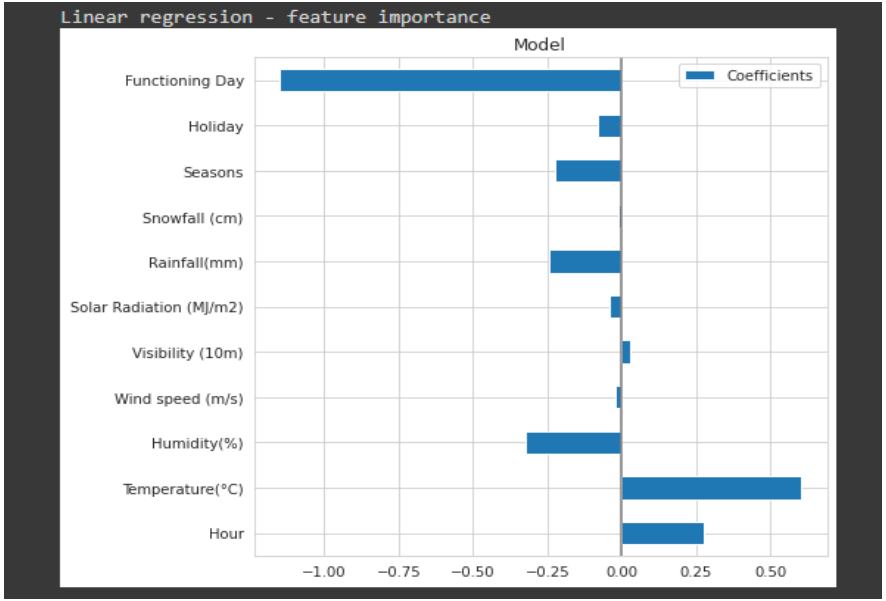
Linear Regression Analysis

Performance

Performance of linear regression

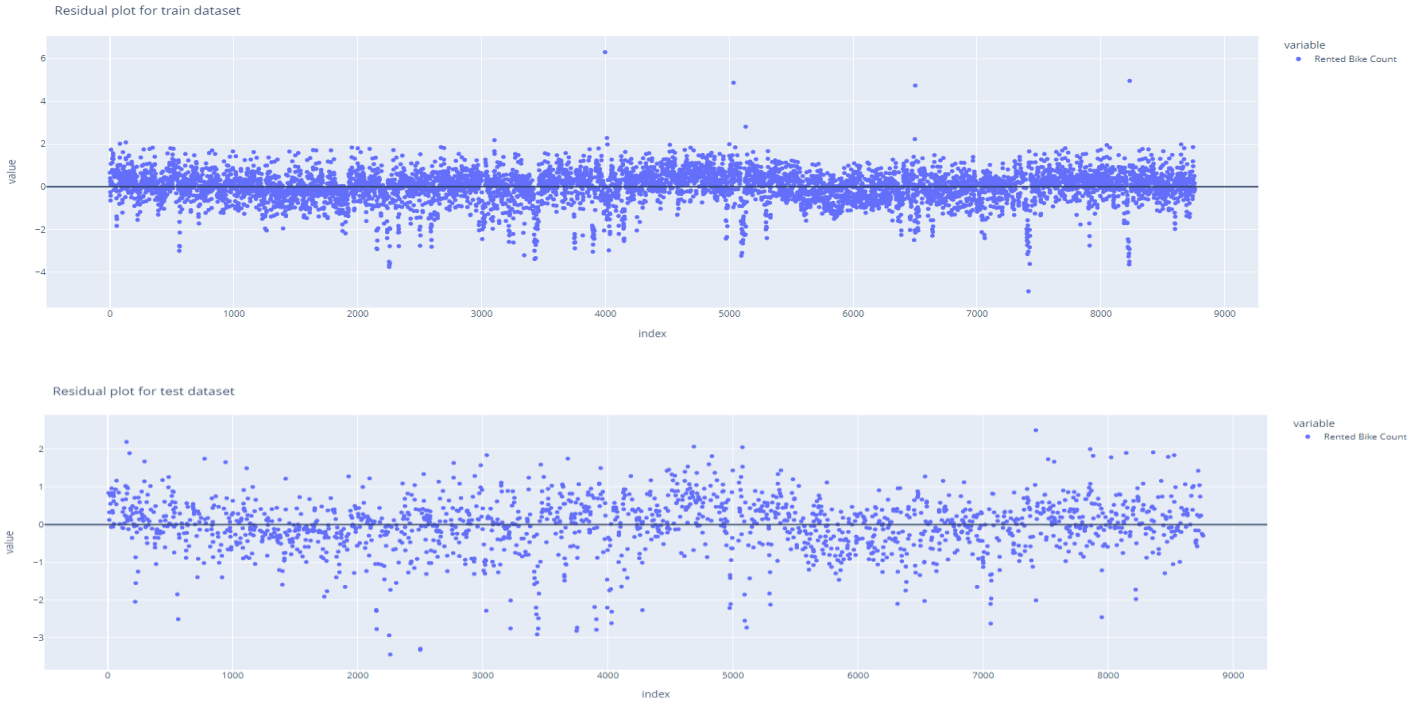
	Metric	Train Score	Test Score
0	MAE	287.44	279.07
1	MSE	202203.58	195114.04
2	RMSE	449.67	441.72
3	r2	0.51	0.53
4	adj_r2	0.51	0.53

Feature Importance



> A model is said to be the best model when the R2 score is close to 1 and the MAE is low.
> With an R2 score of .53, linear regression fails in terms of accuracy.

Residual analysis of linear regression



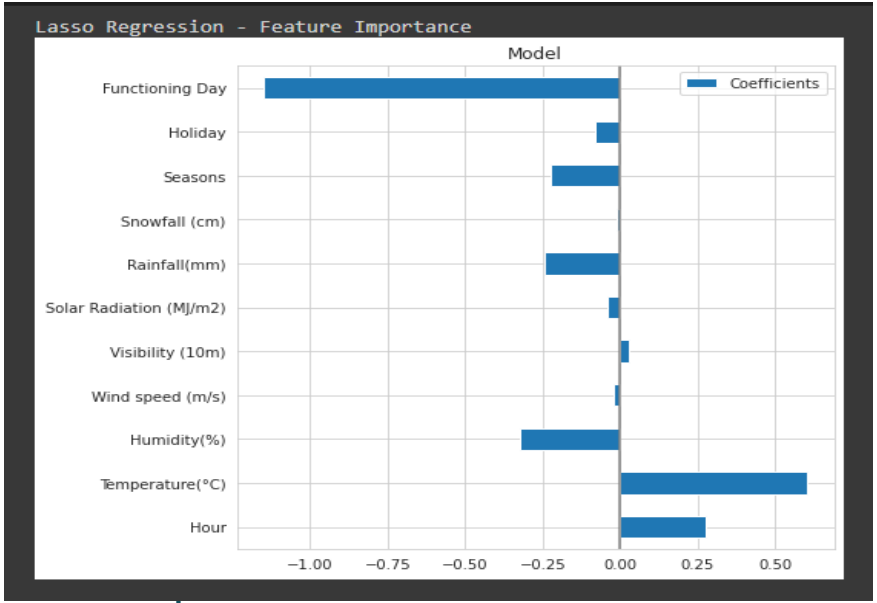
Here we can clearly see that there is homoscadascity in between them for both train & test dataset.

Lasso Regression Analysis

Performance

	Metric	Train Score	Test Score
0	MAE	287.44	279.07
1	MSE	202203.58	195114.04
2	RMSE	449.67	441.72
3	r2	0.51	0.53
4	adj_r2	0.51	0.53

Feature Importance



> The performance of lasso regression is also very poor.

> Even after applying cross validation and hyper parameter tuning there has been no change in the lasso regression score.

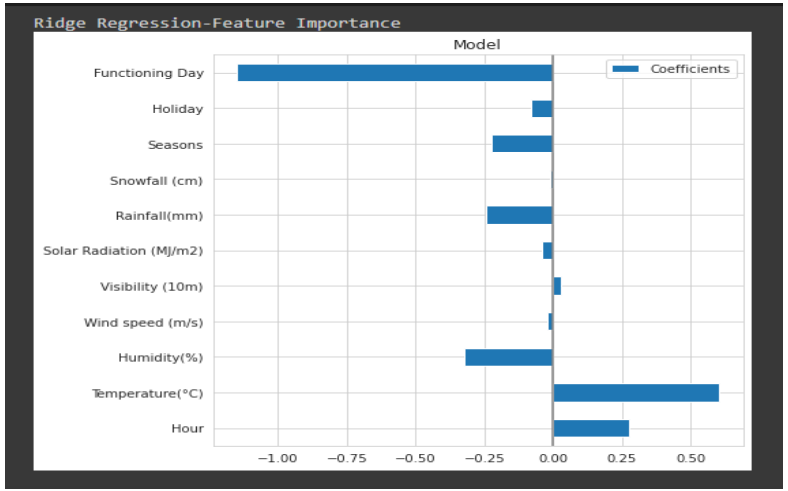
Ridge Regression Analysis

Performance

Performance of Ridge Regression

	Metric	Train Score	Test Score
0	MAE	287.45	279.08
1	MSE	202213.50	195127.98
2	RMSE	449.68	441.73
3	r2	0.51	0.53
4	adj_r2	0.51	0.53

Feature importance



› We can clearly see that the R2 score of linear regression is very poor. Despite using regularization methods like cross validation and hyper parameter tuning there is no change in result, it's still poor.

› Also we know that regularization method helps in fixing over fitting but here our model was not over fitted.

› A possible reason of such low accuracy using the linear model would be low linear relationship & low correlation between target variable & independent variables.

RANDOM FOREST

Random Forest Scores

	Metric	Train Score	Test Score
0	MAE	54.01	140.15
1	MSE	9310.75	55430.98
2	RMSE	96.49	235.44
3	r2	0.98	0.87
4	adj_r2	0.98	0.87

performance of Random Forest after cross validation and hyperparameter tuning

	Metric	Train Score	Test Score
0	MAE	173.31	177.37
1	MSE	82263.64	87740.31
2	RMSE	286.82	296.21
3	r2	0.80	0.79
4	adj_r2	0.80	0.79

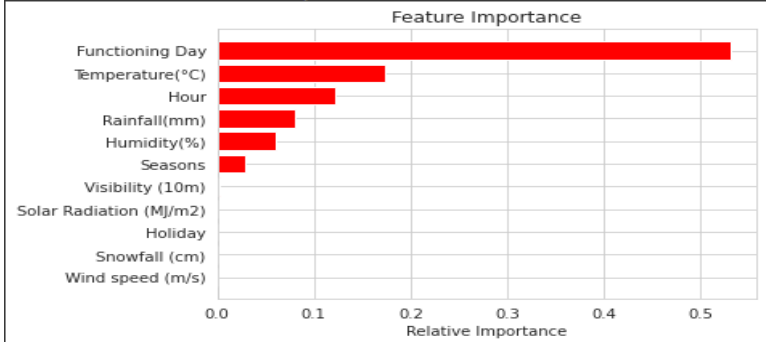
So here we can see the drastic change in the r2 & adj_r2 score as compared to linear regression.

Random forest regressor gives r2 & adj_r2 score of 0.87, 0.86.

Also we can see that r2 and adj_r2 both have different train and test score.

Here we can clearly see that before applying cross validation and hyper parameter tuning train score and test score were dissimilar as they had very big difference but after applying CV & HPT both score are almost identical.

Random Forest-Feature Importance



In random forest the feature variable 'functioning day' has the highest impact on the dependent variable "Rented Bike count". The second most important feature is "Temperature".

Gradient Boosting Machine

Evaluation metrics for GBM

	Metric	Train Score	Test Score
0	MAE	196.04	196.02
1	MSE	104082.42	108884.33
2	RMSE	322.62	329.98
3	r2	0.75	0.74
4	adj_r2	0.75	0.74

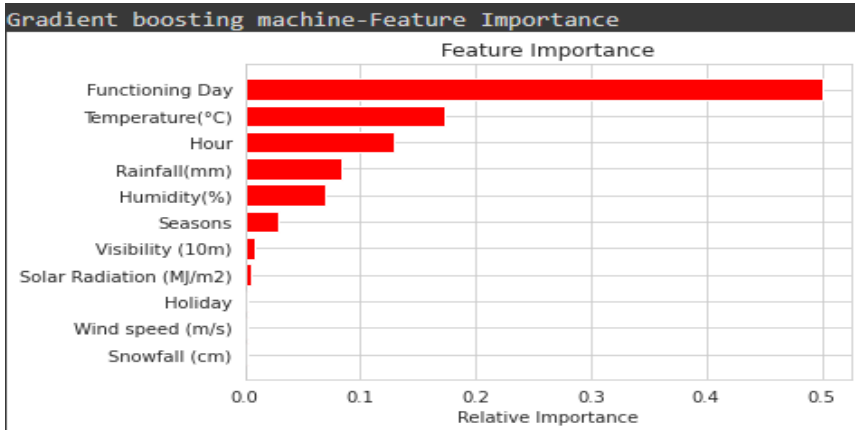
Here the train and test r2 and adjusted r2 are almost identical/same.

Scores of GBM after cross-validation and hyperparameter tuning

	Metric	Train Score	Test Score
0	MAE	126.79	141.29
1	MSE	43053.83	55873.83
2	RMSE	207.49	236.38
3	r2	0.90	0.87
4	adj_r2	0.90	0.87

After applying cross validation and hyper parameter tuning we can see the change in the r2 and adjusted r2 score for train and test data.

In GBM also the feature variable 'functioning day' has the highest impact on the dependent variable "Rented Bike count" and the second most important feature is "Temperature".



XGBoost Analysis

Evaluation metrics for XGBoost

	Metric	Train Score	Test Score
0	MAE	120.88	146.87
1	MSE	39415.53	59618.05
2	RMSE	198.53	244.17
3	r2	0.91	0.86
4	adj_r2	0.90	0.86

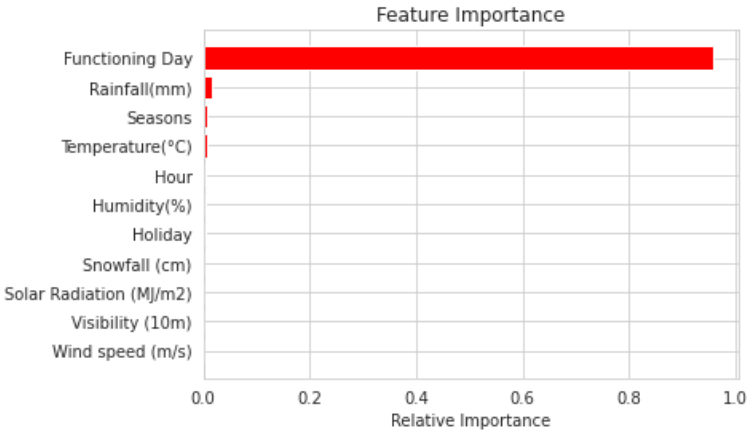
Evaluation metrics for XGBoost after cross-validation and hyperparameter tuning

	Metric	Train Score	Test Score
0	MAE	75.20	139.62
1	MSE	16533.00	54594.92
2	RMSE	128.58	233.66
3	r2	0.96	0.87
4	adj_r2	0.96	0.87

As we can see that XGBOOST gives the best r2 and adj_r2 score among all other models.

In GBM also the feature 'functioning day' has the highest impact on the dependent variable "Rented Bike count" and the second most important feature here is "Rainfall".

XGBoost-Feature Importance



Scores of all different ML Models after cross validation & hyper parameter tuning.

Evaluation metrics for different ML algorithms:

	Model	Train MAE	Test MAE	Train MSE	Test MSE	Train RMSE	Test RMSE	Train r2	Test r2	Train adj r2	Test adj r2
0	Linear	287.44	279.07	202203.58	195114.04	449.67	441.72	0.51	0.53	0.51	0.53
1	Lasso	287.44	279.07	202203.58	195114.04	449.67	441.72	0.51	0.53	0.51	0.53
2	Ridge	287.45	279.08	202213.50	195127.98	449.68	441.73	0.51	0.53	0.51	0.53
3	Random Forest CV	173.31	177.37	82263.64	87740.31	286.82	296.21	0.80	0.79	0.80	0.79
4	GBMCV	126.79	141.29	43053.83	55873.83	207.49	236.38	0.90	0.87	0.90	0.87
5	XGboost CV	75.20	139.62	16533.00	54594.92	128.58	233.66	0.96	0.87	0.96	0.87

1. R2 score in Linear Regression is 0.51 for the train data and 0.53 for the test data. Thus the Linear Regression model fails in this case.
2. Comparing the R2 score of all the models, one can see that XGBoost performs better than other models.
3. Gradient Boosting Machine has a test accuracy of 86%, making it the second-best model.
4. Random Forest is also found to perform well on the data.

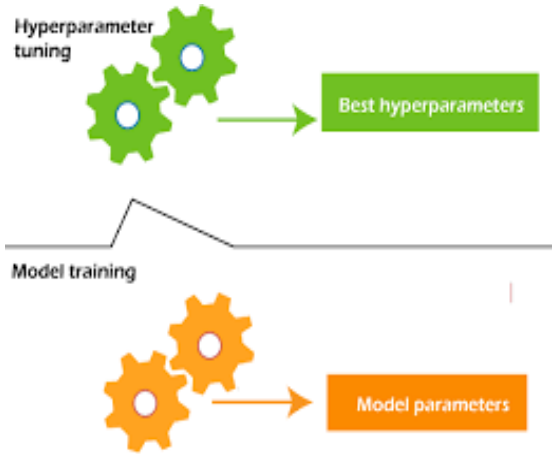
Best Parameters

Random Forest

‘n_estimators’ : 80
‘min_samples_split’ : 50
‘min_samples_leaf’ : 40
‘max_depth’ : 8

GBM

‘n_estimators’ : 100
‘min_samples_split’ : 50
‘min_samples_leaf’ : 60
‘max_depth’ : 8



XGBoost

‘eval_metric’ : ‘rmse’
‘n_estimators’ : 500
‘objective’ : ‘reg:squarederror’
‘max_depth’ : 6

OBSERVATIONS

* The feature variable Functioning Day has the highest impact on the dependent variable Rented Bike Count.

* In Random Forest and GBM, Temperature is making an impact while Rainfall is the second most important factor in XGBOOST.

* Random Forest & GBM gives importance to 6-7 features while XGBOOST considers only the top 3-4 features and almost neglects all the rest.

CONCLUSION

* The project focuses on predicting bike sharing demand using the Seoul dataset.

* In this prediction we used different models and after getting the evaluation score we came to a conclusion that out of all the models used XGBOOST performs the best with the train R2 score of 0.96 & the test R2 of 0.87.

* XGboost gives the least MAE among the models. The most important features for predicting the dependent variable (number of hired bikes) for XGBoost are functioning day, rainfall, season, and temperature.

* This project will be helpful for the company to predict the hourly bike demand and enhance the user experience.

Thank You