Capstone Project - 2

Supervised ML - Regression

Topic - Bike Sharing Demand Prediction

By - Rahul Shah

CONTENTS OF THE PRESENTATION

Introduction

Seoul city in south korea has a rental bike sharing program. The common public can pick up and drop the rental bike in many different bike stands. It is an un-manned rental system that can be used anywhere, anytime by anyone.

Bike sharing is an innovative approach to urban mobility, it was designed to resolve the issues of traffic congestion, air pollution and high oil prices in seoul and to build a healthier society while enhancing the quality of life for its citizens.

Bike sharing systems are a means of renting bicyles where the process of obtaining membership, rental and bike return is automated via a network of kiosk locations throughout a city.

Problem 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.

The system manager for such a program would ideally like to predict the demand to make sure the city doesn't rent less bikes than it requires. Eventually, providing the city with a stable supply of rental bikes becomes a major concern.

The aim of this project is to predict the number of rented bikes using different techniques.

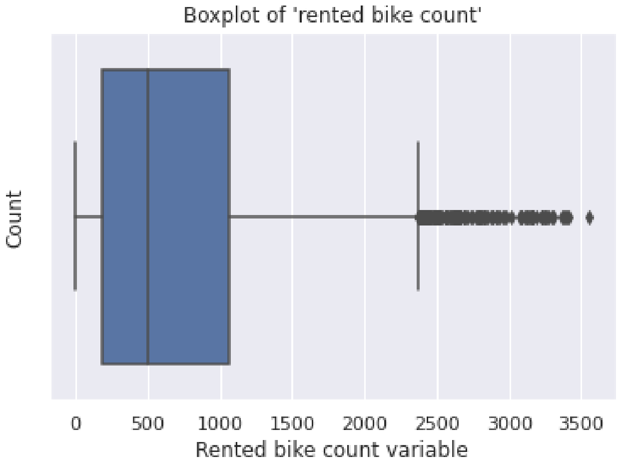
Data Summary

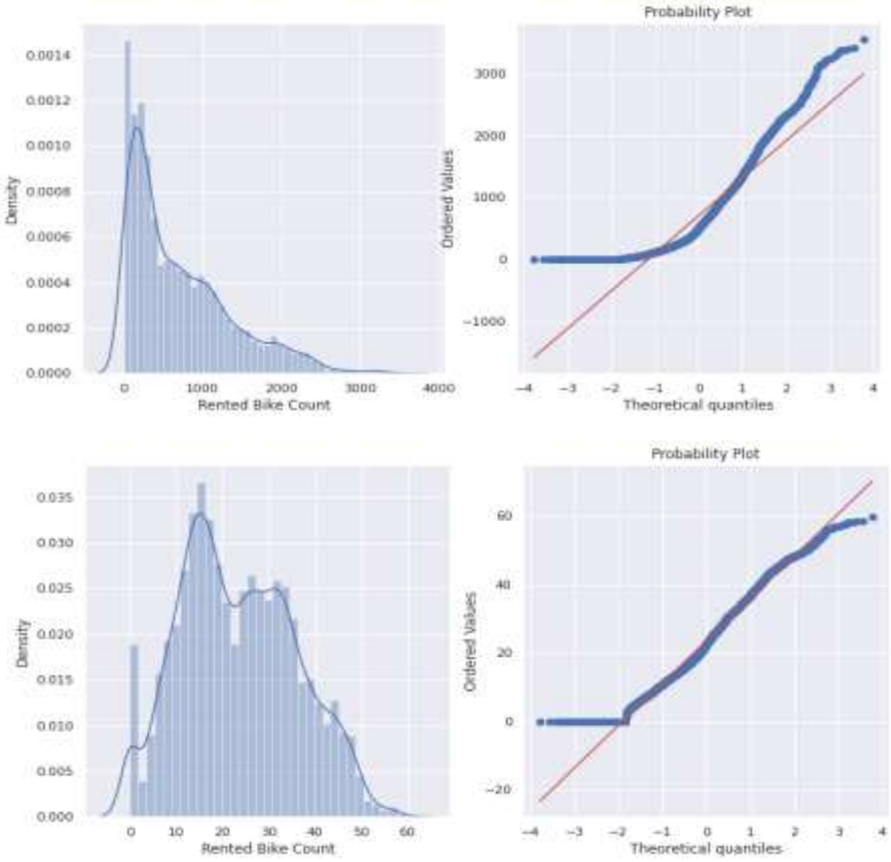
1. Date - This describes the actual date at which bike ride was taken
2. Rented Bike Count - This is the target variable and it tells us the number of bike rides taken by individuals.
3. Hour - It describes us the time or bike rides, we can interpret the peak times.
4. Temperature - It describes us the local temperature of the location during the bike rides.
5. Humidity - This describes the level of humidity in the weather during the ride.
6. Wind speed - This describes the average speed of wind while bike ride was taken
7. Visibility - This describes the outside environment visibility which might be affected due to adverse weather conditions sometimes like fog.
8. Dew Point temperature (in celsius) - This indicates the amount of moisture in the air.
9. Solar radiation - MJ/m2 - This describes us the amount of ultraviolet radiation.
10. Rainfall (mm) - This describes us the measurement of rainfall helps us to check if the rainfall is heavy or light.
11. Snowfall (cm) - This describes us the measurement of snowfall helps us to check if the rainfall is heavy or light.
12. Seasons - It indicates the the type of season like autumn, summer,spring,winter.
13. Holiday - It indicates whether it was official holiday or not
14. Functional Day - It indicates whether the bike ride was during functioning hours or non functioning hours .

EDA - Univariate analysis

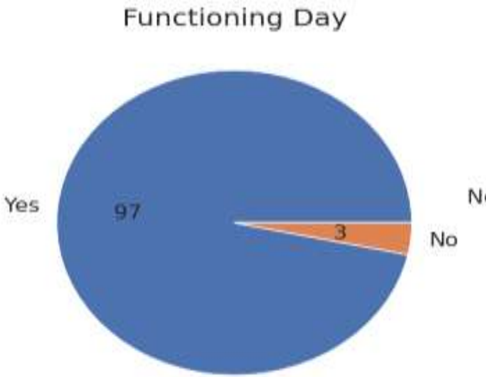
1. **"Rental Bike Count" - Dependent variable(dv)**
   - **Outliers above 2500**
   - **Was moderately skewed, positively**

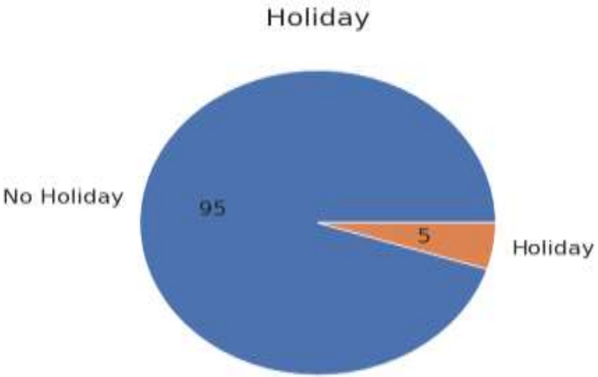
Skewness: 1.153428 Skewness after transformation:0.237362
Kurtosis: 0.853387 Kurtosis after transformation:-0.657201

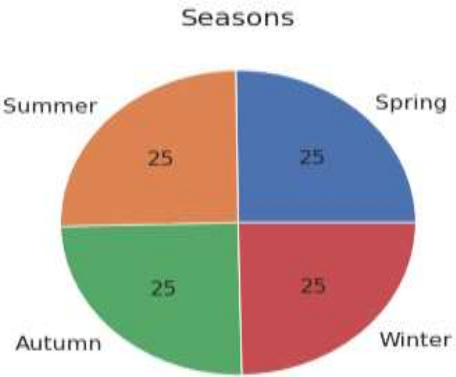


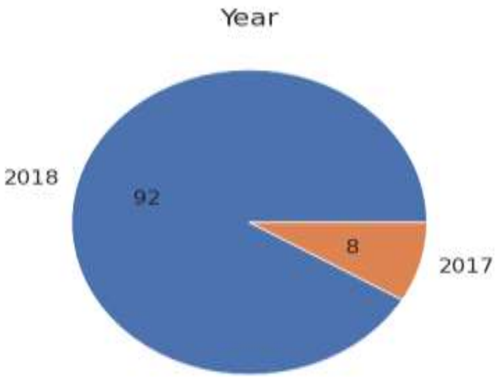


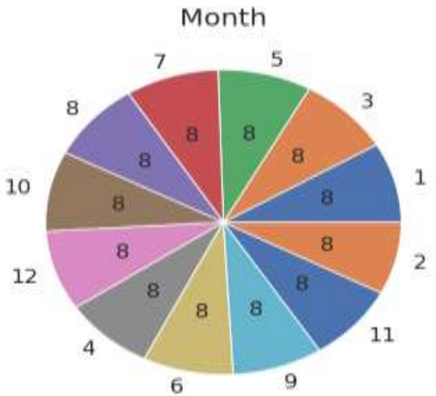
EDA continued…

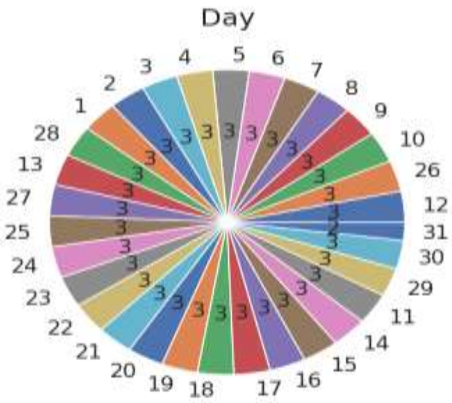






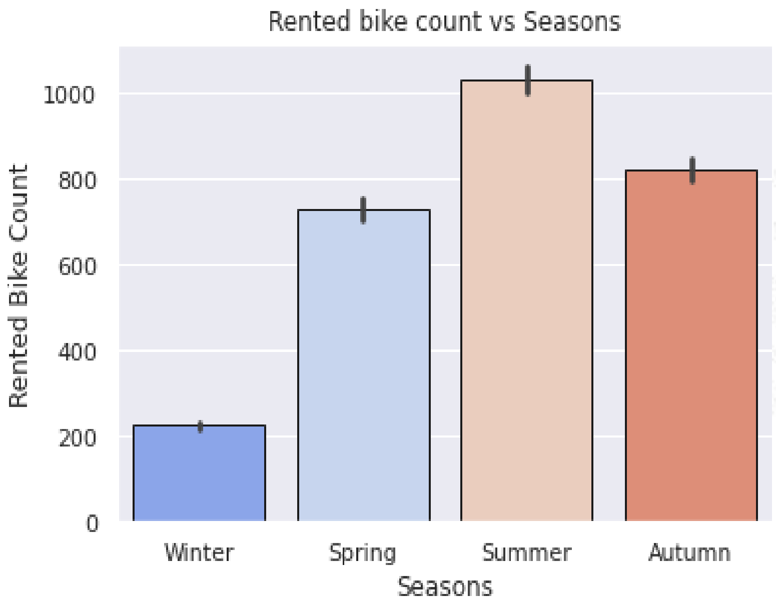


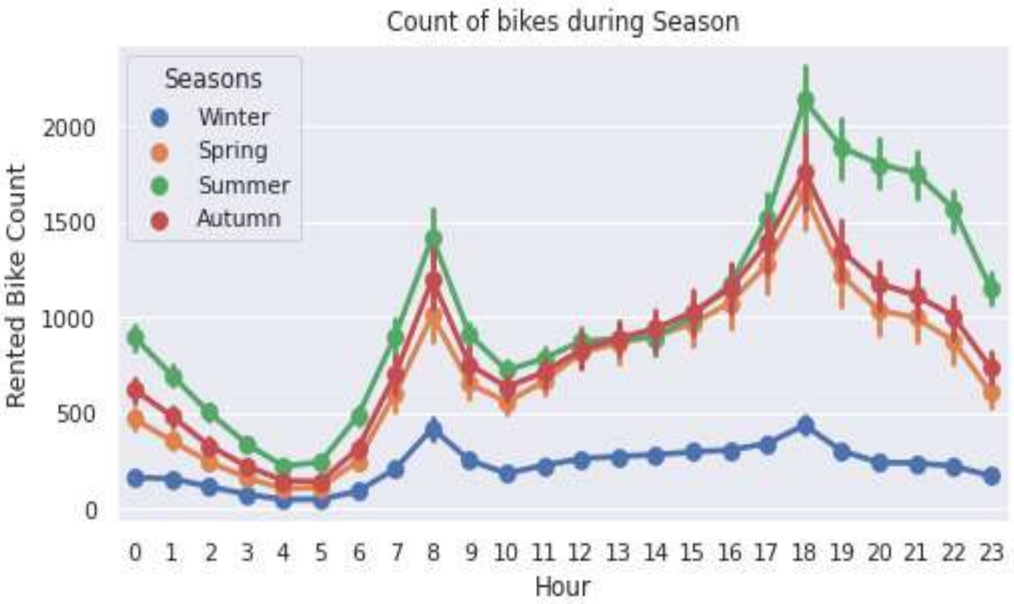




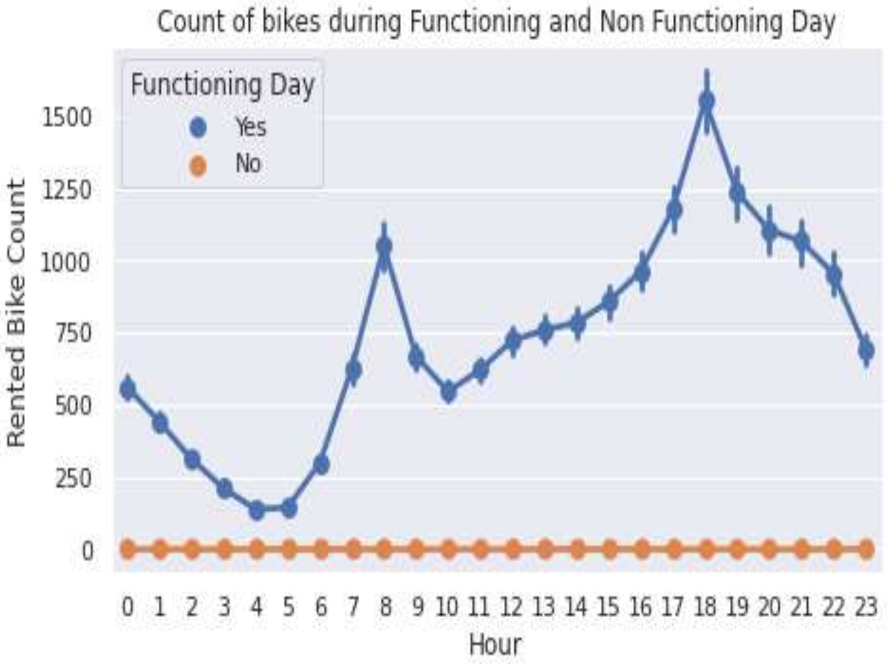
Multivariate analysis

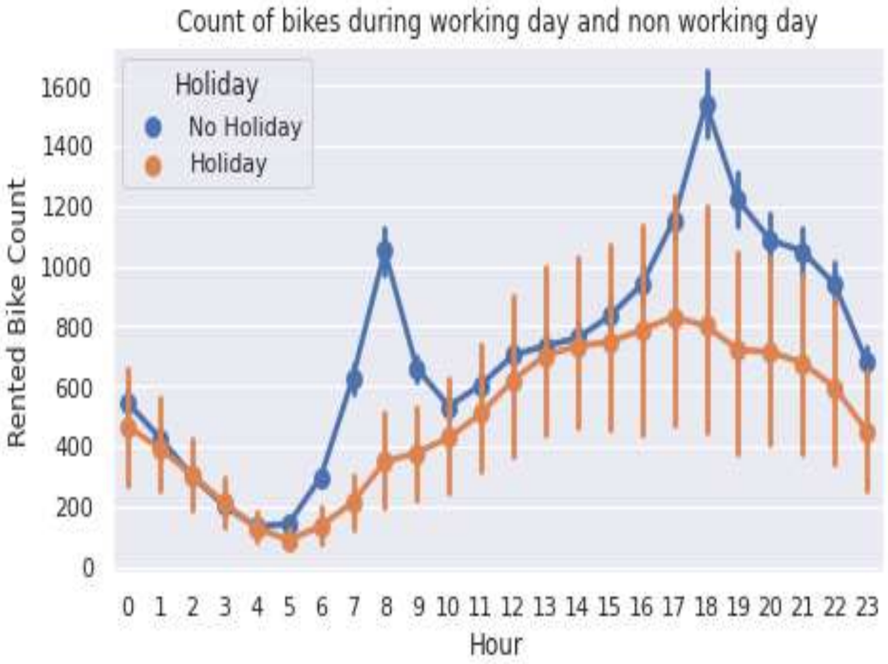
Demand for bikes was high during summer compared to winter



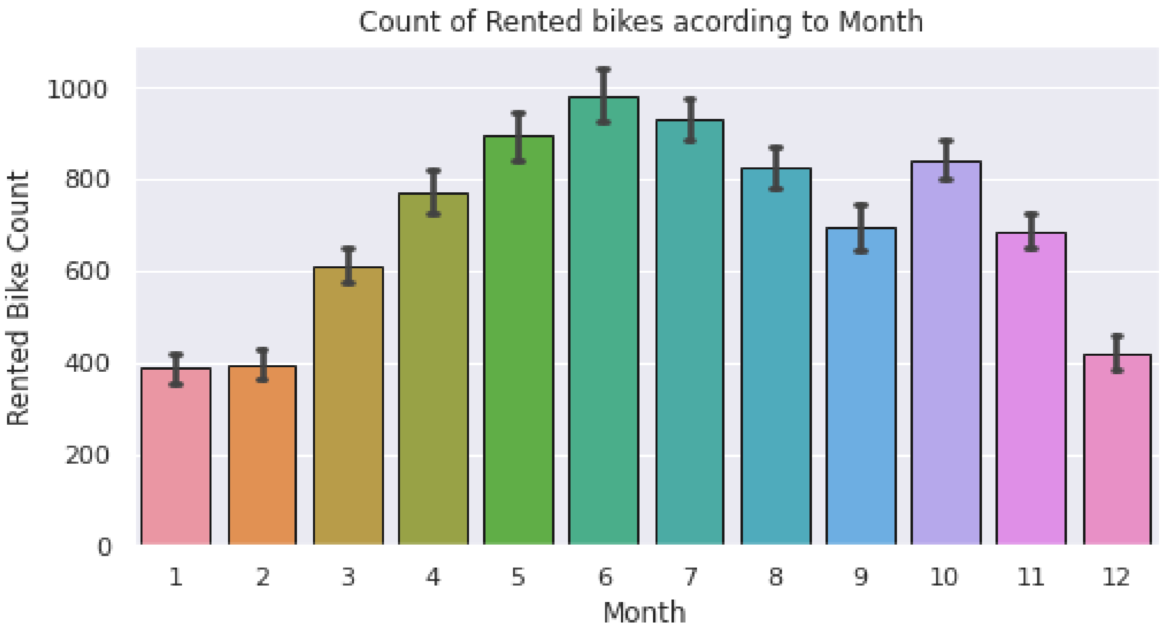


Multivariate analysis



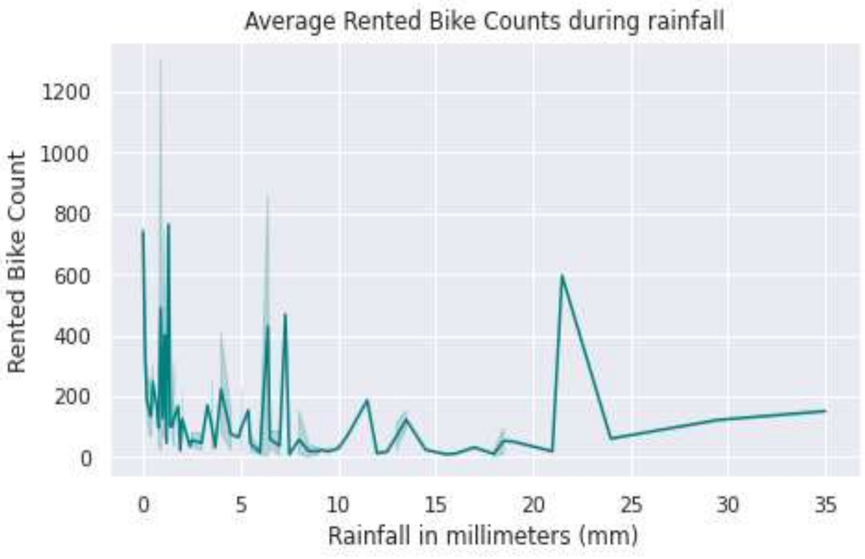


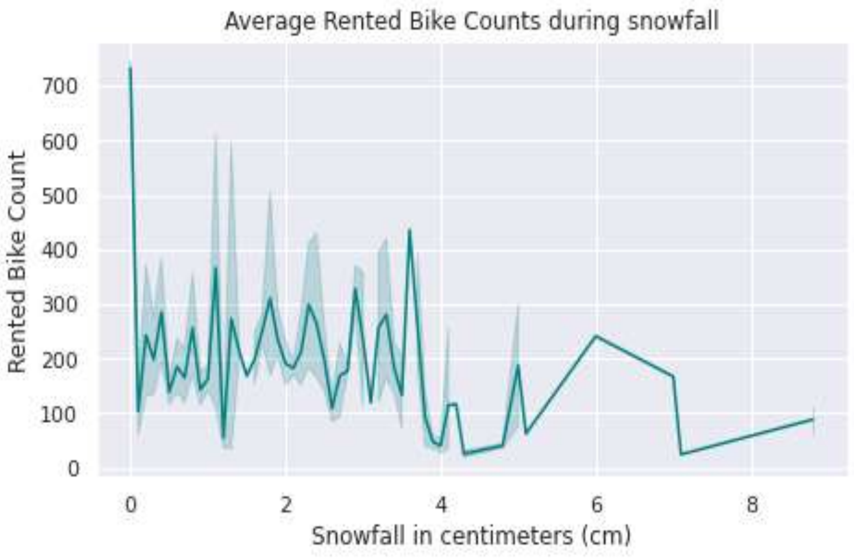
Continued…



The number of rented bikes count is higher in the month of june compared to other months

Continued…

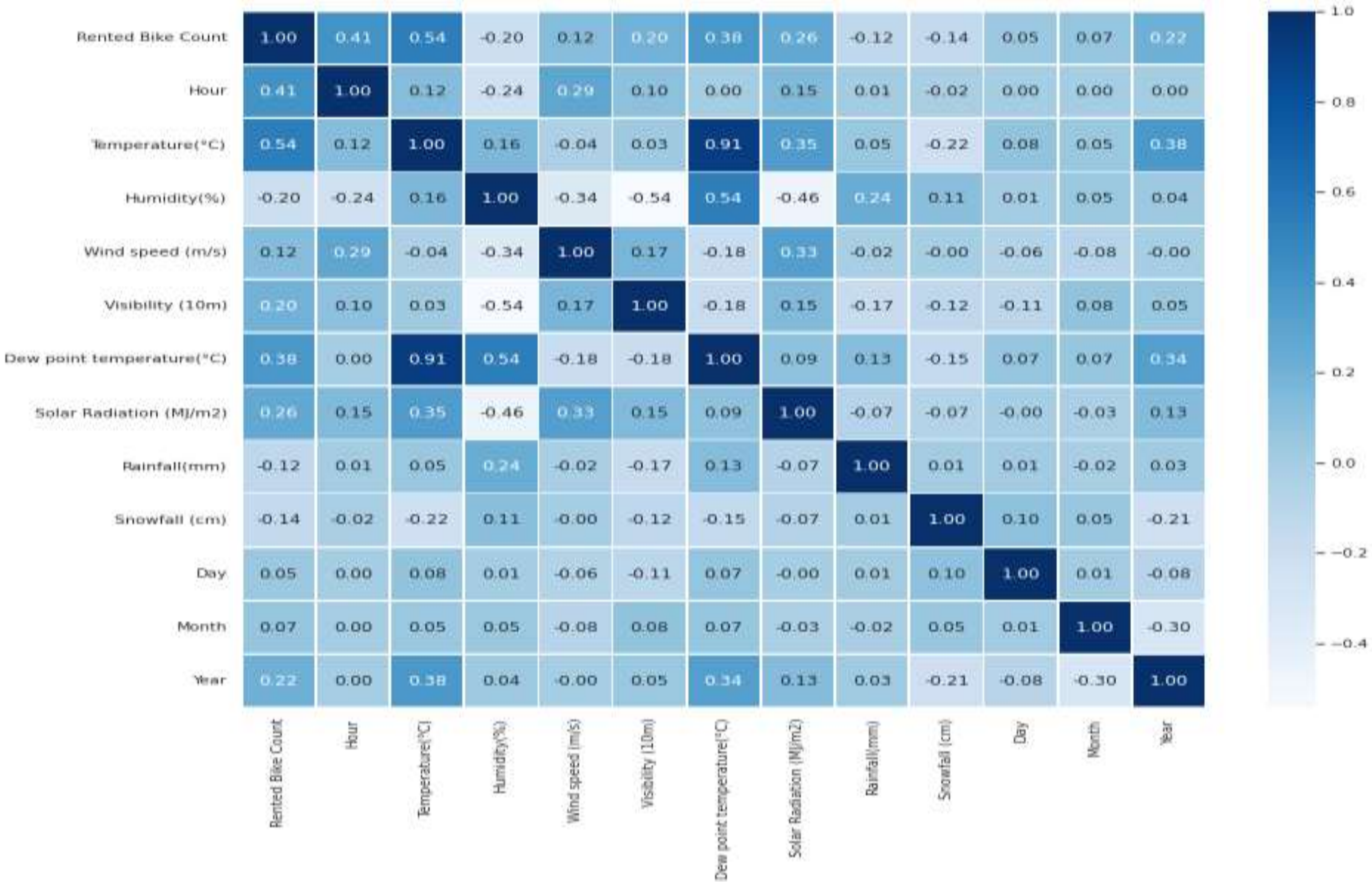




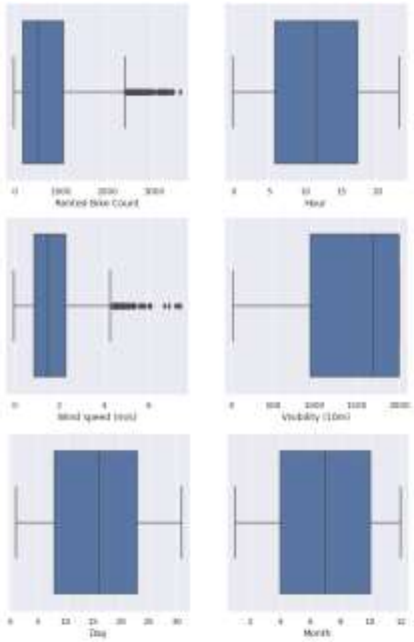
- When the rainfall is less than 8mm people take more bikes on rent. But, we can also see peak in between 20mm to 25mm.
- Demand for rented bikes are high when the snowfall is less than 4 cm.

Multivariate analysis - Correlation

- Dew point temperature and Temperature were highly correlated.
- Linear regression assumes that independent variables must show some linear relationship with dependent variable.
- No such relationship seen here.
- Linear regression might not perform well.

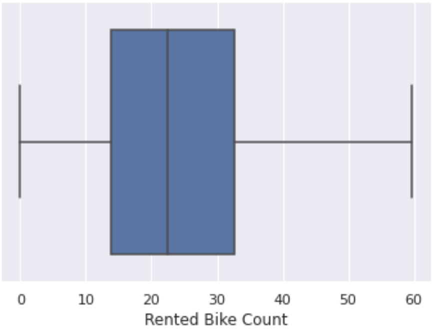


Data Wrangling - missing values and outliers



The no. of missing values in each variable:

Date	0
Rented Bike Count	0
Hour	0
Temperature(°C)	0
Humidity(%)	0
Wind speed (m/s)	0
Visibility (10m)	0
Dew point temperature(°C)	0
Solar Radiation (MJ/m2)	0
Rainfall(mm)	0
Snowfall (cm)	0
Seasons	0
Holiday	0
Functioning Day	0
Day	0
Month	0
Year	0



- No missing values
- Tackled outliers in rented bike count by applying transformation
- Windspeed and solar radiation had outliers, but they were not that far from maximum values.

Feature Modification & Feature Selection

- Converted “Date” column from object data type to DateTime data type.
- Extracted Day, Month and Year from “Date” column.
- One hot encoded feature ‘Seasons’.
- Removed observations where it was “Non-Functional Day” and bike rented count was zero. And removed this column because it had constant values
- Removed features that were not necessary such as ‘Date’ and ‘Year’.
- Removed feature ‘Dew point temperature(°C)’ as it was highly correlated with ‘Temperature’.

Machine Learning Models

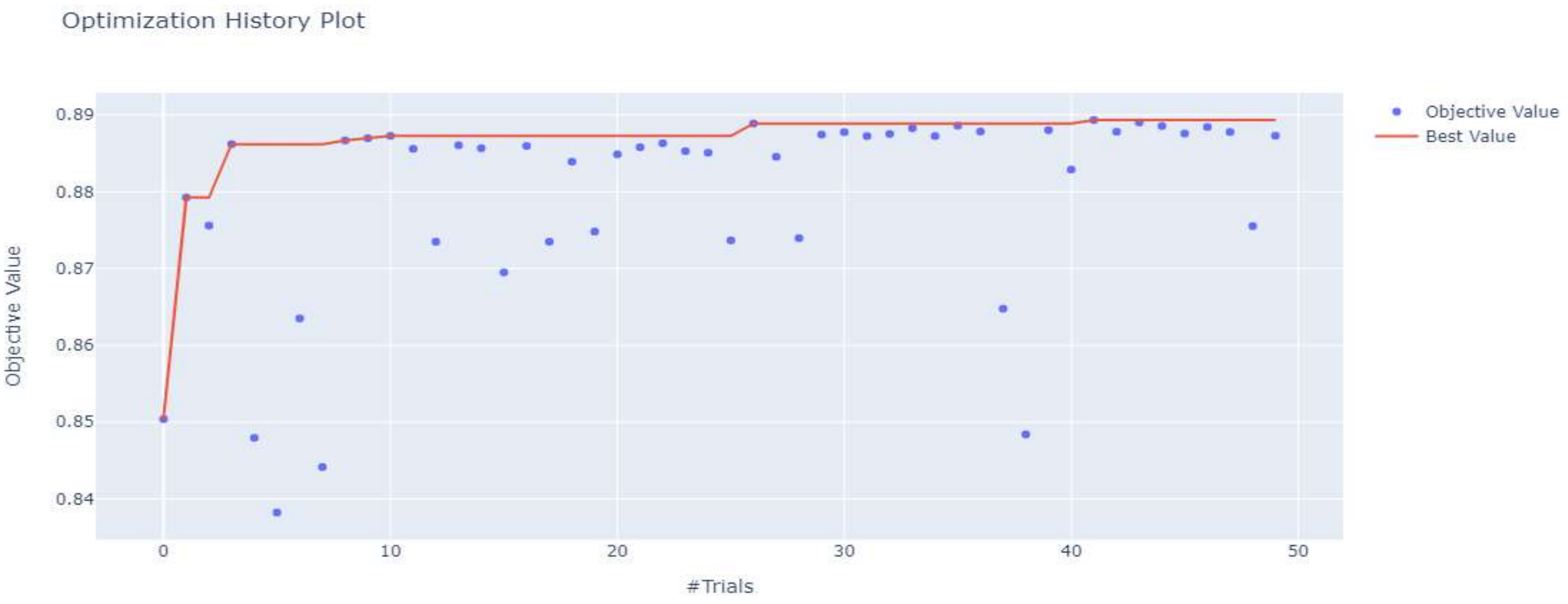
Four models were used : Linear regression, Decision Tree, Random Forest and XGBoost.

The evaluation results are:

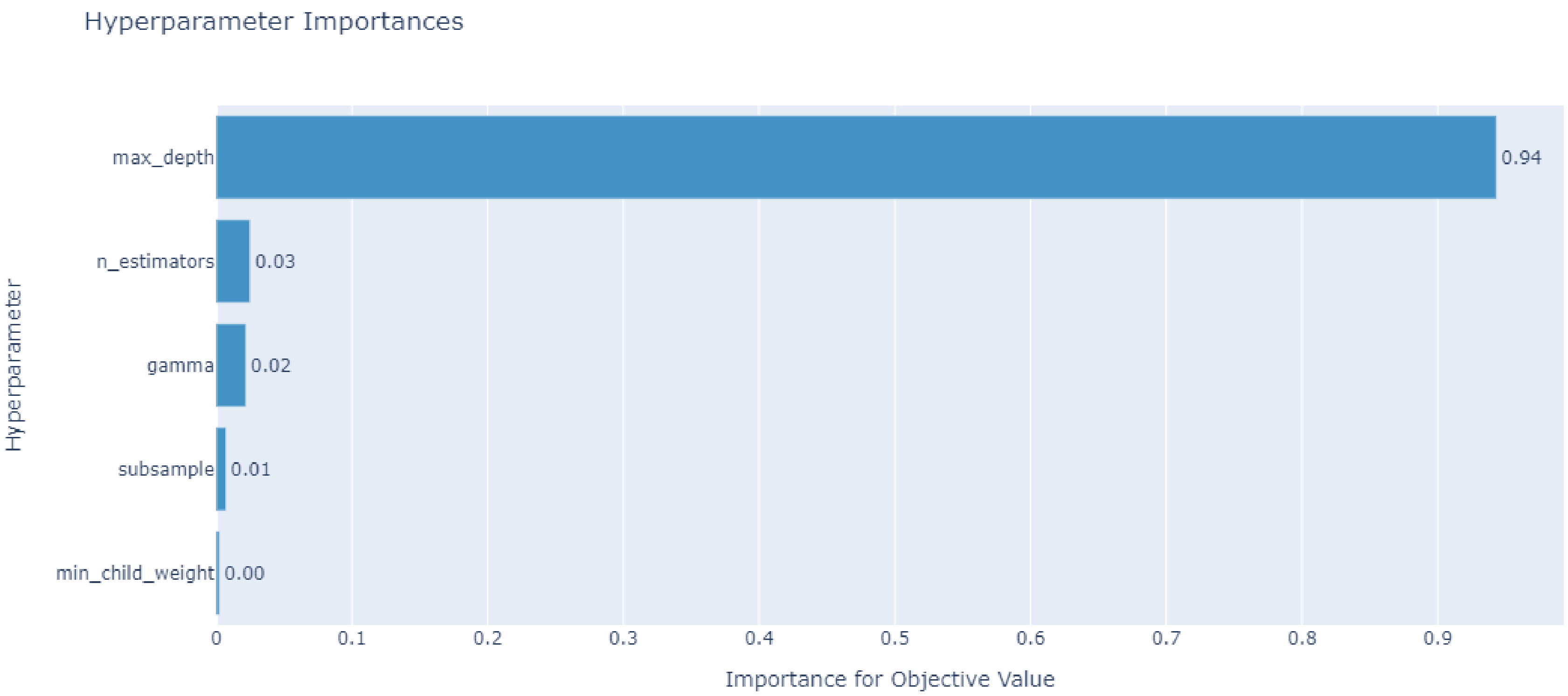
	Model_Name	train_mae	train_mse	train_rmse	train_r2	train_adjr2	test_mae	test_mse	test_rmse	test_r2_	test_adjr2
0	Linear Regression	287.759535	178789.625136	422.835222	0.571559	0.570671	283.514542	172950.803081	415.873542	0.559945	0.556273
1	Decision Tree Regressor	0.000000	0.000000	0.000000	1.000000	1.000000	188.847608	109541.161843	330.970032	0.721284	0.718958
2	Decision Tree Regressor - tuned	132.770212	44908.794869	211.916953	0.892383	0.892160	169.820570	74344.319982	272.661548	0.810839	0.809260
3	Random Forest Regressor	49.894980	7002.793088	83.682693	0.983219	0.983184	140.809651	54450.777221	233.346903	0.861456	0.860300
4	Random Forest Regressor - Tuned	160.071837	62743.059443	250.485647	0.849646	0.849335	178.119963	77830.851253	278.981812	0.801967	0.800315
5	XGBoost Regressor	148.383724	56058.956955	236.767728	0.865664	0.865385	162.653526	66087.857202	257.075587	0.831846	0.830443
6	XGBoost Regressor - tuned	134.744784	46695.918687	216.092385	0.888101	0.887869	152.258234	59919.354810	244.784303	0.847541	0.846269

XGBoost Regressor – Optuna Visualizations

Optuna – Optimization History Plot

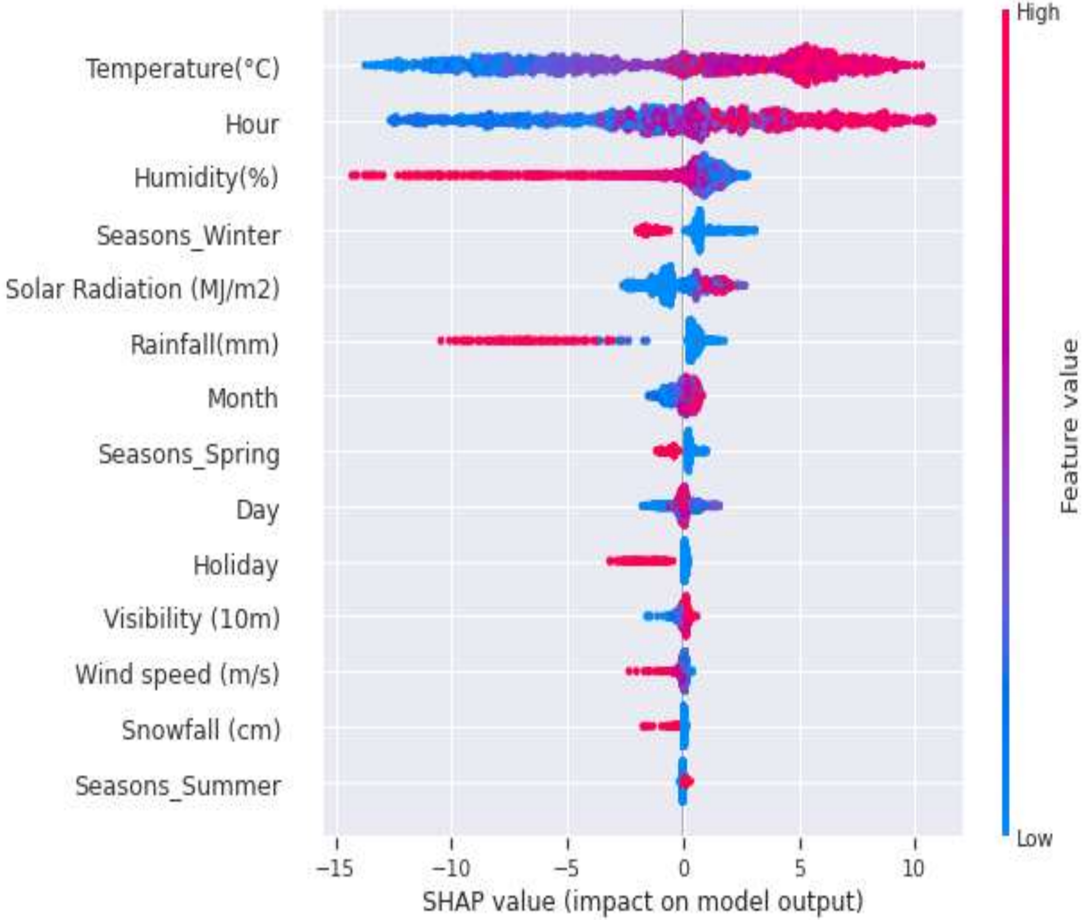


Optuna – Hyperparameter Importances Plot



Model Explanation

The most important features were Temperature, Hour, Humidity, Seasons_Winter



Conclusion

- We found that Linear Regression performed poorly as expected. Decision Tree and Random Forest showed overfitting. The best performance was given by the XGBoost model.
- Hyperparameter Tuning was one of the challenging task.
- We also implemented shap technique to understand the working of our XGBoost model:
    1. Temperature was the most important feature. Demand for bikes was higher when temperature was high.
    2. Hour of the day was the second most important feature. Demand was high during evening hours.
    3. Demand was less in winter season as compared to other seasons.
    4. Demand for bikes increases with increase in solar radiation.
- After a long exercise we concluded that ensemble learning approach such as XGBoost improves the model performance considerably.

THANK YOU