Capstone Project

ML - Supervised learning (Regression)

Title: Bike Sharing Demand Prediction

Team members:

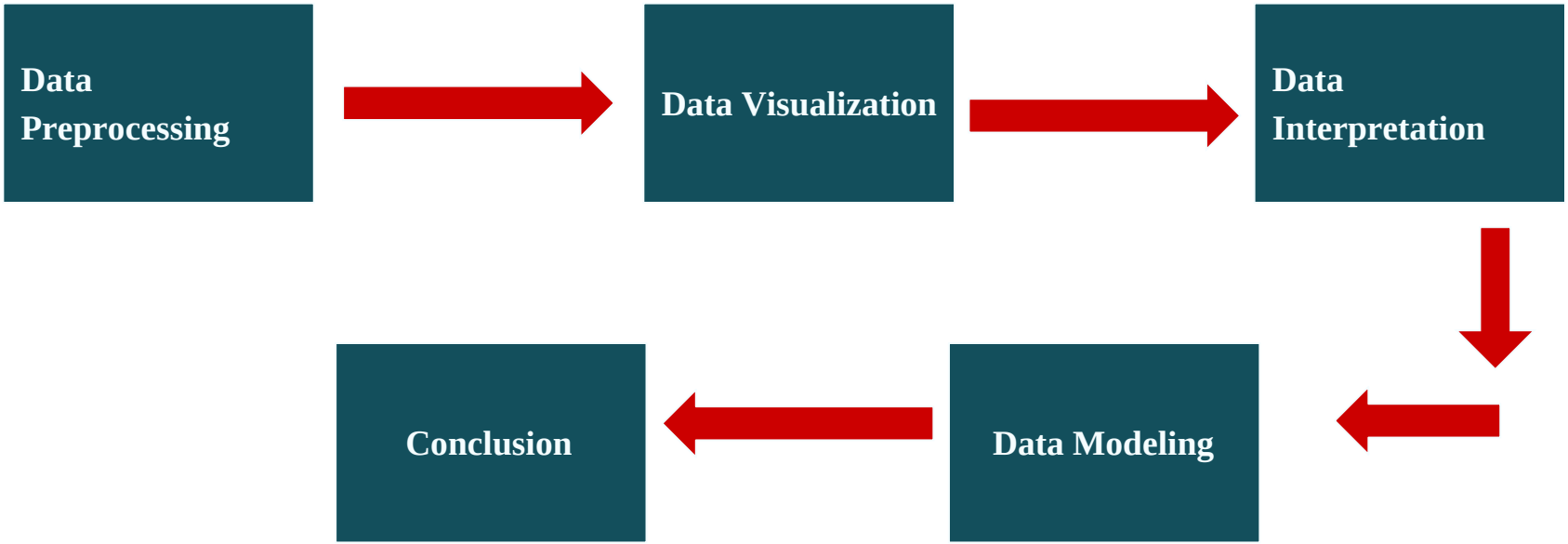
Manjari Lahariya

Manoj Patil

Likhith Ram

Index

Data Pipeline



Topic :- Seoul Bike Sharing Demand Prediction

A bike-sharing system provides people with a sustainable mode of transportation and has beneficial effects for both the environment and the user. And the city-wide accessibility and low cost has exponentially increased its popularity. Nonetheless, the increased usage has led to issues like unavailability of bikes at bike stations. Therefore, the study aims to predict the demand of a bike sharing system using machine learning models. And also analyses the comprehensive effect of the time dependent and inter-station relationship on predicting the demands.

Data Set Overview

- Source dataset is in ‘txt’ format with ‘.csv’.
- Dataset contains 8760 rows and 14 columns.
- Out of the total, 13 Columns are independent variables and 1 is dependent variable.
- There are no missing values for the provided input dataset.
- Rented Bike Count is the number of total bikes rented per hour and this is our dependent variable to be predicted.

```
#Overview of given dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 14 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Date                       8760 non-null   object
 1   Rented Bike Count          8760 non-null   int64
 2   Hour                       8760 non-null   int64
 3   Temperature(°C)            8760 non-null   float64
 4   Humidity(%)                8760 non-null   int64
 5   Wind speed (m/s)           8760 non-null   float64
 6   Visibility (10m)           8760 non-null   int64
 7   Dew point temperature(°C)  8760 non-null   float64
 8   Solar Radiation (MJ/m2)    8760 non-null   float64
 9   Rainfall(mm)               8760 non-null   float64
 10  Snowfall (cm)              8760 non-null   float64
 11  Seasons                    8760 non-null   object
 12  Holiday                    8760 non-null   object
 13  Functioning Day            8760 non-null   object
```

Data Set Summary

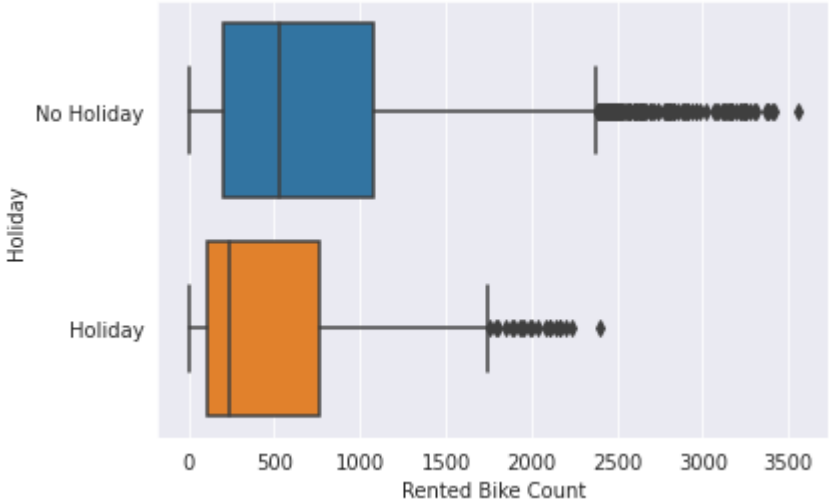
- **Date :(string)**: Date in year-month-day format
- **Rented Bike count :(integer)**: Number of rented bikes per hour which is the target
- **Hour :(integer)**: Hour of the day
- **Temperature(°C): (Float)** -Temperature per hour in Celsius
- **Humidity(%): (integer)** - Humidity in the air in %
- **Wind speed (m/s): (Float)** - Speed of the wind in m/s
- **Visibility (m): (integer)** - Visibility in m
- **Dew point temperature(°C): (Float)** - Temperature at the beginning of the day in Celsius

contd…

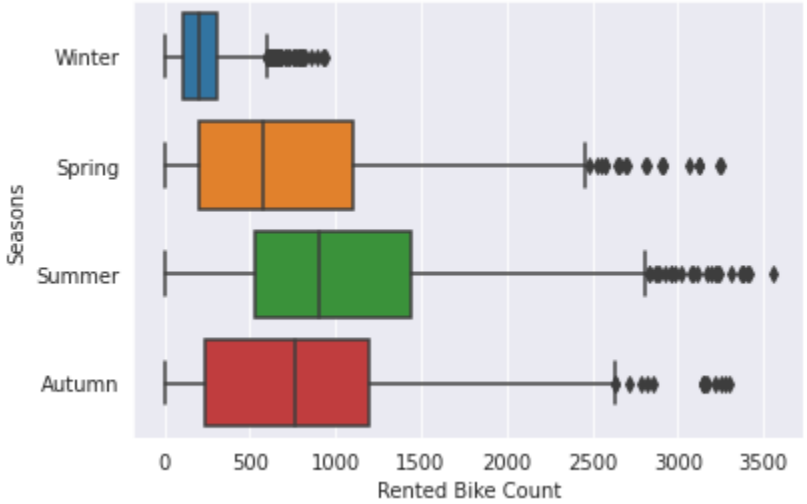
- **Solar Radiation (MJ/m2): (Float)** - Incoming radiation from the Sun
- **Rainfall(mm): (Float)** - Amount of rainfall in mm
- **Snowfall (cm): (Float)** - Amount of snowfall in cm
- **Seasons: (string)** - Season of the year - Winter, Spring, Summer, Autumn
- **Holiday: (string)** - If it is holiday or No holiday
- **Functioning Day(string):** If it is a Functioning Day

Rented Bike count is our Target Variable which we will be predicting using Machine Learning model.

BoxPlot Inference



- **The number of bikes rented is more on working days as compared to on holidays.**

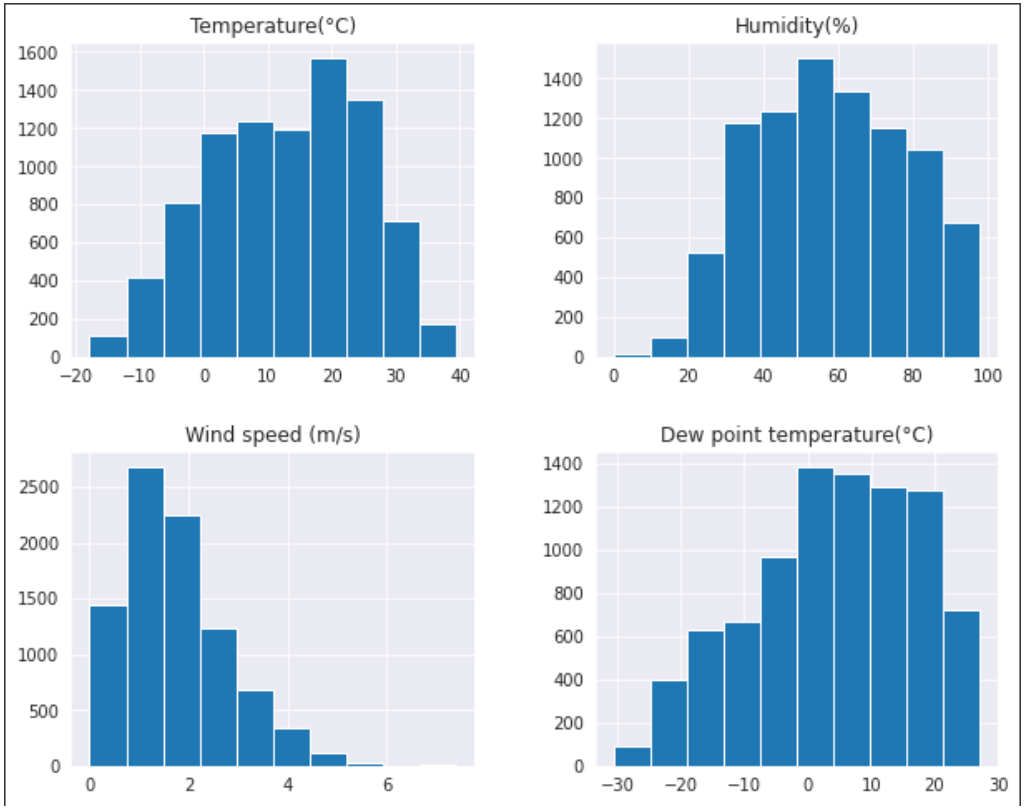


- **The number of Rented Bikes is more in Summer Season followed by Autumn & Spring and Winter being least of all**

Distribution of Features

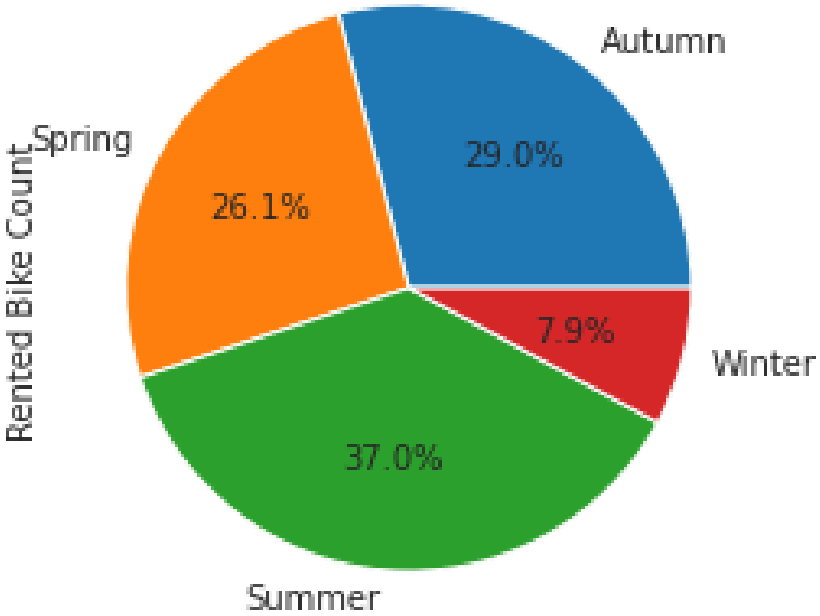
In the plotted histograms in this slide, we can actually look at how our features are distributed in the given dataset.

- Distribution of temperature is approximately normally distributed.
- Distribution of Humidity and Dew point temperature are slightly negatively skewed.
- Distribution of Wind Speed is positively skewed.



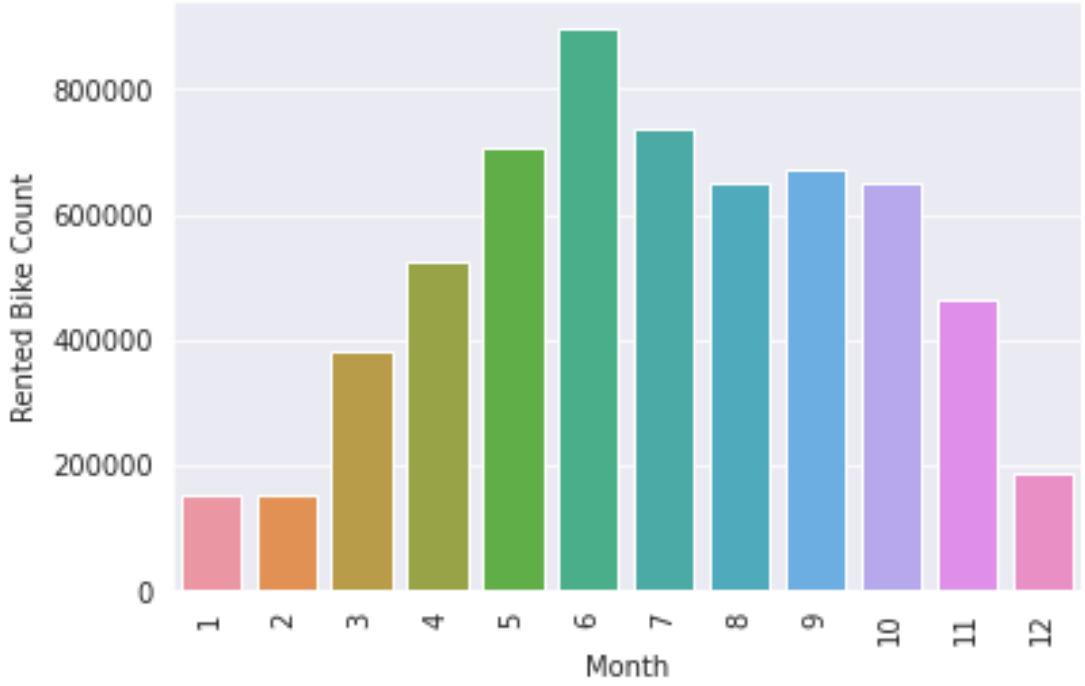
Season wise demand

From the produced pie chart, we can infer that the demand of bikes is huge during summer and it’s the other way during winter.



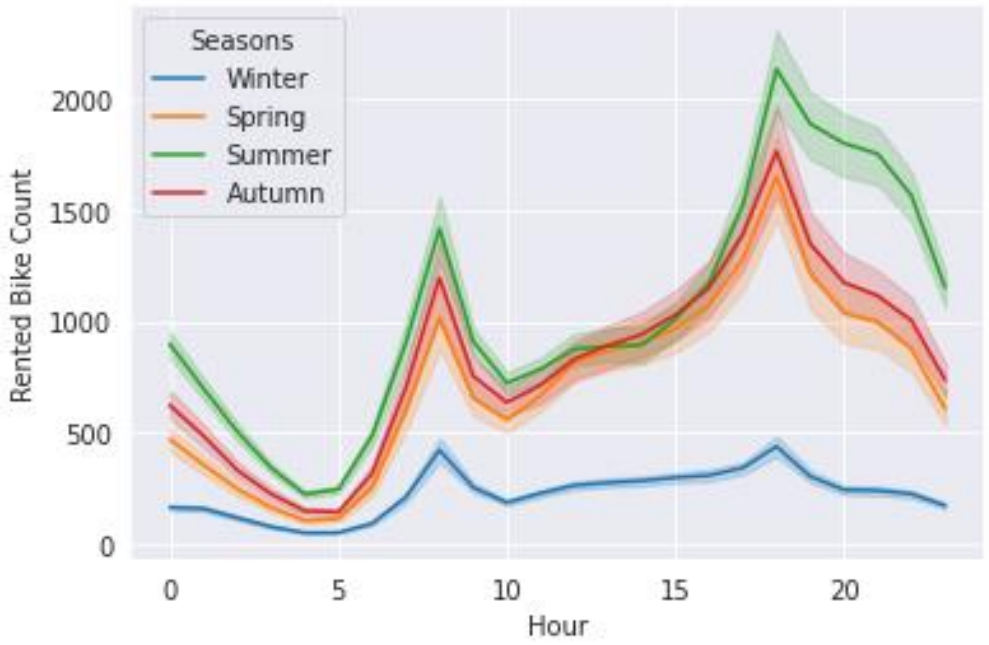
Month Wise demand

Here we can see that demand for bikes is huge in the months of May, June and July and there's a drop in the months of December, January and February.

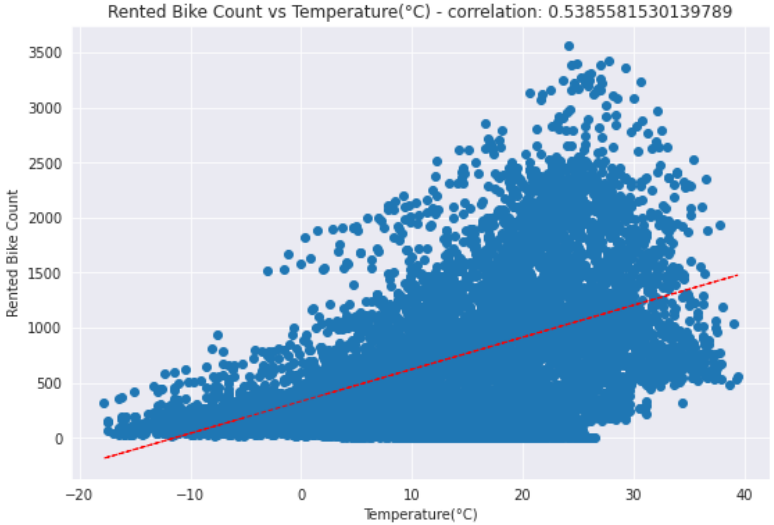


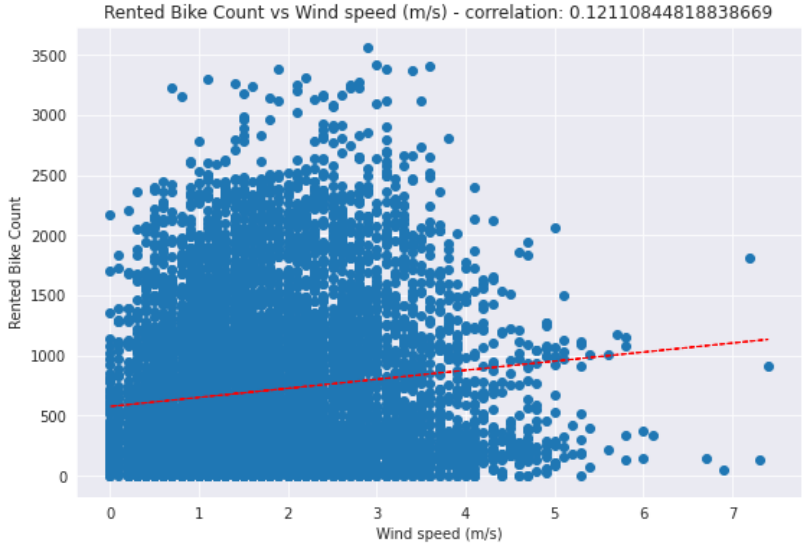
Hour wise demand

Plotting the line graph which includes both hour and the season, we can clearly see how it actually affects the demand of rental bikes and it clearly shows the demand is high during the morning and evening hours.



Scatter plots for few features

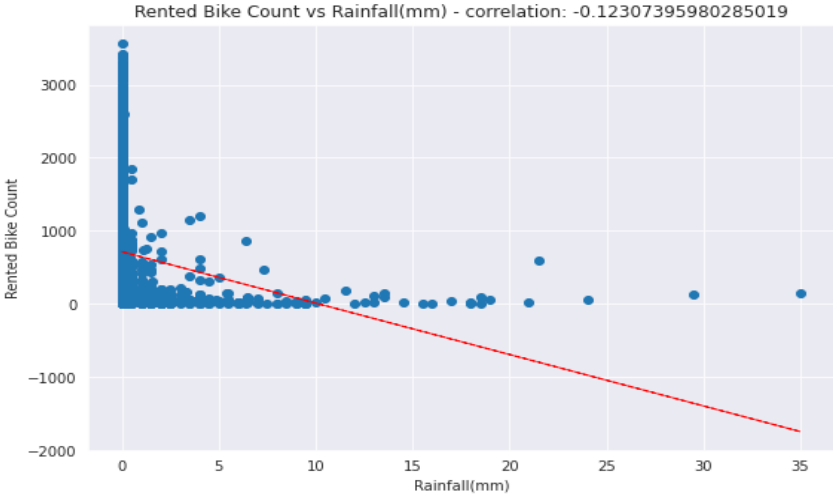


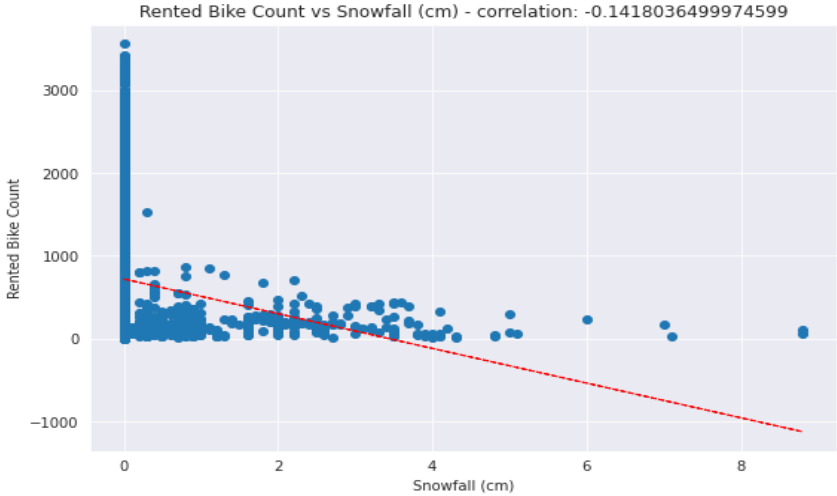


In the first plot of Temperature vs Bike count, we can infer that the Rental Bike Count is spread in form of a cloud from -20 to 40 degree celsius and is denser around the region of 20 to 30 degree celsius.

In the second plot of Wind speed vs Bike count, the data points form a prominent cloud around the Wind Speed lying between 0-5 m/s. However, the distribution starts to fade into secluded clusters till breeze of 6m/s. Post that speed, the data seems to be isolated without any effect from the available data bunch. These points would be considered as potential outliers.

Contd..





In the first plot of Rainfall vs Bike count, we know that as commuting through bike is in open places, Rainfall, almost, will have an inverse relation with the Rental count. The Scatter Plot above also suggests that a significant number of counts lies along the dates when Rainfall was equal to 0 mm. The scattered clusters above 10 mm to 35 mm show a minimal rise of Rental Bikes from zero. The single data point at 35mm remains an exception and suggests orders for recreational purposes or any other relevant cause. And the same goes with snowfall as well.

Correlation

In this heat map we can see that the temperature and the dew point temperature are correlated.

Linear Regression

We have applied Linear Regression Model for given data set. We can see that RMSE as 444 and Adjusted R square being 0.52 which says the accuracy is low. Hence this model is not the best fit for given problem statement

```
MAE -: 328.42337356142406
MSE -: 197729.14962130532
RMSE -: 444.6674595934644
R-Squared accuracy -: 0.5325325441531739
Adj R^2 -: 0.5243790420163106
```

Decision Tree Regressor

We have applied Decision Tree Model for given dataset. We can see that RMSE as 259 which is less than that of Linear Regression and Adjusted R square being 0.83 which is closer to 1 as compared to linear regression. Hence this model is moderate fit for given problem statement.

```
MAE -: 156.21345726508397
MSE -: 67230.66993157589
RMSE -: 259.2887771030129
R-Squared accuracy -: 0.8410545420946618
Adj R^2 -: 0.8382822375963128
```

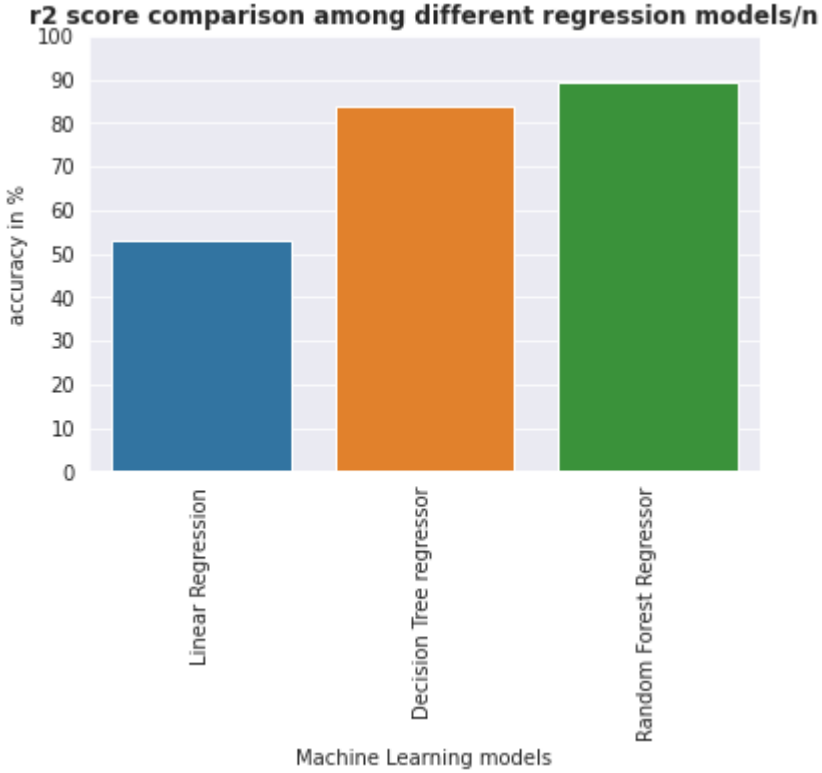
Random Forest Regressor

We have applied Random Forest Regressor for given dataset.We can see that RMSE is 212 which is lesser than that of Decision Tree and Adjusted R square being 0.89 which is closer to 1. Hence this model is good fit for given problem statement

```
MAE -: 126.92366438356163
MSE -: 45170.68192066211
RMSE -: 212.533954747617
R-Squared accuracy -: 0.8932083418314418
Adj R^2 -: 0.8913456966308274
```

Models Comparison

After a detailed comparison of how our models are working, we can see that Random forest is producing better results when compared to the other ones for the given dataset.



Conclusion

- The number of bikes rented is more on working days as compared to on holidays.
- The number of Rented Bikes is more in Summer Season followed by Autumn & Spring and Winter being least of all
- The Demand for bikes is huge in the months of May, June and July
- The demand of bikes is huge during summer and it’s the other way during winter.
- The demand is high during the morning and evening hours.

Thank You!