Capstone Project

Bike Sharing Demand Prediction

Akash Salmuthe

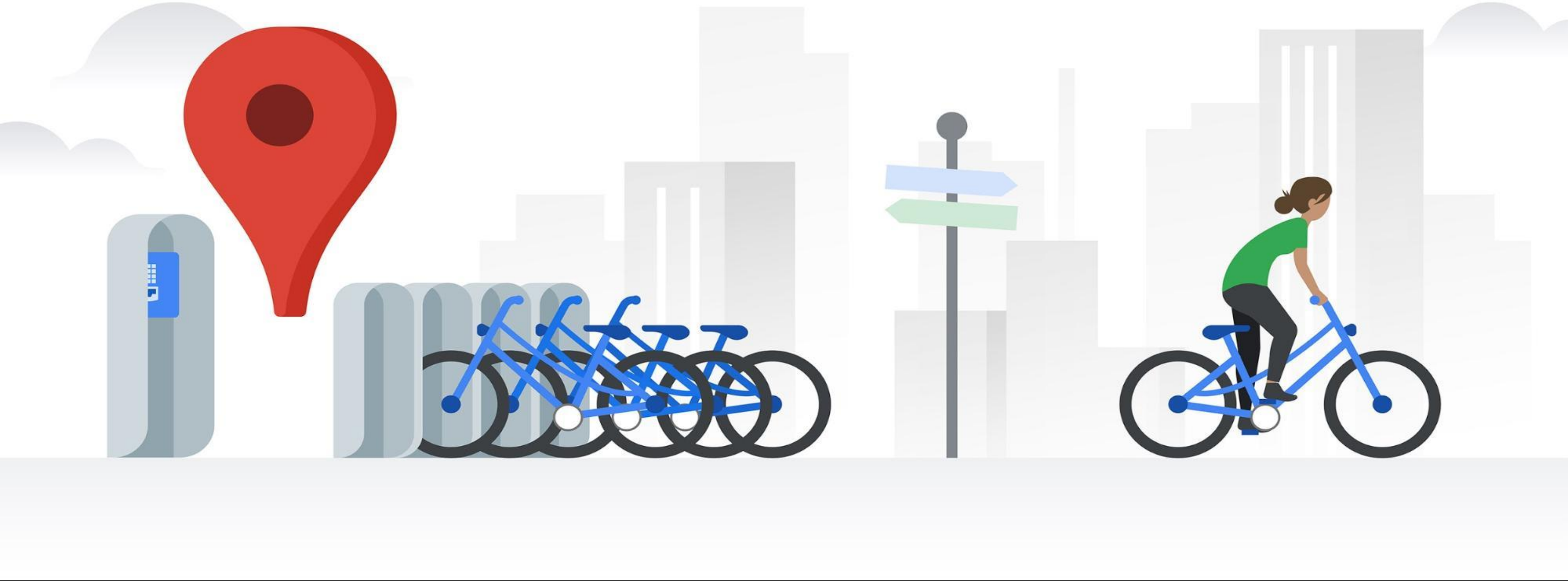
Content

- Problem Description
- Concept - “What is Bike Sharing?”
- Data Summary
- Data Description
- Exploratory Data Analysis
- Supervised Learning Models
- Model Validation & Selection
- Feature Importance
- Challenges
- Conclusion

Problem Description

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.

What is Bike Sharing?



What is Bike Sharing?

- 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. Docks are special bike racks that lock the bike, and only release it by computer control. The user enters payment information, and the computer unlocks a bike. The user returns the bike by placing it in the dock, which locks it in place.

Data Summary

- This Dataset contains 8760 lines and 14 columns.

- Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.

- One Datetime features ‘Date’.

- We have some numerical type variables such as temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall which tells the environment conditions at that particular hour of the day.

Data Description

- 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 - NoFunc(Non Functional Hrs),Fun(Functional Hrs)

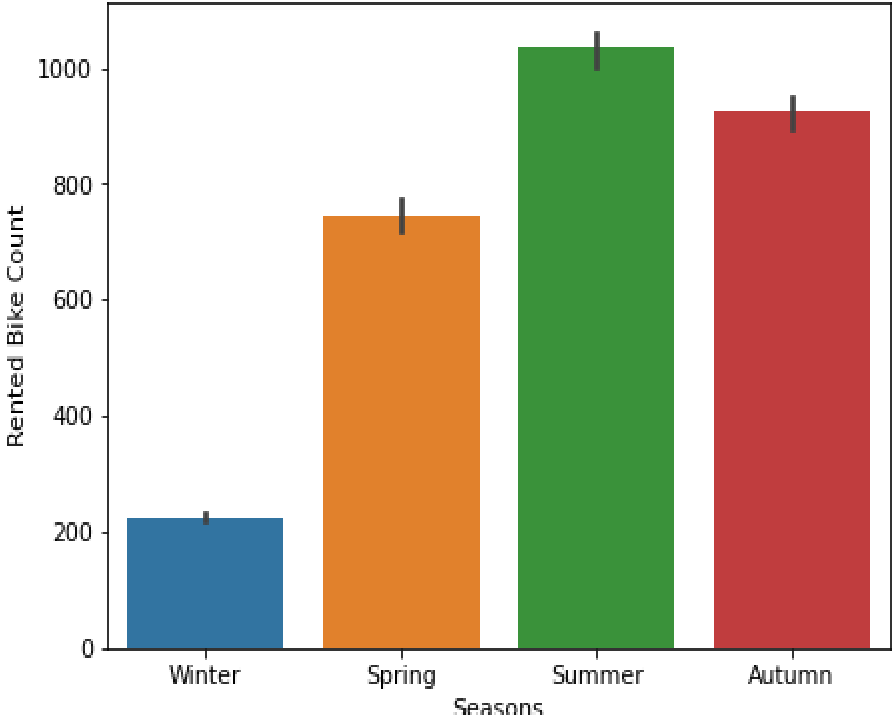
Insights From Dataset

- There are No Missing Values, Duplicate values and No null values present
- We have 'rented bike count' variable which we need to predict for new observations
- The dataset shows rental data for one year (1 December 2017 to 31 November(2018)(365 days)
- We change the name of some features for our convenience

Exploratory Data Analysis

Rented Bike vs Season

- **From the data we can conclude that Summer has highest demand followed by Spring and Autumn Season**

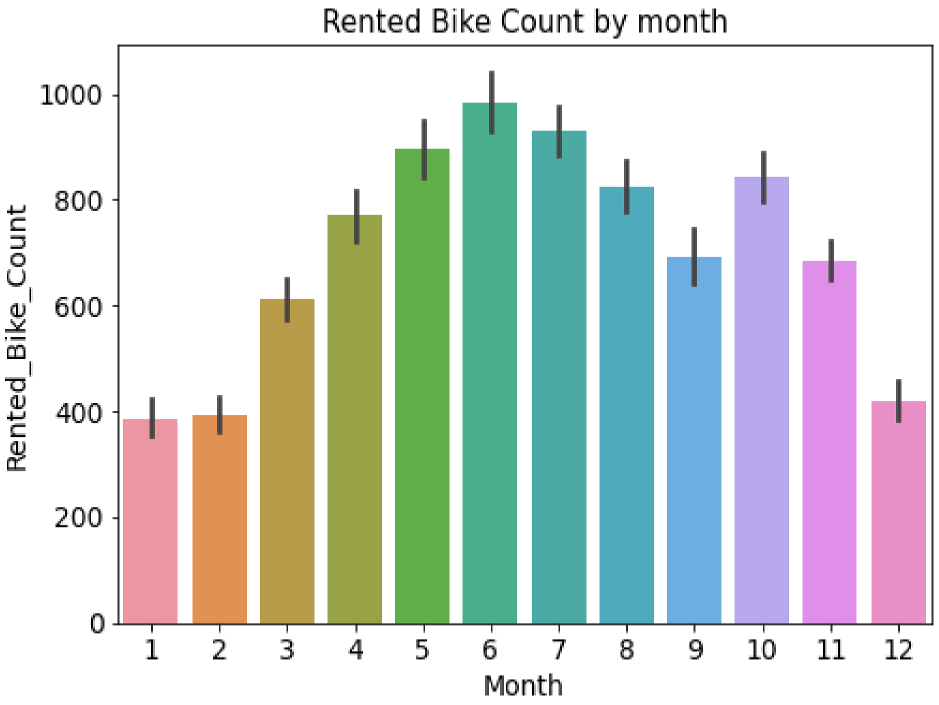


Rented Bike vs Month

- We can see that there is less demand in the month of December, January and February (Winter Season).

- Bike demand is increases from March and the June is at peak level through out the year.

- Demand get started to decrease from July where January is at lowest position in terms of bike demand

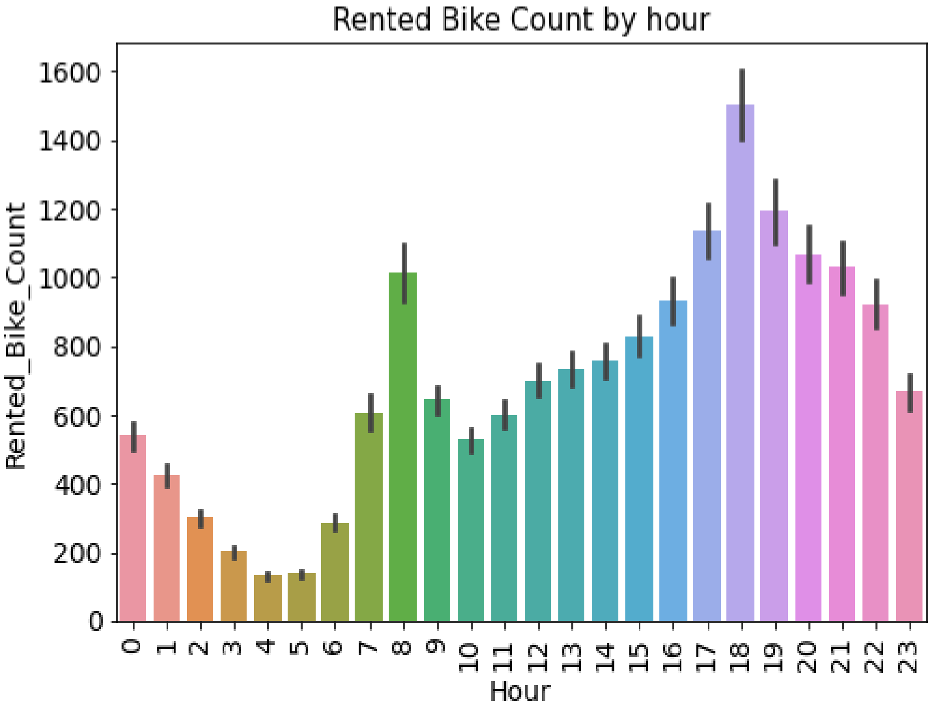


Rented Bike vs Hour

- There is surge of high demand in the morning 8 am and in the evening at 6 pm

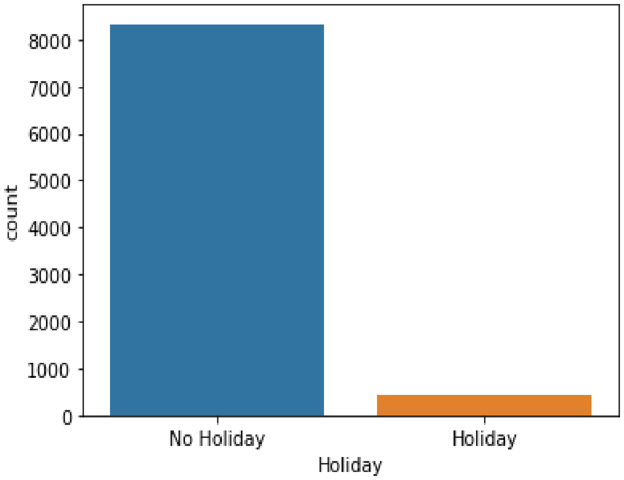
- It is because of work time stared in morning and ends at evening

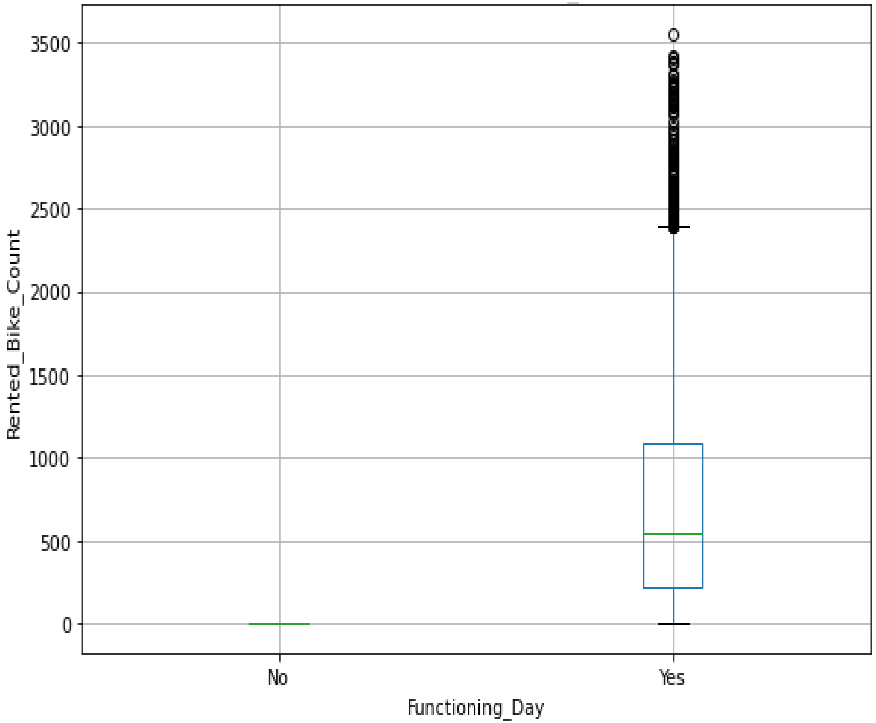
- Mostly we can see this on working day



Rented Bike vs Functioning Day

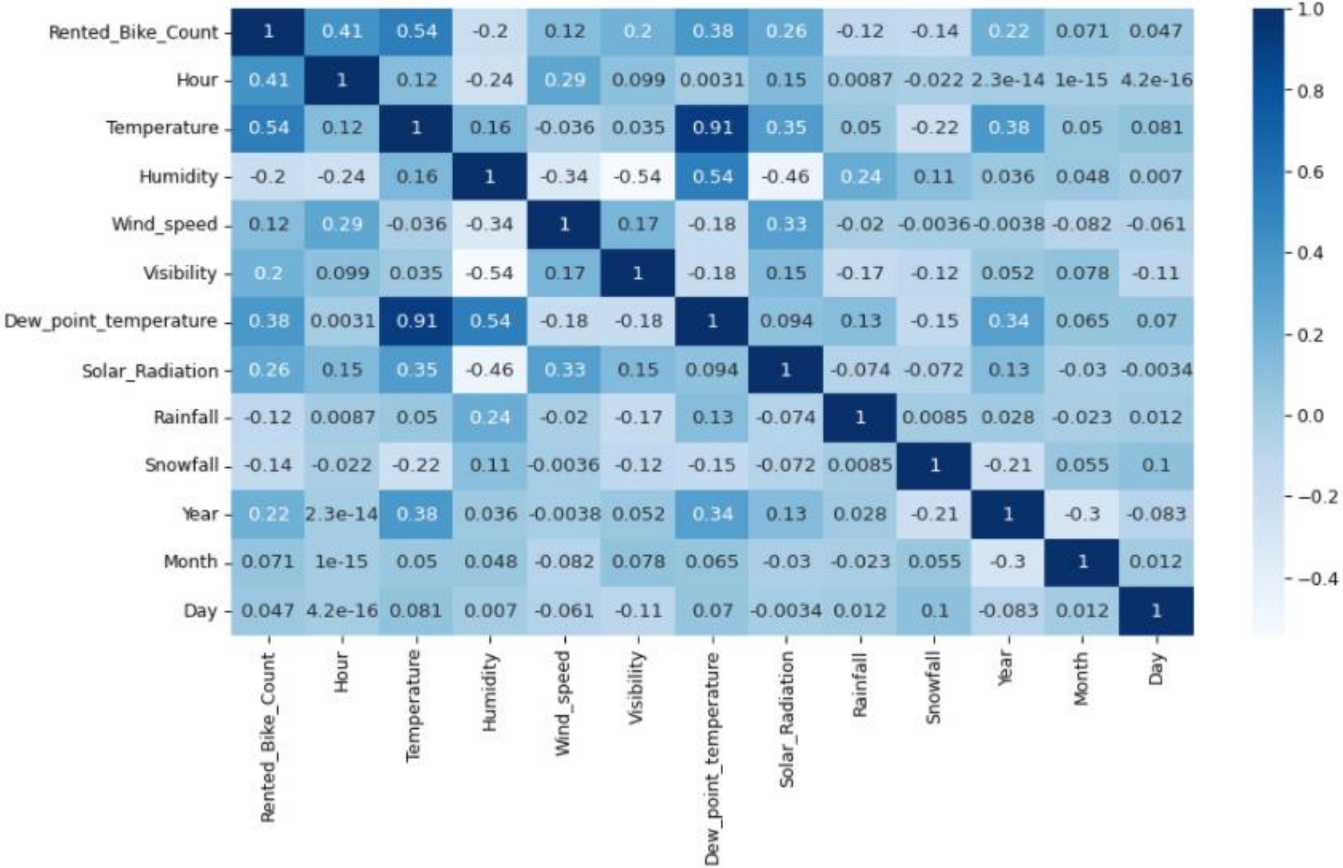
- We can clearly see that there is almost no demand on holiday for bikes





Heatmap

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff964372250>
```

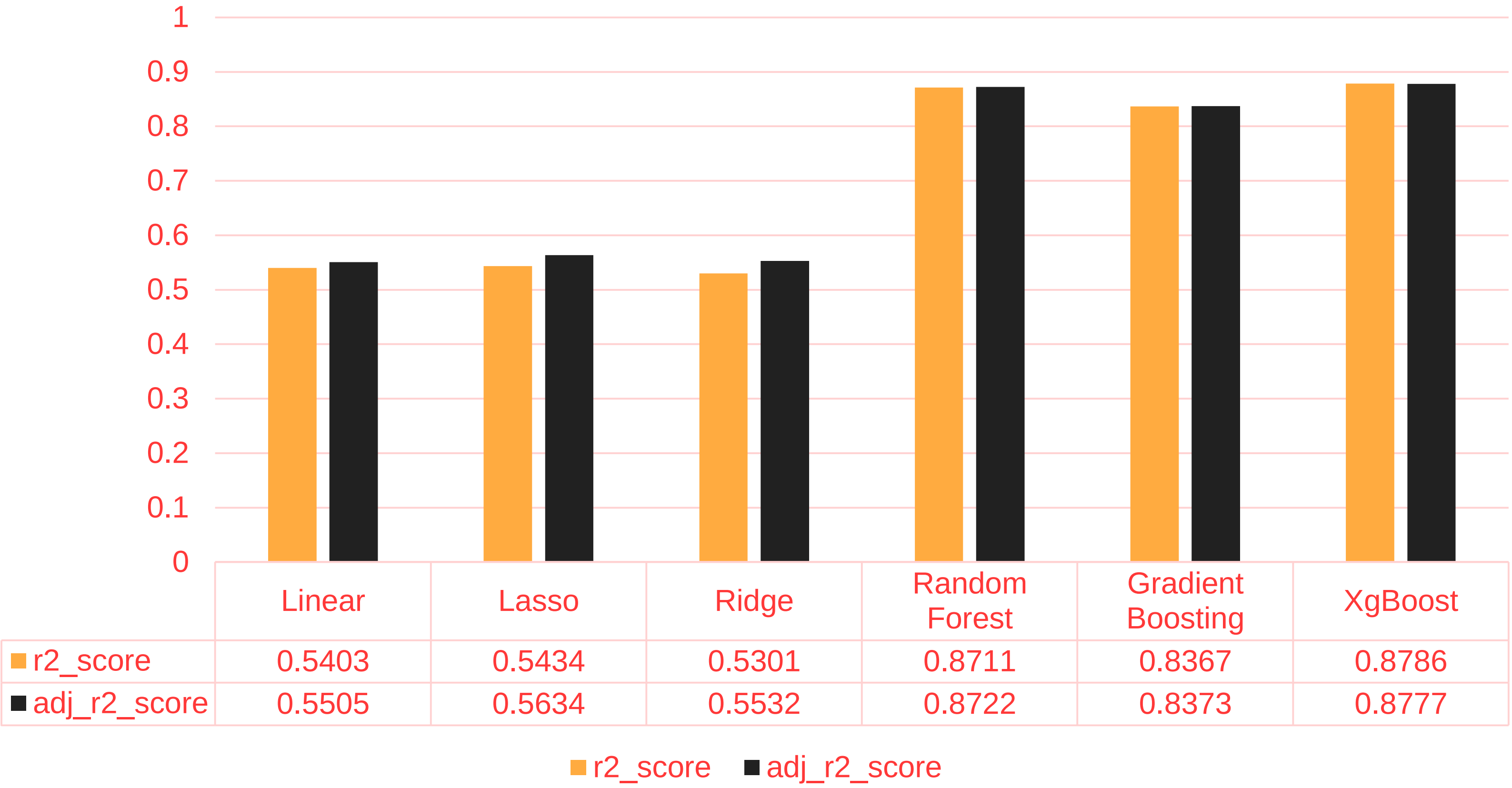


	Rented_Bike_Count	Hour	Temperature	Humidity	Wind_speed	Visibility	Dew_point_temperature	Solar_Radiation	Rainfall	Snowfall	Year	Month	Day
Rented_Bike_Count	1	0.41	0.54	-0.2	0.12	0.2	0.38	0.26	-0.12	-0.14	0.22	0.071	0.047
Hour	0.41	1	0.12	-0.24	0.29	0.099	0.0031	0.15	0.0087	-0.022	2.3e-14	1e-15	4.2e-16
Temperature	0.54	0.12	1	0.16	-0.036	0.035	0.91	0.35	0.05	-0.22	0.38	0.05	0.081
Humidity	-0.2	-0.24	0.16	1	-0.34	-0.54	0.54	-0.46	0.24	0.11	0.036	0.048	0.007
Wind_speed	0.12	0.29	-0.036	-0.34	1	0.17	-0.18	0.33	-0.02	-0.0036	-0.0038	-0.082	-0.061
Visibility	0.2	0.099	0.035	-0.54	0.17	1	-0.18	0.15	-0.17	-0.12	0.052	0.078	-0.11
Dew_point_temperature	0.38	0.0031	0.91	0.54	-0.18	-0.18	1	0.094	0.13	-0.15	0.34	0.065	0.07
Solar_Radiation	0.26	0.15	0.35	-0.46	0.33	0.15	0.094	1	-0.074	-0.072	0.13	-0.03	-0.0034
Rainfall	-0.12	0.0087	0.05	0.24	-0.02	-0.17	0.13	-0.074	1	0.0085	0.028	-0.023	0.012
Snowfall	-0.14	-0.022	-0.22	0.11	-0.0036	-0.12	-0.15	-0.072	0.0085	1	-0.21	0.055	0.1
Year	0.22	2.3e-14	0.38	0.036	-0.0038	0.052	0.34	0.13	0.028	-0.21	1	-0.3	-0.083
Month	0.071	1e-15	0.05	0.048	-0.082	0.078	0.065	-0.03	-0.023	0.055	-0.3	1	0.012
Day	0.047	4.2e-16	0.081	0.007	-0.061	-0.11	0.07	-0.0034	0.012	0.1	-0.083	0.012	1

Supervised Learning Models

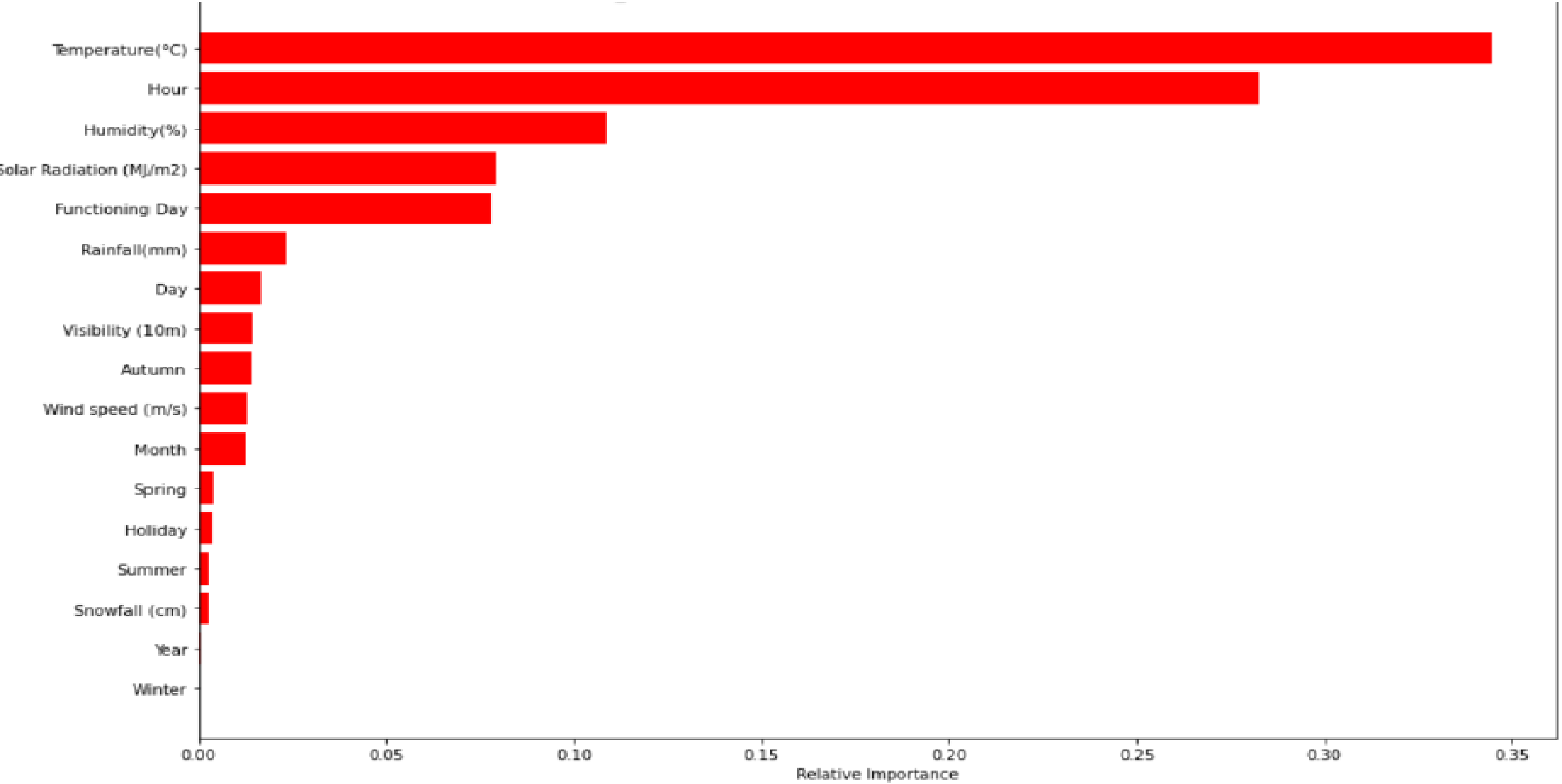
- Linear Regression
- Lasso Regression
- Ridge Regression
- Random Forest
- Gradient Boosting
- XgBoost

Models Performed with r2 and Adj_r2

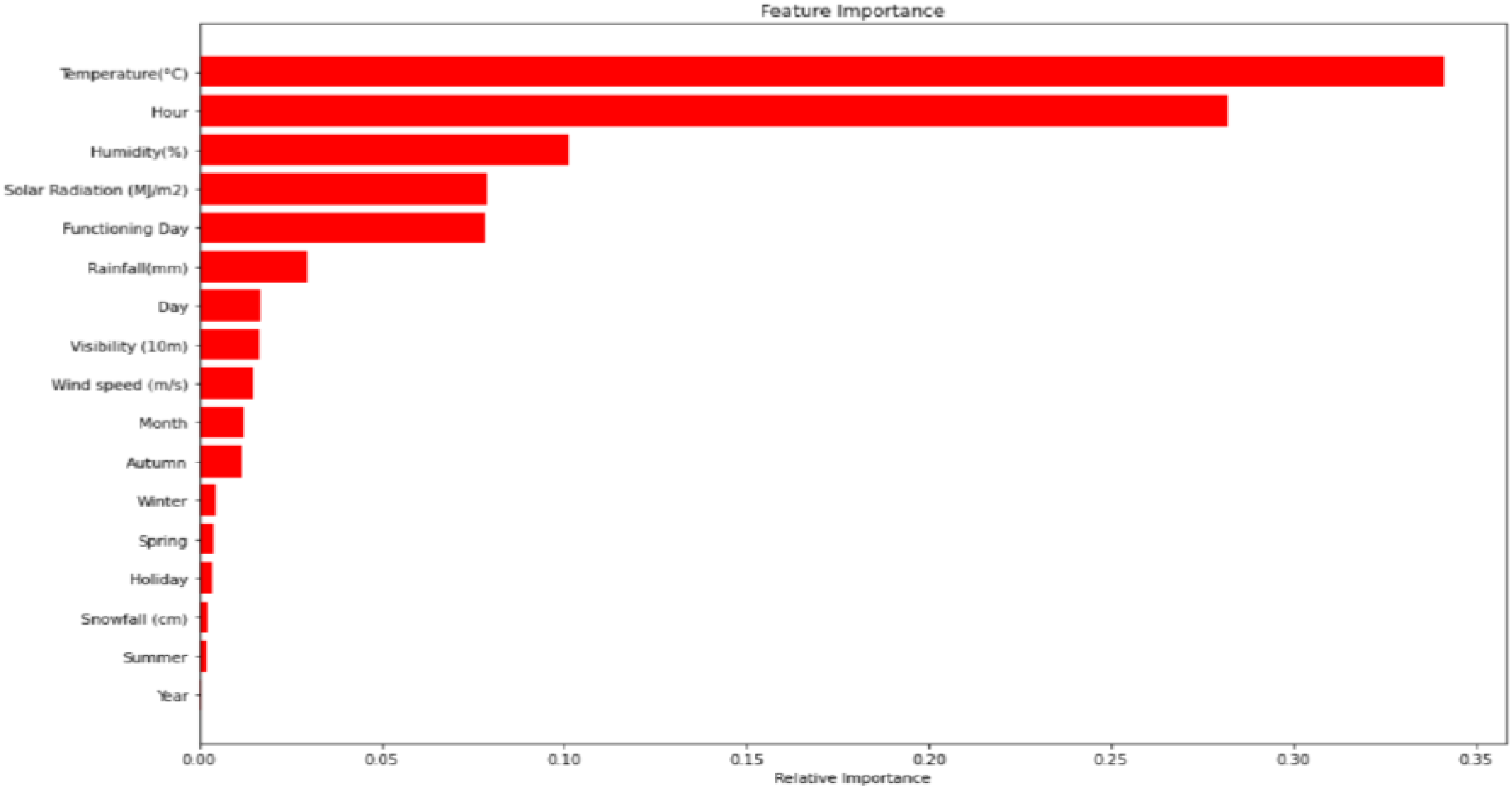


	Linear	Lasso	Ridge	Random Forest	Gradient Boosting	XgBoost
🟧 r2_score	0.5403	0.5434	0.5301	0.8711	0.8367	0.8786
⬛ adj_r2_score	0.5505	0.5634	0.5532	0.8722	0.8373	0.8777

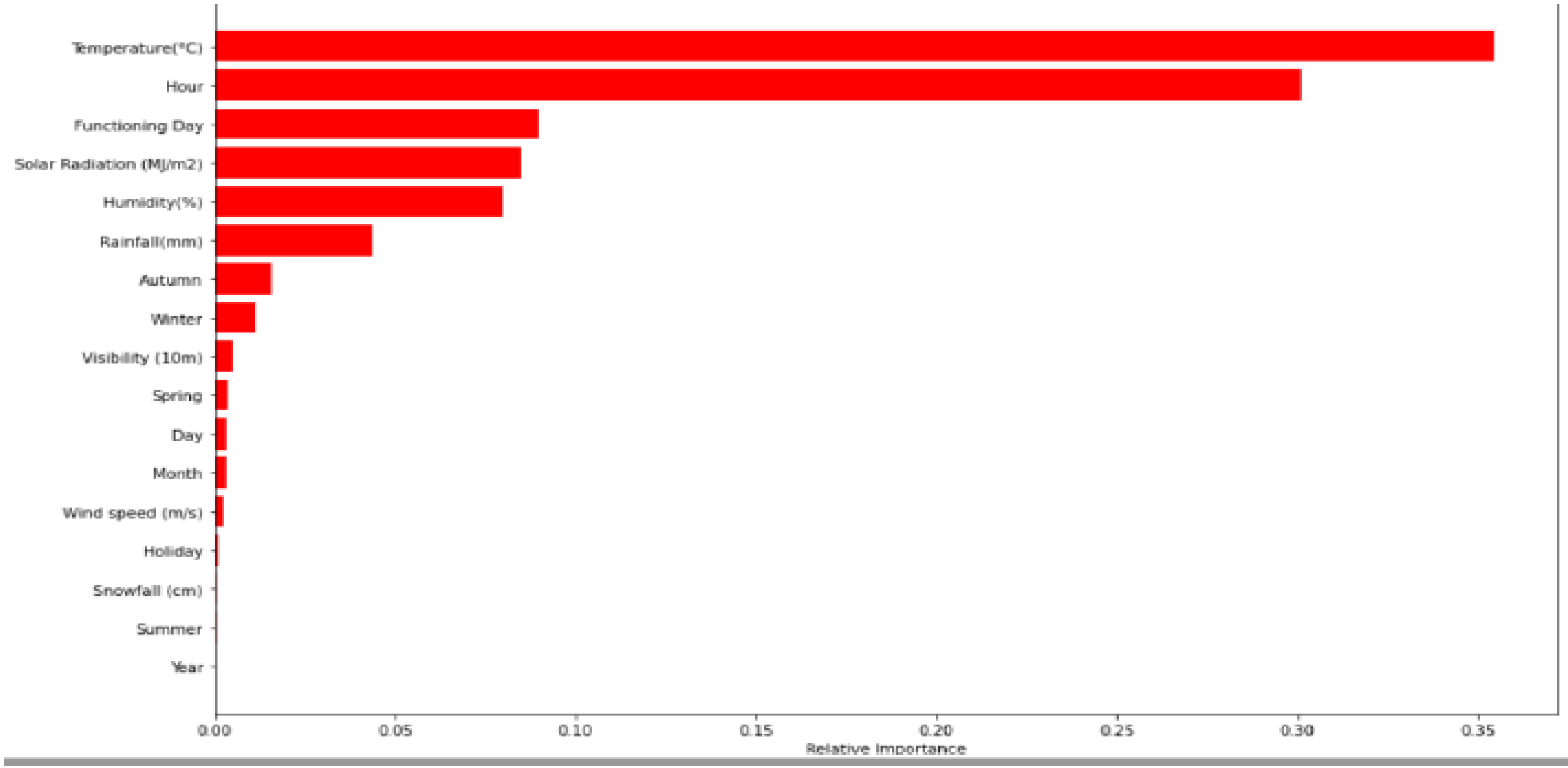
Feature Importance For Decision Tree



Feature Importance For Gradient Boosting



Feature Importance For Random Forest



Model Validation and selection

1. As we can see that Linear, Lasso and Ridge performs almost same but it's not good enough for good model

2. Random forest, Decision Tree and Xgboost performs well in terms of r2 and adjusted_ r2 score.

3. For our bike rental solution we can use either Random Forest or Decision Tree or XgBoost

Challenges

- This is the large data set to handle.
- While selecting feature need to identify feature which has most impact on bike rental
- It’s south Korean City (Seoul) data set we must think according with their perspective and lifestyle

Conclusion

- On holiday there is very less demand in rental bikes
- People usually preferred more rental bikes in Morning and evening
- Hour and functioning Day is most dominant feature in all the data set
- Bike rental count is mostly correlated with the time of the day
- XGBoost model has less root mean squared error and mean absolute error, ending with the accuracy of 87%