Capstone Project-2

Project Title

Seoul Bike Sharing Demand Prediction

BY : AYUSH

AlmaBetter

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. 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.

CONTENT

- Data Pipeline
- Data Description
- Exploratory Data Analysis
- Models performed
- Model Validation & Selection
- Evaluation Matrix of All the models
- Challenges
- Conclusion

DATA PIPELINE

- Exploratory Data Analysis (EDA): In this part we have done some EDA on the features to see the trend.

- Data Processing: In this part we went through each attributes and encoded the categorical features.

- Model Creation: Finally in this part we created the various models. These various models are being analyzed and we tried to study various models so as to get the best performing model for our project.

DATA DESCRIPTION

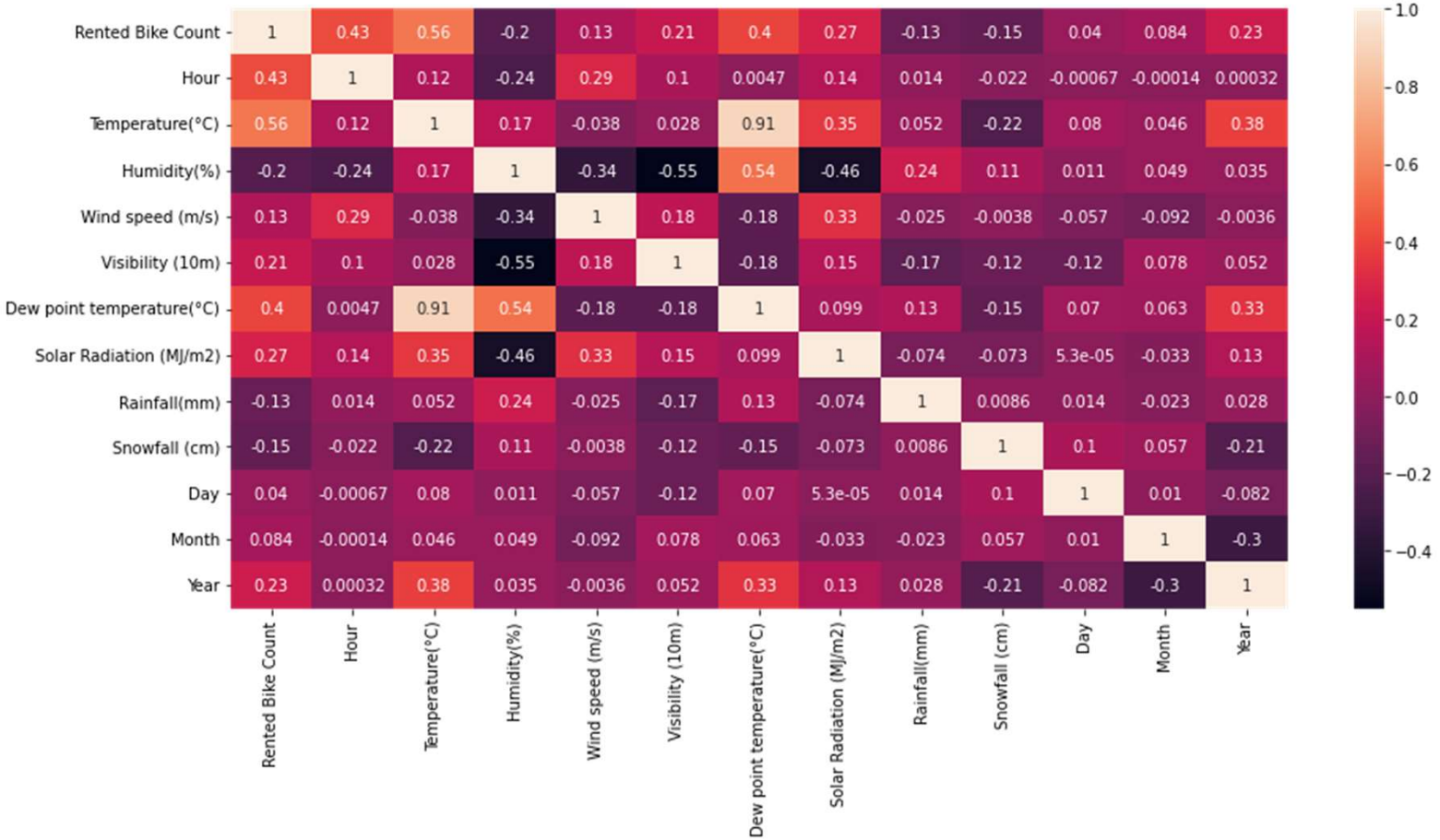
Dependent variable:

- Rented Bike count - Count of bikes rented at each hour

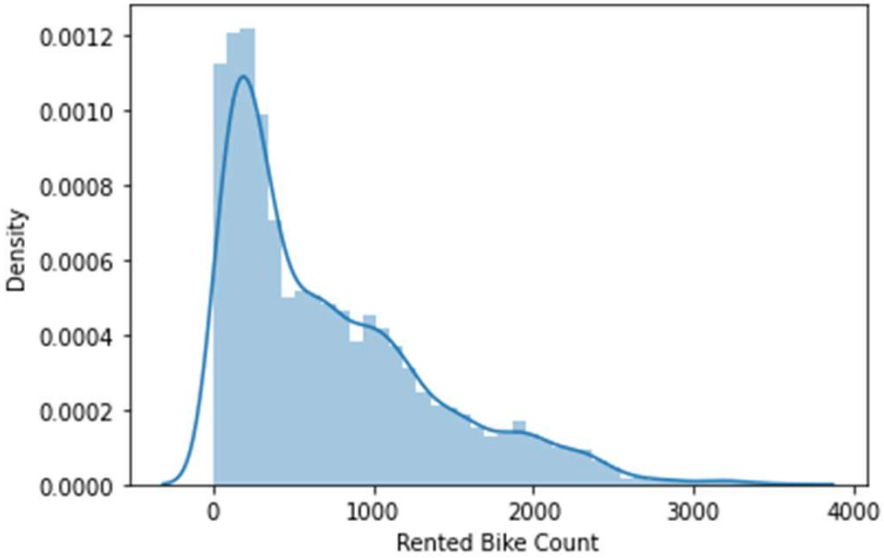
Independent variables:

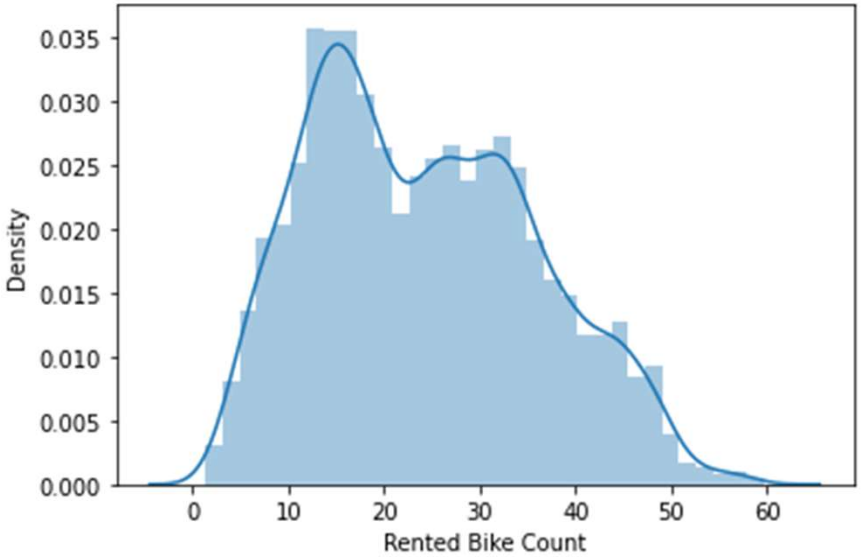
- Date : year-month-day
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10 m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

EDA - FEATURE CORRELATION

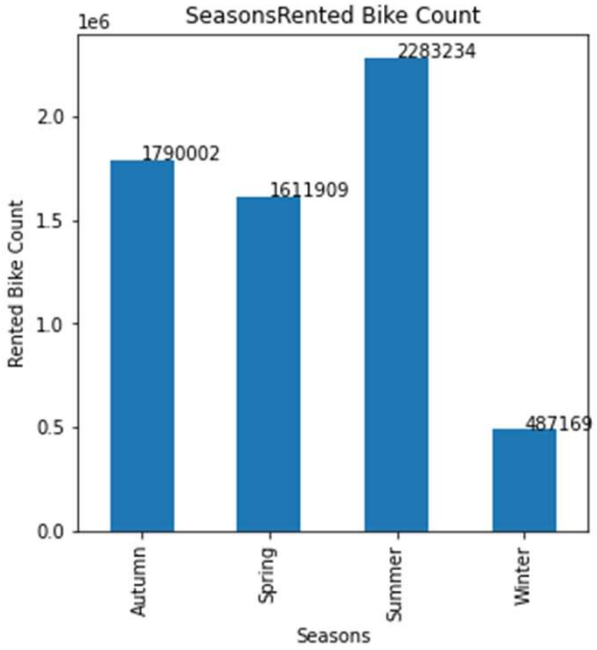


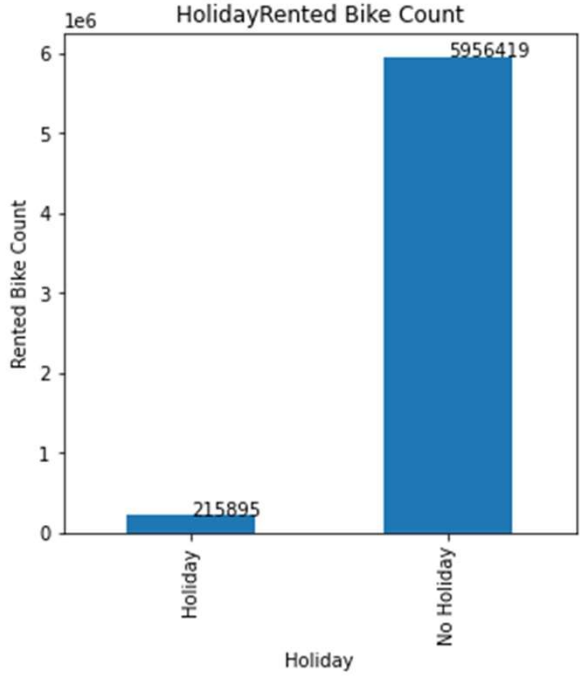
EDA (contd...)


0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Density
0
1000
2000
3000
4000
Rented Bike Count


0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
0
10
20
30
40
50
60
Rented Bike Count

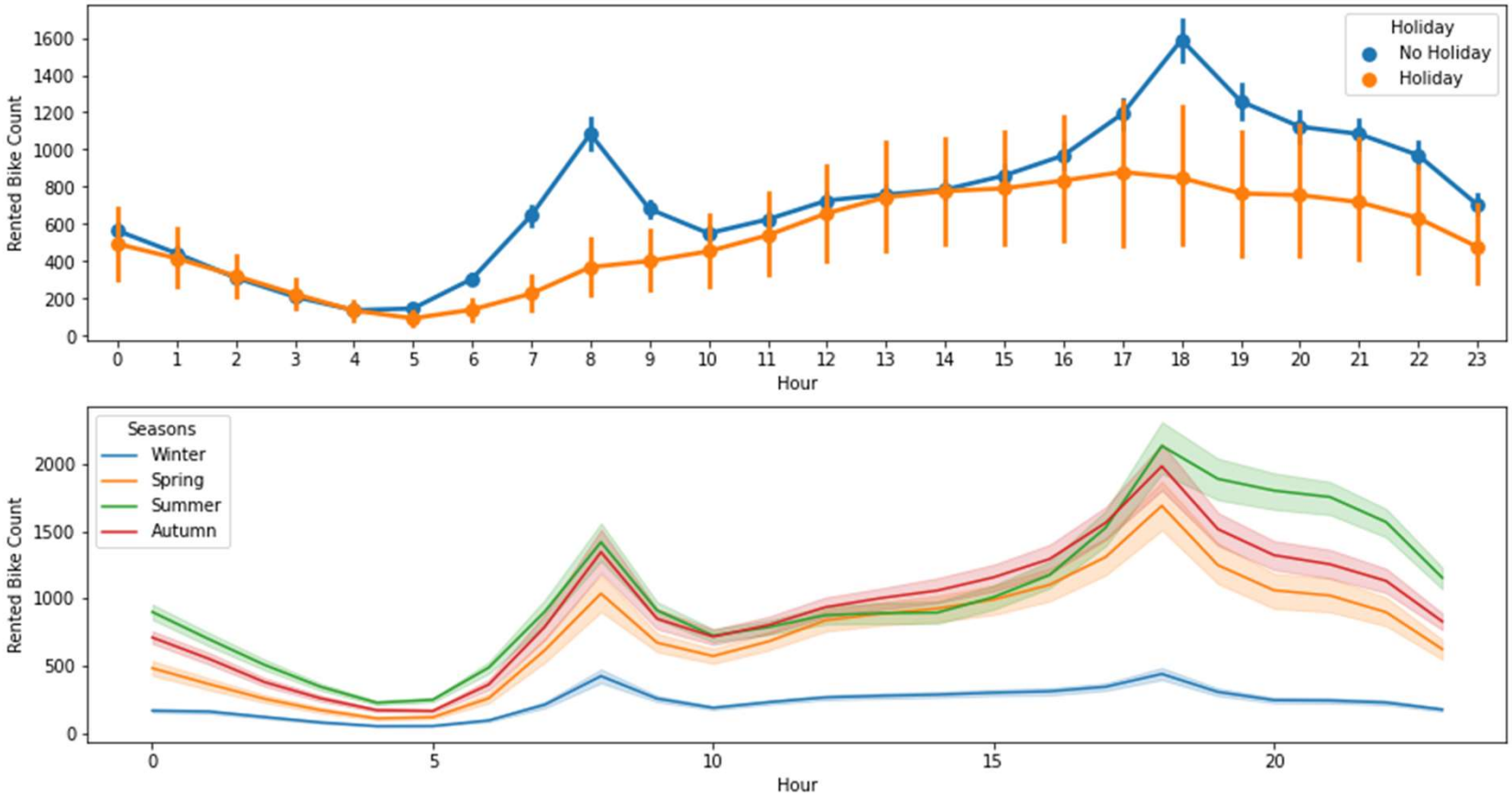
EDA (contd...)





- Less demand on winter seasons
- Slightly Higher demand during Non holiday

EDA (contd...)


Holiday
No Holiday
Holiday
Rented Bike Count
Hour
Seasons
Winter
Spring
Summer
Autumn
Rented Bike Count
Hour

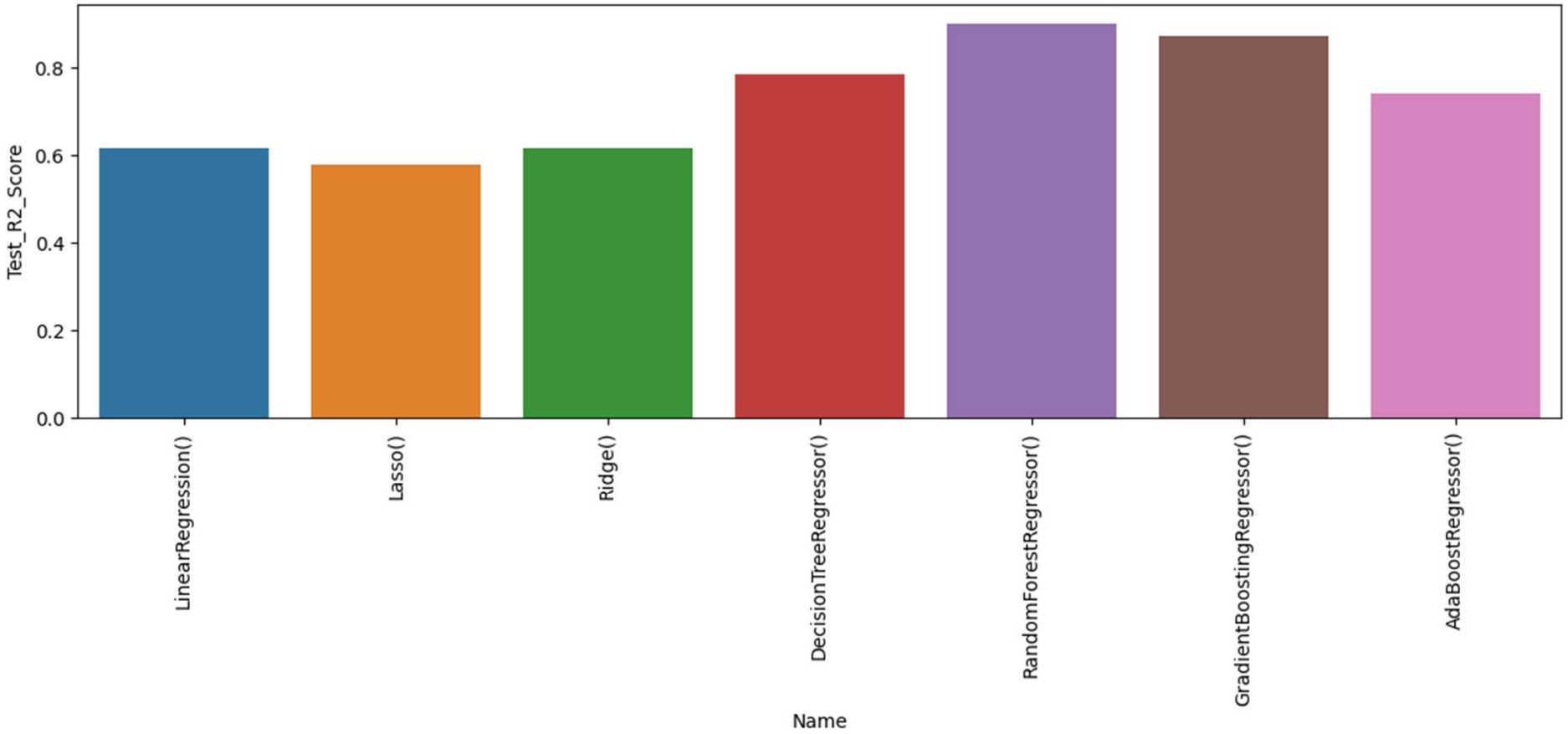
Model's Performed

- Linear Regression with regularizations(ridge and lasso)
- Decision tree
- Random forest
- Gradient Boosting regressor
- AdaBoost

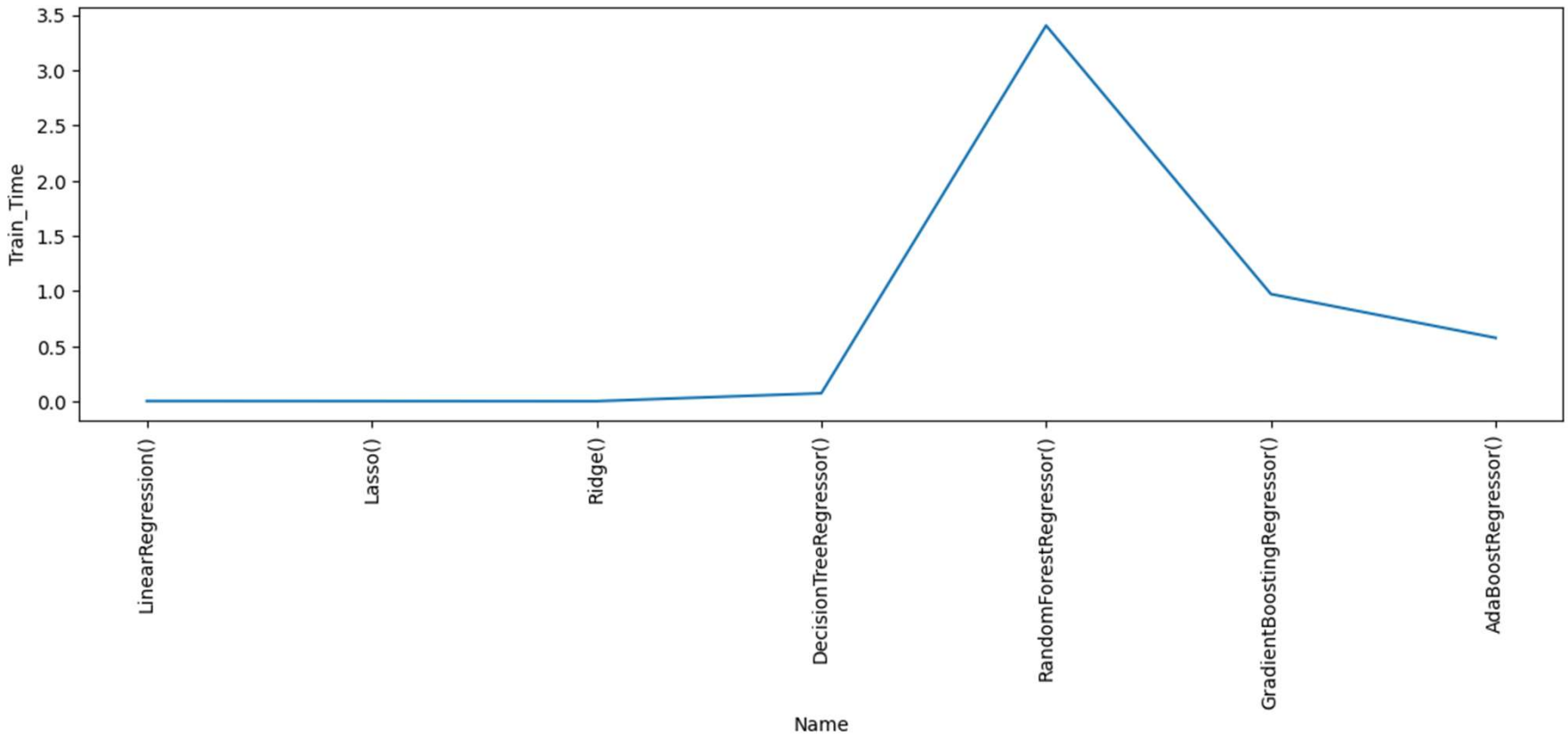
Model’s Evaluation Matrices

	Name	Train_Time	Train_R2_Score	Test_R2_Score	Test_RMSE_Score
0	LinearRegression()	0.003352	0.614822	0.616963	7.499827
1	Lasso()	0.002688	0.577747	0.577491	7.876790
2	Ridge()	0.001956	0.614821	0.616957	7.499891
3	DecisionTreeRegressor()	0.073488	1.000000	0.786056	5.605067
4	RandomForestRegressor()	3.403994	0.984821	0.900992	3.812997
5	GradientBoostingRegressor()	0.972058	0.883240	0.874190	4.298212
6	AdaBoostRegressor()	0.575346	0.745059	0.742617	6.147814

MODEL PERFORMANCE R2 SCORE



MODEL PERFORMANCE ON TRAINING TIME



Model Selection

- **Observation 1:** As seen in the Model Evaluation Matrices table, Linear Regression, Ridge and Lasso is not giving great results.
- **Observation 2:** Random forest & GBR have performed equally good in terms of r2 score.
- **Observation 3:** We are getting the best results from lightGBM and RandomForestRegressor and GBR
- **Observation 4:** RandomForestRegressor takes a lot of training time so we use GBR instead as there is very less difference between there r2 score

CHALLENGES

- A huge amount of data was given to me so first I need to understand the data and its distribution and make it clean to be given to the ML models to make predictions which is the main task
- As the dataset was very big so even a small change may lead to a very different model

conclusion

- It is quite evident from the results that GBM and Random forest regressor is the best model that can be used for the Bike Sharing Demand Prediction since the performance metrics (rmse) shows lower and (r2) show a higher value for the GBM and Random forest regressor models.

- So, we can use either GBM or Random forest regressor model for the above problem but I prefer GBM as it takes less training time with good accuracy and r2 score