Capstone Project II

(SUPERVISED ML – REGRESSION)

Seoul Bike Sharing Demand Prediction

Rajakumaran S
Cohort - Azaadi
Alma Better

~ UNDER THE GUIDANCE OF ALMABETTER TEAM

CONTENT

- PROBLEM STATEMENT
- ROAD MAP
- INTRODUCTION OF PROJECT
- DATA DESCRIPTION
- PRE PROCESSING THE DATA
- EDA
- FEATURE SELECTION
- FITTING VARIOUS MODEL
- MODEL PERFORMANCE COMPARISION
- MODEL VALIDATION
- MODEL EXPLAINABILITY
- CONCLUSION

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.

Road Map

- **EDA**
  - Data distribution of features
  - Deal with multicollinearity
  - Separate dependent and independent features
- **Model building**
  - Data transformation
  - Fitting
  - Prediction
  - Evaluation matrices
- **Model validation**
  - Model selection
  - Feature importance
  - Conclusion

INTRODUCTION

The basic idea of this capstone project is to use the Supervised Machine Learning - Regression to predict the bikes going for rent per hour. We have several seasons, whether conditions, day-wise data for every hours in a day.

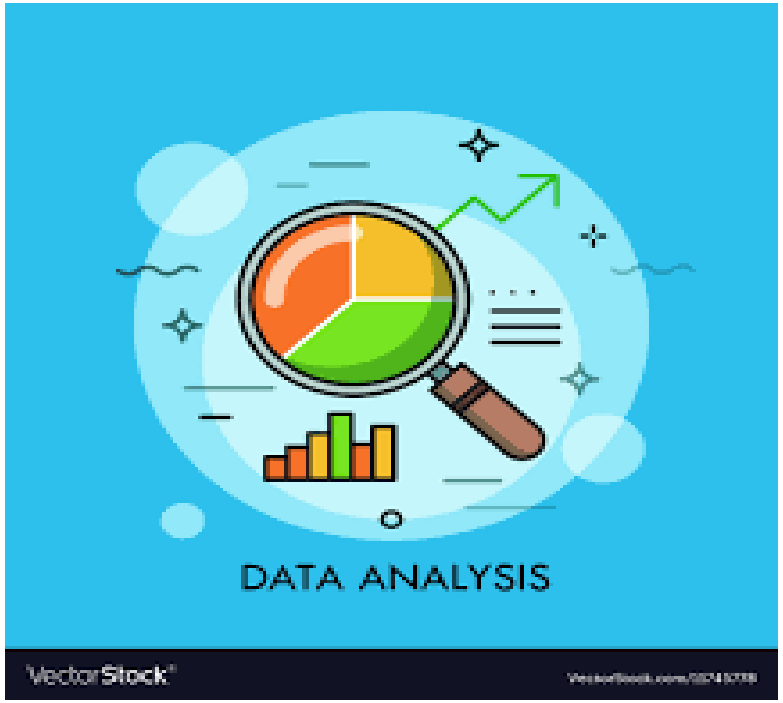
Based on these features we will be predicting our target variable i.e. rented bikes per hour. By using concepts like model validation, we will came to know which features are important and how much they contribute to our target variable.

DATA DESCRIPTION

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

Attribute Information:

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - 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 Hours), Fun(Functional hours)



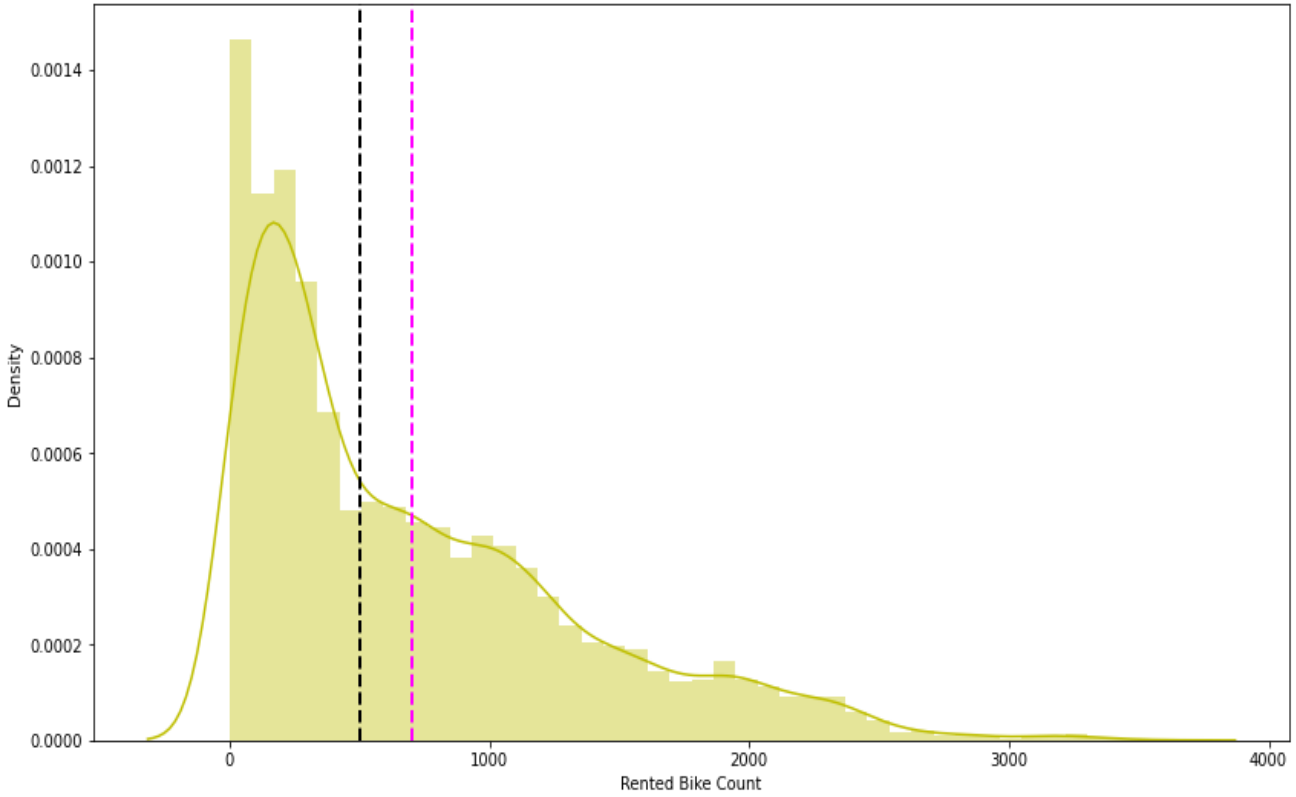
Pre-processing the data:

- In The Dataset having 8760 rows and 14 Columns(Features) including Target Variable Rented Bike Count.
- There is no Null Values in the Dataset.
- There is no Duplicate Values in The Dataset.
- The Data contains a years of 2017 December to 2018 November.

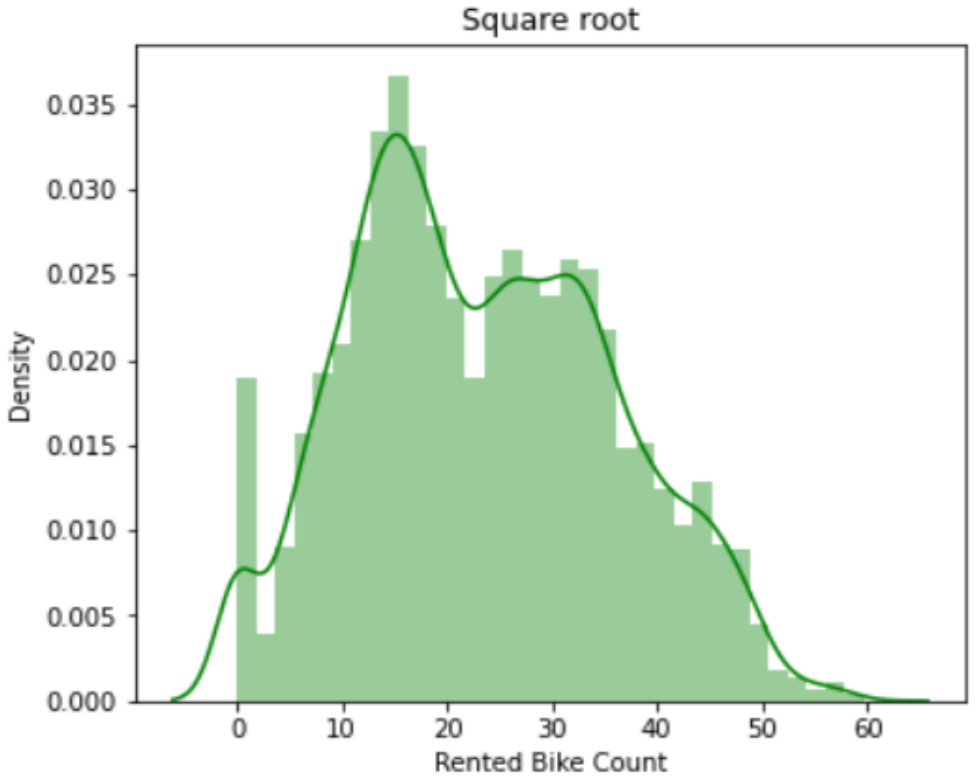
EDA

Data distribution of target variable

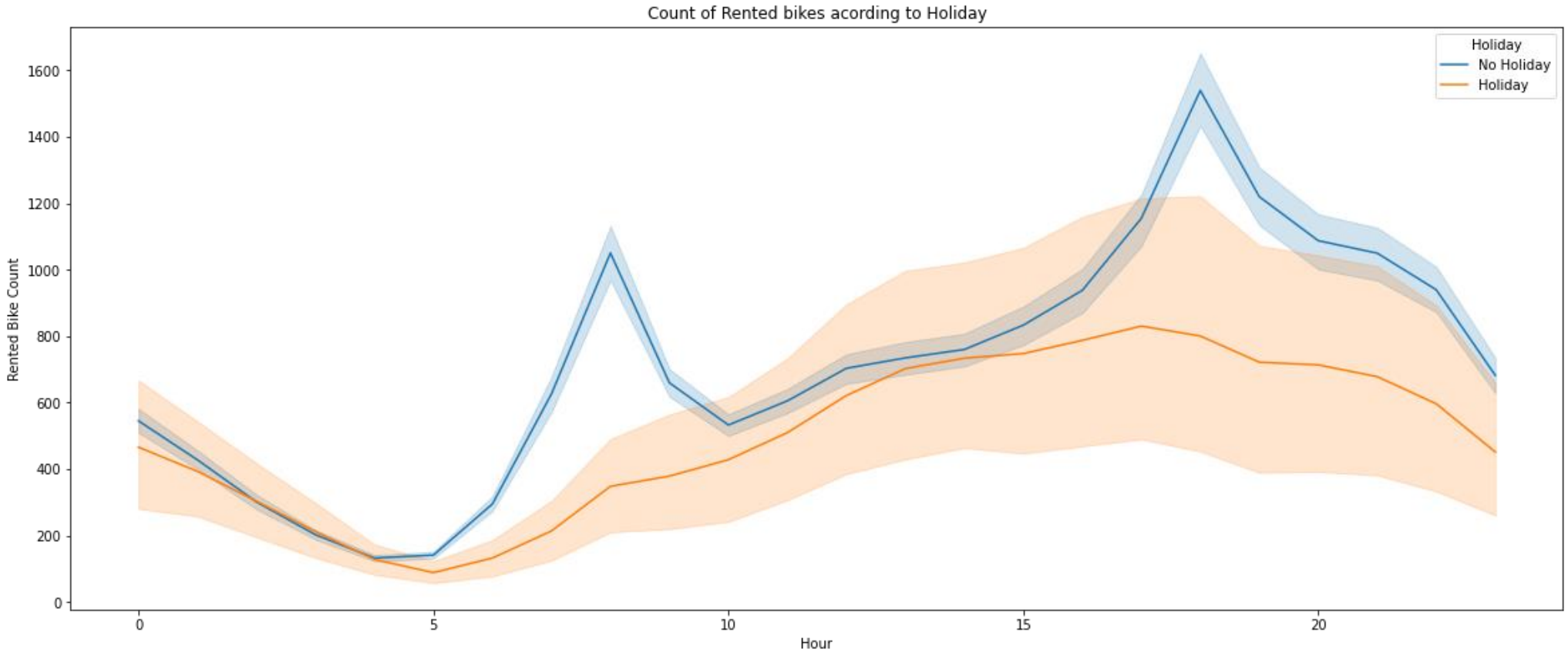
Before transformation



After using sqrt transformation

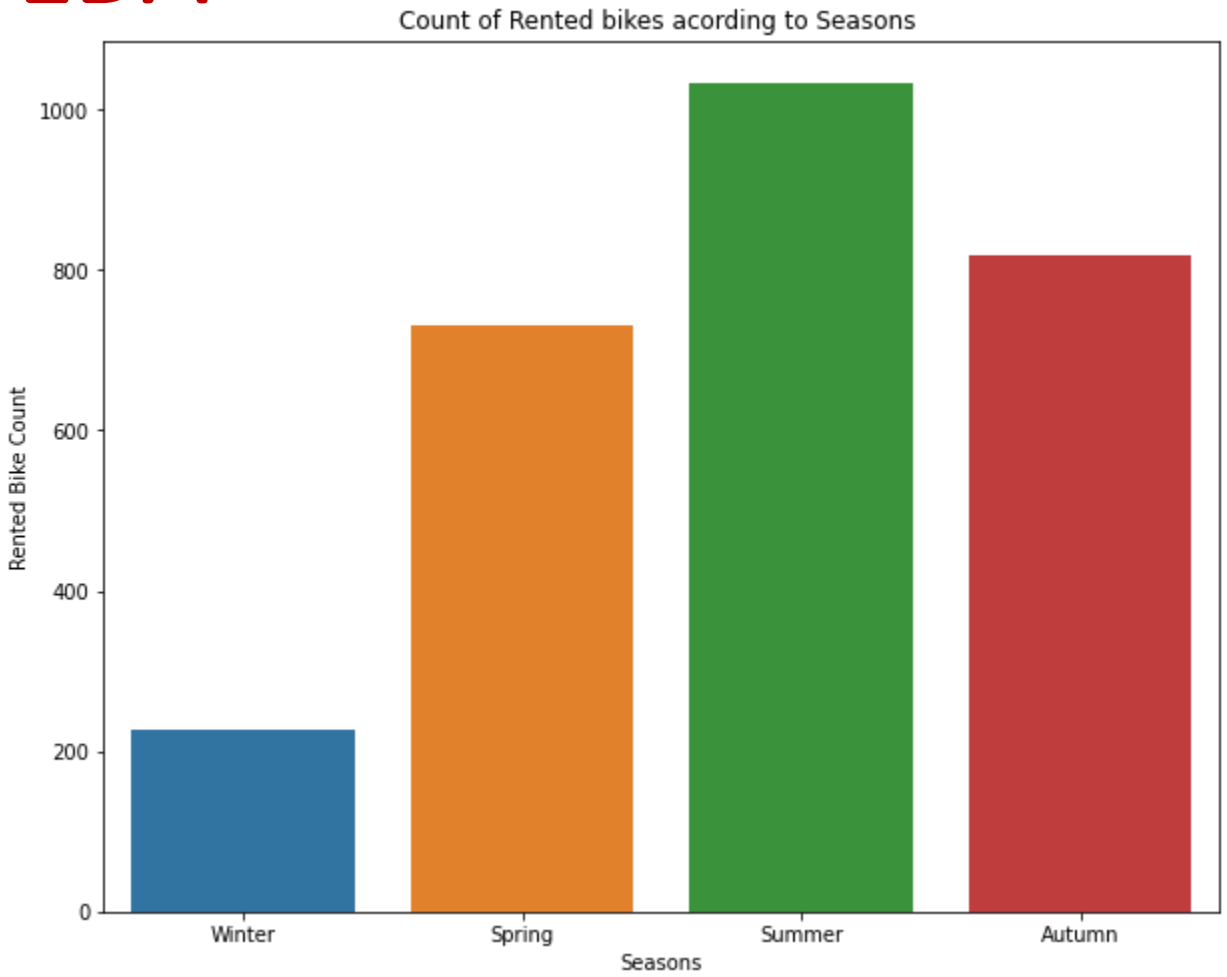


EDA

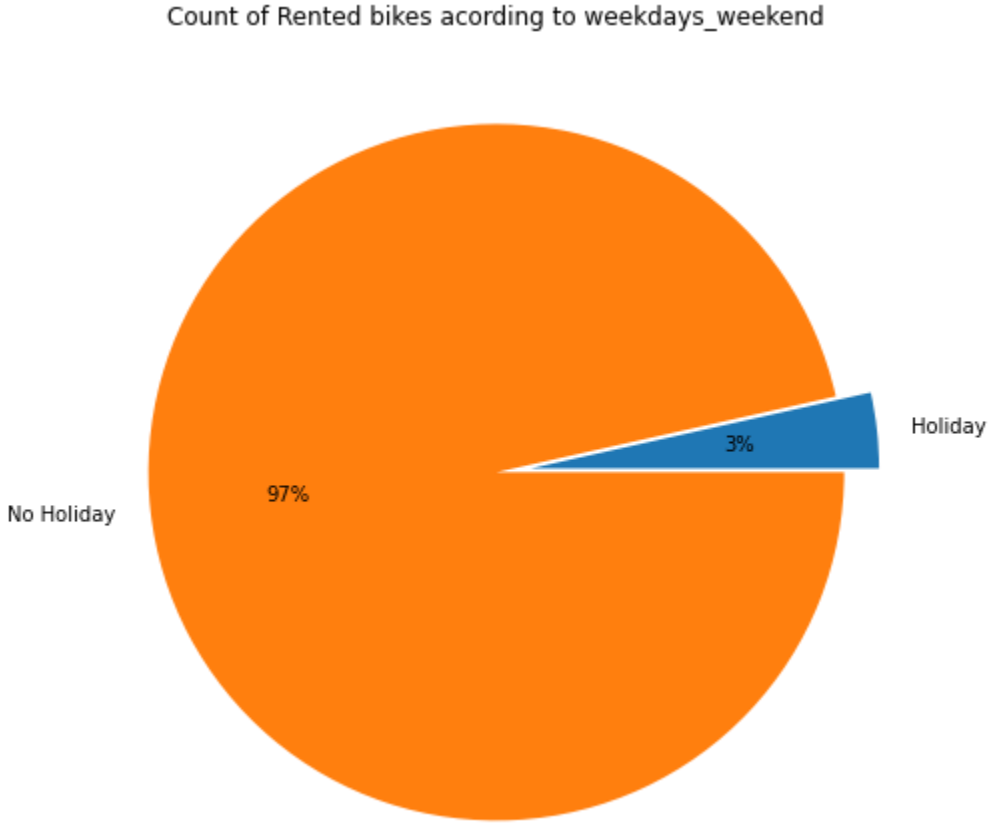


- ❑ High demand on morning 8 AM and Evening 5- 6 PM
- ❑ Quite good counts in day time afternoon to evening after decreased

EDA

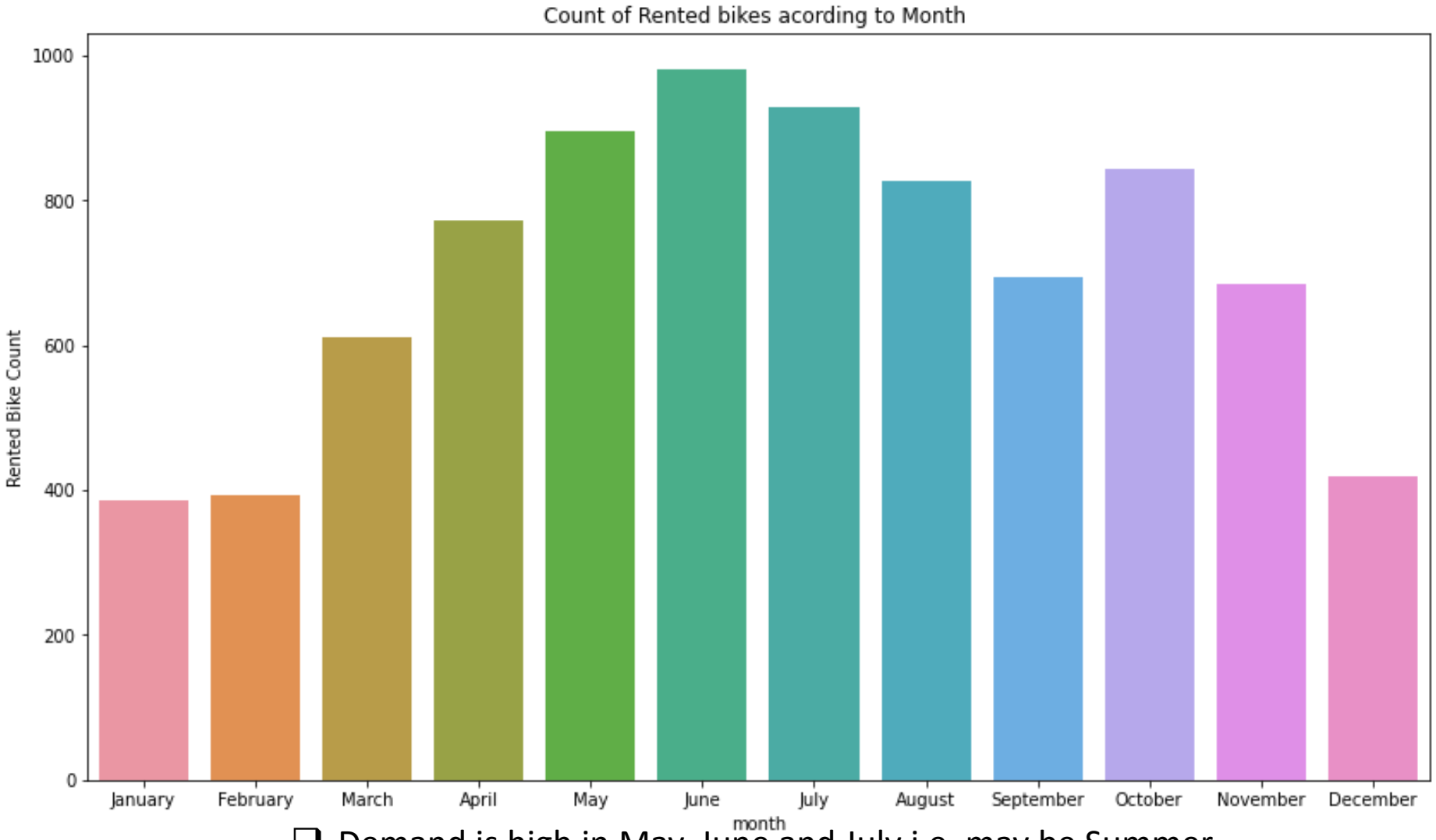


- More bikes are being rented in Summer season
- Less bikes are being rented in Winter Season



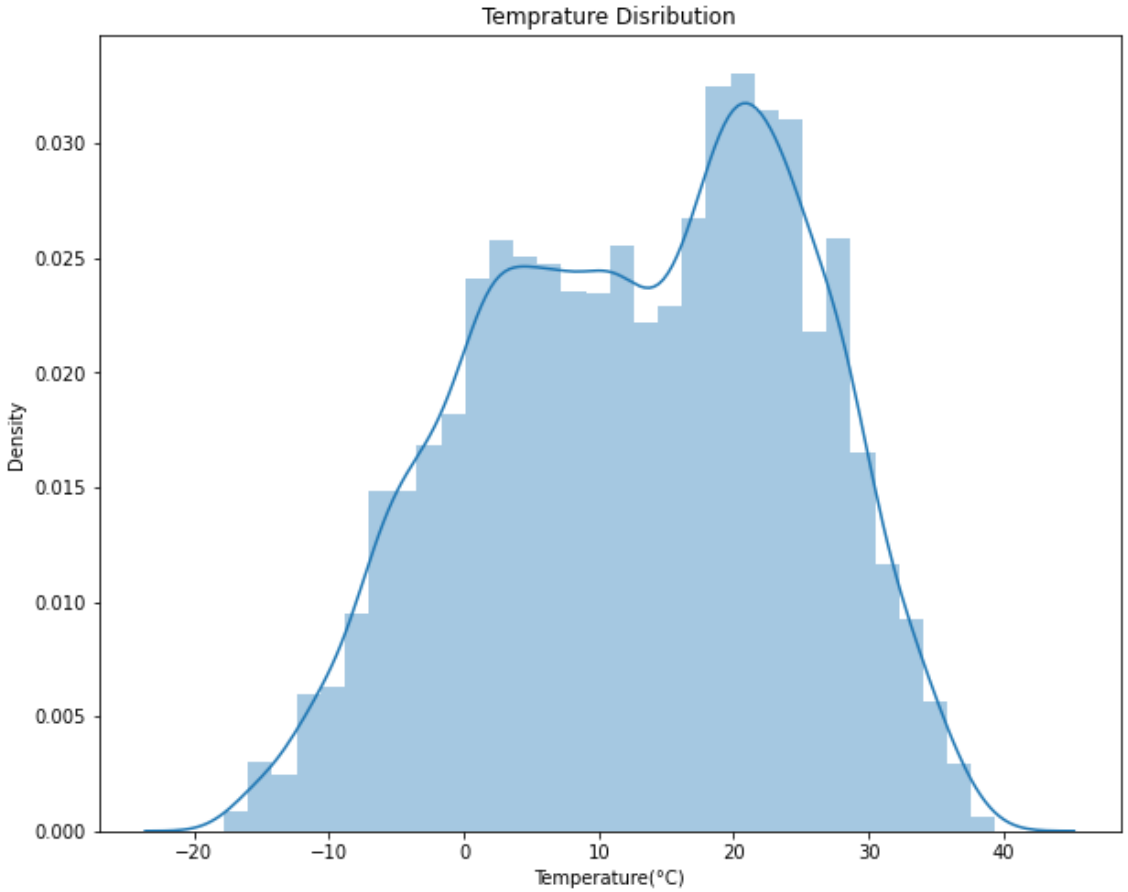
- More bikes are being rented in No Holiday

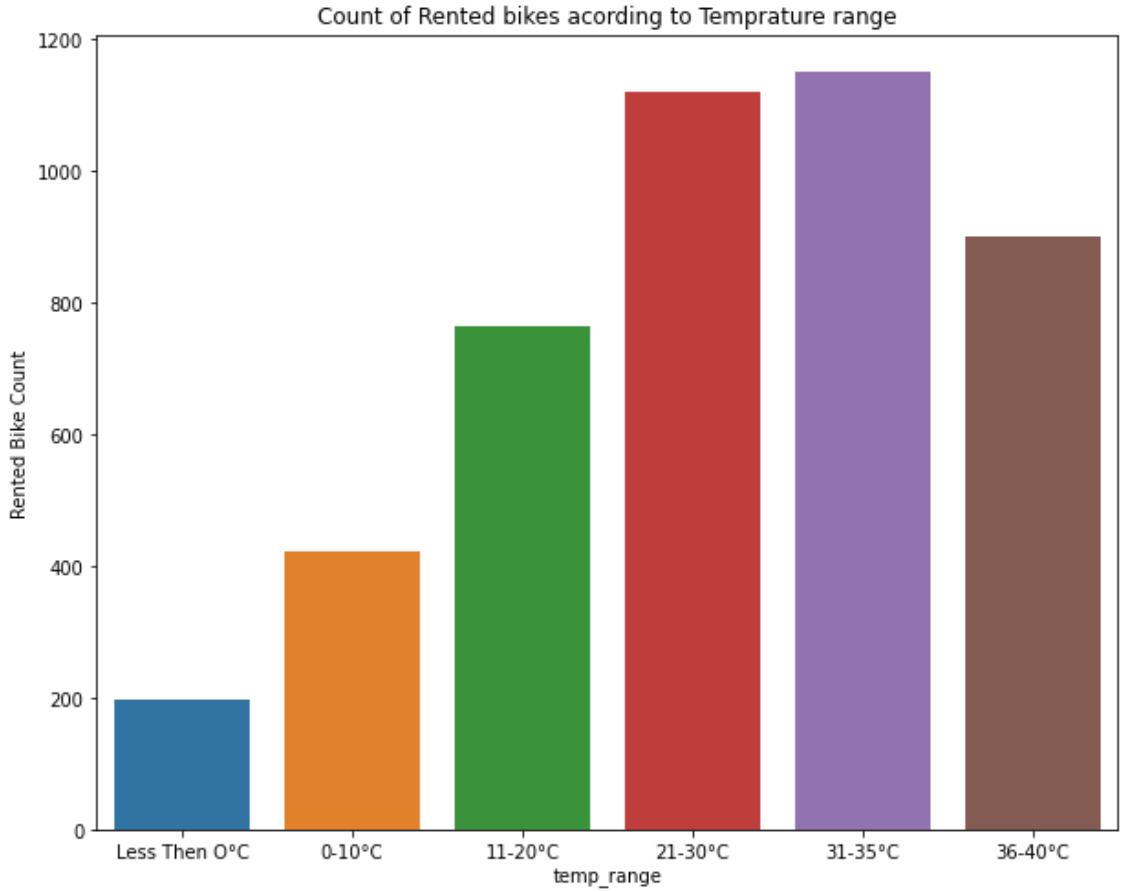
EDA



- Demand is high in May, June and July i.e. may be Summer Seasons

EDA

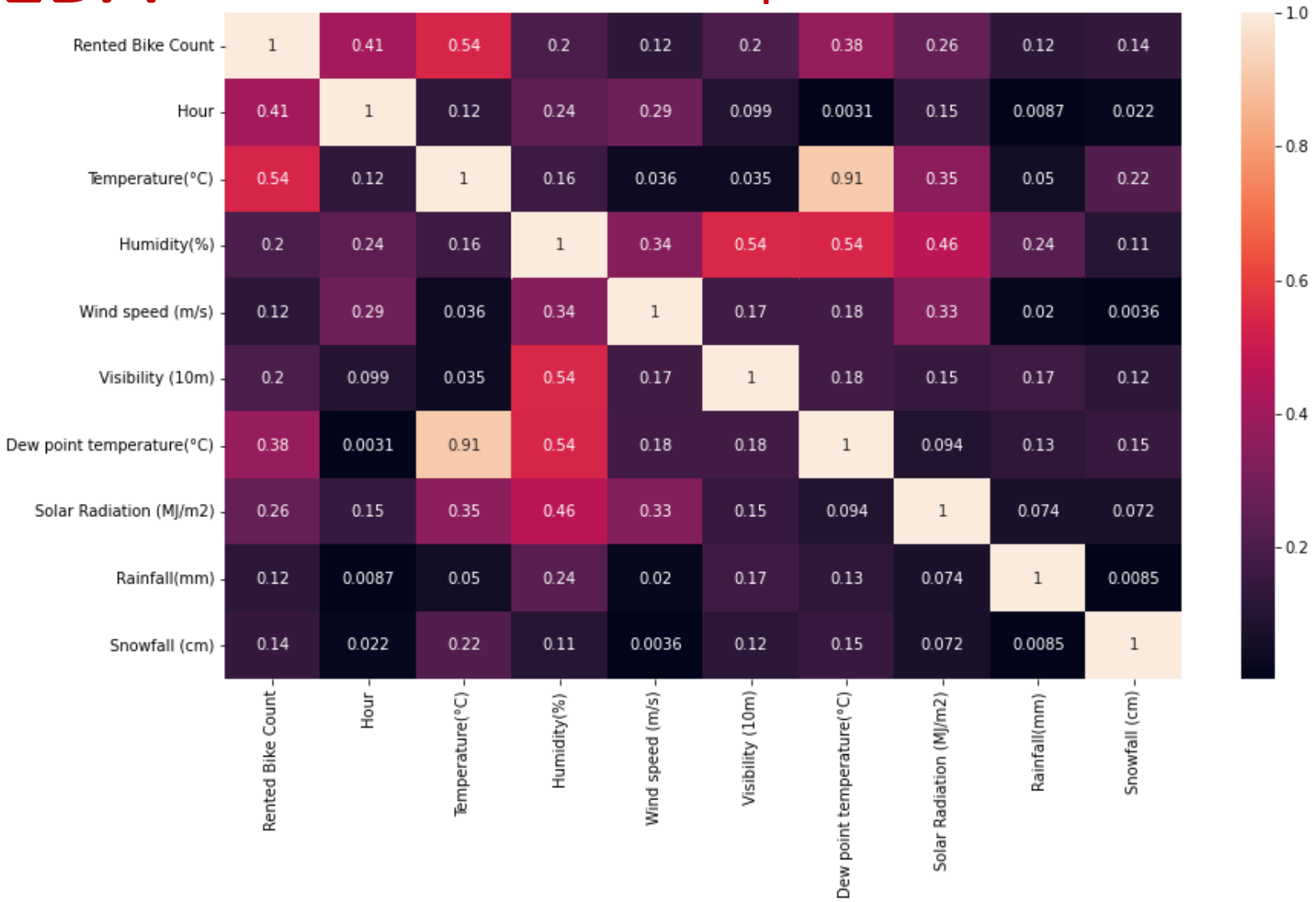




- Bikes are mostly rented between 21-35 Degree Celcious

EDA

Correlation Map



Multicollinearity

	variables	VIF
0	Hour	4.418242
1	Temperature(°C)	33.385256
2	Humidity(%)	5.371996
3	Wind speed (m/s)	4.805364
4	Visibility (10m)	9.085977
5	Dew point temperature(°C)	17.126199
6	Solar Radiation (MJ/m2)	2.881590
7	Rainfall(mm)	1.081567
8	Snowfall (cm)	1.120833

- There is 91% of correlation between Temperature and Dew point temperature feature as well as we can see VIF value of between them. We can remove one of the feature.
- Temperature is correlated with target variable with 54%

FEATURE SELECTION

After doing Exploratory Data Analysis, some Feature Engineering, finding correlation and multicollinearity , we filtered out the features that should be taken for model execution.

- Removing Temperature feature because High VIF value.

14 Final Independent Variable:-

Humidity(%), Wind speed (m/s), Visibility (10m), Solar Radiation (MJ/m2), Rainfall(mm), Snowfall (cm), Dew point temperature, Hour, Holiday, Functioning Day, season_Autumn, season_Spring', season_Summer', season_Winter

FITTING VARIOUS MODEL

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Elastic net Regression
5. Decision trees
6. Random Forest
7. Gradient Boosting
8. Extreme Gradient Boosting

MODEL PERFORMANCE COMPARISION

Evaluation matrices for all the models without Hyper parameter

	name	MAE_train	MAE_test	r2_score_train	r2_score_test	RMSE score_train	RMSE score_test
0	LinearRegression:	5.59	5.66	0.65	0.65	0.81	0.80
1	Lasso:	8.50	8.69	0.25	0.23	0.50	0.48
2	Ridge:	5.60	5.65	0.65	0.65	0.81	0.81
3	ElasticNet:	9.43	9.55	0.15	0.13	0.38	0.37
4	DecisionTreeRegressor:	0.00	3.69	1.00	0.80	1.00	0.90
5	RandomForestRegressor:	1.02	2.84	0.99	0.89	0.99	0.94
6	GradientBoostingRegressor:	3.32	3.51	0.87	0.86	0.93	0.92
7	XGBRegressor:	3.27	3.47	0.87	0.86	0.93	0.93

MODEL PERFORMANCE COMPARISION

Evaluation matrices for all the models with Hyper parameter Tuning

	name	MAE_train	MAE_test	r2_score_train	r2_score_test	RMSE score_train	RMSE score_test
0	LinearRegression:	5.59	5.66	0.65	0.65	0.81	0.80
1	Lasso:	8.50	8.69	0.25	0.23	0.50	0.48
2	Ridge:	5.60	5.65	0.65	0.65	0.81	0.81
3	ElasticNet:	9.43	9.55	0.15	0.13	0.38	0.37
4	DecisionTreeRegressor:	0.00	3.69	1.00	0.80	1.00	0.90
5	RandomForestRegressor:	1.02	2.84	0.99	0.89	0.99	0.94
6	GradientBoostingRegressor:	3.32	3.51	0.87	0.86	0.93	0.92
7	XGBRegressor:	3.27	3.47	0.87	0.86	0.93	0.93
8	Lasso_Hyper_tuned:	5.87	6.00	0.62	0.62	0.79	0.79
9	Ridge_Hyper_tuned:	5.59	5.66	0.65	0.65	0.81	0.81
10	RandomForestReg_Hyper_tuned:	2.63	3.21	0.91	0.86	0.95	0.93
11	XGBRegressor_Hyper_tuned:	1.55	2.74	0.97	0.89	0.98	0.95
12	GradientBoostingReg_Hyper_tuned	2.57	3.00	0.91	0.89	0.96	0.94

MODEL VALIDATION

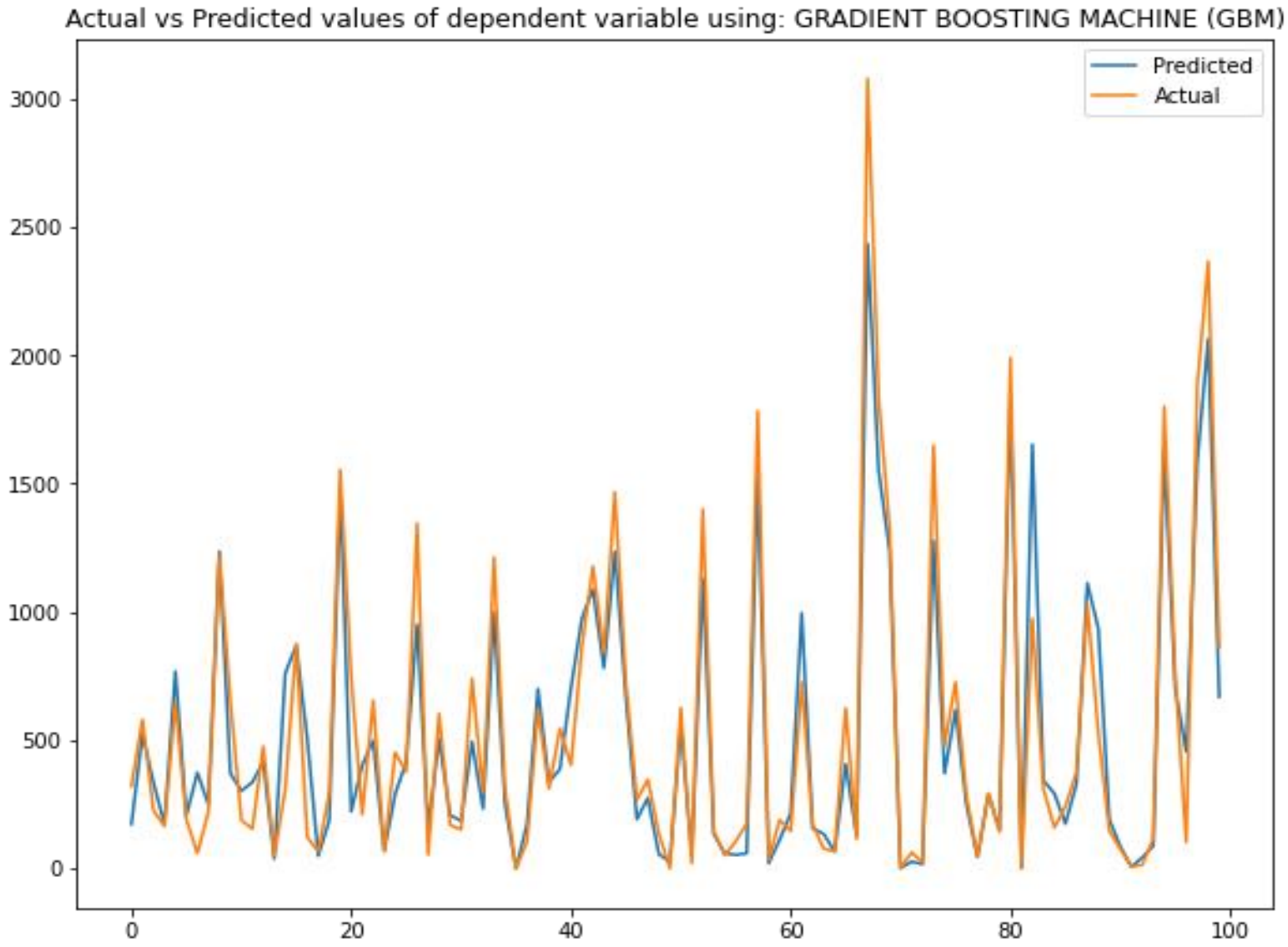
- By observing Evaluation matrices for all the models-

  - ❑Linear Regression, Lasso, Ridge and ElasticNet are not at its best

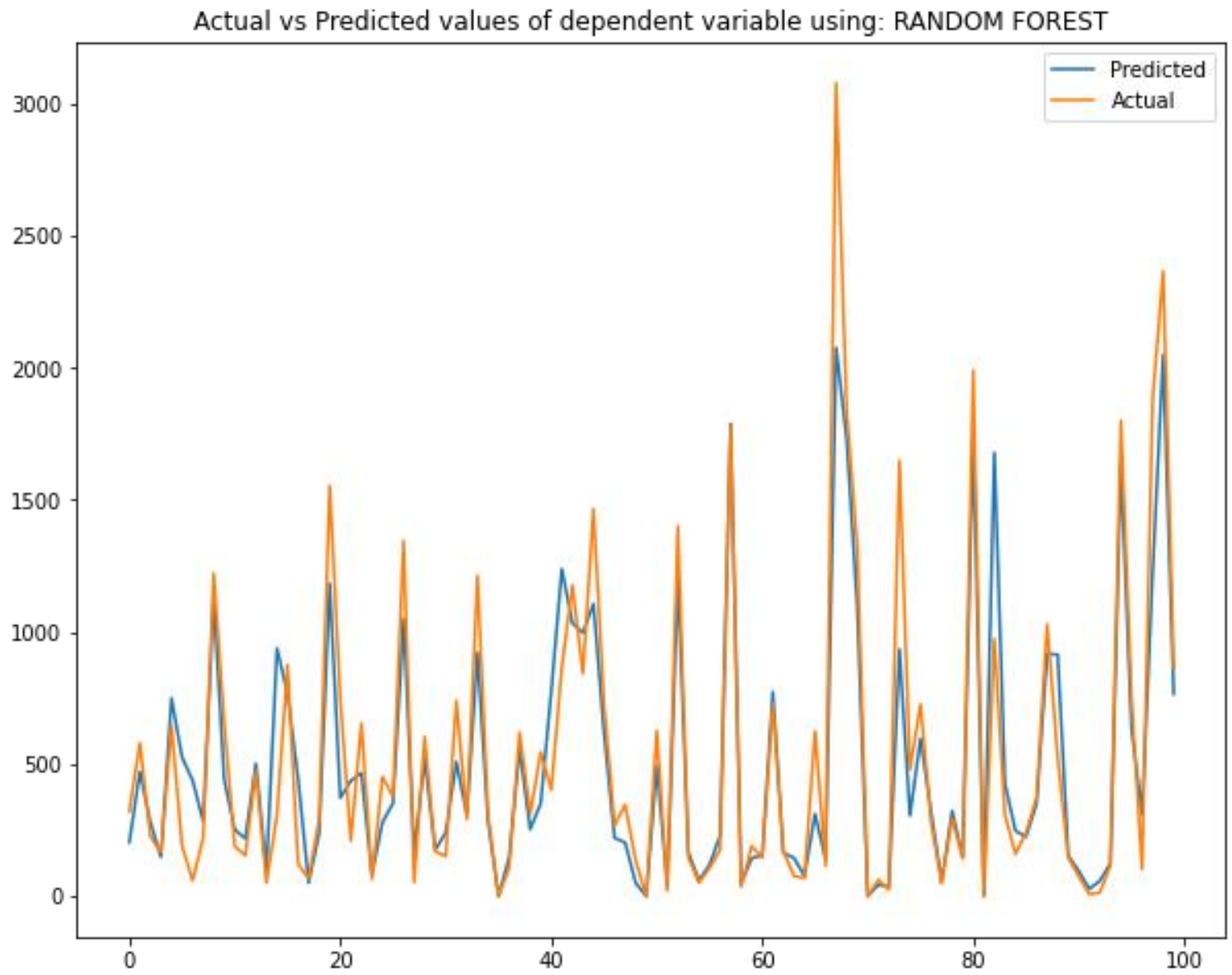
  - ❑Decision Trees, Random Forest are Over Fitting with Training Model and Gradient boosting , XGB gives better results.

  - ❑When we tuning hyperparameter with Grid SearchCV, Gradient boosting and Random Forest giving better results.

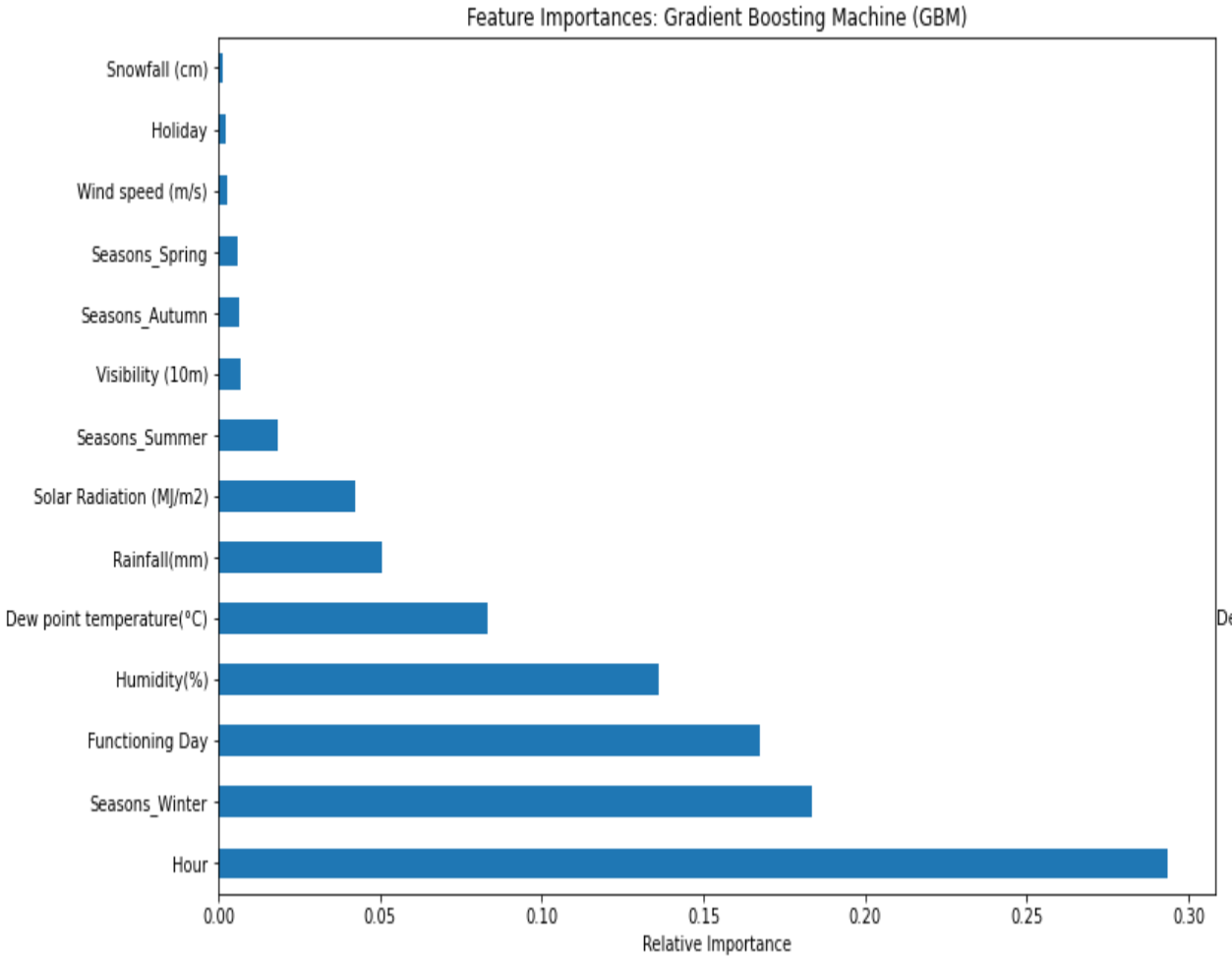
GBM Actual and Predicted Behaviour

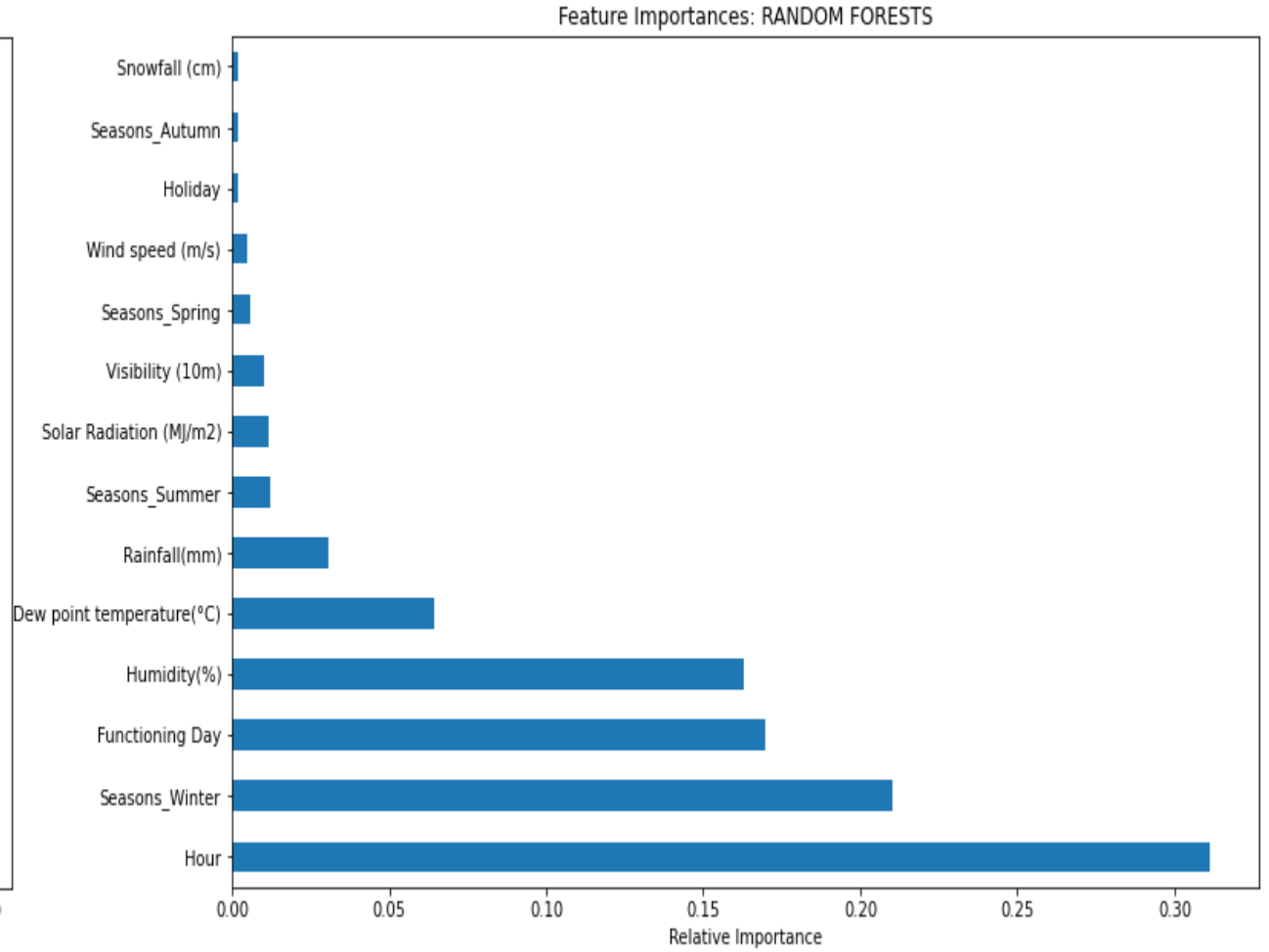


Random Forest - Actual and Predicted Behaviour



Features Importance





CONCLUSION

- From the applying some of the Regression models, we got some evident that GBM will perform better among all the models for the Bike Sharing Demand Prediction, since the evaluation matrices was best for this model(r2 Score).

- Hours Season_Winter, Functioning day and Humidity, These features contributes heavily to predict our target variable.


THANK
YOU