Capstone Project

(SUPERVISED ML – REGRESSION)

Seoul Bike Sharing Demand Prediction

TEAM

- IQBAL BABWANE
- SAMEER ANSARI
- LUKMAN HAIDER KHAN

~ UNDER THE GUIDANCE OF TEAM ALMABETTER

CONTENT

- PROBLEM STATEMENT
- METHODOLOGY
- INTRODUCTION OF PROJECT
- DATA DESCRIPTION
- 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.

METHODOLOGY

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.

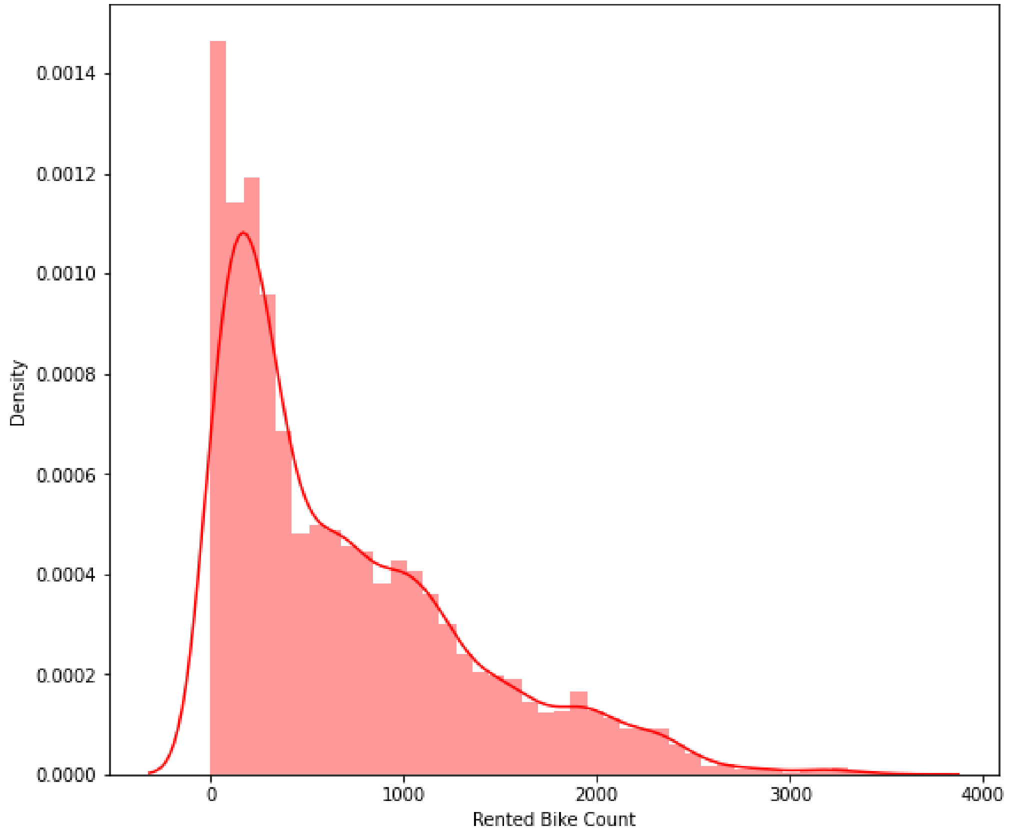
Attribute Information:

- 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)

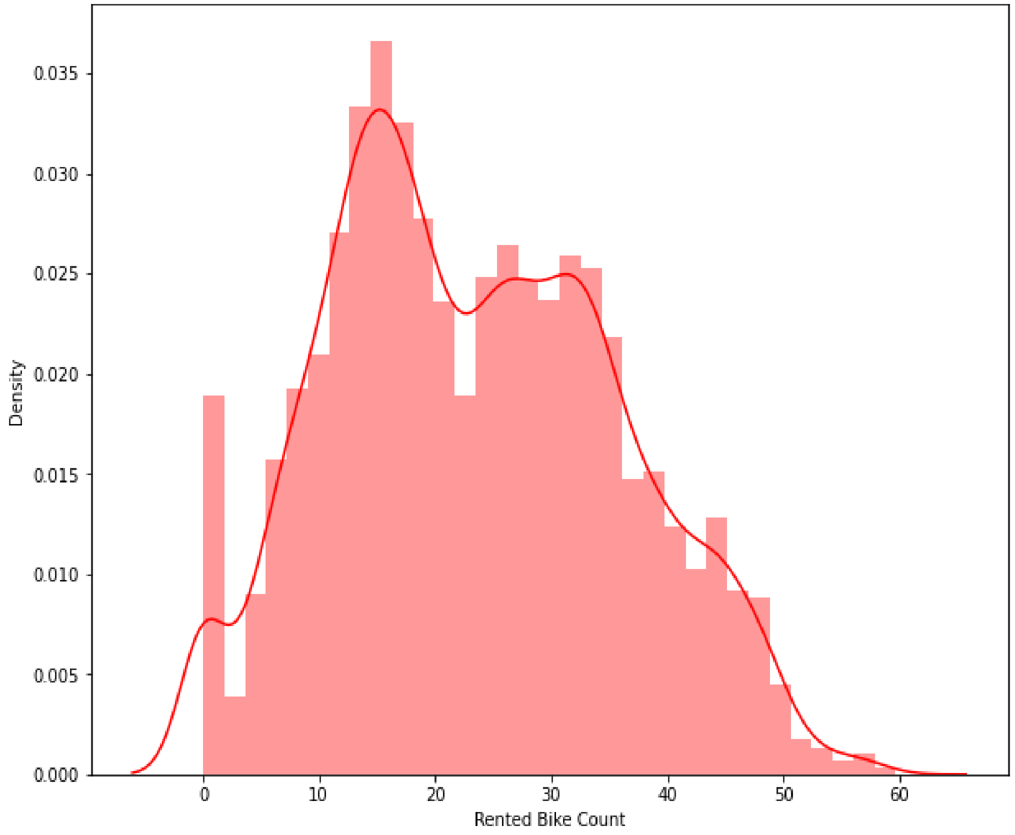
EDA

Data distribution of target variable

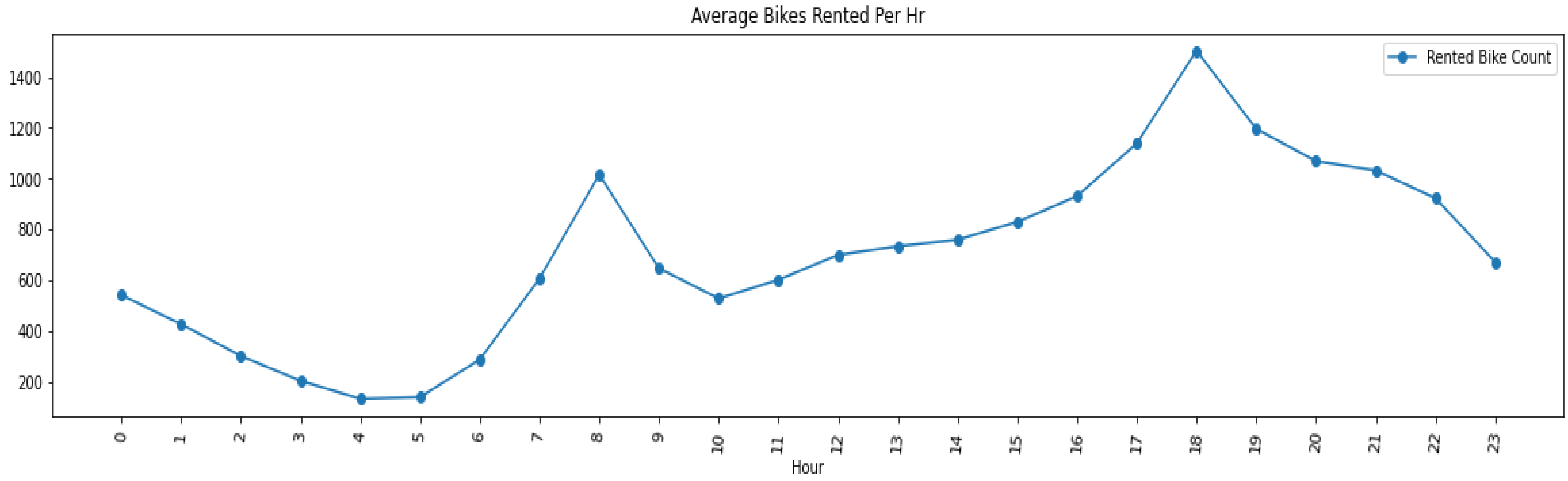
Before transformation



After using sqrt transformation

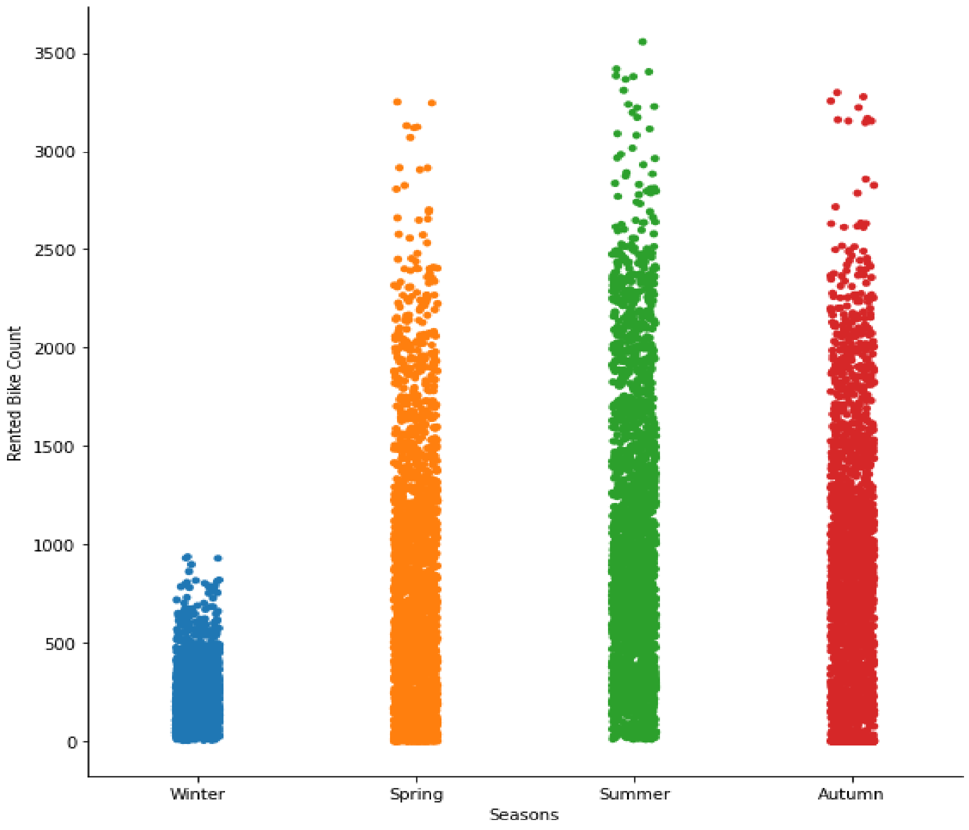


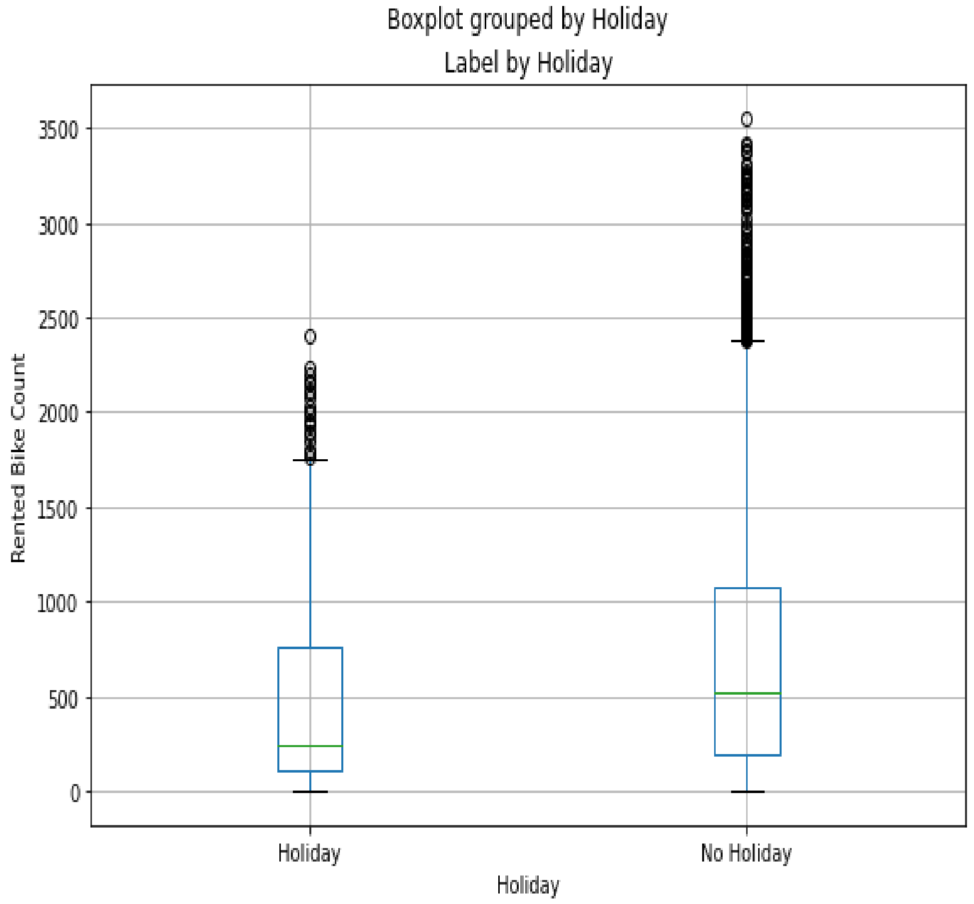
EDA



- ❑ High demand on morning 8 AM and Evening 6 PM
- ❑ Quite good counts in afternoon and evening as well

EDA

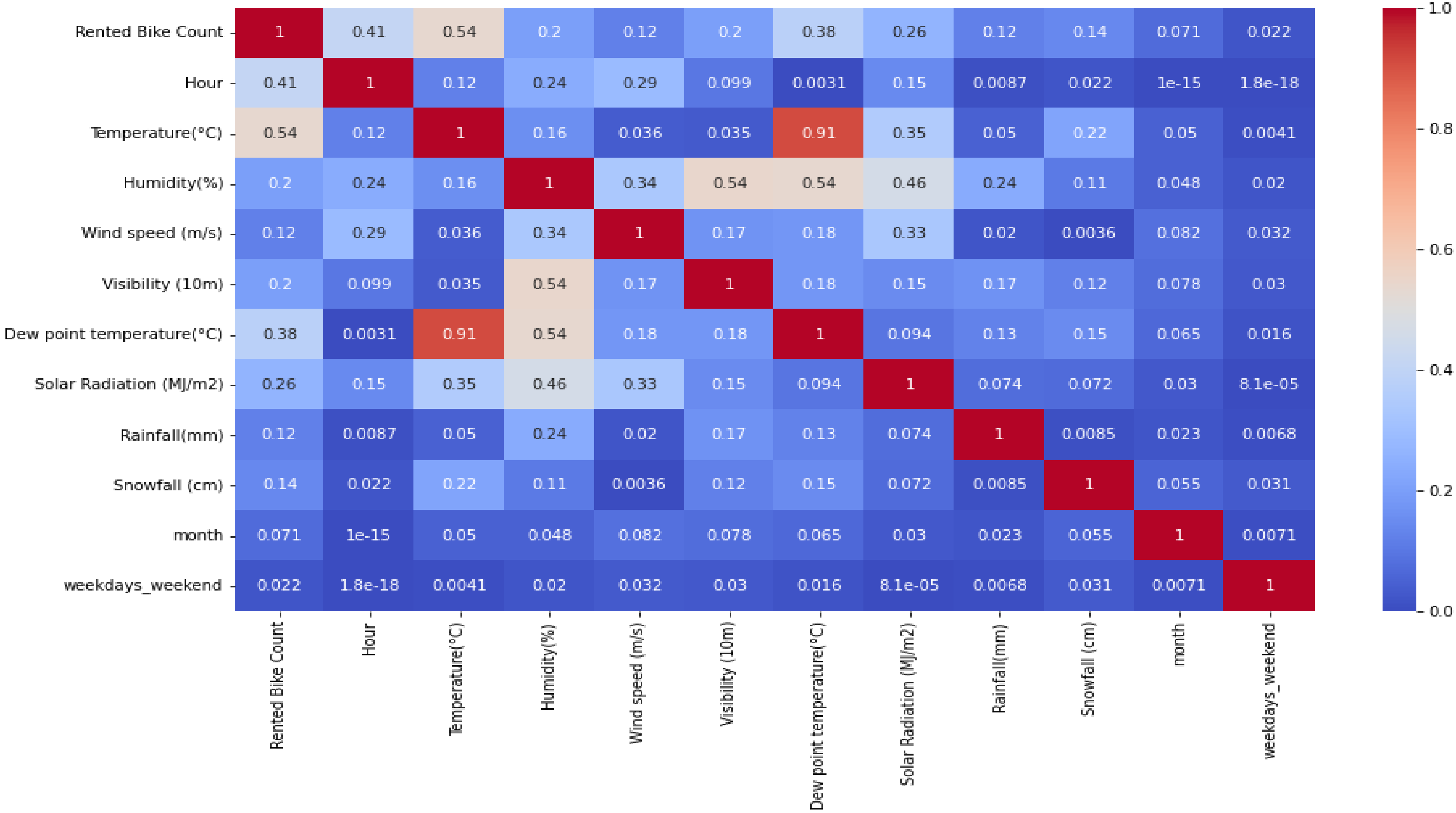




- ❑ Less bikes are being rented in winter season
- ❑ Bikes are mostly rented on working days i.e when there is no holiday

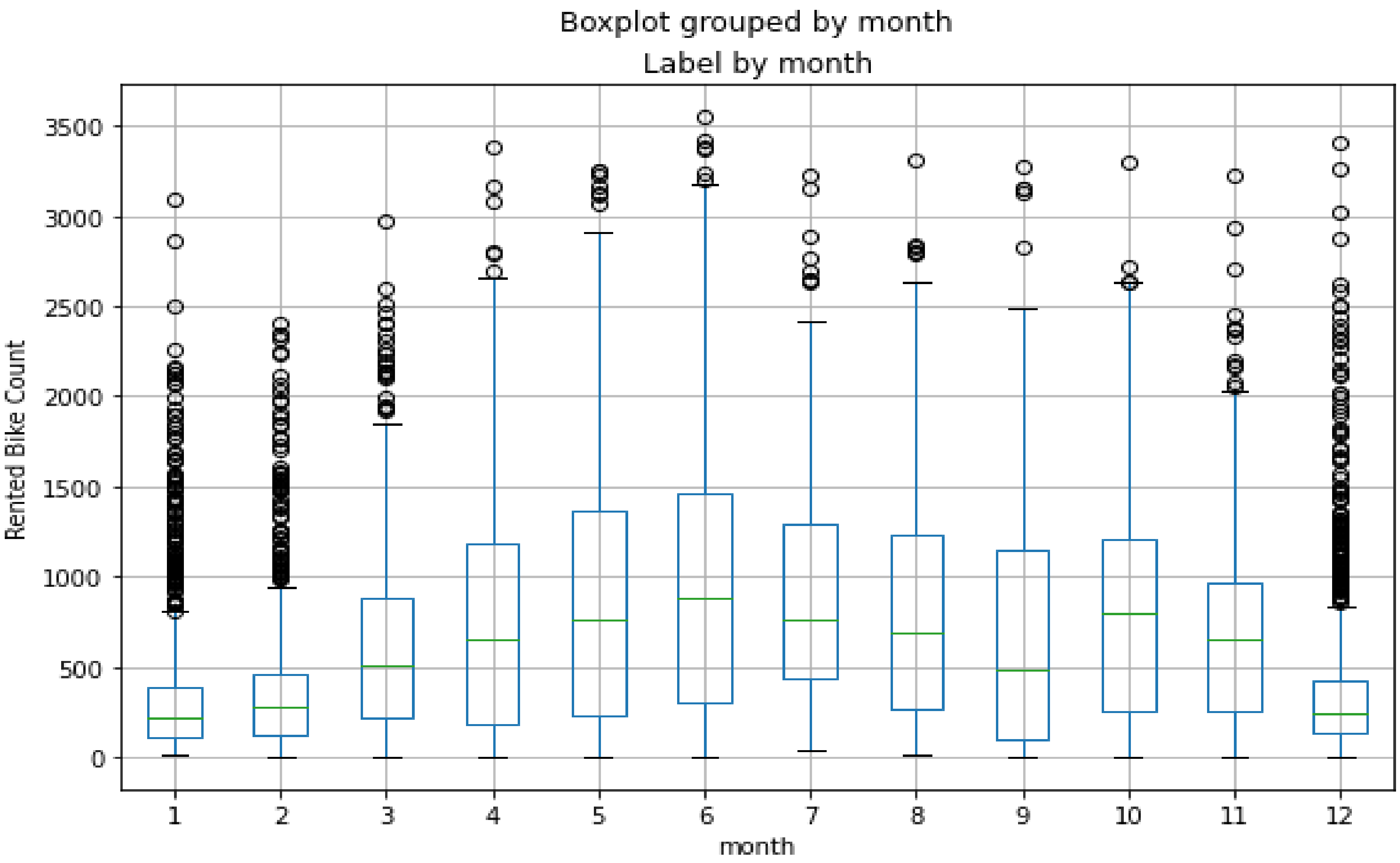
EDA

Multicollinearity



- There is 91% of collinearity between Temperature and Dew point temperature feature
- Temperature is correlated with target variable with 54%

EDA



- ❑ Demand is high in April, May, June i.e. Summer Seasons

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.

Final features :-

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

FITTING VARIOUS MODEL

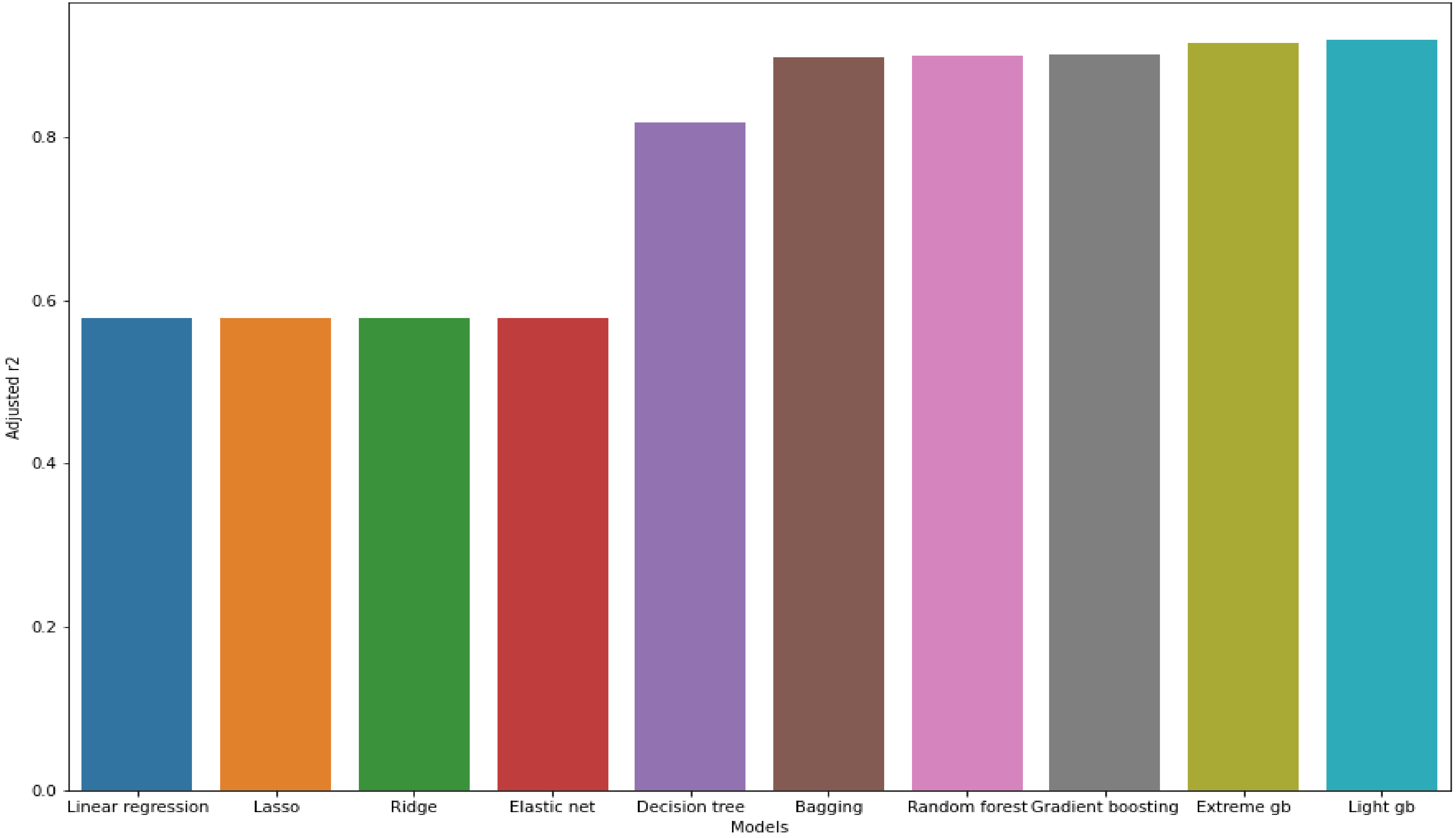
1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Elastic net Regression
5. Decision trees
6. Bagging Regressor
7. Random Forest
8. Gradient Boosting
9. Extreme Gradient Boosting
10. Light Gradient Boosting Machine

MODEL PERFORMANCE COMPARISION

Evaluation matrices for all the models

	Linear regression	Lasso	Ridge	Elastic net	Decision tree	Bagging	Random forest	Gradient boosting	Extreme gb	Light gb
MSE	174696.237393	175057.634536	174771.535577	175198.731562	75617.154680	42047.132786	41666.587882	40838.874664	34939.040485	33568.364275
RMSE	417.966790	418.398894	418.056857	418.567476	274.985735	205.053975	204.123952	202.086305	186.919877	183.216714
r2	0.582588	0.581725	0.582408	0.581388	0.819324	0.899534	0.900444	0.902421	0.916518	0.919793
Adjusted r2	0.578739	0.577867	0.578557	0.577527	0.817657	0.898608	0.899526	0.901521	0.915748	0.919054

MODEL PERFORMANCE COMPARISION

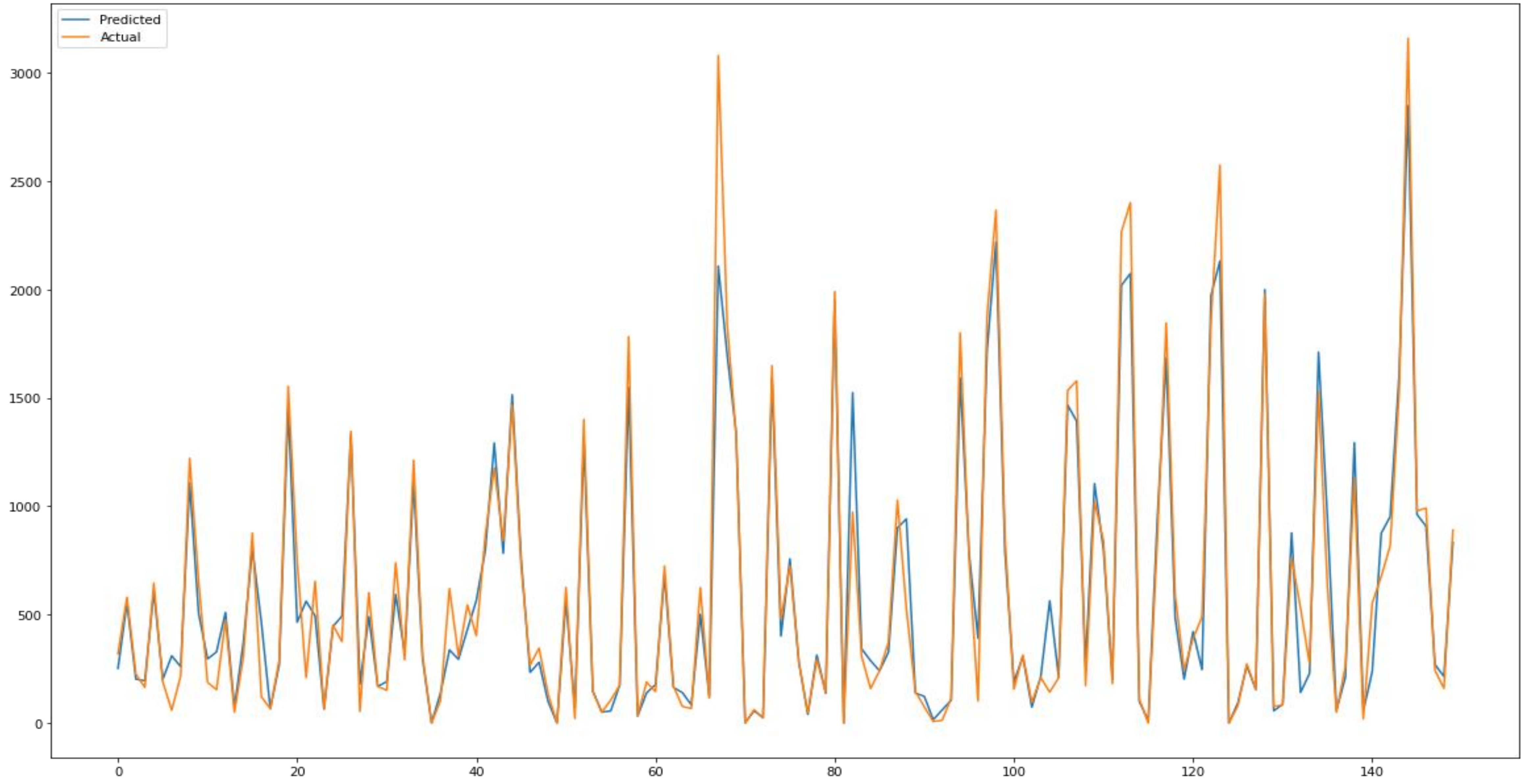


Adjusted R2 Matrix score for all the corresponding models

MODEL VALIDATION

- By observing Adjusted R2 Scores for all the models-
  - ❑Linear Regression, Lasso and Ridge are not at its best
  - ❑Decision Trees, Bagging, Random Forest are quit good with linear models, but they are not giving optimum prediction
  - ❑Gradient boosting type models are giving better results, while Light GBM is performing well among all having 91% score, so we can use Light GBM as our final model for prediction

Light GBM Actual and Predicted Behaviour

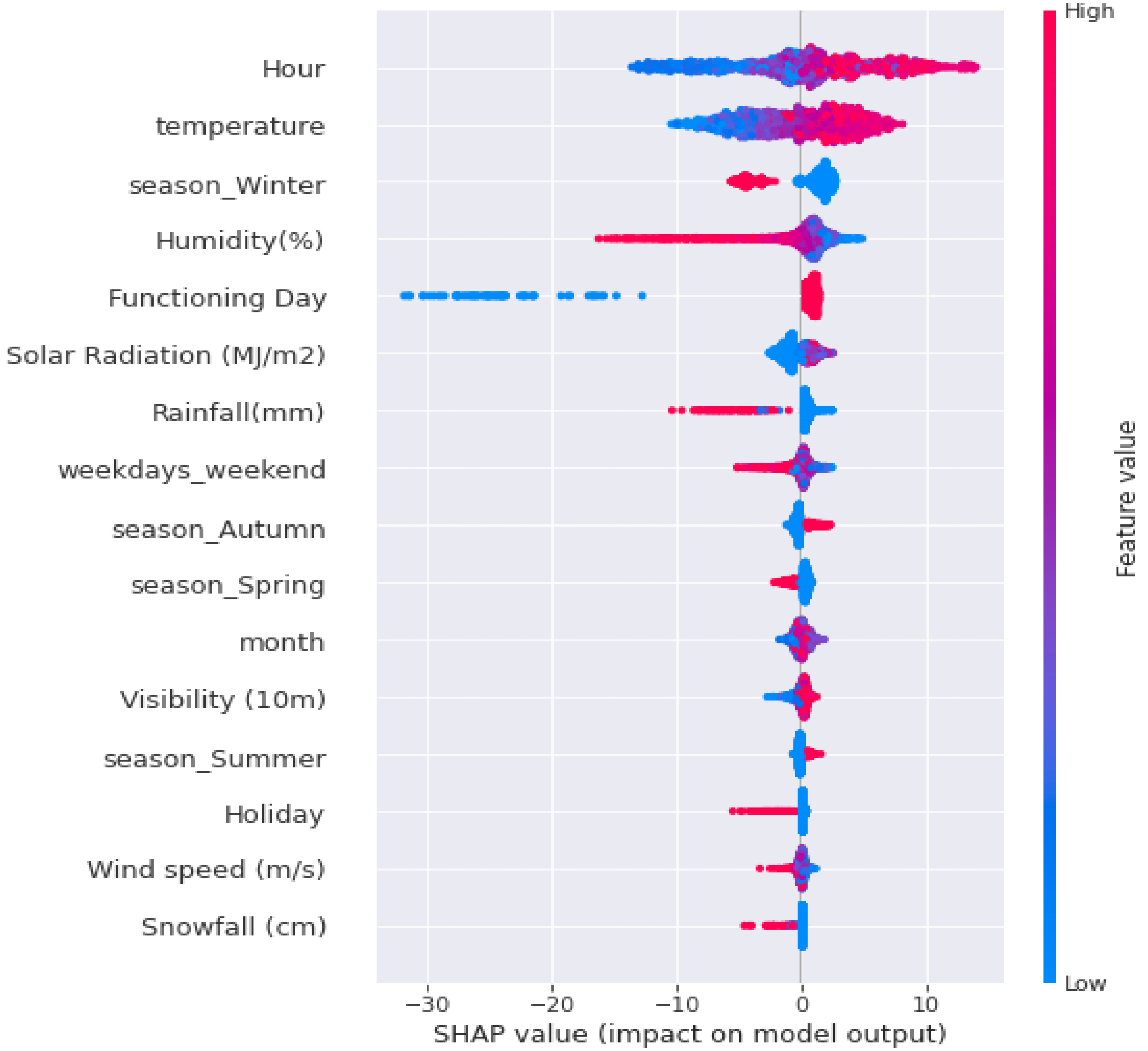


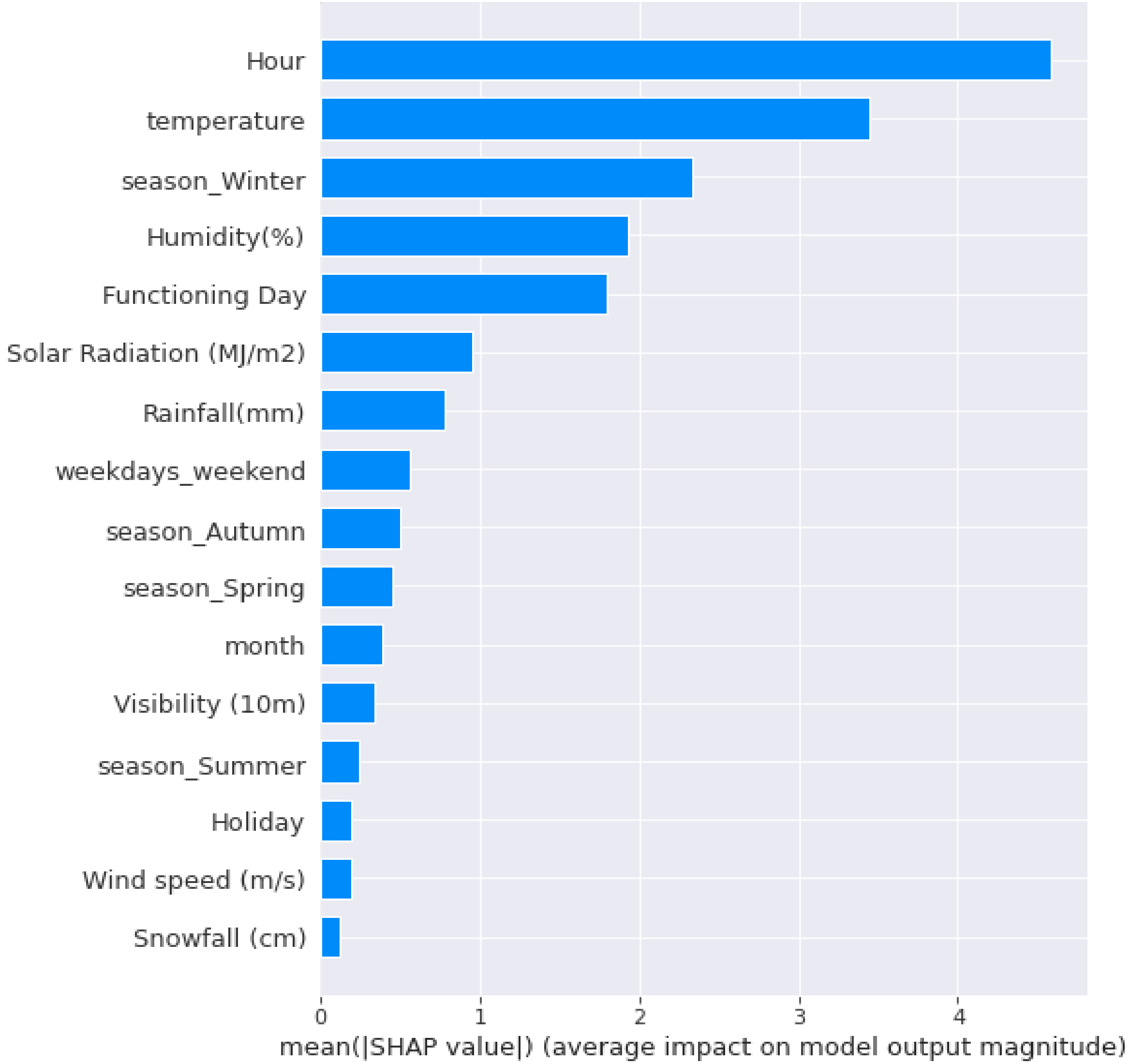
MODEL EXPLAINABILITY

For a data point in Light GBM, we got-

- ❑ Base value = 23.55
- ❑ Output value = 24.64

Top features which helping to make our prediction





CONCLUSION

- From the previous slides we got some evident that Light GBM will perform better among all the models for the Bike Sharing Demand Prediction, since the evaluation matrices was best for this model.

- Hours and temperatures, both the features contributes heavily to predict our target variable.


THANK
YOU