Capstone Project

"Bike Sharing Demand Predictions"

Team : Zizzle Stark

Appendix

- Problem Statement
- Data Summary
- Exploratory Data Analysis (Inferences)
- Correlation Check
- Splitting the Data
- Modelling the Data
- Insights of Evaluation of Matrix of all Models
- 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 reduces 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.

Summary of the Data

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

- Focusing on the ultimate goal, which comprehends to combine the actual and past bike usage pattern with season in order to forecast Bike Sharing Demand

Exploratory Data Analysis (Inferences)

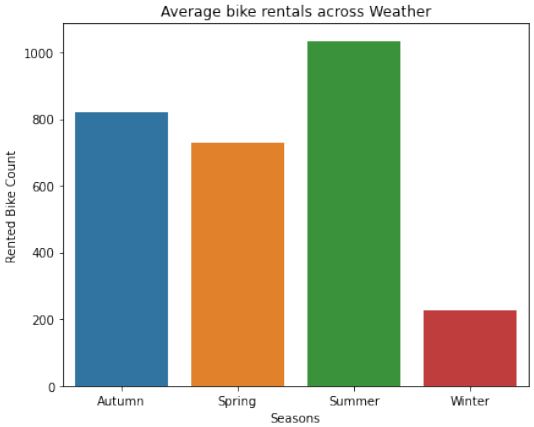
- Multicollinearity Check
- Regression plot Check
- Inference Received :

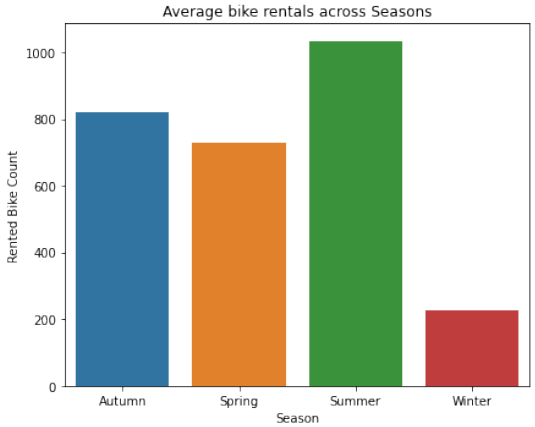
->During Winters the Demand of the Bike Sharing is Low compared to other Seasons, Where Summers seasons aces the demands of rental bikes,

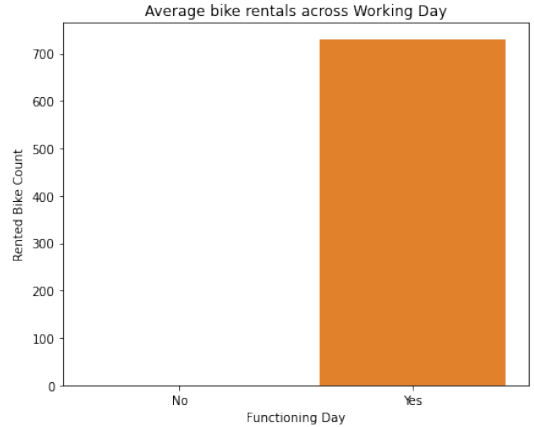
-> Moreover Demands of Rental Bikes are Slightly Higher in Non holidays,

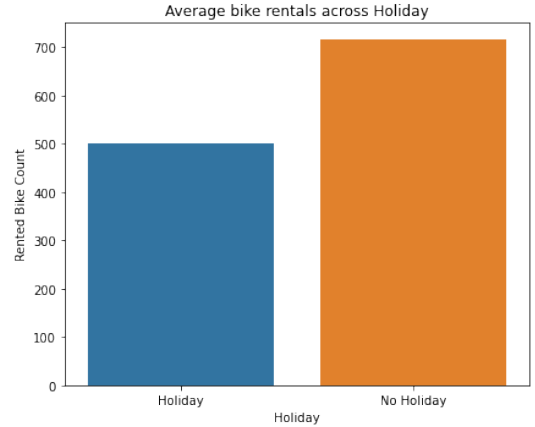
-> Almost Zero demand on Non functional Day.

Average Bike rental Count v/s Categorical Columns

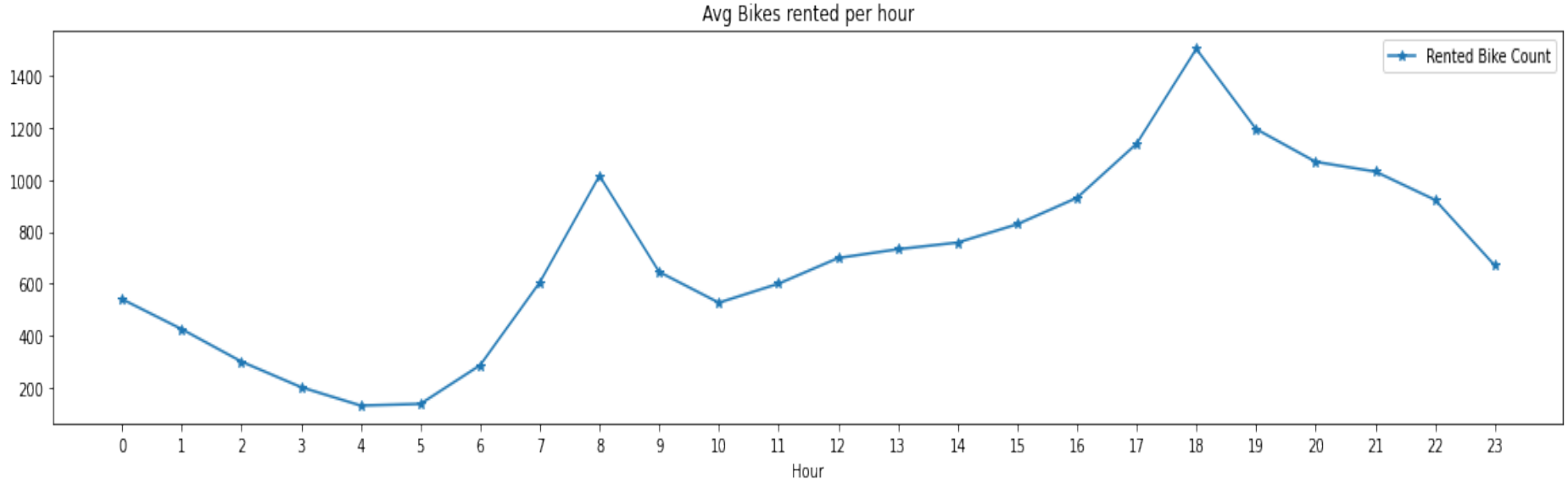






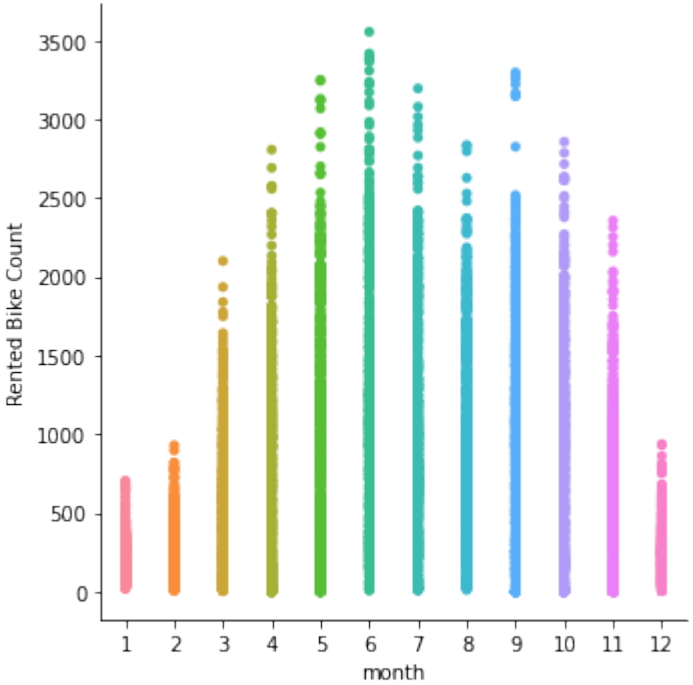


Average Rented Bikes per hour



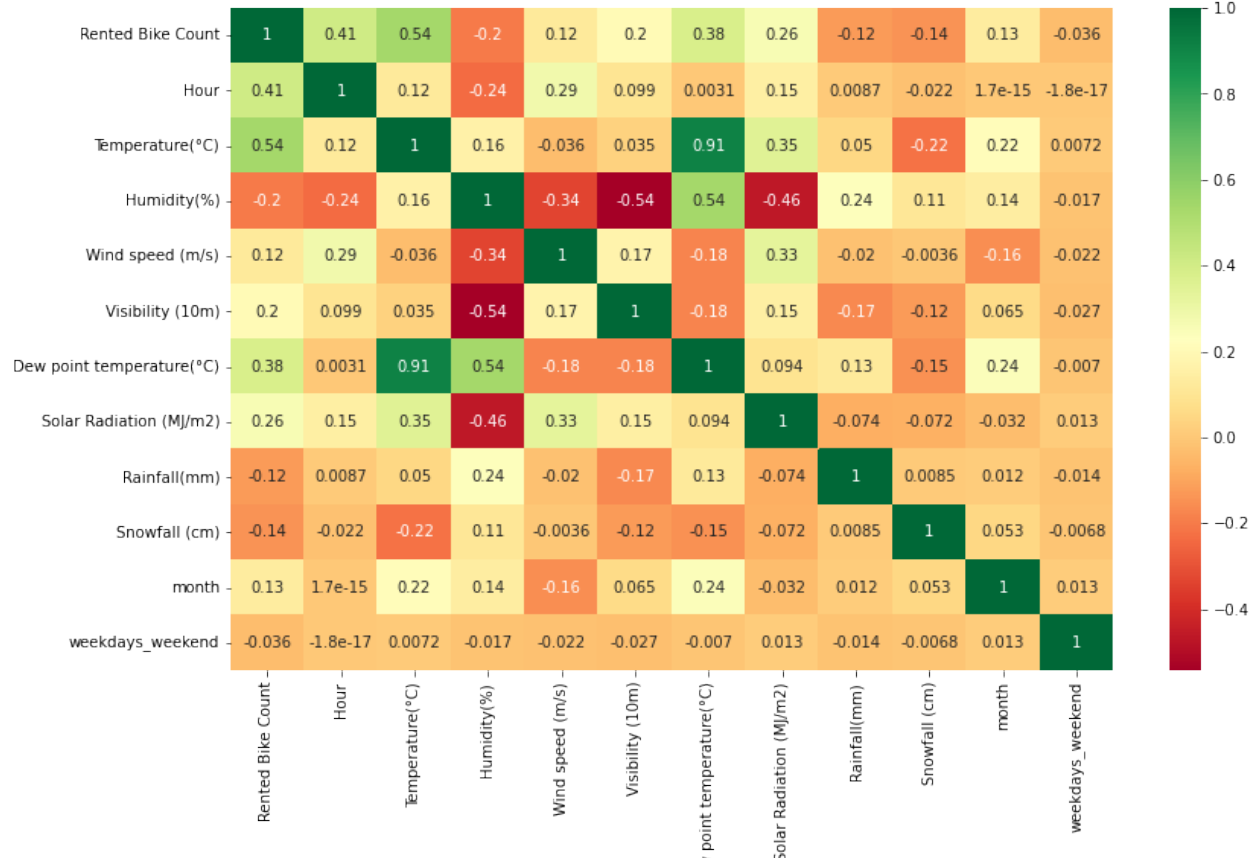
In the comparison between "**hour vs rented count of bikes**" we can clearly notice a high demand in the rush hour of 8:00 am to 9:00 pm.

Rented Bike Count v/s Month



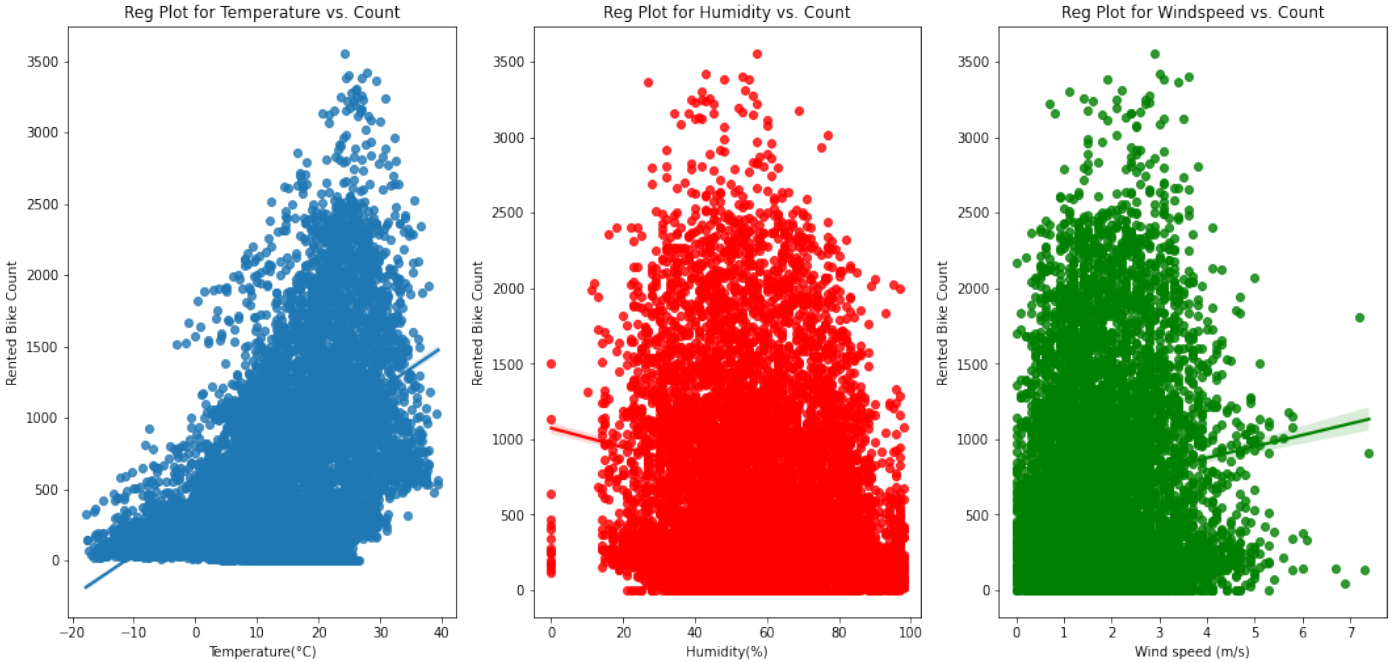
Clearly, we get a notion from the graph, That In the Winter Months the Demand decreases i.e, December, January, February, However demand spikes up in Summer months i.e May, June , July..

Correlation Check



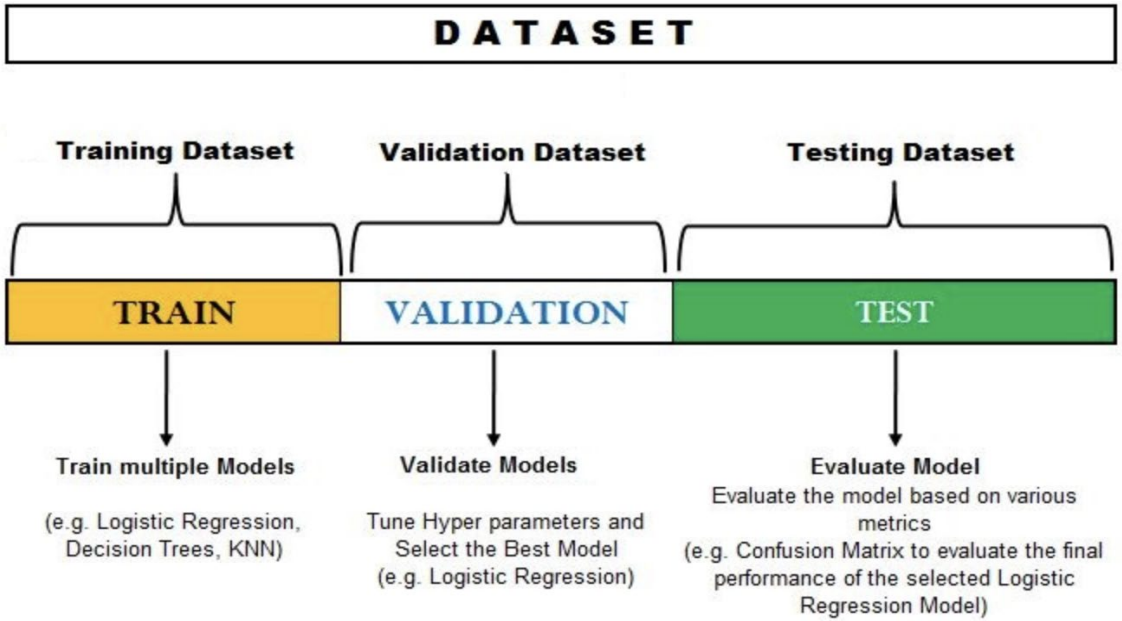
We get the notion that all features are strongly correlated with each other.

Correlation Check



The regplot indicates a positive correlation of count with temperature and windspeed and a negative correlation with humidity

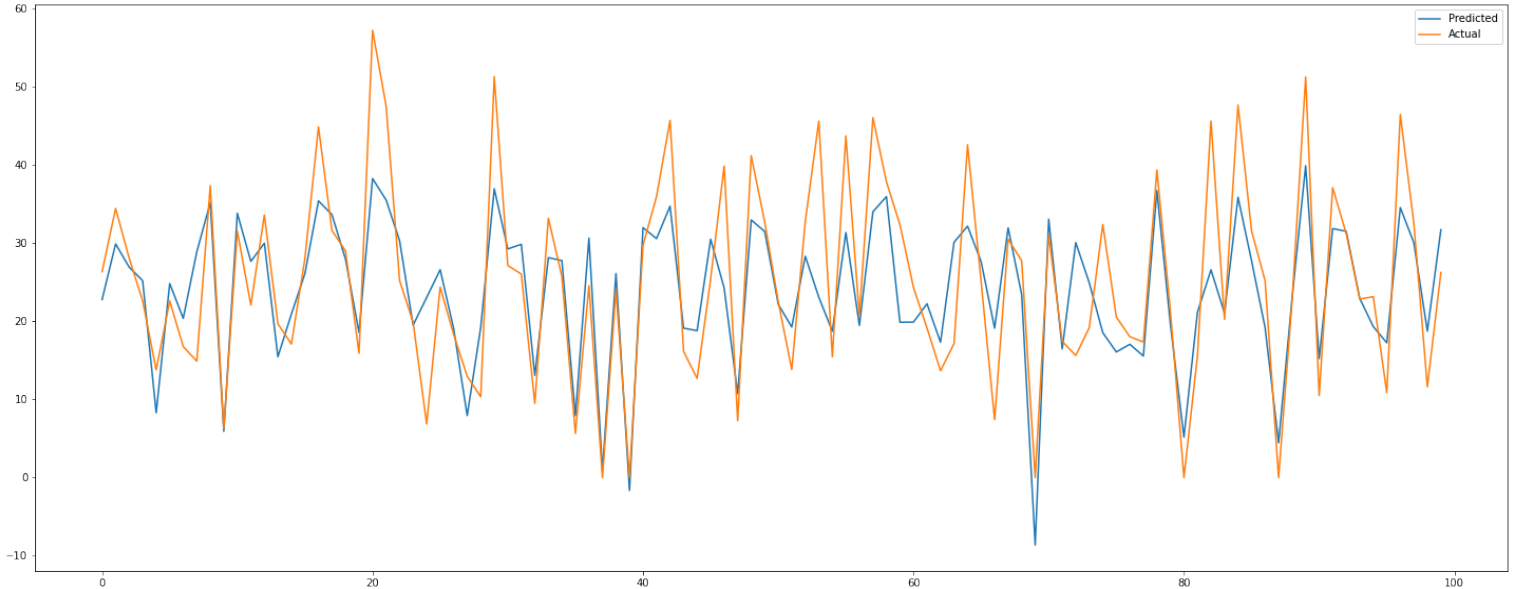
Splitting the Data



Dataset splits into these training , validation and test dataset, we'll take the test data for analysing
the performance of training data set and apply modelling to it.

Modelling the Data

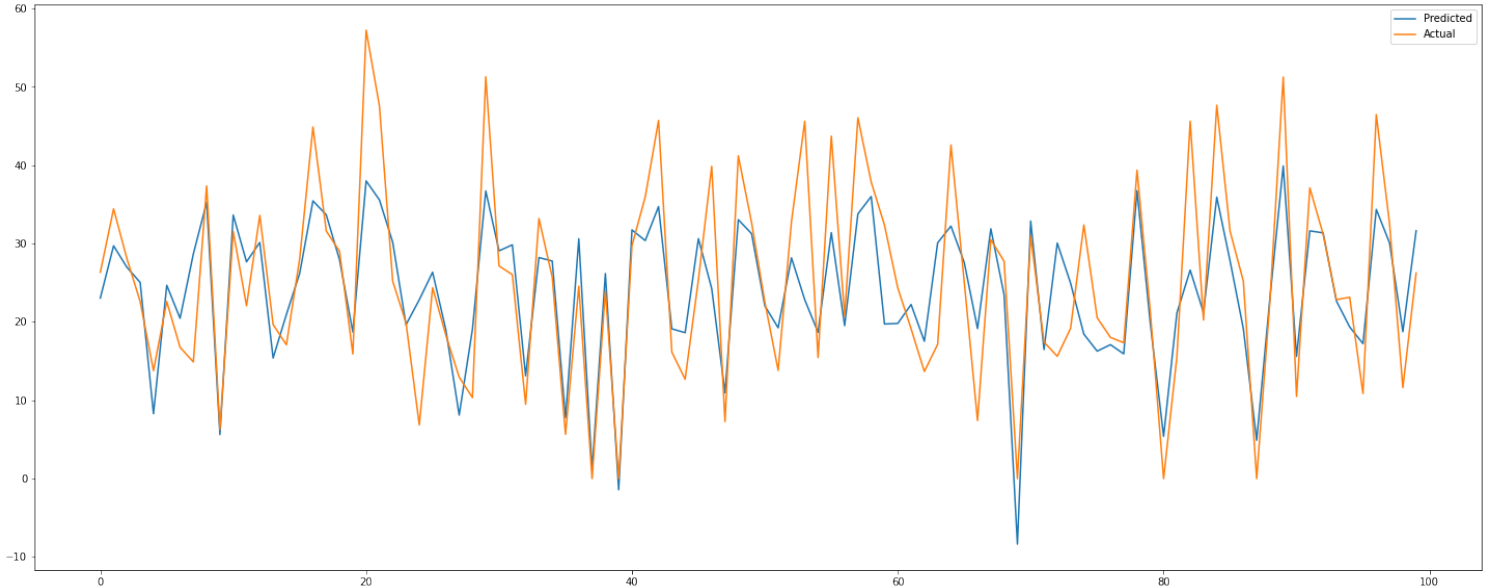
Linear Regression : Linear regression is a basic and commonly used type of predictive analysis.



R2 value and adjusted R2 value are significantly moderate so the model is not performing well in this scenario.

Modelling the Data

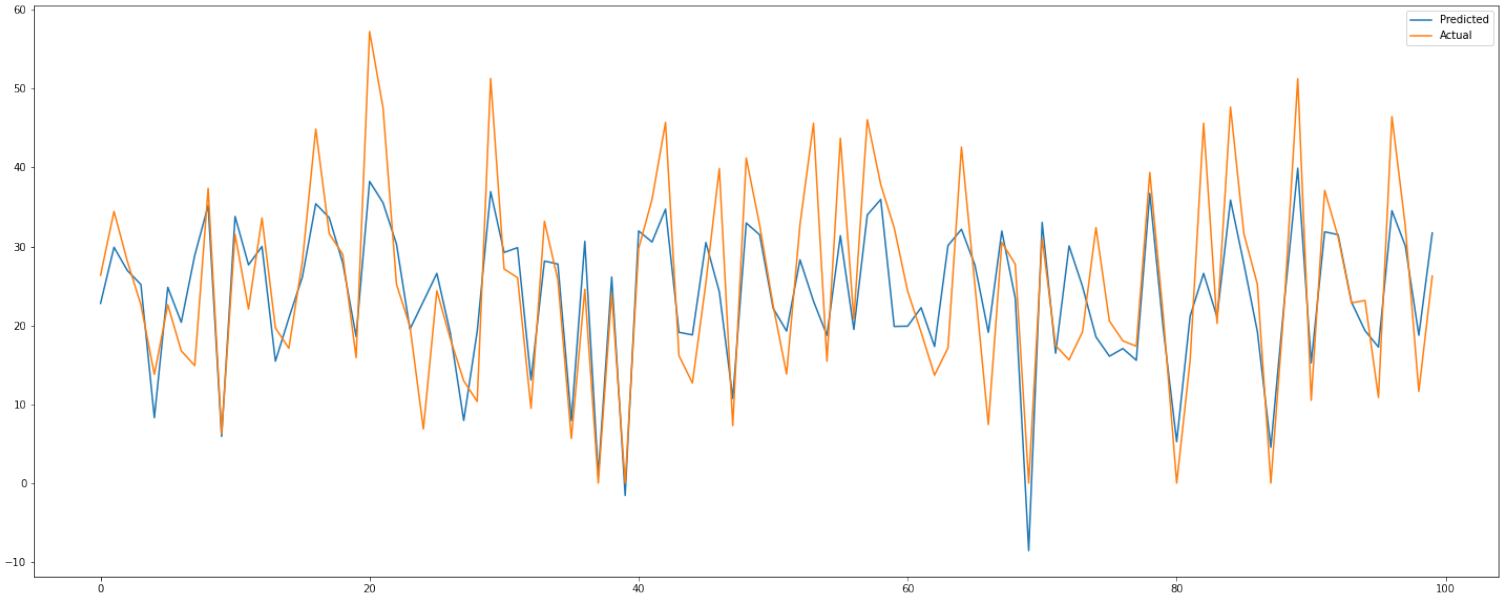
Lasso Regression : Lasso regression is a type of linear regression that uses shrinkage.



R2 value and adjusted R2 value are significantly moderate so the model is not performing well in this scenario, This states that our linear model is not performing well

Modelling the Data

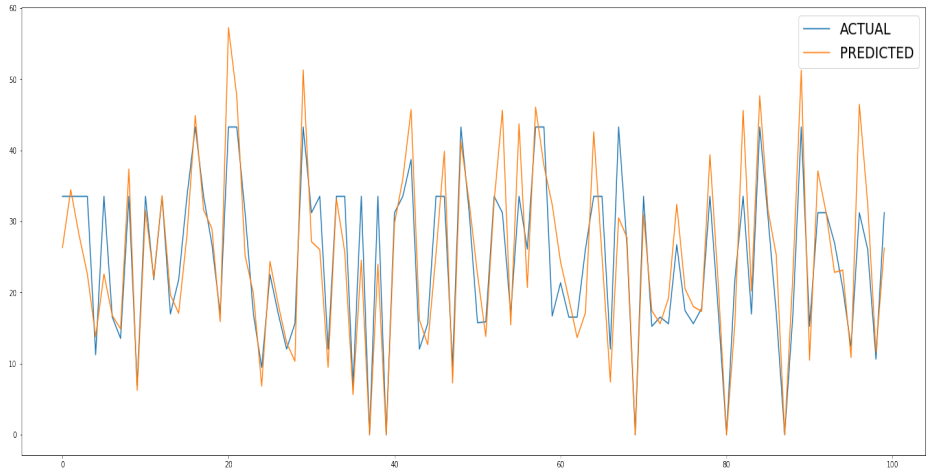
ELASTIC NET REGRESSION.



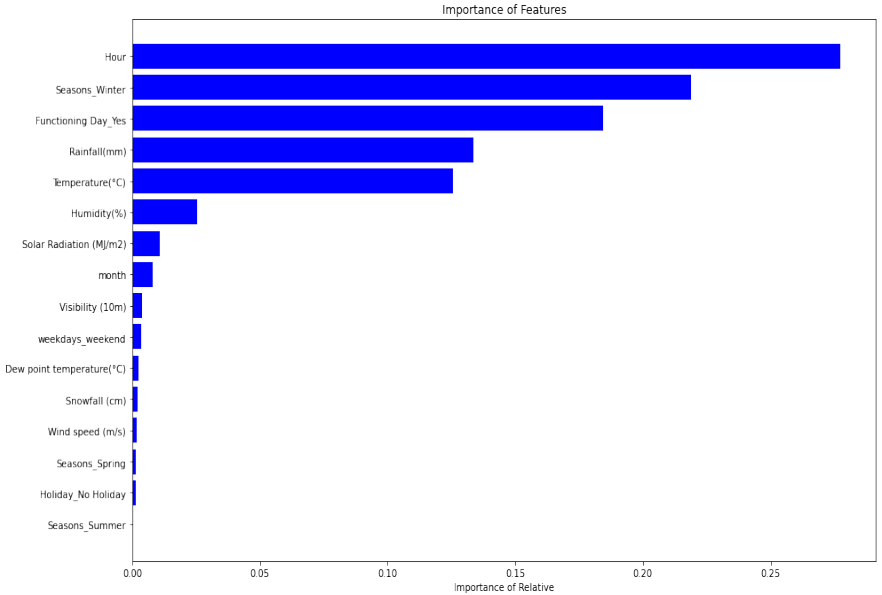
R2 value and adjusted R2 value are significantly moderate so the model is also not performing well in this scenario, let's try Decision Tree...

Modelling the Data

Decision Tree Regression



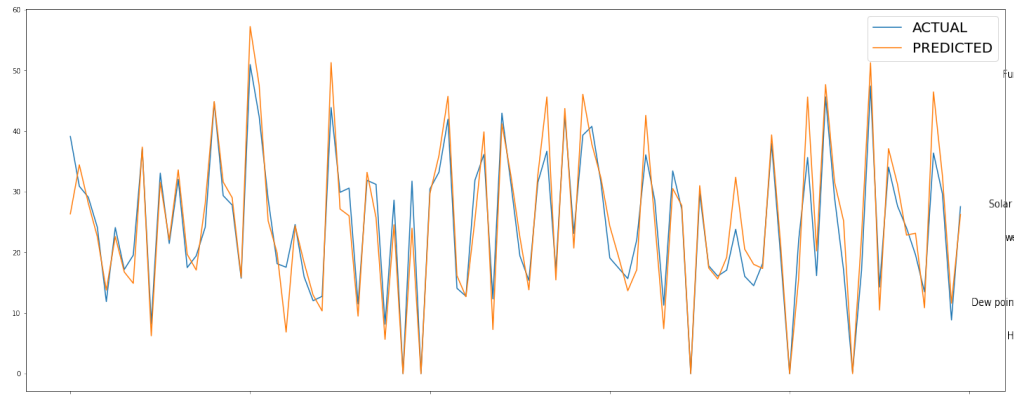
Feature Importances of Decision Tree



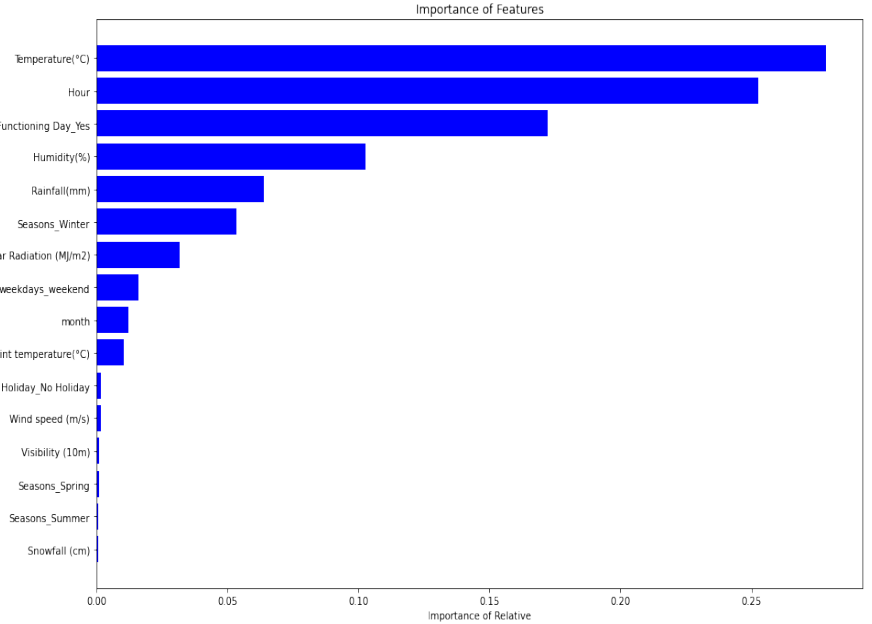
Looks like our r2 score value is 0.76 that means our model is able to capture most of the data variance. Lets save it in a dataframe for later comparisons.
Feat Imp Max - Hour

Modelling the Data

Random Forest



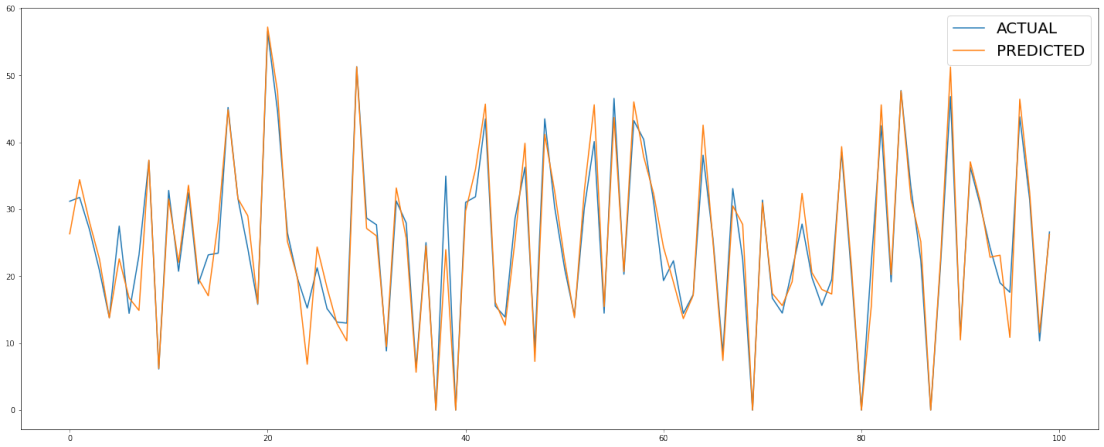
Feature Importances of Random Forest



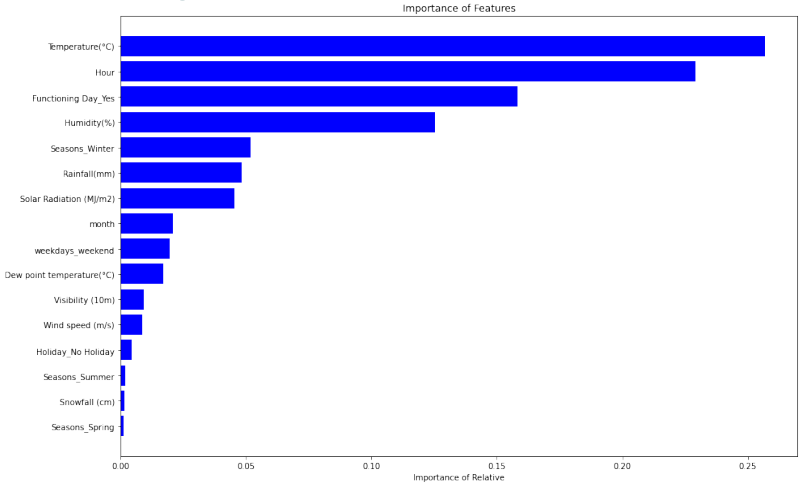
Looks like our r2 score value is 0.87 that means our model is able to capture most of the data variance than Decision Tree Regression . Lets save it in a dataframe for later comparisons, Feature Importance Max – Temperature.

Modelling the Data

Gradient Boosting



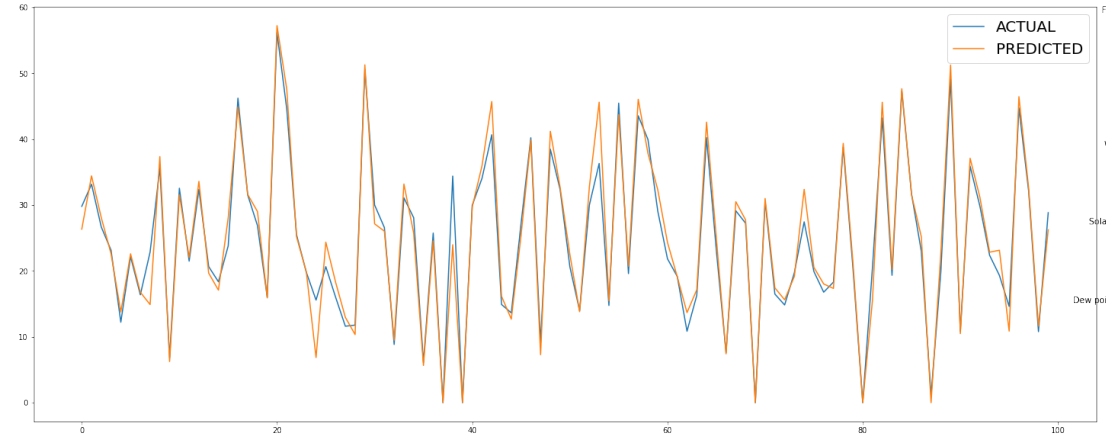
Feature Importances of Gradient Boosting



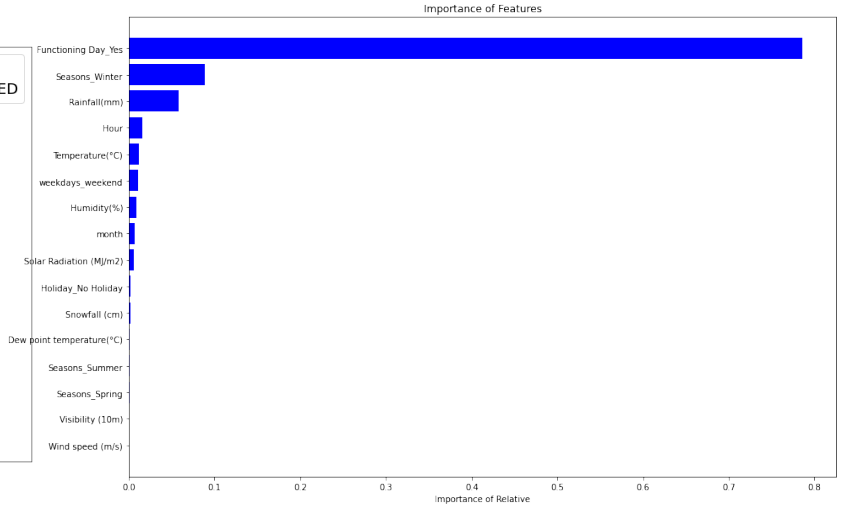
Looks like our r2 score value is 0.91that means our model is working perfectly and accurate. Feature Importance Max - Temperature

Modelling the Data

eXtreme Gradient Boosting

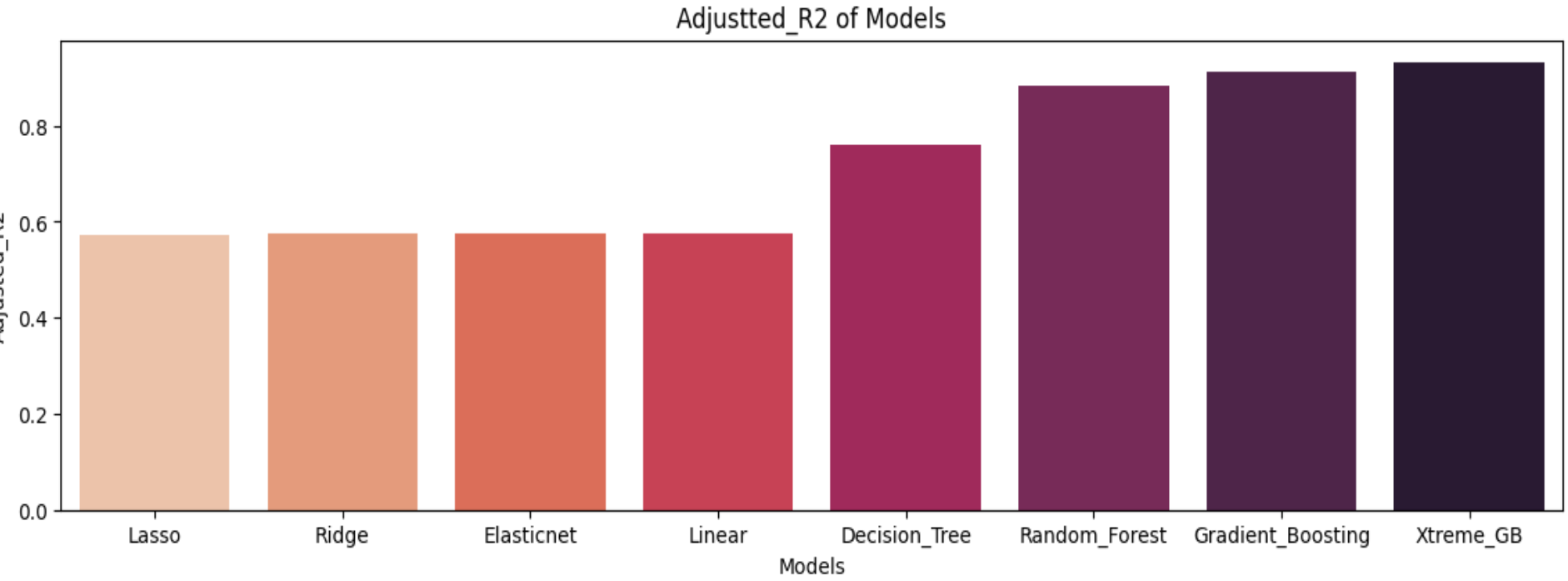


Feature Importances of XGBoost



Looks like our r2 score value is 0.93 that means our model is working perfectly and accurate. Feature Importance Max – Functioning day yes

Combined Evaluation of Matrix of ALL the MODELS



Insight of Evaluation Matrix

- No overfitting is seen.
- Random Forest Regressor, Gradient Boosting and Xtreme GB gives the highest R2 score of 92%, 95% and 99% respectively
- We can deploy this model.
- However, this is not the ultimate end, as the data is time dependent for variables like temperature, windspeed, solar radiation etc., will not always be consistent. Therefore, there will be scenarios where the model might not perform well.

Conclusion

- Data Exploration Conclusions

- In the comparison between "hour vs rented count of bikes" we can clearly notice a high demand in the rush hour of 8:00 am to 9:00 pm.

- In the comparison between "holiday-non holiday vs rented count of bikes" we get the notion of high demand of bikes during non holiday i.e working days compared to holidays i.e non working days

- Demand of Rented Bike gradually decreases with increase in rainfall.

- Same pattern of decrease in demand is observed with the increase in snowfall.

- In the Winter Months the Demand decreases i.e, December, January, February, However demand spikes up in Summer months i.e May, June , July.

Conclusion

- **Modeling Conclusions**

- **We used 8 Regression Models to predict the bike rental count at any hour of the day - 'Linear','Lasso','Ridge','Elasticnet','Decision_Tree','Random_Forest','Gradient_Boosting','Xtreme_GB'.**

- **Using the predictions made by these level 1 individual models as features, we trained 4 level 2 stacking algorithms (Linear Regression, Random Forest Gradient Boost and Xtreme Gradient Boosting) to make more refined predictions.**

- **Of all the models, we found a simple XGBoost Model providing the best/lowest RMSE score and the adjusted_r2 of 99% which made the model deployable.**