Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Individually Contributed By: Akash.K (akash.tup@gmail.com)
Please paste the GitHub Repo link.
Github Link:- https://github.com/AkashKarthikeyan/Bike_Sharing_Prediciton_Capstone_Project
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)
The bike sharing demand is calculated using various variables like 'Date', 'Rented Bike Count', 'Hour', 'Temperature(°C)', 'Humidity(%)', 'Wind speed (m/s)', 'Visibility (10m)', 'Dew point temperature(°C)', 'Solar Radiation (MJ/m2)', 'Rainfall(mm)', 'Snowfall (cm)', 'Seasons', 'Holiday', and Functioning Day. These values are treated for null values at first, then the EDA part is carried out to enhance the normality of the data, to determine the correlation between the variables and finally the multicollinearity is checked between the variables. Each of the variables are defined below: **Date :** year-month-day **Rented Bike count -** Count of bikes rented at each hour **Hour -** Hour of he 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 Hours), Fun(Functional hours)

The problem has been carried out by handling the null values and Exploratory Data Analysis. In these steps I used algorithms like ExtraTree classifier to check the results of each feature i.e which feature is more important compared to our model and which is of less importance. Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment. Then the problem is modelled using different models such as Linear Regression, Decision Tree Regressor, Random Forest Classifier, Gradient Boosting Machine and XGBoost Regressor with Cross validators as GridSearchCV, RandomSearchCV and Bayesian Optimizer.

The best R-Squared value is estimated to be 0.922 in the OLS model and 0.906 in the XGBoost Model.

The most important features are estimated with the help of SHAP library in the XGBoost model with Bayesian Optimisation and those are shown in the below image: