**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Team Name : Team Datavengers  Team Members:   1. Kunal Mahadik   Email id : kunalmahadik0811@gmail.com  Contribution:   1. Data Wrangling 2. Data Preparation 3. Data Cleaning 4. Data Preprocessing 5. Implementation of Linear Regression model 6. Implementation of Lasso and Ridge Regression model 7. Aashruti Agarwal   Email id : aaashruti@gmail.com  Contribution:   1. Data Wrangling 2. Data Preparation 3. Data Cleaning 4. Data Preprocessing 5. Implementation of Decision Tree model 6. Implementation of Random Forest Regressor 7. Raneev K   Email id : raneevk36@gmail.com  Contribution:   1. Data Wrangling 2. Data Preparation 3. Data Cleaning 4. Data Preprocessing 5. Implementation of Gradient Boost Regressor 6. Implementation of Gradient Boost Regressor with Grid Search CV |
| **Please paste the GitHub Repo link.** |
| Github Link:- https://github.com/Raneevk/Seoul-Bike-Sharing-Demand-Prediction |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| 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. In this project we are going to create a best machine learning algorithm to predict the required bike count at each hour.  At first we loaded our dataset and did data some data wrangling & EDA to clean our dataset and to understand the features thoroughly. Thankfully there was no missing and duplicate values were present in the dataset. Some of the feature names were quite lengthy so we renamed some features and checked for any wrong data types. “Date” was given as object datatype so we converted it into “date time” format. “Hour” was given as “int64” datatype but we don’t need it as continuous variable so we converted it into categorical variable.  Then in EDA part first of all we plot the distribution of target variable, unfortunately the distribution was slightly skewed towards right side and there was some outliers also. We did squire root transformation to normalize the distribution. We plot charts of different categorical variables and numerical variables against our target variable and found that in summer season rented bike count is maximum also around 8 am at morning and 6pm at evening there is a peak in rented bike count. We can say that most of the working professionals depends on rented bike for commute. When temperature is extremely low the rented bike count is also less and it follows a linear relationship upto 25-28 degree celcius.  After plotting correlation heatmap we found that “temperature” and “dewpoint temperature” are highly correlated each other, so we merge them together and create a new feature to remove multicollinearity.  After all these data preprocessing we implemented different machine learning algorithms like “linear regression”, ”lasso”, ”ridge”, “elastic net”, ”decision Tree”, “random forest”, “Random forest regression with GridsearchCV” and “Gradient Boosting gridsearchcv”. We found that   1. Almost all algorithms performed really well on both training dataset and   testing dataset so we can say that variance is less and no issues of overfittings are present.   1. Both "Random forest regression" and "Gradient Boosting regression(gridsearch cv) has highest R2 score. 2. Performance on "Decision Tree" algorithm is comparatively less with an R2 score of 68%. 3. For decision tree we got highest feature importance for "temperature\_and\_dewpoint temperature",incase of Random Forest "season\_winter" is the most important feature and for Gradient boost "temperature\_and\_dewpoint\_temperature" has highest feature importance value.   We know that this data is time dependent, the values for variables like temperature, solar\_radiation, wind\_speed etc., will not always be consistent. Therefore, there will be scenarios where the model might not perform well. As Machine learning is an exponentially evolving field, we will have to be prepared for all contingencies and also keep checking our model from time to time. Therefore, having a quality knowledge and keeping pace with the ever evolving ML field would surely help one to stay a step ahead in future |