**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Akshay Kumar Saini (Ak663348@gmail.com):  1.1. Descriptive Analysis  1.1.1. Data frame description  1.1.2. Data frame shape  1.2. Data cleaning, transformation, and Analysis:  1.2.1. Extracting the information from the categorical variable  1.2.2. Rename the column name for better understanding of data set  1.3. Data Wrangling  1.3.1 Extracting the information from date string for further data analysis  1.3.2 Finding the correlation between the various variables of the data set and drop the more correlated column by finding the VIF.  1.4 Data Visualization  1.4.1 Relationship between ‘Bike Booking count’ and ‘year’  1.4.2 Relationship between ‘Bike Booking count’ and ‘months’  1.4.3 Relationship between ‘Bike Booking count’ and ‘season’  1.4.4 Relationship between ‘Bike Booking count’ and ‘weekdays’  1.4.5 Relationship between ‘Bike Booking count’ and ‘weather conditions’  1.4.6 Relationship between ‘Bike Booking count’ and ‘Functioning’  1.4.7 Relationship between ‘Bike Booking count’ and ‘holidays’  1.5 Feature Engineering and Dummy variable  1.5.1 Dropping Temperature and creating new feature respect to dew point  1.5.2 Doing one hot coding for different seasons  1.6 Machine learning Model Analysis  1.6.1 Fitting data on Linear Regression, Lasso, Ridge regression, Random Forest, XGBoost |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/AkshaySaini25/Bike-Sharing-Demand-Predection> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (20)00-400 words)** |
| Bike-Sharing, as a new green public transportation mode, has been developed in several western cities and most of the developing countries are on the path of following the western model of Bike Sharing Systems. In this project, we analyze the data of the Seoul bike sharing system deeply and make the prediction of the bike count required at each hour for the stable supply of rental bikes. Initially we load the data and search for null values and duplicate rows in the dataset. No duplicate and null values are found in our dataset.  After that we did EDA on the dataset. First, we convert the column ‘Date’ (datatype ‘string’) into datetime and then extract all valuable information from that. Thereafter we extract all possible information about the ‘rented bike count’ (dependent variable) with the help of all independent variables. After that, we check the distribution of a dependent variable which is not normally distributed so we apply square root transformation to convert it into normal distribution. After that we plot correlation heatmap and found that temperature, dew point temperature and humidity are highly correlated with each other. To remove this multicollinearity, we find the VIF score of independent variables and the VIF score of each variable comes under 10 when we remove the dew point temperature column. After that we drop all the unnecessary columns like date, year, weekday.  After extracting useful data from the dataset, we did data encoding by converting categorical variables ‘Season’, ’Holiday’ and ‘Functional day’ into numeric variables and made the dataset ready for fitting in various machine learning models.  We fit our data in 5 models namely Linear regression, Lasso regression, Ridge regression, Decision tree regression and XGBboost regression. After fitting data in all the models, we found that XGBboost regression outperforms all the other models with an r-square value of 0. 894637.  We can conclude that most of the bike booking counts are maximum in the month of summer and most of the booking are from the functioning days and weather parameters like rainfall, solar radiation, wind speed, snowfall affects the bike sharing booking count. |