Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
1. Mohammad Jibran jibranmd9@gmail.com 1. **Data wrangling** 2. **EDA** 3. **Visualization** 4. **Decision Tree Regressor Implementation** 5. **Random Forest Regressor Implementation** **2. Siddhi H Thakur** siddhi.thakur04@gmail.com 1. **Data wrangling** 2. **EDA** 3. **Visualization** 4. **Linear Regression with regularisation Implantation** 5. **Polynomial Regression Implantation**
Please paste the GitHub Repo link.
Github Link:- **https://github.com/DSJibran/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)

Summary of Seoul Bike Sharing Demand Prediction

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 become a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

The dataset has two years of record of bike rented 2017 and 2018 and some other information like functioning day, humidity, temperature level, our project is we deal with various weather-related info to make accurate prediction without fail. In the datasets, there are 8760 entries and 14 columns.

We pre-processed and cleaned the data to check null values as well as duplicates etc. After pre-processing we perform EDA to visualise data insight in better ways.

With categorical features we tried to observe the impact of different given factors such as rainfall, snowfall, holidays, weeks, months on our dependent variables.

And with numerical features we observed that some of our columns are right skewed and some are left skewed. We also determined that the skewed columns have skewed mean and median too. We checked the linearity as well between dependent and independent columns to know whether they are positive or negative with the target variable.

We also found the high collinearity between temperature column and dew point and after checking variance inflation factor, we removed the dew point as it is less correlated with target.

At the end we performed the different regression algorithms to determine which one is performing better among them, we used Linear regression with regularization, Polynomial regression, Decision Tree Regressor and Random Forest regressor.

We used evaluation matrix to compare the performance of all these models and found that Random Forest Regressor is the model which is working better with training data as well as testing data followed by Decision tree while Polynomial regression is working better with training data but with testing dataset, which made it poor model among all we used.

THANK YOU