Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Mohd Sharik livisliquoro999@gmail.com 1. EDA, Code and ML algorithm a. Data Wrangling b. Code Comments, Code Quality Check c. Univariate analysis d. Bi-Variate analysis e. Feature engineering f. Model implementation g. Data transformation for algorithm application h. Notebook summarization 2. Documentation a. Presentation b. Project Summary Documentation c. Technical documentation

Please paste the GitHub Repo link.

https://github.com/MohdSharik99/Regression-

Please paste drive link.
https://drive.google.com/drive/folders/1IwK_4hVrid-lNo_b57IxdIORlSVEQUne?usp=sharing

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

About the Dataset:
The dataset that we have chosen is Seoul Bike data and has 8760 and 14 columns.these columns are mixed between numerical and categorical columns**.** In this dataset we will analyze all the possible columns which we can use in the prediction of rented bike counts

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 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.

Approach taken:
Our approach can be explained in these steps-

- Acquire and loading data
- Understanding and summarizing the variables
- Cleaning dataset
- Exploring and Visualizing data
- Analyzing relationships between variables
- Predicting the target column
- Summarizing the whole work

Challenges faced:

- In order to use a parametric algorithm and upon analyzing features we found that deciding the transformation was not so easy as we have noticed various distributions.
- Deciding a non-parametric algorithm was a challenge as we were not getting very good results using some of the algorithms.
- Bi-Variate analysis section was a bit difficult as there was lot of importance and we were supposed get best of it.

Overall Summary: -

in this section we will summarize everything that we have done so far.

section 1 : we simply imported our library and loaded our data in google colab and then we have got overview of the data and finally we have described our data.

Section 2 : we have focused on distributions of each feature in the dataset and found that this data is randomly distributed across different feature

Section 3 : in this section we moved to the next step focused on bi-variate distribution specially between target column and other features

Section 4 : in this section first we have transformed categorical features into numeric features then we used decision tree keeping explain ability in mind however the were not very good so we have chosen a black box model to get better predictions. and this model was xgboost regressor which gave us 89% accurate results. then final task was plot the feature importance of the dataset. overall we have performed analysis on our data and predicted our target column and these predictions are ready to be presented to the senior management who can decide many strategies.