Capstone Project II Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Individual :
Name: Sonika Baheti
Email: bahetisonika50@gmail.com
Contribution:

- Framing problem statements and goals.
- Data understanding and manipulation.
- Exploratory data analysis.
- Preparation of dataset for modelling
- Modelling and evaluation
- Technical document and ppt
- Deriving conclusions.

Please paste the GitHub Repo link.

Sonika Baheti: https://github.com/sonika-07/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)

In the majority of cases, a Data Science project will have to go through five key stages: **defining a problem, data processing, modelling, evaluation and deployment**. In Seoul Bike Sharing Demand Prediction, I tried to gain power along with money. As **Data is Money but information is power**. The dataset contains 8760 rows and 14 columns.

Before applying any machine learning model, it is a good practice to first look at the data and understand it well, so that we can have complete information about the datasets. Applying algorithms without knowing the data can result in wrong conclusions. Hence I first overviewed my dataset and find out that '**Rental Bike Count**' should be response variable while temperature, rainfall, season, snowfall etc are some other variables that can affect rented bikes.

Then did some data wrangling, created new columns (month, weekends) to make analysis deeper and insightful.

Then comes the indispensable part of data analysis, i.e., EDA, **Exploratory Data Analysis**. Here I found certain relationship of variables among each other and dependent variable and bought out some insightful information out of it. For example, Rented bikes are less used in winters as compared to summers and Functioning day is another important factor of bikes on rent, checked correlations etc.

After EDA, find out some columns containing outliers. Then did **one hot encoding** for categorical columns to make it predictable. And checked for Multicollinearity for applying linear models.

Then transformed variables to **standard scalar** (also called z- score transformation) to make analysis more optimized.

Splitting the dataset to train and test with 0.25 as test data. Then first applied linear models (Multiple linear, Ridge, Lasso, Elastic Net, Polynomial) and then Tree based models(Decision Tree, Random Forest and XGB).

Evaluated Model using **adjusted R2** value and **RMSE** and found that adjusted R2 score for XGB is maximum while RMSE is minimum in XGB.

But tree based models are black box models, Hence used model explainability to know how an individual value is being predicted by these models.

So concluding that **XGB is a good fit and best option among other models to deploy.**