Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

❖ **Kirtesh Verma** Email : kirteshverma12345@gmail.com

Contribution:

- Data understanding
- Handling null or missing values
- Performing EDA
- Removing outliers
- Linear Regression Model
- Random Forest Model
- Hyperparameter Tuning on random forest

❖ **Pravin Bejjo** Email: praveen.bejo.pb@gmail.com

Contribution:

- Data understanding
- Data visualization
- Multivariate analysis
- Ridge Regression Model
- Gradient Boosting Model
- Hyperparameter Tuning on Gradient Boosting

❖ **Sahil Pardeshi** Email: 8623879021.sp@gmail.com

Contribution:

- Data understanding
- Data visualization
- Multivariate analysis
- Lasso Regression Model
- Decision Tree Model
- Hyperparameter Tuning Decision Tree

Please paste the GitHub Repo link

GitHub Link: https://github.com/KirteshVerma/Bike-Sharing-Demand

Drive Link: - https://drive.google.com/drive/folders/1iCHbZErKSHBxfDf9hPiGLgKfYDSprWqo

Seoul Bike Sharing Demand

PROJECT SUMMARY:

Bike sharing system is a means of renting bicycles, where the process of obtaining membership, rental, and bike return is automated. Bike sharing system has been booming recently as it includes transport flexibility, reductions to vehicle emission, health benefits and most importantly financial savings for individuals. The given dataset is of Seoul city, consisting data of rented bike counts of the year 2017 and 2018. It has 8760 entries and 14 columns. Seasons, Holiday, Functioning day and Date are four categorical columns present in the dataset. Other columns like rainfall, snowfall, humidity, dew point temperature is of numeric type.

Our objective is to predict the count of bikes required at each hour throughout the day so that bikes are accessible to the public at right time and lessening the waiting time. For this we first performed the Exploratory Data analysis so to understand the important features that governs the rented bike demand. We looked at the distribution of different features and understood the relationship of all the features with our dependent feature i.e., 'Rented Bike Count'.

The second step included the data preprocessing. There were no null and duplicated values in the given dataset. However, there were outliers. We treated the outliers by removing them and filling the nan values with corresponding column, along with this we also did the label encoding for the categorical columns and also extracted the year, month and day of the week from the Date column for a better insight.

In the third step we tried different machine learning model on the clean and standardized dataset. We used different linear and tree-based models namely; linear regression, Lasso Regression, Decision Tree and Random Forest. We did the hyperparameters tunning of the features and evaluated the performance of the model through various metrics. The best performance was given by Random Forest where the r2 score for the training and test set was 86 and 83 respectively. We calculated the feature importance of the best performed model. The feature that came out to be most important were 'Hour' followed by 'Temperature'.

We did a thorough analysis of the data and understood the features and factors that impact the bike sharing demand. Some of the concluded points are most bikes are rented during summer season, most bikes are rented during the office time, on holiday's, nonfunctioning day, and weekends least bikes are rented. Our model performed well in this case but as the features that impacts the bike sharing demand are time dependent like temperature, windspeed, climate etc. will not always be consistent so for that we have to keep updating our model from time to time and keep checking the performance by including or updating the values of various columns.