Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1). **Ankit Patil**
E-mail: ankit.patil67@gmail.com

- Data Cleaning
- Data Analysis
- Error Handling
- One hot encoding
- Feature Engineering
- Normalization
- Correlation Analysis
- Tree Based Model Selection
- Model Deployment
- Hyperparameteric Tuning
- Shap Summary
- PPT presentation
- Technical Documentation

2). **Chukkapalli Naga Sai**
E-mail: nagasaichowdary1111@gmail.com

- Debugging Error
- Data Sorting
- Technical Documentation
- ppt Presentation
- Approach Towards Plan
- Seaborn,matplotlib
- Heatmap
- Linear Model Selection
- Evaluation Matrix

3). **Shreyash Movale**
E-mail: shreyash9m@gmail.com

- Data Sorting
- Matplotlib
- ppt Presentation
- Data Visualization
- Technical Documentation
- Approach toward Plan
- Line Plot,Barplot,Histogram
- Heatmap,VIF factor
- Linear Model Selection
- Evaluation Matrix
- Data Preparation

4). **Saugata Deb**
E-mail: saugatad56@gmail.com

- Framework establishment
- Line Plot

- Data Manipulation
- Data Preprocessing
- Feature Engineering
- Correlation Analysis
- Tree Based Model Selection
- Model Deployment
- Feature Importance
- SHapley Additive exPlanations
- Shap Summary
- PPT presentation
- Technical Documentation

Please paste the GitHub Repo link.

Github Link:- https://github.com/Ankit-Patil1/Bike-Sharing-Prediction

Please paste the Google Drive link

Drive Link: https://drive.google.com/drive/folders/1YZQRt95SSYDw1HfcrpPKQl8TchjpoZTQ?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)

We have dataset of number of rented bikes in Seoul city segregated by hours in each day for 365 days of the year. We checked for duplicate, null values, did exploratory data analysis of the dataset. In exploratory data analysis, we found that in the month of June the rented bike count was the highest and in the month of January it was least i.e. highest in summer season and lowest in winter season. We found that on weekdays count of rented bikes is high between 7-8 A.M and it is the highest between 5-6 P.M. We plotted rented bikes vs numerical variables, found out that, dependent feature is linearly dependent on features.

The data had multicollinearity, we eliminated it by taking reference from Variance Inflation Factor. The dependent variable was skewed, we reduced the skewness by transformation. Converted categorical variables into numerical data by one hot encoding.

After that we split the data into train and test set. We used MinMax scalar on independent variables since some of the features had high values and some had small values.

We applied Linear regression, we got **R2** score of **0.779** and **0.774** on train and test data respectively, it meant that model is moderately fit for train as well as test data and **no overfitting** is seen. Therefore, we didn't go for Regularized Linear Regression.

After that, we applied polynomial Regression with **degree=2**, we got much better results of **0.93** and **0.90** on train and test data respectively. it meant that model is fit well for train as well as test data and no overfitting is seen. **MAE** ad **MSE** was also noticeably lower for polynomial regression

Similarly, to check if we can get even better accuracy, we applied Decision Tree Regressor with **max depth as 10**, we got **r2** score of **0.83** and **0.80** for train as well as test data.

After We applied Random Forest Regressor with certain parameters. It performed better than Decision Tree but not as good as polynomial regression.

Finally, we applied**, Gradient Boost** with **Grid search CV**. After putting in best parameters form GridsearchCV, we got r2 score of **0.96** and **0.93** for train and test data respectively with lowest **MAE** and **MSE**.

We would recommend to use **Gradient boost** for prediction of bikes, but since it needs higher computational time**, polynomial regression** can also be used since it is faster with slightly lower accuracy.