Capstone Project Submission

Team Member's Name, Email and Contribution:

Nivya.T - nivyathiruvoth@gmail.com
Manjusree.K.C - manjusreemarar@gmail.com
Ammar aNjum - ammararq8@gmail.com
Harsh Jain - harshjain15998@gmail.com

Contributor Roles:-

1. **Nivya.T**
    - Exploratory Data Analysis
        a. EDA on Numerical Features
        b. EDA on Categorical Features
    - Random Forest Regression
    - Random Forest Regression: Cross Validation and Hyperparameter Tuning

2. **Manjusree.K.C**
    - Exploratory Data Analysis
        a. EDA on Numerical Features
        b. EDA on Categorical Features
    - GBM Regression
    - GBM: Cross Validation and Hyperparameter Tuning

3. **Ammar Anjum**
    - Exploratory Data Analysis
        a. EDA on Numerical Features
        b. EDA on Categorical Features
    - XGBoost Regression
    - XGBoost Regression: Cross Validation and Hyperparameter Tuning

4. **Harsh Jain**
    - Exploratory Data Analysis
        a. EDA on Numerical Features
        b. EDA on Categorical Features
    - Linear Regression
    - Regularization: Lasso and Ridge with Cross validation and Hyperparameter Tuning

Please paste the GitHub Repo link.

Github Link:- https://github.com/nivyathiruvoth/Seoul-Bike-Sharing-Demand-Prediction

PROJECT SUMMARY

Rental bikes are currently being introduced in many cities to enhance mobility. It is essential to make the rental bike available to the public on time to reduce waiting. Ultimately, providing a stable supply of rental bikes to the city becomes a primary concern. The crucial part is predicting the number of bikes needed per hour for a steady supply of rental bikes.

The inevitable part of any modeling technique is making the dataset model ready. Problem statement comprehension, target variable identification, and determining the type of ML problem play a crucial role in Machine learning. Since our target variable, 'Rented bike count' is continuous, we should apply regression models, and we aim to predict the number of bikes that will rent at any given hour in the future.

We started with basic inspections. The dataset provided contains 8760 bike-rented details. None of them are duplicates or null values, which made our task easier. The dependent variable's distribution was skewed, so we applied log(1+x) transformation to make it follow a normal distribution.

As a part of exploratory data analysis, we examined the distribution of each independent variable and its relationship with the dependent variable. Heat map and VIF analysis approach helped to check multicollinearity, based on which features are selected. We used boxplot to identify outliers. We handled categorical variables using label encoding.

We scaled the data for better results using the standard scaler as a final data preparation stage. We have also applied z-score to the features. Then we divided the dataset between training and testing. We used 80% of the data for training and 20% for testing the model.

MAE, MSE, RMSE, R2-score, and adjusted R2-score are the evaluation metrics used to measure model performance. We created modular functions to compare the evaluation metric scores and determine the significance of the features in each model.

For linear, lasso, and ridge regression, the R2-scores were poor, indicating that these models lack accuracy. So we applied random forest, GBM, and XGBoost. We used cross-validation and hyperparameter tuning to generalize the results and avoid overfitting.

Among the models with the least MAE and highest R2_score, XGBoost is the best model. The most important features for predicting the dependent variable are the functioning day, rainfall, seasons, and temperature. This project will be helpful for the company to predict the hourly bike demand and enrich the user experience.