Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email, and Contribution:
• Nidhi Pandey (np5603817@gmail.com) ◦ Exploratory Data Analysis ◦ Data Cleaning ◦ Discussing insight from EDA ◦ Selected Plotting graph ◦ Performed various models ▪ Linear Regression (regularization) ▪ KNN Regression ▪ Decision Tree ▪ Random Forest ▪ CatBoost ▪ lightGBM ◦ Creating Functions for all model's performed ◦ Model's performed evaluation ◦ Model Explainability
Please paste the GitHub Repo link.
Github Link:- https://github.com/treasure823/Bike_Sharing_Demand_Prediction_Capstone_Project
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches, and your conclusions. (200-400 words)
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. In this study, we train a model to forecast the availability of bikes at any hour of the year based on the weather. The data collection, which included historical bike usage patterns and weather data spanning two years, was collected from the Capital Bikeshare programme in Washington, D.C. The data set is first subjected to exploratory data analysis. We search for missing data values (none were identified) and outliers, then change them as necessary. Additionally, we use correlation analysis to isolate the most crucial and pertinent feature set. Later, we use feature engineering to change a few already-existing columns and eliminate irrelevant ones. Then, we examine a number of well-known individual models, ranging from straightforward ones like Linear Regressor and Regularization Models (Ridge and Lasso) to more intricate ensemble models, such as Random Forest, Gradient Boost, and Catboost. A single unified model for working and non-working days, among other choices for model development, were tested. 2. There are two distinct models for workdays

and non-workdays. Utilizing the provided categorical features, 3. using OneHotEncoding to obtain binary vector representations of the categorical features, and 4. To further improve the predicting skills, we also explored stacking algorithms (Linear Regressor, Random Forest, and Gradient Boost) where the predictions from the level 1 individual models were incorporated as meta-features into a second level model. A portion of the provided training data set (the first 14 days of each month) was used to tweak the hyperparameters using GridSearchCV cross validation using 5 folds. To evaluate the effectiveness of our model, we examined the remaining data (15th to 19th of each month). With train and test scores of 0.36 and 0.427, respectively, we identified the Random Forest Ensemble approach employing a Single Model and Categorical Feature set to be the best option out of all the methods and models. The total train+test observation size is 10871, and the training and testing times for the selected model are pretty reasonable (23 seconds).