**Capstone Project-2 Submission**

|  |
| --- |
| * **Team Member’s Name, Email and Contribution:** |
| * **Name**: - Aniket Abhit Jadhav: * **Email ID:-** aniketjadhavv28@gmail.com |
| * **GitHub Repo link:** |
| * Link : <https://github.com/Aniketj777/Bike_Sharing_Demand_Prediction> |
| * **In this project we word have Seoul Bike Sharing Datset.** * ***In this dataset we have columns such as:***   (Date, Rented Bike Count, Hour, Temperature(c), Humidity(%), High speed(m/s), Visibility (10m), Dew point temperature(c), Solar Radiation(MJ/m2), Rainfall(mm), Snowfall(cm), Seasons, Holiday, Functioning Day) |
| * **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| * Bike sharing systems are a means of renting bikes where the process of obtaining memberships, rental, and bike return is automated throughout a city. Using these systems, People are able to rent a bike from one location and return it to different place. * The dataset “Seoul bike data” contained “Rented bike count” column which is our dependent variable and others columns are independent columns. Our aim is to predict the count of rented bike per hour using different machine learning techniques. * Going through the dataset, some preprocessing required before using that data like filling/removing the missing value but looking at the data there is no missing/null values present in the data. * Further step is EDA, while changing the datatype we observe that the hour column is a numerical column it is a timestamp so we have to treat hour as a categorical feature if we don’t change then, while doing further analysis we didn’t get the accurate outcomes. * Multicollinearity, our columns temperature and dev point temperature are having high correlation so we need to remove one of those columns then it will not affect the outcome. * Model training is done, we split the dataset into training set and testing dataset to improve the accuracy of the model. * We define evaluation different matrices like MSE, RMSE, MASE, R-Squared, Adjusted R-Squared, Best-params and Best-score for evaluation of different models. * Linear Regression is used as a base model here despite knowing that linear regression wouldn’t work here. We fit other models as a reference to check how good the other models are performing. * Regularization is a technique to get the best and optimal results and avoid overfitting using lasso and ridge regressor. Then applied multiple Linear regression by which we get better results. * Other regressors like Decision Tree are used because the multicollinearity doesn’t affect the tree-based models. Random Forest Regressors models are fit for getting the best results using hyper parameter techniques. * Random forest works best in predicting the count of rented bike as its R2 score is maximum from the tried model. * Some of the hyperparameter techniques used to improve the model performance are Gradient Boosting Regressor, and Gradient Boost Regressor giving the best results. |