Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Name: Anish Johnson. **Email: anishjohnson05@gmail.com.** **Contribution: Individual project.**
Please paste the GitHub Repo link.
GitHub Link: - https://github.com/anishjohnson/NYC_Taxi_Trip_Time_Prediction
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches, and your conclusions. (200-400 words)
A Taxi firm often faces a common problem of efficiently assigning the cabs to passengers so that the service is smooth and hassle free. Our objective is to build a model that predicts the total ride duration of taxi trips in New York City. For this we were given a primary dataset released by the NYC Taxi and Limousine Commission, which included pickup time, geo-coordinates, number of passengers, and several other variables. We began by checking if the data contained any null or duplicate values which might cause problems along the way. The next step was to start EDA (Exploratory Data Analysis) to understand the relationship between the dependent and the independent variables better and identifying the necessary trends in them to predict the trip durations. Using the information obtained from the EDA we were able to detect and remove the outliers from the data and perform the required feature engineering. The data was then split into two groups the train, and the test sets. The train set contained 80% of the data and test set contained 20% of the data. This data was transformed using the standard scaler before fitting it into the models. Once the data was prepared, we trained and tested it on the selected algorithms which included **Linear Regression, Gradient Boosting, XG Boost, and Hist Gradient Boosting Regressor.** Out of the above-mentioned algorithms **Hist Gradient Boosting** provided the least error (Root Mean Squared Error) **3.726250** and the highest R2 score **0.686527**. We continued with this model for the hyperparameter tuning and observed a reduced error of **3.367436** and increased R2 score of **0.743991**. Hence, we can proceed with the Hist Gradient Boosting Regressor for future predictions of trip duration of NYC taxis.