**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Shashwat Kulkarni  Mail id: shashkul302@gmail.com  Activities performed:   * Preview Data * Check total number of entries and column types * Check the null values * Plot distribution of numeric data * Plot distribution of categorical data * Remove the outliers * Correlation through heatmap * Building the model * Linear regression * XgBOOST * Decision tree * Gradient boosting * Model interpretation * Conclusion |
| **Please paste the GitHub Repo link.** |
| Shashwat Kulkarni Github Link: https://github.com/Shashwat-23/NYC\_taxi-trip-duration-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 company faces a common problem of efficiently assigning the cabs to passengers so that the service is smooth and hassle free. So, the topic for the supervised machine learning capstone project is NYC taxi-trip time prediction in this project our target variable is trip time prediction so our goal to predict when the cab will be free for the next trip.**  **The initial step in our project is to prepare the dataset for our machine learning models. Upon loading the dataset, I started on Data Cleaning, which includes checking for Nan values and duplicated values. Subsequently, then conducted an Exploratory Data Analysis where we compared our target variable, the trip duration, with other independent variables. This comparison was visualized using the seaborn and matplotlib libraries, aiding us in understanding various aspects and relationships between the target and independent variables. Further steps will include discarding unnecessary columns and performing one-hot encoding on the necessary columns.**  **After data handling and performing EDA on it we get the important feature for our machine learning model then we fit our Machine learning models like Linear regression,LightGBM the data. After applying the ML Model, we determine the key feature of the data set and perform cross-validation and hyperparameter tuning so as to find out the optimal parameter at which the error would be less for the training and testing dataset and the model performance is high and with the help of ML Evaluation metrices like R2 score, RMSE, MSE, Adjusted R2 score we decide that which machine learning model is the best fit for our dataset. We are mostly concerned with the information of pick-up latitude and longitude and drop off latitude and longitude, to get the distance of the trip.**  **After applying various algorithm, it is found that LightGBM perform the best in predict the trip duration for a particular taxi.** |