Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
NAME :- DSN NAVEEN KUMAR MAIL ID :- dsnnaveen8@gmail.com Contribution • Preview of Data • Check total number of entries and column types • Data Cleaning • EDA • Feature engineering and Feature Selection • Distribution check for dependent and independent features numeric data • Categorical values – One hot encoding • Plot distribution of categorical data • Plot distribution of numeric data • Outlier detection and elimination • Correlation • Building and evaluating the models Linear Regression XgBoost Light GBM • Conclusion
Please paste the GitHub Repo link.
Github Link:- https://github.com/dsnnaveen/NYC-Taxi-Trip-Time-Preciction
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. One of main issue is determining the duration of the current trip so it can predict when the cab will be free for the next trip. MY first step is to prepare dataset for models to learn. After loading the data set performed Exploratory Data Analysis by comparing target variable that is trip duration with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. I done certain steps like dropping unnecessary columns and do the one hot encoding for the required columns. Once after exploring the data, started to check out the null values present in the each column of the dataset. After which I went for the visualization part to get the insights of each variable. After data handling I fit the Machine learning models like Linear regression, Xgboost, Lightgbm to the data .I evaluate the ML metrices like r2 score, RMSE I decide that which machine learning model is the best fit for the dataset. I mostly concerned

with the information of pickup latitude and longitude and drop off latitude and longitude, to get the distance of the trip. LightGBM will be the best model to predict the trip duration for a particular taxi.