Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

NAME: Vivek Kumar Soni
EMAIL ID: vivekkumar75251@gmail.com

Contribution

- 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.

Vivek's Github Link: https://github.com/vivek7525/NYC_Taxi_Trip_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.

Our first task is to prepare dataset for our machine learning models. After loading the dataset, we started with the Data Cleaning (it involves Nan value Checking and Duplicated value checking) after this we performed Exploratory Data Analysis by comparing our target variable that is trip duration with other independent variables and visualized it using seaborn and matlibplot library. This process helped us figuring out various aspects and relationships among the target and the independent variables. We will do certain steps like dropping unnecessary columns and do the one hot encoding for the required 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, XGBOOST, 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.