Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Name : Bhaskar Subanji
Email : bysubanji@gmail.com

- **Data understanding**
  Data Analysis
  Data pre processing and exploration

- **Feature Analysis**
  Univariate
  Bivariate and multivariate analysis

- **Feature engineering**
  Null value check
  Feature creation , etc

- **Data visualisation**

- **Algorithm implementation**
  Linear Regression
  Lasso Regression
  Ridge Regression
  Decision tree Regressor
  XGBoost Regressoin

- **Research Analytics**
  Technical documents

Please paste the GitHub Repo link.

Github Link:- https://github.com/bysubanji/Nyc_taxi_ride_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)

PROBLEM

Task is to build a model that predicts the total ride duration of taxi trips in New York City. Your primary dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables.

APPROACH

- In first step, imported the data set to carry out the descriptive analysis over the data set to understand the information of data available.
- Checked for missing and repetition of values in the data set provided.
- Exploring all the variables of the data set (such as Vendor-Id, Pickup_point, Dropoff_point,Distance, Speed, Storage_frwd, etc) with respect to trip duration, to determine the factors and understand factors of trip duration.
- Used data visualization with different kinds of plots to explore the correlation with Trip duration and different variables.
- Encoding of categorical columns and fitting the different models, we used below algorithms :-
  - Lasso Regression
  - Ridge Regression
  - Decision tree Regressor
  - XGBoost Regressor
- Then tuning into Hyperparameters and perfomance evaluation to identify best fit Model.

CONCLUSION

- There is not much distinction between Decision Tree and XGBoost Regressor during training and testing time as we can detect it on that MSE and RMSE which are the metrics used to evaluate the performance of regression models and R^2 is about the identical during training and Testing time.
- The performance of the linear model is low or not good compared to Xgbooster models.
- To predict the trip duration for a particular taxi, from the above table we can conclude that XGBoost Regressor is the most suitable model as compared to the other models.