Capstone Project Summary

Team Member's Name, Email and Contribution:

1. Tito Varghese

   Email - tito.varghese1992@gmail.com

   - Data Preprocessing
   - Data Cleaning
   - Exploratory Data Analysis
   - Feature Engineering
   - Feature Selection

2. Lakshmi Keerthana

   Email - keerthana826@gmail.com

   - Model Building
   - Comparing Different Models
   - Hyper Parameter Tuning
   - Evaluation Metrics
   - Conclusion

Please paste the GitHub Repo link.

GitHub Profile Link: - https://github.com/Keerthana826

GitHub Repository Link: - https://github.com/Keerthana826/EDA-Capstone.git

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

NYC Taxi Trip Prediction
Summary

Taxi cabs serve as a quick and easy means of transportation across the New York City. The downside with having an abundance of cabs in the city results in traffic. We shall predict the trip duration of the taxi for both users and drivers and make it easier to understand the trip duration, pricing range and route to prefer.

We will first start by preparing the dataset for our machine learning models. After loading the dataset, we then get to Data Cleaning, we perform the tasks such as handling missing values, handling duplicate values, handling outliers and handling Irrelevant records.

Then we started with the EDA, where by visualizing using seaborn and matplotlib, we have checked the No of taxi trip passengers, days where taxi trip demand is high, when there are busy hours and the months where there is a highest and lowest number of trips.

We then get to the process of Feature Engineering where we have derived new features from independent feature like trip distance, speed and trip direction. We later used log transform to handle skewness on the target variable in features distribution, we then checked for linearity in independent and target feature and later did the process of feature encoding where we used hot encoding techniques. Method of feature selection was done to reduce the input variable to the model by using relevant data.

We then build a machine learning model as we split the dataset into train and test data. The regression models that we have used were linear regression model, lasso and ridge regularized regression models, decision tree model, random forest, XGBoost

Out of all the five models, we have got the highest accuracy in Xgboost Model RMSE value 0.28 and r2 score of 0.84 in train and 0.83 in test data since we have had a good relation between target variable (trip_duration) and our independent features (trip_distance and trip_speed). Our second-best model is Random Forest based with a r2 score of 0.7485 accuracy in training and 0.7475 accuracy in test data. The RMSE score of random forest is 0.350

Based on negative mean square error metrics, we can say that ridge and lasso model was the best model and second best model is XGBoost followed by Random Forest Model giving third least negative mean square error. The Decision Tree Model has given the highest negative mean square error and hence underperformed model compared to other regression models. We perform hyper parameter tuning so as to tune the parameters in a model to optimize the model and give best accuracy score by selecting the best parameter using cross validation. By Evaluation Metrics like MAE, MSE, RMSE, R-square we have decided which machine learning model is the best fit for our dataset.