Supervised ML Regression
Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Name: Gauri Agarwal
Email: gauriagarwal4444@gmail.com

Contributions:

- Data Preparation and Cleaning
  1. Read the data from csv file and parsed the datetime columns.
  2. Data description
  3. Data Cleaning
  4. Outliers removal through quantile method

- Feature Engineering
  1. Feature Engineering on datetime column
  2. One hot encoding of categorical variables

- Exploratory Data Analysis and visualizations
  1. Heatmap for analyzing correlations.
  2. Bar plot for categorical feature
  3. Scatter plot of total_distance v/s trip_duration with different categorical attributes as hue

- Supervised Machine learning
  1. Linear Regression
  2. Ridge and Lasso regression
  3. Gradient Boosting regressor
  4. Model validation and hyper parameter tuning

Name: Saurav Kumar
Email: bhadanisaurav7@gmail.com

Contributions:

- Data Preparation and Cleaning
  1. Data description
  2. Outliers removal through z-score method

3. Removal of outliers from speed column

- Feature Engineering
    1. Feature Engineering to find total distance

- Exploratory Data Analysis and visualizations
    1. Distribution of dependent variable
    2. Distplot for distribution of other numeric columns
    3. Bar plot showing relationship between categorical variables and trip_duration
    4. Boxplots for various categorical attributes.

- Supervised Machine learning
    1. Model estimator
    2. Decision tree
    3. XGboost regressor
    4. Model validation and hyper parameter tuning

Please paste the GitHub Repo link.

Github Link:- https://github.com/gauriagarwal18/NYC-Taxi-Trip-Time-Prediction
Drive Link:
https://drive.google.com/drive/folders/1_X9DH8Eu9uNab6_1dX8fzkz20dl7YtdG?usp=sharing

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

We have data of nyc taxi trip for six months of year of 2016, it was cleaned and there were no null values, data have many outliers but we have removed them, we note that there are information about two vendor ids through which booking is done among which vendor 2 is more preferred for booking for long duration trips. Then we have pickup and dropoff latitude and longitude which shows the trip's initial and final location, then we have store and fwd flag, it is a biased attribute and mostly taxis do not have store and fwd flag, we also have the date for starting and ending of trip, and one dependent attribute trip duration which represents duration of the trip which we have to predict.

Then after cleaning the data, we focus on our final task which is to analyze what features leads to less trip duration and what type of taxi a person would prefer, also we have to predict the trip duration.

First part of our project is feature engineering in which we observed various columns and changed them as our requirement, we dropped the columns which are not required, and we divided datetime column in different attributes like weekday, is_weekend, pickup shift, month and many more. We calculated distance between initial and final geographic positions through haversine formula, we deleted the rows in which the number of passengers are zero and similarly added and removed some other important features as well.

Then next part is analysis in which we observed trends, visualize all the new and already existing columns and note some useful trends like a taxi with store and fwd flag is preferred for long durations, there are very less bookings during midnight time, the duration of trip mainly depends on the distance and then it also depends on starting location, also from scatter plots we note important factors which affects trip duration, like trip duration is more on non-weekdays and many other important conclusions and then we used these conclusion for improving our model to predict duration also it helps the company to provide better and required facilities to its customers.

Then we finally reached to our final task that is to predict the trip duration here we tried some regression models and then finally chose a grey box model gradient boosting as it was better in terms of various evaluation matrices, we get an r2 score of 0.79 for our test data and 0.82 for our train data, and then used this model for our predictions. We have selected a grey box model because it was giving better results also our task is to predict duration only and not very interested in how a trip reached that duration, as for the knowledge of factors which increases the duration we have already performed analysis and found correlations and drawn heatmaps to know that.

So we finally concluded our project with some useful analysis such as which vendor is preferred by peoples, on which day of a week the duration of taxi trip is more, when we have a higher chance of getting a booking and which time of a day and on which day we are receiving less bookings.

We have performed all these analysis and prediction because it will help passengers to know which taxi to select and when they need to have an advance booking, when the taxi is easily available, which vendor to choose for taxi booking and finally they can know the range of duration in which they can reach to their destination.
Also taxi companies can know in what all fields they are lagging behind and what are the factors which affects trip duration, like on which weekday taxis usually takes more time, they can also predict when they can receive more bookings, and by predicting trip duration they can predict when a taxi is available and when they can take advance bookings, they can promise their customers a maximum time in which they can reach their destination, so these are all the reasons for doing all these analysis and predictions on the nyc taxi data.