Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1 Aamir khan

Email:ak41552010@gmail.com

- EDA
  A. Dropping duplicates
  B. Handling null and missing values
  C. Handling Outliers with specific range value

- Feature Creation
  A. Extracted pickup day from pickupdatetime column
  B. Extracted drop off day from pickup datetime column

- Univariate Analysis
  A. Trip duration using log transformation
  B. Pick up hours
  C. Passenger count
  D. Trips per day

- Bivariate Analysis
  A. Trip duration per vendor
  B. Trip duration per weekday
  C. Distance per weekday
  D. Distance and trip duration

- Feature encoding/ Feature Scaling
  A. One hot encoding
  B. Drop unwanted columns
  C. Applying Z score on independent variables

- Correlation analysis between independent/dependent variables

- Splitting Appropriate dependent and independent features

- Train test split on dependent and independent features

- ML regression algorithms used
  A. Recurssive Feature Elimination (RFE) with hyper parameter tuning
  B. Lasso Regression with hyper parameter tuning
  C. Ridge Regression with hyper parameter tuning
  D. Decision Trees with hyper parameter tuning
  E. XGBoost with hyper parameter tuning

2. Saurabh Daund
Email:sudaund@mitaoe.ac.in

- EDA
  A. Dropping duplicates
  B. Handling Null/nan values
  C. Handling Outliers with IQR

- Feature Creation
  A. Extracted pickup hour from pickup datetime column
  B. Extracted drop hour from pickup datetime column

- Univariate analysis
  A. For loop on all numeric columns
  B. Log transformation on trip duration
  C. Trip counts using trip duration in slabs
  D. Trip count per hour

- Bivariate Analysis
  A. Trip duration per hour
  B. Distance by per vendor
  C. Distance according to hours
  D. Distance and trip duration

- Feature encoding/ Feature Scaling
  A. One hot encoding
  B. Drop unwanted columns
  C. Standardization using Standard Scaler

- Correlation analysis between independent/dependent variables

- Splitting Appropriate dependent and independent features

- Train test split on dependent and independent features

- ML regression algorithms used
  A. Linear Regression with hyper parameter tuning
  B. Lasso regression with hyper parameter tuning
  C. Ridge regression with hyper parameter tuning
  D. Decision trees with hyper parameter tuning
  E. Gradient boost with hyper parameter tuning

3. Mouleena Jaiswal

Email:mouli14112000@gmail.com

- EDA

  A. Dropping duplicates
  B. Handling Null/nan values
  C. Handling Outliers with mean values

- Feature Creation
  A. Extracted pickup month for EDA analysis
  B. Calculated distance using longitude and latitude

- Univariate Analysis

  A. Trip duration using log transformation
  B. Passenger count
  C. Store and forward flag analysis
  D. Trips per month

- Bivariate Analysis

  A. Trip duration per vendor
  B. Trip duration per store and forward flag
  C. Trip duration per month
  D. Distance covered per month

- Feature encoding/ Feature Scaling

  A. One hot encoding
  B. Drop unwanted columns
  C. Applying Z score on independent variables

- Correlation analysis between independent/dependent variables

- Splitting dependent and independent features

- Train test split on dependent and independent features

- ML regression algorithms used

  A. Recursive feature elimination with hyper parameter tuning
  B. Lasso regression with hyper parameter tuning
  C. Ridge regression with hyper parameter tuning
  D. Decision tree regression using hyper parameter tuning

4. Het Kothari
Email:het.k123@gmail.com

- EDA

  A. Dropping Duplicates
  B. Handling Null values
  C. Handling nan values
  D. handling outliers and dropping the outliers

- Feature Creation

  A. Calculated distance using longitude and latitude
  B. Speed using distance and trip duration

- Univariate analysis

  A. Passenger count
  B. Trip duration using log transformation
  C. Distance travelled
  D. Speed according to trip count

- Bivariate analysis

  A. Distance and vendor
  B. Distance and store and forward flag
  C. Distance and hour
  D. How distance vary according to weekday
  E. How distance vary according to trip duration

- Feature encoding/ Feature Scaling

  A. One hot encoding
  B. Drop unwanted columns
  C. Applying Z score on independent variables

- Correlation analysis between independent/dependent variables

- Splitting dependent and independent features

- Train test split on dependent and independent features

- ML regression algorithms used

  A. Recursive feature elimination with hyper parameter tuning
  B. Lasso regression with hyper parameter tuning
  C. Ridge regression with hyper parameter tuning
  D. Decision tree regression using hyper parameter tuning
  E. XGBoost using hyper parameter tuning

5. Kamya Malhotra
Email:malhotra.kamya11@gmail.com

- EDA

  A. Dropping duplicates
  B. Handling null values
  C. Handling nan values
  D. Handling outliers with IQR range

- Feature Creation

  A. Created 4 time zones (morning, afternoon, evening , late night) using trip duration hours
  B. Object date time to date type
  C. Distance using longitude and latitude

- Univariate Analysis

  A. Log transformation on dependent variable
  B. Time zone plot according to count
  C. Store and forward flag
  D. Passenger count
  E. Vendor Id

- Bivariate Analysis

  A. Trip duration per vendor
  B. How distance vary according to trip duration
  C. Trip duration per week day
  D. Distance and store and forward flag
  E. Distance and hour

- Feature encoding/Feature scaling

  A. One hot encoding
  B. Dropping unwanted columns
  C. Standardization using Standard Scaler

- Correlation analysis between dependent/independent variables

- Splitting dependent and independent features

- Train test split on dependent and independent features

- ML Regression algorithms used

  A. Linear Regression with hyper parameter tuning
  B. Lasso Regression with hyper parameter tuning
  C. Ridge Regression with hyper parameter tuning
  D. Decision tree with hyper parameter tuning

During the whole project we continuously discussed all the approach and methods we implemented. Our daily google meets helped us clear our concepts and improved our confidence in group discussions.

Please paste the GitHub Repo link.

Github Link:- https://github.com/mouleenajaiswal1/CapstoneProject2

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

In this project, we have been provided with quite a big dataset of NYC taxi data with different attributes which helped us predict the time duration of a particular trip. Features in NYC taxi data were 'id' , 'vendor_id' , 'passenger_count', 'pickup_datetime', 'dropoff_datetime' , ' pickup_ longitude' , 'pickup_ latitude', 'dropoff_longitude', 'dropoff_latitude' and 'store_and_fwd_flag'.

We started with data overview in which we first looked into null/Nan values, fortunately there were no null values in our dataset it was quite clean. We also changed data types of pickup datetime and drop-off datetime.

Next we moved to feature creation in which we extracted hour, month and day from pickup and drop off datetime. After that with the help of longitude and latitude we got the distance which was one of the most important feature for our further analysis. We also extracted the speed which was quite useful for our EDA analysis. From extracted pick hour, we created the four time zones morning, afternoon, evening, late night which gave us clear idea of fares with respect to time. So after feature creation we handled the outliers using IQR and mean values.

After this, we proceed to EDA (Exploratory Data Analysis). In EDA, we implement univariate analysis and plotted different graphs which eventually helped us find some useful insights about the dataset. In univariate analysis of trip duration we transformed the highly right skewed dependent variable to normalization by log transformation. Additionally we performed bivariate analysis with respect to time duration and distance.

In the next step, we moved to feature engineering in which we performed feature encoding, feature selection and scaling. We used one hot encoding for feature encoding, to get dummies of some of the categorical columns .Next we dropped features which were not important and also with the help of heat map we dropped the columns which were highly correlated.
Then we scaled all the independent values by applying Z score, which got us all the values with mean 0 and std dev= 1

Next we stored all independent features in X and dependent feature i.e. trip duration in y. After this we applied train test split on our X and y. We applied different regression model such as Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Gradient Boost, and XGBoost.

Finally, we evaluated our regression models using mean squared values, root mean squared values , r2 score and adjusted r2 score which helped comparing accuracy of each model. In the end we concluded that XGboostregressor is best for prediction of time duration.