Capstone Project Submission

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| 1. **VISHAKHA KUMARI (**[**vishakhakumari0598@gmail.com**](mailto:vishakhakumari0598@gmail.com)**)**   **Contribution:**   * **Outlining of the project.** * **Data processing.** * **Creating features.** * **Analysis of model evaluation.** * **Conclusion.**  1. **ABHINAV AKOTKAR (**[**akotkarabhinaw@yahoo.com**](mailto:akotkarabhinaw@yahoo.com)**)**   **Contribution:**   * **Finding the target data.** * **Creating Dummies** * **Hot encoding** * **Map visualization.**  1. **DURGESH SHUKLA (**[**@yahoo.com**](mailto:akotkarabhinaw@yahoo.com)**)**   **Contribution:**   * **Univariate Analysis** * **Finding out the outliers and cleaning the data set.** * **Data visualization.** * **Splitting the data into train and test set of data points.** * **Checking the overfitting and underfitting of our model.**  1. **SHIVAM (**[**shivamself1997@gmail.com**](mailto:shivamself1997@gmail.com)**)**   **Contribution:**   * **Feature Engineering** * **Correlation heatmap** * **Methods** * **Model analysis of linear regression, XGBoost.** * **Find the model best fit line.** |
| **Please paste the GitHub Repo link @ drive link** |
| Github Link:- https://github.com/shivam0070/NYc-taxi-project-ml  Drive Link:- https://drive.google.com/drive/folders/1lGCpETlQtLxZx9yzHs4VEMelU4poIt0A |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| Taxi trip time Prediction is a predicting which execute total ride duration of taxi trips, times, numbers of taxies, passengers, pickup, dropoff data in various cities so consider New York City to find taxi trip duration. In this project we used to build a model that predicts the total ride duration.Primary dataset is one released by the NYC Taxi and Limousine Commission including several variables. The dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform.The training set (contains 1458644 trip records) in given dataset. First step is to Importing some important libraries and Installing haversine for distance calculation, mount google drive & Assigning the dataset to a variable using pandas.  Converting pickup and dropoff time to DateTime format.  **Analysis number of Passengers Per cab by** checking what are the minimum and maximum number of passengers travel per Taxi, and to check whether there are any outliers in our dataset.  From the sns box plot we observed that number of passengers per taxi, number of most trip passenger, minimum number of passengers per taxi. It looks little concerning as a trip cannot be done without a passenger.  Graphical representation shows the most of the trips were taken by passenger,Trip distance skewed, outlier, **Number of Trips Taken per second, Minute and hours,** pickup and dropoff hour vs number of rides, **trips per day, average speed.**  In Bivariate Analysis considering two variables and try to find out some usefull relation between them, most likely **Trip Duration Per Hour,Trip Duration per Weekday, Trip Duration per Month, Trip Duration per Vendor,** Trip Distance and Duration Distribution.  Plotting **Location Visualization On Map of** pickup and dropoff location according to their latitude and longitude. We can notice the map of the streets of NYC very clearly.  **In Correlation Heat map we notice that**there isn't much relation between the variables except distance-trip duration and dropoff longitude and pickup longitude.The relation between pickup and dropoff longitude might be due to the outliers, because it can't be same except the trip is a round trip. But we suspect there is surely not that much of trip which are round trip. **Splitting the Data into Train and Test inModel Selection and** Errors in **Linear Regression** assume that the covariance matrix of the errors is correctly specified. Defining the parameters for lasso regression which shows Fitting 3 folds for each of 10 candidates, totaling 30 fits  Defining the parameters for smooth running of the XG boost which shows Fitting 3 folds for each of 1 candidates, totaling 3 fits  At the end compare different parameters of all models and conclusion of model was XGBoost was the best for our Prediction. |
|  |