**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Contributor - Neha Pasi  Email ID – nehapasi05@gmail.com |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/Link/to/Repo> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| For this project I have the New York City Taxi Trip dataset with approximately 1458644 rows and 11 columns.  First I checked null values and found that there are no numerical column with missing data.  The no. of records with passenger count 0, 9 and 7 where very small compared to the entire data. So, I dropped them.  Then I have checked for data type of column names and found that the pickup\_datetime and dropoff\_datetime is object type.  I converted them into datetime format.  I created some extra features in our data [name - pickup\_timezone, dropoff\_timezone, pickup\_hour, dropoff\_hour, pickup\_month, dropoff\_month, pickup\_day, dropoff\_day] for Exploratory Data Analysis.  I checked for any outlier in our dependent variable trip\_duration using box plot and four observations were far from the rest of the observations. So, I removed them.  I have done EDA (exploratory Data Analysis) on the features of our data and to find the relationship between trip duration (Dependent variable) and all other features of data.  I created a function which returns the distance between a pair of latitudes and longitudes using haversine formula and calculated ‘distance’ column.  I found that there were 5868 rows of distance which are having value 0, instead of dropping them I replaced them with the mean value of distance column. Then I created ‘speed’ and ‘log\_distance’ column.  After EDA I done feature selection for training the model. And transformed categorical column into numerical using one hot coding.  Then I used train test split to split the data into training and testing dataset.  In the data of features selected, I fitted four models [model name - Linear Regression, Random Forest Algorithm, GradiantBoosting Algorithm , XG Boost ]  And used R-squared value, Mean squared vale and Mean absolute value to know the accuracy of every model.  I reached to the conclusions below.  Trip Duration varies a lot ranging from few seconds to more than 20 hours  Most trips are taken on Friday, Saturday and Thursday.  The average duration of a trip is most on Thursday and Friday.  The average duration of trips started in between 14 hours and 17 hours is the largest.  Vendor 2 mostly provides the longer trips.  The flag was stored only for short duration trips and for long duration trips the flag was less stored.  XG Boost model has given the highest accuracy ending with accuracy of 99%. So, XG Boost model is the best model for predicting trip duration count on daily basis. |