**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Name: Amol Rakhunde  Email: amolsr92@gmail.com |
| **Please paste the GitHub Repo link.** |
| GitHub Link: - https://github.com/Amolrakhunde/NYC-Taxi-Trip-TIme-Prediction |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| * My task is to build a model that predicts the total ride duration of taxi trips in New York City. My primary dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables. * My approach to the problem starts with understanding of the problem. For that, I have created a set of business metric questionnaires that’ll help me to better understand the problem. Those questionnaires are given below:  1. What exactly is the Trip duration?  * It is the time taken by the taxi to reach from one location to another location within New York City.  1. How the given dataset looks like?  * The above business metric gives me an idea about the data. i.e. whether the data is supervised or not.  1. What is the industry average trip duration?  * This will help the company to compare with its competitors in the market.  1. How the target variable ‘trip duration’ related to other features variables?  * Here I’m looking for the segment of the features variable which is helpful while making predictions Any trends? * Here I’m looking for the value (metric) which will help us to get an idea about the trend of trip duration with other variables. * Based on the above business metric first I’ll look at the data to check whether there are missing values or not. Then I did a visualization of features variables to check their distribution and correlation among them (features variable). To get information about the metric which are responsible for ‘trip duration’, I did a univariate (gives an idea how different features variable impacting the labeled feature) and bivariate (gives idea how combined effect of different features variable impacting the labeled feature) analysis. After that I’ll go for the data modeling part where I start with multiple linear regression (as data suggest it’s a regression problem) and will check their (linear regression) assumption and based on that I had modified the model. After all the statistical assumptions of linear regression, I moved to the Decision tree, gradient boosting and Random Forest model. Finally, I’ll compare all of the model bases on evaluation metrics which is RMSE (root mean square error) and R^2. |
|  |