**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| 1. Susovan Patra(susovanpatra061@gmail.com) 2. Turab Ahmed ([thurab455@gmail.com](mailto:thurab455@gmail.com))   We two as a team contributed the followings on the NYC Taxi trip duration,   * Performed EDA. * Removed outliers. * Checked Multicollinearity. * Trained 80% and Tested 20% with the data. * Applied various models to the training and testing data to predict trip duration. * Applied GridCV cross validation to figure out the learning rate and regression strength of the model to come up with a good R2 score. |
| **Please paste the GitHub Repo link.** |
| Turab’s Github Link:- [Thurab/NYC-Taxi-Trip-Duration-Regression: Predicting the taxi trip duration from NYC data (github.com)](https://github.com/Thurab/NYC-Taxi-Trip-Duration-Regression)  Susovan’s Github link :- [SUSOVAN-061/nyc-taxi-trip-ML: Predicting models for total trip duration (github.com)](https://github.com/SUSOVAN-061/nyc-taxi-trip-ML) |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| **Problem Statement: -**  A typical taxi company faces a common problem of efficiently assigning the cabs to passengers so that the service is smooth and hassle free. One of main issue is determining the duration of the current trip so it can predict when the cab will  be free for the next trip. We have New York City Taxi trip data and to come up with a model to predict trip duration.  **Approach :-**  Below are the approaches to predict dependent variable (Trip\_duration),   * Performed EDA. * Removed outliers. * Checked Multicollinearity. * Trained 80% and Tested 20% with the data. * Applied various models to the training and testing data to predict trip duration. * Applied GridCV cross validation to figure out the learning rate and regression strength of the model to come up with a good R2 score. |
| **Conclusions :-**   * From the EDA, the pickup and drop-off of the trip is peak at 5PM-7PM and most of the trips are taken on Saturdays * We applied Linear regression model to our training and testing data, where the R2 score result is 42.09% which is not good for predicting and it doesn’t show any sign of homoscedasticity. * We applied Ridge regression with lambda=0.01, and the model gives R2 score of 42.09%. Then using GridCV cross validation we found that best regression strength to apply Ridge regression is 60, we trained the model with lambda=60,but still the result stayed same with 42.09% * Next, we applied Lasso regression with lambda=0.005 with maximum iteration of 3000 which yield a result of R2 score=42.09%, then we apply Grid CV for Lasso and it recommended to apply with 0.001, the result came out to be 42.09%, which is no improvement * Then we applied ElasticNet regression which combines both Ridge and Lasso, this model gives us only 40.98% so we moved to next model * Finally, we applied XGBoost Regressor and this model gives 64.08% of R2 score, for further improvements we applied Grid CV which gives learning rate=0.9. When we apply the learning rate(eta)=0.9 there was no improvement on the result which stayed at 64.08%. |