**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| 1. **Shreesha K**   **e-mail-** [**shreeshasrisha@gmail.com**](mailto:shreeshasrisha@gmail.com)  **Contributions:**   * **Data uploading** * **Data cleaning** * **Data exploring**   **Data Visualization**   * **Month vs trip duration** * **Part of the day vs distance travelled** * **Day-wise distribution of trip duration**   **Machine Learning Models**   * **Linear Regression** * **Decision Tree** * **Lasso and Ridge Regularization** * **XGBoost**   **Evaluating the Models**   * **Mean Squared Error** * **Root Mean Squared Error** * **R2 Score** * **Adjusted R2 Score** * **Mean Absolute Error** * **Power point presentation** * **Documentation** |
| **Please paste the GitHub Repo link.** |
| Github Link:- [https://github.com/shreesha2304/ NYC-Taxi-Trip-Time-Prediction](https://github.com/shreesha2304/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)** |
| **With the ever-increasing population and number of vehicles, travel duration between two points can be unpredictable. Many controllable and uncontrollable factors contribute to the variation in trip duration. It was our task to build a Machine Learning model that can efficiently predict the average time required for a taxi to complete a journey in different conditions. We loaded the dataset, performed a general checkup, and checked for NULL or missing values. We converted some of the columns into different datatypes and created some more variables by feature engineering. We observed the distribution of the different variables and tried to remove outliers. By performing the Exploratory Data Analysis, we could draw some important observations regarding the data. We found out the months with the maximum number of trips, the part of the day in which the time taken for the journey is more, and the distribution of different variables involved in the dataset. We used five different ML algorithms to build the model. These are Linear Regression, Lasso and Ridge Regression, Decision Tree, and XGBoost. To get the best parameters for each of the models, we performed hyperparameter tuning and Cross-Validation techniques. For some models, the time taken for Gridsearch Cross Validation was more, in such a scenario we used HalvingGridsearchCV. We evaluated different models by using evaluation matrices like MSE, RMSE, R2 Score, and MAE. The importance given to each feature in a model was also calculated. Finally, we came to know that among the five models, the Linear models had not performed well compared to the other two models. The decision tree performed fairly better, while XGBoost provided the best results with the least possible errors and maximum R2 Score. It was concluded that XGBoost is the most suitable method for making predictions with the least error.** |