**Capstone Project Submission**

| **TEAM: Data Science Hustlers**  **Data: NYC Taxi Data** |
| --- |
| **Contributor Roles:**  **Kajal Mahajan :(**[**kajalmahajan1996@gmail.com**](mailto:kajalmahajan1996@gmail.com)**)**   1. **Data Wrangling and Visualization.** 2. **Total trip per hour and total trip per weekdays.** 3. **Feature selection based heatmap** 4. **Linear Regression**   **MrityunjaySarkar: (**[**mrityu1488@gmail.com**](mailto:mrityu1488@gmail.com)**)**   1. **Data Wrangling and Visualization** 2. **Finding Unique Values** 3. **Trip duration per hour and trip duration per weekdays** 4. **Feature extraction based heatmap.** 5. **XGBoost Regression**   **Akash Choudhury :(**[**akashchoudhury545@gmail.com**](mailto:akashchoudhury545@gmail.com)**)**   1. **Data Wrangling and Visualization** 2. **Finding Null Values** 3. **Trip duration per month and trip duration per vendor.** 4. **Random Forest Regression**   **Adarsh Gaurav :(**[**adarshgaurav18@gmail.com**](mailto:adarshgaurav18@gmail.com)**)**   1. **Data Cleaning and Visualization** 2. **Handling Missing Values** 3. **Distance per hour and distance per weekdays** 4. **LASSO Regression.**   **Vivek Raikwar :(**[**raikwarvivek13@gmail.com**](mailto:raikwarvivek13@gmail.com)**)**   1. **Data Wrangling and Visualization** 2. **Data Exploration** 3. **Average speed per hour and average speed per weekday.** 4. **RIDGE Regression.** |
| **Please paste the GitHub Repo link.** |
| GithubLink:- https://github.com/akashbzs/NYC\_Taxi\_Trip\_Time |
| **Capstone Project Summery, Problem Statement, Our approach, and Conclusion** |
| Our task is to build a model that predicts the total ride duration of taxi trips in New York City. Our 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. This dataset had around 1458644 observations with 11 columns namely *id, vendor id, pickup datetime, dropoff datetime, passenger count, pickup longitude, pickup latitude, dropoff longitude, dropoff latitude, store and fwd flag,* and *trip duration.*  Here we did Data wrangling and after that basic EDA is designed to give answer of the following questions   * Hour and weekly basis total trips. * Trip duration per hour, per weekday, and per month. * Trip duration per vendor. * Distance per hour and distance per weekdays. * Average speed per hour and average speed per weekday.   Next we have done feature selection, and applied various regression measures to assess which model is best fit for future assessments. The various regression measures are: Linear, Lasso, Ridge, XGBoost, and Random forest. Each of us have taken a Regression model and tested it. What we have found that, XGBoost is performing best, followed by Random forest, Linear, Lasso, and Ridge. Conclusion:1. Taxi pickup is generally more in weekdays, that too in daytime.2. In winter the pickup is lowest, which gradually increases in the following months.3. Average taxi speed is maximum at dawn, and it’s the time when a taxi travels maximum distance.4. Application of feature selection significantly reduces the dimension of the dataset.5. Because of poor linear dependency between predicted feature (time) and other predictors, the linear regression gives a bad result. 6. Similar argument can be followed for non linear dependencies too. |