**NYC TAXI TRIP TIME PREDICTION**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Team members and their contributions:   1. Md Nawab Ali:   Email: nawabali.7410@gmail.com   1. Understand the dataframe 2. Data cleaning: 3. Exploratory Data Analysis 4. Multicollinearity studies 5. Dataset preparation for modeling 6. Machine learning 7. Conclusion 8. technical documentation 9. powerpoint presentation 10. Jahnavi Jaolekar:   Email: jaolekarjahnavi@gmail.com   1. Understand the dataframe 2. Data cleaning 3. Exploratory Data Analysis 4. Multicollinearity Studies 5. Dataset preparation for modeling 6. Machine Learning 7. Conclusion 8. technical documentation 9. powerpoint presentation 10. Kaustubh Amare:   Email: kaustubhamare197@gmail.com   1. Understand the dataframe: 2. Data Cleaning: 3. Exploratory Data Analysis 4. Multicollinearity studies 5. Data preparation for modelling 6. Machine learning 7. Conclusion 8. technical documentation 9. powerpoint presentation |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/mdnawabali/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)** |
| **PROBLEM STATEMENT**  To Build a machine learning model that predicts the duration of NYC taxi trip using the dataset which includes pickup time, geo-coordinates, the number of passengers, and several other variables.  **APPROACHES**  **Step 1:**  We began by loading the required libraries, mounting the drive, and saving the data in variables in order to derive relevant insights. Viewing and cleaning the data was our first step. Since the presence of null values could have led to issues in the next steps, the presence of any null values was evaluated.  **Step 2:**  The following steps were preprocessing and feature engineering. We converted two columns from timestamp format to datetime format. We used geodesic distance to construct an additional column of "distance" as part of the feature engineering process.  **Step 3:**  The next phase was data analysis and visualisation, during which we identified the busiest times for pickups and drops-offs, the months that saw the most demand for taxi services, as well as additional information on weekdays that saw a spike in taxi use, etc. Additionally, we used the quartile approach to deal with outliers in our data set. We also used heatmaps and VIF to do a multicollinearity correlation evaluation.  **Step 4:**  The last phase was training our model using five distinct algorithms (linear regression, decision tree, xg boost, gradient boost, and random forest). We evaluated the evaluation metrics for all five algorithms and, based on those results, chose one to utilise as our final approach.  **CONCLUSION**  (EDA)  We discovered that the months of April and March are the busiest for using taxi services, and that Friday and Saturday are the busiest weekdays for using taxi services, which may indicate that people are travelling for weekend celebrations or parties. We also discovered that most pickups and drop-offs occur in the early mornings around 10 am and in the late evenings, which suggests that people use taxi services for When compared to vendor 1, vendor 2 had more bookings.  (Model Training)  In order to train a model, we tested 5 techniques (Linear Regression, Decision Tree, XG Boost, Gradient Boost, and Random Forest) and compared the assessment metrics for each (R2, Adjusted R2, MSE, RMSE). Out of 5 algorithms, we discovered that only XG Boost provided the best accuracy score (71%), the lowest MSE scores, and Random Forest came in second. We then tried taking the optimal parameter to prevent overfitting of our model. |