NYC Taxi Trip Prediction

Team Member's Name, Email and Contribution:

1. Aayush Sharma (Team Leader)
sharma919ayush@gmail.com
- Basic colab work for package imports and data inspection.
- Data summary work for slides and documentation.
- Outlier Dealing for target variable and latitude and longitude data.
- Feature engineering and one hot encoding.
- Final colab commenting work and document review.

2. Rishika Rai
rishikarai70@gmail.com
- Data exploration with DataPrep library. Detailed report creation and briefing data insights and univariate analysis.
- Model training and model cross validation.
- Slide work and documentation work for all trained models.
- Final slide work review.

3. Aman Guleria
amansingguleria@gmail.com
- Finding out missing values. Data cleaning and plotting comparison graph.
- Data Preprocessing and setting matrices for score evaluation.
- Creating function for model and error plot.
- Hyperparameter tuning with randomised search cv.

4. Saurabh Aradwad
saurabhdilip95@gmail.com
- Importing a csv file and github admin work.
- Work in abstraction, Data inspection and visualisation with Seaborn.
- Model comparison, finalising model and drafting results.
- Slide work - for plots and their explanations. Work on document submission & Colab final review.

GitHub Repo link.

🔗 Link:- SaurabhAradwad/NYC-Trip-Time-Prediction-Project: Supervised Machine Learning Project (github.com)

Summary

The Taxi and Limousine Commission (TLC), established in 1971, is in charge of licencing and regulating New York City's Medallion (Yellow) taxi cabs, for-hire vehicles (community-based liveries, black cars, and luxury limousines), commuter vans, and paratransit vehicles.Every day, approximately 1,000,000 trips are completed by over 200,000 TLC licensees. Drivers who work for hire must first pass a background check, have a clean driving record, and complete 24 hours of driver training. TLC-licensed vehicles are inspected at TLC's Woodside Inspection Facility for safety and emissions.

The dataset is based on data from the 2016 NYC Yellow Cab trip records, which were made available in Big Query on Google Cloud Platform. The NYC Taxi and Limousine Commission first made the data public (TLC). For the purposes of this project, the data was sampled and cleaned.

Our task was to build a model that predicts the total ride duration of taxi trips in New York City. 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.

Final Conclusions

- LGBM models have the highest accuracy of 93% when compared to other models.
- The least accurate solutions for this problem are linear, lasso, and ridge regulation.
- Gradient Boost took a moderate time of 8 minutes and 7 seconds, while LGBM took the shortest time of
- LGBM's low learning rate increases accuracy while decreasing percentage error.
- If we want to reduce the training time of our model, we can use hyperparameter tuning to select tuned parameters. It will produce similar results in much less time.
- Logarithm treatment of trip duration is important, due to skewed data.
- Speed and distance calculation. We can confirm this with correlation heatmap and skewed.
- Project is very helpful for improving consumer experience by telling them their arrival time early.