Capstone Project Submission (II)

Summary of 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:
1. Prerit Tyagi (prerittyagip2@gmail.com) **Contributor Roles:** ❖ **Interpretation of dataset** ❖ **Data Cleaning** ❖ **Data Preparation** ❖ **EDA and Data Processing** ❖ **Methodology: -** **1: - Linear Regression** **2: - Lasso Regression** ❖ **Correlation** ❖ **Optimal Model for Testing & Evaluation** ❖ **Challenges** ❖ **Scope of Improvement** ❖ **Conclusion** **2. Saurabh Dattatray Waghmare (sdwaghmare1998@gmail.com)** **Contributor Roles:** ❖ **Interpretation of dataset** ❖ **Data Cleaning** ❖ **Data Preparation** ❖ **EDA and Data Processing** ❖ **Methodology: -** **3: - Ridge Regression** **4: - Decision Tree Regressor** ❖ **Correlation** ❖ **Optimal Model for Testing & Evaluation** ❖ **Challenges** ❖ **Scope of Improvement**

- **Conclusion**

3. Shivraj Yadavraj Saude(shivraj.saude@gmail.com)

Contributor Roles:

- **Interpretation of dataset**
- **Data Cleaning**
- **Data Preparation**
- **EDA and Data Processing**
- **Methodology: -**

  5: - Random Forest Regressor

  6: - LGBMR Regressor
- **Correlation**
- **Optimal Model for Testing & Evaluation**
- **Challenges**
- **Scope of Improvement**
- **Conclusion**

GitHub Repo link.

Github Link:- https://github.com/Preritp2?tab=repositories
Drive Link:
https://drive.google.com/drive/folders/1zZLXglUNkPTt5RDzmkZyAg52lzIhpT1_?usp=sharing

Summary

Machine learning has been of significant help as it has helped businesses in abundant ways. We willcreate a model which will predict taxi trip time duration.

Most on-demand taxi platforms require a way to know the estimated time which driver will be occupied. It is important to predict how long a driver will have his taxi occupied. If a dispatcher or systemgot estimates about the taxi driver's current ride time, they could better recognize which driver to allocate for each pickup request which results to be less waiting time which means less cancellation from the client side and an increase in profit margin, as well as customer base.

The first thing we noticed the dataset has more than 14 lakhs of data, so we've implemented a function that will convert the dtypes of pandas dataset to NumPy.int format to reduce memories. We've also noticed the target variable has positive extreme right-skewed distribution; we've applied log transformation to the target variable. We've plotted graphs, done both univariate and bivariate analysis, and tried to extract a generaloverview and stories of the dataset. Also, we've created some new fields like distance, month and day name, etc. We've split the dataset into 3 parts, train, test, and validation. Then, we applied StandardScaler bysklearn to normalize the data. We've tried Linear, Lasso and Ridge Regression, DecisionTreeRegressor, RandomForestRegressor, and LGBMRegressor. Out of these, we found the best r2_score to be ~0.75 on the validation dataset using the LGBM model. Though the evaluation metrics on prediction are not that accurate, we can use some external features like rating by user, driver experience level, driver rating ortraffic conditions, etc., these can significantly boost our model's performance.