Capstone Project Submission

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

New York City, with a relatively large number of populations and busy streets require taxi with some good algorithms to move from one place to another without any delay. A lot of streets and roads in New York city are quite busy due to traffic jams, construction or roadblockage etc. Therefore, it is very important to **predict the trip duration of taxi** so that the user will know how much time it will take tocommute from one place to other.

In this project, through our exhaustive analysis and feature engineering we have trained the model to predict the value of trip duration using various models such as Linear Regression, Random Forest, XG Boost and Light GBM. We started with understanding the **problem statement and studying the dataset**. We found to have approx. 14 lakhs of observations but only 11 features. So, we started with first exploring the dataset and looking for nulls and duplicates. To our surprise there were no null values as well as duplicates were found.

Thereafter, **exploratory data analysis** was performed by closely looking and understanding each of the given features. During this we found that number of passengers were ranging from 0-9. So, we removed the observation with 0 and 7-9 as these number of passengers are not possible to take trip in taxi. It was also found that vendor id and store and fwd flag had slight variations with trip duration. The dataset also had categorical features such as pickup datetime and drop off datetime, so we converted them to datetime and extracted various features such as day, month, hour, minute and seconds. These extracted features were found to have close relation with trip duration and thus they were of great importance.

As **feature engineering** is the main task to be performed while modelling any dataset, therefore we created a new feature distance with help of pickup and dropoff latitude and longitude using distance formula. After looking at distance of trips we excluded trips with distance less than 50m and greater than 180km as both the cases seems to be suspicious data. We also ordered week day and time of day to three categories according to traffic density. Further and most important we dropped the rows with trip duration less than 15s and more than twice the standard deviation. We did this feature engineering and outlier removal basically on two features i.e distance and trip duration. Thus, after completing feature engineering, we plotted distance and trip duration we found to have cone shaped graph but with lots of error.

Further we chose features passenger_count, pickup_longitude, pickup_latitude,dropoff_longitude, dropoff_latitude, store_and_fwd_flag, pickup_month,pickup_tps,pickup_dayn, pickup_timeofday, and distance for **training our models**. Initially we looked at the baseline scores and then we performed Random Forest regressor, XGBoost Regressor and finally LightGBM. After training these models, **LightGBM** showed the best score of **81.1% on test data** and therefore we performed hyperparameter tuning using random search cross validation technique to further tune our model, but couldn't achieved the score better than this.

Lastly, we **studied our model** using SHAP and found that feature: distance is creating maximum impact on our output: trip duration. We also performed a check on Kaggle to check the authenticity of our model and scored 0.46. Thus, after creating our model we are ready to predict the trip duration with the provided features in dataset. In future we can improve the score by performing more feature and engineering and hyperparameter tuning.

Team Member's Name, Email and Contribution:

Cohort Kaimur- SSP

Contributor roles:

A. Ganeshkumar Patel
Email- ***ganeshkumarpatel452@gmail.com***

- **a) Data Wrangling**
- **b) Exploratory Data Analysis**
- **c) Feature Extraction**
- **d) Feature Engineering**
- **e) Model Explainability**
- **f) Achieving Kaggle score, writing inferences on Colab notebook and making PPT.**

B. Akanksha Agarwal
Email- ***akn.agarwal@gmail.com***

- **a) Data Wrangling**
- **b) Exploratory Data Analysis**
- **c) Feature Extraction**
- **d) Feature Engineering**
- **e) Hyperparameter Tuning**
- **f) Writing inferences on Colab notebook, drafting technical document and summary**

Please paste the GitHub Repo link.

Github Link:- https://github.com/Akn-ag/NYC-Taxi-Trip-Time-Prediction