**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 fairly huge variety of populaces as well as active roads call for taxi with some great algorithms to move from one area to another with no delay. A lot of roads as well as roadways in New York city are fairly active because of traffic jams, building and construction or roadway blockage and so on. Consequently, it's crucial to forecast the journey period of taxi to ensure that the individual will certainly understand just what does it cost? time it will certainly require to commute from one area to various 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.  Afterwards, exploratory information evaluation was carried out by very closely looking and also recognizing each of the provided attributes. Throughout this we discovered that variety of passengers were varying from 0-9. So, we eliminated the observation with 0 and also 7-9 as these variety of passengers are not feasible to take journey in taxi. It was additionally discovered that vendor id and also store and fwd flag had actually small variations with trip duration. The dataset additionally had actually categorical attributes such as pickup\_datetime and dropoff\_datetime, so we transformed them to datetime and also extracted different attributes such as day, month, hour, minute and also secs. These extracted attributes were found to have close relationship with trip duration and also hence they were of terrific significance.  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. |
| **Team Member’s Name, Email and Contribution:** |
| **Cohort Kaimur- SSP** |
| **Contributor roles:**     1. **Ganeshkumar Patel**   **Email-** [*ganeshkumarpatel452@gmail.com*](mailto:ganeshkumarpatel452@gmail.com)   1. **Data Wrangling** 2. **Exploratory Data Analysis** 3. **Feature Extraction** 4. **Feature Engineering** 5. **Model Explainability** 6. **Achieving Kaggle score, writing inferences on Colab notebook and making PPT.** 7. **Akanksha Agarwal**   **Email-** [*akn.agarwal@gmail.com*](mailto:akn.agarwal@gmail.com)   1. **Data Wrangling** 2. **Exploratory Data Analysis** 3. **Feature Extraction** 4. **Feature Engineering** 5. **Hyperparameter Tuning** 6. **Writing inferences on Colab notebook, drafting technical document and summary** |
| **GitHub** |
| Link:  <https://github.com/GaneshkumarPatel/NYC-trip-duration-prediction/blob/main/SSP_NYC_Taxi_Trip_Time_Prediction_Capstone_Project.ipynb> |
| **Google drive:** |
| Link:-  <https://drive.google.com/drive/folders/17wKJt4mENsLlK8TgyB9S8MWhjX-2g-vw?usp=sharing> |