Exploratory Data Analysis
Executive Summary
NYC Taxi Dataset

Authors : dvasani,pphadke

New York is one of the busiest cities in the world. Hence, analysing trip duration is an interesting task. We tried to find trends in trip duration based on factors like number of passengers, hour of the day, month and so on. From the feature engineered data that contained 18 columns, the most important factor in the dataset is Distance in Kilometers. It drives the model greatly and also makes sense that longer distances take more time. That can be observed in the picture below on the left. The model trained only on Distance to predict the Trip duration performs decent with an accuracy of ~72%

The other columns do not add much value to the model and the accuracy drops to ~49% if only the other columns are used.

Hence, there are factors which are not present in the data that need to be accounted for. External factors like traffic rate, blockages, road work and then train a model to better predict the trip duration.

In the figure on the right, we can see that Ground Truth(original Trip duration values) and Predictions are close. The dataset is clearly imbalanced since it has more than half data in the short trip category. And it definitely makes the model biased hence the model misses ~30% of the predictions and gives approximately ~72% accuracy.