**NYC Taxi Trip Time Prediction**

**Ankush**

**AlmaBetter, Bangalore**

**Introduction**

### The dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this project. Based on individual trip attributes, you should predict the duration of each trip in the test set.

Our experiment can help understand the reason for a taxi time trip by feature selection, data analysis and prediction with python.

**Problem Statement**

### My task is to build a model that predicts the total ride duration of taxi trips in New York City. Mine's primary task dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables.

**About data**

So, in our dataset we have

* id - a unique identifier for each trip
* vendor\_id - a code indicating the provider associated with the trip record
* pickup\_datetime - date and time when the meter was engaged
* dropoff\_datetime - date and time when the meter was disengaged
* passenger\_count - the number of passengers in the vehicle (driver entered value)
* pickup\_longitude - the longitude where the meter was engaged
* pickup\_latitude - the latitude where the meter was engaged
* dropoff\_longitude - the longitude where the meter was disengaged
* dropoff\_latitude - the latitude where the meter was disengaged
* store\_and\_fwd\_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
* trip\_duration - duration of the trip in seconds.

**Steps involved:**

* **Exploratory Data Analysis**

After loading the dataset we use python, sql and excel to analyze our data to get our results. This process helped us figuring out various aspects and relationships among the target and the independent variables.

* **Import some important library**

After loading our dataset we import some important libraries too like pandas, numpy, seaborn etc. These libraries help us alot to analyze and to get the results.

.

* **Visualization**

I use different types of graphs like bar graph, pie chart , box plot graph and many more so that we can visualize data in a better way. It also helps to understand the statements.

* **Create Model**

I make models of linear regression because model training and prediction is fast. Make another model of decision tree and random forest to overcome several problems.

**Conclusion**

That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA , an important library and visualization.

* Most of the trip consists of 1 or 2 passengers.
* Vendor 2 is evidently more famous among the population.
* Most pickups are between 5 PM to 9 PM.
* Fridays and Saturdays are those days in a week when people prefer to roam in the city.
* Average distance traveled is approx 3.5 kms.
* Most of the trips are limited to the range of 1-10 kms.
* Most of the rides are 10 min.
* There were very few trips (0.55%) of which the records were stored in memory.
* Average trip duration is lowest at 6 AM when there is minimal traffic on the roads.
* Most of the taxi pickups were done in the Manhattan area as compared to the other areas in NYC.
* Trip distance is highest during early morning hours.
* Possible reason behind this is Outstation trips taken during the weekends.
* Trip distance is fairly equal from morning till the evening
* Sunday being at the top may be due to outstation trips or night trips.

**References-**

* Stack overflow
* GeeksforGeeks