**Abstraction**

Building a model that predicts the total duration of taxi trips in New York City. We are working on a dataset released by the New York City Taxi and Limousine Commission, which includes pickup time, geographic coordinates, number of passengers, and many other variables, and we will talk about the details during the report.

**Motivations**

The motivation behind building a model to predict the duration of a taxi trip within New York City is the great impact it has on both parties, namely the service provider and the customers ( passengers ) .

**It provides the customer with all the following** :

1- Planning and improving the trip schedule

2- Adherence to appointments

3- In emergency situations, such as going to the hospital

**It provides the service provider with all the following:**

1- Correctly allocate human resources

2- The ability to contribute to the development of the city by providing information that is useful to

1. Planning roads and infrastructure for the future
2. traffic management
3. Improving transportation

**Dataset**

* **Description**

The competition 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 playground competition. Based on individual trip attributes, participants should predict the duration of each trip in the test set.

**Data cleaning**

Data cleaning is one of the important parts of machine learning. It plays an important role in building the model. The success or failure of the project depends on proper data cleaning. Professional data scientists usually invest a very large portion of their time in this step.

**“Better data beats cooler algorithms.”**

**We dealt with our own data set and carried out the most important steps in the process of cleaning the data from it.**

**1- The private data set does not contain missing values**

**2- The private data set does not contain duplicate values**

**3- fixing structure error**

We modified pickup\_datetime ,dropoff\_datetime from object to datetime so that we can perform some operations on it.

**4- Handle outlets**

After reviewing the statistical summary of the data set, we noticed that there are some values that are not possible for some columns, for example :

* trip\_duration We found that some trips lasted for more than 22 hours, which raises doubts as it is an impossible value.
* passenger\_count Some flights had no passengers.

some trips that have many passengers (7 passengers)

**5- Data transformation**

I also applied Common scaling methods like Min-Max scaling and Standardization (Z-score scaling).

**EDA Summary**

**dataset**

* training dataset have 1000000 rows and 11 columns
* 7 columns include numerical values.
* 4 columns include object(text, date) values.
* Don’t have any null values or duplicated values**.**

**vendor id**

* vendor id has two value 1,2
* vendor ID 2 has made more trips than vendor id 1
* vendor ID 2 made 54% of trips.
* The different vendors did not affect the average trip duration, so their values were very close

**Number of passengers**

* 71% of trips had 1 passenger and 14 % trips had 2 passengers and 5 % of trips had 3 passengers.
* There are two trips have 7 passengers.
* some trips that have zero passenger (38 trips)

**Longitude and latitude**

* From longitude and latitude of start and end point can  
  extract new features distance
* Distance has clearly outliers ,some trips have distance 771 kilo meters that is impossible.
* Trip duration and distance have high correlation, can you profit from it in your model.

**Note :**

Now have distance and trip duration ,can you extract new feature speed (speed = distance / duration )

**store\_and\_fwd\_flag:**

* 90 % of the trip not a store and forward trip.

**Speed:**

* Speed is computed by divide distance by trip duration.
* Can not used speed on your model but can get sum intuitions about correlated it with other features.

**date features:**

* The number of trips increases at noon and in the evening than at any other time.
* The number of flights increases on Friday, Saturday, and Thursday of each week.
* The speed of vehicles increased between night and late at night.
* The speed of vehicles increased on Sunday and Saturday
* The speed of vehicles increased in the months of 5 and 6
* The speed of vehicles increased on holiday days than on other days.

**add new feature from The Open-Source Routing Machine (OSRM) dataset.**

* There are many columns that can help him predict like total\_distance , total\_travel\_time , number\_of\_steps
* You can also verify this by merging with the main dataset by column ID.

**Total distance :**

* total\_distance has negative skew
* most distance is between 0 and 5000 metre

**total travel time:**

* total travel time has negative skew.
* most total travel time is between 0 and 16 minutes.

**number\_of\_steps:**

* number\_of\_steps have negative skew.
* most number\_of\_steps is between 1.5 and 2.5 steps.

**Modeling**

|  |  |  |
| --- | --- | --- |
| model | Train R2-score | validation R2-score |
| Ridge | **0.7298** | **0.7311** |
| Random forest | **0.8351** | **0.7972** |
| XGBRegressor | **0.8108** | **0.8056** |