**NYC Taxi Trip Duration**

**Abstract:**

New York City taxi rides form the core of the traffic in the city of New York. The many rides taken every day by New Yorkers in the busy city can give us a great idea of traffic times, road blockages, and so on. Predicting the duration of a taxi trip is very important since a user would always like to know precisely how much time it would require of him to travel from one place to another. Given the rising popularity of app-based taxi usage through common vendors like Ola and Uber, competitive pricing has to be offered to ensure users choose them. Prediction of duration and price of trips can help users to plan their trips properly, thus keeping potential margins for traffic congestions. It can also help drivers to determine the correct route which in-turn will take lesser time as accordingly. Moreover, the transparency about pricing and trip duration will help to attract users at times when popular taxi app-based vendor services apply surge fares. Thus in this research study, we used real-time data which customers would provide at the start of a ride, or while booking a ride to predict the duration and fare. This data includes pickup and drop-off point coordinates, the distance of the trip, start time, number of passengers, and a rate code belonging to the different classes of cabs available such that the rate applied is based on a regular or airport basis. Hereafter, we applied different machine learning Perceptron models to find out which one of them provides better accuracy and relationships between real-time variables. At last, a comparison of the two mentioned algorithms facilitates us to decide that is more fitter and efficient than Multi-Layer Perceptron for taxi trip duration-based predictions.

**1.Problem Statement**

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

## **2. 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.

### NYC Taxi Data.csv - the training set (contains 1458644 trip records)

### Data fields

#### 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

3.Approach

Machine learning has become a very important tool in making important business decisions and even people with no coding knowledge or domain experience can develop models with libraries such as data prep, and sklearn. Scikit learn is one of the most powerful machine learning libraries out there. It is used by major corporations around many more.

The package is named scikit-learn therefore you can do

pip install scikit-learn

however, inside your python file, you’d have to do

import sklearn

Import the libraries that we would need.

import pandas as pd

import date time as dt

import matplotlib. pyplot as plt

import seaborn as sns

from sklearn. linear\_model import Linear Regression

import numpy as np

 We will import the libraries and download the data from the source mentioned and load the data as a pandas dataframe.

Next, we explore the data set and the fields available. In summary, we have fare and distance fields available for the yellow cabs but not for for-hire cabs. So as a fun exercise we would try to compute the total fares for-hire cabs (assuming they are similar to yellow cabs; which is not the case).

we convert the data and develop basic plots using pd.datetime.

next is to develop a machine learning model (linear regression) to predict the time for-hire cabs based on Pick Up and Drop Location IDs.

Once model is built on train data prediction has done on test data. Based on train and test data we can evaluate different evaluations parameters and based on that we can make decision which model fits better to take decision.

Analysis has done based on pca and without pca approach. As target variable is dependent we have built supervised models also target variable is continuous so tried linear regression model, decision tree, random forest.

**4.Analysis**

Once file has read we have to perform EDA and pre-processing on it once data is cleaned and ready we can have built different models on it. By checking nulls in this data set there are no nulls present here. So data is cleaned after that converting non-numeric to numeric by using label encoder. So by all these process data is ready so we can perform different visualisation to summarize data. So by this EDA we can summarize data and we get idea what exactly data want to tells us before building model.so by EDA we came to know daily trips duration, weekday trips, monthly trips analysis.

**Summary of Univariate Analysis-**

**1.**Vendor **2** hassignificantly more number of trips than Vendor 1.

2. Around **73%** of the trips have only one passenger with some anomalies of 0, 7, 9 passengers.

3. Negligible percentage **(0.55%)** of rides have a problem with Uploading the Trip Details right away.

4. Most of the trips were less than 10kms and were are of short duration as well (10**-**14minutes).

5.Evenings had the maximum number of taxi trips whereas it was the least during

Early Mornings.

6. Increasingtrend is observed in the number of trips from MondaytoFriday and it decreases on the weekends.

7. March had the highest number of trips.

**Summary of Bivariate Analysis-**

**Vendor 2** has more average trip duration as well as higher average distance than Vendor

Replacing the outliers/anomalies shows that as the passengercounts increase the duration too increases.

**2.Vendor 1** is preferred for longer duration trips by 1 passenger count whereas for others 2 is preferred.

3.Weekdays have longer duration with a peak on Thursday and Friday.

Early morningtrips are faster as expected and noon travels around 1PMto5PM are longer. Office hours are duration is lesser than noon and evening hours. Increasing trend is observed on duration per month

Most of the trips are of shorterduration (1-10 kms).

4. We also see few anomalies with zero distance travelled and huge trip duration.Shorterduration trips face connection issues. Only Vendor 1 had this connectionissue, this might be due to poor GPS machine used by them.

6. The correlation heat map shows that there is notmuchcorrelation among the independent and target variables, except for slight correlation among latitude and longitudes.

**Linear Regression Model:**

As target variable trip duration is continuous we can have built linear regression model.

By passing the PCA Transformed data in our Machine Learning Regression Algorithms. To begin with, Linear Regression is a good approach, by splitting our Data into Training and Testing (30%). Our evaluation metric is RMSLE, not R-squared. We can also hyper tune our Parameters to minimize the loss (RMSLE). We will also calculate Null RMSLE, which we can set as a benchmark for our Models RMSLE.

**Evalution Results:**

With PCA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithms** | **Training Score** | **Validation Score** | **Cross Validation Score** | **R2-Score** | **RMSLE** |
| Linear Regression | 0.0424 | 0.0438 | -0.0485 | -23.10 | **-** |
| Decision Tree | 0.9258 | 0.9169 | 0.9137 | 0.9104 | **0.037** |
| Random Forest | 0.9304 | 0.9245 | 0.9233 | 0.9177 | **0.035** |

Another approach we could go with is without PCA, just Standard Scaling Dataset and applying our Algorithms.

The approach can give us better idea of what works better for us.

This approach might take great amount of computational resources and time

**Recommended Approach:**

* Apply Standard Scaling on the Dataset to Normalize the values.
* Further, Apply PCA to reduce dimensions, as you’ll extract features from our primary Date Time Feature. Those additional features might lead our model to suffer from “Curse of dimensionality” and could drastically affect performance.
* Pass the PCA Transformed data in our ML Regression Algorithms and Evaluate results.

**Insights:**

Observed which taxi service provider is most Frequently used by New Yorkers.

Found out few trips which were going from 528 Hours to 972 Hours, possibly Outliers.

With the help of Tableau, we’re able to make good use of Geographical Data provided in the Dataset to figure prominent Locations of Taxi’s pickup / drop-off points.

Also, found out some Trips of which pickups / drop-off point ended up somewhere in North Atlantic Sea.

Passenger count Analysis showed us that there were few trips with Zero Passengers.

Monthly trip analysis gives us a insight of Month – March and April marking the highest number of Trips while January marking lowest, possibly due to Snowfall.

In a day, we could observe that 5pm to 10pm is the time when New Yorkers Rush too much.

Observations says that Friday’s and Saturday’s are those days in a week when New Yorkers prefer to get out of their home.