FINAL PROJECT

TLC TRIP RECORD DATA

Done by:
Arwa Alolyani Maitha Alqahtani

Submit to: Mr.Ali El-kassas

Abstract

New York City has a huge population that moves from place to place, which means taxi rides are a major part of the traffic. Almost half of the people of New York rely on public transportation, and it is estimated that 54 percent don't own a car or personal vehicle. As a matter of fact, 200 million taxi trips are made annually.The amount of taxi we will need to pay will only be known after we arrive at our destination, and our awkwardness makes us curious about how we can predict this amount, and an accurate estimation will help us control our budget.

Design

This project is one of the T5 Data Science BootCamp requirements. Data provided by from :

- the New York City Taxi and Limousine Commission (TLC).
- NYC Taxi Zones data from NYC OpenData.
- Central Park weather data from the National Climatic Data Center.

Data

we selected a sample of basic data of New York City Taxi and Limousine Commission (TLC) containing more than 6 million rows and created a million-row sample and 18 columns and This data is lacking coordinates for each location, so we downloaded from NYC OpenData as the (GeoJSON) file containing 263 columns and 9 rows as well weather data for jan month in 2020 from Central Park weather data from the National Climatic Data Center containing 4595 columns and 9 rows.

Algorithms

1.Download and read data

Choosing 1,000,000 as a random sample

Divide the data into 3 sections: Training, Verification and Test

2. Quick Look at the Data Structure

3. Prepare the Data for Machine Learning Algorithms

Delete null values

Deleted all values with less than 0 passengers

Deleted all values below 2.5 in the fare amount

Remove all points outside of New York

Deleted all values less than 0 in the distances

4. Feature Engineering

Extract longitude and latitude from geometry columns using geopandas

Calculation of actual distance by latitude and longitude

Convertir date columns type to date type

Extract the day, hour and time from the date columns

Merging weather data for New York on January 1, 2020

5.Models

Looking for Correlations

Create linear regression model for predict fare amount

Training and Evaluating on the Training Set

Evaluate the Test Set

Tools

Library :

Pandas , Numpy,Math ,Matplotlib.pyplot,Sklearn ,Scipy, Sklearn.linear_model ,Statsmodels.api and statsmodels.formula.api, yellowbrick.regressor, geopandas,folium.

Program :

Python , Jypter,Visme,Github.