Trip Duration report

Project title: New York Taxi Trip duration

Author: Ahmed diab

Date: September- 2025

Abstract:

This NYC-trip duration project inspired by Kaggle competition aims to predict the duration for trips in New York by given 10 features and 1Millioon example for train and 200000 for val after doing feature engineering and feature extraction and use model like linear regression, ridge, neural network, xgboost we end up with best model from xgboost regressor with f1-score: 76% for val & 72% for test

Introduction:

Accurately predict the trip duration will reduce the road traffic and help drivers, by deployment as an api the driver will be able to predict the road duration by giving some features

Dataset Descriptions:

Sources: inspired by Kaggle project with some changes

Size: 10 features with 1 million example train, 200000 examples for val and 625134 for test

Features:

- 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
- 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

From EDA:

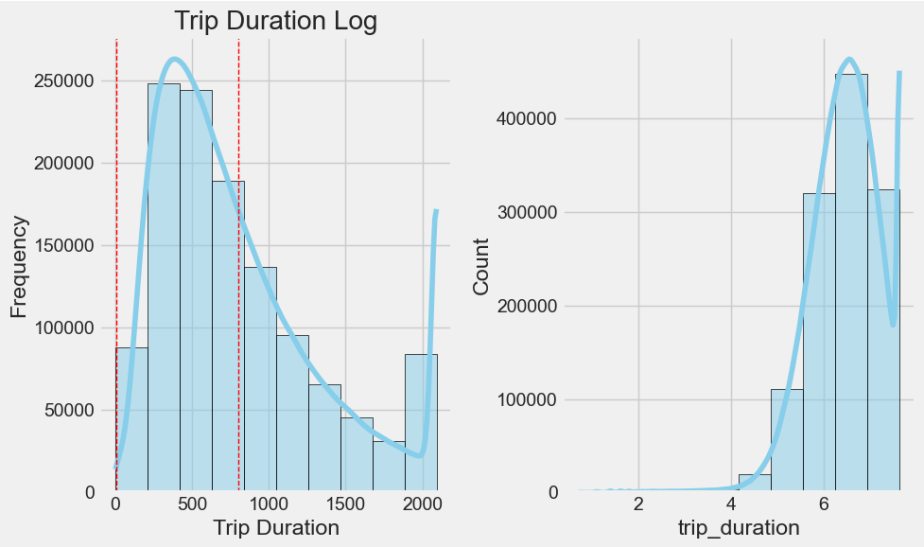
- drop ID useless feature

- data are clean and just 6 duplicate examples

- target feature

- before remove outlier, very high skewness, upnormal plot

- after remove outlier, the data in normal range, and almost guassian plot



- most of the passengers are 1 and 2 but their high outlier like 7 and 8

- we calculate haversine distance from longitude&latitude and got very good correlation with target feature

- from datetime feature varation

- Friday and Saturday are the days that there are trips count

- most of the trips are after 18 hours (maybe because this is after working hours)

- other variations are not so important

- the top 8 features with target are [haversine_distance, dropoff_longitude , pickup_longitude ,pickup_datetime, month, dropoff_latitude, pickup_latitude, season_encoder]

Methodology:

Data processing:

- data are clean no missing data

- There are 6 dublicate examples, drop them
- there is outlier in data, we drop them
- We calculate haversine distance from longitude and latitude and give us high correlation with target trip duration
- We extract from pickup datetime
  - Year
  - Month
  - Hour
  - Day of week (1-7)
  - Day name (ex: Wednesday)
  - Season name (ex: Sprint)

  After that we drop date time feature

- Apply log on trip duration target
- Encode store_and_fwd_flag to int
- Apply scaling (MinMaxScaling, StandardScaler, etc)
- Apply PolynomialFeature

Models:

- linear regression as linear model
- Ridge regression as to prevent overfitting
- Neural network, ml regressor model
- Xgboost , boosting technique

Metrices:

We focus on r2-score and mean square error (mse)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \qquad \text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

Results:

Metric	R2-score (val)	MSE (val)
Linear regression	0.6	0.226
Ridge	0.59	0.232
Neural Network	0.586	0.234
Xgboost	0.76	0.134

The best model we choose for testing XGBoost that give in val: r2-score: 0.76% and MSE: 0.134

And for test: r2score: 0.725 and MSE: 0.174

Conclusion:

We save the val and test model as pkl file now we can install requirements and config file and try the model as an api to predict the ride trip duration.