New York City Taxi Trip Duration

Domain Background

I will be working on Kaggle competition "New York City Taxi Trip Duration" where in I will build a model that predicts the total ride duration of taxi trips in New York City. This problem statement is relevant to transportation industry and can be applied in different contexts like ride sharing services, food/grocery delivery. Being able to predict ride duration accurately is extremely important for both the entities (service provider & customers) involved in the transaction. From service provider standpoint, it's important as ride duration would dictate pricing that could be charged to a customer. In addition, predictability (duration) also provides good estimate of the number of drivers who would be present in certain region at any given time which is important so that customers could be matched with drivers/service provider in an efficient manner. From customer standpoint, predicting duration could help them make a decision as to when is the optimal time to start their commute.

With the advent of technology based cab services, this problem has become particularly important for the drivers and the customers. A good prediction mechanism can be instrumental for drivers in optimizing their returns, while also saving the customers from the uncertainties attached to a trip.

Problem Statement

Objective is to predict the time duration (in secs./mins.) of a New York taxi ride as a function of independent attributes like pick up and drop off location, time, volume etc. I will study the impact of various features and also attempt to find the best possible ways to leverage those features. I will explore Ensemble (decision tree/random forest) models to do the prediction.

Datasets and Inputs

The data provides the details of taxi rides in the New York City from Jan 2016 to June 2016. This data is provided by the NYC Taxi and Limousine Commission (downloaded from Kaggle). Each trip records fields:

#	Field	Description
1	id	a unique identifier for each trip
2	vendor_id	a code indicating the provider associated with the trip record
3	pickup_datetime	date and time when the meter was engaged
4	dropoff_datetime	date and time when the meter was disengaged
5	passenger_count	the number of passengers in the vehicle (driver entered value)
6	pickup_longitude	the longitude where the meter was engaged
7	pickup_latitude	the latitude where the meter was engaged
8	dropoff_longitude	the longitude where the meter was disengaged
9	dropoff_latitude	the latitude where the meter was disengaged
10	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)
11	trip_duration	duration of the trip in seconds

Solution Statement

I will use the provided dataset (see above table) which will be further processed to extract additional features like year, month, day, weekday, hour and minute from the date and time of each ride, as well as the speed, rides in an hour, distance among others. I will explore decision tree/ random forest models to model the nonlinearities of traffic and location effect. For the purpose of training models, tuning hyperparameters and reporting the results, I will split the data into three sets, training set (60%), validation set (20%) and test set (20%).

Benchmark Model

I will run linear regression to do the prediction and use the output as a baseline for comparing it against the actual model.

Evaluation Metrics

I will use Root Mean Square Error (RMSE) as a metric to measure model's accuracy. To validate the usefulness of model, RMSE values for baseline model and the actual model developed would be compared on the test set.

Project Design