New York City Taxi Trip Duration

Domain Background

The goal of this project is to predict trip durations for taxi rides. Predicting the duration of taxi rides is useful, because it can improve the efficiency of electronic taxi dispatching systems. VHF-radio dispatch systems have widely been replaced by electronic dispatch systems. Each taxi vehicle has a data terminal that records GPS and taximeter information as the trip progresses. A taxi driver receives ride requests from people hailing taxis on the street, as well as pickup requests from dispatchers. In addition to the change in dispatch system, the method in which dispatch messages are made has changed. Unicast-based messages, where a dispatcher communicates directly with a driver, rather than broadcast-based messages, where a dispatcher announces a pick up request to multiple drivers, are now the norm.

Since dispatchers are now communicating one-on-one with drivers, each dispatcher needs to decide which driver he/she should send to a pick up location. However, this is not straightforward because dispatchers are not informed of the drop off locations for a request, and the taxi driver does not input this information into the vehicle's data system. Thus, if the duration of a taxi's ride could be predicted, a dispatcher would be able to make better assignments of a pickup request to a taxi. This problem has been previously studied in this Kaggle competition, where the destinations of taxi trips are predicted based on their initial trajectories, and in this Kaggle competition, where the total travel times of taxi trips are predicted based on their initial partial trajectories.

Problem Statement

In this report, we look at a Kaggle competition with data from the NYC Taxi and Limousine Commission, which asks competitors to predict the total ride time of taxi trips in New York City. I will use several machine learning methods for the prediction task, and the models will be evaluated using the root mean squared error.

Datasets and Inputs

The main dataset for this project is data from the NYC Taxi and Limousine Commission (TLC), that was published as 2016 NYC Yellow Cab trip record data on Google Cloud Platform. For the Kaggle competition, the data was cleaned and sampled, so that the training dataset contains 1,458,644 trip records, and the testing dataset contains 625,134 trip records. Each trip record contains the following attributes (descriptions taken from the Kaggle competition):

- 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

I will augment the provided dataset with data for the weather in New York City in 2016. This data was obtained by the National Weather Service, who recorded the weather at a station in Central Park, New York City for the first half of 2016. The features of the dataset are the daily low, high, and average temperatures, precipitation, snowfall, and accumulated snow depth.

Additionally, I will use a dataset that provides the fastest routes for each trip, based on maps from the Open Source Routing Machine, OSRM. For each trip in the Kaggle NYC taxi dataset, this dataset provides the total distance between the trip start and end calculated from the fastest possible route, the trip duration for this route and the steps of the route (turns taken).

Solution Statement

Using the three datasets, I will make a prediction of the trip duration for each route in the testing set. I will use the machine learning models described in the Project Design section to predict the trip duration, and compare the performance to the performance of the Benchmark Model, where performance is measured by the root mean squared error. Since the datasets are publicly available, my solution can be replicated.

Benchmark Model

As the benchmark model, I will use the dataset for the fastest routes for each trip from the OSRM, to simply predict that each trip takes the fastest possible time.

Evaluation Metric

The competition's evaluation metric is the root mean squared logarithmic error:

$\epsilon = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(\log(p_i + 1) - \log(a_i + 1))^2}$, where n is the number of observations in the dataset, p_i is the predicted value for trip duration, and a_i is the actual value for trip duration.

Project Design

- **Programming Language and Libraries**
  - Python 3
  - Scikit-learn: Python's open source machine learning library
  - XGBoost: Python package for XGBoost model
- **Data Cleaning**
  - I will first perform an exploratory data analysis and remove outliers from the dataset. For instance, I will check whether there are trips with nonsensical travel times, like trips with 0 travel time, or trips longer than 24 hours. Since the data is for trips in New York City, I will also check whether the latitude and longitude for each trip in the dataset is actually in the New York City area.
- **Models**

  I will use the following machine learning models to predict trip time:

  - Linear Regression
  - KMeans clustering
  - Random forest
  - XGBoost

  The root mean squared logarithmic error will be reported for each model, as well as for the benchmark model.

References:

1. **http://www.ecmlpkdd2015.org/discovery-challenge/learning-taxi-gps-traces**
2. **https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i**
3. **https://www.kaggle.com/c/pkdd-15-taxi-trip-time-prediction-ii**
4. **https://www.kaggle.com/c/nyc-taxi-trip-duration**