IBM Applied Data Science Capstone Project

Analysis of taxi trip data in the Mexico City Metropolitan Area

Fabiano Corsetti

February 6, 2021

1 Introduction

1.1 Background

The Mexico City Metropolitan Area (MCMA) is one of the largest metropolitan areas in the world, with an estimated population of over 20 million. It comprises Mexico City itself, made up of 16 municipalities, plus 41 municipalities in the neighbouring states of Mexico and Hidalgo. Taxis are one of the many modes of transportation available throughout the city. There are several different classes of taxi and taxi-like services in operation:

- taxis libres (free taxis), that are hailed from the street;
- taxis de sitio (stand taxis), that are linked to particular taxi stands;
- radio taxis, that are ordered through a centralized service;
- ride-sharing companies such as Uber and Cabify.

These services all differ in their use, coverage, cost, fare model, modes of payment, regulation, and safety. We can therefore expect these differences to be reflected in real-world taxi trip data.

1.2 Problem

Our aim is to analyze a dataset of individual taxi trips collected in the MCMA within a data science methodology. Using machine learning (ML) techniques, we wish to answer the following questions:

- What clusters emerge from the data, and do they align with the classes listed in Sec. 1.1?
- What are the main features of each cluster?
- Can a ML model be constructed to accurately predict the class of taxi for a given trip from the other data fields?

1.3 Interest

Our analysis is aimed towards city officials, urban development agencies, and other groups that are interested in improving the city's transport network, e.g., making it more sustainable and effective for inhabitants and visitors. In order to do so, an important first step is to analyze and understand current usage patterns, as we propose to do for the case of the taxi network. The results of our analysis will therefore help inform on how, when and where different classes of taxis are being used, and might potentially uncover useful insights such as holes in coverage that could be addressed in future.

2 Data

2.1 Data sources

The analysis will be conducted using a publicly available, open dataset published on the Kaggle website (1). The dataset contains taxi trip data collected between June 2016 and July 2017 for numerous cities in several countries, notably Quito, Mexico City, and Bogota. There are more than 12,000 individual trips recorded for Mexico City.

The data was collected by users of the EC Taximeter mobile app (2). The app is designed to estimate the cost of the ride based on the phone's GPS location data, so that users can verify that they are being charged fairly. The data fields included from the app in the dataset are:

- the class of taxi (taxi libre, taxi de sitio, radio taxi, and four types of Uber rides);
- date and time for pick up and drop off;
- GPS coordinates for pick up and drop off;
- the total trip distance;
- the total trip duration;
- the wait time (time the taxi was stationary during the trip).

All data is anonymized; the anonymization process and privacy implications have been discussed in detail in previous studies (3).

In order to assist with the taxi trip data analysis, we will also use geospatial data of the territorial subdivisions of the city. These are available on two levels, the largest being the 16 municipalities mentioned in Sec. 1.1, which can then be further subdivided into smaller colonias (neighbourhoods). The government of Mexico City maintains a public database of 1812 colonias covering the entire city with the boundary data for each (4); the boundaries for the municipalities as a whole are also available on another personal data analysis website (5).

2.2 Data cleaning and wrangling

A challenge in preparing the data in the taxi trip dataset for later analysis is to decide how to extract the trips occurring within the MCMA. The single dataset contains trips from several Mexican cities; although some are in other parts of country and are easily distinguishable (e.g., Merida, Veracruz, Torreón), the bulk of the data is in the MCMA but extends continuously to other neighbouring cities (e.g., Toluca). Our geospatial boundary data only covers Mexico City itself but not the MCMA, although it is important to include this larger area in the analysis. Furthermore, we can expect some trips to start within the MCMA and end outside it or vice

versa. In order not to restrict the analysis more than necessary, we will use a radial cut-off from the city centre with a radius equal to the distance to Toluca; this is large enough to cover the entire MCMA. Trips either starting or ending within this cut-off will be included.

Once the trips are selected, we can then use the boundary data to assign the start and end coordinates either to Mexico City or the surrounding metropolitan area. For points within Mexico City we can further assign both a municipality and a colonia.

References

[1] Mario Navas. Taxi routes of Mexico City, Quito and more. `https://www.kaggle.com/mnavas/taxi-routes-for-mexico-city-and-quito`. Accessed: February 5, 2021.

[2] `http://www.ectaximeter.com`. Accessed: February 5, 2021.

[3] Giancarlo Camilo. Demand analysis and privacy of floating car data. Master's thesis, University of Victoria, 2019. `http://hdl.handle.net/1828/11150`.

[4] Delimitación territorial de las colonias de la Ciudad de México. `https://datos.cdmx.gob.mx/dataset/coloniascdmx`. Accessed: February 5, 2021.

[5] `https://hoyodecrimen.com/api/v1/municipios/geojson`. Accessed: February 5, 2021.