Who Dominates Taxi Business?

STAT 432 FA2019 - Team LRHK

Eric Jong Bum Kim (jekim3)
Israel Reyes (ir2)
Jihee Hwang (jhwang55)
Hyun Suk Lee ()

November 15, 2019

Abstract

Coming Soon

Introduction

Uber and Lyft have became one of the most dominant taxi-service business around the major cities around the world, especially big cities in United States. However, when one visits big cities such as New York or Chicago, one realizes that there are plenty of traditional yellow cabs around the city. With this in mind, our group decided to study the traditional taxi market in the big city. With major taxi service companies still providing means of transportation, can we analyze the traditional taxi market from 2013 to present day.

In an attempt to construct a meaningful and effective way to detect the taxi service provider company in Chicago, various statistical learning techniques will be applied to Chicago Taxi Service data from 2013 to present. The main goal of our model is to see if a certain taxi service company can be mapped according to the financial and trip logistics of each fares in Chicago. If such is possible, this could provide an arguments about those that criticizes Uber and Lyft as two monopolistic companies in taxi service industries. Possibly decide if staying in the traditional taxi business is worth it or not.

Background Information On Data

The dataset originates from the Chicago Digital Hub [1] and was accessed through Kaggle [2]. The safety and quality of taxi rides in Chicago are guaranteed through BACP [3] (Department of Business Affairs and Consumer Protection) where they are authorized to collect information on taxi rides. This data is reported periodically through two major payment processors in which not all trips are reported but is believed to cover most of the taxi rides in Chicago. Based on this data reporting process, the dataset contains over 100 million taxi rides in Chicago from 2013.

Due to private concerns, following measures were implemented:

- Each trip appears with delay in the data, long after the completion of the ride
- The Taxi ID is consistent for a given taxi medallion number, but was created specifically for the dataset
- Location is provided only at the Census Tract [4] and Community Area levels [5]

Other corrections applied to the data include:

- Trip times less than zero or greater than 86,400 seconds are removed.
- Trip lengths less than zero or greater than 3,500 miles are removed.
- If any component of the trip cost is less than $0 or greater than $10,000, all components of the trip cost are removed.

Data Description

There are 23 columns of information about each trips. There are trip logistics information including trip starting and ending time stamps, trip duration and miles, pick up and drop off locations, latitude, and longitude, trip fare, tips, number of tolls, extra charges, total trip costs, and payment methods, and information about taxi and its company. More explanation of each columns can be found in the appendix of the file at the end of the project.

To further breakdown the data set, there are various categories that each variables can be classifeid into.

- taxi & company descriptions : *unique_key*, *tax_id*, *company* are used to distinguish each unique taxi trips.

[1] Digital Hub: Chicago Taxi Data Released
[2] Chicago Taxi Trips Data
[3] City of Chicago: BACP
[4] Chicago Data Portal: Census Tracts
[5] Chicago Data Portal: Community Areas

- geographical locations : *pickup_latitude*, *pickup_longitude*, *pickup_location*, *dropoff_latitude* , *dropoff_longitude*, *dropoff_location*, *pickup_community*, and *dropoff_community* are used to decribe a phyiscal location of some aspect of the trip.
- trip logistics : *trip_start_timestamp*, *trip_end_timestap*, *trip_seconds*, *trip_miles*, and *tolls* are used to describe the trip logitics in terms of time and miles travled
- finanical-related : *tripi_total*, *extras*, *tips*, and *payment_type* are used to decribe the fare-related portion of the trips.

Statistical Learning Task

That being said, the statistical learning task to accomplish this goal will be with the means of classification models. Given certain features of the dataset like the geographical location, the length of the trip, cost and many others we will be classifying the company that corresponding trip applies to. It is immediately apparent that binary classification models are out of the question unless we would like to target a specific company to make it the positive class meanwhile the others are grouped into another classification. The problem with this however is that could potentially make the dataset uneven and although there are methods to combat this problem, we would have to model depending how many unique companies are in the dataset. Multiclass classification would have to be the main method for our goal with use of K-Nearest Neighbors and Random Forests along with cross validation, where we will be using 2 metrics to maximize being accuracy and sensitivity.

Data Loading

Hyun Suk

Data Modeling

Hyun Suk