TEAM 11

Date: 26/10/2019

Project: Final project

**Main Question**

|  |
| --- |
| How can we order in an efficient way the physical files in a transit office warehouse to reduce the pickup time of the folders as a first step; keeping the order through time to guarantee the efficiency of the warehouse? |

Supplementary Questions

1. What factors influence the citizen transit requests? For example:
   1. Vehicle type
   2. Engine size
   3. Register Date
   4. Others.
2. What history records impact the folder location:
   1. If the vehicle is being sold
   2. Ownership transfer
   3. The last record of the vehicle
   4. Scrapping.
   5. Accidents
   6. Taxes
   7. Others.
3. How questions 1 and 2 are related.
4. What action methods could help to shorten the filing time?
5. What ML algorithm should be used?

**IMPROVING FOLDER PROCESSING TIMES AT TRANSIT OFFICES IN COLOMBIA**

Currently in Colombia there are more than 200 entities in charge of handling citizens' requests related to transit activities, such as license plates registration, driver's license issuance, ownership certificates, fines, etc. Those entities are called Transit Offices and are part of the Ministry of Transportation.

One of the requirements that Transit Offices have is to securely manage and store the physical documents provided by the citizens and those generated for each transit request, and this is where we see an opportunity to apply data science. Every request implies that a folder has to be taken out of the storage location so the new documents are attached and then the folder is taken back to the same location. This activity demands time and resources, sometimes impacting citizens’ time because they have to wait while the folders are updated to their location.

We are going to focus this project on one of the top 5 transit offices in Colombia, located in Cundinamarca, which currently holds more than 200,000 folders for license plates related documents only. Each folder is stored in X200 type boxes that could hold between 25 to 30 folders each and there are 10,800 storage locations available at the warehouse; those storage locations are classified in 3 categories depending on the ease of access to the boxes: 1. Easy access; 2. Medium access; 3. Difficult access, those locations can be seen in the attached layout.

The objective is to develop an algorithm capable of predicting when the next citizen's request is going to come for a specific folder and define the best location for it, making the folder’s pickup, processing and storage times shorter; also, once the folder is updated, it should instruct where to store it after. Several variables affect the need to move a specific folder, such as time since registration, vehicle type, engine size and previous requests, among others, and we have data available from 2014 to 2019 that we will used to calculate the predictions for the best possible storage location.

When a customer goes to a transit office, there are certain requirements he/she has to fulfill, including physical documentation; then, the transit office representative validates that all the documentation is correct, scans the citizen’s biometrics and processes the request through an in-house system that is connected to the national transit database (known as RUNT) and a local ERP system that handles accounting and invoicing; then, the invoice, license or certification of the request are printed out and all documents scanned; finally, those documents are physically stored at a centralized location.

Project: Datathon

**Main Question**

|  |
| --- |
| If I am a cab driver and I leave a passenger, which is the best borough to go around and increase the probability of taking another one? |

If a cab driver is available in New York, he or she must decide where to go in order to find their next passenger. We want to give them a tool that will recommend to which borough they should go to increase their chances of picking up a new passenger.

The recommendation is going to be based on:

* The location where the driver is at. The application is going to recommend the location where the driver should go.
* The date time at that moment where the driver is looking for their next passenger. The amount of passengers available in the different boroughs change according to the date time.
* Historical data of the performance of yellow cabs, green cabs, uber, and mta in the nearest places (i.e. boroughs) related to the location of the driver. This performance is going to feed the model
* Socio demographic data of the boroughs related to the location of the cab driver.

We are going to use the following external data in the model

* Calculated Distance of every trip (we are working out on how precise to make this calculation).
* Data about the estimated fare

To answer the question proposed in our Datathon project we are planning on using all of the databases in order to estimate a score per borough, where it will propose the best one to move to. Below we will explain why and how:

* Green and yellow trips: These databases will help us estimate the number of passenger per period of time and to be able to estimate in the future how much people will be there waiting to take a cab.
* The mta and uber trips: will be used as additional factors to estimate the score and to try to improve it.
* Zones and Geographic: These data files will help estimating the zones in New York, and also it will help estimating where the cab driver is and which boroughs are near.
* Demographics: With this data set the score will be feed and eventually will explain the estimated revenue of the near boroughs.
* Weather: We will use this dataset to see if the weather could impact in the probability of getting a passenger in any borough.