UNDERSTANDING AND PREDICTING JOURNEYS OF ATLANTA CITIZENS

EY NextWave Data Challenge

Caterina Selman and Mrad Azoury

Table of Contents

INTRODUCTION **2**

PROPOSED APPLICATIONS IN SMART CITIES **3**

MODELLING METHODOLOGY **4**

I) DATA LOADING AND PRE-PROCESSING 4

II) DATA ANALYSIS AND EXPLORATION 6

III) MODELLING 8

Introduction

Given the growth and expansion of cities around the world, public authorities are facing new challenges of making sure the increasing number of citizens in different cities does not inhibit the quality of life in them. Governments and private companies are trying to face these problems with new solutions and ideas that are reshaping the way citizens live, work and travel. With an increasing amount of available data and new tools and techniques available to understand and make use of this data these challenges can now be faced with different strategies and increased value can be brought to citizens through new solutions.

The problem presented for this EY Data Science Challenge involved predicting the ending location of trajectories of the citizens of Atlanta. Specifically, for each trajectory taking place between 15:00 and 16:00 hr. we predicted whether the trajectory's ending location was inside or outside the city borders.

The model built to generate these predictions is a prototype that if adapted to work with additional data sources and different cities can be expanded to help authorities and companies tackle the challenges of big cities in a data-driven manner.

In the following sections the possible use cases for such model are explored. Additionally, the methodology involved in developing this model to produce predictions of trajectory ending locations is described in detail.

Proposed Applications in Smart Cities

This problem presented for this Data Challenge and the solution and model implemented and described below could be adopted to help decision-making in cities with the necessary data available. The solution could also be flexible to additional sources of data and details to further improve the performance of the model and expand its possible use cases. Below is an outline of two possible use cases for such model.

1) **Use predictions of citizen locations and movements throughout the day to efficiently allocate ride-hailing and ride-sharing resources.**
   Ride-hailing and ride-sharing mobile applications are re-shaping how people travel and commute in big cities. During the past recent years traditional cab rides can have been replaced and expanded by ride-hailing application, car-pooling options, and easy access to shared bikes and scooters. These companies provide cost-effective and efficient options to citizens and, more importantly, they have the potential of decreasing the number cars on the road by providing alternative options and decreasing car idle waiting times through efficient planning. This has the potential to alleviate traffic and decrease air pollution. These ride-sharing services could be optimized by predicting where citizens will be moving at certain times of the day and therefore redirect drivers, cars, bikes, scooters, etc. throughout the city to those areas of the city that have the highest predicted demand.

   With the current model, a preliminary implementation could be designed to predict whether more resources would be needed inside or outside the city at different hours of the day. If, however, more granular data becomes available, a more specific and higher performing implementation could be designed. For example, if more specific areas or districts are defined with latitude and longitude borders, then more exact predictions could be made of where citizens are going to be at specific time. Additionally, if the date of each trajectory is available in the training data set useful trends of how descriptive features of each day (weather, day of the week, holidays, etc.) affect the movement of users. Finally, if the dataset was able to keep the same device ID for a specific citizen across multiple days then trends could be formed on the different type of ride-sharing car users (commuters, leisure travelers, etc.). These supplementary data details could potentially lead to a more useful model to make ride-sharing car allocation and scheduling more efficient.

2) **Use of predictions of citizens locations throughout the day to improve city systems and infrastructure.**
   There is also a potential value for the public sector in understanding citizen trajectories and movements and being able to predict these under certain circumstances. For example, addressing the problem of alleviating traffic and keeping automobile drivers safe. With predictions of the quantity of citizens that will be in the city center or outside at a specific point in time (and ideally even more exact locations) authorities could adapt traffic signals, highway directions, speed controls, etc. to alleviate traffic and mitigate accidents.

   If this model was expanded to include streaming data from IoT devise such as highway and automobile sensors, decisions on high way controls could be adapted in real-time and therefore improve the quality of mobility in cities across the world.

Modelling Methodology

I) Data Loading and Pre-Processing

In order to develop a model that predicted the ending locations of trajectories at a certain time of the day, a training data set with historical geographic trajectories was provided. This train dataset described how Atlanta citizens moved around the city in a given day with features such as the longitude and latitude entry points, the entry and exit times of the trajectory, as well as velocity details (if available). Additionally, a similar dataset was provided to test the performance of the model developed and in this dataset the ending location of the citizen's last trajectory was hidden. Both of these training sets were loaded in their raw CSV format and joined before processing.

A limited amount of data exploration was done at the point the data was loaded and joined. Firstly, it became obvious that the target variable that would need to be predicted (whether or not a citizen was inside or outside city center limits at a specific time) was not explicitly set in the training dataset and therefore would need to be calculated based on city center borders provided. Additionally, it was clear that a number of trajectories in the dataset could belong to a single location tracking device, or citizen. To understand the trends and relationships of each citizen throughout the day extensive pre-processing was therefore required prior to significant explanatory analysis or findings was done.

Initial Feature Creation

As a first pre-processing step, a number of features were created to further describe each trajectory.

A list of new features created for each trajectory is outlined here:

- *Entry in Town:* whether or not the entry point of the trajectory lies inside town borders (based on latitude and longitude points and town borders given)
- *Exit in Town:* whether or not the exit point of the trajectory lies inside town borders (based on latitude and longitude points and town borders given)
- *Time Entry Seconds: Time* of start of trajectory in seconds from midnight
- *Time Exit Seconds:* Time of end of trajectory in seconds from midnight
- *Time Difference:* Time duration of trajectory in seconds
- *Distance Traveled:* Euclidean distance between trajectory longitude and latitude start and end points
- *Moved into town:* Whether this trajectory moved the device from out of town to inside town borders
- *Moved out of town:* Whether this trajectory moved the device from inside town to outside town borders
- *Changed:* Whether this trajectory moved the device across town borders
- *Calculated velocity:* Average velocity of trajectory based on distance traveled and time difference of trajectory
- *Distance from city center:* Euclidean distance from trajectory starting point (lat/lon) to city center point

- *Percentage of square around that changed:* Calculates the percentage of other trajectories that starting from the same area (1000 lat/lon points away from entry point) changed into or out of town

Aggregating

Given the dataset included multiple trajectories for the same device on a given day and each trajectory had only a limited number of descriptive features, additional features were developed to describe a specific trajectory based on that device's previous trajectories for that day. This was done by grouping data per device and aggregating certain features of each device's previous trajectories.

The following new features were created through this process:

- Sum of previous trajectories that started inside town
- Sum of previous trajectories that ended inside town
- Sum of previous trajectories that moved out of town
- Sum of previous trajectories that moved into town
- Average distance traveled in previous trajectories
- Average time elapsed in previous trajectories
- Average calculated velocities of previous trajectories

II) Data Analysis and Exploration

After these initial data pre-processing steps some data exploration was done to understand trends in the data that could help in the modelling phase. The most significant of these findings are described in this section.

The first finding that became clear through data analysis was that the relative number of trajectories that moved devices from inside to outside the town, or the vice-versa, was low, as shown in **Figure 1** below. Specifically, this number of trajectories that '*change*' their in-town status is only around 6% of the train set. This could imply that the starting location of each trajectory would have some significance in our models.

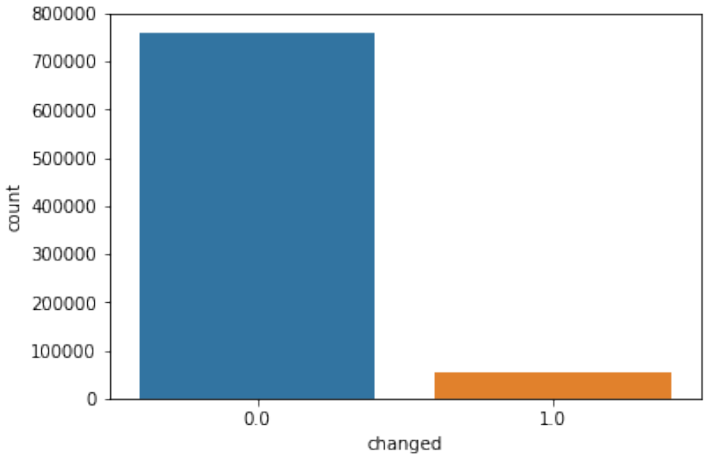

Figure 1

Additionally, looking at the ranges of times elapsed for each trajectory, a significant number of trajectories with a '*time difference*' of 0 was noticed (as shown in **Figure 2**).

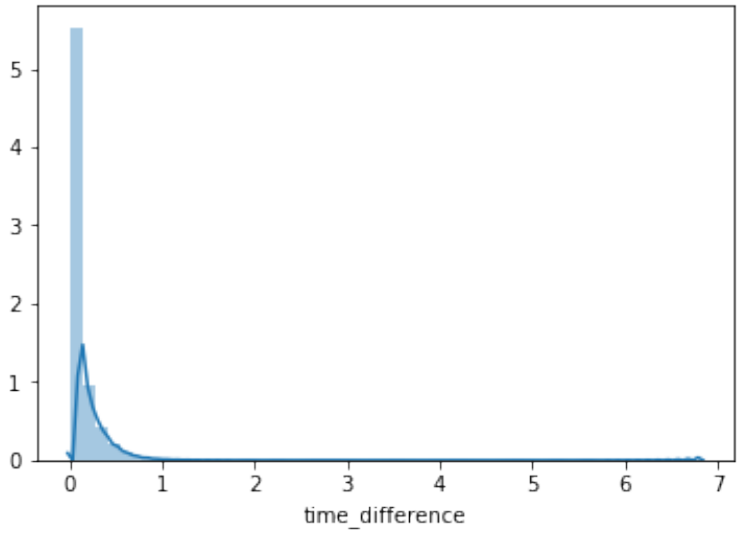

Figure 2

In all of the records of the train set where '*time difference'* was 0 no displacement of the device took place. In other words, the latitude and longitude entry points where the same as the exit coordinate points. These cases, which represent around half of the training data set, therefore did not need to be part of the model. Instead, the descriptive feature '*entry in town'* value for these trajectories could be used as the target variable '*exit in town'* for the predictions.

After filtering out these cases where '*time difference'* was 0 from the train set the starting locations of those trajectories that '*changed'* in-town status was explored through plotting the entry coordinate points as showing below in **Figure 3**. The patterns show of starting locations of those trajectories that moved to or from the city center could be helpful to predict the ending location of an Atlanta citizen trajectory.

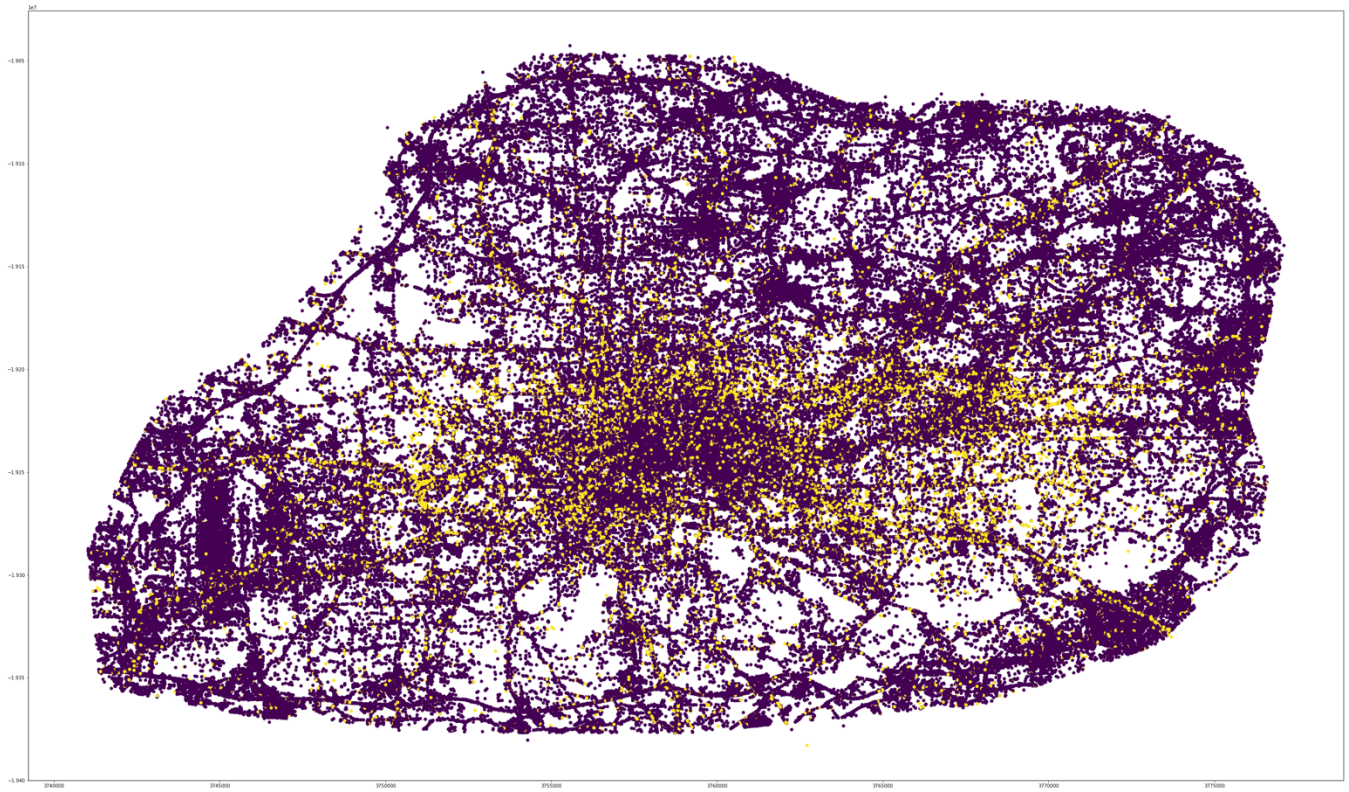

Figure 3

III) Modelling

A number of distinct algorithms and approaches were attempted throughout the model-building phase using the train set provided and the 17 additional features that had been generated for each trajectory through the pre-processing and feature engineering stages. The model that predicted ending locations of trajectories with the highest performance will be explained in this section in detail.

- To generate an additional set of descriptive features for each trajectory a **Denoising Autoencoder Neural Network** was modeled to represent and fix noise present in the train and test datasets.
    - A Neural Network was modeled where the input was our raw data with noise and the output was our original dataset.
    - This middle layer of this Neural Network represented features that could help fix the noise in our data and therefore could be added to our data set before training and predicting a whether each trajectory was inside or out of town for the last trajectory of the day. The middle layer of this Neural Network resulted in 32 additional features.
- To reduce the number of the additional features created **Principal Component Analysis** was applied to the 32 new features. This resulted in 5 components representing the variability in the 32 features extracted in the previous step.
- Finally, this resulted in a final data set of 22 features that could be modeled to predict the final location of Atlanta citizens.

Given the nature of the predictions required (whether Atlanta citizen's trajectories between 15:00 and 16:00 hr. resulted in an ending location inside or outside the city center) a **Binary Classifier** was used to model the problem.

As mentioned above, although a number of algorithms were attempted, the best performing one was a **Gradient Boosting Tree Classifier (XGBClassifier).**

In accordance to the metric that used to measure the prediction submissions through the EY platform, the F1 score was also used to fine-tune the model and increase performance.