Building a better working world

MILTON VASCONCELOS DA GAMA NETO

EY NEXTWAVE DATA SCIENCE COMPETITION 2019

GLOBAL URBANIZATION IS RISING

Population living in urban areas:

54% nowadays

66% in 2050

*Source: United Nations (UN), 2015

MOBILITY

- One of the elements that suffer from population growth
- Government responsibility to ensure the infrastructure and urban service
- Locomotion difficulty in public spaces
- Impact on quality of life

PROPOSAL FOR SMART CITY THRIVE

- Understanding of movement
- Adjust the availability of public transport according to demand
- Encourage increased of alternatives and shared transport
- Adjust the traffic signals to adapt to the movement and predicted flow
  - The prediction can be more dynamic and accurate than data analytics

DATA PREPARATION

- Raw data in this challenge
- Needs insights, creativity and human knowledge
- It takes a lot of the working time

DATA QUALITY → GOOD RESULTS

DATA PREPARATION

- Main approach:
  - The grain in data is a trajectory, a user (hash) has n trajectories.
  - The dataset create for modeling stage has the grain in user.
- Trajectories become “columns”, like a time series forecasting.
- This approach means a sequence of trajectories

HASH	TRAJ_n	…	TRAJ_2	TRAJ_1
123	traj_123_n	…	traj_123_2	traj_123_1
456	traj_456_n	…	traj_456_2	traj_123_1

DATA PREPARATION

- Main approach:
  - To be more exact, a number k of window size is defined. The order is descending, representing the k last trajectories.
- The intention is to maintain the order of trajectories in relation to the last ones.

HASH	TRAJ_LAST(k)	TRAJ_{k-1}	...	TRAJ_1
123	traj_123_20	traj_123_19	...	traj_123_1
456	traj_456_2	traj_456_1	...	Null

DATA PREPARATION

Trajectory types:

- traj_last: Last trajectory, the exit point is unknown.
- traj_i: For all trajectories the points are known.

traj_i has more features than traj_last

DATA PREPARATION

Target:

Binary classification

From the last trajectory (after 15:00)

$$\begin{cases} 1, & 3750901.5068 \le x \le 3770901.5068 \\ & \text{and } -19268905.6133 \le y \le -19208905.6133 \\ 0, & \text{Otherwise} \end{cases}$$

DATA PREPARATION

Standard features from raw data:

- Vmax
- Vmin
- Vmean
- x_entry
- y_entry
- x_exit
- y_exit

DATA PREPARATION

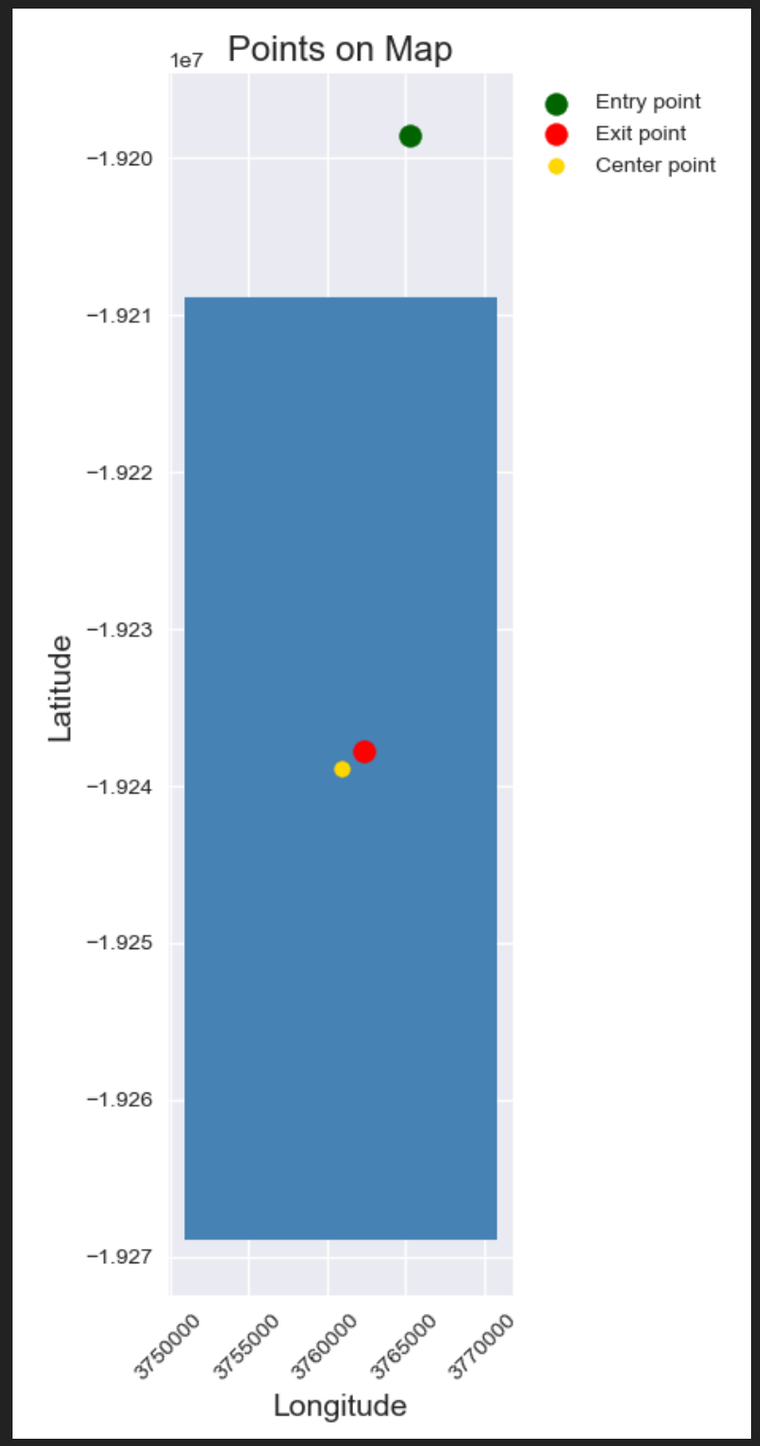
Time features:

- Duration of the trajectory
- Hour and minute
- Period of the day
  - night, early morning, morning, afternoon

DATA PREPARATION

Working with the points:

- Is the center?
- Travelled distance
- Distance from the center
- Distance from the boundary center
- Distance from the center point
- Approach to the center
- Approach to the center point



DATA PREPARATION

Velocity: My_Vmean

With the previously calculated variables:

Velocity mean = Distance ÷ Time

So,

My_Vmean = Travelled distance ÷ Duration of the trajectory

DATA PREPARATION

Aggregation features:

- Number of trajectories
- Average duration
- Average distance

DATA PREPARATION

Missing Values:

- Zero
- Median by group
- Predict Missing Value

MODELING

After the data preparation phase, the final dataset had 632 features

Mainly algorithms used:

- Extreme Gradient Boosting (xgboost)
- Light Gradient Boosting Machine (lgbm)
- Stacking (ensemble)

The final version was submitted with xgboost

FUTURE WORKS

Opportunities to improve results

- Use Cross Validation for measure better
- Tuning k parameter correctly
- Improve fine tuning and use optmization

EY
Building a better working world

THANK YOU!