Building Smarter Cities

from analyzing geolocation records

Jin Cong Ho <jincongho@gmail.com>

EY | Datascience

Overview

Part 1 Methodology for Building Prediction Model

Part 2 Use Cases

Understanding Data

Feature Engineering

Model Finetuning

Findings and Applications

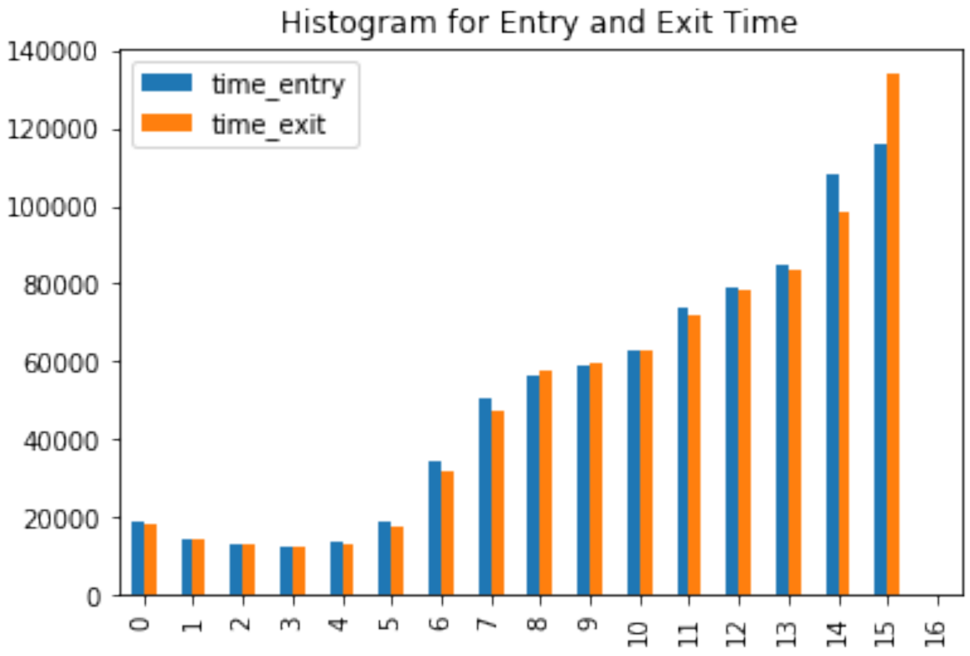
Part 1 Methodology for Building Prediction Model

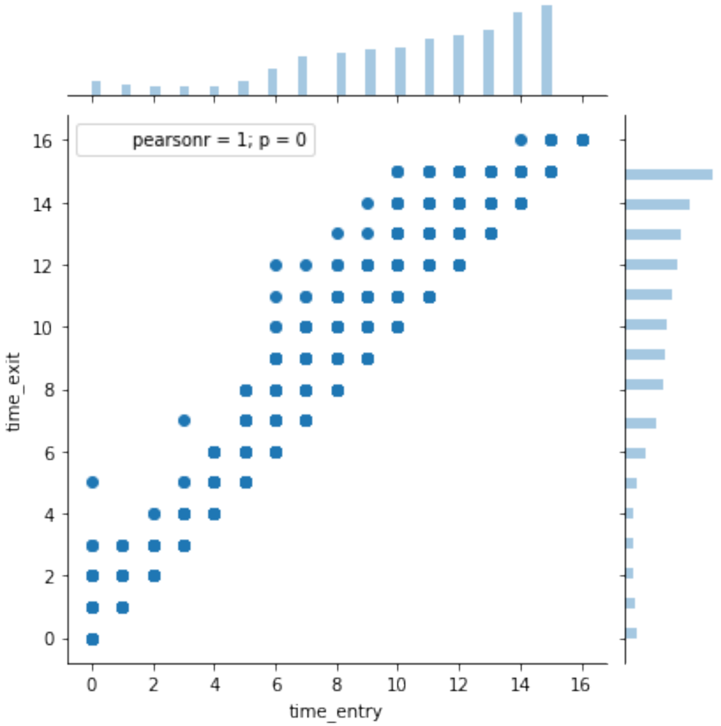
Setup: aws c4.8xlarge instance with 36 CPUs
Libraries: pandas, matplotlib, seaborn, xgboost

1 Understanding Data - Dataset

Features	Example	
hash	0000a8602cf2def930488dee7cdad104_1	Identifications
trajectory_id	traj_0000a8602cf2def930488dee7cdad104_1_0	
time_entry	15:00:32	Time
time_exit	15:29:48	
vmax	1.149404	Velocity
vmean	1.149404	
vmin	1.149404	
x_entry	3.749088e+06	Geolocation
y_entry	-1.926605e+07	
x_exit	3.749610e+06	
y_exit	-1.926594e+07	

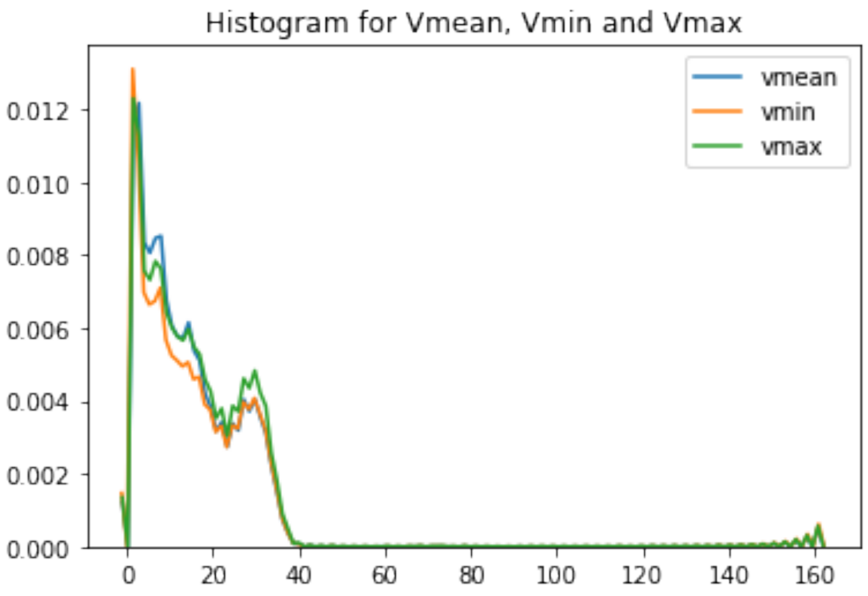
1 Understanding Data - Time

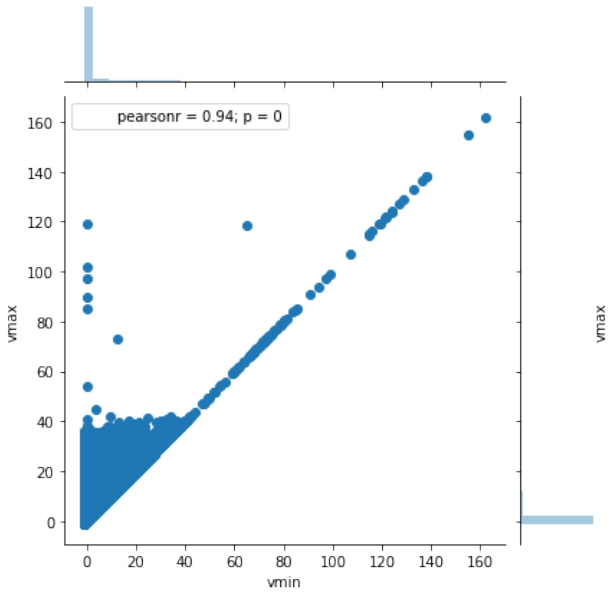


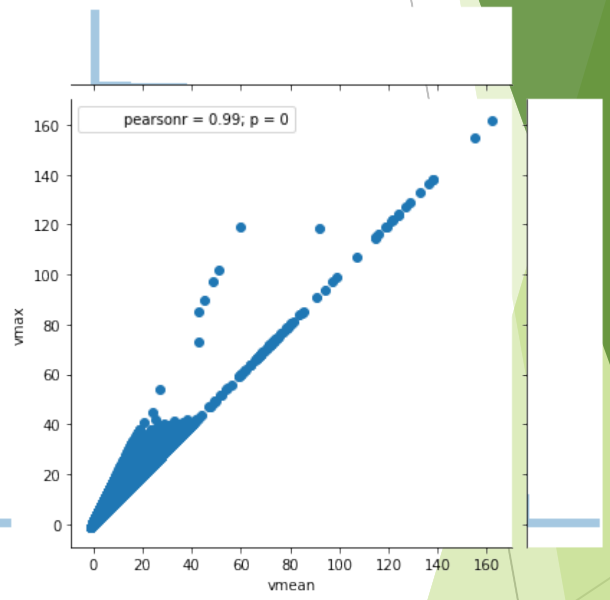


- The trend is going upward, more data in the afternoon
- Both histograms are very close (hour of entry is close to the hour of exit for each pt?)

1 Understanding Data - Velocity

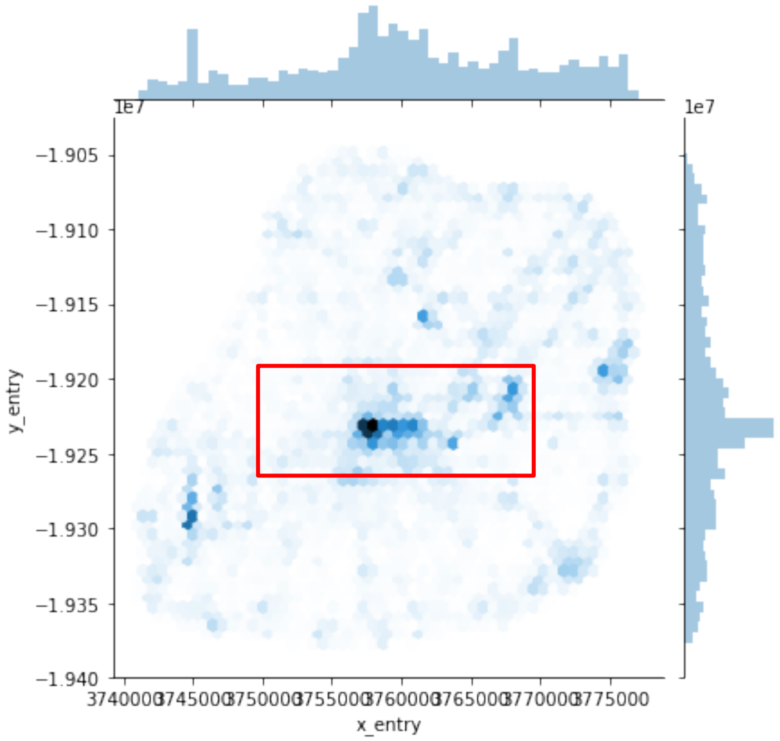




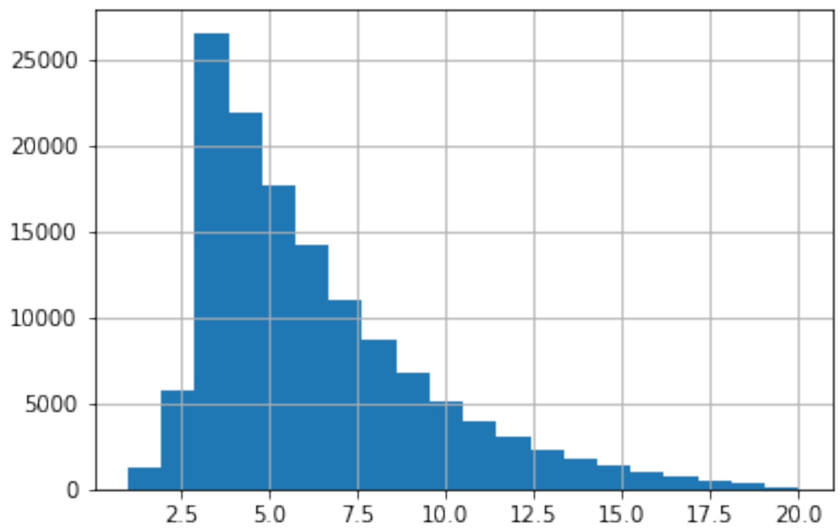


- Most trajectories have velocity < 20
- When vmin/vmean > 40, vmax will most likely be the same

1 Understanding Data - Geolocation



Histogram for Length of Journey of Each Device



- Maximum trajectories for a single device is 20

2 Feature Engineering

- Target = whether the device is in the city centre between 15:00-16:00
  - Modelling the problem as coordinate prediction ~75% accuracy
- Time
  - Duration = Time Exit – Time Entry (Seconds)
  - Time_Entry into entry_hour, entry_ minute, entry_seconds
  - Time_Exit into exit_hour, exit_ minute, exit_seconds
  - Entry/Exit_Quarter_Hour = *hour**4 + ceil(*minute*/15)
- Distance = Euclidean distance of entry and exit point
- Speed = Distance / Duration

93.60% Local Acc (Last Traj)

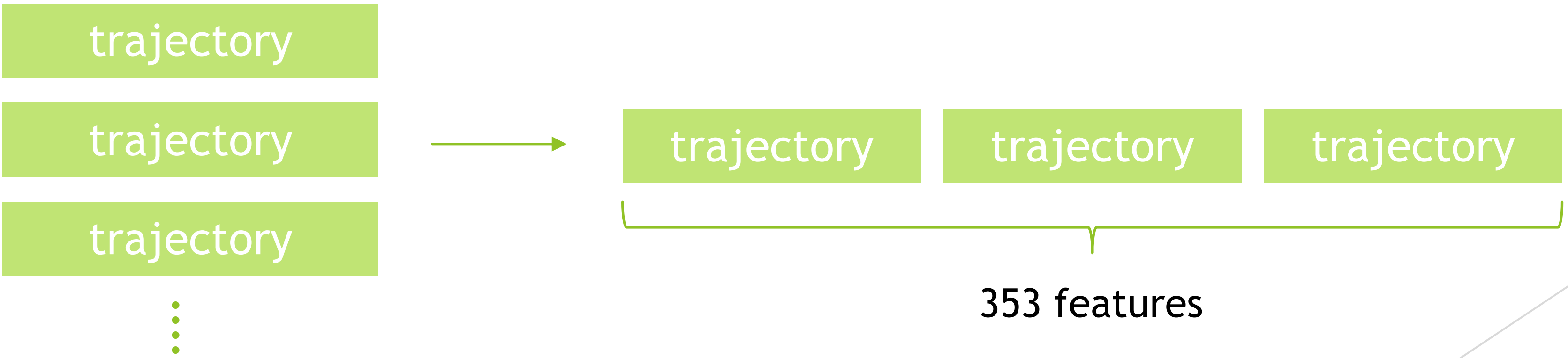
2 Feature Engineering

- Fill missing values as 0
- Concatenate trajectories in a day into a row

94.11% Local Acc

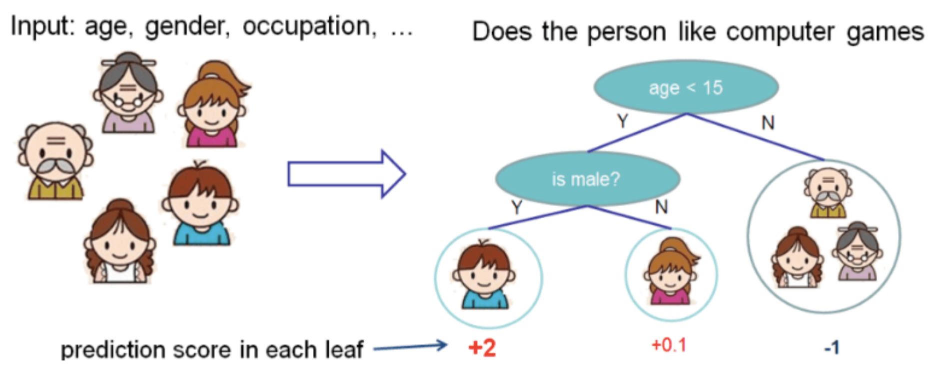
- Standardize all values by removing means and scaling to unit variance
  - z = (x - u) / s, where u = mean and s = std. deviation
- If a device has less than 20 trajectories, then we pad the last entry point as the entry and exit point of all previous trajectories

94.19% Local Acc



3 Model Finetuning - XGBoost

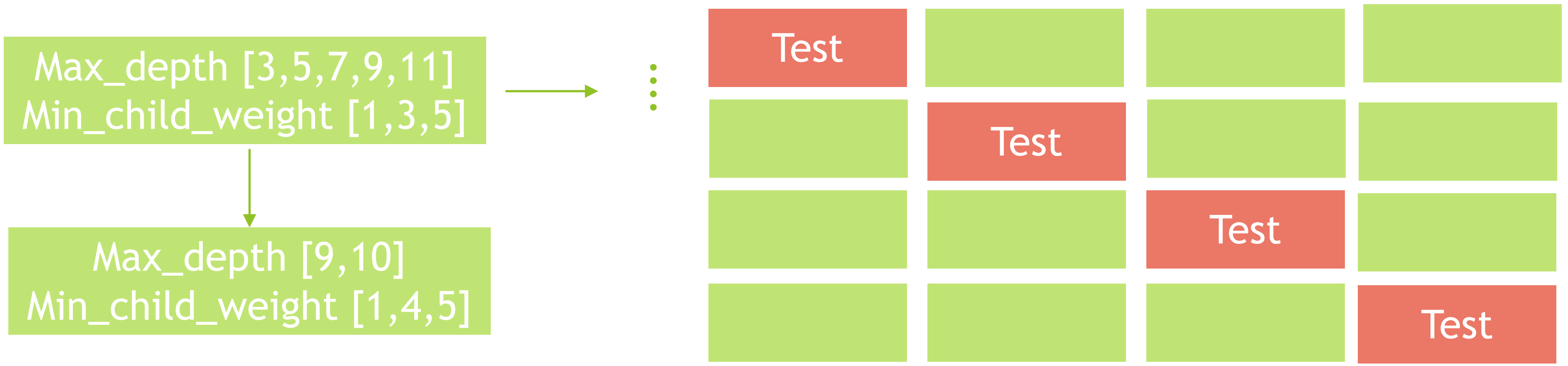
- **eXtreme Gradient Boosting**, uses the concept of **boosting** = an ensemble method of "weak" classifiers that sequentially adds predictors (trees) to corrects previous models
  - **Fast and Scalable:** parallelize tree construction using distributed computing
  - **High Explainability:** can calculate the importance of a feature by its impact to final prediction in tree



Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016.

3 Model Finetuning - XGBoost

- Grid search parameters from coarse to fine with 4-folds CV 94.25% Local Acc



- Improvement can be done by stacking more models, eg: Neural Network (92% acc)

3 Model Finetuning – Feature Importance

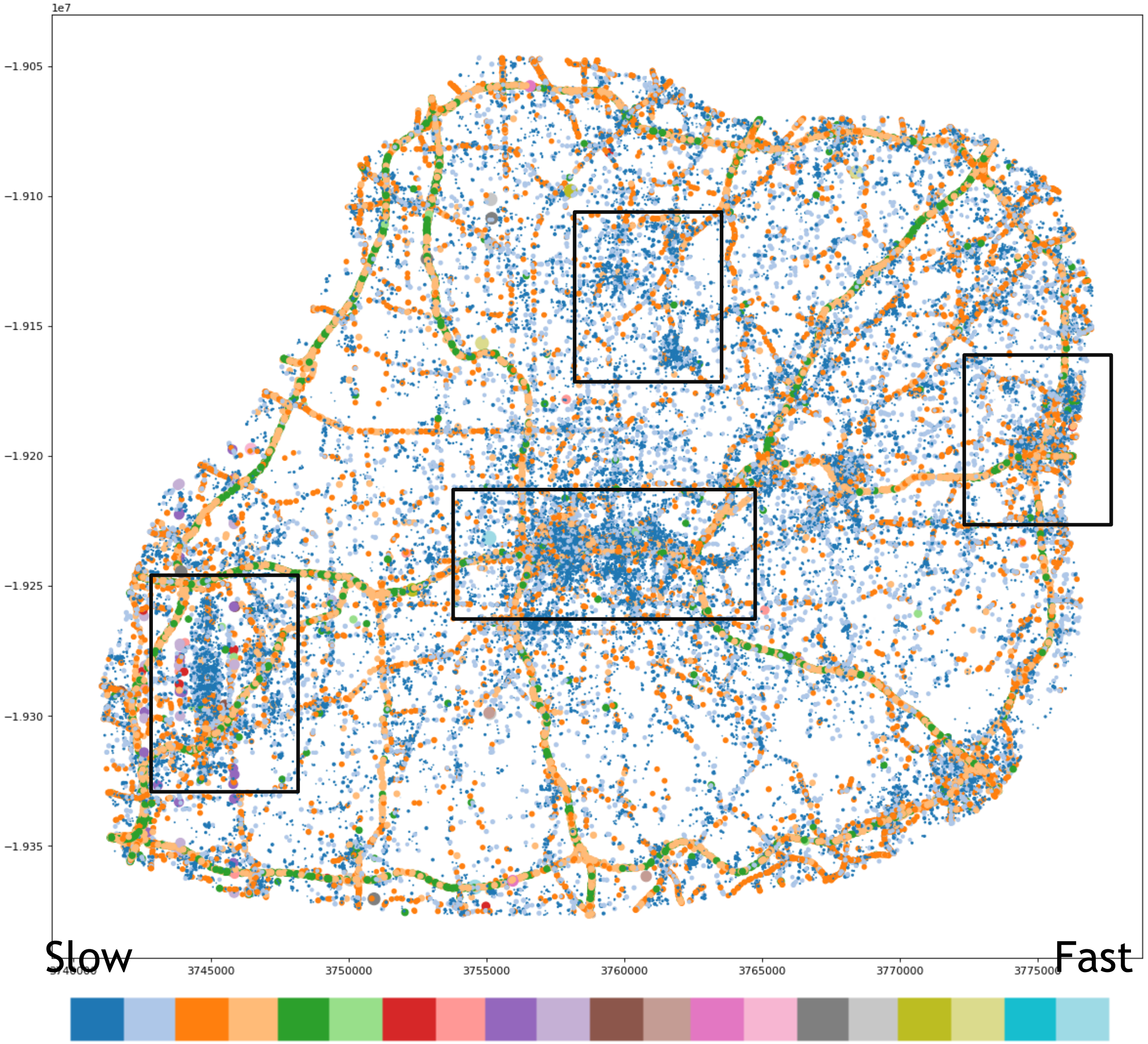
	Top 10 Features	Importance
1	y_entry_19	0.116769
2	x_entry_19	0.041110
3	duration_19	0.025248
4	entry_hour_19	0.024366
5	y_entry_2	0.019652
6	y_entry_18	0.019642
7	y_exit_18	0.014287
8	y_exit_1	0.014023
9	vmax_9	0.010111
10	x_entry_2	0.009191

- Traj18 & 19 contributed most and X & Y coordinates are the most useful features
- Quarter hour feature almost contributed 0%

Part 2 Use Cases

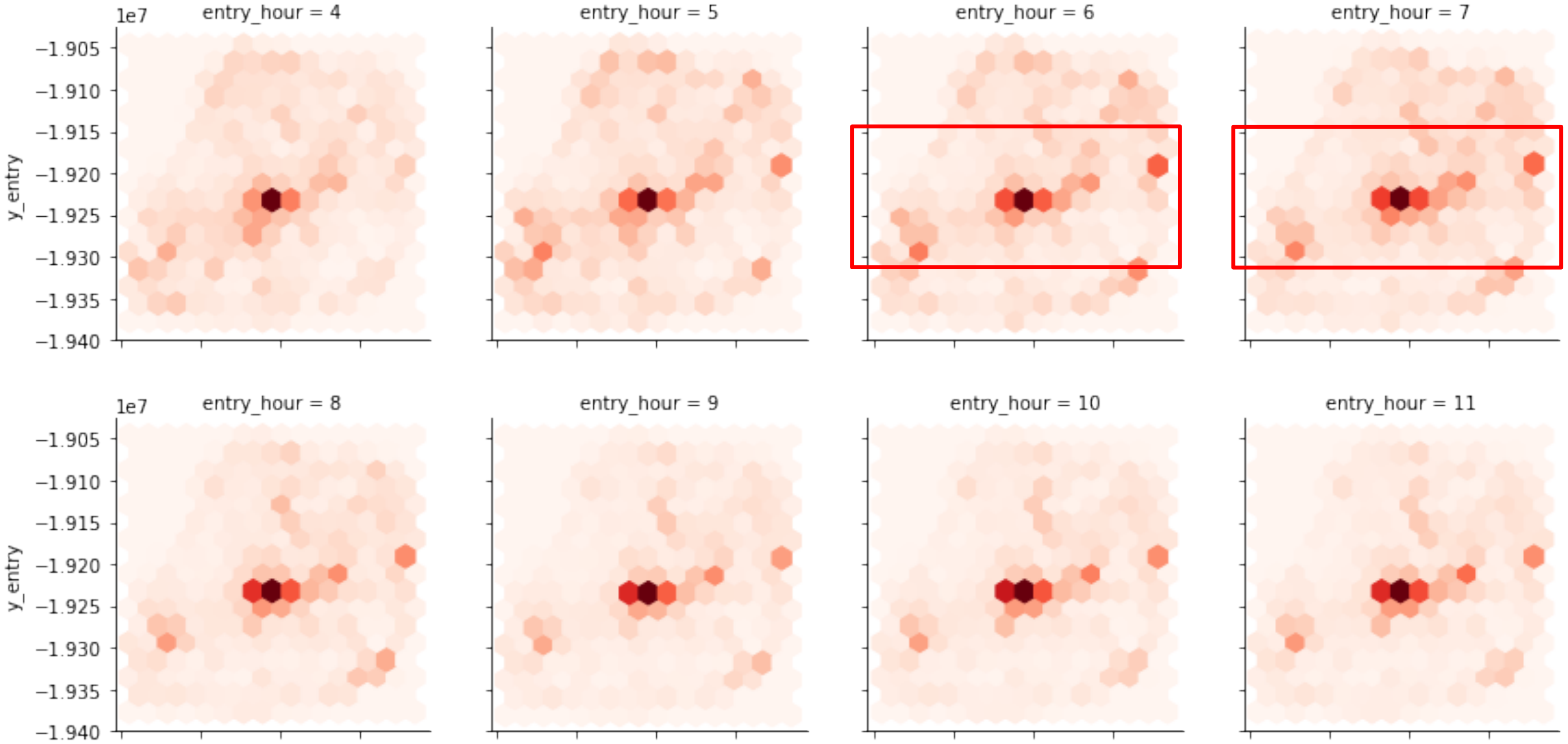
From findings to applications

Use Case 1: Finding Congested Area with Velocity



- Coordinates on the map with mean velocity as color

Use Case 2: Travel Demand by Area & Time



- Travel from the surroundings of the city centre increases between 6 am and 8 am from east and west direction

Applications

- Traffic Volume Analysis and Prediction (combine with weather data?)
  - congested area
  - travel demand by area & time
  - land use & investment decisions
  - infrastructure maintenance time & frequency
  - A data-driven approach to measuring the impact of new projects
- Other aspects of smart cities: pollution

Thank You! Any Question?