ANALYSIS AND VISUALIZATION OF THE FRENCH TRANSPORTATION NETWORK BASED ON MASSIVE PASSIVE DATA

FAB BIG DATA, SNCF
STAGNANCY POINTS & TRIPS DETECTION
USE CASE
Train trips history → Train delay prediction for coming week
USAGE N°1: USER EXPERIENCE

Departure on time

Arrival likely delay
USAGE N°2: DOMAIN EXPERT ANALYSIS
STAGNANCY POINTS & TRIPS DETECTION: DATA

- GPS Tracks
- Bookings
- Favorite places & stations
- Train trips history
- Roads
- Train Lines
- Stations

Reverse geocoding

Project

Stagnancy Points

Trips
GPS tracks:
- One point every 10-15 minutes
- 24/7 background acquisition
- 500 K users
- 50 M geolocations per day

Train travel history:
- 25 K travels per day

Maps:
- 1 M shapes of roads
- 1 M shapes of railway lines
- 5 K railway stations

API:
- 2 M requests per day

Jobs:
- Daily for J-2
DEALING WITH NOISY GPS TRACKS

Train Trips
Consolidation
Maps
GPS Tracks
Bookings
Trips

FAB BIG DATA SNCF, MAAS PROJECT, TRANSITDATA 2019
8 – 10/07/2019
TRIPS DETECTION ACCURACY

Test sample:
- 30 users during a few months
- About 600 travels
- Train & Road travels
  - Biased & Small sample

Long distance trains:
TGV, TER, Intercité, Eurostar, etc.

<table>
<thead>
<tr>
<th>Task</th>
<th>Score</th>
<th>Score Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did the user take the train?</td>
<td>93%</td>
<td>F1 Score</td>
</tr>
<tr>
<td>At which station?</td>
<td>83%</td>
<td>Précision</td>
</tr>
</tbody>
</table>

Short distance trains:
Transilien, RER:

<table>
<thead>
<tr>
<th>Task</th>
<th>Score</th>
<th>Score Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did the user take the train?</td>
<td>81%</td>
<td>F1 Score</td>
</tr>
<tr>
<td>At which station?</td>
<td>57%</td>
<td>Précision</td>
</tr>
</tbody>
</table>

Road:

<table>
<thead>
<tr>
<th>Task</th>
<th>Score</th>
<th>Score Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did the user take the road?</td>
<td>63%</td>
<td>F1 Score</td>
</tr>
</tbody>
</table>
DASHBOARDS FOR DOMAIN EXPERTS
DASHBOARDS FOR DOMAIN EXPERTS

FAB BIG DATA SNCF, MAAS PROJECT, TRANSITDATA 2019
Train trips history

Holidays

Weather predictions

Train delay prediction

- Binary classification
  Positive class: delay > 2 min 30 sec

- 7 next days
- New predictions every morning
  Retraining every week
ALGORITHM

Train travel history:
- Ile de France region
- 13 train lines, 500 stations
- ~50k circulations per day
- 30 months of data

Data set:
- Temporal windows
- Statistical features

Weather predictions:
- Next 3 days
- Every hour of every day
- On each train station

Modeling:
- Spark
- XGBoost

Convenient for big data but harder to inspect
PERFORMANCES

Test set consists of 40 entire days of operation unseen by the model

<table>
<thead>
<tr>
<th>Score</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>0.73</td>
</tr>
<tr>
<td>Precision</td>
<td>0.36</td>
</tr>
<tr>
<td>Recall</td>
<td>0.45</td>
</tr>
</tbody>
</table>

14% of positive class

Uncertainty due to next day delay unpredictability

Work days on Gare du Nord, line RER D
Line = median, Areas = interquartile
THANK YOU!