Estimation des retards des lignes de train en Pologne

Quentin Navarre, Nicolas Carval, Bruno Pincet, Edouard Bertrand

Objectif

Construire un modèle d'IA capable de mieux prédire en avance les retards que l'outil actuel

Aucune connaissance sur le fonctionnement du modèle de prédiction actuel

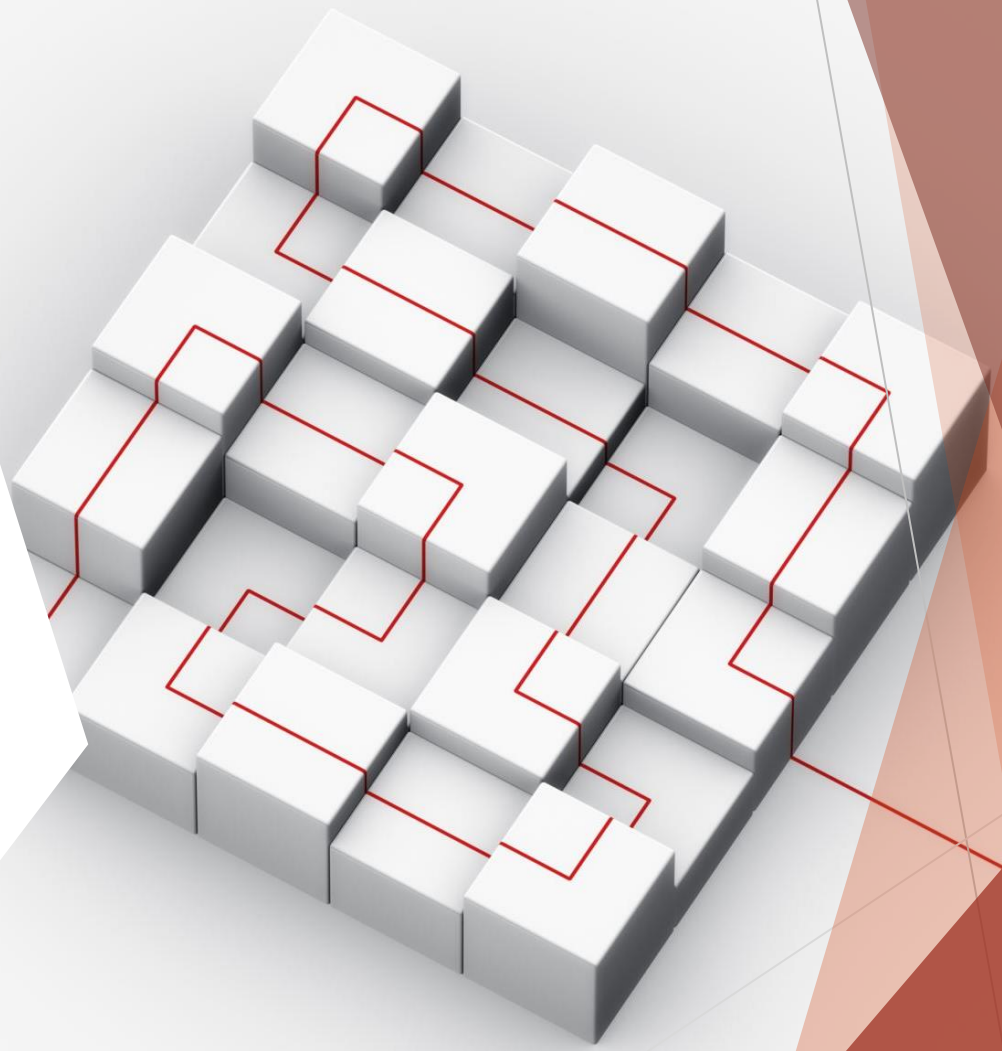
Non-connaissance de l’arrivée exacte du train

Historique de 2 semaines seulement

Non utilisation des prédictions intermédiaires du prestataire actuel

Etapes projet

- Mise en forme et nettoyage des données
- Feature preprocessing
- Choix de la loss function
- Choix et construction du modèle
- Comparaison de nos résultats avec le modèle de prédiction actuel



Mise en forme des données

Création d'un identifiant unique des trains

- Utilisation d'un regex afin d'extraire les différentes parties de l'id
- Similarité des voyages en se basant sur la première partie

Identification de cycles de voyages

- Parcours d'un train pouvant varier même si celui-ci conserve le même id
- Cas d'arrêts du scraping avant la gare d'arrivée

Retard au passage de la gare

- Identifié en utilisant la dernière ligne du dataset avant changement de gare
- On part du principe qu'elle correspond au retard réel

Mise en forme des données

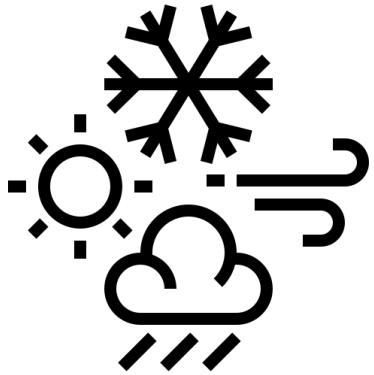
- Conserver seulement les informations d'un train au moment où celui-ci est en gare
  - Id_train
  - Transporteur
  - Date_depart
  - Depart_destination
  - Arrivee_prevue
  - Nom_station

Carte des arrivées et départs

Feature preprocessing

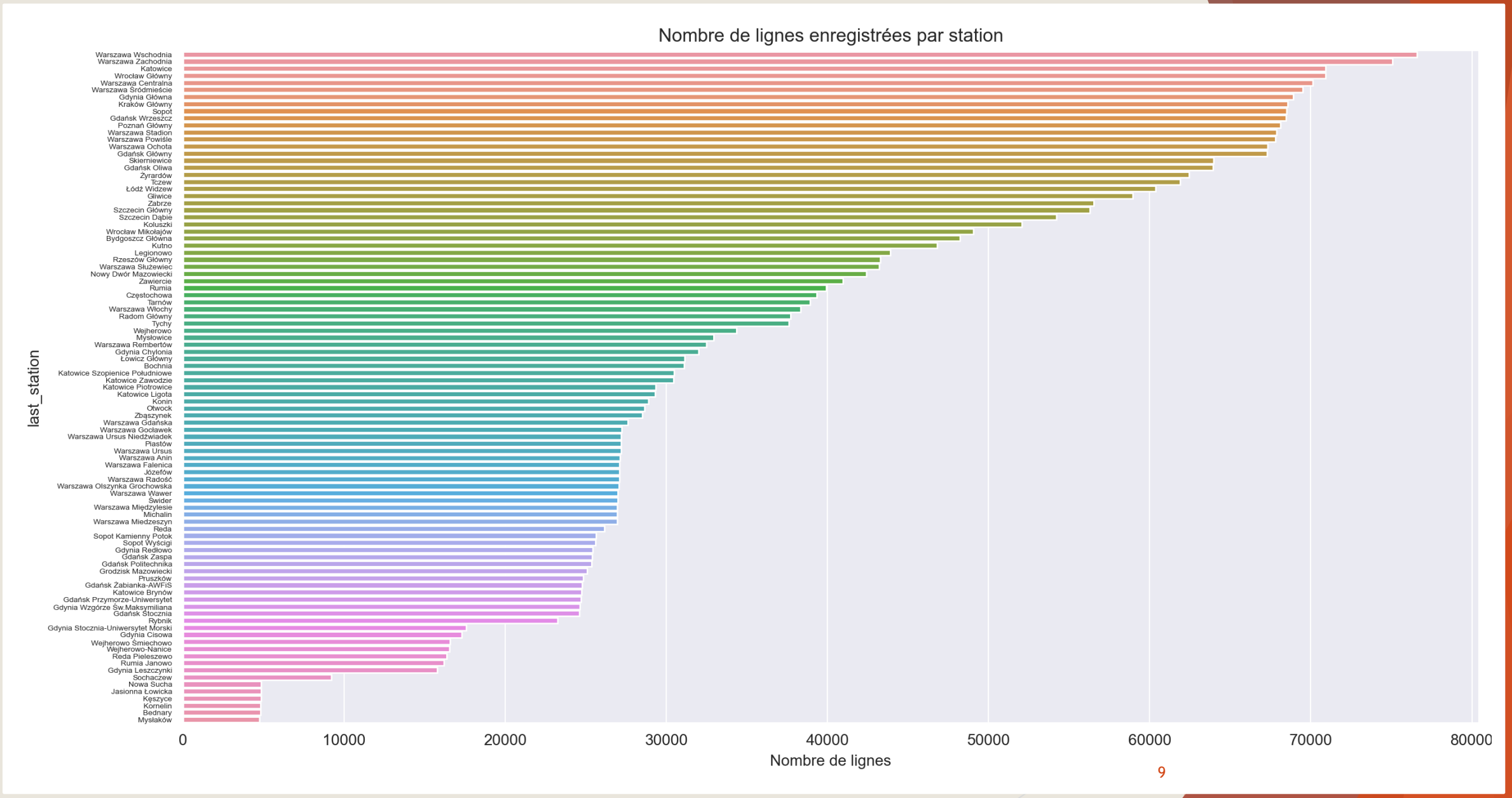
- Format des données :
  - Informations propres au train
    - Identifiant du train
    - Nom du transporteur
  - Informations propres à la station de prédiction
    - Nom de la station
    - Heure d'arrivée prévue
    - Météo
    - Localisation
    - Nombre de trains ayant traversé la station pendant l'heure et la journée
  - Informations sur la dernière station parcourue
    - Nom de la station
    - Heure d'arrivée prévue
    - Météo
    - Localisation
    - Nombre de trains ayant traversé la station pendant l'heure et la journée
    - Retard
  - Informations relatives au parcours restant
    - Durée du trajet entre les deux stations
    - Distance entre les deux stations

Feature preprocessing



- Extraction départ – destination
  - Utilisation d'API pour obtenir la longitude et la latitude
  - Utilisation d'API pour obtenir la météo de la journée à cette localisation
  - Si données indisponibles , météo de la grande ville la plus proche

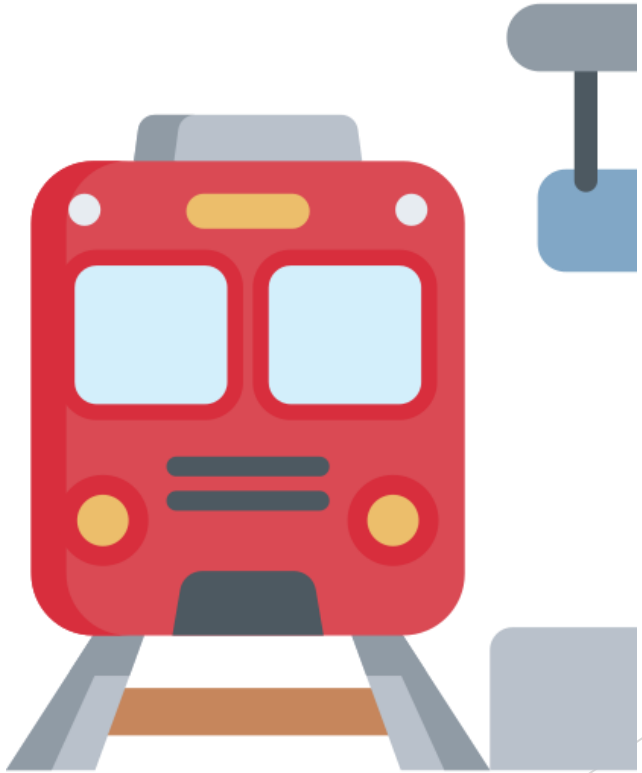
time	tavg	tmin	tmax	prcp	snow	wdir	wspd	wpgt	pres	tsun	gare
2022-05-16	15.6	4.9	24.5	0.0	0.0	68.0	4.7	14.8	1019.7	0.0	Opole Wschodnie
2022-05-17	15.7	11.5	18.5	0.0	0.0	24.0	6.8	22.2	1018.9	0.0	Opole Wschodnie
2022-05-18	13.0	6.2	19.6	0.0	0.0	74.0	6.5	16.7	1027.9	0.0	Opole Wschodnie
2022-05-19	16.0	3.3	25.5	0.0	0.0	155.0	9.8	27.8	1024.9	0.0	Opole Wschodnie
2022-05-20	21.6	14.3	28.4	0.0	0.0	217.0	8.9	29.6	1018.3	0.0	Opole Wschodnie

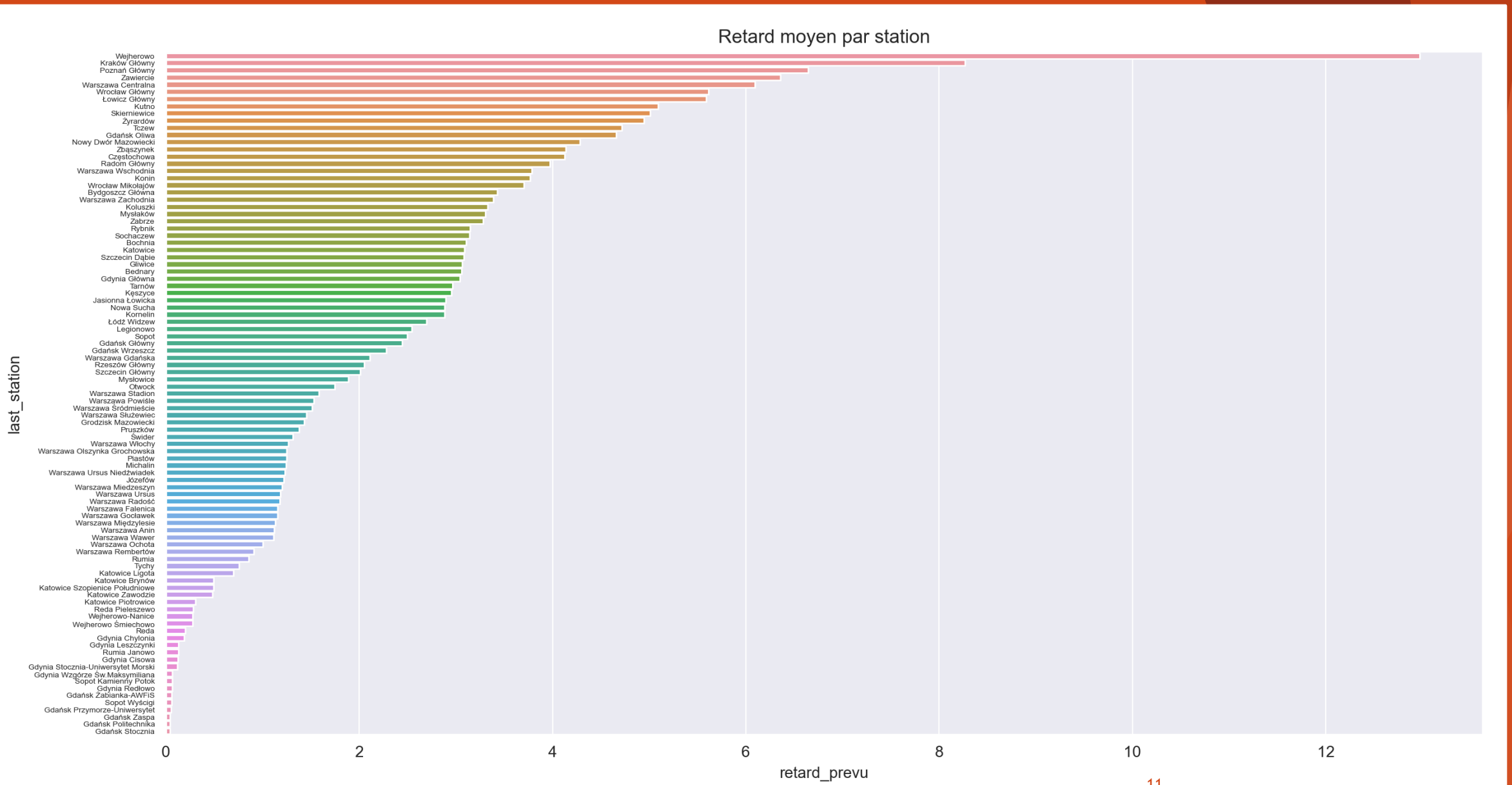

Nombre de lignes enregistrées par station
last_station
Warszawa Wschodnia
Warszawa Zachodnia
Katowice
Wrocław Główny
Warszawa Centralna
Warszawa Śródmieście
Gdynia Główna
Kraków Główny
Sopot
Gdańsk Wrzeszcz
Poznań Główny
Warszawa Stadion
Warszawa Powiśle
Warszawa Ochota
Gdańsk Główny
Skierniewice
Gdańsk Oliwa
Żyrardów
Tczew
Łódź Widzew
Gliwice
Zabrze
Szczecin Główny
Szczecin Dąbie
Koluszki
Wrocław Mikołajów
Bydgoszcz Główna
Kutno
Legionowo
Rzeszów Główny
Warszawa Służewiec
Nowy Dwór Mazowiecki
Zawiercie
Rumia
Częstochowa
Tarnów
Warszawa Włochy
Radom Główny
Tychy
Wejherowo
Mysłowice
Warszawa Rembertów
Gdynia Chylonia
Łowicz Główny
Bochnia
Katowice Szopienice Południowe
Katowice Zawodzie
Katowice Piotrowice
Katowice Ligota
Konin
Otwock
Zbąszynek
Warszawa Gdańska
Warszawa Gocławek
Warszawa Ursus Niedźwiadek
Piastów
Warszawa Ursus
Warszawa Anin
Warszawa Falenica
Józefów
Warszawa Radość
Warszawa Olszynka Grochowska
Warszawa Wawer
Świder
Warszawa Międzylesie
Michalin
Warszawa Miedzeszyn
Reda
Sopot Kamienny Potok
Sopot Wyścigi
Gdynia Redłowo
Gdańsk Zaspa
Gdańsk Politechnika
Grodzisk Mazowiecki
Pruszków
Gdańsk Żabianka-AWFiS
Katowice Brynów
Gdańsk Przymorze-Uniwersytet
Gdynia Wzgórze Św.Maksymiliana
Gdańsk Stocznia
Rybnik
Gdynia Stocznia-Uniwersytet Morski
Gdynia Cisowa
Wejherowo Śmiechowo
Wejherowo-Nanice
Reda Pieleszewo
Rumia Janowo
Gdynia Leszczynki
Sochaczew
Nowa Sucha
Jasionna Łowicka
Kęszyce
Kornelin
Bednary
Mysłaków
0
10000
20000
30000
40000
50000
60000
70000
80000
Nombre de lignes

Feature preprocessing

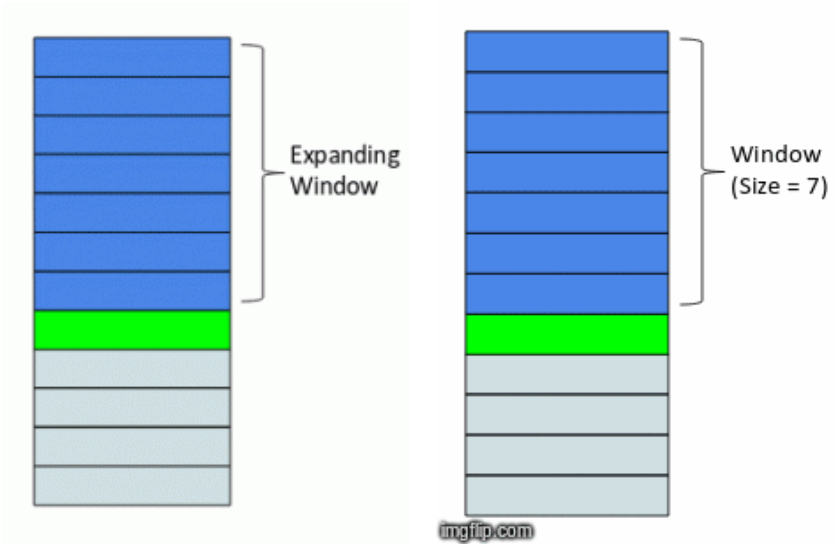
- Identification des 3 gares précédents la gare concernée du trajet
  - Retard dans chacune de ces gares
  - Distance km depuis ces gares (vol d'oiseau)
  - Distance temporelle depuis ces gares

- Gare concernée
  - Nombre de passage de train dans cette gare depuis le début de la journée
  - Nombre de passage de train dans cette gare depuis le début de l'heure
  - Nombre de passage de ce train dans cette gare depuis le début de la journée




Retard moyen par station
last_station
Wejherowo
Kraków Główny
Poznań Główny
Zawiercie
Warszawa Centralna
Wrocław Główny
Łowicz Główny
Kutno
Skierniewice
Żyrardów
Tczew
Gdańsk Oliwa
Nowy Dwór Mazowiecki
Zbąszynek
Częstochowa
Radom Główny
Warszawa Wschodnia
Konin
Wrocław Mikołajów
Bydgoszcz Główna
Warszawa Zachodnia
Koluszki
Mysłaków
Zabrze
Rybnik
Sochaczew
Bochnia
Katowice
Szczecin Dąbie
Gliwice
Bednary
Gdynia Główna
Tarnów
Kęszyce
Jasionna Łowicka
Nowa Sucha
Kornelin
Łódź Widzew
Legionowo
Sopot
Gdańsk Główny
Gdańsk Wrzeszcz
Warszawa Gdańska
Rzeszów Główny
Szczecin Główny
Mysłowice
Otwock
Warszawa Stadion
Warszawa Powiśle
Warszawa Śródmieście
Warszawa Służewiec
Grodzisk Mazowiecki
Pruszków
Świder
Warszawa Włochy
Warszawa Olszynka Grochowska
Piastów
Michalin
Warszawa Ursus Niedźwiadek
Józefów
Warszawa Miedzeszyn
Warszawa Ursus
Warszawa Radość
Warszawa Falenica
Warszawa Gocławek
Warszawa Międzylesie
Warszawa Anin
Warszawa Wawer
Warszawa Ochota
Warszawa Rembertów
Rumia
Tychy
Katowice Ligota
Katowice Brynów
Katowice Szopienice Południowe
Katowice Zawodzie
Katowice Piotrowice
Reda Pieleszewo
Wejherowo-Nanice
Wejherowo Śmiechowo
Reda
Gdynia Chylonia
Gdynia Leszczynki
Rumia Janowo
Gdynia Cisowa
Gdynia Stocznia-Uniwersytet Morski
Gdynia Wzgórze Św.Maksymiliana
Sopot Kamienny Potok
Gdynia Redłowo
Gdańsk Żabianka-AWFiS
Sopot Wyścigi
Gdańsk Przymorze-Uniwersytet
Gdańsk Zaspa
Gdańsk Politechnika
Gdańsk Stocznia
0
2
4
6
8
10
12
retard_prevu

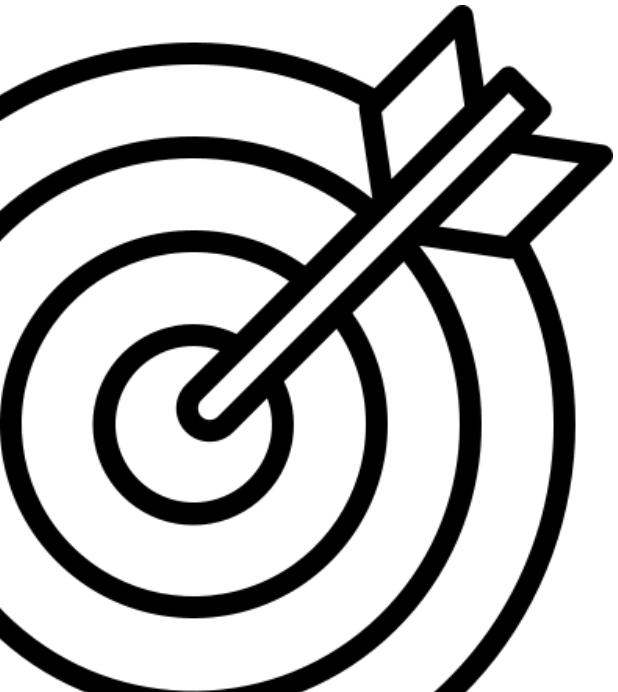
Feature preprocessing



- Retard Moyen dans la gare concernée
  - Expanding windows (moyenne sur historique total)
  - Rolling windows (moyenne locale)

- Heure
- Minute
- Jour de la semaine
- Numéro du jour dans le mois

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$



Loss function

- MAE : moyenne des écarts absolus entre nos prédictions et les retards réels
- Résultat que l'on souhaite minimiser

Stratégie 1 ▸ Prédiction du retard une station à l'avance

Prédictions très précises:

- Information très récente
- Peu de changements d'une station à une autre

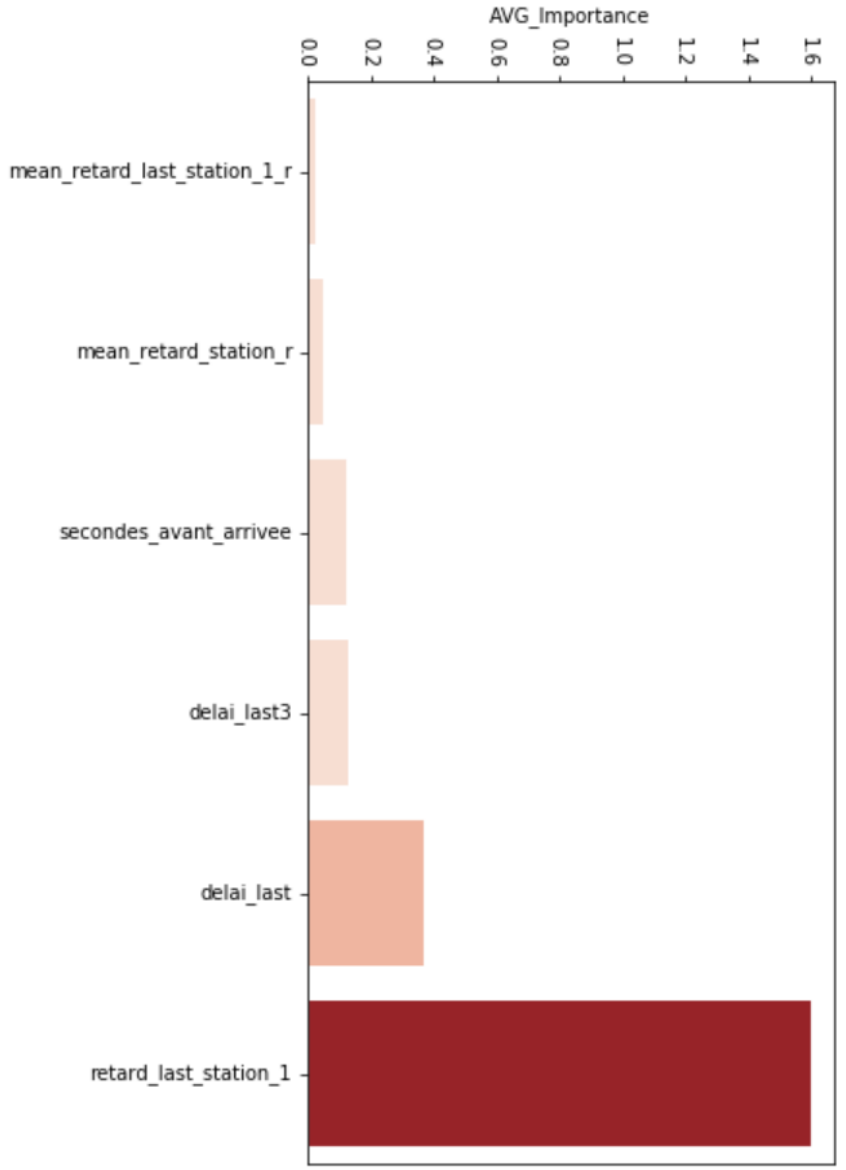
Prédictions parfois tardives:

- Prochaine station dans quelques minutes
- Trajets courts

Résultats Stratégie 1

Gradient Boosting

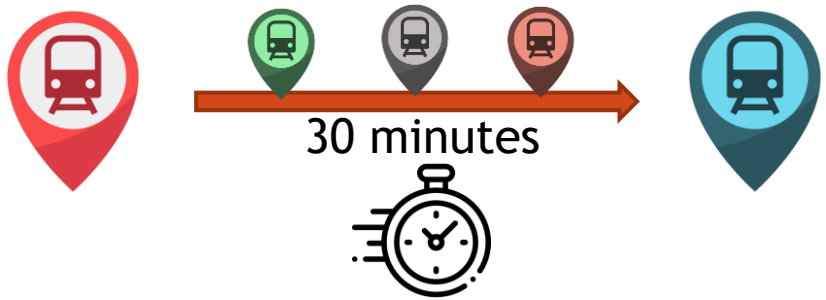
Random Forest



- Gradient boosting
  MAE : 1.3971694139680957
  Accuracy 0.783970508254172

- Random Forest
  MAE : 1.002044627400733
  Accuracy : 0.755176315036189 9

Stratégie 2 ▸ Prédiction du retard 30 minutes à l'avance



Prédiction à l'avance pour le client:

- Possibilité de généraliser à d'autres fenêtres de temps
- Plus proche du système actuel

Prédictions moins précises à court terme:

- Moins d'informations sur le trajet en cours

Résultats Stratégie 2

Deep Learning

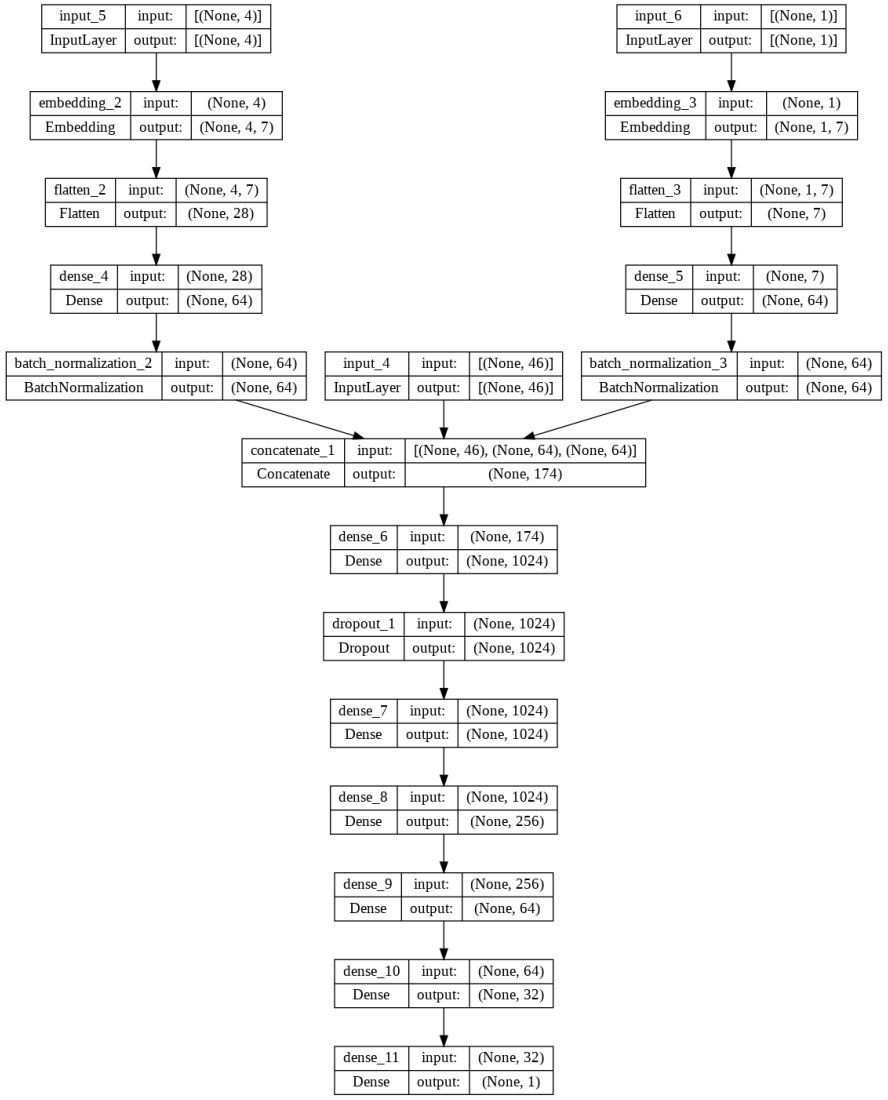
DataFrame de 200 000 lignes

Split train-test de 75-25

51 features

```
X_train shape :  (158365, 51)
y_train shape :  (158365,)
X_test shape :  (52789, 51)
y_test shape :  (52789,)
```

```
Total params: 1,530,861
Trainable params: 1,530,605
Non-trainable params: 256
```



Résultats à 30 minutes

```
model.evaluate([X_test_rest_scaled,X_test_e1,X_test_e2], y_test)
```

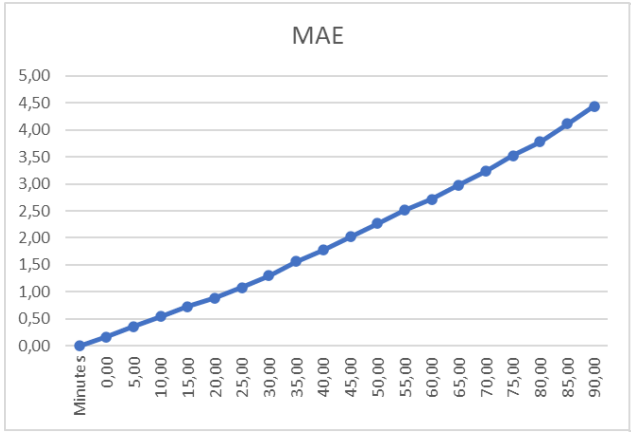
```
1650/1650 [==============================] - 5s 3ms/step - loss: 1.1887 - mse: 30.3691
[1.1886541843414307, 30.36908721923828]
```

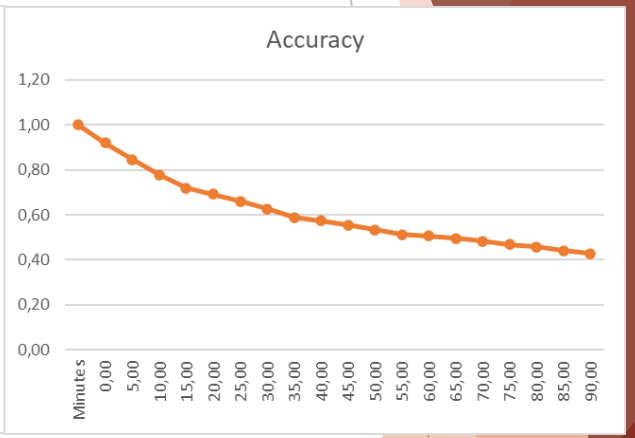
Baseline - Résultats de l'ancien prestataire

- Données enregistrées toutes les 5 minutes pour chaque station à partir du départ du train.

  MAE : 3.57624923918526 56

  Accuracy : 0.7920202173032963





- En prévoyant 5 * *n* minutes à l'avance :

Minutes	MAE IA	MAE Pologne	Accuracy	Minutes	MAE IA	MAE Pologne	Accuracy
0		0.00	1.00	50.00	1.15	2.02	0.55
5		0.17	0.92	55.00		2.27	0.53
10		0.36	0.85	60.00	1.21	2.51	0.51
15		0.54	0.78	65.00		2.72	0.51
20		0.72	0.72	70.00		2.98	0.50
25		0.88	0.69	75.00		3.23	0.48
30	1.18	1.08	0.66	80.00		3.53	0.47
35		1.30	0.63	85.00		3.78	0.46
40	1.21	1.56	0.59	90.00		4.11	0.44
45		1.78	0.57	95.00		4.44	0.43

Pistes d'amélioration

- Feature preprocessing
  - Ajout de variables :
    - Données sur les stations parcourues
    - Nombre de stations parcourues
- Modèles :
  - LSTM
  - Optimisation des hypers paramètres
- Analyse des erreurs