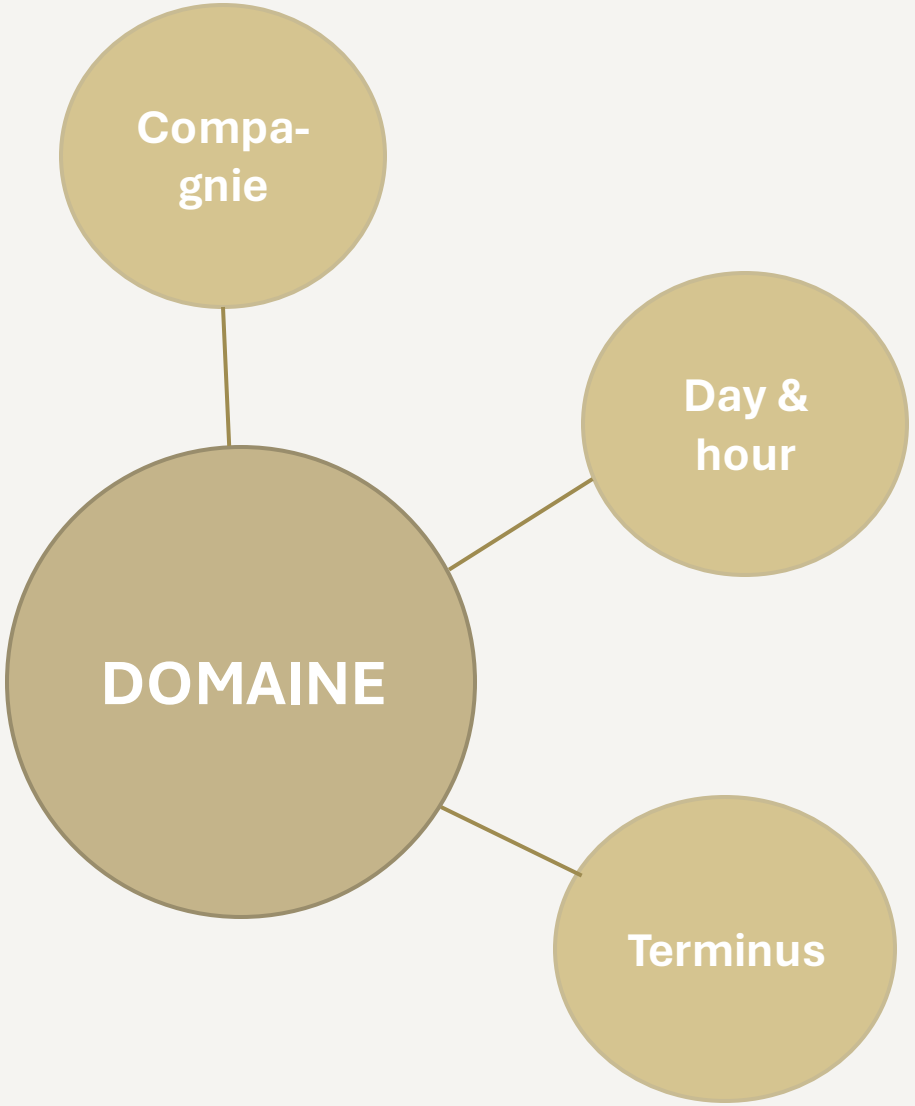
Présentation sprint : Compagnie de trains polonaise

PAR MAUD TISSOT, ARTHUR SCELLES, ALEXANDRE SMADJA, AURÉLIEN POUXVIEL

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Peut-on prévoir un jour à l'avance le retard d'un train ?



1. Traduction du problème en code

Après traitement des données nous avons donc :
- 241 916 lignes
- 24 variables

TARGET

- retard OUI/NON si >10min

PREPROCESS

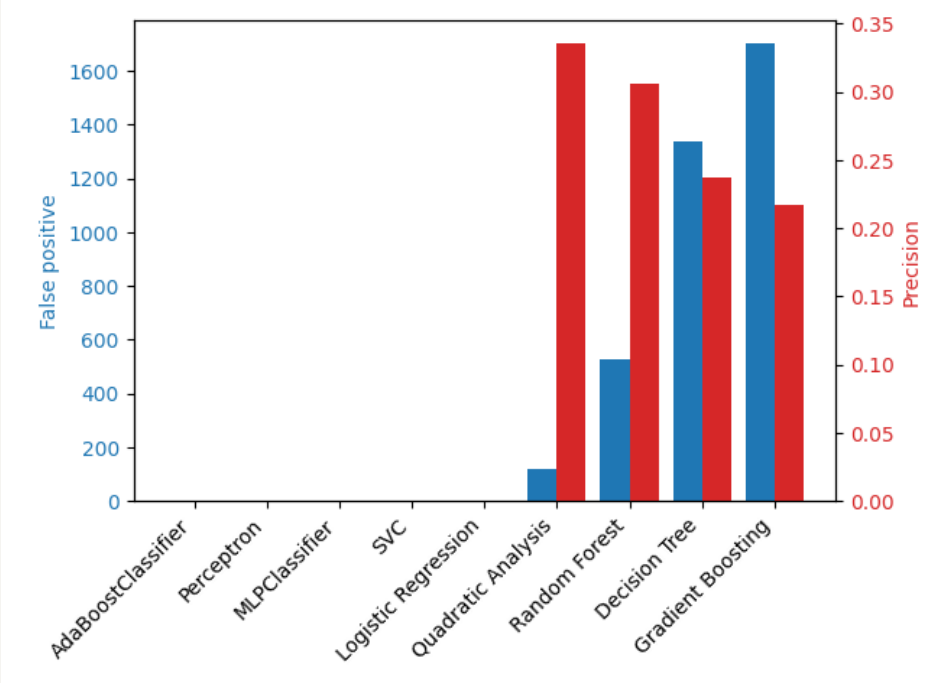
- Filtre 1 trajet = 1 ligne
- New features

TRAIN / TEST SLIT

- Test = 1 jour
- Train = dataset avant j-2

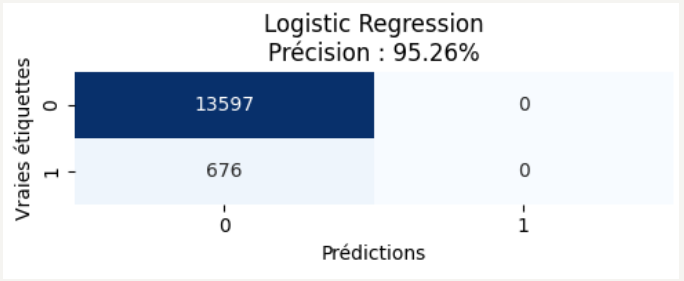
2. Prédiction du retard un jour à l'avance

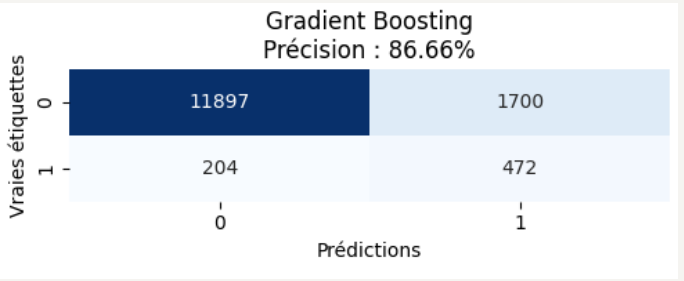
- Nous avons essayé différents modèles de différents types

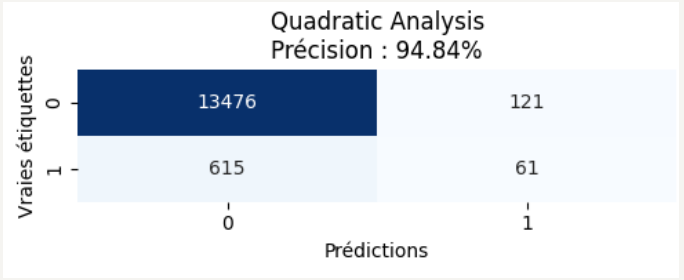


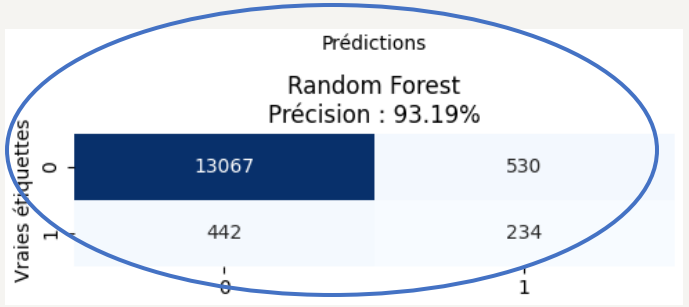
$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

- Choix du modèle





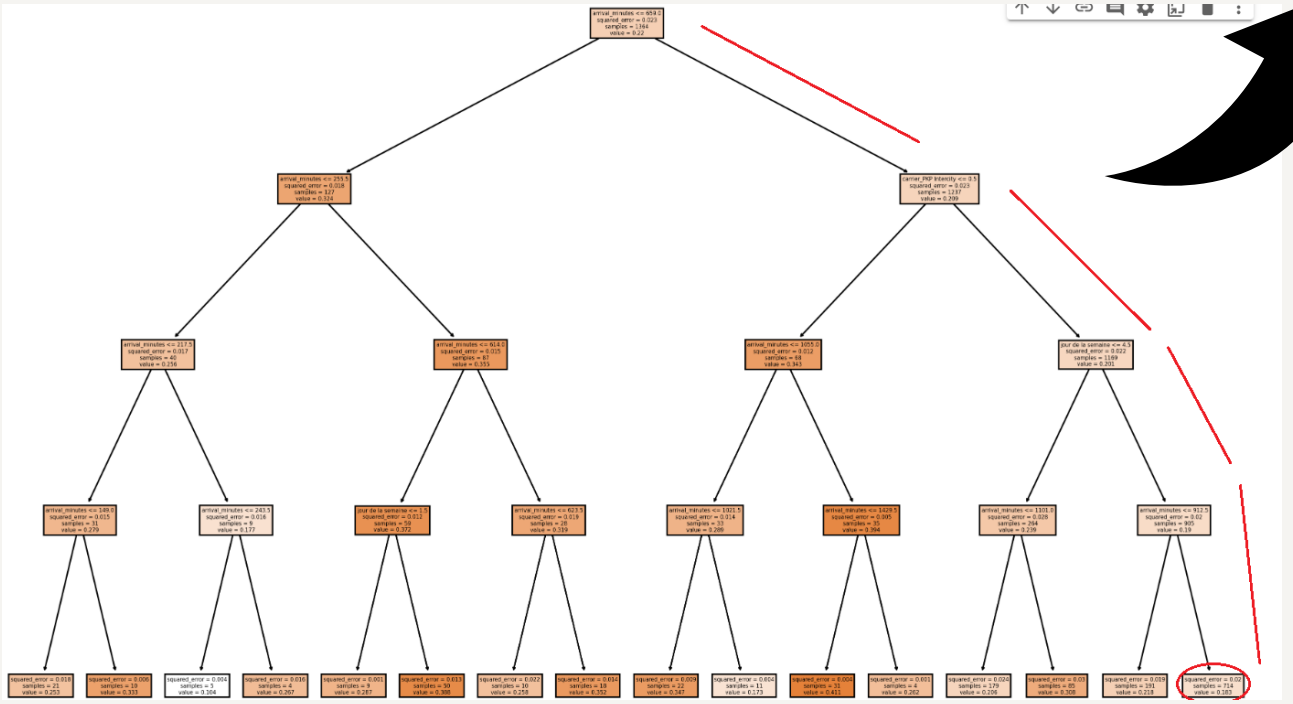


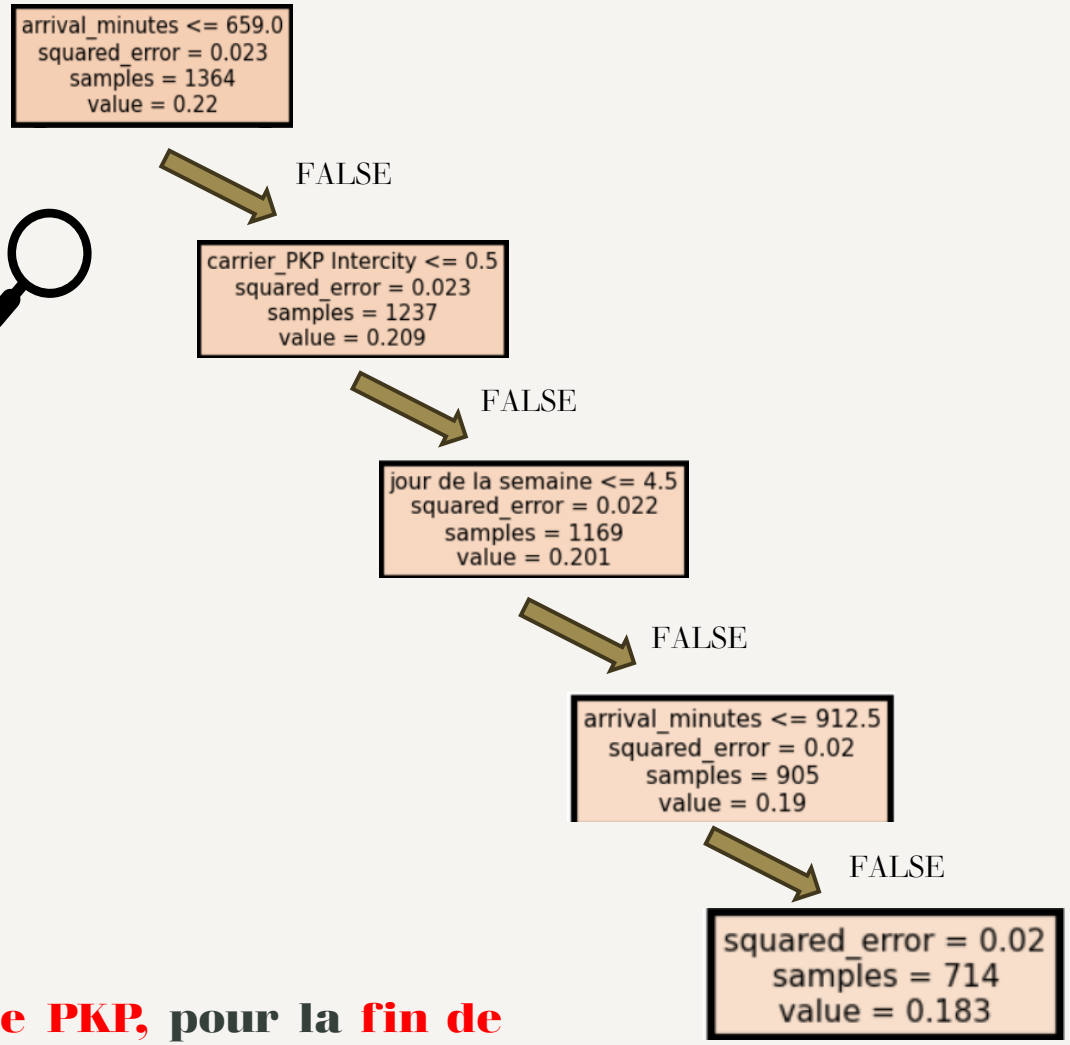


Justement, on ne se fie pas aveuglément à la precision et on ne prend pas de modèle 'naif'

3. Arbre 1 - Identifier les domaines pour lequels le retard est prévisible

- Calcul de l'erreur
- 1ere arbre sur l'horaire et les jours = Périodicité des trains

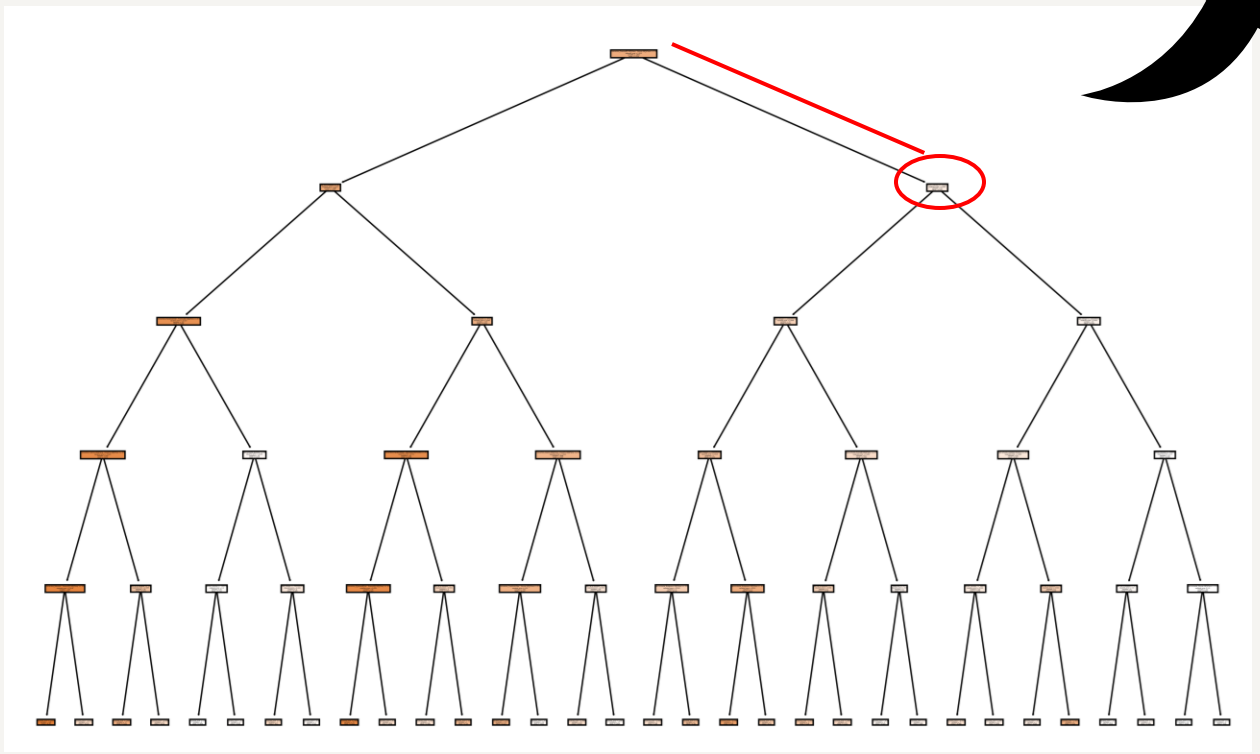


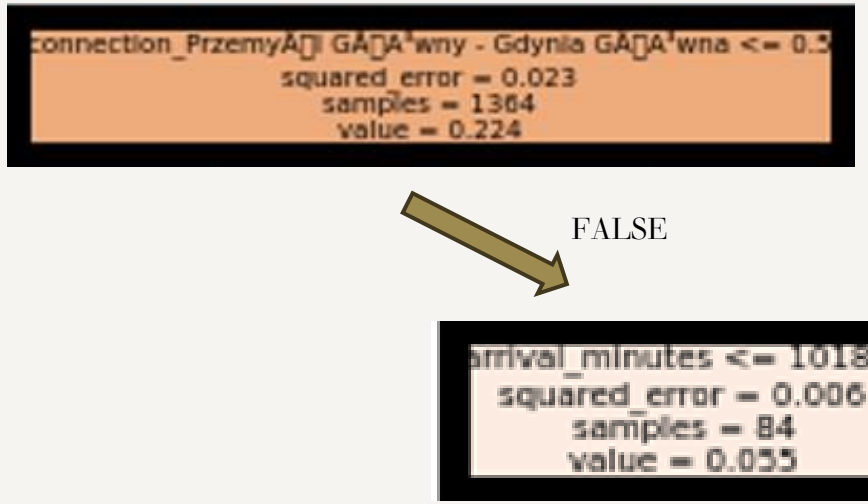


= **Trains à partir de 15 h de PKP, pour la fin de semaine (vendredi, samedi, dimanche). 70% de précision (714 / 1100 retards)**

3. Arbre 2 - Géographie (gares, connections)

- Trop de profondeurs = pas toujours pertinent





= **Trains de la connexion Przemyśl Główny - Gdynia Główna**

202 trains.
97 retards
84 prédits sur 97 = 86% de précision

Conclusion

Prédiction large

Nous sommes capables par exemple de proposer à une entreprise comme PKP intercity de prédire leurs retards en fin de journée 24h à l'avance avec une précision de 70%

Proposition d'un domaine bien plus précis

La ligne **Przemyśl Główny - Gdynia Główna :**

- 202 trajets en une semaine
- 1 train sur 2 est en retard

Nous proposons de prédire 24h à l'avance les retards de cette ligne avec une précision de 86%

Critiques et perspectives

Augmenter les données (spécifiques à la date, 95% de train à l'heure)

Variabilité des informations (bruit)

Visualisations des données et interprétations (faux positifs et faux négatifs, métriques etc...)

Comparaison des domaines identifiés selon les modèles

Gestion des données

Nouveaux modèles

Analyse des domaines

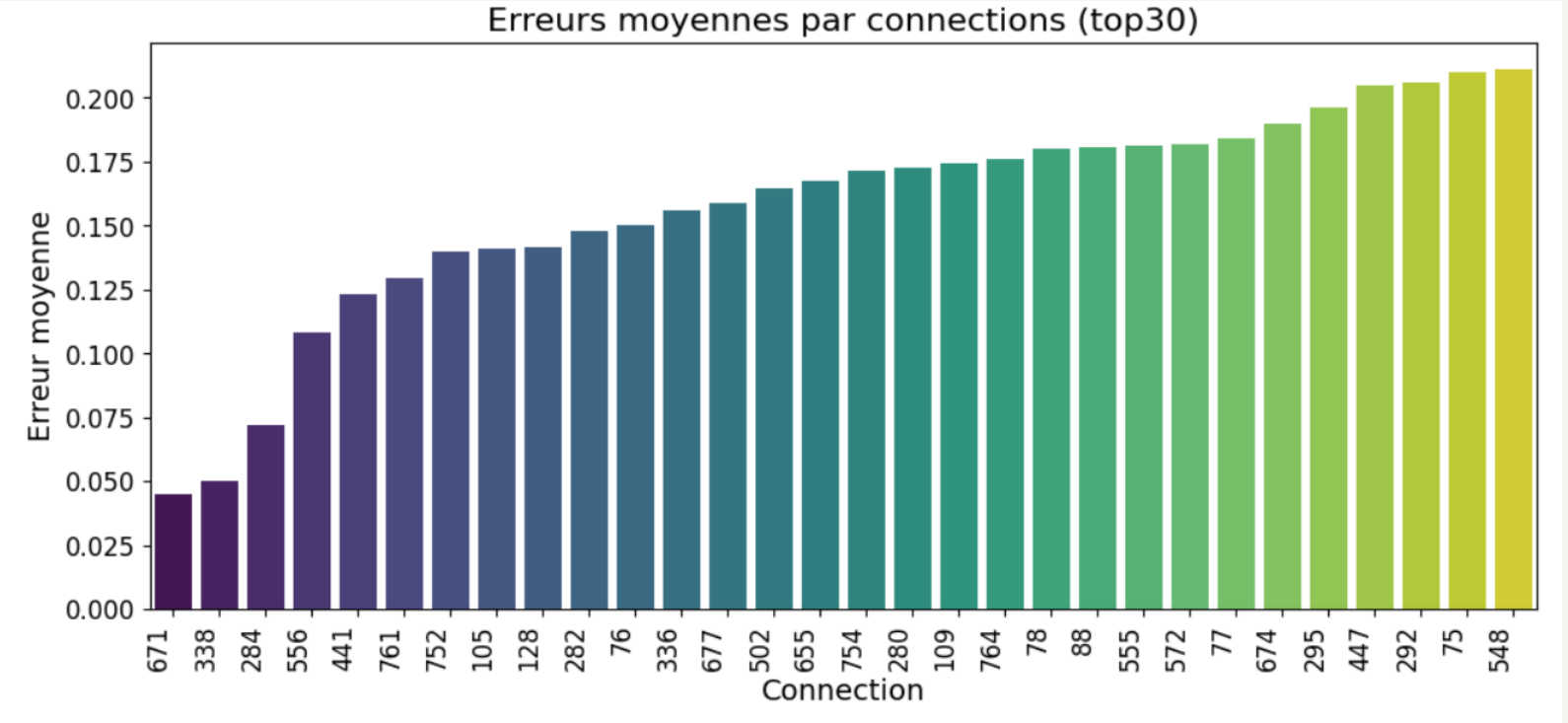
Prediction du retard (features, clustering, regression)

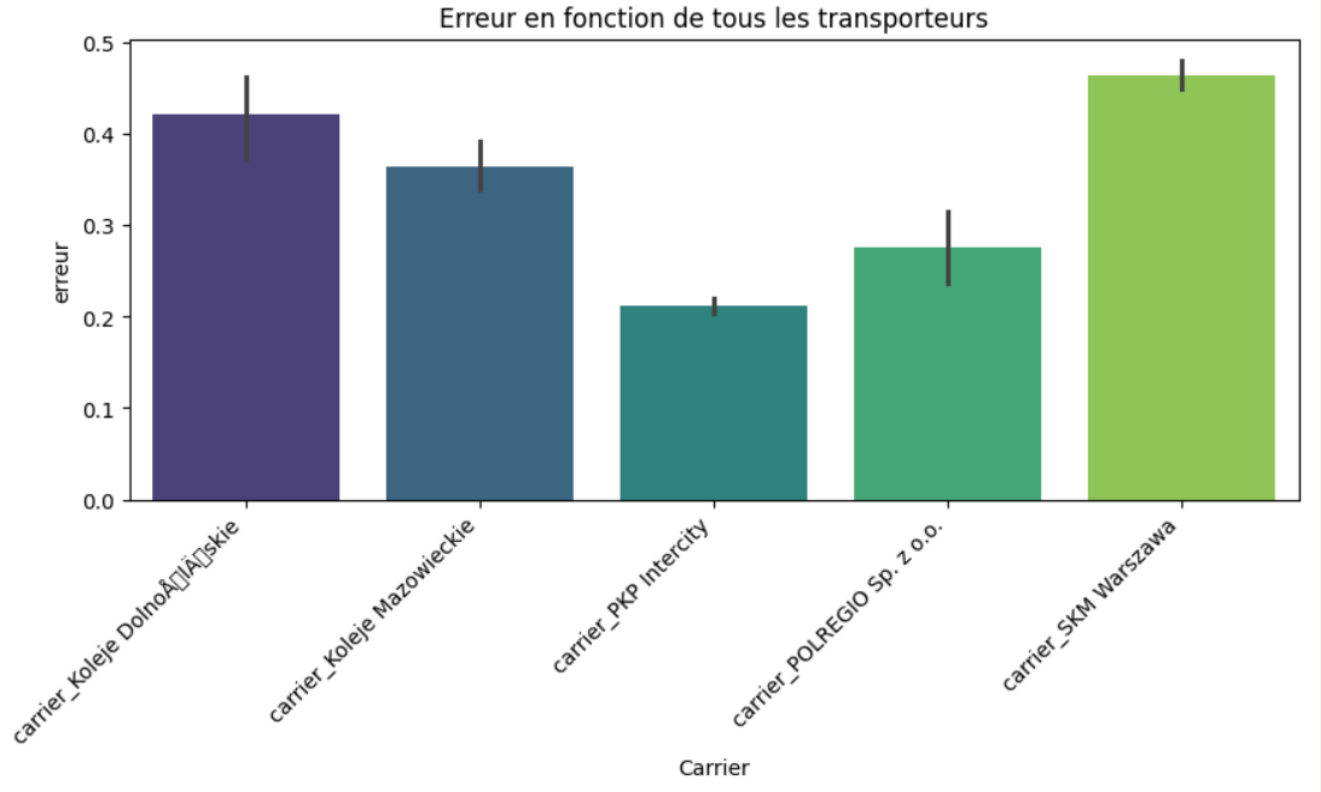
Nouvelles techniques d'identification des domaines (clustering)

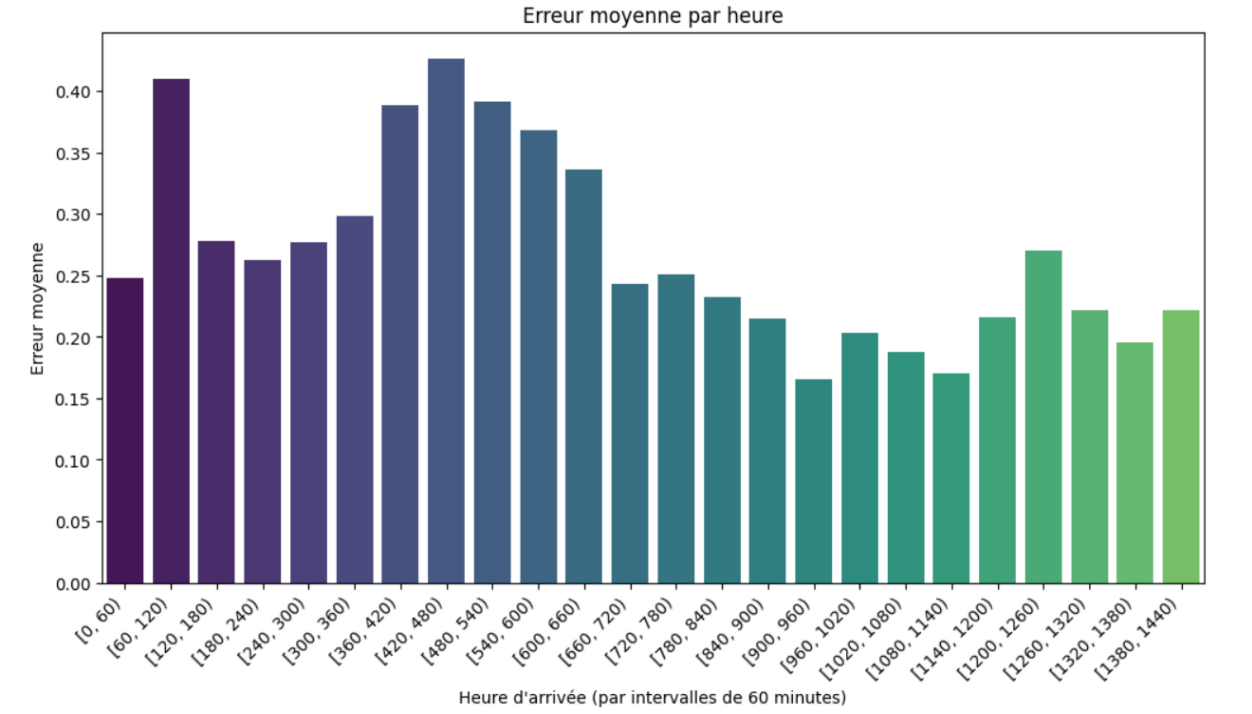
Threshold a determiner avec l'entreprise

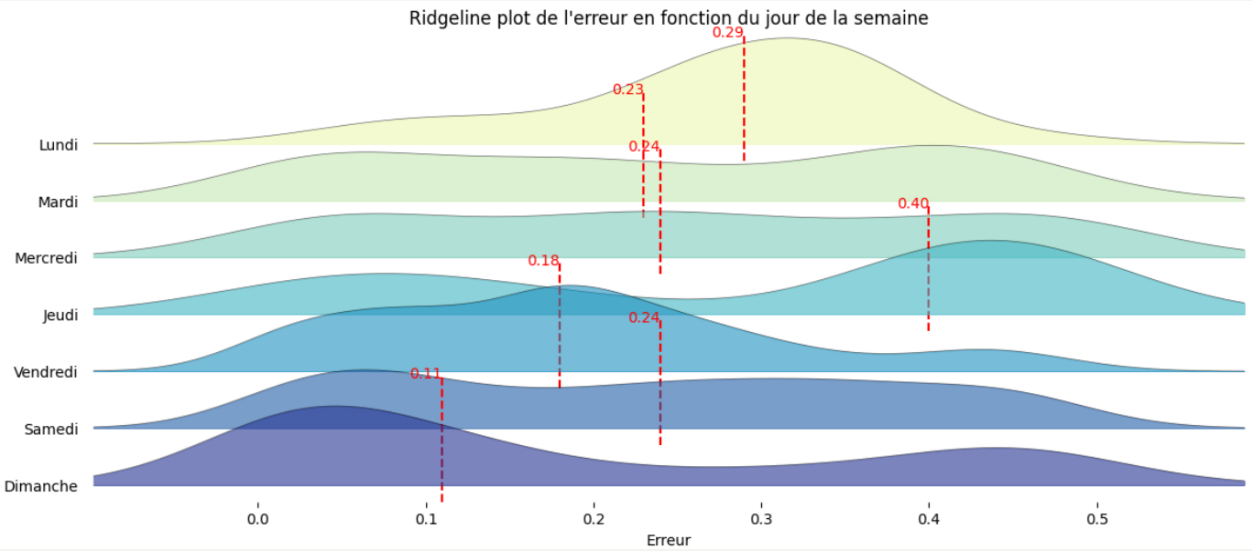
Merci de votre écoute

ANNEXES


Erreurs moyennes par connections (top30)
Erreur moyenne
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
671
338
284
556
441
761
752
105
128
282
76
336
677
502
655
754
280
109
764
78
88
555
572
77
674
295
447
292
75
548
Connection


Erreur en fonction de tous les transporteurs
0.5
0.4
0.3
0.2
0.1
0.0
erreur
carrier_Koleje DolnoÅ□lÄ□skie
carrier_Koleje Mazowieckie
carrier_PKP Intercity
carrier_POLREGIO Sp. z o.o.
carrier_SKM Warszawa
Carrier


Erreur moyenne par heure
Erreur moyenne
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
[0, 60)
[60, 120)
[120, 180)
[180, 240)
[240, 300)
[300, 360)
[360, 420)
[420, 480)
[480, 540)
[540, 600)
[600, 660)
[660, 720)
[720, 780)
[780, 840)
[840, 900)
[900, 960)
[960, 1020)
[1020, 1080)
[1080, 1140)
[1140, 1200)
[1200, 1260)
[1260, 1320)
[1320, 1380)
[1380, 1440)
Heure d'arrivée (par intervalles de 60 minutes)


Ridgeline plot de l'erreur en fonction du jour de la semaine
Lundi
Mardi
Mercredi
Jeudi
Vendredi
Samedi
Dimanche
0.29
0.23
0.24
0.40
0.18
0.24
0.11
0.0
0.1
0.2
0.3
0.4
0.5
Erreur