PRÉDICTION DES RETARDS DE TRAINS

Cindy Hua, Rémi Boutonnier, Romain Ageron

1. INTRODUCTION
2. ANALYSE DES DONNÉES
3. FEATURE ENGINEERING
4. SÉLECTION DES MODÈLES
5. OUVERTURE

INTRODUCTION

1 INTRODUCTION

Contexte

SNCF

6 mois

1 INTRODUCTION

Variables

Variables explicatives

- date
- service
- gare_depart
- gare_arrivee
- duree_moyenne
- nb_train_prevu
- Longitude_gare_depart (ajoutée)
- Lattitude_gare_depart (ajoutée)
- Longitude_gare_arrivee (ajoutée)
- Lattitude_gare_arrivee (ajoutée)

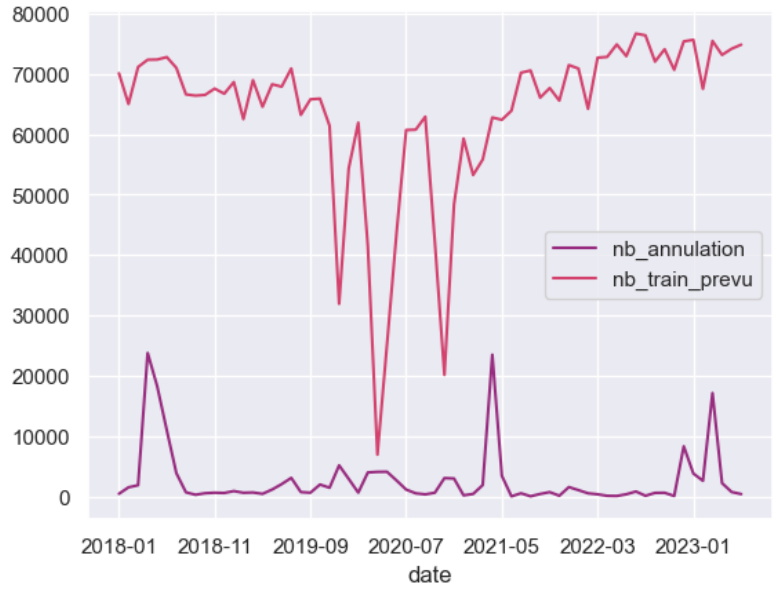
Variables cibles

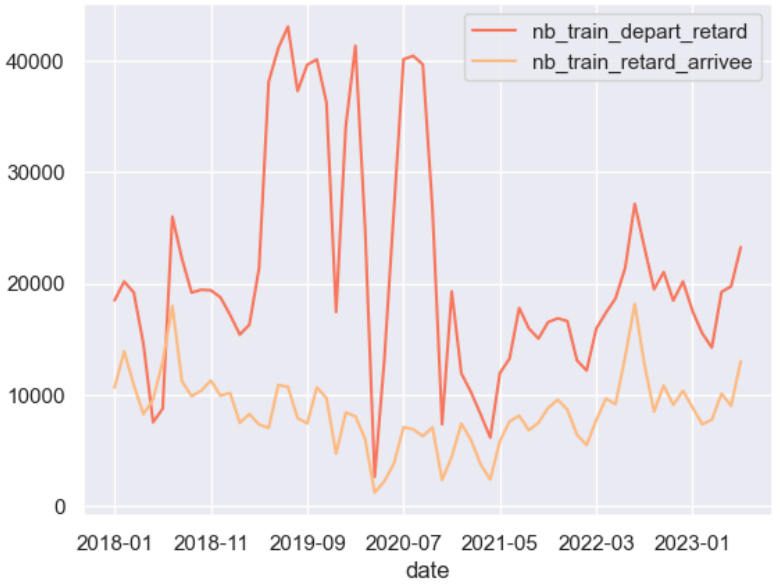
- retard_moyen_arrivee
- prct_cause_externe
- prct_cause_infra
- prct_cause_gestion_trafic
- prct_cause_materiel_roulant
- prct_cause_gestion_gare
- prct_cause_prise_en_charge_voyageurs

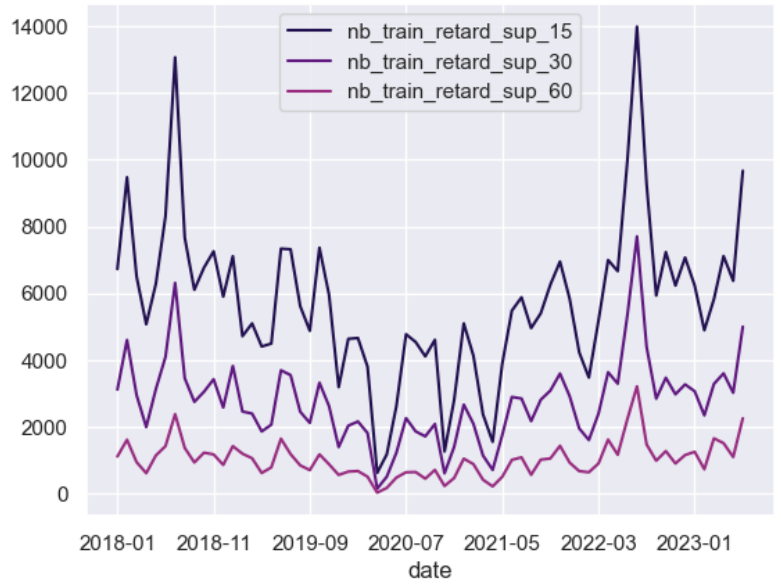
ANALYSE DES DONNÉES

2 ANALYSE DES DONNÉES

Analyse temporelle des variables

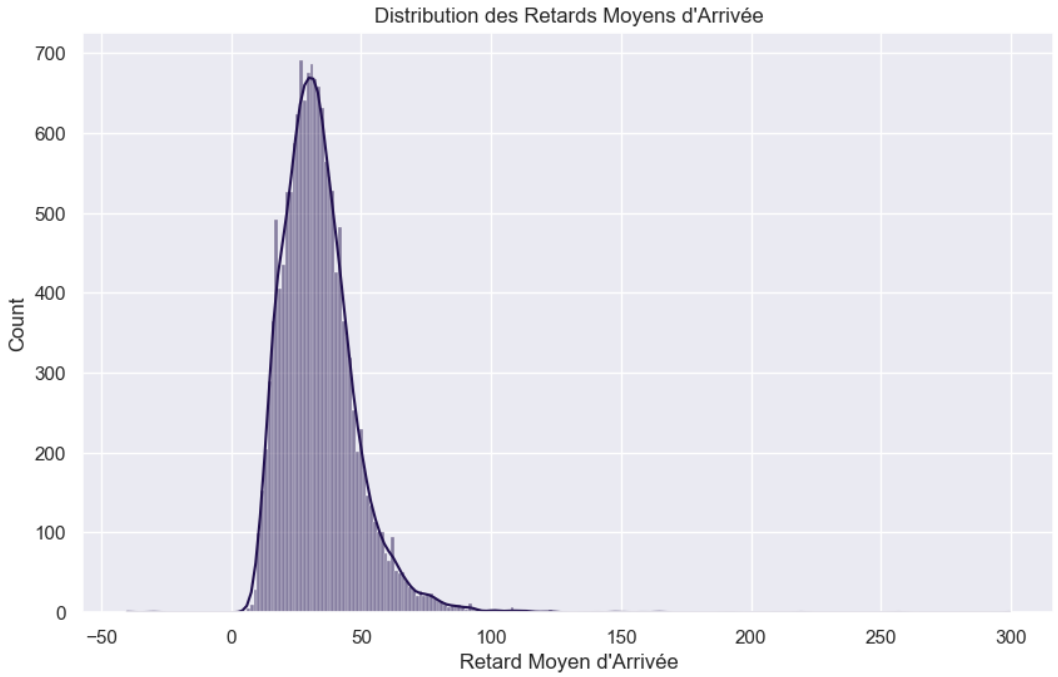


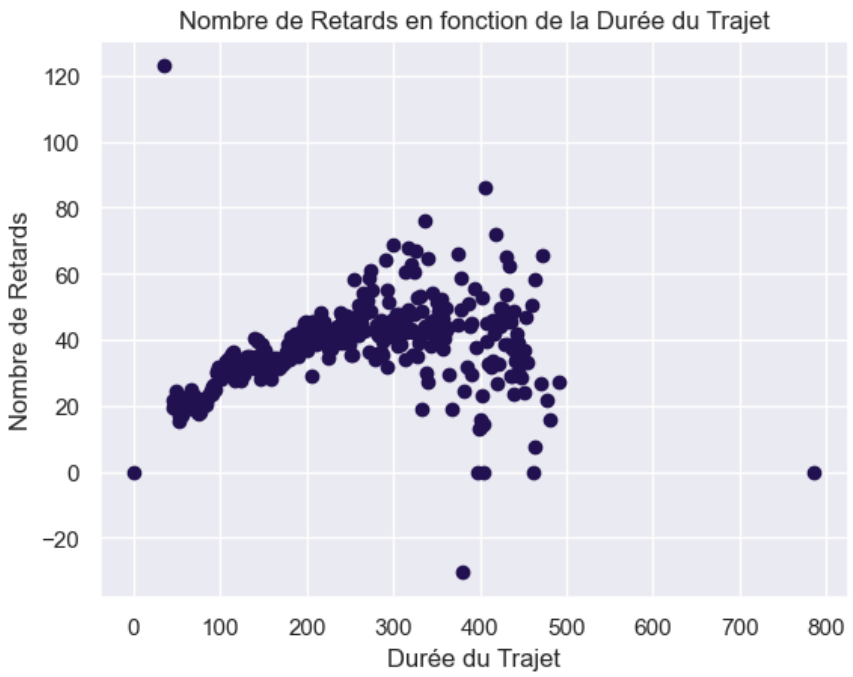




2 ANALYSE DES DONNÉES

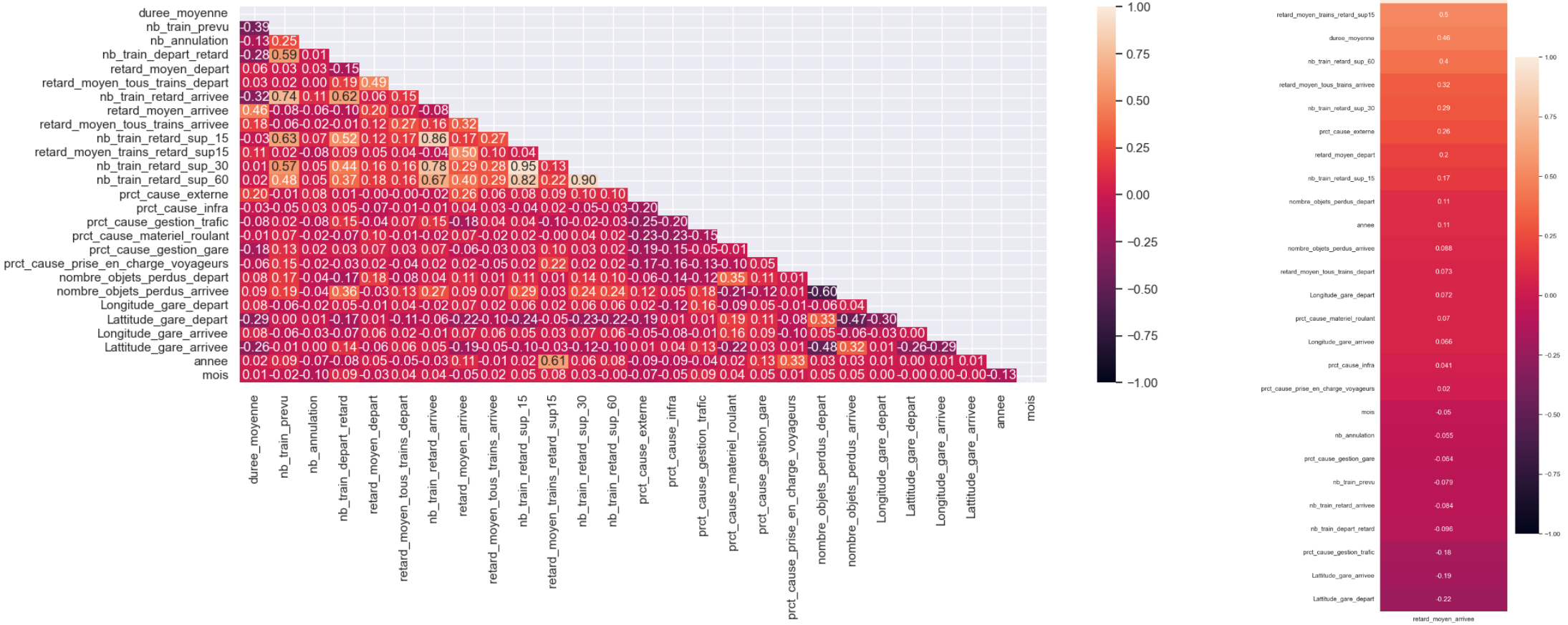
Analyse de la répartition des variables





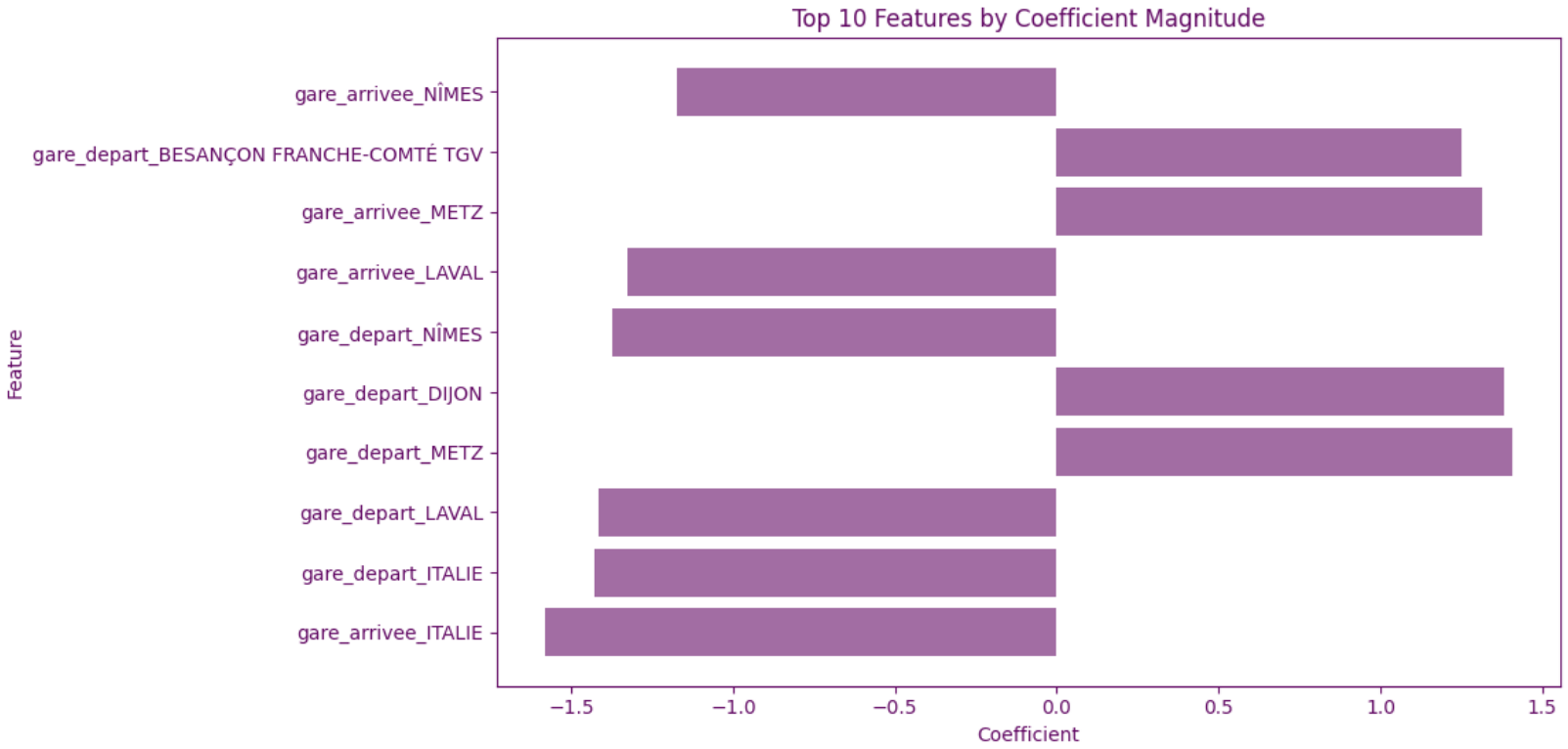
2 ANALYSE DES DONNÉES

Matrice de corrélation



2 ANALYSE DES DONNÉES

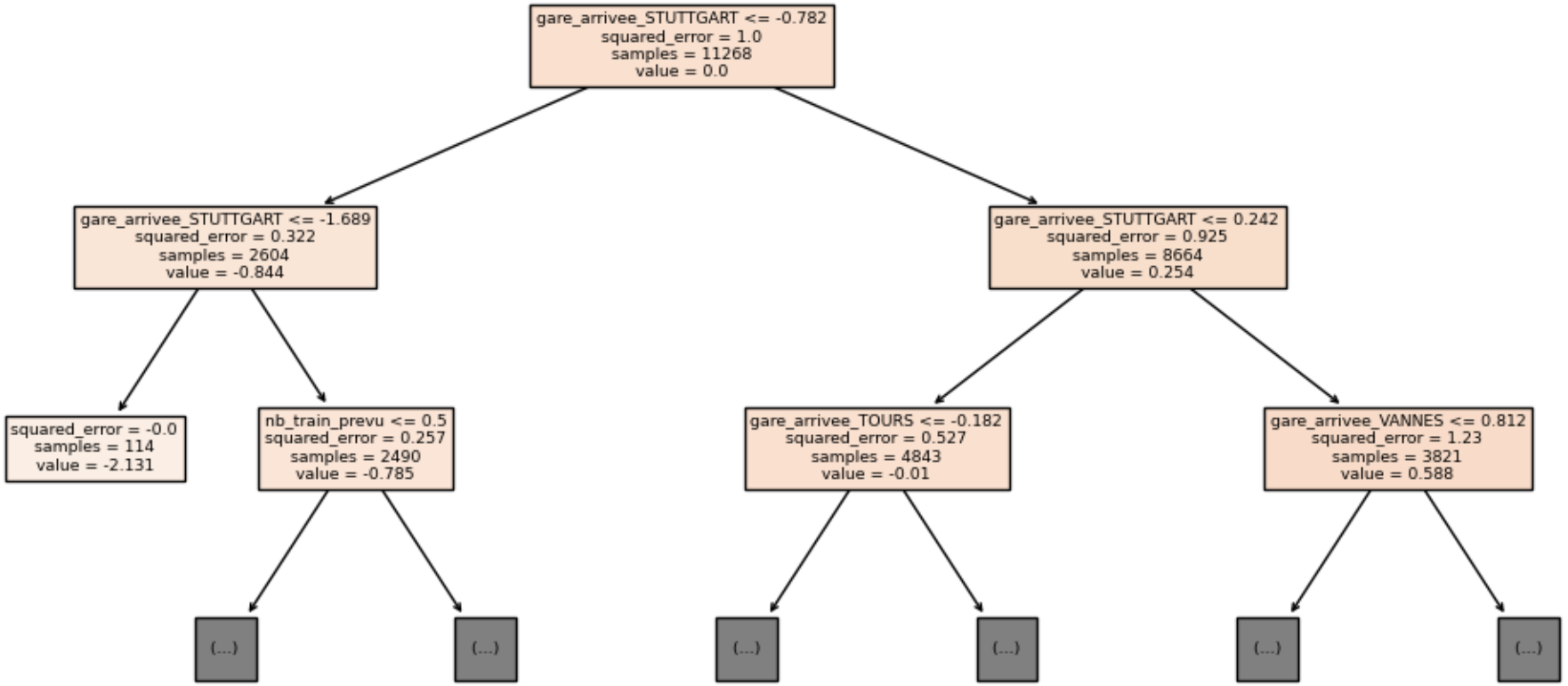
Utilisation de paramètres de modèles



Poids d'une régression linéaire

ANALYSE DES DONNÉES

Utilisation de paramètres de modèles



Nœuds d'un arbre de décision

FEATURE ENGINEERING

3 FEATURE ENGINEERING

Preprocessing

FEATURES DIRECTEMENT PRÉSENTES DANS LES DONNÉES

Features numériques

- Nombre de trains : centré-réduit
- Durées moyennes : centrées-réduites
- Coordonnées GPS : centrées-réduites
- Pourcentages : normalisés entre 0 et 1
- Dates : encodage affine
  - janvier 2018 = -1, décembre 2022 = 1

Features catégorielles

- Service : one-hot encoding
- Gare de départ : one-hot encoding
- Gare d’arrivée : one-hot encoding

3 FEATURE ENGINEERING

Distance

DISTANCE

Calculée à l'aide des coordonnées GPS des gares d'arrivée et départ

Distance à la surface de la sphère

Centrée-réduite

3

FEATURE ENGINEERING

Mois – Informations implicites

MOIS

Informations implicites

Météo

vague de chaleur
→ déformation des rails

Comportement des utilisateurs

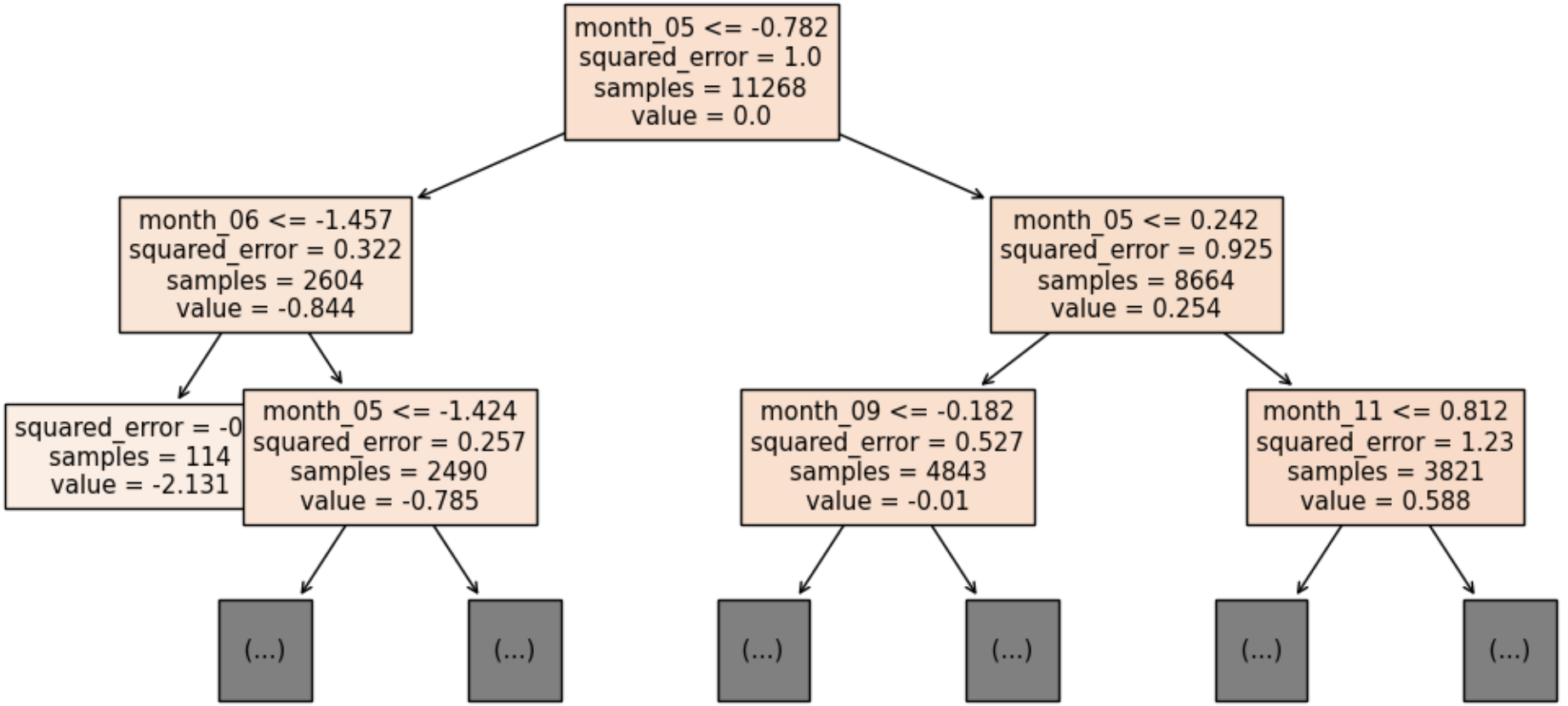
vacances de Noël

≈10%

Réduction sur l'erreur

FEATURE ENGINEERING

Mois – Utilisation par les modèles



Nœuds d'un arbre de décision
(avec les mois)

FEATURE ENGINEERING

Mois – One-hot encoding

MOIS – ONE-HOT ENCONDING

Janvier	(	1	0	0	0	0	0	0	0	0	0	0	)
Février	(	0	1	0	0	0	0	0	0	0	0	0	)
…													
Décembre	(	0	0	0	0	0	0	0	0	0	0	1	)

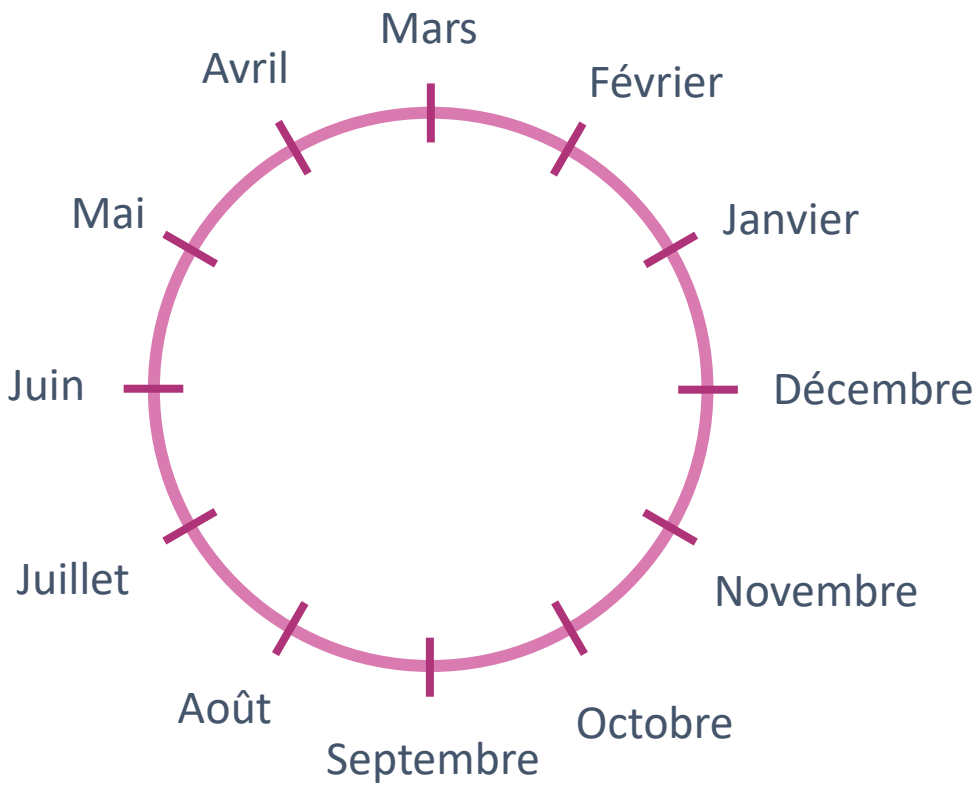
- Facile de prendre une décision sur le mois
- Linéairement indépendants
  - (utile pour la régression linéaire)

- Tous les mois sont équi-distants
  - (impact seulement le KNN)
- Une corrélation à la saison doit être apprise pour chaque mois de la saison
- Utilise 12 dimensions

FEATURE ENGINEERING

Mois – Embedding sur le cercle unité

MOIS – EMBEDDING SUR LE CERCLE UNITÉ



- Distance en lien avec la proximité temporelle
  - (impact seulement le KNN)
- Corrélation saison facilement exprimable
- Utilise 2 dimensions

- Isoler un mois nécessite plus de travail
  - ex. arbre de decision : 2 noeuds
- Linéairement dépendants
  - Mauvais pour la régression linéaire

SIMILARITÉ DES LIAISONS – UTILISATION DES LABELS

Proriété d'un bon embedding f pour les liaisons:

Plus deux liaisons x_1 et x_2 sont similaires, plus $\|f(x_1) - f(x_2)\|$ est petit

$$f(x) = moyenne(y_M | M \in training) \in \mathbb{R}^7$$

label de x au mois M

SÉLECTION DES MODÈLES

4 SÉLECTION DES MODÈLES

Grid Search avec Cross-Validation

Evaluation plus précise des modèles

Plus lente qu’une grid search sans cross validation

Utilisation d’un ensemble de test séparé

4

SÉLECTION DES MODÈLES

Grid Search results

One-Hot Encoding

Low Dimensional Embedding

	Dimension	
	138	16
	Grid Search Time	
	90 minutes	20 minutes

Model	RMSE One-Hot	RMSE Low Dim
Linear Regression	0,295	0,286
Ridge	0,295	0,286
Lasso	0,320	0,287
KNeighbors	0,281	0,280
Support Vector Regression	0,303	0,289
Decision Tree	0,300	0,293
Random Forest	0,325	0,279
ExtraTrees	0,314	0,276
AdaBoost	0,295	0,286
XGBoost	0,293	0,279

OUVERTURE

5 OUVERTURE

Prise en compte de l'aspect réseau

ASPECT RÉSEAU

Partage d'infrastructures communes entre:

Gares

Voies

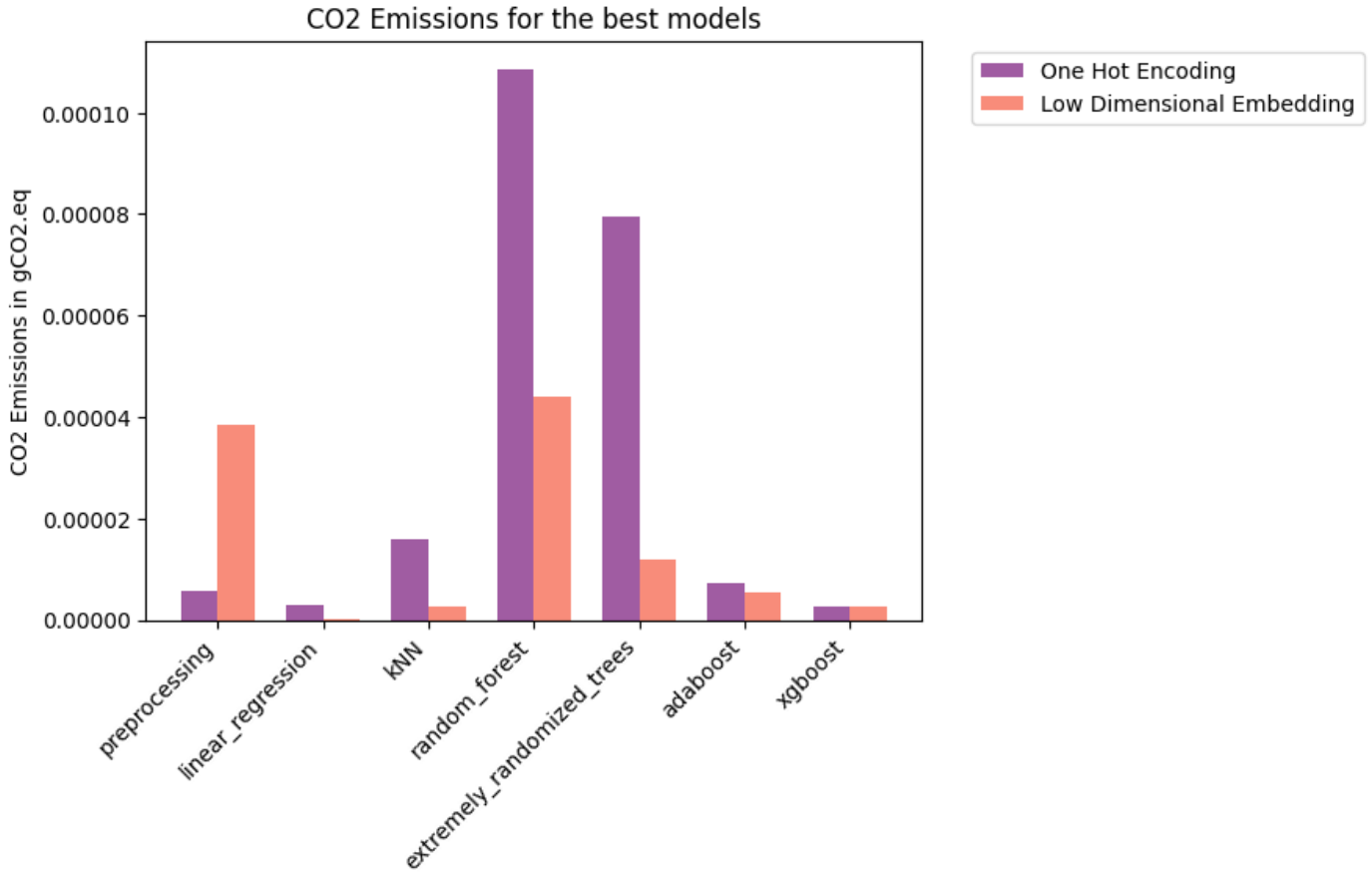
Centres de maintenance

Prise en compte dans:

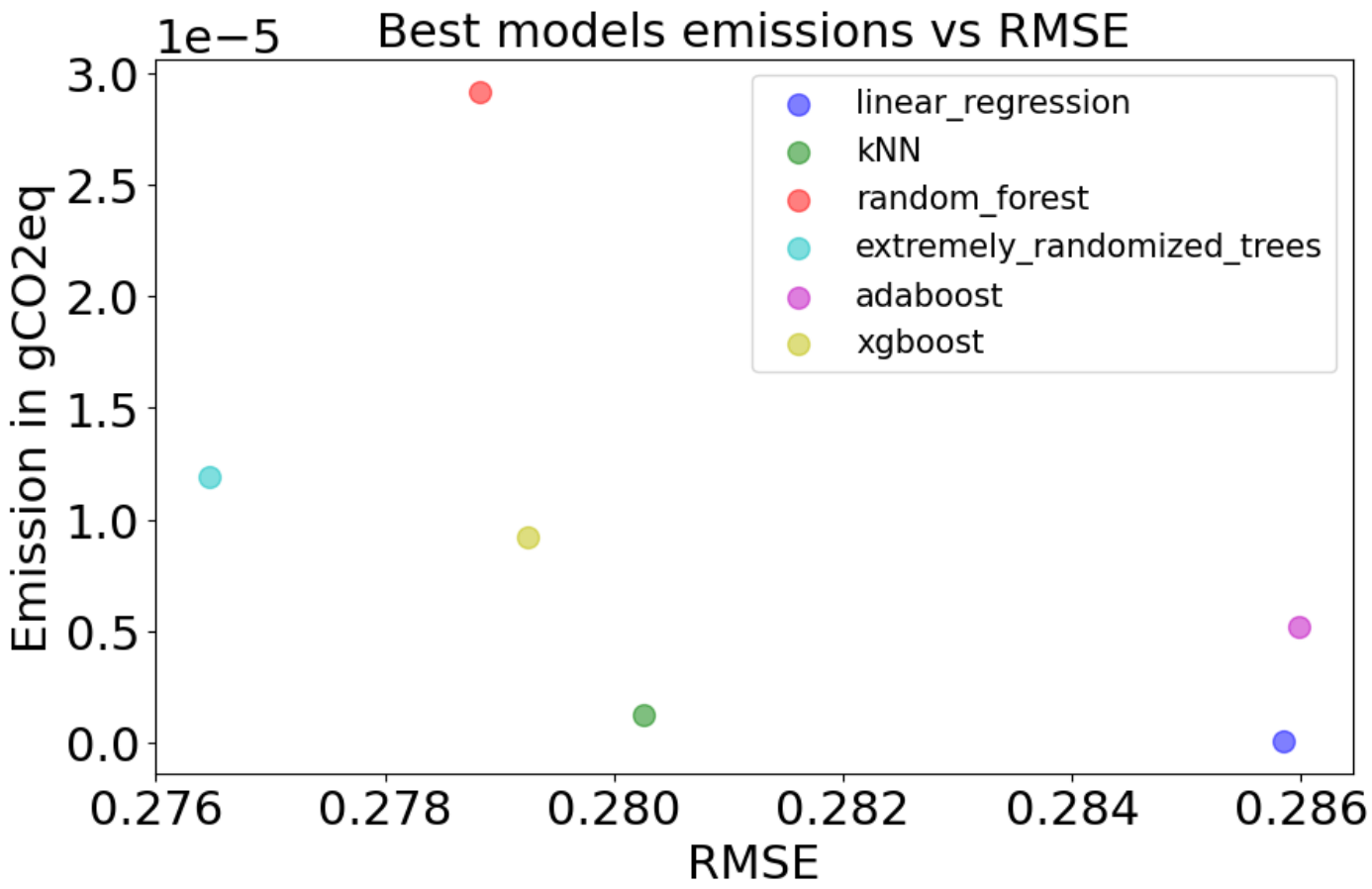
Features

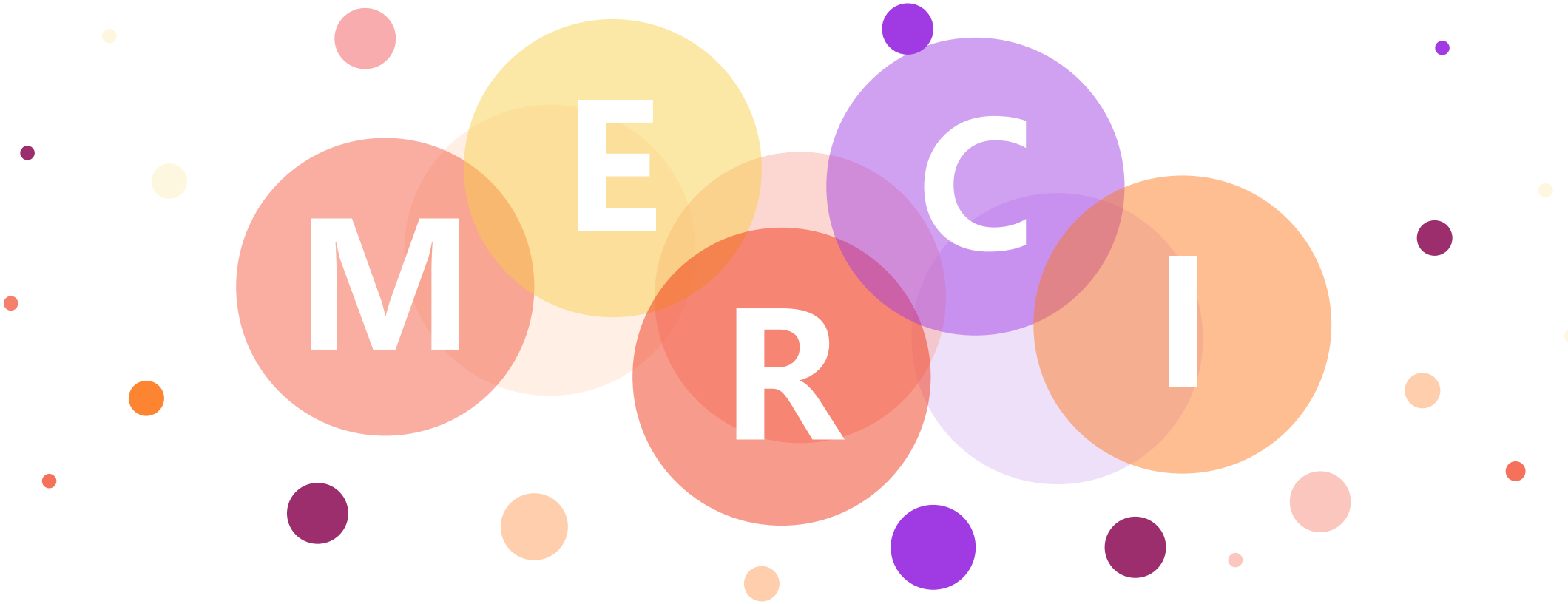
Modélisation (ML+graphes)

ASPECT ENVIRONNEMENTAL



COMPROMIS PERFORMANCES VS ÉMISSIONS




M
E
R
C
I

Annexes

Similarité des liaisons – Importance des gares

[Applicable seulement au KNN]

Avoir la même gare de départ et/ou d'arrivée a une forte influence sur les quantités à prédire

→ Prendre une contribution λ apprise plutôt que 1 dans la distance au carré

→ Remplacer $one - hot(gare)$ par $\lambda\ one - hot(gare)$

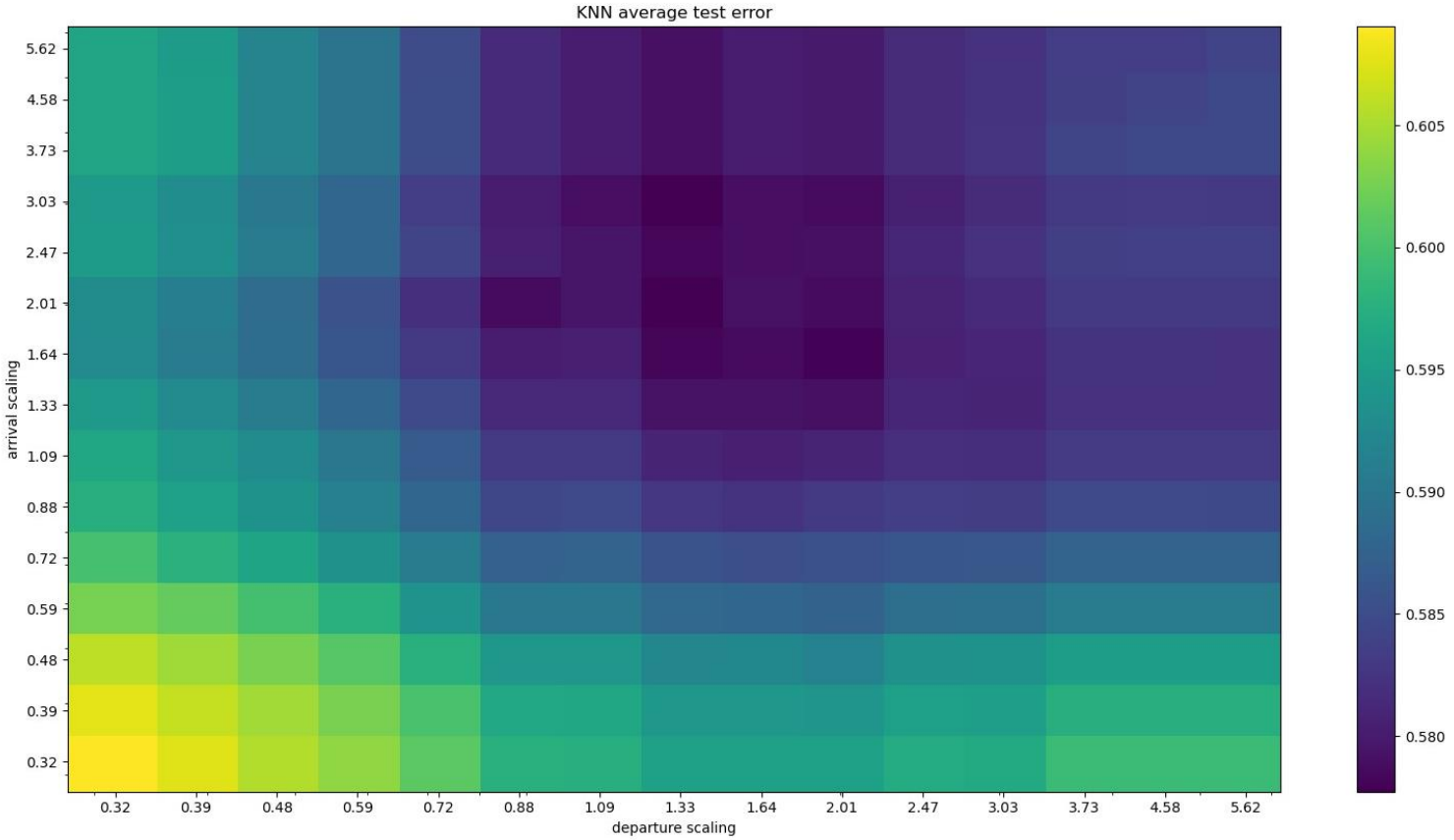
Plus généralement,
Avoir la même gare de départ n'a pas la même influence qu'avoir la même gare d'arrivée

→ Prendre deux contributions α, β apprises plutôt que 1 dans la distance au carré

→ $\alpha\ one - hot(départ)$ et $\beta\ one - hot(arrivée)$

Similarité des liaisons – Importance des gares

[Applicable seulement au KNN]



- Les gares ont plus d'influence que les autres features.
- La gare d'arrivée a plus d'influence que la gare de départ.
- $\alpha = 4/3$ et $\beta = 2$