

OPENCLASSROOMS

Projet 4

-

Anticipez les besoins en consommation de bâtiments



CentraleSupélec

Pierrick BERTHE

Formation Expert en Data Science
Openclassrooms – CentraleSupélec

août 2023 → avril 2024



Sommaire



I – Problématique

II – Présentation du jeu de données

III - Nettoyage des données

IV – Analyses exploratoires

V – Feature engineering

VI – Modèle de prédiction - Energie totale

VII – Modèle de prédiction - Emission CO₂

VIII - Conclusion



Problématique

La **ville de Seattle** étudie ses émissions des bâtiments non destinés à l'habitation puisqu'ils génèrent 33% des émissions de gaz à effet de serre de la ville.

La ville effectue des relevés annuels des bâtiments de la ville pour suivre l'évolution de leurs performances énergétiques depuis 2013. => coûteux.

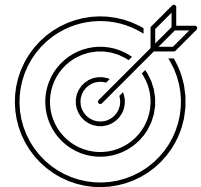
Nous devons tenter de prédire les émissions de CO2 et la consommation totale d'énergie des bâtiments non destinés à l'habitation et non-mesurés à partir du relevé de l'année 2016.

Missions :

1. Réaliser une courte analyse exploratoire.
2. Tester différents modèles de prédiction pour prédire la consommation totale d'énergie.
3. Tester différents modèles de prédiction pour prédire les émissions de CO2.
4. Evaluer l'intérêt de l'ENERGY STAR Score pour les prédictions



City of Seattle





Sommaire



I – Problématique

II – Présentation du jeu de données

III - Nettoyage des données

IV – Analyses exploratoires

V – Feature engineering

VI – Modèle de prédiction - Energie totale

VII – Modèle de prédiction - Emission CO₂

VIII - Conclusion



Présentation du jeu de données



2016_Building_Energy_Benchmarking.csv



3_376 lignes

46 colonnes

➤ Descriptif des bâtiments :

- Des informations **administratives**: type de bâtiment, type d'occupation, etc.
- Des informations **structurelles**: nombre de bâtiment, surface, etc.
- Des informations **géographiques** : longitude, latitude, etc.
- Des informations **énergétiques** : quantité de consommation, type de source énergétique, etc.
- Des informations de **pollution** : quantité totale et relative de gaz à effet de serre.

➤ Valeurs manquantes :

- 13 % de NaN
- 26 / 46 colonnes concernées

➤ Doublons

Pas de doublons sur la colonne de l'identifiant unique des bâtiments.



Sommaire



I – Problématique

II – Présentation du jeu de données

III - Nettoyage des données

IV – Analyses exploratoires

V – Feature engineering

VI – Modèle de prédiction - Energie totale

VII – Modèle de prédiction - Emission CO₂

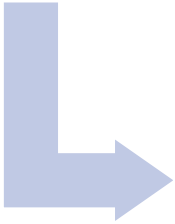
VIII - Conclusion



Nettoyage des données



1. Filtrage méthode entonnoir



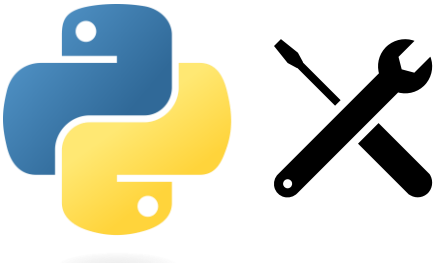
2. Imputation valeurs manquantes



3. Transformation logarithmique
features cibles



4. Suppression des outliers





Nettoyage des données



Nom	Utilisation	Fonctions spécifiques
Anaconda	Gestion de package Gestion d'environnement virtuel	Conda : installation de package via le terminal
Visual Studio Code 1.84.2	Structurer la démarche Exécuter code par étape Expliquer la démarche (markdown)	
Python 3.11.6	Appel aux librairies Boucles for pour générer plusieurs calculs et graphiques	Boucles, listes, dictionnaires, librairies, méthodes
Pandas 2.1.1	Manipulation de données Représentation des données	Manipulation de Dataframe : création, copie, filtres, tris, description, concaténation
Matplotlib 3.8.0 Seaborn 0.13.0	Génération de graphiques de visualisation	Barplot, scatterplot, lineplot, distplot, heatmap
Numpy 1.26.0	Manipulation de matrices et fonctions mathématiques	Histogram, argmax, arange, object, number
Missingno 0.5.2	Représentation graphique pour valeurs manquantes	Matrice de NaN
Sklearn 1.3.1	Apprentissage automatique et modélisation statistique	SimpleImputer, KNNImputer, StandardScaler, PCA
Scipy 1.11.3	Calculs de mathématiques complexes ou de problèmes scientifiques	Stats, chi2_contingency, shapiro, kruskal



Nettoyage des données



1/ Filtrage méthode « entonnoir »

a) Filtrage remplissage feature

=> ~~features < 50% de remplissage~~



Nettoyage des données



1/ Filtrage méthode « entonnoir »

a) Filtrage remplissage feature

=> ~~features < 50% de remplissage~~



b) Filtrage features redondantes

=> Quantité d'énergie (kBtu / kWh)

~~Localisation géographique (longitude & latitude / adresse / n° de parcelle / Zip Code)~~



Nettoyage des données



1/ Filtrage méthode « entonnoir »

a) Filtrage remplissage feature

=> ~~features < 50% de remplissage~~

b) Filtrage features redondantes

=> Quantité d'énergie (kBtu / kWh)

Localisation géographique (longitude & latitude / ~~adresse / n° de parcelle / Zip Code~~)

c) Filtrage features inutiles

=> « ~~Date~~year » / « ~~city~~ » / « ~~State~~ »



Nettoyage des données



1/ Filtrage méthode « entonnoir »

a) Filtrage remplissage feature

=> ~~features < 50% de remplissage~~

b) Filtrage features redondantes

=> Quantité d'énergie (kBtu / kWh)

~~Localisation géographique (longitude & latitude / adresse / n° de parcelle / Zip Code)~~

c) Filtrage features inutiles

=> « ~~Date~~year » / « ~~city~~ » / « ~~State~~ »

d) Gestion des features de consommation

=> proportion des sources d'énergie (électricité / gaz naturel / vapeur)
~~consommation énergie et émission CO2~~



Nettoyage des données



1/ Filtrage méthode « entonnoir »

a) Filtrage remplissage feature

=> ~~features < 50% de remplissage~~

b) Filtrage features redondantes

=> Quantité d'énergie (kBtu / kWh)

~~Localisation géographique (longitude & latitude / adresse / n° de parcelle / Zip-Code)~~

c) Filtrage features inutiles

=> « ~~Date~~year » / « ~~city~~ » / « ~~State~~ »

d) Gestion des features de consommation

=> proportion des sources d'énergie (électricité / gaz naturel / vapeur)
~~consommation énergie et émission CO2~~

e) Filtrage par type d'utilisation

=> Conservation des bâtiments non destinés à l'habitation



Nettoyage des données



1/ Filtrage méthode « entonnoir »

a) Filtrage remplissage feature

=> ~~features < 50% de remplissage~~

b) Filtrage features redondantes

=> Quantité d'énergie (kBtu / kWh)

~~Localisation géographique (longitude & latitude / adresse / n° de parcelle / Zip Code)~~

c) Filtrage features inutiles

=> « ~~State~~year » / « city » / « State »

d) Gestion des features de consommation

=> proportion des sources d'énergie (électricité / gaz naturel / vapeur)
~~consommation énergie et émission CO2~~

e) Filtrage par type d'utilisation

=> Conservation des bâtiments non destinés à l'habitation

f) Filtrage bâtiments atypiques

=> ~~bâtiments labélisés « outliers »~~



Nettoyage des données

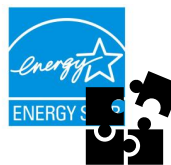
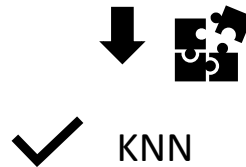


2. Imputation valeurs manquantes



1 2 3 4 5
6 7 8 9 0
+ - * ÷

Features quantitatives

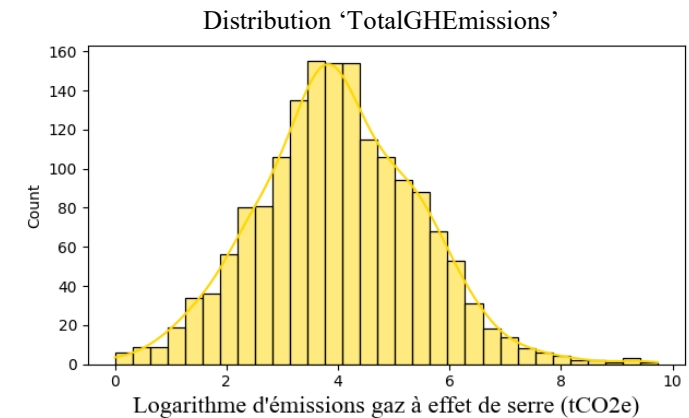
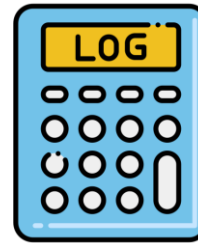
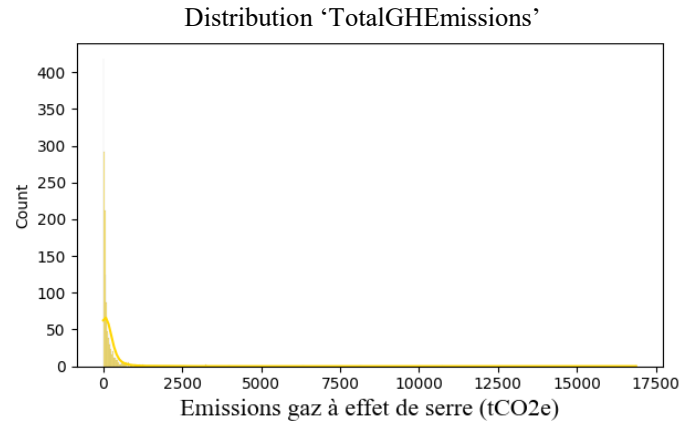




Nettoyage des données



3. Transformation logarithmique features cibles



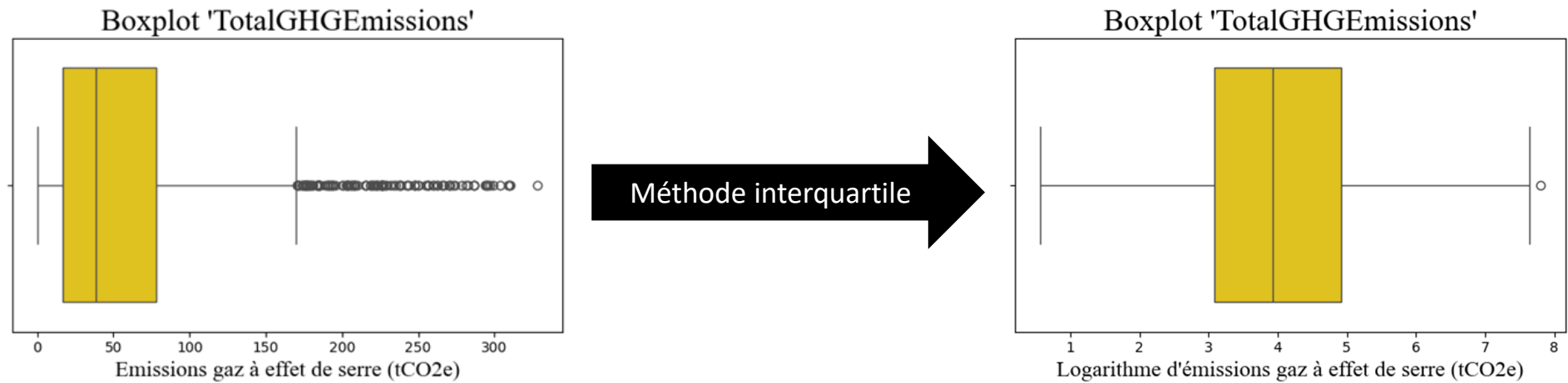
- ✓ Consommation annuelle totale d'énergie
- ✓ Emission de CO₂



Nettoyage des données



4. Gestion des outliers



- ✓ Consommation annuelle totale d'énergie
- ✓ Emission de CO₂



Sommaire



- I – Problématique
- II – Présentation du jeu de données
- III - Nettoyage des données
- IV – Analyses exploratoires**
- V – Feature engineering
- VI – Modèle de prédiction - Energie totale
- VII – Modèle de prédiction - Emission CO₂
- VIII - Conclusion

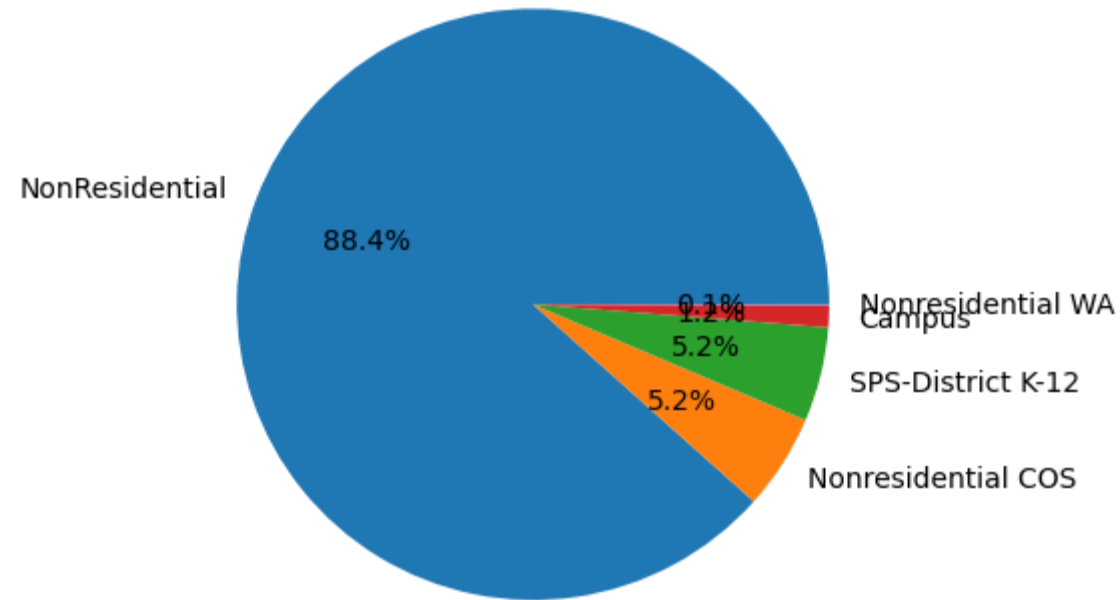


Analyses exploratoires



=> Analyse univarié

Répartition des types de batiment



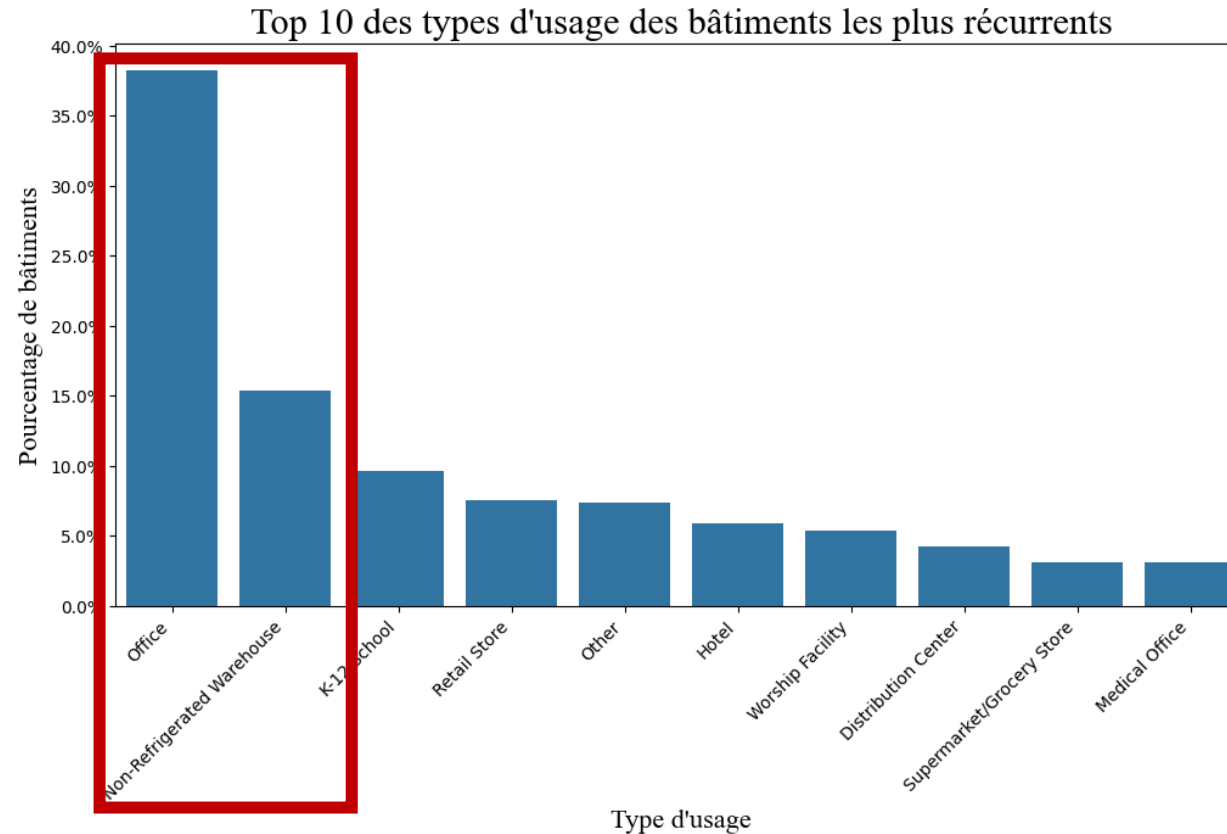
Bâtiments non résidentiels ✓



Analyses exploratoires



=> Analyse univari 



> 50% des b timents

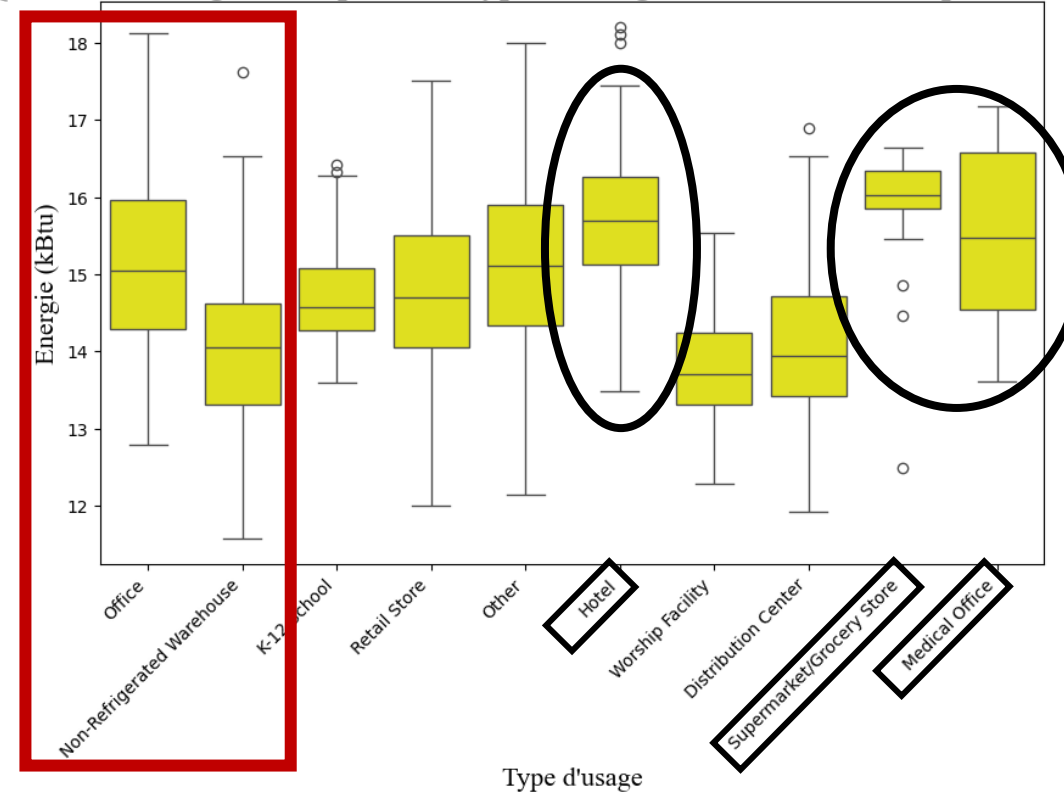


Analyses exploratoires



=> Analyse univari 

Quantit  d' nergie du top 10 des types d'usage des b timents les plus r currents



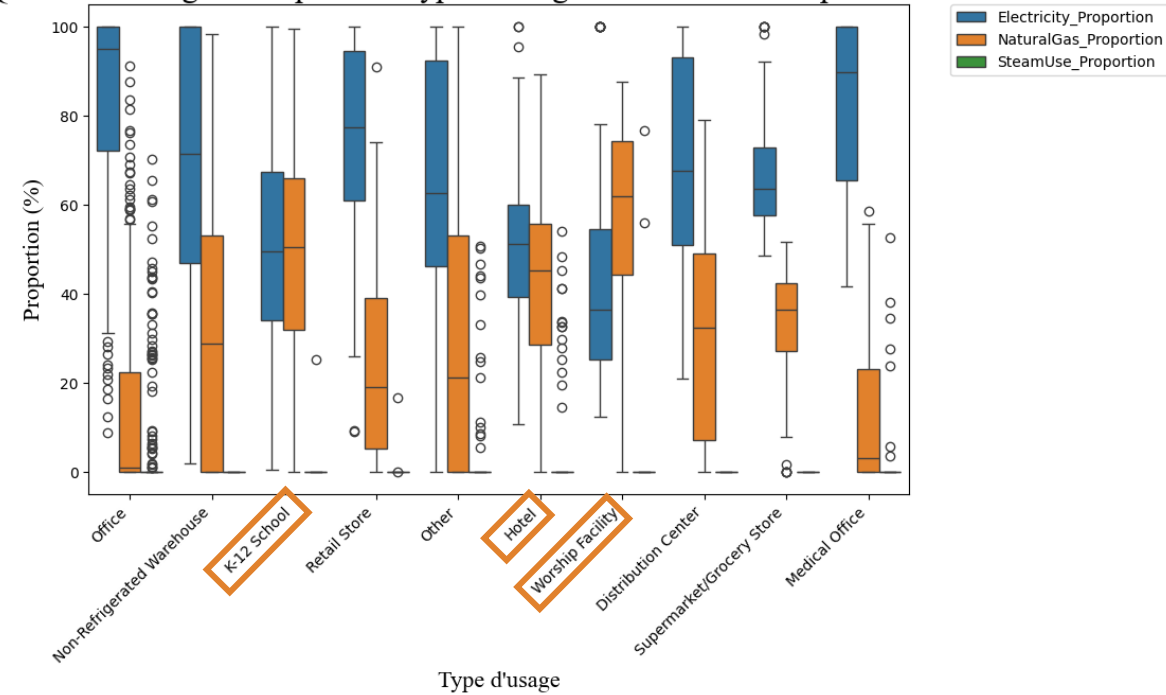


Analyses exploratoires



=> Analyse univarié

Quantité d'énergie du top 10 des types d'usage des bâtiments les plus récurrents



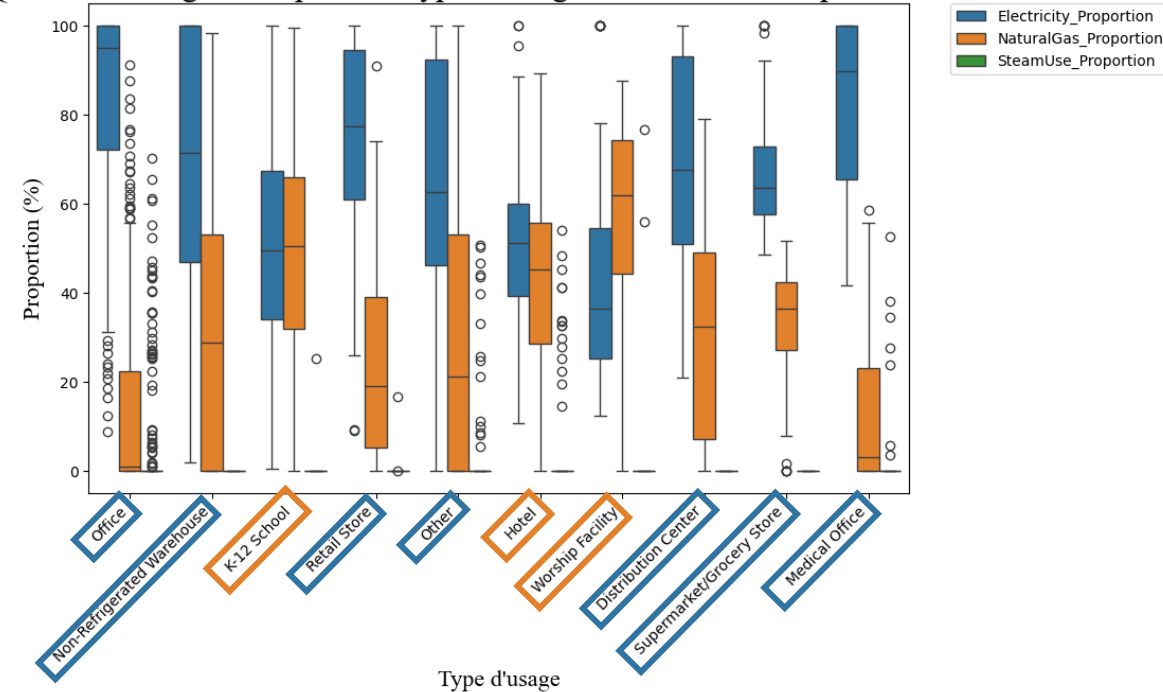


Analyses exploratoires



=> Analyse univarié

Quantité d'énergie du top 10 des types d'usage des bâtiments les plus récurrents



L'électricité est la principale source d'énergie



Analyses exploratoires

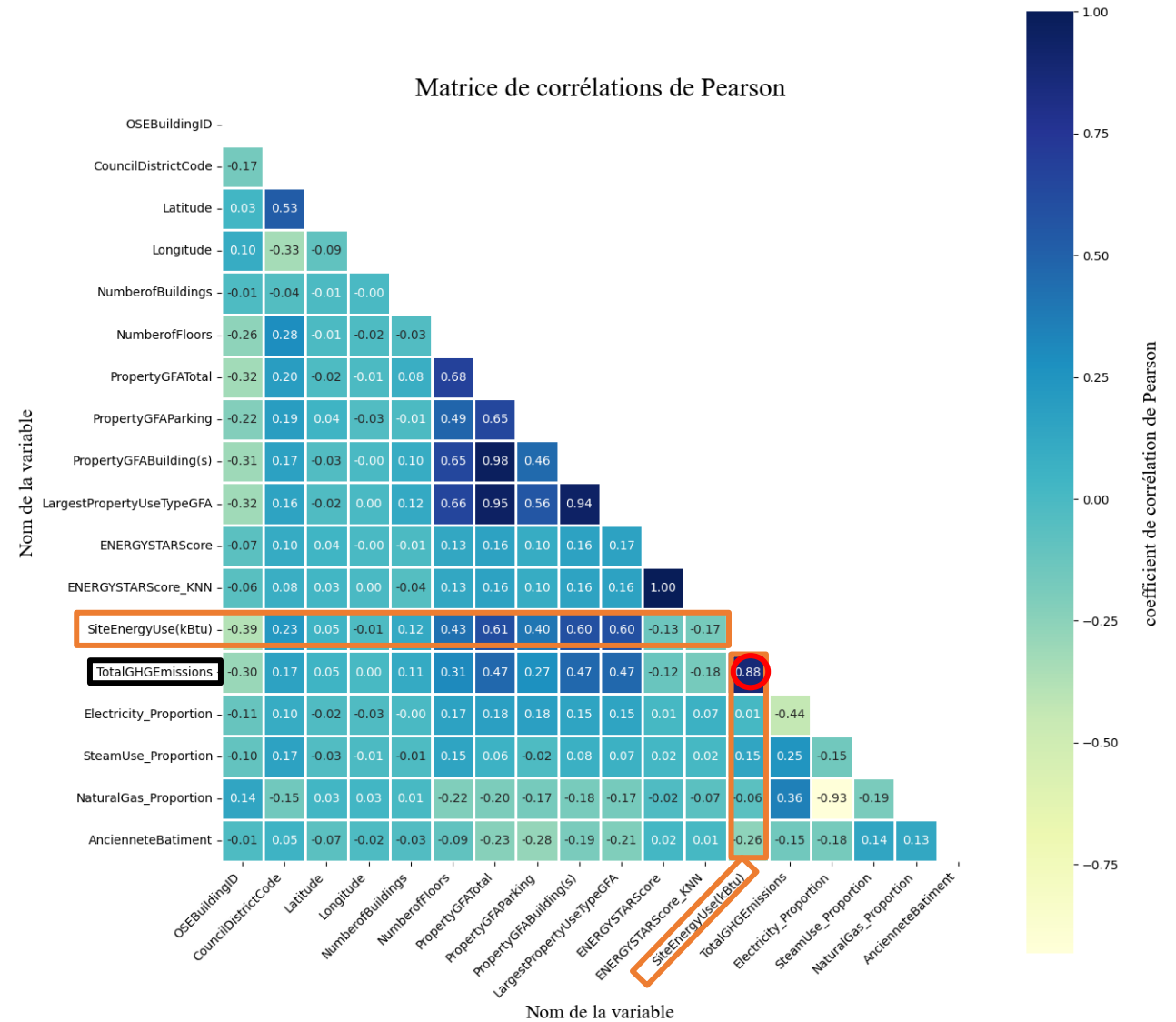


=> Analyse bivarié



$R_{(Pearson)} > 50\%$

Matrice de corrélations de Pearson



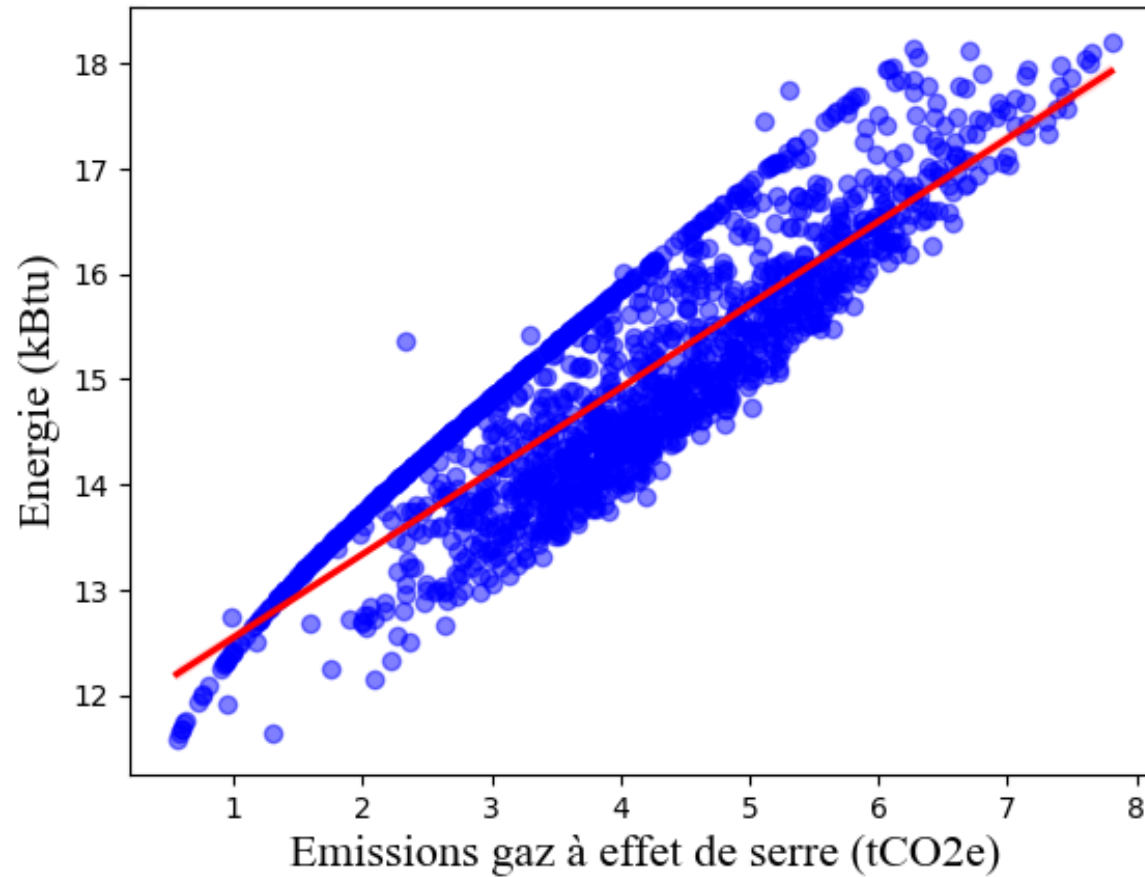


Analyses exploratoires



=> Analyse bivarié

Corrélation de Pearson sans les outliers



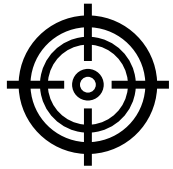
Corrélation Pearson (r) = 0.88



Analyses exploratoires

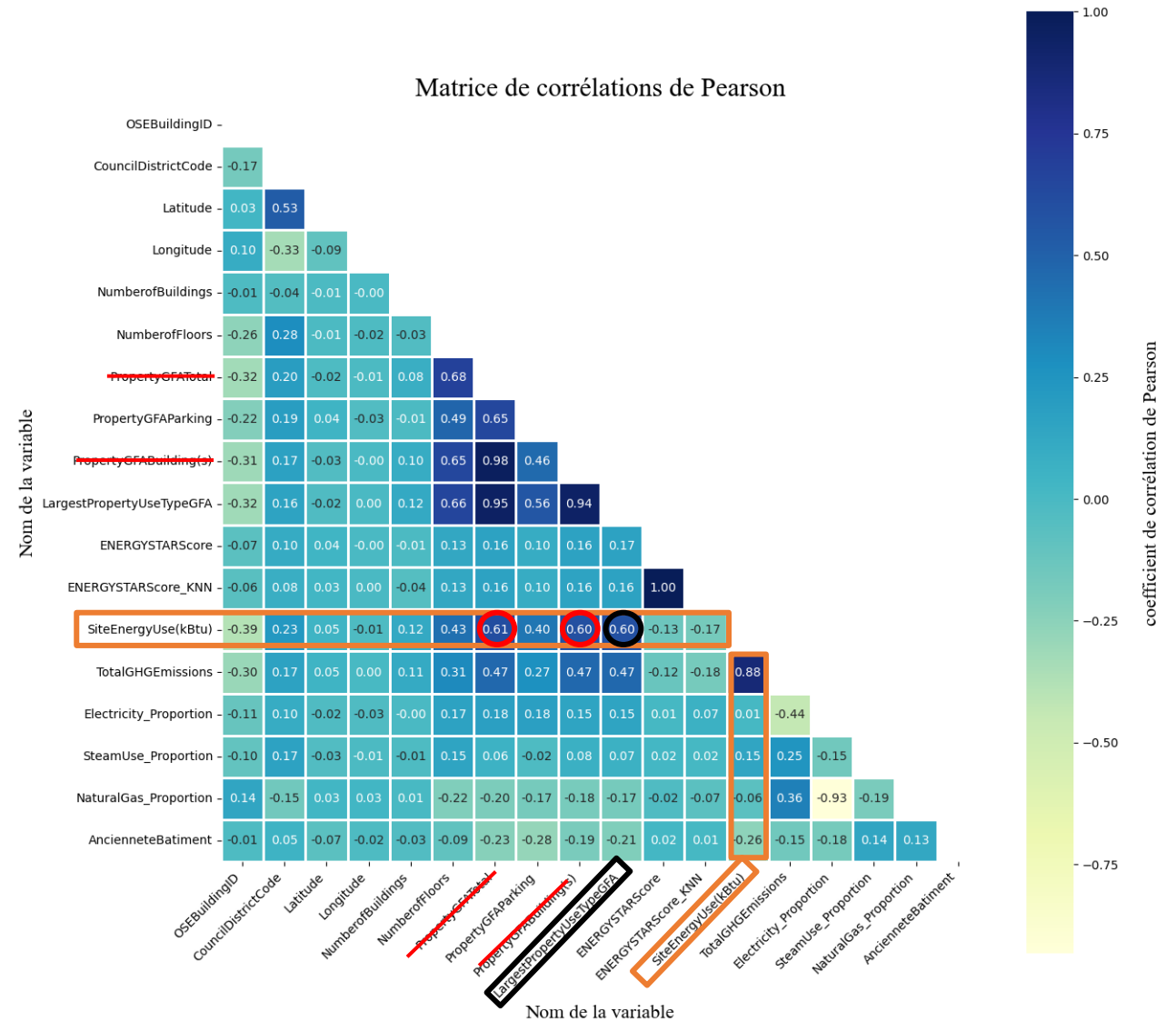


=> Analyse bivarié



$R_{(Pearson)} > 50\%$

Matrice de corrélations de Pearson





Analyses exploratoires

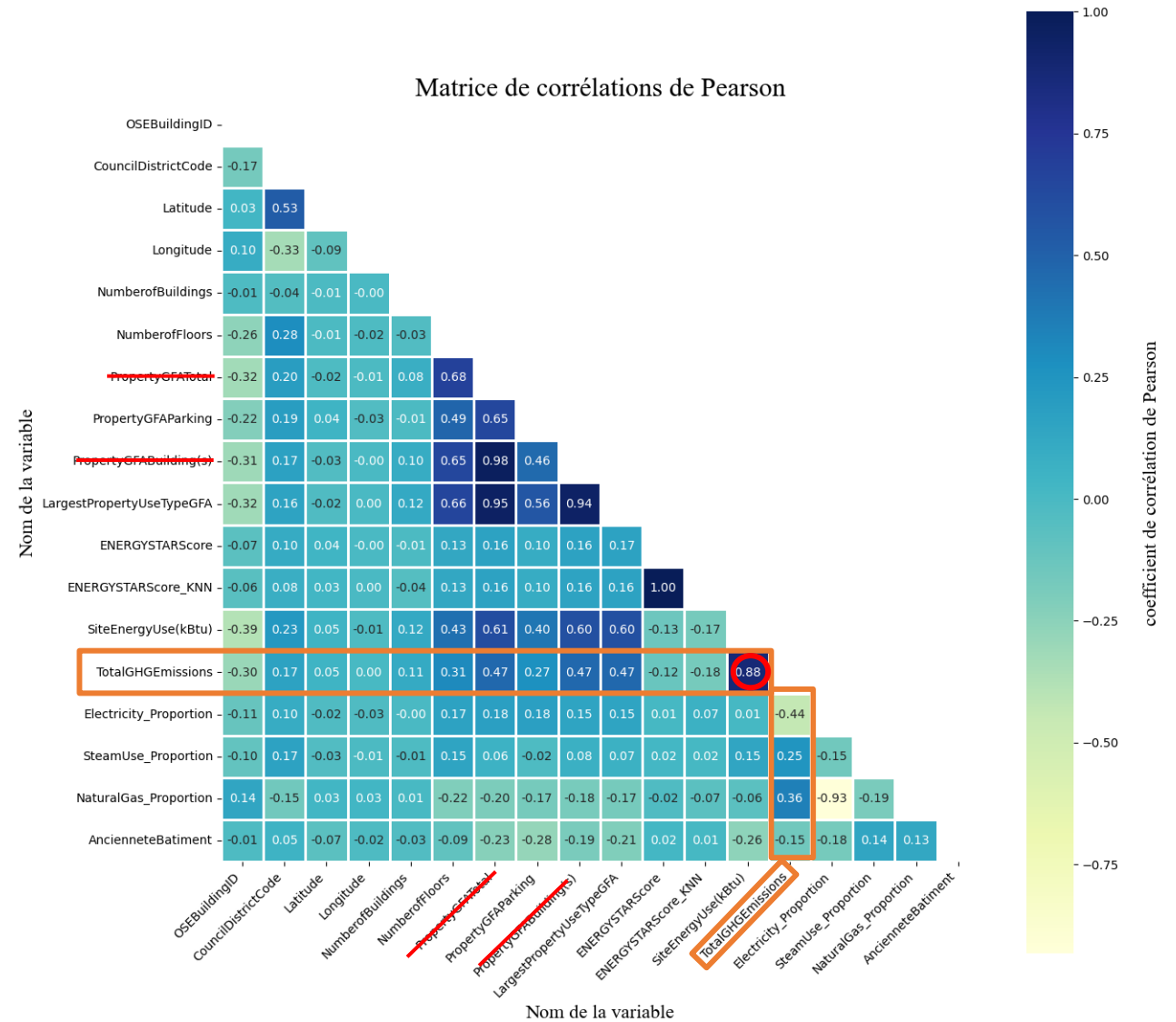


=> Analyse bivarié



$R_{(Pearson)} > 50\%$

Matrice de corrélations de Pearson

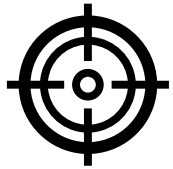




Analyses exploratoires

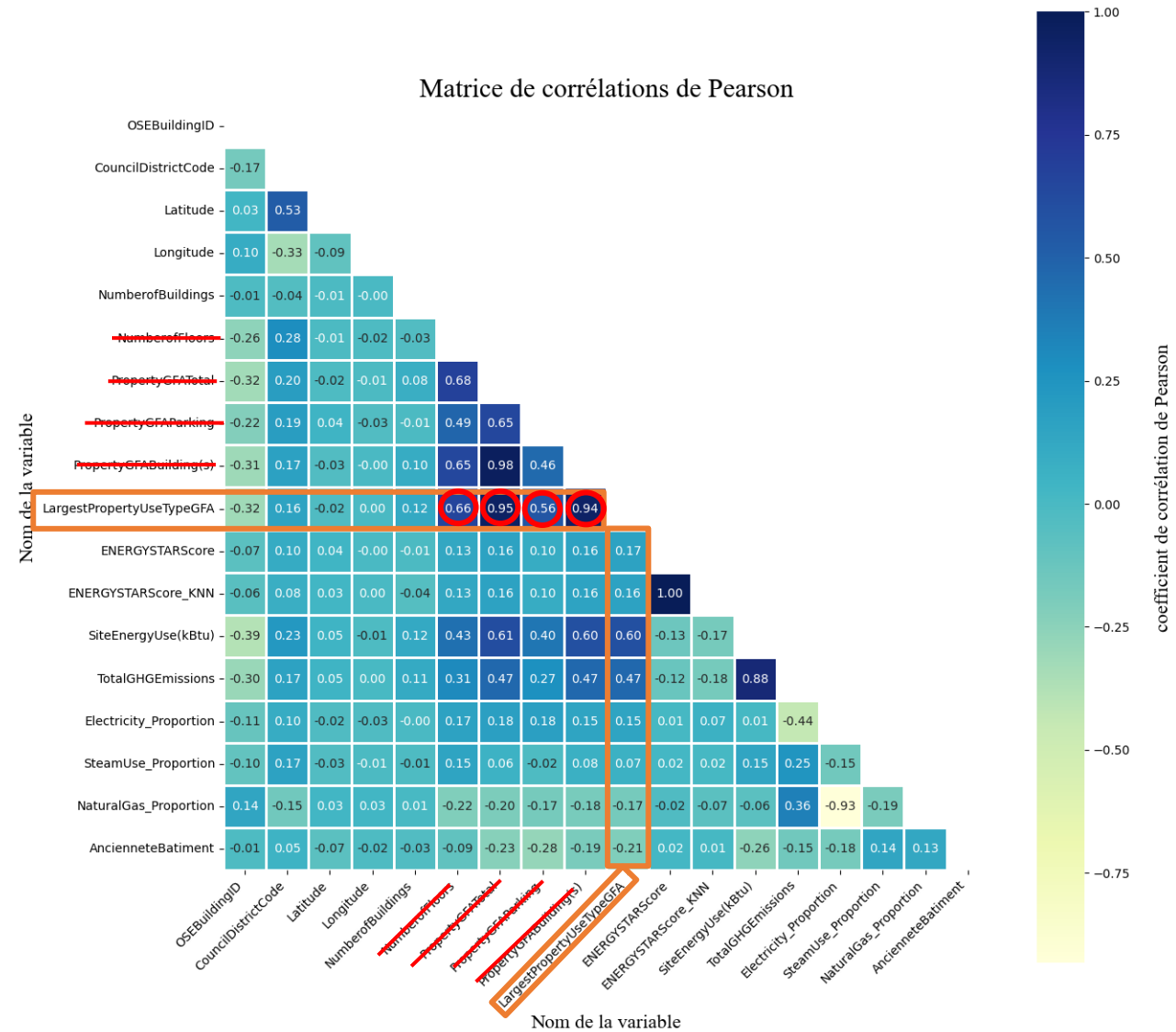


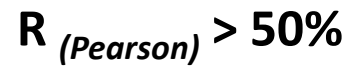
=> Analyse bivarié



$R_{(Pearson)} > 50\%$

Matrice de corrélations de Pearson



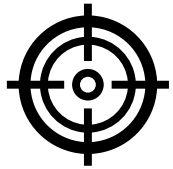




Analyses exploratoires



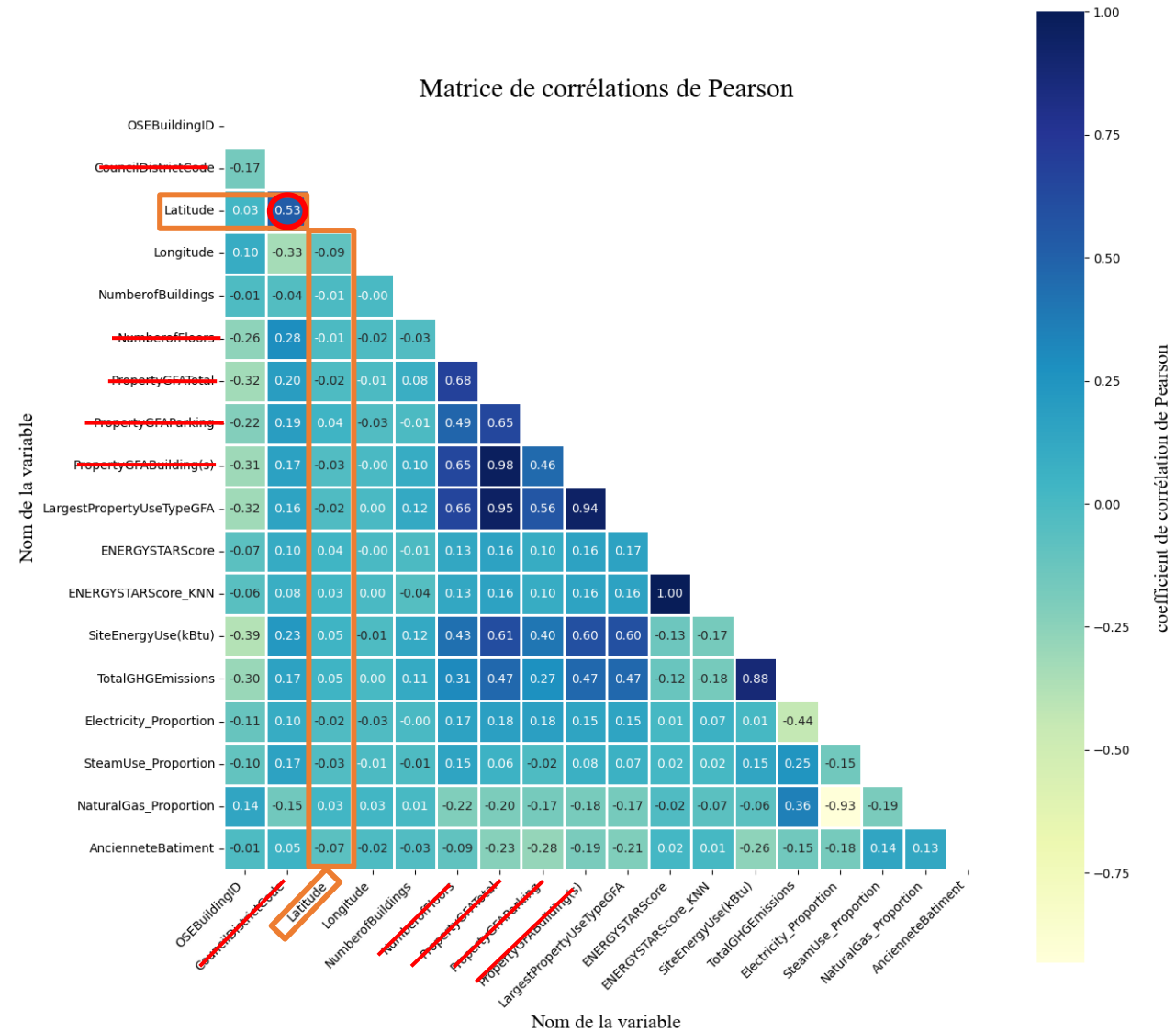
=> Analyse bivarié



$R_{(Pearson)} > 50\%$

-5 features

Matrice de corrélations de Pearson





Analyses exploratoires



=> **Variance Inflation Factor (VIF)**

	feature	VIF
2	Longitude	982253.28997
1	Latitude	981765.89052
6	Electricity_Proportion	478.49544
8	NaturalGas_Proportion	127.99405
7	SteamUse_Proportion	8.35286
5	ENERGYSTARScore_KNN	6.81848
9	AncienneteBatiment	4.13828
0	OSEBuildingID	2.75987
3	NumberofBuildings	2.04125
4	LargestPropertyUseTypeGFA	1.75981



Analyses exploratoires

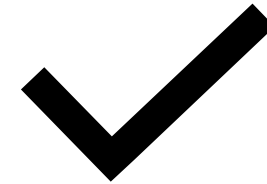


=> Variance Inflation Factor (VIF)

	feature	VIF
2	Longitude	982253.28997
1	Latitude	981765.89052
6	Electricity_Proportion	478.49544
8	NaturalGas_Proportion	127.99405
7	SteamUse_Proportion	8.35286
5	ENERGYSTARScore_KNN	6.81848
9	AncienneteBatiment	4.13828
0	OSEBuildingID	2.75987
3	NumberofBuildings	2.04125
4	LargestPropertyUseTypeGFA	1.75981

-2 features

	feature	VIF
4	Electricity_Proportion	7.85350
3	ENERGYSTARScore_KNN	6.79006
7	AncienneteBatiment	4.07662
6	NaturalGas_Proportion	3.28454
0	OSEBuildingID	2.72844
1	NumberofBuildings	2.03883
2	LargestPropertyUseTypeGFA	1.75427
5	SteamUse_Proportion	1.20706





Analyses exploratoires



2016_Building_Energy_Benchmarking.csv
2016_Building_Energy_Benchmarking_cleaned.csv

~~3 376 lignes~~
1 610 lignes



46 colonnes
17 colonnes

➤ Descriptif des bâtiments :

- Des informations **administratives**: ancienneté du bâtiment, type de commerce, etc.
- Des informations **structurelles**: nombre de bâtiment, surface, etc.
- Des informations **géographiques** : quartier.
- Des informations **énergétiques** : quantité de consommation d'énergie et proportion des sources d'énergie.
- Des informations de **pollution** : quantité d'émission de CO2.

➤ Valeurs manquantes :

- ~~13 % de NaN~~
- 2 % de Nan
- ~~26 / 46 colonnes concernées~~
- 1 / 17 colonnes concernées



Analyses exploratoires

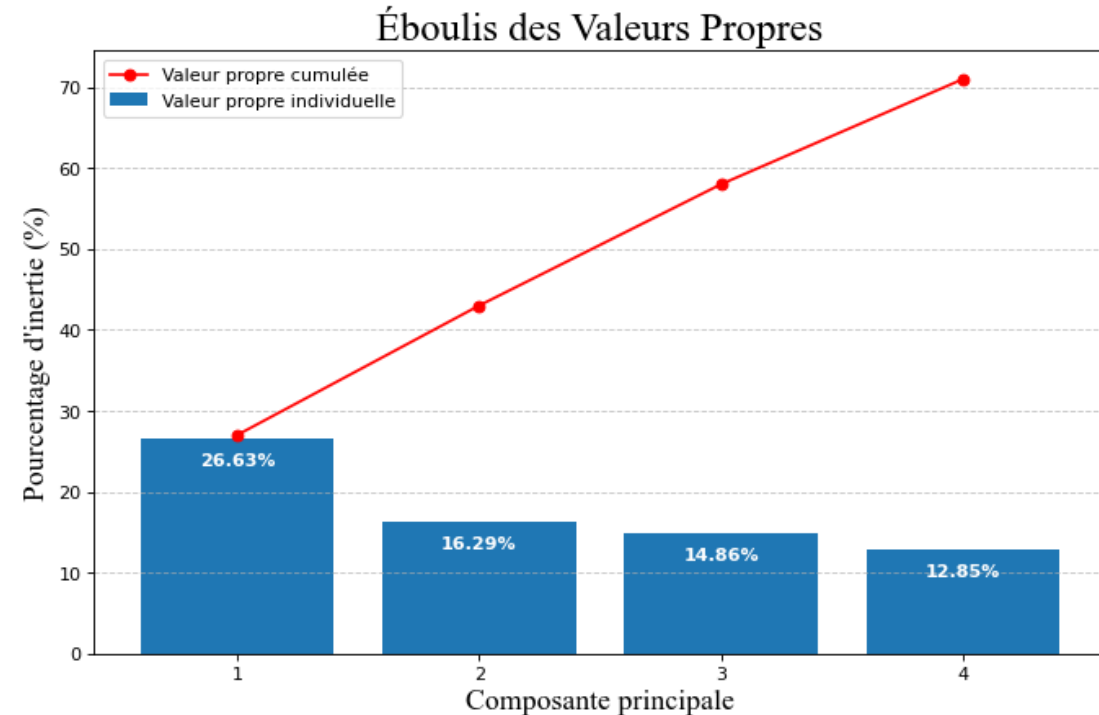


Analyse en Composantes Principales (ACP)



8 composantes expliquent 100% des données

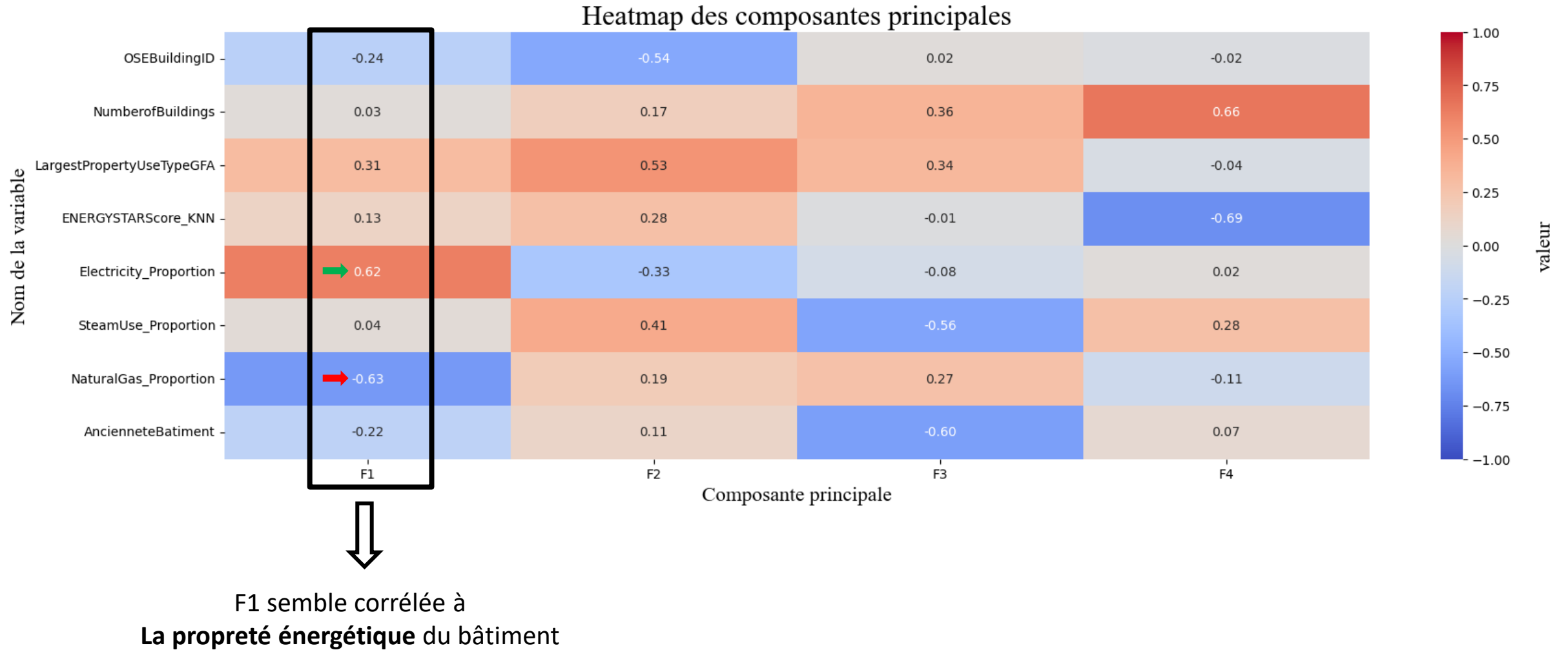
4 composantes > seuil de Kaiser



=> Les 4 premières composantes expliquent > 70% de l'inertie totale

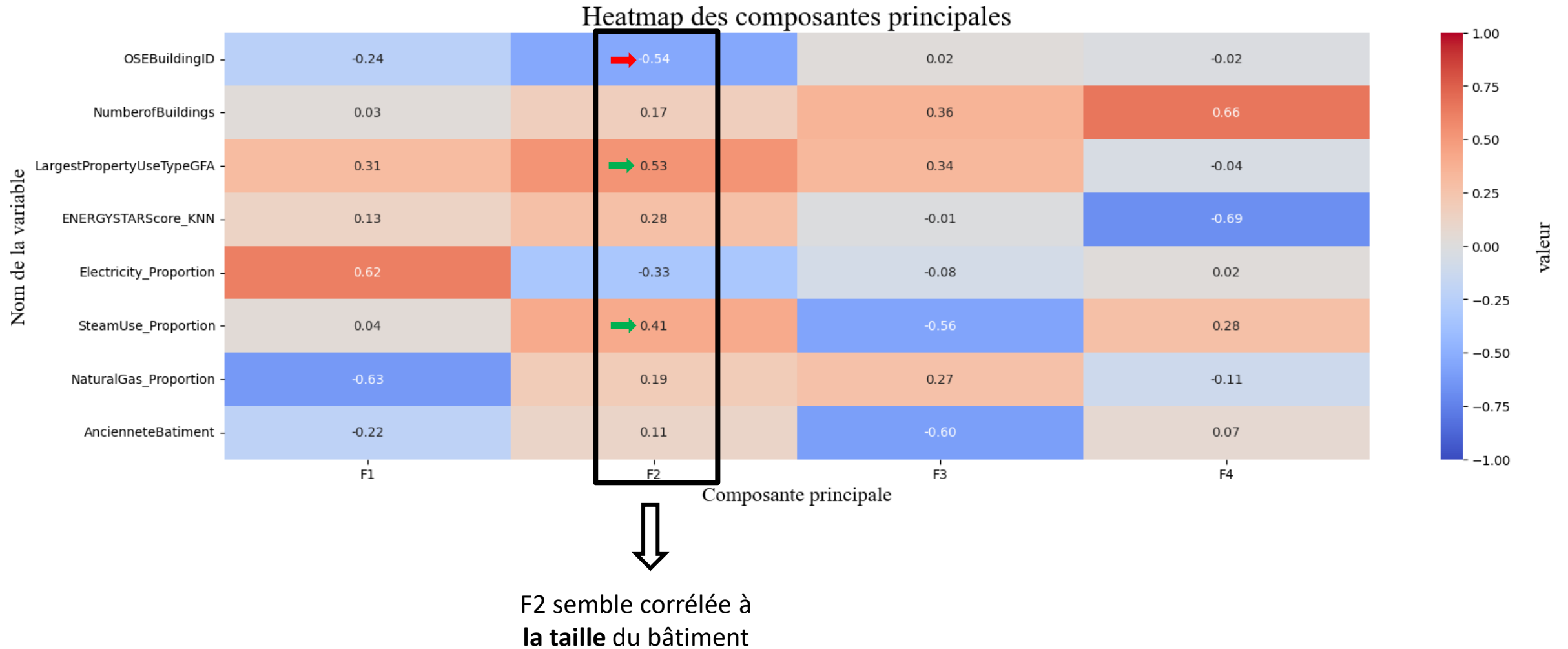


Analyses exploratoires



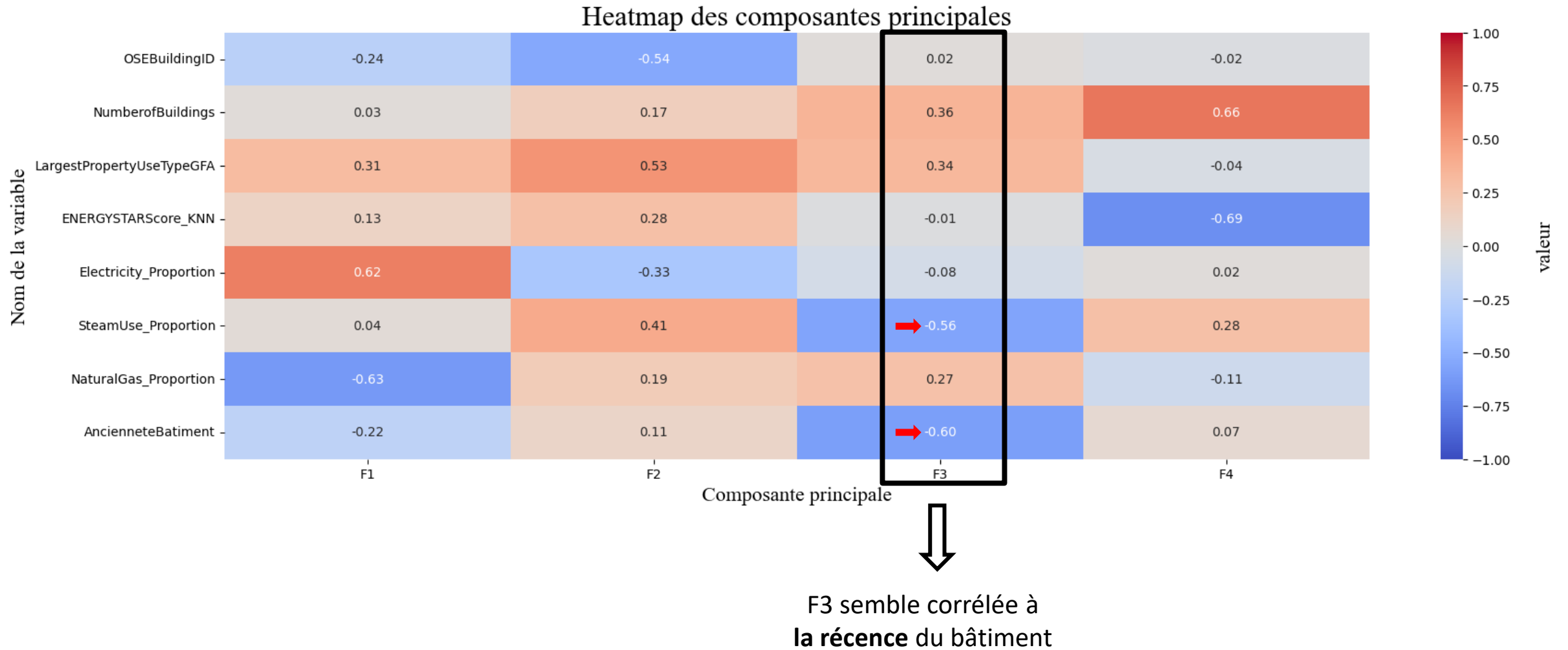


Analyses exploratoires



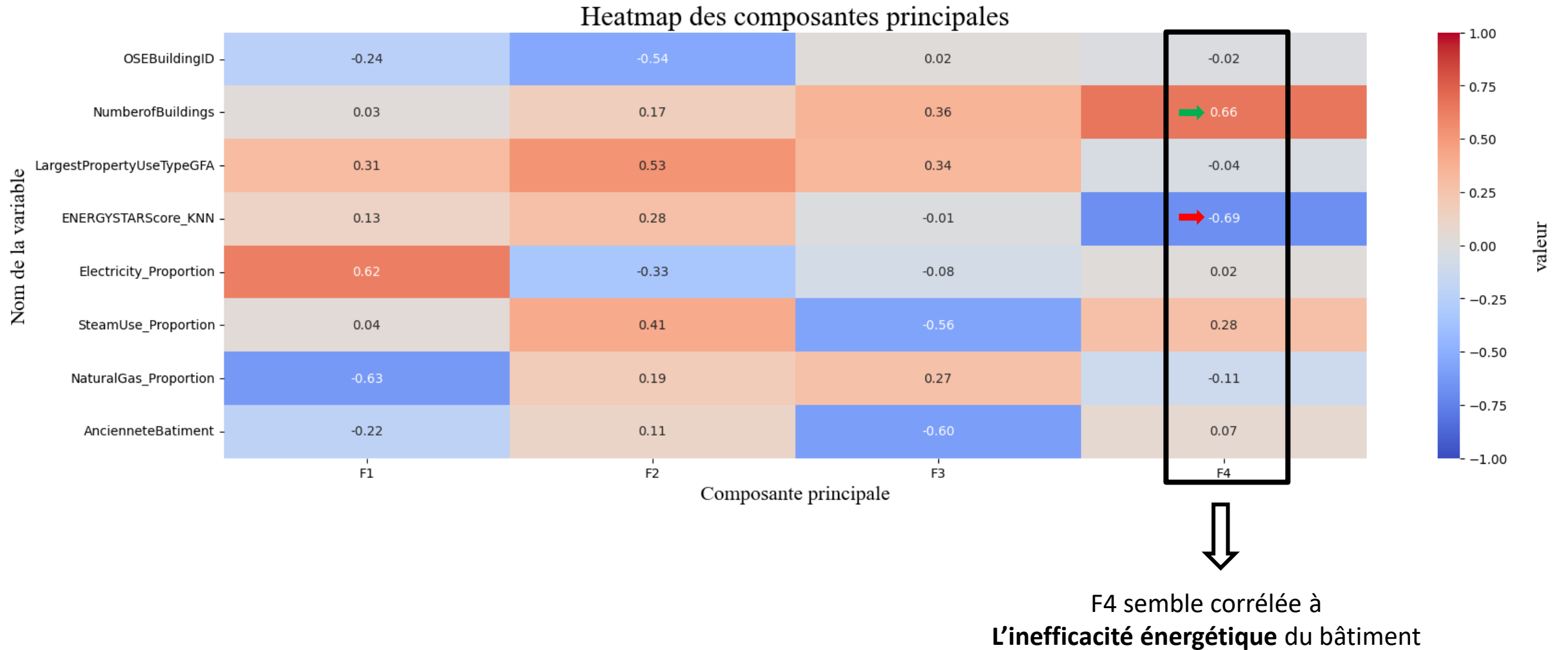


Analyses exploratoires





Analyses exploratoires





Analyses exploratoires



Selon l'ACP, > **70% de l'inertie totale** est expliqué par certaines caractéristiques des bâtiments :

- La propreté énergétique
- La taille
- L'âge
- L'efficacité énergétique



Sommaire



- I – Problématique
- II – Présentation du jeu de données
- III - Nettoyage des données
- IV – Analyses exploratoires
- V – Feature engineering**
- VI – Modèle de prédiction - Energie totale
- VII – Modèle de prédiction - Emission CO₂
- VIII - Conclusion



Feature engineering



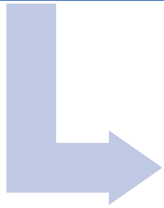
1. Renommer feature cible



Feature engineering



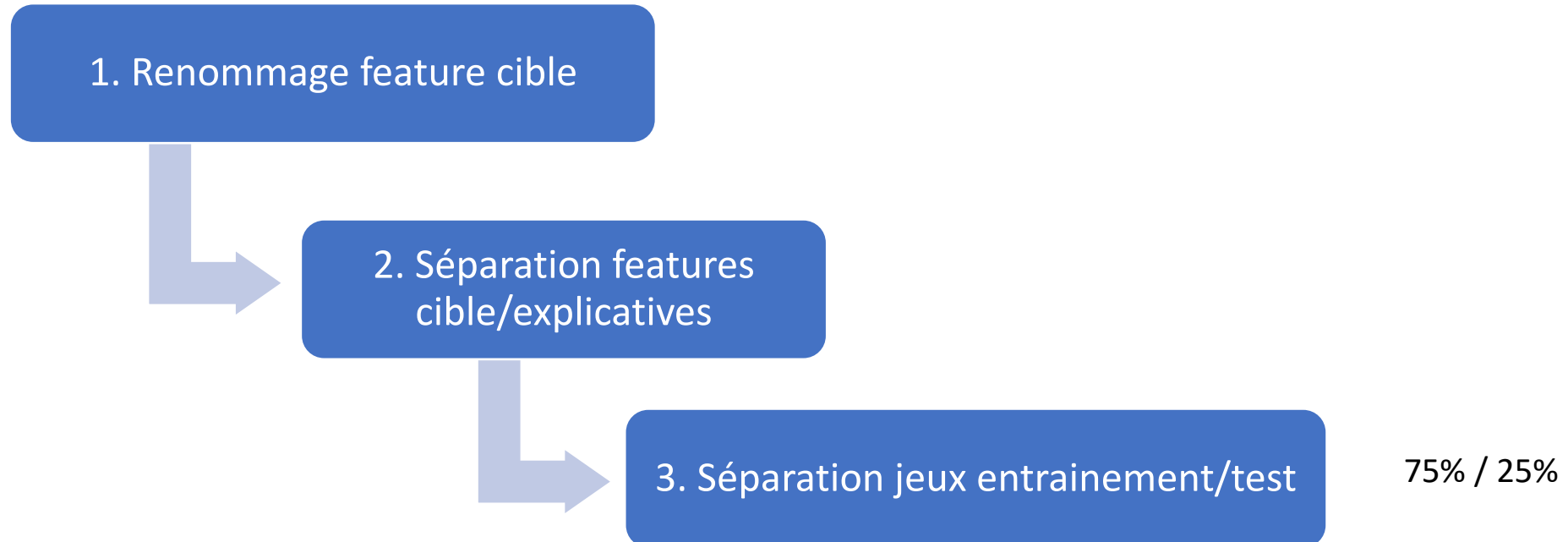
1. Renommage feature cible



2. Séparation features
cible/explicatives

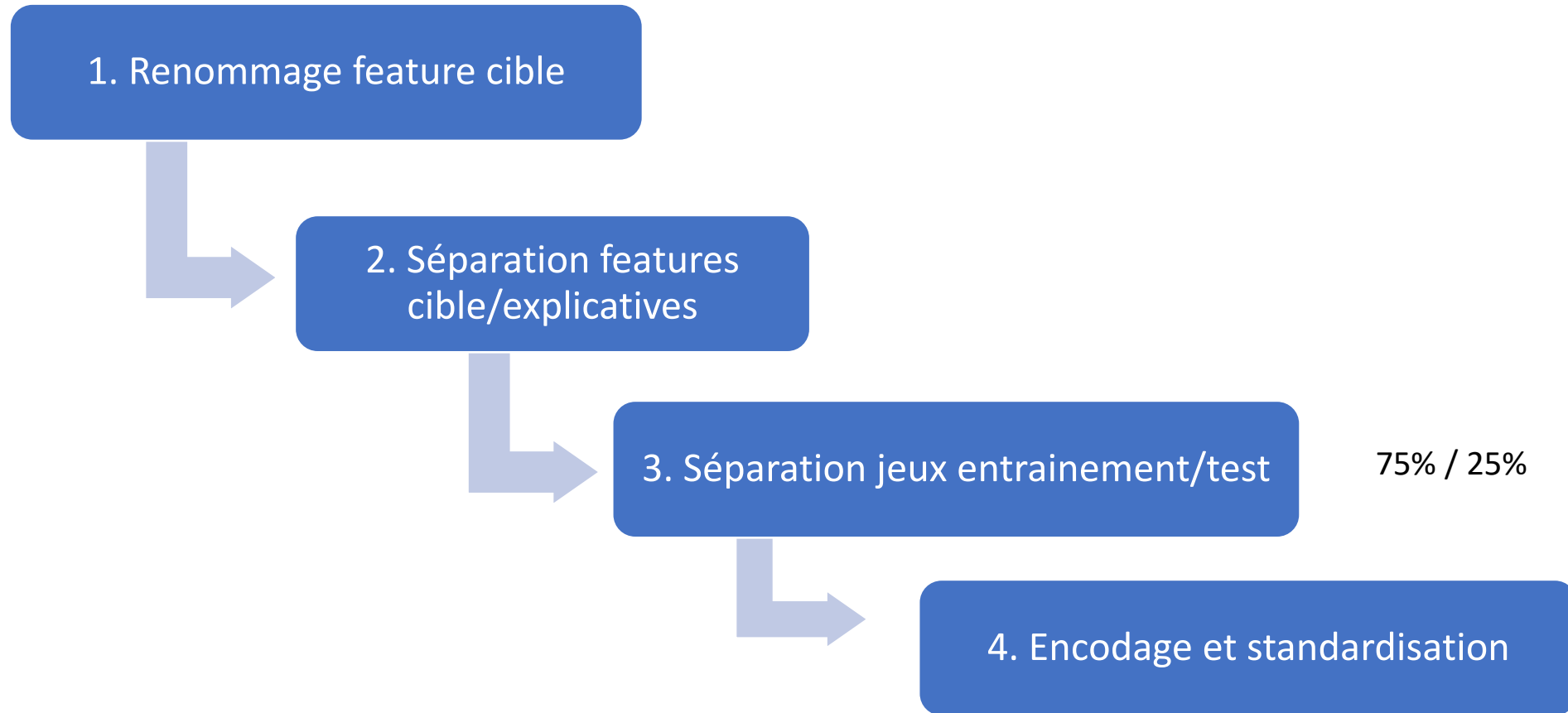


Feature engineering





Feature engineering



Qualitatif => SimpleImputer (valeur la + fréquente) + OneHotEncoder

Quantitatif => SimpleImputer (médiane) + StandardScaler



Sommaire



I – Problématique

II – Présentation du jeu de données

III - Nettoyage des données

IV – Analyses exploratoires

V – Feature engineering

VI – Modèle de prédiction - Energie totale

VII – Modèle de prédiction - Emission CO₂

VIII - Conclusion



Modèle de prédiction – Energie totale ⚡

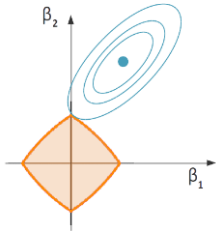


Modèle x 5



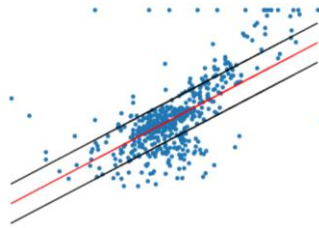
Dummy regressor
(moyenne)

VS



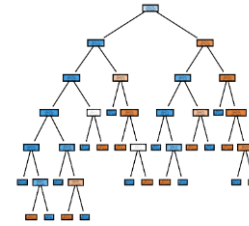
ElasticNet

VS



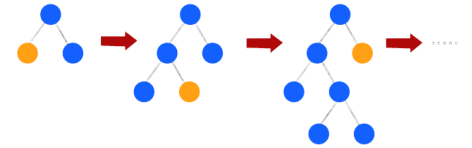
SVR

VS



RandomForest

VS



GradientBoosting

GridSearchCV

	0.01	0.1	1.0	10.0
3				
5				
7				
9				

Performance

ERROR

R^2



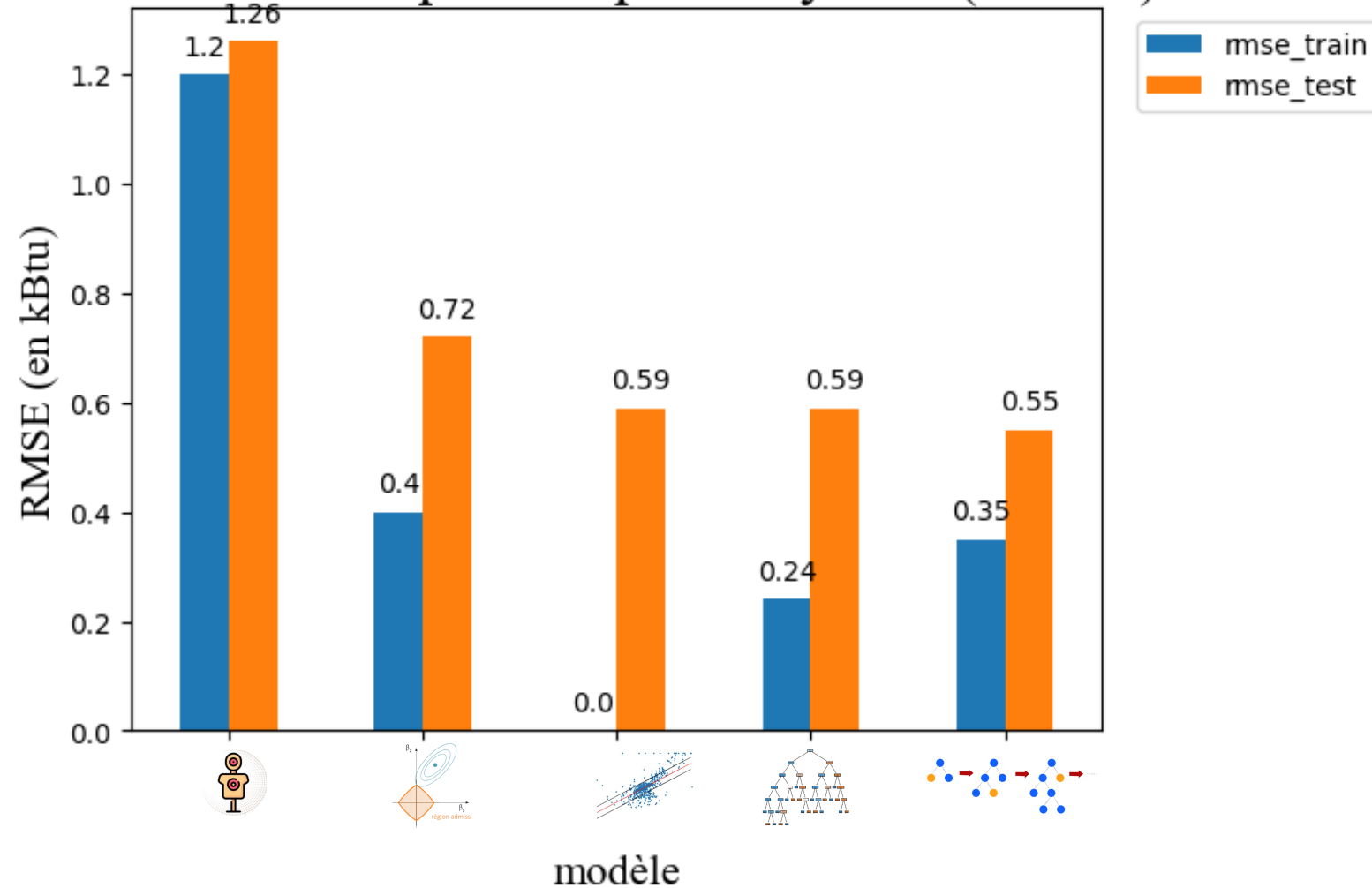


Modèle de prédiction – Energie totale ⚡



Comparaison de l'erreur quadratiques moyenne (RMSE) de modèle

ERROR



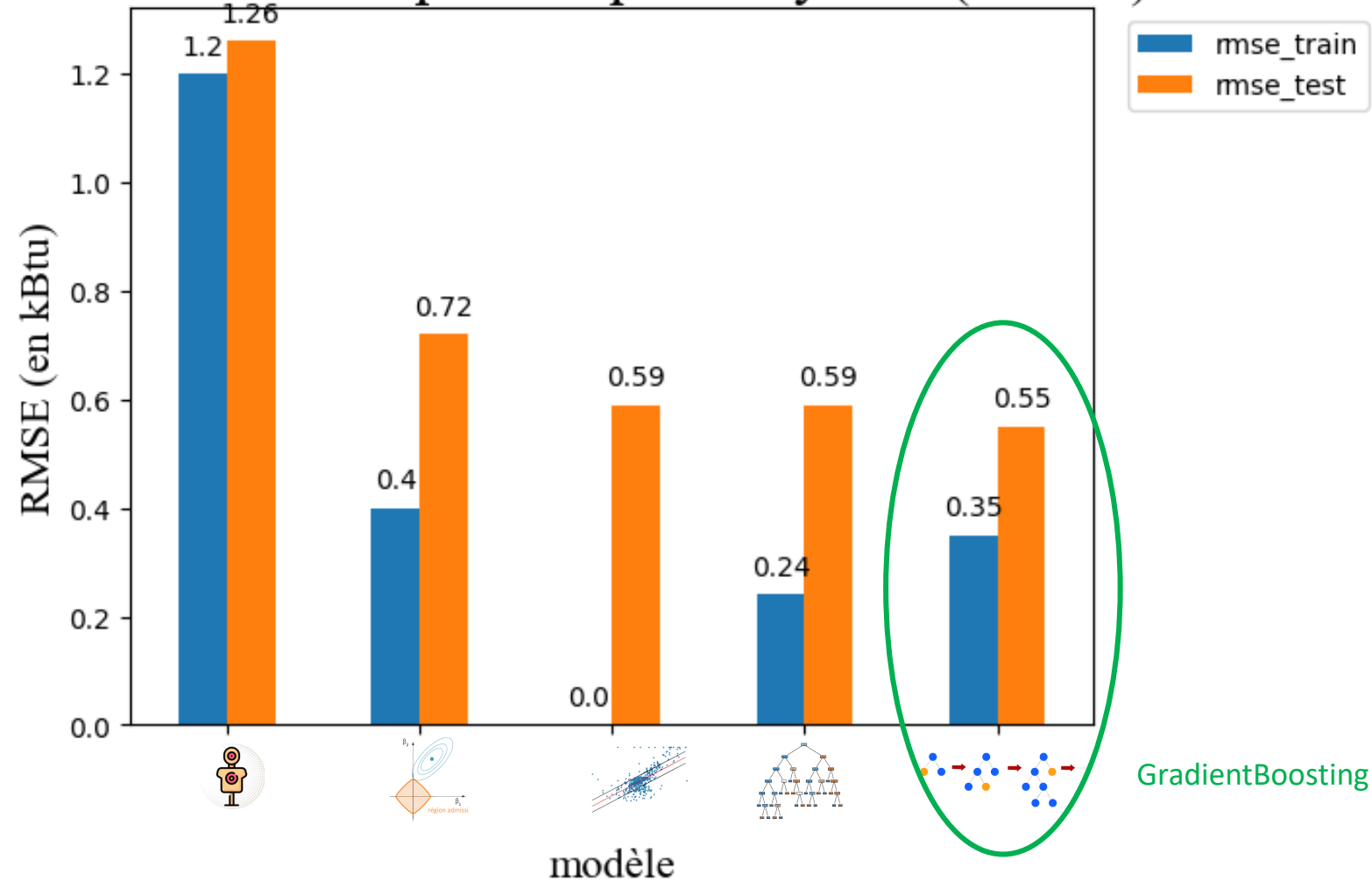


Modèle de prédiction – Energie totale ⚡



Comparaison de l'erreur quadratiques moyenne (RMSE) de modèle

ERROR



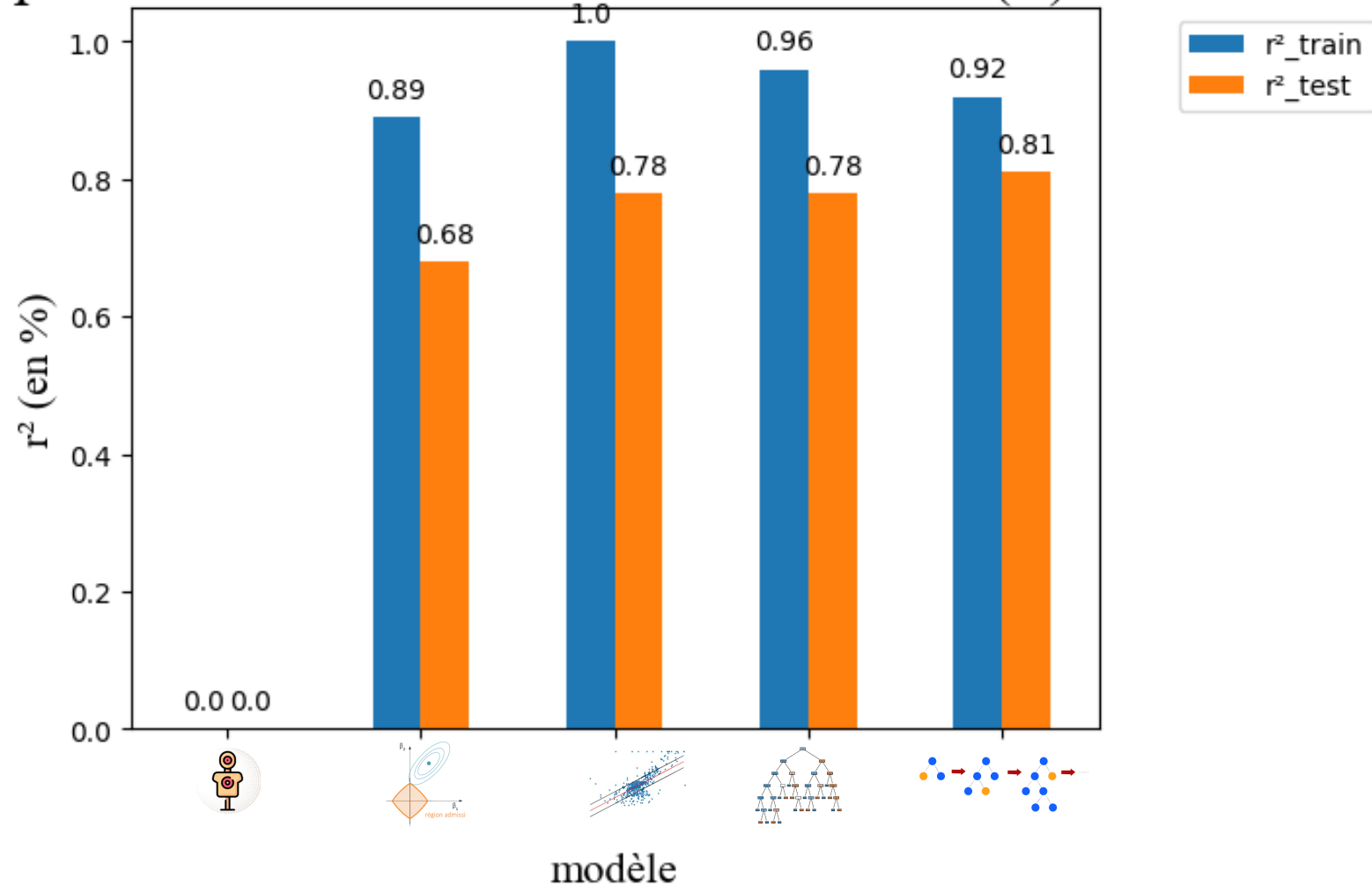


Modèle de prédiction – Energie totale ⚡



R^2

Comparaison du coefficient de détermination (r^2) de modèle



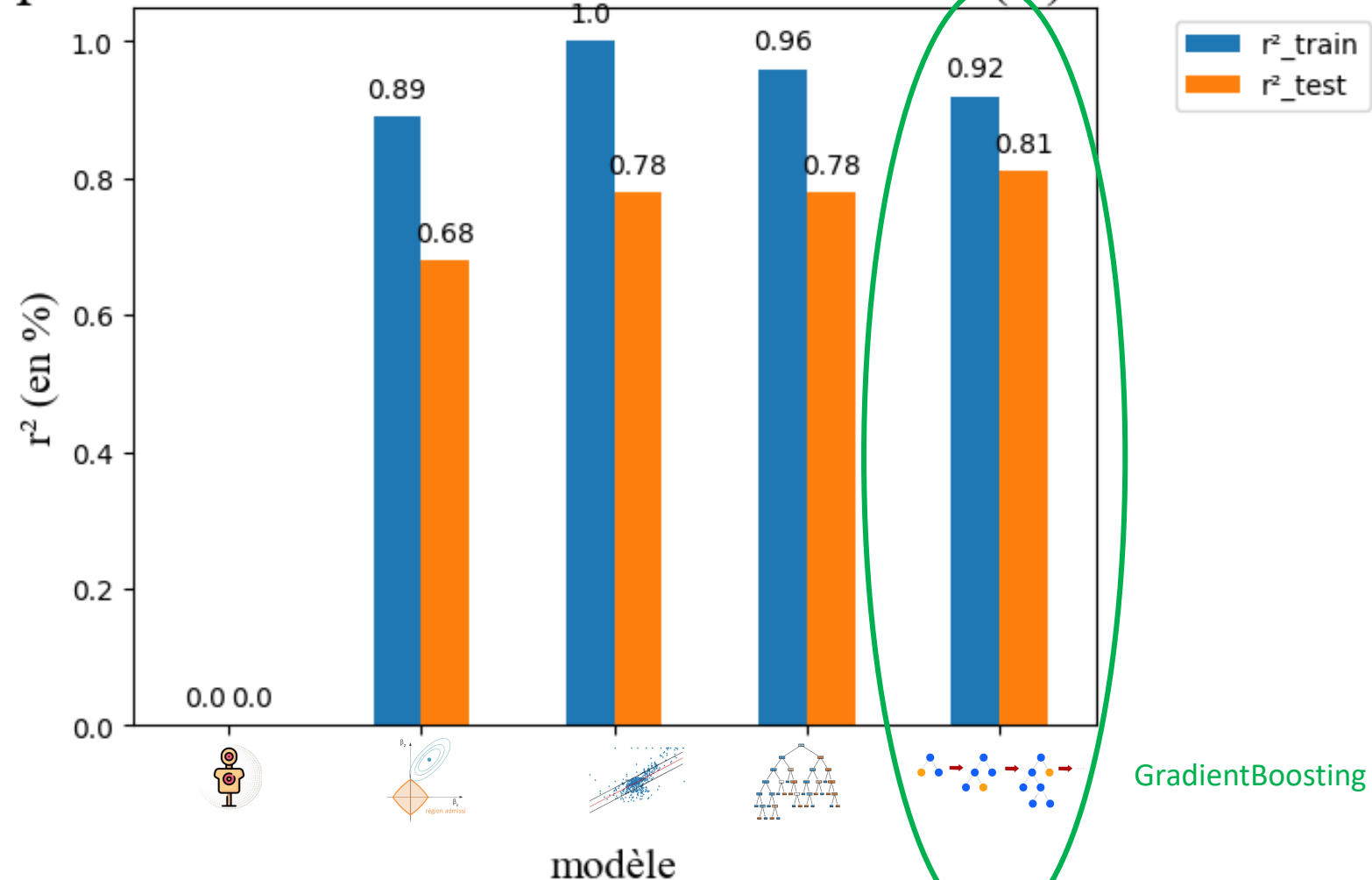


Modèle de prédiction – Energie totale ⚡



R^2

Comparaison du coefficient de détermination (r^2) de modèle

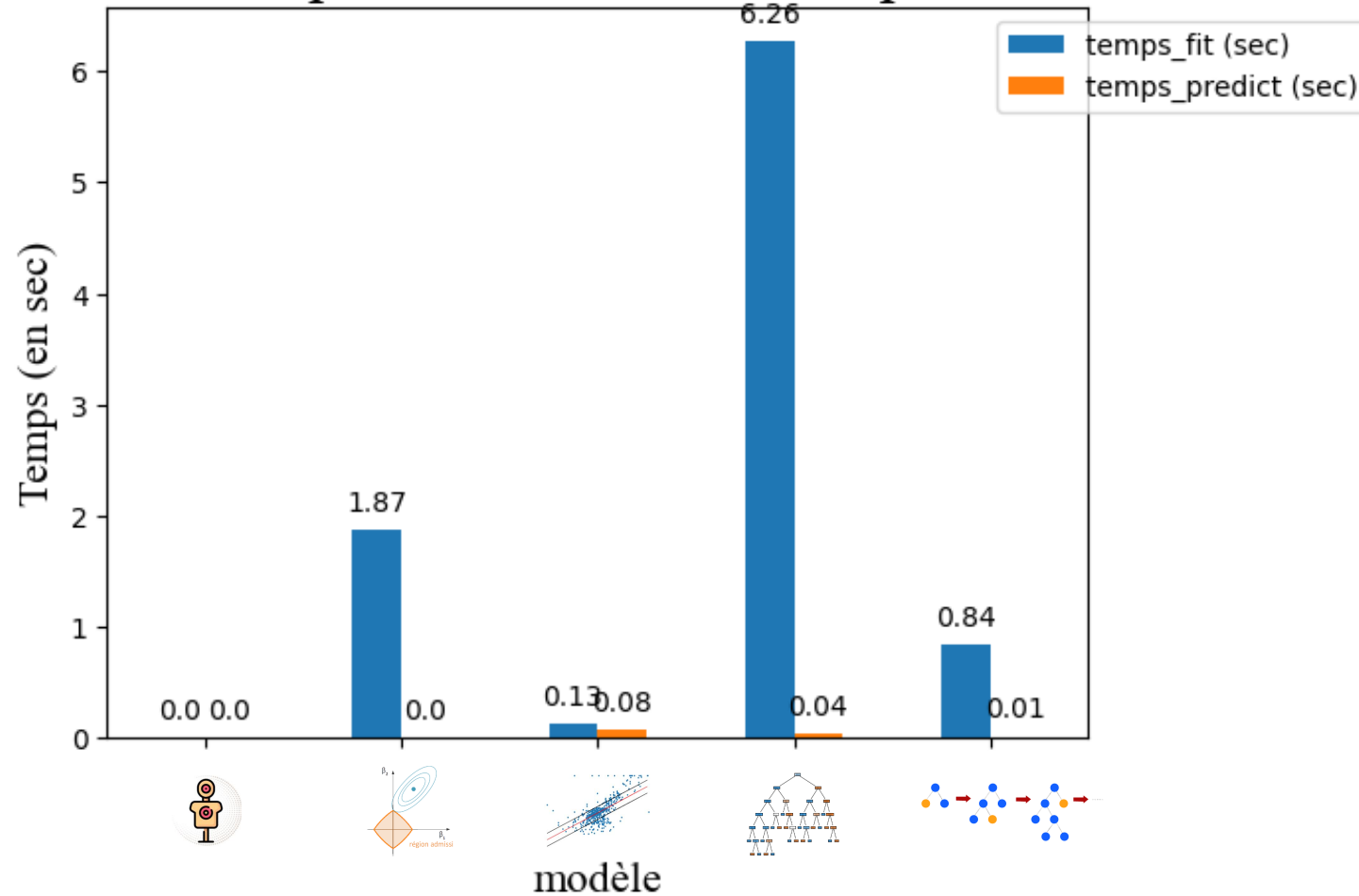




Modèle de prédiction – Energie totale ⚡



Comparaison du temps d'entrainement et de prédiction de modèle

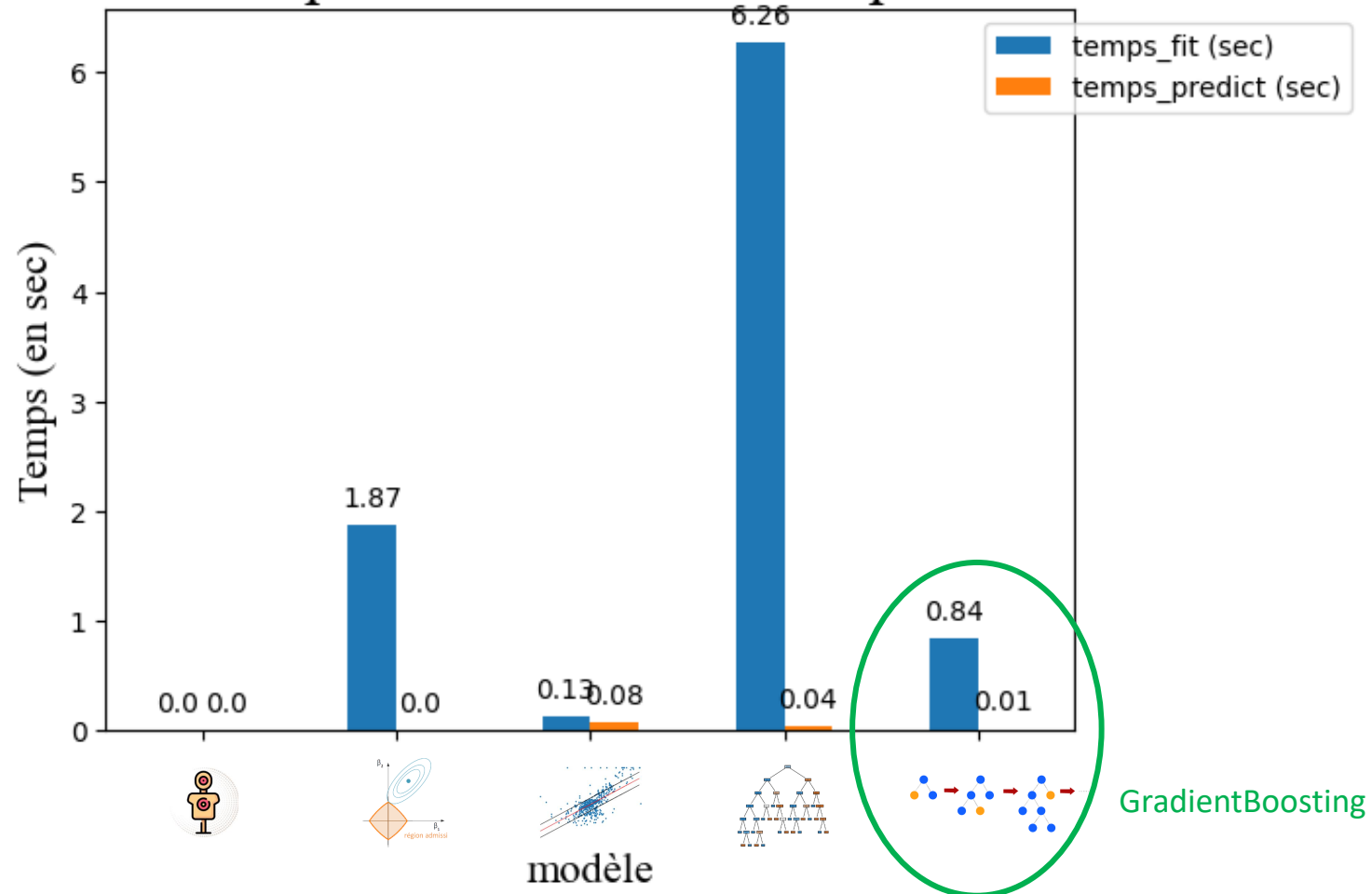




Modèle de prédiction – Energie totale ⚡

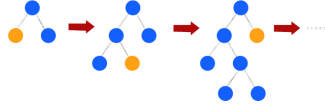


Comparaison du temps d'entrainement et de prédiction de modèle





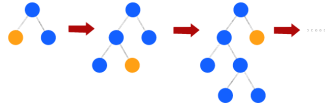
Modèle de prédiction – Energie totale



Feature importance



Modèle de prédiction – Energie totale ⚡

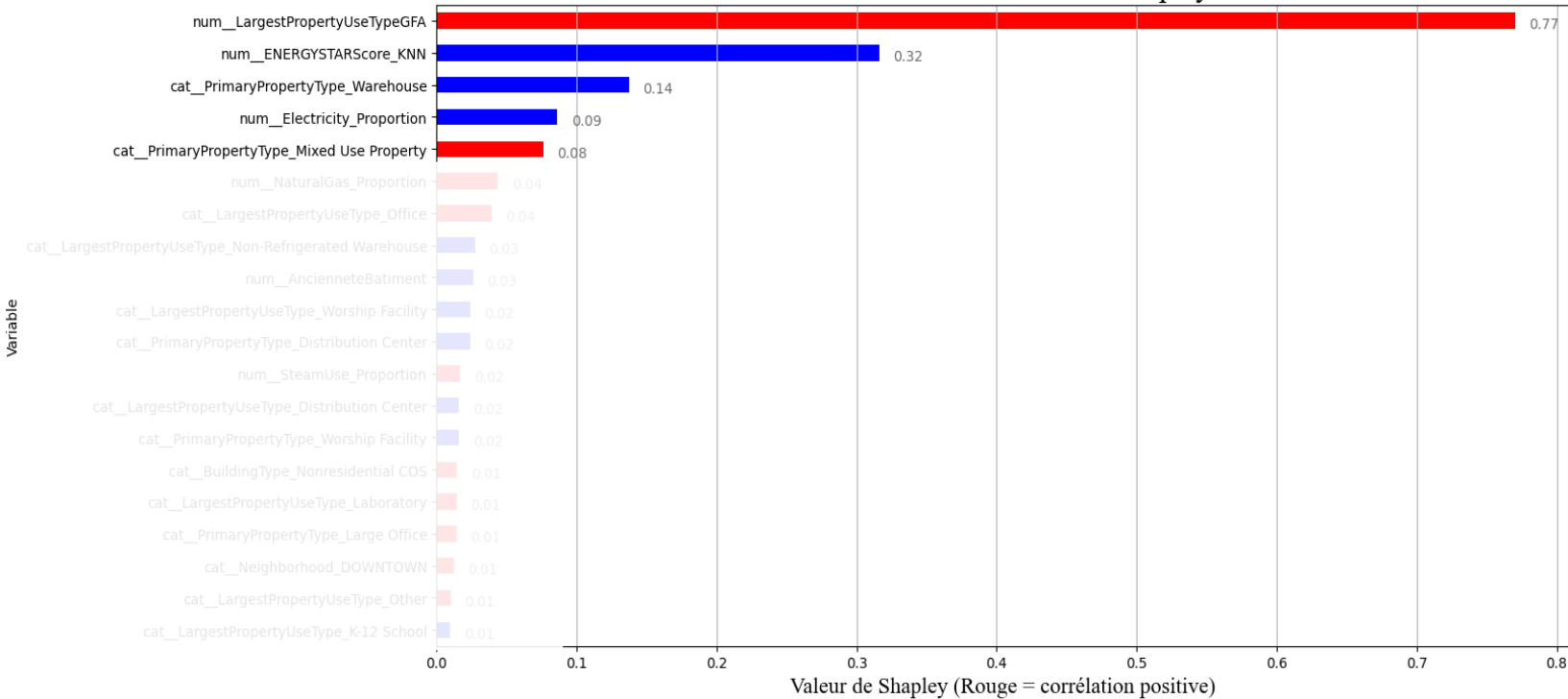


Feature importance

Globale

Locale

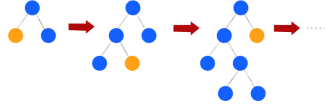
Corrélation des valeurs de Shapley




feature	importance
num_LargestPropertyUseTypeGFA	1.09185
num_ENERGYSTARScore_KNN	0.19605
cat_PrimaryPropertyType_Warehouse	0.06851



Modèle de prédiction – Energie totale



Apport =>  ?



Vs



Vs

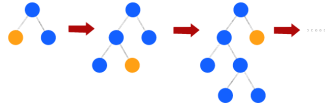



Vs





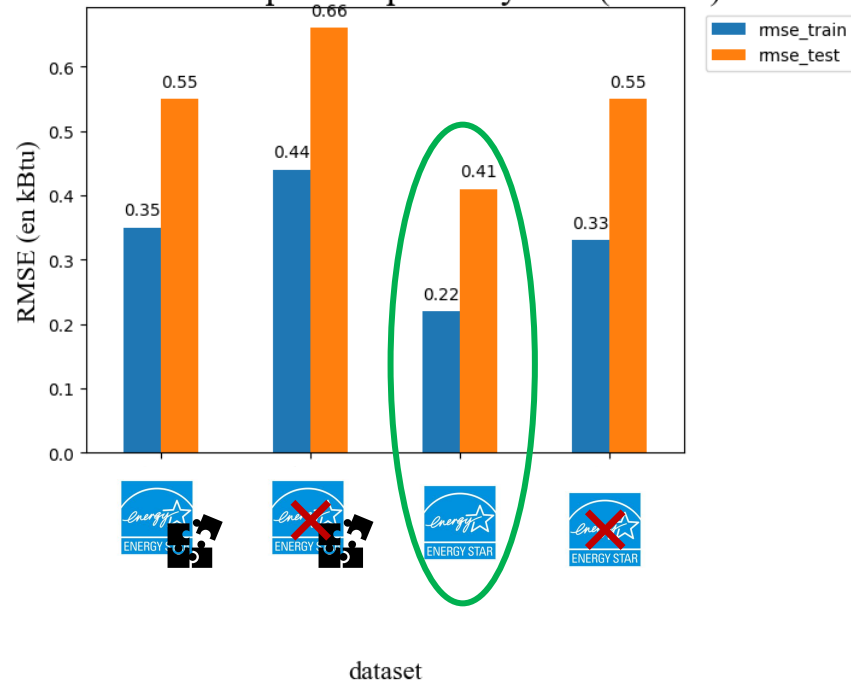
Modèle de prédiction – Energie totale ⚡



Apport =>  ?

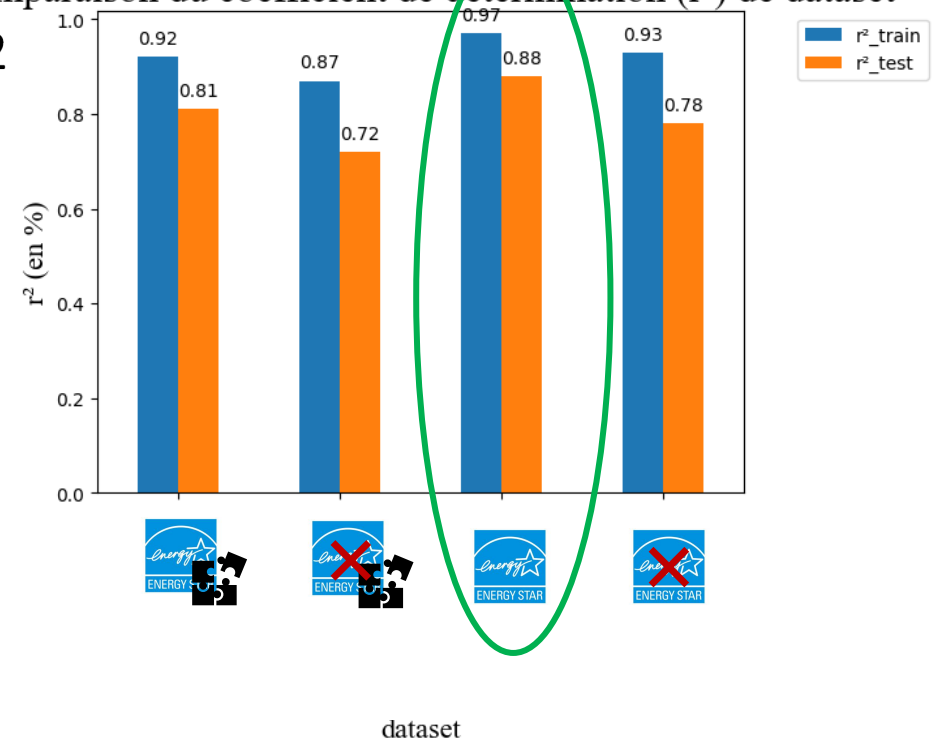
Comparaison de l'erreur quadratique moyenne (RMSE) de dataset

ERROR



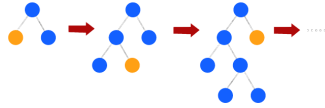
Comparaison du coefficient de détermination (r^2) de dataset


R^2





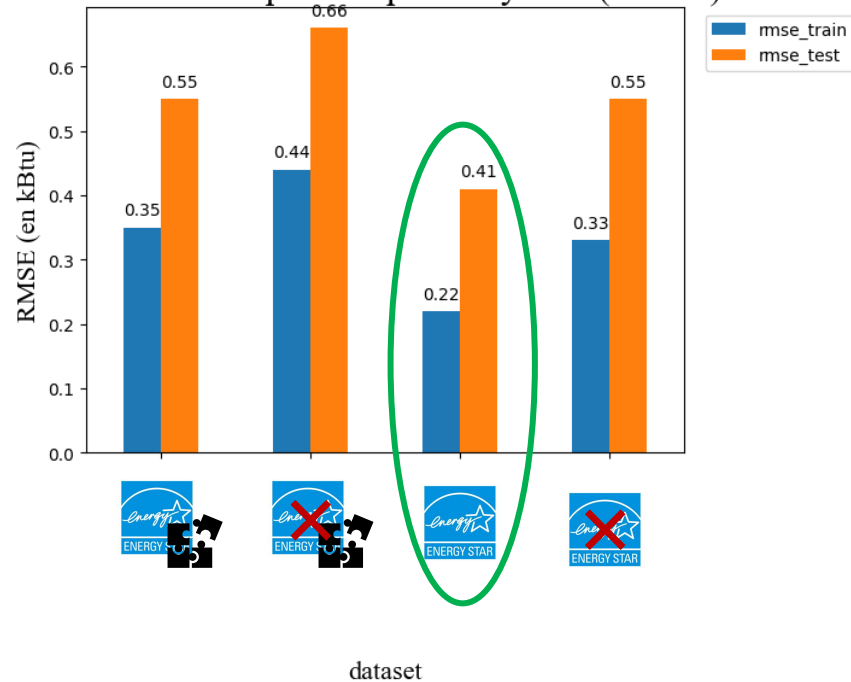
Modèle de prédiction – Energie totale ⚡



Apport =>  ?

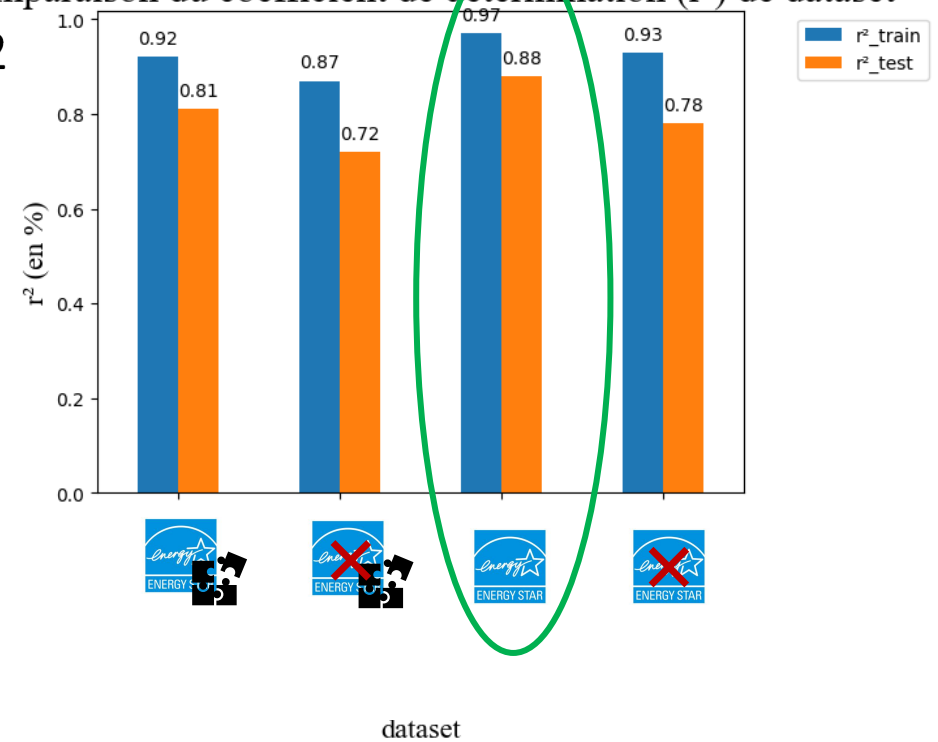
Comparaison de l'erreur quadratique moyenne (RMSE) de dataset

ERROR



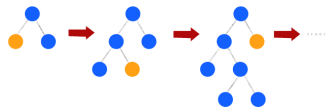
Comparaison du coefficient de détermination (r^2) de dataset

R^2





Modèle de prédiction – Energie totale ⚡



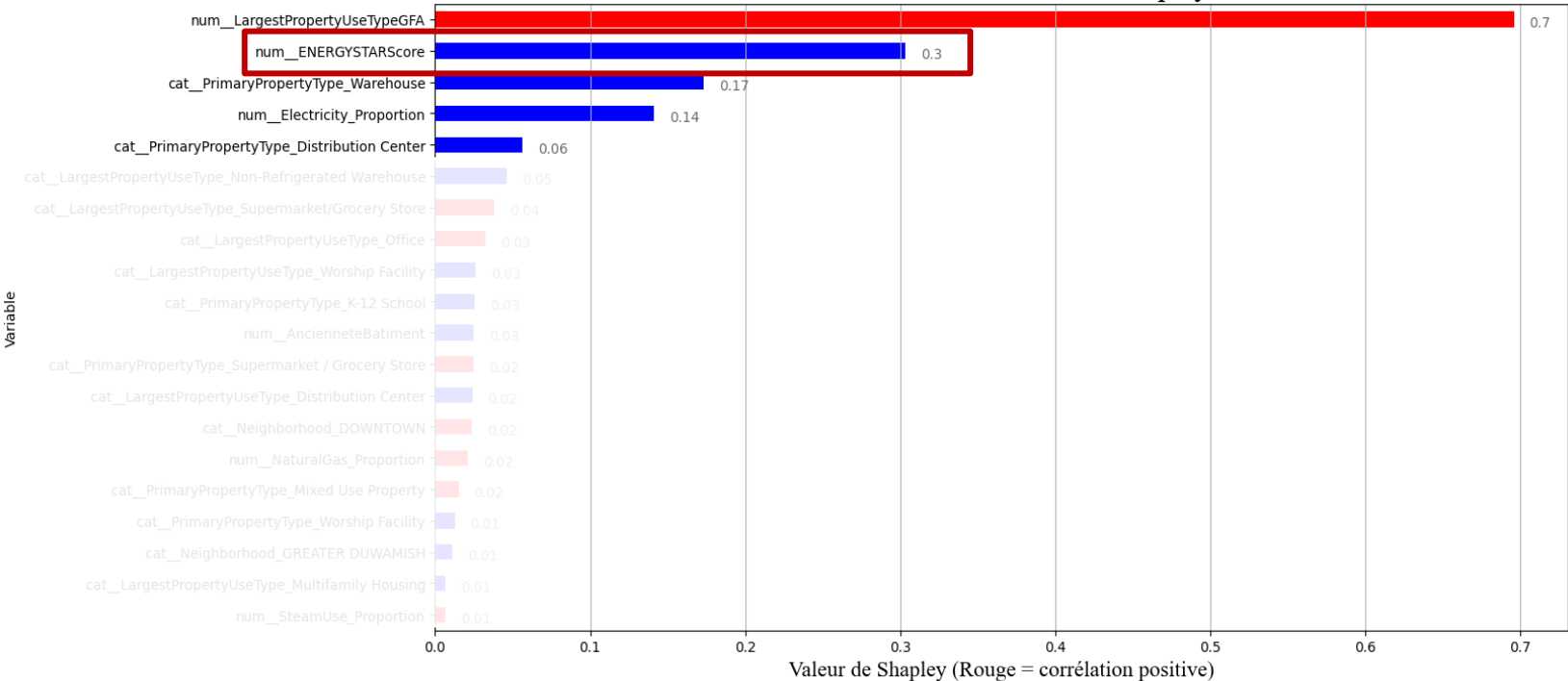
Feature importance

Globale



Locale

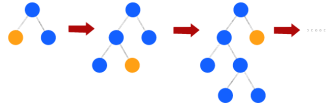
Corrélation des valeurs de Shapley



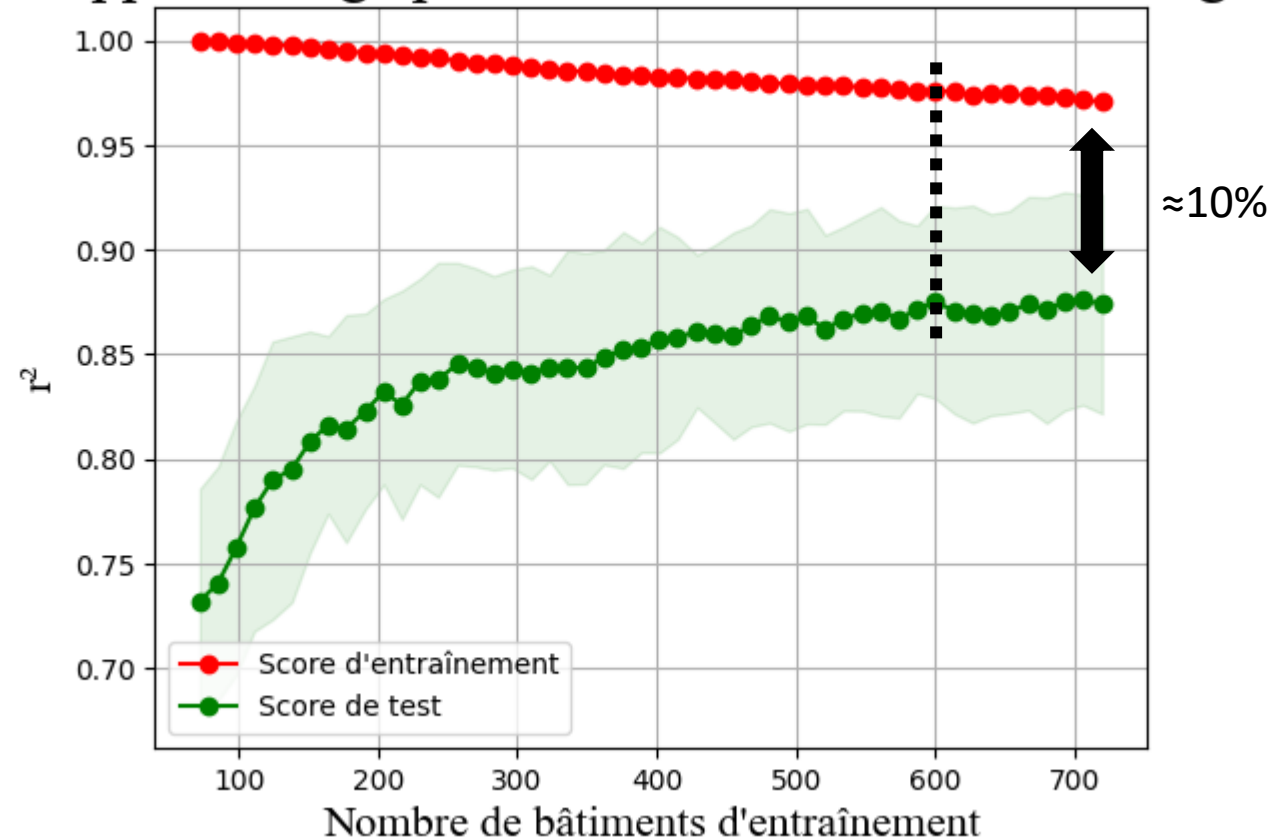
feature	importance
num_LargestPropertyUseTypeGFA	1.10980
num_ENERGYSTARScore	0.23139
cat_PrimaryPropertyType_Warehouse	0.09489
num_Electricity_Proportion	0.06138



Modèle de prédiction – Energie totale ⚡



Courbe d'apprentissage pour le modèle GradientBoostingRegressor





Sommaire



I – Problématique

II – Présentation du jeu de données

III - Nettoyage des données

IV – Analyses exploratoires

V – Feature engineering

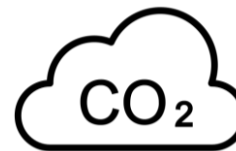
VI – Modèle de prédiction - Energie totale

VII – Modèle de prédiction - Emission CO₂

VIII - Conclusion



Modèle de prédiction – Emission

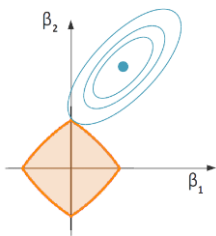


Modèle x 5



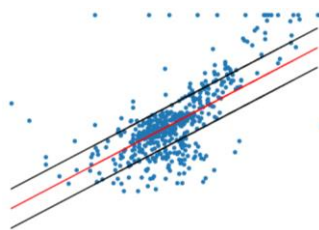
Dummy regressor
(moyenne)

VS



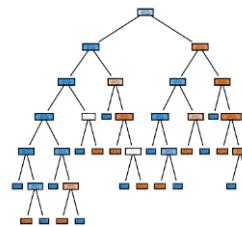
ElasticNet

VS



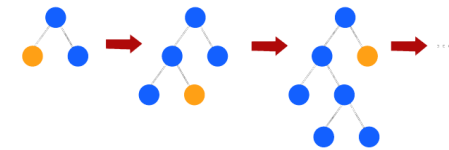
SVR

VS



RandomForest

VS



GradientBoosting

GridSearchCV

	0.01	0.1	1.0	10.0
3				
5				
7				
9				

Performance

ERROR

R^2

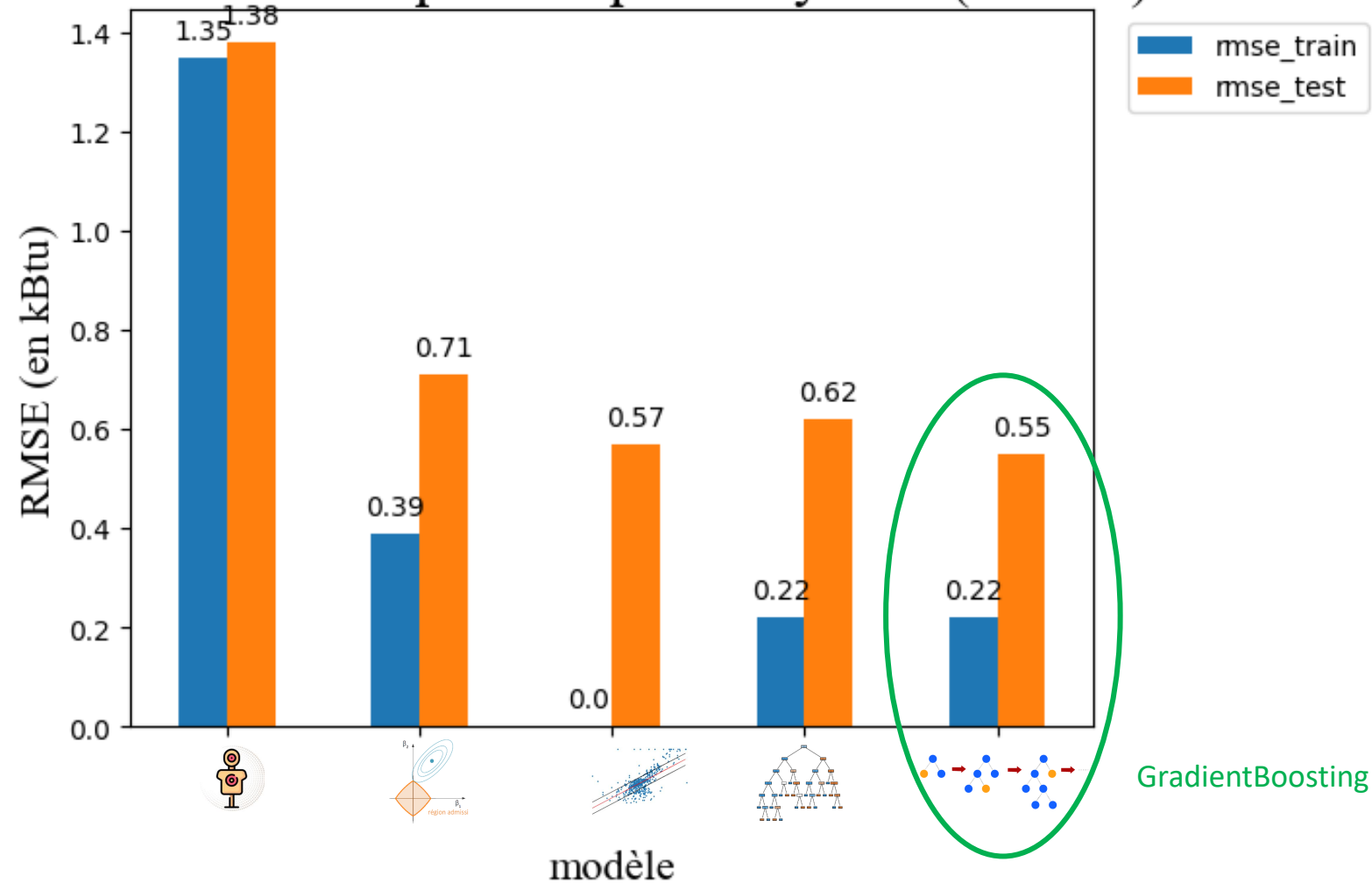




Modèle de prédiction – Emission CO₂



Comparaison de l'erreur quadratiques moyenne (RMSE) de modèle



ERROR

GradientBoosting

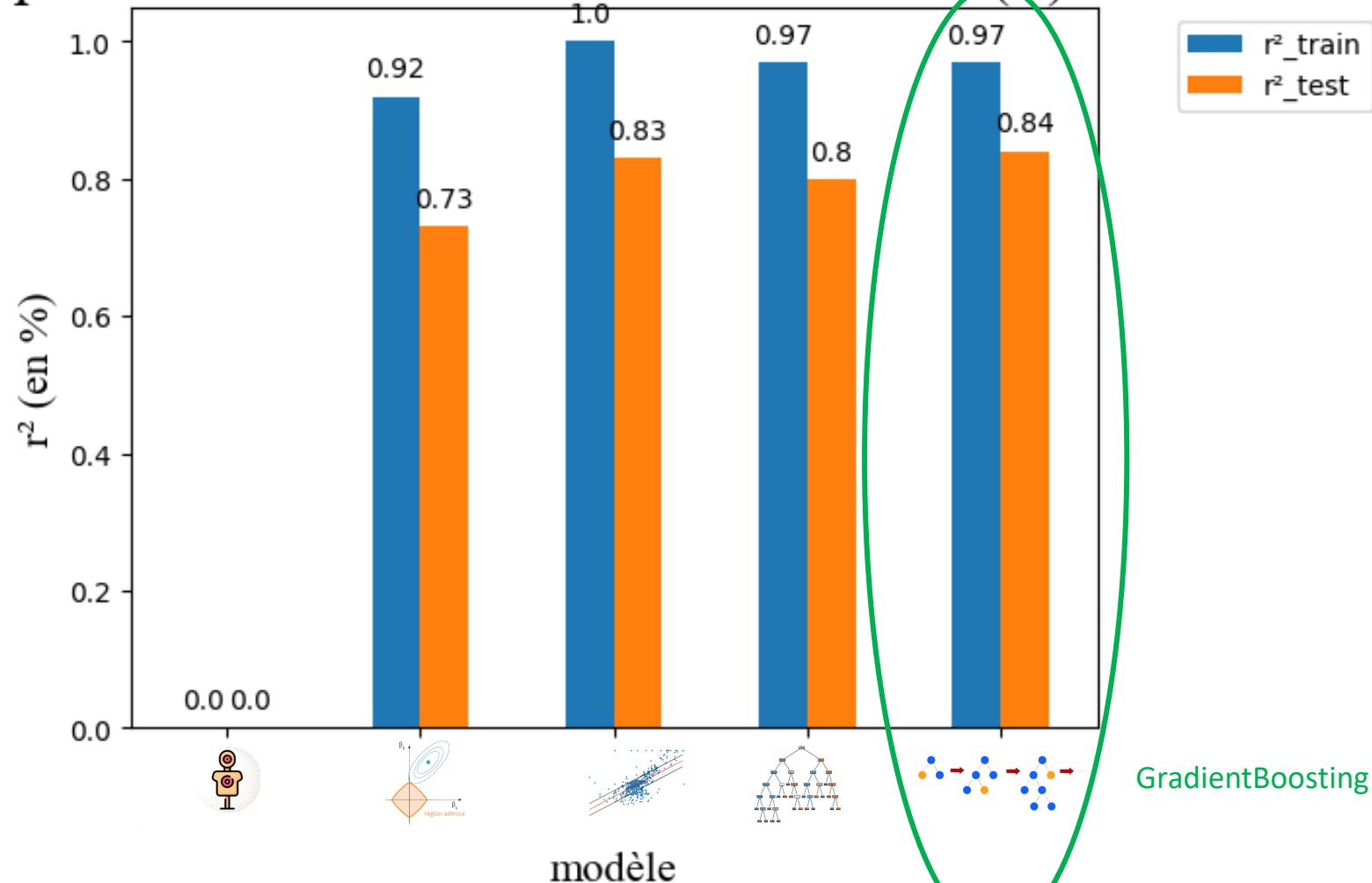


Modèle de prédiction – Emission



R^2

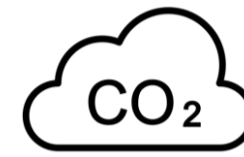
Comparaison du coefficient de détermination (r^2) de modèle



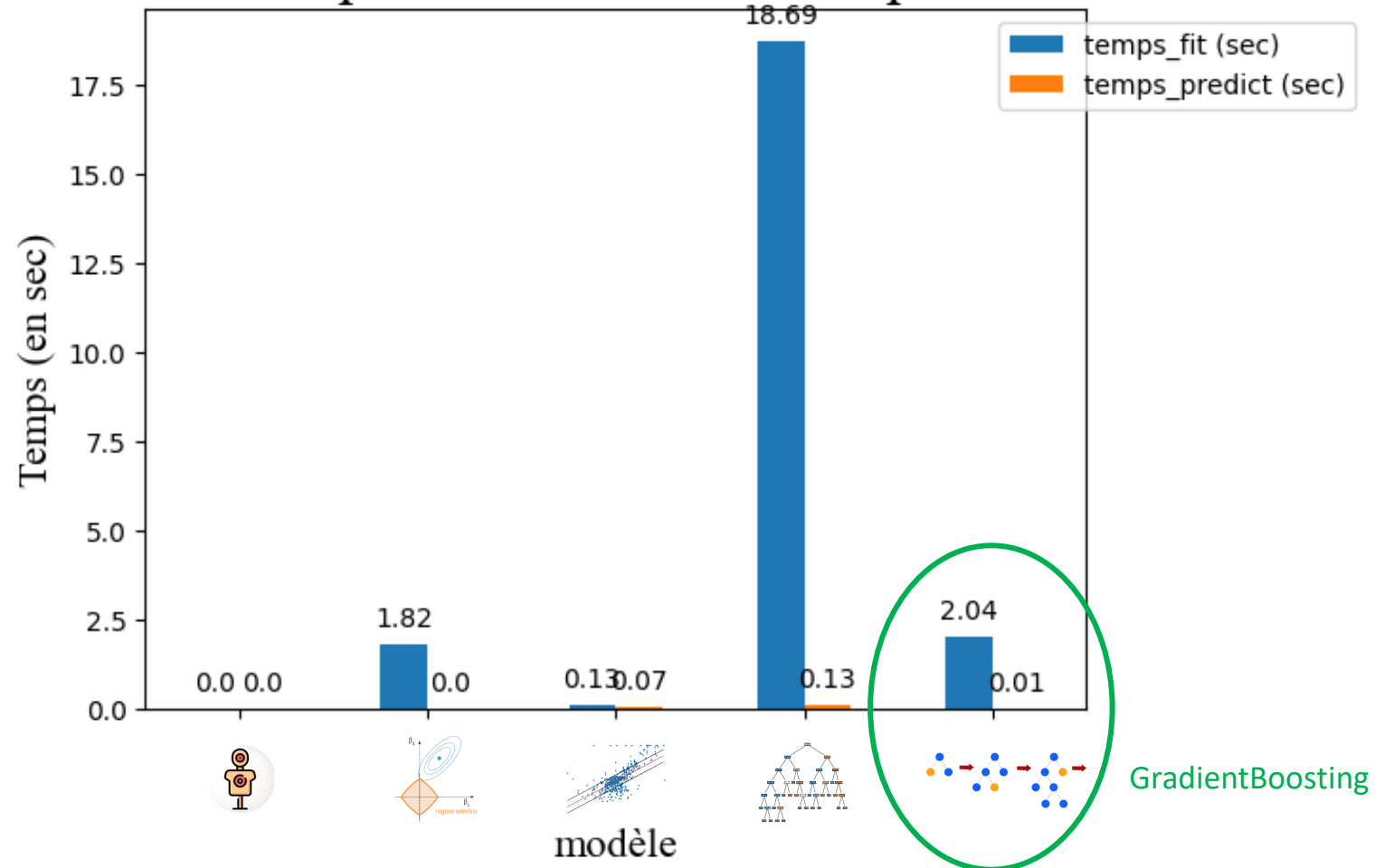
GradientBoosting



Modèle de prédiction – Emission

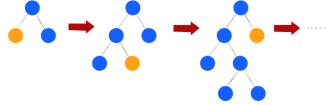


Comparaison du temps d'entrainement et de prédiction de modèle





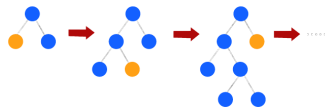
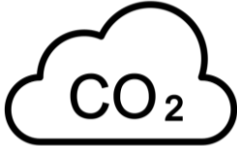
Modèle de prédiction – Emission



Feature importance



Modèle de prédiction – Emission

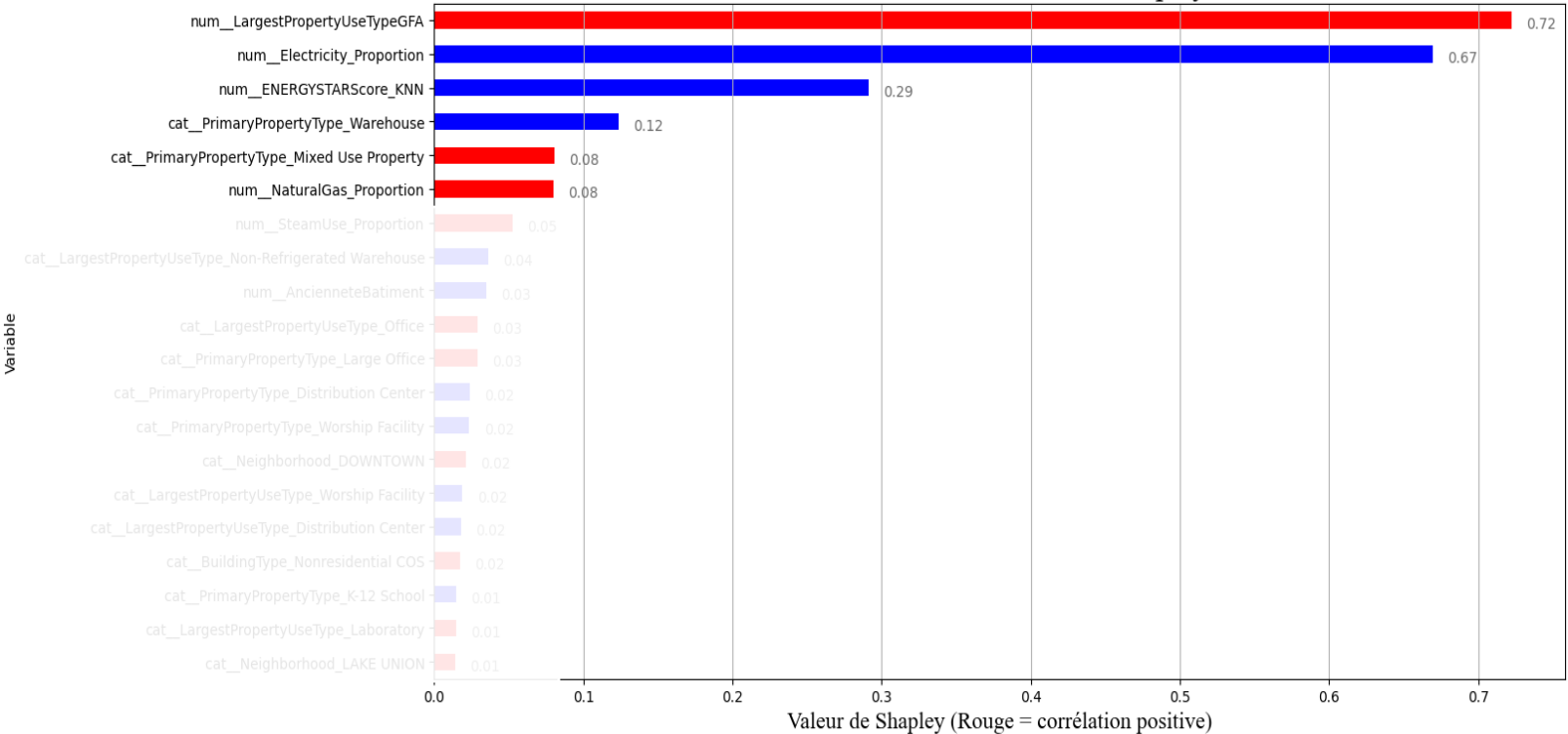


Feature importance

Globale

Locale

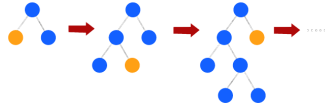
Corrélation des valeurs de Shapley




feature	importance
num_LargestPropertyUseTypeGFA	0.81642
num_Electricity_Proportion	0.57925
num_ENERGYSTARScore_KNN	0.14221



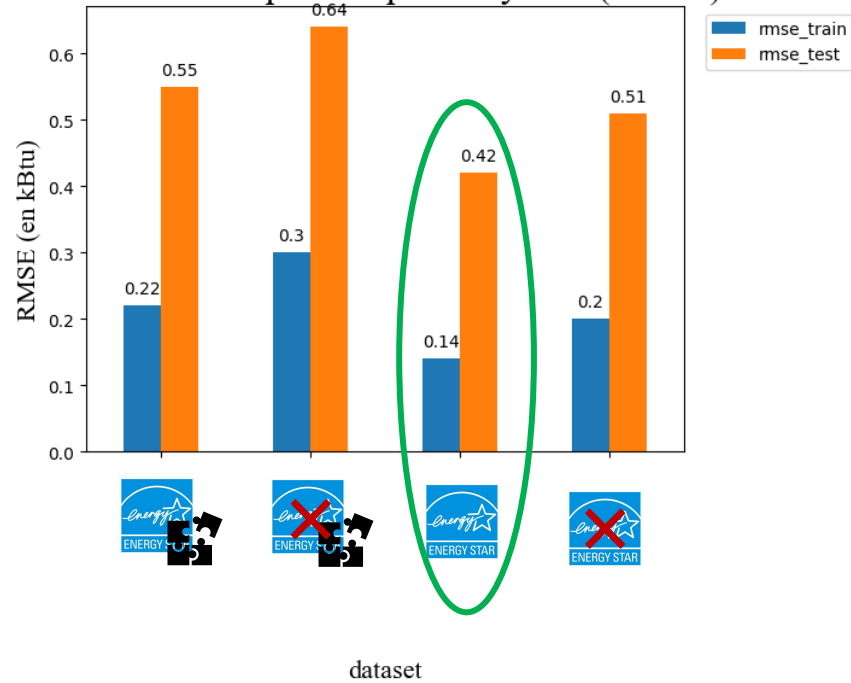
Modèle de prédiction – Emission



Apport =>  ?

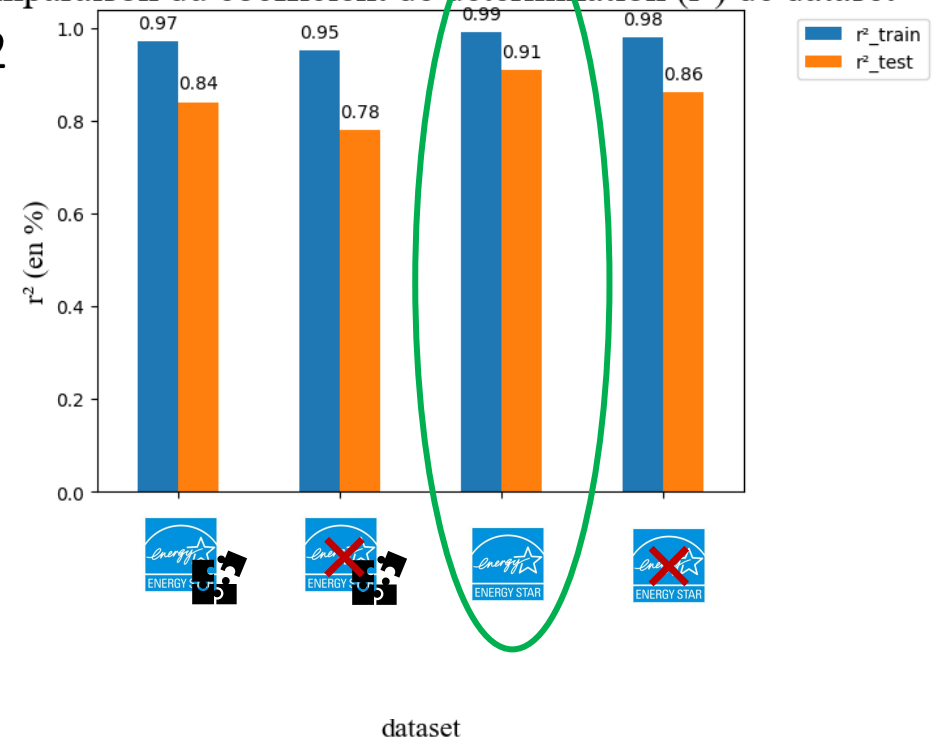
Comparaison de l'erreur quadratique moyenne (RMSE) de dataset

ERROR



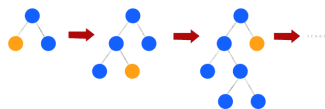
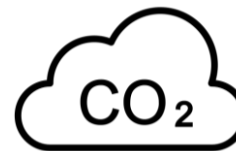
Comparaison du coefficient de détermination (r^2) de dataset

R²





Modèle de prédiction – Emission



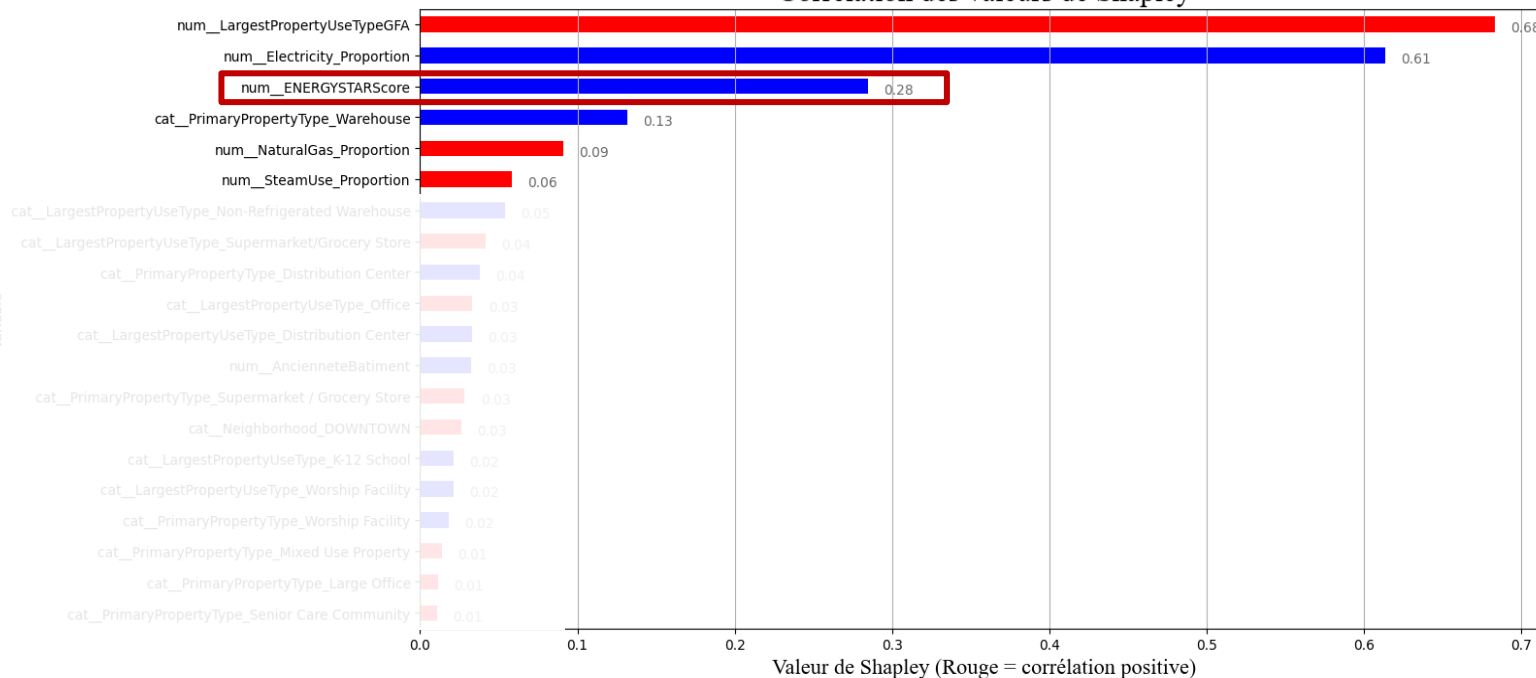
Feature importance

Globale



Locale

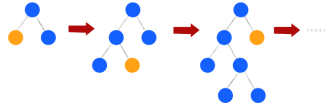
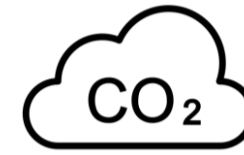
Corrélation des valeurs de Shapley



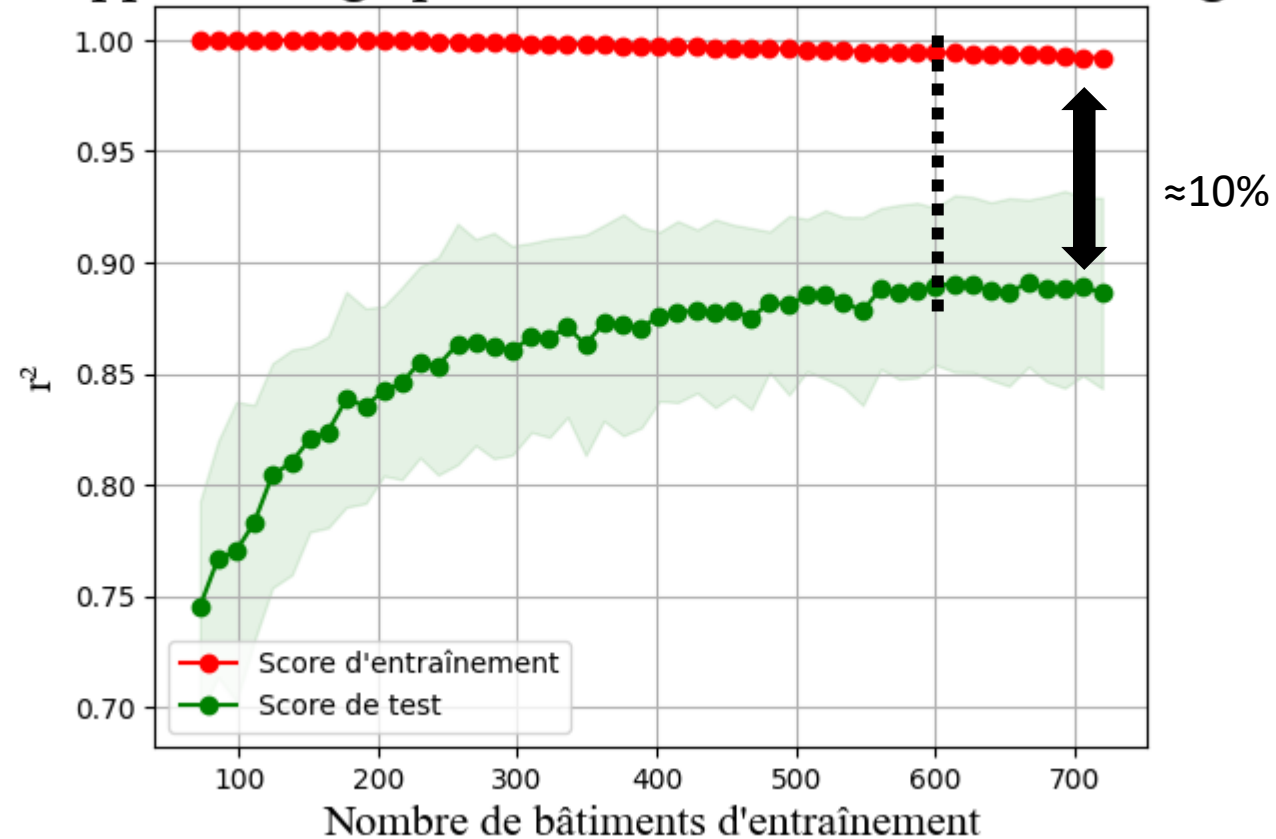
feature	importance
num_LargestPropertyUseTypeGFA	0.76019
num_Electricity_Proportion	0.54434
num_ENERGYSTARScore	0.13505



Modèle de prédiction – Emission



Courbe d'apprentissage pour le modèle GradientBoostingRegressor





Sommaire



- I – Problématique
- II – Présentation du jeu de données
- III - Nettoyage des données
- IV – Analyses exploratoires
- V – Feature engineering
- VI – Modèle de prédiction - Energie totale
- VII – Modèle de prédiction - Emission CO₂
- VIII - Conclusion**

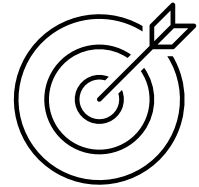


Conclusion



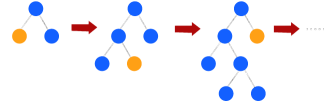
Missions :

1. Réaliser une courte analyse exploratoire. ✓
2. Tester différents modèles de prédiction pour prédire la consommation totale d'énergie. ✓
3. Tester différents modèles de prédiction pour prédire les émissions de CO2. ✓
4. Evaluer l'intérêt de l'ENERGY STAR Score pour les prédictions ✓





Conclusion



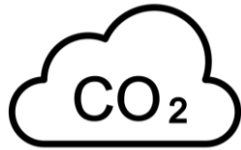
Le modèle **GradientBoosting** démontre les meilleures performances



R^2 0,88

ERROR

$\pm 0,41$ kBtu



R^2 0,90

ERROR








$\pm 0,42$ tCO2e



Conclusion



Features importantes pour les 2 prédictions :








- Surface d'usage principale  
- Proportion d'énergie provenant de l'électricité  
- ENERGY STAR Score non-imputé  
- Usage principal du bâtiment comme entrepôt 



Conclusion



Features importantes pour les 2 prédictions :

- Surface d'usage principale  
- Proportion d'énergie provenant de l'électricité  
- ENERGY STAR Score non-imputé  
- Usage principal du bâtiment comme entrepôt 





Conclusion



Limites :

- Il existe des algorithmes de ML qui n'ont pas été testé.
- Une seule méthode d'optimisation (GridSearchCV) a été testée.

OPENCLASSROOMS

Merci pour votre attention



CentraleSupélec

Pierrick BERTHE

Formation Expert en Data Science
Openclassrooms – CentraleSupélec

août 2023 → avril 2024