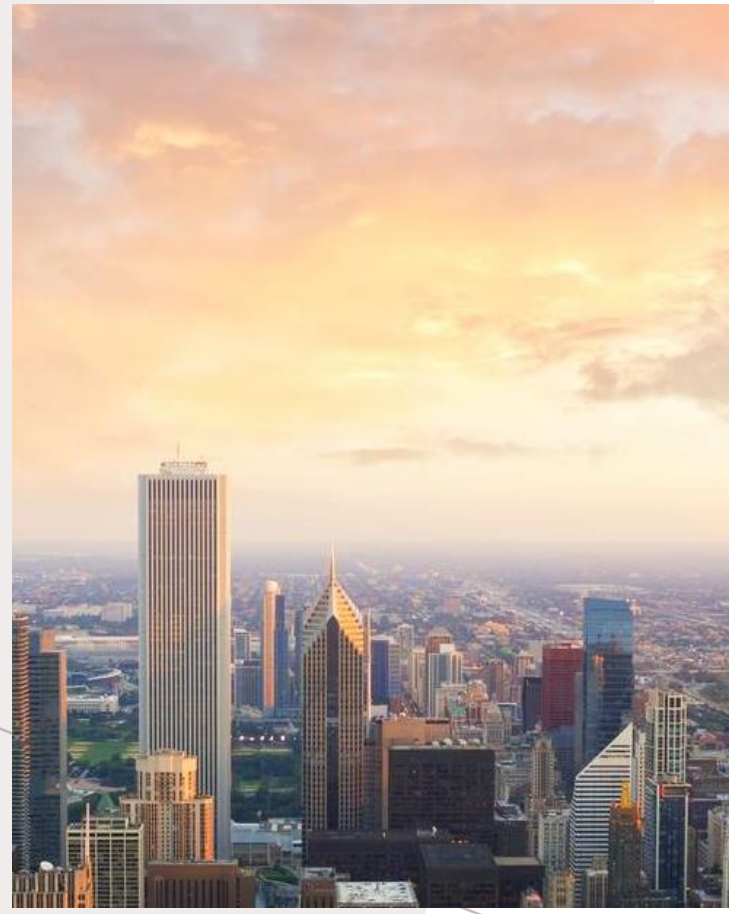
ANTICIPEZ LES BESOINS EN CONSOMMATION DES BÂTIMENTS

PROJET: #4

FORMATION: DATA SCIENTIST

MENTOR: MEDINA HADJEM

AGENDA



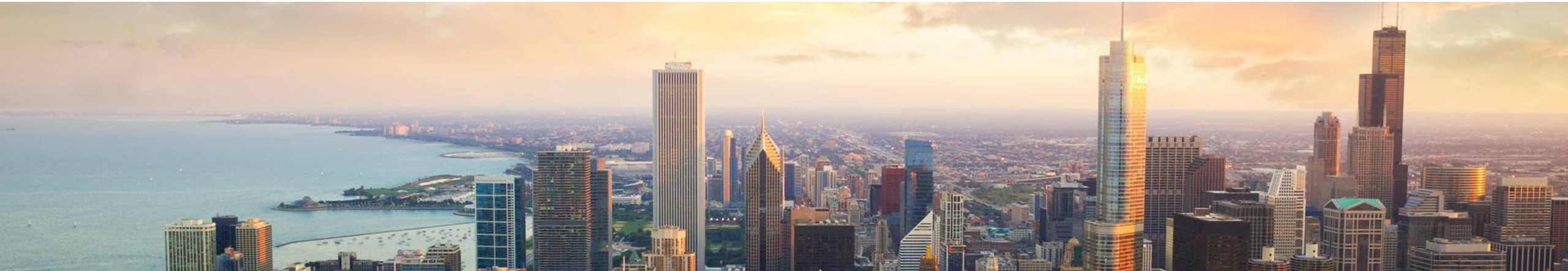
A) INTRODUCTION

B) PROBLÉMATIQUE

C) FEATURE ENGINEERING

E) APPROCHE DE MODÉLISATION ET RESULTATS

INTRODUCTION



LA VILLE DE SEATTLE

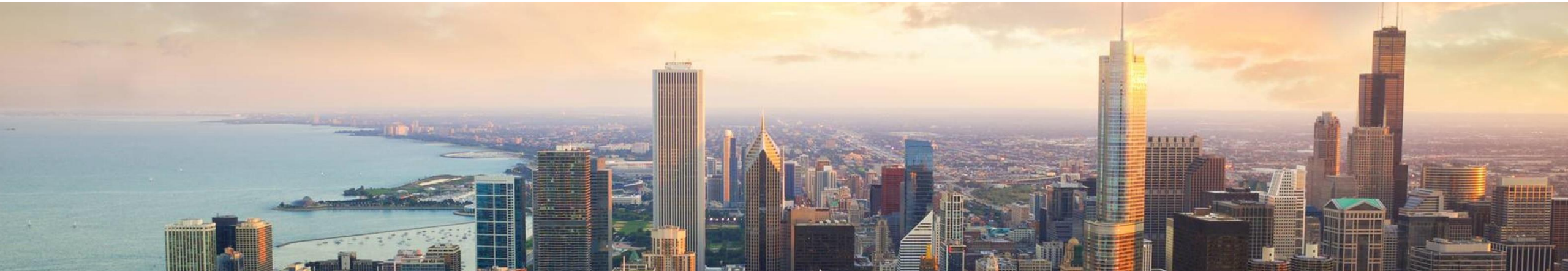
La ville de Seattle souhaite atteindre son objectif de neutralité carbone d'ici **2050**.

L'analyse et la prédiction sur des bâtiments **non résidentiels**, des consommations :

- **Consommation énergétique**
- **Emissions de CO_2**

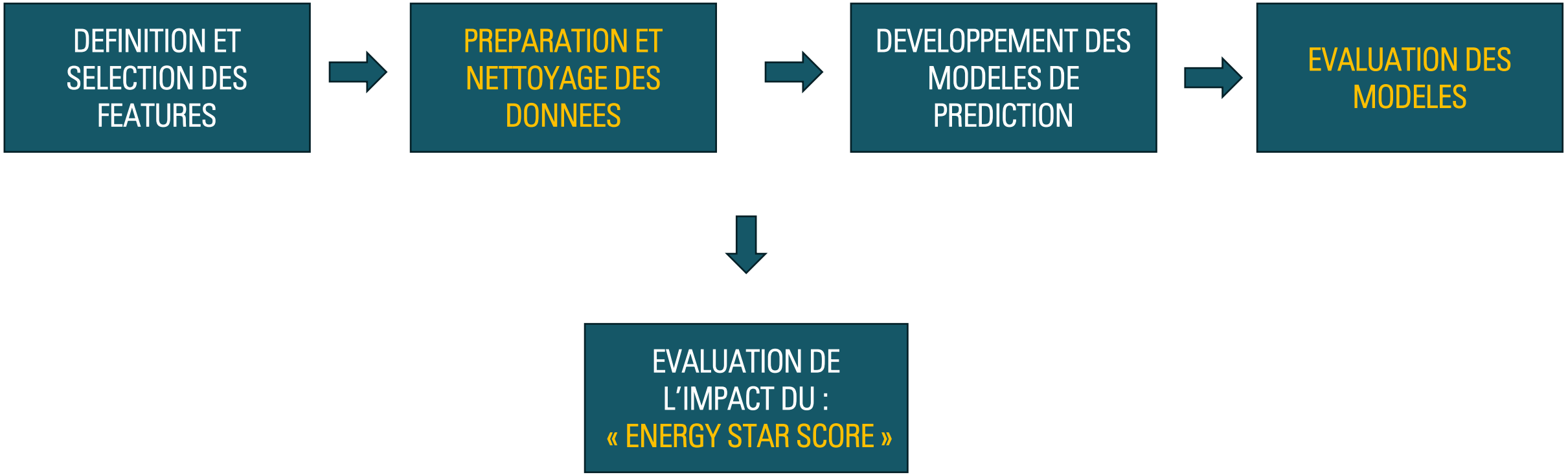
Aidera la ville de Seattle à mesurer son impact carbone et à se positionner concernant son taux de neutralité carbone d'ici 2050.

PROBLÉMATIQUE

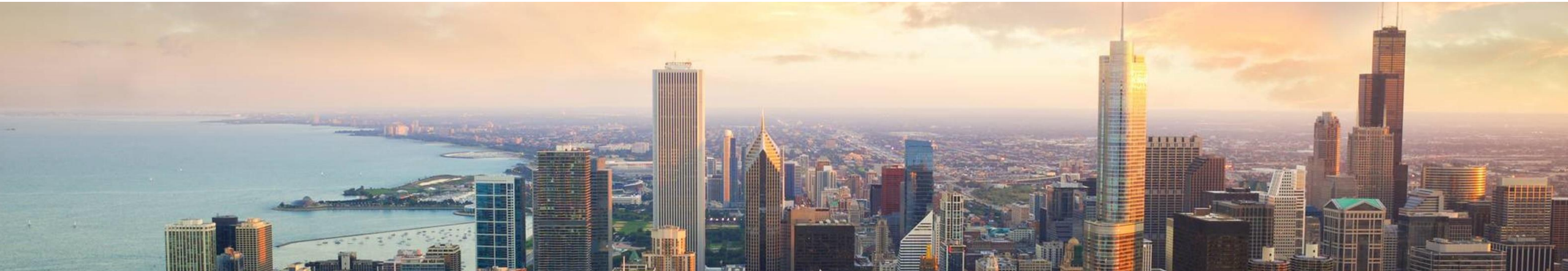


ATTEINDRE LA NEUTRALITÉ CARBONE D'ICI 2050

PLAN



FEATURE ENGINEERING



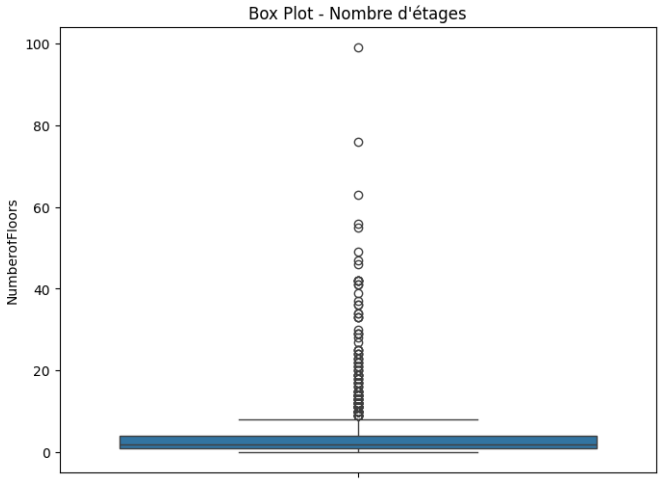
PRÉPARATION ET ANALYSE DES DONNÉES

CONTENU DE LA BASE DE **DONNEES**

- EMPLACEMENT DES BATIMENTS
- LES TYPES DE BATIMENTS
- CARACTERISTIQUES PHYSIQUES
- INDICATEURS ENERGETIQUES
- INDICATEURS DE CONFORMITE

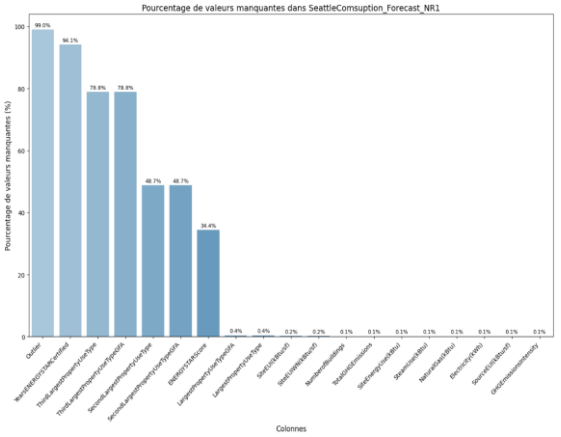
1

Box plot pour détecter et supprimer les valeurs abérrantes



2

Affichage et traitement des valeurs manquantes pour chaque feature



Le « **KNN Imputer** » remplace les valeurs manquantes en utilisant la moyenne des valeurs de caractéristiques similaires (voisines) dans les données.

Le « **Simple Imputer** » remplit les valeurs manquantes en utilisant la moyenne, la médiane, ou la valeur la plus fréquente de la colonne."

PREPARATION ET ANALYSE DES DONNEES

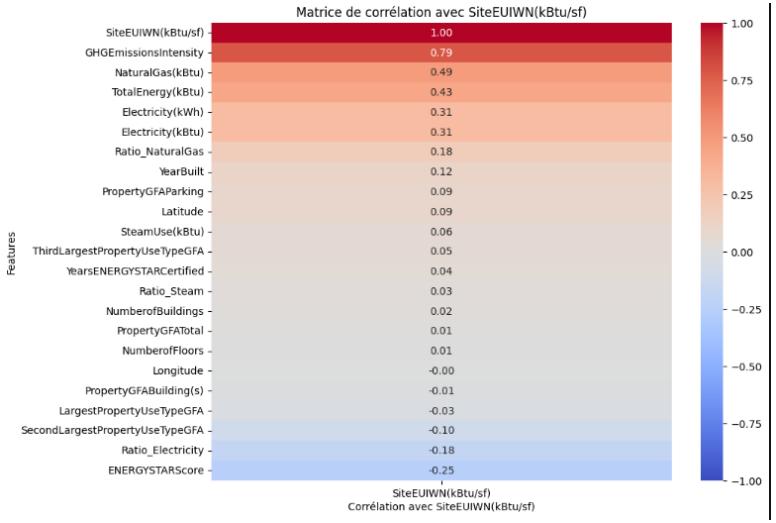
- Sélection des bâtiments non résidentiels

-

- Sélection des features qui vont nous permettre de prédire:

LA CONSOMMATION ENERGETIQUE
TARGET (Y) SiteEUIWN(kBtu/sf)

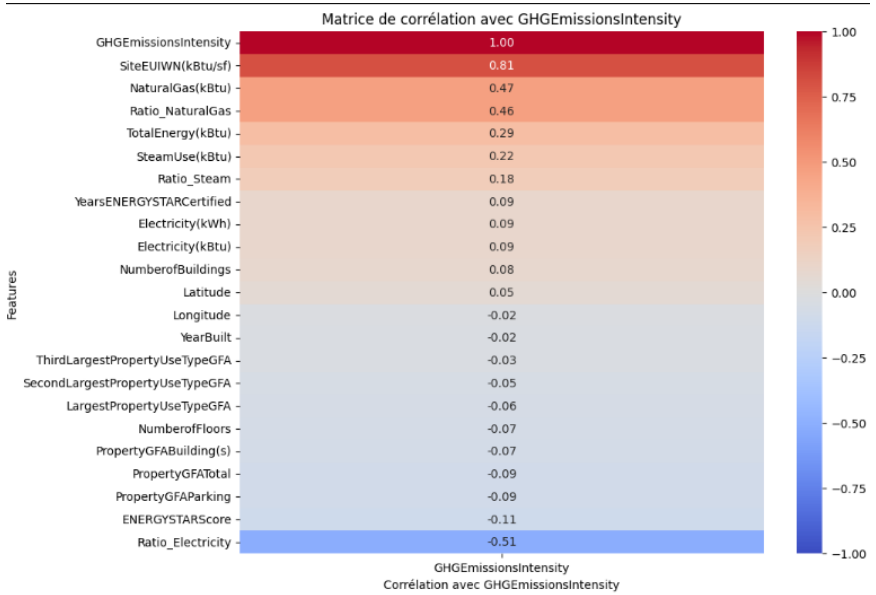
Intensité énergétique du site ajustée aux conditions météorologiques (en kBtu par pied carré).



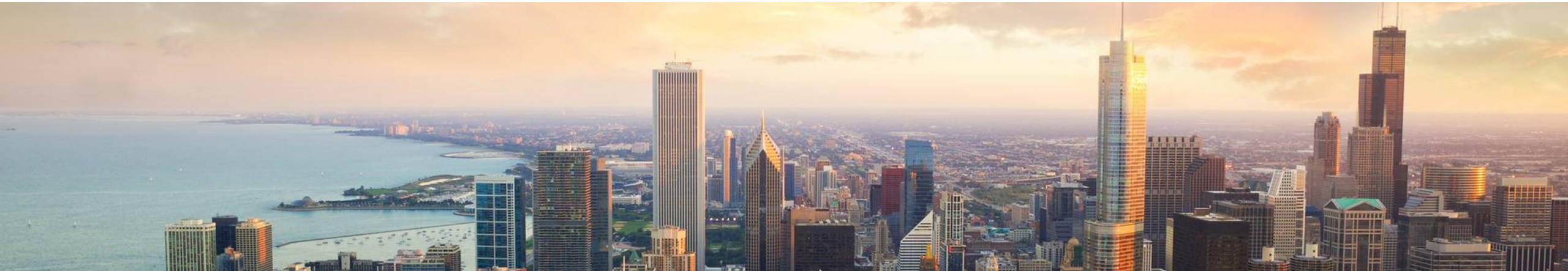
Les deux variables à prédire sont des variables normalisées (les moyennes de chacune des variables sont autour de 0)

LA CONSOMMATION CO2
TARGET (Y) GHEmissionsIntensity

Intensité des émissions de gaz à effet de serre (en kg de CO2 par pied carré).



APPROCHE DE MODELISATION ET RESULTATS



FEATURE ENGINEERING

LES NOUVELLES CARACTERISTIQUES

Nous intégrons des features dans notre dataframe afin d'augmenter la taille des variables X qui nous aideront à prédire nos targets.

NOM DU FEATURE	DEFINITION DU FEATURE	FORMULE DE CALCULE
Ratio_Electricity	Ce ratio représente la part de l'électricité dans la consommation énergétique totale. TotalEnergy(kBtu) = Electricity(kBtu) + NaturalGas(kBtu) + SteamUse(kBtu)	Electricity(kBtu) / TotalEnergy(kBtu)
Ratio_NaturalGas	Ce ratio représente la proportion de gaz naturel dans la consommation énergétique totale. TotalEnergy(kBtu) = Electricity(kBtu) + NaturalGas(kBtu) + SteamUse(kBtu)	NaturalGas(kBtu) / TotalEnergy(kBtu)
Ratio_Steam	Ce ratio indique la part d'utilisation de vapeur dans la consommation totale d'énergie. TotalEnergy(kBtu) = Electricity(kBtu) + NaturalGas(kBtu) + SteamUse(kBtu)	SteamUse(kBtu) / TotalEnergy(kBtu)

APPROCHE DE MODÉLISATION

Dans un premier temps nous avons testé **4 modèles pour prédire** la consommation ENERGETIQUE

1

REGRESSION LINEAIRE SIMPLE *MODELE DE BASE*

Modèle statistique simple qui établit une **relation linéaire** entre les variables **indépendantes** et la variable **dépendante**.

2

BAGGING BOSTRAP AGGREGATING

Technique d'ensemble qui **améliore la stabilité** et la précision des modèles en combinant **plusieurs prédictions issues d'échantillons de données différents.**

3

GRADIENT BOOSTING REGRESSOR

Modèle d'ensemble qui **optimise les performances en réduisant les erreurs résiduelles à chaque itération**, en construisant des **arbres successifs**.

4

STACKING REGRESSOR

Modèle d'ensemble qui combine **plusieurs modèles** de base et utilise un **modèle méta pour prédire**, optimisant ainsi la performance globale.

METRIQUES RMSE ET R2

RMSE (Root Mean Square Error)

Plus la valeur de la **RMSE** est **faible**, plus le modèle est précis.

-

Cela signifie que les prédictions sont proches des valeurs réelles.

RMSE = > 1

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

y : Les valeurs observées (réelles) du jeu de données.

$\hat{y}$: Les valeurs prédites par le modèle.

$n_{samples}$: Le nombre total d'échantillons dans le jeu de données.

y_i et $\hat{y}_i$: La valeur réelle et la valeur prédite pour le i-ième échantillon, respectivement.

Structure de la formule :

- La **somme des erreurs quadratiques** est calculée pour chaque échantillon en faisant la différence entre y_i et $\hat{y}_i$, puis en élevant cette différence au carré : $(y_i - \hat{y}_i)^2$.
- La **moyenne des erreurs quadratiques** est obtenue en divisant la somme par $n_{samples}$, soit le nombre d'échantillons.
- Enfin, on prend la **racine carrée** de cette moyenne pour obtenir la RMSE.

R2 (coefficient de détermination)

Une valeur R2 **positive** (proche de 1) indique:
un modèle qui explique bien la variabilité des données.

-

Une valeur **négative** indique que le modèle est pire qu'une simple moyenne des valeurs.

R2 = > 0

Interprétation de R^2 :

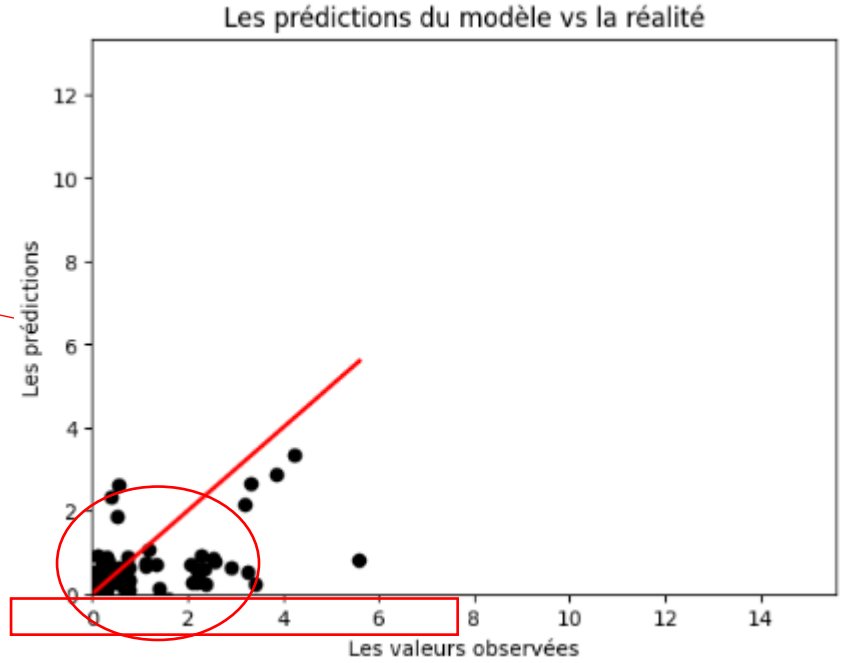
1. **Intervalle de valeurs** :
   - R^2 varie généralement entre **0** et **1**. Plus R^2 est proche de 1, mieux le modèle explique la variance des données observées.
   - Dans certains cas, notamment si le modèle est inadéquat, R^2 peut être négatif, indiquant que le modèle prédit moins bien que la moyenne des données.
2. **Signification des valeurs de R^2** :
   - $R^2 = 1$: Le modèle explique parfaitement toute la variance des données observées (valeurs prédites identiques aux valeurs observées).
   - $R^2 = 0$: Le modèle ne parvient pas à expliquer la variance des données. Dans ce cas, il est aussi performant qu'un modèle constant qui prédirait la moyenne des valeurs observées.
   - R^2 négatif : Indique que le modèle est pire que la moyenne, c'est-à-dire qu'il produit des prédictions plus éloignées des valeurs observées que la moyenne elle-même.
3. **Exemple d'interprétation** :
   - Si $R^2 = 0.85$, cela signifie que **85 % de la variance des données observées** est expliquée par le modèle, et 15 % de la variance reste inexpliquée par le modèle.

LA RÉGRESSION LINÉAIRE SIMPLE

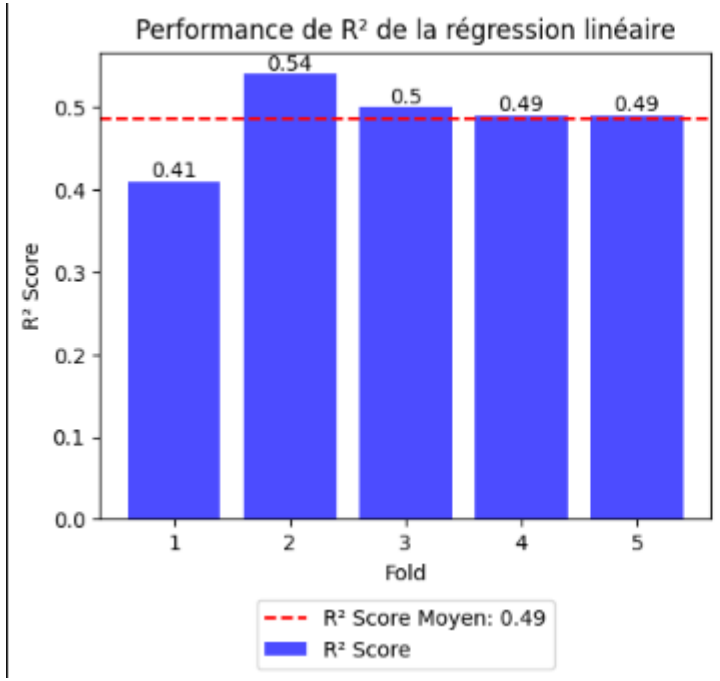
RMSE : 0.71 et R2 0.41

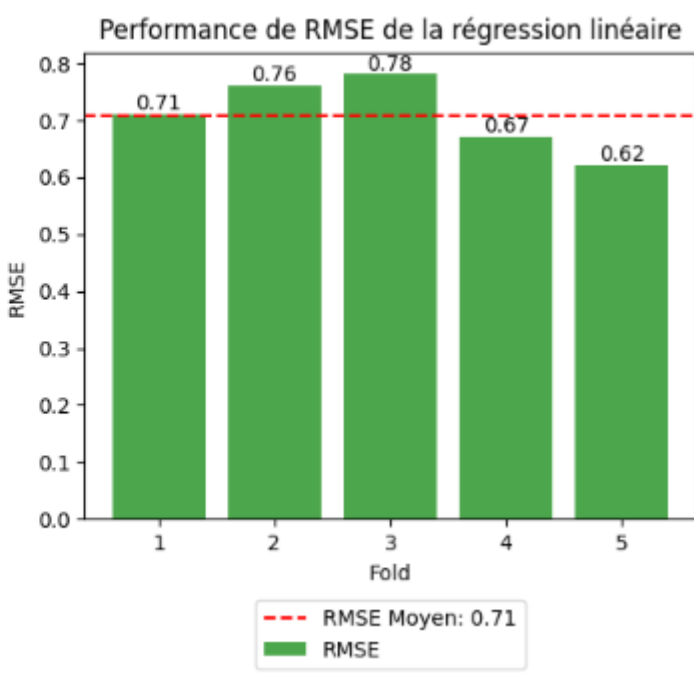
Le graphique montre **la comparaison** entre les **valeurs observées (axe des abscisses)** et les **prédictions du modèle (axe des ordonnées)**.

- **La ligne rouge** représente la **relation parfaite**.
- La majorité des points étant éloignés de la ligne rouge, le modèle sous-estime certaines valeurs élevées.
- Les valeurs se concentrent principalement entre **0** et **6** avec quelques valeurs plus hautes, indiquant une possible asymétrie



LA RÉGRESSION LINÉAIRE SIMPLE





RMSE : 0.71 et R2 0.41

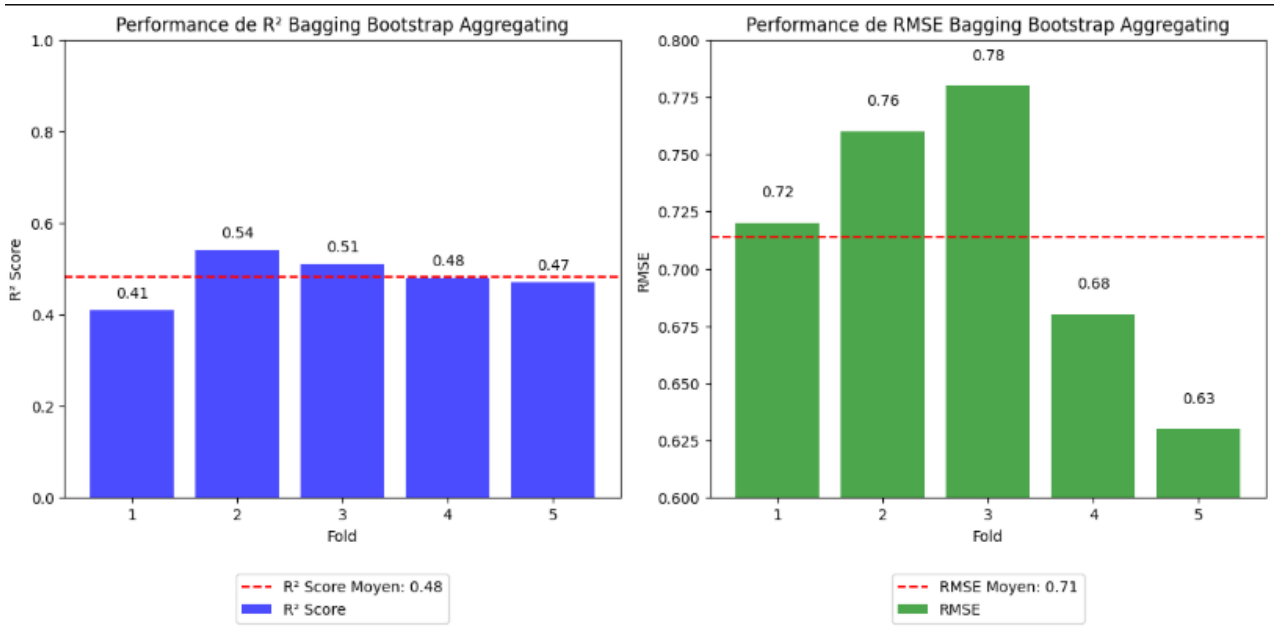
Le **R² score** varie entre **0.41** et **0.54** avec une moyenne de **0.49** indiquant une **capacité explicative** du modèle peu performante

Le **RMSE** reste relativement haut, autour de **0.71**, ce qui suggère une **précision modeste** dans les prédictions du modèle, avec des erreurs de prédiction.

LE BAGGING BOOSTRAP AGGREGATING

BAGGING BOSSTRAP AGGREGATING

Le modèle Bagging Bootstrap Aggregating ne présente pas d'amélioration significative par rapport à la régression linéaire simple.



Le R² moyen est de 0.48

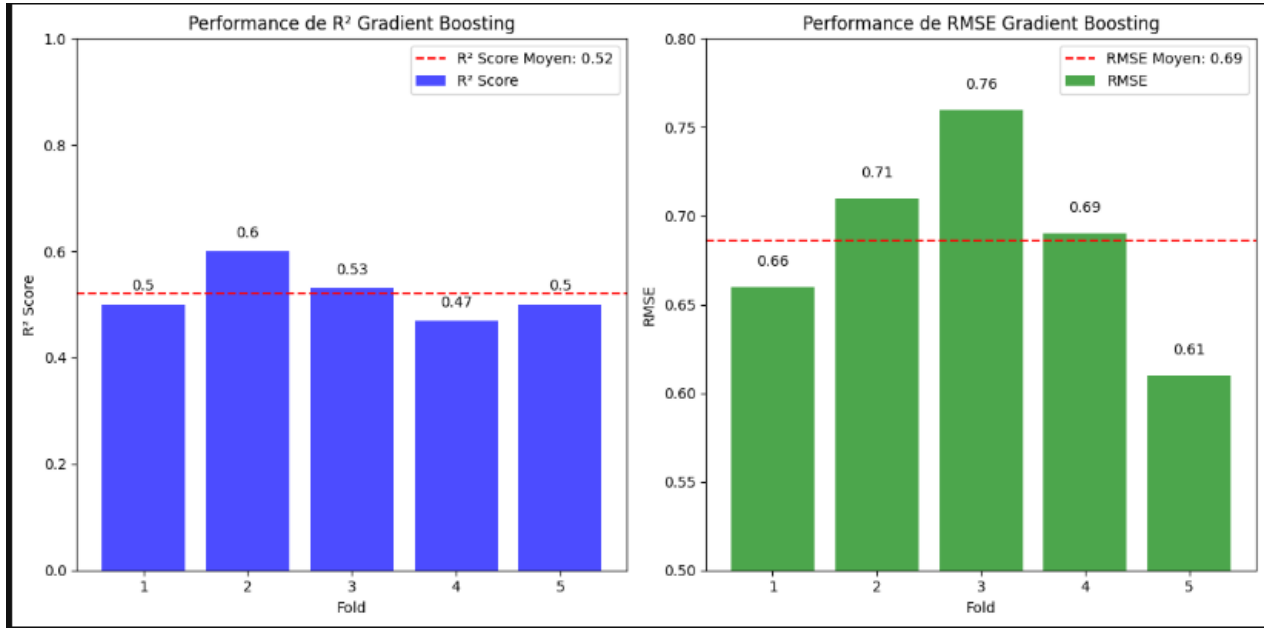
Le RMSE moyen est de 0.71, montrant une erreur comparable à celle du modèle précédent.

RMSE : 0.72 et R2 0.40

LE GRADIENT BOOSTING REGRESSOR

GRADIENT BOOSTING REGRESSOR

Le modèle Gradient Boosting Regressor surpasse le modèle Bagging Bootstrap Aggregating en termes de précision et de fiabilité prédictive, grâce à un meilleur score R² et un RMSE plus faible.



RMSE : 0.66 et R2 0.50

R² moyen: 0.52 / indique une meilleure capacité explicative du modèle.

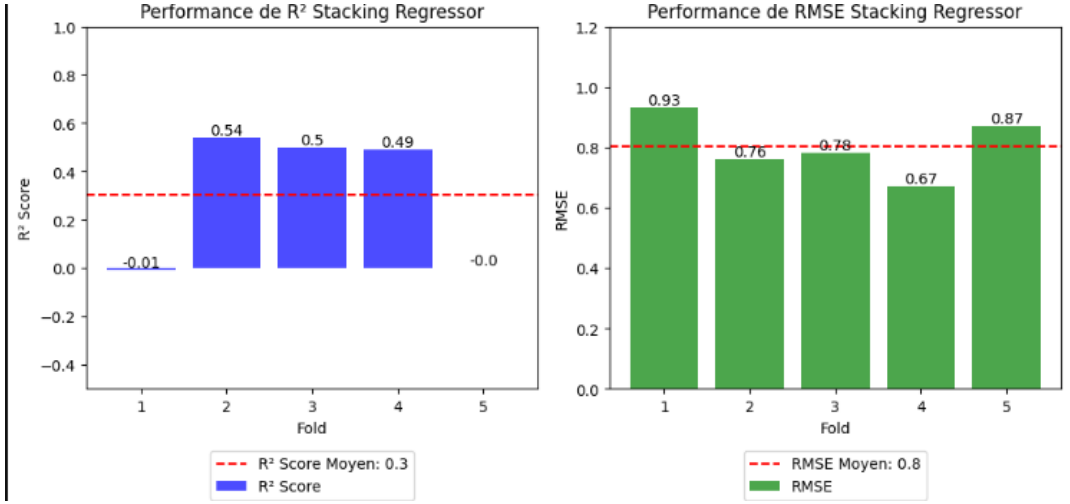
Il capture en partie la variance des données et offre une prédiction fiable.

RMSE moyen: 0.66/ Erreur de prédiction qui peut être améliorée

LE STACKING REGRESSOR

STACKING REGRESSOR

Le modèle Stacking Regressor reste inférieur aux modèles Bagging et Gradient Boosting. Avec un RMSE moyen de 0.80, l'erreur de prédiction élevée limite sa fiabilité globale.



Un R^2 moyen (0.30)

le RMSE moyen (0.80)

RMSE : 0.71 et R2 0.41

- Régression linéaire : RMSE = **0.71** R^2 = **0.41** - (précision limitée malgré une capacité explicative modérée.
- Bagging Bootstrap Aggregating : RMSE = **0.72** R^2 = **0.40** - (capacité explicative raisonnable).
- Gradient Boosting Regressor : RMSE = **0.66** R^2 = **0.50** - (bonne capacité explicative).
- Stacking Regressor : RMSE = **0.71** R^2 = **0.41** - (précision et capacité explicative faibles).

Le modèle de prédiction Gradient Boosting Regressor est le meilleur modèle globalement, car il offre la meilleure précision **(RMSE le plus faible)** et la meilleure capacité explicative (R^2 élevé), surpassant les autres modèles.

FEATURE ENGINEERING

Nous intégrons de nouvelles features dans notre dataframe et relançons le modèle le plus performant

LES NOUVELLES CARACTERISTIQUES

NOM DU FEATURE	DEFINITION DU FEATURE	FORMULE DE CALCULE
BUILDING AGE	Age actuel du bâtiment. Cela reflète l'efficacité énergétique et l'état des infrastructures, influençant la consommation d'énergie.	Année actuelle (2024) - Building Age
ParkingGFA_Ratio	Superficie totale de stationnement divisée par la superficie brute de plancher (GFA) totale. Il indique l'impact indirect de la surface de stationnement sur la consommation d'énergie.	PropertyGFAParking / PropertyGFATotal
BuildingGFA_Ratio	La superficie de la partie principale du bâtiment divisée par la superficie totale brute de plancher (GFA). Cela aide à distinguer les usages intérieurs et leurs effets énergétiques.	PropertyGFABuilding(s) / PropertyGFATotal
GFA_Per_Floor	La surface brute de plancher par étage est obtenue en divisant la superficie brute totale par le nombre d'étages du bâtiment. Cela indique la densité spatiale, influençant l'efficacité énergétique.	PropertyGFATotal» / NumberofFloors

PERFORMANCE: GRADIENT BOOSTING REGRESSOR

IMPACT POSITIF DU FEATURE ENGINEERING

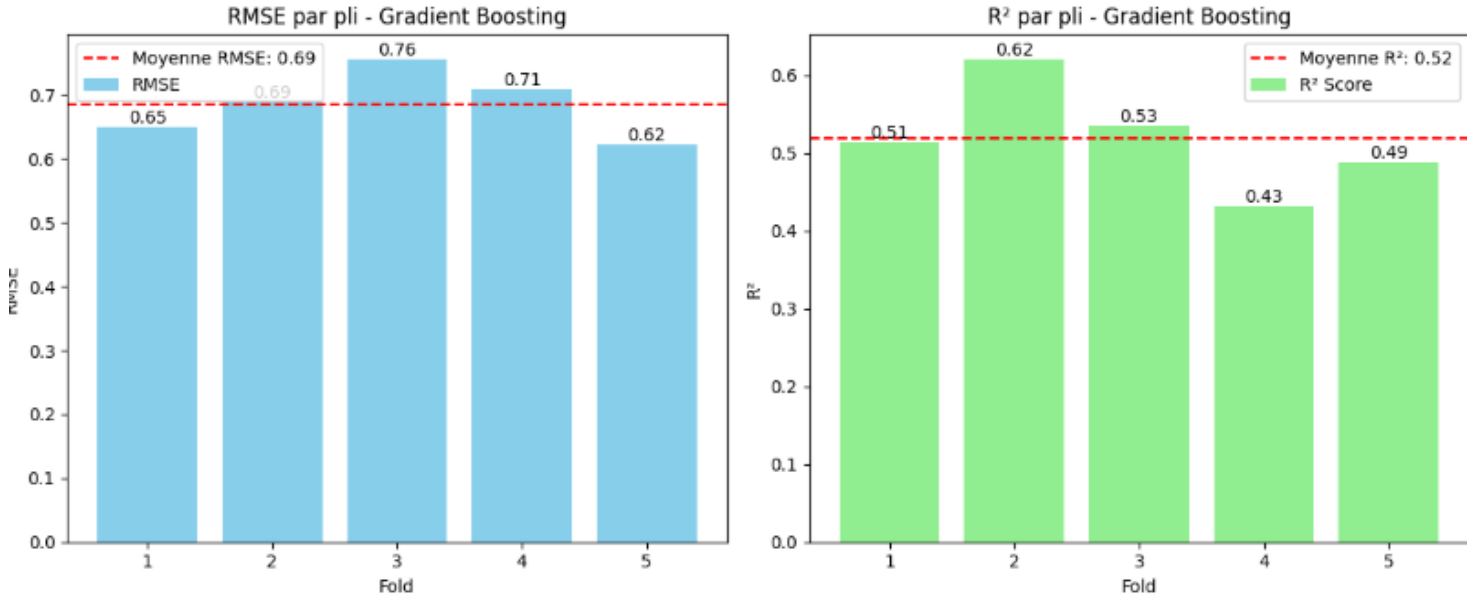
- **Optimisation des hyperparamètres**

Utilisation de GridSearchCV pour ajuster les paramètres du Gradient Boosting, améliorant la précision.

- **Validation croisée à 5 plis** :

Évaluation robuste du modèle, avec **RMSE faible** et **R² élevé**, montrant une meilleure prédiction et explication des variations.

RMSE moyen : 0.69 et R2 moyen: 0.52

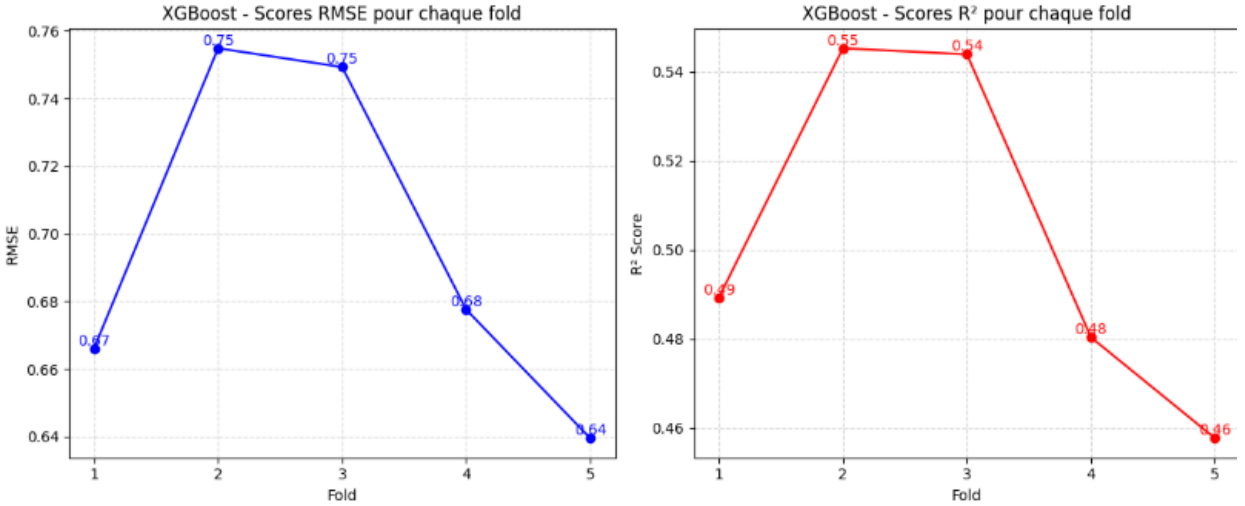


Le modèle **Gradient Boosting Regressor** montre une amélioration significative avec un **RMSE moyen de 0.69** et un **R² moyen de 0.52**, indiquant une meilleure capacité explicative.

Cette performance est atteinte grâce à l'intégration de nouvelles caractéristiques et à l'optimisation des hyperparamètres, qui ont permis au modèle de mieux capturer les variations des données.

XG BOOST - RANDOM FOREST

RMSE moyen : 0.70 et R2 moyen: 0.50

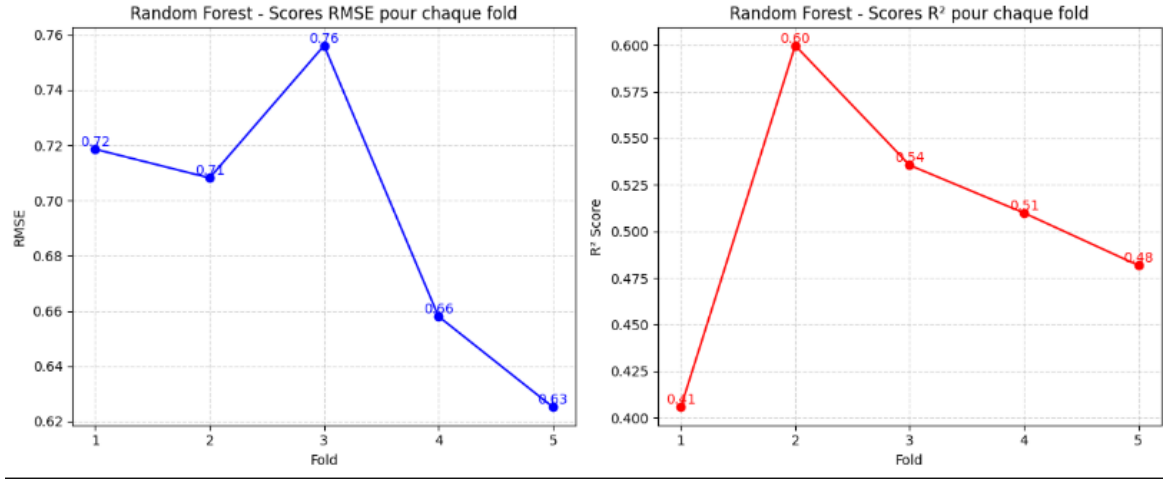


Le RMSE moyen de 0.70 et le R2 moyen de 0.50 indiquent que le modèle a une capacité prédictive modérée.

/

La diminution du RMSE à travers les folds démontre une amélioration en précision. La stabilité R2 montre également une bonne cohérence explicative à travers les différentes partitions de validation croisée.

RMSE moyen : 0.69 et R2 moyen: 0.51

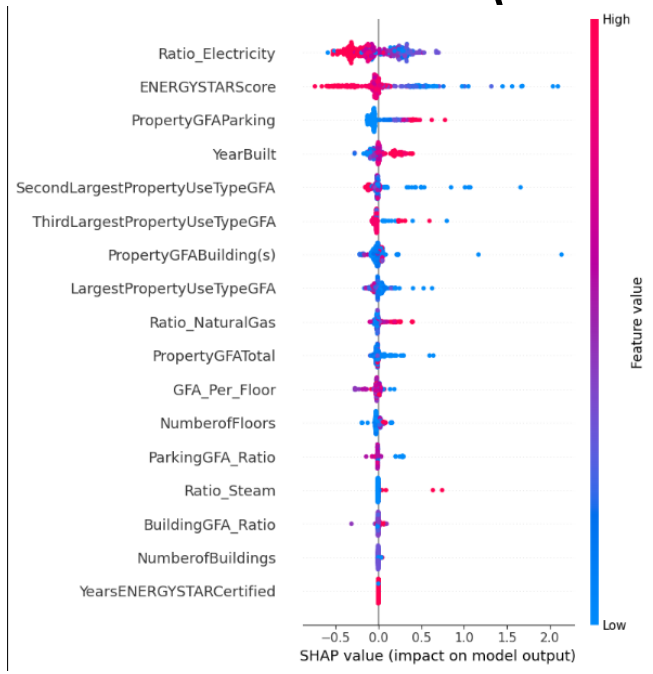


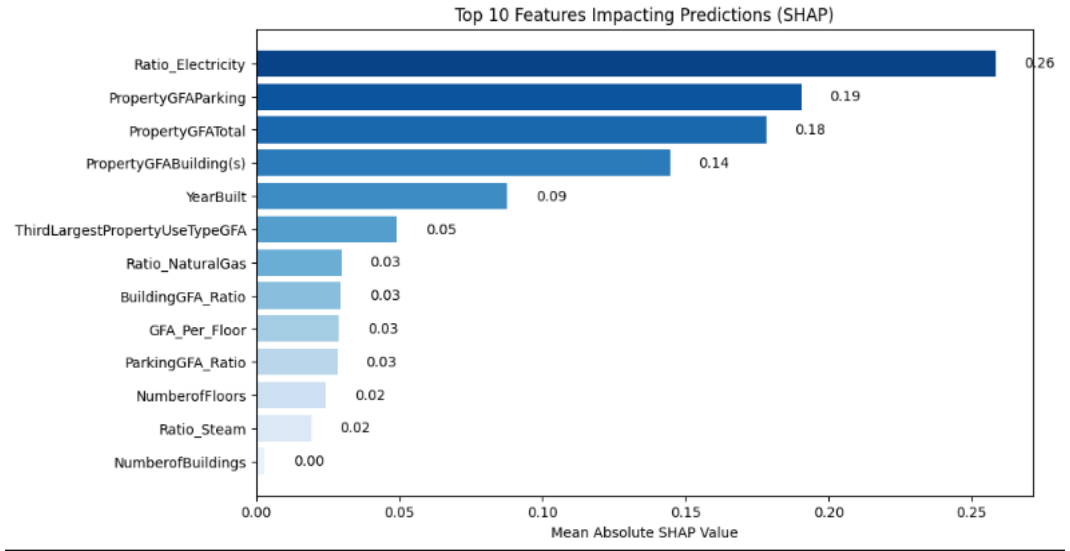
Le modèle Random Forest affiche un RMSE moyen de 0.69 et un R2 moyen de 0.51, montrant des performances comparables à celles de XGBoost.

/

La stabilité des scores RMSE et R2 à travers les folds indique que le modèle est robuste et bénéficie du feature engineering appliqué, tout en offrant une explication de la variance des données acceptable.

FEATURE IMPORTANCE
(SHAP) ADDITIVE EXPLANATION GLOBALE





Valeurs de Shapley :

SHAP utilise les valeurs de Shapley pour attribuer à chaque caractéristique un poids quantitatif, mesurant ainsi son impact sur la prédiction finale..

Interprétabilité locale et globale :

Locale : SHAP explique une prédiction en visualisant la contribution de chaque caractéristique.

Globale : En analysant toutes les prédictions, SHAP montre l'impact moyen de chaque caractéristique.

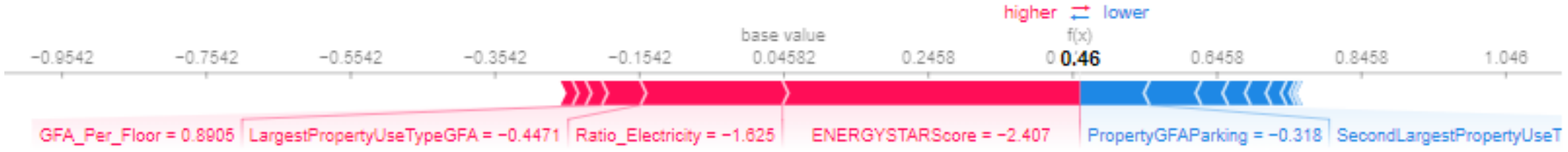
Additivité et cohérence :

SHAP assure que les contributions cumulées des caractéristiques donnent la prédiction finale, garantissant une évaluation équitable.

Visualisations intuitives :

SHAP propose divers graphiques (barres, nuages de points, etc.) pour faciliter la compréhension de l'impact des caractéristiques.

FEATURE IMPORTANCE
(SHAP) ADDITIVE EXPLANATION LOCALE



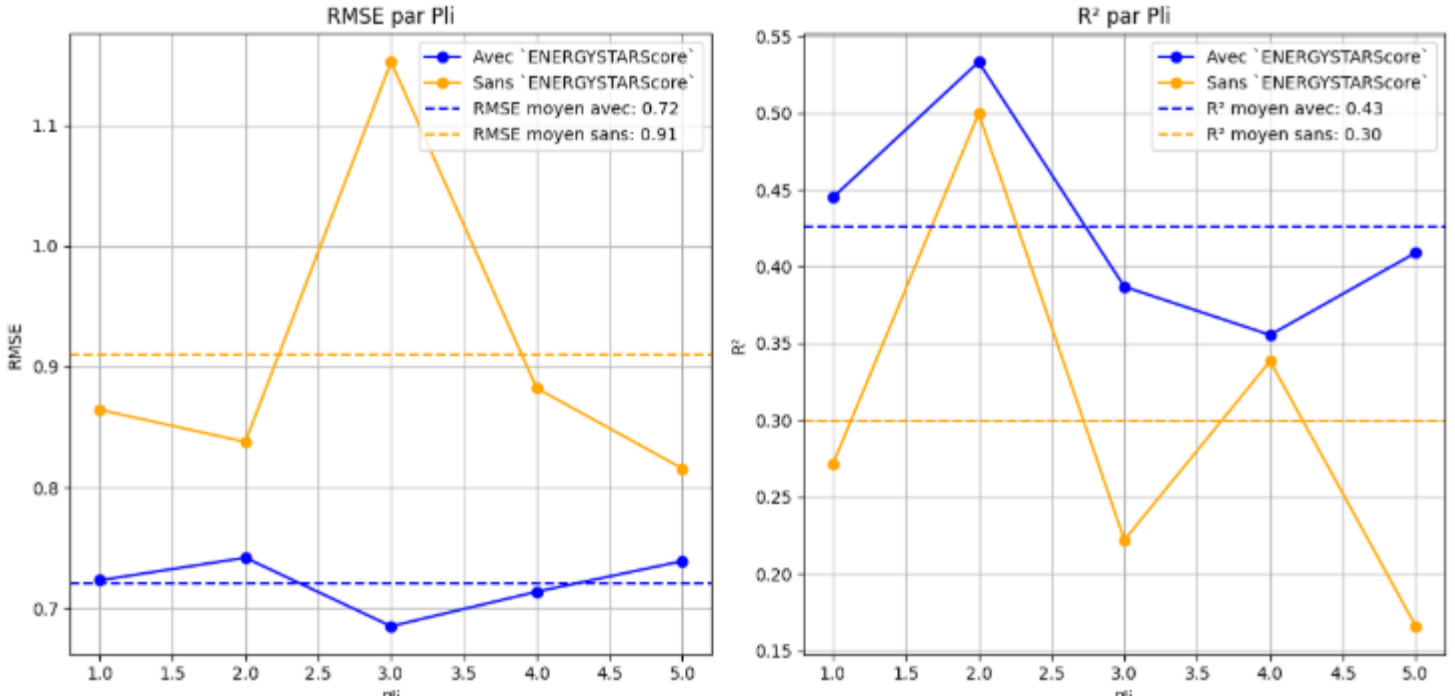
Les caractéristiques en **rouge** (LargestPropertyUseTypeGFA, Ratio_Electricity poussent la prédiction vers une valeur plus élevée.

Leur contribution **positive** signifie qu'elles augmentent la prédiction de cette observation par rapport à la valeur de base.

Les caractéristiques en **bleu** (par exemple, Property GFAParking, SecondLargestPropertyUseType ont un effet inverse, abaissant la prédiction finale.

Prédiction finale (0.46) : La valeur finale est la somme de la valeur de base et des contributions de chaque caractéristique. La prédiction du modèle **GRADIENT BOOSTING REGRESSOR** pour cet échantillon est de 0.46

ENERGYSTARSCORE IMPACT DU FEATURE MODELE: GRADIENT BOOSTING REGRESSOR

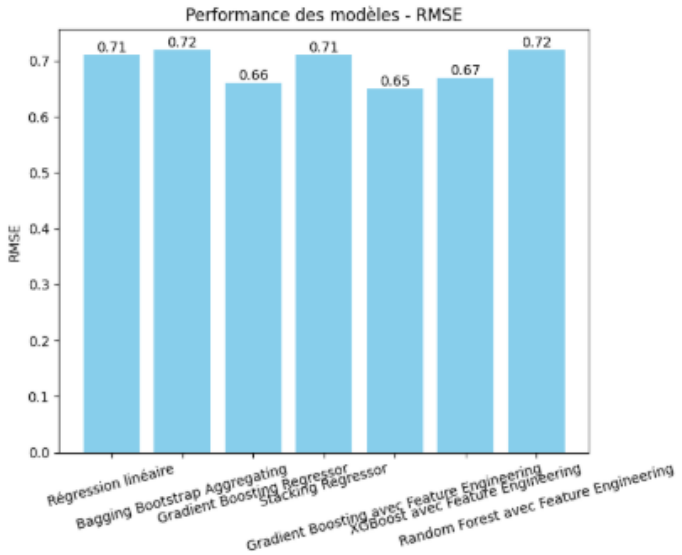


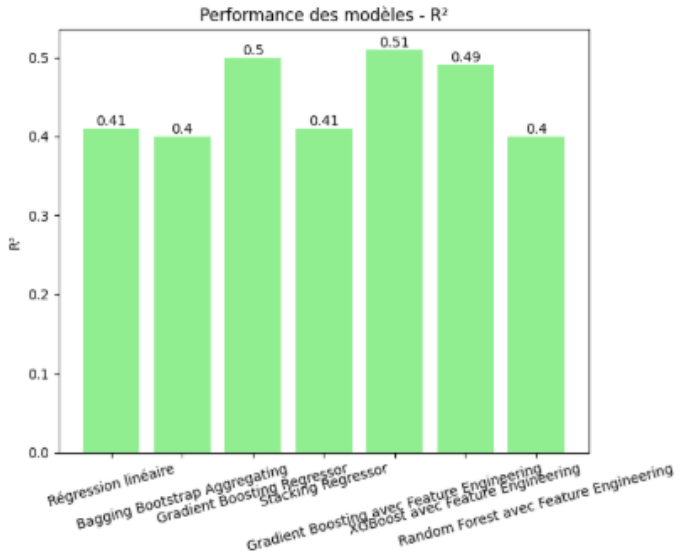
- RMSE par pli : La performance du modèle avec et sans ENERGYSTARScore est relativement different, l'ajout de ENERGYSTARScore apporte une légère amélioration de la précision du modèle, bien que cet impact soit modeste.

- R^2 par pli : Le score R^2 reste également très proche entre les deux configurations. Cela montre que le pouvoir explicatif du modèle diffère lorsque l'on inclut l'ENERGYSTARScore.

CONCLUSION

Modèles prédictifs de l'Energie





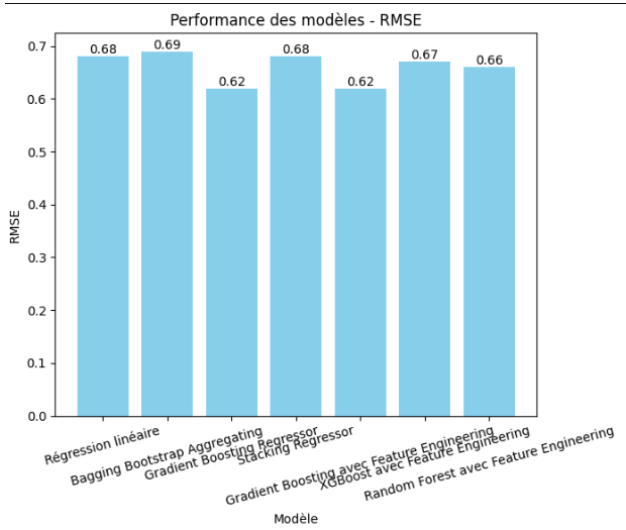
Le modèle prédictif **Gradient Boosting Regressor** avec le feature engineering s'est révélé être le modèle le plus performant, surpassant les autres modèles en termes de précision et de capacité explicative.

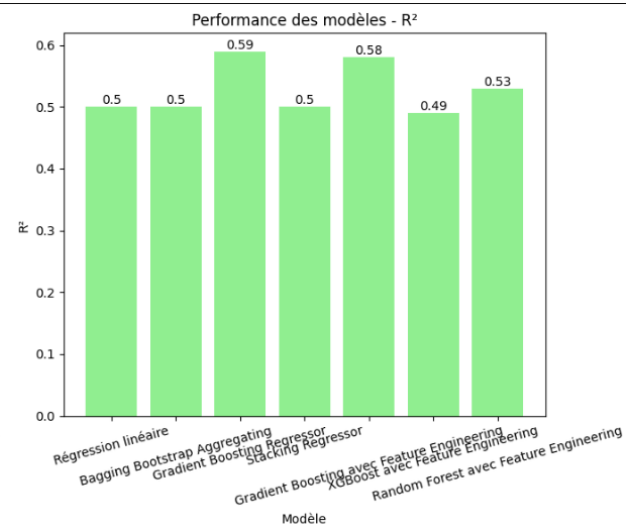
La combinaison de transformations avancées, d'optimisation des hyperparamètres et d'un choix judicieux des caractéristiques a permis d'obtenir un modèle robuste et fiable pour prédire la consommation énergétique.

Modèles prédictifs de CO2

Les résultats montrent que les modèles avancés avec ingénierie des caractéristiques, comme le **Gradient Boosting** et **XGBoost**, offrent une précision supérieure pour la prédiction des émissions de CO_2.

Le **Gradient Boosting** avec feature engineering atteint un RMSE de 0,62 et un R^2 de 0,58, surpassant la Régression linéaire simple (RMSE de 0,68 et R^2 de 0,50). Ces résultats soulignent l'impact positif des techniques avancées et du feature engineering sur la précision des prédictions.





MERCI DE VOTRE ATTENTION