



SEATTLE

Anticiper les besoins en
consommation des batiments

AVRIL 2024
LOKMAN AALIOUI

Feuille de route



INTRODUCTION



PRÉPARATION DES DONNÉES



ANALYSE PRÉLIMINAIRE



MODÉLISATION



CONCLUSION





Contexte

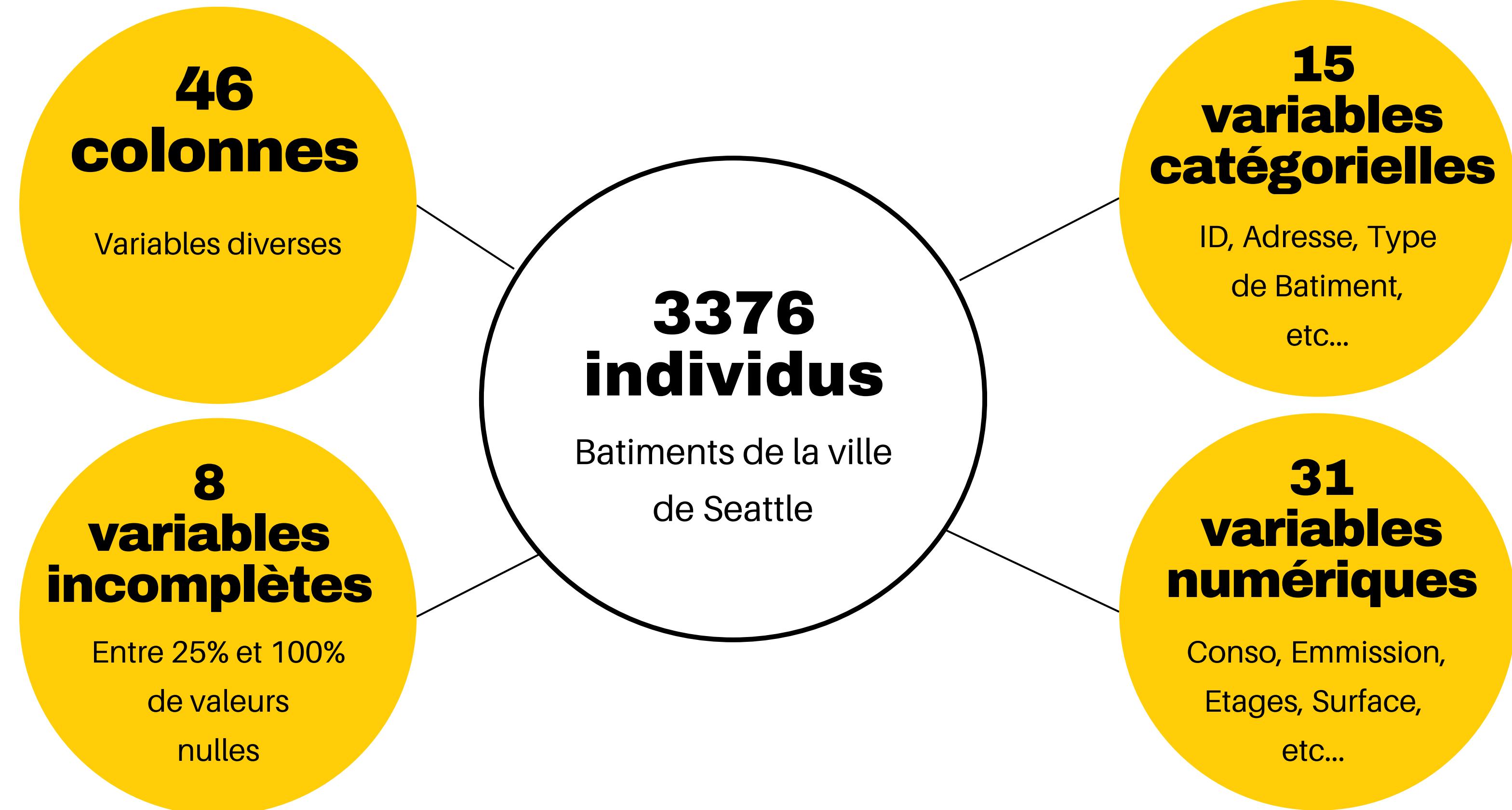
- Mesurer la performance énergétique des bâtiments est cruciale dans la lutte contre le changement climatique
- Seattle s'engage vers la neutralité carbone d'ici 2050, évaluer cette performance est impératif pour réduire les émissions de CO2 et préserver l'environnement
- Dans ce contexte, des relevés très coûteux ont été réalisés dans un échantillon de bâtiments



Mission

- Prédire les émissions de CO₂ et la consommation énergétique pour les bâtiments non relevés
- Questionner la pertinence de l'indicateur ENERGY STAR Score

Les données





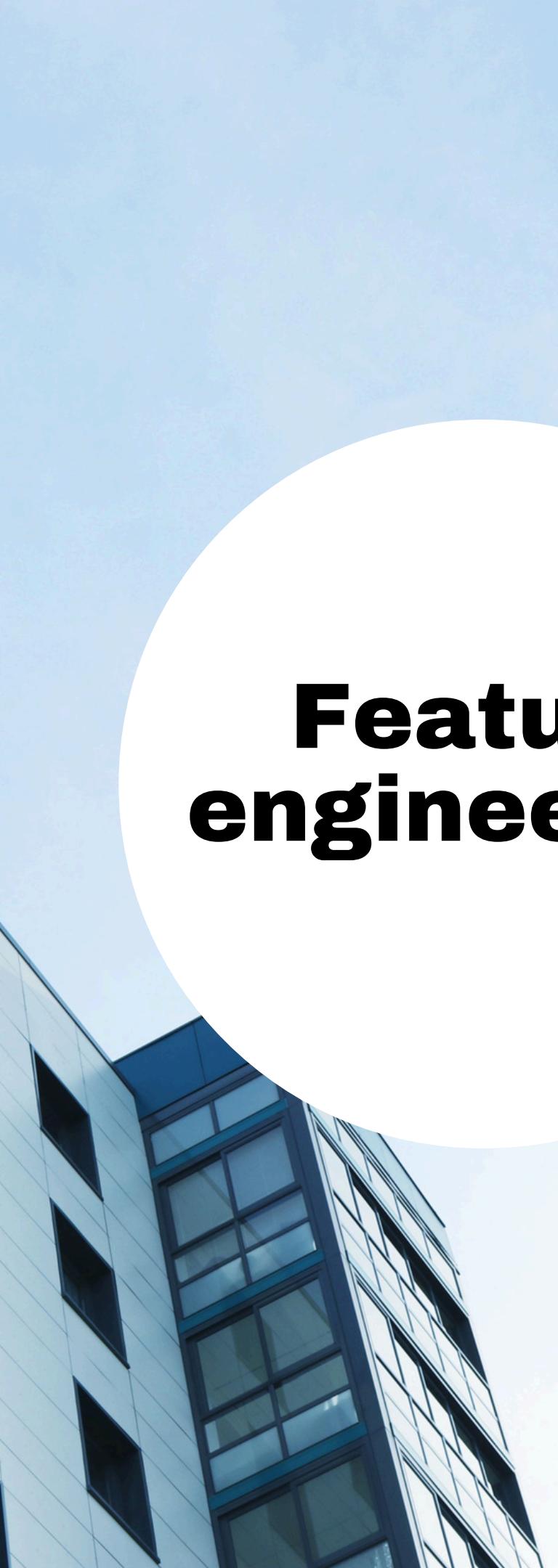
Traitement des données

Approche qualité

- Correction des types
- Suppression des colonnes incomplètes à plus de 50% (-6)

Approche métier

- Suppression des bâtiments résidentiels (-1708)
- Suppression des individus dont les valeurs énergétiques sont nulles (-121)
- Suppression des colonnes inutiles City, State, DataYear et Longitude et Latitude (-5)
- Suppression des individus marqués comme Outlier (-17) et des individus marqués comme Non-Compliant (-15)
- Suppression des batiments dont le nombre d'étage est >76 (-1)
- Suppression des batiments dont les émissions ou la consommation est négative (-1)



Feature engineering

Création de variables

- Ancienneté du bâtiment
- Ratio de consommation de gaz
- Ratio de consommation d'électricité
- Pourcentage de parking par rapport à la surface total

Regroupement des valeurs

- Regroupement des valeurs de LargestPropertyUseType en 13 catégories



Choix des variables

Variables qualitatives

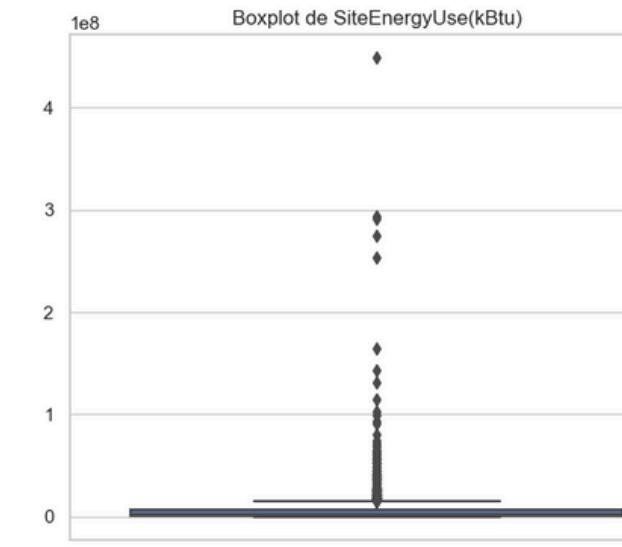
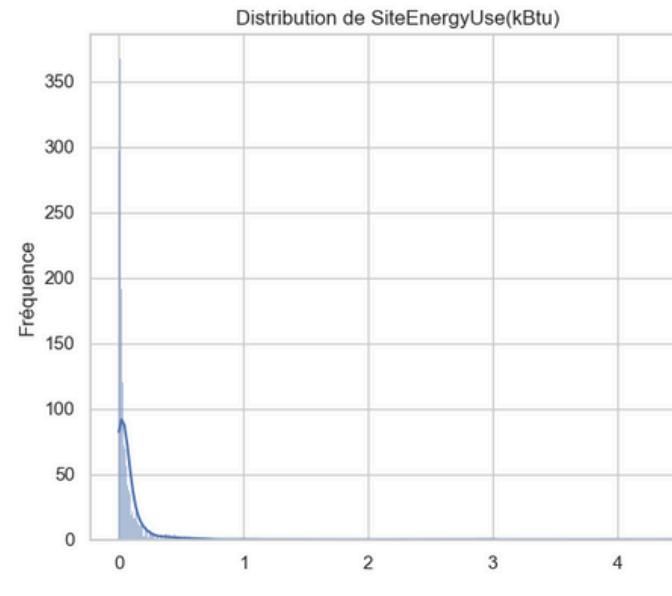
- LargestPropertyUseType
- Neighborhood

Variables quantitatives

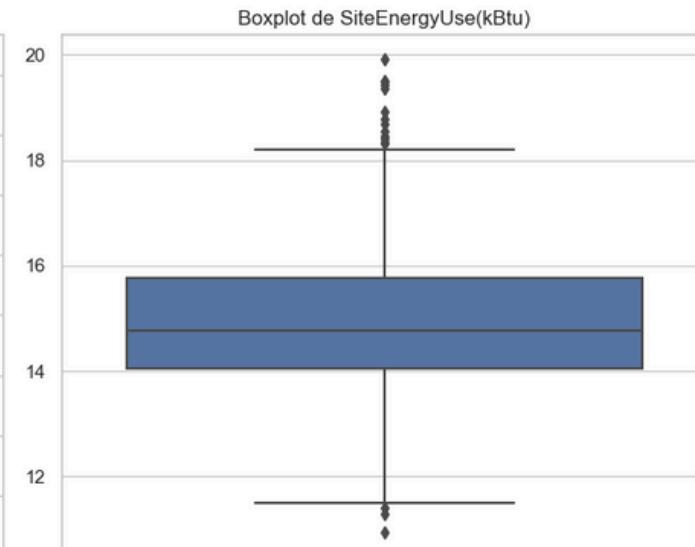
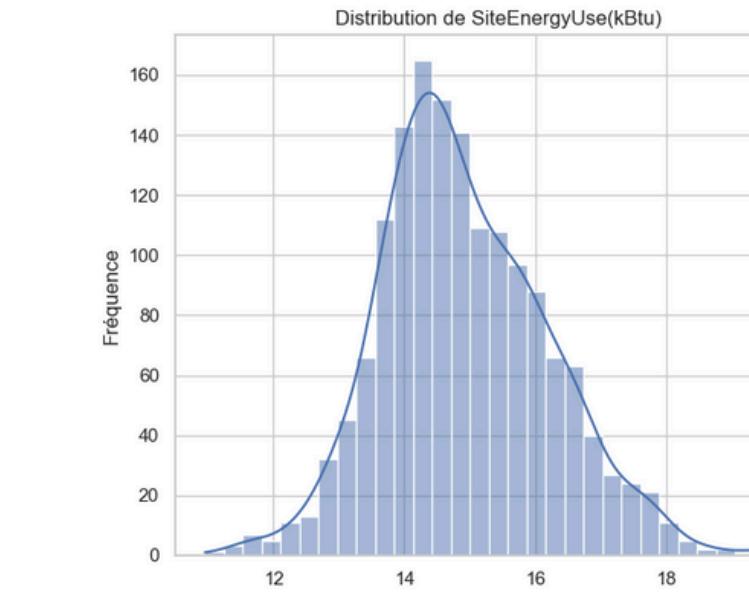
- NumberofFloors
- NumberofBuildings
- PropertyGFATotal
- BuildingAge
- GasUseRatio
- ElectricityUseRatio
- ParkingGFARatio

Observation des variables cibles

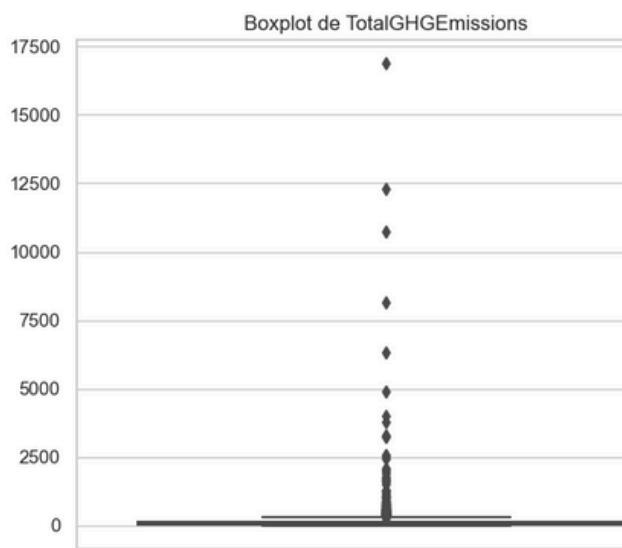
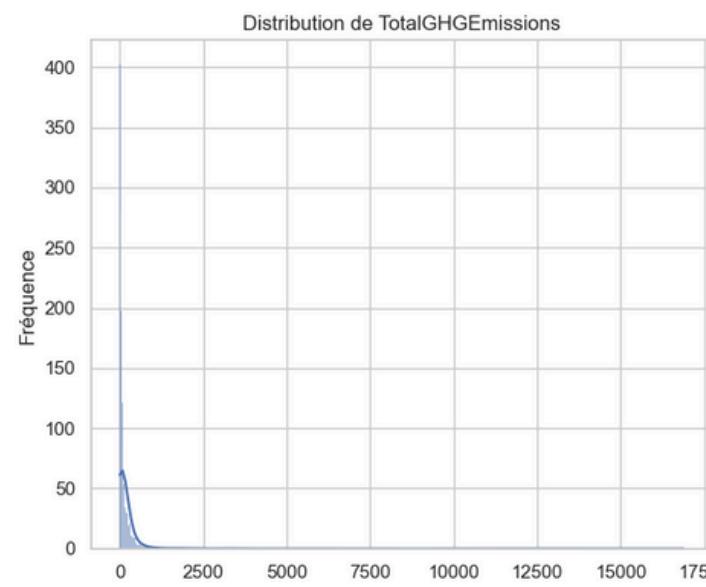
- 'SiteEnergyUse(kBtu)'



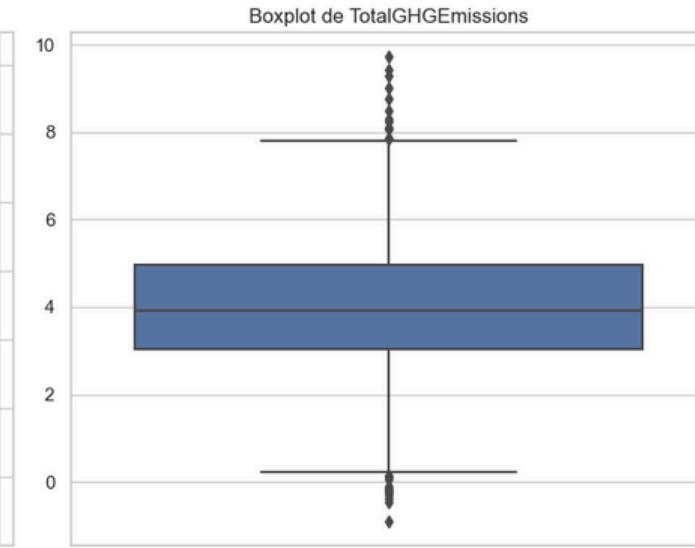
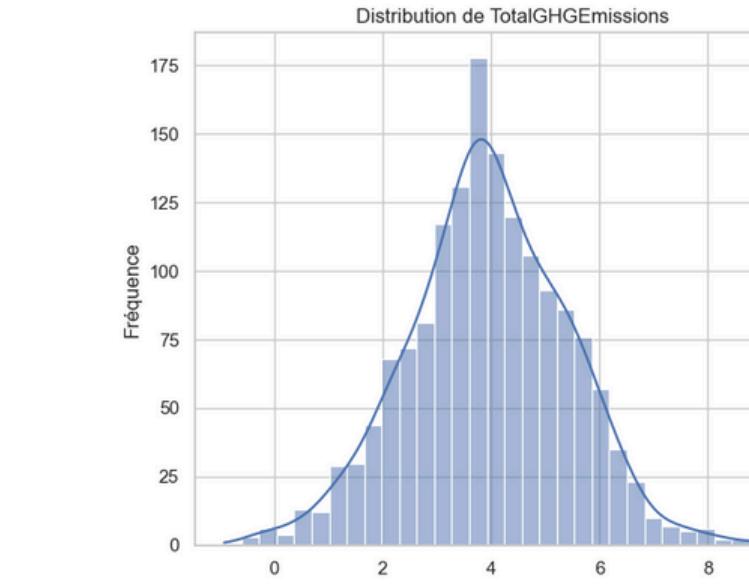
- Avec transformation logarithmique



- 'TotalGHGEmissions'

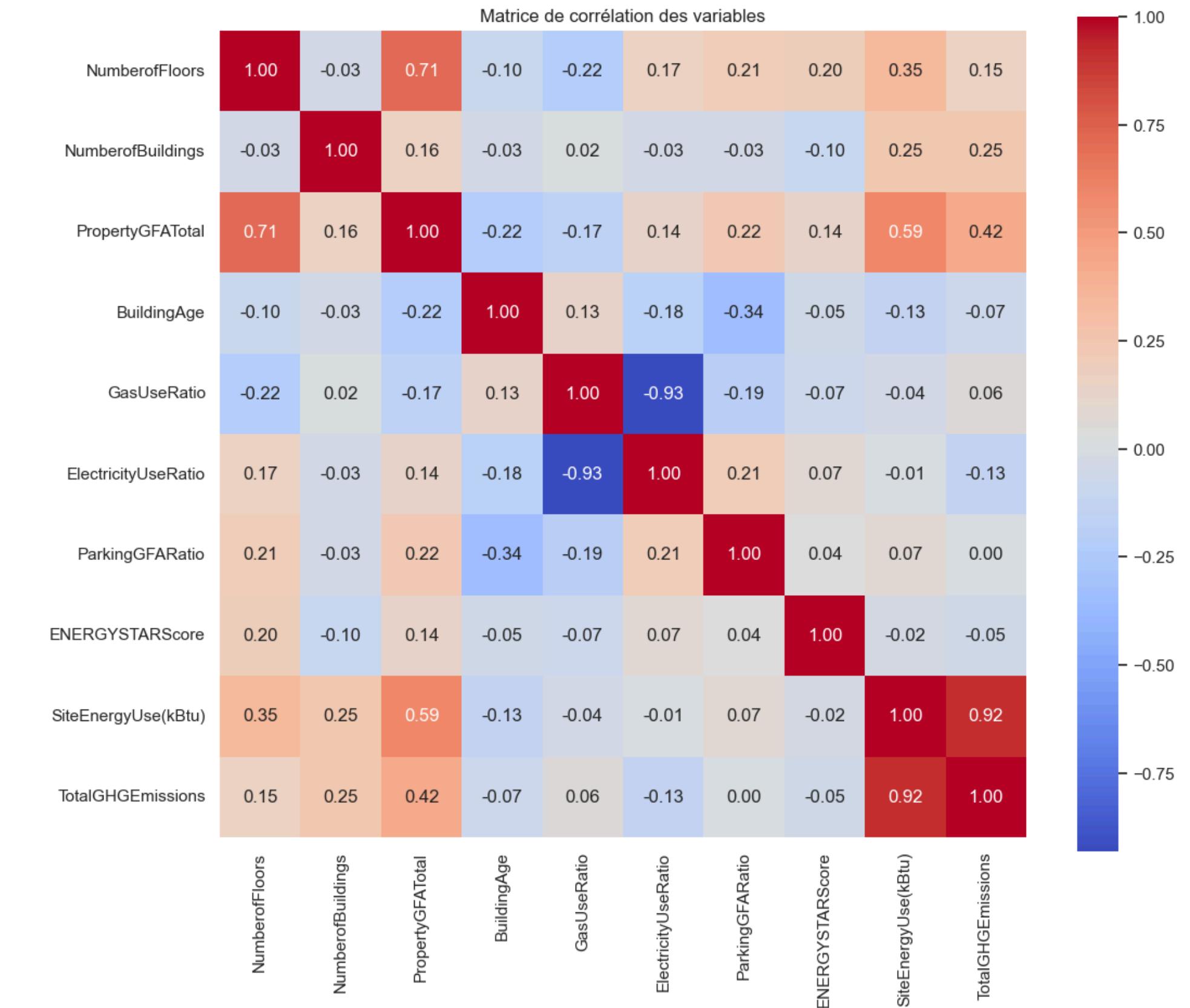


- Avec transformation logarithmique



Analyse des corrélations

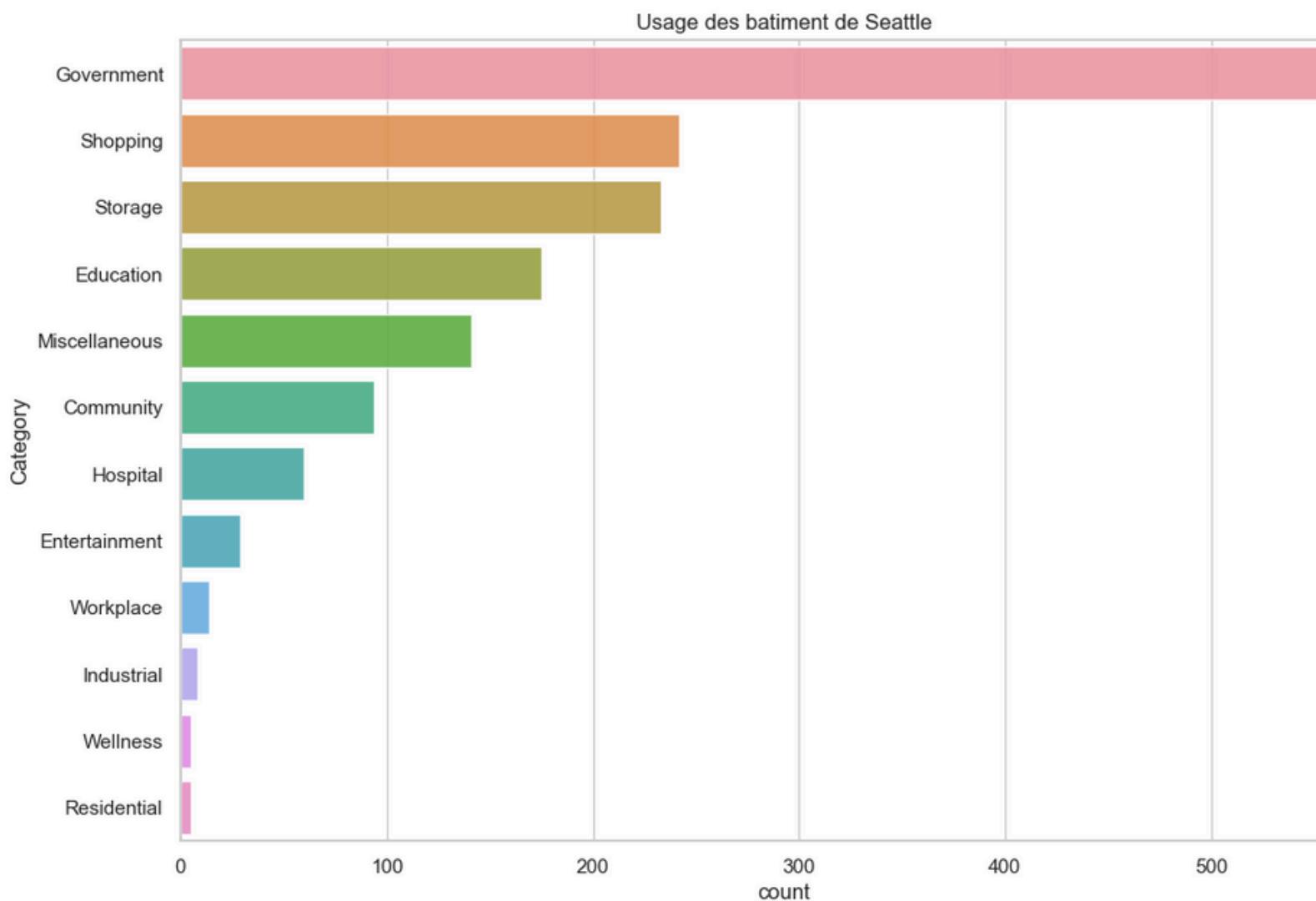
- Afin d'éviter le risque de sur-apprentissage, nous ne prenons pas en compte les variables trop corrélées



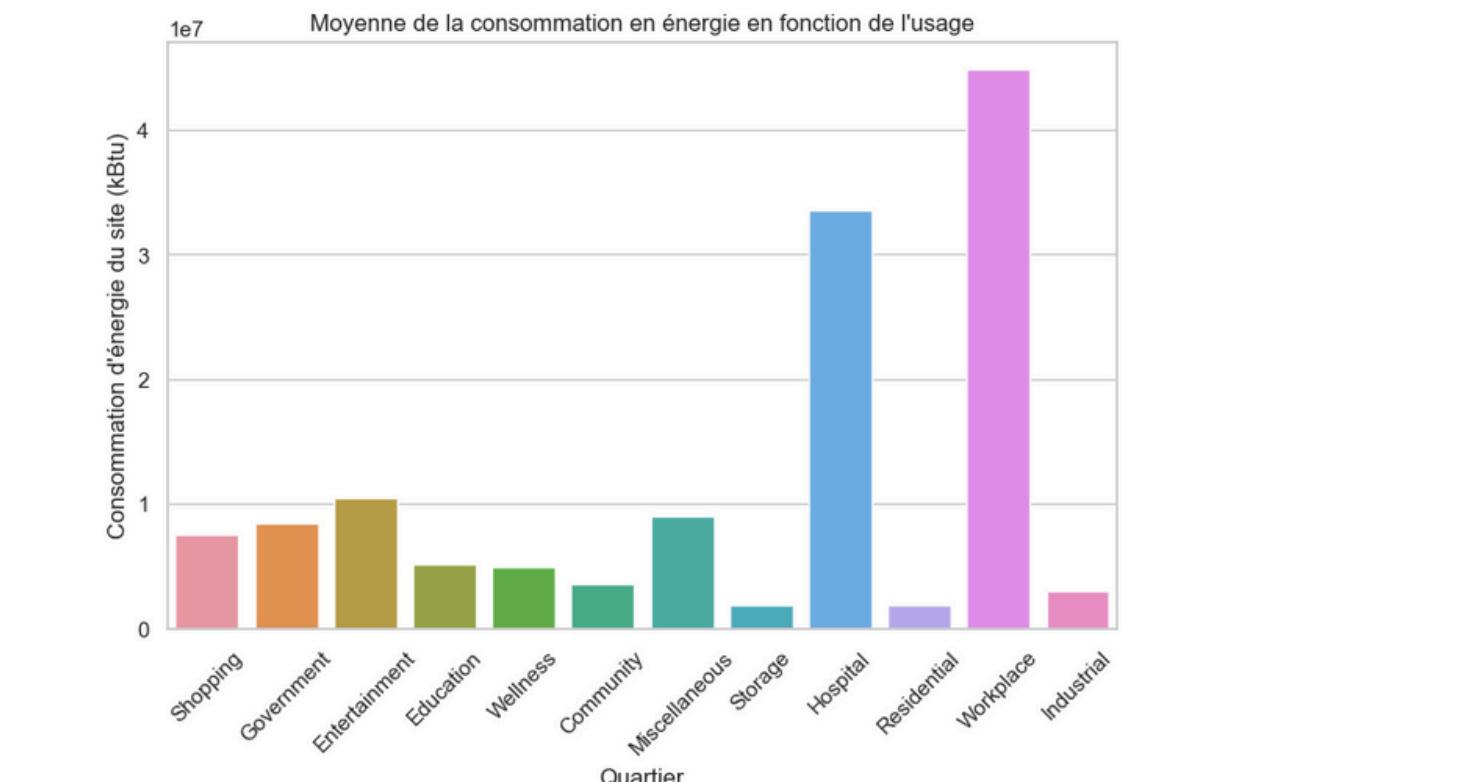
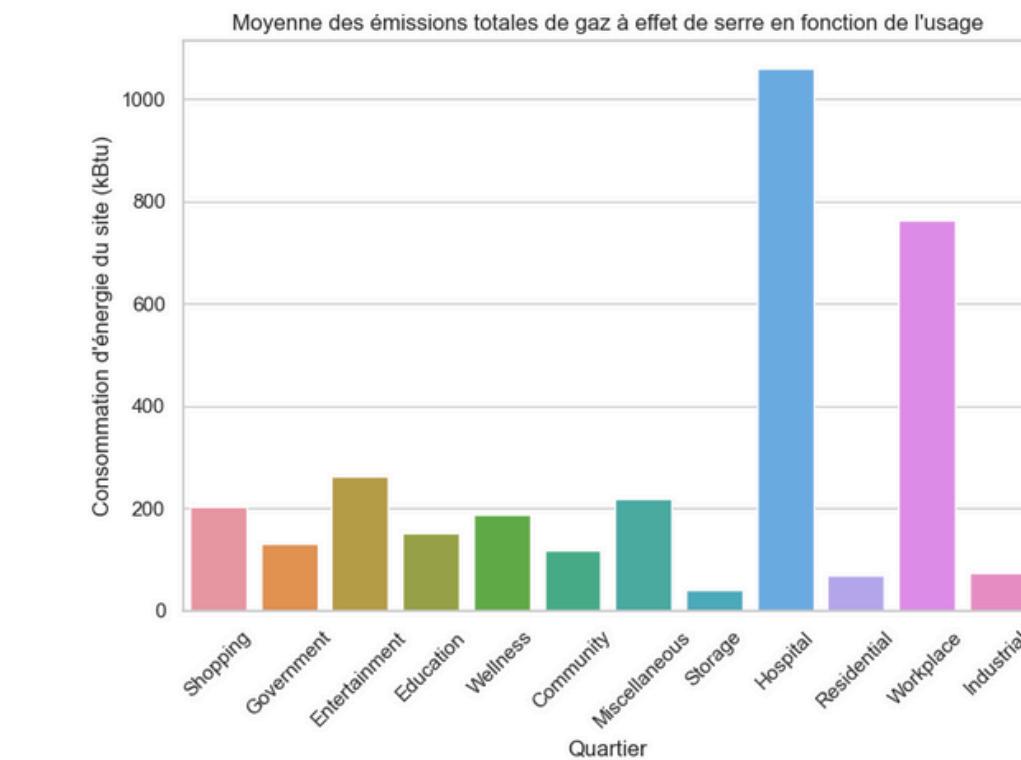
Analyse des variables qualitatives : LargestPropertyUseType

- Emission de gaz moyenne par type

- Décompte des bâtiments par type

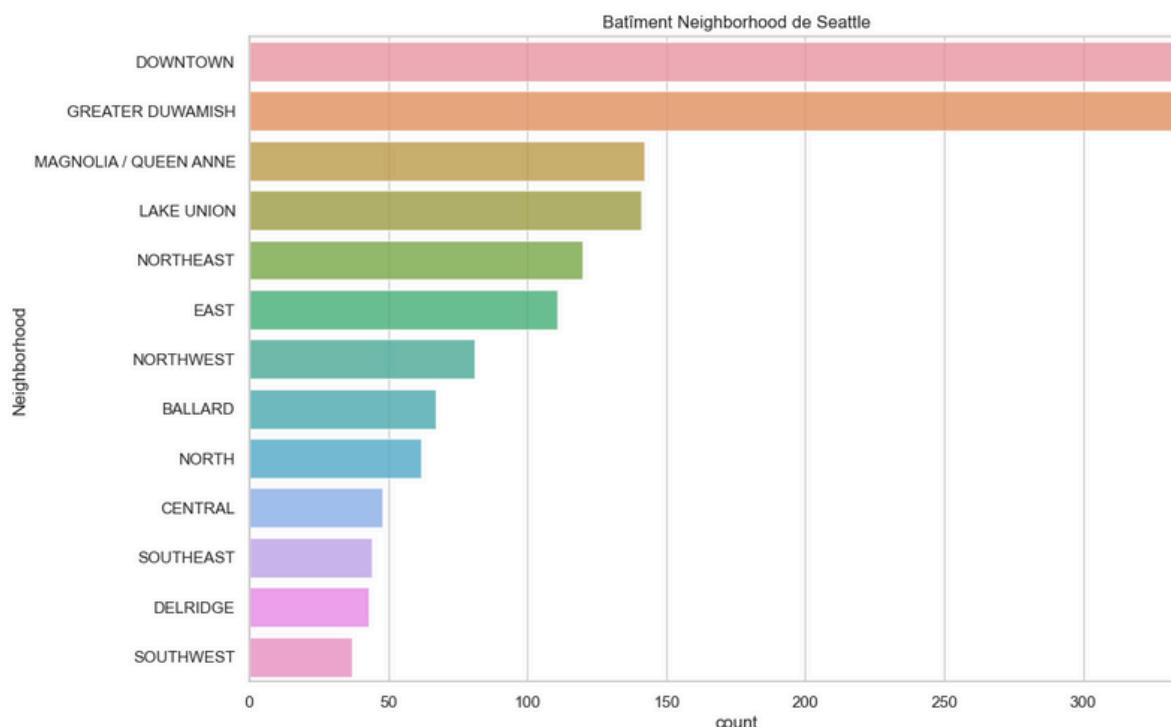


- Consommation énergétique moyenne par type

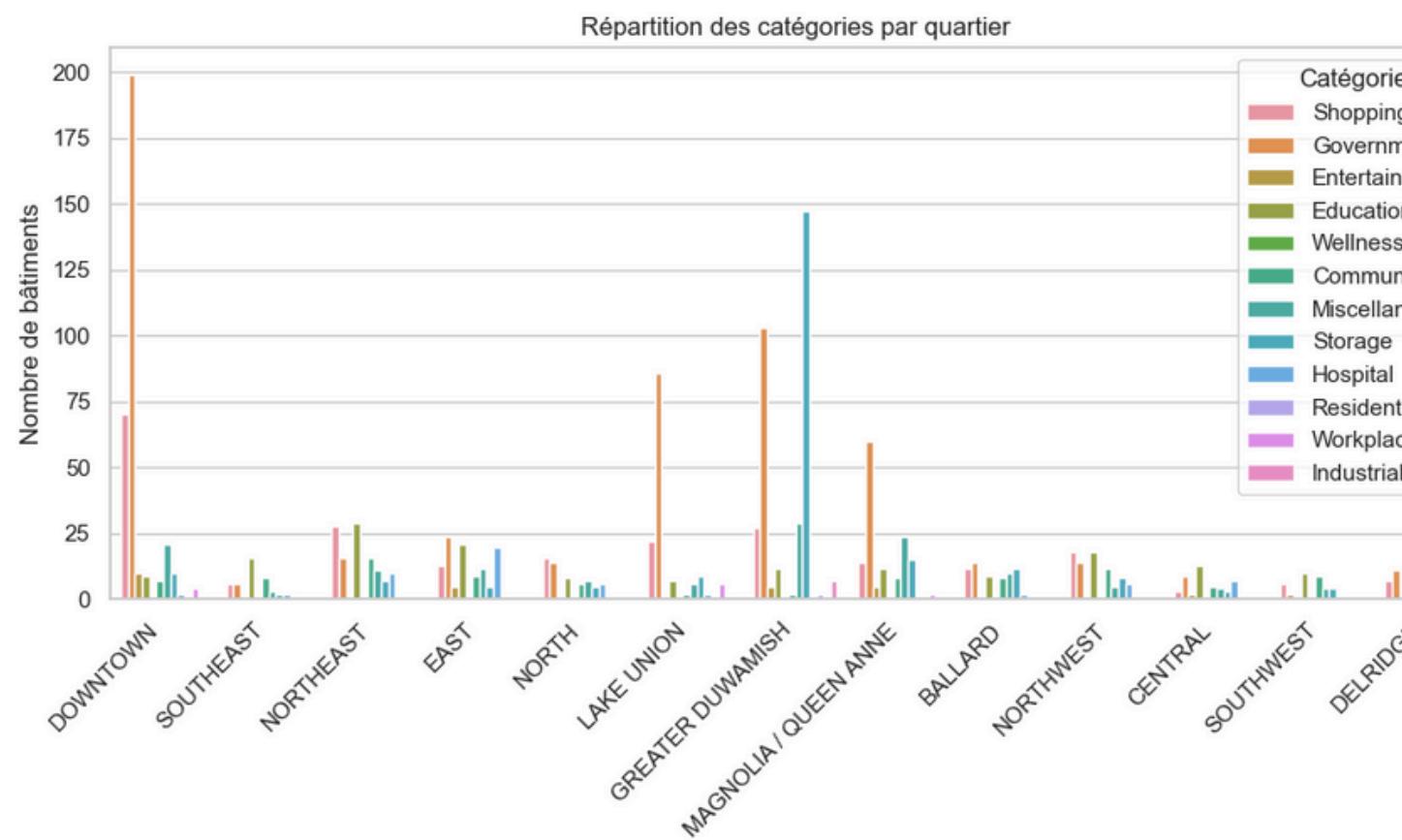


Analyse des variables qualitatives : Neighborhood

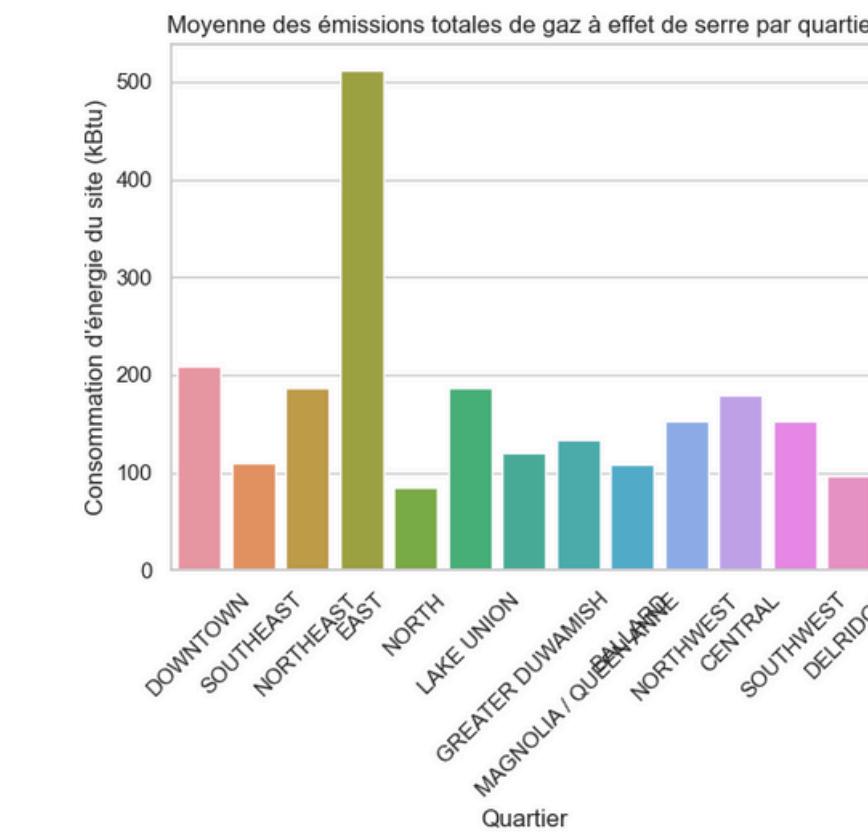
- Décompte des bâtiments par zone



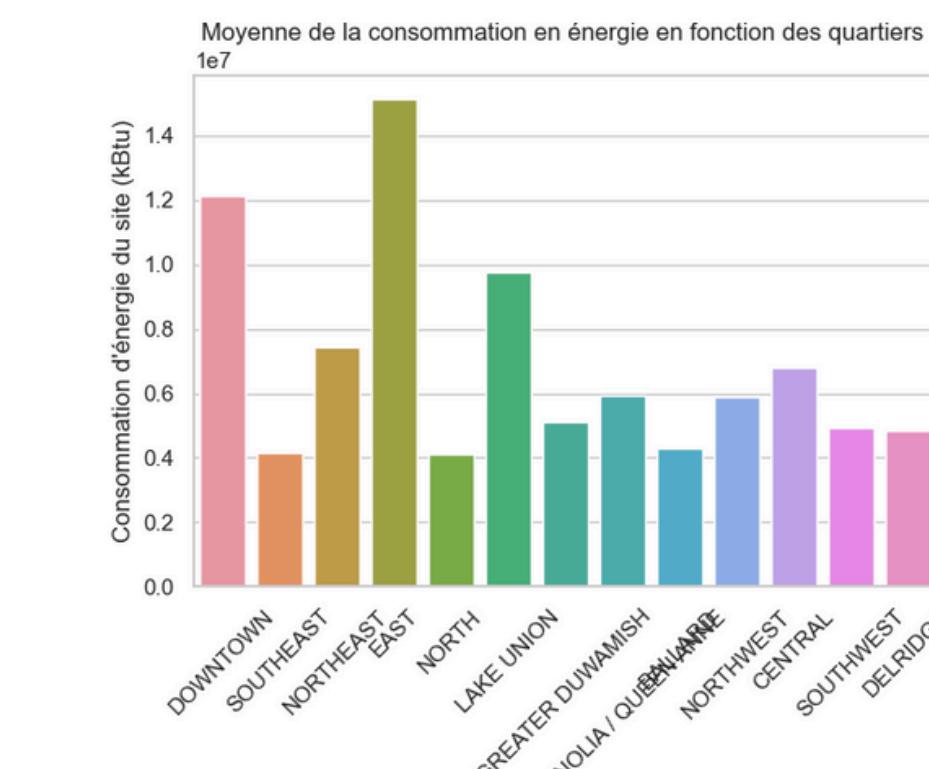
- Décompte du type de bâtiment par zone



- Emission de gaz moyenne par zone



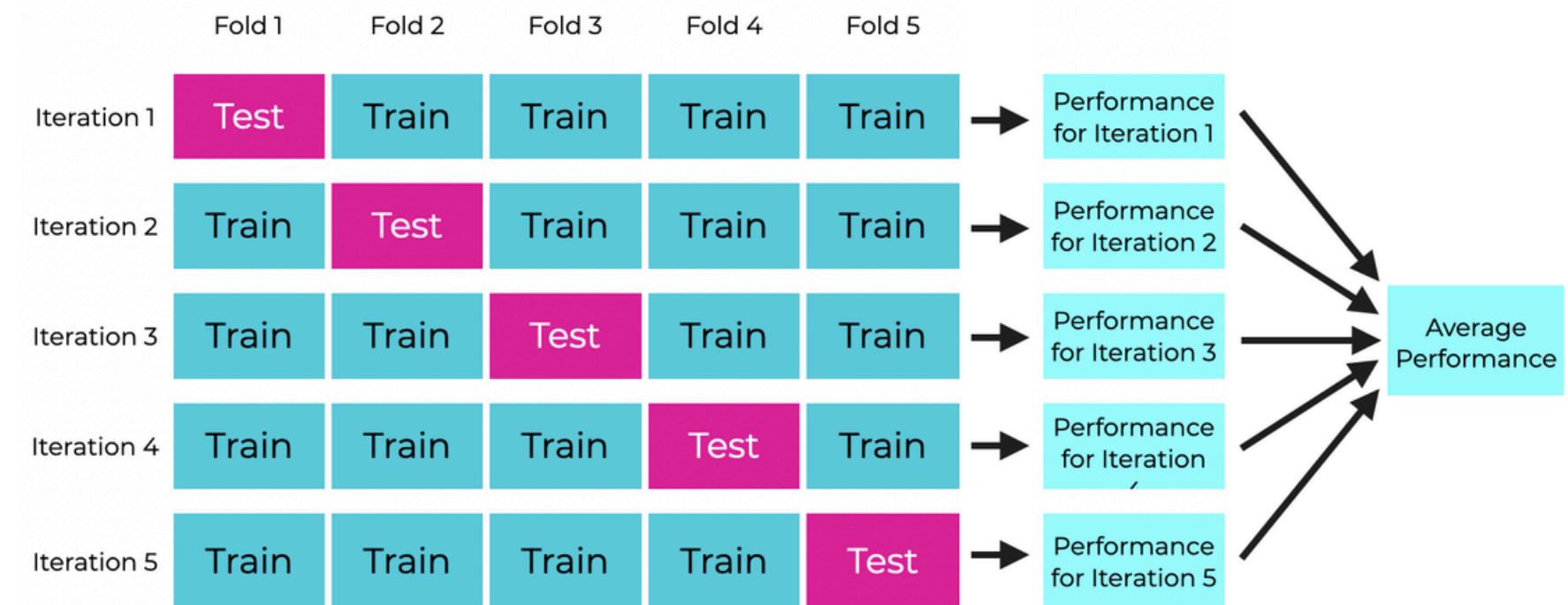
- Consommation énergétique moyenne par zone



Préparation des données

- Division des données en données de test et d'entraînement
- Standardisation des variables numériques via StandardScaler
- Encodage des variables catégorielles via OneHotEncoder
- Transformation logarithmiques des variables numériques via FunctionTransformer

Validation croisée



Modélisation

Modèles linéaires

- Régression Linéaire
- Dummy Regressor
- Régression Lasso
- Régression Ridge

Métriques utilisées

- R^2 : Coefficient de détermination
- MAE : Erreur absolue moyenne
- RMSE : Erreur quadratique moyenne
- Temps d'exécution

Modèles ensemblistes

- Random Forest
- Bagging Regressor
- Gradient Boost Regressor

Les modèles testés

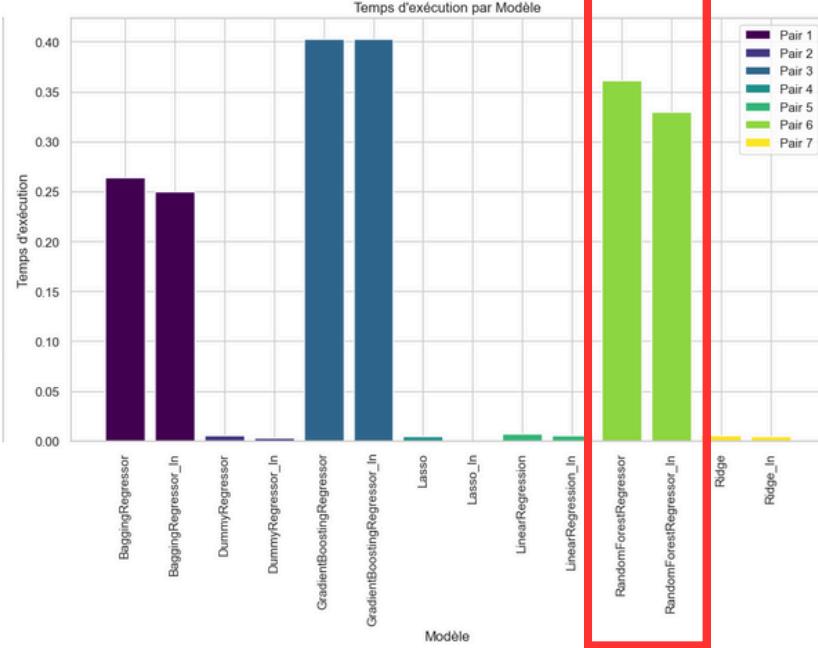
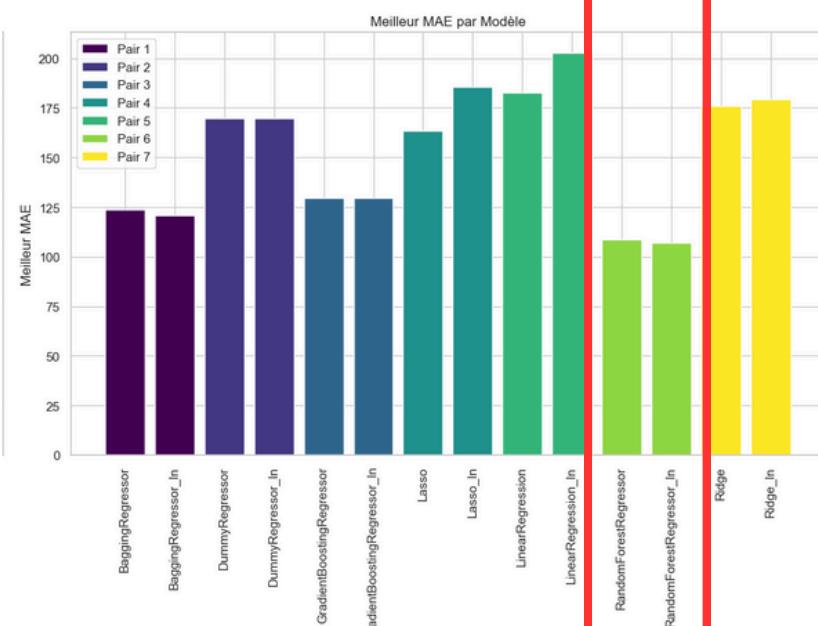
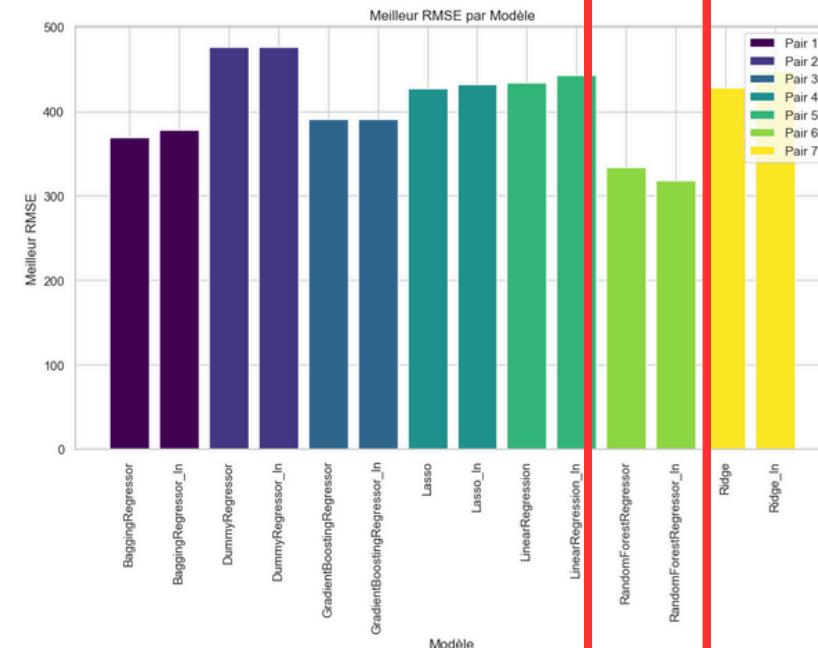
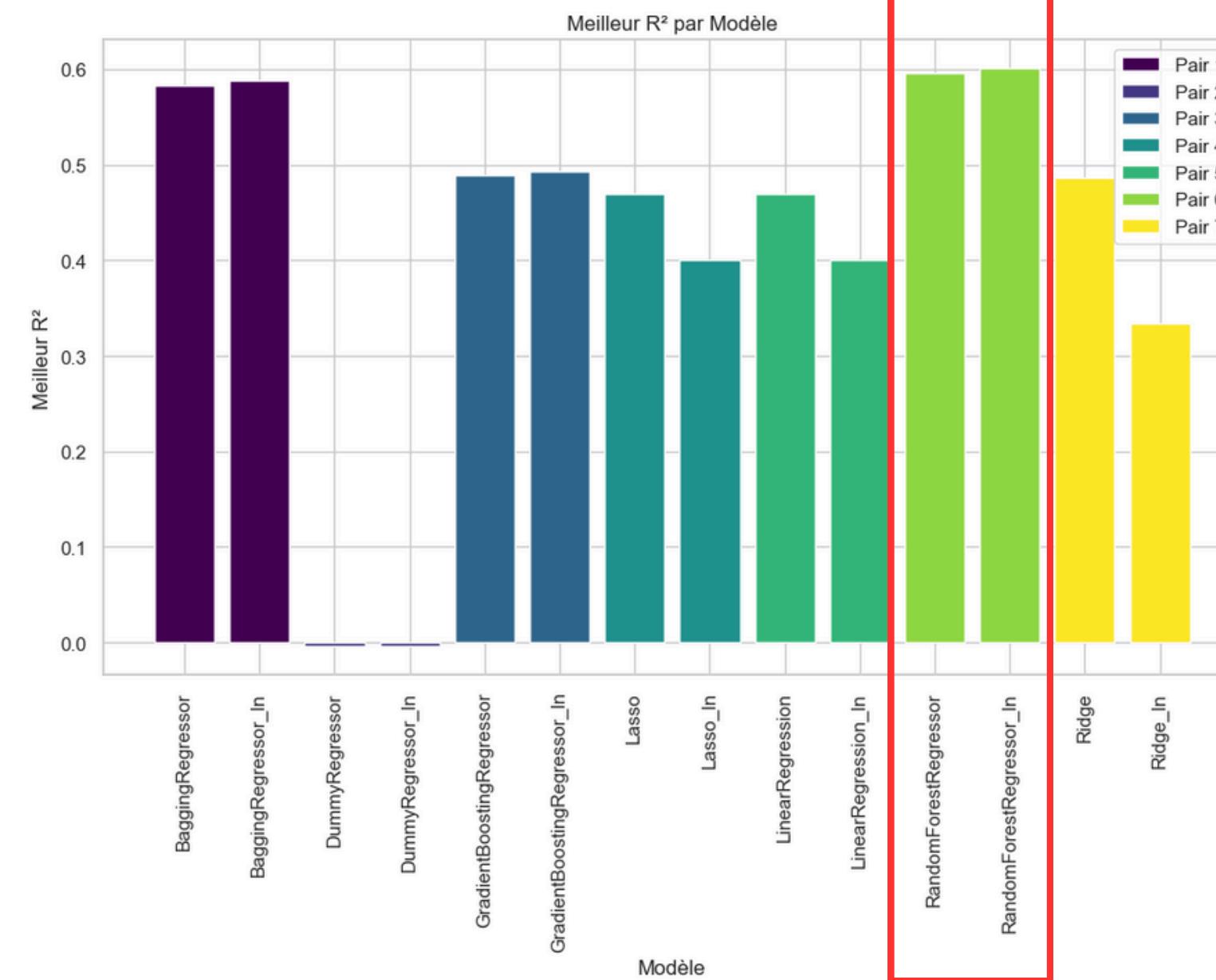


Choix du modèle le plus performant

Random Forest

- **R² : 0.62** > Le modèle explique près de 62% de la variance des données, indiquant une adéquation modérée.
- **MAE : 92.10** > Une erreur absolue moyenne de 106.89 montre une moyenne des erreurs de prédictions relativement élevée.
- **RMSE : 320.45** > Un RMSE de 318.18 suggère des écarts importants par rapport aux valeurs réelles, indiquant une précision potentiellement faible.
- **TE : 0.32** > Un temps d'exécution de 0.32 secondes indique une bonne rapidité de prédiction du modèle.

'TotalGHGEmissions'

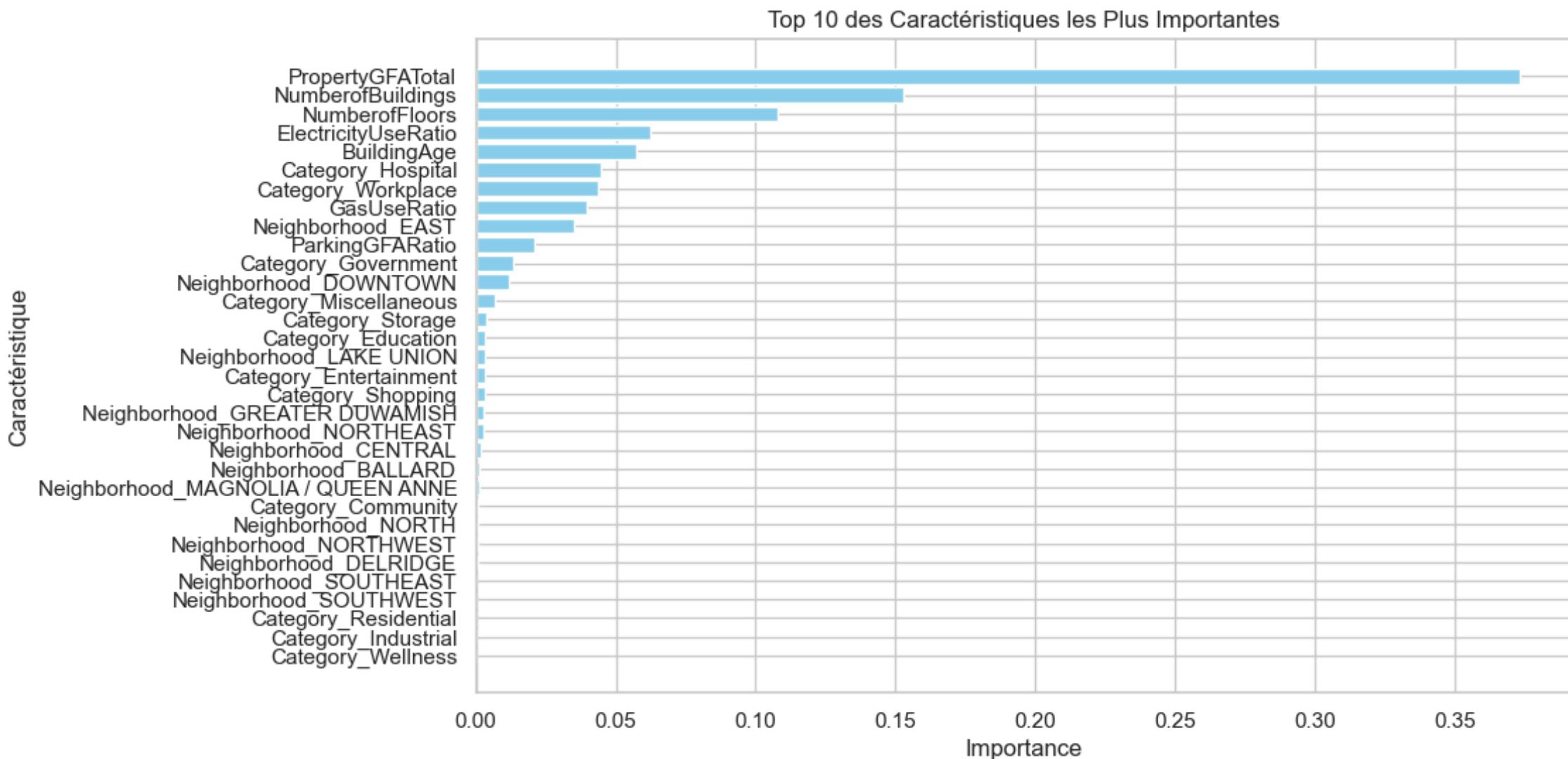


Optimisation des hyperparamètres

'TotalGHGEmissions'

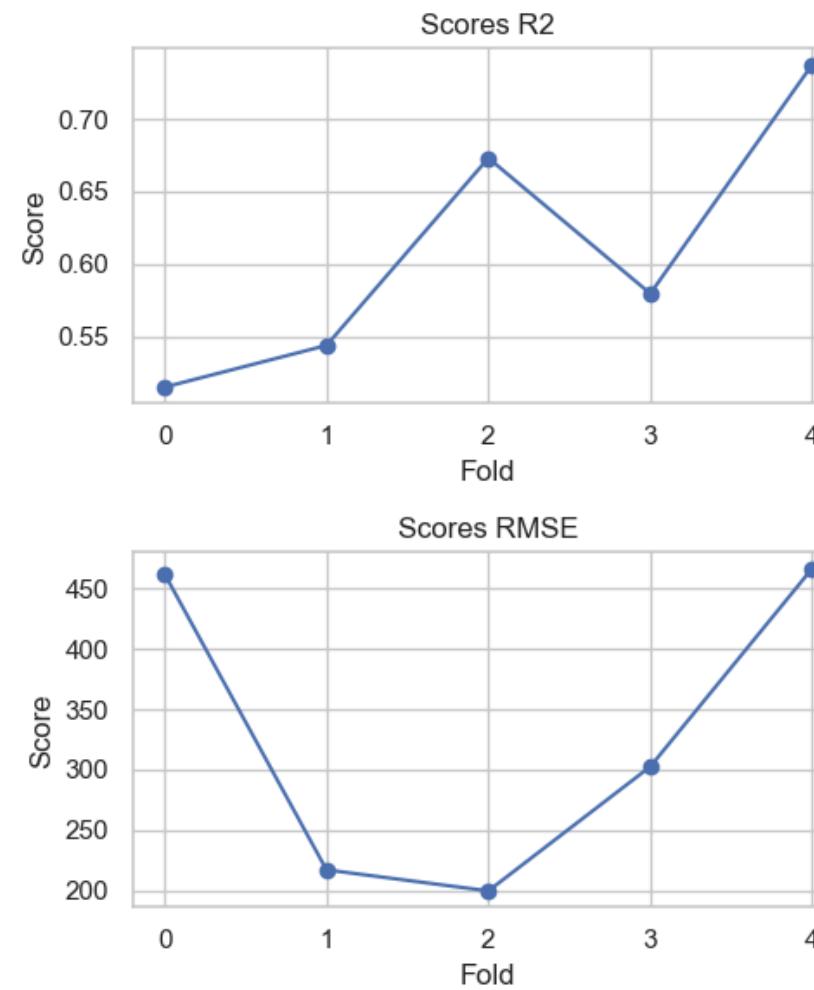
- GridSearchCV
- 1. model__bootstrap: [False]** > chaque arbre de la forêt est construit sur un échantillon de données différent
 - 2. model__max_depth: [15]** > profondeur maximale des arbres est limitée à 25 noeuds
 - 3. model__max_features: [0.8]** > on considère lors de la recherche 80% des variables
 - 4. model__min_samples_leaf: [1]** > nombre d'échantillon minimum pour créer un noeud est égal à 1
 - 5. model__min_samples_split: [2]** > nombre minimum d'échantillons requis pour diviser un noeud interne est de 2

- Feature Importance



Analyse du modèle

'TotalGHGEmissions'



- Scores des plis de la validation croisée
- Comparaison entre donnée d'entraînement et de test

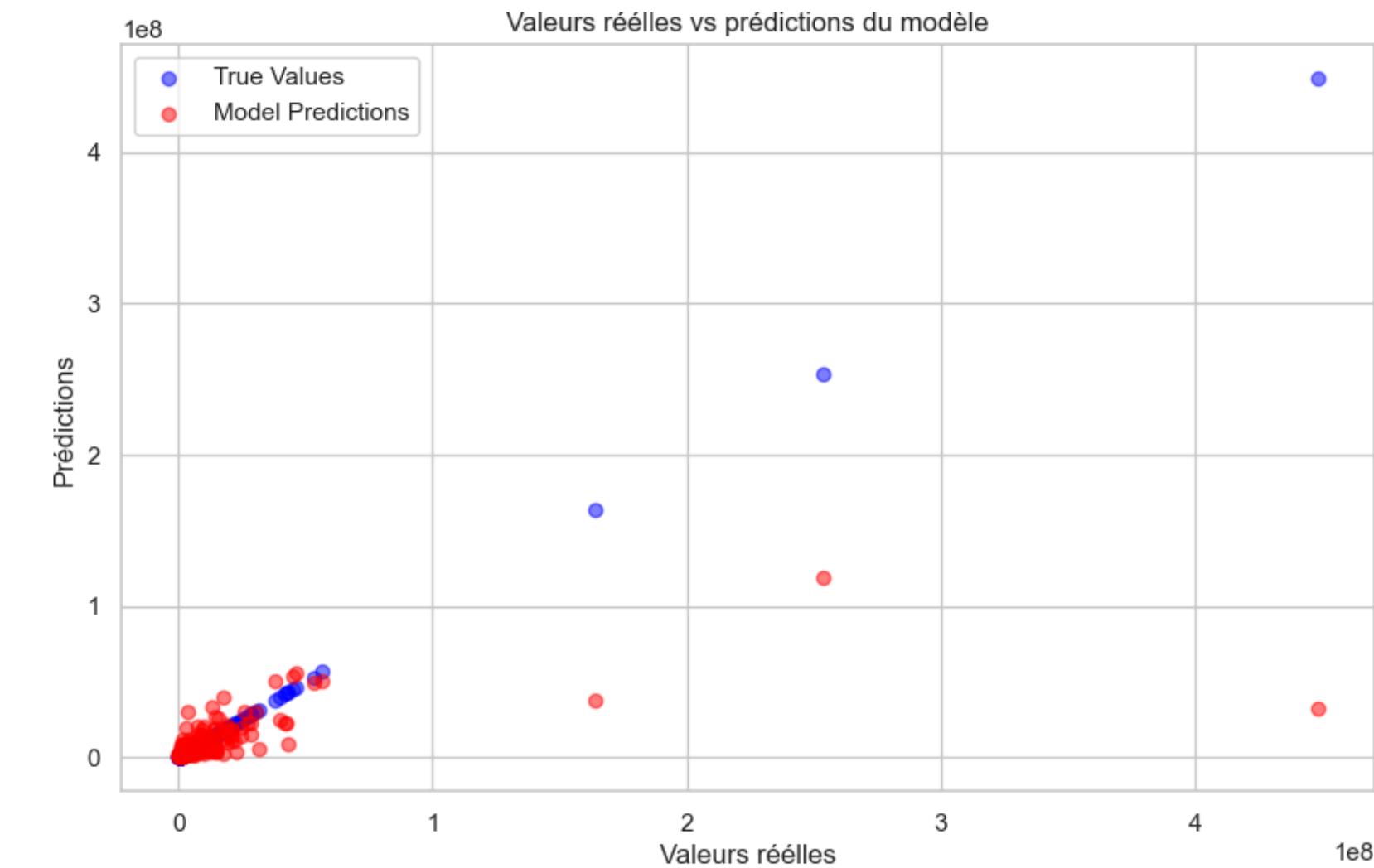
Train

$R^2 : 0.62$

Test

$R^2 : 0.56$

- Comparaison entre prédictions et valeurs réelles

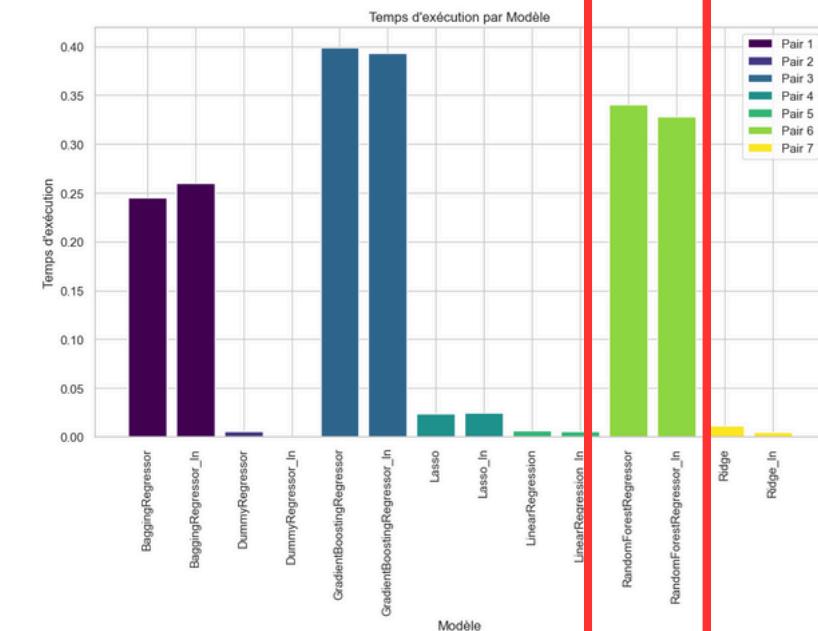
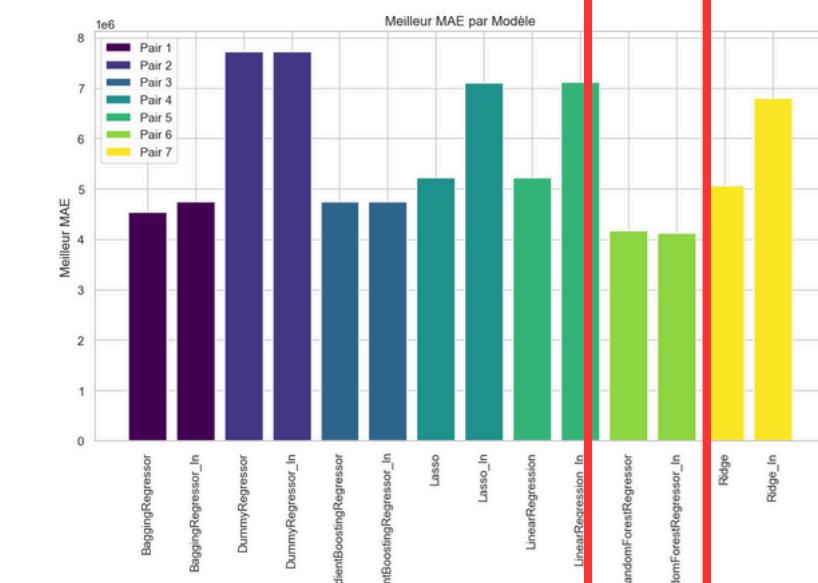
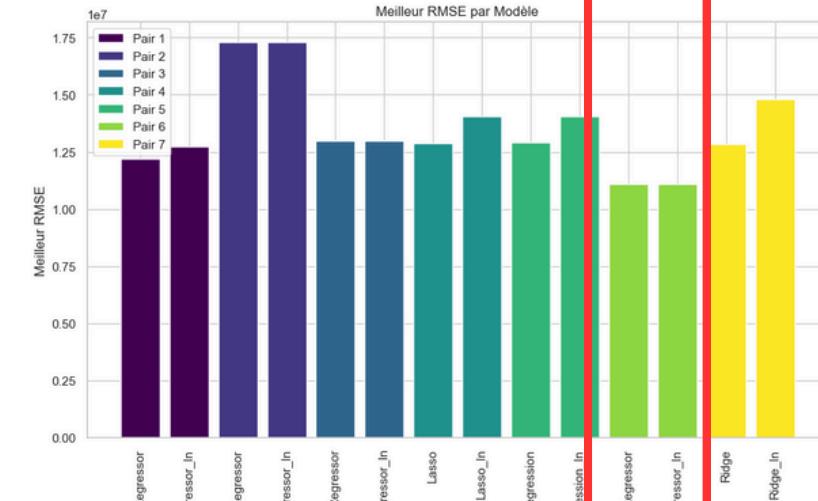
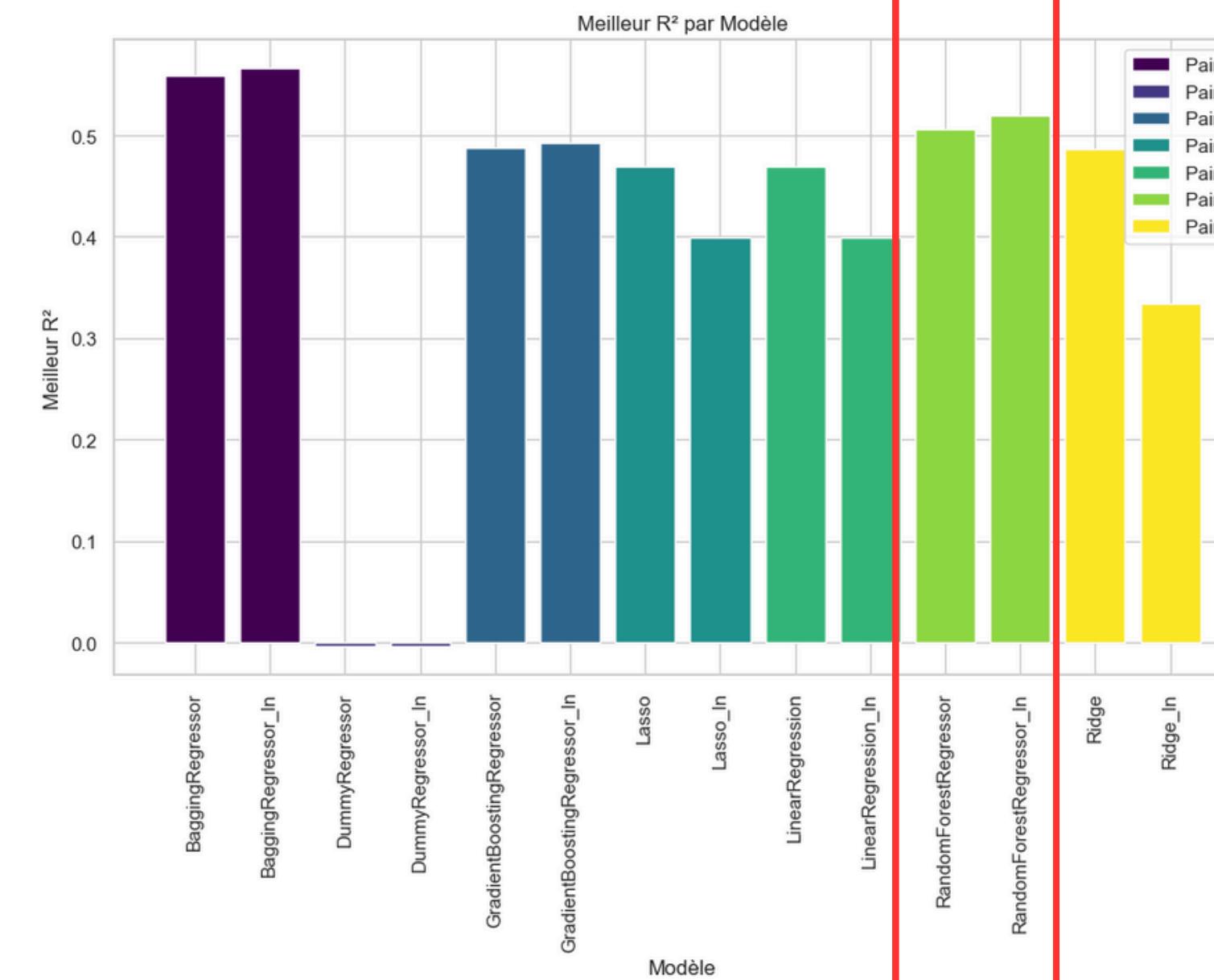


Choix du modèle le plus performant

Random Forest

- **R² : 0.60** > Le modèle explique près de 60% de la variance des données, indiquant une adéquation modérée.
- **MAE : 379.84** > Une erreur absolue moyenne de 379.84 montre une moyenne des erreurs de prédictions relativement élevée.
- **RMSE : 110.17** > Un RMSE de 110.17 suggère des écarts importants par rapport aux valeurs réelles, indiquant une précision potentiellement faible.
- **TE : 0.42** > Un temps d'exécution de 0.42 secondes indique une bonne rapidité de prédiction du modèle.

'SiteEnergyUse(kBtu)



Optimisation des hyperparamètres

'SiteEnergyUse(kBtu)

- GridSearchCV

1. **model__bootstrap: [False]** > chaque arbre de la forêt est construit sur un échantillon de données différent

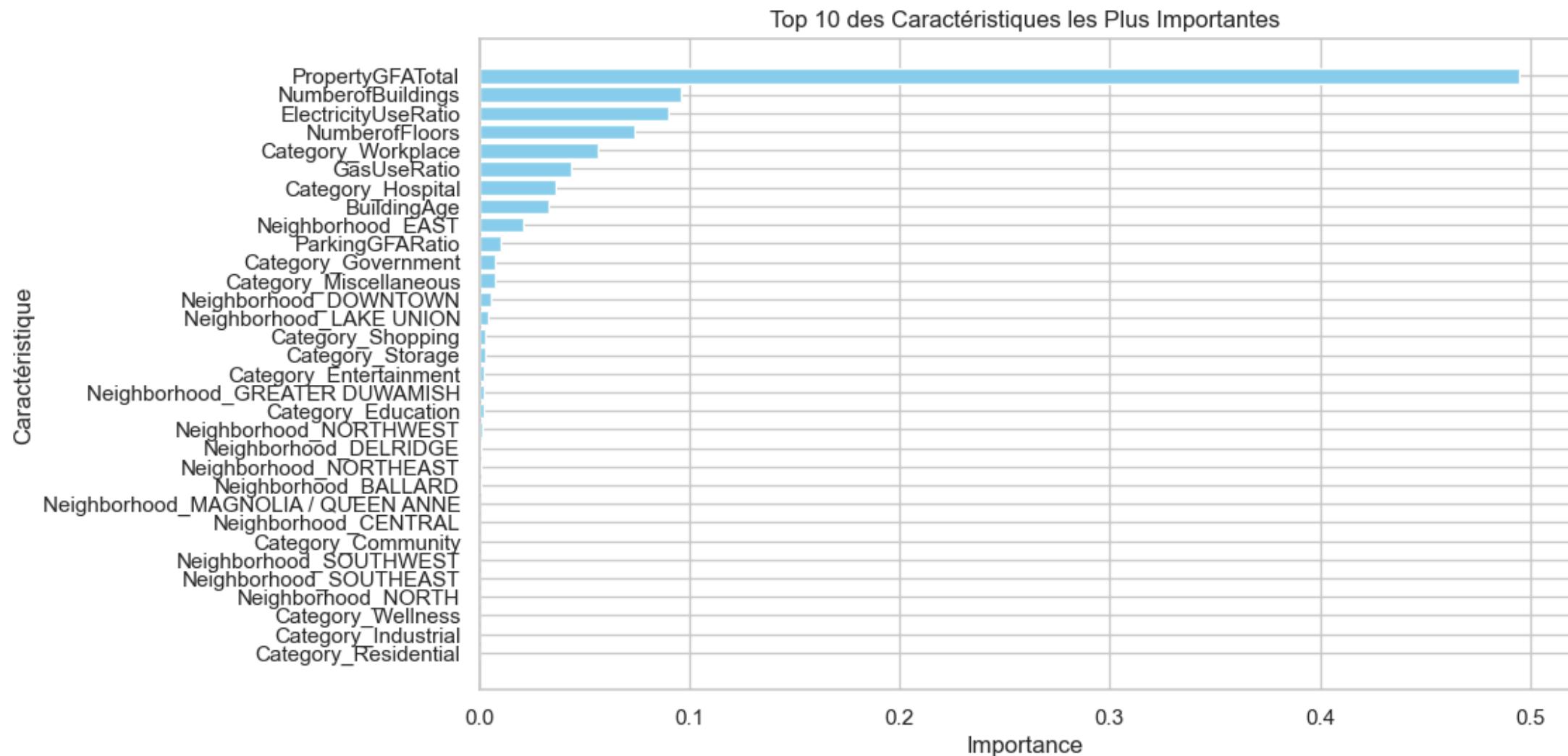
2. **model__max_depth: [15]** > profondeur maximale des arbres est limitée à 50 noeuds

3. **model__max_features: [0.8]** > on considère lors de la recherche de la meilleure division la 2ème échelle logarithmique du nombre total de fonctionnalités

4. **model__min_samples_leaf: [1]** > nombre d'échantillon minimum pour créer un noeud est égal à 1

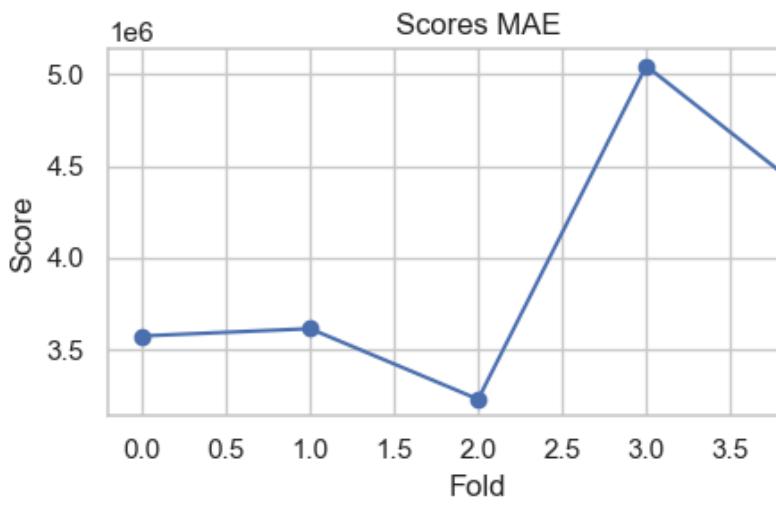
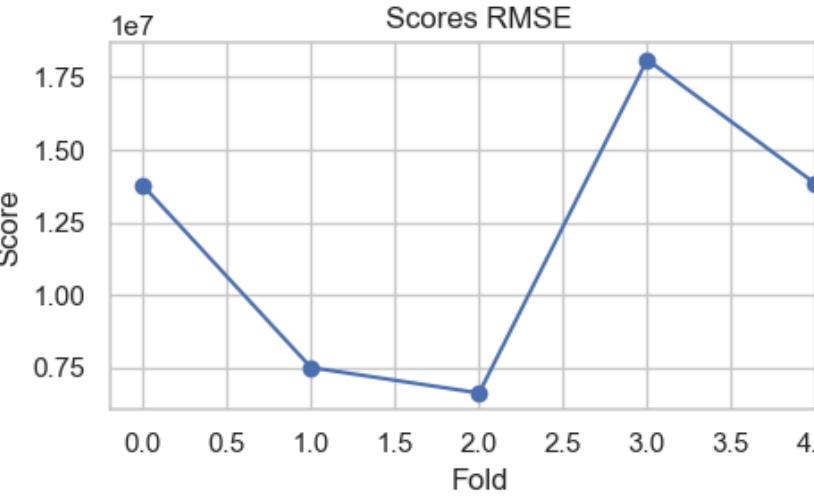
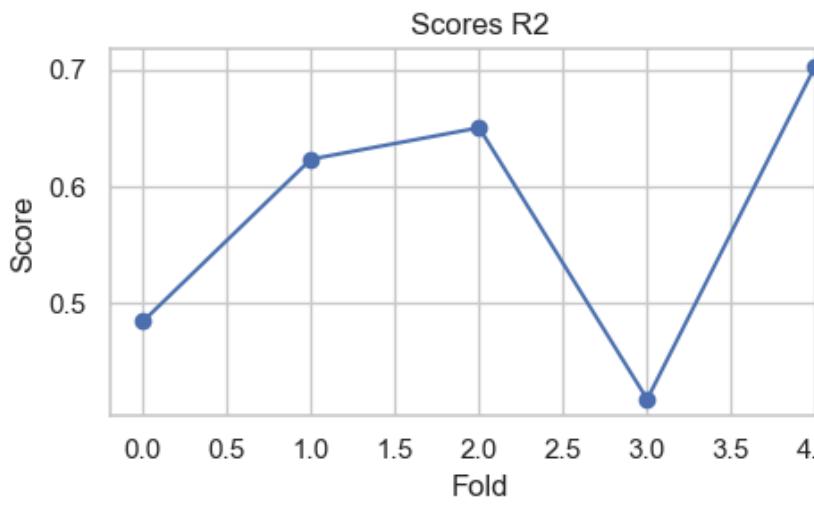
5. **model__min_samples_split: [2]** > nombre minimum d'échantillons requis pour diviser un noeud interne est de 2

- Feature Importance



Analyse du modèle

'SiteEnergyUse(kBtu)



- Scores des plis de la validation croisée

- Comparaison entre donnée d'entraînement et de test

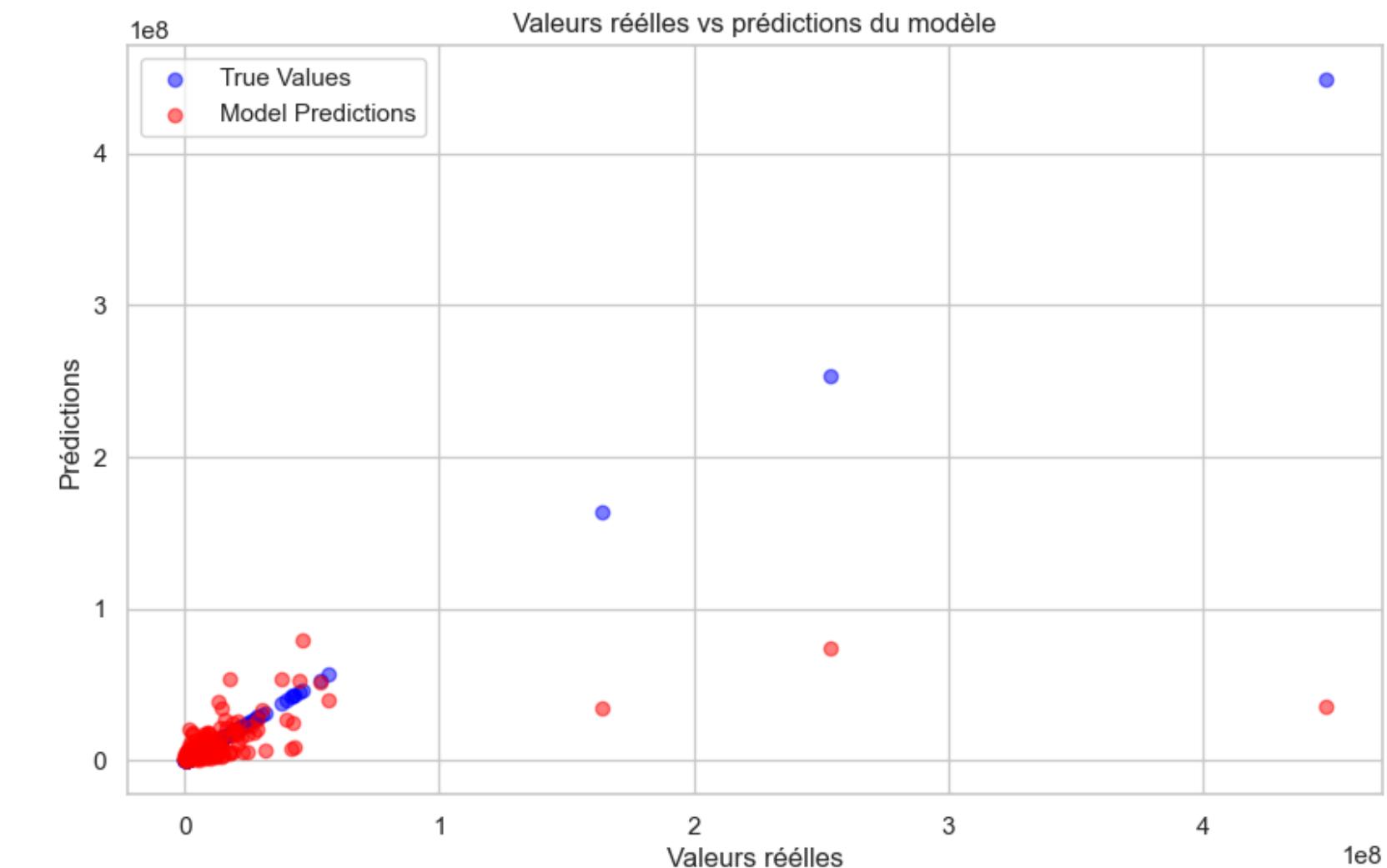
Train

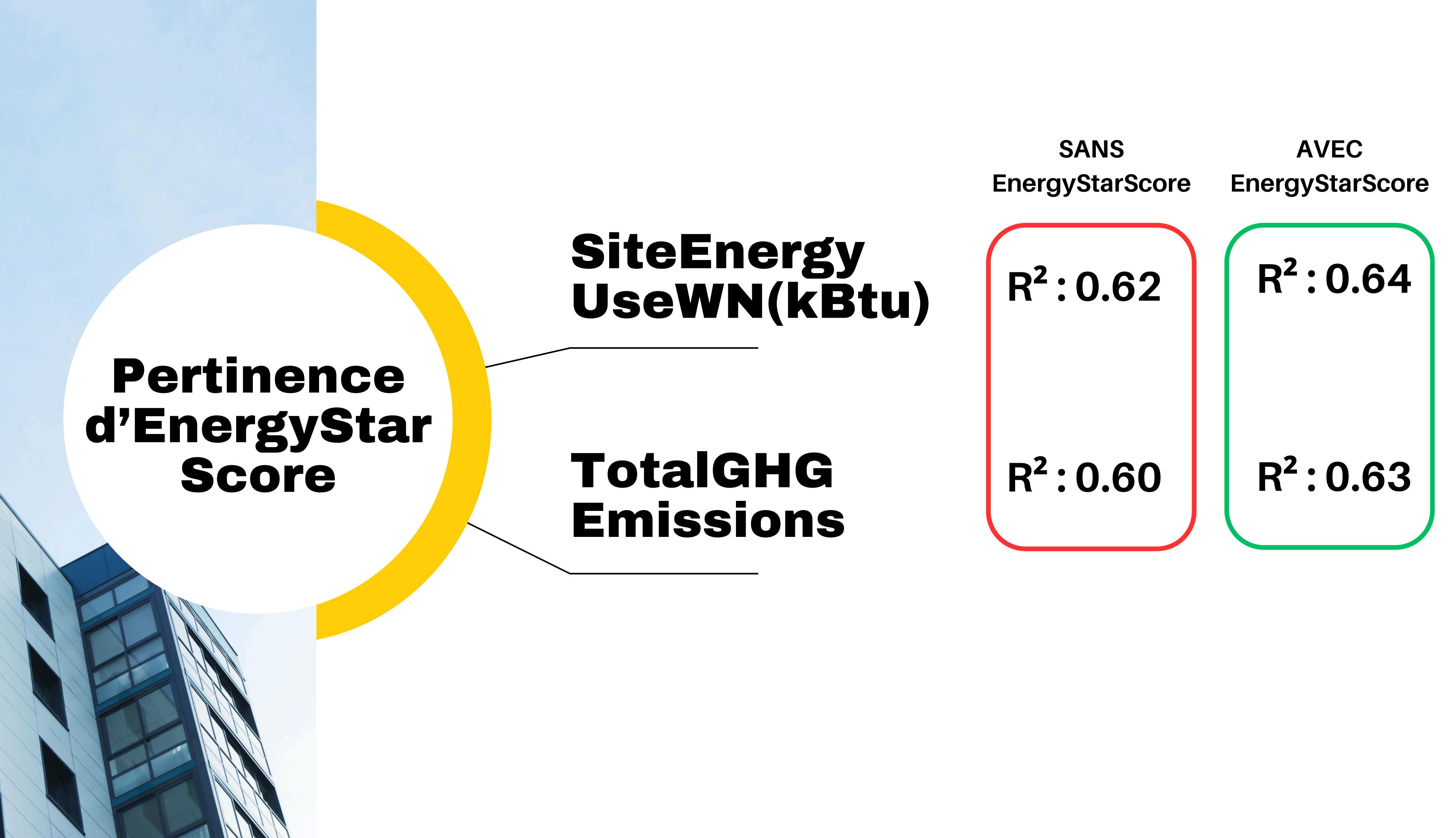
$R^2 : 0.60$

Test

$R^2 : 0.54$

- Comparaison entre prédictions et valeurs réelles







Conclusion

- La sélection des features et l'affinage des hyperparamètres ont joué un rôle primordial dans l'amélioration des performances du modèle
- La Random Forest s'est avérée être le modèle le plus pertinent dans notre cas grâce à ses performances satisfaisantes en termes de prédiction
- L'intégration de l'ENERGY STAR Score dans la modélisation a apporté une légère précision aux résultats du modèle



MERCI
pour votre attention

AVRIL 2024
LOKMAN AALIOUI