Seattle

Analyse et Modélisation des Émissions de CO_2 et de la Consommation d'Énergie

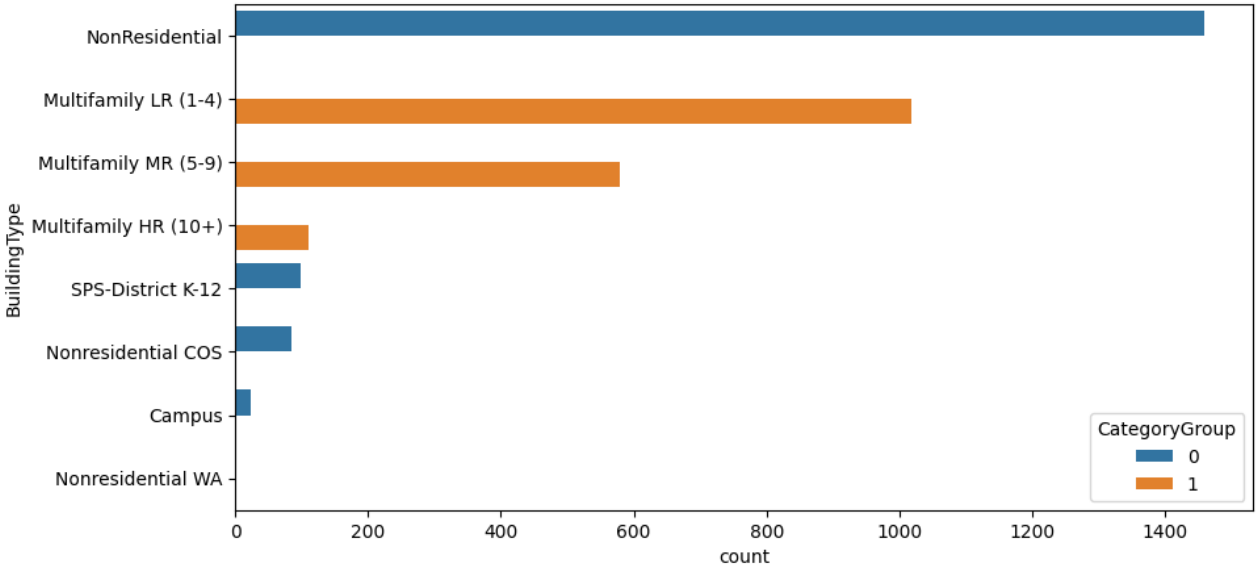
Sommaire

Anticiper les Besoins Énergétiques

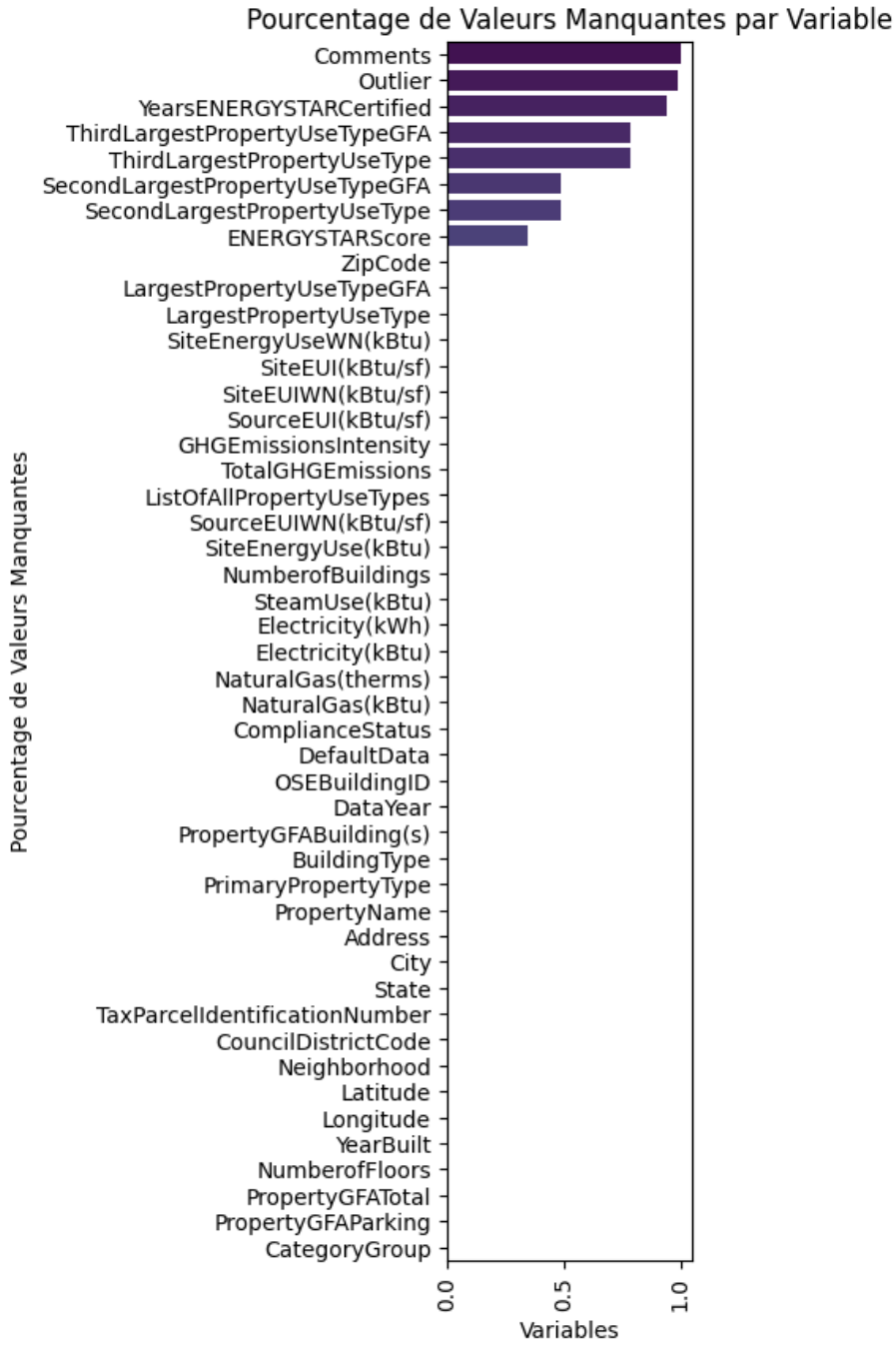
- **Mission :**
  - Prédiction des **émissions de CO_2** et de la **consommation totale d'énergie** des **bâtiments non résidentiels** à Seattle.
- **Contexte :**
  - Les **relevés** d'émissions et de consommation sont **coûteux** et **chronophage**s.
- **Données Structurelles des Bâtiments :**
  - Exploitation des relevés existants pour développer des modèles prédictifs basés sur les caractéristiques structurelles des bâtiments.
  - Pour tout nouveau bâtiment, un premier relevé de référence sera effectué la première année.
- **Objectif à Long Terme :**
  - Contribuer à l'objectif de neutralité carbone de Seattle d'ici 2050.

Description du jeu de données

- **Nombre de Bâtiments :**
  - Total de bâtiments analysés : 3 376.
- **Types de Bâtiments :**



- **Variables cibles** : Émissions de CO_2 et consommation d'énergie.
- **Variables structurelles :** Catégorielles (p. ex., BuildingType) et Numériques (p. ex., superficie, année de construction).



Qualité des Données :

Valeurs manquantes :

Bien renseignée.

ENERGYSTARSscore.

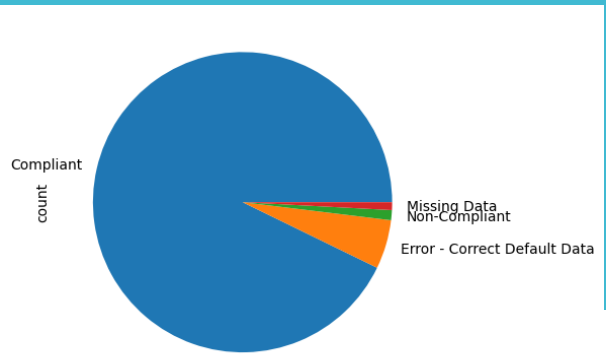
Outliers :

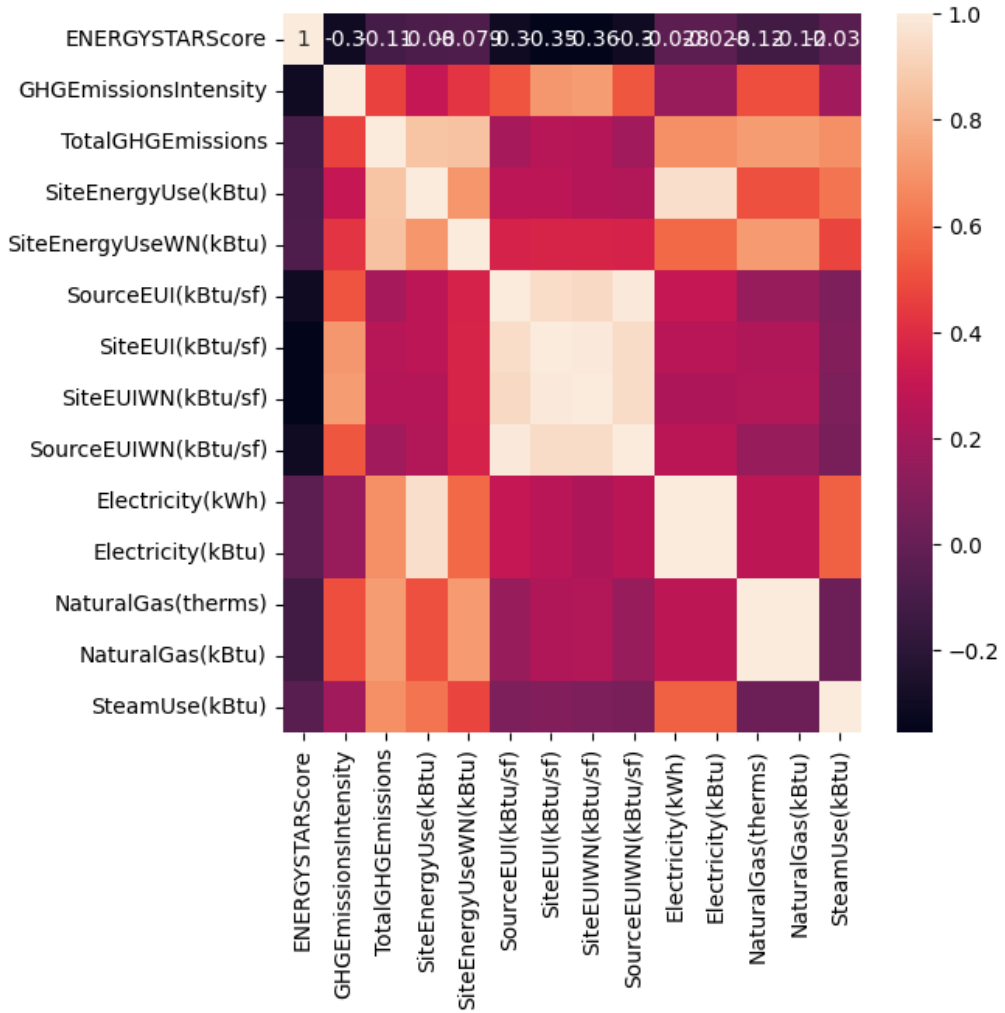
High : Data center.

Low : Office.

ComplianceStatus :

1 548 bâtiments.





Variables cibles

Consommation d'électricité vs. Energie sur site: Corrélation très forte (Pearson: 0,95).

Consommation d'énergie vs. Émissions CO_2: Corrélation forte, plus marquée sans normalisation.

Choix pour l'analyse:

"SiteEUIWN(kBtu/sf)" et "GHGEmissionsIntensity".

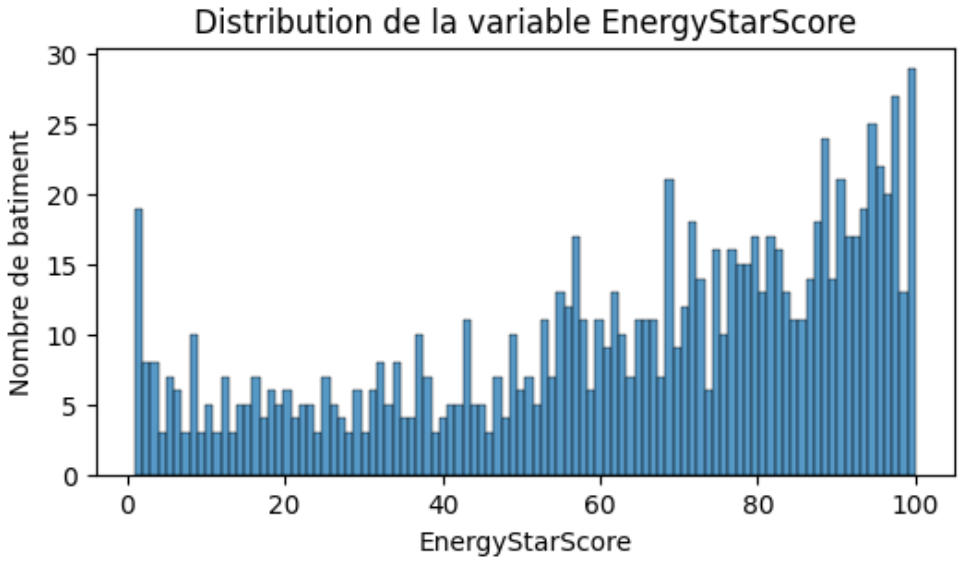
Comparaisons sur une base équitable.

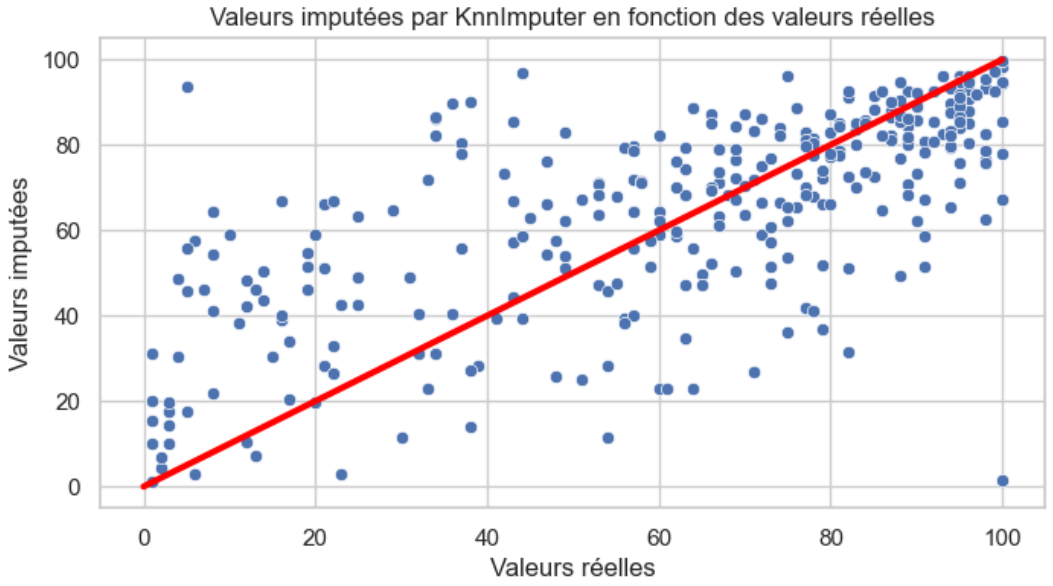
ENERGYSTARScore :

Note de performance de 1 à 100.
Calculé à partir de la consommation, l'utilisation du type d'immeuble, l'occupation, la superficie…

Corrélation modeste avec variables cibles : -0,30

KNNImputer pour les valeurs manquantes.
MAE=15,5



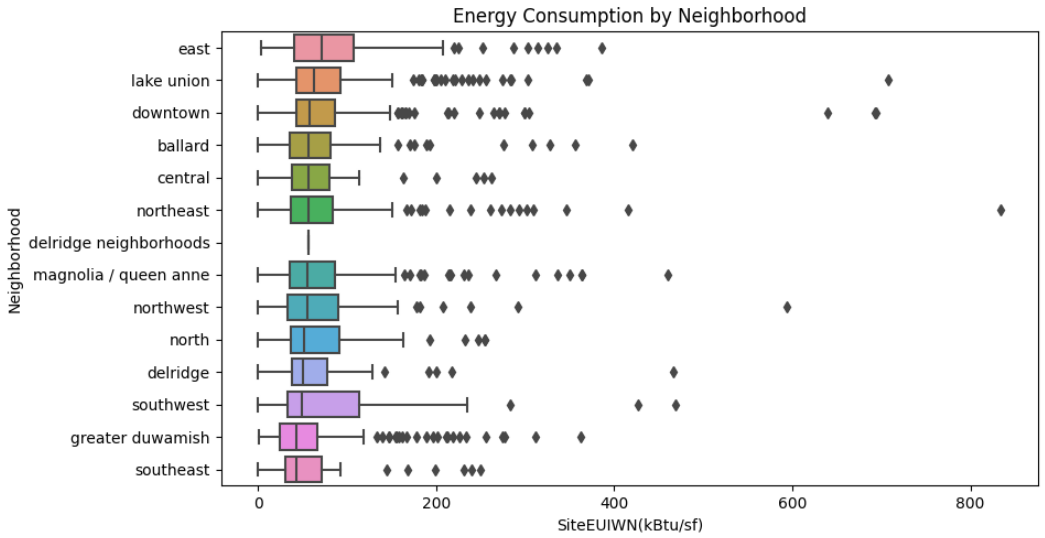


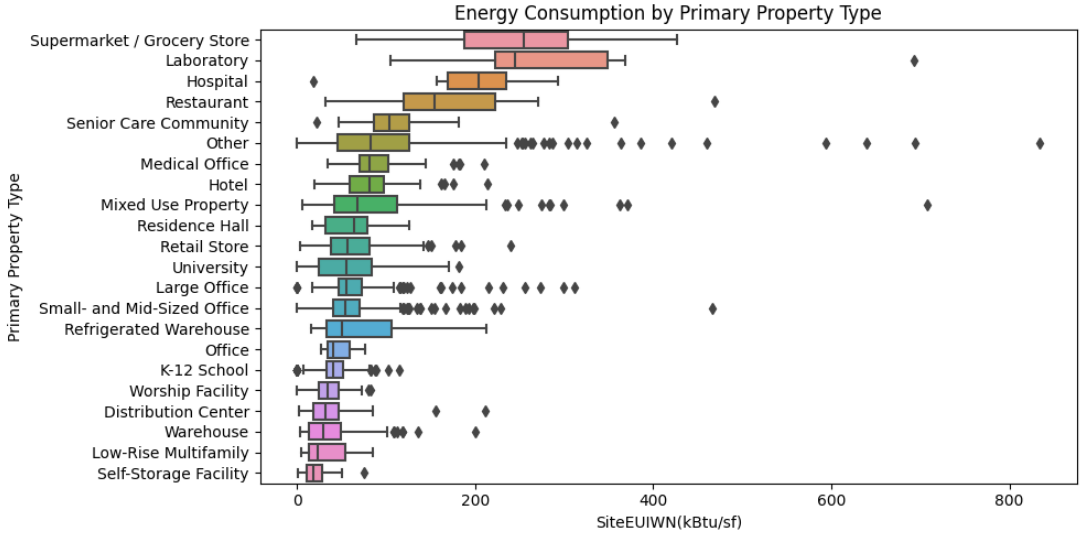
Variables structurelles Catégorielles

Type de construction.
Usage du bâtiment.
Quartiers.

Simplification pour les variables avec un grand nombre de catégories en 3 niveaux d'émissions : Haute / Moyenne/Faible

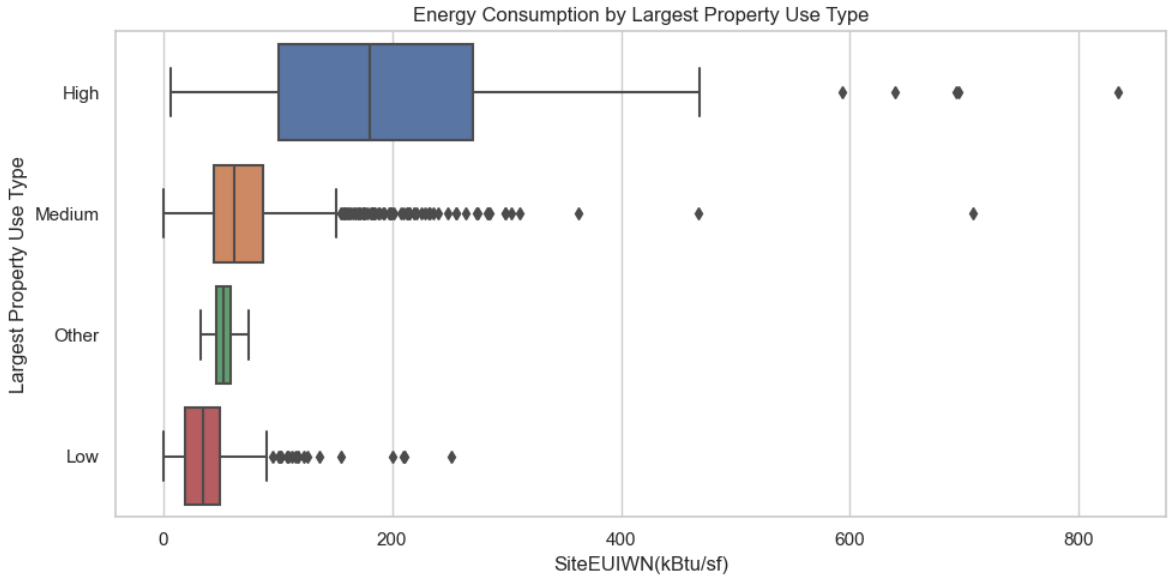
Réduire la complexité du modèle et potentiellement à améliorer sa généralisation et performance.

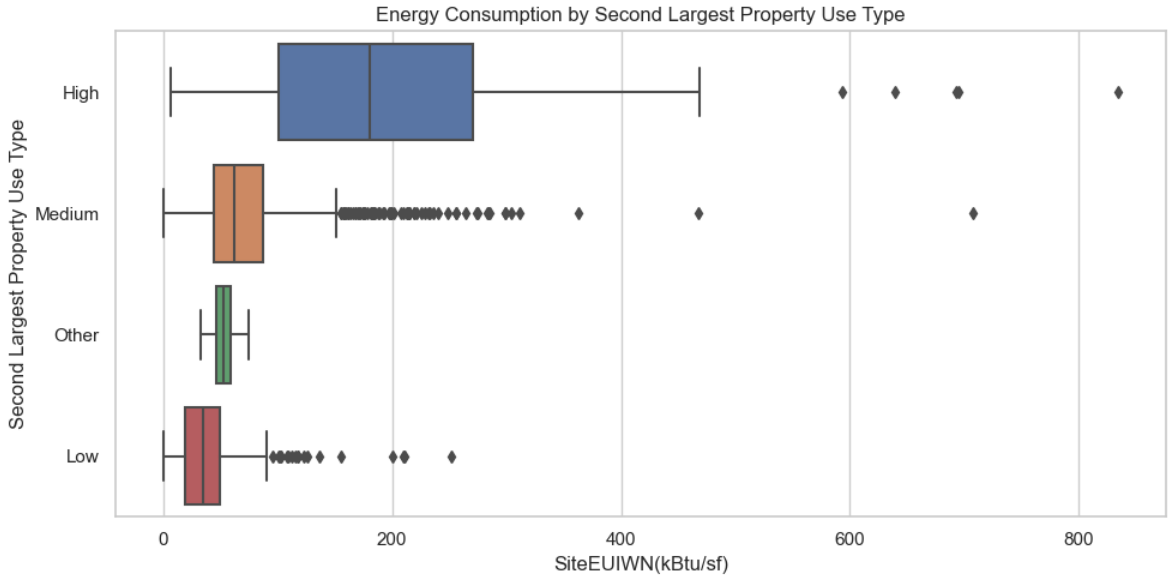




Variables structurelles catégorielles

LPUT et SLPUT :
Division en 3 groupes de catégorie pour réduire le nombre de variables au moment de l'encodage :
Low, Medium, High





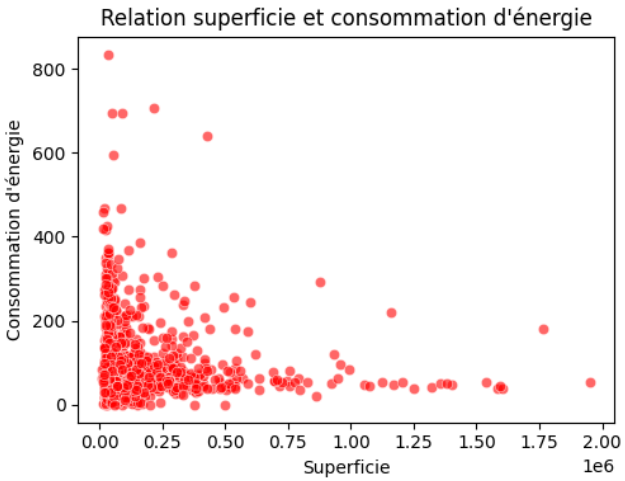
Variables structurelles

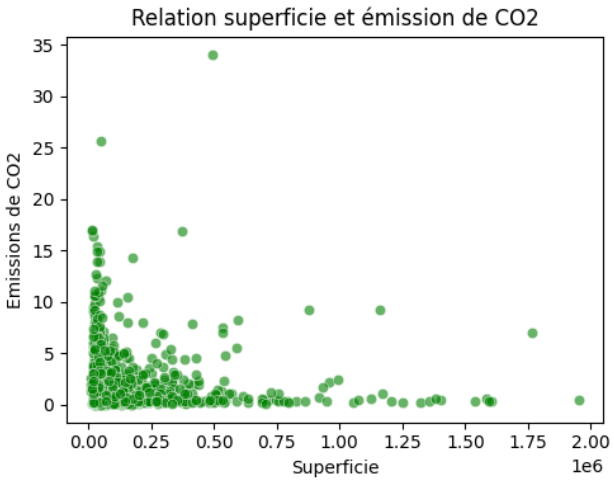
Date de Construction
Superficie
Nombre d'Étages

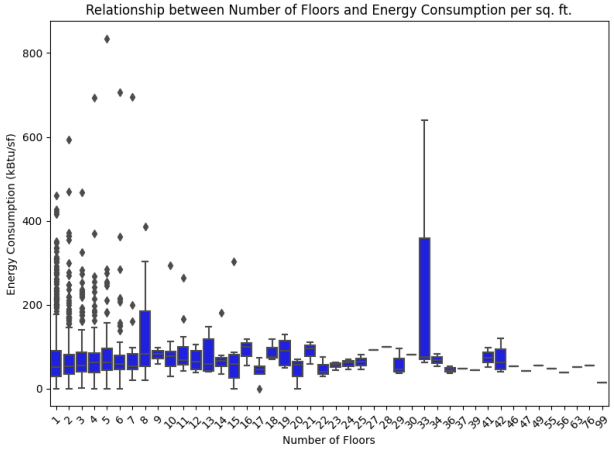
Gestion des Superficies :

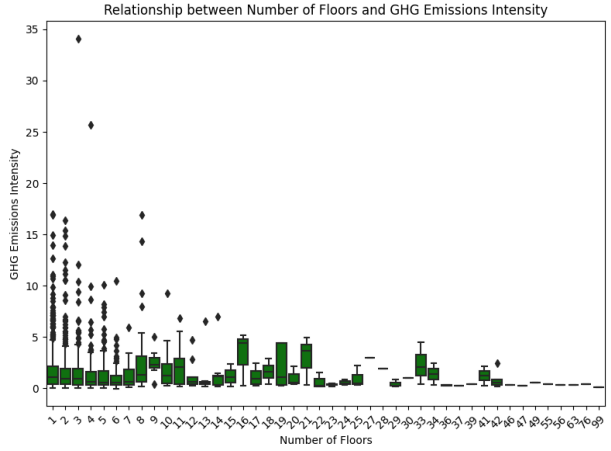
Superficie du Bâtiment / Superficie du Parking : Considérée séparément pour refléter l'espace non chauffé.

Autres variables à tester :
Identifiant Taxe.
ZipCode.





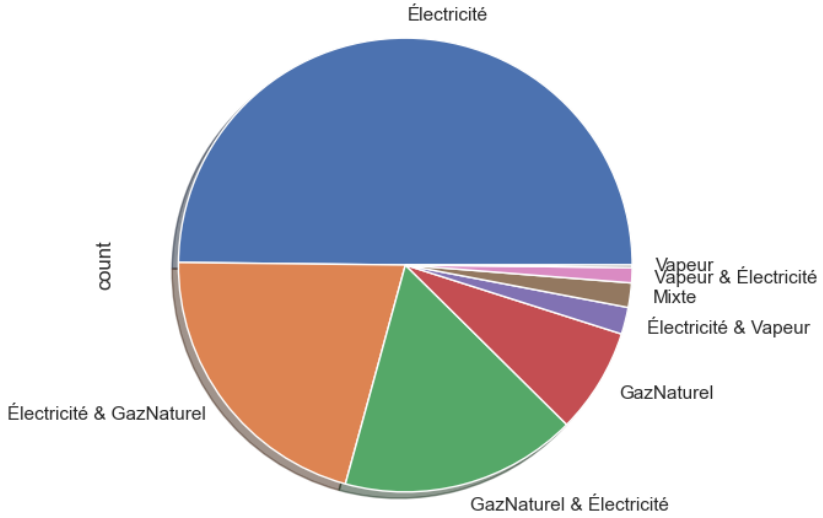


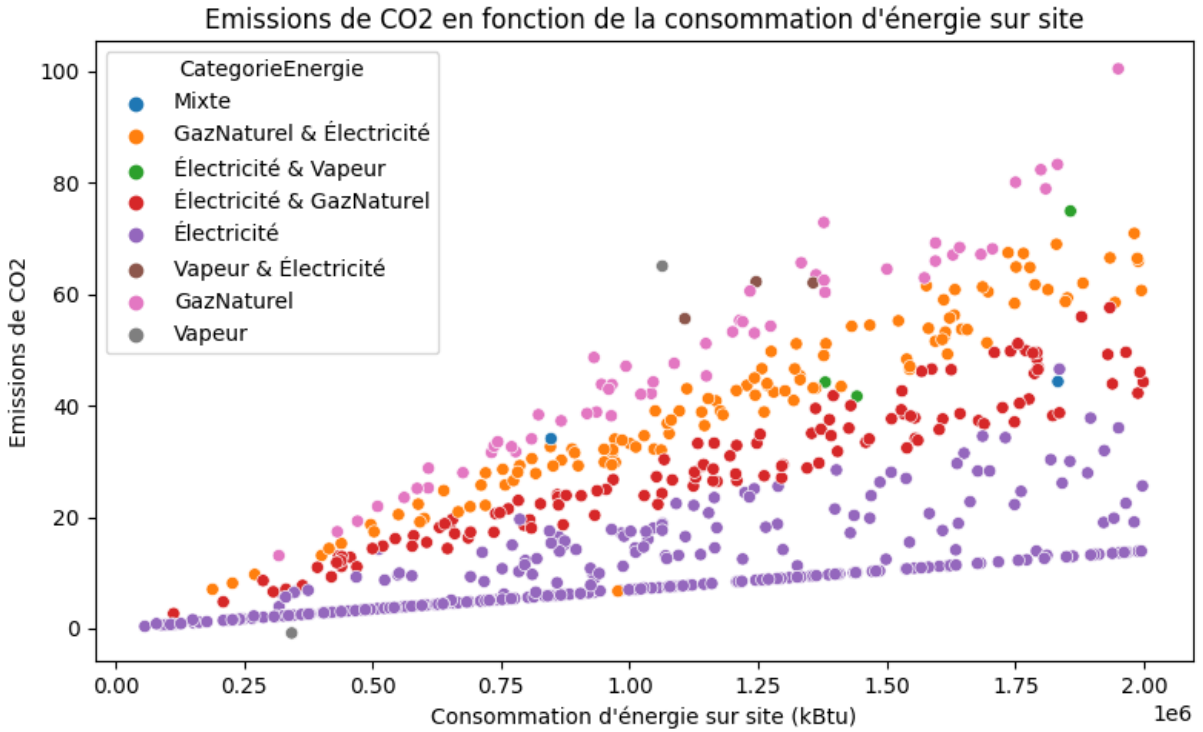


Création de variables

Types d'Énergie :
Identifier l'impact des différentes sources d'énergie.

Méthode de Création :
Calcul de la proportion de chaque type d'énergie par rapport à la consommation totale.





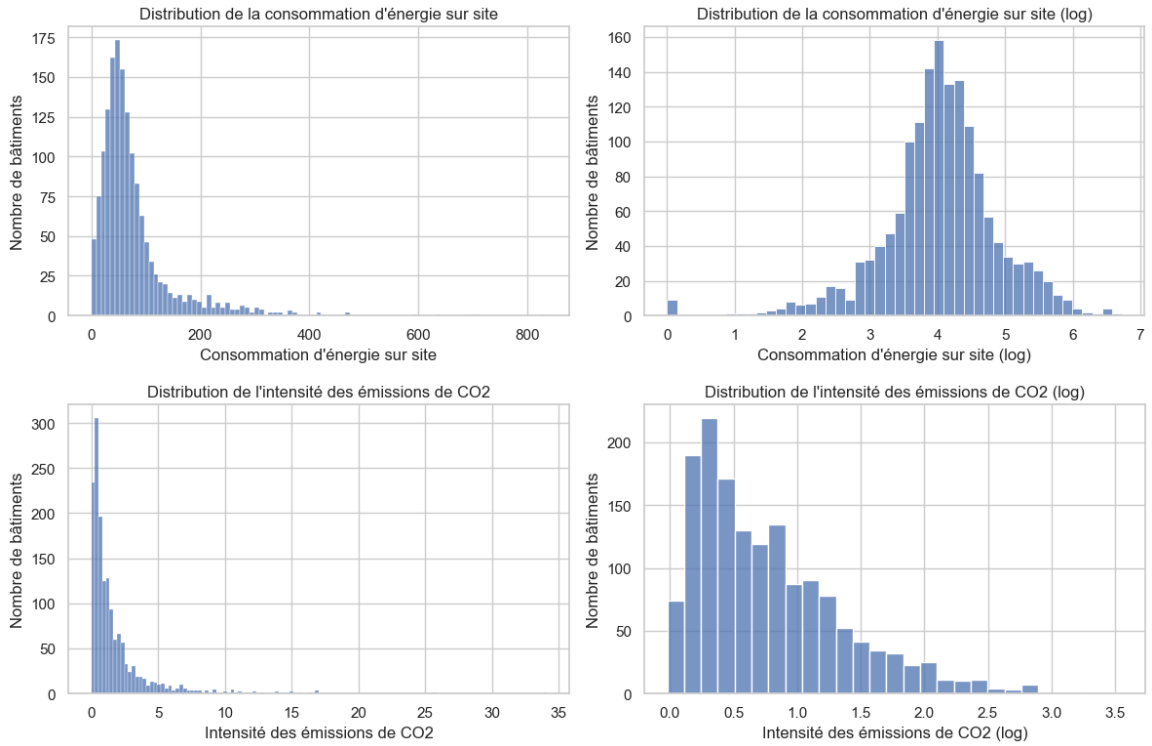
Encodage des variables catégorielles

- **Séparation** avant encodage pour **éviter le data leakage**.
- **Processus d'Encodage OneHot :**
- **Étape 1 : Préparer l'encodeur**
  - OneHotEncoder pour les variables catégorielles.
  - drop='first' pour éviter les variables redondantes.
- **Étape 2 : Appliquer l'encodage**
  - *fit_transform* sur (X_train) pour apprendre les catégories et les transformer.
  - *transform* sur (X_test) pour appliquer les mêmes transformations sans apprendre de nouvelles informations.
- **Étape 3 : Intégration avec les données originales**
  - Alignement des indexes.

Transformation des variables

Objectif :
Améliorer la linéarité et la normalité des variables.
Réduire l'impact des valeurs extrêmes ou des échelles disparates sur le modèle.

- Normalisation de 'ENERGYSTARScore'
- Transformation np.log1p pour ajuster la distribution et gérer les valeurs de 0.
- Retransformation des prédictions des variables cibles (np.expm1) pour les ramener à l'échelle originale pour interpréter les résultats dans un contexte métier significatif.



Méthodologie de l'Optimisation des Hyperparamètres

- Utilisation de **GridSearchCV** pour une exploration systématique de l'espace des hyperparamètres.
- **Validation Croisée** à 4 Folds : Division des données d'entraînement en 4 sous-ensembles pour évaluer la stabilité et la fiabilité des performances.
  - **Robustesse :** minimise le risque de surajustement.
  - **Représentativité :** Chaque observation est utilisée à la fois comme donnée d'entraînement et de validation.
- **R2,** mesure de performance choisie :
  - quantifie la quantité de variance de la variable cible que le modèle est capable d'expliquer.
  - Permet de comparer l'efficacité des différents modèles sur une base comparable et interprétable.
- **MAE :** l'erreur absolu moyen pour évaluer la différence entre valeur réelle et prédite dans l'unité de la variable cible.

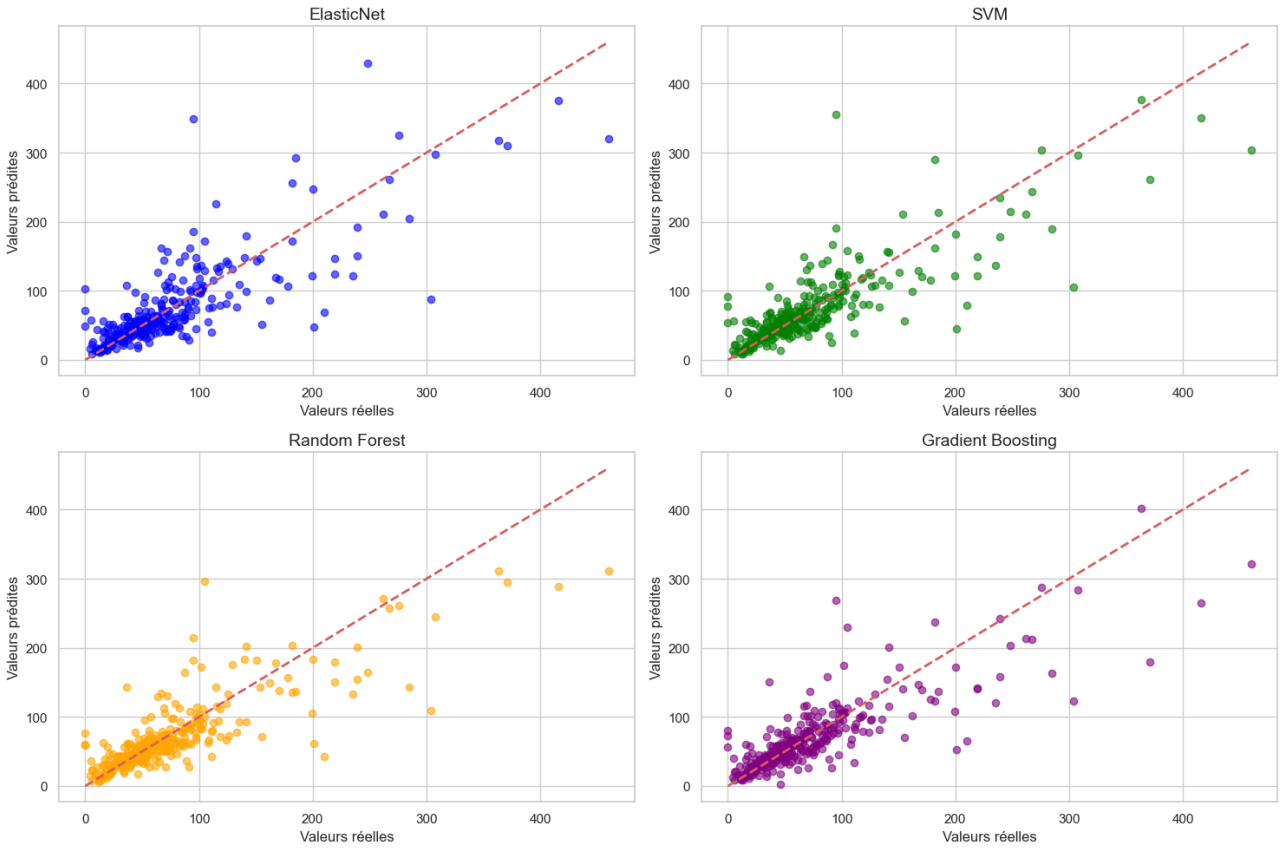
Comparaison des résultats:

- ElasticNet a bien progressé après optimisation.
- SVM et Random Forest ont montré des performances robustes.
- Gradient Boosting se démarque par sa consistance et sa précision.

Modèle	R^2 (Test)	MAE (Test)	R^2 par Fold	Hyperparamètres
ElasticNet	0.63	23.75	0.768, 0.649, 0.673, 0.533	Alpha: 0.00259, L1_ratio: 0.0707
SVM (SVR)	0.684	21.61	0.796, 0.627, 0.718, 0.551 (avec kernel)	C: 10 (sans kernel trick), Degree: 2, Epsilon: 0.01, Gamma: 'scale', Kernel: 'rbf' (avec kernel)
Random Forest	0.693	21.40	0.782, 0.567, 0.703, 0.519	Max Depth: 8, Min Samples Leaf: 6, Min Samples Split: 3, N Estimators: 100
Gradient Boosting	0.684	21.61	0.778, 0.668, 0.712, 0.665	Learning Rate: 0.08, Max Depth: 4, Max Features: 'sqrt', Min Samples Leaf: 1, Min Samples Split: 4, N Estimators: 165

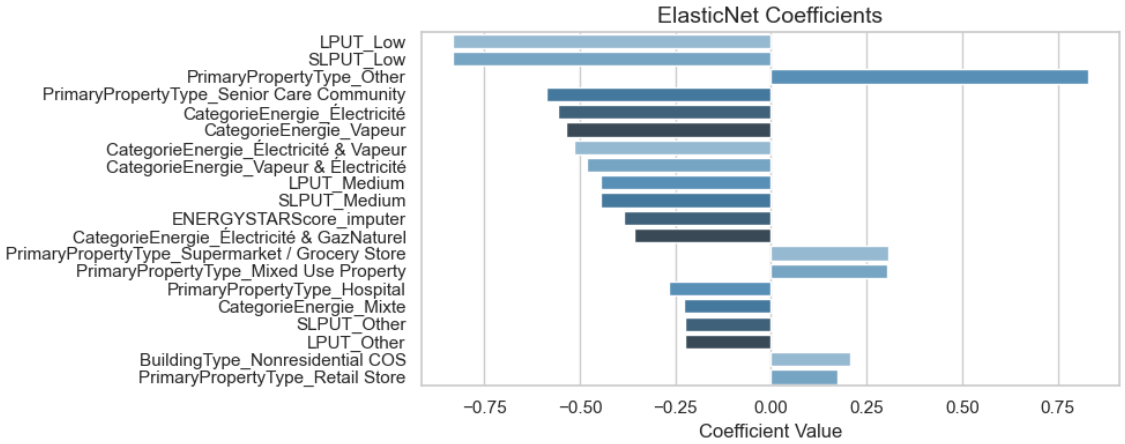
Valeurs réelles vs valeurs prédites

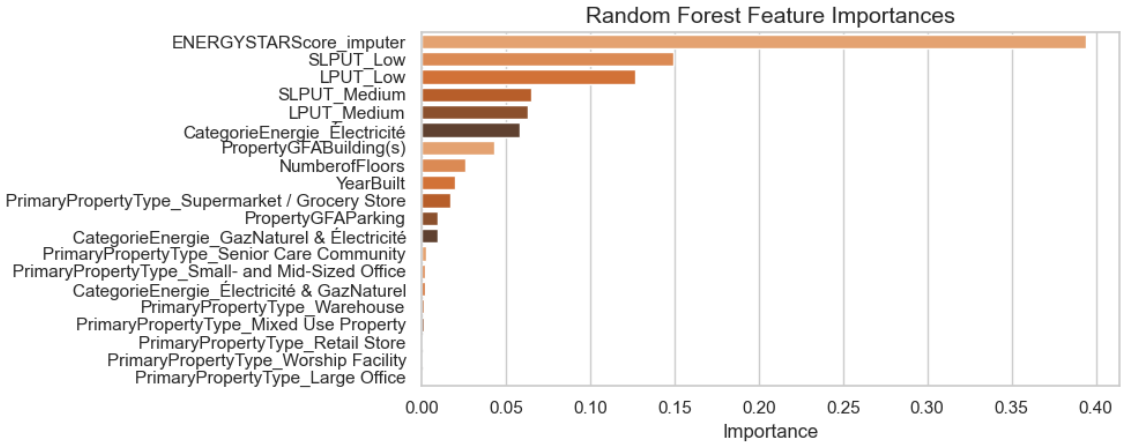
Dispersion plus importante pour les valeurs supérieures à 200

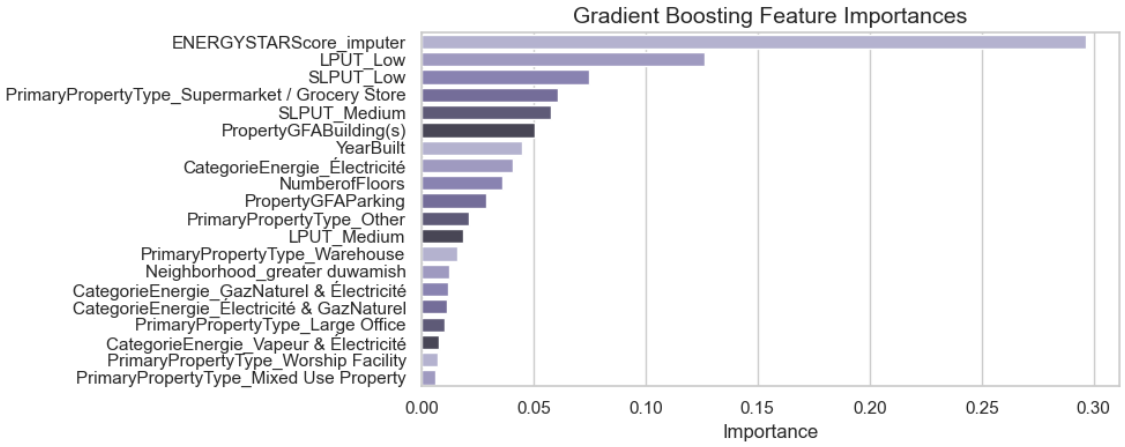


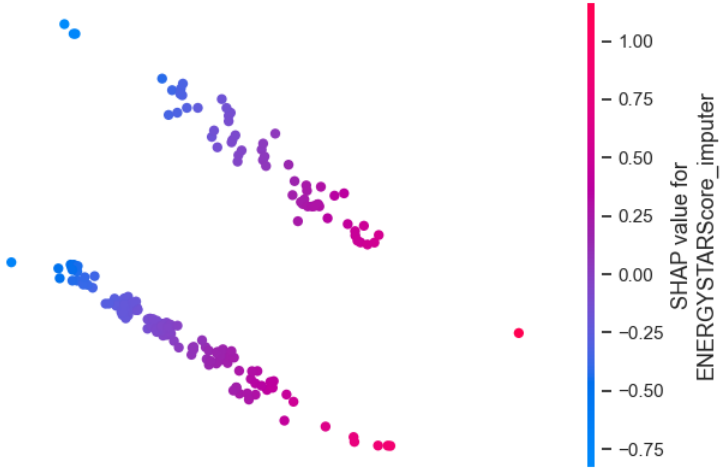
Importances des variables

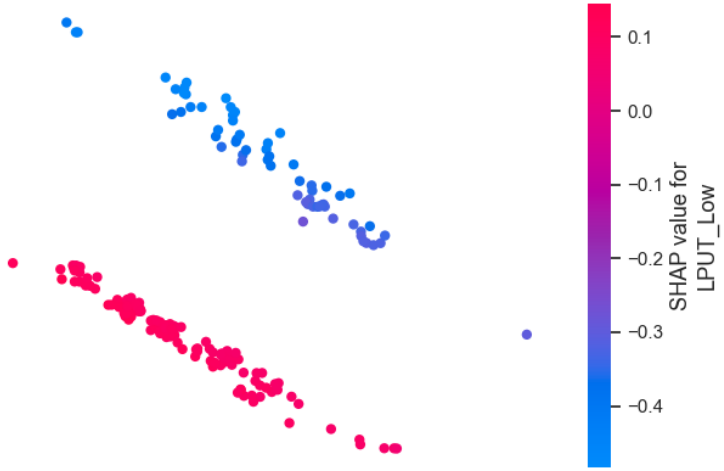
Forte similarité entre les modèles pour la hiérarchie des variables











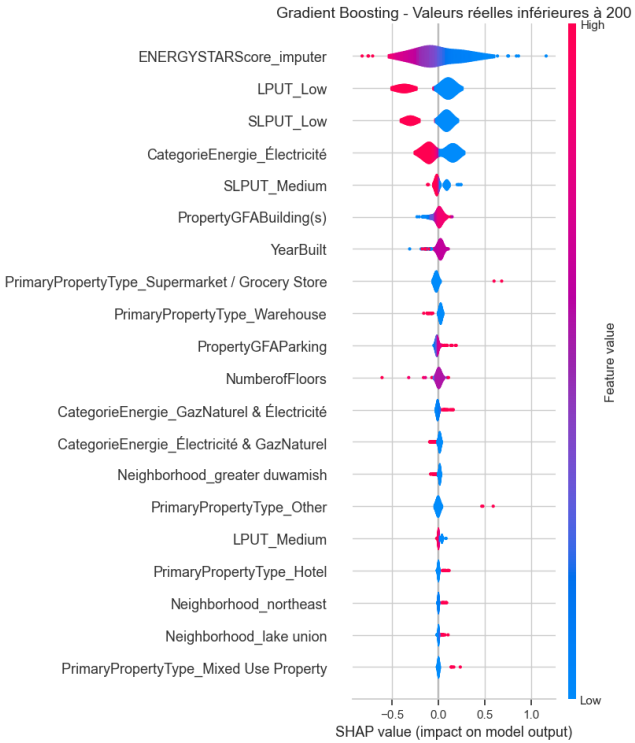
Interprétation des prédictions avec Shap Value

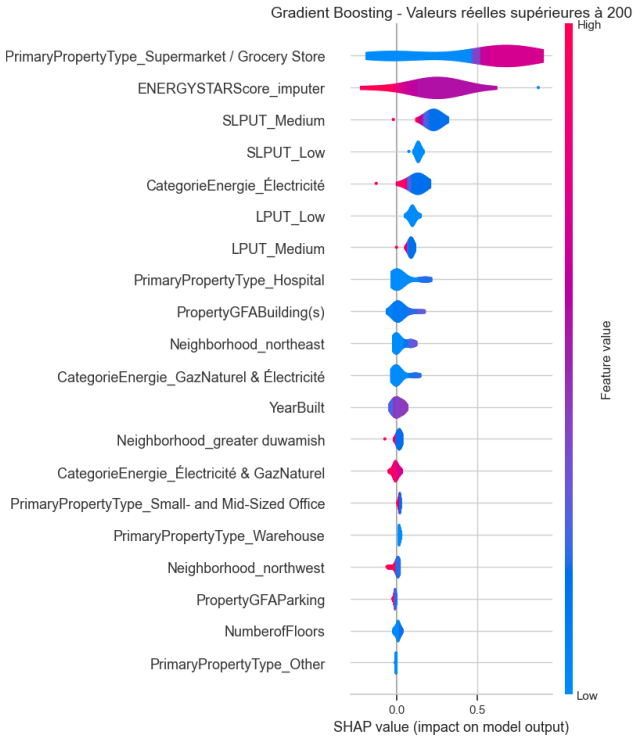
Embedding Plot des variables les plus importante:

- CP1: ENERGYSTARScore
- CP2: Type de bâtiment à consommation basse à élevée

Comparaison de l'importance des variables en fonction des valeurs de la variable cible pour Gradient Boosting

Changement important pour les bâtiments à forte consommation



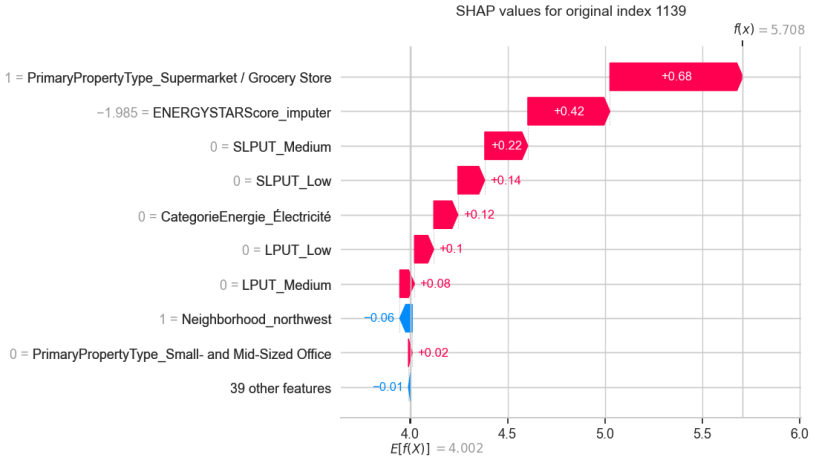


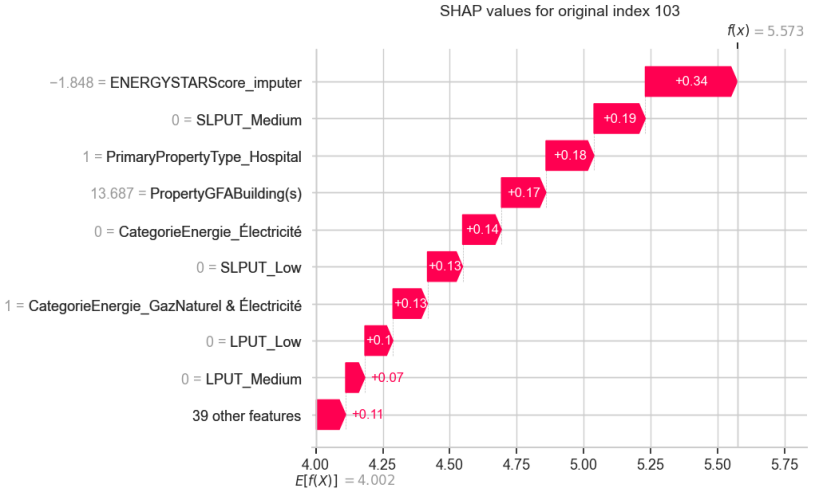
Interprétation des prédiction du Gradient Boosting pour des instances remarquable

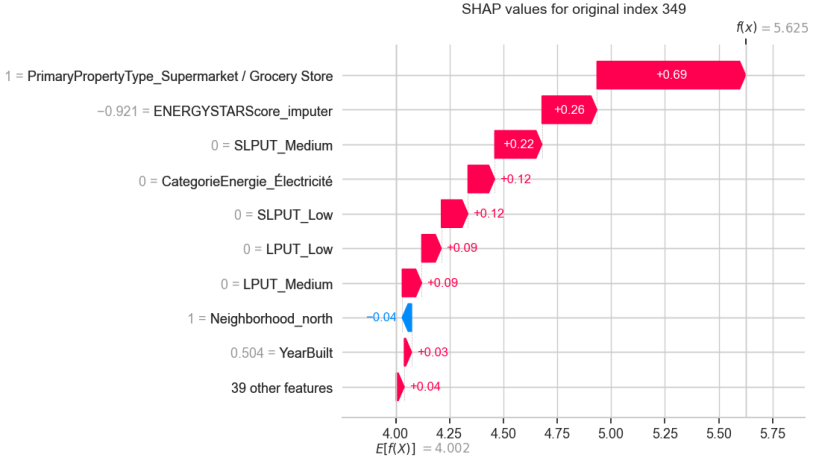
Index 1139: prédiction largement supérieure à la valeur réelle

Index 103: prédiction largement inférieure à la valeur réelle

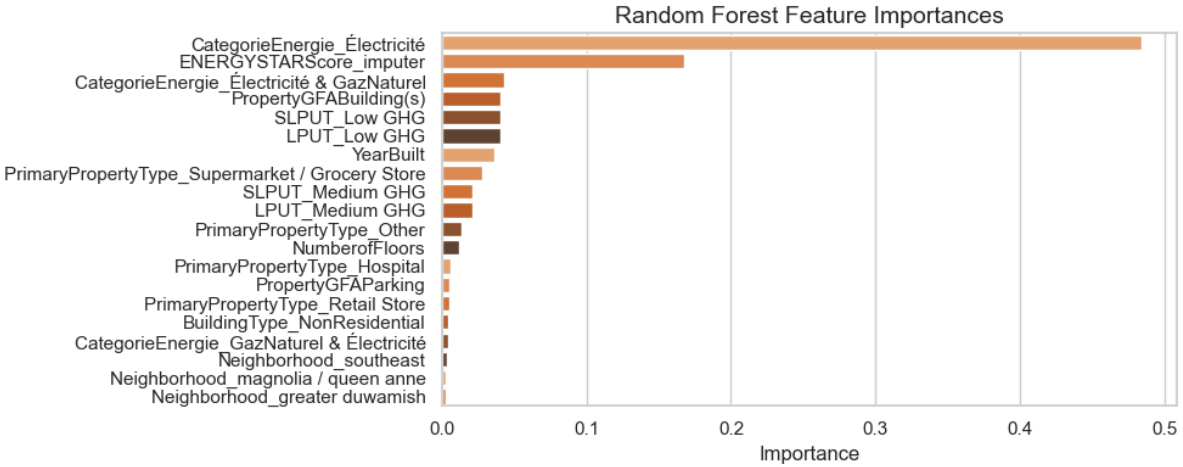
Index 349: prédiction proche de la valeur réelle pour un bâtiment à consommation élevée

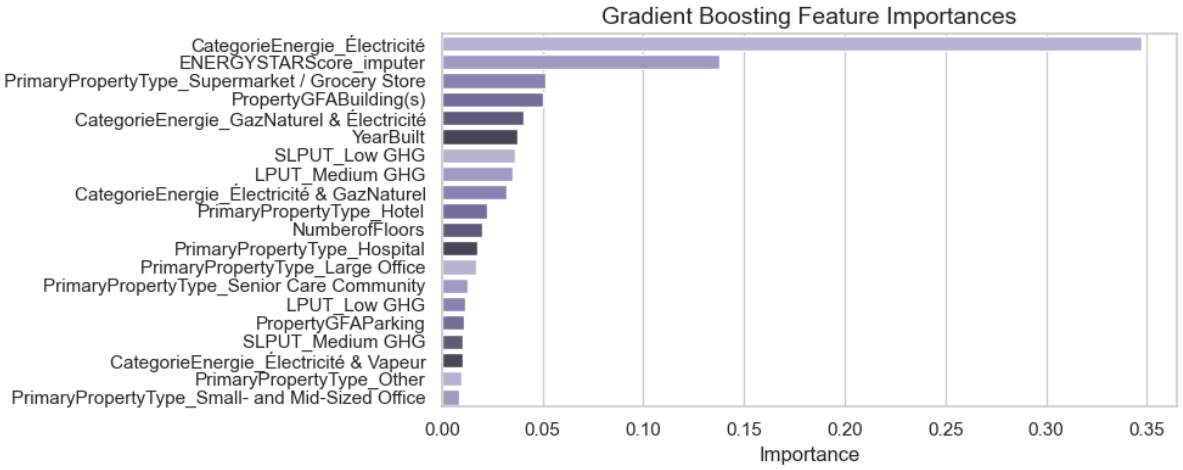






Résultats des prédictions de CO_2

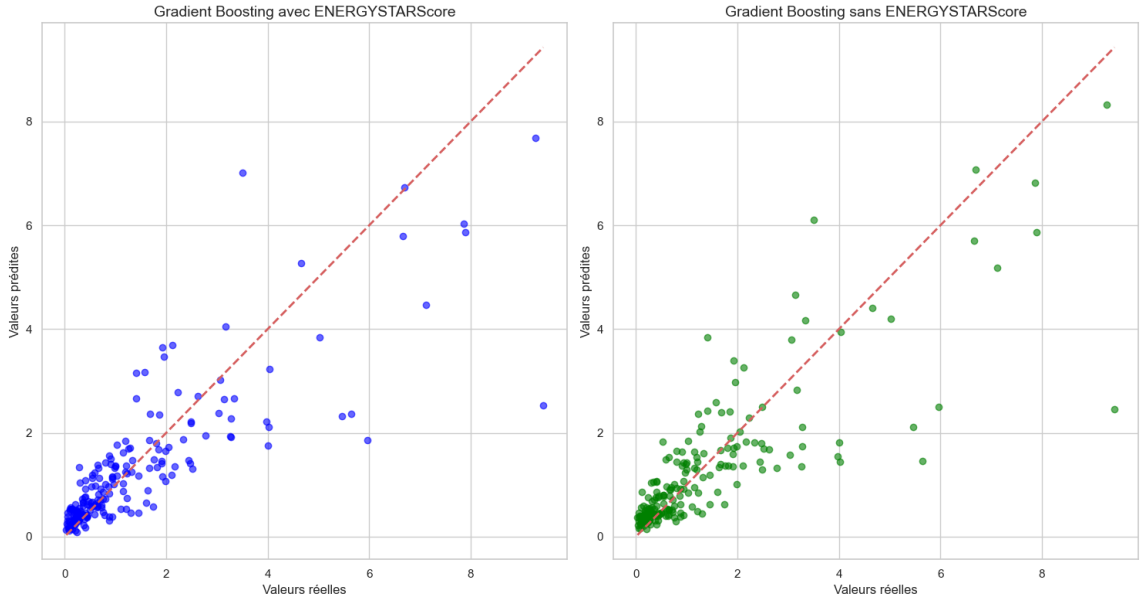




Performance post-optimisation pour le Gradient Boosting

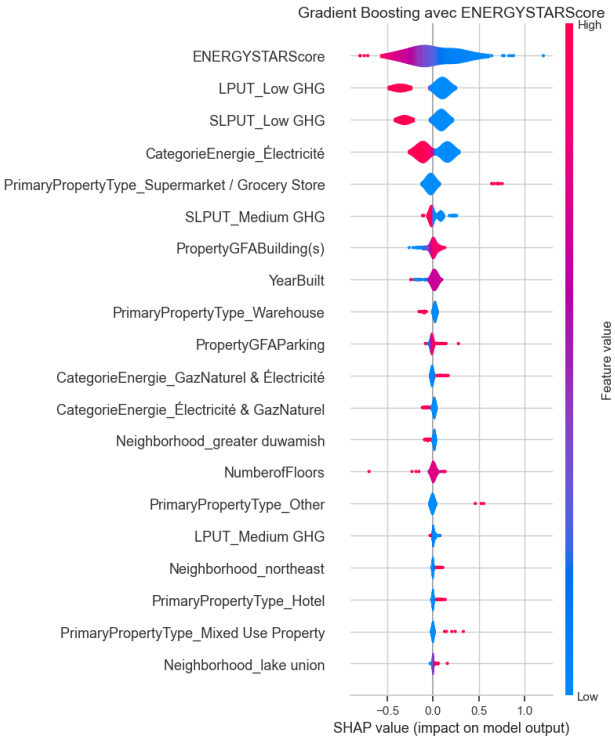
- R^2 pour chaque fold : 0.7501, 0.8118, 0.6583, 0.7519
- R^2 sur l'ensemble de test : 0.727
- *MAE* sur l'ensemble de test : 0.501

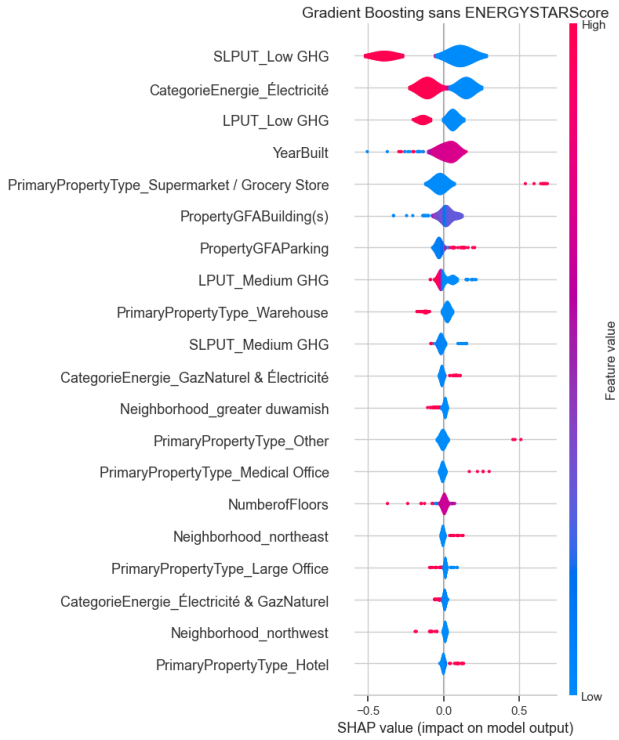
Importance de la variable ENERGYSTARScore pour la prédiction d'émission de CO_2



Critère	Gradient Boosting avec ENERGYSTARScore	Gradient Boosting sans ENERGYSTARScore
Meilleurs Paramètres	learning_rate: 0.05, max_depth: None, max_features: 'log2', min_samples_leaf: 2, min_samples_split: 5, n_estimators: 120	learning_rate: 0.09, max_depth: 5, max_features: 'sqrt', min_samples_leaf: 2, min_samples_split: 6, n_estimators: 120
Score R^2 pour chaque Fold	0.7659, 0.6660, 0.7492	0.6897, 0.5694, 0.7034
Score R^2 (optimisé)	0.6881	0.6812
MAE (optimisé)	0.5345	0.5496

Importance de la variable ENERGYSTARScore pour la prédiction d'émission de CO_2





Conclusion

- **Modèles Testés** : Gradient Boosting se démarque pour sa précision en consommation et émissions de CO_2.
- **Variable Clé** : La catégorie d'énergie est essentielle pour prédire les émissions de CO_2.
- **Variabilité des Prédictions :** Importance des variables varie entre bâtiments à forte et faible consommation.
- **ENERGYSTARScore :** Sa non-utilisation est envisageable, avec un léger compromis sur le R^2.