Projet 3 : Anticipez les besoins en consommation de bâtiments

02/07/2024

Soukaina GUAOUA ELJADDI

Parcours Data Scientist
OpenClassrooms

Plan:

- ❏ Problématique et présentation du jeu de données
- ❏ Traitement, exploration des données et feature engineering
- ❏ Modélisation avec testing des hyperparamètres
- ❏ Évaluation des performances et choix du modèle final
- ❏ Analyse de la "feature importance" globale et locale
- ❏ Analyse de l'influence de l'EnergyStarScore
- ❏ Conclusion

Problématique et présentation du jeu de données

Problématique

Seattle

Contexte : La **ville de Seattle** vise la **neutralité carbone** en **2050**.

Objectif : Réduire les **émissions de CO2** et la **consommation énergétique** des bâtiments non résidentiels.

Missions : Prédire les **émissions de CO2** et la **consommation d'énergie** des bâtiments **non destinés à l'habitation** en utilisant leurs données structurales.

Évaluer l'intérêt de **l'ENERGY STAR Score** dans ces prédictions.

Présentation du jeu de données

2016_Building_Energy_Benchmarking.csv :

- Données créés le 15 mars 2018 par Seattle sur des relevés de **2016**
- **3376** observations et **46** variables décrivant les propriétés (géographiques, architecturales, usage, consommations, émissions)

GÉOGRAPHIQUES	ARCHITECTURAUX	USAGE	EMISSIONS/CONSO
• Longitude • Latitude • Code de District (District Code) • Quartier (Neighborhood)	• Nombre de bâtiments • Nombre d'étages • Année de construction • Surface totale (propertyGFATotal)	• Type de propriété principale	•TotalGHGEmmissions • SiteEnergyUse

Présentation du jeu de données

2 Variables Cibles:

- **TotalGHGEmissions** : total des émissions de gaz à effet de serre (GES)
- **SiteEnergyUse(kBtu)** : quantité totale d'énergie consommée annuellement par bâtiment.

TotalGHGEmissions	SiteEnergyUse(kBtu)
- Dioxyde de carbone (CO2) - Méthane (CH4) - Protoxyde d'azote (N2O) - en tonnes métriques d'équivalent dioxyde de carbone (CO2e)	- toutes sources d'énergie - en kilo-British Thermal Units (kBtu)

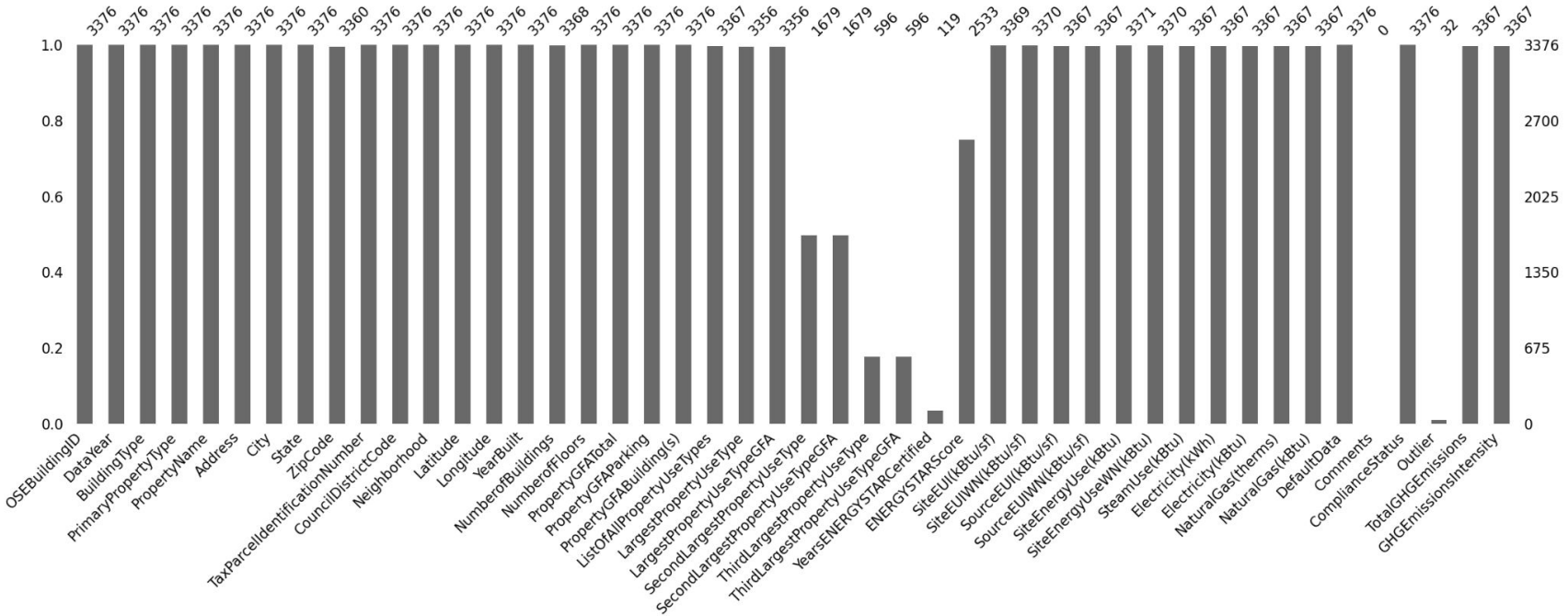
Présentation du jeu de données

Feature à considérer (ou non) : ENERGYSTARScore

- Mesure la performance énergétique globale d'un bâtiment,
- Calculée par l'Agence de Protection de l'Environnement (EPA) des États-Unis,
- Score : 1 à 100
- Calculé selon ces facteurs : - Le climat - L'usage du bâtiment - Les opérations
- Fastidieux à calculer ⇒ l'Intégrer dans la modélisation et analyser sa pertinence pour la prédiction des émissions de CO2 et de la consommation énergétique

Partie 1 : Traitement, exploration des données et feature engineering

Partie 1 : Traitement, exploration des données et feature engineering



2016_Building_Energy_Benchmarking.csv : 3376 Lignes, 46 Colonnes

Suppression de la colonne vide “Comments” ⇒ 3376 Lignes, 45 Colonnes

Partie 1 : Traitement, exploration des données et feature engineering

Suppression des colonnes possédant une valeur unique:

Colonne	Valeur unique
DataYear	2016
City	Seattle
State	WA

3376 Lignes, 45 Colonnes ⇒ 3376 Lignes, 42 Colonnes

Partie 1 : Traitement, exploration des données et feature engineering

Récupération des bâtiments **non destinés à l'habitation** :
'NonResidential', 'Nonresidential COS', 'SPS-District K-12',
'Campus', 'Nonresidential WA'

Suppression des bâtiments destinés à l'habitation : 'Multifamily MR (5-9)', 'Multifamily LR (1-4)', 'Multifamily HR (10+)'

3376 Lignes, 42 Colonnes ⇒ 1668 Lignes × 42 Colonnes

Partie 1 : Traitement, exploration des données et feature engineering

- Suppression des lignes ne possédant pas de valeurs pour la variable 'ENERGYSTARScore'

  1668 Lignes × 42 Colonnes ⇒ 1094 Lignes × 42 Colonnes

- Suppression des colonnes 'PropertyName', 'OSEBuildingID' et 'TaxParcelIdentificationNumber'

  1094 Lignes × 42 Colonnes ⇒ 1094 lignes × 39 colonnes

Partie 1 : Traitement, exploration des données et feature engineering

- Suppression des types propriétés “PrimaryPropertyType” et “ListOfAllPropertyUseTypes” déjà remplacés par “LargestPropertyUseType”, “SecondLargestPropertyUseType” et “ThirdLargestPropertyUseType”

  1094 lignes × 39 colonnes ⇒ 1094 lignes × 37 colonnes

- Suppression des variables métriques géométriques inutiles pour l'analyse "Address", "ZipCode" et "CouncilDistrictCode"

  1094 lignes × 37 colonnes ⇒ 1094 lignes × 34 colonnes

Partie 1 : Traitement, exploration des données et feature engineering

- Suppression des lignes avec des valeurs par défaut 'DefaultData'

  1094 lignes × 34 colonnes ⇒ 1006 lignes × 34 colonnes

- Suppression de la colonne unique 'DefaultData'

  1006 lignes × 34 colonnes ⇒ 1006 lignes × 33 colonnes

Partie 1 : Traitement, exploration des données et feature engineering

- Suppression de la colonne 'NaturalGas(therms)' redondante avec ‘NaturalGas(kBtu)’

  1006 lignes × 33 colonnes ⇒ 1006 lignes × 32 colonnes

- Suppression de la colonne 'Electricity(kWh)' redondante avec 'Electricity(kBtu)'

  1006 lignes × 32 colonnes ⇒ 1006 lignes × 31 colonnes

Partie 1 : Traitement, exploration des données et feature engineering

- Suppression des colonnes ‘SiteEUI(kBtu/sf’, ‘SiteEUIWN(kBtu/sf)’, ‘SourceEUI(kBtu/sf)’ et ‘SourceEUIWN(kBtu/sf)’, car notre cible est 'SiteEnergyUse(kBtu)'

  1006 lignes × 31 colonnes ⇒ 1006 lignes × 27 colonnes

- Suppression de la colonne 'SiteEnergyUseWN(kBtu)' redondante avec 'SiteEnergyUse(kBtu)'

  1006 lignes × 27 colonnes ⇒ 1006 lignes × 26 colonnes

Partie 1 : Traitement, exploration des données et feature engineering

Suppression de la colonne 'GHGEmissionsIntensity' car notre cible est 'TotalGHGEmissions'

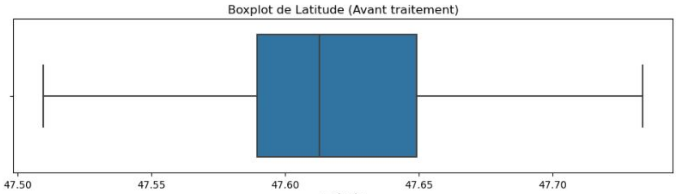
1006 lignes × 26 colonnes ⇒ 1006 lignes × 25 colonnes

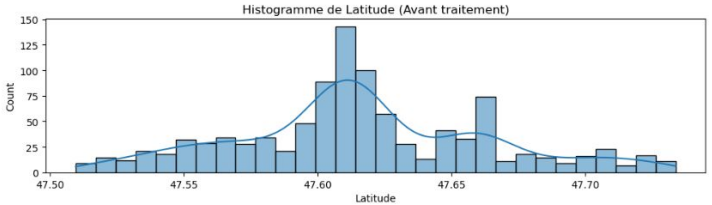
Suppression de la colonne 'YearsENERGYSTARCertified'

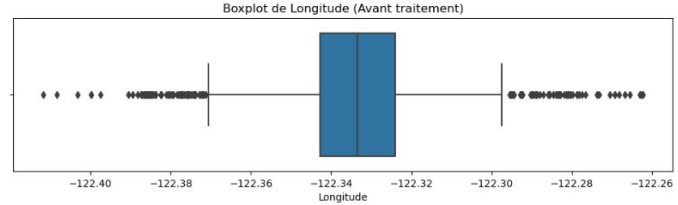
1006 lignes × 25 colonnes ⇒ 1006 lignes × 24 colonnes

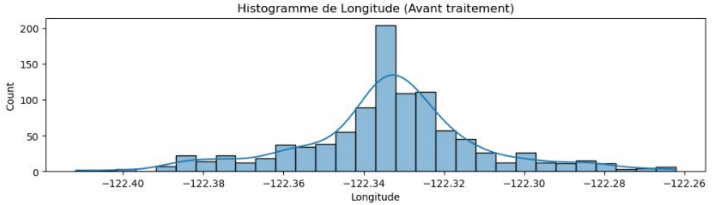
Partie 1 : Traitement, exploration des données et feature engineering

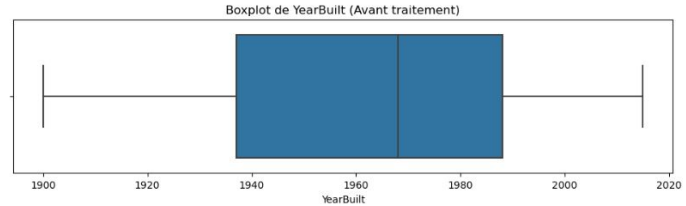
Identification et traitement des valeurs aberrantes :

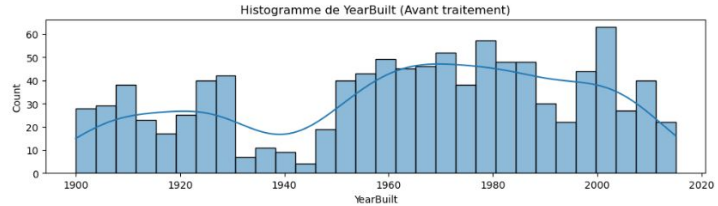












Partie 1 : Traitement, exploration des données et feature engineering

Identification et traitement des valeurs aberrantes :

- Suppression des valeurs aberrantes basées sur la colonne "Outlier"
- Suppression des bâtiments non conformes si "ComplianceStatus" indique non-conformité

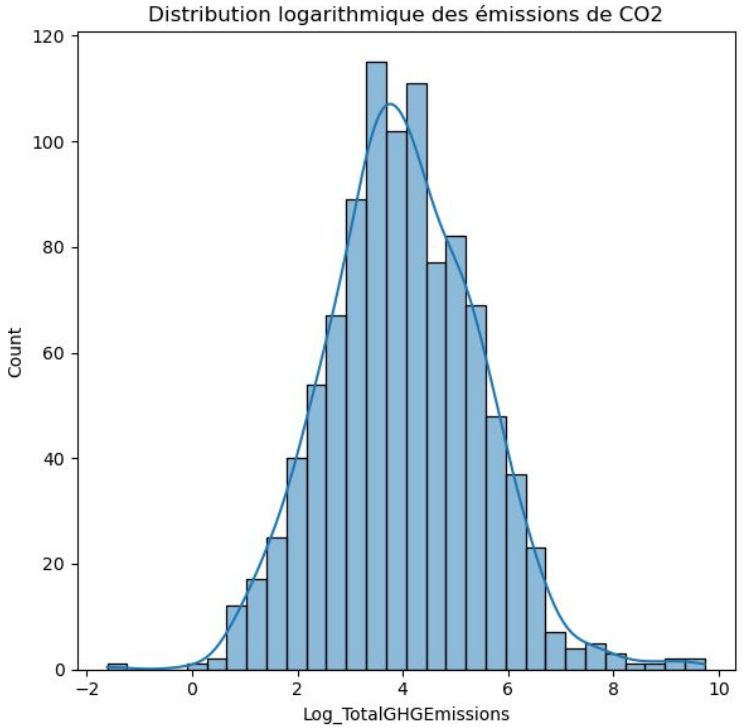
1006 lignes × 24 colonnes ⇒ 997 lignes × 24 colonnes

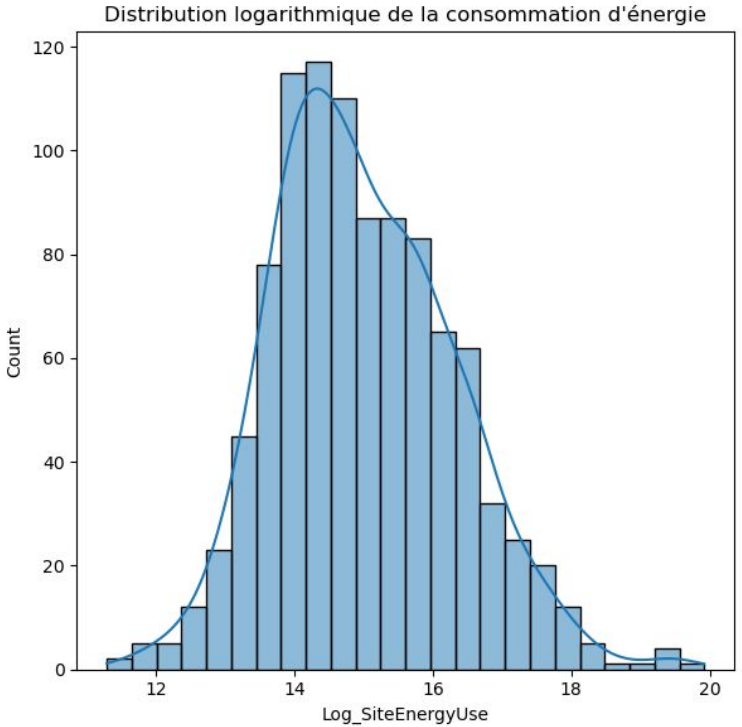
- Suppression des colonnes 'ComplianceStatus' et 'Outlier' avec des valeurs uniques ‘Compliant’ et ‘NaN’

997 lignes × 24 colonnes ⇒ 997 lignes × 22 colonnes

Partie 1 : Traitement, exploration des données et feature engineering

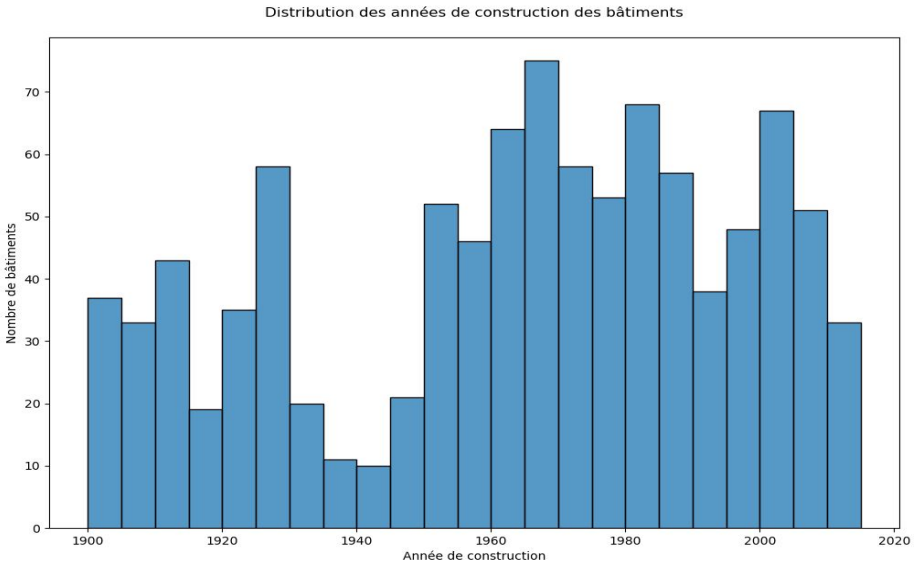
Visualisation des distributions logarithmiques des variables cibles

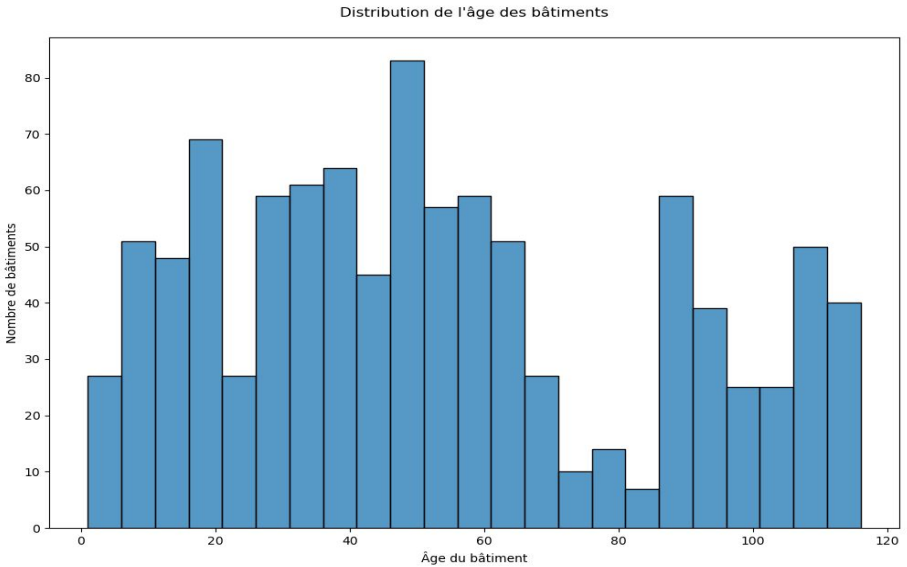




Partie 1 : Traitement, exploration des données et feature engineering

feature engineering: Année de construction (YearBuilt) ⇒ Âge du bâtiment (BuildingAge)



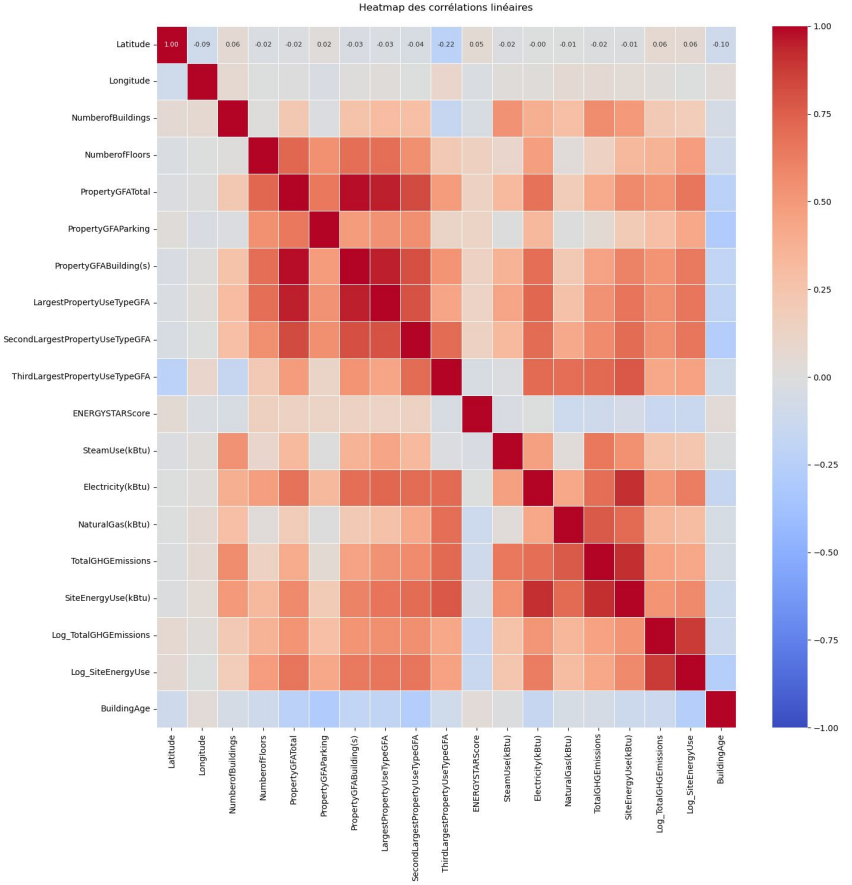


Partie 1 : Traitement, exploration des données et feature engineering

Suppression des variables avec degré de corrélation > 0.7 :

'LargestPropertyUseTypeGFA', 'SecondLargestPropertyUseType', 'SecondLargestPropertyUseTypeGFA', 'ThirdLargestPropertyUseType', 'ThirdLargestPropertyUseTypeGFA', 'PropertyGFAParking', 'PropertyGFABuilding(s)'

⇒ 997 lignes × 17 colonnes



Partie 1 : Traitement, exploration des données et feature engineering

Suppression des variables à forte corrélation avec les variables cibles et qui peuvent causer

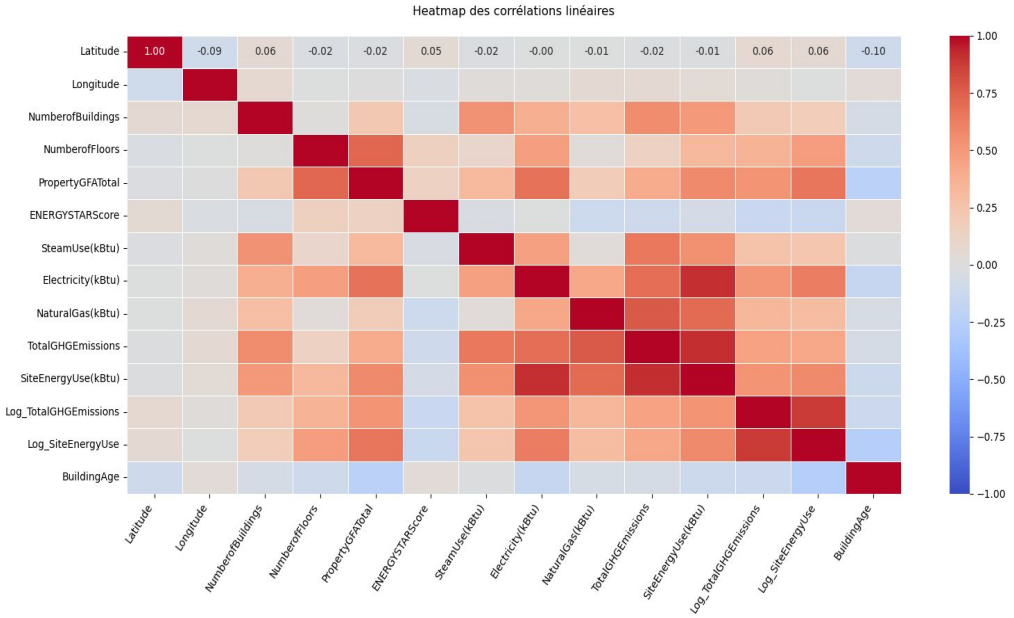
un problème de data-leakage:

SteamUse(kBtu),

Electricity(kBtu) et

NaturalGas(therms)

⇒ 997 lignes × 14 colonnes



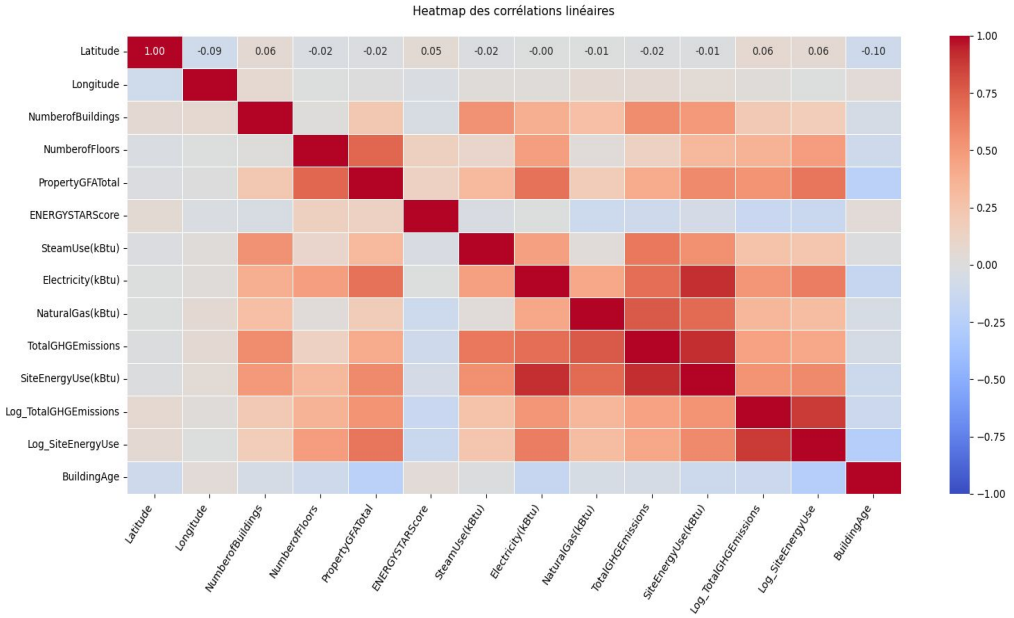
Partie 1 : Traitement, exploration des données et feature engineering

- Suppression des variables non corrélés avec les variables cibles:

➔ 'Latitude',
➔ 'Longitude' et
➔ 'BuildingAge'

- Suppression des lignes avec des valeurs manquantes pour 'LargestPropertyUseType'

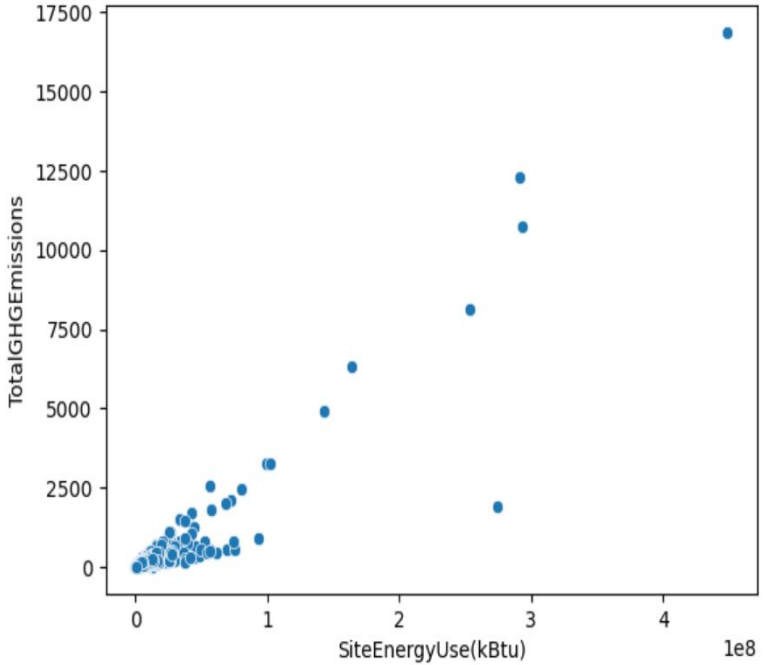
⇒ 995 lignes × 11 colonnes



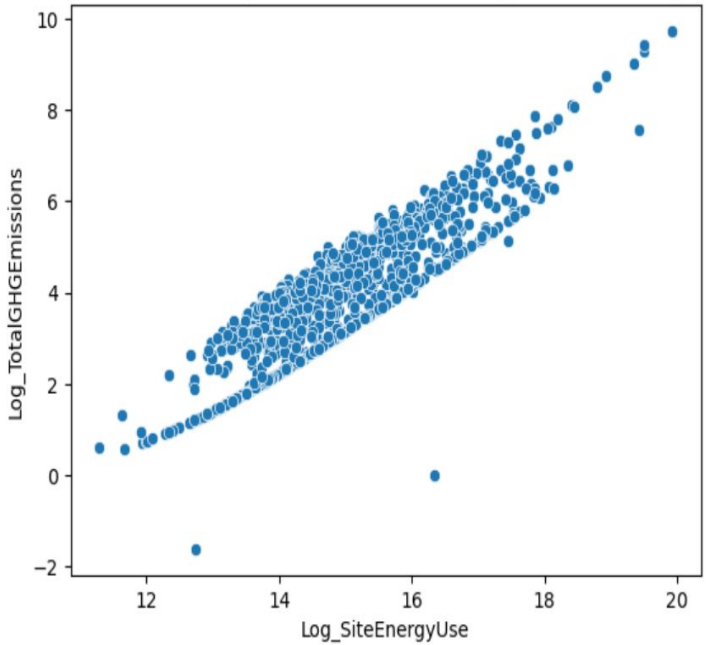
Partie 1 : Traitement, exploration des données et feature engineering

Analyse bivariée: corrélation entre les 2 variables cibles

Émissions de CO2 en fonction de la quantité annuelle d'énergie consommée



Émissions de CO2 en fonction de la quantité annuelle d'énergie consommée par la propriété, en log



Partie 1 : Traitement, exploration des données et feature engineering

Transformation des variables catégorielles en variables factices ou indicatrices:

'BuildingType',

'Neighborhood',

'LargestPropertyUseType'

⇒ 995 lignes × 48 colonnes

Partie 2 : Modélisation avec testing des hyperparamètres

Partie 2 : Modélisation avec testing des hyperparamètres

Développement et simulation d'un 1er modèle

Développement des modèles de **régression linéaire** pour les logarithmes des variables cibles :

1. Définition des caractéristiques (features) et des cibles (targets)

2. Création des modèles de régression linéaire

3. Validation croisée

Explication des mesures de performance (R^2)

Le coefficient de détermination R^2 indique la proportion de la variance des variables cibles qui est expliquée par les caractéristiques. Un R^2 de 1 signifie que le modèle explique parfaitement la variance des données, tandis qu'un R^2 de 0 indique que le modèle ne fait pas mieux qu'une prédiction moyenne.

Partie 2 : Modélisation avec testing des hyperparamètres

Résultats des modèles de régression linéaire

- Pour la prédiction du logarithme des émissions de CO2 :

⇒ R^2 scores : [0.16588005 0.09020311 0.2037925 0.20585722 0.36608608]

⇒ R^2 moyen : 0.2063637937661316 2

- Pour la prédiction du logarithme de la consommation d'énergie :

⇒ R^2 scores : [0.38493455 0.27213622 0.30419405 0.29174352 0.56730451]

⇒ R^2 moyen : 0.3640625688553746

Partie 2 : Modélisation avec testing des hyperparamètres

Évaluation du modèle

1. Division des données en ensembles d'entraînement 75% et de test 25%

2. Ajustement des modèles sur les données d'entraînement

3. Prédictions sur les données de test

4. Calcul des métriques de performance

- **MAE (Mean Absolute Error)** : Erreur absolue moyenne.
- **RMSE (Root Mean Squared Error)** : Erreur quadratique moyenne racine.
- **R^2 (Coefficient de détermination)** : Mesure la proportion de la variance expliquée par le modèle.

Partie 2 : Modélisation avec testing des hyperparamètres

Résultats d'Évaluation du modèle des émissions de CO2:

	émissions de CO2
MAE	0.7737
RMSE	0.9617
R^2	0.5348

MAE à 0.7737 signifie qu'en moyenne, les prédictions de logarithme des émissions de CO2 s'écartent des valeurs réelles de 0.7737 unités.

RMSE à 0.9618, les erreurs de prédiction plus élevée que le MAE, suggère qu'il y a quelques grandes erreurs de prédiction qui augmentent la moyenne quadratique.

R^2 à 0.5349 indique que le modèle explique environ 53.5% de la variance des données de logarithme des émissions de CO2. Près de la moitié de la variance des émissions de CO2 n'est pas expliquée par le modèle.

Résultats d'Évaluation du modèle de la consommation d'énergie :

	consommation d'énergie
MAE	0.5966
RMSE	0.7208
R^2	0.6880

MAE à 0.5967 signifie que Les prédictions de logarithme de la consommation d'énergie sont, en moyenne, à 0.5967 unités des valeurs réelles.

RMSE à 0.7208 indique que les erreurs de prédiction sont de l'ordre de 0.7208 unités. la RMSE plus élevée que le MAE, il y a quelques grandes erreurs de prédiction.

R^2 à 0.6881 indique que le modèle explique environ 68.8% de la variance des données de logarithme de la consommation d'énergie. Cela montre une meilleure performance par rapport au modèle des émissions de CO2, mais il y a encore environ 31.2% de la variance non expliquée par le modèle.

Partie 2 : Modélisation avec testing des hyperparamètres

Amélioration du Feature Engineering

1. Standardisation/Normalisation des caractéristiques
2. Réduction de dimension avec PCA
3. Entraînement et évaluation du modèle avec les nouvelles caractéristiques

Amélioration du Feature Engineering

Résultats de l'ACP:

Variance expliquée par chaque composante principale: [0.33222696 0.19148344 0.1580441 0.05219617 0.04286809 0.03420705 0.02539937 0.01719506 0.01553708 0.01373998 0.01329913 0.01214451 0.01069186 0.01022748 0.00869696 0.0074915 0.00635766]

Nombre de composantes principales: **17**

⇒ Les trois premières composantes principales expliquent ensemble environ 68.5% de la variance totale des données.

⇒ 17 composantes principales ont été retenues pour expliquer 95% de la variance totale des données.

Partie 2 : Modélisation avec testing des hyperparamètres

Amélioration du Feature Engineering

Entraînement et évaluation du modèle avec les nouvelles caractéristiques

	émissions de CO2		consommation d'énergie	
	avant ACP	après ACP	avant ACP	après ACP
MAE	0.7737	0.8032	0.5966	0.6363
RMSE	0.9617	1.0140	0.7208	0.8100
R^2	0.5348	0.4829	0.6880	0.6060

Après application de l'**ACP**, les erreurs MAE et RMSE ont augmentés et la capacité explicative R^2 a diminué, pour les émissions de CO2 et la consommation d'énergie.

⇒ L'application de l'ACP pour réduire la dimensionnalité n'a pas amélioré les performances des modèles de régression linéaire pour les deux cibles.

Partie 3 : Évaluation des performances et choix du modèle final

Partie 3 : Évaluation des performances et choix du modèle final

Simulation d'autres modèles et choix d'un modèle final : Test des modèles :

Régression linéaire :

- Modèle qui prédit une variable cible en trouvant une relation linéaire entre cette variable et des variables prédictives.

Régression Ridge :

- Variante de la régression linéaire qui ajoute une pénalité pour la taille des coefficients pour éviter le surapprentissage.

Partie 3 : Évaluation des performances et choix du modèle final

Simulation d'autres modèles et choix d'un modèle final : Test des modèles :

Régression Lasso :

- Variante de la régression linéaire qui ajoute une pénalité pour la taille des coefficients, pouvant supprimer certains d'entre eux pour la sélection de variables.

Random Forest Regressor :

- Modèle d'ensemble utilisant plusieurs arbres de décision pour améliorer la précision des prédictions.

Partie 3 : Évaluation des performances et choix du modèle final

Simulation d'autres modèles et choix d'un modèle final : Test des modèles :

Gradient Boosting Regressor :

- Modèle d'ensemble qui construit des arbres de décision séquentiellement pour corriger les erreurs des arbres précédents, améliorant ainsi la précision.

Partie 3 : Évaluation des performances et choix du modèle fina

	émissions de CO2		consommation d'énergie	
	RMSE	R^2	RMSE	R^2
linéaire	1.1229	0.2064	0.8452	0.3641
Ridge	1.1044	0.2318	0.8346	0.3785
Lasso	1.5064	-0.4505	1.3881	-0.7235
Random Forest	0.9509	0.4345	0.5438	0.7431
Gradient Boosting	0.9068	0.4850	0.5010	0.7790

⇒ Le Gradient Boosting est le meilleur modèle avec le RMSE le plus faible et le R^2 le plus élevé, pour les 2 cibles émissions de CO2 et consommation d'énergie.

Partie 3 : Évaluation des performances et choix du modèle final

Optimisation des hyperparamètres avec du Gradient Boosting Regressor : RandomizedSearchCV.

	n_estimators	min_samples_split	min_samples_leaf	max_depth	learning_rate
émissions de CO2 / consommation d'énergie	**200**	**5**	**1**	**3**	**0.05**

Les hyperparamètres optimaux identifiés indiquent une préférence pour un modèle de Gradient Boosting avec une complexité modérée (200 estimateurs, profondeur maximale de 3), un taux d'apprentissage prudent (0.05), et une régularisation suffisante (min_samples_split de 5 et min_samples_leaf de 1).

Partie 3 : Évaluation des performances et choix du modèle final

Évaluation des modèles optimisés : validation croisée

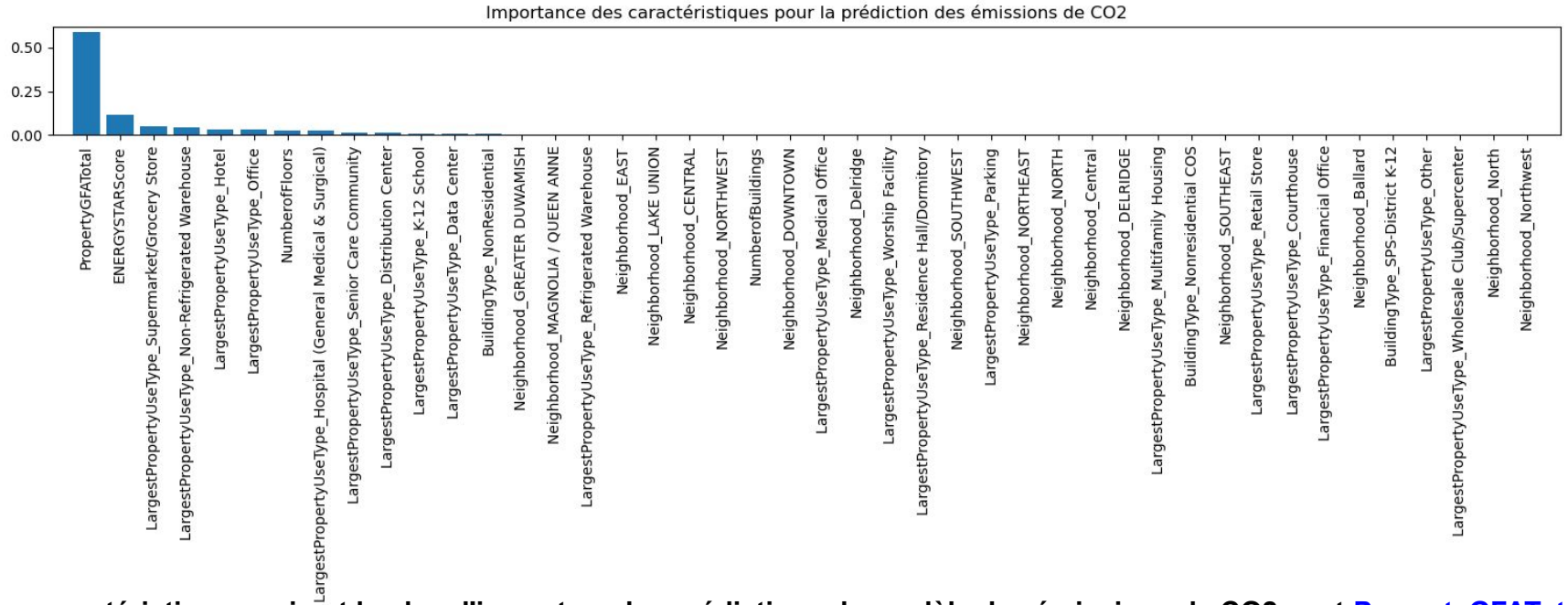
	émissions de CO2	consommation d'énergie
RMSE	0.9056	0.4943
R^2	0.4858	0.7843

⇒ Le modèle de prédiction de la consommation d'énergie fonctionne bien et explique une grande partie de la variance des données,
⇒ Le modèle de prédiction des émissions de CO2 a une performance modérée

Partie 4 : Analyse de la "feature importance" globale et locale

Partie 4 : Analyse de la "feature importance" globale et locale

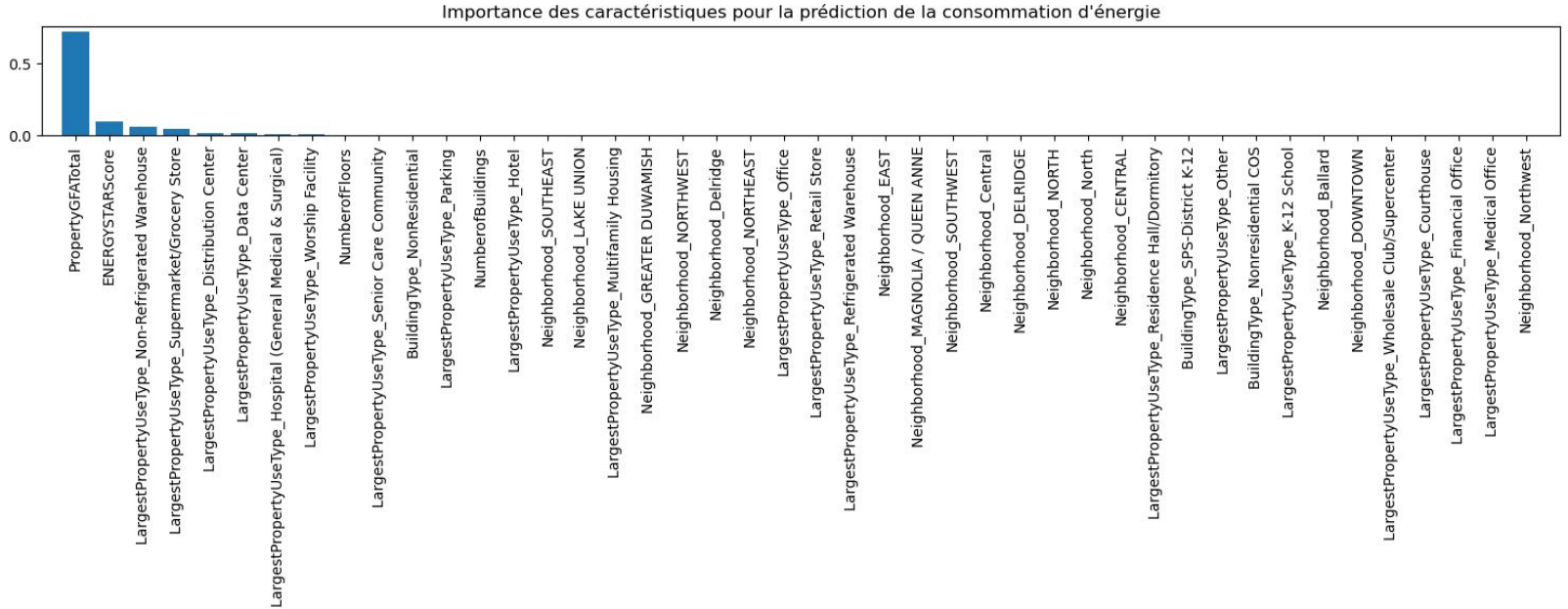
Importance des caractéristiques globale



Les caractéristiques qui ont le plus d'impact sur les prédictions du modèle des émissions de CO2 sont PropertyGFATotal suivi de ENERGYSTARScore puis de LargestPropertyUseType_Supermarket/Grocery Store et de LargestPropertyUseType_Non-Refrigerated Warehous

Partie 4 : Analyse de la "feature importance" globale et locale

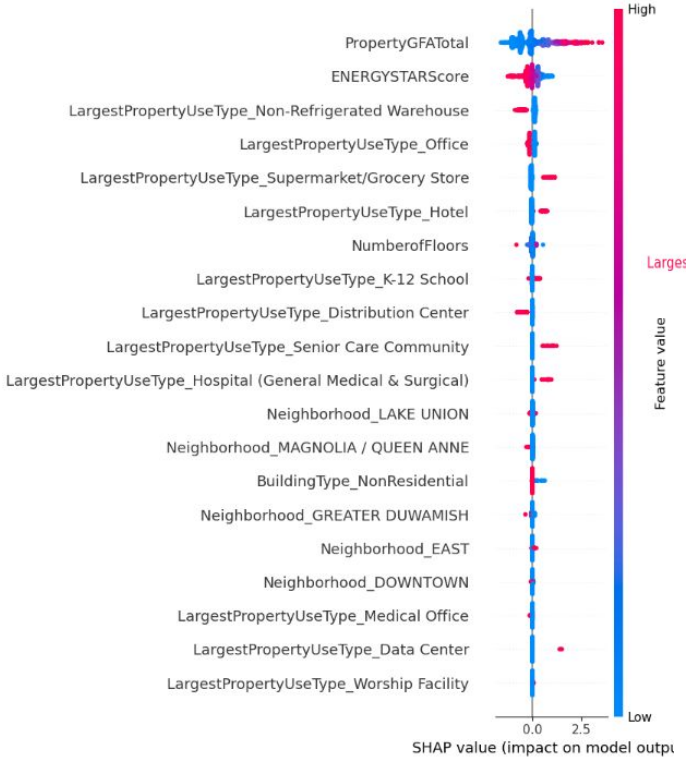
Importance des caractéristiques globale

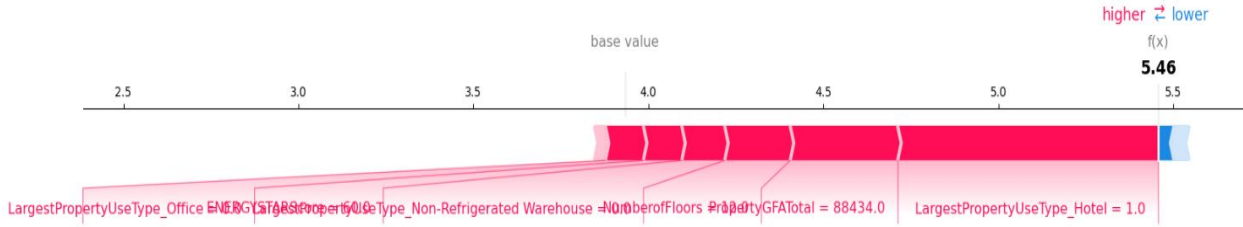


Les caractéristiques qui ont le plus d'impact sur les prédictions du modèle des émissions de CO2 sont PropertyGFATotal suivi de ENERGYSTARScore puis de LargestPropertyUseType_Non-Refrigerated et de WarehousLargestPropertyUseType_Supermarket/Grocery Store

Partie 4 : Analyse de la "feature importance" globale et locale

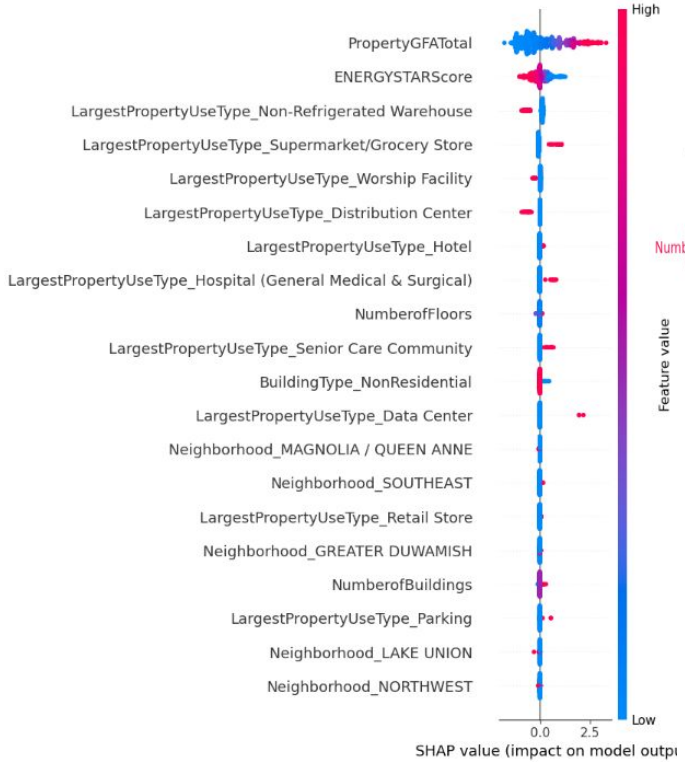
Analyse d'importance locale et globale avec SHAP : émissions de CO2

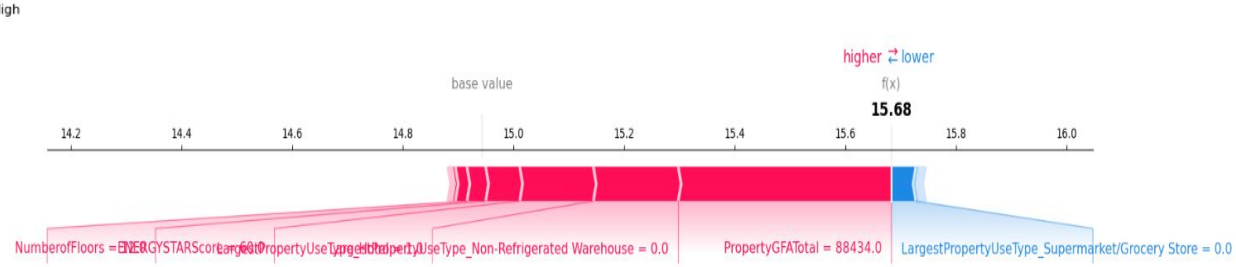




Partie 4 : Analyse de la "feature importance" globale et locale

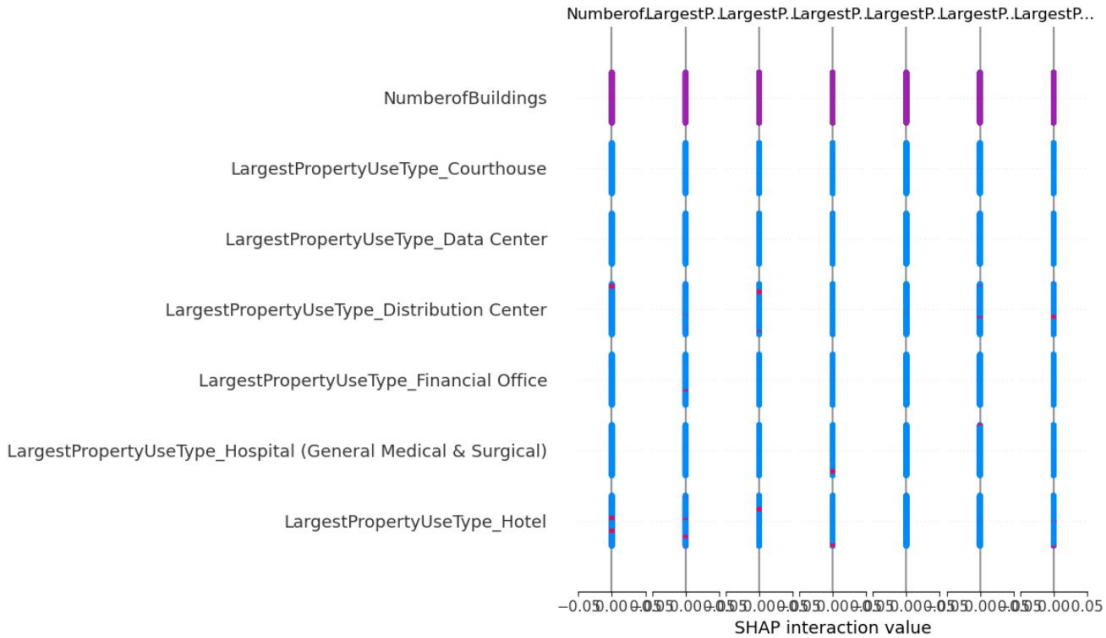
Analyse d'importance locale et globale avec SHAP : consommation d'énergie





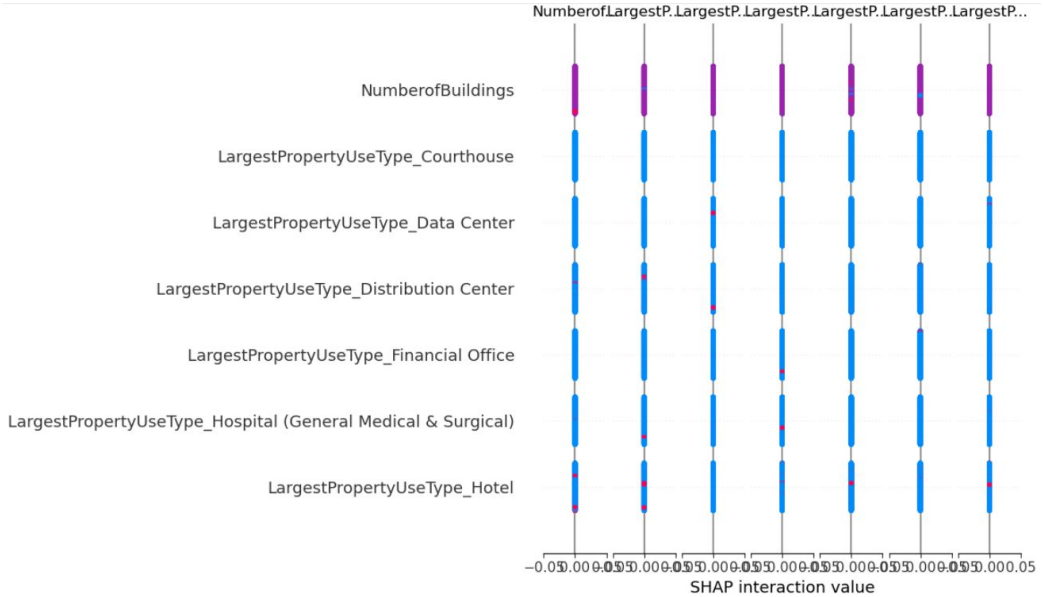
Partie 4 : Analyse de la "feature importance" globale et locale

Représentation graphique de l'importance des caractéristiques avec SHAP Interaction Values : émissions de CO2



Partie 4 : Analyse de la "feature importance" globale et locale

Représentation graphique de l'importance des caractéristiques avec SHAP Interaction Values : consommation d'énergie



Partie 5 : Analyse de l'influence de l'EnergyStarScore

Partie 5 : Analyse de l'influence de l'EnergyStarScore

	Avec ENERGYSTARScore		Sans ENERGYSTARScore	
	RMSE	R^2	RMSE	R^2
émissions de CO2	0.9071	0.4845	0.9844	0.3924
consommation d'énergie	0.5007	0.7786	0.6240	0.6523

⇒ ENERGYSTARScore joue un rôle significatif dans le modèle, améliorant à la fois la précision des prédictions (RMSE réduit) et la capacité explicative (R^2 augmenté) pour les 2 cibles, émissions de CO2 et consommation d'énergie.

Conclusion

Conclusion

Cible 1 : TotalGHGEmissions

Modèle de Régression : Gradient Boosting et peut être amélioré avec XGBoost

Cible 2 : SiteEnergyUse(kBtu)

Modèle de Régression : Gradient Boosting

Intérêt de la feature ENERGYSTARScore?

Intérêt bénéfique