Projet numéro 4 :

Anticipez les besoins en consommation de bâtiments pour la ville de Seattle

Thomas Zilliox

Janvier 2025

Présentation du Projet

Seattle

Contexte

- La ville de Seattle veut être neutre en émissions de carbone en 2050.
- Des relevés très coûteux ont été réalisé dans un échantillon de bâtiment

Mission

- Prédire les émissions de CO2 et la consommation totale d'énergie des bâtiments non destinés à l'habitation.
- Évaluer l'importance de l'ENERGY STAR Score.

Sommaire

Résumé des étapes du projet :

1) Exploration des données et premier feature engineering

2) Algorithme de régression linéaire

3) Algorithmes plus complexes : RandomForest, SVM, etc…

4) Feature importance global et local

Première étape : Exploration des données et premier feature engineering

Deux parties :

- Premier filtrage du jeu de données

- Analyse exploratoire plus approfondie

Première étape : Exploration des données et premier feature engineering

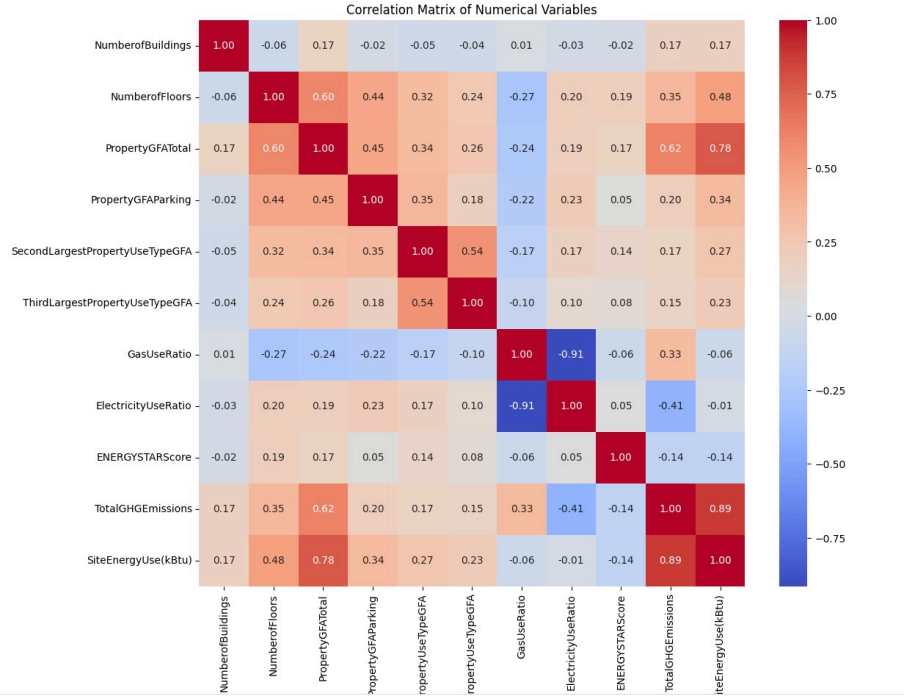
Jeu de données initial : **3376** lignes et **46** colonnes

- Étude sur les bâtiments non-résidentielles (colonne 'BuildingType')
- Filtrage des données aberrantes : colonnes 'Outlier' et 'ComplianceStatus'
- Suppression des lignes dont les valeurs au sein de la colonne 'LargestPropertyUseType' sont nulles (20 lignes)
- Remplacement des valeurs nulles au sein des colonnes SecondLargestPropertyUseType et ThirdLargestPropertyUseType par les valeurs « Non renseigné »
- Remplacement des valeurs nulles au sein des colonnes SecondLargestPropertyUseTypeGFA et ThirdLargestPropertyUseTypeGFA par 0
- Suppression des colonnes non utiles

Jeu de données final : **1544** lignes et **31** colonnes

Première étape : Exploration des données et premier feature engineering

Corrélations entre les variables numériques :



Suppression des colonnes en doublons :

- 'Electricity(kWh)',
- 'NaturalGas(therms)',
- 'SiteEUIWN(kBtu/sf)',
- 'SourceEUIWN(kBtu/sf)',
- 'SiteEnergyUseWN(kBtu)'

Suppression des colonnes trop corrélées avec d'autres colonnes :

- 'SiteEUI(kBtu/sf)' avec SourceEUI(kBtu/sf)

- 'PropertyGFABuilding(s)' et 'LargestPropertyUseTypeGFA' avec 'PropertyGFATotal'

Première étape : Exploration des données et premier feature engineering

Création de nouvelles variables :

Âge du bâtiment, Ratio de consommation de gaz et Ratio de consommation d'électricité.

Regroupement des valeurs :

de 'LargestPropertyUseType' en 13 catégories.

Nouvelles suppressions des colonnes :

'SourceEUI(kBtu/sf)', 'SteamUse(kBtu)', 'Electricity(kBtu)', 'NaturalGas(kBtu)'et 'GHGEmissionsIntensity' car on doit se passer des relevés futurs.

Deuxième étape : Sélection des variables et premier algorithme

Trois parties :

- Séparation des variables en deux catégories distinctes : catégorielles et numériques
- Encodage des variables catégorielles (OneHotEncoder, OrdinalEncoder)
- Standardisation des données pour optimiser l'algorithme (RobustScaler)

Deuxième étape : Feature Engineering et premier algorithme

Catégorisation des variables

Variables catégorielles utilisées :

- 'PrimaryPropertyType'
- 'Neighborhood'
- 'LargestPropertyUseType'
- 'ZipCode'
- 'SecondLargestPropertyUseType'
- 'ThirdLargestPropertyUseType'

Variables numériques utilisées :

- 'NumberofBuildings'
- 'NumberofFloors'
- 'PropertyGFATotal'
- 'PropertyGFAParking'
- 'SecondLargestPropertyUseTypeGFA'
- 'ThirdLargestPropertyUseTypeGFA'
- 'ENERGYSTARScore'

Traitements des variables pour l'algorithme de prédiction :

Variables numériques → RobustScaler

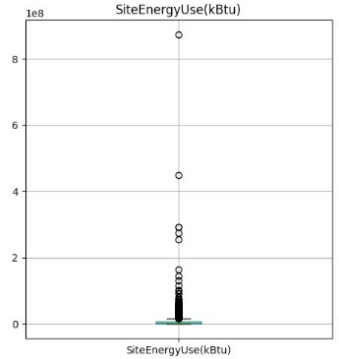
Variables catégorielles → OrdinalEncoder, OneHotEncoder

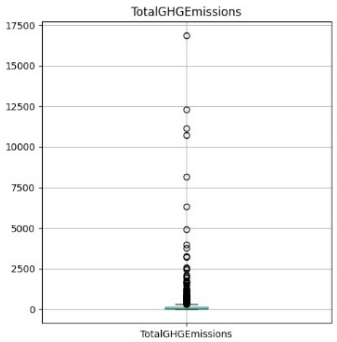
Deuxième étape : Feature Engineering et premier algorithme

Suppression des valeurs extrêmes pour optimiser notre algorithme :

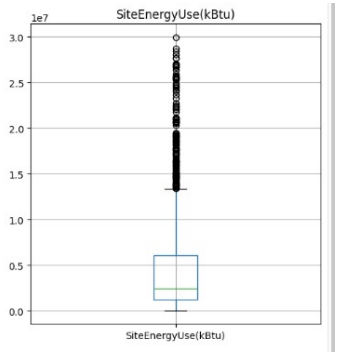
- TotalGHGEmissions > 1000 : 41 lignes supprimées (-3%)
- SiteEnergyUse(kBtu) > 30000000 : 84 lignes supprimées (5,5%)

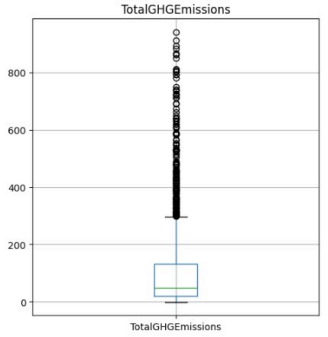
Avant suppression des valeurs extrêmes :





Après suppression des valeurs extrêmes :





Deuxième étape : Feature Engineering et premier algorithme

Premier essai : Régression linéaire

	Scores de l'algorithme	
TotalGHGEmissions	Root Mean Squared Error (RMSE):	131.370
	R^2 Score (Accuracy):	0.269
SiteEnergyUse (kBtu)	Root Mean Squared Error (RMSE):	3630815.252
	R^2 Score (Accuracy):	0.488

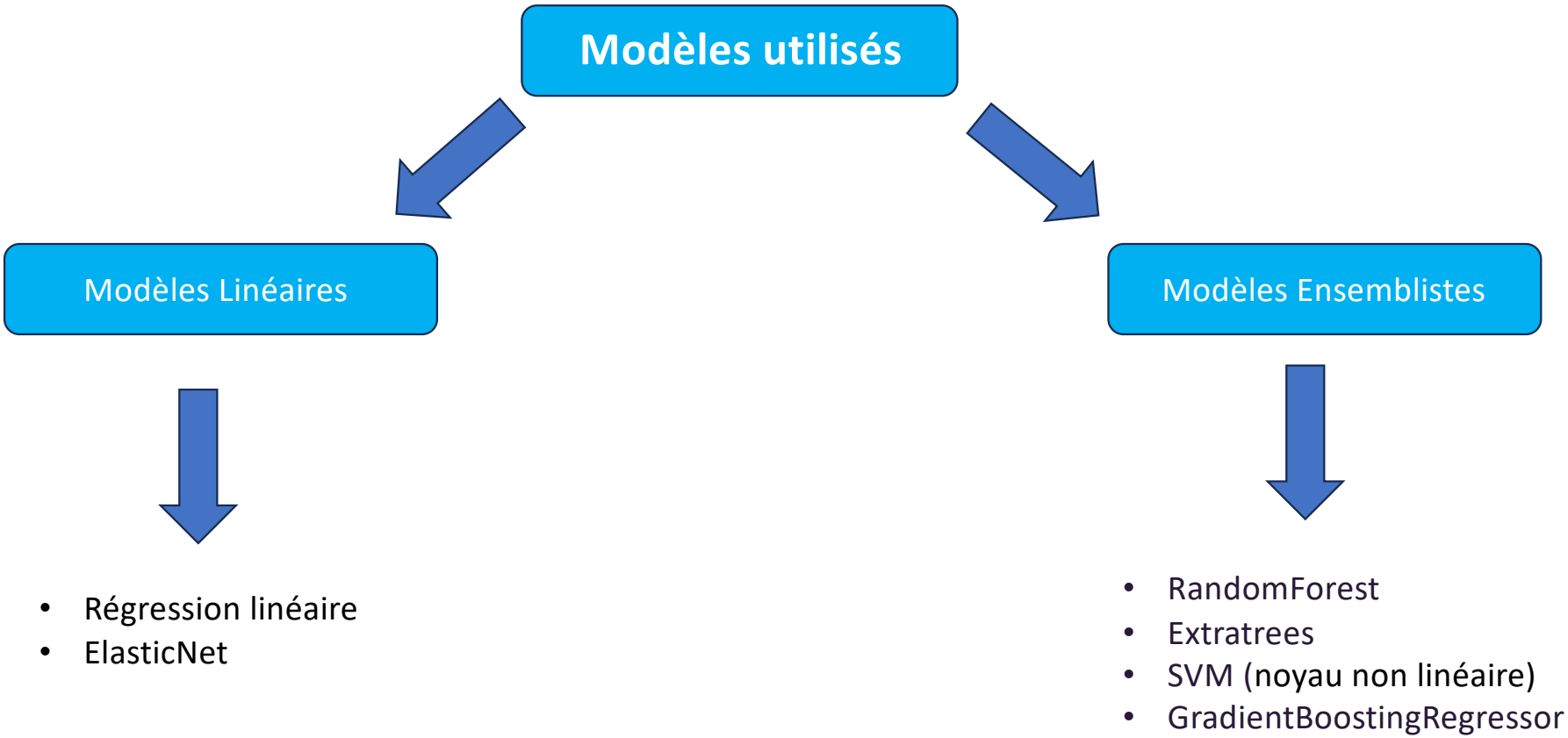
Troisième étape : Algorithmes plus complexes et validation croisée

Trois parties :

- Essai d'algorithmes plus complexes : RandomForest, Extratrees, ElasticNet, SVM et GradientBoostingRegressor

- Validation croisée pour optimiser les algorithmes de prédictions

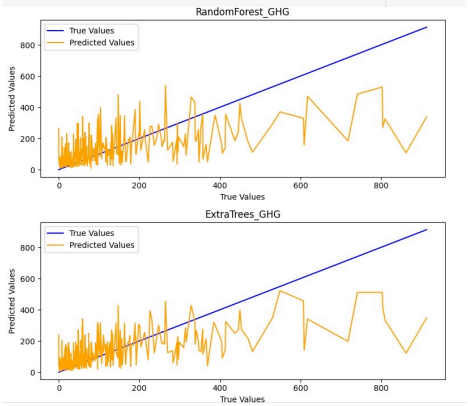
Troisième étape : Algorithmes plus complexes et validation croisée

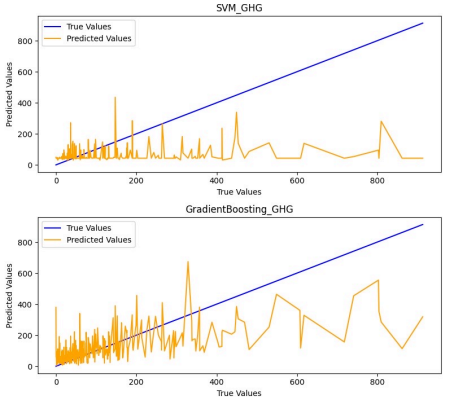


Troisième Étape : Approches des modélisations étudiées

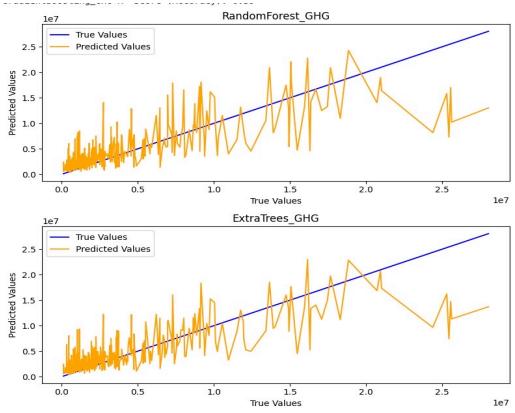
Graphique des prédictions sans l'ENERGYSTARSCORE :

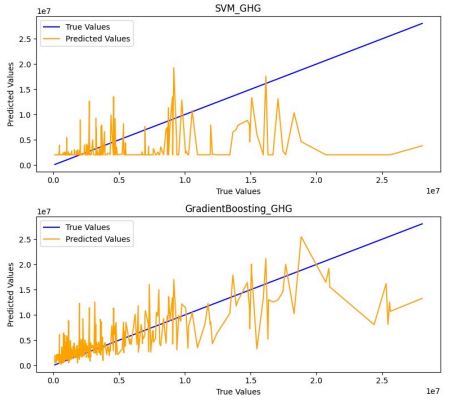
→ TotalGHGEmissions





→ SiteEnergyUse (kBtu)

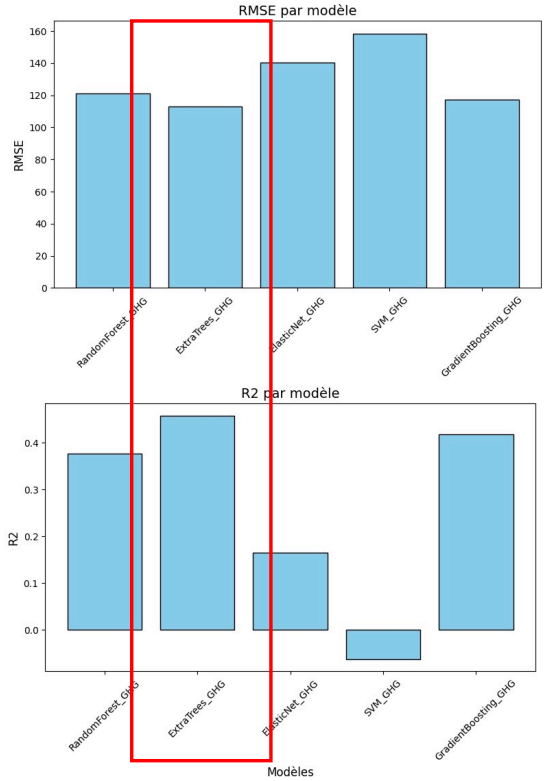




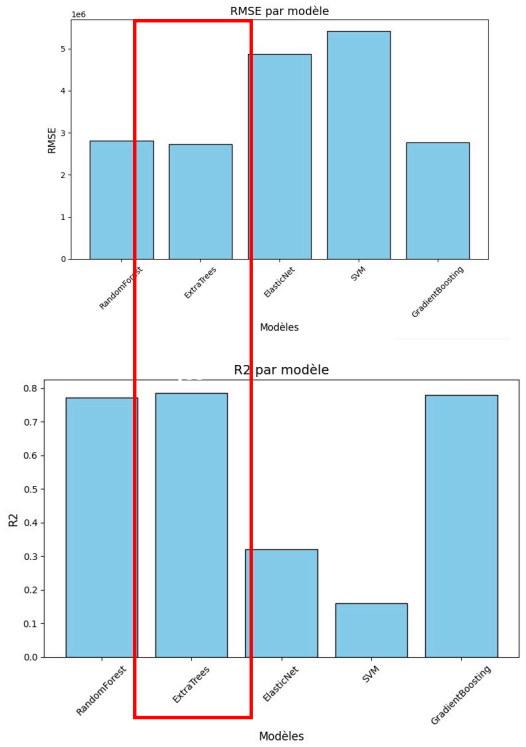
Troisième Étape : Approches des modélisations étudiées

Graphique des prédictions sans l'ENERGYSTARSCORE :

TotalGHGEmissions



SiteEnergyUse (kBtu)



Troisième étape : Algorithmes plus complexes et validation croisée

Optimisation des hyperparamètres :

Meilleurs résultats :

ExtraTreesRegressor
GradientBoostingRegressor

Score pour le modèle ExtraTreesRegressor

	31
param_model__max_depth	10
param_model__min_samples_split	15
param_model__n_estimators	100
mean_test_score	0.529133
mean_train_score	0.801352

dtype: object
Score pour le modèle GradientBoostingRegressor

	6
param_model__max_depth	3
param_model__min_samples_split	10
param_model__n_estimators	50
mean_test_score	0.553816
mean_train_score	0.827409

TotalGHGEmissions

Modèles testés :

- `ExtraTreesRegressor`
- `GradientBoostingRegressor`

5 validations croisées

Train Score

Test Score

SiteEnergyUse (kBtu)

`ExtraTreesRegressor`

5 validations croisées

param_model__max_depth	30
param_model__min_samples_split	15
param_model__n_estimators	50
mean_test_score	0.663781
mean_train_score	0.865345

Modèle le plus robuste possible : train_score et test_score qui ne soient pas trop éloignés

Troisième Étape : Approches des modélisations étudiées

Application de la validation croisée (GridSearchCV) :

Grille des paramètres :

ExtraTreesRegressor	GradientBoostingRegressor
'model__n_estimators': [10,20, 50, 100],	'model__n_estimators': [10,20, 50, 100],
'model__max_depth': [None, 10, 20,30],	'model__max_depth': [None, 10, 20,30],
'model__min_samples_split': [2, 5, 10,15]	'model__min_samples_split': [2, 5, 10,15]

Grille des paramètres :
'model__n_estimators': [10,20, 50, 100],
'model__max_depth': [None, 10, 20,30],
'model__min_samples_split': [2, 5, 10,15]

Meilleurs résultats :

Quatrième Étape : Features importances globale et locale

Deux parties :

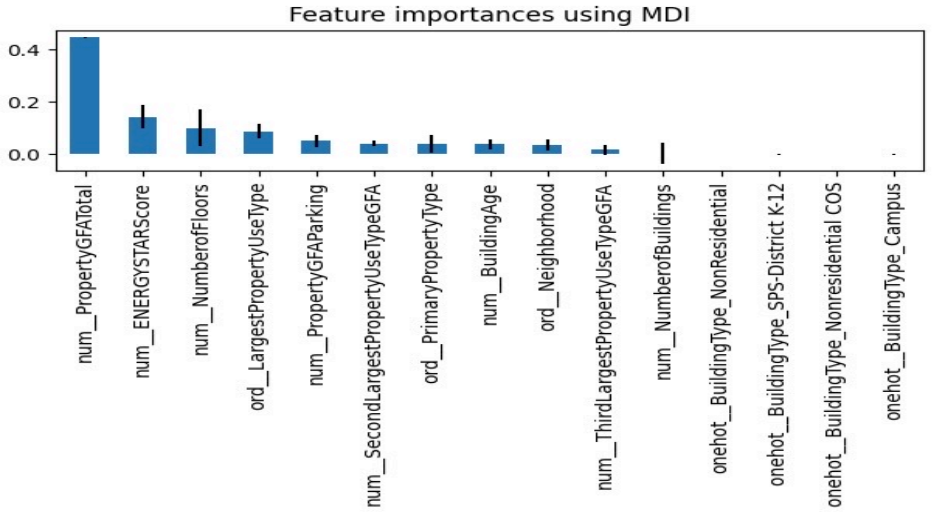
- Feature Importances globale

- Feature Importances local

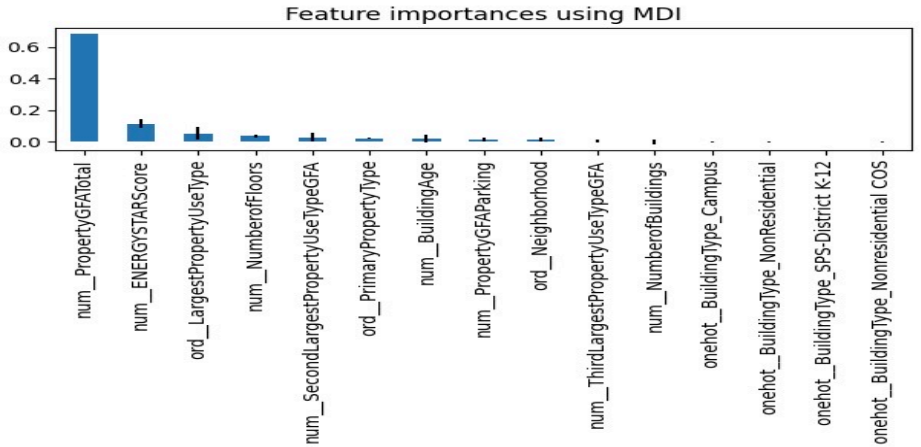
Quatrième Étape : Features importances globale et locale

Feature importance global :

TotalGHGEmissions :



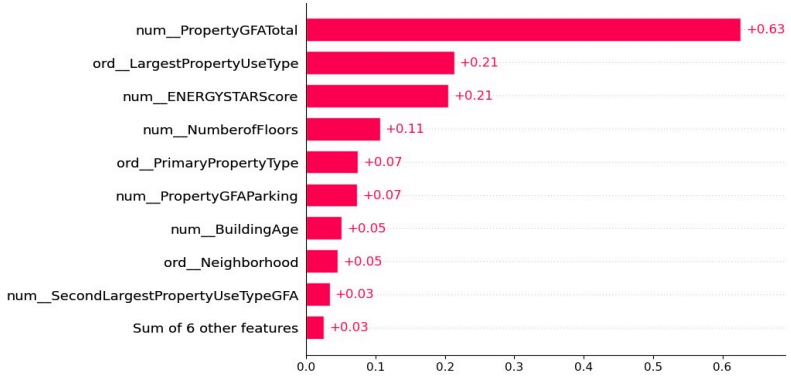
SiteEnergyUse(kBtu) :



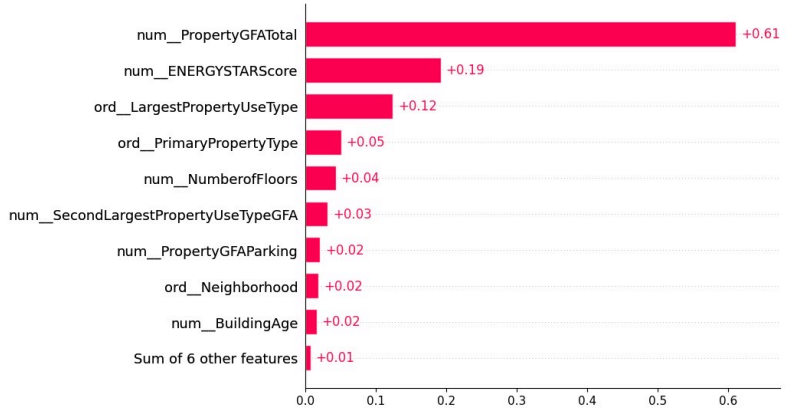
Quatrième Étape : Features importances globale et locale

Feature importance locale (méthode SHAP) :

TotalGHGEmissions :



SiteEnergyUse(kBtu) :



Conclusion

Pour notre algorithme de prédiction, avec l'aide de la validation croisée et du GridSearchCV, les meilleurs algorithmes de prédiction à utiliser sont :

- Le ExtraTreesRegressor et le GradientBoostingRegressor pour la prédiction d'émissions de CO2
- Le ExtraTreesRegressor pour la consommation d'énergie

Sans les relevés énergétiques futurs, les variables les plus importantes pour les prédictions de notre modèle sont :

- PropertyGFATotal, ENERGYSTARSCORE et Numberoffloors pour la prédiction d'émissions de CO2
- PropertyGFATotal, ENERGYSTARSCORE et LargestPropertyUseType pour la prédiction de consommation d'énergie

Perspectives d'amélioration :

- Collecte de données supplémentaires (Type de système de chauffage et de refroidissement, Isolation et matériaux de construction.),
- Essai de modèles avancés (XGBoost / LightGBM, Deeplearning),
- Segmentation des modèles et augmentation des données

Merci pour votre attention.