Projet 4

Anticipez les besoins en consommation électrique de bâtiments

BOUKHEZAR Yani

SOMMAIRE

1- Présentation de l'objectif

Problématique

1.Réaliser une courte analyse exploratoire.
2.Tester différents modèles de prédiction afin de répondre au mieux à la problématique.

Objectifs

Prédire la consommation énergétique des bâtiments

Prédire la consommation de CO2 des bâtiments

Evaluer l'intérêt de la variable « ENERGYSTARScore »

Bâtiments non destinés à l'habitation

2- Présentation du documents

Contient des informations sur les bâtiments de la ville de Seattle

- Information sur la consommation des bâtiments
- Information sur le type de bâtiment
- Information sur l'emplacement des bâtiments

19952 données manquantes

3376 lignes et 46 colonnes

3- Nettoyage et analyse exploratoire

Sélections des bâtiments non destinés à l'habitation

```
array(['NonResidential', 'Nonresidential COS', 'Multifamily MR (5-9)',
       'SPS-District K-12', 'Campus', 'Multifamily LR (1-4)',
       'Multifamily HR (10+)', 'Nonresidential WA'], dtype=object)
```

- Apres une recherche internet, SPS-District K-12 est Seattle public school District. c'est une école qui n'est donc pas destiné a l'habitation.
- Les habitation contenant 'Multifamily' sont donc a supprimer.

Data frame est presque diviser par 2

1668 lignes et 46 colonnes

3- Nettoyage et analyse exploratoire

Suppression des colonnes avec un taux de valeur manquante supérieur a 75%

5 colonnes sont donc supprimées.

	Variable	Taux_de_Null
0	Comments	100.000000
1	Outlier	98.980815
2	YearsENERGYSTARCertified	94.124700
3	ThirdLargestPropertyUseType	78.836930
4	ThirdLargestPropertyUseTypeGFA	78.836930
5	SecondLargestPropertyUseType	48.741007
6	SecondLargestPropertyUseTypeGFA	48.741007
7	ENERGYSTARScore	34.412470
8	ZipCode	0.959233
9	LargestPropertyUseType	0.359712

Outliers

Certaines valeurs sont incohérentes ou aberrantes, il faut donc les corriger.

	YearBuilt	NumberofFloors	PropertyGFATotal	Latitude	Longitude	NumberofBuildings	EN
count	1668.000000	1668.000000	1.668000e+03	1668.000000	1668.000000	1666.000000	
mean	1961.913669	4.121103	1.188427e+05	47.616054	-122.332908	1.168667	
std	32.741755	6.563407	2.973622e+05	0.048168	0.024580	2.931409	
min	1900.000000	0.000000	1.128500e+04	47.499170	-122.411820	0.000000	
25%	1930.000000	1.000000	2.947775e+04	47.585458	-122.343280	1.000000	
50%	1965.000000	2.000000	4.928950e+04	47.612340	-122.332935	1.000000	
75%	1989.000000	4.000000	1.053250e+05	47.649675	-122.321675	1.000000	
max	2015.000000	99.000000	9.320156e+06	47.733870	-122.258640	111.000000	

Une recherche sur Internet permet de vérifier que le bâtiment le plus grand de Seattle fait 76 étages.

1 bâtiment a plus de 76 étages, il doit donc être supprimé.

Le nombre d'étages et de bâtiments ne peut être égale à 0.

Remplacer par la médiane

Valeurs négatives pour la variable TotalGHGEmissions remplacé par 0

BuildingType	0
PrimaryPropertyType	0
YearBuilt	0
NumberofFloors	0
PropertyGFATotal	0
LargestPropertyUseType	6
Neighborhood	0
Latitude	0
Longitude	0
NumberofBuildings	2
ENERGYSTARScore	574
SiteEUI(kBtu/sf)	3
SourceEUI(kBtu/sf)	2
SiteEnergyUse(kBtu)	2
SteamUse(kBtu)	2
Electricity(kBtu)	2
NaturalGas(kBtu)	2
TotalGHGEmissions	2
GHGEmissionsIntensity	2

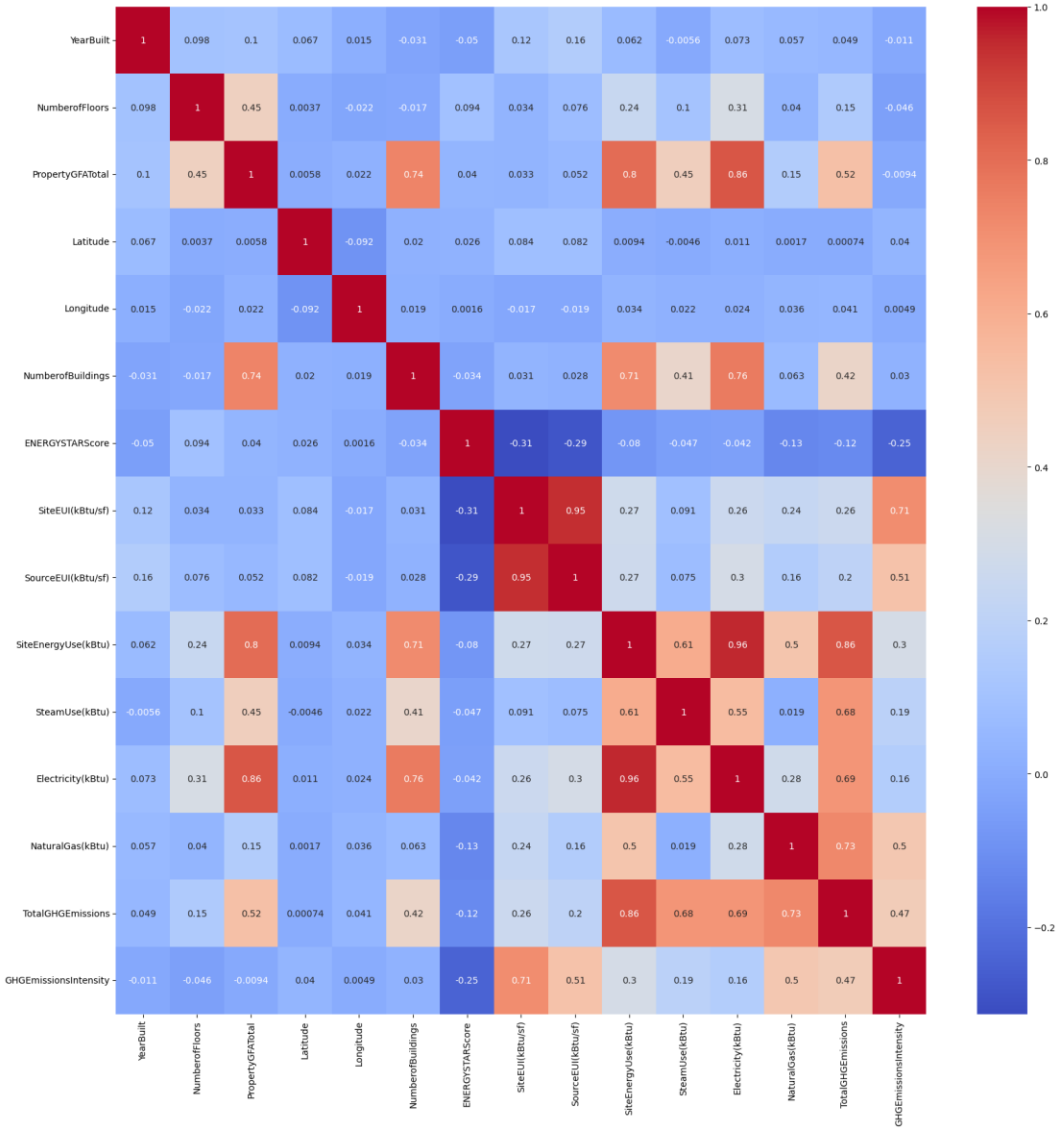
dtype: int64

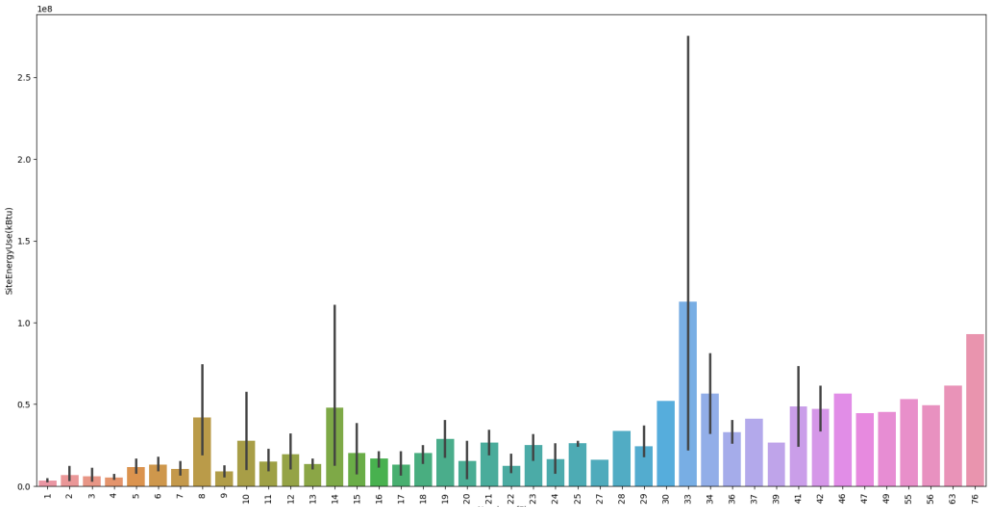
Peu de valeur manquante sauf pour ENERGYSTARScore

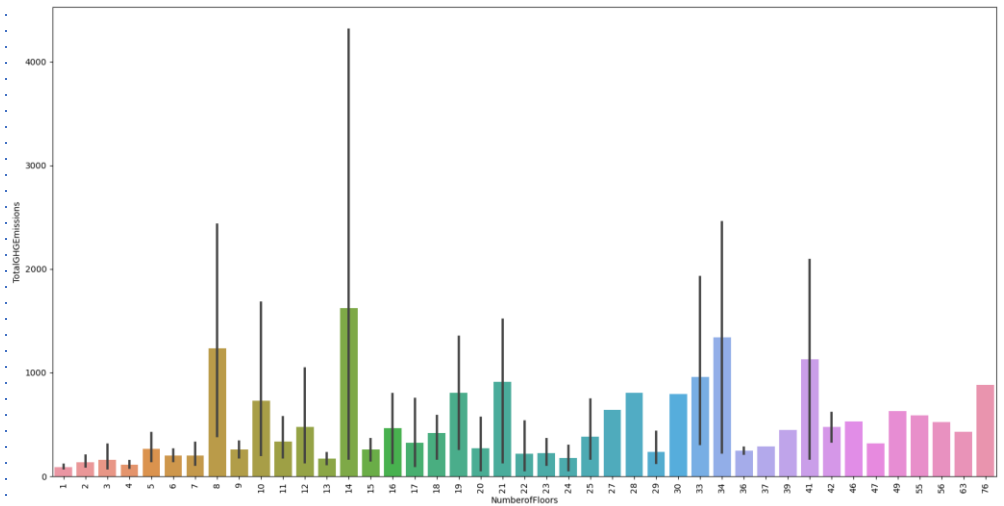
Supprimer les lignes avec des valeurs manquantes sauf ENERGYSTARScore

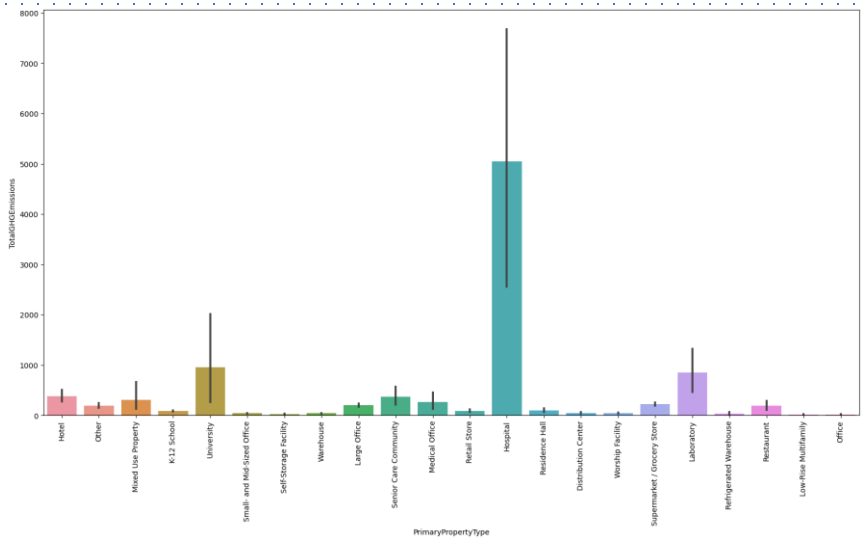
Une corrélation élevée entre deux variables, peut biaiser les prévisions.
Cela est dû à la colinéarité, qui se produit lorsque deux variables sont fortement corrélées et ont une forte dépendance les unes des autres. Dans ce cas, il peut être difficile de déterminer avec précision la contribution individuelle de chaque variable à la variabilité dans les données.

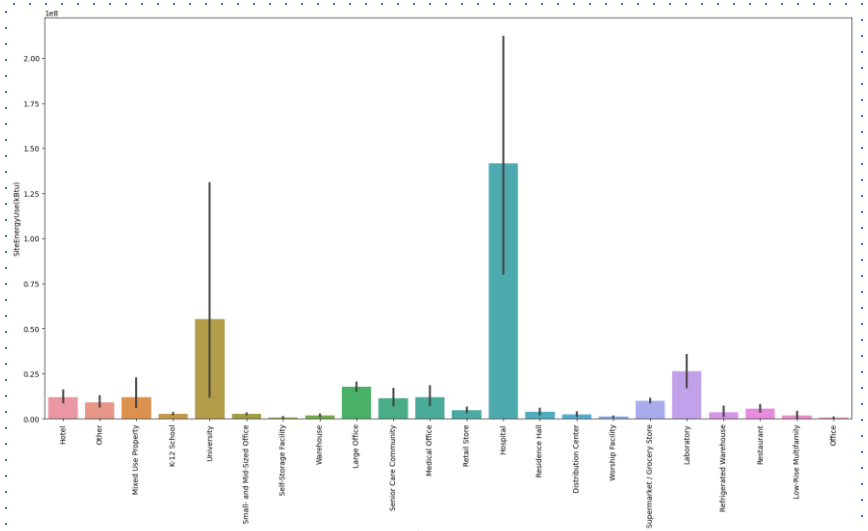
Seuil a 0,75

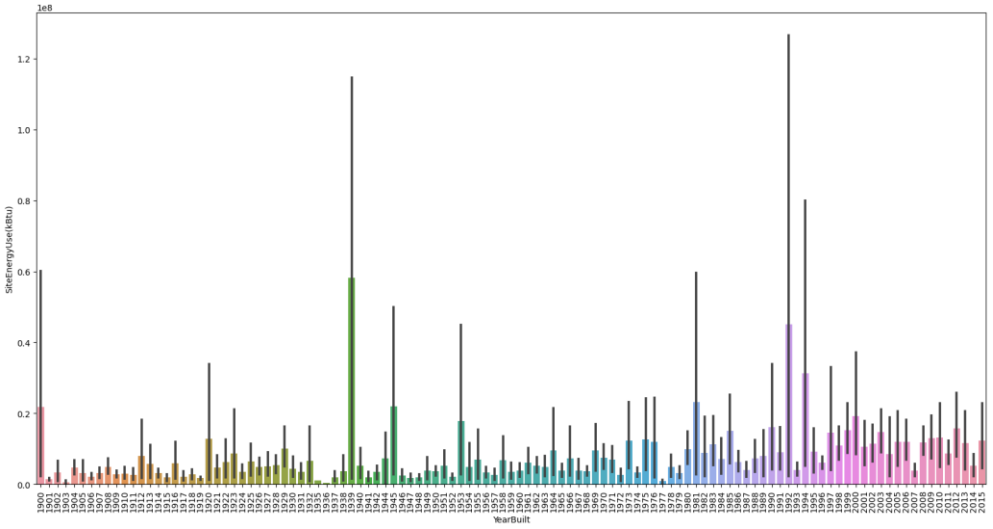


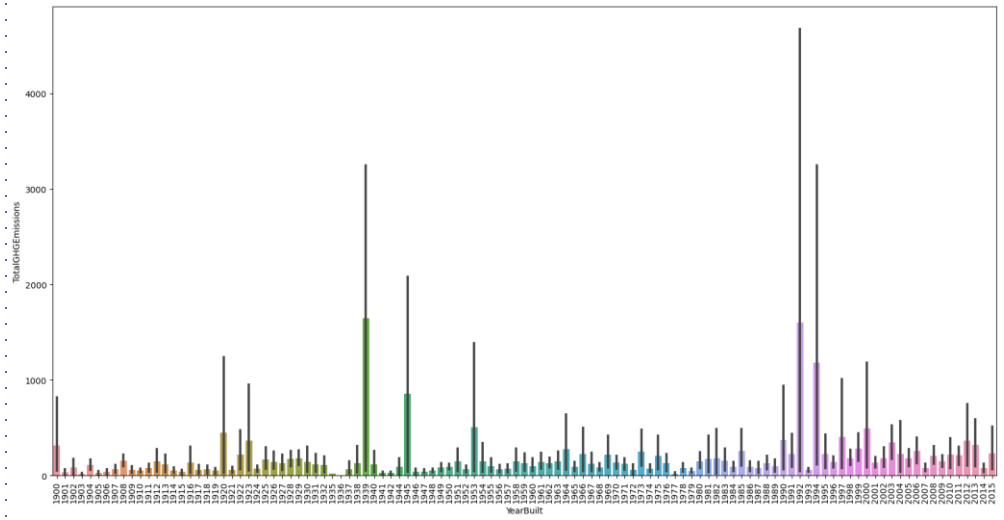

1e8
SiteEnergyUse(kBtu)
NumberofFloors

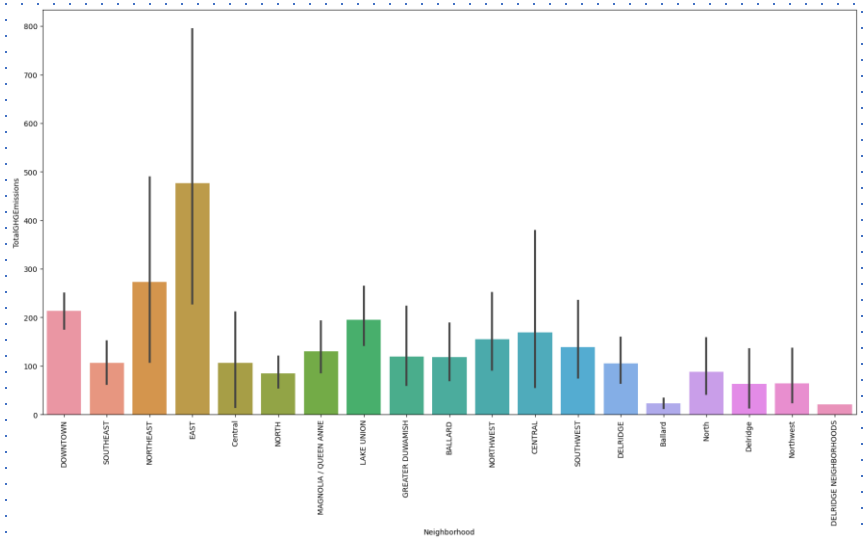

TotalGHGEmissions
NumberofFloors

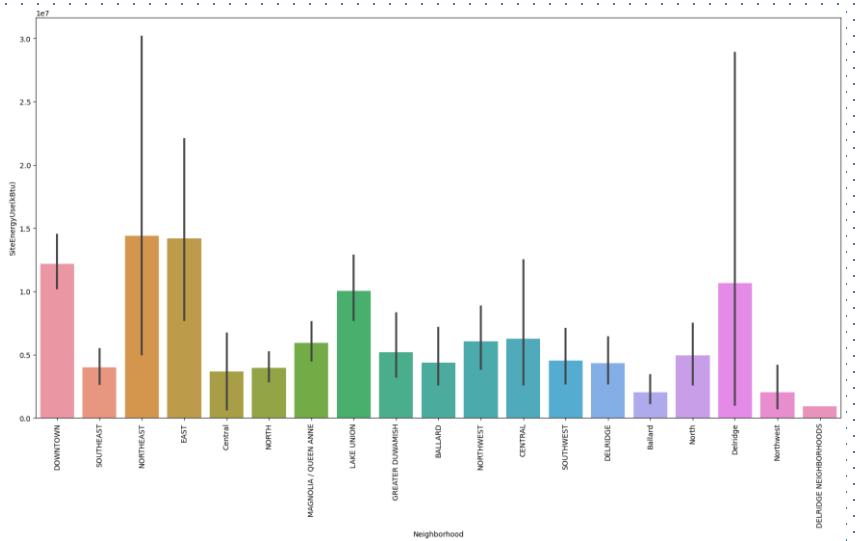

TotalGHGEmissions
Hotel
Other
Mixed Use Property
K-12 School
University
Small- and Mid-Sized Office
Self-Storage Facility
Warehouse
Large Office
Senior Care Community
Medical Office
Retail Store
Hospital
Residence Hall
Distribution Center
Worship Facility
Supermarket / Grocery Store
Laboratory
Refrigerated Warehouse
Restaurant
Low-Rise Multifamily
Office
PrimaryPropertyType

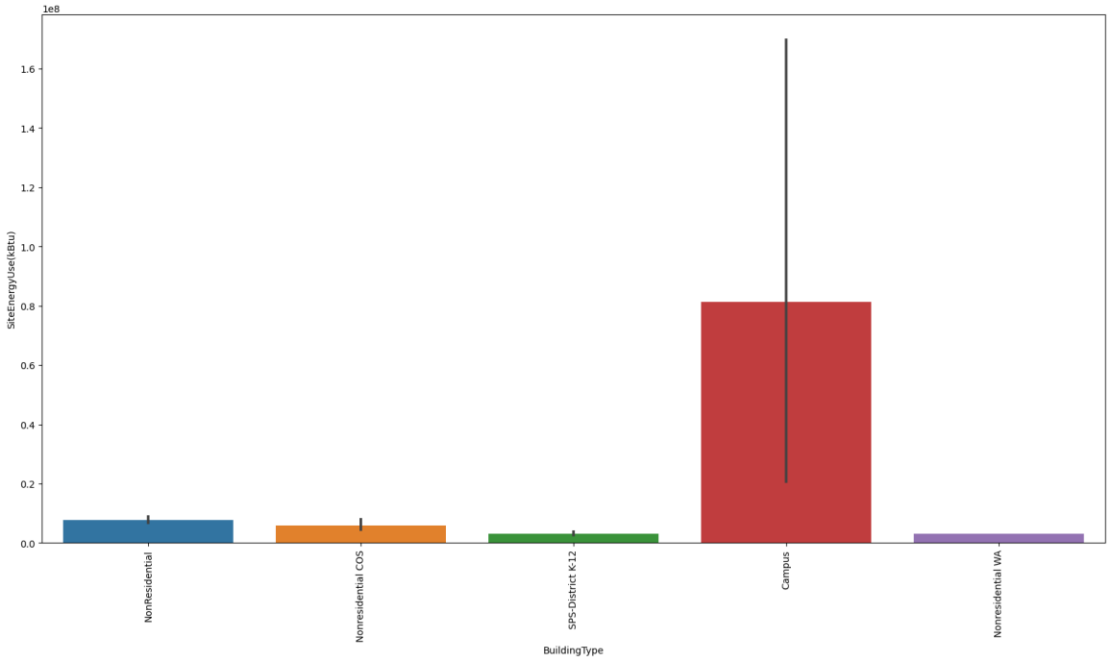

1e8
SiteEnergyUse(kBtu)
Hotel
Other
Mixed Use Property
K-12 School
University
Small- and Mid-Sized Office
Self-Storage Facility
Warehouse
Large Office
Senior Care Community
Medical Office
Retail Store
Hospital
Residence Hall
Distribution Center
Worship Facility
Supermarket / Grocery Store
Laboratory
Refrigerated Warehouse
Restaurant
Low-Rise Multifamily
Office
PrimaryPropertyType

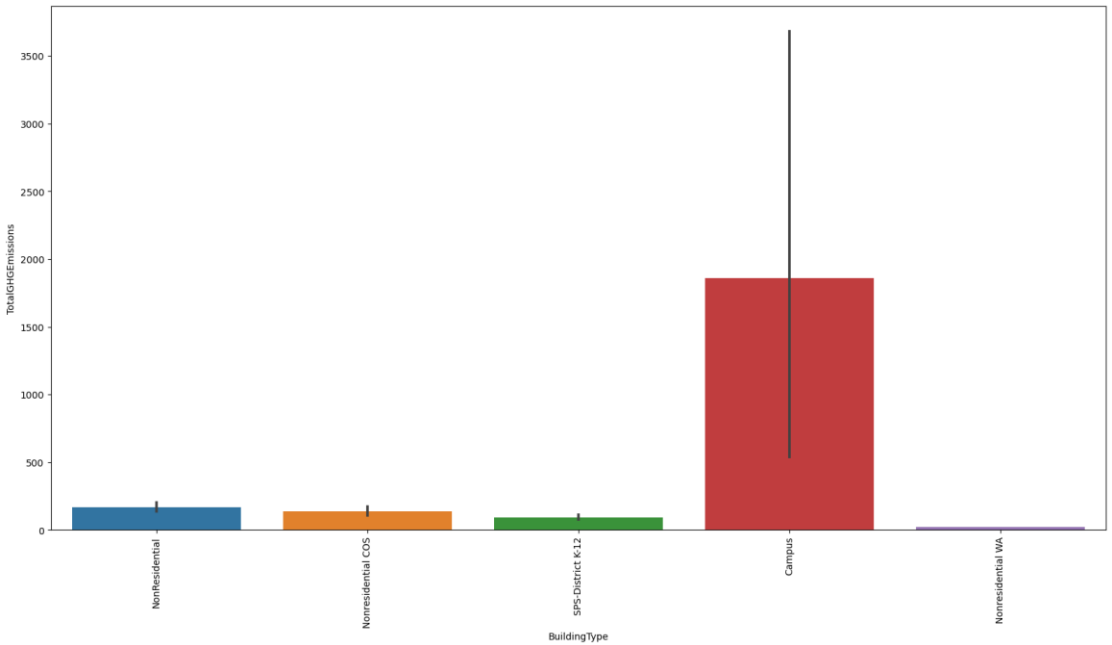

1e8
SiteEnergyUse(kBtu)
YearBuilt


TotalGHGEmissions
YearBuilt


TotalGHGEmissions
DOWNTOWN
SOUTHEAST
NORTHEAST
EAST
Central
NORTH
MAGNOLIA / QUEEN ANNE
LAKE UNION
GREATER DUWAMISH
BALLARD
NORTHWEST
CENTRAL
SOUTHWEST
DELRIDGE
Ballard
North
Delridge
Northwest
DELRIDGE NEIGHBORHOODS
Neighborhood


1e7
SiteEnergyUse(kBtu)
DOWNTOWN
SOUTHEAST
NORTHEAST
EAST
Central
NORTH
MAGNOLIA / QUEEN ANNE
LAKE UNION
GREATER DUWAMISH
BALLARD
NORTHWEST
CENTRAL
SOUTHWEST
DELRIDGE
Ballard
North
Delridge
Northwest
DELRIDGE NEIGHBORHOODS
Neighborhood


1e8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
SiteEnergyUse(kBtu)
NonResidential
Nonresidential COS
SPS-District K-12
Campus
Nonresidential WA
BuildingType


3500
3000
2500
2000
1500
1000
500
0
TotalGHGEmissions
NonResidential
Nonresidential COS
SPS-District K-12
Campus
Nonresidential WA
BuildingType

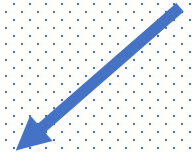
Ajout de variables

La création d'une colonne distance à partir de la longitude et de la latitude.

Distance entre les bâtiments et le centre de Seattle

Création de colonne à partir de LargestPropertyUseType

OneHotEncoder sur la variable



- Il permet de convertir des variables catégorielles en variables numériques qui peuvent être utilisées dans des modèles de machine learning.
- Il préserve les informations sur les différentes valeurs possibles d'une variable catégorielle, ce qui peut améliorer les performances du modèle.

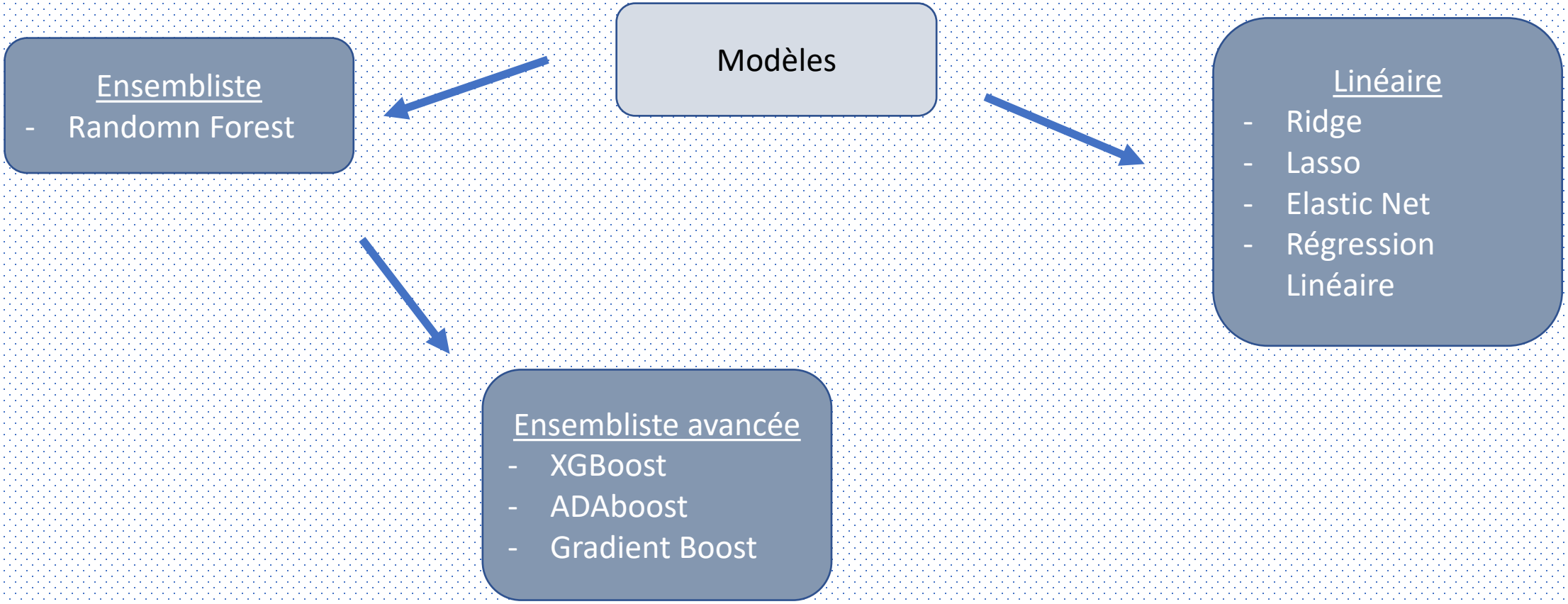
Dataframe final

```
Index(['BuildingType', 'PrimaryPropertyType', 'YearBuilt', 'NumberofFloors',
       'PropertyGFATotal', 'LargestPropertyUseType', 'Neighborhood',
       'NumberofBuildings', 'ENERGYSTARScore', 'SiteEnergyUse(kBtu)',
       'TotalGHGEmissions', 'distance', 'Autre', 'Bureau', 'Commercial',
       'Hotel / Restaurant', 'Industrial', 'Loisirs', 'Public', 'Residential',
       'Santé', 'Scolaire', 'Service', 'industrial'],
      dtype='object')
```

1659 lignes et 24 colonnes

Prediction des données

4- Modèle de prédiction



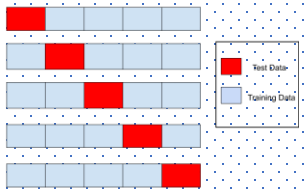
Test d'un modèle

scaling des données

Séparation des données en ensembles d'entraînement et de tests : les données doivent être divisées en deux ensembles distincts, l'un pour l'entraînement du modèle et l'autre pour tester les performances du modèle.

Choix du modèle

Validation croisée



Évaluation des performances du modèle

GridSearchCV pour l'optimisation des hyperparamètres

Évaluer la performance d'un modèle

R^2

Le coefficient de détermination, est une mesure de la qualité de l'ajustement d'un modèle de régression linéaire. Il indique la proportion de la variance totale de la variable dépendante qui peut être expliquée par le modèle. Le R^2 est compris entre 0 et 1.

RMSE

Root Mean Square Error, qui est une mesure de l'erreur quadratique moyenne entre les valeurs prédites et les valeurs réelles. Le RMSE est une mesure de l'erreur de prédiction qui est utilisée pour évaluer la précision d'un modèle de régression. Plus le RMSE est faible, meilleure est la qualité de prédiction du modèle.

MAE

Mean Absolute Error, qui est une mesure de l'erreur absolue moyenne entre les valeurs prédites et les valeurs réelles. Le MAE mesure de l'erreur de prédiction qui est utilisée pour évaluer la précision d'un modèle de régression. Contrairement au RMSE, le MAE ne donne pas plus de poids aux grosses erreurs, il est donc moins sensible aux valeurs aberrantes (ou outliers). Cependant, il ne tient pas compte des différences de magnitude entre les prédictions et les observations.

5- Comparaison des Modèles de prédictions

- Le temps d'exécution d'un modèle n'est pas un facteur décisif, mais il reste important.

	Variable	Modèle	R^2	RMSE	MAE	Temps d'exécution
0	SiteEnergyUse(kBtu)	Régression linéaire	0.500481	1.609443e+07	5.405087e+06	1.833003
0	SiteEnergyUse(kBtu)	Elastic Net	0.500555	1.609324e+07	5.351732e+06	0.094998
0	SiteEnergyUse(kBtu)	XGBoost	0.359914	1.821876e+07	6.291897e+06	138.974074
0	SiteEnergyUse(kBtu)	AdaBoost	0.316612	1.882493e+07	6.044865e+06	0.987000
0	SiteEnergyUse(kBtu)	Lasso	0.500481	1.609443e+07	5.405078e+06	0.072000
0	SiteEnergyUse(kBtu)	Ridge	0.500574	1.609294e+07	5.363923e+06	0.055000
0	SiteEnergyUse(kBtu)	Random Forest	0.540173	1.544177e+07	5.184939e+06	29.454016
0	SiteEnergyUse(kBtu)	Gradient Boost	0.606456	1.428553e+07	5.128797e+06	43.613023

	Variable	Modèle	R^2	RMSE	MAE	Temps d'exécution
0	TotalGHGEmissions	Régression linéaire	0.300589	631.807813	169.042789	0.084002
0	TotalGHGEmissions	Elastic Net	0.261558	649.197663	156.736822	0.096000
0	TotalGHGEmissions	XGBoost	0.292015	635.668632	137.227572	140.305071
0	TotalGHGEmissions	AdaBoost	0.269598	645.653932	184.585218	0.864001
0	TotalGHGEmissions	Lasso	0.295148	634.260451	159.588335	0.059999
0	TotalGHGEmissions	Ridge	0.300361	631.910894	166.691642	0.054000
0	TotalGHGEmissions	Random Forest	0.452407	559.046047	157.157276	30.735017
0	TotalGHGEmissions	Gradient Boost	0.363720	602.619322	163.875961	43.158022

Après avoir effectué les différents modèles de prédiction, on observe que Gradient Boost obtient une meilleure évaluation pour **SiteEnergyUse(kBtu)**.
Random Forest obtient une meilleure évaluation pour **TotalGHGEmissions.**

6 - ENERGYSTARScore

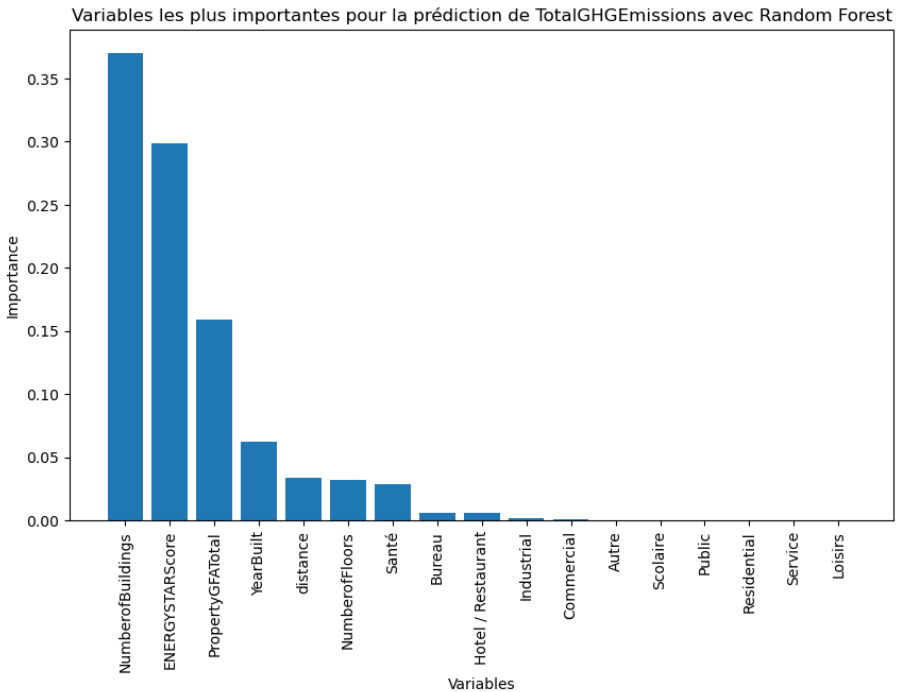
	Variable	Modèle	R^2	RMSE	MAE	Temps d'exécution
0	TotalGHGEmissions	Random Forest	0.526545	612.251537	159.214586	24.665013

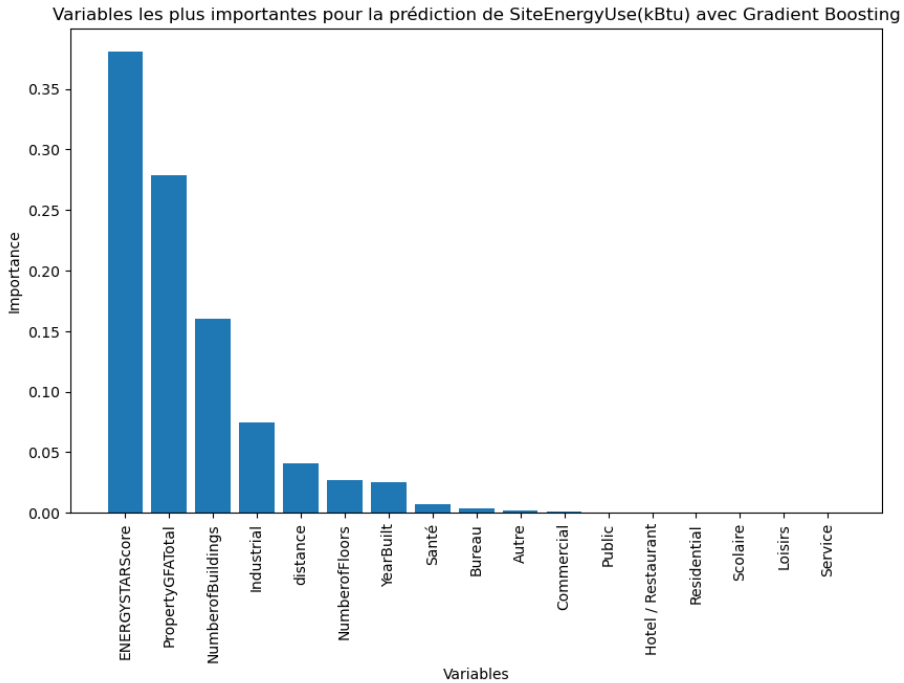
On observe que l'ajoute de la variable « ENERGYSTARScore » améliore uniquement R^2 résultats pour la variable TotalGHGEmissions.
Le modèle de prédiction n'est pas améliorer

	Variable	Modèle	R^2	RMSE	MAE	Temps d'exécution
0	SiteEnergyUse(kBtu)	Gradient Boost	0.70363	1.244035e+07	5.363923e+06	34.910018

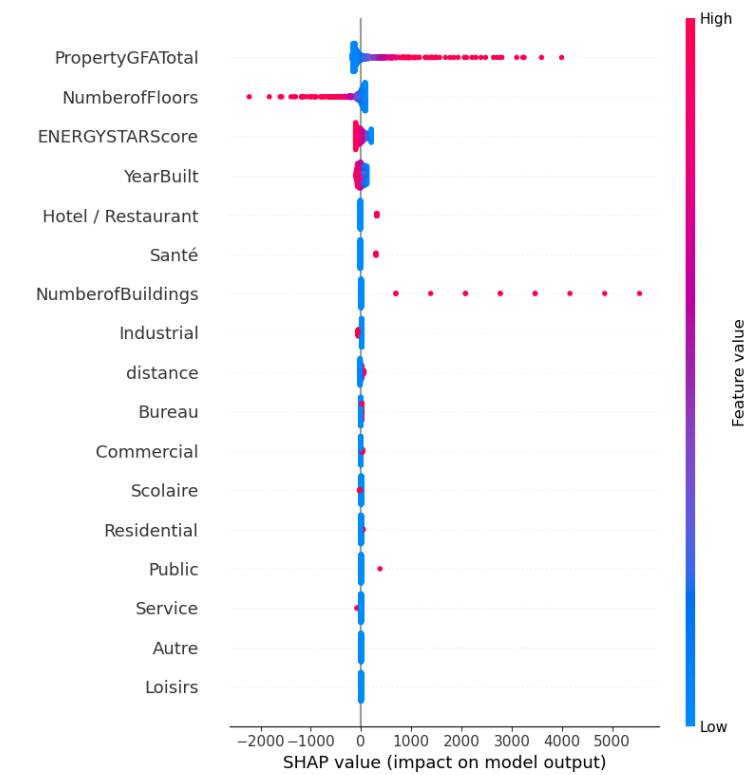
On observe que l'ajoute de la variable « ENERGYSTARScore » améliore les résultats pour la variable SiteEnergyUse(kBtu).

Importance des variables

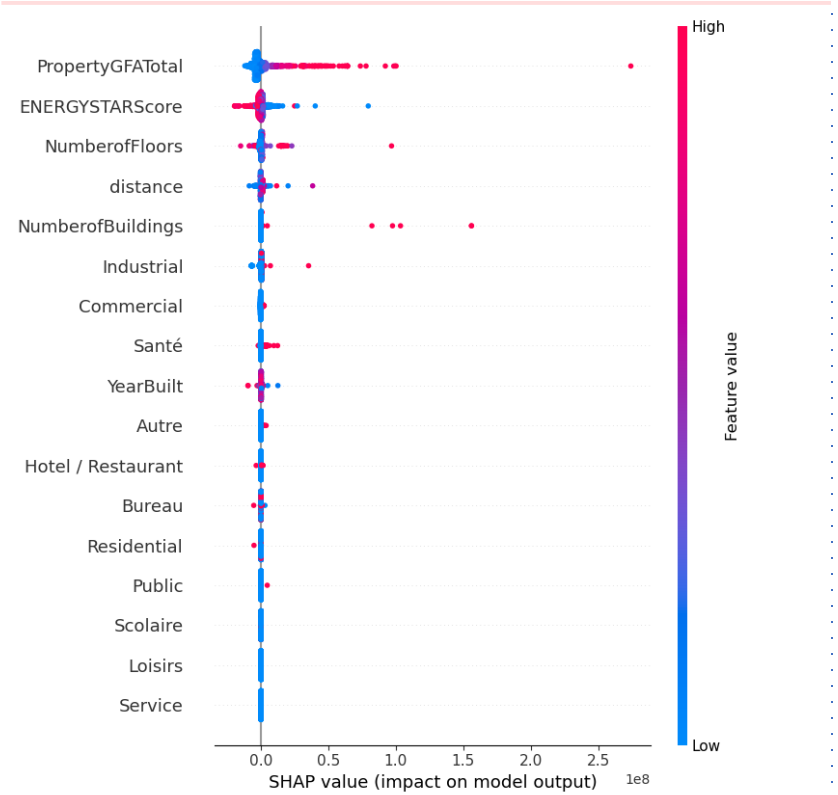




TotalGHGEmissions



SiteEnergyUse(kBtu)



Les graphiques de sommaires SHAP (SHAP summary plots) montrent l'impact moyen de chaque variable sur la sortie du modèle de prédiction.
Nous pouvons donc voir quelles variables ont l'impact le plus important sur les prédictions du modèle et comment leur impact varie en fonction de la valeur de la variable.

Conclusion

- Comparaison de différents modèles de prédiction, certains ont des résultats proches.
- Le Grading Boost Regressor est meilleur pour la prédiction de la variable « SiteEnergyUse(kBtu) »
- Le modèle Random Forest est le meilleur pour la prédiction de la variable « TotalGHGEmissions »

On observe que l'ajout de la variable « ENERGYSTARScore » améliore les résultats pour la variable SiteEnergyUse(kBtu) mais pas pour TotalGHGEmissions.

Axe d'amélioration :

Améliorer les valeurs manquantes de la variable ENERGYSTARScore