INGENIEUR MACHINE LEARNING
PROJET 3

Anticipez les besoins en consommation de bâtiments

Seattle

Anticipez les besoins en consommation de bâtiments

Objectifs

- Prédire les émissions de CO_2
- Prédire la consommation totale d’énergie
- Évaluer l’intérêt de l’"ENERGY STAR Score"

Présentation du jeu de données

data.seattle.gov/dataset/2016-Building-Energy-Benchmarking

1 - Informations sur les 46 paramètres

Location : adresse, code postal, quartier, ville, latitude et longitude

Structure : année de construction, nombre d'étage, nombre de bâtiments

Types des différents biens et leurs superficies

Relevés des consommations énergétiques

Relevés des émissions de CO_2

3376
observations

Présentation du jeu de données

2 - Informations sur les 3376 observations

Aucun doublon

Aucun erreur lexicale ou de formatage

Peu de valeurs manquantes : 12,8 %

Certaines valeurs aberrantes déjà identifiées (32)

Observation comportant des valeurs d'énergies négatives (1)

3343
observations

Présentation du jeu de données

3 - Sélection des observations et des variables

Suppression des catégories "Multifamily" de la variable "BuildingType"

Suppression identique pour la variable "LargestPropertyUseType"

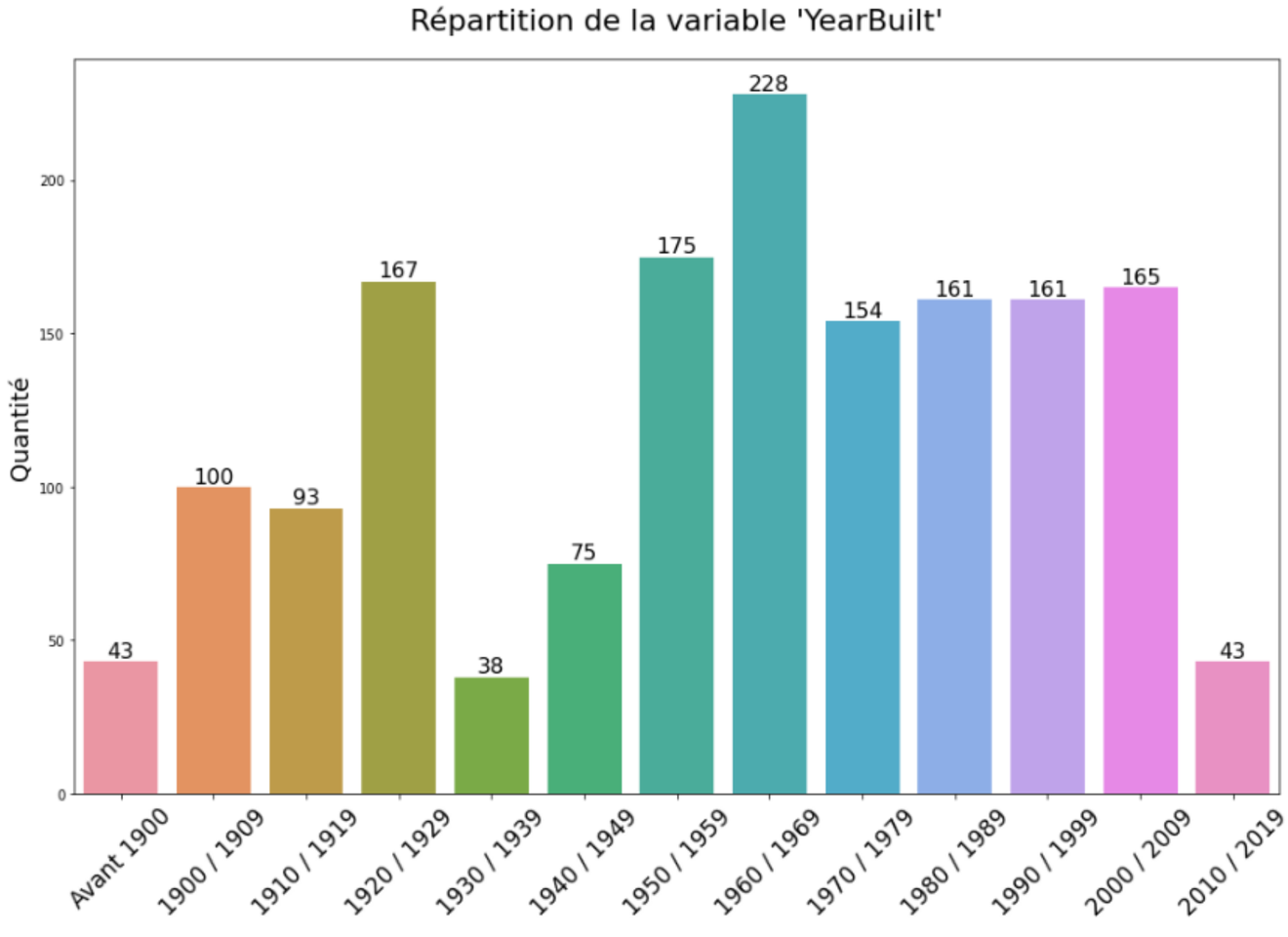
Traitement des variables par type

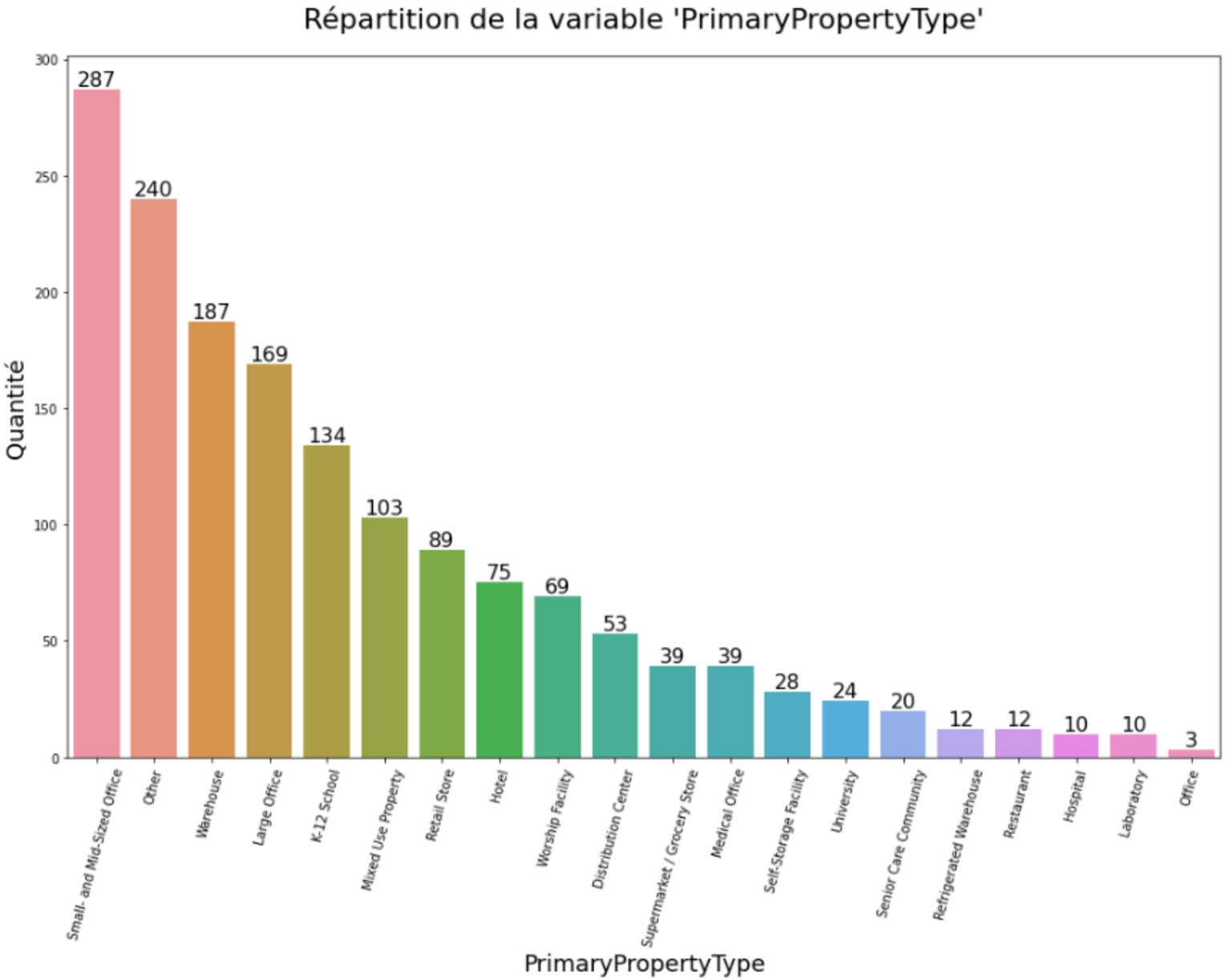
Valeurs manquantes restantes : 1,5 % ("ENERGYSTARScore")

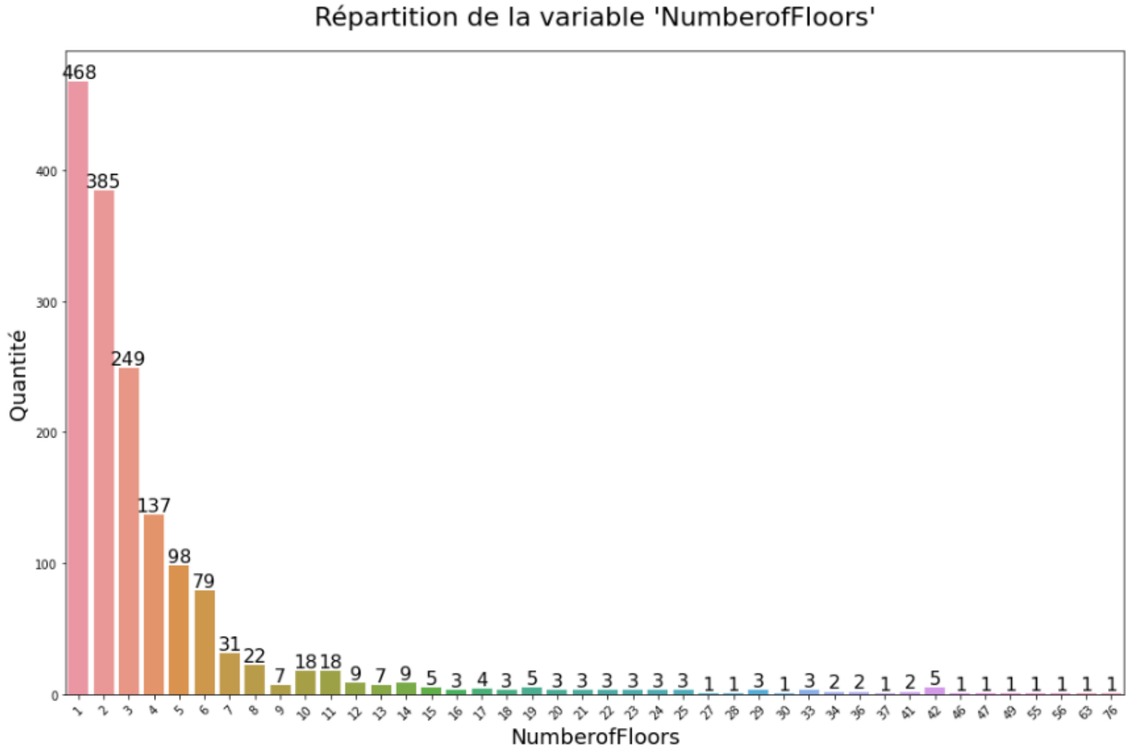
1603
observations

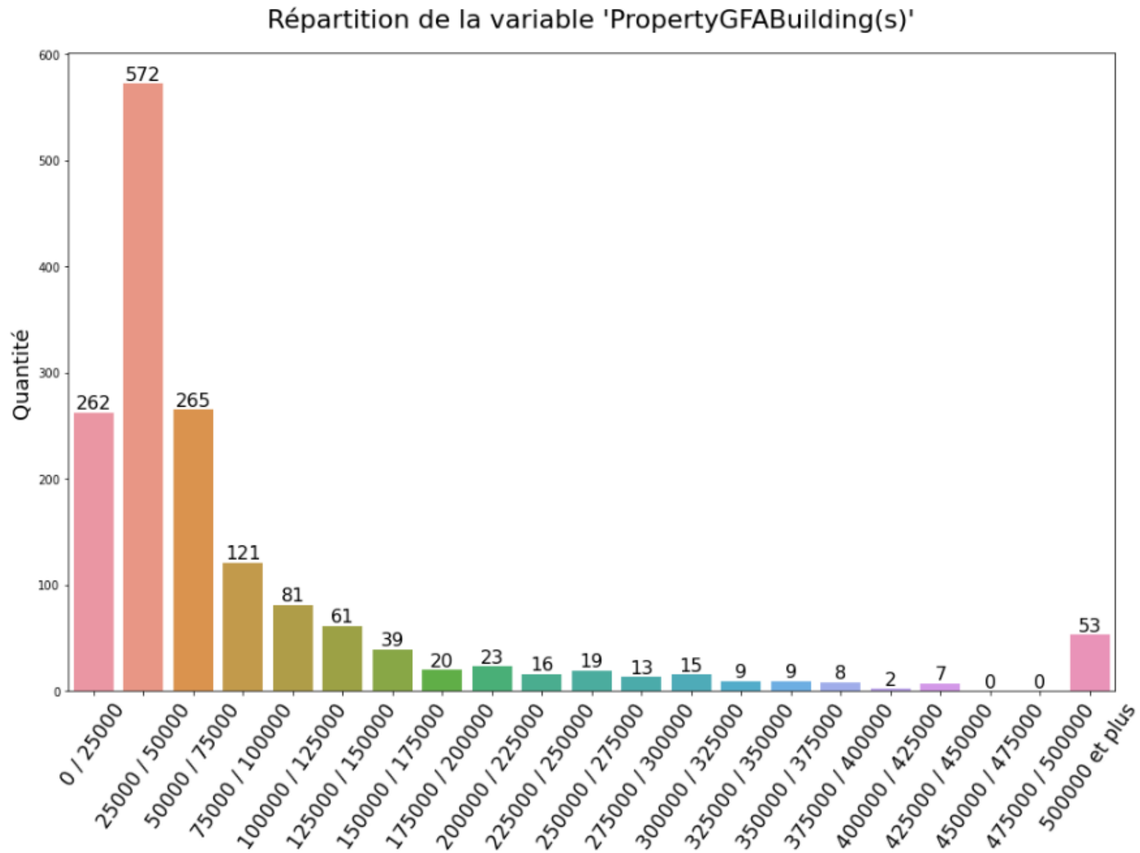
Présentation du jeu de données

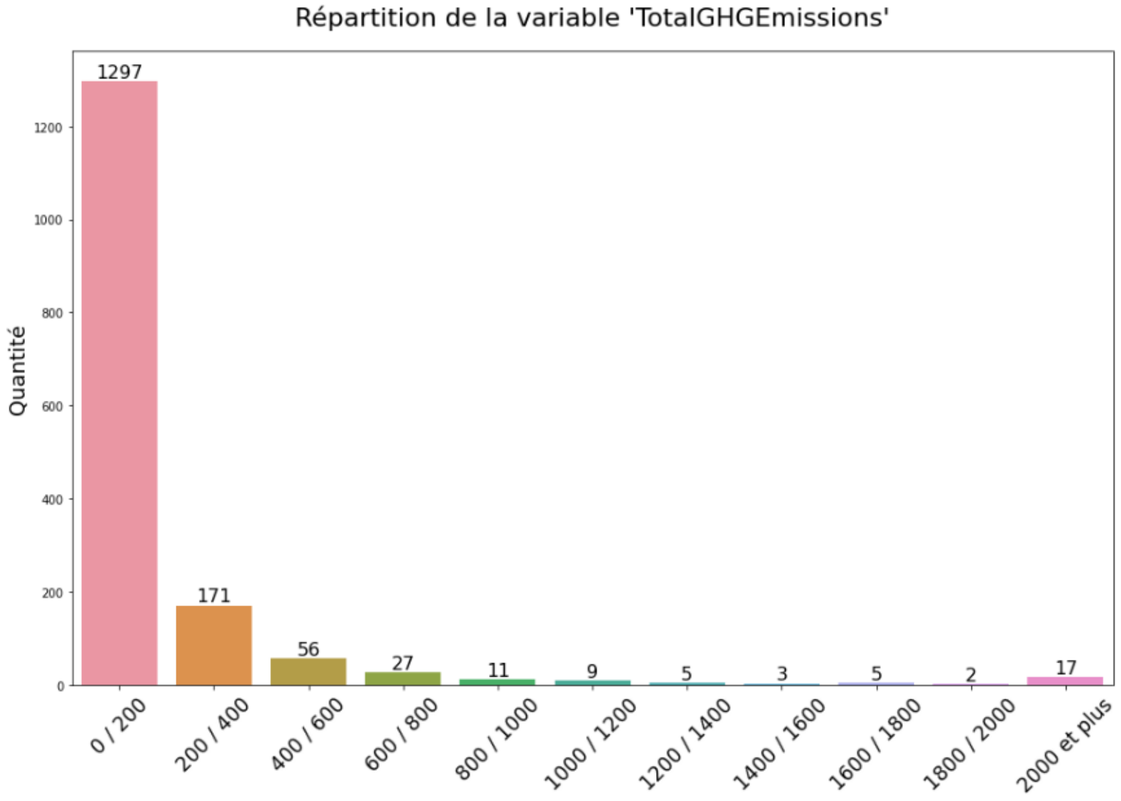
4 - Analyse univariée

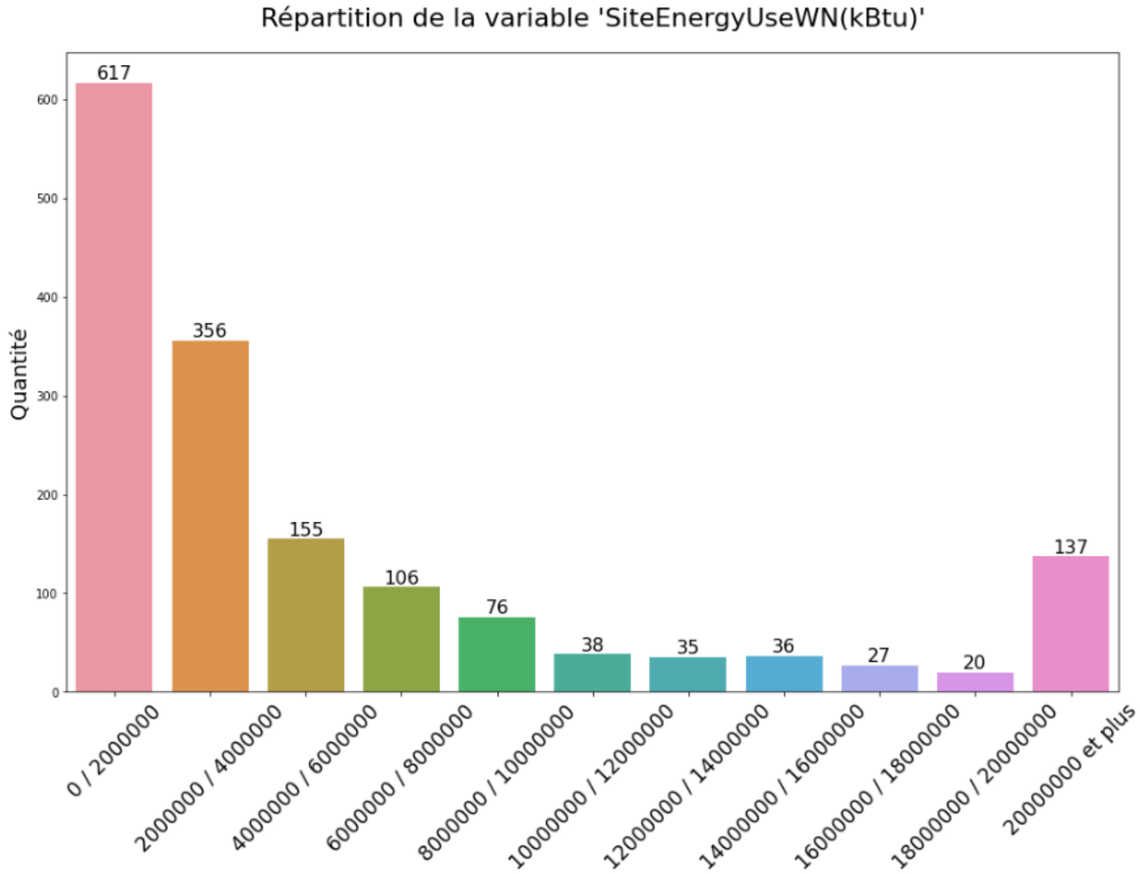



Répartition de la variable 'PrimaryPropertyType'
Quantité
PrimaryPropertyType
0
50
100
150
200
250
300
287
240
187
169
134
103
89
75
69
53
39
39
28
24
20
12
12
10
10
3
Small- and Mid-Sized Office
Other
Warehouse
Large Office
K-12 School
Mixed Use Property
Retail Store
Hotel
Worship Facility
Distribution Center
Supermarket / Grocery Store
Medical Office
Self-Storage Facility
University
Senior Care Community
Refrigerated Warehouse
Restaurant
Hospital
Laboratory
Office

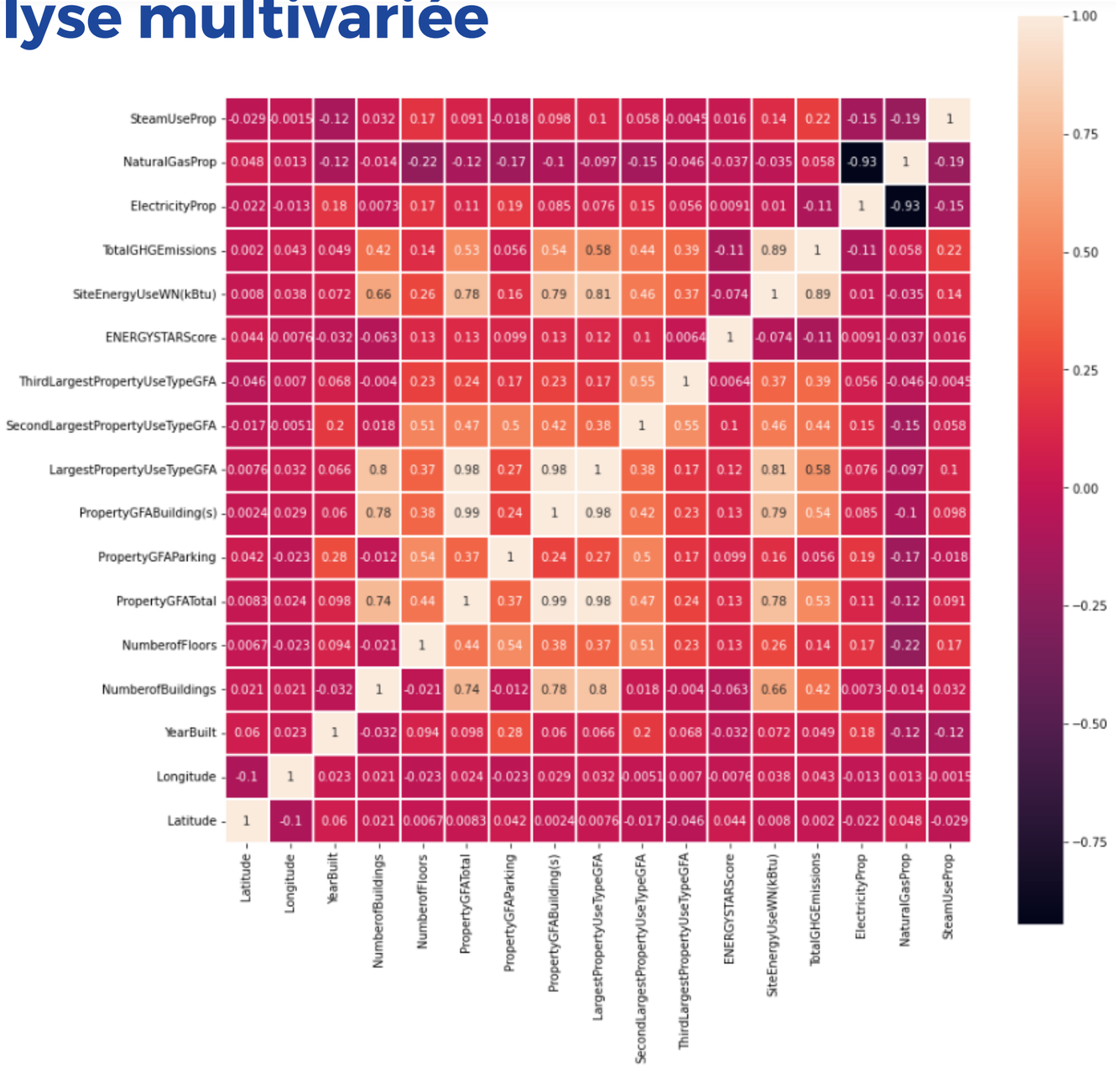

Répartition de la variable 'NumberofFloors'
Quantité
NumberofFloors
468
385
249
137
98
79
31
22
7
18
18
9
7
9
5
3
4
3
5
3
3
3
3
3
3
1
1
3
1
3
2
2
1
2
5
1
1
1
1
1
1
1


Répartition de la variable 'PropertyGFABuilding(s)'
Quantité
262
572
265
121
81
61
39
20
23
16
19
13
15
9
9
8
2
7
0
0
53
0 / 25000
25000 / 50000
50000 / 75000
75000 / 100000
100000 / 125000
125000 / 150000
150000 / 175000
175000 / 200000
200000 / 225000
225000 / 250000
250000 / 275000
275000 / 300000
300000 / 325000
325000 / 350000
350000 / 375000
375000 / 400000
400000 / 425000
425000 / 450000
450000 / 475000
475000 / 500000
500000 et plus


Répartition de la variable 'TotalGHGEmissions'
Quantité
1297
171
56
27
11
9
5
3
5
2
17
0 / 200
200 / 400
400 / 600
600 / 800
800 / 1000
1000 / 1200
1200 / 1400
1400 / 1600
1600 / 1800
1800 / 2000
2000 et plus

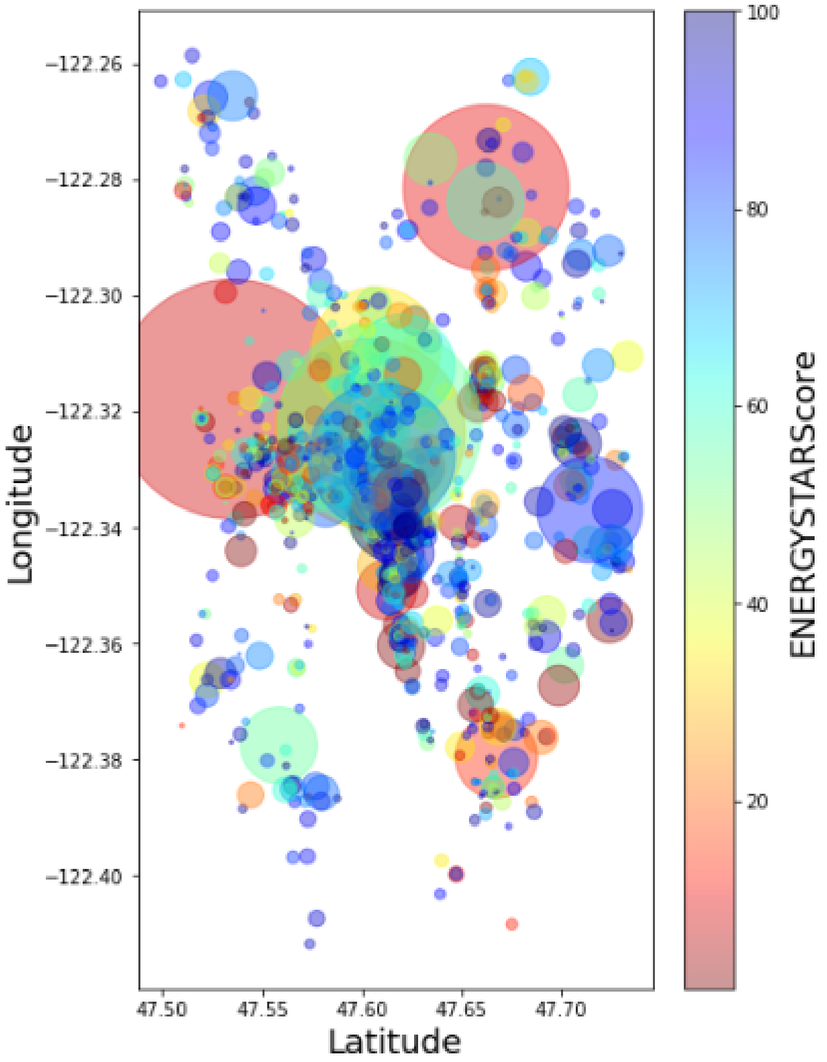

Répartition de la variable 'SiteEnergyUseWN(kBtu)'
Quantité
617
356
155
106
76
38
35
36
27
20
137
0 / 2000000
2000000 / 4000000
4000000 / 6000000
6000000 / 8000000
8000000 / 10000000
10000000 / 12000000
12000000 / 14000000
14000000 / 16000000
16000000 / 18000000
18000000 / 20000000
20000000 et plus

5 - Analyse multivariée

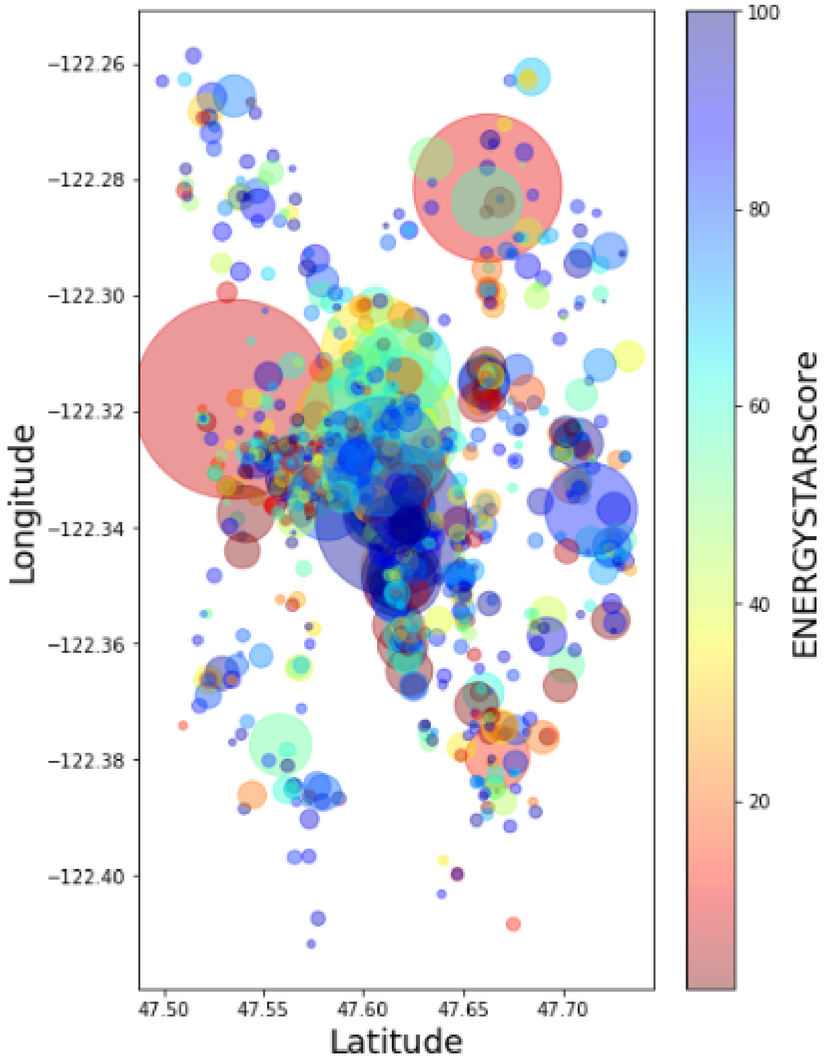


5 - Analyse multivariée

TotalGHGEmissions / ENERGYSTARScore

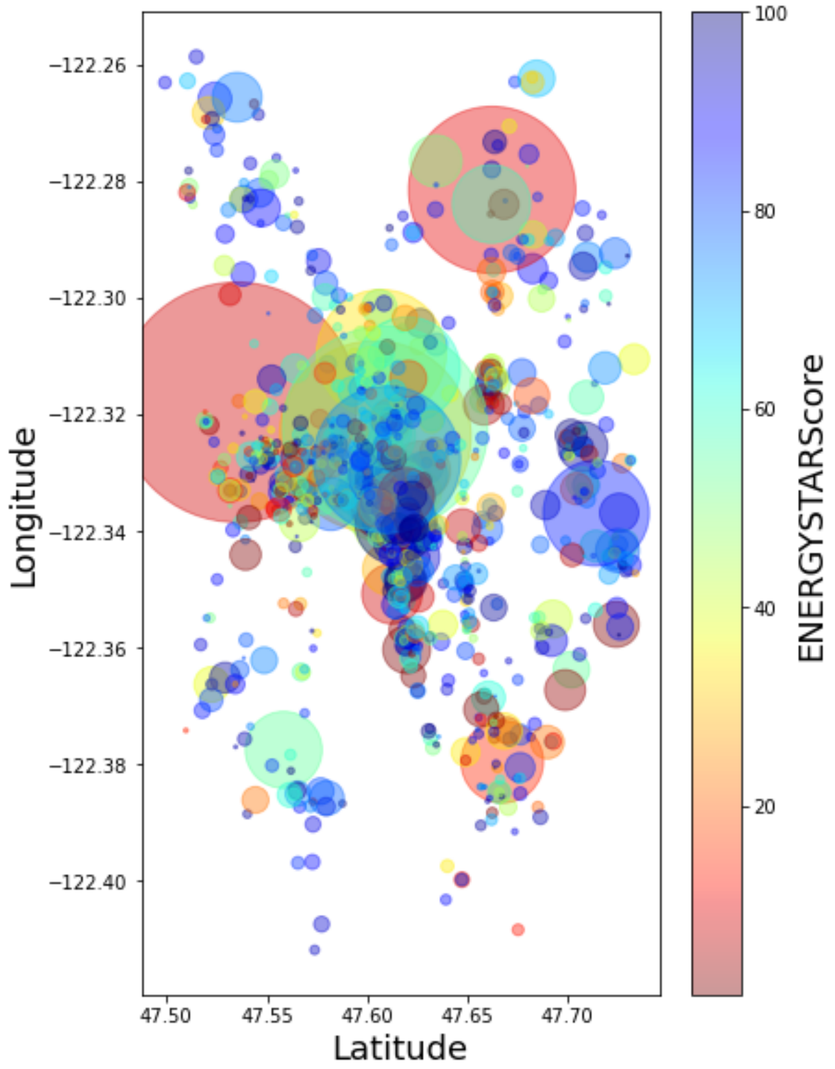


SiteEnergyUseWN(kBtu) / ENERGYSTARScore

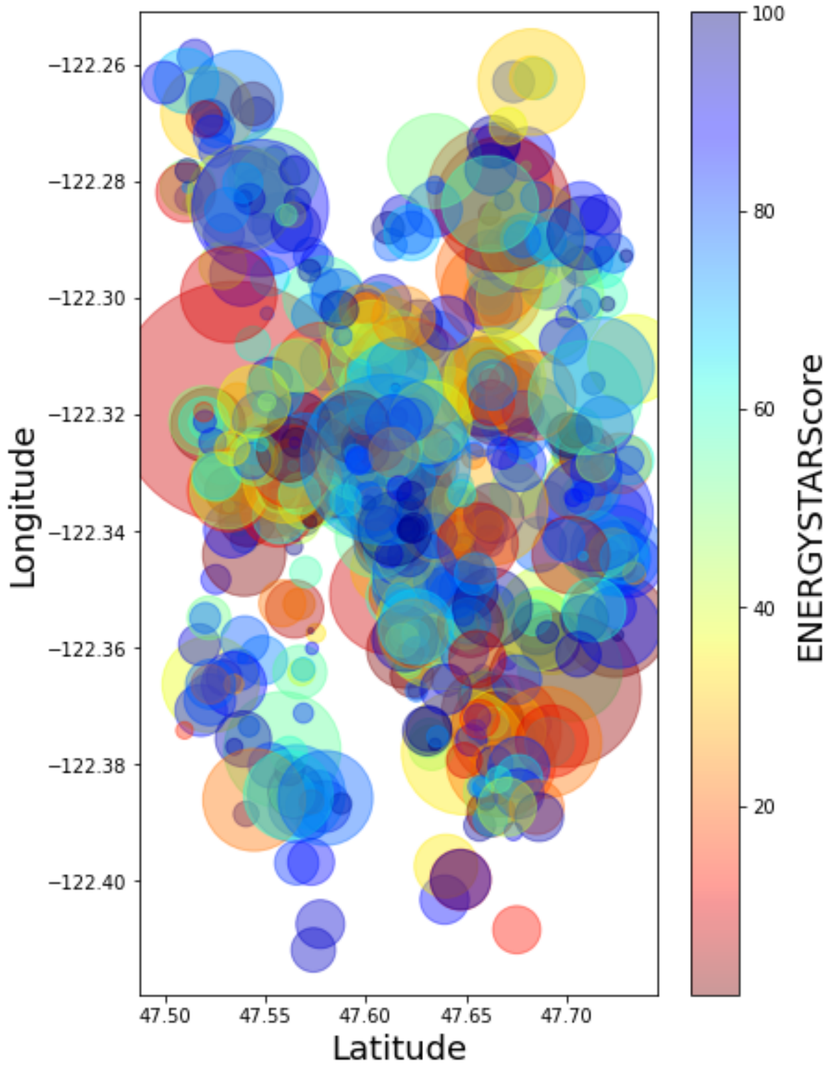


5 - Analyse multivariée

TotalGHGEmissions / ENERGYSTARScore



GHGEmissionsIntensity / ENERGYSTARScore



Présentation du Feature Engineering

1 - Variable "YearBuilt"

Remplacement par l'âge du bâtiment en années

2 - Variable "NumberofFloors"

Passage au log : application de la fonction np.log(1 + x)

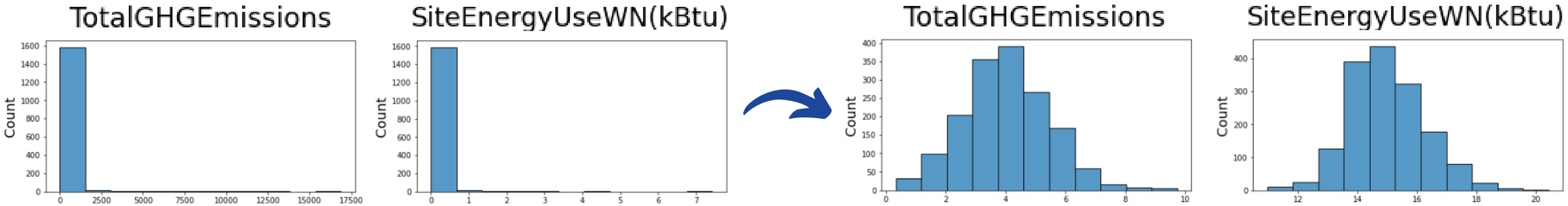
3 - Suppression de certaines variables

"NumberofBuildings"
"Neighborhood"
"BuildingType"
"Latitude" et "Longitude"

Présentation du Feature Engineering

4 - Variables énergétiques

Application de la fonction logarithme



Calcul des proportions des sources d'énergie utilisées :

- "ElectricityProp"
- "NaturalGasProp"
- "SteamUseProp"

Présentation du Feature Engineering

5 - Variables surfaciques

Standardisation des surfaces en divisant par la surface totale

Création d'un encodage des surfaces normalisées pour :
- "LargestPropertyUseType"
- "SecondLargestPropertyUseType"
- "ThirdLargestPropertyUseType"

Objectif : créer de nombreuses colonnes qui correspondent aux différents types de biens et y affecter pour chaque observation le rapport entre la surface utilisée pour ce bien et la surface totale

Hotel
1.000000
0.809918
0.791220
1.000000
0.703070

Prédiction des émissions de CO_2

1 - Modélisations

StandardScaler sur l'ensemble des variables quantitatives

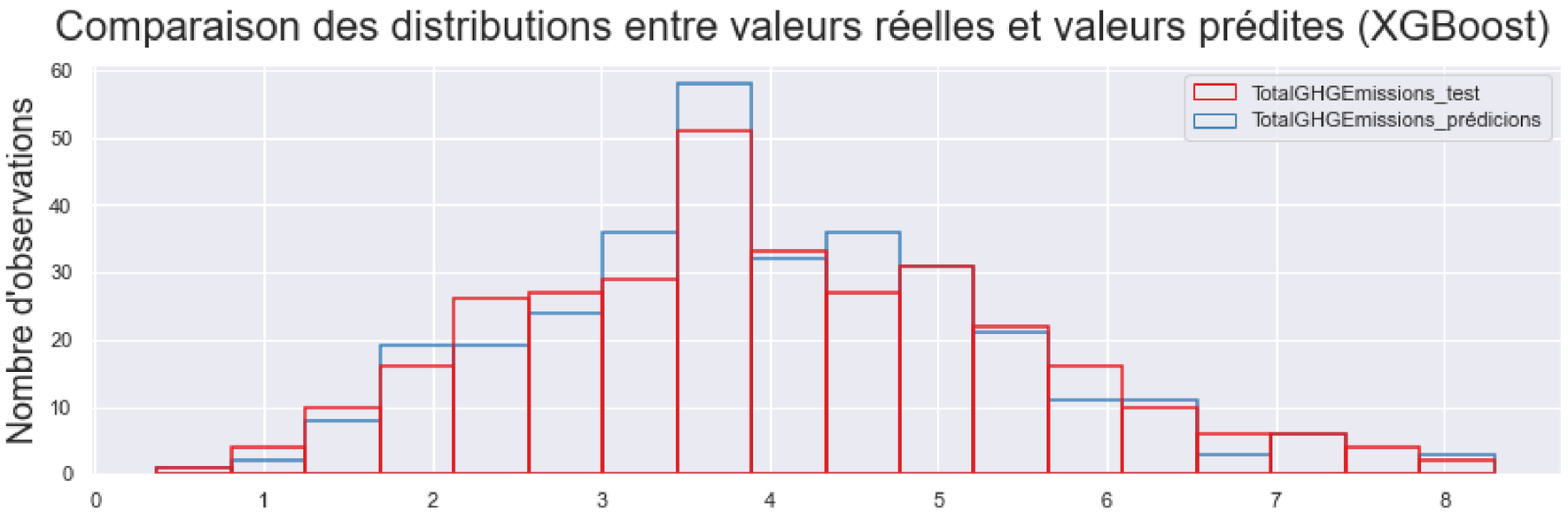
Régressions sans CV et sans réglage des hyperparamètres avec :

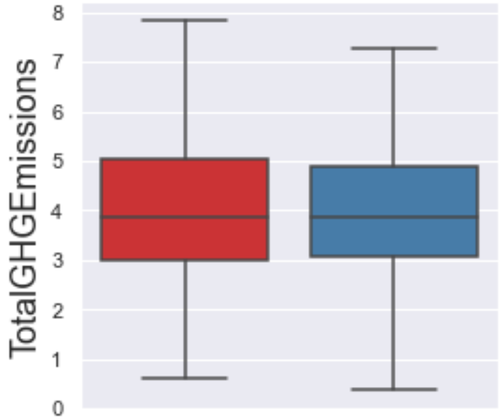
- DummyRegressor
- RandomForestRegressor
- GradientBoostingRegressor
- ElasticNet
- ExtraTreesRegressor
- SVR
- AdaBoostRegressor
- XGBRegressor
- KerasRegressor

Prédiction des émissions de CO_2

1 - Modélisations

	DummyRegressor	RandomForest	GradientBoosting	ElasticNet	ExtraTrees	SupportVector	AdaBoost	XGBoost	NeuralNetwork
R^2 Train	0.000	0.973	0.901	0.256	1.000	0.857	0.749	0.987	0.945
R^2 Test	-0.000	0.801	0.820	0.263	0.809	0.780	0.707	0.833	0.801
M.A.E.	1.176	0.498	0.480	1.006	0.487	0.489	0.624	0.465	0.491
R.M.S.E.	1.466	0.653	0.621	1.258	0.641	0.687	0.793	0.599	0.654

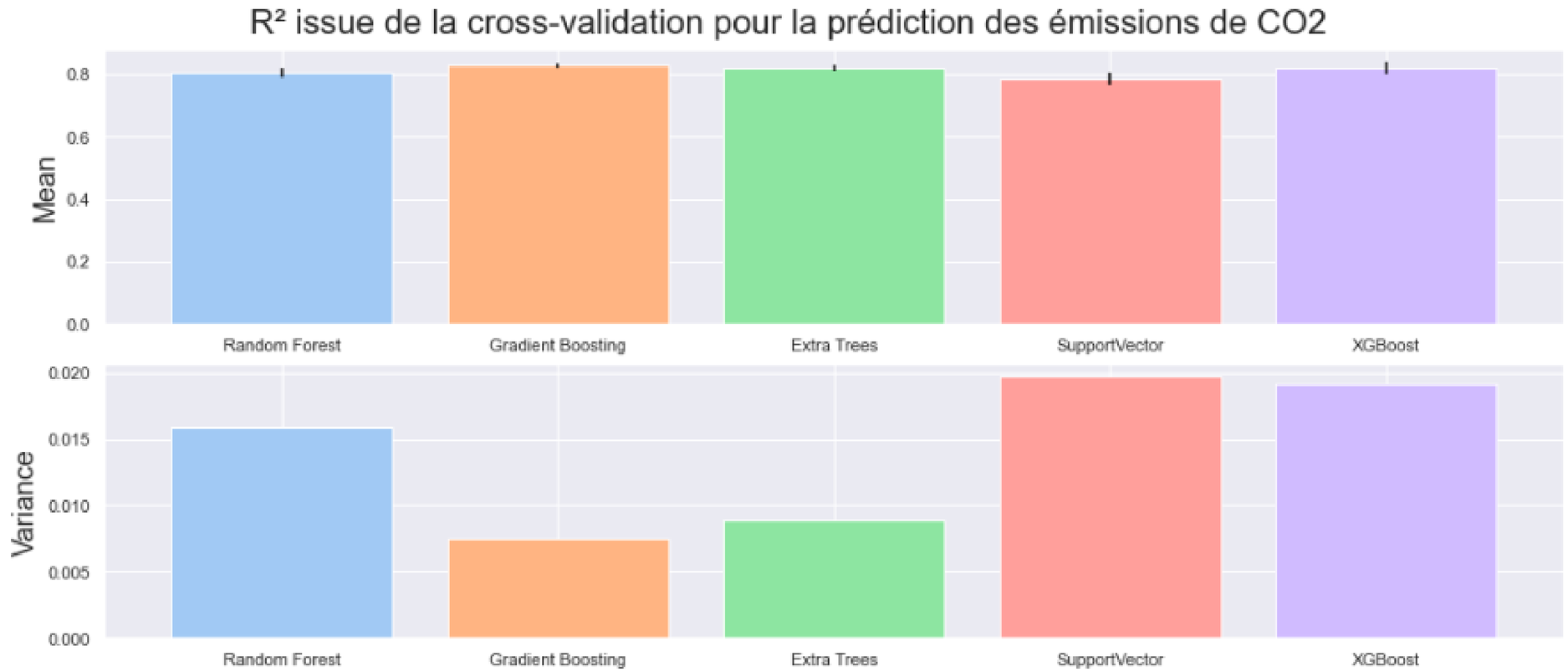




Prédiction des émissions de CO_2

1 - Modélisations

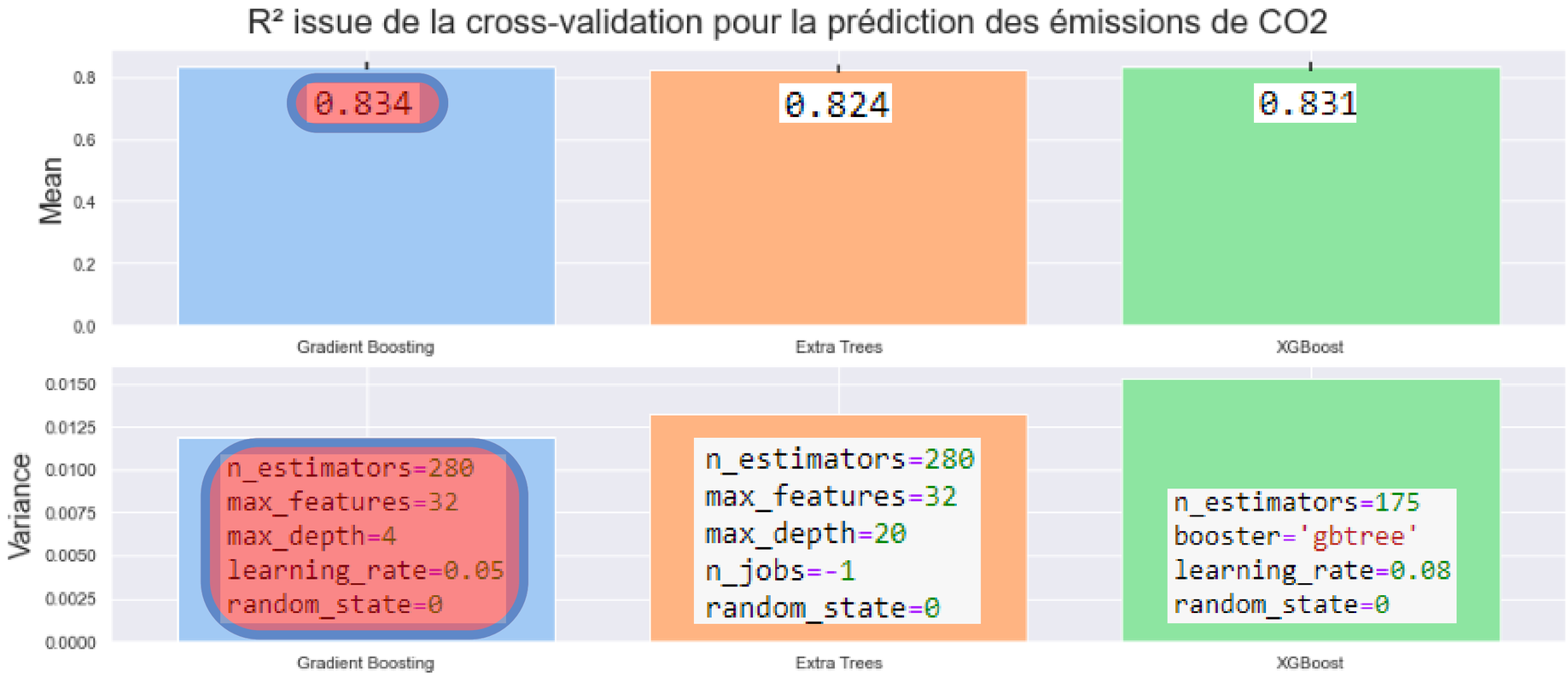
Cross-Validation avec 5 modèles



Prédiction des émissions de CO_2

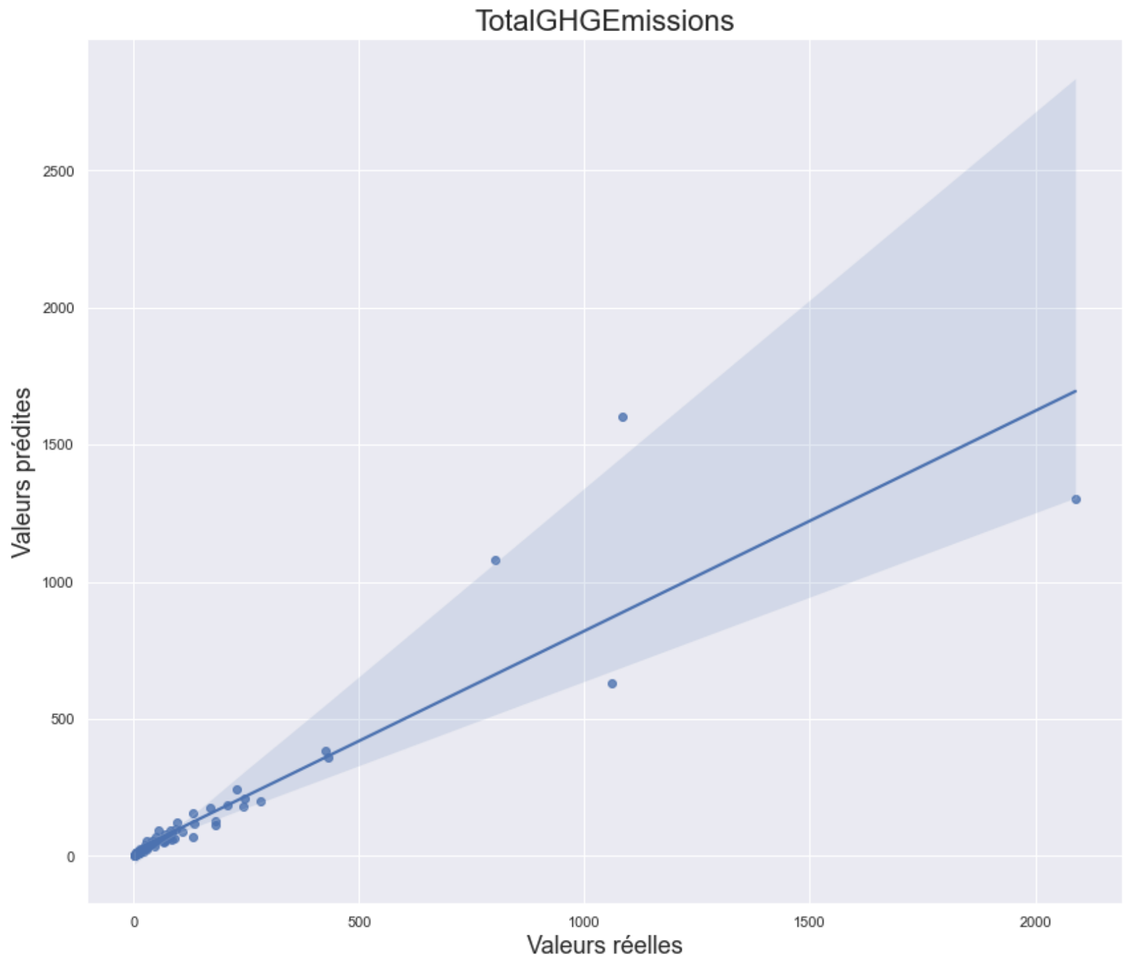
1 - Modélisations

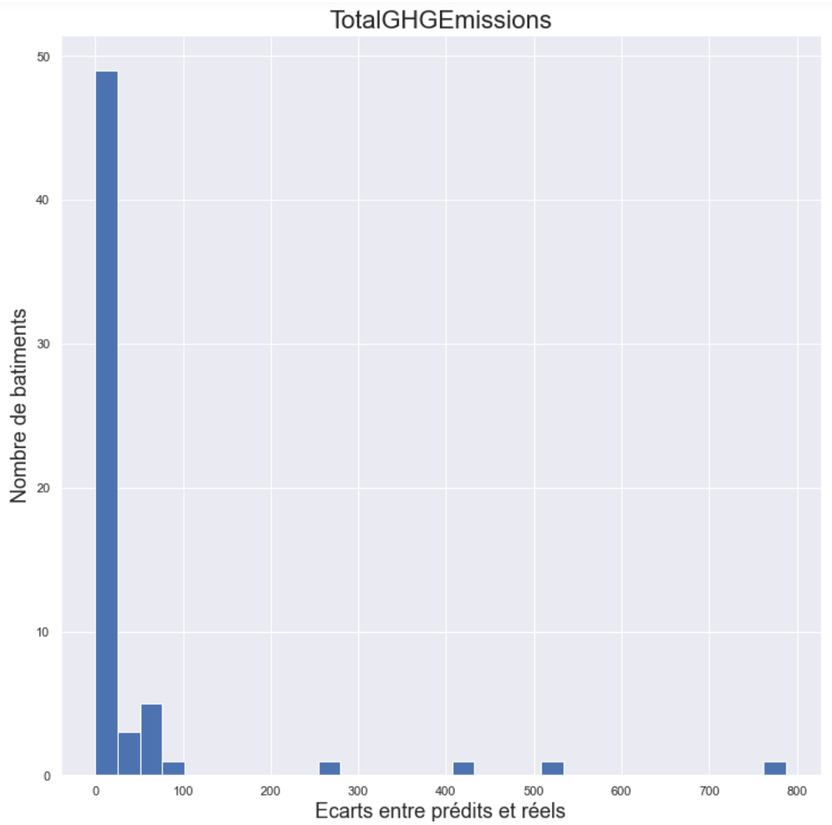
Hyperparamétrage de trois modèles



Prédiction des émissions de CO_2

2 - Présentation des résultats





Modèle performant pour les bâtiments à faible émissions

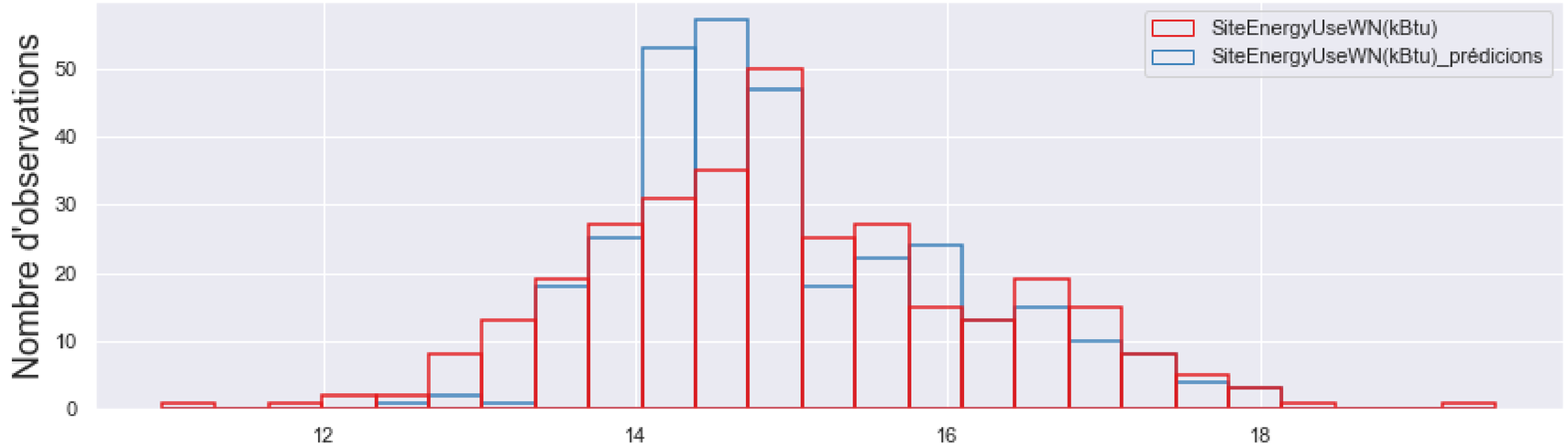
Prédiction de la consommation totale d'énergie

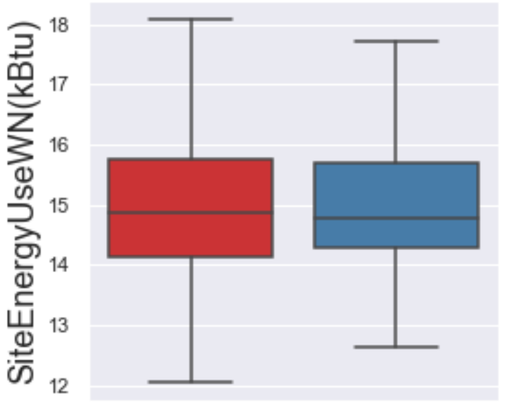
1 - Modélisations

Même amorçage que pour la prédiction des émissions de CO_2

	DummyRegressor	RandomForest	GradientBoosting	ElasticNet	ExtraTrees	SupportVector	AdaBoost	XGBoost
R² Train	0.000	0.965	0.866	0.338	1.000	0.817	0.701	0.984
R² Test	-0.000	0.724	0.748	0.321	0.714	0.733	0.630	0.713
M.A.E.	0.999	0.503	0.480	0.819	0.496	0.474	0.594	0.494
R.M.S.E.	1.274	0.669	0.639	1.050	0.681	0.659	0.775	0.682

Comparaison des distributions entre valeurs réelles et valeurs prédites (GradientBoosting)

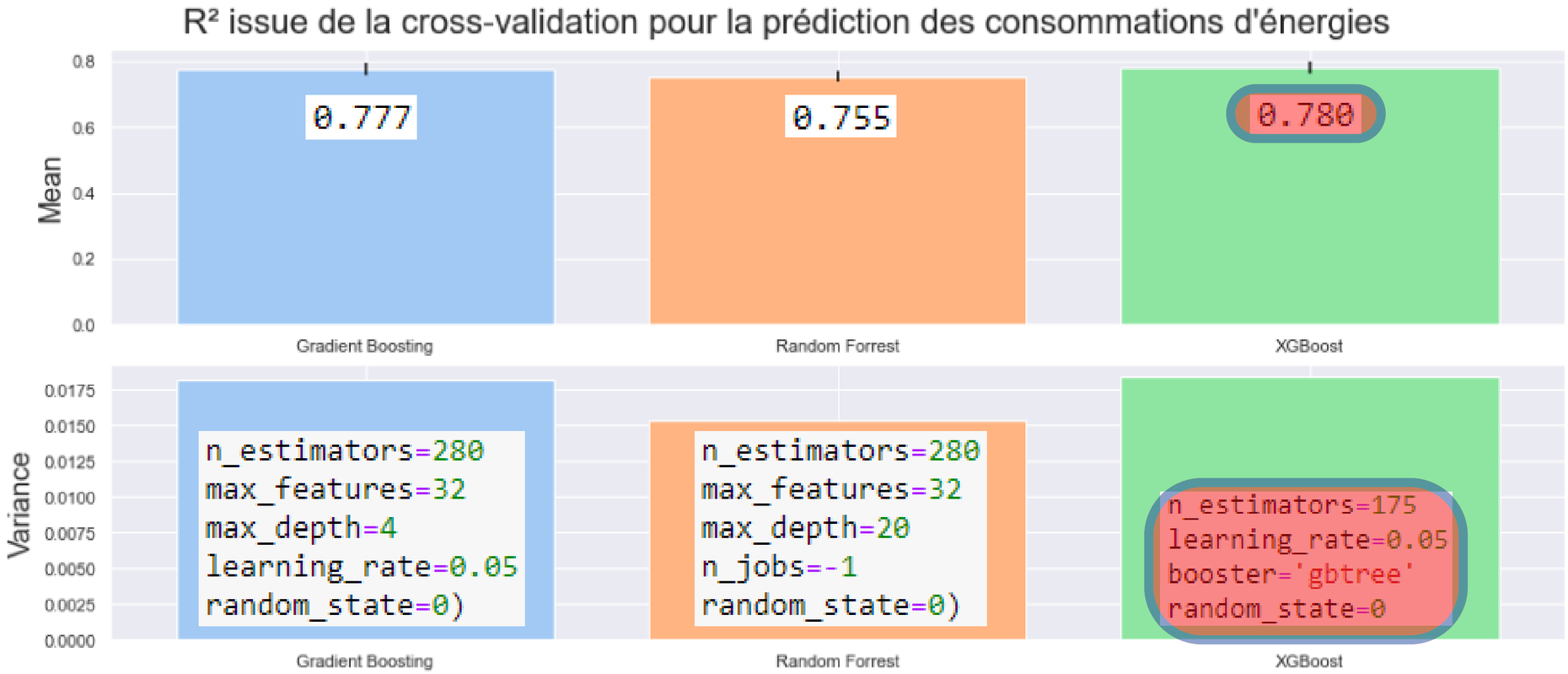




Prédiction de la consommation totale d'énergie

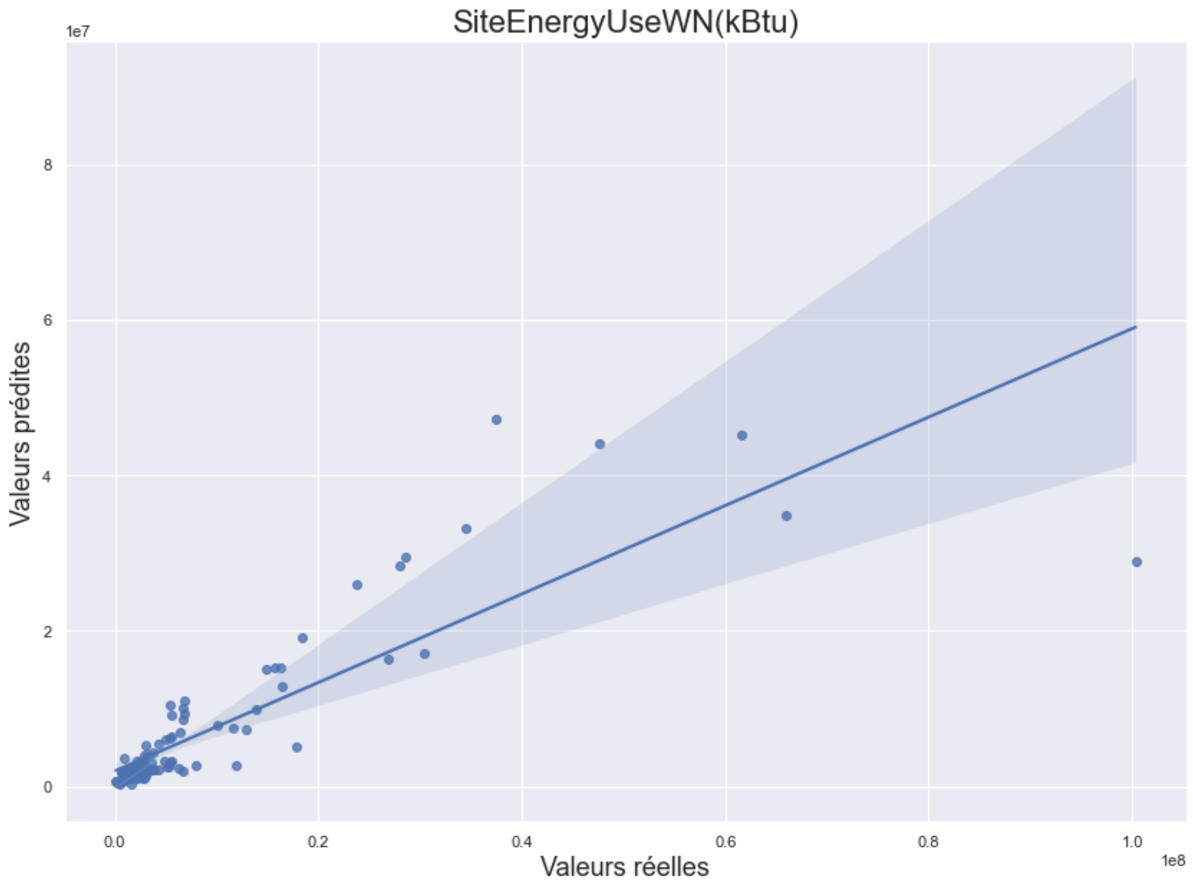
1 - Modélisations

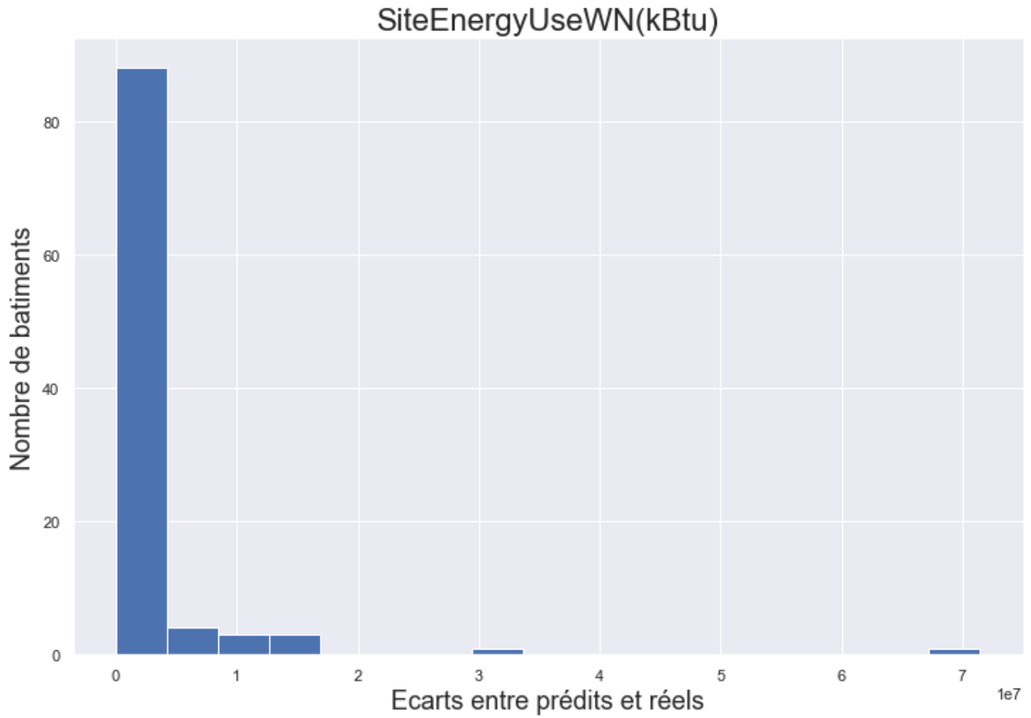
Après Cross-Validation et hyperparamètrage sur 3 modèles :



Prédiction de la consommation totale d’énergie

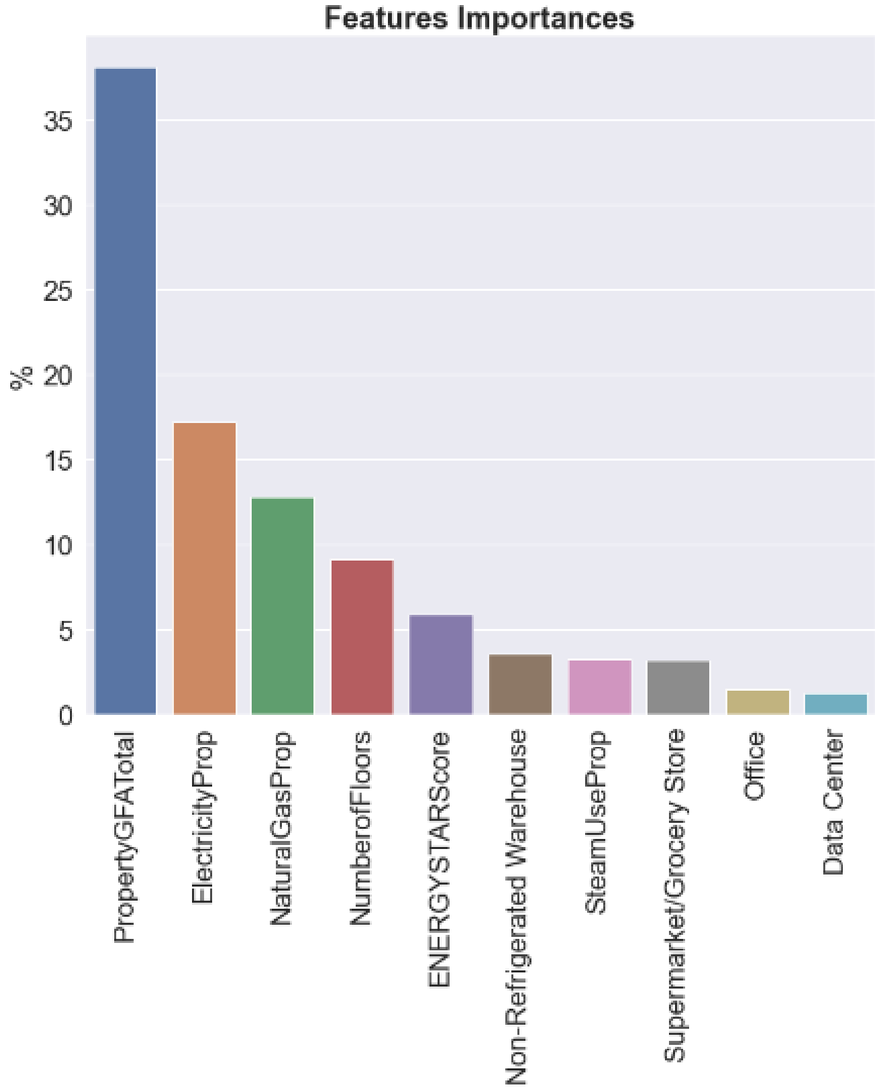
2 - Présentation des résultats

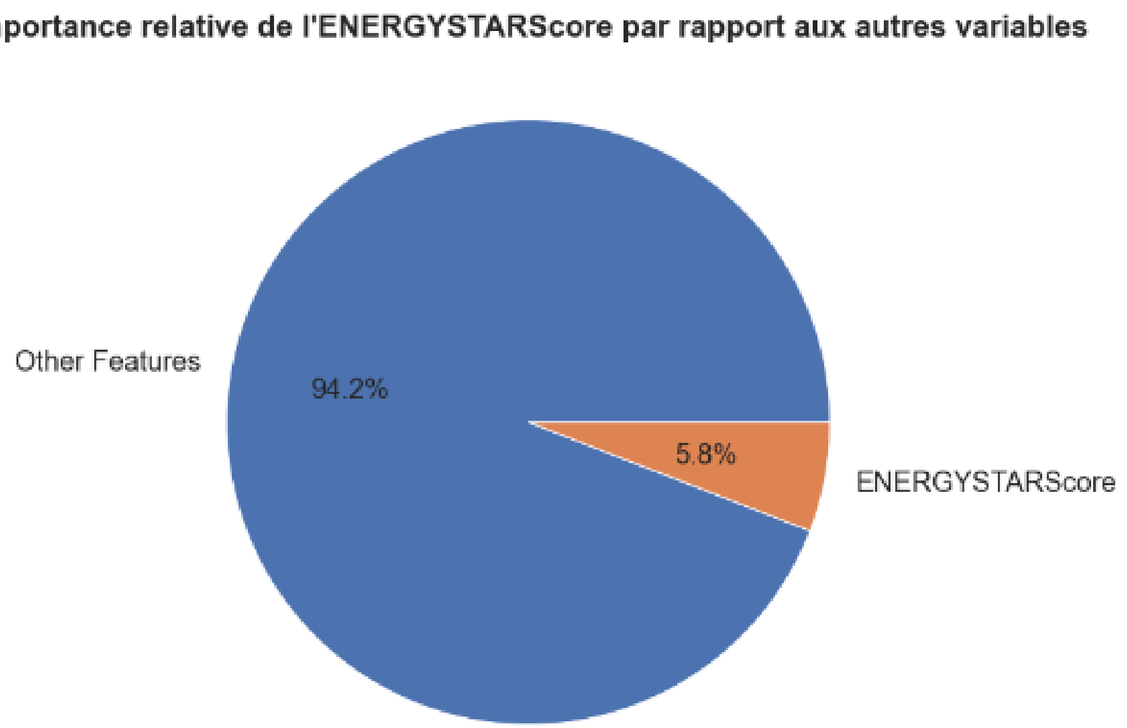




Modèle performant pour les bâtiments à faible consommation

Évaluation de l’intérêt de l’"ENERGY STAR Score"





Évaluation de l’intérêt de l’"ENERGY STAR Score"

SHAP : SHapley Additive exPlanations