Projet 4 : Anticipez les besoins en consommation de bâtiments

Oumeima EL GHARBI

OpenClassrooms – Data Scientist

Soutenance : 25/09/2022

Plan

Introduction

- Problématique
- Présentation du jeu de données

1. Preprocessing

- Nettoyage
- Analyse exploratoire
- Feature engineering

2. Modélisation Energie

- Baseline : Dummy, Régression Linéaire
- Modèles linéaires
- Méthodes ensemblistes
- Optimisation des hyperparamètres
- Evaluation

3. Modélisation CO2

- Baseline
- Modèles linéaires
- Méthodes ensemblistes
- Optimisation des hyperparamètres
- Evaluation

Conclusion

- Intêret de l'Energy Star Score

Introduction

Problématique :

« Votre équipe s'intéresse de près à la consommation et aux émissions des bâtiments non destinés à l'habitation.

Ces relevés sont coûteux à obtenir, et à partir de ceux déjà réalisés, vous voulez tenter de prédire les émissions de CO2 et la consommation totale d'énergie de bâtiments non destinés à l'habitation pour lesquels elles n'ont pas encore été mesurées.

Votre prédiction se basera sur les données structurelles des bâtiments (taille et usage des bâtiments, date de construction, situation géographique, ...) »

Vous cherchez également à évaluer l'intérêt de **l'ENERGY STAR Score** pour la prédiction d'émissions, qui est fastidieux à calculer avec l'approche utilisée actuellement par votre équipe. Vous l'intégrerez dans la modélisation et jugerez de son intérêt.

Implémentation :

Cadre : apprentissage supervisé, étiquettes connues.

Problème de **régression** : les étiquettes sont variables numériques

Modèles linéaires : Régression Linéaire, Régression Ridge, LASSO, Elastic Net

Méthodes ensemblistes : Forêts aléatoires, Gradient Boosting

Optimisation des **hyper-paramèteres** : GridSearch, RandomSearch

Validation Croisée

Evaluation : MSE, RMSE, MAE, R2 score

Introduction

Jeu de données :

Dataset de la ville de Seattle : 2016_Building_Energy_Benchmarking.csv

3376 bâtiments (uniques), 46 colonnes / features.

Stratégie :

- Prédire à l'aide des caractéristiques des bâtiments :
  - leurs consommations en énergie : SiteEnergyUseWN(kBtu)
  - leurs émissions en CO2 : TotalGHGEmissions

- On prédit à nouveau le CO2 en ajoutant la variable ENERGY STAR Score : observation de son effet sur les predictions.

Preprocessing

Nettoyage

- Retirer les bâtiments résidentiels
- Correction du nombre de bâtiments (si nul)
- Retirer colonnes pas utiles
- Retirer les bâtiments non conformes
- Imputation des colonnes PropertyUseType
- Retirer bâtiments avec des valeurs manquantes
- Catégorisation des variables
- Mapping des chaînes de caractères
- Traitement de la variable « Neighborhood »
- Mapping de BuildingType
- Retirer les bâtiments ayant des valeurs énergétiques négatives
- Vérification de PropertyGFATotal
- Traitement des Outliers (0.5% extrémités)

Exploration

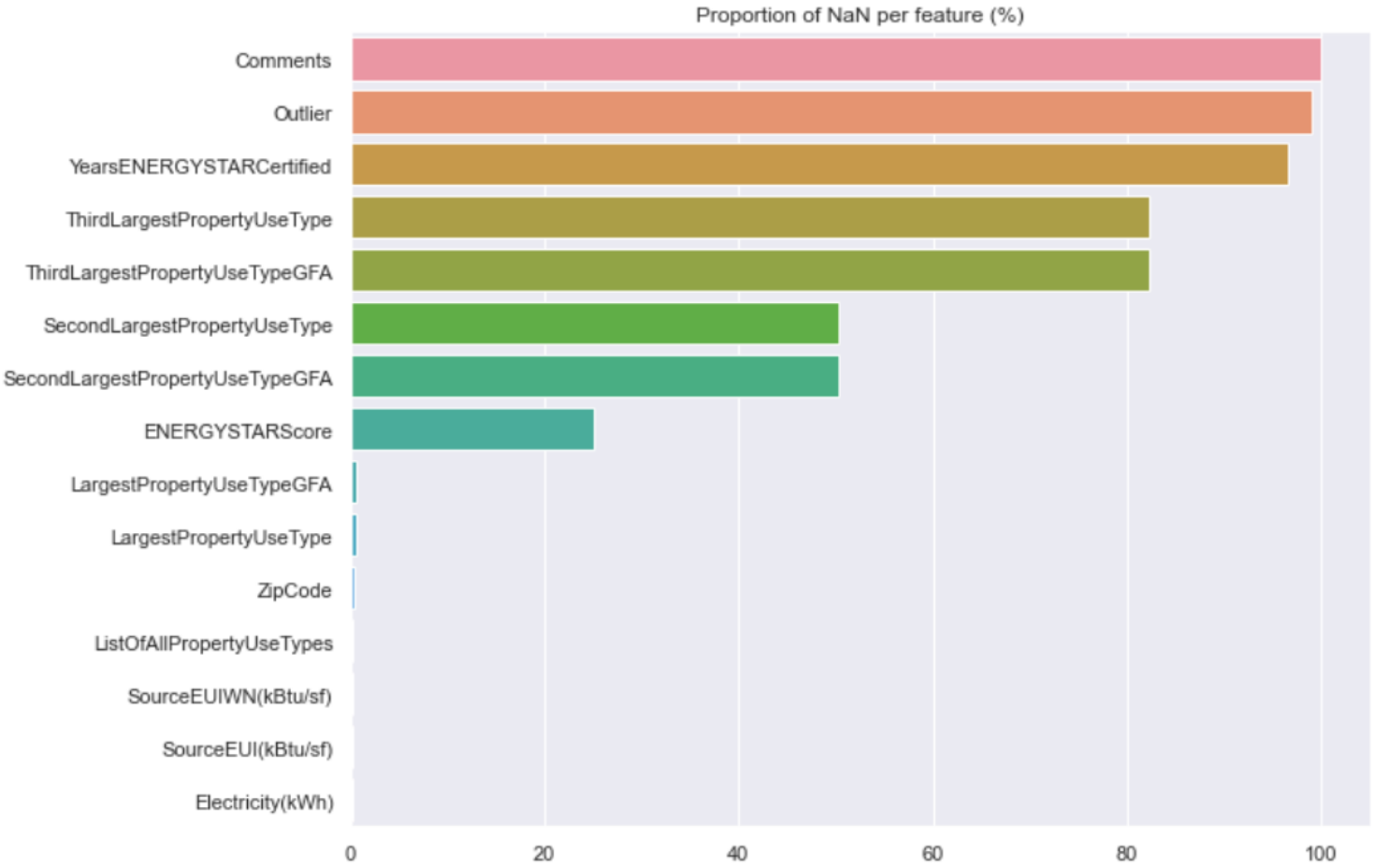
Feature engineering

- Choix des variables pour la modélisation avec matrice de corrélation
- Transformation Logarithmique
- Sélection des variables pour la modélisation

- Standardisation des variables numériques
- Encodage des variables catégorielles en valeurs binaires

Preprocessing
1) Nettoyage

Retirer colonnes pas utiles : Comments, YearsENERGYSTARCertified, etc



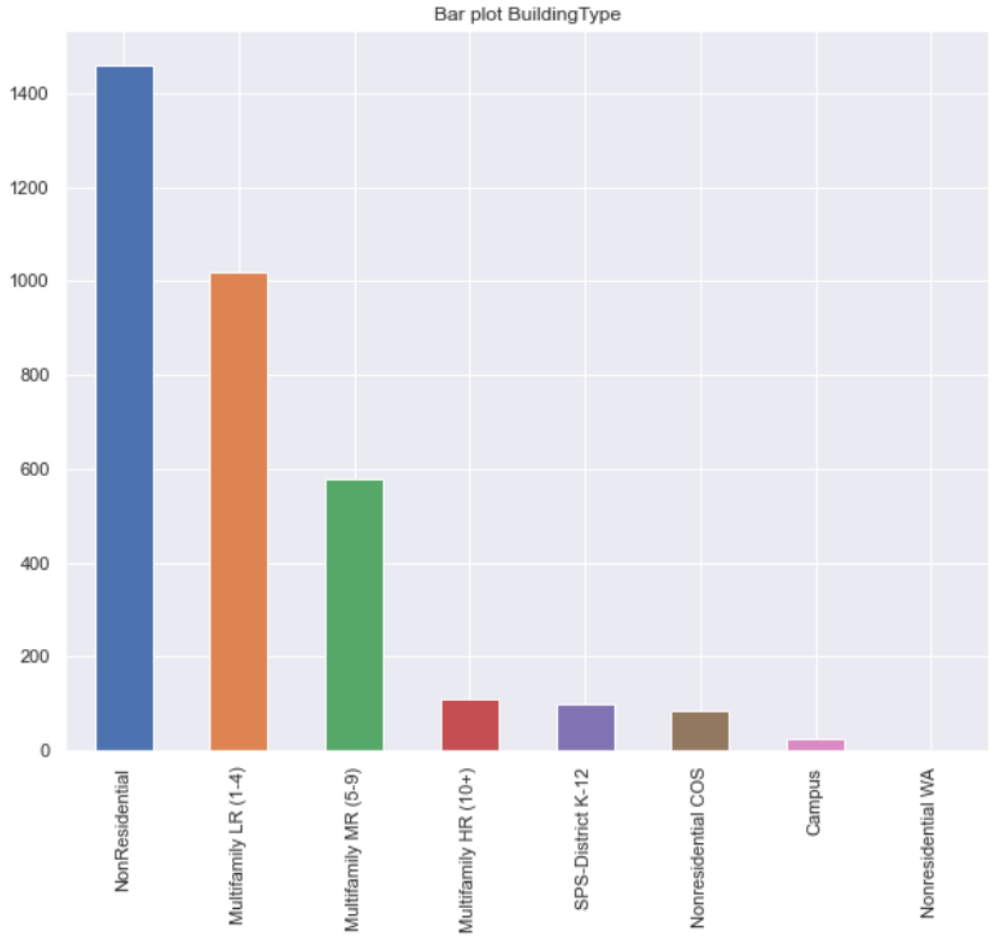
Preprocessing
1) Nettoyage

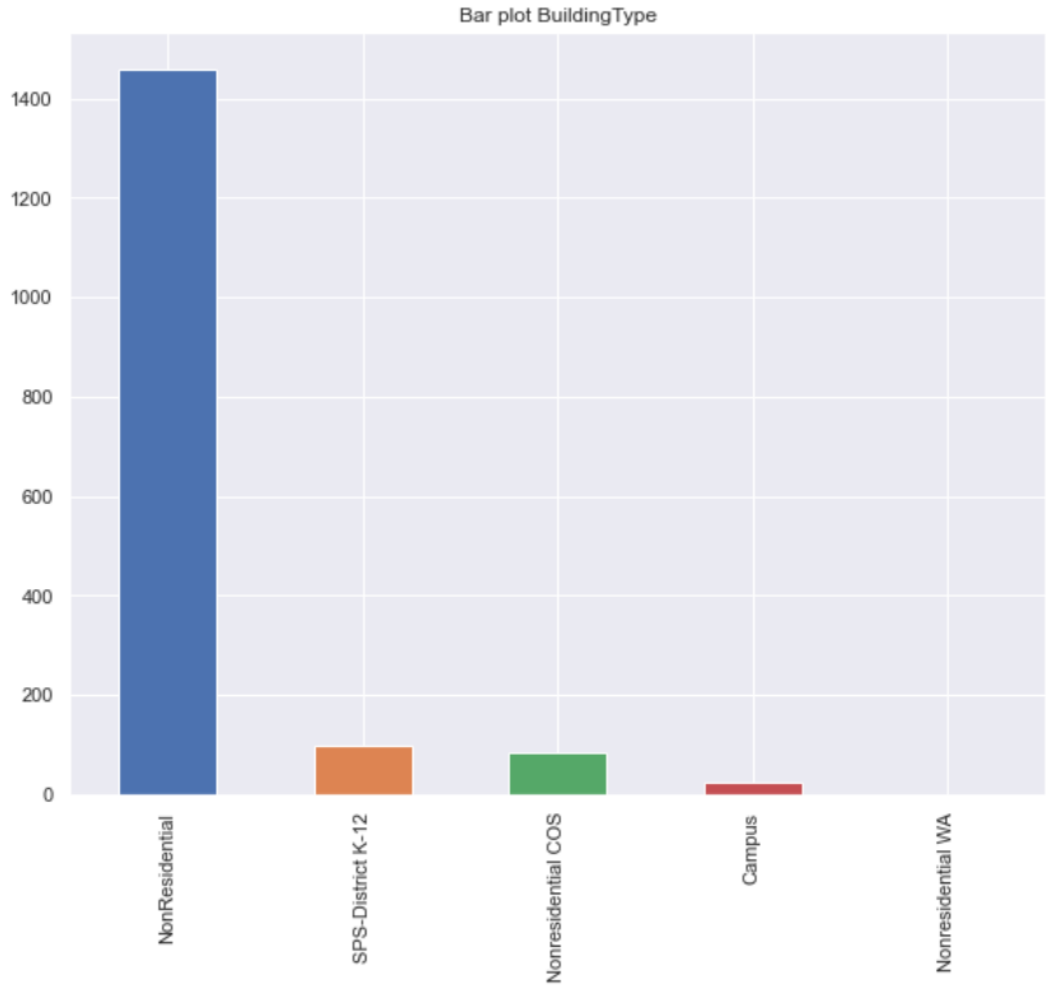
- Valeurs énergétiques négatives : valeurs aberrantes, on supprime ces bâtiments.

	count	mean	std	min	25%	50%	75%	max
Electricity(kBtu)	3367.00	3707612.16	14850656.14	-115417.00	639487.00	1177583.00	2829632.50	657074389.00
Electricity(kWh)	3367.00	1086638.97	4352478.36	-33826.80	187422.95	345129.91	829317.84	192577488.00
Longitude	3376.00	-122.33	0.03	-122.41	-122.35	-122.33	-122.32	-122.22
SourceEUIWN(kBtu/sf)	3367.00	137.78	139.11	-2.10	78.40	101.10	148.35	2620.00
TotalGHGEmissions	3367.00	119.72	538.83	-0.80	9.50	33.92	93.94	16870.98
GHGEmissionsIntensity	3367.00	1.18	1.82	-0.02	0.21	0.61	1.37	34.09
NaturalGas(kBtu)	3367.00	1368504.54	6709780.83	0.00	0.00	323754.00	1189033.50	297909000.00
NaturalGas(therms)	3367.00	13685.05	67097.81	0.00	0.00	3237.54	11890.33	2979090.00
SteamUse(kBtu)	3367.00	274595.90	3912173.39	0.00	0.00	0.00	0.00	134943456.00
SiteEnergyUseWN(kBtu)	3370.00	5276725.71	15938786.48	0.00	970182.23	1904452.00	4381429.12	471613856.00
SiteEnergyUse(kBtu)	3371.00	5403667.29	21610628.63	0.00	925128.59	1803753.25	4222455.25	873923712.00
SourceEUI(kBtu/sf)	3367.00	134.23	139.29	0.00	74.70	96.20	143.90	2620.00
SiteEUIWN(kBtu/sf)	3370.00	57.03	57.16	0.00	29.40	40.90	64.28	834.40
SiteEUI(kBtu/sf)	3369.00	54.73	56.27	0.00	27.90	38.60	60.40	834.40
ThirdLargestPropertyUseTypeGFA	596.00	11738.68	29331.20	0.00	2239.00	5043.00	10138.75	459748.00
PropertyGFAParking	3376.00	8001.53	32326.72	0.00	0.00	0.00	0.00	512608.00
NumberofFloors	3376.00	4.71	5.49	0.00	2.00	4.00	5.00	99.00
NumberofBuildings	3368.00	1.11	2.11	0.00	1.00	1.00	1.00	111.00
SecondLargestPropertyUseTypeGFA	1679.00	28444.08	54392.92	0.00	5000.00	10664.00	26640.00	686750.00
ENERGYSTARScore	2533.00	67.92	26.87	1.00	53.00	75.00	90.00	100.00
CouncilDistrictCode	3376.00	4.44	2.12	1.00	3.00	4.00	7.00	7.00

Preprocessing
1) Nettoyage

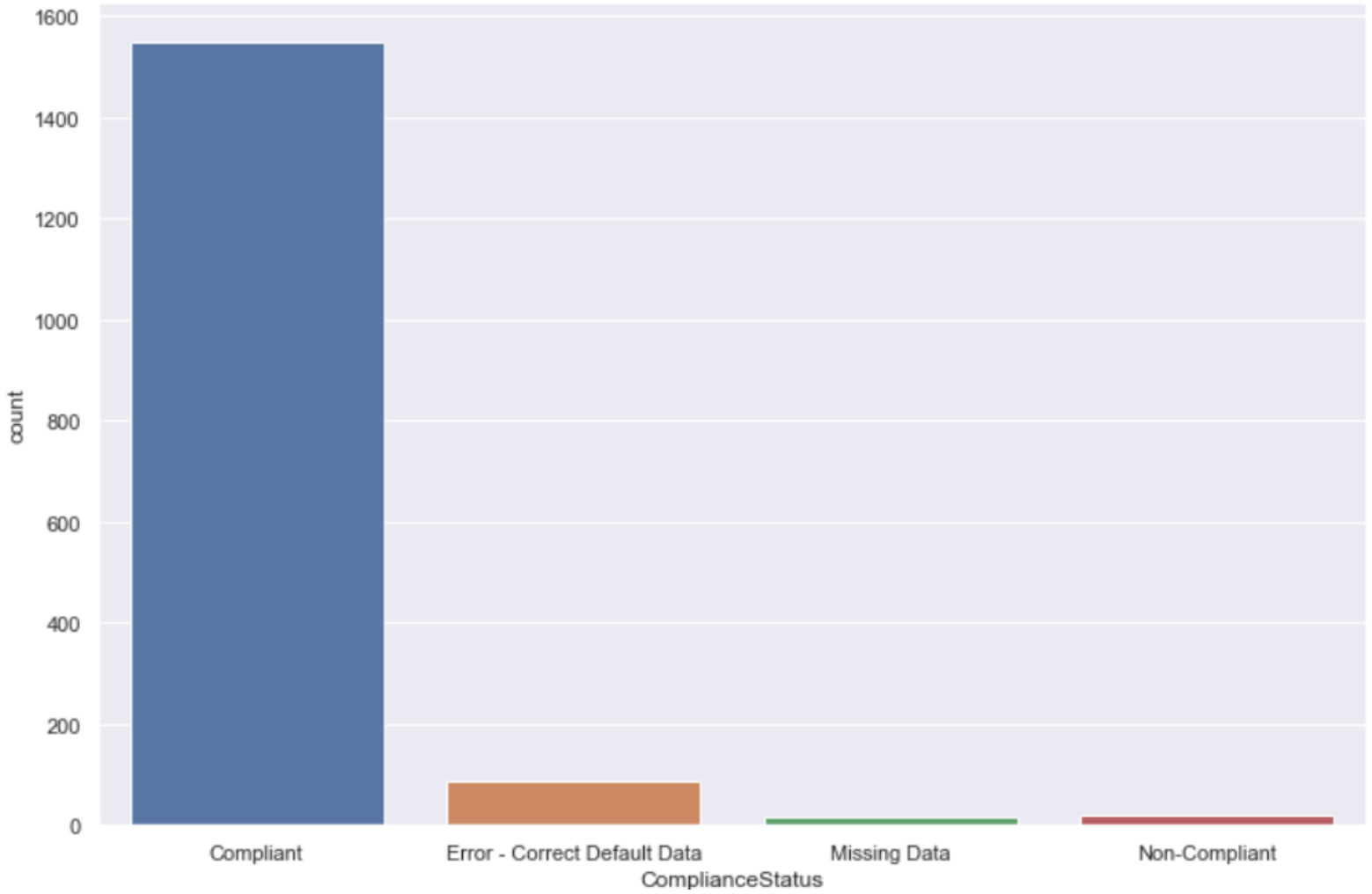
- On retire les bâtiments résidentiels : il reste 1668 bâtiments





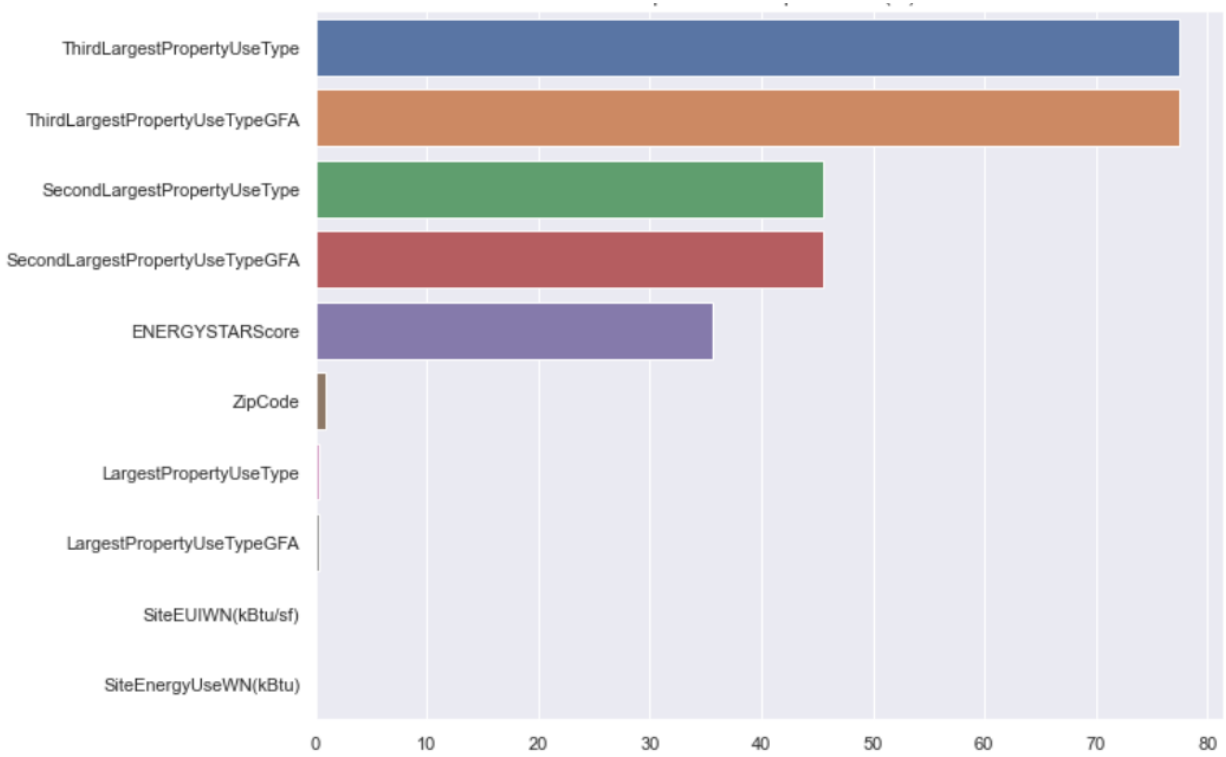
Preprocessing
1) Nettoyage

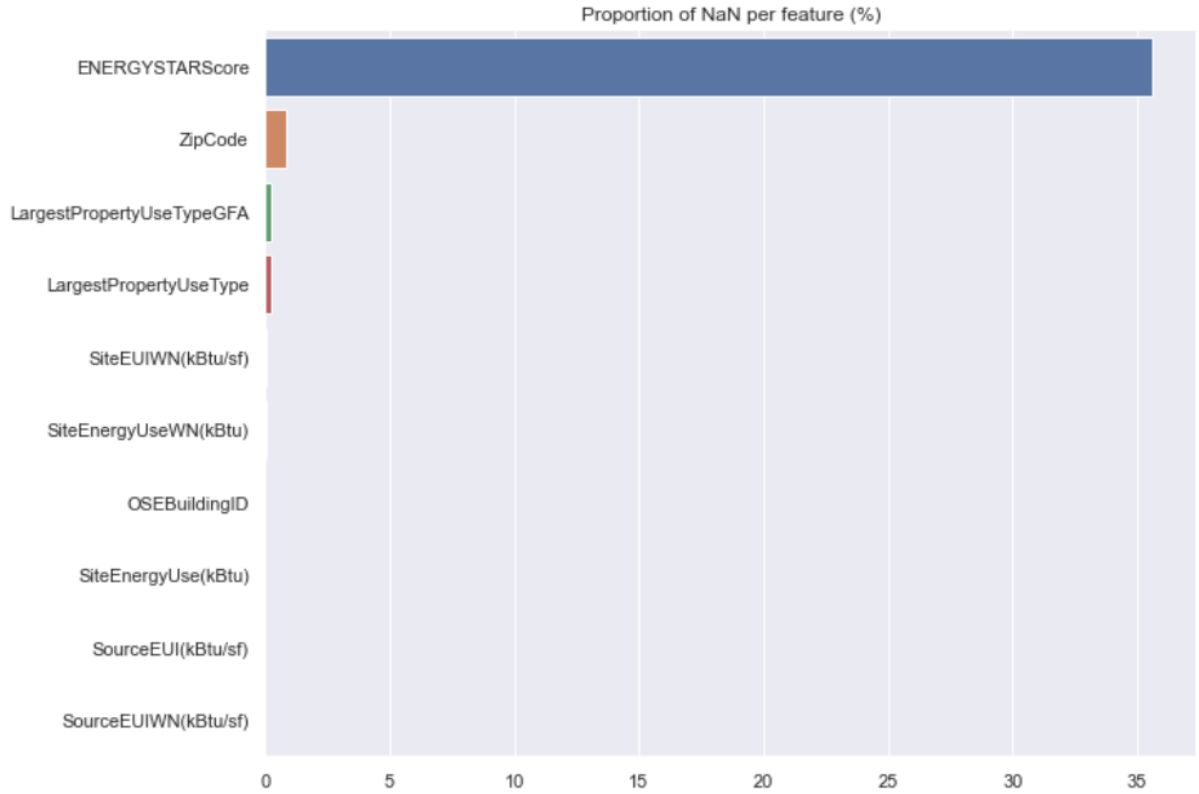
Bâtiments non conformes dont 32 outliers : suppression



Preprocessing
1) Nettoyage

Imputation des variables Second / ThirdLargestPropertyUseType(GFA)





Preprocessing
1) Nettoyage

- On calcule le nombre d’années depuis la construction de chaque bâtiment

```
___Computing the years since the buildings were built___
Before : (1508, 32)
After : (1508, 33)
2022 <class 'int'>
```

	YearSinceBuilt	YearBuilt
0	95	1927
1	26	1996
2	53	1969
3	96	1926
4	42	1980
...	...	...
3339	93	1929
3340	9	2013
3347	7	2015
3356	7	2015
3359	60	1962

Preprocessing
1) Nettoyage

- On retire les bâtiments dont le SiteEnergyUseWN(kBtu) ou TotalGHGEmissions est supérieur au quantile 0.995 ou inférieur au quantile 0.05
- =>> On retire 0.5% des bâtiments aux extrémités.

- Avant :

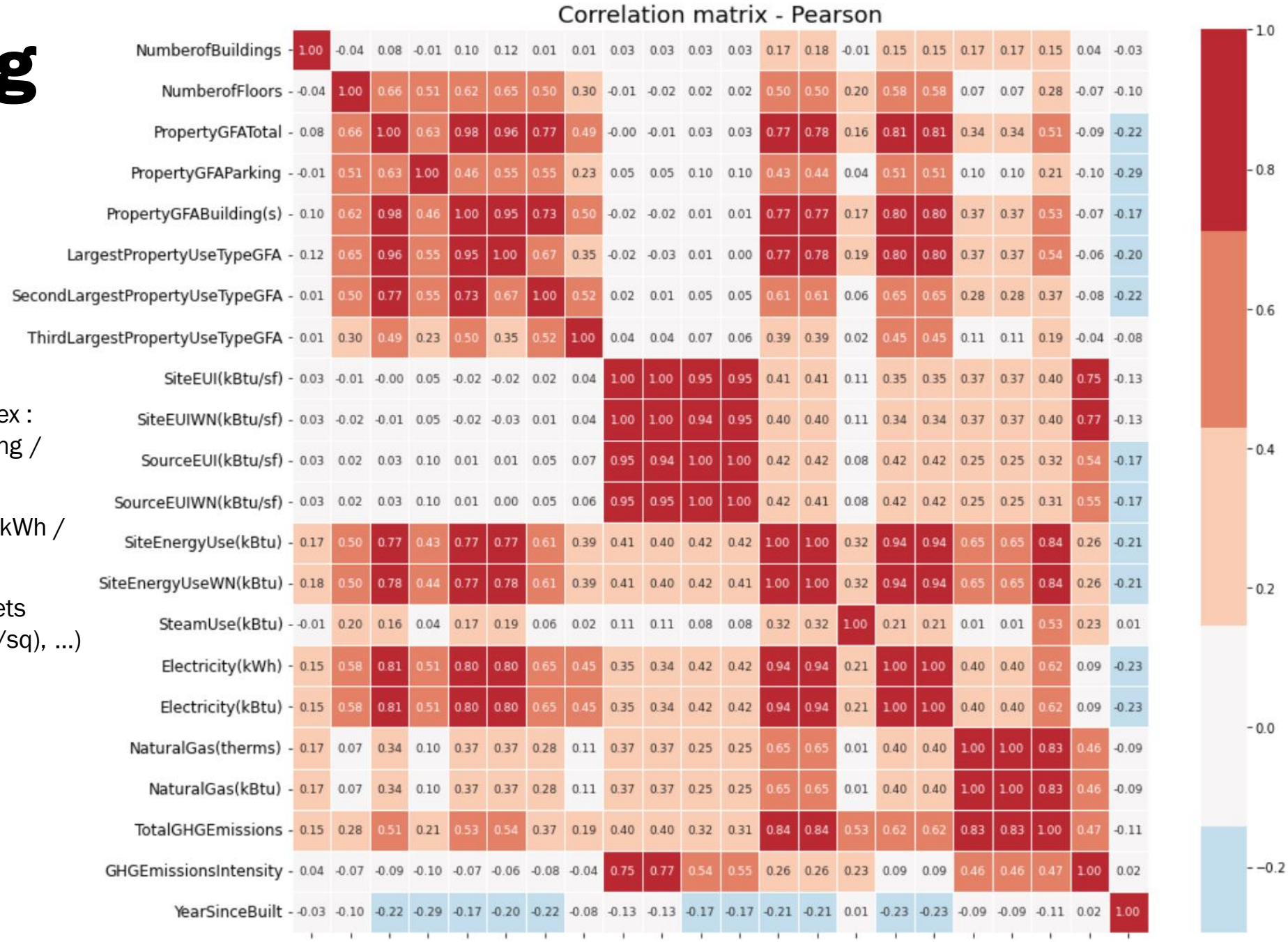
	count	mean	std	min	25%	50%	75%	max
TotalGHGEmissions	1508.00	188.16	736.64	0.40	20.42	49.94	146.95	16870.98
SiteEnergyUseWN(kBtu)	1508.00	8500286.98	23012684.18	0.00	1331696.16	2829517.25	7541073.12	471613856.00

- Après :

	count	mean	std	min	25%	50%	75%	max
TotalGHGEmissions	1484.00	143.74	288.95	0.89	20.76	49.94	143.61	3278.11
SiteEnergyUseWN(kBtu)	1484.00	7191649.20	11971723.61	116642.50	1356709.62	2833792.75	7450241.00	123205560.00

Preprocessing 2) Analyse exploratoire

- Variables peu corrélées aux targets (NumberofBuildings)
- Variables très corrélées entre elles ; ex : PropertyGFATotal / PropertyGFAParking / PropertyGFABuilding(s)
- Suppression variables redondantes (kWh / kBtu / therms)
- Suppression variables liées aux targets (Electricity, NaturalGas, SiteEUI(kBtu/sq), ...)
- Targets : TotalGHGEmissions et SiteEnergyUseWN(kBtu)

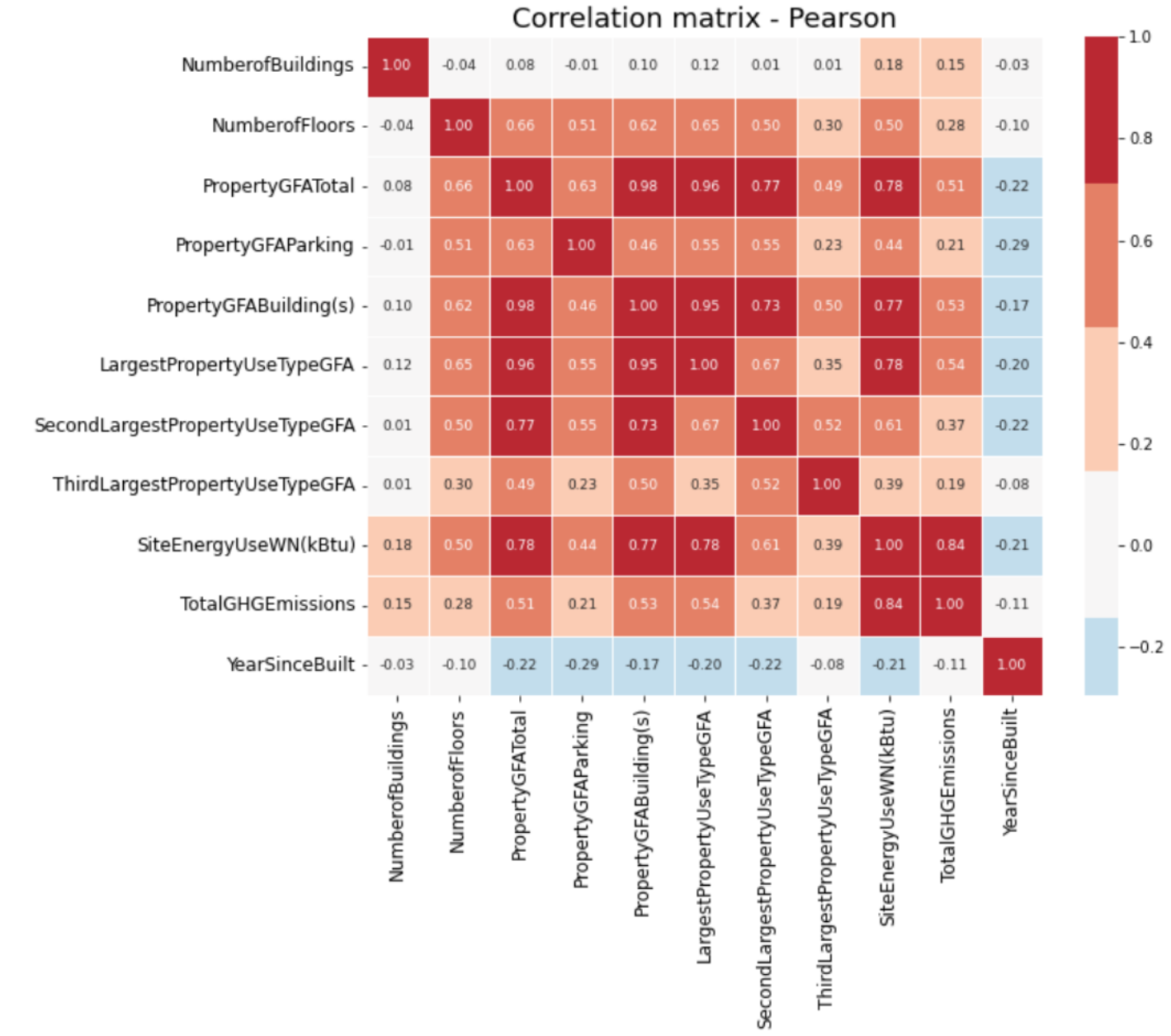


Preprocessing 2) Analyse exploratoire

On veut garder des variables structurelles des bâtiments qui sont :

- peu corrélées entre elles
- et très corrélées aux variables à prédire (SiteEnergyUseWN(kBtu) et TotalGHGEmissions

- Au final, on ne garde pas NoumberofBuildings, PropertyGFAParking et YearSinceBuilt (les prédictions ont été testées avec et sans ces features).



Preprocessing
2) Analyse exploratoire

Stratégie :

- Préparation d'un dataset d'entraînement de de test pour prédire « Log-SiteEnergyUseWN(kBtu) »

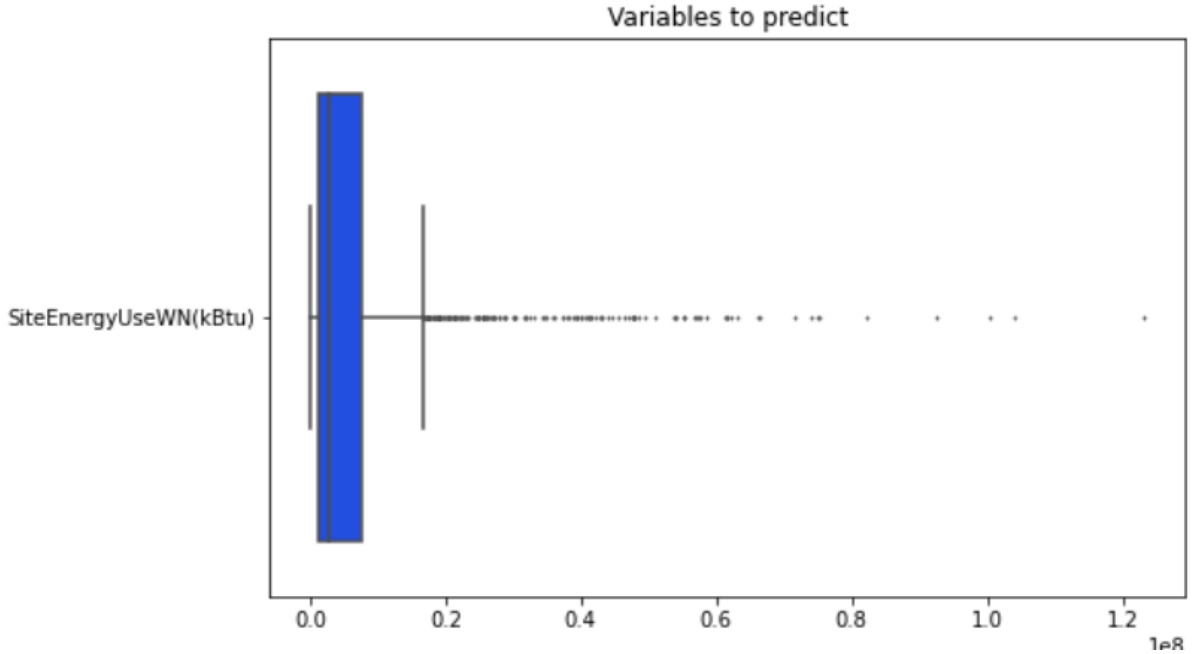
Ouput : train_energy.csv et test_energy.csv

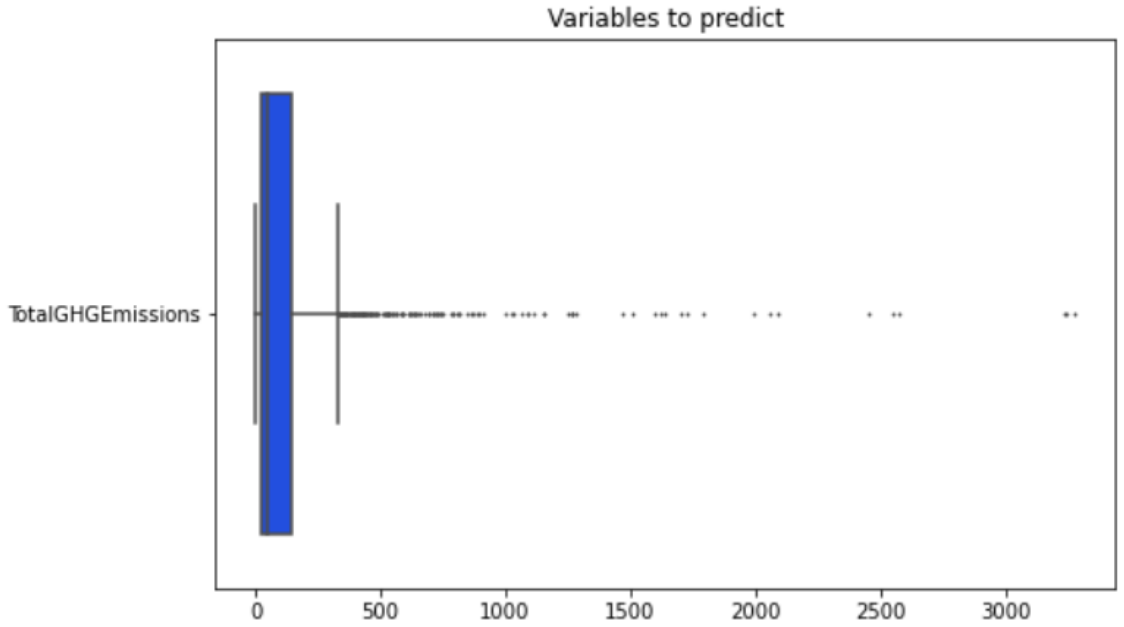
- Préparation d'un dataset d'entraînement de de test pour prédire « Log-TotalGHGEmissions »

Ouput : train_CO2.csv et test_CO2.csv

- Préparation d'un dataset d'entraînement de de test pour prédire « Log-TotalGHGEmissions » avec l'ENERGYSTARScore

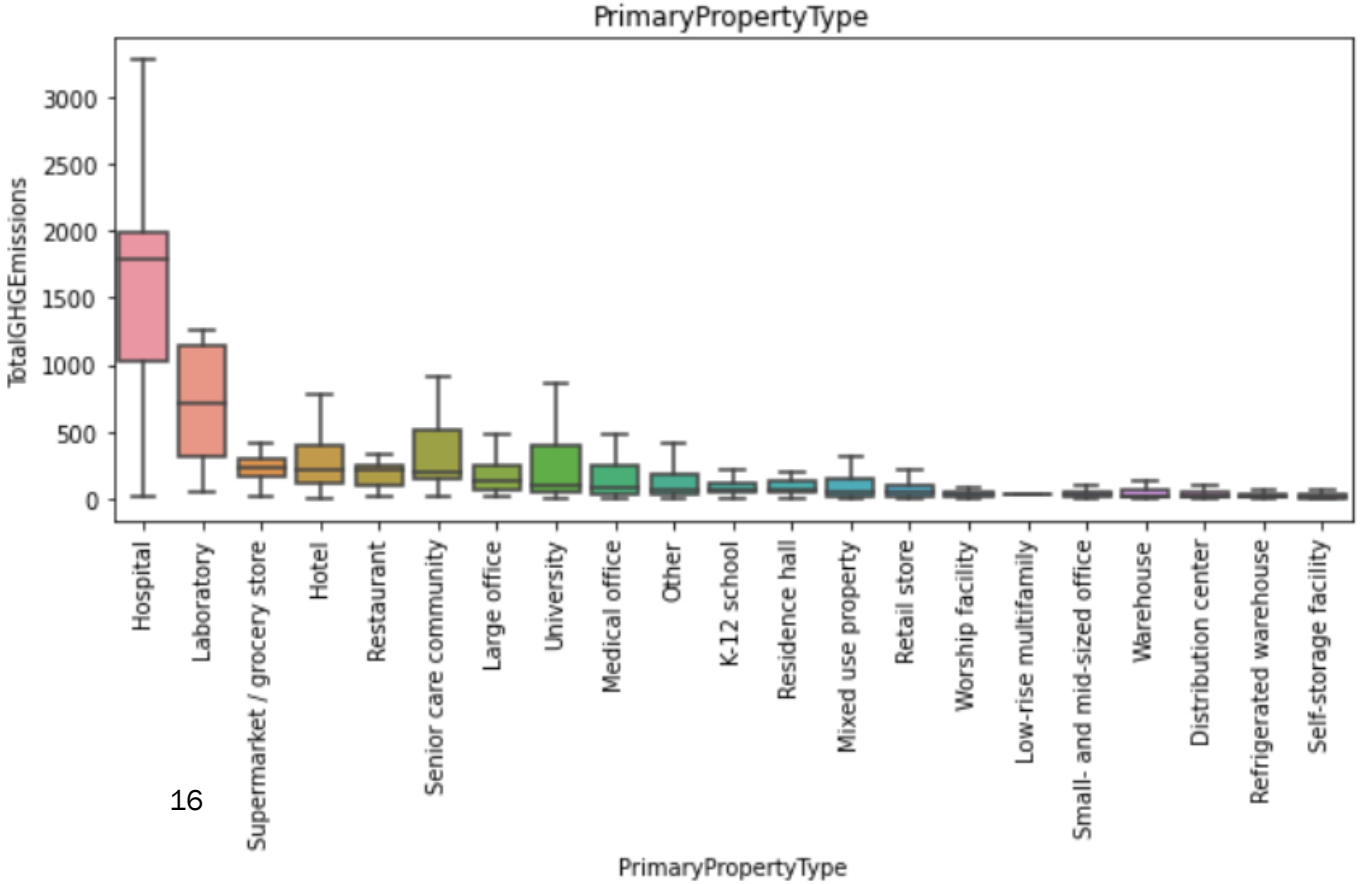
Ouput : train_ENERGYSTARScore.csv et test_ENERGYSTARScore.csv

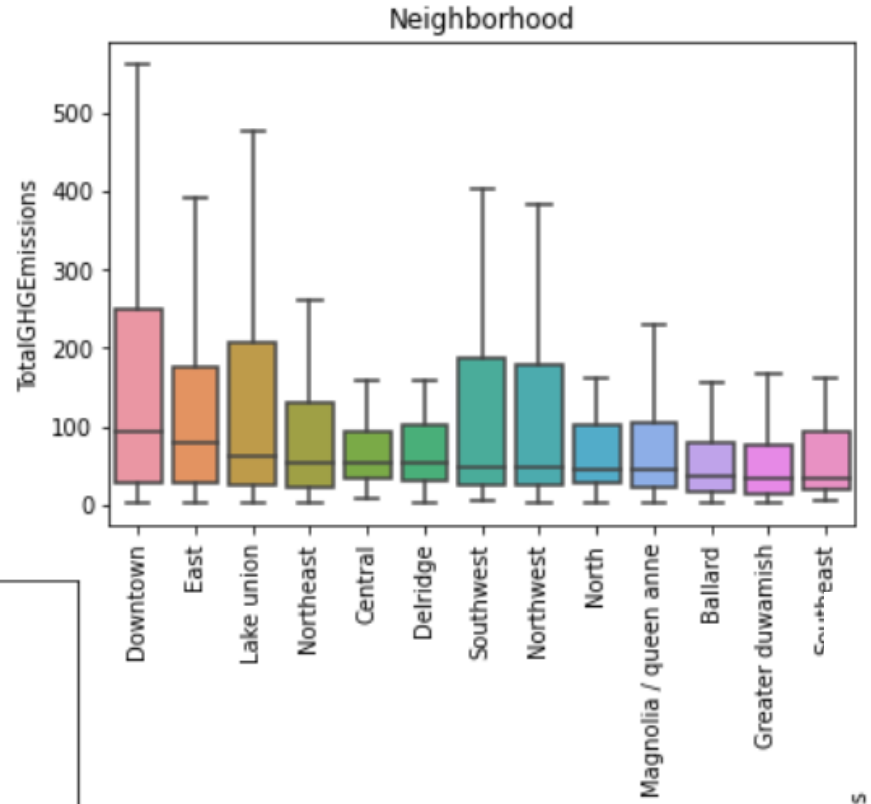


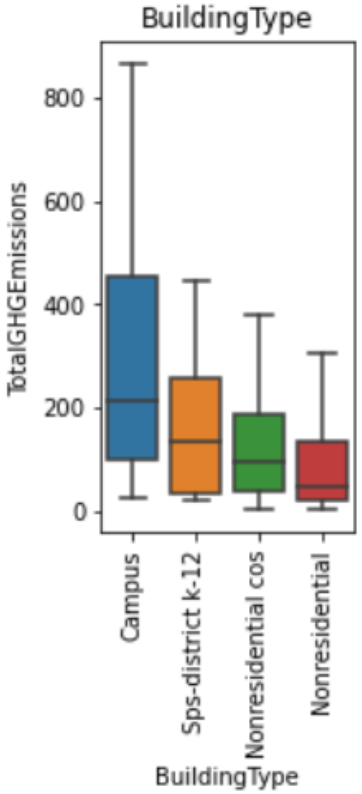


Preprocessing
2) Analyse exploratoire

Distribution du taux de CO2 en fonction des variables catégorielles

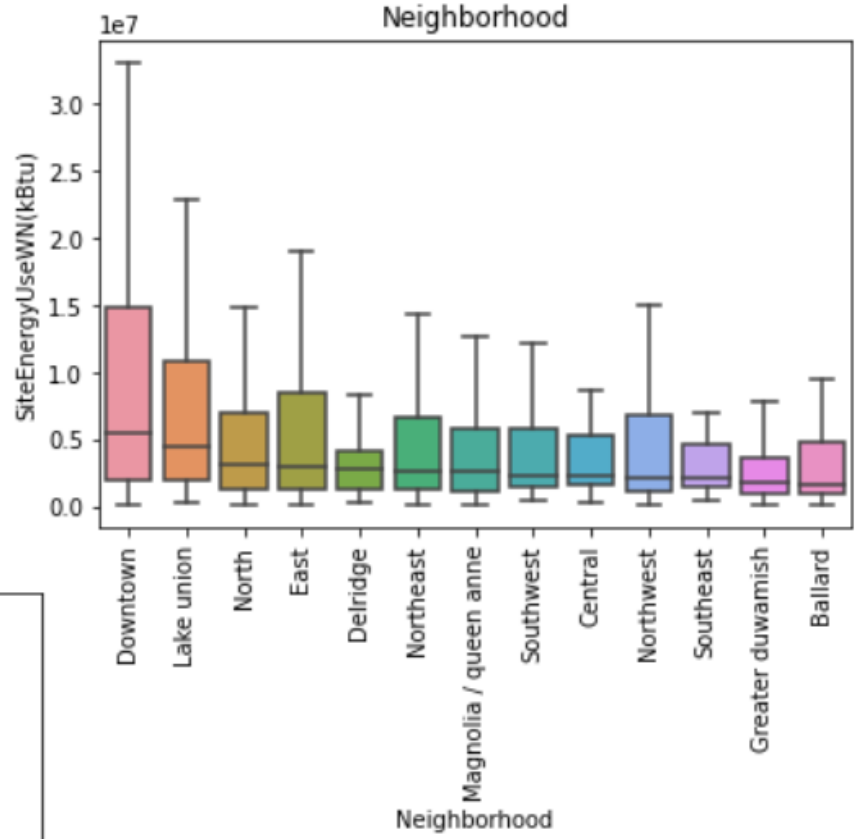


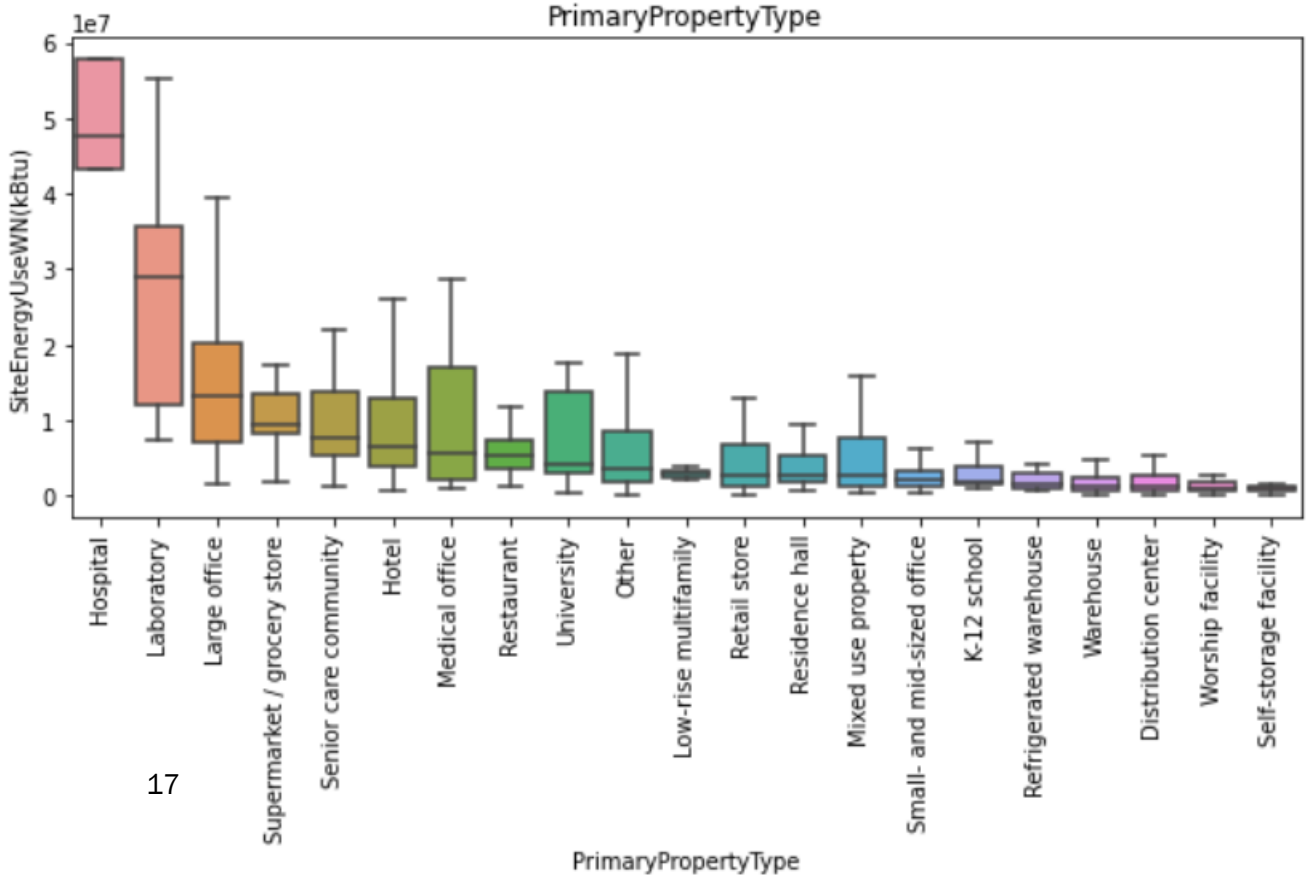


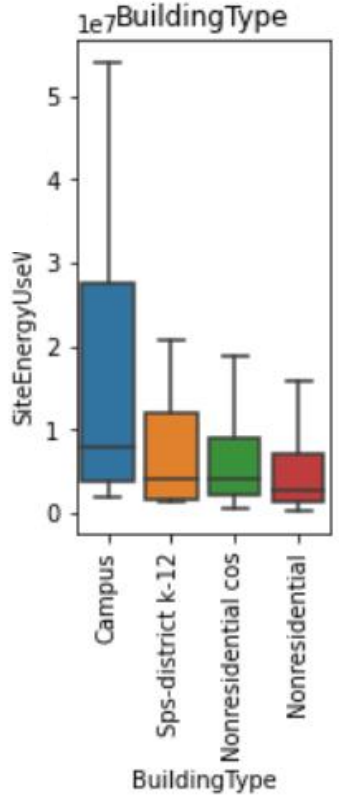


Preprocessing
2) Analyse exploratoire

Distribution de l'énergie totale en fonction des variables catégorielles





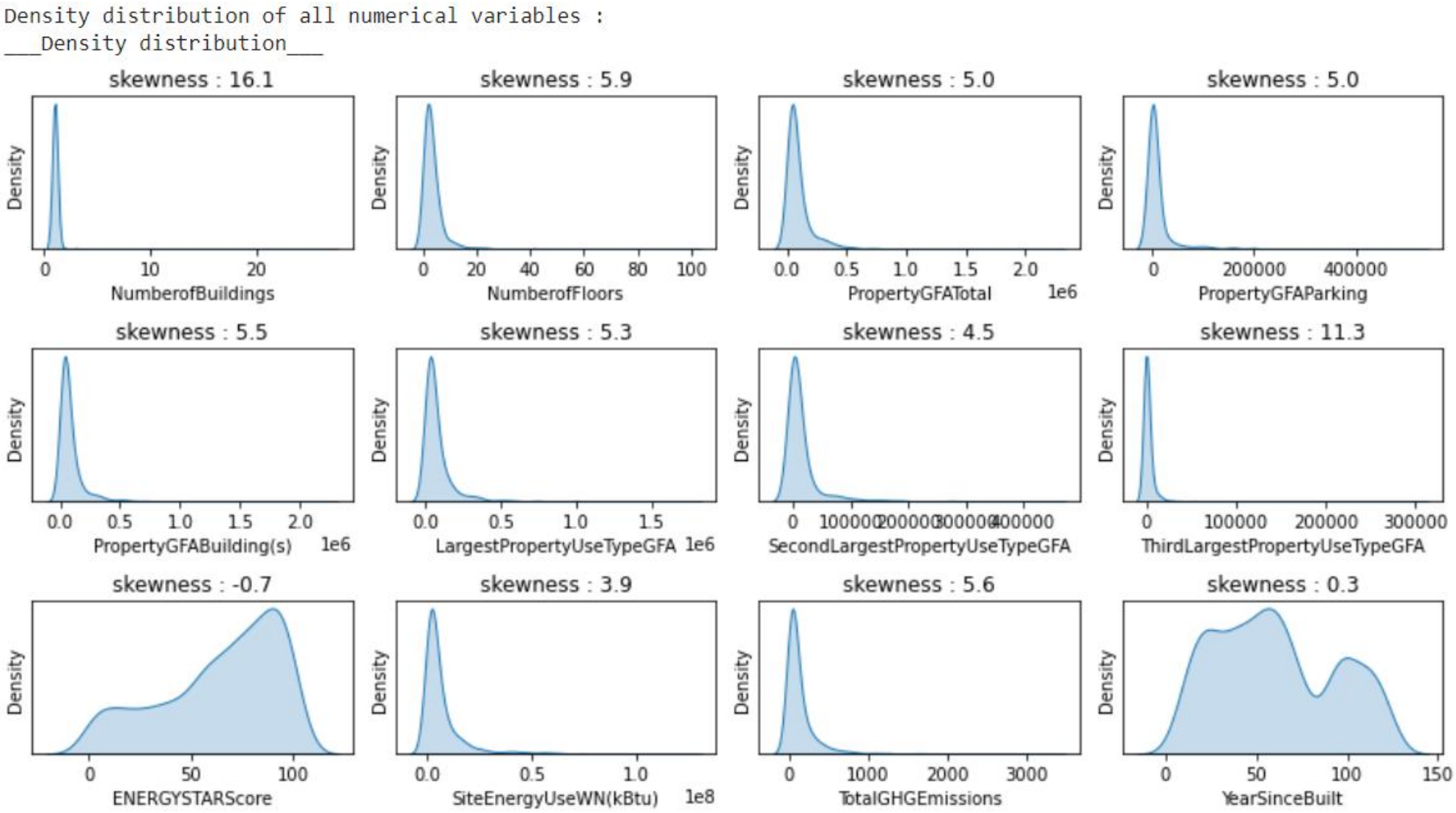


Preprocessing
2) Analyse exploratoire

Avant transformation logarithmique

Courbes de densité de toutes les variables numériques

On veut avoir une distribution normale (coefficient de Skewness < 2)

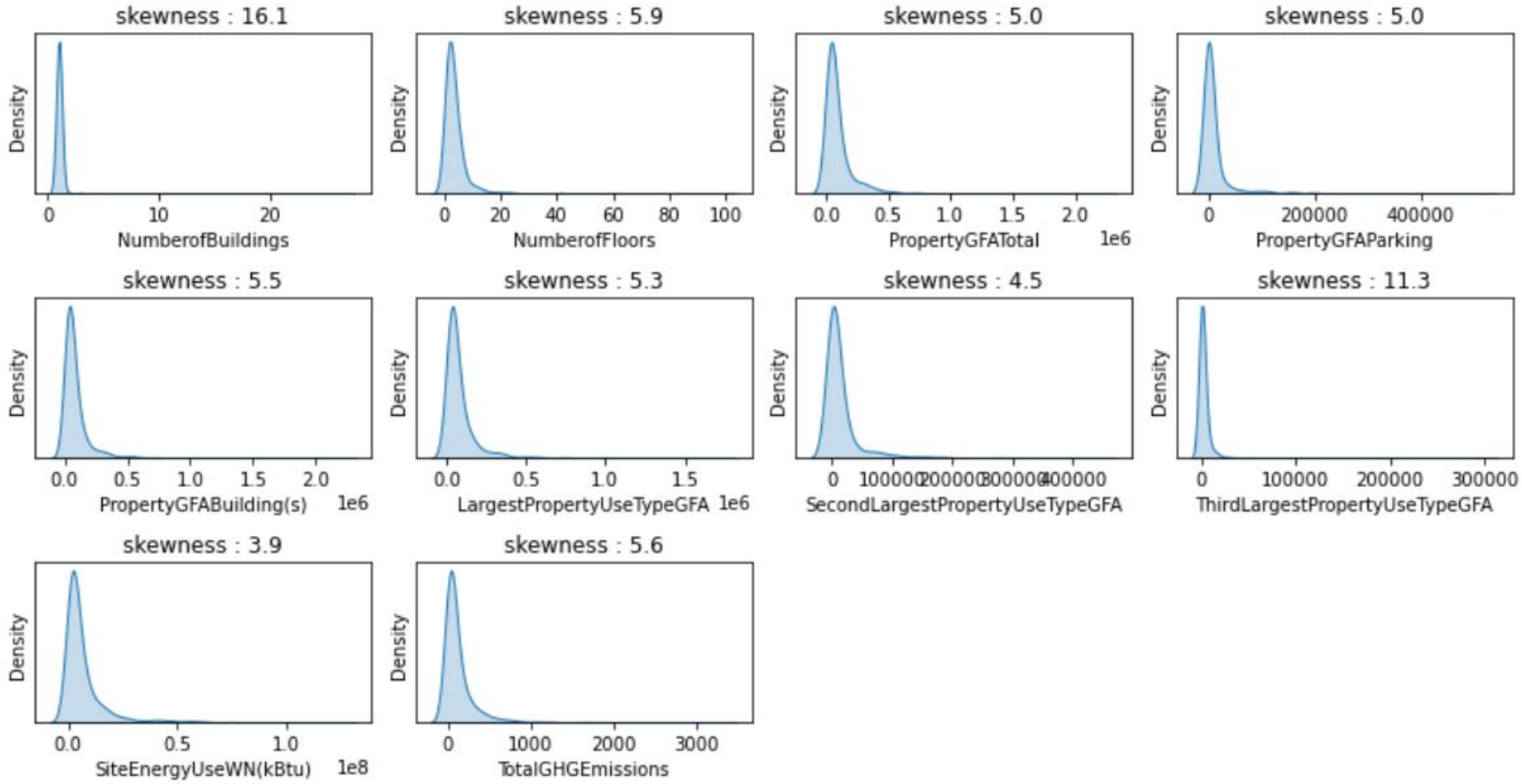


Preprocessing
2) Analyse exploratoire

Avant transformation logarithmique

Toutes les variables dont skewness > 2

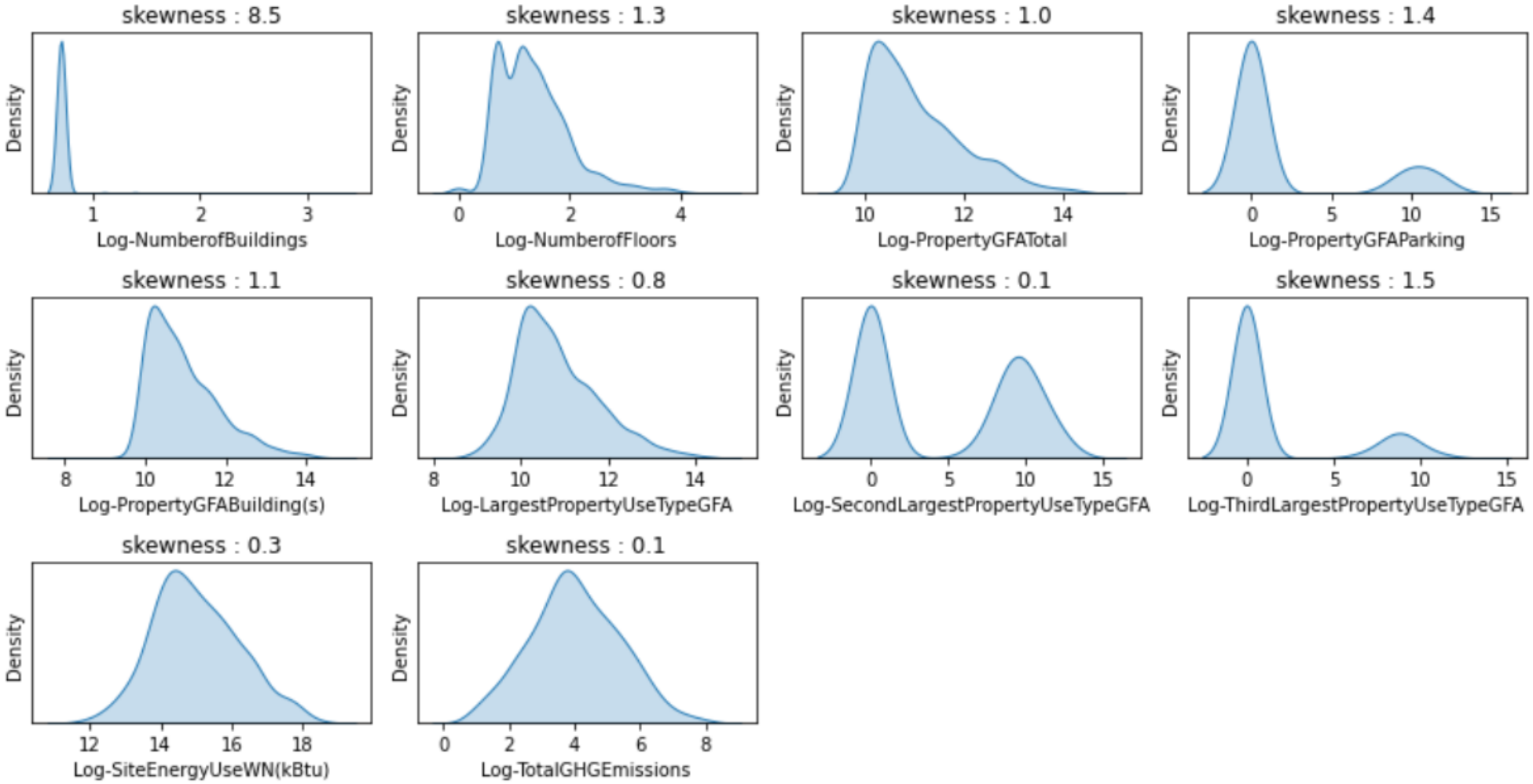
Before :
___Density distribution___



Preprocessing
2) Analyse exploratoire

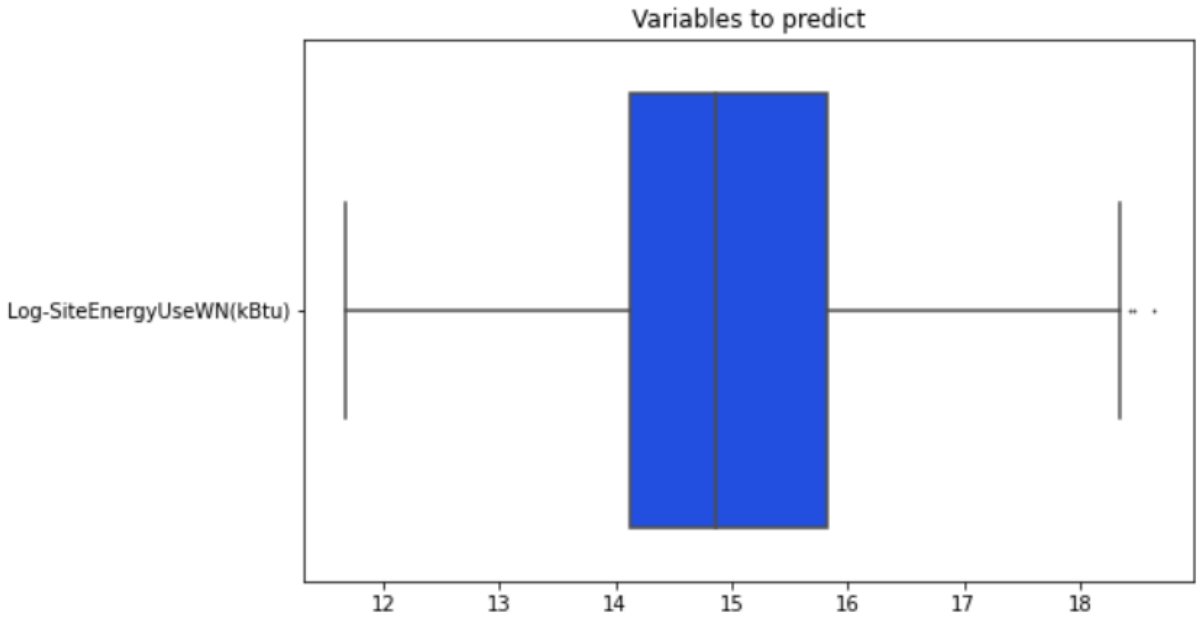
Après transformation logarithmique si skewness > 2

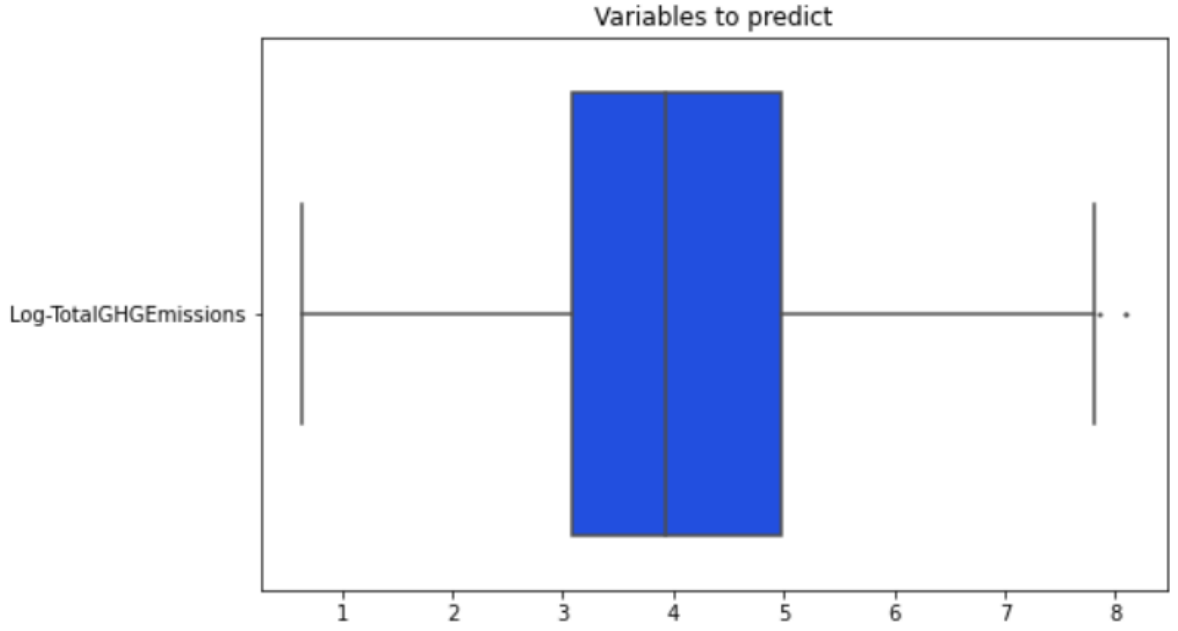
___Logarithmic transformation of features___
Log-transformation of the variables to predict.
___Density distribution___



Preprocessing
2) Analyse exploratoire

Distribution des variables à prédire après transformation logarithmique

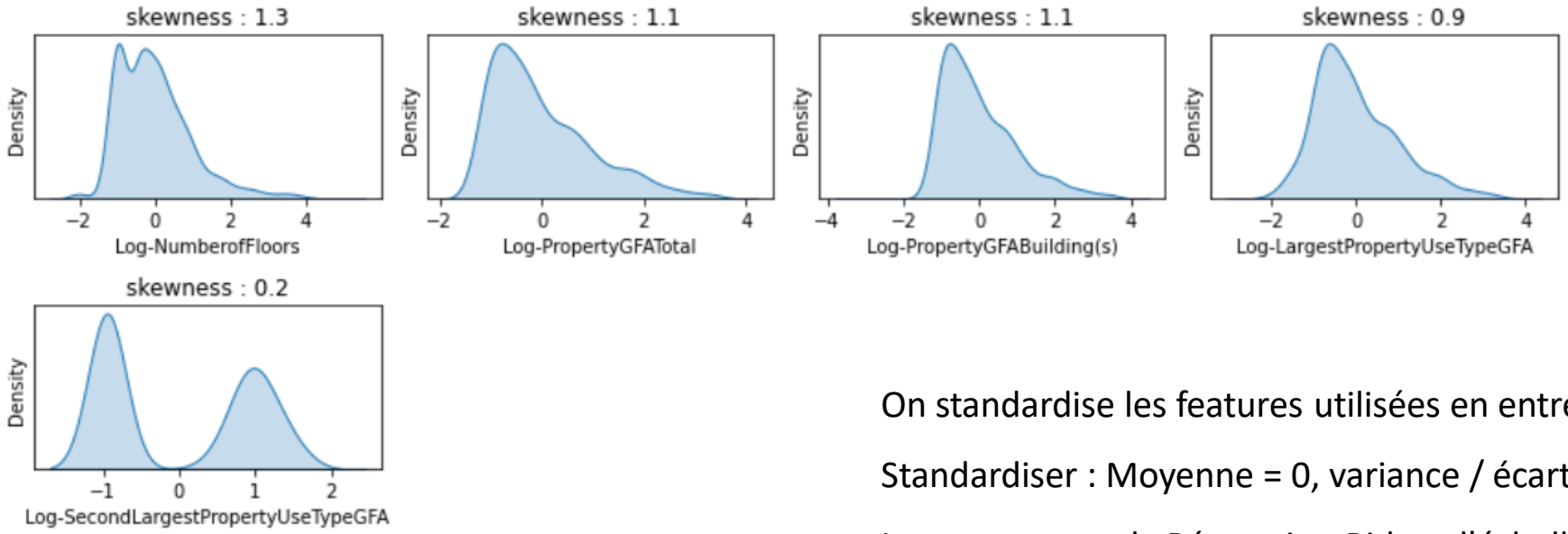




Preprocessing
3) Feature Engineering

Standardisation des variables numériques pour prédire la consummation totale en énergie et le CO2.

We can check that the numerical variables have a Standard Normal distribution.
___Density distribution___



$$X_j^i \leftarrow \frac{X_j^i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_j^i - \frac{1}{n}\sum_{i=1}^{n} X_j^i)^2}}.$$

On standardise les features utilisées en entrée.

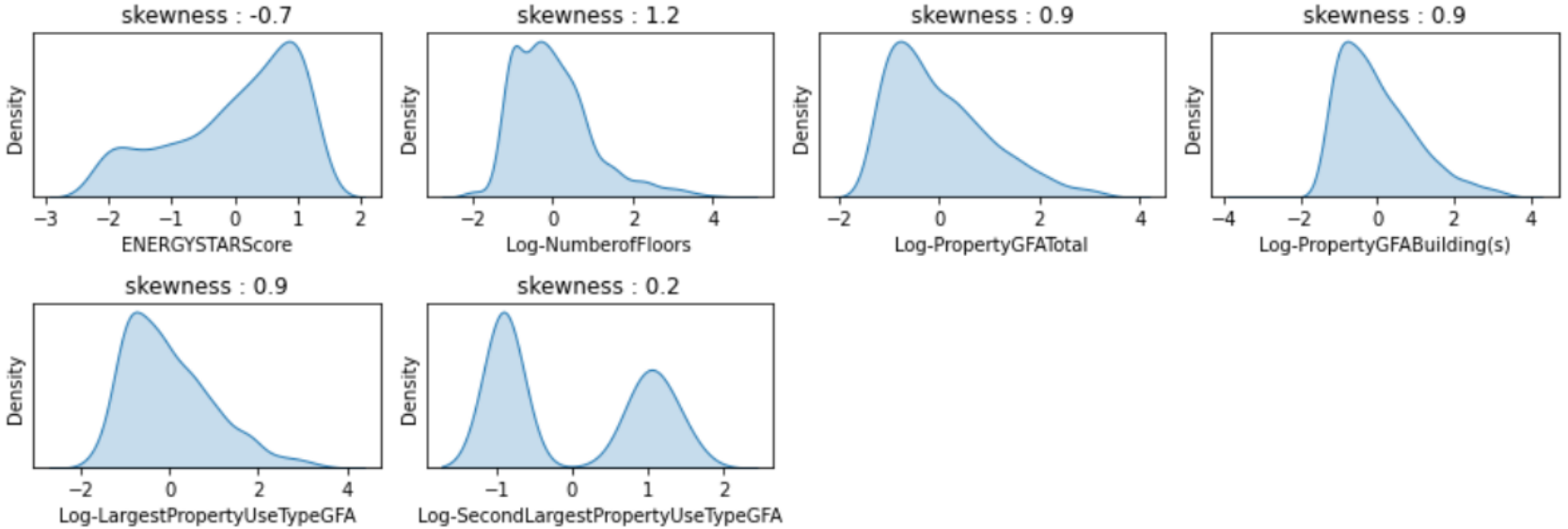
Standardiser : Moyenne = 0, variance / écart-type = 1.

Important pour la Régression Ridge : l'échelle de la plage des valeurs prises par les différentes variables a un impact sur le résultat de la régression ridge.

Preprocessing
3) Feature Engineering

Standardisation des variables numériques pour prédiret le CO2 à l'aide de l'ENERGY STAR Score

We can check that the numerical variables have a Standard Normal distribution.
___Density distribution___



$$X_j^i \leftarrow \frac{X_j^i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_j^i - \frac{1}{n}\sum_{i=1}^{n} X_j^i)^2}}.$$

Preprocessing
3) Feature Engineering

Encodage des variables catégorielles : One Hot Encoder

3 variables catégorielles avec 4 + 13 + 20 modalités donc 37 nouvelles colonnes.

	Neighborhood	BuildingType	PrimaryPropertyType
654	Downtown	Nonresidential	Other
1012	Greater duwamish	Nonresidential	Distribution center
1152	Greater duwamish	Nonresidential	Small- and mid-sized office
242	East	Nonresidential	Medical office
1315	East	Nonresidential cos	Mixed use property
...	...	...	...
715	Greater duwamish	Nonresidential	Small- and mid-sized office
905	Magnolia / queen anne	Nonresidential	Small- and mid-sized office
1096	Greater duwamish	Nonresidential	Warehouse
235	Downtown	Nonresidential	Large office
1061	Greater duwamish	Nonresidential	Warehouse

1038 rows × 3 columns

```
We have indeed : 37 labels after encoding the categorical variables.
```

	Neighborhood_Ballard	Neighborhood_Central	Neighborhood_Delridge	Neighborhood_Downtown	Neighborhood_East	Neighborhood_Greater duwamish	N
0	0.0	0.0	0.0	1.0	0.0	0.0	
1	0.0	0.0	0.0	1.0	0.0	0.0	
2	0.0	0.0	0.0	1.0	0.0	0.0	
4	0.0	0.0	0.0	1.0	0.0	0.0	
5	0.0	0.0	0.0	1.0	0.0	0.0	
...	...	...	...	...	...	...	
1479	0.0	0.0	0.0	0.0	0.0	1.0	
1480	0.0	0.0	0.0	0.0	0.0	0.0	
1481	0.0	0.0	0.0	0.0	0.0	1.0	
1482	0.0	0.0	0.0	0.0	0.0	0.0	
1483	0.0	0.0	0.0	0.0	0.0	0.0	

1038 rows × 37 columns

Modélisation Energie

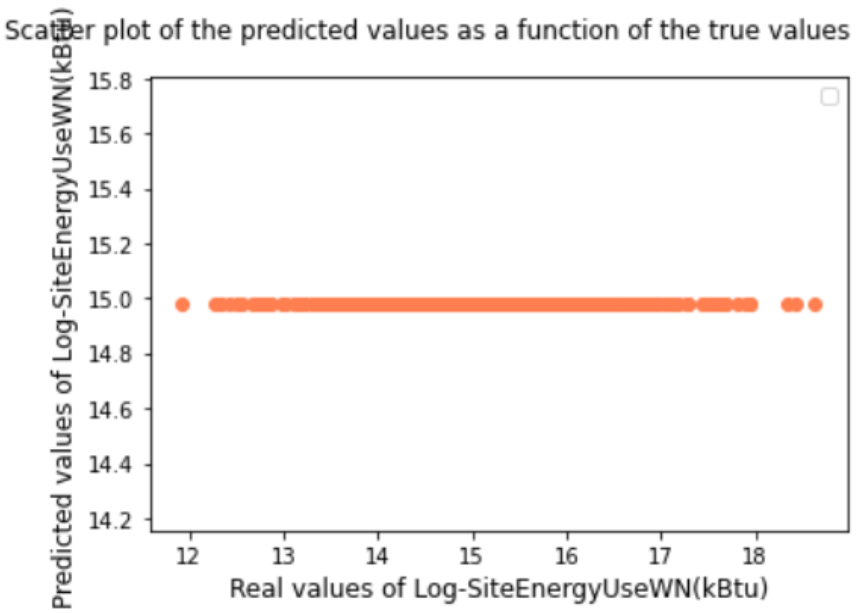
- Baseline : Dummy, Régression Linéaire
- Modèles linéaires
- Méthodes ensemblistes
- Optimisation des hyperparamètres
- Evaluation

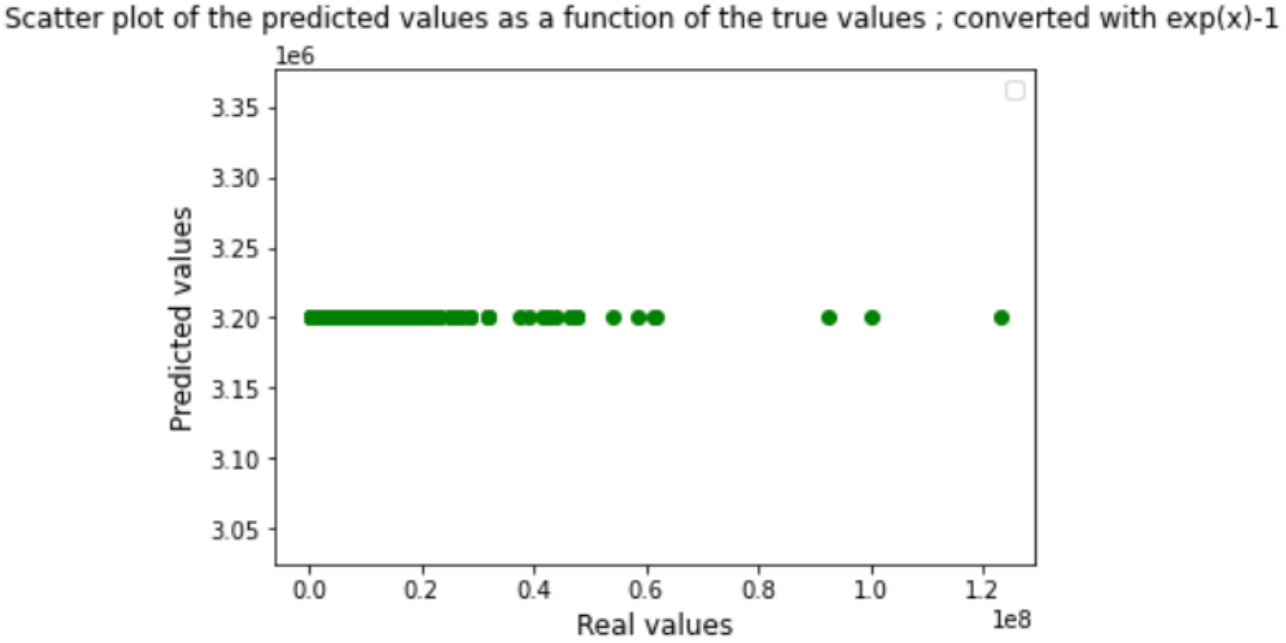
Modélisation énergie
Baseline : Dummy Regressor

Dummy Regressor : prédit toujours la moyenne

R2 score = 0

	Model	RMSE	MSE	MAE	Median Absolute Error	R^2 = 1 - RSE
0	Dummy Regressor	1.21260	1.47039	0.98044	0.85753	-0.00208

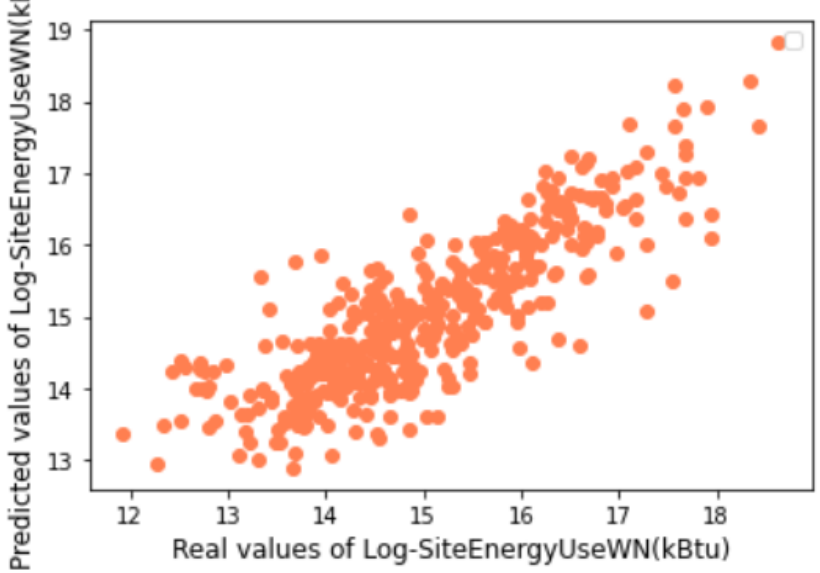




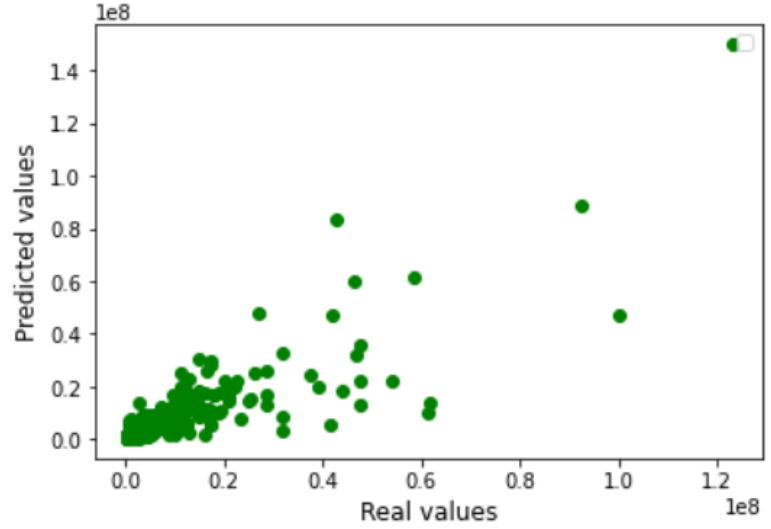
Modélisation énergie
Baseline : Régression linéaire

	Model	RMSE	MSE	MAE	Median Absolute Error	R^2 = 1 - RSE
0	Linear Regression	0.64607	0.41741	0.47804	0.37301	0.71553

Scatter plot of the predicted values as a function of the true values ; ln(1+x)



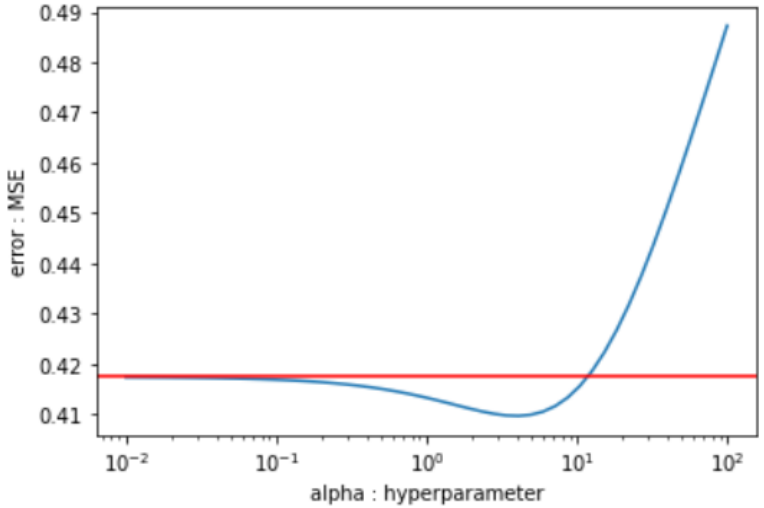
Scatter plot of the predicted values as a function of the true values ; converted with exp(x)-1

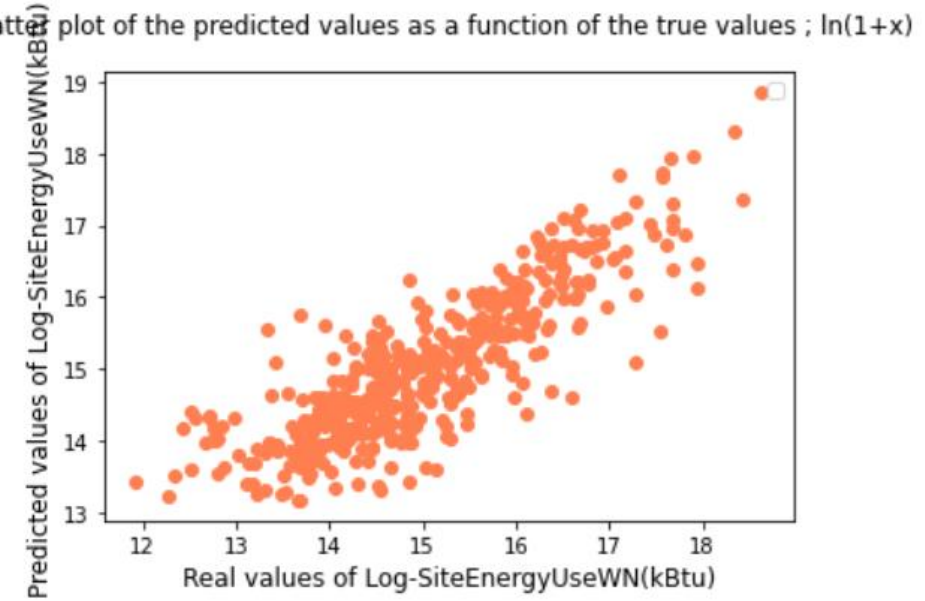


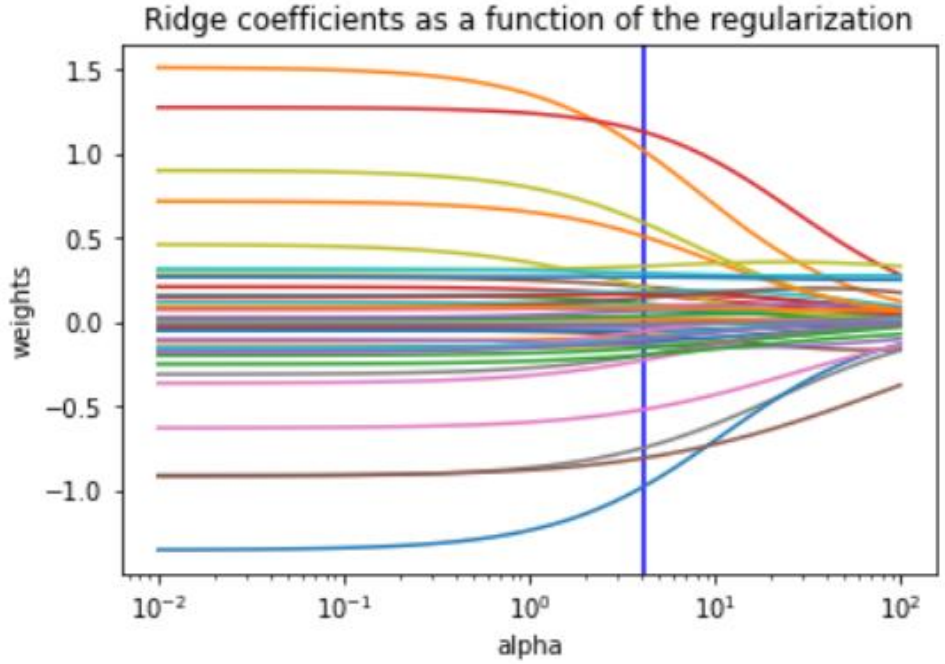
Modélisation énergie
1) Modèle linéaire : Régression Ridge

$$\arg\min_{\beta \in \mathbb{R}^{p+1}} ||y - X\beta||_2^2 + \lambda ||\beta||_2^2$$

```
Minimum Mean Squared Error for Ridge Regression : 0.40969075722747716
Best alpha for that minimal MSE :  4.094915062380423
(0.00630957344480193,
 158.48931924611142,
 0.40581123876044894,
 0.4911606450350697)
```

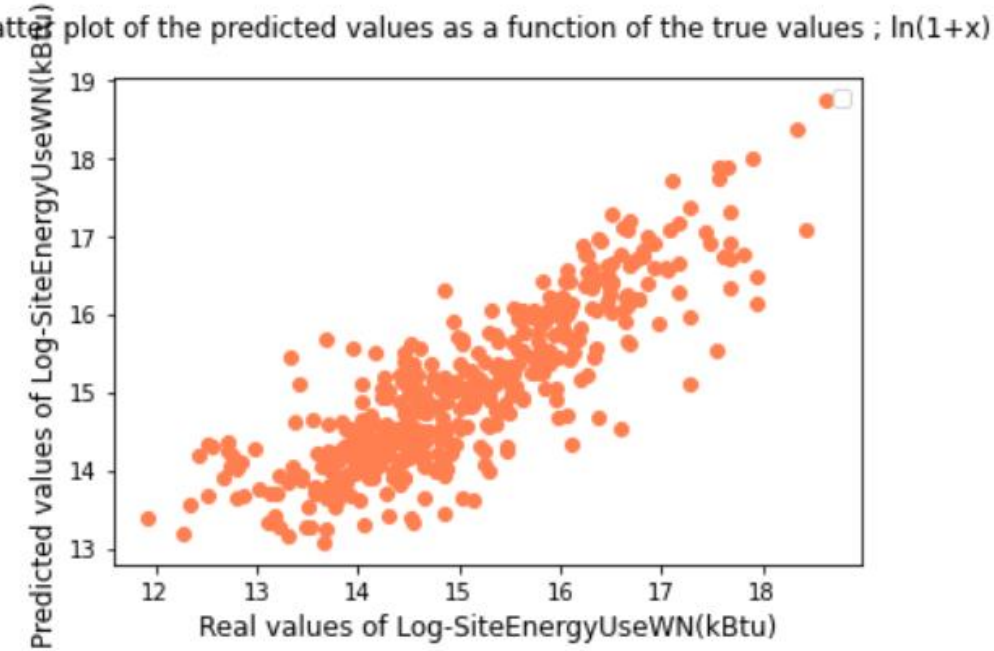




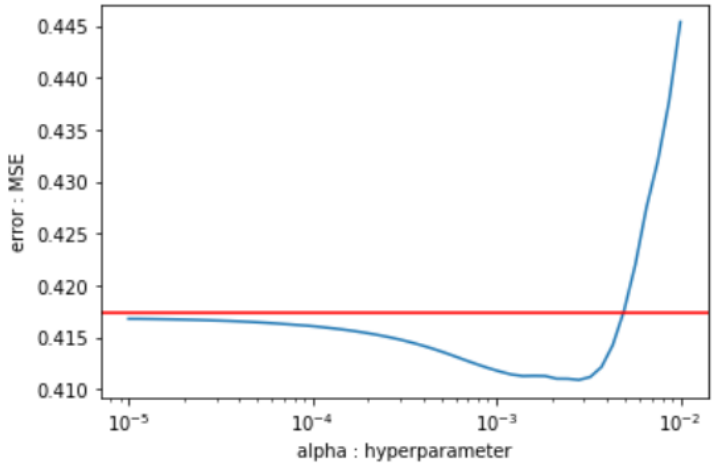


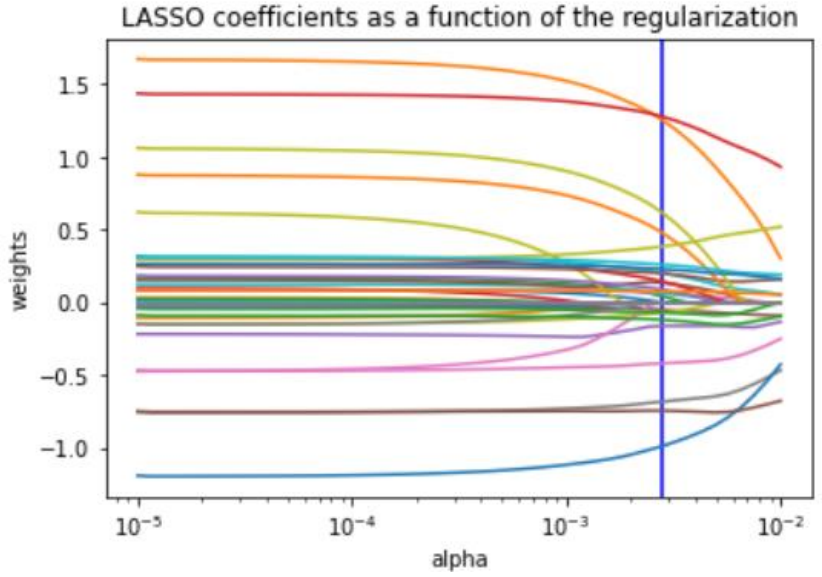
Modélisation énergie
2) Modèle linéaire : LASSO

$$\arg\min_{\beta \in \mathbb{R}^{p+1}} ||y - X\beta||_2^2 + \lambda ||\beta||_1 .$$




```
Minimum Mean Squared Error for LASSO Regression : 0.4108929072570126
Best alpha for that minimal MSE :  0.002811768697974231
(7.079457843841373e-06,
 0.01412537544622754,
 0.4091695266948739,
 0.4470838990619259)
```





Modélisation énergie
3) Modèle linéaire : Elastic Net

$$\arg\min_{\beta \in \mathbb{R}^{p+1}} ||y - X\beta||_2^2 + \lambda \left((1-\alpha)||\beta||_1 + \alpha||\beta||_2^2 \right)$$

```python
n = 100
# or : l1_ratio = [i / n for i in range(n)]
l1_ratio = np.arange(start=0, stop=1, step= 1/n)

a, b , n_alphas = -5, 5, 1000

param_grid_elastic = [{
    "alpha": np.logspace(a, b, n_alphas), # penalty intensity (5 values between 10⁻³ and 10¹)
    "l1_ratio": l1_ratio # mixing parameter for l1 and l2 penalties
}]
```

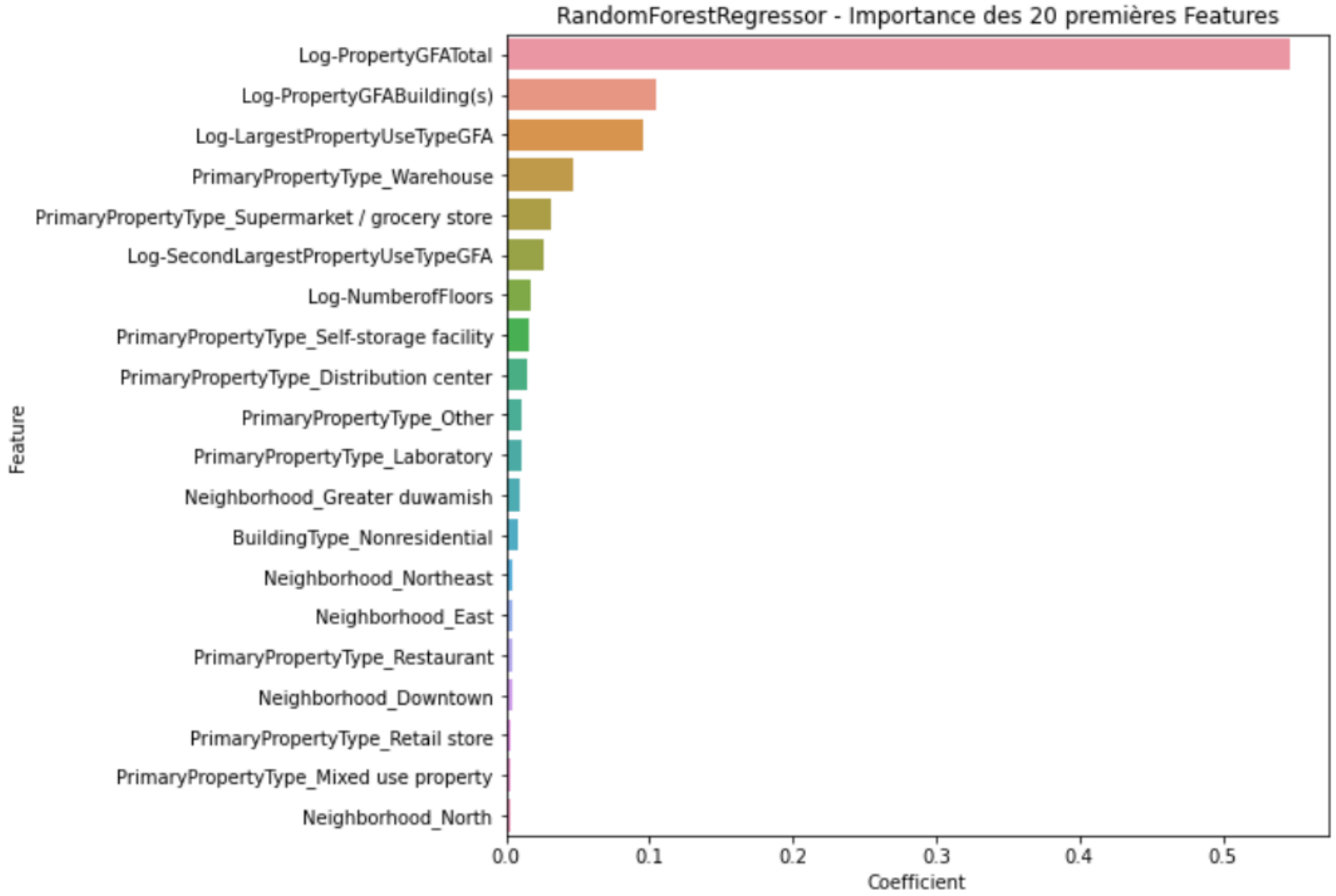
Hyperparamètres alpha et l1_ratio

Modélisation énergie
4) Méthodes ensemblistes
Méthodes parallèles : Forêt Aléatoire

```
param_grid_forest = [{
    "n_estimators": n_estimators
}]
```

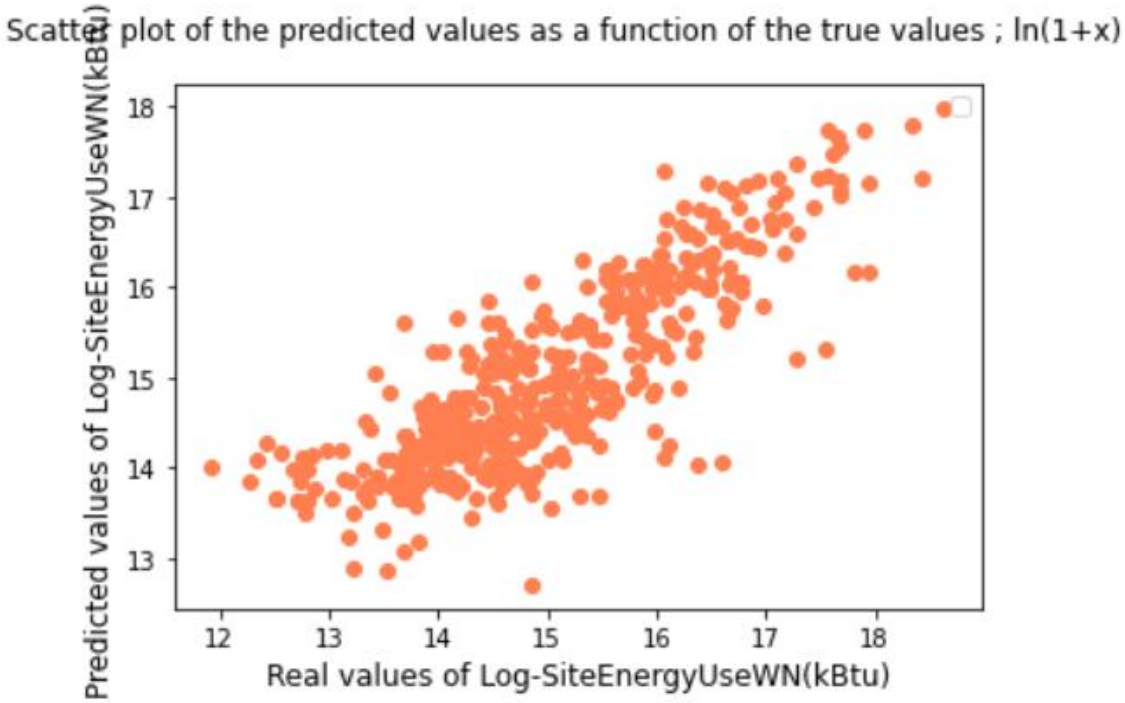
Nombre d'arbres de décision :

[10, 50, 100, 300, 500]



Modélisation énergie
5) Méthodes ensemblistes
Méthodes séquentielles : Gradient Boosting

```
parameters = {
    'n_estimators' : [100,500,1000,2000],
    "learning_rate": (0.05, 0.10, 0.15),
    "gamma":[ 0.0, 0.1, 0.2]
}
```



Modélisation énergie Evaluation

Résultats de tous les modèles testés.

Validation Croisée avec 5 folds.

Conclusion : le meilleur estimateur pour prédire Log-SiteEnergyUseWN(kBtu) est **Elastic Net** que l'on sauvegarde sous le nom de « best_model_energy.

On a une légère amélioration du RMSE et du R2 score par rapport à notre baseline.

Prediction for : Log-SiteEnergyUseWN(kBtu)

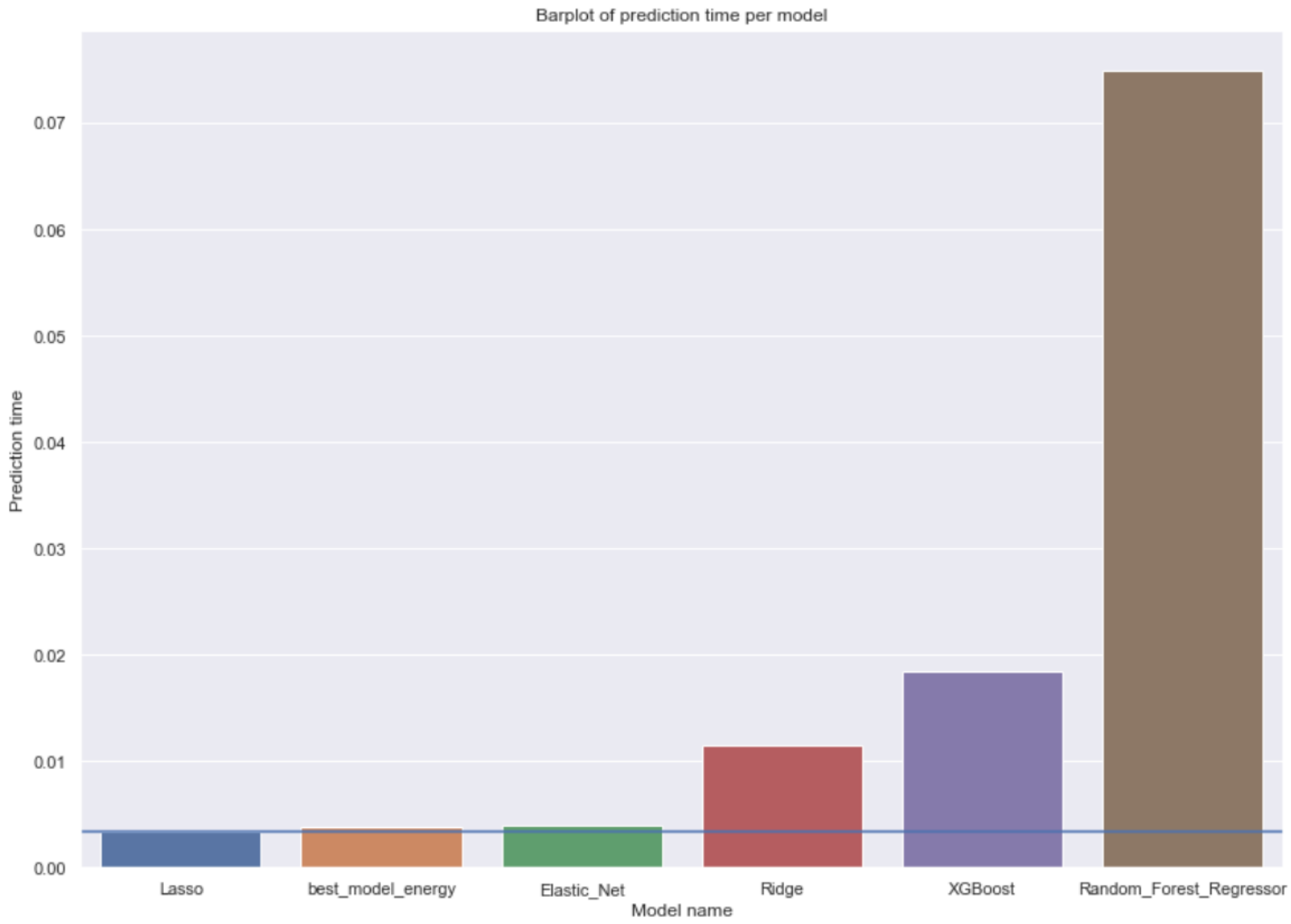
	Model	RMSE	MSE	MAE	Median Absolute Error	R^2 = 1 - RSE
0	Ridge manually	0.64007	0.40969	0.47571	0.36154	0.72079
0	LASSO manually	0.64101	0.41089	0.47862	0.36180	0.71997
0	Elastic Net GridSearchCV	0.64115	0.41108	0.47479	0.36454	0.71985
0	Elastic Net RandomSearchCV	0.64134	0.41132	0.47491	0.36561	0.71968
0	LASSO GridSearchCV	0.64163	0.41169	0.47502	0.36603	0.71943
0	Ridge GridSearchCV	0.64218	0.41240	0.47488	0.36423	0.71895
0	Linear Regression	0.64607	0.41741	0.47804	0.37301	0.71553
0	XGBoost GridSearchCV	0.64700	0.41861	0.48646	0.36554	0.71472
0	Random Forest GridSearchCV	0.67649	0.45763	0.51510	0.43265	0.68812

Results Cross-Validated

	Model	Mean CV R^2
0	LASSO Regression CV	0.71222
0	Elastic Net GridSearchCV	0.71222
0	Elastic Net RandomSearchCV	0.71221
0	Ridge Regression CV	0.71186
0	Linear Regression CV	0.71043
0	XGBoost CV	0.69327
0	Random Forest CV	0.68171

Choix meilleur estimateur

Temps de prédiction par modèle en secondes



Modélisation CO2

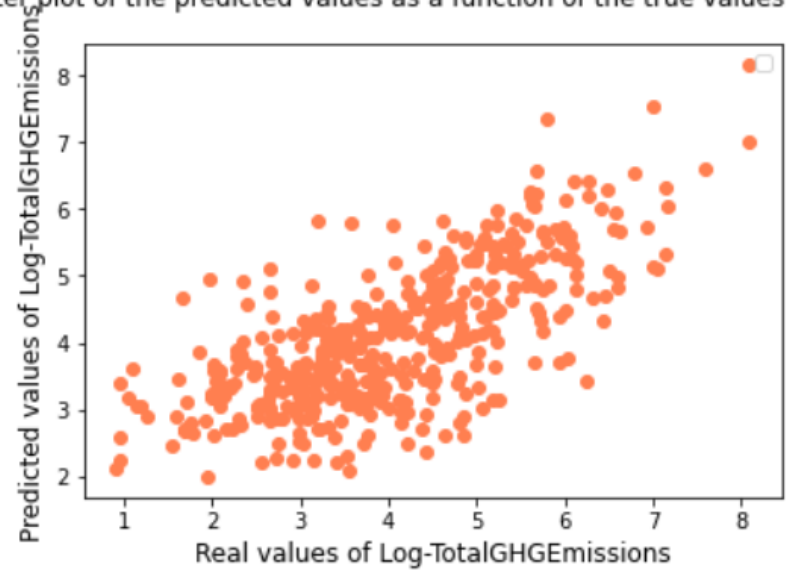
- Baseline : Dummy, Régression Linéaire
- Modèles linéaires
- Méthodes ensemblistes
- Optimisation des hyperparamètres
- Evaluation

Modélisation CO2
Baseline : Dummy et Régression Linéaire

R2 : 0.49676232050197566
Prediction for : Log-TotalGHGEmissions

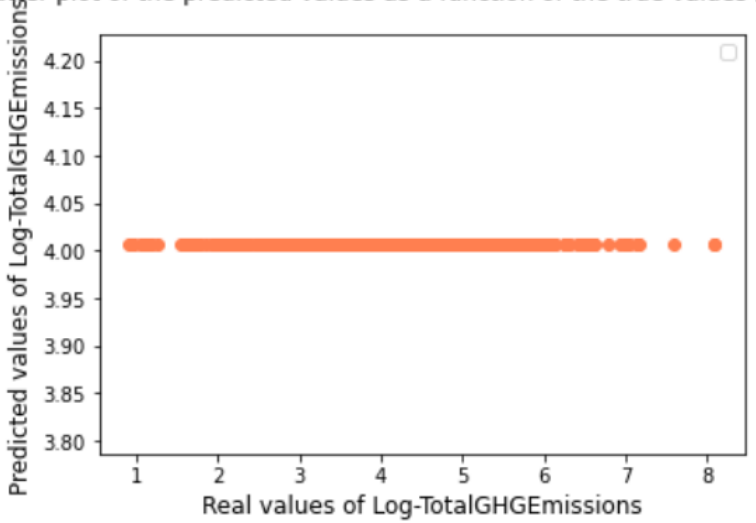
	Model	RMSE	MSE	MAE	Median Absolute Error	R^2 = 1 - RSE
0	Linear Regression	0.96314	0.92764	0.75974	0.60130	0.49676

Scatter plot of the predicted values as a function of the true values ; ln(1+x)



	Model	RMSE	MSE	MAE	Median Absolute Error	R^2 = 1 - RSE
0	Dummy Regressor	1.35810	1.84442	1.11076	0.98457	-0.00058

Scatter plot of the predicted values as a function of the true values ; ln(1+x)



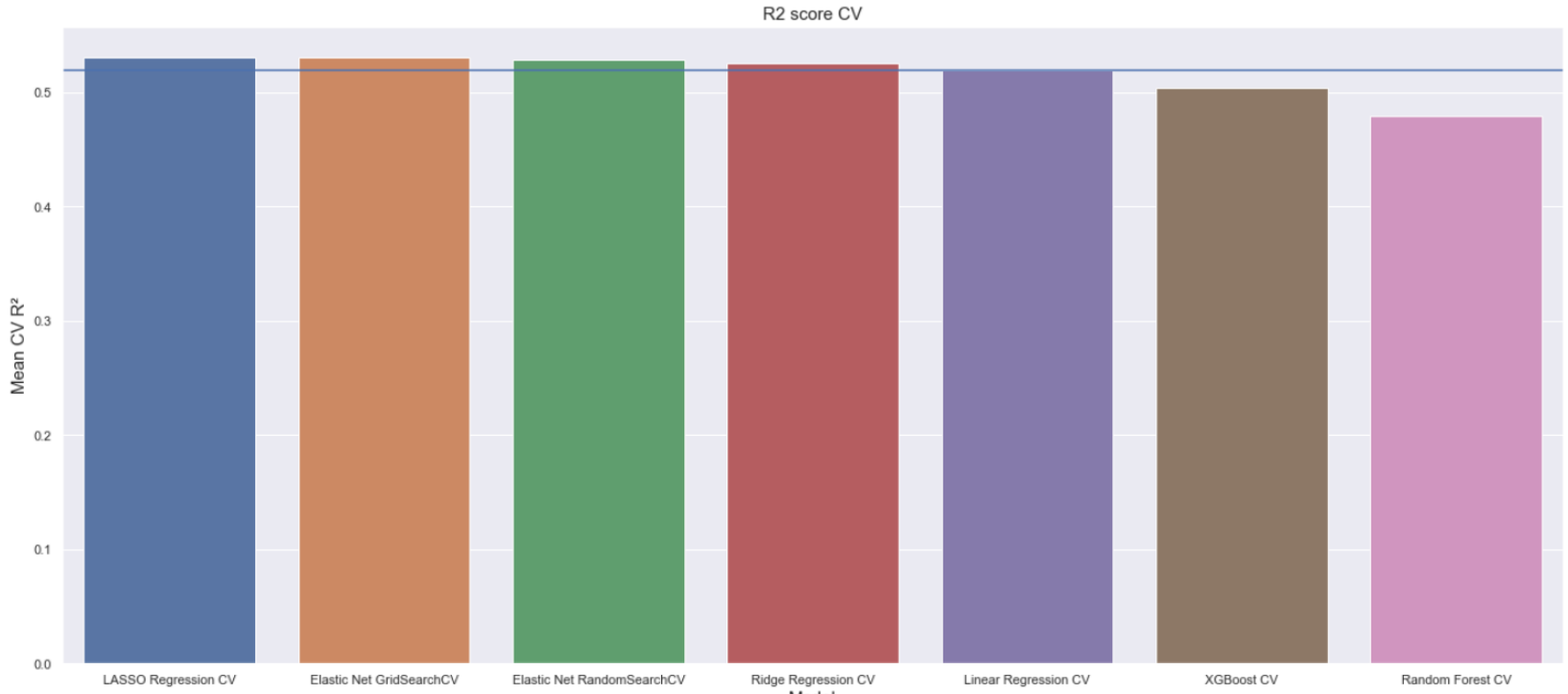
Modélisation CO2

Prediction for : Log-TotalGHGEmissions

	Model	RMSE	MSE	MAE	Median Absolute Error	R^2 = 1 - RSE
0	XGBoost GridSearchCV	0.95307	0.90835	0.75929	0.63840	0.50723
0	Ridge manually	0.95323	0.90865	0.75327	0.57506	0.50706
0	LASSO manually	0.95433	0.91075	0.75633	0.60501	0.50593
0	Elastic Net GridSearchCV	0.95499	0.91201	0.75516	0.58423	0.50525
0	Ridge GridSearchCV	0.95522	0.91244	0.75347	0.58518	0.50501
0	LASSO GridSearchCV	0.95573	0.91342	0.75581	0.58488	0.50448
0	Elastic Net RandomSearchCV	0.95670	0.91528	0.75477	0.58435	0.50347
0	Linear Regression	0.96314	0.92764	0.75974	0.60130	0.49676
0	Random Forest GridSearchCV	0.98889	0.97790	0.78273	0.62643	0.46950

Results Cross-Validated

	Model	Mean CV R^2
0	LASSO Regression CV	0.53114
0	Elastic Net GridSearchCV	0.53069
0	Elastic Net RandomSearchCV	0.52855
0	Ridge Regression CV	0.52575
0	Linear Regression CV	0.51895
0	XGBoost CV	0.50379
0	Random Forest CV	0.47928

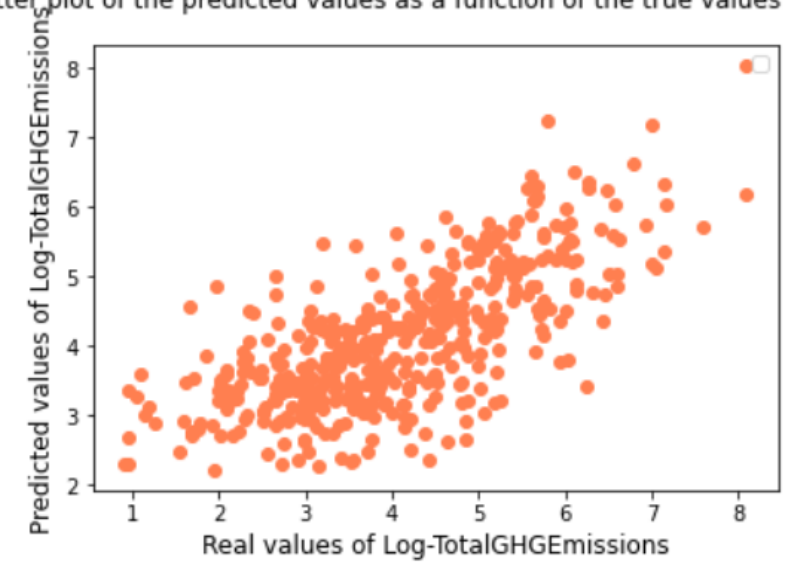


Modélisation CO2

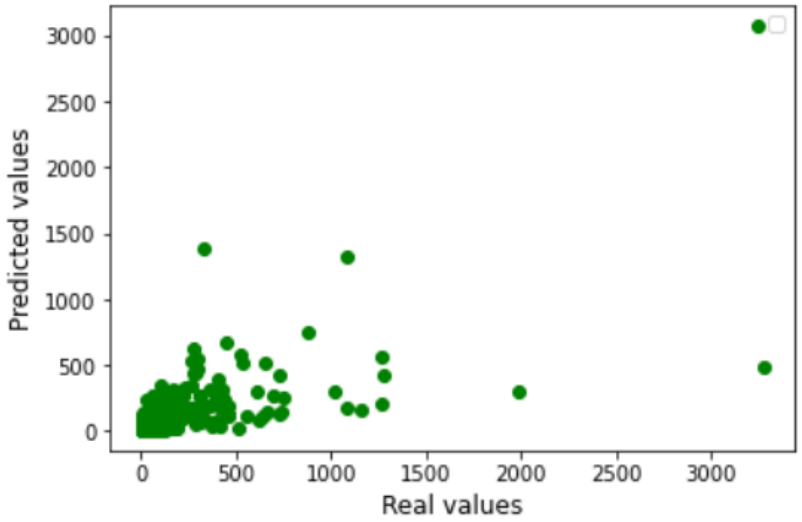
Conclusion : le meilleur estimateur pour prédire Log-TotalGHGEmissions est le LASSO que l'on sauvegarde sous le nom de « best_model_CO2

Erreur de prédiction avec le LASSO (meilleur modèle pour prédire le CO2)

Scatter plot of the predicted values as a function of the true values ; ln(1+x)



Scatter plot of the predicted values as a function of the true values ; converted with exp(x)-1



Modélisation CO2 / ENERGY STAR Score

- Baseline : Dummy, Régression Linéaire
- Modèles linéaires
- Méthodes ensemblistes
- Optimisation des hyperparamètres
- Evaluation

Modélisation ENERGY STAR Score Baseline

Baseline plus élevée que celle pour la prédiction de CO2 sans Energy Star Score

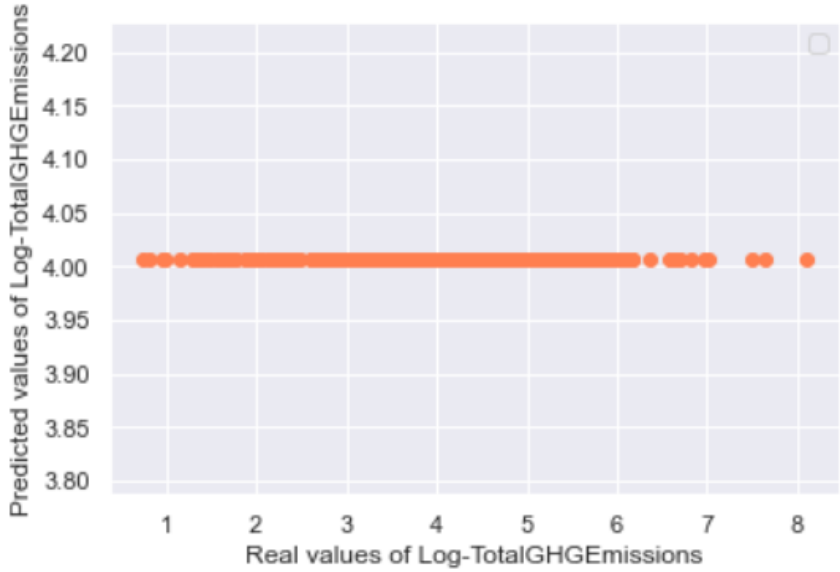
```
R2 : 0.663966910042564
Prediction for :  Log-TotalGHGEmissions
```

	Model	RMSE	MSE	MAE	Median Absolute Error	R^2 = 1 - RSE
0	Linear Regression	0.83635	0.69948	0.67449	0.57991	0.66397

Scatter plot of the predicted values as a function of the true values ; ln(1+x)

	Model	RMSE	MSE	MAE	Median Absolute Error	R^2 = 1 - RSE
0	Dummy Regressor	1.44475	2.08730	1.19178	1.03728	-0.00276

Scatter plot of the predicted values as a function of the true values ; ln(1+x)



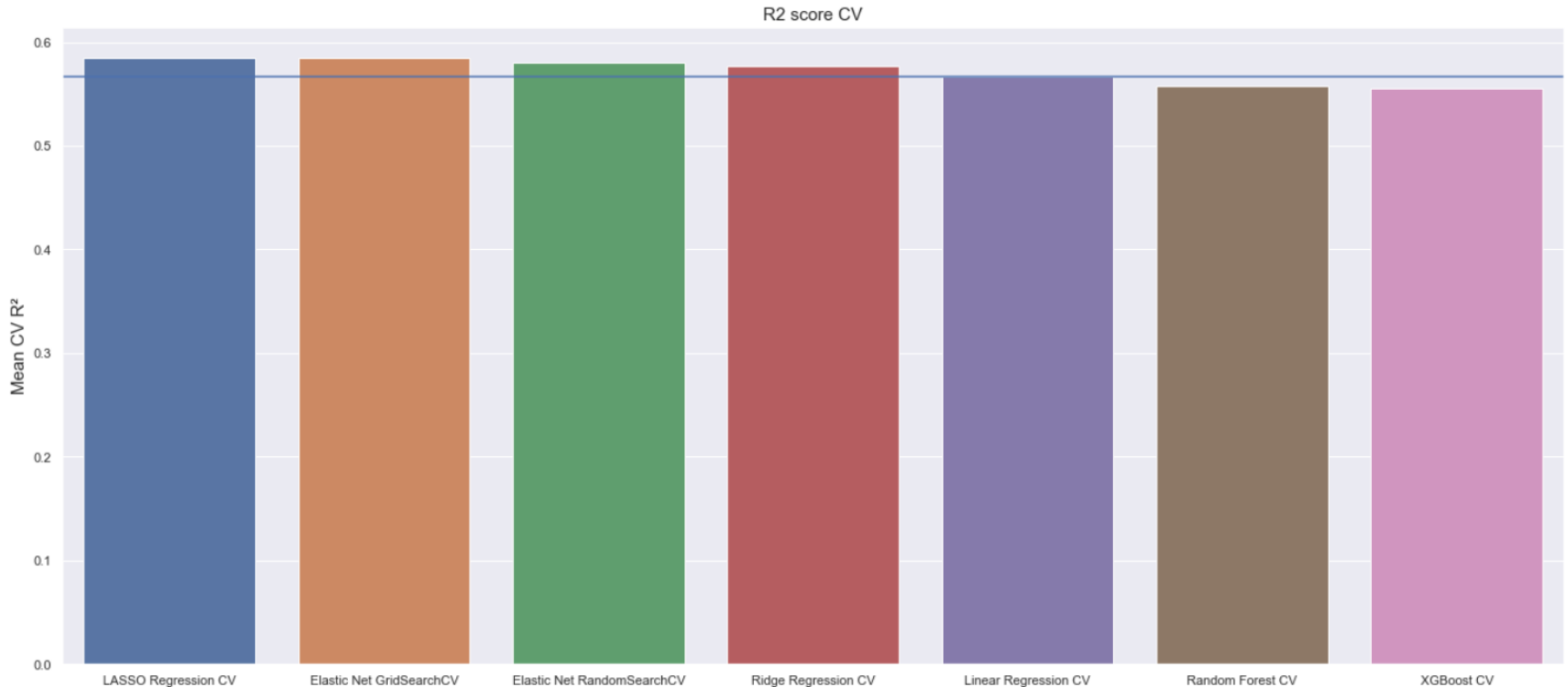
Modélisation ENERGY STAR Score

Prediction for : Log-TotalGHGEmissions

	Model	RMSE	MSE	MAE	Median Absolute Error	R^2 = 1 - RSE
0	LASSO manually	0.83016	0.68916	0.68353	0.60217	0.66892
0	LASSO GridSearchCV	0.83114	0.69080	0.68683	0.60924	0.66814
0	Ridge manually	0.83249	0.69304	0.67793	0.59134	0.66706
0	Elastic Net GridSearchCV	0.83375	0.69514	0.68985	0.60974	0.66605
0	Ridge GridSearchCV	0.83573	0.69845	0.68562	0.60294	0.66446
0	Linear Regression	0.83635	0.69948	0.67449	0.57991	0.66397
0	Elastic Net RandomSearchCV	0.84450	0.71319	0.69991	0.61366	0.65738
0	Random Forest GridSearchCV	0.85609	0.73289	0.68553	0.57373	0.64791
0	XGBoost GridSearchCV	0.85874	0.73744	0.68595	0.56115	0.64573

Results Cross-Validated

	Model	Mean CV R^2
0	LASSO Regression CV	0.58523
0	Elastic Net GridSearchCV	0.58480
0	Elastic Net RandomSearchCV	0.58008
0	Ridge Regression CV	0.57638
0	Linear Regression CV	0.56651
0	Random Forest CV	0.55771
0	XGBoost CV	0.55489



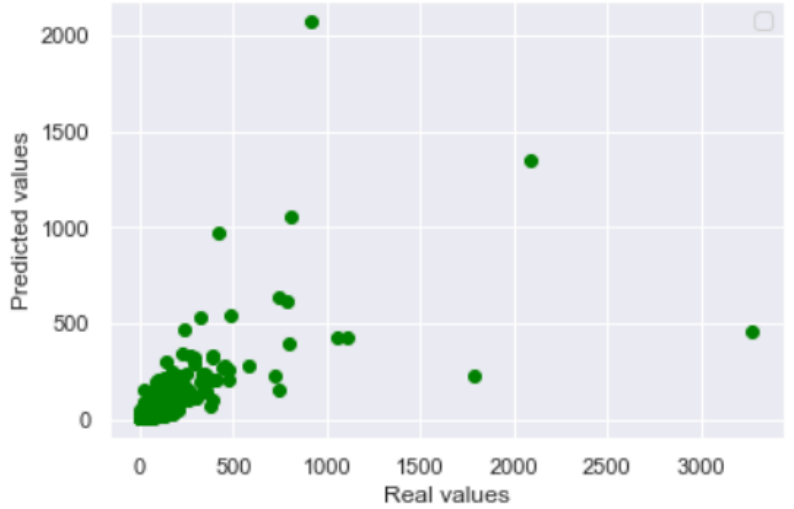
Modélisation ENERGY STAR Score

Conclusion : le meilleur estimateur pour prédire Log-TotalGHGEmissions est le LASSO que l'on sauvegarde sous le nom de « best_model_ENERGYSTARScore

Erreur de prédiction avec le LASSO (meilleur modèle pour prédire le CO2)

Scatter plot of the predicted values as a function of the true values ; ln(1+x)

Scatter plot of the predicted values as a function of the true values ; converted with exp(x)-1



Conclusion

Intérêt de l'ENERGY STAR Score

R2 score pour le LASSO avec CV = 0.58 avec ENERGY STAR Score

contre R2 = 0.53 sans ENERGY STAR Score.

Donc, l'ENERGY STAR Score améliore les prédictions de CO2.

Cependant, le calcul de l'ENERGY STAR Score étant fastidieux, il est peu intéressant de le calculer pour l'amélioration des prédictions possibles.

Merci !