Seattle

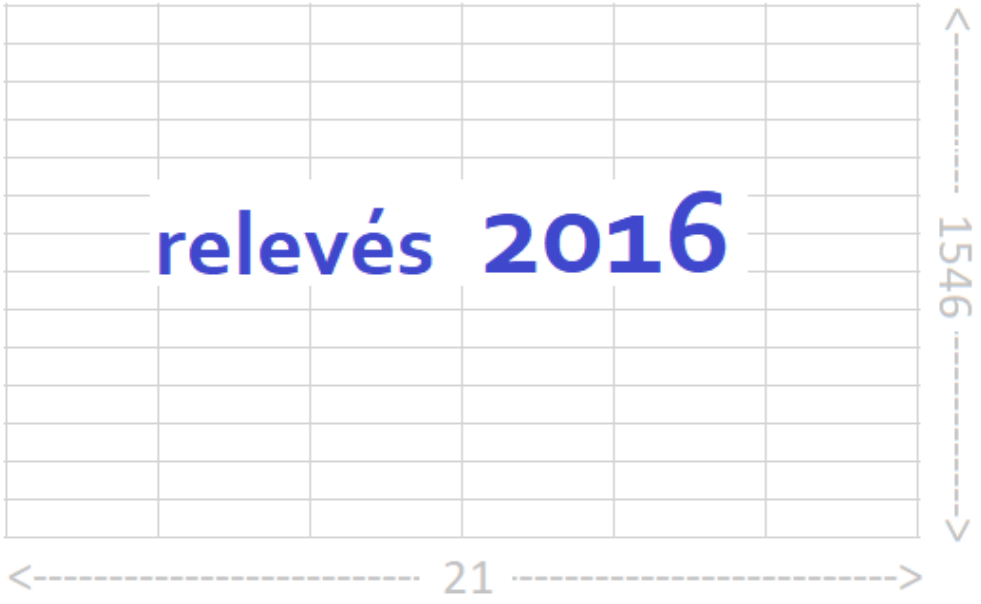
* émissions de CO2
* consommation totale d’énergie

Sommaire

1. Objectifs
2. Données
   a. Source
   b. Features retenues
   c. Qualité
   d. Nettoyage – Complétion – Feature engineering
3. Modélisation
4. Optimisation
5. Interprétation des résultats

Source des Données

Ville de **SEATTLE**



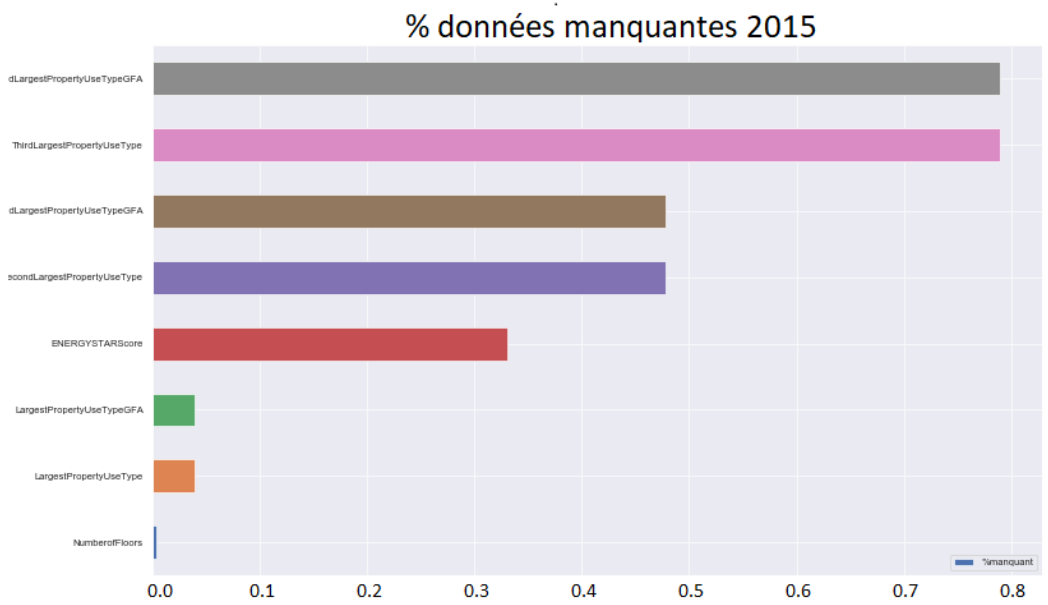
Features retenues

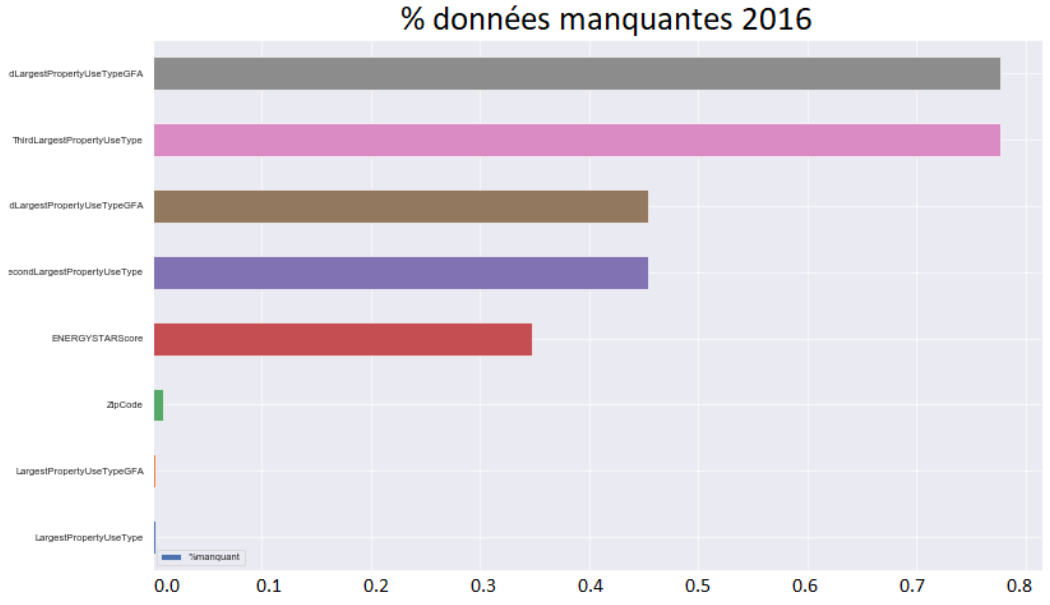
BuildingId	clé	
age		age du bâtiment
NumberofFloors		Nombre d'étages
CouncilDistrictCode Neighborhood ZipCode Latitude - Longitude		localisation
PropertyGFATotal PropertyGFABuildings PropertyGFAParking LargestPropertyUseType LargestPropertyUseTypeGFA SecondLargestPropertyUseType SecondLargestPropertyUseTypeGFA ThirdLargestPropertyUseType ThirdLargestPropertyUseTypeGFA	X	Surface
ENERGYSTARScore		
SiteEnergyUse	y	énergie consommée
GHGEmissionsCO2		émission de CO2

X = features

y = cibles

Qualité des Données





NumberofFloors	0.45
LargestPropertyUseTypeGFA	3.99
LargestPropertyUseType	3.99
ENERGYSTARScore	33.02
SecondLargestPropertyUseType	47.77
SecondLargestPropertyUseTypeGFA	47.77
ThirdLargestPropertyUseType	78.89
ThirdLargestPropertyUseTypeGFA	78.89

% données manquantes

0.25	LargestPropertyUseType
0.25	LargestPropertyUseTypeGFA
1.03	ZipCode
34.80	ENERGYSTARScore
45.45	SecondLargestPropertyUseType
45.45	SecondLargestPropertyUseTypeGFA
77.66	ThirdLargestPropertyUseType
77.66	ThirdLargestPropertyUseTypeGFA

Nettoyage

1. (‘SiteEnergyUse’ = 0) ou ('GHGEmissionsCO2’ = 0)
2. Doublon (‘BuildingId’)
3. (‘PropertyGFABuildings’ < 0)
4. ('LargestPropertyUseTypeGFA’ = 0 & 'PropertyGFATotal’ > 0)

? ('NumberofFloors’] = 0) ?

? (‘PropertyGFATotal’ =! ‘LargestPropertyUseTypeGFA’ +
‘SecondLargestPropertyUseTypeGFA’ + ‘ThirdLargestPropertyUseTypeGFA’) ?

Complétion

2015

ZipCode
ENERGYSTARScore
LargestPropertyUseType
LargestPropertyUseTypeGFA
SecondLargestPropertyUseType
SecondLargestPropertyUseTypeGFA
ThirdLargestPropertyUseType
ThirdLargestPropertyUseTypeGFA

2016

ZipCode	0
ENERGYSTARScore	29
LargestPropertyUseType	0
LargestPropertyUseTypeGFA	0
SecondLargestPropertyUseType	2
SecondLargestPropertyUseTypeGFA	2
ThirdLargestPropertyUseType	2
ThirdLargestPropertyUseTypeGFA	3

nombre de complétions réalisées

Feature engineering

- Ajout de la variable :
  ‘PercentParking’ = ‘PropertyGFAParking’ / 'PropertyGFATotal’

- Modification : ‘age’ = ‘YearBuilt’ - 2020

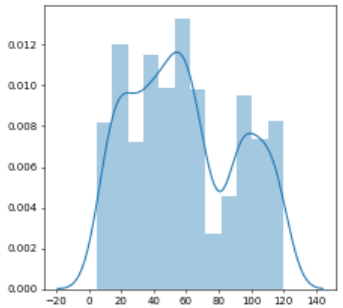
- Encodage des variables catégorielles :
  ‘CouncilDistrictCode’, ‘Neighborhood’, ‘ZipCode’,
  ‘LargestPropertyUseType’,
  ‘SecondLargestPropertyUseType’,
  ‘ThirdLargestPropertyUseType’ .

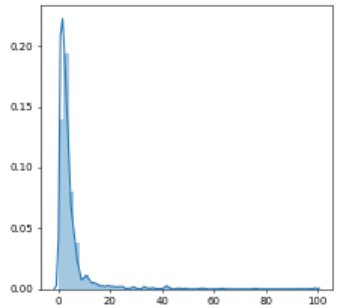
Statistiques Produits

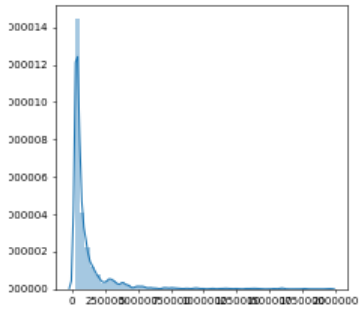
	Latitude	Longitude	age	NumberofFloors	PropertyGFATotal	PropertyGFAParking	PropertyGFABuildings	LargestPropertyUseTypeGFA
count	1540.000000	1540.000000	1540.000000	1540.000000	1.540000e+03	1540.000000	1.540000e+03	1.540000e+03
mean	47.616116	122.333380	58.834416	4.287662	1.111861e+05	13876.837662	9.730930e+04	8.983416e+04
median	47.61	122.33	55	2	4.72260e+04	0.0000	4.50750e+04	4.1291e+04
min	47.509590	122.261800	5.000000	0.000000	1.128500e+04	0.000000	3.636000e+03	5.656000e+03
max	47.733870	122.411820	120.000000	99.000000	1.952220e+06	512608.000000	1.765970e+06	1.680937e+06

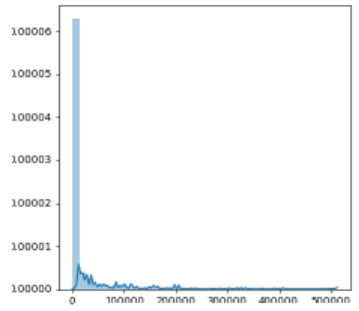
	SecondLargestPropertyUseTypeGFA	ThirdLargestPropertyUseTypeGFA	ENERGYSTARScore	SiteEnergyUse	GHGEmissionsCO2	PercentParking
	1540.000000	1540.00000	1540.000000	1.540000e+03	1540.000000	1540.000000
	19176.971621	3014.24896	42.627922	7.653865e+06	165.175396	0.066027
	0.0000	0.0000	44.00	2.615611e+06	48.51	0.0000
	0.000000	0.00000	0.000000	1.680890e+04	0.120000	0.000000
	441551.000000	303910.00000	100.000000	2.930908e+08	12307.160000	0.895023

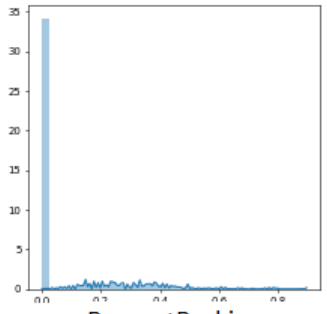
Analyse univariée

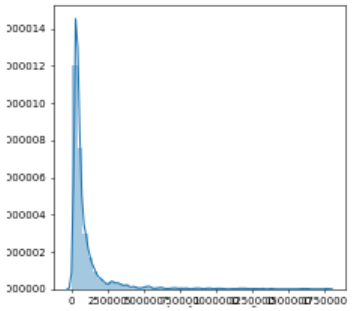


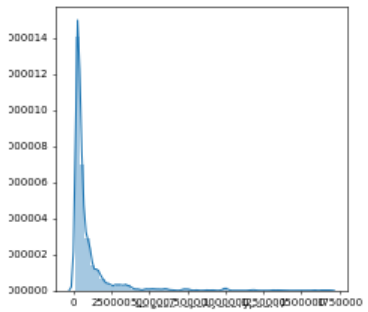


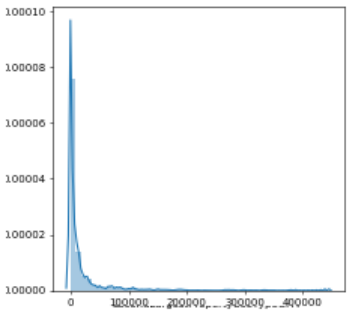


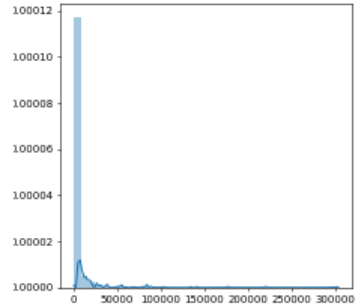


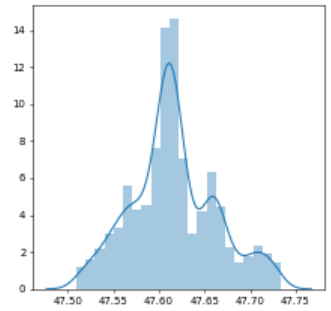


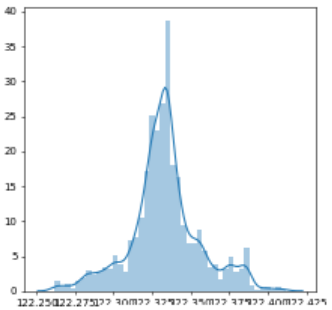


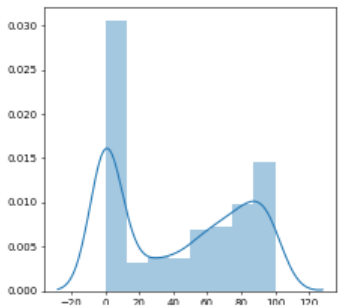


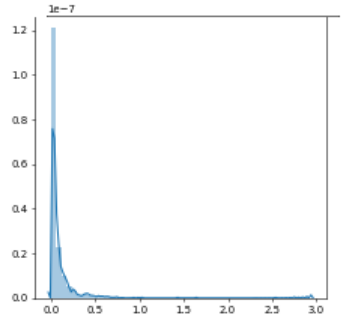


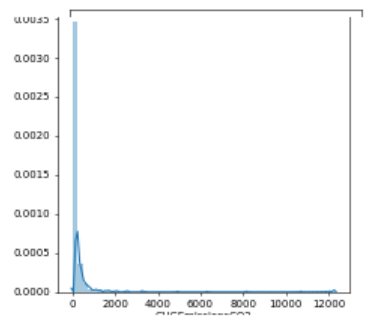












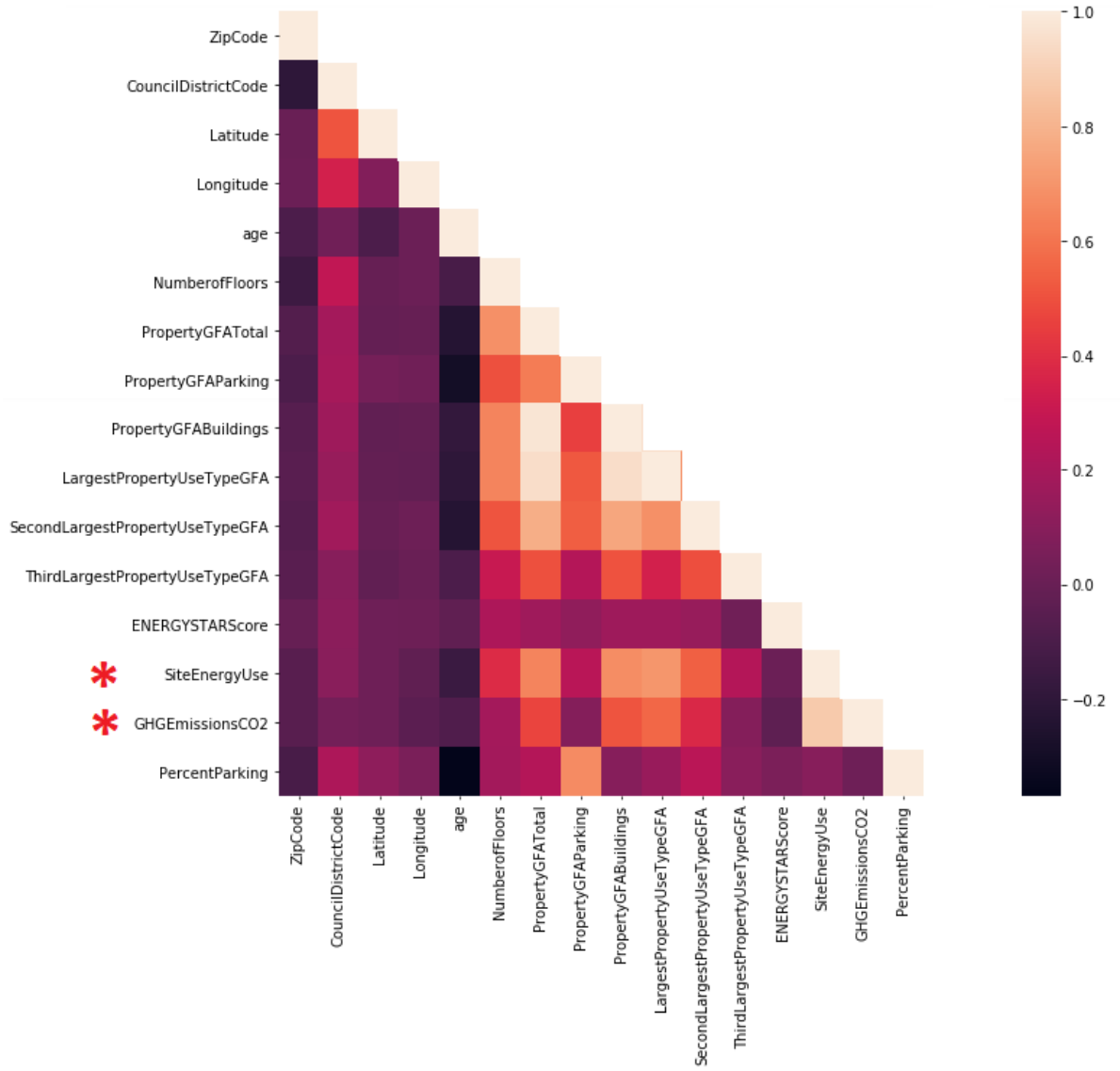
Analyse bivariée – Corrélation linéaire de ‘SiteEnergyUse’

Variable	Coef. Pearson
'GHGEmissionsCO2'	0.881
'LargestPropertyUseTypeGFA'	0.703
'PropertyGFABuildings'	0.675
'PropertyGFATotal'	0.649
'LargestPropertyUseType_Hospital (General Medical & Surgical)'	0.579
'SecondLargestPropertyUseTypeGFA'	0.541
'NumberofFloors'	0.390
'PropertyGFAParking'	0.256
'LargestPropertyUseType_Data Center'	0.255
'ThirdLargestPropertyUseTypeGFA'	0.238
'SecondLargestPropertyUseType_Parking'	0.196
'ZipCode_98101.0'	0.147
'ThirdLargestPropertyUseType_Other/Specialty Hospital'	0.129
...	

Analyse bivariée – Corrélation linéaire de ‘GHGEmissionsCO2’

Variable	Coef. Pearson
'SiteEnergyUse'	0.881
'LargestPropertyUseType_Hospital (General Medical & Surgical'	0.691
'LargestPropertyUseTypeGFA'	0.565
'PropertyGFABuildings'	0.508
'PropertyGFATotal'	0.465
'SecondLargestPropertyUseTypeGFA'	0.377
'NumberofFloors'	0.188
'Neighborhood_EAST'	0.151
'CouncilDistrictCode_3'	0.139
'ThirdLargestPropertyUseType_Other/Specialty Hospital'	0.137
'SecondLargestPropertyUseType_Parking'	0.126
'ZipCode_98101.0'	0.105
'LargestPropertyUseType_Laboratory'	0.104
...	

Analyse bivariée



coefficient de corrélation (r) entre deux variables aléatoires réelles X et Y :

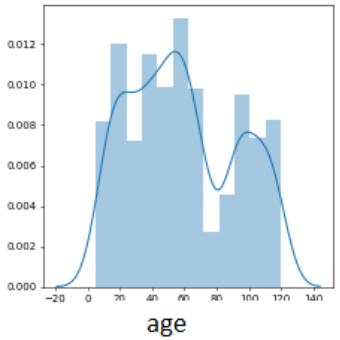
$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

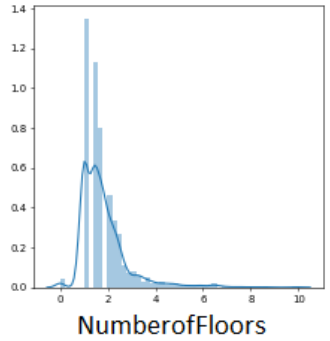
$\mathrm{Cov}(X,Y)$: covariance des variables X et Y

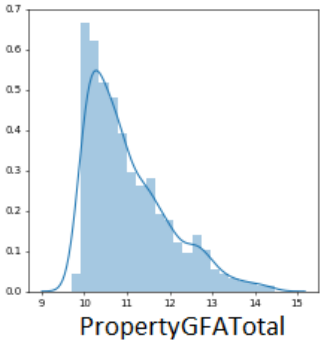
σ_X et σ_Y : leurs écarts types.

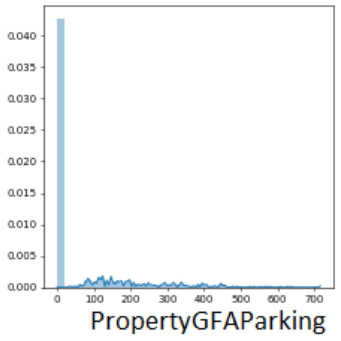
Corrélation	Négative	Positive
Faible	-0,5 à 0,0	0,0 à 0,5
Forte	−1,0 à −0,5	0,5 à 1,0

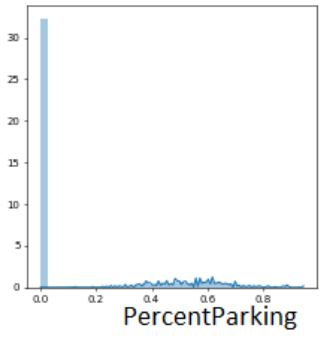
Analyse univariée – après traitement

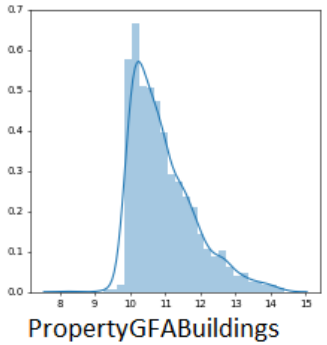

age

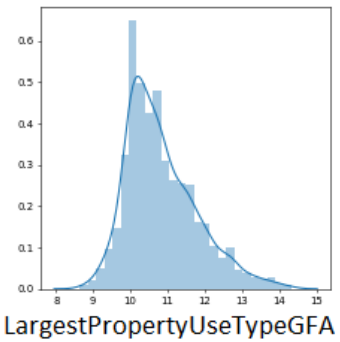

NumberofFloors

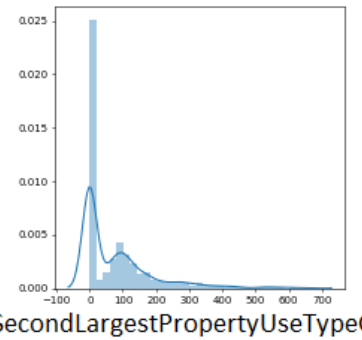

PropertyGFATotal

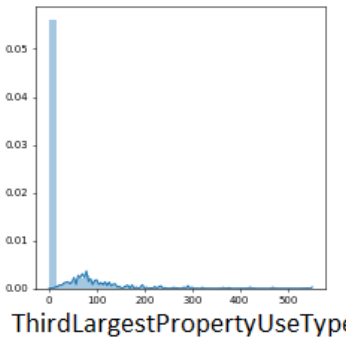

PropertyGFAParking

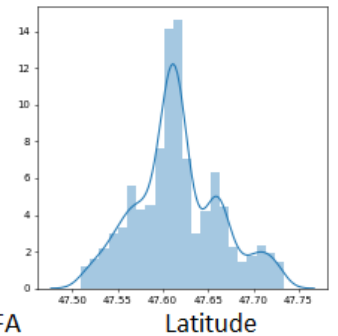

PercentParking

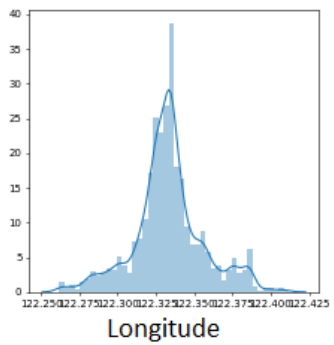

PropertyGFABuildings

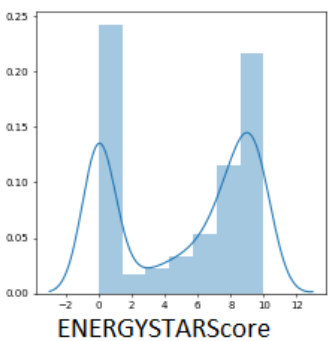

LargestPropertyUseTypeGFA

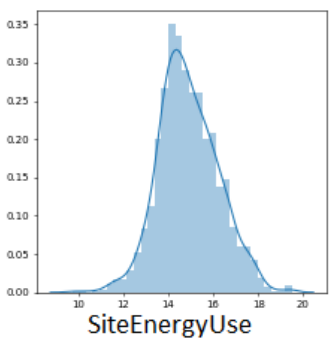

SecondLargestPropertyUseTypeGFA

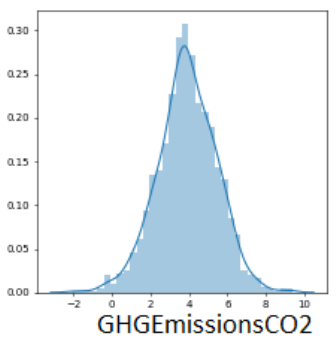

ThirdLargestPropertyUseTypeGFA


Latitude


Longitude


ENERGYSTARScore


SiteEnergyUse

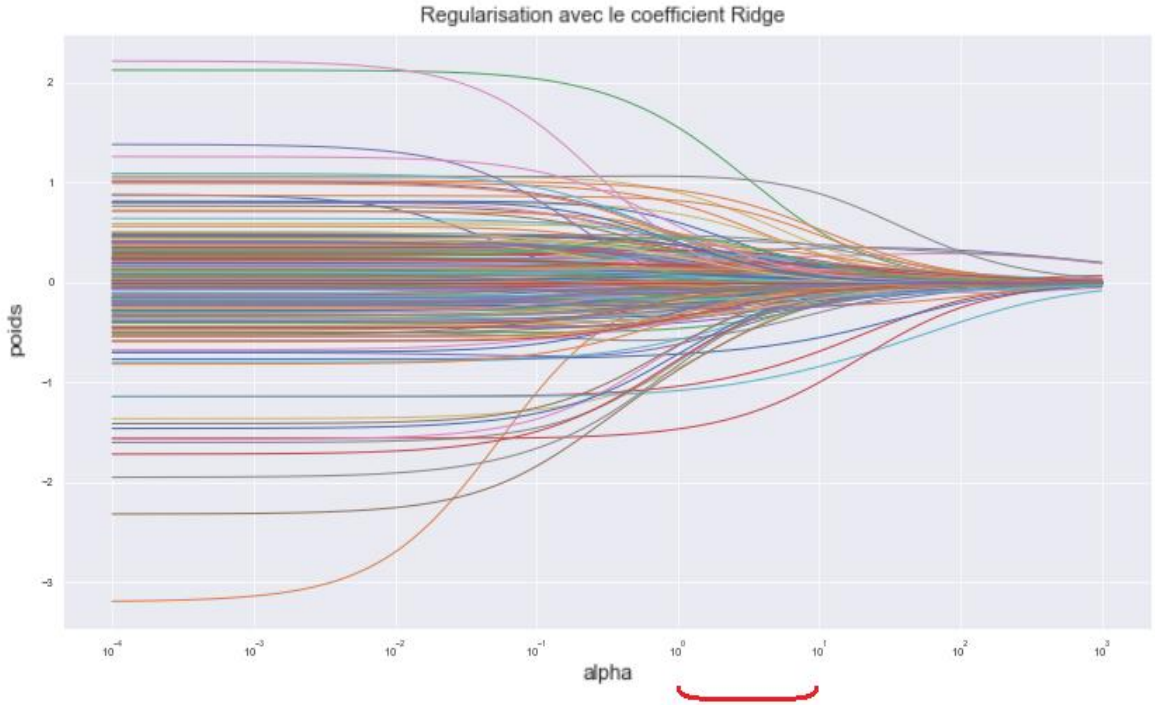

GHGEmissionsCO2

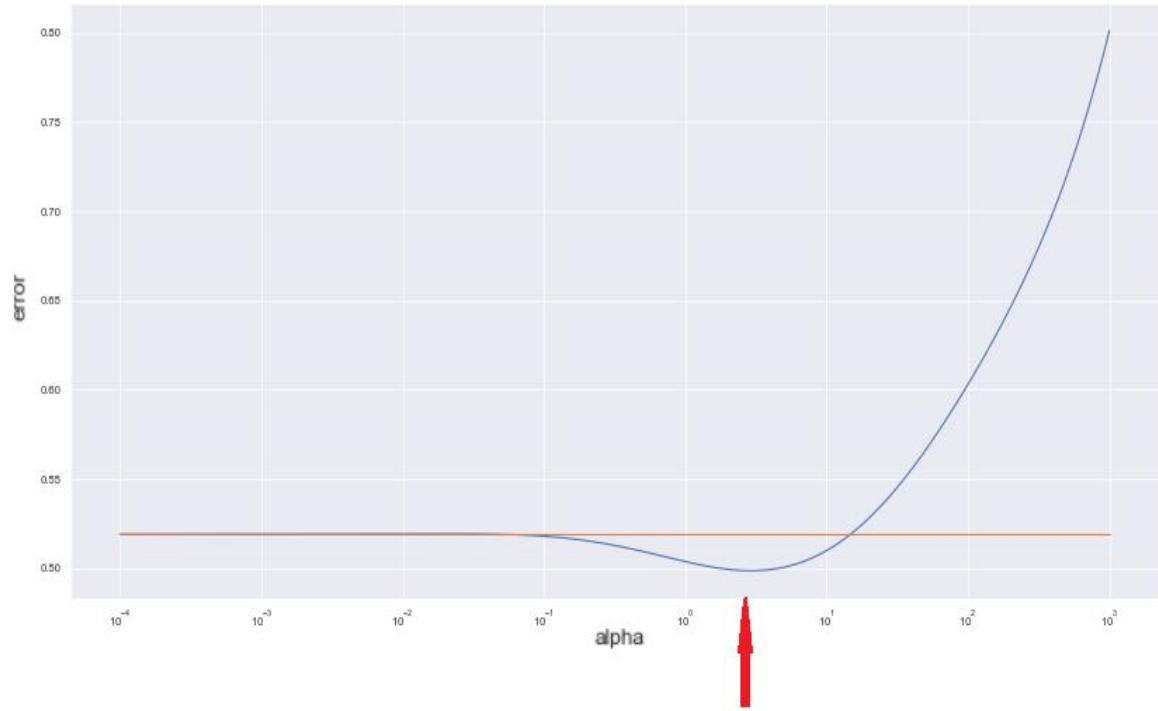
Modélisation

- Ridge
- Lasso
- ElasticNet
- Arbres de décision
- Bagging
- Forêts aléatoires

- Boosting
  - AdaBoost
  - GradientBoosting
  - XGBoost

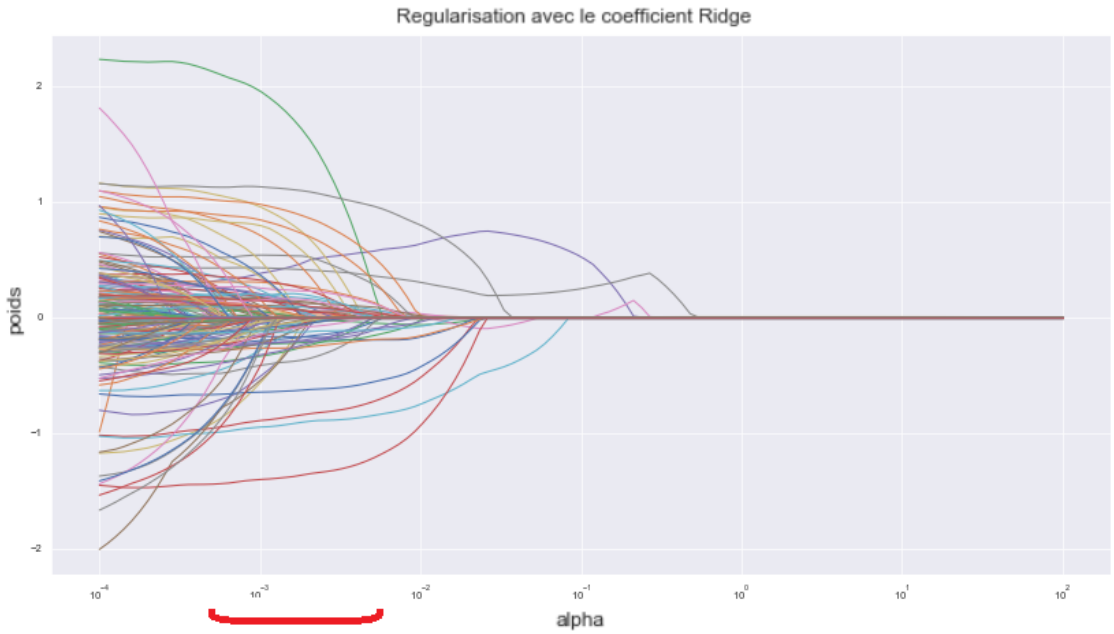
Ridge

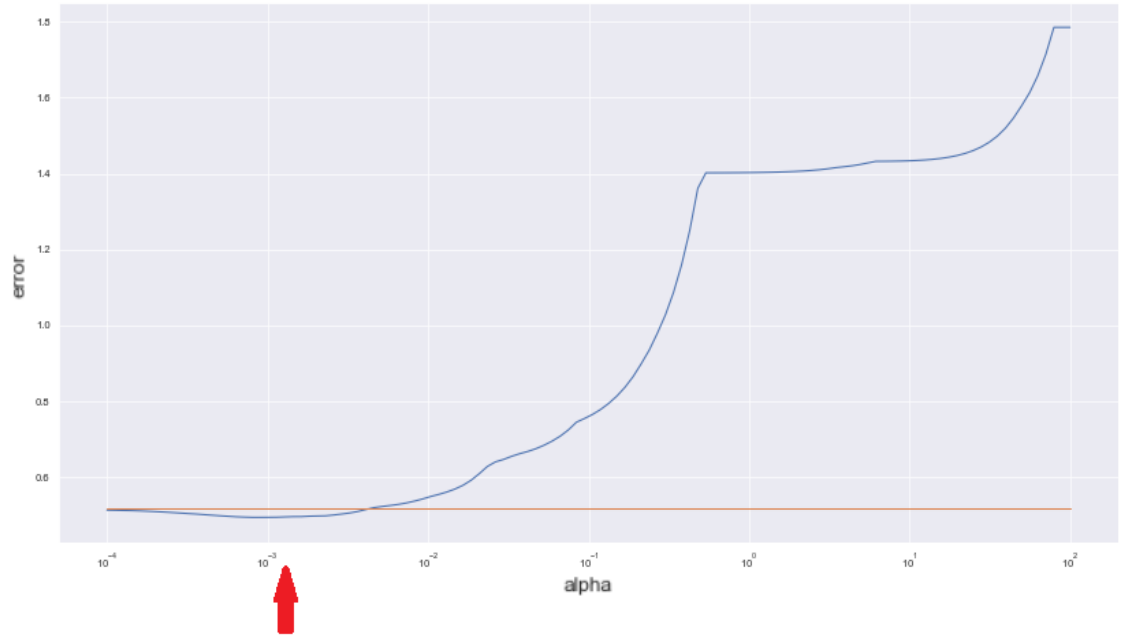





```
R2  = 0.7205
MSE = 0.4989
Meilleur alpha = 3.6061
```

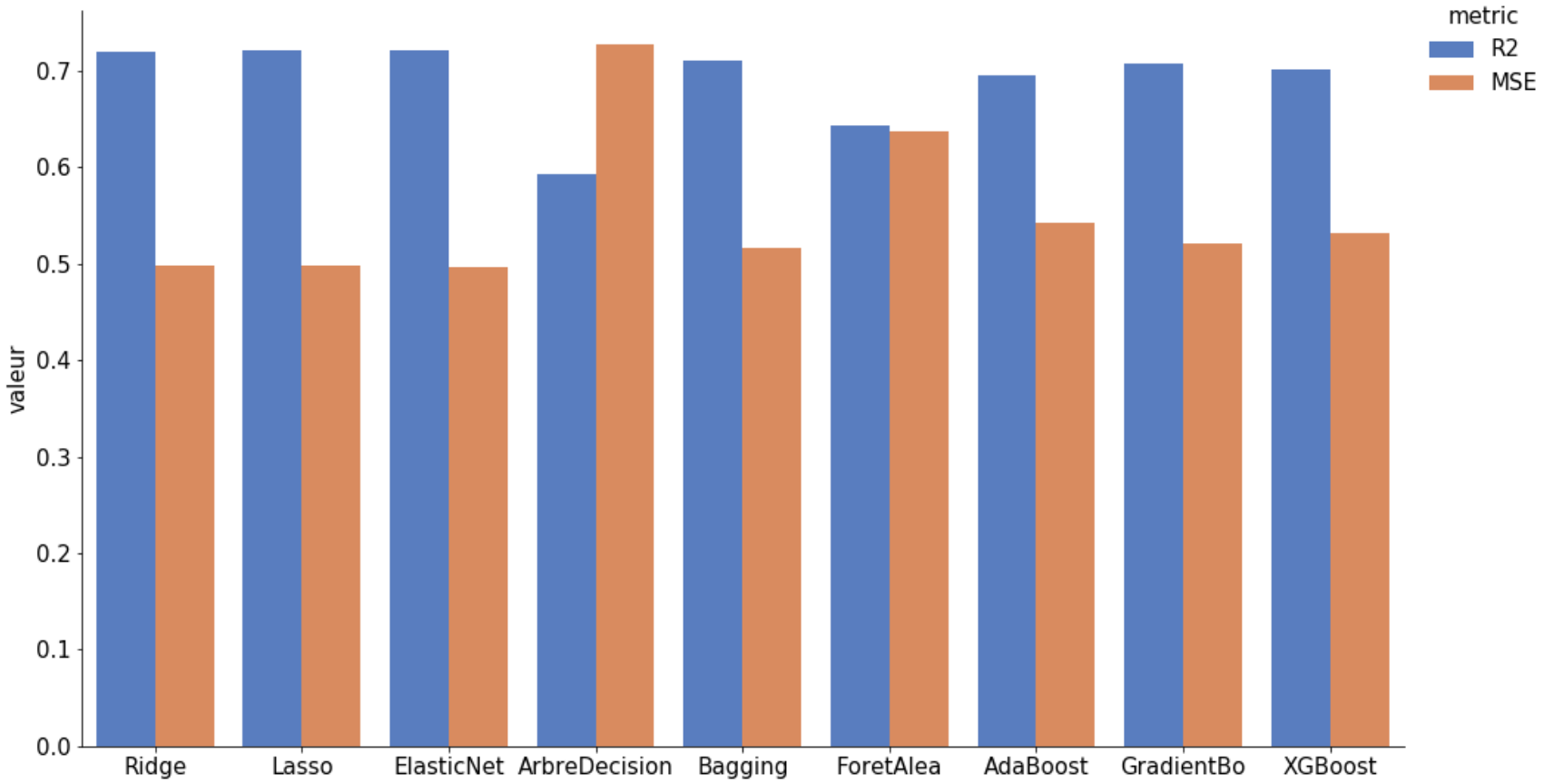
Lasso






```
R2  = 0.7211
MSE = 0.4978
Meilleur alpha = 0.0018
```

Résultats modélisation – SiteEnergyUse (1)



Résultats modélisation – SiteEnergyUse (2)

Modèle	R2	MSE	Hyper-paramètres
Ridge	0.7205	0.4989	(Meilleur alpha = 3.606)
Lasso	0.7211	0.4978	(Meilleur alpha = 0.018)
ElasticNet	0.7220	0.4962	(Meilleur alpha = 0.0013, Meilleur ratio = 1.7)
Arbre de décision	0.5927	0.7269	(max_depth=6, min_samples_leaf=10, min_samples_split=12)
Bagging	0.7106	0.5165	(n_estimators=90, max_samples=0.99, max_features=0.90)
Foret Alea	0.6433	0.6367	(n_estimators=100, max_features=0.9, max_samples=0.99, min_samples_leaf=20, min_samples_split=8)
AdaBoost	0.6962	0.5422	(n_estimators=60, learning_rate=1.0, loss='square')
GradientBo	0.7081	0.5211	(n_estimators=60, learning_rate=0.1, loss='ls')
XGBoost	0.7022	0.5317	{'colsample_bytree': 0.7, 'learning_rate': 0.1, 'max_depth': 5, 'min_child_weight': 4, 'n_estimators': 60, 'nthread': 4, 'objective': 'reg:linear', 'subsample': 0.7}

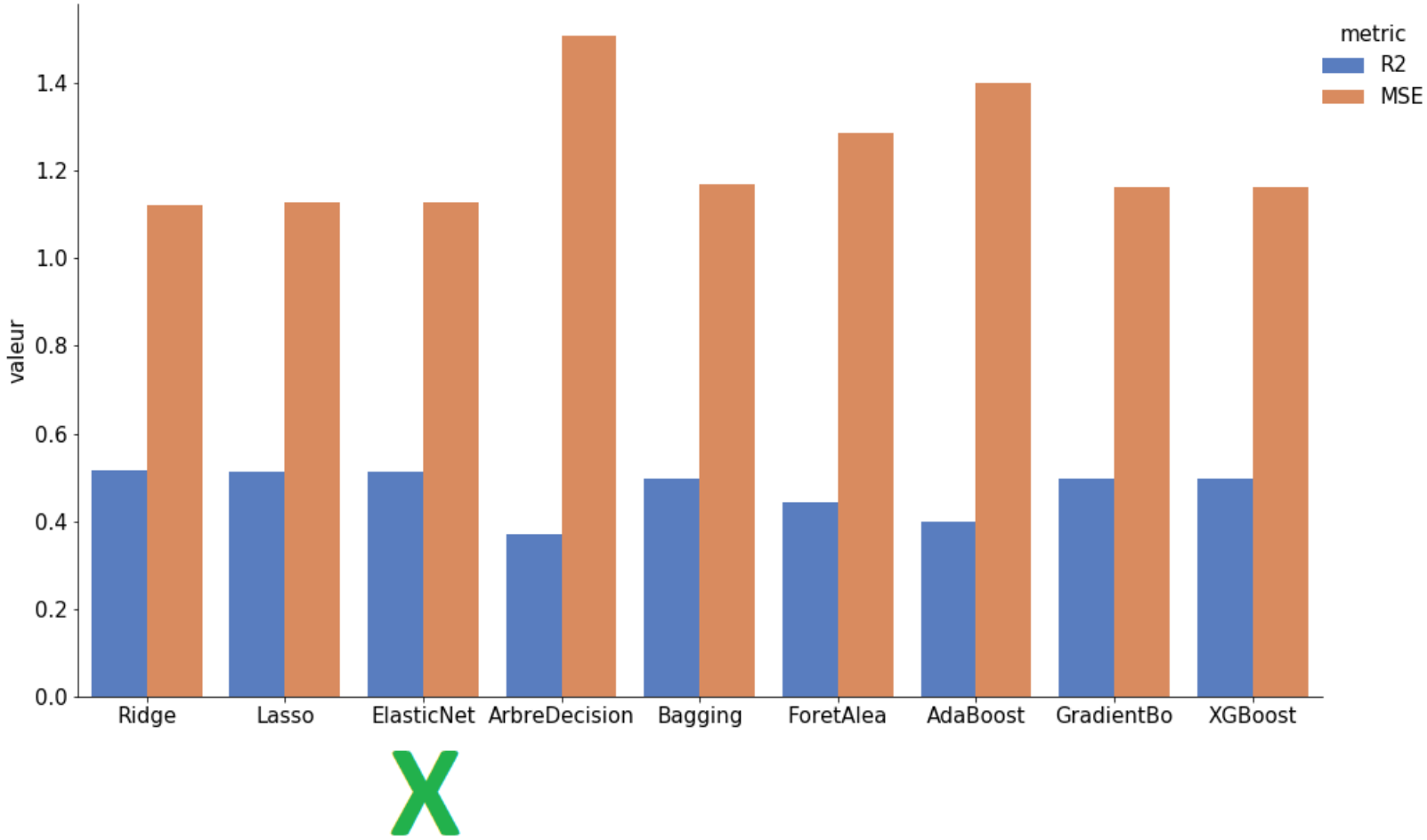
Optimisation du modèle - SiteEnergyUse

Modèle	R2	MSE	
ElasticNet	0.7220	0.4962	(Meilleur alpha = 0.00133, Meilleur ratio = 1.7)
seuil= 0.5	0.7050	0.5266	12 variables
+ENERGYSTARScore	0.7430	0.4588	13 variables
+ENERGYSTARScore non null	0.8113	0.3099	

Features sélectionnées - SiteEnergyUse

Feature	Coefficient
LargestPropertyUseType_Data Center	2.8106043525109135
LargestPropertyUseType_Supermarket/Grocery Store	1.097720455680584
PropertyGFATotal	1.0434163162244234
LargestPropertyUseType_Hospital (General Medical & Surgical	0.8660382865404561
LargestPropertyUseType_Laboratory	0.0
LargestPropertyUseType_Other - Recreation',	0.0
LargestPropertyUseType_Restaurant',	0.0
LargestPropertyUseType_Urgent Care/Clinic/Other Outpatient	0.0
ENERGYSTARScore	-0.16832379173296957
LargestPropertyUseType_Worship Facility	-0.5179834868009998
LargestPropertyUseType_Distribution Center	-0.7676170751353568
LargestPropertyUseType_Non-Refrigerated Warehouse	-0.9580233060870447
LargestPropertyUseType_Self-Storage Facility	-3.807395449887726

Résultats modélisation - GHGEmissionsCO2 (1)



Résultats modélisation - GHGEmissionsCO2

Modèle	R2	MSE	Hyper-paramètres
Ridge	0.5162	1.1234	(Meilleur alpha = 6.1936)
Lasso	0.5146	1.1272	(Meilleur alpha = 0.0031)
ElasticNet	0.5144	1.1277	(Meilleur alpha = 0.0032, Meilleur ratio = 1.0)
Arbre de décision	0.3706	1.5072	(max_depth=6, min_samples_leaf=10, min_samples_split=3)
Bagging	0.4962	1.1699	(n_estimators=50, max_samples=0.95, max_features=0.99)
Foret Alea	0.4424	1.2875	(n_estimators=80, max_features=0.99, max_samples=0.99, min_samples_leaf=20, min_samples_split=10)
AdaBoost	0.40	1.40	(n_estimators=60, learning_rate=1.0, loss='exponential')
GradientBo	0.4984	1.1649	(n_estimators=60, learning_rate=0.1, loss='ls')
XGBoost	0.4989	1.1637	{'colsample_bytree': 0.7, 'learning_rate': 0.1, 'max_depth': 6, 'min_child_weight': 4, 'n_estimators': 60, 'nthread': 4, 'objective': 'reg:linear', 'subsample': 0.7}

Optimisation du modèle - GHGEmissionsCO2

Modèle	R2	MSE	
ElasticNet	0.5144	1.1277	(Meilleur alpha = 0.0032, Meilleur ratio = 1.0)
seuil= 0.3	0.5028	1.1547	18 colonnes
+ENERGYSTARScore	0.5395	1.0693	19 colonnes
+ENERGYSTARScore non null	0.5946	0.9190	

Features sélectionnées- GHGEmissionsCO2

Feature	Coefficient
LargestPropertyUseType_Supermarket/Grocery Store	1.153
LargestPropertyUseType_Hospital (General Medical & Surgical	0.990
LargestPropertyUseType_Hotel	0.751
LargestPropertyUseTypeGFA	0.669
LargestPropertyUseType_Senior Care Community	0.643
PropertyGFABuildings	0.353
LargestPropertyUseType_Laboratory	0.0
LargestPropertyUseType_Multifamily Housing	-0.0
LargestPropertyUseType_Other - Recreation	0.0
LargestPropertyUseType_Parking	-0.0
LargestPropertyUseType_Restaurant	0.0
LargestPropertyUseType_Self-Storage Facility	-0.0
ENERGYSTARScore	-0.166
LargestPropertyUseType_Retail Store	-0.250
SecondLargestPropertyUseType_Parking	-0.325
LargestPropertyUseType_Office	-0.461
SecondLargestPropertyUseType_	-0.535
LargestPropertyUseType_Distribution Center	-0.813
argestPropertyUseType_Non-Refrigerated Warehouse	-1.137

Analyse bivariée – Corrélation non linéaire de ‘SiteEnergyUse’

Variable	Coef. Spearman
'PropertyGFATotal'	0.737
'PropertyGFABuildings'	0.719
'LargestPropertyUseTypeGFA'	0.702
'NumberofFloors'	0.452
'PropertyGFAParking'	0.370
'SecondLargestPropertyUseTypeGFA'	0.356
'PercentParking'	0.335
'SecondLargestPropertyUseType_Parking'	0.312
'CouncilDistrictCode_7'	0.242
'Neighborhood_DOWNTOWN'	0.231
'ZipCode_98101.0'	0.230
...	

Analyse bivariée – Corrélation non linéaire de ‘GHGEmissionsCO2’

Variable	Coef. Spearman
'PropertyGFABuildings'	0.568
'PropertyGFATotal'	0.568
'LargestPropertyUseTypeGFA'	0.551
'NumberofFloors'	0.308
'SecondLargestPropertyUseTypeGFA'	0.247
'PropertyGFAParking'	0.221
'LargestPropertyUseType_Hotel'	0.220
'ZipCode_98101.0'	0.204
'SecondLargestPropertyUseType_Parking'	0.201
'PercentParking'	0.185
'CouncilDistrictCode_7'	0.179
'ThirdLargestPropertyUseTypeGFA'	0.171
...	