Objectif Neutralité Carbone 2050

Anticiper les besoins en consommation de bâtiments

Cédric - Octobre 2024

Déroulement de la mission

- Objectifs de l'étude
- Analyse exploratoire & premier feature engineering
- Création d'un modèle Baseline
- Amélioration du feature engineering
- Simulation de différents modèles et choix du modèle final
- Analyse de la feature importance
- Influence de l'Energie Star Score

Objectifs de l'étude

Neutralité carbone de Seattle en 2050

Prédire les émissions de CO2 et la consommation totale d'énergie de bâtiments non destinés à l'habitation

Evaluer l'intérêt de l'Energy Star Score pour les prédictions

Se passer des relevés annuels très couteux après un relevé de référence

Analyse exploratoire & feature engineering

Source: Open Data Program de la ville de Seattle (https://data.seattle.gov/)

DataYear	BuildingType	PrimaryPropertyType	PropertyName	Address	City	State	ZipCode	TaxParcelIdentificationNumber	...	Electricity(kWh)	Electricity(kBtu)	NaturalGas(therms)	NaturalGas(kBtu)	DefaultData	Comments	ComplianceStatus
2016	NonResidential	Hotel	Mayflower park hotel	405 Olive way	Seattle	WA	98101.0	0659000030	...	1.156514e+06	3946027.0	12764.52930	1276453.0	False	NaN	Compliant
2016	NonResidential	Hotel	Paramount Hotel	724 Pine street	Seattle	WA	98101.0	0659000220	...	9.504252e+05	3242851.0	51450.81641	5145082.0	False	NaN	Compliant
2016	NonResidential	Hotel	5673-The Westin Seattle	1900 5th Avenue	Seattle	WA	98101.0	0659000475	...	1.451544e+07	49526664.0	14938.00000	1493800.0	False	NaN	Compliant
2016	NonResidential	Hotel	HOTEL MAX	620 STEWART ST	Seattle	WA	98101.0	0659000640	...	8.115253e+05	2768924.0	18112.13086	1811213.0	False	NaN	Compliant
2016	NonResidential	Hotel	WARWICK SEATTLE HOTEL (ID8)	401 LENORA ST	Seattle	WA	98121.0	0659000970	...	1.573449e+06	5368607.0	88039.98438	8803998.0	False	NaN	Compliant

Données de consommation énergétiques de 2016
46 Variables
3376 Individus
Pas de doublons dans les relevés
12,85% Valeurs NaN

Identification des variables cibles

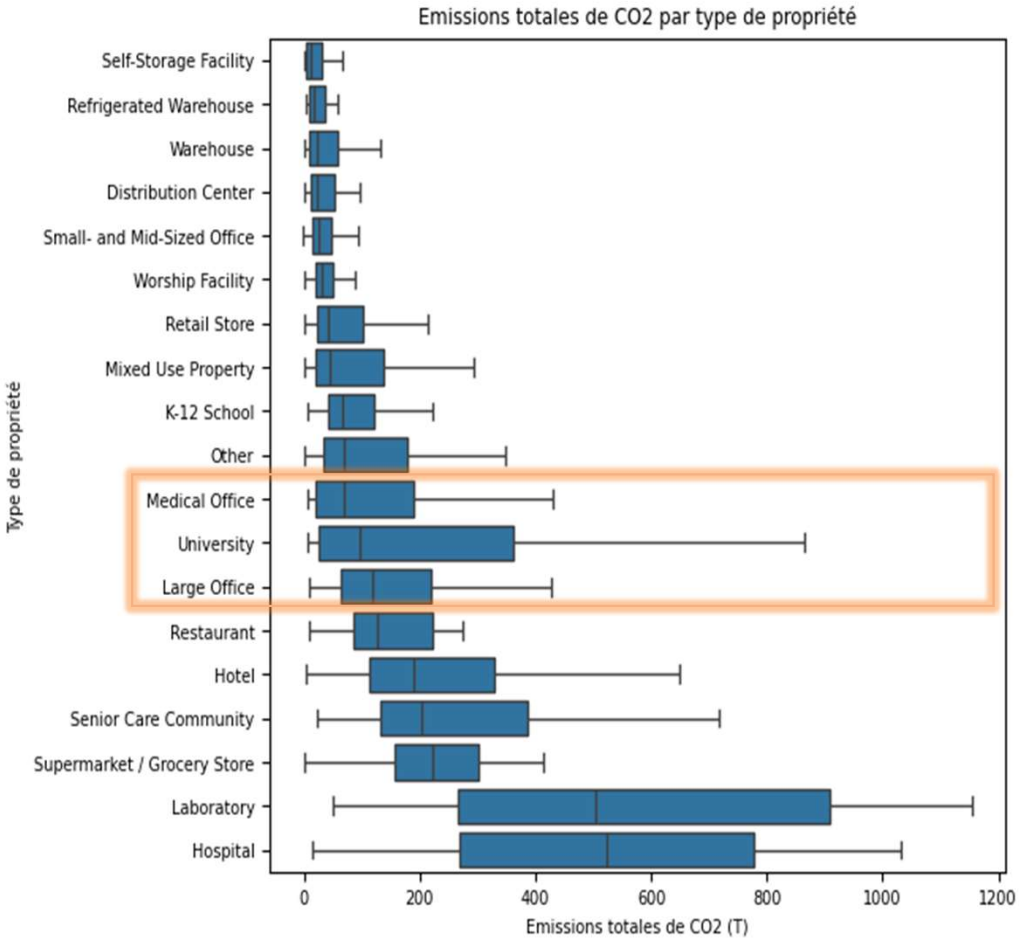
Emissions de CO2 : TotalGHGEmissions

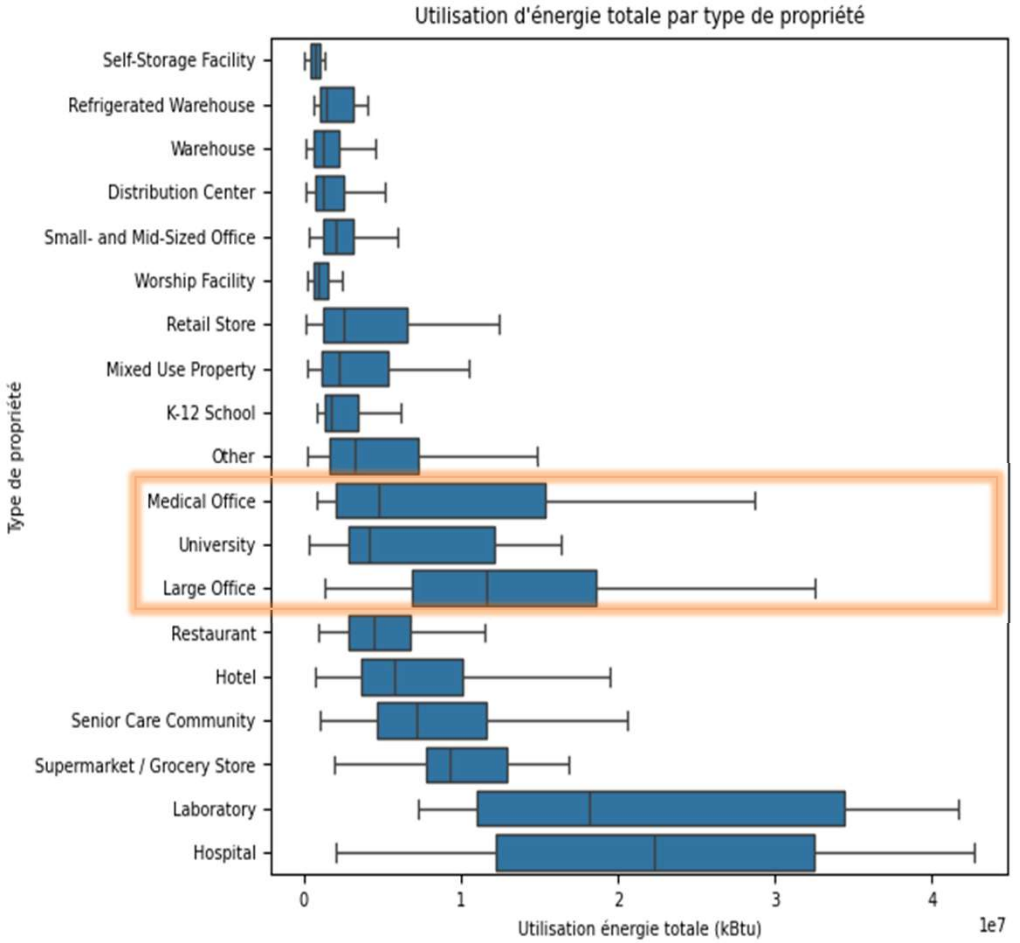
Consommation totale d'énergie : SiteEnergyUse(kBtu)

	OSEBuildingID	ZipCode	CouncilDistrictCode	Latitude	Longitude	YearBuilt	NumberofBuildings	NumberofFloors	PropertyGFATotal	PropertyGFAParking	PropertyGFABuilding(s)	LargestPropertyUseTypeGFA	SecondLargestPropertyUseTypeGFA
count	1512.000000	1499.000000	1512.000000	1512.000000	1512.000000	1512.000000	1512.000000	1512.000000	1.512000e+03	1512.000000	1.512000e+03	1.512000e+03	1512.000000
mean	16236.841931	98116.643763	4.416005	47.615949	-122.333957	1961.551587	1.101190	4.257275	1.142097e+05	14100.925265	1.001087e+05	9.270012e+04	19435.530620
std	13649.698851	18.234523	2.205122	0.047110	0.023078	32.810266	1.176518	6.389669	1.967924e+05	44181.698581	1.736075e+05	1.623371e+05	51004.157677
min	1.000000	98006.000000	1.000000	47.509590	-122.411820	1900.000000	0.000000	1.000000	1.128500e+04	0.000000	3.636000e+03	0.000000e+00	0.000000
25%	596.750000	98104.000000	2.000000	47.586950	-122.343290	1930.000000	1.000000	1.000000	2.884700e+04	0.000000	2.793675e+04	2.499750e+04	0.000000
50%	21147.500000	98109.000000	4.000000	47.612335	-122.333275	1965.000000	1.000000	2.000000	4.812550e+04	0.000000	4.606350e+04	4.170550e+04	0.000000
75%	24589.500000	98125.000000	7.000000	47.647960	-122.323215	1988.000000	1.000000	4.000000	1.044605e+05	0.000000	9.441750e+04	9.000000e+04	13623.250000
max	50226.000000	98199.000000	7.000000	47.733870	-122.261800	2015.000000	27.000000	76.000000	2.200000e+06	512608.000000	2.200000e+06	1.719643e+06	639931.000000

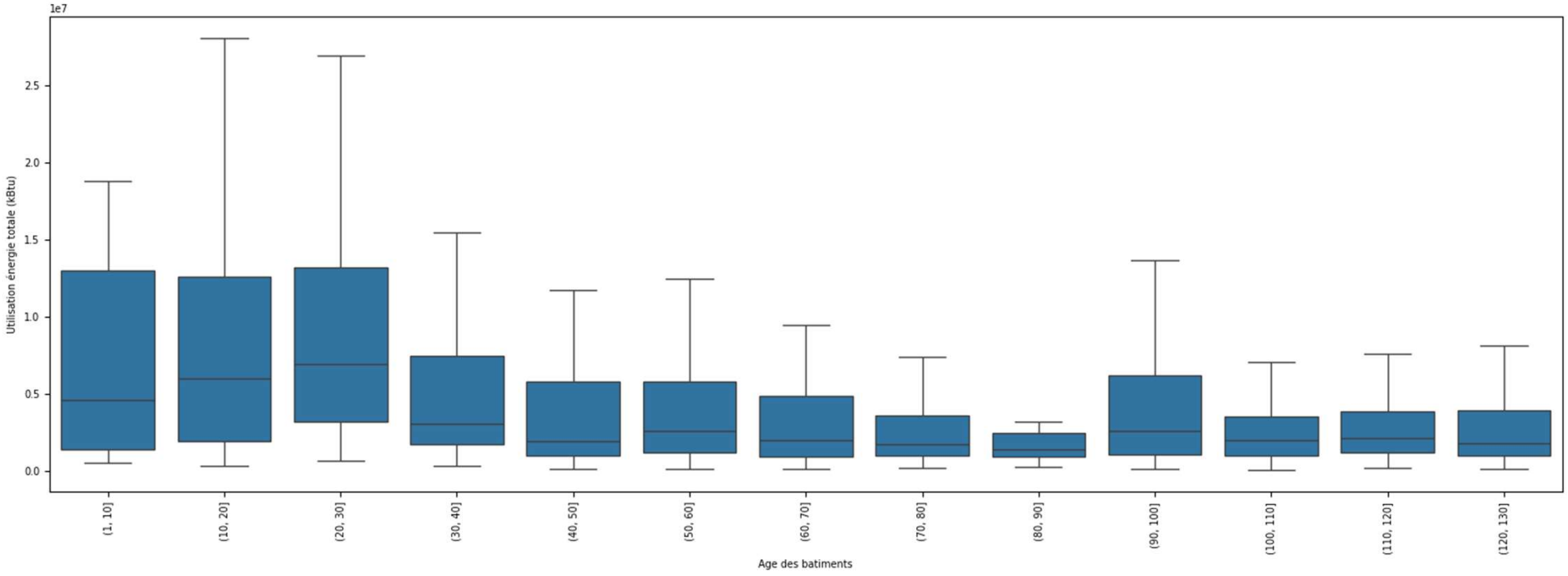
	ThirdLargestPropertyUseTypeGFA	ENERGYSTARScore	SiteEUI(kBtu/sf)	SiteEUIWN(kBtu/sf)	SourceEUI(kBtu/sf)	SourceEUIWN(kBtu/sf)	SiteEnergyUse(kBtu)	SiteEnergyUseWN(kBtu)	SteamUse(kBtu)	Electricity(kBtu)	NaturalGas(kBtu)	TotalGHGEmissions	GHGEmissionsIntensity
count	1512.000000	970.000000	1512.000000	1511.000000	1512.000000	1512.000000	1.512000e+03	1.511000e+03	1.512000e+03	1.512000e+03	1.512000e+03	1512.000000	1512.000000
mean	3311.012168	63.772165	75.350728	77.844342	184.042923	186.297817	8.264211e+06	8.405905e+06	4.908315e+05	5.698424e+06	2.029341e+06	185.390569	1.667269
std	18624.697685	28.912962	75.932488	76.937132	189.802122	189.785588	2.242735e+07	2.293617e+07	5.348316e+06	1.380116e+07	9.798744e+06	734.366023	2.426781
min	0.000000	1.000000	1.400000	0.000000	0.000000	-2.100000	5.713320e+04	0.000000e+00	0.000000e+00	-1.154170e+05	0.000000e+00	-0.800000	-0.020000
25%	0.000000	44.000000	34.900002	37.099998	81.099998	83.875000	1.242775e+06	1.312356e+06	0.000000e+00	7.288942e+05	0.000000e+00	20.297500	0.360000
50%	0.000000	71.000000	53.599998	56.099998	138.849998	142.149994	2.711055e+06	2.810298e+06	0.000000e+00	1.711462e+06	4.761960e+05	49.485000	0.880000
75%	0.000000	89.000000	85.249998	88.100002	213.249996	215.675003	7.267172e+06	7.442882e+06	0.000000e+00	5.185815e+06	1.528624e+06	144.962500	1.950000
max	459748.000000	100.000000	834.400024	834.400024	2620.000000	2620.000000	4.483853e+08	4.716139e+08	1.349435e+08	2.745325e+08	2.979090e+08	16870.980000	34.090000

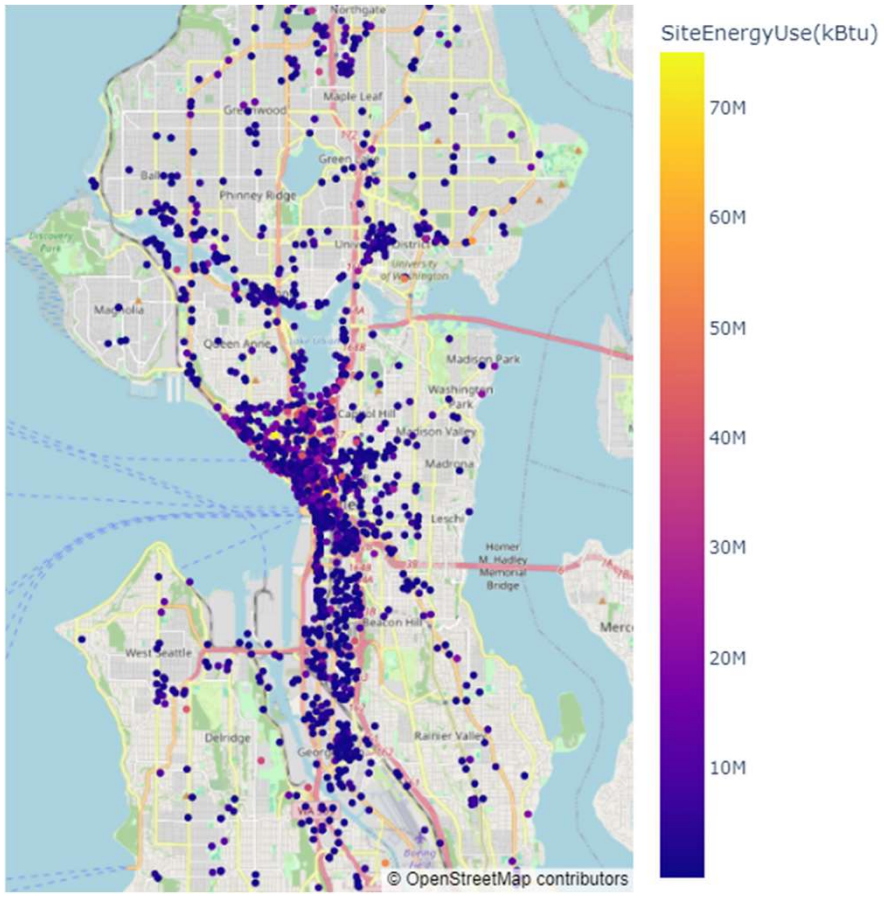

SeattleCity


Emissions totales de CO2 par type de propriété
Self-Storage Facility
Refrigerated Warehouse
Warehouse
Distribution Center
Small- and Mid-Sized Office
Worship Facility
Retail Store
Mixed Use Property
K-12 School
Other
Medical Office
University
Large Office
Restaurant
Hotel
Senior Care Community
Supermarket / Grocery Store
Laboratory
Hospital
Type de propriété
0
200
400
600
800
1000
1200
Emissions totales de CO2 (T)


Utilisation d'énergie totale par type de propriété
Self-Storage Facility
Refrigerated Warehouse
Warehouse
Distribution Center
Small- and Mid-Sized Office
Worship Facility
Retail Store
Mixed Use Property
K-12 School
Other
Medical Office
University
Large Office
Restaurant
Hotel
Senior Care Community
Supermarket / Grocery Store
Laboratory
Hospital
Type de propriété
0
1
2
3
4
1e7
Utilisation énergie totale (kBtu)

Visualisation consommation énergie d'après âge bâtiment





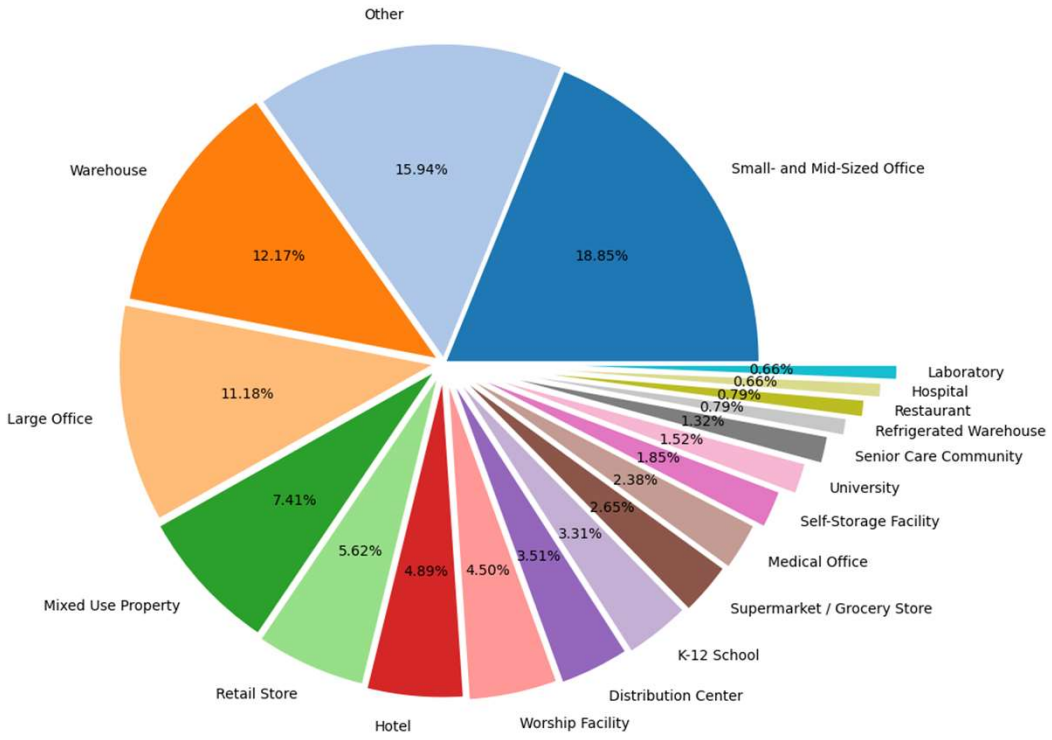
Localisation des bâtiments utilisant le plus d'énergie

Identification de plusieurs foyers :

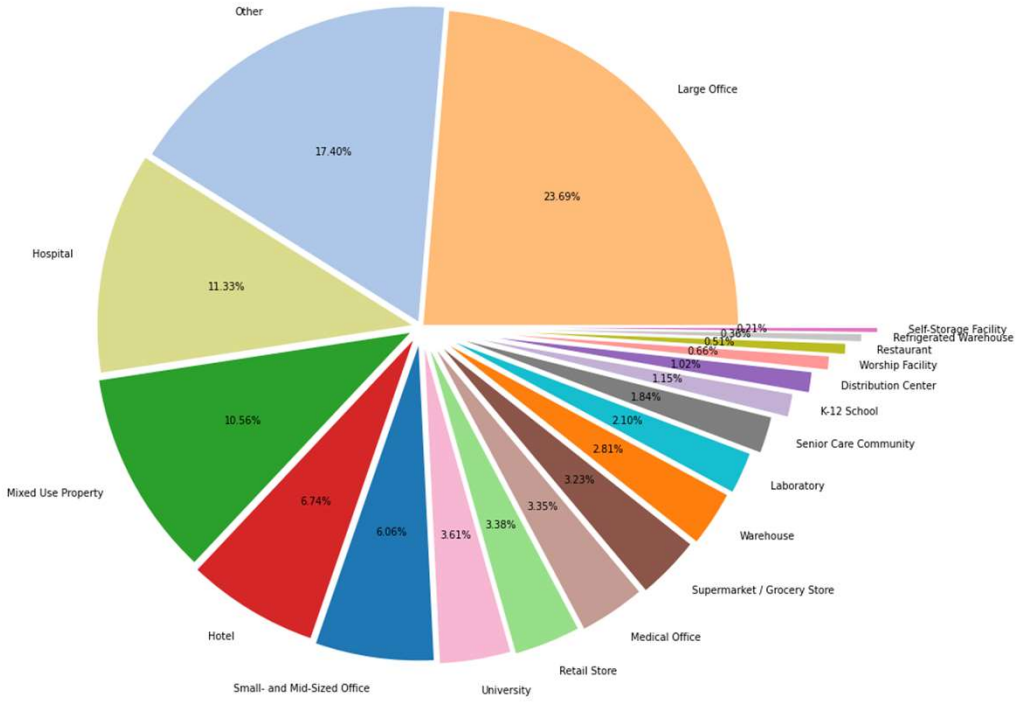
Le cœur de ville (hôtels, musées, salles spectacles, magasins)

Le site du campus

Individus par type de catégories des batiments non résidentiels



Répartition de la consommation des batiments non résidentiels



Traitement des données

Imputation des codes postaux avec un Knn

Sélection des bâtiments non résidentiels (1644)

Effacement des features inutiles (DataYear, City, State, DefaultData, Comments)

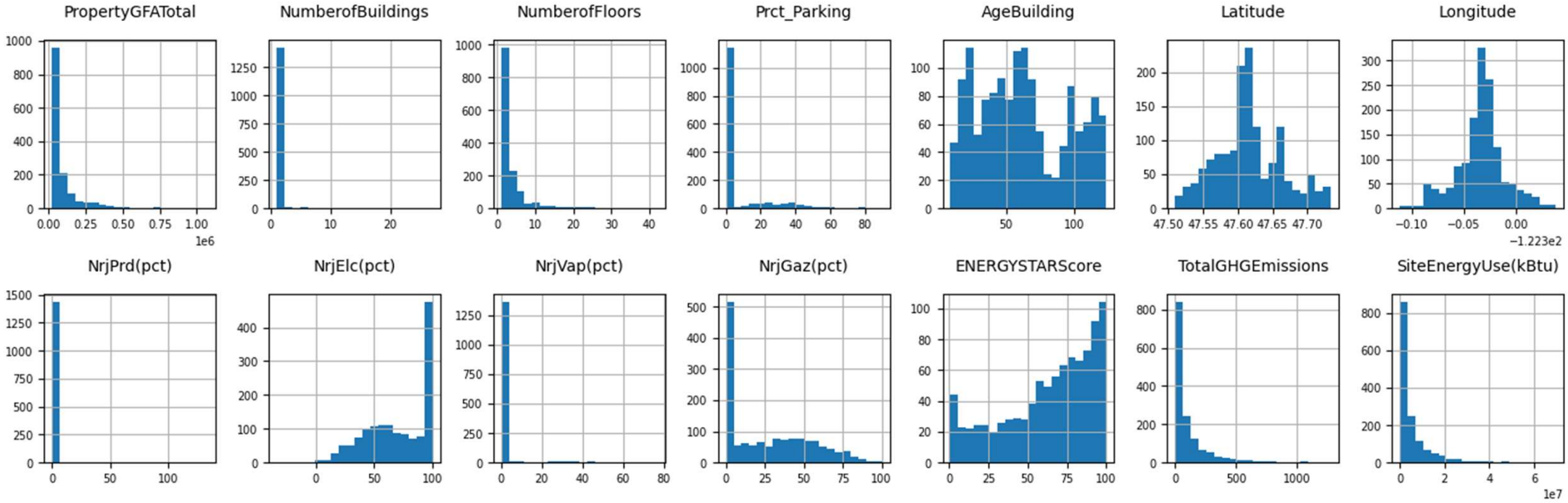
Filtrage des individus indiqués dans la feature «outliers» et «ComplianceStatus» (119)

Effacement des complexes de 0 étages et > 76 (13)

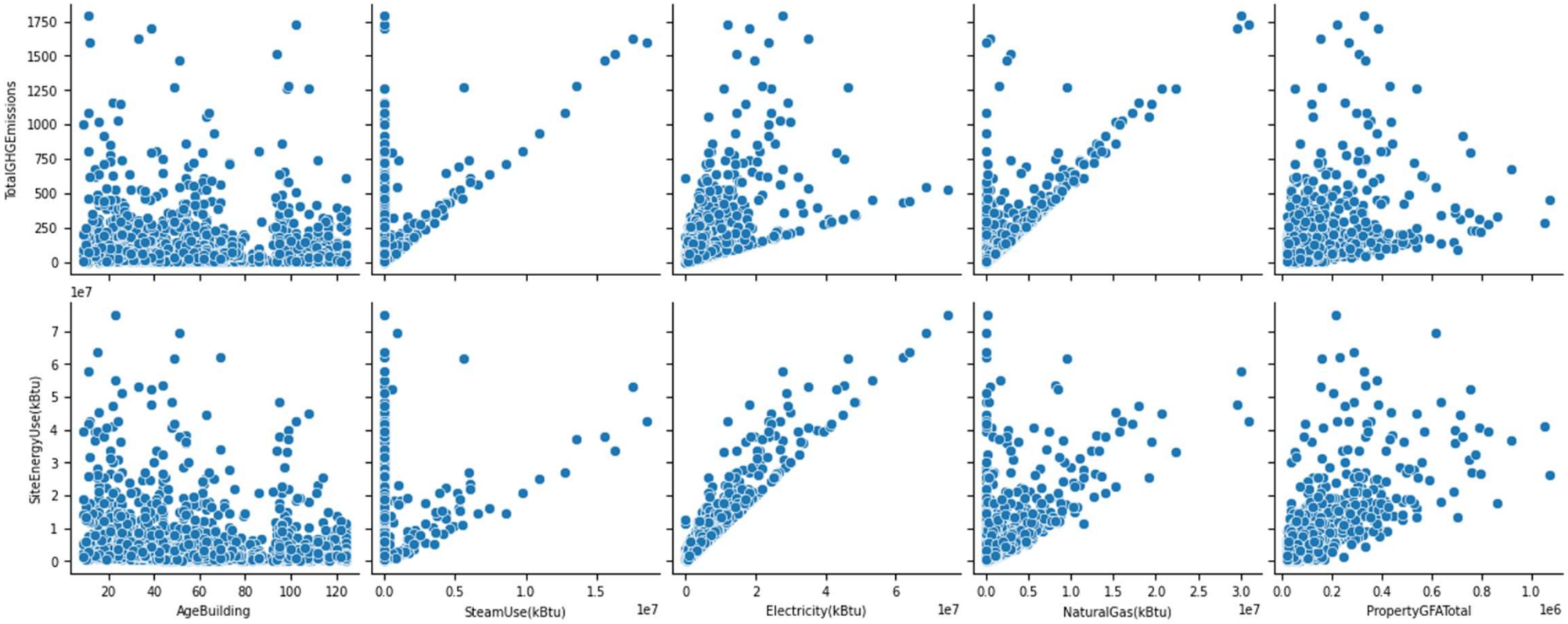
Modification du nombre de bâtiment à 1 quand 0 (50)

BuildingType	PrimaryPropertyType	
NonResidential	Small- and Mid-Sized Office	288
	Other	185
	Warehouse	180
	Large Office	168
	Mixed Use Property	102
SPS-District K-12	K-12 School	96
NonResidential	Retail Store	91
	Hotel	76
	Worship Facility	71
Nonresidential COS	Other	56
NonResidential	Distribution Center	51
	Supermarket / Grocery Store	40
	K-12 School	39
	Medical Office	38
	Self-Storage Facility	28
	Senior Care Community	20
	University	17
	Refrigerated Warehouse	12
	Restaurant	11
	Laboratory	10

Distribution des features retenues



Visualisation des variables cibles en fonction de certaines features

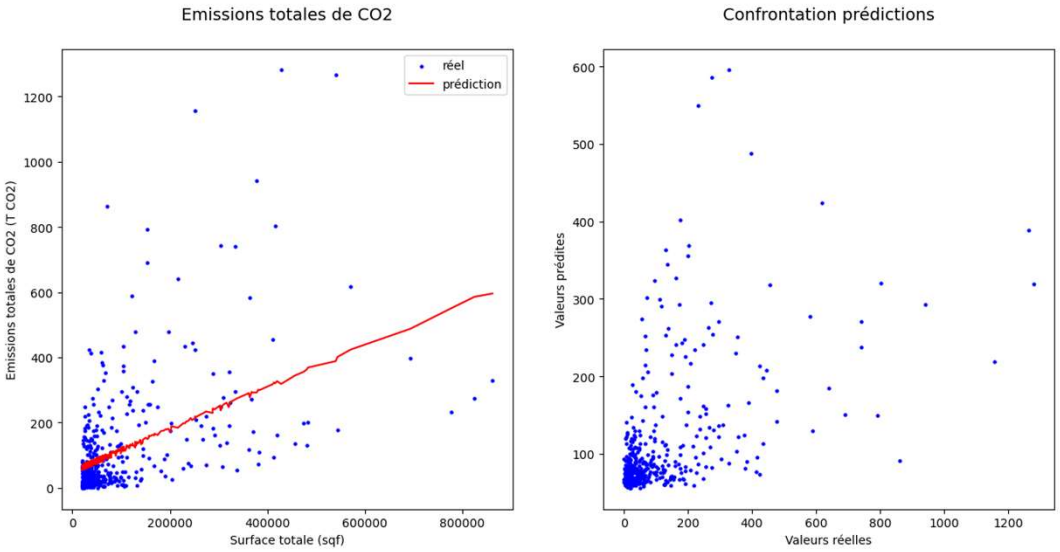


Création d'un modèle Baseline

Régression linéaire basique
Peu de features (Surface totale, Age du bâtiment)
Découpage données (entrainement/test) 70%/30%

Score train	: 0.23423
Score test	: 0.25477
R2	: 0.25477
Mse	: 21788.41917
Rmse	: 147.60901

Obtention d'une référence pour la modélisation finale



Préprocessing des données

Appliqué à la totalité des modèles

Homogénéité du traitement pour test et évolutions futures
avec utilisation de « Pipelines »

Découpage données (entrainement/test) 70%/30%

Features numériques (11) :

- imputation des valeurs manquantes par la moyenne
- standardisation des données (centrées en 0 et écart type de 1)

Features catégorielles (3) :

- imputation des valeurs manquantes par les plus fréquentes
- encodage One-Hot

Amélioration du feature engineering

Limitation des corrélations entre les features

Création

Production énergie (% énergie totale) (Delta entre énergie consommée et utilisée)

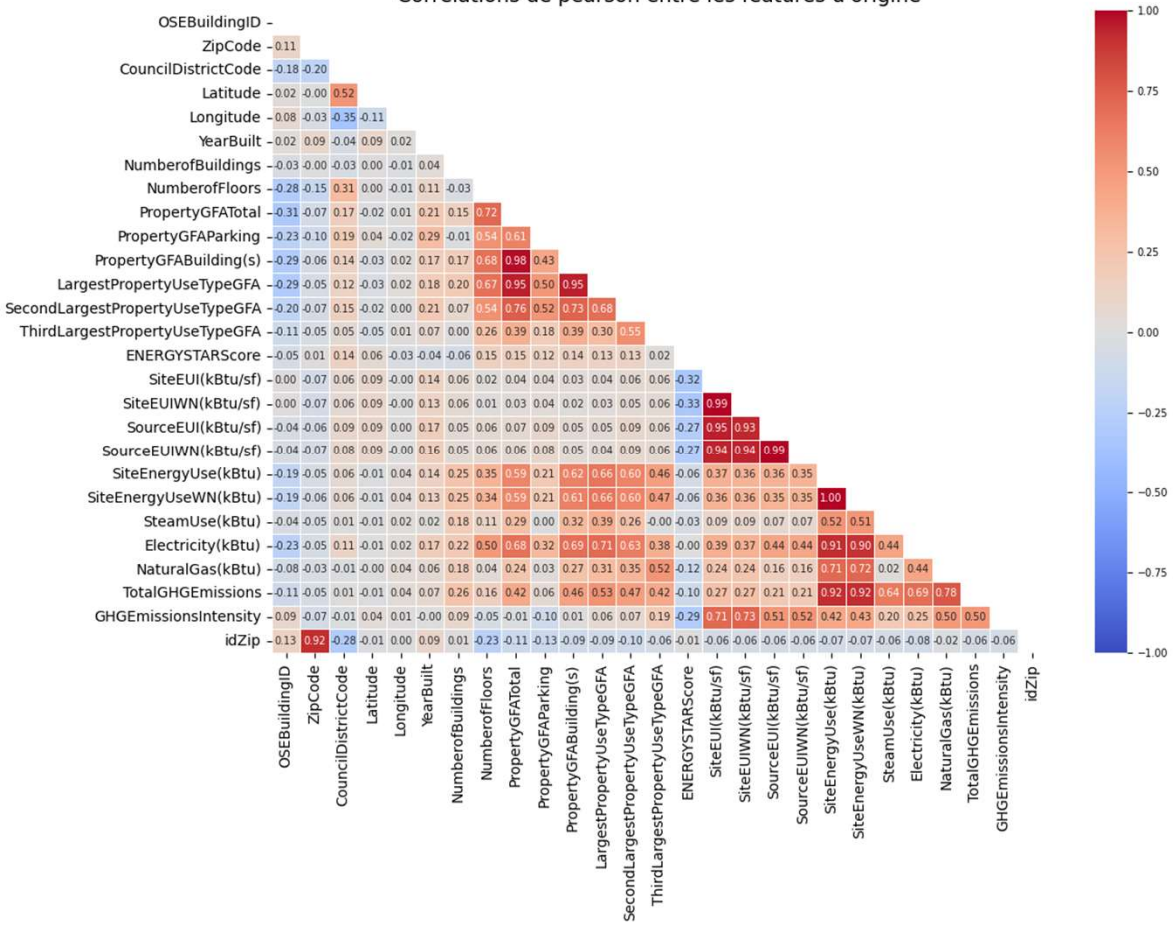
Modification

surface de parking	> parking en % de surface totale
dates de construction	> âge des bâtiments
mix énergétique (kBtu)	> mix énergétique (% énergie totale)

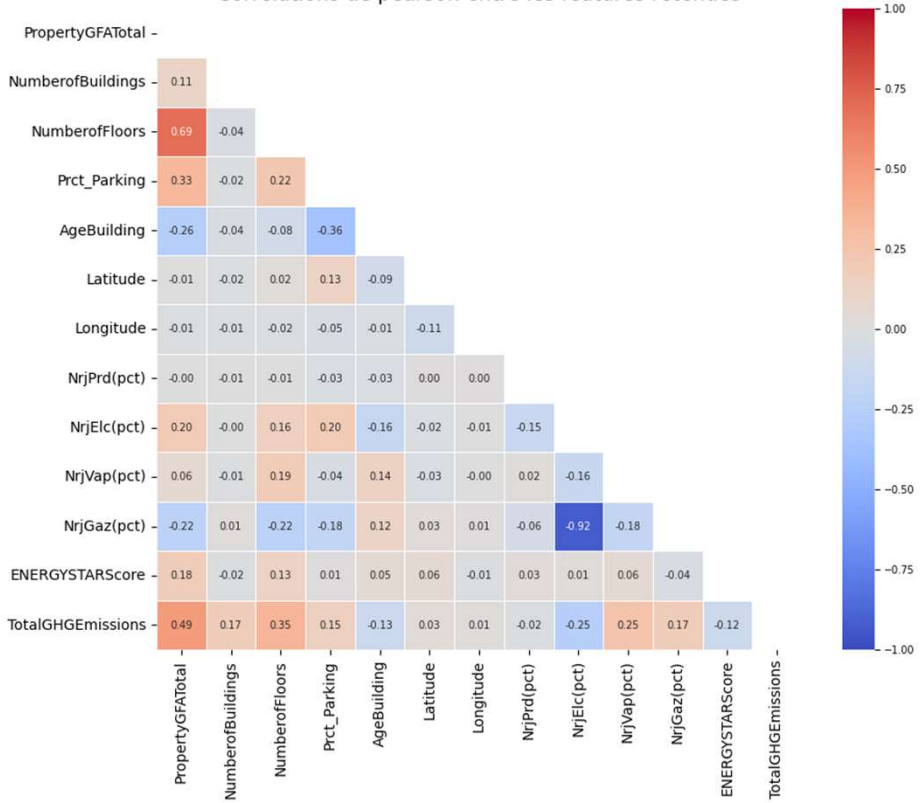
Suppression

du premier et dernier percentiles émissions CO2 (32)
des producteurs d'énergie négatives (3)
des consommations bâtiments > 400 kBtu/sf (9)

Corrélations de pearson entre les features d'origine



Corrélations de pearson entre les features retenues



Simulation de modèles et choix du modèle final

Modèles comparés :

Simples : LinearRegression
SVR
Lasso
Ridge
ElasticNet

Ensemblistes :
GradientBoostingRegressor
RandomForestRegressor

Performances comparées :
R^2, Mse, Rmse, Temps d'appentissage

Trois méthodes appliquées à chaque modèle :

Base (aucun ajustement de caractéristique)
Validations croisées pour ajustement Hyperparamètres :
GridSearchCV
RandomizedSearchCV

```
# définition des modèles testés et de leurs hyperparamètres
Modeles2Test=[[LinearRegression(),{
                'linearregression__fit_intercept': [True, False],
                'linearregression__copy_X': [True, False]}],
            [SVR(),{
                'svr__gamma' : ['auto','scale'],
                'svr__epsilon' : [0.001, 0.01, 0.1, 1],
                'svr__C' : [0.001, 0.01, 0.1, 1, 10],
                'svr__tol':[0.001]}],
            [Lasso(),{
                'lasso__alpha':[1,0.1,0.01,0.001],
                'lasso__max_iter':[1000],
                'lasso__random_state':[Id_Random],
                'lasso__tol':[0.001]}],
            [Ridge(),{
                'ridge__alpha':[1,0.1,0.01,0.001],
                'ridge__max_iter':[1000],
                'ridge__random_state':[Id_Random],
                'ridge__tol':[0.001]}],
```

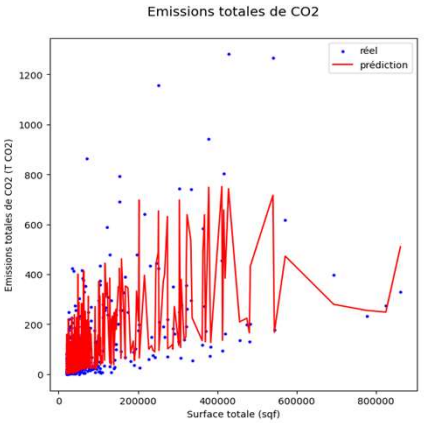
Prédiction émission CO2

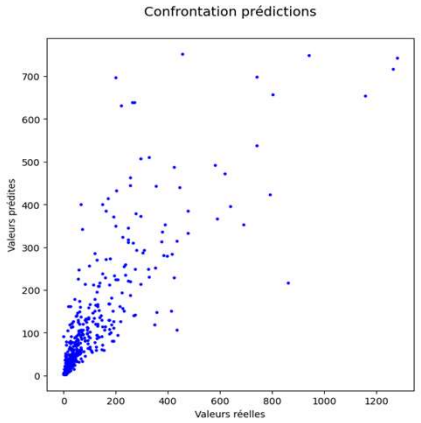
Meilleur modèle parmi les 7 testés : RandomForestRegressor

Pas de différences de temps entre méthodes
Grid ou Random scores très proche

Pas de convergence ou mauvais scores pour le SVR

Surapprentissage du GradientBoosting

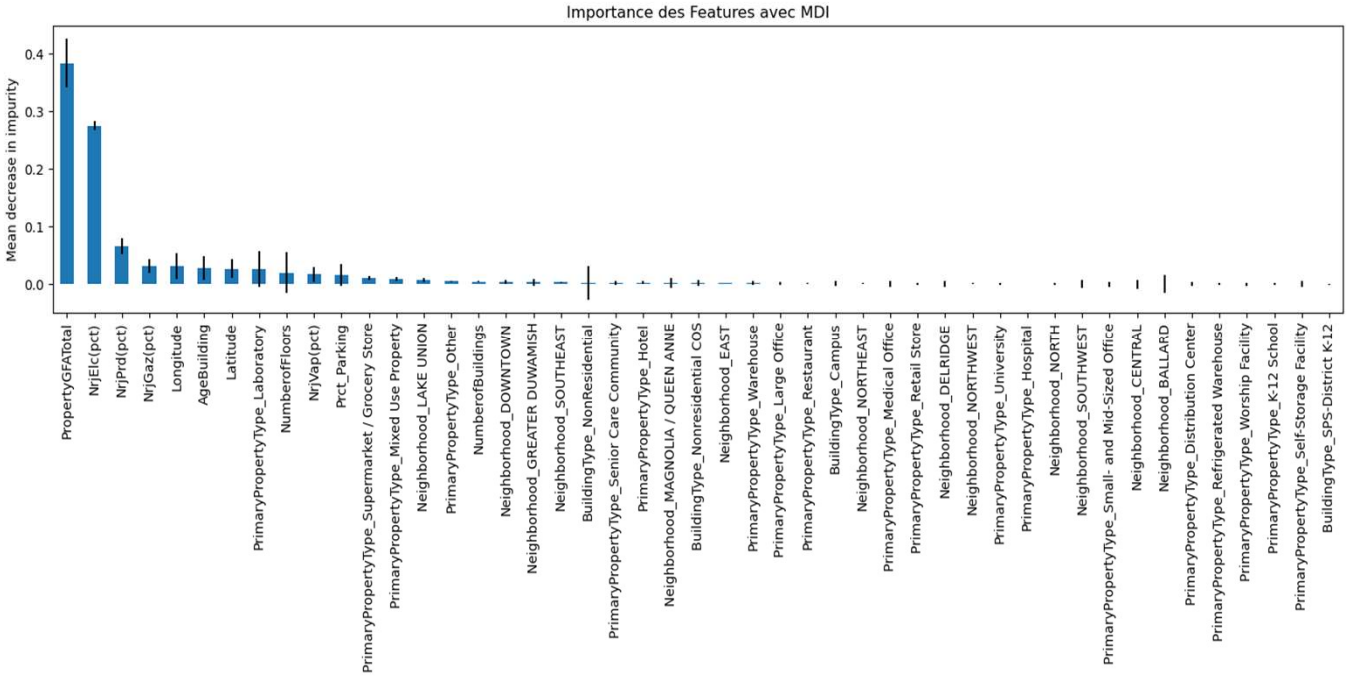


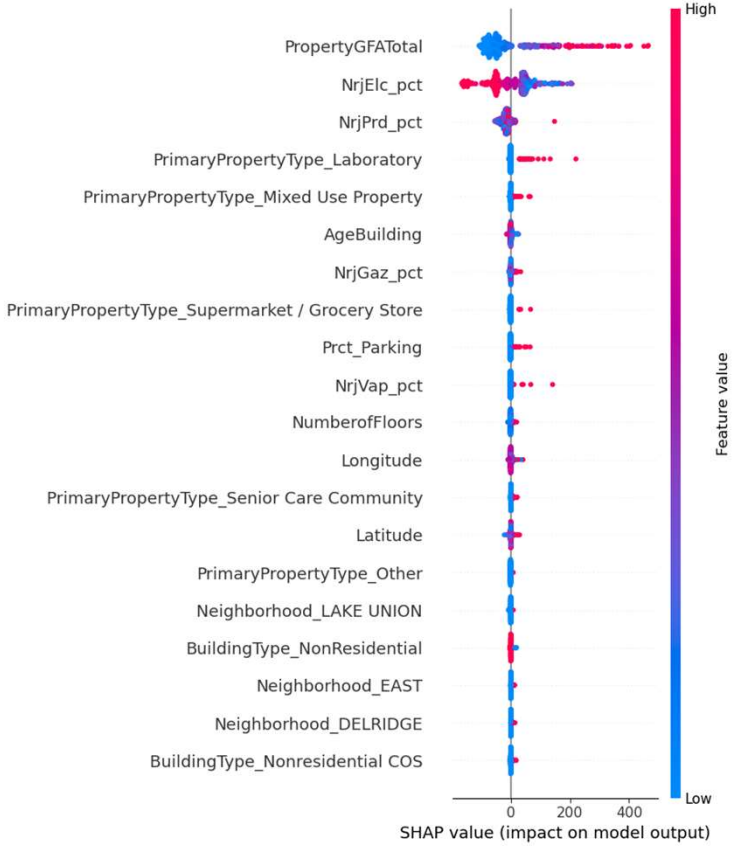


Modèle	BaseScoreTrain	GridScoreTrain	RandScoreTrain	BaseScoreTest	GridScoreTest	RandScoreTest	BaseMse	GridMse	RandMse	BaseRmse	GridRmse	RandRmse	BaseTps	GridTps	RandTps
LinearRegression	0.55867	0.55867	0.55864	0.47098	0.47098	0.47066	15466.91054	15466.91054	15476.32731	124.36603	124.36603	124.40389	0.013	1.334	0.268
SVR	-0.00439	0.24567	-0.00439	-0.01306	0.23172	-0.01306	29618.88361	22462.27161	29618.88361	172.10138	149.87419	172.10138	0.051	2.409	0.790
Lasso	0.52998	0.52995	0.52995	0.47686	0.47675	0.47675	15295.14572	15298.28332	15298.28332	123.67354	123.68623	123.68623	0.010	0.137	0.462
Ridge	0.55694	0.55694	0.55861	0.47705	0.47705	0.47155	15289.60338	15289.60338	15450.41876	123.65114	123.65114	124.29971	0.019	0.100	0.294
ElasticNet	0.40013	0.54451	0.54451	0.37800	0.48360	0.48360	18185.48736	15098.01732	15098.01732	134.85358	122.87399	122.87399	0.015	0.161	0.417
GradientBoostingRegressor	0.87373	1.00000	1.00000	0.68168	0.38723	0.40978	9306.87096	17915.67071	17256.33441	96.47213	133.84943	131.36337	0.237	8.515	25.312
RandomForestRegressor	0.94888	0.94938	0.94310	0.65118	0.66155	0.66195	10198.43181	9895.22942	9883.66920	100.98729	99.47477	99.41664	0.677	16.999	15.079

Analyse feature importance émission CO2

Les features « superficies totales »,
Utilisation et production d'électricité
impactent le plus fortement les résultats





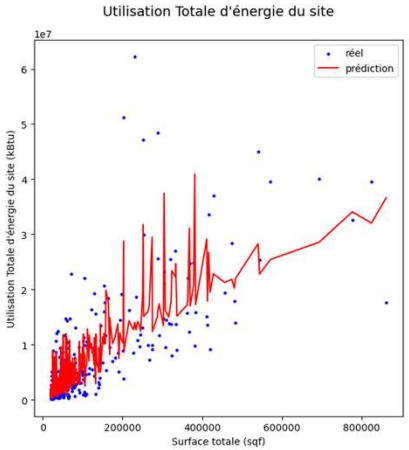
Prédiction consommation énergie

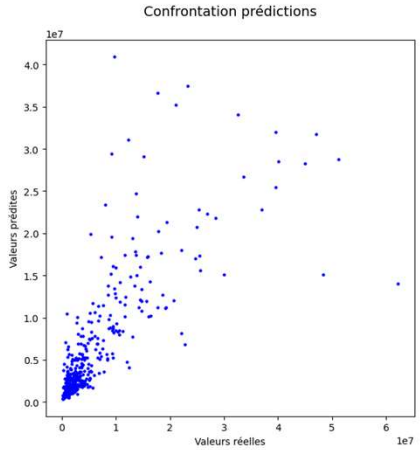
Meilleur modèle parmi les 7 testés : RandomForestRegressor

Différences significatives de temps entre méthodes
Grid ou Random scores très proche

Pas de convergence ou mauvais scores pour le SVR

Surapprentissage du GradientBoosting

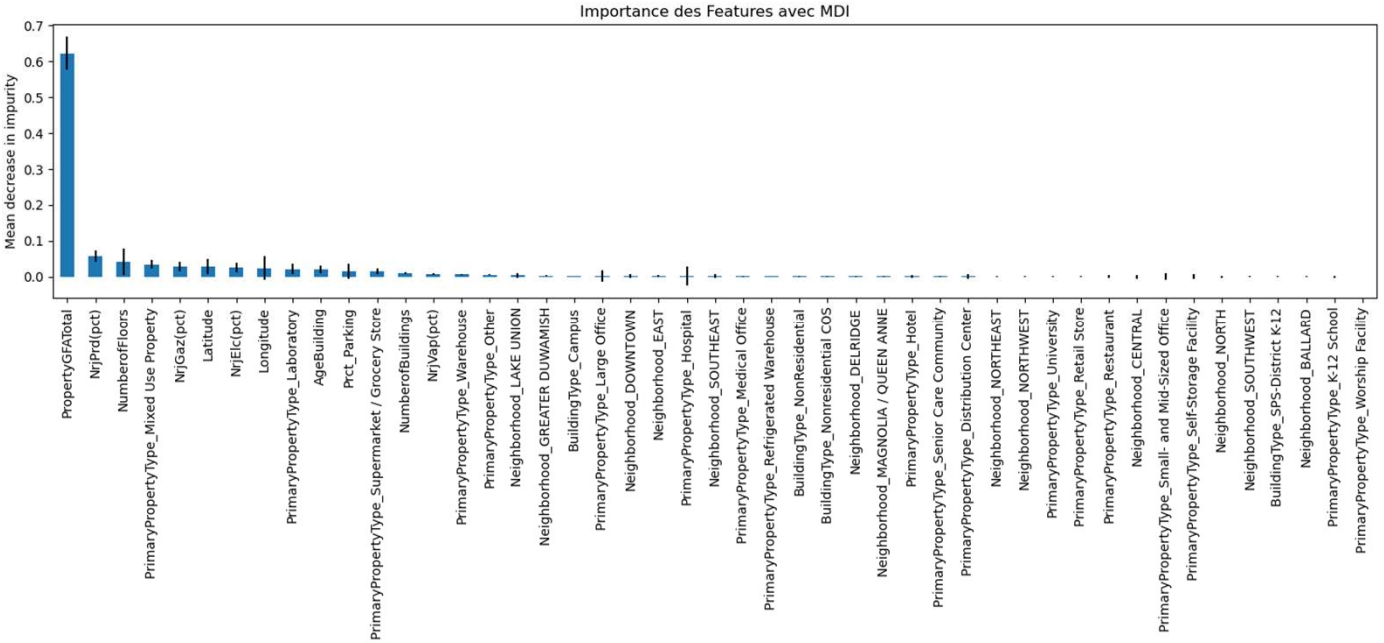


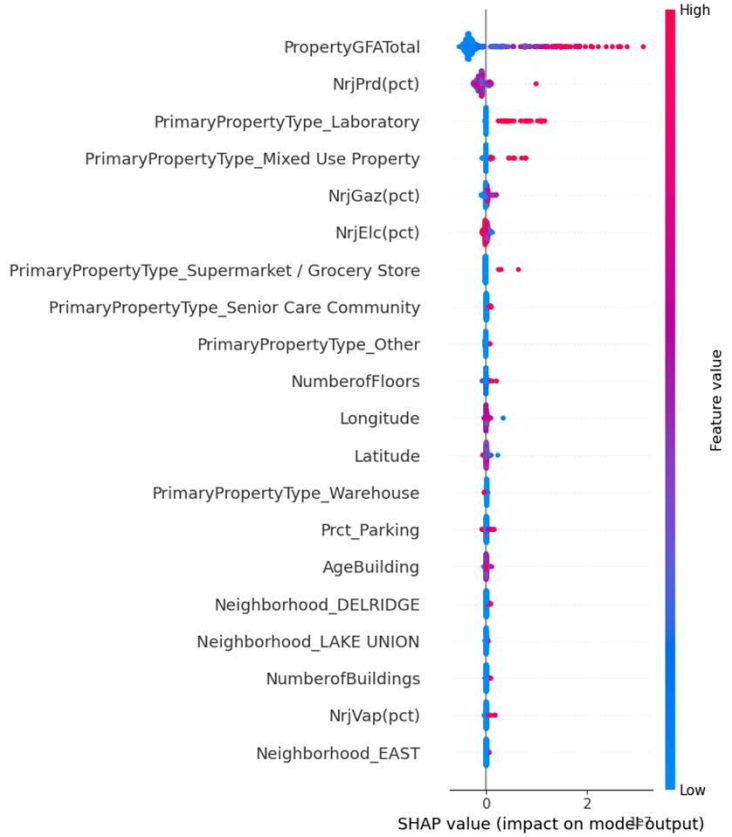


Modèle	BaseScoreTrain	GridScoreTrain	RandScoreTrain	BaseScoreTest	GridScoreTest	RandScoreTest	BaseMse	GridMse	RandMse	BaseRmse	GridRmse	RandRmse	BaseTemps	GridTemps	RandTemps
LinearRegression	0.64961	0.64961	0.64961	0.58733	0.58733	0.58733	2.841071e+13	2.841059e+13	2.841059e+13	5.330170e+06	5.330159e+06	5.330159e+06	0.020	1.311	0.298
SVR	-0.14580	-0.14574	-0.14575	-0.14608	-0.14603	-0.14604	7.890374e+13	7.889980e+13	7.890066e+13	8.882778e+06	8.882556e+06	8.882604e+06	0.042	2.209	0.716
Lasso	0.64961	0.64961	0.64961	0.58733	0.58733	0.58733	2.841068e+13	2.841068e+13	2.841068e+13	5.330167e+06	5.330167e+06	5.330167e+06	0.174	0.795	2.732
Ridge	0.64867	0.64861	0.64861	0.59110	0.59134	0.59134	2.815144e+13	2.813509e+13	2.813509e+13	5.305793e+06	5.304252e+06	5.304252e+06	0.015	0.135	0.352
ElasticNet	0.54662	0.64193	0.64193	0.50556	0.59524	0.59524	3.404044e+13	2.786603e+13	2.786603e+13	5.834418e+06	5.278828e+06	5.278828e+06	0.016	1.145	2.290
GradientBoostingRegressor	0.88101	1.00000	1.00000	0.62814	0.28976	0.29503	2.560141e+13	4.889729e+13	4.853460e+13	5.059783e+06	6.992659e+06	6.966678e+06	0.330	32.057	83.389
RandomForestRegressor	0.95021	0.95266	0.95152	0.62458	0.61113	0.60786	2.584657e+13	2.677235e+13	2.699712e+13	5.083953e+06	5.174200e+06	5.195876e+06	1.504	28.890	15.706

Analyse feature importance consommation énergie

La feature « superficie totale »
impacte le plus fortement les résultats

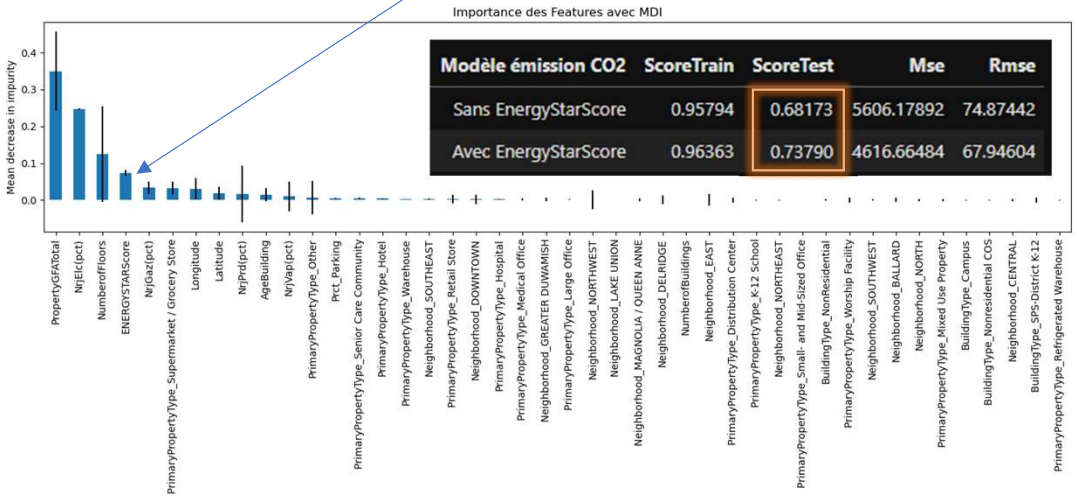




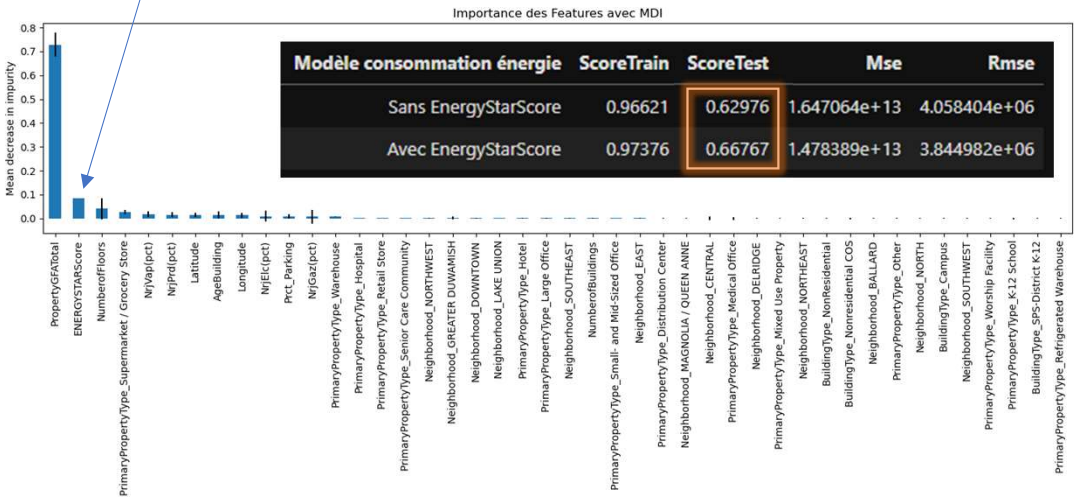
Influence de l'EnergieStarScore sur les modèles

Echantillon modifié à 929 individus présenté aux 2 modèles avec et sans EnergieStarScore
Observation d'une amélioration dans les deux cas

Quatrième variable importante sur la prédiction de CO2
Deuxième sur la prédiction de consommation d'énergie



Modèle émission CO2	ScoreTrain	ScoreTest	Mse	Rmse
Sans EnergyStarScore	0.95794	0.68173	5606.17892	74.87442
Avec EnergyStarScore	0.96363	0.73790	4616.66484	67.94604



Modèle consommation énergie	ScoreTrain	ScoreTest	Mse	Rmse
Sans EnergyStarScore	0.96621	0.62976	1.647064e+13	4.058404e+06
Avec EnergyStarScore	0.97376	0.66767	1.478389e+13	3.844982e+06

Conclusion

Traitement de la problématique avec une régression

Tests et optimisation de 7 algorithmes avec validation croisée

Modèles retenus RandomForest

Gain important de l'EnergyStarScore sur la prédiction des données

Optimisation nécessaire pour amélioration des prédictions avant mise en production

Merci pour votre attention