Projet 4 : Anticipez les besoins en consommation de bâtiments

Parcours Data Scientist – OpenClassrooms

Soutenance

Sommaire

I – PROBLÉMATIQUE
ET JEU DE DONNÉES

II – FEATURE
ENGINEERING

III – MODÉLISATION
ET RÉSULTATS

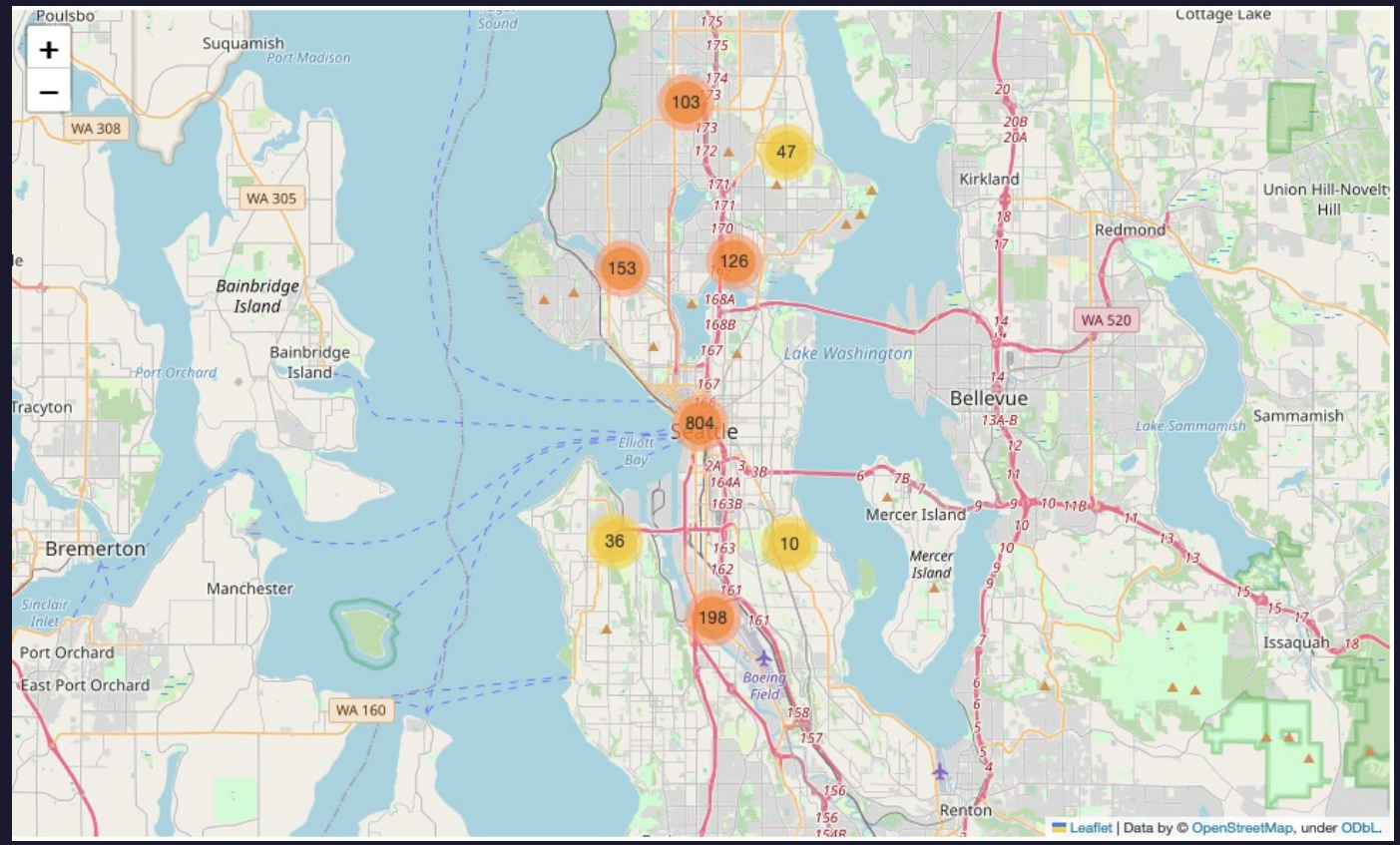
I – Problématique et jeu de données

Problématique

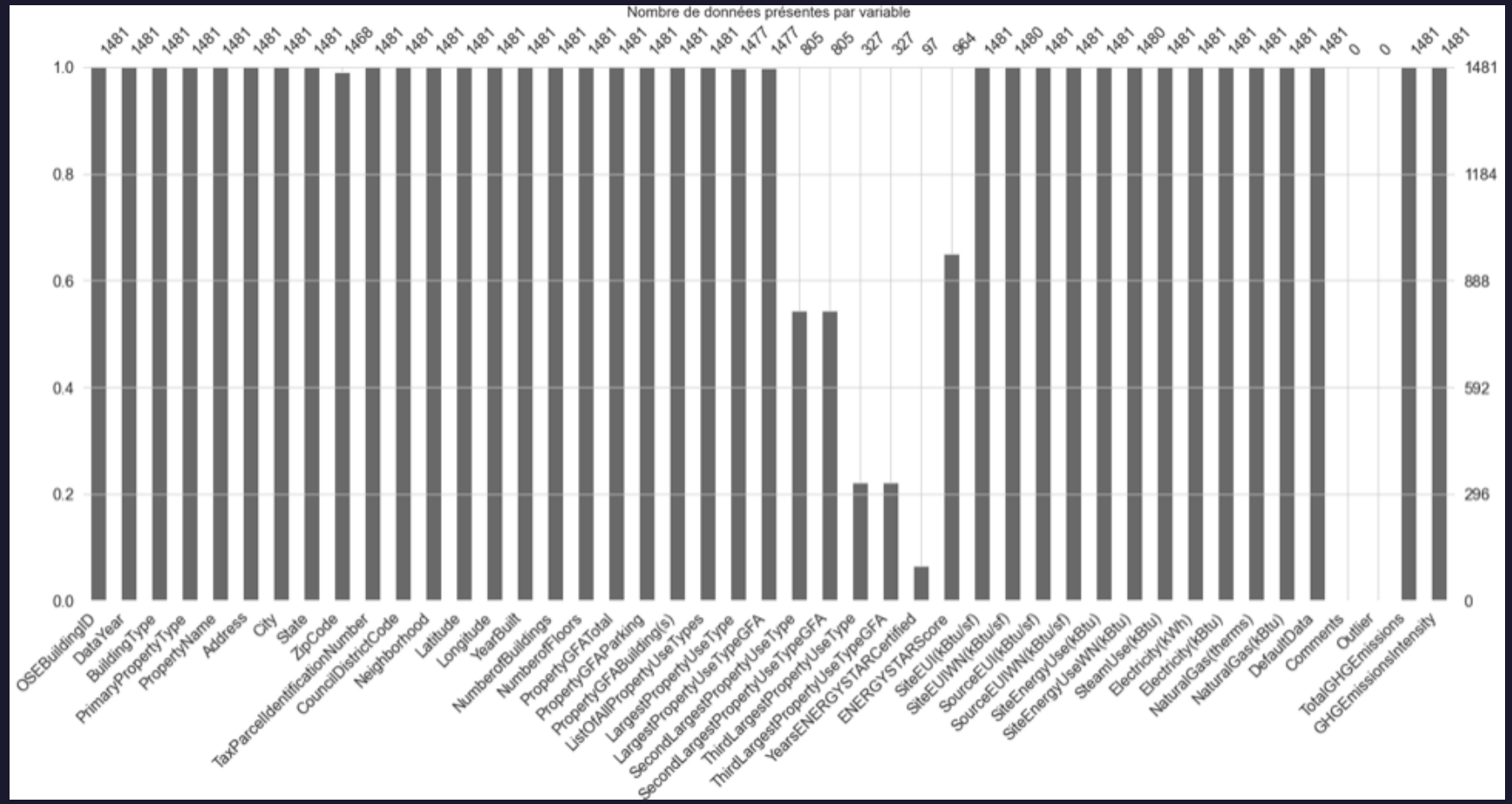
- Objectif neutre en émissions de carbone en 2050
- Consommation et aux émissions des bâtiments non destinés à l'habitation
- Évaluer l'intérêt de l'ENERGY STAR Score
- Relevés en 2016

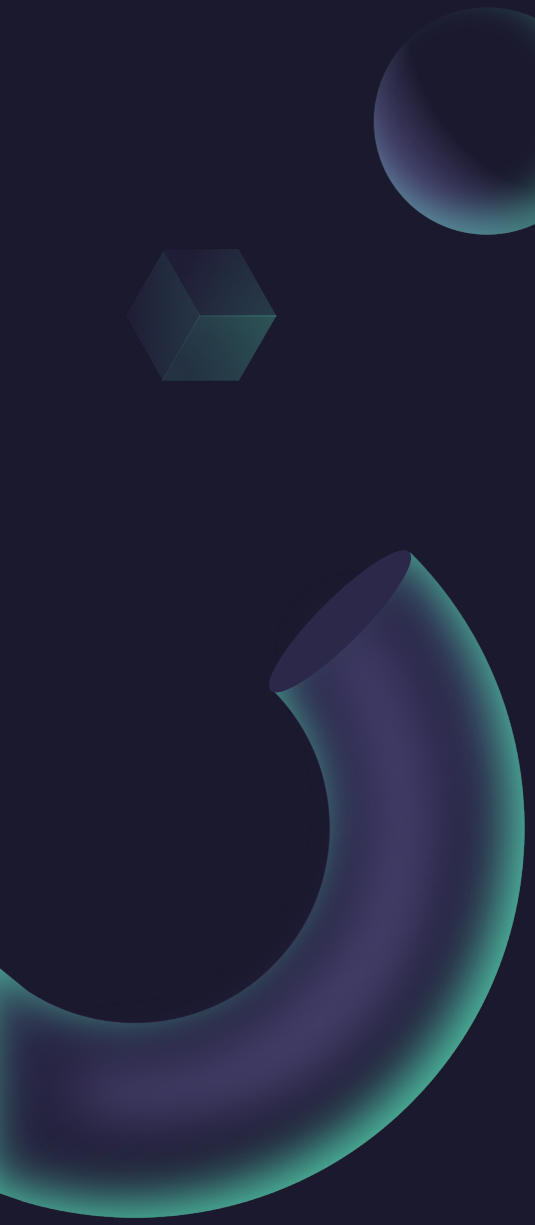
Jeu de données

0	OSEBuildingID	3376 non-null	int64
1	DataYear	3376 non-null	int64
2	BuildingType	3376 non-null	object
3	PrimaryPropertyType	3376 non-null	object
4	PropertyName	3376 non-null	object
5	Address	3376 non-null	object
6	City	3376 non-null	object
7	State	3376 non-null	object
8	ZipCode	3360 non-null	float64
9	TaxParcelIdentificationNumber	3376 non-null	object
10	CouncilDistrictCode	3376 non-null	int64
11	Neighborhood	3376 non-null	object
12	Latitude	3376 non-null	float64
13	Longitude	3376 non-null	float64
14	YearBuilt	3376 non-null	int64
15	NumberofBuildings	3368 non-null	float64
16	NumberofFloors	3376 non-null	int64
17	PropertyGFATotal	3376 non-null	int64
18	PropertyGFAParking	3376 non-null	int64
19	PropertyGFABuilding(s)	3376 non-null	int64
20	ListOfAllPropertyUseTypes	3367 non-null	object
21	LargestPropertyUseType	3356 non-null	object
22	LargestPropertyUseTypeGFA	3356 non-null	float64
23	SecondLargestPropertyUseType	1679 non-null	object
24	SecondLargestPropertyUseTypeGFA	1679 non-null	float64
25	ThirdLargestPropertyUseType	596 non-null	object
26	ThirdLargestPropertyUseTypeGFA	596 non-null	float64
27	YearsENERGYSTARCertified	119 non-null	object
28	ENERGYSTARScore	2533 non-null	float64
29	SiteEUI(kBtu/sf)	3369 non-null	float64
30	SiteEUIWN(kBtu/sf)	3370 non-null	float64
31	SourceEUI(kBtu/sf)	3367 non-null	float64
32	SourceEUIWN(kBtu/sf)	3367 non-null	float64
33	SiteEnergyUse(kBtu)	3371 non-null	float64
34	SiteEnergyUseWN(kBtu)	3370 non-null	float64
35	SteamUse(kBtu)	3367 non-null	float64
36	Electricity(kWh)	3367 non-null	float64
37	Electricity(kBtu)	3367 non-null	float64
38	NaturalGas(therms)	3367 non-null	float64
39	NaturalGas(kBtu)	3367 non-null	float64
40	DefaultData	3376 non-null	bool
41	Comments	0 non-null	float64
42	ComplianceStatus	3376 non-null	object
43	Outlier	32 non-null	object
44	TotalGHGEmissions	3367 non-null	float64
45	GHGEmissionsIntensity	3367 non-null	float64



Jeu de données





II – Feature Engineering

Feature Engineering

- *BuildingAge* (int): détermine l'âge du bâtiment en fonction de son année de construction et de l'année des données

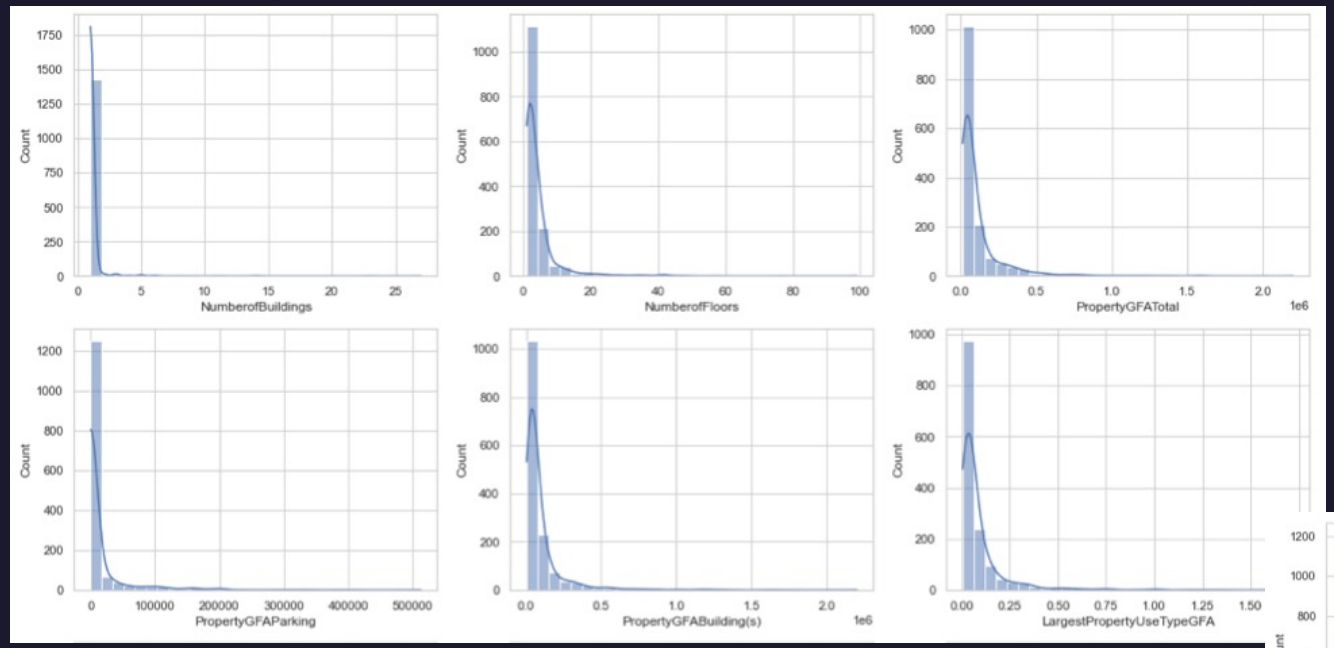
- *ElectricityUse* (bool): détermine si le bâtiment consomme oui ou non de l'électricité
- *SteamUse* (bool): détermine si le bâtiment consomme oui ou non de la vapeur
- *NaturalGasUse* (bool): détermine si le bâtiment consomme oui ou non du gaz naturel

- *ElectricityUseMost* (bool) : détermine si l'électricité est l'énergie la plus utilisée par le bâtiment
- *SteamUseMost* (bool) : détermine si la vapeur est l'énergie la plus utilisée par le bâtiment
- *NaturalGasUseMost* (bool) : détermine si le gaz naturel est l'énergie la plus utilisée par le bâtiment

- *ParkingRatioGFA* (float) : ratio de la surface de parking en fonction de la surface totale

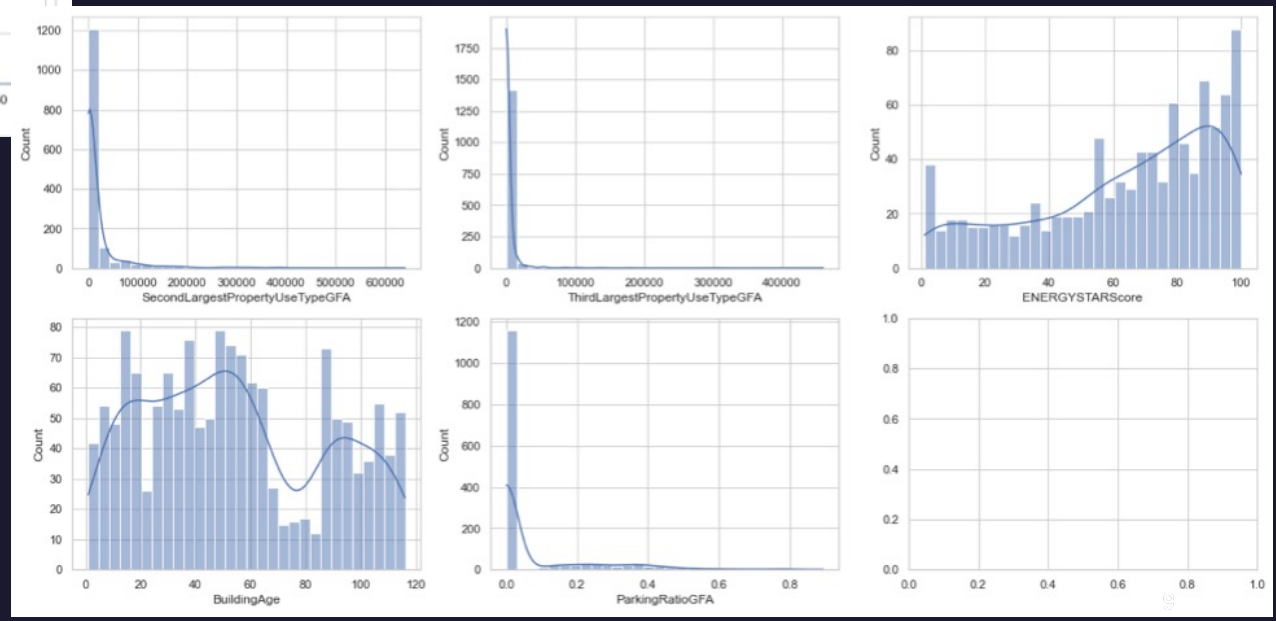
- *MoreThan1Building* (bool) : détermine si il y a plus d'1 bâtiment ou non

Analyse Exploratoire



Tests de normalité pour la variable TotalGHGEmissions.

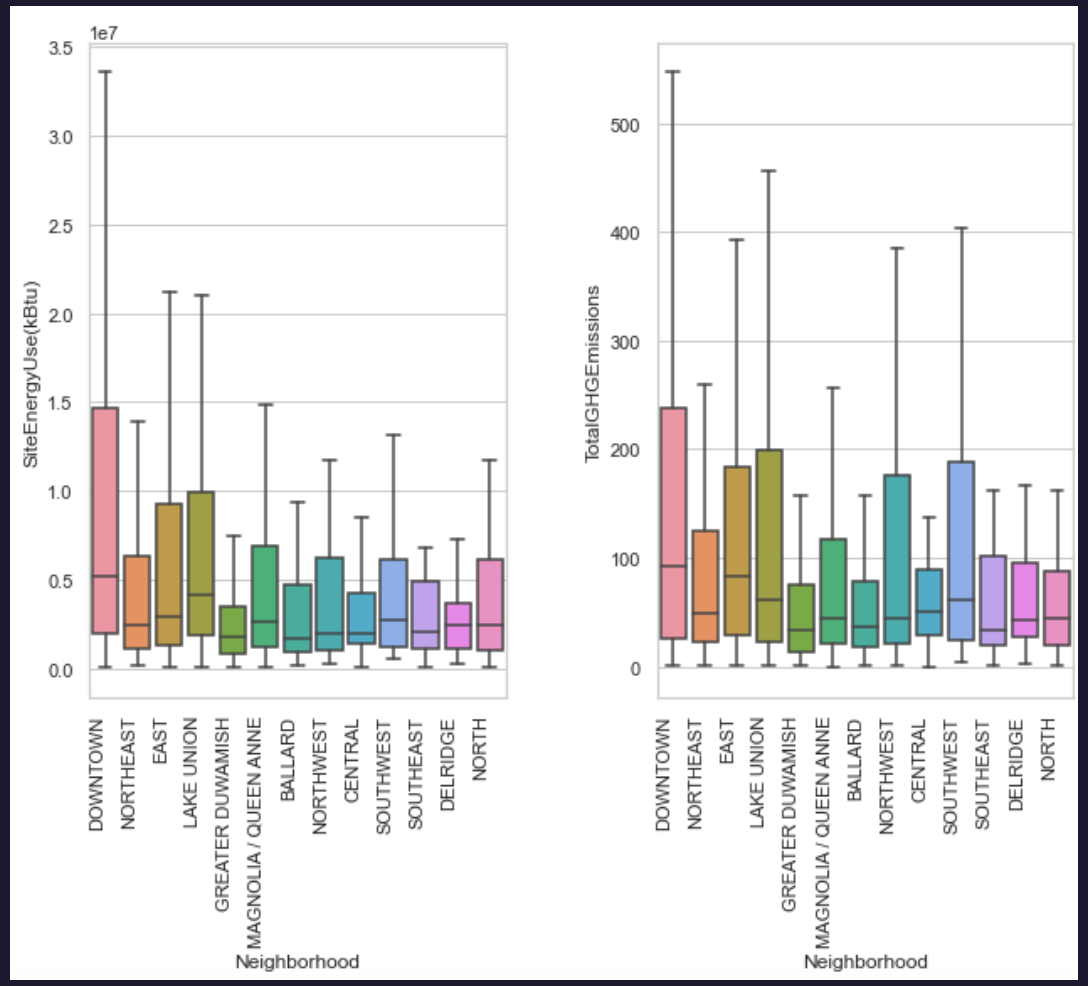
	Stat	p-value	Resultat
Shapiro Wilk	0.17853	0.0	H1
Anderson-Darling	366.44347	NaN	H1
K2 de D'Agostino	2854.804645	0.0	H1
Kolmogorov-Smirnov	0.969876	0.0	H1

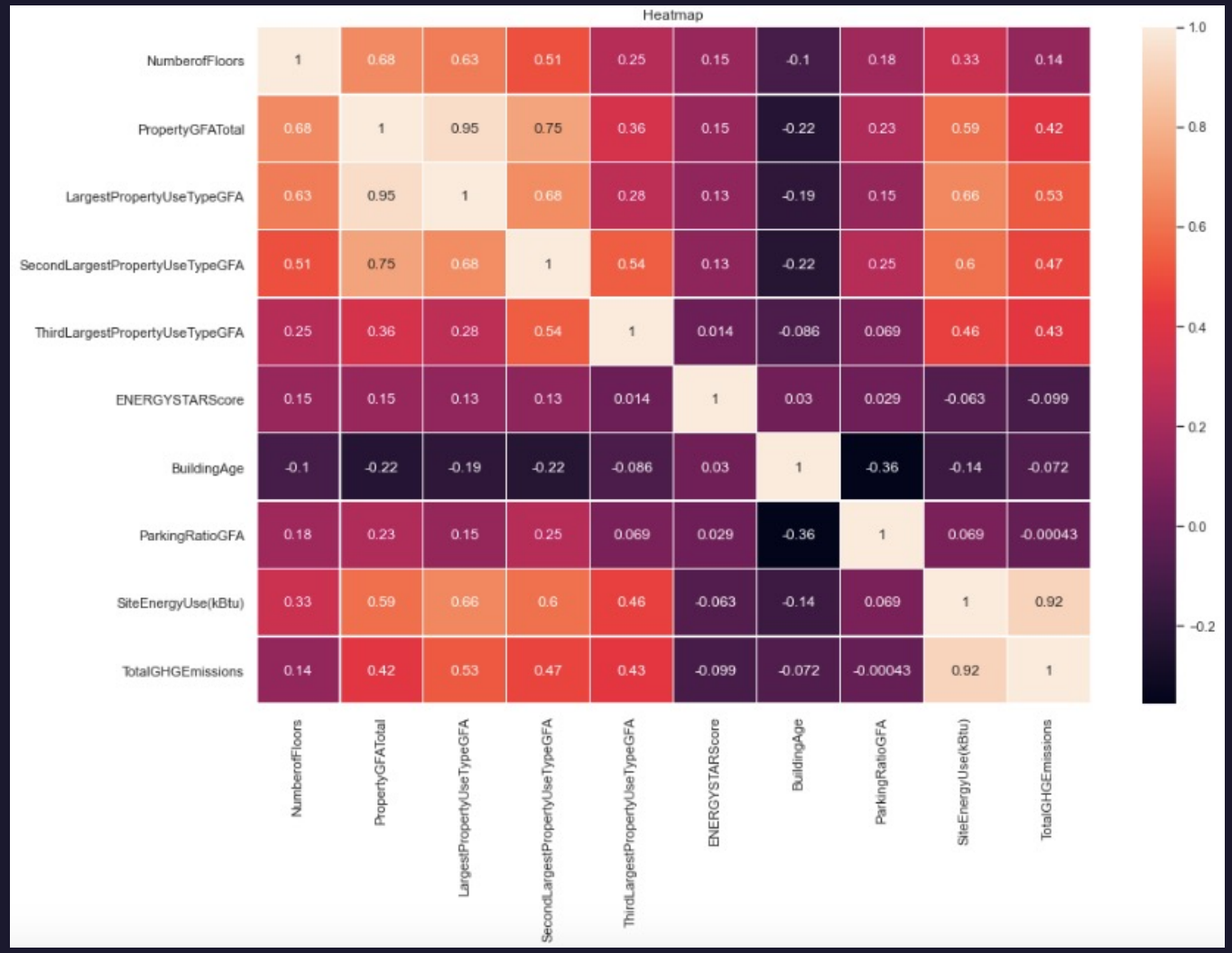


Tests de normalité pour la variable SiteEnergyUse(kBtu).

	Stat	p-value	Resultat
Shapiro Wilk	0.289505	0.0	H1
Anderson-Darling	293.954419	NaN	H1
K2 de D'Agostino	2450.681246	0.0	H1
Kolmogorov-Smirnov	1.0	0.0	H1

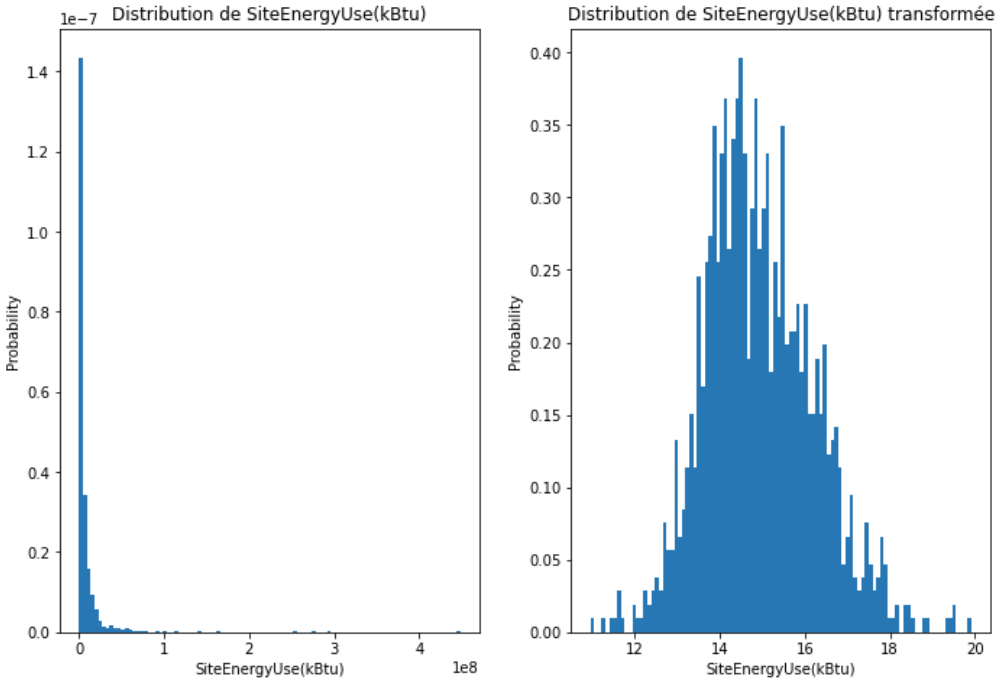
Analyse Exploratoire





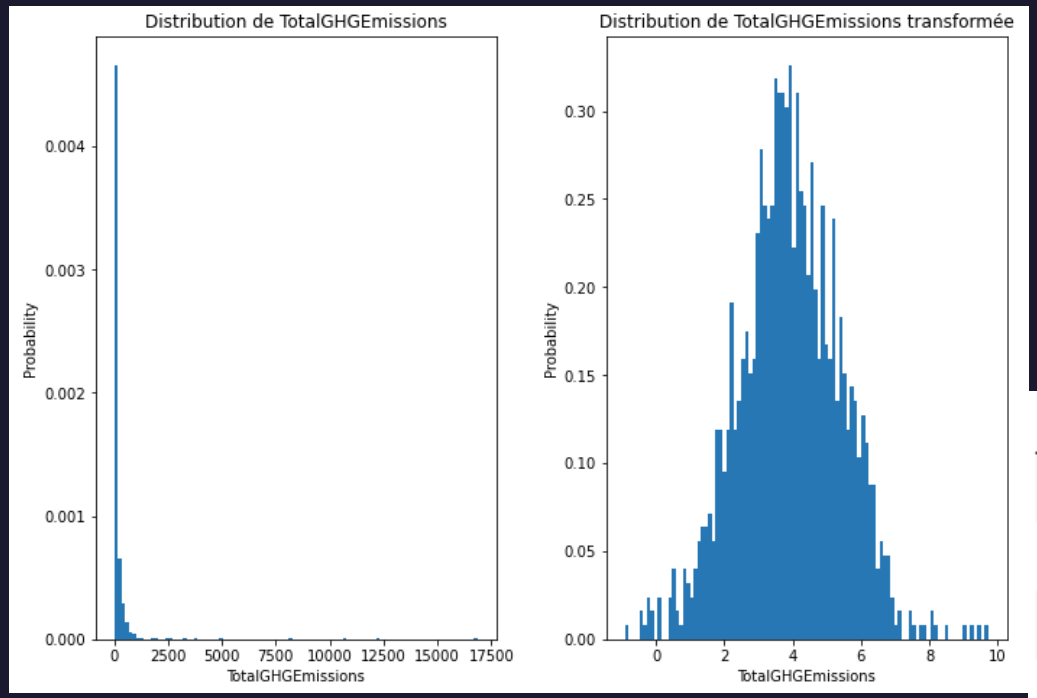
III – MODÉLISATION ET RÉSULTATS

Preprocessing et Modèles de base SiteEnergyUse(kBtu)



	RMSE	R2	MAE	MedAE	FIT_TIME	SCORE_TIME
dum	1.287039e+00	-2.470693e-03	1.026176e+00	0.875725	0.020022	0.009478
lr	1.665601e+10	-2.563512e+20	2.319685e+09	0.574233	0.017959	0.007368
ridge	8.286052e-01	5.822809e-01	6.496803e-01	0.553595	0.012779	0.007278
lasso	1.287039e+00	-2.470693e-03	1.026176e+00	0.875725	0.012695	0.007392
ElasticNet	1.175119e+00	1.645134e-01	9.500257e-01	0.821818	0.013293	0.008287
RandomForest	6.816806e-01	7.173551e-01	5.131139e-01	0.386688	0.732019	0.019686
KNR	8.441212e-01	5.691695e-01	6.479010e-01	0.517739	0.015617	0.016878
SVR	7.095940e-01	6.947184e-01	5.412912e-01	0.435238	0.117691	0.061906
XGBR	6.780001e-01	7.208610e-01	5.112961e-01	0.408970	0.594588	0.013957

Preprocessing et Modèles de base TotalGHGEmissions



	RMSE	R2	MAE	MedAE	FIT_TIME	SCORE_TIME
dum	1.491073e+00	-1.518917e-02	1.169503e+00	0.954433	0.019423	0.012079
lr	3.398999e+10	-1.342673e+21	5.184132e+09	0.634092	0.018480	0.007810
ridge	9.022880e-01	6.274067e-01	7.139222e-01	0.613338	0.012910	0.007394
lasso	1.491073e+00	-1.518917e-02	1.169503e+00	0.954433	0.012705	0.007624
ElasticNet	1.424688e+00	7.296317e-02	1.119958e+00	0.923315	0.013088	0.007850
RandomForest	7.932584e-01	7.114382e-01	6.007466e-01	0.459774	0.610457	0.018981
KNR	9.328690e-01	6.011874e-01	7.279353e-01	0.594728	0.016503	0.018248
SVR	8.033684e-01	7.043327e-01	6.169765e-01	0.516830	0.103894	0.047677
XGBR	7.675397e-01	7.299258e-01	5.910919e-01	0.469872	0.429936	0.012721

Optimisation des modèles

Séparation du jeu de données (train/test)

Définition de la grille d'hyperparamètres

Preprocessing :
- Standard Scaler (variables quantitatives)
- One Hot Encoder (variables qualitatives)

Grid Search CV :
- Teste toutes les combinaisons d'hyperparamètres
- Cross-validation
- Training set
- Trouve les hyperparamètres optimaux

Pipeline

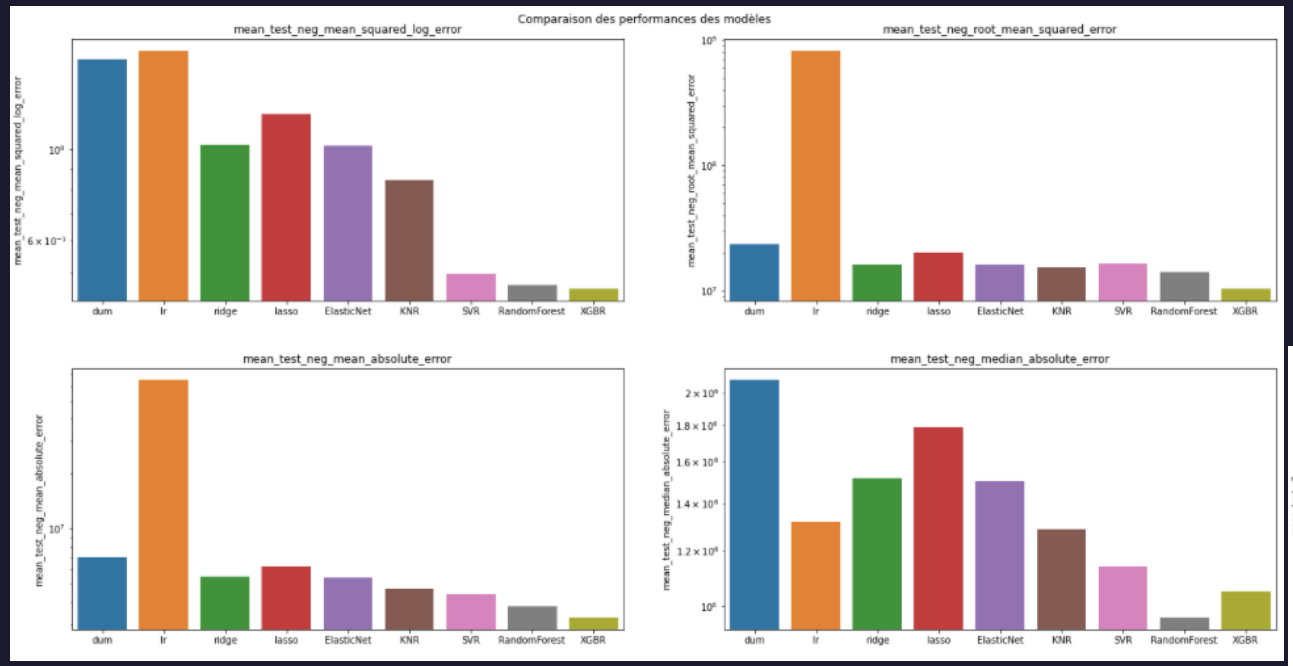
Résultat de l'optimisation pour la prédiction de la consommation d'énergie

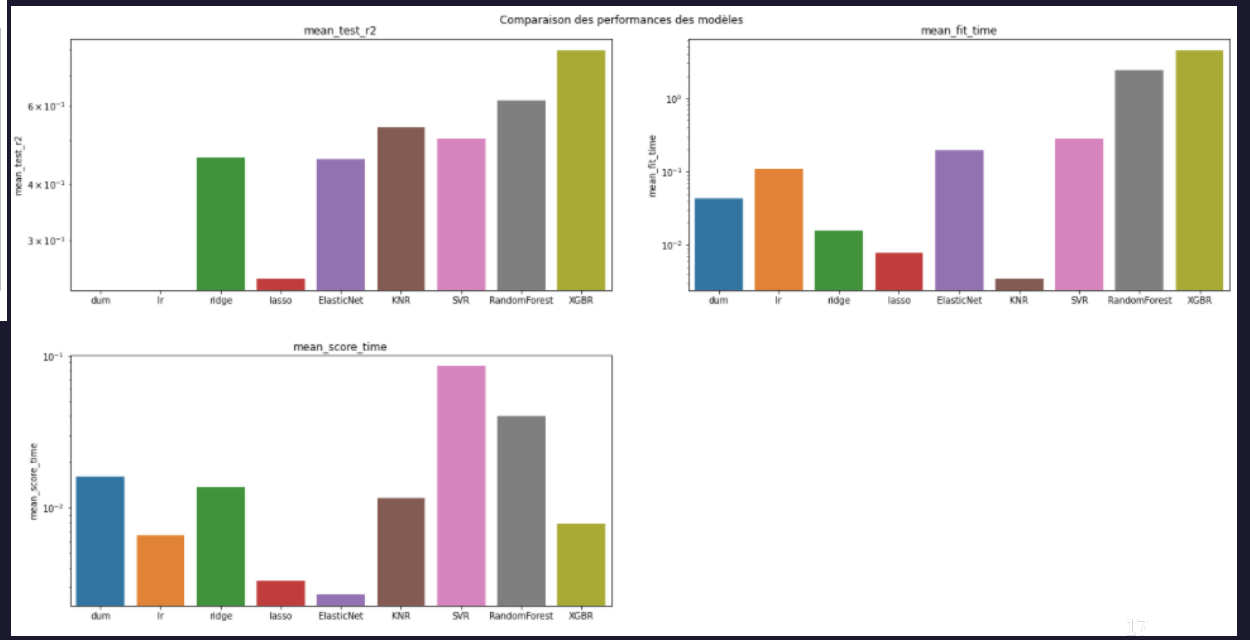
Modèle	Paramètres				
Dummy moyenne	strategy: mean				
Régression linéaire	fit_intercept : False	Positive: True			
Régression Ridge	Alpha : 1100				
Regression Lasso	Alpha : 0.416				
Elastic Net	Alpha : 1.099	l1_ratio : 0.0			
KNR	n_neighbors : 2	Weights : distance	leaf_size : 0	p : 2	
SVR	C : 10	Degree : 1	Epsilon : 0.001	Gamma : 0.1	Kernel : rbf
Random Forest	max_depth : None	n_estimators : 200			
XGB Regressor	Gamma : 0.001	learning_rate : 0.2	max_depth : 4	n_estimators : 400	

Résultat de l'optimisation pour la prédiction de l'émission de CO2

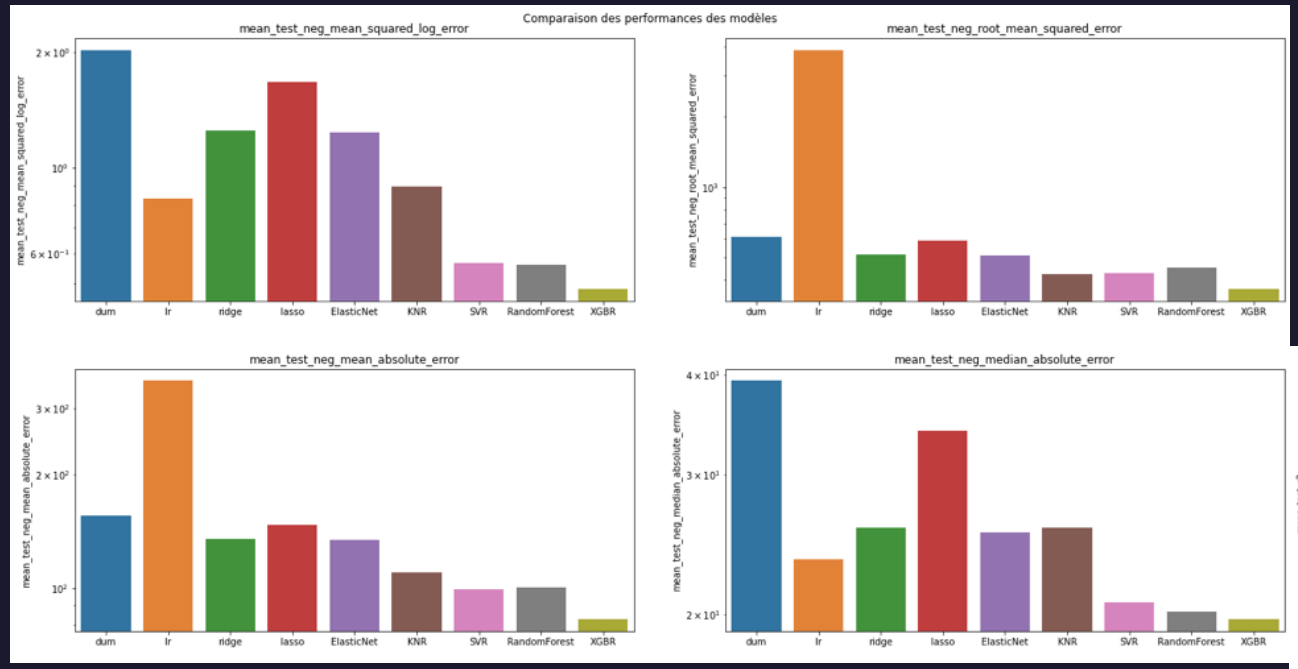
Modèle	Paramètres				
Dummy moyenne	strategy: mean				
Régression linéaire	fit_intercept : False	Positive: True			
Régression Ridge	Alpha : 1000				
Regression Lasso	Alpha : 0.395				
Elastic Net	Alpha : 1.014	l1_ratio : 0.0			
KNR	n_neighbors : 2	Weights : distance	leaf_size : 0	p : 1	
SVR	C : 10	Degree : 1	Epsilon : 0.1	Gamma : 0.01	Kernel : rbf
Random Forest	max_depth : None	n_estimators : 800			
XGB Regressor	Gamma : 0	learning_rate : 0.1	max_depth : 4	n_estimators : 300	

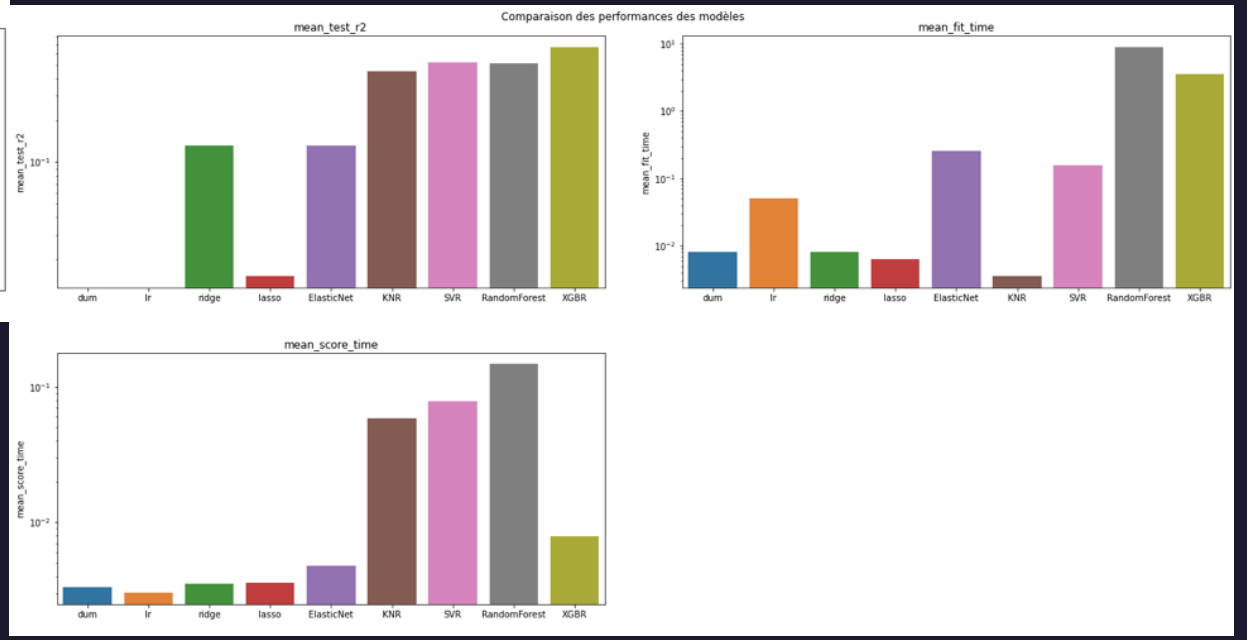
Comparaison des modèles pour la prédiction de la consommation d'énergie





Comparaison des modèles pour la prédiction de l'émission de CO2





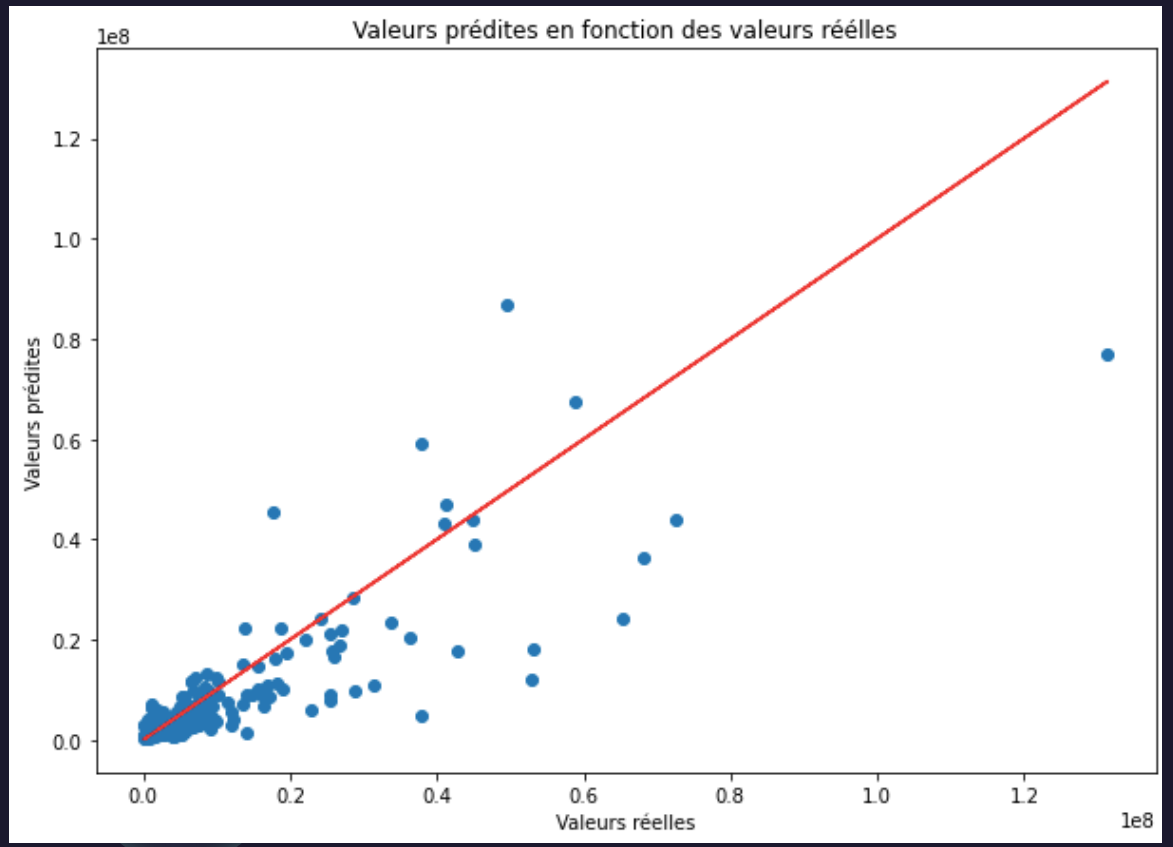
Généralisation des modèles pour la prédiction de la consommation d'énergie

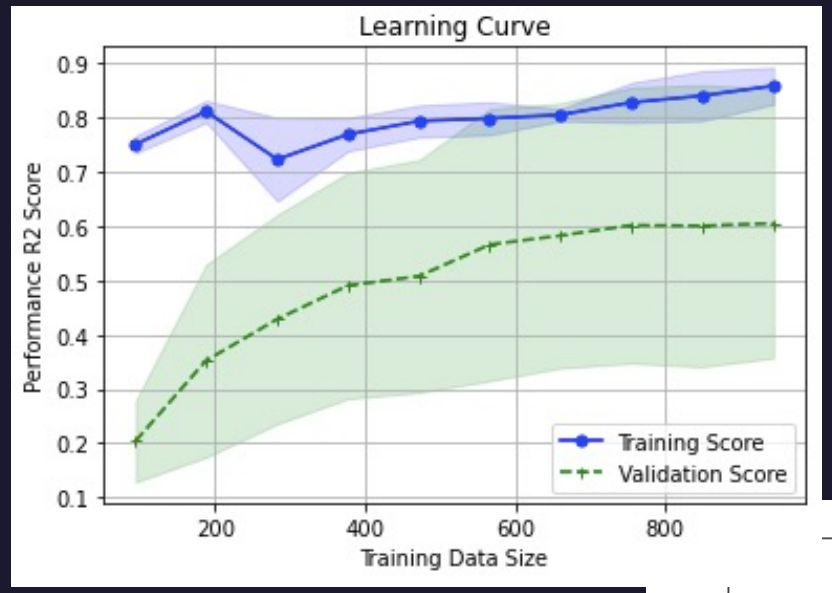
Modèle	Best R2 grid search	R2 test	RMSLE
Dummy moyenne	-0.06	-0.11	1.87
Régression linéaire	-	-	2.66
Régression Ridge	0.46	-0.20	1.21
Regression Lasso	0.25	0.11	1.42
Elastic Net	0.45	-0.05	1.24
KNR	0.54	0.39	0.80
SVR	0.50	0.61	0.48
Random Forest	0.62	0.69	0.56
XGB Regressor	0.80	0.64	0.50

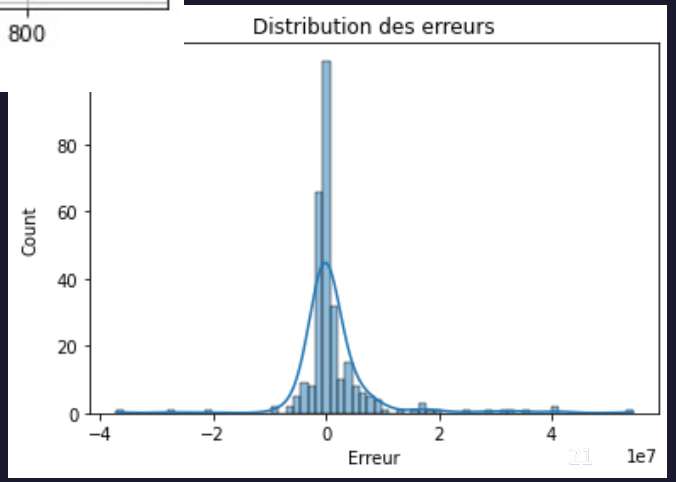
Généralisation des modèles pour la prédiction de l'émission de CO2

Modèle	Best R2 grid search	R2 test	RMSLE
Dummy moyenne	-0.07	-0.08	2.31
Régression linéaire	-	-	0.89
Régression Ridge	0.13	0.05	1.38
Regression Lasso	0.02	0.07	1.83
Elastic Net	0.13	0.06	1.44
KNR	0.45	0.24	1.02
SVR	0.52	0.61	0.49
Random Forest	0.51	0.52	0.50
XGB Regressor	0.67	0.64	0.46

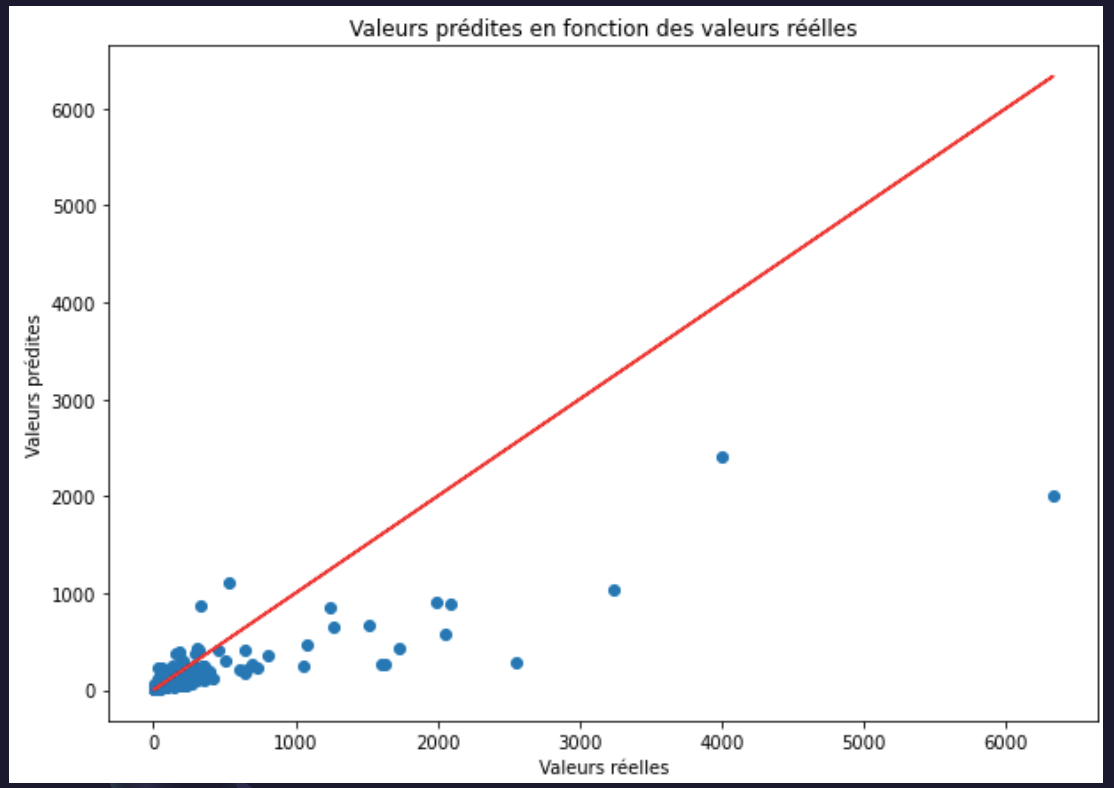
Choix du modèle pour la prédiction de la consommation d'énergie

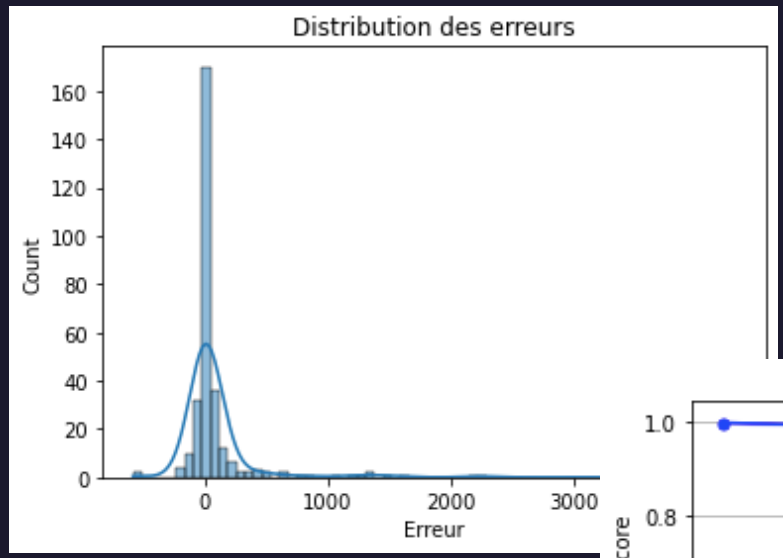


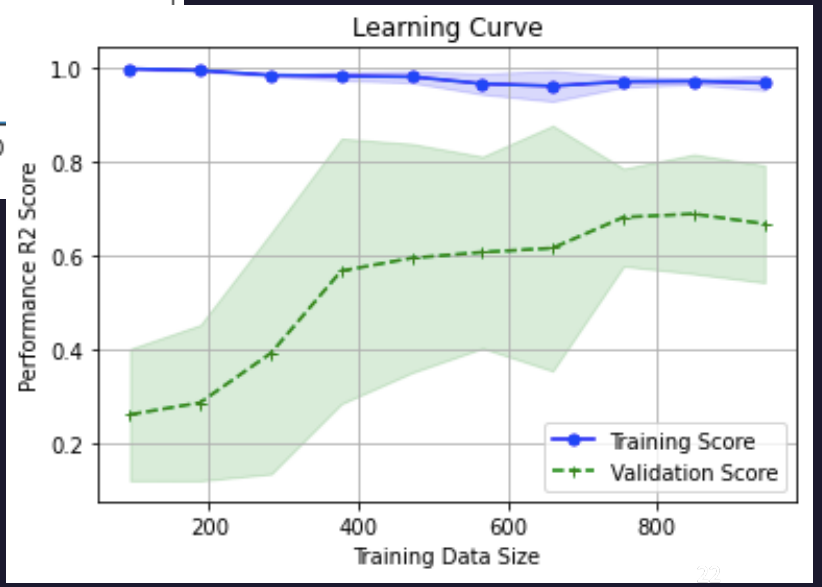




Choix du modèle pour la prédiction de l'émission de CO_2







Impact de l'EnergyStarScore

Consommation d'énergie :
RandomForestRegressor

		RMSLE	R2
Sans ESS	**0**	0.559049	0.688671
Avec ESS	**0**	0.411317	0.720652
Avec ESS imputée	**0**	0.405252	0.727552

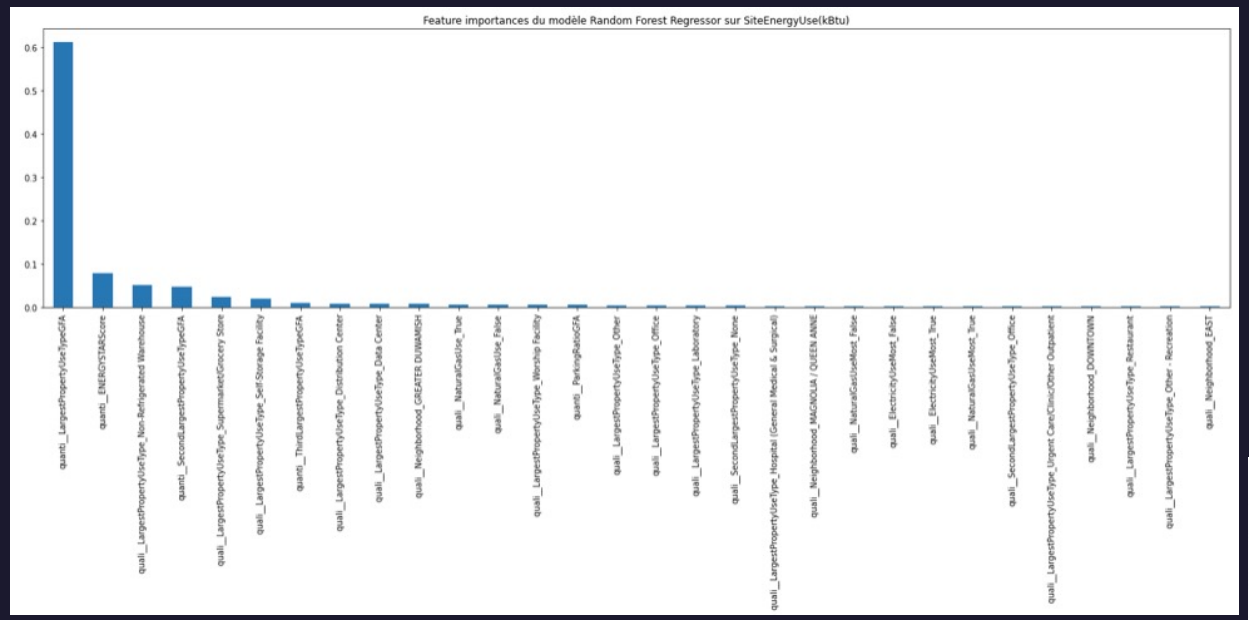
```
Meilleurs paramètres {'regressor__max_depth': None, 'regressor__n_estimators': 500}
```

Emission de CO2 :
XGBRegressor

		RMSLE	R2
Sans ESS	**0**	0.457038	0.644576
Avec ESS	**0**	0.176498	0.652233
Avec ESS imputée	**0**	0.280578	0.698308

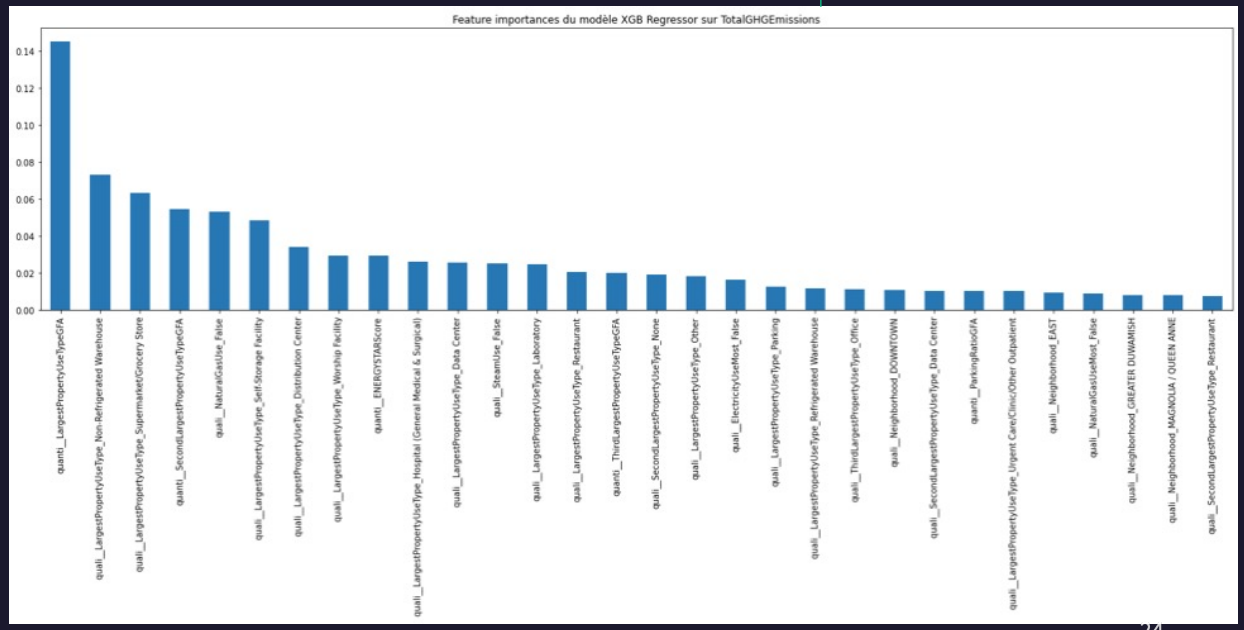
```
{'regressor__gamma': 0.01,
 'regressor__learning_rate': 0.2,
 'regressor__max_depth': 2,
 'regressor__n_estimators': 200}
```

Feature importance



- *LargestPropertyUseTypeGFA*,
- *EnergyStarScore*,
- *SecondLargestPropertyUseTypeGFA*,
- *LargestPropertyUseType*,
- *ThirdLargestPropertyUseTypeGFA*,
- *Neighborhood*,
- *NaturalGasUse*,
- *ParkingRatioGFA*
- *SecondLargestPropertyUseType*,
- *ElectricityUseMost*,
- *NaturalGasUseMost*

- LargestPropertyUseTypeGFA,
- LargestPropertyUseType,
- SecondLargestPropertyUseTypeGFA,
- NaturalGasUse,
- ENERGYSTARScore,
- SteamUse,
- ThirdLargestPropertyUseTypeGFA,
- SecondLargestPropertyUseType,
- ElectricityUseMost,
- Neighborhood,
- ParkingRatioGFA,
- NaturalGasUseMost.

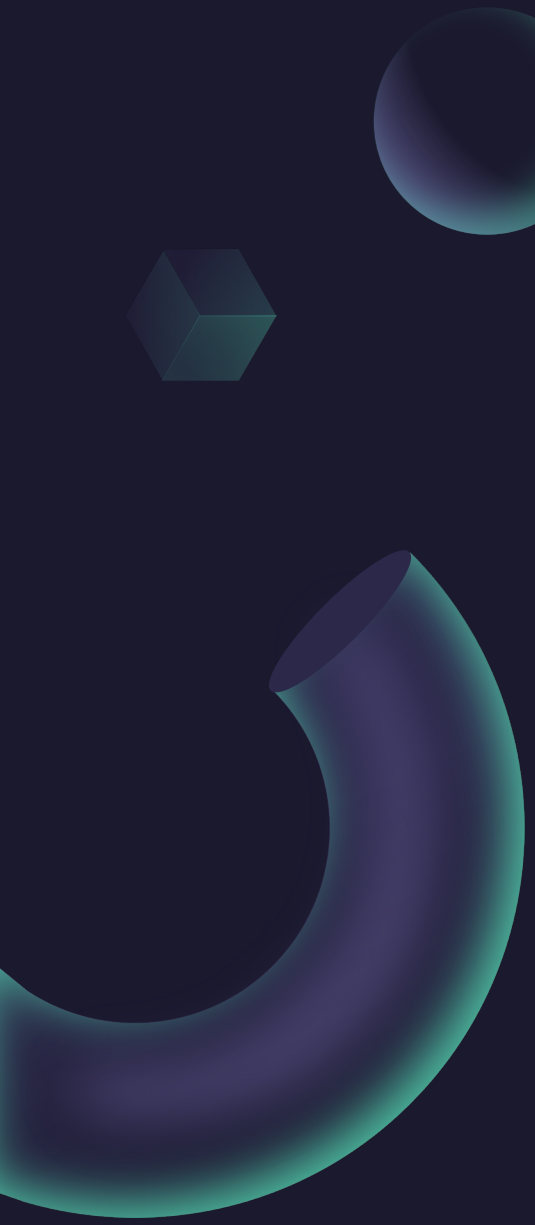


Conclusion

SiteEnergyUse(kBtu) → Random Forest Regressor

TotalGHGEmissions → XGB Regressor

Corrélation entre les 2 targets (0.92)



Merci de votre attention