OpenClassrooms

Projet 4 :

Anticipez les besoins en consommation de bâtiments

DJIAB Aymen

Sommaire

Présentation de la mission

On travaille pour la ville de **Seattle**.

- Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, notre équipe s’intéresse de près à la consommation et aux émissions des bâtiments non destinés à l’habitation.

- La mission est de tenter de prédire les émissions de CO_2 et la consommation totale d’énergie de bâtiments non destinées à l’habitation pour lesquels elles n’ont pas encore été mesurées.

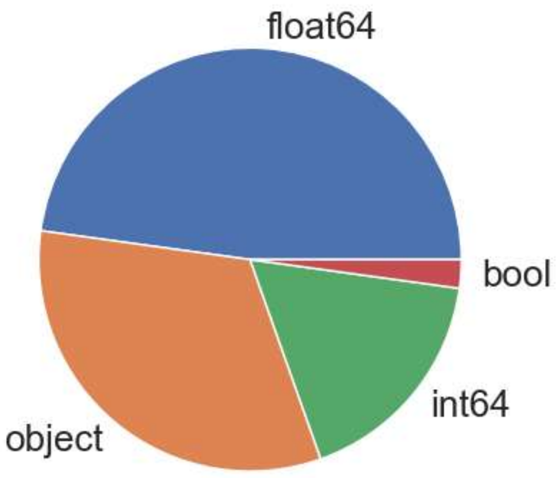
- Notre prédiction se basera sur les données structurelles des bâtiments (taille et usage des bâtiments, date de construction, situation géographique, ...)

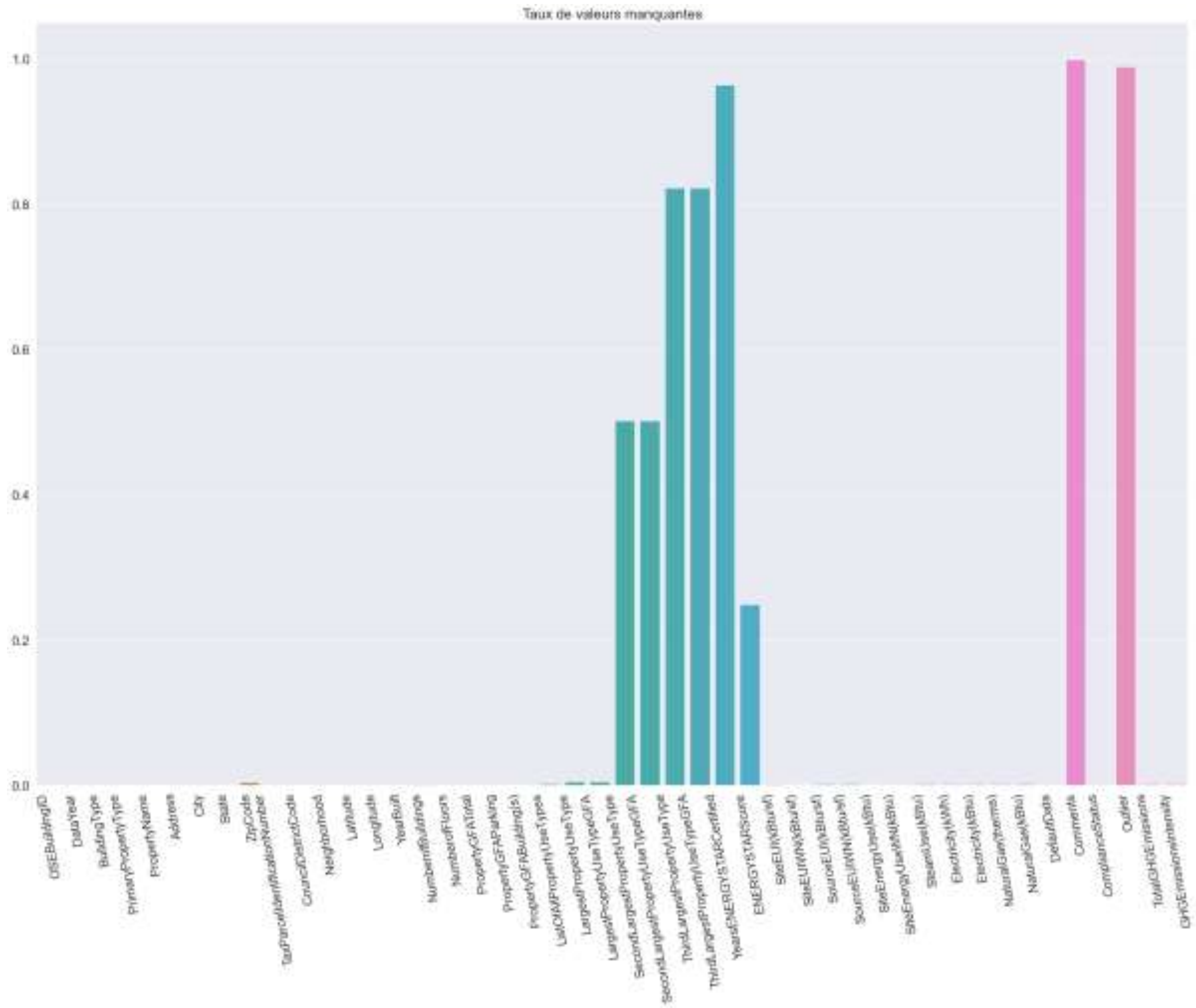
- On cherche également à évaluer l’intérêt de ‘’ l’ENERGY STAR SCORE ‘’ pour la prédiction d’émissions, qui est fastidieux à calculer, nous l'intégrerons dans modélisation et jugerons de son intérêt

Jeu de données (analyse exploratoire)

"2016_Building_Energy_Benchmarking.csv" :

- 3376 lignes
- 46 colonnes
- 19952 NAN au total
- Sois 7,78 % de NAN dans le fichier





Jeu de données (analyse exploratoire)

regroupement des colonnes en catégorie pour plus de simplicité → suppression des données des bâtiments résidentiel et des NAN → pour plus de facilité, création d'une colonne de distance entre les bâtiments et le centre-ville → suppression des colonnes numérique qui sont fortement corrélés → Analyse uni et bivariée entre les colonnes numériques

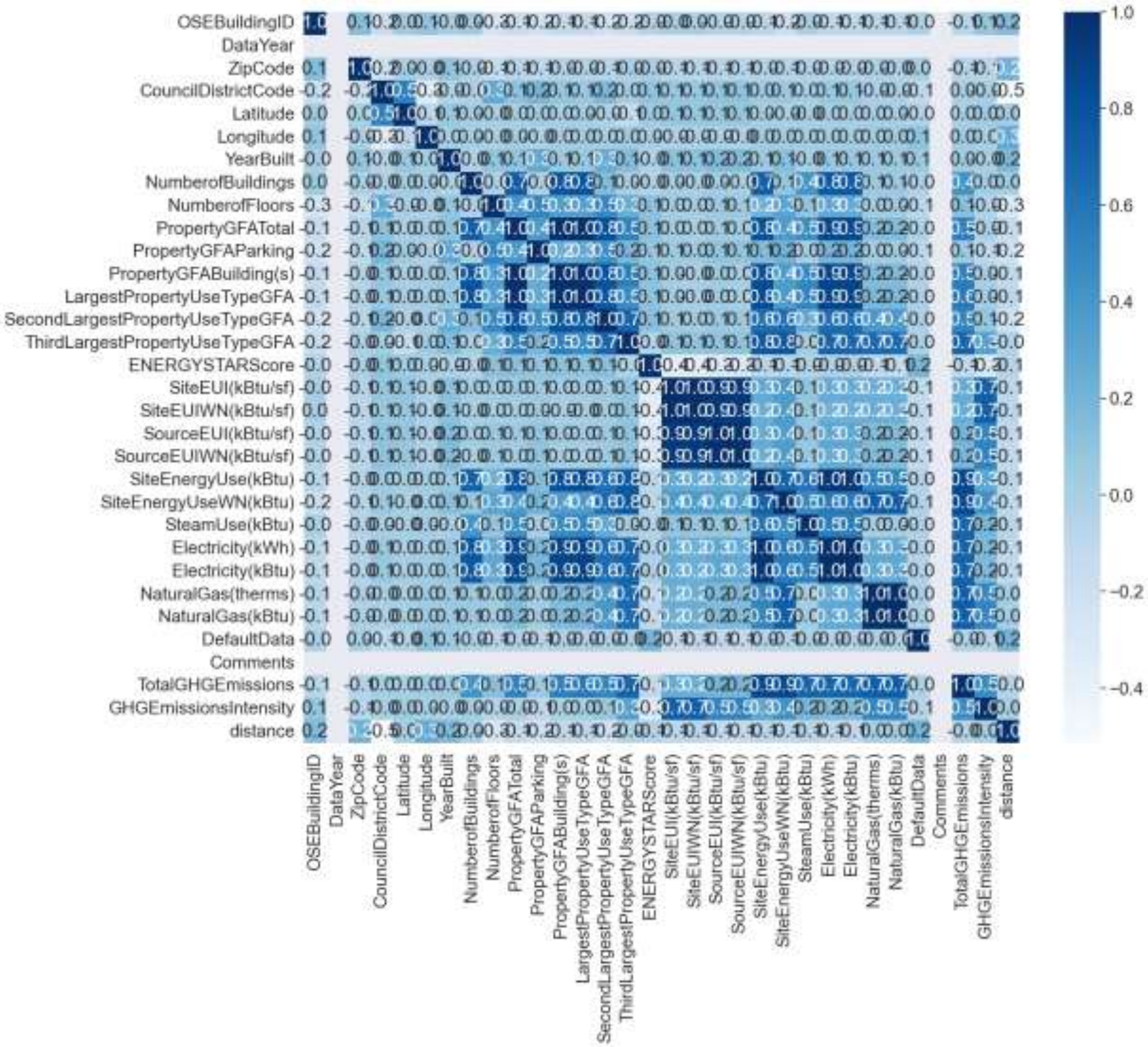
```
hébergements & bien-être    1824
stockage & service           604
bureau                       515
Education                    183
divertissement               162
soins                         68
Name: LargestPropertyUseType, dtype: int64
```

```
stockage & service          1274
bureau                       239
divertissement                94
hébergements & bien-être      27
soins                         22
Education                     13
Swimming Pool                 10
Name: SecondLargestPropertyUseType, dtype: int64
```

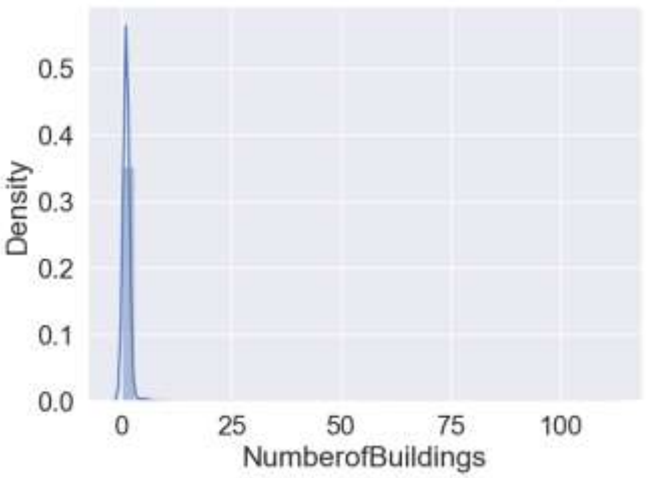
```
stockage & service            278
bureau                        119
divertissement                111
Swimming Pool                  29
hébergements & bien-être       23
soins                          20
Education                      14
Other - Technology/Science      2
Name: ThirdLargestPropertyUseType, dtype: int64
```

```
Low-Rise Multifamily           987
Mid-Rise Multifamily           564
stockage & service             440
Small- and Mid-Sized Office    293
hébergements & bien-être       227
Warehouse                      187
Large Office                   173
Education                      164
Mixed Use Property             133
divertissement                  83
soins                           59
Supermarket / Grocery Store     40
Residence Hall                  23
bureau                           3
Name: PrimaryPropertyType, dtype: int64
```

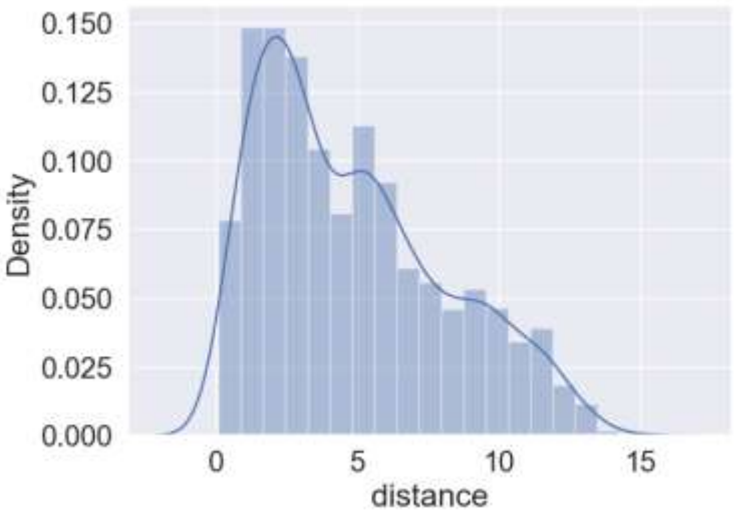
	distance
0	1.36889
1	1.54735
2	1.24055
3	1.31908
4	1.11532
...	...
1663	6.64022
1664	3.44794
1665	1.83618
1666	10.45376
1667	9.95580

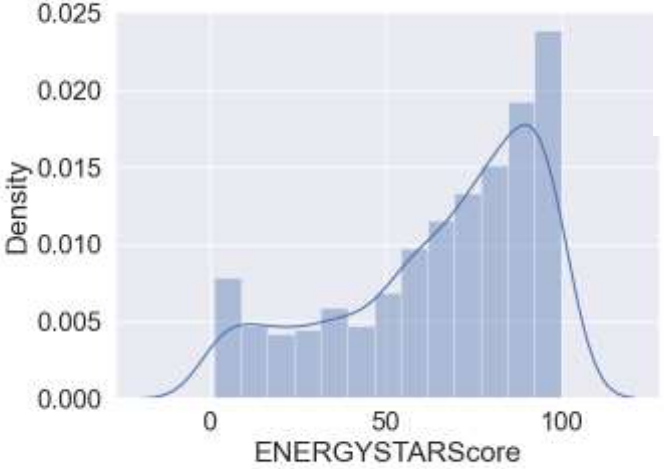


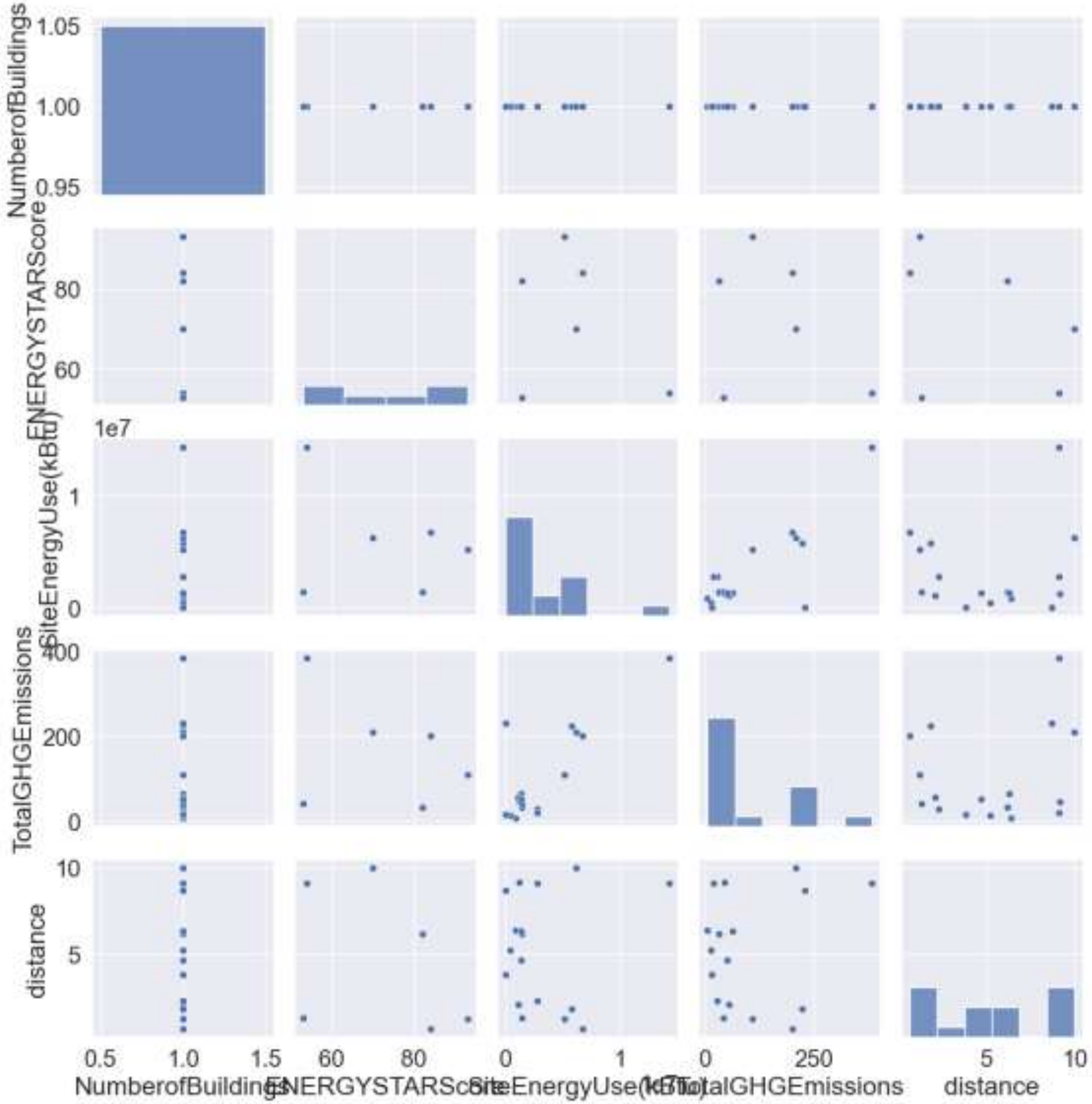
Jeu de données (analyse exploratoire)



Analyse Univariée







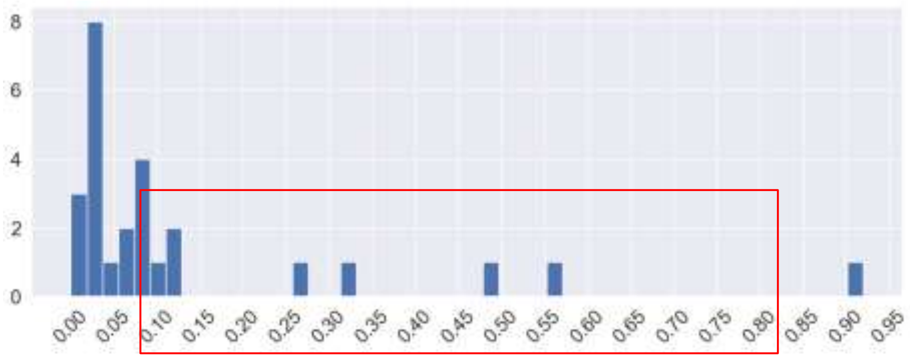
Analyse Bivariée

Jeu de données (analyse exploratoire)

ENERGY

	col_name	corr
0	TotalGHGEmissions	0.92000
1	PropertyGFATotal	0.57000
2	NumberofBuildings	0.48000
3	NumberofFloors	0.33000
4	type_8	0.27000
5	YearBuilt	0.12000
6	code_4	0.12000
7	code_5	0.09000
8	type_7	0.03000
9	type_5	0.03000
10	code_2	-0.00000
11	code_10	-0.00000
12	code_7	-0.01000
13	type_1	-0.02000
14	code_1	-0.02000
15	code_9	-0.03000
16	code_3	-0.03000
17	type_2	-0.03000
18	type_3	-0.03000
19	code_8	-0.04000
20	code_6	-0.07000
21	type_6	-0.07000
22	type_4	-0.08000
23	ENERGYSTARScore	-0.08000
24	type_9	-0.08000
25	distance	-0.10000

One Hot Encoding les colonnes catégoriels → Standard Scaling des colonnes non-cibles → Sélection des features pour les 2 cibles séparément avec méthode de corrélation entre la colonne cible (Energy ou GHG) et les autres colonnes → Sélection des features avec Seuil 0.1< seuil >0.8



	PropertyGFATotal	NumberofBuildings	NumberofFloors	type_8	YearBuilt	code_4	SiteEnergyUse(kBtu)
0	-0.18113	-0.04707	0.90919	0.00000	-1.15292	1.00000	-0.04651
1	-0.10695	-0.04707	0.78181	0.00000	1.01960	1.00000	0.00045
2	4.07237	-0.04707	4.60336	0.00000	0.16948	1.00000	2.59553
3	-0.31405	-0.04707	0.65442	0.00000	-1.18440	1.00000	-0.06396
4	0.24607	-0.04707	1.67350	0.00000	0.51582	1.00000	0.23428
...	...	...	...	...	...	...	...
1647	-0.49215	-0.04707	-0.36466	0.00000	-1.08995	0.00000	-0.15992
1649	-0.39405	-0.04707	-0.23727	0.00000	1.61782	0.00000	-0.28501
1658	-0.54768	-0.04707	-0.49204	0.00000	-0.36578	0.00000	-0.31829
1661	-0.53917	-0.04707	-0.49204	0.00000	-0.11389	0.00000	-0.32294
1663	-0.55438	-0.04707	-0.49204	0.00000	0.83068	0.00000	-0.30426

1092 rows × 7 columns

Jeu de données (analyse exploratoire)

GHG

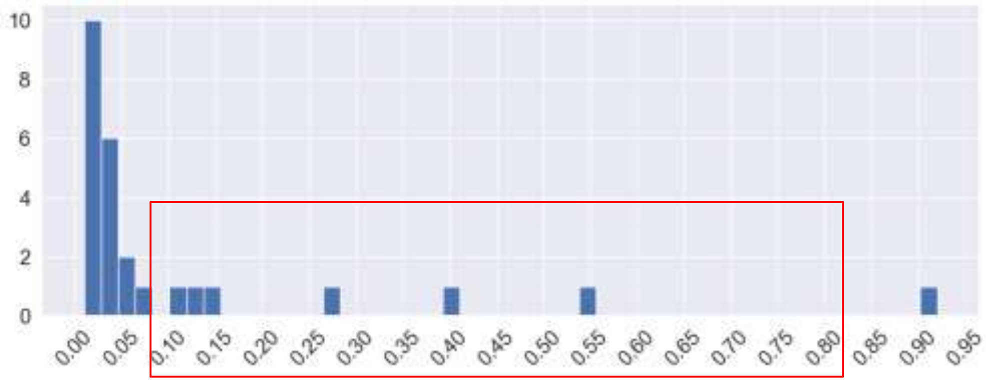
	col_name	corr
0	SiteEnergyUse(kBtu)	0.92000
1	NumberofBuildings	0.54000
2	PropertyGFATotal	0.40000
3	type_8	0.28000
4	NumberofFloors	0.14000
5	code_5	0.13000
6	YearBuilt	0.06000
7	type_7	0.06000
8	code_4	0.02000
9	code_10	0.01000
10	code_2	0.01000
11	code_1	-0.01000
12	type_3	-0.01000
13	type_1	-0.01000
14	type_2	-0.02000
15	code_7	-0.02000
16	code_9	-0.02000
17	code_3	-0.02000
18	code_6	-0.03000
19	distance	-0.03000
20	code_8	-0.04000
21	type_4	-0.04000
22	type_5	-0.04000
23	type_6	-0.04000
24	type_9	-0.07000
25	ENERGYSTARScore	-0.11000

One Hot Encoding les colonnes catégoriels

Standard Scaling des colonnes non-cibles

Sélection des features pour les 2 cibles séparément avec méthode de corrélation entre la colonne cible (Energy ou GHG) et les autres colonnes

Sélection des features avec Seuil 0.1< seuil >0.8



	NumberofBuildings	PropertyGFATotal	type_8	NumberofFloors	code_5	ENERGYSTARScore	TotalGHGEmissions
0	-0.04707	-0.18113	0.00000	0.90919	0.00000	-0.19108	0.07920
1	-0.04707	-0.10695	0.00000	0.78181	0.00000	-0.15597	0.13488
2	-0.04707	4.07237	0.00000	4.60336	0.00000	-0.78784	2.31137
3	-0.04707	-0.31405	0.00000	0.65442	0.00000	-0.33149	0.12344
4	-0.04707	0.24607	0.00000	1.67350	0.00000	0.33548	0.38871
...	...	...	...	...	...	...	...
1647	-0.04707	-0.49215	0.00000	-0.36466	0.00000	-1.98136	-0.06058
1649	-0.04707	-0.39405	0.00000	-0.23727	0.00000	0.40569	-0.21296
1658	-0.04707	-0.54768	0.00000	-0.49204	0.00000	0.33548	-0.21992
1661	-0.04707	-0.53917	0.00000	-0.49204	0.00000	0.96734	-0.21472
1663	-0.04707	-0.55438	0.00000	-0.49204	0.00000	-0.68253	-0.19876

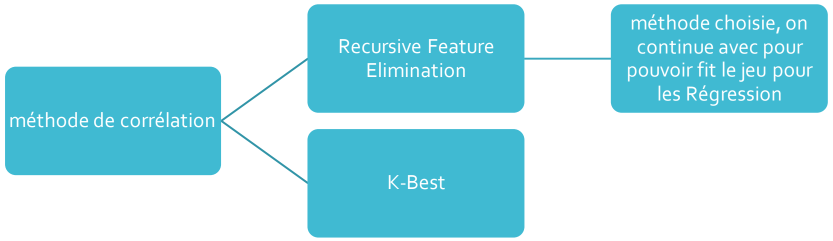
1092 rows × 7 columns

Les modèles de prédiction (modélisation)

Pour notre modélisation nous avons enregistré notre DataFrame « df.csv »

On applique aussi le Recursive Feature Elimination et le K-Best de scikit-learn

On optimise les performances en appliquant des transformations simples aux variables cibles « passage au logarithme »



ENERGY sans ENERGYSTAR SCORE

	PropertyGFATotal	YearBuilt	NumberofFloors	distance	code_1	code_6	type_5	type_7	type_8	type_9
0	-0.18113	-1.15292	0.90919	-1.06756	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
1	-0.10695	1.01960	0.78181	-1.01397	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
2	4.07237	0.16948	4.60336	-1.10610	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
3	-0.31405	-1.18440	0.65442	-1.08252	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
4	0.24607	0.51582	1.67350	-1.14370	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
...	...	...	...	...	...	...	...	...	...	...
1087	-0.49215	-1.08995	-0.36466	1.71342	0.00000	1.00000	1.00000	0.00000	0.00000	0.00000
1088	-0.39405	1.61782	-0.23727	-0.00160	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000
1089	-0.54768	-0.36578	-0.49204	-0.74040	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000
1090	-0.53917	-0.11389	-0.49204	1.01940	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000
1091	-0.55438	0.83068	-0.49204	0.51533	0.00000	1.00000	1.00000	0.00000	0.00000	0.00000

1092 rows × 10 columns

ENERGY avec ENERGYSTAR SCORE

	PropertyGFATotal	YearBuilt	NumberofFloors	ENERGYSTARScore	distance	code_1	code_6	type_5	type_7	type_8	type_9
0	-0.18113	-1.15292	0.90919	-0.19108	-1.06756	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
1	-0.10695	1.01960	0.78181	-0.15597	-1.01397	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
2	4.07237	0.16948	4.60336	-0.78784	-1.10610	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
3	-0.31405	-1.18440	0.65442	-0.33149	-1.08252	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
4	0.24607	0.51582	1.67350	0.33548	-1.14370	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
...	...	...	...	...	...	...	...	...	...	...	...
1087	-0.49215	-1.08995	-0.36466	-1.98136	1.71342	0.00000	1.00000	1.00000	0.00000	0.00000	0.00000
1088	-0.39405	1.61782	-0.23727	0.40569	-0.00160	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000
1089	-0.54768	-0.36578	-0.49204	0.33548	-0.74040	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000
1090	-0.53917	-0.11389	-0.49204	0.96734	1.01940	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000
1091	-0.55438	0.83068	-0.49204	-0.68253	0.51533	0.00000	1.00000	1.00000	0.00000	0.00000	0.00000

1092 rows × 11 columns

Les modèles de prédiction (modélisation)

GHG avec ENERGYSTAR SCORE

	PropertyGFATotal	YearBuilt	NumberofFloors	ENERGYSTARScore	distance	code_6	code_7	type_1	type_5	type_7	type_9
0	-0.18113	-1.15292	0.90919	-0.19108	-1.06756	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
1	-0.10695	1.01960	0.78181	-0.15597	-1.01397	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
2	4.07237	0.16948	4.60336	-0.78784	-1.10610	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
3	-0.31405	-1.18440	0.65442	-0.33149	-1.08252	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
4	0.24607	0.51582	1.67350	0.33548	-1.14370	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
...	...	...	...	...	...	...	...	...	...	...	...
1087	-0.49215	-1.08995	-0.36466	-1.98136	1.71342	1.00000	0.00000	0.00000	1.00000	0.00000	0.00000
1088	-0.39405	1.61782	-0.23727	0.40569	-0.00160	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000
1089	-0.54768	-0.36578	-0.49204	0.33548	-0.74040	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
1090	-0.53917	-0.11389	-0.49204	0.96734	1.01940	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
1091	-0.55438	0.83068	-0.49204	-0.68253	0.51533	1.00000	0.00000	0.00000	1.00000	0.00000	0.00000

1092 rows × 11 columns

GHG sans ENERGYSTAR SCORE

	PropertyGFATotal	YearBuilt	NumberofFloors	distance	code_6	code_7	type_1	type_5	type_7	type_9
0	-0.18113	-1.15292	0.90919	-1.06756	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
1	-0.10695	1.01960	0.78181	-1.01397	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
2	4.07237	0.16948	4.60336	-1.10610	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
3	-0.31405	-1.18440	0.65442	-1.08252	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
4	0.24607	0.51582	1.67350	-1.14370	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
...	...	...	...	...	...	...	...	...	...	...
1087	-0.49215	-1.08995	-0.36466	1.71342	1.00000	0.00000	0.00000	1.00000	0.00000	0.00000
1088	-0.39405	1.61782	-0.23727	-0.00160	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000
1089	-0.54768	-0.36578	-0.49204	-0.74040	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
1090	-0.53917	-0.11389	-0.49204	1.01940	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000
1091	-0.55438	0.83068	-0.49204	0.51533	1.00000	0.00000	0.00000	1.00000	0.00000	0.00000

1092 rows × 10 columns

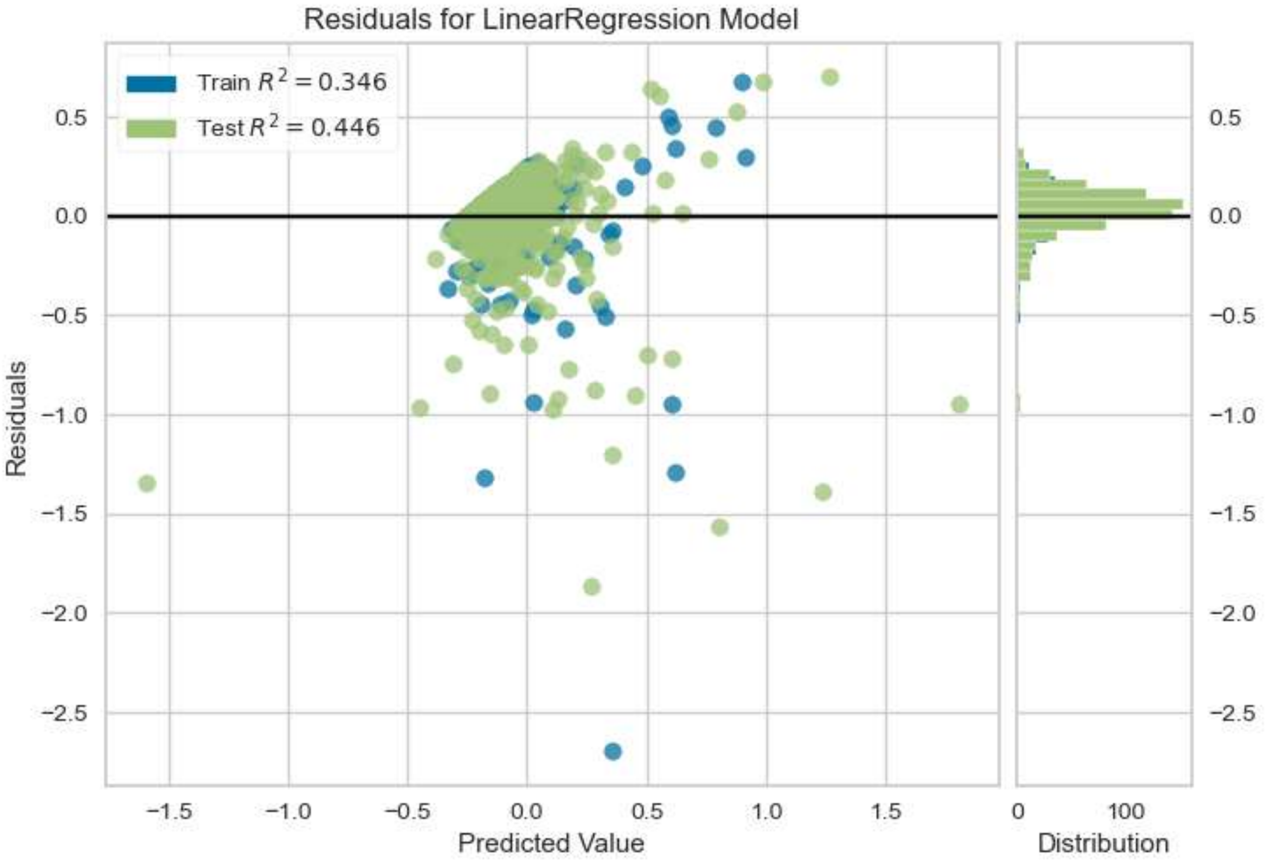
Les modèles de prédiction (modélisation)

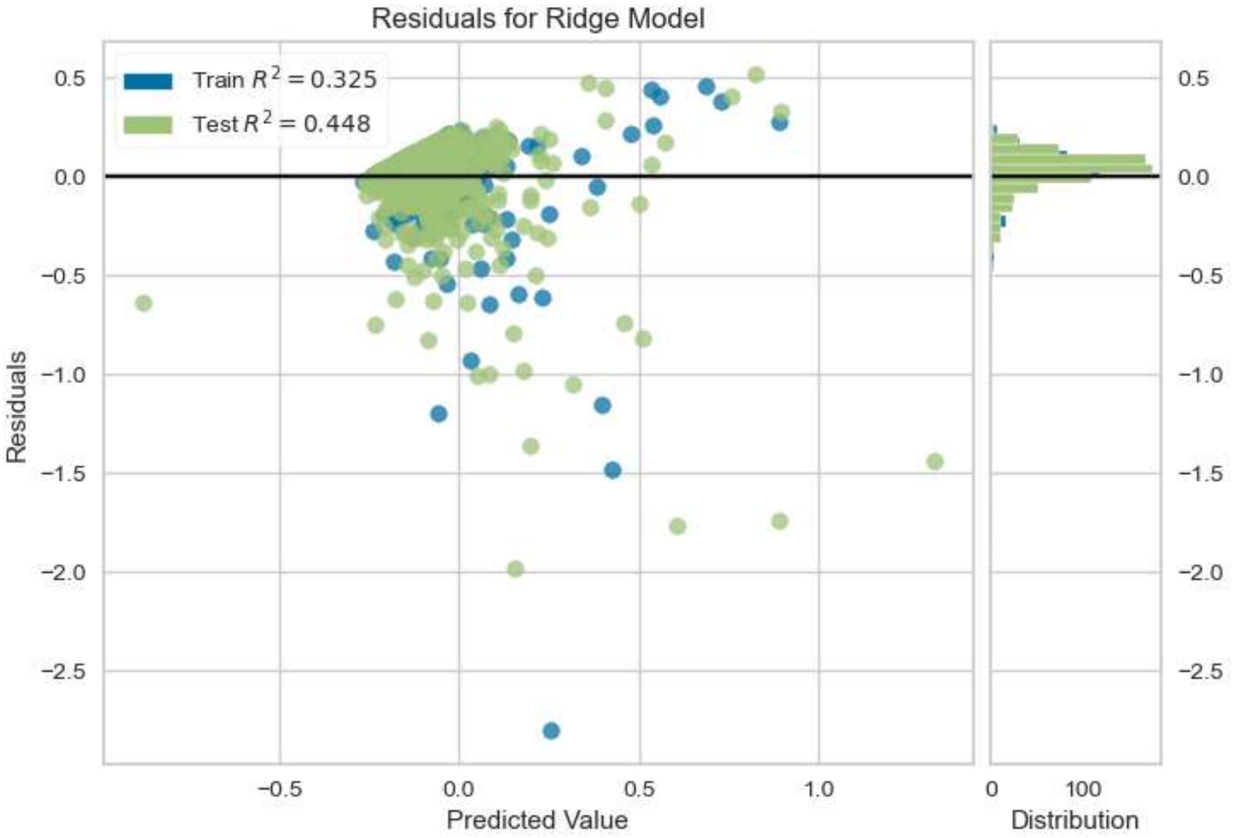
- Régression
  - Simple (Linéar)
  - Ridge
  - Lasso
  - ElasticNet
  - RandomForestRegressor
  - GradientBoostingRegressor

Linéar	• ENERGY et GHG avec ENERGY STAR SCORE • ENERGY et GHG sans ENERGY STAR SCORE
Ridge	• ENERGY et GHG avec ENERGY STAR SCORE • ENERGY et GHG avec ENERGY STAR SCORE
Lasso	• ENERGY et GHG avec ENERGY STAR SCORE • ENERGY et GHG avec ENERGY STAR SCORE
Lasso	• ENERGY et GHG avec ENERGY STAR SCORE • ENERGY et GHG sans ENERGY STAR SCORE
ElasticNet	• ENERGY et GHG avec ENERGY STAR SCORE • ENERGY et GHG sans ENERGY STAR SCORE
RandomForestRegressor	• ENERGY et GHG avec ENERGY STAR SCORE • ENERGY et GHG sans ENERGY STAR SCORE
GradientBoostingRegressor	• ENERGY et GHG avec ENERGY STAR SCORE • ENERGY et GHG sans ENERGY STAR SCORE

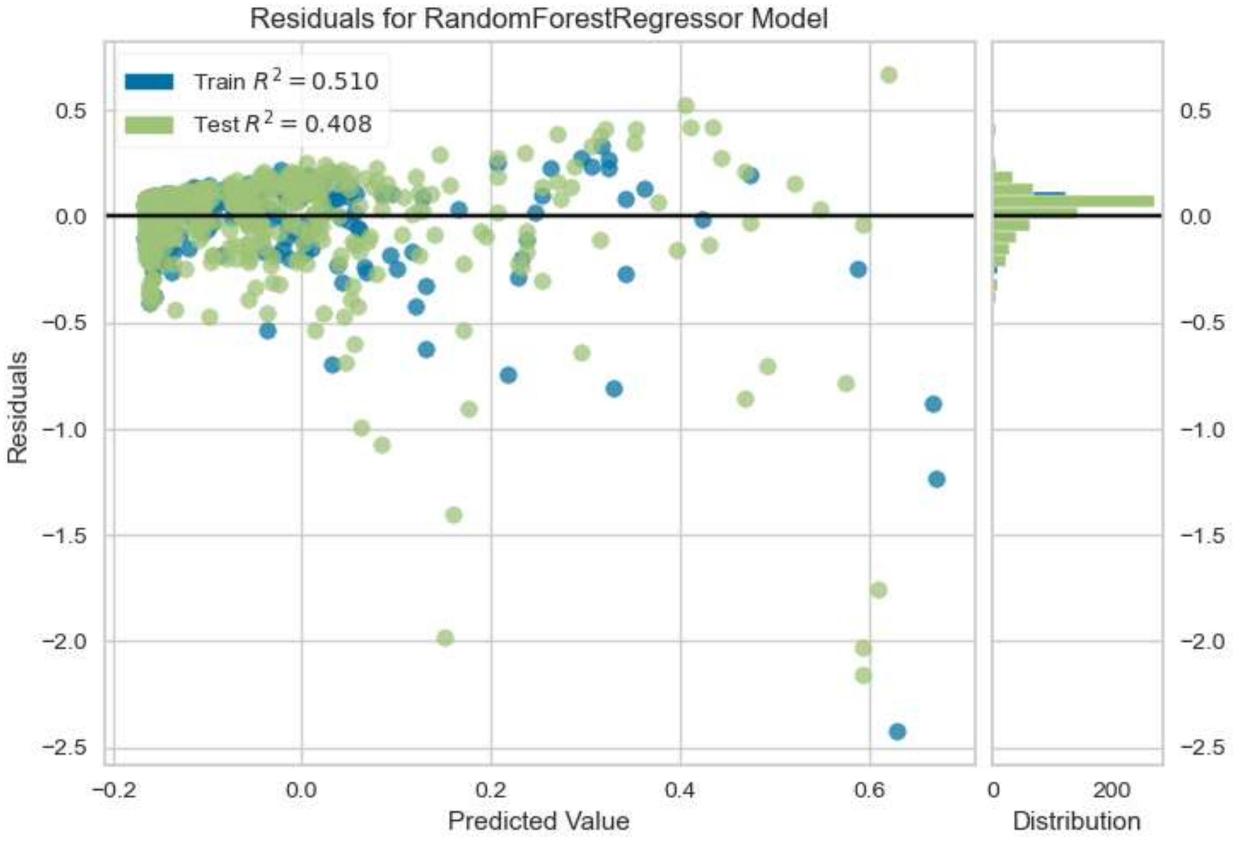
Les modèles de prédiction (modélisation)

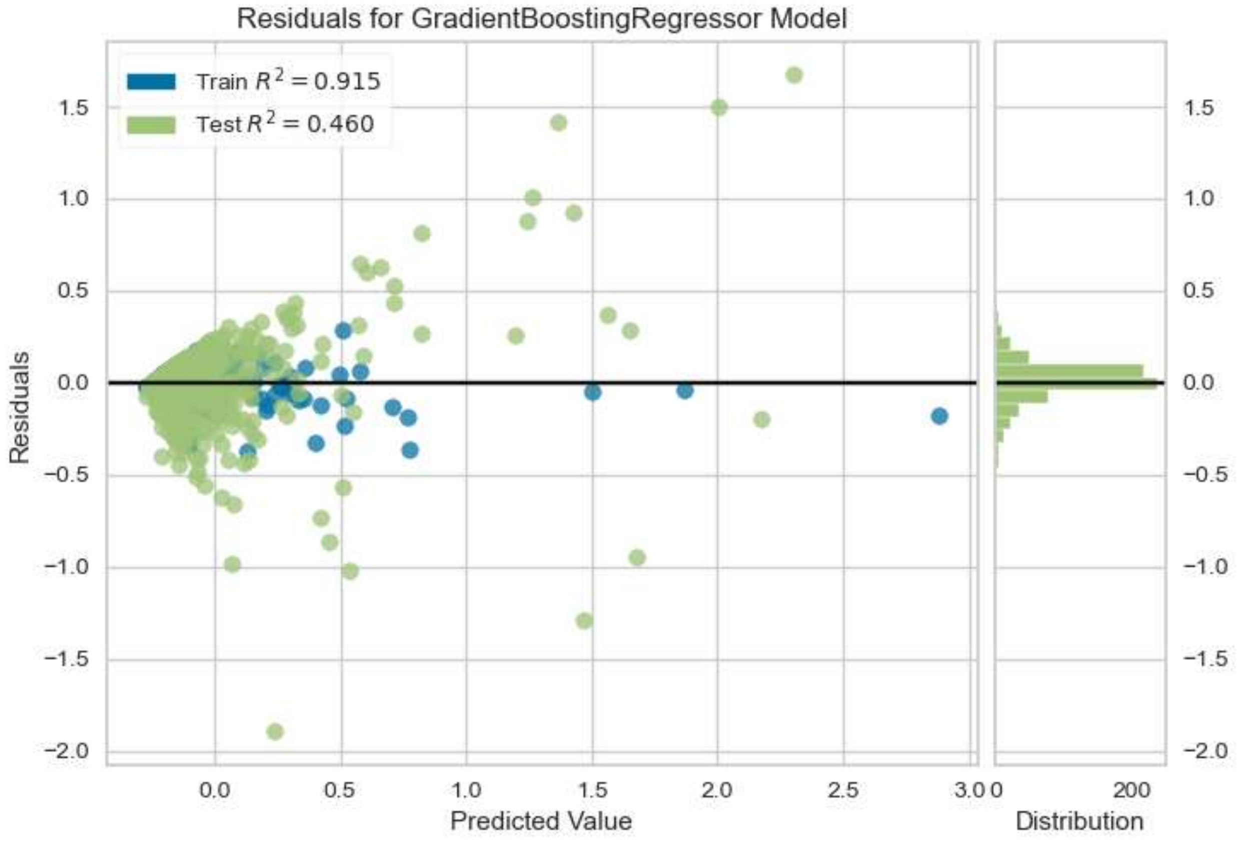
GHG avec Energy Star Score





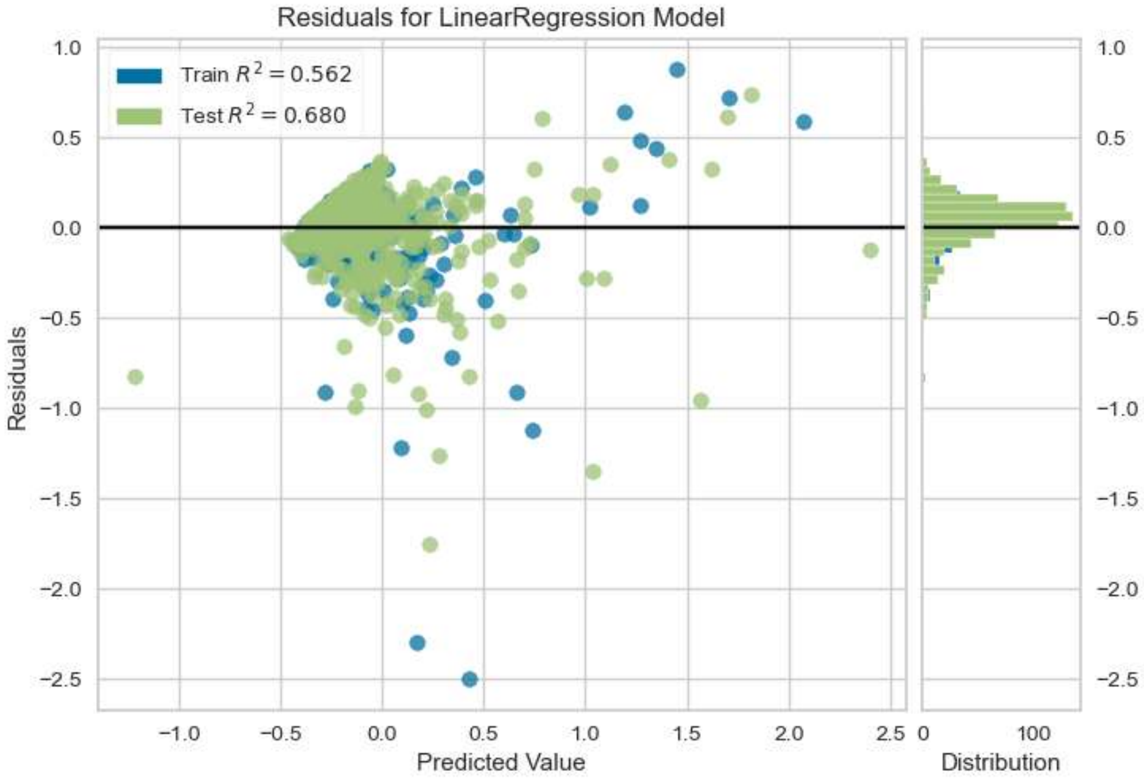
Les modèles de prédiction (modélisation)

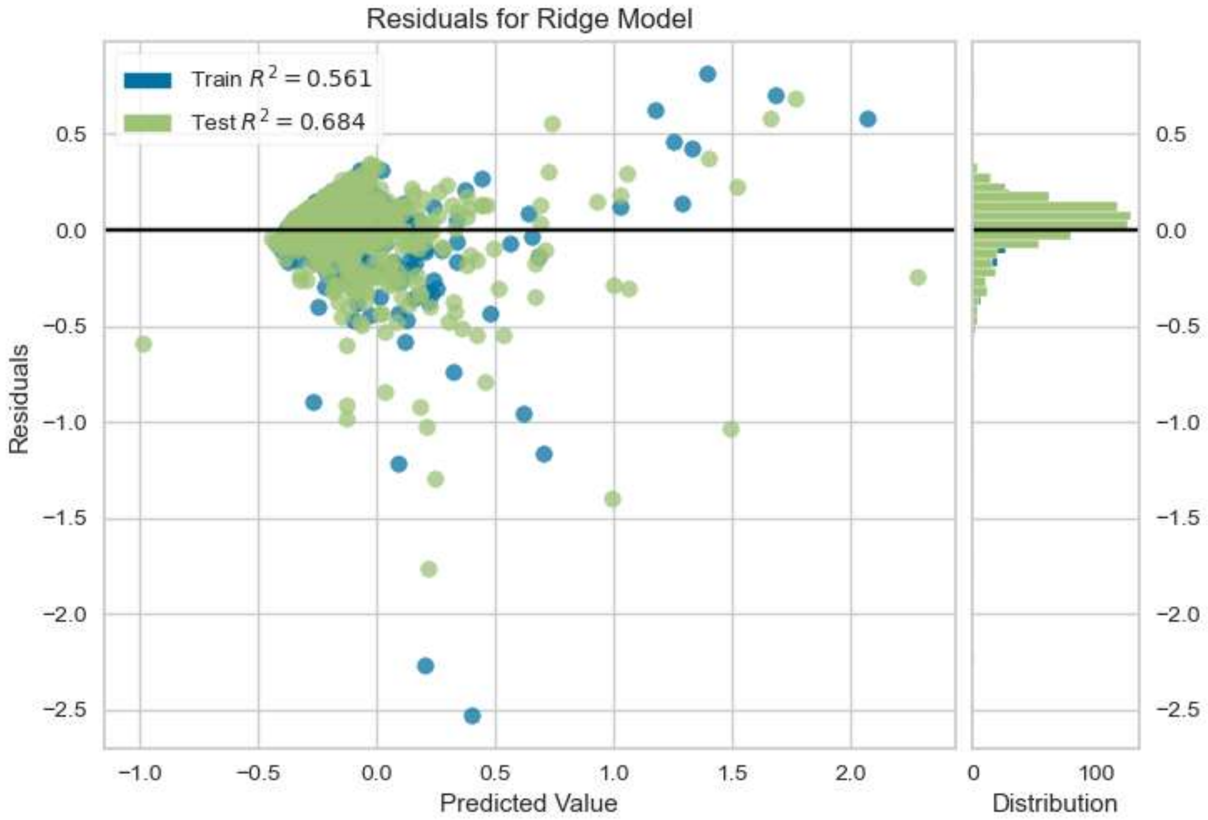




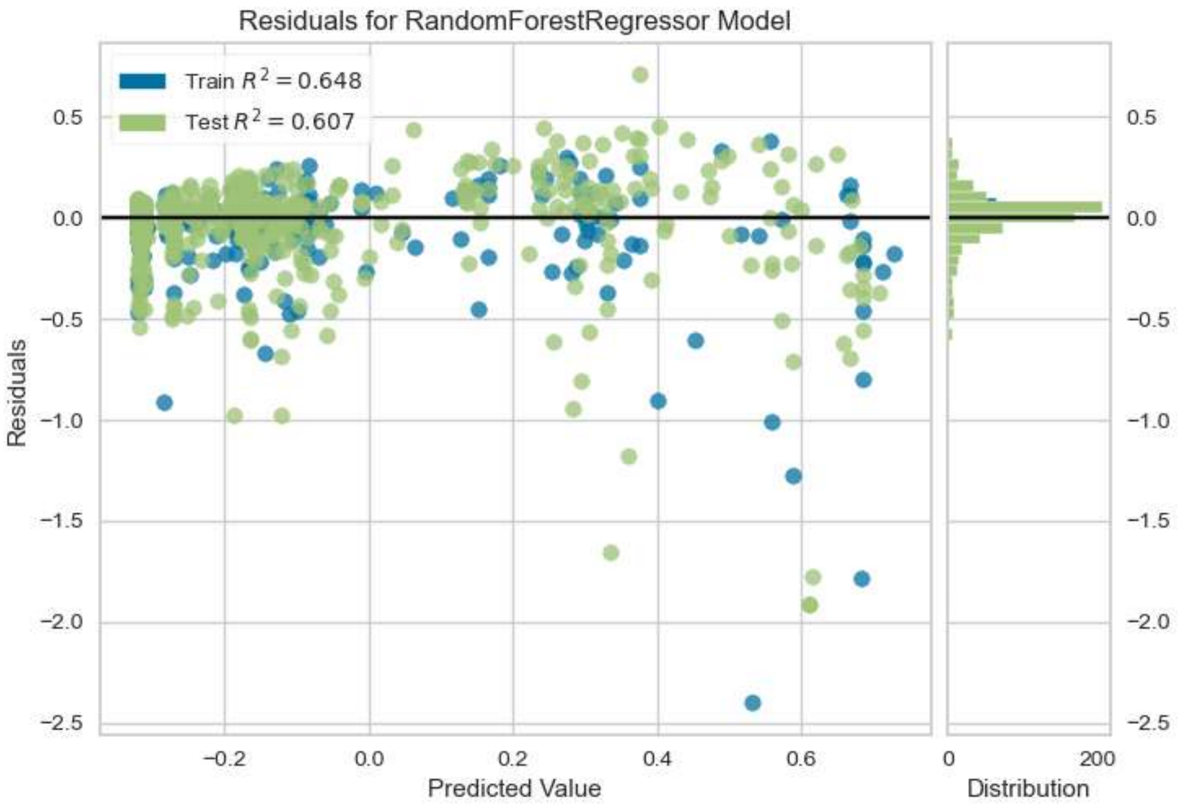
Les modèles de prédiction (modélisation)

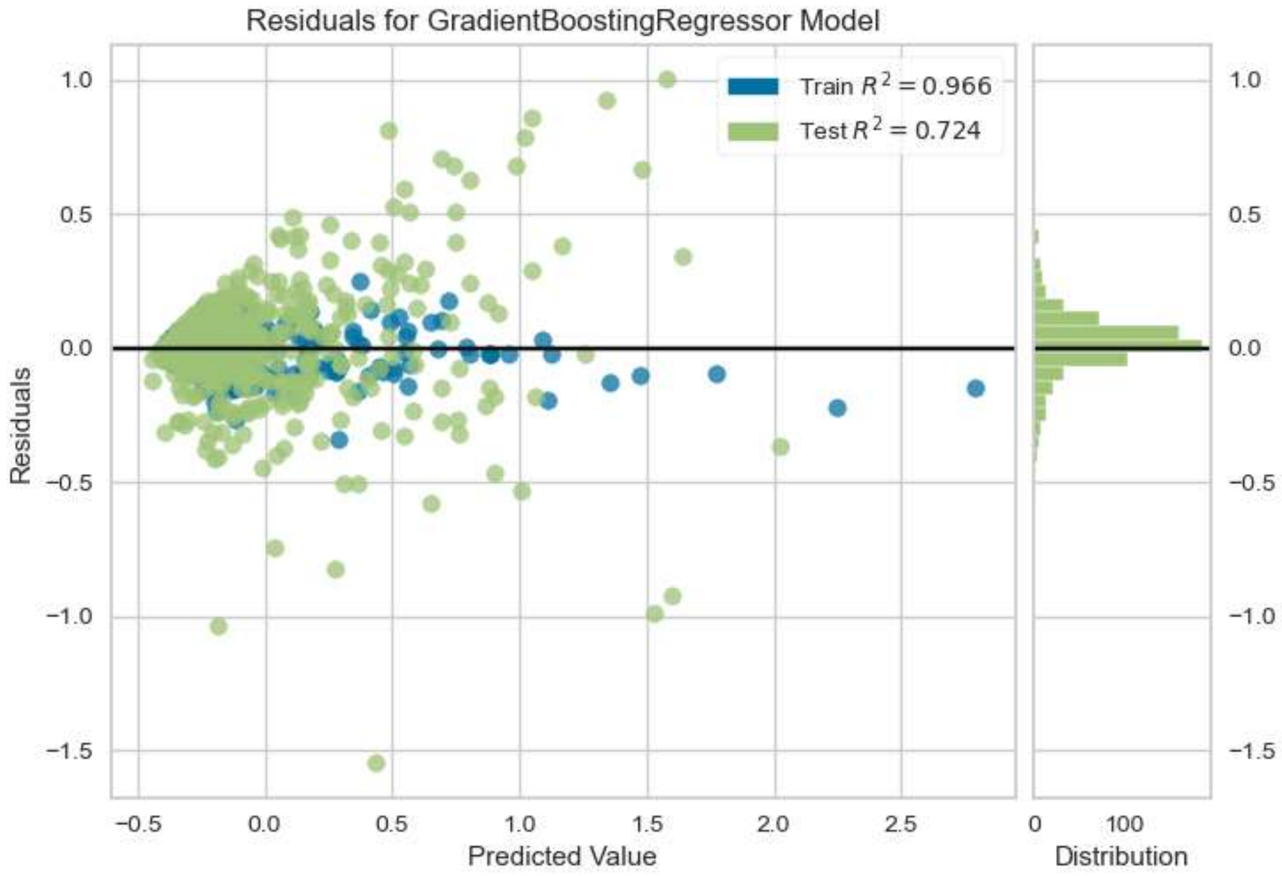
ENERGY avec Energy Star Score





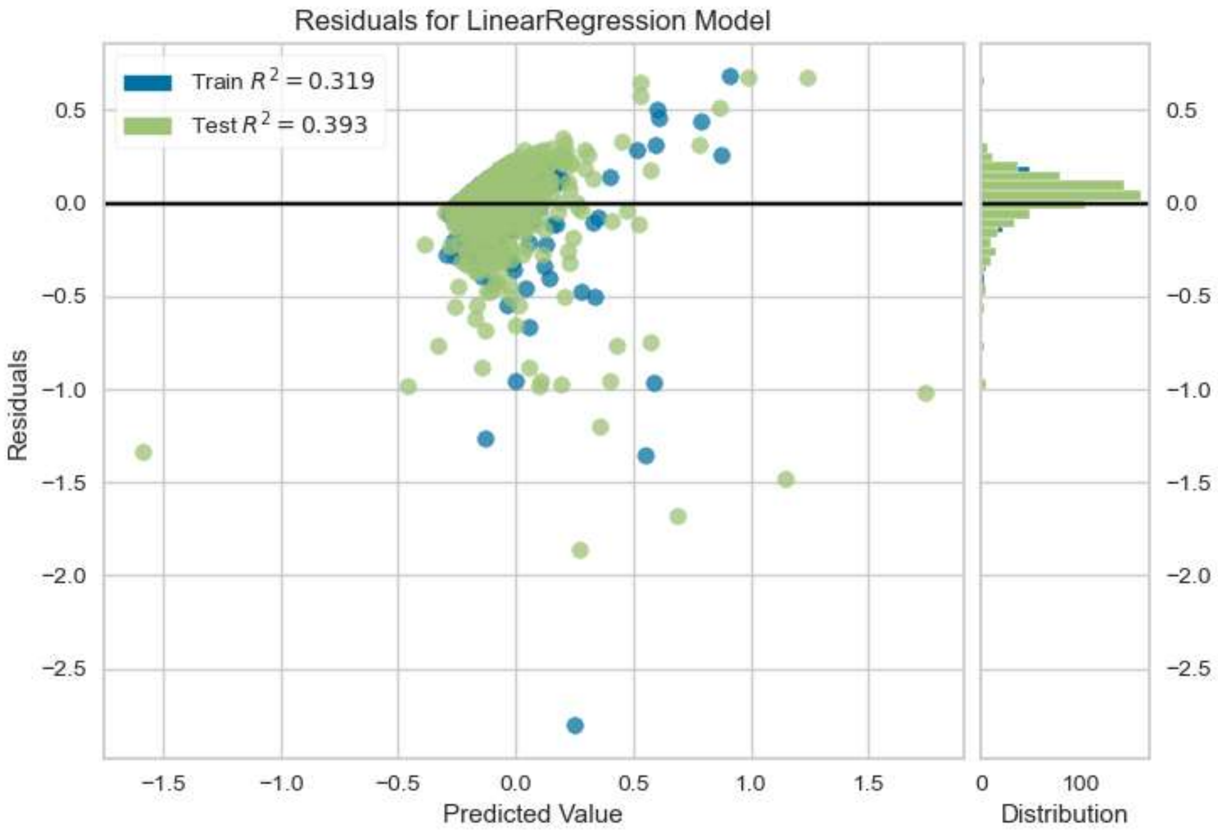
Les modèles de prédiction (modélisation)

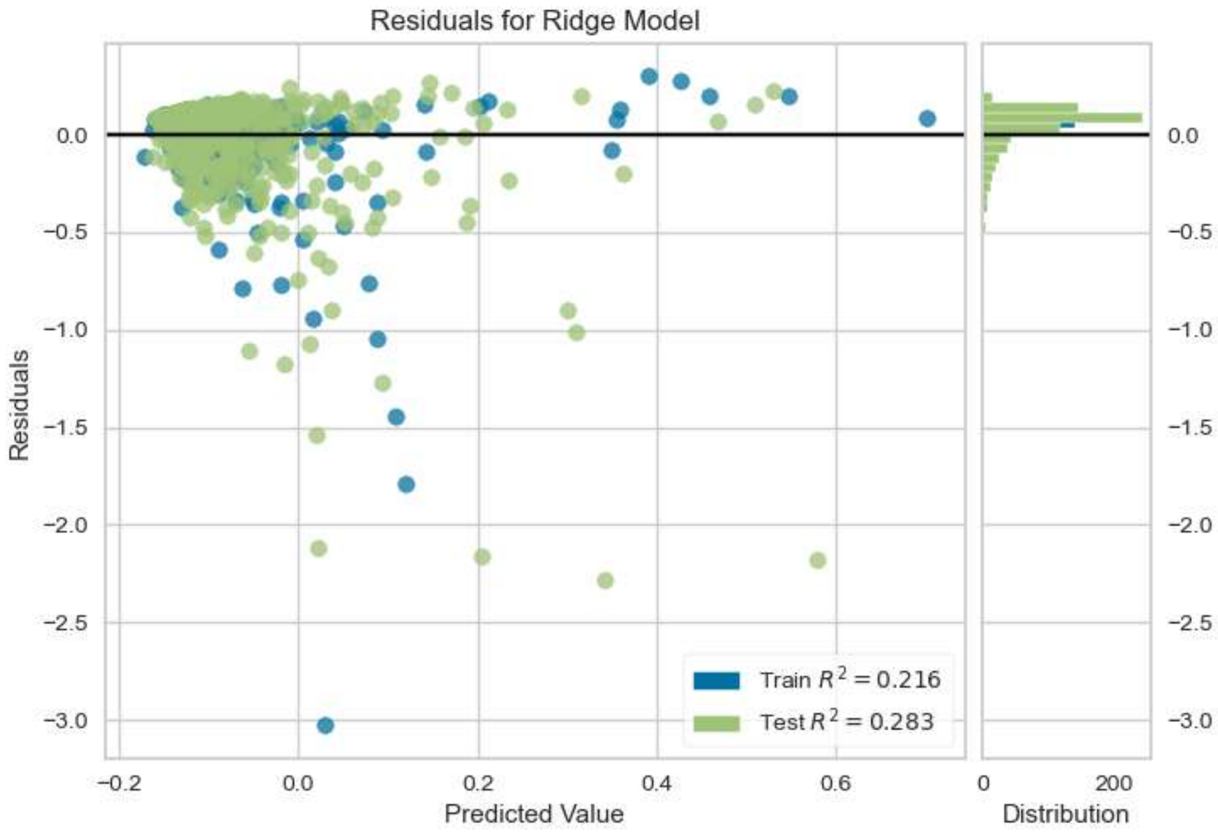




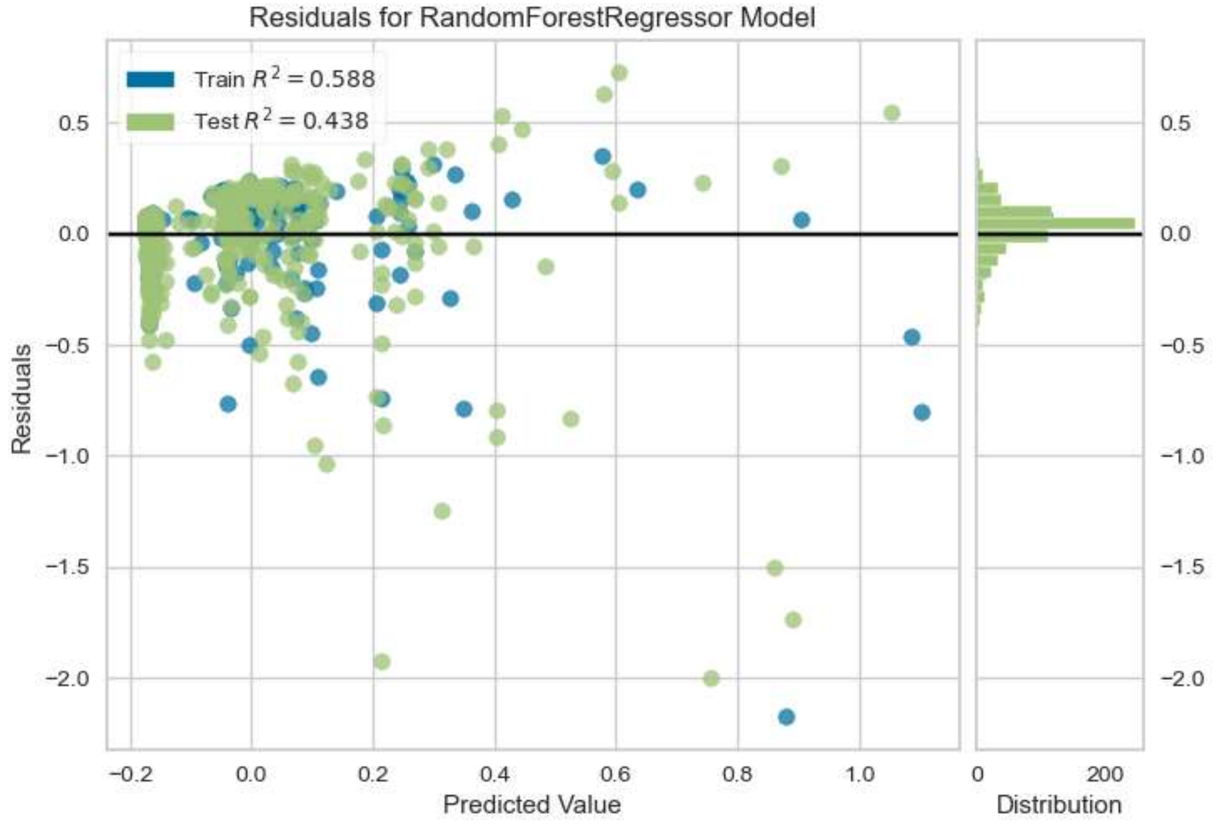
Les modèles de prédiction (modélisation)

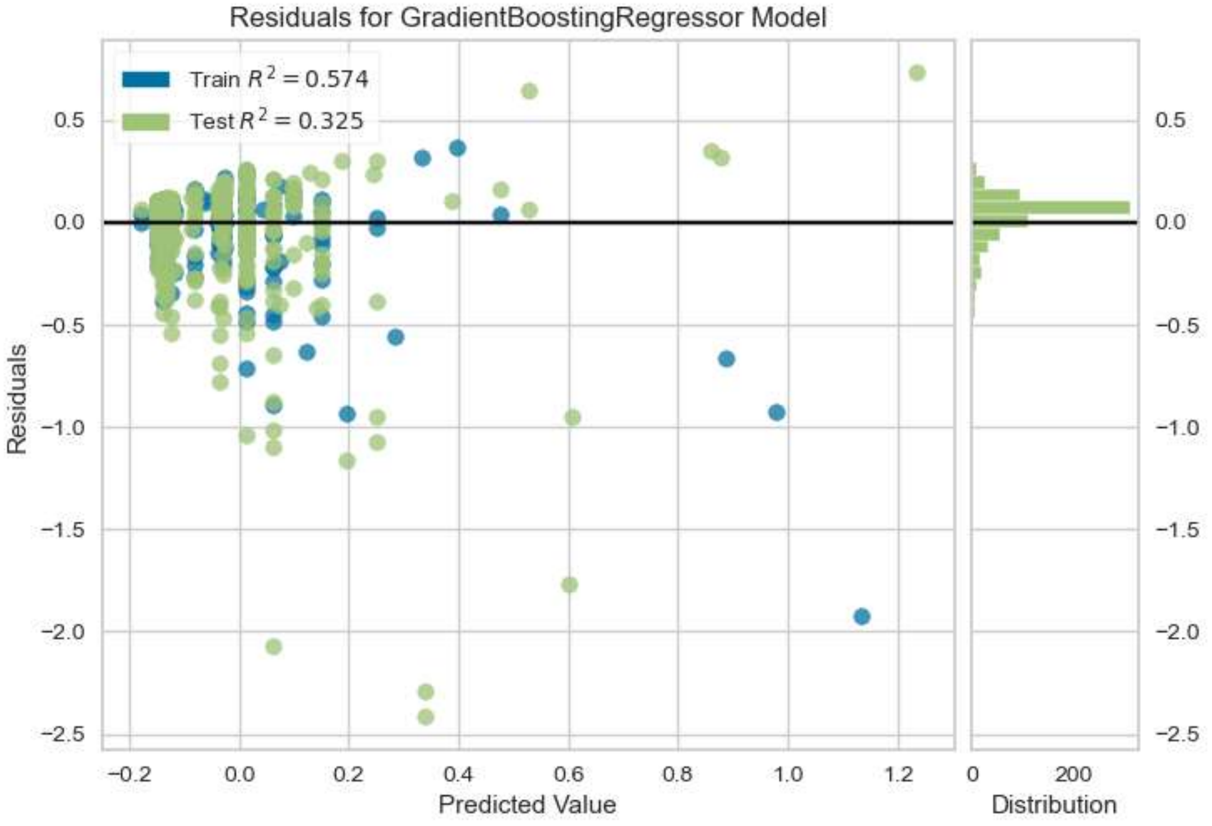
GHG sans Energy Star Score





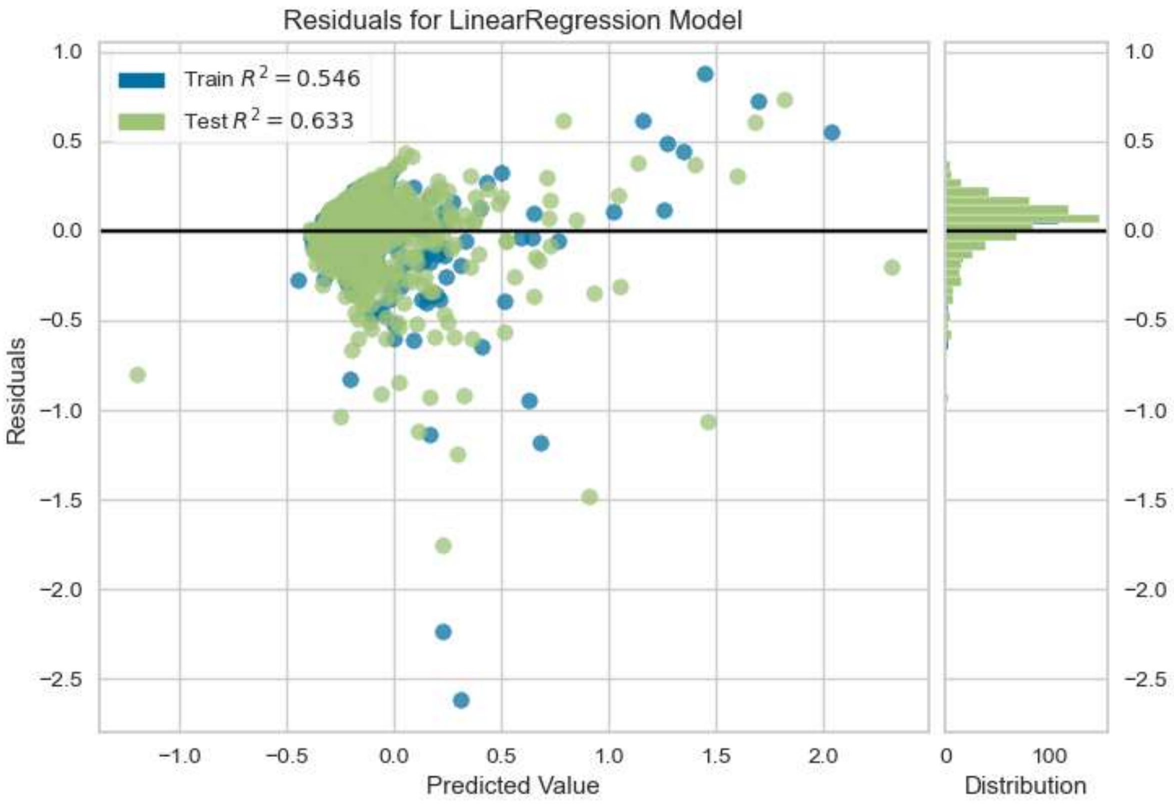
Les modèles de prédiction (modélisation)

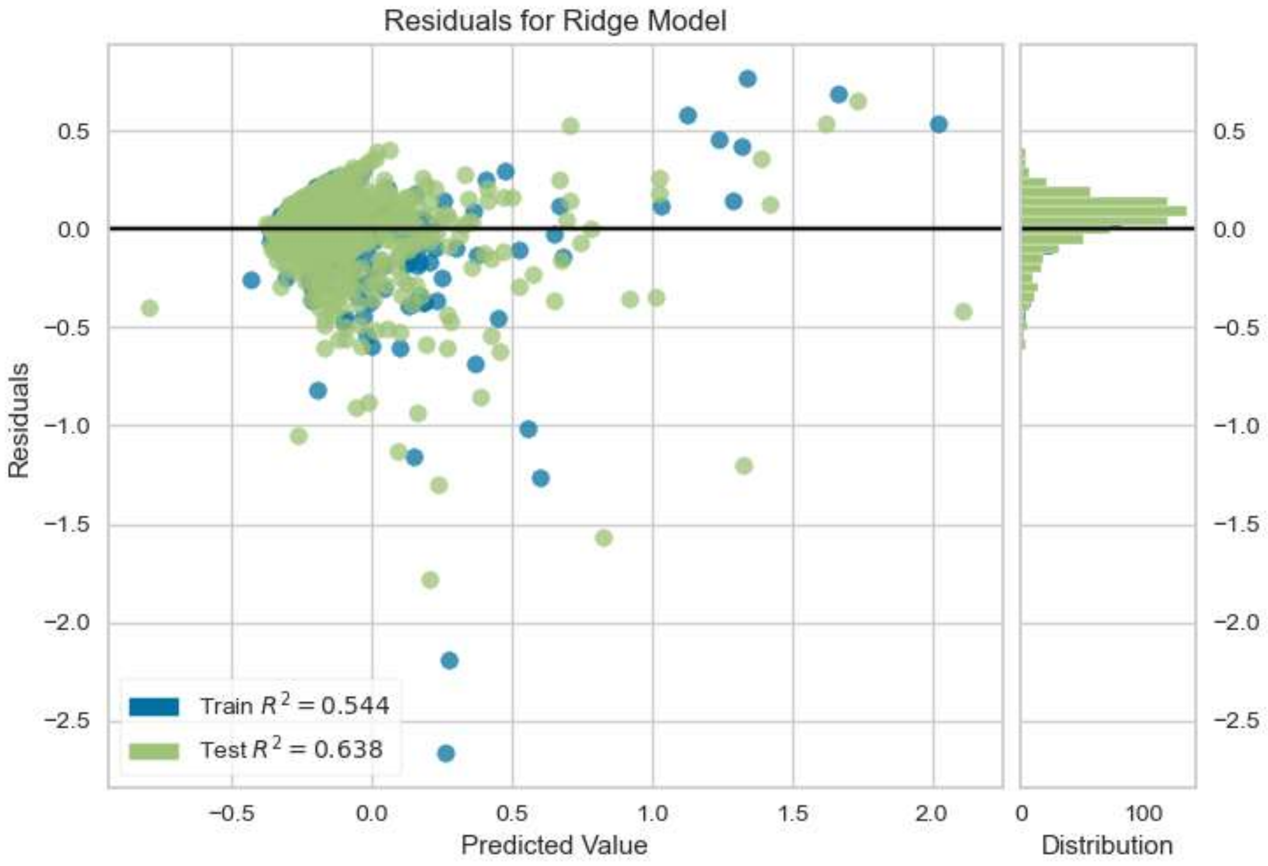




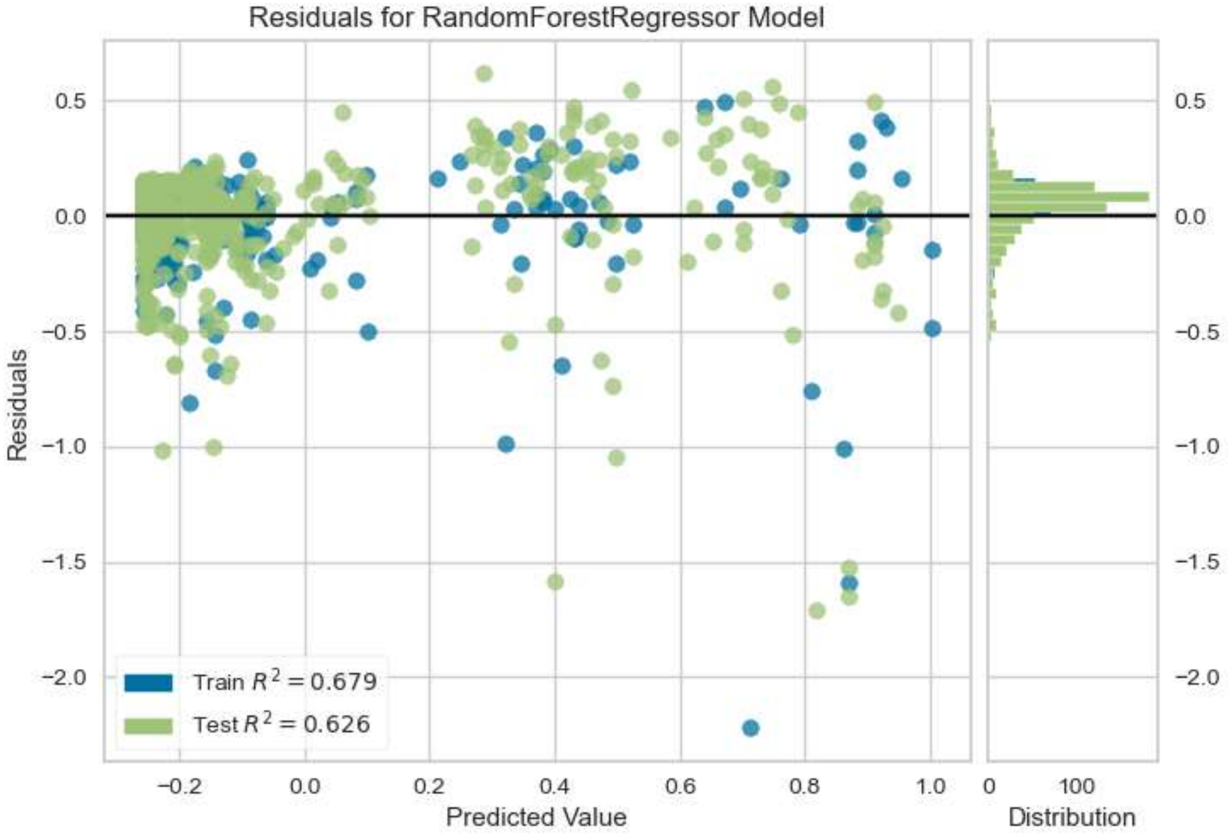
Les modèles de prédiction (modélisation)

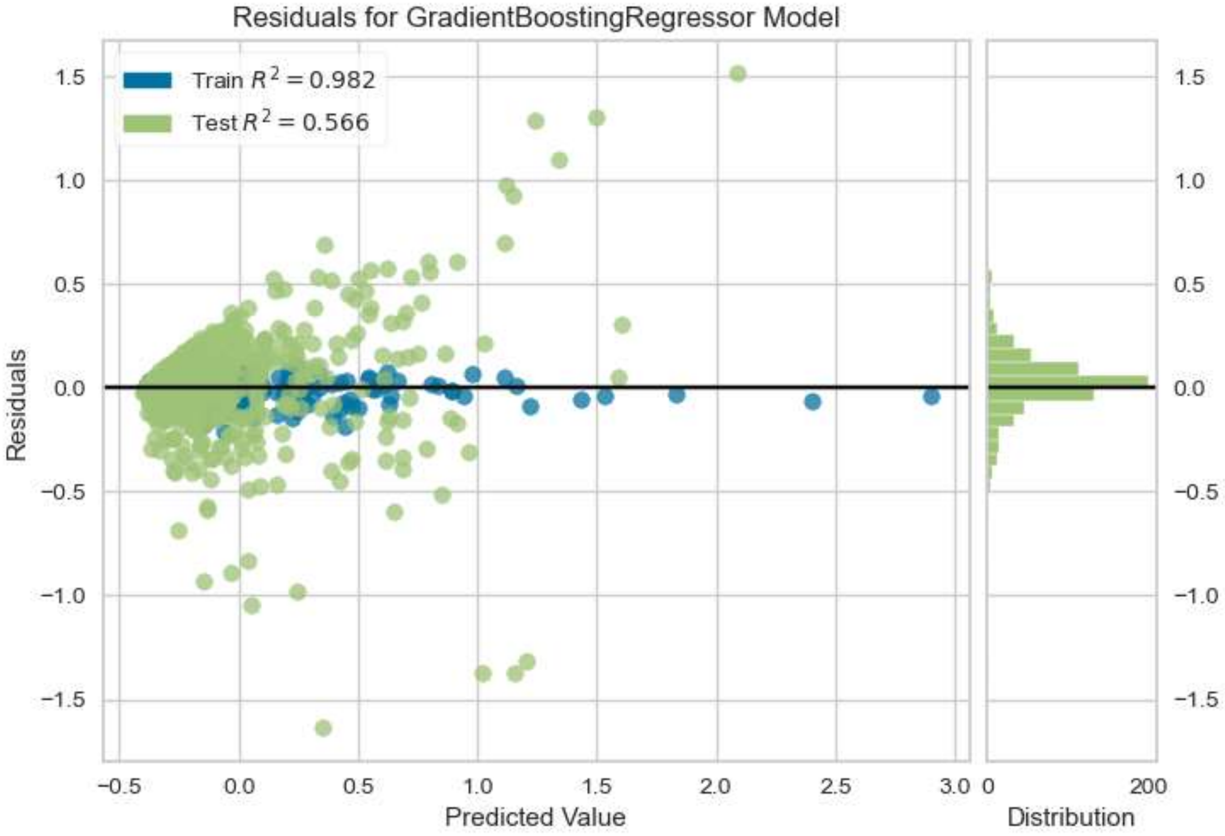
ENERGY sans Energy Star Score





Les modèles de prédiction (modélisation)





Les modèles de prédiction (modélisation)

Modèle avec ENERGY STAR SCORE

Modele pour GHG avec ENERGYSTARSCORE	R² test	R² train	RMSE test	RMSE train	Prediction time	Best parameters	Best score
LinearRegression()	0.44625	0.34552	0.04645	0.05545	0.01071	{}	0.21635
Ridge()	0.44849	0.32504	0.04627	0.05718	0.00613	{'alpha': 26.0865362}	0.25859
Lasso()	0.43321	0.30488	0.04755	0.05889	0.00622	{'alpha': 0.008183}	0.22902
ElasticNet()	0.44887	0.34529	0.04623	0.05547	0.00582	{'l1_ratio': 0.8, 'alpha': 0.0005392}	0.22078
RandomForestRegressor(random_state=10)	0.40791	0.50988	0.04967	0.04152	0.02812	{'n_estimators': 45, 'min_samples_split': 7, 'min_samples_leaf': 5, 'max_features': 6, 'max_depth': 2, 'criterion': 'squared_error'}	0.35480
GradientBoostingRegressor()	0.46021	0.91497	0.04528	0.00720	0.01175	{'subsample': 0.75, 'random_state': 42, 'n_estimators': 100, 'max_depth': 2, 'learning_rate': 0.1}	0.53265

Pour le modèle optimale pour GHG avec ENRGYSTAR SCORE c'est le **GradientBoostingRegressor** car

- R2> RMSE
- prédiction time < ∀ modèle prédiction time
- best score(meilleurs score de validation croisée) > ∀ best score modèle

Modele pour energy avec ENERGYSTARSCORE	R² test	R² train	RMSE test	RMSE train	Prediction time	Best parameters	Best score
LinearRegression()	0.68001	0.56162	0.04193	0.07193	0.01110	{}	0.48351
Ridge()	0.68356	0.56090	0.04146	0.07205	0.00615	{'alpha': 4.9624449}	0.49507
Lasso()	0.68766	0.55170	0.04093	0.07356	0.00564	{'alpha': 0.0062057}	0.50000
ElasticNet()	0.62749	0.51450	0.04881	0.07966	0.00544	{'l1_ratio': 0.6, 'alpha': 0.0580449}	0.50116
RandomForestRegressor(random_state=10)	0.60716	0.64762	0.05147	0.05782	0.02039	{'n_estimators': 35, 'min_samples_split': 6, 'min_samples_leaf': 4, 'max_features': 8, 'max_depth': 2, 'criterion': 'absolute_error'}	0.58794
GradientBoostingRegressor()	0.72429	0.96570	0.03613	0.00563	0.00856	{'subsample': 0.5, 'random_state': 42, 'n_estimators': 50, 'max_depth': 4, 'learning_rate': 0.1}	0.68031

Pour le modèle optimale pour ENERGY avec ENRGYSTAR SCORE c'est le **GradientBoostingRegressor** car

- R2> RMSE
- prédiction time < ∀ modèle prédiction time
- best score (meilleurs score de validation croisée) > ∀ best score modèle

Les modèles de prédiction (modélisation)

Modèle sans ENERGY STAR SCORE

Modele pour GHG sans ENERGYSTARSCORE	R² test	R² train	RMSE test	RMSE train	Prediction time	Best parameters	Best score
LinearRegression()	0.39320	0.31864	0.05090	0.05773	0.00694	{}	0.18585
Ridge()	0.28252	0.21624	0.06019	0.06640	0.00570	{'alpha': 212.4845352}	0.18994
Lasso()	0.32581	0.22772	0.05656	0.06543	0.00547	{'alpha': 0.0205737}	0.22134
ElasticNet()	0.36018	0.25142	0.05367	0.06342	0.00561	{'l1_ratio': 0.2, 'alpha': 0.0554299}	0.22591
RandomForestRegressor(random_state=10)	0.43815	0.58833	0.04713	0.03488	0.02067	{'n_estimators': 35, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_features': 9, 'max_depth': 2, 'criterion': 'squared_error'}	0.35715
GradientBoostingRegressor()	0.32517	0.57423	0.05661	0.03607	0.00707	{'subsample': 0.5, 'random_state': 42, 'n_estimators': 10, 'max_depth': 2, 'learning_rate': 0.1}	0.42751

Pour le modèle optimale pour GHG sans ENRGYSTAR SCORE c'est le **GradientBoostingRegressor** car

- R2> RMSE
- prédiction time < ∀ modèle prédiction time
- best score (meilleurs score de validation croisée) > ∀ best score modèle

Modele pour ENERGY sans ENERGYSTARSCORE	R² test	R² train	RMSE test	RMSE train	Prediction time	Best parameters	Best score
LinearRegression()	0.63326	0.54608	0.04806	0.07448	0.00898	{}	0.47164
Ridge()	0.63798	0.54382	0.04744	0.07485	0.00576	{'alpha': 9.9082281}	0.48739
Lasso()	0.64363	0.53924	0.04670	0.07560	0.00560	{'alpha': 0.004599}	0.48140
ElasticNet()	0.64098	0.53324	0.04704	0.07659	0.00634	{'l1_ratio': 0.6, 'alpha': 0.0115628}	0.48567
RandomForestRegressor(random_state=10)	0.62612	0.67928	0.04899	0.05262	0.01659	{'n_estimators': 25, 'min_samples_split': 8, 'min_samples_leaf': 6, 'max_features': 7, 'max_depth': 2, 'criterion': 'squared_error'}	0.59543
GradientBoostingRegressor()	0.56589	0.98152	0.05688	0.00303	0.01482	{'subsample': 0.75, 'random_state': 42, 'n_estimators': 100, 'max_depth': 4, 'learning_rate': 0.1}	0.66131

Pour le modèle optimale pour ENERGY sans ENRGYSTAR SCORE c'est le **GradientBoostingRegressor** car

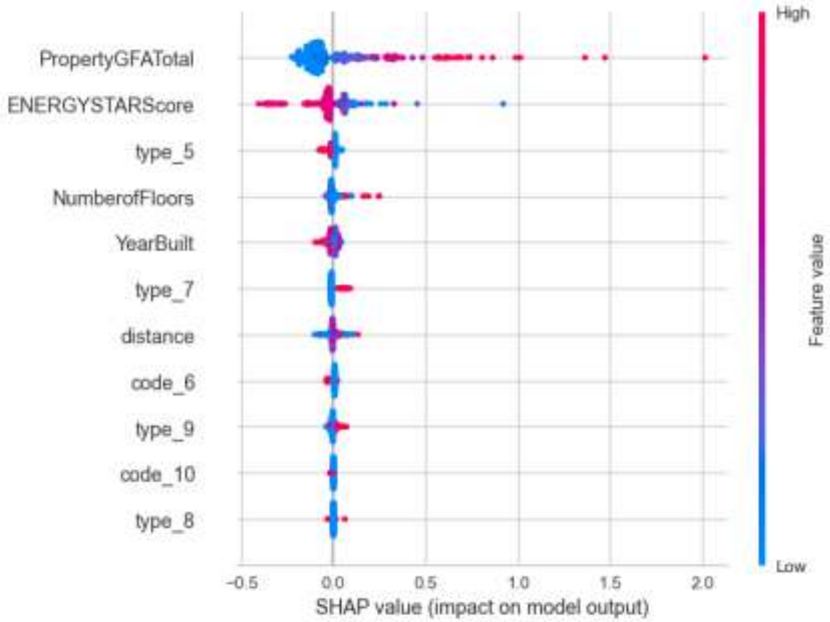
- R2> RMSE
- prédiction time < ∀ modèle prédiction time
- best score (meilleurs score de validation croisée) > ∀ best score modèle

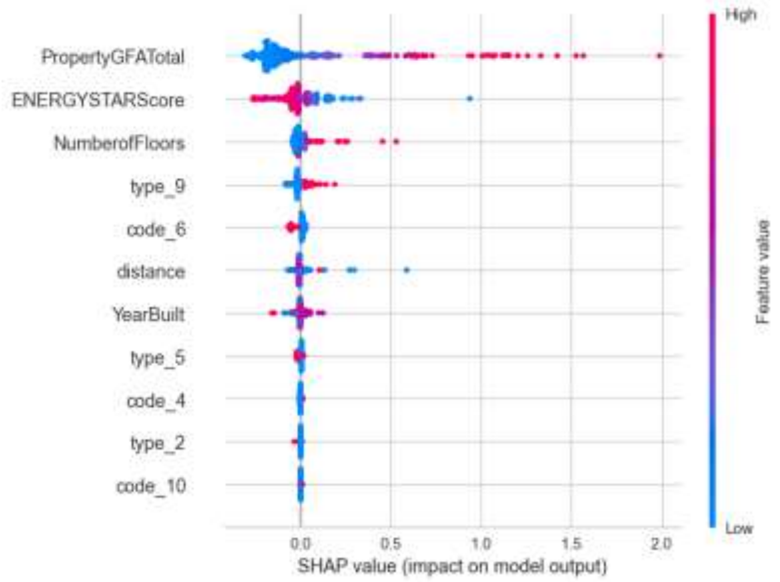
Les modèles de prédiction (modélisation, shap value)

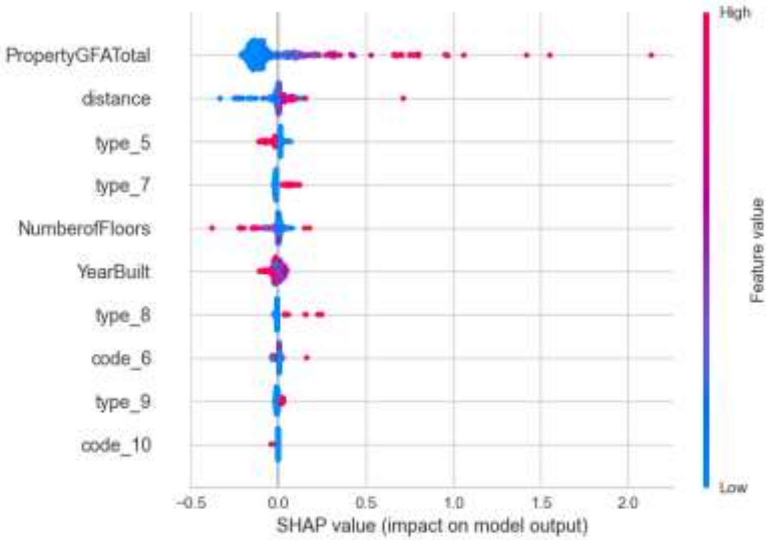
Shap Value / features importante

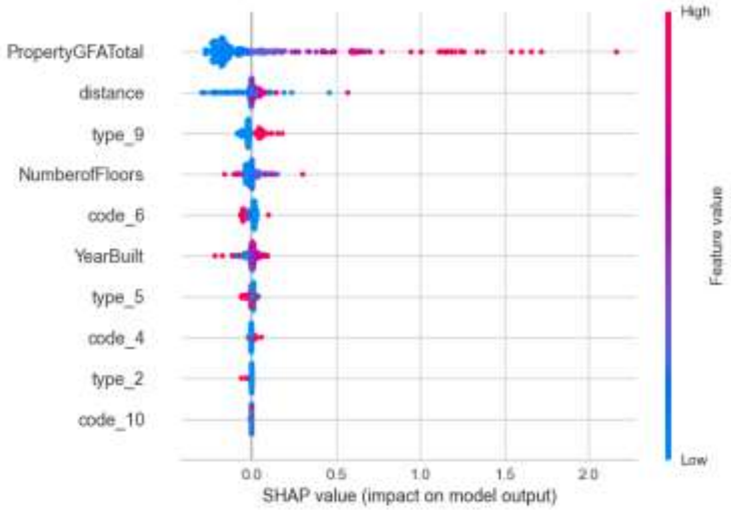
Le graphique récapitulatif SHAP montre l'effet moyen de chaque variable sur la sortie du modèle de prédiction.

Ainsi, nous pouvons voir quelles variables ont le plus d'influence sur les prédictions du modèle et comment leur influence varie en fonction des valeurs des variables.









Conclusion

Nous pouvons voir qu'après avoir fait plusieurs prétraitements testés, afin d’identifier le plus performant est avoir choisir le Recursive Features Elimination.

Après avoir tester plusieurs types de modèles de Régression :
linéaire ; Ridge; Lasso; ElasticNet; Random Forest; Gradient Boosting

On peux donc conclure qu’un modèle de boosting optimisé comme le « Gradient Boosting » est meilleur pour prédire la consommation et les émissions même si l’écart avec les validations croisées sont supérieur à 10%, cela donne un modèle sur entrainée. L’intérêts de « l’ENERGY STAR Score » sur les modèles de performance nous permettent de comprendre qu’il faudrait le prendre en compte car les meilleurs score fournit via les modèles de prédiction sont avec la prise en compte de « l’ENERGY STAR Score ».

Seattle

Seattle

Merci