ANTICIPEZ LES BESOINS EN CONSOMMATION DE BÂTIMENTS

Sommaire

Problématique

L'équipe s’intéresse à la consommation et aux émissions des bâtiments non destinés à l’habitation.

L'équipe va tenter de prédire les émissions de CO2 et la consommation totale d’énergie des bâtiments non destinés à l’habitation

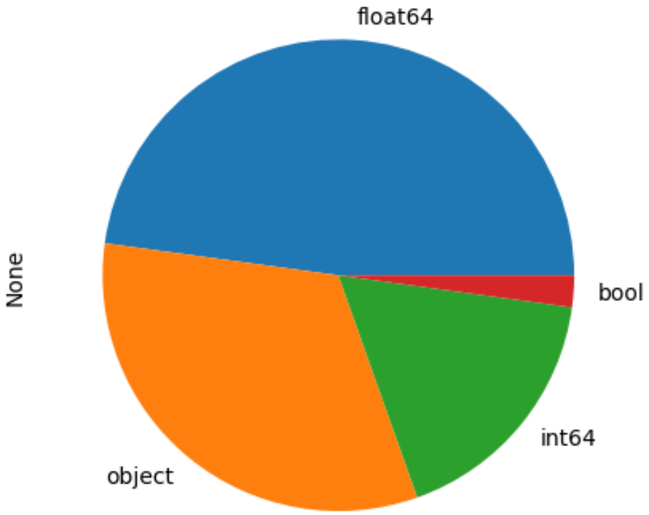
On veut évaluer l’intérêt de l’ENERGY STAR Score pour la prédiction d’émissions

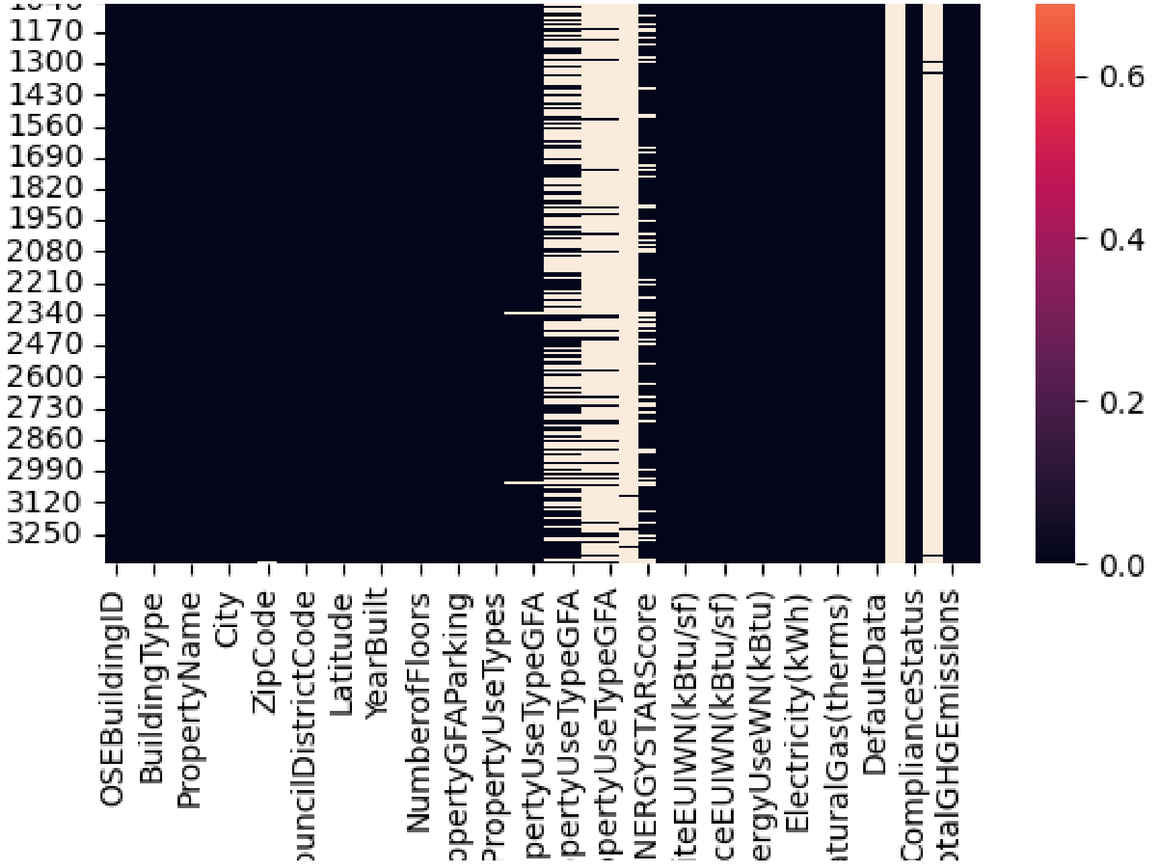
Les données

'2016_Building_Energy_Benchmarking.csv'

Présentation du dataset

- Dimensions (3376 lignes , 46 colonnes)
- Valeurs manquantes: 19952 au total





NÉTTOYAGE DES DONNÉES

features ingeneering

- Sélection variables pertinentes pour répondre à notre objectif
- Création des features : 'BuildingAge', "Surface" et 'energytype_count'

---

les outliers

- Suppression des valeurs negatives, et des batiments ayant la valeur 0,
- Methode k-Nearest Neighbors pour les traitements des outliers
- Regroupement des catégories d'usage des bâtiments

Les valeurs manquantes

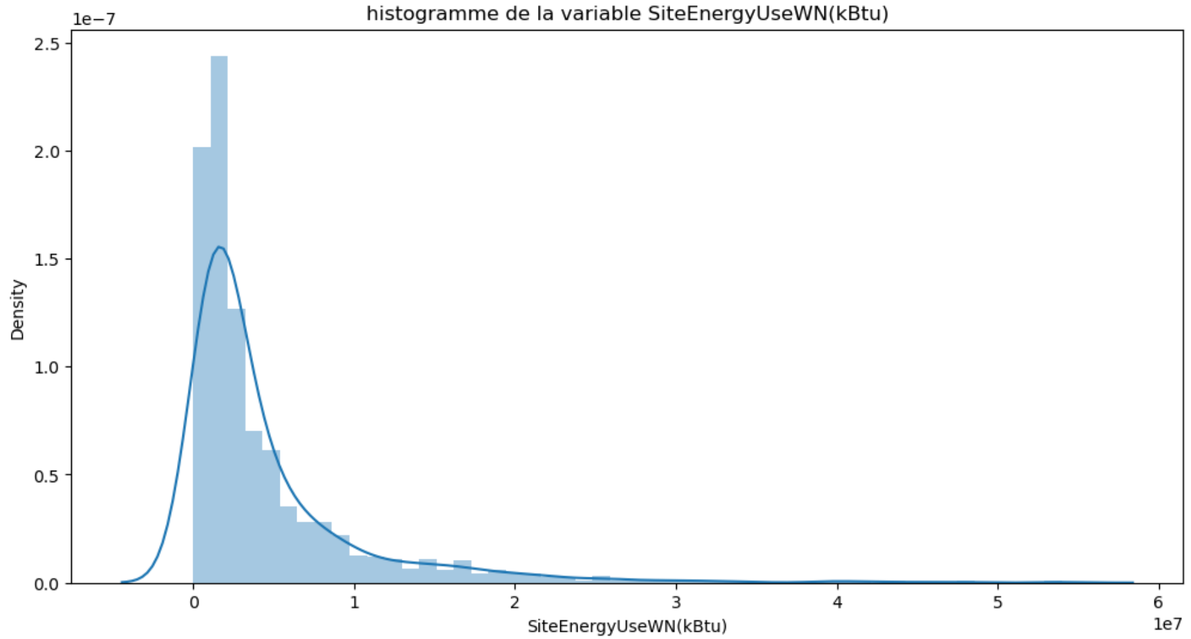
- Imputation des Nan par la médiane du type de bâtiment
- Imputation des Nan des variables catégorielles par 'Unknow'

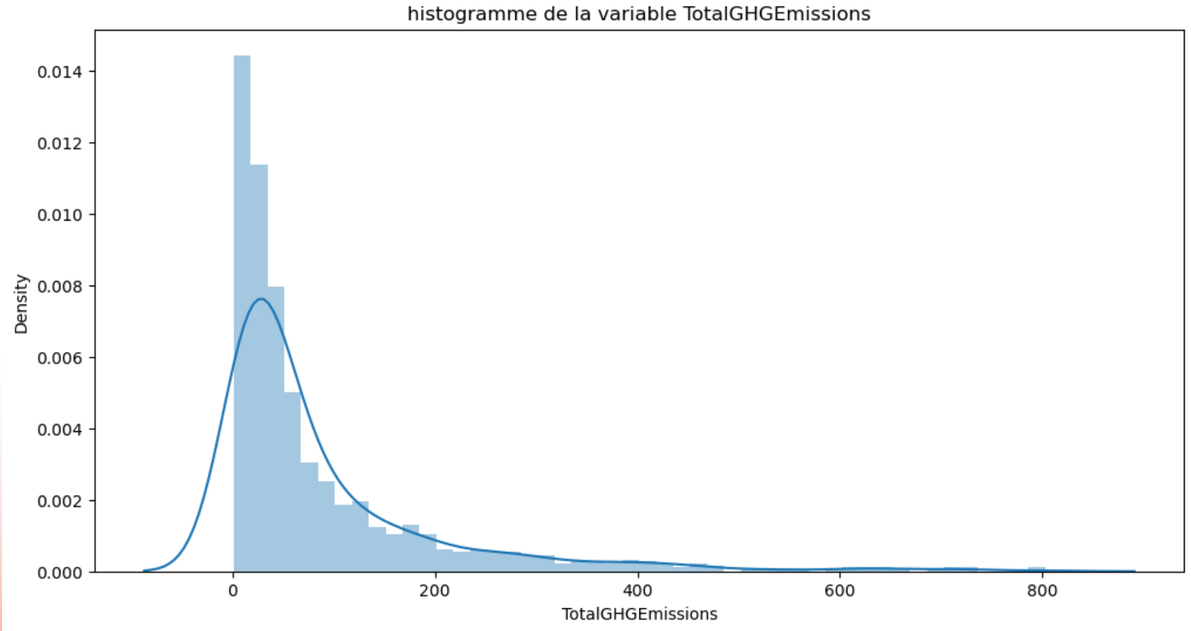
T **Targets**

- SiteEnergyUseWN(kBtu) : consommation normalisée « Weather Normalized »
- 'TotalGHGEmissions': les émissions de CO2

Analyse univariée

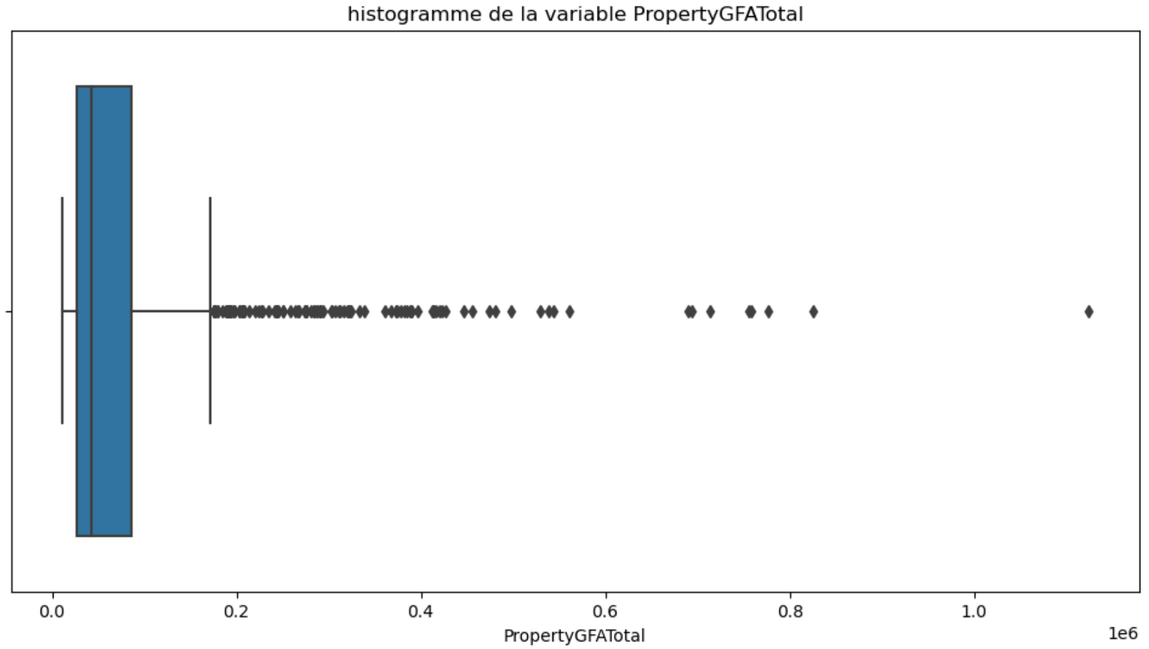
Distribution des variables targets

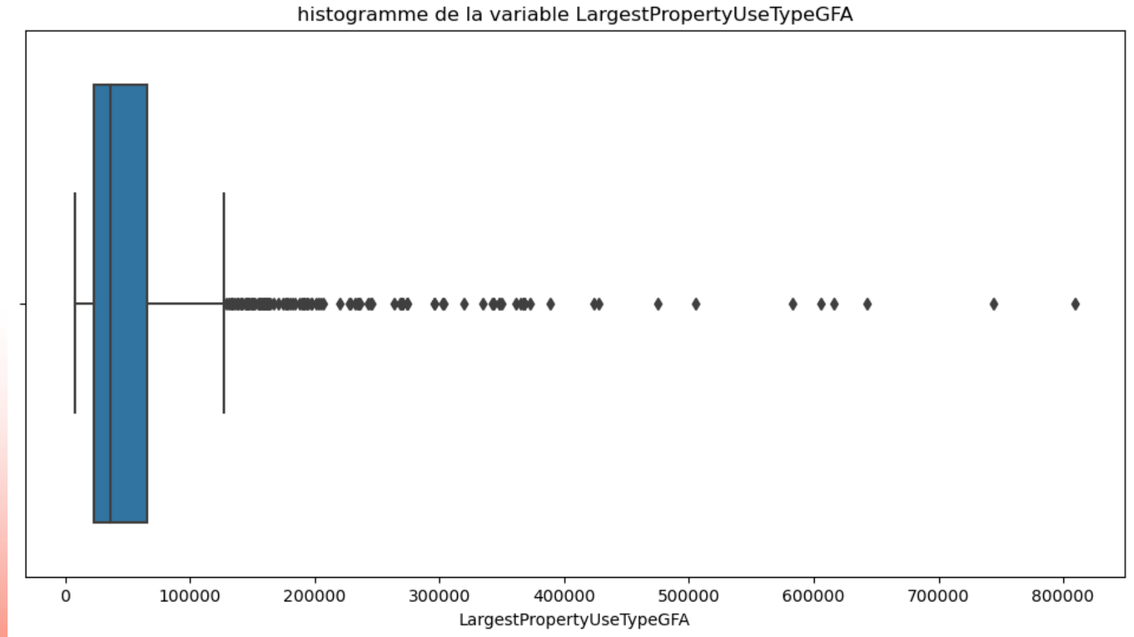




Analyse univariée

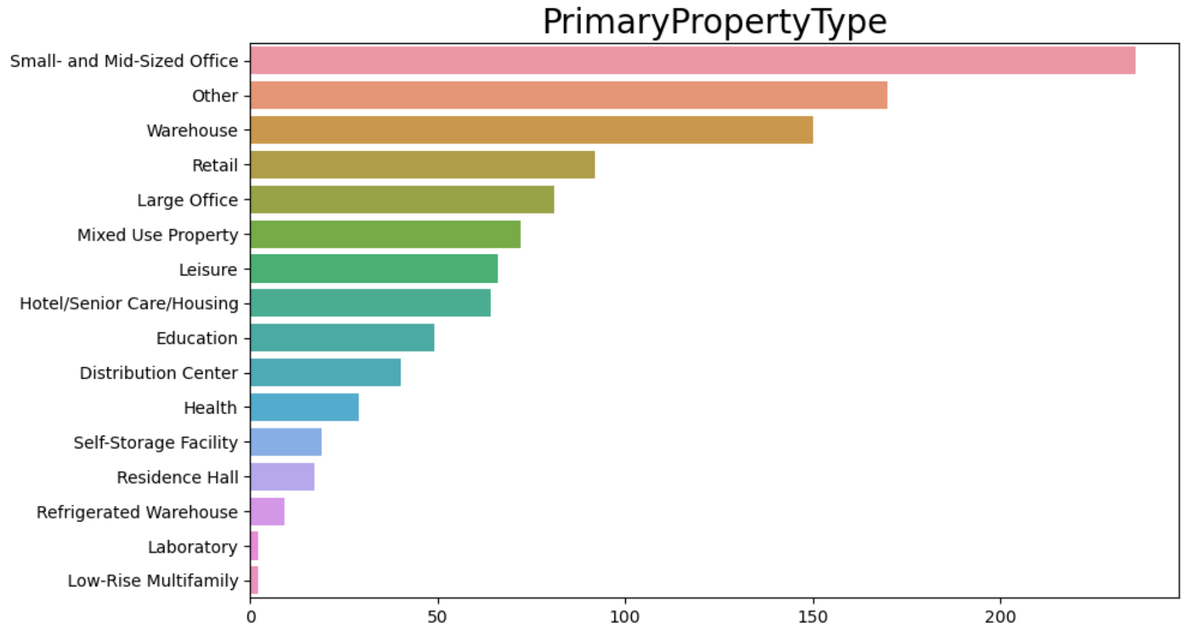
Boxplot des variables continues

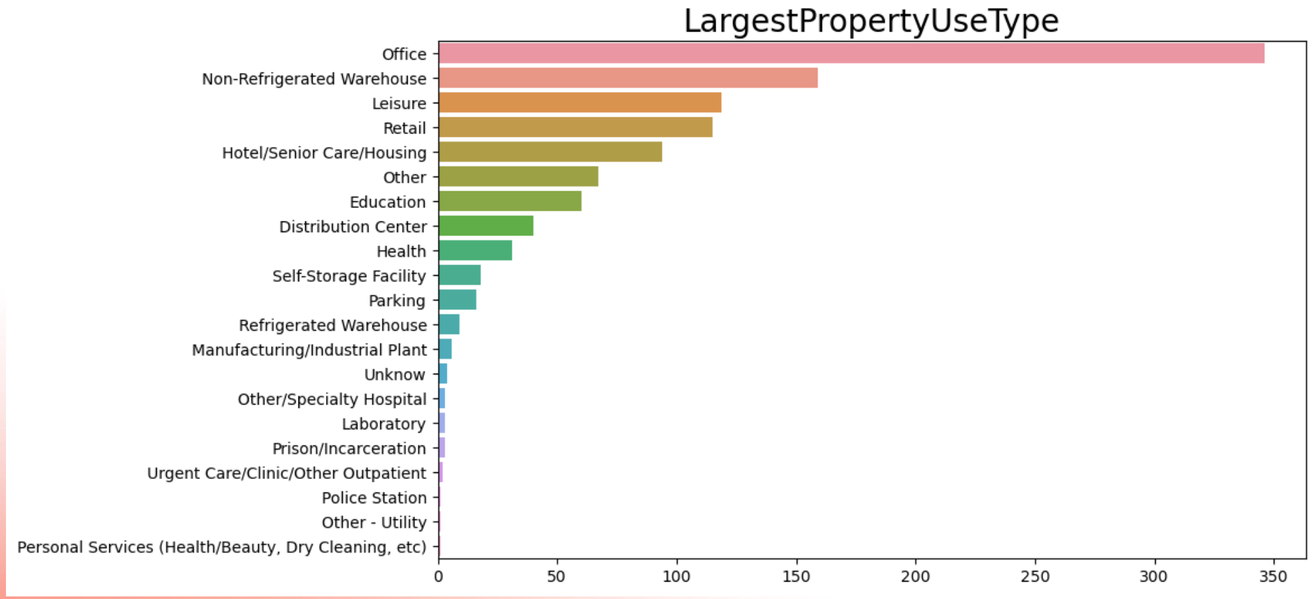




Analyse univariée

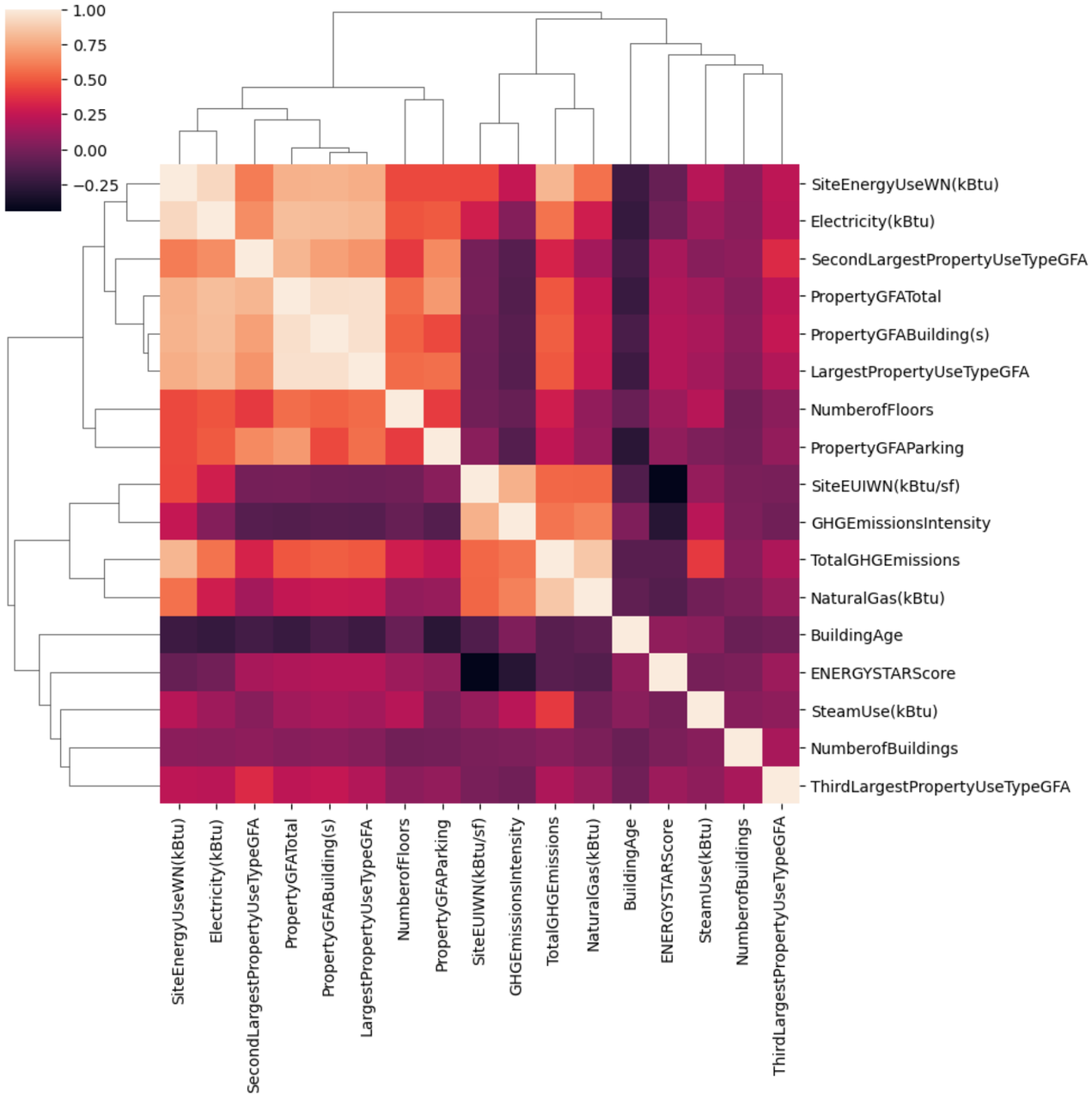
Distribution de variables catégorielles





Analyse Bivariée

La matrice de correlation



Modélisation

SÉPARATION DU JEU DE DONNÉES

Model selection: KFold, Train, Test

PREPROCESSING

StandardScaler, OneHotEncoder, Normalisation

BASELINE

Entrainement de différents modèles

HYPERPARAMETRES TUNING

GridSearchCV

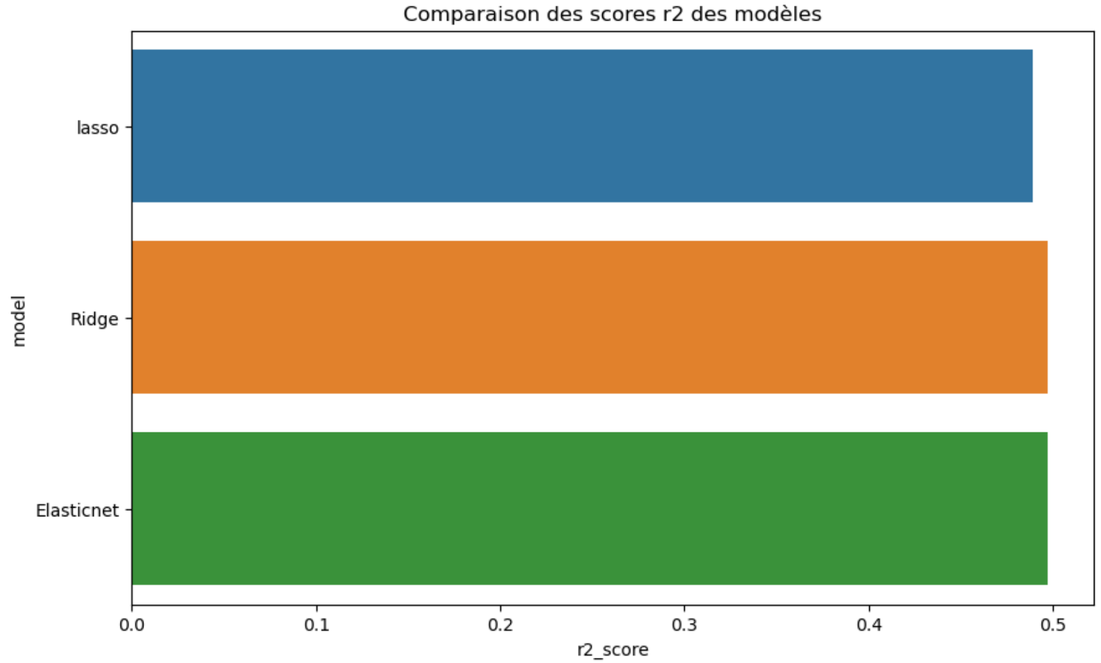
COMPARAISON

- Erreur quadratique moyenne : RMSE
- Le coefficient de determination : *R2*

Modélisation target 'SiteEnergyUseWN(kBtu)'

Preprocessing

- StandardScaler
- OneHotEncoder



Modèles linéaires

	model	Score_RMSE	r2_score
0	lasso	0.698524	0.488958
1	Ridge	0.692582	0.497616
2	Elasticnet	0.692895	0.497162

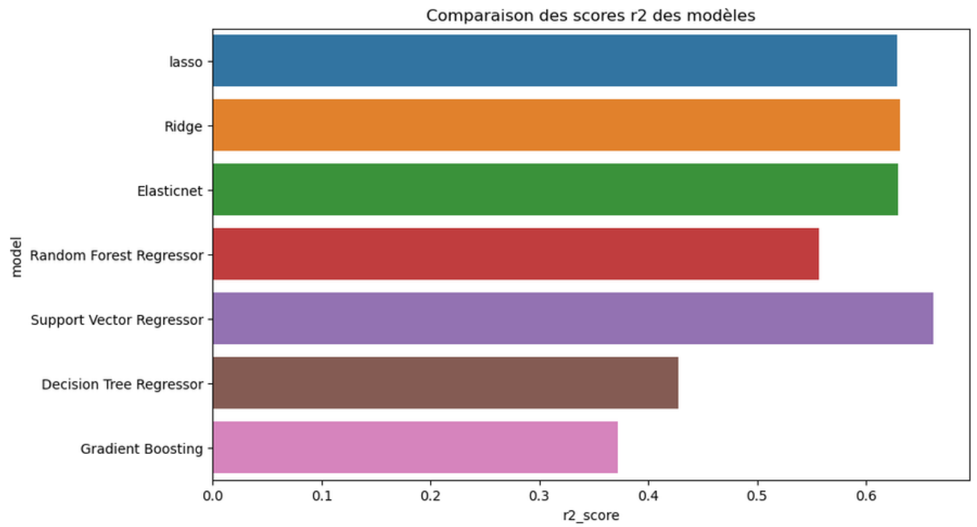
Modélisation target 'SiteEnergyUseWN(kBtu)'

Amélioration du feature engineering:

- Choix de nouvelles variables
- Transformation en log
- OneHotEncoder

```
features= ['BuildingAge',
    'Surface',
 'energytype_count',
'PrimaryPropertyType',
'LargestPropertyUseType',
 'LargestPropertyUseTypeGFA',
 'SecondLargestPropertyUseType',
 'SecondLargestPropertyUseTypeGFA',
 'ThirdLargestPropertyUseType',
 'ThirdLargestPropertyUseTypeGFA',
 'SiteEnergyUseWN(kBtu)']
```

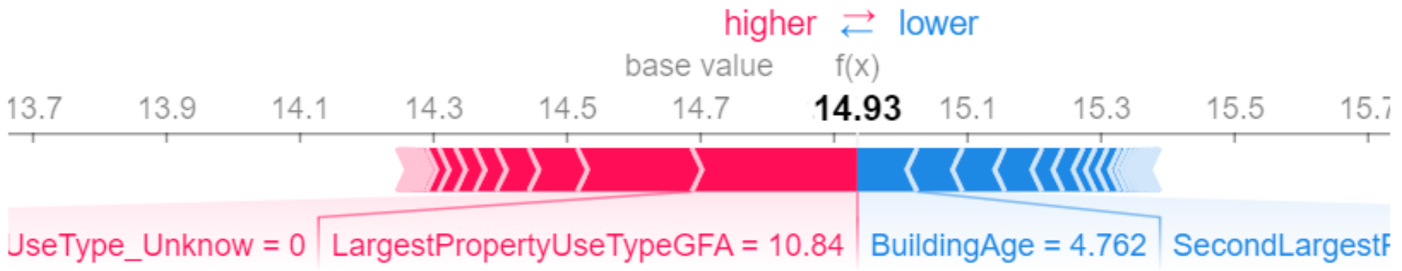
	model	Score_RMSE	r2_score
0	lasso	0.595613	0.628446
1	Ridge	0.593442	0.631149
2	Elasticnet	0.594851	0.629395
3	Random Forest Regressor	0.650602	0.556673
4	Support Vector Regressor	0.567731	0.662418
5	Decision Tree Regressor	0.739274	0.427593
6	Gradient Boosting	0.774070	0.372441

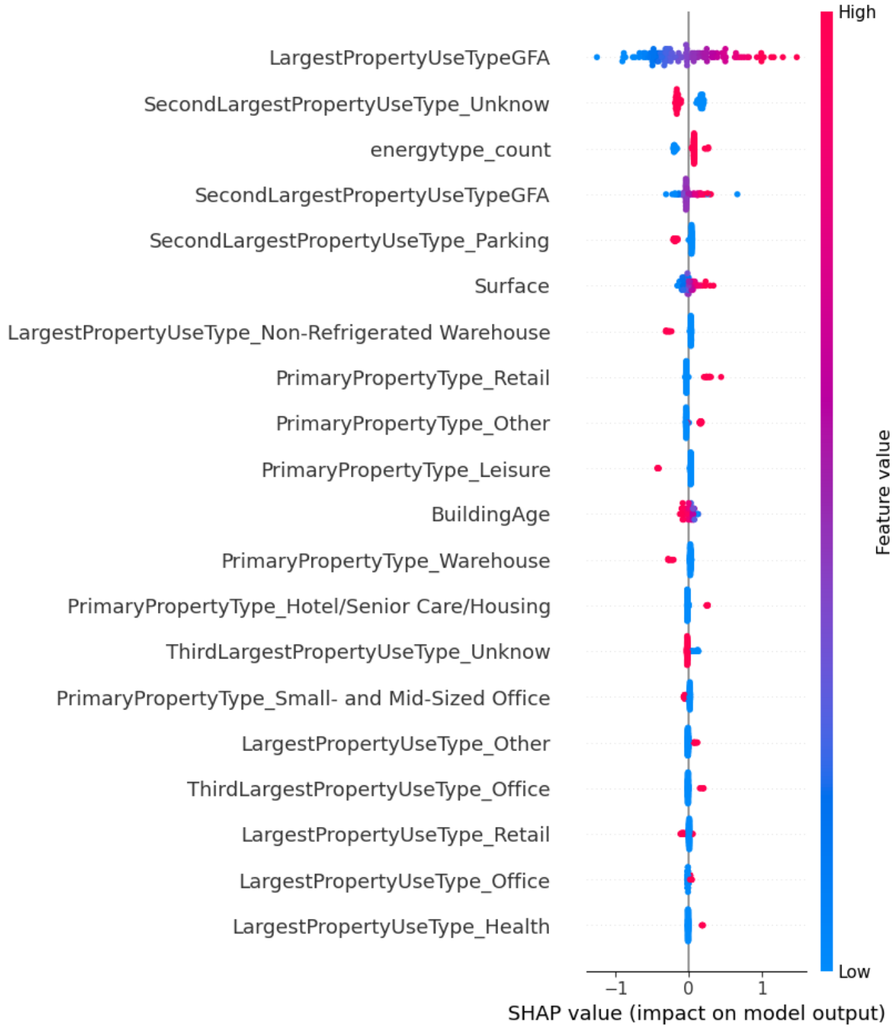


Modélisation target 'SiteEnergyUseWN(kBtu)'

Feature importance SVR

```
i = 4
shap.force_plot(explainer.expected_value, shap_values[i], features=X_te
```





Modélisation target 'SiteEnergyUseWN(kBtu)'

L'influence de l'EnergyStarScore

```
#Modélisation
model_svr_en = SVR(C=10, epsilon=0.1, gamma=0.01)
model_svr_en.fit(X_train_en, y_train_en)
print(model_svr_en.score(X_test_en, y_test_en))
```

```
0.7826089809626751
```

```
#Modélisation sans 'ENERGYSTARScore'
model_svr_en = SVR(C=10, epsilon=0.1, gamma=0.01)
model_svr_en.fit(X_train_en, y_train_en)
print(model_svr_en.score(X_test_en, y_test_en))
```

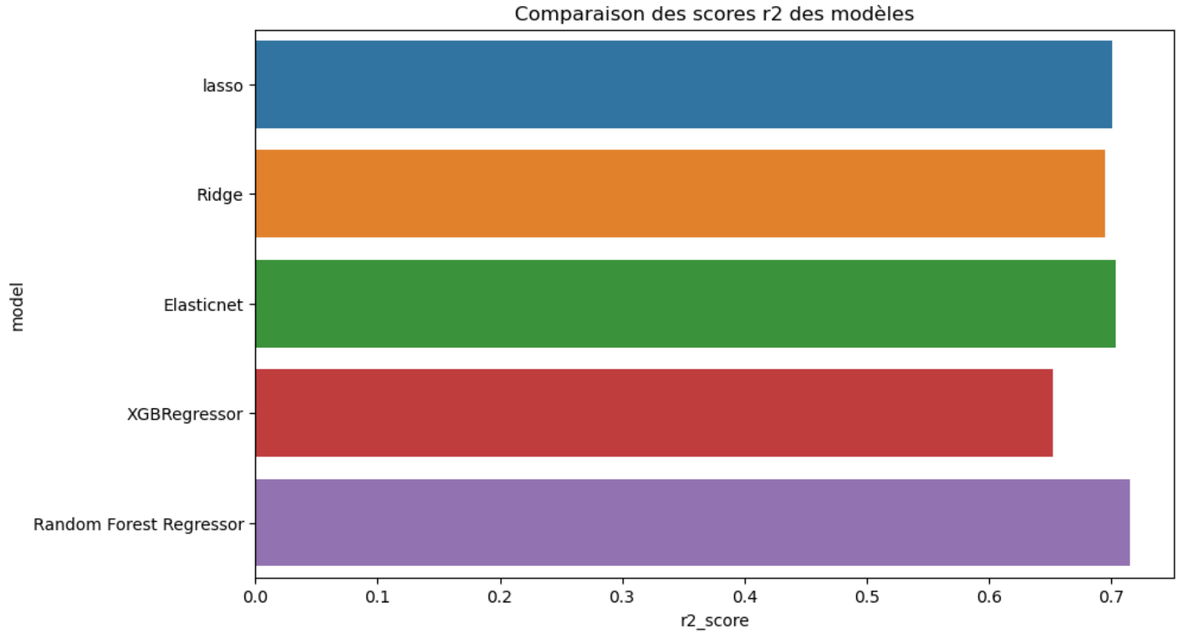
```
0.6992541688535057
```

Modélisation target 'TotalGHGEmissions'

Preprocessing

- Transformation en log
- OneHotEncoder

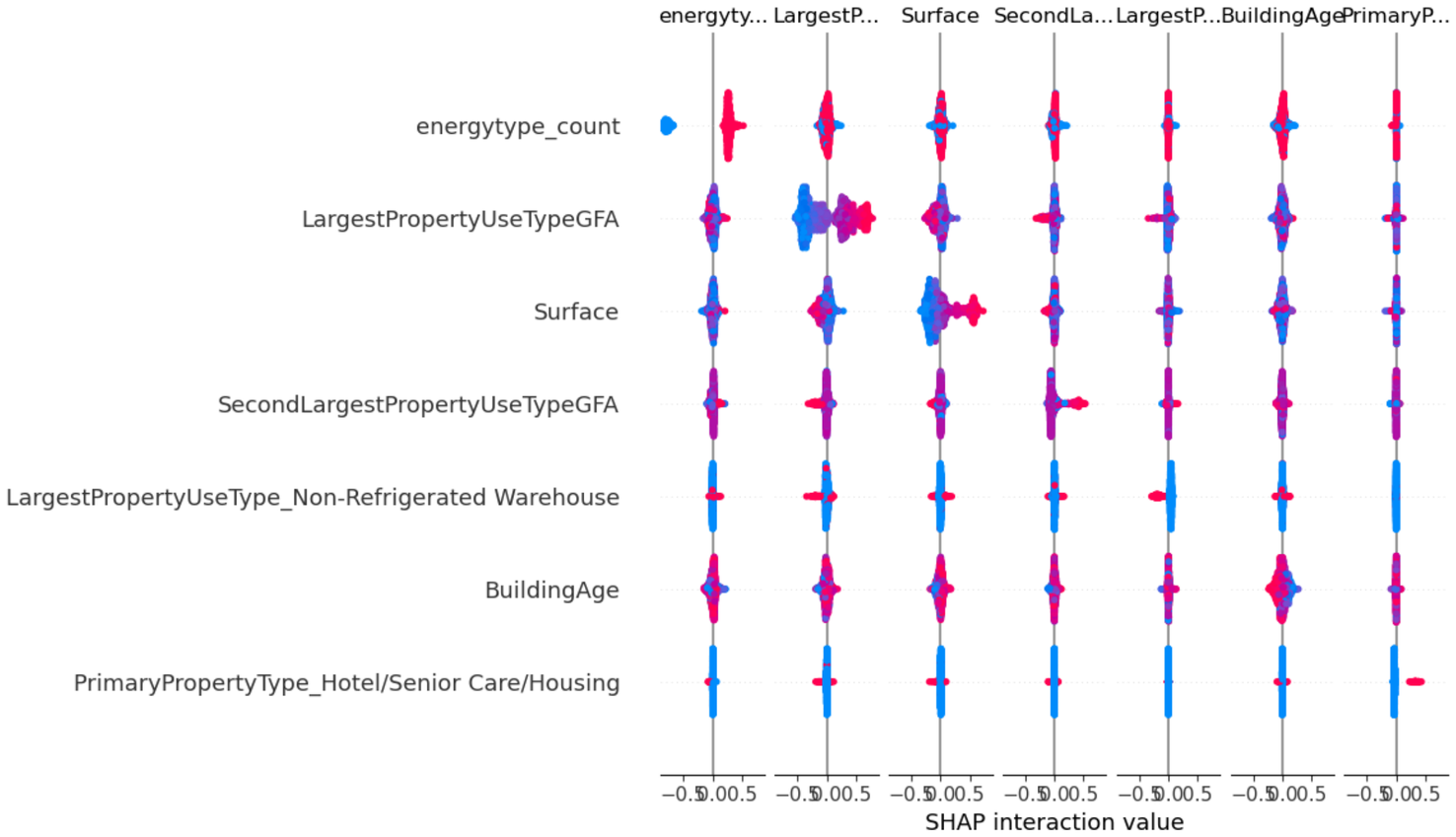
```
features= ['energytype_count',
    'Surface',
    'BuildingAge',
    'PrimaryPropertyType',
    'LargestPropertyUseType',
 'LargestPropertyUseTypeGFA',
 'SecondLargestPropertyUseType',
 'SecondLargestPropertyUseTypeGFA',
 'ThirdLargestPropertyUseType',
 'ThirdLargestPropertyUseTypeGFA',
    'TotalGHGEmissions']
```

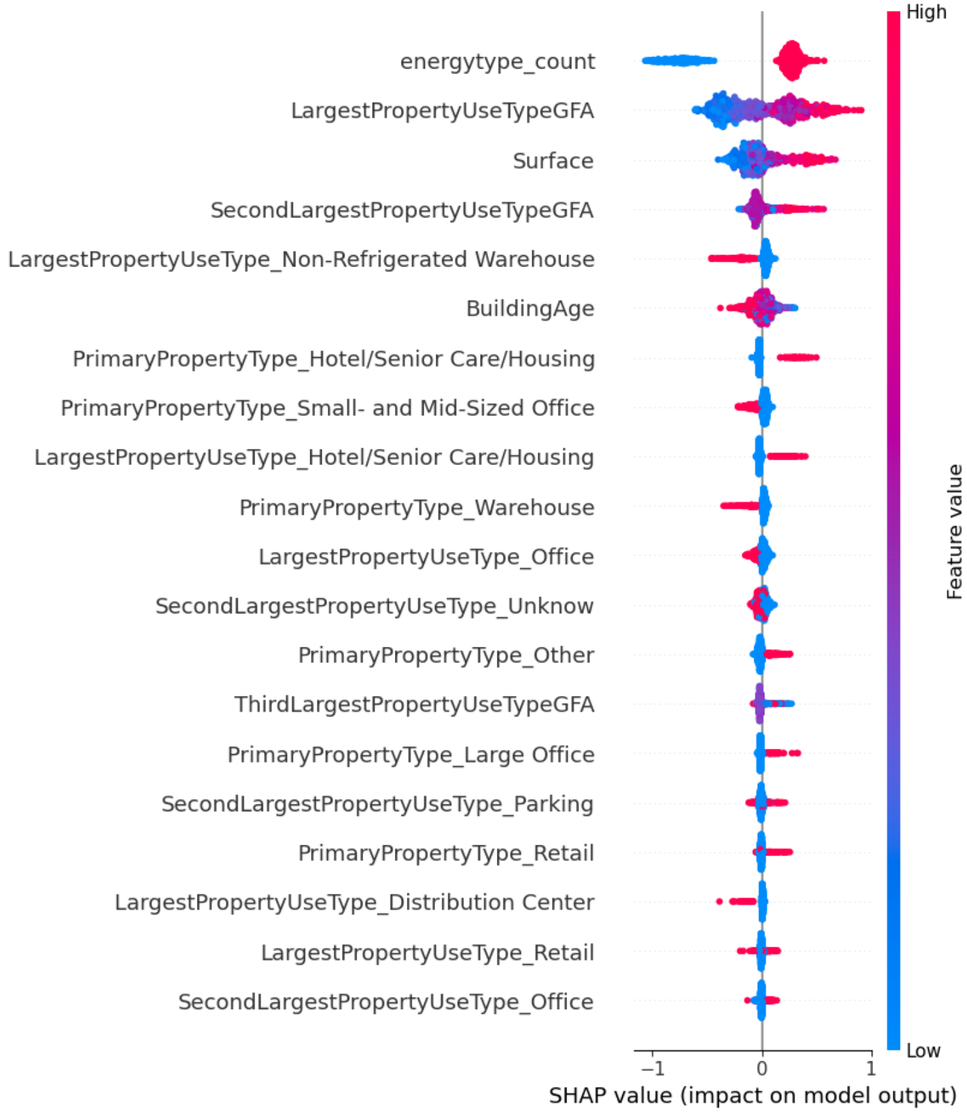


	model	Score_RMSE	r2_score
0	lasso	0.672705	0.701439
1	Ridge	0.680004	0.694925
2	Elasticnet	0.669923	0.703903
3	XGBRegressor	0.725891	0.652362
4	Random Forest Regressor	0.656081	0.716013

Modélisation target 'TotalGHGEmissions'

Feature importance RFR





Modélisation target 'TotalGHGEmissions'

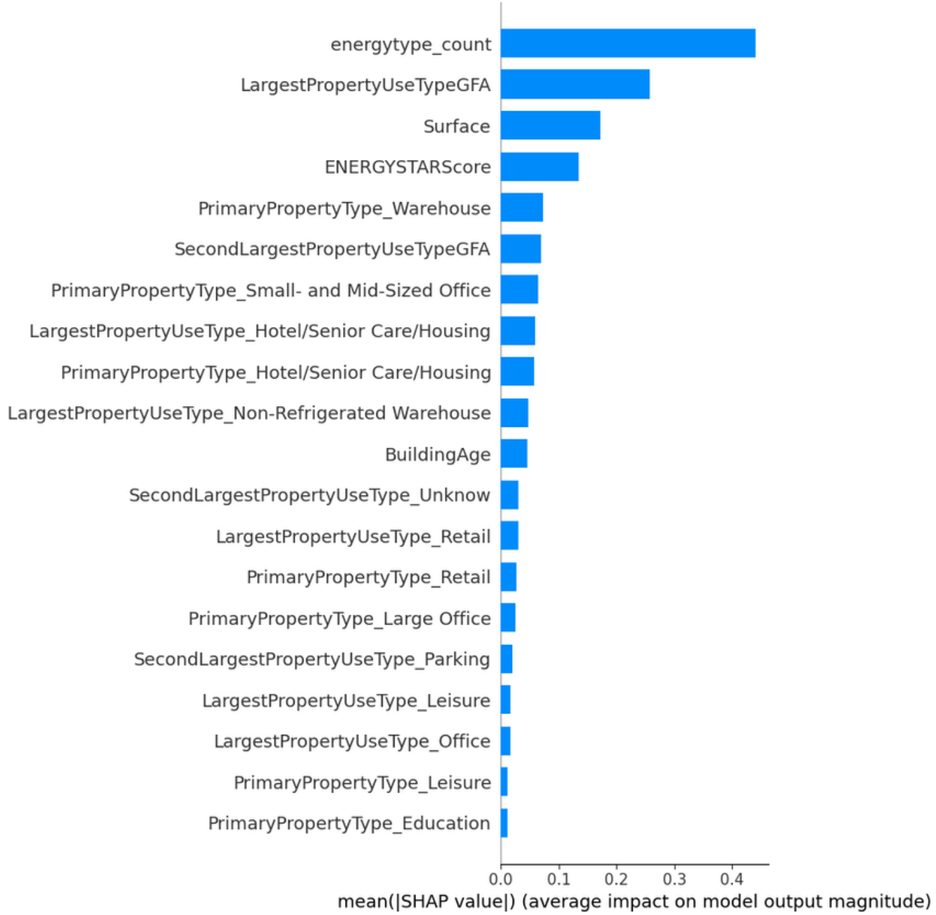
L’influence de l’EnergyStarScore

```
model_rdr = RandomForestRegressor(max_features='sqrt',min_samples_leaf=
model_rdr.fit(X_train_en, y_train_en)
print("Le score de prédiction avec 'ENERGYSTARScore' est de " ,model_rd
```

```
Le score de prédiction avec 'ENERGYSTARScore' est de  0.6695335567157
392
```

```
model_rdr = RandomForestRegressor(max_features='sqrt',min_samples_leaf=
model_rdr.fit(X_train_en, y_train_en)
print("Le score de prédiction sans 'ENERGYSTARScore' est de ",model_rdr
```

```
Le score de prédiction sans 'ENERGYSTARScore' est de  0.5615848168874
576
```



CONCLUSION

Sur les prédictions de la consommation d'énergie, les résultats sont décevants. Ce qui est dû aux données.

ENERGY STAR Score :

- Les prédictions sont meilleures avec la variable,
- la variable comporte plusieurs données manquantes, ce qui limite son utilisation

MERCI !