

# Anticiper les besoins en consommation de bâtiments

4

*Essai de prédiction : émissions CO2  
et consommation totale d'énergie*

Parcours Data Scientist.

*Présentation : Dai TENSAOUT*

Plan

Contexte

État des lieux de la base de données

Filtrage et Analyses

Feature engineering

Modélisation

Conclusion



# Contexte



**Seattle**

- 1- Objectif de la ville de devenir neutre en émissions de carbone.
- 2 -Émissions des bâtiments non résidentiels.
- 3- Utilisation des données structurelles des bâtiments et se passer des relevés de consommation.
- 4- Évaluer l'intérêt de l'ENERGY STAR Score



# Etat des lieux de la base de données

Taille data brut:	3376 Lignes 46 Colonnes
Taux de remplissage:	23,78 %
Doublon:	0
Colonnes vides	1

Taux de NaN  
supérieur à 50 %



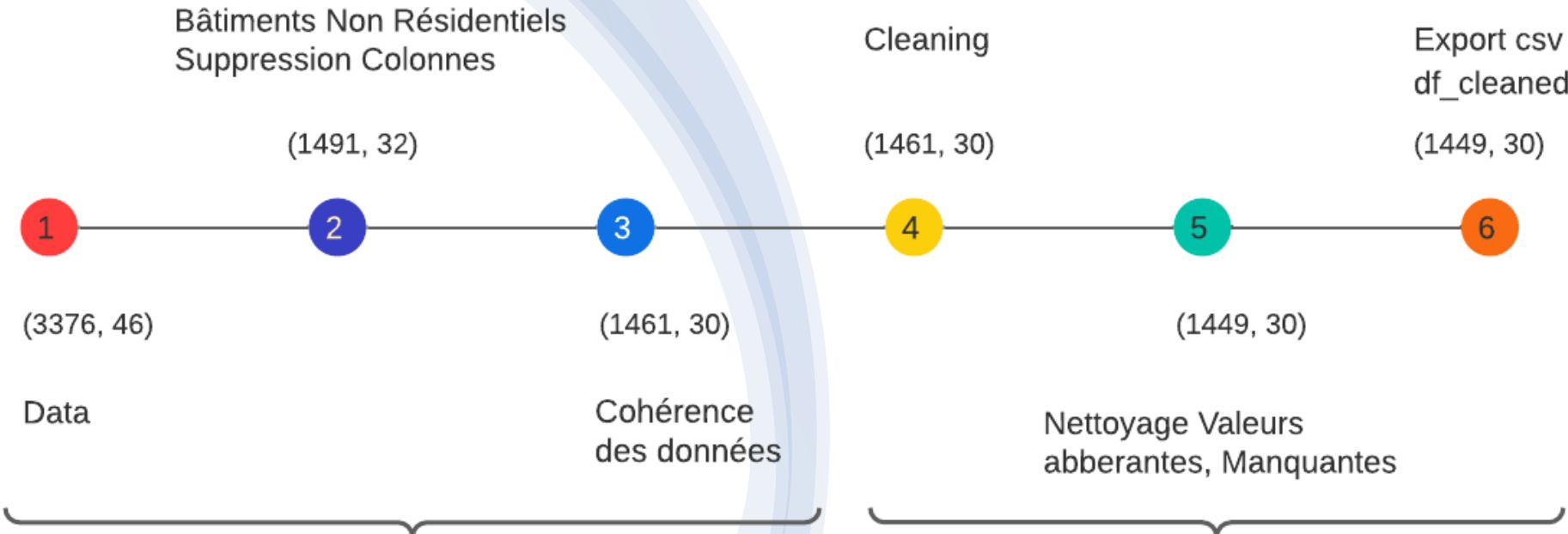
- Géographiques : adresse, position, quartier...
- Types d'usage : bureaux, commerces, santé...
- Données structurelles : superficies, nombre d'étages...
- Profil de consommation : électricité, gaz ou vapeur.
- Quantité de CO2émise.

Source de données

Définitions des features

Taux de NaN  
Inférieur à 0.6 %

# Filtrage et analyses

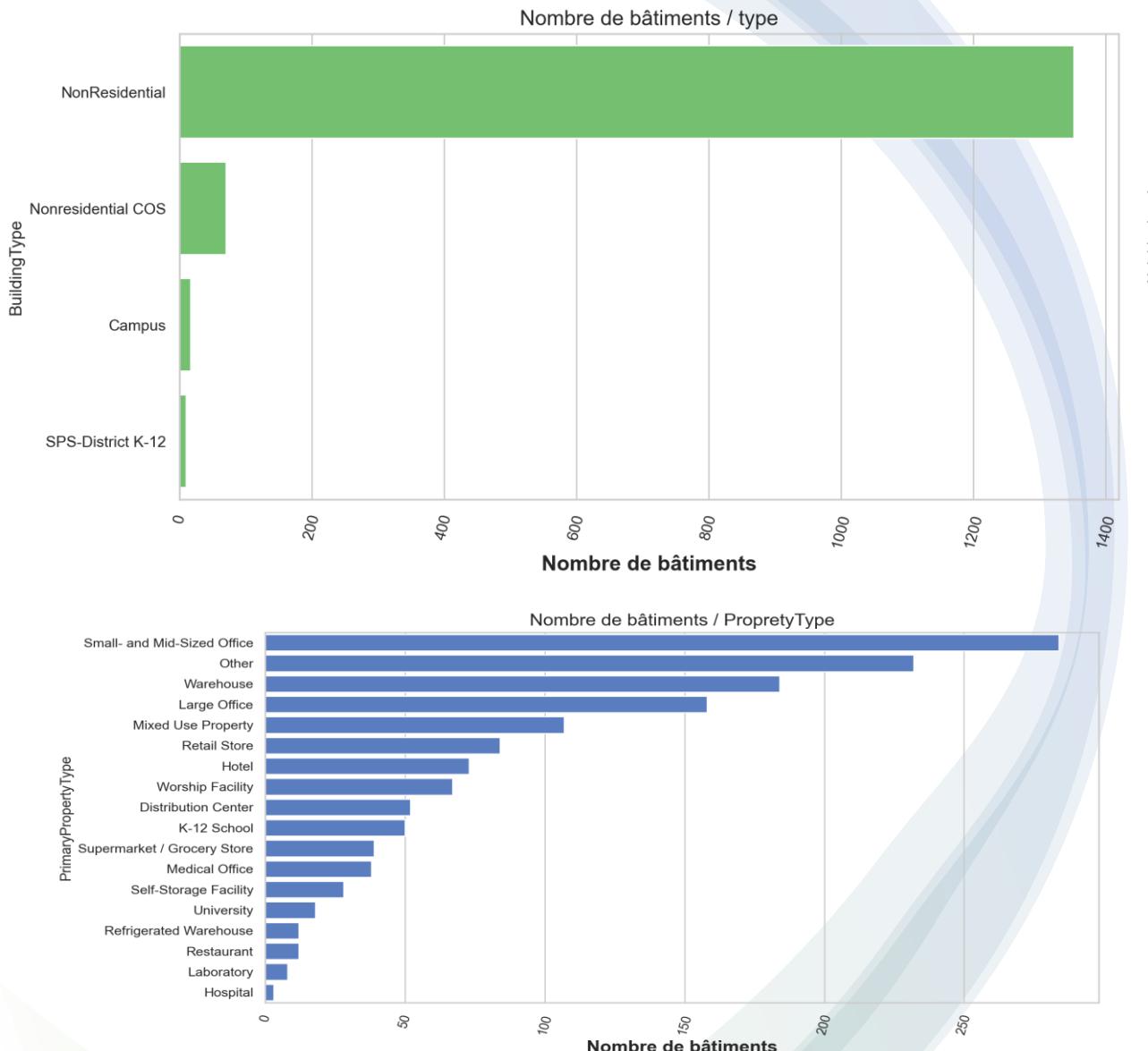


- Traitements des valeurs abberantes.
- Imputation valeurs manquantes.



# EDA

## distribution des bâtiments

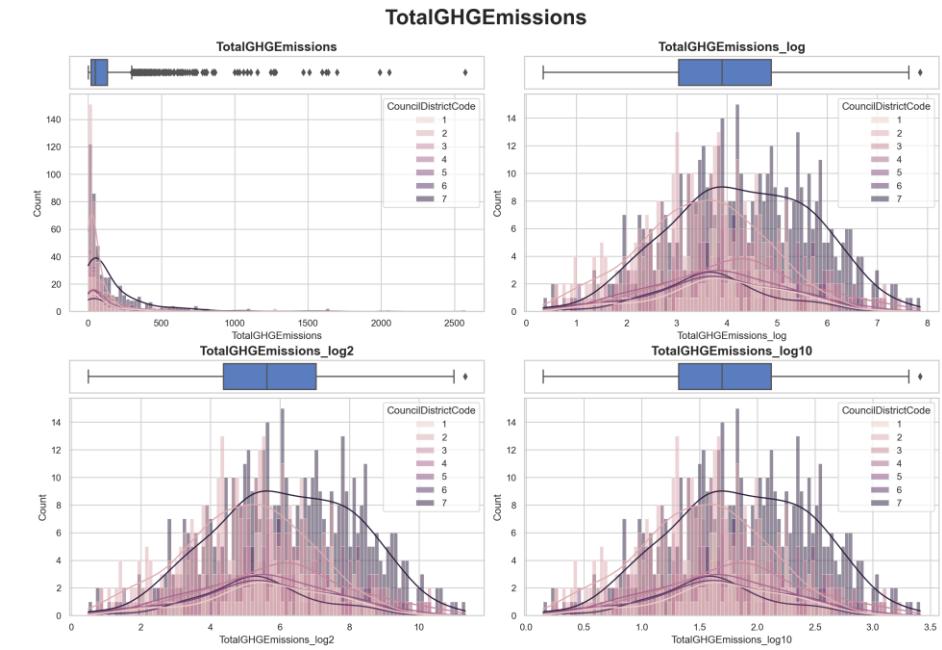
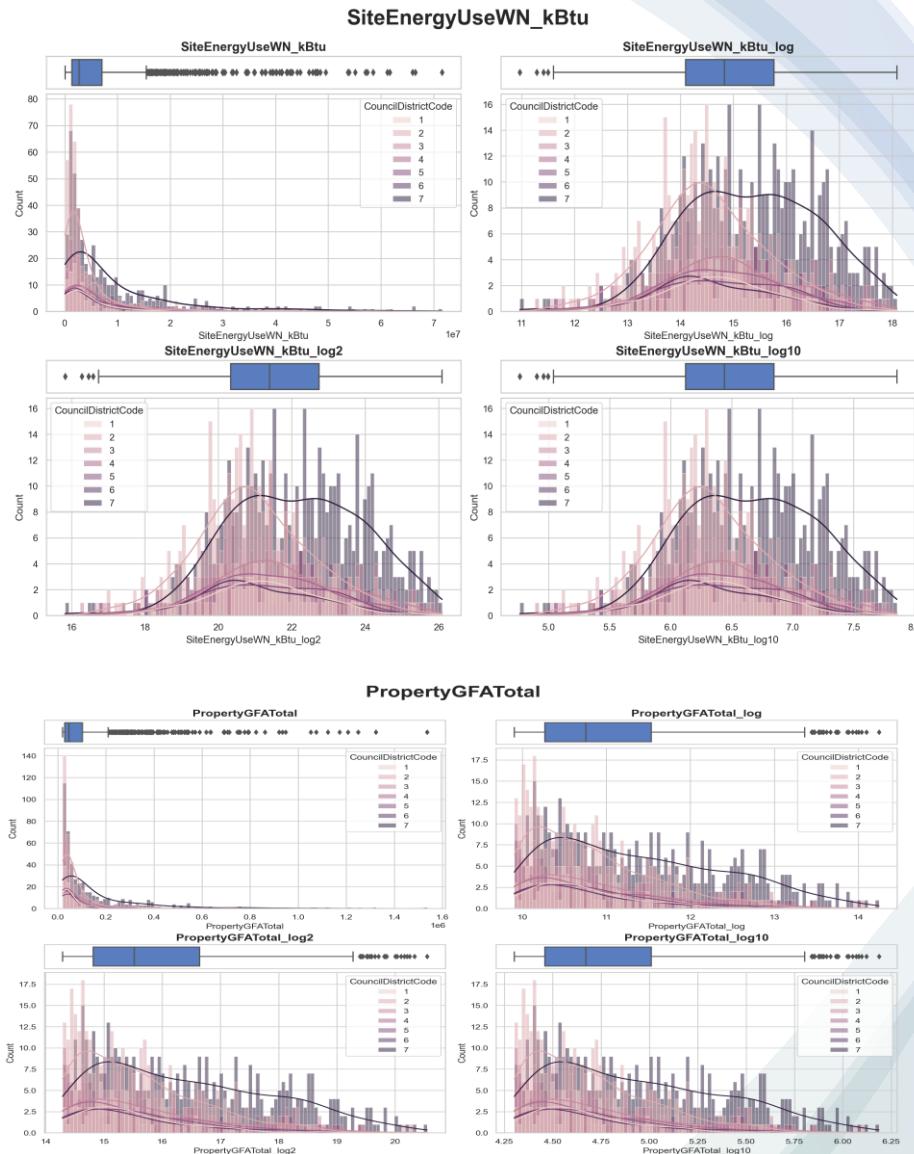


La *majeur* partie des bâtiments sont de type NonResidential.  
Les *différentes catégories* de bâtiments sont réparties d'une façon *non homogènes*.  
La *catégorie la plus représentée* est « *bureau de petite et moyenne taille* »



# EDA

## distribution Target et gfa\_total

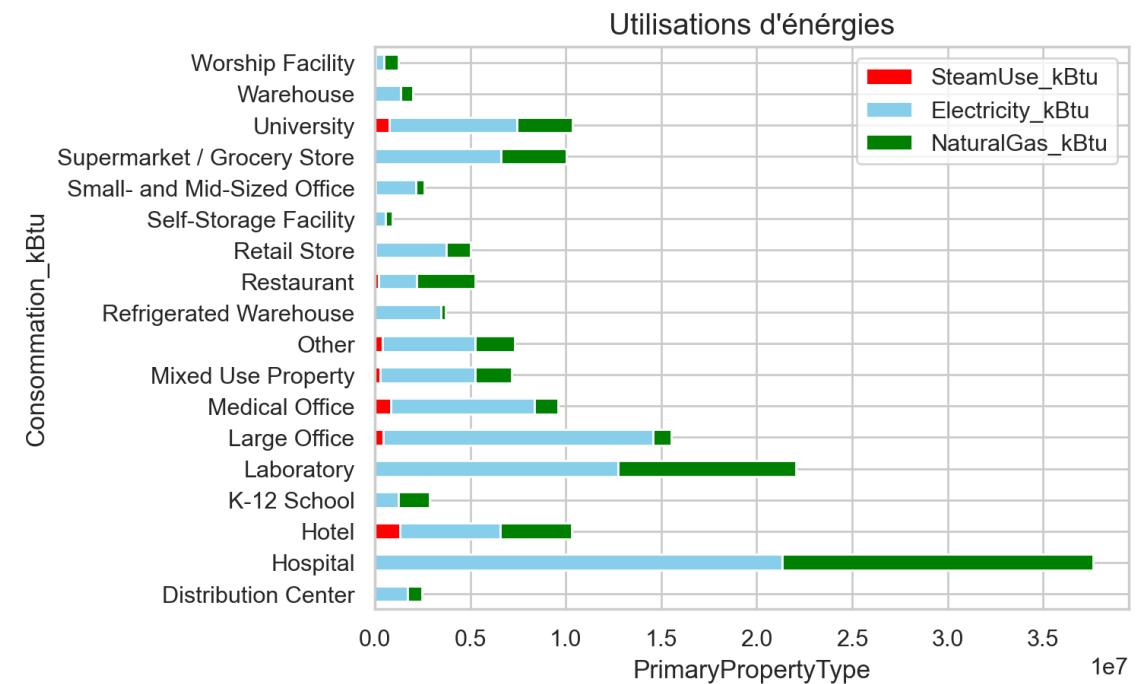
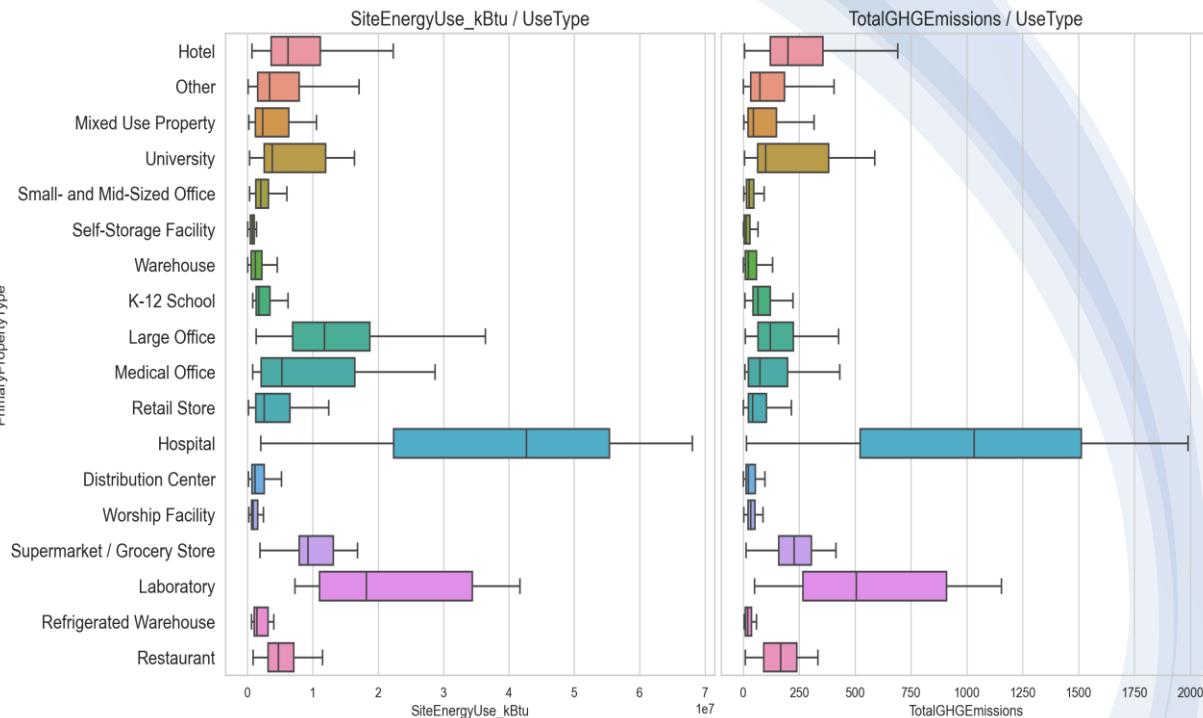


Les distributions des deux Target « SiteEnergyUse » et « TotalGHGEmissions » ne sont pas normalement distribuées, une transformation est utile dans notre cas.  
Même constat pour la colonne « PropretyGFATotal »



# EDA

# consommation et émission



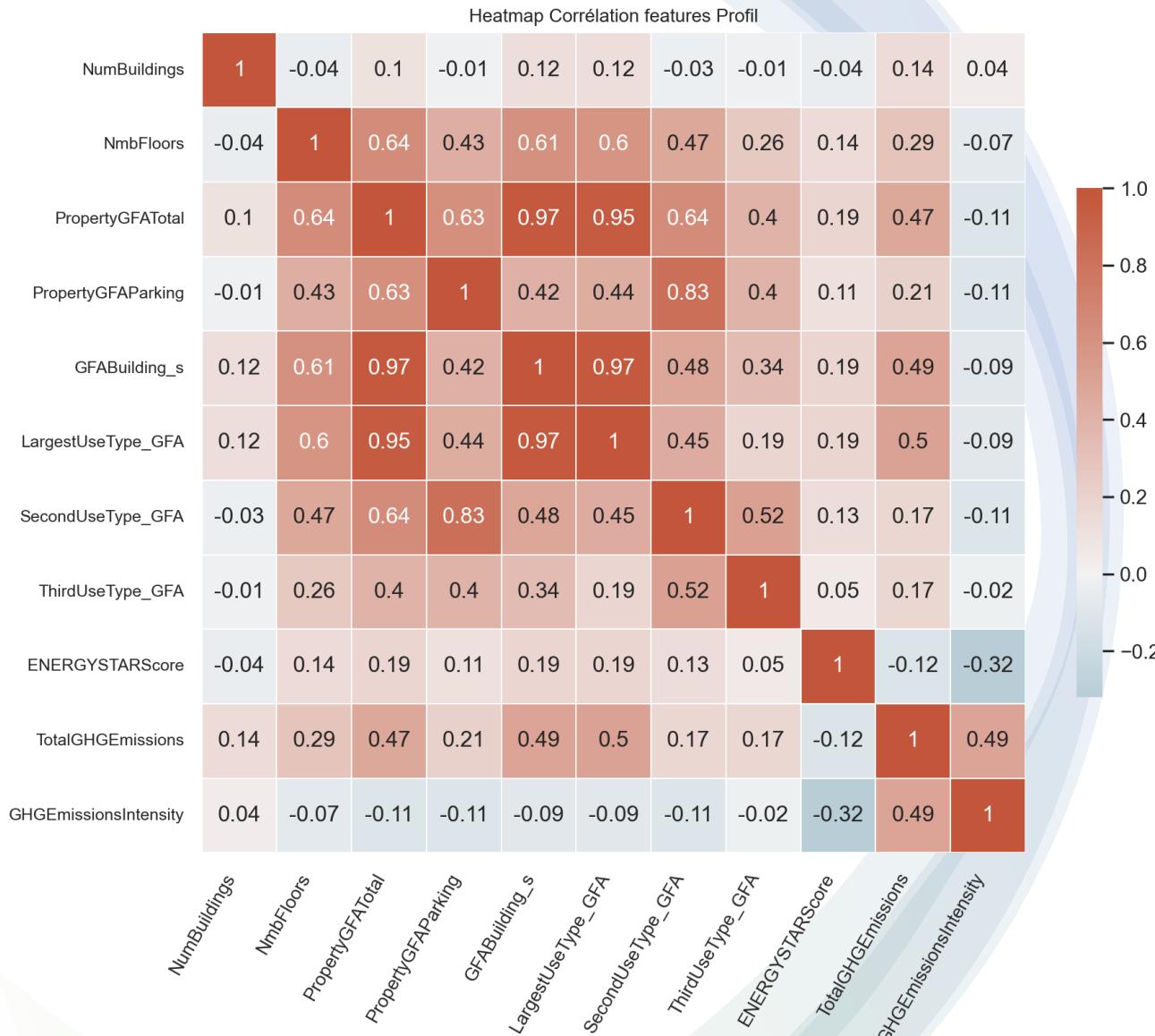
Les Hôpitaux : bâtiments qui consomment et rejettent le plus de CO<sub>2</sub>.

Types d'énergies : la distribution de la consommation d'énergie selon le type d'énergie n'est pas homogène dans chacun des types de bâtiments, l'électricité est la source d'énergie la plus utilisée.

L'utilisation de la vapeur est négligeable par rapport aux deux autres sources l'électricité et le gaz naturel.



## corrélations vs Target



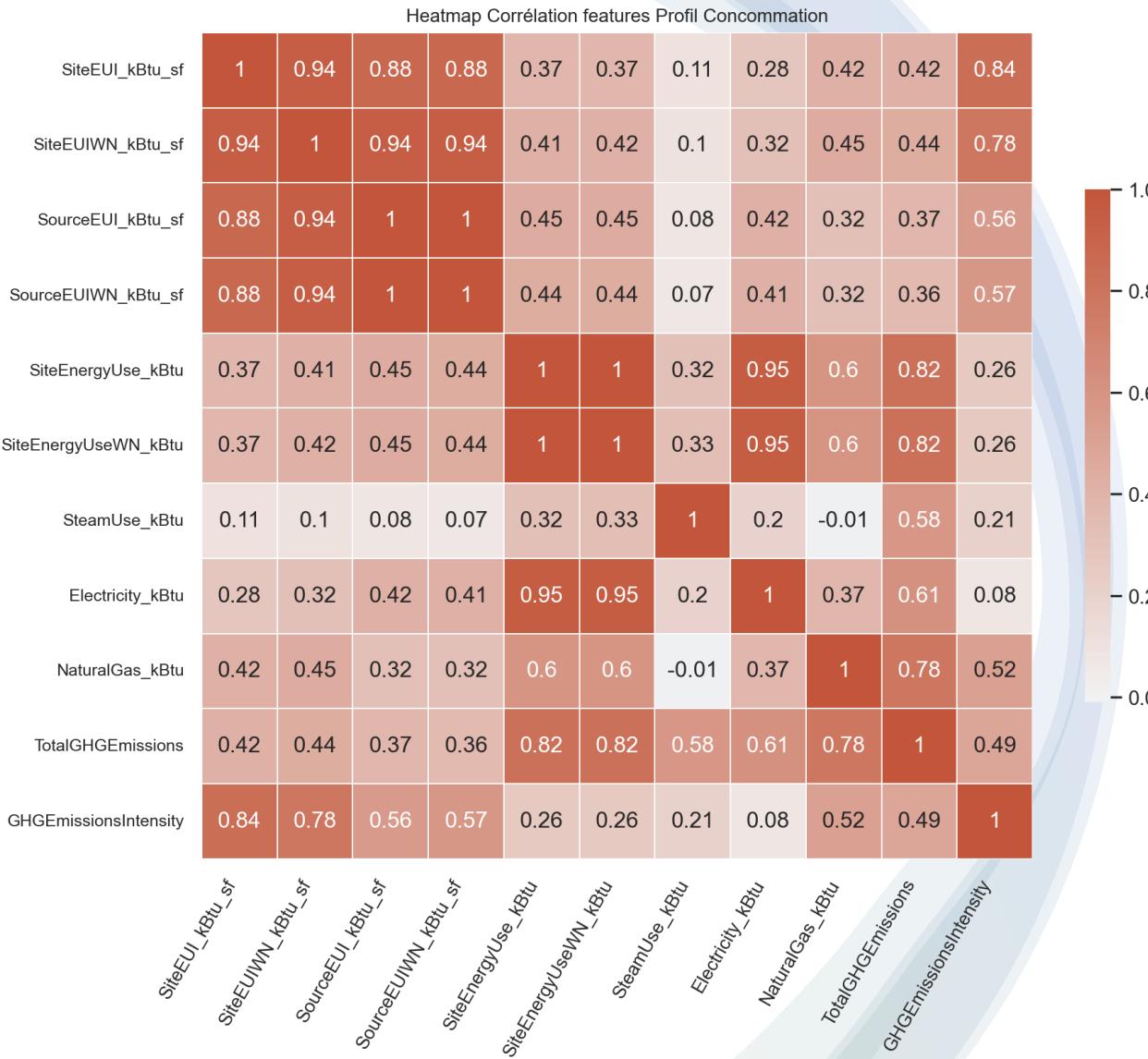
	level_0	level_1	corr_coeff
16	PropertyGFATotal	GFABuilding_s	0.97
18	LargestUseType_GFA	GFABuilding_s	0.97
14	PropertyGFATotal	LargestUseType_GFA	0.95
12	SecondUseType_GFA	PropertyGFAParking	0.83
8	PropertyGFATotal	SecondUseType_GFA	0.64
10	SecondUseType_GFA	PropertyGFATotal	0.64
6	PropertyGFAParking	PropertyGFATotal	0.63
4	NmbFloors	GFABuilding_s	0.61
2	NmbFloors	LargestUseType_GFA	0.60
0	ThirdUseType_GFA	SecondUseType_GFA	0.52

La Target « TotalGHGEmission » : est corrélée positivement avec toutes les *features* en relation avec la superficie.  
 Les corrélations les plus fortes sont entre les *features* qui rapportent les superficies, bâtiments, parking, nombre d'étage...



# EDA

## corrélations vs Target

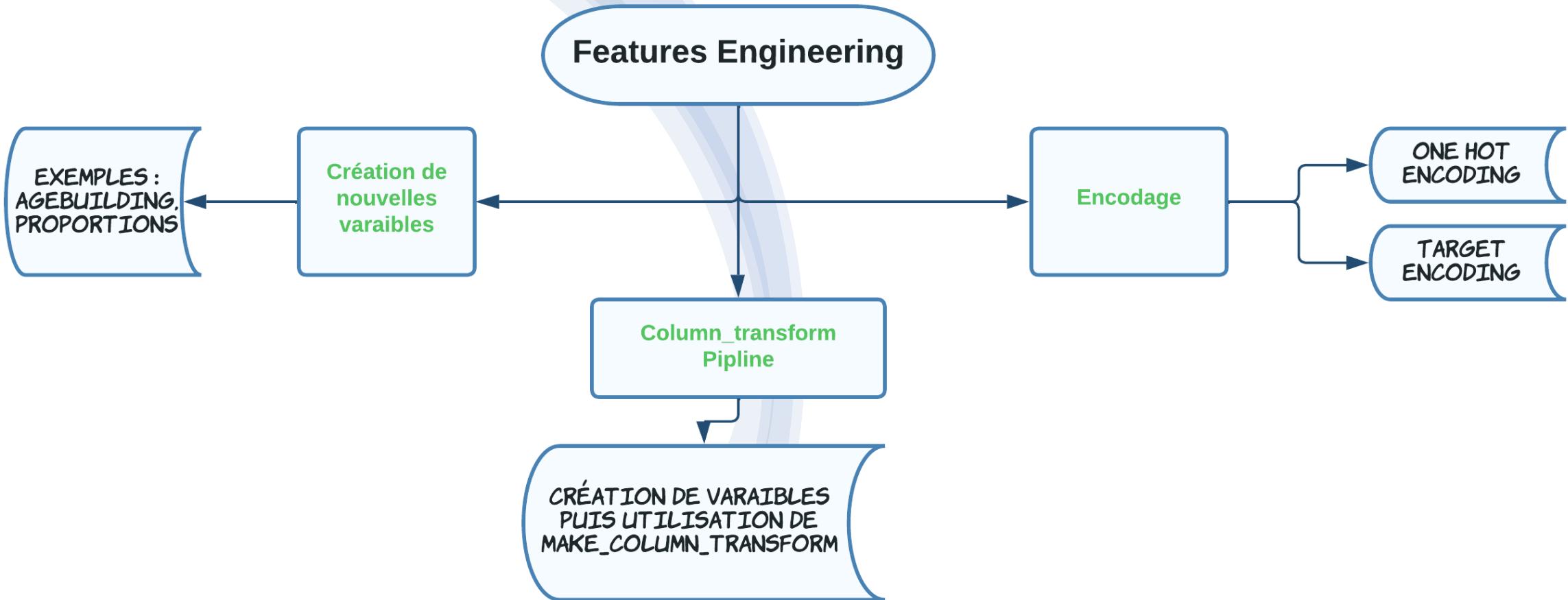


	level_0	level_1	corr_coeff
48	SourceEUI_kBtu_sf	SourceEUIWN_kBtu_sf	1.00
44	SiteEnergyUse_kBtu	SiteEnergyUseWN_kBtu	1.00
36	SiteEnergyUseWN_kBtu	Electricity_kBtu	0.95
34	Electricity_kBtu	SiteEnergyUse_kBtu	0.95
32	SiteEUIWN_kBtu_sf	SourceEUI_kBtu_sf	0.94
30	SourceEUI_kBtu_sf	SiteEUIWN_kBtu_sf	0.94
28	SiteEUI_kBtu_sf	SiteEUIWN_kBtu_sf	0.94
26	SourceEUIWN_kBtu_sf	SiteEUI_kBtu_sf	0.88
24	SourceEUI_kBtu_sf	SiteEUI_kBtu_sf	0.88
22	GHGEmissionsIntensity	SiteEUI_kBtu_sf	0.84
20	SiteEnergyUse_kBtu	TotalGHGEmissions	0.82
18	TotalGHGEmissions	SiteEnergyUse_kBtu	0.82
16	SiteEUIWN_kBtu_sf	GHGEmissionsIntensity	0.78
14	GHGEmissionsIntensity	SiteEUIWN_kBtu_sf	0.78
12	TotalGHGEmissions	Electricity_kBtu	0.61
10	NaturalGas_kBtu	SiteEnergyUseWN_kBtu	0.60
8	NaturalGas_kBtu	SiteEnergyUse_kBtu	0.60
6	SteamUse_kBtu	TotalGHGEmissions	0.58
4	SourceEUIWN_kBtu_sf	GHGEmissionsIntensity	0.57
2	SourceEUI_kBtu_sf	GHGEmissionsIntensity	0.56
0	NaturalGas_kBtu	GHGEmissionsIntensity	0.52

De forte corrélation positives entre les *features* consommation d'énergie et la Target « SiteEnergyUse ».



# Features engineering



# Features engineering

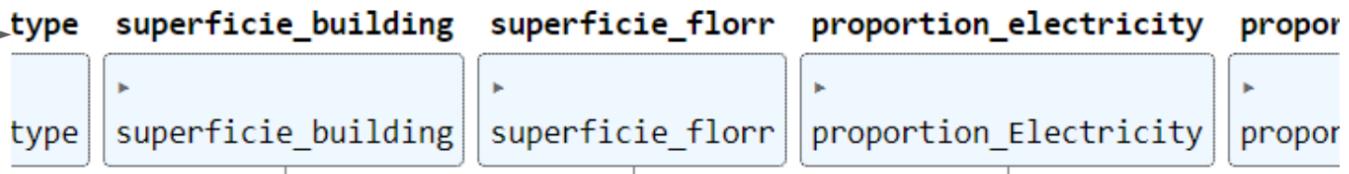
## Features Engineering

Création de nouvelles variables

```
class AgeBuilding(BaseEstimator, TransformerMixin):  
  
    def fit(self, X, y=None):  
        return self  
  
    def transform(self, X):  
        X = X.copy()  
        X['Age_Building'] = X['DataYear'] - X['YearBuilt']  
        del X['DataYear']  
        del X['YearBuilt']  
        return X  
  
    def get_feature_names_out(self):  
        return ['Age_Building']
```

Column\_transform Pipeline

ColumnTransformer

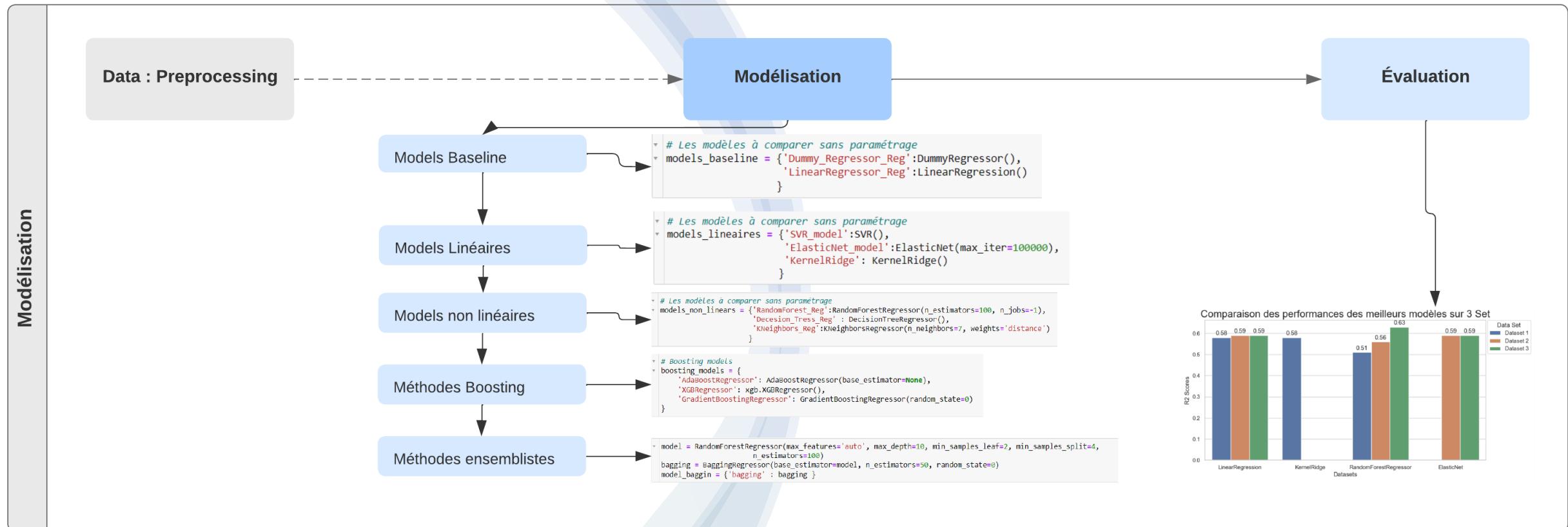


Encodage

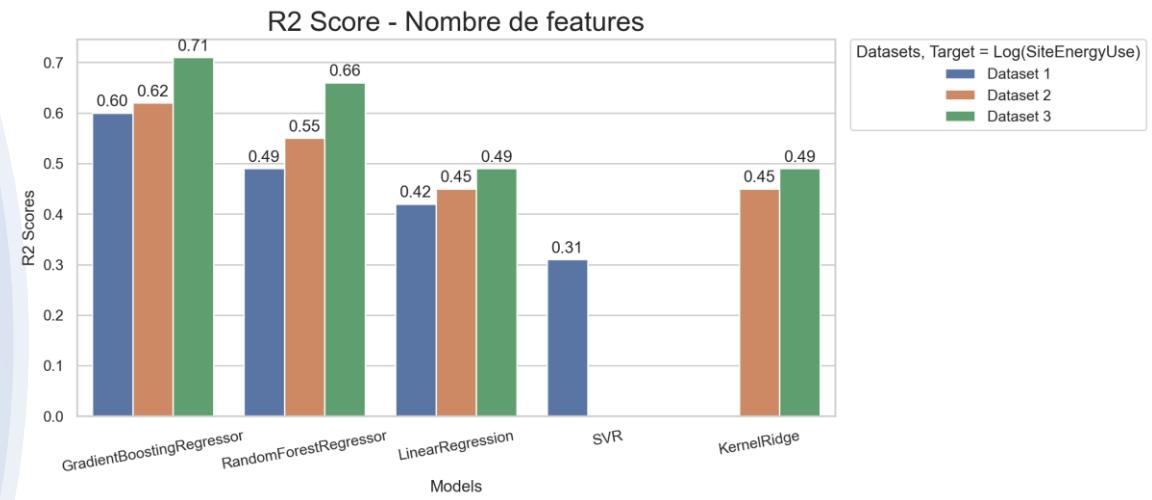
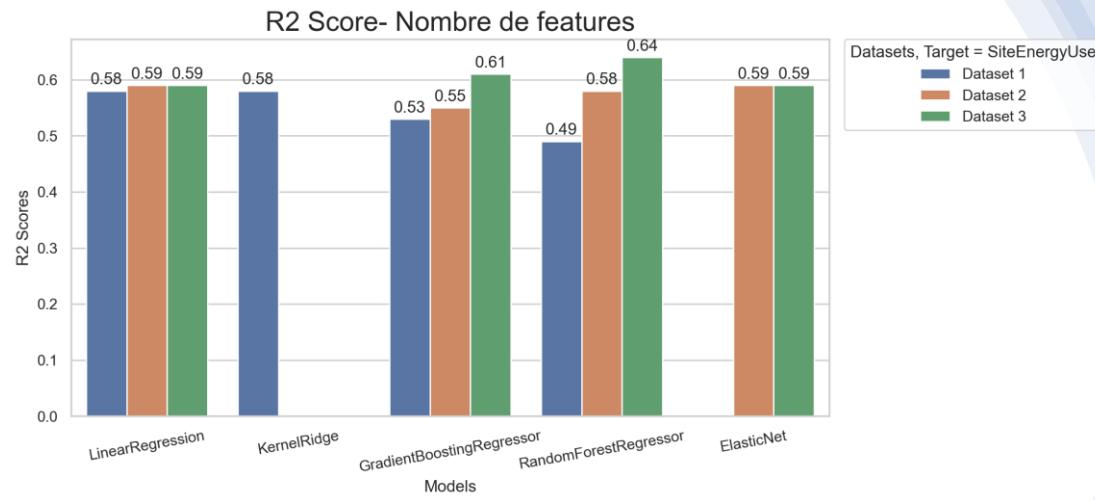
```
# Encodage LabelEncoder()  
le = LabelEncoder()  
df_transformed['Age_Building_groupe_L'] = le.fit_transform(df_transformed['Age_Building_groupe'])
```

# Modélisation

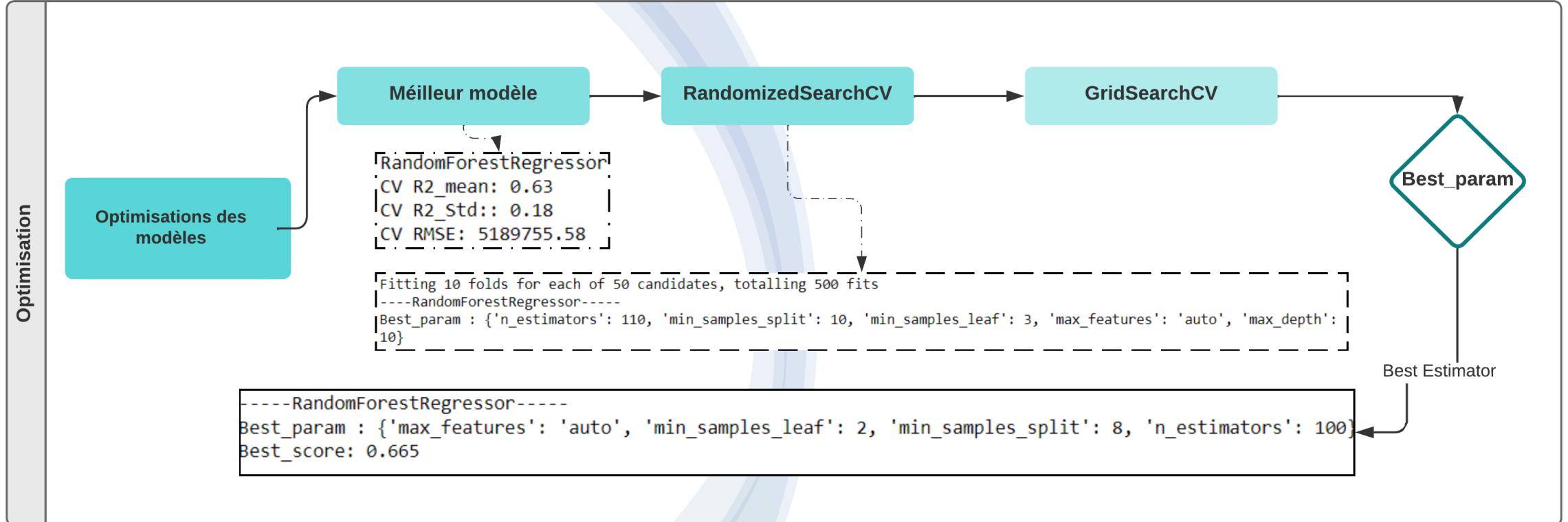
## Processus de modélisation - évaluation - validation - prévision



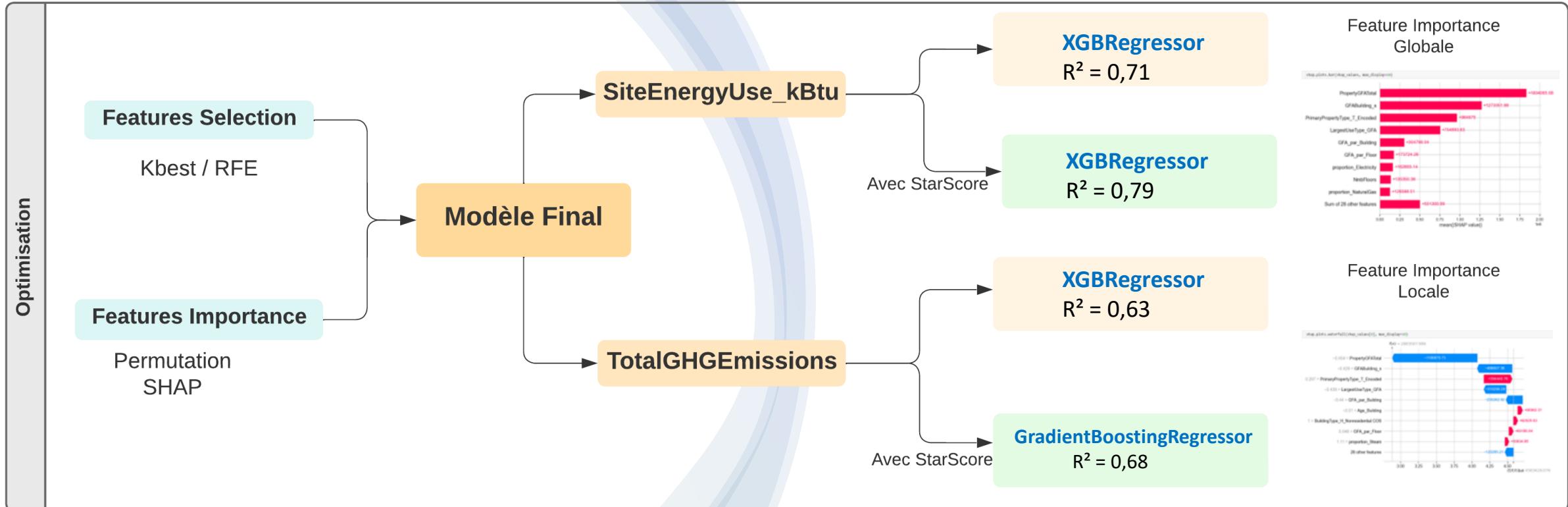
# Résultats de modélisation



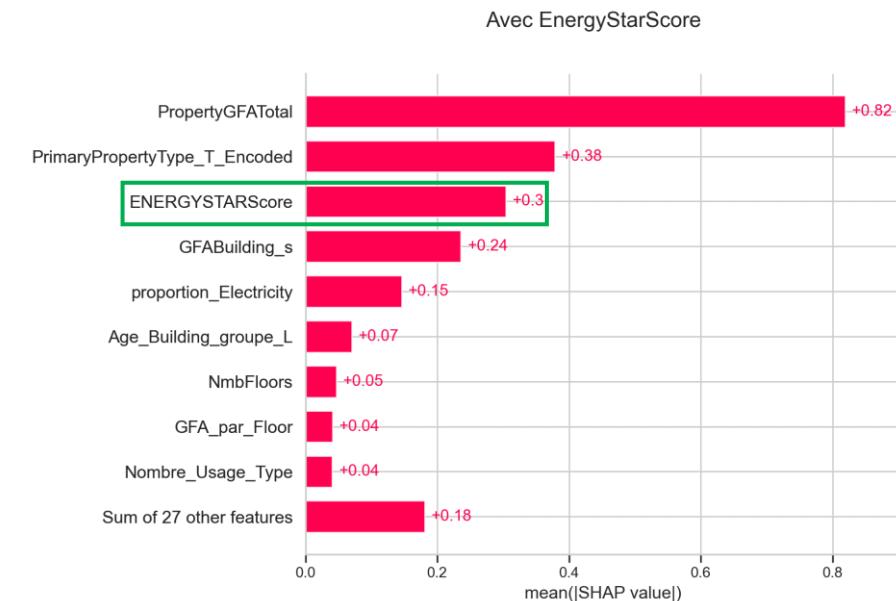
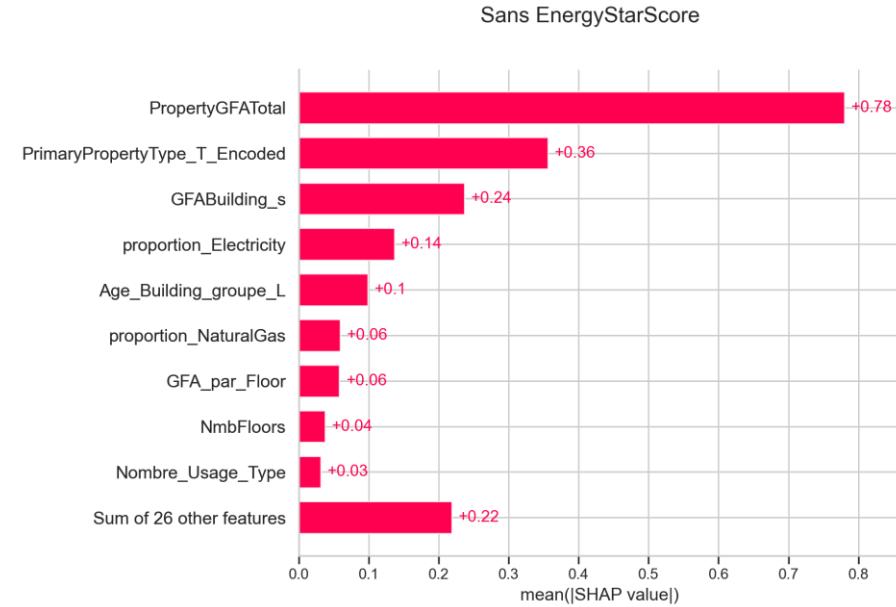
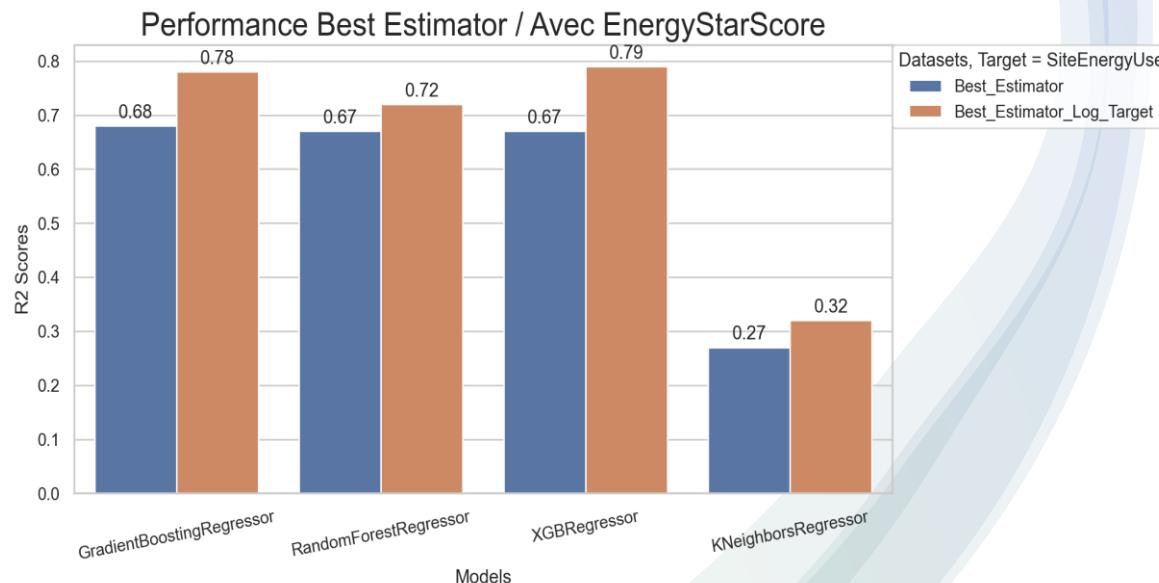
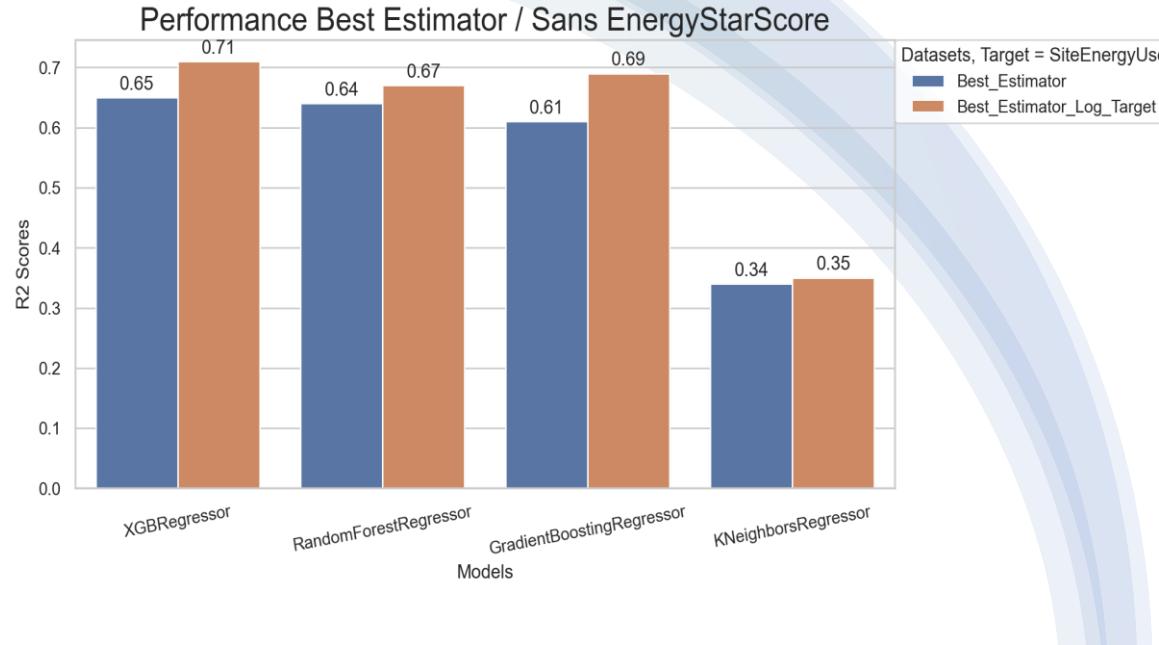
# Optimisation



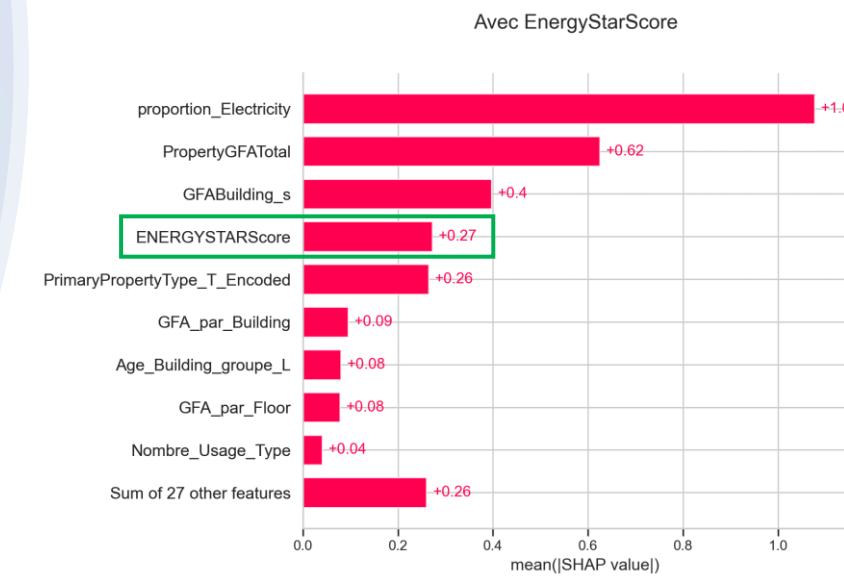
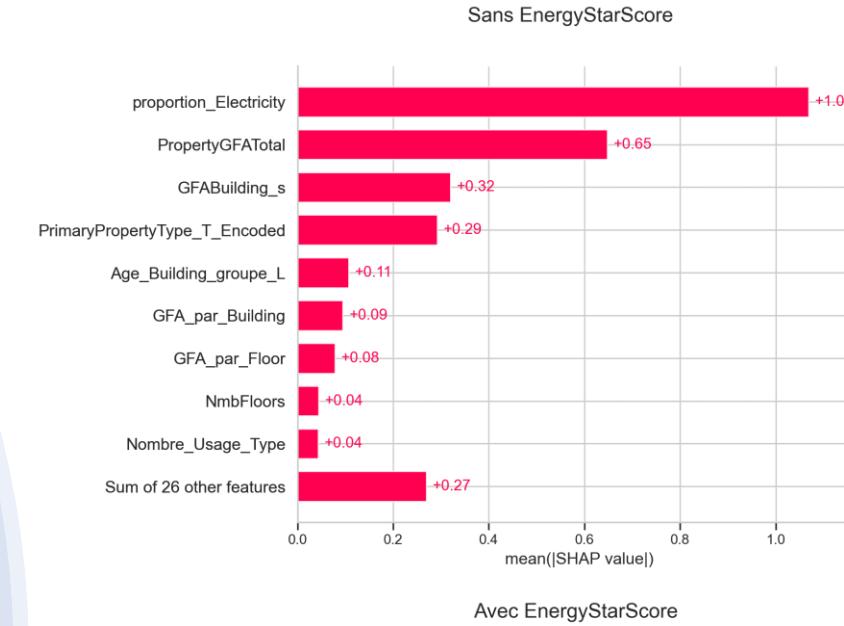
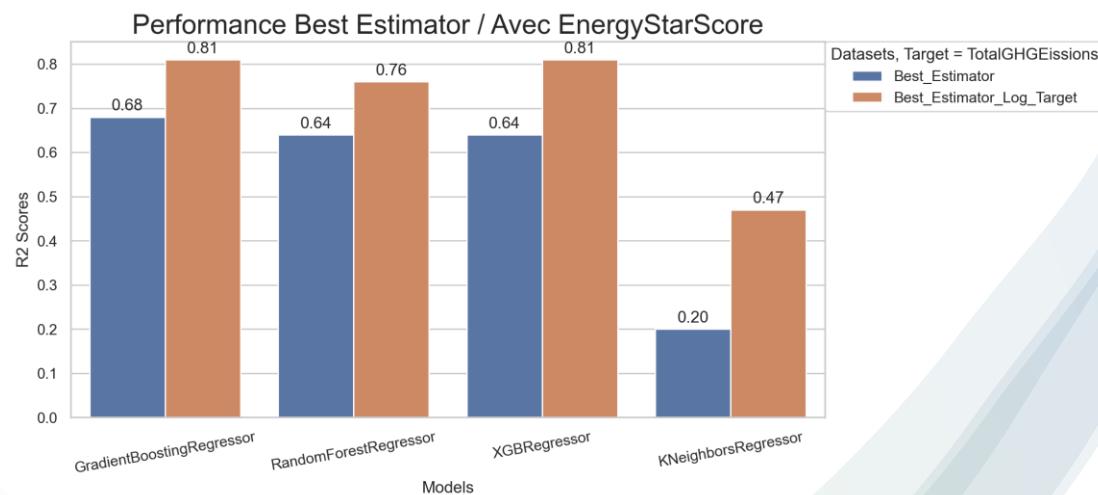
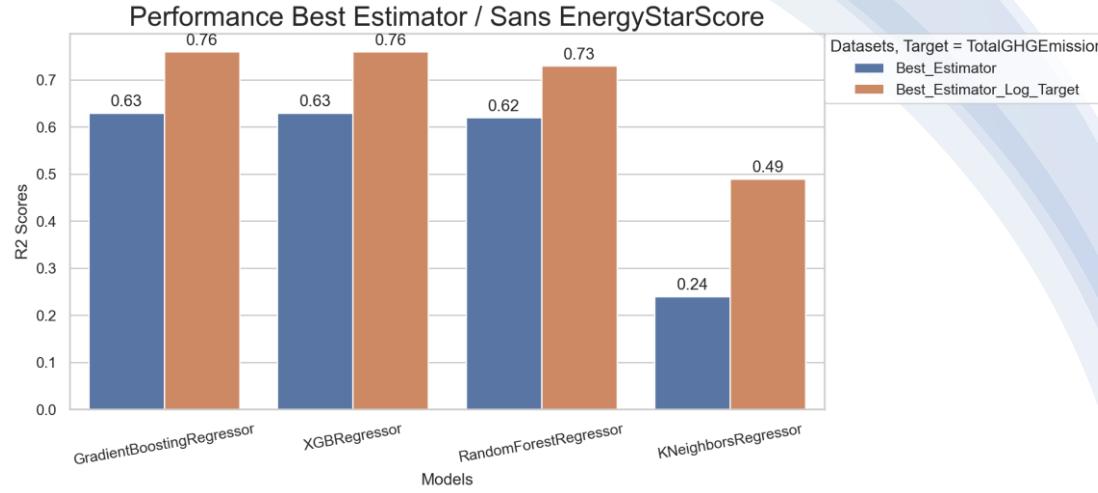
# Choix du modèle Final



# Modélisation – Consommation d'énergie-



# Modélisation





# Conclusions

Le ***feature engineering*** a un **impact** considérable sur la performance des différents modèles, et sur la performance du modèle final.

La **transformation logarithmique** des **Target** augmente la performance du modèle choisis.

L'introduction de **L'EnergyStarScore** dans la modélisation de la **consommation d'énergie** augmente la performance du meilleur modèle de **0,08 pts**, et de **0,05 pts** lors de la modélisation des **émissions de CO<sub>2</sub>**.

*Anticipation de la consommation  
d'énergies des bâtiments et les  
émissions CO2*

Parcours Data Scientist.

Merci de votre attention

Présenté par Mr Dai TENSAOUT