CONSOMMATION D'ÉNERGIE DE SEATTLE

OBJECTIFS

1. Prédire la consommation énergétique globale et le total des émissions de gaz à effet de serre (GES) des bâtiments **non résidentiels** à Seattle en fonction de leurs **caractéristiques structurelles**.

2. Analyser l'influence de la variable Energy STAR Score sur les prédictions des variables cibles.

LES SOURCES

- Données issues du programme de référencement énergétique des bâtiments de Seattle du bureau du développement durable et de l'environnement,
- Données de 2016,
- Bâtiments non résidentiels (tous les bâtiments type logement sont exclus de l'analyse),
- Analyse sur 1610 observations
- Analyse sur 1063 observations avec ES

VARIABLES CIBLES

VARIABLES CIBLES:

- SITEENERGYUSEWN(KBTU)
  Consommation totale d'énergie ajustée aux conditions météorologiques moyennes.

- TOTALGHGEMISSIONS: Quantité totale d'émission de GES (en tonnes)

VARIABLES EXPLICATIVES

6 variables qualitatives

- Primarypropertytype
- Neighborhood
- Buildingtype
- Largestpropertyusetype
- Secondlargestpropertyusetype
- Thirdlargestpropertyusetype

8 variables quantitatives:

- Propertygfatotal
- Yearbuilt
- Numberoffloors
- Numberofbuildings
- Largestpropertyusetypegfa
- Secondlargestpropertyusetypegfa
- Thirdlargestpropertyusetypegfa
- Energystarscore[2]

*[2]: est un système d'évaluation créé par l'**Environmental Protection Agency (EPA)** pour évaluer l'efficacité énergétique des bâtiments. Il attribue une note de 1 à 100*, indiquant l'économie d'énergie d'un bâtiment par rapport à d'autres similaires. Un score élevé signifie une meilleure performance énergétique.

GESTION DES VALEURS ABERRANTES

- Pour les variables quantitatives, je vérifie si les années, le nombre d'étages et le nombre de bâtiments semblent raisonnables. Je m'assure que l'EnergyStarScore est compris entre 1 et 100.

- En ce qui concerne les variables de superficie et les cibles, il est essentiel de s'assurer qu'il n'y a pas de valeurs négatives et qu'elles ne présentent pas de problèmes d'échelle.

- Pour les variables qualitatives, j'ai effectué une harmonisation de l'orthographe des quartiers (tous en majuscules et suppression des doublons); je supprime les bâtiments considérés comme résidentiels s'il en reste.

```
Neighborhood
BALLARD                    64
Ballard                     6
CENTRAL                    49
Central                     5
DELRIDGE                   40
DELRIDGE NEIGHBORHOODS      1
DOWNTOWN                  355
Delridge                    4
EAST                      119
GREATER DUWAMISH          340
LAKE UNION                147
MAGNOLIA / QUEEN ANNE     149
NORTH                      58
NORTHEAST                 126
NORTHWEST                  80
North                       9
Northwest                   5
SOUTHEAST                  46
SOUTHWEST                  41
Name: OSEBuildingID, dtype: int64
```

GESTION DES VALEURS MANQUANTES

1. **Traitement manuel** lorsque les données manquantes sont peu nombreuses.
2. **Suppression** de l'observation remplies de valeurs NaN, sans possibilité d'imputation.
3. **Déductions pour les secondes et troisièmes utilisations du bâtiment** : Par exemple, si la superficie totale = superficie de l'utilisation principale → pas d'autres utilisations.
4. **Imputation des variables cibles** : Les bâtiments de même taille et type d'usage ont des niveaux de consommation énergétique et des quantités d'émissions de GES comparables. Création de groupes de bâtiments similaires (type & quantile de superficie). Remplacement par la première valeur rencontrée dans chaque groupe.

FEATURE ENGINEERING

1. Suppression de la variable ListOfAllPropertyUseTypes,

2. Création des variables représentant le pourcentage d'utilisation par type d'énergie:

- Pourcentage d'utilisation de l'électricité :

  Electricité consommée / consommation totale

- Pourcentage d'utilisation du gaz naturel :

  Gaz consommé / consommation totale

- Pourcentage d'utilisation de l'énergie thermique à vapeur :

  Energie thermique / consommation totale

FEATURE ENGINEERING

3. Création de la variable Ancienneté du bâtiment:

Année actuelle – YearBuilt

4. Transformation des variables quantitatives:

La Skewness évalue l'asymétrie d'une distribution par rapport à une distribution normale, tandis que le Kurtosis mesure le degré d'aplatissement.

- Skewness ≈ 0 indique une distribution symétrique.
- Skewness > 0 signifie une asymétrie à droite, >2 très forte asymétrie.
- Skewness < 0 indique une asymétrie à gauche.
- Kurtosis ≈ 3 indique une distribution normale.
- Kurtosis > 3 révèle de nombreuses valeurs extrêmes.
- Kurtosis < 3 suggère peu de valeurs extrêmes.

Variable	Skewness	Kurtosis
NumberofBuildings	10.183519	116.277327
NumberofFloors	5.116773	34.441585
PropertyGFATotal	4.786011	29.172849
LargestPropertyUseTypeGFA	5.250706	36.513714
SecondLargestPropertyUseTypeGFA	5.201632	36.733839
ThirdLargestPropertyUseTypeGFA	9.340637	117.588919
Electricity_%	-0.339141	-0.980219
NaturalGas(kBtu)_%	0.424373	-1.113223
SteamUse(kBtu)_%	4.708219	22.202180
BuildingAge	0.301209	-1.025280

TRANSFORMATION FEATURES

Les transformations suivantes seront appliquées

- **Transformation en log**:

NumberofBuldings,

NumberofFloors,

PropertyGFATotal,

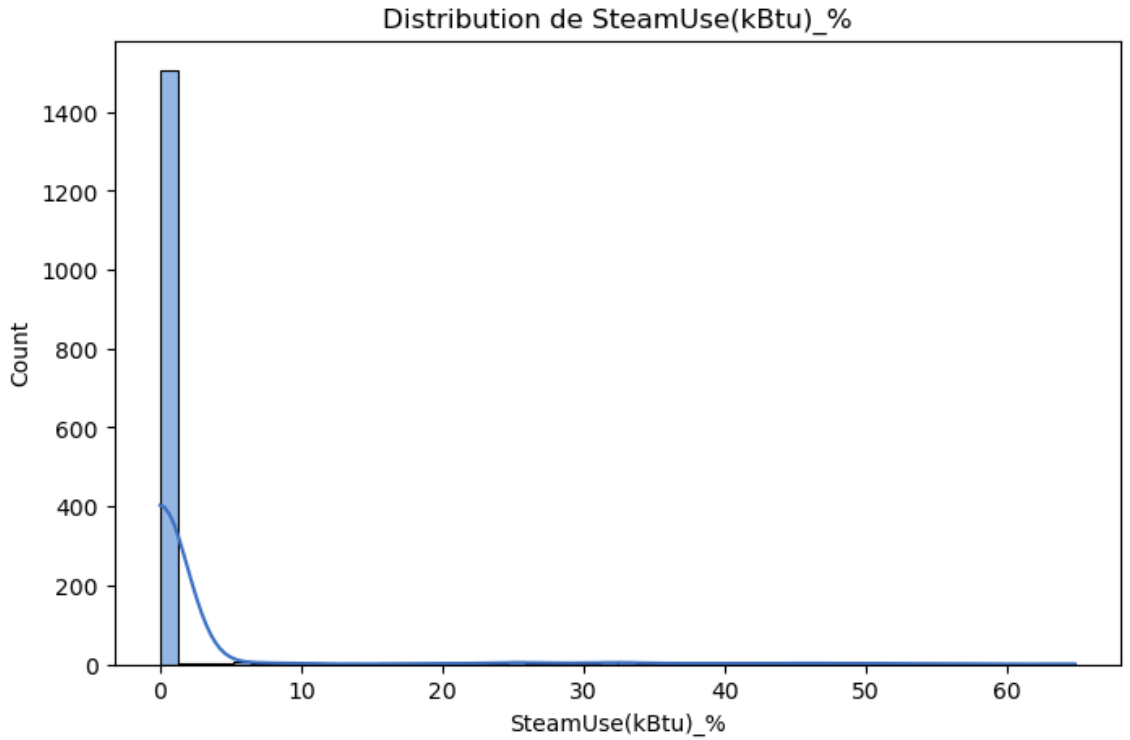
LargestPropertyUseTypeGFA,

SecondLargestPropertyUseTypeGFA,

ThirdLargestPropertyUseTypeGFA.

- **Transformation en racine carrée:**

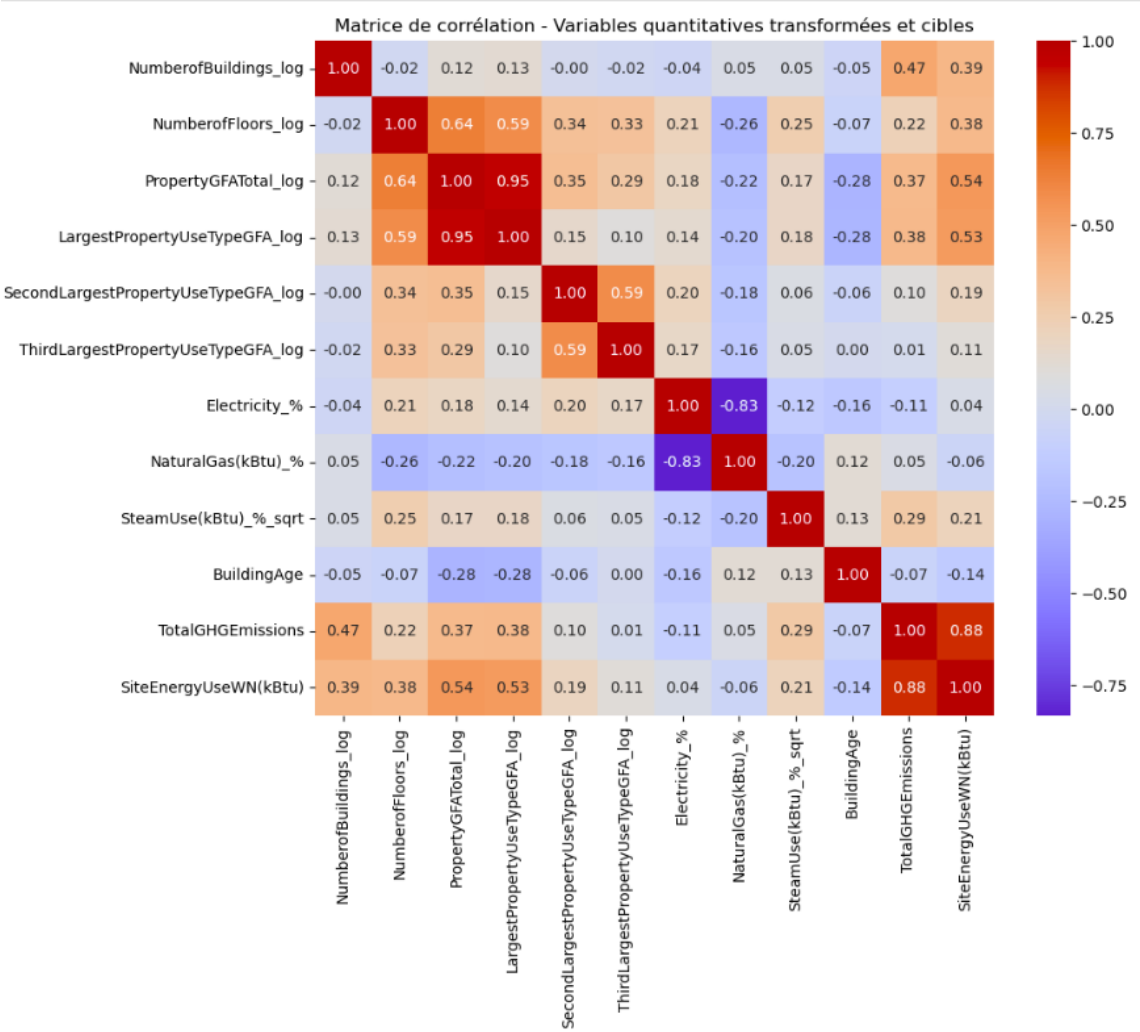
StreamUse(kBtu)_%



CORRÉLATION

Variables quantitatives : corrélation de Pearson

- Le nombre de bâtiment, la superficie totale et la superficie de l'utilisation principale ont des corrélation positives modérées avec l'émission de GES.
- La superficie totale et la superficie de l'utilisation principale ont des corrélations positives avec la consommation d'énergie globale.



TEST CHI²

Pour la cible GES : toutes les variables qualitatives, à l'exception de ThirdLargestPropertyUseType, sont significativement dépendantes à la cible.

Pour la cible consommation d'énergie : toutes les variables sont significativement liées à cette cible, avec PrimaryPropertyType montrant la plus forte association.

Variables qualitatives : test du Chi²:

- Hypothèse nulle (H0) : Les variables sont indépendantes.
- (H1) : Les variables ne sont pas indépendantes.

```
Chi2 pour BuildingType et TotalGHGEmissions: p-value = 0.0003648353188218136, Chi² = 38.99919693533765
Chi2 pour PrimaryPropertyType et TotalGHGEmissions: p-value = 5.959313639535879e-84, Chi² = 749.025448257909
Chi2 pour Neighborhood et TotalGHGEmissions: p-value = 1.0633379939437723e-07, Chi² = 169.27961325374508
Chi2 pour LargestPropertyUseType et TotalGHGEmissions: p-value = 3.438572037313877e-41, Chi² = 778.7794433154602
Chi2 pour SecondLargestPropertyUseType et TotalGHGEmissions: p-value = 0.006331813195740377, Chi² = 272.40121118459894
Chi2 pour ThirdLargestPropertyUseType et TotalGHGEmissions: p-value = 0.12895818263791767, Chi² = 218.54952042487767

Chi2 pour BuildingType et SiteEnergyUseWN(kBtu): p-value = 6.476526256542697e-09, Chi² = 67.08505578303263
Chi2 pour PrimaryPropertyType et SiteEnergyUseWN(kBtu): p-value = 5.121537745840572e-139, Chi² = 1048.9649563916232
Chi2 pour Neighborhood et SiteEnergyUseWN(kBtu): p-value = 5.990758404080496e-17, Chi² = 240.05557090573757
Chi2 pour LargestPropertyUseType et SiteEnergyUseWN(kBtu): p-value = 9.17249755603484e-59, Chi² = 907.3606460697424
Chi2 pour SecondLargestPropertyUseType et SiteEnergyUseWN(kBtu): p-value = 7.309861562223303e-13, Chi² = 398.6468929729858
Chi2 pour ThirdLargestPropertyUseType et SiteEnergyUseWN(kBtu): p-value = 5.509892202682574e-06, Chi² = 295.5309040530953
```

ENCODAGE ET ACP

Modèles sans ES:
Features explicatives : 16

Features OHE: 148

ACP: 78

Modèles avec ES:
Features explicatives: 17

Features OHE: 98

ACP: 47

Le "One Hot Encoding" permet de convertir les variables catégorielles en variables numériques. Chaque catégorie est représentée par une colonne binaire (0 ou 1), où une seule colonne est égale à 1 pour indiquer la présence de la catégorie, tandis que les autres affichent 0.

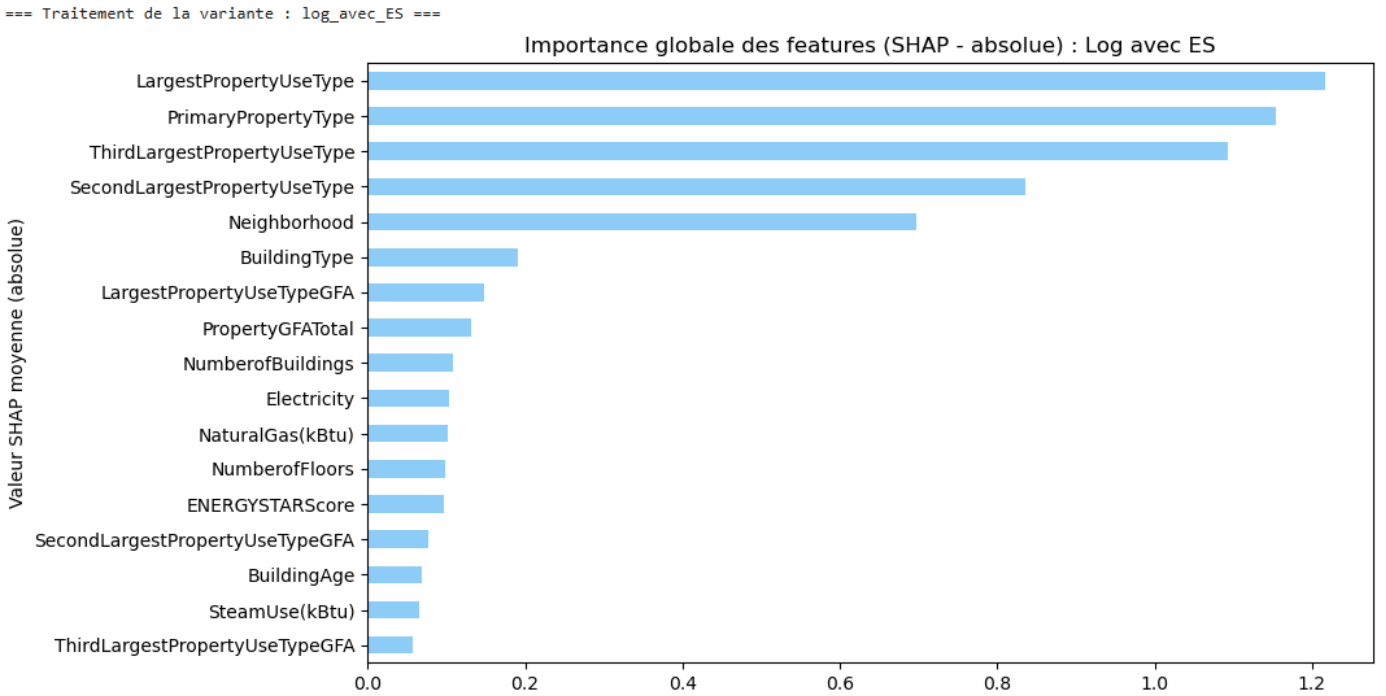
- Chaque catégorie est indépendante (pas de poids implicite)
- Aucune perte d'informations
- Augmente la dimensionnalité

ACP permet de contrebalancer cet effet en transformant ces variables explicatives en composantes principales.

MODÈLES: ÉMISSIONS DE GES

Modèles testés	Base sans ES	Base avec ES	Log / features transformées sans ES	Log / features transformées avec ES
RandomForest	R² train: 0,29 R² test: 0,24 Overfitting: 0,05 Timer: 40 secondes	R² train: 0,32 R² test: 0,29 Overfitting: 0,02 Timer: 22 secondes	R² train: 0,68 R² test: 0,49 Overfitting: 0,18 Timer: 40 secondes	R² train: 0,77 R² test: 0,63 Overfitting: 0,14 Timer: 16 secondes
ElasticNet	R² train: 0,50 R² test: 0,49 Overfitting: 0,01 Timer: 0,29 secondes	R² train: 0,57 R² test: 0,53 Overfitting: 0,03 Timer: 0,19 secondes	**R² train: 0,69** **R² test: 0,55** **Overfitting: 0,14** **Timer: 0,28 secondes**	R² train: 0,70 R² test: 0,67 Overfitting: 0,03 Timer: 0,32 secondes
GradientBoosting	**R² train: 0,74** **R² test: 0,76** **Overfitting: -0,02** **Timer: 52** secondes	**R² train: 0,66** **R² test: 0,57** **Overfitting: 0,08** **Timer: 25 secondes**	R² train: 0,92 R² test: 0,61 Overfitting: 0,3 Timer: 78 secondes	**R² train: 0,97** **R² test: 0,72** **Overfitting: 0,2** **Timer: 32 secondes**
KNN	R² train: 0,43 R² test: 0,43 Overfitting: 0,005 Timer: 1,7 secondes	R² train: 0,99 R² test: 0,70 Overfitting: 0,30 Timer: 1,3 secondes	R² train: 0,99 R² test: 0,49 Overfitting: 0,5 Timer: 1,6 secondes	R² train: 0,99 R² test: 0,56 Overfitting: 0,4 Timer: 1,3 secondes

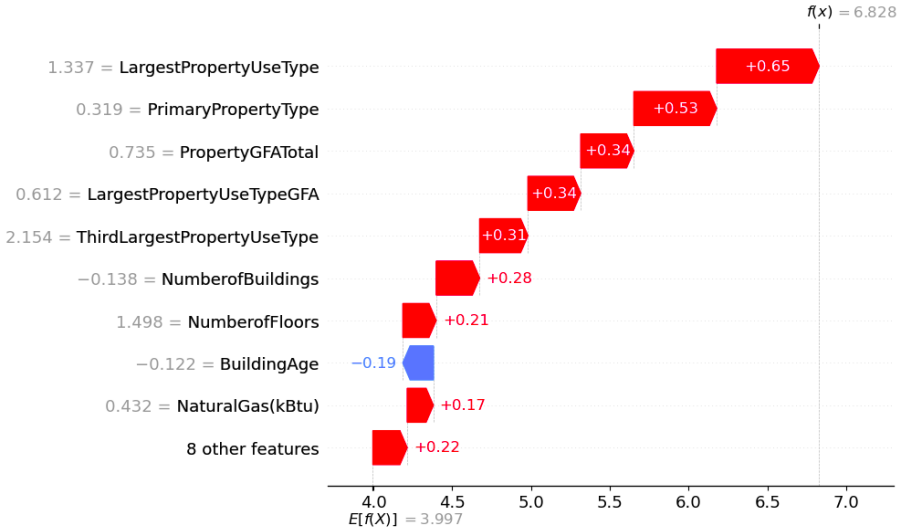
FEATURE IMPORTANCE 1/2



L'utilisation des valeurs SHAP: indiquent combien chaque feature augmente ou diminue la prédiction par rapport à la valeur moyenne du modèle. Ici ce sont les valeurs SHAP approximatives, puisque les valeurs SHAP réelles sont calculées sur les composantes principales.

Le barplot montre que les variables « types d'utilisation » sont les variables qui impactent le plus la prédiction. La variable ENERGYSTARScore a un impact faible.

FEATURE IMPORTANCE 2/2

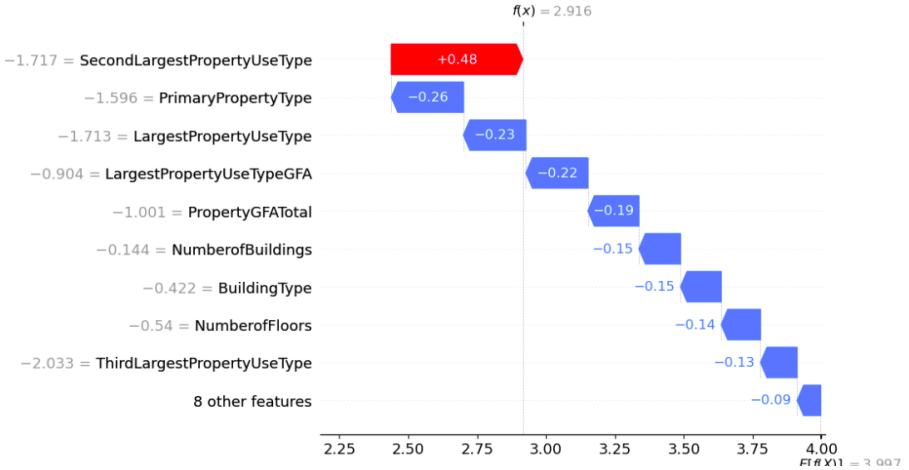


Les graphiques en cascade illustrent la feature importance locale pour chaque observation.

Observation 1 : (gauche) un hôtel de 12 étages datant de 1927. Ici, toutes les variables explicatives, sauf « BuildingAge », ont une influence positive sur la prédiction de la cible. L'Energystarscore appartient aux variables qui ont un impact le plus faible.

Observation 78 (droite); bureau financier et parking (2nd utilisation)

La plupart des variables affectent négativement la prédiction de la cible, sauf pour la variable SecondLargestPropertyUseType

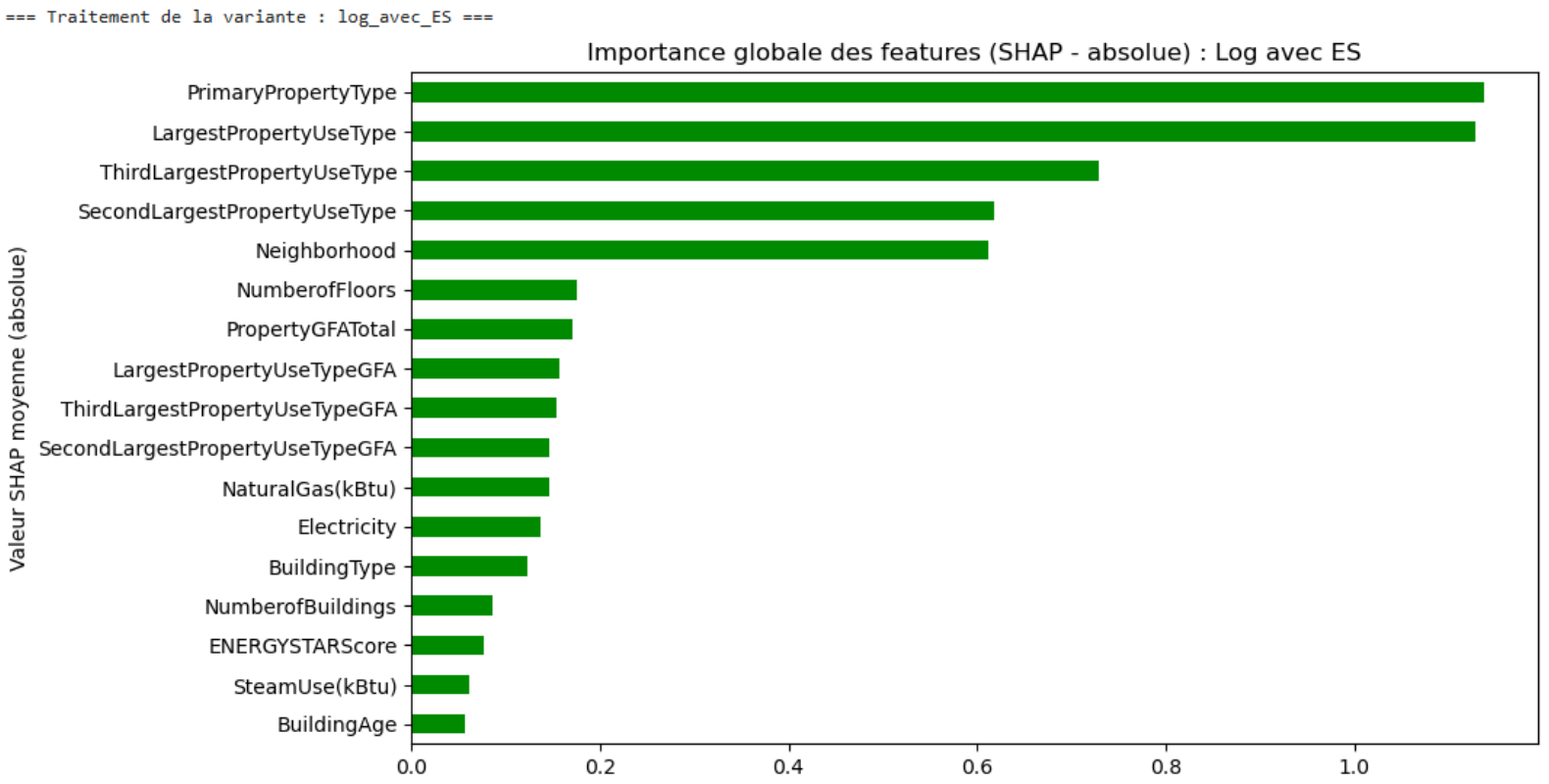


MODÈLES: CONSOMMATION D'ÉNERGIE

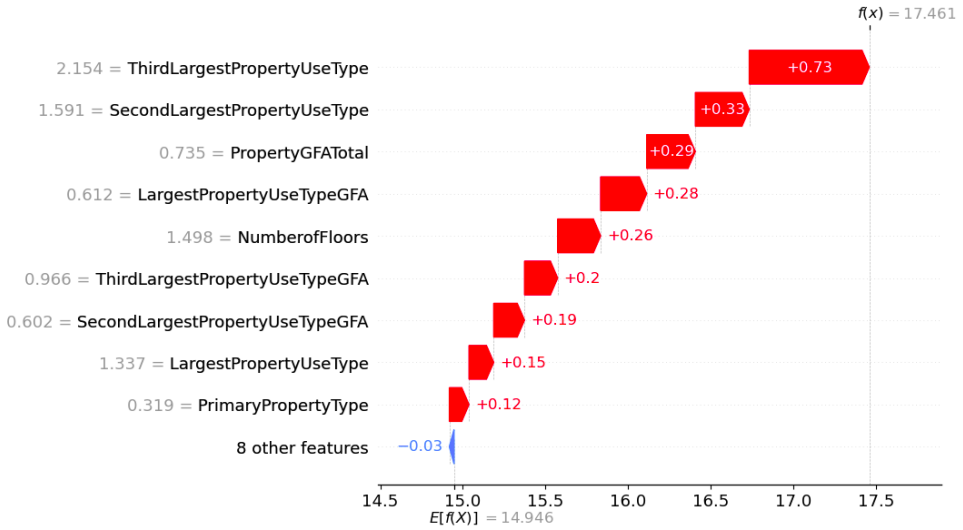
Modèles testés	Base	Base avec ES	Log / features transformées sans ES	Log / features transformées
RandomForest	R² train: 0,50 R² test: 0,32 Overfitting: 0,2 Timer: 51 secondes	R² train: 0,38 R² test: 0,38 Overfitting: 0,0 Timer: 35 secondes	R² train: 0,45 R² test: 0,62 Overfitting: -0,1 Timer: 34 secondes	R² train: 0,36 R² test: 0,15 Overfitting: 0,2 Timer: 32 secondes
ElasticNet	**R² train: 0,64 R² test: 0,46 Overfitting: 0,2 Timer: 0,3 seconde**	**R² train: 0,56 R² test: 0,60 Overfitting: -0,04 Timer: 0,3 secondes**	**R² train: 0,32 R² test: 0,65 Overfitting: -0,3 Timer: 0,24**	R² train: 0,28 R² test: 0,15 Overfitting: 0,1 Timer: 0,3 secondes
GradientBoosting	R² train: 0,69 R² test: 0,43 Overfitting: 0,2 Timer: 52 secondes	R² train: 0,25 R² test: 0,26 Overfitting: -0,01 Timer: 33 secondes	R² train: 0,90 R² test: 0,42 Overfitting: 0,5 Timer: 44 secondes	R² train: 0,95 R² test: 0,25 Overfitting: 0,7 Timer: 68 secondes
KNN	R² train: 0,99 R² test: 0,45 Overfitting: 0,6 Timer: 1,7 secondes	R² train: 0,46 R² test: 0,53 Overfitting: -0,07 Timer: 1,7 secondes	R² train: 0,99 R² test: 0,46 Overfitting: 0,5 Timer: 1,7 secondes	R² train: 0,99 R² test: 0,22 Overfitting: 0,7 Timer: 1,41 secondes

FEATURE IMPORTANCE 1/2

Les types d'utilisation du bâtiment sont les variables ayant le plus d'impact sur la cible.

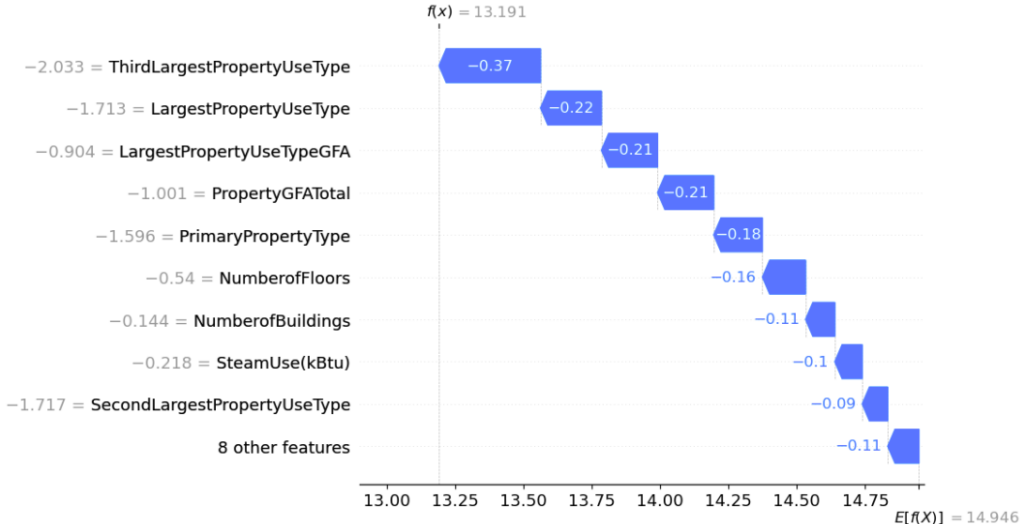


FEATURE IMPORTANCE 2/2



La plupart des variables ont un impact positif sur la prédiction de la feature pour l'observation 1, sauf les 8 variables qui ont un impact très faible.

Pour cette observation, toutes les variables explicatives ont un impact négatif sur la prédiction de la cible,



CONCLUSION

- Le modèle le plus performant a été sélectionné en fonction du R^2 le plus élevé, d'un overfitting faible et d'un temps d'exécution le plus rapide.
- Les modèles Gradient Boosting et ElasticNet se sont imposés comme les modèles les plus performants.
- L'ajout de la variable « EnergyStarScore » n'a pas d'impact significatif sur la prédiction de la quantité totale d'émission de GES
- Le type d'utilisation des bâtiments est le facteur le plus influent sur les prédictions de la consommation d'énergie et de la quantité d' émission des GES

MERCI

Antonine LAROZE-CERVETTI