



# ANTICIPEZ LES BESOINS EN CONSOMMATION DE BÂTIMENTS

OPEN CLASSROOMS 2022 - 2023  
PROJET 3

# OBJECTIFS

Ville de Seattle

Prévisions pour 2050

Emissions de CO<sub>2</sub>

Consommation d'énergie





# MISSIONS

Préparer jeu de données

Exploration des données

Entrainer des modèles de prédiction :  
Regression

Améliorer les performances

Analyser la variable ENERGYSTARScore

# DATASET

Dataset de taille moyenne (3400 lignes et 42 colonnes)

Différents types de données (float, int, object, booléen)

Visualisation globale

Variables intérêts : SiteEnergyUse et TotalGHGEmissions

Etude de la variable ENERGYSTARScore

The monitor displays two main windows. The top window is a terminal showing system monitoring data with columns for CPU usage, memory usage, tasks, load average, and uptime. The bottom window is a database table with columns for Year, Building Type, Primary Property Type, Property Name, and Address.

Year	Building Type	Primary Property Type	Property Name	Address
2016	NonResidential		Mayflower park hotel	405 Olive way
2016	NonResidential		Paramount Hotel	724 Pine street
2016	NonResidential		5673-The Westin Seattle	1900 5th Avenue
2016	NonResidential	HOTEL MAX	STEWART ST	620
2016	NonResidential	WARWICK SEATTLE HOTEL (ID8)	LENORA ST	401



# FILTRE TECHNIQUE

- Suppression colonnes et lignes vides
- Suppression colonnes avec information unique
- Suppression outliers
- Seuil d'acceptation valeurs manquantes (30%)
- Conservation des bâtiments non destinés à l'habitation
- Nouvelles dimensions (1500x36)

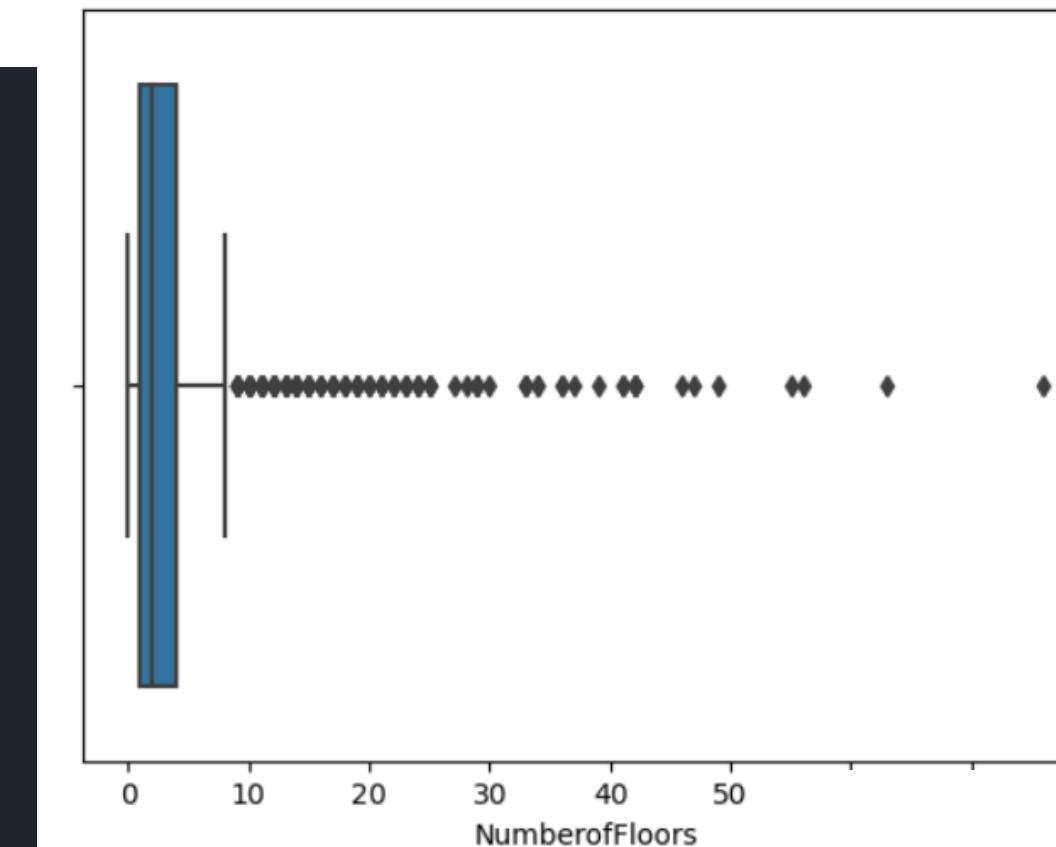
# FILTRE MÉTIER



- Sélection variables d'intérêt (année construction, nombre étages, surface...) : 8
- Eviter Data Leakage
- Suppression variables ENERGYSTARScore
- Nouvelles dimensions (1500x7)
- Matrice de corrélation

# VALEURS ABERRANTES/MANQUANTES

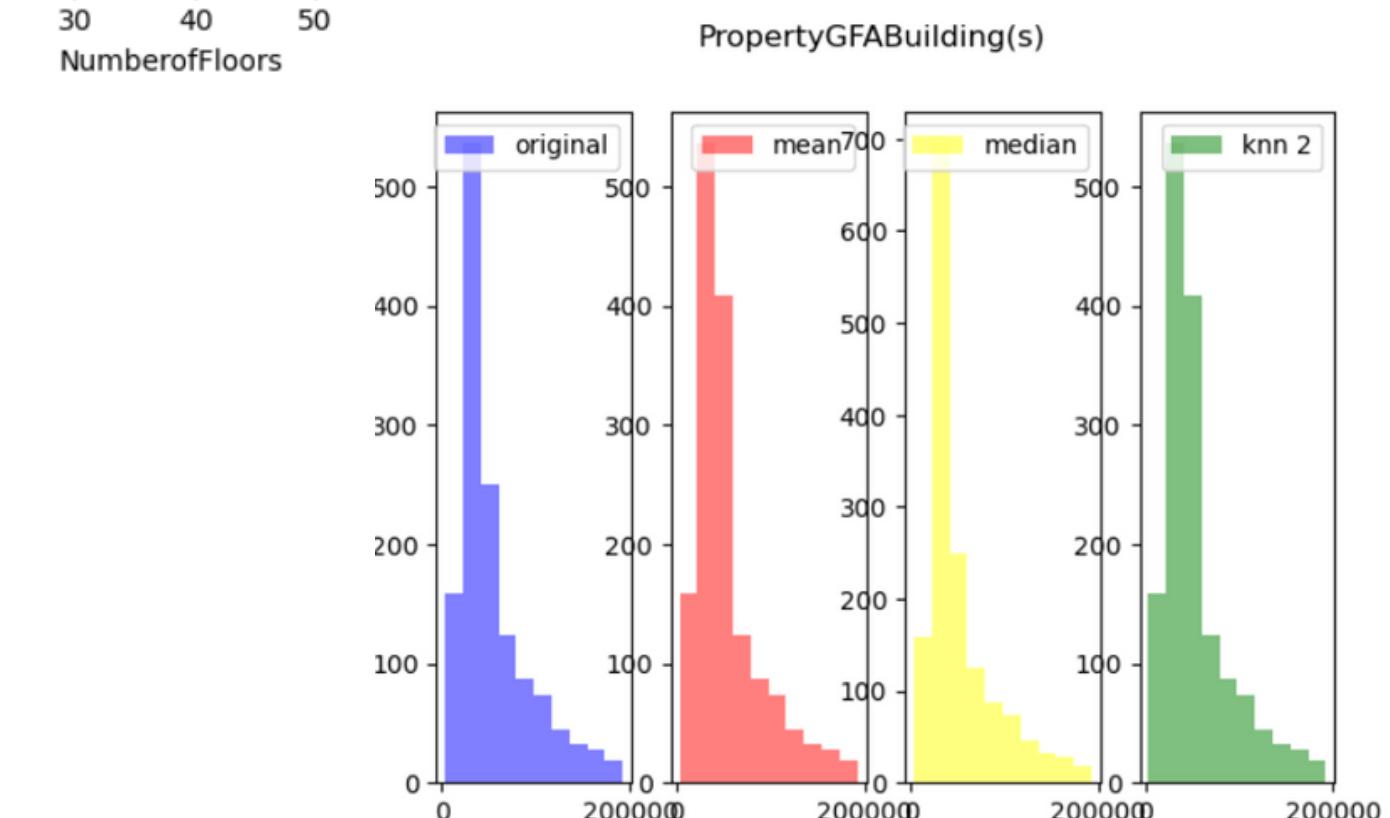
Detection valeurs aberrantes (~10% par variables)



Remplacement par NAN par méthode interquartile

Imputation par médiane ou KNN

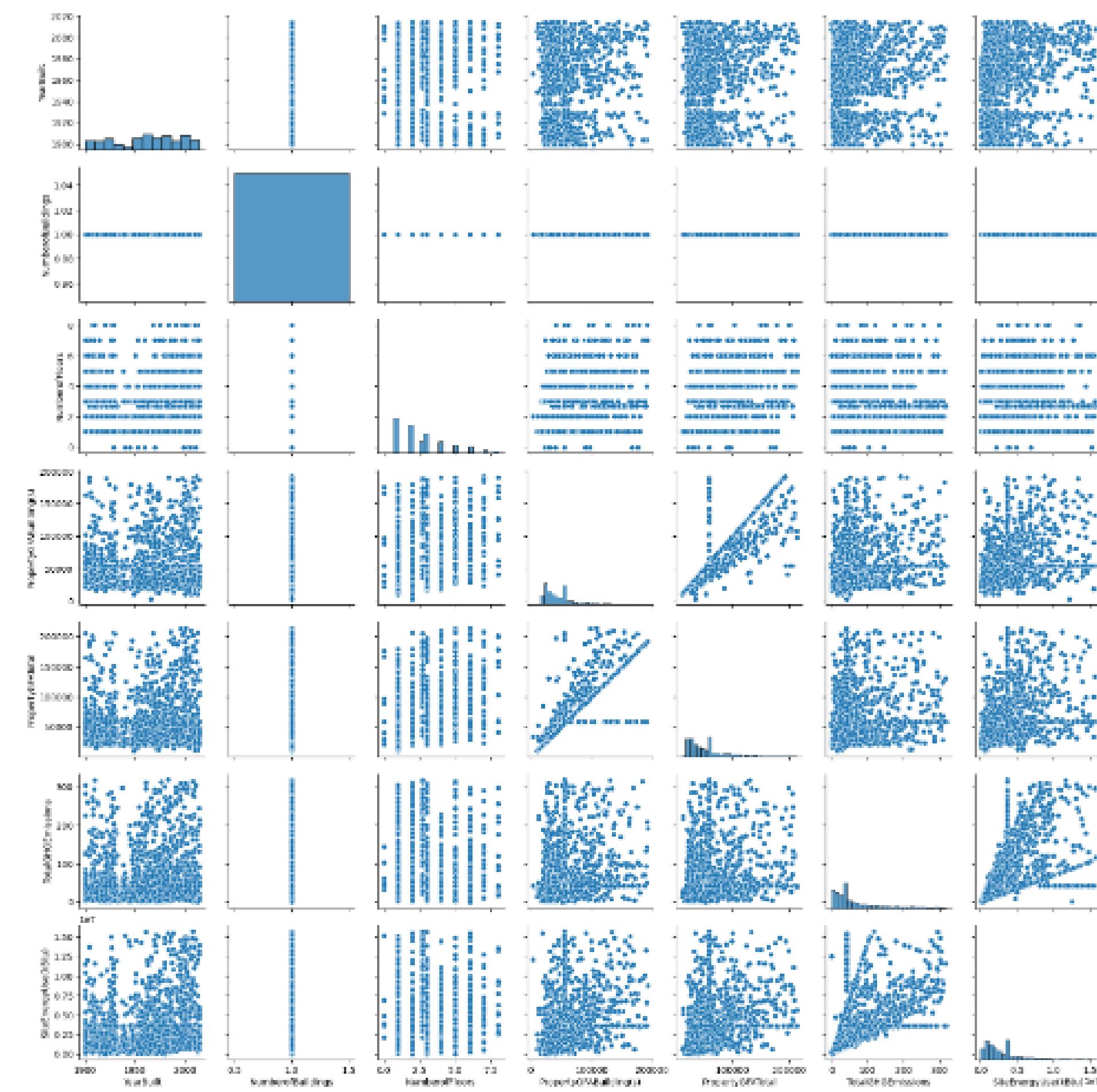
Déduction visuelle



# EDA

## Scatterplot

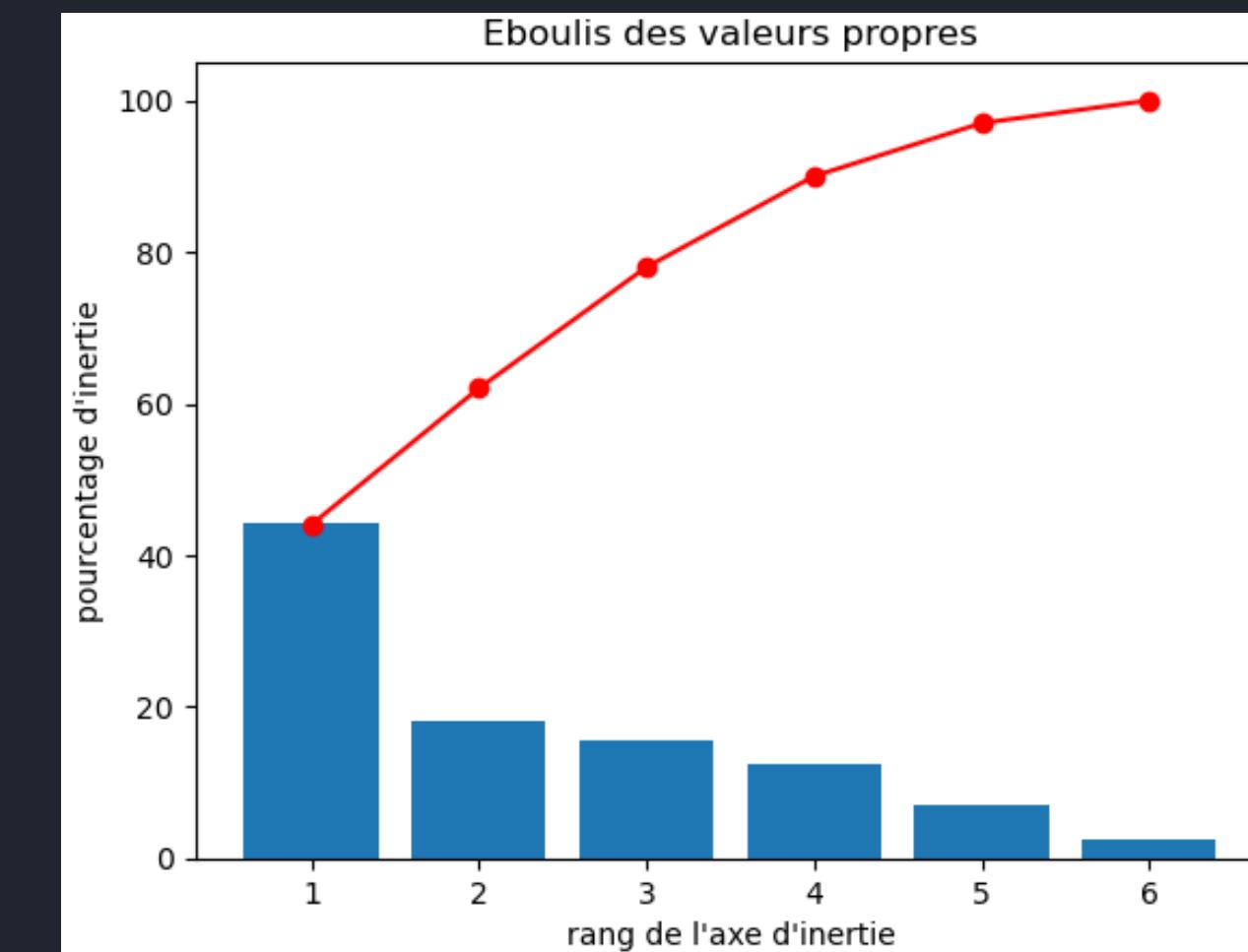
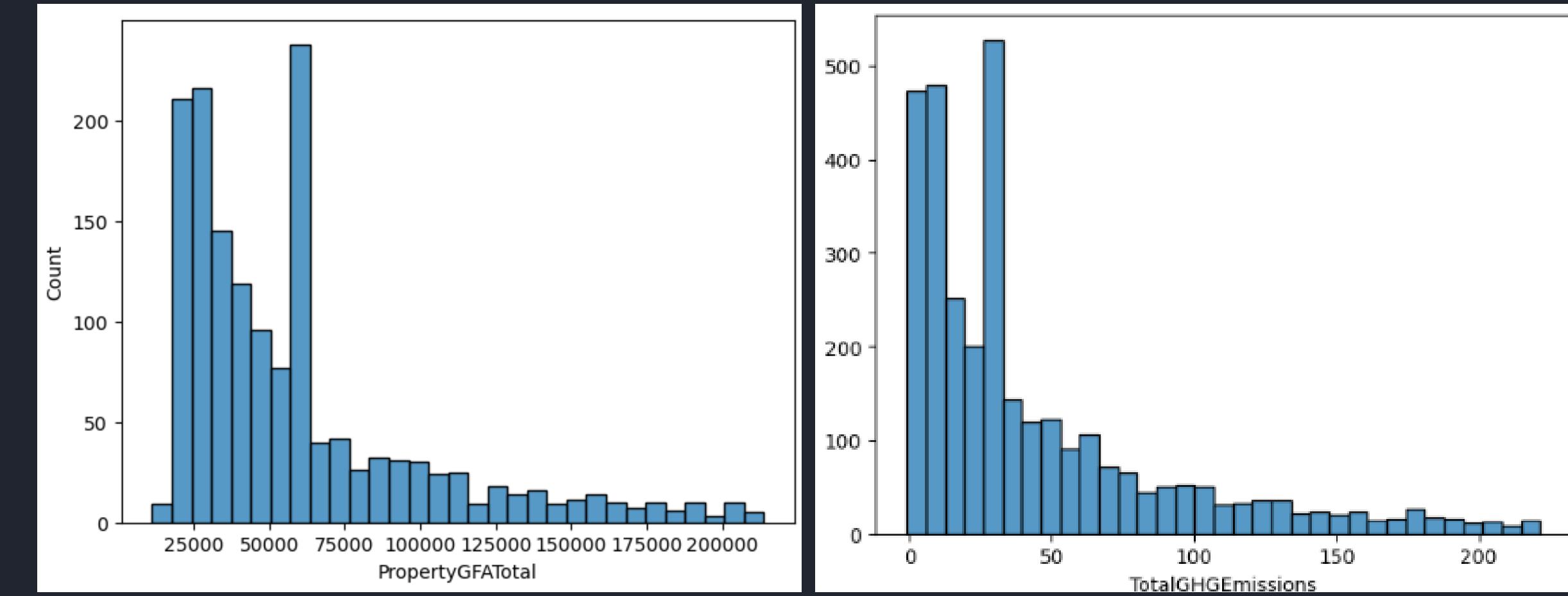
## Statistiques descriptives



	YearBuilt	NumberofBuildings	NumberofFloors	PropertyGFABuilding(s)	PropertyGFAutoTotal
count	1513.000000	1513.0	1513.000000	1513.000000	1513.000000
mean	1961.532056		1.0	2.720407	55045.339734
std	32.825806		0.0	1.698770	36387.865870
min	1900.000000		1.0	0.000000	3636.000000
25%	1930.000000		1.0	1.000000	27800.000000
50%	1965.000000		1.0	2.000000	45680.000000
75%	1988.000000		1.0	3.000000	63888.000000
max	2015.000000		1.0	8.000000	192259.000000

# DISTRIBUTION/ ACP

- Distribution  $\neq$  loi normale
- Test de shapiro pour valider
- ACP ne semble pas adaptée



# EMISSIONS CO<sub>2</sub>

## Data preparation

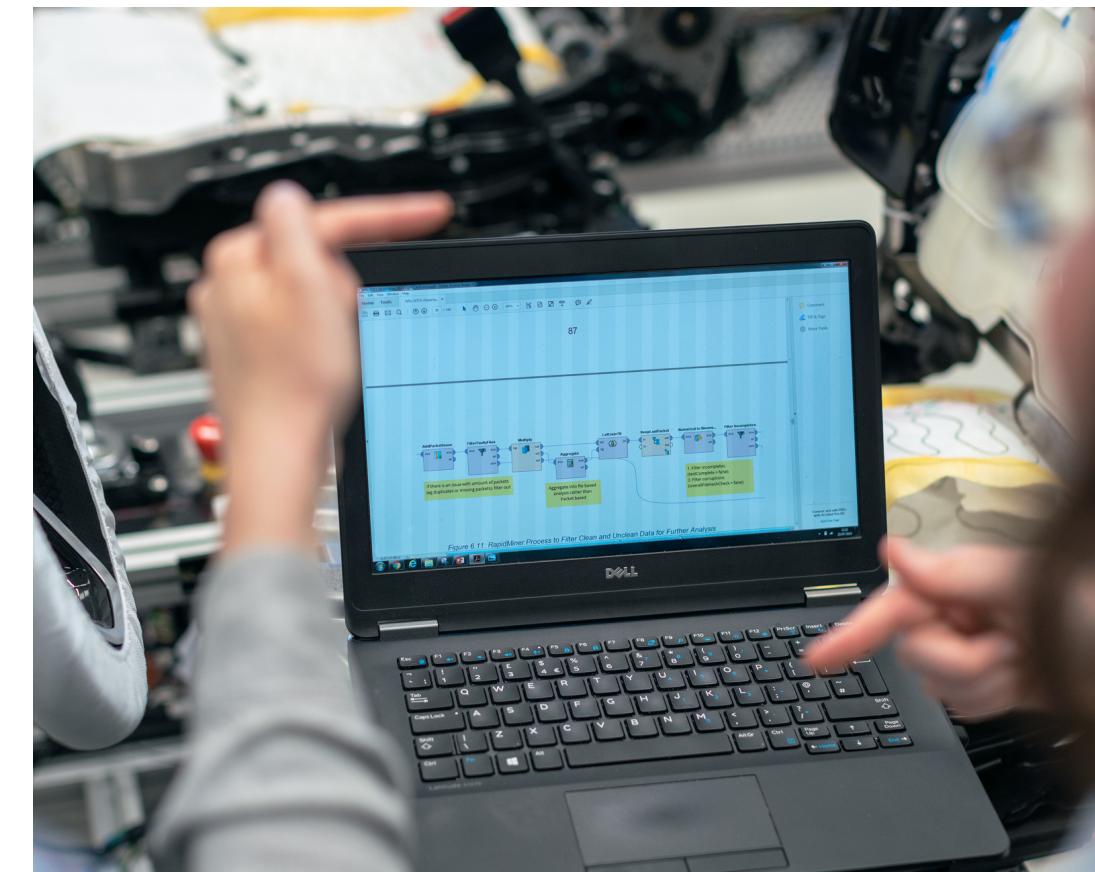
Suppression de la 2e target

Séparation en 2 jeu : Variables + target

Séparation en jeu d'entraînement et de test

Utilisation de LazyRegressor

Validation croisée



# EMISSIONS CO<sub>2</sub>

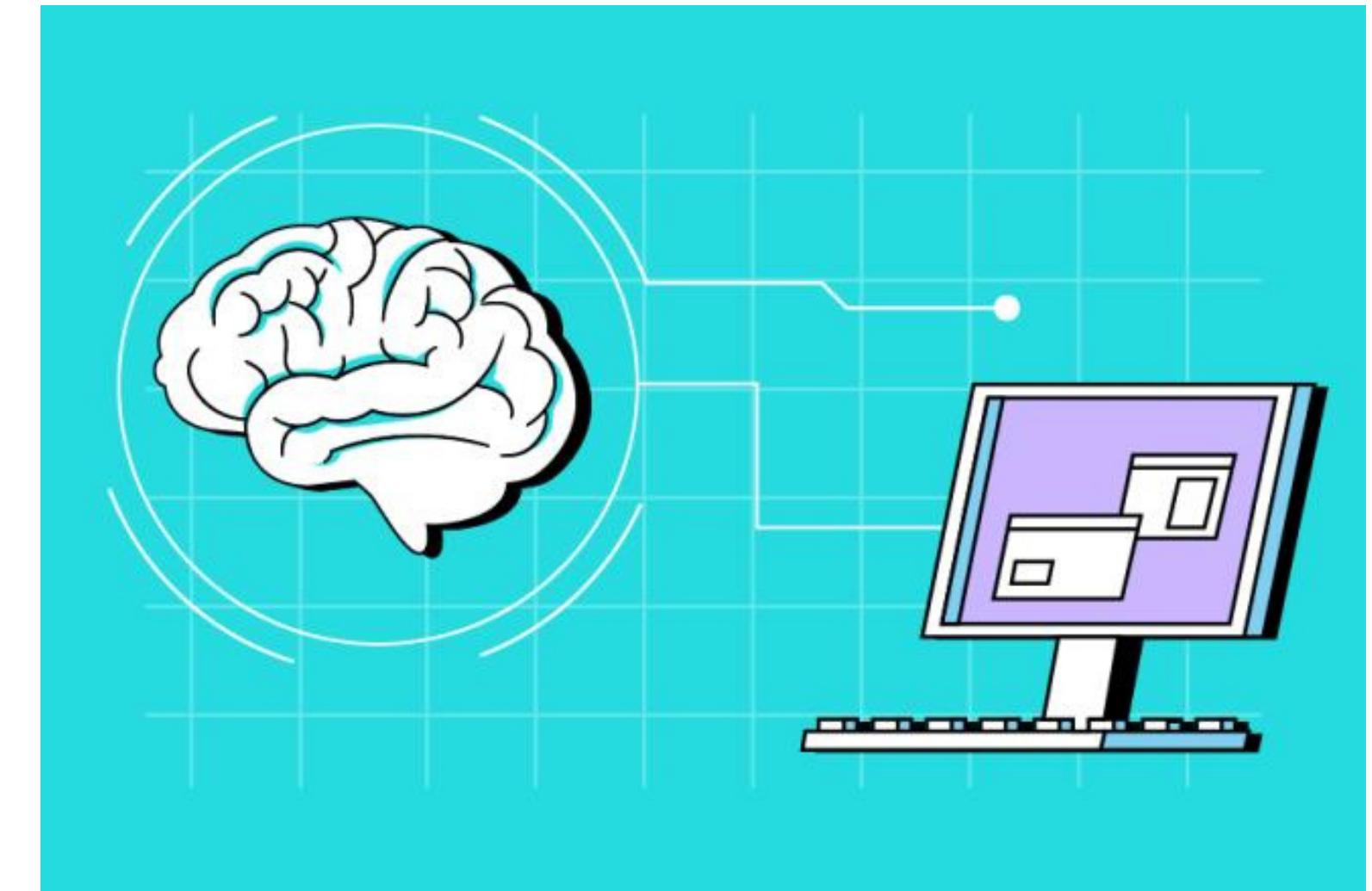
## Modelisation

Dummy regressor

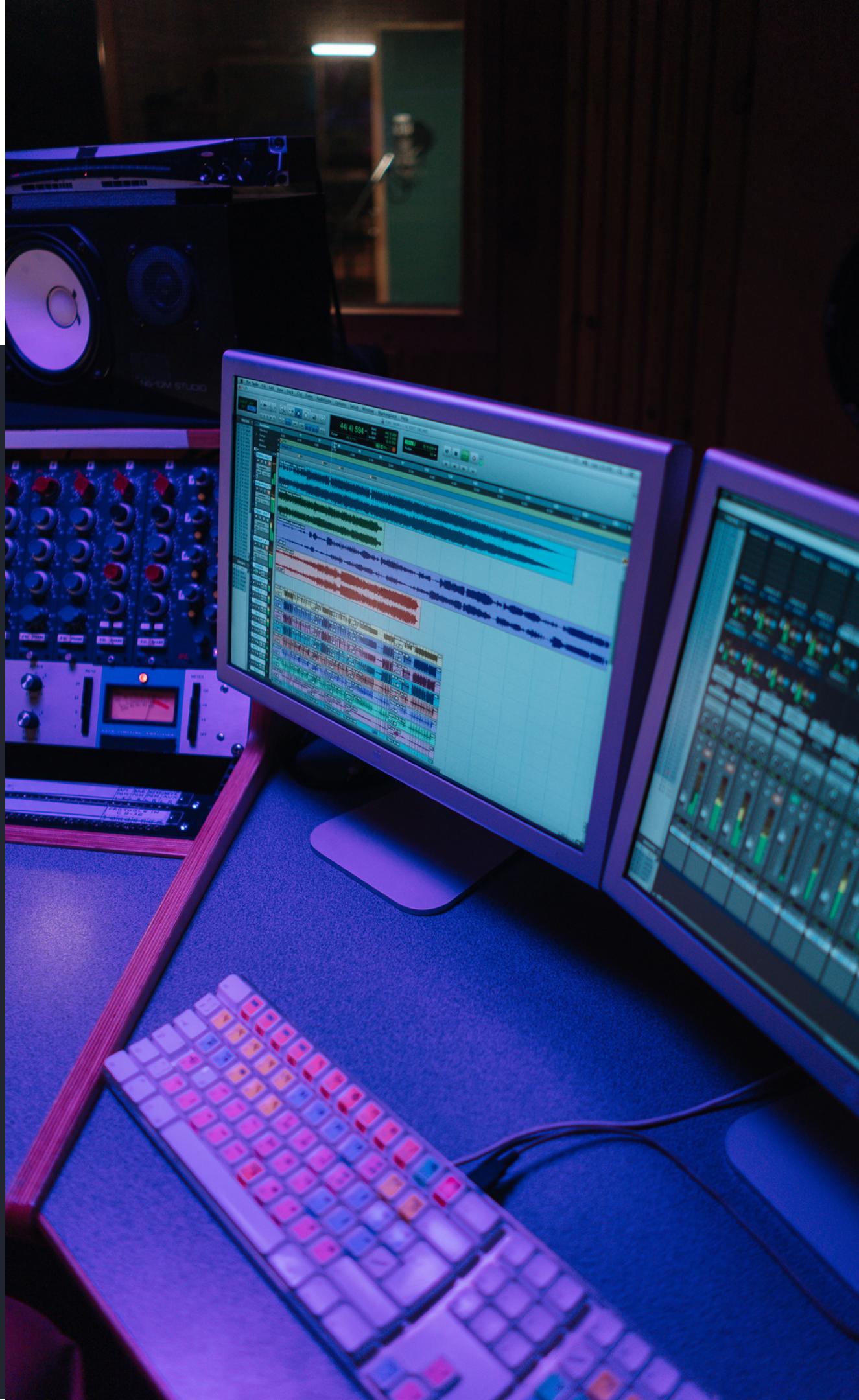
Regression linéaire

Random forest regressor

Métriques de comparaison : MSE, RMSE,  
MAE, R<sup>2</sup>



MSE : 5576.848 || RMSE : 74.678 || MAE : 45.864 || r<sup>2</sup> : -0.124



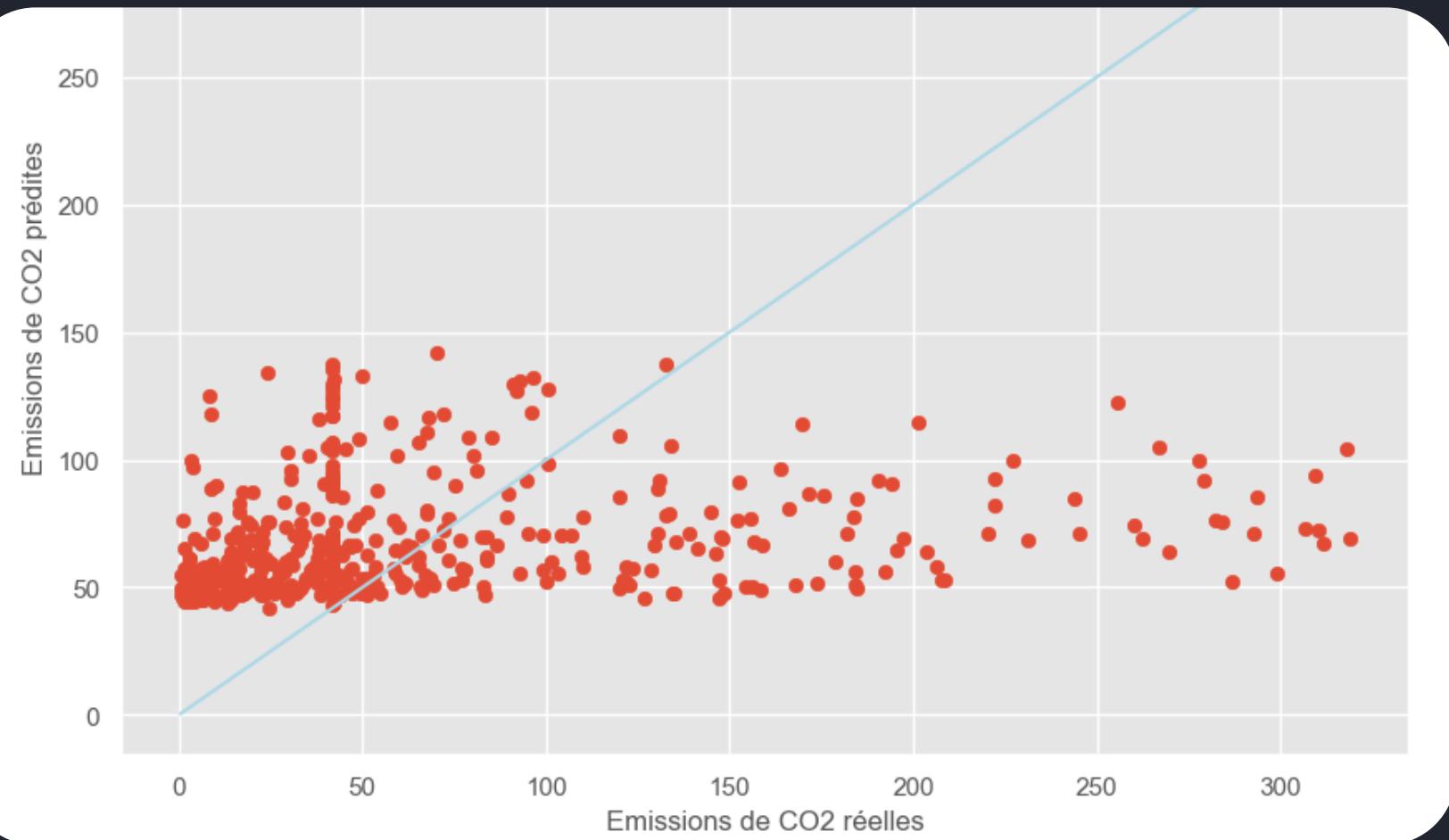
# Dummy Regressor

- Modèle de base
- Strategy mean et median
- Performances faibles

MSE : 4959.626 || RMSE : 70.425 || MAE : 52.409 ||  $r^2$  : -0.0

MSE : 5576.848 || RMSE : 74.678 || MAE : 45.804 ||  $r^2$  : -0.124

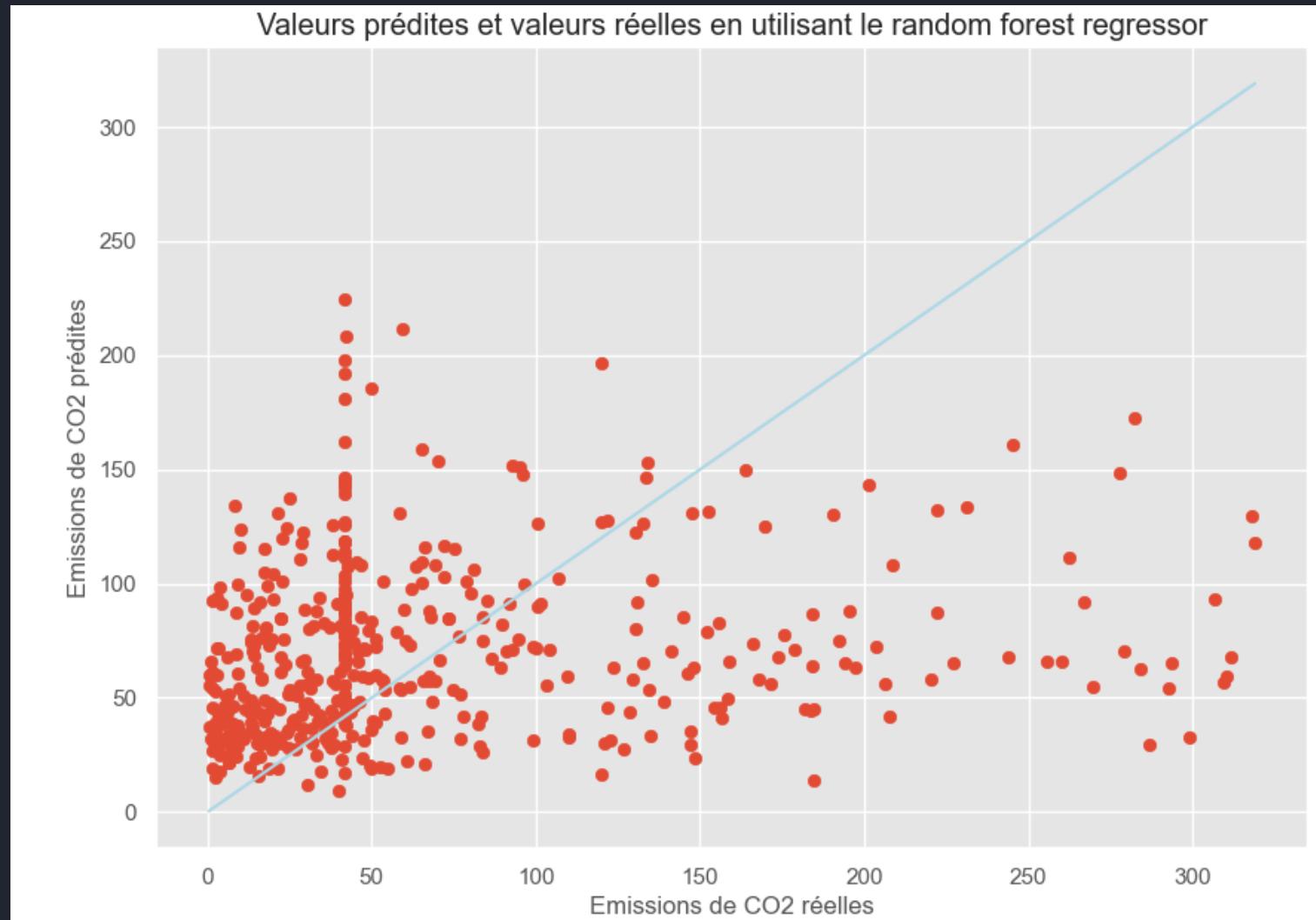
# Régression linéaire



- Meilleures performances
- Standardisation
- Représentation graphique

MSE : 4742.275 || RMSE : 68.864 || MAE : 51.21 ||  $r^2$  : 0.044

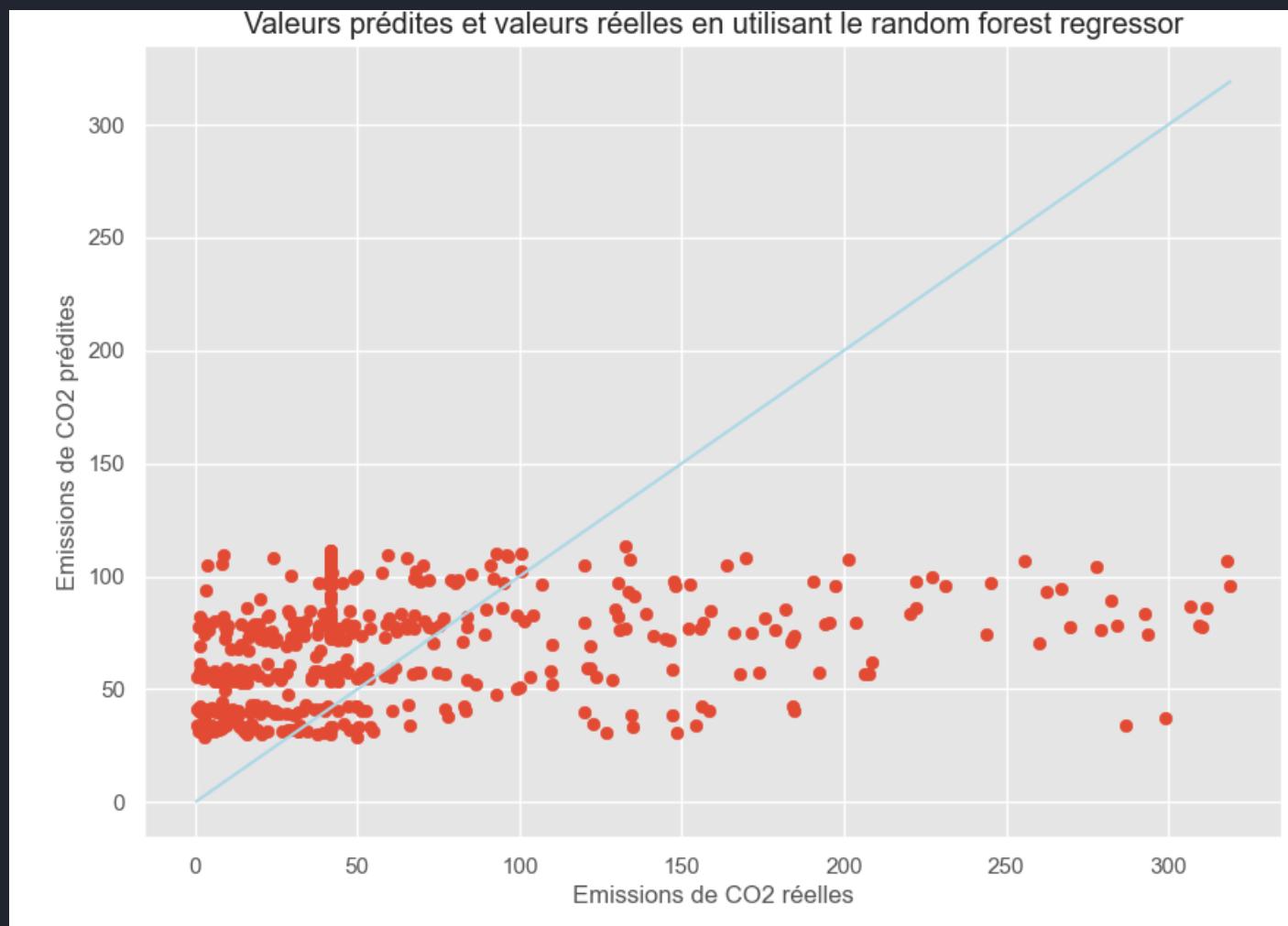
# Random forest Regressor



- Moins bonnes performances que régression linéaire
- Standardisation
- Représentation graphique

MSE : 5340.459 || RMSE : 73.078 || MAE : 52.98 ||  $r^2$  : -0.077

# Optimisation des performances



- Regression linéaire : grid search
- Random forest : grid search, random search, approche bayésienne
- Approche bayésienne retenue

MSE : 4541.514 || RMSE : 67.391 || MAE : 49.805 ||  $r^2$  : 0.084

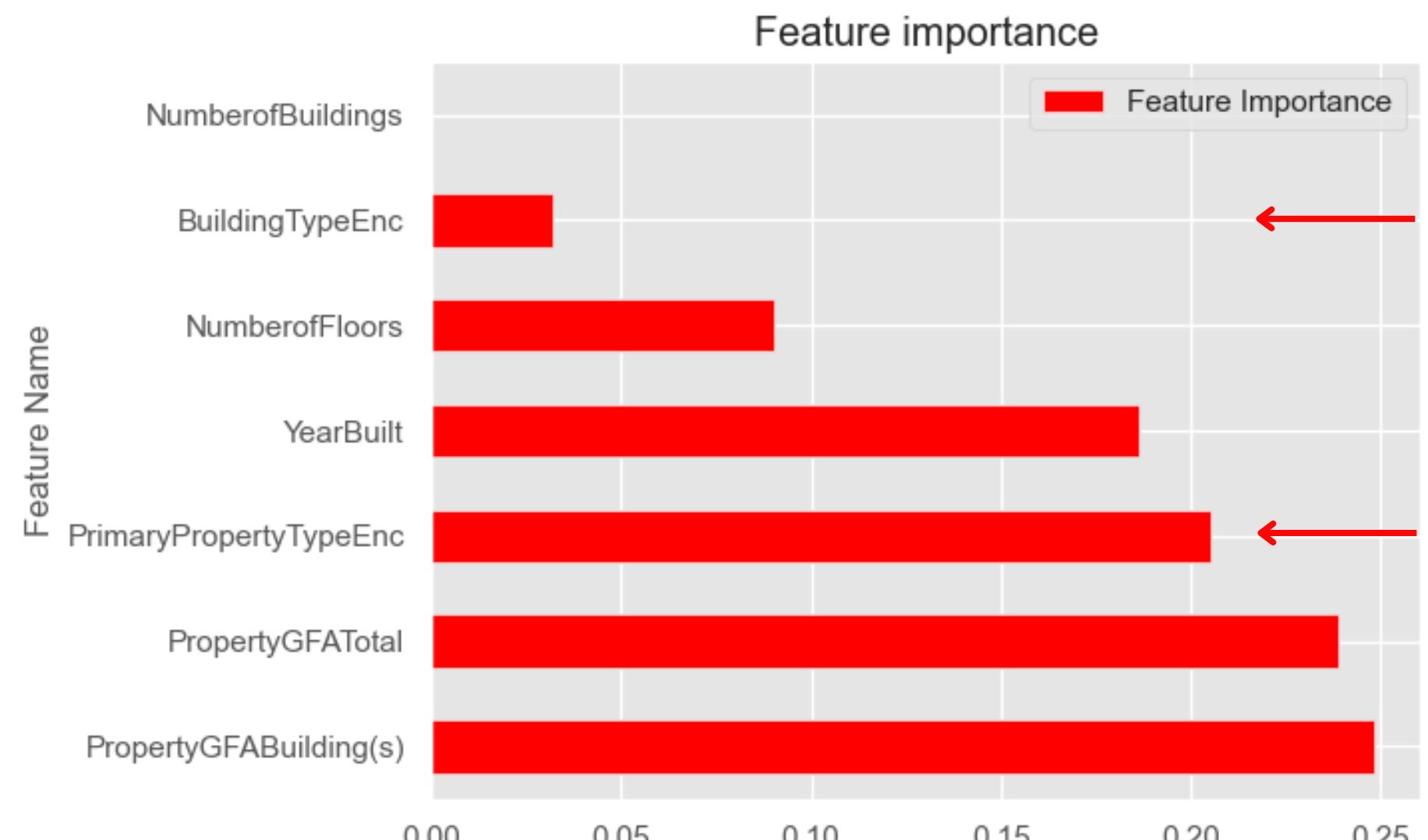
# Ajout de variables catégorielles

2 nouvelles variables : OneHotEncoder

Regression linéaire : pas de meilleure performance mais importance

Random forest regressor : meilleure performance et importance

Choix du modèle final : random search



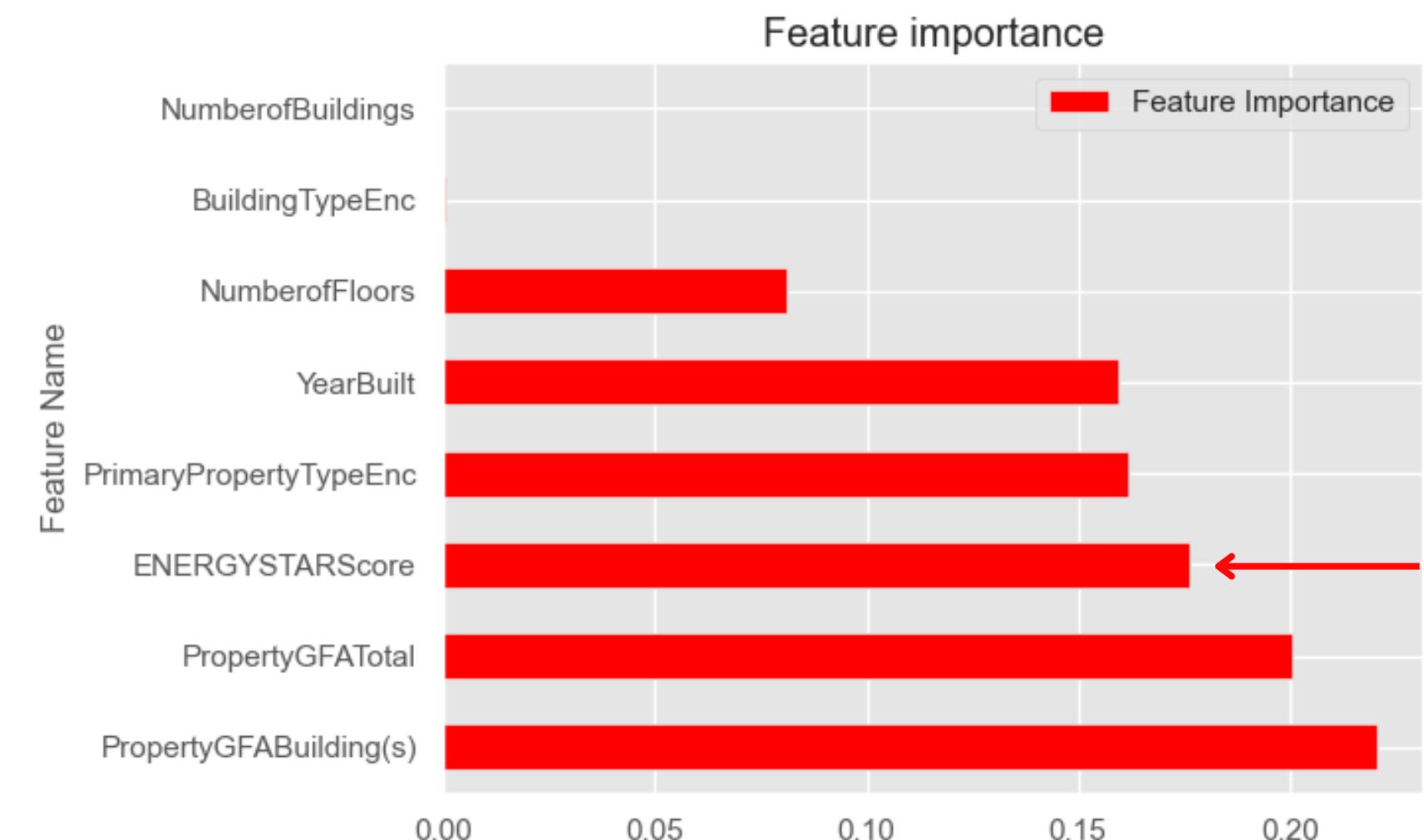
MSE : 4247.445 || RMSE : 65.172 || MAE : 46.662 ||  $r^2$  : 0.134

# Variable ENERGYSTARScore

Test avec et sans la variable

Amélioration des performances

Importance dans le modèle



Avec variable cible :

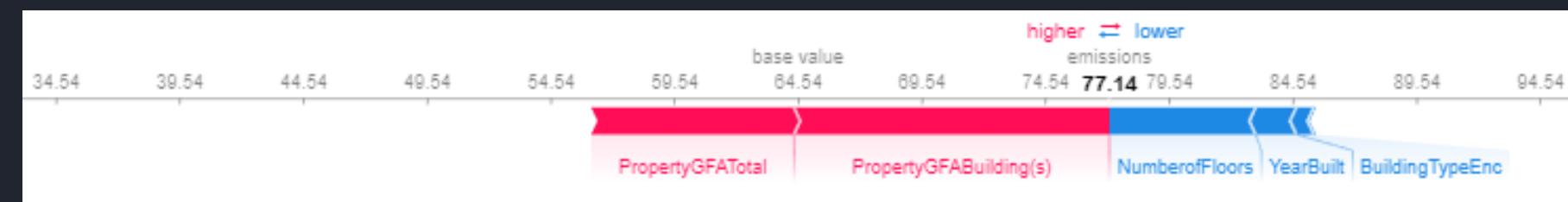
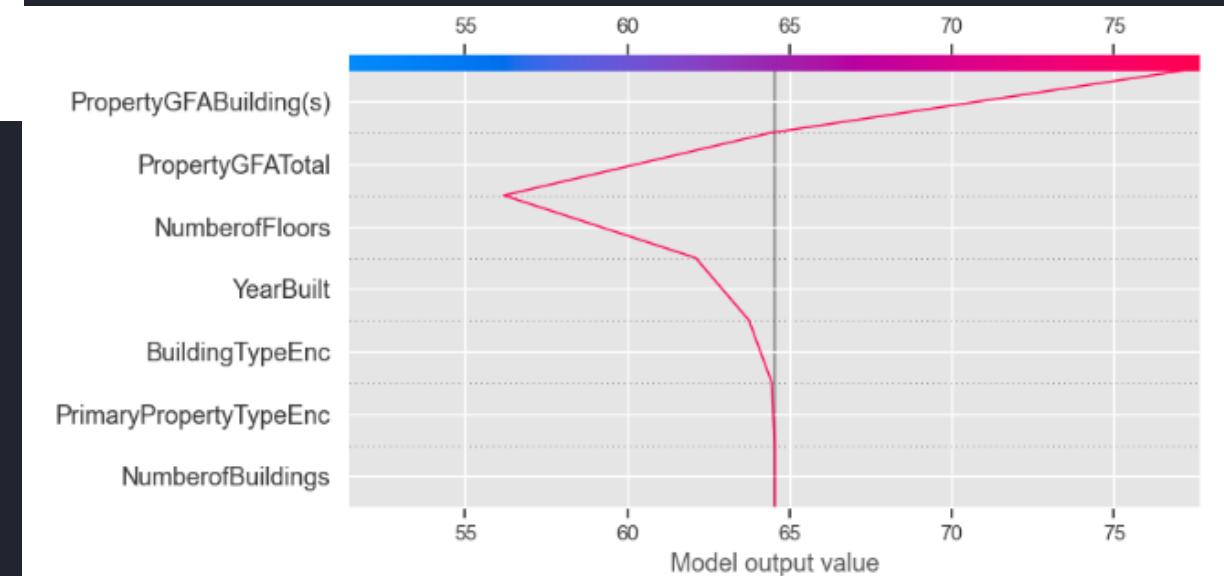
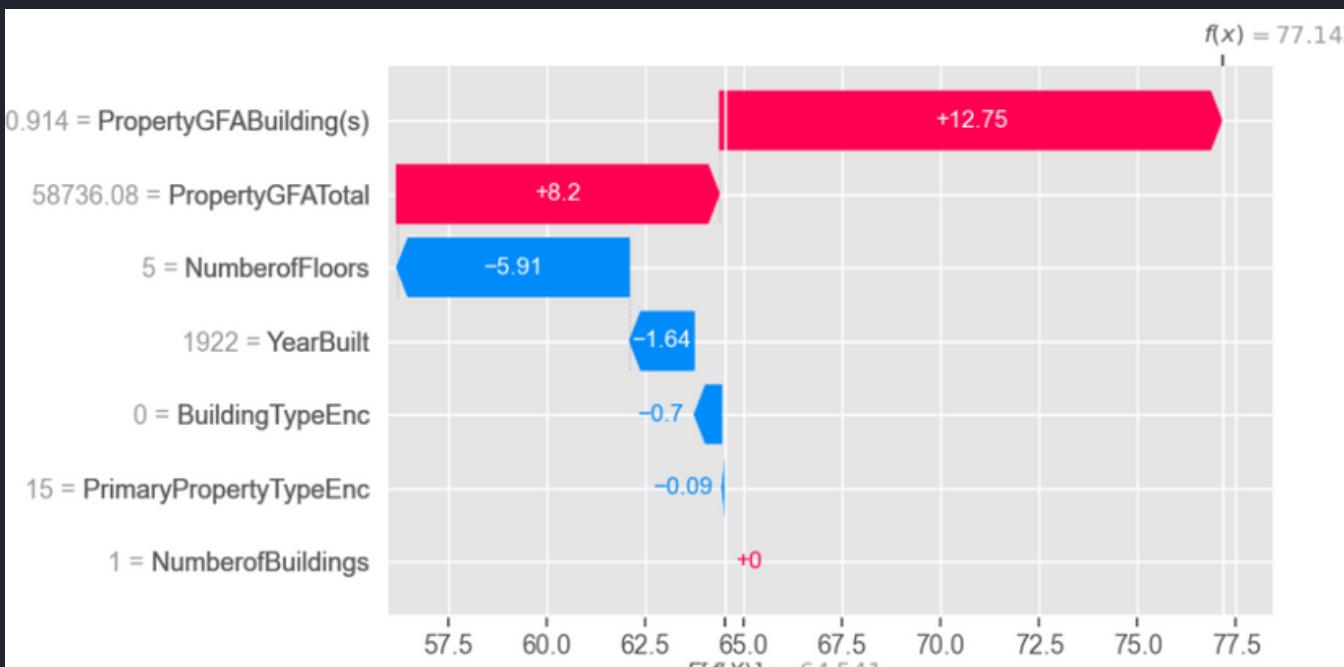
MSE : 3720.649 || RMSE : 60.997 || MAE : 44.862 ||  $r^2$  : 0.146

Sans variable cible :

MSE : 3793.585 || RMSE : 61.592 || MAE : 44.99 ||  $r^2$  : 0.129

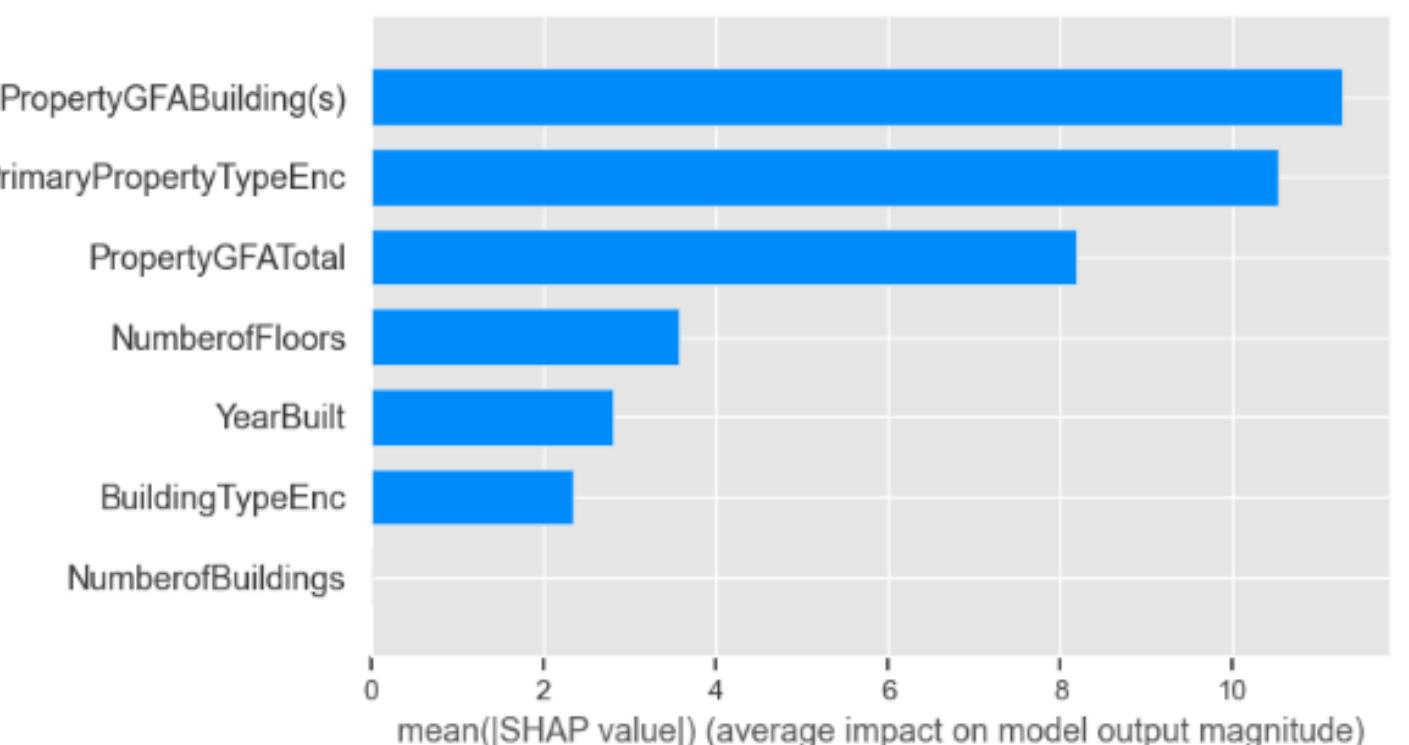
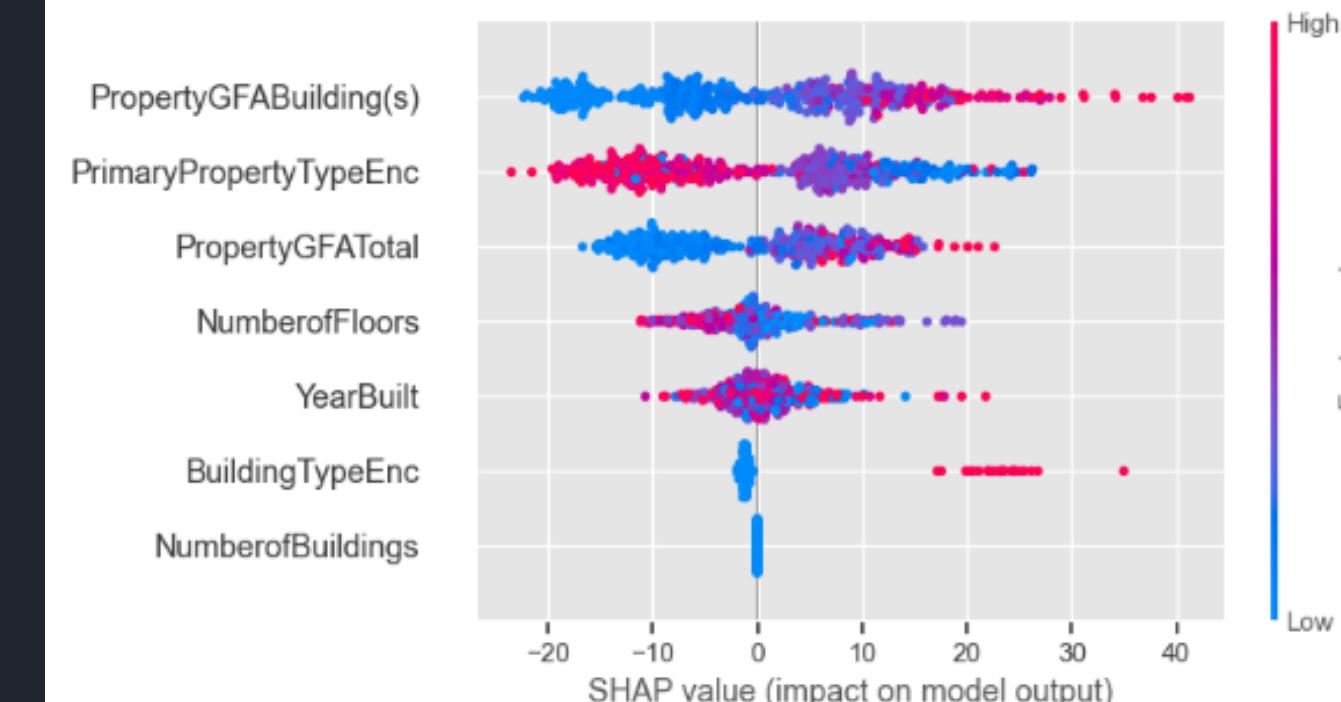
# IMPORTANCE VARIABLES AVEC SHAP

- Explication d'une instance
- Impact sur la prédiction
- Complémentarité des informations



# IMPORTANCE VARIABLES AVEC SHAP

Variables les plus importantes  
Amplitude de l'impact sur le modèle  
Légèrement différent de notre modèle



# CONSOMMATION ENERGIE

## Préparation et modélisation

Similaire à la première target

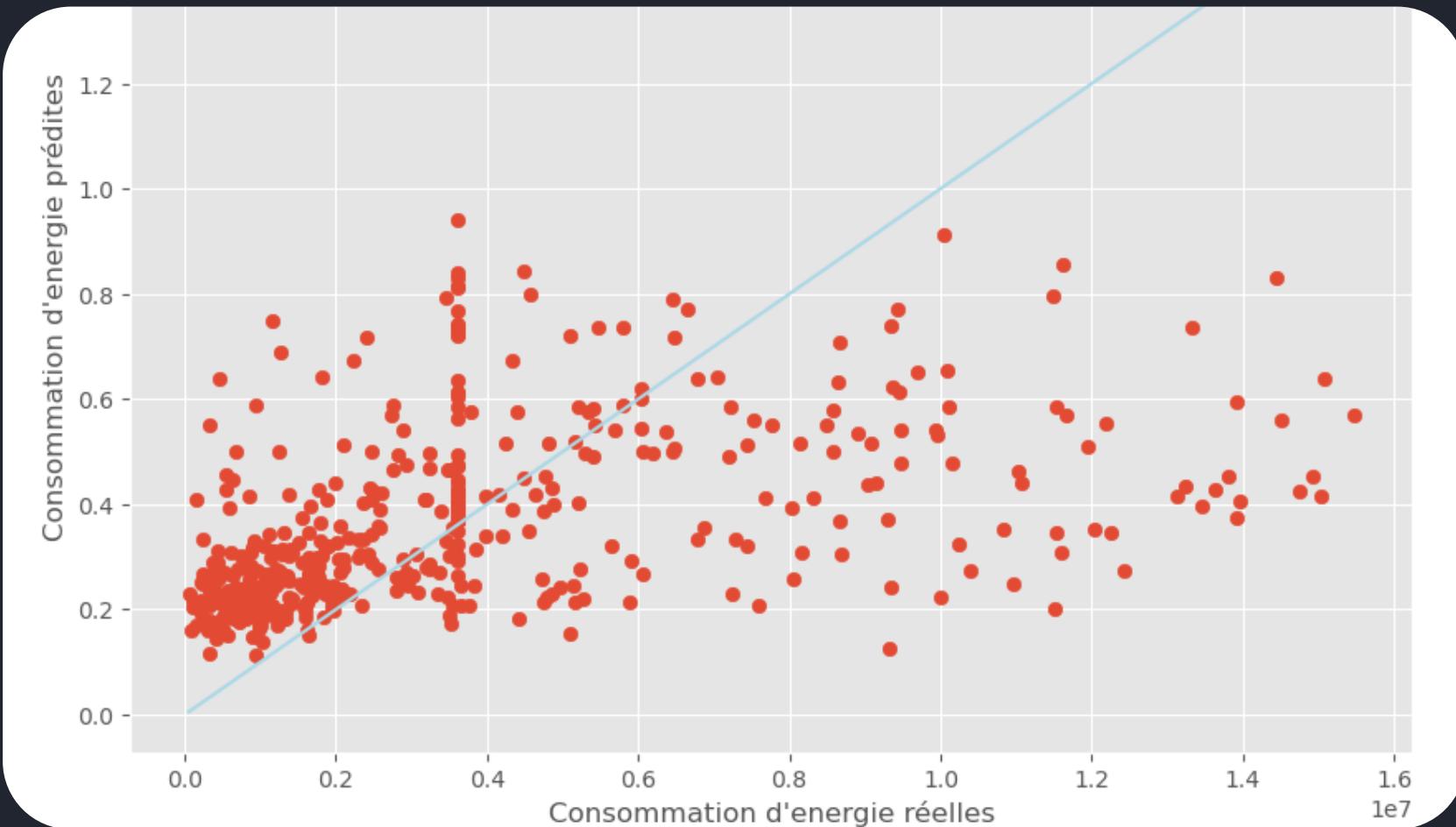
Validation croisée

Dummy regressor, Regression linéaire,  
Random forest regressor

Métriques de comparaison : MSE, RMSE,  
MAE, R<sup>2</sup>

LR: 8448178347465.656250 (1984991834265.125488)  
RF: 8869682367085.746094 (1673125402201.313721)  
DRMean: 11053126054081.332031 (1975385505488.013428)  
DRMed: 11928978583275.466797 (2430266779679.644531)

# Régression linéaire

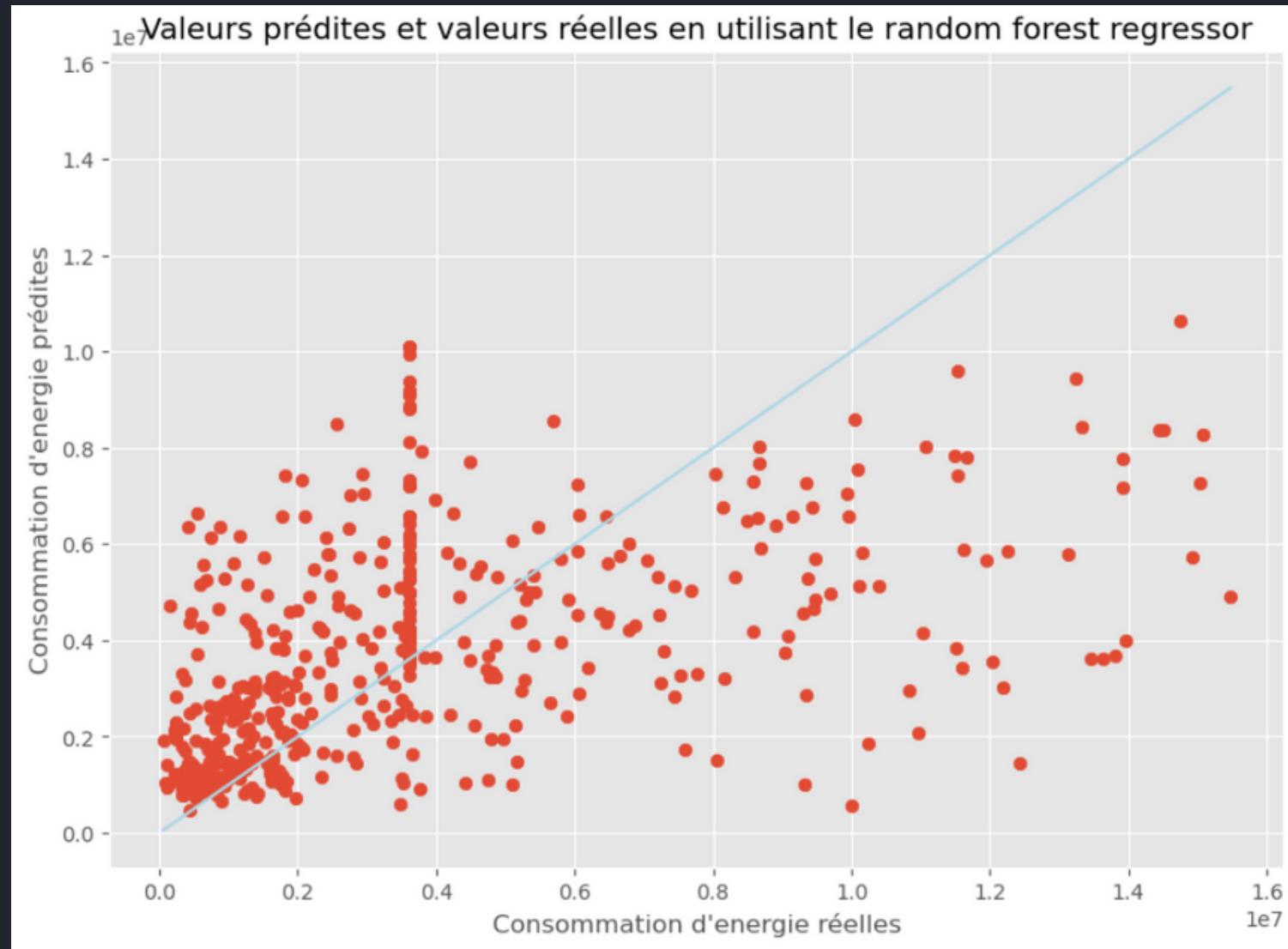


- Meilleures performances
- Standardisation
- Représentation graphique

MSE : 12600529628358.217 ||| RMSE : 3549722.472 ||| MAE : 2626273.028 |||  $r^2$  : -0.004

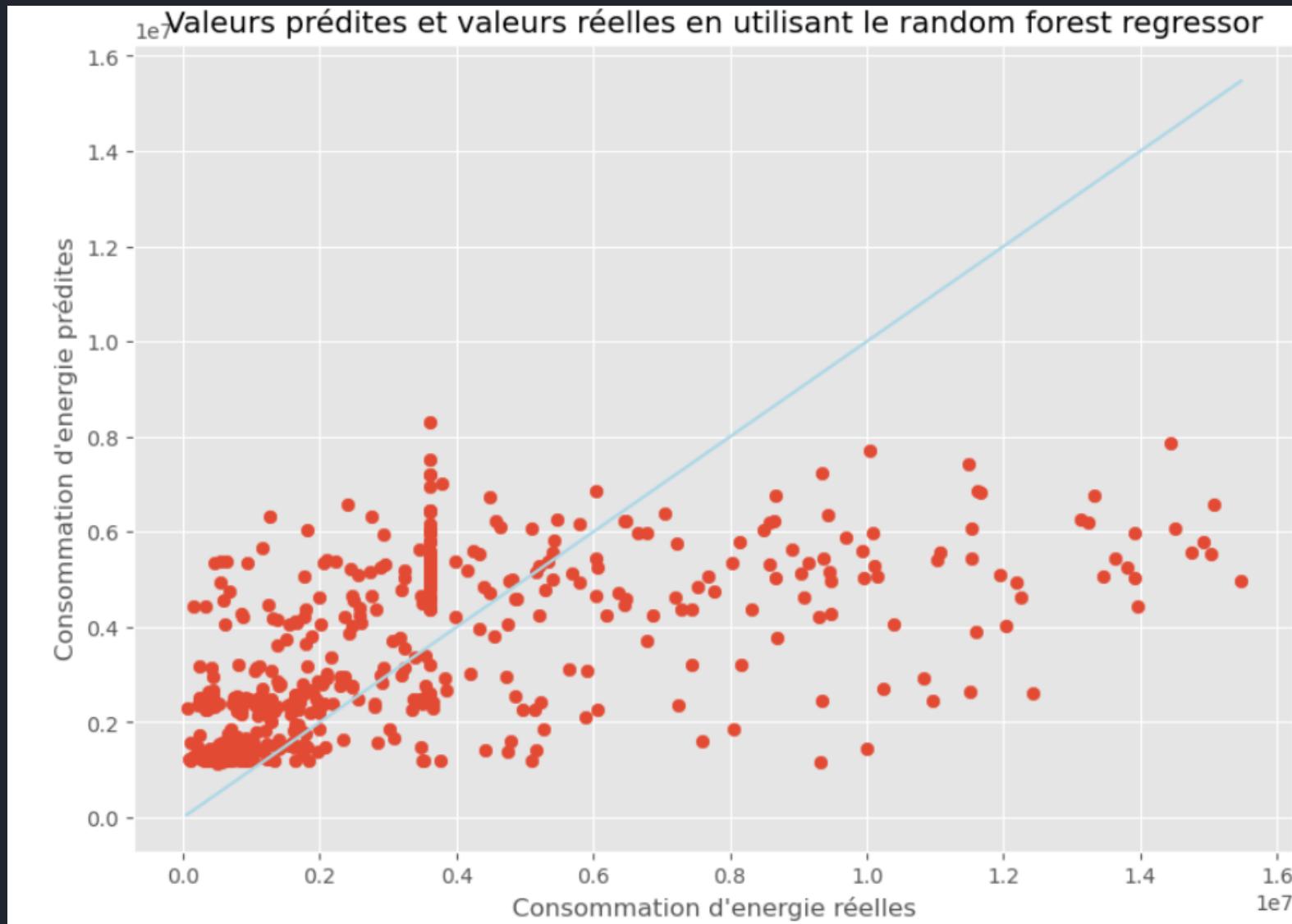
MSE : 9793757276633.795 ||| RMSE : 3129497.927 ||| MAE : 2219833.589 |||  $r^2$  : 0.219

# Random forest Regressor



- Meilleures performances que les 2 autres modèles
- Standardisation
- Représentation graphique

# Optimisation des performances



- Regression linéaire : grid search
- Random forest : grid search, random search, approche bayésienne
- Random search retenu

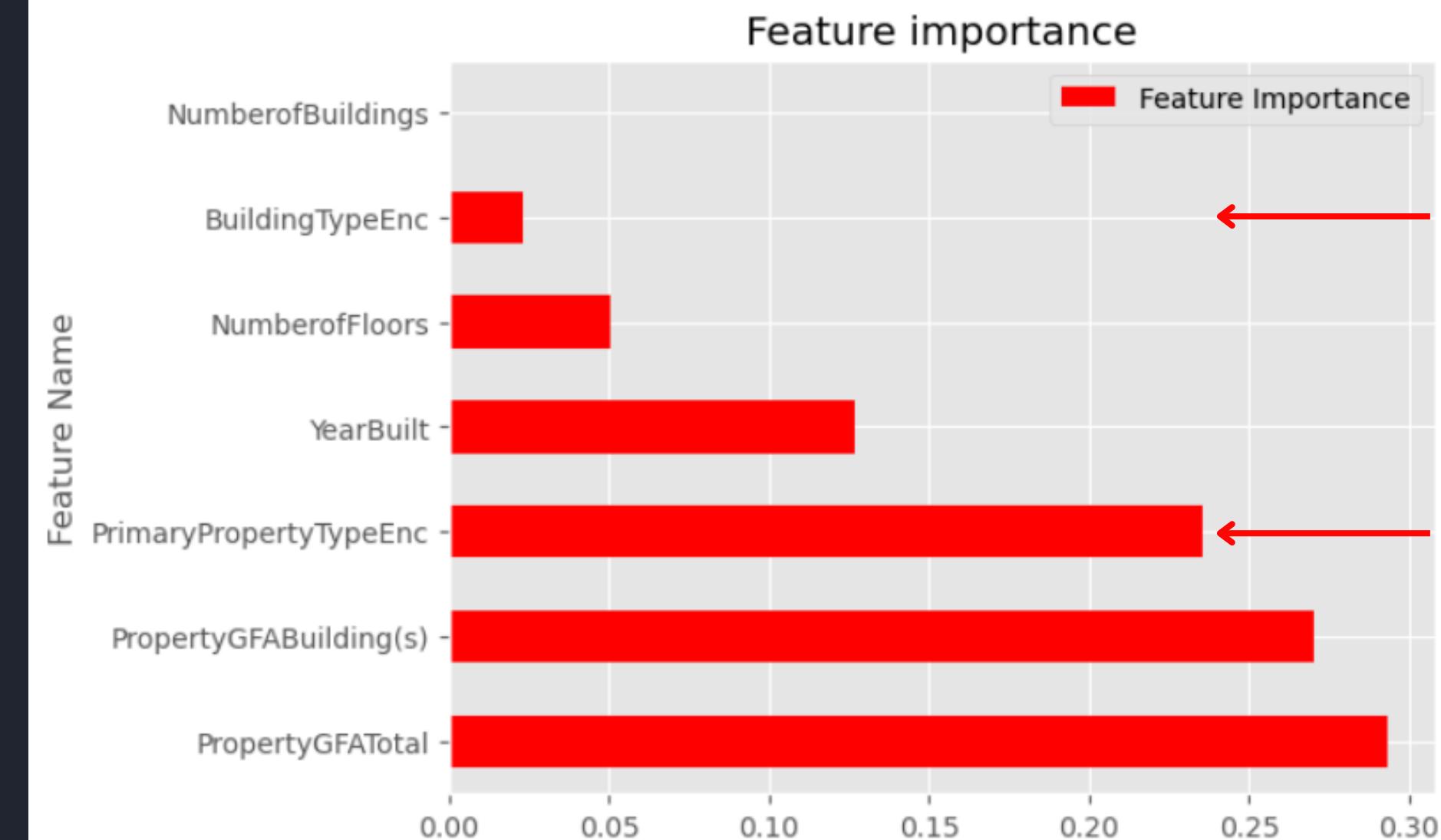
# Ajout de variables catégorielles

2 nouvelles variables : OneHotEncoder

Regression linéaire : meilleure performance et grande importance

Random forest regressor : meilleure performance et importance

Choix du modèle final : approche bayésienne



MSE : 5922203119348.72 || RMSE : 2433557.708 || MAE : 1677564.583 ||  $r^2$  : 0.406

# Variable ENERGYSTARScore

Test avec et sans la variable  
Amélioration des performances  
Importance moyenne dans le modèle



**Avec variable cible :**

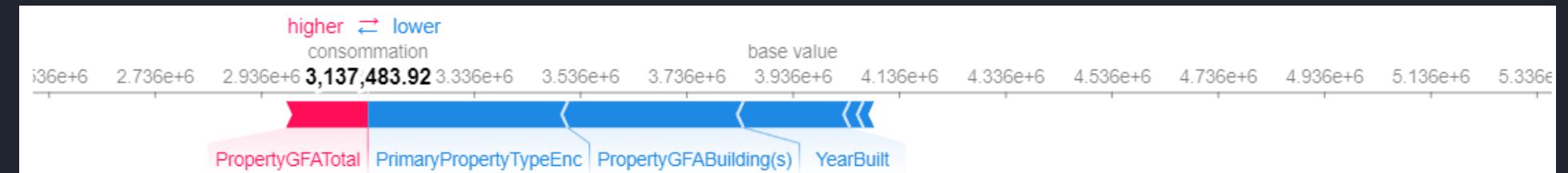
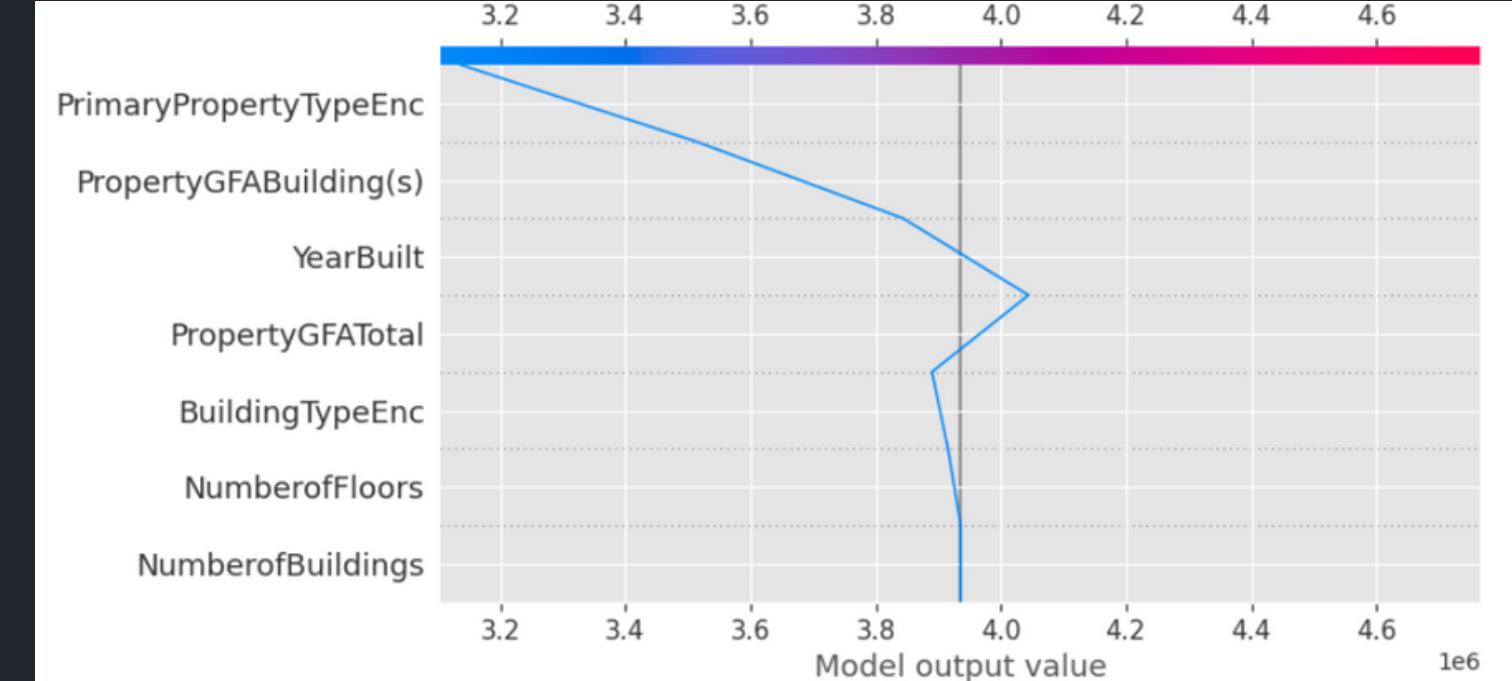
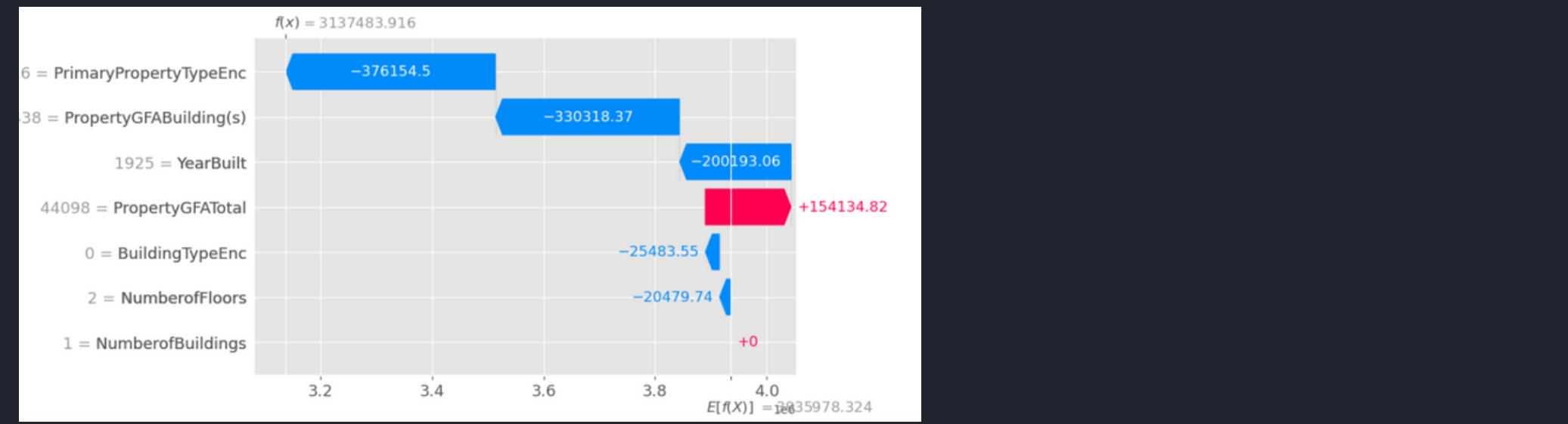
MSE : 5555930227237.476 || RMSE : 2357102.082 || MAE : 1559127.716 ||  $r^2$  : 0.517

**Sans variable cible :**

MSE : 5935368808447.453 || RMSE : 2436261.236 || MAE : 1652060.403 ||  $r^2$  : 0.484

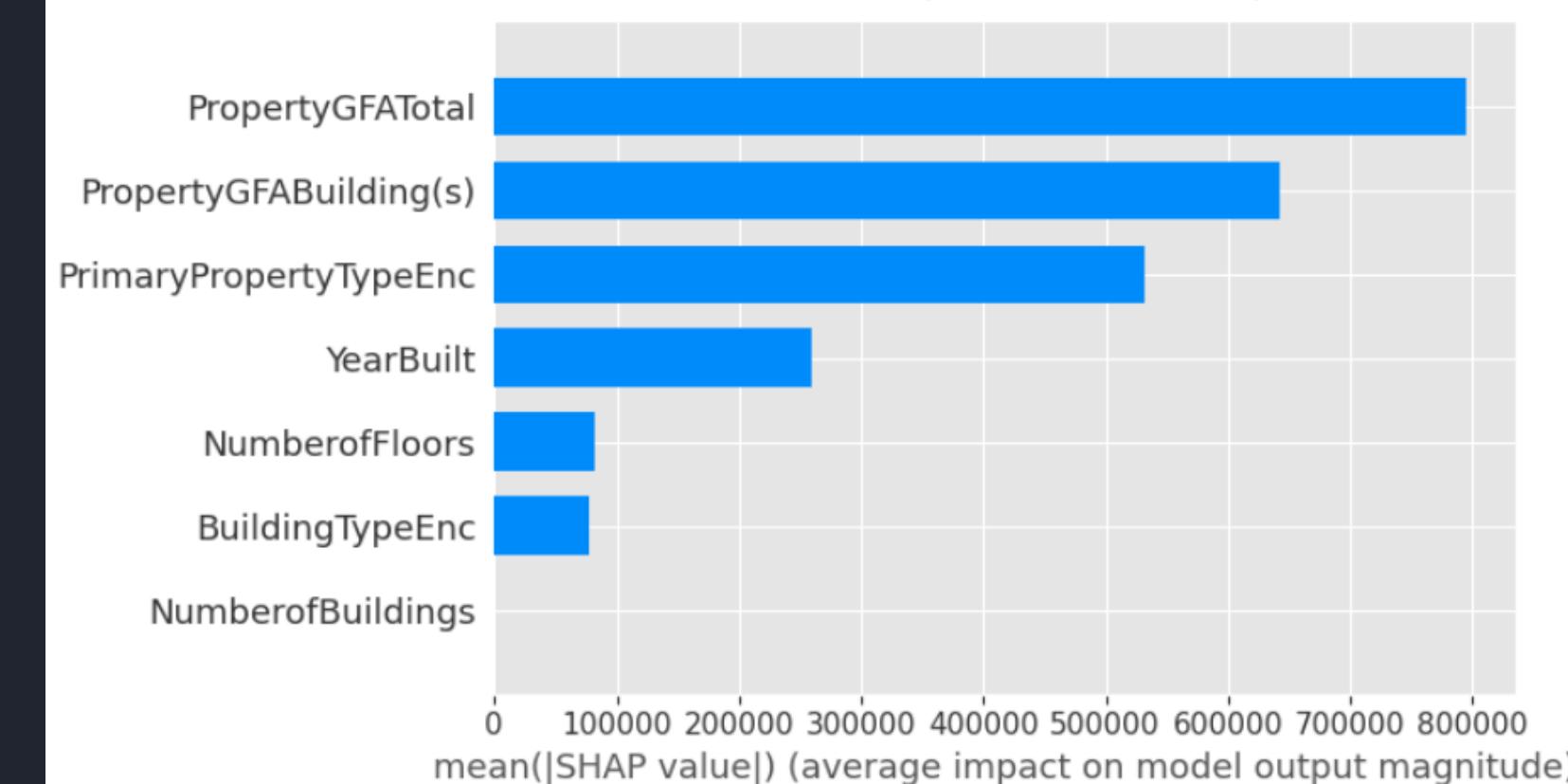
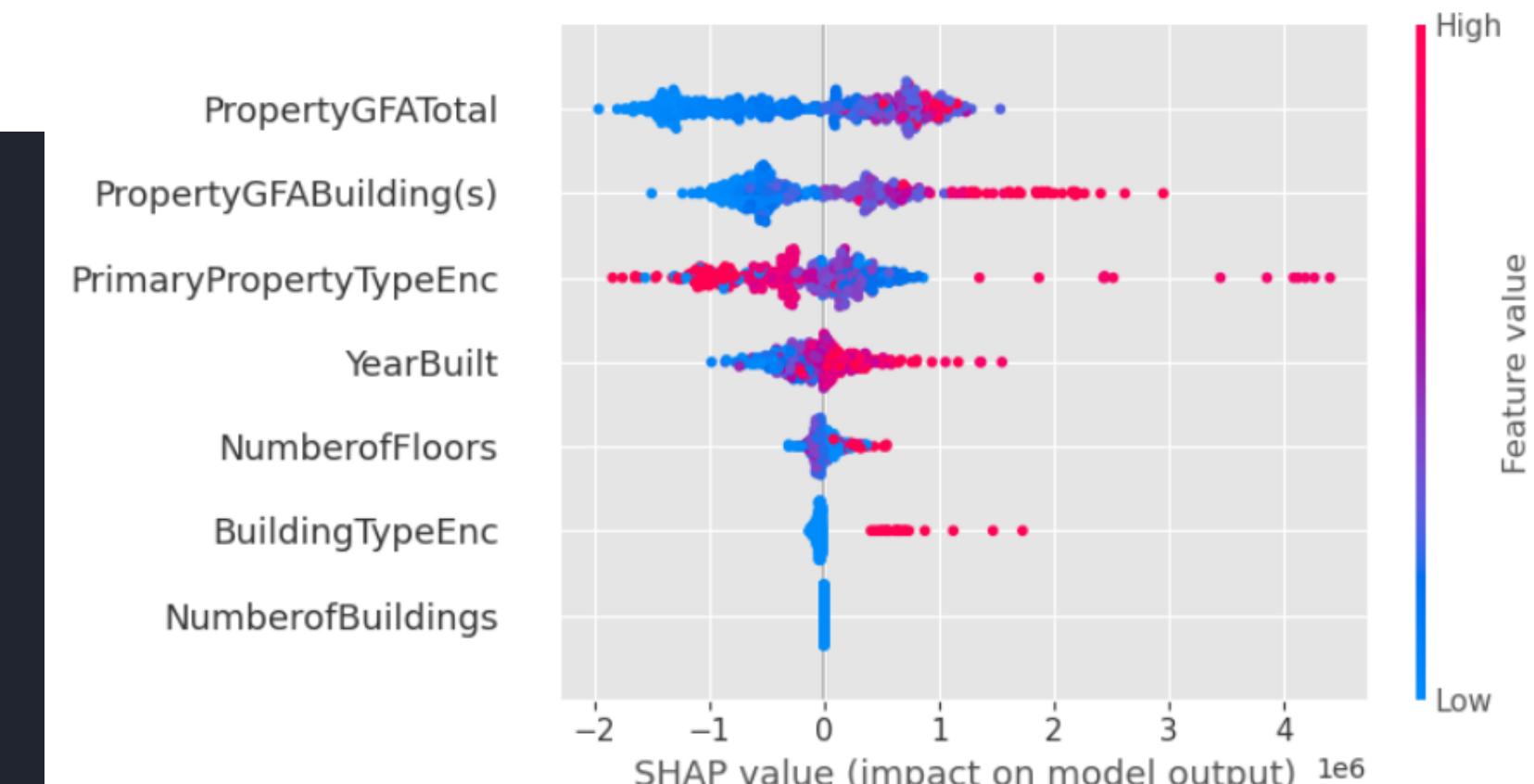
# IMPORTANCE VARIABLES AVEC SHAP

- Explication d'une instance
- Impact sur la prédiction
- Complémentarité des informations
- Impact principalement négatif



# IMPORTANCE VARIABLES AVEC SHAP

Variables les plus importantes  
Amplitude de l'impact sur le modèle  
Légèrement différent de notre modèle



A large, modern building with a glass and steel facade, viewed from a low angle looking up. The building has a curved, angular design with many windows. The sky is overcast.

**MERCI POUR VOTRE  
ATTENTION**