Modélisation des besoins en consommation de bâtiments de la ville de Seattle

Présenté par Gabriel Chehade

Problématique

Contexte :

En 2016, des agents de la ville de Seattle ont effectué des relevés minutieux de la consommation et des émissions de CO2 des bâtiments.

Inconvénient : ces relevés sont coûteux à obtenir.

Problématique : Peut-on prévoir ces données pour les bâtiments qui n'ont pas été étudiés à partir des données existantes?

Objectifs :

- Modéliser la consommation énergétique et les émissions de gaz à effet de serre (GES) des bâtiments de la ville de Seattle non destinés à l'habitation.
- Évaluer l'intérêt de l'ENERGY STAR Score

Présentation du jeu de données

Source des données :
https://data.seattle.gov/dataset/2016-Building-Energy-Benchmarking/2bpz-gwpy

Dimensions : 3376 lignes × 46 colonnes. Chaque ligne représente un bâtiment.

On va retrouver des informations sur :

- Le lieu (adresse, position géographique, district)
- Les types d'usage du bâtiment (Bureaux, hôtel, entrepôt, ...)
- La taille (nombre d'étages, surface)
- La consommation énergétique
- La quantité de GES émise
- La nature des sources d'énergie (électricité, gaz, vapeur)

Nettoyage du jeu de données

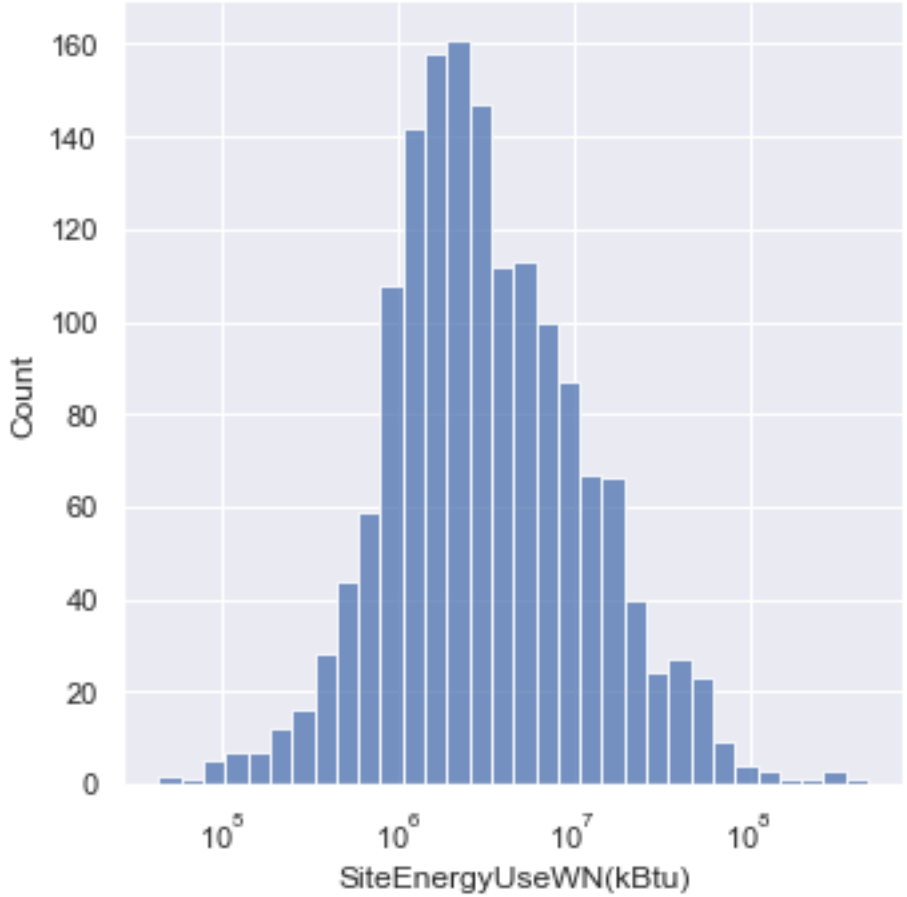
- Suppression des variables :
  - constantes (3 colonnes)
  - vides à plus de 90%
  - répétées (Electricity(kWh), Electricity(kBtu))
  - non normalisées WN*
  - Fortement corrélées à d'autres variables

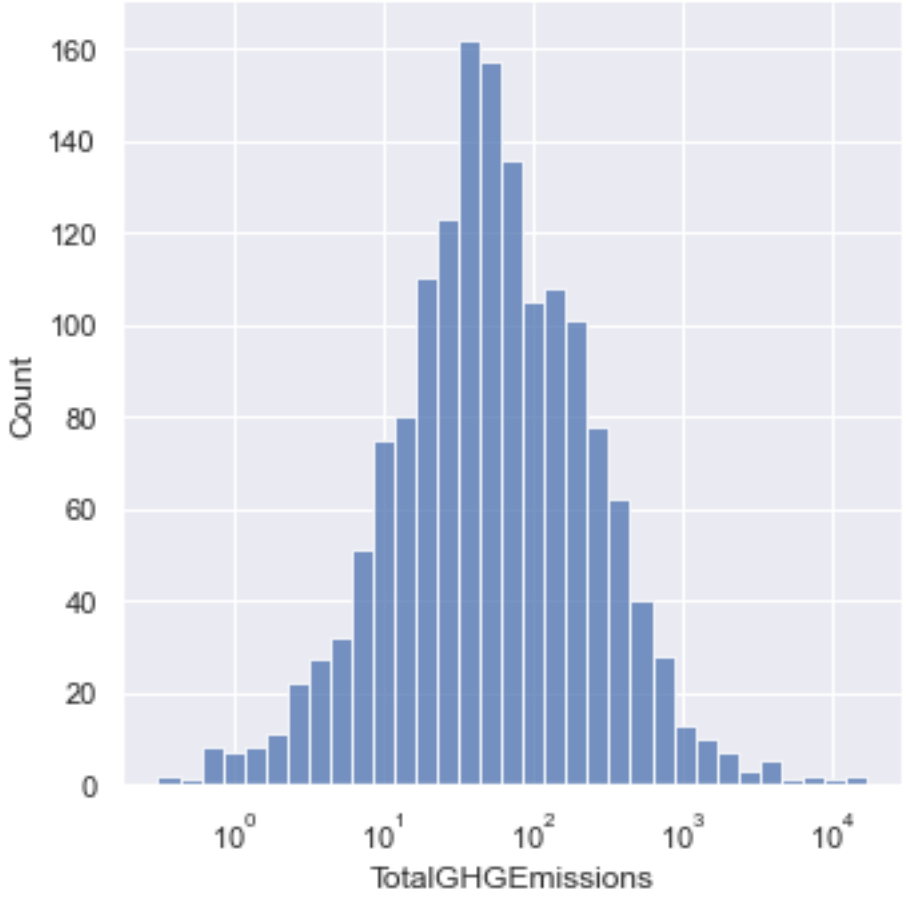
- Suppression des bâtiments destinés à l'habitation
- Imputation des valeurs manquantes sauf pour l'ENERGYSTARScore

→ 1578 lignes, 17 colonnes

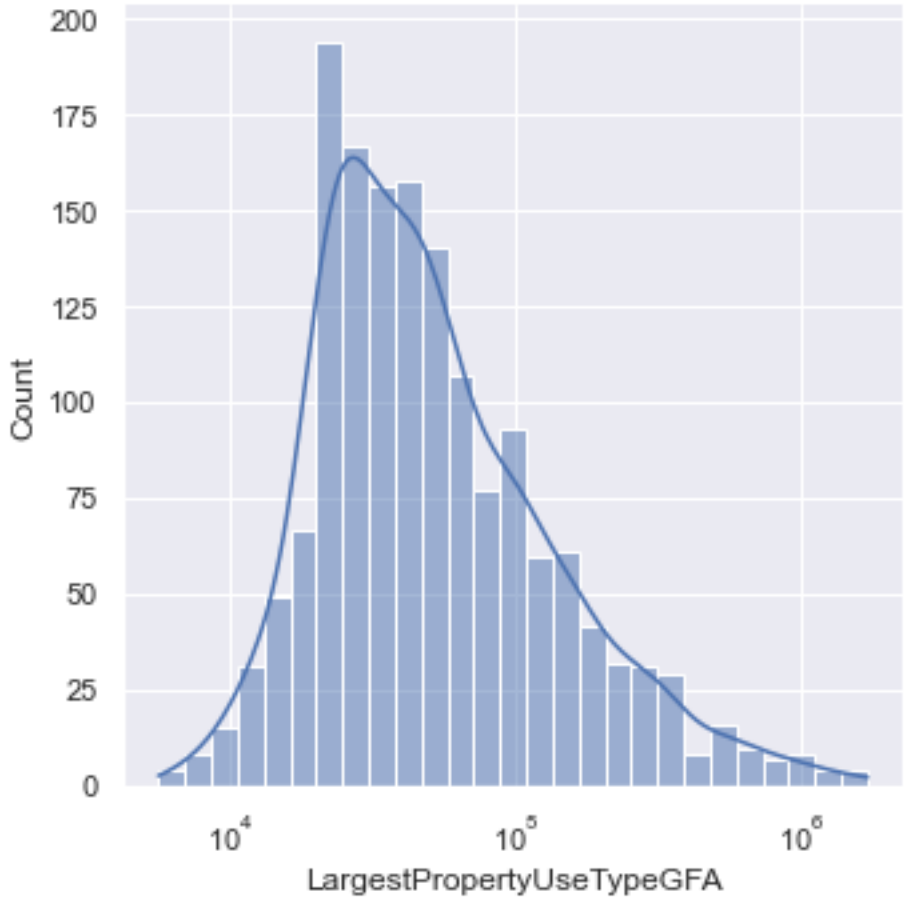
*WN : Weather-Normalized. Valeur moyenne que prendrait la variable sur une période de 30 ans.

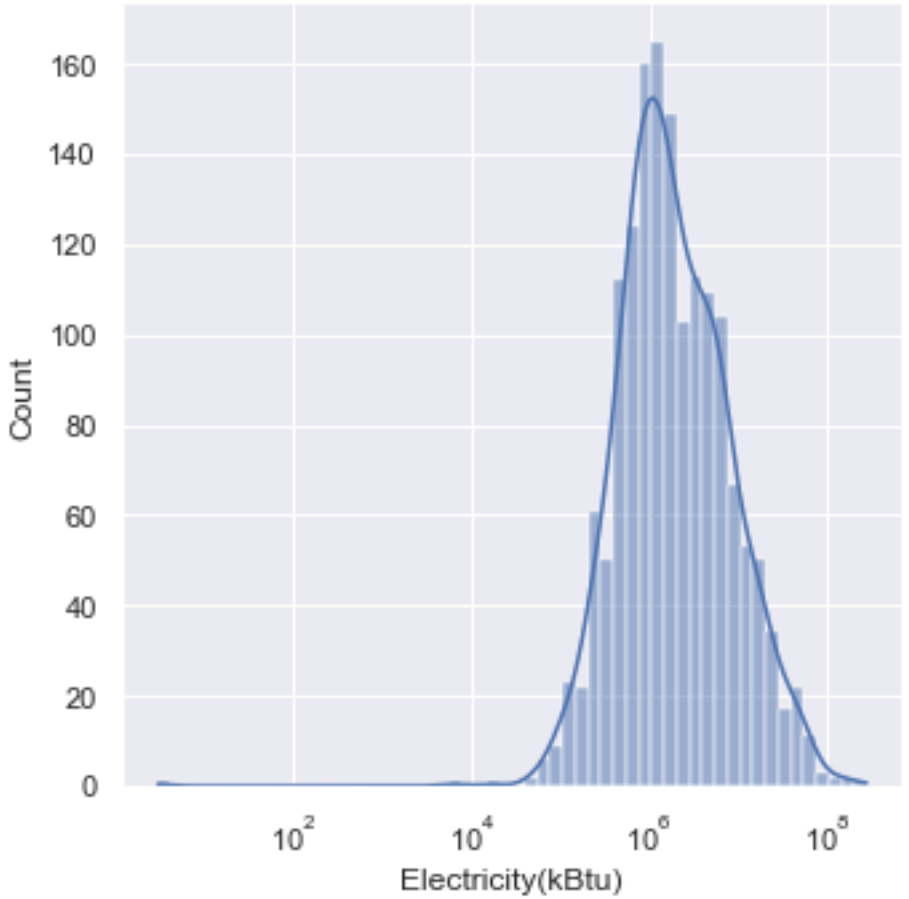
Variables à modéliser (targets)





Asymétrie des variables continues





Feature Engineering

- Suppression des variables énergétiques
- l'ENERGYSTARScore : dropna ou fillna par la médiane ou la moyenne
- Regroupement de types d'usage du bâtiment
- Encodage : OneHotEncoding, TargetEncoding
- Si OneHotEncoding : Regroupement des variables de surface en une seule nommée TotalGFA
- Sélection de variables (RFE ou SelectKBest)
- Transformation en log(1+X) pour les variables continues
- Scaling : StandardScaler ou MinMaxScaler

Feature Engineering

- Suppression des variables énergétiques
- l'ENERGYSTARScore : dropna ou fillna par la médiane ou la moyenne
- Regroupement de types d'usage du bâtiment
- Encodage : **OneHotEncoding**, TargetEncoding
- Si OneHotEncoding : Regroupement des variables de surface en une seule nommée TotalGFA
- Sélection de variables (RFE ou SelectKBest)
- Transformation en log(1+X) pour les variables continues
- Scaling : StandardScaler ou MinMaxScaler

Usage Principal	Usage Principal GFA	Usage secondaire	Usage Secondaire GFA
Bureau	400 000	Parking	100 000

Bureau	Parking
0.80	0.20

Feature Engineering

- Suppression des variables énergétiques
- l'ENERGYSTARScore : dropna ou fillna par la médiane ou la moyenne
- Regroupement de types d'usage du bâtiment
- Encodage : OneHotEncoding, **TargetEncoding**
- Si OneHotEncoding : Regroupement des variables de surface en une seule nommée TotalGFA
- Sélection de variables (RFE ou SelectKBest)
- Transformation en log(1+X) pour les variables continues
- Scaling : StandardScaler ou MinMaxScaler

LargestPropertyUseType
Warehouse
Retail Store
Office
Office
Retail Store
K-12 School
Self-Storage Facility
Office

LargestPropertyUseType
2.29e+06
4.26e+06
9.13e+06
9.13e+06
4.26e+06
3.63e+06
6.65e+06
9.13e+06

Feature Engineering

- Suppression des variables énergétiques
- l'ENERGYSTARScore : dropna ou fillna par la médiane ou la moyenne
- Regroupement de types d'usage du bâtiment
- Encodage : OneHotEncoding, TargetEncoding
- **Si OneHotEncoding : Regroupement des variables de surface en une seule nommée TotalGFA**
- Sélection de variables (RFE ou SelectKBest)
- Transformation en log(1+X) pour les variables continues
- Scaling : StandardScaler ou MinMaxScaler

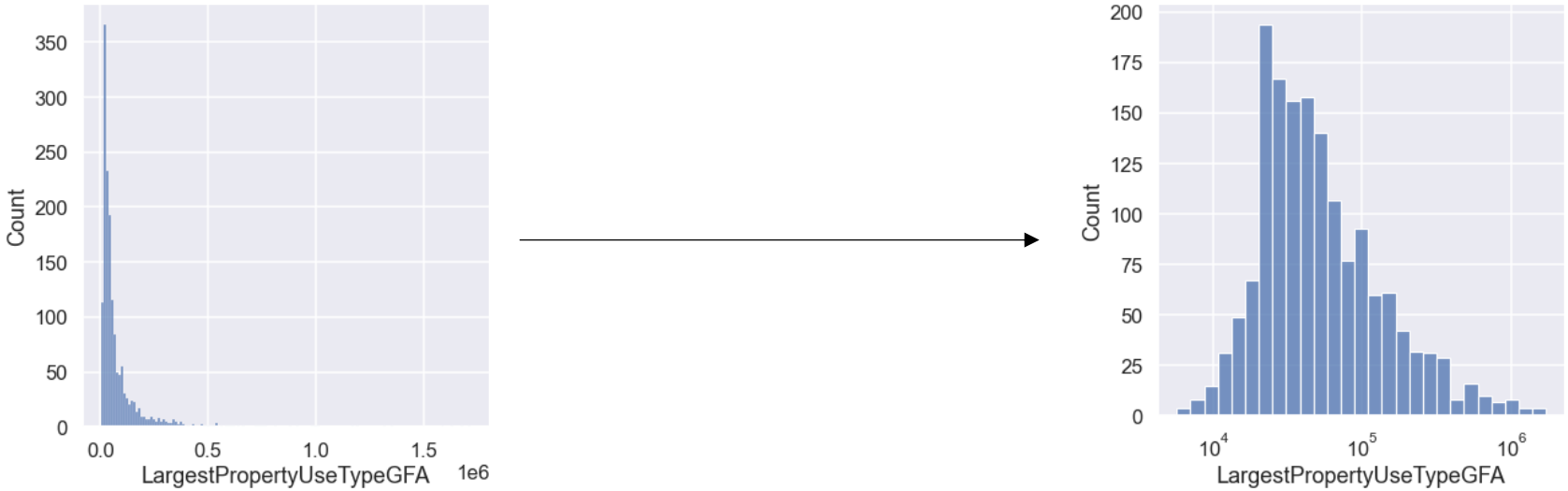
Usage Principal GFA	Usage Secondaire GFA	Usage tertiaire GFA
900 000	90 000	10 000

→

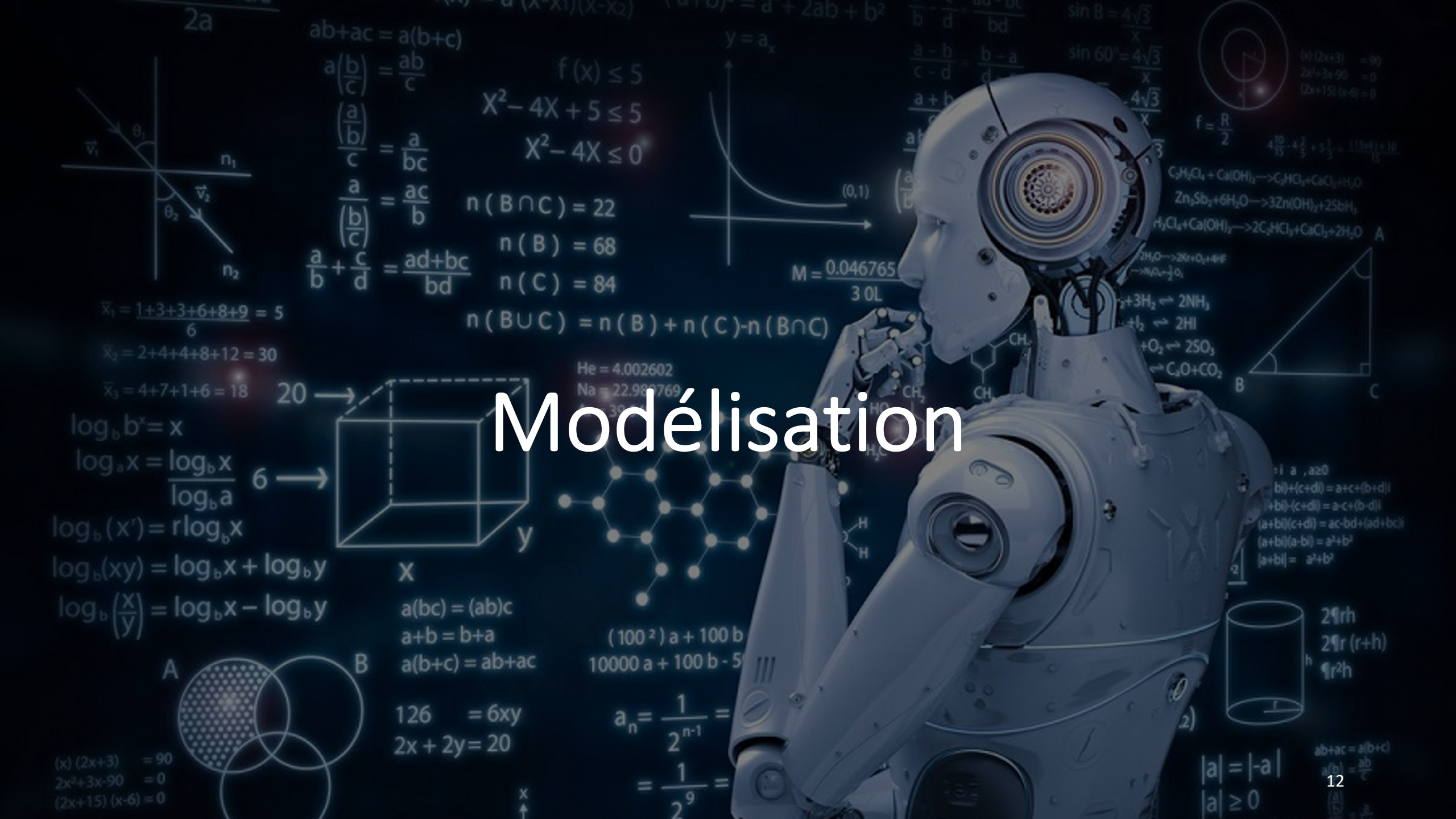
TotalGFA
1 000 000

Feature Engineering

- Suppression des variables énergétiques
- l'ENERGYSTARScore : dropna ou fillna par la médiane ou la moyenne
- Regroupement de types d'usage du bâtiment
- Encodage : OneHotEncoding, TargetEncoding
- Si OneHotEncoding : Regroupement des variables de surface en une seule nommée TotalGFA
- Sélection de variables (RFE ou SelectKBest)
- **Transformation en log(1+X) pour les variables continues**
- Scaling : StandardScaler ou MinMaxScaler



Modélisation



Modèles

Modèles de base :

- KNN (KNeighboorsRegressor)
- SGD (SGDRegressor)
- SVR
- KernelRidge

Méthodes ensemblistes :

- RandomForest
- Bagging (BaggingRegressor)
- Boosting (Adaboost et GradientBoosting)
- Voting

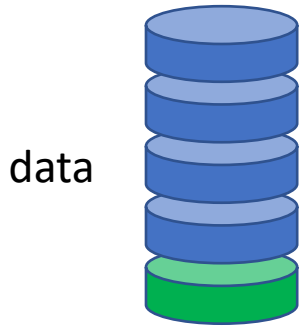
Méthode

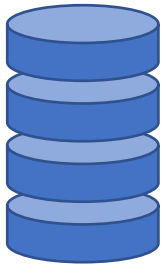
GridSearchCV

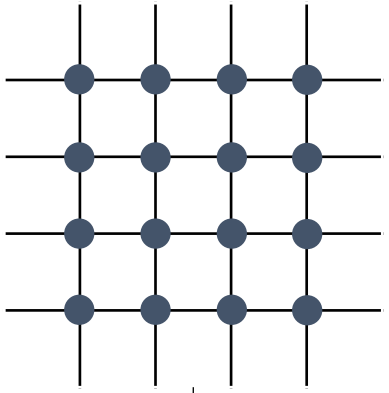
Cross-Validation

Train set

data







Test set

model.predict(X_{test})

Learning Curve

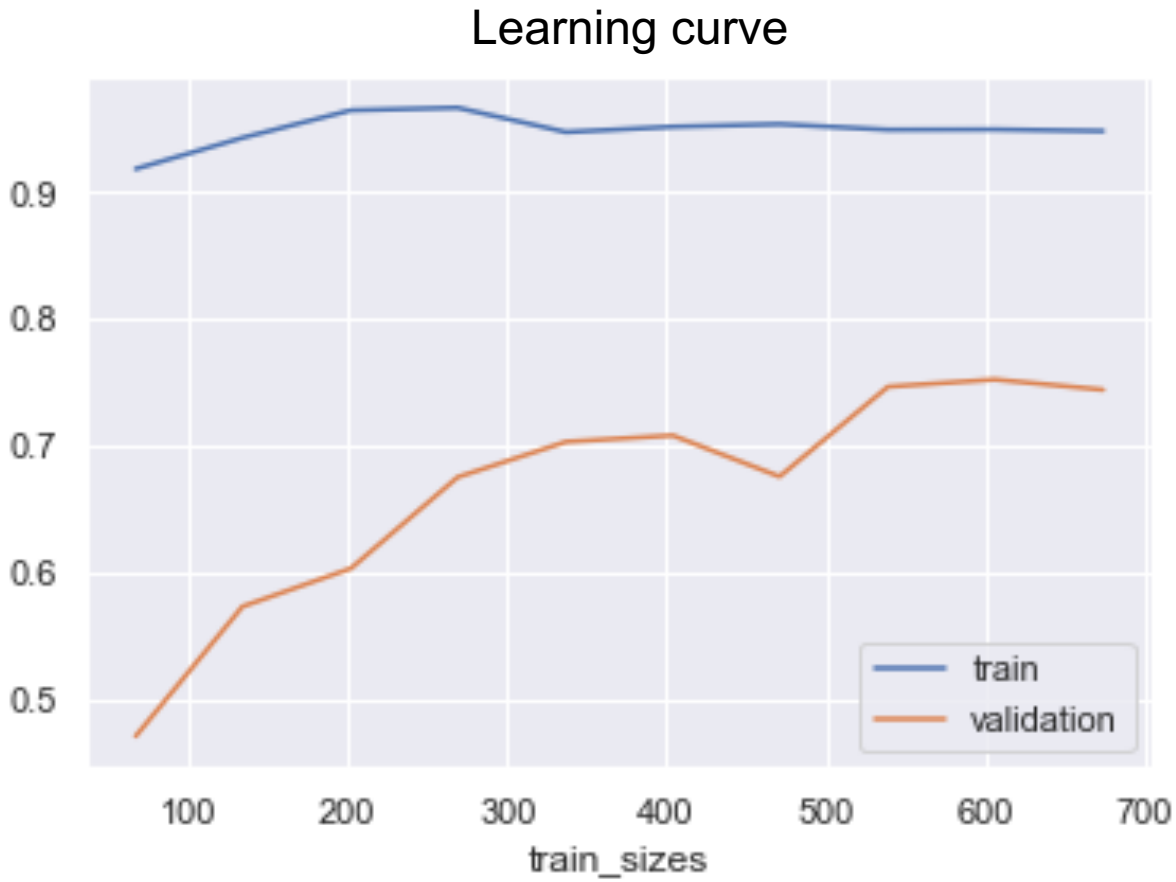
Target 1 : Consommation d'énergie

Feature Engineering :
- dropna()
- OneHotEncoding
- StandardScaler

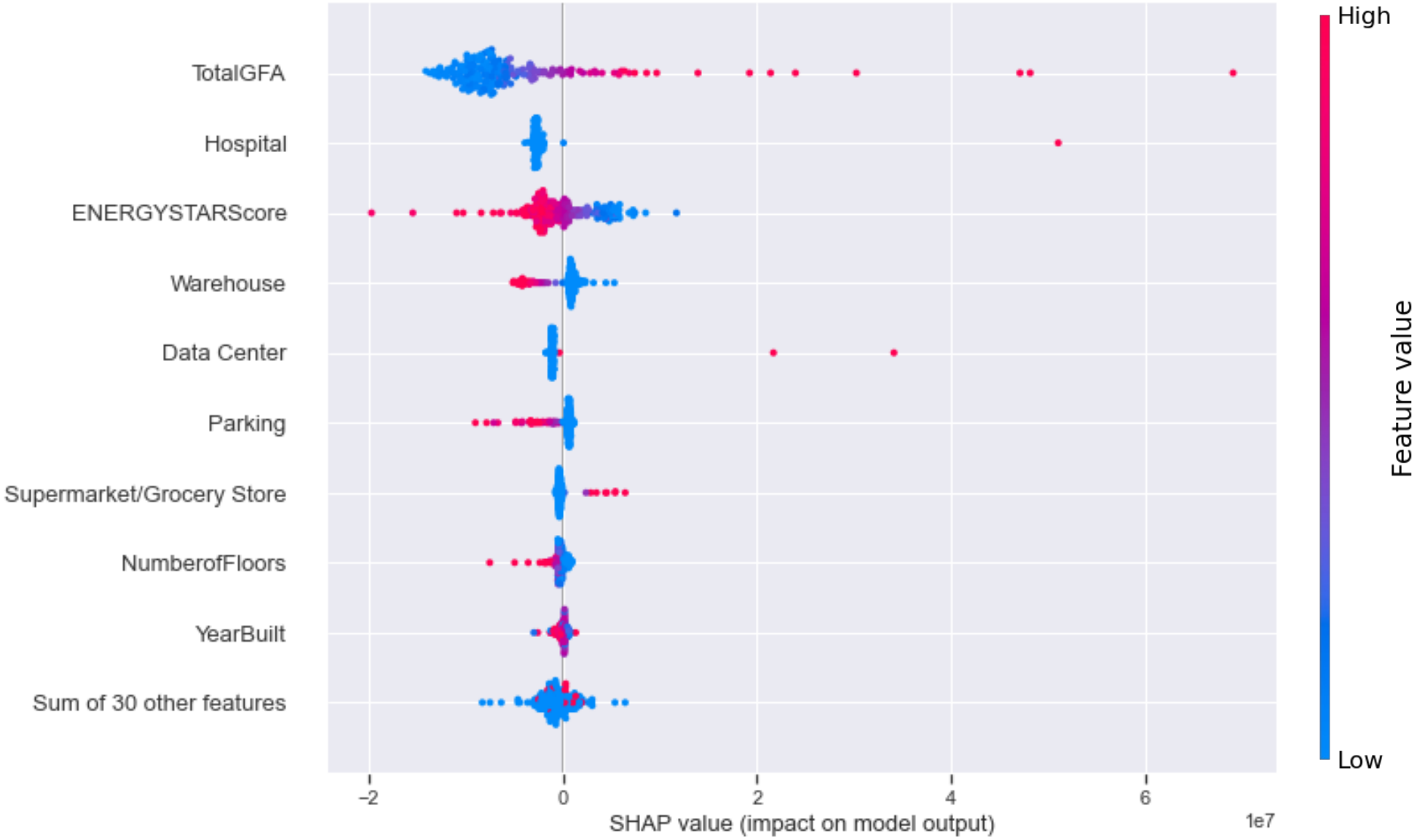
Modèle final : **BaggingRegressor**
- base_estimator = **KernelRidge**
  - alpha = 0.01
  - kernel = 'poly'
  - degree = 2
- n_estimators = 50

Résultats
- Cross-validation :
  - R^2 moyen : **0.74**
  - Écart-type : 0.13

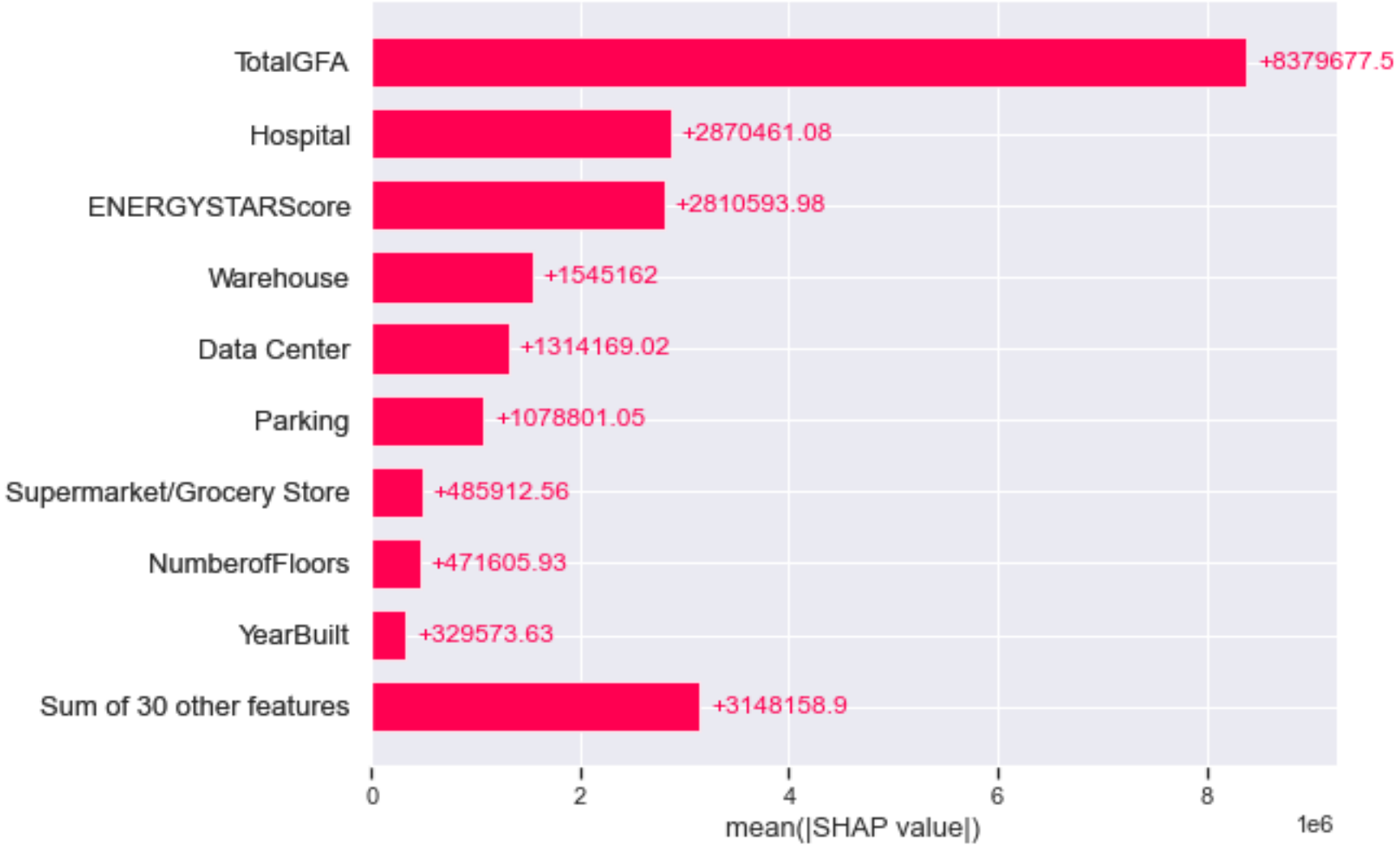
- Test set : **0.76**



Feature Importance



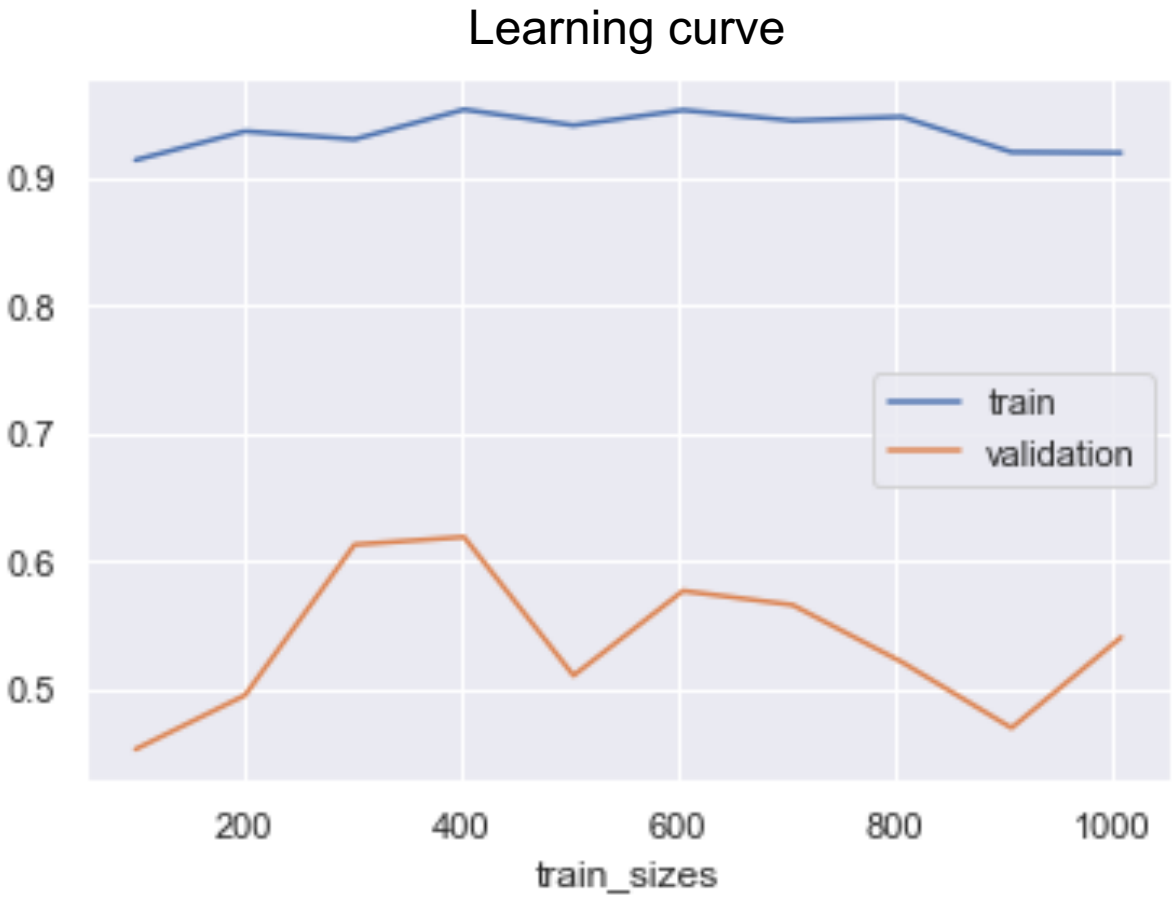
Feature Importance



Sans l'ENERGYSTARScore

En reprenant le même modèle :

- Cross-validation :
  - R^2 moyen : **0.59**
  - Écart-type : 0.25
- Test set : **0.73**



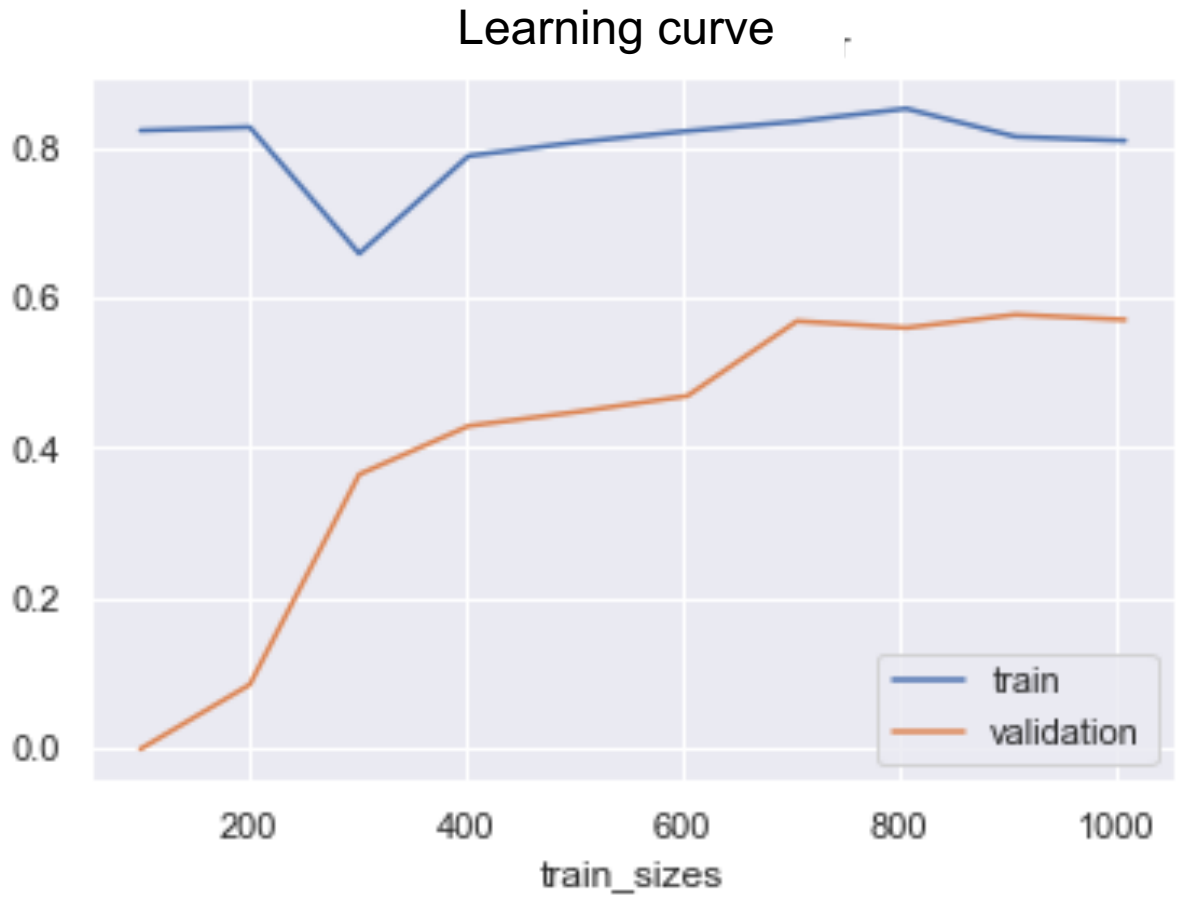
Sans l'ENERGYSTARScore

Feature Engineering :
- OneHotEncoding
- MinMaxScaler

Modèle final : **RandomForestRegressor**
- min_samples_leaf = 2
- min_samples_split = 2
- n_estimators = 10

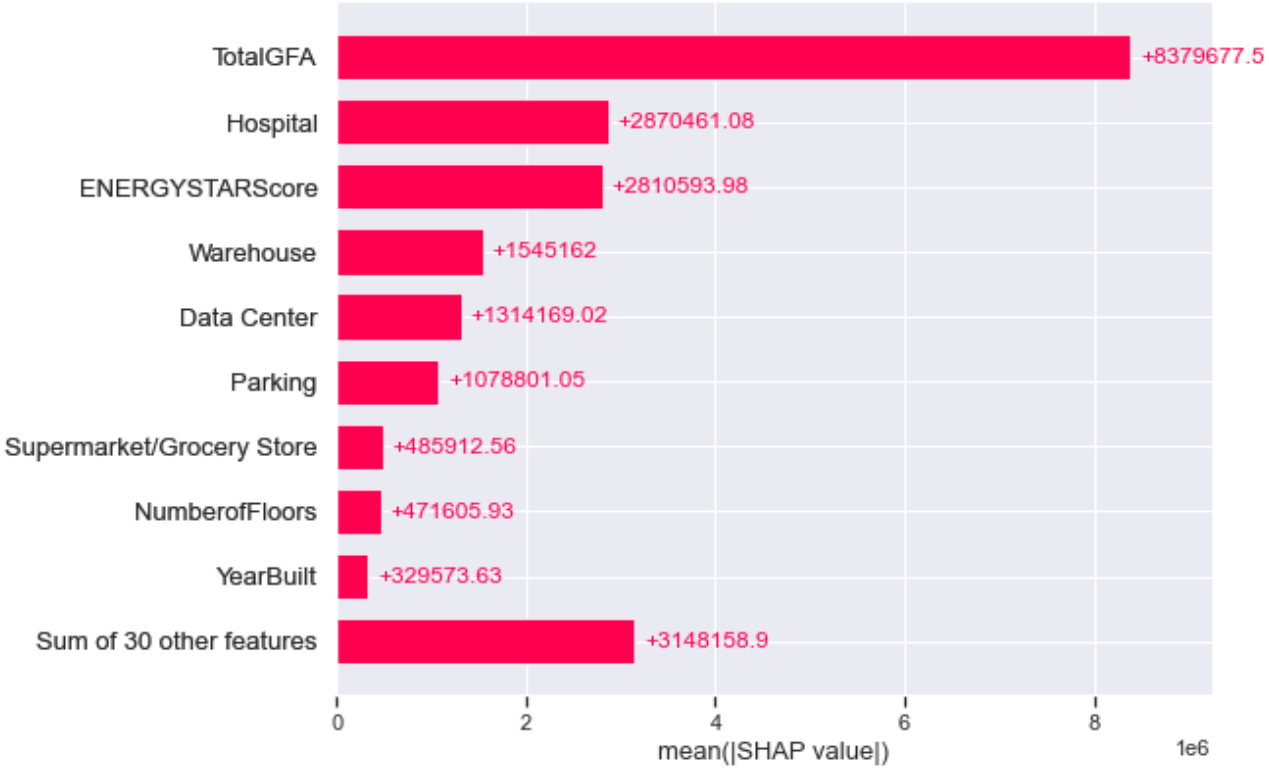
Résultats
- Cross-validation :
  - R^2 moyen : **0.65**
  - Écart-type : 0.09

- Test set : **0.66**

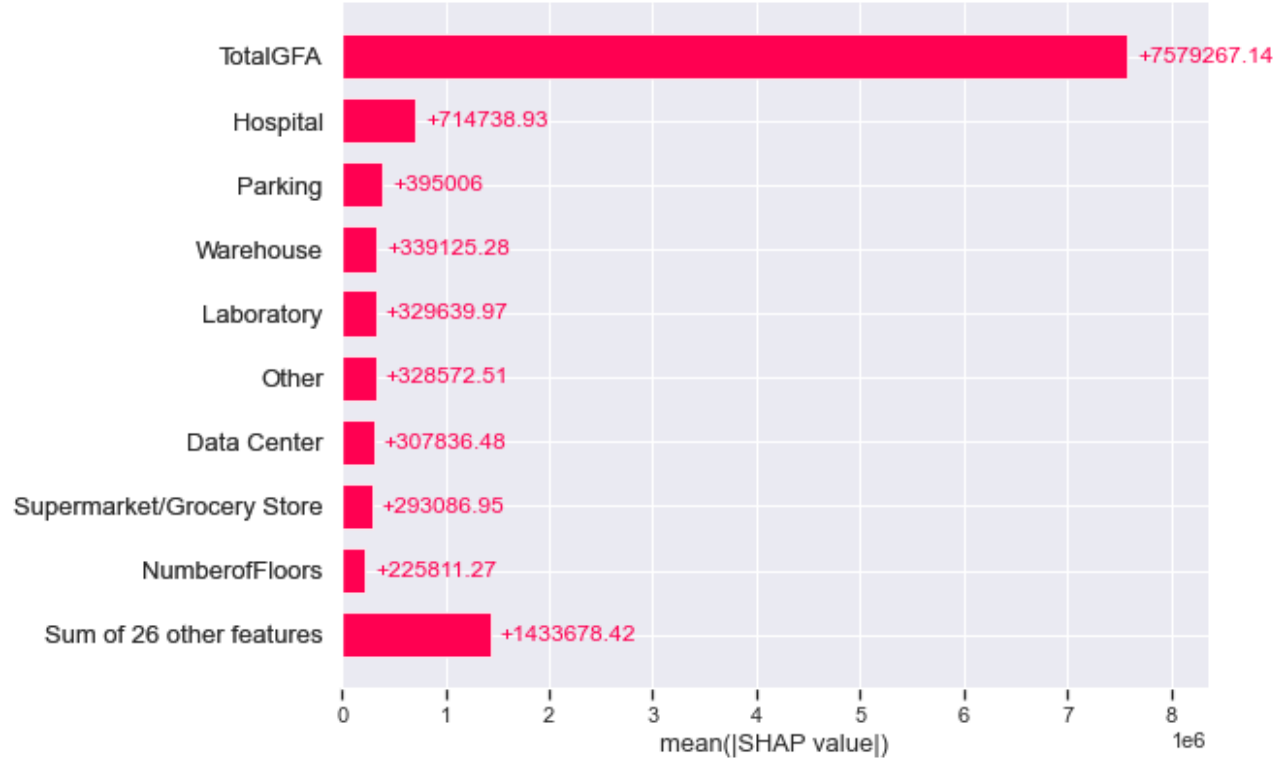


Feature Importance

Avec l'ENERGYSTARScore



Sans l'ENERGYSTARScore



Target 2 : Émission des GES

Avant le Feature Engineering :

- Suppression des outliers à l'aide d'une IsolationForest

Feature Engineering :

- Transformation des variables Electricity, NaturalGas et SteamUse en pourcentages

Electricity (kBtu)	NaturalGas (kBtu)	SteamUse (kBtu)
9000	900	100

Electricity	NaturalGas	SteamUse
0.90	0.09	0.01

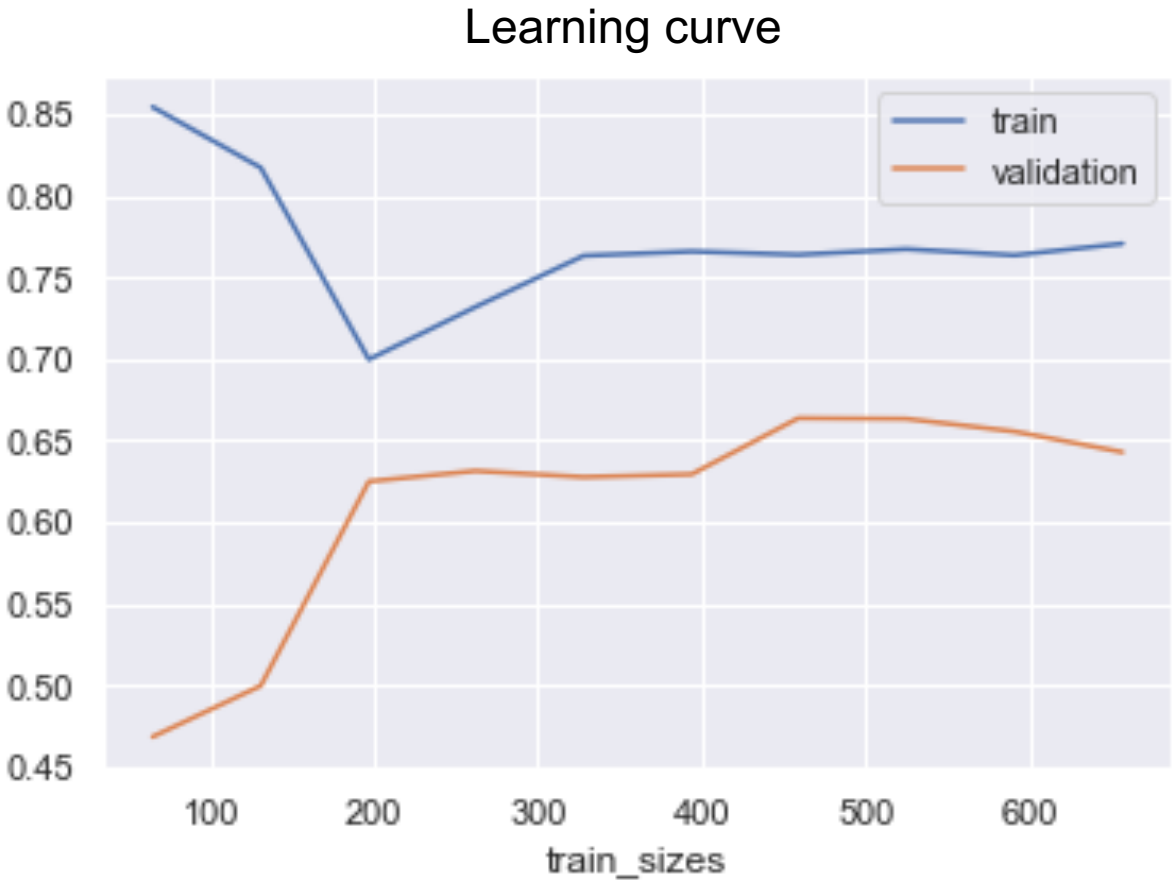
Target 2 : Émission des GES

Feature Engineering :
- dropna()
- TargetEncoding
- StandardScaler

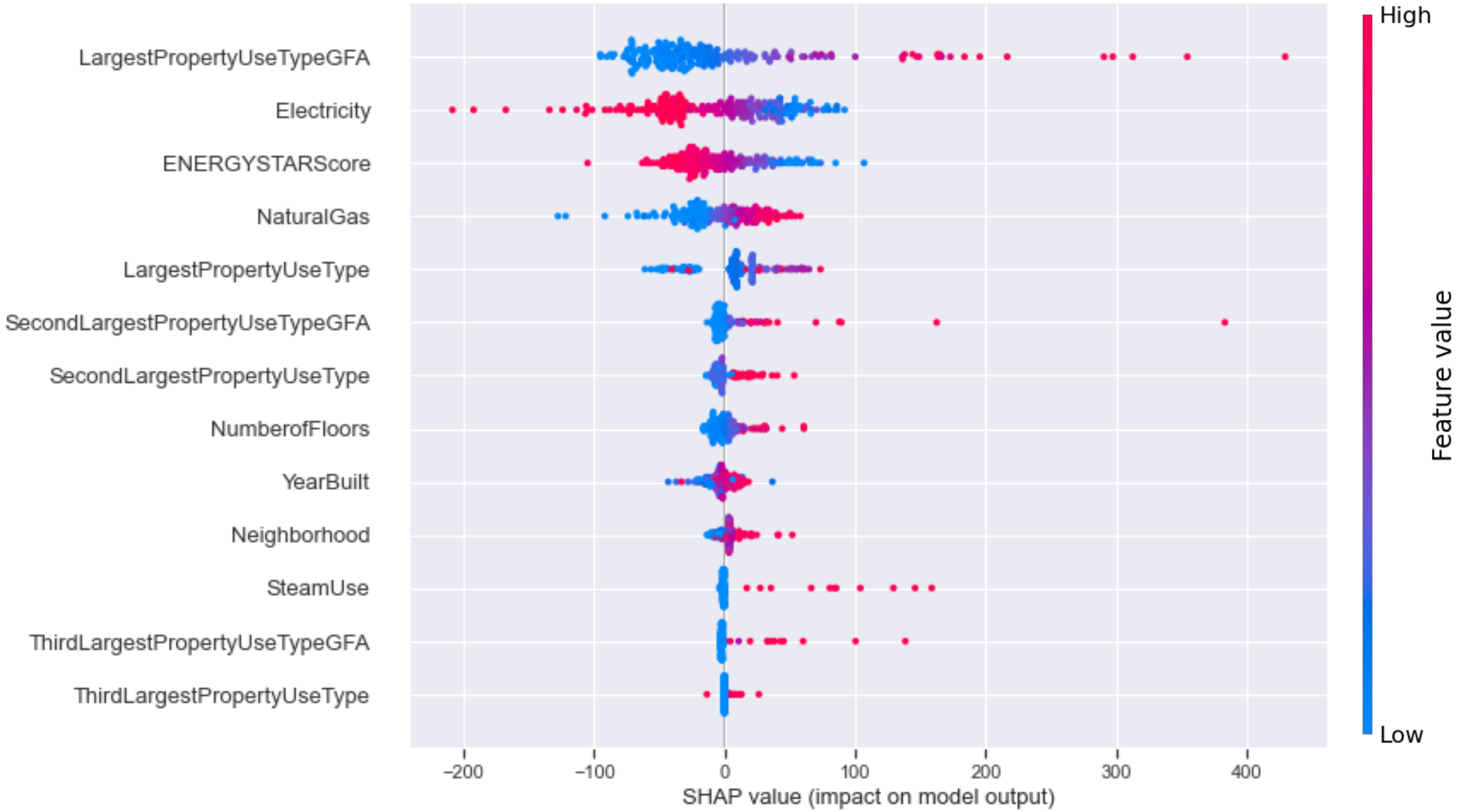
Modèle final : **BaggingRegressor**
- base_estimator = KernelRidge
  - alpha = 1
  - kernel = 'poly'
  - degree = 2
- max_features = 0.75
- n_estimators = 30

Résultats
- Cross-validation :
  - R^2 moyen : **0.65**
  - Écart-type : 0.09

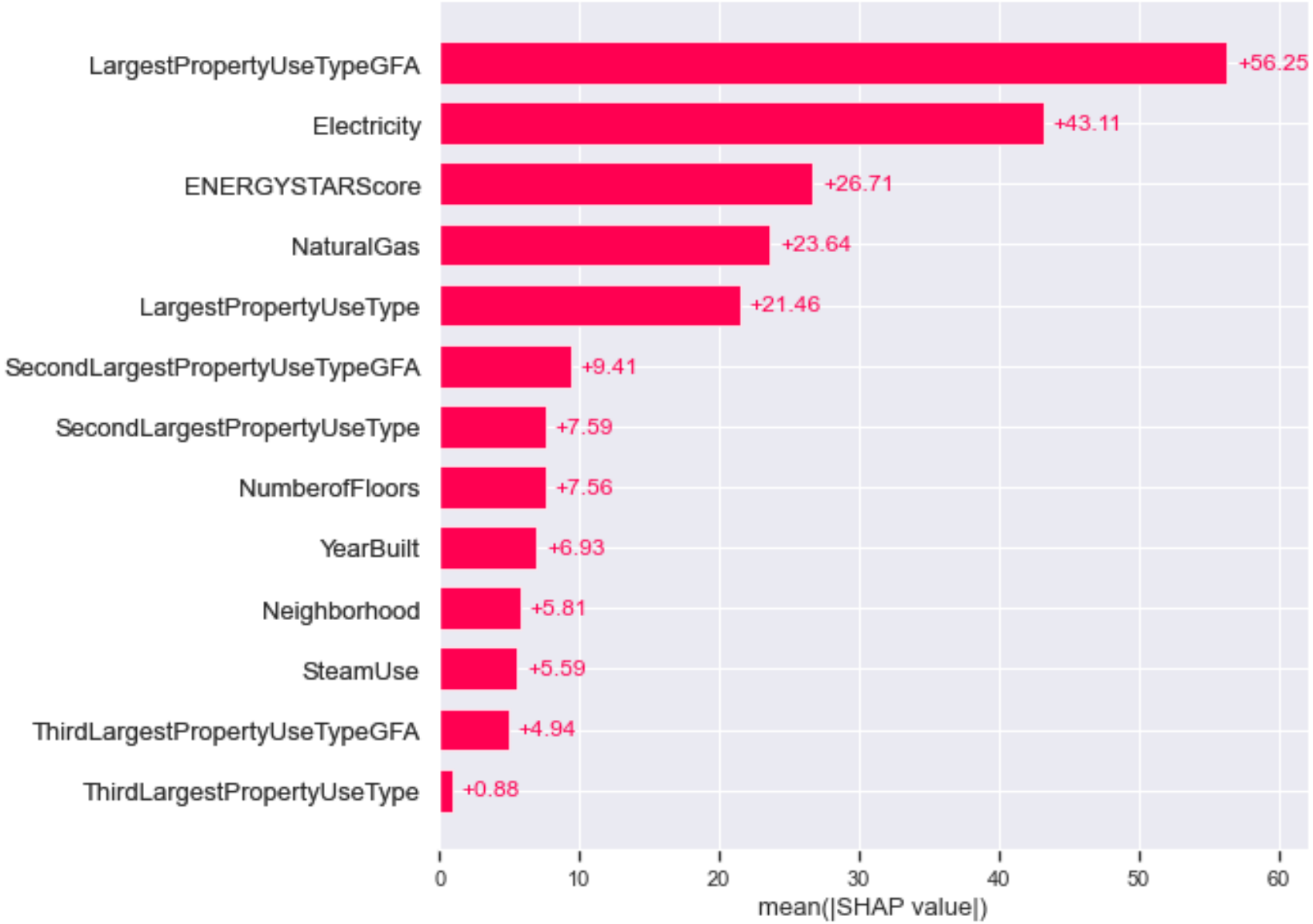
- Test set : **0.64**



Feature Importance



Feature Importance



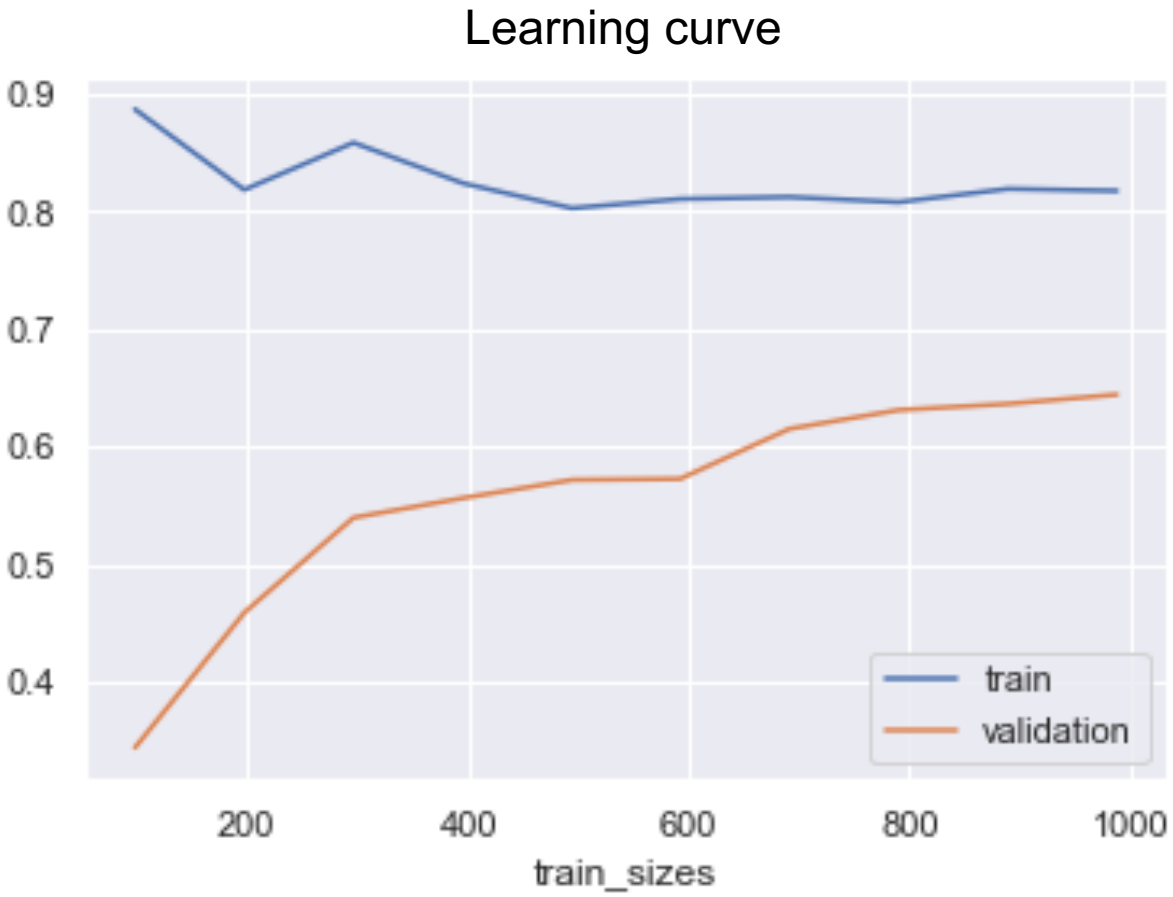
Sans l'ENERGYSTARScore

Feature Engineering :
- TargetEncoding
- MinMaxScaler

Modèle final : **SVR**($C=10^4$)

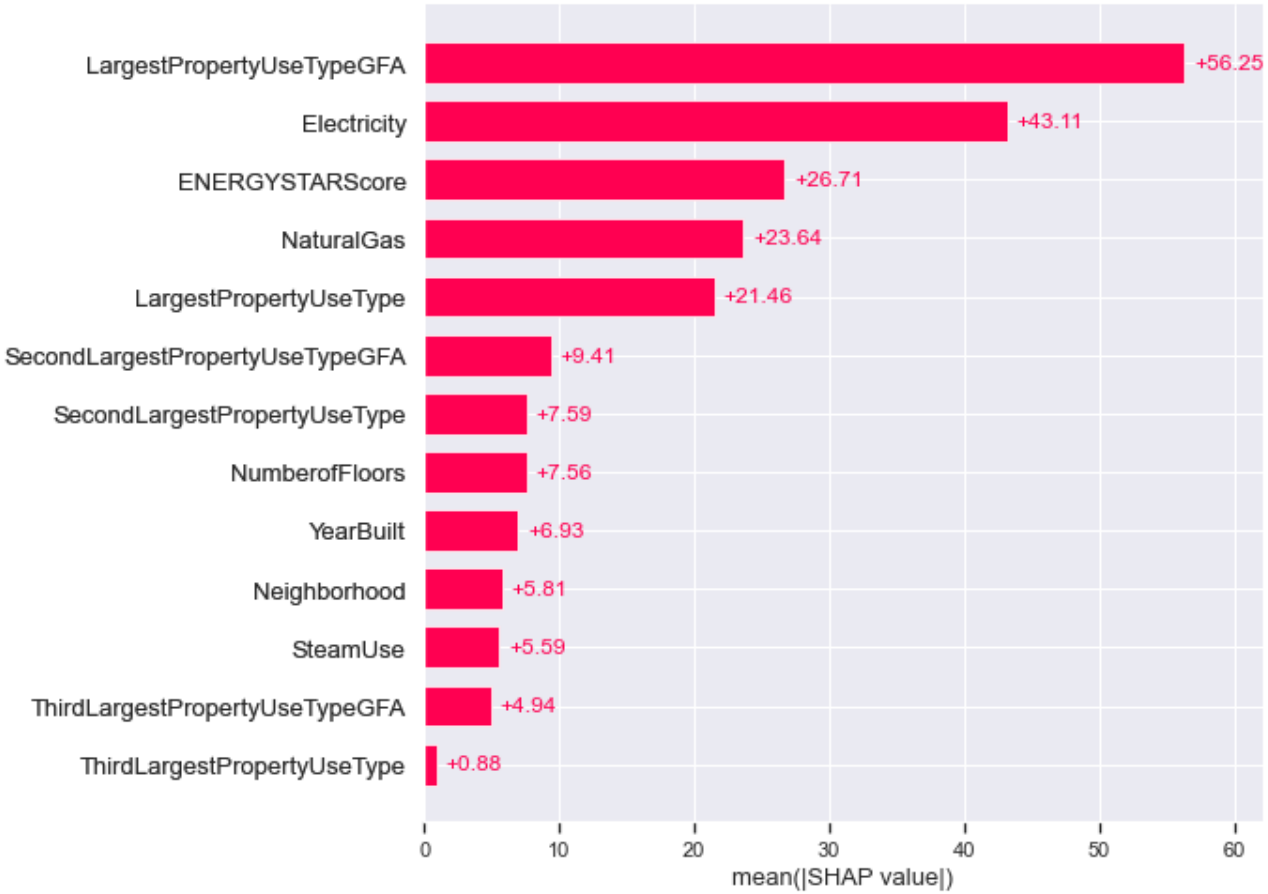
Résultats
- Cross-validation :
  - R^2 moyen : 0.60
  - Écart-type : 0.03

- Test set : 0.51

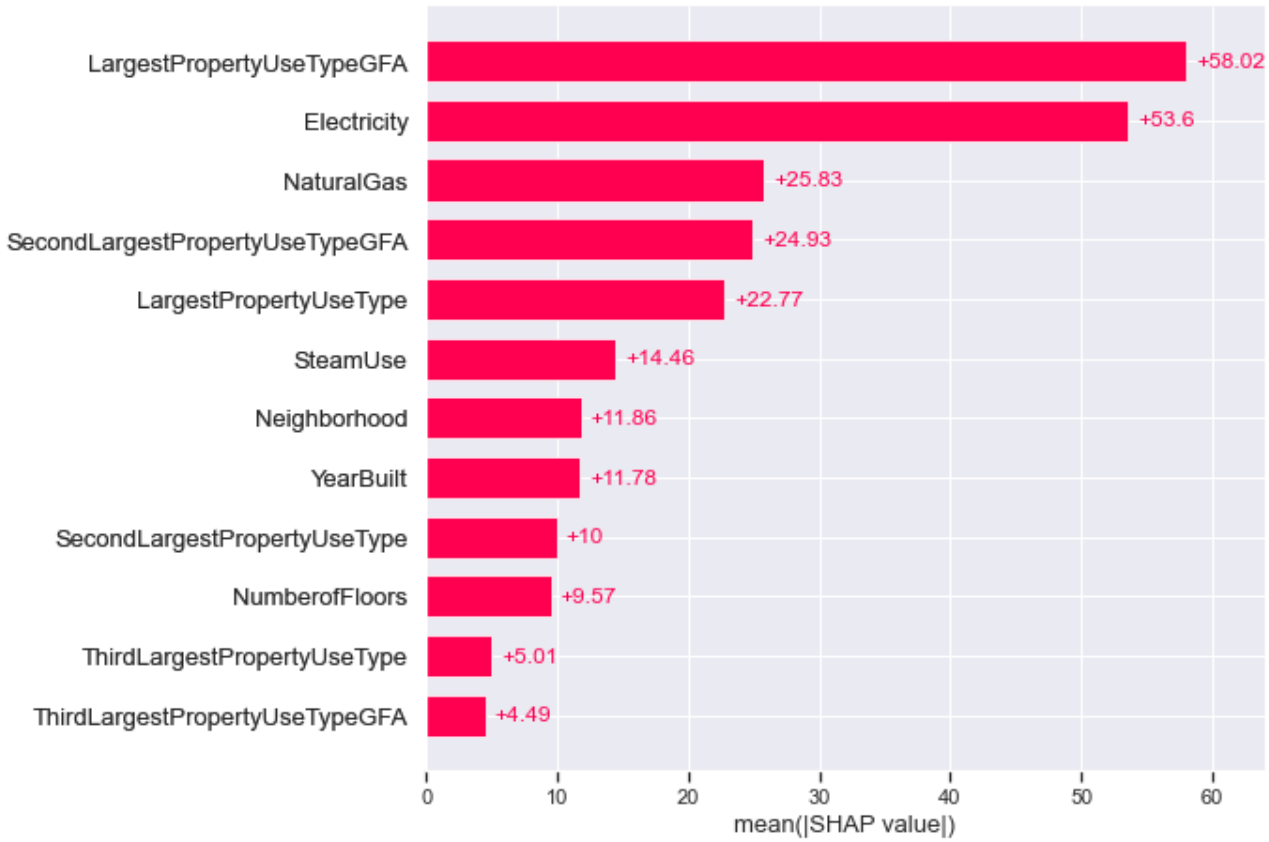


Feature Importance

Avec l'ENERGYSTARScore



Sans l'ENERGYSTARScore



Conclusion

- Modélisation **avec** l'ENERGYSTARScore :
  - Consommation : R^2 ~ 0.75
  - Émissions : R^2 ~ 0.65
- Modélisation **sans** l'ENERGYSTARScore :
  - Consommation : R^2 ~ 0.65
  - Émissions : R^2 ~ 0.60

- Variables les plus importantes :
  - Taille (surface, étages)
  - ENERGYSTARScore
  - Type d'usage
  - Sources d'énergies

- Il est possible de prédire la consommation avec une bonne précision

- Pour les émissions en revanche, il faudrait peut-être plus de données pour améliorer le modèle

MERCI POUR VOTRE ATTENTION