# Anticipez les besoins en consommation électrique de bâtiments

Seattle

Neutralité carbone 2050

# Sommaire

# Sommaire

# Problématique - Contexte

## Objectif de la ville de Seattle :

**Neutralité carbone** en 2050.
33% des émissions par bâtiments non résidentiels
➔ connaître leurs **consommation** en énergie et **émission.**

## Problème :

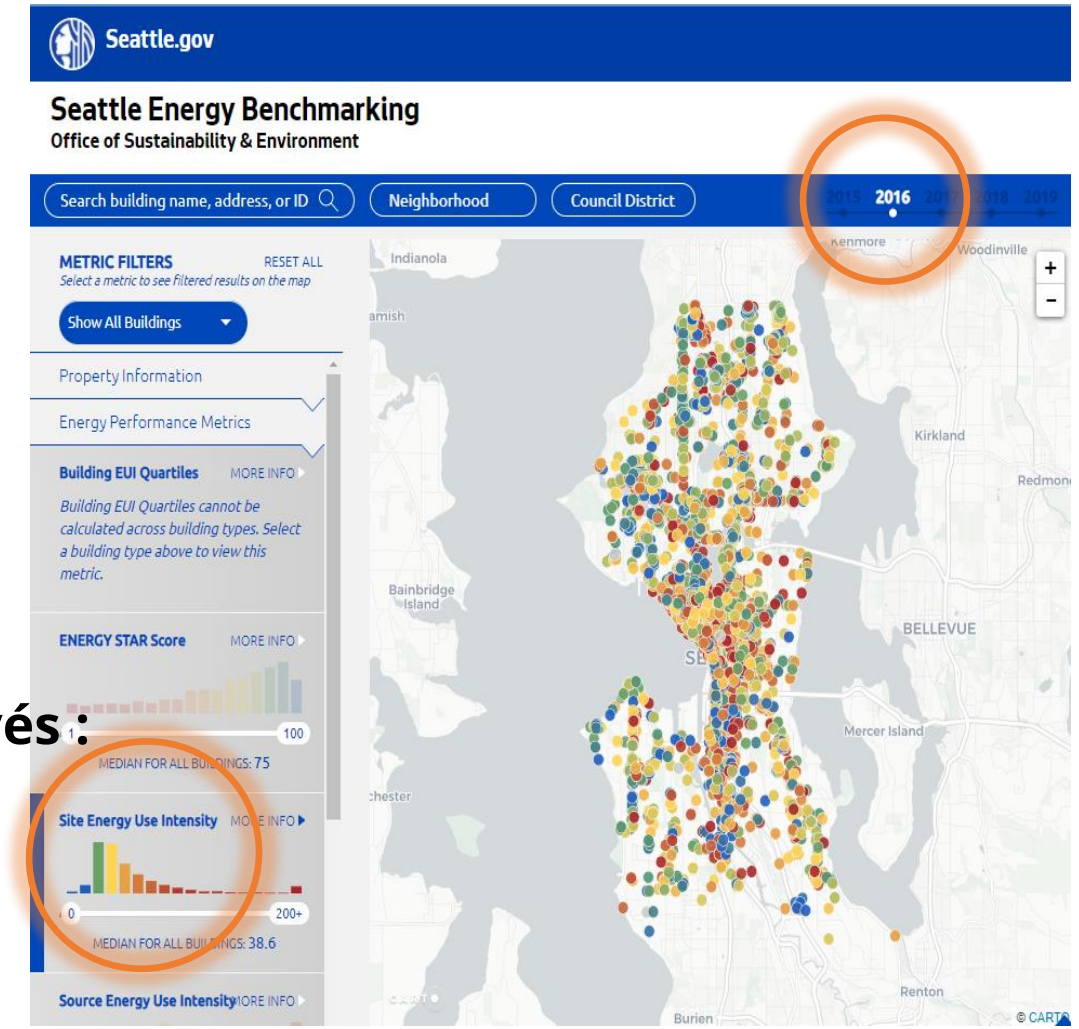
Des bilans **fastidieux** et **coûteux** effectués en 2015 et 2016

## Missions :

A partir des données récoltées et **sans nouveaux relevés :**

**(1) Prédire** la **consommation** totale d'énergie
**(2) Prédire** l'**émission** en GES
pour des nouveaux bâtiments ou non mesurés.

**(3)** Evaluer l'intérêt de **l'ENERGY STAR Score** pour la **prédiction** d'**émission.**




Source

# Problématique – Interprétation/pistes

**2 jeux de données (kaggle)** :

Similarité ? Doublons? Assembler

**Sélectionner les targets** :

Total/Intensité?
Source/Site
Site/SiteWN

**Sélectionner les variables indépendantes** :

caractéristiques intrinsèques des bâtiments
➔ pas les variables d'énergie

**Bâtiments non résidentiels** :

filtrer les bâtiments multi-familly?

**Modélisation** :

2 variables cibles quantitatives à prédire
➔ 2 modélisations de régression

**SiteEnergyUseWN(kBtu)**
➔ modèle consommation d'énergie

**TotalGHGEmissions**
➔ modèle émission de GES

Intérêt de **EnergyStar Score**
➔2 modèles à comparer (un avec EnergyStar Score et un sans)

# Problématique - Données

**2 jeux de données de 2015 et 2016 (kaggle)**

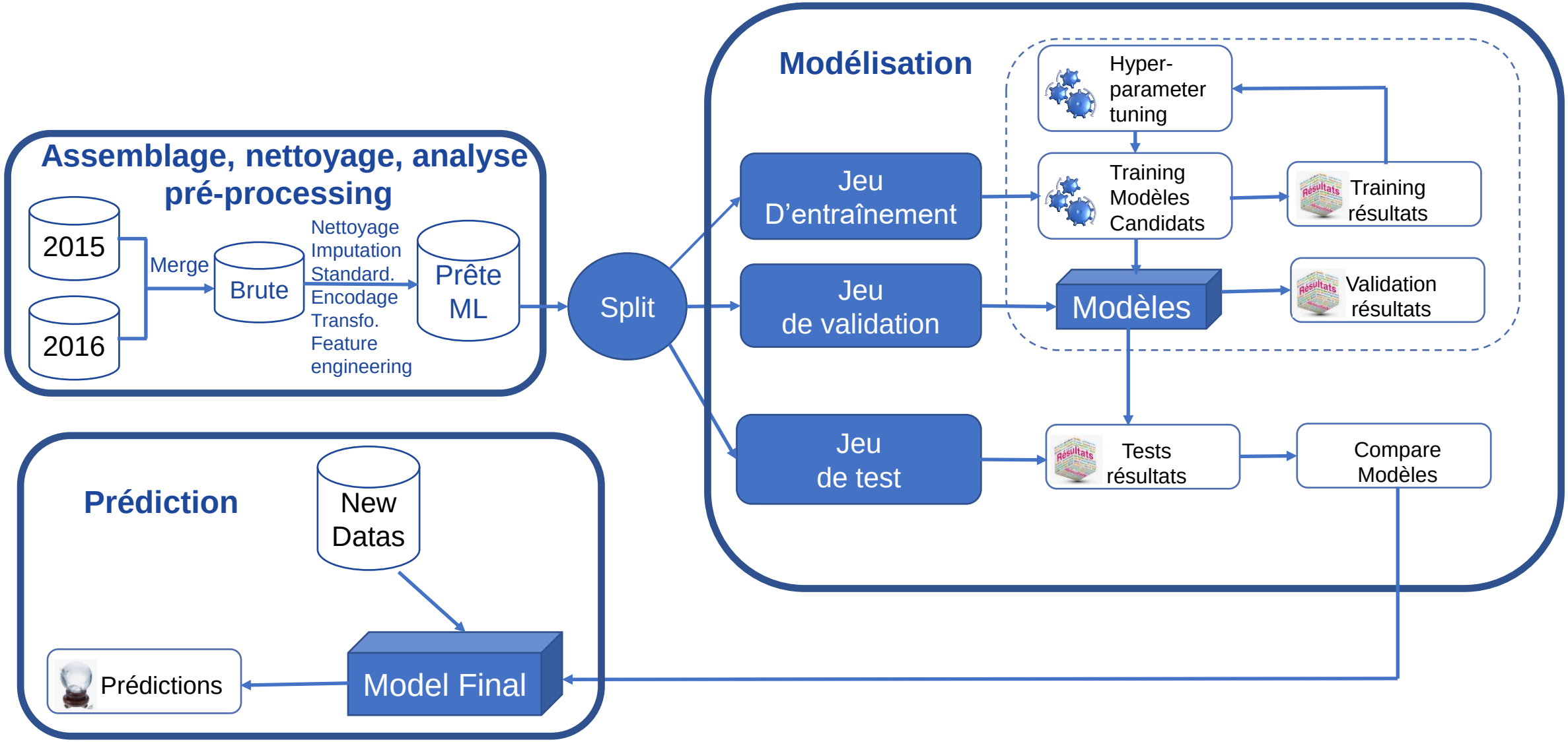
**Informations :** identification, localisation, usage, construction, consommation énergétique, émissions de GES.

| | | Données 2015 | Données 2016 |
|---|---|---|---|
| Lignes | Taille | 3340 | 3376 |
| Variables | Nombre | 47 | 46 |
| | Similaires | Location | Latitude, Longitude, Address, City, ZipCode, State |
| | | GHGEmissions(MetricTonsCO2e) | TotalGHGEmissions |
| | | GHGEmissionsIntensity(kgCO2e/ft2) | GHGEmissionsIntensity |
| | | Comment | Comments |
| | En plus | OtherFuelUse(kBtu), SPD Beats, Seattle Police Department Micro Community Policing Plan Areas, City Council Districts, Zip Codes, 2010 Census Tracts | |

**Variables cibles :** SiteEnergyUseWN(kBtu)
TotalGHGEmissions

**EnergyStar Score :** ENERGYSTARScore

# Problématique - Démarche



# Sommaire

# Données

**Métier**

- Compréhension du métier
- Assemblage des jeux de données

**Nettoyage**

- Suppression des données inutiles, filtre
- Valeurs manquantes, aberrantes
- Harmonisation maj/min, gestion des doublons

**Analyse**

- Analyse univariée
- Analyse bivariée / multi-variée

**Pré processing**

- Feature engineering
- Imputation
- Type des variables, transformation cible

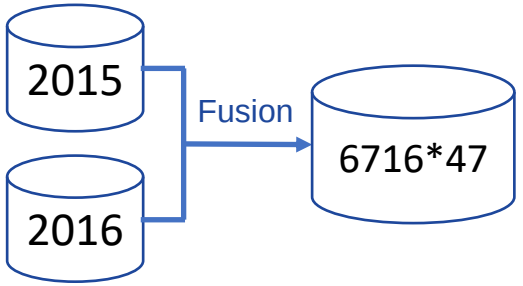
# Données - Assemblage

## Compréhension métier

| Données 2015 | Données 2016 |
|---|---|
| 3340 | 3376 |
| 47 | 46 |
| Location | **Latitude, Longitude, Address, City, ZipCode, State** |
| GHGEmissions(MetricTonsCO2e | **TotalGHGEmissions** |
| GHGEmissionsIntensity(kgCO2e/ft2) | **GHGEmissionsIntensity** |
| Comment | **Comments** |
| **OtherFuelUse(kBtu)**, SPD Beats, Seattle Police Department Micro Community Policing Plan Areas, City Council Districts, Zip Codes, 2010 Census Tracts | |

| | Nombres de valeurs manquantes | % de valeurs manquantes |
|---|---|---|
| City Council Districts | 3127 | 93.62000 |
| 2010 Census Tracts | 3116 | 93.29000 |
| OtherFuelUse(kBtu) | 10 | 0.30000 |
| SPD Beats | 2 | 0.06000 |
| Seattle Police Department Micro Community Policing Plan Areas | 2 | 0.06000 |

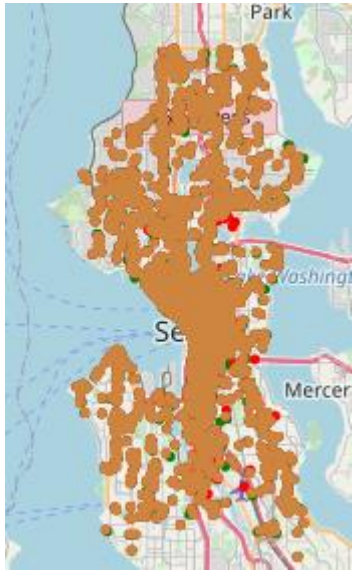
## Assemblage




3284 doublons

Stratégie
3511 bâtiments

- +92 **2016**
- **PARTIE FIXE 2016** 3284
- **2015** +56
- **2015 / 2016** Var>100% +79

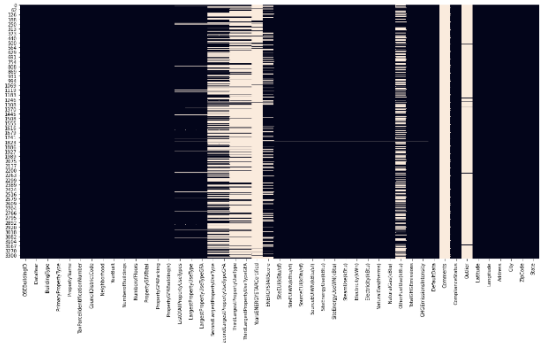



NumberofBuildings
PropertyGFAParking
Targets

# Données - Nettoyage

## De…

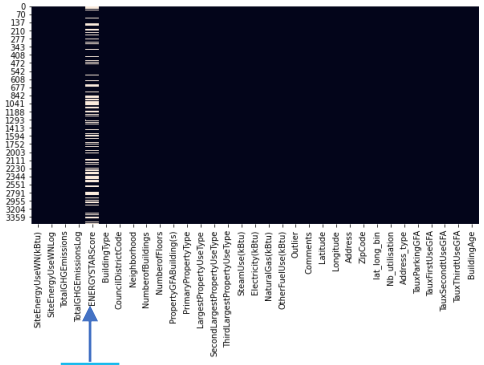
**6716 lignes, 47 variables**



**Assemblage**

## … À

**1722 lignes, 33 variables**



**Nettoyées**

**valeurs manquantes**

| | Nb lignes | Nb var. |
|---|---|---|
| Fusion des jeux de données | **6716** | **47** |
| Suppression des variables inutiles | 6716 | 23 |
| Harmonisation du contenu des variables catégorielles<br>Majuscules, suppression des accents ➔ doublons | 6716 | 23 |
| Réduction des modalités des variables catégorielles | 6716 | 23 |
| Filtre des bâtiments (2016 + sup 2015 + var>100%) | 3511 | 23 |
| Filtre des bâtiments non résidentiels | 1760 | 23 |
| Gestion des valeurs manquantes - imputation | 1722 | 23 |
| Gestion des valeurs aberrantes | 1722 | 23 |
| Feature engineering | 1722 | 23 |
| Types (catégorielle en object, float/int64 en 32) | **1722** | **33** |

+10

# Données - Nettoyage - Suppression variables

| Variables | Raison |
|---|---|
| City Council Districts, Zip Codes, 2010 Census Tracts, Seattle Police Department Micro Community Policing Plan Areas, SPD Beats, | Seulement en 2015, abandonnées avant fusion |
| PropertyName, TaxParcelIdentificationNumber, OSEBuildingID YearsENERGYSTARCertified, DefaultData, ComplianceStatus | Inutiles pour notre problématique ou trop de valeurs manquantes |
| Electricity(kWh), NaturalGas(therms) | Autres unités de mesure d'énergie |
| SiteEUI(kBtu/sf), SiteEUIWN(kBtu/sf), SourceEUI(kBtu/sf), SourceEUIWN(kBtu/sf), | Unités en fonction de la surface en pieds carrés |
| SiteEnergyUse(kBtu), GHGEmissionsIntensity | Redondantes avec les cibles |
| DataYear, YearBuilt, PropertyGFAParking, LargestPropertyUseTypeGFA, SecondLargestPropertyUseTypeGFA, ThirdLargestPropertyUseTypeGFA | Après Feature engineering |
| ListOfAllPropertyUseTypes, PropertyGFATotal | Après imputation |
| City (SEATTLE), State (WA) | 1 seule valeur |

INUTILES

REDONDANCES

PRÉ-PROCESSING

# Données - Nettoyage - Modalités

Nombreuses **modalités** avec redondances, espace en plus, minuscules, majuscules

➔ **regroupements**

| | BuildingType | PrimaryPropertyType | Neighborhood | SecondLargestPropertyUseType | ThirdLargestPropertyUseType |
|---|---|---|---|---|---|
| unique | 8 | 32 | 19 | 50 | 45 |

| Variable | Modalités | Regroupement |
|---|---|---|
| BuildingType | 8 | 6 |
| Neighborhood | 19 | 13 |
| PrimaryPropertyType | 32 | 10 |
| LargestPropertyUseType | 57 | 10 |
| SecondLargestPropertyUseType | 50 | 10 |
| ThirdLargestPropertyUseType | 45 | 10 |

Residential, Public_Services, Hobbies, Others, Education, Retails, Office, Technology_Science, Healthcare, Foods_Sales

| Neighborhood |
|---|
| BALLARD |
| Ballard |
| CENTRAL |
| Central |
| DELRIDGE |
| DELRIDGE NEIGHBORHOODS |
| DOWNTOWN |
| Delridge |
| EAST |
| GREATER DUWAMISH |
| LAKE UNION |
| MAGNOLIA / QUEEN ANNE |
| NORTH |
| NORTHEAST |
| NORTHWEST |
| North |
| Northwest |
| SOUTHEAST |
| SOUTHWEST |

| PrimaryPropertyType | n |
|---|---|
| Low-Rise Multifamily | 1985 |
| Mid-Rise Multifamily | 1103 |
| Small- and Mid-Sized Office | 590 |
| Other | 514 |
| Large Office | 344 |
| K-12 School | 275 |
| Mixed Use Property | 259 |
| High-Rise Multifamily | 208 |
| Retail Store | 191 |
| Warehouse | 187 |
| Non-Refrigerated Warehouse | 187 |
| Hotel | 150 |
| Worship Facility | 143 |
| Senior Care Community | 88 |
| Medical Office | 82 |
| Distribution Center | 55 |
| Distribution Center\n | 51 |
| Supermarket / Grocery Store | 40 |
| Supermarket/Grocery Store | 36 |
| Self-Storage Facility | 29 |
| Self-Storage Facility\n | 27 |
| University | 25 |
| Refrigerated Warehouse | 25 |
| Residence Hall | 23 |
| College/University | 22 |
| Hospital | 20 |
| Residence Hall/Dormitory | 15 |
| Restaurant | 13 |
| Laboratory | 11 |
| Restaurant\n | 11 |
| SPS-District K-12 | 4 |
| Office | 3 |

# Données - Nettoyage - Valeurs manquantes

## Valeurs manquantes

| | Nombres de valeurs manquantes | % de valeurs manquantes |
|---|---|---|
| Comments | 1758 | 99.89000 |
| OtherFuelUse(kBtu) | 1669 | 94.83000 |
| ThirdLargestPropertyUseType | 1391 | 79.03000 |
| ThirdLargestPropertyUseTypeGFA | 1391 | 79.03000 |
| SecondLargestPropertyUseType | 868 | 49.32000 |
| SecondLargestPropertyUseTypeGFA | 868 | 49.32000 |
| ENERGYSTARScore | 603 | 34.26000 |
| LargestPropertyUseTypeGFA | 17 | 0.97000 |
| LargestPropertyUseType | 17 | 0.97000 |
| ListOfAllPropertyUseTypes | 13 | 0.74000 |
| SiteEnergyUseWN(kBtu) | 4 | 0.23000 |
| SteamUse(kBtu) | 3 | 0.17000 |
| Electricity(kBtu) | 3 | 0.17000 |
| NaturalGas(kBtu) | 3 | 0.17000 |
| TotalGHGEmissions | 3 | 0.17000 |
| GHGEmissionsIntensity | 3 | 0.17000 |
| SiteEnergyUse(kBtu) | 3 | 0.17000 |

## Imputation

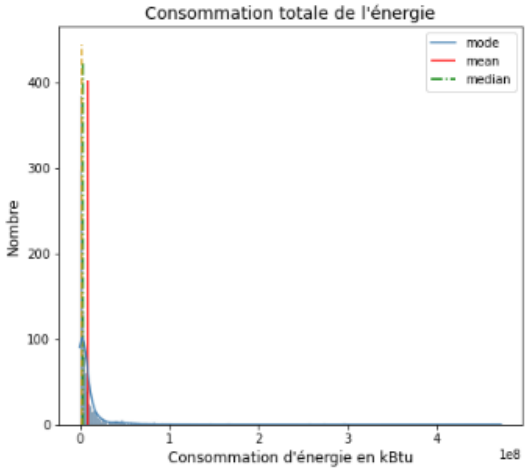
| | Variable | Imputation |
|---|---|---|
| Manuelle | ZipCode (17, 2016) | A la main en cherchant l'adresse dans Google Map |
| | NumberofBuildings | A la main, 1 au lieu de 0 après recherche dans Google Map |
| | NumberofFloors | A la main, 2 au lieu de 99 : église moderne de 2 étages maximum |
| | Outlier | fillna('not'), 'low outlier' => 'low', 'high outlier': 'high'} |
| Fillna(0) | SecondLargestPropertyUseTypeGFA<br>ThirdLargestPropertyUseTypeGFA<br>OtherFuelUse(kBtu) | fillna(0) |
| Fillna('constante') | SecondLargestPropertyUseType<br>ThirdLargestPropertyUseType | fillna('Pas utilisation') |
| | Comments | fillna('Sans commentaire') |
| Autre variable | LargestPropertyUseTypeGFA | Récupération donnée de PropertyGFATotal |
| | LargestPropertyUseType | A partir de PrimaryPropertyType |
| Non imputée | ENERGYSTARScore | Non imputée |
| | SiteEnergyUseWN(kBtu)<br>TotalGHGEmissions | Suppression sans information |

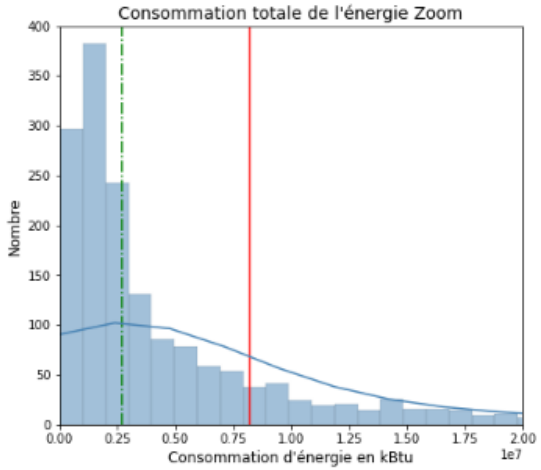
# Données - Nettoyage – Valeurs aberrantes

| | NumberofBuildings | NumberofFloors | PropertyGFABuilding(s) | LargestPropertyUseTypeGFA | SecondLargestPropertyUseTypeGFA | ThirdLargestPropertyUseTypeGFA | ENERGYSTARScore | Electricity(kBtu) | TotalGHGEmissions | SiteEnergyUse(kBtu) |
|---|---|---|---|---|---|---|---|---|---|---|
| type | int32 | int32 | int32 | float32 | float32 | float32 | float32 | float32 | float32 | float32 |
| nb_nan | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 580.0 | 0.0 | 0.0 | 0.0 |
| %_nan | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 33.68177 | 0.0 | 0.0 | 0.0 |
| count | 1722.0 | 1722.0 | 1722.0 | 1722.0 | 1722.0 | 1722.0 | 1142.0 | 1722.0 | 1722.0 | 1722.0 |
| mean | 1.13298 | 4.08943 | 99870.05865 | 93009.8125 | 18901.0332 | 3079.55737 | 65.91856 | 5454258.5 | 181.87529 | 7969176.0 |
| std | 1.13283 | 6.48667 | 171570.91025 | 159220.46875 | 53464.19922 | 17698.26367 | 28.34305 | 13329824.0 | 708.27771 | 21663752.0 |
| min | 1.0 | 0.0 | 3636.0 | 5656.0 | 0.0 | 0.0 | 1.0 | -115417.0 | -0.8 | 11441.0 |
| 25% | 1.0 | 1.0 | 28311.75 | 25480.25 | 0.0 | 0.0 | 49.0 | 720540.75 | 19.805 | 1231865.21875 |
| 50% | 1.0 | 2.0 | 46970.0 | 43525.5 | 0.0 | 0.0 | 73.5 | 1598629.5 | 49.265 | 2522355.75 |
| 75% | 1.0 | 1.0 | 94471.5 | 91616.75 | 12678.0 | 0.0 | 89.0 | 4782748.75 | 138.72501 | 6854863.875 |
| max | 27.0 | 99.0 | 2200000.0 | 1719643.0 | 686750.0 | 459748.0 | 100.0 | 274532480.0 | 16870.98047 | 448385312.0 |

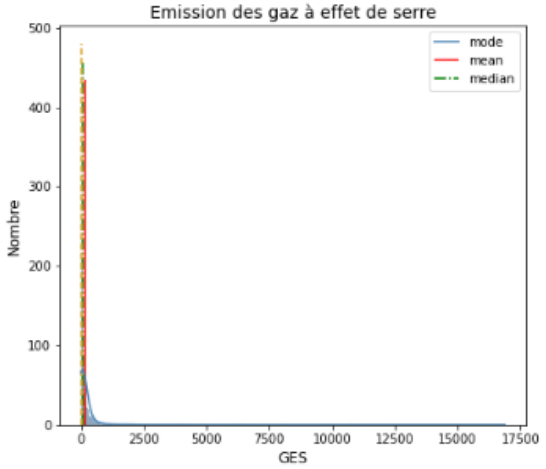
| | Variable | Aberrante? |
|---|---|---|
| **Aberrante** | NumberofFloors | A la main, 2 au lieu de 99 : église moderne de 2 étages maximum (google street), tour la plus haute 93 étages |
| **Outliers** | NumberofBuildings | >27? Pas aberrant, campus, hôpitaux |
| | TotalGHGEmissions Electricity(kBtu) | Négative? Pas aberrant : pour l'adresse du bâtiment : It's a zero waste building and is the greenest commercial building in the US |
| | TotalGHGEmissions SiteEnergyUse(kBtu) | Max aberrant? Non, les bâtiments = hôpitaux, campus |

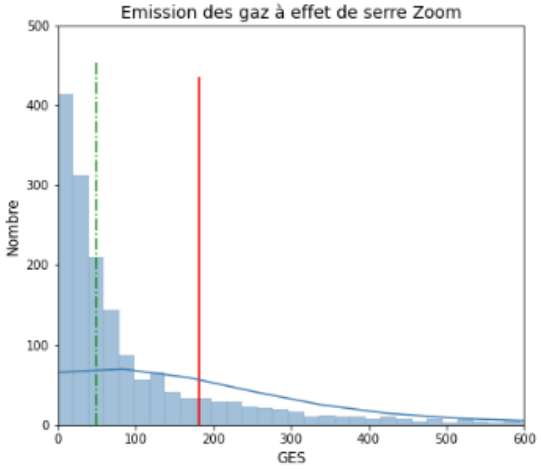
# Données – Analyse Univariée



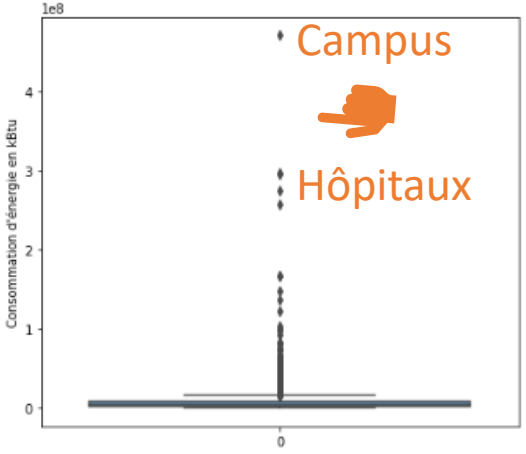



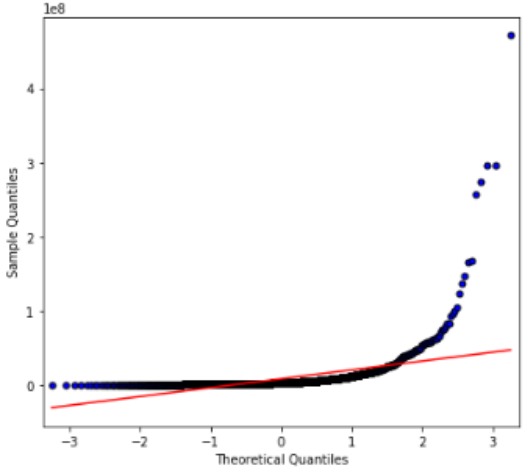

| Desc | SiteEnergyUseWN(kBtu) |
|---|---|
| mean | 8202849.50000 |
| median | 2697353.25000 |
| var | 492555842093056.00000 |
| std | 22193600.00000 |
| skew | 10.93242 |
| kurtosis | 168.72591 |
| mode | 0 2127889.25 |
| Min | 11441.00000 |
| Max | 471613856.00000 |



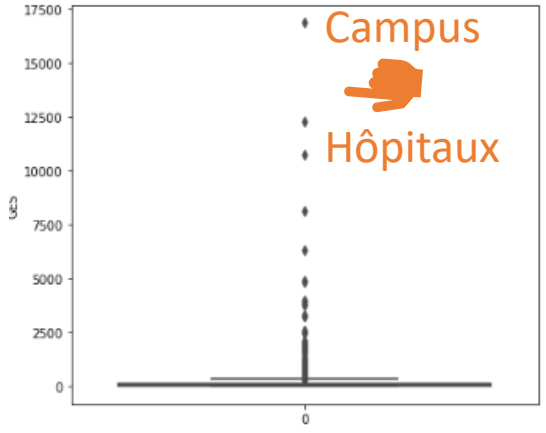

**Consommation totale de l'énergie**

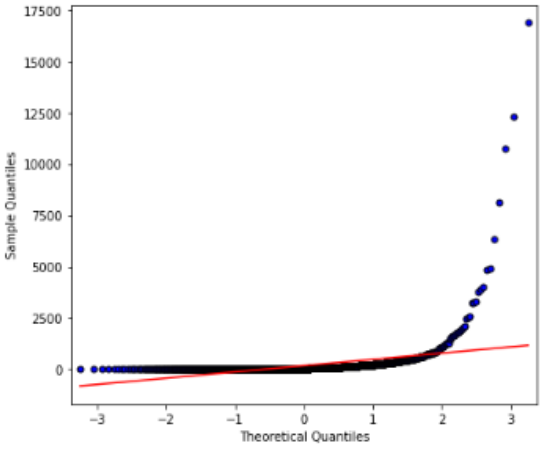



| Desc | TotalGHGEmissions |
|---|---|
| mean | 181.87529 |
| median | 49.26500 |
| var | 501365.15625 |
| std | 708.07141 |
| skew | 14.57867 |
| kurtosis | 272.83044 |
| mode | 0 6.16 1 6.30 2 20.94 |
| Min | -0.80000 |
| Max | 16870.98047 |

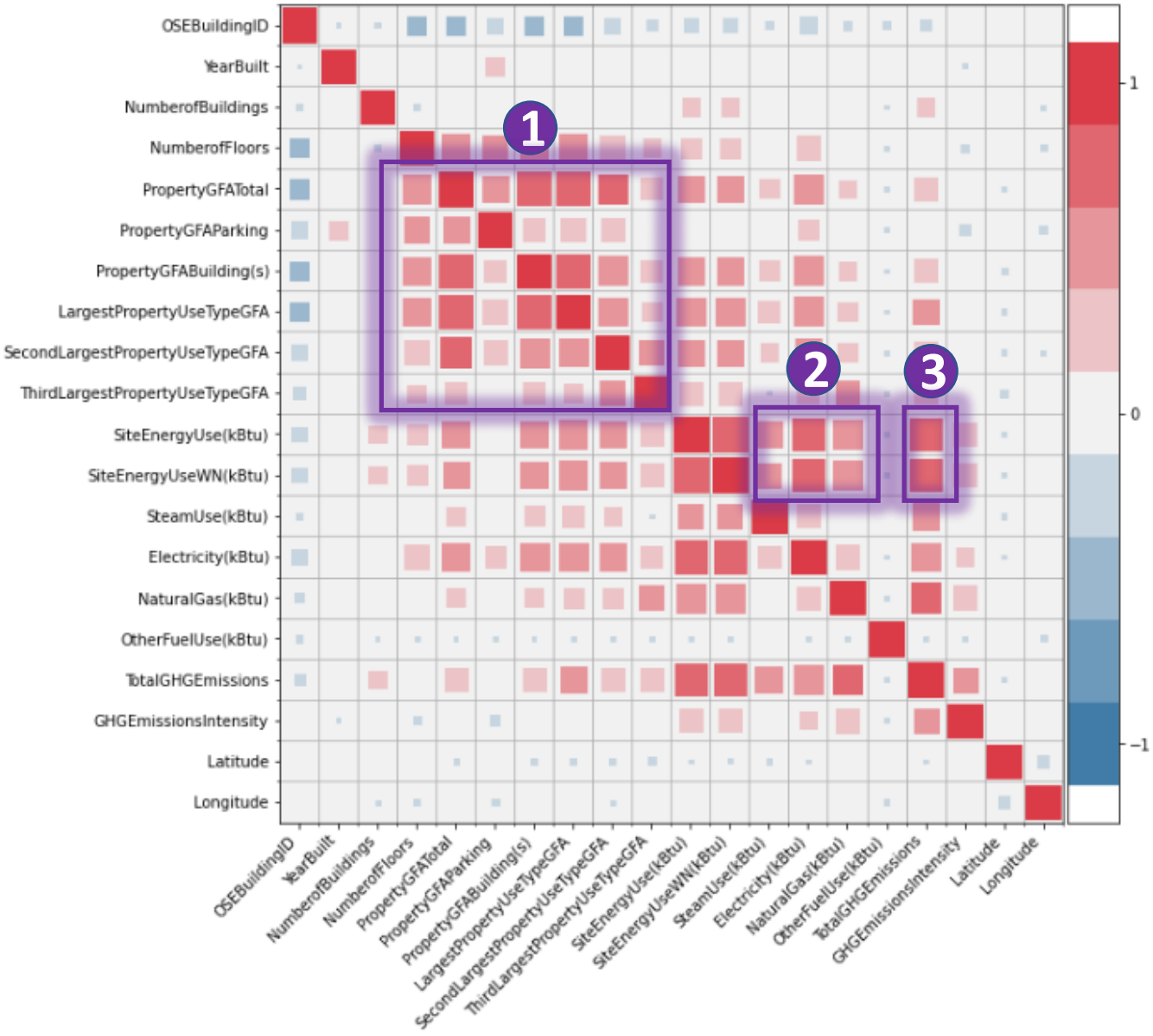
**Emission des gaz à effet de serre**




Skewness>1 → tranformation logarithmique

# Données – Analyse Multi-variée

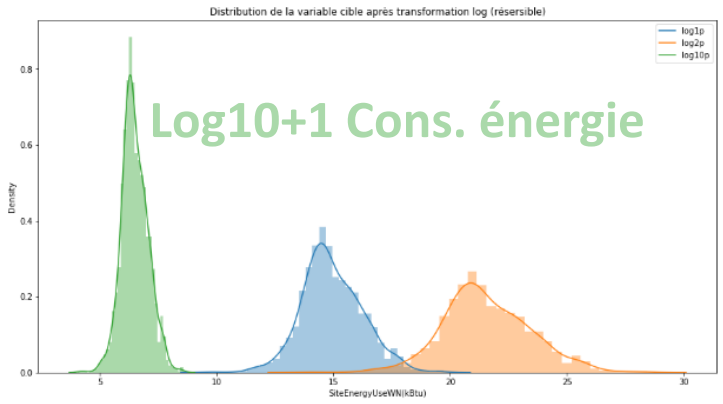


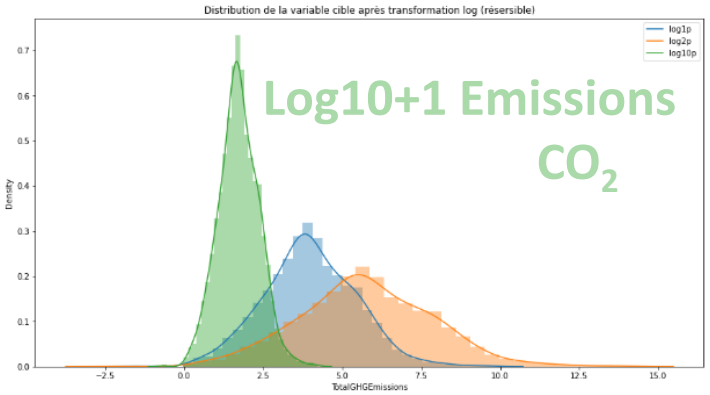

1. Features Engineering : nouvelles variables

2. Features Engineering : Seule information dans Le permis de construire : les sources d'énergie

3. Cibles très corrélées

# Données – Nettoyage - Feature engineering

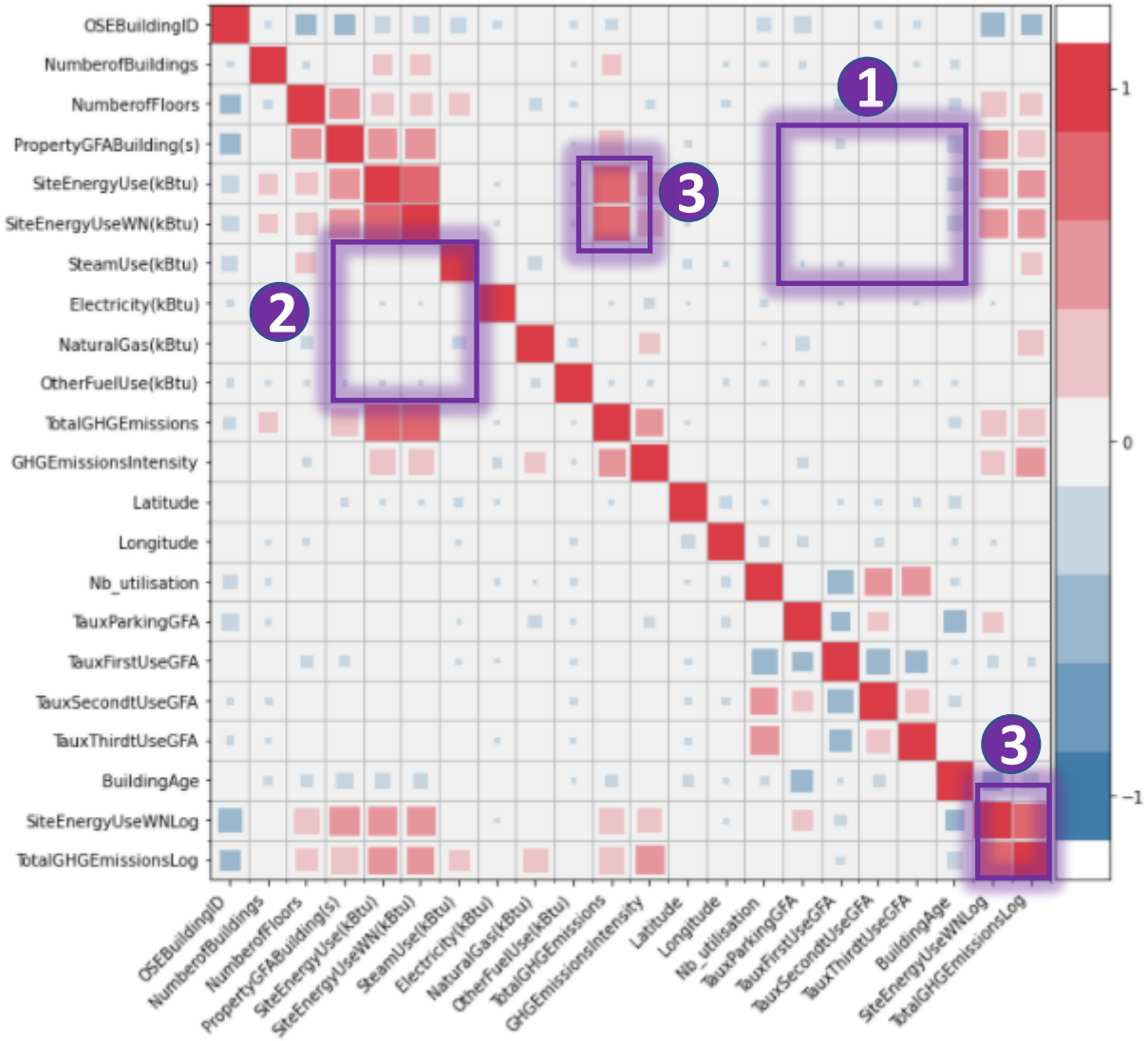



| | Variables | Description |
|---|---|---|
| **Localisation** | ListOfAllPropertyUseTypes | Compte le nombre de type de propriété pour chaque bâtiment |
| | Latitude/Longitude | Cartographie des bâtiments en binérisant la latitude et la longitude et en faisant la somme |
| | Address | Influence si le bâtiment est dans une rue, avenue, chemin? ➔ WAY, AVENUE ou STREET |
| **Construction** | DataYear<br>YearBuilt | L'âge du bâtiment ou de la dernière rénovation. |
| | PropertyGFAParking<br>PropertyGFATotal<br>LargestPropertyUseTypeGFA<br>SecondLargestPropertyUseTypeGFA<br>ThirdLargestPropertyUseTypeGFA | Ratio de la surface du parking sur la surface totale<br>Ratio de la surface de la permière (2nde, 3ième) sur la surface totale |
| **Energie, Emission** | SteamUse(kBtu), Electricity(kBtu), NaturalGas(kBtu), OtherFuelUse(kBtu) | 0 : n'utilise pas cette énergie,<br>1 : utilise cette énergie. |
| | SiteEnergyUseWN(kBtu) | Transformation en log10 + 1 |
| | TotalGHGEmissions | Transformation en log10 + 1 |

*Annexe 1- Nettoyage - Feature engineering*

# Données – Analyse Multi-variée




## APRES FEATURE ENGINEERING

1. Corrélation fortement réduite
2. Corrélation fortement réduite
3. Cibles très corrélées

# Sommaire

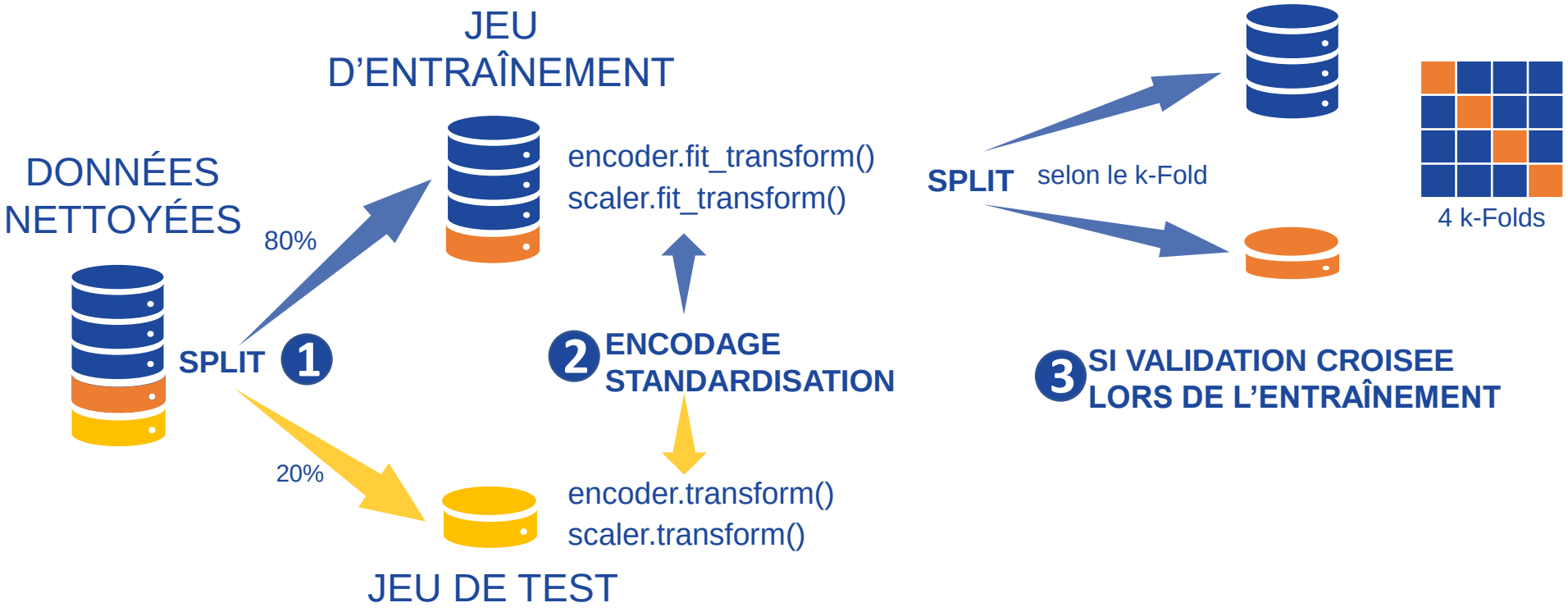
# Sommaire



## Consommation Totale d'électricité

# Modélisation - Consommation d'énergie

**SPLIT - ENCODAGE/STANDARDISATION** : préparation des données au machine learning



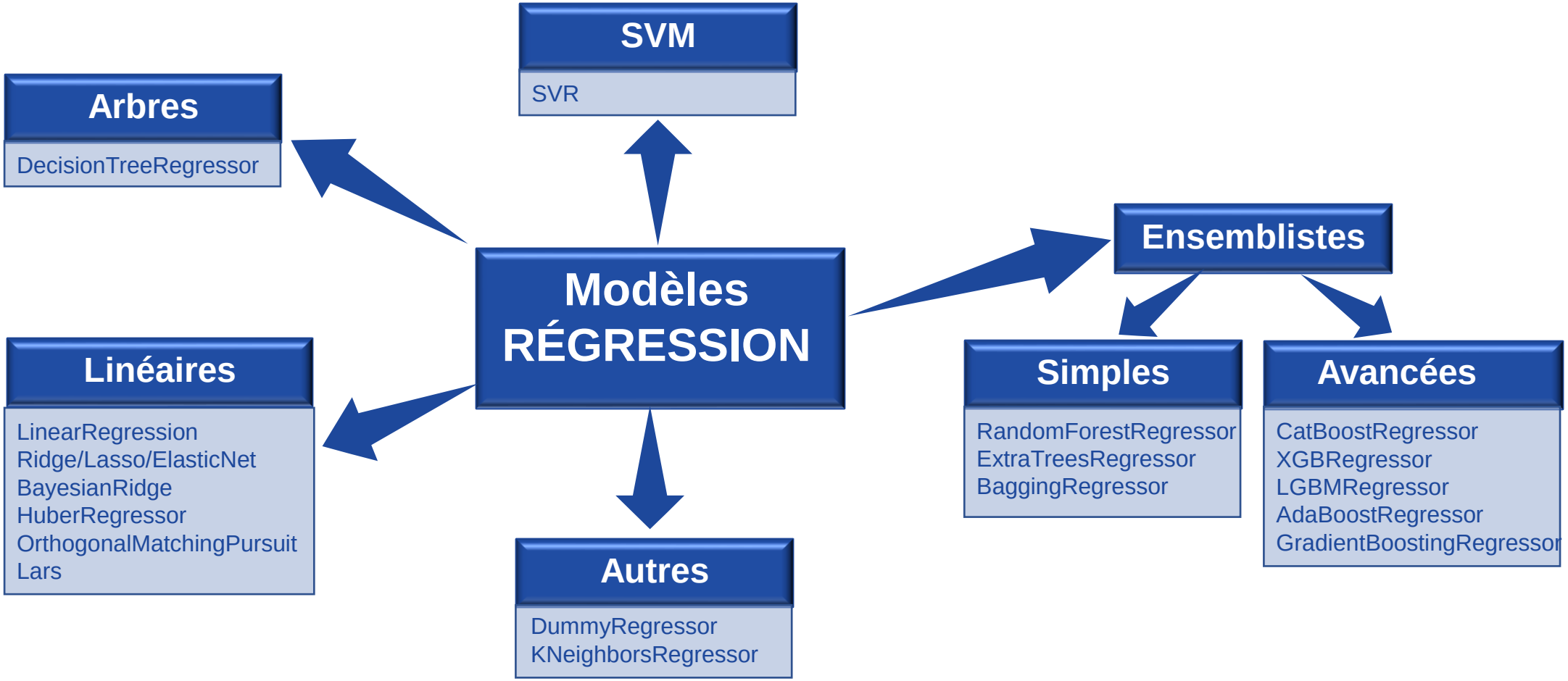

**Variables catégorielles : encodage** avec encoder = **TargetEncoder**

**Variables numériques : standardisation** avec scaler = **RobustScaler**

# Modélisation - Consommation d'énergie

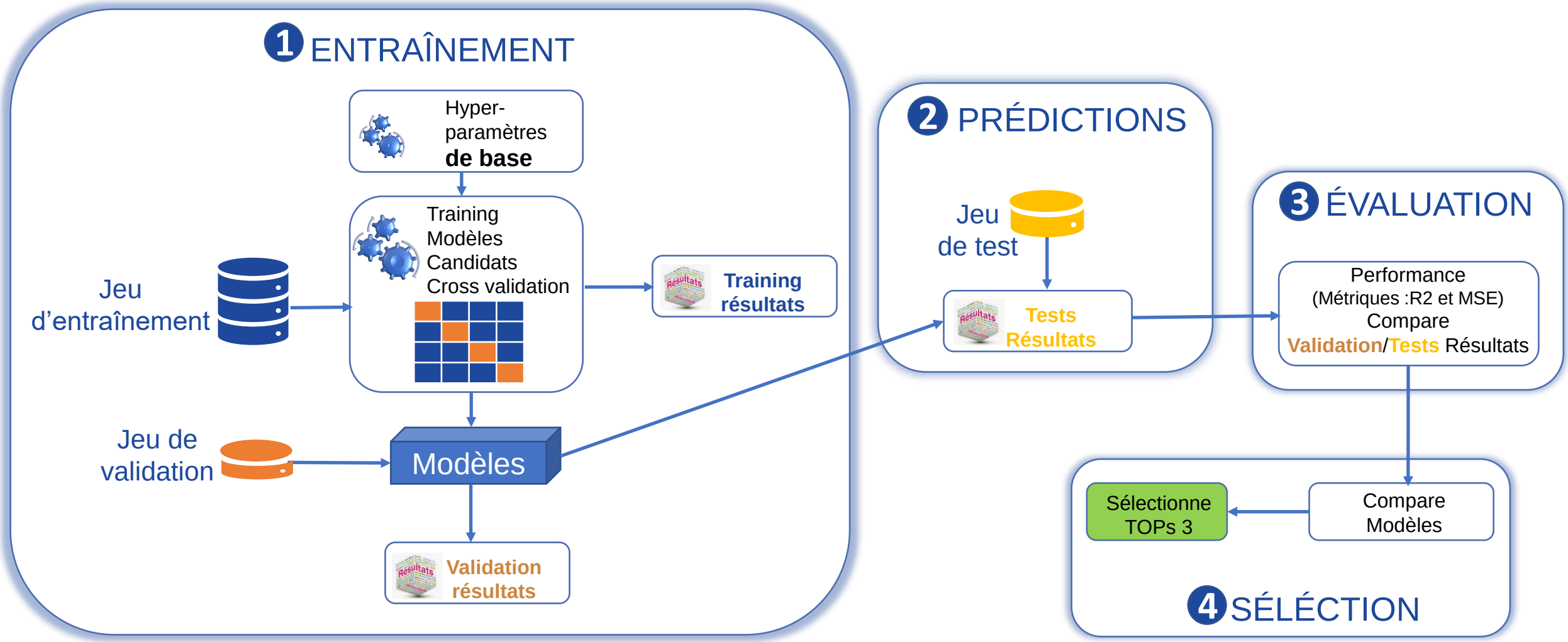
## SÉLECTION MODÈLES DE BASE

Cible **SiteEnergyUseWNLog** numérique ➔ RÉGRESSION



# Modélisation - Consommation d'énergie

## SÉLECTION MODÈLES DE BASE

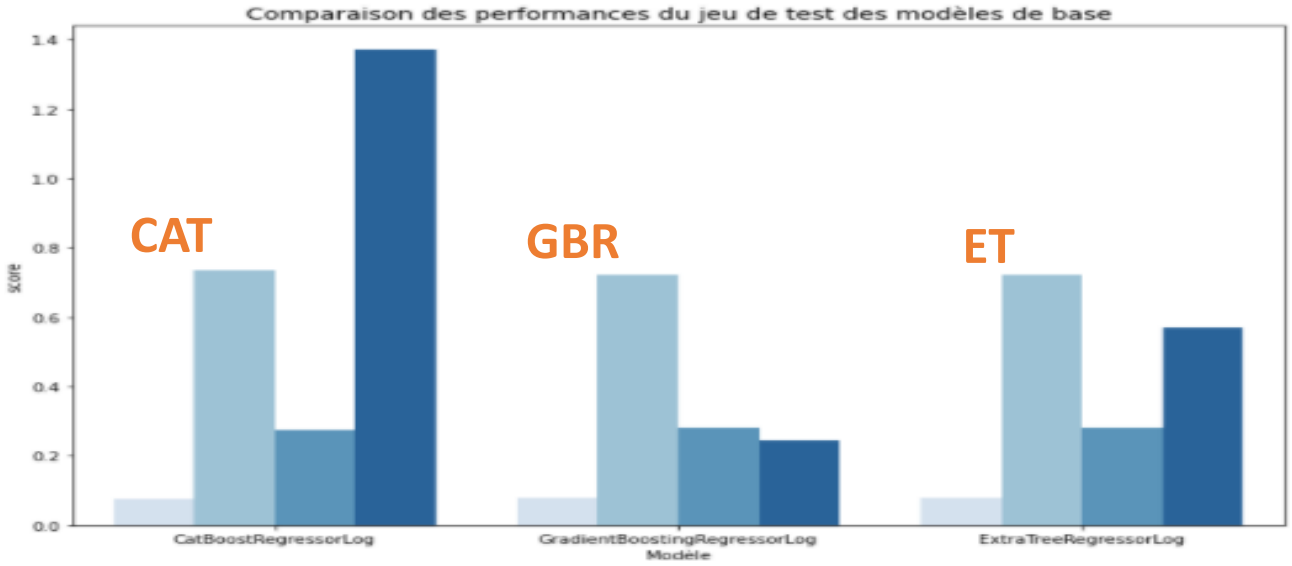


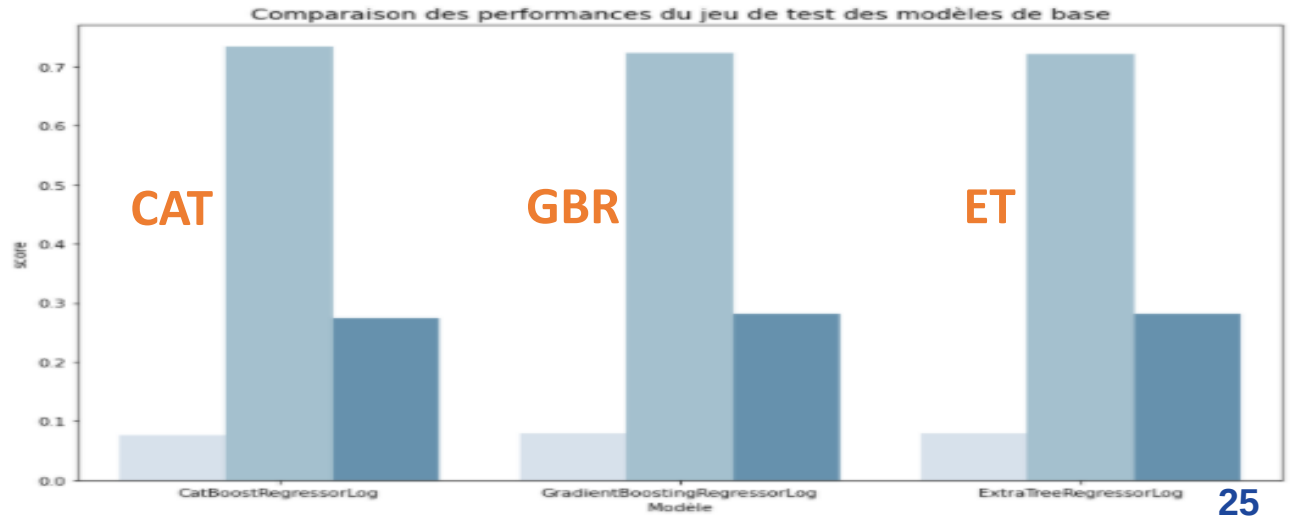
# Modélisation - Consommation d'énergie

## PERFORMANCE - COMPARAISON

| Modèle | MSE | RMSE | R2 | Durée | MAE |
|---|---|---|---|---|---|
| CatBoostRegressorLog | 7.58033e-02 | 2.75324e-01 | 0.73443 | 1.37266 | 2.06013e-01 |
| GradientBoostingRegressorLog | 7.92383e-02 | 2.81493e-01 | 0.72239 | 0.24534 | 2.09047e-01 |
| ExtraTreeRegressorLog | 7.93401e-02 | 2.81674e-01 | 0.72204 | 0.56947 | 2.06397e-01 |
| RandomForestRegressorLog | 7.96759e-02 | 2.82269e-01 | 0.72086 | 0.77492 | 2.10061e-01 |
| BaggingRegressorLog | 8.86998e-02 | 2.97825e-01 | 0.68924 | 0.08477 | 2.21992e-01 |
| LightGradientBoostingMachineLog | 9.03060e-02 | 3.00510e-01 | 0.68362 | 0.05884 | 2.19114e-01 |
| XGBoostRegressorLog | 9.58520e-02 | 3.09600e-01 | 0.66419 | 0.11270 | 2.34563e-01 |
| AdaBoostRegressorLog | 1.00760e-01 | 3.17427e-01 | 0.64699 | 0.13265 | 2.43468e-01 |
| SupportVectorRegressorRbfLog | 1.07384e-01 | 3.27695e-01 | 0.62379 | 0.07978 | 2.45054e-01 |
| KNRegressorLog | 1.08584e-01 | 3.29521e-01 | 0.61958 | 0.01995 | 2.49657e-01 |
| LinearRegressionLassoLog | 1.23257e-01 | 3.51080e-01 | 0.56818 | 0.00299 | 2.76302e-01 |
| LinearRegressionElasticnetLog | 1.23257e-01 | 3.51080e-01 | 0.56818 | 0.00299 | 2.76302e-01 |
| BayesianRidgeLog | 1.23556e-01 | 3.51505e-01 | 0.56713 | 0.00399 | 2.76846e-01 |
| LinearRegressionRidgeLog | 1.24145e-01 | 3.52342e-01 | 0.56507 | 0.00356 | 2.75945e-01 |
| LassoLeastAngleRegressionLog | 1.24263e-01 | 3.52509e-01 | 0.56465 | 0.00672 | 2.78138e-01 |
| LinearRegressionLog | 1.24263e-01 | 3.52509e-01 | 0.56465 | 0.00399 | 2.78138e-01 |
| HubertRegressorLog | 1.30794e-01 | 3.61655e-01 | 0.54177 | 0.05286 | 2.75836e-01 |
| DecisionTreeRegressorLog | 1.69905e-01 | 4.12195e-01 | 0.40475 | 0.01396 | 2.96683e-01 |
| OrthogonalMatchingPursuitLog | 1.81107e-01 | 4.25567e-01 | 0.36550 | 0.00369 | 3.11450e-01 |
| DummyRegressorMedianLog | 2.86412e-01 | 5.35174e-01 | -0.00343 | 0.00100 | 4.12294e-01 |
| LinearRegressionLassoLog | 2.86581e-01 | 5.35333e-01 | -0.00402 | 0.00299 | 4.16658e-01 |
| DummyRegressorMeanLog | 2.86581e-01 | 5.35333e-01 | -0.00402 | 0.00000 | 4.16658e-01 |

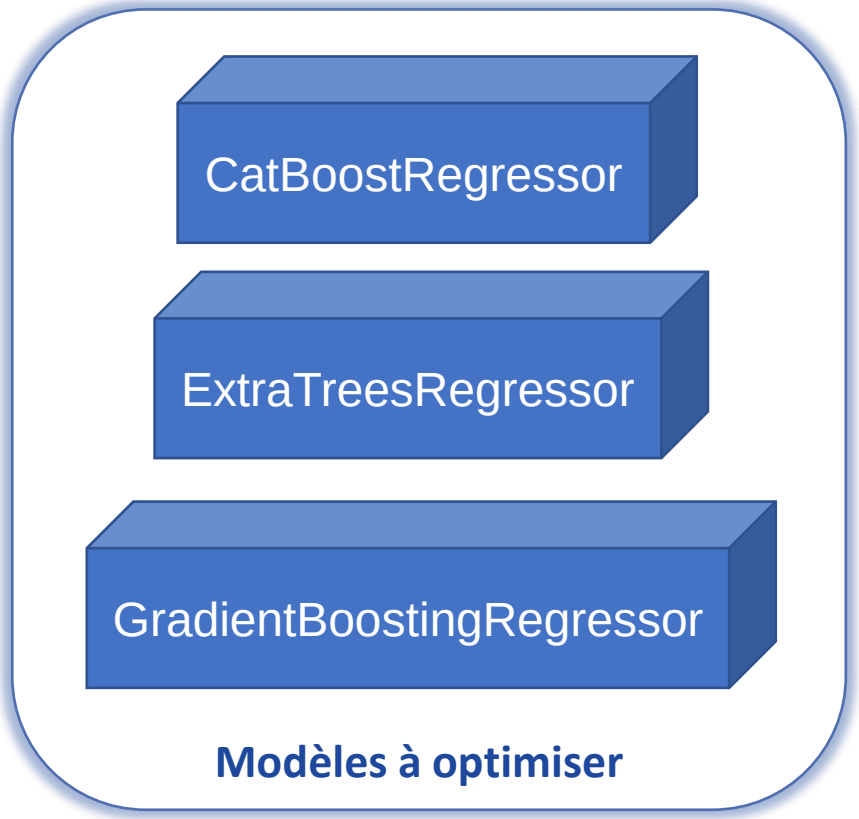
Tops 3



# Modélisation - Consommation d'énergie

## OPTIMISATION




OPTIMISATION

Hyperparamètres

**TUNING**

Recherche manuelle

Recherche automatique

- Randomized Search CV
- Grid Search CV
- Bayésienne hyperopt

# Modélisation - Consommation d'énergie

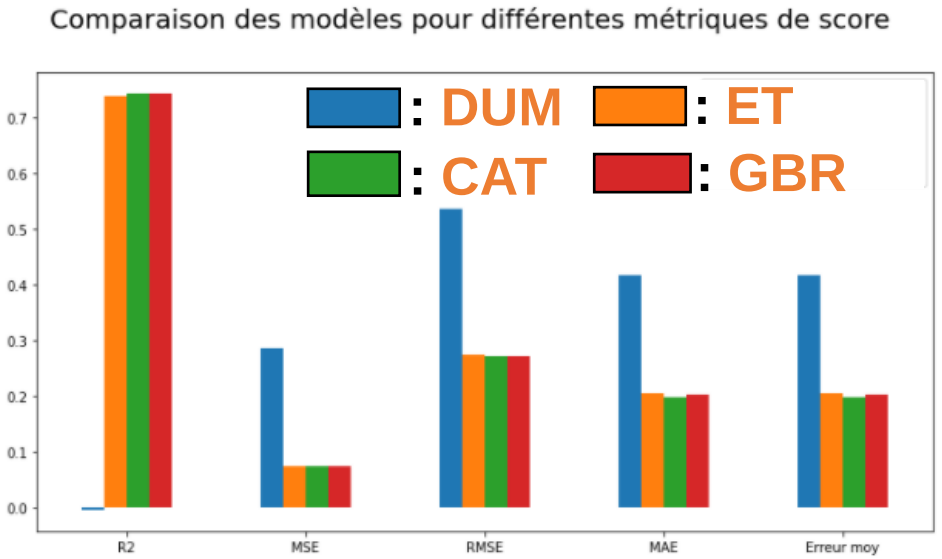
## RÉGLAGE DES HYPERPARAMÈTRES

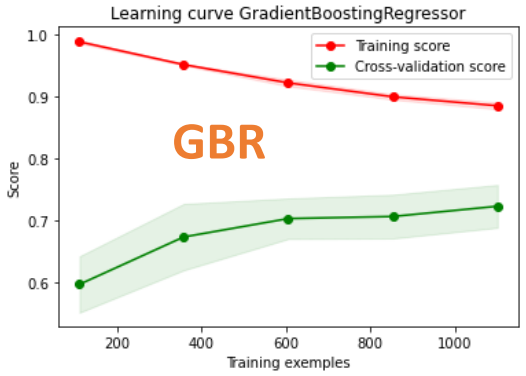
| Modèle | Hyperparamètre | Défaut | Grille de recherche | Meilleure performance |
|---|---|---|---|---|
| ExtraTreesRegressor | n_estimators | 100 | [260, 270, **280**, 290, **300, 310**] | **270** |
| | max_features | auto | ['auto', None] | **auto** |
| | max_depth | None | [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None] | **50** |
| | min_samples_split | 2 | [1, 2, 3, 4, 5, 6, 10, **15, 16, 18, 19, 20, 21**] | **15** |
| | min_samples_leaf | 1 | [1, **2**, 3, **4**, 5, 6, 10, 15] | **2** |
| CatBoostRegressor | depth | 5 | [4, 5, 6] | **6** |
| | iterations | 1000 | [1000, **1100**] | **1100** |
| | l2_leaf_reg | 3 | [1, 2, **3, 4**] | **3** |
| | learning_rate | 0,03 | [**0.03, 0.04**]} | **0,03** |
| GradientBoostingRegressor | n_estimators | 100 | [**107,108**, 109,110,111] | **107** |
| | min_samples_split | 2 | [**19, 21**] | **21** |
| | min_samples_leaf | 1 | [1, 2, **3, 4**] | **1** |
| | max_depth | 3 | [**3, 4**, 5, 6] | **4** |
| | learning_rate | 0,1 | [**0.05, 0.1**] | **0,1** |

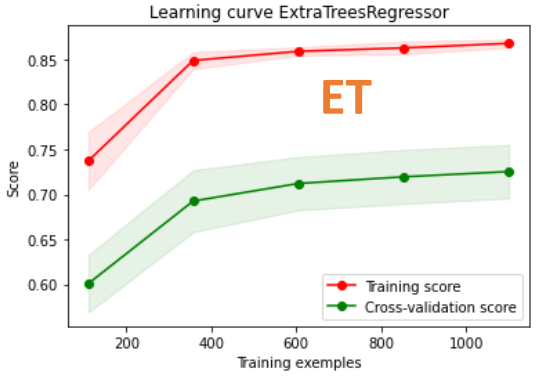
# 3 Modélisation - Consommation d'énergie

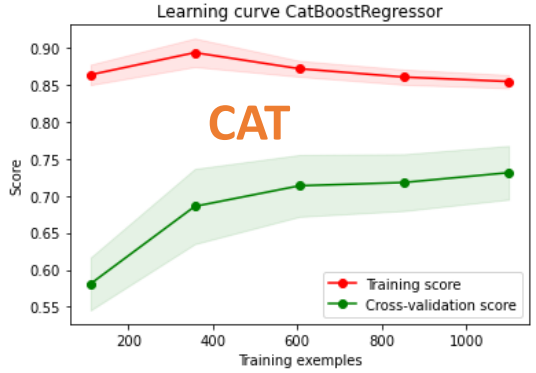
## PERFORMANCE - COMPARAISON

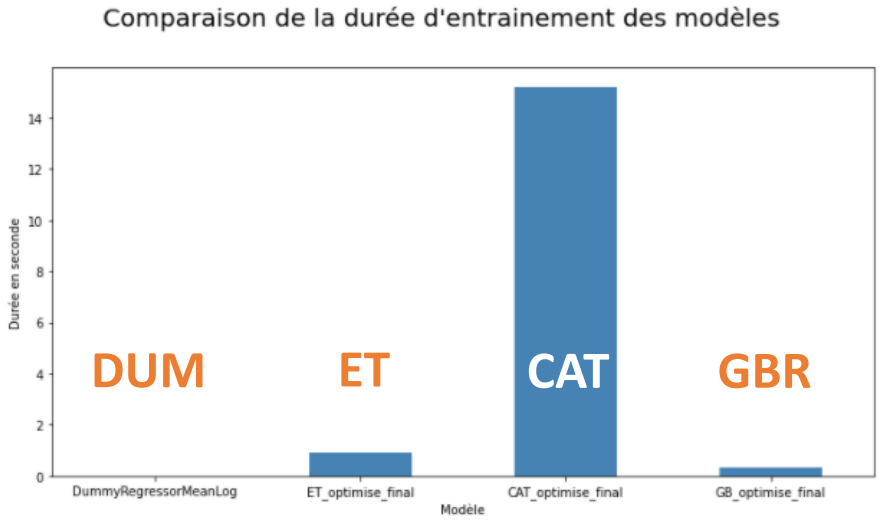
| | Modèle | R2 | MSE | RMSE | MAE | Erreur moy | Précision | Durée |
|---|---|---|---|---|---|---|---|---|
| | DummyRegressorMeanLog | -0.00402 | 0.28658 | 0.53533 | 0.41666 | 0.41666 | 93.45041 | 0.00000 |
| | ET_optimise_final | 0.73639 | 0.07524 | 0.27430 | 0.20467 | 0.20467 | 96.77937 | 0.86869 |
| | CAT_optimise_final | 0.74109 | 0.07390 | 0.27185 | 0.19929 | 0.19929 | 96.87987 | 16.78995 |
| Best | GB_optimise_final | 0.74184 | 0.07369 | 0.27146 | 0.20258 | 0.20258 | 96.82105 | 0.33780 |



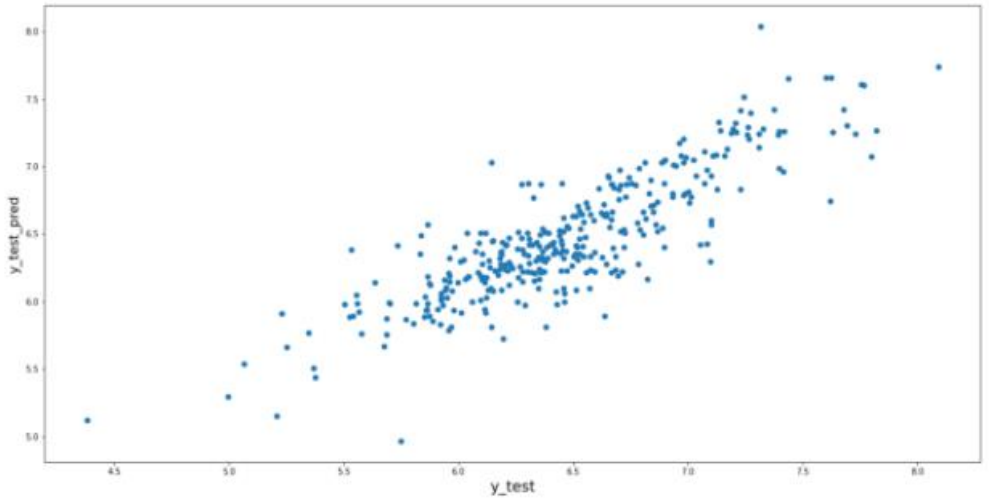



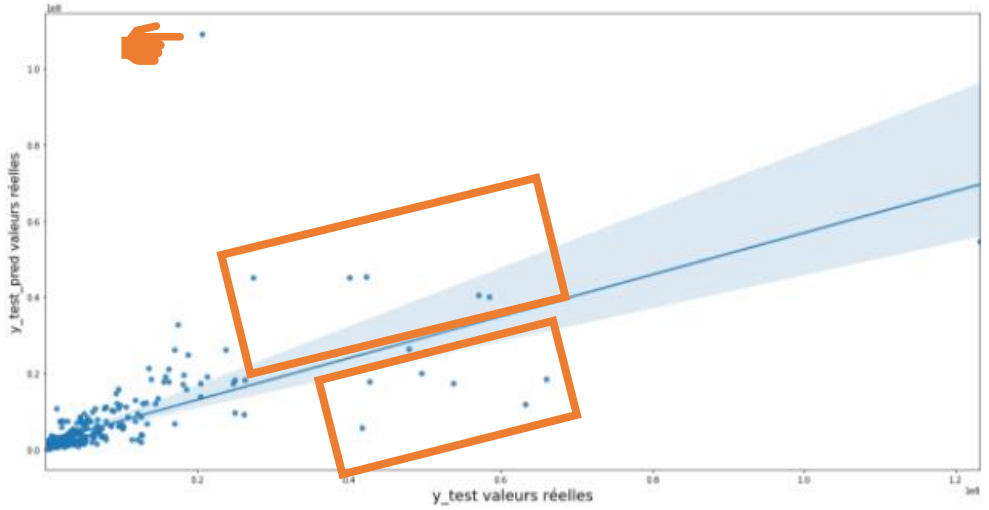

## Modèle FINAL : GradientBoostingRegressor

# 3 Modélisation - Consommation d'énergie

## PRÉDICTIONS du modèle FINAL




y_test valeurs réelles vs y_test_pred valeurs réelles

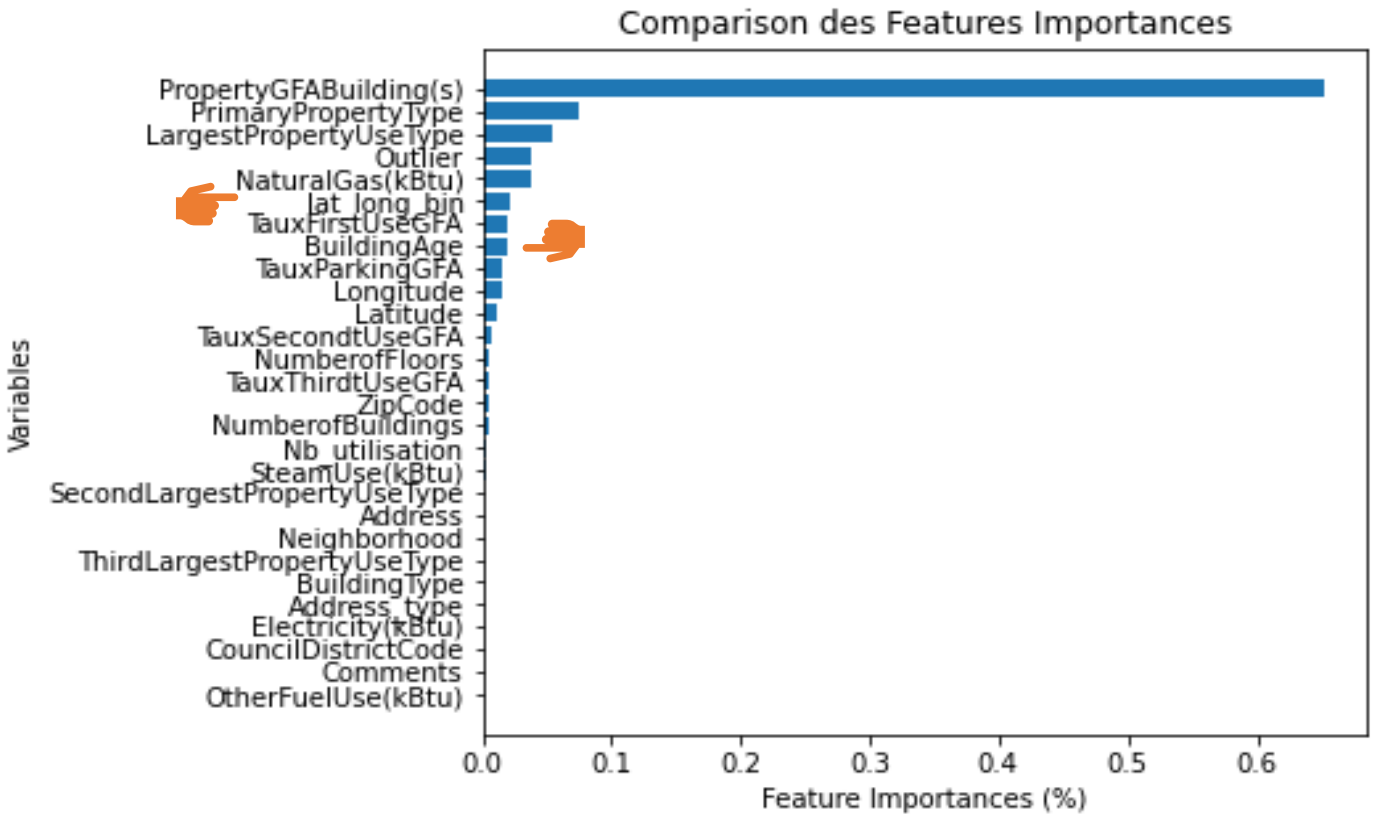



### Plus grandes erreurs

| y_test_pred | y_test_log | erreur_abs | erreur_sens | |
|---|---|---|---|---|
| 6.74595 | 7.61993 | 0.87398 | -0.87398 | Data Center |
| 6.39230 | 5.53212 | 0.86018 | 0.86018 | Ecole |
| 6.99002 | 6.14491 | 0.84511 | 0.84511 | Entrepôt de stockage |
| 6.26628 | 7.09736 | 0.83108 | -0.83108 | Location de bureau |
| 4.98415 | 5.75011 | 0.76596 | -0.76596 | Centre commercial |
| 5.11307 | 4.38213 | 0.73094 | 0.73094 | Ensemble commerce, bureau |
| 6.46380 | 5.73574 | 0.72806 | 0.72806 | Commerce |
| 6.56835 | 5.86602 | 0.70232 | 0.70232 | Salle en location |
| 5.93447 | 6.63624 | 0.70178 | -0.70178 | Centre de service |
| 5.92126 | 5.22908 | 0.69218 | 0.69218 | Entrepôt de stockage |

# Modélisation - Consommation d'énergie

## FEATURES IMPORTANCE




## PERMUTATION IMPORTANCE

| Weight | Feature |
|---|---|
| 0.7858 ± 0.0821 | PropertyGFABuilding(s) |
| 0.0913 ± 0.0371 | PrimaryPropertyType |
| 0.0761 ± 0.0242 | Outlier |
| 0.0566 ± 0.0264 | NaturalGas(kBtu) |
| 0.0207 ± 0.0138 | LargestPropertyUseType |
| 0.0144 ± 0.0094 | NumberofFloors |
| 0.0127 ± 0.0062 | lat_long_bin |
| 0.0078 ± 0.0099 | BuildingAge |
| 0.0064 ± 0.0018 | TauxThirdtUseGFA |
| 0.0060 ± 0.0106 | TauxParkingGFA |
| 0.0055 ± 0.0023 | Nb_utilisation |
| 0.0045 ± 0.0013 | TauxSecondtUseGFA |
| 0.0029 ± 0.0032 | ZipCode |
| 0.0029 ± 0.0006 | BuildingType |
| 0.0024 ± 0.0023 | Neighborhood |
| 0.0024 ± 0.0048 | CouncilDistrictCode |
| 0.0023 ± 0.0032 | Latitude |
| 0.0014 ± 0.0022 | SecondLargestPropertyUseType |
| 0.0014 ± 0.0020 | Longitude |
| 0.0010 ± 0.0014 | SteamUse(kBtu) |

*Annexe 2 - RFECV – ACP – Cons. Énergie*

# Sommaire




Émissions De CO2

# 3 Modélisation – Émissions de $CO_2$

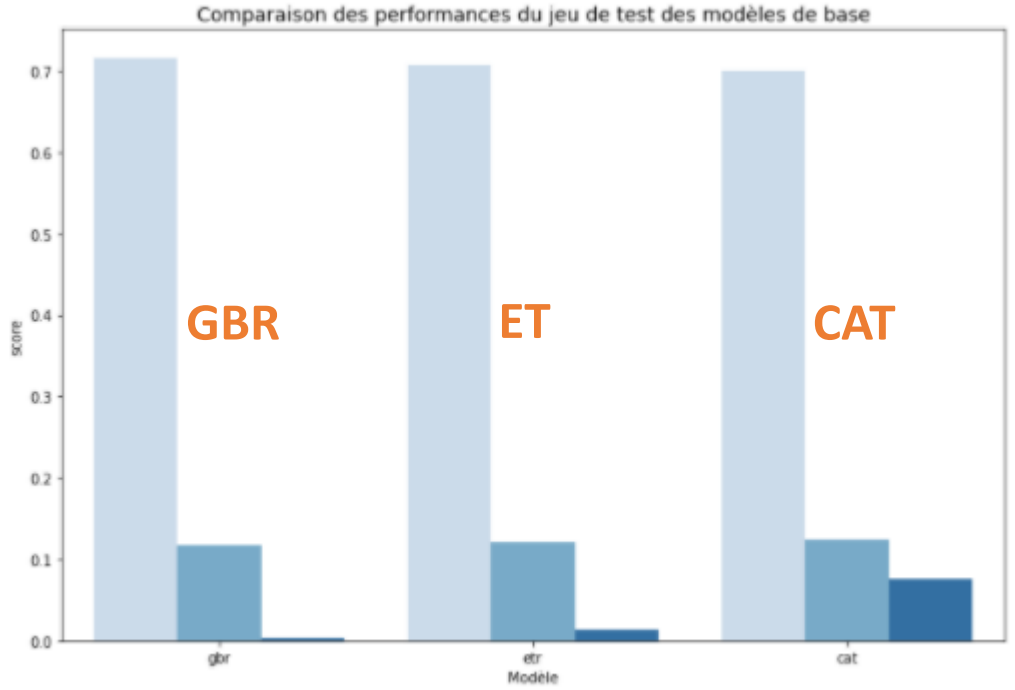
## SÉLECTION MODÈLES DE BASE

| | Modèle | Fit time | Durée | Test R2 CV | Test R2 +/- | Test MSE CV | Train R2 CV | Train R2 +/- | Train MSE CV |
|---|---|---|---|---|---|---|---|---|---|
| Best | gbr | 0.22147 | 0.00265 | 0.71669 | 0.04637 | 0.11801 | 0.82283 | 0.00401 | 0.07404 |
| | etr | 0.78122 | 0.01412 | 0.70708 | 0.04260 | 0.12084 | 1.00000 | 0.00000 | -0.00000 |
| | cat | 2.64042 | 0.07527 | 0.69965 | 0.04757 | 0.12501 | 0.93829 | 0.00231 | 0.02579 |
| | rfr | 0.89302 | 0.01295 | 0.69790 | 0.04749 | 0.12527 | 0.95740 | 0.00118 | 0.01780 |
| | lgbm | 0.06913 | 0.00150 | 0.67604 | 0.05112 | 0.13417 | 0.94571 | 0.00103 | 0.02269 |
| | xgb | 0.11877 | 0.00200 | 0.67454 | 0.04474 | 0.13508 | 0.99320 | 0.00083 | 0.00284 |
| | bag | 0.07096 | 0.00224 | 0.67193 | 0.04704 | 0.13593 | 0.94137 | 0.00198 | 0.02450 |
| | svr | 0.06921 | 0.00331 | 0.66410 | 0.05541 | 0.13982 | 0.72633 | 0.00538 | 0.11437 |
| | ada | 0.12550 | 0.00381 | 0.63138 | 0.05374 | 0.15314 | 0.68191 | 0.00906 | 0.13293 |
| | br | 0.00401 | 0.00100 | 0.59252 | 0.04940 | 0.16943 | 0.62581 | 0.00531 | 0.15637 |
| | ridge | 0.00222 | 0.00155 | 0.59141 | 0.05259 | 0.17011 | 0.61633 | 0.00574 | 0.16034 |
| | lars | 0.00606 | 0.00101 | 0.58741 | 0.05064 | 0.17136 | 0.62710 | 0.00526 | 0.15583 |
| | lin | 0.00347 | 0.00150 | 0.58741 | 0.05064 | 0.17136 | 0.62710 | 0.00526 | 0.15583 |
| | hr | 0.04377 | 0.00290 | 0.57716 | 0.06972 | 0.17588 | 0.61556 | 0.00657 | 0.16065 |
| | sgd | 0.00290 | 0.00320 | 0.56926 | 0.04960 | 0.17941 | 0.59081 | 0.01142 | 0.17101 |
| | knr | 0.00456 | 0.00814 | 0.52896 | 0.04808 | 0.19566 | 0.69275 | 0.00620 | 0.12839 |
| | dt | 0.01273 | 0.00073 | 0.42959 | 0.10199 | 0.23544 | 1.00000 | 0.00000 | -0.00000 |
| | omp | 0.00283 | 0.00061 | 0.40297 | 0.07668 | 0.24781 | 0.41417 | 0.00848 | 0.24482 |
| | en | 0.00286 | 0.00201 | 0.14867 | 0.04181 | 0.35443 | 0.15461 | 0.01131 | 0.35329 |
| | lasso | 0.00206 | 0.00181 | -0.00577 | 0.00582 | 0.41845 | -0.00000 | 0.00000 | 0.41791 |
| | dum_mean | 0.00070 | 0.00030 | -0.00577 | 0.00582 | 0.41845 | 0.00000 | 0.00000 | 0.41791 |
| | dum_med | 0.00060 | 0.00050 | -0.00767 | 0.00902 | 0.41916 | -0.00208 | 0.00055 | 0.41878 |

Cible **TotalGHGEmissionsLog** numérique ➔ RÉGRESSION

Même démarche : split, encodage, standardisation, modèle de base

**: Test R2 CV : Test MSE CV : Durée**




Modèle à **OPTIMISER** : **GradientBoostingRegressor**

# Modélisation – Émissions de $CO_2$

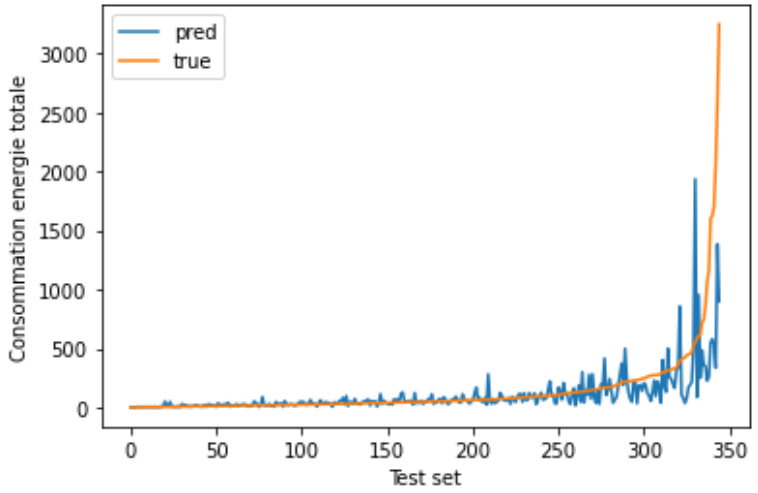
## OPTIMISATION

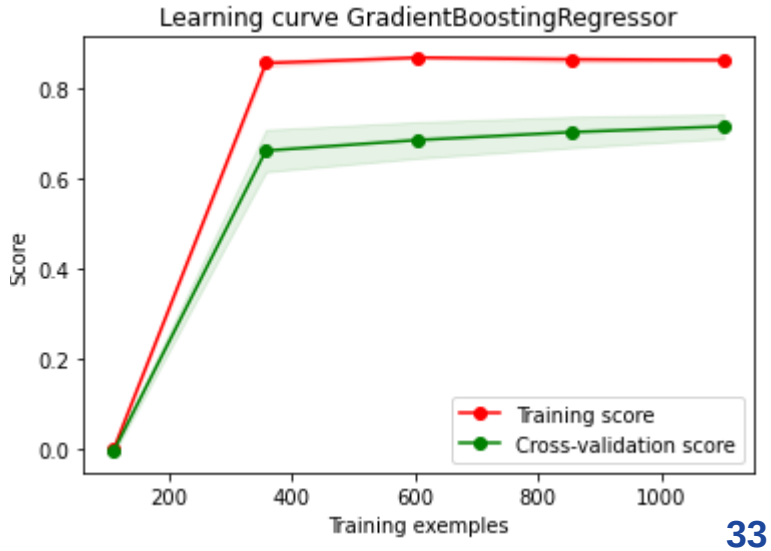
Recherche manuelle itérative avec GridSearch CV, automatique par Randomized Search CV et hyperopt

| Modèle | Hyperparamètre | Défaut | Grille de recherche | Meilleure performance |
|---|---|---|---|---|
| GradientBoostingRegressor | n_estimators | 100 | range(1000, 2000, 100) | **1400** |
| | min_samples_split | 2 | range(1, 200, 10) | **101** |
| | min_samples_leaf | 1 | [1, 2, 4] | **1** |
| | max_depth | 3 | [5, 6, 7, 8, 9, 10] | **7** |
| | learning_rate | 0,1 | [0.001, 0.005, 0.01, 0.02, 0.03] | **0.005** |
| | subsample | 1,0 | [0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 1] | **0,65** |

## PERFORMANCE du modèle FINAL

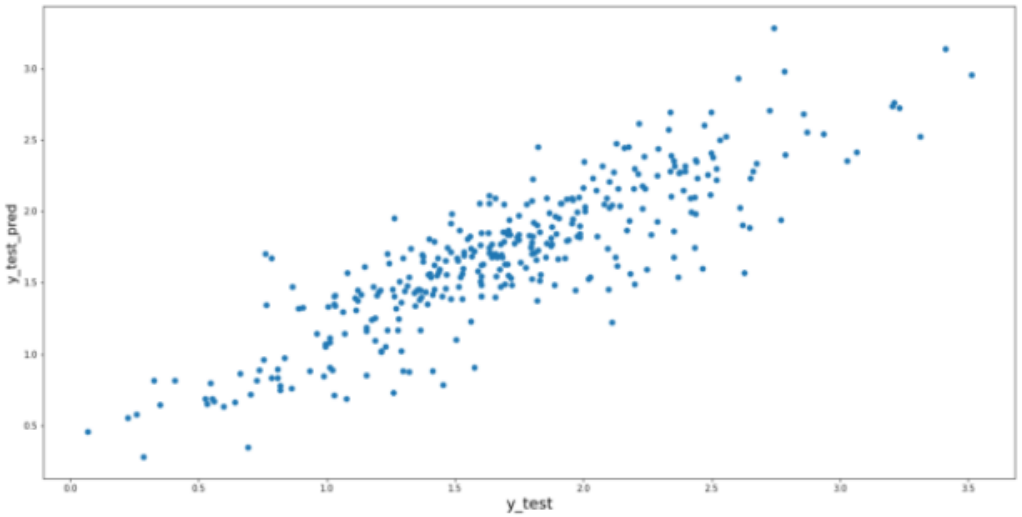
| Modèle | R2 | MSE | RMSE | MAE | Durée |
|---|---|---|---|---|---|
| GB_opt_man_sans_ess_total | 0.74531 | 0.09244 | 0.30403 | 0.23057 | 4.55528 |



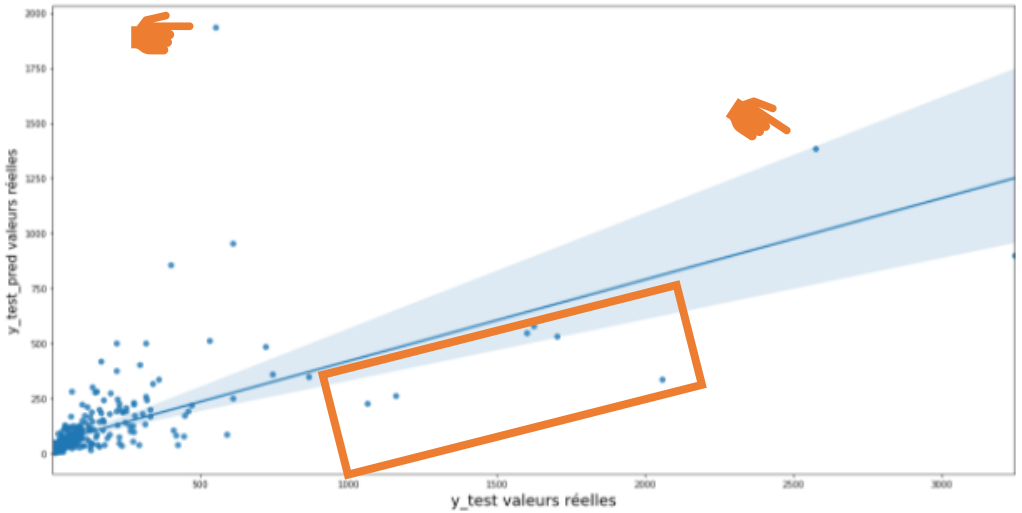



# Modélisation – Émissions de $CO_2$

## PRÉDICTIONS




y_test valeurs réelles vs y_test_pred valeurs réelles

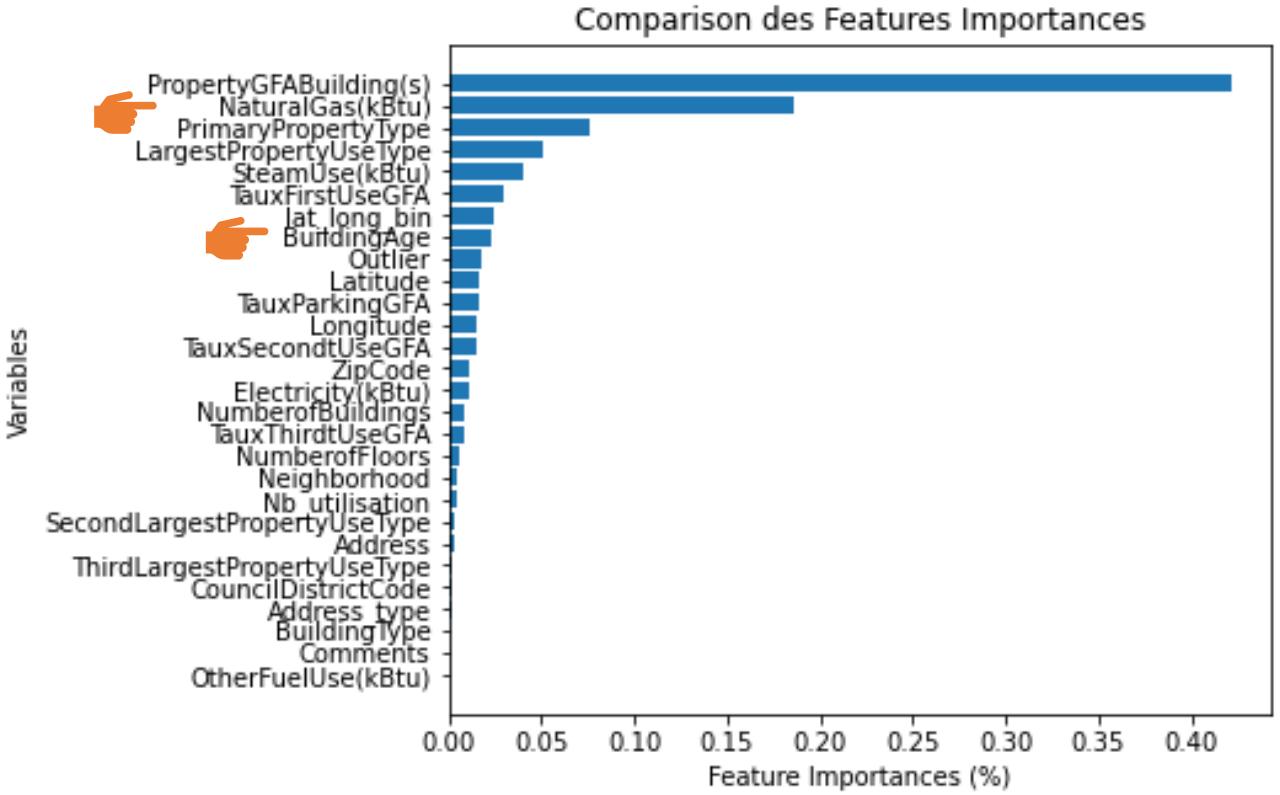



### Plus grandes erreurs

| y_test_pred | y_test_log | erreur_abs | erreur_sens | |
|---|---|---|---|---|
| 1.57763 | 2.62680 | 1.04918 | -1.04918 | Location de bureau * |
| 1.70396 | 0.76118 | 0.94278 | 0.94278 | Ecole * |
| 1.71497 | 0.78462 | 0.93035 | 0.93035 | Centre d'alimentation thermique |
| 1.20124 | 2.11083 | 0.90959 | -0.90959 | Ensemble commerce, bureau * |
| 1.88873 | 2.76950 | 0.88076 | -0.88076 | Centre de recherche |
| 1.62515 | 2.46474 | 0.83959 | -0.83959 | Data center * |
| 1.54777 | 2.36758 | 0.81981 | -0.81981 | Concessionnaire véhicules |
| 2.52917 | 3.31320 | 0.78402 | -0.78402 | Centre de recherche |
| 1.90159 | 2.64634 | 0.74476 | -0.74476 | Bureau |
| 1.63977 | 2.35129 | 0.71153 | -0.71153 | Piscine |

# Modélisation – Émissions de $CO_2$

## FEATURES IMPORTANCE




## PERMUTATION IMPORTANCE

| Weight | Feature |
|---|---|
| 0.5817 ± 0.0685 | PropertyGFABuilding(s) |
| 0.3677 ± 0.0580 | NaturalGas(kBtu) |
| 0.1024 ± 0.0224 | PrimaryPropertyType |
| 0.0493 ± 0.0154 | SteamUse(kBtu) |
| 0.0259 ± 0.0113 | LargestPropertyUseType |
| 0.0225 ± 0.0101 | Outlier |
| 0.0180 ± 0.0041 | lat_long_bin |
| 0.0128 ± 0.0047 | BuildingAge |
| 0.0078 ± 0.0082 | TauxParkingGFA |
| 0.0035 ± 0.0012 | SecondLargestPropertyUseType |
| 0.0029 ± 0.0052 | TauxFirstUseGFA |
| 0.0023 ± 0.0038 | NumberofFloors |
| 0.0020 ± 0.0013 | Nb_utilisation |
| 0.0017 ± 0.0034 | Latitude |
| 0.0013 ± 0.0021 | Longitude |
| 0.0013 ± 0.0010 | Neighborhood |
| 0.0012 ± 0.0015 | ThirdLargestPropertyUseType |
| 0.0012 ± 0.0009 | CouncilDistrictCode |
| 0.0007 ± 0.0003 | BuildingType |
| 0.0003 ± 0.0007 | Address |

*Annexe 3 – RFECV - Émissions de $CO_2$*

# Sommaire

1 Problématique

2 Données

3 Modélisation

4 Conclusion

Intérêt de l'EnergyStar sur l'émission de $CO_2$

# Modélisation – Intérêt

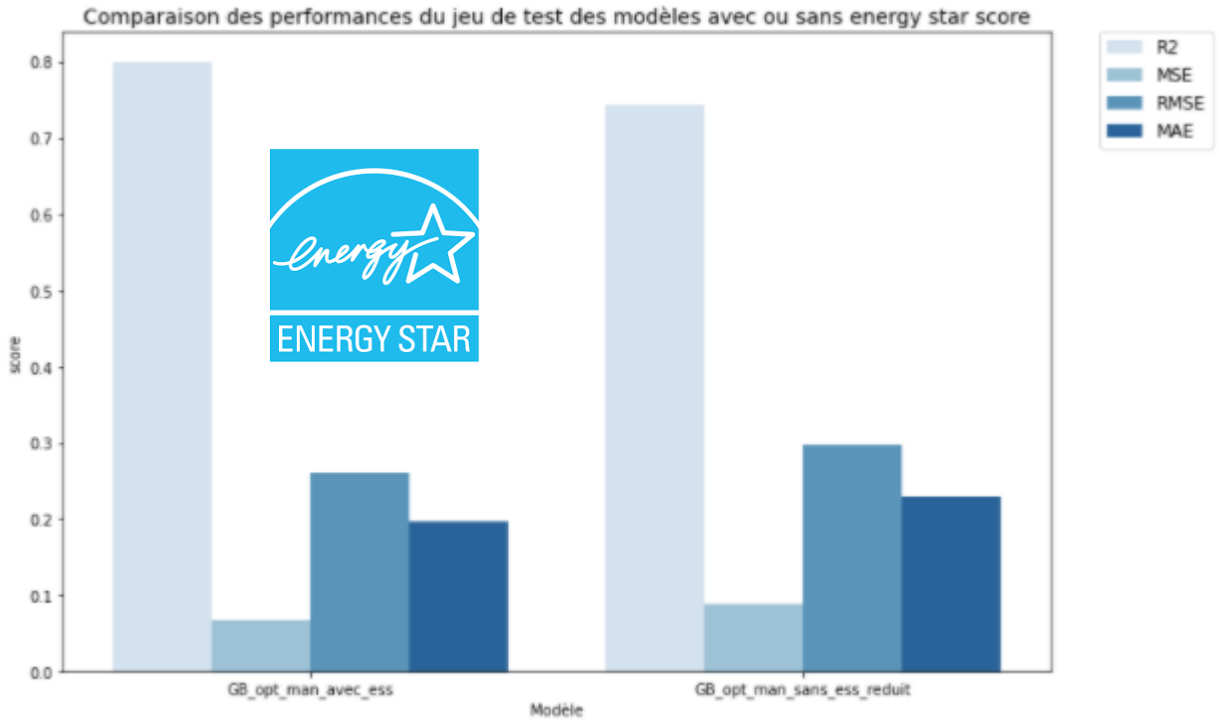
## COMPARAISON

Avec ou sans **ENERGYSTARScore**
Même démarche : split, encodage, standardisation, modèle réduit (36% de valeurs manquantes)
Optimisation d'un modèle GradientBoostingRegressor réduit sans ESS et un autre avec ESS

| Modèle | R2 | MSE | RMSE | MAE | Durée |
|---|---|---|---|---|---|
| GB_opt_man_avec_ess | 0.80009 | 0.06834 | 0.26143 | 0.19720 | 0.60037 |
| GB_opt_man_sans_ess_reduit | 0.74292 | 0.08789 | 0.29646 | 0.23019 | 5.31093 |

**Hausse : 7% R2**

**Baisse de : 28% MSE**

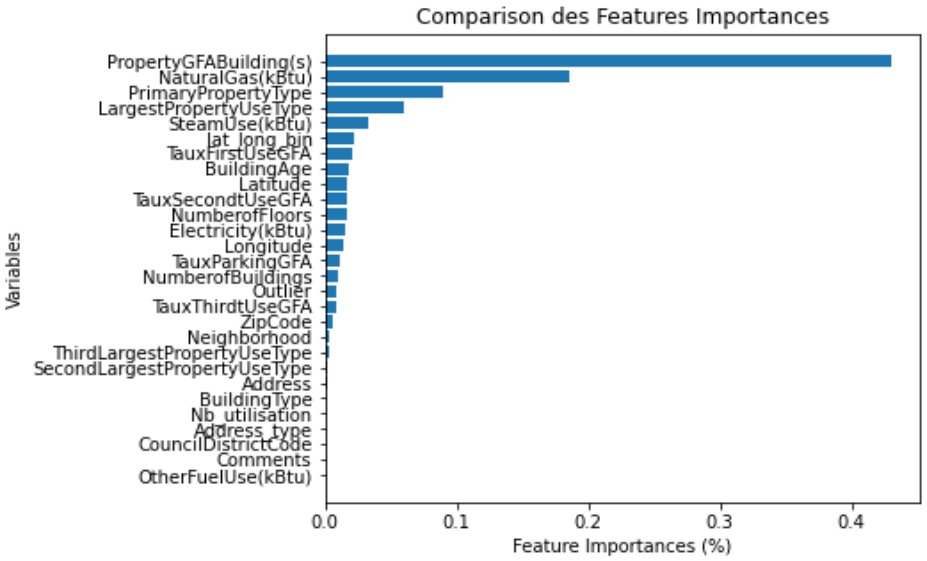


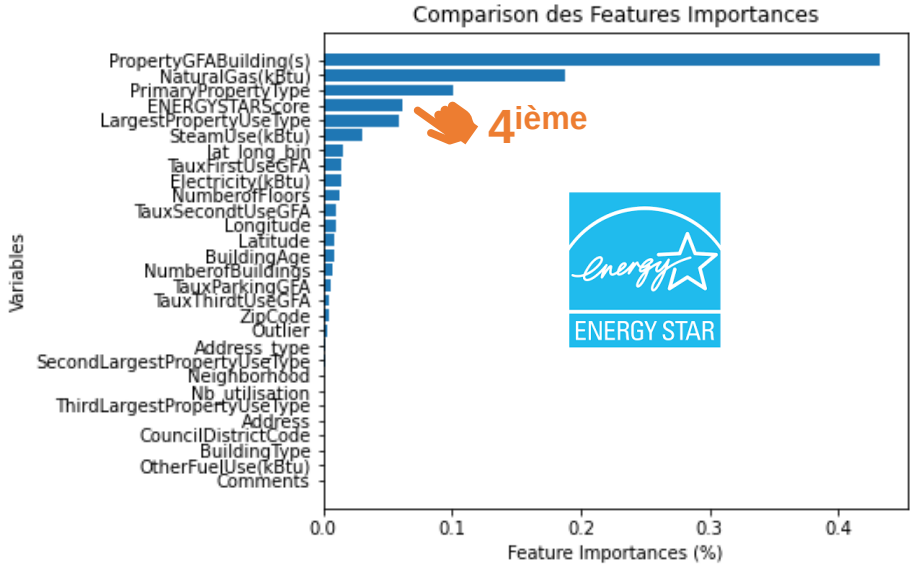

**ENERGYSTARScore** améliore le modèle

Mais **coûteuse** et très **difficile** à obtenir
➔ Arbitrage à faire

# 3 Modélisation – Intérêt

## FEATURES IMPORTANCE




4ième

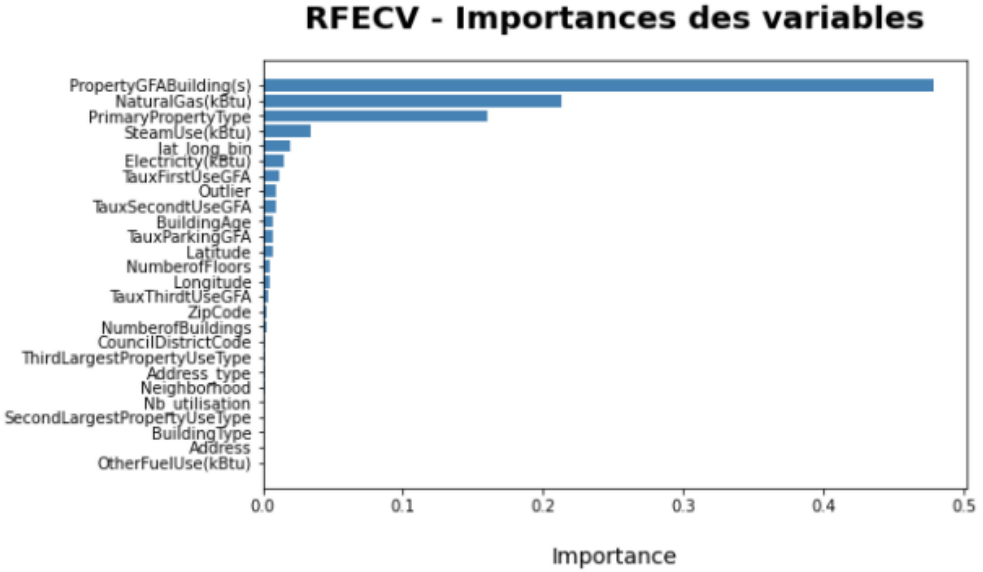
## PERMUTATION IMPORTANCE

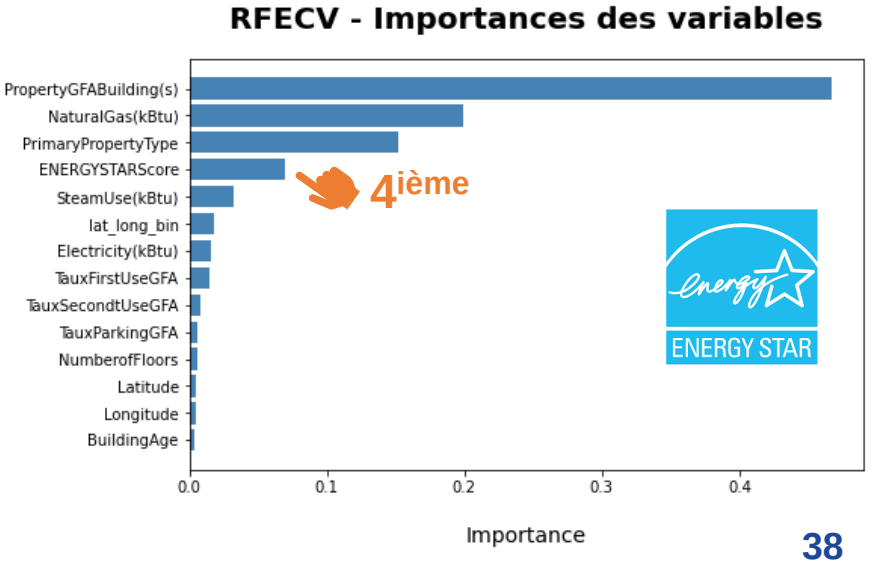
| Weight | Feature |
|---|---|
| 0.6576 ± 0.1222 | PropertyGFABuilding(s) |
| 0.2771 ± 0.0878 | NaturalGas(kBtu) |
| 0.2112 ± 0.0585 | PrimaryPropertyType |
| 0.0841 ± 0.0209 | LargestPropertyUseType |
| 0.0392 ± 0.0159 | Outlier |
| 0.0298 ± 0.0138 | SteamUse(kBtu) |
| 0.0113 ± 0.0043 | BuildingAge |
| 0.0084 ± 0.0078 | NumberofBuildings |
| 0.0067 ± 0.0040 | TauxParkingGFA |
| 0.0037 ± 0.0048 | Longitude |
| 0.0030 ± 0.0014 | ZipCode |
| 0.0027 ± 0.0041 | Neighborhood |
| 0.0021 ± 0.0021 | TauxThirdtUseGFA |
| 0.0016 ± 0.0061 | NumberofFloors |
| 0.0013 ± 0.0011 | BuildingType |
| 0.0007 ± 0.0010 | SecondLargestPropertyUseType |
| 0.0006 ± 0.0008 | ThirdLargestPropertyUseType |
| 0.0005 ± 0.0007 | Address_type |
| 0.0003 ± 0.0005 | CouncilDistrictCode |
| 0.0002 ± 0.0013 | Nb_utilisation |

| Weight | Feature |
|---|---|
| 0.7121 ± 0.0654 | PropertyGFABuilding(s) |
| 0.2851 ± 0.0835 | NaturalGas(kBtu) |
| 0.2118 ± 0.0326 | PrimaryPropertyType |
| 0.1291 ± 0.0302 | ENERGYSTARScore |
| 0.0868 ± 0.0183 | LargestPropertyUseType |
| 0.0357 ± 0.0134 | SteamUse(kBtu) |
| 0.0094 ± 0.0052 | Outlier |
| 0.0062 ± 0.0063 | NumberofBuildings |
| 0.0050 ± 0.0038 | BuildingAge |
| 0.0031 ± 0.0025 | ZipCode |
| 0.0025 ± 0.0032 | NumberofFloors |
| 0.0018 ± 0.0049 | Longitude |
| 0.0013 ± 0.0032 | TauxFirstUseGFA |
| 0.0010 ± 0.0010 | Neighborhood |
| 0.0006 ± 0.0007 | CouncilDistrictCode |
| 0.0006 ± 0.0004 | ThirdLargestPropertyUseType |
| 0.0006 ± 0.0030 | TauxThirdtUseGFA |
| 0.0004 ± 0.0017 | Latitude |
| 0.0002 ± 0.0003 | BuildingType |
| 0.0001 ± 0.0010 | Nb_utilisation |

13%

## RFECV




4ième

# Sommaire

# 4 Pistes d'amélioration

## 1 Jeu de données

Discussion client : bâtiments non résidentiels
récolte des données : site internet
arbitrage 'EnergyStar score'

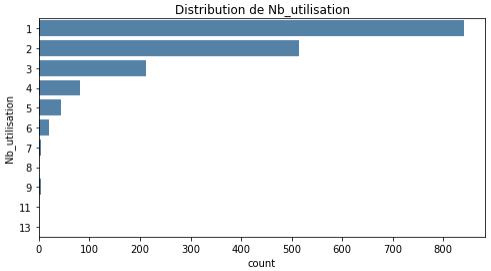
## 2 Modélisation

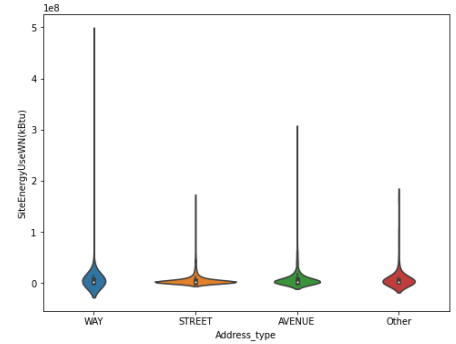
Influence des outliers? 2 modèles ?
ACP : en utilisant moins de composantes (11 éboulis)?
Hyperparamètre : optimisation bayésienne optuna ?
Tester avec les réseaux de neurones?
XGBoost paramètres de base ?

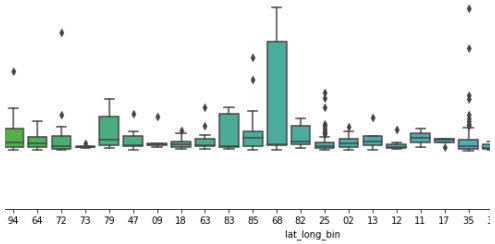
# Annexes

# Annexe 1- Nettoyage - Feature engineering

| Variables | Nouvelle variable | Description |
|---|---|---|
| ListOfAllPropertyUseTypes | Nb_utilisation | Compte le nombre de type de propriété pour chaque bâtiment |
| Latitude/Longitude | lat_long_bin | Cartographie des bâtiments en binérisant la latitude et la longitude et en faisant la somme |
| Address | Address_type | Influence si le bâtiment est dans une rue, avenue, chemin? ➔ WAY, AVENUE ou STREET |
| SteamUse(kBtu),<br>Electricity(kBtu),<br>NaturalGas(kBtu),<br>OtherFuelUse(kBtu) | SteamUse(kBtu),<br>Electricity(kBtu),<br>NaturalGas(kBtu),<br>OtherFuelUse(kBtu) | 0 : n'utilise pas cette énergie,<br>1 : utilise cette énergie. |
| DataYear<br>YearBuilt | BuildingAge | L'âge du bâtiment ou de la dernière rénovation. |

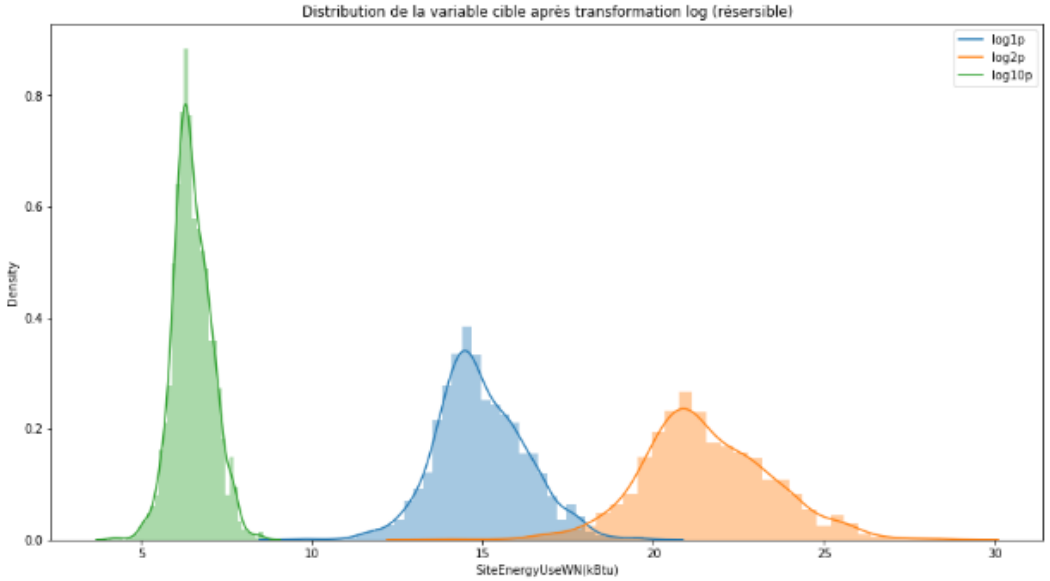


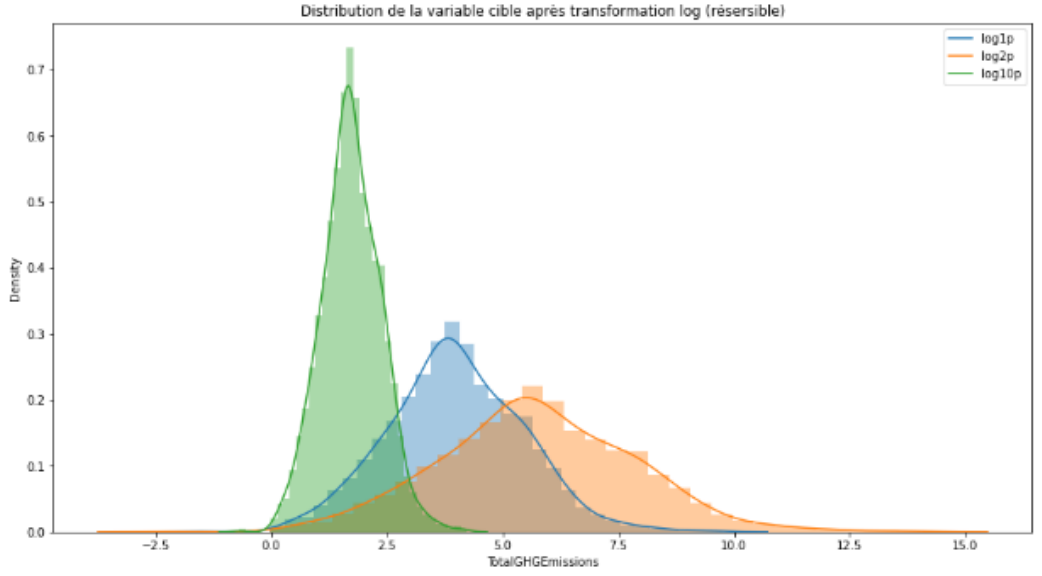



# Annexe 1- Nettoyage - Feature engineering

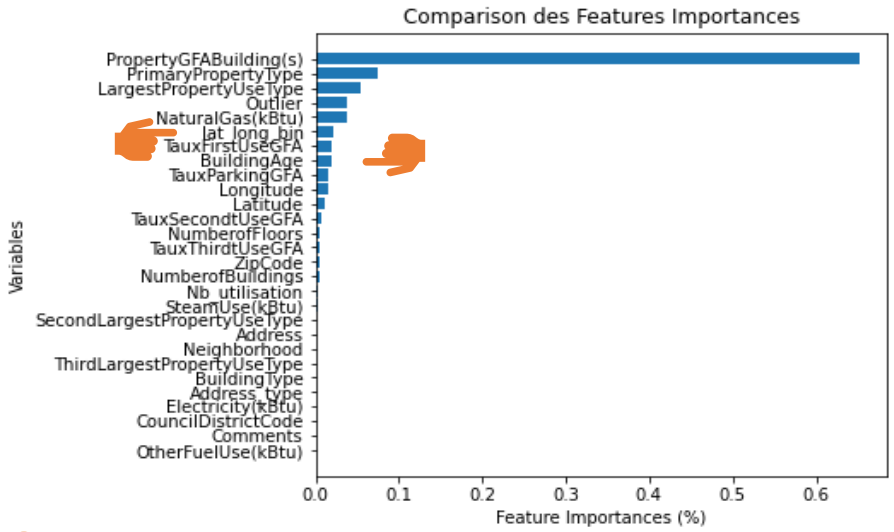
| Variables | Nouvelle variable | |
|---|---|---|
| PropertyGFAParking<br>PropertyGFATotal<br>LargestPropertyUseTypeGFA<br>SecondLargestPropertyUseTypeGFA<br>ThirdLargestPropertyUseTypeGFA | TauxParkingGFA<br>TauxFirstUseGFA<br>TauxSecondtUseGFA<br>TauxThirdtUseGFA | Ratio de la surface du parking sur la surface totale<br>Ratio de la surface de la permière (2nde, 3ième) sur la surface totale |
| SiteEnergyUseWN(kBtu) | SiteEnergyUseWNLog | Transformation en log10 + 1 |
| TotalGHGEmissions | TotalGHGEmissionsLog | Transformation en log10 + 1 |



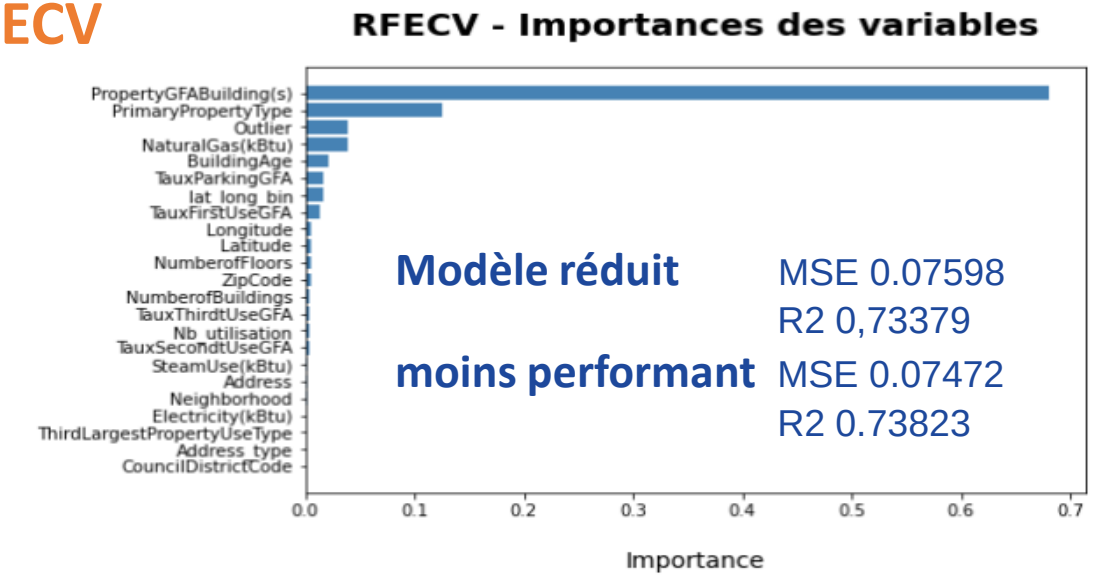



# Annexe 2 - RFECV – ACP – Cons. Énergie

## FEATURES IMPORTANCE



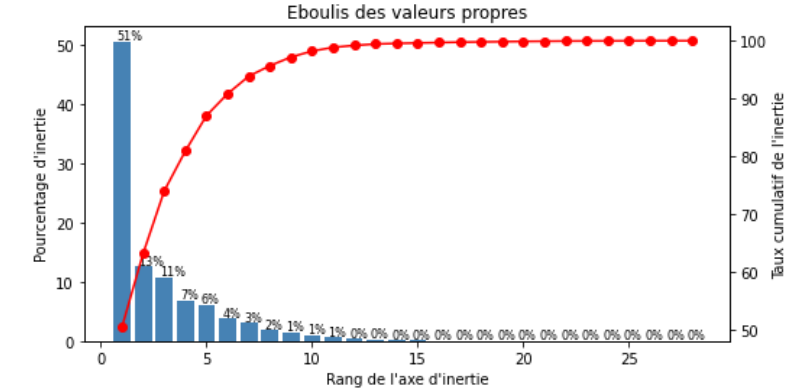

## RFECV




**Modèle réduit** MSE 0.07598 R2 0,73379

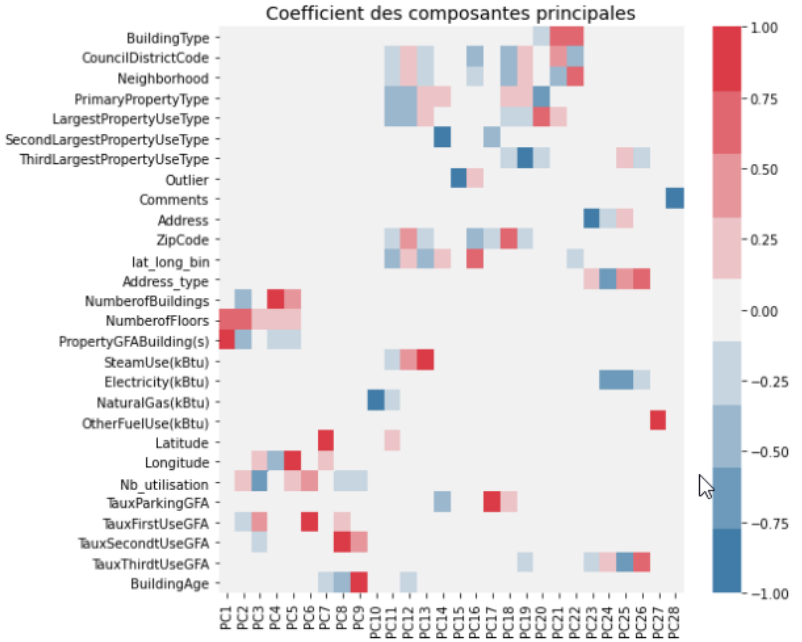
**moins performant** MSE 0.07472 R2 0.73823

## ACP



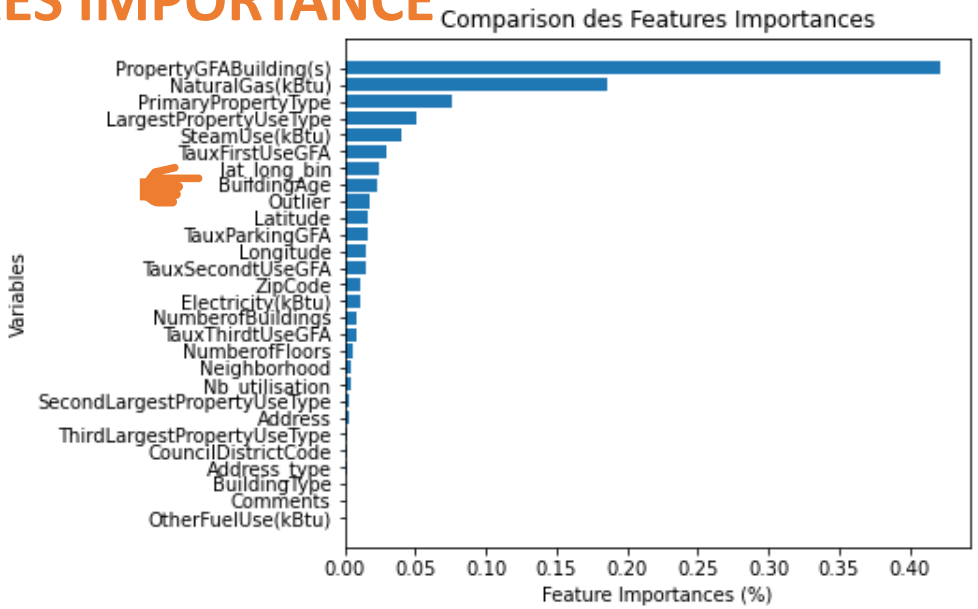

**Modèle ACP** MSE 0.10508 R2 0,634

**moins performant** MSE 0.07472 R2 0.738



# Annexe 3 – RFECV - Émissions de $CO_2$

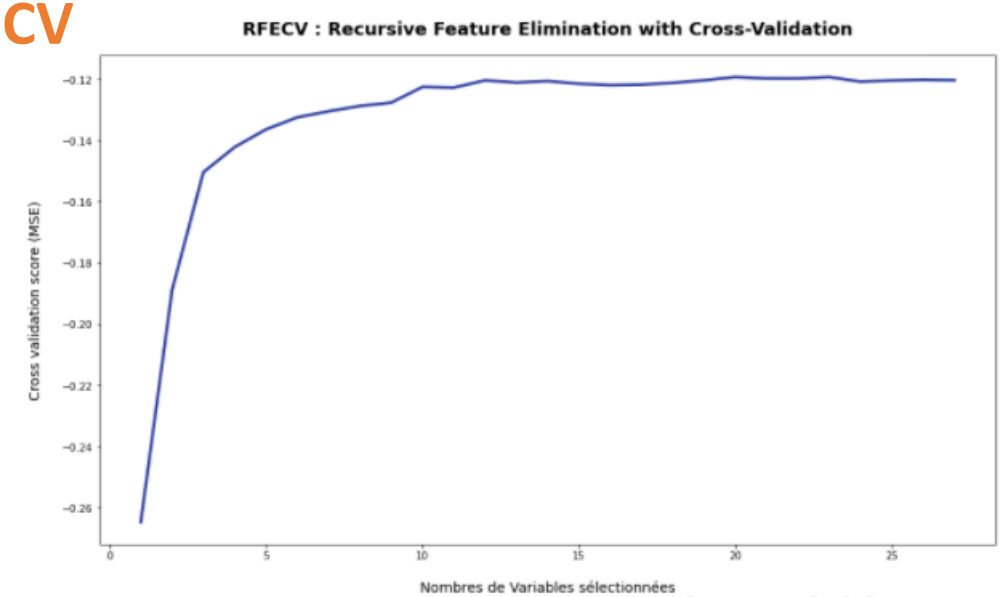
## FEATURES IMPORTANCE

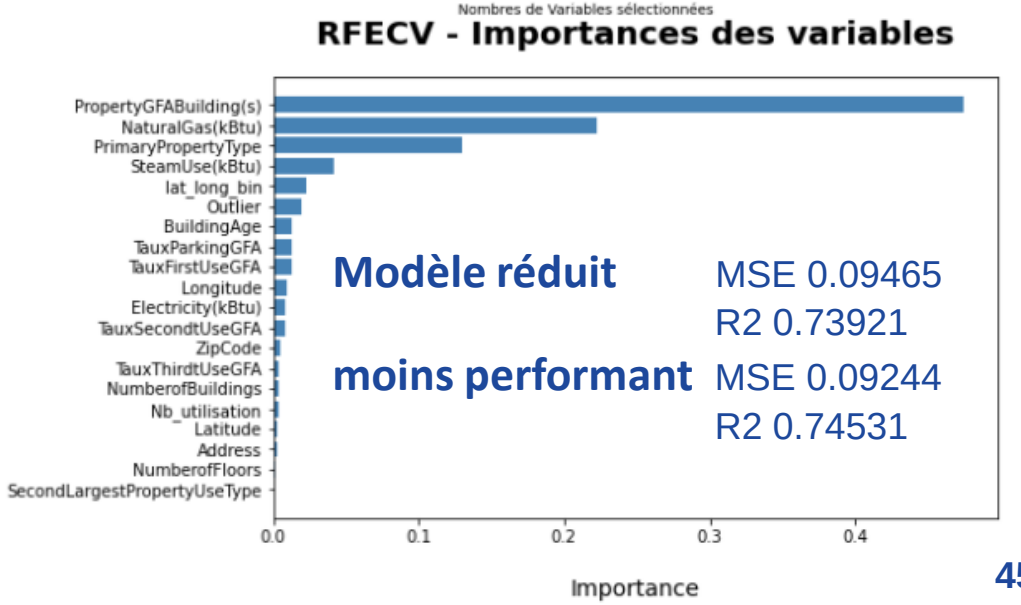



## PERMUTATION IMPORTANCE

| Weight | Feature |
|---|---|
| 0.5817 ± 0.0685 | PropertyGFABuilding(s) |
| 0.3677 ± 0.0580 | NaturalGas(kBtu) |
| 0.1024 ± 0.0224 | PrimaryPropertyType |
| 0.0493 ± 0.0154 | SteamUse(kBtu) |
| 0.0259 ± 0.0113 | LargestPropertyUseType |
| 0.0225 ± 0.0101 | Outlier |
| 0.0180 ± 0.0041 | lat_long_bin |
| 0.0128 ± 0.0047 | BuildingAge |
| 0.0078 ± 0.0082 | TauxParkingGFA |
| 0.0035 ± 0.0012 | SecondLargestPropertyUseType |
| 0.0029 ± 0.0052 | TauxFirstUseGFA |
| 0.0023 ± 0.0038 | NumberofFloors |
| 0.0020 ± 0.0013 | Nb_utilisation |
| 0.0017 ± 0.0034 | Latitude |
| 0.0013 ± 0.0021 | Longitude |
| 0.0013 ± 0.0010 | Neighborhood |
| 0.0012 ± 0.0015 | ThirdLargestPropertyUseType |
| 0.0012 ± 0.0009 | CouncilDistrictCode |
| 0.0007 ± 0.0003 | BuildingType |
| 0.0003 ± 0.0007 | Address |

## RFECV




**Modèle réduit** MSE 0.09465 R2 0.73921

**moins performant** MSE 0.09244 R2 0.74531