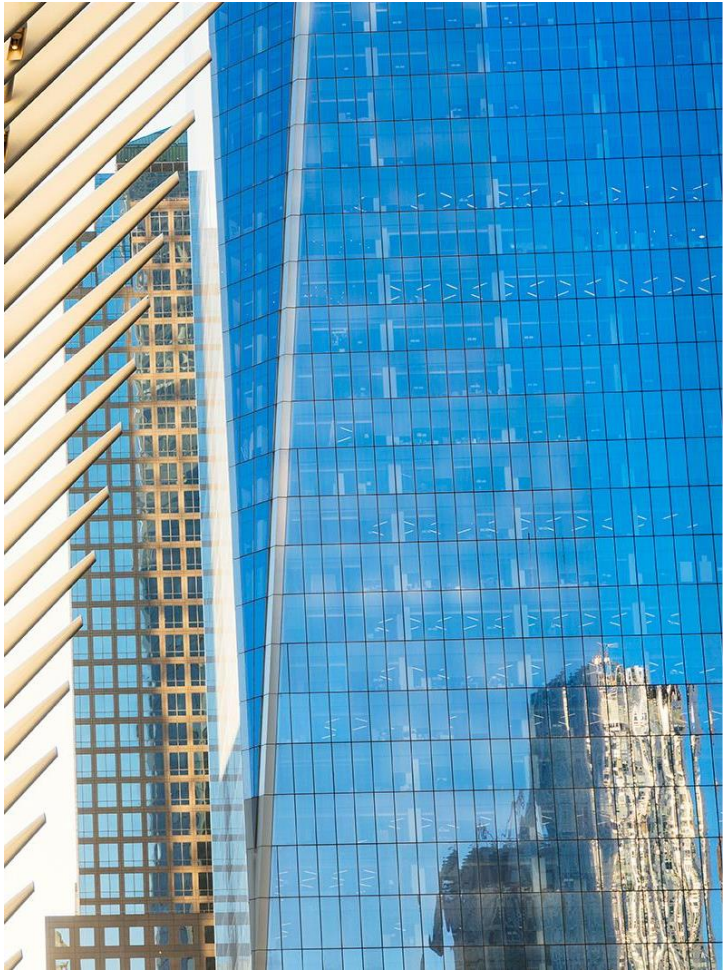
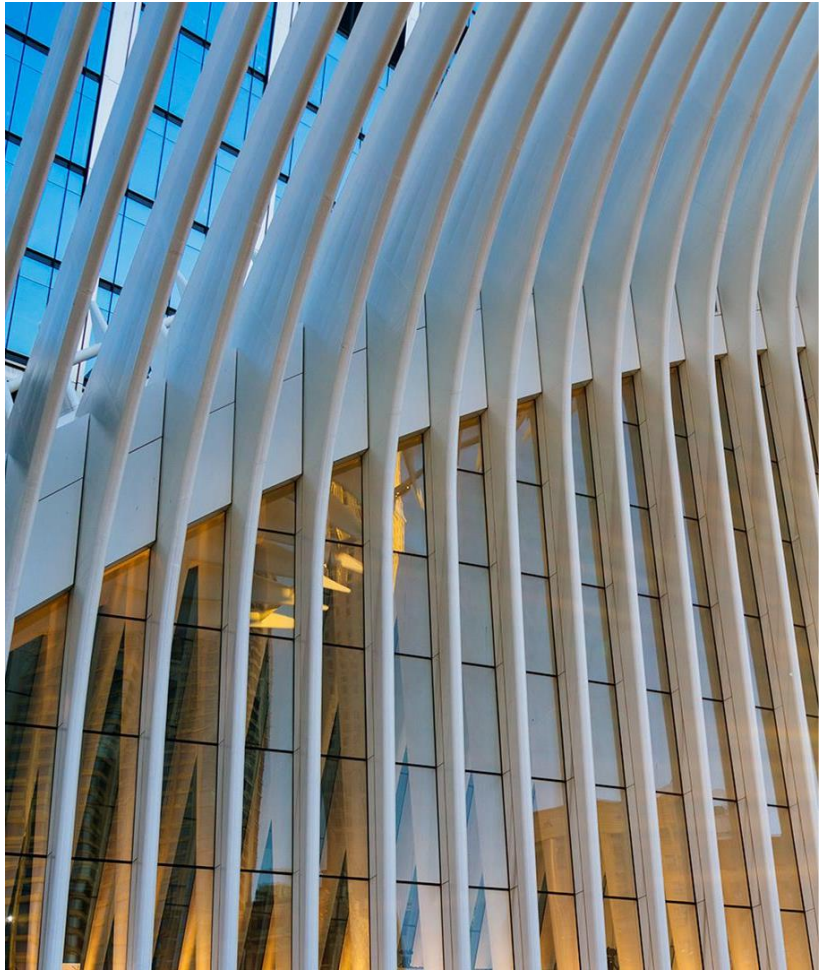


# Anticiper les besoins en consommation électrique de bâtiments

Ville de Seattle

# Sommaire





# 01
# Rappels du besoin

# Le besoin

## Entraîner des modèles

Consommation d'énergie et émissions de CO2

## Evaluer la variable

ENERGYSTARScore coûteuse à calculer

# Choix des variables cibles

| Description | Energie | $CO_2$ |
|---|---|---|
| Variable brûte | X | X |
| Variable ramenée à l'unité d'espace | X | X |
| Variable brûte normalisée par la météo | X | |
| Plusieurs variables par échelle | | X |

| | |
|---|---|
| Energie | Variable brute normalisée (SiteEnergyUseWN(kBtu)) |
| CO2 | Variable brute GHGEmissions(MetricTonsCO2e) |

# Les données

Des relevés de consommation sur deux années : 2015 et 2016

| | Nombre observations | Observations exclusives | Moyenne SiteEnergyUseWN(kBtu) | Ecart type SiteEnergyUseWN(kBtu) | Moyenne GHGEmissions(MetricTonsCO2e) | Ecart type GHGEmissions(MetricTonsCO2e) |
|---|---|---|---|---|---|---|
| 2015 | 3 340 | 56 | $5{,}178.10^{6}$ | $1{,}38.10^{7}$ | 109,08 | 403,19 |
| 2016 | 3 376 | 92 | $5.141.10^{6}$ | $1{,}39.10^{7}$ | 111,27 | 418,14 |

# 02
# Nettoyage des données

# Valeurs aberrantes

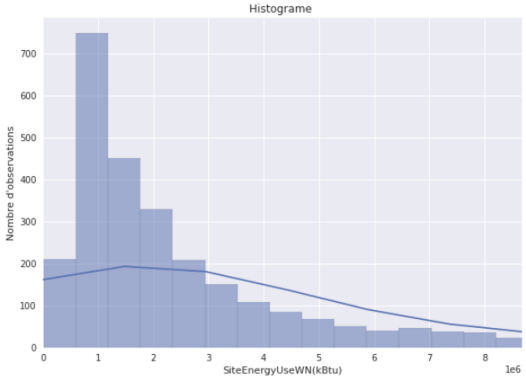
| Description | Observations restantes | Supprimées |
|---|---|---|
| Etat initial | 3318 | 0 |
| Colonne outliers | 3236 | 92 |
| Surface activité principale > surface totale | 2938 | 298 |
| Surface de parking négative | 2936 | 2 |
| Energie totale < somme des types d'énergie | 2934 | 2 |
| Nombre d'étages aberrants | 2933 | 1 |

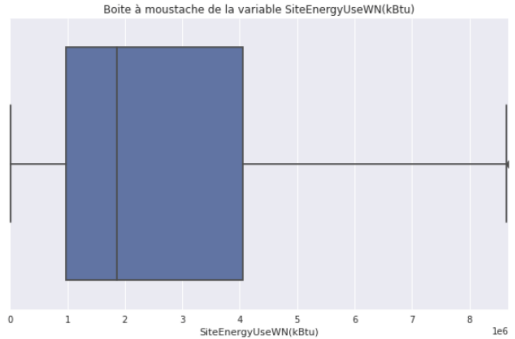
# Premier filtrage de variables

| Variable | Taux de complétion |
|---|---|
| DataYear | 100% |
| Seattle Police Departement Micro Community Policing Plan Areas | 99,94% |
| SPD Beats | 99,94% |
| 2010 Census Tracts | 6,71% |
| City Council Districts | 6,38% |
| YearsENERGYSTARCertified | 3,29% |
| Comment | 0,39% |

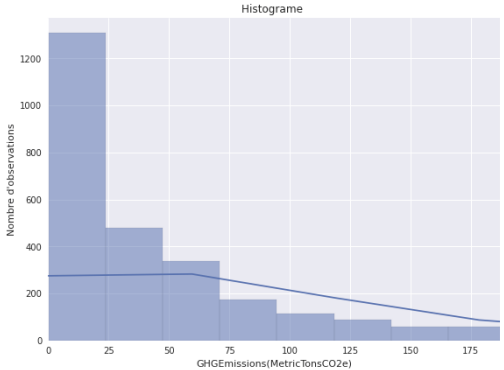
# 03
# Exploration des données

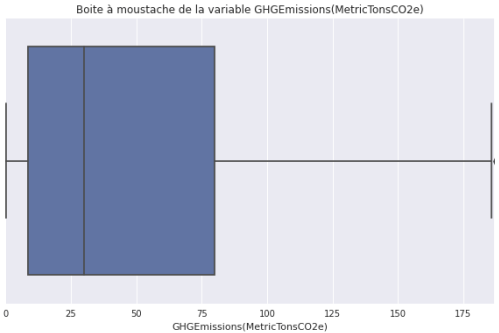
# Distribution des variables cibles



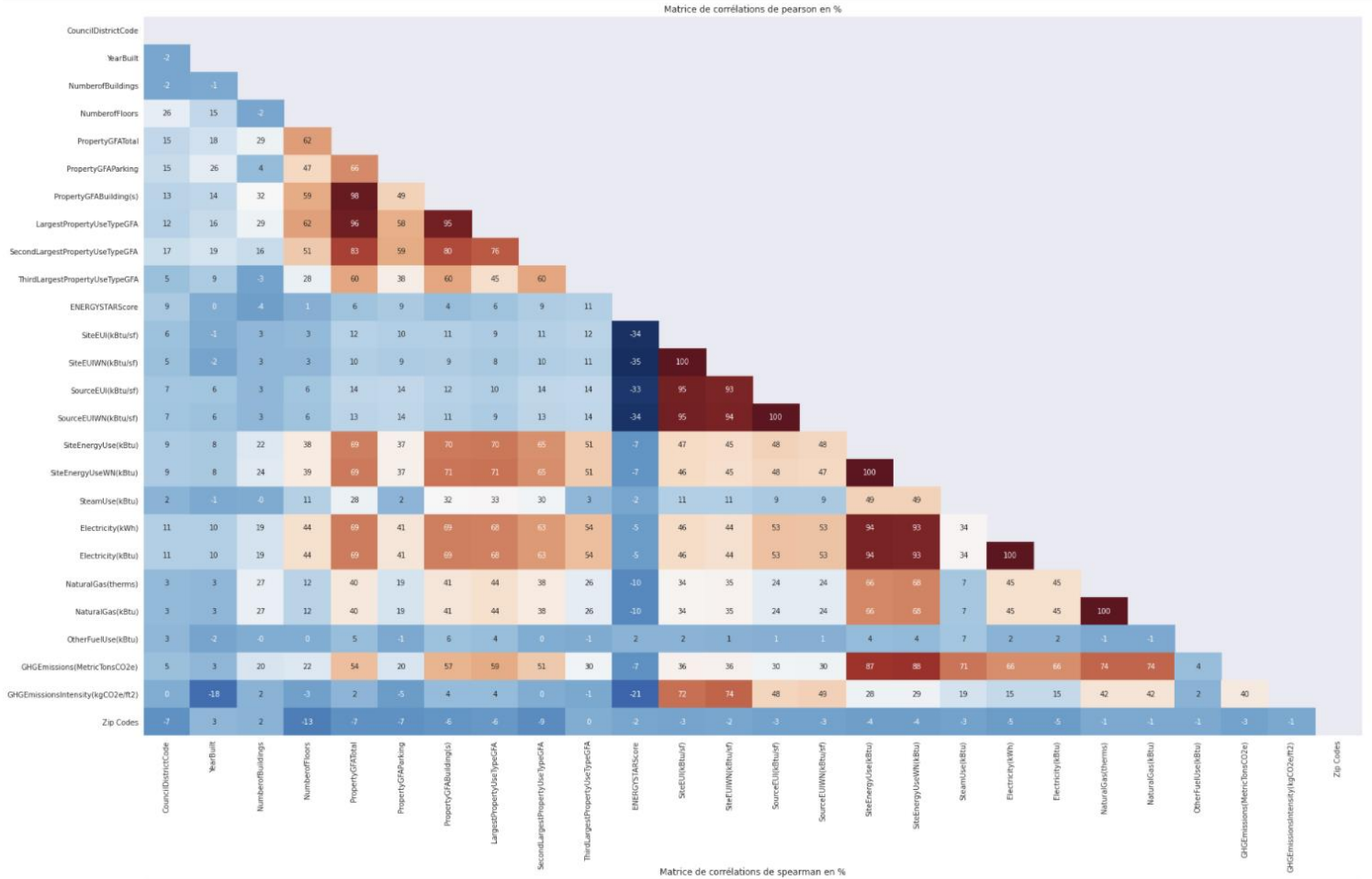

```
count    2.933000e+03
mean     4.885725e+06
std      1.310088e+07
min      1.144100e+04
25%      9.738990e+05
50%      1.868740e+06
75%      4.055974e+06
max      2.927463e+08
Name: SiteEnergyUseWN(kBtu), dtype: float64
```



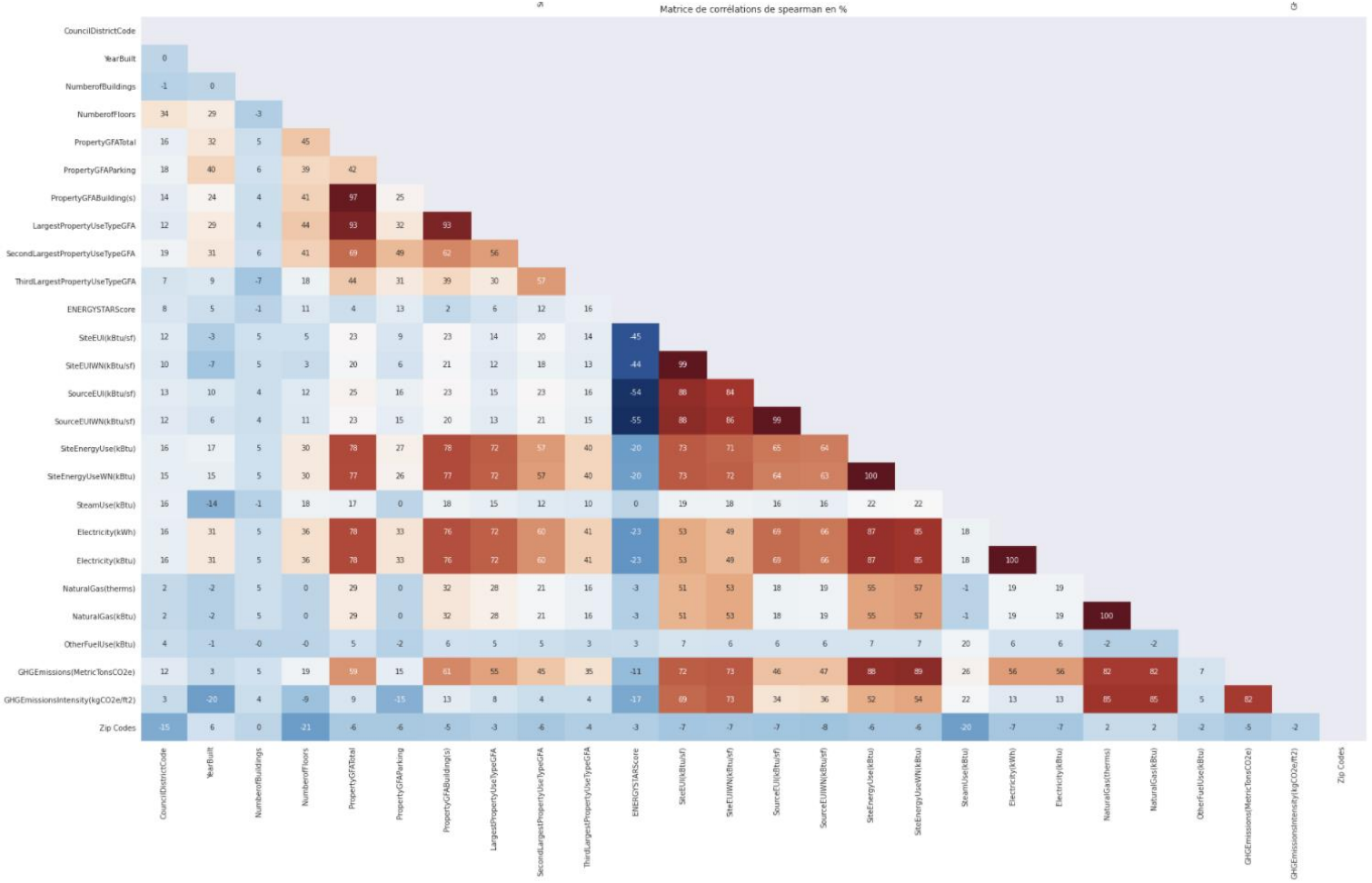




```
count     2933.000000
mean        98.579236
std        360.138301
min          0.080000
25%          8.700000
50%         30.190000
75%         79.980000
max      11824.890000
Name: GHGEmissions(MetricTonsCO2e), dtype: float64
```

# Corrélations linéaires



# Corrélations de rangs

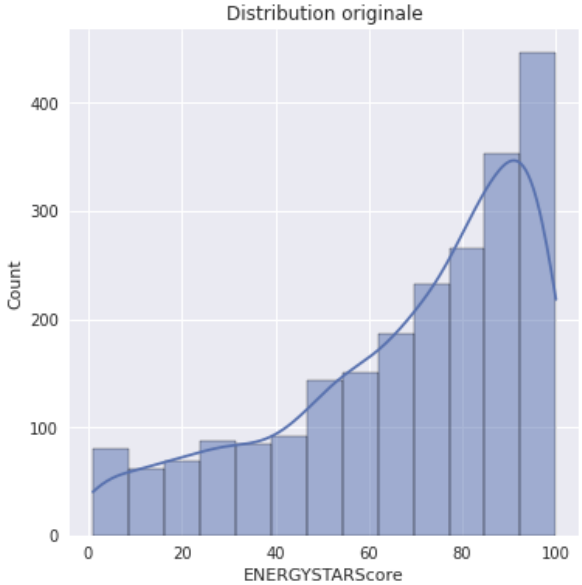


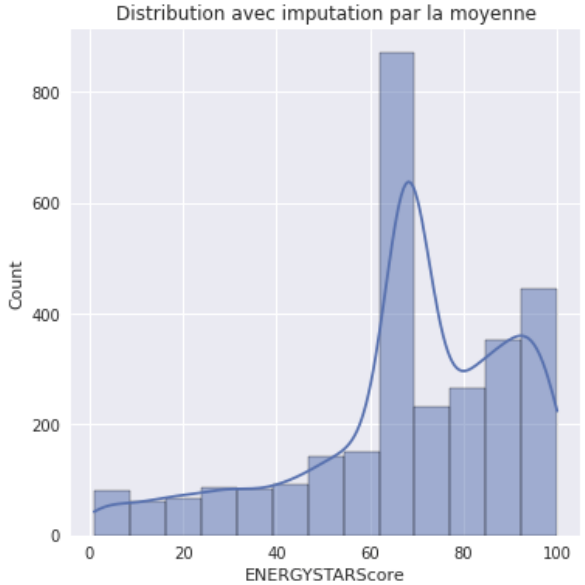
# 04
# Feature engineering

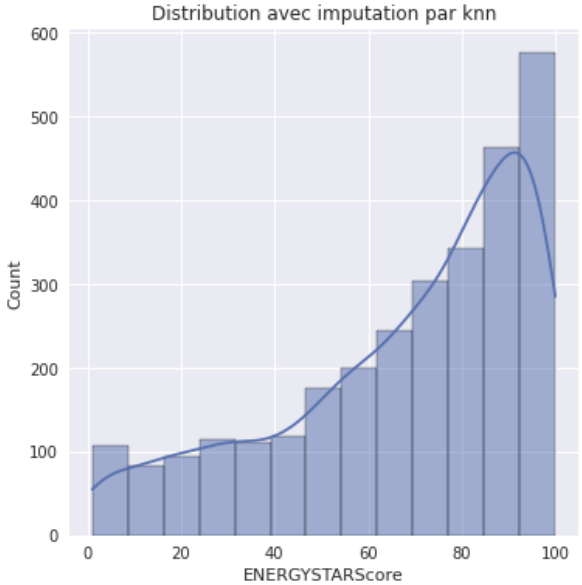
# Variables crées / transformée

| Transformation | Traitements |
|---|---|
| La liste d'activités en un nombre d'activités | Comptage de liste |
| Les trois activités principales en proportion de surface par activité significative | 1. One hot encoding<br>2. Ou logique<br>3. t-test de student<br>4. imputation de valeurs manquantes<br>5. calcul de proportions |
| Les sources d'énergie en proportion d'énergie sur l'énergie totale | Calcul de proportion |
| Le nombre d'étages en surface par étage | 1. Imputation de valeurs manquantes<br>2. Calcul de rapport |
| Le nombre de bâtiments en variable binaire (plus d'un bâtiment ou non) | Test logique |

# Imputation de l'ENERGYSTARScore



# Variables / types de variables retenues

| Type de variable | Variable |
|---|---|
| Année de construction | YearBuilt |
| Surface totale | PropertyGFATotal |
| Surface par étage | GFA_by_floor |
| Plus d'un bâtiment: oui / non | more_one_building |
| Nombre d'activités | PropertyUseTypes |
| Proportion de surface pour des activités significatives | Multifamily Housing_prop, Non-refrigerated Warehouse_prop, ... (15 variables) |
| Proportion de la source d'énergie sur la consommation totale | Electricity(kBtu)_prop, NaturalGas(kBtu)_prop, SteamUse(kBtu)_prop |
| Efficacité énergétique au pied carré | ENERGYSTARScore |
| Variables à prédire | SiteEnergyUseWN(kBtu), GHGEmissions(MetricTonsCO2e) |

# 05
# Préprocessing

# Stratificaion et transformations

| Transformation | Statut |
|---|---|
| partitionnement avec un jeu d'entraînement à 80% | Retenu |
| Stratification lors du partitionnement | Retenu sur une discrétisation de la surface totale (6 classes) |
| RobustScaler (équivalent centrage réduction avec médiane et écart interquartiles) | Retenu sur tous les inputs (hormis more_one_building) et variables à prédire |
| MinMaxScaler | Retenu sur tous les input (hormis more_one_building) et variables à prédire |
| Transformation log | Rejeté sur les inputs à distribution log normal et variables à prédire |
| Transformatoion BoxCox | Rejeté sur les inputs à distribution log normal et variables à prédire |
| Transformation Yeo-Johson | Rejeté sur les inputs à distribution log normal et variables à prédire |

# 06
# Entrainement des modèles

# Modèles prédiction énergie:  sans / avec ENERGYSTARScore

| Type de modèle | R2 validation croisée sans | R2 validation croisée avec |
|---|---|---|
| Prédiction par la moyenne | -0.003 | - |
| Régression linéaire non pénalisée | 57.0 | 57.9 |
| Régression linéaire Ridge | 61.2 | 62.1 |
| Régression linéaire Lasso | 57.7 | 59.1 |
| SVM Linéaire | 61.1 | 61.9 |
| SVM à noyeau | - 57.7 | -59.6 |
| Random Forest | 74.6 | 75.0 |
| Gradient Boosting | 80.4 | 77.8 |

# Modèles prédiction $CO_2$ : sans / avec ENERGYSTARScore

| Type de modèle | R2 validation croisée sans | R2 validation croisée avec |
|---|---|---|
| Prédiction par la moyenne | -0.005 | - |
| Régression linéaire non pénalisée | 57.0 | 57.9 |
| Régression linéaire Ridge | 42.2 | 43.3 |
| Régression linéaire Lasso | 37.6 | 38.3 |
| SVM Linéaire | 41.8 | 58.0 |
| SVM à noyeau | - 1.39 | -0.65 |
| Random Forest | 71.4 | 74.4 |
| Gradient Boosting | 54.7 | 77.1 |

# Retour des régressions Linéaires aves StatsModels

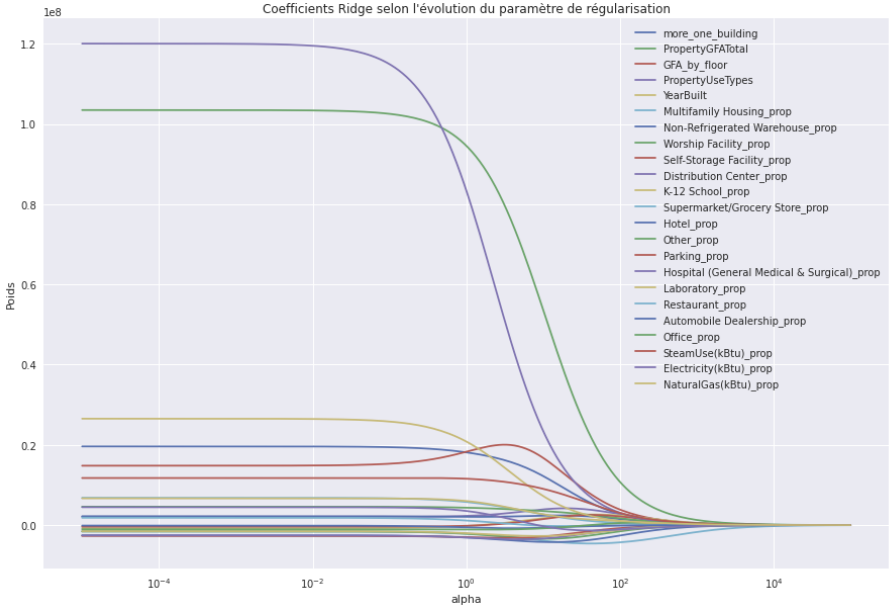
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -2.726e+05 | 2.51e+06 | -0.109 | 0.913 | -5.19e+06 | 4.64e+06 |
| more_one_building | 1.932e+07 | 2.36e+06 | 8.188 | 0.000 | 1.47e+07 | 2.39e+07 |
| PropertyGFATotal | 1.042e+08 | 2.3e+06 | 45.254 | 0.000 | 9.97e+07 | 1.09e+08 |
| GFA_by_floor | 1.425e+07 | 3.57e+06 | 3.987 | 0.000 | 7.24e+06 | 2.13e+07 |
| PropertyUseTypes | 2.025e+06 | 1.12e+06 | 1.802 | 0.072 | -1.79e+05 | 4.23e+06 |
| YearBuilt | -8.783e+05 | 5.03e+05 | -1.746 | 0.081 | -1.86e+06 | 1.08e+05 |
| ENERGYSTARScore | -3.276e+06 | 4.68e+05 | -7.005 | 0.000 | -4.19e+06 | -2.36e+06 |
| Multifamily Housing_prop | -1.253e+06 | 4.16e+05 | -3.009 | 0.003 | -2.07e+06 | -4.36e+05 |
| Non-Refrigerated Warehouse_prop | -2.587e+06 | 6.76e+05 | -3.829 | 0.000 | -3.91e+06 | -1.26e+06 |
| Worship Facility_prop | -1.545e+06 | 1.02e+06 | -1.521 | 0.128 | -3.54e+06 | 4.47e+05 |
| Self-Storage Facility_prop | -2.094e+06 | 1.58e+06 | -1.328 | 0.184 | -5.19e+06 | 9.98e+05 |
| Distribution Center_prop | -2.66e+06 | 1.15e+06 | -2.313 | 0.021 | -4.92e+06 | -4.05e+05 |
| K-12 School_prop | -9.032e+05 | 8.12e+05 | -1.112 | 0.266 | -2.5e+06 | 6.89e+05 |
| Supermarket/Grocery Store_prop | 6.16e+06 | 1.38e+06 | 4.471 | 0.000 | 3.46e+06 | 8.86e+06 |
| Hotel_prop | 2.091e+06 | 1.13e+06 | 1.845 | 0.065 | -1.32e+05 | 4.31e+06 |
| Other_prop | 4.517e+06 | 8.93e+05 | 5.057 | 0.000 | 2.77e+06 | 6.27e+06 |
| Parking_prop | 3.157e+05 | 9.85e+05 | 0.321 | 0.749 | -1.62e+06 | 2.25e+06 |
| Hospital (General Medical & Surgical)_prop | 1.191e+08 | 4.11e+06 | 28.993 | 0.000 | 1.11e+08 | 1.27e+08 |
| Laboratory_prop | 2.508e+07 | 3.06e+06 | 8.201 | 0.000 | 1.91e+07 | 3.11e+07 |
| Restaurant_prop | 1.402e+06 | 2.06e+06 | 0.680 | 0.497 | -2.64e+06 | 5.45e+06 |
| Automobile Dealership_prop | -1.94e+04 | 3.37e+06 | -0.006 | 0.995 | -6.63e+06 | 6.59e+06 |
| Office_prop | -7.971e+05 | 5.63e+05 | -1.416 | 0.157 | -1.9e+06 | 3.06e+05 |
| SteamUse(kBtu)_prop | 1.175e+07 | 2.16e+06 | 5.436 | 0.000 | 7.51e+06 | 1.6e+07 |
| Electricity(kBtu)_prop | 4.315e+06 | 2.86e+06 | 1.510 | 0.131 | -1.29e+06 | 9.92e+06 |
| NaturalGas(kBtu)_prop | 6.695e+06 | 2.9e+06 | 2.312 | 0.021 | 1.02e+06 | 1.24e+07 |

Prédiction énergie

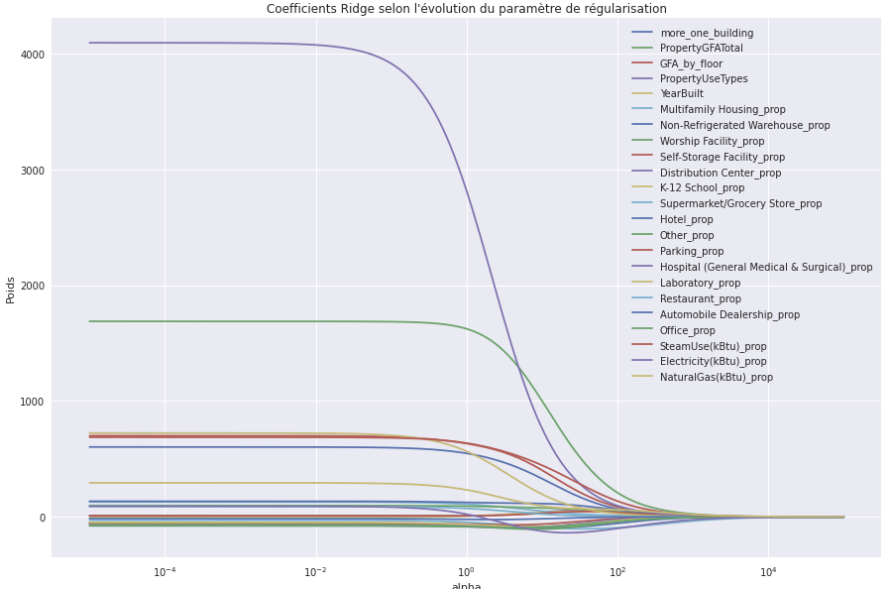
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -87.9800 | 80.798 | -1.089 | 0.276 | -246.424 | 70.464 |
| more_one_building | 603.3527 | 76.654 | 7.871 | 0.000 | 453.035 | 753.670 |
| PropertyGFATotal | 1688.6171 | 74.722 | 22.599 | 0.000 | 1542.088 | 1835.146 |
| GFA_by_floor | 700.5179 | 116.081 | 6.035 | 0.000 | 472.884 | 928.152 |
| PropertyUseTypes | 9.3566 | 36.526 | 0.256 | 0.798 | -62.270 | 80.983 |
| YearBuilt | -4.8087 | 16.339 | -0.294 | 0.769 | -36.850 | 27.233 |
| Multifamily Housing_prop | -29.0812 | 13.456 | -2.161 | 0.031 | -55.468 | -2.694 |
| Non-Refrigerated Warehouse_prop | -55.9884 | 21.954 | -2.550 | 0.011 | -99.040 | -12.937 |
| Worship Facility_prop | -63.2694 | 32.998 | -1.917 | 0.055 | -127.978 | 1.439 |
| Self-Storage Facility_prop | -52.5800 | 51.163 | -1.028 | 0.304 | -152.911 | 47.751 |
| Distribution Center_prop | -76.7057 | 37.372 | -2.052 | 0.040 | -149.991 | -3.420 |
| K-12 School_prop | -48.5121 | 26.237 | -1.849 | 0.065 | -99.962 | 2.938 |
| Supermarket/Grocery Store_prop | 128.6487 | 44.667 | 2.880 | 0.004 | 41.057 | 216.241 |
| Hotel_prop | 134.7024 | 36.835 | 3.657 | 0.000 | 62.470 | 206.935 |
| Other_prop | 97.8124 | 29.021 | 3.370 | 0.001 | 40.902 | 154.723 |
| Parking_prop | 11.1006 | 31.840 | 0.349 | 0.727 | -51.336 | 73.538 |
| Hospital (General Medical & Surgical)_prop | 4094.7720 | 133.377 | 30.701 | 0.000 | 3833.221 | 4356.323 |
| Laboratory_prop | 724.0854 | 99.139 | 7.304 | 0.000 | 529.675 | 918.496 |
| Restaurant_prop | 89.2655 | 67.033 | 1.332 | 0.183 | -42.186 | 220.717 |
| Automobile Dealership_prop | -14.5382 | 109.469 | -0.133 | 0.894 | -229.205 | 200.129 |
| Office_prop | -77.5175 | 18.251 | -4.247 | 0.000 | -113.307 | -41.728 |
| SteamUse(kBtu)_prop | 687.5713 | 70.246 | 9.788 | 0.000 | 549.820 | 825.323 |
| Electricity(kBtu)_prop | 91.7899 | 92.837 | 0.989 | 0.323 | -90.261 | 273.841 |
| NaturalGas(kBtu)_prop | 293.8760 | 94.097 | 3.123 | 0.002 | 109.354 | 478.398 |

Prédiction CO2

# Evolution des coefficients Ridge

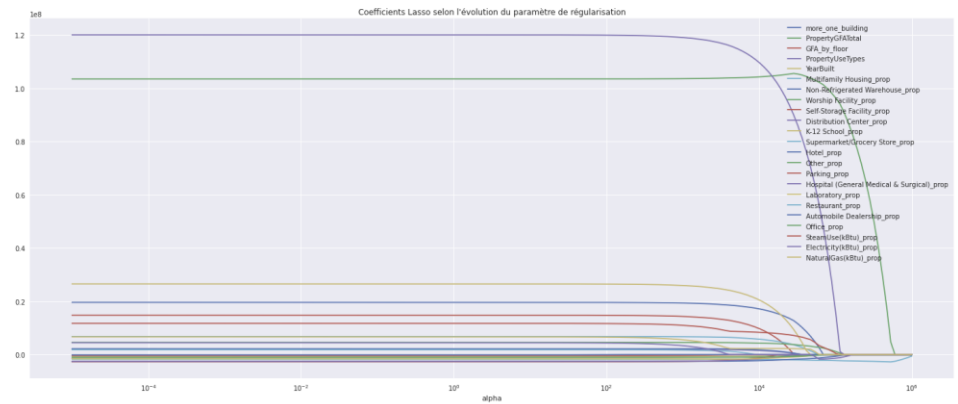



Prédiction énergie

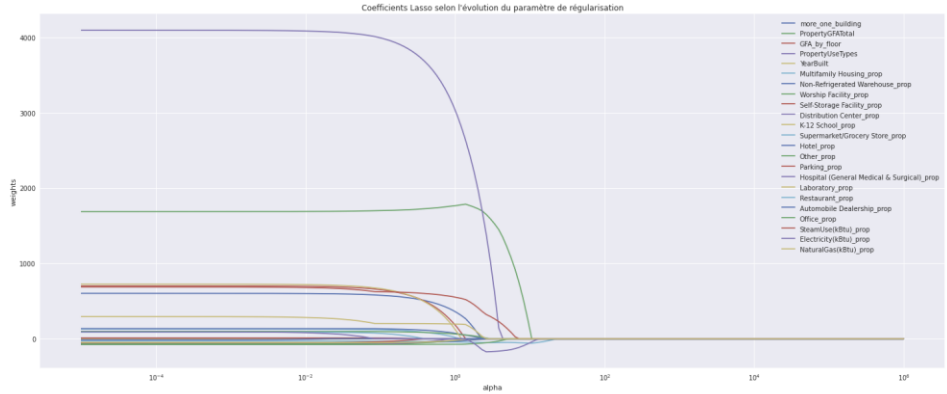



Prédiction CO2

# Evolution des coefficients Lasso




Prédiction énergie




Prédiction CO2

# Energie : feature Importance de la random forest

| Variable | Score |
|---|---|
| PropertyGFATotal | 0.501945 |
| Hospital (General Medical & Surgical)_prop | 0.236309 |
| Electricity(kBtu)_prop | 0.075688 |
| Other_prop | 0.031854 |
| NaturalGas(kBtu)_prop | 0.023104 |
| SteamUse(kBtu)_prop | 0.020881 |
| PropertyUseTypes | 0.020833 |
| YearBuilt | 0.017230 |
| GFA_by_floor | 0.015079 |
| Laboratory_prop | 0.014357 |
| more_one_building | 0.011589 |

| Variable | Score |
|---|---|
| Multifamilly Housing_prop | 0.010244 |
| Parking_prop | 0.008262 |
| Supermarket/GroceryStrore_prop | 0.003706 |
| Office_prop | 0.003239 |
| Restaurant_prop | 0.001953 |
| Hotel_prop | 0.001525 |
| Non-Refrigerated Warehouse_prop | 0.001439 |
| Distribution Center_prop | 0.000497 |
| Self-Storage Facility_prop | 0.000098 |
| K-12 School_prop | 0.000095 |
| Worship Facility_prop | 0.000095 |
| Automobile Dealership_prop | 0.000014 |

# Co2 : feature Importance de la random forrest

| Variable | Score |
|---|---|
| PropertyGFATotal | 4.255723e-01 |
| Electricity(kBtu)_prop | 2.892538e-01 |
| Hospital (General Medical & Surgical)_prop | 1.479773e-01 |
| SteamUse(kBtu)_prop | 4.710764e-02 |
| NaturalGas(kBtu)_prop | 2.573871e-02 |
| GFA_by_floor | 1.919814e-02 |
| YearBuilt | 8.960775e-03 |
| Other_prop | 7.437989e-03 |
| Multifamily Housing_prop | 4.635946e-03 |
| PropertyUseTypes | 4.635946e-03 |
| Parking_prop | 4.302727e-03 |

| Variable | Score |
|---|---|
| Laboratory_prop | 3.238330e-03 |
| Office_prop | 2.496743e-03 |
| Supermarket/GroceryStrore_prop | 2.257894e-03 |
| Restaurant_prop | 2.042214e-03 |
| Hotel_prop | 9.422415e-04 |
| more_one_building | 4.353137e-04 |
| Non-Refrigerated Warehouse_prop | 1.113923e-04 |
| K-12 School_prop | 5.610112e-05 |
| Worship Facility_prop | 4.091153e-06 |
| Distribution Center_pro | 2.590996e-06 |
| Self-Storage Facility_prop | 4.889443e-07 |
| Automobile Dealership_prop | 0.000000e+00 |

# Prédiction énergie: généralisation des modèles

| Type de modèle | R2 validation croisée jeu d'entrainement | R2 jeu de test | Temps d'entrainement |
|---|---|---|---|
| Prédiction par la moyenne | -0.003 | - | - |
| Régression linéaire non pénalisée | 57.0 | 54.2 | 1.34s |
| Régression linéaire Ridge | 61.2 | 47.9 | 1.24s |
| Régression linéaire Lasso | 57.7 | 41.2 | 1.42s |
| SVM Linéaire | 61.1 | 39.0 | 1.48s |
| SVM à noyeau | - 57.7 | - | - |
| Random Forest | 74.6 | 50.2 | 9.14s |
| Gradient Boosting | 80.4 | 57.2 | 821 ms |

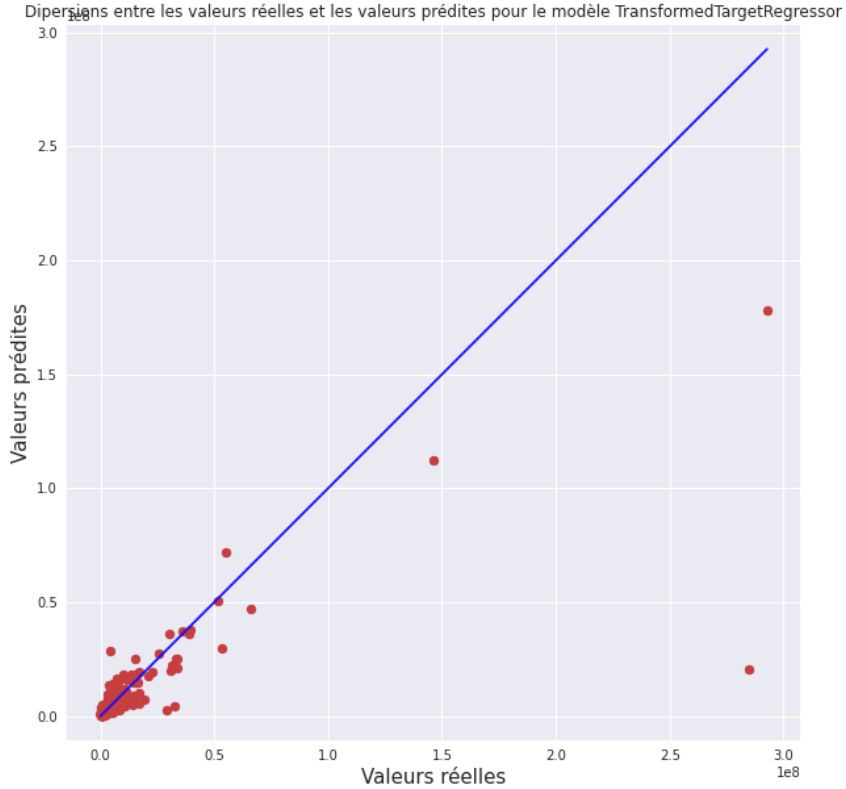
# Prédiction CO2: généralisation des modèles

| Type de modèle | R2 validation croisée jeu d'entraînement | R2 jeu de test | Temps d'entraînement |
|---|---|---|---|
| Prédiction par la moyenne | -0.003 | - | - |
| Régression linéaire non pénalisée | 57.0 | 66.7 | 1.29s |
| Régression linéaire Ridge | 42.2 | 44.6 | 1.27s |
| Régression linéaire Lasso | 37.6 | 37.6 | 1.5s |
| SVM Linéaire | 41.8 | 37.6 | 991ms |
| SVM à noyeau | - 1.39 | - | - |
| Random Forest | 71.4 | 47.4 | 7,91s |
| Gradient Boosting | 54.7 | 78.0 | 3,9s |

# Modèle retenu pour la prédiction d'énergie

Gradient Boosting

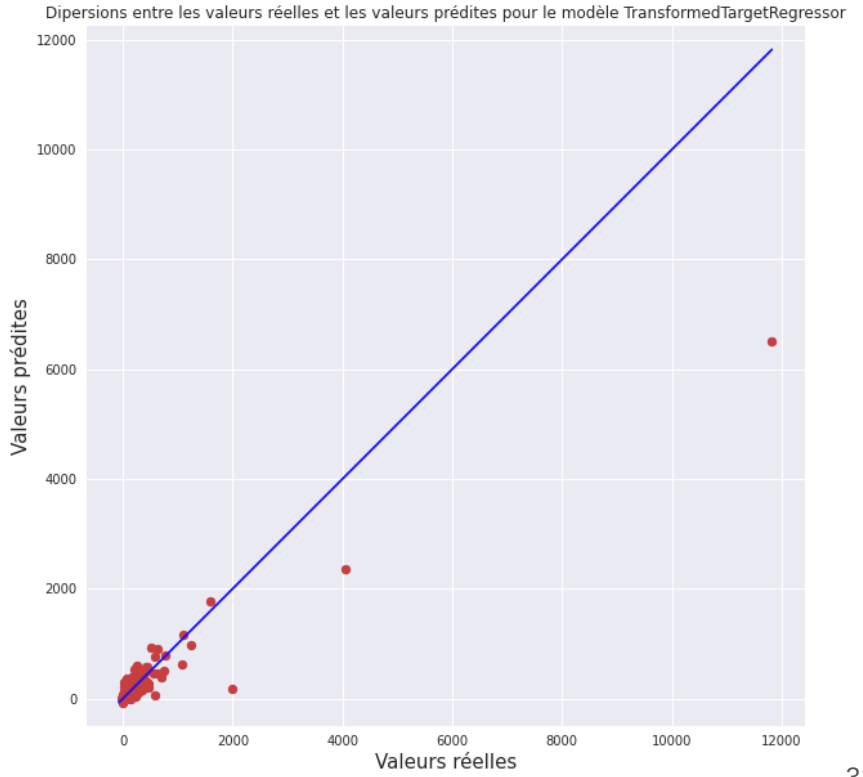
| Hyperparamètre | Valeur |
|---|---|
| learning rate | 0.1 |
| max depth | 4 |
| n_estimators | 200 |



# Modèle retenu pour la prédiction de $CO_2$

## Gradient Boosting

| Hyperparamètre | Valeur |
|---|---|
| learning rate | 0.2 |
| max depth | 3 |
| n_estimators | 1400 |



# Résultats 2015 vs 2016

| Variable à prédire | R2 jeu de test 2015 | R2 jeu de test 2016 |
|---|---|---|
| SiteEnergyUseWN(kBtu) | 57.2 | 77.9 |
| GHGEmissions(MetricTonsCO2e) | 78.0 | 85.5 |

# Merci!

Avez-vous des questions?

cedricsoares@me.com
06 09 25 47 45