# Projet 4 : Anticipez les besoins en consommation de bâtiments

Laurent Cagniart

# Problématique (1/2)

- Vous travaillez pour la ville de Seattle. Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, votre équipe s'intéresse de près à la consommation et aux émissions des **bâtiments non destinés à l'habitation**.
- Les **relevés sont coûteux à obtenir**, et vous voulez tenter de **prédire les émissions de CO2 et la consommation totale d'énergie** de bâtiments pour lesquels elles n'ont pas encore été mesurées.

# Problématique (2/2)

- Vous cherchez également à évaluer l'intérêt de l' « ENERGY STAR Score » , indicateur de performance énergétique, qui est fastidieux à calculer. Vous l'intégrerez dans la modélisation et jugerez de son intérêt.
- Voici un récapitulatif de votre mission :
  - ➢ Réaliser une courte analyse exploratoire.
  - ➢ Tester différents modèles de prédiction afin de répondre au mieux à la problématique.

# Jeu de données (1/2)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3376 entries, 0 to 3375
Data columns (total 46 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   OSEBuildingID                     3376 non-null   int64
 1   DataYear                          3376 non-null   int64
 2   BuildingType                      3376 non-null   object
 3   PrimaryPropertyType               3376 non-null   object
 4   PropertyName                      3376 non-null   object
 5   Address                           3376 non-null   object
 6   City                              3376 non-null   object
 7   State                             3376 non-null   object
 8   ZipCode                           3360 non-null   float64
 9   TaxParcelIdentificationNumber     3376 non-null   object
 10  CouncilDistrictCode               3376 non-null   int64
 11  Neighborhood                      3376 non-null   object
 12  Latitude                          3376 non-null   float64
 13  Longitude                         3376 non-null   float64
 14  YearBuilt                         3376 non-null   int64
 15  NumberofBuildings                 3368 non-null   float64
 16  NumberofFloors                    3376 non-null   int64
 17  PropertyGFATotal                  3376 non-null   int64
 18  PropertyGFAParking                3376 non-null   int64
 19  PropertyGFABuilding(s)            3376 non-null   int64
 20  ListOfAllPropertyUseTypes         3367 non-null   object
 21  LargestPropertyUseType            3356 non-null   object
 22  LargestPropertyUseTypeGFA         3356 non-null   float64
 23  SecondLargestPropertyUseType      1679 non-null   object
 24  SecondLargestPropertyUseTypeGFA   1679 non-null   float64
 25  ThirdLargestPropertyUseType       596 non-null    object
 26  ThirdLargestPropertyUseTypeGFA    596 non-null    float64
 27  YearsENERGYSTARCertified          119 non-null    object
 28  ENERGYSTARScore                   2533 non-null   float64
 29  SiteEUI(kBtu/sf)                  3369 non-null   float64
 30  SiteEUIWN(kBtu/sf)                3370 non-null   float64
 31  SourceEUI(kBtu/sf)                3367 non-null   float64
 32  SourceEUIWN(kBtu/sf)              3367 non-null   float64
 33  SiteEnergyUse(kBtu)               3371 non-null   float64
 34  SiteEnergyUseWN(kBtu)             3370 non-null   float64
 35  SteamUse(kBtu)                    3367 non-null   float64
 36  Electricity(kWh)                  3367 non-null   float64
 37  Electricity(kBtu)                 3367 non-null   float64
 38  NaturalGas(therms)                3367 non-null   float64
 39  NaturalGas(kBtu)                  3367 non-null   float64
 40  DefaultData                       3376 non-null   bool
 41  Comments                          0 non-null      float64
 42  ComplianceStatus                  3376 non-null   object
 43  Outlier                           32 non-null     object
 44  TotalGHGEmissions                 3367 non-null   float64
 45  GHGEmissionsIntensity             3367 non-null   float64
dtypes: bool(1), float64(22), int64(8), object(15)
memory usage: 1.2+ MB
```

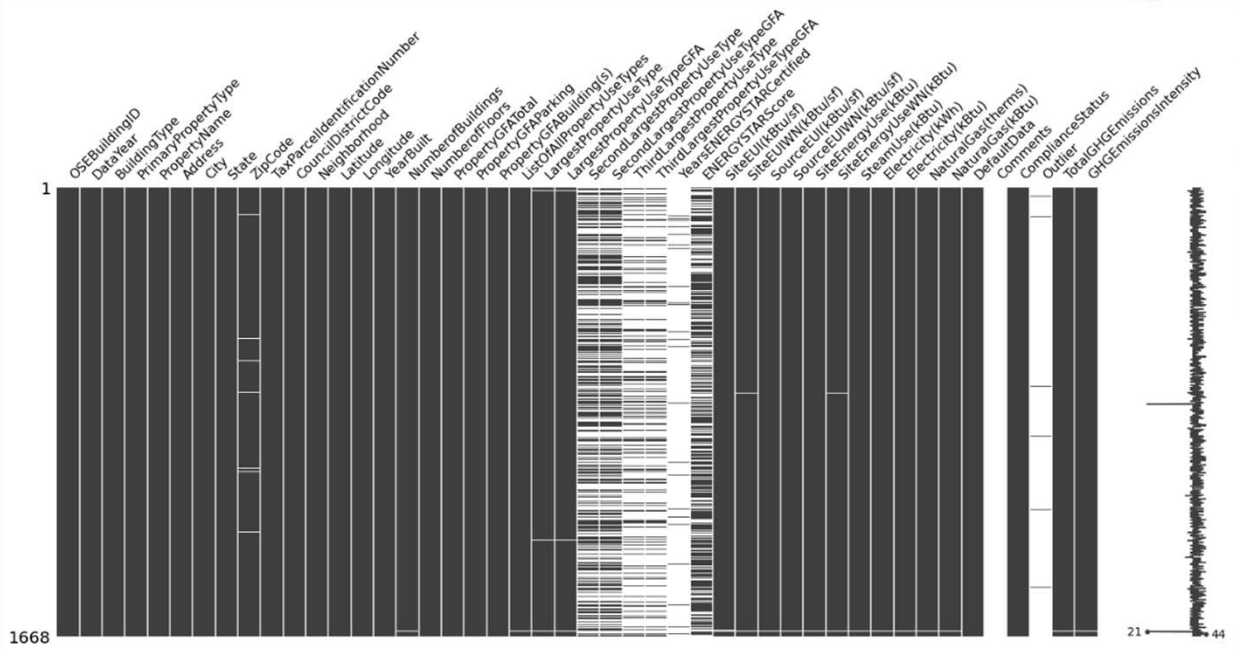
```
BuildingType
Campus                    24
Multifamily HR (10+)     110
Multifamily LR (1-4)    1018
Multifamily MR (5-9)     580
NonResidential          1460
Nonresidential COS        85
Nonresidential WA          1
SPS-District K-12         98
Name: OSEBuildingID, dtype: int64
```

- Conformément aux critères de la mission, nous excluons tous les bâtiments de type « Multifamily »

3376 → -1708 → 1668

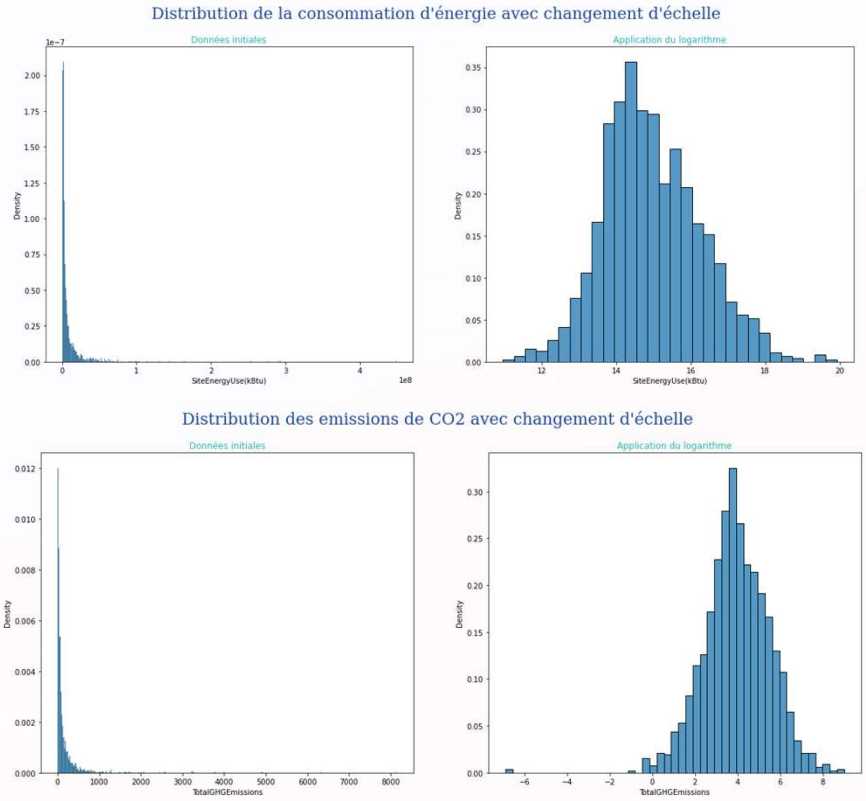
- 46 variables/features

# Jeu de données (2/2)




- Les taux de remplissage sont élevés, à l'exception des second et third property use (type et GFA) et deux features supprimables (comments et outlier dont on supprimera également les observations)

# Feature engineering (1/3)

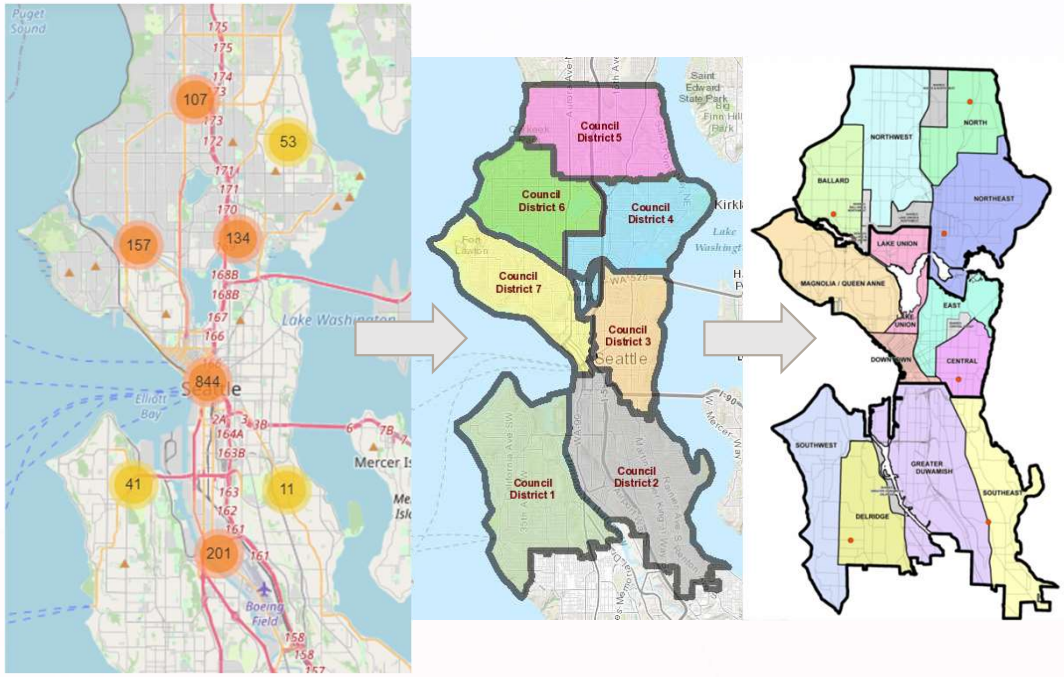



Nos features à prédire sont :

- la consommation d'énergie, à savoir « SiteEnergyUse(kBtu) »
- Et, les émissions de gaz à effet de serre (en CO2e), « TotalGHGEmissions »

On remarque sur les graphiques suivants que le passage au logarithme nous permet d'obtenir une distribution de type gaussienne

# Feature engineering (2/3)

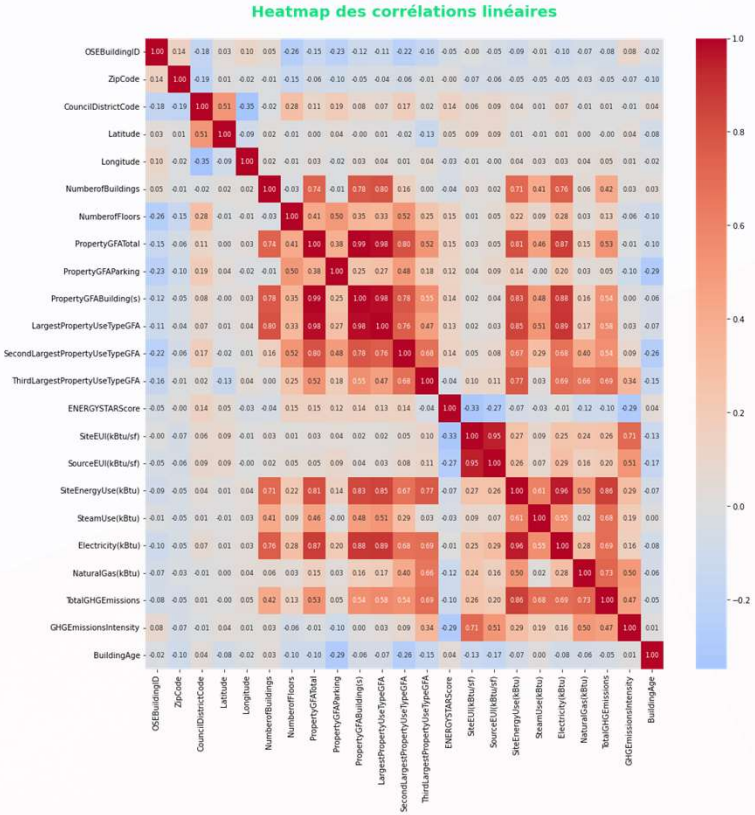



Pour les features liées à l'identification et la localisation géographique, on retiendra :

- « OSEBuildingID »
- « Neighborhood » (bon compromis au niveau des valeurs possibles)

Au niveau des features 'DataYear' et 'YearBuilt', nous les combinons afin d'obtenir la feature 'building age'

# Feature engineering (3/3)




Pour le bon fonctionnement de nos modèles, on étudie les corrélations linéaires entre les variables numériques

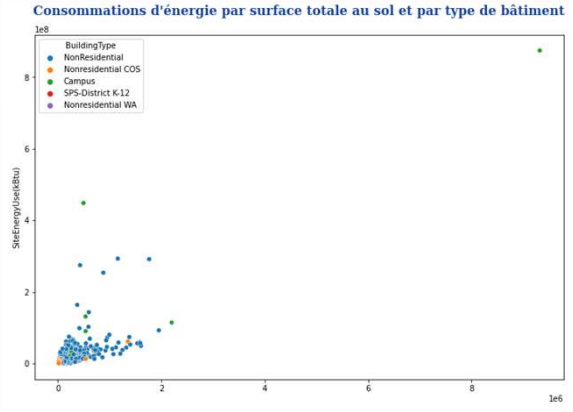
- Suppressions des données liées aux superficies (/sf, GFA…)
- Plutôt que la superficie totale, on conserve les superficies ‘building’ et ‘parking’
- Le but est de supprimer les relevés coûteux pour les années à venir => exclusion de toutes les données de relève

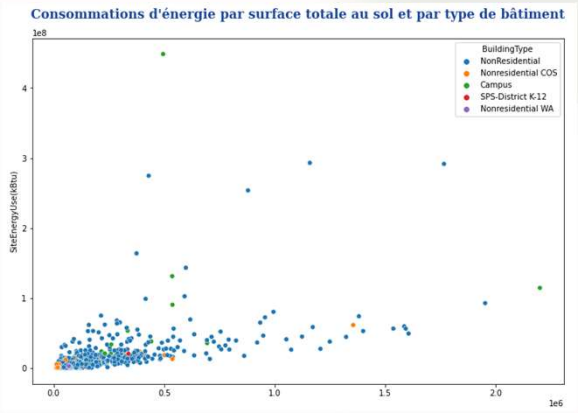
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1548 entries, 0 to 3375
Data columns (total 13 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   OSEBuildingID           1548 non-null   int64
 1   BuildingType            1548 non-null   object
 2   PrimaryPropertyType     1548 non-null   object
 3   Neighborhood            1548 non-null   object
 4   NumberofBuildings       1548 non-null   float64
 5   NumberofFloors          1548 non-null   int64
 6   PropertyGFAParking      1548 non-null   int64
 7   PropertyGFABuilding(s)  1548 non-null   int64
 8   LargestPropertyUseType  1548 non-null   object
 9   ENERGYSTARScore         997 non-null    float64
 10  SiteEnergyUse(kBtu)     1548 non-null   float64
 11  TotalGHGEmissions       1548 non-null   float64
 12  BuildingAge             1548 non-null   int64
dtypes: float64(4), int64(5), object(4)
memory usage: 201.6+ KB
```

# Modélisation(1/15)

| | OSEBuildingID | NumberofBuildings | NumberofFloors | PropertyGFAParking | PropertyGFABuilding(s) | ENERGYSTARScore | SiteEnergyUse(kBtu) | TotalGHGEmissions | BuildingAge |
|---|---|---|---|---|---|---|---|---|---|
| count | 1548.000000 | 1548.000000 | 1548.000000 | 1548.000000 | 1.548000e+03 | 997.000000 | 1.548000e+03 | 1548.000000 | 1548.000000 |
| mean | 16497.944444 | 1.212532 | 4.286822 | 13842.337209 | 1.074898e+05 | 63.635908 | 8.860058e+06 | 193.609426 | 54.355943 |
| std | 13827.877766 | 3.031517 | 6.774923 | 43721.822291 | 2.926272e+05 | 28.825309 | 3.130568e+07 | 779.105149 | 32.886918 |
| min | 1.000000 | 1.000000 | 1.000000 | 0.000000 | 3.636000e+03 | 1.000000 | 5.713320e+04 | 0.001000 | 1.000000 |
| 25% | 602.750000 | 1.000000 | 1.000000 | 0.000000 | 2.793675e+04 | 44.000000 | 1.251083e+06 | 20.655000 | 27.000000 |
| 50% | 21180.500000 | 1.000000 | 2.000000 | 0.000000 | 4.608400e+04 | 71.000000 | 2.732167e+06 | 49.845000 | 50.500000 |
| 75% | 24609.000000 | 1.000000 | 4.000000 | 0.000000 | 9.556825e+04 | 88.000000 | 7.294487e+06 | 147.227500 | 86.000000 |
| max | 50226.000000 | 111.000000 | 99.000000 | 512608.000000 | 9.320156e+06 | 100.000000 | 8.739237e+08 | 16870.980000 | 116.000000 |

Il semble que nous ayons des outliers à retraiter pour nos variables à prédire



# Modélisation(2/15)

1-1) Encodage et standardisation :

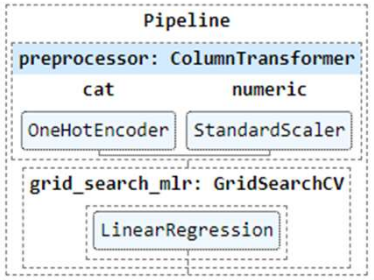
```
categorical_features.nunique()
```

```
BuildingType              5
PrimaryPropertyType      21
Neighborhood             13
LargestPropertyUseType   55
dtype: int64
```

- En plus des valeurs numériques vues précédemment, voici la liste des features catégorielles qu'il faudra encoder

- Nous utiliserons OneHotEncoder pour les features catégorielles et StandardScaler pour les features numériques

Ex :



# Modélisation(3/15)

1-2) Préparation des jeux d'entrainement et de test:

Jeu de test (20%)

```
Entrainement: 1237 lignes,
Test: 310 lignes.
```

2-1) Modèle Baseline: DummyRegressor (moyenne)

| | Métrique | Baseline |
|---|---|---|
| 0 | MAE | 9.452627e+06 |
| 1 | MSE | 8.366226e+14 |
| 2 | RMSE | 2.892443e+07 |
| 3 | $R^2$ | -8.019255e-04 |

2-2) Modèle Baseline: Régression linéaire multivariée

```
Meilleur score MAE : -11263938.077
Meilleur Score R2 : -43.582
Meilleurs paramètres : {'regressor__fit_intercept': False, 'regressor__normalize': True}
Temps moyen d'entrainement : 35.59s
```

# Modélisation(4/15)

3) Modèles linéaires (avec GridSearch/Validation croisée)

3-1) Modèle : ElasticNet (Combinaison des 2 régularisations Ridge et Lasso)

```
Meilleur score MAE : -5948750.728
Meilleur Score R2 : 0.14
Meilleurs paramètres : {'regressor__alpha': 1.0, 'regressor__l1_ratio': 0.1, 'regressor__max_iter': 10}
Temps moyen d'entrainement : 66.72s
```

3-2) Modèle : Support Vector Regression (SVR)

```
Meilleur score MAE : -6531295.329
Meilleur Score R2 : 0.007
Meilleurs paramètres : {'regressor__C': 0.01, 'regressor__epsilon': 0, 'regressor__loss': 'epsilon_insensitive', 'regressor__ma
x_iter': 1000}
Temps moyen d'entrainement : 7.19s
```

# Modélisation(5/15)

4) Modèles non-linéaires (avec GridSearch/Validation croisée)

4-1) Modèle RandomForestRegressor

```
Meilleur score MAE : -3653371.261
Meilleur Score R2 : 0.658
Meilleurs paramètres : {'regressor__bootstrap': False, 'regressor__max_depth': 25, 'regressor__max_features': 'sqrt', 'regresso
r__min_samples_leaf': 1, 'regressor__min_samples_split': 2}
Temps moyen d'entrainement : 263.54s
```

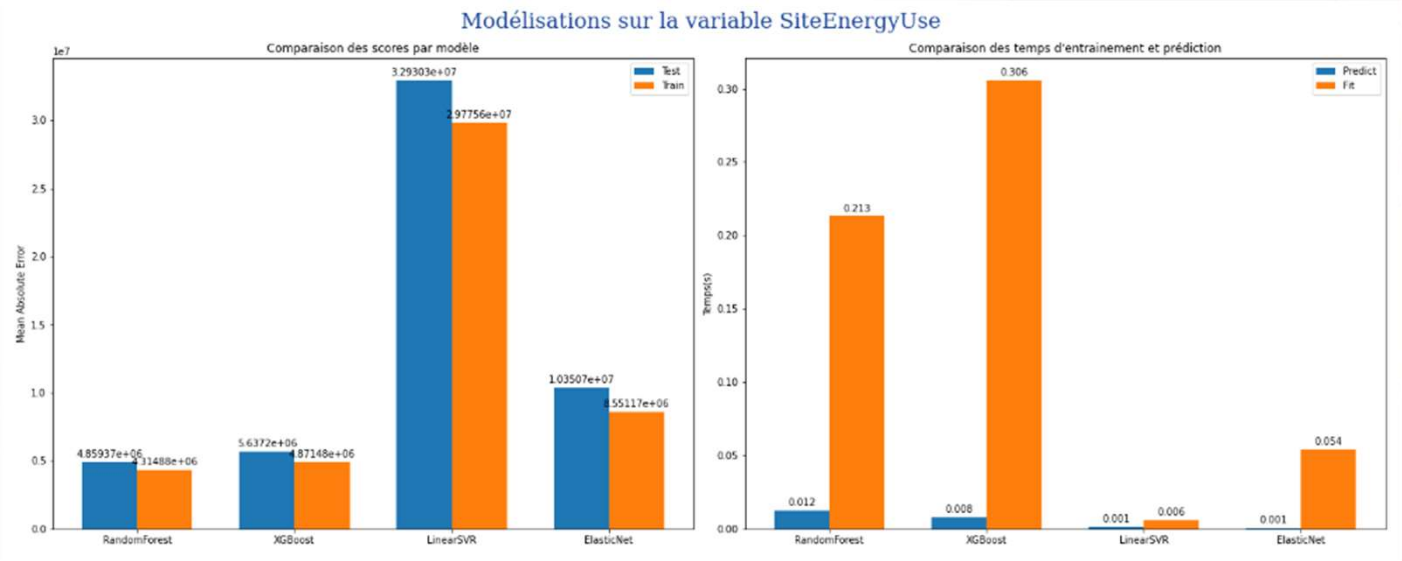
4-2) Modèle XGBoost (eXtreme Gradient Boosting)

```
Meilleur score MAE : -3554694.26
Meilleur Score R2 : 0.651
Meilleurs paramètres : {'regressor__n_estimators': 1000, 'regressor__min_child_weight': 1.0, 'regressor__max_depth': 15, 'regre
ssor__learning_rate': 0.01, 'regressor__gamma': 0.25}
Temps moyen d'entrainement : 378.14s
```

# Modélisation(6/15)

5) Sélection du meilleur modèle

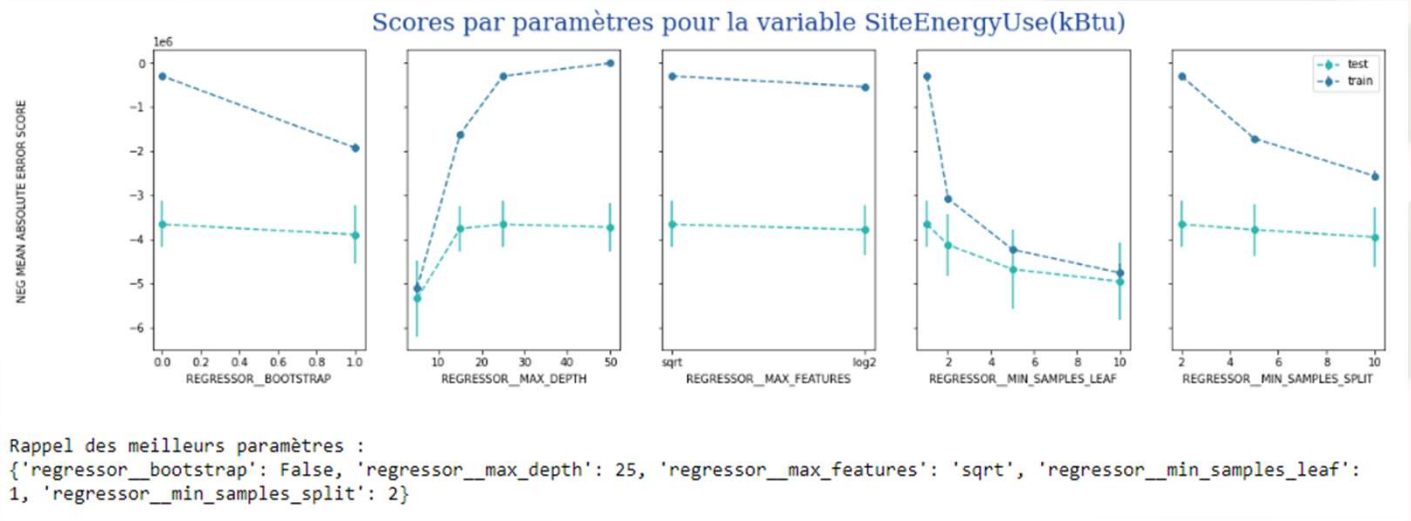
5-1-1) Modèle de prédiction des consommations d'énergie




Pour la variable SiteEnergyUse, le modèle RandomForest offre le meilleur compromis scores MAE et temps d'entrainement et de prédiction

# Modélisation(7/15)

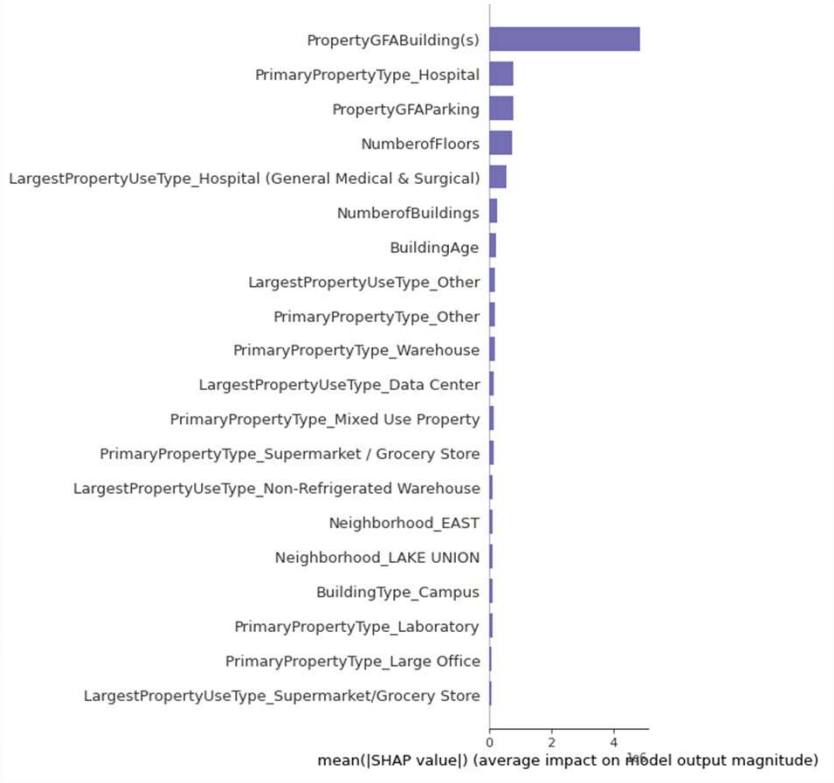
Visualisation des impacts des hyperparamètres de la GridSearch

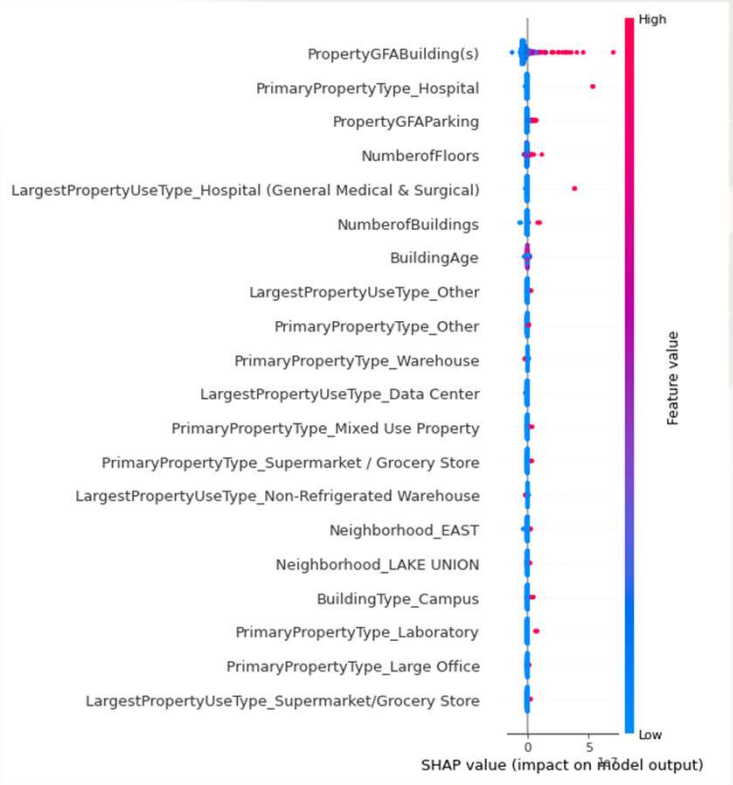



```
Rappel des meilleurs paramètres :
{'regressor__bootstrap': False, 'regressor__max_depth': 25, 'regressor__max_features': 'sqrt', 'regressor__min_samples_leaf':
1, 'regressor__min_samples_split': 2}
```

# Modélisation(8a/15)

Visualisation de l'importance des variables dans notre modèle de forêts aléatoires (global)

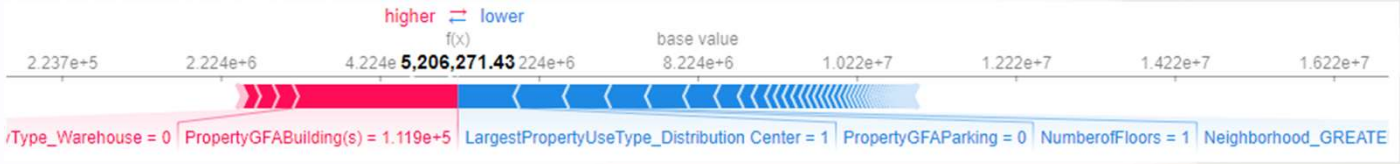


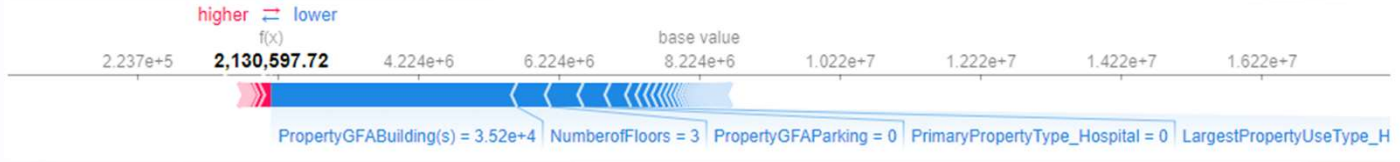
# Modélisation(8b/15)

Visualisation de l'importance des variables dans notre modèle de forêts aléatoires pour des observations

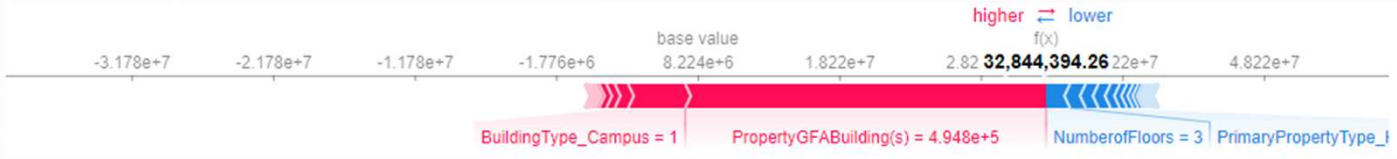
Observation 20




Observation 2



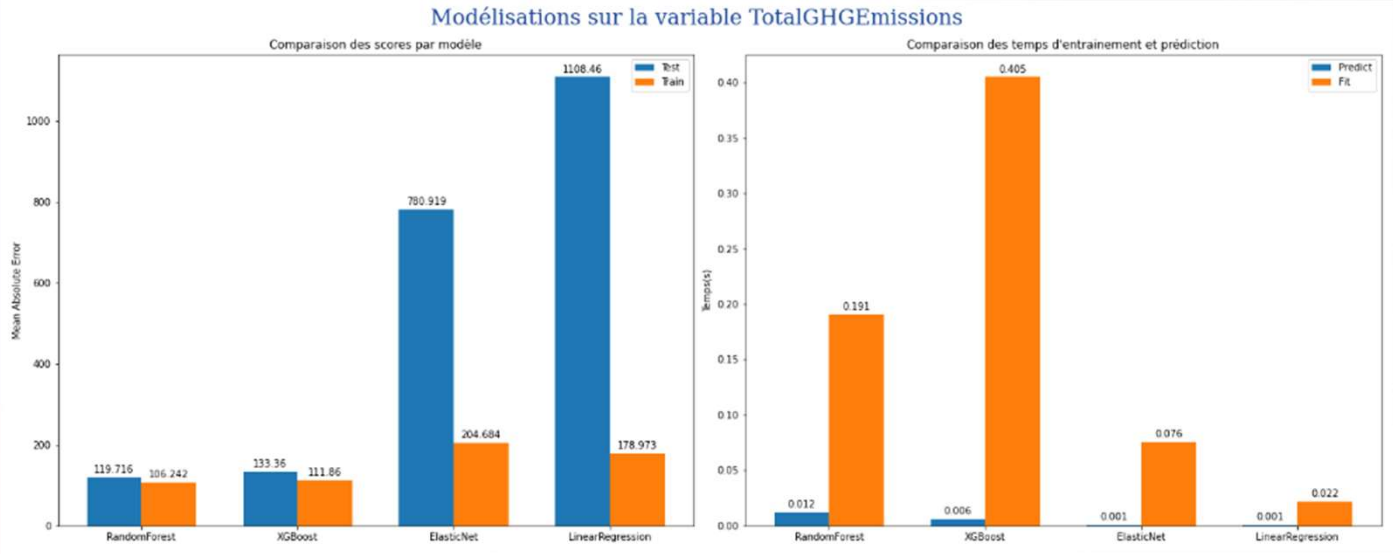

Observation 14




On voit ici les features qui ont un impact positif ou négatif , expliquant la variation par rapport à la base value

# Modélisation(9/15)

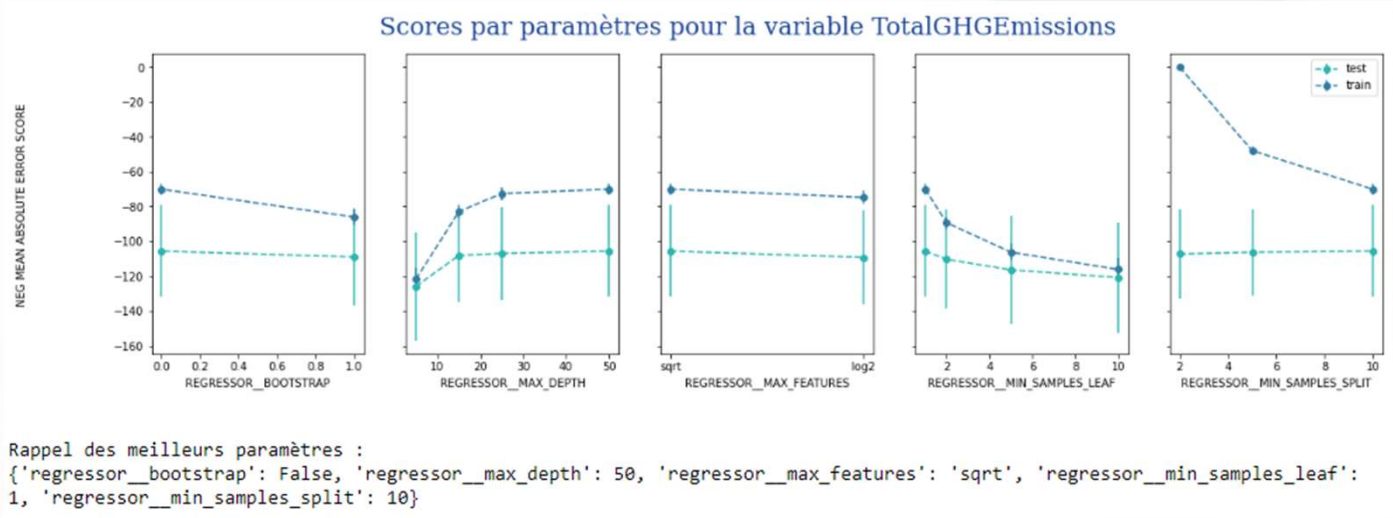
5-1-2) Modèle de prédiction des émissions de gaz à effet de serre




Pour la variable TotalGHGEmissions, le modèle RandomForest offre le meilleur compromis scores MAE et temps d'entrainement et de prédiction

# Modélisation(10/15)

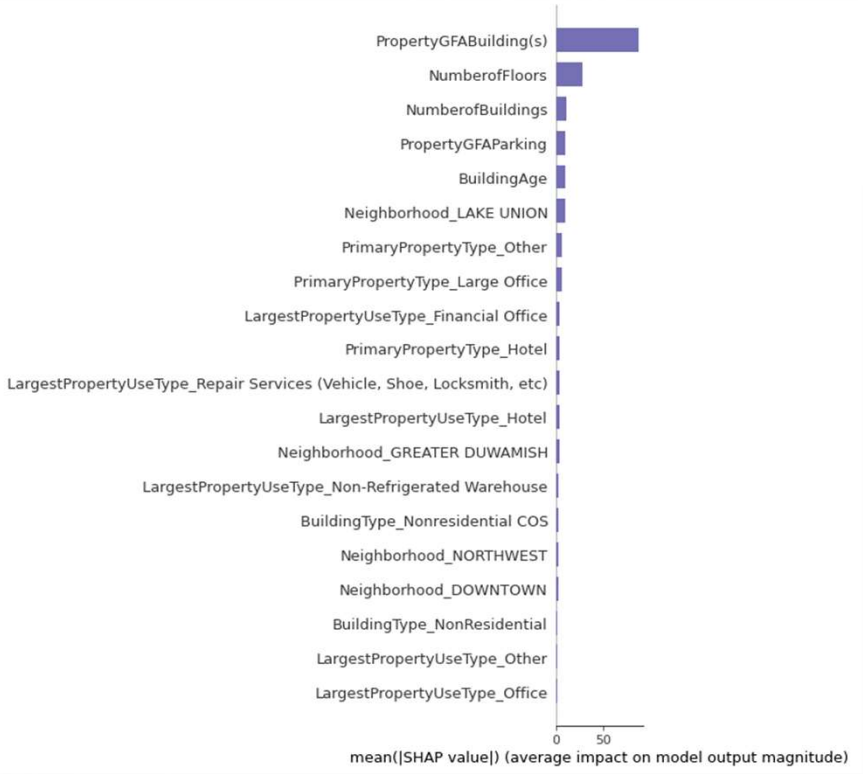
Visualisation des impacts des hyperparamètres de la GridSearch

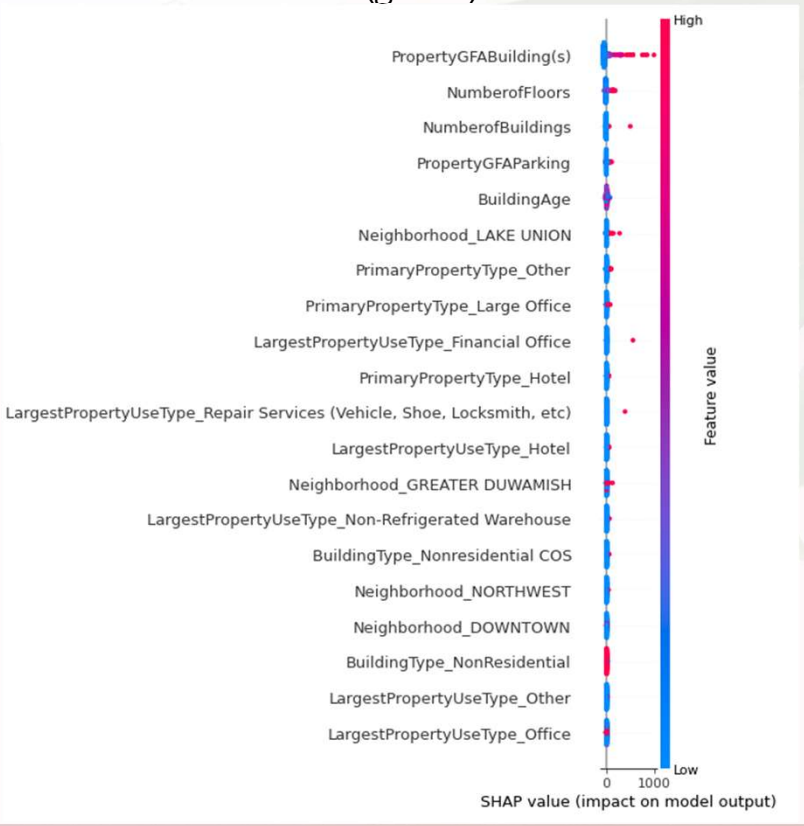



```
Rappel des meilleurs paramètres :
{'regressor__bootstrap': False, 'regressor__max_depth': 50, 'regressor__max_features': 'sqrt', 'regressor__min_samples_leaf':
1, 'regressor__min_samples_split': 10}
```

# Modélisation(11a/15)

Visualisation de l'importance des variables dans notre modèle de forêts aléatoires (global)

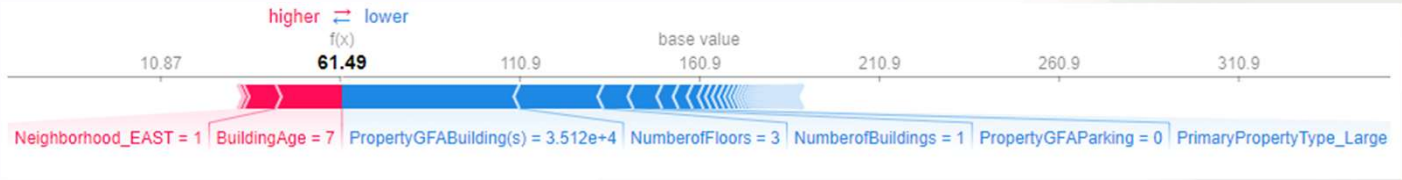


# Modélisation(11b/15)

Visualisation de l'importance des variables dans notre modèle de forêts aléatoires pour des observations

Observation 20



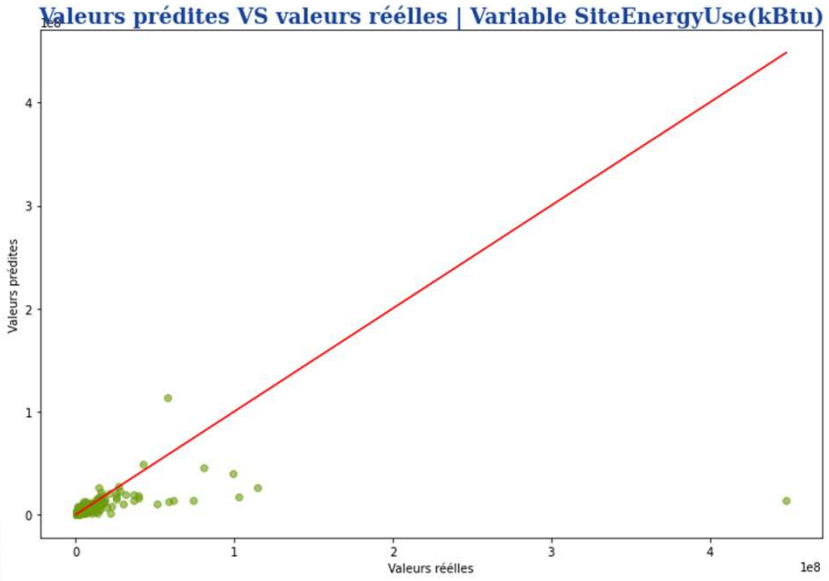

Observation 2

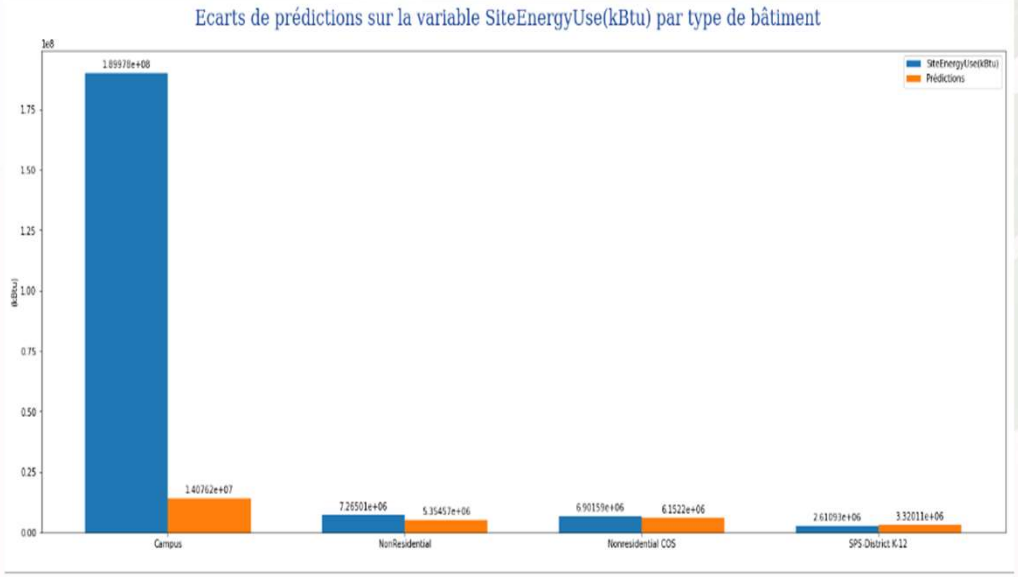

Observation 4


On voit ici les features qui ont un impact positif ou négatif , expliquant la variation par rapport à la base value

# Modélisation(12/15)

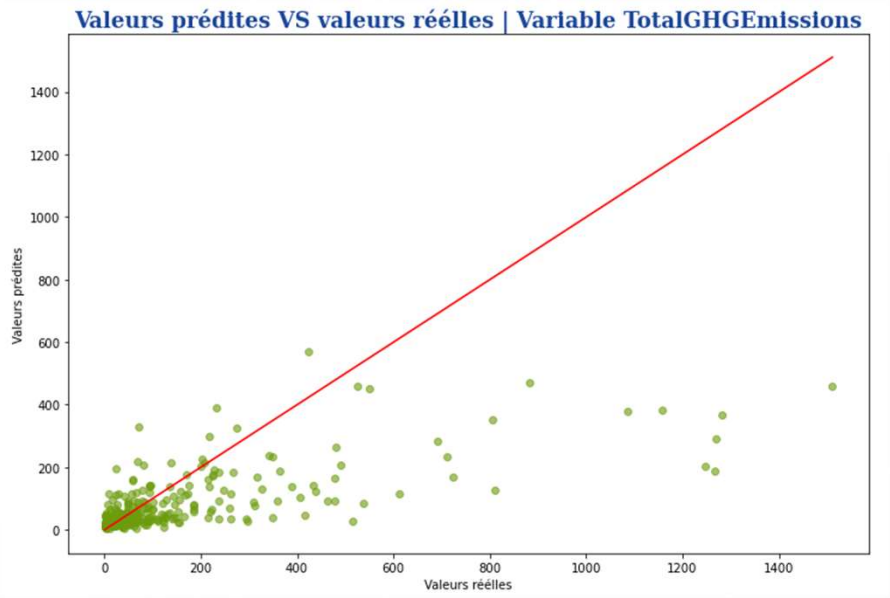
6) Test des modèles sélectionnés

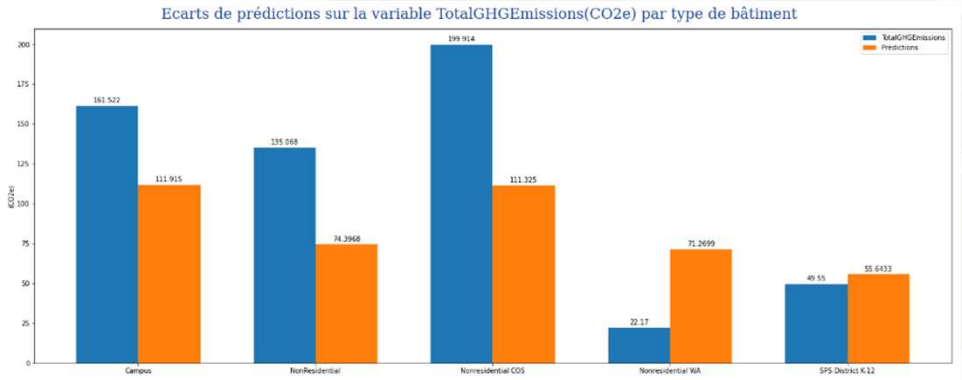



L'écart est très important sur la catégorie "Campus" qui est faiblement représentée dans le jeu de données mais qui présente les plus grandes consommations.

# Modélisation(13/15)

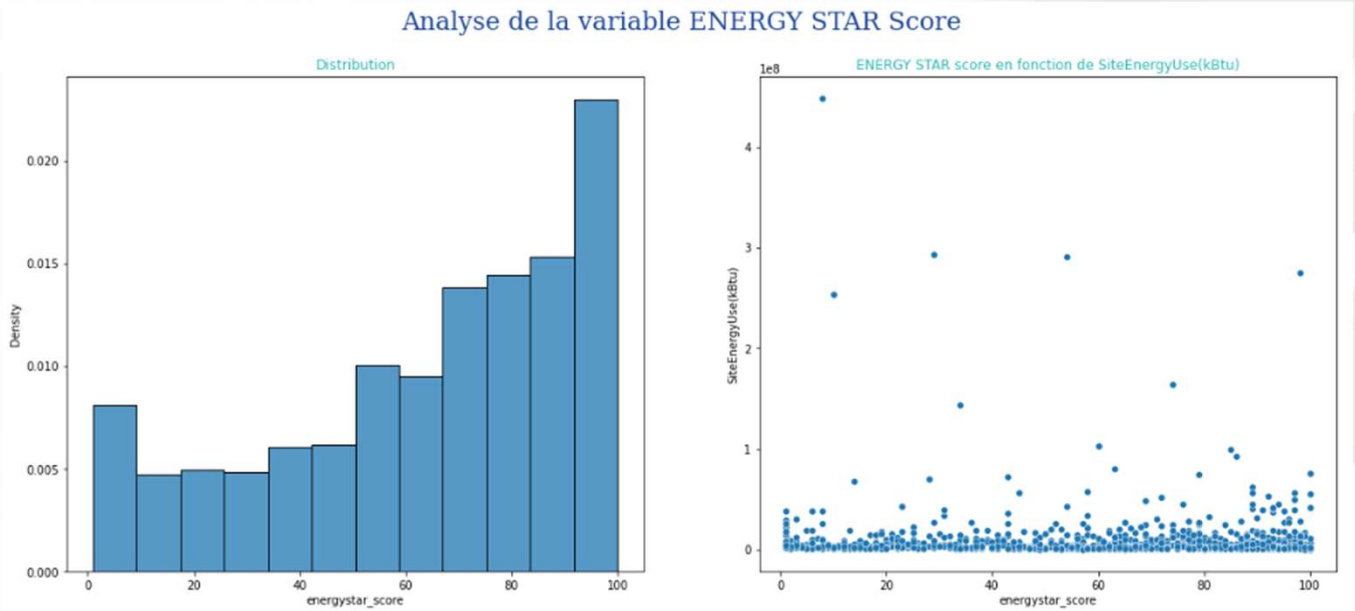



Les écarts de prédiction ne se concentrent pas sur un type de bâtiment comme la feature SiteEnergyUse
Sous-évaluation pour 3 catégories et sur-évaluation pour une autre, la 5e est relativement proche

# Modélisation(14/15)

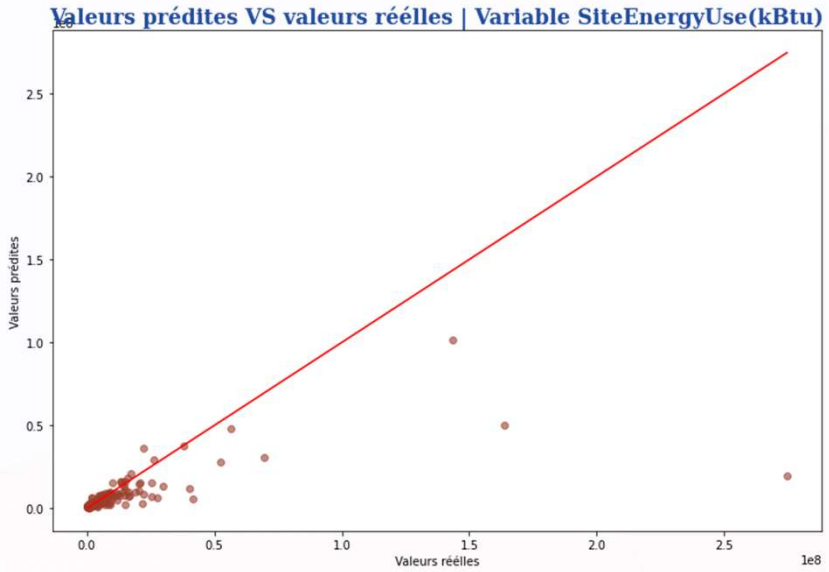
7) Influence du score ENERGY STAR




On remarque ici que le score ENERGY STAR ne semble pas avoir de corrélation importante avec la consommation d'énergie. La distribution ne suit pas de loi normale et la majorité des bâtiments a un score supérieur à 50 (de bonne qualité voir de très bonne qualité).

# Modélisation(15/15)

7) Influence du score ENERGY STAR




Consommation d'énergie

| | Métrique | Sans ENERGY STAR | Avec ENERGY STAR |
|---|---|---|---|
| 0 | MAE | 5.295473e+06 | 4.668833e+06 |
| 1 | $R^2$ | 1.238042e-01 | 3.465774e-01 |

Emissions de gaz à effet de serre

| | Métrique | Sans ENERGY STAR | Avec ENERGY STAR |
|---|---|---|---|
| 0 | MAE | 88.619471 | 111.550684 |
| 1 | $R^2$ | 0.345278 | 0.298050 |

Pour le coût de collecte, l'amélioration n'apparaît pas comme significative

# Merci

Laurent Cagniart