Secteur du bâtiment non résidentiel. Prédiction consommation énergétique et émissions de CO2

Tests d'algorithmes en Machine Learning.

Sommaire.

Contexte de l'étude.

Le secteur du bâtiment représente 44 % de l'énergie consommée en France, loin devant le secteur des transports (31,3%)

Chaque année, le secteur du bâtiment émet plus de 123 millions de tonnes de C02, ce qui en fait l'un des domaines clé dans la lutte contre le réchauffement climatique et la transition énergétique.

Pour rendre le bâtiment plus économe en énergie, il faut rénover massivement l'existant et développer des normes plus strictes en termes de consommation d'énergie pour les bâtiments neufs. C'est l'objet de la politique de l'énergie dans les bâtiments. (https://www.ecologie.gouv.fr/energie-dans-batiments).

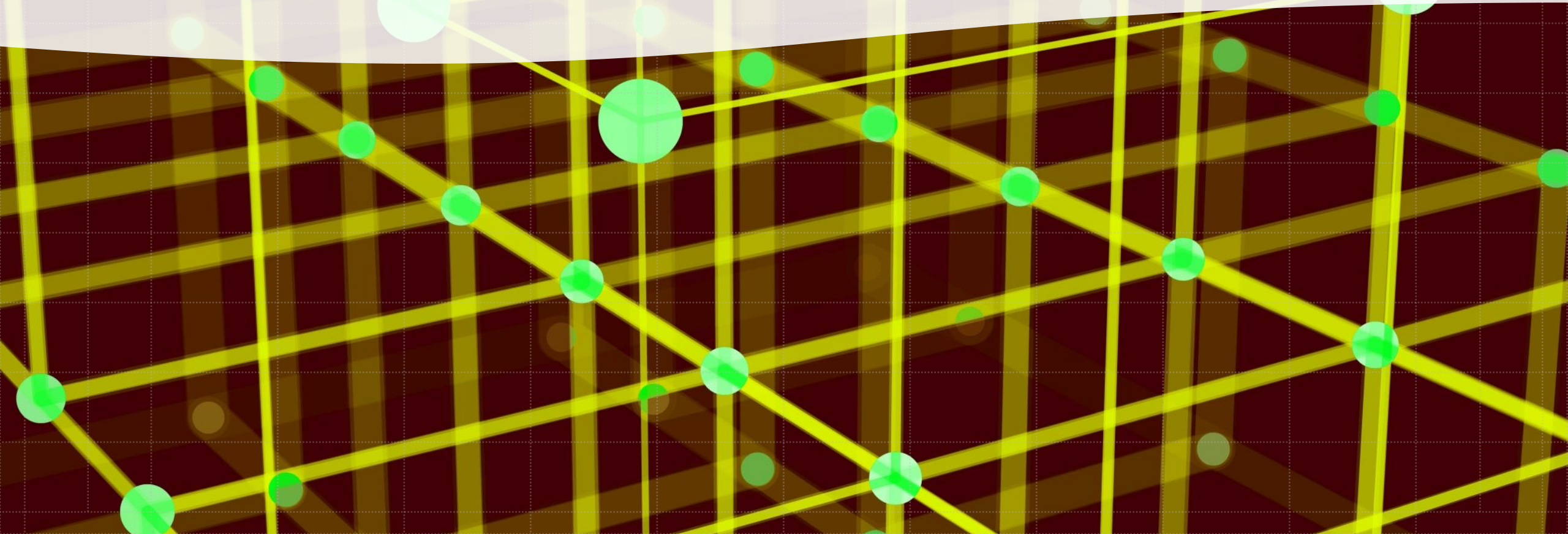
Dans ce contexte de réduction drastique de la consommation énergétique des bâtiments et de la réduction tout aussi sévère des émissions de CO2, ainsi que du renchérissement du coût de l'énergie la ville a depuis plusieurs années procédé à des relevés de mesures.

Celles-ci sont couteuses. A partir de ces relevés, existe t-il des techniques fiables permettant d'estimer la consommation et les niveaux d'émissions pour des bâtiments non destinés à l'habitation et pour lesquels on n'a pas de mesures ?

Les relevés fournis concernent des données structurelles des bâtiments (taille et usage des bâtiments, date de construction, situation géographique, ...)

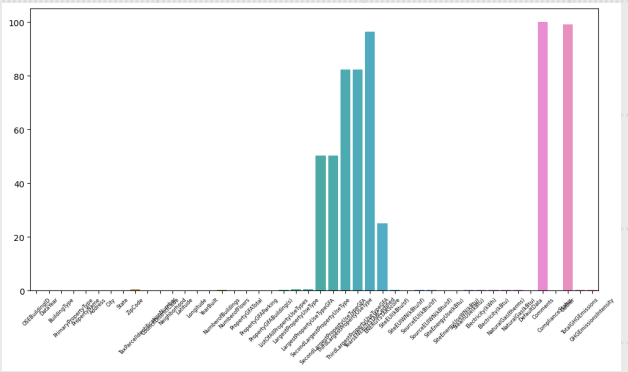
Dans cette étude, on s'intéresse également à la mesure "Energy star score" pour la prédiction d'émissions. Mesure d'une approche fastidieuse et l'on cherchera à juger de son intérêt.

Phase exploratoire.



Phase exploratoire.

- Profil dataset.
  - 3376 lignes, 46 colonnes.
  - Format des variables:
    - Numérique-float64: 22
    - Numérique-int64: 8
    - object: 15
    - Boolean: 1
  - Valeurs manquantes.
    - 8 variables avec taux > 20%.

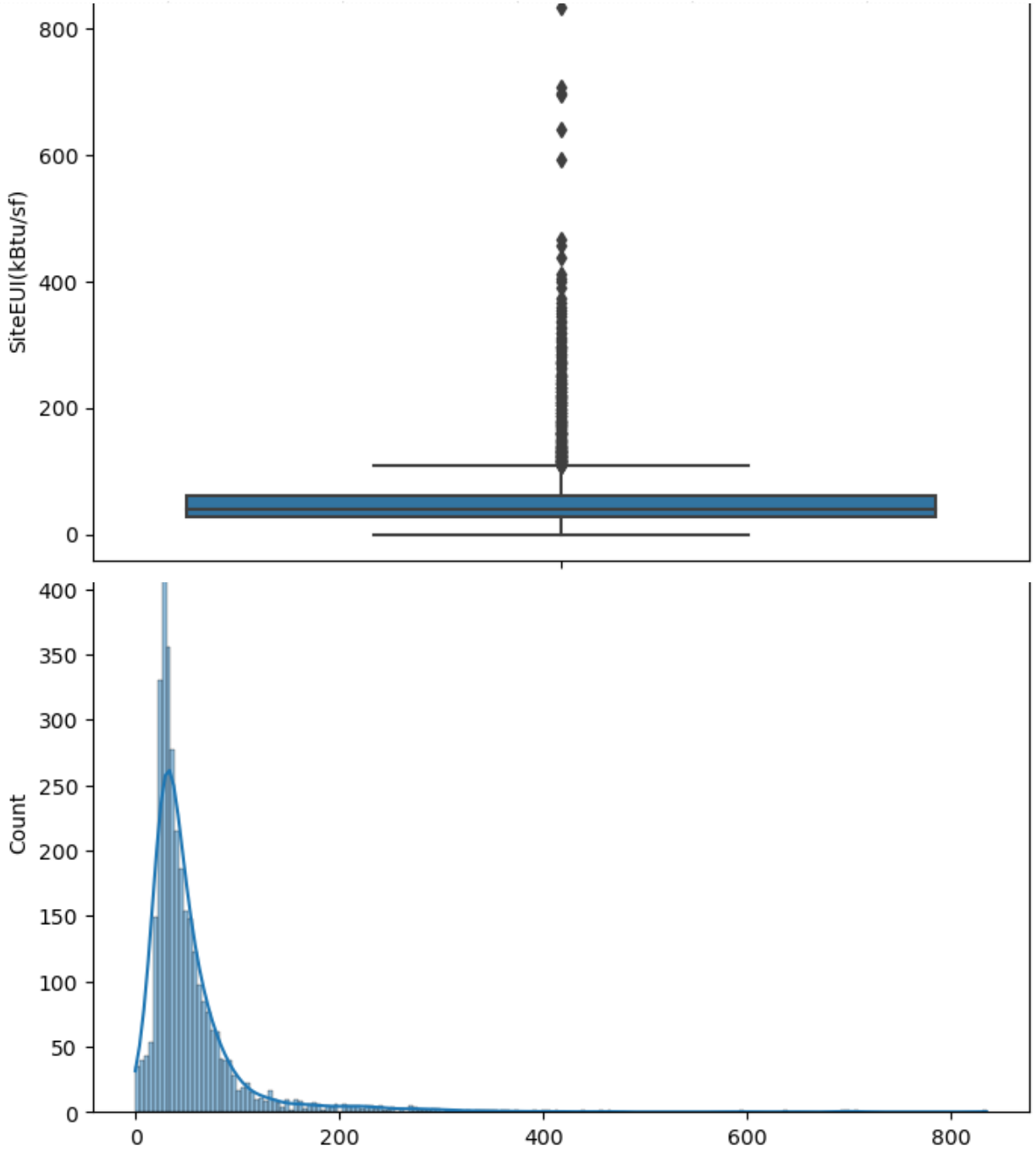


ZipCode	0.473934
TaxParcelIdentificationNumber	0.000000
CouncilDistrictCode	0.000000
Neighborhood	0.000000
Latitude	0.000000
Longitude	0.000000
YearBuilt	0.000000
NumberofBuildings	0.236967

SecondLargestPropertyUseType	50.266588
SecondLargestPropertyUseTypeGFA	50.266588
ThirdLargestPropertyUseType	82.345972
ThirdLargestPropertyUseTypeGFA	82.345972
YearsENERGYSTARCertified	96.475118
ENERGYSTARScore	24.970379

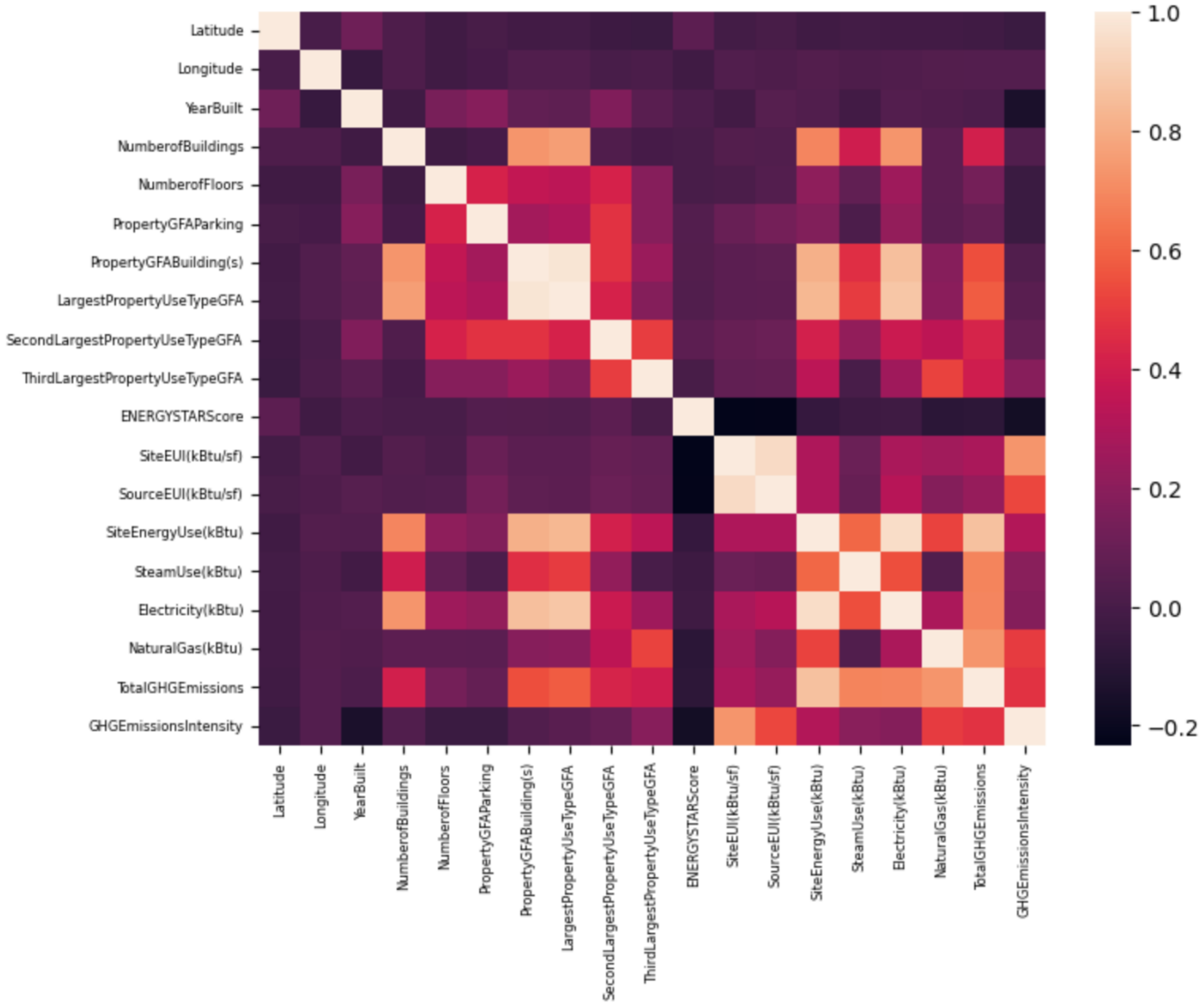
Label: densité de probabilité.

- Label: valeurs remarquables.
  - Label, val. moy: 54.732116399095844
  - Label, val. médiane: 38.59999847
  - Label, val. max: 834.4000244
  - Label, val. min: 0.0
  - Label, val. variance: 3166.664495211066
  - Label, val. écart type: 56.27312409322114
- Distribution.
  - Forte variance (« Nan »)
  - Outliers



Numerical features.

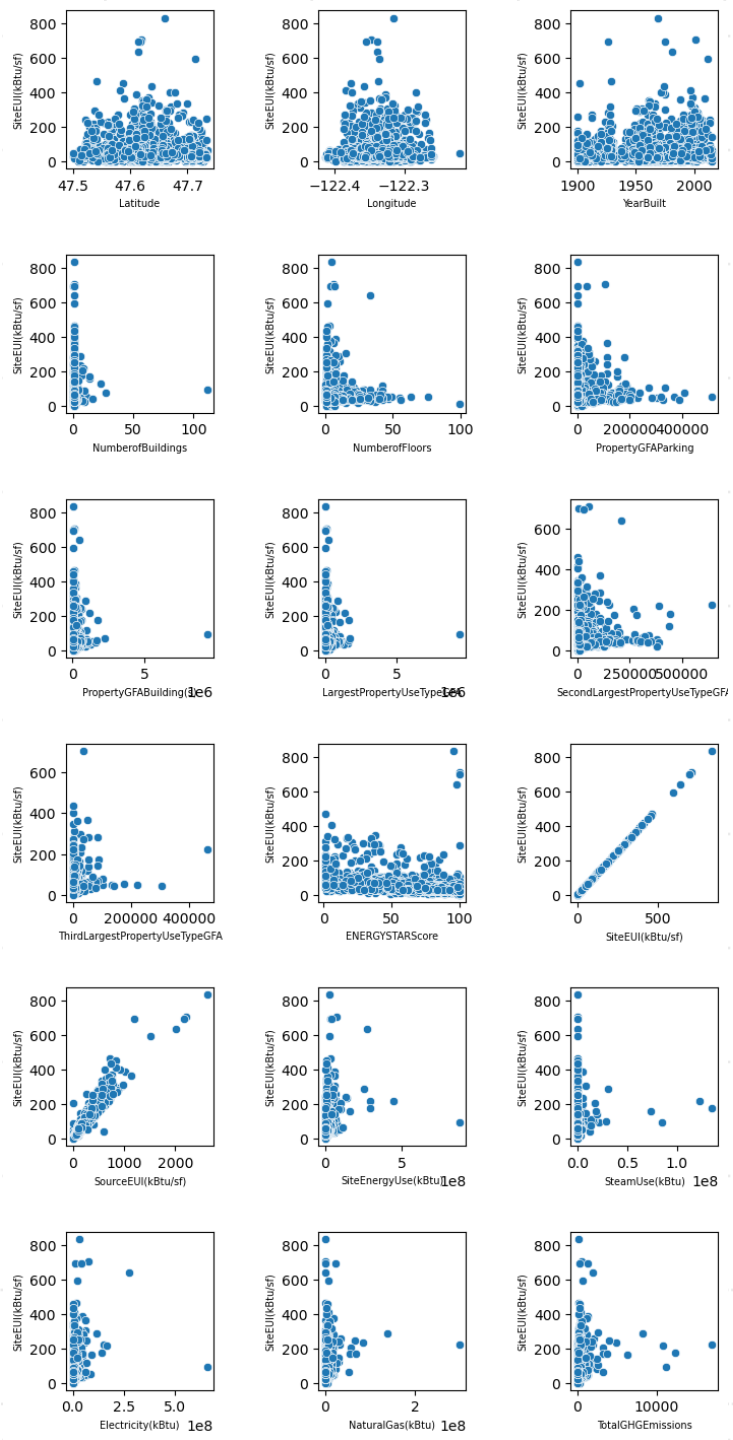
- Plusieurs variables explicatives corrélées.
- Exclusion des variables suivantes:
  - 'PropertyGFATotal'
  - 'SiteEUIWN(kBtu/sf)'
  - 'SourceEUIWN(kBtu/sf)'
  - 'SiteEnergyUseWN(kBtu)'
  - 'Electricity(kWh)'
  - 'NaturalGas(therms)'



Numerical features (suite).

Visualisation des relations entre variables explicatives et variable expliquée.

Pas de relation évidente entre variables indépendantes et label.

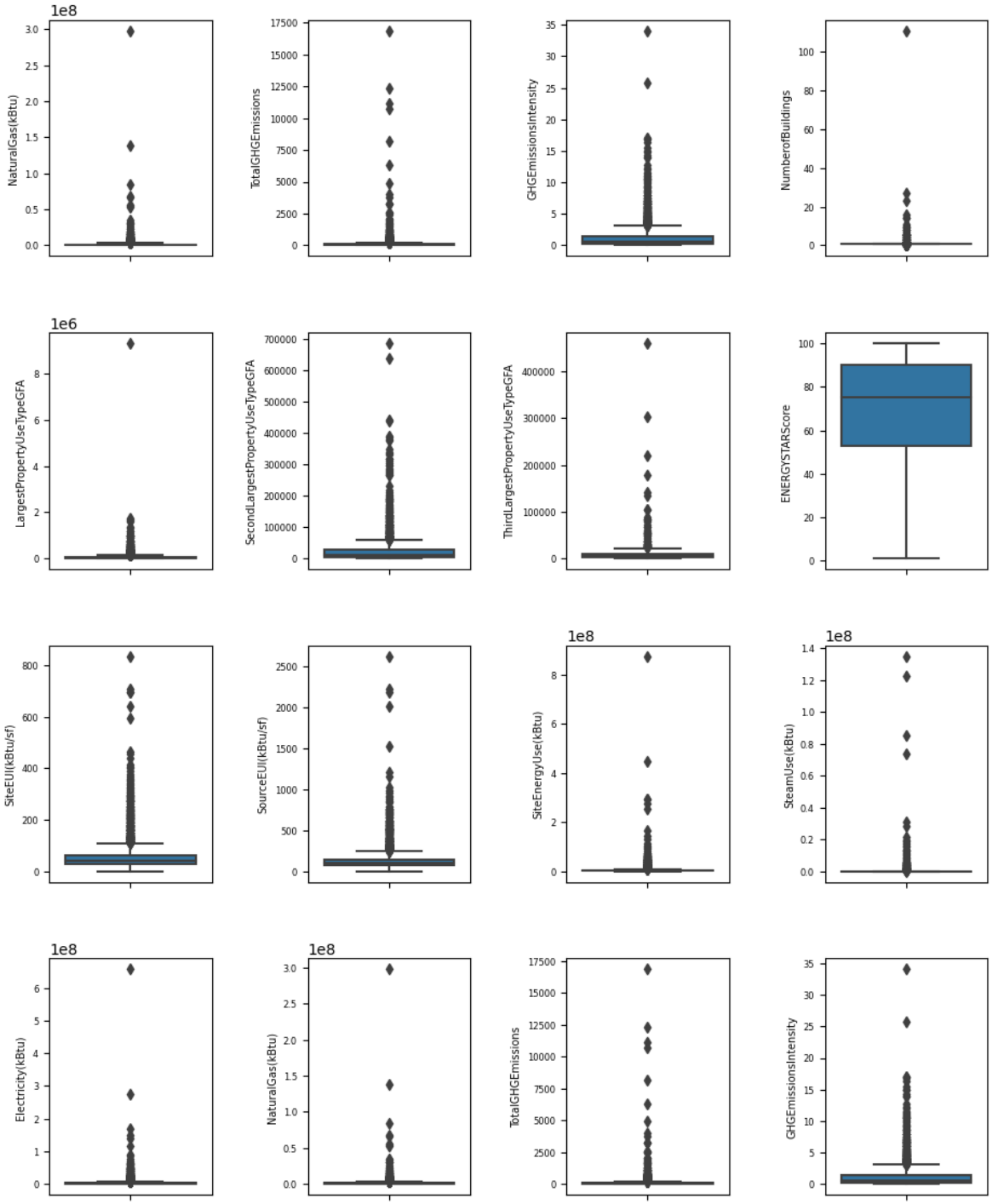


Numerical features (suite).

- Distribution.
  - D'une façon générale, données non centrées.
  - Remplacement valeurs manquantes par médiane.

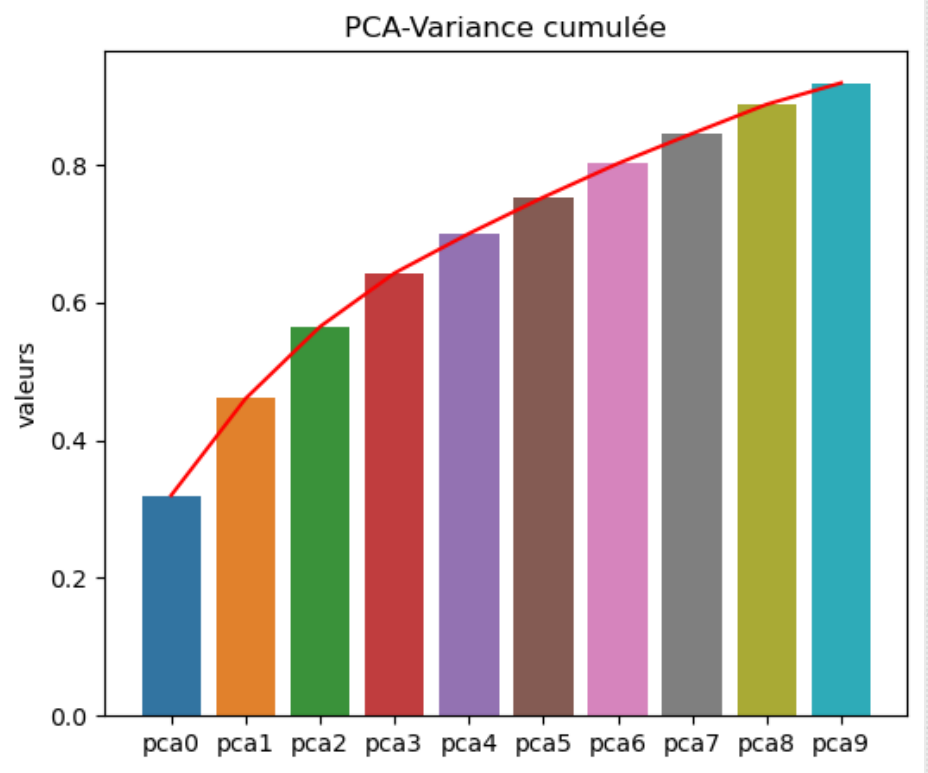
```
numerical_features.isna().mean()*100
```

```
Latitude                          0.000000
Longitude                         0.000000
YearBuilt                         0.000000
NumberofBuildings                 0.236967
NumberofFloors                    0.000000
PropertyGFAParking                0.000000
PropertyGFABuilding(s)            0.000000
LargestPropertyUseTypeGFA         0.592417
SecondLargestPropertyUseTypeGFA  50.266588
ThirdLargestPropertyUseTypeGFA   82.345972
ENERGYSTARScore                  24.970379
SiteEUI(kBtu/sf)                  0.207346
SourceEUI(kBtu/sf)                0.266588
SiteEnergyUse(kBtu)               0.148104
SteamUse(kBtu)                    0.266588
Electricity(kBtu)                 0.266588
NaturalGas(kBtu)                  0.266588
TotalGHGEmissions                 0.266588
GHGEmissionsIntensity             0.266588
dtype: float64
```



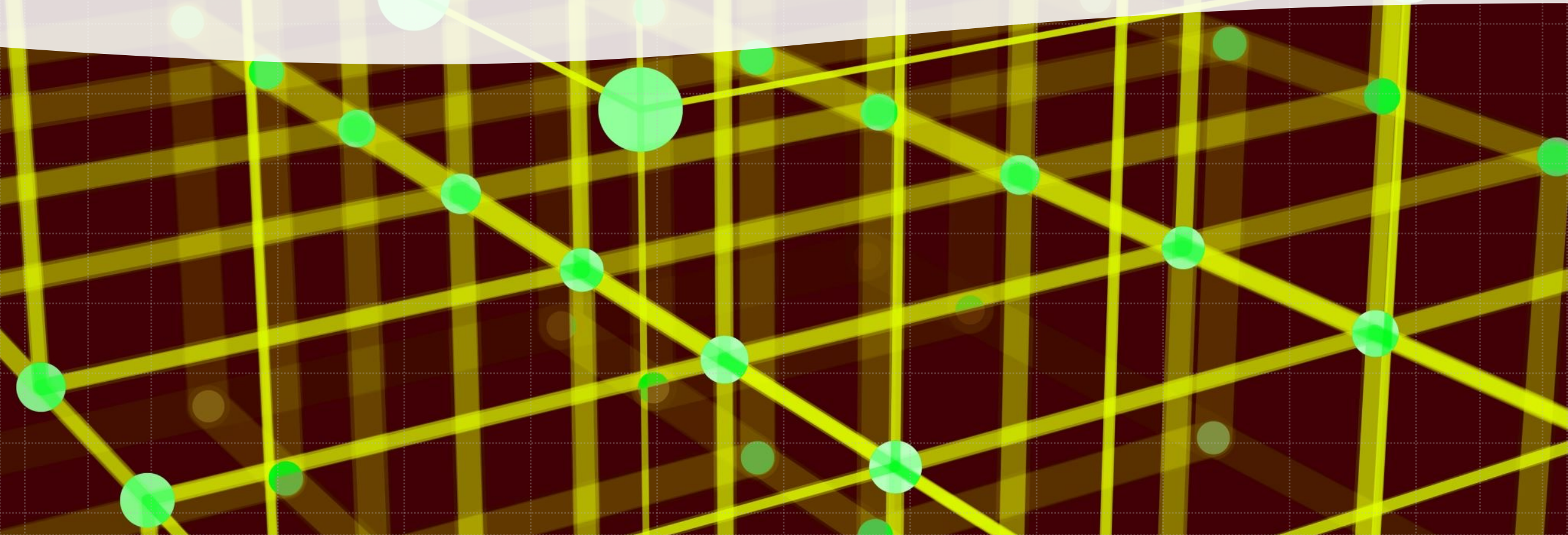
Réduction dimension.

L'utilisation de la méthode PCA est intéressante à plus d'un titre:	Intérêt dans notre contexte ?
• Réduction dimension en ne gardant que les variables « informatives ». • Éliminer les variables « non informatives ». Plus de variables explicatives corrélées.	• A ce stade, l'exclusion des variables corrélées est faite. • Réduction dimension avec objectif de conserver 90% de la variance initiale.



Pour atteindre 90% de la variance initiale, il faut 9 composantes principales. On pourrait retenir cette méthode de sélection des features. Mais on va explorer d'autres méthodes pour pouvoir connaitre les features qui exercent une influence sur le label. On préfère faire une sélection de features en se basant soit sur VarianceThrehold ,soit sur KBest. On étudie ci-après les deux méthodes.

Intérêt de différentes méthodes de selection de variables. Performance du modèle.



Variance Threshold.

Par cette méthode, on peut retenir les features qui ont une variance supérieure à un seuil donné.

Détermination seuil: valeur médiane des variances moyennes de chaque variable.

Output:

- Quatre features retenues.
- Selected features: ['SiteEnergyUse(kBtu)' 'SteamUse(kBtu)' 'Electricity(kBtu)' 'NaturalGas(kBtu)']
- Quatorze features non retenues.
- No selected features: ['Latitude' 'Longitude' 'YearBuilt' 'NumberofBuildings' 'NumberofFloors' 'PropertyGFAParking' 'PropertyGFABuilding(s)' 'LargestPropertyUseTypeGFA' 'SecondLargestPropertyUseTypeGFA' 'ThirdLargestPropertyUseTypeGFA' 'ENERGYSTARScore' 'SourceEUI(kBtu/sf)' 'TotalGHGEmissions' 'GHGEmissionsIntensity']

Conclusion.

- On voit que certaines features sont absentes. Il aurait été intéressant d'avoir l'année de construction, le nombre de bâtiments, d'étages par exemple dont on peut supposer qu'elles ont un impact sur la target. Dans une deuxième approche on essaye la méthode Kbest.

SelectKBest, SVR.

- Constitution des sets X_num et X_cat.
  - Exclusion des variables catégorielles (jugées non informatives): 'City', 'State', 'TaxParcelIdentificationNumber', 'ListOfAllPropertyUseTypes', 'YearsENERGYSTARCertified', 'PropertyName', 'Address', 'ComplianceStatus'.
- Exclusion des catégories pour lesquelles le nombre d'échantillons est égal à 1 (pénalisant lors du train_test_split).
- Subset Label = 'SiteEUI(kBtu/sf)'.

```
X_cat.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3376 entries, 0 to 3375
Data columns (total 6 columns):
 #   Column                          Non-Null Count  Dtype 
---  ------                          --------------  ----- 
 0   BuildingType                    3376 non-null   object
 1   PrimaryPropertyType             3376 non-null   object
 2   Neighborhood                    3376 non-null   object
 3   LargestPropertyUseType          3356 non-null   object
 4   SecondLargestPropertyUseType    1679 non-null   object
 5   ThirdLargestPropertyUseType     596 non-null    object
dtypes: object(6)
memory usage: 158.4+ KB
```

SelectKBest, SVR.

- Subset X.
  - Dimensions: (3345, 23).
  - numerical_features, categorical_features.
  - Construction pipelines.
    - numerical_pipeline.
      - Remplacement Nan: médiane.
      - Centrer, scaling: RobustScaler()
    - categorical pipeline.
      - Remplacement Nan: + fréquente
      - OneHotEncoder() + infrequent_if_exist
  - Transformations.
  - Construction « Train » et « Test » sets.
    - Taille échantillon Test: 20%.
  - Construction modèle.
    - Méthode Kbest. f_regression, k=10
    - Support vestor regressor. Noyau Rbf.
  - Entrainement modèle.
  - **Performance modèle: 0.79 (baseline, coeff. détermination)**

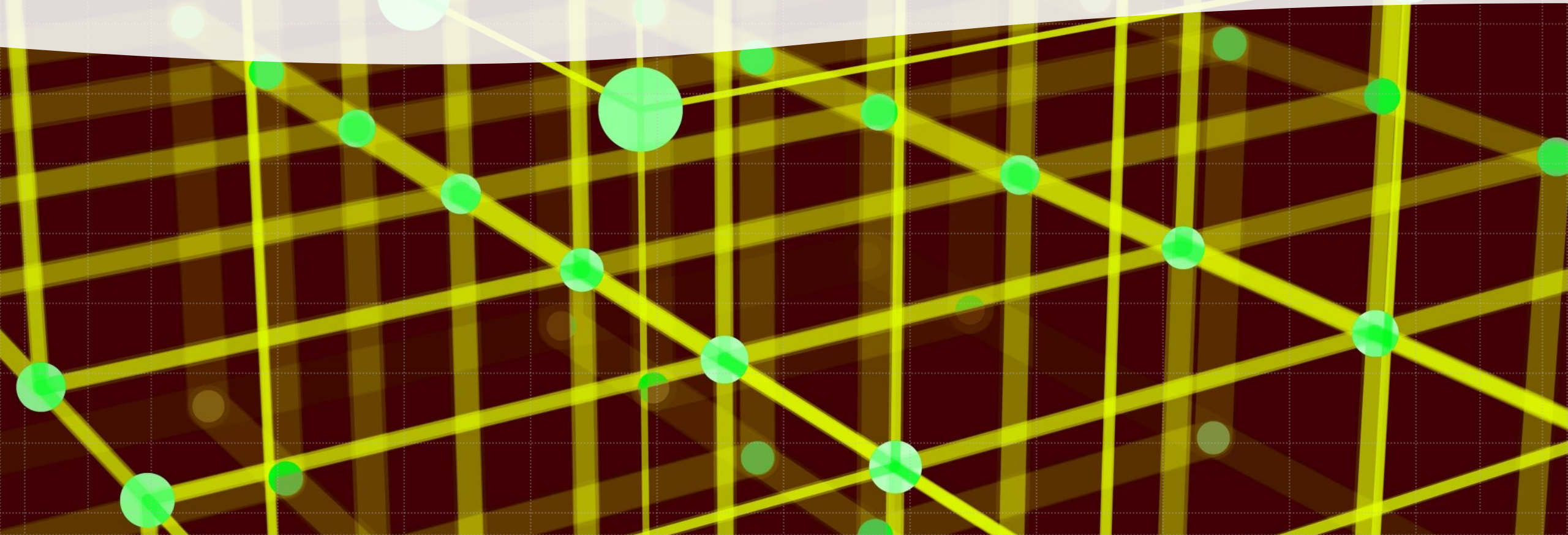
```
X_wo_Y.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3345 entries, 0 to 3344
Data columns (total 23 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Latitude                         3345 non-null   float64
 1   Longitude                        3345 non-null   float64
 2   YearBuilt                        3345 non-null   int64
 3   NumberofBuildings                3345 non-null   float64
 4   NumberofFloors                   3345 non-null   int64
 5   PropertyGFAParking               3345 non-null   int64
 6   PropertyGFABuilding(s)           3345 non-null   int64
 7   LargestPropertyUseTypeGFA        3345 non-null   float64
 8   SecondLargestPropertyUseTypeGFA  3345 non-null   float64
 9   ThirdLargestPropertyUseTypeGFA   3345 non-null   float64
 10  ENERGYSTARScore                  3345 non-null   float64
 11  SourceEUI(kBtu/sf)               3345 non-null   float64
 12  SiteEnergyUse(kBtu)              3345 non-null   float64
 13  SteamUse(kBtu)                   3345 non-null   float64
 14  Electricity(kBtu)                3345 non-null   float64
 15  NaturalGas(kBtu)                 3345 non-null   float64
 16  TotalGHGEmissions                3345 non-null   float64
 17  GHGEmissionsIntensity            3345 non-null   float64
 18  BuildingType                     3345 non-null   object
 19  PrimaryPropertyType              3345 non-null   object
 20  Neighborhood                     3345 non-null   object
 21  LargestPropertyUseType           3325 non-null   object
 22  SecondLargestPropertyUseType     1658 non-null   object
dtypes: float64(14), int64(4), object(5)
```

SelectKBest, PolynomialFeatures, SVR.

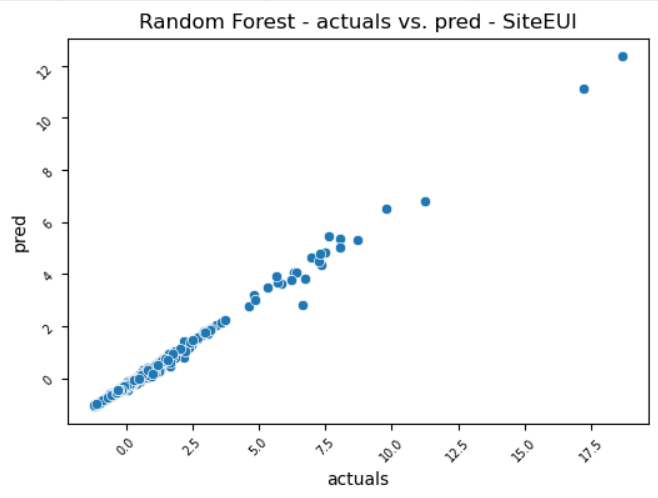
- Features input: variables sélectionnées par méthode selectKBest().
- Transformation des features par application d'une fonction linéaire polynomiale (mieux segmenter les variables à forte variance des autres).
- Centrage, scaling.
- Recherche optimisation paramètre 'degré polynome': utilisation méthode GridSearchCV()
- Application estimateur LinearRegression().
- Performance modèle: voisin de 0.
- NB: On peut conclure que le modèle polynomial n'est pas satisfaisant puisque le score est quasiment nul, ce qui traduit une absence de relation entre y et les feature. NB: ici (contexte de limitation de puissance de calcul ==> restriction du périmètre features à celui de selectKBest)

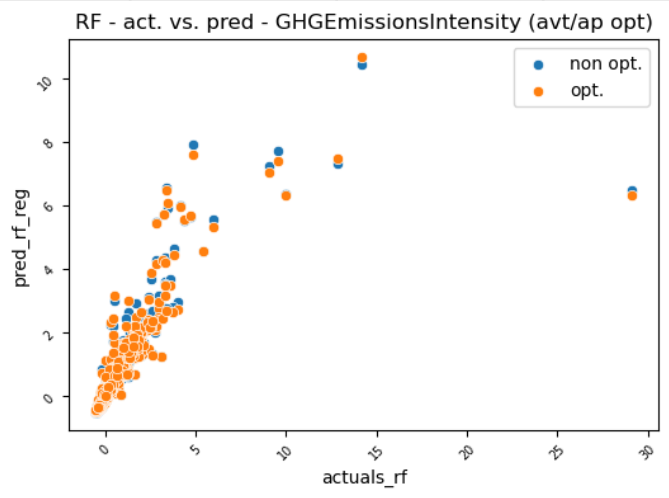
Tests des méthodes ensemblistes.

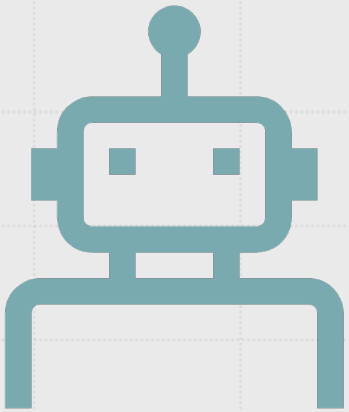


Bagging avec Random Forest.

- Réutilisation preprocesseur.
- Algorithme RandomForestRegressor. Population 1000 arbres décision.
- Entrainement.
- Performance
  - label *('SiteEUI(kBtu/sf')*: 0.94 (score: coeff. détermination)
  - Label *('GHGEmissionsIntensity')*: 0.67/0.66. Avant/après optimisation (score: coeff. détermination)

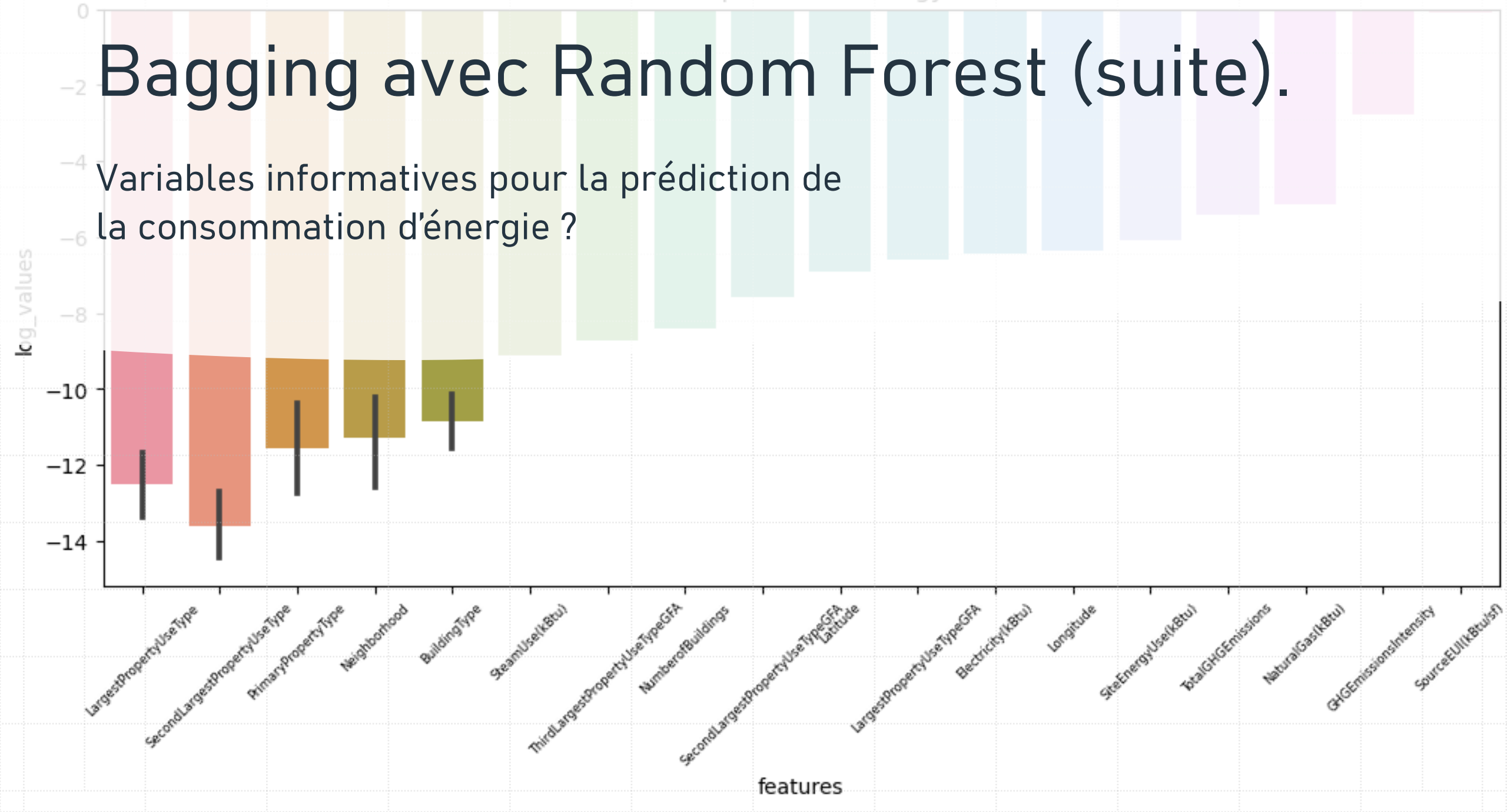






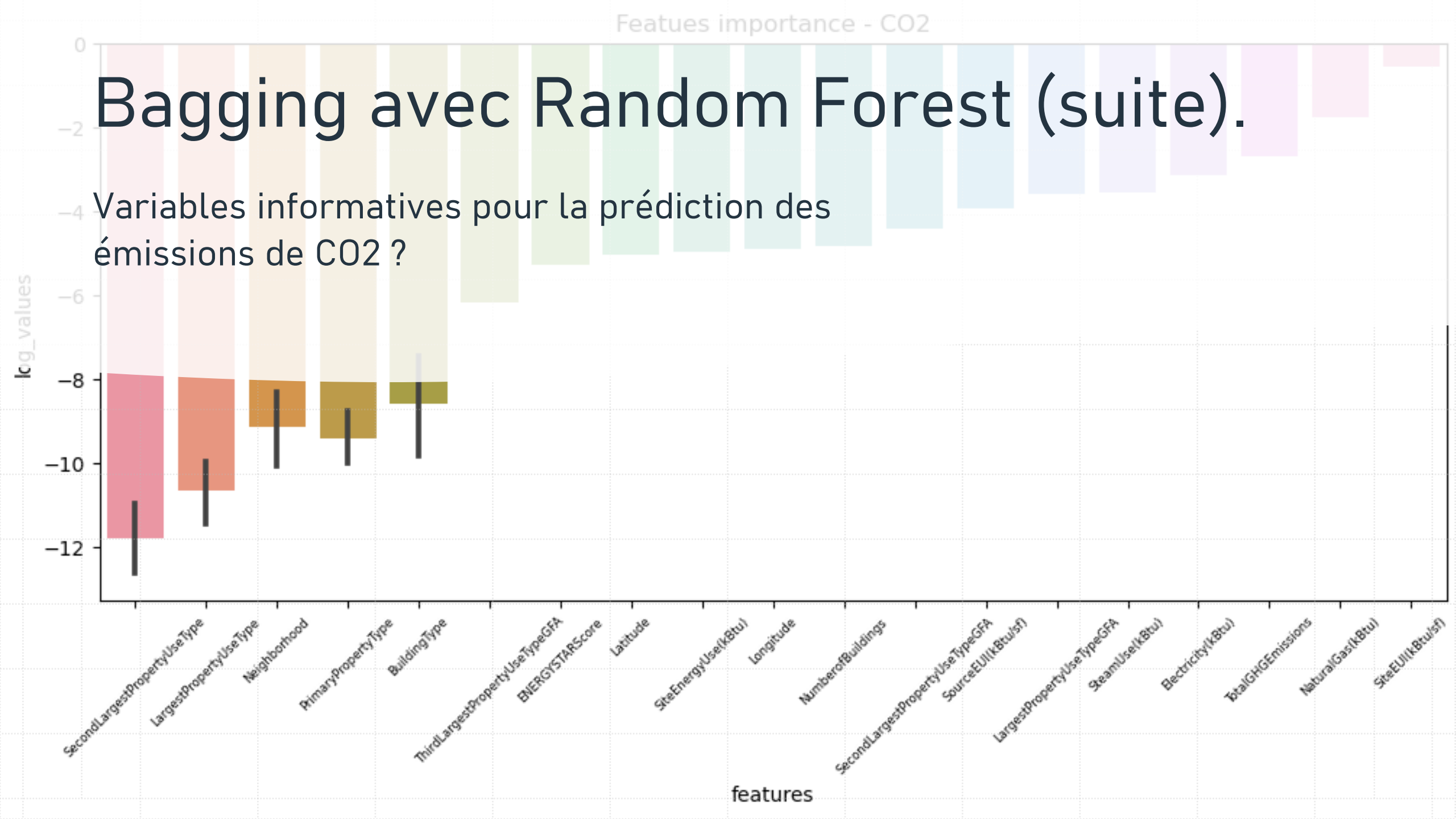
Bagging avec Random Forest (suite).

Variables informatives pour la prédiction de
la consommation d'énergie ?



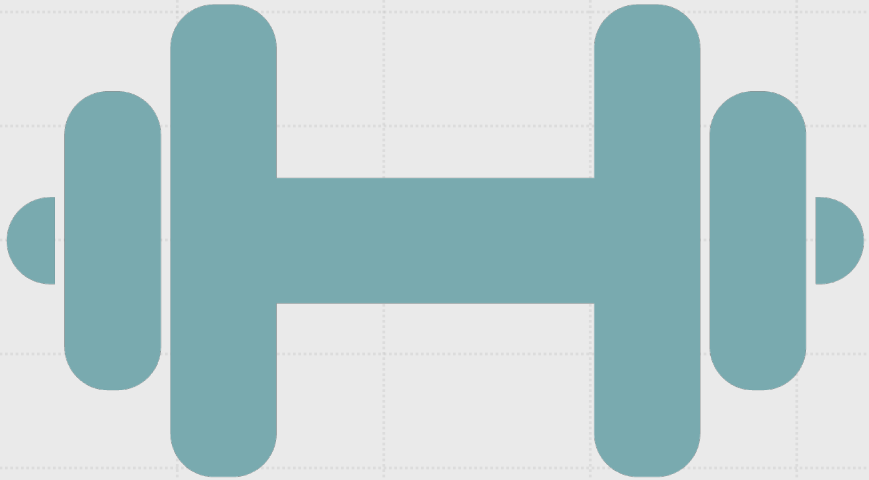
Bagging avec Random Forest (suite).

Variables informatives pour la prédiction des émissions de CO2 ?



Boosting, Adaboost.

- Réutilisation preprocesseur.
- Algorithme AdaBoost. Population 1000 arbres décision.
- Entrainement.
- Performance.
  - label *('SiteEUI(kBtu/sf')*: 0.86 (score: MSE)
  - Label *('GHGEmissionsIntensity')*: 0.24/0.35. avant/après opt. (score: coeff. détermination)



Première conclusion.

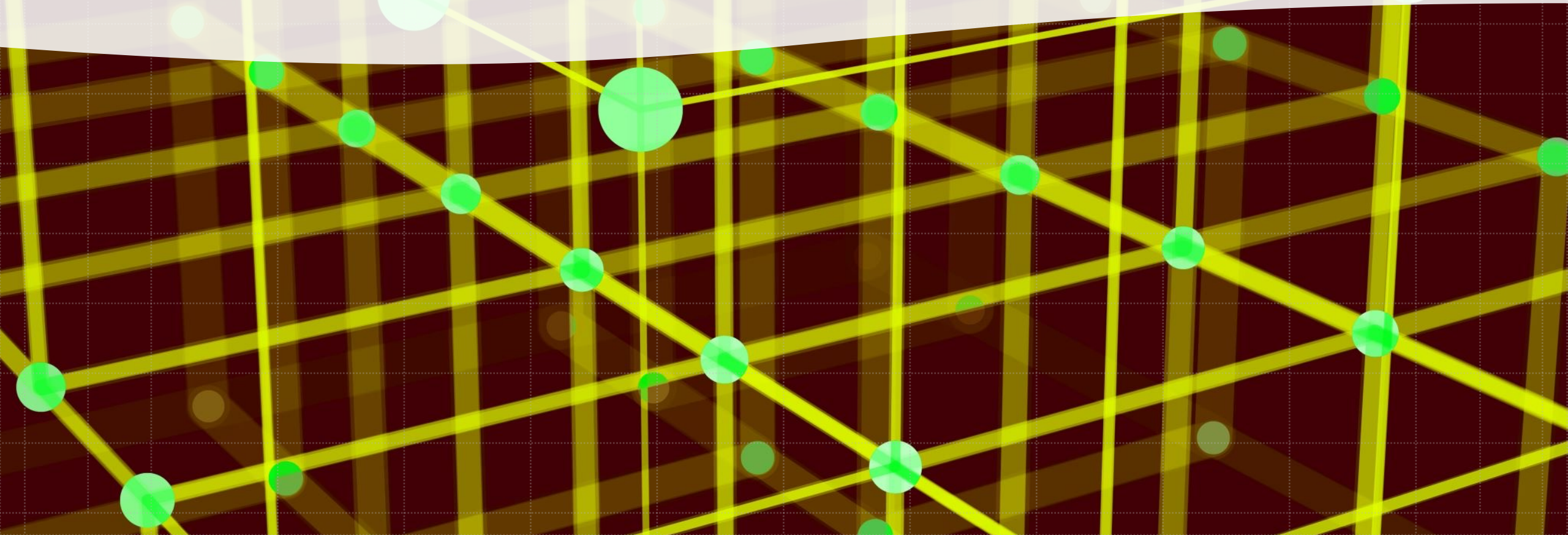
- **A ce stade de l'analyse, on peut tirer une première conclusion: Le modèle le plus performant dans le contexte dans lequel on se trouve est Random Forest. La technique de "Bagging" (analyse en parallèle) est ici plus performante que la technique concurrente du "Boosting » (série).**

- **L'objectif ici est de s'intéresser à la consommation et aux émissions des bâtiments non destinés à l'habitation. Or pour le moment le dataset contient aussi bien des bâtimments d'habitation que des bâtiments non destinés à l'habitation. Il faut donc maintenant exlure du dataset les objets non destinés à l'habitation.**

- **Limitation du dataset au domaine du non résidentiel, en utilisant le modèle le plus performant(RF), donne une performance de 0.95.**

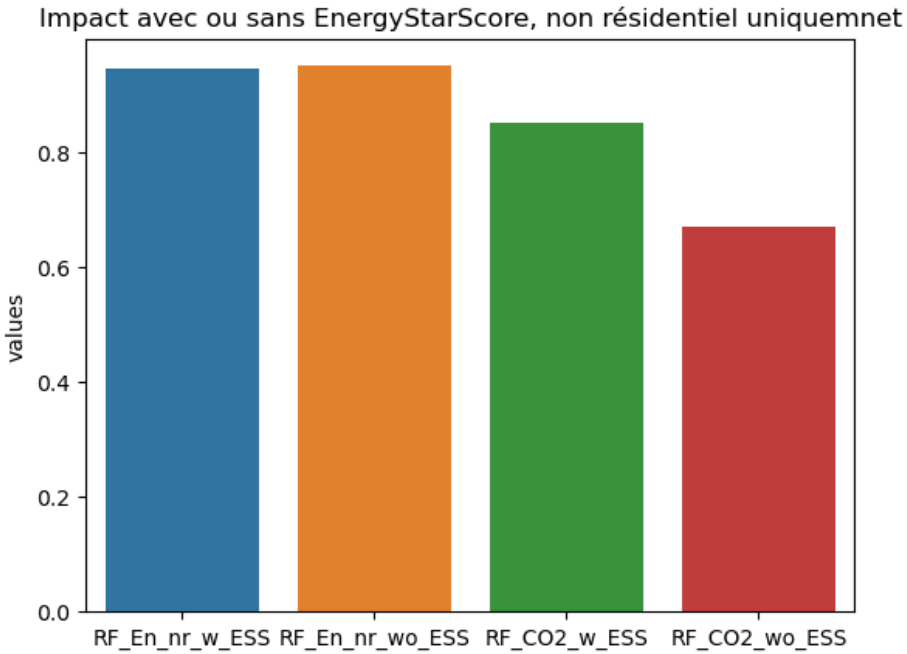
- **Le modèle Random Forest donne de bons résultats pour la prédiction de consommation énergétique, mais est moins performant pour la prédiction des émissions de CO2, et l'optimisation n'a pas délivré la performance attendue.**

Intérêt de l'Energy Star Score.



Intérêt Energy Star Score.

- Dataset utilisé ci-contre. (ENERGYStarScore exclue).
- Analyse basée sur RF.
- Score.
  - Consommation énergétique: 0.94
  - Émissions de CO2:
- Conclusion:
  - Influence sur prédiction consommation énergétique: **pas d'influence**, scores de 0.945 et 0.949 (avec/sans la variable ENERGYStarScore).
  - Influence sur prédiction émissions de CO2: **influence**, scores de 0.85 et 0.67 (avec/sans la variable ENERGYStarScore).
- Conclusion: EnergyStarScore= variable fabriquée, entachée d'erreur, devrait être exclue des analyses.



Secteur du bâtiment non résidentiel. Prédiction consommation énergétique et émissions de CO2

FIN.