#4 Anticipez les besoins en consommation de bâtiments

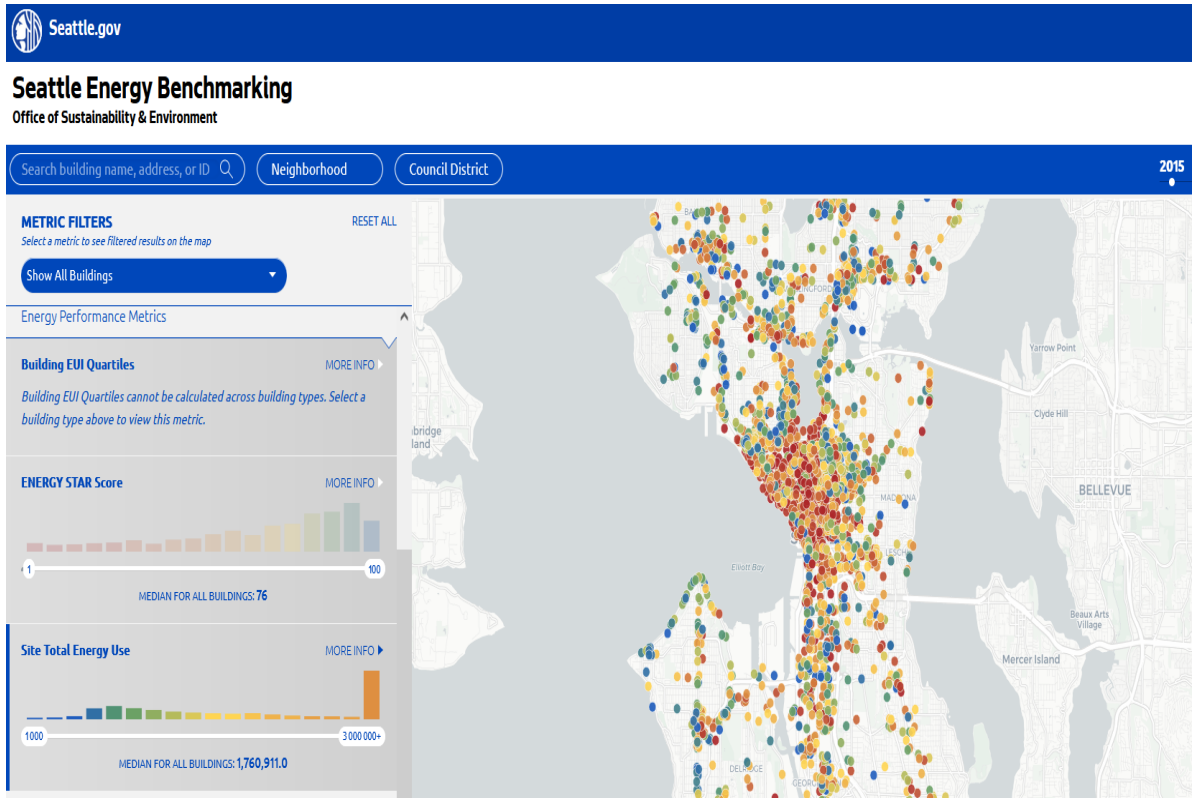
Soutenance Emilie Groschêne le 07/06/2022
Evaluateur: Ababacar Ba
Mentor: Lea Naccache

Seattle

Sommaire

I. PRESENTATION DE LA PROBLEMATIQUE

I. Présentation de la problématique



Mise à disposition par la ville de Seattle d'un **benchmark de la performance énergétique des bâtiments**

Ambition: ville **neutre** en **émissions de carbone en 2050**

Problématique:

- des **relevés manuels** ont été effectués par les agents de la ville en 2015 et 2016
- ils sont **coûteux** et **fastidieux** à obtenir

Mission:

- **prédire les émissions de CO2 et la consommation totale d'énergie,**
- des bâtiments **non destinés à l'habitation,**
- et non encore mesurés,
- en se **passant des relevés manuels,**
- et en évaluant **l'intérêt de l'Energy Star Score** pour les prédictions d'émissions de CO2

I. Présentation de la problématique

Interprétation:

- **Données à prédire (targets):** émissions de CO2 (**Total GHGEmissions**) et consommation totale d'énergie (**SiteEnergyUse(kBtu)**)

- **Features:** données déclaratives du **permis d'exploitation commerciale** et non des relevés de consommation

- **EnergyStarScore:** Modélisation avec et sans cette variable pour juger de son intérêt dans la prédiction de CO2

TotalGHGEmissions	The total amount of greenhouse gas emissions, including carbon dioxide, methane, and nitrous oxide gases released into the atmosphere as a result of energy consumption at the property, measured in metric tons of carbon dioxide equivalent. This calculation uses a GHG emissions factor from Seattle Clty Light's portfolio of generating resources. This uses Seattle City Light's 2015 emissions factor of 52.44 lbs CO2e/MWh until the 2016 factor is available. Enwave steam factor = 170.17 lbs CO2e/MMBtu. Gas factor sourced from EPA Portfolio Manager = 53.11 kg CO2e/MBtu.
SiteEnergyUse(kBtu)	The annual amount of energy consumed by the property from all sources of energy.

II. PREPARATION DU JEU DE DONNEES

II. Préparation du jeu de données – Présentation

2 datasets distincts	2015	2016
1 ligne par bâtiment	3 340	3 376
Nombre de features	47	46
Features différentes entre les datasets	10	9
% données manquantes	16,9%	12,9%

- **Peu d'observations** car de nombreux bâtiments identiques pour les 2 années
- Données parfois **complémentaires** entre les 2 années
- Relativement **peu de valeurs manquantes**

Identification des bâtiments	OSEBuildingID PropertyName TaxParcelIdentificationNumber
Localisation	Address City State ZipCode CouncilDistrictCode Neighborhood Latitude Longitude
Données déclaratives du permis d'exploitation commerciale	YearBuilt BuildingType PrimaryPropertyType NumberofBuildings NumberofFloors PropertyGFATotal PropertyGFAParking PropertyGFABuilding(s) ListOfAllPropertyUseTypes LargestPropertyUseType LargestPropertyUseTypeGFA SecondLargestPropertyUseType SecondLargestPropertyUseTypeGFA ThirdLargestPropertyUseType ThirdLargestPropertyUseTypeGFA
Relevés manuels	YearsENERGYSTARCertified ENERGYSTARScore SiteEUI(kBtu/sf) SiteEUIWN(kBtu/sf) SourceEUI(kBtu/sf) SourceEUIWN(kBtu/sf) **SiteEnergyUse(kBtu)** SiteEnergyUseWN(kBtu) SteamUse(kBtu) Electricity(kWh) Electricity(kBtu) NaturalGas(therms) NaturalGas(kBtu) **TotalGHGEmissions** GHGEmissionsIntensity
Autres données	DataYear DefaultData Comments ComplianceStatus Outlier

II. Préparation du jeu de données - Cleaning

- **Renommer** les variables identiques
- **Créer** de nouvelles variables
- **Supprimer** des variables

- Filtre sur le champ **PrimaryPropertyType** qui correspond à l'utilisation principale d'une propriété
- Eléments supprimés: High-Rise Multifamily, Low-Rise Multifamily, Mid-Rise Multifamily

- Les émissions de CO2 et la consommation d'énergie de bâtiments ne peuvent être égales ou inférieures à 0
- Le nombre de bâtiments ne peut être inférieur ou égal à 0
- La somme des sources d'énergie doit être proche de la valeur de la variable SiteEnergyUse(kBtu)
- la surface des bâtiments + parkings (PropertyGFATotal) est bien égale à la surface des bâtiments (PropertyGFABuilding(s)) et du parking (PropertyGFAParking)
- la surface des bâtiments la plus importante (LargestPropertyUseTypeGFA) ne peut être supérieure à la surface des bâtiments + parkings (PropertyGFATotal)

Fusion des 2 datasets

Harmonisation des modalités des variables catégorielles

Filtre sur les bâtiments non destinés à l'habitation

Gestion des doublons

Gestion des valeurs aberrantes

Gestion des valeurs manquantes

Suppression des variables provenant des relevés manuels + trop corrélées aux targets (data leak)

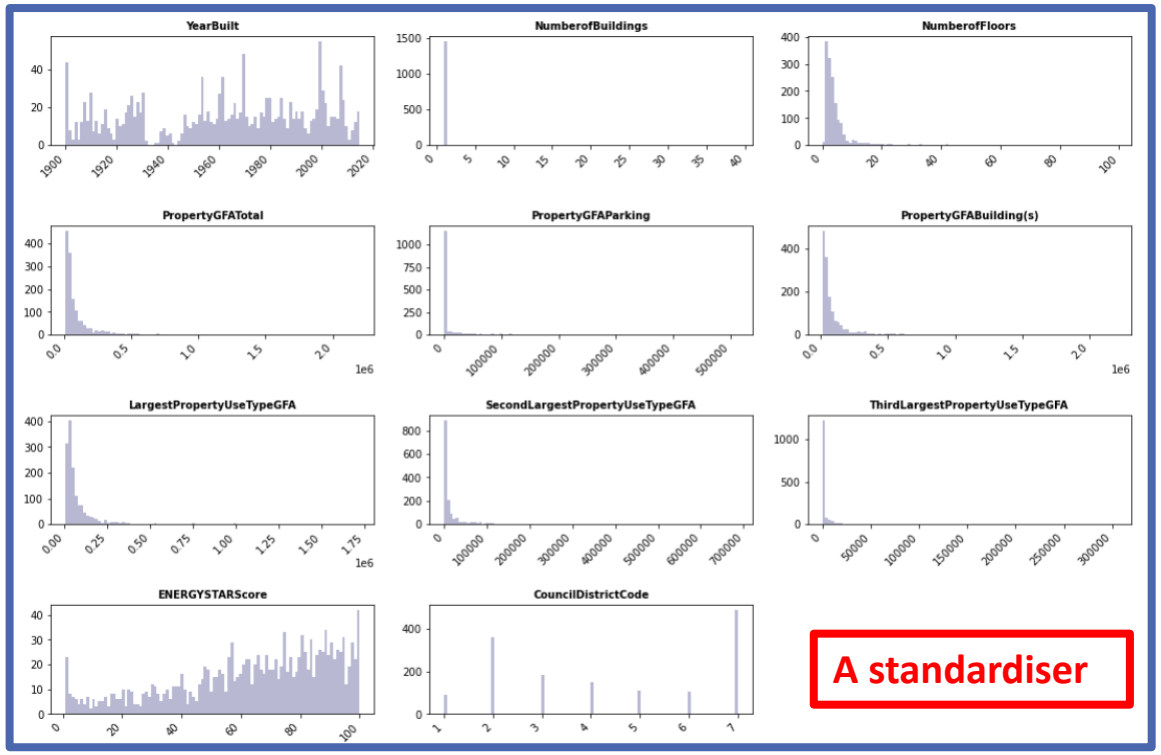
```
***BuildingType***
Nonresidential          2921
Multifamily Lr (1-4)    2047
Multifamily Mr (5-9)    1134
Multifamily Hr (10+)     217
Sps-District K-12        197
Nonresidential Cos       153
Campus                    46
Nonresidential Wa          1
```

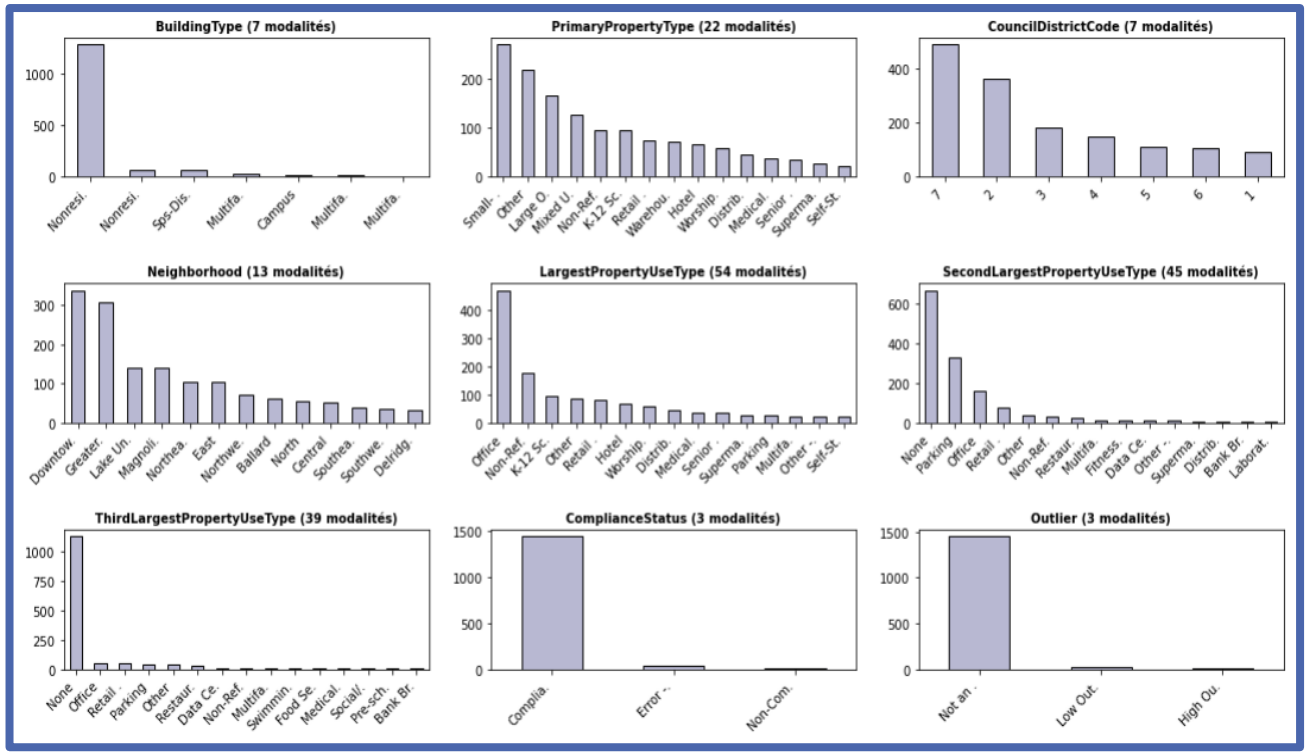
- Suppression des observations ayant le moins de valeurs manquantes lorsque les **OSEBuildingID** sont identiques
- **Enrichissement des données** si possible

- **Suppression** des variables DefaultData et ZipCode
- Constante '**Not an Outlier**" pour la variable Outlier
- Constante '**None**' pour les variables sur le type des bâtiments
- **0** pour les variables superficie associées
- **K Nearest Neighboors** pour la variable ENERGYSTARScore

II. Préparation du jeu de données – Exploration

Variables indépendantes





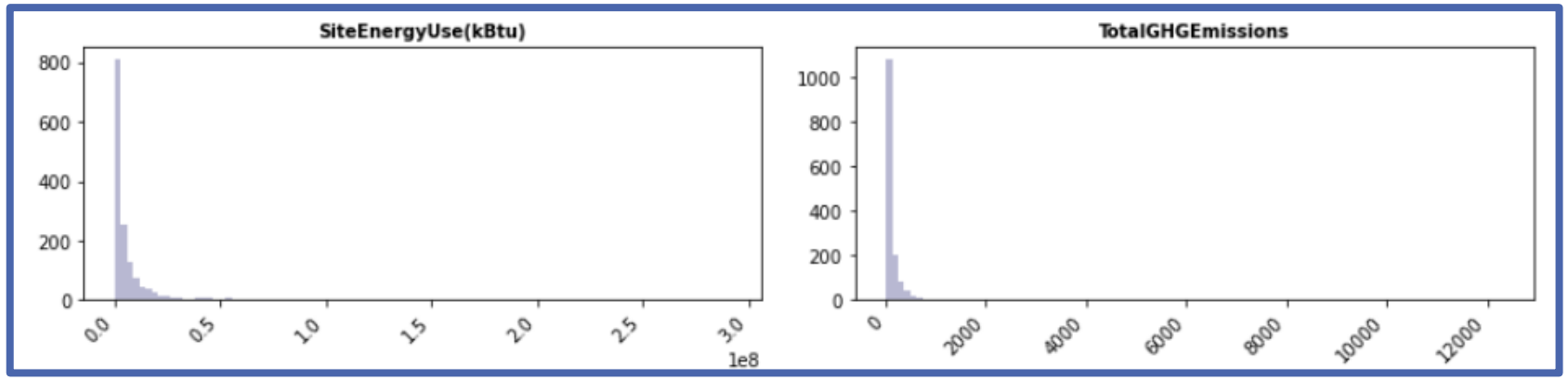
Le test de Shapiro_Wilk rejette l'hypothèse nulle selon laquelle les échantillons suivent une distribution Normale.
L'hypothèse nulle n'est pas rejetée pour la variable ENERGYSTARScore.

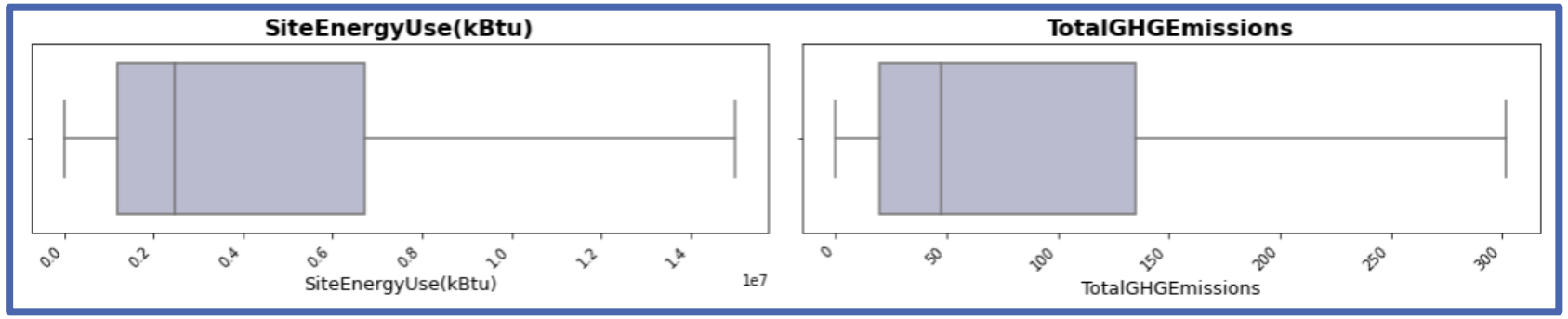
Certaines variables telles que le PrimaryPropertyType, ou celles sur l'usage du bâtiment ont une **cardinalité forte** (>20 modalités). La plupart des **modalités** ne sont **pas équilibrées**.

II. Préparation du jeu de données – Exploration

Variables cibles

A passer au log





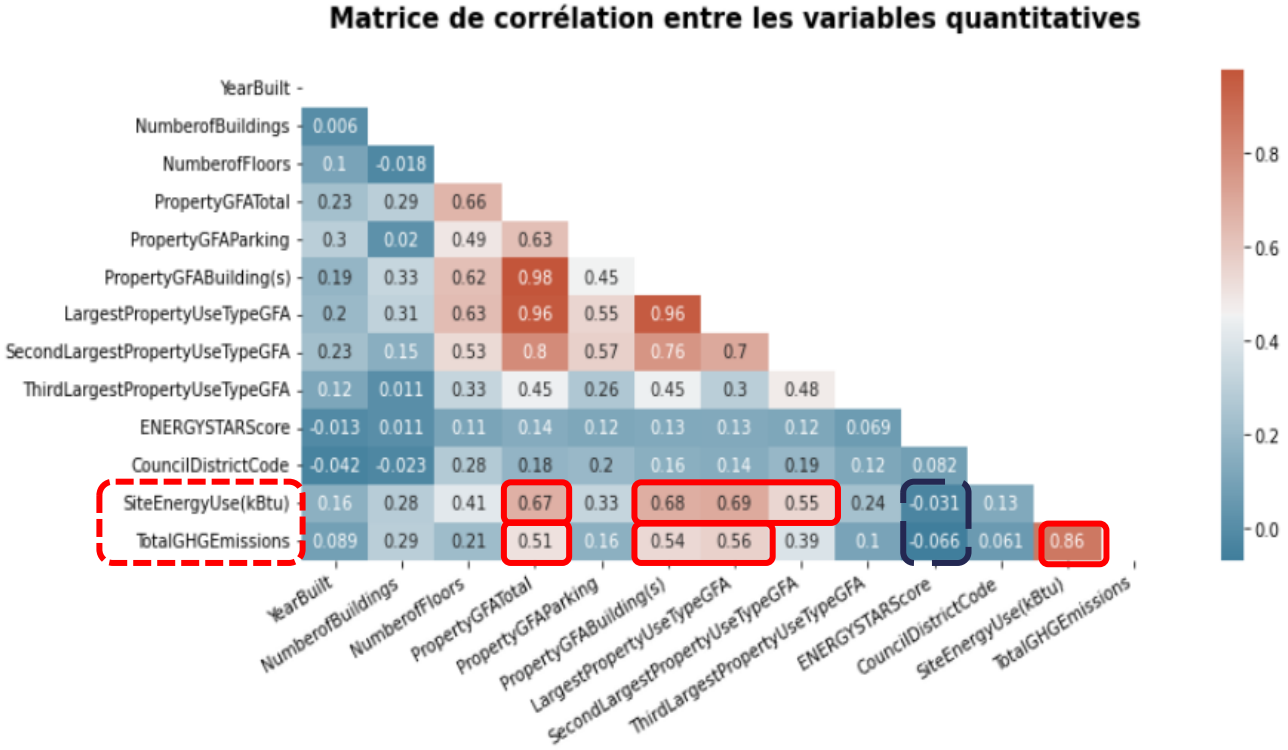

```
Test de Shapiro-Wilk:

Variable SiteEnergyUse(kBtu): On rejette H0, distribution non Normale (pvalue = 0.0 < 0.05)
Variable TotalGHGEmissions: On rejette H0, distribution non Normale (pvalue = 0.0 < 0.05)
```

Ces deux distributions ont une apparence similaire avec de nombreuses observations sur les valeurs les plus faibles.
Lorsque les distributions sont regroupées sur les valeurs les plus faibles et peu étalées la transformation log est le plus souvent utilisée.

Le test de Shapiro_Wilk rejette l'hypothèse nulle selon laquelle les variables target suivent une distribution Normale.

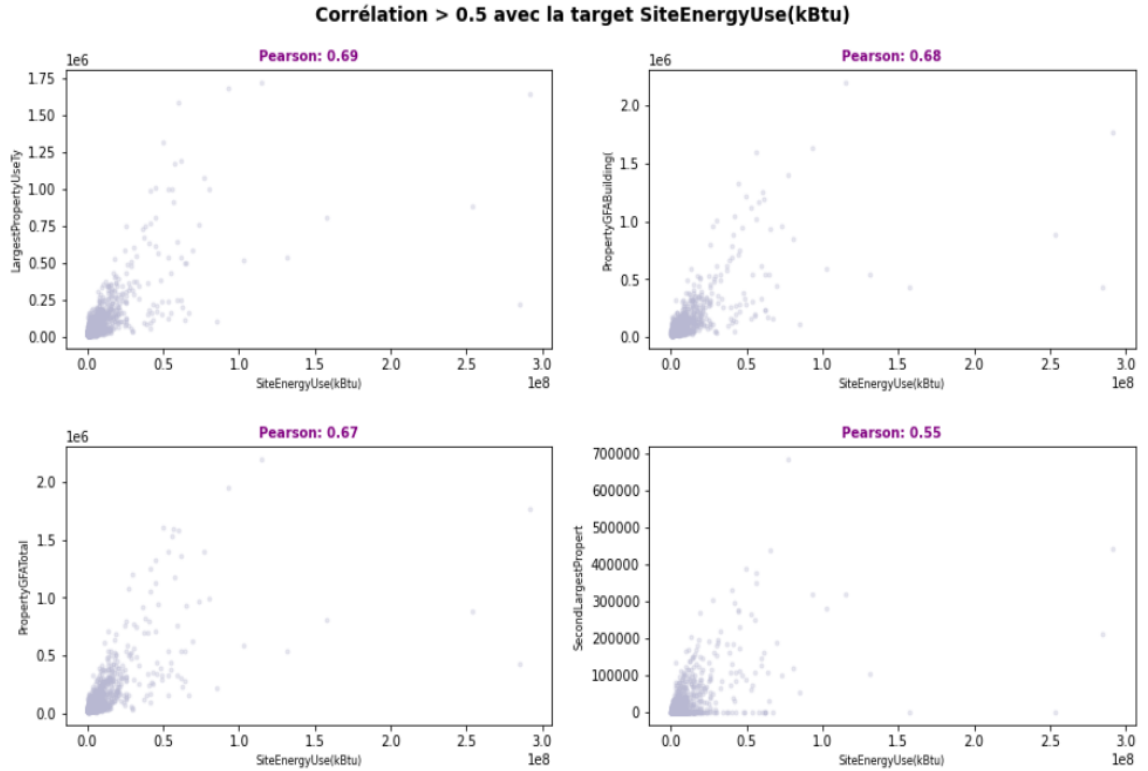
II. Préparation du jeu de données – Exploration

Corrélation avec les targets



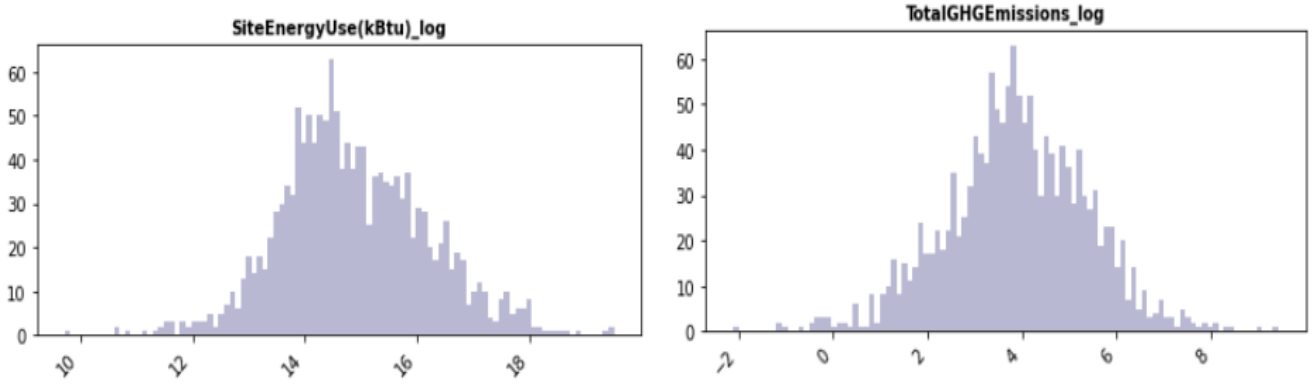
Les features les plus corrélées entre nos deux targets sont les mêmes. L'EnergyStarScore n'a pas vraiment de corrélation avec nos variables targets. Nous testerons nos modèles avec et sans cette variable pour vérifier son intérêt sachant qu'elle est fastidieuse à calculer.

Afin que le modèle soit plus robuste et bien que pas forcément aberrantes, certaines valeurs atypiques pourraient diminuer la capacité de nos modèles à généraliser. Nous ne conserverons que les observations pour lesquelles la target SiteEnergyUse(kBtu) est inférieure à 125 000 000 kBu.

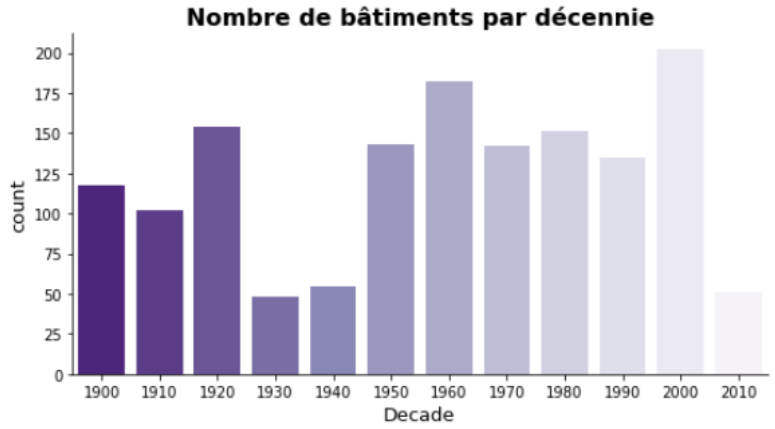


II. Préparation du jeu de données - Feature Engineering

✓ **Transformation des targets en log**



✓ **Regroupement des YearBuilt en décennies pour pallier à certaines années sous-représentées**



✓ **Proportion de chaque type d'énergie utilisé par tranche** **Non utilisé – connaissance métier nécessaire**

	%Electricity_Used	%Gas_Used	%Steam_Used
0	medium	low	low
2	medium	medium	low
4	high	low	low
7	medium	low	low
8	medium	medium	low

✓ **Nombre de bâtiments**

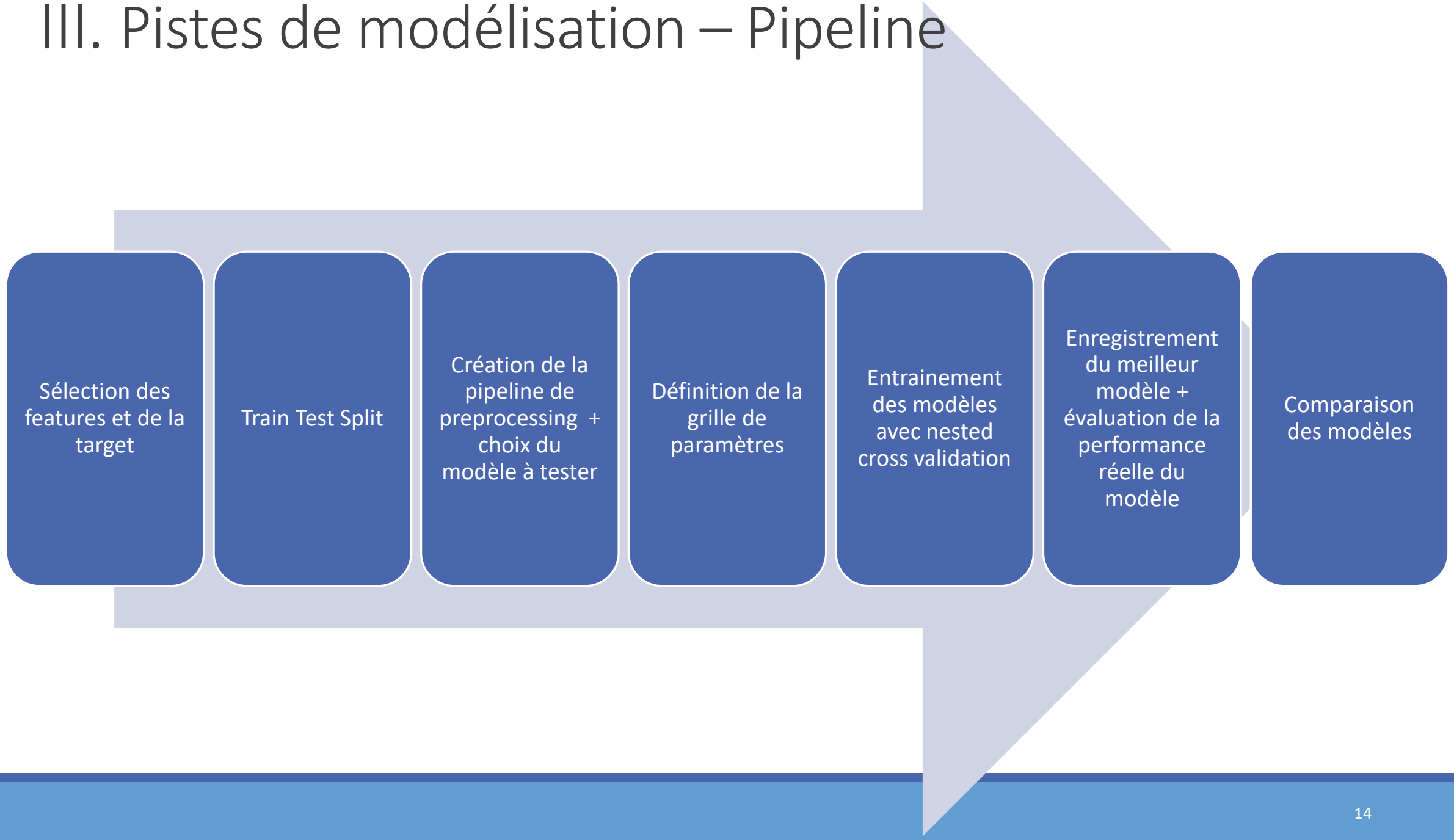
one building	1458
more than one building	25

✓ **Nombre d'étages**

(-0.001, 1.0]	396
(1.0, 2.0]	323
(5.0, 99.0]	266
(2.0, 3.0]	250
(3.0, 5.0]	248

III. PISTES DE MODELISATION

III. Pistes de modélisation – Pipeline



III. Pistes de modélisation – Choix des features

Les variables suivantes seront utilisées dans les modélisations:

- **Données déclaratives du permis d'exploitation commerciale corrélées fortement aux targets en ne prenant pas celles qui sont corrélées entre elles à plus de 0,7:** Decade, LargestPropertyUseTypeGFA
- **Emplacement du bâtiment:** CouncilDistrictCode, Neighborhood
- **Type des bâtiments:** BuildingType, PrimaryPropertyType
- **Usage des bâtiments:** LargestPropertyUseType
- **Construction des bâtiments:** NumberofBuildings, NumberofFloors

Test des modèles avec et sans la variable ENERGYSTARScore

Les variables suivantes seront écartées car non exploitables, fastidieuses à récupérer ou présentant un risque de data leak (liées directement aux variables cibles):

- **Identification des bâtiments:** OSEBuildingID, PropertyName, TaxParcelIdentificationNumber
- **Localisation:** Address, City, State, ZipCode, Latitude, Longitude
- **Relevés manuels:** YearsENERGYSTARCertified, SiteEUI(kBtu/sf), SiteEUIWN(kBtu/sf), SourceEUI(kBtu/sf), SourceEUIWN(kBtu/sf), SiteEnergyUseWN(kBtu), SteamUse(kBtu), Electricity(kWh), Electricity(kBtu), NaturalGas(therms), NaturalGas(kBtu), GHGEmissionsIntensity
- **Autres données:** DataYear, DefaultData, Comments, ComplianceStatus, Outlier
- **Données corrélées entre elles à plus de 0,7:** PropertyGFATotal, PropertyGFAParking, PropertyGFABuilding(s), SecondLargestPropertyUseTypeGFA, ThirdLargestPropertyUseTypeGFA

III. Pistes de modélisation – Modèles testés

- **Apprentissage supervisé (données étiquetées)**
- **Problème de régression (targets numériques)**

Modèles testés:

1) Dummy Regressor
2) Régression Linéaire
3) Régression Ridge
4) SVR
5) Random Forest
6) XGBoost

Modèle baseline

Modèles linéaires

Modèle bagging

Modèle boosting

Critères de performance:

- MAE

$$MAE = \frac{1}{m}\sum|y_{vrai} - y_{pred}|$$

- Temps d'entrainement

III. Pistes de modélisation – Grille de paramètres

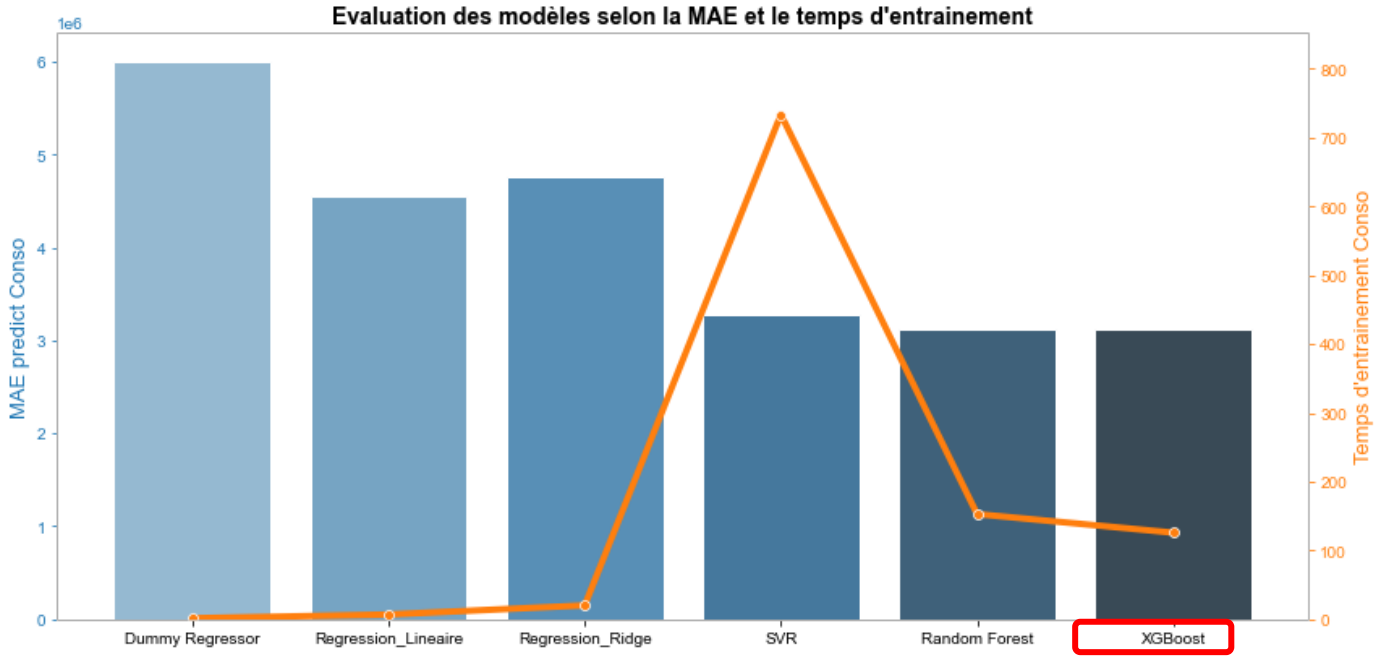
Hyperparameters tuning:

Régression Ridge	SVR	Random Forest	XGBoost
GridSearchCV	GridSearchCV	RandomizedSearchCV	RandomizedSearchCV
alpha: np.logspace(-2, 2, num=10)	C : [1, 10, 100, 1000]	n_estimators : [1, 2, 5, 10, 20, 50, 100, 200, 500]	learning_rate : [0.001, 0.01, 0.1, 0.2, 0,3]
	epsilon : [0, 0.5, 1]	max_depth : [5, 15, 25, 50]	gamma: [0, 0.25, 0.5, 1.0]
	degree : [2,3,4,5]	max_leaf_nodes : [2, 5, 10, 20, 50, 100]	max_depth: [6, 10, 15, 20]
	gamma: [0.0001, 0,001, 0.1]		min_child_weight : [0.5, 1.0, 3.0, 5.0, 7.0, 10.0]
			n_estimators: [25, 50, 100, 500, 1000]

III. Pistes de modélisation – Résultats obtenus

Consommation d'énergie:

- 3 modèles se distinguent par leur performance: **SVR**, **RandomForest**, **XGBoost**
- Le **SVR** sera **écarté**: temps d'entrainement beaucoup plus long pour une moindre performance
- Nous retiendrons le **XGBoost** qui a une performance à peu près équivalente au Random Forest avec un temps d'entrainement un peu plus faible



Modèle	MAE predict Conso	Temps d'entrainement Conso
Dummy Regressor	6,005,540.69	1.39
Regression_Lineaire	4,557,802.61	6.86
Regression_Ridge	4,768,445.72	20.14
SVR	3,279,287.97	733.65
Random Forest	3,116,314.88	152.88
XGBoost	3,117,841.17	125.81

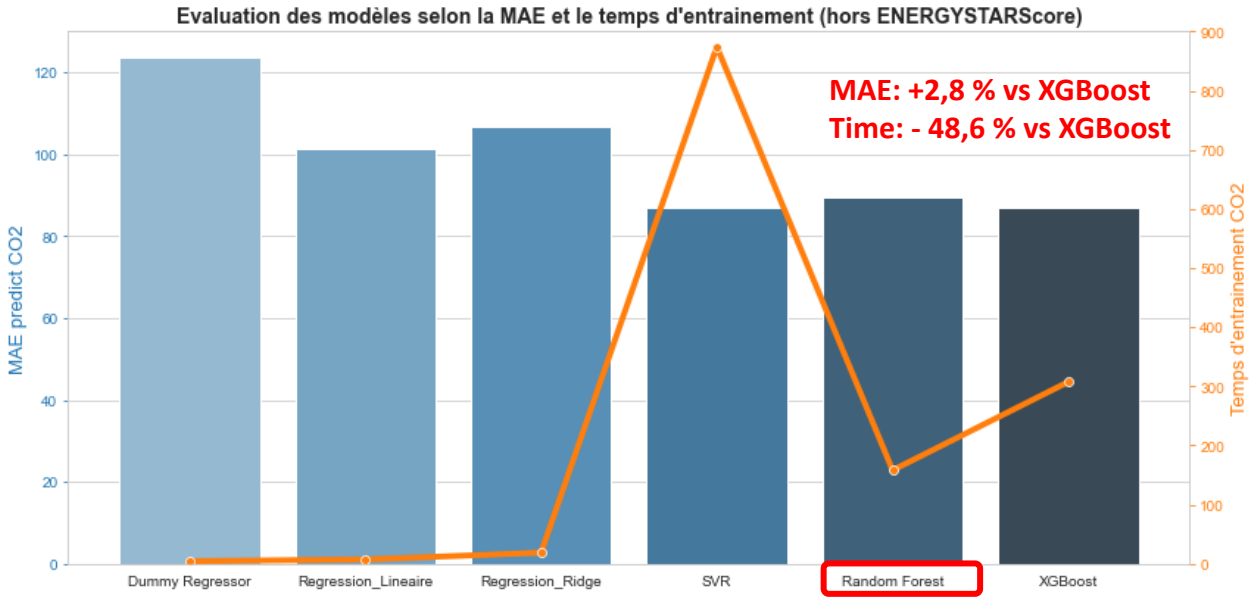
III. Pistes de modélisation – Résultats obtenus

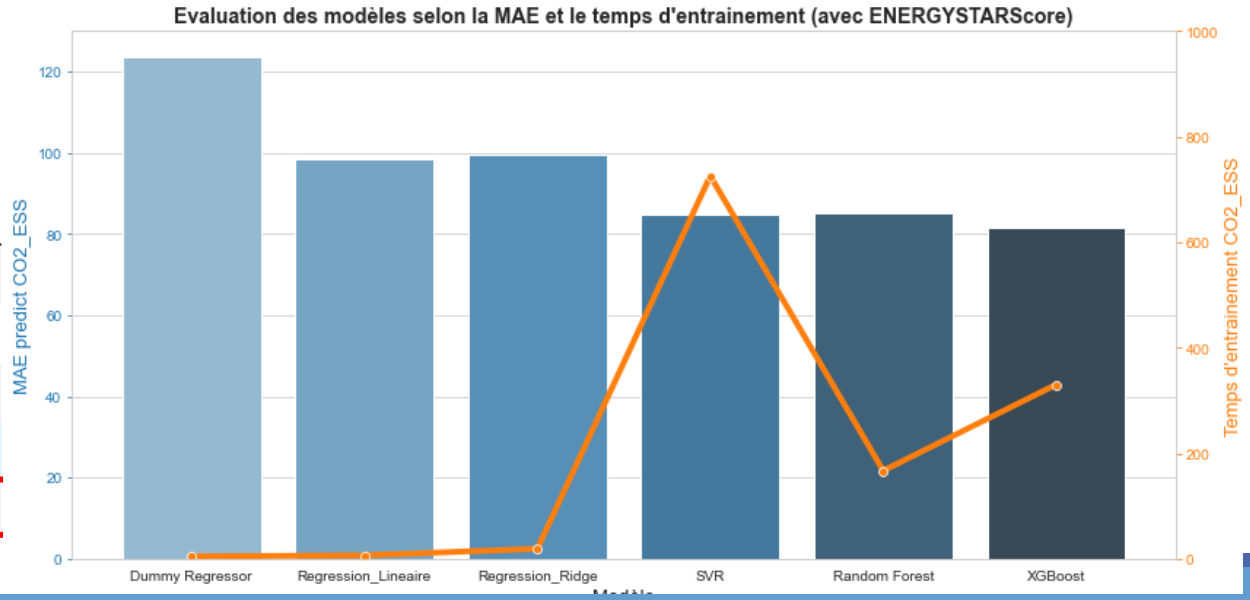
Emissions de CO2 :

- 3 modèles se distinguent par leur performance quasi similaire: **SVR**, **Random Forest** et **XGBoost**
- Le **SVR** sera **écarté**: temps d'entrainement beaucoup trop long
- Le **XGBoost** serait choisi en termes de **performances** mais le **Random Forest** a un **temps d'entrainement plus faible**

- L'ajout de la variable **EnergyStarScore améliore sensiblement la performance de nos modèles**
- Un arbitrage doit être réalisé car l'EnergyStarScore est **complexe** et **fastidieux** à calculer. De plus, nous avons complété les valeurs manquantes pour bénéficier de plus d'observations (bruit suppl.)

Modèle	MAE predict CO2	Temps d'entrainement CO2	MAE predict CO2_ESS	Temps d'entrainement CO2_ESS
Dummy Regressor	123.76	4.41	123.76	4.41
Regression_Lineaire	101.46	7.28	98.57	6.21
Regression_Ridge	106.64	18.75	99.70	18.92
SVR	86.91	874.04	84.91	725.56
Random Forest	89.54	158.18	85.25	166.85
XGBoost	87.06	307.54	81.38	330.04





IV. MODELE FINAL

IV. Modèle final – Consommation d'énergie - XGBoost

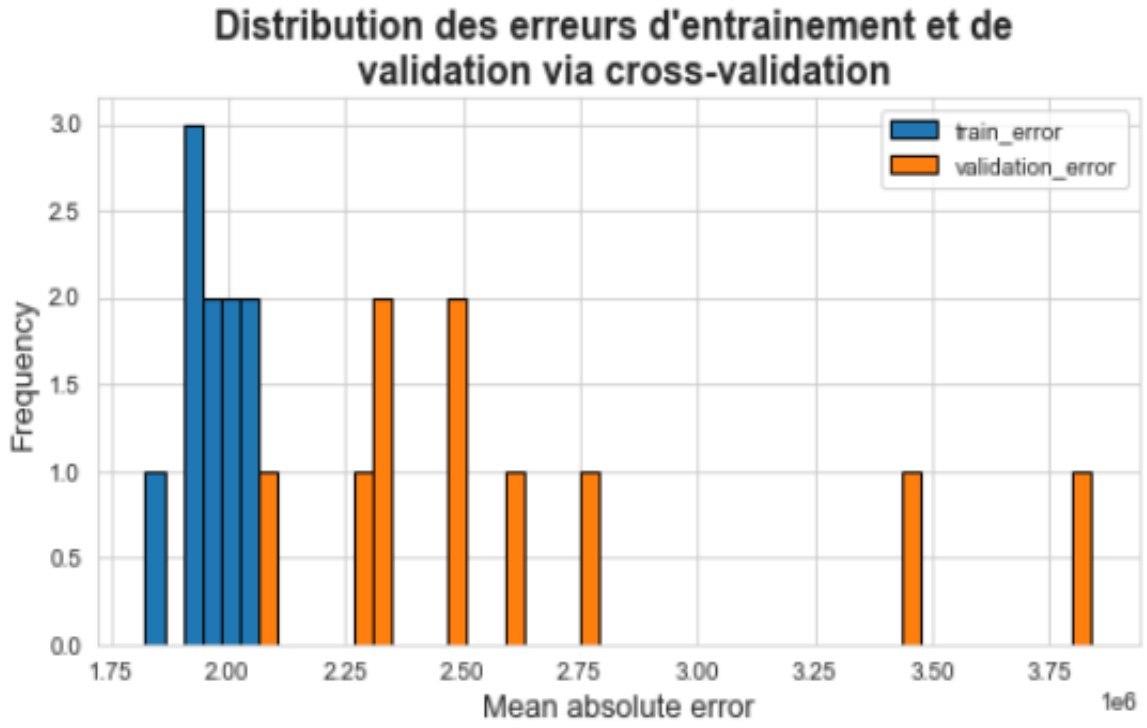
MAE: 2 991 878 kBtu

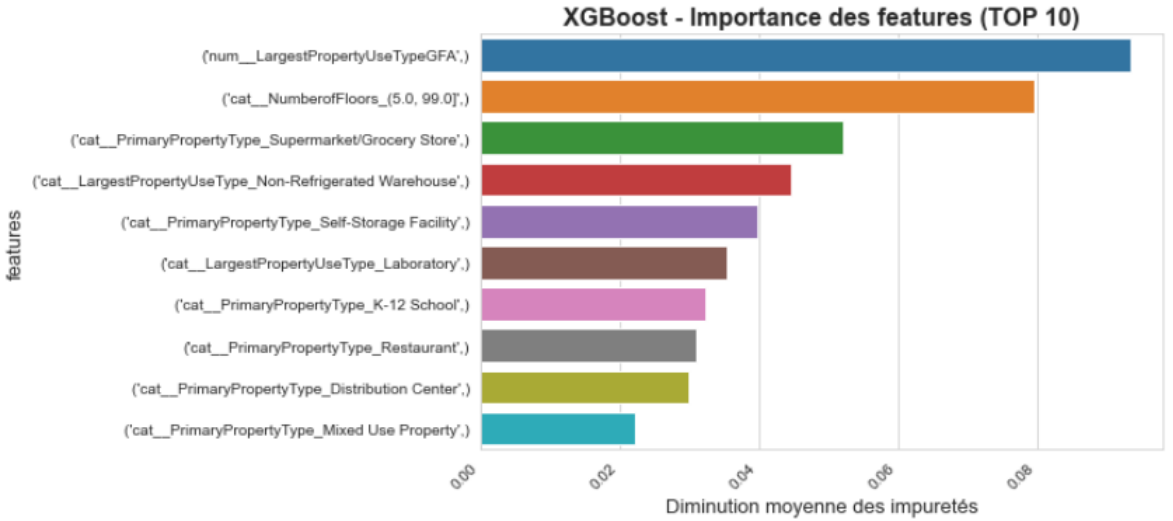
- **Preprocessing des features**:
  - One Hot Encoder
  - StandardScaler

- **Paramétrage des hyperparamètres avec nested cross-validation et RandomizedSearchCV:**
  - **Gamma**: régule la profondeur des arbres (+ élevé, arbres moins profonds)
  - **Learning_rate**: chaque nouvel arbre est multiplié par le taux d'apprentissage (plus taux faible, plus convergence vers un optimal lente et vice versa)
  - **Max_depth**: profondeur maximale d'un arbre
  - **Min_Child_Weight**: seuil min du nombre d'individus présents dans un nœud
  - **N_estimators**: nombre d'arbres

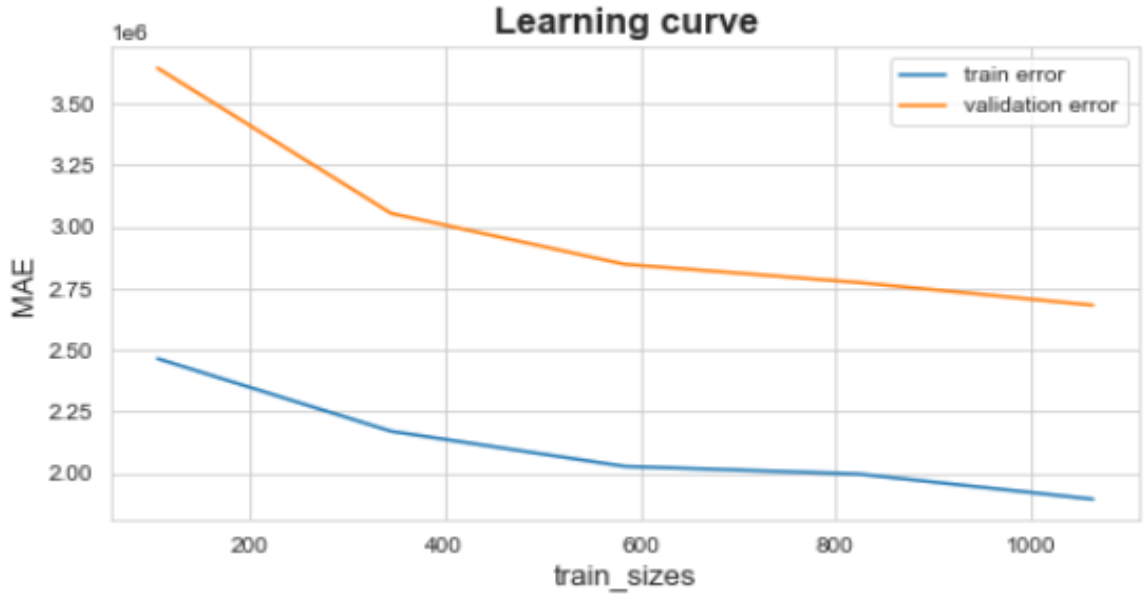
- **GridSearchCV** sur différents paliers de sélection de variables avec les meilleurs hyperparamètres trouvés à l'étape précédente => modèle plus performant avec toutes les features

IV. Modèle final – Consommation d'énergie - XGBoost

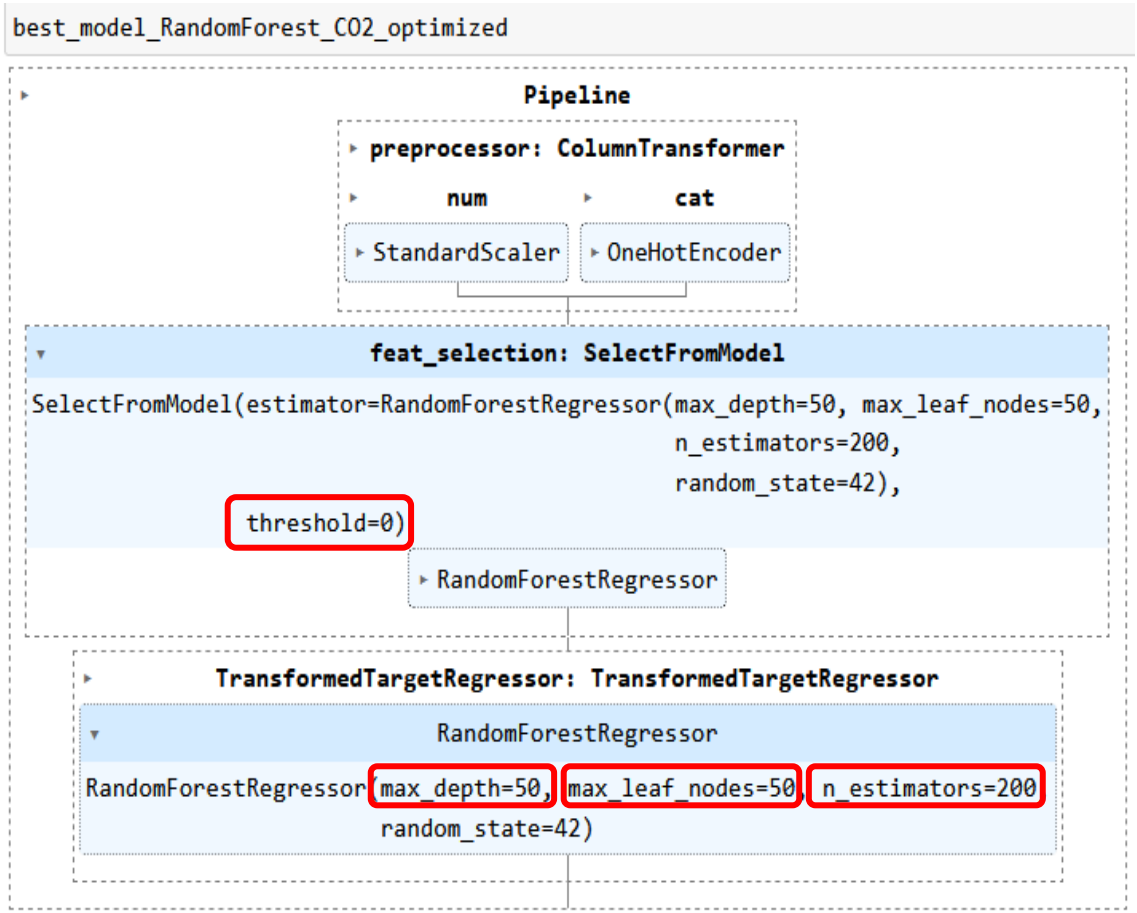





```
MAE train set: 1964081.8848 +/- 74040.5311
MAE validation set: 2671253.8975 +/- 553180.8358
Temps d'entrainement: 368.13
```



IV. Modèle final – Emissions de CO2 – Random Forest



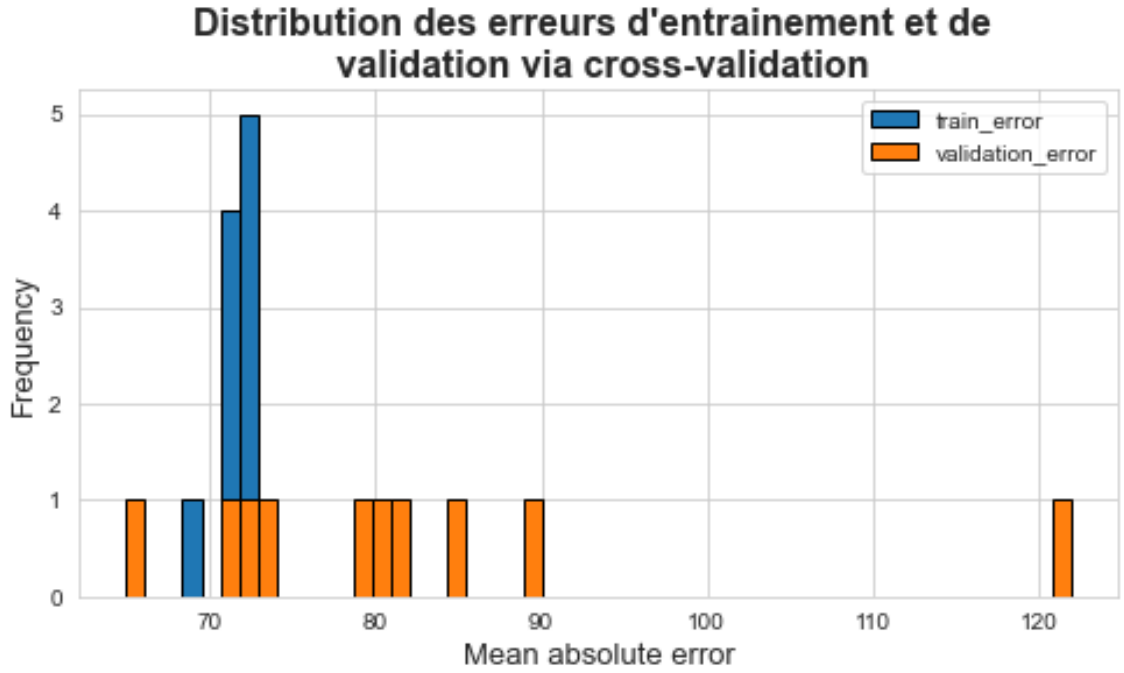
MAE: 90,64

- **Preprocessing des features**:
  - One Hot Encoder
  - StandardScaler

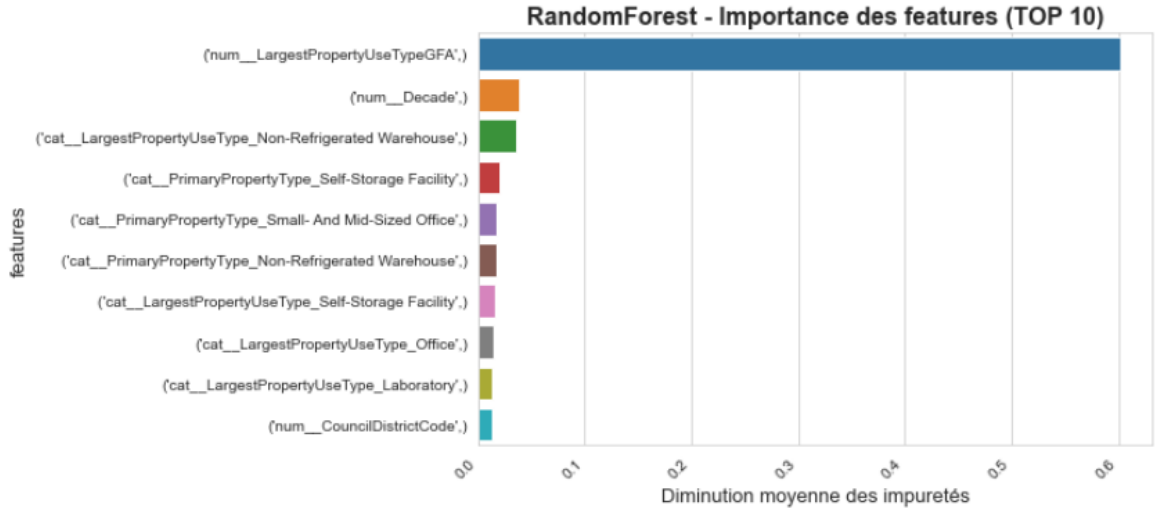
- **Paramétrage des hyperparamètres avec nested cross-validation et RandomizedSearchCV:**
  - **Max_depth**: profondeur maximale d'un arbre (arbre plus symétrique)
  - **Max_leaf_nodes**: seuil minimal quant au nombre d'individus présents dans un nœud
  - **N_estimators**: nb d'arbres. En général, plus il y a d'arbres dans la forêt, meilleure sera la performance de généralisation mais plus ce sera coûteux en temps et puissance de calcul. L'objectif est d'équilibrer le temps de calcul et les performances de généralisation lors de la définition du nombre d'estimateurs.

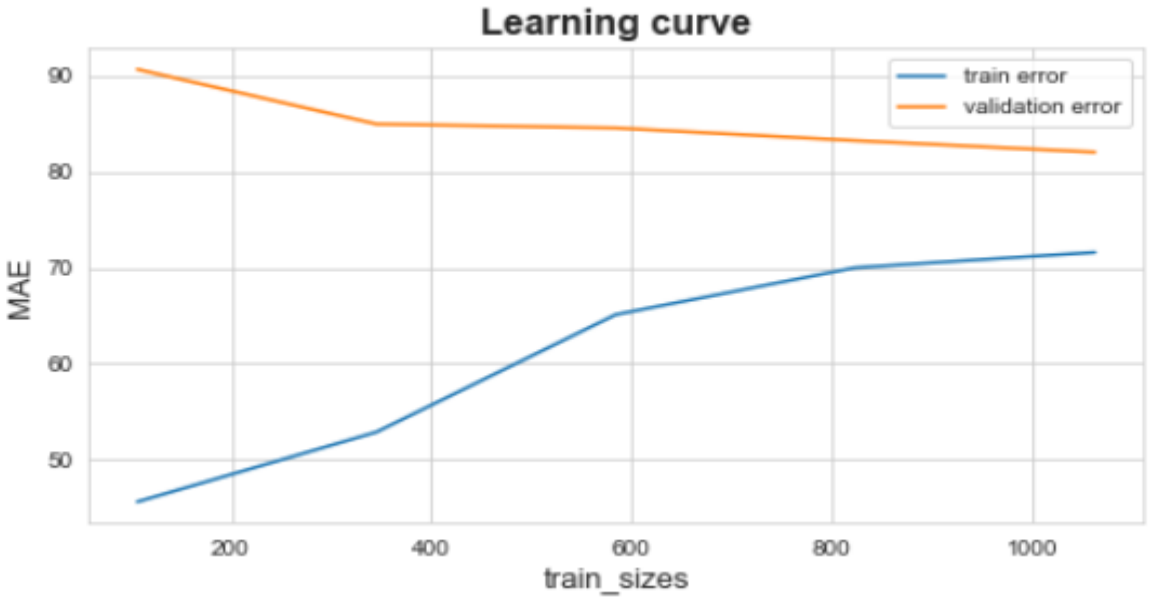
- **GridSearchCV** sur différents paliers de sélection de variables avec les meilleurs hyperparamètres trouvés à l'étape précédente => modèle plus performant avec toutes les features

IV. Modèle final – Emissions de CO2 – Random Forest



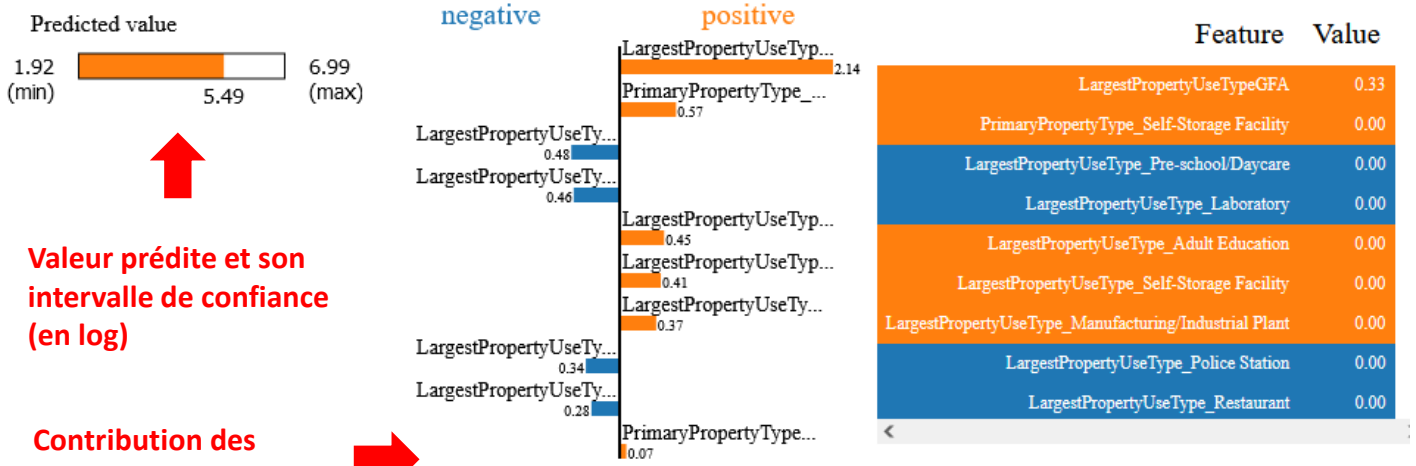

```
MAE train set: 71.5203 +/- 1.2971
MAE validation set: 81.9665 +/- 15.7410
Temps d'entrainement: 184.70
```





IV. Modèle final – Emissions de CO2 – Random Forest

Features importance locales avec LIME:



Valeur prédite et son intervalle de confiance (en log)

Contribution des features

Valeurs des features

```
explanation.as_list()
```

```
[('LargestPropertyUseTypeGFA > 0.02', 2.1388360159409245),
 ('PrimaryPropertyType_Self-Storage Facility <= 0.00', 0.5658906131185732),
 ('LargestPropertyUseType_Pre-school/Daycare <= 0.00', -0.483970520540503),
 ('LargestPropertyUseType_Laboratory <= 0.00', -0.46007027686281454),
 ('LargestPropertyUseType_Adult Education <= 0.00', 0.44689822174761473),
 ('LargestPropertyUseType_Self-Storage Facility <= 0.00', 0.4137120681875528),
 ('LargestPropertyUseType_Manufacturing/Industrial Plant <= 0.00',
  0.3662216117561406),
 ('LargestPropertyUseType_Police Station <= 0.00', -0.3376039352142964),
 ('LargestPropertyUseType_Restaurant <= 0.00', -0.28215457806491856),
 ('PrimaryPropertyType_Office <= 0.00', 0.06726677662536437)]
```

Local Interpretable Model-agnostic Explanations

- Lime prend en entrée n'importe quel modèle d'apprentissage automatique et génère des explications sur la contribution des caractéristiques à la prédiction.
- Il suppose qu'il s'agit d'un modèle boîte noire, ce qui signifie qu'il ne connaît pas le fonctionnement interne des modèles

V. CONCLUSIONS

V. CONCLUSIONS

- Les **émissions de CO2** sont **fortement corrélées** avec la **consommation d'énergie** et les **mêmes features** peuvent être utilisées dans les deux modèles

- Il serait intéressant d'avoir à notre disposition d'autres informations telles que le **type d'isolation**, si des travaux **d'amélioration énergétique** ont été faits, l'utilisation **d'énergies renouvelables** etc

- L'étude de la learning curve nous indique qu'avoir **plus d'observations** en entrée permettrait d'améliorer la performance de nos modèles

- La variable **EnergyStarScore** améliore sensiblement la performance de nos modèles alors qu'elle n'est quasiment pas corrélée aux targets. Un **arbitrage performance / coût** devrait être réalisé car cet indicateur est fastidieux à calculer

VI. QUESTIONS / REPONSES

MERCI