



PROJET 4

LA CONSOMMATION D'ÉNERGIE DE BÂTIMENTS

EVA BOOKJANS



Seattle

Objectif

Prédire

pour des bâtiments:

- **Consommation d'Énergie**
- **Émission de CO₂**

Les Données :

- déclarées pour le permis d'exploitation commerciale
- Taille
- Usage de Bâtiments
- Date de Construction
- Emplacement
- ...
- ENERGYSTAR Score

Le Jeu de Données

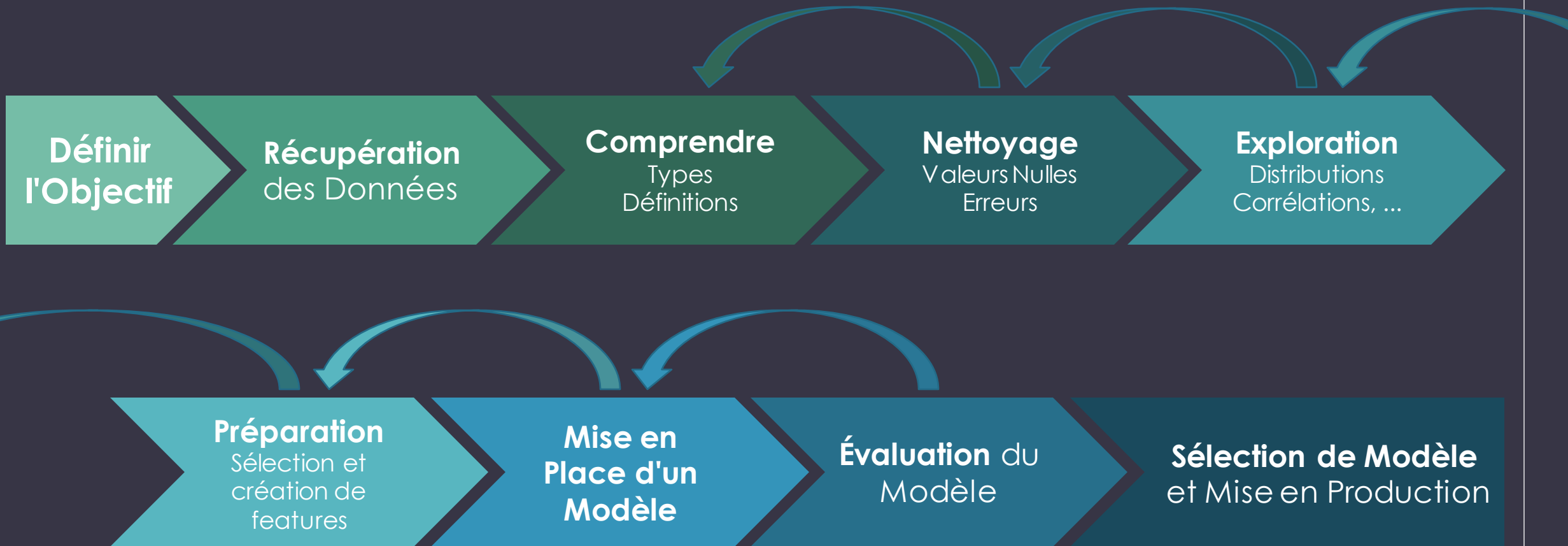


Seattle

- **Données collectées par la ville de Seattle** avec l'objectif à réduire la consommation d'énergie et les émissions de gaz à effet de serre des bâtiments existants.
- **Obligatoire pour les bâtiments avec plus de 20 000 Sf (1 858 m²)**
- Années: **2015 et 2016**
- Analyse pour les **bâtiments non-résidentielles**

Ce Projet

Chemin Général



Les Variables

Paramètres et d'Aide

Target

Consommation d'Énergie

- **En-soi**
- Les relevés
 - électricité, gaz, vapeur
 - ~~avec différents unités~~
- Normalisée
 - par la surface / le météo

Émissions de CO₂

- **Calculées des relevés**
- Normalisées
 - par la surface / le météo

Identificateur

- Numéro Id
- ~~Nom~~
- ~~Numéro d'impôt~~

Emplacement

- **Latitude, Longitude**
- ~~Adresse, Code postale~~
- Quartier
- ~~Districts...~~

Date de Construction

ENERGYSTAR Score

- Score
- ~~Années de certification~~

Taille

- Nombre de bâtiments
- Nombre d'étages
- **Surface** (Totale)
 - Parking + **Bâtiment**

Usage de Bâtiment

- Type de bâtiment
- **Usages de bâtiment**
- Premiers 3 types d'usage
 - Surfaces correspondantes

Conformité

- Conforme (ou pas)
- Données par défaut
- Valeurs aberrantes

Nettoyage

Détecter (et corriger si possible) des erreurs

- valeurs absurdes (e.g. une surface négative)
- comparaison des données des années 2015 et 2016

Imputer des valeurs nulles avec

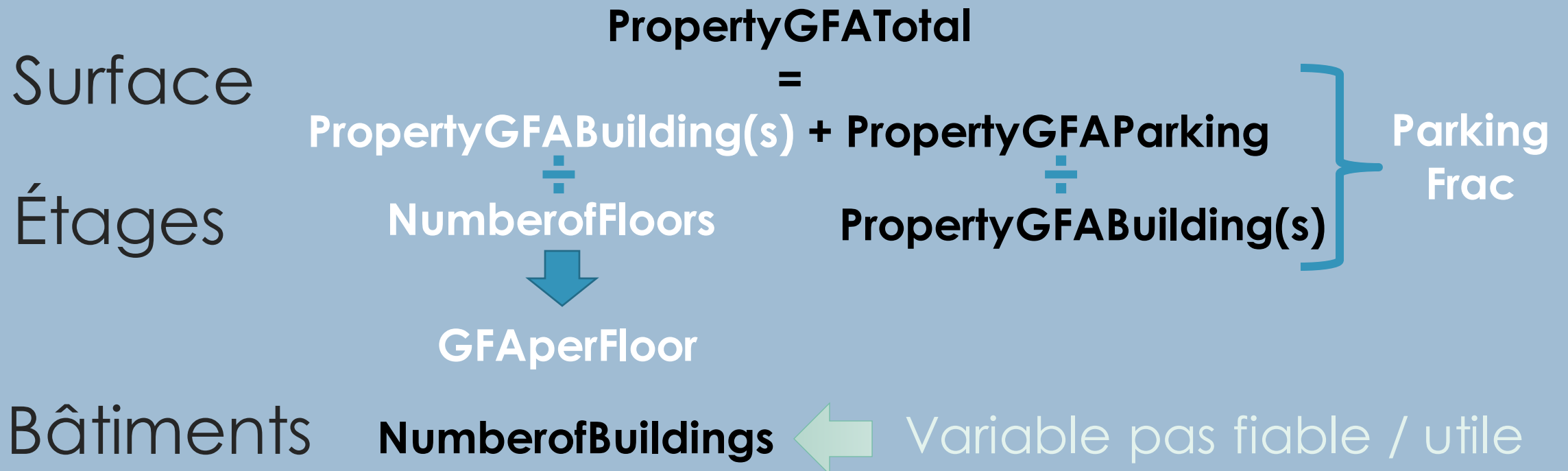
- la valeur de l'autre année
- une valeur par défaut ou une estimation raisonnable

Écarter des données

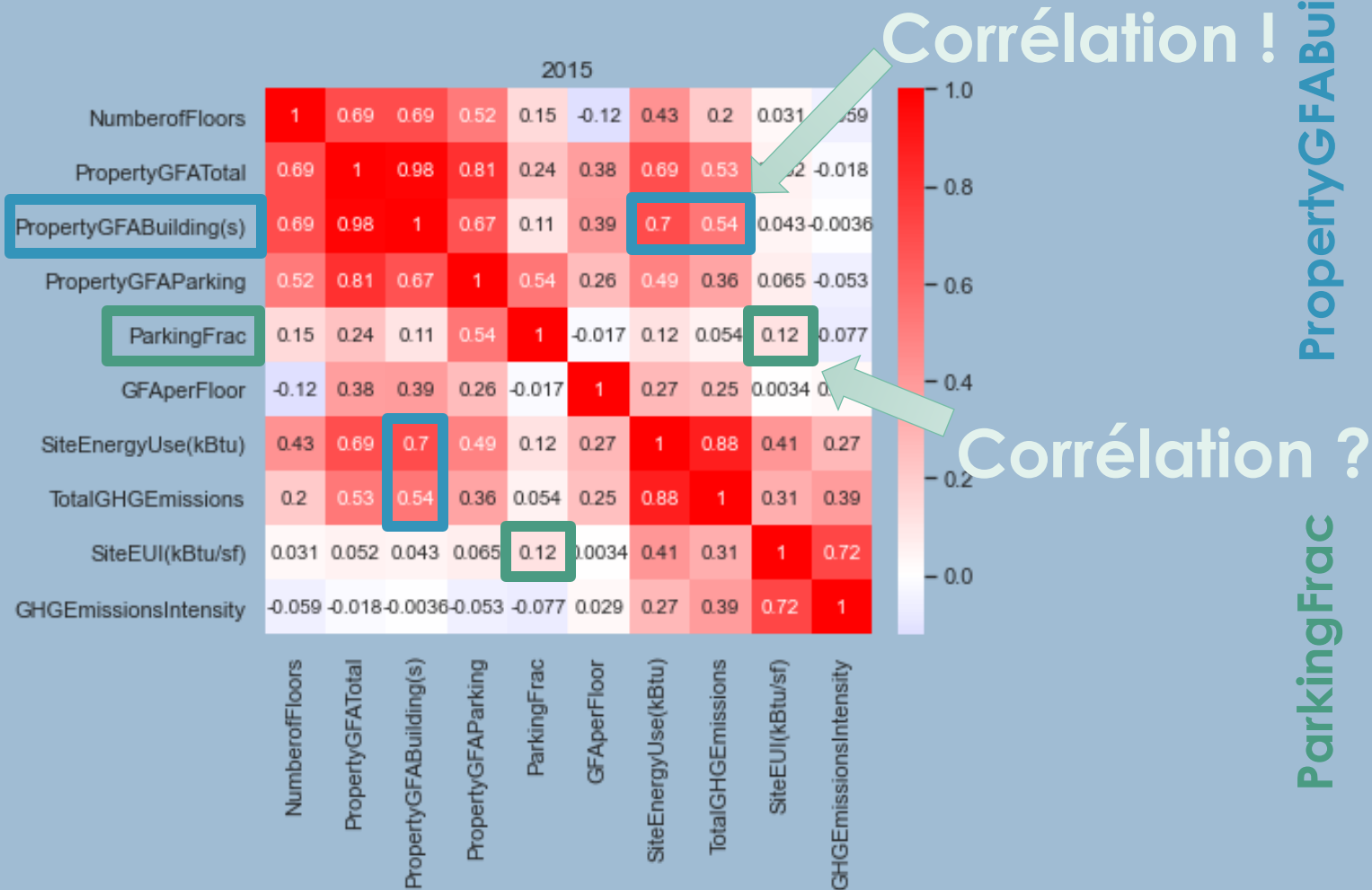
- non-conformes ou pas fiables
- sans rapport au projet (les bâtiment résidentiels)
- avec des variables targets nulles

Feature Engineering et Exploration

Taille du Bâtiment



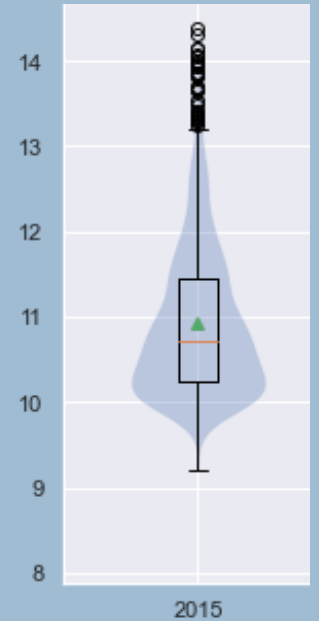
Taille du Bâtiment



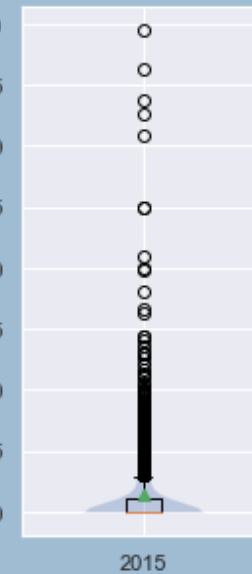
PropertyGFABuilding(s)



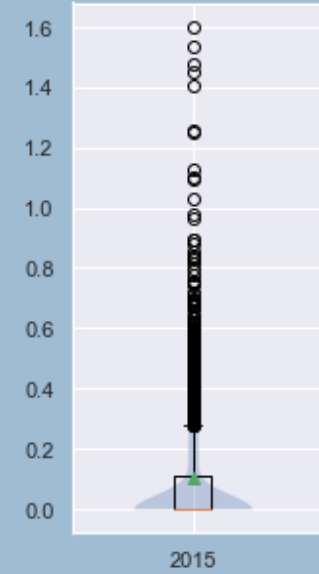
$\ln(x)$



ParkingFrac

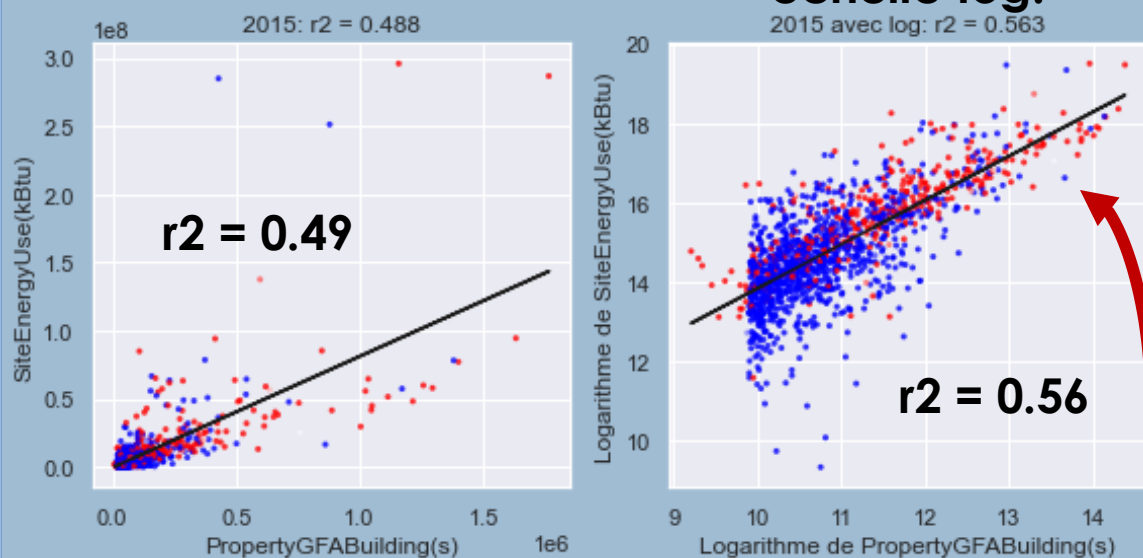


$\ln(1+x)$



Régression Linéaire - un simple modèle

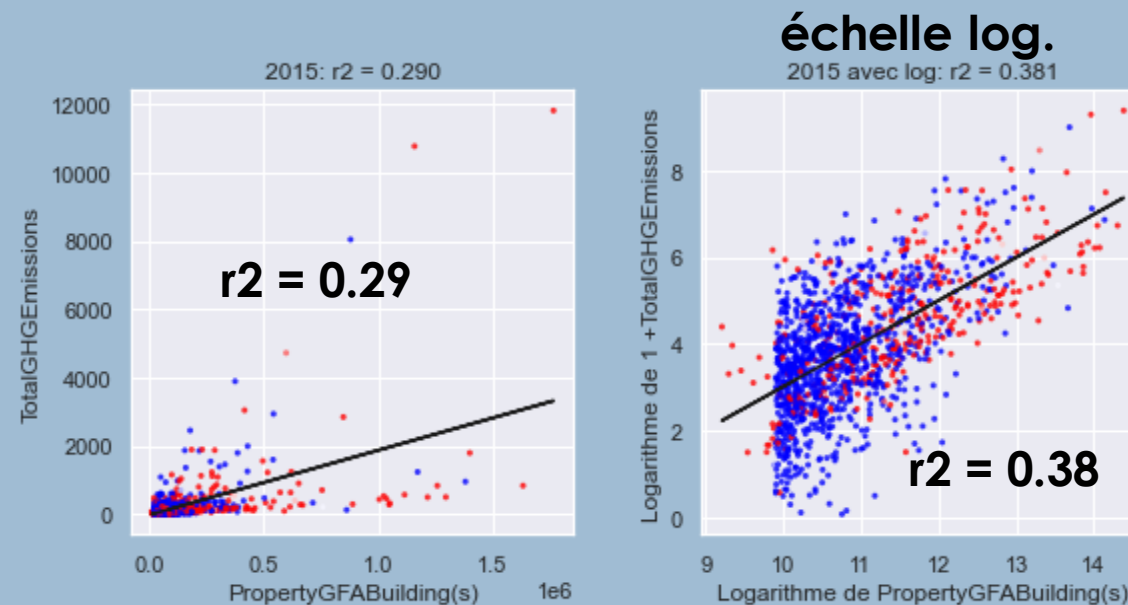
SiteEnergyUse(kBtu)



versus

PropertyGFABuilding(s)

TotalGHGEmissions



ParkingFrac > 0.1



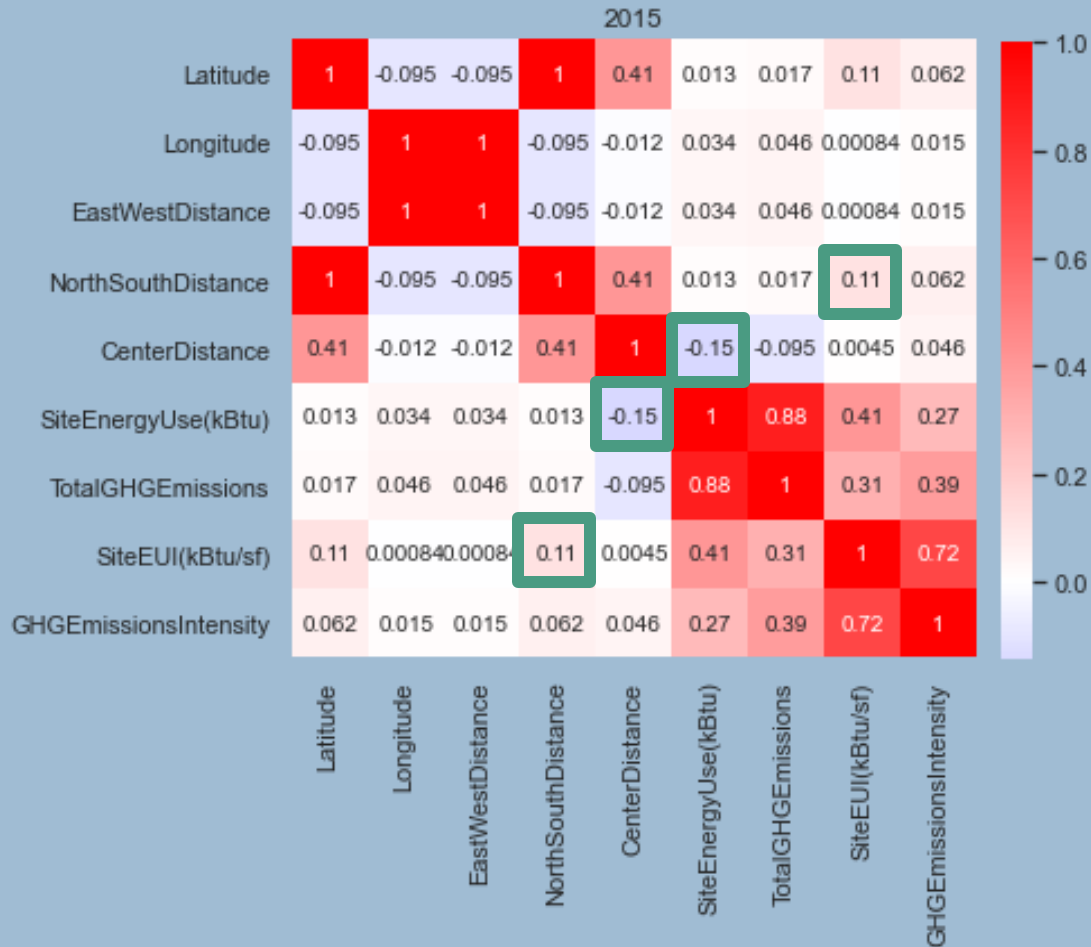
plus de points rouges au-dessous la ligne de régression

Emplacement

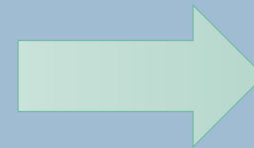
Latitude
Longitude

→
référéncé au
centre-ville

EastWestDistance
NorthSouthDistance
CenterDistance



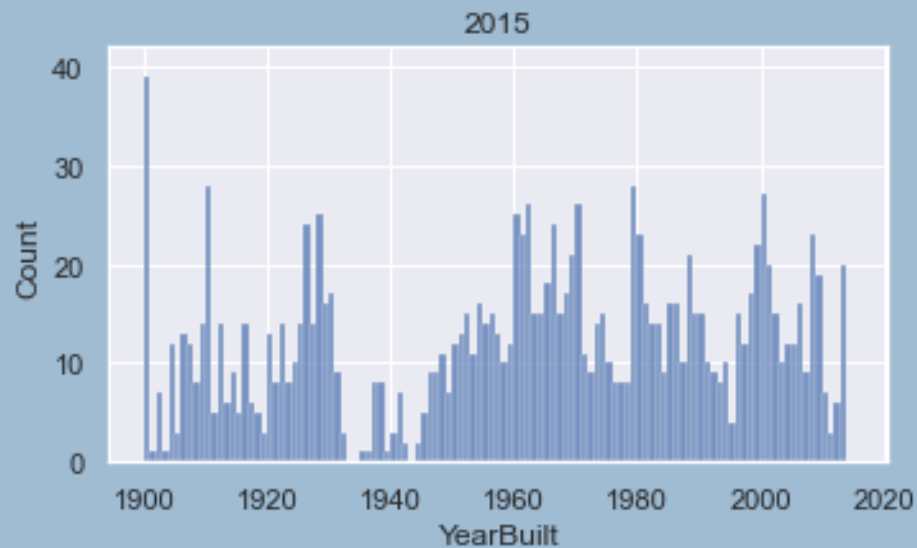
Neighborhood	Eta-carré
SiteEnergyUse(kBtu)	0.042
TotalGHGEmissions	0.033
SiteEUI(kBtu/sf)	0.041
GHGEmissionsIntensity	0.051



Faible Corrélation ?

Âge du Bâtiment

DataYear - YearBuilt ➡ BuildingAge




DecadeBuilt/BuildingAge	Eta-carré	Corrélation
SiteEnergyUse(kBtu)	0.042	-0.17
TotalGHGEmissions	0.033	-0.09
SiteEUI(kBtu/sf)	0.041	-0.15
GHGEmissionsIntensity	0.051	-0.02




Faible Corrélation ?


Type et Usage de Bâtiment



BuildingType : ~~Campus, MultifamilyHR (10+), MultifamilyLR (1-4), MultifamilyMR (5-9), NonResidential, Nonresidential COS, Nonresidential WA, SPS-District K-12~~

 **PublicBuilding** : 0 1

(A blue bracket groups the first four categories under '0' and the remaining four under '1')

PrimaryPropertyType : 20 catégories (sans 'Multifamily')  **19 One-Hot Variables**

ListofPropertyUseTypes :  **NumberofPropertyUseTypes**

- Contient 'Swimming Pool' ?  **HasSwimmingPool : Oui = 1, Non = 0**
- Contient 'Data Center' ?  **HasDataCenter : Oui = 1, Non = 0**

Type et Usage de Bâtiment

Eta-carré	Primary PropertyType	Largest PropertyUseType	SecondLargest PropertyUseType	ThirsLargest PropertyUseType
SiteEnergyUse(kBtu)	0.51	0.53	0.05	0.16
TotalGHGEmissions	0.55	0.55	0.03	0.14
SiteEUI(kBtu/sf)	0.49	0.55	0.07	0.09
GHGEmissionsIntensity	0.40	0.43	0.09	0.06

Corrélation !



NumberofPropertyUseTypes	Corrélation
SiteEnergyUse(kBtu)	0.04
TotalGHGEmissions	0.03
SiteEUI(kBtu/sf)	0.04
GHGEmissionsIntensity	0.05

Type et Usage de Bâtiment

LargestPropertyUseType
SecondLargestPropertyUseType
ThirdLargestPropertyUseType

68 catégories

grouper

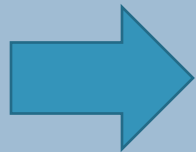
25 catégories (sans 'Parking')

LargestPropertyUseTypeGFA
SecondLargestPropertyUseTypeGFA
ThirdLargestPropertyUseTypeGFA

÷

\sum nLargestPropertyUseTypeGFA
si (nLargestPropertyUseType != Parking)

nLargestPropertyUseTypeGFAFrac



24

Frac-Hot Variables

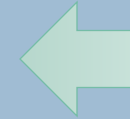
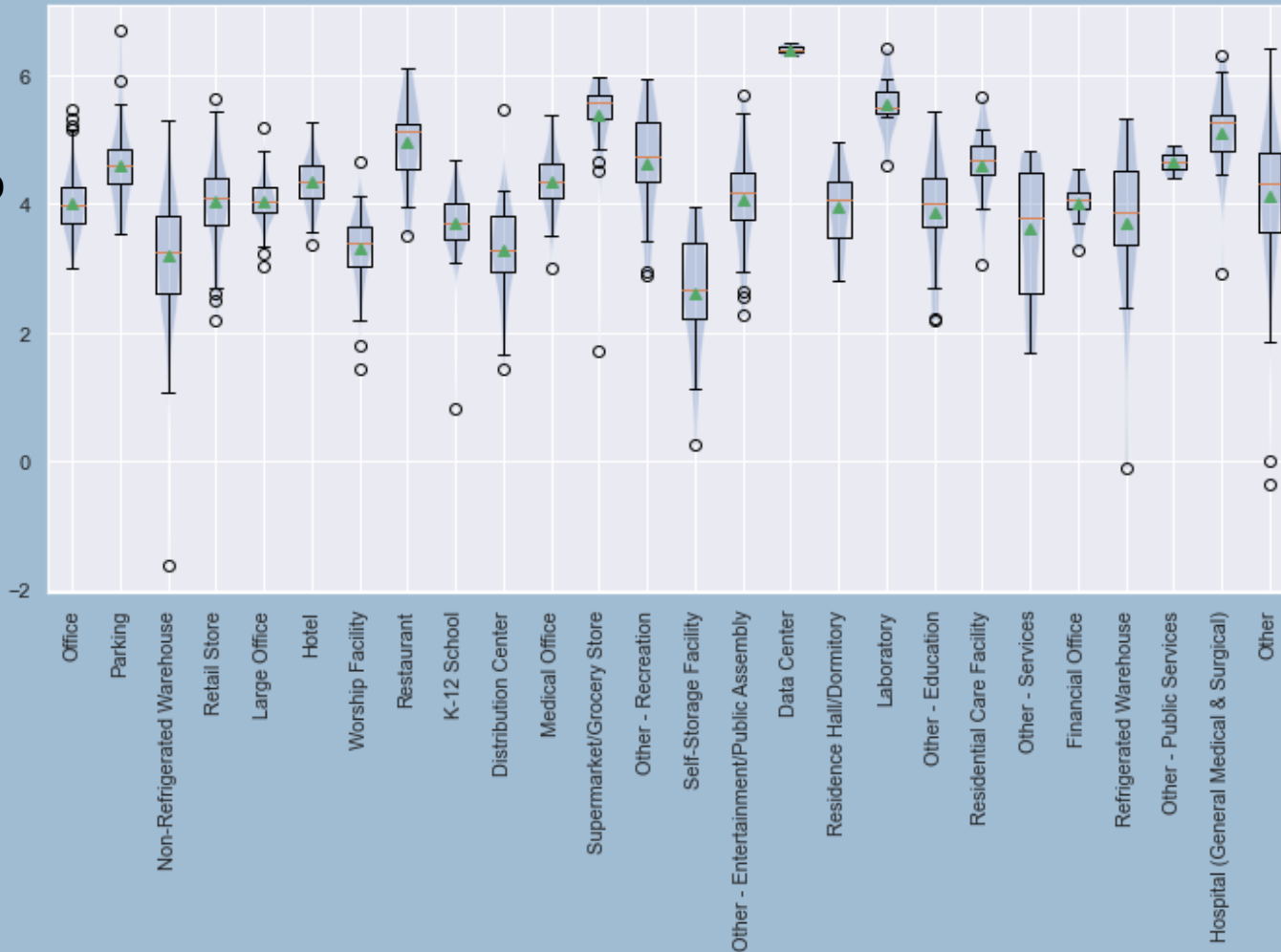
avec valeurs entre 0 et 1

indiquent la fraction de la surface du bâtiment utilisée pour chaque catégorie d'utilisation

Type et Usage de Bâtiment

SiteEUI(kBtu/sf)

échelle log.



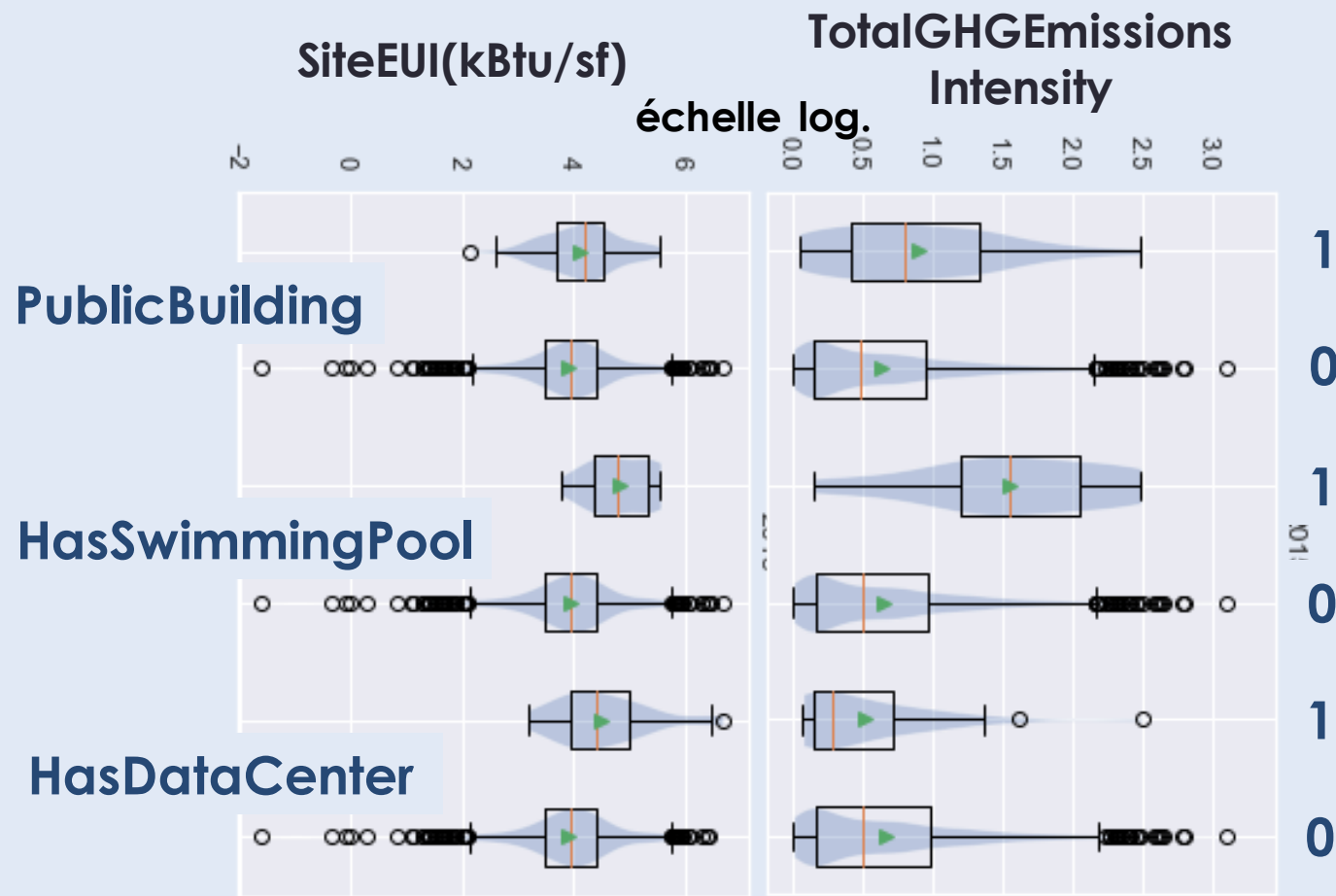
Catégorie
Différente

=

Distribution
Différente

Attention aux
petits échantillons !

Type et Usage de Bâtiment



	Public Building	Has Swimming Pool	Has Data Center
0	1310	1380	1361
1	86	16	35

Petits échantillons

Corrélation ?

Sources d'Énergie

Electricity(kBtu)
NaturalGas(kBtu)
Steam(kBtu)



les compteurs connectés sont connus,
mais pas les relevés en détail

if != 0 : 1

else : 0



UseofElectricity
UseofNaturalGas
UseofSteam



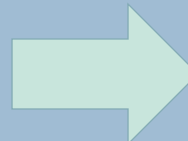
Tous les bâtiments
ont l'air d'utiliser
de l'électricité



4 catégories

SourcesEnergy	Eta-carré
SiteEnergyUse(kBtu)	0.058
TotalGHGEmissions	0.089
SiteEUI(kBtu/sf)	0.029
GHGEmissionsIntensity	0.150

Corrélation ?



Interdépendances !

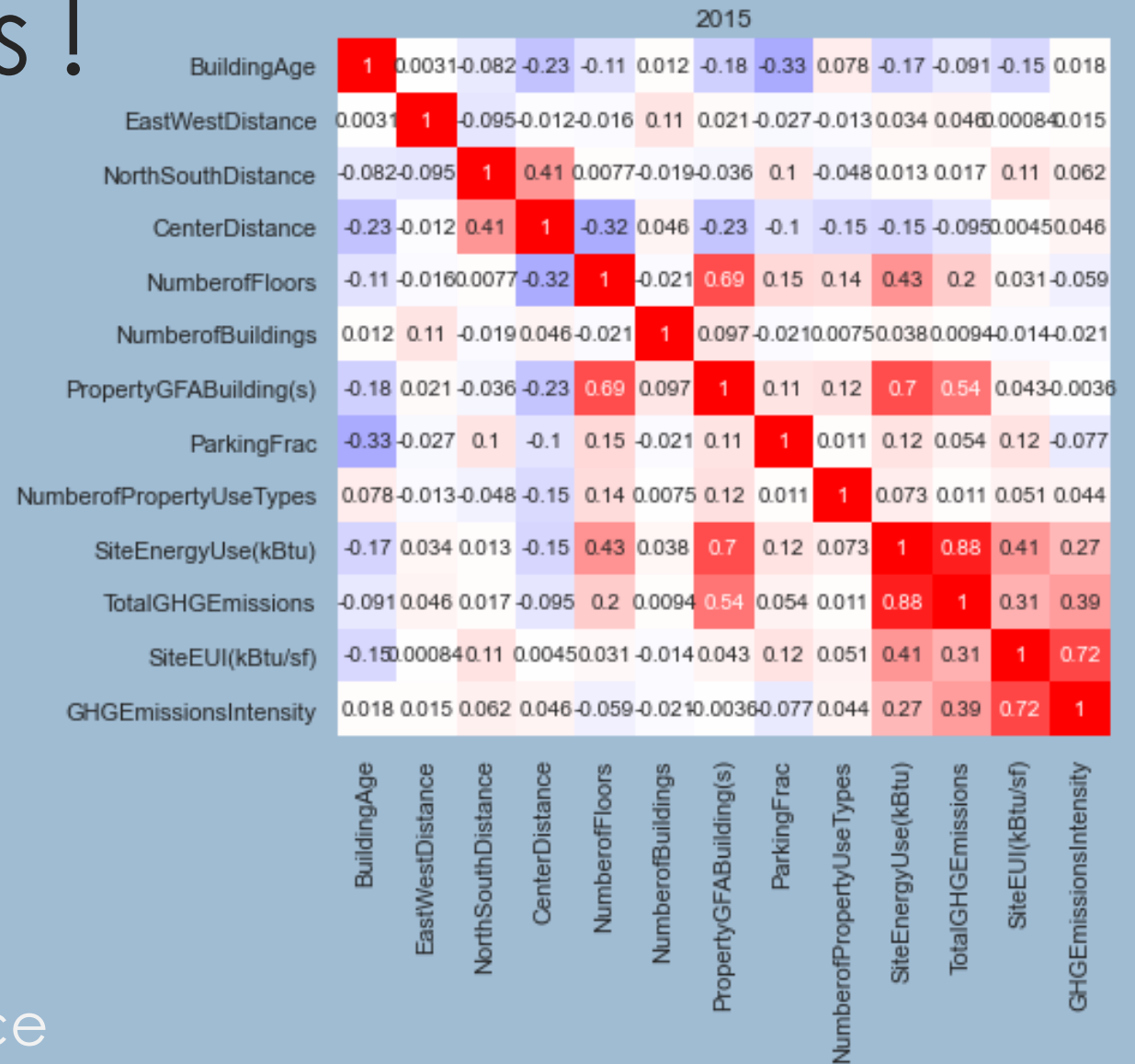
LargestPropertyUseType	Eta-carré
PropertyGFABuilding(s)	0.36
ParkingFrac	0.42
NumberofFloors	0.36
BuildingAge	0.14

Bâtiments en centre-ville :

- plus haute et plus de surface
- plus vieux

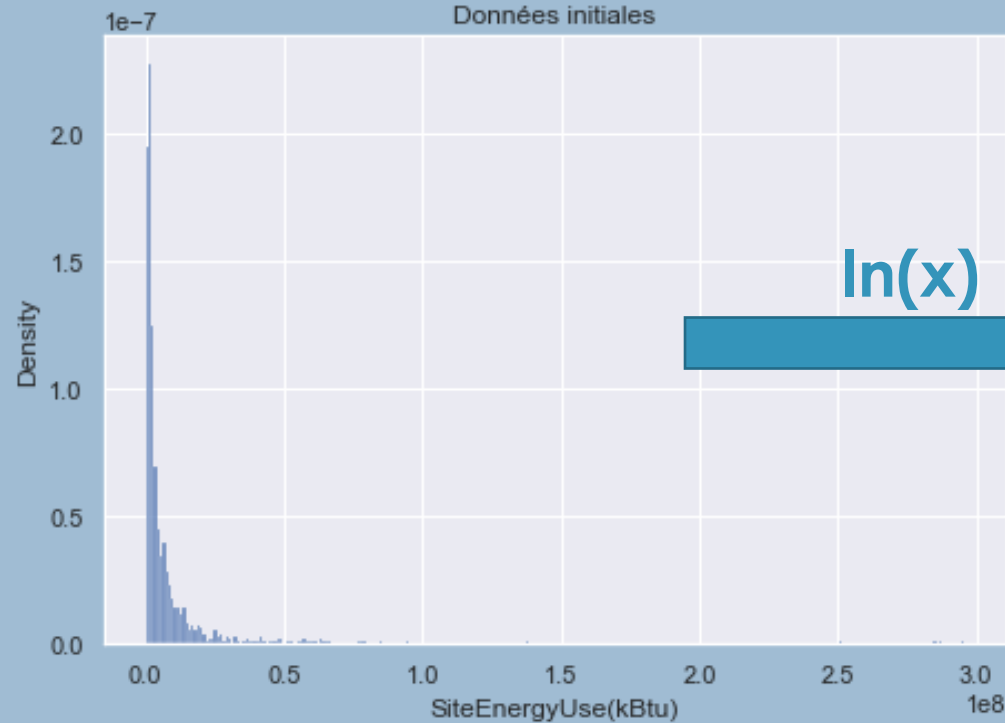
Bâtiments plus vieux :

- moins de % de parking
- moins haute et moins de surface

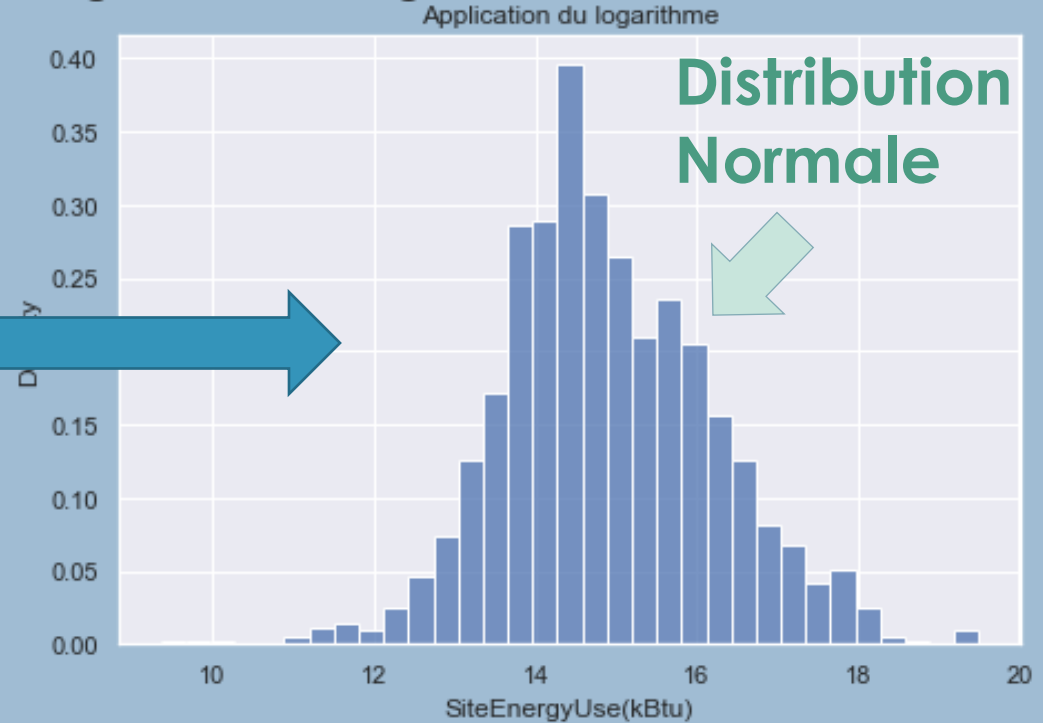


Consommation d'Énergie

Distribution de la Consommation d'Énergie avec Changement d'Echelle

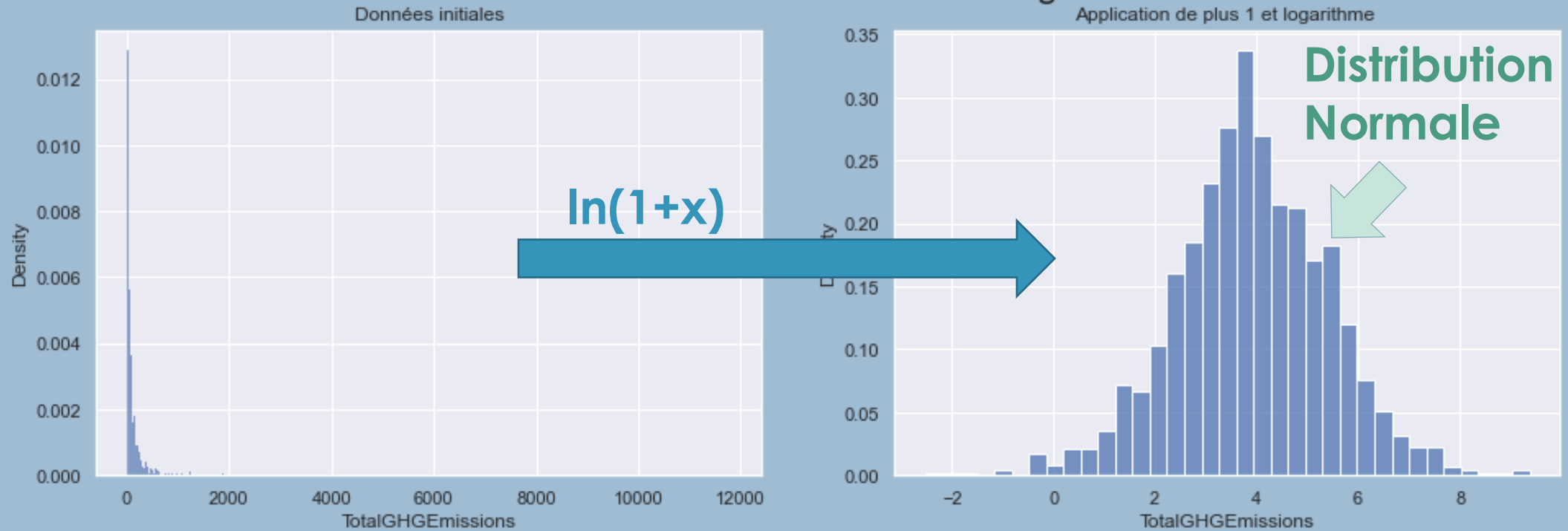


$\ln(x)$



Émissions de CO2

Distribution des Emissions de CO2 avec Changement d'Echelle



Combinaison des Datasets

1325 propriétés avec des données pour 2015 et 2016

- PrimaryPropertyType et nLargestPropertyUseType ne changent pas:
 - la **moyenne** des années pour les variables numériques
 - la valeur de 2016 pour les variables catégorielles
- PrimaryPropertyType et nLargestPropertyUseType changent :
 - Ce sont 59 des 1325 bâtiments : prend l'**entrée de 2016**



1539 Propriétés
pour construire le
modèle

Construire un Modèle

- **Features**
- Train-Test Division
- Prétraitement
- Modèle
- Pipeline
- Entraînement
- Optimisation
- Évaluation

Version 0 :

- PrimaryPropertyType
- PropertyGFABuilding(s)
- ParkingFrac



Version
Minimaliste

Version 1 :

- PrimaryPropertyType
- PropertyGFABuilding(s), **NumberOfFloors**
- ParkingFrac
- **BuildingAge, EastWestDistance, NorthSouthDistance**

Version 2 :

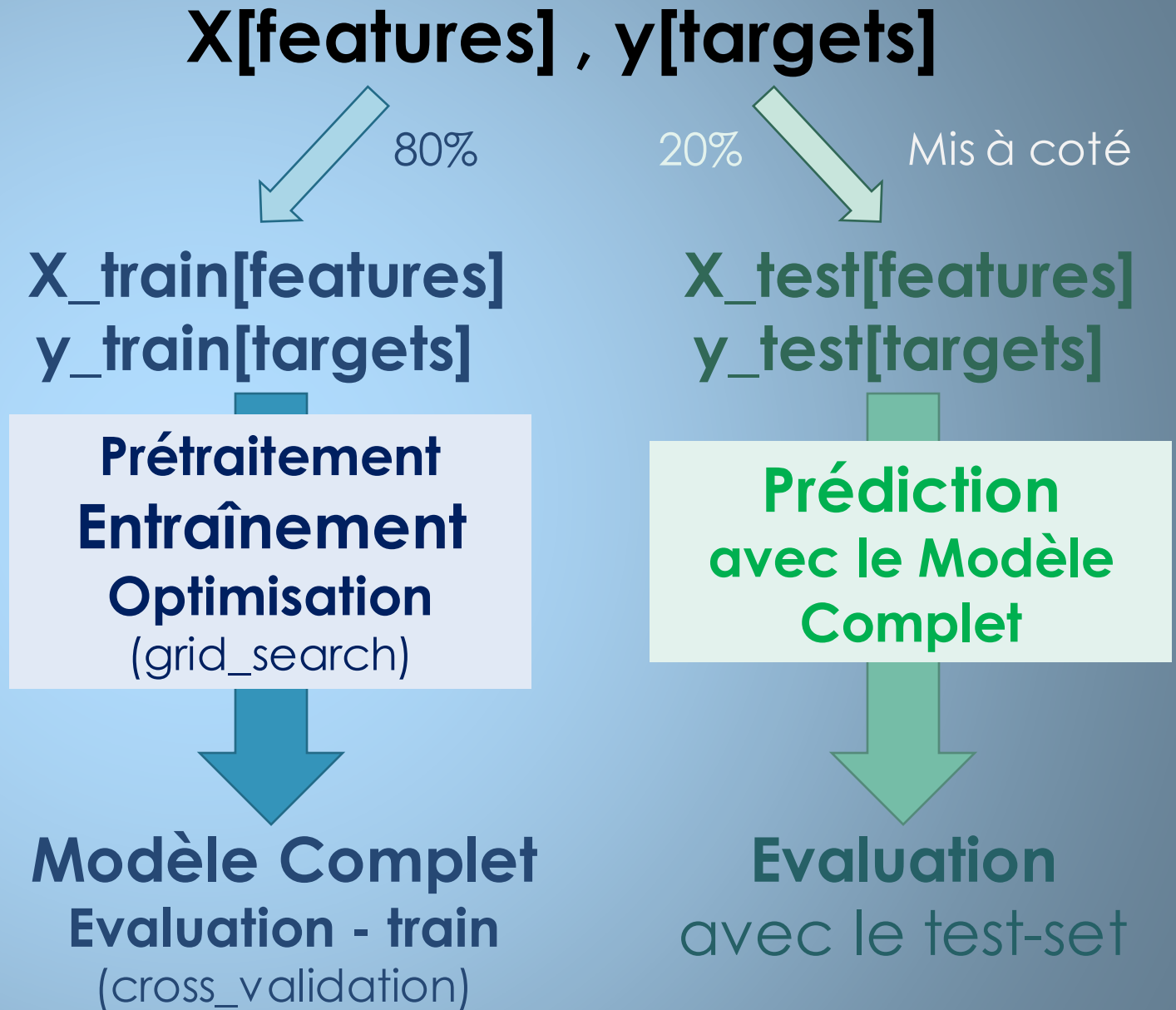
- **nLargestPropertyUseTypes**
- PropertyGFABuilding(s),
- ParkingFrac

Version 3 :

-

Construire un Modèle

- Features
- **Train-Test Division**
- Prétraitement
- Modèle
- Pipeline
- Entraînement
- Optimisation
- Évaluation



Construire un Modèle

- Features
- Train-Test Division
- **Prétraitement**
- Modèle
- Pipeline
- Entraînement
- Optimisation
- Évaluation

Conversion des catégories en numérique

Feature Engineering

- one-hot / frac-hot variables
 - PrimaryPropertyType, nLargestPropertyUseType
 - PublicBuilding, HasSwimmingPool, HasDataCenter, UseofSteam, UseofNaturalGas

Mise à l'échelle

Préprocesseurs pour features et targets

- StandardScaler
 - BuildingAge, EastWestDistance, NorthSouthDistance, CenterDistance
- Logarithme (+ StandardScaler)
 - **SiteEnergyUse(kBtu)**
 - PropertyGFABuilding(s), NumberofFloors, GFAperFlors
- Logarithme de 1+ (+ StandardScaler)
 - **TotalGHGEmissions**
 - ParkingFrac

Construire un Modèle

- Features
- Train-Test Division
- Prétraitement
- **Modèle**
- **Pipeline**
- Entraînement
- Optimisation
- Évaluation

Mise à l'échelle

+

Modèle

- DummyRegressor
- LinearRegressor
- Ridge
- Lasso
- LinearSVR
- SVR
- GradientBoostingRegressor
- KNeighborsRegressor
- RandomForestRegressor

Pipeline



Simplification du
processus de
modélisation

Création d'un
modèle complet

Construire un Modèle

- Features
- Train-Test Division
- Prétraitement
- Modèle
- Pipeline
- **Entraînement**
- **Optimisation**
- Évaluation

Entraînement / Optimisation



GridSearchCV

- Optimise les hyperparamètres
- Retourne le meilleur version de modèle
- Retourne une évaluation du modèle (cross validation avec 5 divisions)

Le Métrique - R^2

une mesure statistique représentant la proportion de la variance d'une variable dépendante qui est expliquée par une ou plusieurs variables indépendantes dans un modèle de régression

- $R^2 = 1$ modèle explique la variance parfaitement
- $R^2 = 0$ le modèle performe également à la moyenne

Construire un Modèle

- Features
- Train-Test Division
- Prétraitement
- Modèle
- Pipeline
- Entraînement
- Optimisation
- **Évaluation**

Évaluation

Cross Validation
avec le Train set



Évaluation
avec le Test set

Comparaison
Surapprentissage ?

Les Métriques

- **R²** - coefficient de détermination
- **MAE** – erreur absolue moyenne
- **RMSE** – racine de l'erreur quadratique moyenne
- **Temps d'entraînement**

Comparaisons des Modèles

- la Consommation d'Énergie (EnergieSiteUse(kBtu))

	Train R2	Test R2	Train nMAE xe6	Test nMAE xe6	Train nRMSE	Test nRMSE	Train time (ms)
Dummy0	-0.069	-0.104	-6.59	-5.55	-19.89	-12.53	18.2
SVR0	0.774	0.621	-3.03	-3.10	-8.85	-7.35	64.3
GradientBoostingR0	0.751	0.630	-3.17	-3.12	-9.02	-7.26	81.1
SVR1	0.764	0.648	-3.01	-3.00	-8.89	-7.08	47.7
Ridge2	0.782	0.603	-2.93	-3.16	-8.84	-7.51	9.8
Lasso2	0.783	0.626	-2.93	-3.11	-8.80	-7.29	9.3
LinearSVR2	0.776	0.491	-2.95	-3.34	-9.00	-8.50	14.0
SVR2	0.817	0.621	-2.85	-3.11	-7.94	-7.34	39.0
GradientBoostingR2	0.753	0.744	-3.05	-2.82	-8.97	-6.04	91.2
RandomForestR2	0.770	0.670	-3.16	-3.12	-8.88	-6.85	57.9
LinearSVR3	0.755	0.369	-3.05	-3.32	-9.45	-9.47	22.4
SVR3	0.788	0.602	-2.96	-3.09	-8.25	-7.52	57.2
RandomForestR3	0.759	0.604	-3.13	-3.24	-9.21	-7.50	55.0
SVR4	0.784	0.630	-2.99	-3.06	-8.39	-7.25	63.6

Comparaisons des Modèles

- la Consommation d'Énergie (EnergieSiteUse(kBtu))

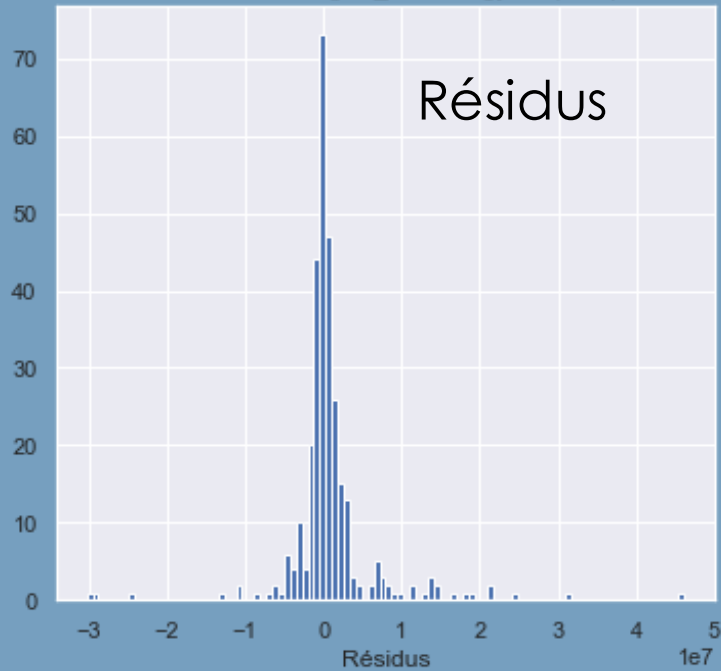
- Les **modèles de Version 2** ont l'air de performer le mieux
 - nLargestPropertyUseType (frac-hot variables)
 - PropertyGFABuilding(s)
 - ParkingFrac
- Les **classes de modèle** le plus performantes:
 - GradientBoostingRegressor
 - RandomForestRegressor
 - SVR
- **LinearSVR** surajoute les données
- **GradientBoostingR2** est le plus performant modèle
 - Learning rate = 0.1778 (hyperparamètre optimisé)
 - Temps d'entraînement : 91 ms (presque 10 fois le temps que pour le Ridge ou Lasso)

Comparaisons des Modèles

- la Consommation d'Énergie

GradientBoostingR2

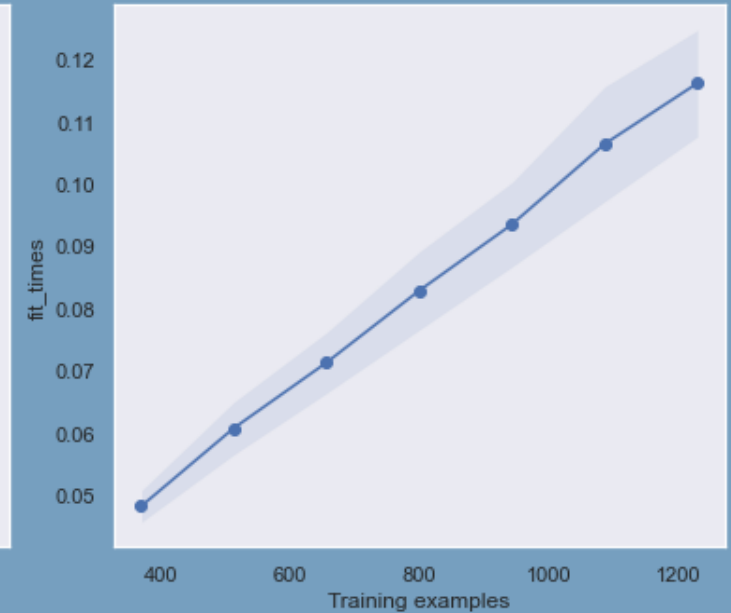
GradientBoostingR2_SiteEnergyUse(kBtu)



Courbe d'Apprentissage



Scalability of the model



Comparaisons des Modèles

- la Consommation d'Énergie

GradientBoostingR2

28 Data Centers:

- 2 à 100 %
- 3 > 30 %

20 Hôpitaux:

- 18 à 100 %

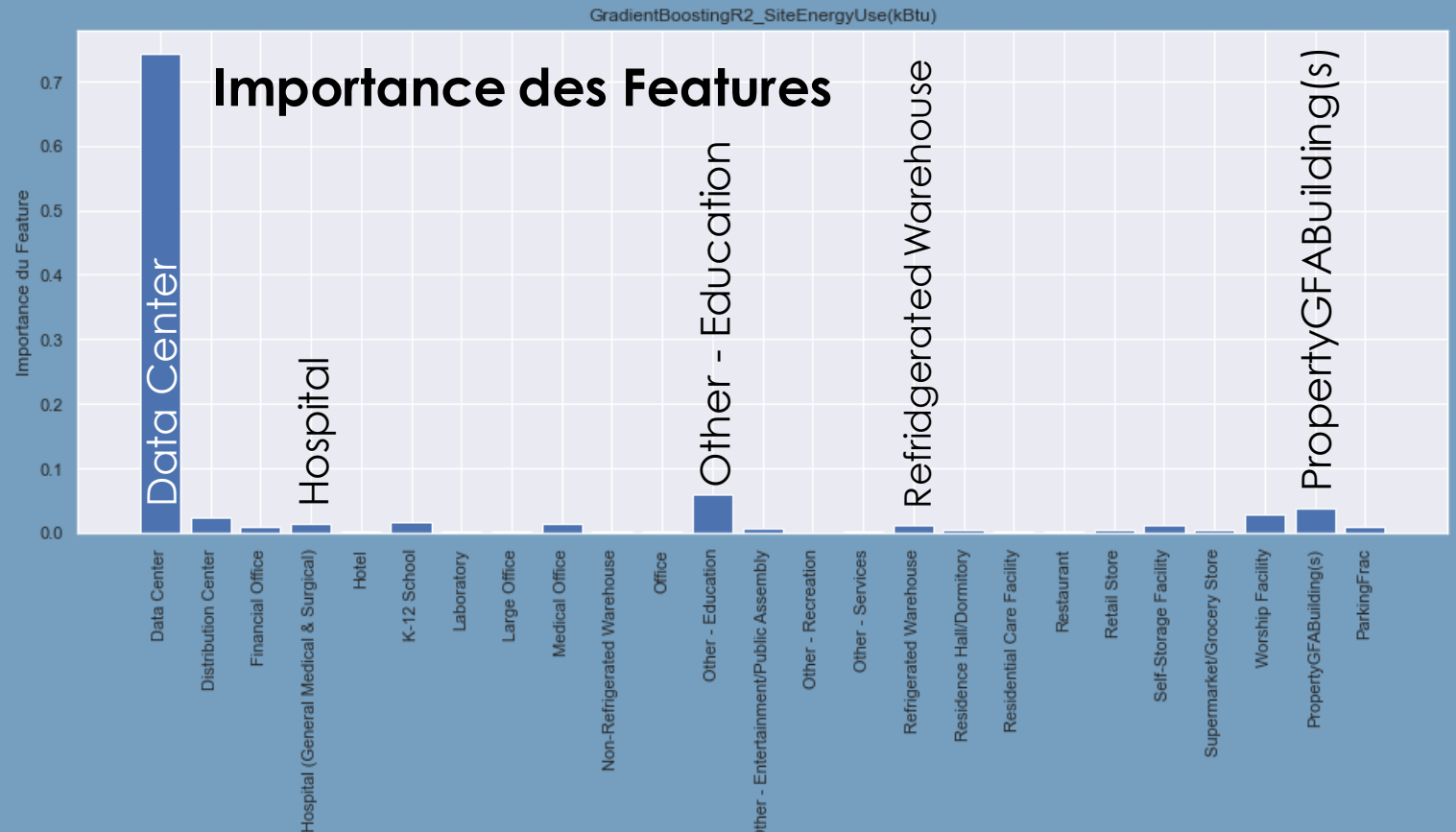
16 Refrigerated Warehouses

- 9 > 90 %

31 Other- Education

- 19 > 98%

Petits échantillons



Comparaisons des Modèles

- Émissions de CO2 (TotalGHGEmissions)

	Train R2	Test R2	Train nMAE	Test nMAE	Train nRMSE	Test nRMSE	Train time (ms)
Dummy0	-0.001	-0.011	-193.6	-165.5	-591.5	-286.5	11.2
GradientBoostingR0	0.768	0.215	-95.4	-116.9	-239.0	-252.5	76.8
KNR0	0.736	0.361	-108.9	-107.2	-284.2	-227.7	6.34
RandomForestR0	0.775	0.314	-103.2	-115.1	-258.2	-235.9	32.0
GradientBoostingR1	0.752	0.078	-98.8	-113.6	-255.4	-273.5	156.0
SVR2	0.735	0.481	-97.3	-96.2	-301.0	-205.2	7486.7
GradientBoostingR2	0.794	0.344	-92.8	-106.7	-242.1	-230.8	89.7
KNR2	0.743	0.541	-104.0	-98.1	-271.0	-193.0	7.90
RandomForestR2	0.808	0.345	-94.3	-107.2	-249.6	-230.6	32.4
SVR3	0.713	0.555	-101.0	-88.5	-310.6	-190.1	56262.8
GradientBoostingR3	0.753	0.511	-98.8	-103.1	-263.0	-199.3	151.9
RandomForestR3	0.782	0.511	-88.0	-96.1	-252.0	-233.0	127.2
GradientBoostingR4	0.772	0.474	-89.2	-97.5	-254.0	-206.6	208.5
RandomForestR4	0.737	0.374	-92.7	-98.3	-289.8	-225.4	61.7

Comparaisons des Modèles

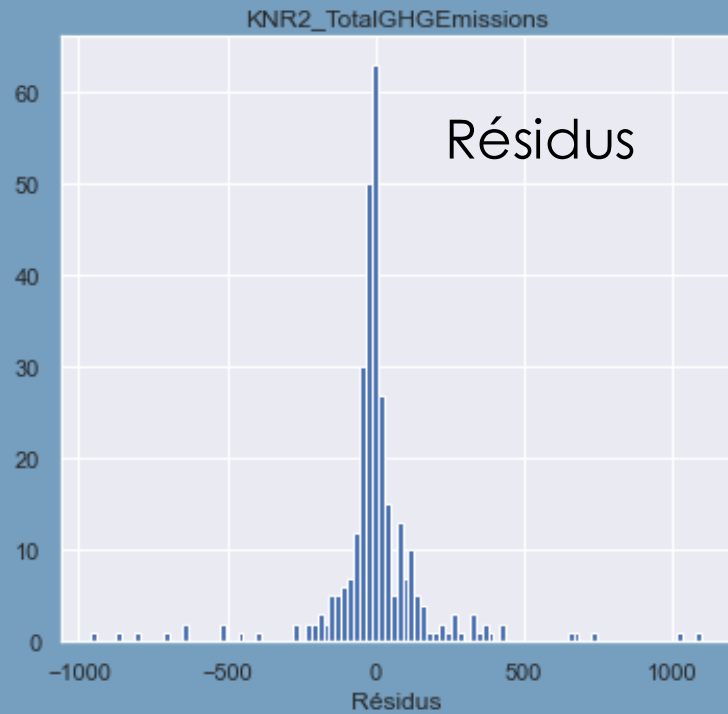
- Émissions de CO2 (TotalGHGEmissions)

- **Les classes de modèle** le plus performantes:
 - GradientBoostingRegressor
 - RandomForestRegressor
- **SVR** est performant mais a **un temps d'entraînement trop longue**
 - **Version 0** : 3.7s / **Version 1** : 4.1s / **Version 2** : 7.4s / **Version 3** : 56s !!
- **LinearSVR ne converge pas**
- **KNR2** est le plus performant modèle
 - Leaf size : 1, N neighbors : 2, Weights: 'uniform' (hyperparamètres optimisés)
 - **Temps d'entraînement : 7.9 ms**

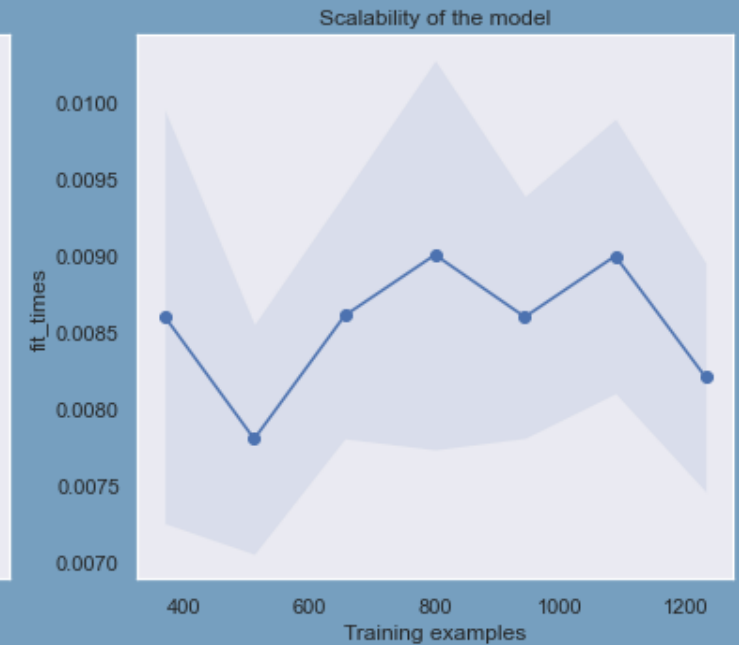
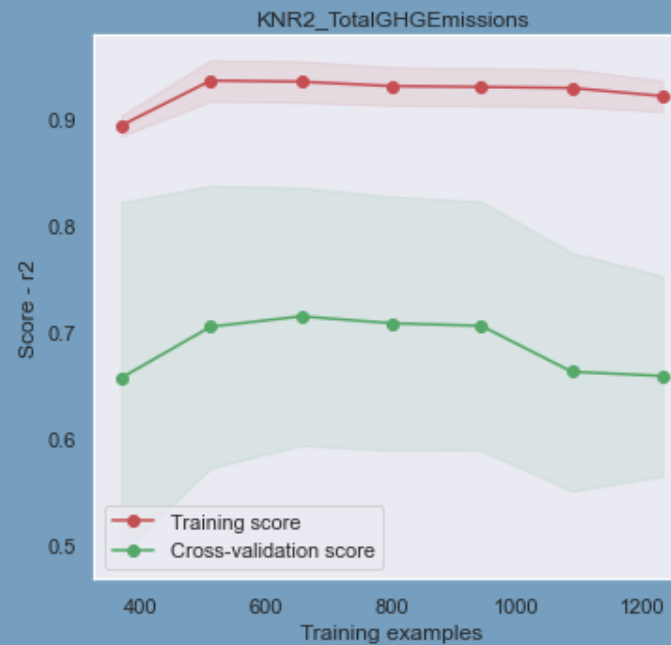
Comparaisons des Modèles

- Émissions de CO2 (TotalGHGEmissions)

KNR2



Courbe d'Apprentissage



Comparaisons des Modèles

- Émissions de CO2 (TotalGHGEmissions)

RandomForestR2

28 Data Centers:

- 2 à 100%
- 3 > 30 %

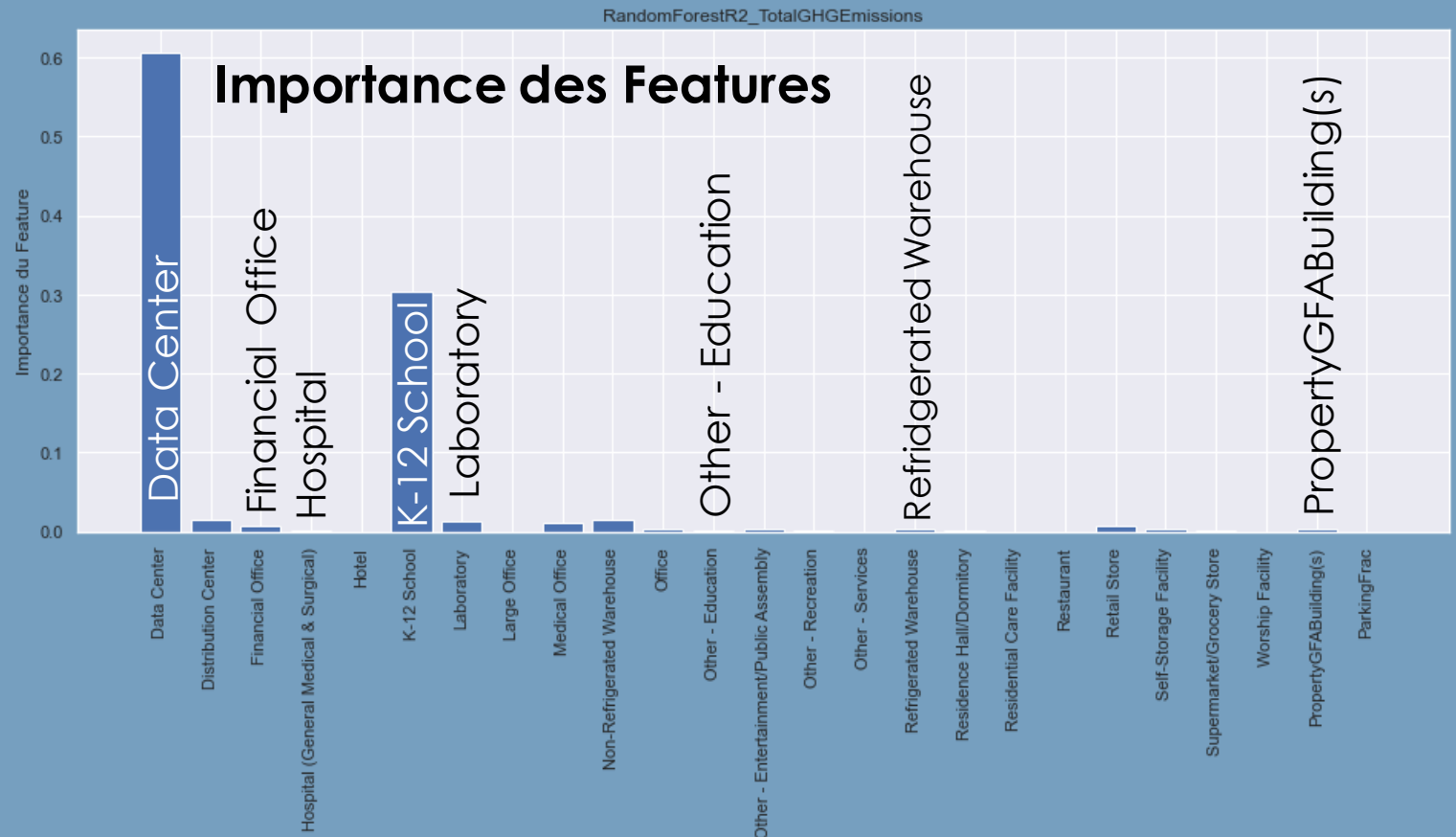
23 Laboratories :

- 7 > 99%
- 7 > 50%

25 Financial Offices :

- 9 > 75%

Petits échantillons



Comparaisons des Modèles

- Émissions de CO2 (TotalGHGEmissions)

- **GradientBoostingR3** et **RandomForestR3** sont aussi performants
 - Temps d'entraînement : 152 ms et 127ms (plus que 15 fois plus longue que KNR2)
- **Version 3:**
 - nLargestPropertyUseTypeFrac
 - PropertyGFABuilding(s)
 - ParkingFrac
 - GPAperFloor
 - BuildingAge, CenterDistance
 - UseofSteam, UseofNaturalGas
 - PublicBuilding, HasSwimmingPool, HasDataCenter

} **Version 2**

Émissions de CO₂ - modèles avancés

Prédire la Consommation d'Energie

- avec GradientBoostingR2 (considéré comme le plus performant modèle)

Utiliser le ENERGYSTARScore

- imputer les valeurs nulles avec 0

Modéliser avec KNR2, GradientBoostingR3, RandomForestR3

- avec la prédiction de la consommation d'énergie
- avec l'ENERGYSTARScore
- avec les deux variables de plus

Émissions de CO2 - modèles avancés

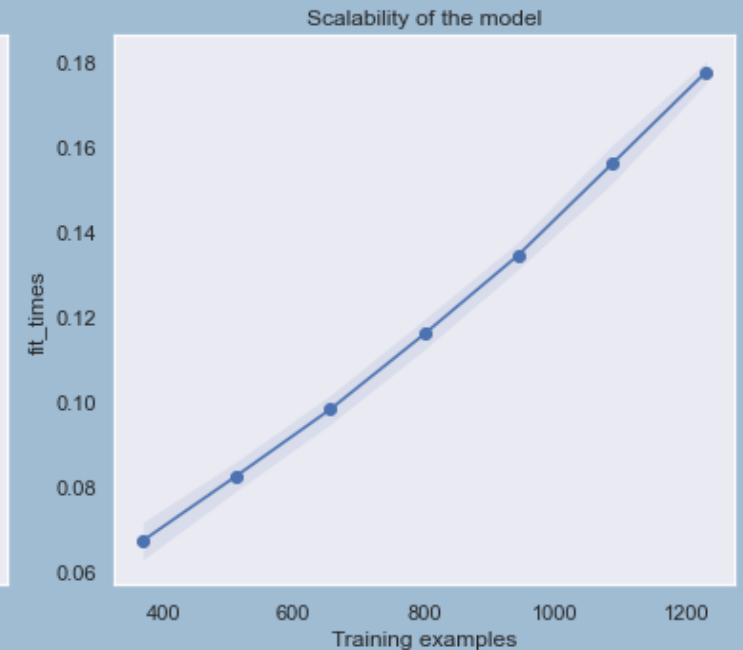
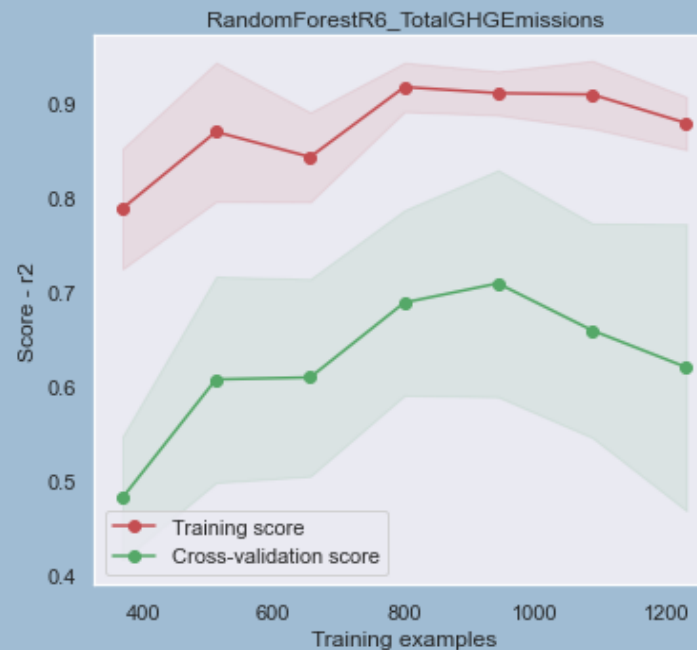
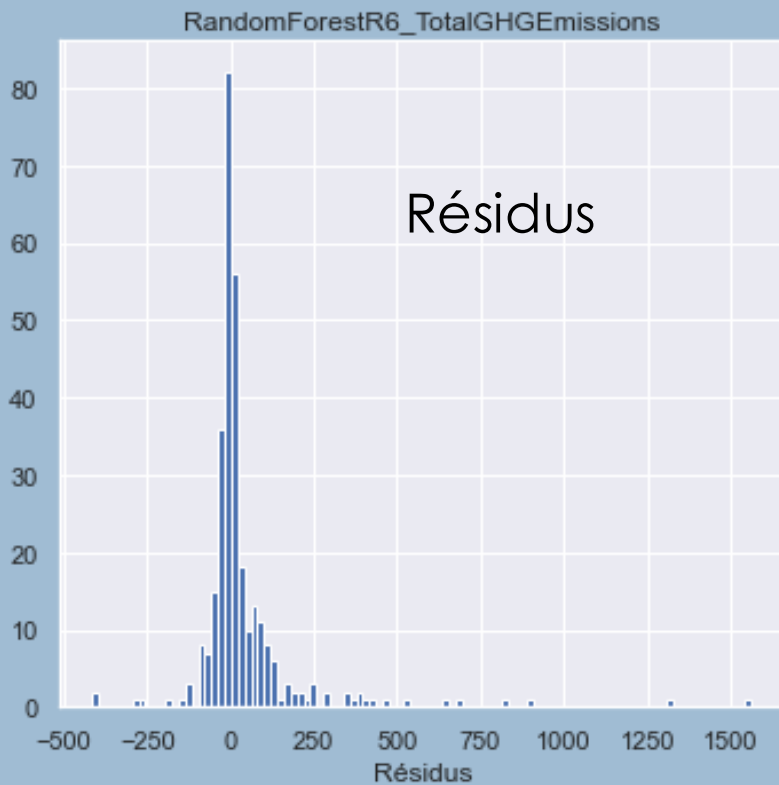
	Train R2	Test R2	Train nMAE	Test nMAE	Train nRMSE	Test nRMSE	Train time (ms)
KNR2	0.743	0.541	-104.0	-98.1	-271.0	-193.1	7.9
GradientBoostingR2	0.793	0.344	-92.8	-106.7	-242.0	-230.8	89.7
RandomForestR2	0.808	0.345	-94.3	-107.2	-249.6	-230.6	32.4
KNR3	0.558	-0.254	-105.3	-101.7	-370.9	-319.0	9.7
GradientBoostingR3	0.753	0.511	-98.8	-103.1	-263.0	-199.3	151.9
RandomForestR3	0.782	0.331	-88.0	-96.1	-252.0	-233.0	127.2
KNR5	0.790	0.509	-86.4	-91.2	-255.3	-199.6	*14.4
GradientBoostingR5	0.843	0.180	-69.0	-83.7	-214.7	-257.9	*284.9
RandomForestR5	0.873	0.411	-68.4	-84.5	-200.1	-218.7	*38.4
KNR6	0.603	0.240	-114.5	-105.9	-322.0	-248.4	17.8
GradientBoostingR6	0.726	0.387	-81.3	-72.9	-282.6	-223.1	265.3
RandomForestR6	0.752	0.578	-81.8	-74.0	-286.2	-185.0	142.0
GradientBoostingR7	0.844	0.146	-69.0	-79.1	-221.7	-263.3	*297.9
RandomForestR7	0.882	0.553	-66.7	-76.8	-191.8	-190.5	*48.1

Émissions de CO₂ - modèles avancés

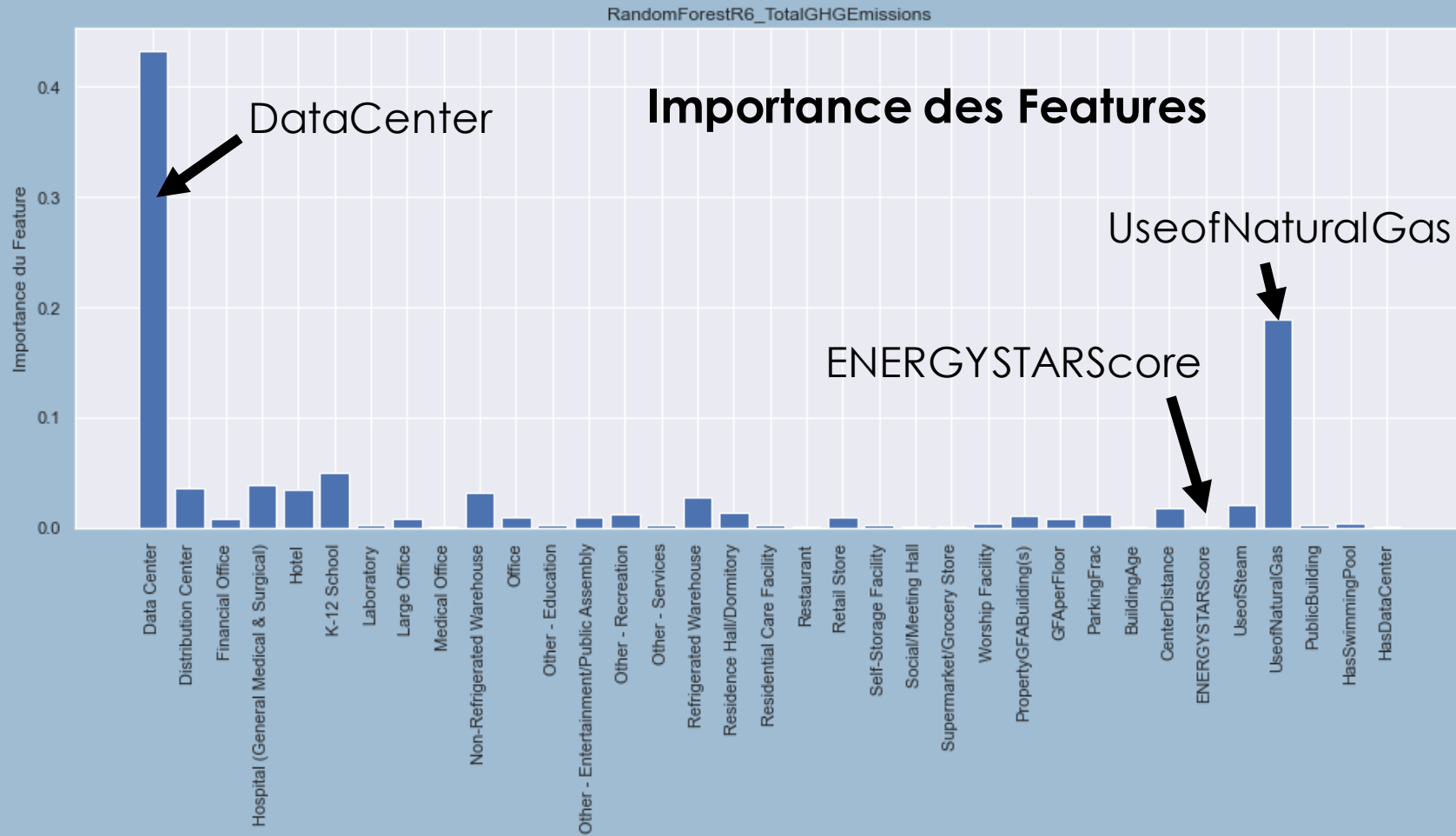
- **La Prédiction de la Consommation d'Énergie**
 - améliore **GradientBoostingR3** et **RandomForestR3**
 - entraîne un surapprentissage pour KNR2
- **ENERGYSTARScore**
 - améliore **GradientBoostingR3** et **RandomForestR3**
 - entraîne un surapprentissage pour KNR2
- **KNR a une tendance vers le surapprentissage**
- **RandomForestR3 + ENERGYSTARScore** est le modèle le plus performant
 - Max depth: 50, N estimators: 30
 - Temps d'entraînement : 142 ms

Émissions de CO2 - RandomForestR6

Courbe d'Apprentissage



Émissions de CO2 - RandomForestR6



Résumé

La Consommation d'Énergie:

➤ GradientBoostingR2

- Train R2 : 0.75, Test R2 : 0.74
- Temps d'entraînement : 91 ms

Les Émissions de CO₂:

- Le plus performante:
 - RandomForestR6 = RandomForestR3 + ENERGYSTARScore
 - Train R2 : 0.75, Test R2 : 0.58
 - Temps d'entraînement : 142 ms
- Modèle rapide et simple:
 - KNR2
 - Train R2 : 0.74, Test R2 : 0.54
 - Temps d'entraînement : 8 ms

Résumé

Variables paramètres le plus importantes:

- Taille du bâtiment : PropertyGFABuilding(s), ParkingFrac
- Usage du bâtiment : PrimaryPropertyType / nLargestPropertyUseType

Pistes d'Améliorations:

- Meilleure Catégorisation d'usage de bâtiment
 - essayer d'éviter les groupes avec peu de représentants
 - envisager d'écarter des cas particuliers (e.g. Data Center)
 - identifier des groupes d'observations similaires (e.g. clustering)
- Considérer la Non-linéarité (e.g. Polynomial Features)
- Sélection des Features
 - Différentes features ou collection de plus de données (e.g. panneaux solaires)
 - Analyser leur importance (e.g. SequentialFeatureSelector)
- Affiner les hyperparamètres du modèle / tester autres modèles