Projet 4

Anticipez les besoins en consommation électrique de bâtiments

22 Octobre 2021

MBIADOU SALEU G. Raïssa

PLAN

01

CONTEXTE ET JEU DE DONNÉES

CONTEXTE

🎯OBJECTIF : VILLE NEUTRE EN CARBONE D'ICI 2050

- Difficulté : coût important d'obtention des relevés / fastidieuses à collecter
- Mission :
  - ❑ Prédire les émissions de CO2 et la consommation totale d'énergie sans les relevés annuels (en particulier pour les bâtiments non destinés à l'habitation)
  - ❑ Evaluer l'intérêt de l'Energy Star Score pour la prédiction d'émissions

LES DONNÉES

- Relevés détaillés années 2015 et 2016
  - ✓ 3000+ bâtiments référencés
  - ✓ 46+ indicateurs
  - ✓ couvrant 3 types d'informations:
    - ➢ Les informations générales sur le bâtiment : nom, localisation, type d'utilisation
    - ➢ Les informations de consommation énergétique
    - ➢ Les informations d'émissions de CO2

FICHIER	NB DE LIGNES	NB DE COLONNES
Energy Data 2015	3340	47
Energy Data 2016	3376	46

02

TRAITEMENT ET NETTOYAGE DES DONNÉES

Environnement de développement

Anaconda

Installation d'Anaconda: plateforme de disbution python la plus populaire

environement virtuel

Mise en place d'un environnement virtuel dédié au projet

Installation des paquets

Installation des paquets nécessaires (numpy, pandas,matplotlib, seaborn, sklearn, scipy) avec la commande **pip install**

Fusion et pré-sélection des données

Harmonisation des colonnes et fusion des données de 2015 et 2016
(Noms de colonnes)

→ 6716 lignes
44 colonnes

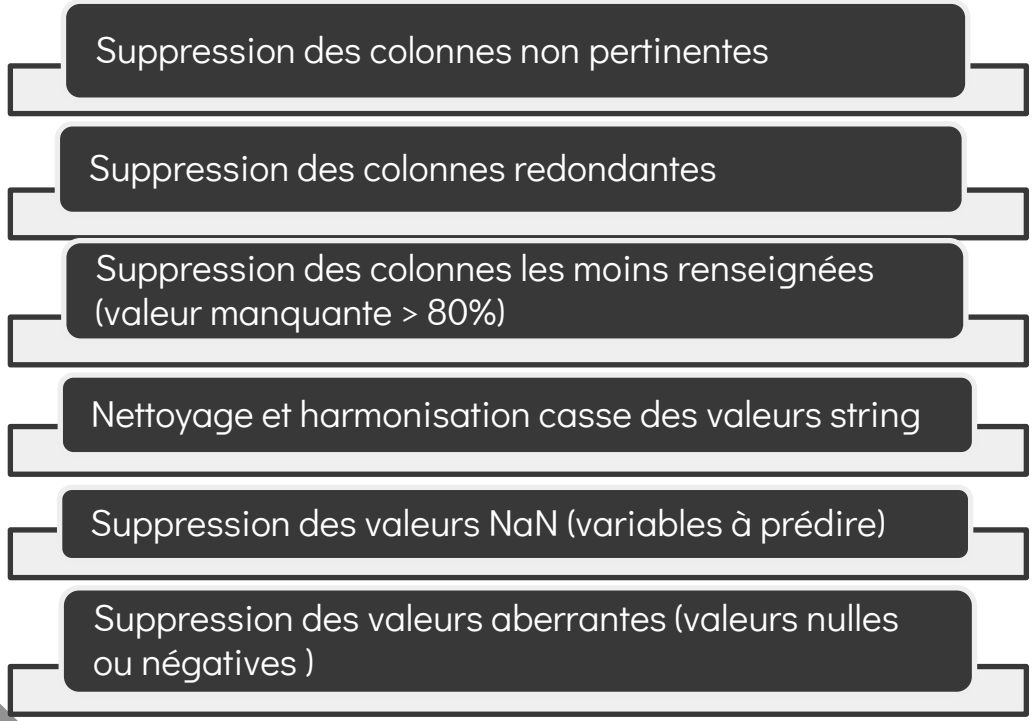
Sélection des bâtiments non résidentiels
(Suppression des lignes avec BuildingType=Multifamily)

→ **3318 lignes**
44 colonnes

Suppression des bâtiments dupliqués
(moyenne des variables sur les 2 années)

→ **1698 lignes**
44 colonnes

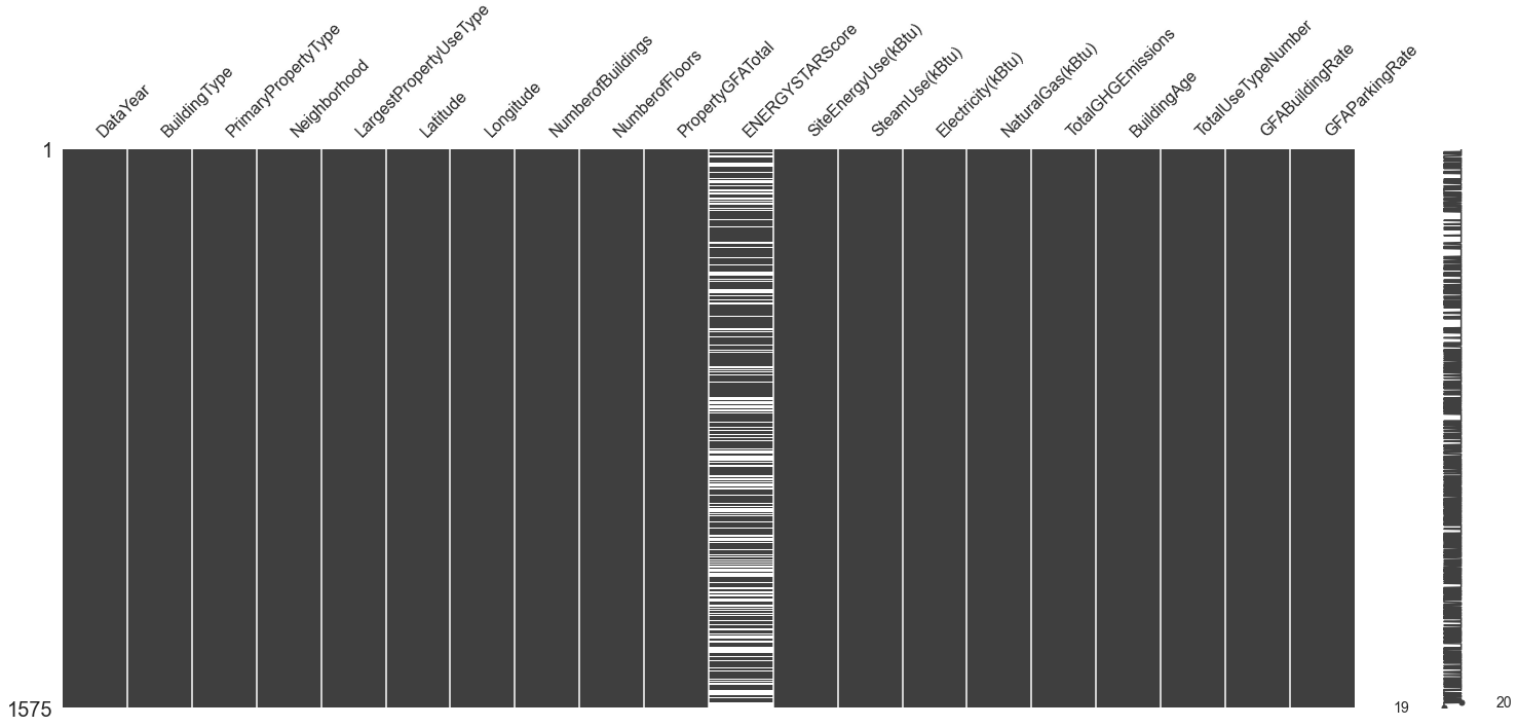
Nettoyage et features engineering



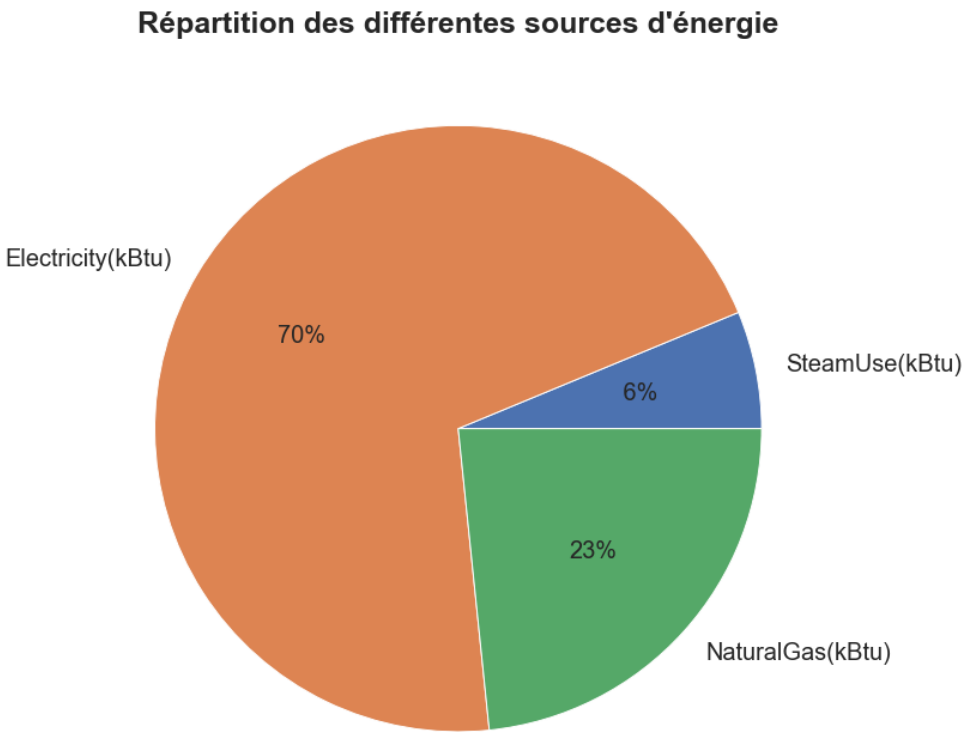
Nouvelles colonnes

- **BuildingAge** (DataYear – YearBuilt)
- **TotalUseTypeNumber** : nombre d'utilisations possible du bâtiment
- **GFABuildingRate** et **GFAParkingRate** : conversion des différentes surfaces (Buildings et Parking) en pourcentage de la surface totale

Données nettoyées



03
ANALYSE EXPLORATOIRE



PROPORTION DES DIFFÉRENTES SOURCES D'ÉNERGIE

GLOBAL

- L'électricité est la source d'énergie la plus largement utilisée.

	Mode
BuildingType	NONRESIDENTIAL
PrimaryPropertyType	SMALL- AND MID-SIZED OFFICE
LargestPropertyUseType	OFFICE
Neighborhood	DOWNTOWN
TotalUseTypeNumber	1

	Médiane	Moyenne	Variance	Écart-type	Coeff de variation
GFABuildingRate	1.000000e+00	9.330954e-01	2.045818e-02	1.430321e-01	0.153288
GFAParkingRate	0.000000e+00	6.690461e-02	2.045818e-02	1.430321e-01	2.137851
PropertyGFATotal	4.817900e+04	1.209164e+05	9.249315e+10	3.041269e+05	2.515182
NumberofBuildings	1.000000e+00	1.159048e+00	8.510107e+00	2.917209e+00	2.516902
NumberofFloors	2.000000e+00	4.256508e+00	4.534839e+01	6.734122e+00	1.582077
ENERGYSTARScore	6.950000e+01	6.302159e+01	8.083552e+02	2.843159e+01	0.451140
SiteEnergyUse(kBtu)	2.679420e+06	8.858683e+06	9.716993e+14	3.117209e+07	3.518818
SteamUse(kBtu)	0.000000e+00	5.473360e+05	3.080063e+13	5.549832e+06	10.139716
Electricity(kBtu)	1.756366e+06	6.149722e+06	4.582923e+14	2.140776e+07	3.481095
NaturalGas(kBtu)	4.821830e+05	2.041839e+06	9.501722e+13	9.747677e+06	4.773969
TotalGHGEmissions	5.093500e+01	1.935611e+02	5.969983e+05	7.726567e+02	3.991796
Latitude	4.761275e+01	4.761649e+01	2.194970e-03	4.685050e-02	0.000984
Longitude	-1.223332e+02	-1.223335e+02	5.415020e-04	2.327019e-02	-0.000190
BuildingAge	5.100000e+01	5.449524e+01	1.087810e+03	3.298197e+01	0.605227

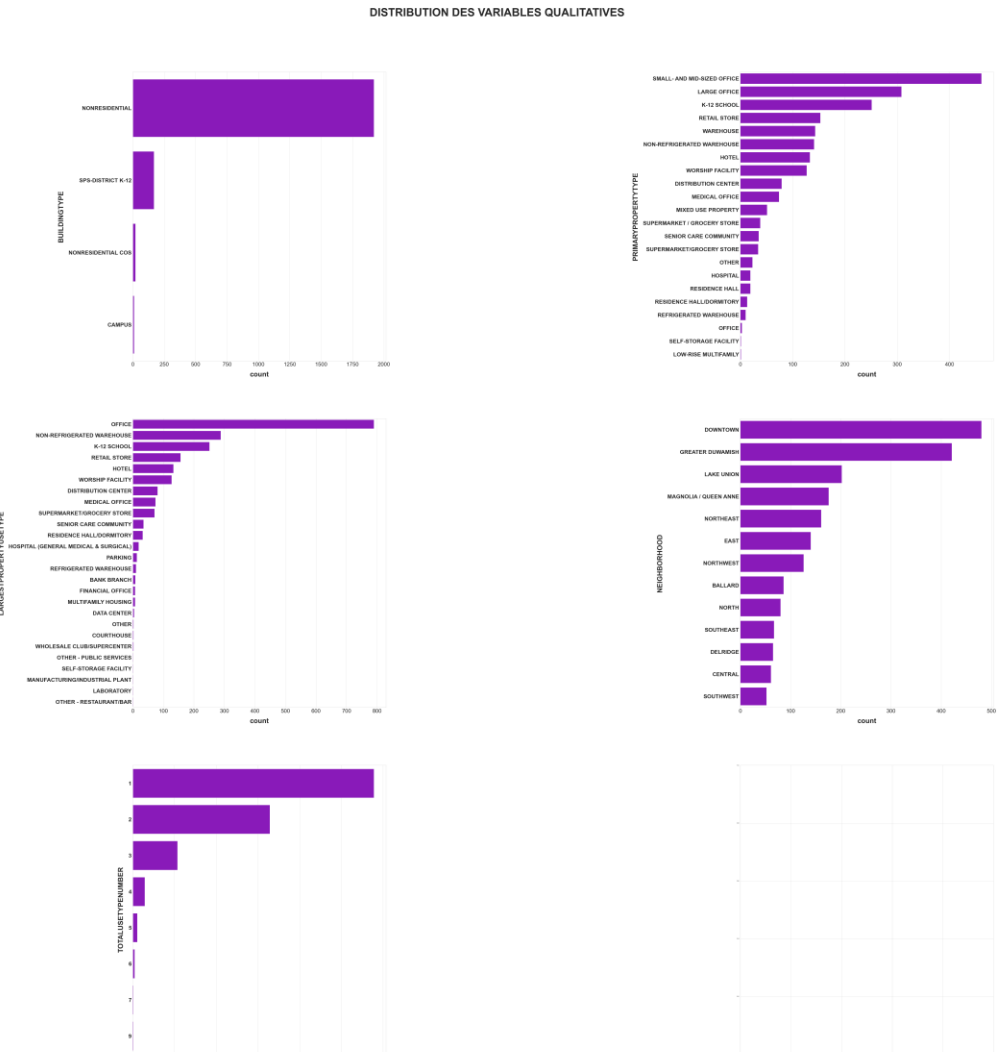
MESURES DE TENDANCE CENTRALE

VARIABLES QUALITATIVES

- Les bâtiments du jeu de données sont essentiellement des constructions situées au centre ville et abritant des bureaux de petite et moyenne taille.

VARIABLES QUANTITATIVES

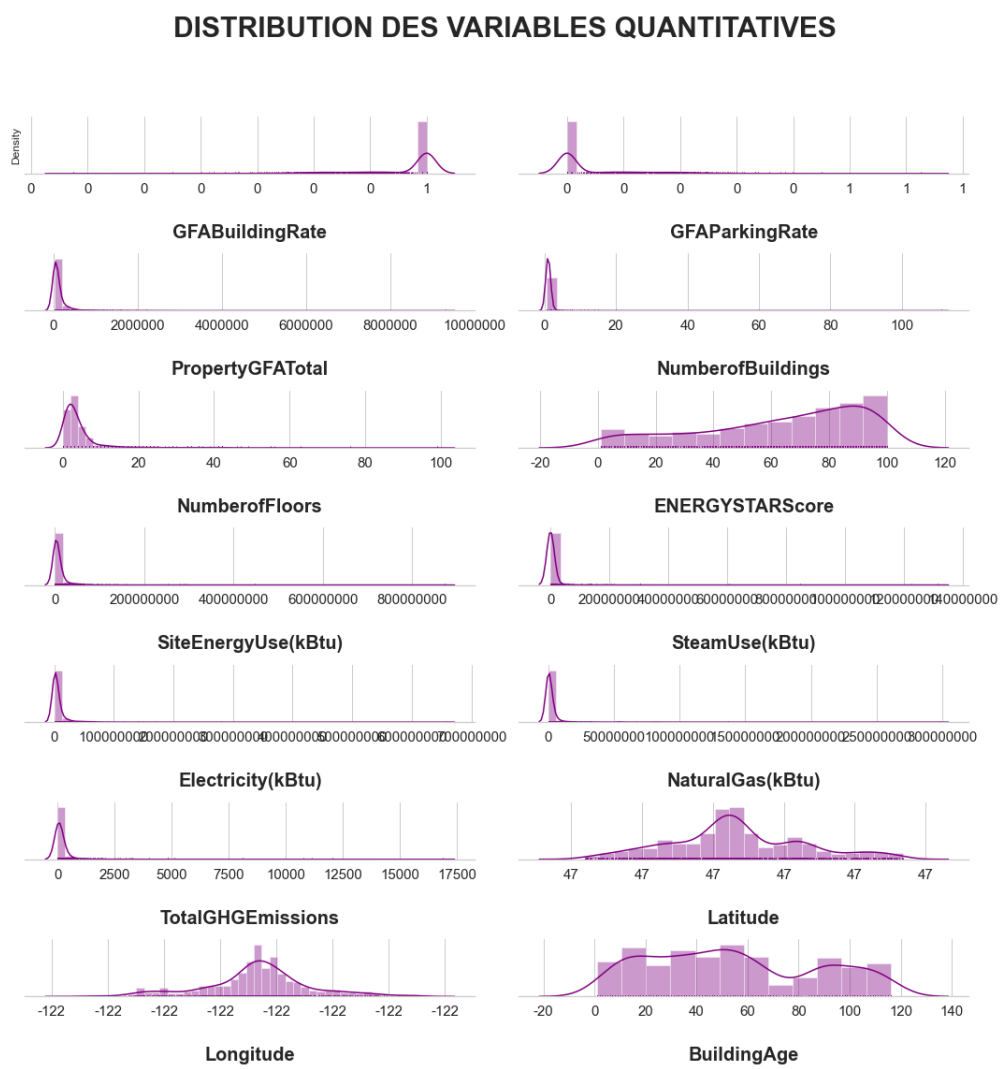
- Une grande variabilité excepté pour certaines variables
- Les constructions du jeu de données :
- Comportent en moyenne 1 bâtiment
- Possèdent en moyenne 4 étages



MESURES DE DISPERSION

VARIABLES QUALITATIVES

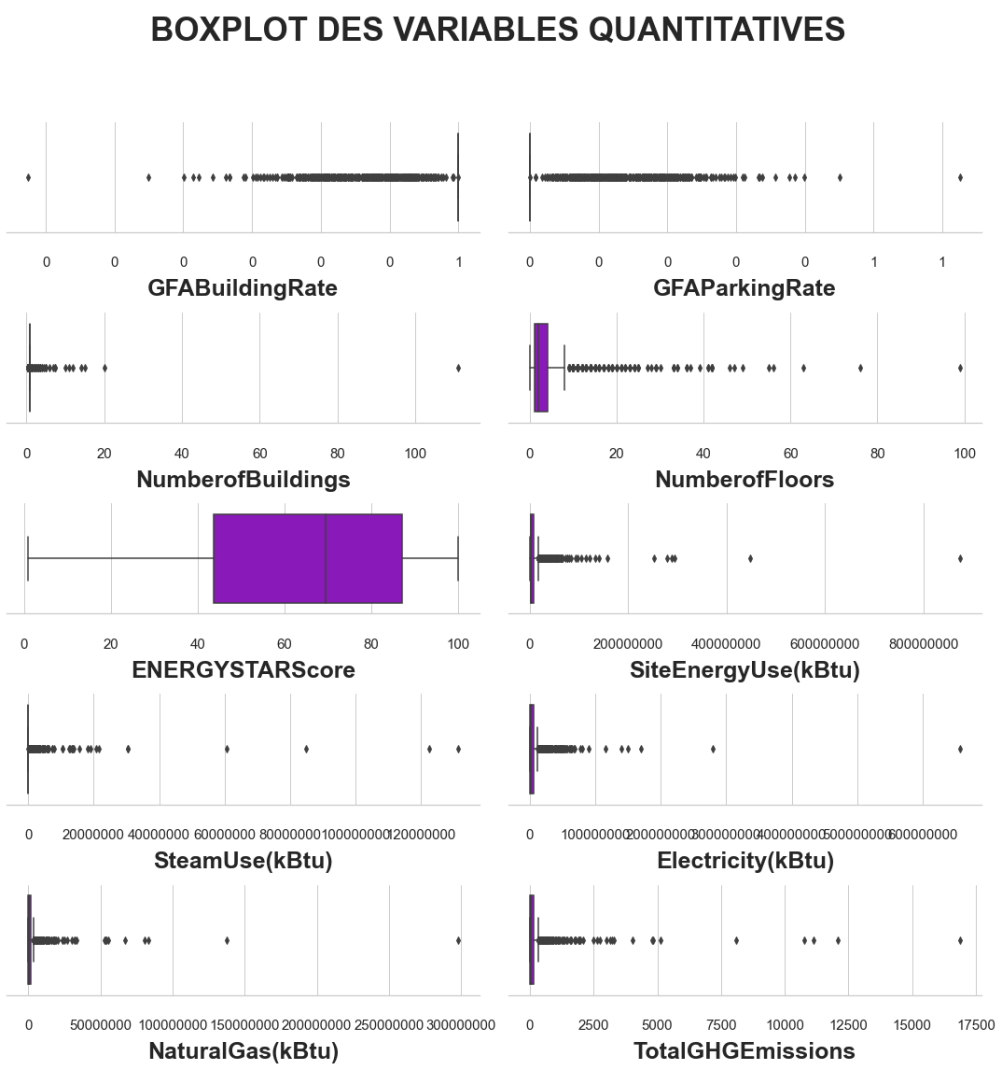
- Distribution peu équilibrée de certaines variables au sein de l'échantillon.



MESURES DE DISPERSION

VARIABLES QUANTITATIVES

- Caractère globalement inégalitaire des distributions des variables



Mesures de dispersion

VARIABLES QUANTITATIVES

Les 2 variables à prédire ont des valeurs très dispersées. Il sera certainement intéressant de les passer en logarithme

VARIABLES QUALITATIVES et SiteEnergyUse

	Coeff de corrélation
LargestPropertyUseType	0.330274
PrimaryPropertyType	0.325209
TotalUseTypeNumber	0.072962
BuildingType	0.053605
Neighborhood	0.026348

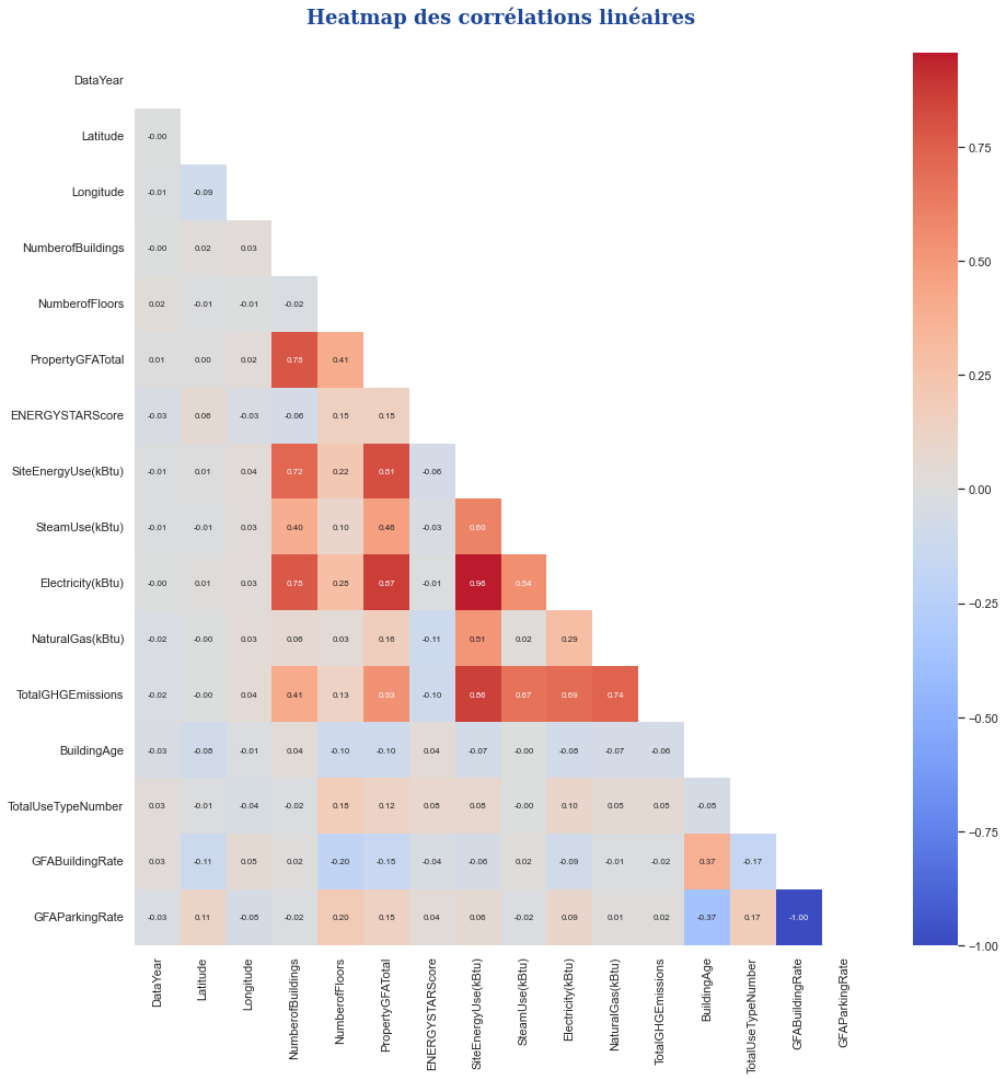
VARIABLES QUALITATIVES et TotalGHGEmissions

	Coeff de corrélation
PrimaryPropertyType	0.335679
LargestPropertyUseType	0.322778
BuildingType	0.074237
TotalUseTypeNumber	0.032696
Neighborhood	0.020766

Corrélations

Le rapport de corrélation éta carré montre que pour les 2 variables à prédire, les variables qualitatives les plus corrélées avec elles sont:

- PrimaryPropertyType
- LargestPropertyUseType



CORRÉLATIONS

- Les 2 variables à prédire sont corrélées entre-elles
- Les variables à prédire sont également corrélées avec les différentes sources d'énergie, le nombre de bâtiments, le nombre d'étages ainsi que les surfaces au sol.

Conclusions de l'analyse

- Les variables présentent des distributions peu équilibrées
- Plusieurs variables présentent des corrélations avec nos 2 variables d'intérêt, la consommation énergétique et les émissions de CO2.
- Il devrait être possible de créer des modèles permettant de prédire ces variables.
- Toutefois, le déséquilibre dans la distribution de certaines variables pourrait impacter la qualité des modèles.

04

MODELISATION

Encodage des variables

Variables catégorielles :

```
1  categorical_features.nunique()
```

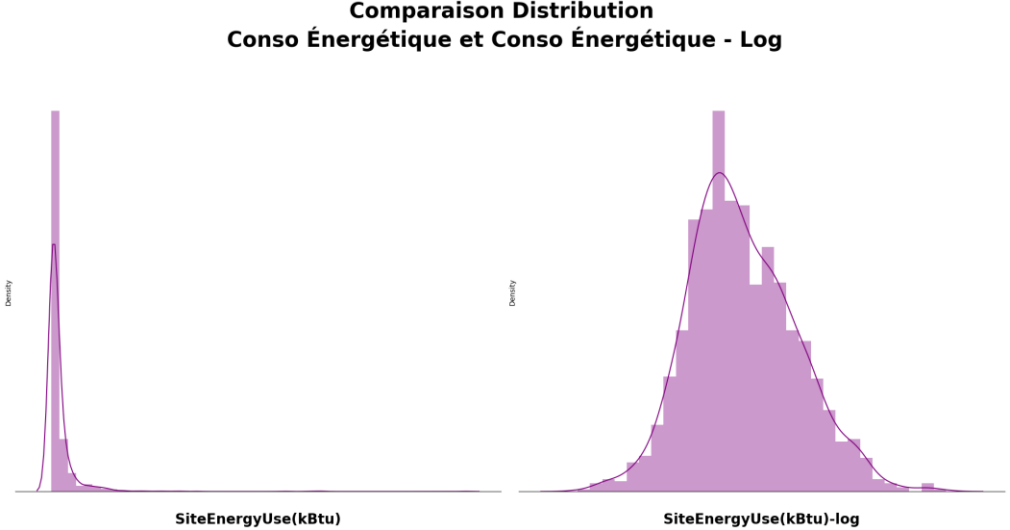
```
BuildingType              4
PrimaryPropertyType      22
Neighborhood             13
LargestPropertyUseType   26
dtype: int64
```

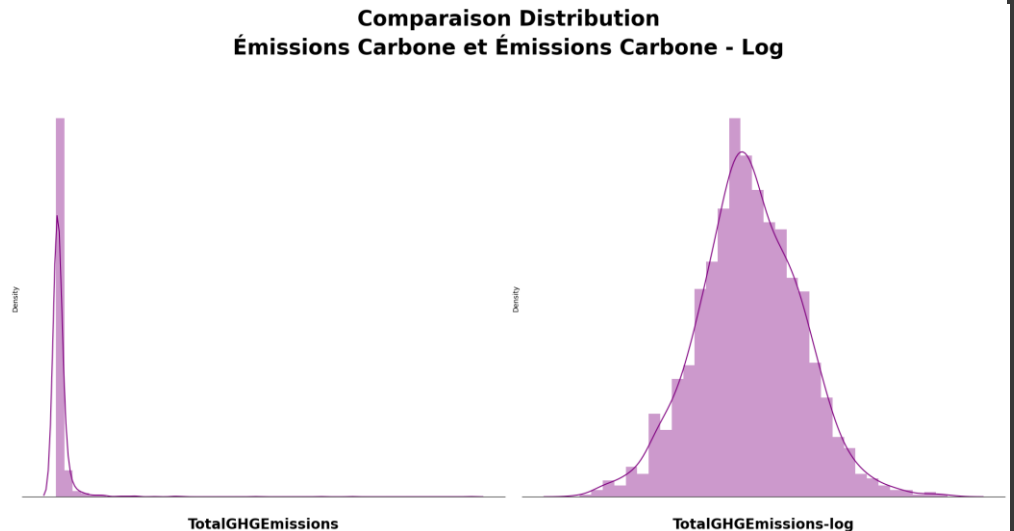
Variables numériques : *RobustScaler*

Target Encoding
(librairie Category_Encoders)

- Groupe les données par chaque catégorie et compte de nombre d'occurrences de chaque cible.
- Calcul de la probabilité que chaque cible se produise pour chaque groupe spécifique.

NB: Nous avons supprimé toutes les données de relève énergétiques (en dehors des 2 variables à prédire) de notre dataset pour éviter une fuite de données





IMPACT DU PASSAGE À L'ÉCHELLE LOGARITHMIQUE SUR LES DISTRIBUTION

- En passant les données à l'échelle logarithmique, nous obtenons une distribution normale des données à prédire.
- Nous allons donc appliquer cette transformation dans notre pipeline grâce à la fonction *TransformedTargetRegressor* de la librairie *Sklearn*.

Modèles et critères de comparaison

Famille de modèles	Modèle choisi
Baseline	Linear Regressor
Linéaires	Elastic Net
Machines à vecteurs de support	Support Vector Regressor (SVR)
Proches Voisins	K Neighbors Regressor
Ensemblistes	Random Forest Regressor

Critères

- R2 : Coefficient de determination
- MAE: Mean Absolute Error

Interprétation

- Plus les valeur MAE est proche de zéro, meilleur est le modèle évalué en terme d'exactitude
- Plus la valeur de R2 est proche de 1, plus les observations sont regroupées autour de la droite de régression, et par conséquent plus les erreurs de prédictions sont faibles.

DÉMARCHE

Séparation jeu de données

Train (80%)

Test (20%)

Définition grille de paramètres

Entrainement des modèles

Paramétrage optimal obtenu via validation croisée

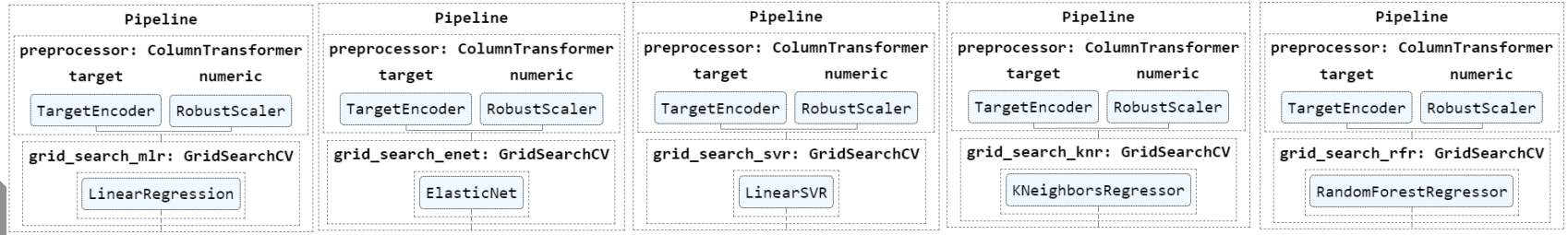
Pipeline: (Preprocessor, GridSearchCv)

(Train, cv=5)

Comparaison des modèles

MAE, R2

(fonction inverse (exp))



Prédiction de la consommation énergétique

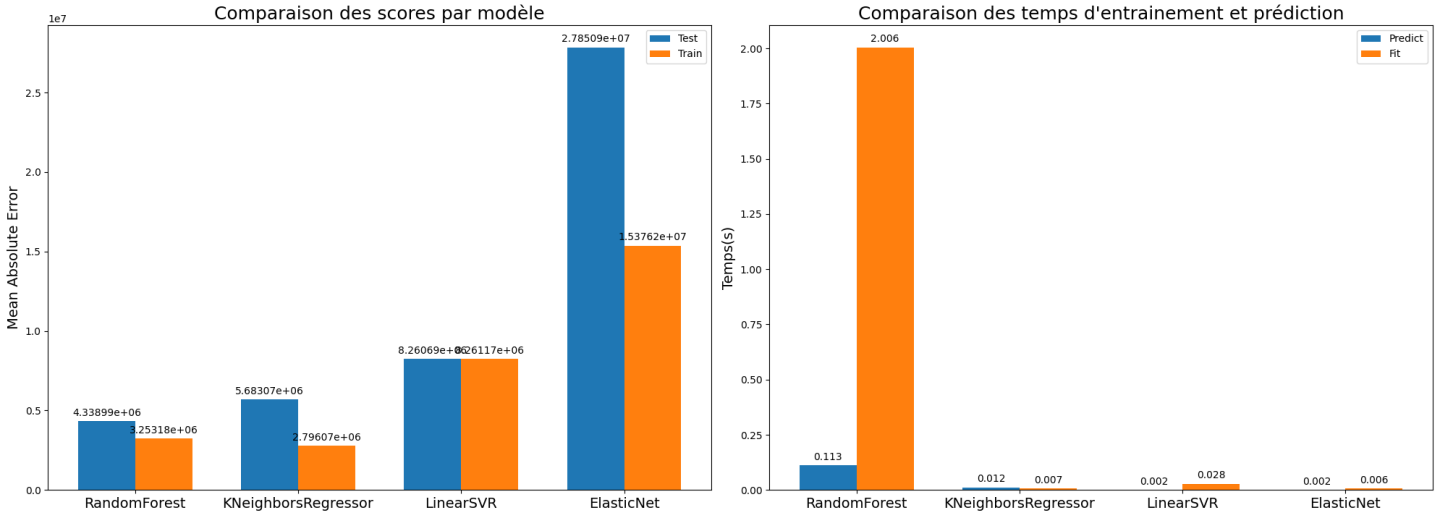
Performances des modèles (Validation croisée)

Meilleurs scores de la GridSearch (SiteEnergyUse(kBtu))

	MAE	R2	Time
RandomForestRegressor	3884568.320819	0.603766	2525.01
KNeighborsRegressor	4982387.01345	0.423037	8.89
LinearSVR	8260690.982534	-0.210575	35.81
ElasticNet	30519333.446128	-82.456825	7.92
LinearRegression	31876086.679196	-346.493345	4.7

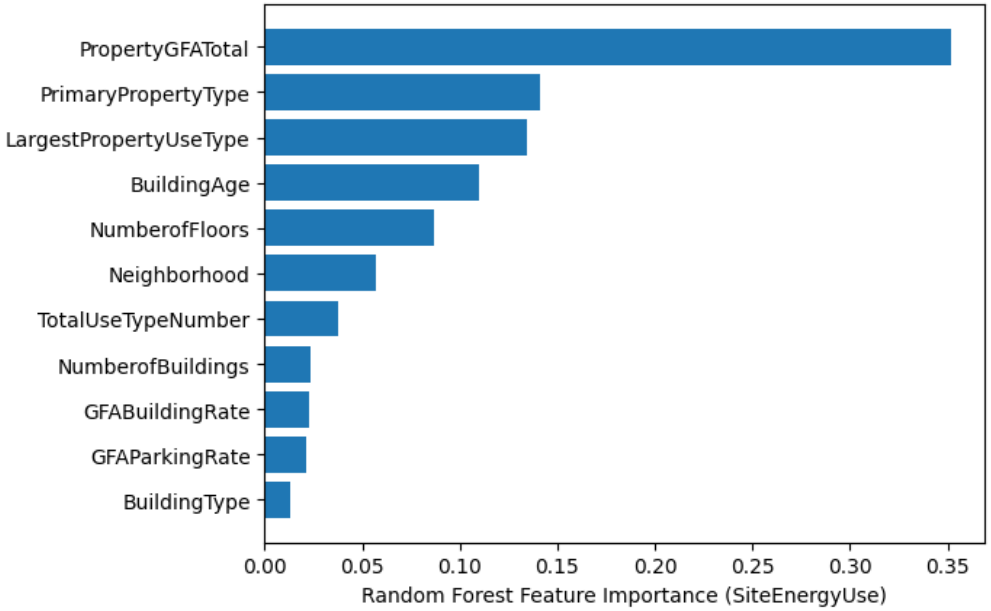
PERFORMANCES DES MODÈLES (VALIDATION CROISÉE)

Modélisations sur la variable SiteEnergyUse

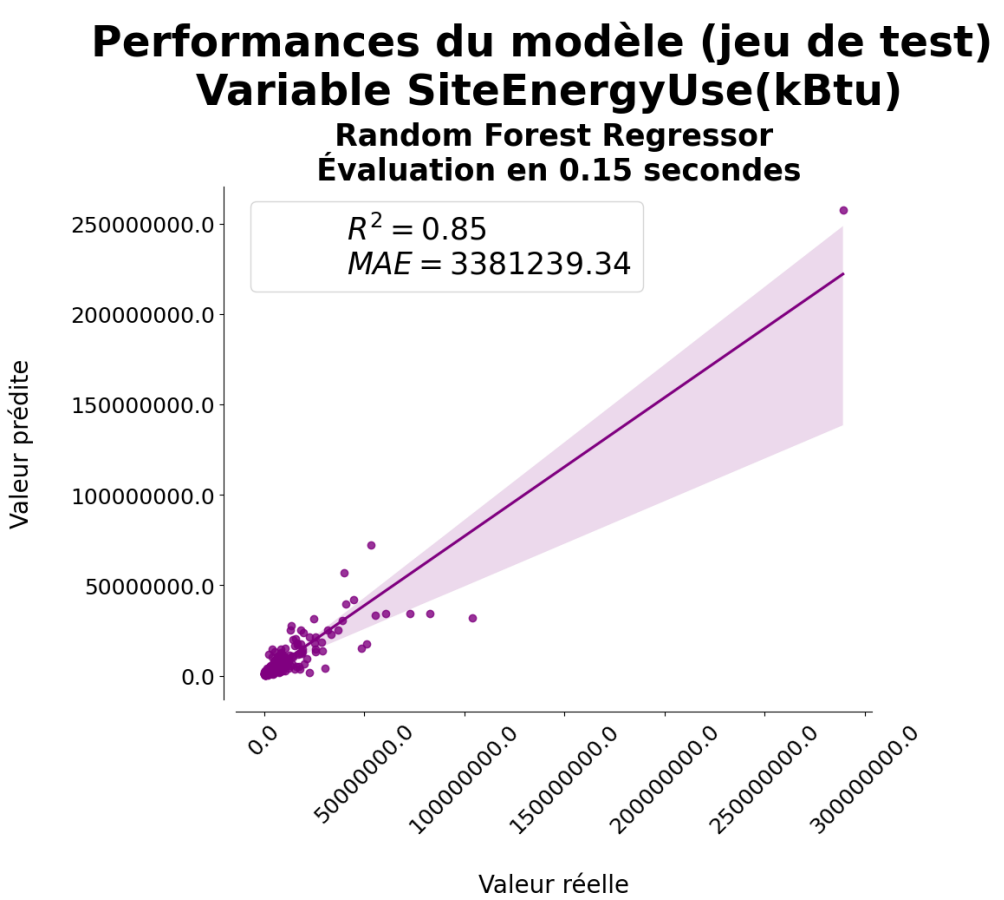


- Le modèle RandomForest offre le meilleur score MAE en terme de prédiction
- Le modèle KNeighborsRegressor offre le meilleur score MAE en terme d'entrainement
- L'entrainement et la prédiction de RandomForest prend plus de temps
- **Modèle sélectionné: RandomForest**

FEATURE IMPORTANCE



- La surface totale de la propriété a une importance bien supérieure aux autres variables.
- Le type de bâtiment a un impact très limité.



	MAE	R2	Time
Train	3.890904e+06	0.607450	2397.470000
Test	3.381239e+06	0.849294	0.151598

EVALUATION DU MODÈLE SUR LE JEU DE TEST

- Les métriques sur le jeu de données de test sont améliorées comparativement aux métriques obtenues avec la GridSearch avec le modèle de RandomForestRegressor
- Pas de sur-apprentissage observé

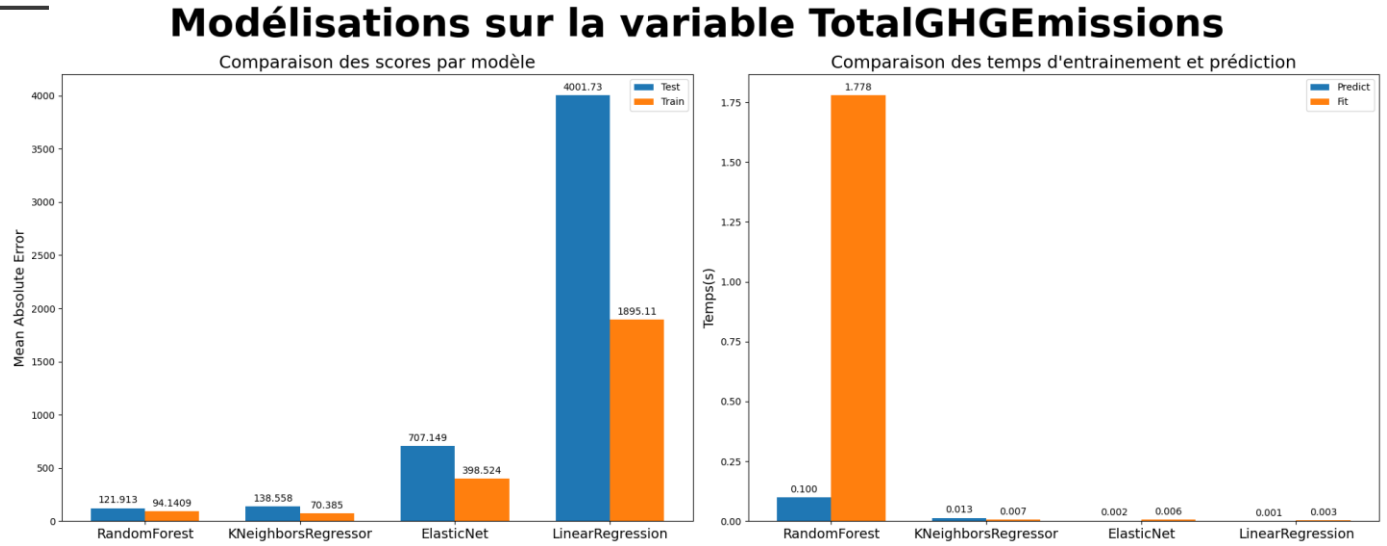
Prédiction des émissions de CO2

Performances des modèles (validation croisée)

Meilleurs scores de la GridSearch (TotalGHGEmissions)

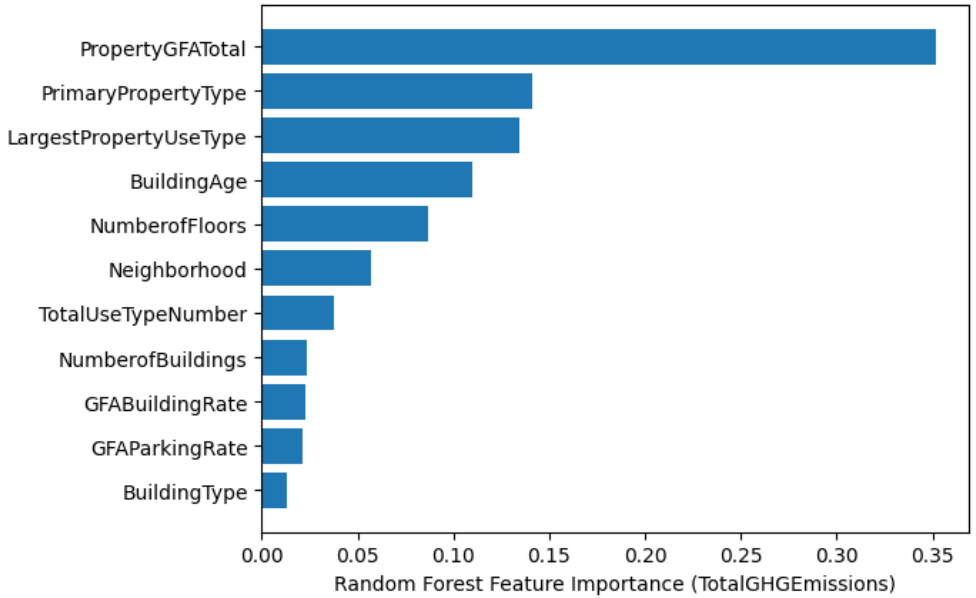
	MAE	R2	Time
RandomForestRegressor	111.662383	0.50954	2238.61
KNeighborsRegressor	132.094119	0.360295	8.88
ElasticNet	826.665054	-26.261464	7.77
LinearRegression	835.673587	-230.401019	3.36
LinearSVR	620539225640.060791	-1137.682772	47.37

Performances des modèles (validation croisée)



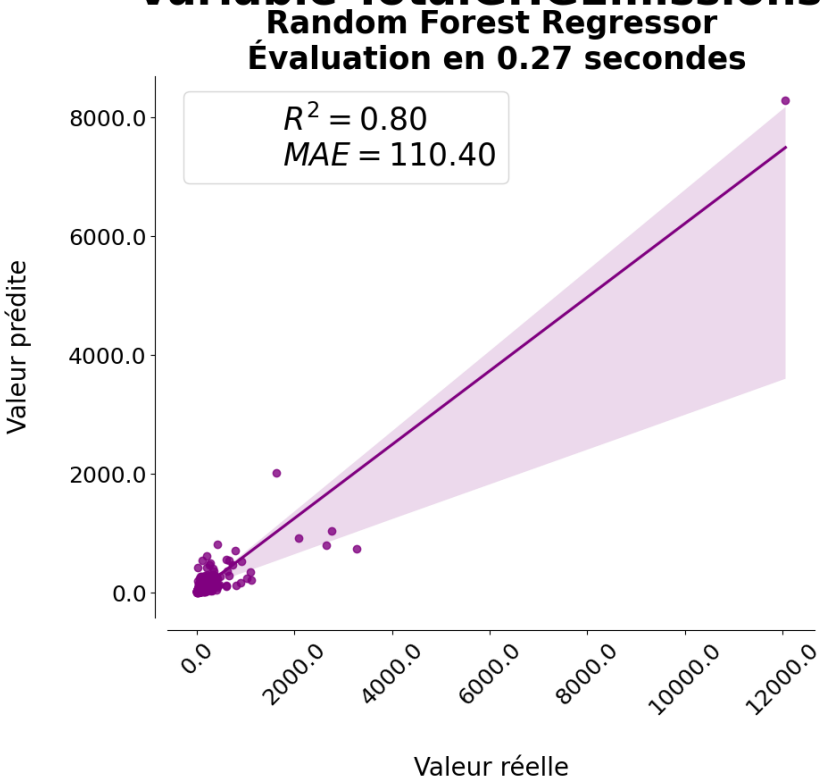
- Le modèle RandomForest offre le meilleur score MAE en terme de prédiction
- Le modèle KNeighborsRegressor offre le meilleur score MAE en terme d'entrainement
- L'entrainement et la prédiction de RandomForest prend plus de temps
- **Modèle sélectionné: RandomForest**

FEATURE IMPORTANCE



- Les surfaces (GFA) de la propriété et le type d'utilisation principale ont un poids plus important dans les décisions de notre modèle.
- En revanche, le type de bâtiment a un impact très limité.

Performances du modèle (jeu de test)
Variable TotalGHGEmissions



	MAE	R2	Time
Train	111.520448	0.509821	2423.450000
Test	110.402443	0.804090	0.269296

EVALUATION DU MODÈLE SUR LE JEU DE TEST

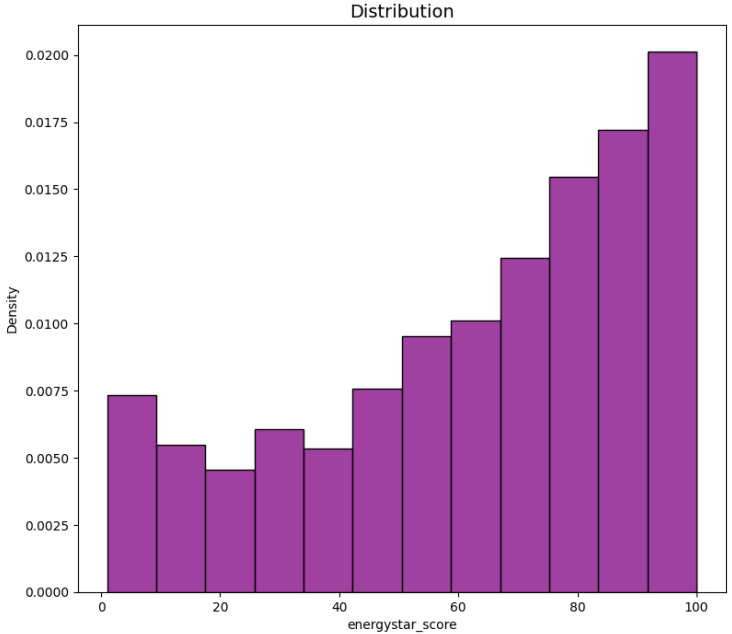
- Les métriques sur le jeu de données de test sont améliorées comparativement aux métriques obtenues avec la GridSearch avec le modèle de RandomForestRegressor
- Pas de sur-apprentissage observé

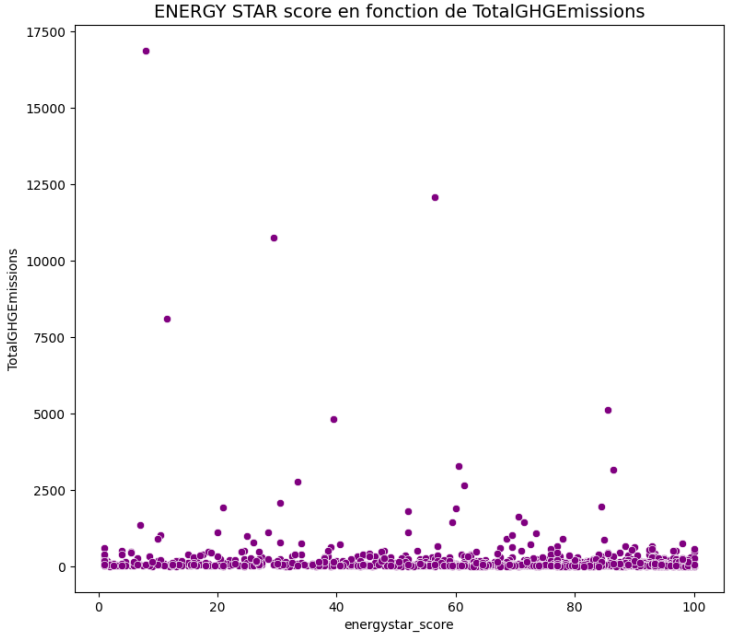
IMPACT DE L'ENERGY STAR SCORE POUR LA PRÉDICTION D'ÉMISSIONS

L'ENERGY STAR SCORE

- Le score ENERGY STAR fournit un aperçu complet de la performance énergétique d'un bâtiment, en tenant compte des actifs physiques, des opérations et du comportement des occupants du bâtiment
- Il est exprimé sur une échelle de 1 à 100 facile à comprendre
- plus le score est élevé, meilleure est la performance énergétique du bâtiment.
- Nous allons évaluer si ce score à un impact significatif sur les performances de notre modélisation

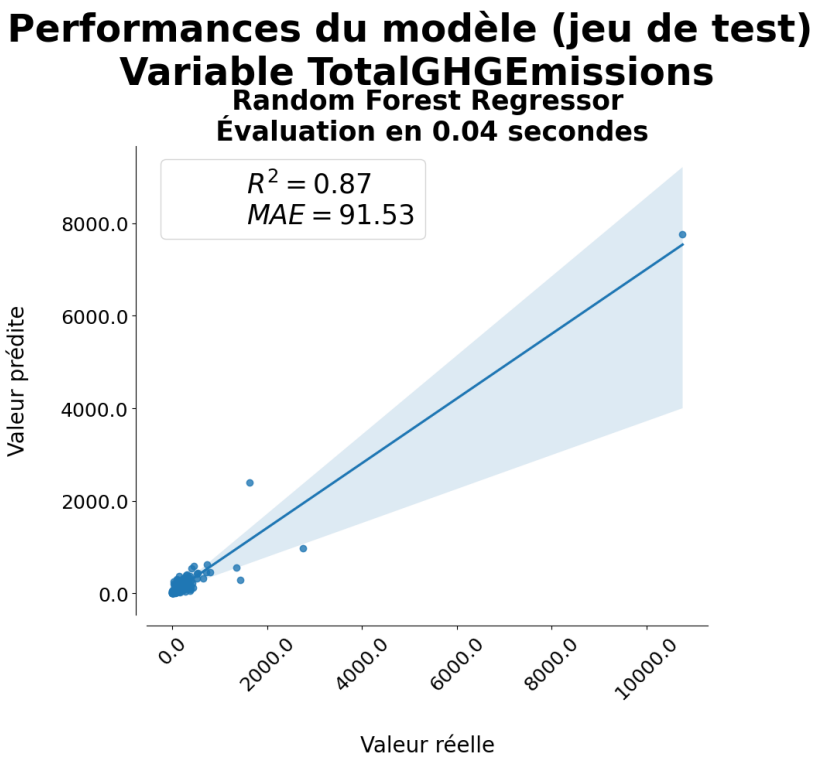
Analyse de la variable ENERGY STAR Score



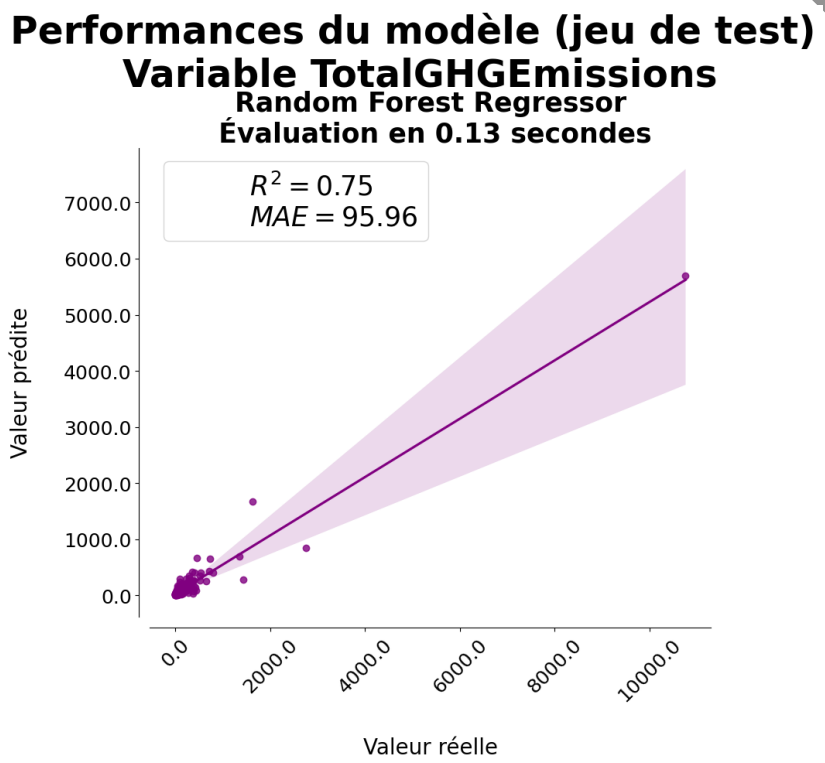


- La distribution ne suit pas de loi normale et la majorité des bâtiments a un score supérieur à 50
- Le score ENERGY STAR ne semble pas avoir de corrélation importante avec les émissions de CO2

SANS ENERGY STAR



AVEC ENERGY STAR



- L'ajout de l'Energy Star score n'améliore pas forcément les performances du modèle

05

Conclusion et perspectives

Conclusion et perspectives

Ce projet nous a permis d'appréhender des notions importantes telles que :

- La mise en place d'un modèle d'apprentissage supervisé adapté au problème métier
- L'évaluation des performances d'un modèle d'apprentissage supervisé
- Le manipulation des paramètres des modèles
- La transformation les variables pertinentes d'un modèle d'apprentissage supervisé

Perspectives:

- Tester d'autres modèles (ex: XGBoost)

MERCI

Des questions ?