City of Seattle

Prédiction de **consommations** et **d'émissions** des bâtiments

Guilhem Berthou - Pierre-Antoine Ganaye (mentor)

Introduction

Problématique : Prédire **la consommation totale d'énergie** et les **émissions de CO2** des bâtiments **non destinés à l'habitation** de la ville de **Seattle**.

Relevés *2015* et *2016*

`SiteEnergyse(kBtu)`

`TotalGHGEmissions`

Variables à prédire

ENERGY STAR Score

⇒ **Interprétation :** Il s'agit d'un problème de **régression**. Nous allons tenter de **prédire** les **variables cibles** (*target*) à partir des **variables présentes dans la base de donnée** (*features)*.

Sommaire

Sommaire

Préparation de la donnée

➢ **Harmonisation des colonnes** des deux datasets

- Différences de **libellés** :
  - Contrôle des **magnitudes similaires** et modification des libellés
- Différences de **formats** :
  - Séparation des adresses (2015) en latitudes et longitudes

➢ **Suppression** des **données non nécessaires**

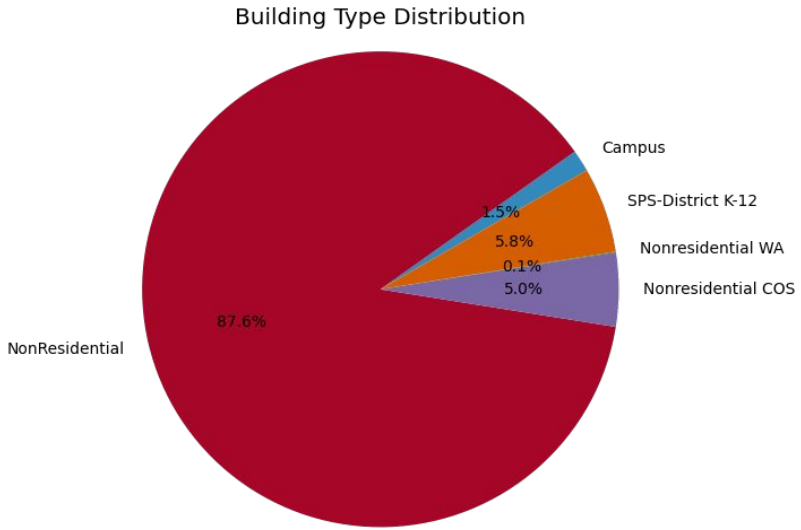
- Suppression **bâtiments résidentiels**
- Suppression des **doublons**
  - Conservation de la dernière valeur disponible
- Suppression des **variables non pertinentes**
  - Variables de description insuffisante (Weather Normalized)
  - Unités redondantes
  - Variables éparses (Comments, outliers)
- Suppression des lignes dont les ***Targets*** sont **manquantes**

⇒ **Conclusion :** Nous obtenons un dataset propre de **1676 lignes** et **39 colonnes** sur lequel nous pouvons commencer l'exploration de la donnée.

Préparation de la donnée

➢ Analyse du **type de bâtiments**

- Les sous-catégories majoritaires sont :
  - Small and Mid-Sized Office (17%)
  - Other (11%)
  - Warehouse (11%)
- Six sous-catégories < 1% :
  - Hôpitaux
  - Laboratoire
  - etc.



⇒ **<u>Conclusion :</u>** Nous supposons de *a priori* que les catégories de bâtiments puissent avoir un impact sur nos cibles. Nous veillerons à ce que certaines données ne soient pas des **outliers**.

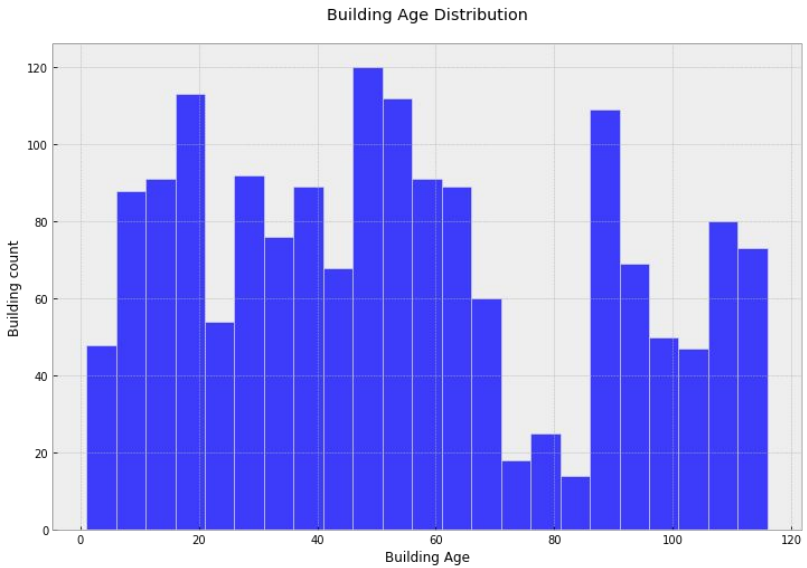
Préparation de la donnée

1. Préparation de la donnée
   a. Nettoyage
   b. **Exploration et Feature Engineering**
   c. Pre-processing

➢ ***Feature Engineering*** :

- Création d'une date d'un **âge de bâtiment**
  - ***BuildingAge***

- Construction d'une variable unique pour prise en compte des **longitudes** et **latitudes** :
  - ***Haversine Distance***

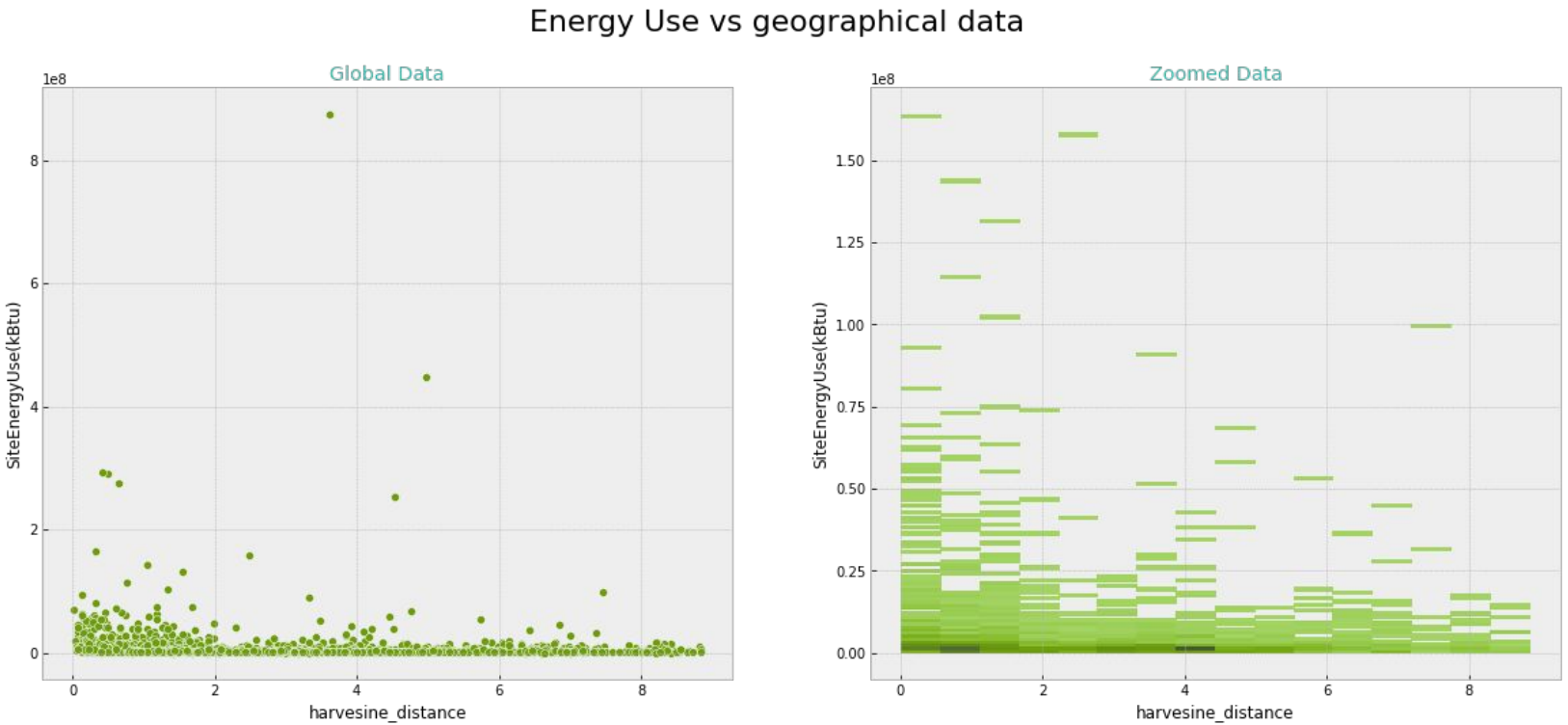


⇒ **<u>Conclusion :</u>** Nous allons maintenant pouvoir analyser nos **variables cibles,** notamment *par rapport à ces variables.*

Préparation de la donnée

➢ ***Analyse bivariée** des **targets** :*

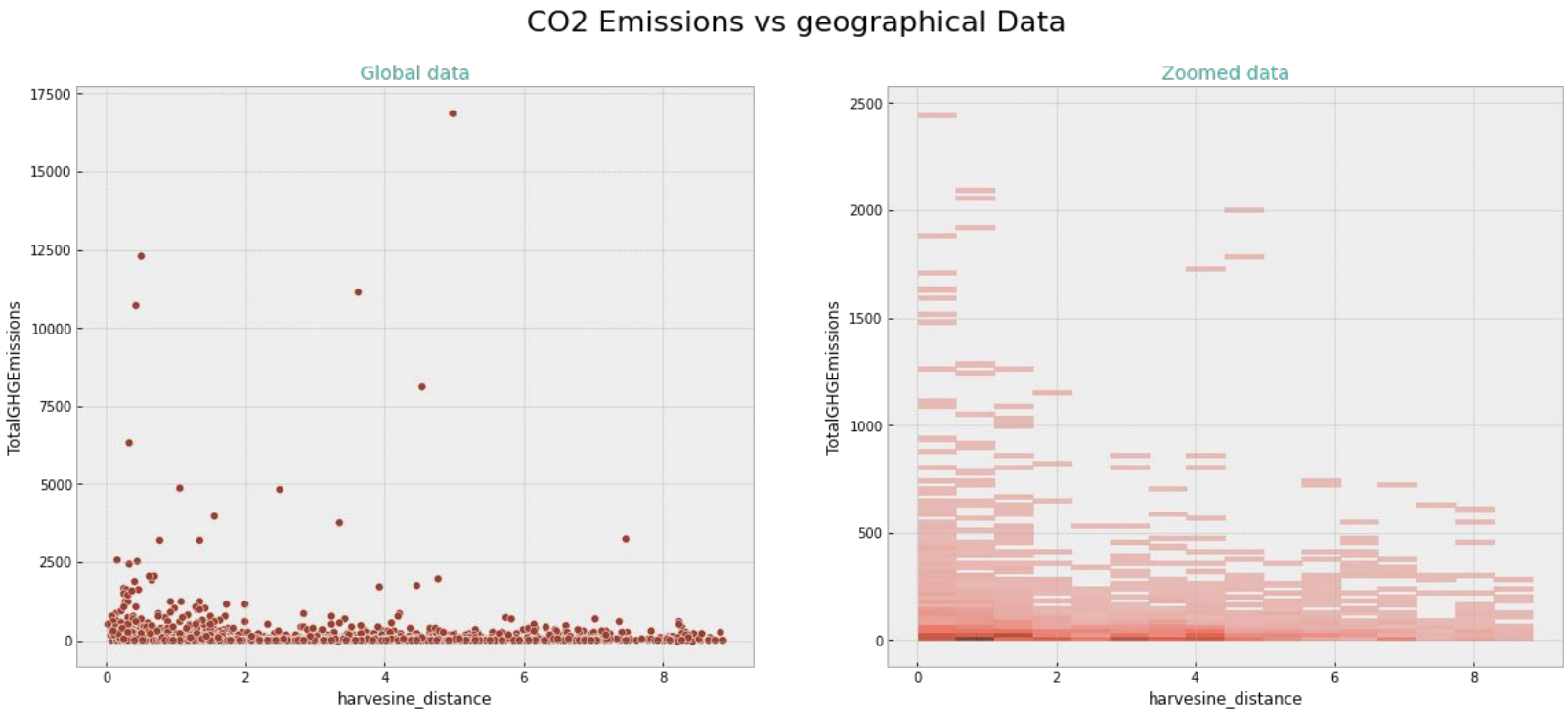
- En fonction de la **Distance de Haversine**



Préparation de la donnée

➢ ***Analyse bivariée** des **targets** :*

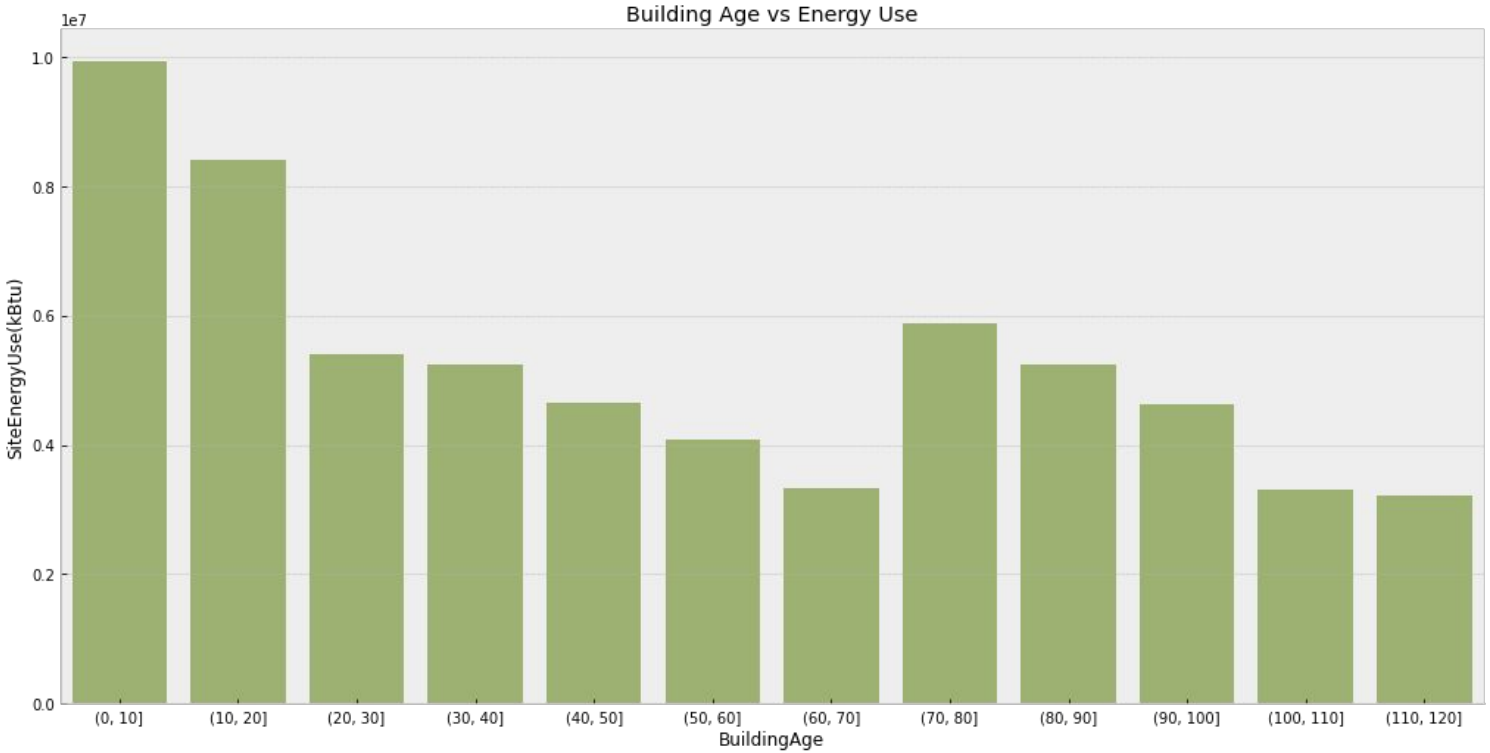
- En fonction de la **Distance de Haversine**



Préparation de la donnée

➢ ***Analyse bivariée** des **targets** :*

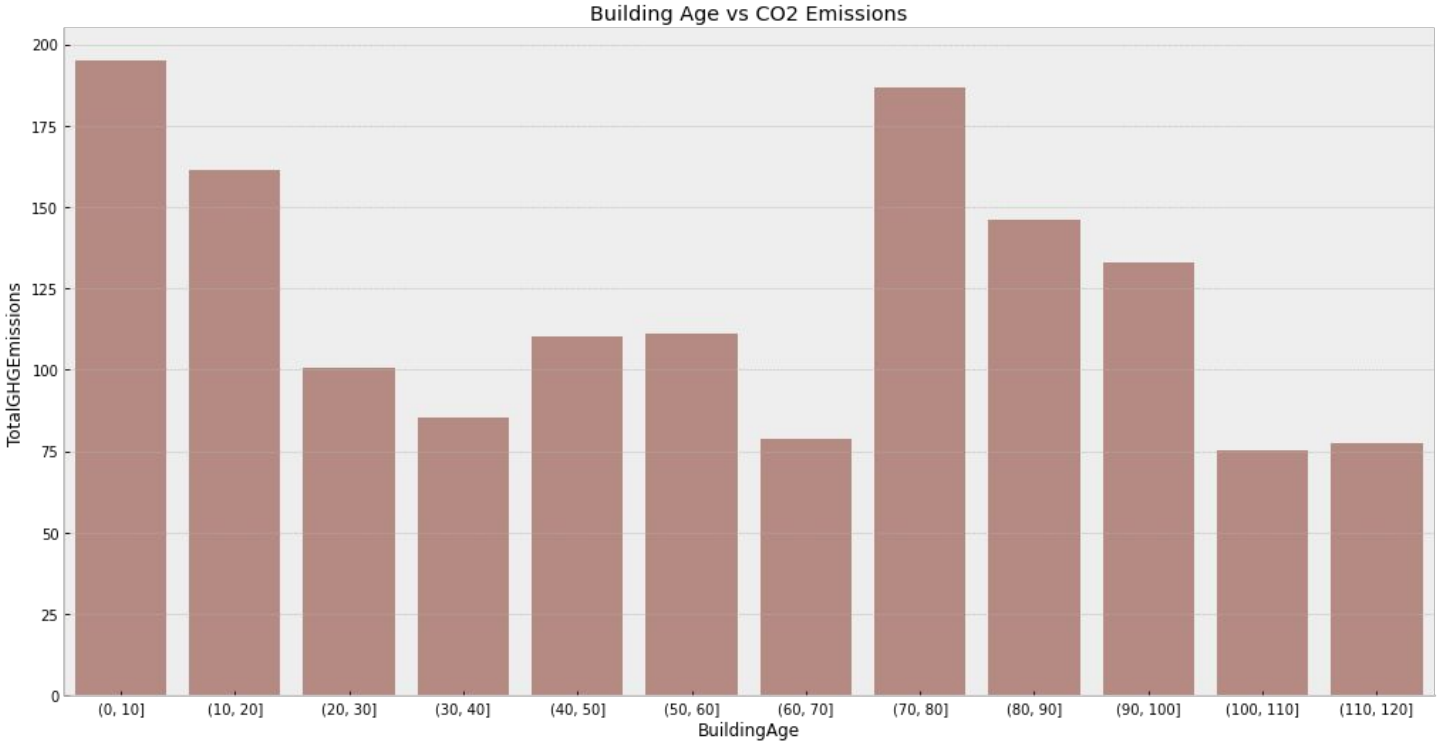
- En fonction de l'âge des bâtiments (**Building Age**)



Préparation de la donnée

➢ ***Analyse bivariée** des **targets** :*

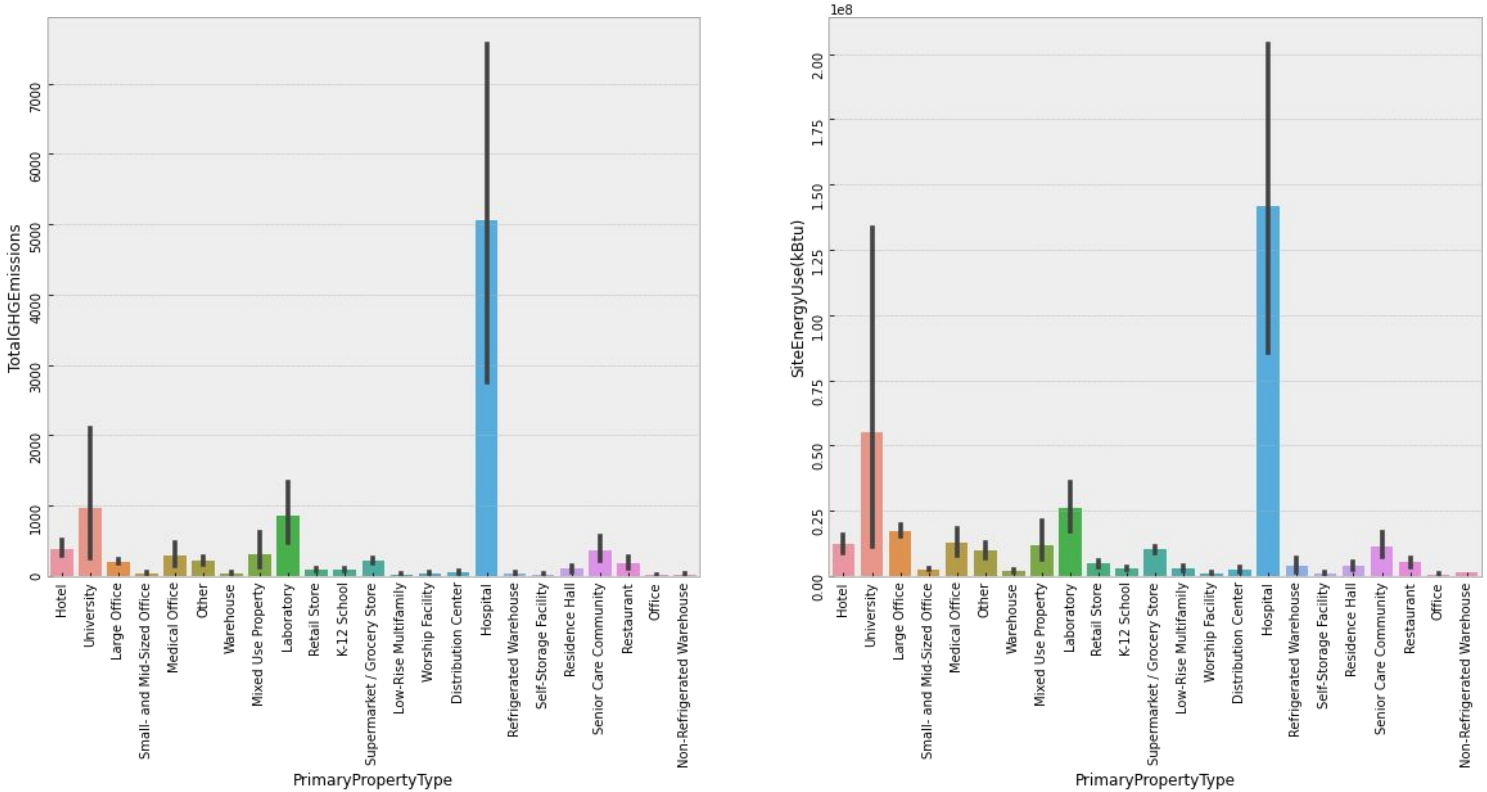
- En fonction de l'âge des bâtiments (**Building Age**)



Préparation de la donnée

➢ ***Analyse bivariée** des **targets** : En fonction du type de bâtiments*

Energy use and CO2 Emissions per PrimaryPropertyType



Préparation de la donnée

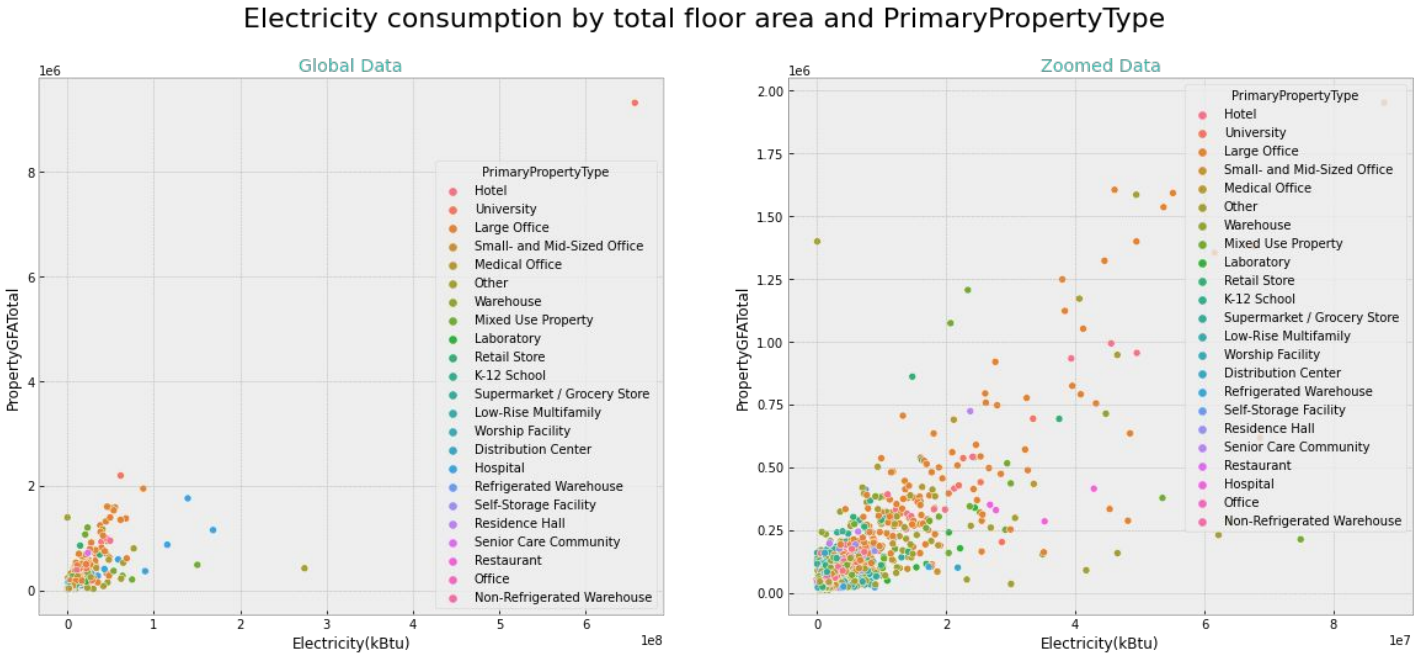
➢ ***Conclusion - Analyse bivariée** des **targets*** :

- **Distance de Harvesine** :
  - *Impact possible, non clairement défini*
- **Building Age :**
  - *Impact possible, non clairement défini*
- **Type de bâtiments :**
  - *Fort Impact de certaines catégories*

⇒ *À prendre en compte dans notre **identification des outliers**.*

Préparation de la donnée

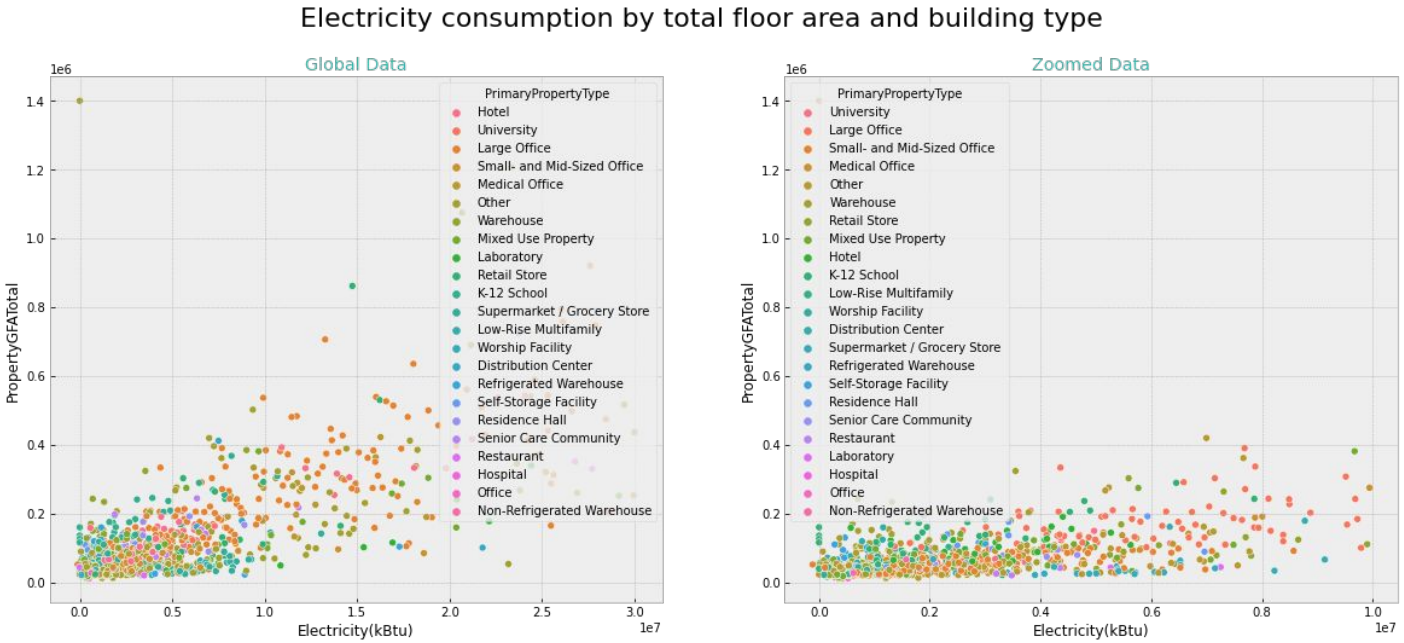
➢ ***Traitement des outliers*** : Analyse Multivariée



⇒ **<u>Conclusion :</u>** Présence d'outliers - nous supprimons les données dont la consommation électrique est > 0.3E8 (kBtu).

Préparation de la donnée

➢ ***Traitement des outliers*** : Analyse Multivariée

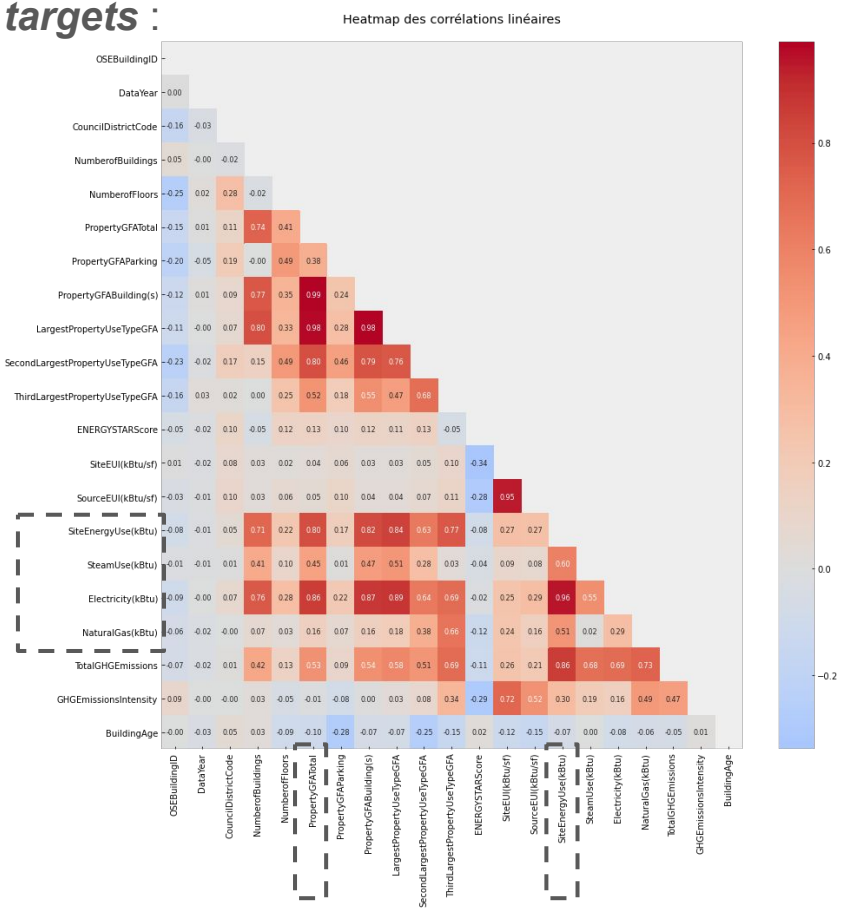


⇒ **<u>Conclusion :</u>** Après suppression des outliers *(54 points - 3% du dataset),* nous obtenons une population plus harmonieuse.

Préparation de la donnée

➢ ***Analyse multivariée** des **targets** :*

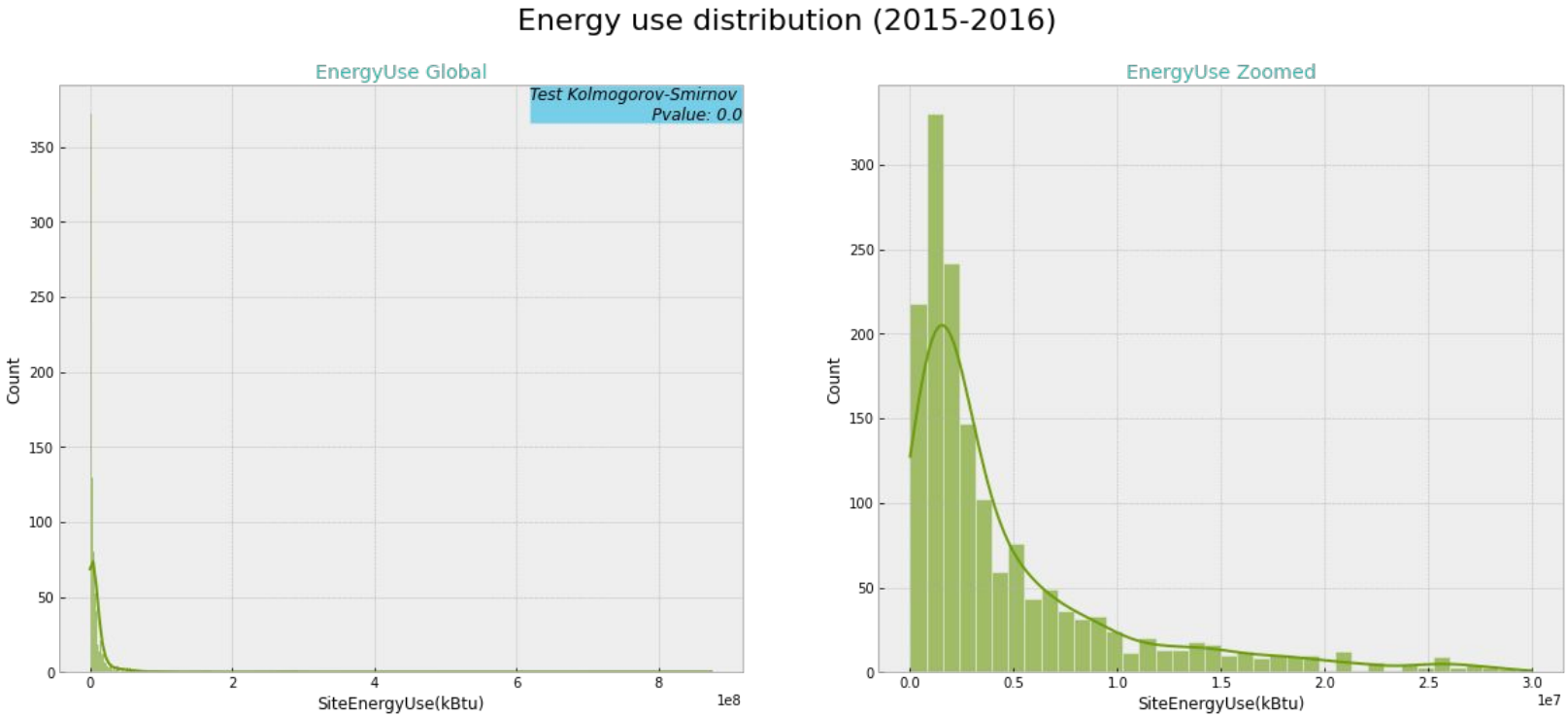
- **Matrice de corrélation**



Préparation de la donnée

➢ ***Analyse univariée** des **targets** :*

- **Test de Normalité** - Kolmogorov-Smirnov

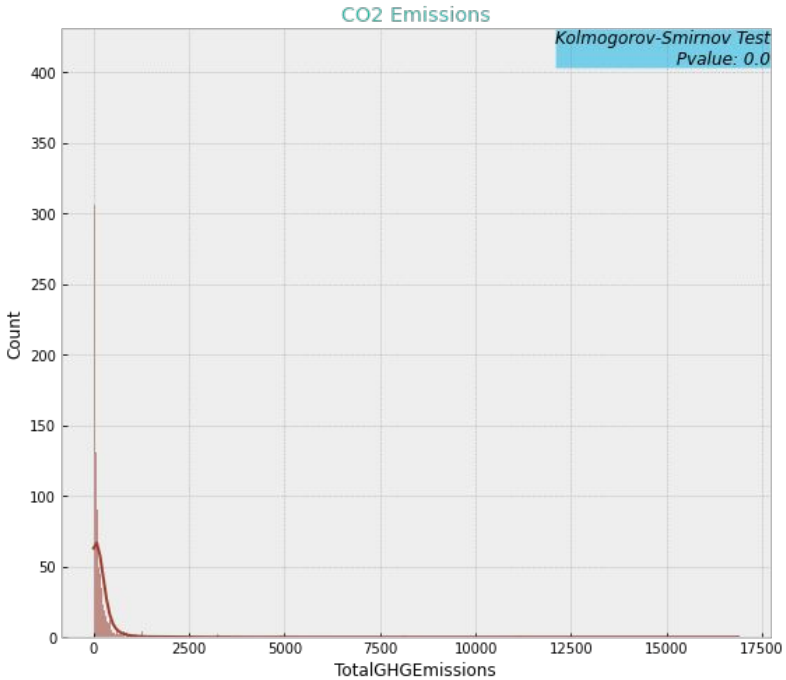


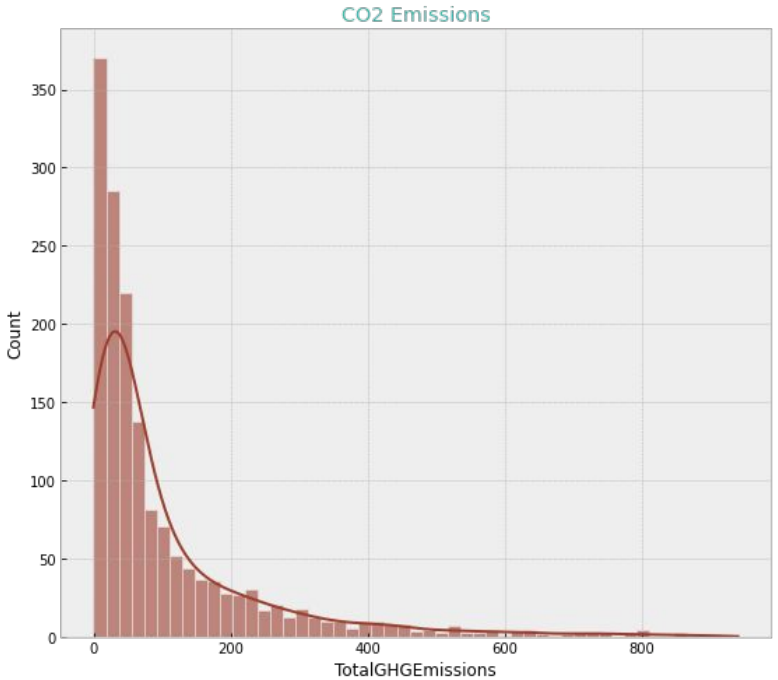
Préparation de la donnée

➢ ***Analyse univariée** des **targets** :*

- **Test de Normalité** - Kolmogorov-Smirnov

CO2 emissions distribution (2015-2016)





Préparation de la donnée

➢ ***Conclusion*** :

- **Analyse multivariée - Matrice de corrélation :**
  - **Suppression des données de relevés** nécessaire pour éviter la **fuite de données.**
  - La **transformation** de ces données en ***variables plus simples*** est à considérer.

- **Analyse univariée - Test de Normalité** :
  - **Standardisation** nécessaire pour obtenir des *features* plus **homogènes**.

⇒ *A prendre en compte dans notre **pre-processing**.*

Préparation de la donnée

➢ ***Standardisation des données*** :

- Distribution **centrées-réduites** via StandardScaler (module preprocessing - scikit-learn)
  - *Modalités détaillées dans la méthodologie de modélisation (cf 2.)*

➢ ***Prévention de la fuite de donnée :***

- **Binarisation** des sources d'énergie
- **Suppression** des **sources d'énergie**

➢ ***Encodage des variables catégorielles :***

- **OneHotEncoder** (Get_dummies)
- **Mean Target** Encoding
  - *Modalités détaillées dans la synthèse des améliorations réalisées (cf 3.)*

⇒ **Conclusion :** Nous obtenons un dataset préparé de **1520 lignes** et **15 colonnes** sur lequel nous pouvons commencer la ***modélisation***.

Sommaire

Méthodologie de modélisation

➢ **Rappel de l'objectif du projet : Prédire** les **variables cibles** (*targets : Consommation d'énergie et Emissions de CO2*) à partir des **variables présentes dans la base de donnée** (*features)*.

➢ **Division de la base** préparée en **jeux d'entraînement / jeux de test :**

- **train_test_split** *(Scikit Learn)*
  - Proportion : **70% - 30%**
  - Harmonisation : stratification via **qcut** (*pandas*)
  - Reproductibilité **: random_state**

➢ **Standardisation des données**

- Données d'entraînement :
  - Apprentissage des paramètres de normalisation sur le *training set*, puis transformation ;
- Données de test :
  - Transformation du *testing set* en appliquant les paramètres appris sur le *training set* ;

➢ **Séparation** de nos ***targets*** en **deux variables uniques**

Méthodologie de modélisation

➢ **Entraînement des modèles :**

- **Modèles linéaires**
  - Régression linéaire
  - Régression Lasso
  - Régression Ridge
- **Modèles non-linéaires**
  - *K-Nearest Neighbours* (KNN)
  - Arbre de décision (*Decision Tree*)
- **Modèles ensemblistes**
  - *Bagging* (*Random Forest*)
  - *Boosting* (LGBM)

Méthodologie de modélisation

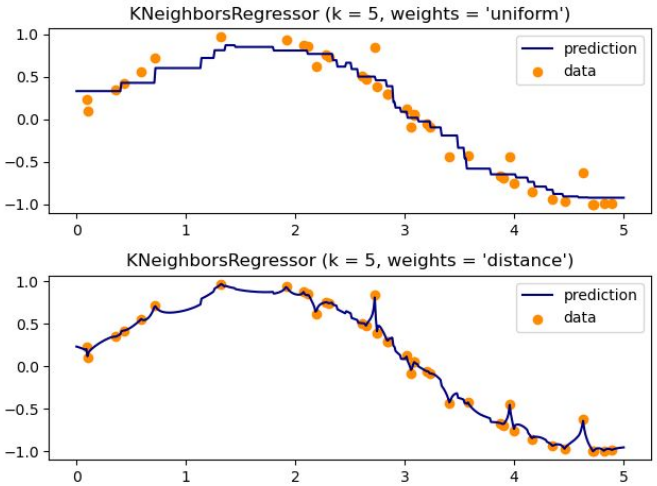
➢ **<u>Entraînement des modèles :</u>**

- **Modèles linéaires**
  - Régression linéaire
    - Explique de manière linéaire, une **variable Y** (variable à expliquer - *target*) en fonction de **variables explicatives X**
  - Régression Ridge
    - Forme de **régularisation** de la régression linéaire (via l'**hyper-paramètre *L2***)
    - Permet d'éviter le sur-apprentissage en **réduisant l'amplitude des coefficients de régression**
      - *<u>Parameters</u> :* **alpha** = facteur de régularisation (*multiplie L2*)
  - Régression Lasso (*Least Absolute Shrinkage and Selection Operator*)
    - Forme de **régularisation** de la régression linéaire (via l'**hyper-paramètre *L1***)
    - Méthode de **sélection des variables** (Modèle parcimonieux)
      - *<u>Parameters :</u>* **alpha** = facteur de régularisation (*multiplie L1*)

Méthodologie de modélisation

➢ **Entraînement des modèles :**

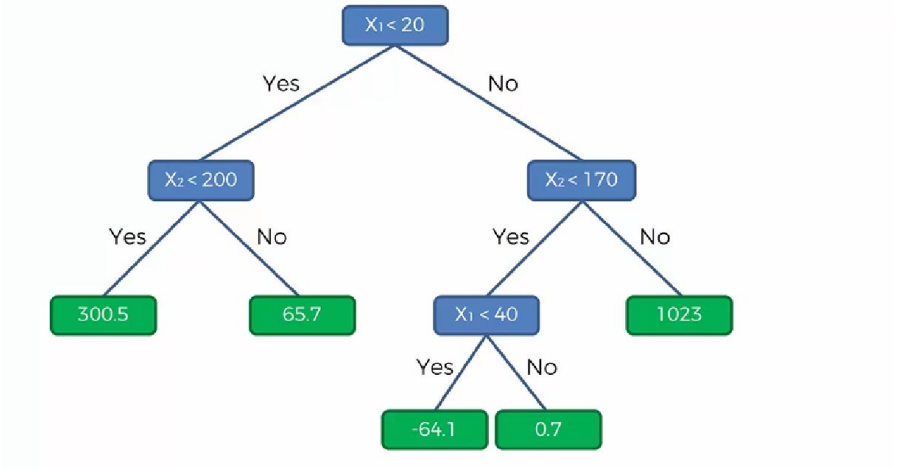
- **Modèles non-linéaires**
  - *K-Nearest Neighbours* (KNN)
    - *Parameters :*
      - *N_neighbors* : nombre de **voisins** considérés pour **l'interpolation locale**
      - *Weights* : (*uniform* vs *distance)* **pondération** des voisins



Méthodologie de modélisation

➢ **Entraînement des modèles :**

- **Modèles non-linéaires**
  - Arbre de décision (*Decision Tree*)
    - *Parameters :*
      - *Max depth*
      - *Min Samples split*
      - *Min Samples leaf*



Méthodologie de modélisation

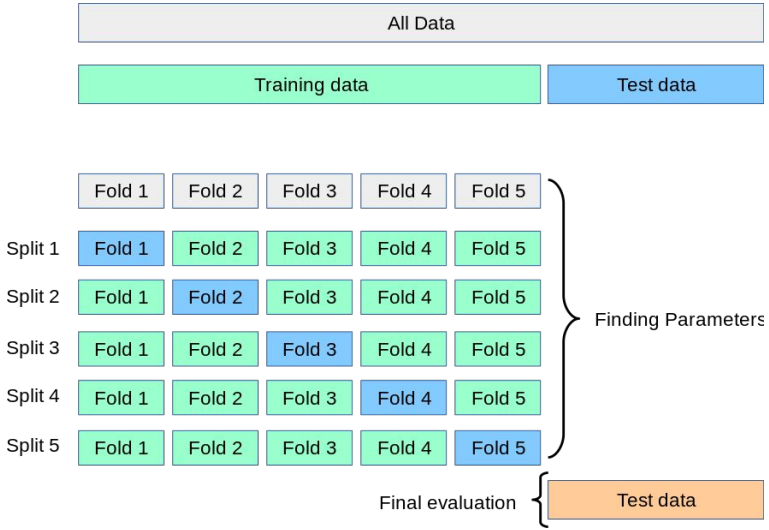
➢ **Entraînement des modèles :**

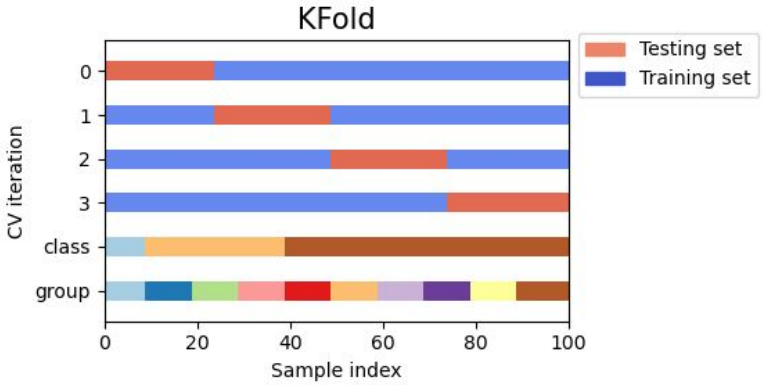
- **Modèles ensemblistes**
  - *Bagging* (*Random Forest*)
    - Créer **plusieurs copies** d'un même modèle en **entraînant** chaque copie sur une **partie *aléatoire*** du dataset (*boostraping*)
  - **Modèles ensemblistes**
    - *Boosting* (LGBM)
      - Entraîner les modèles **à la suite** des autres en leur demandant de **corriger les erreurs** de leurs **prédécesseurs**.

Méthodologie de modélisation

➢ Optimisation des hyperparamètres des modèles :

- Calcul de performance via *Cross Validation*
  - Evaluer la **performance** de **généralisation** d'un modèle

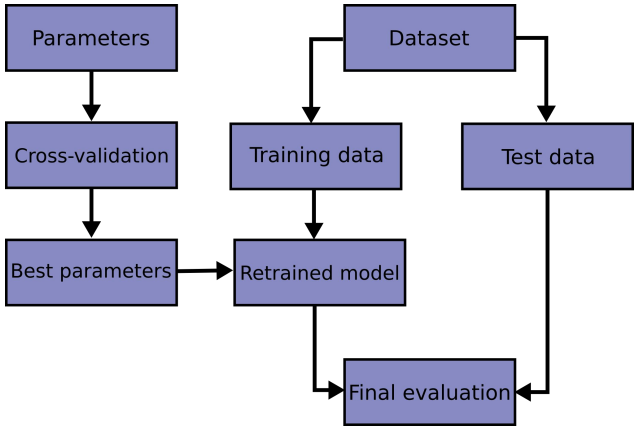




Méthodologie de modélisation

➢ **Optimisation des hyperparamètres des modèles :**

- Recherche des **hyperparamètres** via *GridSearchCV* :
  - Définition d'une **plage de valeurs possibles** pour les **hyperparamètres** (*grid*)
  - Evaluation des **performances** des modèles pour chaque **combinaison d'hyperparamètres**
    - Calcul de la performance réalisé par Cross Validation (nombre de split par défaut, *cv = 5*)



Méthodologie de modélisation

➢ **Métriques d'évaluation de la performance des modèles :**

- Module ***metrics*** de *sklearn*
  - *Mean Absolute Error* **(MAE)** *I Mean Absolute Percentage Error* **(MAPE)**

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

  - *Mean Squared Error (**MSE**) / Root Mean Squared Error (**RMSE**)*

$$MSE = \frac{1}{m}\sum (y_{vrai} - y_{pred})^2$$

Méthodologie de modélisation

➢ **Métriques d'évaluation de la performance des modèles :**

- **Coefficient de détermination (R Squared)**
  - Évalue la performance du modèle par rapport au **niveau de variation** présent dans les **données**
    - Numérateur = ***Erreur quadratique***
    - Dénominateur = ***Variance***

$$R2 = 1 - \frac{\sum(y_{vrai} - y_{pred})^2}{\sum(y_{vrai} - \overline{y_{vrai}})^2}$$

  - **Valeurs possibles**
    - **R2 = 1**
      - **Erreurs** commises par le modèle **<< Variance** des données
    - **R2 = 0**
      - Modèle **prédit** la **moyenne** (indépendant des *features*)
    - **R2 < 0**
      - **Erreurs** commises par le modèle **>> Variance** des données

Sommaire

Méthodologie de modélisation

➢ Comparaison des résultats

Target	Model	r2_test	mae_test	mape_test
SiteEnergyUse(kBtu)	LinearRegression()	0.68	2,193,562.93	0.69
SiteEnergyUse(kBtu)	Ridge()	0.68	2,191,454.62	0.68
SiteEnergyUse(kBtu)	Lasso(max_iter=2000, tol=0.1)	0.68	2,193,559.37	0.69
SiteEnergyUse(kBtu)	DecisionTreeRegressor()	0.68	2,165,096.37	0.60
SiteEnergyUse(kBtu)	KNeighborsRegressor()	0.55	2,193,420.51	0.55
SiteEnergyUse(kBtu)	SVR()	-0.14	4,140,205.72	1.42
SiteEnergyUse(kBtu)	LGBMRegressor()	0.71	1,890,701.11	0.59
SiteEnergyUse(kBtu)	RandomForestRegressor()	0.73	1,903,699.17	0.52

Target	Model	r2_test	mae_test	mape_test
TotalGHGEmissions	Ridge()	0.41	77.37	2.05
TotalGHGEmissions	Lasso(max_iter=2000, tol=0.1)	0.40	78.08	2.10
TotalGHGEmissions	DecisionTreeRegressor()	0.61	62.75	0.98
TotalGHGEmissions	KNeighborsRegressor()	0.30	65.30	0.99
TotalGHGEmissions	SVR()	0.07	77.59	0.97
TotalGHGEmissions	LGBMRegressor()	0.56	57.53	1.14
TotalGHGEmissions	RandomForestRegressor()	0.58	58.65	1.03

➢ **Conclusion :** Le **Random Forest** réalise la **meilleure** qualité de **prédiction**.

La **hiérarchie** des performances des **modèles** nous **alerte** sur la cohérence des résultats.

Méthodologie de modélisation

➢ **Prévention du sur-apprentissage**

- **Validation croisée** *(Cross Validation)*
- **Suppression** des **variables** les **moins importantes***(features selection)*

➢ **Détection du sur-apprentissage**

- Comparaison des performances obtenues sur le *training set* et *testing set* :

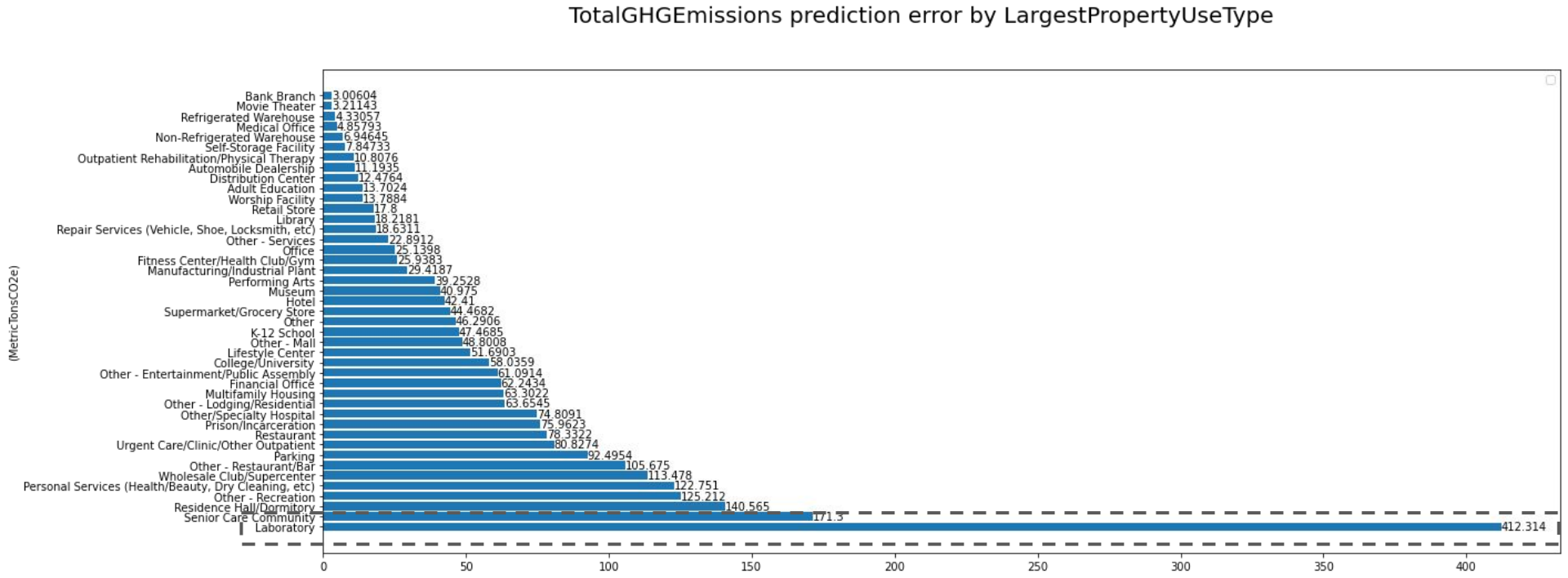
Target	Model	r2_train	r2_test	mae_test	mape_test
SiteEnergyUse(kBtu)	RandomForestRegressor()	0.938373	0.816616	1.514868e+06	0.433147

➢ **Conclusion :** Les performances obtenues sur le testing set sont cohérentes avec celles obtenues sur le jeu d'apprentissage. **L'écart constaté** n'est ***pas trop important*** et ne traduit **pas** un trop fort **sur-apprentissage**.

Méthodologie de modélisation

➢ Sélection du meilleur modèle - Améliorations réalisées

- Analyse d'erreur par type de bâtiments



Méthodologie de modélisation

➢ Sélection du meilleur modèle - Améliorations réalisées

- ***One Hot Encoding***
  - Chaque **modalité** de variable catégorielle est remplacée par une **colonne binaire**

id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding

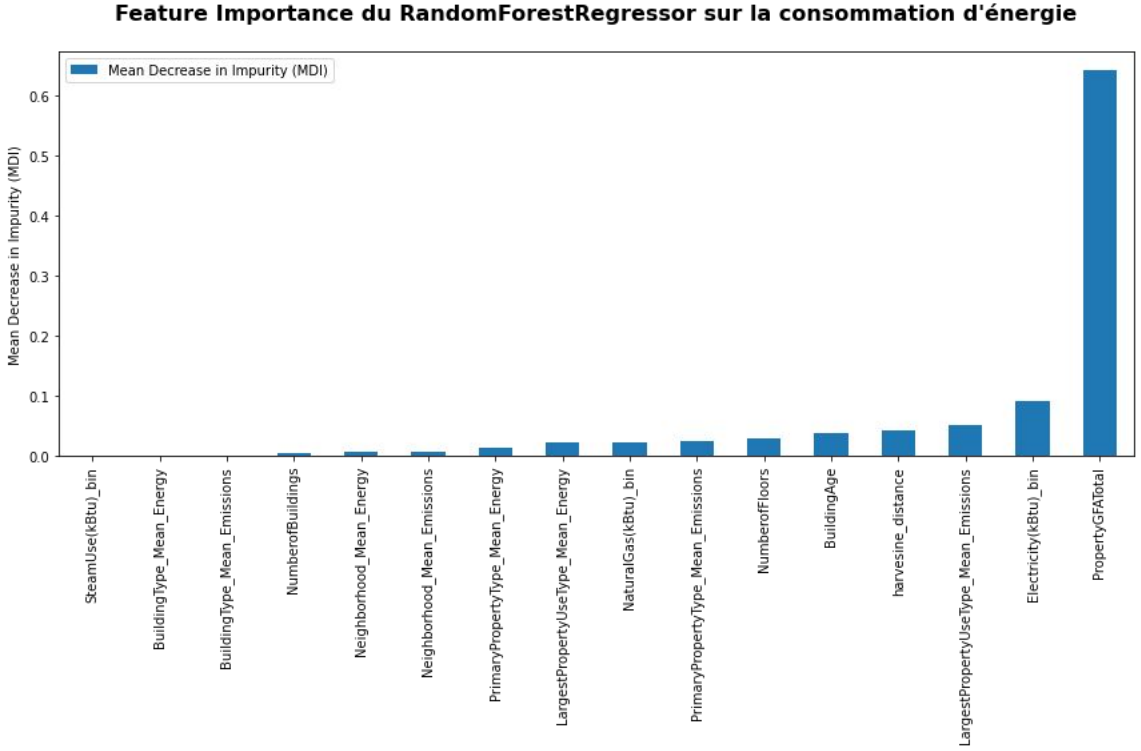
id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

- ***Mean Target Encoding***
  - Chaque variable catégorielle est remplacée par une **moyenne de la *target*** étant donné la **modalité** considérée sur le *training set.*

Méthodologie de modélisation

➢ Sélection du meilleur modèle - Améliorations réalisées

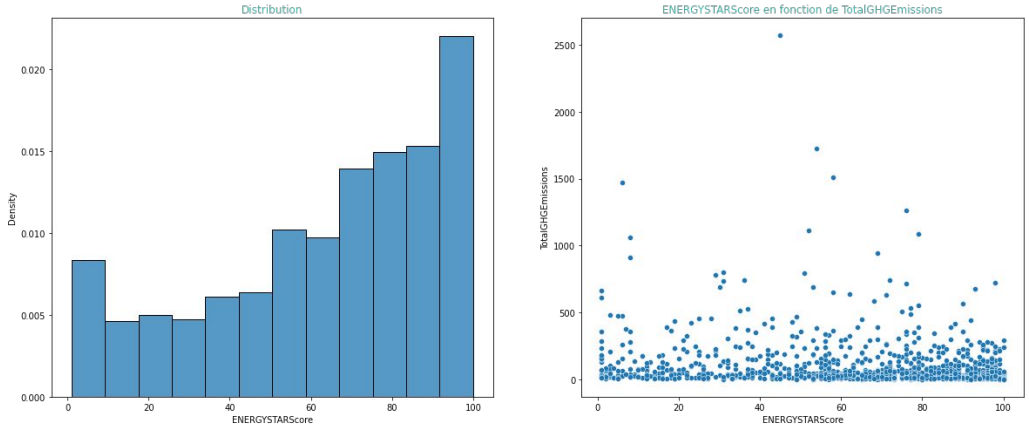
- Analyse de la Feature Importance



Méthodologie de modélisation

➢ Impact de l'ENERGYSTAR Score

Analyse de la variabe ENERGYSTAR Score



- Performances obtenues :

Target	Model	r2_test	mae_test	mape_test
TotalGHGEmissions	RandomForestRegressor()	0.49514	60.289237	1.351087

without ENERGYSTAR Score

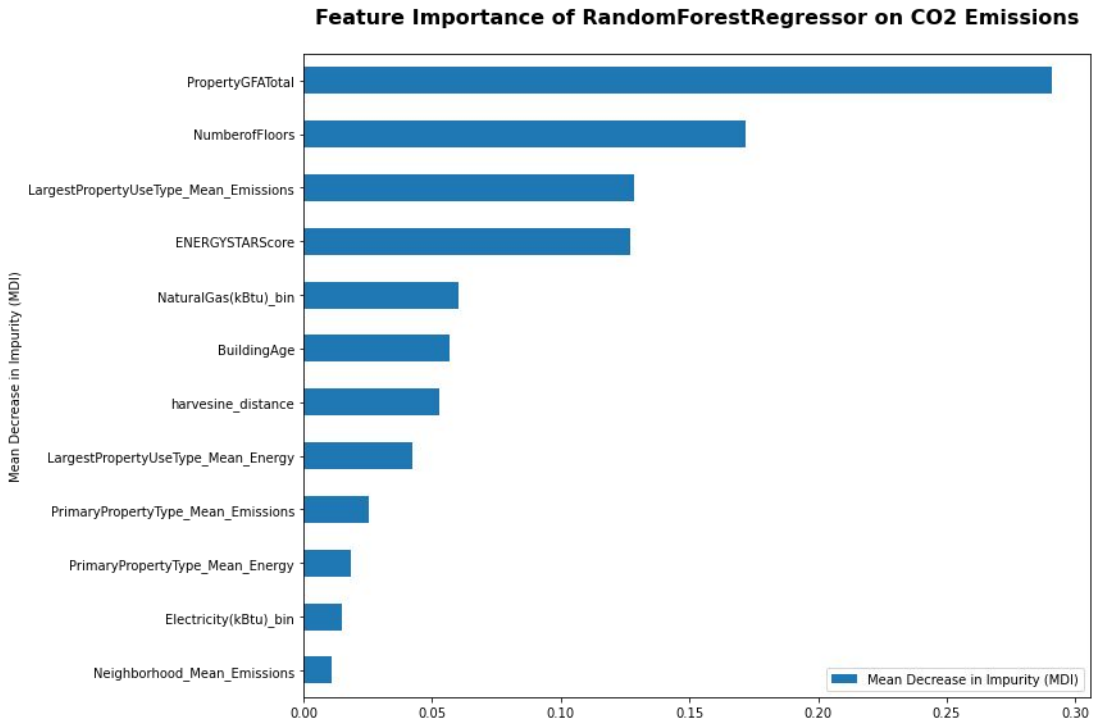
→

Target	Model	r2_test	mae_test	mape_test
TotalGHGEmissions	RandomForestRegressor()	0.506062	59.153222	1.307786

with ENERGYSTAR Score

➢ **Conclusion :** L'ENERGYSTAR Score permet **d'améliorer** la qualité de prédiction du modèle.

Méthodologie de modélisation



➢ **<u>Conclusion</u>** : La modification de l'**importance des features** confirme l'impact positif de l'ENERGYSTAR Score.

Conclusion

- Les **modèles testés** permettent de prédire avec une **précision limitée** les variables cibles

- **L'impact positif** de **l'ENERGYSTAR** Score

- La **hiérarchie des performances** des modèles obtenue nous alerte :
  - Rassembler **plus de données** pour valider l'exactitude des prédictions
  - **Ajouter des métriques** pour analyser avec plus de précision la **génération d'erreurs**

Synthèse des remarques examinateur

- Attention à la **traduction dans l'utilisation finale** lors de la **création de variables** *(Feature Engineering)* :

    - Définir des ***règles de suppression des outliers*** plus **précises** *(ex : type de bâtiments, surface, nombre de bâtiments par catégorie)* afin de pouvoir produire une ***notice explicative claire*** au client
        - Définir des règles évidentes qui *permettraient d'exclure dans tous les cas* des points du dataset
            - *Ex : Le modèle est utilisable tant que les bâtiments ne sont pas des hôpitaux ou des datacenters.*

    - ***Variable binaire*** créée **non-utilisable** en pratique (son calcul nécessite de connaître la distribution de la variable sur la période et donc de faire des relevés) : **A SUPPRIMER**
    -

- **Analyse du sur-apprentissage à réaliser**