Formation de Data Science - Openclassrooms

Formation Ouverte et à Distance – FOAD par Pôle Emploi

Solutions 100% à distance

Projet 4 : Anticipez les besoins en consommation électrique de bâtiments

Étudiant : Maria Daniela Barrios

Mentor : Dan Slama

Février, 2022

Contexte du problème

- ➔ Le projet consiste à créer des stratégies basées sur des données pour la ville de Seattle, aux États-Unis. Pour atteindre l'objectif d'une ville neutre en carbone d'ici 2050, l'équipe porte une attention particulière aux émissions des bâtiments non résidentiels

- ➔ Mission : prédire les **émissions de CO_2** et la **consommation totale d'énergie** de bâtiments pour lesquels elles n'ont pas encore été mesurées.

- ➔ Les données de consommation peuvent être téléchargées à cette adresse : https://www.kaggle.com/city-of-seattle/sea-building-energy-benchmarking#2015-building-energy-benchmarking.csv

- ➔ Les prédictions seront basées sur les données déclaratives du permis d'exploitation commerciale (taille et utilisation des bâtiments, mention de travaux récents, date de construction, etc.)

Stratégie pour réaliser la mission

➔ Phase pré-exploratoire : Analyse générale et découverte des fichiers
- Décrire les informations contenues dans l'ensemble de données: nombre de lignes et de colonnes
- **Sélection et explication d'indicateurs (variables) pertinents**

➔ Analyse exploratoire et nettoyage des données
- Exploration des valeurs manquantes et **nettoyage** des données
- **Création de nouvelles variables - feature engineering**
- Analyse de la corrélation des variables
- Détection des valeurs aberrantes et standardisation des données

➔ Analyse des prédictions
- Test de deux algorithmes de prédiction différents: Ordinary Least Squares Linear Regression et Random Forest Regressor
- Prédiction de la consommation et des émissions de CO_2
- **Déterminer l'influence du ENERGYSTARScore dans les prévisions d'émissions de gaz**

Phase pré-exploratoire : Analyse générale et découverte des fichiers

Deux fichiers contenant :

2015-building-energy-benchmarking.csv :

- **3340** lignes et **47** colonnes

2016-building-energy-benchmarking.csv :

- **3376** lignes et **47** colonnes

Les fichiers de données des années 2015 et 2016 sont différents. Nous pouvons vérifier quelles colonnes sont différentes :

Colonnes communes

```
(Index(['OSEBuildingID', 'DataYear', 'BuildingType', 'PrimaryPropertyType',
        'PropertyName', 'TaxParcelIdentificationNumber', 'CouncilDistrictCode',
        'Neighborhood', 'YearBuilt', 'NumberofBuildings', 'NumberofFloors',
        'PropertyGFATotal', 'PropertyGFAParking', 'PropertyGFABuilding(s)',
        'ListOfAllPropertyUseTypes', 'LargestPropertyUseType',
        'LargestPropertyUseTypeGFA', 'SecondLargestPropertyUseType',
        'SecondLargestPropertyUseTypeGFA', 'ThirdLargestPropertyUseType',
        'ThirdLargestPropertyUseTypeGFA', '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)',
        'DefaultData', 'ComplianceStatus', 'Outlier'],
       dtype='object'),
```

Colonnes présentes dans le fichier 2015 et non dans le fichier 2016 :

```
Index(['2010 Census Tracts', 'City Council Districts', 'Comment',
       'GHGEmissions(MetricTonsCO2e)', 'GHGEmissionsIntensity(kgCO2e/ft2)',
       'Location', 'OtherFuelUse(kBtu)', 'SPD Beats',
       'Seattle Police Department Micro Community Policing Plan Areas',
       'Zip Codes'],
```

Colonnes présentes dans le fichier 2016 et non dans le fichier 2015 :

```
Index(['Address', 'City', 'Comments', 'GHGEmissionsIntensity', 'Latitude',
       'Longitude', 'State', 'TotalGHGEmissions', 'ZipCode'],
      dtype='object'))
```

Phase pré-exploratoire : Analyse générale et découverte des fichiers

Comme la mission du projet concerne les émissions de CO_2, nous devons accorder une attention particulière aux variables :

- **GHGEmissionsIntensity** : total des émissions Greenhouse Gas divisé par la surface brute de la propriété (mesurée en kilogrammes d'équivalent de dioxyde de carbone par pied carré)

- **TotalGHGEmissions** : quantité totale d'émissions de Greenhouse Gases, y compris le dioxyde de carbone, le méthane et l'oxyde nitreux libérés dans l'atmosphère en raison de la consommation d'énergie de la propriété (mesurée en tonnes métriques d'équivalent de dioxyde de carbone)

- **ENERGYSTARScore** : Une note de 1 à 100 qui évalue la performance énergétique globale d'un bien immobilier

- **SiteEUI(kBtu/sf)** : Energy Use Intensity du site (EUI) divisée par sa surface brute

- **SourceEUI(kBtu/sf)** : Energy Use Intensity à la source (EUI) divisée par la surface

- **SiteEnergyUse(kBtu)** : La quantité annuelle d'énergie consommée par la propriété, toutes sources d'énergie comprises

- **Electricity(kBtu)** : La quantité annuelle d'électricité consommée par le bien sur place

- **NaturalGas(kBtu)** : La quantité annuelle de gaz naturel fourni par le distributeur et consommé par la propriété.

https://data.seattle.gov/dataset/2016-Building-Energy-Benchmarking/2bpz-gwpy

Phase pré-exploratoire : Analyse générale et découverte des fichiers

D'autres variables importantes :

- **OSEBuildingID** → Identifiant unique attribué à chaque propriété : utile pour trouver des données doublées
- **DataYear**
- **BuildingType**
- **PrimaryPropertyType**
- **Latitude**
- **Longitude**
- **Address**
- **Neighborhood**
- **YearBuilt**
- **NumberofBuildings**
- **NumberofFloors**
- **PropertyGFATotal**
- **PropertyGFABuilding(s)**
- **PropertyGFAParking**
- **LargestPropertyUseType and LargestPropertyUseTypeGFA**
- **SecondLargestPropertyUseType and SecondLargestPropertyUseTypeGFA**
- **ThirdLargestPropertyUseType and ThirdLargestPropertyUseTypeGFA**

↓

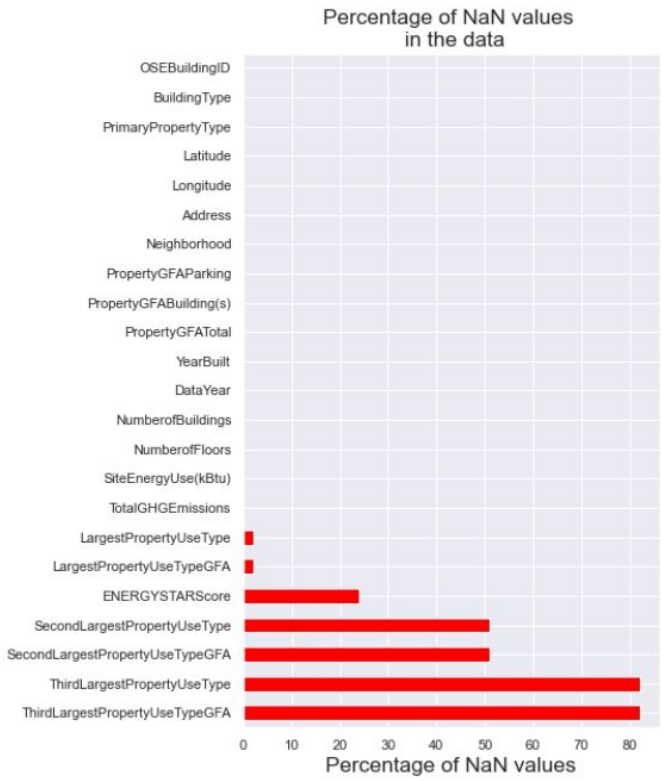
Le nombre de valeurs dupliquées dans les variables "OSEBuildingID" du fichiers pour 2015 et pour 2016 était de 0

Fichier de données unique avec les données des années 2015 et 2016 avec les variables pertinentes contient 6716 lignes et 23 colonnes

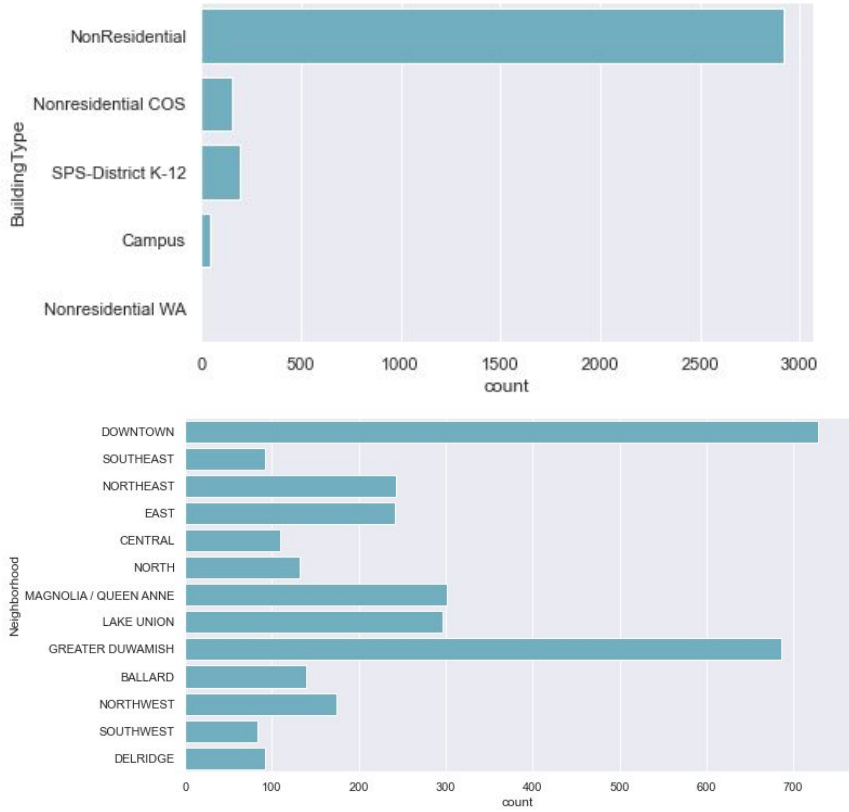
https://data.seattle.gov/dataset/2016-Building-Energy-Benchmarking/2bpz-gwpy

Analyse exploratoire et nettoyage des données

Valeurs manquantes parmi les variables

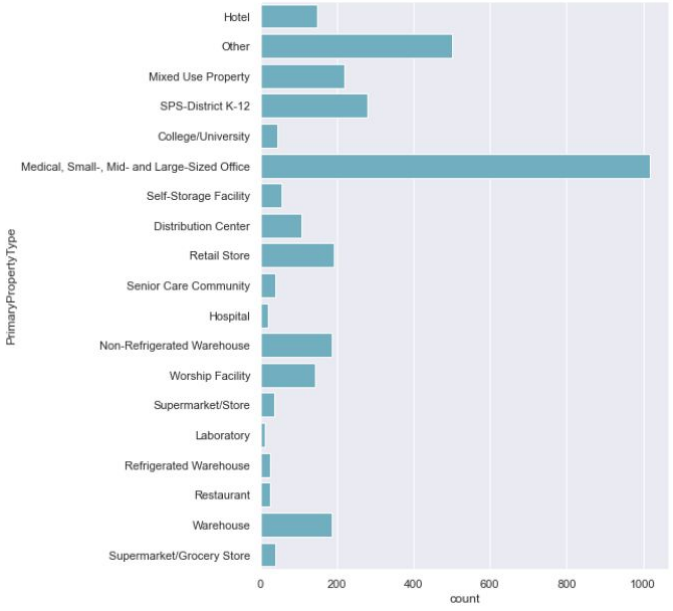


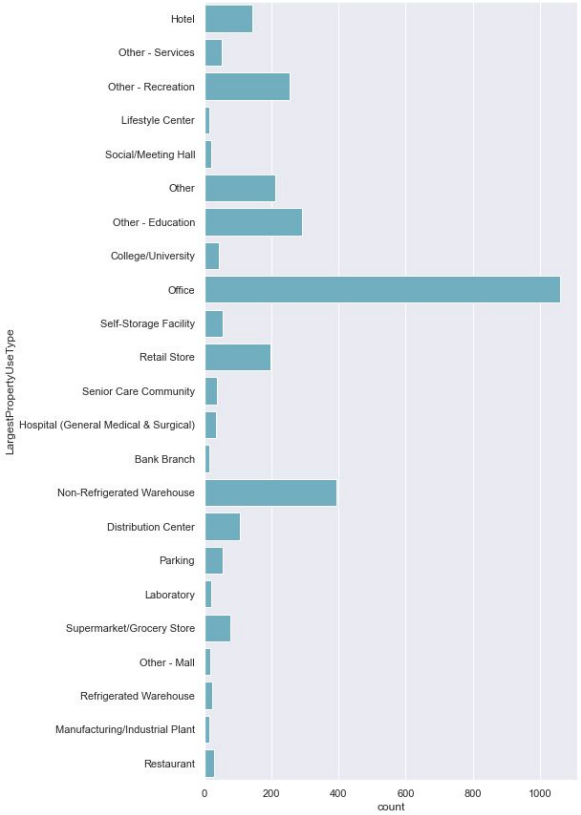
Bâtiments non résidentiels et leurs quartiers

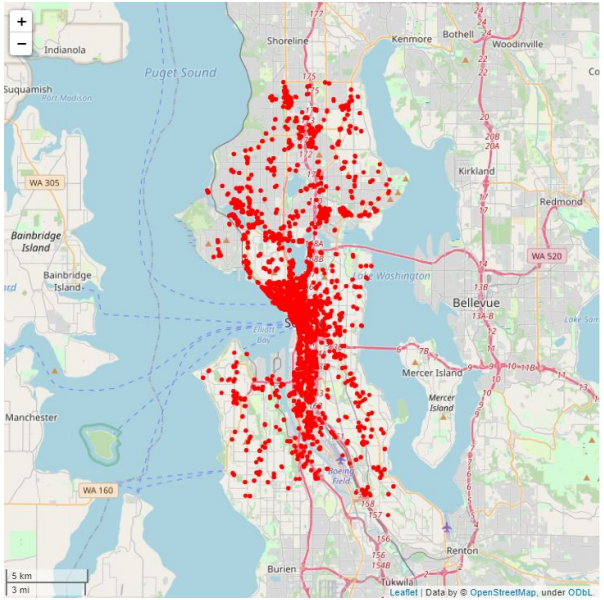


Analyse exploratoire et nettoyage des données

Principal type de propriété, principal type d'utilisation de la propriété et son emplacement sur la carte de Seattle







Analyse exploratoire et nettoyage des données

Création de nouvelles variables - **feature engineering**

- Âge des bâtiments - 'BuildingAge' from ('DataYear' - 'YearBuilt')

- Surface par étage - 'GFAPerFloor' from ('PropertyGFATotal'/'NumberofFloors')

- Surface par bâtiment - 'GFAPerBuilding' from ('PropertyGFATotal'/'NumberofBuildings')

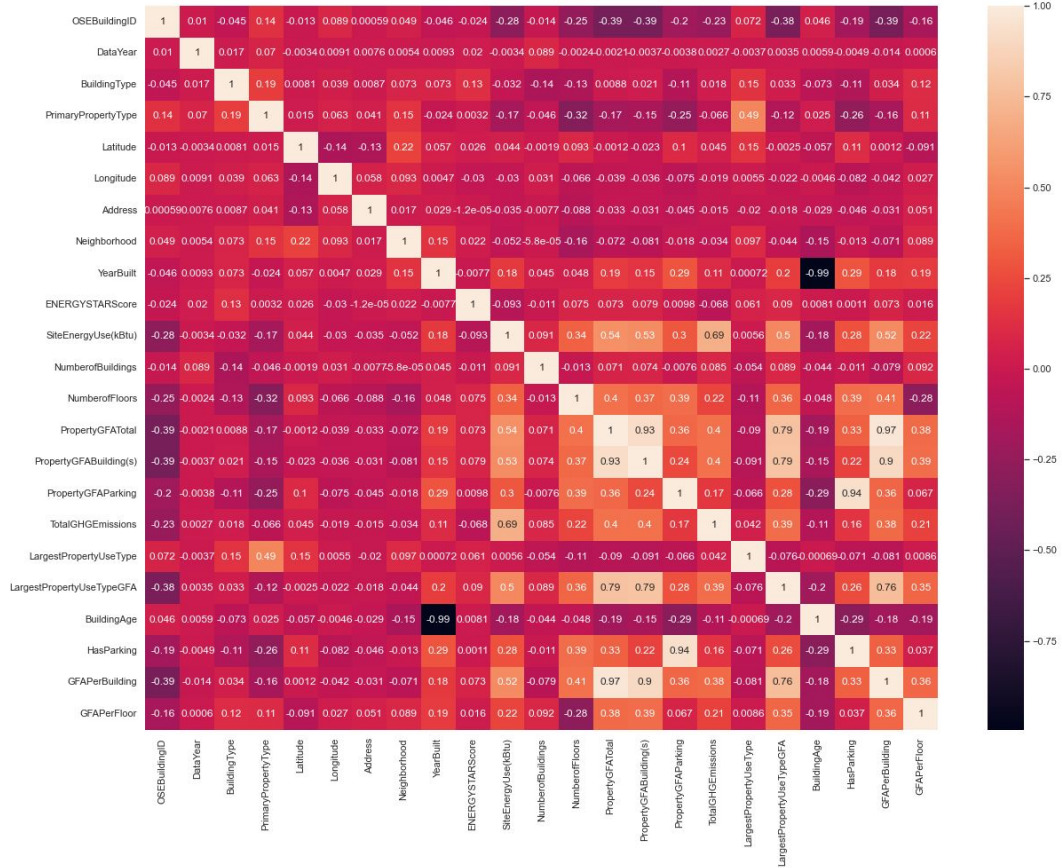
- Y a-t-il un parking ? - 'HasParking' - ('PropertyGFAParking' > 0) ou ('PropertyGFAParking' < 0)

Analyse exploratoire et nettoyage des données

Méthode de Kendall : le coefficient de corrélation mesure la relation monotone entre deux variables. Il n'est pas nécessaire que les variables soient normalement distribuées

Afin d'éviter le sur-apprentissage, nous devons **éliminer** les variables à forte corrélation > 0.75

GFAPerBuilding	~~PropertyGFATotal~~	0.970971
HasParking	~~PropertyGFAParking~~	0.941702
~~PropertyGFABuilding(s)~~	~~PropertyGFATotal~~	0.927388
~~PropertyGFABuilding(s)~~	GFAPerBuilding	0.899975
~~LargestPropertyUseTypeGFA~~	~~PropertyGFABuilding(s)~~	0.793218
~~PropertyGFATotal~~	~~LargestPropertyUseTypeGFA~~	0.788412
GFAPerBuilding	~~LargestPropertyUseTypeGFA~~	0.762522
~~YearBuilt~~	BuildingAge	-0.994586



Analyse exploratoire et nettoyage des données

- Avant de traiter les valeurs aberrantes, nous avons remplacé les valeurs manquantes par la valeur des médianes

- Nous avons également transformé les variables catégorielles en variables numériques

```
Int64Index: 2938 entries, 0 to 6715
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   PrimaryPropertyType     2938 non-null   float64
 1   Latitude                2938 non-null   float64
 2   Longitude               2938 non-null   float64
 3   Address                 2938 non-null   float64
 4   Neighborhood            2938 non-null   float64
 5   ENERGYSTARScore         2938 non-null   float64
 6   SiteEnergyUse(kBtu)     2938 non-null   float64
 7   NumberofBuildings       2938 non-null   float64
 8   NumberofFloors          2938 non-null   float64
 9   TotalGHGEmissions       2938 non-null   float64
 10  LargestPropertyUseType  2938 non-null   float64
 11  BuildingAge             2938 non-null   float64
 12  GFAPerBuilding          2938 non-null   float64
 13  GFAPerFloor             2938 non-null   float64
 14  HasParking              2938 non-null   float64
```

Pourquoi transformer ?

Nous avons aussi transformé les variables qui sont fortement asymétriques en faisant une transformation logarithmique

Pourquoi normaliser ou standardiser ?

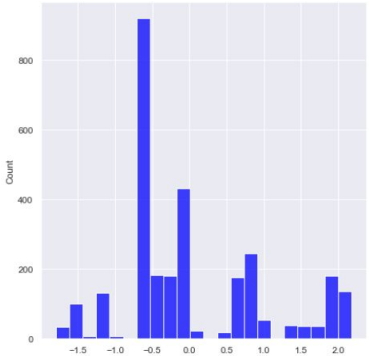
De nombreux algorithmes d'apprentissage automatique tentent de trouver des tendances dans les données en comparant les caractéristiques des points de données. Cependant, un problème se pose lorsque les caractéristiques sont à des échelles radicalement différentes

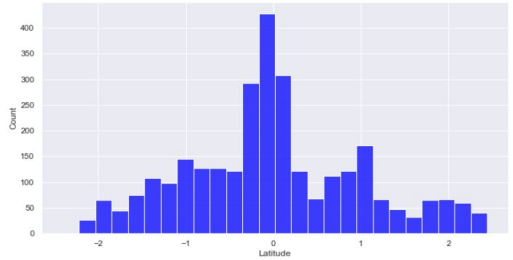
Z-scores pour détecter les valeurs aberrantes

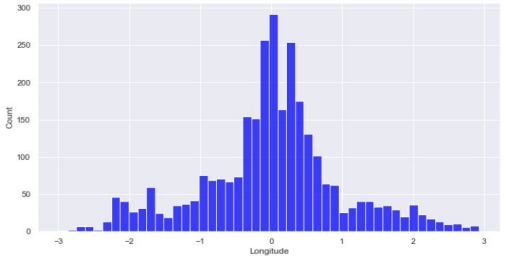
Nous avons utilisé la standardisation des z-scores en considérant une tolérance de +/-3 des z-scores

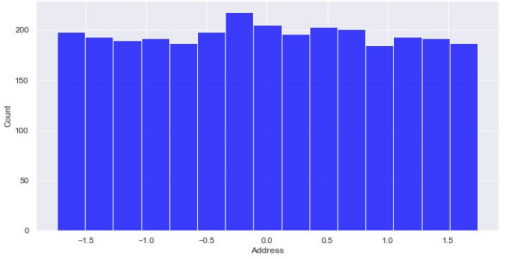
Analyse exploratoire et nettoyage des données

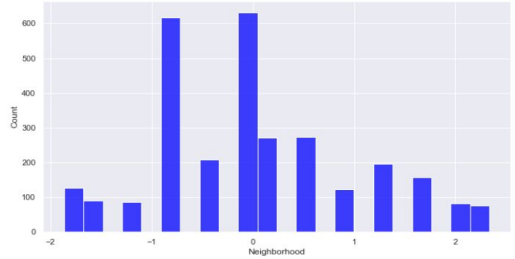
Données transformées et standardisées

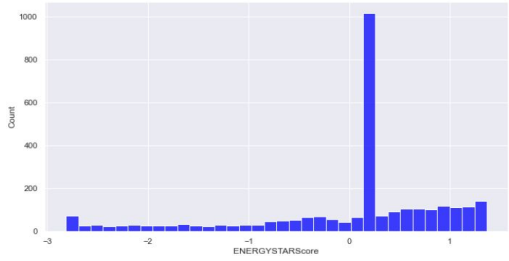


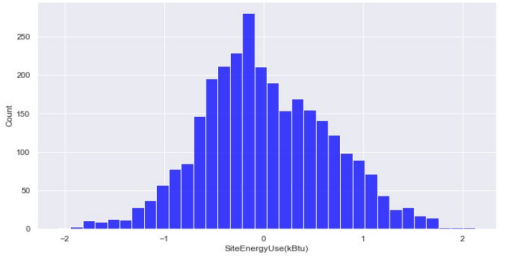


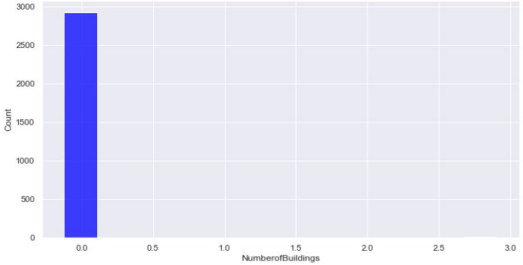


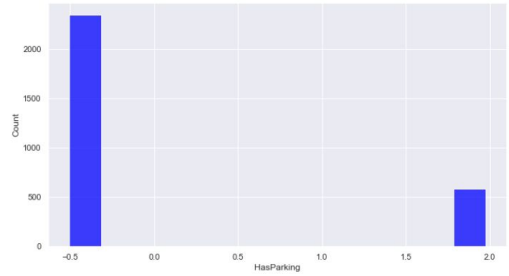






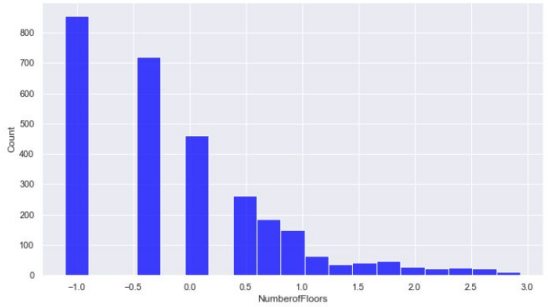


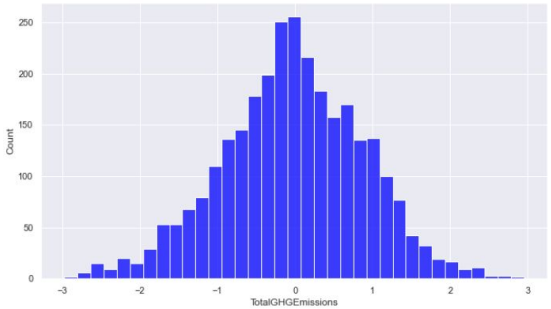


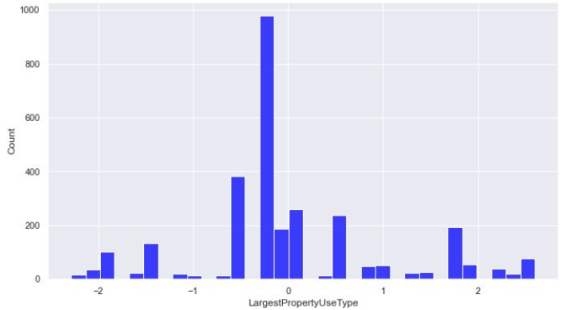


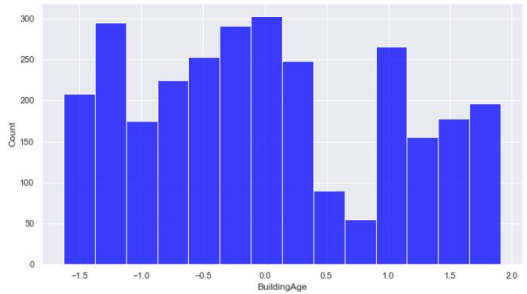
Analyse exploratoire et nettoyage des données

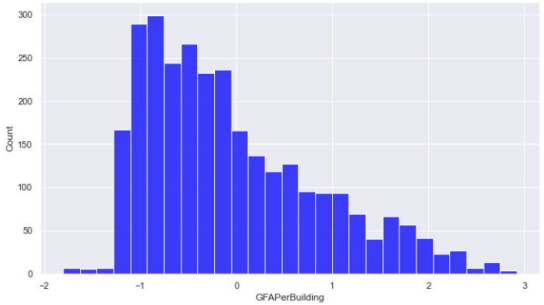
Données transformées et standardisées

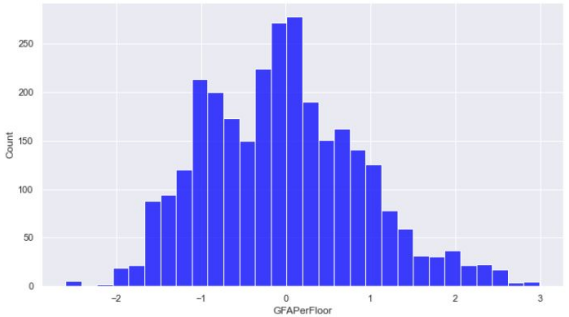




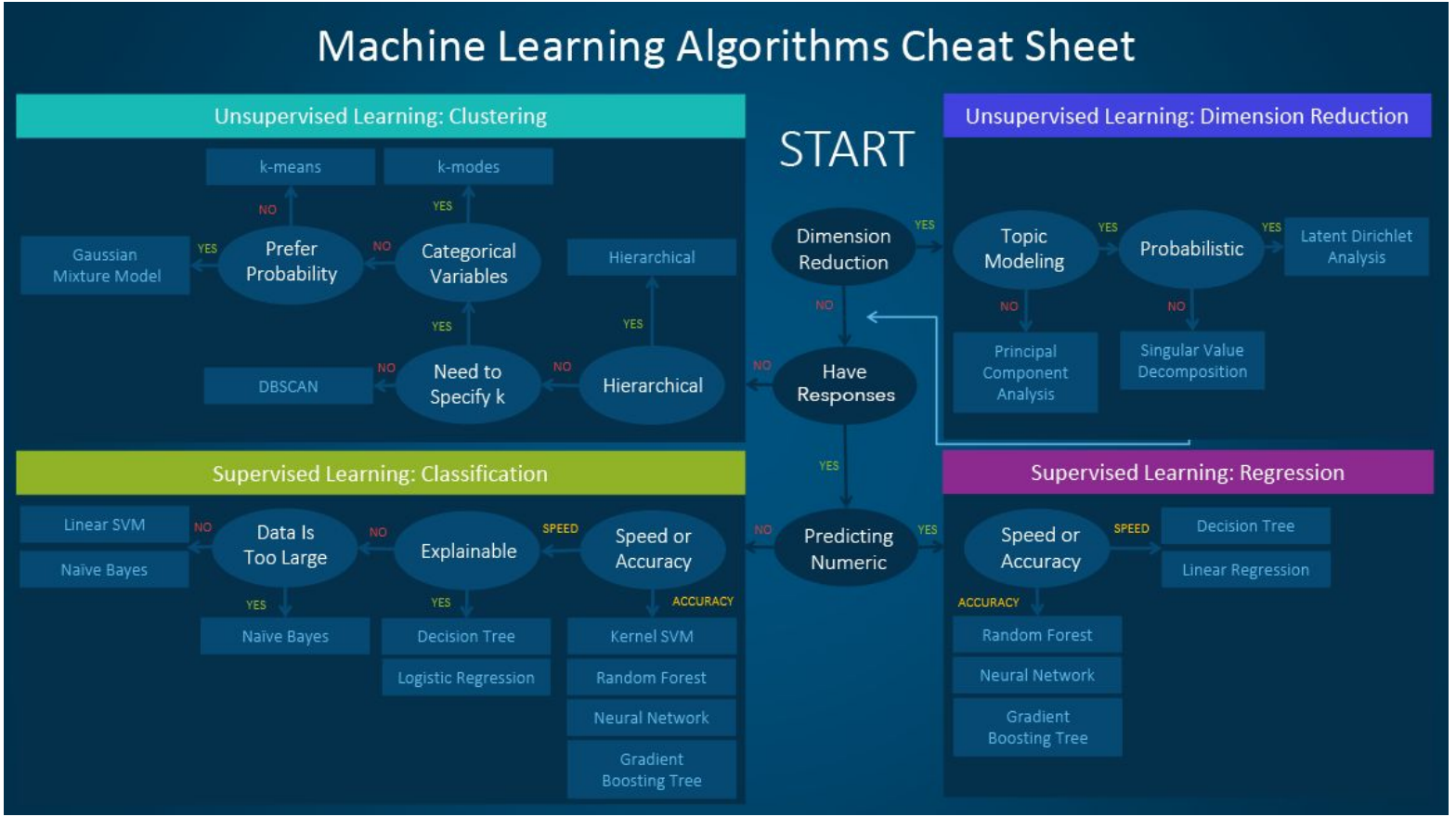








Analyse exploratoire et nettoyage des données



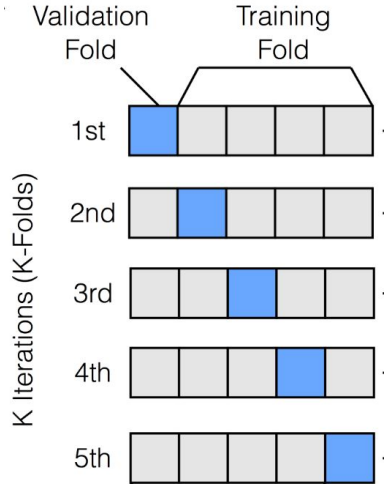
https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/

Analyse des prédictions

- ➔ **Fractionnement des données :** nous utilisons 80 % pour la traning et 20 % pour le test
- ➔ **Nous avons éliminé des données d'entrée les deux variables cibles : 'SiteEnergyUse(kBtu)' et 'TotalGHGEmissions' -** Nous avons d'abord choisi une variable pour tester les algorithmes: **'SiteEnergyUse(kBtu)'**

Analyse par régression linéaire (moindres carrés ordinaires - ordinary least squares)

Nous effectuons une validation croisée K-Fold pour évaluer la performance du modèle de **régression linéaire (moindres carrés ordinaires)**



R^2 moyen = 0.634
R^2 std = 0.042

La valeur moyenne de R^2 parmi les k-folds est une indication de la bonne performance du modèle

Après d'entraîner le modèle de régression linéaire, on teste le modèle dans des données non vues et les résultats sont les suivants

Données de test R^2 = 0.616

MSE = 0.167

MAE = 0.297

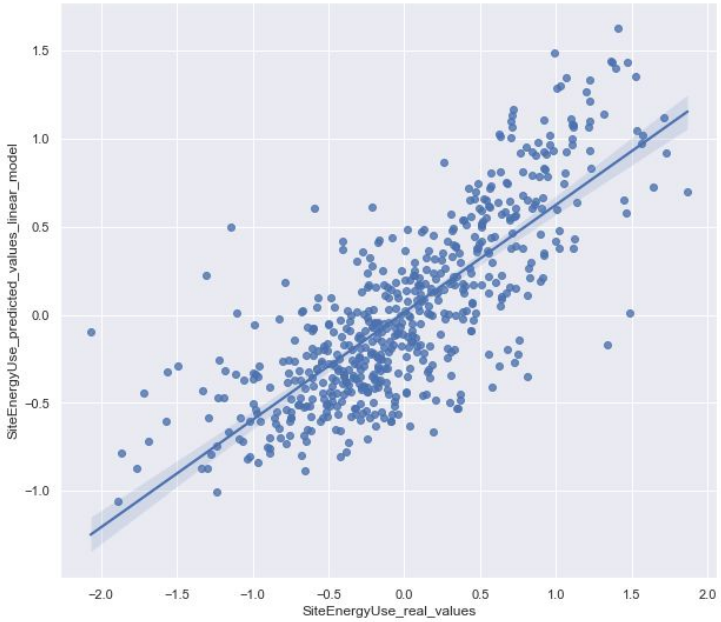
La valeur R^2 dans le sous-ensemble de données de test suggère que le modèle s'adapte modérément bien aux données non vues

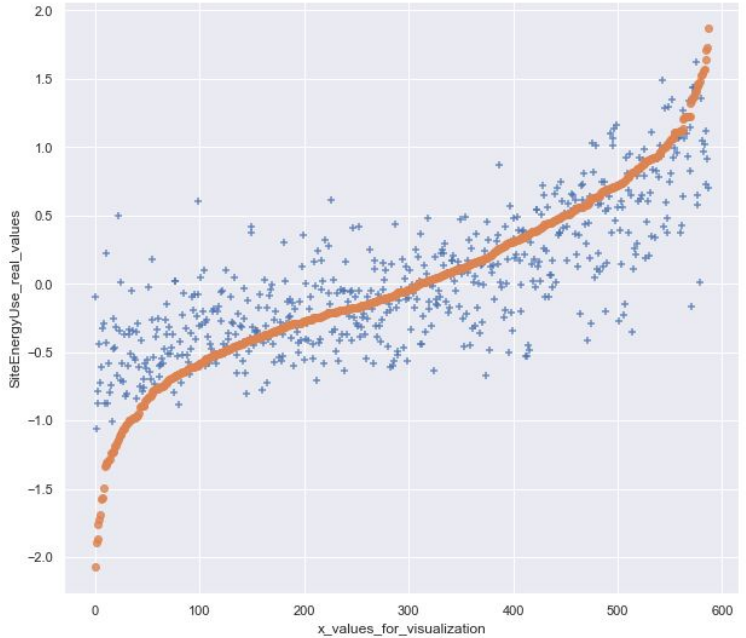
Analyse des prédictions

Analyse par régression linéaire (moindres carrés ordinaires - ordinary least squares)

Nous constatons la tendance à une corrélation linéaire entre les valeurs prédites et les valeurs réelles

Points bleus : données prédites
Points orange : données réelles

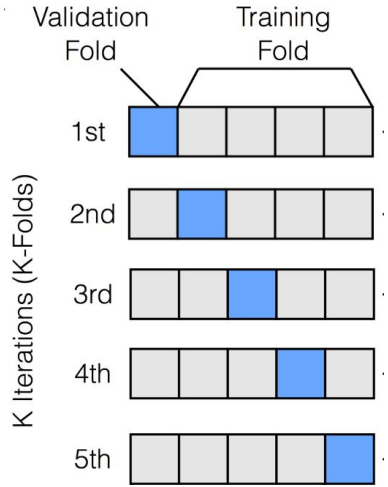




Analyse des prédictions

Analyse utilisant une régression par forêt aléatoire (random forest regressor)

Nous avons effectué aussi une validation croisée K-Fold pour évaluer la performance du modèle de **régression par forêt aléatoire**



R^2 moyen = 0.818
R^2 std = 0.026

La valeur moyenne de R^2 parmi les k-folds est encore mieux que R^2 que du modèle de régression linéaire (moindres carrés ordinaires)

Après avoir optimisé le modèle avec une validation croisée et après l'avoir entraîné, on teste le modèle dans des données non vues et les résultats sont les suivants

Données de test R^2 = 0.800

MSE = 0.087

MAE = 0.201

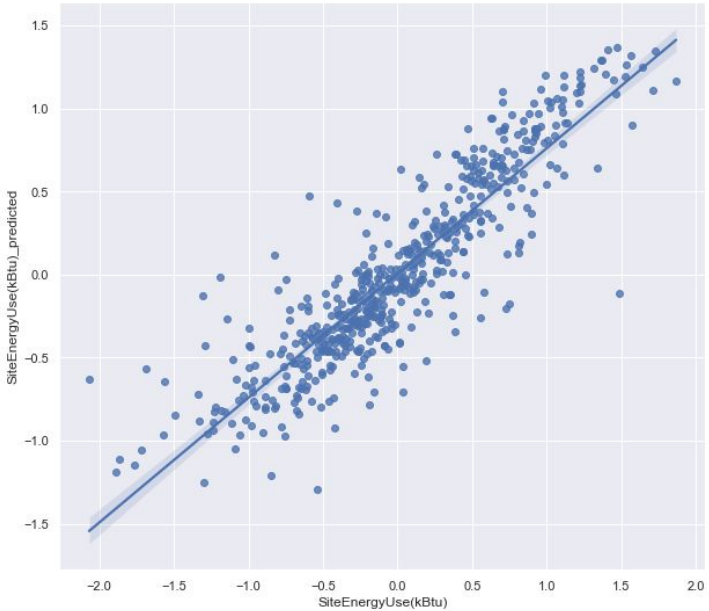
La valeur R^2 dans le sous-ensemble de données de test suggère que le modèle de régression par forêt aléatoire s'est mieux adapté aux données non vues que le modèle de régression linéaire (moindres carrés ordinaires)

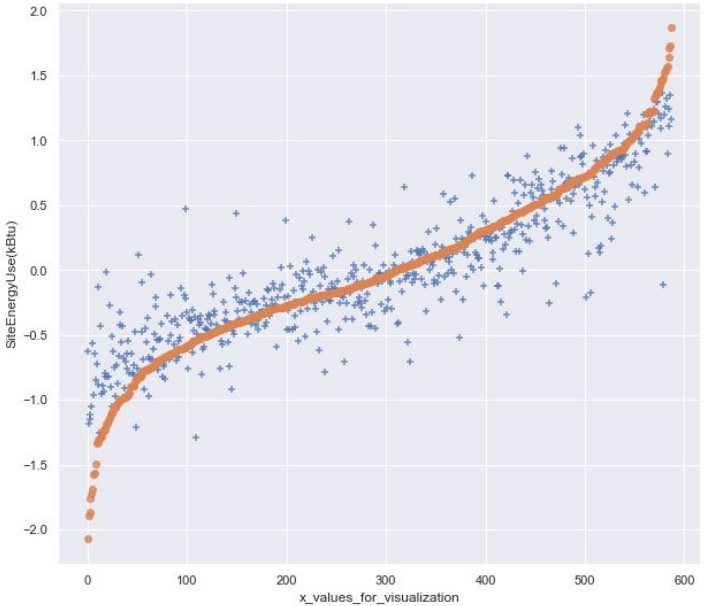
Analyse des prédictions

Analyse utilisant une régression par forêt aléatoire (random forest regressor)

Nous observons que la performance du régresseur de la forêt aléatoire est supérieure à celle de la régression linéaire des moindres carrés ordinaires

Points bleus : données prédites
Points orange : données réelles

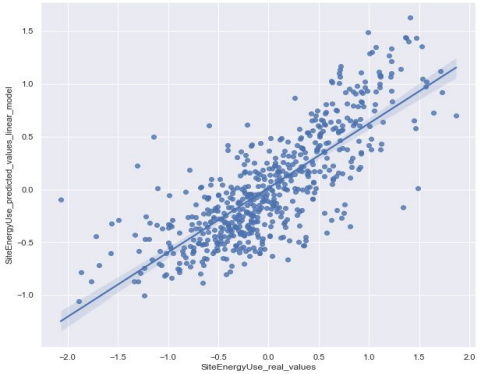




Analyse des prédictions

Comparaison entre les résultats des prédictions utilisant la régression linéaire des moindres carrés et le régresseur de la forêt aléatoire

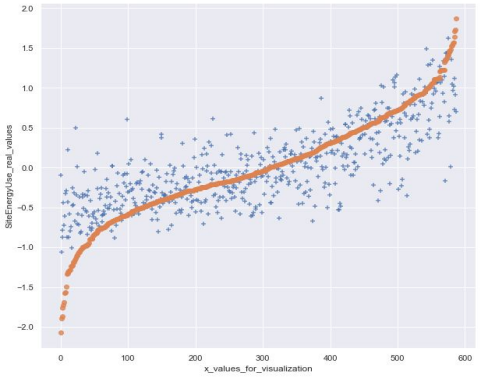
Linear regression- ordinary least squares



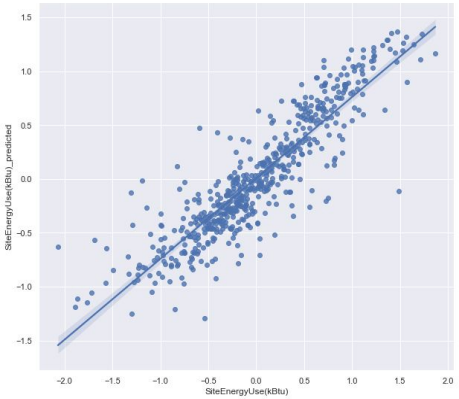
Données de test R^2 = 0.616

MSE = 0.167

MAE = 0.297



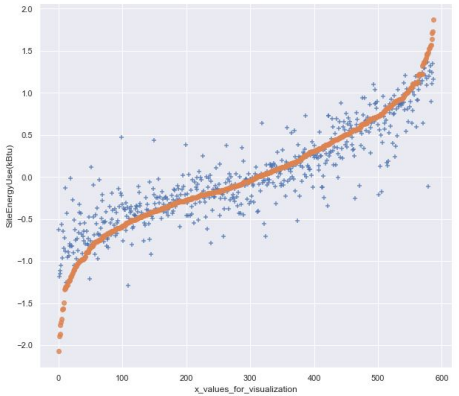
Random forest regressor



Données de test R^2 = 0.800

MSE = 0.087

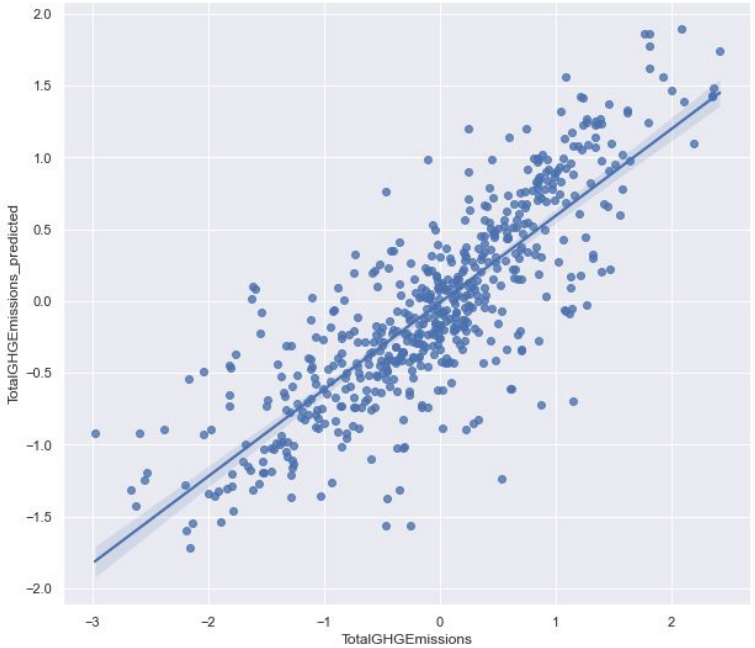
MAE = 0.201

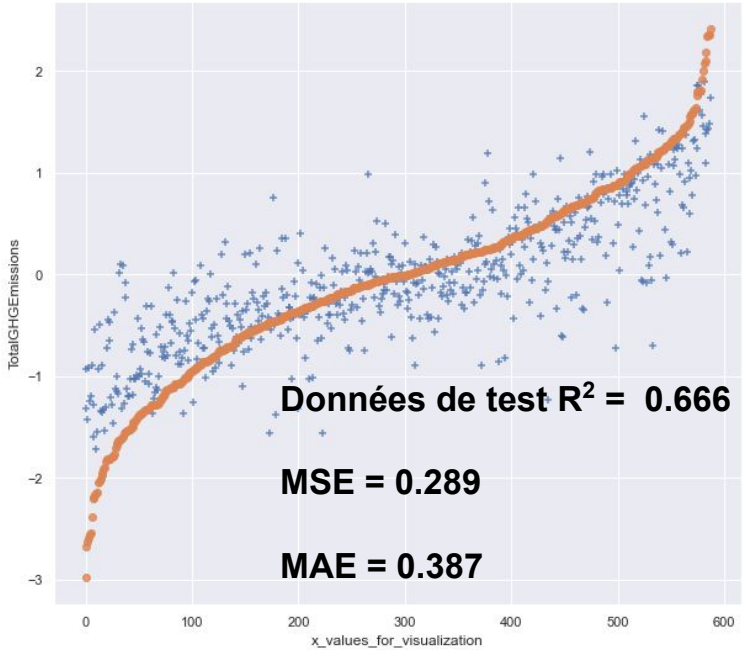


Analyse des prédictions

Analyse utilisant une régression par forêt aléatoire (random forest regressor)

En utilisant le régresseur de la forêt aléatoire, nous allons montrer la prédiction des émissions de gaz à effet de serre en choisissant la variable **'TotalGHGEmissions' - Avec l'influence du ENERGYSTARScore**

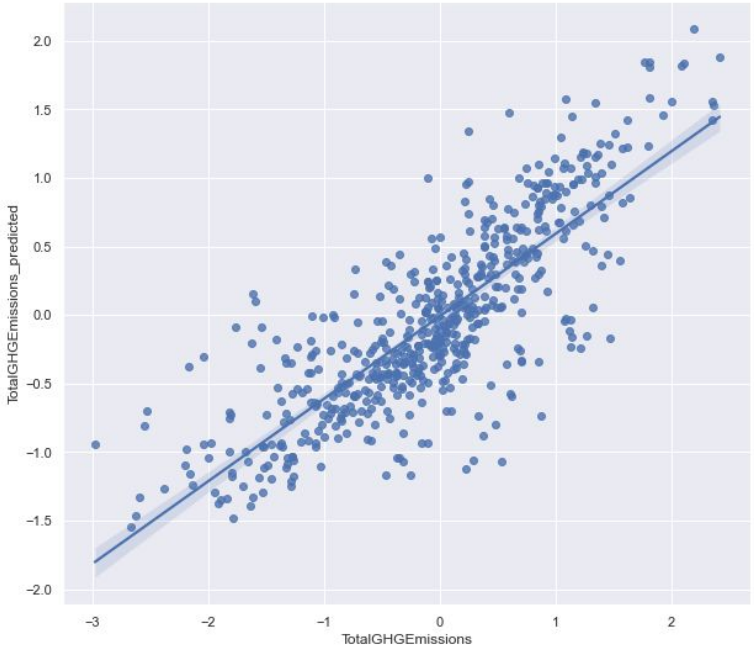


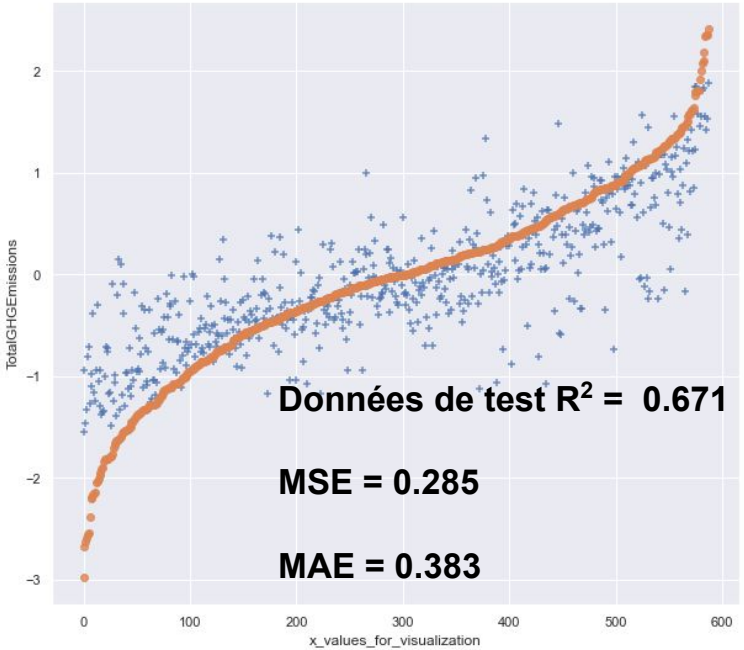


Analyse des prédictions

Analyse utilisant une régression par forêt aléatoire (random forest regressor)

En utilisant le régresseur de la forêt aléatoire, nous allons montrer la prédiction de la consommation d'énergie en choisissant la variable **'TotalGHGEmissions' - Sans l'influence du ENERGYSTARScore**





Conclusions

- ➔ Après une sélection appropriée des variables d'intérêts pour réaliser cette mission, nous avons évalué deux modèles de prédiction, basés sur le besoin de prédictions numériques :
  - ◆ Régression linéaire (moindres carrés ordinaires - ordinary least squares)
  - ◆ Régression par forêt aléatoire (random forest regressor)
- ➔ Nous avons obtenu une meilleure performance du modèle random forest regressor en termes de fidélité des prédictions
  - ◆ Il est important de noter qu'en choisissant ce modèle, on compromet les ressources informatiques en augmentant le temps de calcul
  - ◆ Bien que la régression par forêt aléatoire a eu une performance élevée, nous avons effectué une validation croisée pour optimiser le modèle et nous avons obtenu des résultats proches
- ➔ Nous pouvons utiliser un régresseur de forêt aléatoire pour prédire les valeurs d'ENERGY STAR Score pour la performance énergétique des bâtiments
- ➔ **À partir des valeurs obtenues avec l'influence et sans l'influence du ENERGYSTARScore, on ne voit pas la pertinence d'utiliser cet indicateur pour la prédiction des émissions de gaz.**