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
- Analyse de la corrélation des variables
- Détection des valeurs aberrantes et normalisation des données

➔ Analyse des prédictions
- Test de deux algorithmes de prédiction différents: Linear Regression et Random Forest Regressor
- Prédiction de la consommation et des émissions de CO_2

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 (kilogrammes d'équivalent de dioxyde de carbone par pied carré)

- TotalGHGEmissions : quantité totale d'émissions Greenhouse Gas, y compris le dioxyde de carbone, le méthane et l'oxyde nitreux, rejetés dans l'atmosphère par la consommation d'énergie du bien (tonnes métriques d'équivalent 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

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)**

↓

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

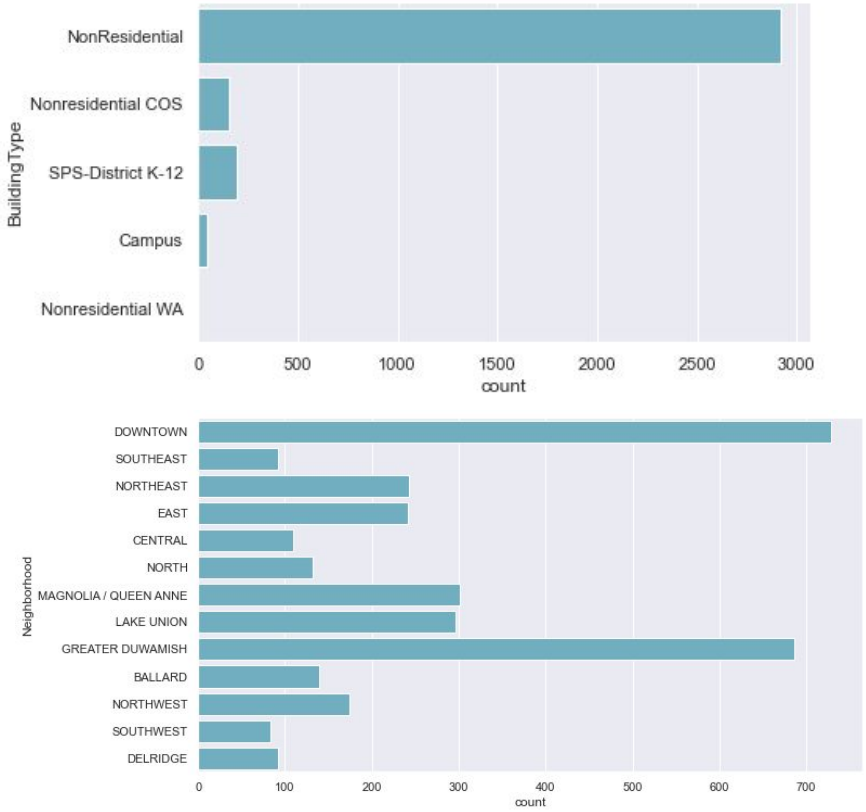
Fichier de données unique avec les données des années 2015 et 2016 avec les variables pertinentes contient 6716 lignes et 19 colonnes

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

Analyse exploratoire et nettoyage des données

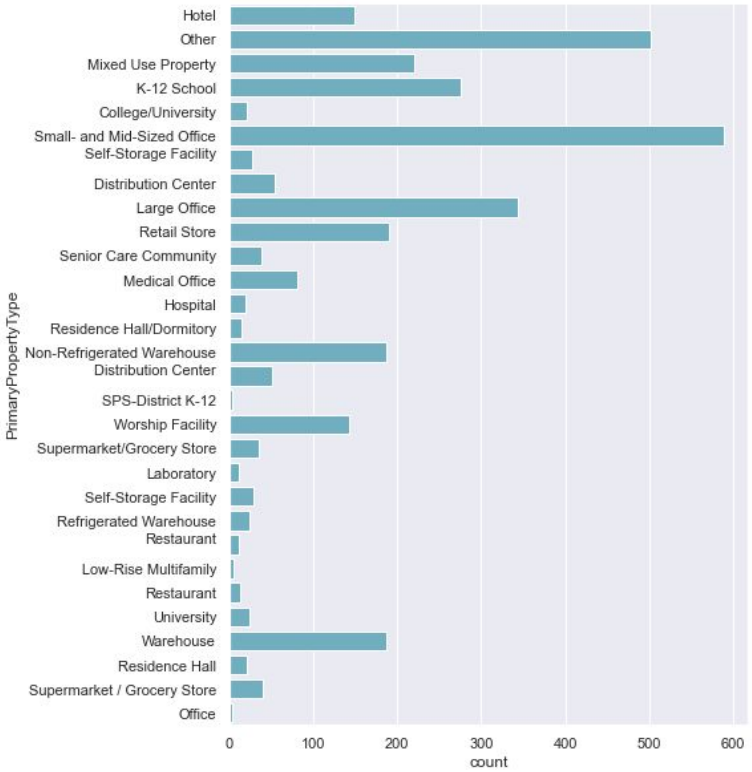
Valeurs manquantes parmi les variables

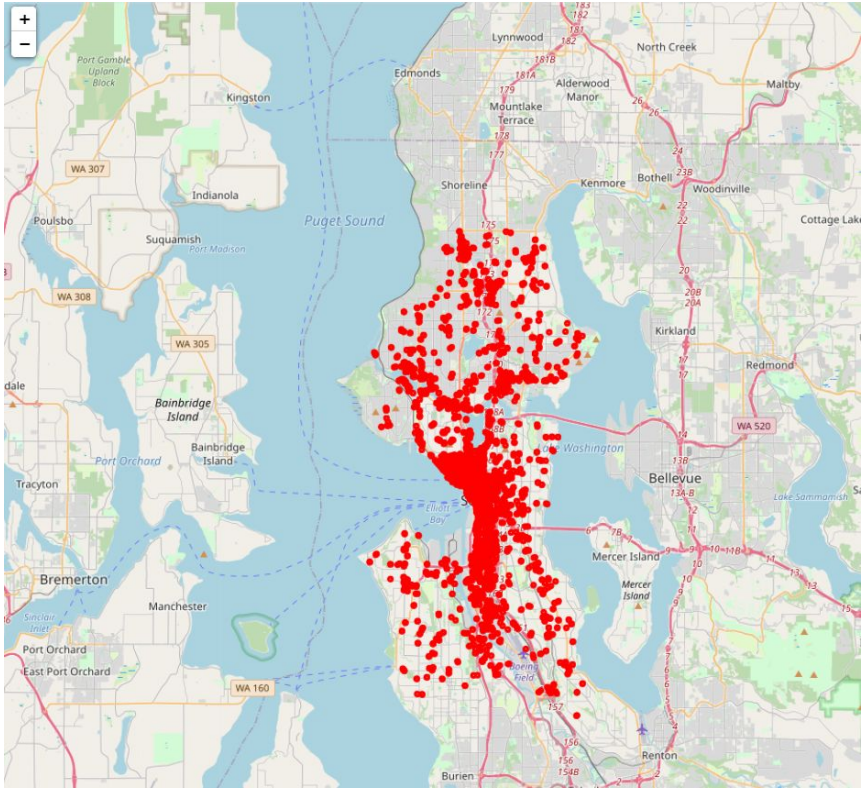
Bâtiments non résidentiels



Analyse exploratoire et nettoyage des données

Bâtiments non résidentiels et son emplacement sur la carte de Seattle





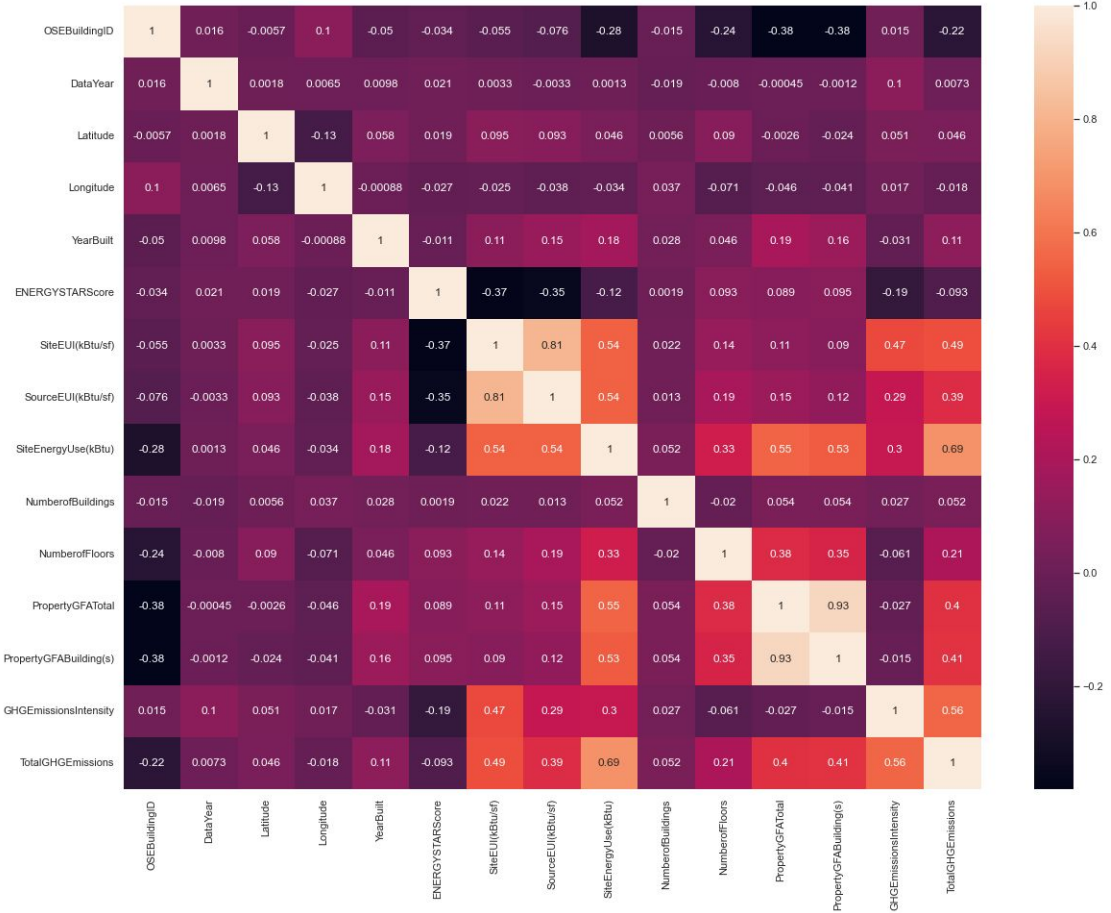
Analyse exploratoire et nettoyage des données

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

PropertyGFABuilding(s)	PropertyGFATotal	0.928091
SiteEUI(kBtu/sf)	SourceEUI(kBtu/sf)	0.807923

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

'SourceEUI(kBtu/sf)' et 'PropertyGFABuilding(s)'



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

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3305 entries, 0 to 6715
Data columns (total 9 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   PrimaryPropertyType    3305 non-null   int8
 1   Neighborhood           3305 non-null   int8
 2   YearBuilt              3305 non-null   int64
 3   SiteEUI(kBtu/sf)       3305 non-null   float64
 4   SiteEnergyUse(kBtu)    3305 non-null   float64
 5   NumberofFloors         3305 non-null   float64
 6   PropertyGFATotal       3305 non-null   int64
 7   GHGEmissionsIntensity  3305 non-null   float64
 8   TotalGHGEmissions      3305 non-null   float64
dtypes: float64(5), int64(2), int8(2)
memory usage: 342.1 KB
```

...et nous avons aussi normalisé les variables et supprimé les valeurs aberrantes.

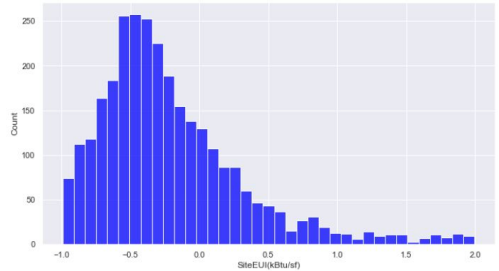
Pourquoi normaliser ?

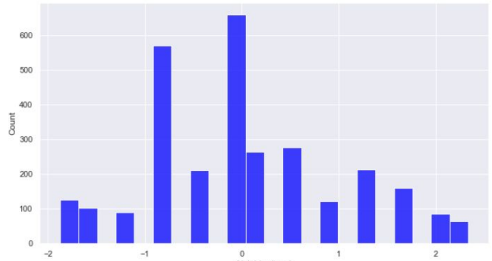
De nombreux algorithmes d'apprentissage automatique tentent de trouver des tendances dans les données en comparant les caractéristiques des points de données. Cependant, un problème se pose lorsque les caractéristiques sont à des échelles radicalement différentes.

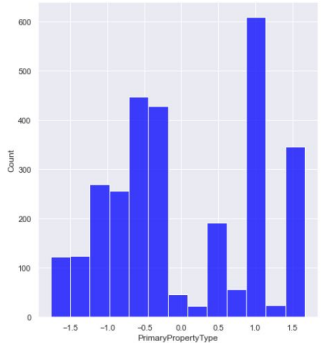
Les données normalisées sans valeurs aberrantes contiennent 2942 lignes et 9 colonnes

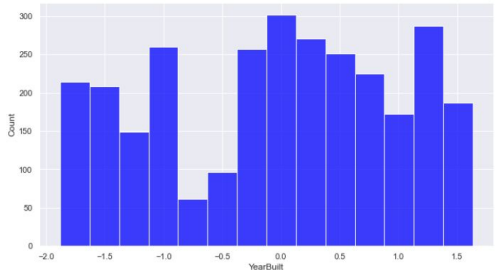
Analyse exploratoire et nettoyage des données

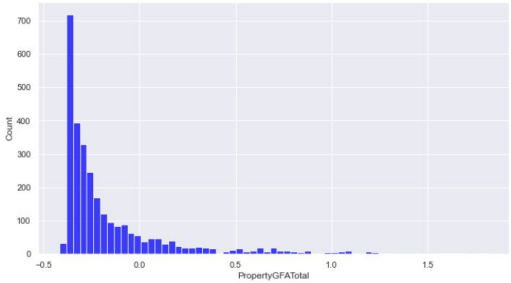
Données normalisées en utilisant Z-scores

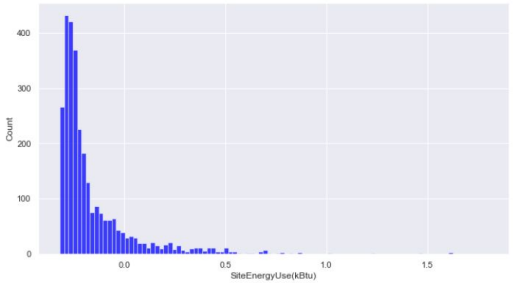


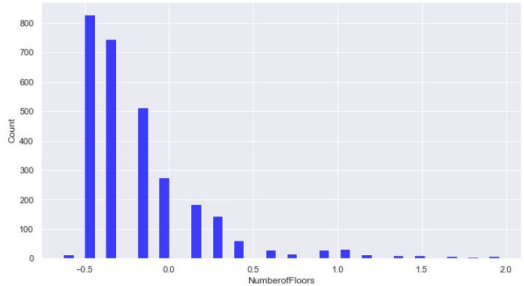


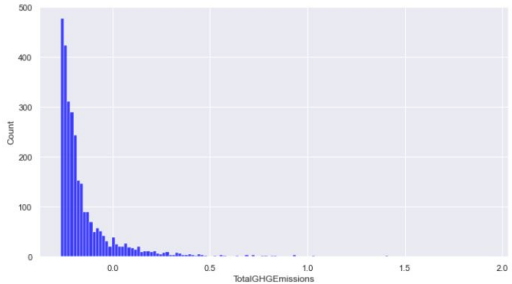


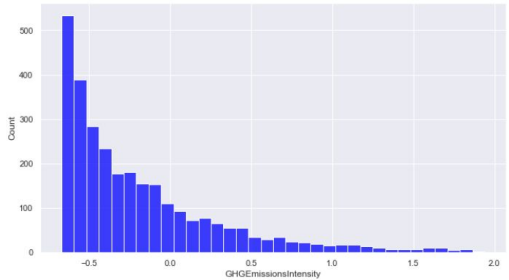




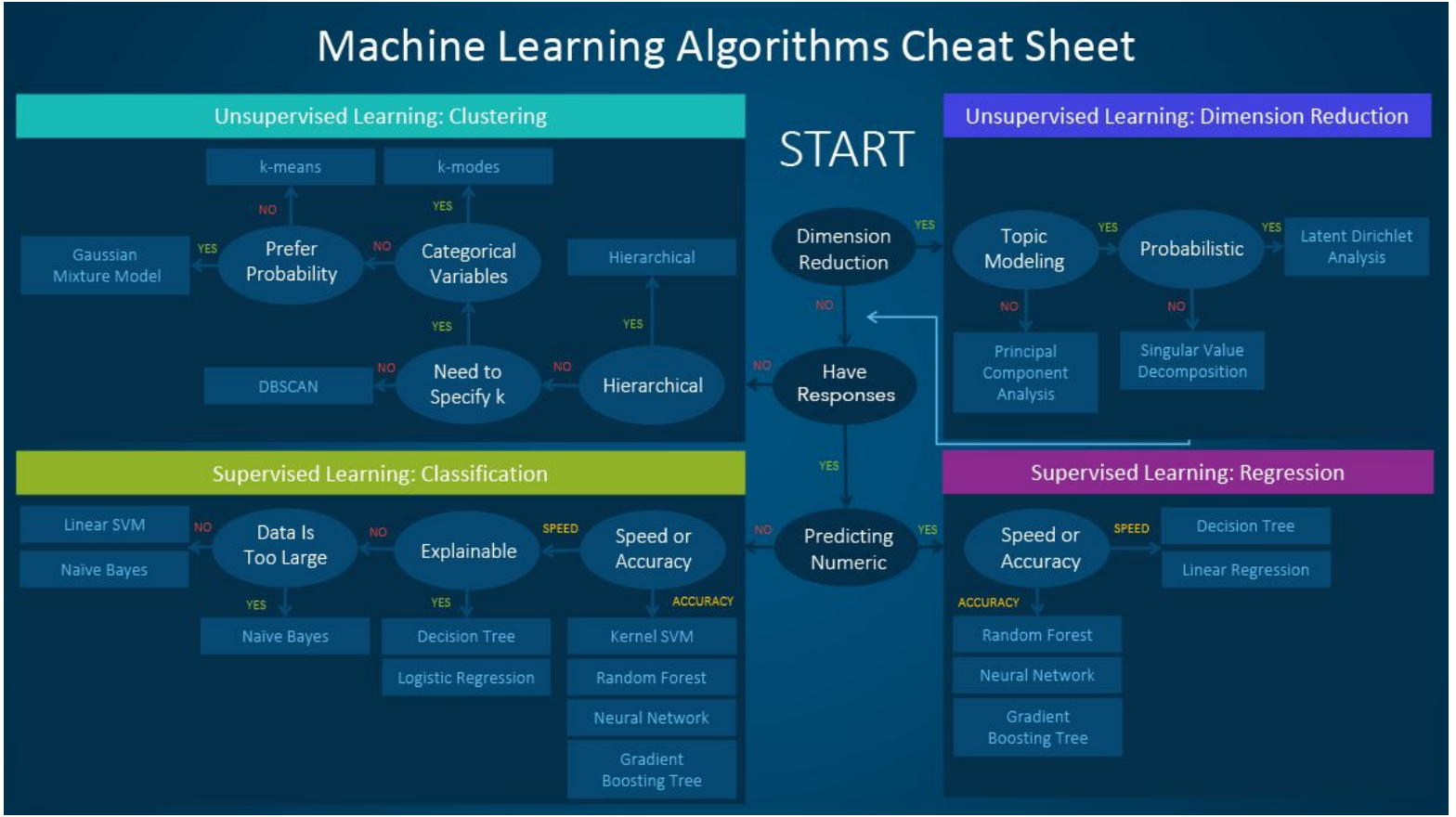








Analyse exploratoire et nettoyage des données



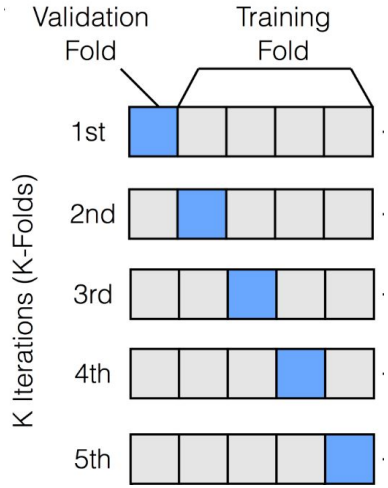
https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/

Analyse des prédictions

Fractionnement des données : nous utilisons 80 % pour la traning et 20 % pour le test
Nous avons d'abord choisi une variable pour tester les algorithmes: **'TotalGHGEmissions'**

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 = 0.843
R^2 std = 0.032

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 = 0.811

MSE = 0.007

MAE = 0.045

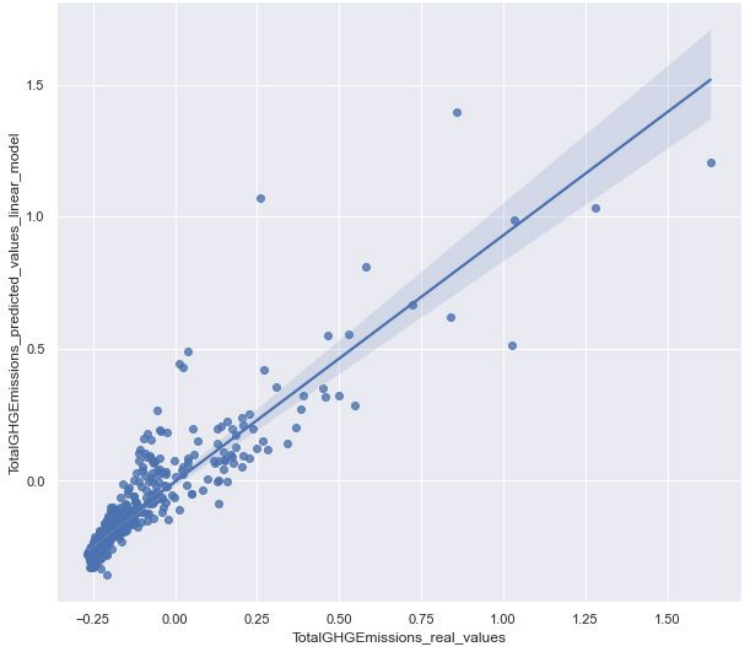
La valeur R^2 dans le sous-ensemble de données de test suggère que notre modèle régressif s'est bien adapté aux données non vues

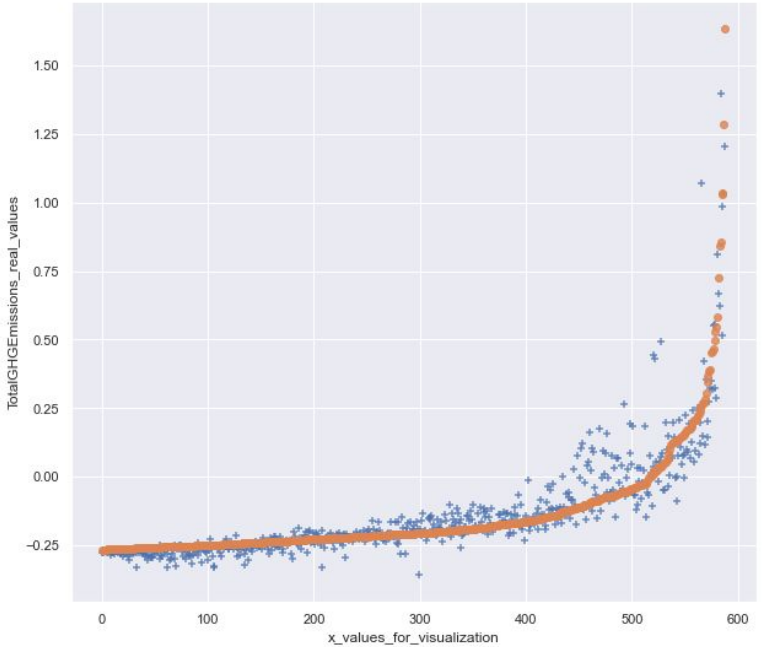
Analyse des prédictions

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

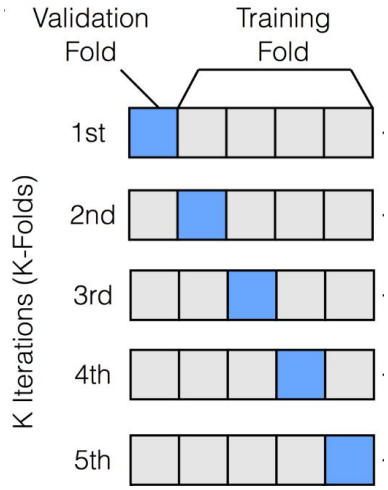




Analyse des prédictions

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 = 0.963
R^2 std = 0.013

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 = 0.962

MSE = 0.002

MAE = 0.014

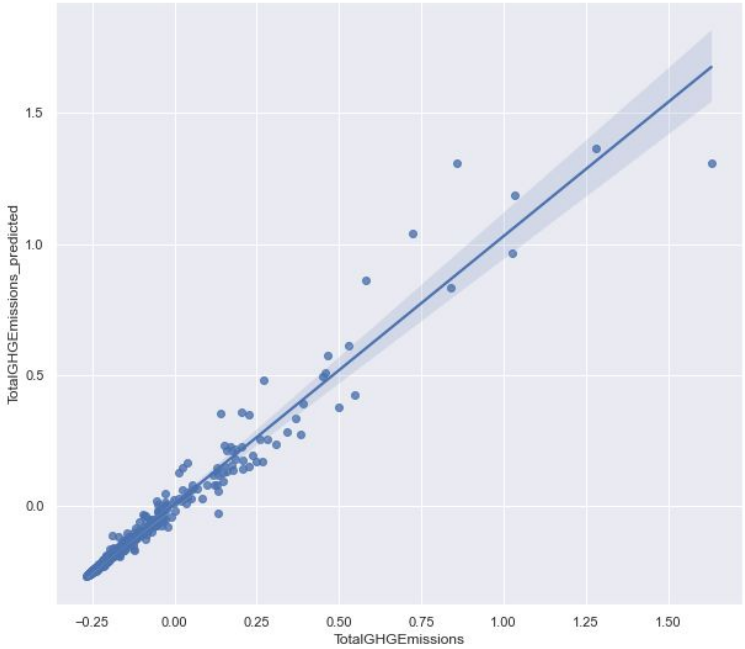
La valeur R^2 dans le sous-ensemble de données de test suggère que le modèle de régression par forêt aléatoire s'est bien adapté aux données non vues est encore mieux que le modèle de régression linéaire (moindres carrés ordinaires)

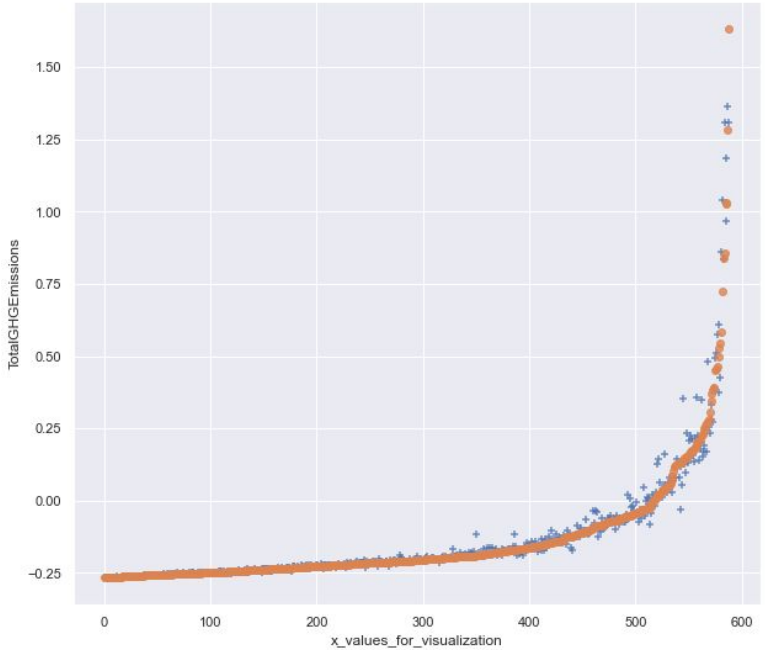
Analyse des prédictions

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

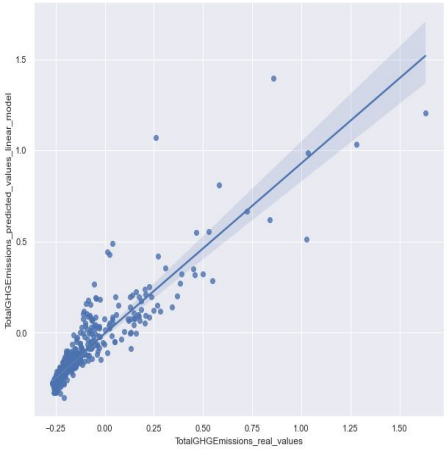




Analyse des prédictions

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

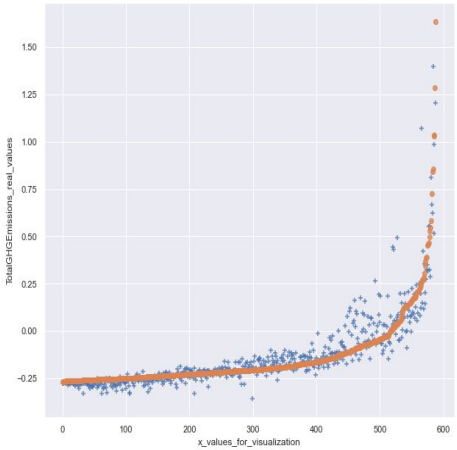
Linear regression- ordinary least squares



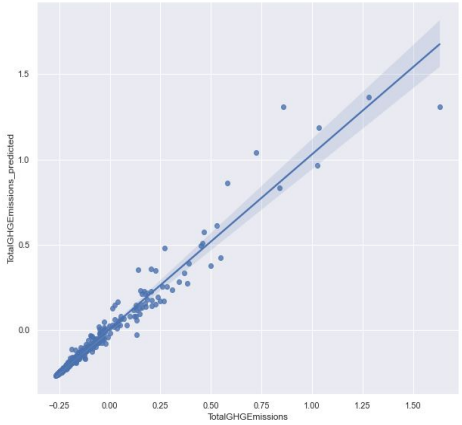
Données de test R^2 = 0.811

MSE = 0.007

MAE = 0.045



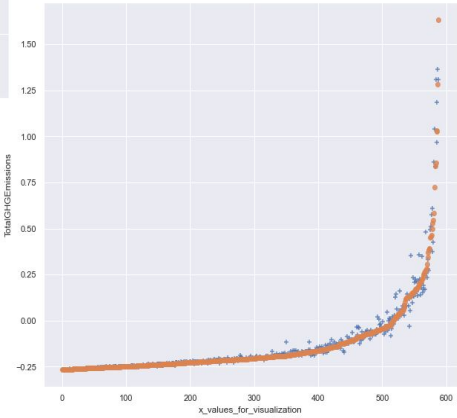
Random forest regressor



Données de test R^2 = 0.962

MSE = 0.002

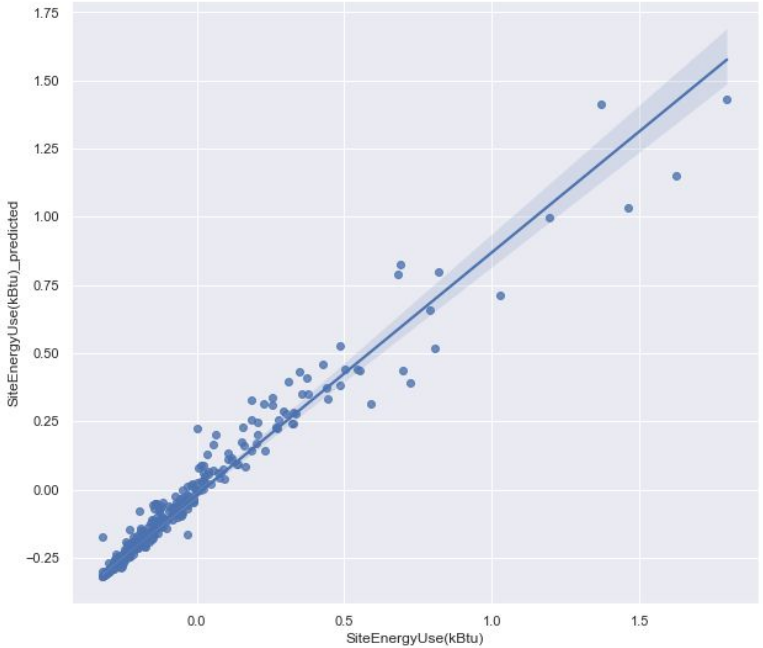
MAE = 0.014

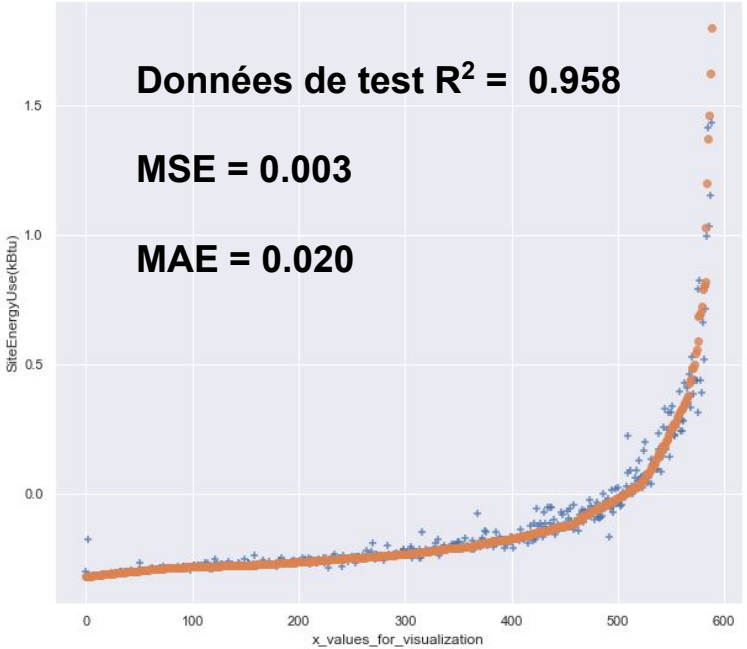


Analyse des prédictions

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 **'SiteEnergyUse(kBtu)'**





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
- ➔ Nous vous recommandons également d'utiliser une approche en utilisant des réseaux neuronaux, qui offrent également une bonne précision de prédiction