Anticipez les besoins en consommation électrique de bâtiments

Data Science | Projet 4

Fırat Yasar

08/10/2021

Sommaire

Présentation

Présentation de la problématique

- L'objectif de la ville de Seattle :
  neutre en émissions de carbone en 2050.
- Relevés manuels minutieux effectués par nos agents en 2015 et 2016.
- Ces relevés sont coûteux à obtenir
- Il reste encore des bâtiments à mesurer.
- Tenter de prédire les émissions de CO2 et la consommation totale d'énergie
- Intérêt de l'indicateur Energy Star Score pour les prédictions d'émissions
- Pour cela, nous avons à notre disposition la base de données :

https://www.kaggle.com/city-of-seattle/sea-building-energy-benchmarking#2015-building-energy-benchmarking.csv

Découverte du jeu de données

Dataset 2015

3340 lignes
47 colonnes

10 colonnes différentes de 2016

Dataset 2016

3376 lignes
46 colonnes

9 colonnes différentes de 2015

Information géographique

- Adresse
- Coordonnées GPS
- Code postale
- etc.

Décrivants caractéristiques

- Le type d'utilisation
- Le nombre d'étages et bâtiments
- La surface brute de plancher
- etc.

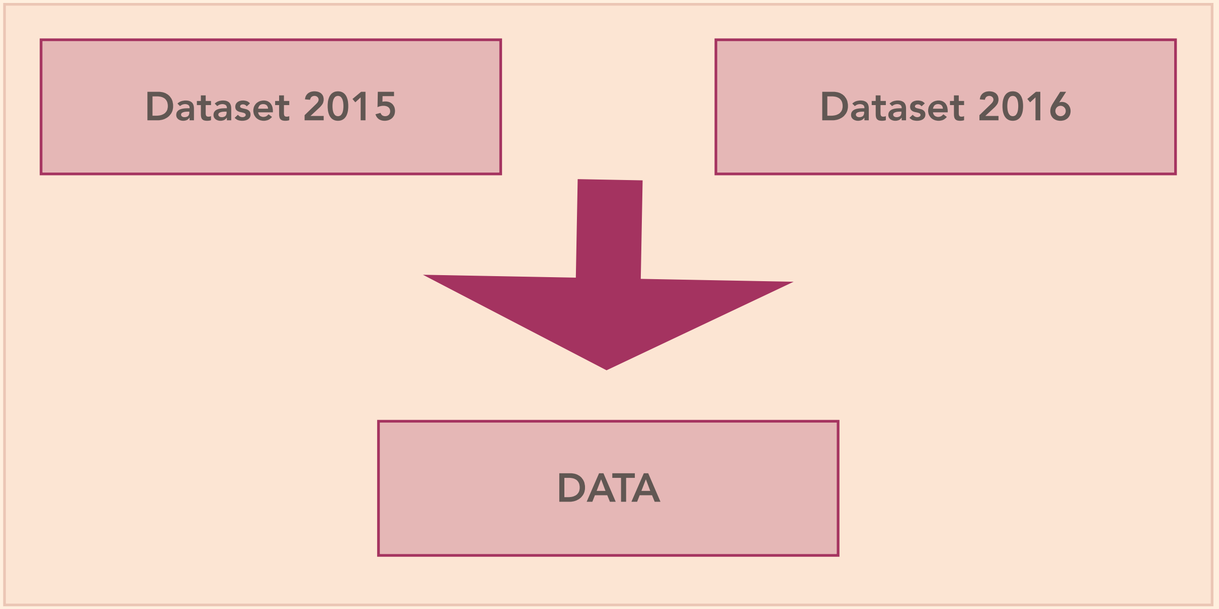
Variables énergétiques

- La consommation électrique,
- La consommation de gaz
- L'émissions de CO2
- etc.

Nettoyage des données

Étape 1 : Création d'un nouveau dataset

- Concaténation des deux dataset
- Remise en forme des colonnes et correction

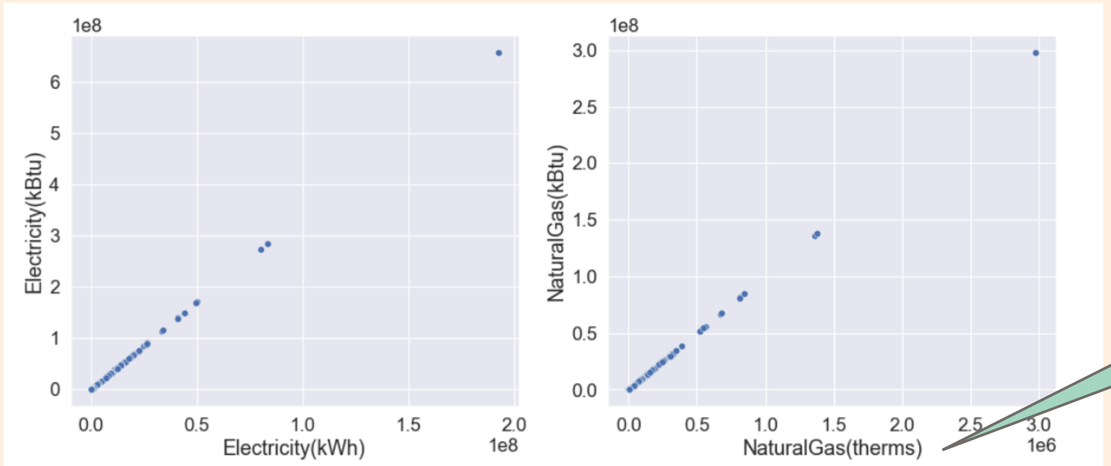



```
============================================================

('=============== Colonnes data1 - data2 ===============',
 0                                                Location
 1                                     OtherFuelUse(kBtu)
 2                          GHGEmissions(MetricTonsCO2e)
 3                       GHGEmissionsIntensity(kgCO2e/ft2)
 4                                                Comment
 5                                          2010 Census Tracts
 6       Seattle Police Department Micro Community Poli...
 7                                  City Council Districts
 8                                               SPD Beats
 9                                               Zip Codes
dtype: string,
 '=============== Colonnes data2 - data1 ===============',
 0                                 Address
 1                                    City
 2                                   State
 3                                 ZipCode
 4                                Latitude
 5                               Longitude
 6                                Comments
 7                          TotalGHGEmissions
 8                       GHGEmissionsIntensity
dtype: string)
```

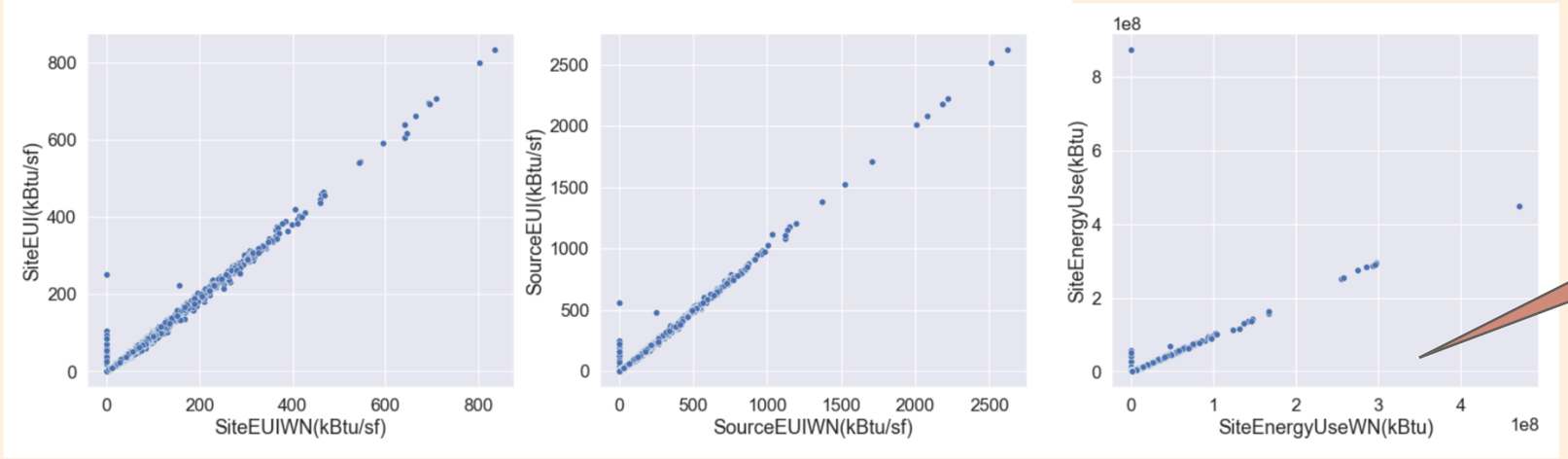
Étape 2 : Suppression des variables redondantes

- Elimination des variables avec le suffixe "**WN**" (weather normalised)
- Suppression des variables qui sont fournies en plusieurs unités



kBtu x 3.412 = kWh

Therms x 99.976129 = kBtu



Effets météorologiques

Étape 3 : Nettoyage divers

- Elimination des outliers
  - Les valeurs négatives (5 lignes)
  - Les valeurs non conformes (120 lignes)
- Données des bâtiments résidentiels
- Traitement des valeurs manquantes

	BuildingType
NonResidential	1487
Multifamily LR (1-4)	1036
Multifamily MR (5-9)	583
Multifamily HR (10+)	110
SPS-District K-12	93
Nonresidential COS	82
Campus	25
Nonresidential WA	1

multifamily

	ComplianceStatus
22	Error - Correct Default Data
28	Missing Data
30	Error - Correct Default Data
31	Missing Data
38	Error - Correct Default Data

Étape 4 : Exclusion de variables des relevés

- Regroupement des variables
- L'étude ne se base pas sur les relevés (électricité, gaz, etc.)

```
17  ThirdLargestPropertyUseTypeGFA
18  ENERGYSTARScore
19  SiteEnergyUse(kBtu)
20  SteamUse(kBtu)
21  Electricity(kBtu)
22  NaturalGas(kBtu)
23  TotalGHGEmissions
24  ZipCode
```

DATA

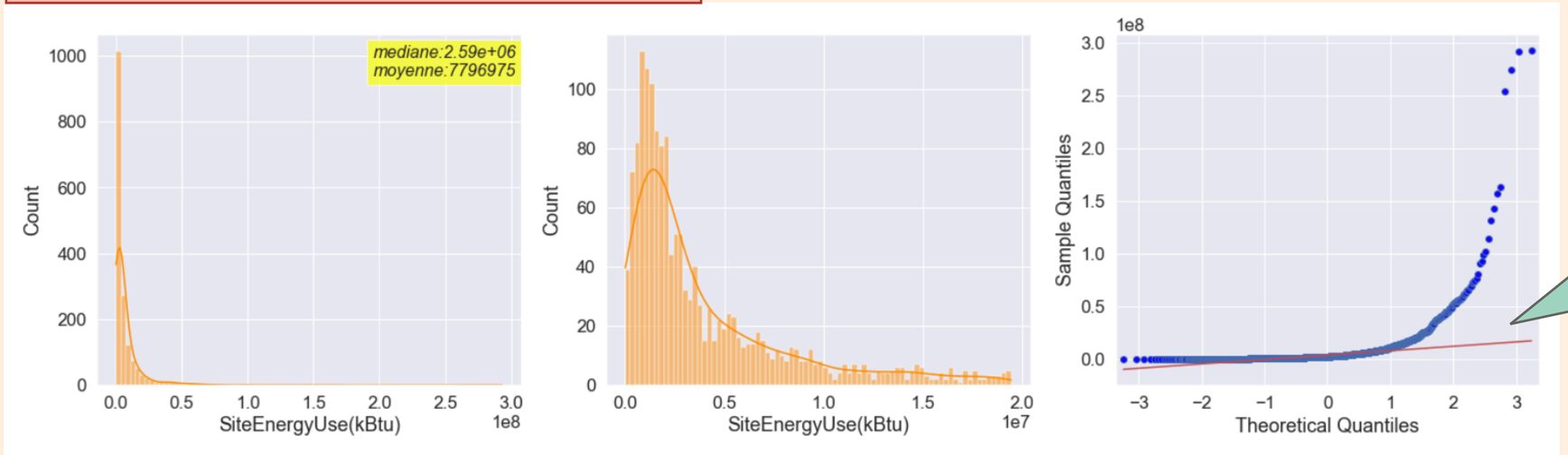
30 colonnes
1688 lignes

1547 lignes (2016)
141 lignes (2015)

Analyse exploratoire

Analyse de la distribution des cibles

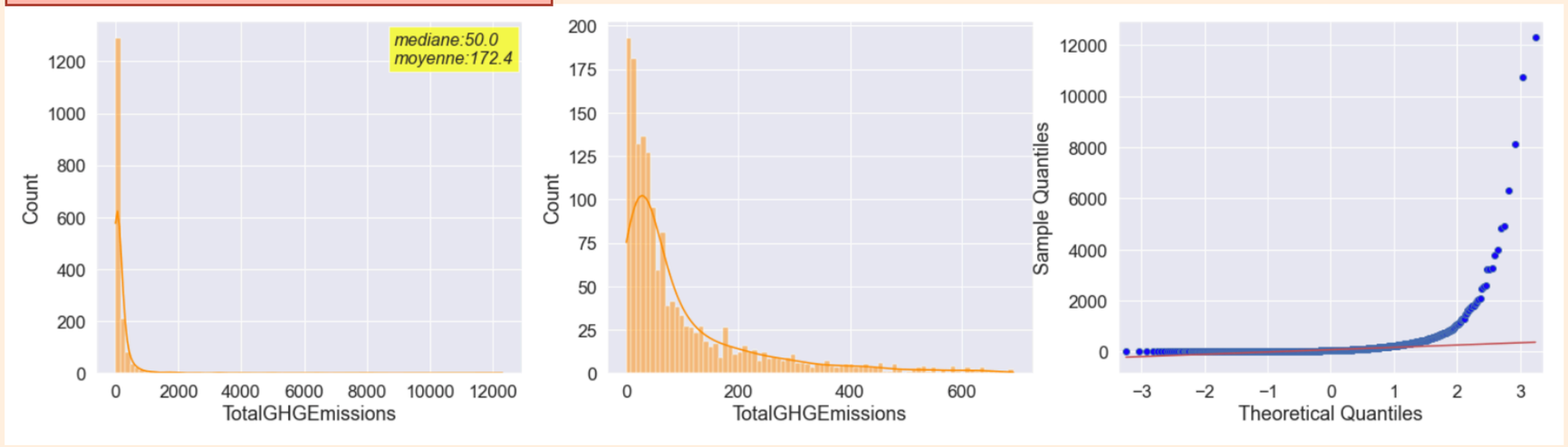
La consommation totale d'énergie : SiteEnergyUse(kBtu)



Les distributions des deux variables ne sont pas Gaussienne

```
TotalGHGEmissions (normaltest):
Statistics=3041.926, p=0.000
Sample does not look Gaussian (reject H0)
==========================================
TotalGHGEmissions (Shapiro):
Statistics=0.229, p=0.000
Sample does not look Gaussian (reject H0)
==========================================
SiteEnergyUse(kBtu) (normaltest):
Statistics=2495.587, p=0.000
Sample does not look Gaussian (reject H0)
==========================================
SiteEnergyUse(kBtu) (Shapiro):
Statistics=0.344, p=0.000
Sample does not look Gaussian (reject H0)
==========================================
```

Les émissions de CO2 : TotalGHGEmissions



Feature engineering

Nouveaux features (attributs)

- Transformation des données brutes en la forme la plus appropriée pour les algorithmes de machine learning

- 3 catégories de nouveaux features

Categorie A.

Les variables qui vérifient le type de consommation

- Electricity_check
- NaturalGas_check
- SteamUse_check

*La valeur de **NaturalGas_check** est 1 si la consommation de gaz est un nombre positif; sinon elle est 0.*

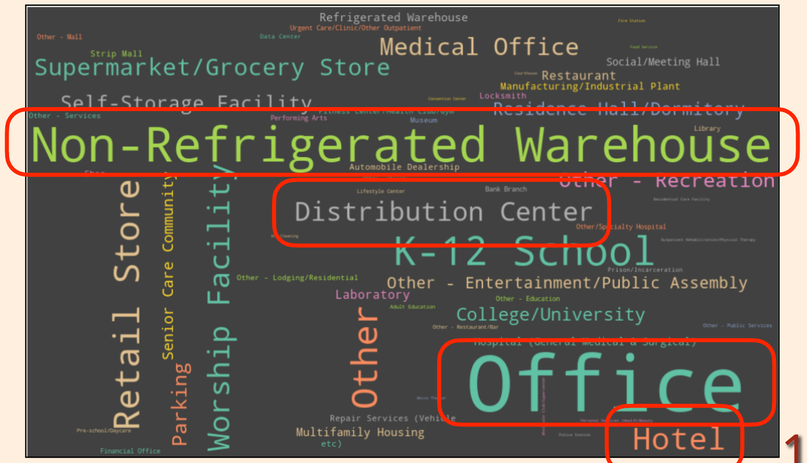
Categorie B.

La proportion des trois premiers plus gros usages immobiliers par la surface totale

- LP_Ratio (LargestPropertyUse)
- SLP_Ratio (SecondLargestPropertyUse)
- TLP_Ratio (ThirdLargestPropertyUse)

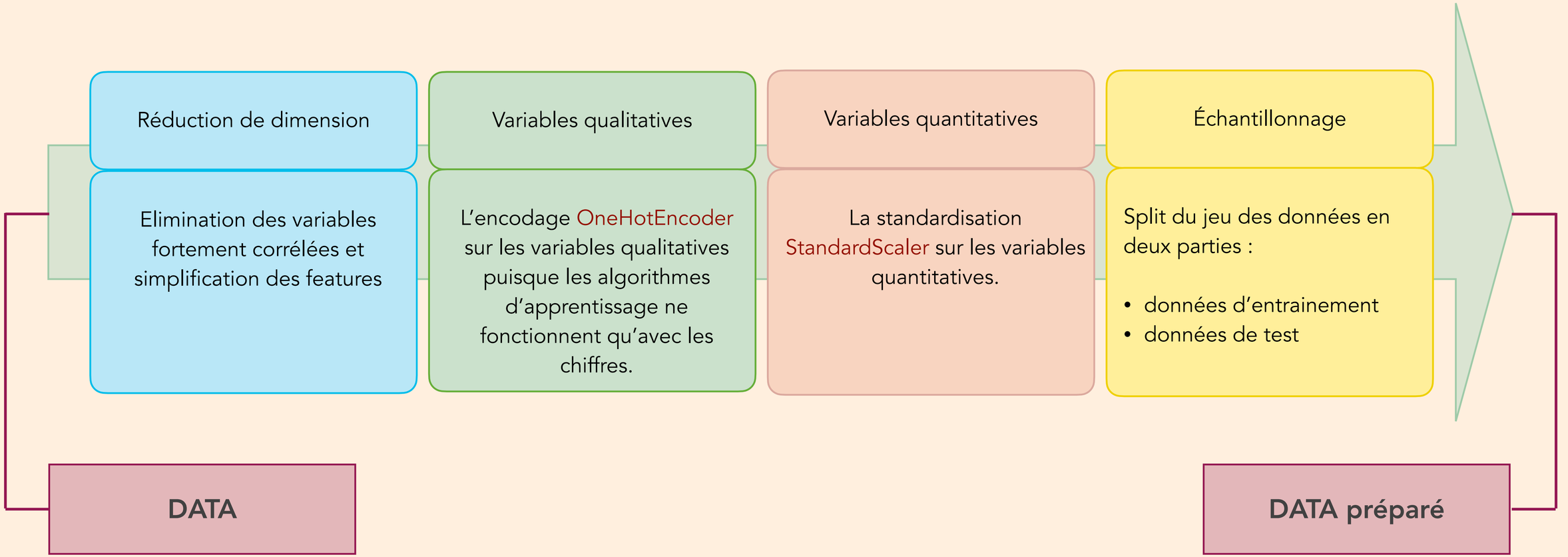
Categorie C.

Analyse de la significativité des types de plus grande utilisation de la propriété **(LargestPropertyUseType)**



Preprocessing

- Préparation des données
  - Encodage
  - Standardisation
  - Échantillonnage



Les mesures de la performance

- R2 — coefficient de détermination

  Meilleur score = plus élevé

- RMSE — Racine de l'erreur quadratique moyenne

  Meilleur score = plus faible

- Temps moyen de calcul

R2 — coefficient de détermination

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Le modèle baseline

```
X = data[['BuildingAge', 'NumberofFloors']]
y = data['SiteEnergyUse(kBtu)'].values
```

Régression linéaire

```
Le score R^2 de la performance du modèle baseline : 0.1407
```

Modélisation

Supervisée

Régression

R2 : Meilleur score = plus élevé

Les modèles testés

Les scores obtenus ont été vérifiés par validation croisée (5 échantillonnages)

Modèles simples

- Linear Regression
- Lasso Regression
- Ridge Regression
- Elastic Net Regression
- Support Vector Regression
- k-Nearest Neighbours
- Decision Tree

Modèles ensemblistes

- Random Forest Regression
- Gradient Boosting Regressor

Avec une procédure de grille de recherche pour choisir les hyperparamètres

Les scores pour SiteEnergyUse(kBtu)

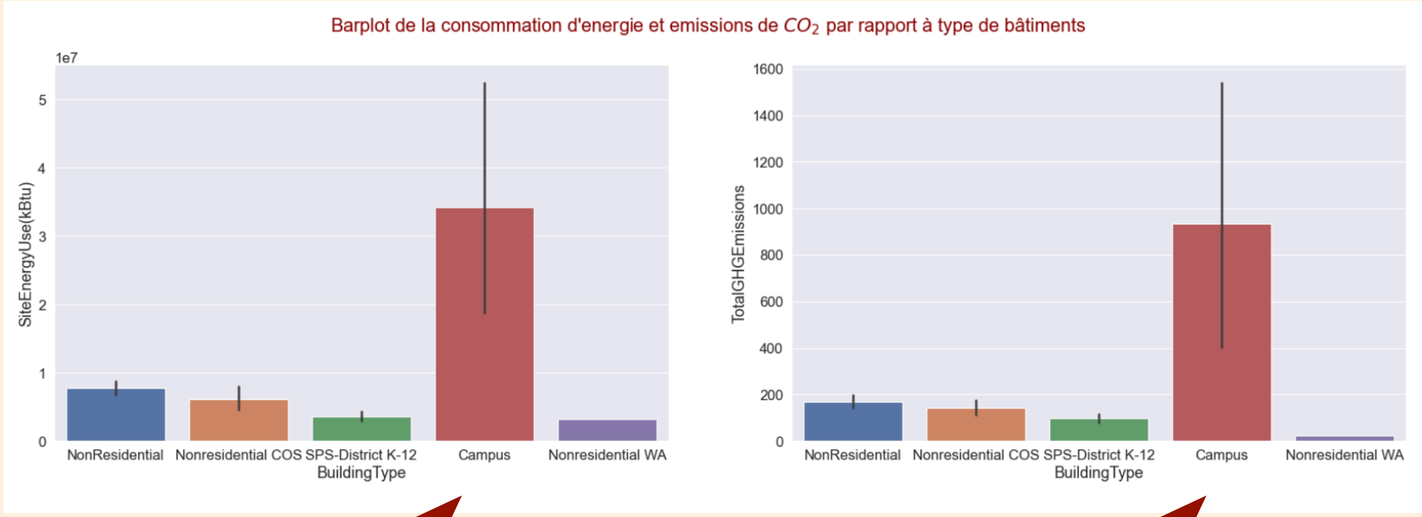
	modele	R2	RMSE	MAE	time
0	Gradient Boosting Regressor	0.752000	9.814537e+06	1174.021	0.596
1	Lasso	0.742000	1.002414e+07	1440.061	0.032
2	Linear Regression	0.741922	1.002083e+07	1438.220	0.003
3	Ridge	0.741000	1.003638e+07	1449.359	0.004
4	Random Forest Regressor	0.735000	1.015959e+07	1129.036	0.437
5	Elastic Net	0.683000	1.110176e+07	1329.593	0.003
6	Decision Tree Regressor	0.617000	1.221106e+07	1449.090	0.006
7	K Neighbors Regressor	0.600000	1.247193e+07	1115.295	0.001
8	SVM	0.350000	1.590580e+07	1032.102	0.002

Les scores pour TotalGHGEmissions

	modele	R2	RMSE	MAE	time
0	Gradient Boosting Regressor	0.886000	217.199	6.019	0.598
1	Random Forest Regressor	0.762000	313.931	5.925	0.472
2	Decision Tree Regressor	0.751000	320.962	7.412	0.006
3	Ridge	0.730000	334.075	7.894	0.003
4	Linear Regression	0.726465	336.200	8.014	0.002
5	Lasso	0.726000	336.231	7.462	0.005
6	K Neighbors Regressor	0.670000	369.322	5.071	0.001
7	Elastic Net	0.545000	433.398	7.321	0.004
8	SVM	0.525000	443.235	5.526	0.078

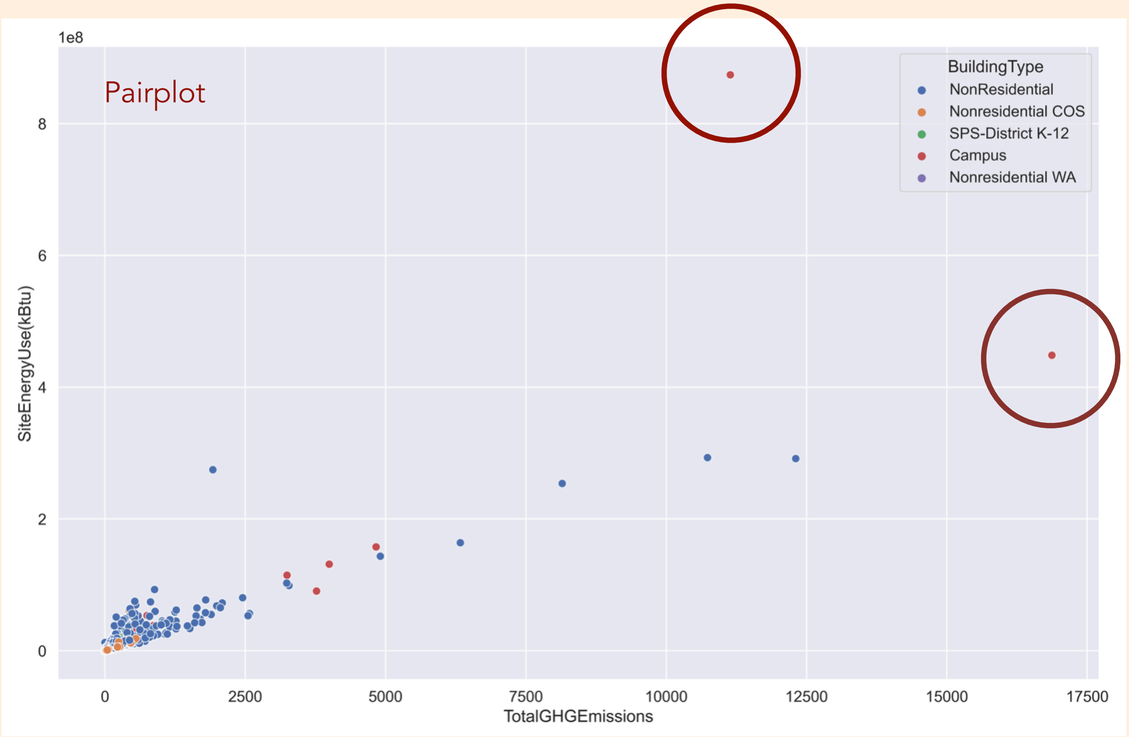
Optimisation

- Elimination des outliers (le problème de nombre de bâtiments)
- La methode ANOVA



1650	49967	Campus	University	University of Washington - Seattle Campus	4	Northeast	1900	111	0
102	172	Campus	University	SSCC MAIN CAMPUS	a	Delridge		27	2
1112	23622	Campus	Other	FT C15 Fishermen's Center	g	Magnolia / queen anne		23	1
158	261	Campus	Large Office	South Park	a	Greater duwamish		14	2
1337	25251	Campus	University	5th Avenue Master Meter	g	Magnolia / queen anne		14	2
124	211	Campus	University	NSCC MAIN CAMPUS	e	Northwest		11	2

Nombre de bâtiments



Le modèle sélectionné

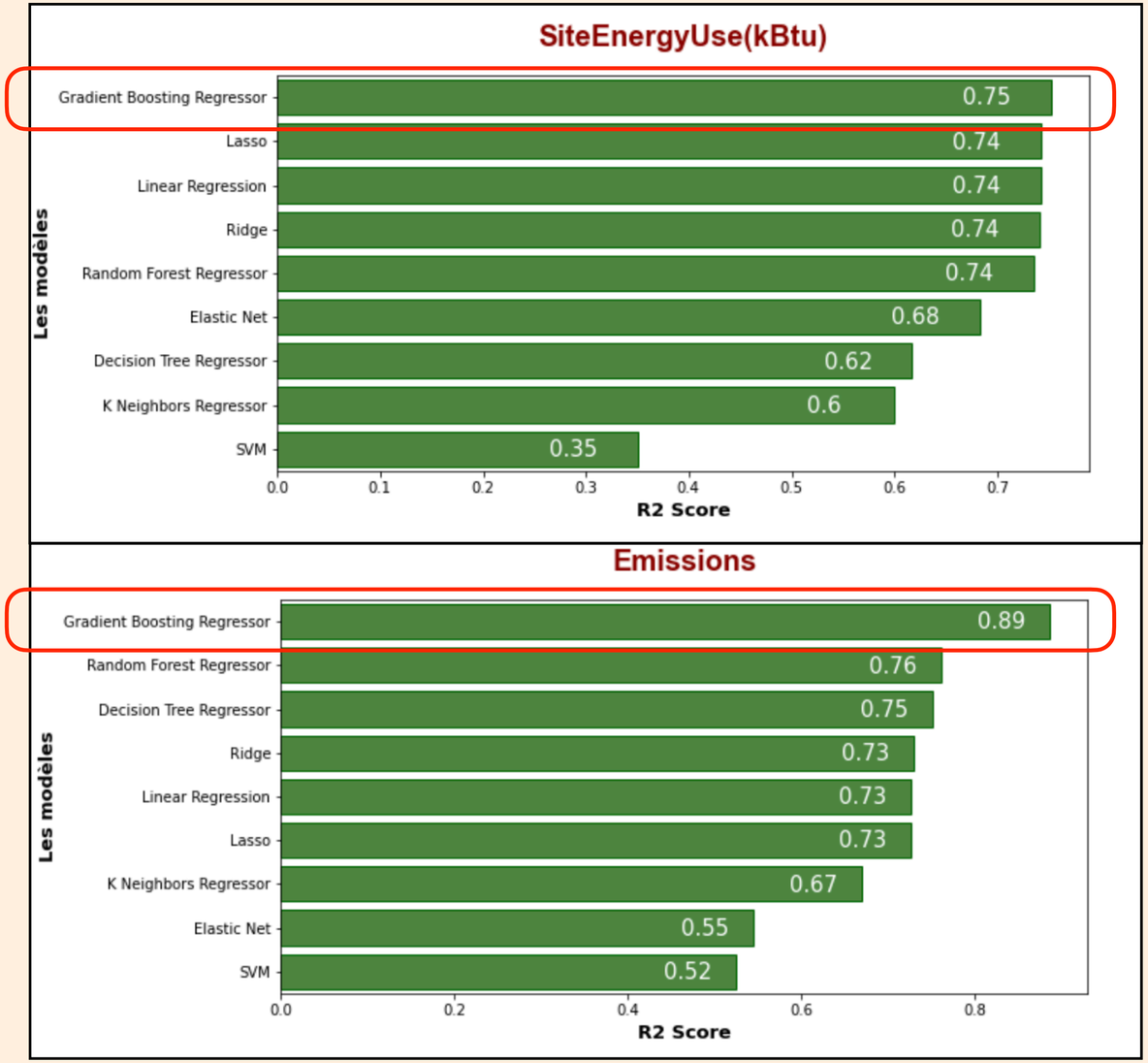
Gradient Boosting Regressor

- max_depth : profondeur maximale des estimateurs de régression individuelles

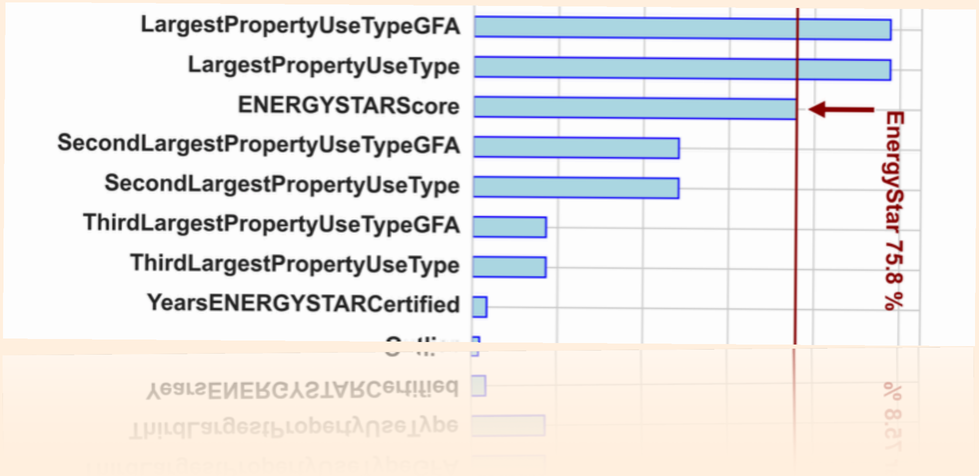
  max_depth = 3

- max_features: le nombre de caractéristiques à prendre en compte lors de la recherche du meilleur split.

  max_features = 'auto'

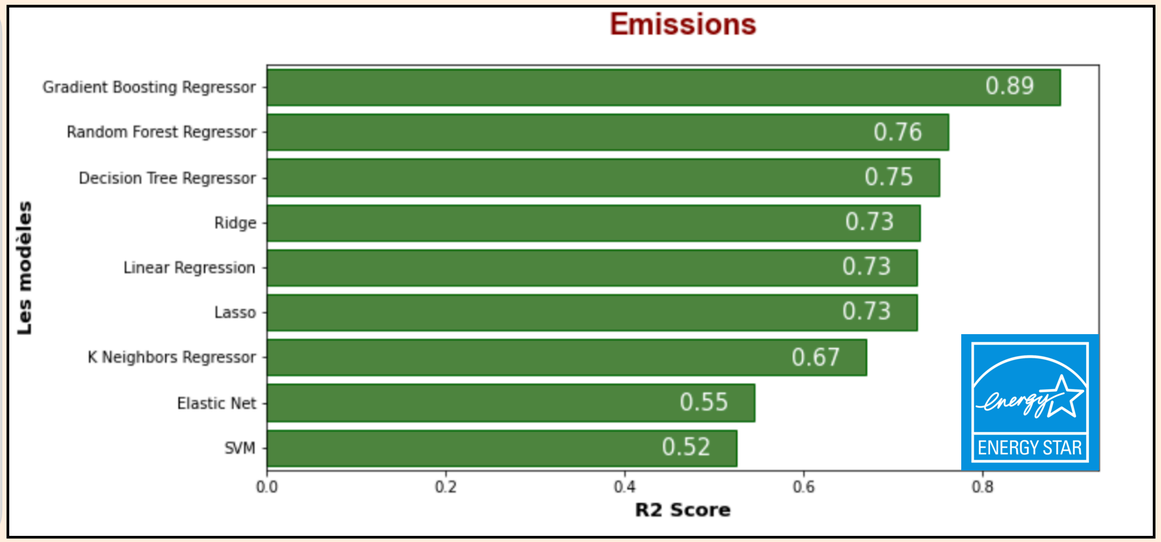


Intérêt de l'Energy Star Score

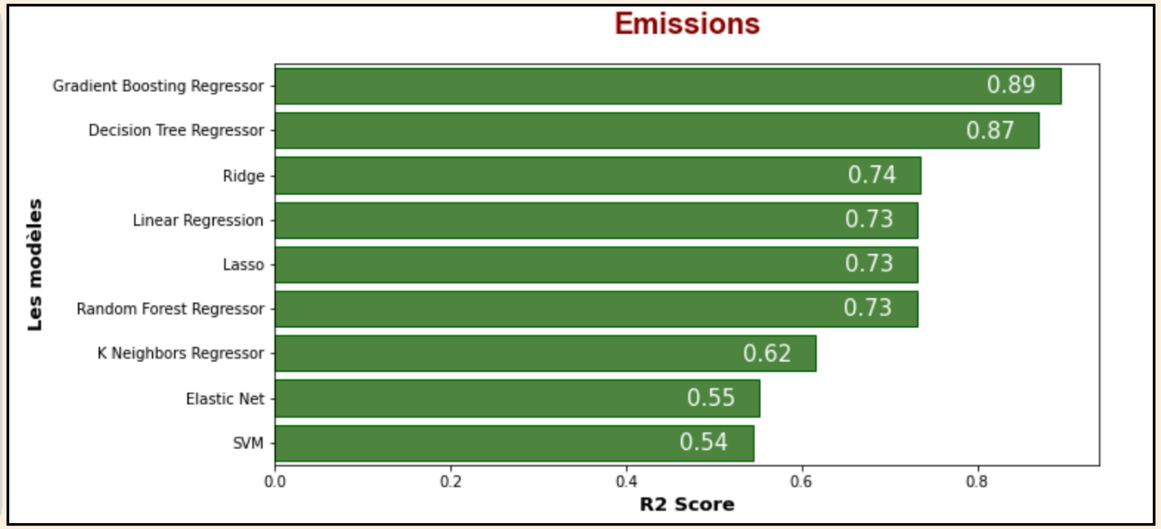


- Une échelle numérique de 0 à 100 (100 étant le meilleur score)
- L'Energy Star Score est un outil de dépistage aidant à évaluer les performances d'émission de CO2 d'un bâtiment par rapport aux établissements similaires
- Les prédictions de la consommation totale d'énergie AVEC la feature Energy Star Score sont légèrement améliorées
- Le feature ne présente que peu d'intérêt

Avec Energy STAR



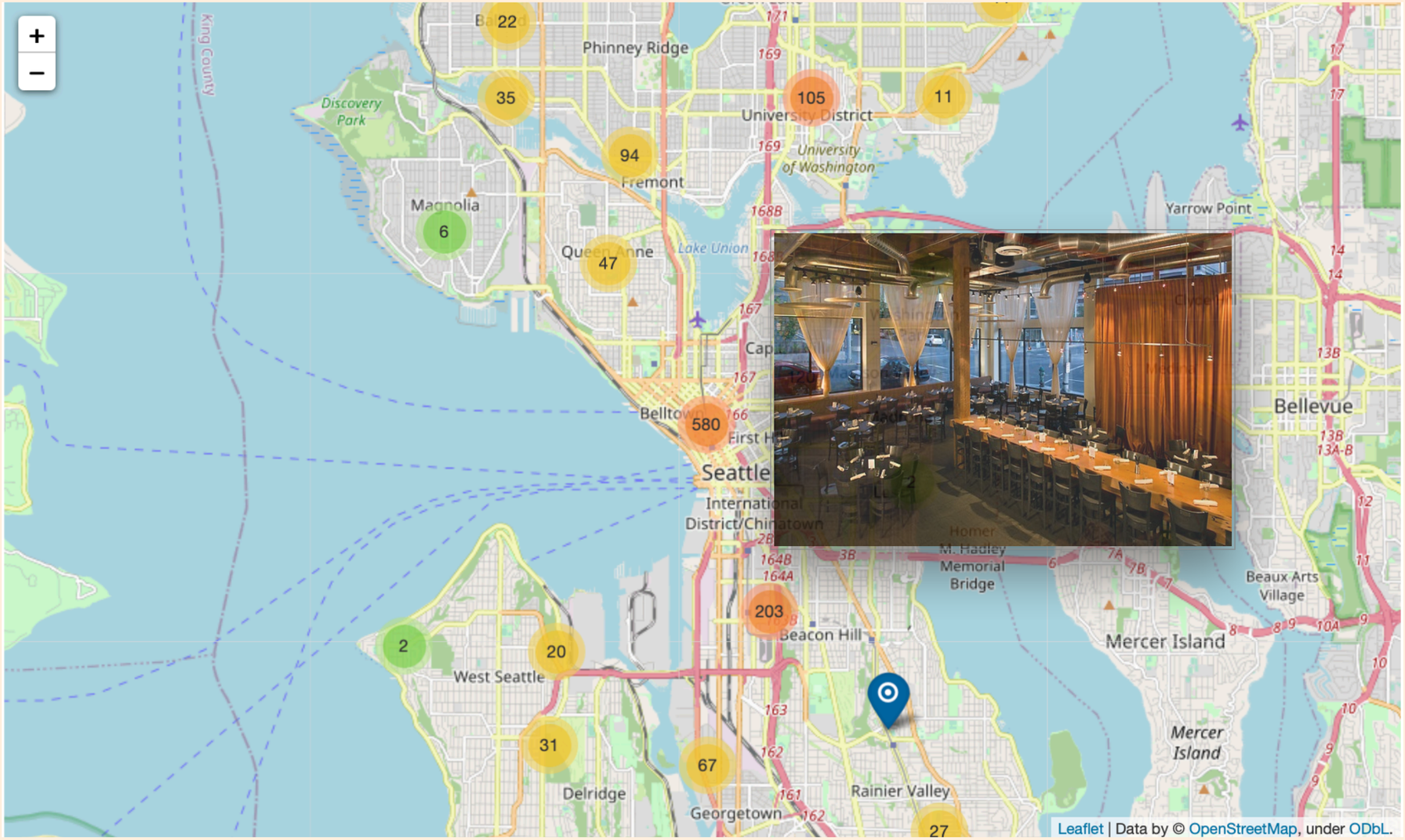
Sans Energy STAR



Une carte interactive de la ville de Seattle

https://yasarigno.github.io/seattle_folium_map.html

- grace à la librarie FOLIUM



Merci de votre attention