

DATA SCIENTIST



***ANTICIPEZ LES BESOINS
EN CONSOMMATION
DE BÂTIMENTS***



Réalisée par: Kchaou Mariem
Encadré par: Mr Wilfried Josset

Août 2022



Présentation



Traitement des données



Feature engineering



Modélisation



Conclusion



Présentation de la problématique



- Objectif de la ville de Seattle:
neutre en émissions de carbone en 2050.
- Relevés manuels minutieux effectués par nos agents en 2016.
- Ces relevés sont coûteux à obtenir.
- Il reste encore des bâtiments à mesurer.
- Tenir 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 une base de données. 3



Découverte du jeu de données

Dataset

3376 lignes
46 colonnes

Information géographique

- Adresse
- Coordonnées GPS
- Code postale
- Etc

Décrivants caractéristiques

- Le type d'utilisation
- Le nombre d'étages et bâtiments
- La surface brute
- Etc

Variables énergétiques

- La consommation électrique
- La consommation de gaz
- L'émissions de CO2
- Etc

Nettoyage de données



OSEBuildingID				
DataYear				
BuildingType				
PrimaryPropertyType				
PropertyName				
Address				
City				
State				
ZipCode				
TaxParcelIdentificationNumber				
CouncilDistrictCode				
Neighborhood				
Latitude				
Longitude				
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				
Comments				
ComplianceStatus				
Outlier				
TotalGHGEmissions				
GHGEmissionsIntensity				



Nettoyage de données

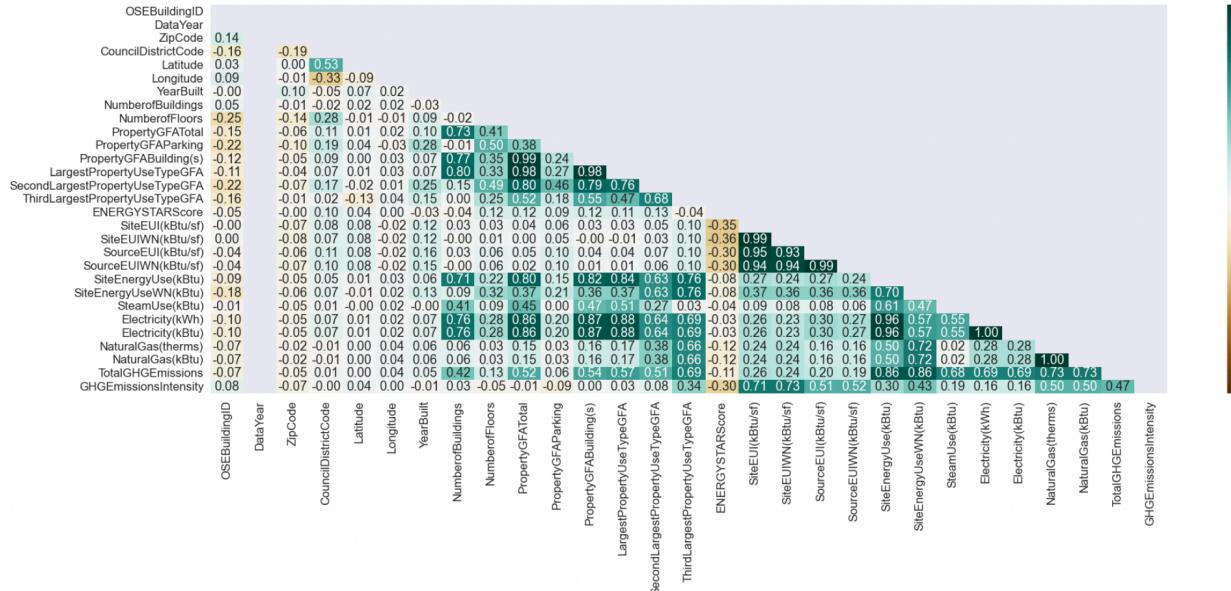
- Elimination des bâtiments résidentiels.
- Suppression des colonnes qui ne sont pas pertinentes.

BuildingType	
NonResidential	1460
Multifamily LR (1-4)	1018
Multifamily MR (5-9)	580
Multifamily HR (10+)	110
SPS-District K-12	98
Nonresidential COS	85
Campus	24
Nonresidential WA	1

Multifamily

Nettoyage de données

Heatmap des corrélations

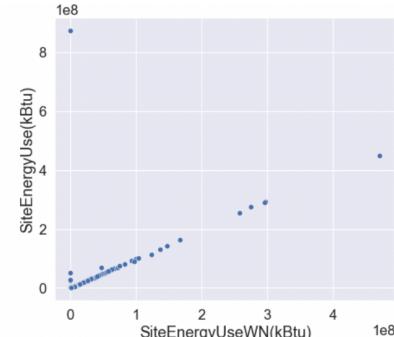
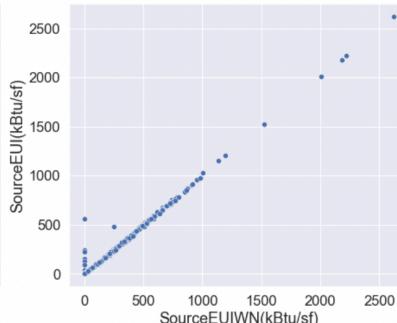
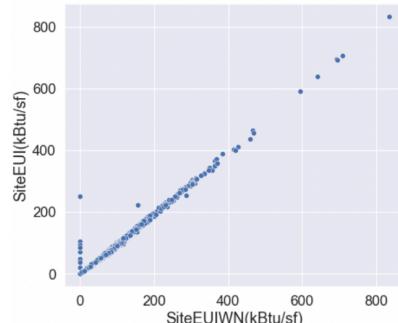
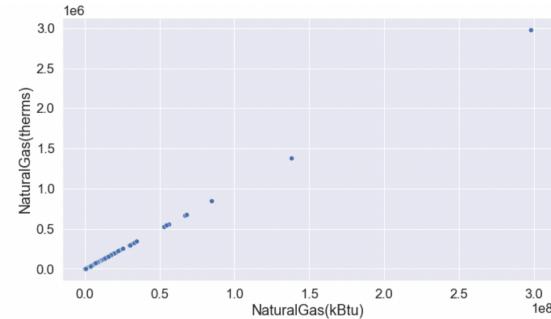
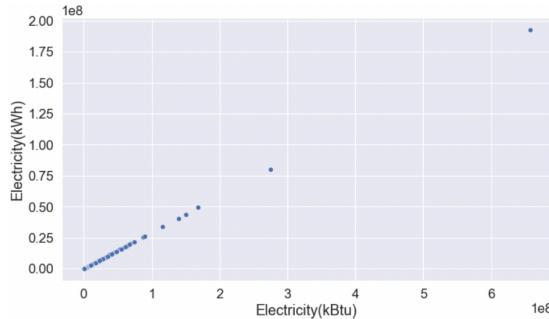


- Suppression des colonnes fortement corrélés.



Nettoyage de données

- Elimination des variables avec le suffixe « WN » (weather normalized).
- Suppression des variables qui sont fournies en plusieurs unités.





Traitement des valeurs manquantes

ThirdLargestPropertyUseTypeGFA	1315
ThirdLargestPropertyUseType	1315
SecondLargestPropertyUseTypeGFA	813
SecondLargestPropertyUseType	813
ENERGYSTARScore	574
ZipCode	16
LargestPropertyUseType	6
LargestPropertyUseTypeGFA	6
SiteEUI(kBtu/sf)	3
ListOfAllPropertyUseTypes	2
TotalGHGEmissions	2
NaturalGas(kBtu)	2
Electricity(kBtu)	2
SteamUse(kBtu)	2
SiteEnergyUse(kBtu)	2
SourceEUI(kBtu/sf)	2
GHGEmissionsIntensity	2
NumberofBuildings	2
DataYear	0
Neighborhood	0
BuildingType	0
ComplianceStatus	0
PrimaryPropertyType	0
PropertyName	0
Address	0
CouncilDistrictCode	0
Latitude	0
PropertyGFABuilding(s)	0
Longitude	0
YearBuilt	0
NumberofFloors	0
PropertyGFATotal	0
PropertyGFAParking	0
OSEBuildingID	0
dtype:	int64

- Complétiion des valeurs manquantes quand applicable.
- Suppression des observation pour lesquelles on a beaucoup de NAN pour conserver un maximum des features.
- Suppression des lignes qui ont un peu des valeurs manquantes.

Traitement des valeurs manquantes

OSEBuildingID	0
ComplianceStatus	0
NaturalGas(kBtu)	0
Electricity(kBtu)	0
SteamUse(kBtu)	0
SiteEnergyUse(kBtu)	0
SourceEUI(kBtu/sf)	0
SiteEUI(kBtu/sf)	0
ThirdLargestPropertyUseTypeGFA	0
ThirdLargestPropertyUseType	0
SecondLargestPropertyUseTypeGFA	0
SecondLargestPropertyUseType	0
LargestPropertyUseTypeGFA	0
LargestPropertyUseType	0
ListofAllPropertyUseTypes	0
TotalGHGEmissions	0
PropertyGFABuilding(s)	0
PropertyGFATotal	0
DataYear	0
BuildingType	0
PrimaryPropertyType	0
PropertyName	0
Address	0
ZipCode	0
PropertyGFAParking	0
CouncilDistrictCode	0
Neighborhood	0
Latitude	0
Longitude	0
YearBuilt	0
NumberofBuildings	0
NumberofFloors	0
GHGEmissionsIntensity	0
ENERGYSTARScore	556
	dtype: int64

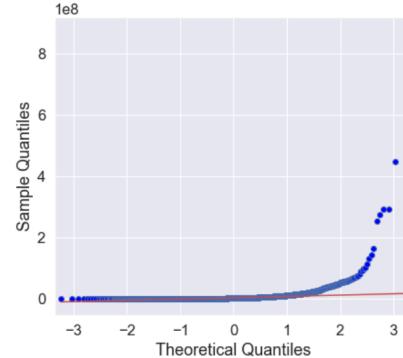
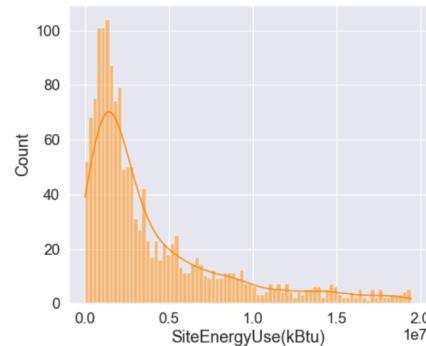
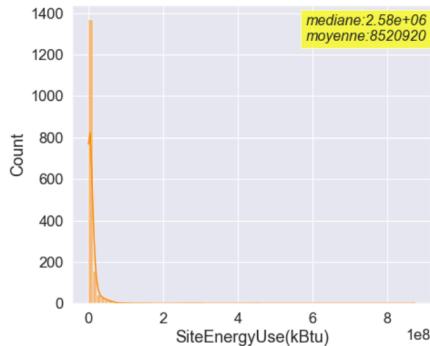
```
data.reset_index(drop=True)
print('Notre jeu des données après nettoyage et imputation compte {} colonnes et {} lignes.'.format(data.shape[1], data.shape[0]))
```

Notre jeu des données après nettoyage et imputation compte 34 colonnes et 1645 lignes.

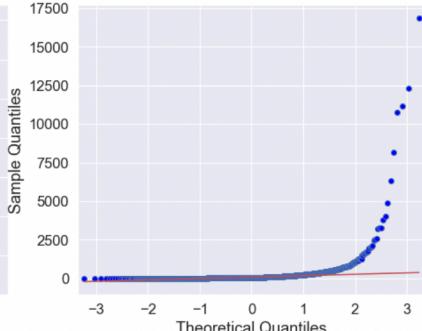
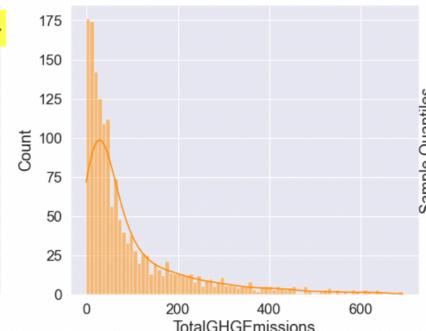
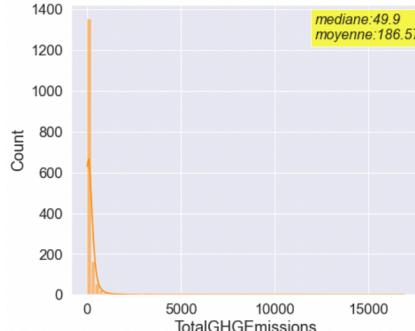


Analyse exploratoire

La consommation totale d'énergie: SiteEnergyUse(kBtu)



Les émissions de CO₂: TotalGHGEmissions



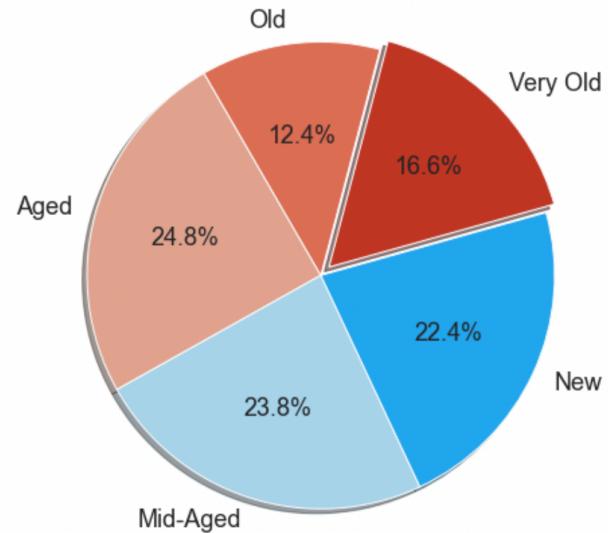
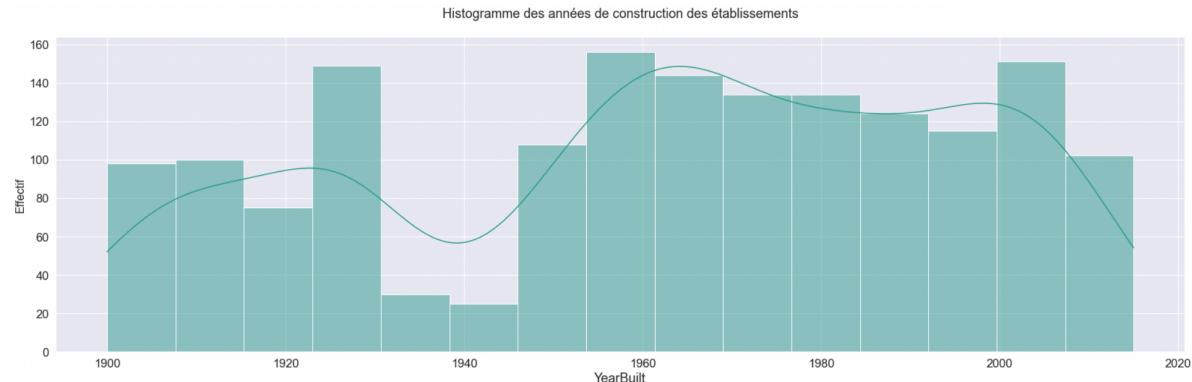
Les distributions des deux variables ne sont pas Gaussiennes.

```
TotalGHGEmissions (normaltest):
Statistics=3119.302, p=0.000
Sample does not look Gaussian (reject H0)
=====
TotalGHGEmissions (Shapiro):
Statistics=0.178, p=0.000
Sample does not look Gaussian (reject H0)
=====
SiteEnergyUse(kBtu) (normaltest):
Statistics=3506.146, p=0.000
Sample does not look Gaussian (reject H0)
=====
SiteEnergyUse(kBtu) (Shapiro):
Statistics=0.195, p=0.000
Sample does not look Gaussian (reject H0)
```

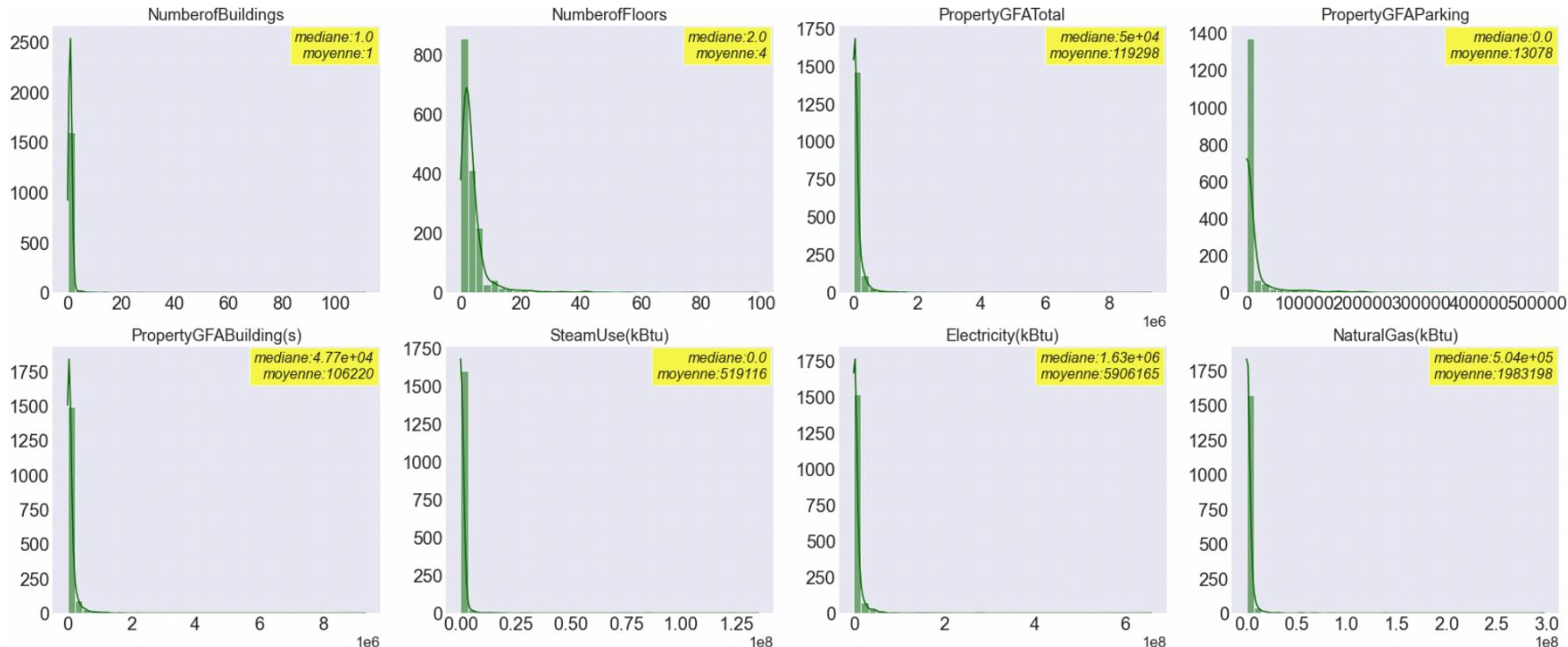


Analyse exploratoire

Les ages des établissements en 5 catégories

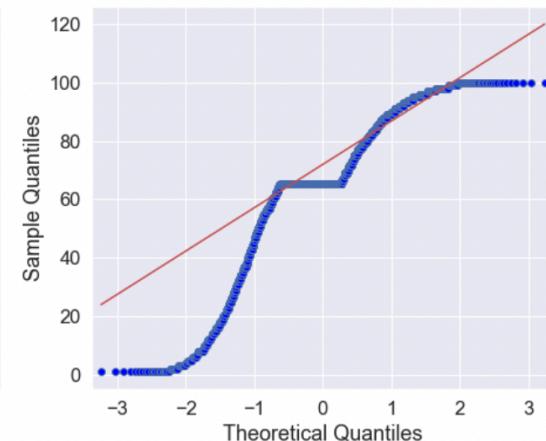
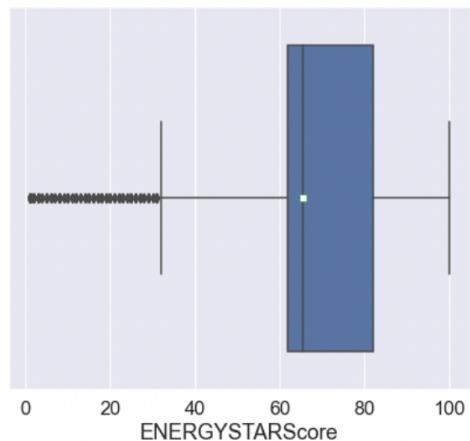
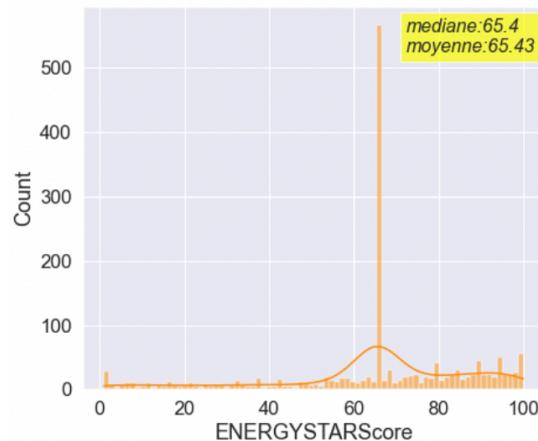


Analyse exploratoire





Analyse exploratoire





Nouveaux features attributs

- Transformation des données en la forme la plus appropriée pour les algorithmes de machine learning.
- 2 catégories de nouveaux features.

Catégorie 1

Les variables qui vérifient le type de consommation:

- Electricity_check
- NaturalGaz_check
- SteamUse_check

La valeur de **NaturalGaz_check** est 1 si la consommation de gaz est un nombre positif, Sinon elle est 0.

Catégorie 2

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

- LP_Ratio (LargestPropertyUse)
- SPL_Ratio (SecondLargestPropertyUse)
- TPL_Ratio (ThirdLargestPropertyUse)



Preprocessing

Préparation des données:

- Encodage
- Standardisation
- Échantillonnage

Réduction de dimension

Variables qualitatives

Variables quantitatives

Échantillonnage

Elimination des variables fortement corrélées et simplification des features.

L'encodage **OneHotEncoder** sur les variables qualitatives puisque les algorithmes d'apprentissage ne fonctionnent qu'avec les chiffres.

La standardisation **StandardScaler** sur les variables quantitatives.

Split du jeu de données en deux parties:
- Données d'entraînement
- Données de test



Les mesures de la performance

- R2 — coefficient de détermination
Meilleur score = plus élevé
- RMSE — Racine de l'erreur quadratique moyenne
Meilleur score = plus faible
- MAE — l'erreur absolue moyenne
Meilleur score = plus faible
- Temps moyen de calcul

R2 — coefficient de détermination

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

Introduction

Traitement des données

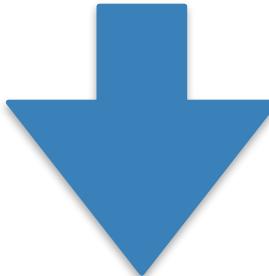
Feature engineering

Modélisation

Conclusion

Le modèle baseline

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



Régression linéaire

Le score R^2 de la performance du modèle baseline : 0.0885

Les modèles testés

Linear Regression

Lasso Regression

Ridge Regression

Elastic Net Regression

Support Vector Regression

K-Nearset Neighbors

Decision Tree Regression

Random Forest Regression

Gradient Boosting Regressor



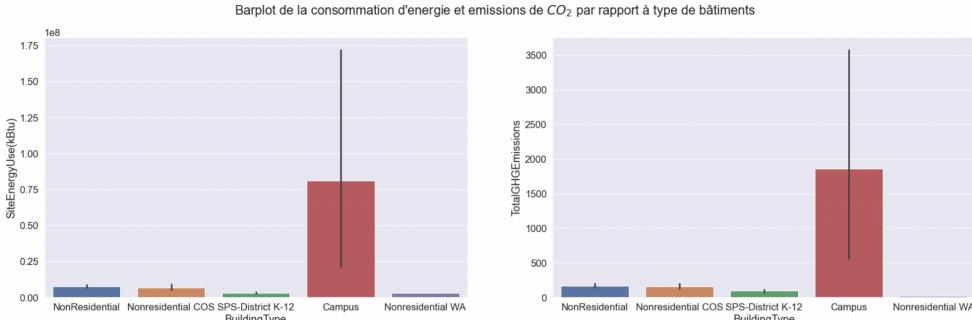
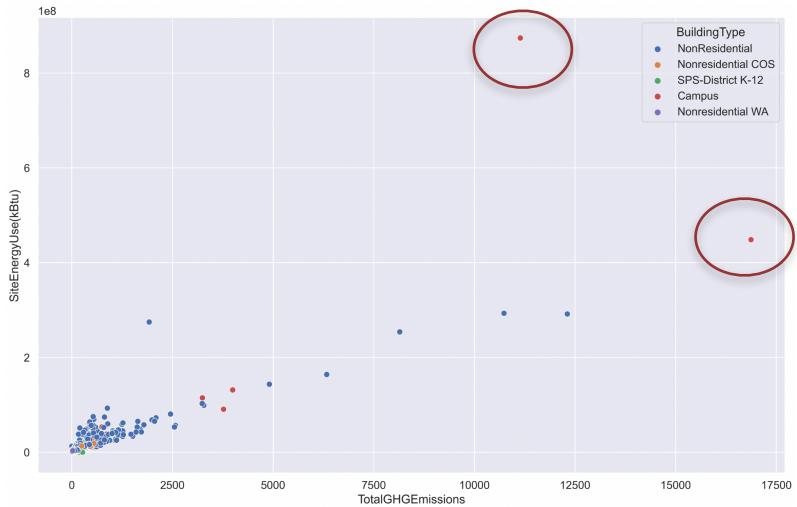
	modele	R2	RMSE	MAE	time
0	Decision Tree Regressor	0.802000	1.022694e+07	1961.977	0.005
1	Ridge	0.738000	1.174816e+07	1579.833	0.003
2	Linear Regression	0.736627	1.178431e+07	1597.547	0.004
3	Lasso	0.736000	1.179202e+07	1597.881	0.035
4	Gradient Boosting Regressor	0.629000	1.397918e+07	1132.154	0.466
5	Elastic Net	0.616000	1.422100e+07	1453.151	0.002
6	Random Forest Regressor	0.607000	1.440373e+07	1255.858	0.362
7	K Neighbors Regressor	0.545000	1.548909e+07	1233.758	0.001
8	SVR	0.210000	2.040832e+07	1001.576	0.003

	modele	R2	RMSE	MAE	time
0	Decision Tree Regressor	0.757000	379.289	5.868	0.005
1	Gradient Boosting Regressor	0.705000	417.629	6.311	0.475
2	Linear Regression	0.690609	427.736	8.738	0.003
3	Lasso	0.678000	436.282	8.638	0.020
4	Ridge	0.672000	440.463	8.959	0.002
5	Random Forest Regressor	0.552000	514.593	5.960	0.354
6	K Neighbors Regressor	0.547000	517.549	5.618	0.001
7	SVM	0.525000	529.956	9.633	0.052
8	Elastic Net	0.239000	670.897	7.952	0.002

Les scores pour SiteEnergyUse(kBtu)

Les scores pour TotalGHGEmissions

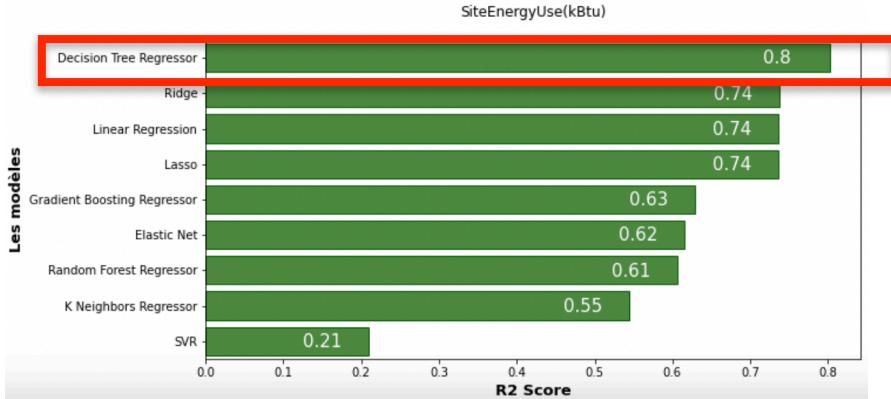
Optimisation



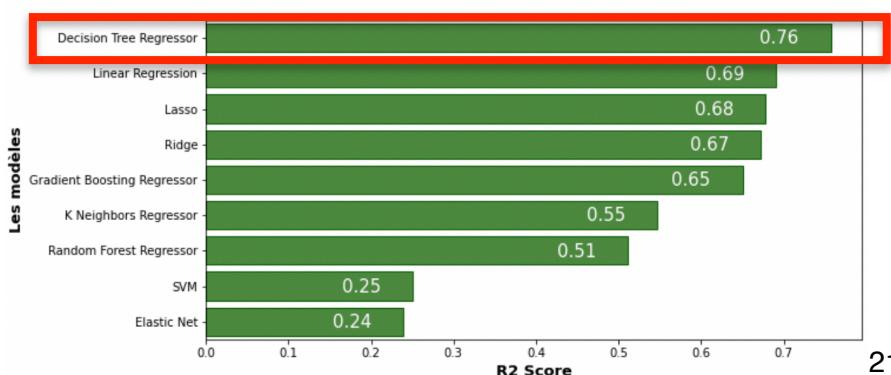
Le modèle sélectionné

Decision Tree Regression

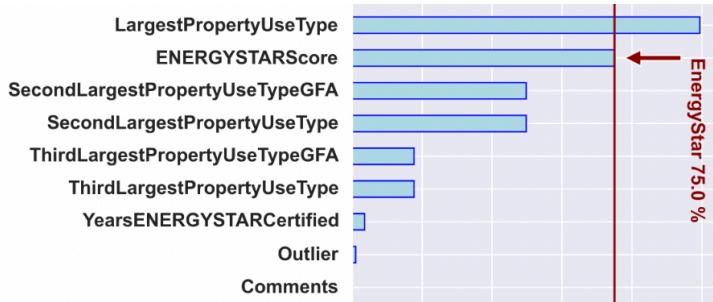
Les modèles



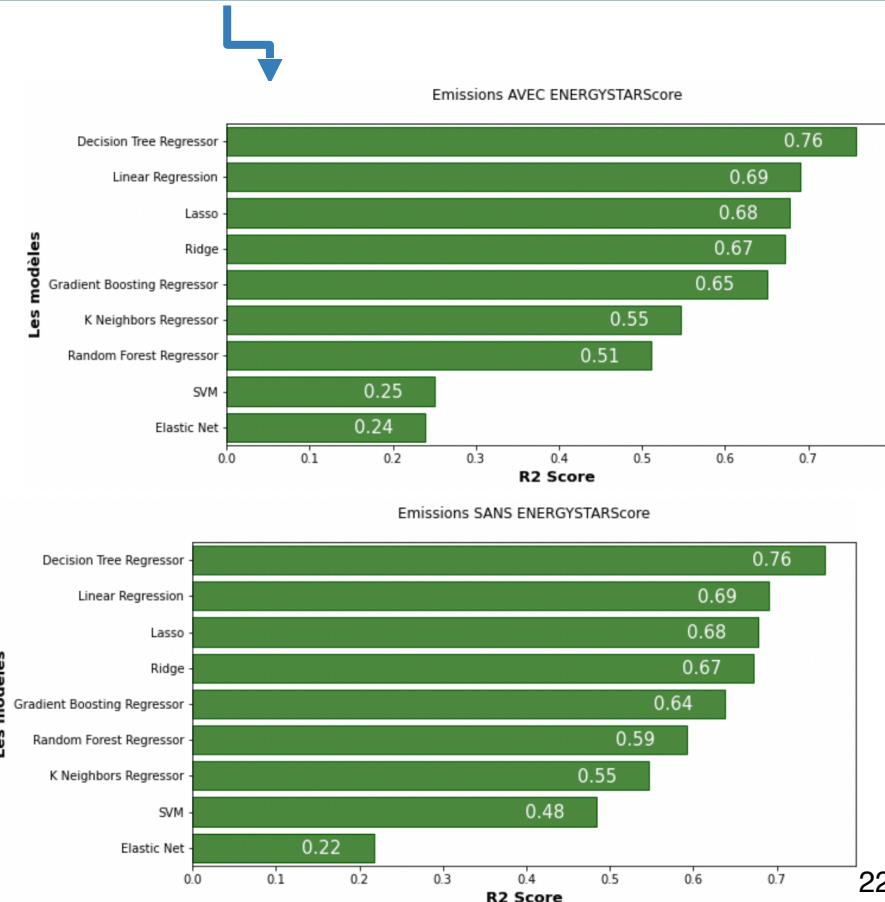
Les modèles



Intérêt de l'Energy Star Score



- Une échelle numérique de 0 à 100 (100 étant le meilleur score).
- L'Energie 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.
- La feature ne présente que peu d'intérêt.

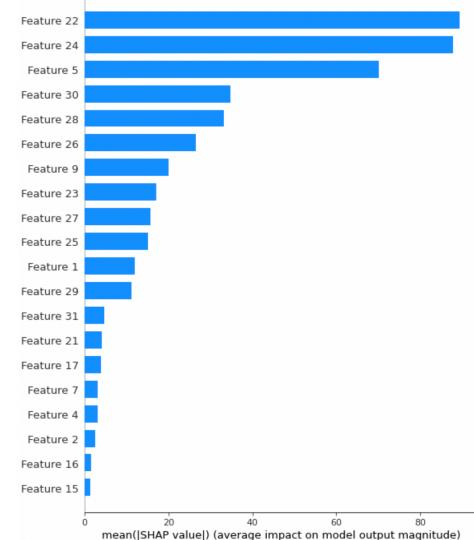
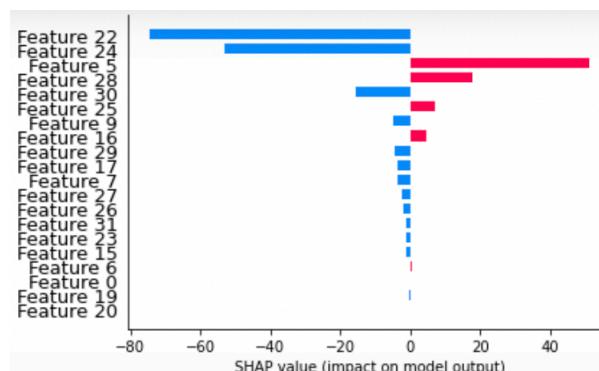




Explication du prédictions à l'aide des valeurs SHAP

Interprétation des prédictions avec SHAP.

Modèle : DecisionTreeRegressor()



Introduction

Traitement des données

Feature engineering

Modélisation

Conclusion



- Les résultats sont globalement pas mal. Cela est en partie dû aux données dont nous disposons en entrée.
- Il serait bien d'avoir quelques informations techniques du type:
 - Travaux de rénovation récents
 - Type d'isolation, d'éclairage (LED,...)
 - Type de chauffage

*Merci pour votre
attention*



Questions