Prédire les émissions de CO2 et la consommation totale d'énergie de bâtiments non destinés à l'habitation à Seattle

Soutenance Projet 4 Data Scientist Agnès Regaud juillet 2023

Sommaire

1.Problématique

Emissions de CO2 et consommation totale d'énergie

- Relevés minutieux effectués en 2016
- Bâtiments non résidentiels à Seattle
- But des modèles proposés :
  - ➢ Prédire la consommation totale d'énergie et les émissions de CO2
  - ➢ Evaluer l'intérêt de l'"ENERGY STAR Score" pour la prédiction d'émissions (et la consommation d'énergie)

2. Etude du jeu de données

Etude préliminaire du fichier de données de la ville de Seattle en Open Source

Fichier source	- Nom du fichier : 2016_Building_Energy_Benchmarking.csv - 3376 lignes (bâtiments) et 46 colonnes (variables) - Pas de ligne/colonne dupliquée - Une colonne entièrement vide : « Comments »
Colonnes « target »	- « SiteEnergyUse(kBtu) » : consommation d'énergie totale - « TotalGHGEmissions » : émissions de CO2
Nettoyage	- Sélection de la catégorie « NonResidential » (non prise en compte de « Nonresidential COS » et « Nonresidential WA » - Suppression de 2 bâtiments sans données sur énergie consommée, émissions de CO2, électricité, gaz

3.Description et analyse de certaines variables

Etude des variables

- **Transformation de la colonne « YearBuilt » en « NumberOfYearsSinceConstruction »,**

- Création d'une variable « energy_source » avec pour valeurs possibles : 'Electricity', 'Gaz', 'Electricity & Gaz', 'Source non connue' et suppression des variables 'Electricity(kWh)' et 'NaturalGas(therms)',

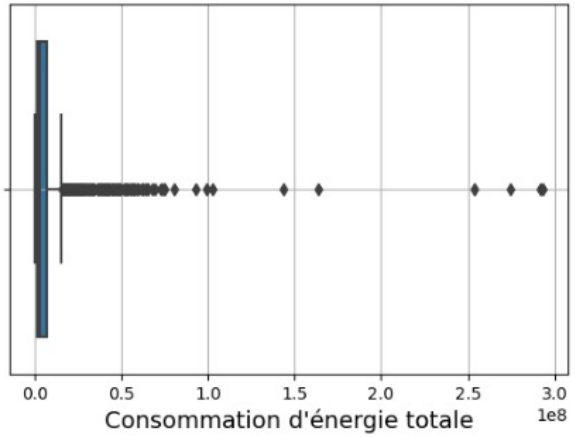
- Sélection des variables pertinentes pour l'étude : "NumberOfYearsSinceConstruction", "NumberofBuildings", "NumberofFloors", "PropertyGFAParking", "ENERGYSTARScore », "PrimaryPropertyType", "Neighborhood", 'energy_source',

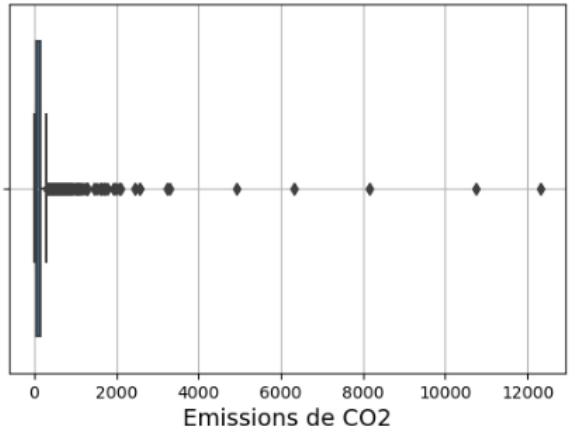
- #target1 : "SiteEnergyUse(kBtu)" est la consommation totale d'énergie
#target2 : "TotalGHGEmissions" est la quantité d'émission de CO2

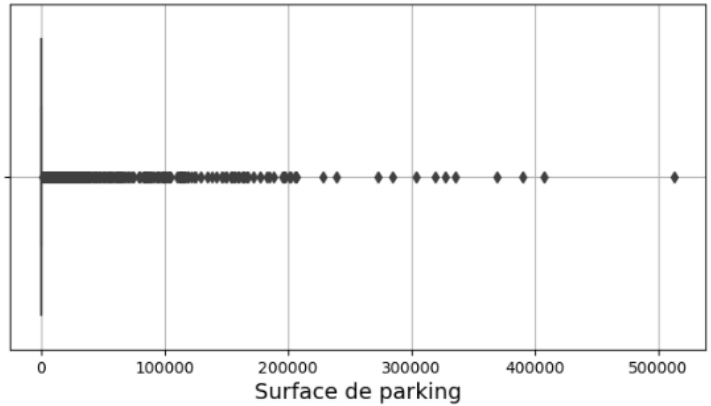
Etude des ouliers pour quelques variables

- **Colonne « Outlier » : suppression des bâtiments « Low outlier » et « High outlier » (16) puis suppression de la colonne**

- **Colonnes « SiteEnergyUse(kBtu) », « TotalGHGEmissions » et « PropertyGFAParking » :**



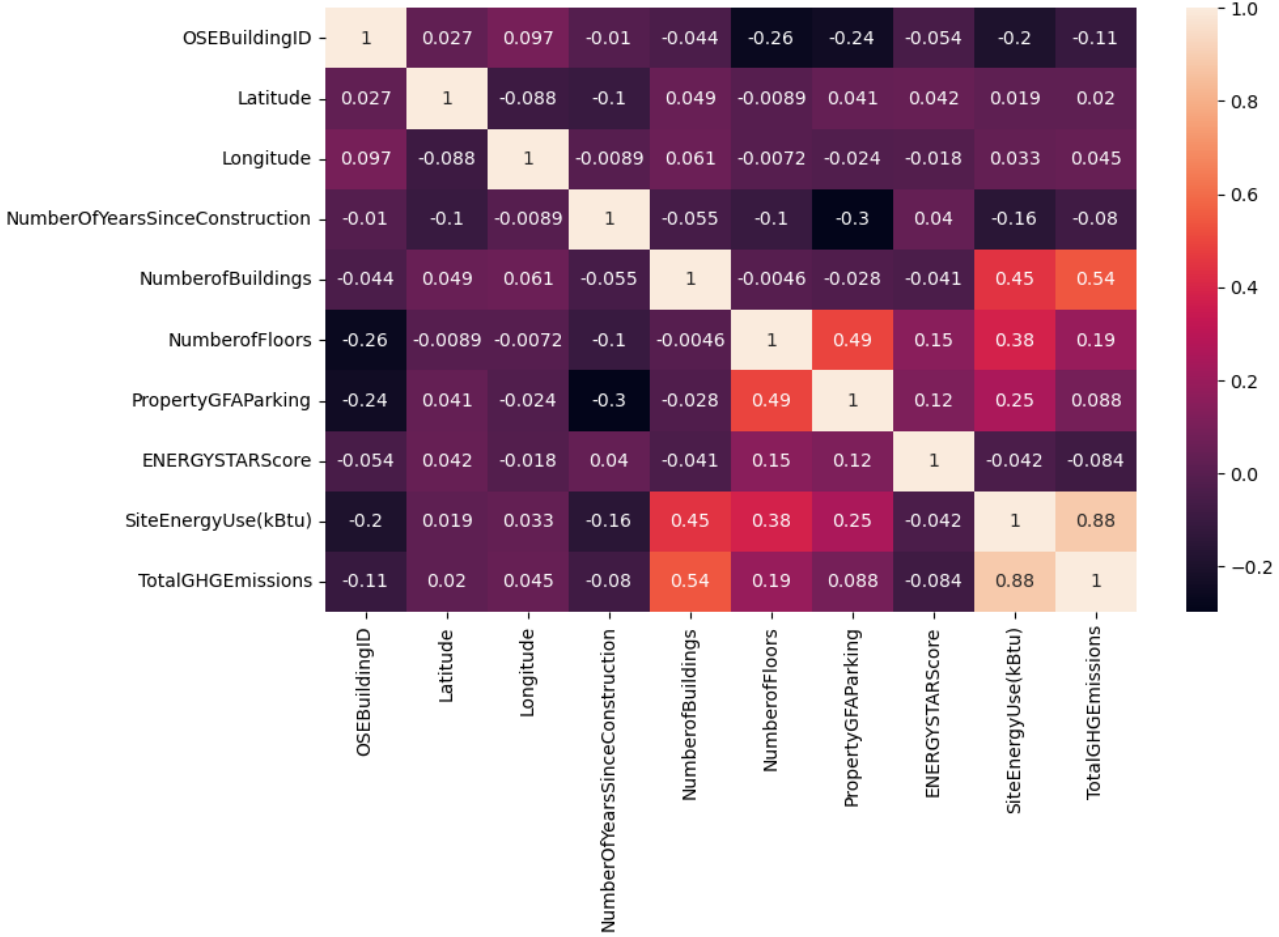




Pour chaque variable, les outliers sont des bureaux, hôtels, ou hôpitaux.
Pas de remarque particulière.

4.Lien entre les variables

Heatmap



On observe une corrélation certaine entre les 2 targets : consommation totale d'énergie (SiteEnergyUse(kBtu)) et émission de CO2 (TotalGHGEmissions)

Et SiteEnergyUse(kBtu) avec : NumberofBuildings, NumberofFloors, PropertyGFAParking,

et TotalGHGEmissions avec NumberofBuildings et NumberofFloors

5.Modèles de prédiction

Traitements et modèles pour les 2 targets

Une variable comporte des valeurs manquantes : "ENERGYSTARScore »

GridSearchCV et **Pipeline** avec :

- ✓ **Préprocessing** :
  - ❑ Variables catégorielles : OneHotEncoder
  - ❑ Variables numériques : SimpleImputer(strategy='median') et StandardScaler

- ✓ **Modèles** :

	Consommation d'énergie	Émissions de CO2
DummyRegressor()	strategy : ['mean', 'median']	strategy : ['mean', 'median']
LinearRegression()	fit_intercept : [True, False] n_jobs : [None, -1]	fit_intercept : [True, False] n_jobs : [None, -1]
Ridge()	alpha : [0,0.5,1]	alpha : [0,0.5,1]
Lasso()	alpha : [0,0.5,1]	alpha : [0,0.5,1]
RandomForestRegressor()	n_estimators : [100,110,120,150] random_state : [0, 42] max_features : [5,7,10]	max_features : [4,5,6] n_estimators :[100,150,200] random_state : [0, 42]
KNeighborsRegressor()	n_neighbors : list(range(1,31) weights : ['uniform', 'distance']	n_neighbors : list(range(4,6)
GradientBoostingRegressor()	learning_rate : [0.086,0.087,0.088, 0.09] max_depth : [4,5,6,7] random_state : [0,42] max_features : [5,6,7]	learning_rate : [0.086,0.087,0.088, 0.09] max_depth : [4,5,6,7] random_state : [0,42] max_features : [5,6,7]
SVR()	kernel : ('linear', 'rbf','poly') C : [1.5, 10] gamma : [1e-7, 1e-4] epsilon : [0.1,0.2,0.5,0.3]	kernel : ('linear', 'rbf','poly') C : [1.5, 10] gamma : [1e-7, 1e-4] epsilon : [0.1,0.2,0.5,0.3]

Meilleurs Hyper paramètres pour : Consommation d'énergie

Modèle	Hyper paramètres	Score Train	Score test
DummyRegressor()	strategy : 'mean'	0.0	-1.2169
LinearRegression()	fit_intercept : False n_jobs : None	0.6215	0.3419
Ridge()	alpha : 0.5	0.6204	0.3465
Lasso()	alpha : 1	0.6215	0.3432
RandomForestRegressor()	max_features : 7 n_estimators : 120 random_state : 0	0.9397	0.4080
KNeighborsRegressor()	n_neighbors : 25 weights : 'distance'	0.9999	0.3506
GradientBoostingRegressor()	learning_rate : 0.086 max_depth : 7 max_features : 6 random_state : 42	0.9727	0.4238
SVR()	kernel : 'linear' C : 10 gamma : 1e-07 epsilon : 0.1	-0.0713	-0.0735

Remarque :
K-NN est en over-fitting

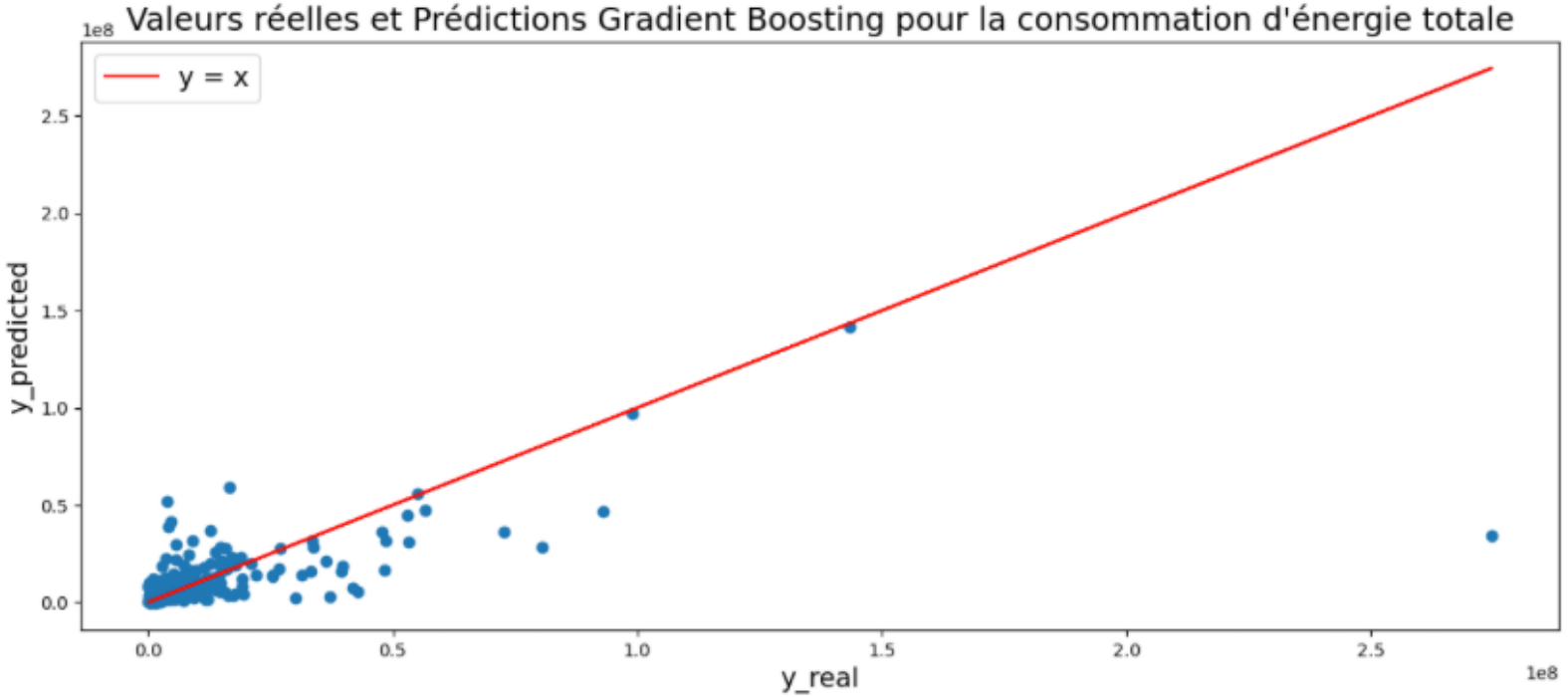
Meilleurs Hyper paramètres pour : Emission de CO2

Modèle	Hyper paramètres	Score Train	Score test
DummyRegressor()	strategy : 'mean'	0.0	-0.0006
LinearRegression()	fit_intercept : False n_jobs : None	0.6385	0.3046
Ridge()	alpha : 0.5	0.6372	0.3183
Lasso()	alpha : 1	0.6358	0.3178
RandomForestRegressor()	max_features : 4 n_estimators : 150 random_state : 0	0.9467	0.5307
KNeighborsRegressor()	n_neighbors : 5	0.6635	0.4365
GradientBoostingRegressor()	learning_rate : 0.09 max_depth : 6 max_features : 6 random_state : 42	0.9759	0.5966
SVR()	kernel : 'linear' C : 10 gamma : 1e-07 epsilon : 0.2	0.0543	0.1324

Performances des validations croisées pour le meilleur modèle : Gradient Boosting (10 premiers résultats)
1) Consommation d'énergie

	mean_test_score	std_test_score	param_gbt__learning_rate	param_gbt__max_depth	param_gbt__max_features	param_gbt__random_state
0	0.196031	0.740505	0.086	4	5	0
1	0.242458	0.730313	0.086	4	5	42
2	0.228552	0.788802	0.086	4	6	0
3	0.270883	0.533537	0.086	4	6	42
4	0.247276	0.633687	0.086	4	7	0
5	0.156181	0.953493	0.086	4	7	42
6	0.167648	0.726739	0.086	5	5	0
7	0.264776	0.642374	0.086	5	5	42
8	0.291310	0.596902	0.086	5	6	0
9	0.309816	0.484283	0.086	5	6	42

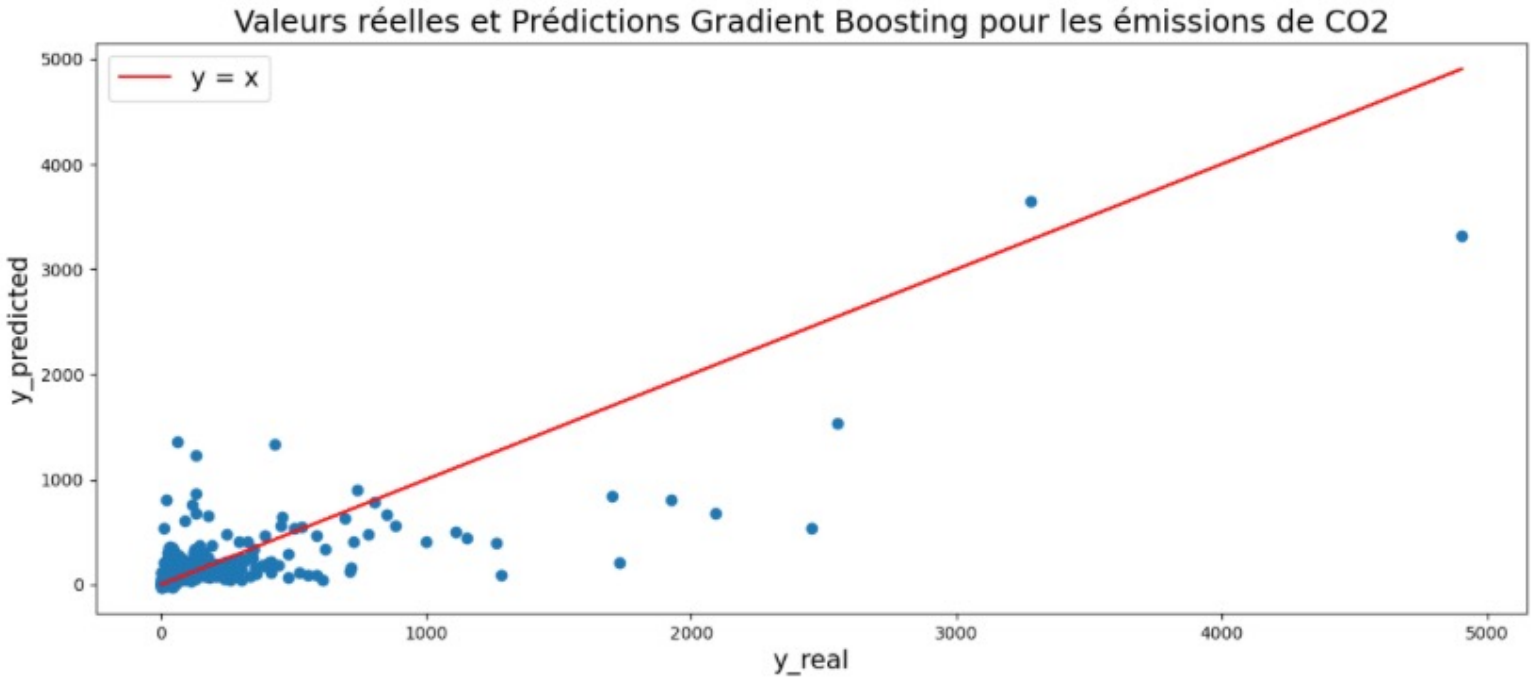
Visualisation performance Gradient Boosting sur consommation énergie totale



Performances des validations croisées pour le meilleur modèle : Gradient Boosting (8 premiers résultats)
2) Emissions de CO2

	mean_test_score	std_test_score	param_gbt__learning_rate	param_gbt__max_depth	param_gbt__max_features	param_gbt__random
0	0.270476	0.732215	0.086	4	5	
1	0.240981	0.894434	0.086	4	5	
2	0.152697	0.960075	0.086	4	6	
3	0.310077	0.641670	0.086	4	6	
4	0.129507	1.004799	0.086	4	7	
5	0.246045	0.704796	0.086	4	7	
6	0.289149	0.721175	0.086	5	5	
7	0.306998	0.711313	0.086	5	5	

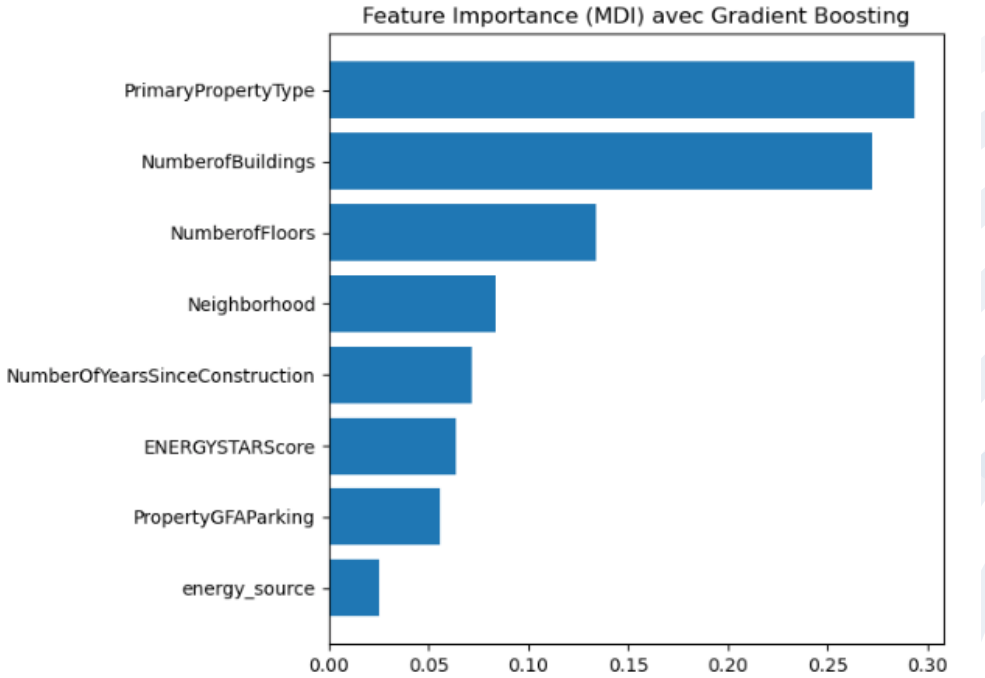
Visualisation performance Gradient Boosting sur émissions de CO2



6.Conclusion sur la variable ENERGY STAR Score

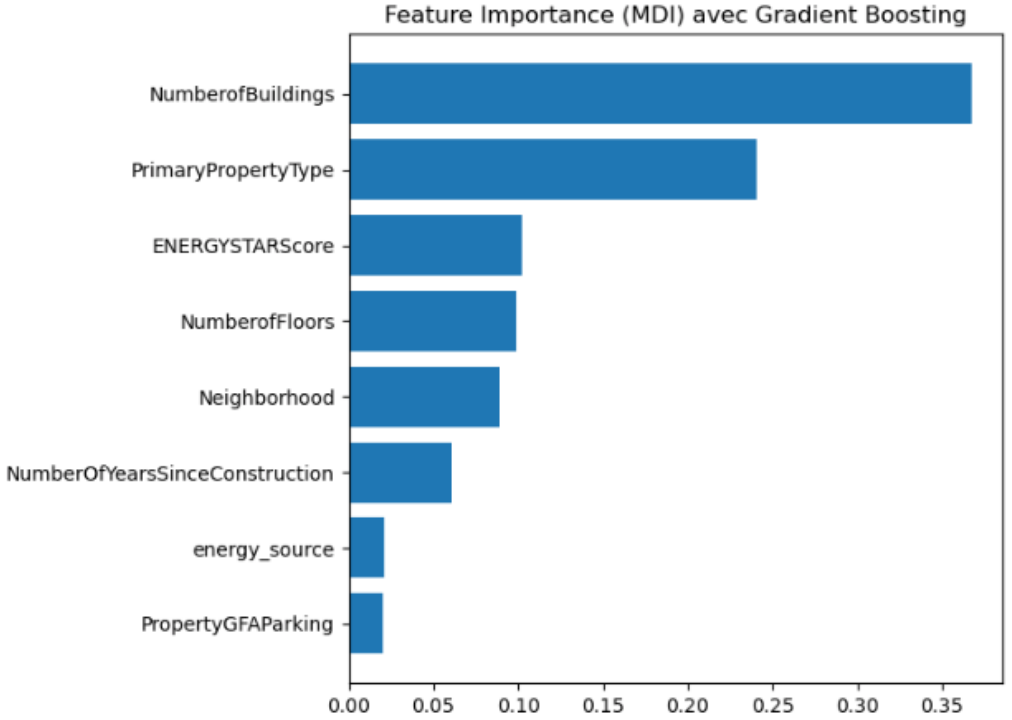
Feature Importance sur :
1) Consommation d'énergie totale

	features	importance
0	PrimaryPropertyType	0.293129
1	NumberofBuildings	0.272428
2	NumberofFloors	0.133927
3	Neighborhood	0.083701
4	NumberOfYearsSinceConstruction	0.071830
5	ENERGYSTARScore	0.063813
6	PropertyGFAParking	0.055882
7	energy_source	0.025290



Feature Importance sur :
2) Emissions de CO2

	features	importance
0	NumberofBuildings	0.366974
1	PrimaryPropertyType	0.240798
2	ENERGYSTARScore	0.102385
3	NumberofFloors	0.098595
4	Neighborhood	0.089300
5	NumberOfYearsSinceConstruction	0.061005
6	energy_source	0.021157
7	PropertyGFAParking	0.019786



SOURCE

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