VILLE DE SEATTLE :

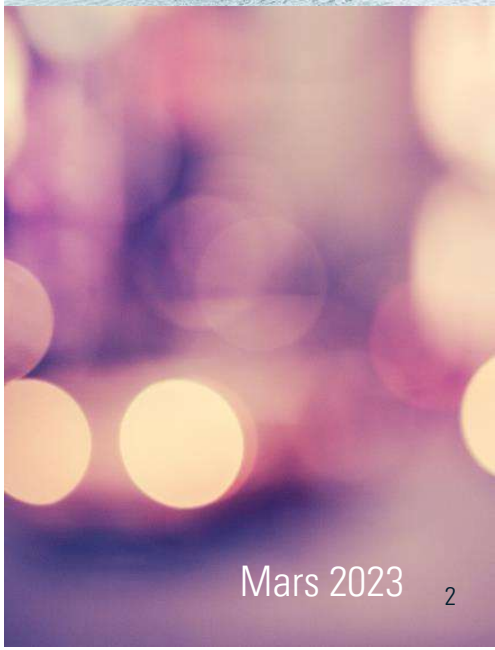
ANTICIPER LES BESOINS EN CONSOMMATION DES BÂTIMENTS NON RESIDENTIELS

Jay Corentin

Mars 2023

VILLE DE SEATTLE

- ✓ Analyse exploratoire
- ✓ Feature engineering
- ✓ Prédiction de la consommation
- ✓ Prédiction des émissions



Jay Corentin

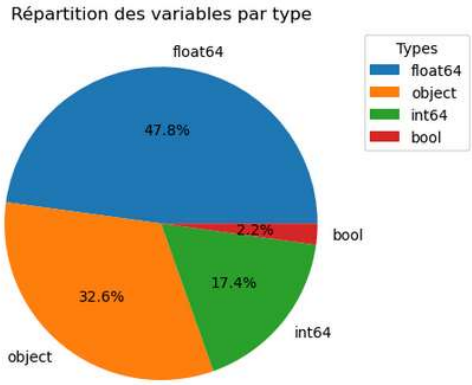
Mars 2023

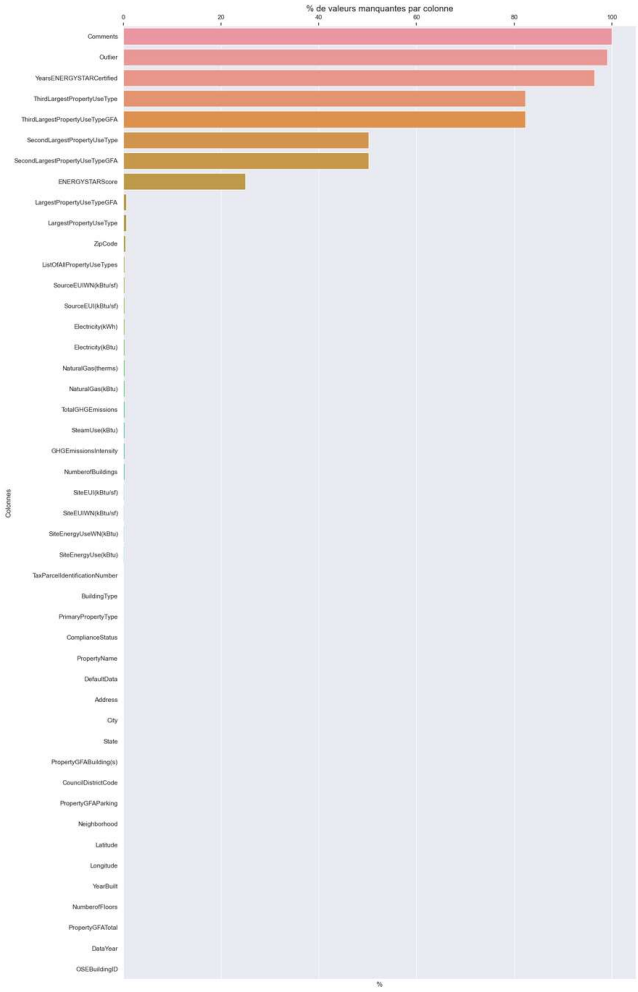
ANALYSE EXPLORATOIRE

Seattle

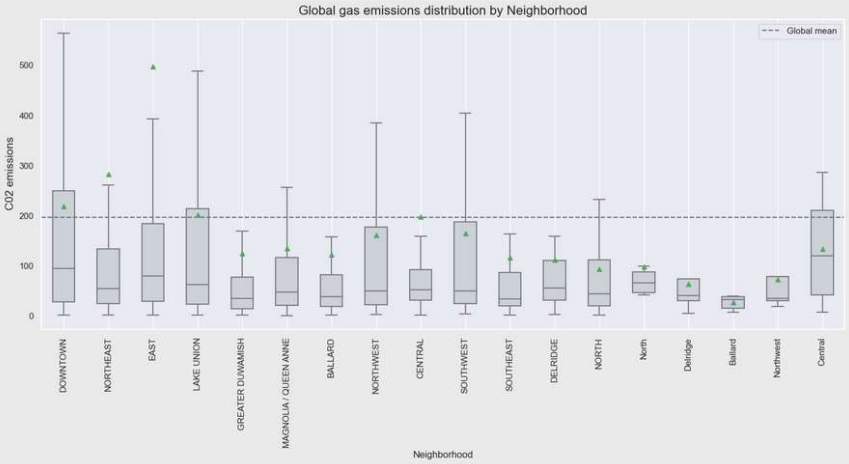
Description du Dataset :

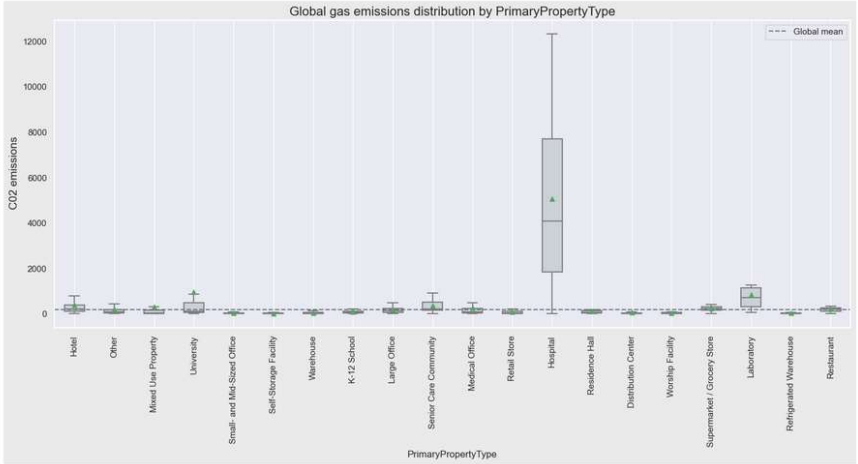
- Données sur l'année 2016
- 3'376 lignes, 46 colonnes
- 13% de valeurs nulles
- Pas de doublons sur les variables
- Taille du fichier : 1,2 Mb





ANALYSE EXPLORATOIRE

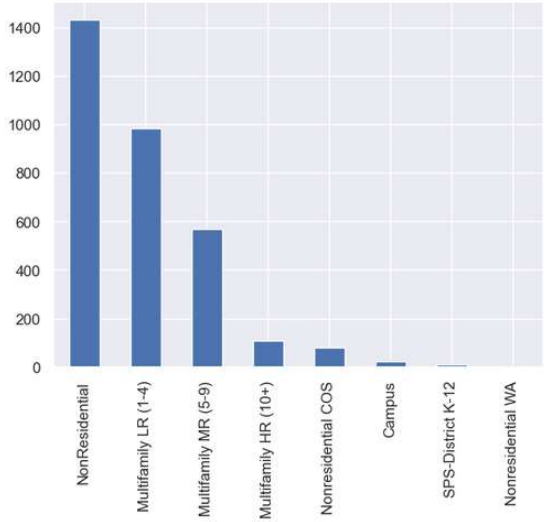


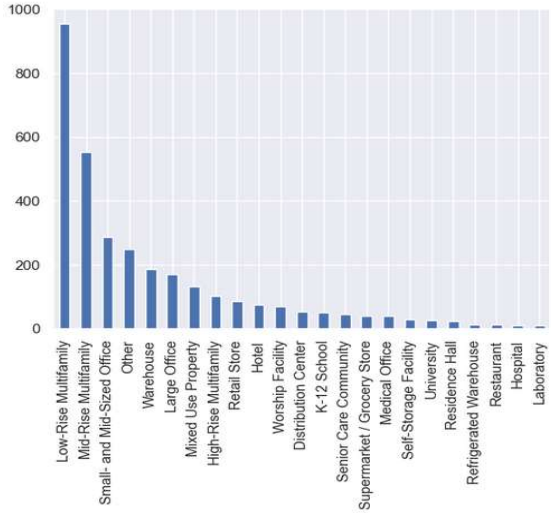


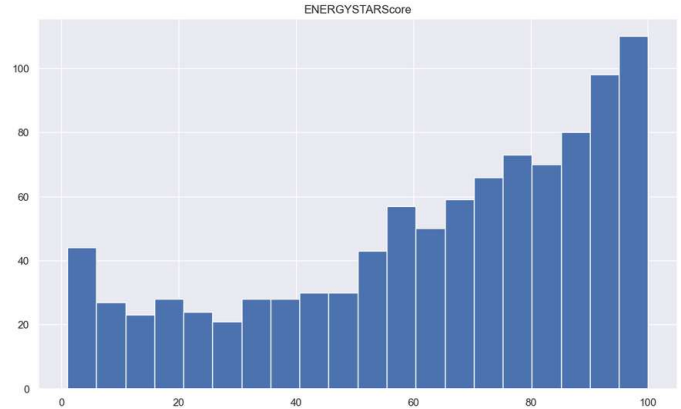
- Certains quartiers présentent des niveaux d'émissions de gaz supérieurs (Downtown, Lake Union, Central).

- Types de bâtiments émettant le plus de gaz : hôpitaux, universités et laboratoires.

ANALYSE EXPLORATOIRE







Analyse des types de buildings :

Multifamily = habitation rédisentielle
Low-rise (LR) / Mid-rise (MR) / High-rise (HR) = taille de l'habitation
Il y a donc trois types d'habitations résidentielles, classées selon leur taille.

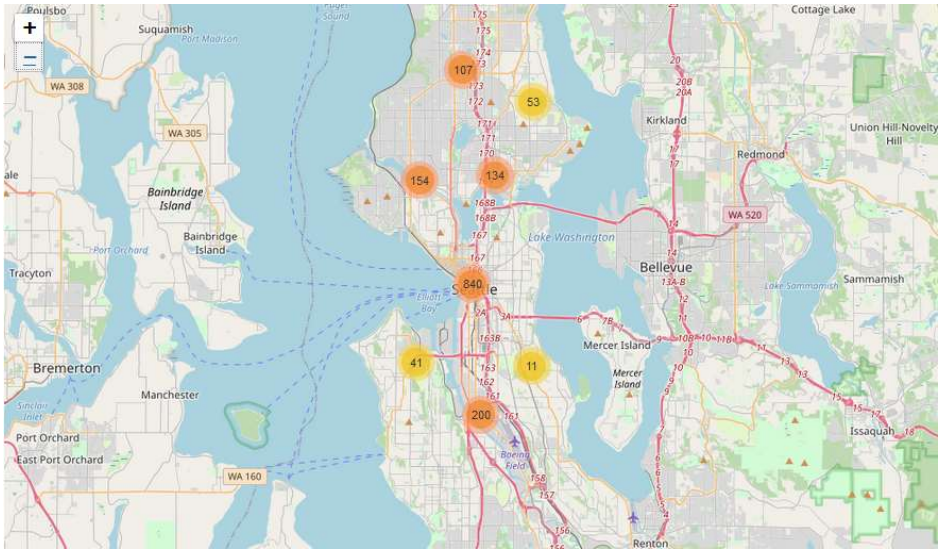
Cleaning du Dataset :

- Suppression des bâtiments résidentiels ('Mulitfamily')
- Suppression des individus 'Outliers'
- Suppression des individus non 'Compliant'
- Suppression des lignes avec plus de 20% de valeurs nulles
- Suppression des doublons sur les individus.
- Nous ne conservons que les variables dans une même unité de mesure (kBtu).

Seattle

Dataset cleané :

- 1’523 individus (bâtiments) restants ;
- 30 variables.



FEATURE ENGINEERING

Quelques définitions utiles :
- GHG: greenhouse gas emissions, correspond aux émissions des gazs à effets de serre.
- OSE: Seattle Office of Sustainability and Environment.
- EUI: Energy Use Intensity.
- kBtu: kilo-British thermal unit, 1 kWh = 3.412 kBtu.
- sf: square feet, 1m² = 10,7639sf.
- WN: weather-normalized, normalisé vis à vis des conditions climatiques.
- GFA: Gross floor area, Surface de plancher brute - La surface de plancher couverte (par un toit, même sans mur) totale contenue dans le bâtiment.
- therm: mesure énergétique 1thm =100000Btu.

Création de variables :

- Âge des bâtiments (2016 – 'YearBuilt')
- Proportion des surfaces des bâtiments par types d'utilisation

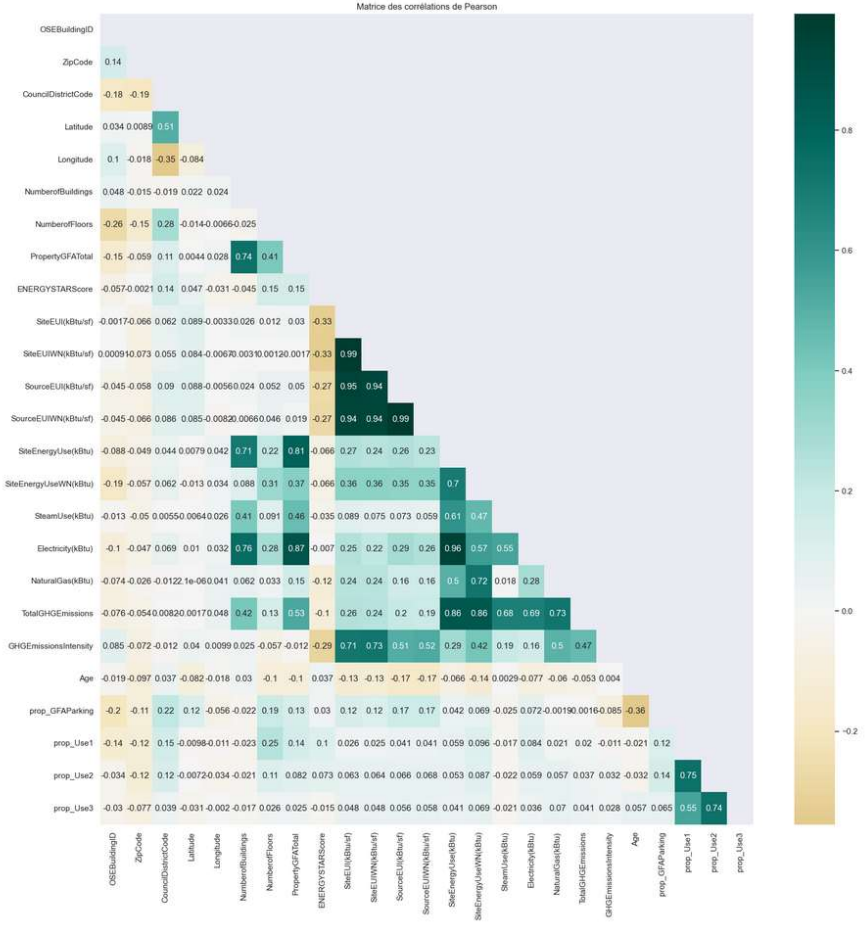
Identification des targets :

- Consommation : SiteEnergyUse(kBtu)
- Emissions : TotalGHGEmissions

Suppression des variables inutiles :

- PropertyName
- Address
- City
- State
- Comments
- DefaultData
- Outlier
- ComplianceStatus

FEATURE ENGINEERING



Analyse des corrélations (Matrice de Pearson) :
Certaines variables présentent de fortes corrélations.

⇒ Suppression des variables relatives aux relevés, afin de ne se baser que sur les paramètres propres aux bâtiments (emplacement, surface, utilisation, nombre d'étages, etc.).

Ex. : SiteEnergyUseWN(kBtu)
Electricity(kBtu)
NaturalGas(kBtu)

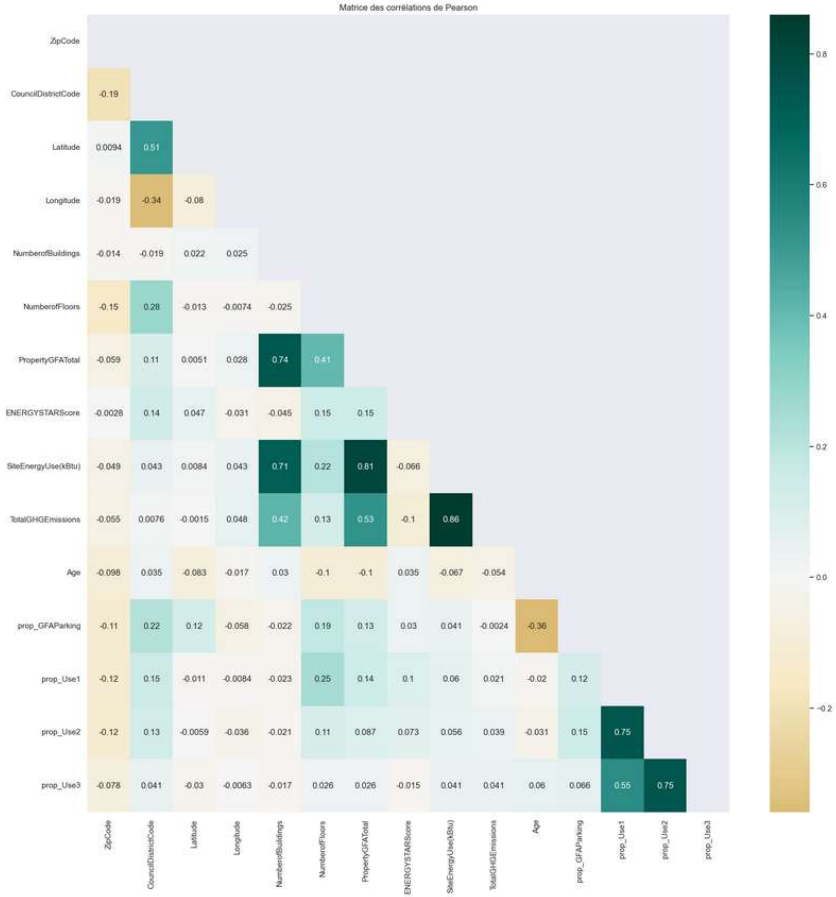
FEATURE ENGINEERING

Seattle

Analyse des corrélations post cleaning :

Nous n'avons plus de variables liées aux relevés effectués sur les bâtiments.

La modélisation peut donc être effectuée sur cet ensemble de features.



FEATURE ENGINEERING

Afin de procéder à la modélisation, une transformation et/ou un encodage sont nécessaires sur les données quantitatives et qualitatives (Preprocessing).

Variables quantitatives		
MinMaxScaler	**RobustScaler**	**StandardScaler**
Echelle de 0 à 1 Sensible aux outliers	Normalise en fonction des quartiles 1 et 3 Peu sensible aux outliers	Normalisation, moyenne de 0 et écart-type de 1

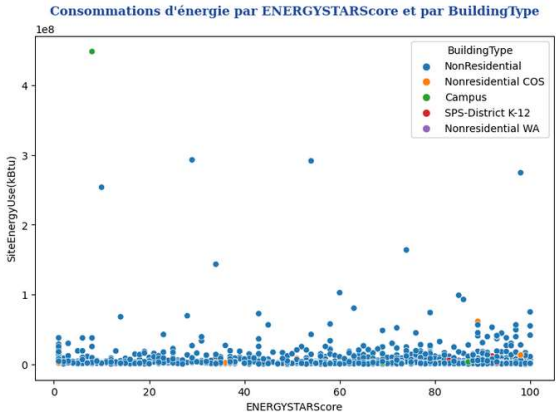
Variables qualitatives		
	Encodage ordinal	**Encodage OneHot**
y (target)	LabelEncoder	LabelBinarizer
X (variables)	OrdinalEncoder	OneHotEncoder
		TagetEncoder

Nous retiendrons :

- Variables quantitatives : MinMaxScaler
- Variables qualitatives : OneHotEncoder

PREDICTION DE LA CONSOMMATION ENERGETIQUE

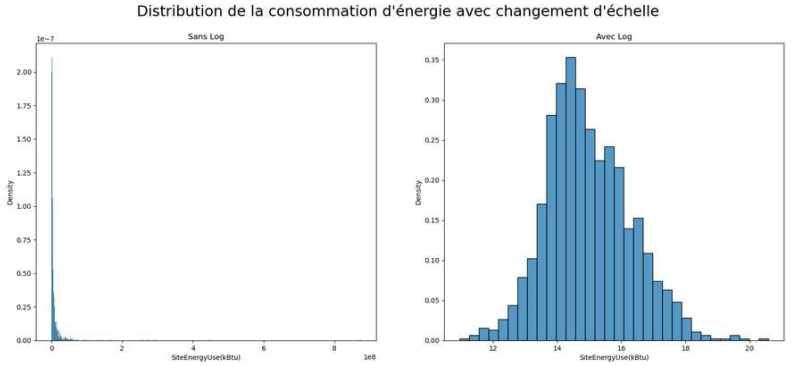
Target : SiteEnergyUse(kBtu)



La target n'est pas corrélé à la variable ENERGYSTARScore

Nous modéliserons sans cette variable, puis en l'incluant afin de sélectionner le modèle le plus pertinent.

Nous passons la target y au Log pour une meilleure représentation.



PREDICTION DE LA CONSOMMATION ENERGETIQUE

MODELISATION

Pipeline : MinMaxScaler + OneHotEncoder

SANS ENERGYSTARScore

	MAE	RMSE	Med Abs error	R2	Chrono (s)
LinearRegression	7.226397	15.529769	4.272332	-125.633669	0.10
Ridge	0.736996	0.976979	0.543991	0.498825	0.01
Lasso	1.094897	1.380175	0.878949	-0.000202	0.01
ElasticNet	1.094897	1.380175	0.878949	-0.000202	0.01
SVR	0.724230	0.965373	0.553423	0.510661	0.07
RandomForest	0.524378	0.711974	0.387452	0.733837	1.75
XGBoost	0.530723	0.731116	0.381481	0.719333	0.09

AVEC ENERGYSTARScore

	MAE	RMSE	Med Abs error	R2	Chrono (s)
LinearRegression	10.361749	24.266686	5.007654	-308.200494	0.10
Ridge	0.703110	0.942276	0.532376	0.533797	0.01
Lasso	1.094897	1.380175	0.878949	-0.000202	0.01
ElasticNet	1.094897	1.380175	0.878949	-0.000202	0.01
SVR	0.689995	0.936942	0.511880	0.539060	0.07
RandomForest	0.471391	0.661777	0.319227	0.770045	1.81
XGBoost	0.473190	0.671356	0.334653	0.763340	0.09

CROSSVALIDATION (GridSearchCV)

Meilleur modélisation retenue :

- XGBoostRegressor
- Avec ENERGYSTARScore
- MinMaxScaler + OneHotEncoder

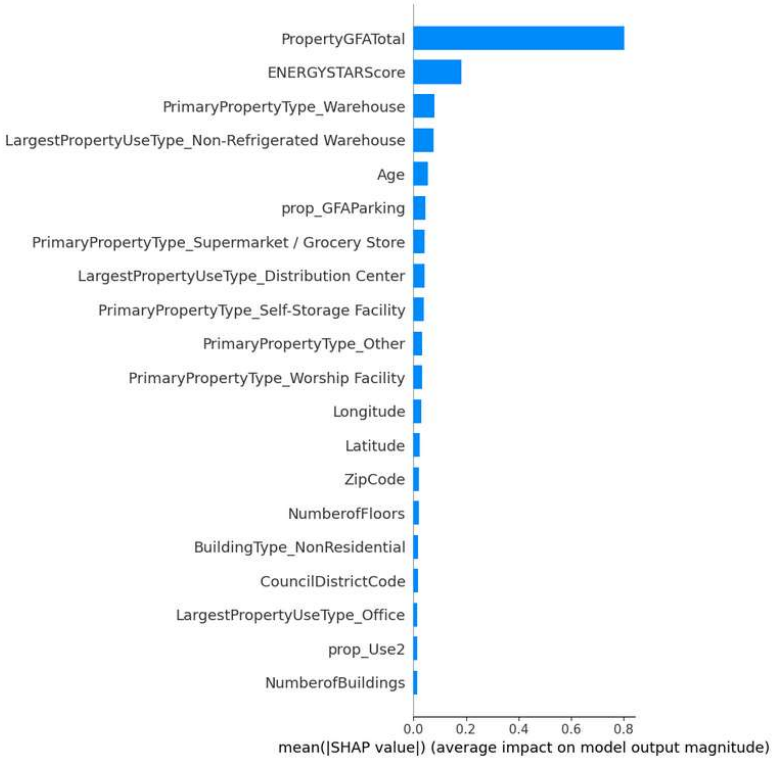
Best score : 0,752
RMSE : 0,461
MAE : 0,663

Hyperparamètres retenus :

Learning_rate = **0,1** parmi [0,1; 0,3; 0,5]
n_estimators = **100** parmi [10; 100; 500]
enable_categorical = False parmi [True; False]

PREDICTION DE LA CONSOMMATION ENERGETIQUE

Feature importance

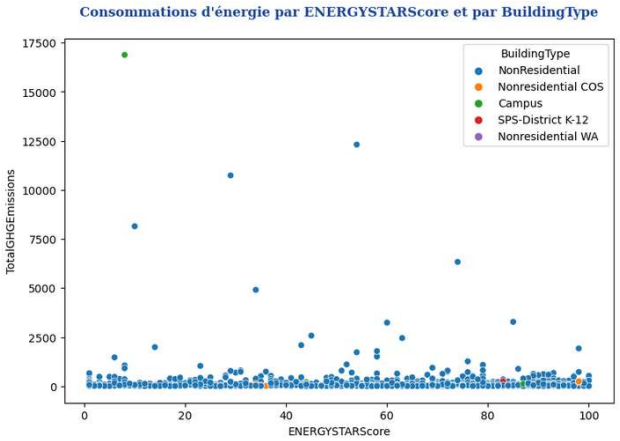


Top5 des variables ayant le plus d'influence sur le modèle :

- Surface Totale
- ENERGYSTARScore
- Type principal = Warehouse
- Principale utilsiation = Warehouse non réfrigérée
- Âge du bâtiment

PREDICTION DES EMISSIONS DE GAZ

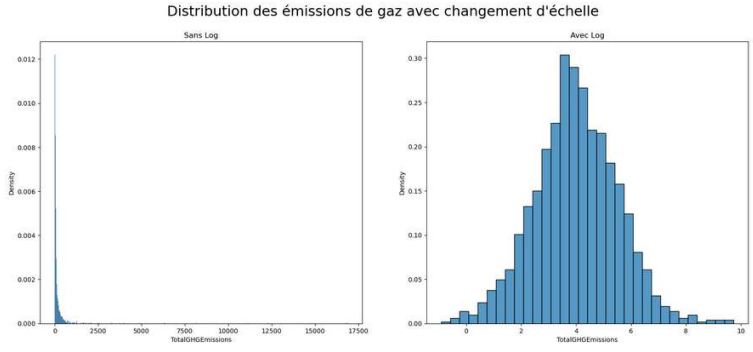
Target : TotalGHGEmissions



La target n'est pas corrélé à la variable ENERGYSTARScore

Nous modéliserons sans cette variable, puis en l'incluant afin de sélectionner le modèle le plus pertinent.

Nous passons la target y au Log pour une meilleure représentation.



PREDICTION DES EMISSIONS DE GAZ

MODELISATION

Pipeline : MinMaxScaler + OneHotEncoder

SANS ENERGYSTARScore

	MAE	RMSE	Med Abs error	R2	Chrono (s)
LinearRegression	9.411928	19.291580	4.935762	-145.104548	0.10
Ridge	0.969038	1.284966	0.730828	0.351796	0.01
Lasso	1.236434	1.596360	1.058407	-0.000438	0.01
ElasticNet	1.236434	1.596360	1.058407	-0.000438	0.01
SVR	0.960642	1.270402	0.680248	0.366407	0.07
RandomForest	0.846134	1.092708	0.683847	0.531255	1.71
XGBoost	0.841282	1.107260	0.652392	0.518687	0.09

AVEC ENERGYSTARScore

	MAE	RMSE	Med Abs error	R2	Chrono (s)
LinearRegression	9.803755	20.463313	5.154687	-163.391748	0.10
Ridge	0.945356	1.254068	0.668115	0.382595	0.01
Lasso	1.236434	1.596360	1.058407	-0.000438	0.01
ElasticNet	1.236434	1.596360	1.058407	-0.000438	0.01
SVR	0.942501	1.249959	0.664262	0.386634	0.07
RandomForest	0.826181	1.055066	0.664458	0.562994	1.79
XGBoost	0.822015	1.068969	0.621908	0.551401	0.09

PREDICTION DES EMISSIONS DE GAZ

CROSSVALIDATION (GridSearchCV)

Meilleur modélisation retenue :

- XGBoostRegressor
- Avec ENERGYSTARScore
- MinMaxScaler + OneHotEncoder

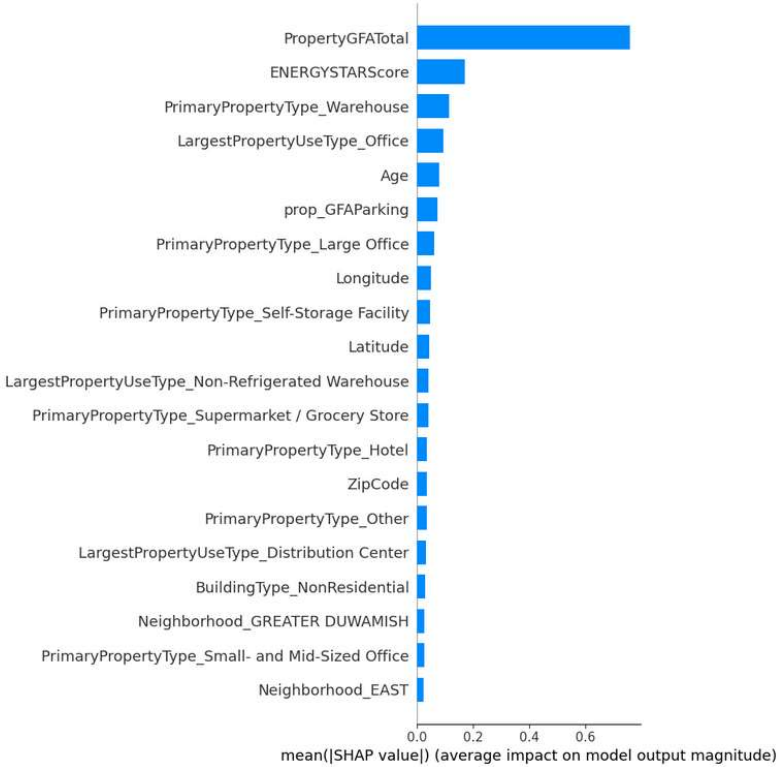
Best score : 0,532
RMSE : 0,83
MAE : 0,663

Hyperparamètres retenus :

Learning_rate = **0,1** parmi [0,1; 0,3; 0,5]
n_estimators = **100** parmi [10; 100; 500]
enable_categorical = False parmi [True; False]

PREDICTION DES EMISSIONS DE GAZ

Feature importance



Top5 des variables ayant le plus d'influence sur le modèle :

- Surface Totale
- ENERGYSTARScore
- Type principal = Warehouse
- Principale utilisation = Bureaux
- Âge du bâtiment

BILAN DE LA MODELISATION

Seattle

CONSOMMATION ENERGETIQUE :

Avec ENERGYSTARScore
Encodage retenu : MinMaxScaler + OneHotEncoder
Modèle retenu : XGBoostRegressor
Hyper paramètres :

- Learning rate = 0,1
- n estimators = 100
- enable_categorical = False

BEST SCORE : 0,752

XGBoost

EMISSIONS DE GAZ :

Avec ENERGYSTARScore
Encodage retenu : MinMaxScaler + OneHotEncoder
Modèle retenu : XGBoostRegressor
Hyper paramètres :

- Learning rate = 0,1
- n estimators = 100
- enable_categorical = False

BEST SCORE : 0,532

XGBoost

Jay Corentin

CONCLUSION

La variable ENERGYSTARScore permet une modélisation plus efficace, que ce soit pour les consommations d'énergie ou les émissions de gaz. Cependant, le gain apporté par cette dernière n'est pas déterminant, et une modélisation pourrait être envisagée également sans cette variable.

La modélisation apportée ici présente une meilleure efficacité concernant la consommation d'énergie (score de 0,752 vs 0,532 pour les émissions de gaz), et nous permet d'identifier les paramètres les plus influents pour les bâtiments non résidentiels.

Seattle

PACKAGES PYTHON :

pandas

seaborn

scikit learn

matplotlib

NumPy

Folium

plotly