Seattle

ANTICIPEZ LES BESOINS EN CONSOMMATION DE BÂTIMENTS

Alexandre Delaguillaumie

PARTIE I

Introduction

Introduction summary

Seattle veut être une ville neutre
En emission de carbone en 2050

En réduisant la consommation et l'émission de gaz
des bâtiments non destinés à l'habitation

Les relevés des bâtiments sont coûteux à obtenir
On cherche donc à s'en passer en les prédisant

Introduction problem

Réaliser un **algorithme** de **prédiction** des **émissions** de gaz à effet de serre et de la **consommation** d'énergie des **bâtiments non destinés à l'habitation** qui n'ont pas encore été mesurés

Introduction few things to know

Attention à la fuite de données
car les variables energetiques ne sont pas connues.

Un 1er relevé de référence est effectué
Pour tous les bâtiments.

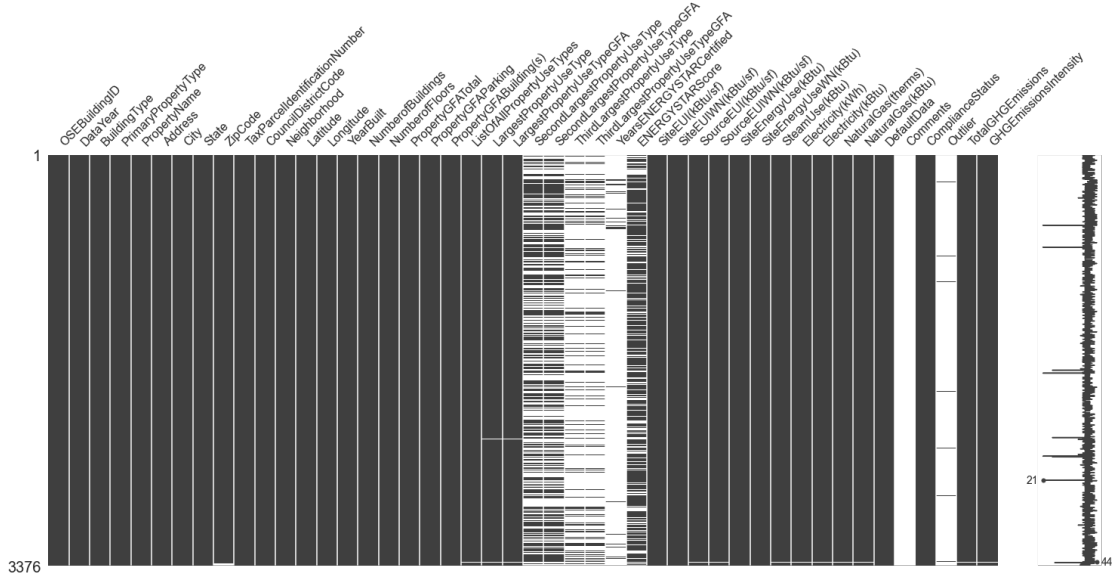
On cherche à évaluer l'intérêt de l'EnergyStarScore
Car il est fastidieux à calculer actuellement.

PARTIE II

Exploration

Exploration overview

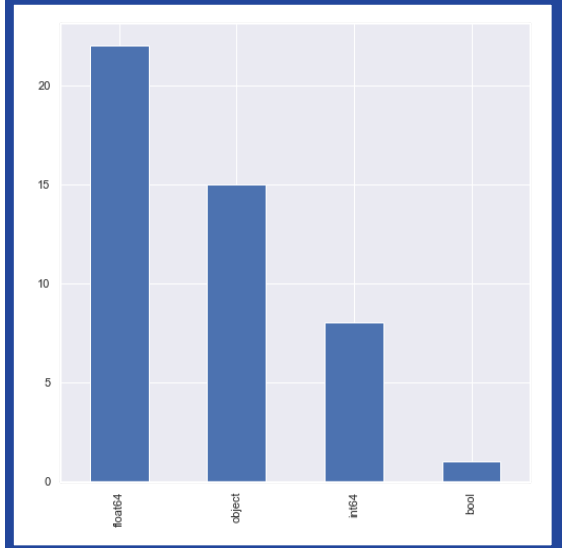
FILL RATE du dataset brut



3376 BÂTIMENTS

Dont 1583 non résidentiels

RÉPARTITION DES VARIABLES par type



45 variables

Dont 20 d'intérêt

Exploration targets & features

Features

Nombre d'étages
Number of Floors

Nombre de bâtiments
Number of Buildings

Superficie brute
PropertyGFABuilding
SecondLPUT GFA
ThirdLPUT GFA

Emplacement
Latitude / Longitude

Ancienneté
YearBuilt

Quartier
Neighborhood

Type de bâtiment
LargestPropertyUseType
SecondLargestPropertyUT
ThirdLargestPropertyUT

Superficie
LargestPropertyUseGFA

Energies utilisées
NaturalGas,
SteamUse,
Electricity

Target 1

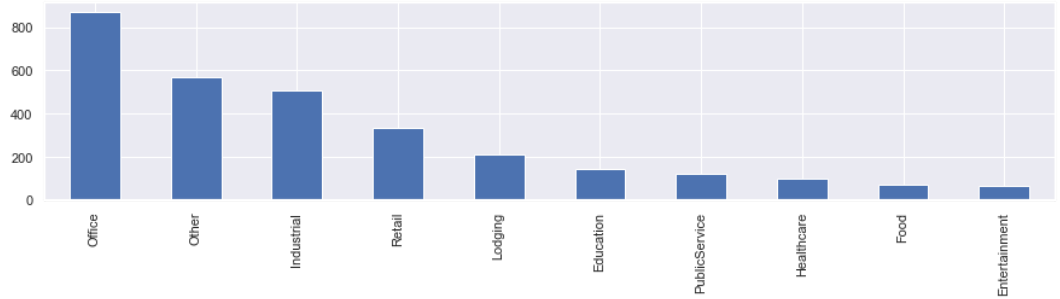
Consommation d'énergie du site
SiteEnergyUse(kBtu)

Target 2

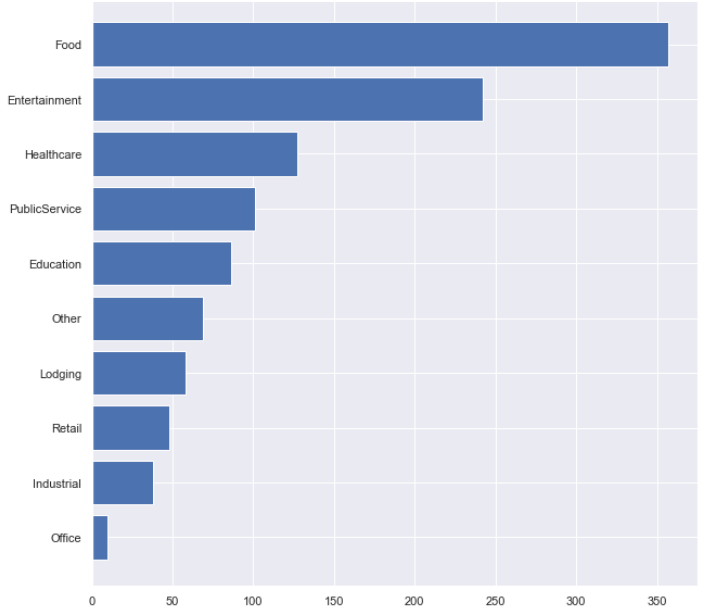
Emissions de gaz à effet de serre
TotalGHGEmissions

Exploration building caracteristics

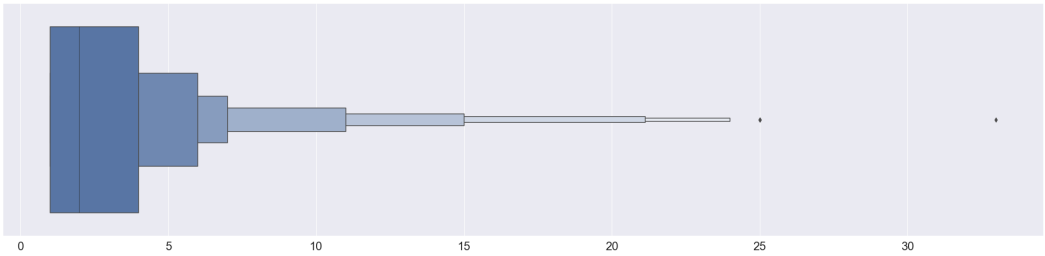
UTILISATION TOUT CONFONDU



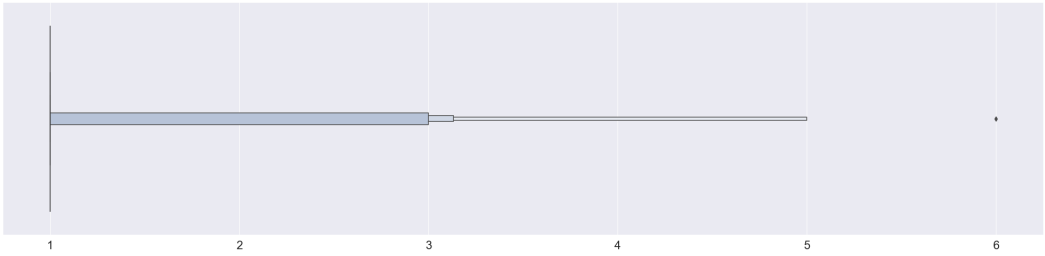
UTILISATION PRINCIPALE



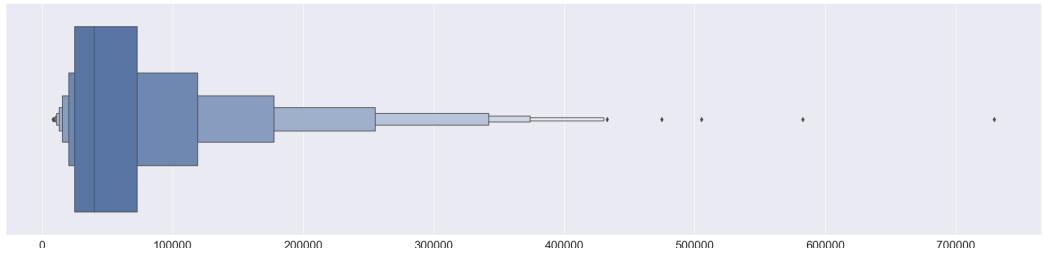
NOMBRE D'ÉTAGES



NOMBRE DE BÂTIMENTS

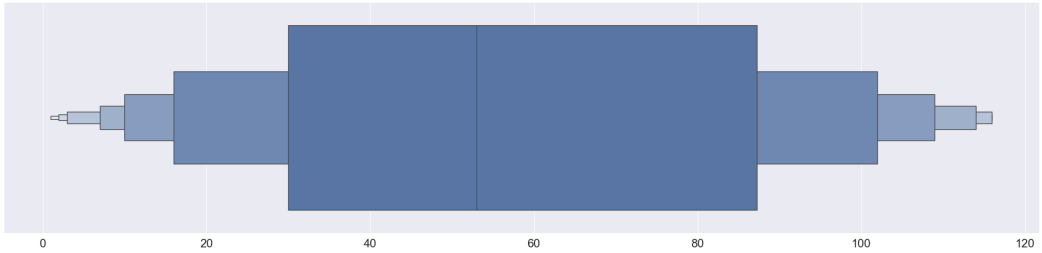


SUPERFICIE DE L'ACTIVITÉ PRINCIPALE



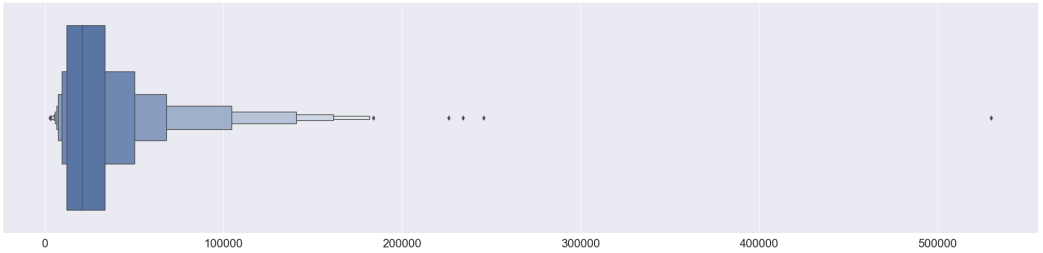
Exploration feature engineering

ÂGE DU BÂTIMENT



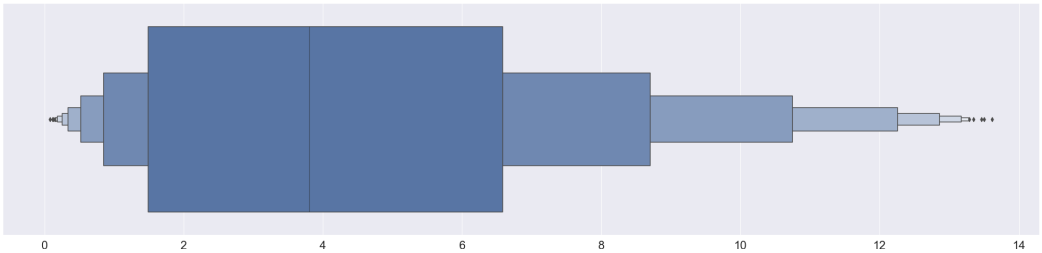

```
# Âge de la construction
data['Age'] = 2016 - data['YearBuilt']
```

SUPERFICIE MOYENNE



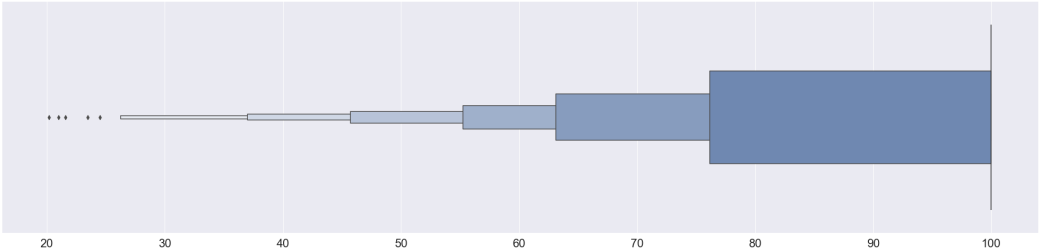

```
# Surface par étage par bâtiment
data['mean_GFA'] = data['PropertyGFATotal']/(data['NumberofFloors']*data['NumberofBuildings']).round(2)
```

DISTANCE DU CENTRE-VILLE (EN KM)



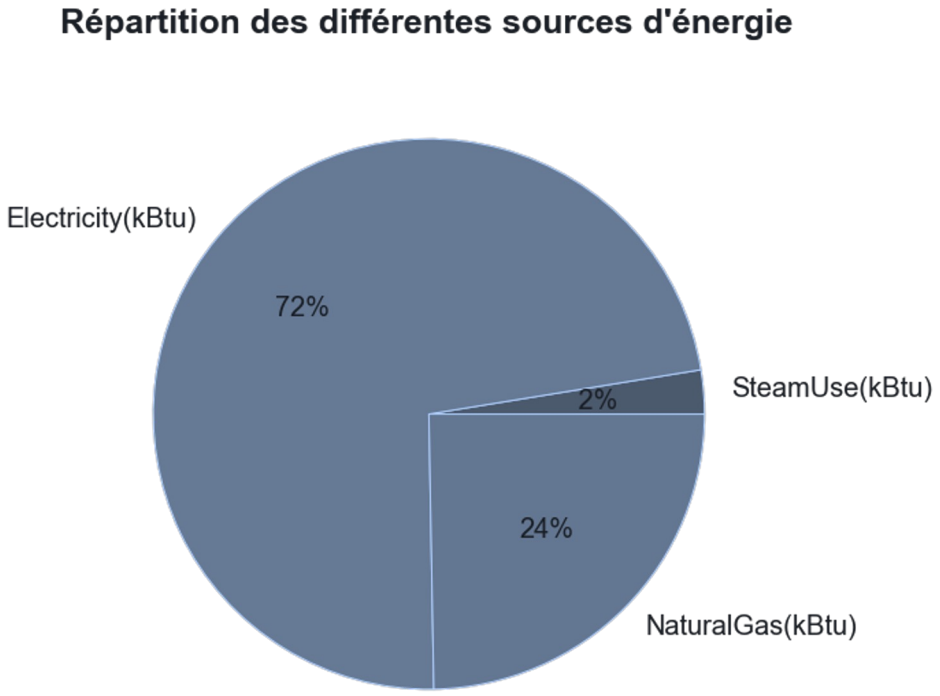

```
# Distance par rapport au centre
downtown = (47.608056,-122.336111)
data['CBD_distance_km'] = data.apply(lambda x: haversine((x['Latitude'], x['Longitude']),downtown), axis=1)
```

PROPORTION DE BATIMENTS

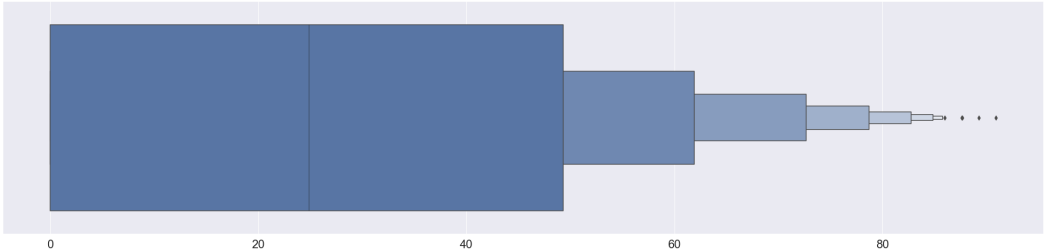



```
# Proportion de surface non-parking
data['GFABuildingRate'] = (round((data['PropertyGFABuilding(s)'].fillna(0)
                                  /data['PropertyGFATotal'].fillna(0)),5))*100
```

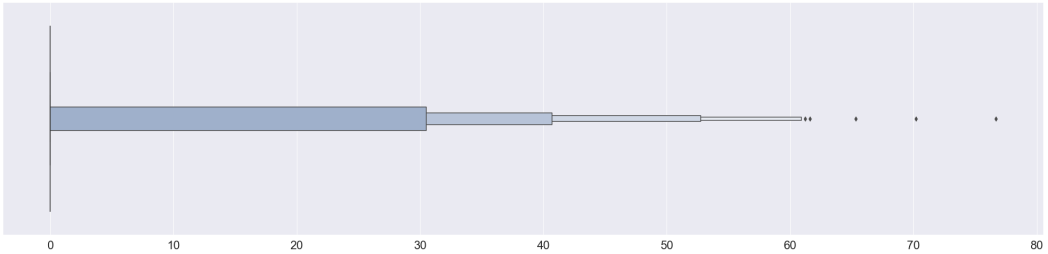
Exploration energy



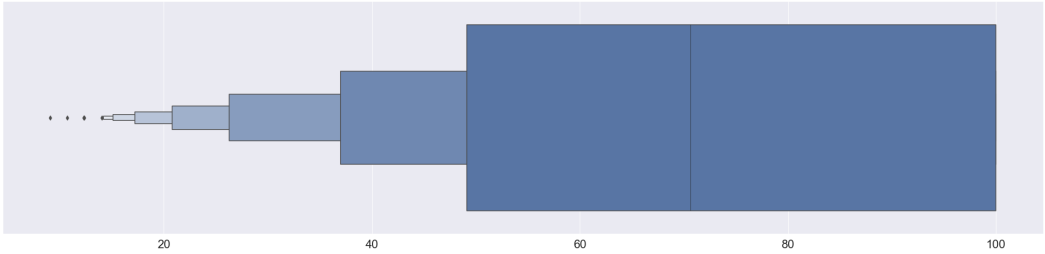
GAZ NATUREL



VAPEUR



ELECTRICITÉ

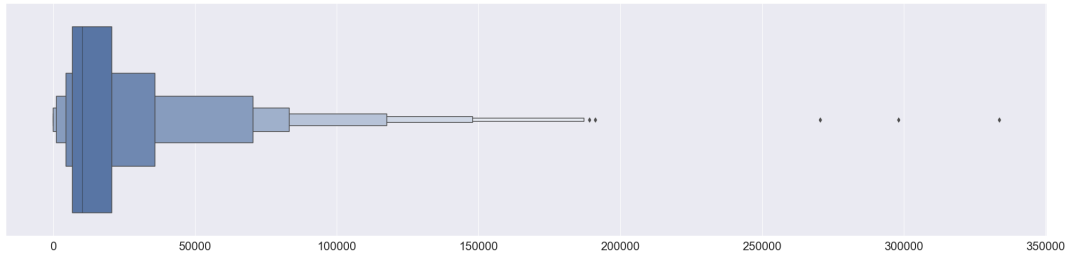


À savoir :

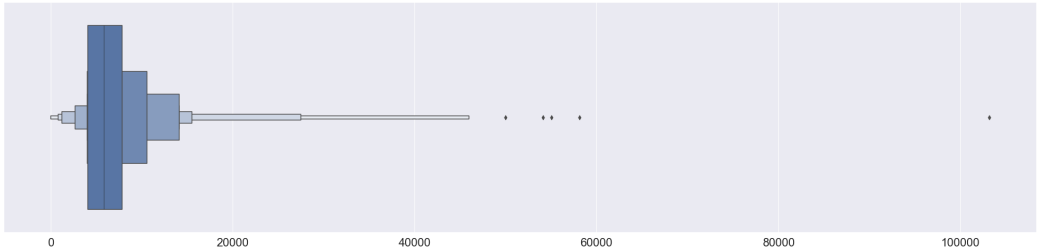
kBtu (kilo-British thermal unit)		kWh (kilo-Watt-hour)
3.142	=	1

Exploration KNN imputer

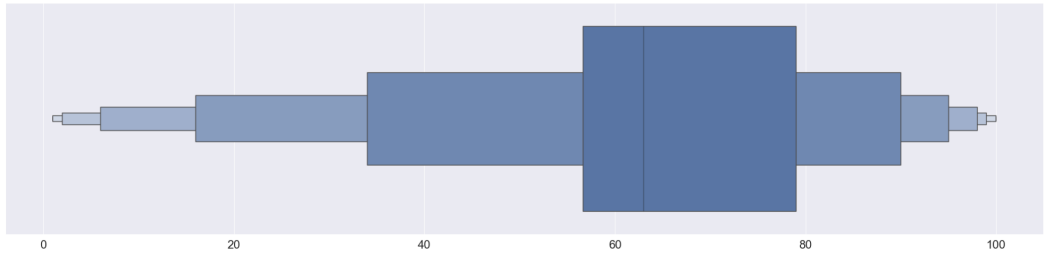
SECOND LARGEST PROPERTY USE TYPE



THIRD LARGEST PROPERTY USE TYPE GFA



ENERGY STAR SCORE



`sklearn.neighbors`.KNeighborsRegressor

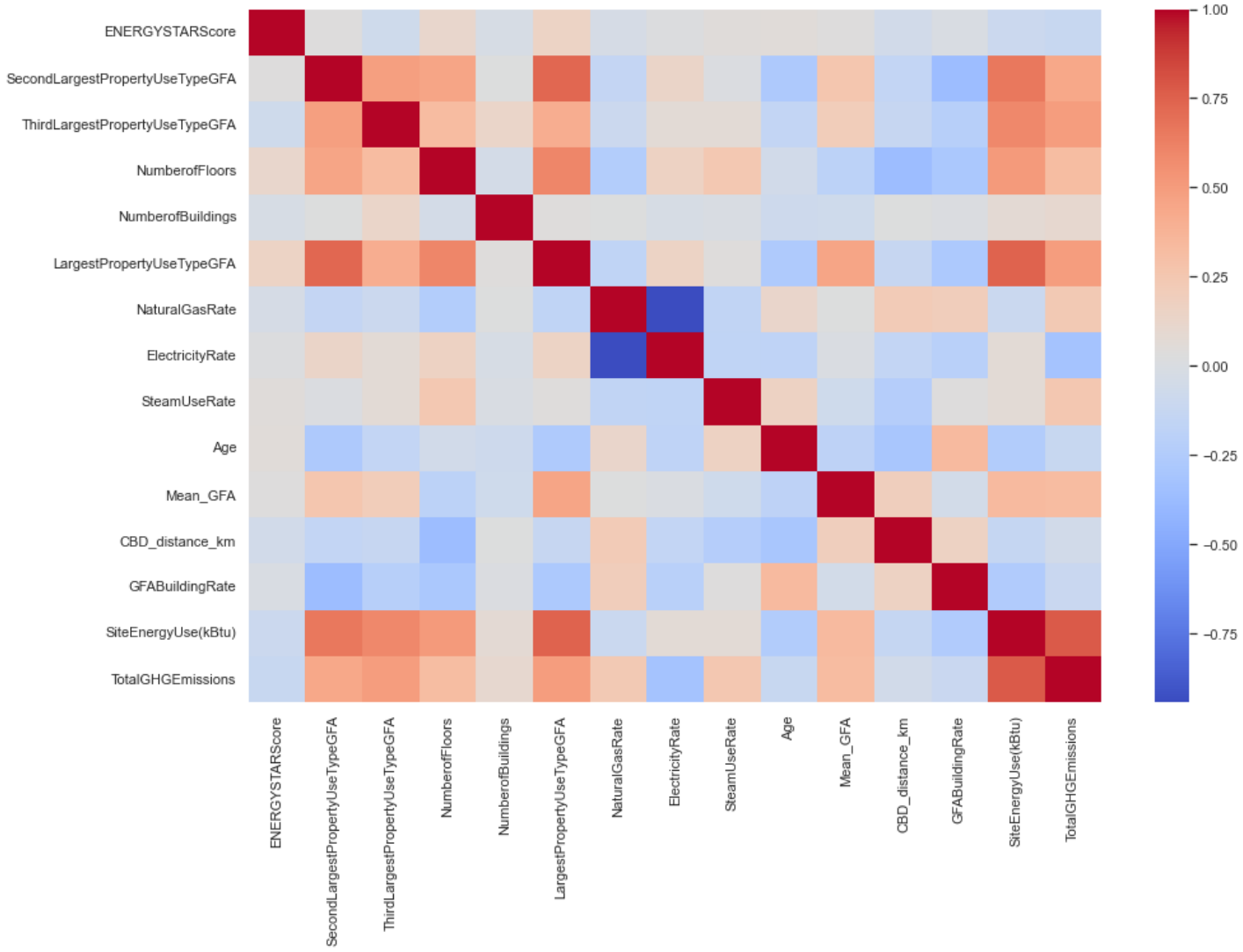
`sklearn.impute`.KNNImputer

GridSearchCV

n_neighbors :

SLPUT GFA	41
TLPUT GFA	54
ESS	138

Exploration correlation



PARTIE III

Preprocessing

Preprocessing GetDummies

1 **Encodage sur les features catégorielles**
pour effectuer des opérations de Machine Learning

2 **Plus simple**
que OneHotEncoder

3 **Effectué sur 4 variables catégorielles**
Neighborhood, SecondLargestPropertyUseType,
ThirdLargestPropertyUseType et LargestPropertyUseType

Preprocessing GetDummies

1 **Encodage sur les features catégorielles**
pour effectuer des opérations de Machine Learning

2 **Plus simple**
que OneHotEncoder

3 **Effectué sur 4 variables catégorielles**
Neighborhood, SecondLargestPropertyUseType,
ThirdLargestPropertyUseType et LargestPropertyUseType

Preprocessing GetDummies

1 **Encodage sur les features catégorielles**
pour effectuer des opérations de Machine Learning

2 **Plus simple**
que OneHotEncoder

3 **Effectué sur 4 variables catégorielles**
Neighborhood, SecondLargestPropertyUseType,
ThirdLargestPropertyUseType et LargestPropertyUseType

Preprocessing MinMaxScaler

Raison 1

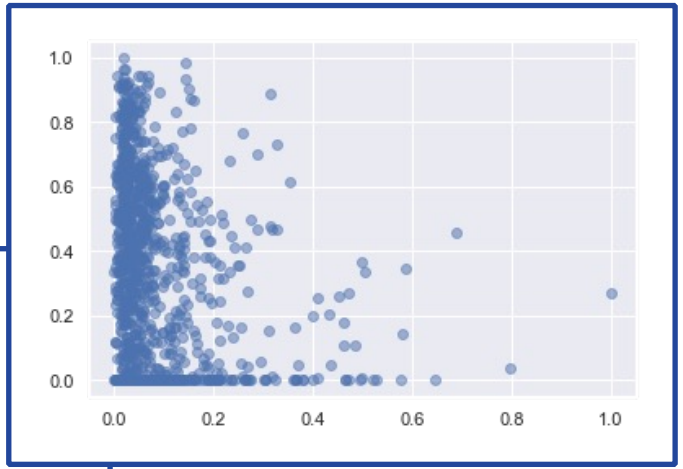
Différentes échelles entre les variables

Une variable peut prendre le dessus lors de la minimisation de la fonction coût

Raison 2

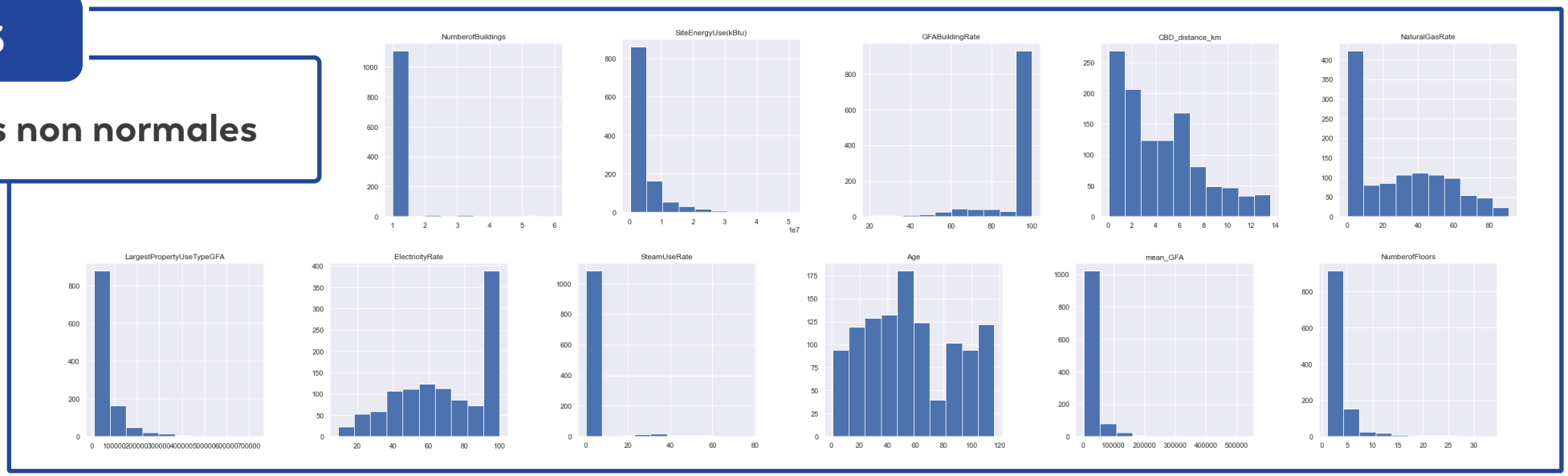
Pas d ’écrasement de données

Variance suffisante lors de la transformation (pas d’outliers)



Raison 3

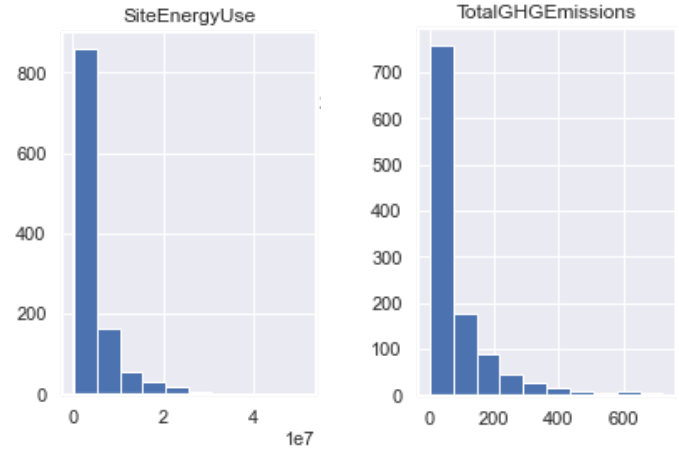
Données non normales



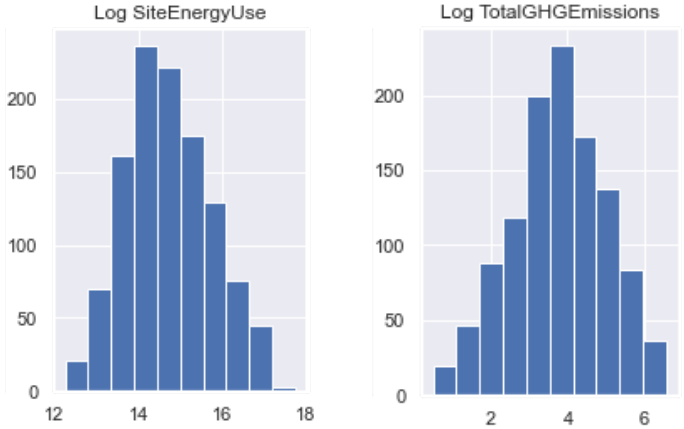
Preprocessing log trick

```
FunctionTransformer(np.log,inverse_func=np.exp,check_inverse=True)
```

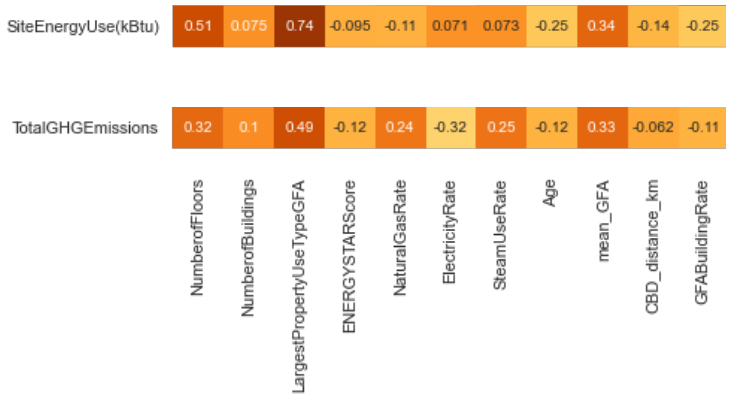
Original



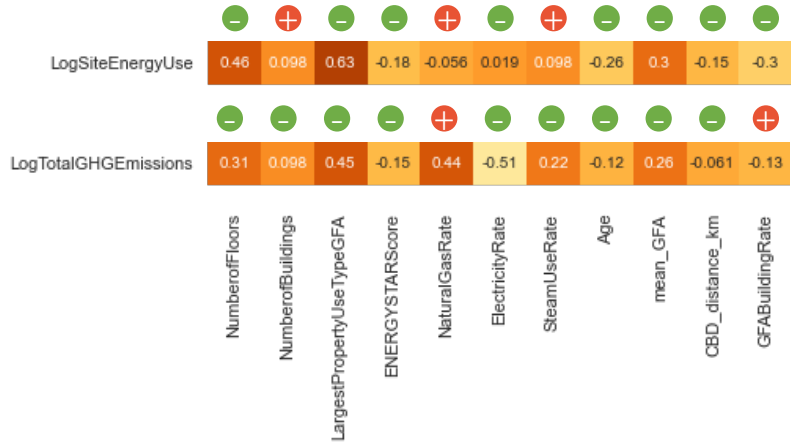
Transformed



Corrélation



Corrélation ajustée




PARTIE IV
Algorithms

Algorithms naive approaches

TARGETS

- GHGEmmissions
- SiteEnergyUse

Tenter de prédire avec des valeurs aléatoires :

```
np.random.randint(np.min(y_log), np.max(y_log), y_test_rand.shape)
```

	GHGEmmissions	SiteEnergyUse
RMSE	368.91	25 731 761
r2	-12	-28

Prédit systématiquement la moyenne du train set:

`sklearn.dummy.DummyClassifier`

	GHGEmmissions	SiteEnergyUse
RMSE	128.28	6 201 804
r2	-0.5730	-0.69

Algorithms hyperparams tuning

TARGETS

- GHGEmmissions
- SiteEnergyUse

sklearn.linear_model.LinearRegression

No hyperparams

sklearn.linear_model.Ridge

GridSearchCV

Alpha : 0.15 | 1.0

Algorithms hyperparams tuning

TARGETS

- GHGEmmissions
- SiteEnergyUse

sklearn.linear_model.Lasso

GridSearchCV

Alpha :	0.001363	0.00534

sklearn.linear_model.ElasticNet

GridSearchCV

Alpha :	0.00129	0.00207
l1 ratio :	1	0.3591

Algorithms hyperparams tuning

TARGETS

- GHGEmmissions
- SiteEnergyUse

sklearn.neighbors.KNeighborsRegressor

GridSearchCV

n_neighbors :	6	7
metrics :	manhattan	manhattan

sklearn.ensemble.GradientBoostingRegressor

GridSearchCV

n_estimators :	93	81

Algorithms hyperparams tuning

TARGETS

GHGEmmissions

SiteEnergyUse

sklearn.ensemble.RandomForestRegressor

RandomizedSearchCV

n_estimators :	99	14
min_samples_split :	6	5
min_samples_leaf :	4	2
max_samples :	0.11	0.48
max_depth :	28	29

Algorithms hyperparams tuning

TARGETS

- GHGEmmissions
- SiteEnergyUse

dmlc XGBoost

RandomizedSearchCV

base_score :	0.3105	0.2789
learning_rate :	0.0578	0.0763
max_depth :	6	6
min_child_weight :	5	2
n_estimators :	145	91

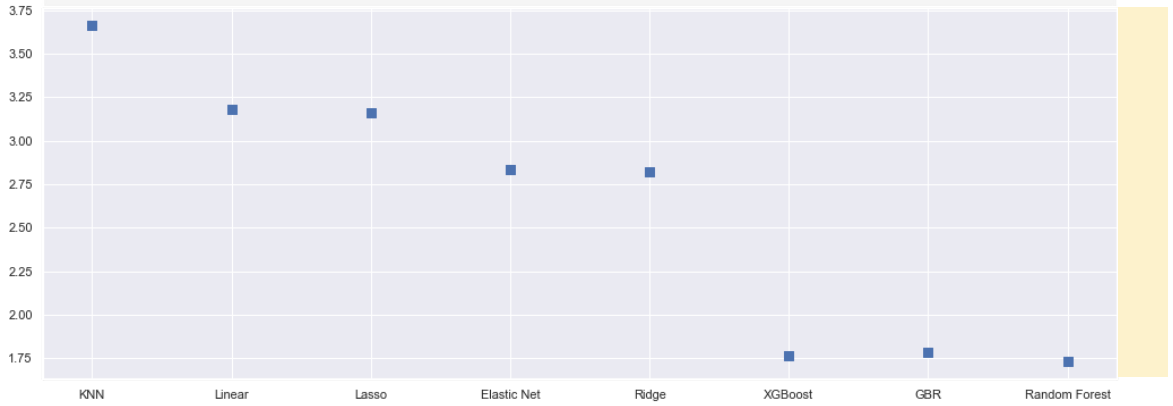
Algorithms RMSE score

TARGETS

GHGEmmissions

SiteEnergyUse

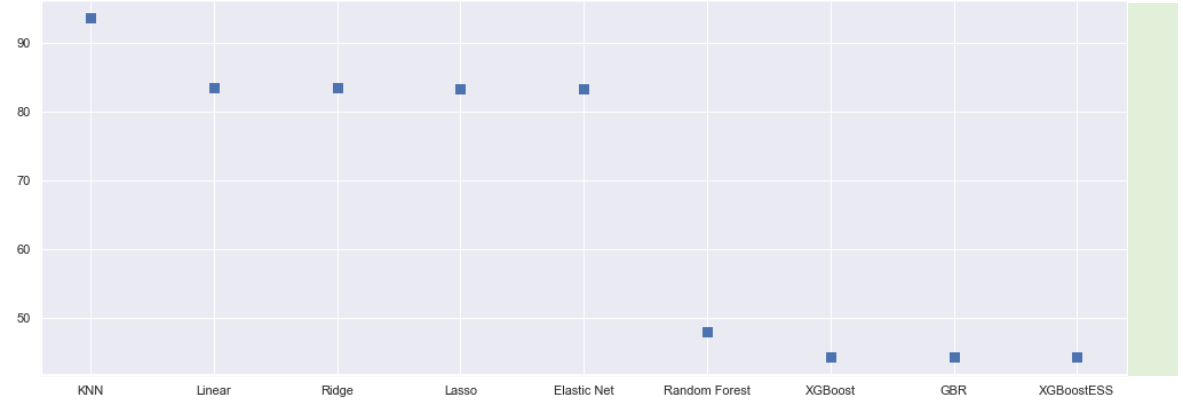
	KNN	Linear	Lasso	Elastic Net	Ridge	XGBoost	GBR	Random Forest
RMSE	3662463.89	3.178658e+06	3160704.30	2831146.13	2823091.78	1763785.96	1785067.75	1729116.46



Prediction
min : 606 895
moy : 3 957 993
max : 32 703 654

Prediction
min : -66.15
moy : 81.18
max : 541.73

	KNN	Linear	Ridge	Lasso	Elastic Net	Random Forest	XGBoost	GBR	XGBoostESS
RMSE	93.590	83.449	83.449	83.354	83.360	48.055	44.490	44.401	44.331



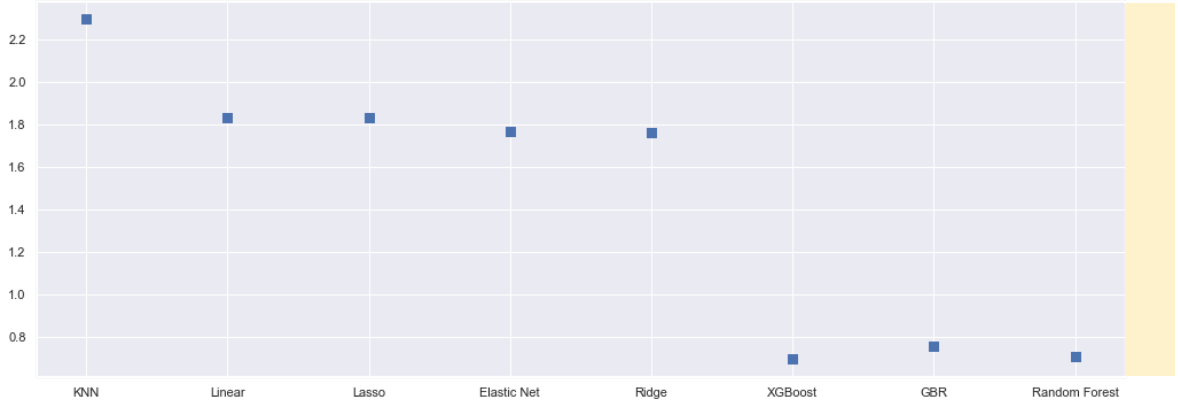
Algorithms MAE score

TARGETS

GHGEmmissions

SiteEnergyUse

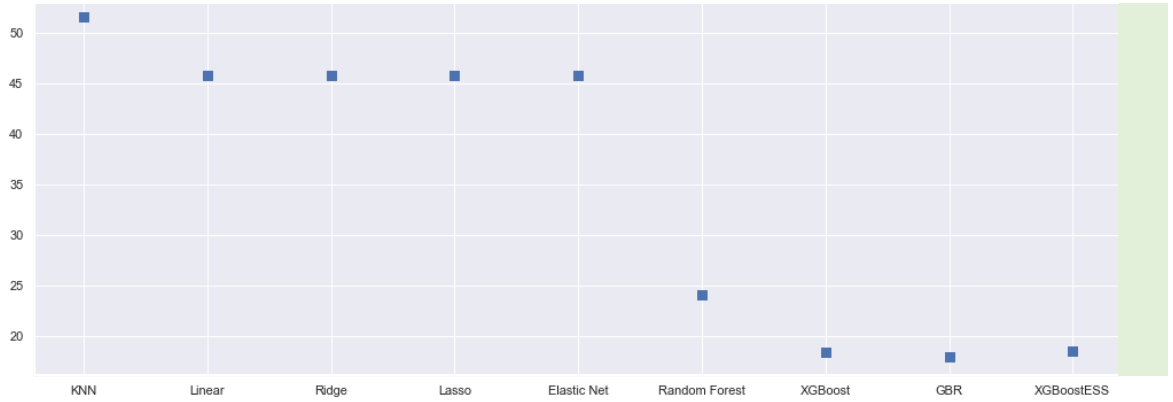
	KNN	Linear	Lasso	Elastic Net	Ridge	XGBoost	GBR	Random Forest
MAE	2295188.39	1.832303e+06	1828928.56	1768381.79	1761744.71	696142.36	759360.13	707462.99



Prediction
min : 606 895
moy : 3 957 993
max : 32 703 654

Prediction
min : -66.15
moy : 81.18
max : 541.73

	KNN	Linear	Ridge	Lasso	Elastic Net	Random Forest	XGBoost	GBR	XGBoostESS
MAE	51.592	45.828	45.828	45.832	45.834	24.036	18.442	17.900	18.576



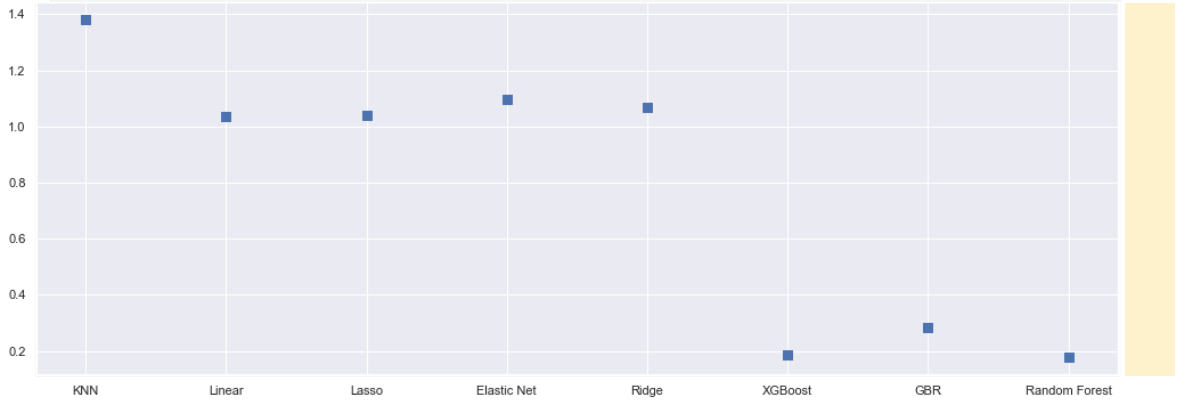
Algorithms MdAE score

TARGETS

GHGEmmissions

SiteEnergyUse

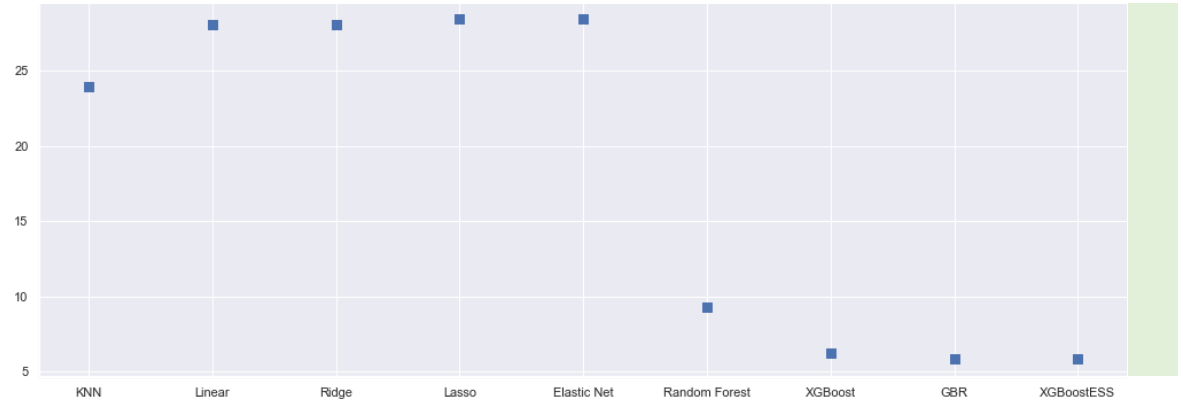
	KNN	Linear	Lasso	Elastic Net	Ridge	XGBoost	GBR	Random Forest
MdAE	1381141.52	1.037427e+06	1040736.29	1097210.23	1069818.64	186013.22	284129.76	176206.26



Prediction
min : 606 895
moy : 3 957 993
max : 32 703 654

Prediction
min : -66.15
moy : 81.18
max : 541.73

	KNN	Linear	Ridge	Lasso	Elastic Net	Random Forest	XGBoost	GBR	XGBoostESS
MdAE	23.957	28.115	28.115	28.469	28.467	9.296	6.224	5.815	5.818



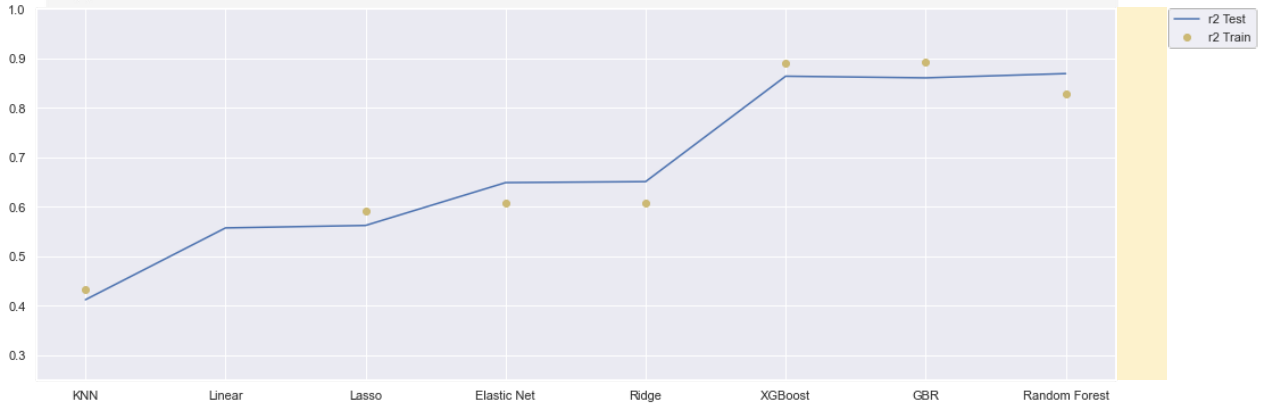
Algorithms r2 score

TARGETS

GHGEmmissions

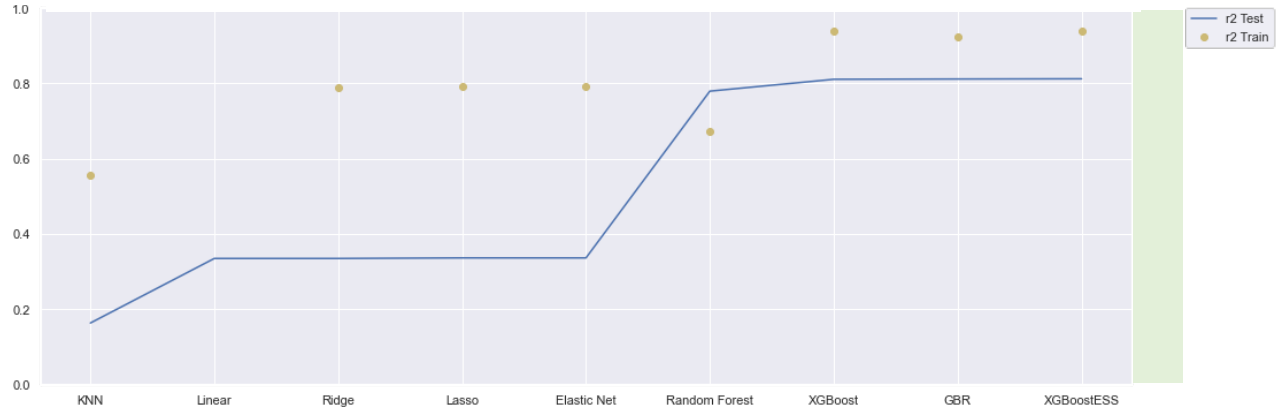
SiteEnergyUse

	KNN	Linear	Lasso	Elastic Net	Ridge	XGBoost	GBR	Random Forest
r2_test	0.41	5.600000e-01	0.56	0.65	0.65	0.86	0.86	0.87



Avec l'ESS
r2 : 0.812
r2 ajusté : 0.803

	KNN	Linear	Ridge	Lasso	Elastic Net	Random Forest	XGBoost	GBR	XGBoostESS
r2_test	0.163	0.334	0.334	0.336	0.336	0.779	0.811	0.812	0.812

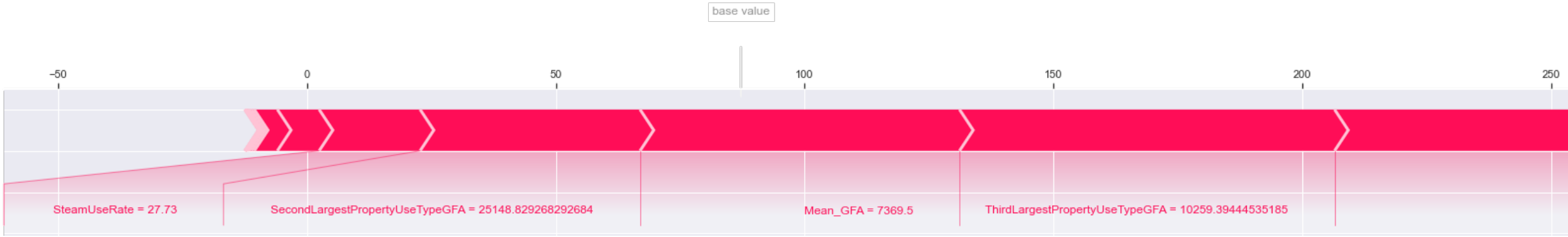


PARTIE V

Interpretation

Interpretation SHAP (Force Plot)

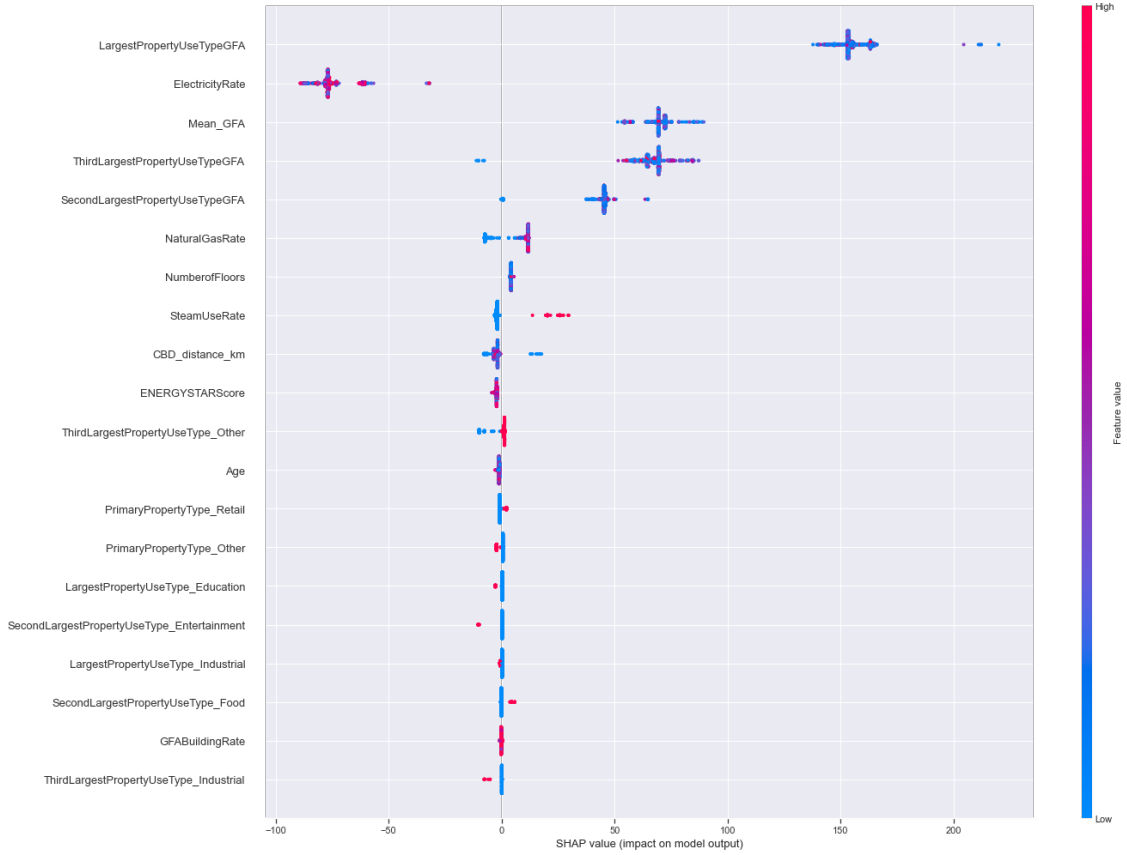
EMISSIONS DE GAZ À EFFET DE SERRE



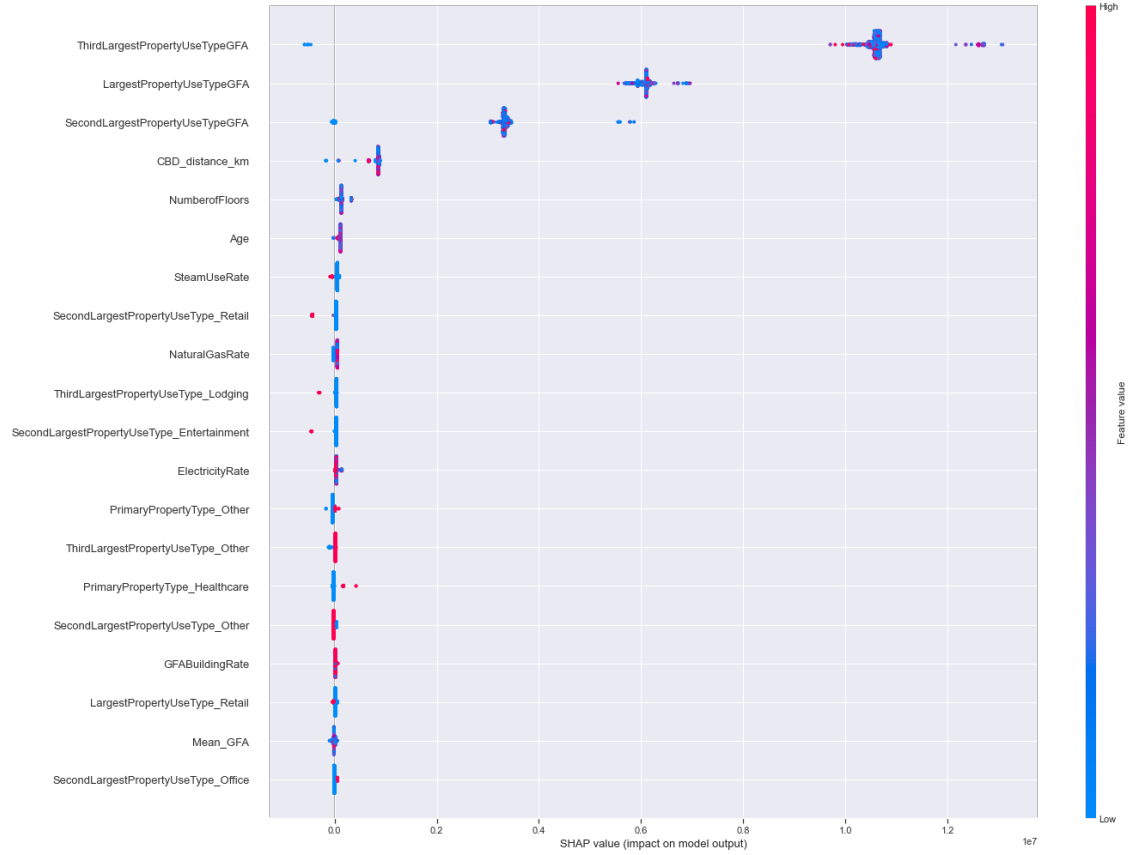
CONSOMMATION D'ENERGIE

Interpretation SHAP (Summary Plot)

EMISSIONS DE GAZ À EFFET DE SERRE



CONSOMMATION D'ENERGIE



Merci

Alexandre Delaguillaumie