Projet P4

Anticipez les besoins en consommation de bâtiments

OPENCLASSROOMS

Introduction

Objectif :

- Prédiction de la consommation totale d'énergie et des émissions de CO2 des bâtiments sur Seattle
- Débattre de l'intérêt de l'ENERGYSTARScore sur ces prédiction

Sommaire

1. Nettoyage des données
2. Analyse pré-exploratoire
3. Feature Engineering
4. Modélisation
5. Conclusion

Nettoyage des données : Dataset

	OSEBuildingID	DataYear	BuildingType	PrimaryPropertyType	PropertyName	Address	City	State	ZipCode	TaxParcelIdentificationNumber	...
0	1	2016	NonResidential	Hotel	Mayflower park hotel	405 Olive way	Seattle	WA	98101.0	0659000030	...
1	2	2016	NonResidential	Hotel	Paramount Hotel	724 Pine street	Seattle	WA	98101.0	0659000220	...
2	3	2016	NonResidential	Hotel	5673-The Westin Seattle	1900 5th Avenue	Seattle	WA	98101.0	0659000475	...
3	5	2016	NonResidential	Hotel	HOTEL MAX	620 STEWART ST	Seattle	WA	98101.0	0659000640	...
4	8	2016	NonResidential	Hotel	WARWICK SEATTLE HOTEL (ID8)	401 LENORA ST	Seattle	WA	98121.0	0659000970	...
...	...	...	...	...	...	...	...	...	...	...	...
3371	50222	2016	Nonresidential COS	Office	Horticulture building	1600 S Dakota St	Seattle	WA	NaN	1624049080	...
3372	50223	2016	Nonresidential COS	Other	International district/Chinatown CC	719 8th Ave S	Seattle	WA	NaN	3558300000	...
3373	50224	2016	Nonresidential COS	Other	Queen Anne Pool	1920 1st Ave W	Seattle	WA	NaN	1794501150	...
3374	50225	2016	Nonresidential COS	Mixed Use Property	South Park Community Center	8319 8th Ave S	Seattle	WA	NaN	7883603155	...
3375	50226	2016	Nonresidential COS	Mixed Use Property	Van Asselt Community Center	2820 S Myrtle St	Seattle	WA	NaN	7857002030	...

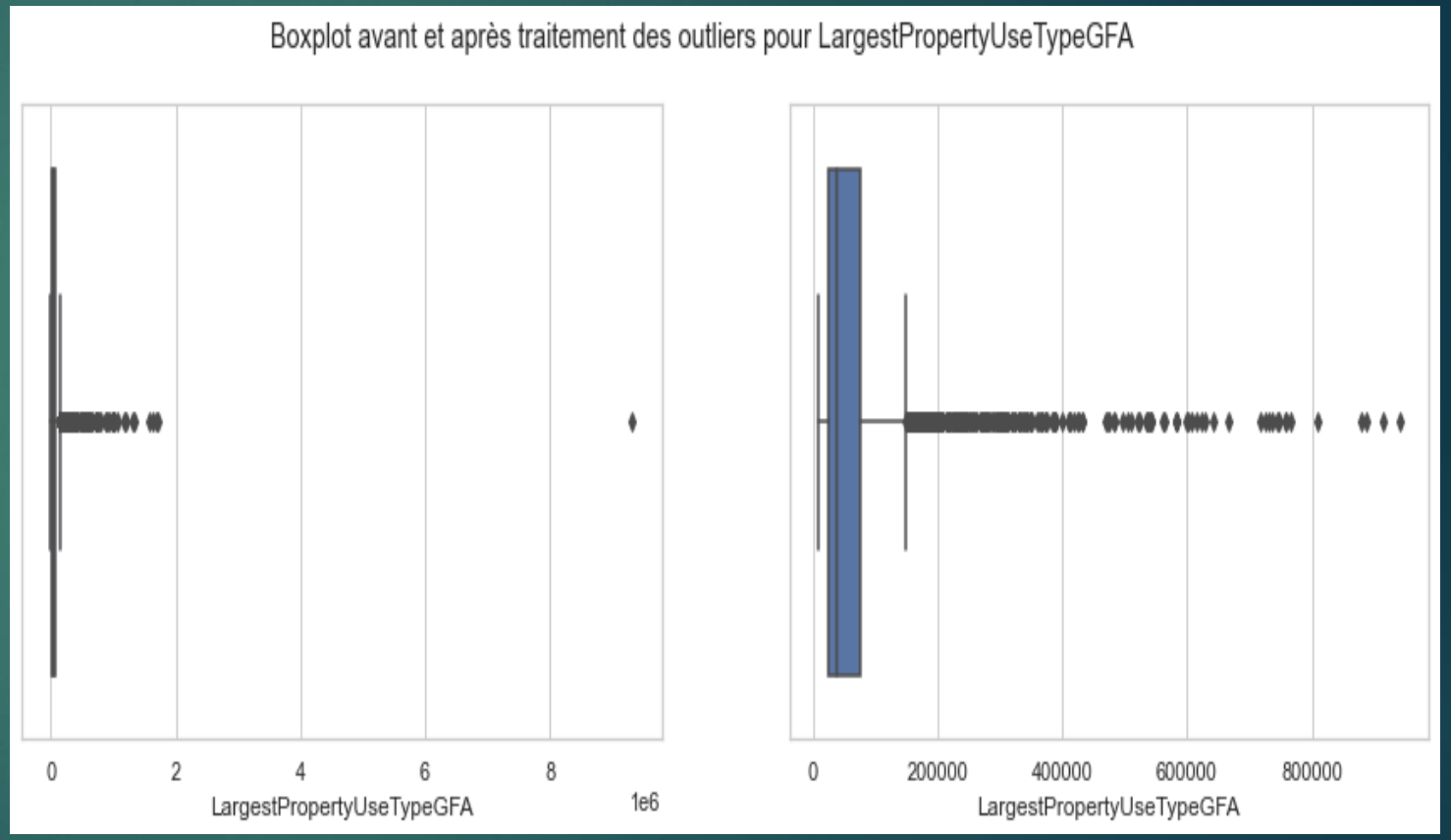
3376 rows × 46 columns

Nettoyage des données

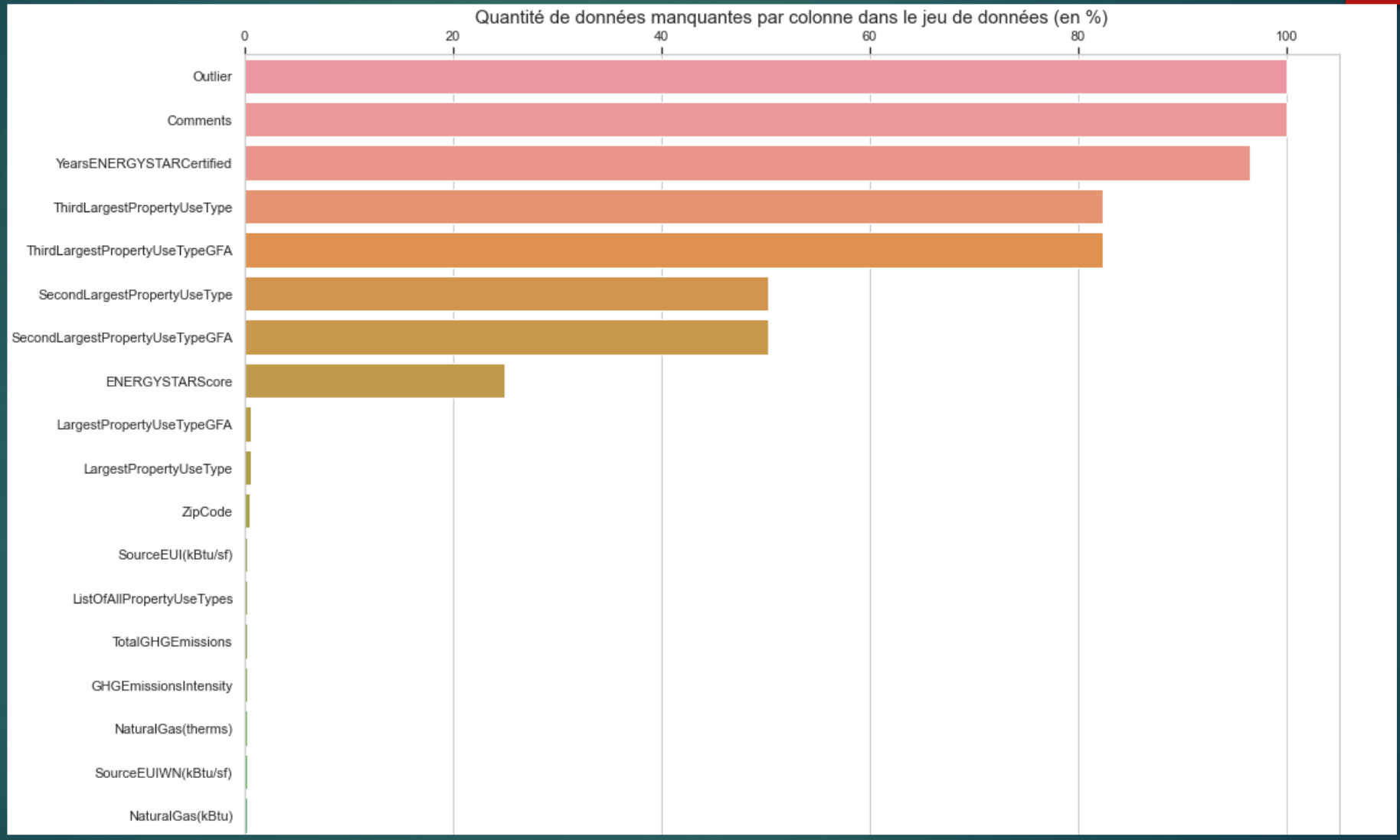
1. Elimination des outliers
2. Contrôle des colonnes en fonction du % de valeurs manquant
3. Suppression des bâtiments résidentiels
4. Imputation des colonnes ‘object’
5. Elimination des colonnes inutiles
6. Imputation via KNN imputer

Nettoyage des données : Supression des Outliers

- Suppression des lignes low et high outliers (32 lignes)
- Suppression des valeurs aberrante en utilisant la méthode du centile extrême



Nettoyage des données : Contrôle des colonnes



Nettoyage des données : Suppression des bâtiments résidentiels

On élimine les lignes pour lesquels BuildingType vaut :

- Multifamily MR (5-9)
- Multifamily HR (10+)
- Multifamily LR (1-4)

1693 lignes supprimé

Nettoyage des données : Imputation colonnes object

Les colonnes de type object et possèdant des valeurs manquantes

- LargestPropertyUseType
- SecondLargestPropertyUseType
- ListOfAllPropertyUseTypes

Remplacement par ‘no information’

Nettoyage des données : Elimination des colonnes inutiles

L'intérêt de ce projet est de prédire la consommation d'énergie ainsi que les émission de CO2. Les variables d'intérêt sont donc :

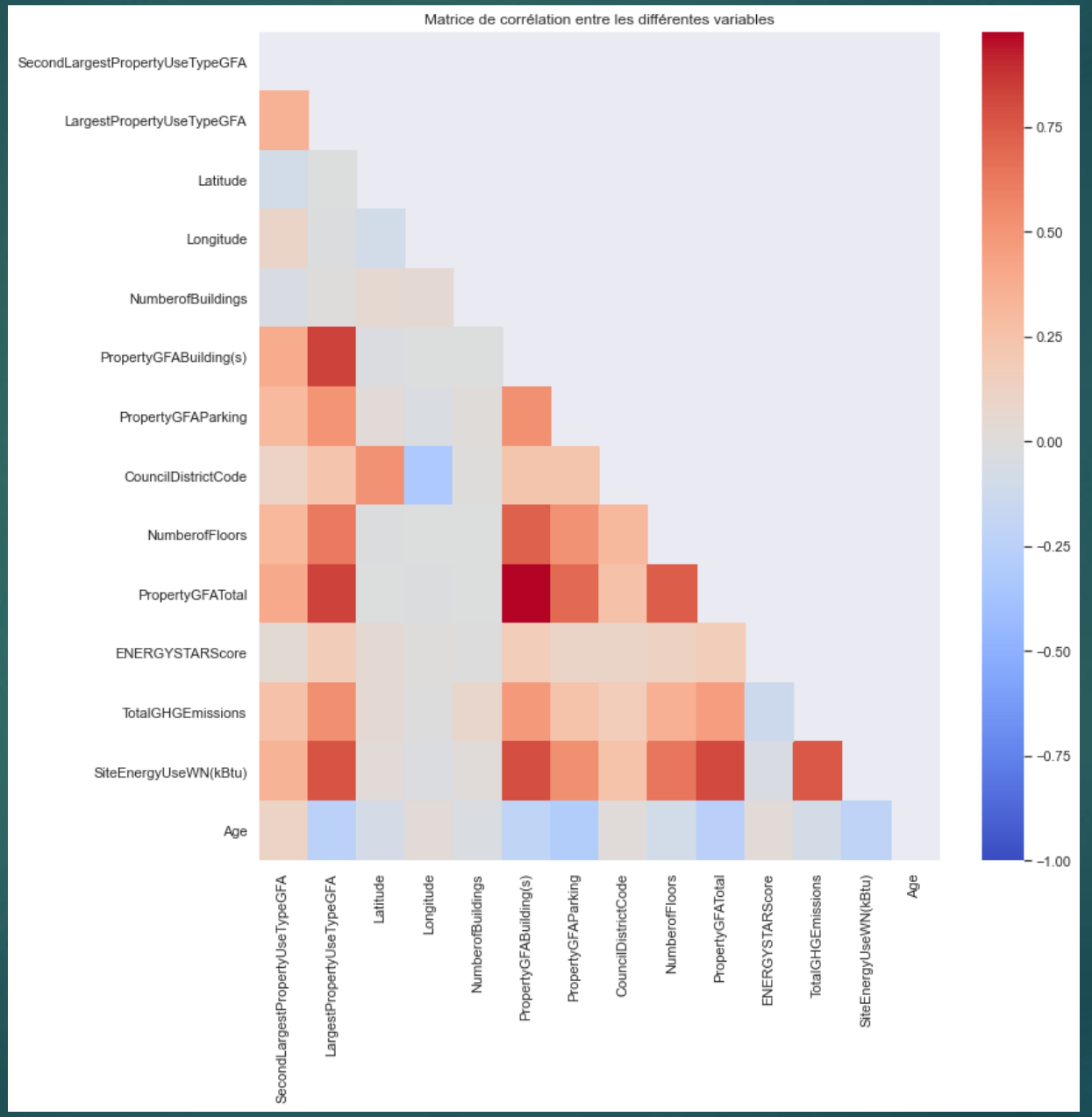
- **SiteEnergyUseWN(kBtu) (consommation totale d'énergie)**
- **TotalGHGEmissions (émission de C02)**

On peut supprimer les variables de relevé de consommation annuelles et les variables peu pertinent (par exemple : adress)

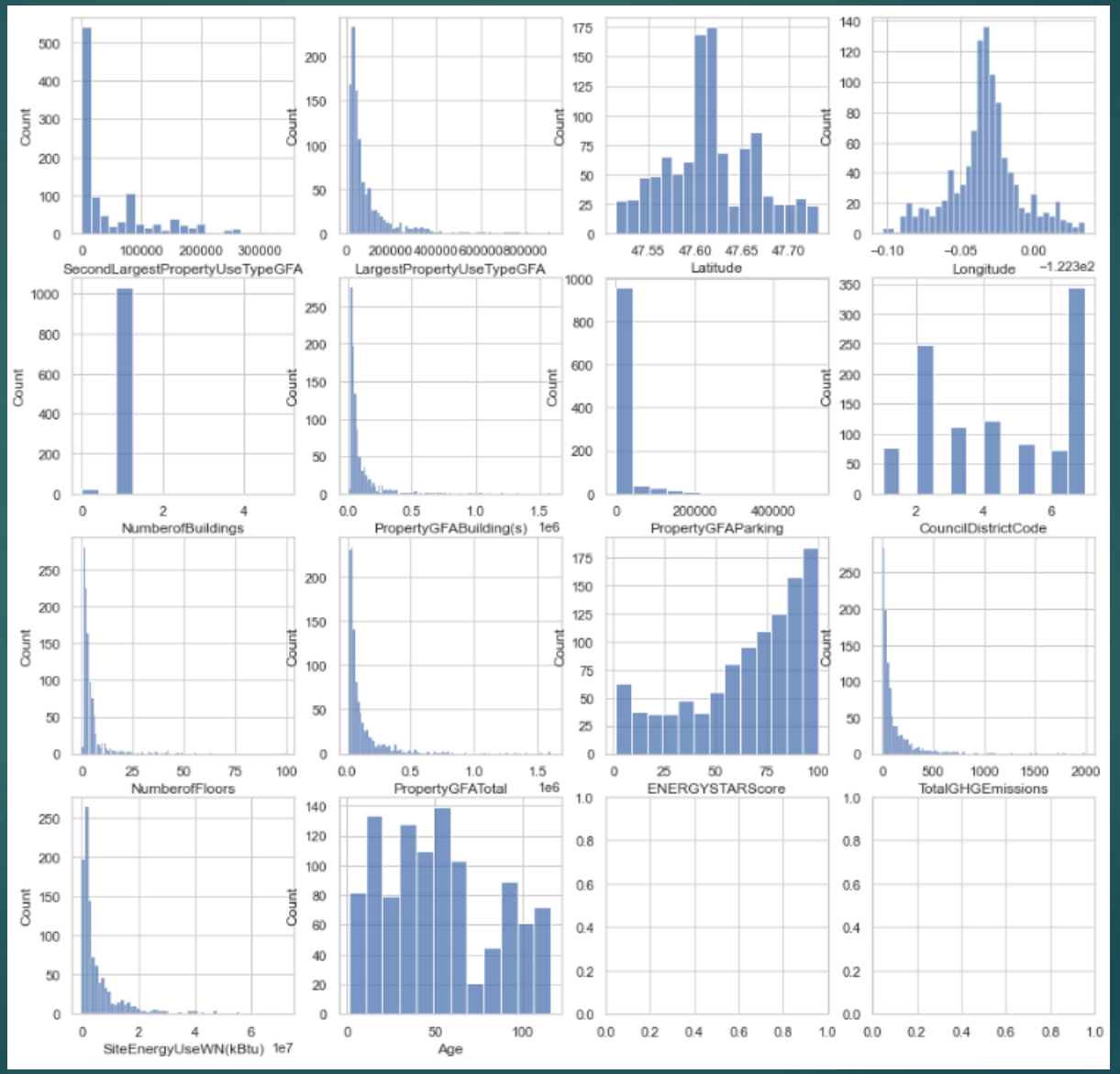
On supprime donc 21 colonnes

On peut ensuite faire un KNNImputer

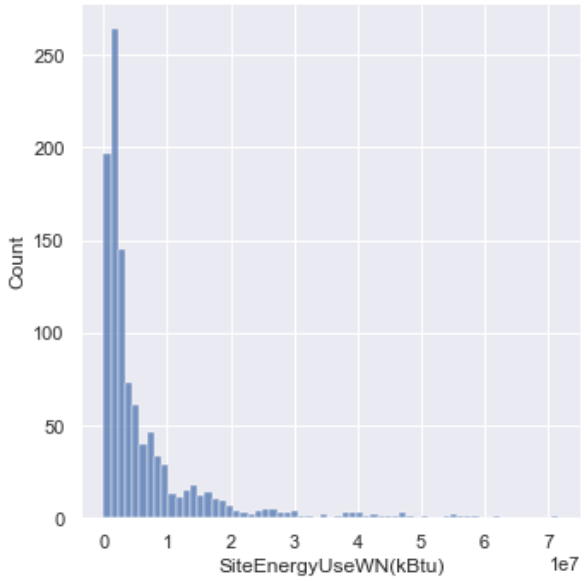
Analyse des données : Analyse Bivariés

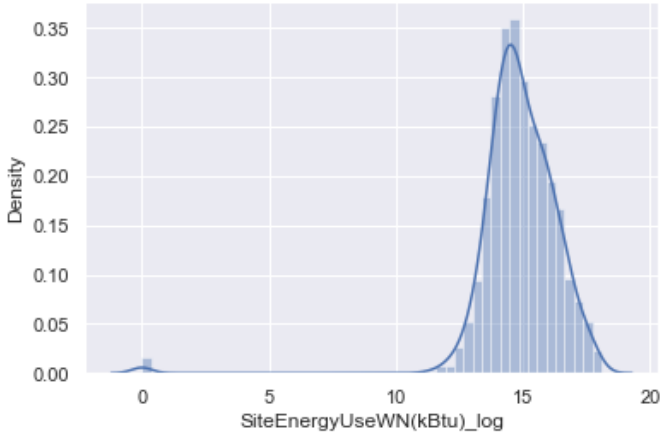


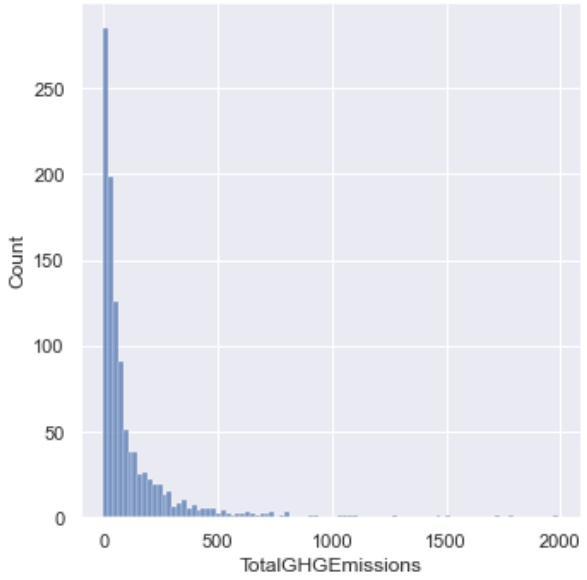
Analyse des données : Analyse Univariés

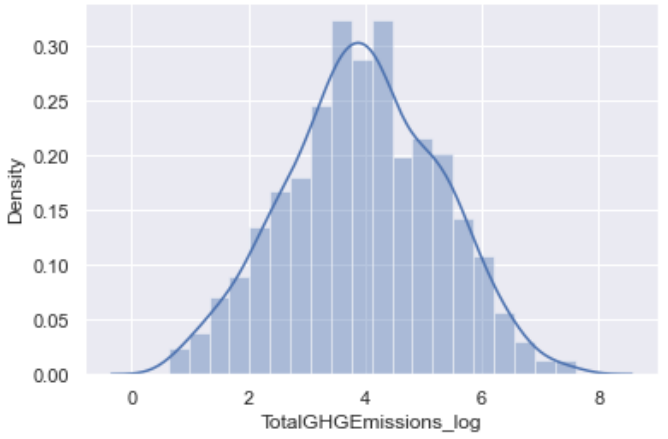


Feature Engineering : Transformation log des variable d'intérêt









Feature Engineering : Réunifications des Property Use

- On créer des colonnes pour chaque valeur des PropertyUse (Largest et SecondLargest)
- On calcule la % de GFA pour chaque valeur :

$$\frac{Largest\ GFA}{Largest\ GFA + SecondLargest\ GFA} \times 100$$

- Exemple :

SecondLargestPropertyUseTypeGFA	LargestPropertyUseTypeGFA	LargestPropertyUseType	SecondLargestPropertyUseType
165161.00	88434.0	Hotel	no information

LargestPropertyUseType_Hotel	.....	SecondLargestPropertyUseType_no information
34,87213865		65,12786135

Modélisation SiteEnergyUseWN(kBtu) : Mise en place des modèles

On va utiliser 6 modèles :

- Dummy Regressor (baseline : modèle simple)
- Linear Regression
- kNN Regressor
- RandomForest Regressor
- Adaboost Regressor
- MultiLayerPerceptron Regressor

Auparavant on a passé au log les variables avec une skewness élevé pour voir s'il améliore la performances des modèles

Pour évaluer les modèles on va utiliser les métrique suivante :

- Root Mean Squared Error
- Mean Absolute Error
- R2

Modélisation SiteEnergyUseWN(kBtu) : Mise en place des modèles

```
numerical_feature = make_column_selector(dtype_include=np.number)
categorical_feature = make_column_selector(dtype_exclude=np.number)
```

executed in 16ms, finished 00:09:27 2022-06-19

```
numeric_transformer = Pipeline(steps=[('scaler', StandardScaler())])
categorical_transformer = Pipeline(steps=[('onehot', OneHotEncoder(handle_unknown='ignore'))])
```

executed in 16ms, finished 00:09:29 2022-06-19

```
preprocessor = ColumnTransformer(transformers=[('num', numeric_transformer, numerical_feature),
                                               ('cat', categorical_transformer, categorical_feature)])
```

executed in 16ms, finished 00:09:32 2022-06-19

```
#baseline
dummy = Pipeline(steps=[('prepa',preprocessor),('dummy',DummyRegressor())])

# Modele à tester

linear = Pipeline(steps=[('prepa',preprocessor),('linear',linear_model.LinearRegression())])
KNN = Pipeline(steps=[('prepa',preprocessor),('kNN',KNeighborsRegressor())])

forest = Pipeline(steps=[('prepa',preprocessor),('forest',RandomForestRegressor(random_state=0))])
ADA = Pipeline(steps=[('prepa',preprocessor),('ADA',AdaBoostRegressor(random_state=0,))])
MLPr = Pipeline(steps=[('prepa',preprocessor),('MLPr',MLPRegressor(random_state=0,max_iter=10000))])
```

executed in 16ms, finished 00:09:35 2022-06-19

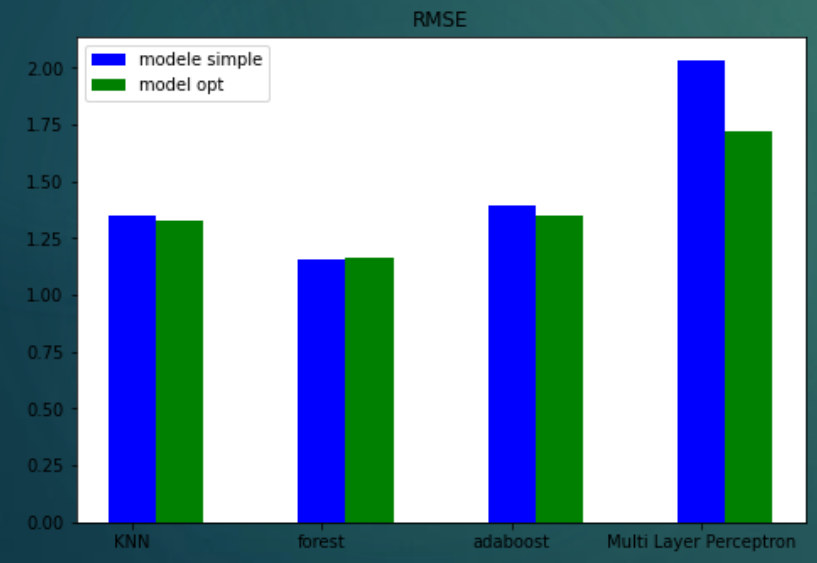
Modélisation SiteEnergyUseWN(kBtu) : Mise en place des modèles

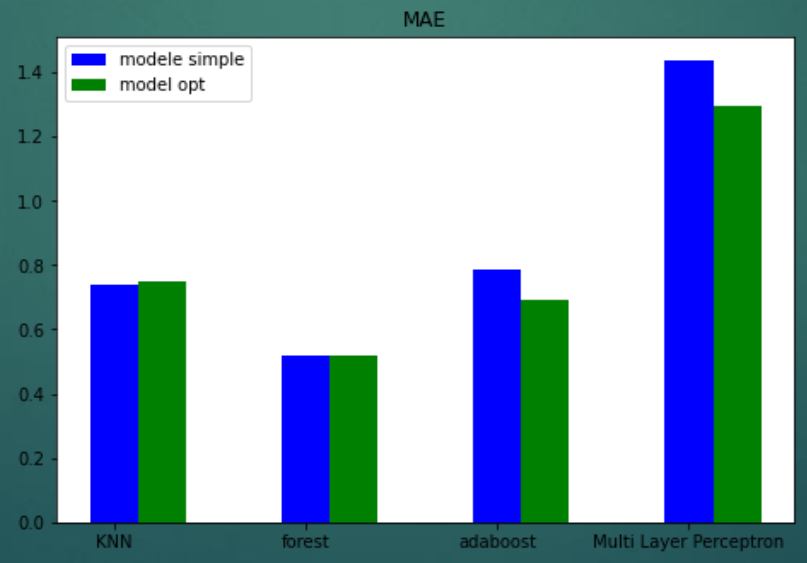
Variable normal

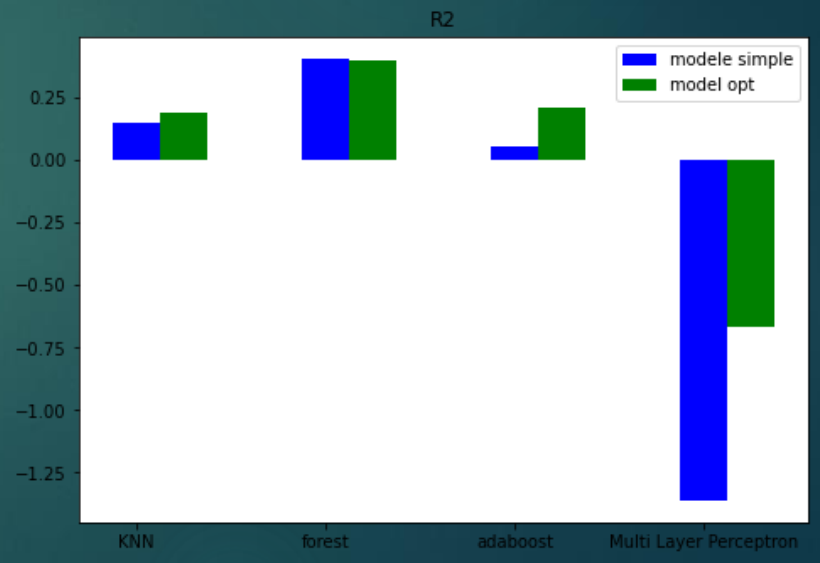
	RMSE	MAE	R2
KNN	1.323633	0.750961	0.186164
forest	1.163364	0.520865	0.394769
adaboost	1.345124	0.691426	0.207337
Multi Layer Perceptron	1.717469	1.295822	-0.666734

Variable log

	RMSE	MAE	R2
KNN	1.350427	0.736835	0.148611
forest	1.156751	0.517939	0.403341
adaboost	1.395702	0.784397	0.052405
Multi Layer Perceptron	2.034153	1.439052	-1.359702







Modélisation SiteEnergyUseWN(kBtu) : Recherche hyperparamètres

Paramètres optimiser pour chaque modèles

- kNN Regressor :
  - N_neigbors : 11
  - Leaf size : 2
  - P (distance) : 1 (distance de Manhattan)

- RandomForest Regressor
  - N_estimators : 138
  - Min_samples_leaf : 1
  - Max_depth : 14

- Adaboost Regressor
  - N_estimators : 1
  - Loss : linear

- MultiLayerPerceptron Regressor
  - Hidden layer size : 95
  - Activation : identity

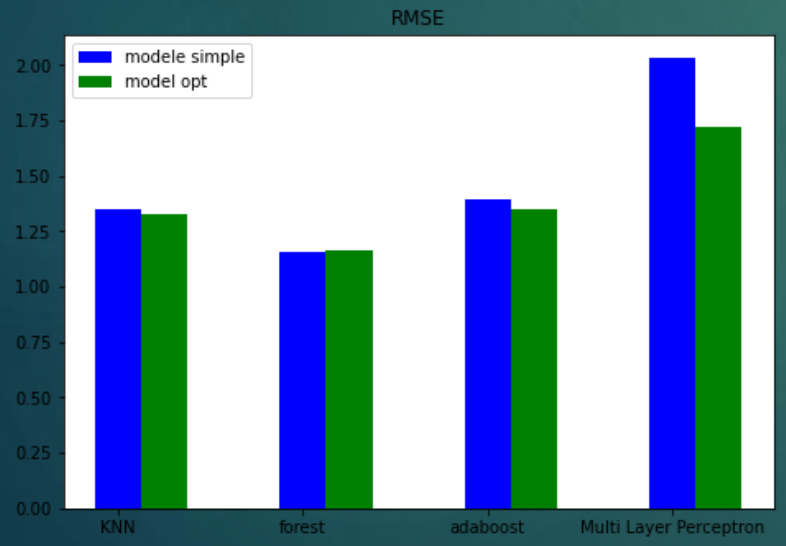
Modélisation SiteEnergyUseWN(kBtu) : Recherche hyperparamètres

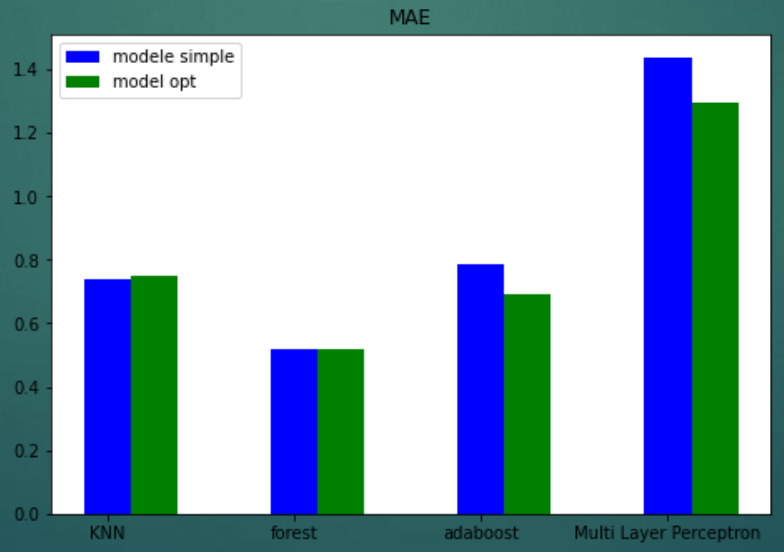
Modèle optimisé

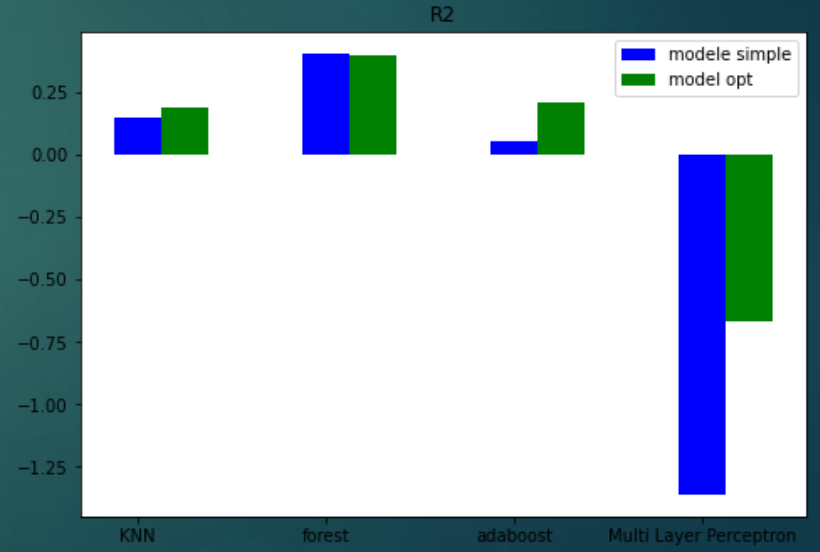
	RMSE	MAE	R2
KNN	1.323633	0.750961	0.186164
forest	1.163364	0.520865	0.394769
adaboost	1.345124	0.691426	0.207337
Multi Layer Perceptron	1.717469	1.295822	-0.666734

Modèles simple

	RMSE	MAE	R2
KNN	1.350427	0.736835	0.148611
forest	1.156751	0.517939	0.403341
adaboost	1.395702	0.784397	0.052405
Multi Layer Perceptron	2.034153	1.439052	-1.359702

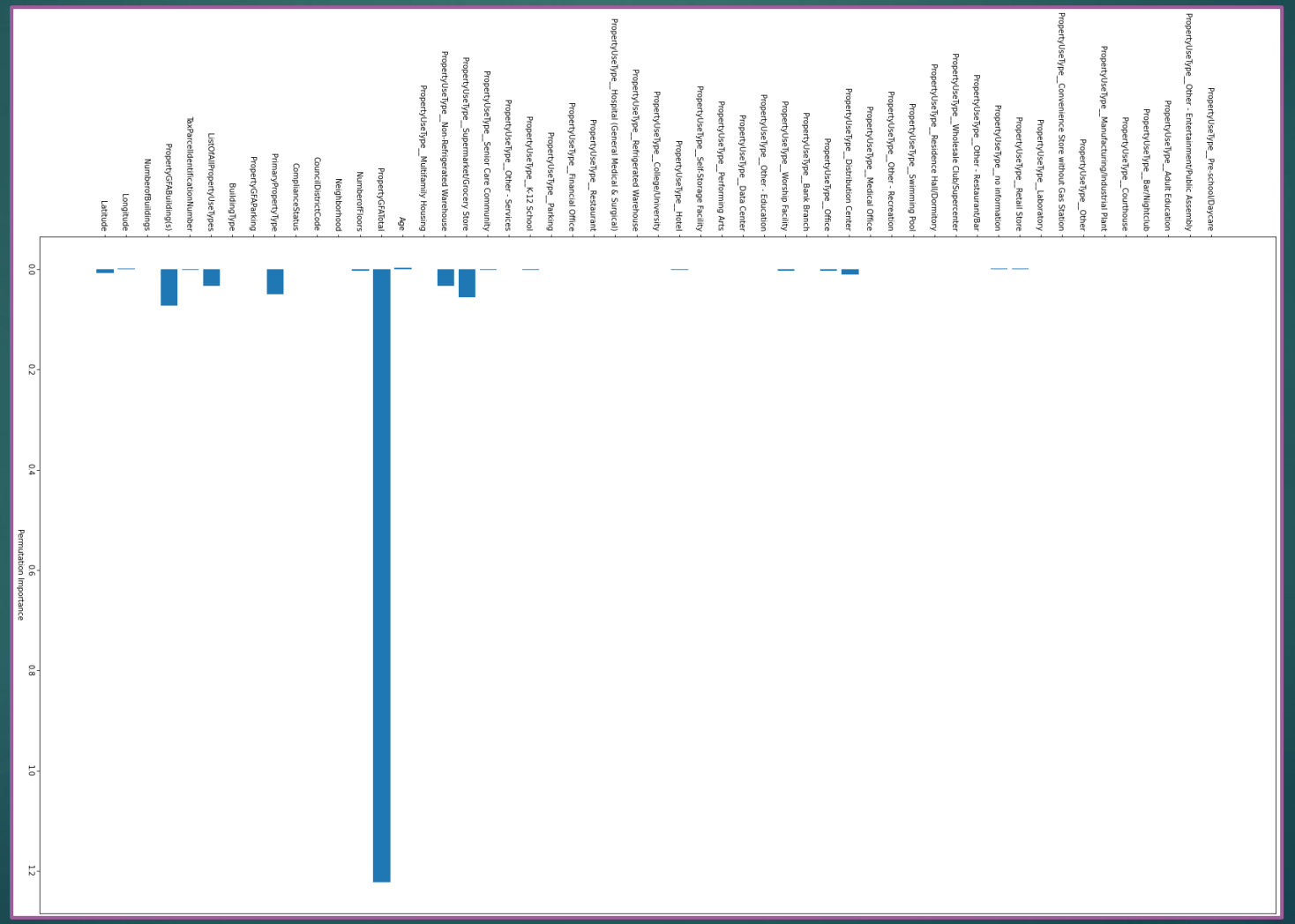






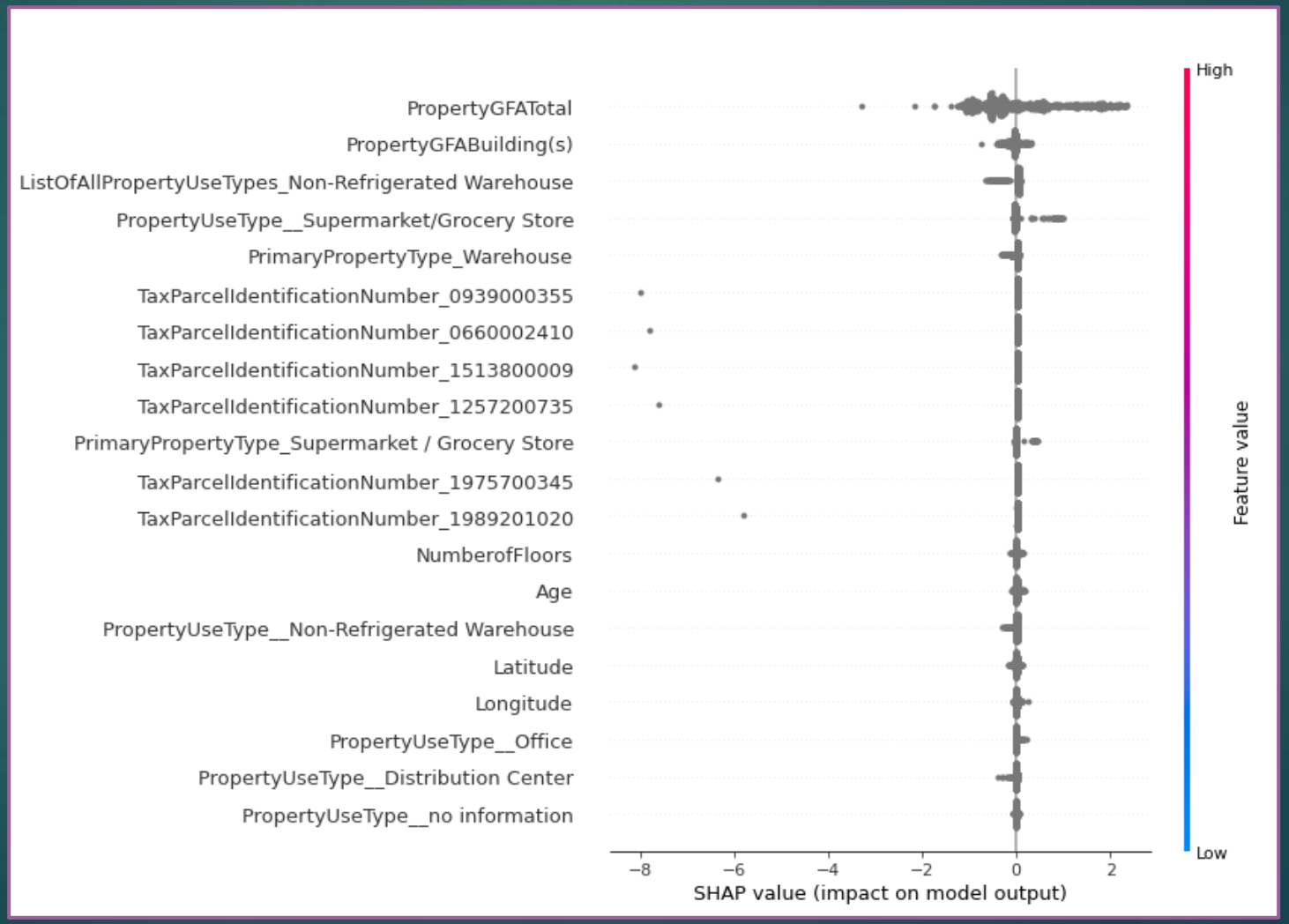
Modélisation SiteEnergyUseWN(kBtu) : Feature importance

Méthode 1 : Permutation importance



Modélisation SiteEnergyUseWN(kBtu) : Feature importance

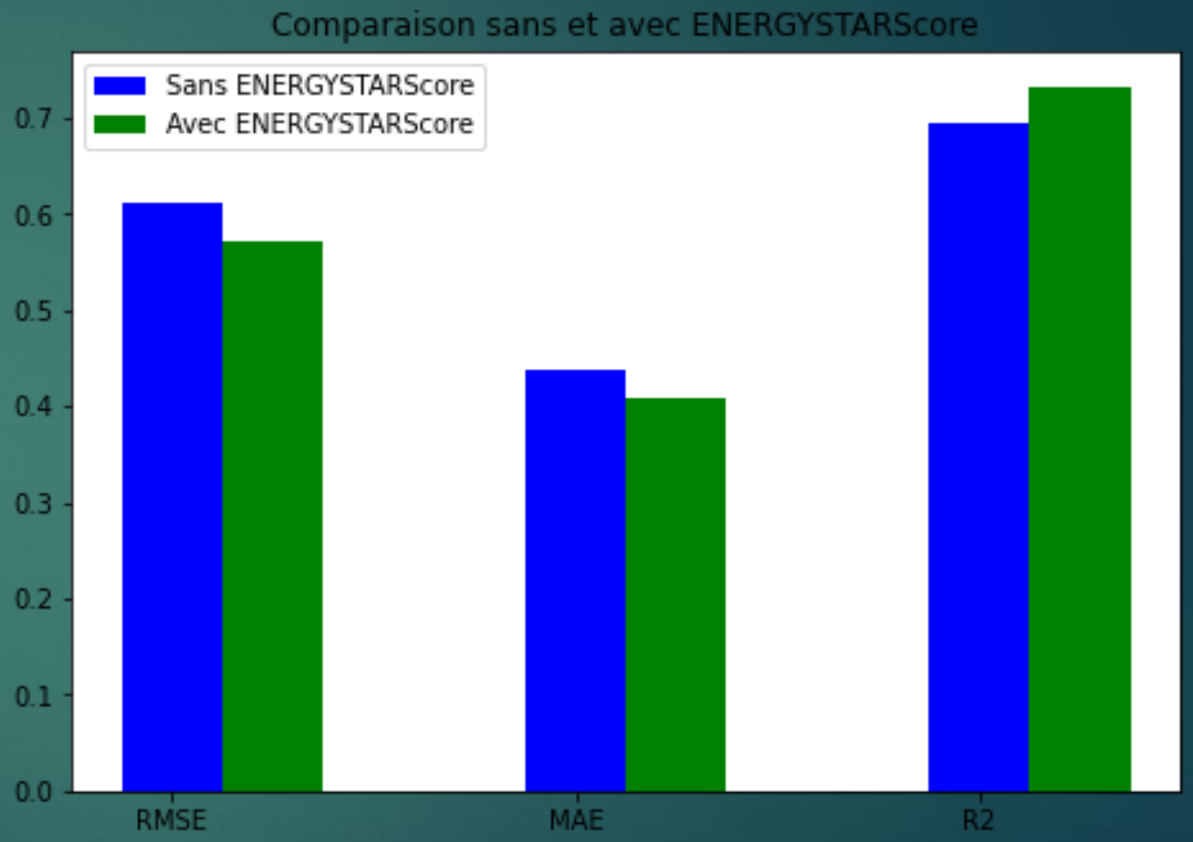
Méthode 2 : Shap



Modélisation SiteEnergyUseWN(kBtu) : Impact de l'Energy Score

	RMSE	MAE	R2
Sans ENERGYSTARScore	0.61	0.44	0.69

	RMSE	MAE	R2
Avec ENERGYSTARScore	0.57	0.4	0.73



Modélisation TotalGHGEmissions

Variables non log

	RMSE	MAE	R2
dummy	1.390852	1.130045	-0.441166
lr	126.295637	72.105794	-24210.560723
KNN	1.114956	0.877215	0.082778
forest	0.899496	0.711367	0.403943
adaboost	0.979836	0.806152	0.297514
Multi Layer Perceptron	1.050659	0.814500	0.195088

Variable log

	RMSE	MAE	R2
dummy	1.390852	1.130045	-0.441166
lr	37.459953	23.794542	-2168.015564
KNN	1.037817	0.817111	0.205443
forest	0.899328	0.711093	0.404227
adaboost	0.983600	0.813147	0.292406
Multi Layer Perceptron	0.901695	0.718799	0.402057

Recherche hyperparamètres

Modèle non optimisé

	RMSE	MAE	R2
KNN	1.037817	0.817111	0.205443
forest	0.899328	0.711093	0.404227
adaboost	0.983600	0.813147	0.292406
Multi Layer Perceptron	0.901695	0.718799	0.402057

Modèle optimisé

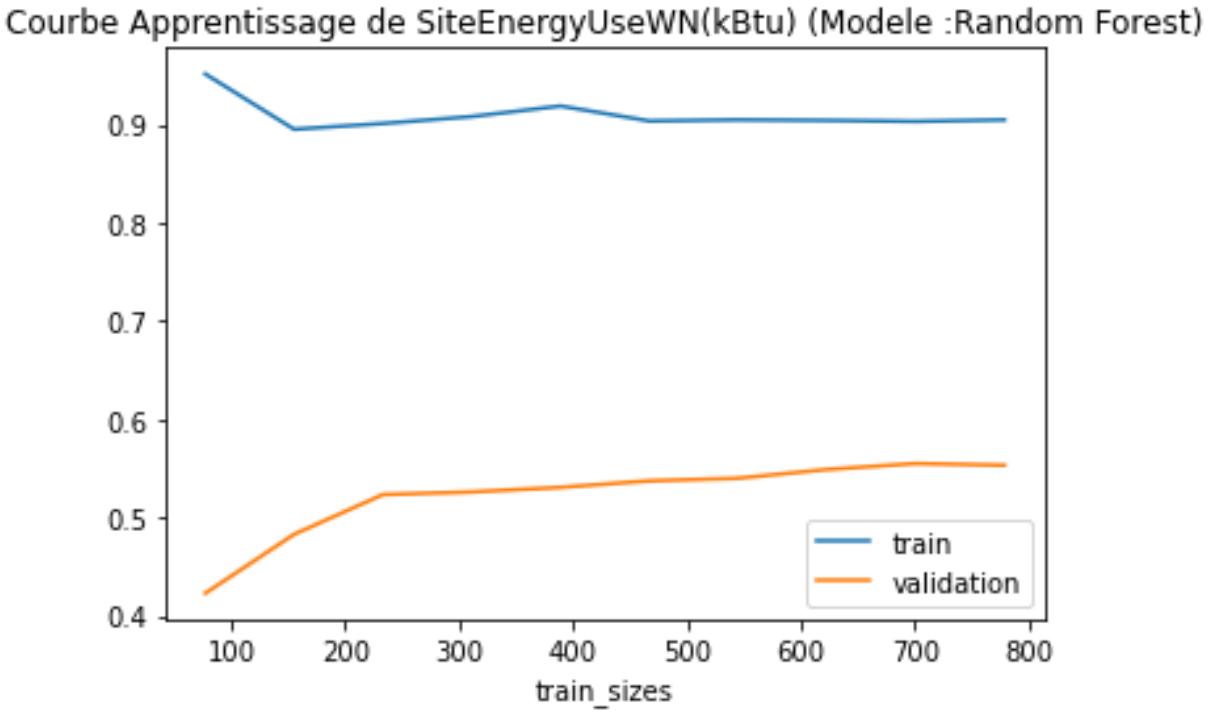
	RMSE	MAE	R2
KNN	1.029463	0.815951	0.217353
forest	0.884420	0.698930	0.404816
adaboost	0.987216	0.816907	0.287614
Multi Layer Perceptron	0.878647	0.706649	0.409120

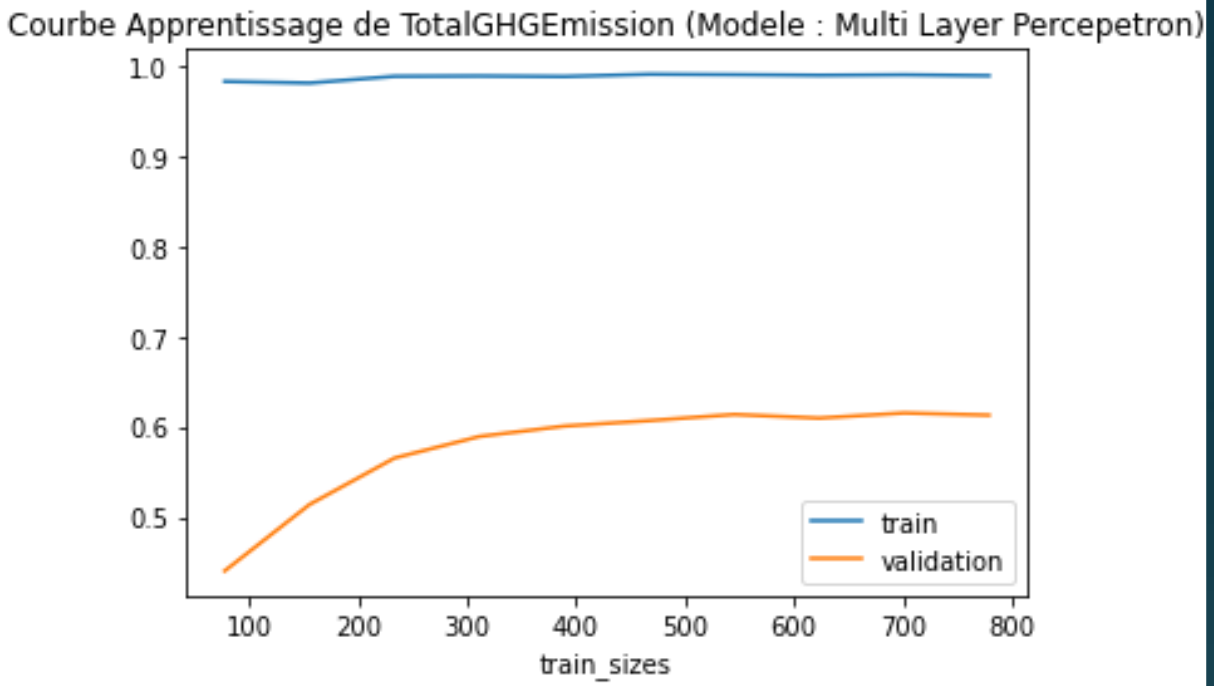
Impact ENERGYSTARScore sous MLP

	RMSE	MAE	R2
Sans ENERGYSTARScore	0.91	0.74	0.51

	RMSE	MAE	R2
Avec ENERGYSTARScore	0.86	0.69	0.56

Courbe d'apprentissage





Conclusion

Modèle pertinent :

- Random Forest pour la consommation d'énergie
  - N_estimators : 138
  - Min_samples_leaf : 1
  - Max_depth : 14

- Multi Layer Perceptron pour les émission de CO2
  - Hidden layer size : 105
  - Activation : logistic

Impact de l'ENERGYSTARScore +

- Pour la consommation d'energie :
  - R2 : 0.51 avec ESS vs 0.56 sans

- Pour les emission de CO2
  - R2 : 0.73 avec ESS vs 0.69 sans

Merci de votre attention