
Anticipez les
besoins en
consommation
de bâtiments

PLAN

1/ Présentation de la problématique
2/ Présentation du jeu de données
3/ Exploration du jeu de données
4/ Exploration et analyse du jeu de données
5/ Modélisation, recherche du meilleur estimateur
6/ Conclusion
7/ Suite du projet

Ville de **Seattle** : ville neutre en émissions carbone en 2050

Objectifs :

- S'intéresser à la consommation et aux émissions des **bâtiments non destinés à l'habitation**

- Evaluer l'intérêt de l'ENERGYSTARScore pour les prédictions

Fichier CSV sur les caractéristiques et les consommations des bâtiments de 2016 (3376 lignes et 46 colonnes)

Lignes :

Bâtiments

Colonnes :

Colonnes textuelles (type de bâtiments, usage du bâtiment, quartier, nom, adresse, code postal)
Colonnes numériques (année de construction, indicateurs de surface, ENERGYSTARScore, relevés de consommation annuels)

- <u>Nettoyage</u>

  Supprimer les bâtiments liés à l'habitation (recherche sur plusieurs variables)

  Supprimer les lignes sans donnée sur les variables numériques

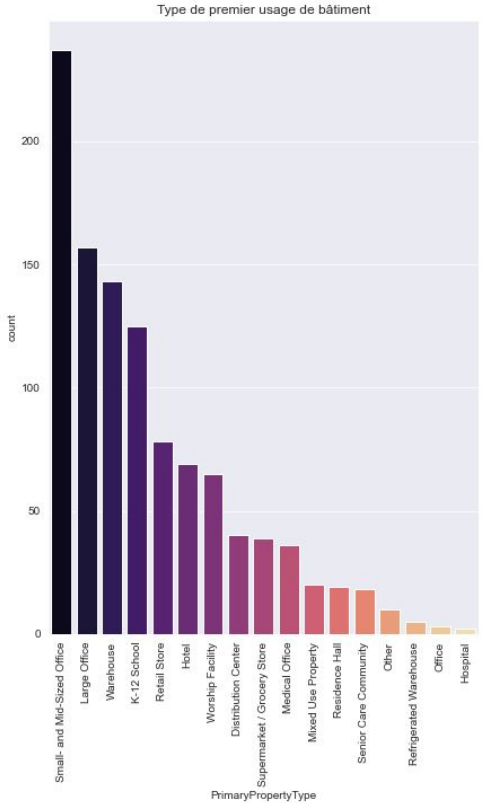
  Supprimer les colonnes avec valeurs manquantes (>45%) : informations sur 2ème et 3ème usage du bâtiment

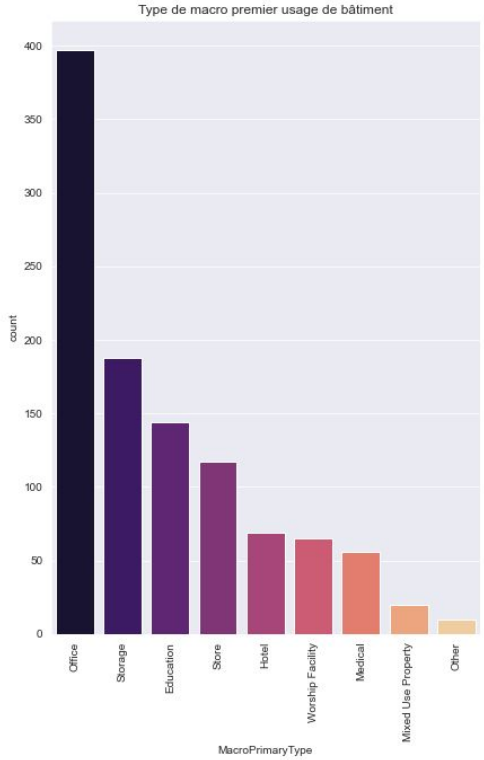
  Réattribuer des informations sur certains bâtiments (bonne catégorie)

  Supprimer les lignes sans ENERGYSTARScore

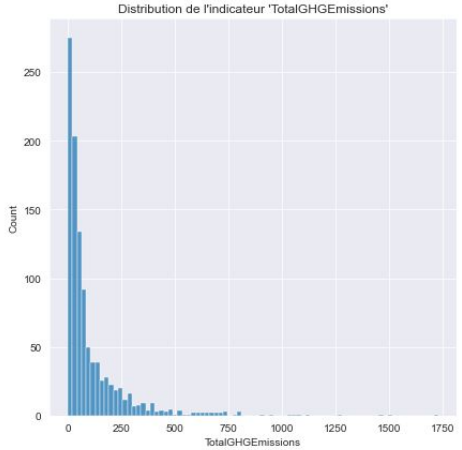
  Garder valeurs < P99 sur variables cibles (supprimer les outliers)

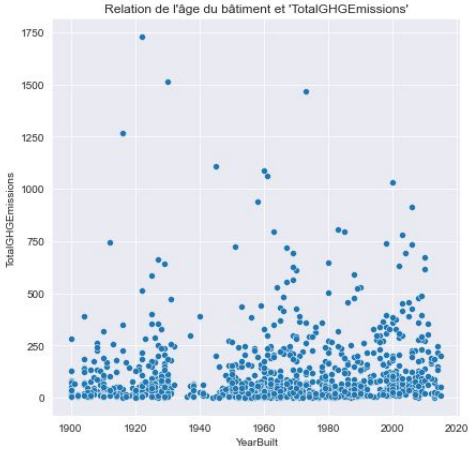
- <u>Exploration visuelle</u>

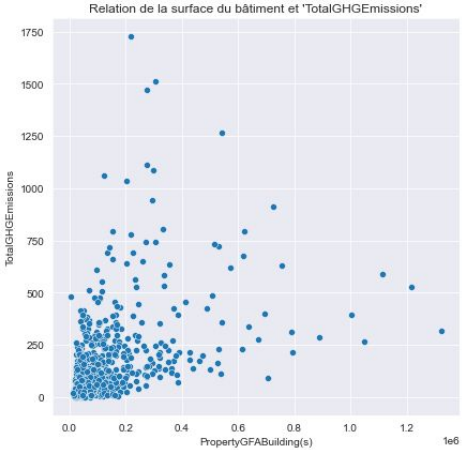


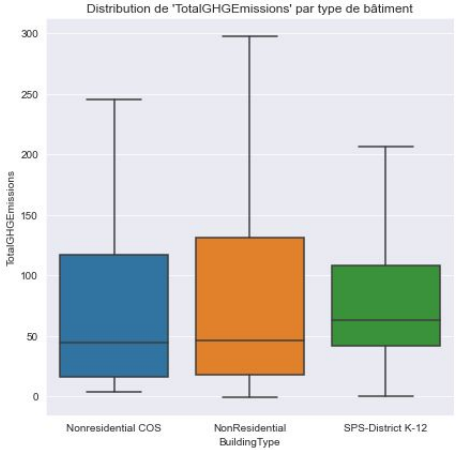


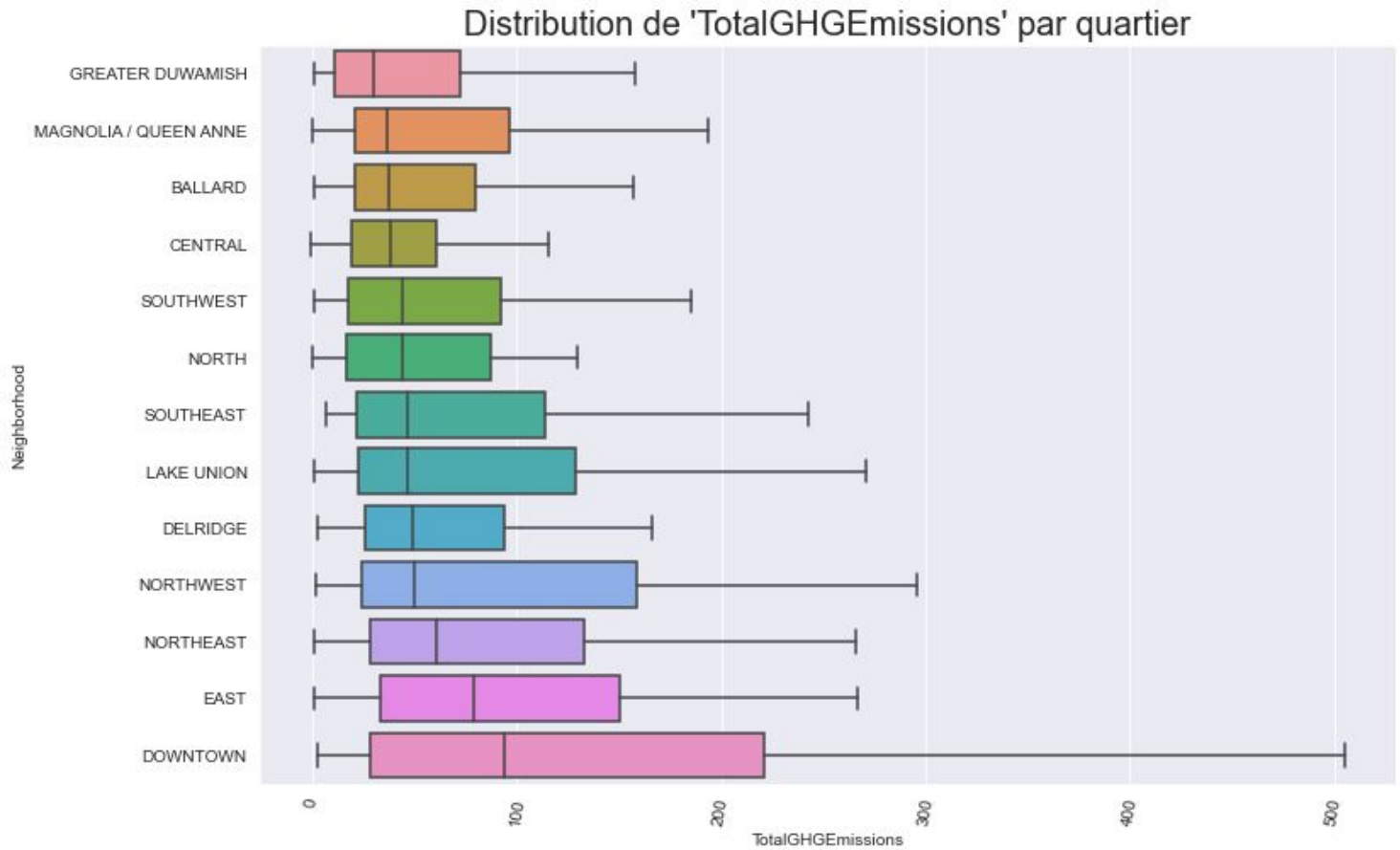
Exploration du jeu de données



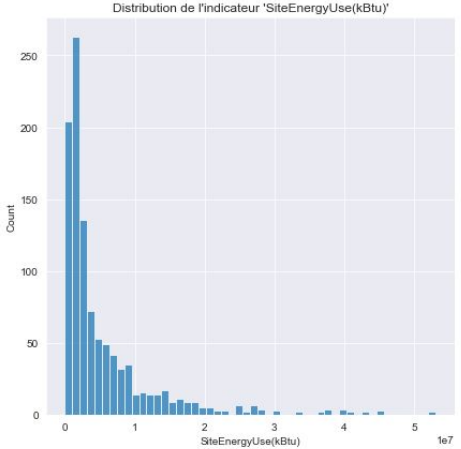


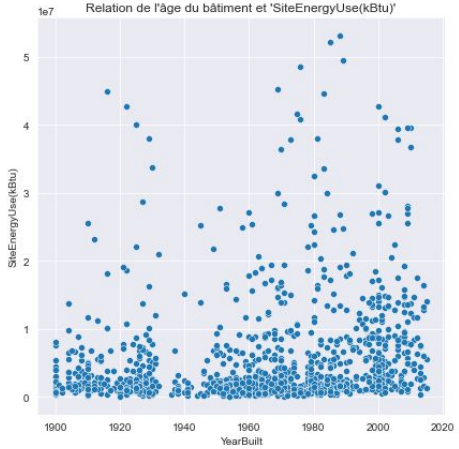


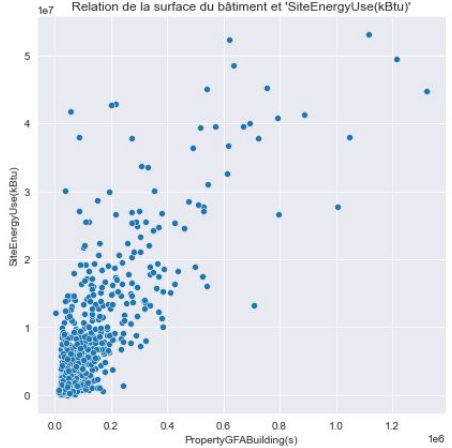


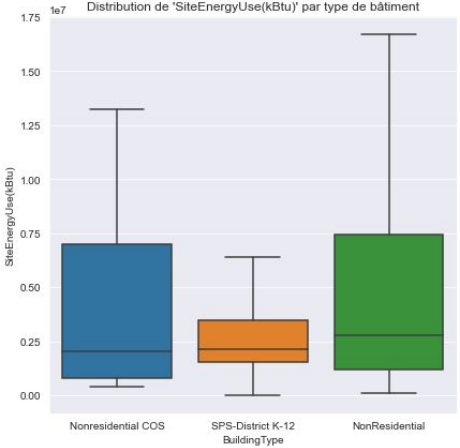

Distribution de 'TotalGHGEmissions' par quartier
GREATER DUWAMISH
MAGNOLIA / QUEEN ANNE
BALLARD
CENTRAL
SOUTHWEST
NORTH
SOUTHEAST
LAKE UNION
DELRIDGE
NORTHWEST
NORTHEAST
EAST
DOWNTOWN
Neighborhood
0
100
200
300
400
500
TotalGHGEmissions

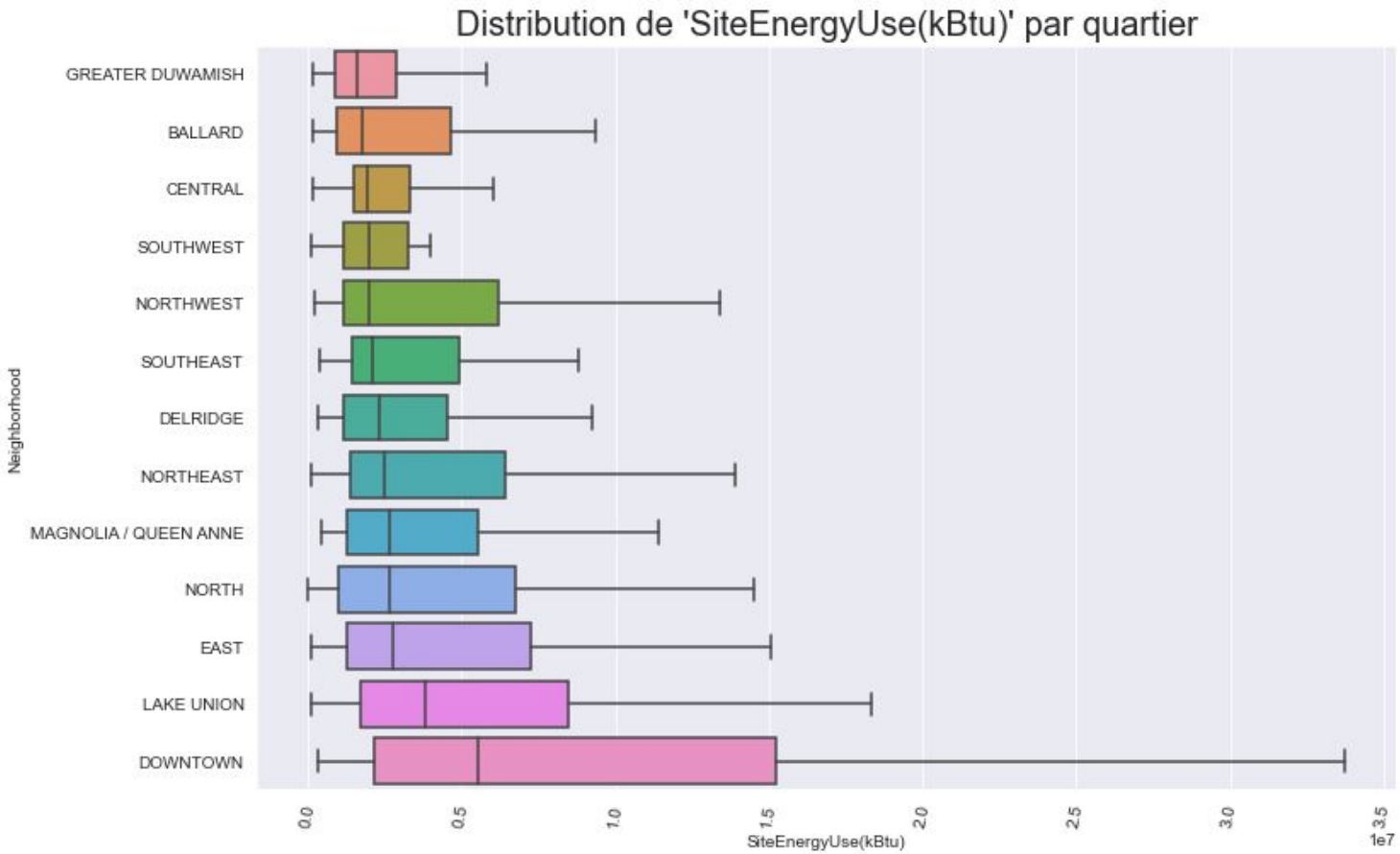
Exploration du jeu de données



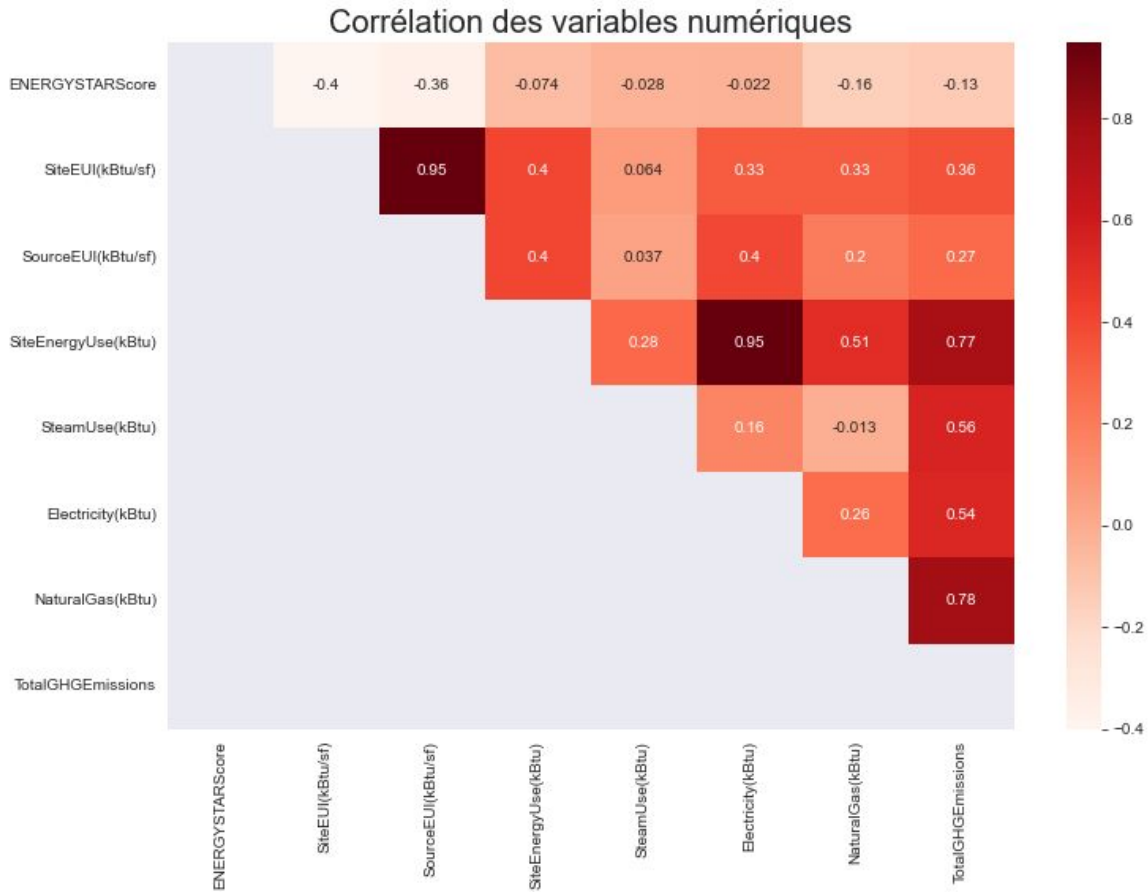







Distribution de 'SiteEnergyUse(kBtu)' par quartier
GREATER DUWAMISH
BALLARD
CENTRAL
SOUTHWEST
NORTHWEST
SOUTHEAST
DELRIDGE
NORTHEAST
MAGNOLIA / QUEEN ANNE
NORTH
EAST
LAKE UNION
DOWNTOWN
Neighborhood
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1e7
SiteEnergyUse(kBtu)

Exploration du jeu de données



- **Feature engineering**

  Créer une macro catégorie de bâtiments à partir de ‘PrimaryPropertyType’ (21 catégories à 10 catégories)

  Créer une variable de volume du bâtiment :
  surface du parking + (nombre d’étages * la surface du batîment)

  Créer une variable de ratio de la surface du parking par rapport à la surface totale

  Jeu de données final de 1066 lignes et 44 colonnes

- **Sélection de variables**

```
['BuildingType', 'PrimaryPropertyType', 'Neighborhood', 'YearBuilt', 'NumberofFloors', 'PropertyGFATotal',
 'PropertyGFAParking', 'PropertyGFABuilding(s)', 'LargestPropertyUseType', 'LargestPropertyUseTypeGFA',
 'ENERGYSTARScore', 'MacroPrimaryType', 'BuildingVolume', 'RatioGFAParking']
```

- **Séparation du jeu de données**
  X : variables sélectionnées
  y : variable cible à prédire

- **Séparation en jeu d'entraînement et de test**
  80/20
  Stratifier sur la variable cible avec KBinsDiscretizer

- **Création d'un column transformer**
  Identification variables catégorielles et numériques

```
cat_var = ['BuildingType', 'PrimaryPropertyType', 'Neighborhood', 'LargestPropertyUseType', 'MacroPrimaryType']
num_var = ['YearBuilt', 'NumberofFloors', 'PropertyGFATotal', 'PropertyGFAParking', 'PropertyGFABuilding(s)',
           'LargestPropertyUseTypeGFA', 'ENERGYSTARScore', 'BuildingVolume', 'RatioGFAParking']
```

OneHotEncocder : variables catégorielles
StandardScaler : variables numériques

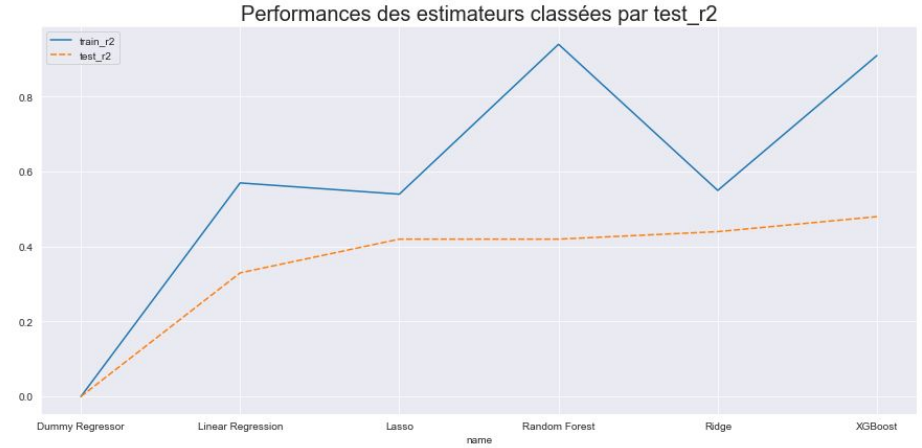
- **Création des estimateurs**
  Liste d'estimateurs avec liste de leurs hyper paramètres
  Dummy Regressor, Linear Regression, Rigde, Lasso, Random Forest et XGBRegressor

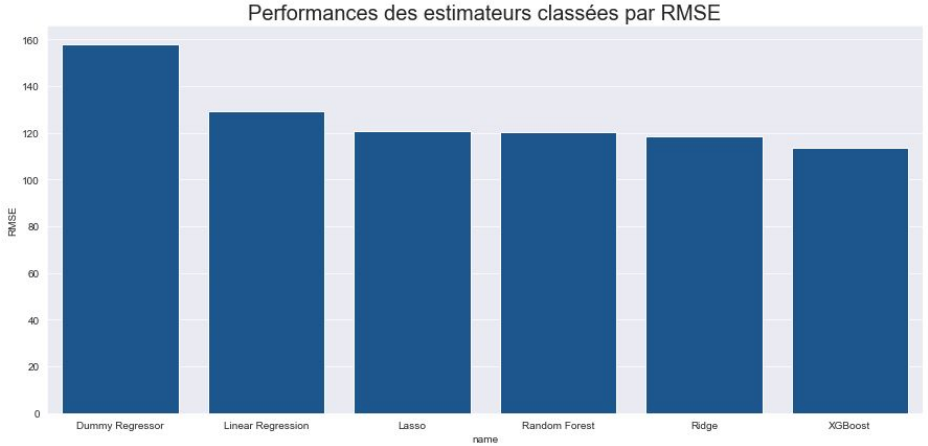
- **Création d'une boucle**
  Recherche des meilleurs hyper paramètres (GridSearchCV)
  Entrainement du meilleur estimateur
  Scores : R^2 (train et test), MAE et RMSE

- **Interprétation des scores**
  Choix de l'estimateur avec les meilleures performances
  R^2 et RMSE (estimateur avec meilleure stabilité)

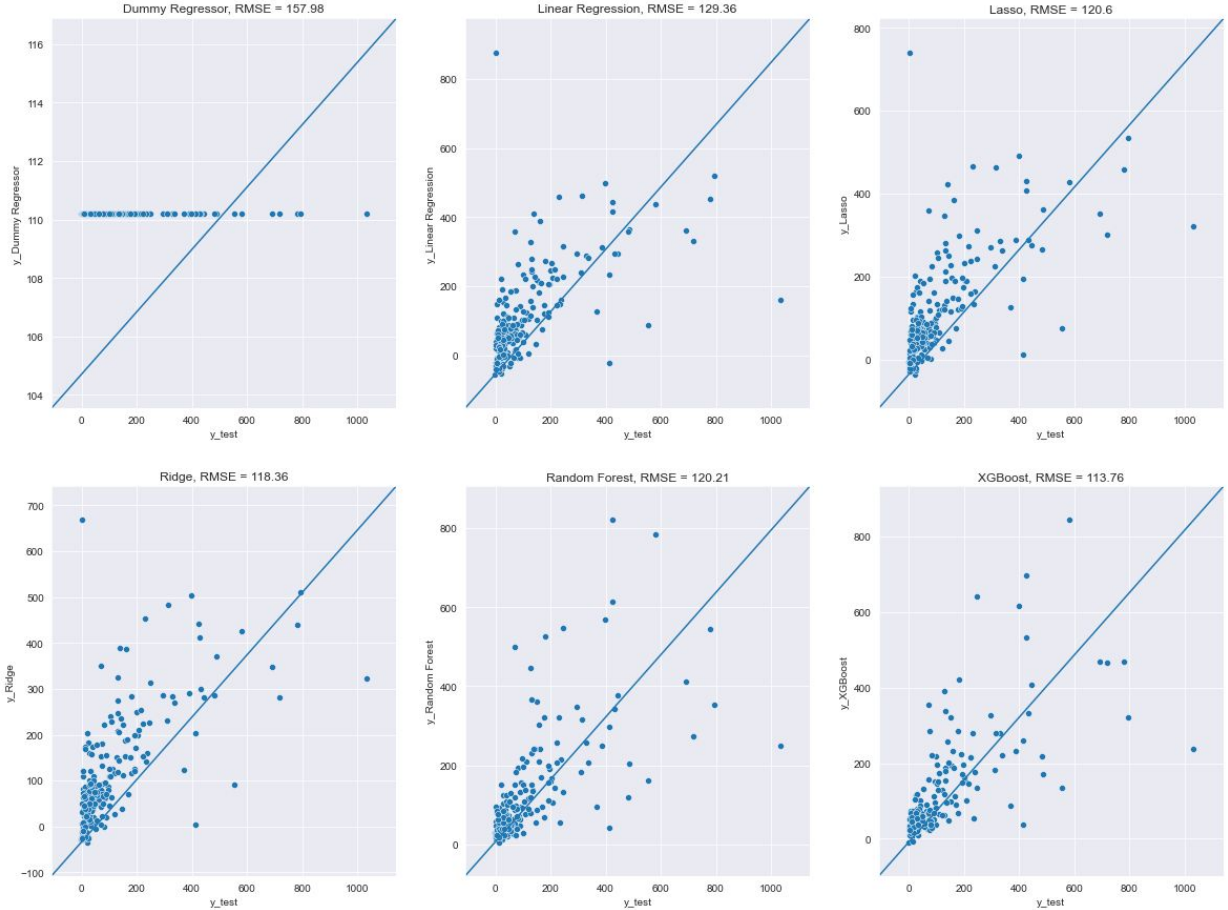
'TotalGHGEmissions'

name	train_r2	test_r2	MAE	RMSE
Dummy Regressor	0.00	-0.00	24957.14	157.98
Linear Regression	0.57	0.33	16733.10	129.36
Lasso	0.54	0.42	14543.17	120.60
Random Forest	0.94	0.42	14450.80	120.21
Ridge	0.55	0.44	14008.92	118.36
XGBoost	0.91	0.48	12941.24	113.76





Modélisation, recherche du meilleur estimateur



- <u>Meilleur estimateur</u>

  XGBRegressor

  Hyper paramètres :

  Nb d'estimateurs = 200

  Learning rate = 0.1

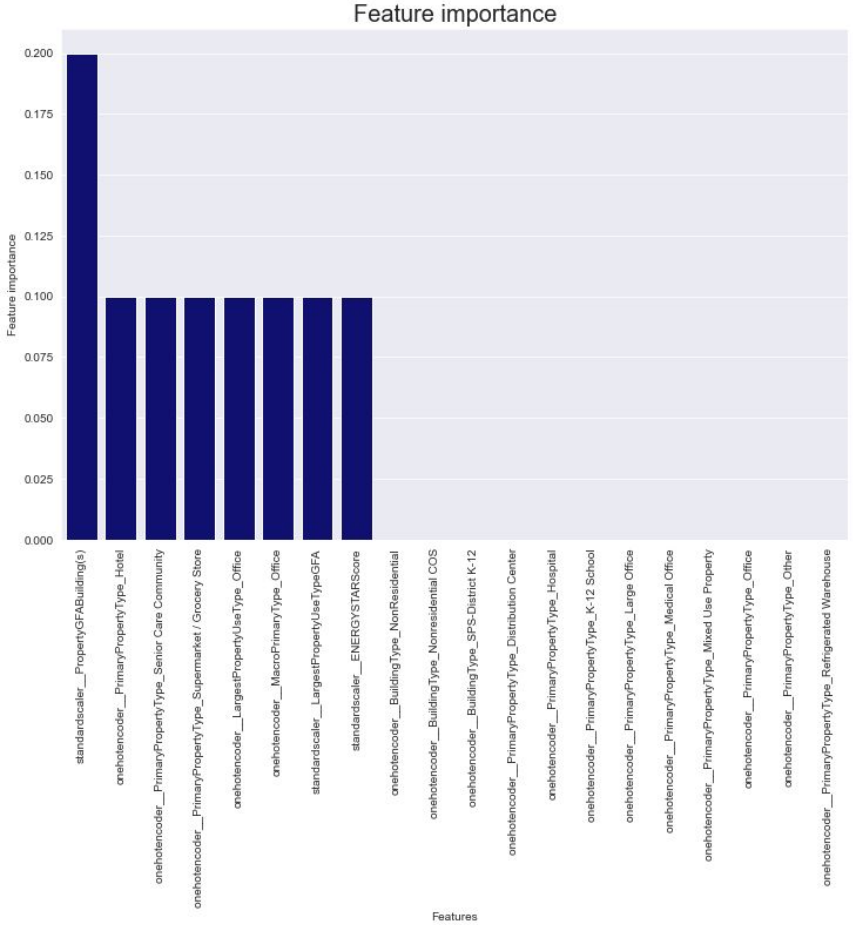
  Max depth = 3

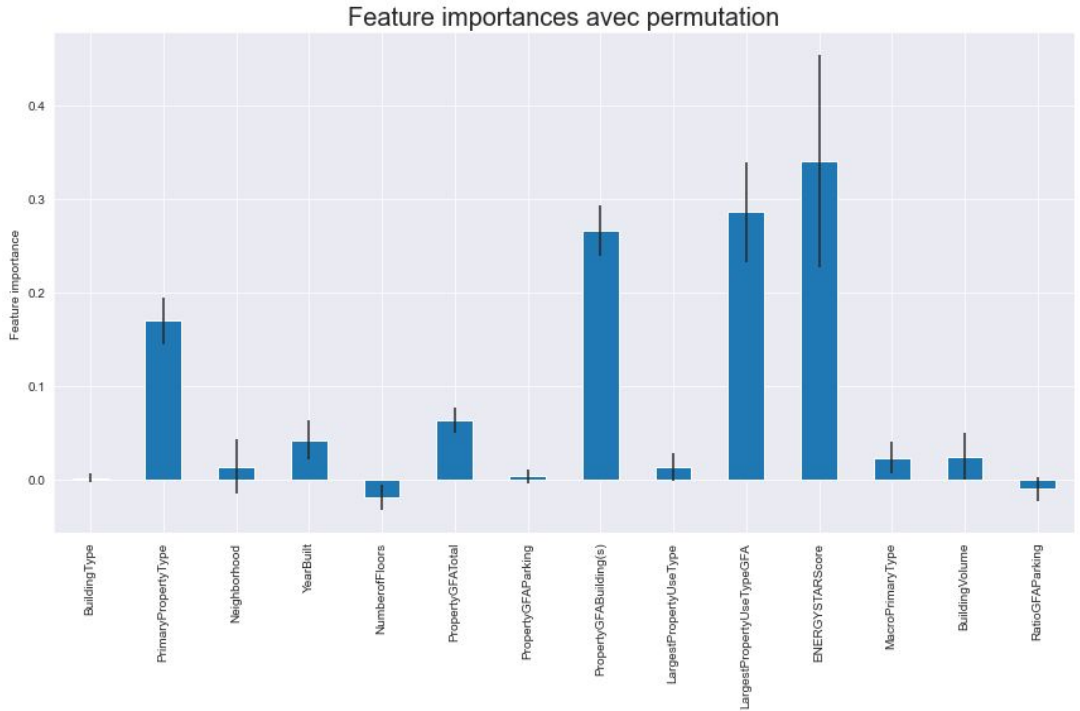
  Régression alpha = 1.2

  Scores :

  R^2 train = 0.95

  R^2 test = 0.48





Suppression des variables à 0 ou négatives :

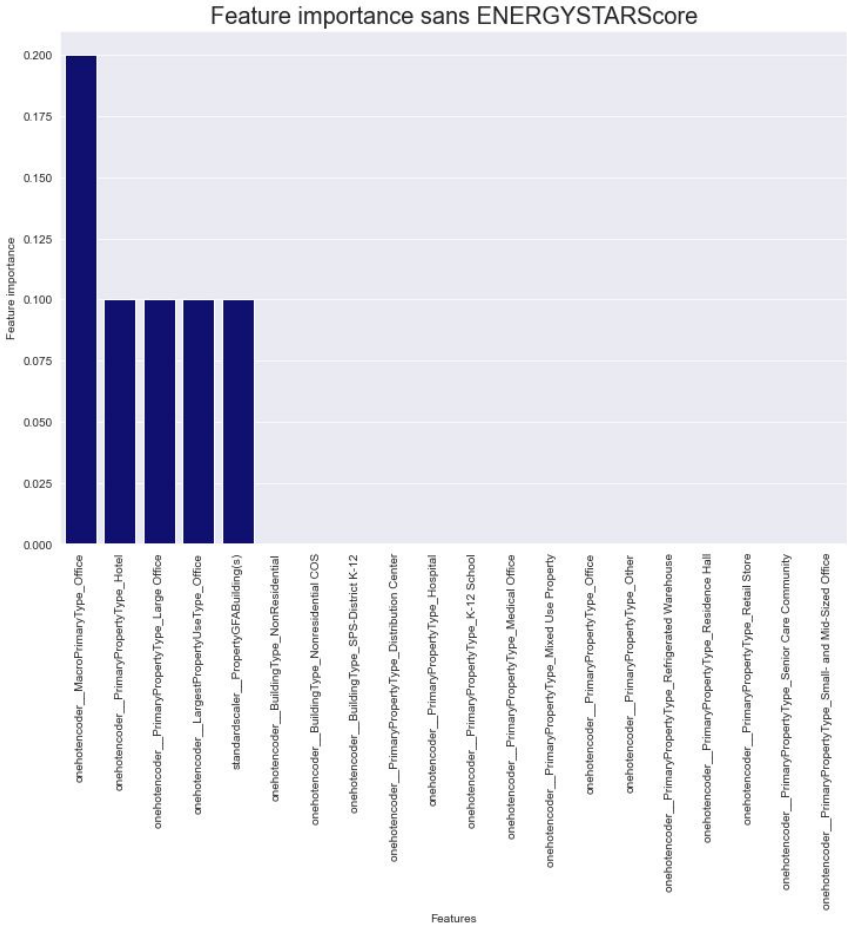
R^2 train = 0.92
R^2 test = 0.33 (perte de 15%) !

- <u>Sans ENERGYSTARScore</u>

  Scores :

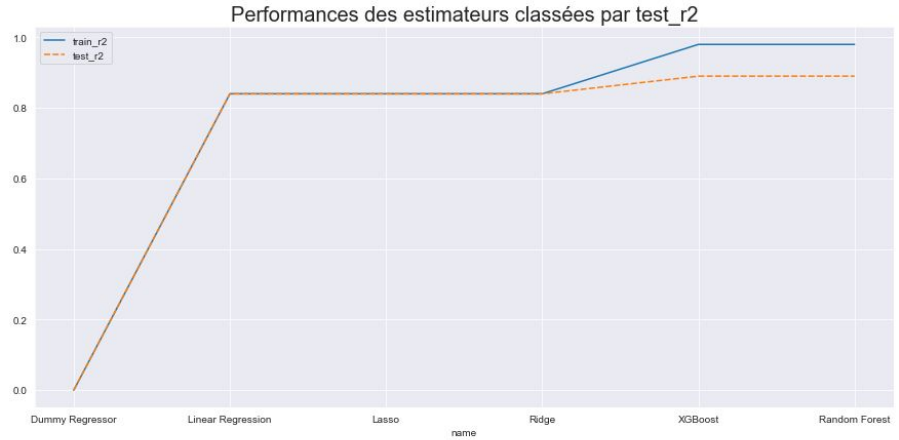
  R^2 train = 0.92

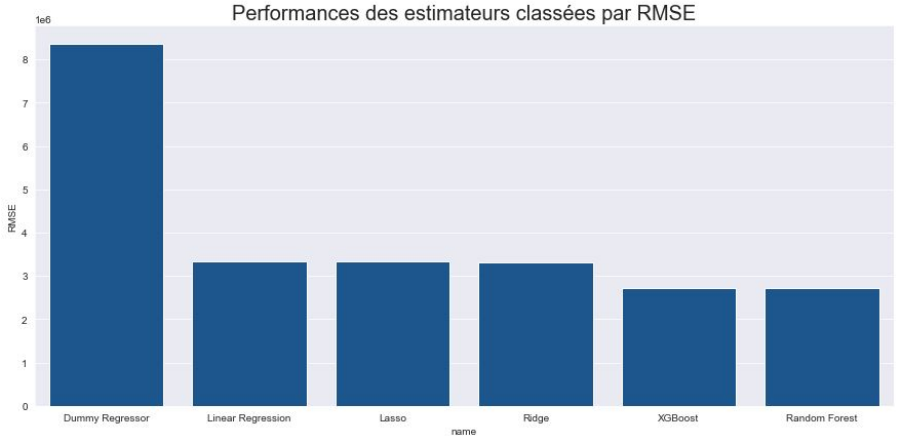
  R^2 test = 0.37 (11% de perte) !



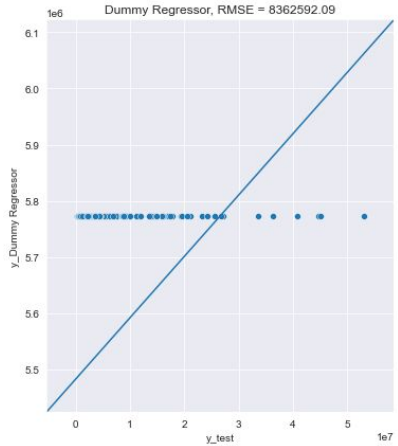
‘SiteEnergyUse(kBtu)’

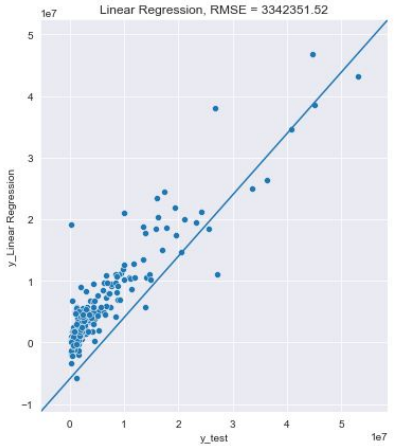
name	train_r2	test_r2	MAE	RMSE
Dummy Regressor	0.00	-0.00	6.993295e+13	8362592.09
Linear Regression	0.84	0.84	1.117131e+13	3342351.52
Lasso	0.84	0.84	1.105409e+13	3324768.61
Ridge	0.84	0.84	1.097932e+13	3313506.36
XGBoost	0.98	0.89	7.405097e+12	2721230.76
Random Forest	0.98	0.89	7.367804e+12	2714369.84

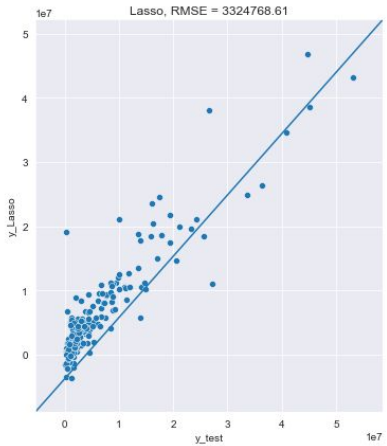


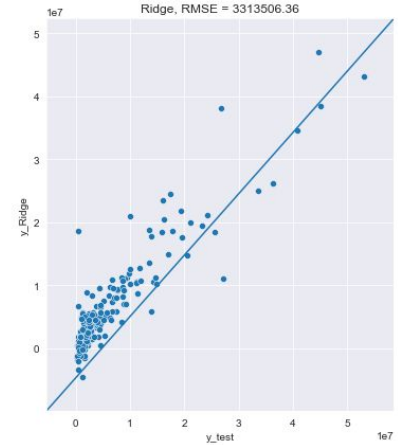


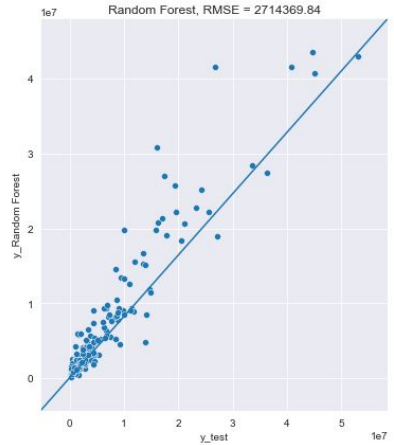
Modélisation, recherche du meilleur estimateur

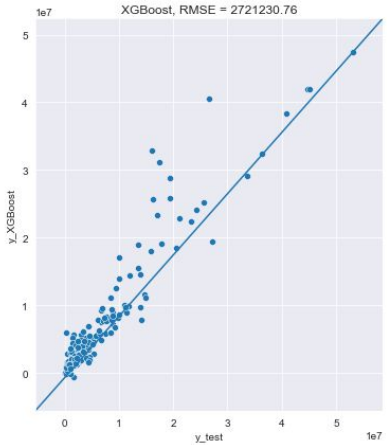












- **Meilleur estimateur**

  RandomForestRegressor

  Hyper paramètres :

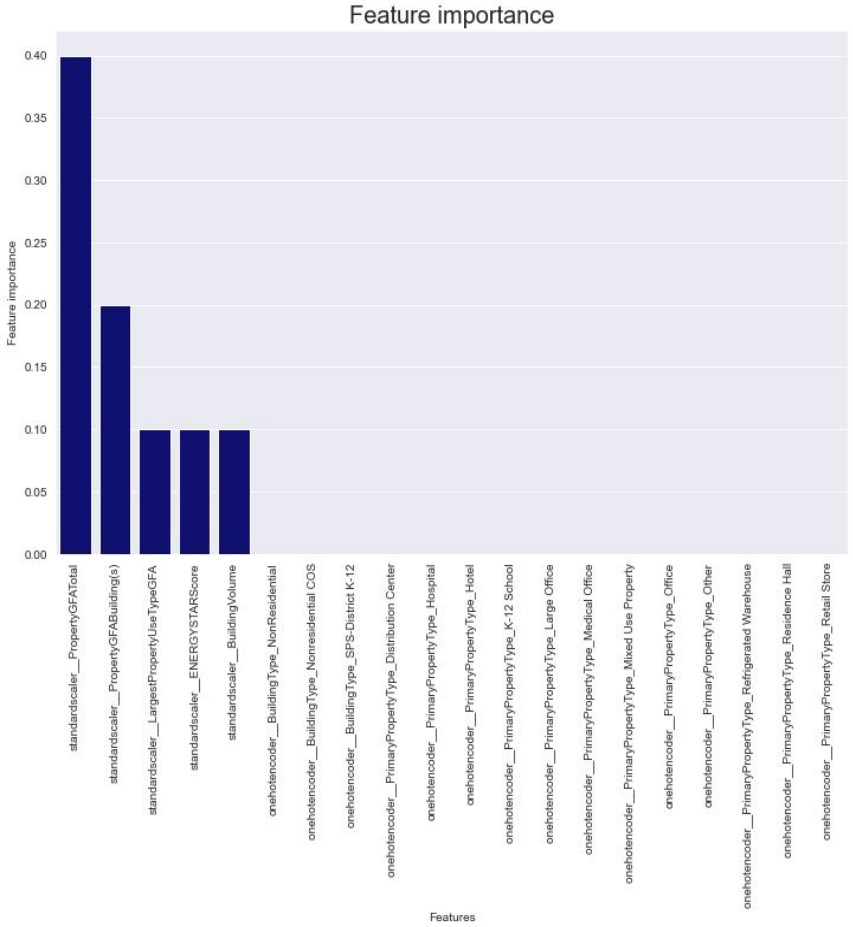
  Nb d'estimateurs = 200

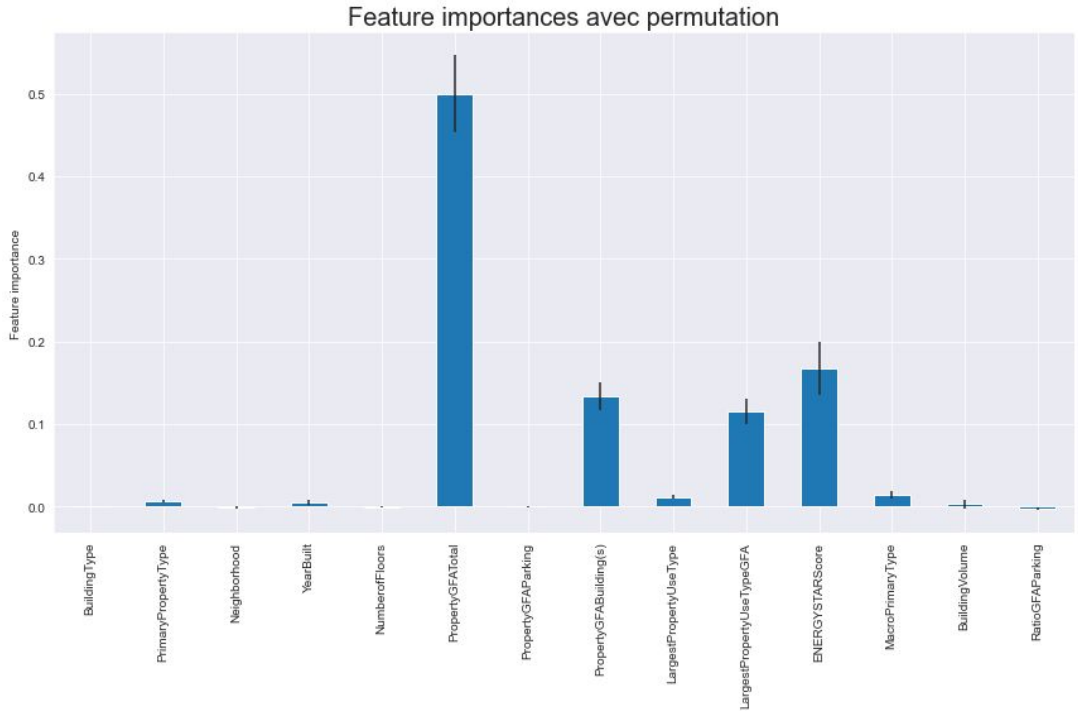
  Max depth = 22

  Scores :

  R^2 train = 0.98

  R^2 test = 0.89



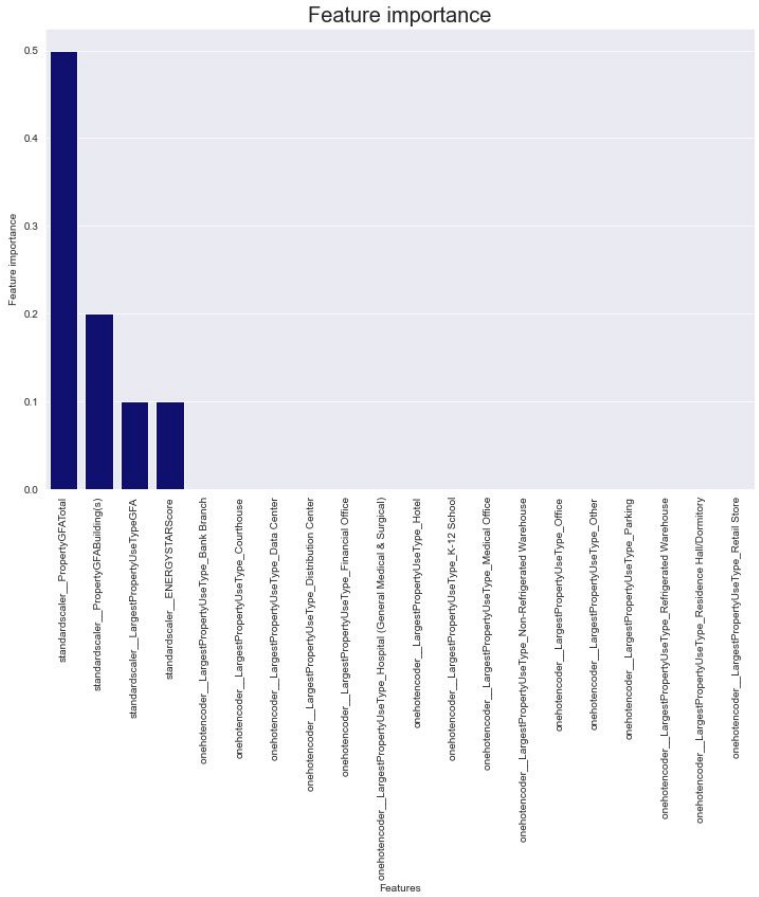


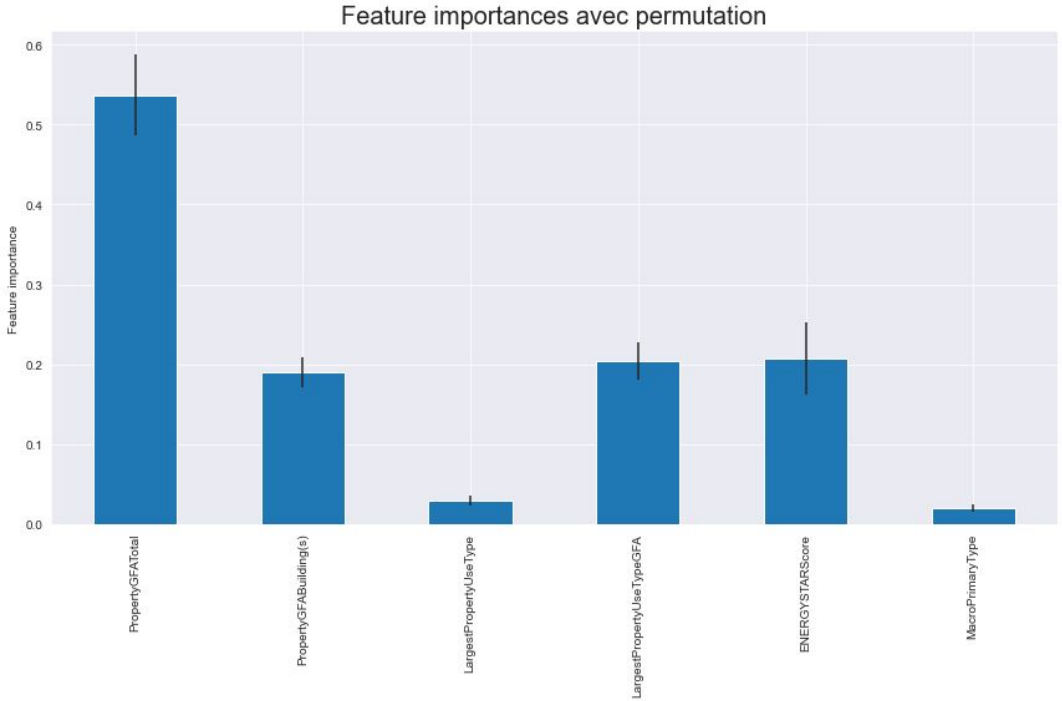
Suppression des variables à 0 ou négatives :

R^2 train = 0.98
R^2 test = 0.90

Gain de précision !

Modélisation, recherche du meilleur estimateur





- **‘TotalGHGEmissions’**

  Variables importantes :

  PropertyGFABuilding(s), PrimaryPropertyType (Hotel, Senior Care Community et Super Market/Grocery store), LargestPropertyUseType (Office), MacroPrimaryType (Office), LargestPropertyUseTypeGFA et ENERGYSTARScore

  ENERGYSTARSCore est une information importante : si on ne la prend pas en compte, **perte de 11%** de précision

- **‘SiteEnergyUse(kBtu)’**

  Variables importantes :
  PropertyGFATotal, PropertyGFABuilding(s), LargestPropertyUseTypeGFA, ENERGYSTARScore et Building Volume

  Après suppression des variables “sans importance”, **gain de 1%** de précision
  Variables importantes :
  PropertyGFATotal, PropertyGFABuilding(s), LargestPropertyUseTypeGFA, ENERGYSTARScore

- **Feature engineering :** créer de nouvelles variables
- **Sélection des variables :** affiner/faire un autre choix de variables
- **Encoder, Scaler :** utiliser d'autres encoder et/ou scaler
- **Hyper paramètres :** affiner davantage les hyper paramètres