Ville de Seattle – Consommation et émissions des bâtiments

- Prédiction des émissions de CO2 et de consommation totale d’énergie

Préambule

- Ville de Seattle
  - Objectif neutralité carbone 2050
  - Souhaite comprendre les consommations et émissions de CO2 de ses bâtiments
  - A fait des relevés dans différents bâtiments de la ville
- Objectif
  - Prédire les consommations et émissions de CO2
  - En utilisant les données structurelles
  - Évaluer l'intérêt de l'ENERGYSTARScore dans la modélisation

Sommaire

Présentation des données

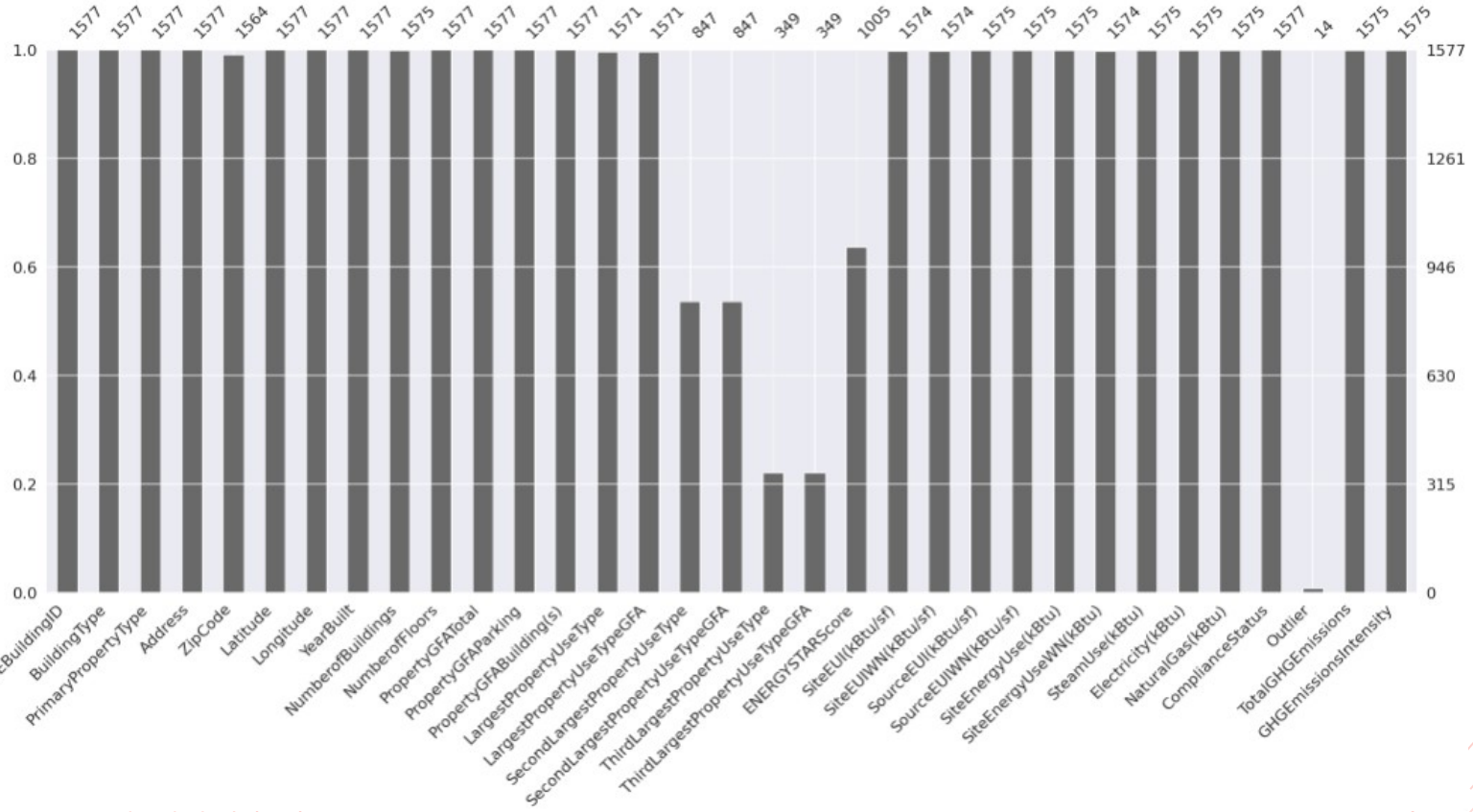
- Premier passage en revue du jeu de données
  - 3380 bâtiments environ
  - 45 caractéristiques
    - Localisation
    - Nombre de bâtiments et étages
    - Surfaces et utilisations
    - Consommations et émissions de CO2
  - Peu de valeurs manquantes
  - Pas de doublons

Extrait du tableau de remplissage des variables

Exploration des données

- Type de bâtiments
  - 1665 bâtiments non destinés à l'habitation
- Simplification des features
  - DataYear
  - City
  - State
  - Neighborhood
  - CouncilDistrictCode
  - Comments
  - PropertyName
  - TaxParcelIdentificationNumber
  - Différentes unités
  - YearsENERGYSTARCertified

Valeurs manquantes



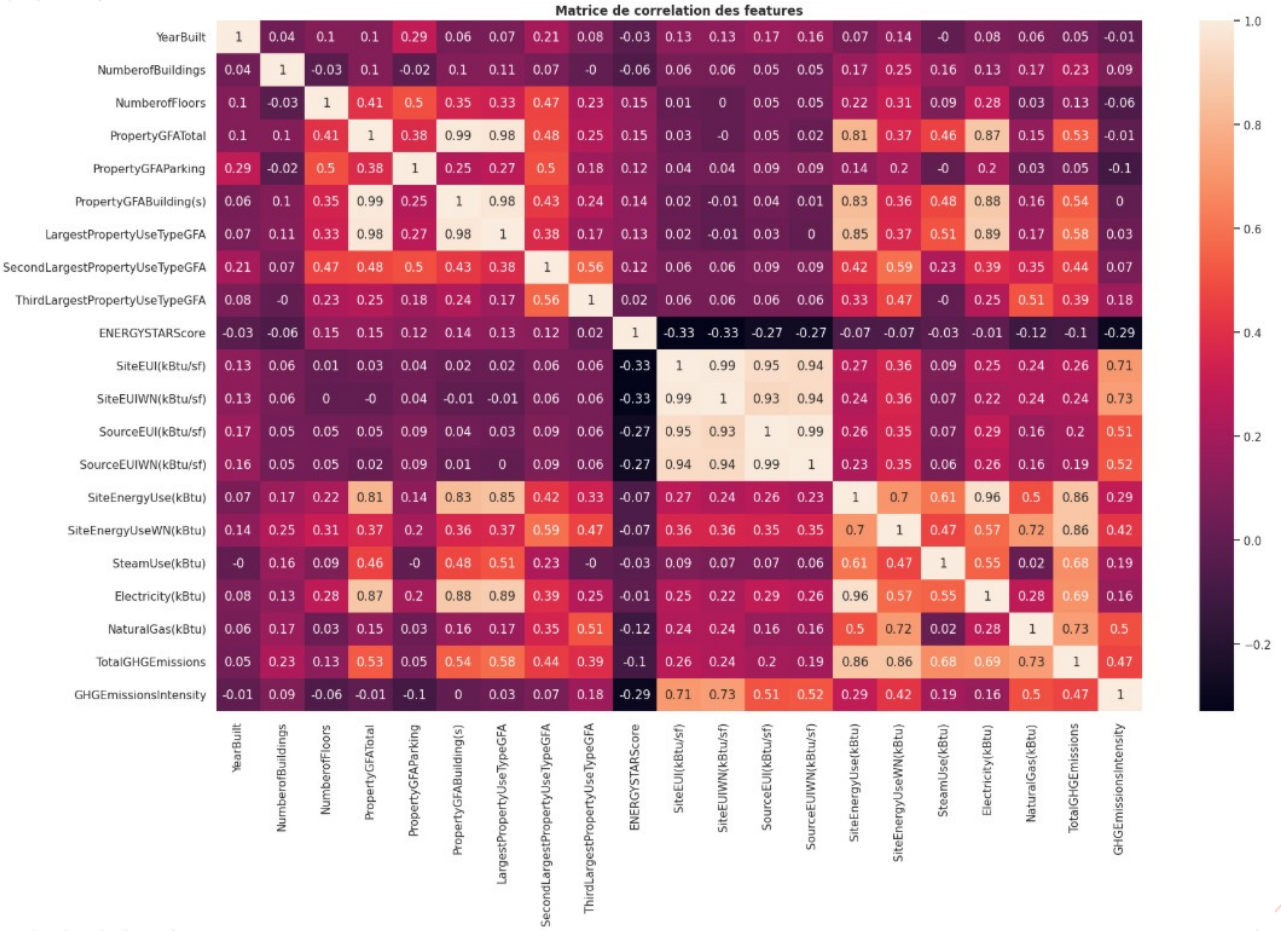
Nombre de valeurs manquantes par feature

Valeurs manquantes

- Renseignement des valeurs lorsque possible
  - Google Street View
  - Recherche adresse
  - …
- Données énergétiques
  - Suppression des bâtiments
- Surfaces et utilisations
  - Largest => Suppression
  - Second & Third =>
    - Surface = 0
    - Type = NotUsed

Variables quantitatives

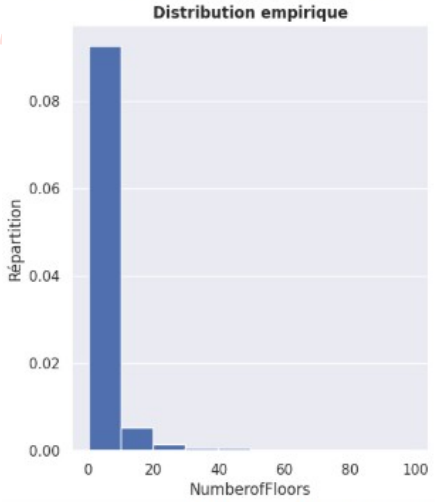
- Analyse des corrélations entre features
  - Suppression des variables peu informatives

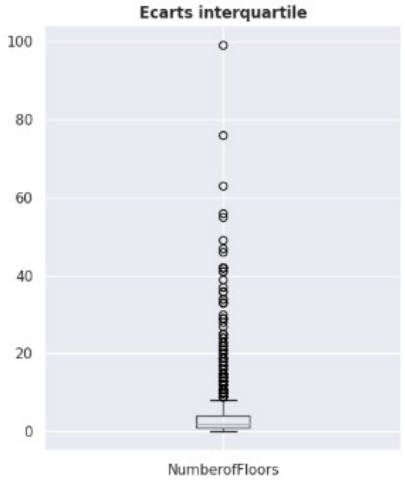


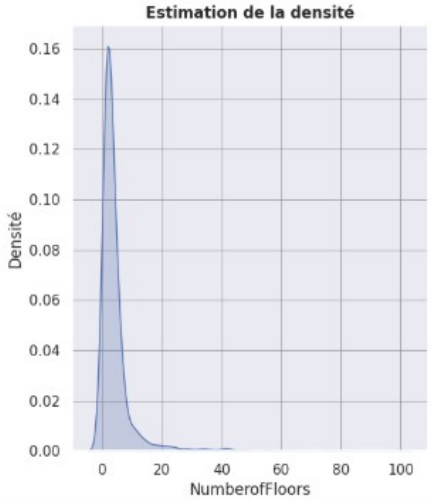
Variables quantitatives

- Analyse statistiques
  - Correction des valeurs aberrantes

```
------------------------------------------------
     Analyse de la feature NumberofFloors
------------------------------------------------
```








```
              Valeurs des statistiques
Min :                         0.000000
Max :                        99.000000
Mode :                        1.000000
Moyenne :                     4.275795
Médiane :                     2.000000
Variance :                   46.167394
Ecart type :                  6.794659
Skewness :                    5.832520
Kurtosis :                   49.018322
```

Variables quantitatives

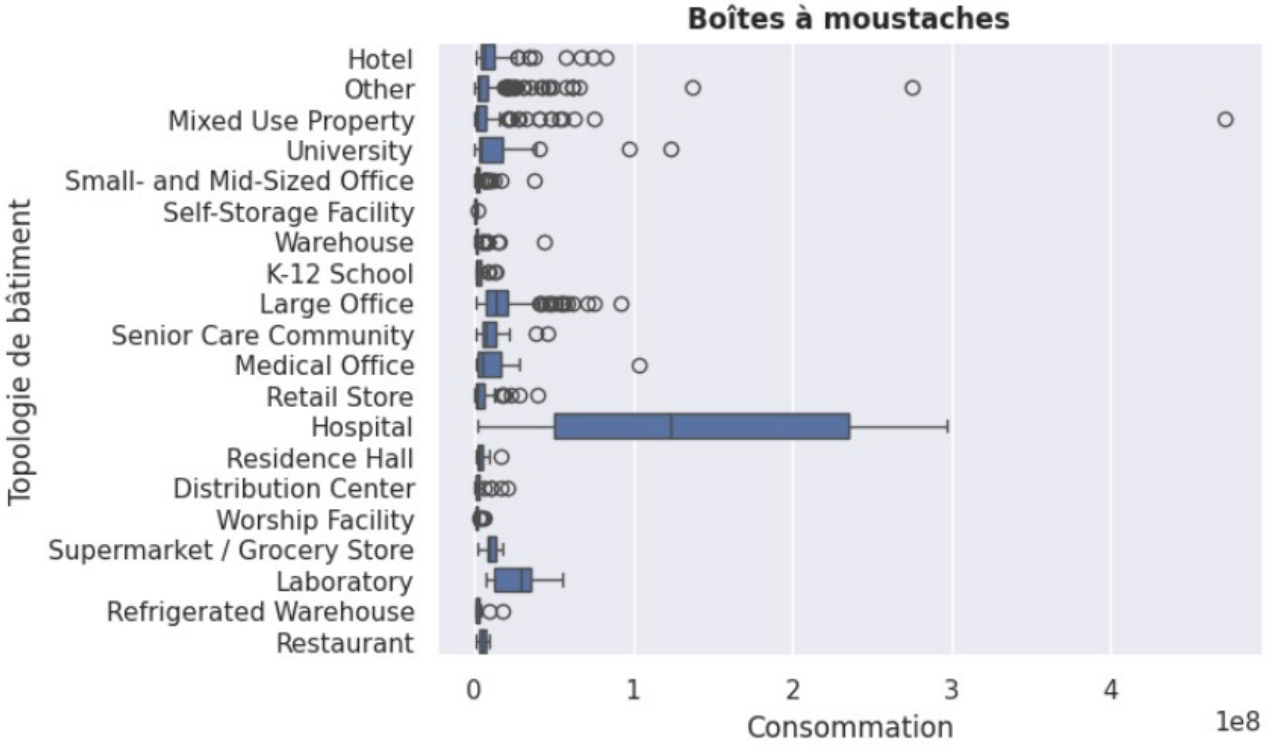
- Création de nouvelles features
  - Répartition des énergies en pourcentage de la consommation totale
  - Pourcentage d'utilisation du bâtiment pour :
    - Bâtiment hors parking
    - Parking
    - Plus grande utilisation
    - Deuxième plus grande utilisation
    - Troisième plus grande utilisation

Variables qualitatives - Localisation

- Carte des bâtiments pour apprécier les variables liées à l'adresse

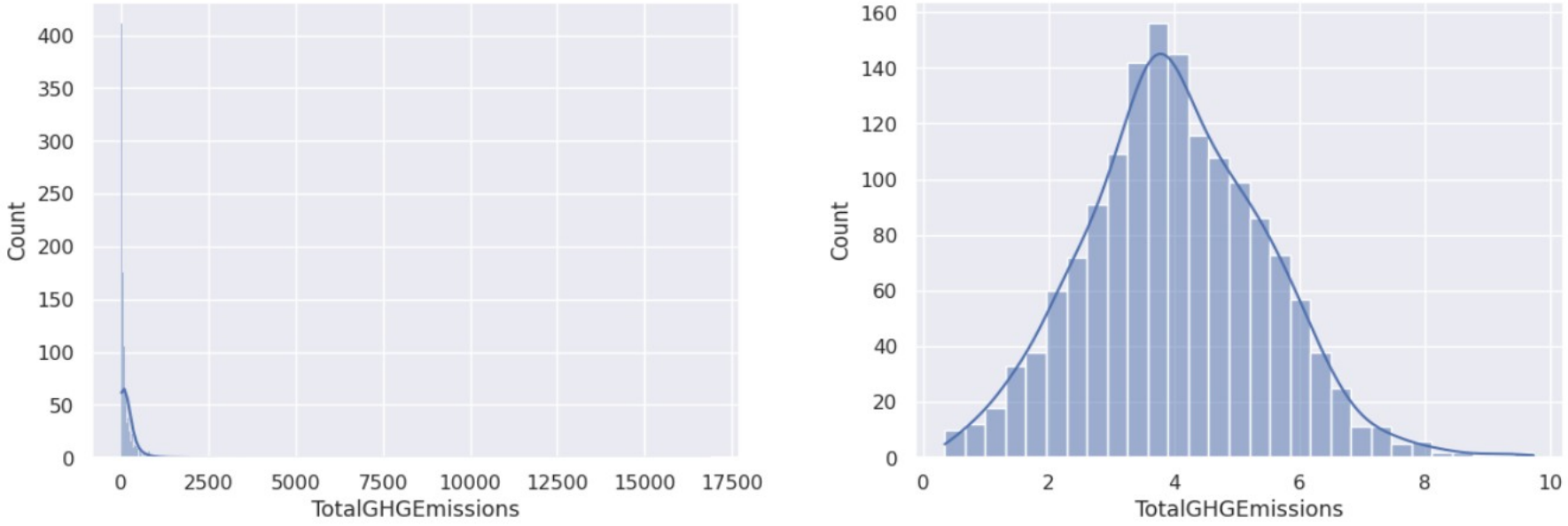
Variables qualitatives – Type de bâtiment

- Analyse ANOVA pour étudier l'impact de la feature
- Différents types de traitement
  - OneHotEncoder si peu de catégories
  - TargetEncoder sinon (pipeline)



Normalisation

- Passage au log
  - Plusieurs variables et la cible ont des distributions très étalées



Distribution avant et après passage au log

Normalisation

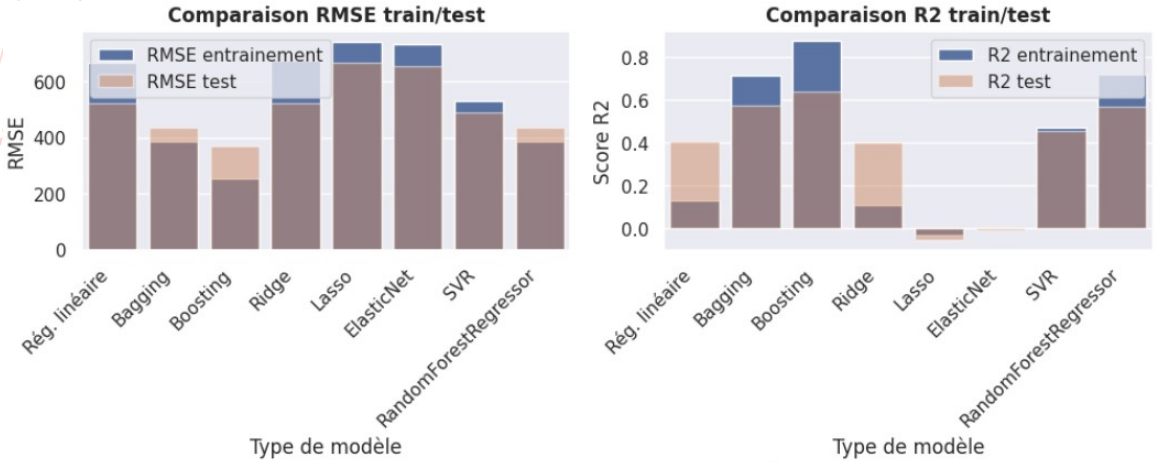
- Préprocesseur
  - Normalisation selon le type de données
    - Quantitative = StandardScaler
    - Qualitative = TargetEncoder
    - Booléenne = MinMaxScaler
  - Sera appliqué dans un pipeline

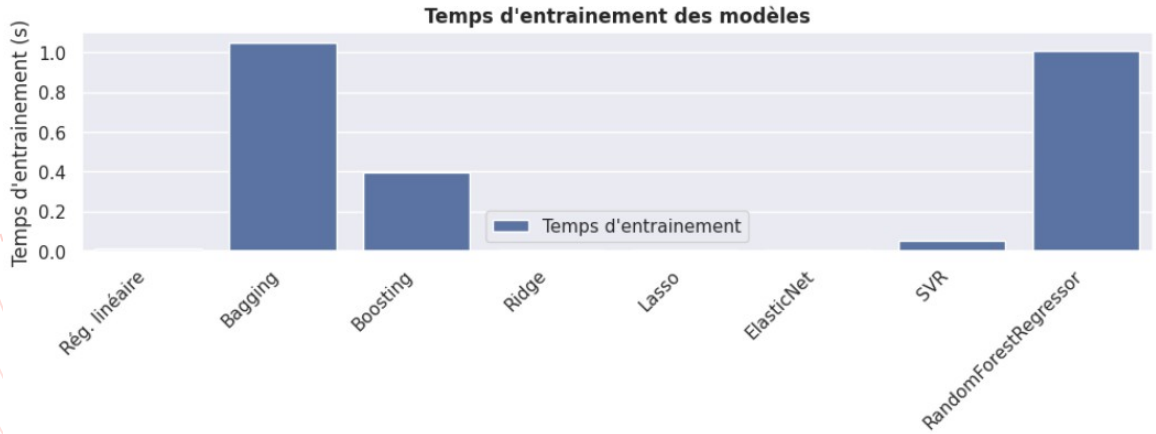
Métriques d'évaluation

- RMSE
  - Grandeur de l'erreur dans l'unité de la cible
  - Les valeurs prédites seront passées à l'exponentiel avant calcul du score
- Coefficient de détermination R2
  - Performance du modèle
  - Les valeurs prédites seront passées à l'exponentiel avant calcul du score
- Temps d'entrainement

Scores avant optimisation

- Sélection algorithmes
  - SVR
  - ElasticNet

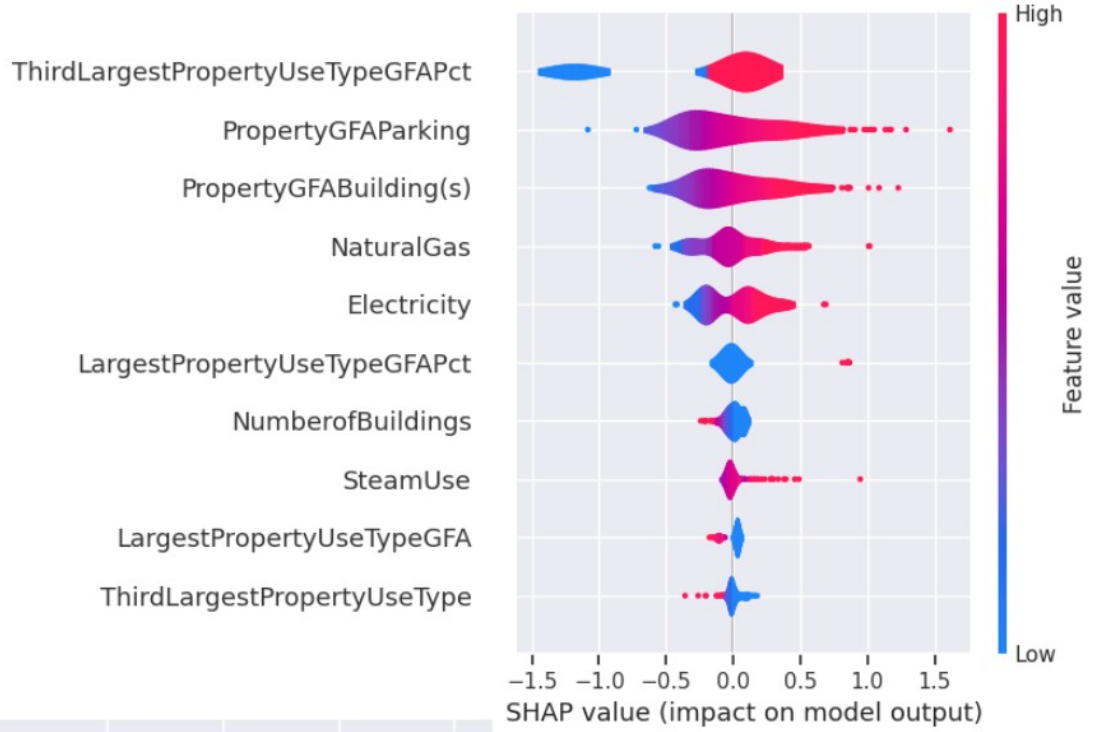


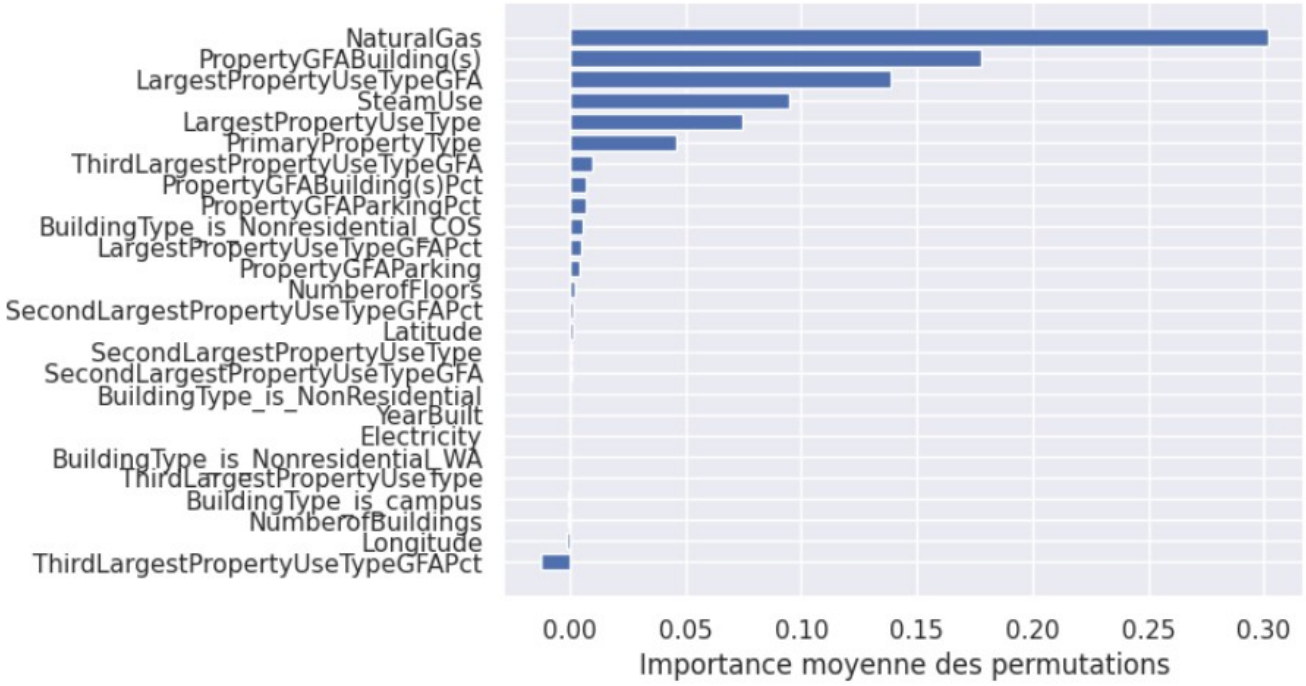


Scores après optimisation

- Scores GridSearchCV :
  - RMSE moyenne entraînement : 327,27
  - R2 moyen entraînement : 0,797
- Scores set de test
  - RMSE : 322,26
  - R2 : 0,706

Feature importance





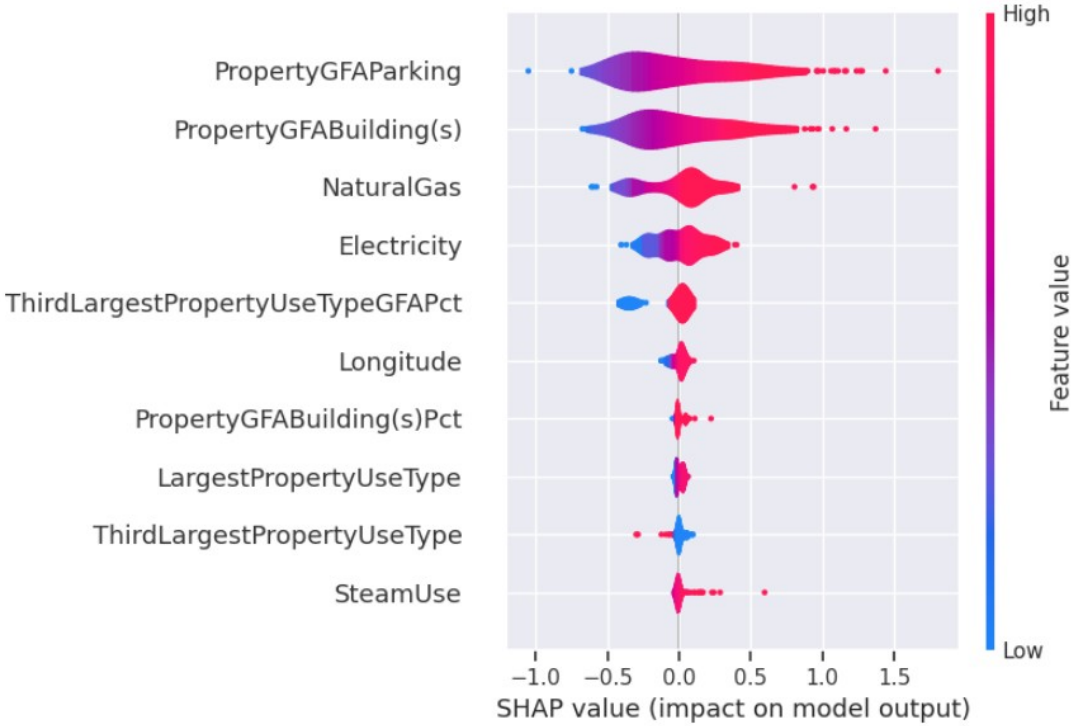
Scores avec ENERGYSTAR Score

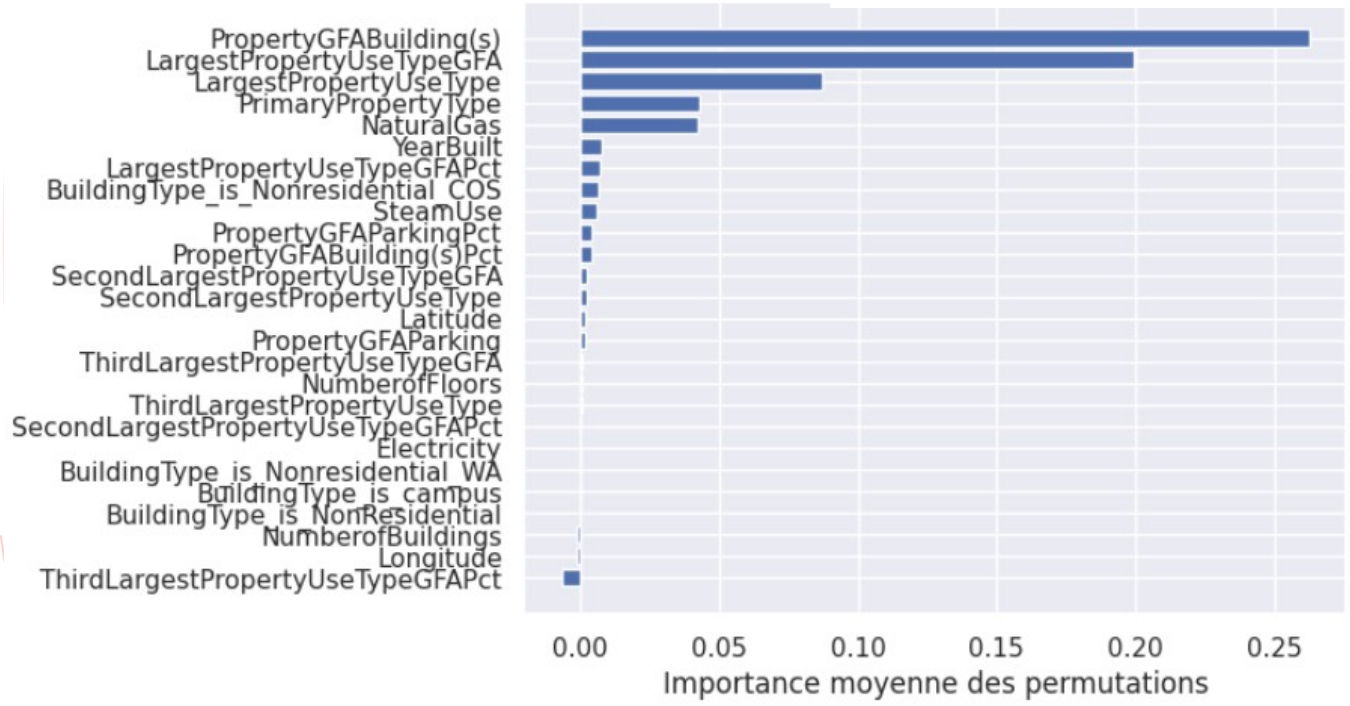
- Scores GridSearchCV :
  - RMSE moyenne entraînement : 340,95
  - R2 moyen entraînement : 0,842
- Scores set de test
  - RMSE : 286,17
  - R2 : 0,743

Scores après optimisation - Consommation

- Scores GridSearchCV :
  - RMSE moyenne entraînement : 9,8M
  - R2 moyen entraînement : 0,821
- Scores set de test
  - RMSE : 8,4M
  - R2 : 0,817

Feature importance - Consommation





Scores avec ENERGYSTAR Score - Consommation

- Scores GridSearchCV :
  - RMSE moyenne entraînement : 10,4M
  - R2 moyen entraînement : 0,839
- Scores set de test
  - RMSE : 10,7M
  - R2 : 0,862