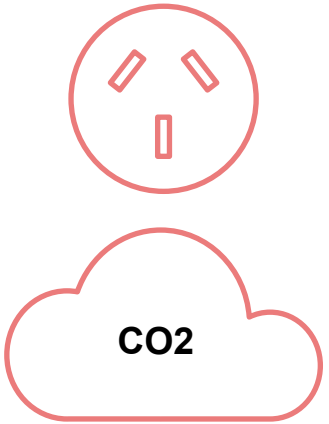
Anticipez les besoins en consommation de bâtiments

A. Monod - Parcours ML - p3

Contexte

Seattle



SOMMAIRE

1. Nettoyage

Suppression variables inutiles

Feature Engineering

2. Modélisation

Modèles

Résultats

1. Nettoyage

1 - Suppression variables et lignes

2 - Feature Engineering

1- 1 Suppression variables inutiles

Sélection des variables cibles et suppression des variables proches :

- SiteEnergyUse(kBtu) (pas normalisé) —> SiteEnergyUseWN(kBtu)
- GHGEmissionsIntensity (information partielle)—> TotalGHGEmissions

Variables inutiles :

- YearsENERGYStarCertified
- Comments
- DataYear

Variables ambiguës et peu discriminantes :

- Outliers
- CompliantState
- DefaultData

Variables de consommation :

- SiteEUI(kBtu/sf)
- SteamUse(kBtu)
- …

Variables géographiques :

- Address
- City
- State
- ZipCode
- TaxParcelIdentificationNumber

1- 2 Suppression lignes inutiles

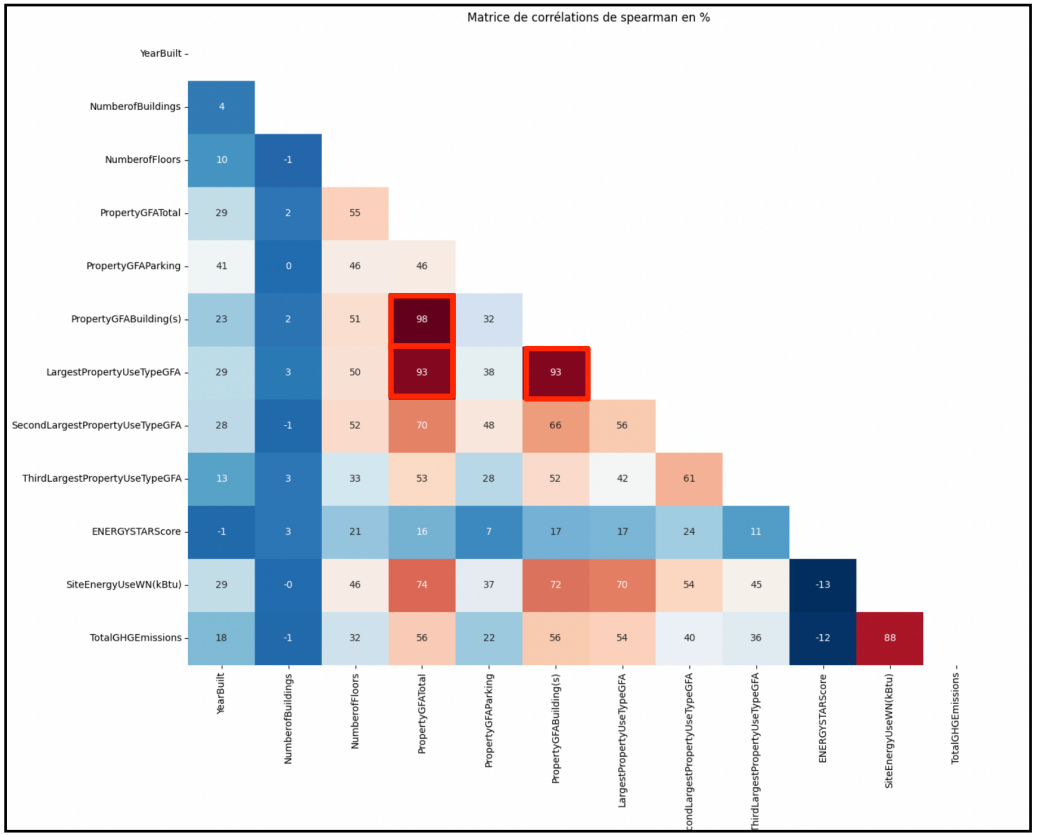
Valeurs inutiles

- variables cibles à 0
- bâtiments non destinés à l'habitation (BuildingType)

Valeurs aberrantes

- Surface de l'activité principale > Surface totale
- Nombre d'étages aberrant

1- 3 Suppression variables corrélées



- PropertyGFATotal
- PropertyGFABuilding(s)
- LargestPropertyUseTypeGFA

2- 1 - Création de variables

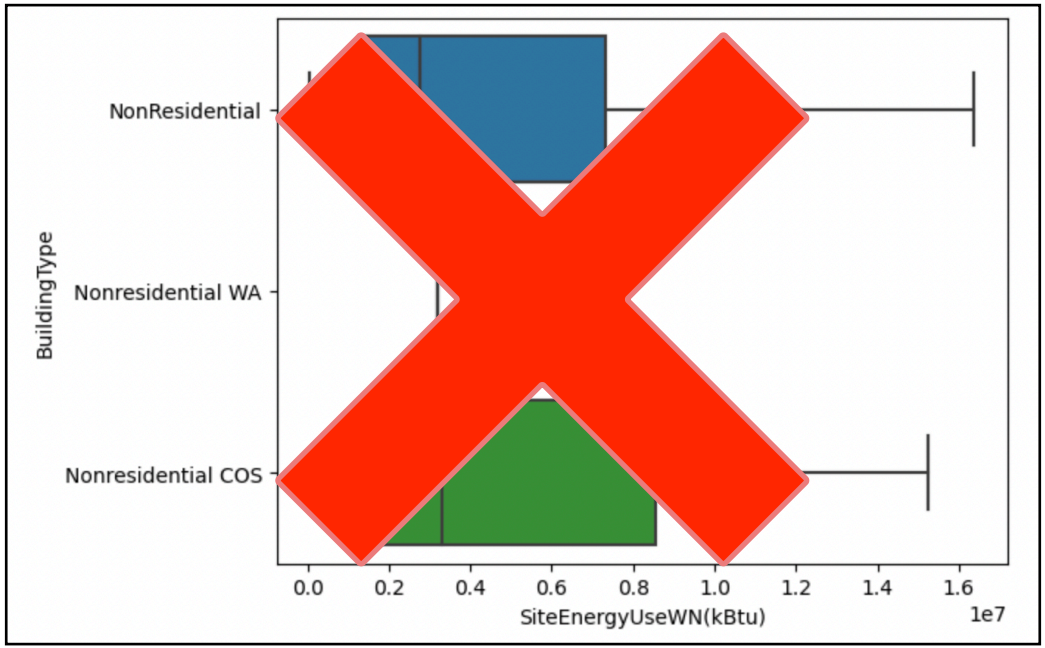
Haversine
- Distance entre la position du bâtiment et le centre de Seattle

Âge des bâtiments
- Transformation de la variable YearBuilt

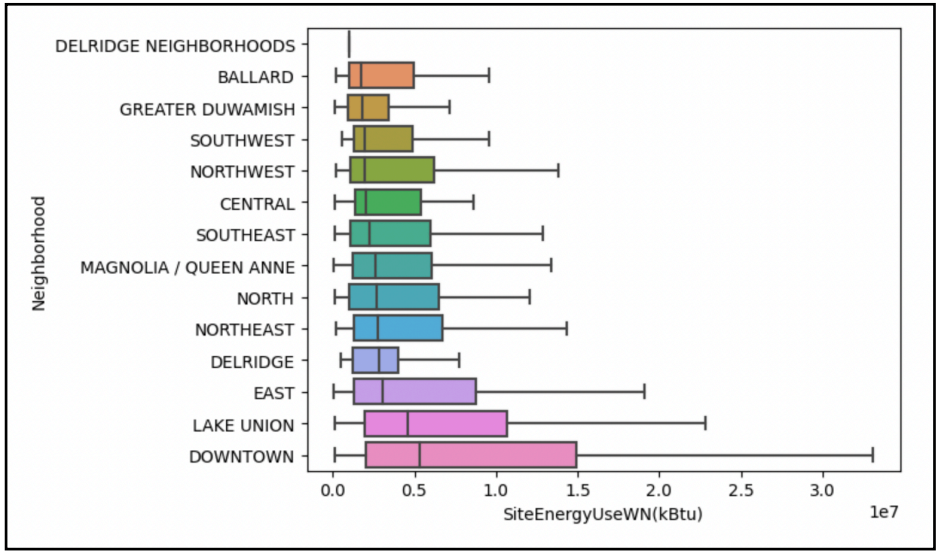
Nb d'usages
- Compilation du nb de PropertyTypes par bâtiment

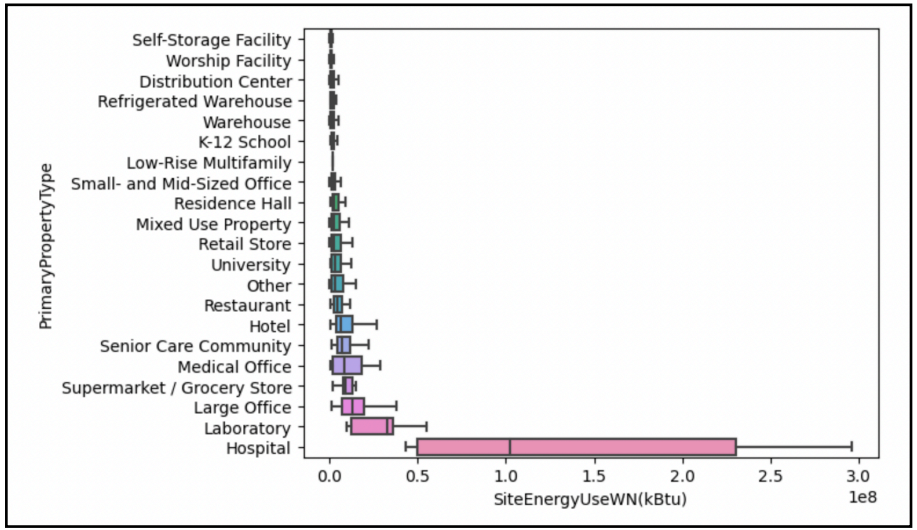
Nb de bâtiments :
- Transformation de la variable NumberofBuildings

2- 2 - Variables catégorielles

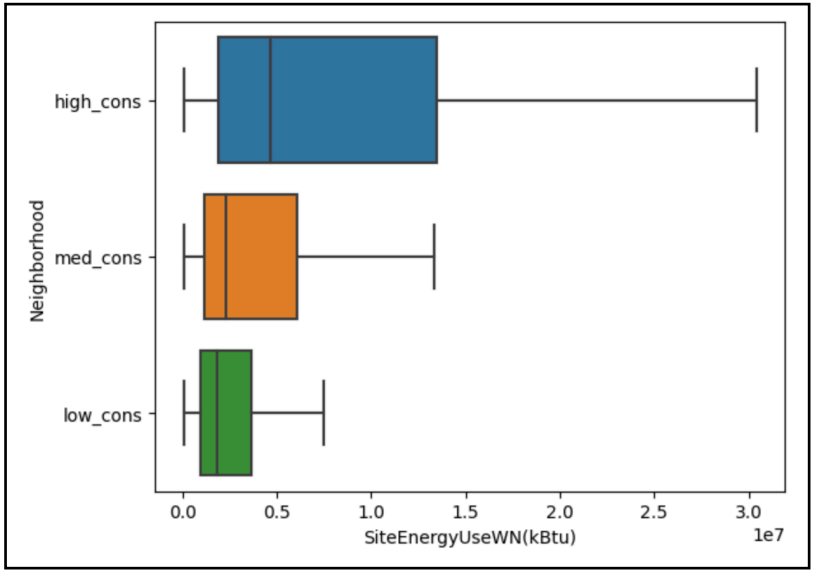


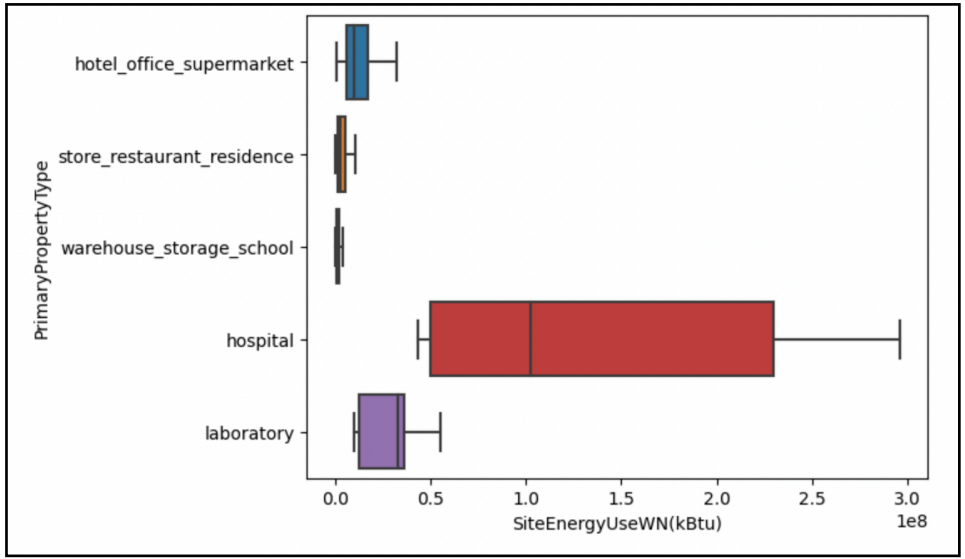
2- 2 - Variables catégorielles





2- 2 - Variables catégorielles





2. Modélisation

1 - Modélisation

2 - Analyse des résultats

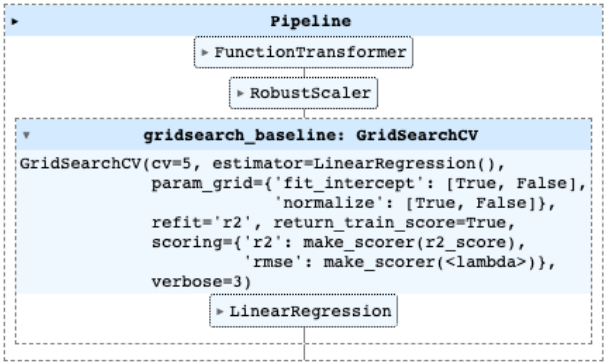
1- 1 Preprocessing

- Copie du dataframe et suppression de la variable ESS dans le premier df

- Split df train/df test 80/20

Construction GridSearch

- FunctionTransformer : passage au log
- RobustScaler
- cv = 5
- Scoring : R2 et RMSE



1- 2 Modélisation

Modèles linéaires

- Baseline : Régression Linéaire
- ElasticNet (Régression pénalisée)
- KNN

Modèles non linéaires

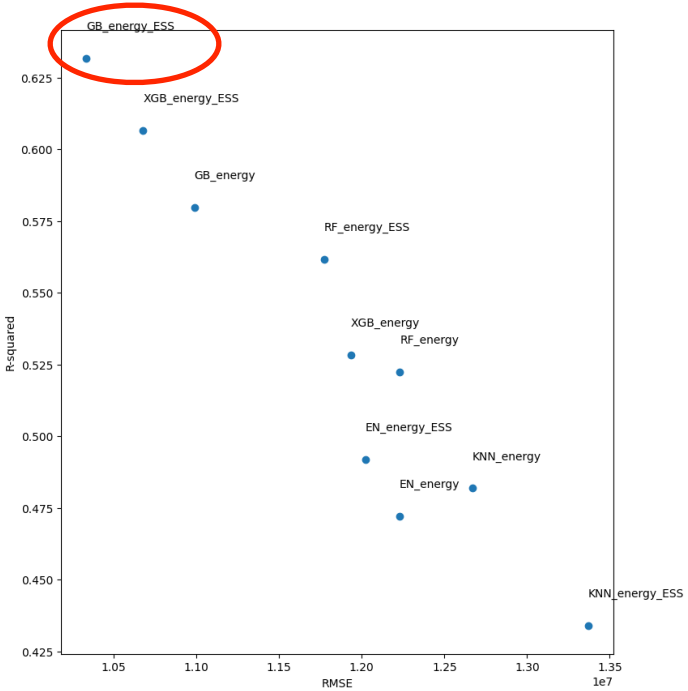
Ensemblistes aléatoires

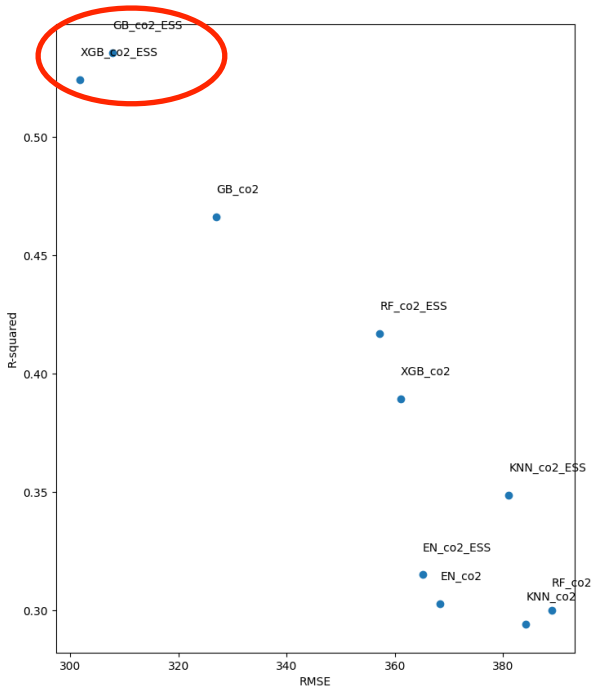
- RandomForrest

Ensemblistes adaptatifs

- GradientBoosting
- XGBoost

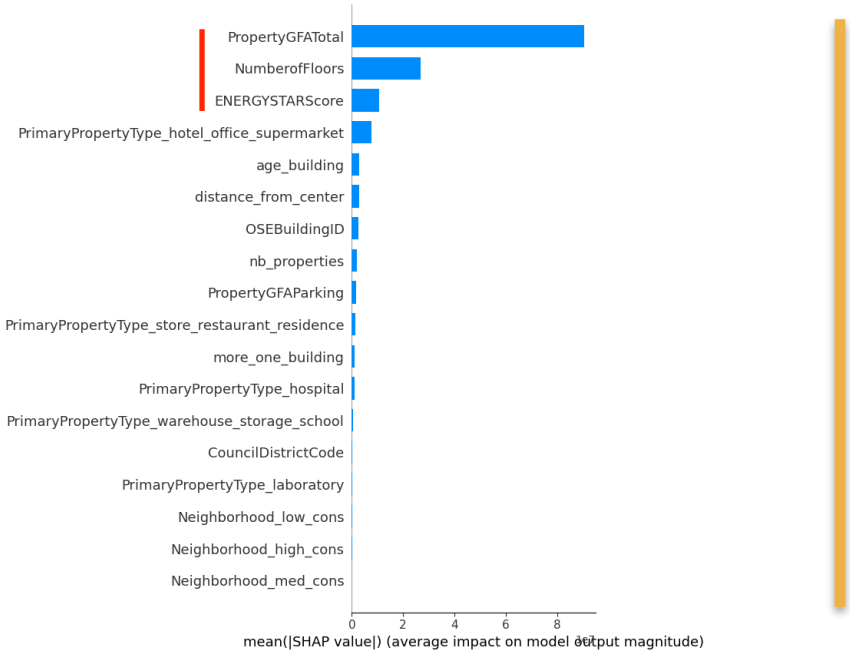
2- 1 Meilleur modèle

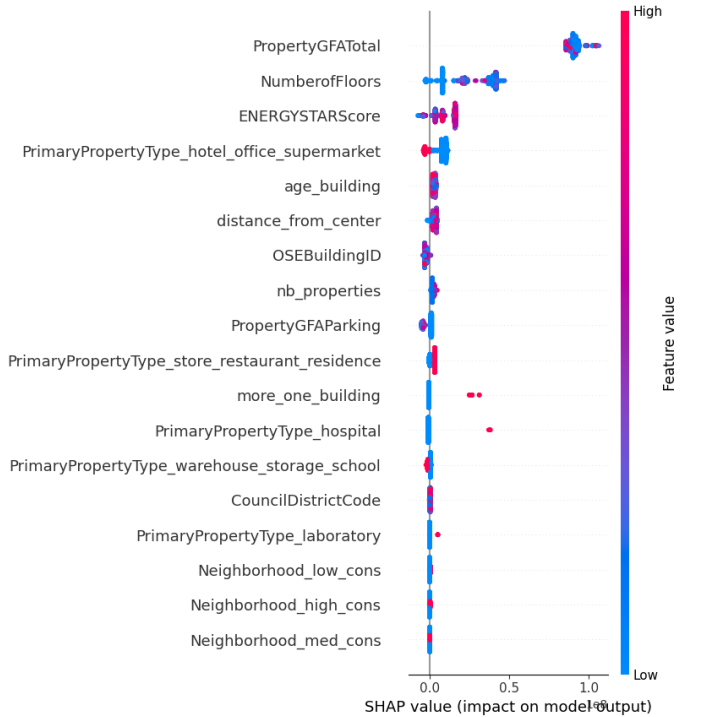




2- 2 Analyse des résultats

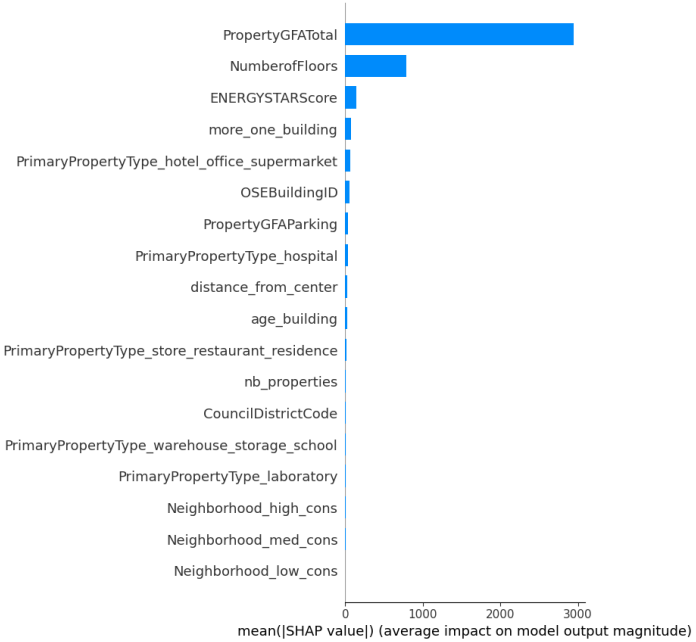
GB Energy ESS





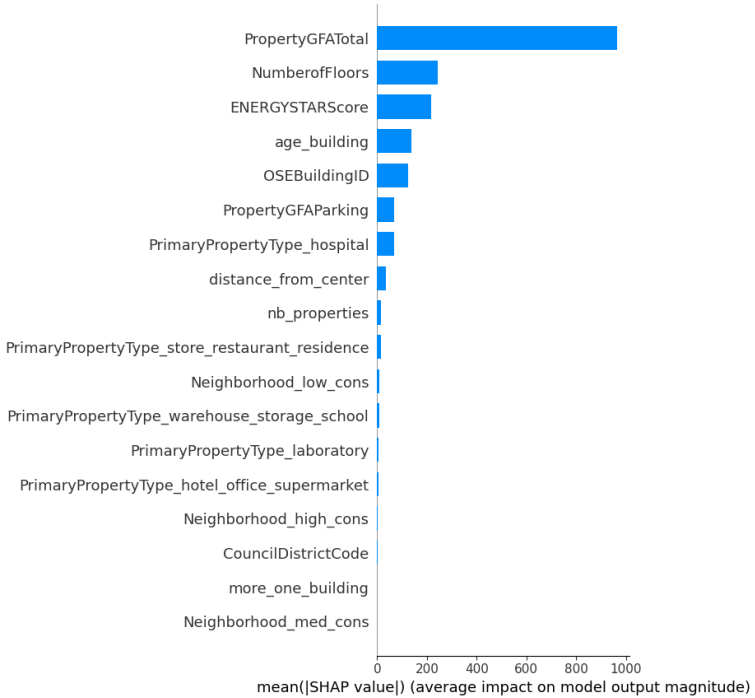
2- 2 Analyse des résultats

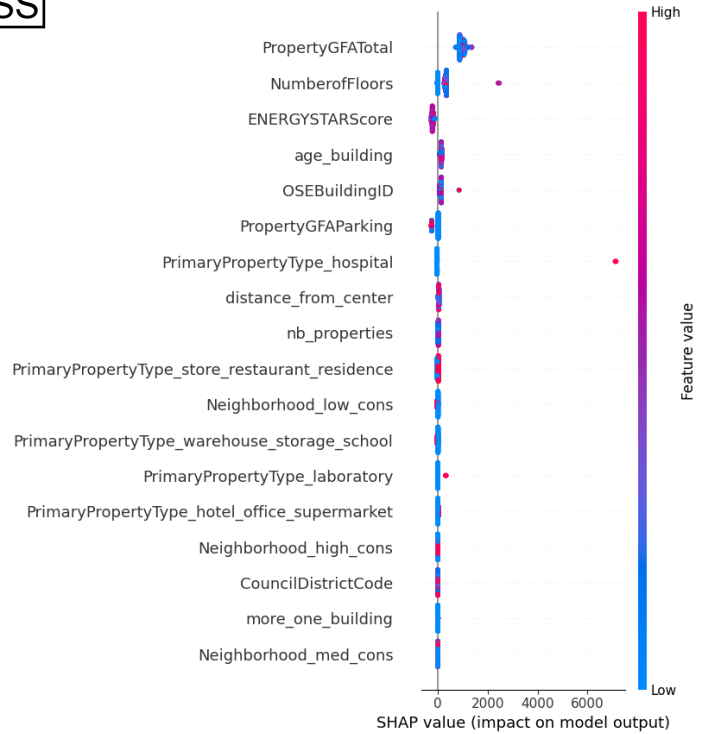
GB CO2 ESS



2- 2 Analyse des résultats

XGB CO2 ESS





CONCLUSION

Conclusions

- Modèle le plus performant : GB_Energy_ESS
- Il évalue à 0,63% la consommation en énergie des bâtiments
- Importance première de la variable PropertyGFA
- Importance secondaire de la variable ESS

Axes d'amélioration

Intégration variables de consommation (variables brutes ou ratio)