Anticipez les besoins en consommation de bâtiments

Entreprise : RES Group

Ecole : OPENCLASSROOM

Divine Tulomba

Année : 2022/2023

SOMMAIRE

- Présentation de la problématique
- Préparation du jeu de données
- Pistes de modélisations
- Présentation des résultats

Rappel de la problématique

Interprétation

Pistes de recherche envisagées

Problématique

Présentation de la problématique

- Données de consommation disponibles pour les bâtiments de la ville de Seattle pour l'année 2016
- Coût important d'obtention des relevés / Fastidieuses à collecter

La mission :

- Prédire les émissions de CO_2 et la consommation totale d'énergie sans les relevés annuels
- Evaluer l'intérêt de l'ENERGY STAR Score
- Mettre en place un modèle de prédiction réutilisable

Présentation de la problématique

Prévision :

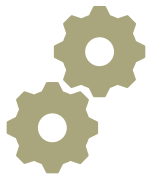
- **Features :** Caractéristiques intrinsèques des bâtiments (hors consommations)

- **Données à prédire à partir de deux modèles différents:**
  - Consommation totale des bâtiments SiteEnergyUseWN(kBtu)
  - Emissions totales des bâtiments TotalGHEmissions

- **Energy Star Score :**
  - Comparaison de son intérêt en essayant de modéliser avec et sans pour le meilleur modèle simplement

II - Préparation du jeu de données

Nettoyage de données



Feature Engineering

Exploration de données

Etude de la dataframe

Etude Générale de notre dataframe :

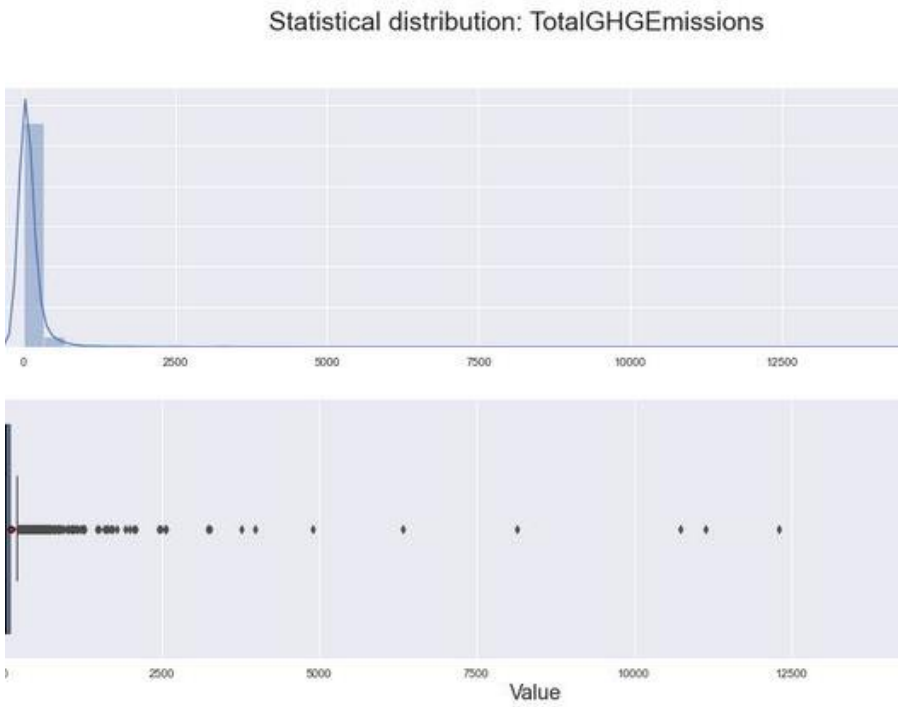
- Le dataframe a 3376 lignes et 46 colonnes
- Le dataframe a un taux de remplissage de 87.15%
- Il y a 19 952 valeurs manquantes sur 115 296

Nettoyage de données

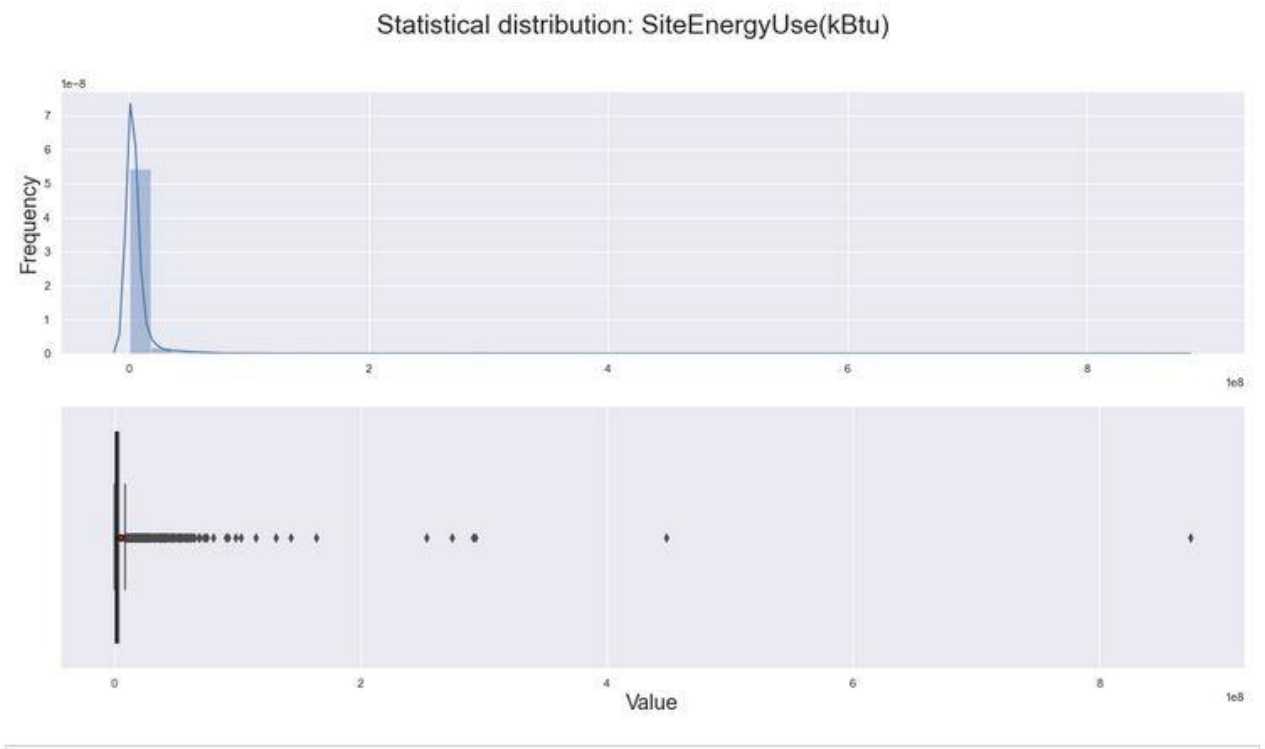
Suppression de colonnes non pertinentes pour notre modèle, tells que :

- Données sans catégorisation possible
- Données avec une unique information (exemple : State)
- Suppression de données non destinés à l'habitation
- Suppression de données issues des relevés de consommation annuels
- Élimination des colonnes trop peu remplies
- Données sans information pertinente pour le modèle :

  - DefaultDate : booléen avec beaucoup de NaN
  - SPD Beats : Informations non utiles à la problématique + beaucoup de NaN
  - Features redondantes (address/zipcode remplacées par latitude et longitude)



Distributions empiriques des variables cibles



Distributions empiriques des variables cibles

Feature engineering

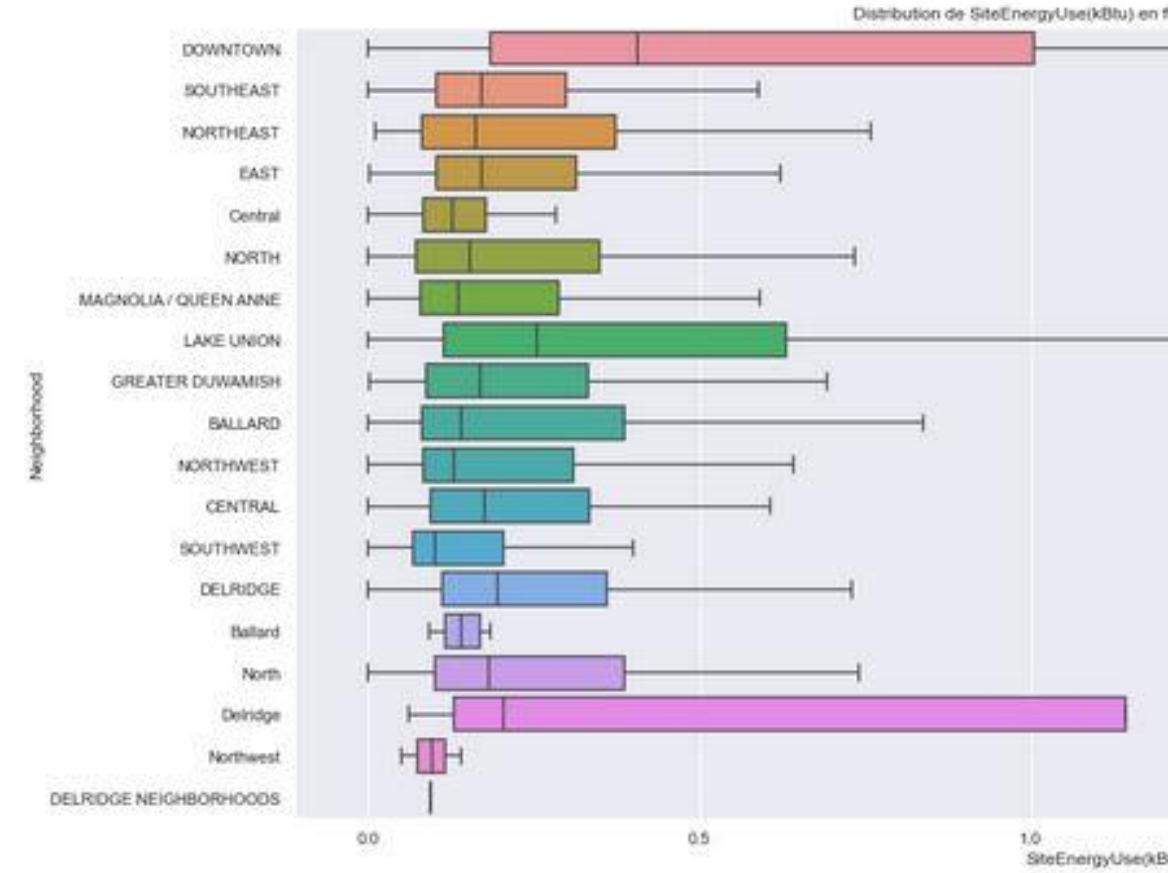
- **Prévision :**

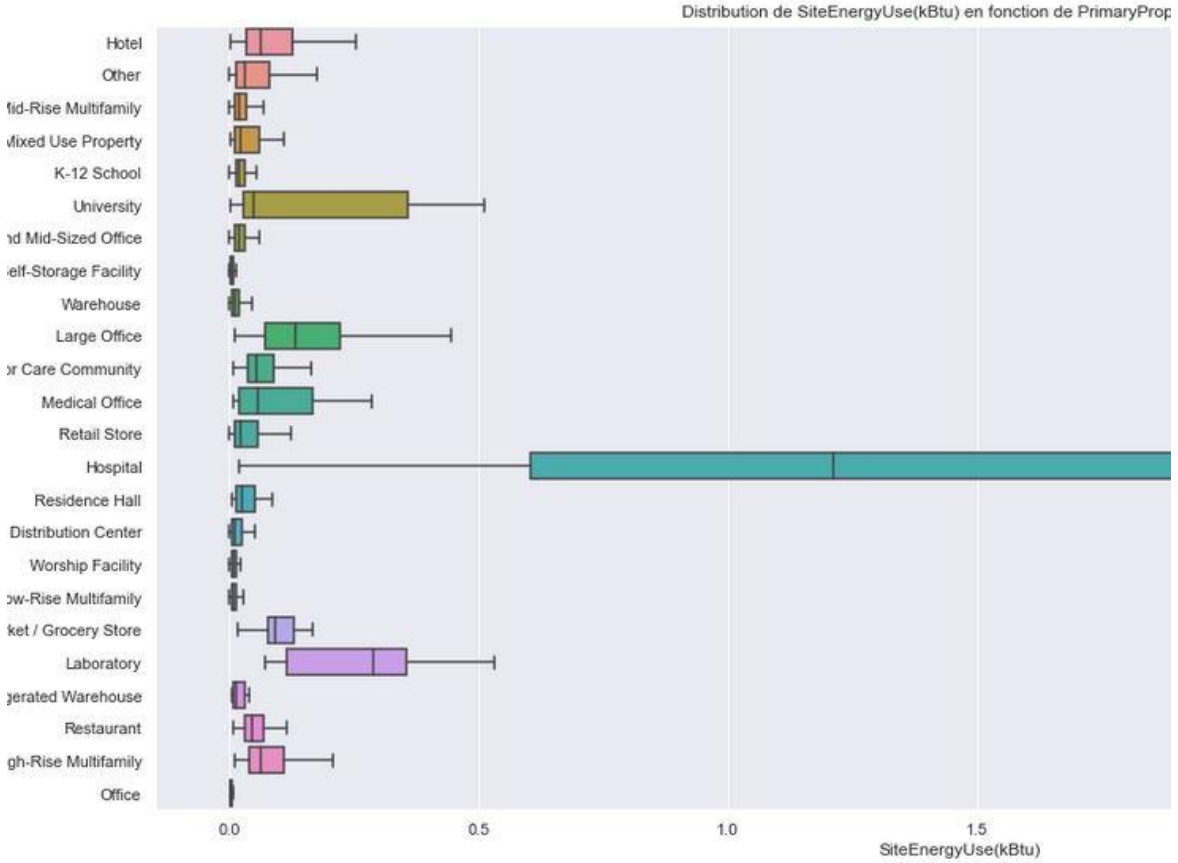
- Idées écartées
  - Features liées à la proportion des sources d'énergie (coûteux à obtenir pour futures données)

Idées retenues :

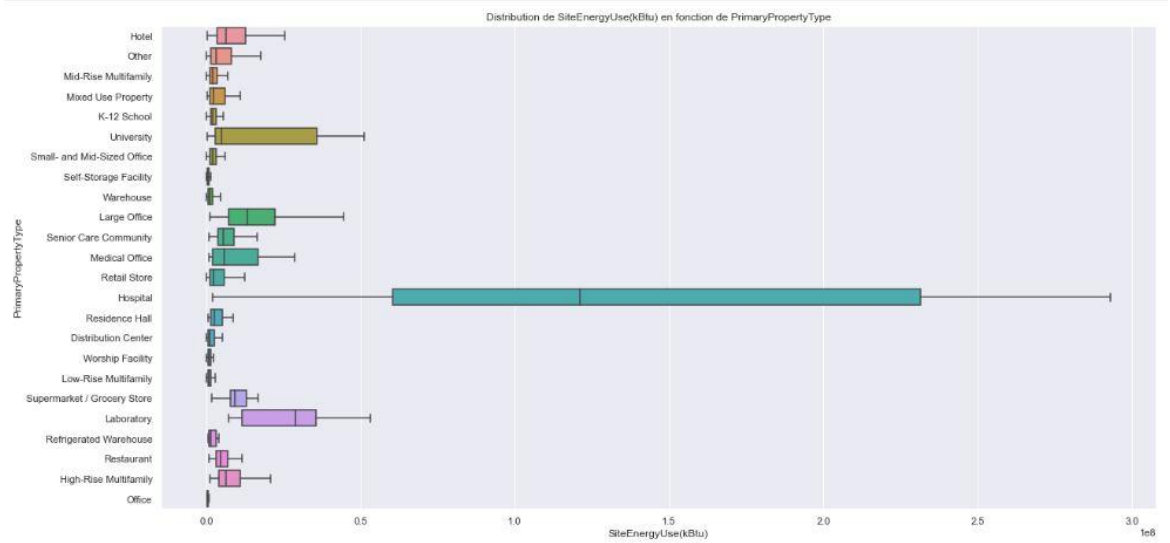
- Suppression des features de consommation (ormis les 2 features qu'on cherche à prédire)
- Utilisation du Energy Star score (mis de côté pour analyse ultérieure)
- Création de nouvelles features (nb_use_types, distance_center, BuildingAge)
- One hot encoding : Transformation d'une feature avec n catégories en n features booléennes.
- Log2-transformation variable de prédiction

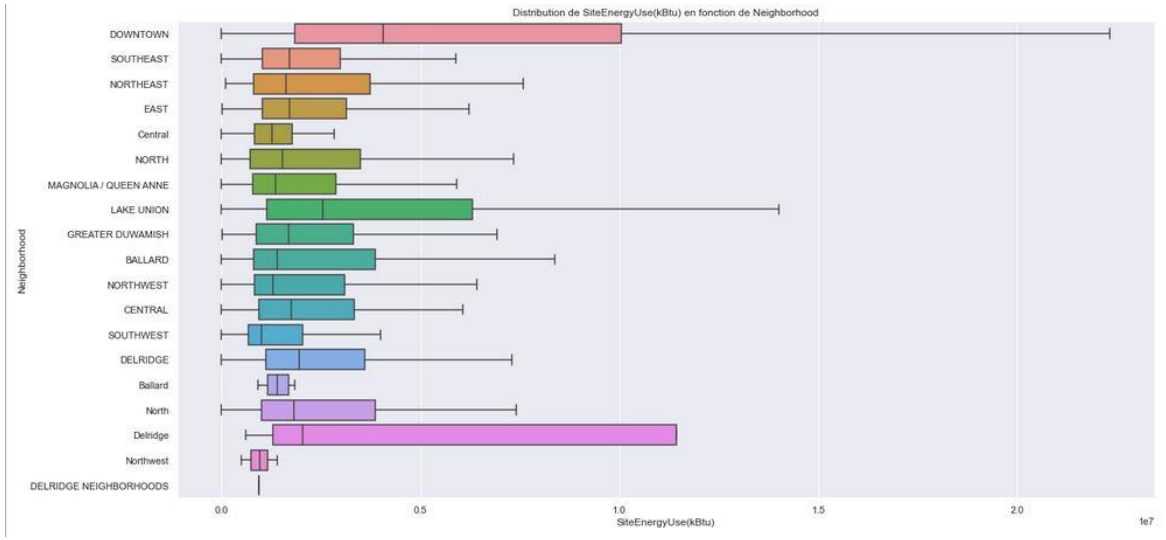
Avant regroupement …





Après regroupement …





Exploration de données

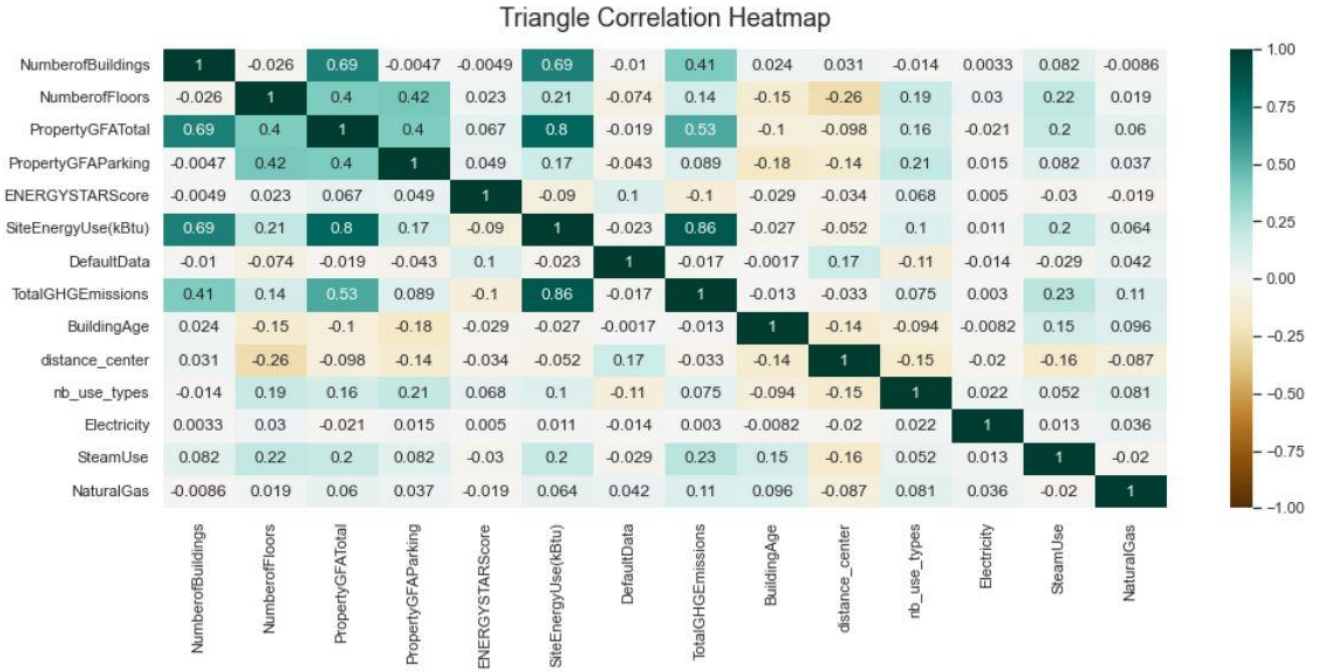
Points Majeurs :

Points Majeurs :

Pour des variables de consommation : Corrélation importante entre :

- PropertyGFATotal et PropertyGFABuildings
- PropertyGFATotal et LargestPropertyUseTypeGFA
- LargestPropertyUseTypeFGA et PropertyGFABuildings

Energy Star Score : pas de corrélation notable



III - Modélisations

Modèle de consommation : démarche

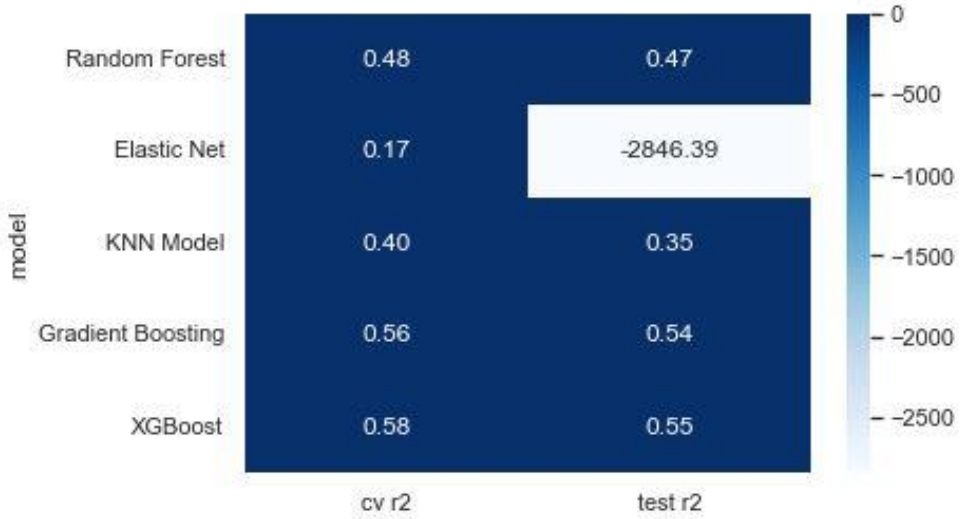
Séparation du jeu de données (train/validation et test)

Pour chaque algorithme

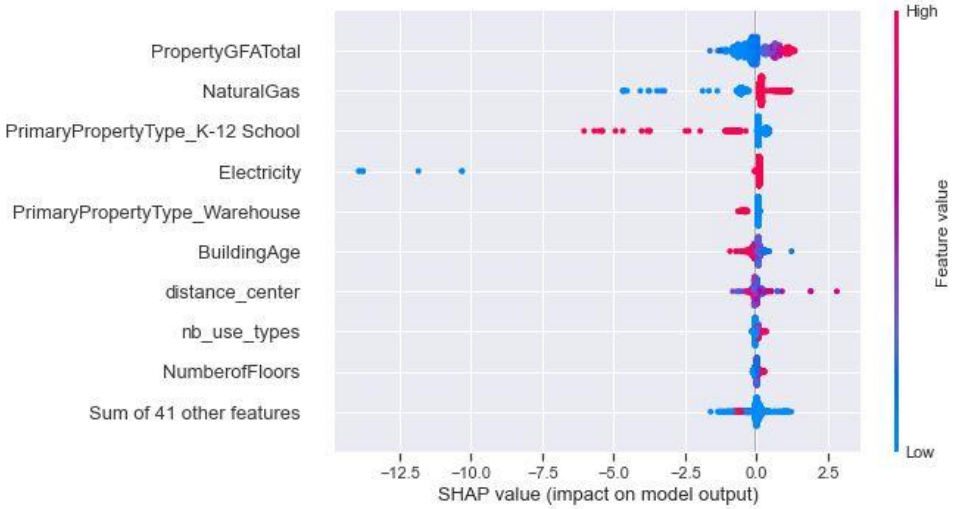
- - Définition grille de paramètres
- - Entrainement des modèles (N modèles, Jeu training, Cross-validation)

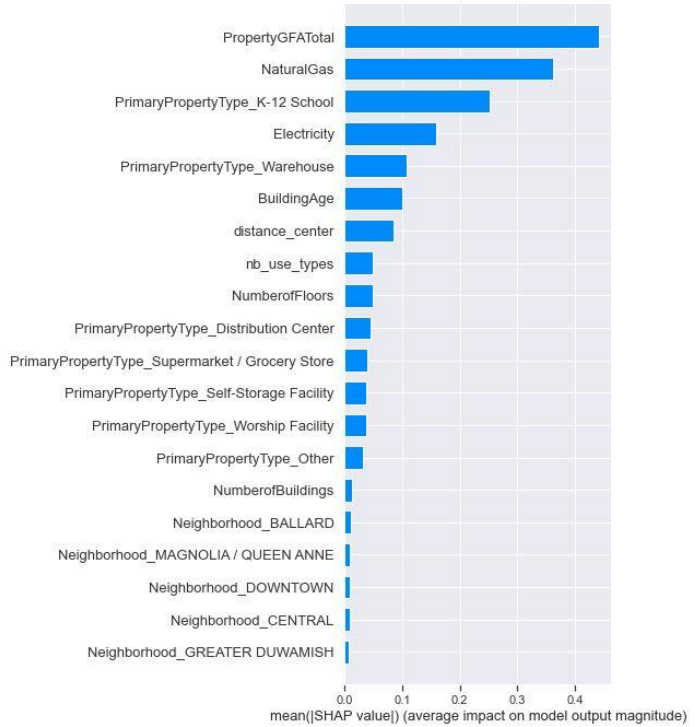
Comparaison des modèles sur la RMSE et le R2 de validation

III-I Modèle avec SiteEnergyUse(kbtu)



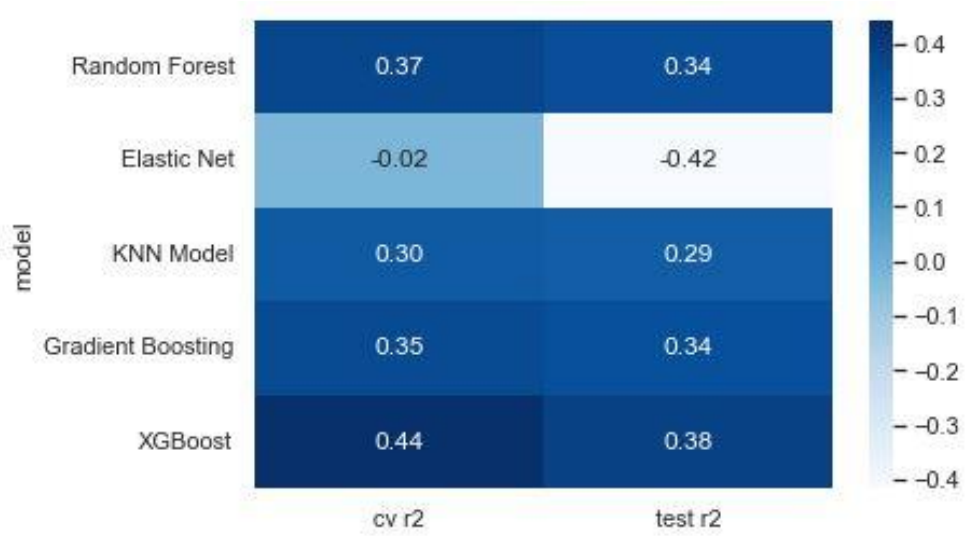
Modèle avec : SiteEnergyUse(kBtu)

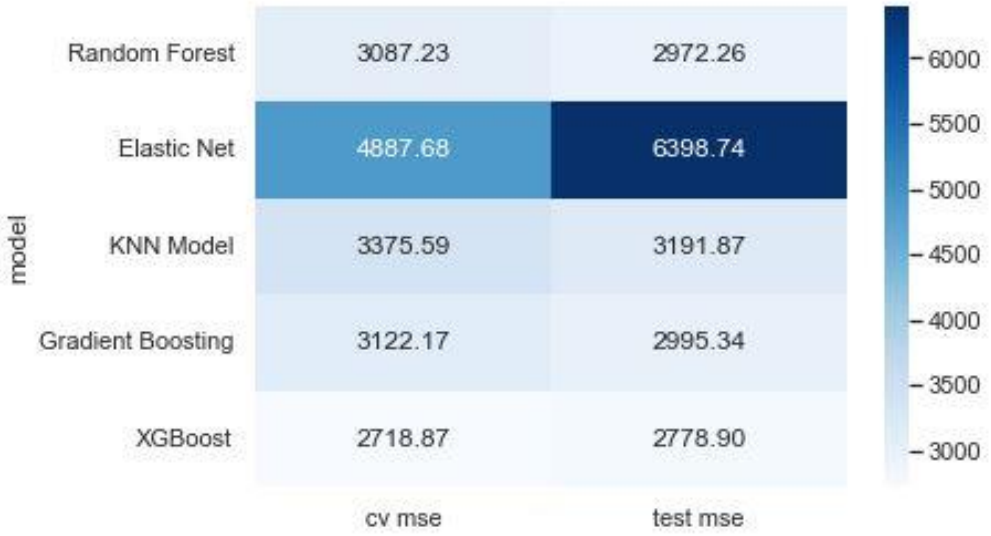




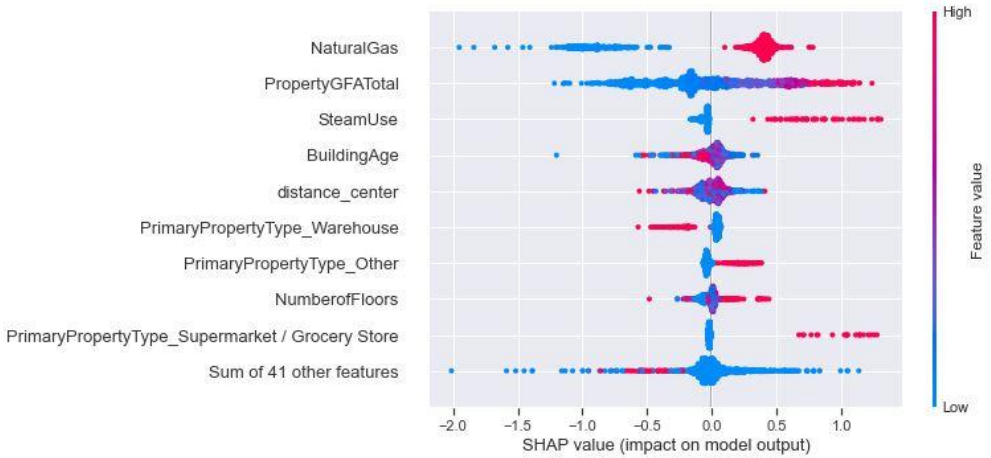
Interprétation globale du modèle avec SiteEnergyUse(kBtu)

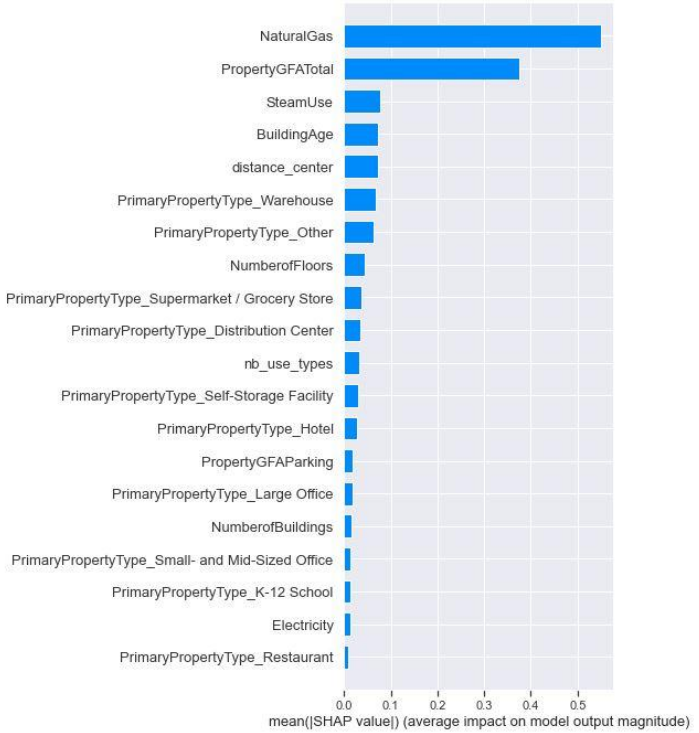
III-III Modèle TotalGHGEmissions sans EnergyStar





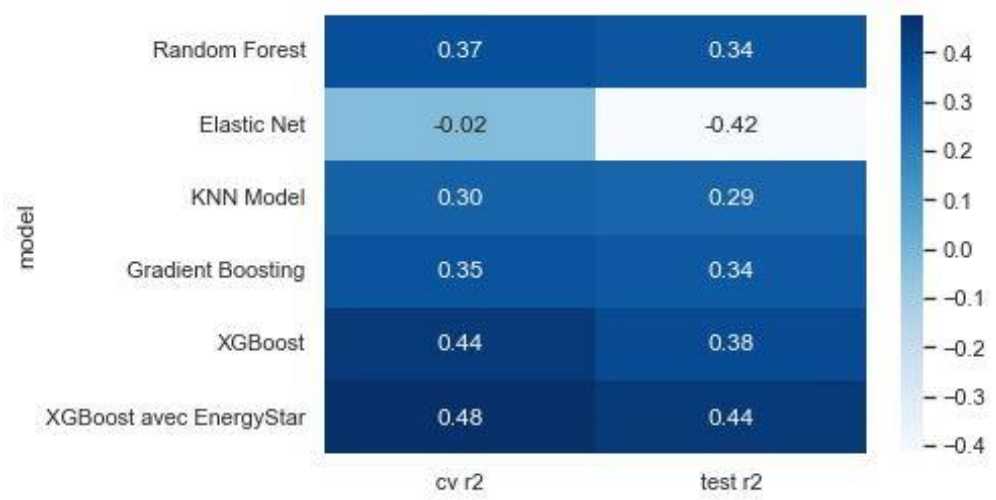
Modèle : SiteEnergyUse(kBtu)



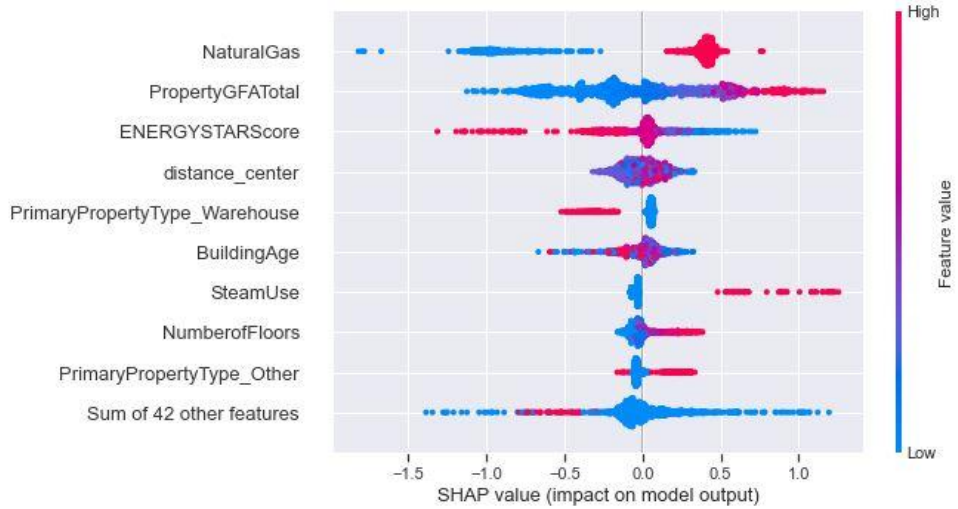


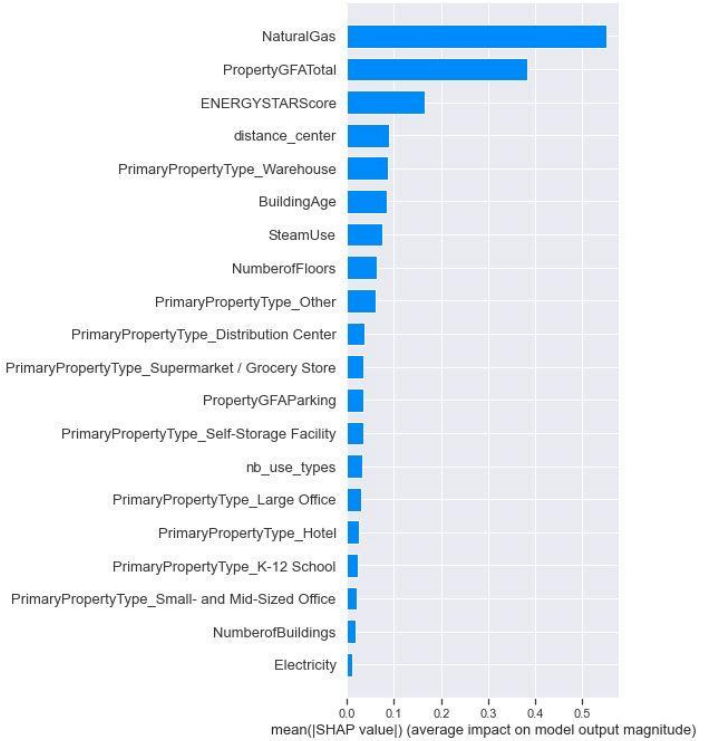
Interprétation globale du modèle avec SiteEnergyUse(kBtu)

III-III Modèle TotalGHGEmissions avec EnergyStar



Modèle : TotalGHGEmissions





Interprétation globale du modèle avec TotalGHGEmissions

Conclusion

- Intérêt du Energy star Score
- Le meilleur modèle

MERCI POUR VOTRE ATTENTION