

Projet 4 : Anticipez les besoins en consommation de bâtiments

Présentation de la problématique:



- Objectif de la ville de Seattle : atteindre la neutralité carbone en 2050 ➡ nécessité de mesurer la quantité de gaz à effet de serre émise.
- Sujet du projet : étude de la consommation énergétique et des émissions en carbone des bâtiments non destinés à l'habitation.
- Les relevés effectués sont très coûteux et fastidieux à collecter.
- **Notre objectif** : Prédire la consommation totale d'énergie et les émissions de CO₂ de bâtiments pour lesquelles elles n'ont pas encore été mesurées.

PLAN d'Etude:

I. Présentation des données relevées en 2015 et 2016

II. Nettoyage des données

III. Analyse exploratoire

IV. Modélisation

V. Comparaison des modèles

VI. Détermination de l'intérêt de l'Energy star score

VII. conclusions

Présentation des données

- Les données sont :
 - disponibles sur le site [kaggle.com](https://www.kaggle.com) (mises à disposition par la ville de Seattle via leur portail web dédié aux données <https://data.seattle.gov/>)
 - 2 fichiers csv correspondant aux relevés effectués en 2015 et 2016. (tailles respectives : (3340 lignes, 47 colonnes) et (3376 lignes, 46 colonnes)).
- Les lignes :
 - les bâtiments avec leurs identifiants.
- Les colonnes :
 - les caractéristiques intrinsèques de l'édifice (sa date de construction, le nombre de bâtiments , le nombre d'étages, sa surface, son usage...)
 - Ses consommations énergétiques et ses émissions carbone
- **Objectif** : regrouper toutes les données au sein d'un seul tableau

Uniformisation et jointure des 2 datasets

- l'uniformisation des 2 datasets est nécessaire pour effectuer leur jointure.
- Modification des noms de certaines colonnes : certaines variables ont des noms différents dans les 2 tableaux de données.
ex : `GHGEmissions(MetricTonsCO2e) / TotalGHGEmissions`.
- Création de nouvelles colonnes :
ex : Dans le dataset de 2016, plusieurs colonnes : Address , ZipCode, Longitude, Latitude, City etc.
Dans le dataset 2015, toutes ces variables sont regroupées dans une seule colonne : Location
- Jointure des 2 datasets par la colonne OSEBuildingID (jointure externe complète pour garder toute les informations des 2 tableaux)

Nettoyage

- NAN: certaines variables ont des valeurs manquantes car le bâtiment ne possède pas la caractéristique correspondante.

Pour les variables SecondLargestPropertyUseTypeGFA et ThirdLargestPropertyUseTypeGFA:



Pour les variables SecondLargestPropertyUseType et ThirdLargestPropertyUseType :



- Suppression de colonnes non utiles à notre étude:
 - Census Tracts, Seattle Police Department Micro Community Policing Plan Areas, City Council Districts, SPD Beats, Zip Codes
 - Données avec une information unique pour tous notre échantillon(exemple: State, City)

Synthèse de toutes les données

- Réalisation d'un tableau qui synthétise toutes nos données:
 - conservation des données des bâtiments présents uniquement dans l'un des 2 datasets de 2015 ou 2016.
 - les bâtiments communs aux 2 datasets:
 - conservation de leurs données intrinsèques car elles sont très similaires
 - réalisation de la moyenne de leurs consommations énergétiques et de leurs émissions carbone.
- Taille du jeu de données final : 3432 lignes et 38 colonnes

Choix des variables pertinentes

- **Objectif** : développement d'un modèle de machine Learning pour la prédiction de la consommation énergétique et des émissions de gaz à effet de serre à partir des caractéristiques intrinsèques d'un bâtiment.
- Variables à prédire (targets) :
 - Consommation d'énergie annuelle totale des bâtiments
SiteEnergyUse(kBtu)
 - La quantité totale des gaz à effet de serre émise
TotalGHGEmissions
- Variables d'entrée (features) :
 - Caractéristiques intrinsèques aux bâtiments hors consommations (la surface, l'activité qu'il abrite, nombre d'étage etc.).

Choix des variables pertinentes

- 2 types de variables :

- variables quantitatives :

NumberofBuildings, NumberofFloors, PropertyGFATotal, PropertyGFAParking etc.

- variables qualitatives :

LargestPropertyUseType, PrimaryPropertyType, SecondLargestPropertyUseType etc.

Nécessité de les transformer en données numériques pour que l'ordinateur puisse les intégrer dans ses calculs = Encodage

Utilisation du transformateur de sklearn : OneHotEncoder (Représentation de façon binaire de chaque catégorie ou classe dans une colonne qui lui est propre)

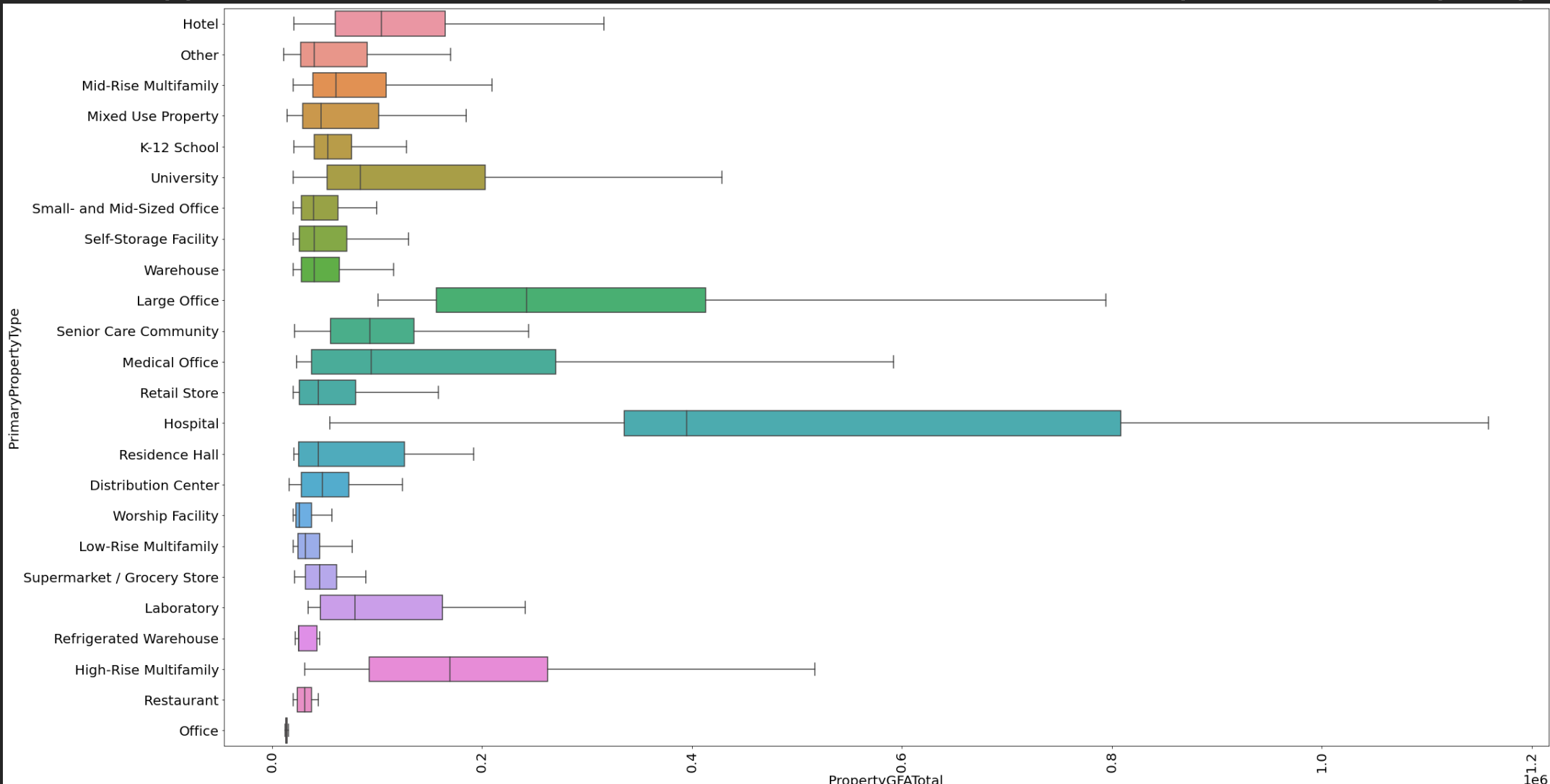
➡ Décomposition de la variable initiale en plusieurs sous-variables (colonnes).

- ENERGY STAR Score:

À la fin de notre projet , l'Energy star score sera ajouté aux données d'entrée de notre modèle pour évaluer son intérêt (comparaison du modèle avec et sans l'Energy star score).

Etude exploratoire

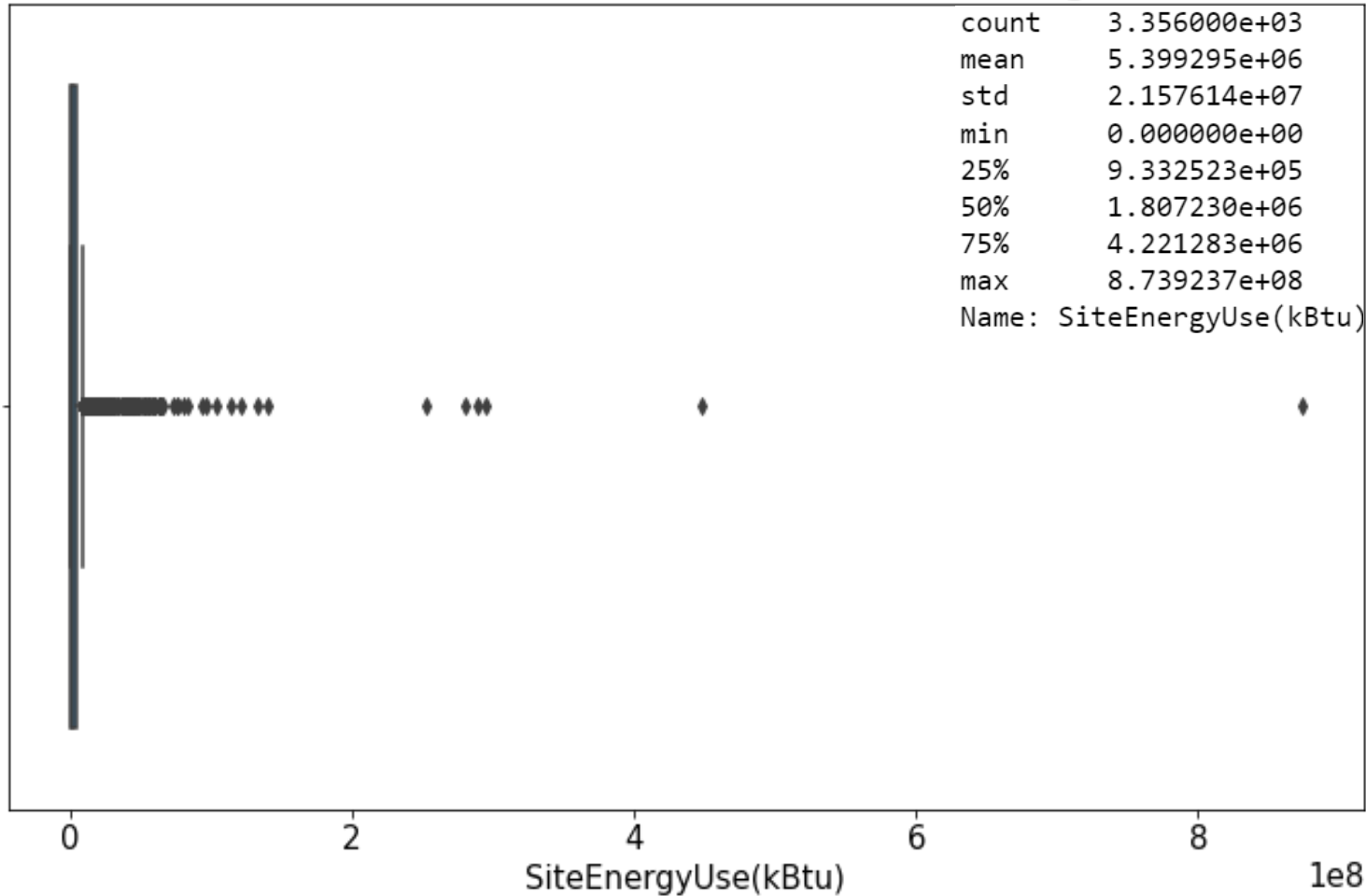
- Les types de bâtiments et la distribution de leurs tailles respectives (PropertyGFATotal)



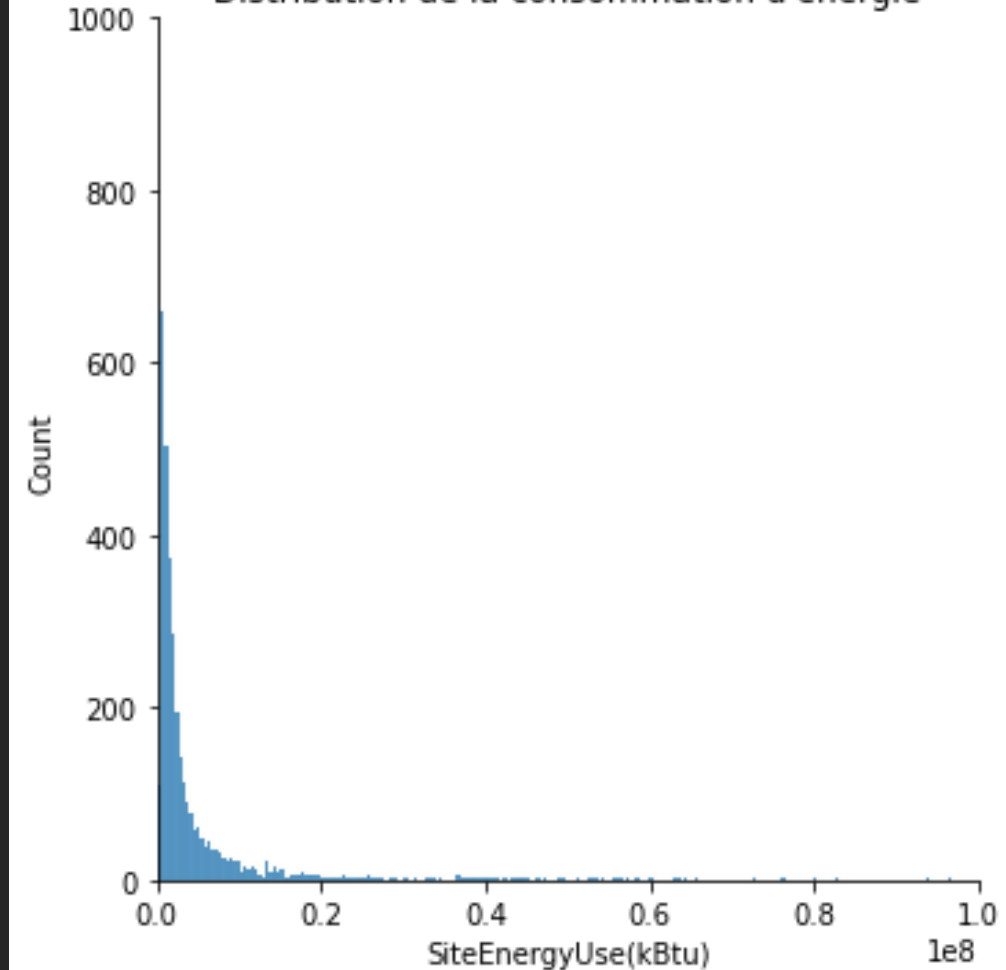
Distribution de la consommation d'énergie

- Dispersion très étendue de la consommation énergétique

Distribution de la consommation d'énergie

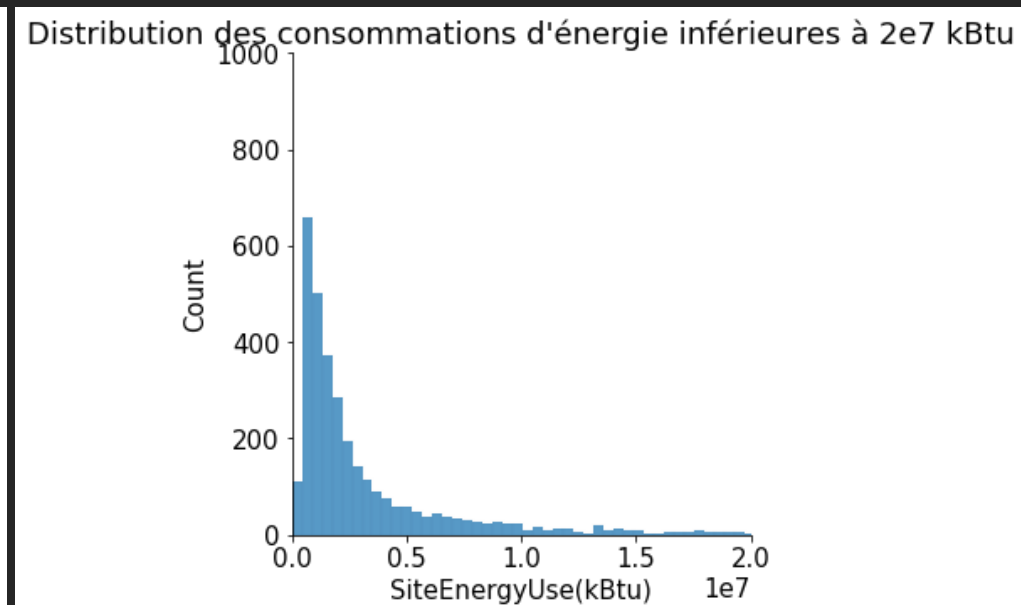
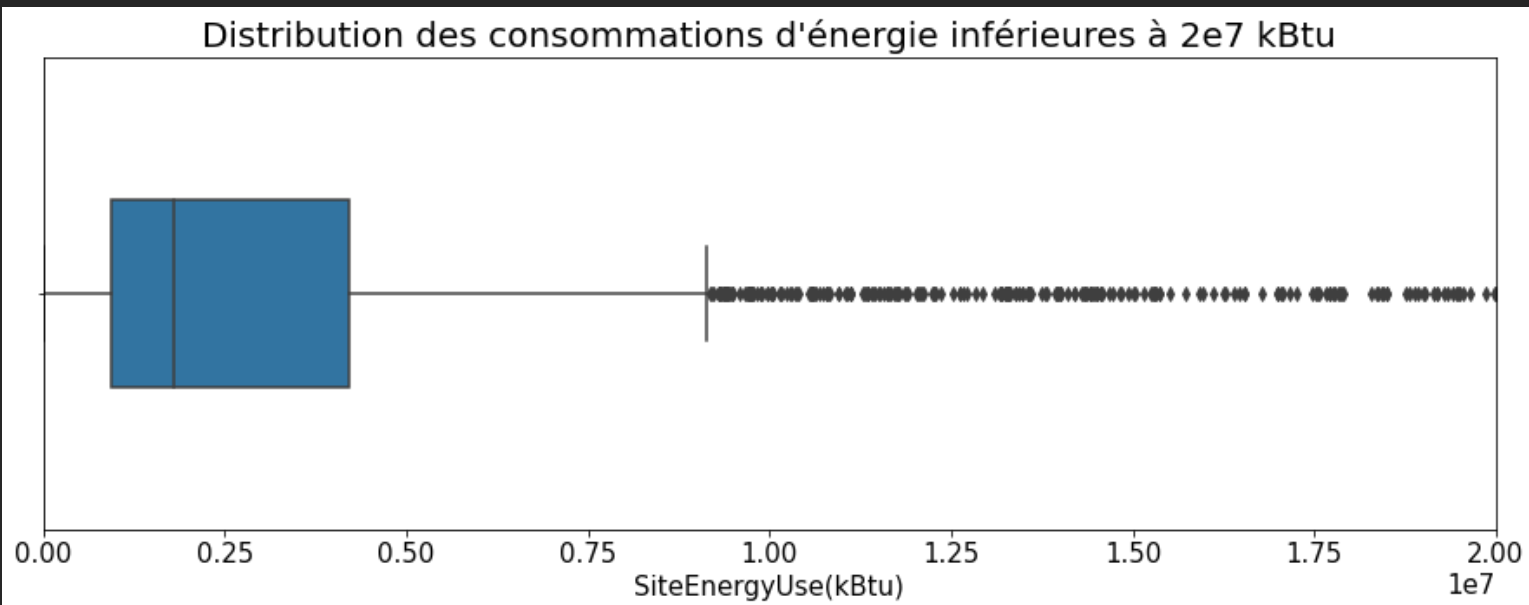


Distribution de la consommation d'énergie



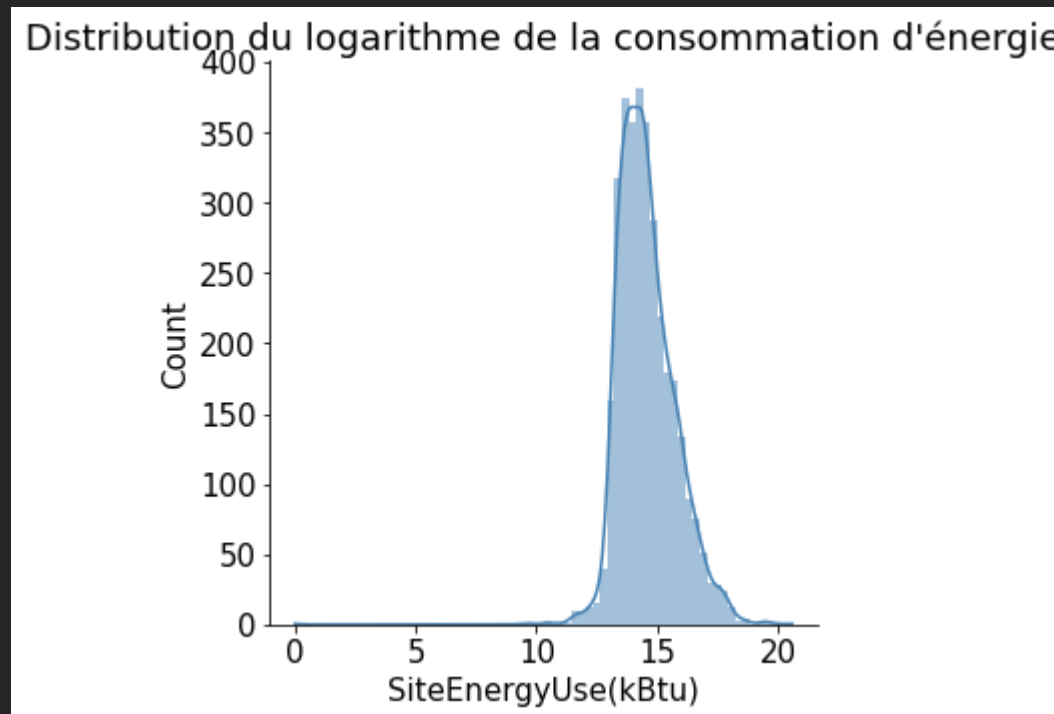
Distribution de la consommation d'énergie:

- Réduction de l'intervalle de représentation pour une meilleure visualisation (Energie < $2 \cdot 10^7$ kBtu)



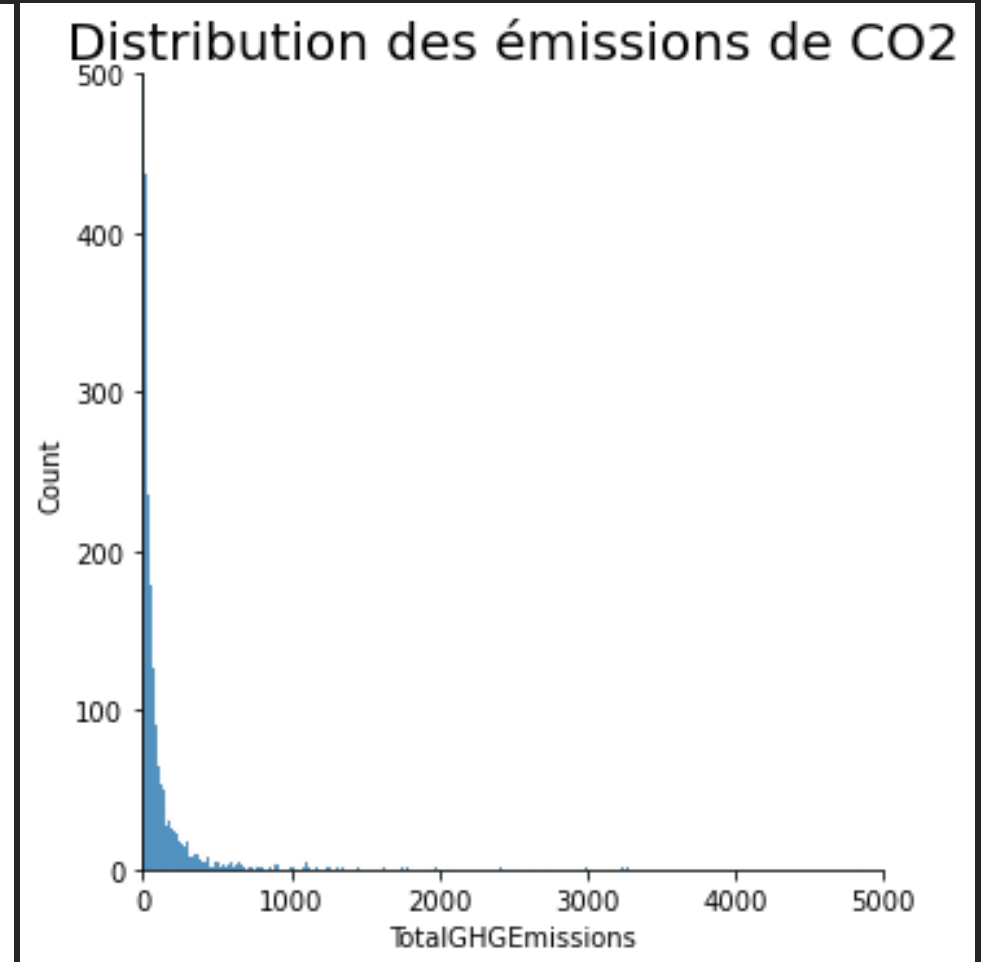
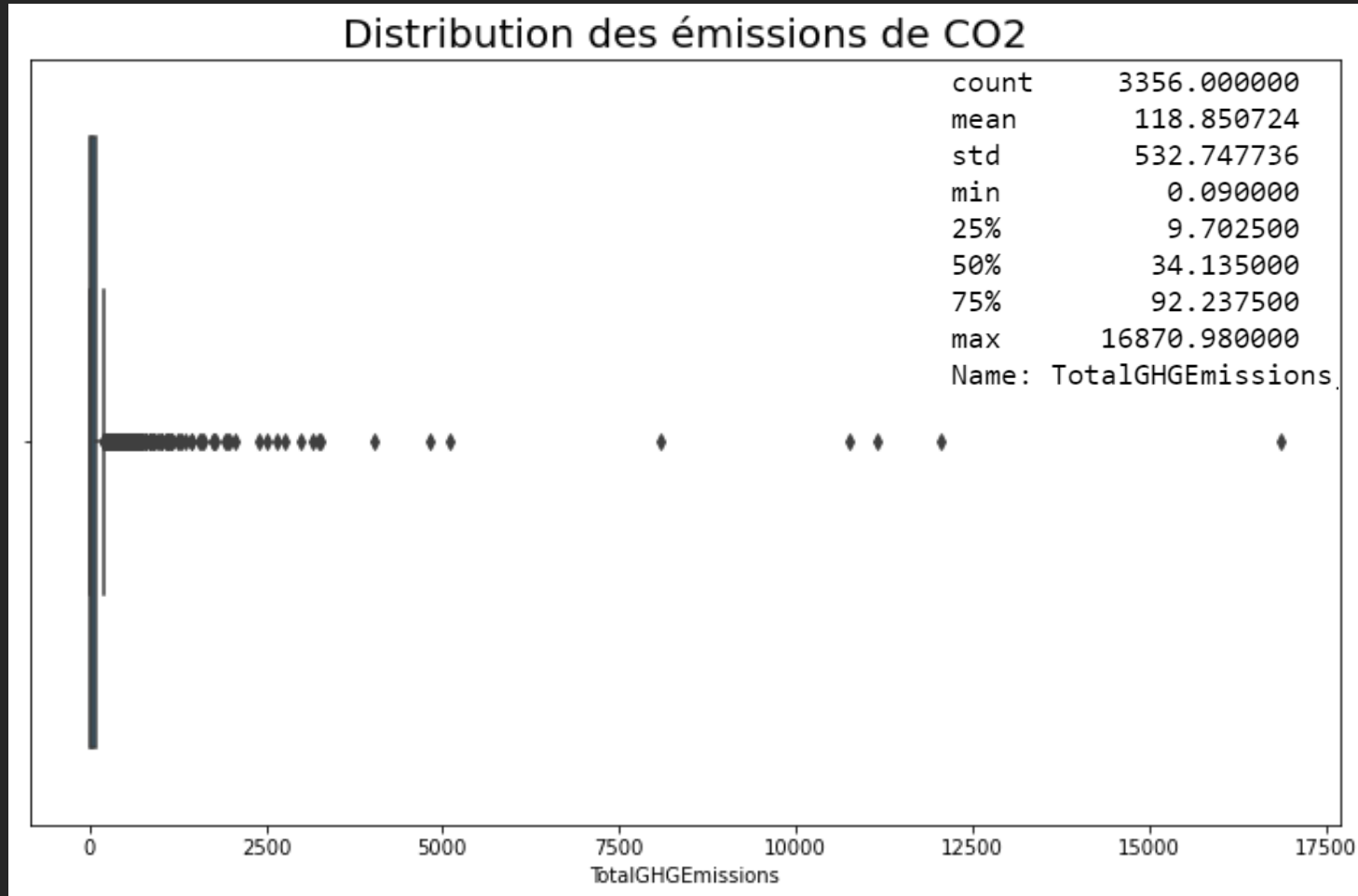
La représentation logarithmique

- Intérêt de l'échelle logarithmique : ■ représentation sur une petite échelle de valeurs très étendues.



Distribution des émissions de CO₂

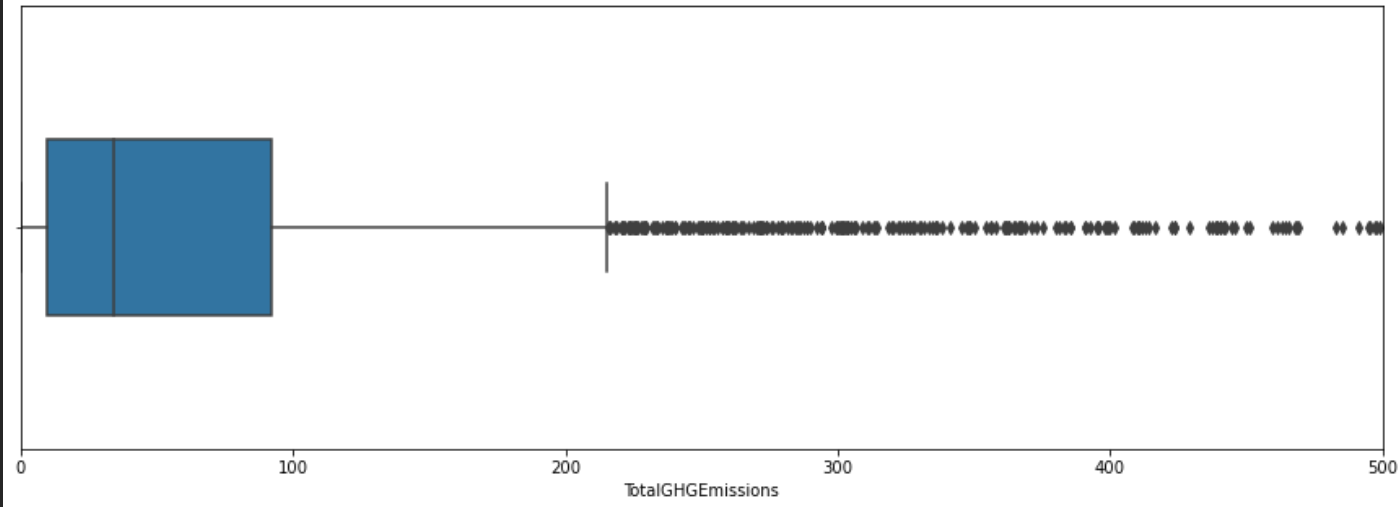
- Dispersion également très étendue des émissions de CO₂



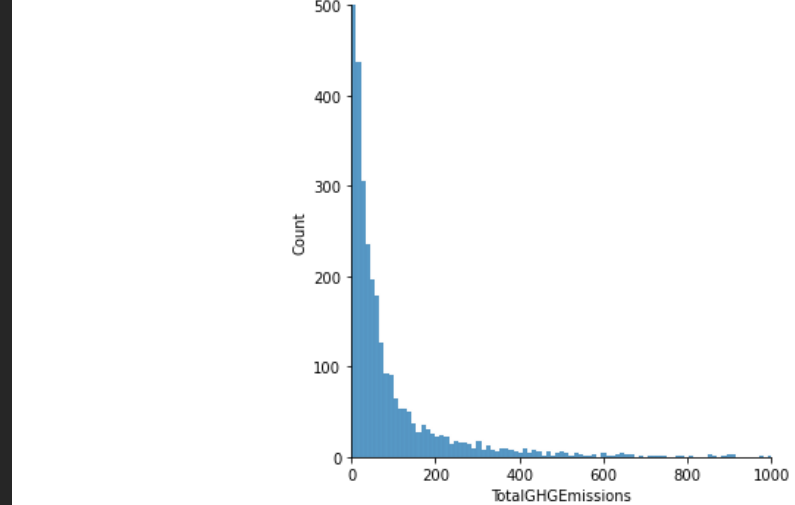
Distribution de la consommation d'énergie:

- Réduction de l'intervalle de représentation pour une meilleure visualisation (Emissions<1000 MetricTonsCO2e)

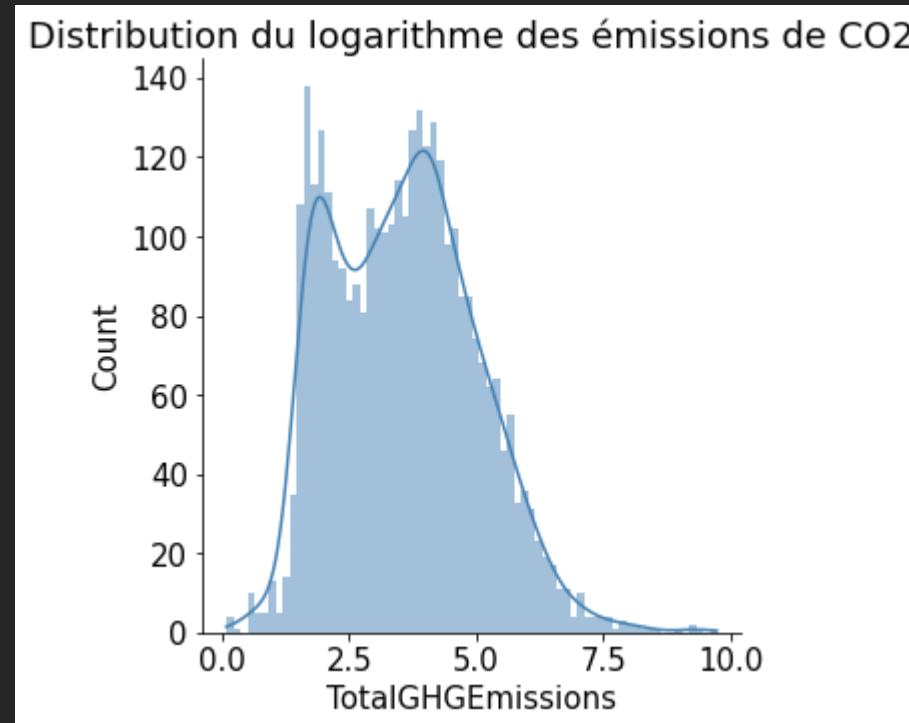
Distribution des émissions de CO2 inférieures à 1000 en MetricTonsCO2e



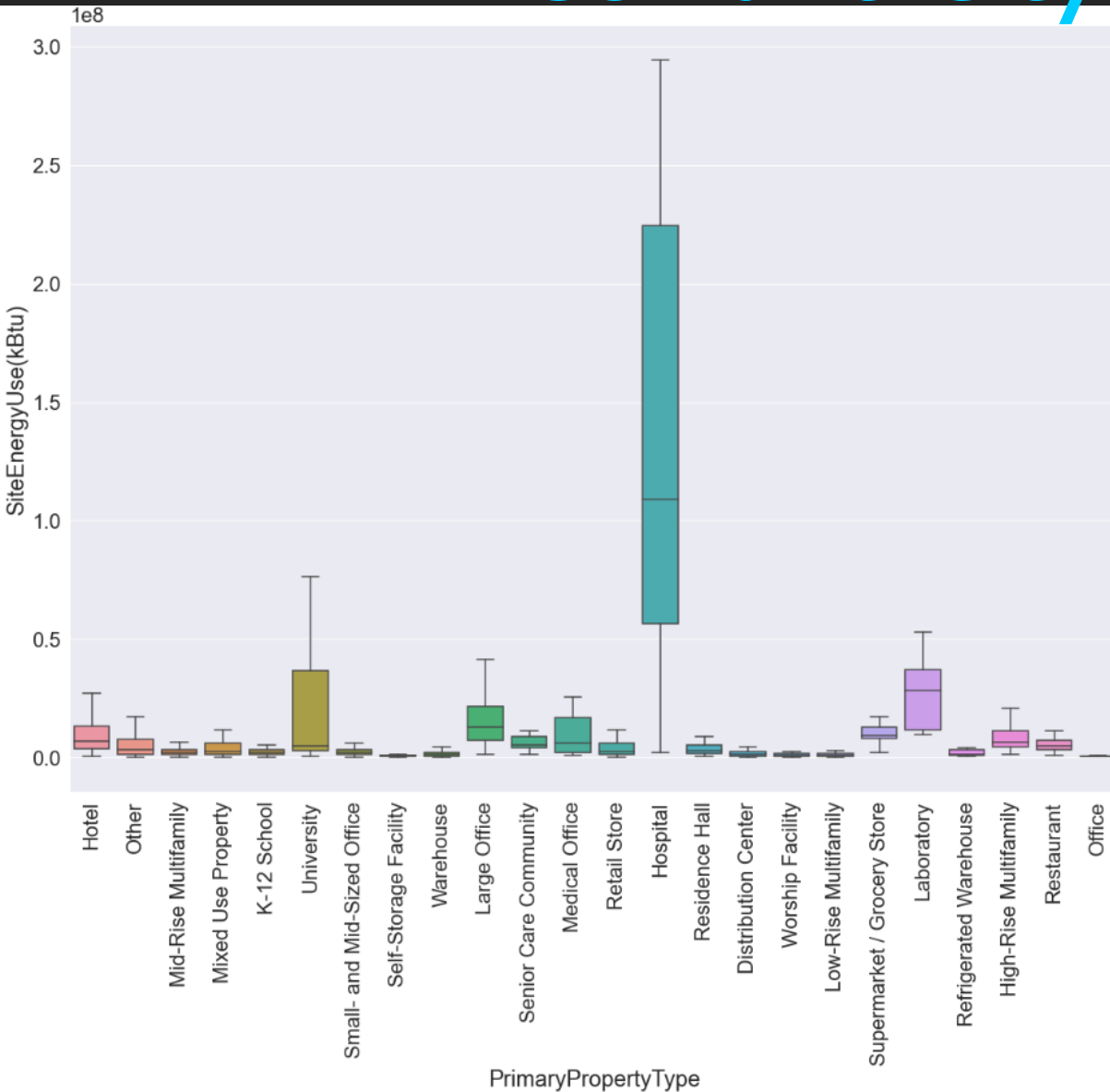
Distribution des émissions de CO2 inférieures à 1000 en MetricTonsCO2e

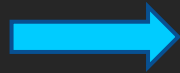


La représentation logarithmique

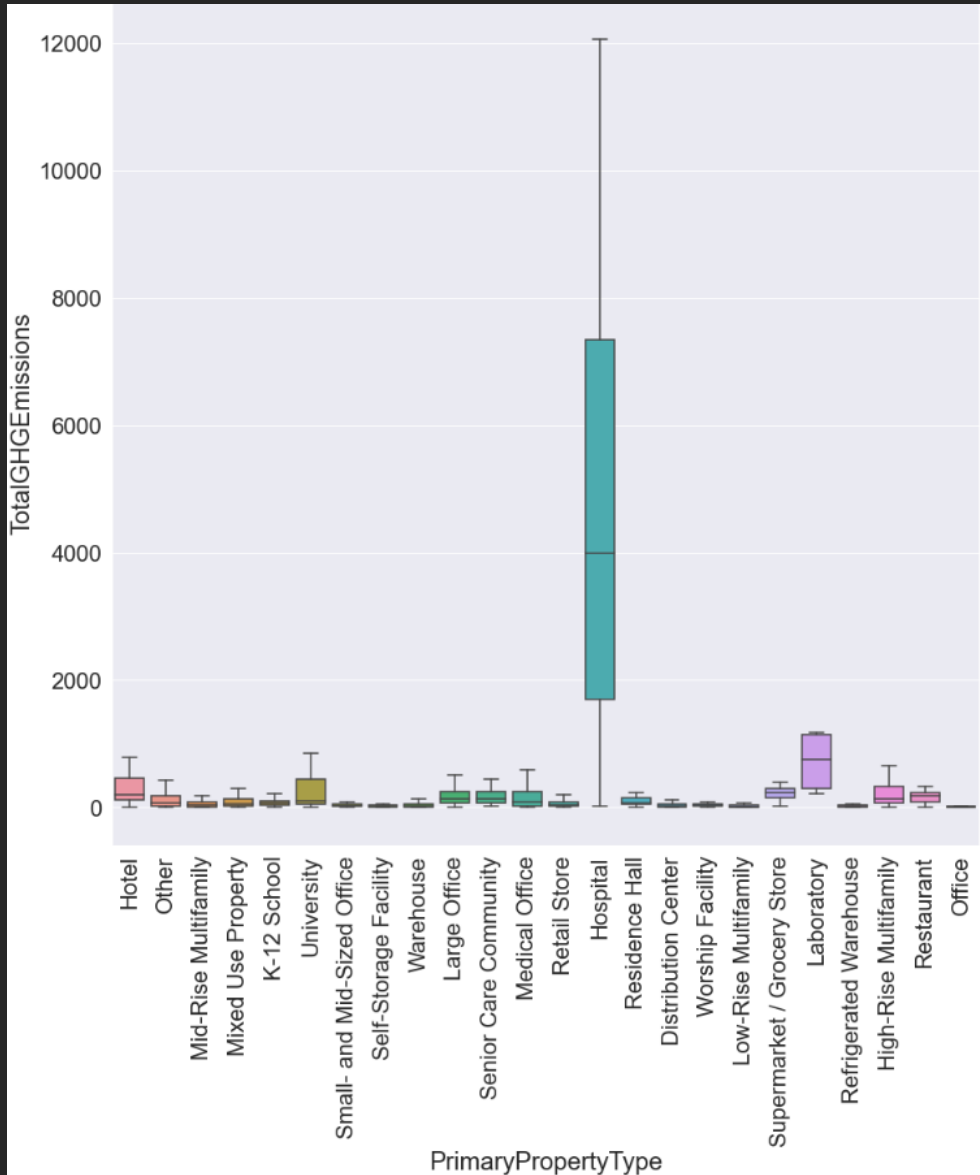


Distribution de la consommation énergétique suivant le type de bâtiment



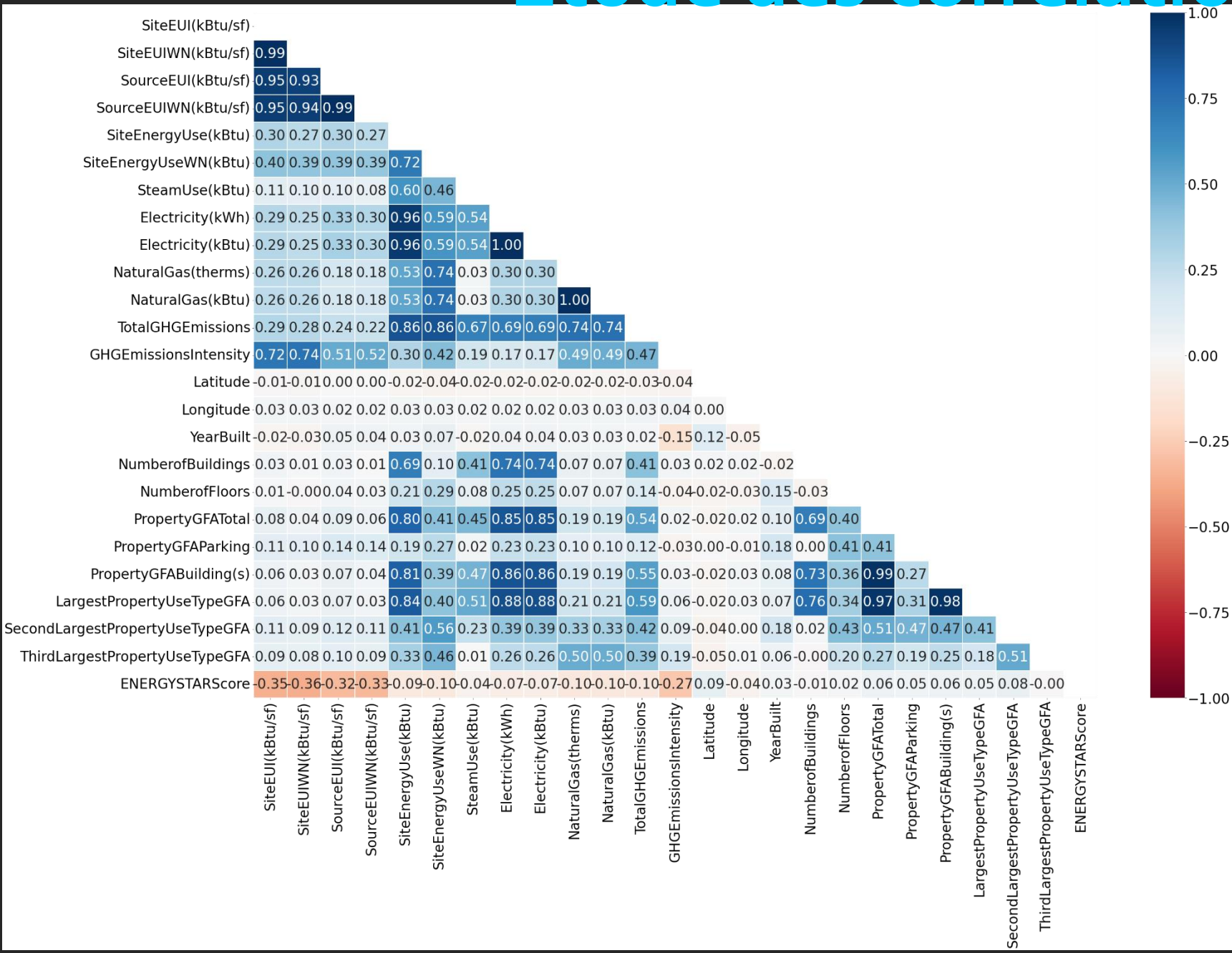
- Les consommations énergétiques dépendent du type de bâtiment.
- Vérification de notre hypothèse intuitive : lien entre la variable **PrimaryPropertyType** et la variable à prédire **SiteEnergyUse**.
-  **PrimaryPropertyType** sera importante dans notre modèle de prédiction.

Distribution des émissions de CO₂ suivant le type de bâtiment



- Même conclusion que précédemment :
 - lien entre la variable **PrimaryPropertyType** et la variable à prédire TotalGHGEmissions .

Etude des corrélations



▪ SiteEnergyUse(kBtu) :

- corrélations importantes avec :
 - LargestPropertyUseTypeGFA
 - PropertyGFABuilding(s)
 - PropertyGFATotal
 - TotalGHGEmissions

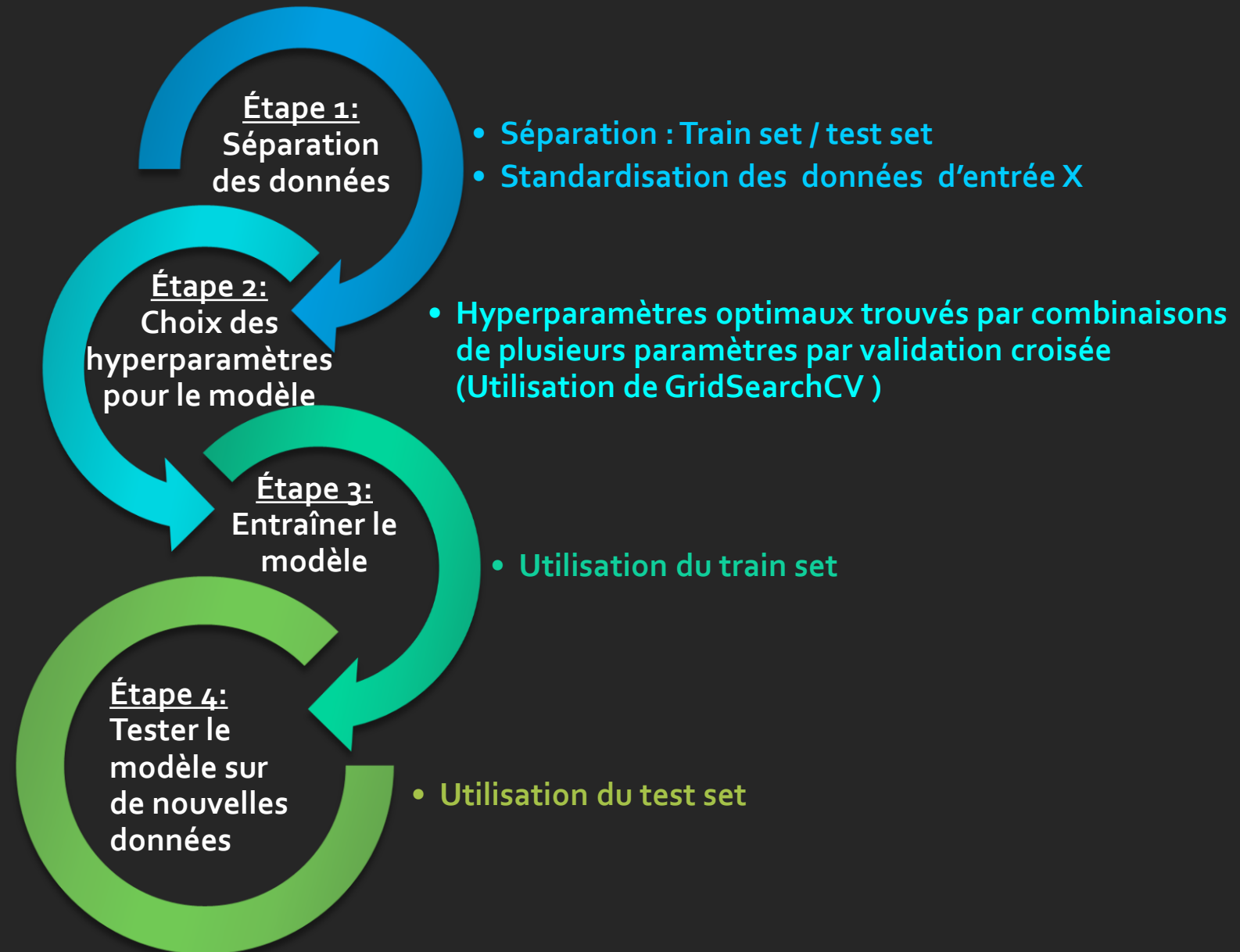
▪ TotalGHGEmissions :

- corrélations plus faibles avec :
 - LargestPropertyUseTypeGFA
 - PropertyGFABuilding(s)
 - PropertyGFATotal
- corrélation importante avec :
 - SiteEnergyUse(kBtu).

▪ ENERGYSTARScore :

- corrélations négatives faibles avec :
 - SiteEUIWN(kBtu/sf)
 - SiteEUI(kBtu/sf)
 - SourceEUIWN(kBtu/sf)
 - SourceEUI(kBtu/sf)
 - GHGEmissionsIntensity

Etapes d'optimisation des modèles



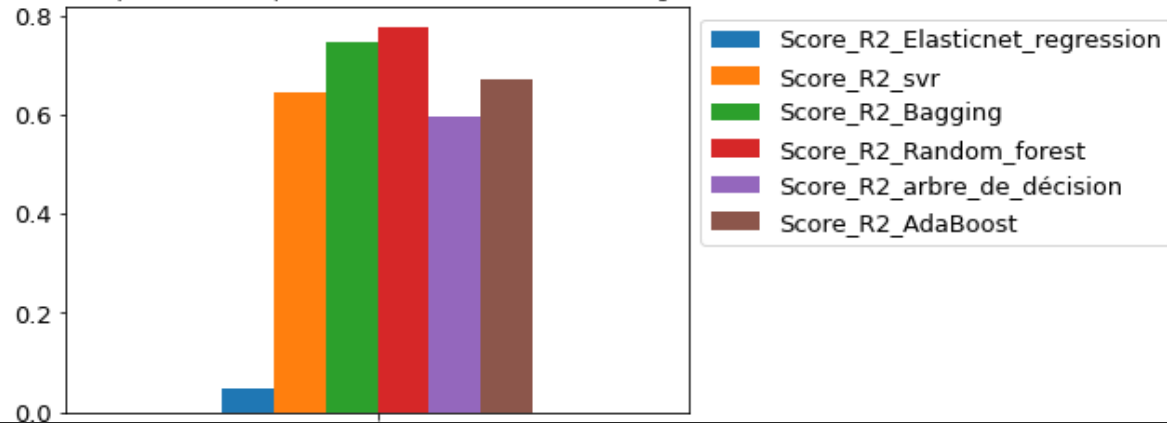
Recherche des meilleurs paramètres pour chaque modèle

Modèles simples		Modèles ensemblistes			
		Méthodes parallèles			Méthodes séquentielles (boosting)
		Bagging (par défaut base_estimator=DecisionTreeRegressor)	Arbres de décision	Forêts aléatoires	Adaboost (par défaut base_estimator=DecisionTreeRegressor)
Elastic net regression	SVR (kernel = <i>linear</i> , <i>poly</i> , <i>rbf</i>)				
$\alpha = 10^{-4}, 10^{-1}, \dots, 10, 10^2$	$C : 10^{-2}, 10^{-1}, 1, 10, 100$	$N_estimators = [10, 50, 100, 300, 500, 1000]$	$'max_depth' = [2, 4, 6, 8, 10, 12]$	$N_estimators = [50, 100, 300, 500, 1000]$	$N_estimators = [50, 100, 500, 1000]$
	$\Gamma : 10^{-4}, 10^{-3}, \dots, 1, 10$				
	$\epsilon : 10^{-4}, 10^{-3}, \dots, 10^{-1}, 1$				
En blanc, les meilleurs hyperparamètres trouvés.					

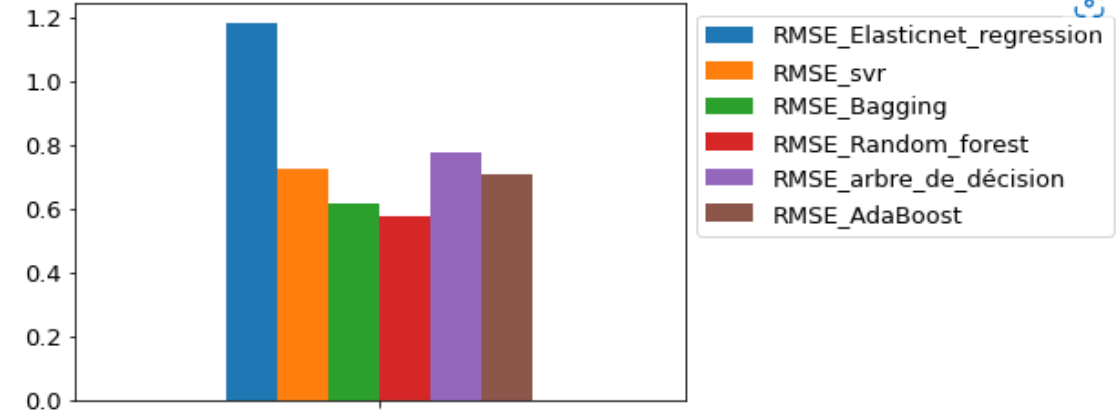
Résultats et comparaison des modèles

■ Prédiction de la consommation d'énergie:

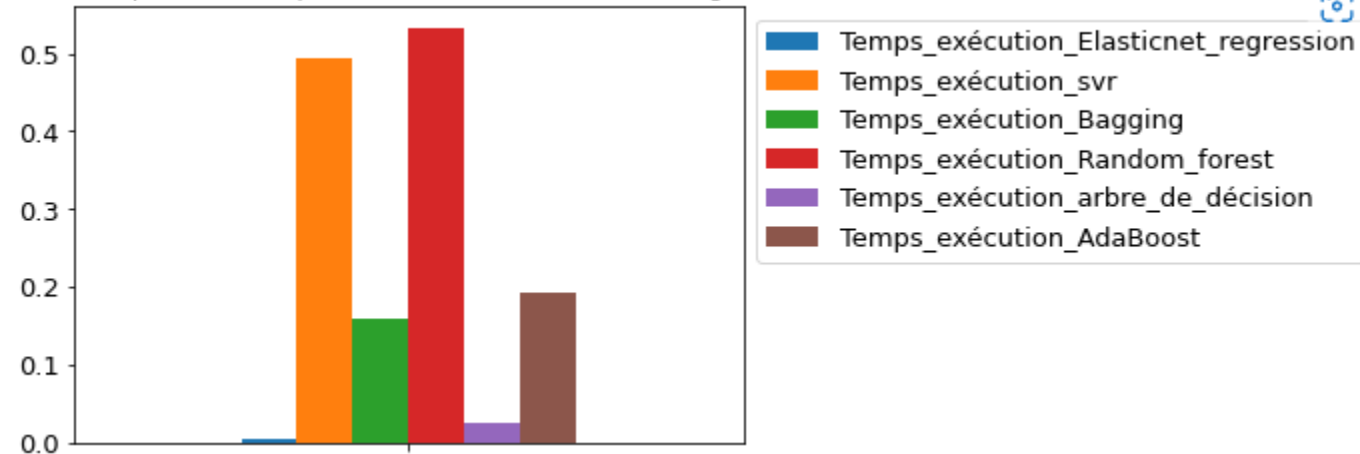
Le score R^2 obtenu par modèle de prédiction de la consommation d'énergie



La RMSE obtenue par modèle de prédiction de la consommation d'énergie

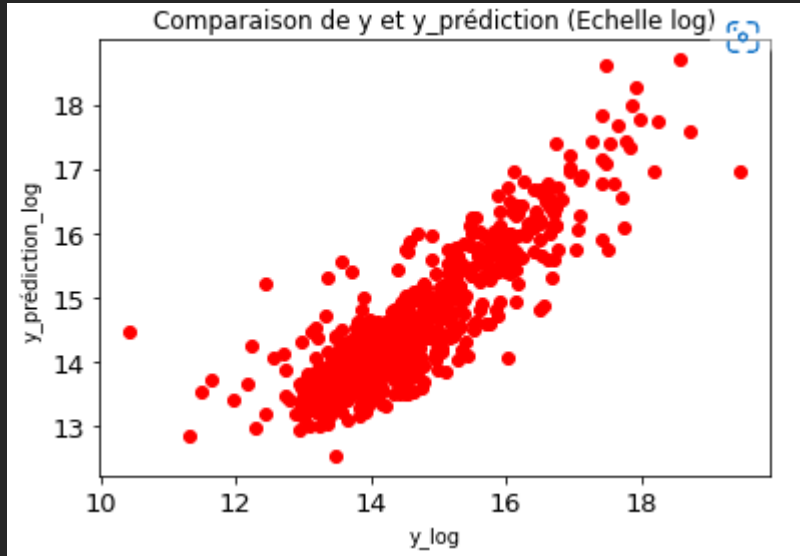


Le temps d'exécution de chaque modèle de prédiction de la consommation d'énergie

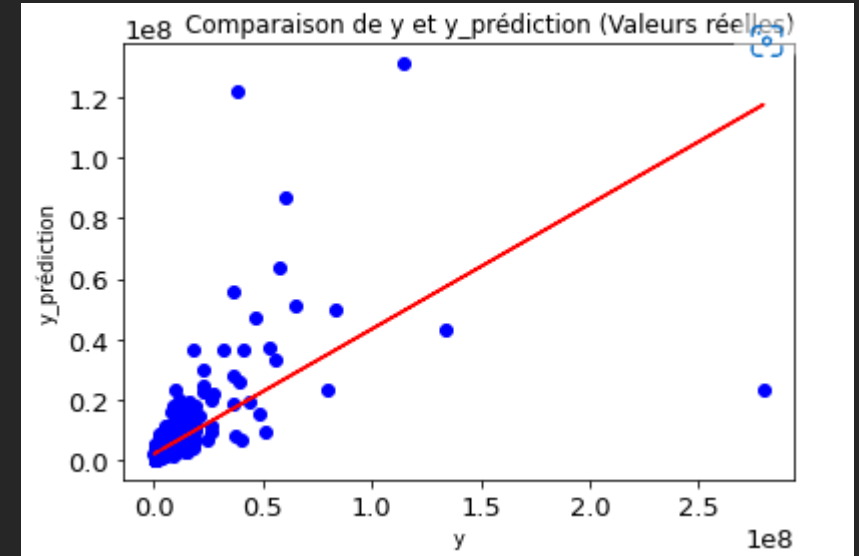


Prédiction par le modèle des forêts aléatoires

- Prédiction de la consommation d'énergie:



exponentielle

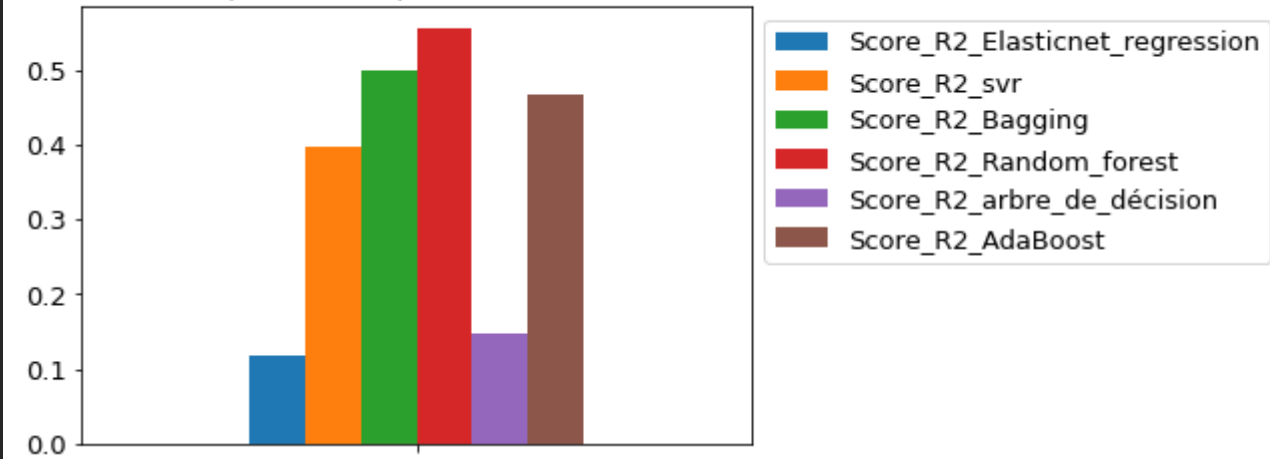


- Limites de notre modèle : Bonnes prédictions pour un certain domaine d'énergie (valeurs inférieures à $0.4 \cdot 10^8$ kBtu)

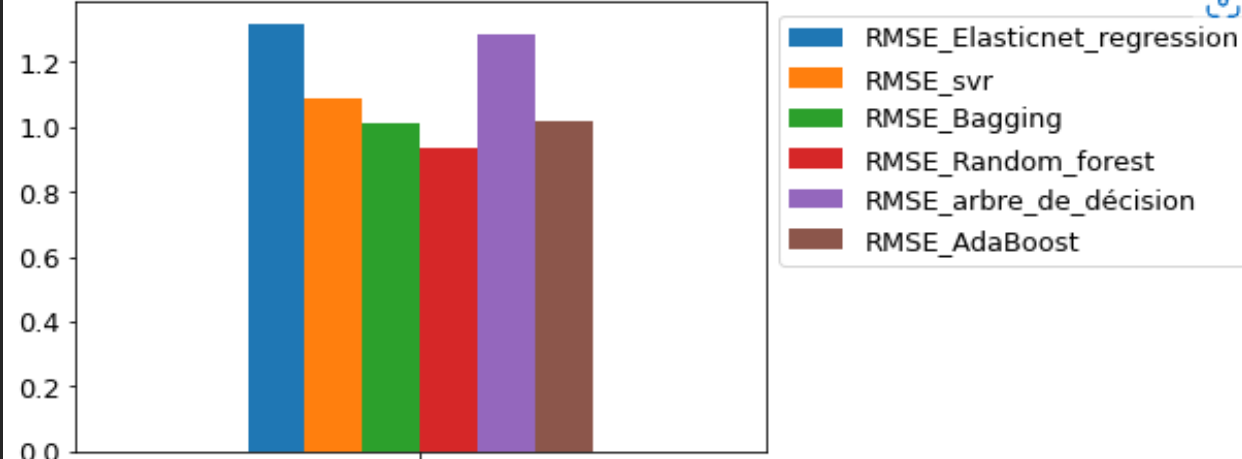
Résultats et comparaison des modèles

■ Prédiction de l'émission de CO2:

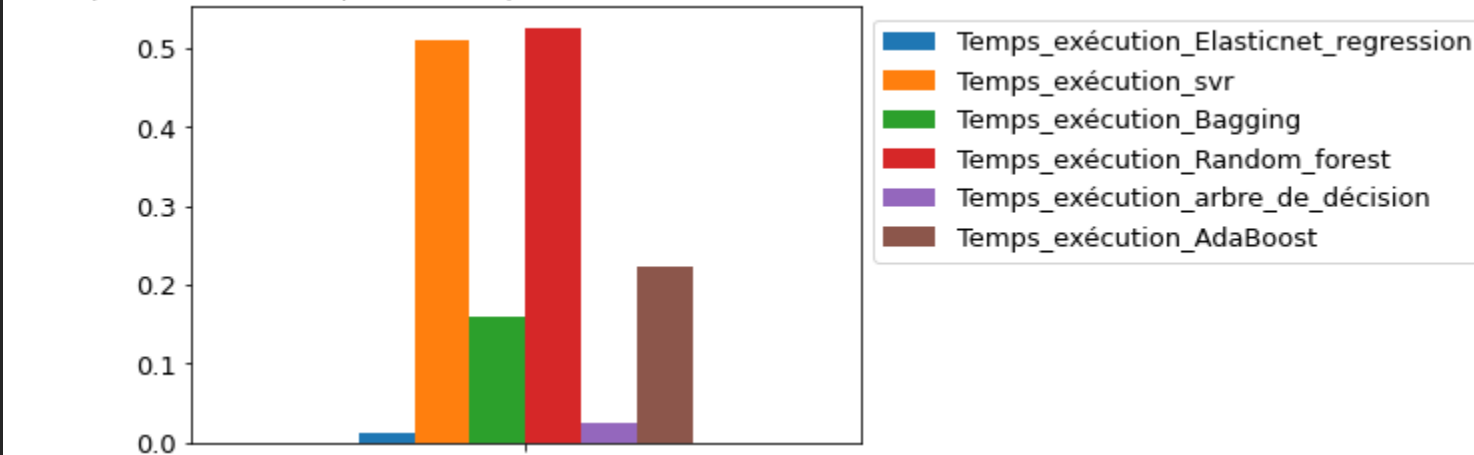
Le score R² obtenu par modèle de prédiction des émissions de CO2



La RMSE obtenue par modèle de prédiction des émissions de CO2

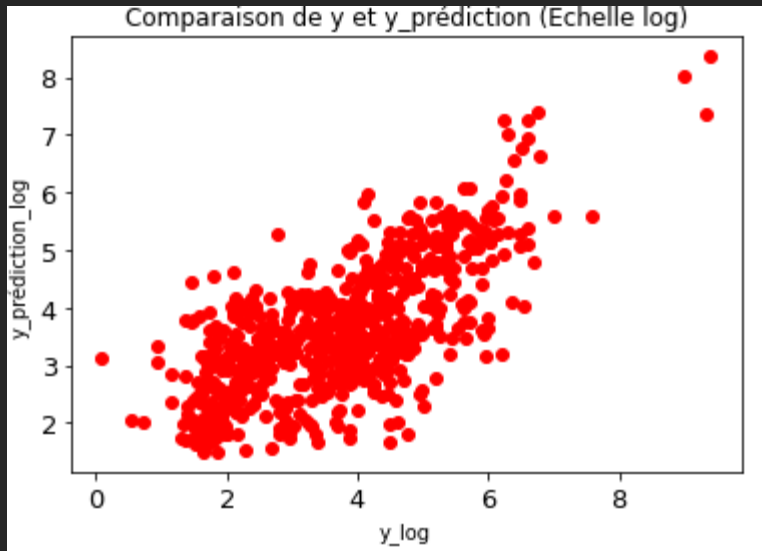


Le temps d'exécution de chaque modèle de prédiction des émissions de CO2

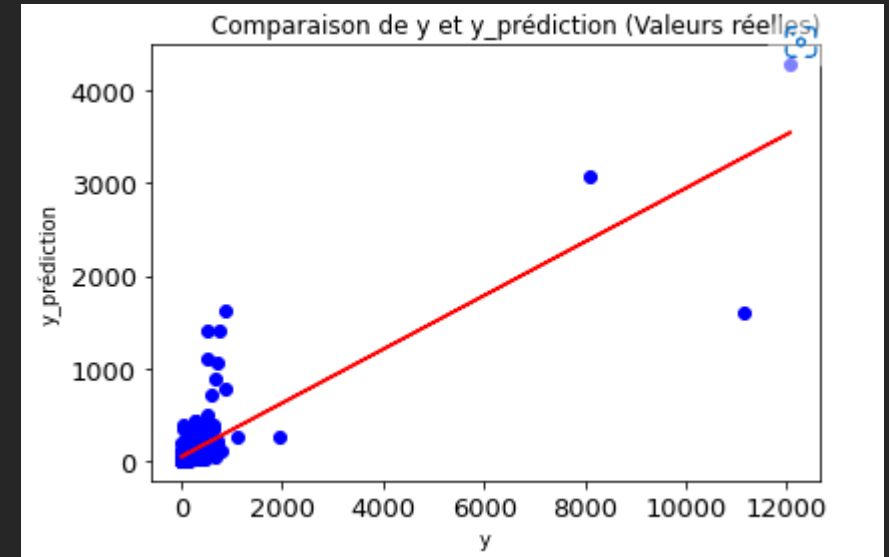


Prédiction par le modèle des forêts aléatoires

- Prédiction des émissions de CO₂:



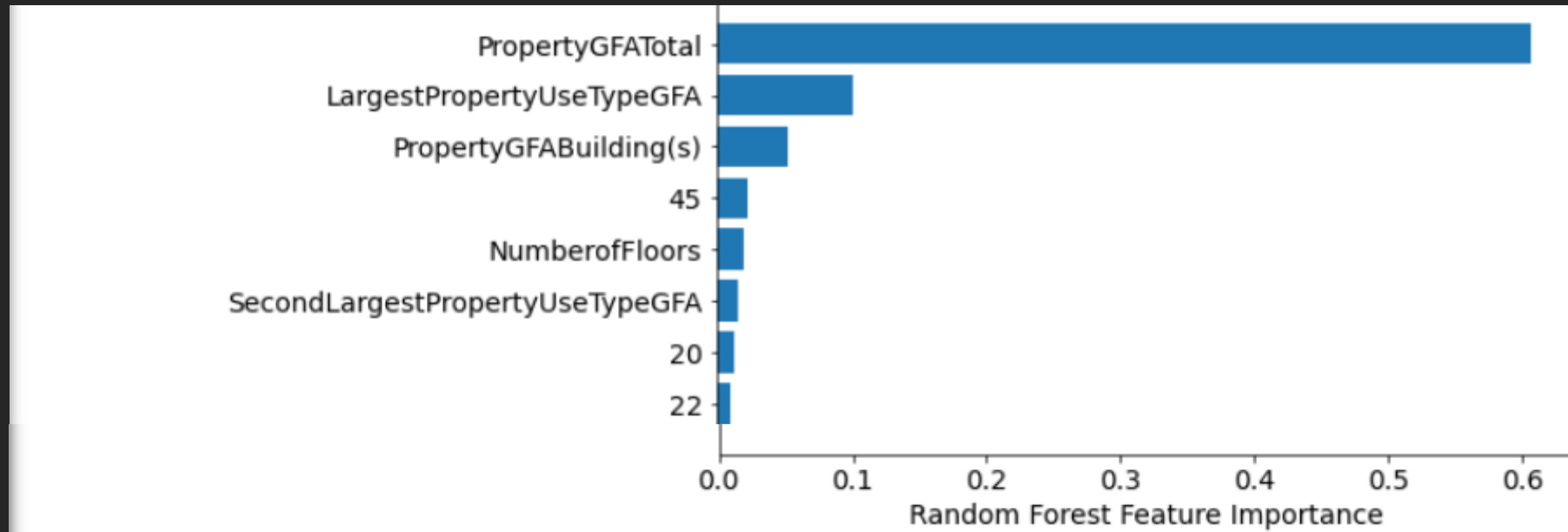
exponentielle



- Limites de notre modèle : Bonnes prédictions pour un certain domaine de quantité de gaz émise (valeurs inférieures à 1000 MetricToneCO₂e)

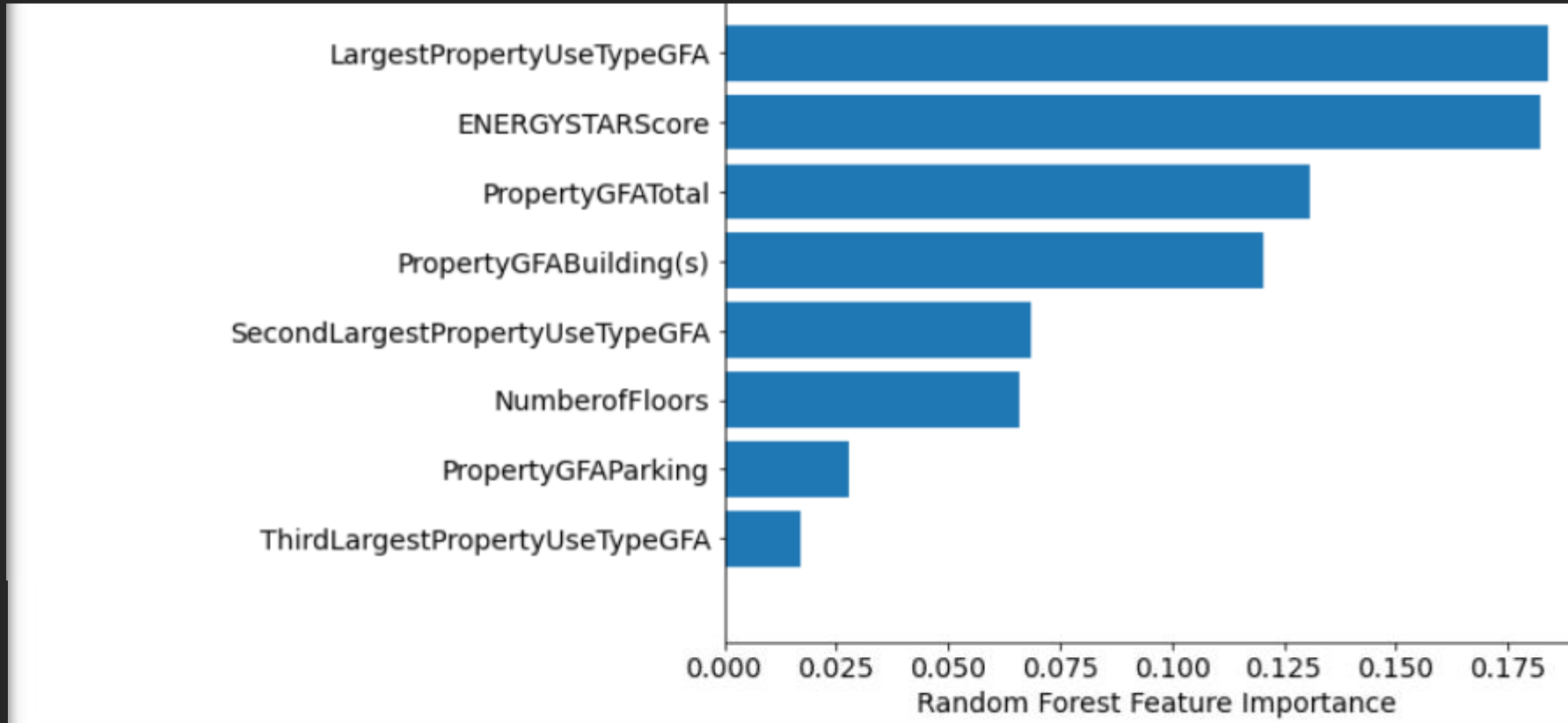
Importance des variables dans notre modèle

- Prédiction des émissions de la consommation énergétique:
 - Modèle sans l' Energy Star Score:



Importance des variables dans notre modèle

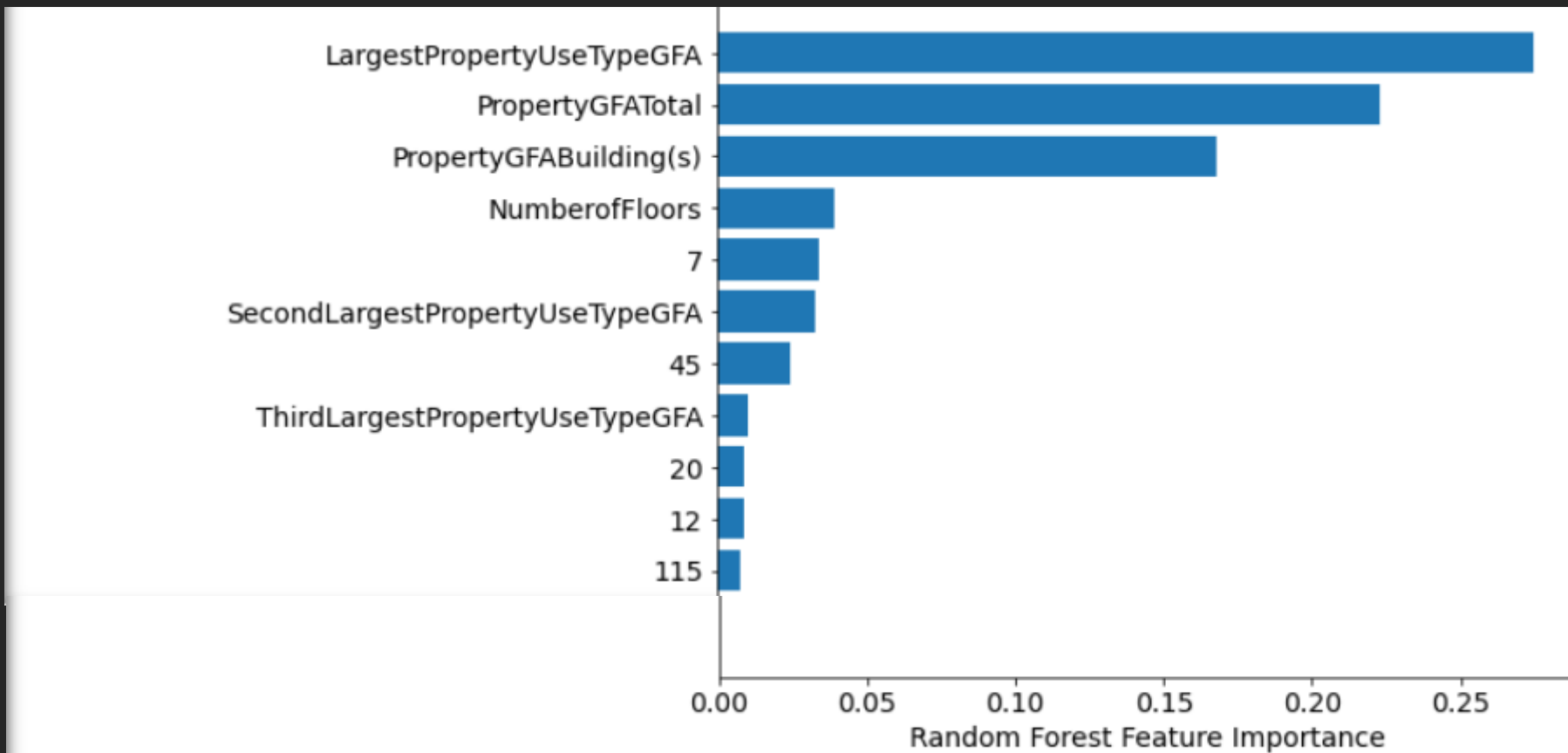
- Prédiction des émissions de la consommation énergétique:
 - Modèle avec l' Energy Star Score:



- Utilisation importante de l'Energy Stare Score par le modèle

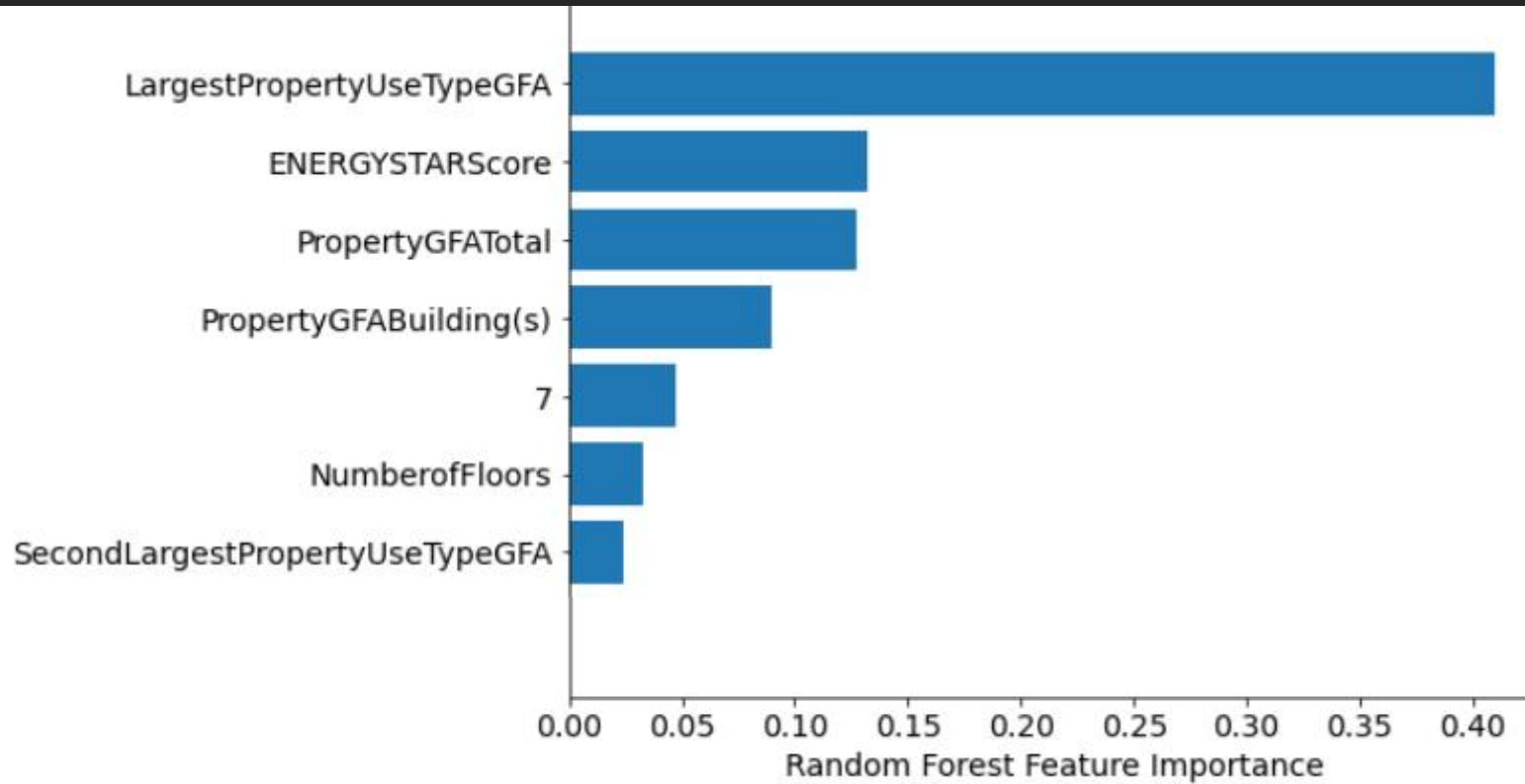
Importance des variables dans notre modèle

- Prédiction des émissions de CO₂:
 - Modèle sans l' Energy Star Score:



Résultats et comparaison des modèles

- Prédiction des émissions de CO₂:
 - Modèle avec l' Energy Star Score:



- Utilisation importante de l'Energy Stare Score par le modèle

Comparaison de la performance du modèle avec et sans l'Energy star score

	Prédiction de l'énergie		Prédiction des émissions de CO ₂	
	Sans l'Energy Star Score	Avec l'Energy Star Score	Sans l'Energy Star Score	Avec l'Energy Star Score
R ²	0.72	0.82	0.57	0.62

Conclusions

- Bonnes prédictions de notre modèle sur un domaine où les valeurs énergétiques et les quantités de carbone émises ne sont pas trop grandes.
- Possibilité d'améliorer le modèle:
 - en affinant davantage les hyperparamètres.
 - en lui fournissant plus données d'autres villes ayant des similitudes avec Seattle...
- Ouvrir d'autres pistes de travail : développer et associer des modèles différents pouvant être performants sur des domaines bien précis.
- L'Energy Star Score est utilisé de manière importante par le modèle et il a permis une légère amélioration de sa performance.