Anticipez les besoins en consommation électrique de bâtiments

Photo by Varadh Jain on Unsplash

Projet 4 du parcours
« Data Scientist » d'OpenClassrooms

Mark Creasey

Sommaire

01 Présentation de la problématique

Mission

À partir des relevés de 2015 et 2016

Prédire pour des nouveaux bâtiments commerciaux

- consommation totale d'énergie
- émissions de CO2

Évaluer l'intérêt de l'ENERGY STAR Score pour la prédiction d'émissions

Contraintes

- basé sur les données déclaratives du permis d'exploitation commerciale
  - taille des bâtiments
  - usage des bâtiments
  - mention de travaux récents,
  - date de construction

Seattle

Interprétation de la problématique

Cibles à prédire

- SiteEnergyUse(kBtu)
- TotalGHGEmissions

Cibles alternatifs

- A. divisé par superficie
  - SiteEUI(kBtu/sf)
  - GHGEmissionsIntensity
- B. normalisé par le météo de chaque année

Variables indépendants

- Localisation (Lat, Lon, Adresse)
- Physique (étages, année construction)
- Types d'usage (1er, 2ème 3ème)
- Superficies pour chaque type d'usage
- ENERGYSTARScore
- Consommations électriques, gaz, vapeur

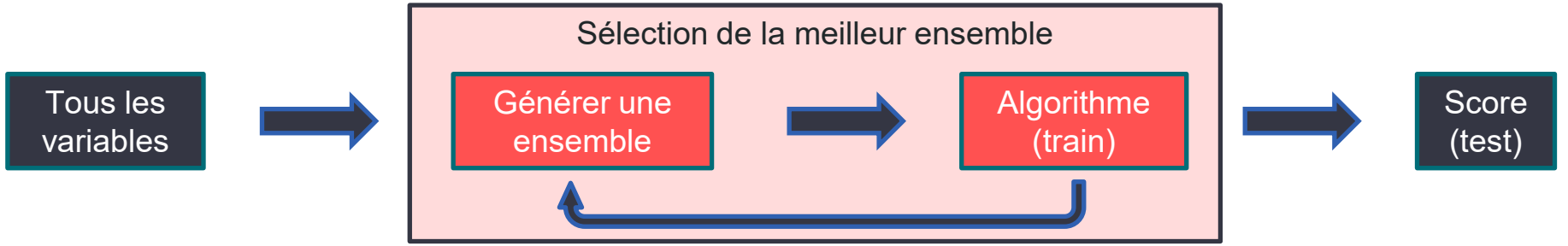
Pistes de recherche envisagées

Régression des variables non-colinéaires via **réduction de dimensions**

- **Filtrer** (Corrélation, VIF, KBest)

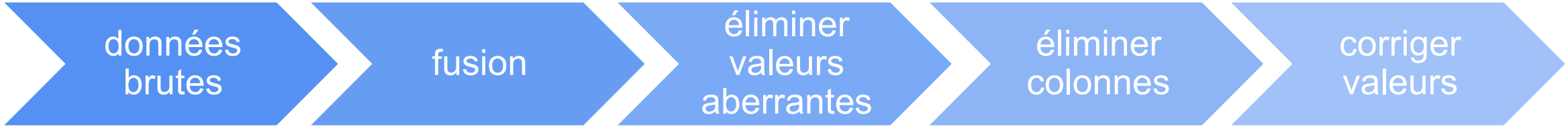
- **Embedded** (régularisation L1,L2, décision tree)

- **Wrapper** (RFE)



02 Nettoyage et analyse exploratoire

Nettoyage des données



(3340, 46) + (3376, 47)

(6568, 51)

(2986,51)

(2986, 27)

(2986, 27)

Données 2015

renommer colonnes split location

> 97% bâtiments en commun

Données 2016

3376 bâtiments

Concaténer données 2015 et 2016

Supprimer bâtiments non communs

Corriger datatypes

Bâtiments résidentiels

non-compliant

> 80% variation de consommation entre années

Données manquantes

non-disponibles

au moment de prédire pour des nouveaux batiments

(consommation électrique, gaz)

non-pertinentes

trop vides

Whitespace

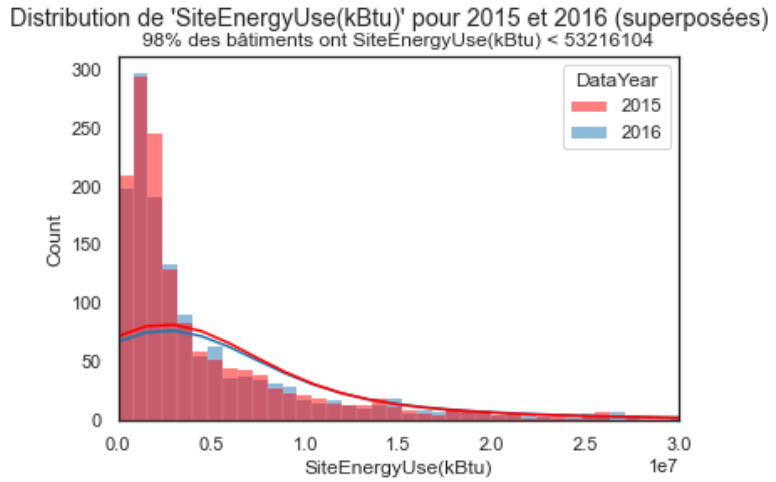
Majuscules

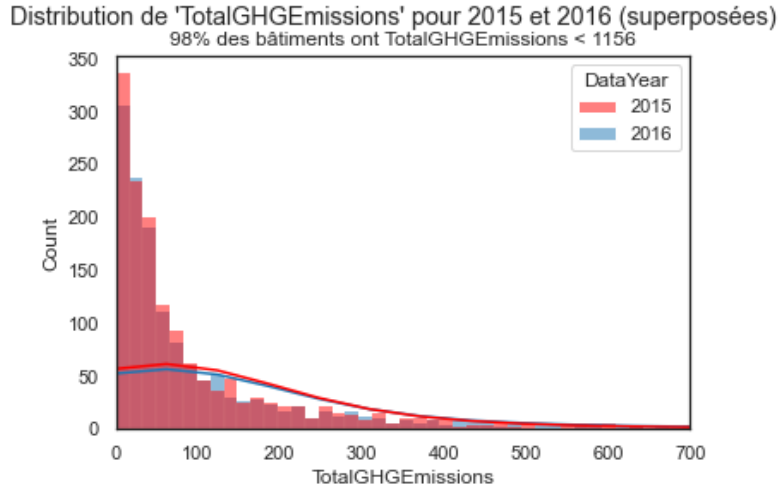
Fillna('None')
Fillna(0)

(Second et Third Largest Use Types)

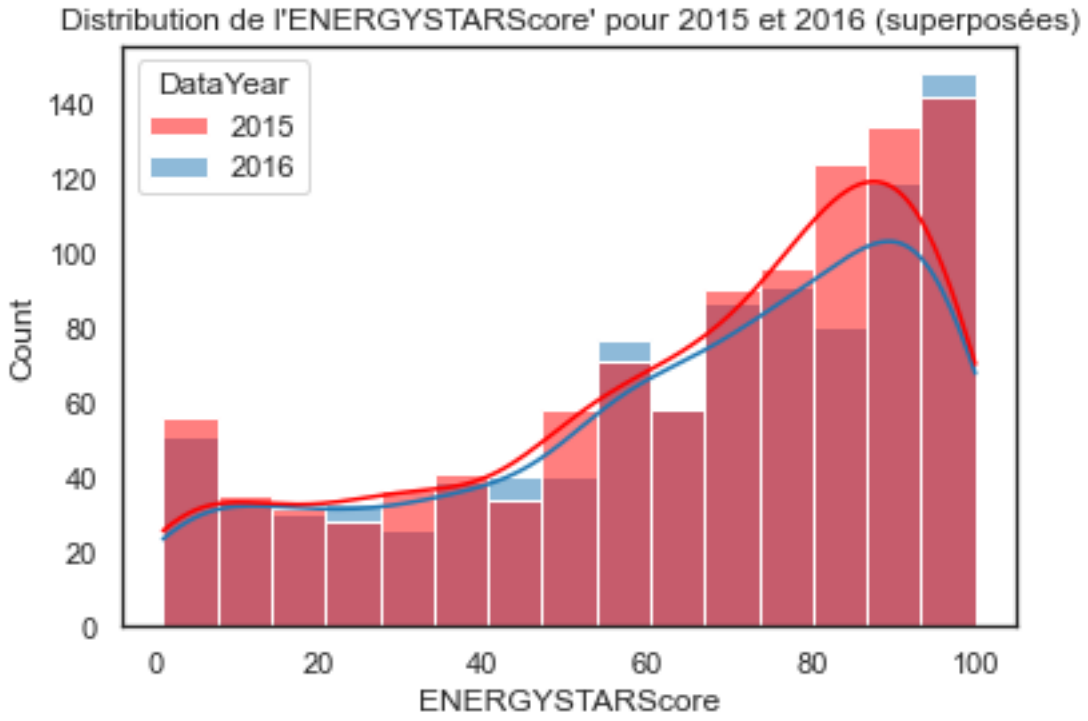
Analyse Exploratoire – non-linearités

- Variables cibles non-linéaires

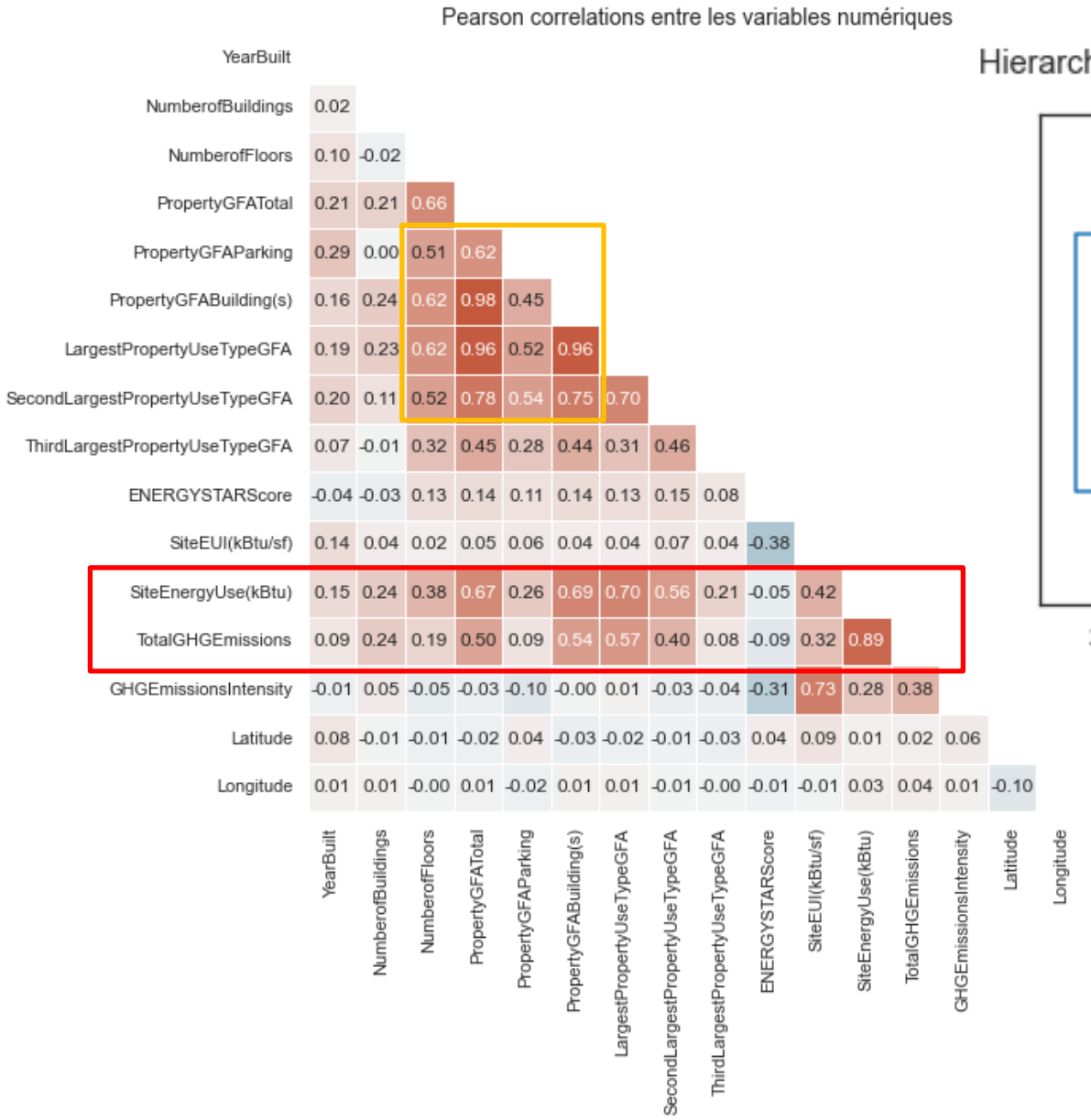


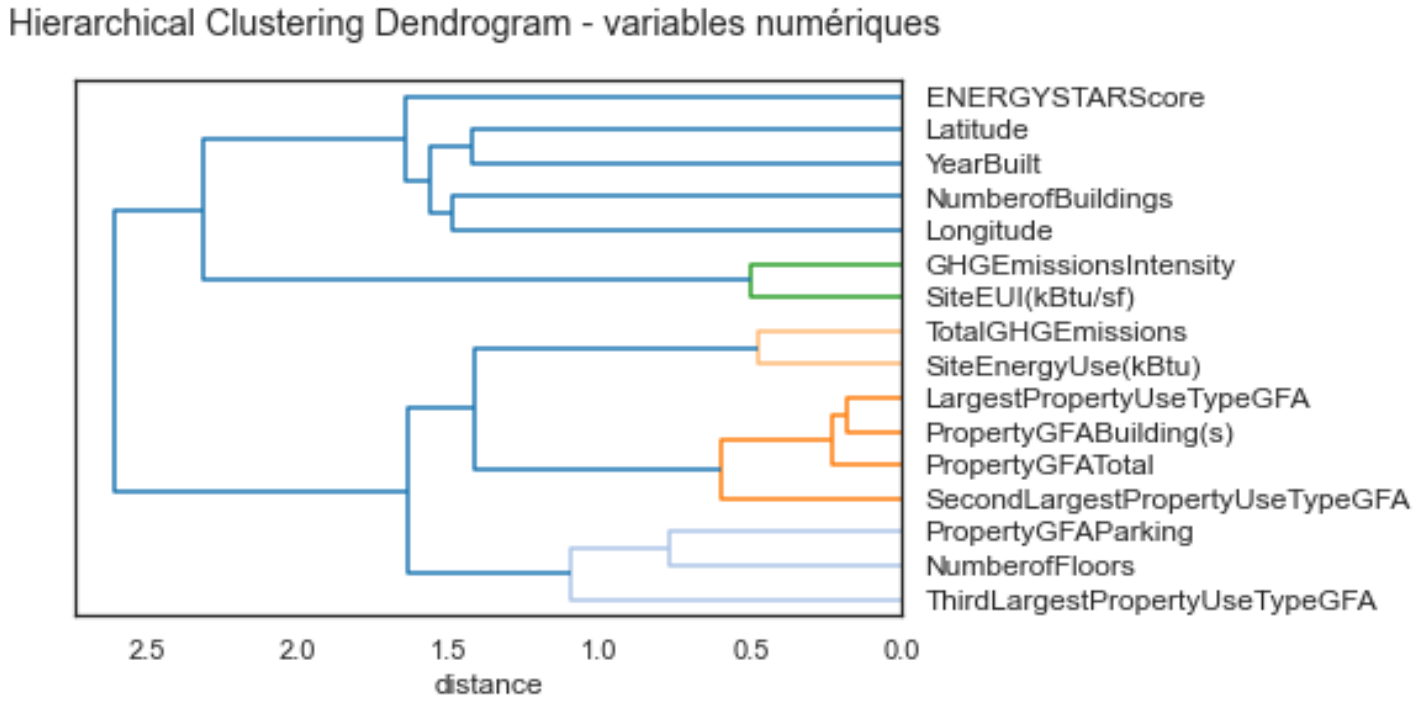


- Variables indépendantes avec distribution loin de normale

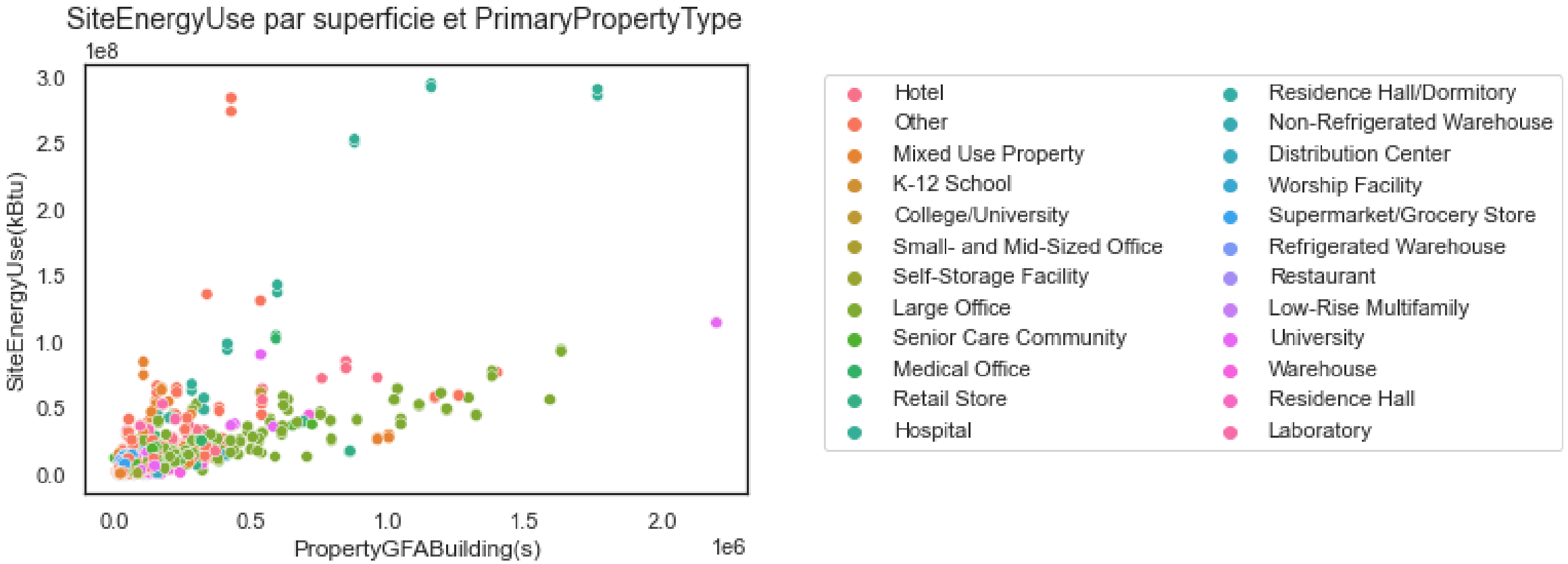


Analyse Exploratoire – Colinéarités entre colonnes



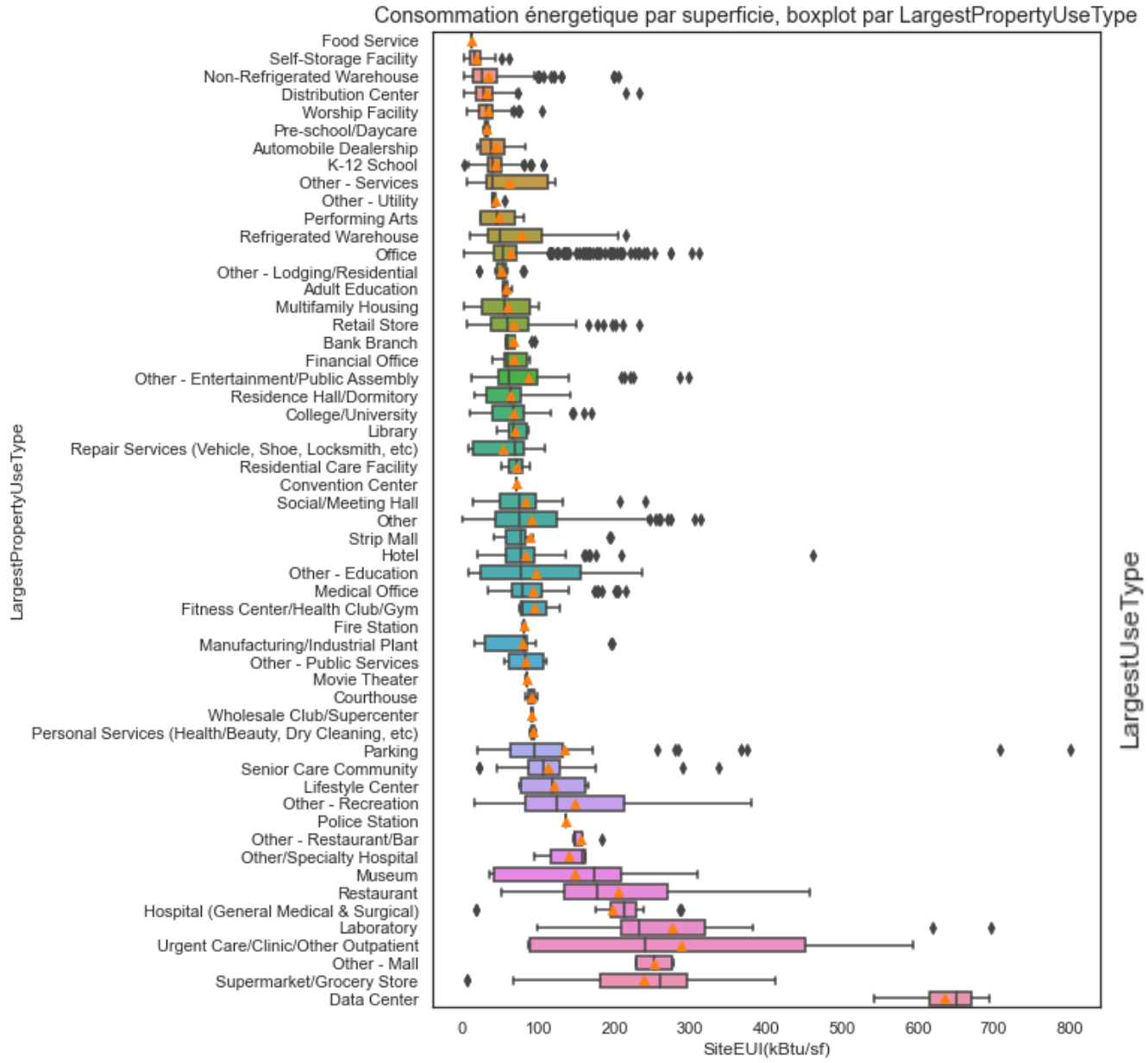


Analyse Exploratoire – Catégories sont importantes

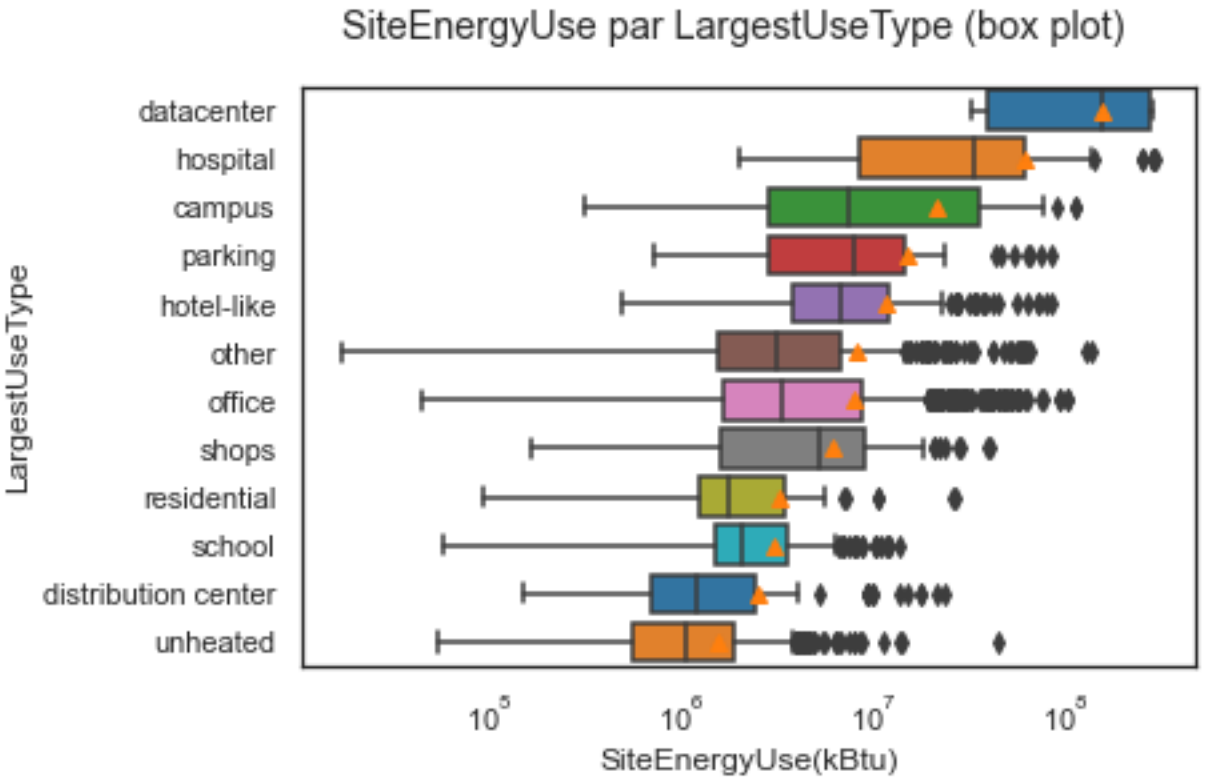


03 Feature Engineering

Réduction de la dimensionnalité des catégories existants :

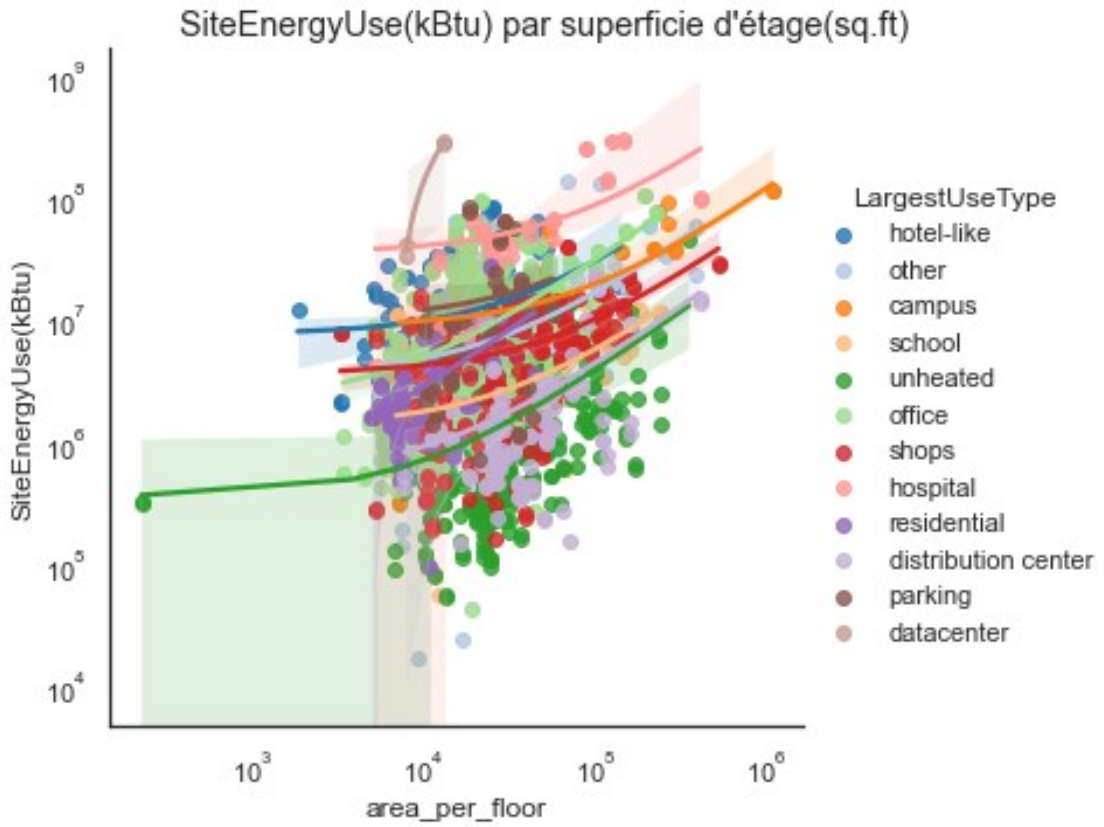


De 55 à 10 PropertyUseType



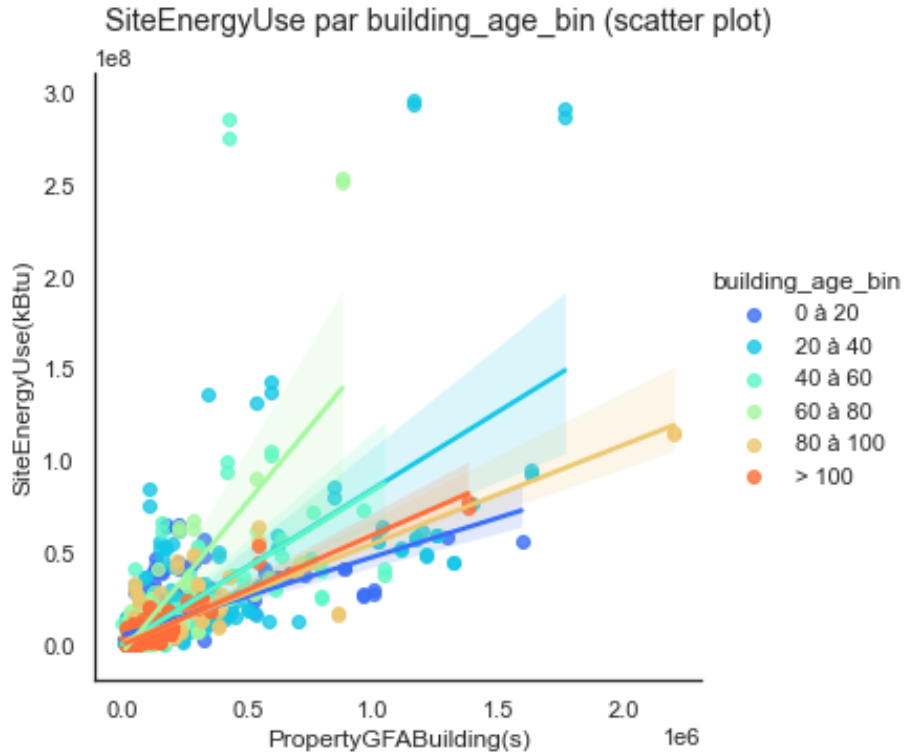
Nouvelles variables non colinéaires

- **Superficie par étage**



Age du bâtiment

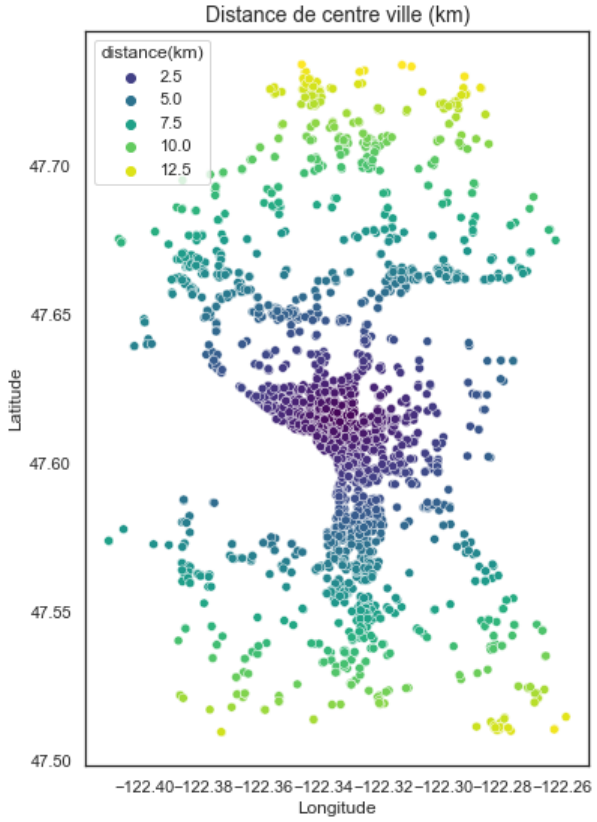
- Consommation énergétique est plus bas pour bâtiments <20 ans et >80 ans

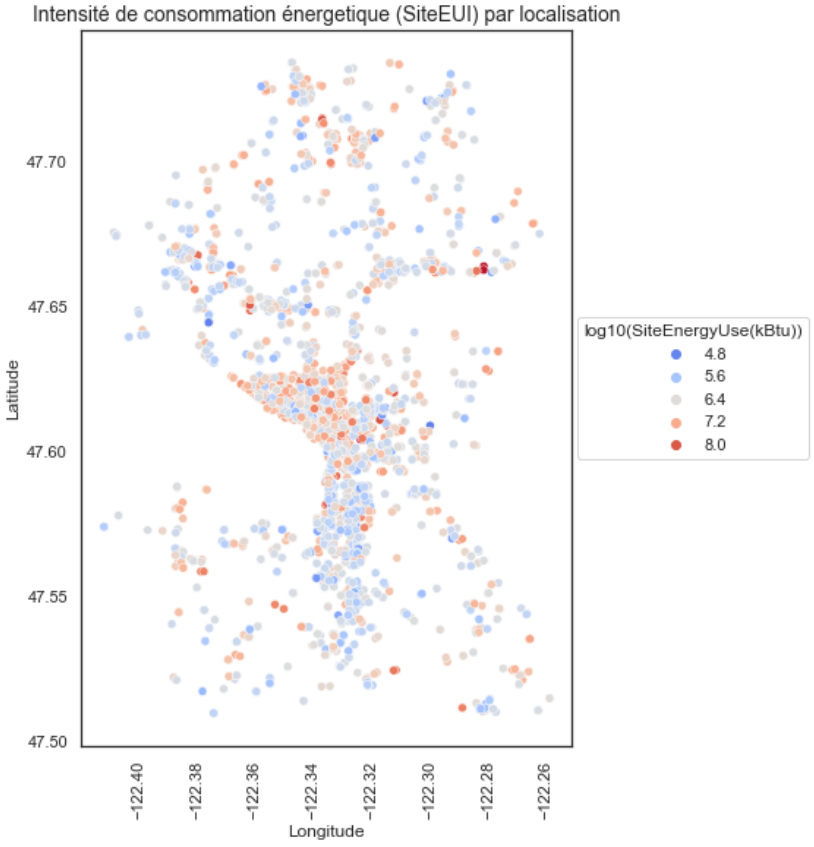


Reduction de dimensionalité de location

Distance de centre ville

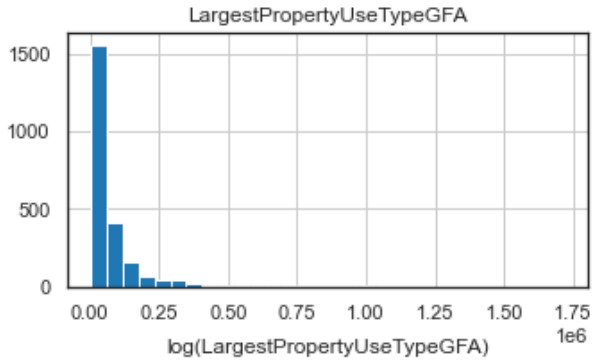
- Proxy pour zip code
- Proxy pour densité des bâtiments

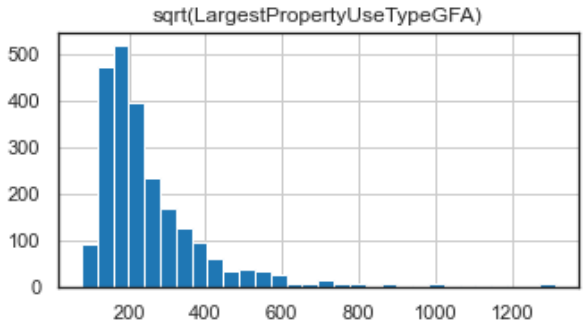


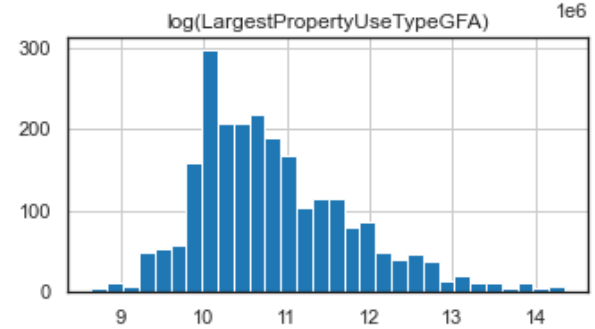


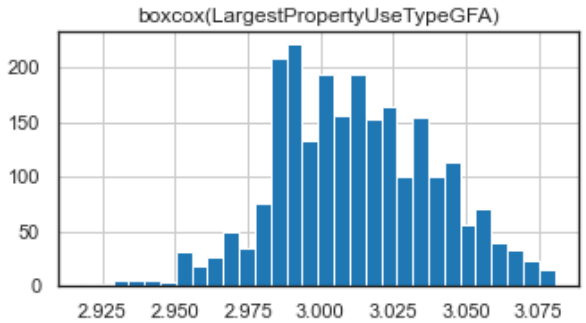
Préprocessing / Feature Sélection

Pre-processing des données - Transformations





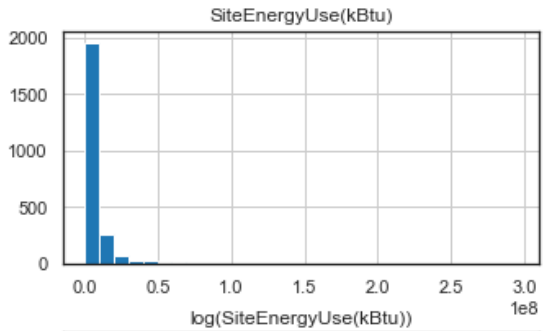


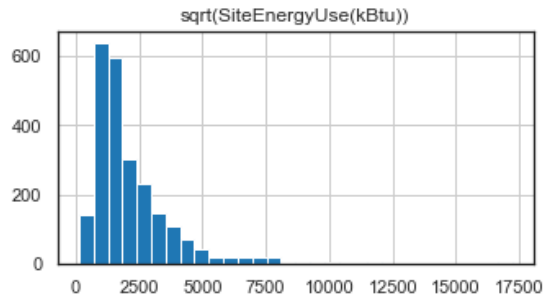


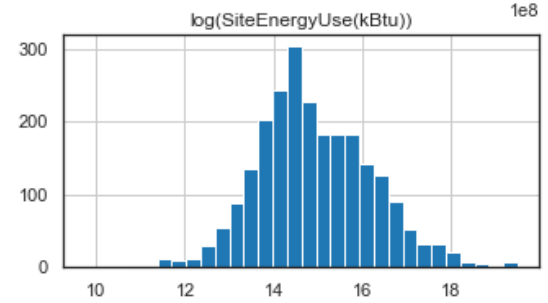
Transformation

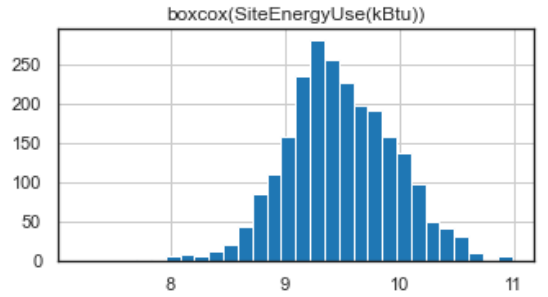
- des variables X
- des cibles Y

	feature.skew()	square	sqrt	log	box-cox
NumberofBuildings	20.118523	33.282573	8.104272	6.608768	2.275358
PropertyGFAParking	5.370409	12.248061	2.637591	1.407786	1.353876
LargestPropertyUseTypeGFA	5.507316	10.613578	2.731488	0.864747	0.032017
SecondLargestPropertyUseTypeGFA	5.229551	13.209216	2.145091	0.149621	0.085003
ThirdLargestPropertyUseTypeGFA	12.904976	24.310708	4.063160	1.572207	1.493188
building_age	0.291364	0.876201	0.127441	0.625899	0.114117
area_per_floor	10.369995	28.811076	3.160541	0.581321	0.100265
distance(km)	0.698303	1.570299	0.168465	0.081293	0.096709



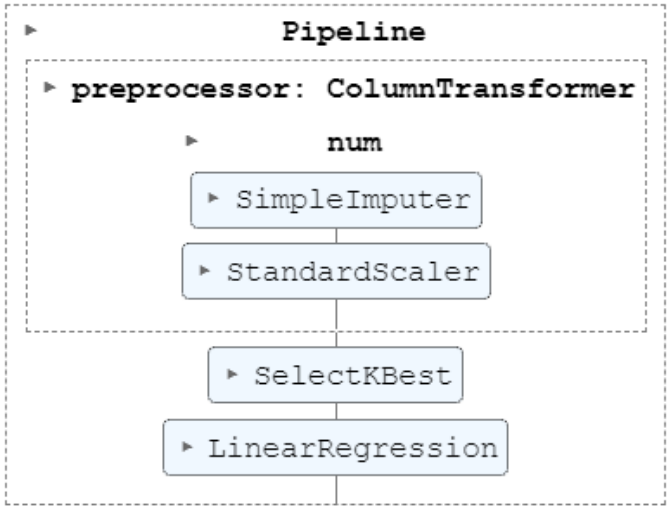




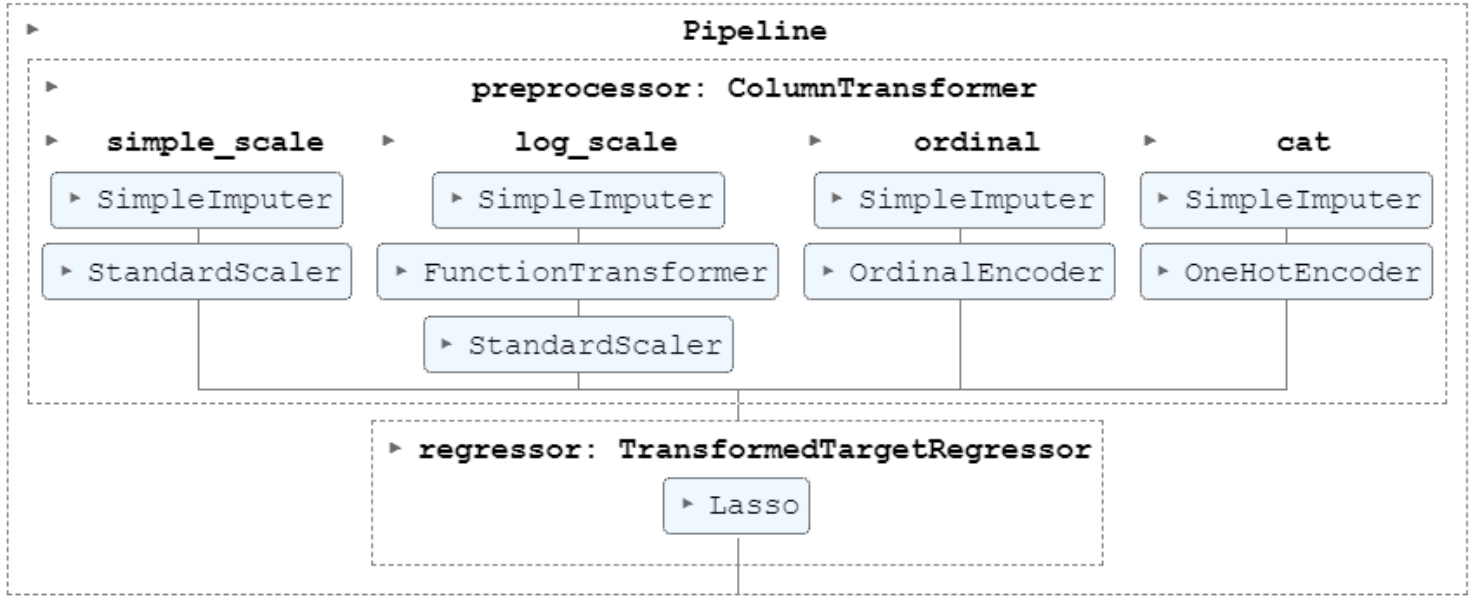


Pre-processing des données

De plus simple ...



... a plus complexe



Techniques de sélection des « features » utilisés

Filter

(pre-processing)
utilise des indicateurs statistiques
est rapide

Entre variables numériques:

- Variance Inflation Factor
- Pearson Corrélation

Entre variables catégoriques

- Cramer's V (Chi-squared)
- Thiel's U (Entropie conditionnel)

Embedded

(sélection par le modèle)

- L1 régularisation (Lasso,
- L2 régularisation (Ridge)
- Feature importance (arbres de décision)

Wrapper

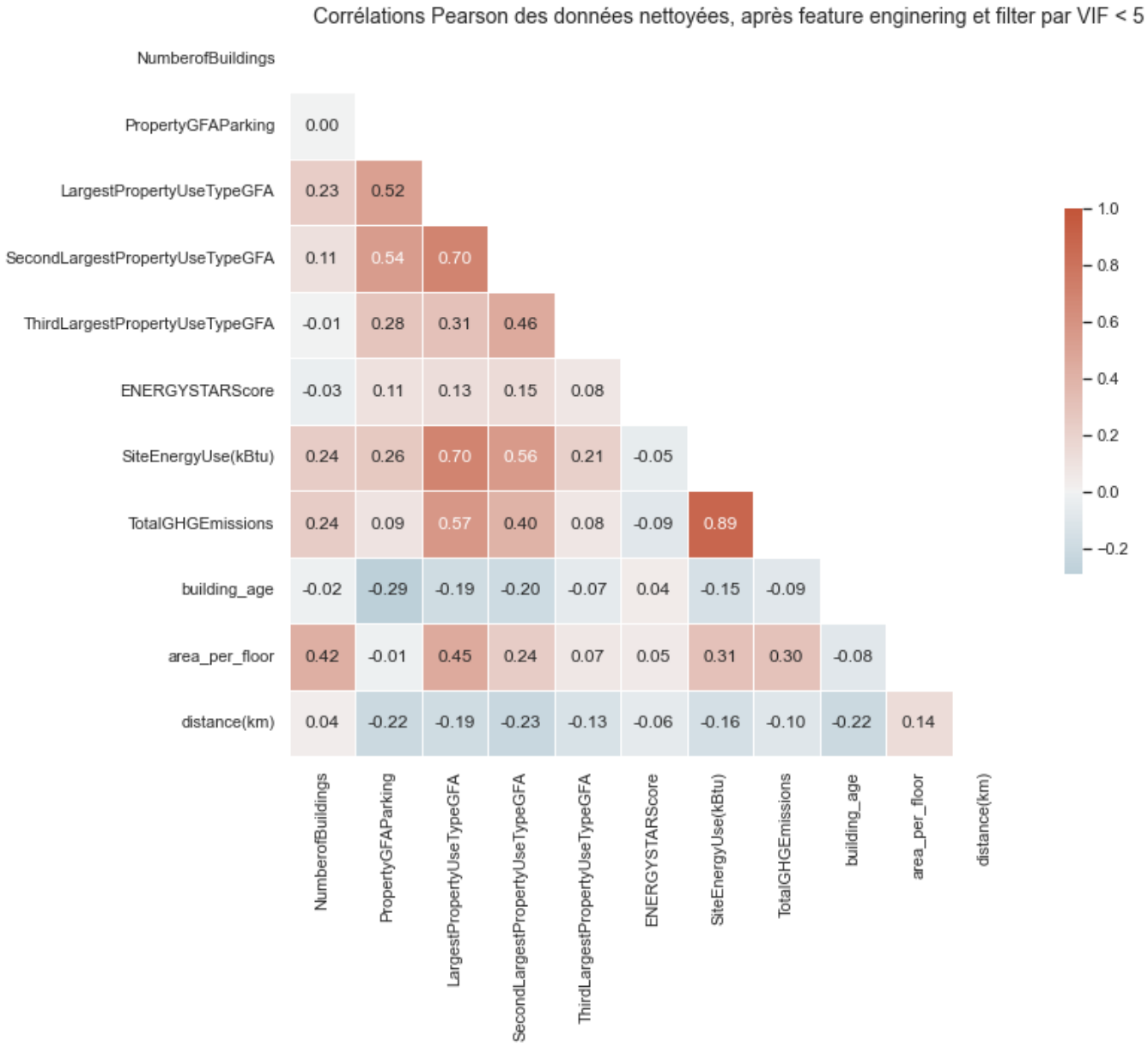
(sélection pendant l'entrainement)

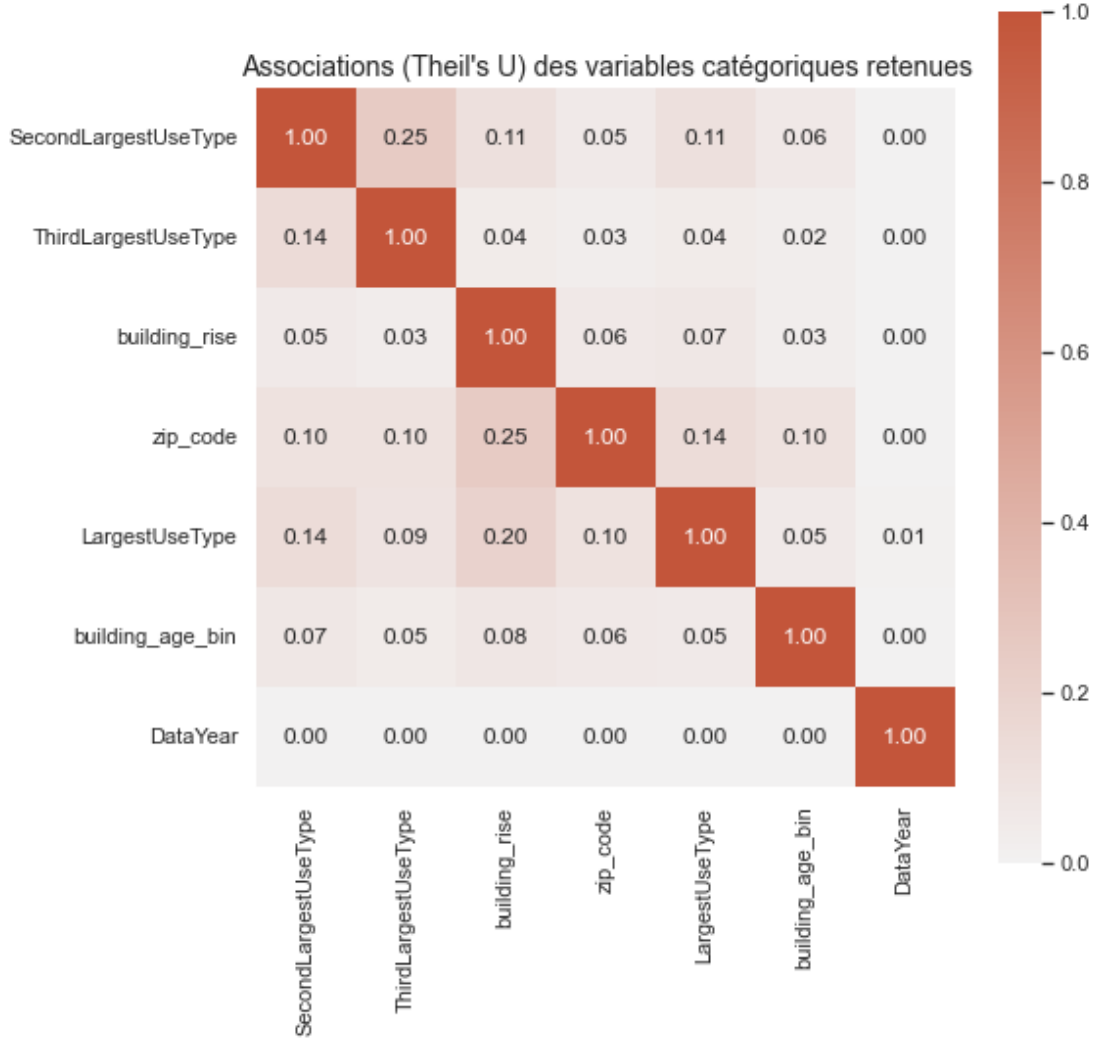
- KBestFeatures
- Recursive Feature Elimination (RFE)

Pourquoi sélectionner les « features »

- Simplification de la modèle (plus facile d’interpreter)
- Amélioration de la confiance de prévision
- Réduction de risque d‘overfit’ (high variance)
- Accélerer le temps d’entrainement

Feature sélection par corrélations



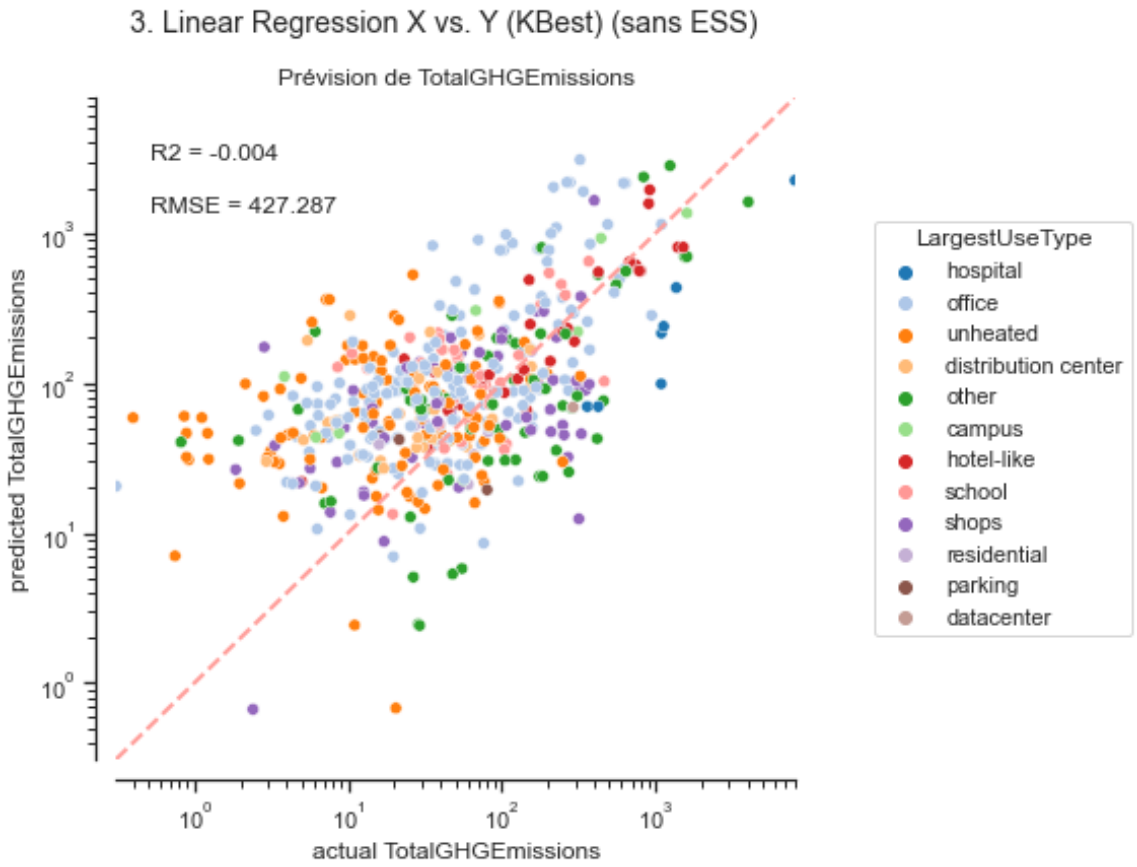


04. Modélisation effectuées

Modèles linéaires : Emissions CO2

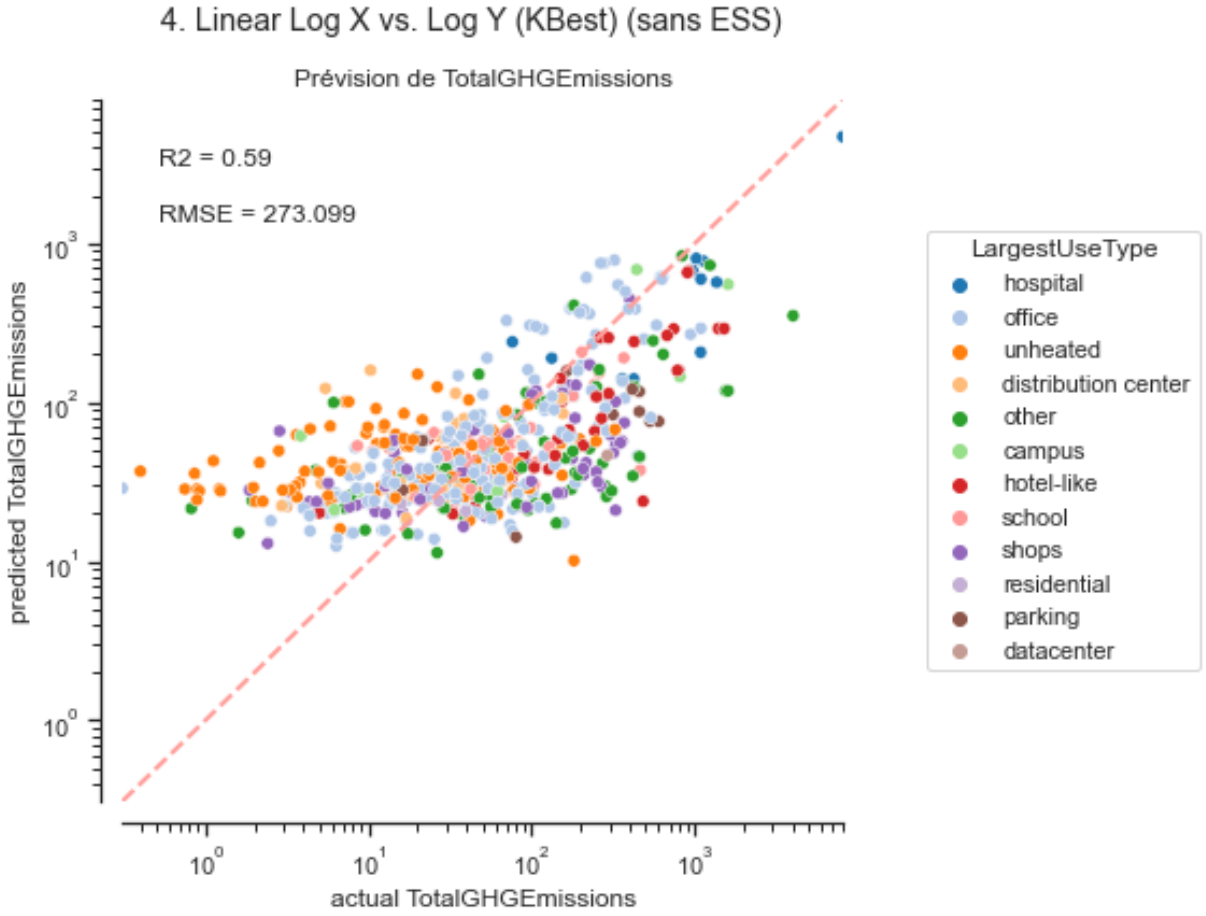
X vs Y

- R2 = 0.0



Log X vs log Y

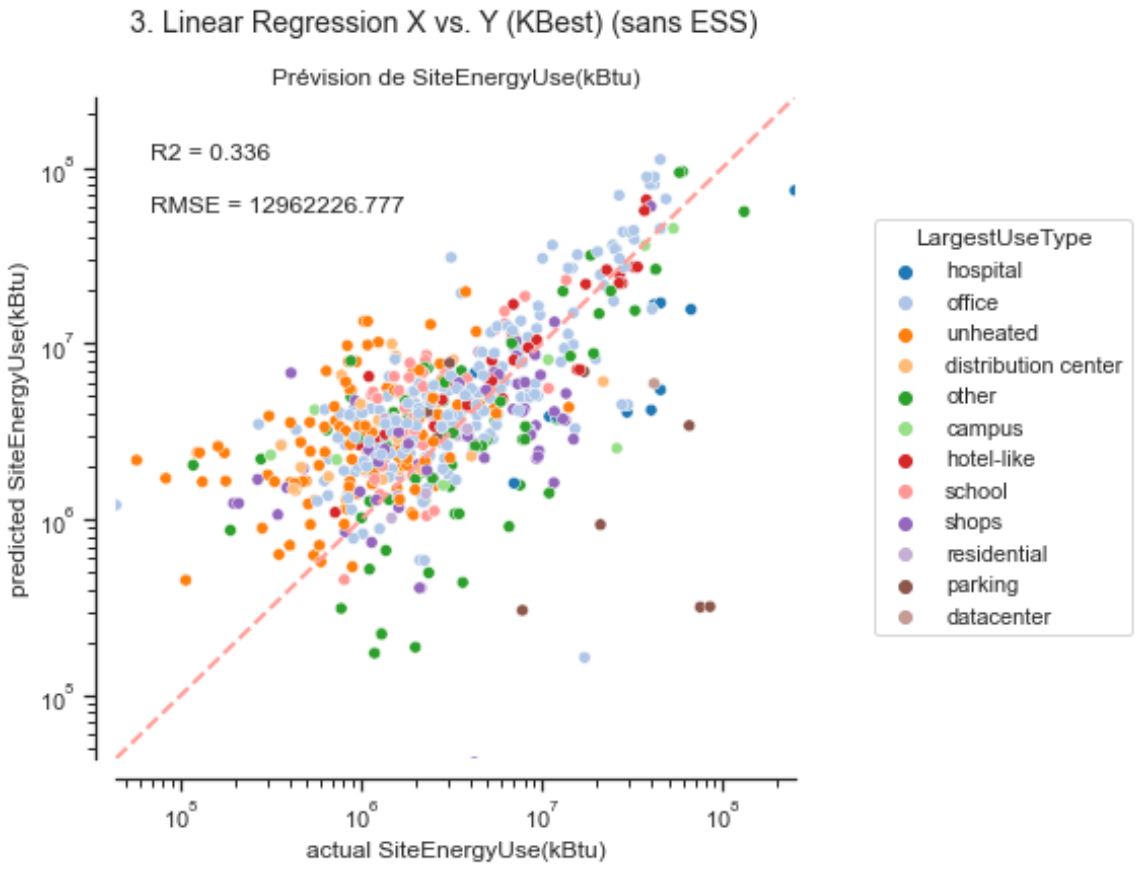
- R2 = 0.59



Modèles linéaires : Consommation Energétique

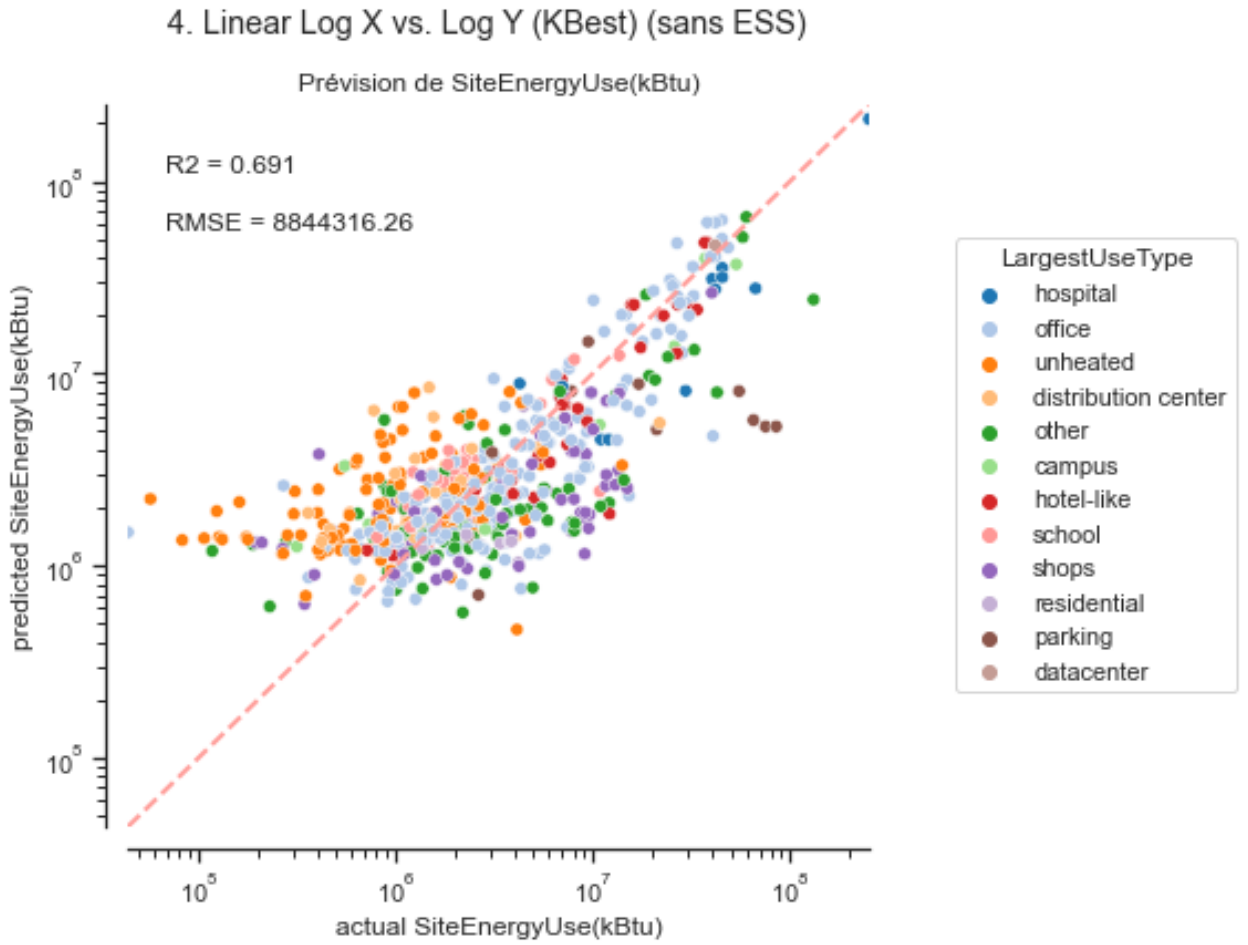
X vs Y

- R2 = 0.34



Log X vs log Y

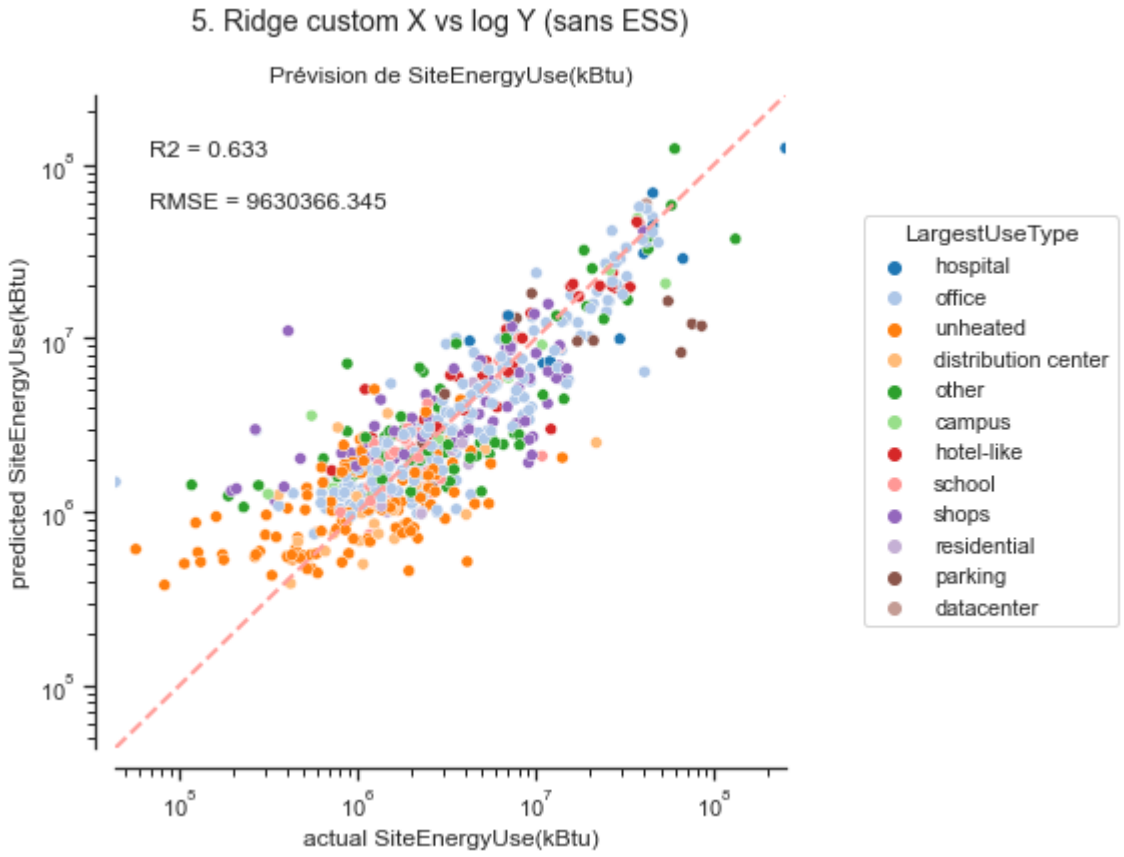
- R2 = 0.69



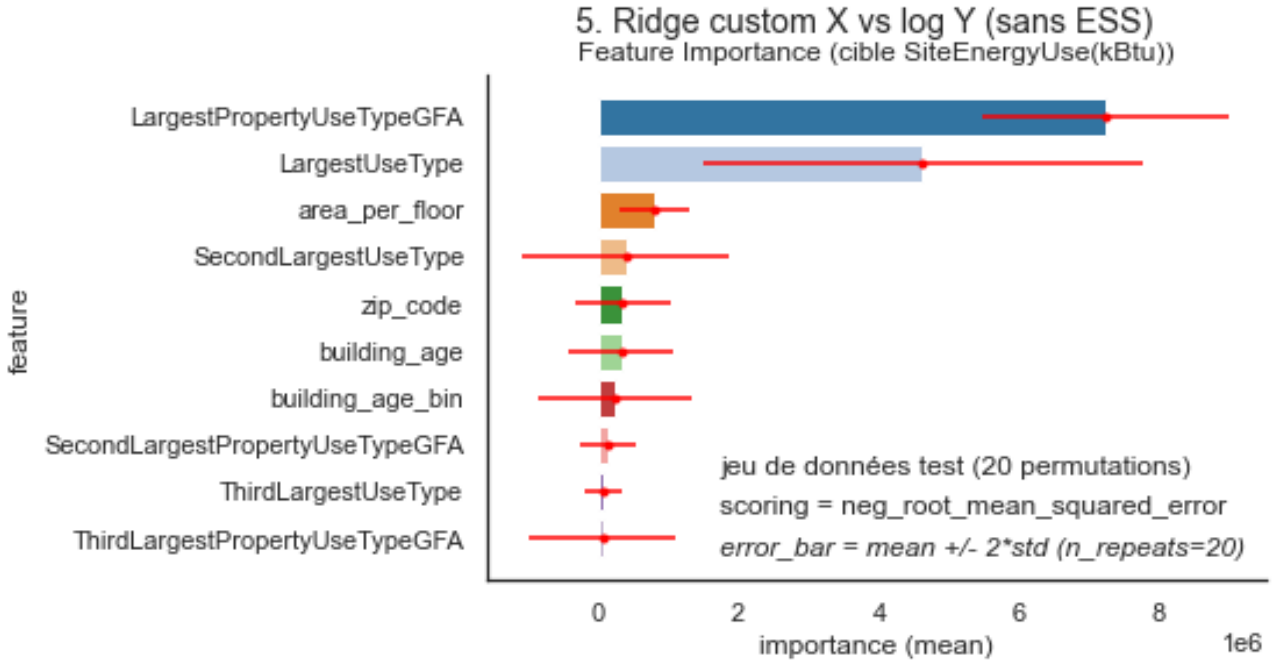
Régularisation L2 (Ridge) et L1 (Lasso)

Custom X vs log Y

- R2 = 0.63



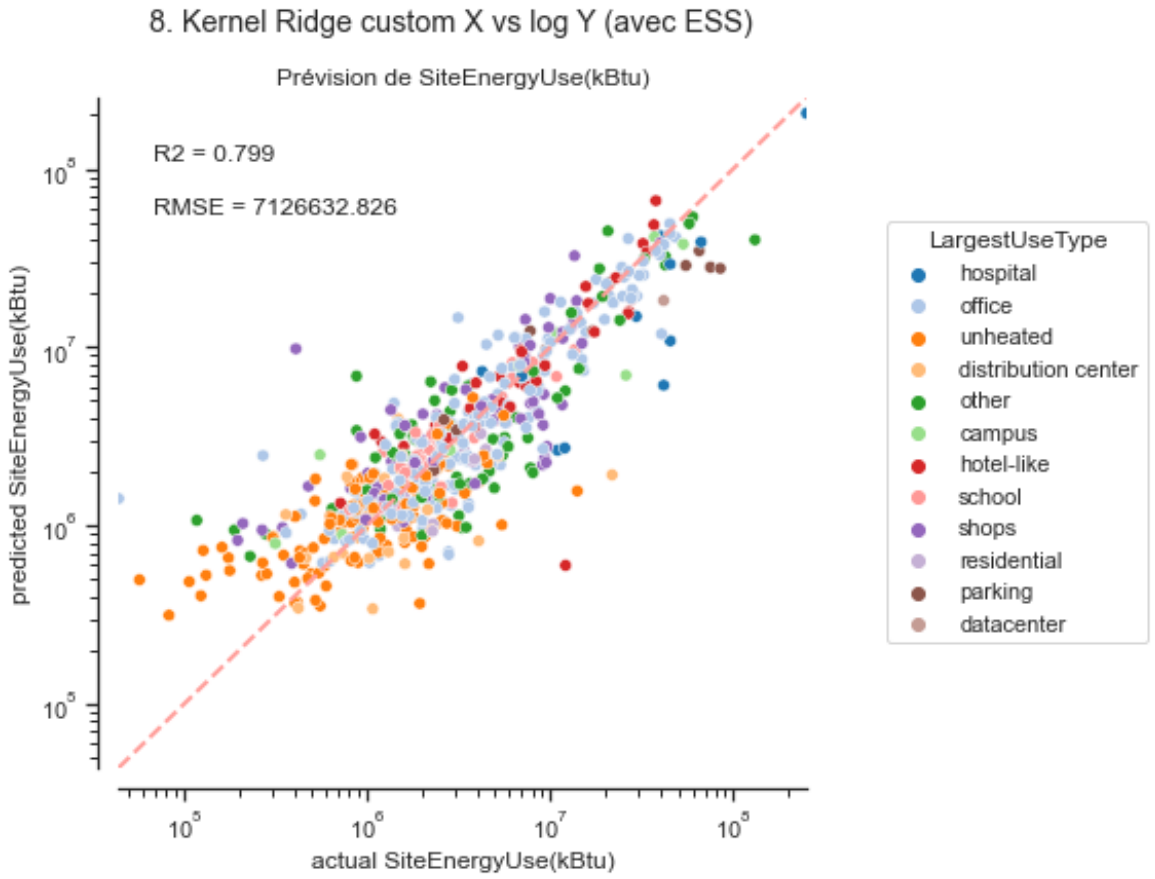
Feature Importance



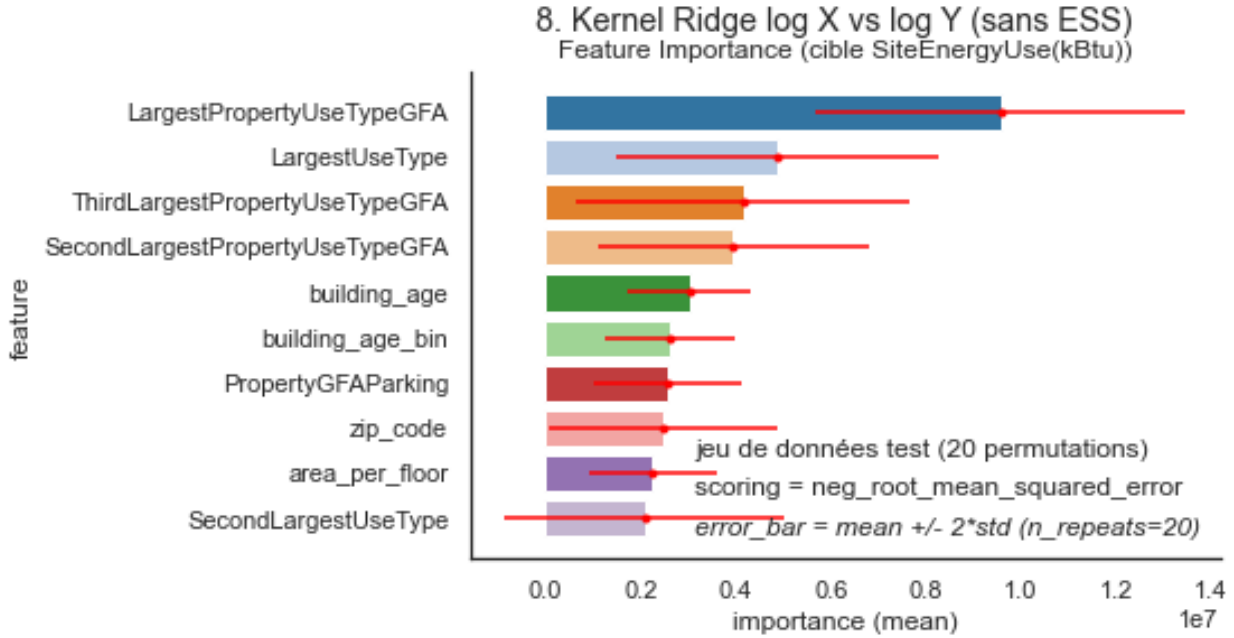
Modèles non-linéaires (SVR, Kernel Ridge)

Log X vs log Y

- R2 = 0.80



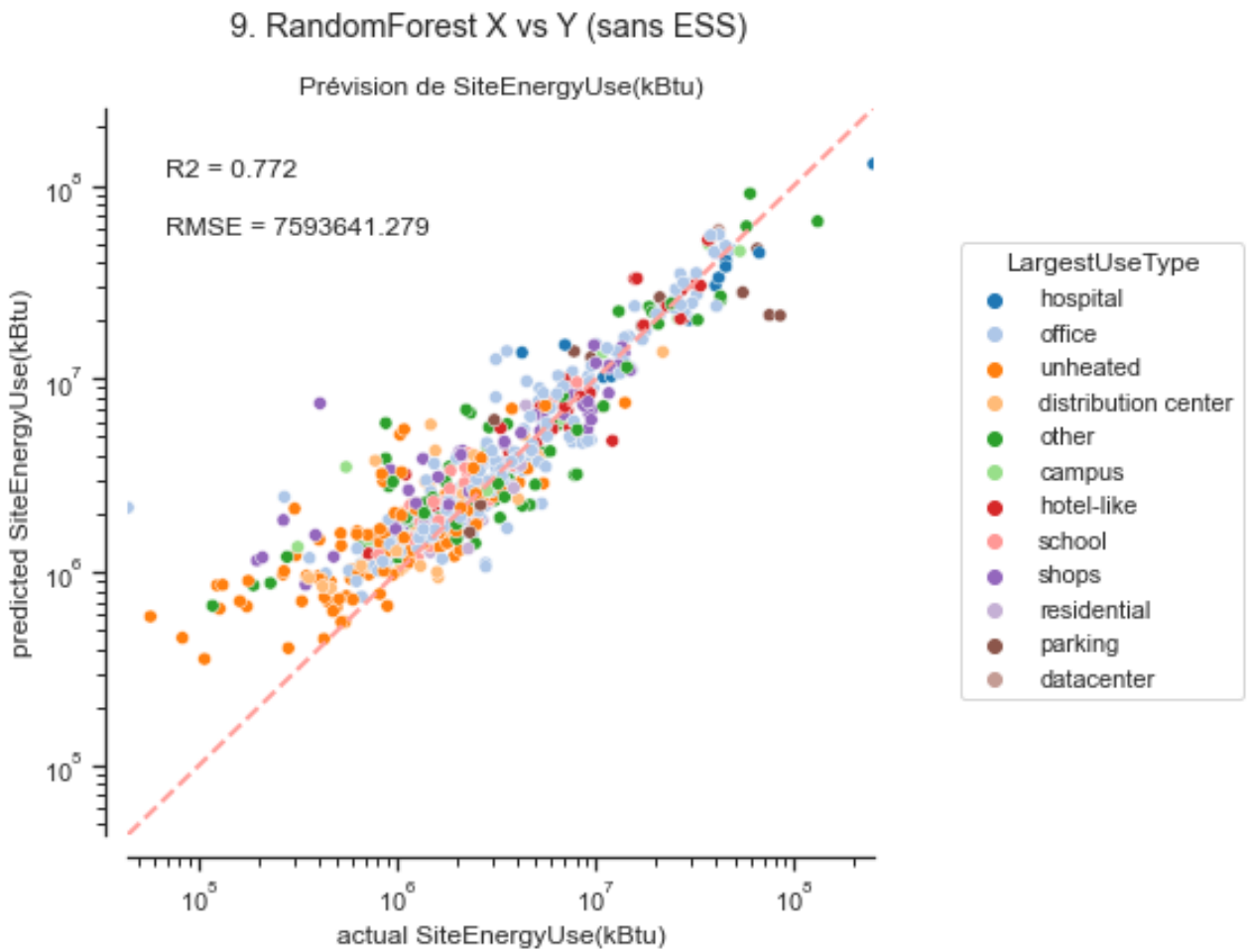
Feature Importance



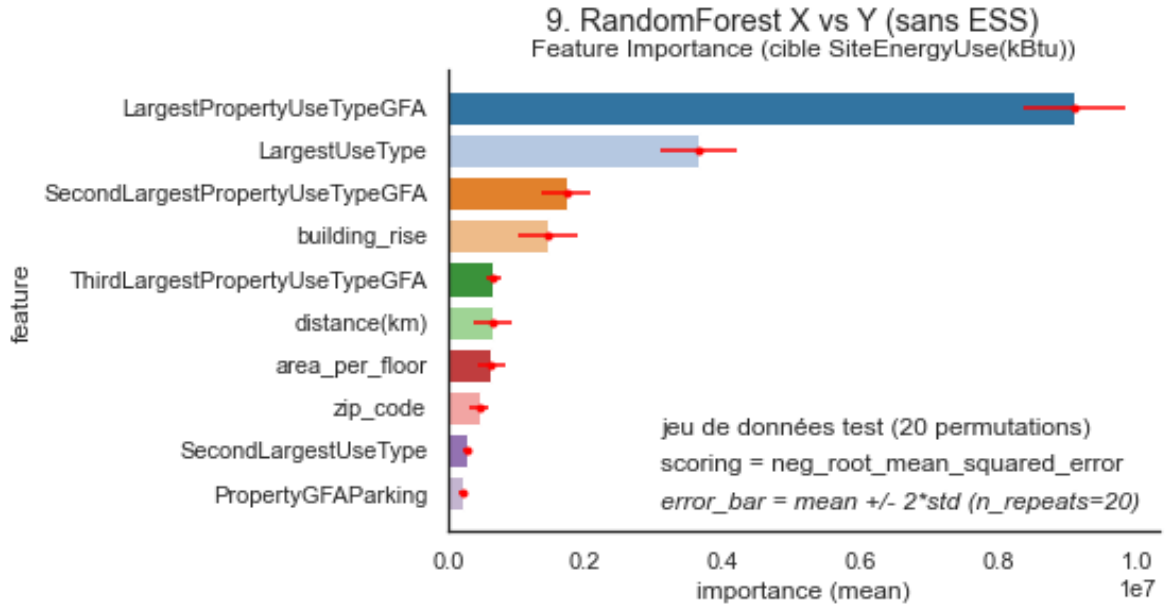
Modèles ensemblistes (RandomForest, Bagging)

Log X vs log Y

- R2 = 0.77

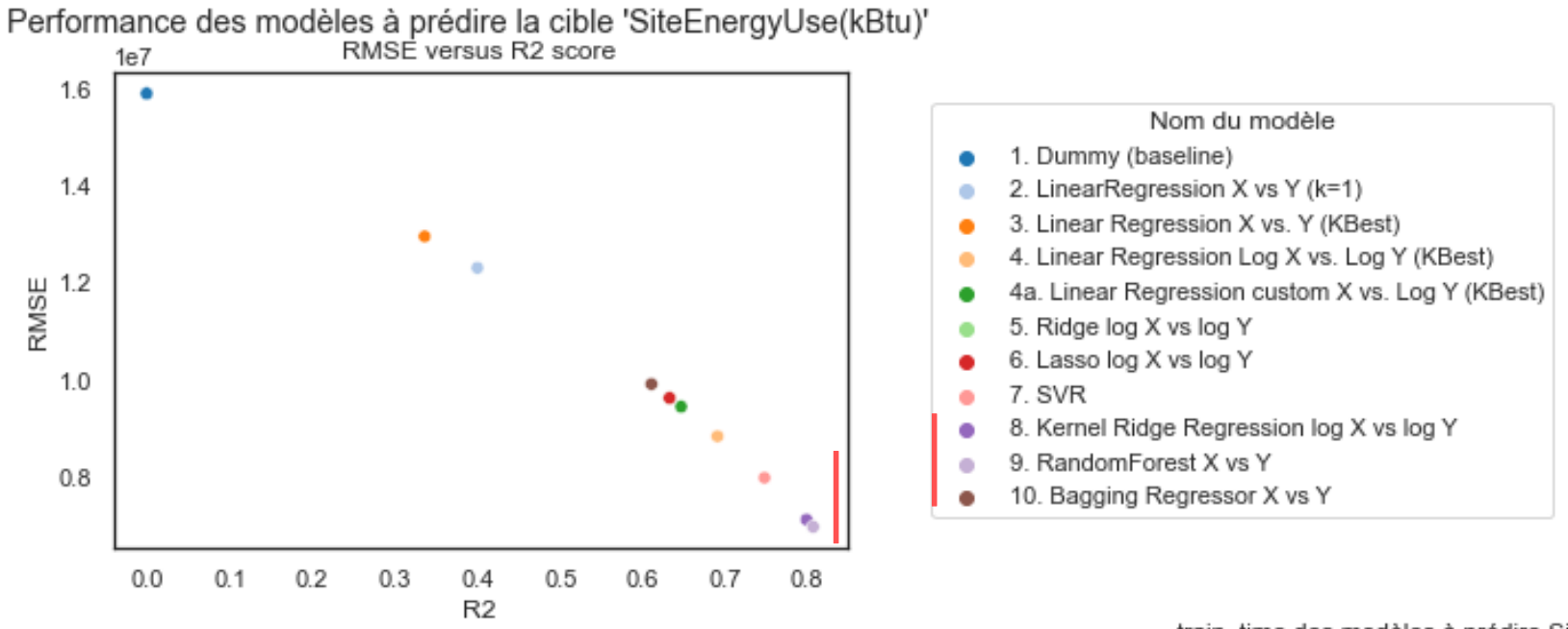


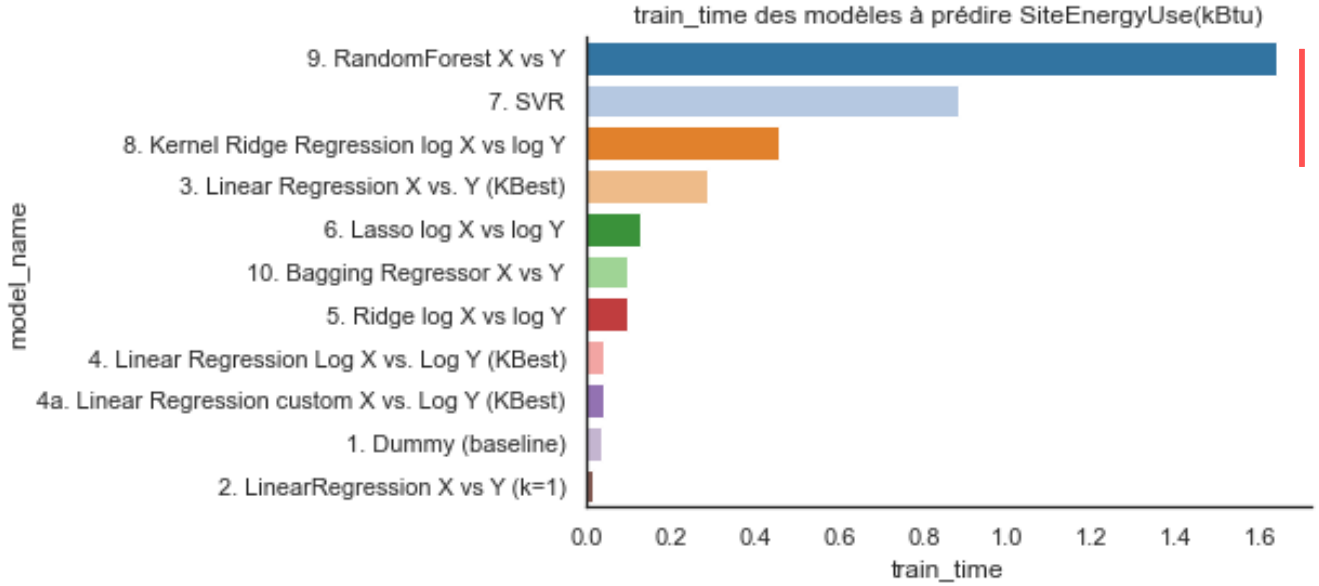
Feature Importance



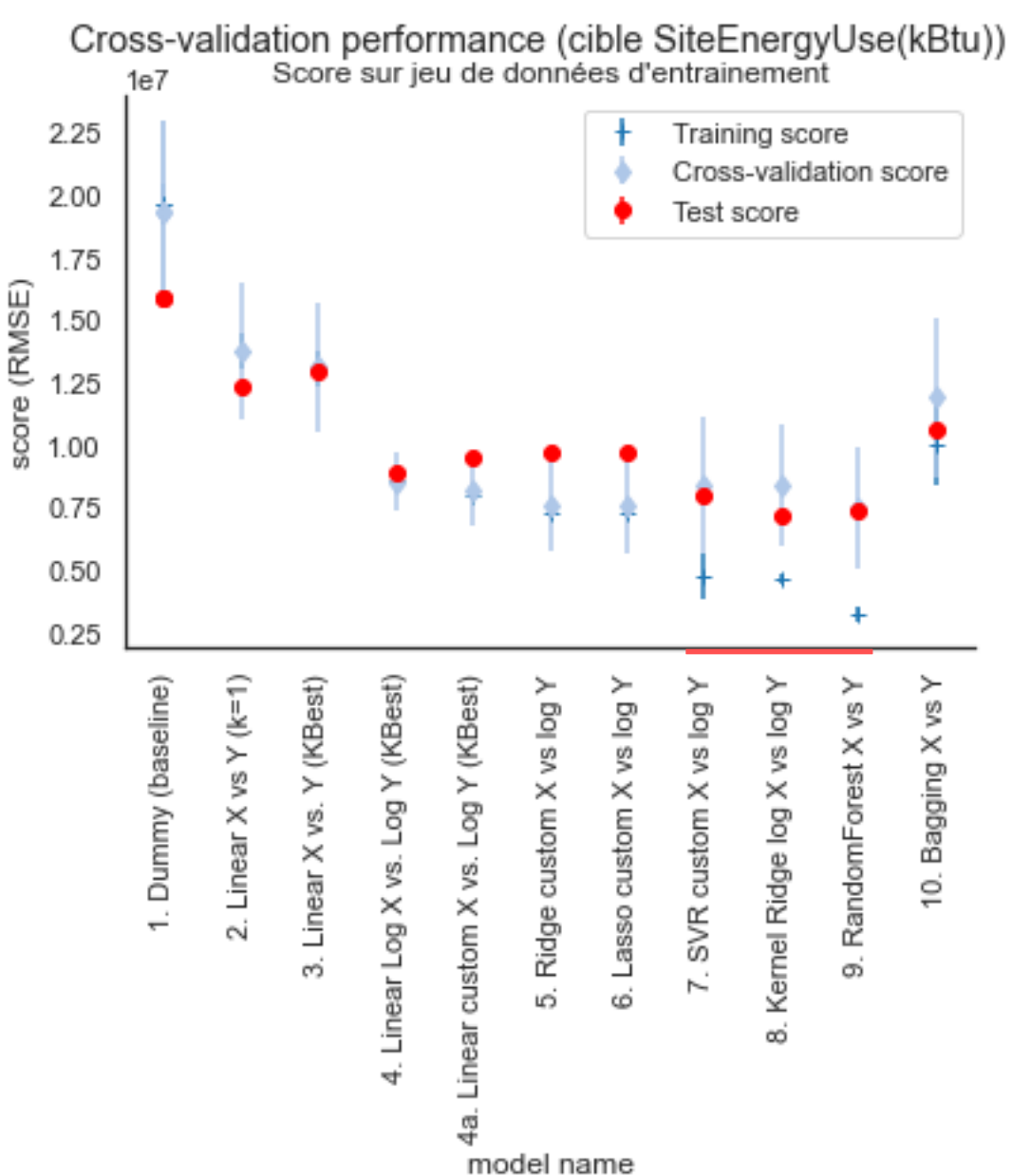
05 Le modèle final sélectionné

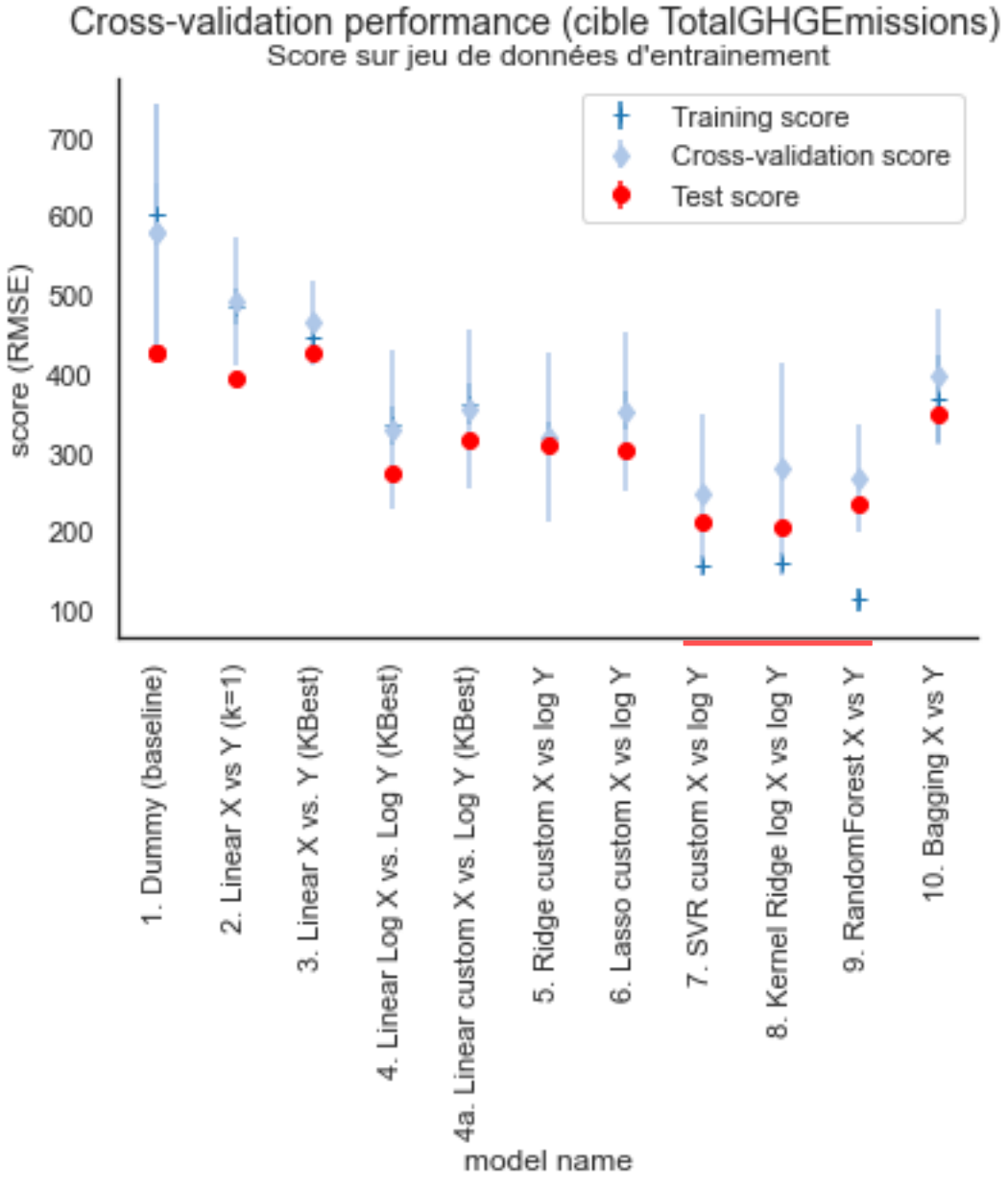
Comparaison des modèles – Consommation Energétique



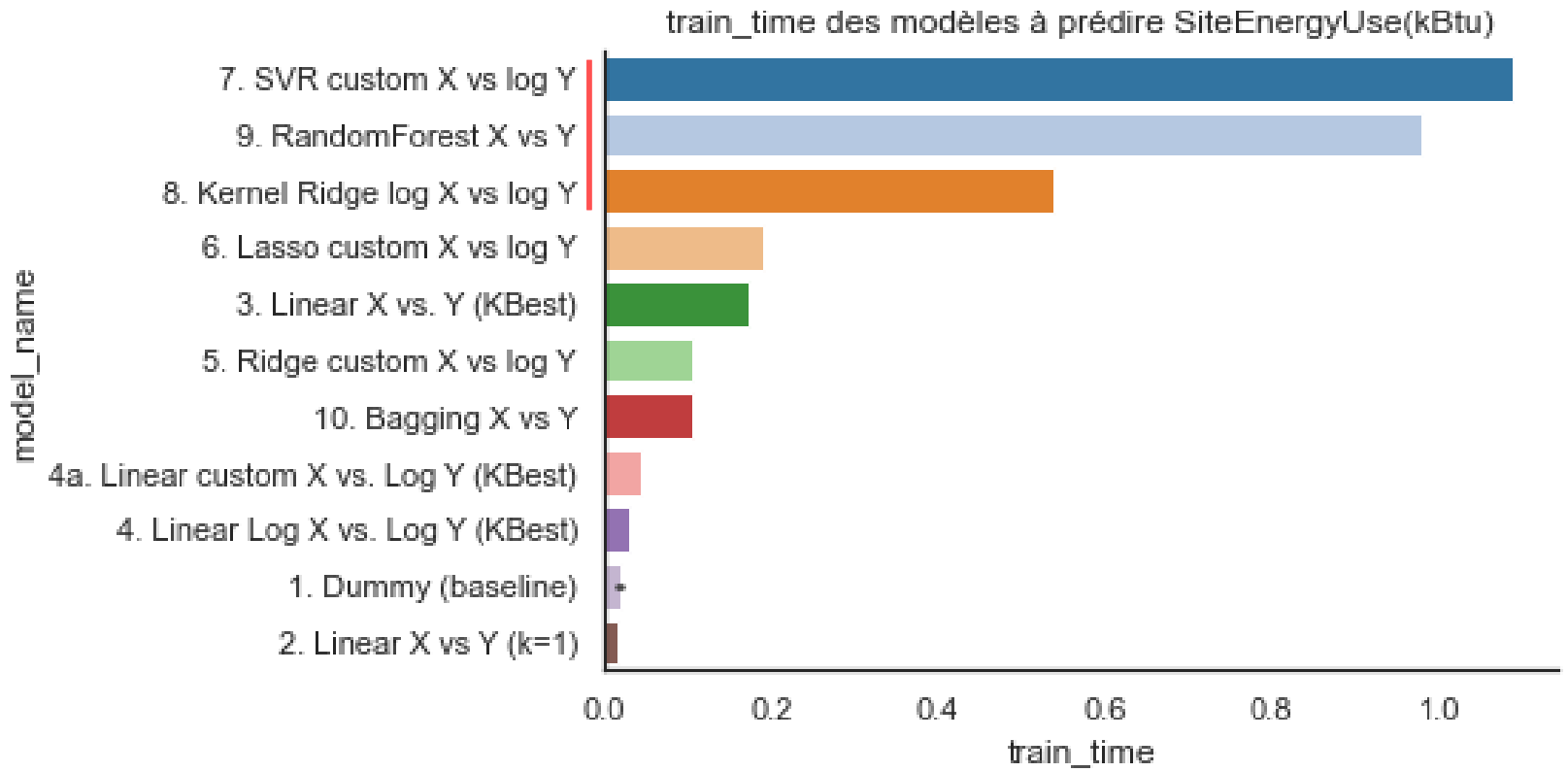


Comparaison des modèles – Scores

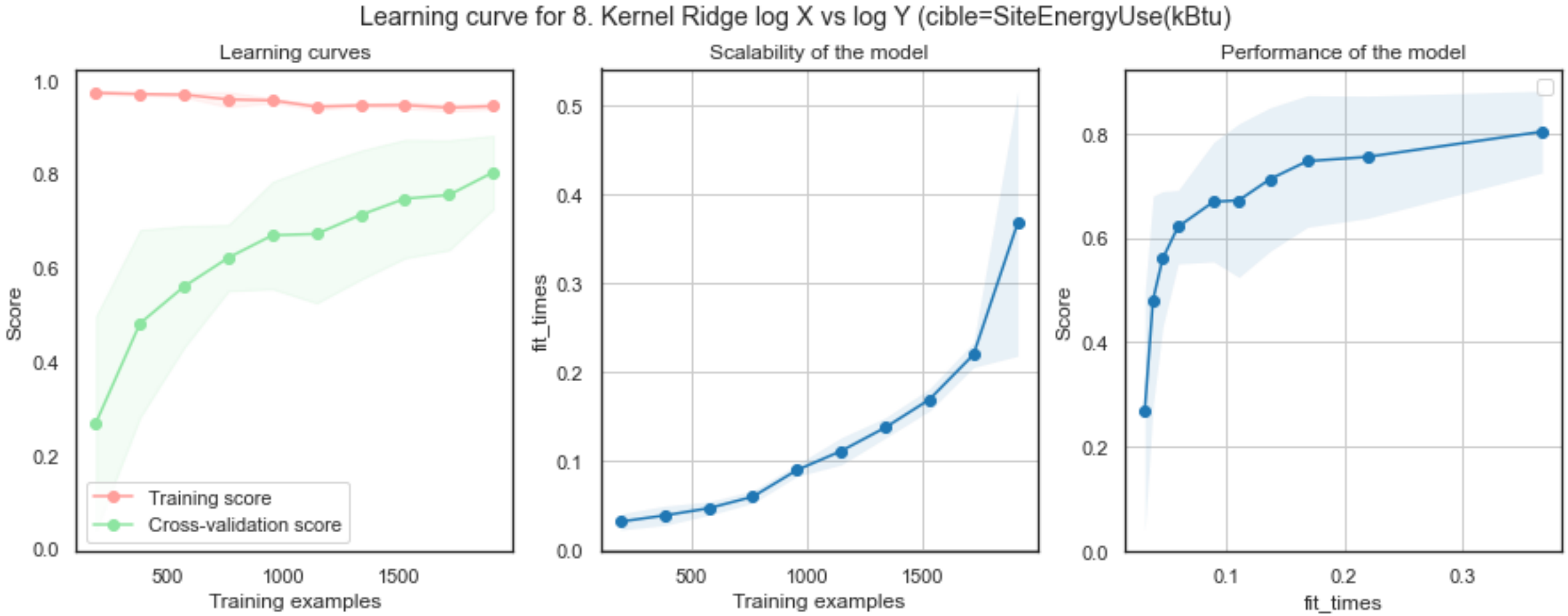




Comparaison des modèles – Temps

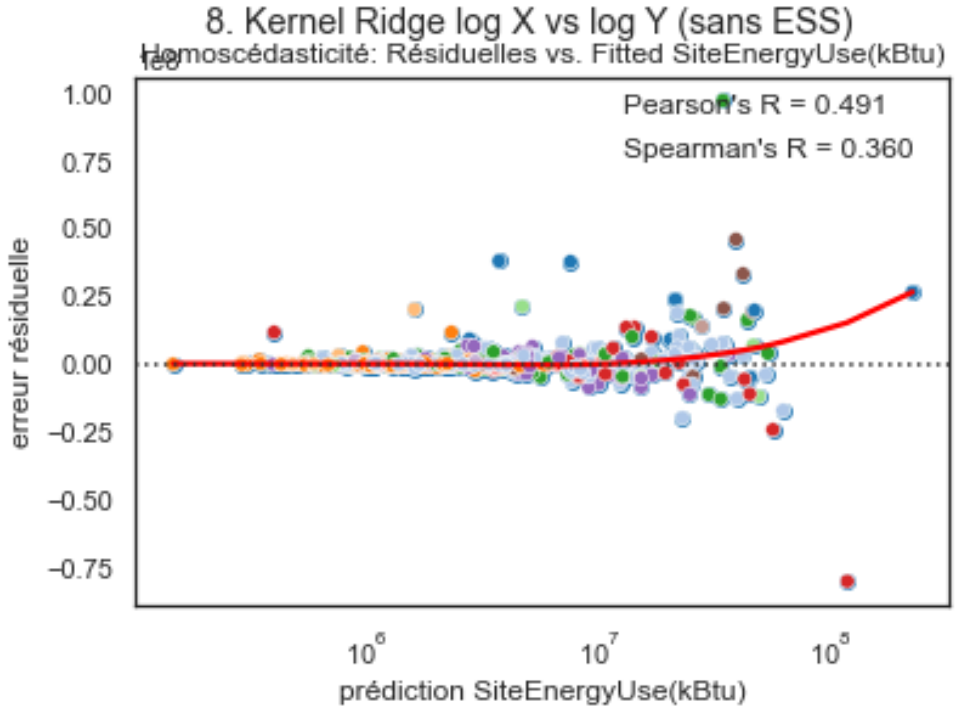


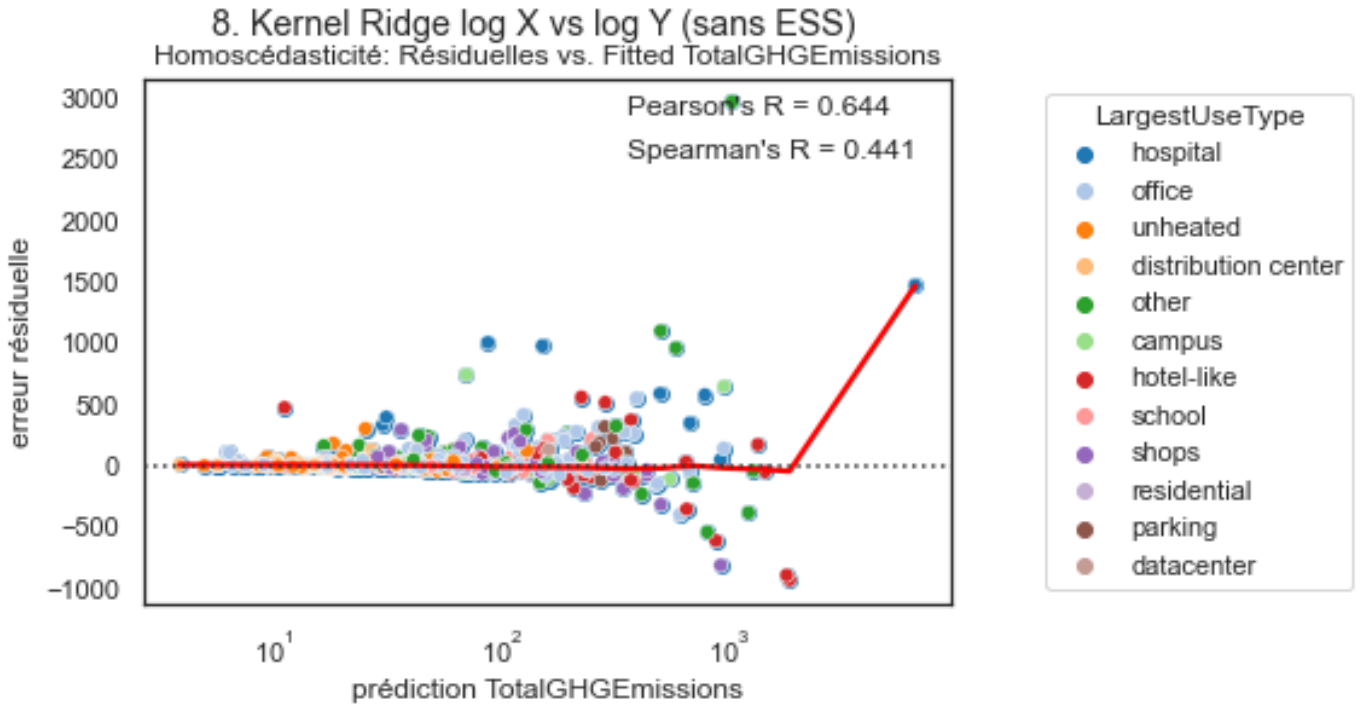
Modèle final (Kernel Ridge): learning curves



Modèle final : Analyse des résiduels

- Sous-estimation du consommation des hôpitaux et data centers
- Manque de homoscédasticité





Modèle final : influence de ENERGY STAR Score

- SiteEnergyUse(kBtu)

- Meilleur résultat **sans** Energy Star Score

	Sans ESS	Avec ESS
RMSE	**7126632**	9630552
R2	**0.799**	0.633

TotalGHGEmissions

- **Aucun effet** sur la performance du modèle

	Sans ESS	Avec ESS
RMSE	205.2	204.8
R2	0.768	0.769

06 Conclusion et améliorations à faire

Conclusions

- La consommation énergétique et émissions CO2 sont non-linéaire
- Meilleurs prédictions avec une modèle non-linéaire
- La transformation Log X et Log Y est nécessaire pour réduire l'influence d'outliers
- L'ENERGY STAR Score n' améliore pas les performances

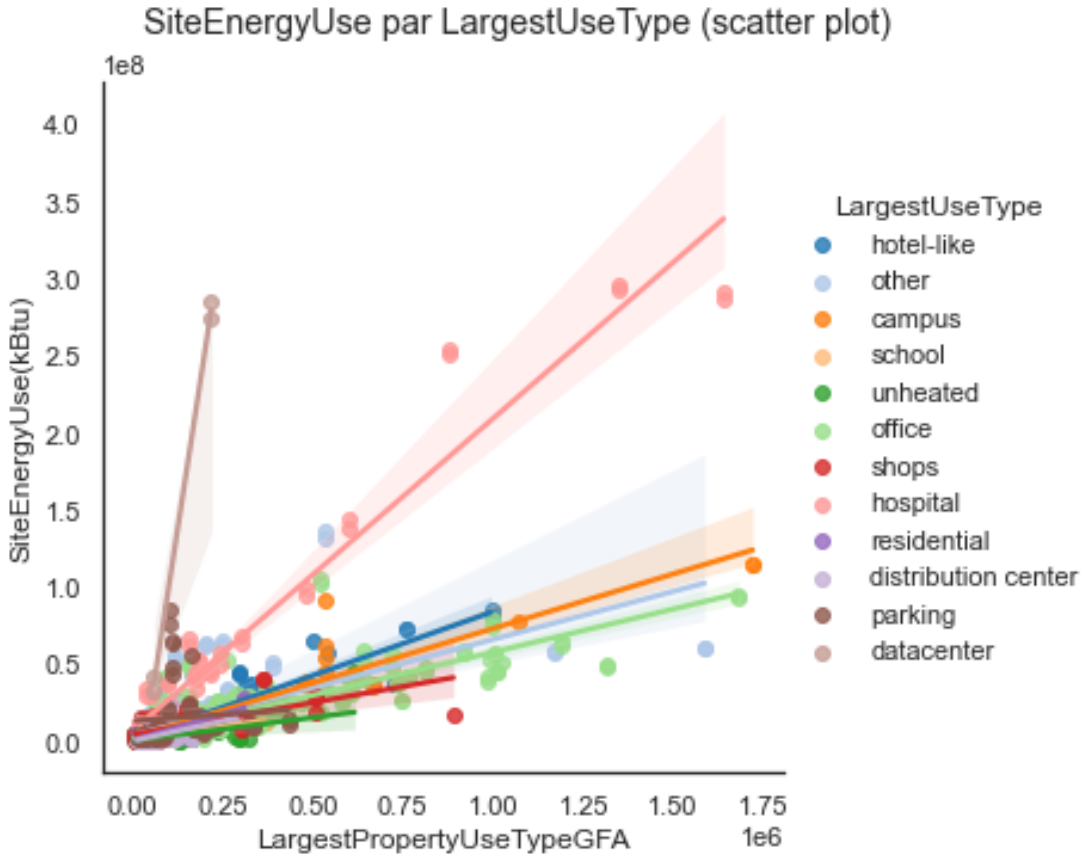
Améliorations à faire

- Recursive Feature Elimination
- Besoin de nouvelles features pour améliorer les résiduelles
- Meilleur interpretabilité avec SHAPely values

Améliorations à faire : Nouvelle feature engineering

- Datacenter_GFA =
  - LargestPropertyUseTypeGFA * (LargestUseType == Datacenter)
  - + SecondLargestPropertyUseTypeGFA * (SecondLargestUseType == Datacenter)
  - + ThirdLargestPropertyUseTypeGFA * (ThirdLargestUseType == Datacenter)

- Hospital_GFA
- Unheated_GFA
- Campus_GFA
- ...



Questions

images: Mark Creasey

- mrcreasey@gmail.com

- Merci !