anticipez la consommation électrique de bâtiments

Soutenance Olivier Legrand
Parcours Data Scientist
Projet P4

Problématique

- Prédiction de la consommation totale d'énergie et d'émissions de GES à partir des caractéristiques de bâtiments:
  - Ground floor area
  - number of buildings
  - building type
  - Largest property use type
  - etc.
- Evaluation de l'ENERGYSTARScore comme prédicteur des émissions de GES:
  - On cherchera à évaluer si ce prédicteur est fortement associé aux émissions de GES
  - On cherchera à évaluer l'impact de cet indicateur dans la qualité des prédictions.

Interprétation

- Jeu de données: Seattle energy benchmarking, pour les années 2015 et 2016
- Identification des cibles
  - Emissions GES: TotalGHGEmissions
  - Consommation d'énergie totale: SiteEnergyUse

Pistes

- Plusieurs problématiques associées au jeu de données:
  - potentielle fuite de données → On s' empêchera d'utiliser toutes les variables "dérivées" (Intensity).
  - pas de données issues des relevés annuels, mais possibilité d'utiliser les nature et proportion d'énergie utilisées → on devra créer de nouvelles variables, mais ne pas utiliser Electricity, NaturalGas, SteamUse
- Comment les variables liées au permis d'exploitation commerciale sont-elles associées aux grandeurs cibles? Type et importance des corrélations
- Exploiter également les associations entre les variables catégorielles et les variables cibles.
- Envisager d'utiliser une prédiction sur une des cibles pour prédire l'autre.

Analyse exploratoire

Nettoyage

1. **Fusion des deux tables:**
   1.1. Transformation de la colonne Location en 6 colonnes: Address, ZipCode, Latitude, Longitude, State, City
2. **Sélection des colonnes pertinentes:**
   2.1. Variables du permis d'exploitation: GFA (PropertyGFAs, LargestPropertyUseTypeGFAs) NumberofFloors/Building, PrimaryPropertyType, LargestPropertyType (and Second-, Third-), DataYear, YearBuilt + Electricity, NaturalGas, Steam, OtherFuelUse
3. **Traitement des valeurs manquantes:**
   3.1. SecondLargestPropertyUseType: on remplace par "None", car l'absence de valeur est cohérente, mais la présence de NaN peut empêcher certains traitements numériques. Idem pour ThirdLargestPropertyUseType.
   3.2. Pour les variables quantitatives, on supprime les lignes incomplètement renseignées - sauf pour ENERGYSTARScore: trop de valeurs manquantes.

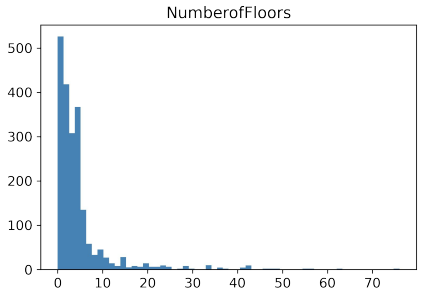
Analyse exploratoire

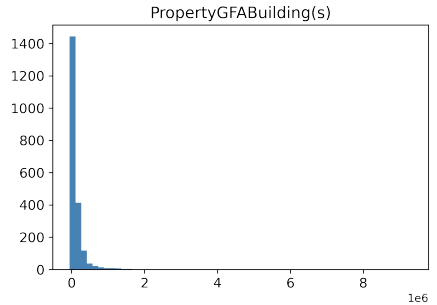
Distribution des variables quantitatives

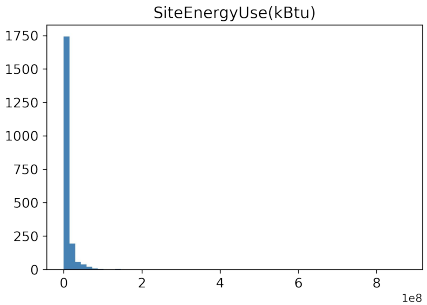
- Non-gaussiennes
- Grandes dispersions
- Grandes différences d'échelles

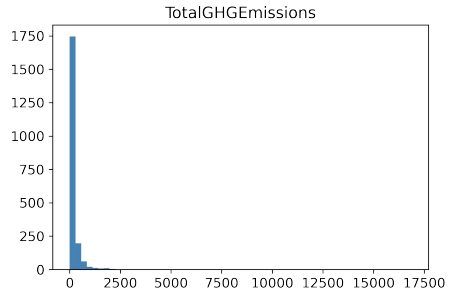
Corrélations importantes pour certains groupes de variables:

- 98% pour PropertyGFATotal, PropertyGFABuilding(s)
- 97% pour PropertyGFATotal, LargestPropertyUseTypeGFA
- 78% pour PropertyGFATotal, SecondLargestPropertyUseTypeGFA
- 94% pour Electricity et SiteEnergyUse, -92% pour Electricity et NaturalGas
- 89 % pour SiteEnergyUse et TotalGHGEmissions









Analyse exploratoire

Cas des variables catégorielles

- Grand nombre de modalités pour chaque variable (DataYear: 2 modalités, mais LargestPropertyUseType:53 et YearBuilt: 112 par exemple)
- Grand nombre de modalités presque vides: source potentielle de bruit
- Corrélations: ANOVA

	η^2	
	SiteEnergyUse	TotalGHGEmissions
PrimaryPropertyType	0.077	0.056
LargestPropertyUseType	0.060	0.067
SecondLargestPropertyUseType	0.042	0.019
BuildingType	0.023	0.013
YearBuilt	0.023	0.007

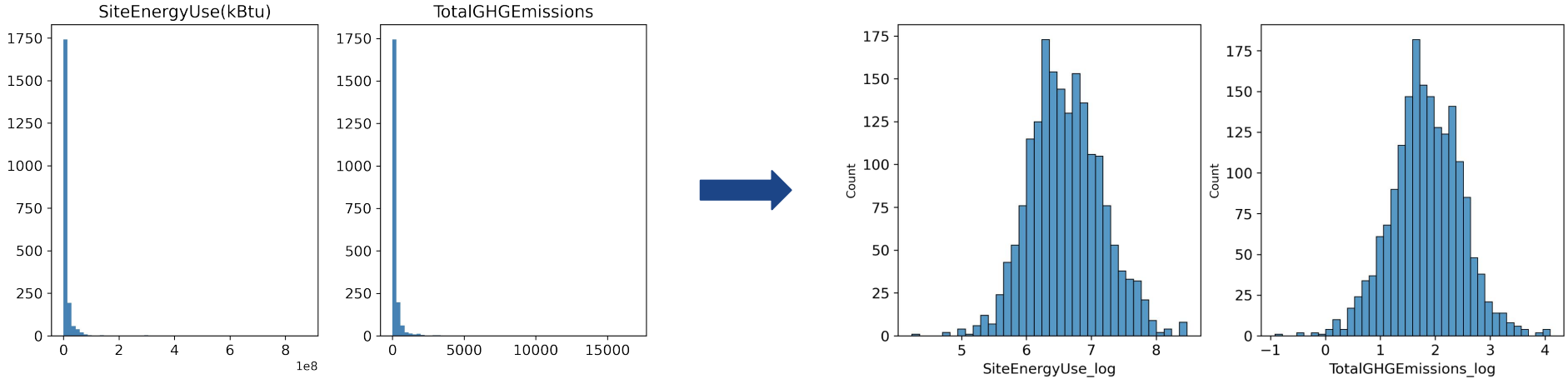
Analyse exploratoire

Corrélations et associations

- Intéressant d'envisager une PCA sur les variables quantitatives corrélées aux cibles, et très corrélées entre elles
- ANOVA → mise à l'écart des v. catégorielles non associées, ou dont l'effet est très faible
- ENERGYSTARScore très peu corrélée aux cibles et aux autres variables
- TotalGHGEmissions et SiteEnergyUse très corrélées entre elles, SiteEnergyUSe plus corrélée aux autres variables que TotalGHGEmissions: modèle séquentiel vs modèle non-séquentiel

Feature engineering

1. Très grande dispersion et non-gaussianité des données:
   a. log10 sur les cibles TotalGHGEmissions et SiteEnergyUse
   b. standardisation des prédicteurs (sur jeu d'entraînement seulement)



Feature engineering

1. Très grande dispersion et non-gaussianité des données:
   a. log10 sur les cibles TotalGHGEmissions et SiteEnergyUse
   b. standardisation des prédicteurs (sur jeu d'entraînement seulement)
2. PCA sur le groupe de variables 'GFA' très corrélées entre elles
3. Création des variables Energy_ratio, NaturalGas_ratio, Steam_ratio
4. One-hot encoding des variables catégorielles, puis réduction du nombre de modalités:
   a. par groupements basés sur des seuils de population et/ou règles métiers et/ou des considérations portant sur les dépendances entre les cibles et les diverses modalités.
   b. YearBuilt groupé en deux catégories: avant 1980, après 1980

Modèles

Structure des modèles

Modèle: pipeline de prétraitement + estimateur.

- Pipeline de prétraitement:
  - Standardisation des variables et PCA
  - one-hot encoding des variables catégorielles
  - optionnel: Transformation des features pour la régression polynomiale

- Estimateurs:
  - Régression Linéaire,
  - KNN,
  - Lasso pour réduire la complexité du modèle polynomial,
  - RandomForest pour réduire la complexité et prédire.

Baseline: régression linéaire

- Jeu de données: On écarte ‘DataYear’ (SEU) ou ‘DataYear’ et ThirdLargestPropertyUseType
- Pipeline: Standardisation, PCA, One-hot encoding
- Méthode: validation croisée 5 folds

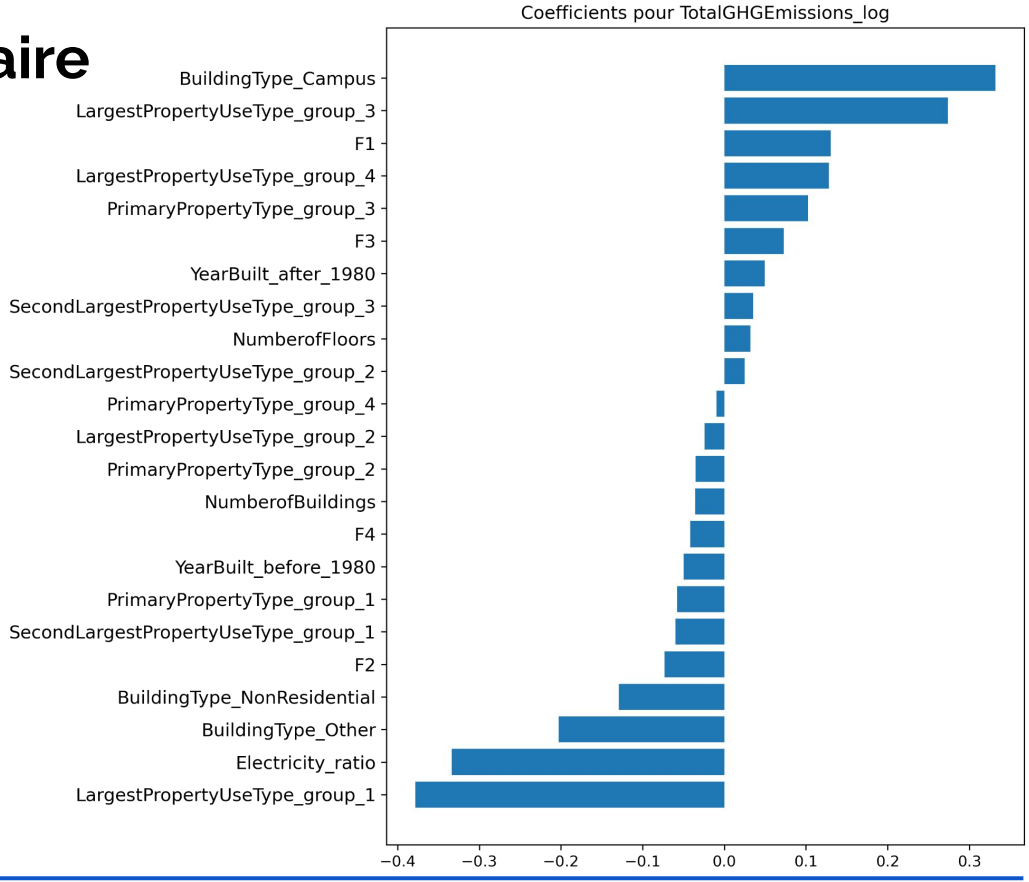
Baseline: régression linéaire

	SiteEnergyUse	TotalGHGEmissions
R2 (entraînement)	0.59 +/- 0.01	0.65 +/- 0.1
R2 (test)	**0.56 +/- 0.04**	**0.62 +/- 0.04**

→ Le modèle semble stable, mais le score indique un possible sous-apprentissage.
Régression polynomiale pour prendre en compte les interactions entre variables.

Baseline: régression linéaire

Principaux prédicteurs:

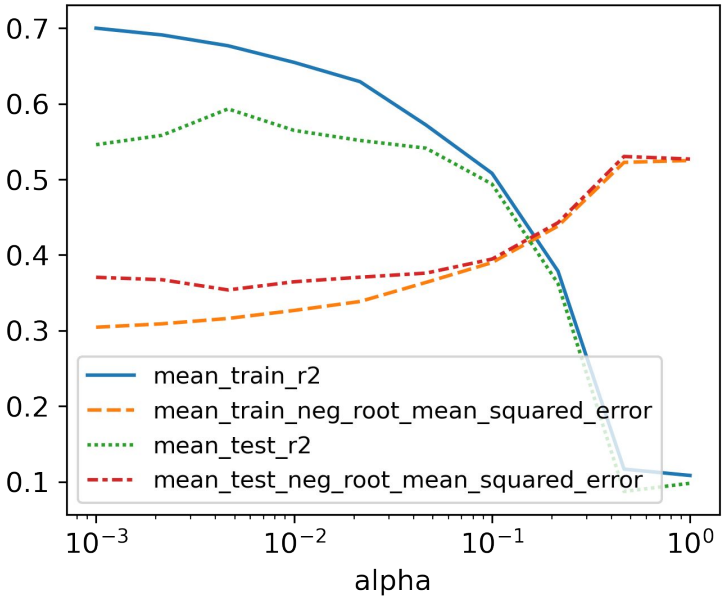


Régression polynomiale avec lasso

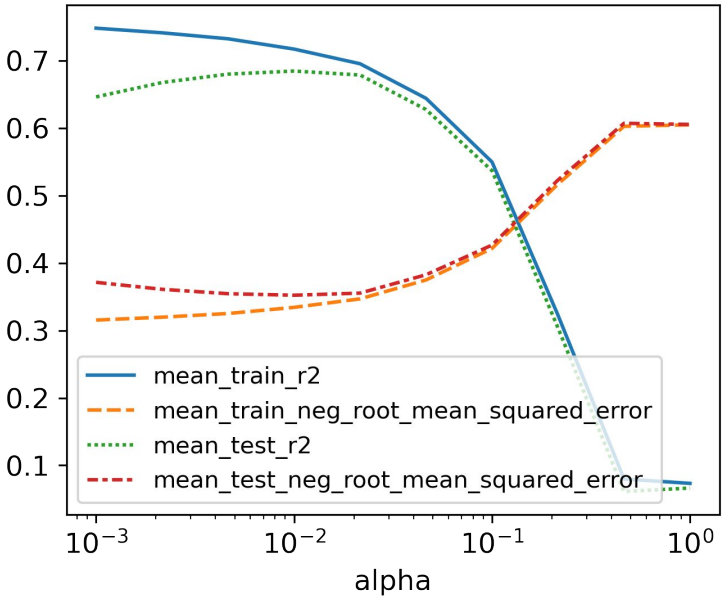
- Jeu de données: On écarte ‘DataYear’, ‘ThirdLargestPropertyUseType’, ‘BuildingType’ car risque de surapprentissage.
- Modèle exploré seulement sur la cible SiteEnergyUse.
- Pipeline: Standardisation, PCA, One-hot encoding, PolynomialFeatures de degré 2, Lasso
- Méthode: GridSearch (5 folds) sur le set d’entraînement pour l’évaluation de alpha

Régression polynomiale avec lasso

SiteEnergyUSe



TotalGHGEmissions



Régression polynomiale avec lasso

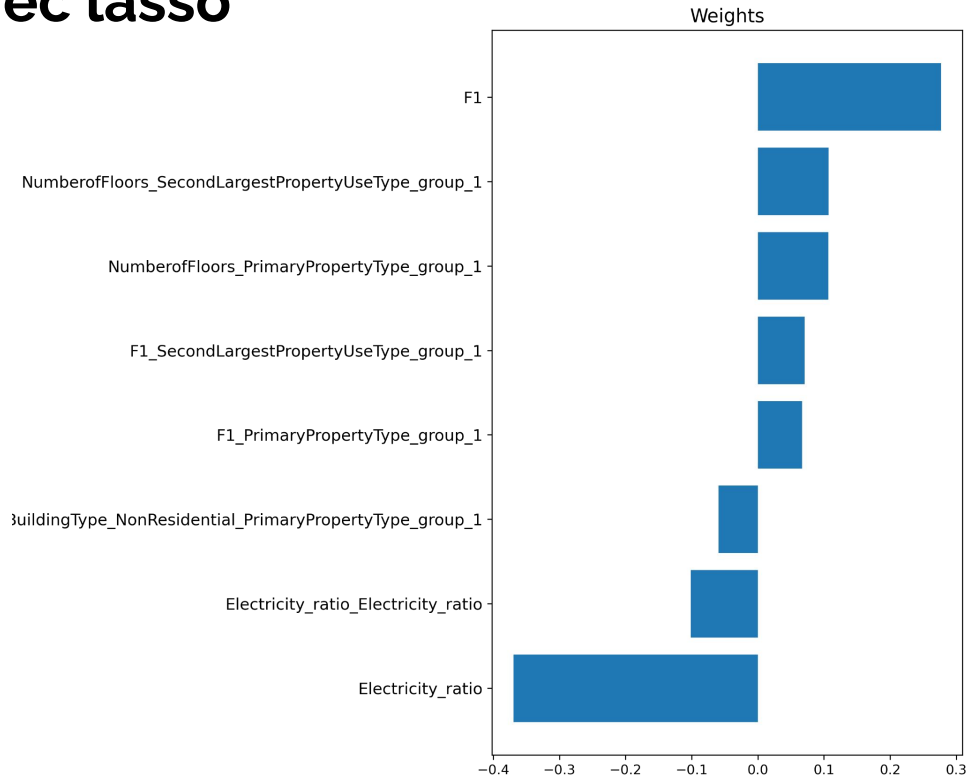
	SiteEnergyUse	TotalGHGEmissions
alpha	0.00717	0.00717
R2 (entraînement)	0.65 +/- 0.01	0.730 +/- 0.002
R2 (test)	**0.59 +/- 0.03**	**0.705 +/- 0.012**

→ Amélioration par rapport à la régression linéaire. Toujours en sous-apprentissage.
Pour aller plus loin: knn, randomforest

Régression polynomiale avec lasso

Principaux prédicteurs:

→ Prise en compte des interactions.

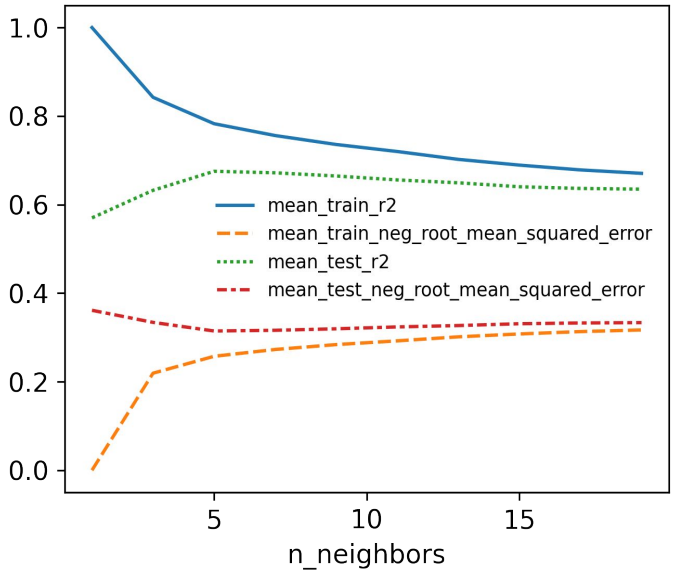


k-NearestNeighbors

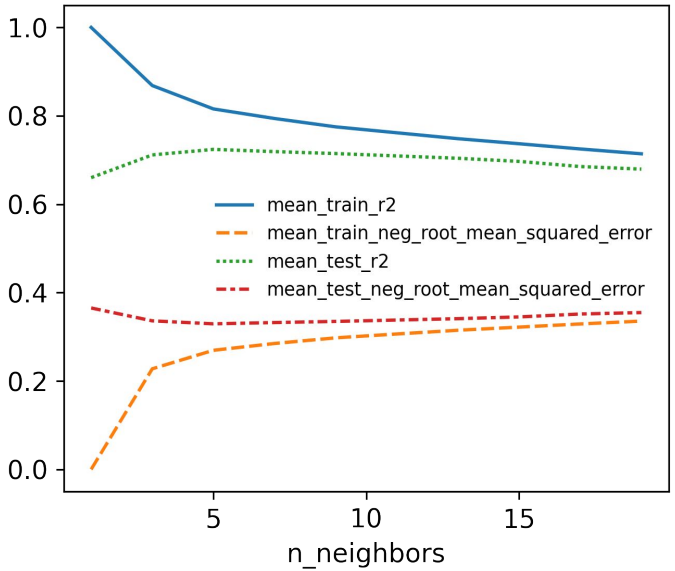
- Jeu de données: On écarte ‘DataYear’, ‘ThirdLargestPropertyUseType’, ‘BuildingType’ (SEU), et ‘DataYear’, ‘ThirdLargestPropertyUseType’, ‘YearBuilt’
- Pipeline: Standardisation, PCA, One-hot encoding
- Méthode: GridSearch pour la détermination du nombre optimal de p.p. voisins

k-NearestNeighbors

SiteEnergyUSe



TotalGHGEmissions



k-NearestNeighbors

	SiteEnergyUse	TotalGHGEmissions
n_neighbors	5	5
R2 (entraînement)	0.783 +/- 0.002	0.816 +/- 0.004
R2 (test)	**0.675 +/- 0.02**	**0.724 +/- 0.018**

Réduction de la complexité avec RandomForest

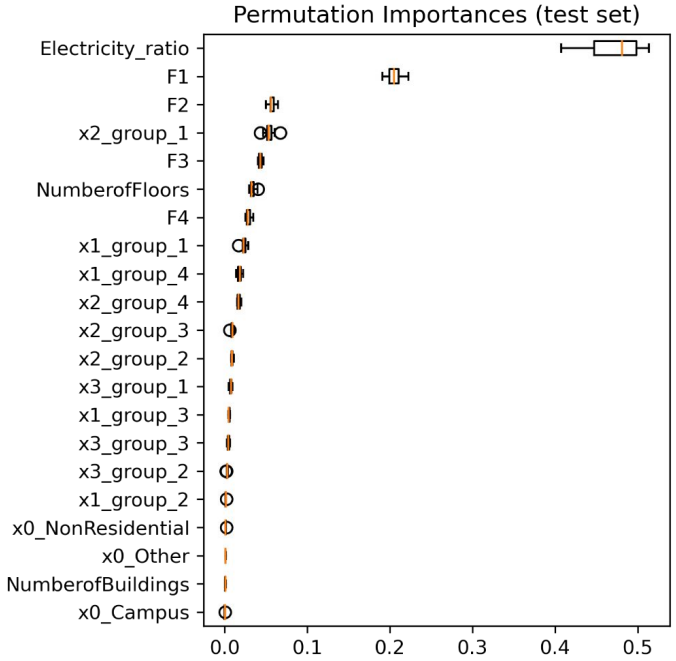
- Jeu de données: Même jeu de données que pour la régression linéaire simple
- Pipeline: Standardisation, PCA, One-hot encoding
- Méthode: GridSearch pour déterminer les valeurs optimales de max_features, max_depth et n_estimators
- + On utilise deux méthodes (feature permutation et feature importance) pour évaluer l'importance des différents prédicteurs et créer un modèle plus simple.

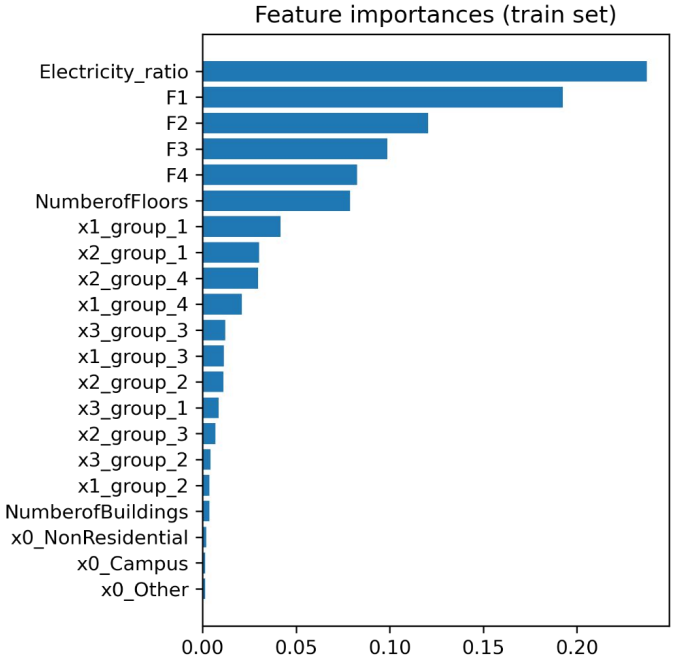
Réduction de la complexité avec RandomForest

	SiteEnergyUse	TotalGHGEmissions
n_estimators, max_depth, max_features	500, 'None', 'sqrt'	500, 'None', 'auto'
R2 (entraînement)	0.971 +/- 0.000	0.975 +/- 0.001
R2 (test)	**0.790 +/- 0.023**	**0.816 +/- 0.027**

Réduction de la complexité avec RandomForest

x0: BuildingType
x1: PrimaryPropertyType
x2: LargestPropertyUseType
x3: SecondLargestPropertyUseType



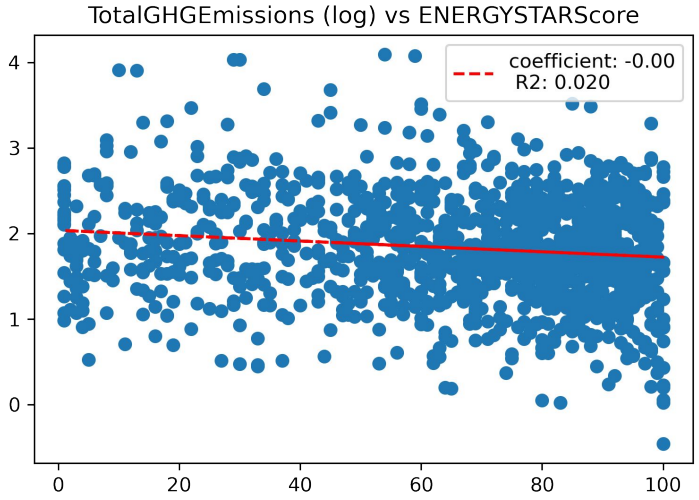


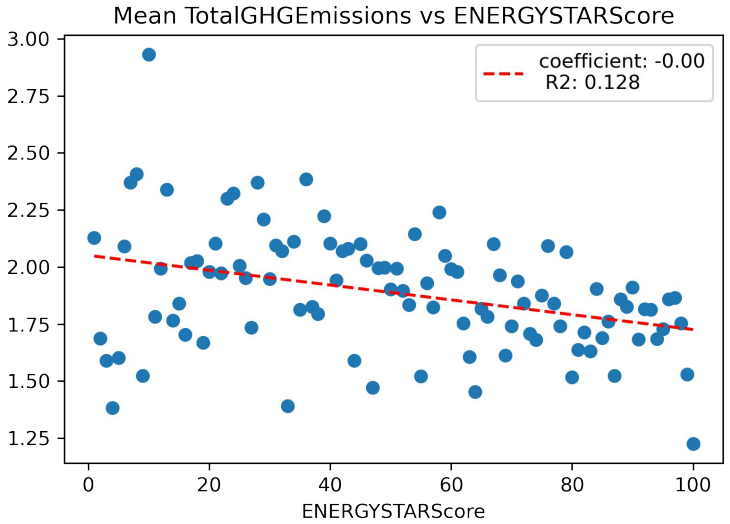
Réduction de la complexité avec RandomForest

- Sélection des variables dont l'importance ≥ 3%
- Variables sélectionnées: 'Electricity_ratio' → 'x3_group_1' (8 variables sur 21)
- Score (GHGEmissions): 0.84
  → perte de performance maîtrisée avec la sélection.
- De plus:
  → moindre temps de calcul

  → amélioration de l'interprétabilité.

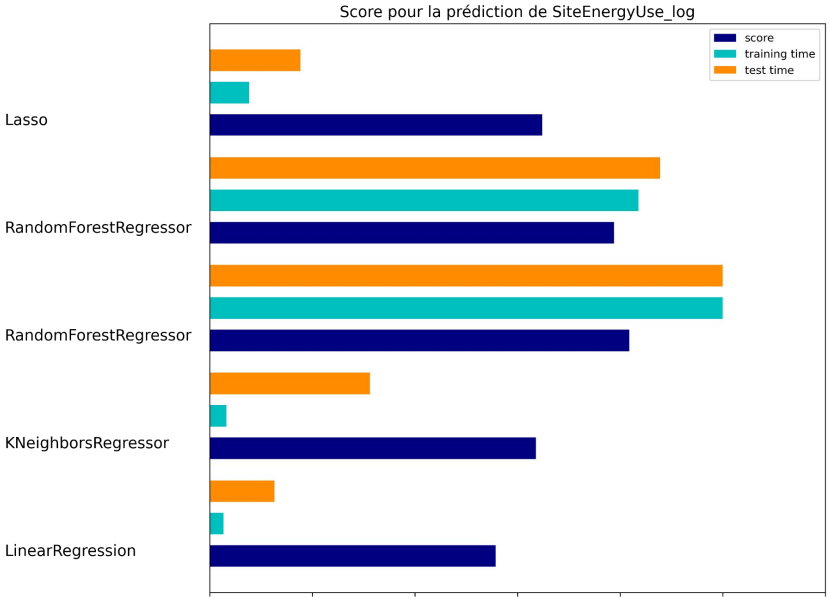
Pertinence d'ENERGYSTARScore pour la prédiction de TotalGHGEmissions

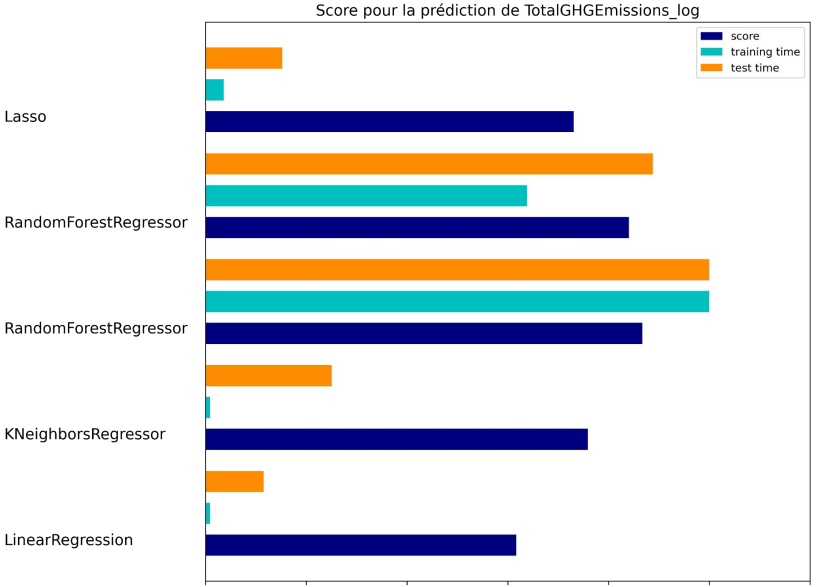




Modèle final

Comparaison des modèles





Modèle final

1. Sélection des inputs
2. PCA, normalisation, Onehot encoding
3. Prédiction de SiteEnergyUse avec RandomForest
4. Deux modèles:
    a. Prédiction de TotalGHGEmissions avec Random Forest, à partir des prédictions de SiteEnergyUse
    b. Prédictions de SiteEnergyUse et TotalGHGEmissions en parallèle

I Modèle séquentiel

1. Sélection des inputs: jeu de données complet (pas de sélection des features)
2. PCA, normalisation, One-hot encoding
3. Prédiction de SiteEnergyUse avec RandomForest
   a. scoring sur train/test = 80/20
   b. prédictions sur train1/test1 = 50/50
4. Prédiction de TotalGHGEmissions avec Random Forest, à partir des prédictions de SiteEnergyUse (train2/test2: 80/20 *(test1))

II Modèle non séquentiel

1. Sélection des inputs: jeux de données complets
2. PCA, normalisation, On-ehot encoding
3. Réduction de la complexité avec SelectFromModel et RandomForest
4. Prédiction de SiteEnergyUse avec RandomForest
5. Prédiction de TotalGHGEmissions avec Random Forest

Réduction de la complexité avec RandomForest

	SiteEnergyUse	TotalGHGEmissions
Modèle séquentiel	**0.83 +/- 0.09**	**0.84 +/- 0.02**
Modèle non-séquentiel	**0.83 +/- 0.01**	**0.83 +/- 0.02**

→ Résultats très similaires. Modèle non-séquentiel plus stable, performances sur chaque variable indépendantes.