anticipez la consommation électrique de bâtiments

Soutenance Olivier Legrand
Parcours Data Scientist
Projet P4

Interprétation de la problématique

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

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

1. Nettoyage
2. Analyse exploratoire
3. Feature Engineering

Nettoyage (1)

1. **Fusion des deux tables:**
   1.1. Transformation de la colonne Location en 6 colonnes: Address, ZipCode, Latitude, Longitude, State, City

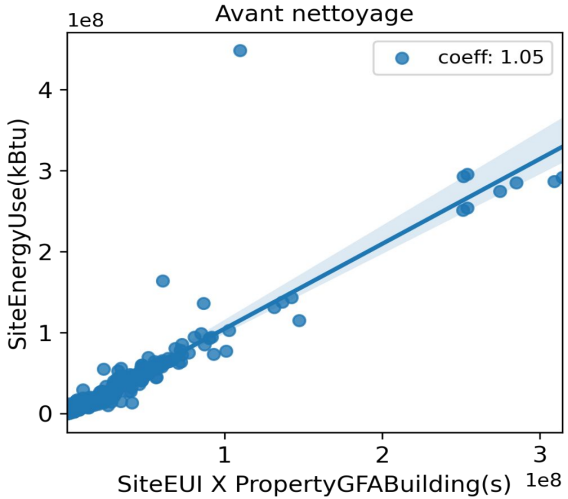
2. **Sélection des colonnes pertinentes:**
   2.1. Variables du permis d'exploitation: GFA (PropertyGFAs, LargestPropertyUseTypeGFAs) NumberofFloors/Building, PrimaryPropertyType, LargestPropertyType (and Second-, Third-), DataYear, YearBuilt + Electricity, NaturalGas, Steam, OtherFuelUse

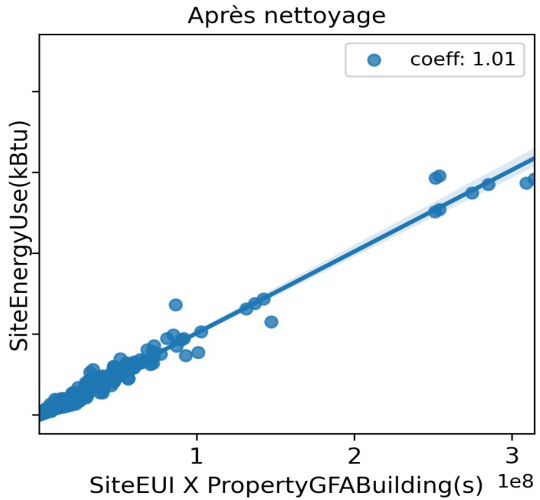
3. **Traitement des valeurs manquantes:**
   3.1. SecondLargestPropertyUseType: on remplace par "None", car l'absence de valeur est cohérente, mais la présence de NaN peut empêcher certains traitements numériques. Idem pour ThirdLargestPropertyUseType.
   3.2. Pour les variables quantitatives, on supprime les lignes incomplètement renseignées - sauf pour ENERGYSTARScore: trop de valeurs manquantes.

Nettoyage (2)

Traitement des outliers

1. Utilisation de la colonne ‘Outliers’
2. Sélection des individus pour lesquels SEU, GHG, NumberofBuildings > 0; 0 < NumberofFloors < 80
3. Correction des valeurs négatives de PropertyGFABuilding(s)
4. Utilisation de la relation linéaire entre SEU et SEUIntensity





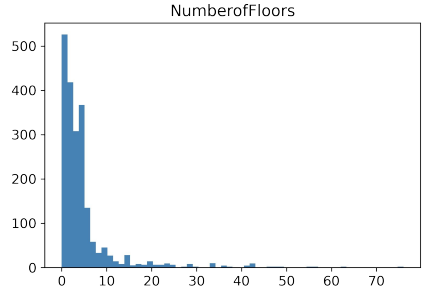
Analyse exploratoire

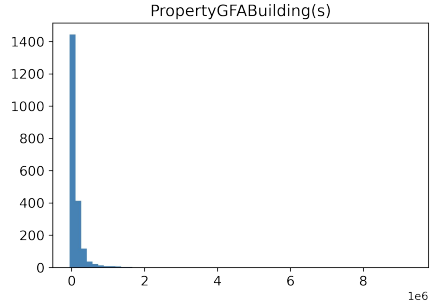
Distributions des variables quantitatives

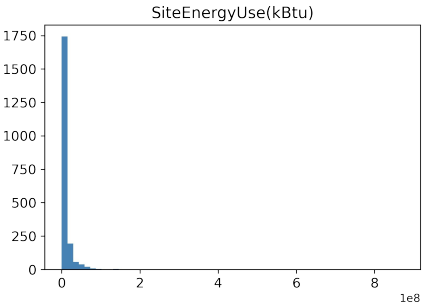
- Non-gaussiennes
- Grandes dispersions
- Grandes différences d'échelles

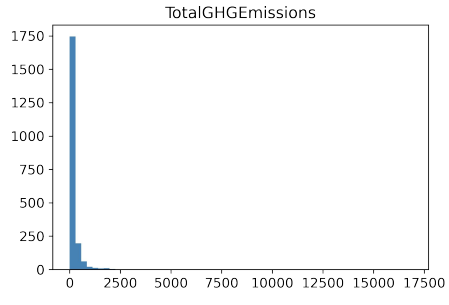
Corrélations des variables quantitatives

- 98% pour PropertyGFATotal, PropertyGFABuilding(s)
- 97% pour PropertyGFATotal, LargestPropertyUseTypeGFA
- 78% pour PropertyGFATotal, SecondLargestPropertyUseTypeGFA
- SiteEnergyUse et TotalGHGEmissions sont hautement corrélées. Peut-on prédire l'une en fonction de l'autre?









Analyse exploratoire

Distributions des variables quantitatives

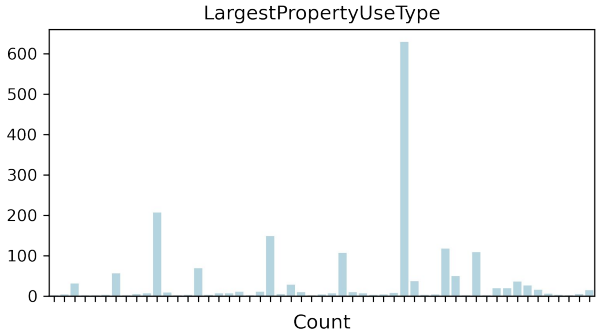
- Non-gaussiennes
- Grandes dispersions
- Grandes différences d'échelles

Corrélations des variables quantitatives

- 98% pour PropertyGFATotal, PropertyGFABuilding(s)
- 97% pour PropertyGFATotal, LargestPropertyUseTypeGFA
- 78% pour PropertyGFATotal, SecondLargestPropertyUseTypeGFA
- SiteEnergyUse et TotalGHGEmissions sont hautement corrélées. Peut-on prédire l'une en fonction de l'autre?

Distributions 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



Analyse exploratoire

Associations v. catégorielles / cibles: ANOVA

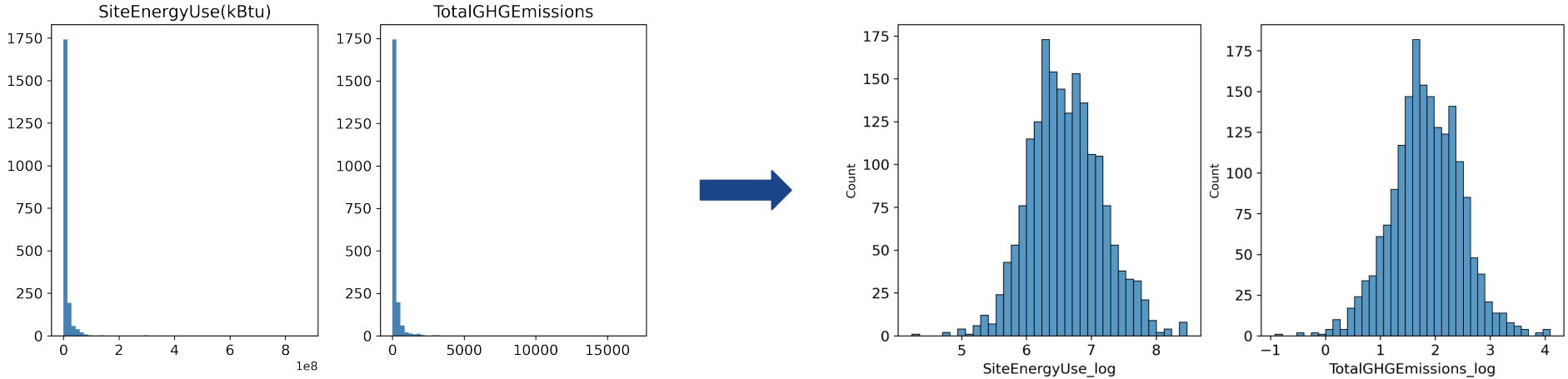
	η^2	
Indicateur	**SiteEnergyUse**	**TotalGHGEmissions**
PrimaryPropertyType	0.077	0.056
LargestPropertyUseType	0.060	0.067
SecondLargestPropertyUseType	0.042	0.019
BuildingType	0.023	0.013
YearBuilt	0.023	0.007
ThirdLargestPropertyUseType	0.012	0.000717
DataYear	0.000046	0.000012

p-val > 5%

Feature engineering

1. Variables quantitatives

a. log10 sur les cibles TotalGHGEmissions et SiteEnergyUse



Feature engineering

1. Variables quantitatives

a. log10 sur les cibles TotalGHGEmissions et SiteEnergyUse

b. Création des variables Energy_ratio, NaturalGas_ratio, Steam_ratio

	Electricity(kBtu)	NaturalGas(kBtu)	SteamUse(kBtu)	OtherFuelUse(kBtu)	Electricity_ratio	NaturalGas_ratio	Steam_ratio	OtherFuel_ratio
0	3686160.0	1272388.0	2023032.0	0.0	0.527995	0.182253	0.289773	0.0
1	3905411.0	4448985.0	0.0	0.0	0.467477	0.532542	0.000000	0.0
2	49762435.0	3709900.0	19660404.0	0.0	0.680459	0.050730	0.268839	0.0
4	6066245.0	8763105.0	0.0	0.0	0.409077	0.590940	0.000000	0.0
5	7271004.0	4781283.0	0.0	0.0	0.603303	0.396722	0.000000	0.0

Feature engineering

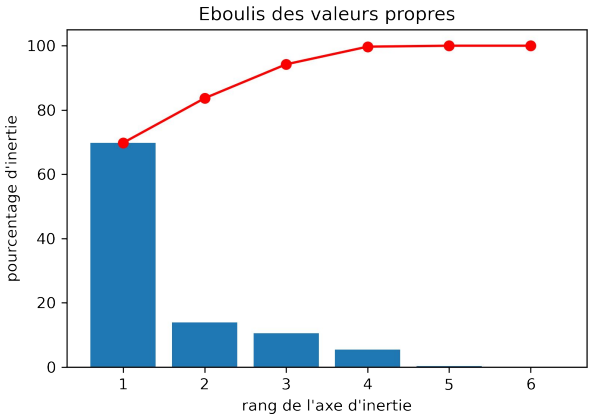
1. Variables quantitatives

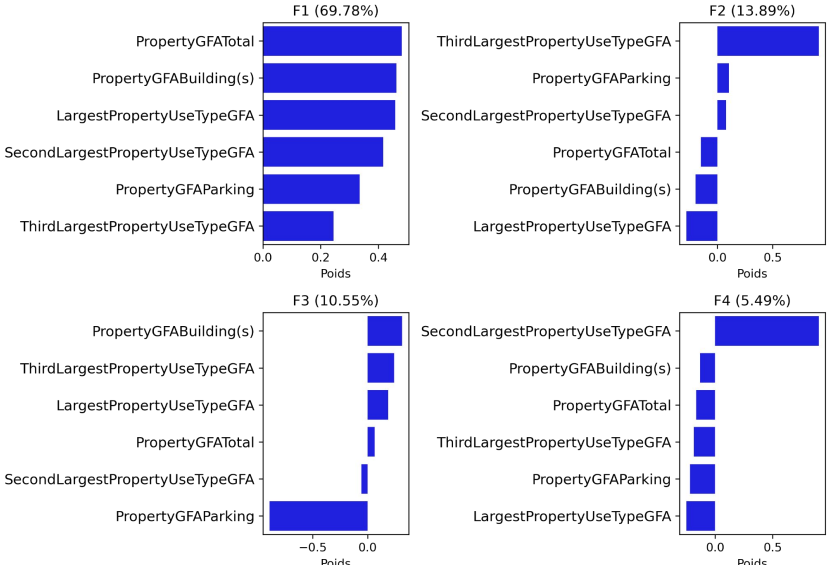
a. log10 sur les cibles TotalGHGEmissions et SiteEnergyUse

b. Création des variables Energy_ratio, NaturalGas_ratio, Steam_ratio

c. standardisation des prédicteurs (sur jeu d'entraînement seulement)

d. ACP sur les variables corrélées





Feature engineering

1. **Variables quantitatives**
   a. log10 sur les cibles TotalGHGEmissions et SiteEnergyUse
   b. Création des variables Energy_ratio, NaturalGas_ratio, Steam_ratio
   c. standardisation des prédicteurs (sur jeu d'entraînement seulement)
   d. ACP sur les variables corrélées
2. **Variables catégorielles**
   a. Réduction du nombre de modalités: 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
   c. Retrait des variables les moins associées aux cibles: DataYear (SEU), DataYear + ThirdLargestPropertyUseType (TGHGE)
   d. One hot encoding

Feature engineering

durant l'analyse exploratoire

intégré au pipeline de prétraitement

1. Variables quantitatives
   a. log10 sur les cibles TotalGHGEmissions et SiteEnergyUse
   b. Création des variables Energy_ratio, NaturalGas_ratio, Steam_ratio
   c. standardisation des prédicteurs (sur jeu d'entraînement seulement)
   d. ACP sur les variables corrélées
2. Variables catégorielles
   a. Réduction du nombre de modalités: 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
   c. Retrait des variables les moins associées aux cibles: DataYear (SEU), DataYear + ThirdLargestPropertyUseType (TGHGE)
   d. One hot encoding

Modèles

1. Structure générale des modèles
2. Baseline: Régression linéaire
3. Régression polynomiale avec Lasso
4. K-NN
5. Random Forest
6. Comparaison des modèles
7. ENERGYSTARScore

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: baseline
  - Régression Polynomiale: PolynomialFeatures + Lasso pour réduire la complexité du modèle polynomial,
  - KNN,
  - RandomForest

Baseline: régression linéaire

- Jeu de données: 7 var. quantitatives, 6 var. catégorielles (SEU), 5 var. catégorielles (TotalGHGE)
- Pipeline: Standardisation, PCA, One-hot encoding
- Méthode: validation croisée 5 folds

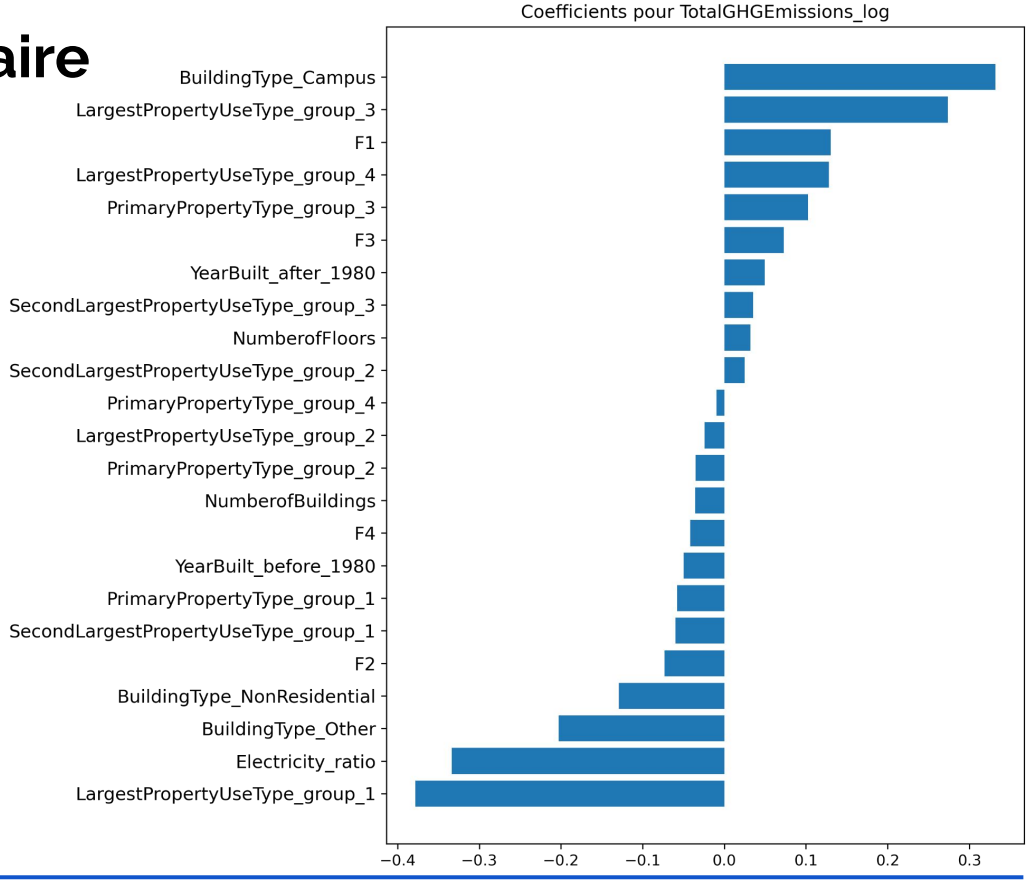
Baseline: régression linéaire

	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

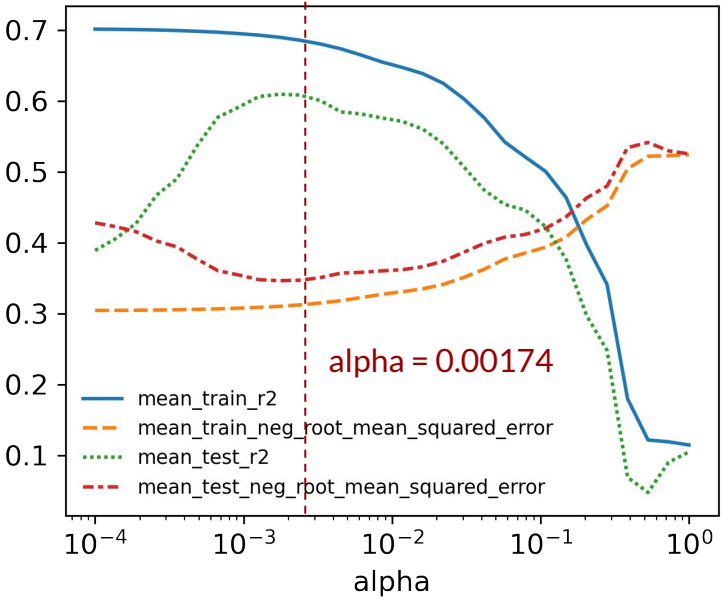
- Jeu de données:
  - risque de surapprentissage → 7 var. quantitatives + 4 var. catégorielles (ANOVA)

- Pipeline: Standardisation, PCA, One-hot encoding, PolynomialFeatures de degré 2, Lasso
  - input 11 variables → 171 variables avant régularisation

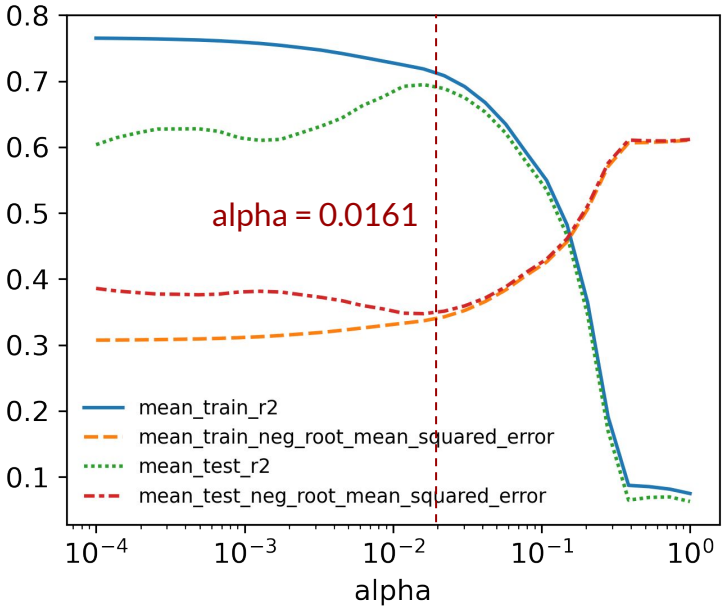
- Méthode: GridSearch (5 folds) sur le set d'entraînement pour l'évaluation de alpha

Régression polynomiale avec lasso

SiteEnergyUSe



TotalGHGEmissions



Régression polynomiale avec lasso

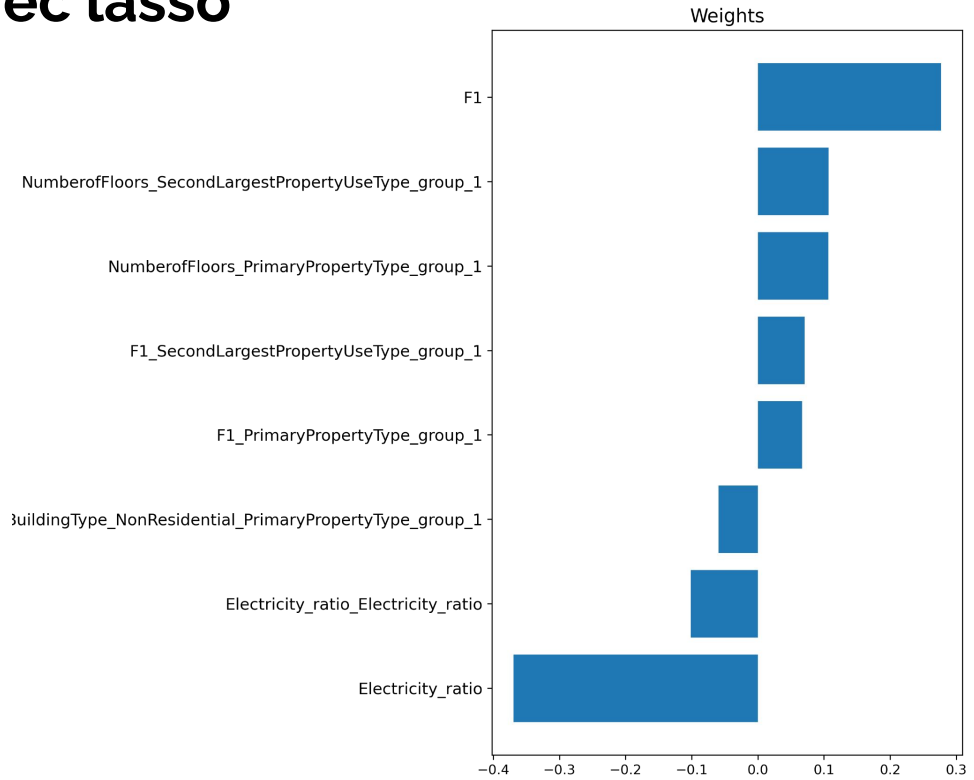
	SiteEnergyUse	TotalGHGEmissions
alpha	0.00174	0.0161
nb de variables après régularisation	31	11
R2 (entraînement)	0.690 +/- 0.005	0.719 +/- 0.007
R2 (test)	**0.609 +/- 0.064**	**0.695 +/- 0.041**

→ 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.

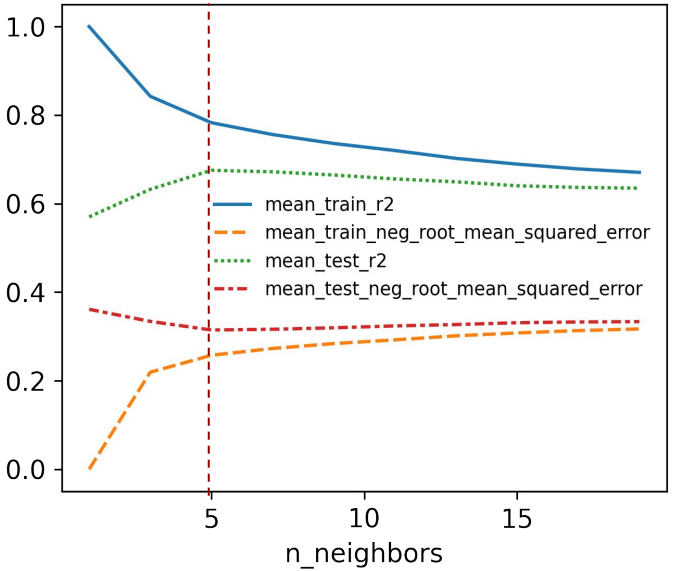


k-NearestNeighbors

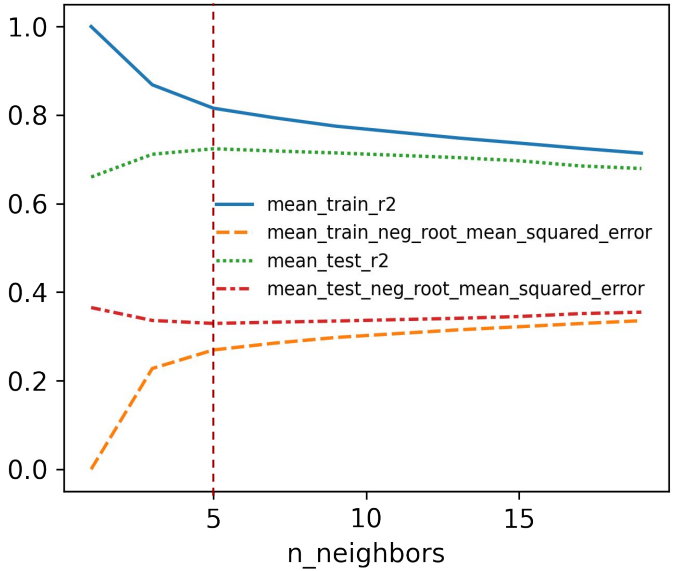
- Jeu de données:
  - 7 var. quantitatives + 4 var. catégorielles (ANOVA) (similaire régression polynomiale)
- Pipeline: Standardisation, PCA, One-hot encoding
- Méthode: GridSearch pour la détermination du nombre optimal de p.p. voisins

k-NearestNeighbors

SiteEnergyUSe



TotalGHGEmissions



k-NearestNeighbors

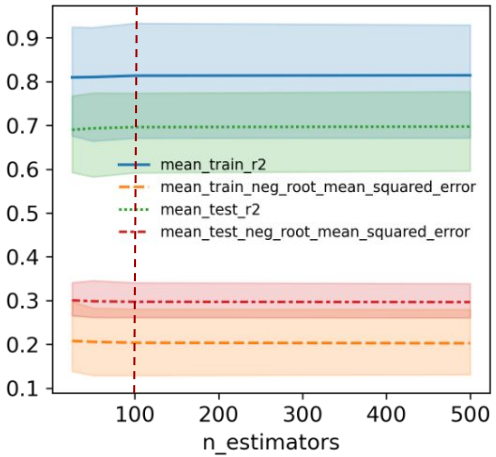
	SiteEnergyUse	TotalGHGEmissions
n_neighbors	5	5
R2 (entraînement)	0.783 +/- 0.002	0.816 +/- 0.004
R2 (test)	**0.675 +/- 0.02**	**0.724 +/- 0.018**

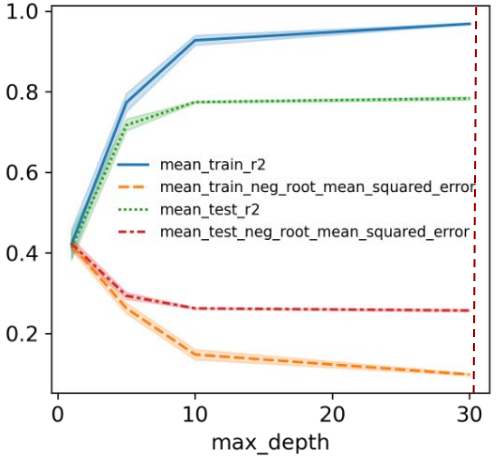
Random Forest

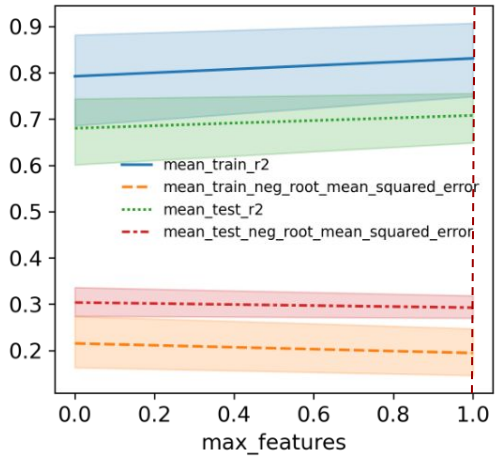
- Jeu de données: 7 var. quantitatives, 6 var. catégorielles (SEU), 5 var. catégorielles (TotalGHGE)
- Pipeline: Standardisation, PCA, One-hot encoding
- Méthode: GridSearch pour déterminer les valeurs optimales de max_features, max_depth et n_estimators
- + On utilise deux méthodes (feature permutation et feature importance) pour évaluer l'importance des différents prédicteurs et créer un modèle plus simple.

Random Forest

SiteEnergyUSe

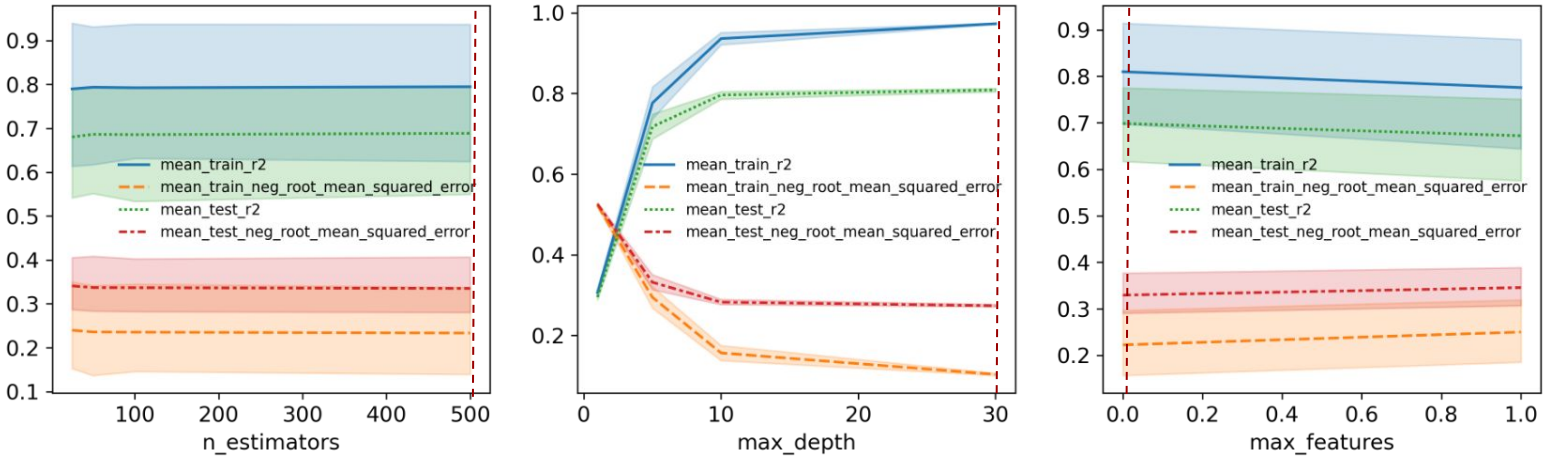






Random Forest

TotalGHGEmissions

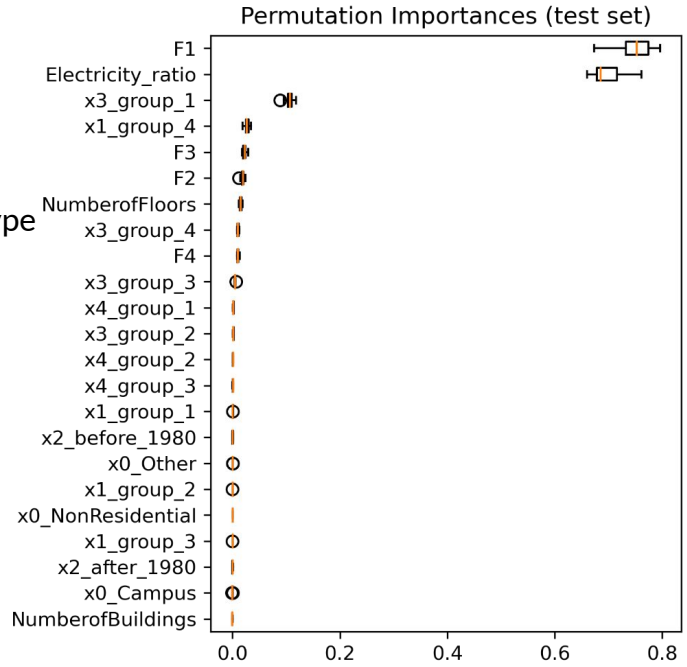


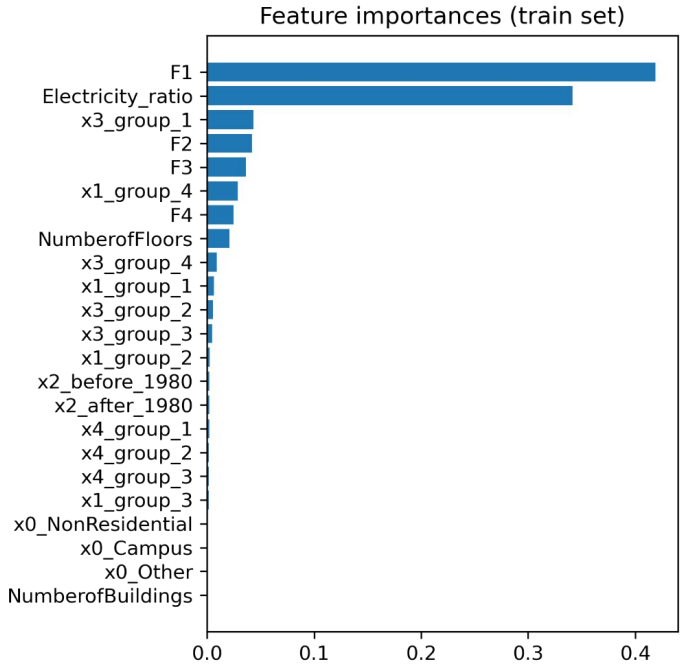
Random Forest

	SiteEnergyUse	TotalGHGEmissions
n_estimators, max_depth, max_features	100, '30', 'sqrt'	500, 'None', 'auto'
R2 (entraînement)	0.970 +/- 0.001	0.975 +/- 0.001
R2 (test)	**0.792 +/- 0.022**	**0.816 +/- 0.028**

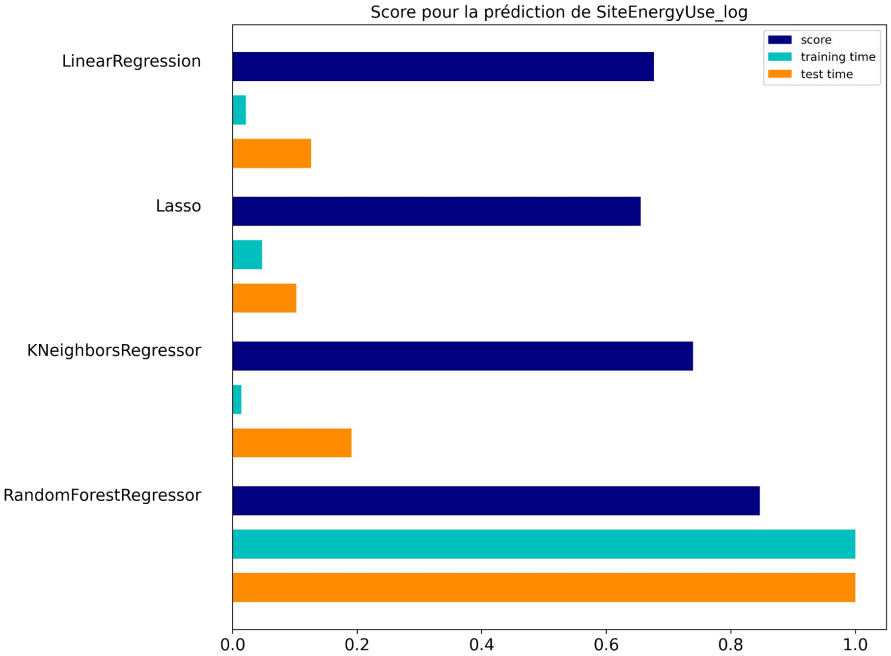
Feature importance avec Random Forest

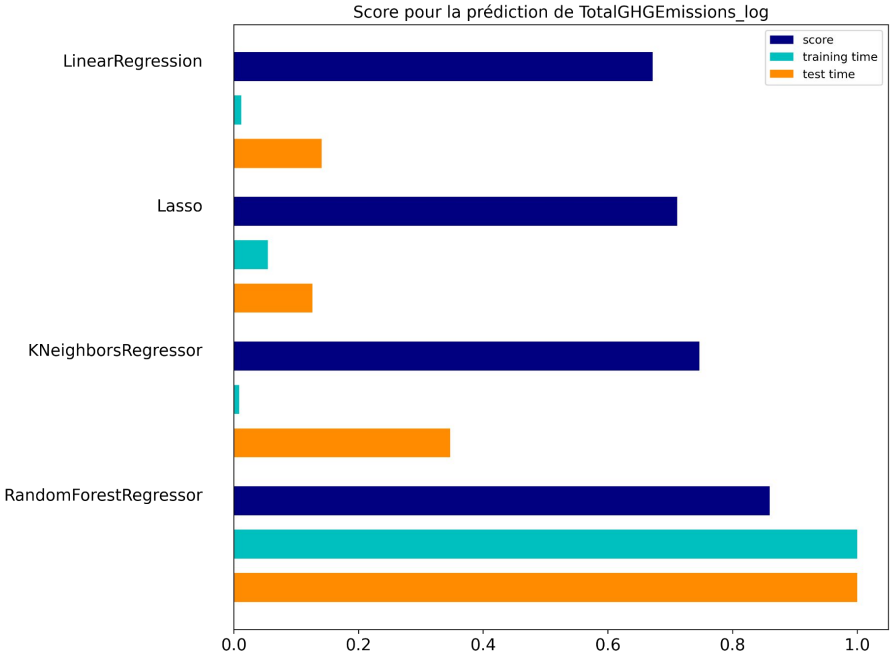
x0: BuildingType
x1: PrimaryPropertyType
x2: LargestPropertyUseType
x3: SecondLargestPropertyUseType



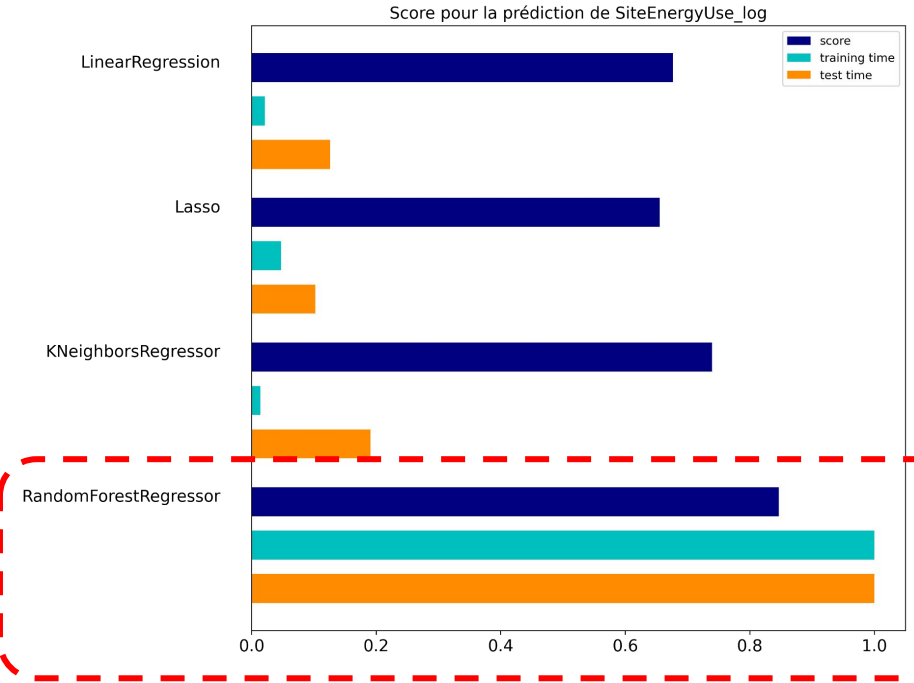


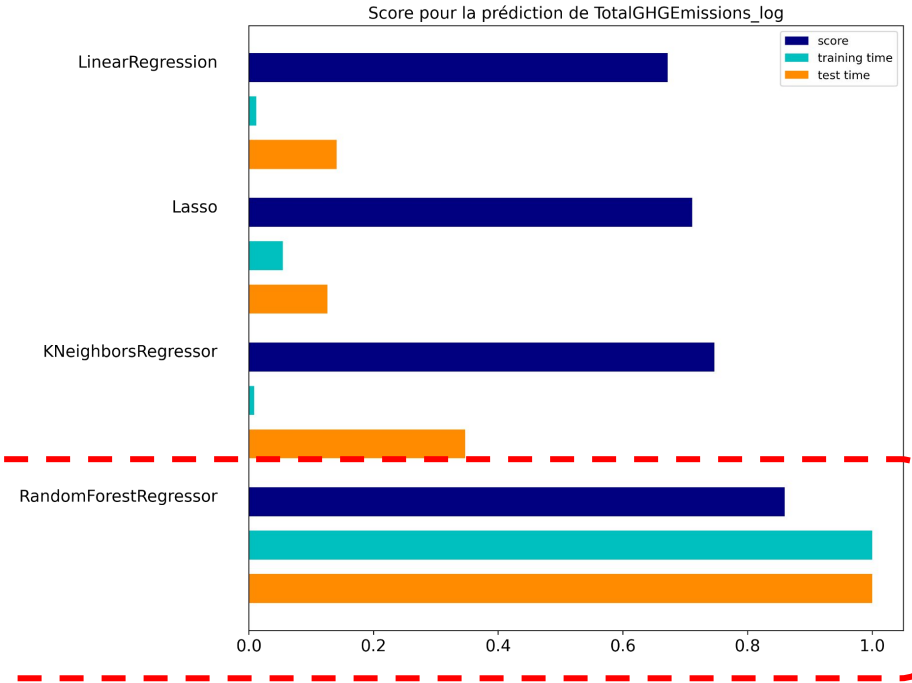
Comparaison des modèles



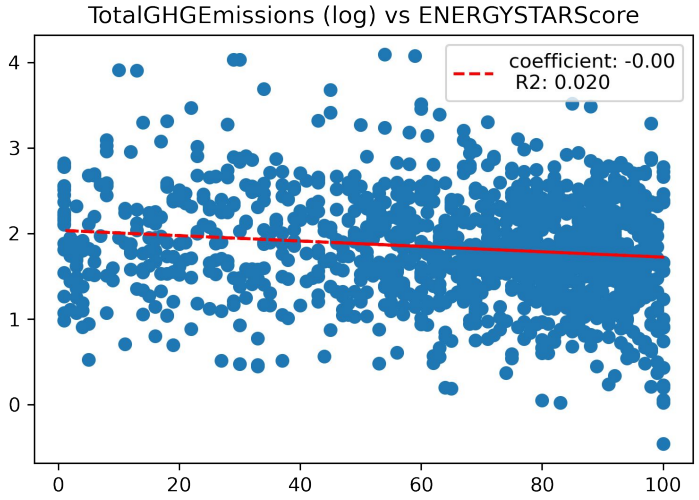


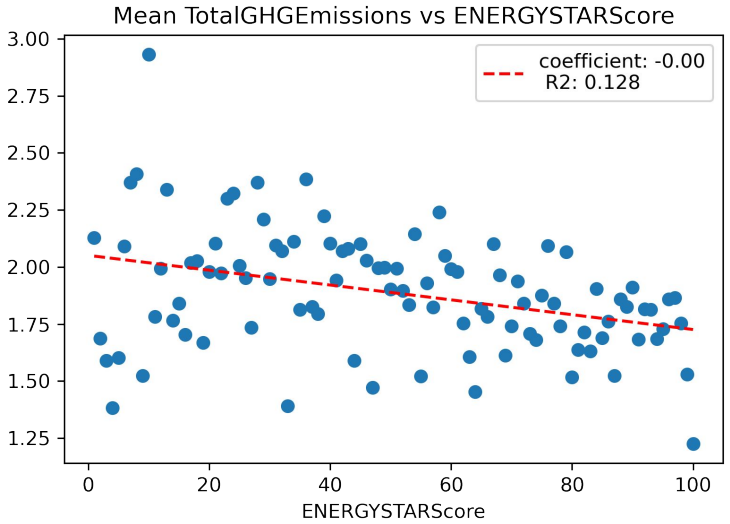
Comparaison des modèles





Pertinence d'ENERGYSTARScore pour la prédiction de TotalGHGEmissions





→ Corrélation présente, mais faible entre ENERGYSTARScore et TotalGHGEmissions.
Evaluation réalisée par la comparaison de la qualité des prédictions avec et sans ESS.

Pertinence d'ENERGYSTARScore pour la prédiction de TotalGHGEmissions

Comparaison de trois cas:

- prédiction s'appuyant sur le jeu de données complet
- prédiction s'appuyant sur le jeu de données restreint aux seuls individus pour lesquels ENERGYSTARScore est renseigné
- ENERGYSTARScore inclus dans l'ensemble des prédicteurs.

	TotalGHGEmissions
jeu de données complet	0.84 +/- 0.02
Jeu de données restreint	0.88 +/- 0.02
Prise en compte d'ENERGYSTARScore	**0.92 +/- 0.01**

Modèle final

1. Modèle final
2. Pistes d’amélioration (1): Sélection des features
3. Pistes d’amélioration (2): modèles séquentiel vs non-séquentiel

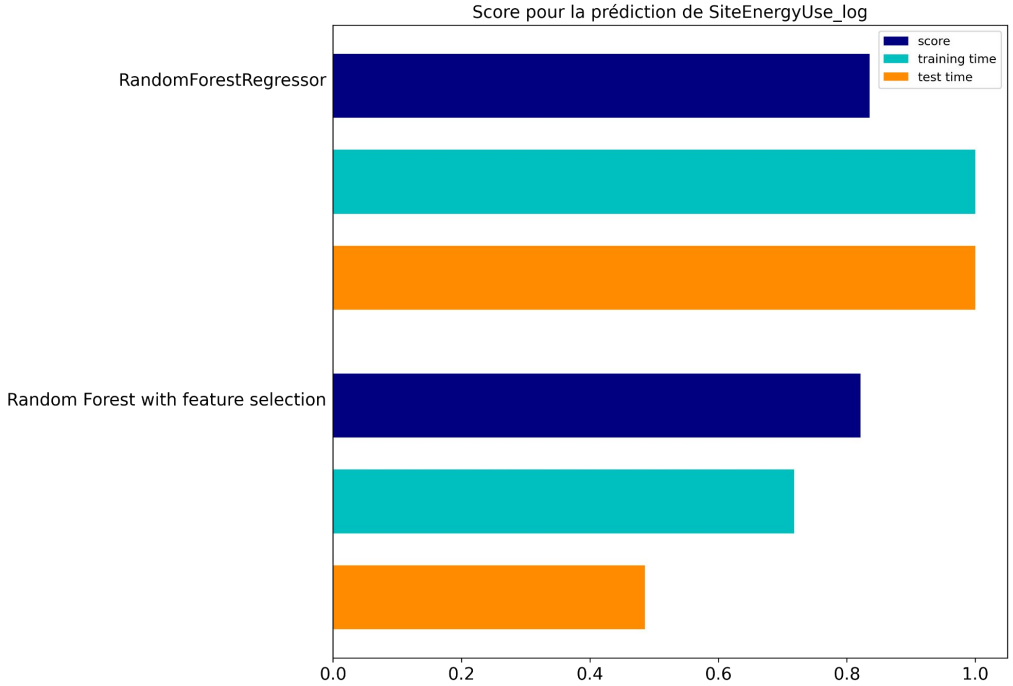
Modèle final

1. Sélection des inputs
2. PCA, normalisation, Onehot encoding
3. Prédictions avec **RandomForest**
4. Pistes d'améliorations:
    a. Sélection des features
    b. Modèle séquentiel vs modèle non-séquentiel

Modèle final - Améliorations (1): Sélection des features

Sélection des features pour lesquelles feature importance > 2.5%

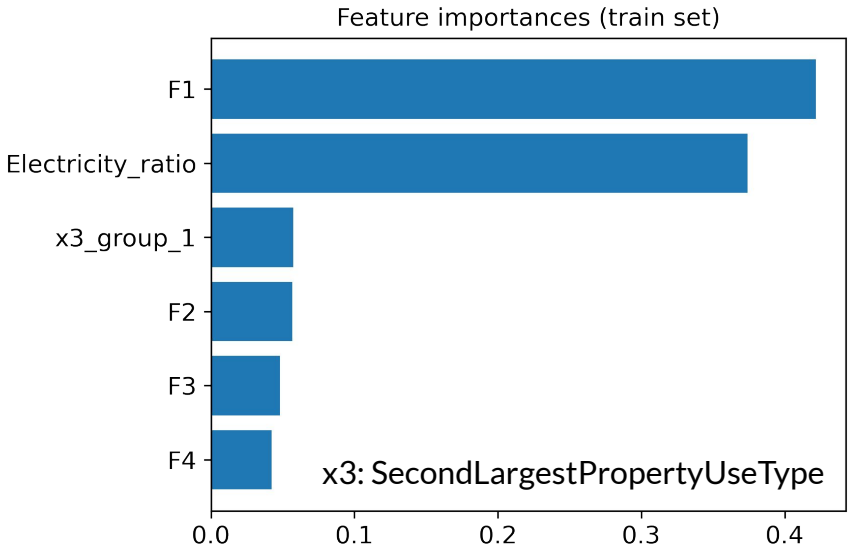
- Amélioration du temps d'entraînement
- Perte de performance maîtrisée



Modèle final - Améliorations (1): Sélection des features

Sélection des features pour lesquelles feature importance > 2.5%

- Amélioration du temps d'entraînement
- Perte de performance maîtrisée
- Meilleure interprétabilité



Modèle final - Améliorations (2): modèle séquentiel vs non-séquentiel

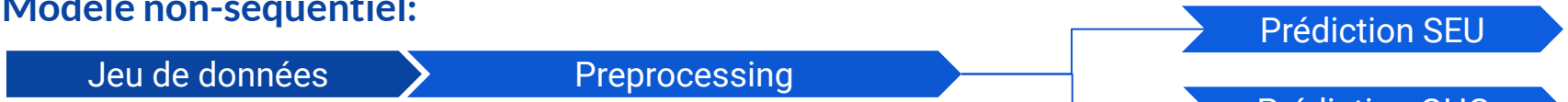
Modèle séquentiel:

jeu de données avec sélection des features "à la main"

prédictions sur train/test = 50/50

Prédiction de TotalGHGEmissions avec Random Forest, à partir des prédictions de SiteEnergyUse: validation croisée 3 folds sur 'test'

Modèle non-séquentiel:



jeu de données complet

- PCA
- standardisation
- One-hot encoding
- Feature selection

validation croisée 5 folds

Résultats finaux

	EnergyStarScore	SiteEnergyUse	TotalGHGEmissions
Modèle séquentiel	sans ESS	0.82 +/- 0.09	0.82 +/- 0.02
	avec ESS	**0.90 +/- 0.04**	**0.92 +/- 0.01**
Modèle non-séquentiel	sans ESS	0.84 +/- 0.03	0.85 +/- 0.02
	avec ESS	**0.91 +/- 0.01**	**0.92 +/- 0.03**

- Prendre en compte ENERGYSTARScore permet d'améliorer les prédictions, à la fois par la mise à l'écart d'éléments introduisant du bruit dans le modèle et la prise en compte d'une variable corrélée avec la cible
- Les résultats sur chaque modèle sont équivalents