Consommations / Emissions des bâtiments dans la ville de Seattle

présentation du 19 juillet 2021

Plan de la présentation

1) Problématique de la ville de Seattle

2) Description des données

3) Nettoyages des données et exploration

4) Intérêt de l'Energy Star Score

5) Modélisation des consommations d'énergie

6) Modélisation des émissions de gaz à effet de serre

7) Conclusion

Problématique de la ville de Seattle

Contexte : réduire les émissions de gaz à effet de serre (GHG).

Objectif de la municipalité: disposer des données « consommation d'énergie » et « émission de GHG » pour les bâtiments non destinés à l'habitation.

Moyens déjà mis en œuvre : relevés minutieux déjà effectués sur plusieurs milliers de bâtiments.
Nombreuses données récoltées pour chaque bâtiment :
- données du permis d'exploitation commerciale : usages des bâtiments, surfaces des principaux usages, données de géolocalisation, …
- relevés de consommation / émission : types d'énergies consommées, émissions de GHG, Energy Star Score.

Problématique de la ville de Seattle

Problème :

- relevés de consommation / émission : coûteux à obtenir.

- données du permis d'exploitation commerciale : plus faciles à acquérir.

Missions confiées :

- prédire les consommations d’énergie et les émissions de GHG sur la base des données déclaratives du permis d'exploitation commerciale

- évaluer l’intérêt de l’Energy Star Score pour la prédiction des émissions.

Présentation des données

Source : www.kaggle.com/city-of-seattle/sea-building-energy-benchmarking
2 fichiers json de métadonnées (2015, 2016) et 2 fichiers csv de données (2015, 2016)

Description des données 2016 :

- 3376 lignes (« properties »).

- 45 colonnes (variables) : localisation, différents usages commerciaux, surfaces, différentes consommations d'énergie (électricité, vapeur, gaz naturel), émissions …

- Variables d'intérêt pour les modèles de prédiction à concevoir :
    - *SiteEnergyUse(kBtu)* : consommation d'énergie par « property » en kBtu / an.
    - *TotalGHGEmissions* : émissions par « property » en équivalent tonne de CO_2 / an.

Description des données 2015 : 3340 properties, 48 variables (similaires à celles de 2016).

Données retenues

On choisit de ne travailler qu'avec les données de 2016.

Raisons :

- union des datasets 2015 et 2016 ⇒ 3432 properties dans le nouveau dataset (gain de seulement 1.7%).

- risque de joindre des données qui ont disparu pour une raison non déterminée entre 2015 et 2016.

Etanchéité des données

Séparation des données en :
- jeu d'entraînement.
- jeu de test,

avant les étapes de nettoyage / exploration pour limiter les fuites de données.

Ratio de split : 4 / 1 → 3376 = 2700 + 676 (lignes)

Etapes de nettoyage réalisées sur le jeu de test après exploration : à l'identique du jeu d'entraînement.

Nettoyages des données

Suppression de variables :
Raisons :
- modalité unique (*DataYear*, *City*, *State*, …),
- modalité différente pour chaque bâtiment (*PropertyName*, *TaxParcelIdentificationNumber*, *Location*, …),
- trop peu de bâtiments renseignés (*YearsENERGYSTARCertified*, *ComplianceStatus*, *Outlier*),
- non pertinent pour notre problème (*Comment , DefaultData*).

Suppression des bâtiments destinés à l'habitation :
Variable *BuildingType* avec les modalités : *Multifamily LR (1-4)*, *Multifamily MR (5-9)*, *Multifamily HR (10+)*
2700 - 1365 = 1335 lignes dans le train set

Suppression des lignes pour lesquelles la valeur à prédire n'est pas renseignée :
- 16 lignes supprimées pour la modélisation de *SiteEnergyUse(kBtu)*,
- 6 lignes supprimées pour la modélisation de *GHGEmissionsIntensity*.

Nettoyages des données

Gestion des valeurs atypiques :
Conservation des properties aux valeurs atypiques (Z-scores élevés).
Principalement des grands hopitaux et des campus universitaires.
Suppression non judicieuse car plus gros consommateurs d'énergie et émetteurs de GHG.

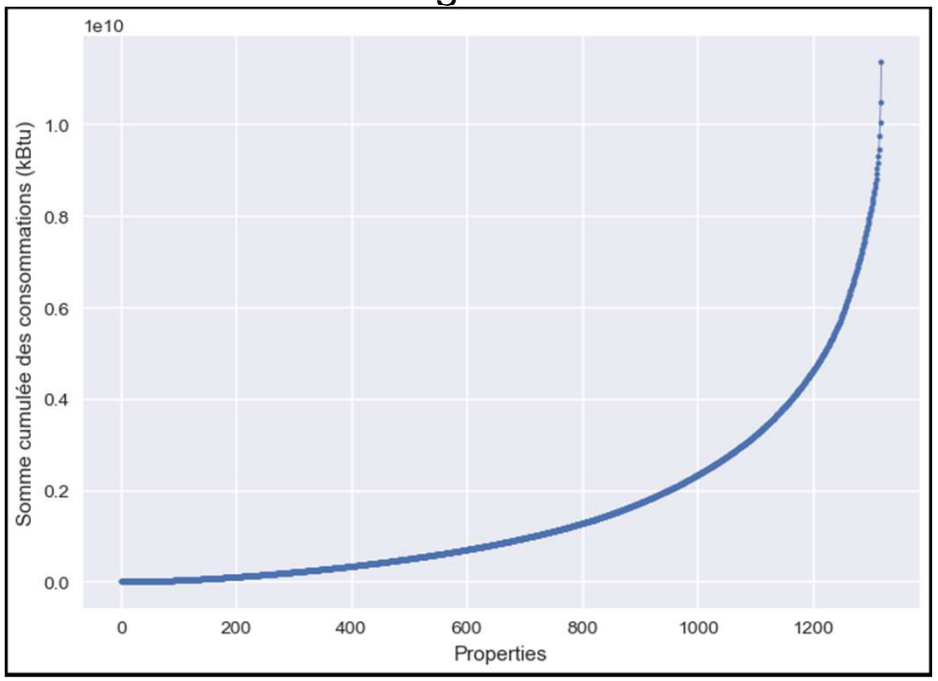
Remplacement de valeurs à zéro :
Valeurs à zéro remplacées par des NaN pour des variables ne pouvant être nulles : *PropertyGFATotal*, *LargestPropertyUseTypeGFA*, *SiteEnergyUse(kBtu)*, *GHGEmissionsIntensity*, …

Conclusion sur le nettoyage :
- 30 variables supprimées ; 15 restantes avant feature engineering.
- Suppression des properties non commerciales (soit la moitié du dataset).
- Très peu de properties supprimées par ailleurs.

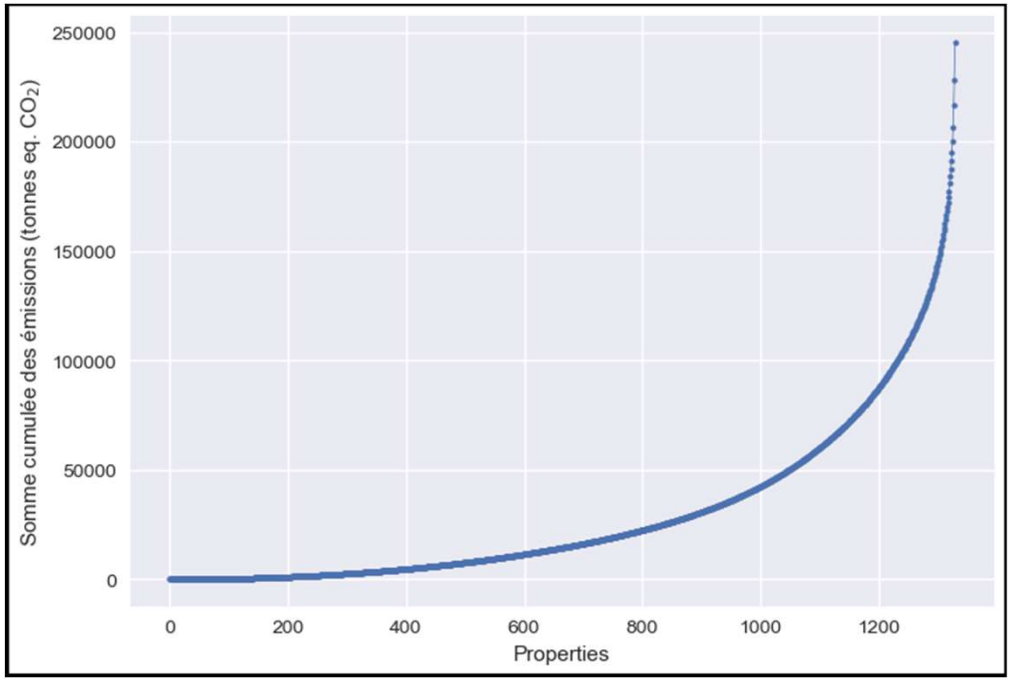
Exploration – variables à prédire

Somme cumulée pour les properties du train set :

Consommation d'énergie :



Emission de GHG :



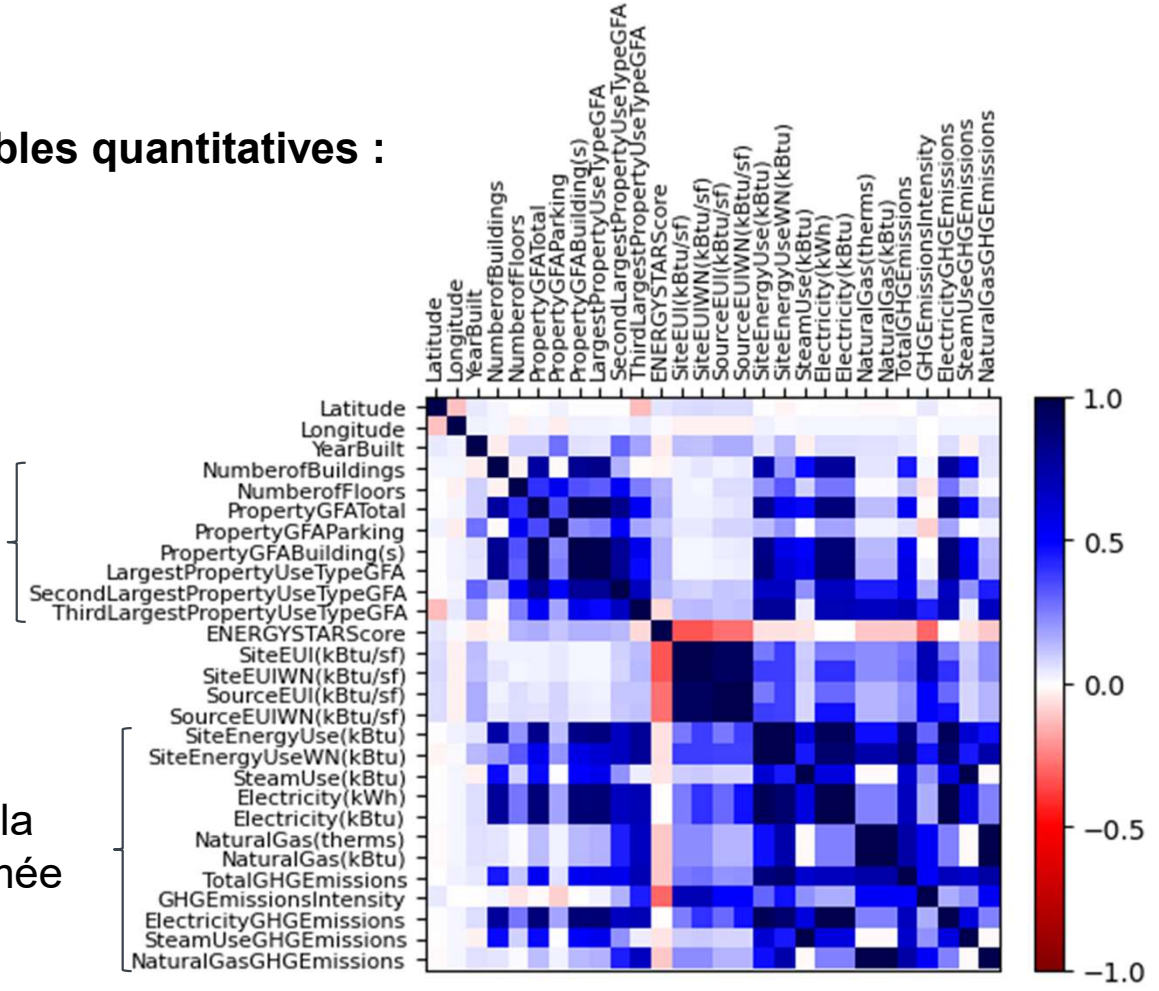
Les 10 principales properties du dataset émettent un quart des GHG.
⇒ pertinence de la conservation des données atypiques

Exploration – Matrice de corrélation linéaire (r, Pearson)

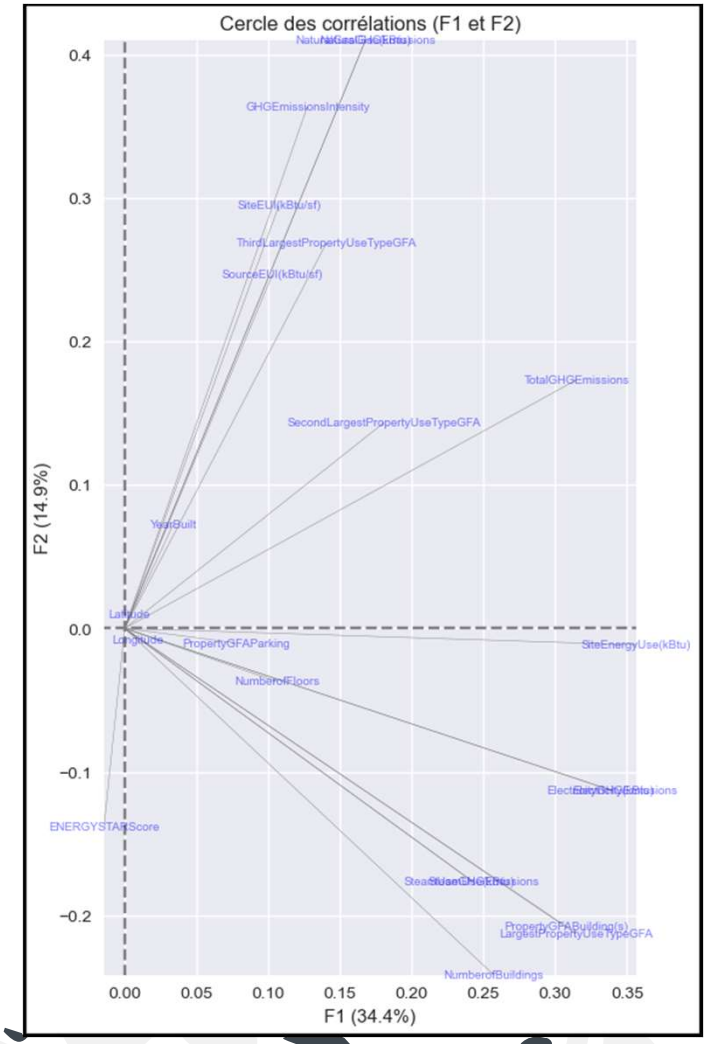
Pour les principales variables quantitatives :

Variables en lien avec la taille de la property

Variables en lien avec la qté d'énergie consommée



Exploration – analyse en composantes principales



ACP réalisée après suppression des variables fortement corrélées (r>0.99).

Interprétation du **premier plan factoriel** :
- les variables les plus corrélées à **F1** sont les variables corrélées à la **quantité d'énergie consommée** par property.
- les variables les plus corrélées à **F2** sont les variables corrélées à la **quantité d'énergie consommée par unité de surface**.

Feature engineering – surfaces des PropertyUseType

Croisement des **64 modalités** des variables de type PropertyUseType :
LargestPropertyUseType, SecondLargestPropertyUseType, ThirdLargestPropertyUseType

avec les **valeurs** des variables PropertyUseTypeGFA :
LargestPropertyUseTypeGFA, SecondLargestPropertyUseTypeGFA, ThirdLargestPropertyUseTypeGFA

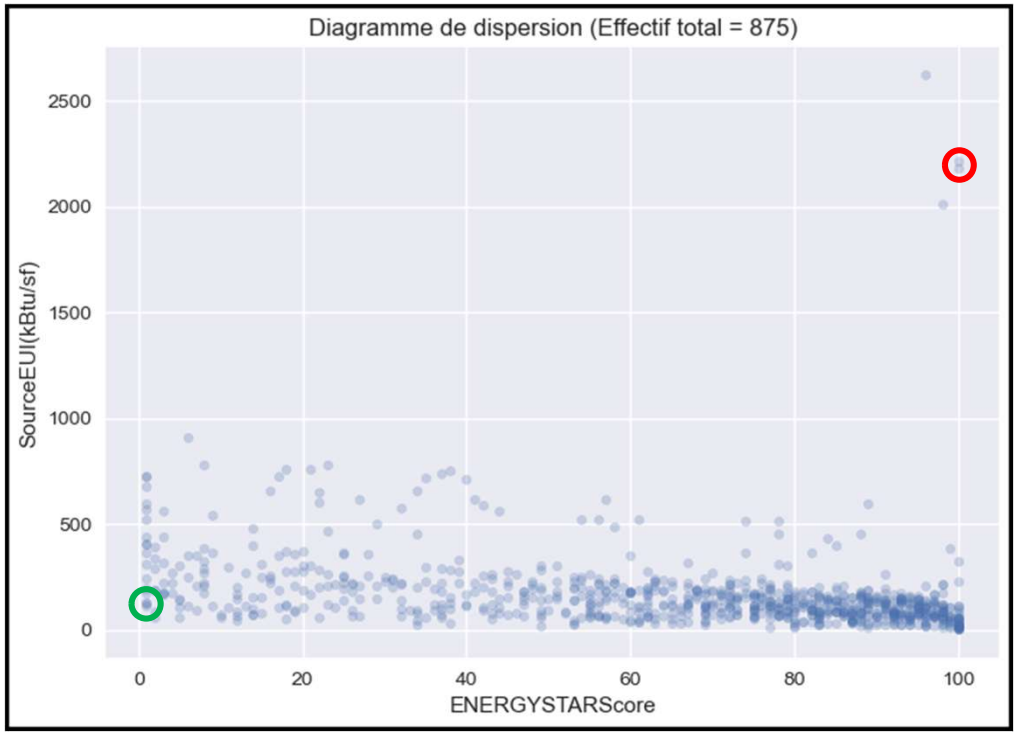
⇒ **Création de 64 variables** qui correspondent à la superficie pour la modalité considérée.

Exemple de variables créées : *OfficeGFA, HotelGFA, RestaurantGFA,* …

Intérêt de l'Energy Star Score

- Score qui a pour objet de mesurer **l'efficacité énergétique d'un bâtiment**

Normalisé à l'échelle des USA, de l'usage du bâtiment, de l'activité et des conditions météo.
Valeur de 1 à 100 correspondant au centile d'efficacité d'un bâtiment par rapport à ses pairs.
Note : il ne considère que l'usage principal du bâtiment (contrairement à nos modèles).



- Corrélation négative entre la variable *ENERGYSTARScore* et la variables *SourceEUI*, mais tendance très faible.

- Properties du premier centile (Score = 1) avec des consommations primaires (SourceEUI) inférieures à celles du dernier centile.

- Exemple extrême du dataset : datacenter (score 100) avec EUI 19 fois supérieur à un lieu de culte (score 1).

⇒ **Energy Star Score non adapté pour les prédictions** d'émissions de GHG

Modélisations - généralités

Module : scikit-learn.

Métrique : choix du R^2 (coefficient de détermination linéaire de Pearson).

Validations croisées / optimisations / régularisations :
- métrique retenue : moyenne des R^2 de n plis.
- d'abord n=10 plis : valeurs de R^2 trop dispersées.
- au final : on a retenu 5 plis.

Imputations des valeurs manquantes :
- variables quantitatives → imputation par la valeur médiane.
- variables qualitatives → imputation par la modalité «ND».

Encodage des variables qualitatives : one-hot encoding.

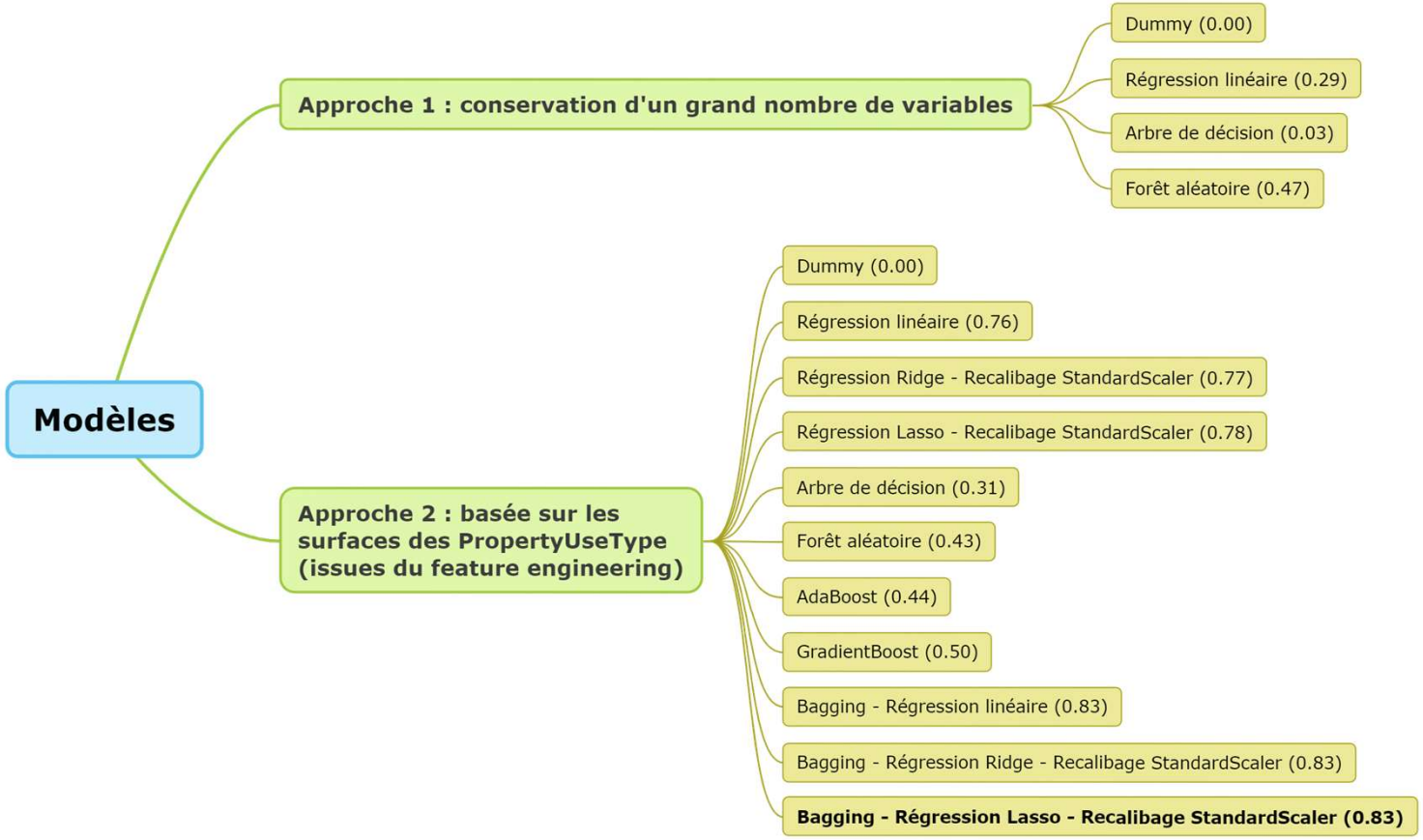
Modélisation des consommations d'énergie

Variable à prédire : *SiteEnergyUse(kBtu)*

Première approche :
- prédiction de la variable proxy : *SiteEUI(kBtu/sf)* (donnée intensive)
- utilisation de toutes les variables du dataset post-nettoyage, à l'exclusion des données extensives (*PropertyGFATotal, LargestPropertyUseTypeGFA*, etc…).
- d'où les variables retenues :
  - qualitatives (8) : *BuildingType, PrimaryPropertyType, ZipCode, CouncilDistrictCode, Neighborhood, LargestPropertyUseType, SecondLargestPropertyUseType, ThirdLargestPropertyUseType*
  - quantitatives (5) : *Latitude, Longitude, YearBuilt, NumberofBuildings, NumberofFloors*

Deuxième approche :
- prédiction directe de *SiteEnergyUse(kBtu)* (donnée extensive).
- utilisation des surfaces des PropertyUseType (issues du feature engineering).
- 64 variables, toutes quantitatives.

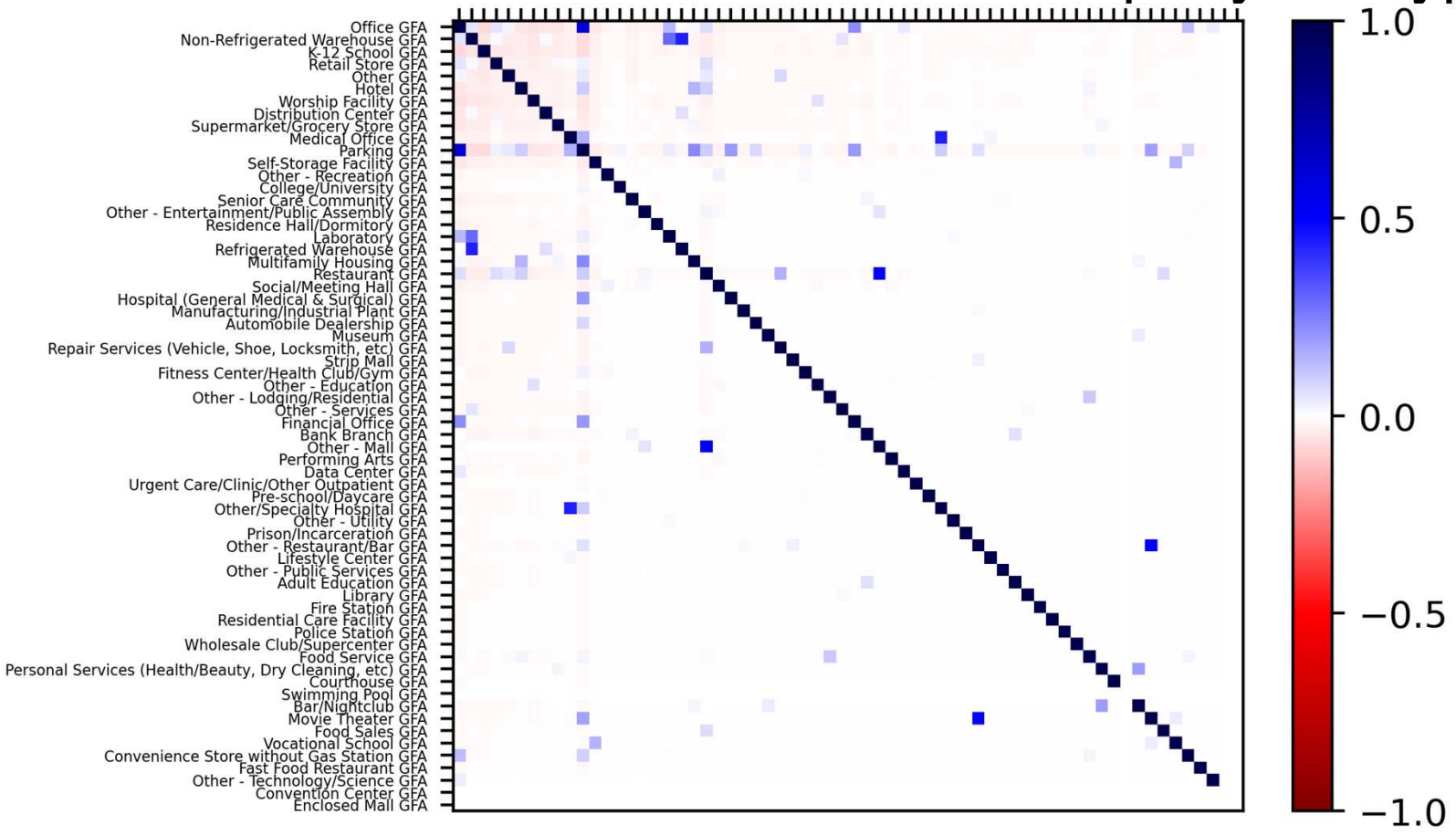
Modélisation des consommations d'énergie



Les chiffres entre parenthèses indiquent le R^2 obtenu (CV 5 plis).

Approche 2 : pas de gain significatif par régularisation (cf. heatmap des variables).

Feature engineering – Matrice de corrélation linéaire (r, Pearson) des surfaces des PropertyUseType



Rares occurrences de corrélations entre ces variables
⇒ impact sur régularisations

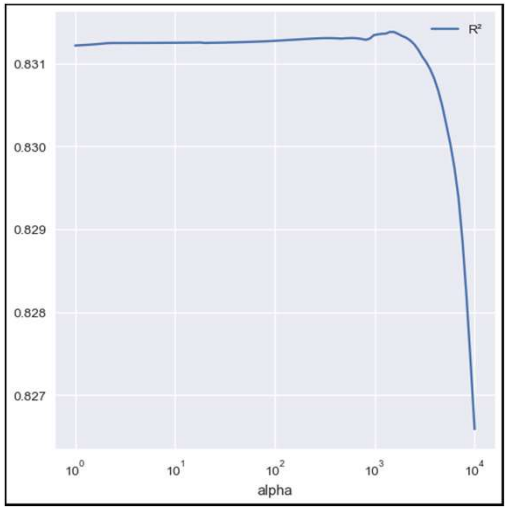
Exemples de corrélation :
- Parking / Office (r=0.6)
- Mall / Restaurant (r=0.5)

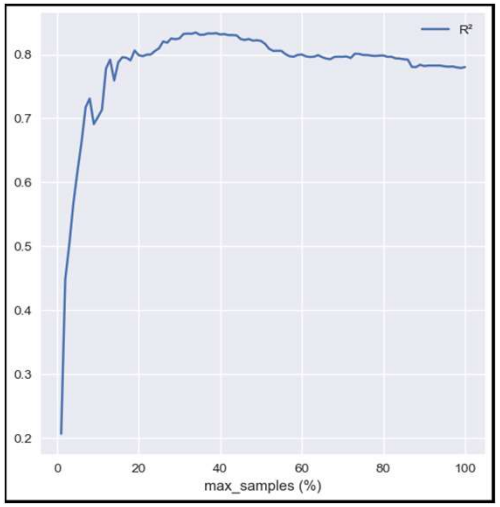
Modélisation des consommations d'énergie

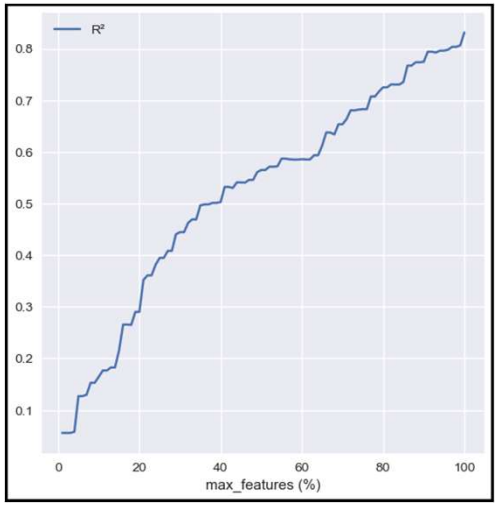
Hyperparamètres du modèle ensembliste retenu (BaggingRegressor – approche directe – $R^2 = 0.83$) :
- Lasso, alpha = 1000,
- bootstrap = True,
- max_samples = 40%,
- max_features = 100%,
- n_estimators = 12.

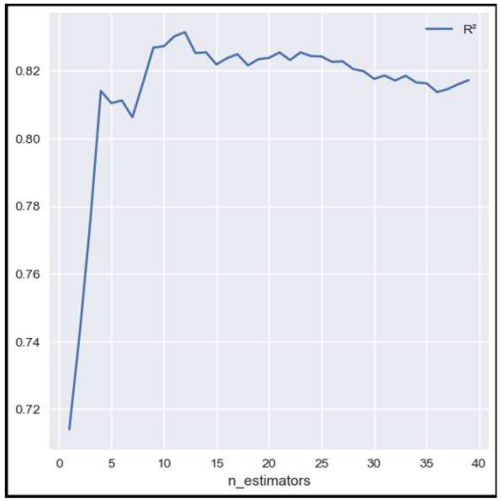
Variabilité des plis :
$R^2 = 0.83$ est la moyenne des R^2 de 5 CV :0.70, 0.78, 0.82, 0.90, 0.96 .

Stabilité des optima des hyperparamètres :





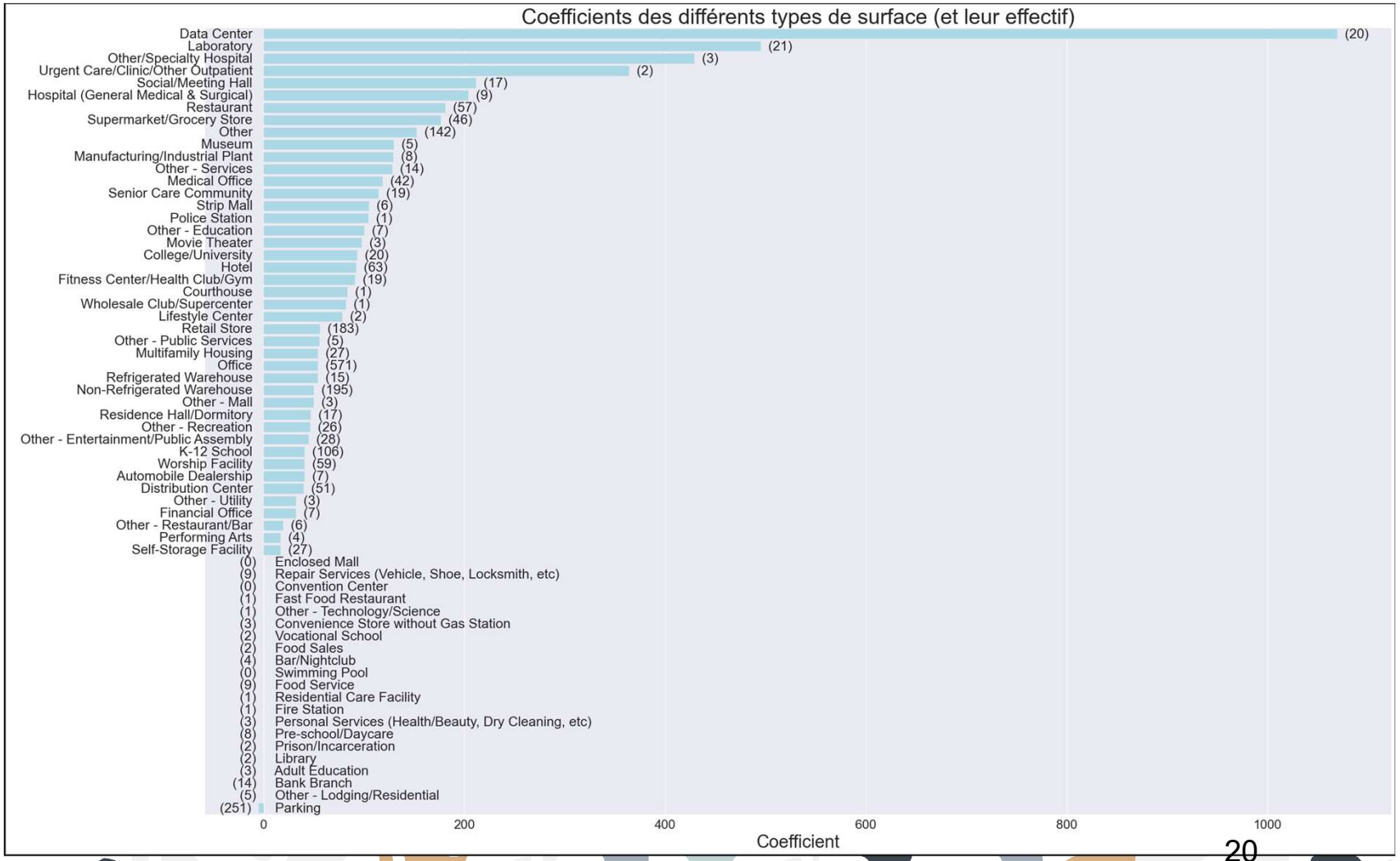




Modélisation des consommations d'énergie

Estimation des coefficients du modèle :

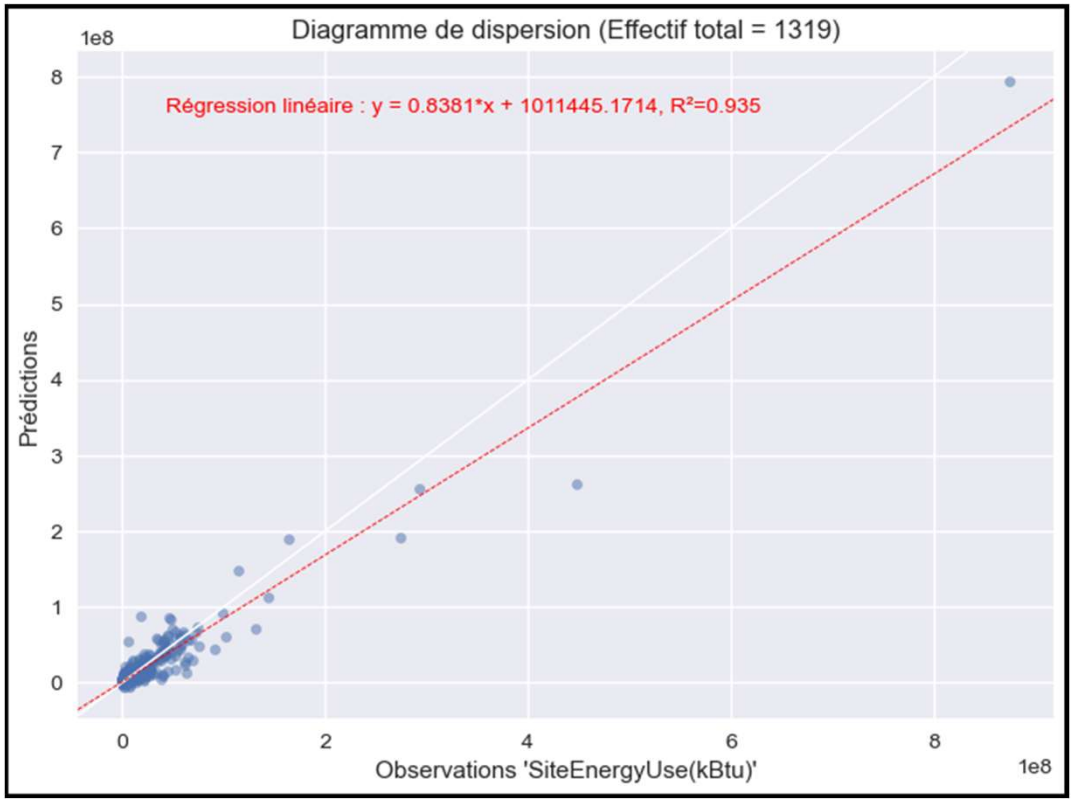
- Utilisation d'un proxy : modèle non ensembliste.
- Tracé avec des données non recalibrées pour avoir l'interprétabilité du modèle.
- Conso E = Σ (surface * coef) avec surface en sq feet, et énergie en kBtu.
- Top 3 des contributeurs par unité de surface : data center, laboratory, other/specialty hospital.



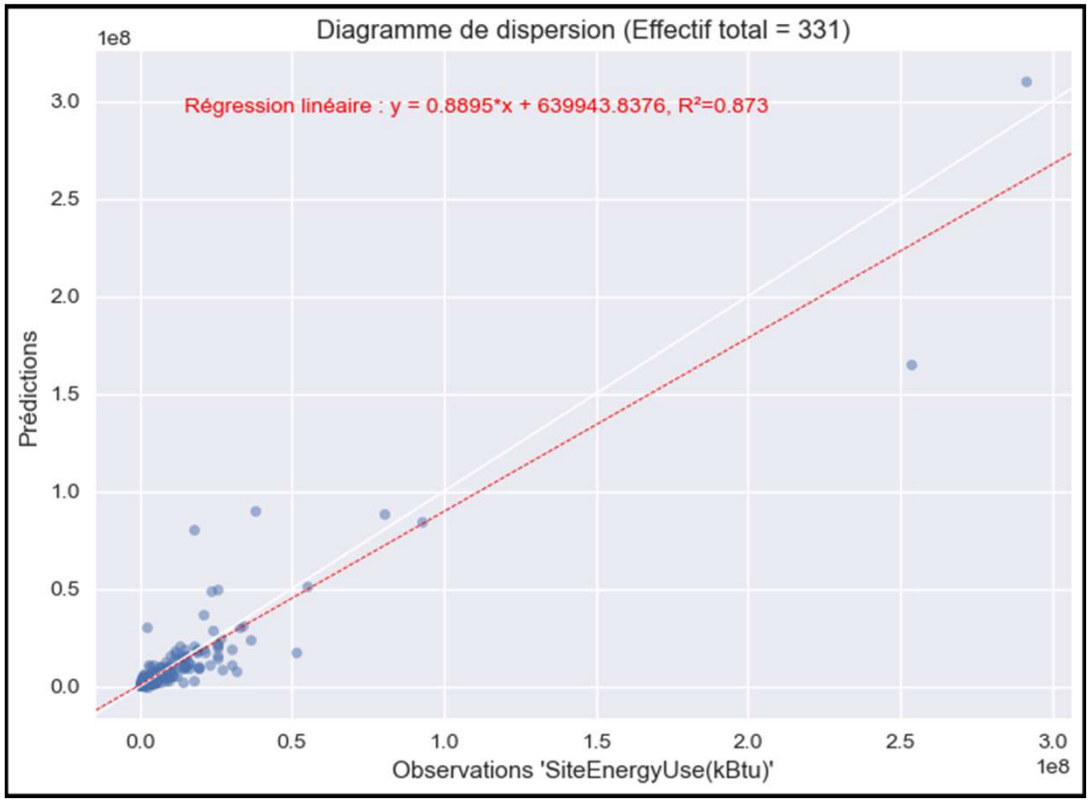
Modélisation des consommations d'énergie

Généralisation du modèle retenu (Bagging – régression Lasso– recalibrage StandardScaler) :
(rappel : score de cross-validation = 0.83)

Données d'entraînement (R^2=0.93) **:** Prédictions = f(Observations)



Données de test (R²=0.87) : Prédictions = f(Observations)



Modélisation des émissions de gaz à effet de serre

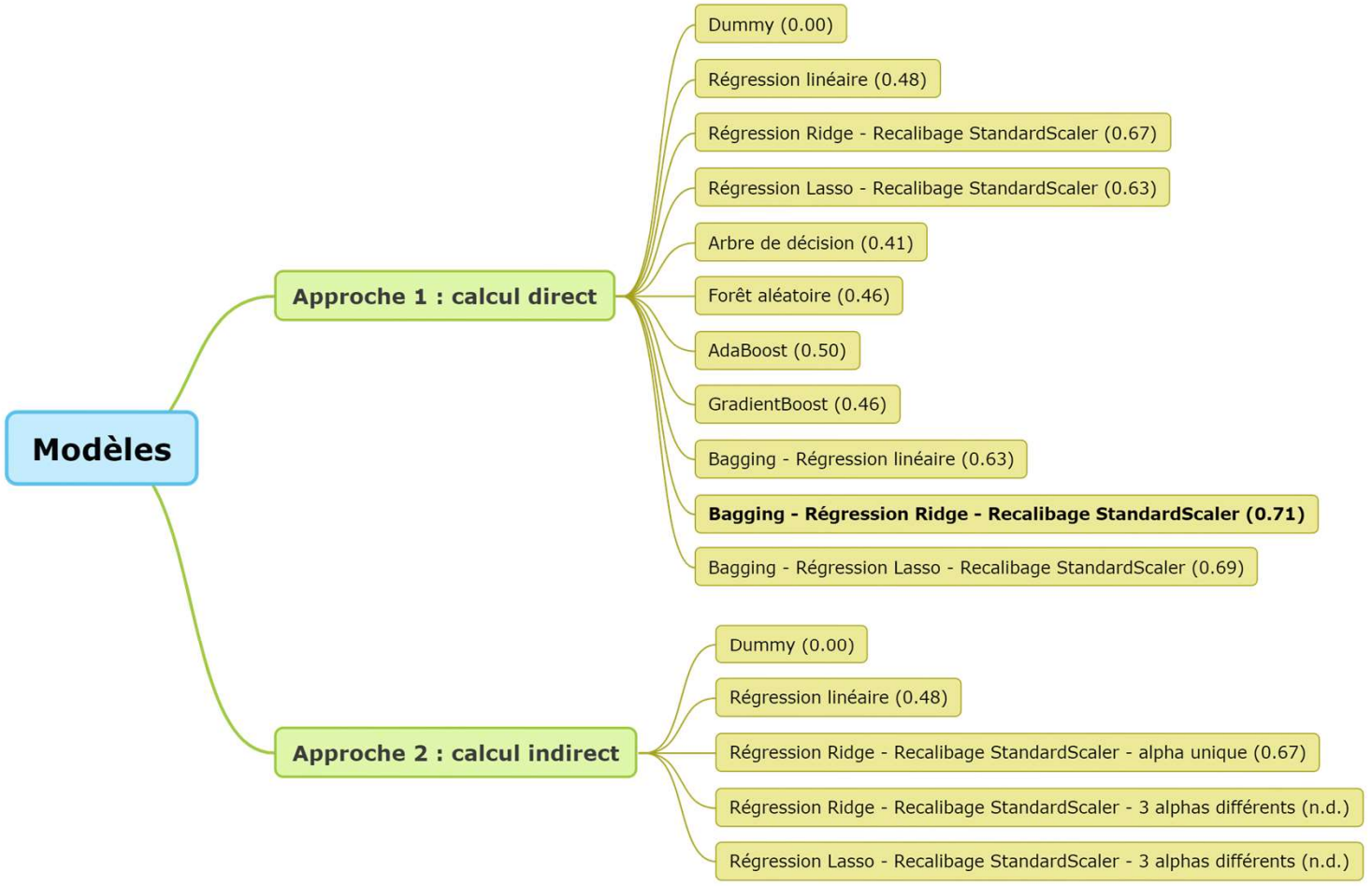
Variable à prédire : *TotalGHGEmissions*.

Variables utilisées : surfaces des PropertyUseType (idem modélisation consommation d'énergie).

Première approche :
- détermination directe de la variable à prédire.

Deuxième approche :
- détermination indirecte de la variable à prédire.
- prédiction des variables (valeurs obtenues par feature engineering des données de consommation) :
  - *ElectricityGHGEmissions*,
  - *SteamUseGHGEmissions*,
  - *NaturalGasGHGEmissions*.
- *TotalGHGEmissions* est la somme de ces 3 variables.

Modélisation des émissions de gaz à effet de serre



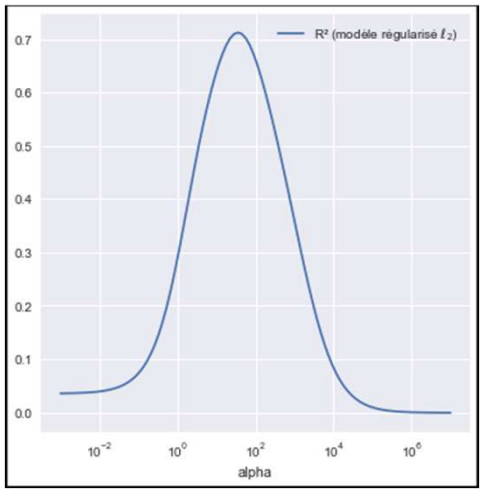
Les chiffres entre parenthèses indiquent le R^2 obtenu (CV 5 plis).

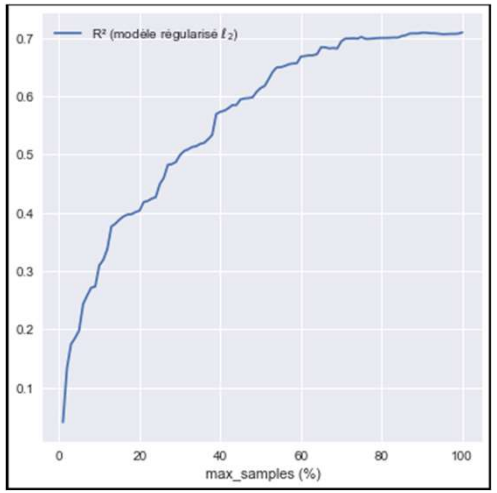
Modélisation des émissions de gaz à effet de serre

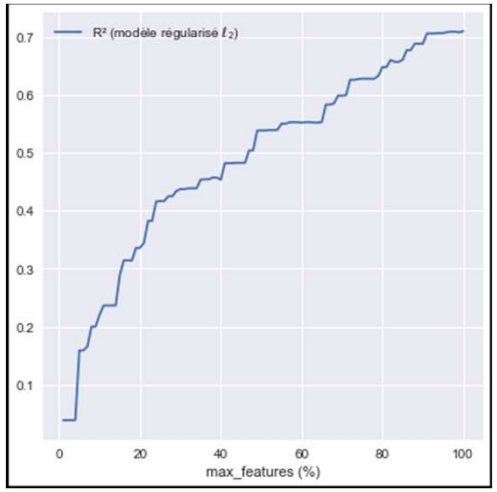
Hyperparamètres du modèle ensembliste retenu (BaggingRegressor – approche directe – $R^2 = 0.71$) :
- Ridge, alpha = 46,
- bootstrap = True,
- max_samples = 100%,
- max_features = 100%,
- n_estimators = 10.

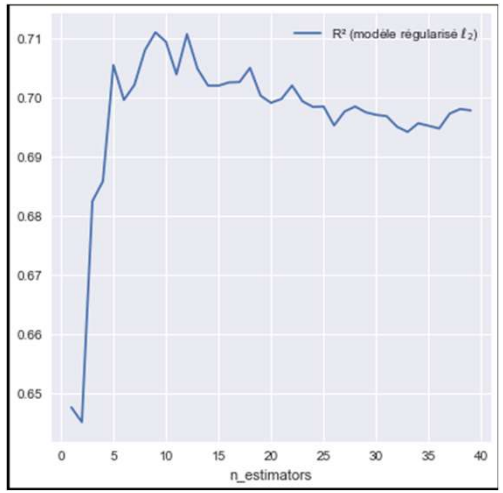
Variabilité des plis :
$R^2 = 0.71$ est la moyenne des R^2 de 5 CV : 0.47, 0.52, 0.84, 0.84, 0.87 .

Stabilité des optima des hyperparamètres :





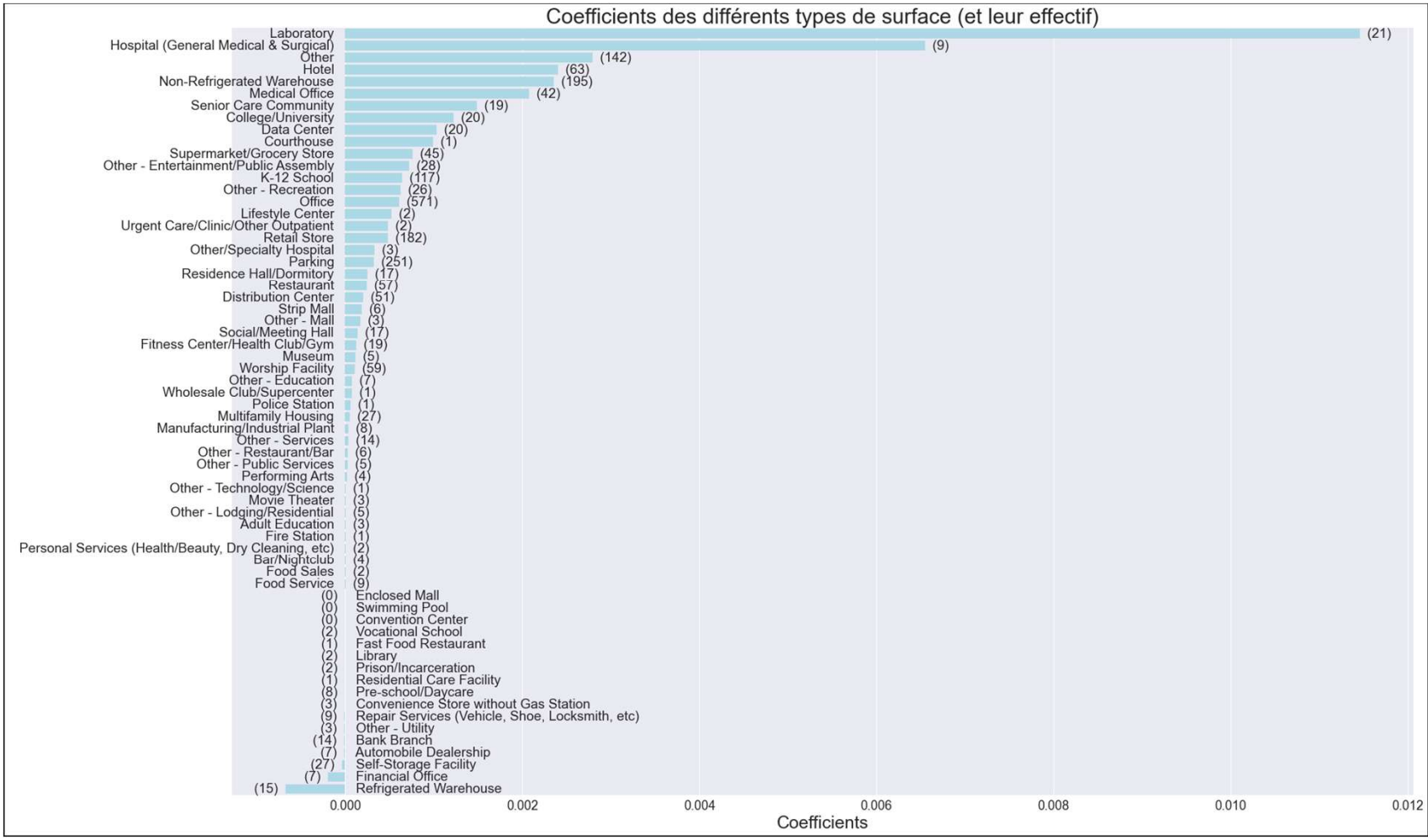




Modélisation des émissions de gaz à effet de serre

Estimation des coefficients du modèle :

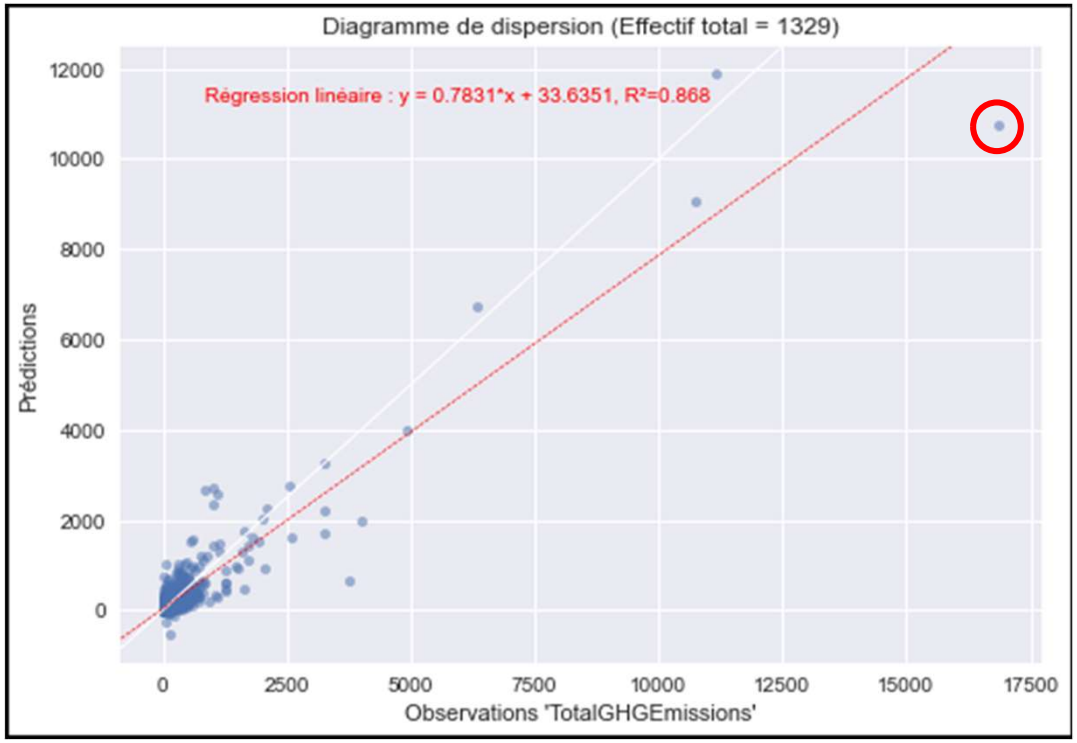
- Utilisation d'un proxy : modèle non ensembliste.
- Tracé avec des données non recalibrées pour avoir l'interprétabilité du modèle.
- Emission = Σ (surface * coef) avec surface en sq feet, et émission en tonne éq. CO_2.
- Top 3 des contributeurs par unité de surface : laboratoires, hopitaux, hotels.



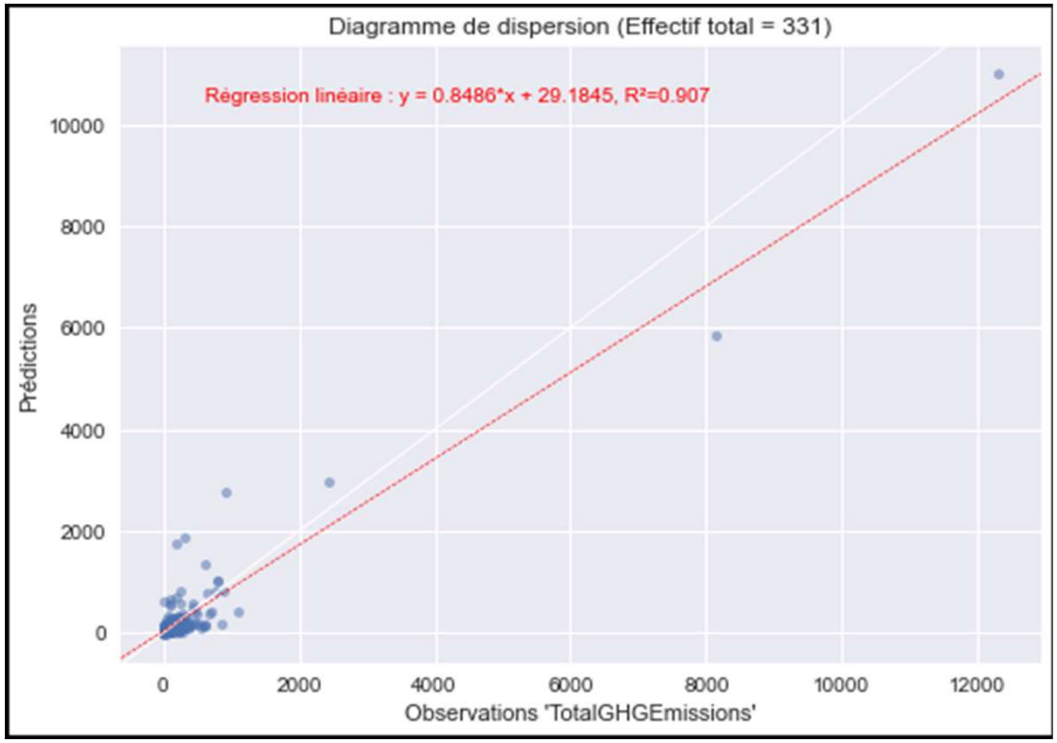
Modélisation des émissions de gaz à effet de serre

Généralisation du modèle retenu (Bagging – régression Ridge – recalibrage StandardScaler) :
(rappel : score de cross-validation = 0.71)

Données d'entraînement (R^2=0.87) : Prédictions = f(Observations)



Données de test (R^2=0.91) : Prédictions = f(Observations)



Conclusion

Modèles retenus :

	Consommation d'énergie	Émission de GHG
Recalibrage	StandardScaler	
Modèle	Ensembliste avec bootstrap	
	Lasso	Ridge
alpha	1000	46
max_samples (%)	40	100
max_samples (%)	100	100
n_estimators	12	10
R^2 généralisation sur training set	**0.87**	**0.91**

Conclusion

Taille du dataset :
Suffisant, mais conduit à une forte variabilité entre les plis lors de la cross-validation.

Régularisation des modèles Ridge / Lasso :
Améliorations peu significatives lors de la régularisation sur alpha, car faible corrélation entre les variables d'entrée (surfaces des PropertyUseType).

Bagging des meilleurs modèles :
Amélioration significative du R^2.

⇒ Les **données déclaratives** du permis d'exploitation commerciale ont permis de mettre au point des **modèles prédictifs** des consommations d'énergie et d'émissions de GHG.

Feature engineering – Emissions de GHG

Emissions de GHG décomposées en :
- ‘ElectricityGHGEmissions’,
- ‘SteamUseGHGEmissions’,
- ‘NaturalGasGHGEmissions’,
- calculées à partir des consommations d’énergie : 'Electricity(kWh)‘, 'SteamUse(kBtu)‘, 'NaturalGas(kBtu)‘.

En faisant la somme de ces 3 émissions, on boucle bien sur ‘TotalGHGEmissions’ :