OPENCLASSROOMS

P4 : Anticipez les besoins en consommation de bâtiments

Seattle City

30/05/2022

DUBART Maxime

Projet - Prédictions conso. bâtiments

Objectif: ville neutre en émissions de carbone en 2050

Prédire la consommation et les émissions des bâtiments non destinés à l'habitation

Evaluation de l'intérêt de l'ENERGIE STAR Score pour prédire ces consommations/émissions

Données

Données Ville Seattle (https://data.seattle.gov/dataset/2016-Building-Energy-Benchmarking/2bpz-gwpy)

- Relevés annuels (2015 & 2016)
  - Données générales (e.g., id, nom, **adresse**)
  - **Type d'utilisation** & **Surfaces** (e.g., bureaux, médical, scolaire, etc.)
  - **Relevés de consommation d'énergie (depuis différentes sources) & émissions GHG**
  - **EnergyStarScore**
  - Appartenance à différents découpages géographiques
- env. 3500 lignes x 45 colonnes

Données

Vérification et nettoyage des données

- Concaténation des deux jeux de données (2015/2016)
  - Structures légèrement différentes : renommage de certaines colonnes
  - Gestion des dupliquas : conserve les valeurs les plus récentes

- Filtres
  - Suppression des bâtiments résidentiels (~ 50%)
  - Filtre des lignes sur certaines variables (Comment, ComplianceStatus, DefaultData, Outliers)
  - Suppression de features non pertinentes (State, City, DataYear, PropertyName, Address, etc.)
  - Nettoyage et sélection des données numériques
    - Données sont souvent données brutes et normalisées par le climat (WN) – ne conserve qu'une seule source d'information (+ qualité inférieure pour les données WN)
    - Données sont parfois données dans différentes unités – conserve la même unité pour toutes les variables (kBtu)

Données

Vérification et nettoyage des données

- Relations entre *features*
  - *Les surfaces*
    - Surface total = Parking + Bâtiments **=> ok**

Données

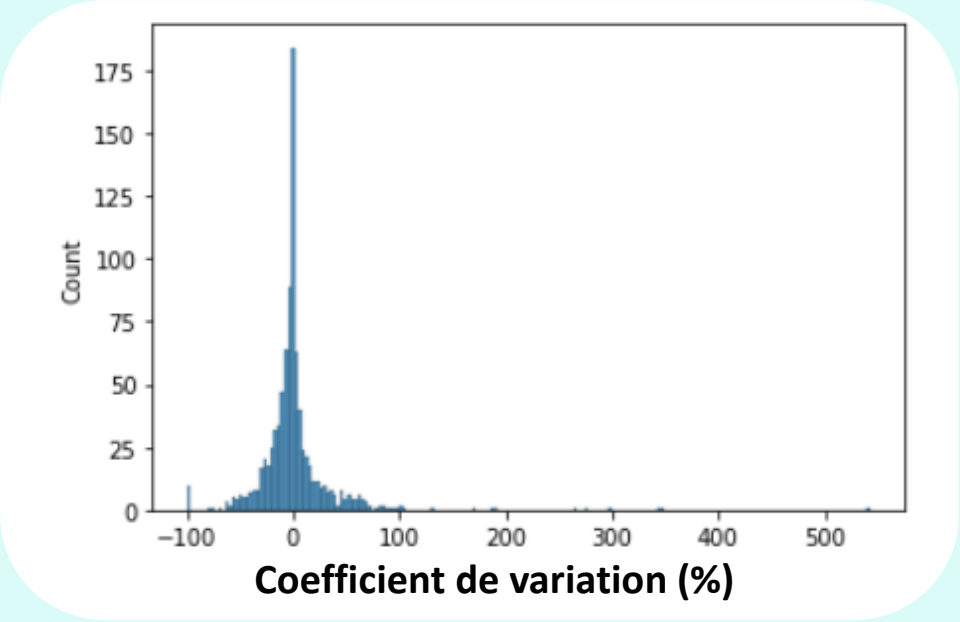
Vérification et nettoyage des données

- Relations entre *features*
  - *Les surfaces*
    - Surface total = Parking + Bâtiments **=> ok**
    - Forte corrélation Surface total ~ Surface bâtiment => voir après

Données

Vérification et nettoyage des données

- Relations entre *features*
  - *Les surfaces*
    - Surface total = Parking + Bâtiments **=> ok**
    - Forte corrélation Surface total ~ Surface bâtiment => voir après
    - Somme(différentes utilisations) = Surface total **=> nok**
      - **Rejet des features 'Type de bâtiment' & surfaces associées – I/II/III**



Données

Vérification et nettoyage des données

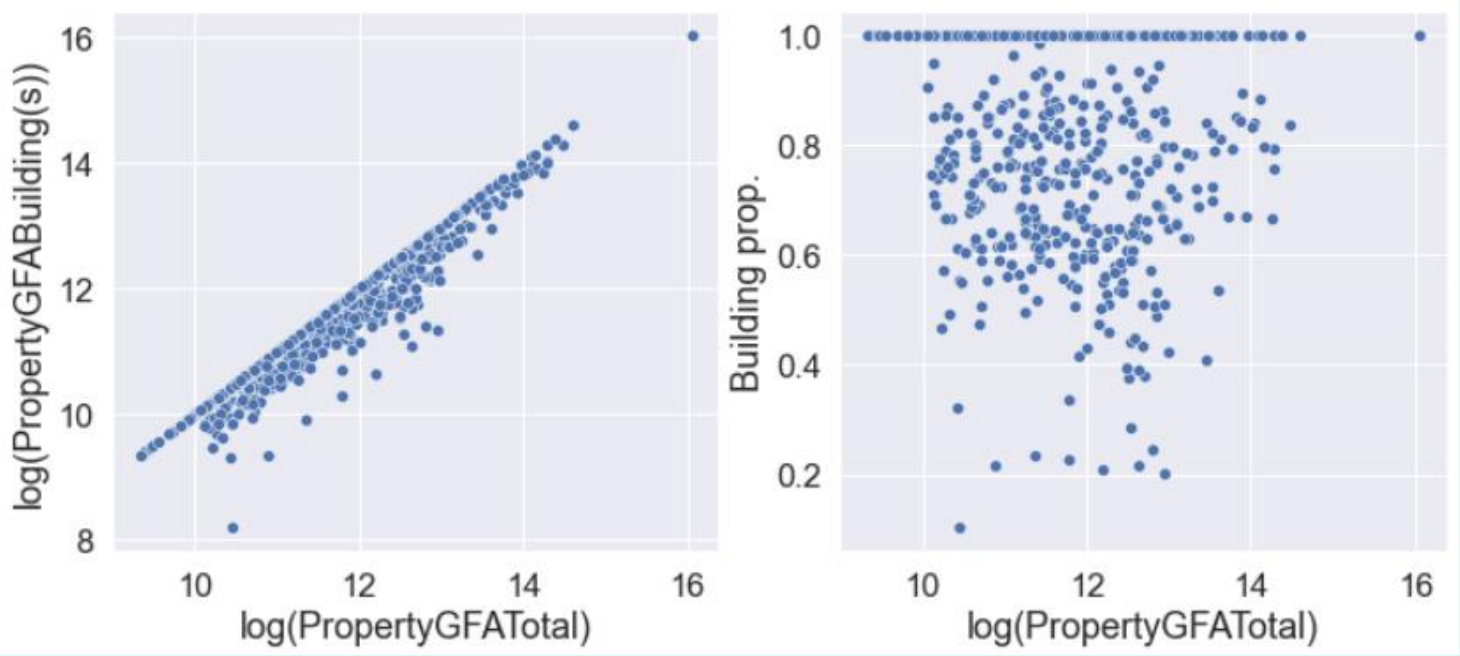
- Relations entre *features*
  - *Les énergies*
    - Energie totale = Electricité + Gaz + Vapeur **=> ok** (<1% avec CV > 1%)
    - Intensité d'utilisation de l'énergie = Energie total / Surface des bâtiments
    - Intensité d'émission de GHG = GHG Total / Surface totale

Données

***Features* transformation**

- Surfaces
  - Forte corrélation *Surface total ~ Surface bâtiments (0.99)*
  - *Degrés de libertés - lié à la relation entre les trois features*

  - Construction d'un nouvelle variable : « *proportion de bâtiment* »

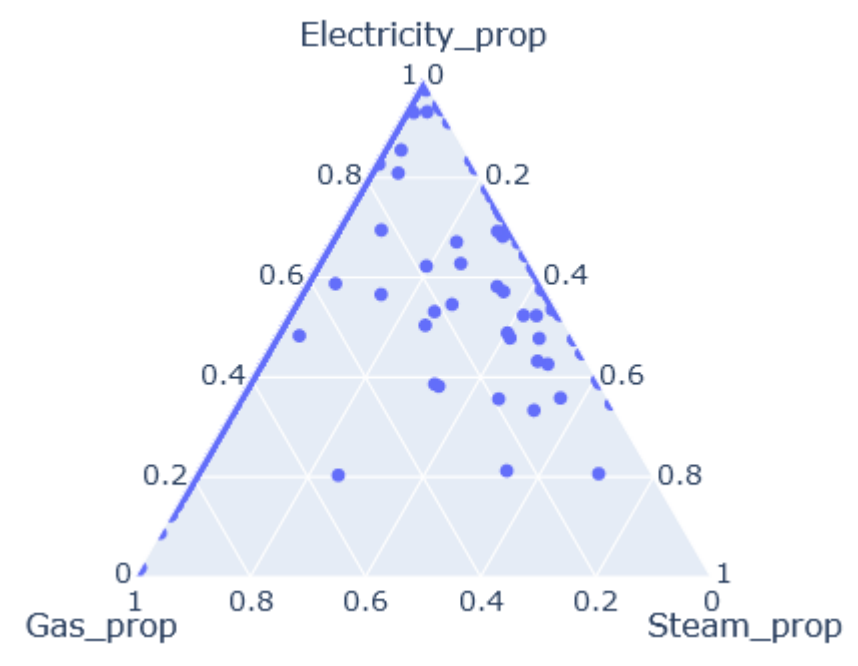


Données

***Features* transformation**

- Mix énergétique (Sources électricité, gaz et vapeur)
  - Degrés de libertés - lié à la relation entre les quatre features
  - Données non utilisables pour notre objectif

  - Construction de trois nouvelles variables *: « proportion de elec/gas/steam source»*



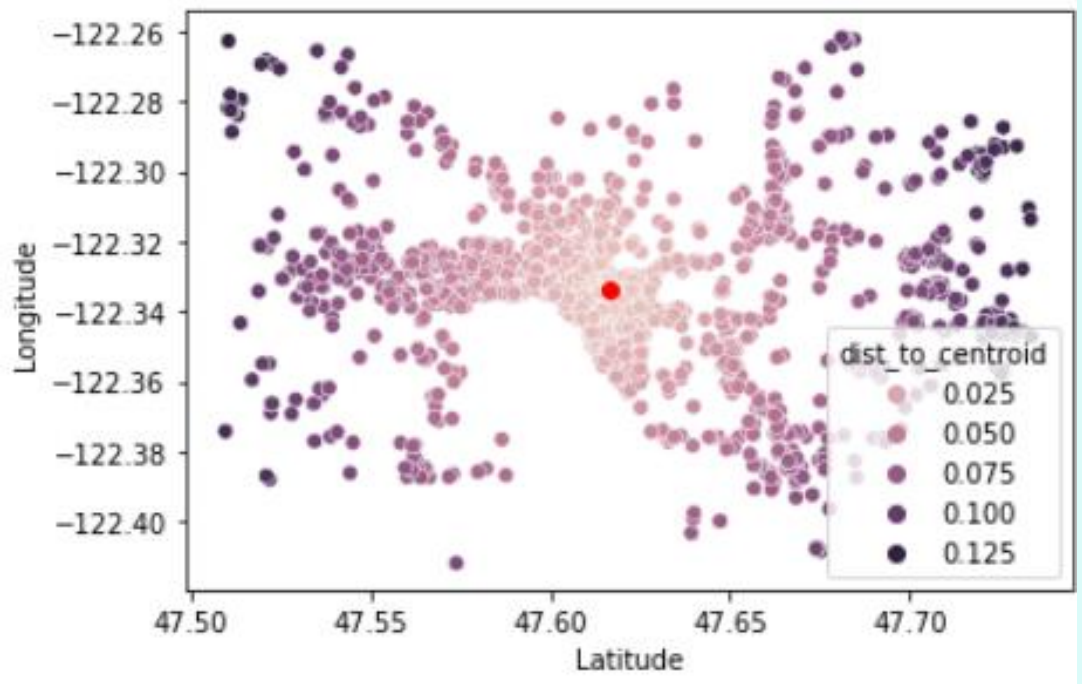
A noter : forte corrélation électricité ~ gaz !

Données

***Features* transformation**

- Types de bâtiments
  - Regroupement par types similaires (e.g., éducation, médical, etc.)

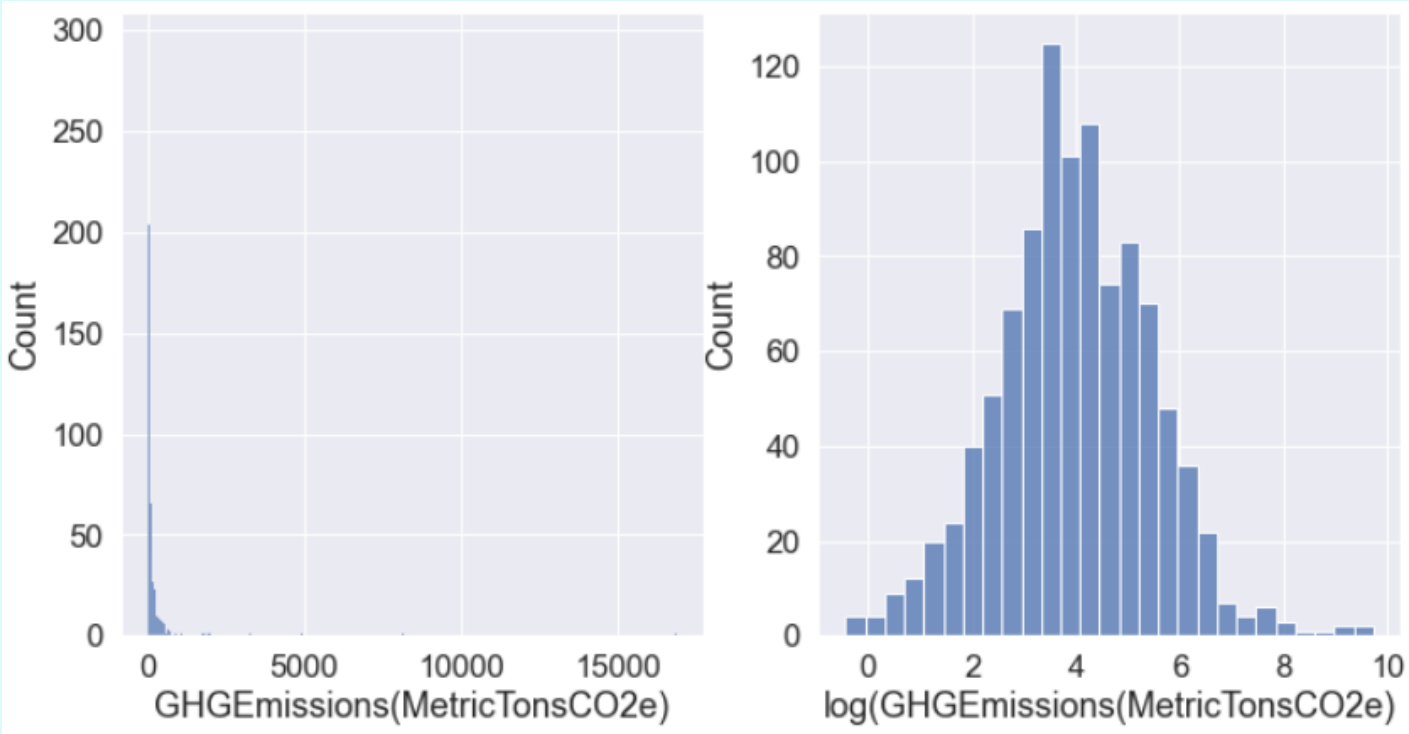
- Positions géographiques
  - Création d'un index de distance au centre-ville



Modèles

Préparation des données

- Passage au log() de la variable réponse



Note: modèles (linéaires) sont donc maintenant multiplicatifs

$$\log(y) = \sum \beta x \Rightarrow y = e^{\Sigma\, \beta x} = \prod e^{\beta x}$$

Modèles

Préparation des données

- Passage au log() de la variable réponse
- Transformation *features* catégorielles (type de bâtiment, et districts) en variables indicatrices
- *Features* numériques sont centrées-réduites

Modèles

Préparation des données

- Passage au log() de la variable réponse
- Transformation *features* catégorielles (type de bâtiment, et districts) en variables indicatrices
- *Features* numériques sont centrées-réduites

- Comparaison de quatre jeux de données :
  - #0 : **Avec** ENERGYSTARScore & **avec** proportions d'énergie
  - #1 : **Avec** ENERGYSTARScore & **sans** proportions d'énergie
  - #2 : **Sans** ENERGYSTARScore & **avec** proportions d'énergie
  - #3 : **Sans** ENERGYSTARScore & **sans** proportions d'énergie

- Les autres features : Année construction, Surface total, Proportion de bâtiments, Distance au centre, type de bâtiment et district.

- Split entrainement, test (5 folds), et validation (20%)

Modèles

Modèles utilisés

- Modèle linéaire
- Modèle kNN
- Modèle Ridge
- Modèle Lasso
- Modèle Kernel ridge (polynomial ou gaussien)

- SVR linéaire
- SVR à kernal (poly. ou gaussien)

- Multilayers perceptron

- Arbre de décision
- Forêt aléatoire
- XGBoost

Modèles

Modèles utilisés

- Modèle linéaire
- Modèle kNN
- Modèle Ridge
- Modèle Lasso
- Modèle Kernel ridge (polynomial ou gaussien)

- SVR linéaire
- SVR à kernal (poly. ou gaussien)

- Multilayers perceptron

- Arbre de décision
- Forêt aléatoire
- XGBoost

Choix des hyperparamètres par recherche sur grille et validation croisée

Score : R2
Folds : 5

Résultats

Emission de gaz à effet de serre

Résultats

Emission de gaz à effet de serre

Choix du jeu de données

- Effet fort de l'inclusion du mix énergétique (e.g., #0 vs #1)
- Effet plus faible de l'inclusion du ENERGYScore (e.g. #0 vs #2)

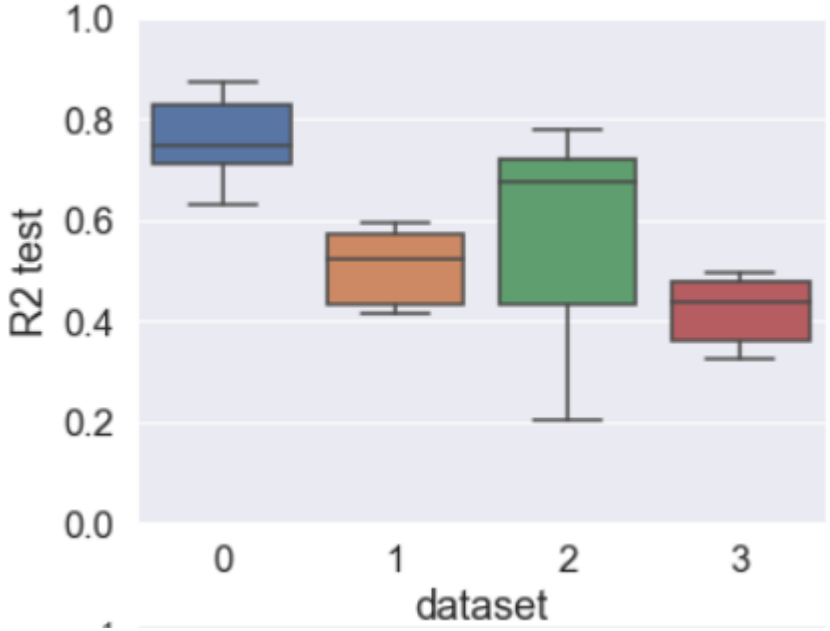
Utilisation du dataset **#0 Energy* + Energy Mix**

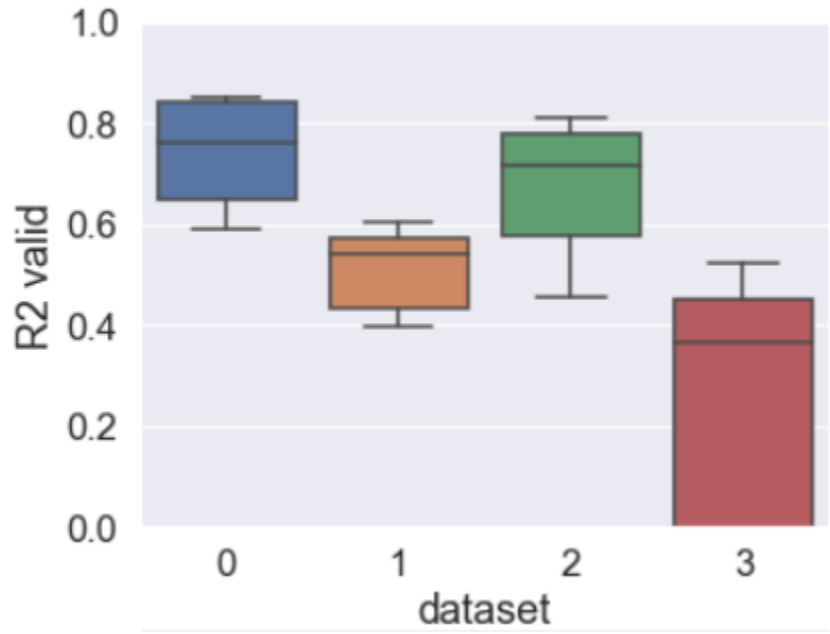
#0 Energy* + Energy Mix

#1 Energy* + ~~Energy Mix~~

#2 ~~Energy*~~ + Energy Mix

#3 ~~Energy*~~ + ~~Energy Mix~~





Résultats

Emission de gaz à effet de serre

Choix du modèle

Classement des modèles par valeur de R2 validation
Les hyperparamètres de chacun ont été optimisés par CV

	dataset	model	mean_fit_time	mean_score_time	train_r2	test_r2_mean	test_r2_sd	val_r2	
5	0	KernelRidge(kernel='rbf')	0.036695	0.008153	0.926829	0.816135	0.053196	0.849283	
4	0	KernelRidge(kernel='poly')	0.035023	0.006128	0.911461	0.845043	0.025633	0.842901	
9	0	MLPRegressor(max_iter=1000)	2.460856	0.003640	0.887993	0.860686	0.017857	0.841252	M
12	0	XGBRegressor(base_score=None, booster=None, ca...	0.394506	0.004424	0.993530	0.874519	0.015360	0.840280	X
11	0	RandomForestRegressor()	2.279515	0.064362	0.965137	0.826503	0.019743	0.831103	R
8	0	SVR(max_iter=20000)	0.010889	0.003217	0.841046	0.745572	0.032935	0.767244	
7	0	SVR(kernel='poly', max_iter=20000)	0.015730	0.000405	0.840193	0.806274	0.028824	0.761129	
10	0	DecisionTreeRegressor()	0.008854	0.000000	0.890697	0.735126	0.036450	0.746681	
3	0	Lasso(max_iter=15000)	0.000000	0.000000	0.736188	0.713459	0.048089	0.648337	L

Résultats

Emission de gaz à effet de serre

Choix du modèle

R2 validations équivalents

Classement des modèles par valeur de R2 validation
Les hyperparamètres de chacun ont été optimisés par CV

	dataset	model	mean_fit_time	mean_score_time	train_r2	test_r2_mean	test_r2_sd	val_r2
5	0	KernelRidge(kernel='rbf')	0.036695	0.008153	0.926829	0.816135	0.05319	0.849283
4	0	KernelRidge(kernel='poly')	0.035023	0.006128	0.911461	0.845043	0.02563	0.842901
9	0	MLPRegressor(max_iter=1000)	2.460856	0.003640	0.887993	0.860686	0.01785	0.841252
12	0	XGBRegressor(base_score=None, booster=None, ca...	0.394506	0.004424	0.993530	0.874519	0.01536	0.840280
11	0	RandomForestRegressor()	2.279515	0.064362	0.965137	0.826503	0.01974	0.831103
8	0	SVR(max_iter=20000)	0.010889	0.003217	0.841046	0.745572	0.032935	0.767244
7	0	SVR(kernel='poly', max_iter=20000)	0.015730	0.000405	0.840193	0.806274	0.028824	0.761129
10	0	DecisionTreeRegressor()	0.008854	0.000000	0.890697	0.735126	0.036450	0.746681
3	0	Lasso(max_iter=15000)	0.000000	0.000000	0.736188	0.713459	0.048089	0.648337

Résultats

Emission de gaz à effet de serre

Classement des modèles par valeur de R2 validation
Les hyperparamètres de chacun ont été optimisés par CV

Choix du modèle

R2 validations équivalents
XGBoost meilleur sur les tests

	dataset	model	mean_fit_time	mean_score_time	train_r2	test_r2_mean	test_r2_sd	val_r2	
5	0	KernelRidge(kernel='rbf')	0.036695	0.008153	0.926829	0.816135	0.053196	0.849283	
4	0	KernelRidge(kernel='poly')	0.035023	0.006128	0.911461	0.845043	0.025633	0.842901	
9	0	MLPRegressor(max_iter=1000)	2.460856	0.003640	0.887993	0.860686	0.017857	0.841252	M
12	0	XGBRegressor(base_score=None, booster=None, ca...	0.394506	0.004424	0.993530	0.874519	0.015360	0.840280	X
11	0	RandomForestRegressor()	2.279515	0.064362	0.965137	0.826503	0.019743	0.831103	R
8	0	SVR(max_iter=20000)	0.010889	0.003217	0.841046	0.745572	0.032935	0.767244	
7	0	SVR(kernel='poly', max_iter=20000)	0.015730	0.000405	0.840193	0.806274	0.028824	0.761129	
10	0	DecisionTreeRegressor()	0.008854	0.000000	0.890697	0.735126	0.036450	0.746681	
3	0	Lasso(max_iter=15000)	0.000000	0.000000	0.736188	0.713459	0.048089	0.648337	L

Résultats

Emission de gaz à effet de serre

Classement des modèles par valeur de R2 validation
Les hyperparamètres de chacun ont été optimisés par CV

Choix du modèle

R2 validations équivalents
XGBoost meilleur sur les tests

Temps calcul (fit plus long pour XGBoost) mais prédictions rapides !

	dataset	model	mean_fit_time	mean_score_time	train_r2	test_r2_mean	test_r2_sd	val_r2
5	0	KernelRidge(kernel='rbf')	0.036695	0.008153	0.926829	0.816135	0.053196	0.849283
4	0	KernelRidge(kernel='poly')	0.035023	0.006128	0.911461	0.845043	0.025633	0.842901
9	0	MLPRegressor(max_iter=1000)	2.460856	0.003640	0.887993	0.860686	0.017857	0.841252
12	0	XGBRegressor(base_score=None, booster=None, ca...	0.394506	0.004424	0.993530	0.874519	0.015360	0.840280
11	0	RandomForestRegressor()	2.279515	0.064362	0.965137	0.826503	0.019743	0.831103
8	0	SVR(max_iter=20000)	0.010889	0.003217	0.841046	0.745572	0.032935	0.767244
7	0	SVR(kernel='poly', max_iter=20000)	0.015730	0.000405	0.840193	0.806274	0.028824	0.761129
10	0	DecisionTreeRegressor()	0.008854	0.000000	0.890697	0.735126	0.036450	0.746681
3	0	Lasso(max_iter=15000)	0.000000	0.000000	0.736188	0.713459	0.048089	0.648337

Résultats

Emission de gaz à effet de serre

Classement des modèles par valeur de R2 validation
Les hyperparamètres de chacun ont été optimisés par CV

Choix du modèle

R2 validations équivalents
XGBoost meilleur sur les tests

Temps calcul (fit plus long pour XGBoost) mais prédictions rapides !

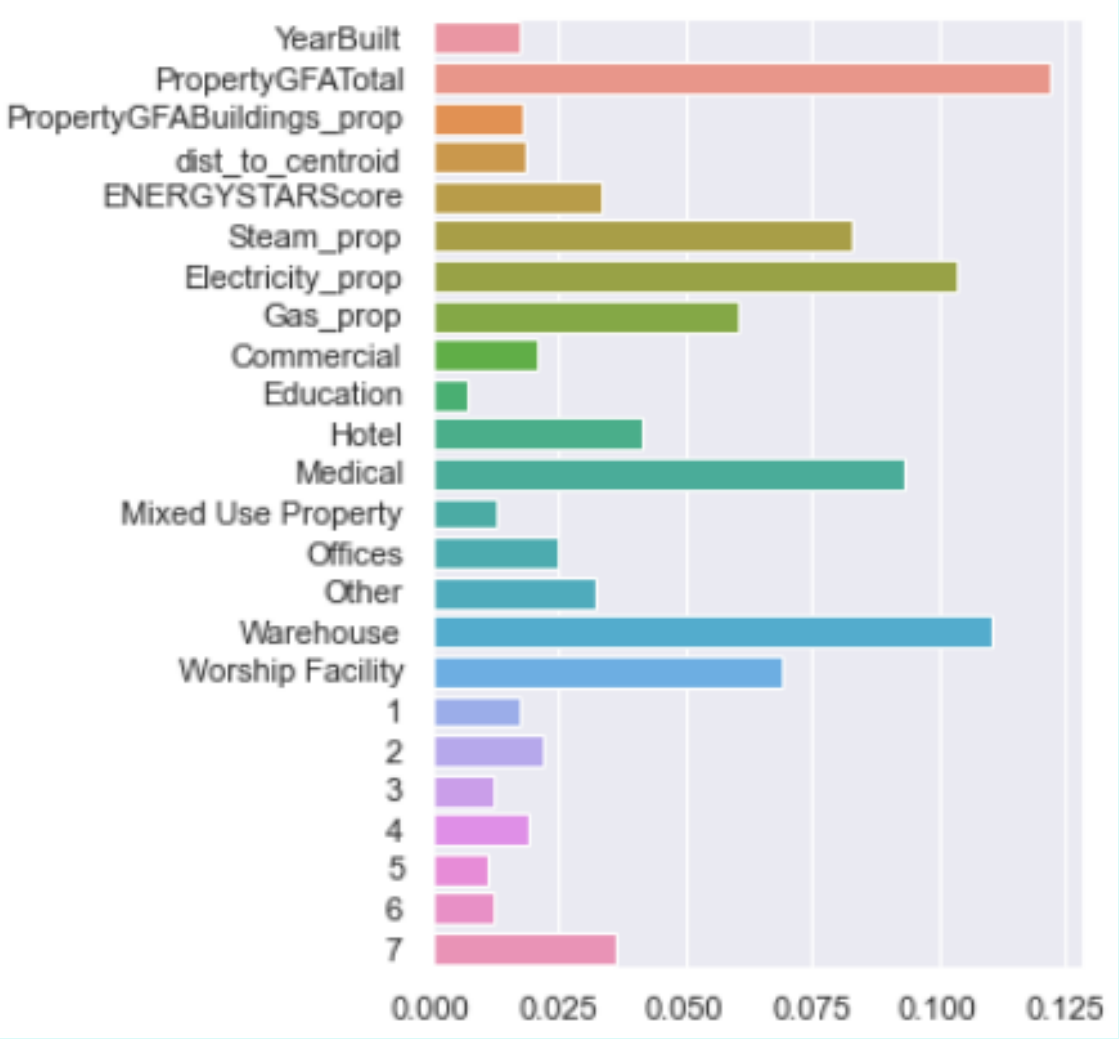
Choix : XGBoost

	dataset	model	mean_fit_time	mean_score_time	train_r2	test_r2_mean	test_r2_sd	val_r2
5	0	KernelRidge(kernel='rbf')	0.036695	0.008153	0.926829	0.816135	0.053196	0.849283
4	0	KernelRidge(kernel='poly')	0.035023	0.006128	0.911461	0.845043	0.025633	0.842901
9	0	MLPRegressor(max_iter=1000)	2.460856	0.003640	0.887993	0.860686	0.017857	0.841252
12	0	XGBRegressor(base_score=None, booster=None, ca...	0.394506	0.004424	0.993530	0.874519	0.015360	0.840280
11	0	RandomForestRegressor()	2.279515	0.064362	0.965137	0.826503	0.019743	0.831103
8	0	SVR(max_iter=20000)	0.010889	0.003217	0.841046	0.745572	0.032935	0.767244
7	0	SVR(kernel='poly', max_iter=20000)	0.015730	0.000405	0.840193	0.806274	0.028824	0.761129
10	0	DecisionTreeRegressor()	0.008854	0.000000	0.890697	0.735126	0.036450	0.746681
3	0	Lasso(max_iter=15000)	0.000000	0.000000	0.736188	0.713459	0.048089	0.648337

Résultats

Emission de gaz à effet de serre

Features importances

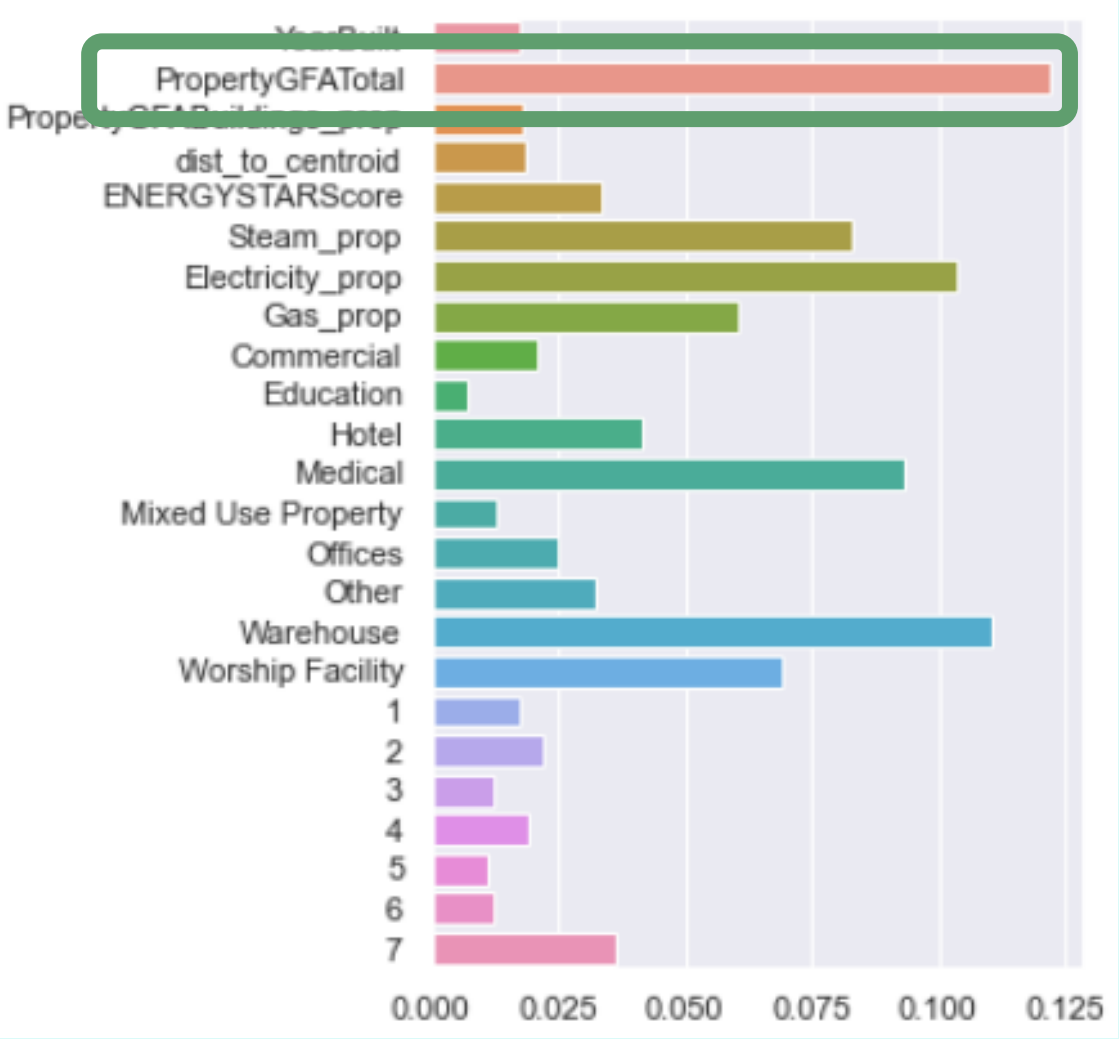


Résultats

Emission de gaz à effet de serre

Surface totale

Features importances



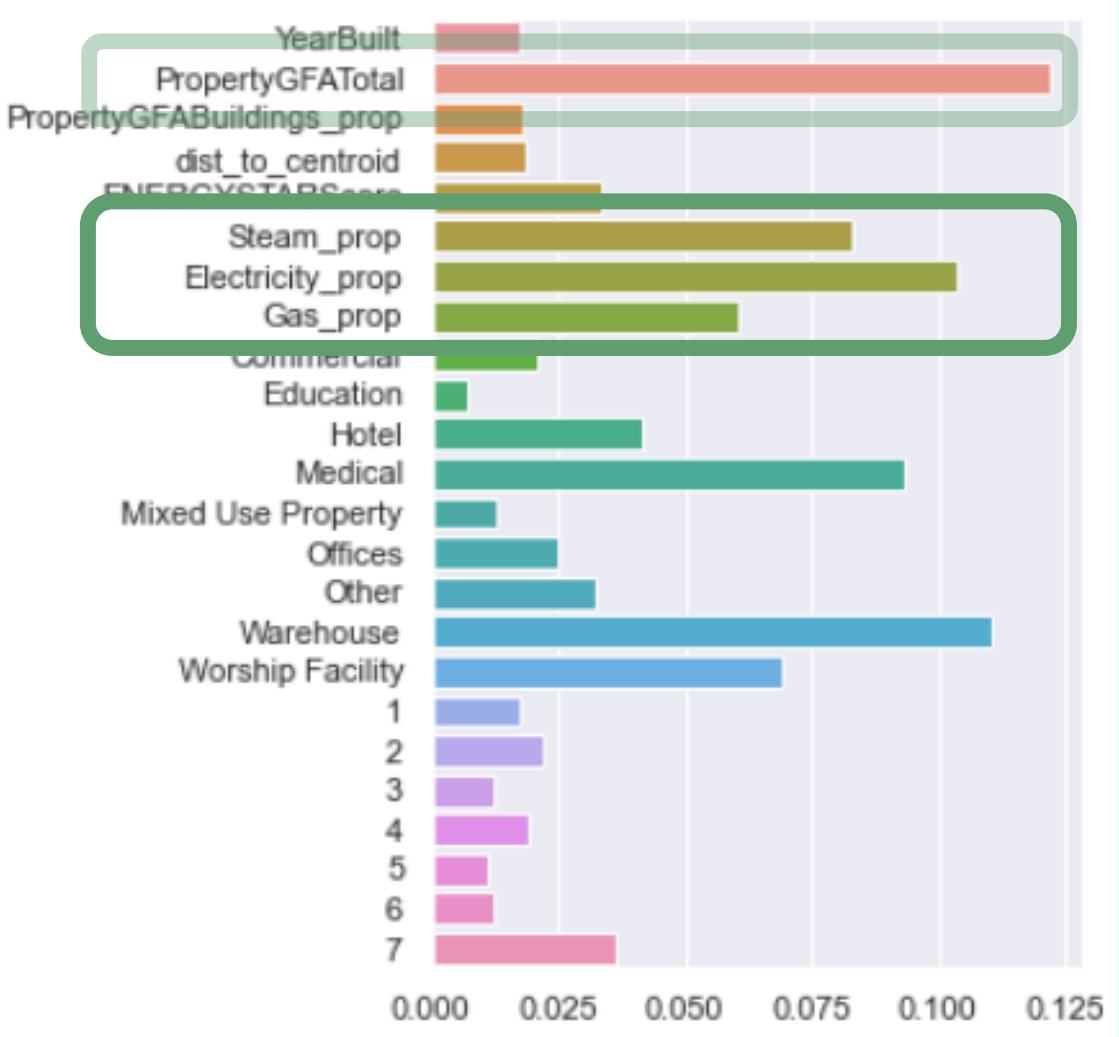
Résultats

Emission de gaz à effet de serre

Surface totale

Mix sources d'énergie

Features importances



Résultats

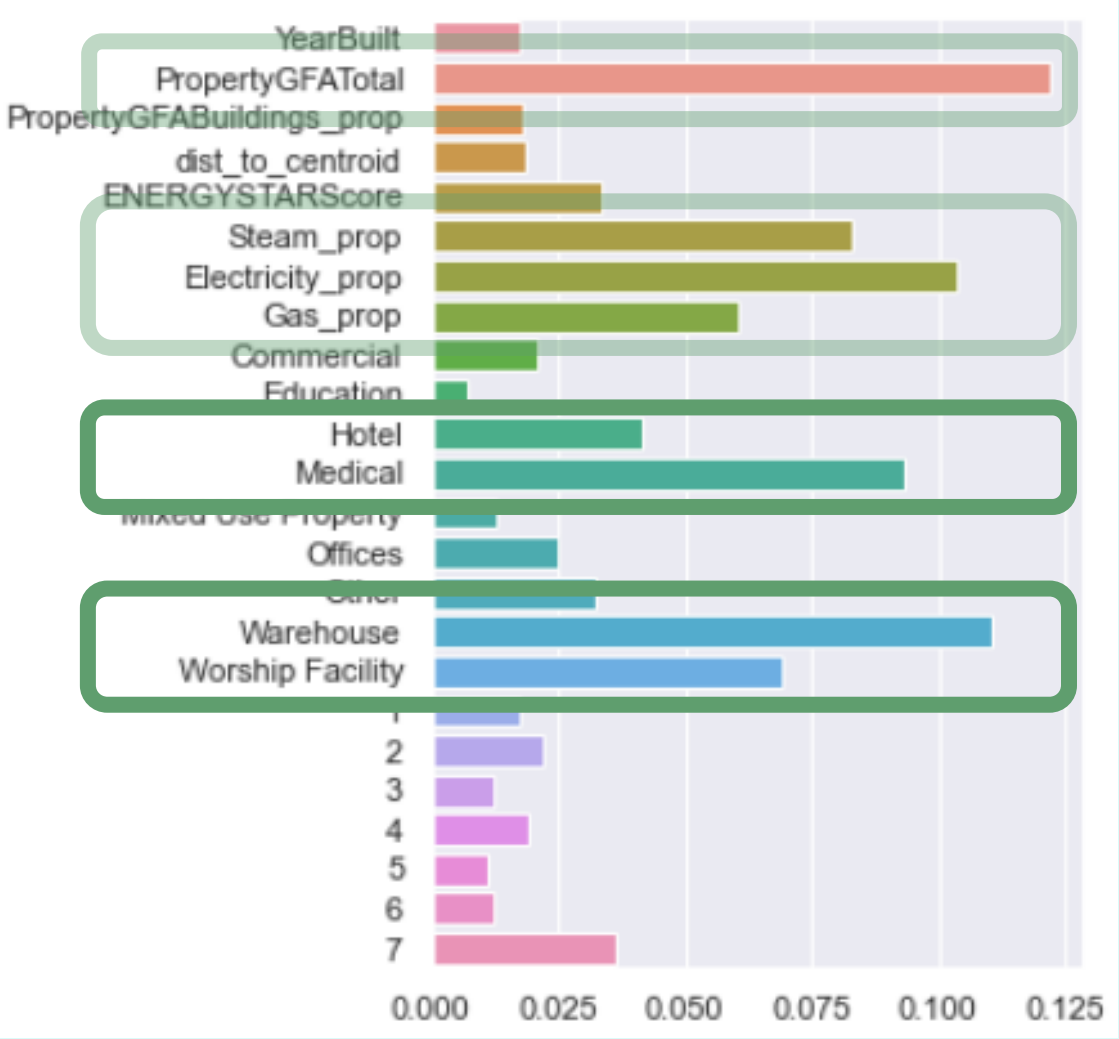
Emission de gaz à effet de serre

Surface totale

Mix sources d'énergie

Certains type de bâtiment

Features importances



Résultats

Emission de gaz à effet de serre

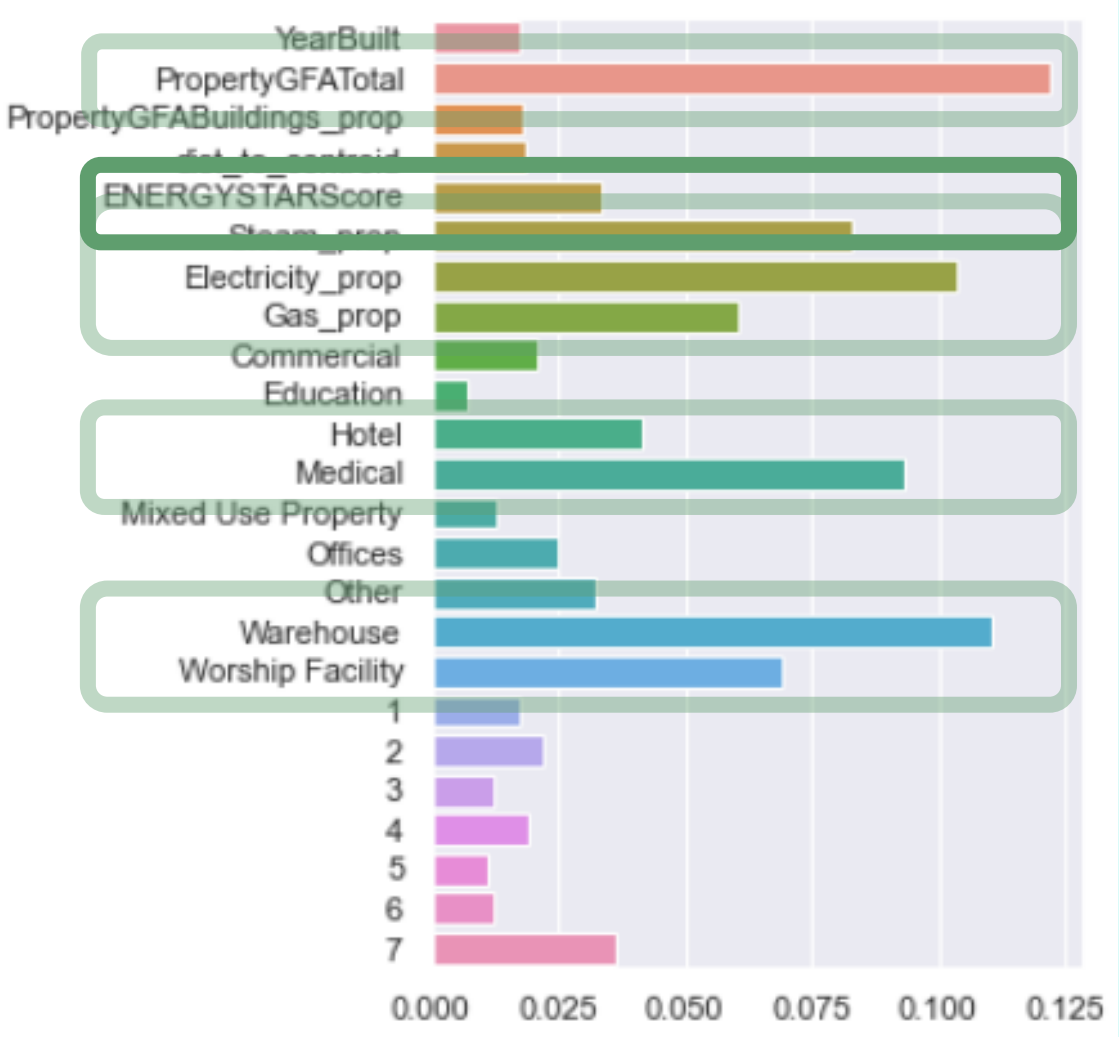
Surface totale

Mix sources d'énergie

Certains type de bâtiment

ENERGYSTARScore : importance marginale

Features importances



Résultats

Emission de gaz à effet de serre

Surface totale

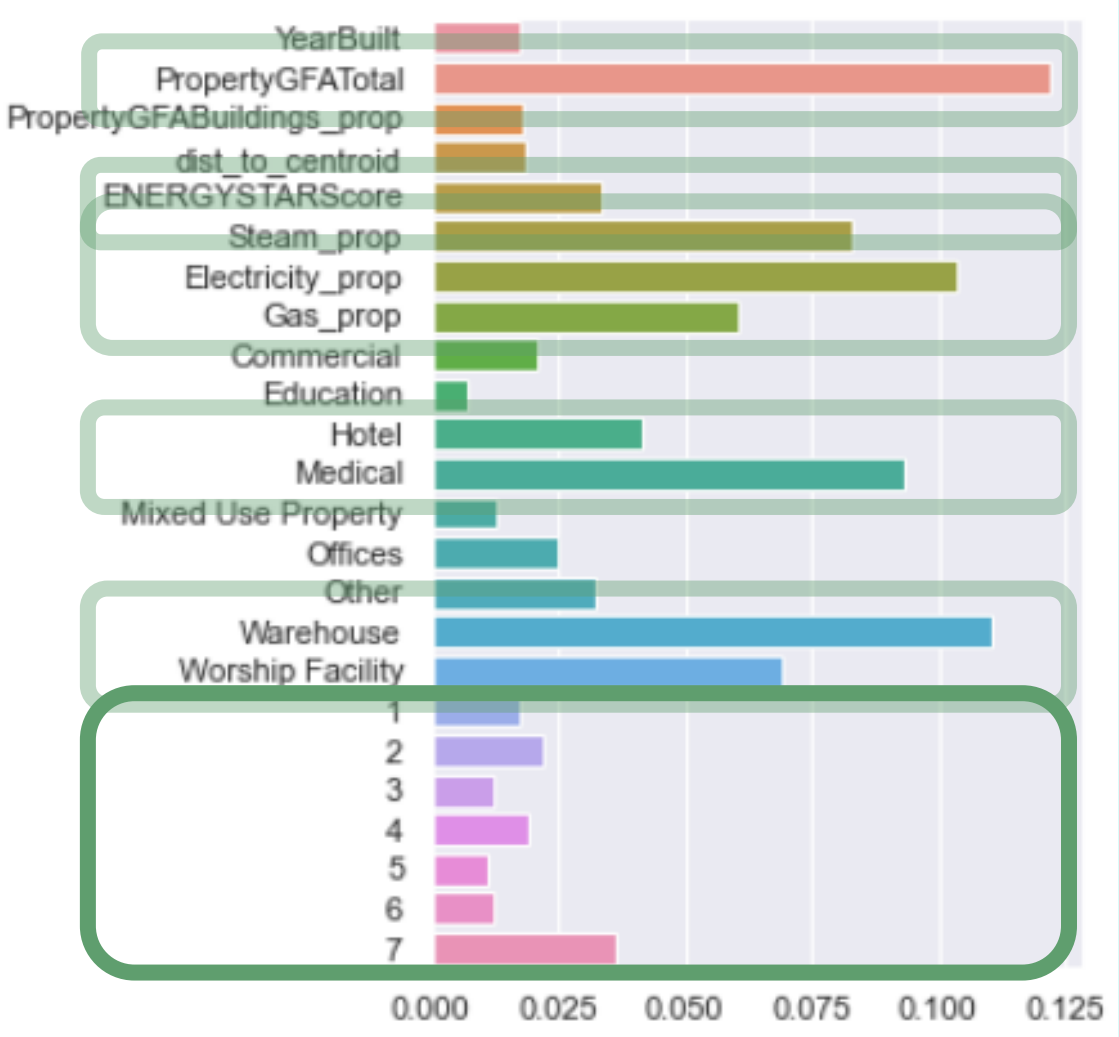
Mix sources d'énergie

Certains type de bâtiment

ENERGYSTARScore : importance marginale

District peu importants

Features importances



Résultats

Emission de gaz à effet de serre

Simplifier le modèle (backward-like selection)

Modèles	R2 validation
Full	0.88
Full – sans district	0.88
Full – sans district, sans Surface Totale	**0.58**
Full – sans district, sans Type de bâtiment	0.80
Full – sans district, sans ENERGYSTARScore	**0.79**
Full – sans district, sans Mix énergétique	**0.53**
Full – sans district, sans Distance au centre	0.88
Full – sans district, sans Année de construction	0.87
Reduced – sans district, ni distance, ni année, ni type de propriété	0.79
Reduced - sans district, ni distance, ni année, ni type de propriété, ni ENERGYSTAR	0.70

Résultats

Emission de gaz à effet de serre

Simplifier le modèle (backward-like selection)

***Features* nécessaires:**

- **Mix énergétique (35pts)**
- **Surface totale (30 pts)**
- **Type de bâtiment (10pts)**
- **ENERGYSTARScore (10pts)**

Modèles	R2 validation
Full	0.88
Full – sans district	0.88
Full – sans district, sans Surface Totale	**0.58**
Full – sans district, sans Type de bâtiment	0.80
Full – sans district, sans ENERGYSTARScore	**0.79**
Full – sans district, sans Mix énergétique	**0.53**
Full – sans district, sans Distance au centre	0.88
Full – sans district, sans Année de construction	0.87
Reduced – sans district, ni distance, ni année, ni type de propriété	0.79
Reduced - sans district, ni distance, ni année, ni type de propriété, ni ENERGYSTAR	0.70

Résultats

Emission de gaz à effet de serre

Simplifier le modèle (backward-like selection)

***Features* nécessaires:**

- **Mix énergétique (35pts)**
- **Surface totale (30 pts)**
- **Type de bâtiment (10pts)**
- **ENERGYSTARScore (10pts)**

A noter :

Pour le mix énergétique, utiliser la proportion d'électricité seule ou la proportion de gaz seule est suffisant !

Modèles	R2 validation
Full	0.88
Full – sans district	0.88
Full – sans district, sans Surface Totale	**0.58**
Full – sans district, sans Type de bâtiment	0.80
Full – sans district, sans ENERGYSTARScore	**0.79**
Full – sans district, sans Mix énergétique	**0.53**
Full – sans district, sans Distance au centre	0.88
Full – sans district, sans Année de construction	0.87
Reduced – sans district, ni distance, ni année, ni type de propriété	0.79
Reduced - sans district, ni distance, ni année, ni type de propriété, ni ENERGYSTAR	0.70

Résultats

Consommation d’énergie

Résultats

Consommation d'énergie

Choix du jeu de données

- Pas d'effet de l'inclusion du mix énergétique (e.g., #0 vs #1)
- Effet fort de l'inclusion du ENERGYScore (e.g. #0 vs #2)

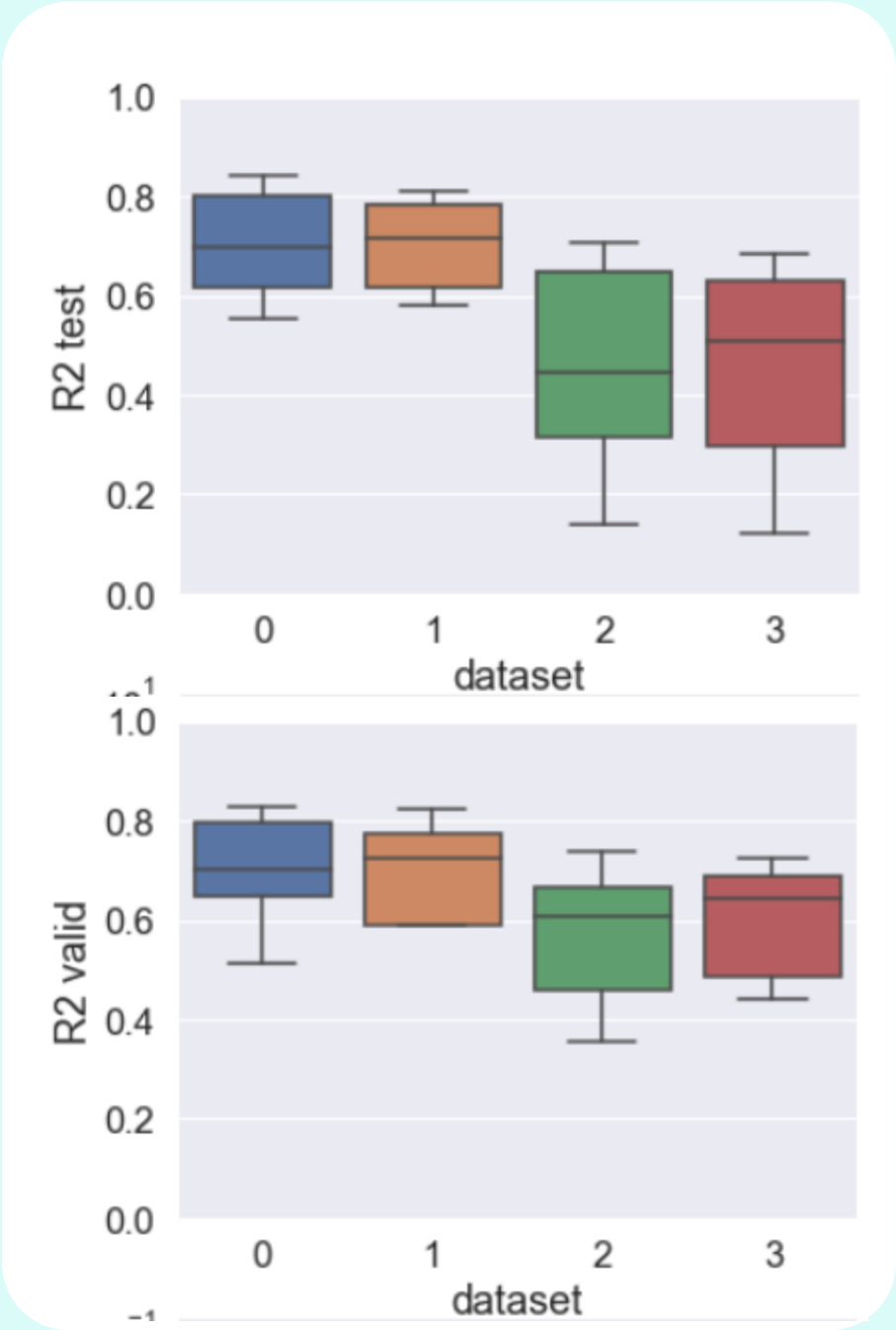
Utilisation du dataset **#1 Energy***

#0 Energy* + Energy Mix

#1 Energy* + ~~Energy Mix~~

#2 ~~Energy*~~ + Energy Mix

#3 ~~Energy*~~ + ~~Energy Mix~~



Résultats

Consommation d’énergie

Choix du modèle

Classement des modèles par valeur de R2 validation
Les hyperparamètres de chacun ont été optimisés par CV

	dataset	model	mean_fit_time	mean_score_time	train_r2	test_r2_mean	test_r2_sd	val_r2
25	1	XGBRegressor(base_score=None, booster=None, ca...	0.477278	0.003136	0.994013	0.806330	0.017770	0.821363
22	1	MLPRegressor(max_iter=1000)	6.600118	0.004070	0.847264	0.809436	0.020163	0.805837
24	1	RandomForestRegressor()	1.189705	0.041086	0.951321	0.786644	0.025701	0.781812
18	1	KernelRidge(kernel='rbf')	0.036284	0.008555	0.879750	0.754071	0.064978	0.773458
17	1	KernelRidge(kernel='poly')	0.039095	0.005666	0.862931	0.781739	0.020652	0.773453
23	1	DecisionTreeRegressor()	0.003126	0.003124	0.845783	0.715934	0.023787	0.769229

Résultats

Consommation d’énergie

Choix du modèle

Choix : XGBoost

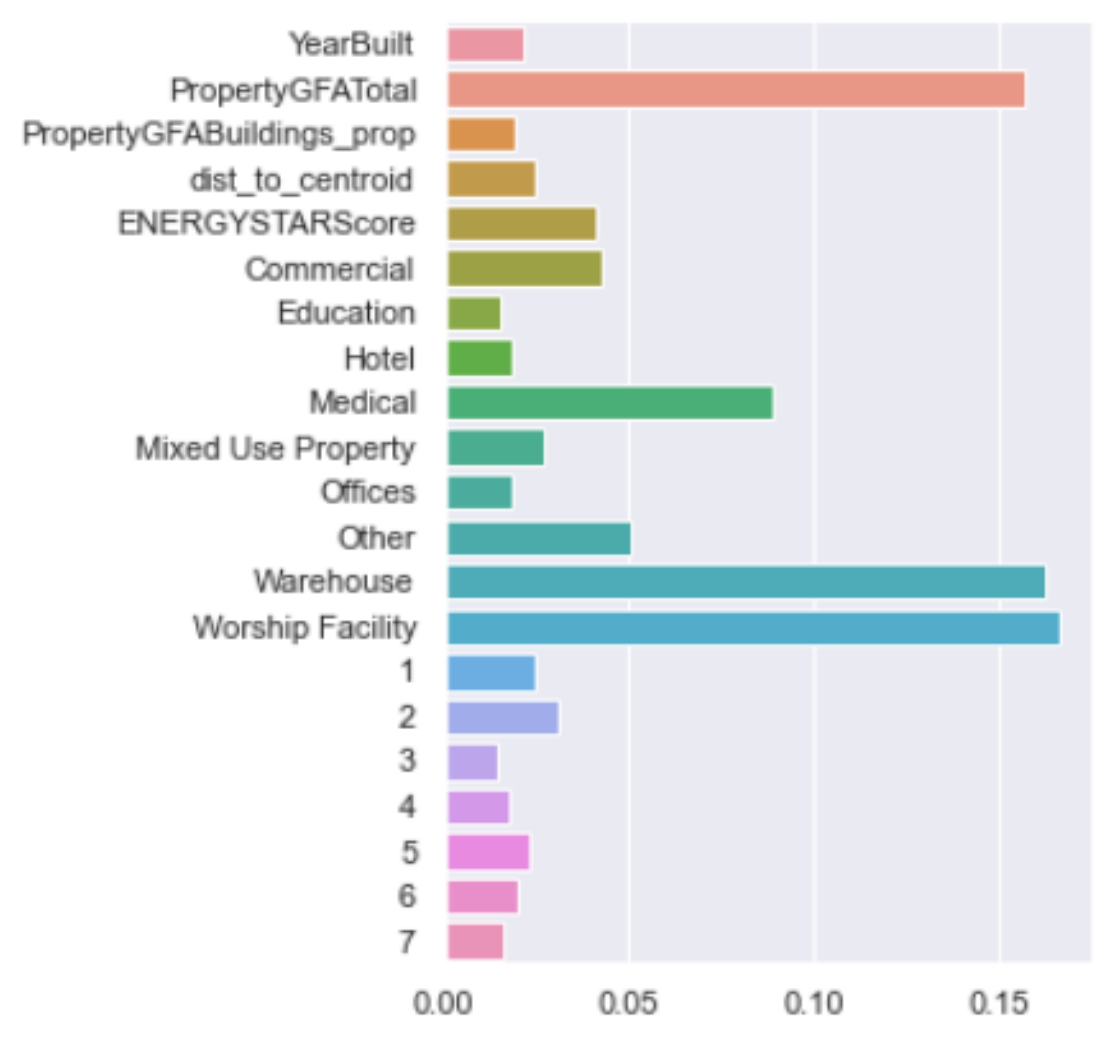
Classement des modèles par valeur de R2 validation
Les hyperparamètres de chacun ont été optimisés par CV

	dataset	model	mean_fit_time	mean_score_time	train_r2	test_r2_mean	test_r2_sd	val_r2
25	1	XGBRegressor(base_score=None, booster=None, ca...	0.477278	0.003136	0.994013	0.806330	0.017770	0.821363
22	1	MLPRegressor(max_iter=1000)	6.600118	0.004070	0.847264	0.809436	0.020163	0.805837
24	1	RandomForestRegressor()	1.189705	0.041086	0.951321	0.786644	0.025701	0.781812
18	1	KernelRidge(kernel='rbf')	0.036284	0.008555	0.879750	0.754071	0.064978	0.773458
17	1	KernelRidge(kernel='poly')	0.039095	0.005666	0.862931	0.781739	0.020652	0.773453
23	1	DecisionTreeRegressor()	0.003126	0.003124	0.845783	0.715934	0.023787	0.769229

Résultats

Consommation d’énergie

Features importances

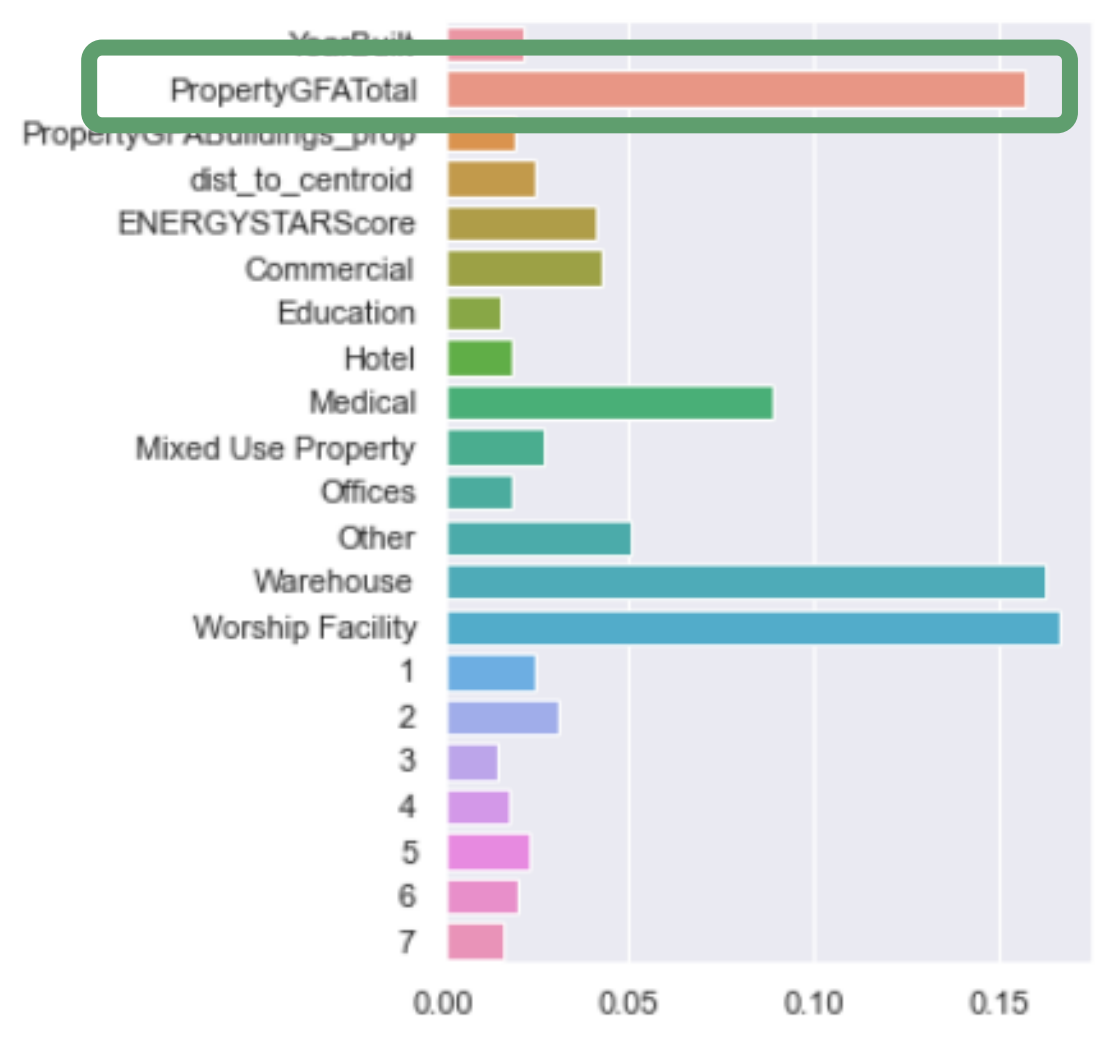


Résultats

Consommation d’énergie

Surface totale

Features importances



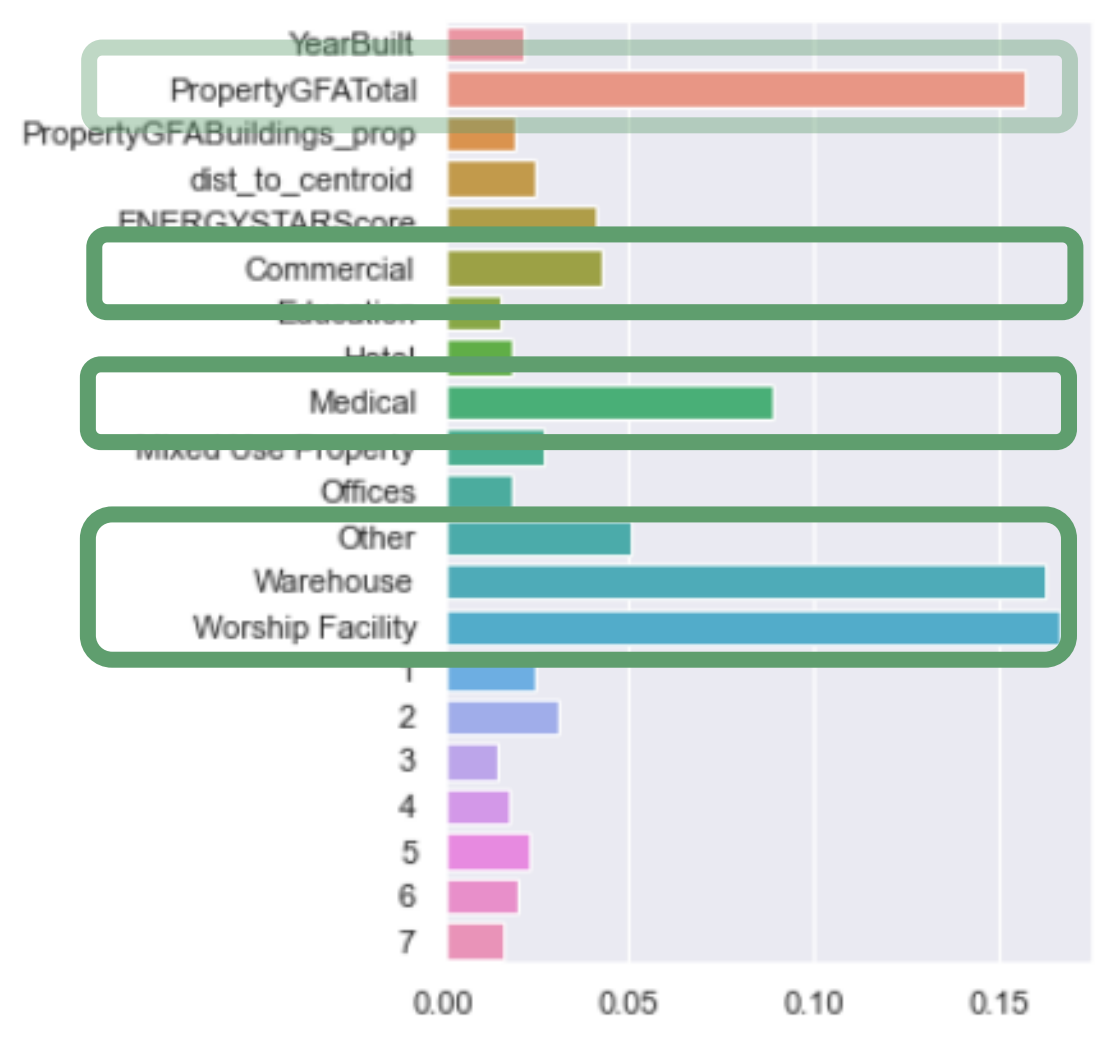
Résultats

Consommation d’énergie

Surface totale

Type de bâtiment

Features importances



Résultats

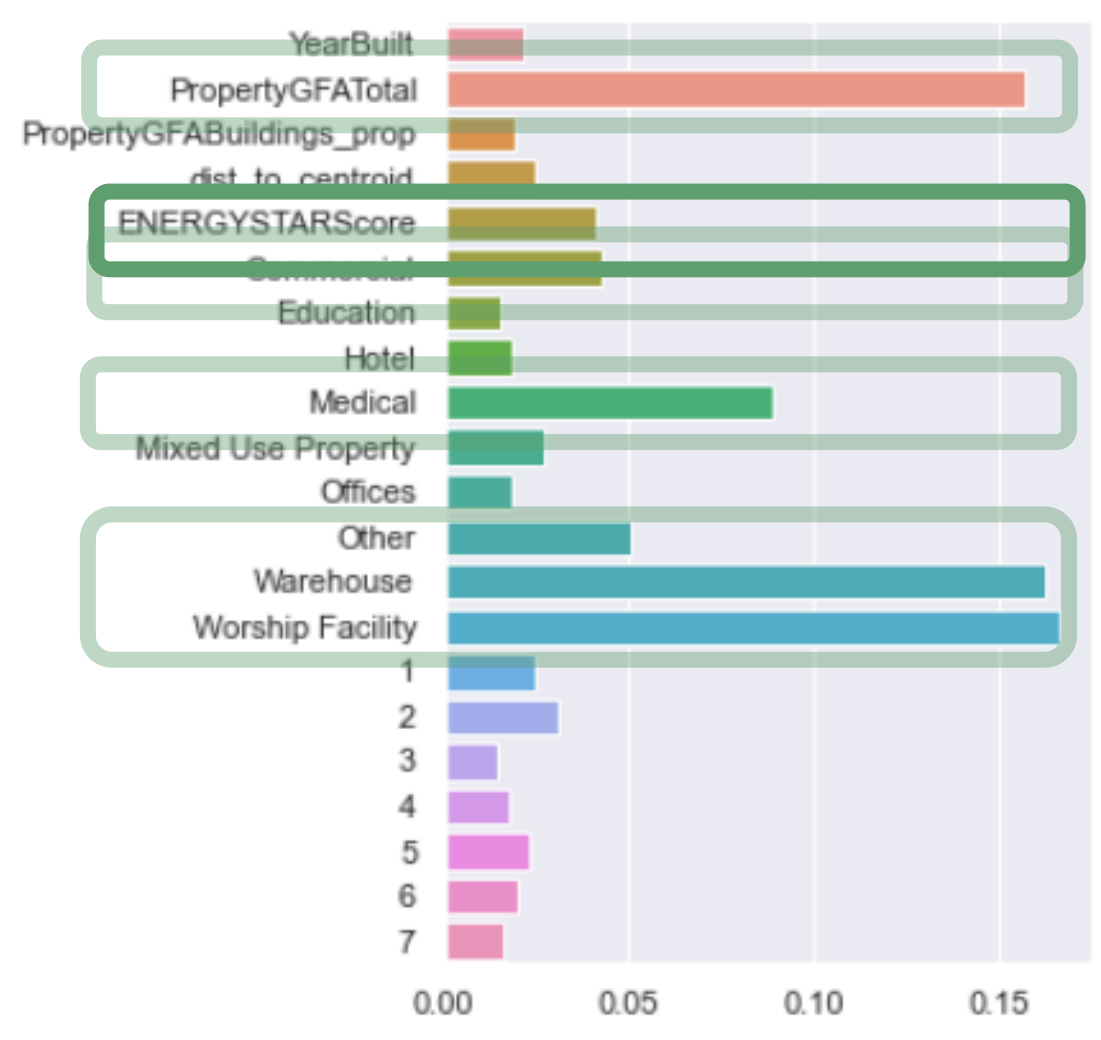
Consommation d’énergie

Surface totale

Type de bâtiment

ENERGYSTARScore ?

Features importances



Résultats

Consommation d'énergie

Simplifier le modèle (backward-like selection)

Modèles	R2 validation
Full	0.81
Full – sans district	0.79
Full – sans district, sans Surface Totale	**0.33**
Full – sans district, sans Type de bâtiment	0.72
Full – sans district, sans ENERGYSTARScore	**0.67**
Full – sans district, sans Distance au centre	0.81
Full – sans district, sans Année de construction	0.79
Reduced – sans district, ni distance, ni année, ni type de propriété	0.69
Reduced - sans district, ni distance, ni année, ni type de propriété, ni ENERGYSTAR	0.61

Résultats

Consommation d'énergie

Simplifier le modèle (backward-like selection)

***Features* nécessaires:**

- **Surface totale (46 pts)**
- **ENERGYSTARScore (12 pts)**
- **~Type de bâtiment (7 pts)**

Modèles	R2 validation
Full	0.81
Full – sans district	0.79
Full – sans district, sans Surface Totale	**0.33**
Full – sans district, sans Type de bâtiment	0.72
Full – sans district, sans ENERGYSTARScore	**0.67**
Full – sans district, sans Distance au centre	0.81
Full – sans district, sans Année de construction	0.79
Reduced – sans district, ni distance, ni année, ni type de propriété	0.69
Reduced - sans district, ni distance, ni année, ni type de propriété, ni ENERGYSTAR	0.61

Conclusion

Prédire la consommation et les émissions des bâtiments non destinés à l'habitation

Possible pour les deux variables (Energie et Gaz à effet de serre) – avec R2 respectifs de env. 0.8 et 0.9.
Variables nécessaires à ces prédictions :

GHG
- Mix énergétique (35pts)
- Surface totale (30 pts)
- Type de bâtiment (10pts)
- ENERGYSTARScore (10pts)

Energie
- Surface totale (46 pts)
- ENERGYSTARScore (12 pts)
- Type de bâtiment (7 pts)

Conclusion

Prédire la consommation et les émissions des bâtiments non destinés à l'habitation

Possible pour les deux variables (Energie et Gaz à effet de serre) – avec R2 respectifs de env. 0.8 et 0.9.
Variables nécessaires à ces prédictions :

GHG
- Mix énergétique (35pts)
- Surface totale (30 pts)
- Type de bâtiment (10pts)
- ENERGYSTARScore (10pts)

Energie
- Surface totale (46 pts)
- ENERGYSTARScore (12 pts)
- Type de bâtiment (7 pts)

Evaluation de l'intérêt de l'ENERGIE STAR Score pour prédire ces consommations/émissions

Ce score améliore les prédictions de env. 10 points dans les deux cas.

Merci