

Anticipez les besoins en consommation de bâtiments



Seattle

E. Campos / Juillet 2023



Contexte, problématique & objectif

Seattle, ville neutre en émissions de carbone en 2050

Contexte

Afin d'atteindre l'objectif de neutralité de la ville, deux variables relatives aux **bâtiments non destinés à l'habitation** font l'objet d'une attention particulière :

- ✓ la **consommation d'énergie**, et
- ✓ les **émissions de CO₂**

Problématique

- ★ Des relevés minutieux ont été effectués par les agents de la ville en 2016, mais ces derniers sont **coûteux**.
- ★ L'Energy Star Score (symbole administré par l'Agence Américaine pour la protection de l'environnement), un indicateur qui fournit un aperçu de la performance énergétique des bâtiments, est **fastidieux** à calculer.



Objectif

- A partir des relevés déjà réalisés et des données structurelles des bâtiments non destinés à l'habitation (surface, activité, type d'énergie consommée...), **prédir les émissions de CO₂ et la consommation totale d'énergie de bâtiments pour lesquels elles n'ont pas encore été mesurées**
 - ➡ Objectif : pouvoir se passer des relevés de consommation dans le futur.
- L'intérêt de l'ENERGY STAR Score pour la prédition d'émissions sera évaluée

Travaux effectués

- Préparation des données (EDA & feature engineering)
- Comparaison des résultats obtenus par différents algorithmes de Machine Learning
- Analyse du modèle retenu

Présentation du jeu de données

- Nos travaux se sont basés sur le « Building Energy Benchmarking » de la ville de Seattle réalisée en 2016
- Ce jeu de données est composé de 3376 bâtiments et 46 attributs. Près de 13% de ses données sont manquantes.

Chaque bâtiment est identifiable par un ID propre, ainsi qu'un certain nombre d'attributs:

- ◆ **Informations générales d'identification** : nom du bâtiment, adresse postale, numéro d'identification fiscal
- ◆ **Informations relative à la localisation** : quartier, code postal, longitude, latitude
- ◆ **Caractéristiques techniques** : type de bâtiment, année de construction, nombre de bâtiments, nombre d'étages, surface occupée...
- ◆ **Caractéristiques liées à l'usage** : liste des différentes activités et des surfaces respectives occupées
- ◆ **Relevés de mesures** : la quantité annuelle totale d'énergie consommée, la quantité annuelle de gaz consommée, la quantité annuelle d'électricité consommée, la quantité annuelle de vapeur consommée, la quantité annuelle de gaz à effet de serre émis, la quantité d'énergie consommé au m², la quantité annuelle de gaz à effet de serre émis au m²
- ◆ **La note Energy Star Score ainsi que la date de certification** (tous les bâtiments n'ont pas de certification)
- ◆ **Données supplémentaires** : commentaires, status de conformité, statut d'outlier, usage de donnée par défaut

- Les deux variables 2 à prédire (« target »):

La consommation annuelle d'énergie consommée par la propriété, quelque soit sa source
« SiteEnergyUse(kBtu) »

Les émissions de gaz à effet de serre du fait de l'énergie consommée par la propriété
« TotalGHGEmissions »

Exploratory Data Analysis

Retraitemet des données

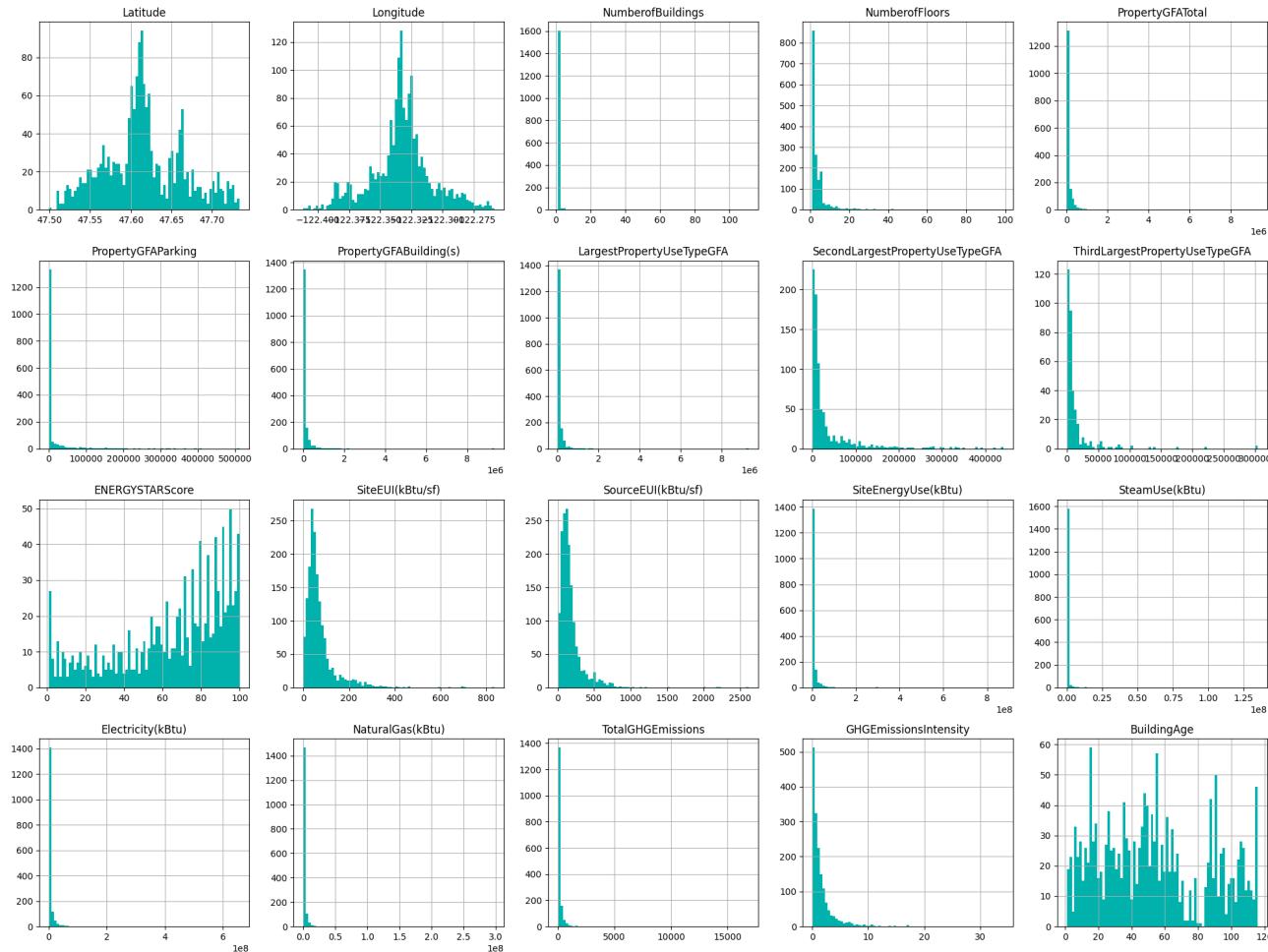
- Suppression des individus **hors scope** (sont conservés uniquement les bâtiments « Non residential ») - 1711 entrées
- Suppression des **outliers** (indiqués comme tels) - 16 entrées
- Suppression des variables « **inutiles** » (ville, état, année de certification, DefaultsData, ComplianceStatus, WN) ou vides (Comments)
- Suppression des variables **redondantes** : on conserve la mesure de la consommation d'énergie en électricité et gaz naturel en kBtu (suppression de kWh et therms)
- Rationalisation des variables (nom de quartier écrit de manière différente)
- Correction de cohérence : la surface brute totale doit être supérieure ou égale à la surface occupée par les différentes activités
- Correction des valeurs nulles : le nombre de bâtiment et d'étage minimal est fixé à 1
- Suppression des entrées avec des valeurs négatives aberrantes - 7 entrées
- Suppression des entrées susceptibles de fausser les calculs
 - Bâtiments avec consommation nulle sans détail du type de d'énergie - 5 entrées
 - Bâtiments avec une émission de CO2 nulle, mais une consommation d'énergie non nulle - 1 entrée
- Recalcule des valeurs nulles :
 - Cas où l'énergie total consommée est nulle alors qu'il existe des relevés de mesure pour différent type d'énergie ; nous avons appliquer la somme
- Transformation d'une variable : age du bâtiment au moment de la mesure plutôt que l'année de construction

Gestion des données manquantes

- Suppression des entrées avec des mesures manquantes - 2 entrées
- Cas des données « ZipCode » manquantes
 - ✓ Complété sur la base du zip code le plus fréquent de bâtiments présent dans le même Council District
- Cas des données « LargestPropertyUseType » manquantes
 - ✓ Complété avec la première activité citée dans la variable « ListOfAllPropertyUseTypes »
 - ✓ La surface totale de la propriété est imputée à la surface occupée par cette activité dans la mesure où les entrées concernées n'avait qu'une seule activité
- Suite à ces imputations, les données manquantes concernent uniquement
 - les entrées relatives aux Energy Star Score (bâtiments non certifiés),
 - les entrées relatives aux activités secondaires et tertiaires (bâtiment qui ont une unique activité)

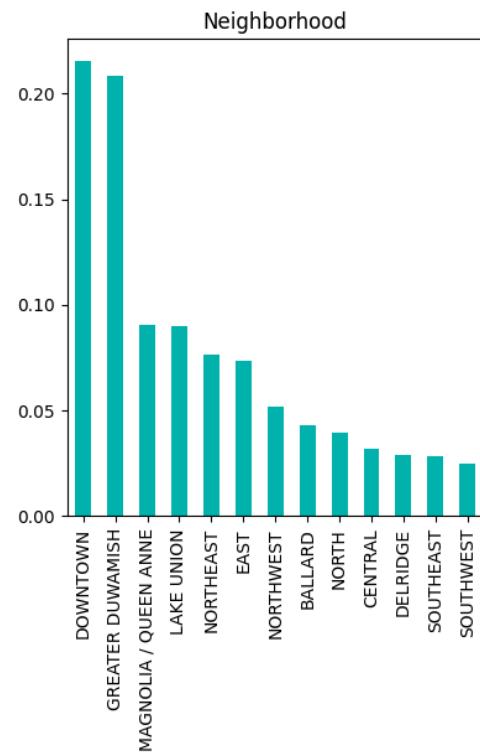
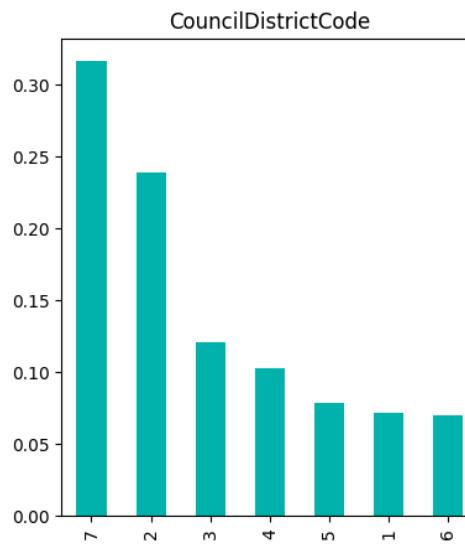
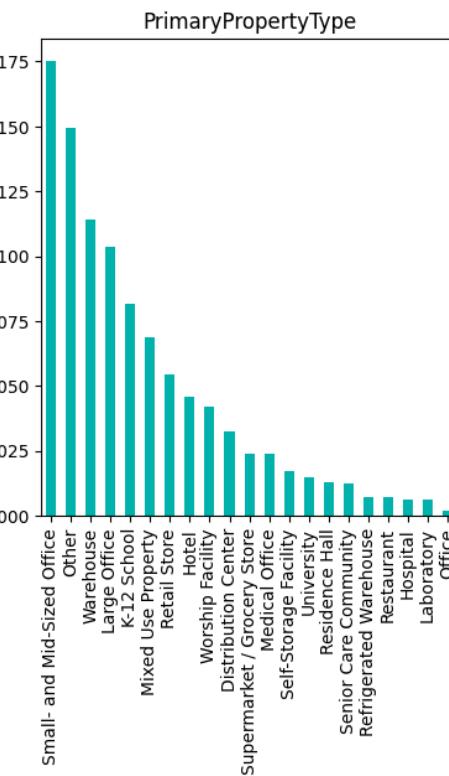
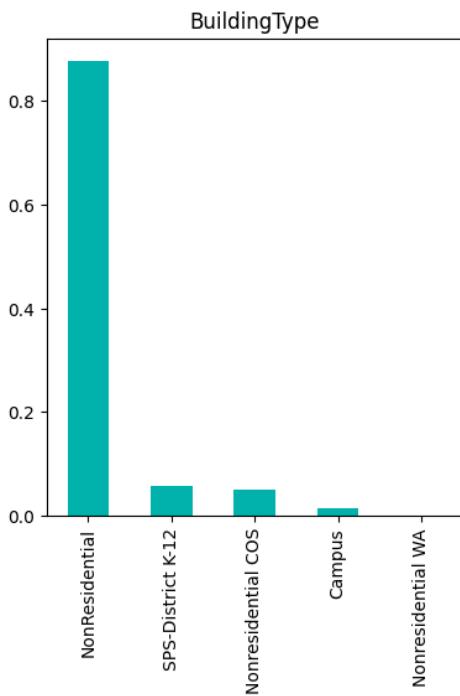
Exploratory Data Analysis

Analyse numérique univariée



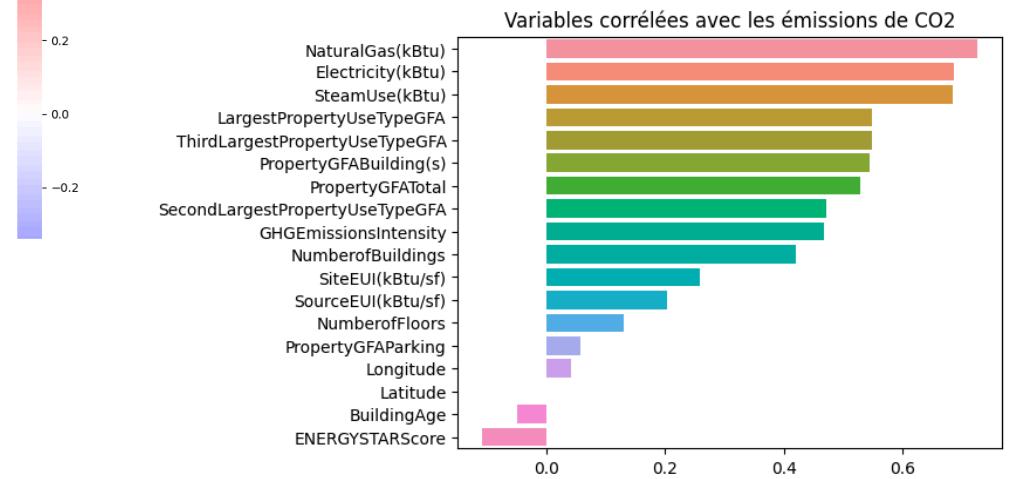
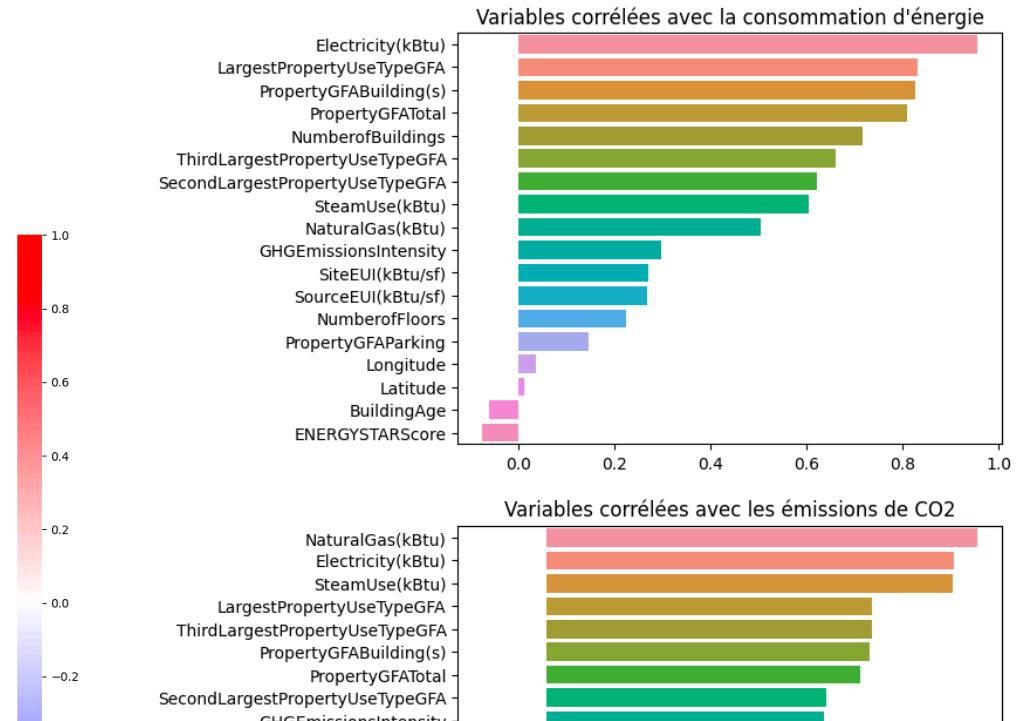
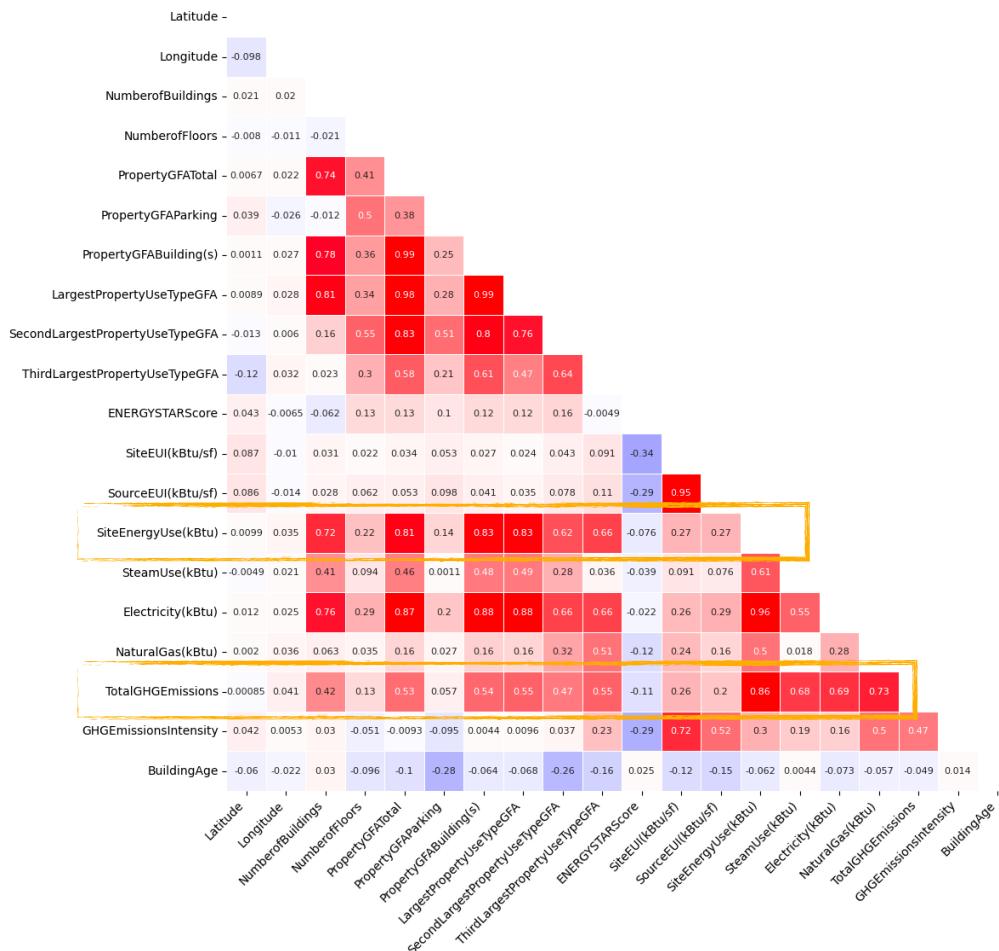
Exploratory Data Analysis

Analyse catégorielle univariée



Exploratory Data Analysis

Analyse multivariée



Feature selection & engineering (1/2)

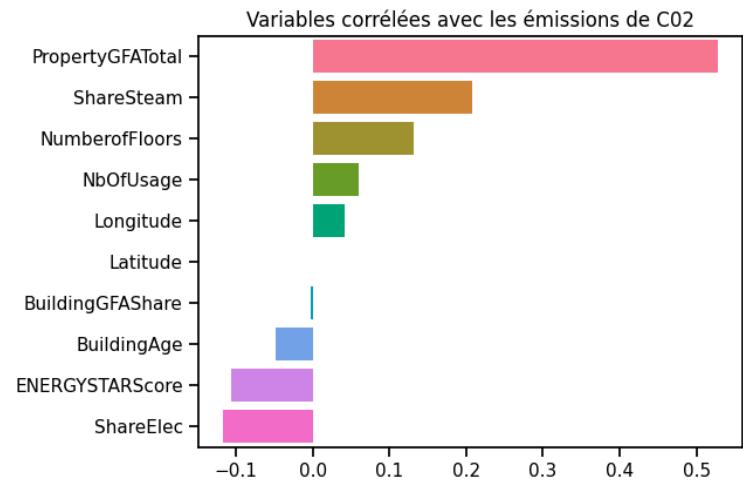
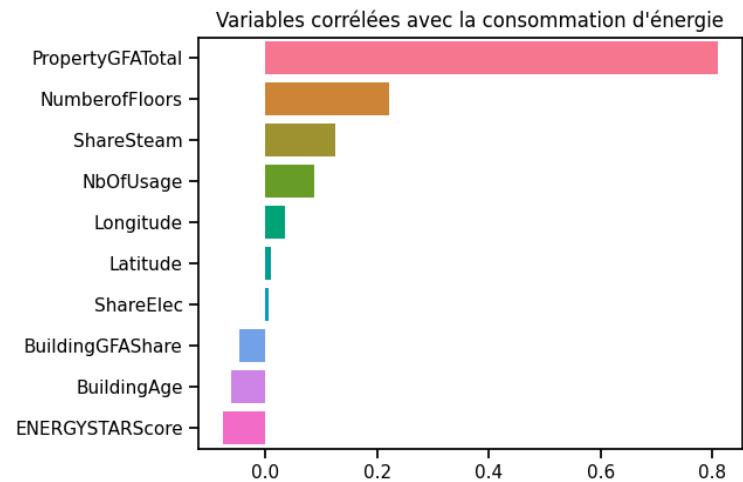
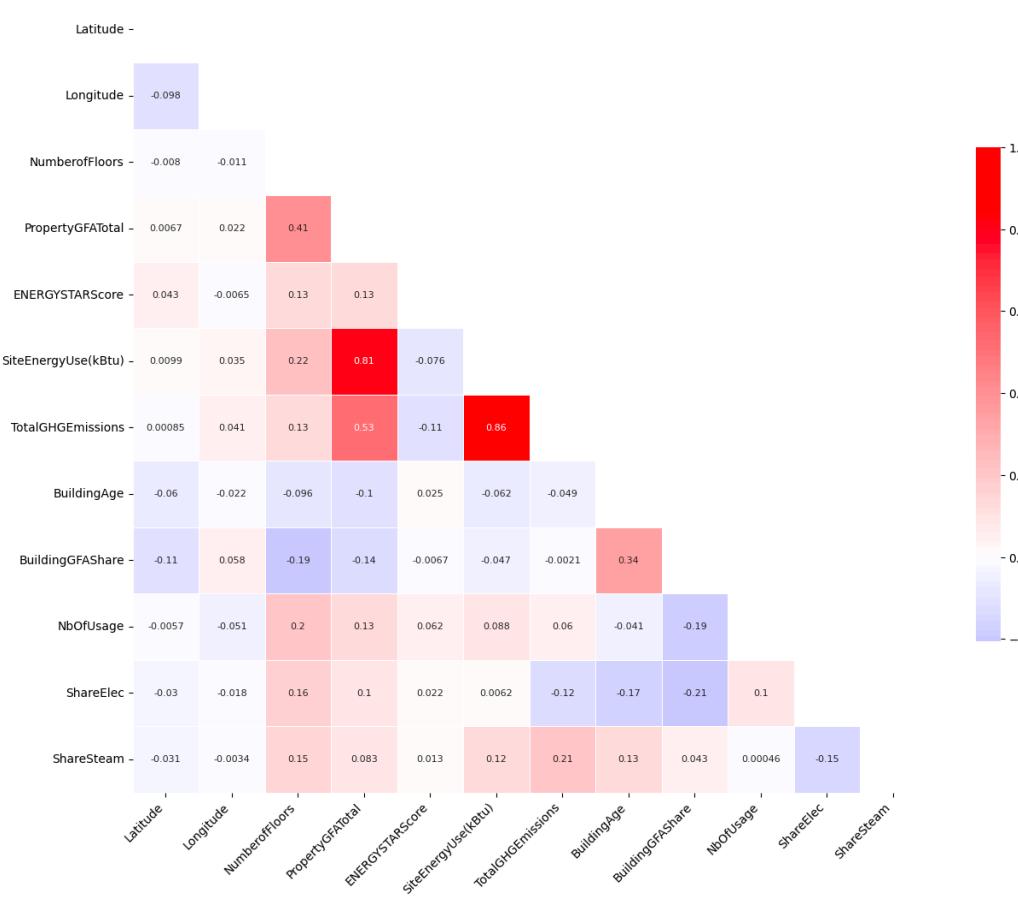
- Afin d'améliorer la performance des algorithme de machine learning, nous allons supprimer les variables trop corrélées entre elles (information redondante):
 - Limiter le nombre de variables participe à améliorer les performances des modèles de ML (limite l'overfitting et les temps d'entraînement)
 - La multicolinéarité peut engendrer des solutions numériquement instable et non interprétable

PropertyGFATotal	PropertyGFABuilding(s)	0.990361
PropertyGFABuilding(s)	LargestPropertyUseTypeGFA	0.985486
PropertyGFATotal	LargestPropertyUseTypeGFA	0.981180
SiteEnergyUse(kBtu)	Electricity(kBtu)	0.956358
SiteEUI(kBtu/sf)	SourceEUI(kBtu/sf)	0.949626
LargestPropertyUseTypeGFA	Electricity(kBtu)	0.882557
PropertyGFABuilding(s)	Electricity(kBtu)	0.878956
PropertyGFATotal	Electricity(kBtu)	0.868159
SiteEnergyUse(kBtu)	TotalGHGEmissions	0.860211
LargestPropertyUseTypeGFA	SiteEnergyUse(kBtu)	0.832423
PropertyGFABuilding(s)	SiteEnergyUse(kBtu)	0.825914
PropertyGFATotal	SecondLargestPropertyUseTypeGFA	0.825794
	SiteEnergyUse(kBtu)	0.809647
NumberofBuildings	LargestPropertyUseTypeGFA	0.805369
PropertyGFABuilding(s)	SecondLargestPropertyUseTypeGFA	0.803908

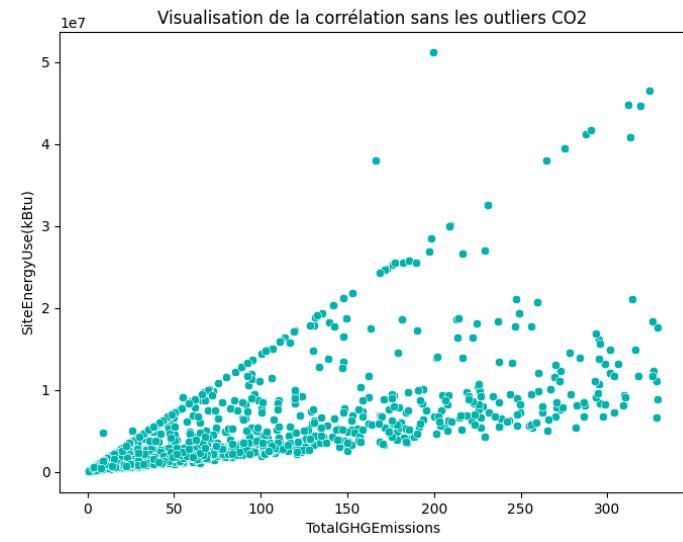
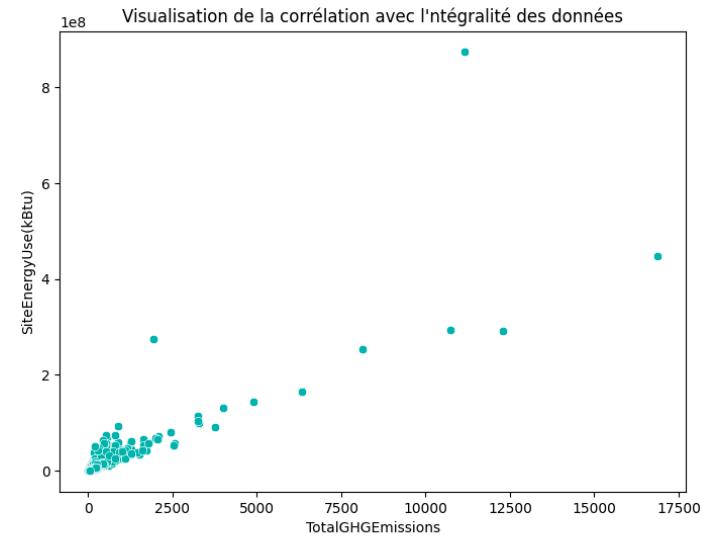
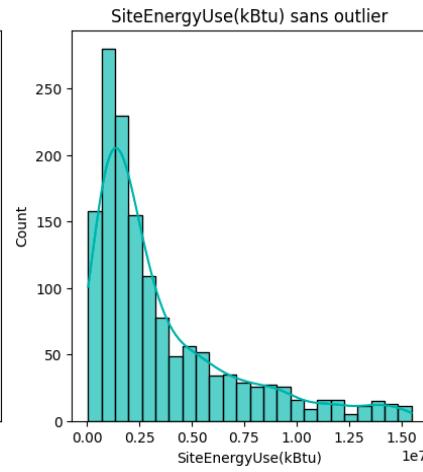
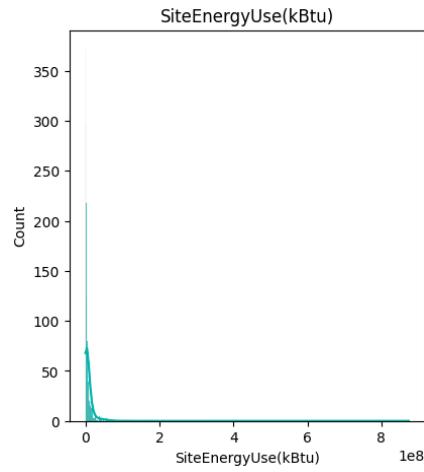
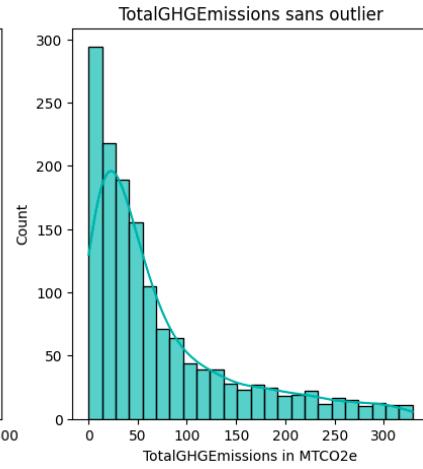
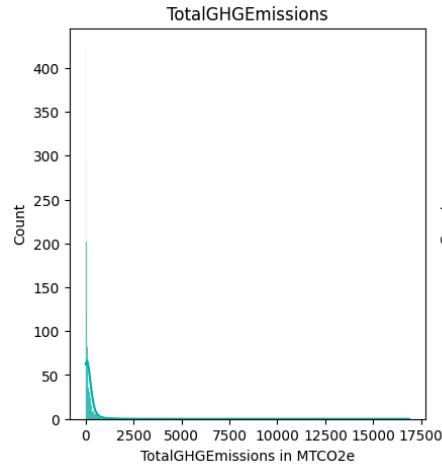
- La surface occupée par le bâtiment sur la propriété est remplacée par la part (en %) occupée par le bâtiment dans la surface totale : « **BuildingGFAShare** »
 - * *La variable relative à la surface occupée par le parking est supprimée : elle peut être déduite de la variable relative au bâtiment*
- Suppression de la variable relative aux nombres de bâtiment qui est très corrélée à la surface totale de la propriété sans apporter d'information supplémentaire
- La liste des activités est remplacée par le nombre d'activités : « **NbOfUsage** »
- Suppression des informations relatives aux activités secondaires et tertiaires (activité et surface)
- Suppression des mesures en kBtu/sf qui correspondent aux mesures par étage
- Suppression des données « intensity » qui correspondent aux consommations d'énergie et émission par m²
- Le détail des mesures de consommation d'électricité, de gaz et de vapeur est supprimé et remplacé par la part d'électricité et la part de vapeurs dans la consommation d'énergie totale : « **ShareElec** » et « **ShareSteam** »
 - * *La variable relative à la part du gaz n'est pas créée car elle peut être déduite des deux autres*

SiteEnergyUse(kBtu)	TotalGHGEmissions	0.860211
PropertyGFATotal	SiteEnergyUse(kBtu)	0.809647
	TotalGHGEmissions	0.527824
NumberofFloors	PropertyGFATotal	0.411103
BuildingAge	BuildingGFAShare	0.339437

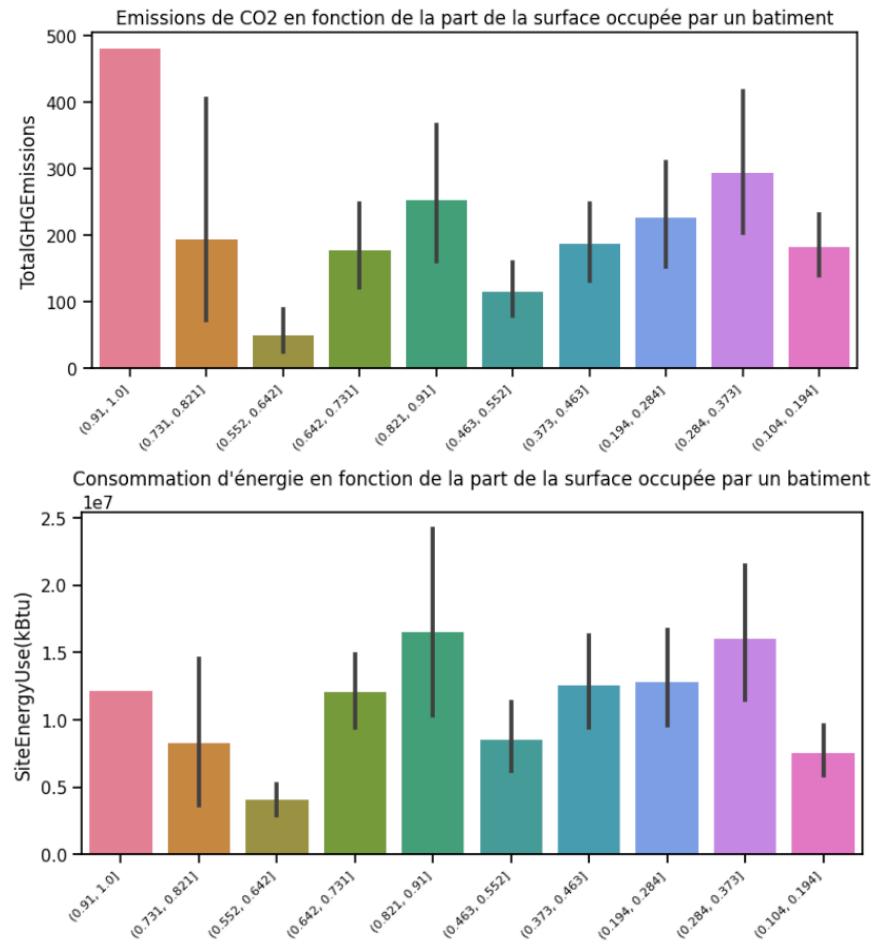
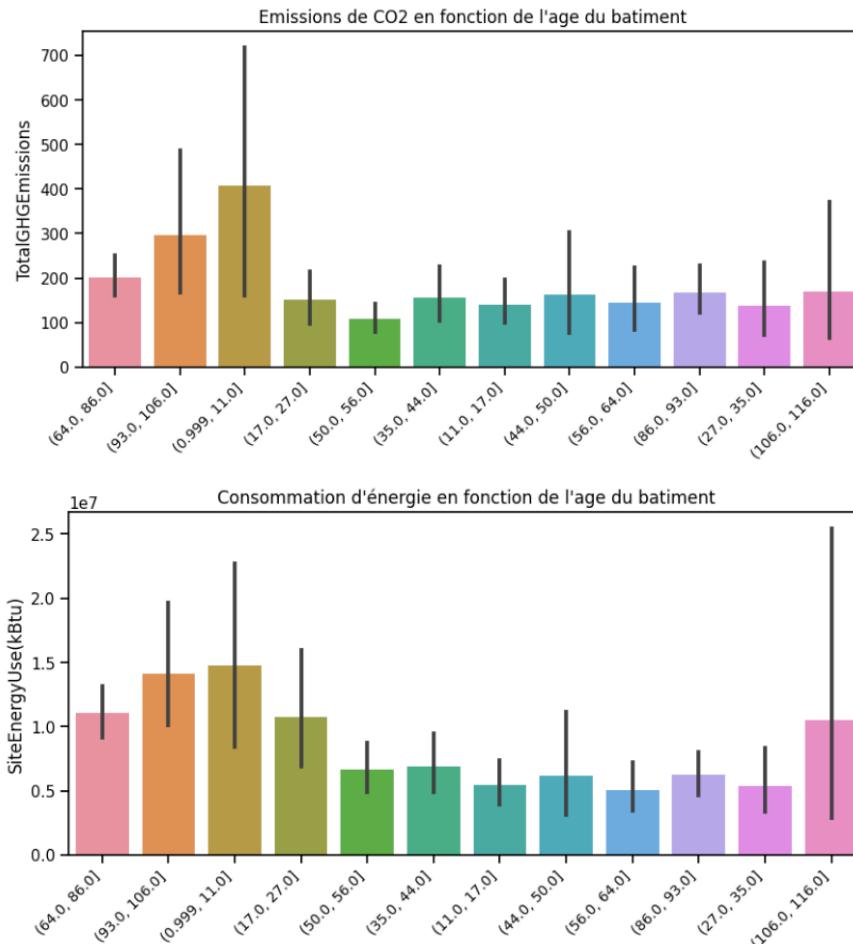
Feature selection & engineering (2/2)



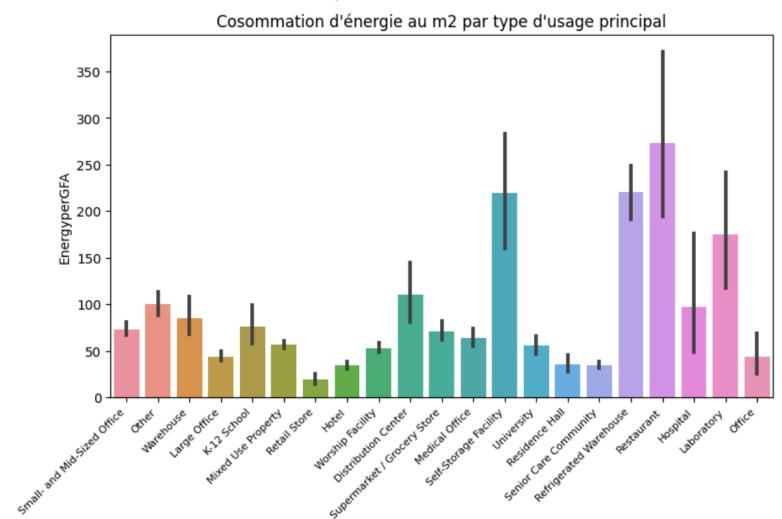
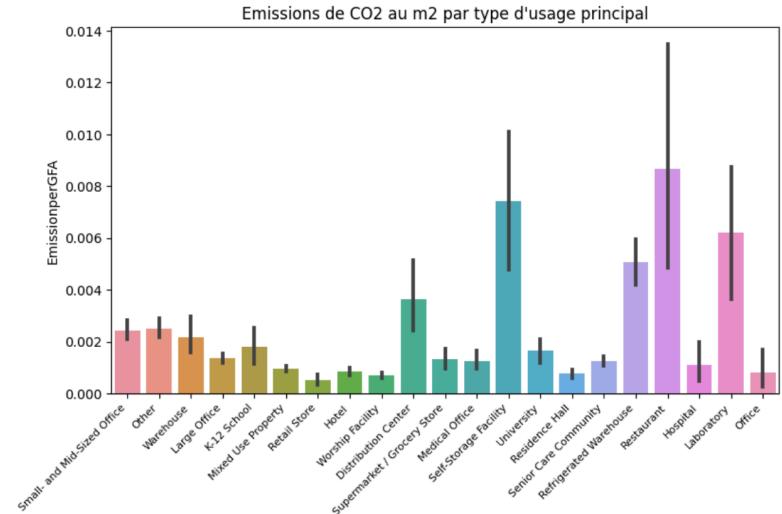
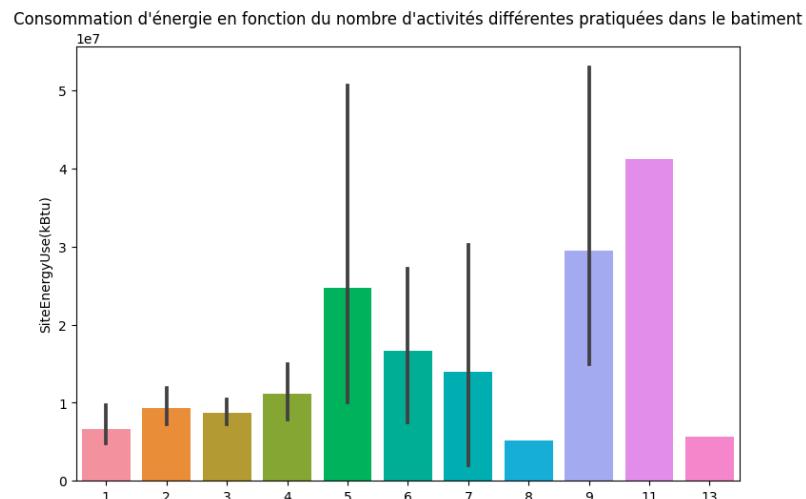
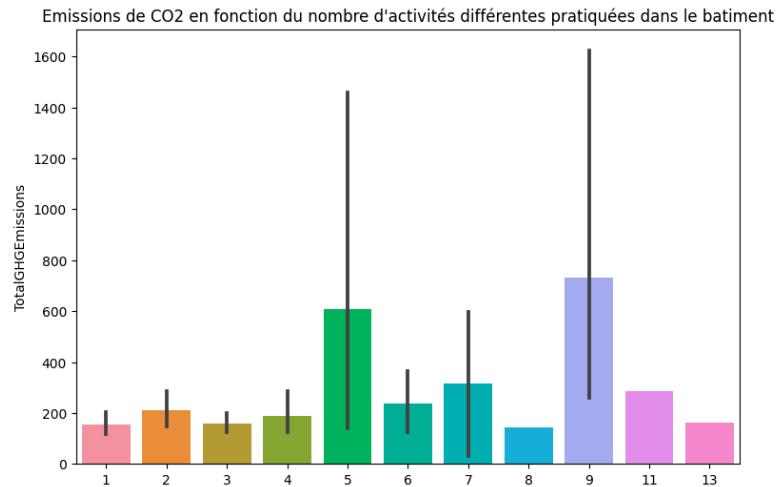
Analyse des targets



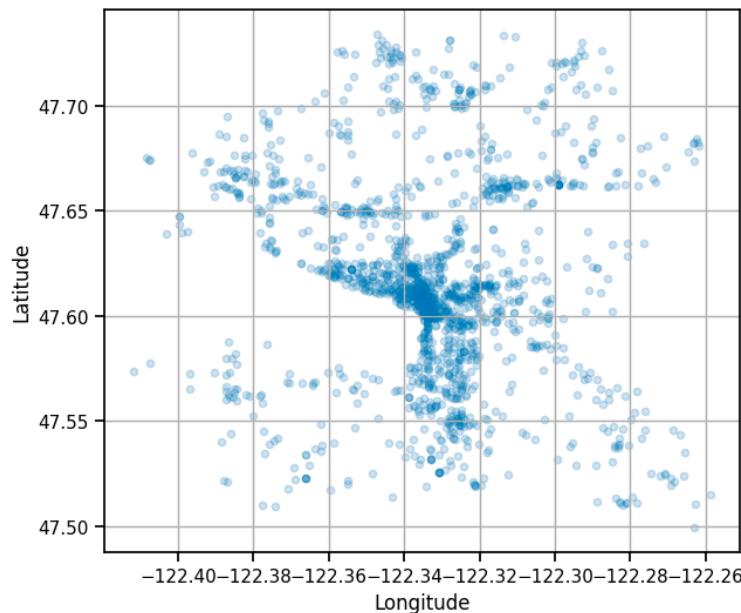
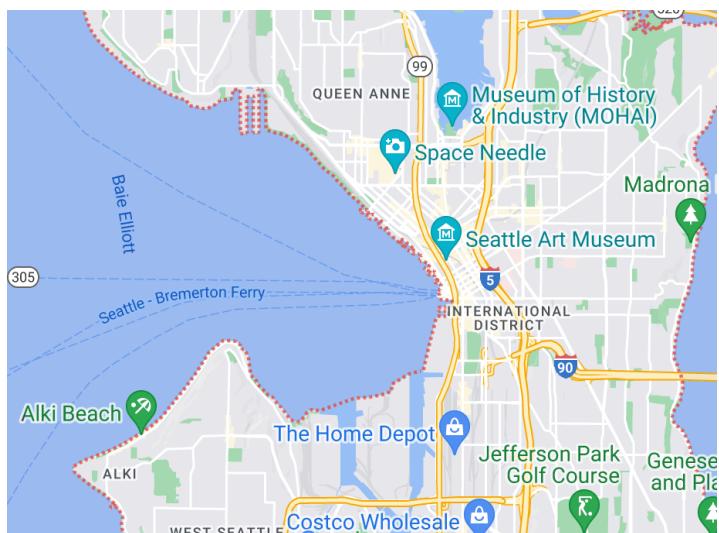
Targets en fonction des caractéristiques structurelles du bâtiment



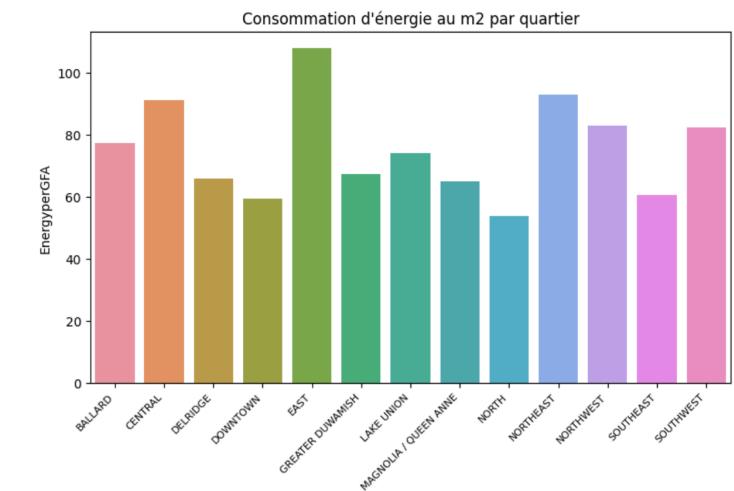
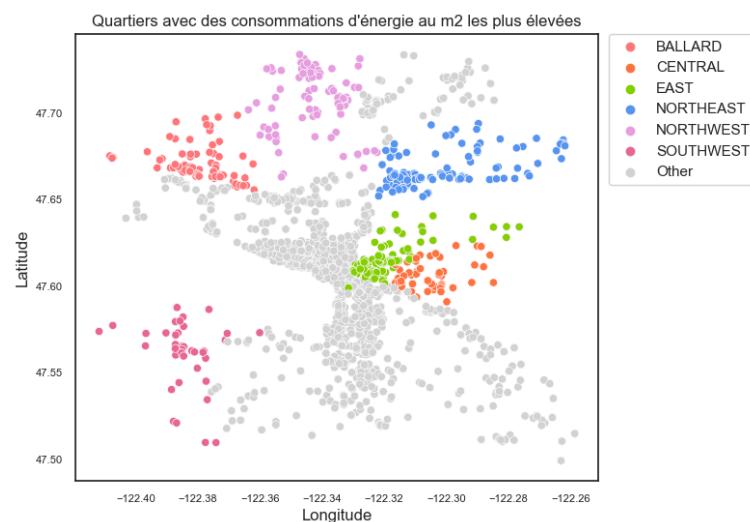
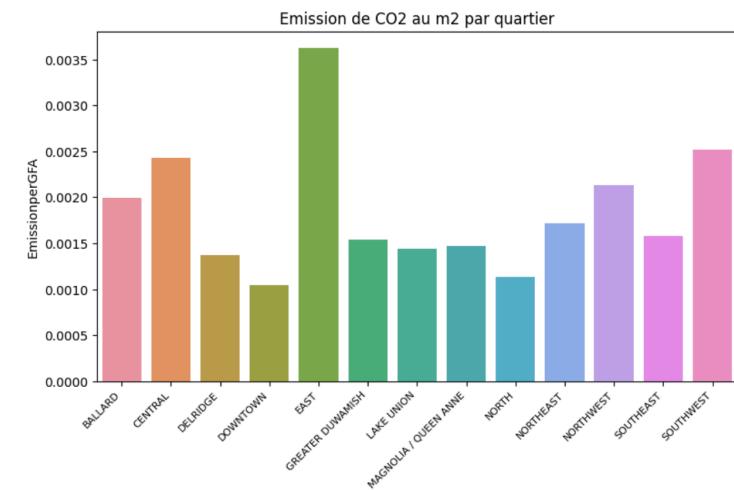
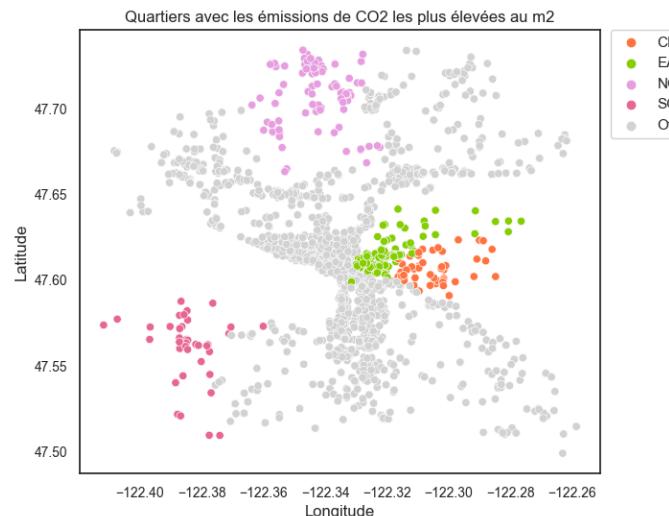
Targets en fonction des activités exercées sur la propriété



Analyse préliminaire de l'impact de la situation géographique



Analyse préliminaire de l'impact de la situation géographique



Feature preprocessing

- Choix de variables finales :

PropertyGFATotal	NumberofFloors	ShareSteam	NbOfUsage	ShareElec	BuildingGFAShare	BuildingAge	PrimaryPropertyType
88434	12	0.277302	1	0.546060	1.000000	89	Hotel
103566	11	0.000000	3	0.386609	0.854547	20	Hotel
956110	41	0.297113	1	0.682307	0.794252	47	Hotel
61320	10	0.325913	1	0.407519	1.000000	90	Hotel
175580	18	0.000000	3	0.378802	0.646885	36	Hotel

A noter :
Energy Star Score pas pris en compte dans un premier temps

- La variable catégorielle est « One hot » encodé :

Refrigerated Warehouse	Residence Hall	Restaurant	Retail Store	Self-Storage Facility	Senior Care Community	Supermarket / Grocery Store	University	Warehouse	Worship Facility
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

- Les variables numériques dont le skew est supérieur à 0.9 sont soumises à une transformation « log » np.log1p (pour leur faire approcher une distribution normale) et les autres sont scalées (« Standard Scaler »)

```
preprocessing = ColumnTransformer(
    transformers=[
        ('log', FunctionTransformer(np.log1p, validate=False), log_features),
        ("scale", StandardScaler(), scale_features)],
    remainder='passthrough')
```

- Les targets, qui ne suivent pas une loi normale, sont également soumises à une transformation log

Modélisation

Présentation des modèles

- Machine Learning **Supervisé** : le jeu de donnée sur lequel s'entraîne l'algorithme est labellisé
- Problème de **regression** : l'objectif est de prédire une valeur continue

Modèles linéaires

Linear Regression

Modèles qui cherchent les meilleurs paramètres (poids) de façon à « fitter » au mieux les données du training set en minimisant la fonction de coût (somme des carrés des erreurs)

Modèles régularisés

Ridge

Alpha
[0, inf]

Elastic Net

Alpha
[0, inf]

I1_ratio
[0,1]

Modèles non linéaires - Instance based

Modèles qui apprennent les données du training set et généralisent à des nouvelles données sur la base d'une mesure de similarité

KNeighbors Regressor

n_neighbors
[1, inf]

SVR

kernel

C
[0, inf]

epsilon
[0, inf]

Modèles ensemblistes

Agrégation de plusieurs modèles « faibles », entraînés sur les mêmes jeux de données pour trouver des prédictions plus performantes.

Random Forest

n_estimators
[1, inf]

max_depth

min_sample_leaf

Gradient boosting

n_estimators
[1, inf]

max_depth

learning_rate
[0, inf]

Modélisation

Présentation de la méthodologie

- **GridSearchCV** pour identifier les meilleurs hyperparamètres
 - ◆ Scoring = 'r2' et cv= KFold de 5 splits
- **Cross validation** afin de mettre en évidence les meilleurs résultats des modèles avec les hyperparamètres issues de la GridsearchCV
- **Calcul des scores** (R2, RMSE et MAE) des train et test set avec les hyperparamètres de la GridSearch pour évaluer les performances du modèle

La performance des algorithmes de régression sont évaluées à l'aune des indicateurs suivants:

- **R2** : coefficient de détermination (carré de la corrélation de Pearson entre les vrais valeurs et les valeurs prédictives). Il indique à quel point les valeurs prédictives sont corrélées aux vraies valeurs (valeur comprise entre 0 et 1)
- **MAE (Mean Absolute Error)** : erreur de prédiction moyenne (sans prendre en compte si positive ou négative)
 - Moins impacté par les outliers que MSE
 - Même unité de mesure que la target
 - Moins interprétable (moyenne de valeur absolue)
- **RMSE (Root Mean Squared Error)** : racine carré de l'erreur quadratique moyenne
 - Plus sensible aux outliers
 - Pénalise les grandes erreurs, donc particulièrement pertinent dans les cas où les erreurs importantes sont indésirables



- **Finetuning** des hyperparamètres pour aboutir

- ✓ à de meilleures scores - objectif : améliorer le R2 ou diminuer le MAE ou RMSE
- ✓ à une meilleure learning curve - objectif : réduire l'écart entre les courbes de training et de validation set (objectif : modèle généralisable)

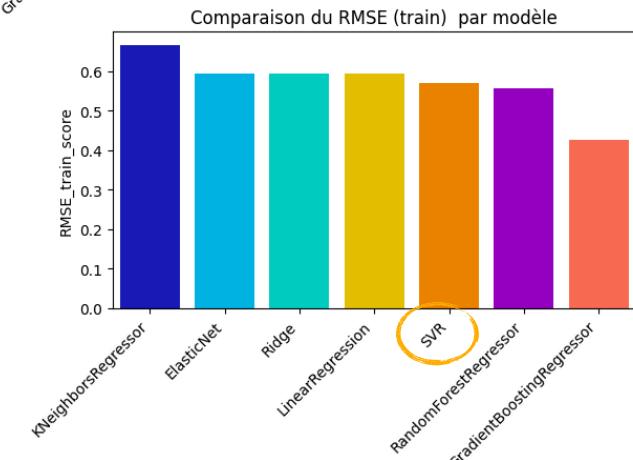
Modélisation

Comparaison des résultats - Emissions de CO2

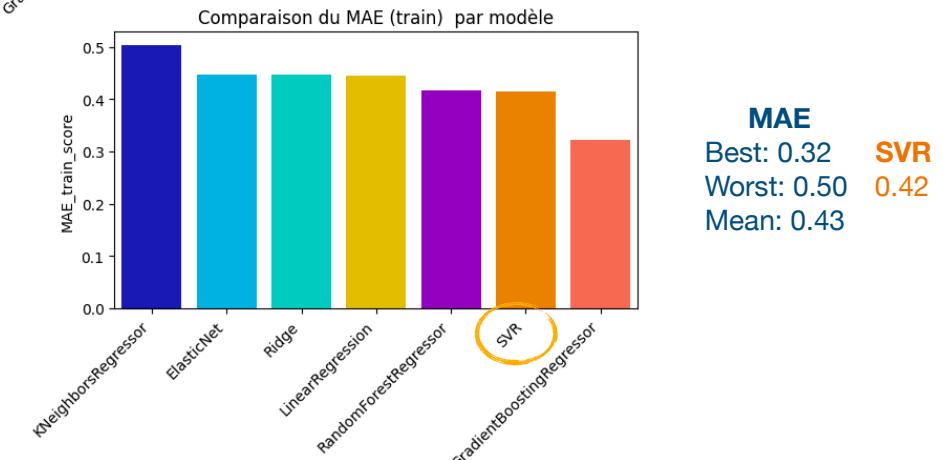
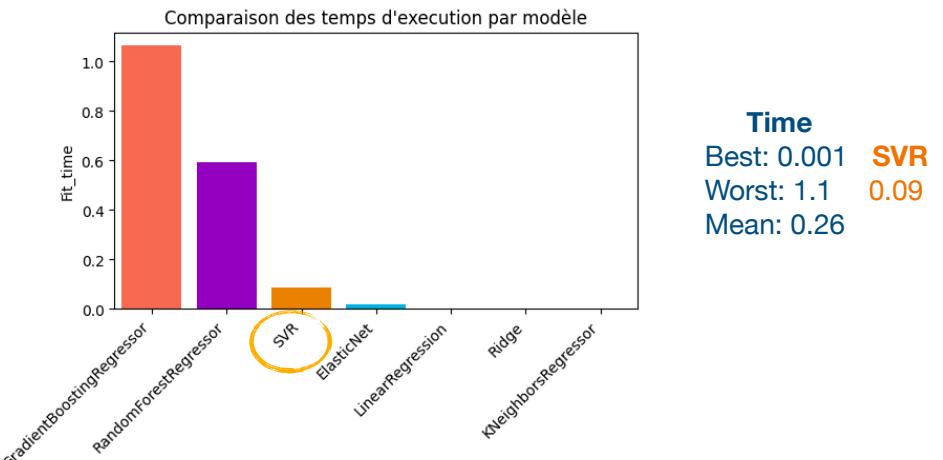
R2
Best: 0.91 **SVR**: 0.84
Worst: 0.78 **SVR**: 0.84
Mean: 0.84



RMSE
Best: 0.43 **SVR**: 0.57
Worst: 0.67 **SVR**: 0.57
Mean: 0.57



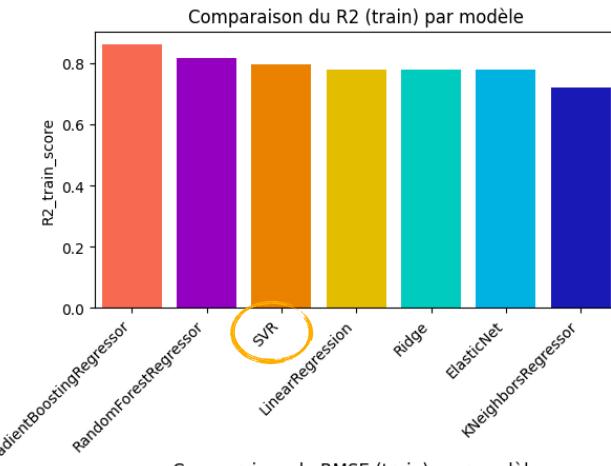
Time
Best: 0.001 **SVR**: 0.09
Worst: 1.1 **SVR**: 0.09
Mean: 0.26



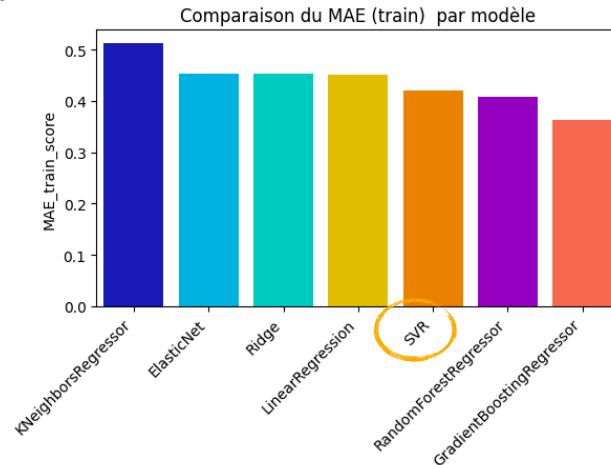
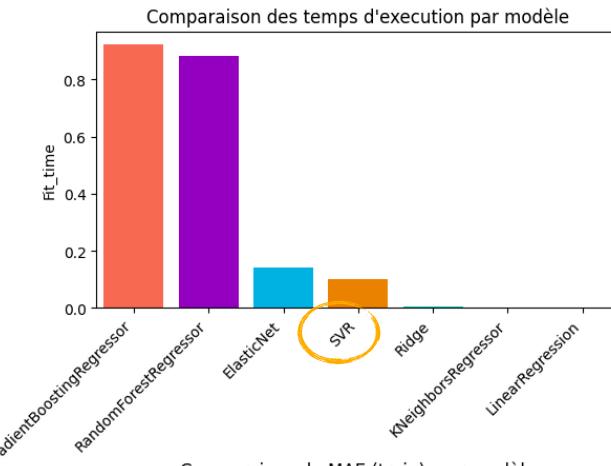
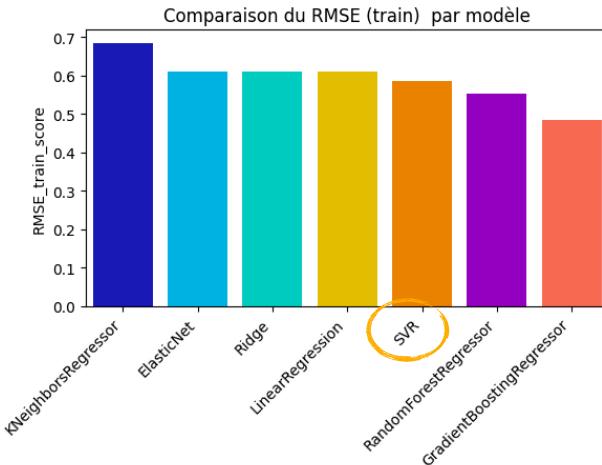
Modélisation

Comparaison des résultats - Consommation d'énergies

R2
Best: 0.86 **SVR**: 0.80
Worst: 0.72 Mean: 0.79



RMSE
Best: 0.48 **SVR**: 0.58
Worst: 0.68 Mean: 0.59



Time
Best: 0.002 **SVR**: 0.12
Worst: 1.08 Mean: 0.34

Modèle retenu : SVR

Détail des résultats

Emission CO2

--SVR--

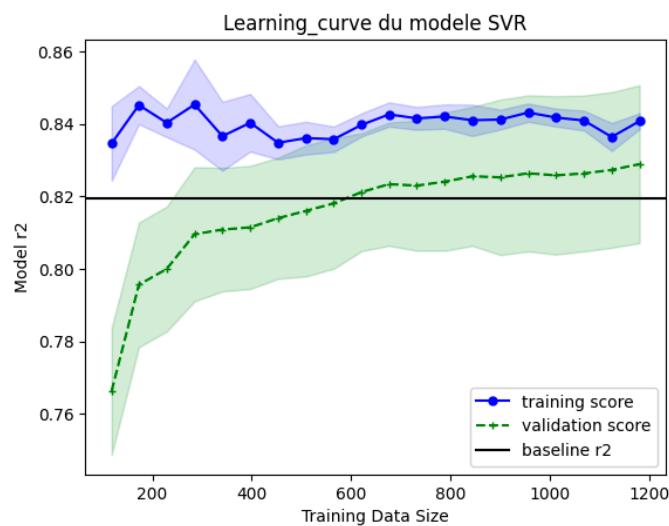
Le meilleur score obtenu avec la GridSearch est un r2 de 0.82 avec les paramètres {'C': 10, 'degree': 2, 'epsilon': 0.021544346900318822, 'kernel': 'rbf'}.

Le R2 moyen de la cross validation est de 0.82 avec un écart type de 0.03.
Le RMSE moyen de la cross validation est de 0.60 avec un écart type de 0.23.
Le MAE moyen de la cross validation est de 0.44 avec un écart type de 0.02.

R2
Training score:0.84
Test score:0.80

RMSE
Training score:0.57
Test score:0.63

MAE
Training score:0.42
Test score:0.45



Consommation d'énergie

--SVR--

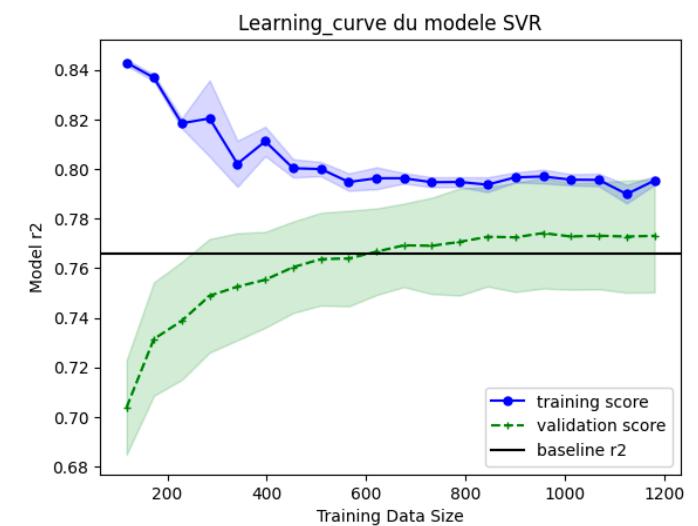
Le meilleur score obtenu avec la GridSearch est un r2 de 0.77 avec les paramètres {'C': 37.27593720314938, 'degree': 2, 'epsilon': 0.05994842503189409, 'kernel': 'rbf'}.

Le R2 moyen de la cross validation est de 0.77 avec un écart type de 0.04.
Le RMSE moyen de la cross validation est de 0.62 avec un écart type de 0.22.
Le MAE moyen de la cross validation est de 0.45 avec un écart type de 0.02.

R2
Training score:0.80
Test score:0.73

RMSE
Training score:0.58
Test score:0.67

MAE
Training score:0.42
Test score:0.46

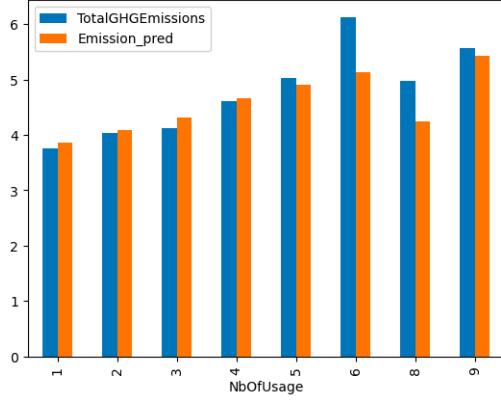


Modèle retenu : SVR

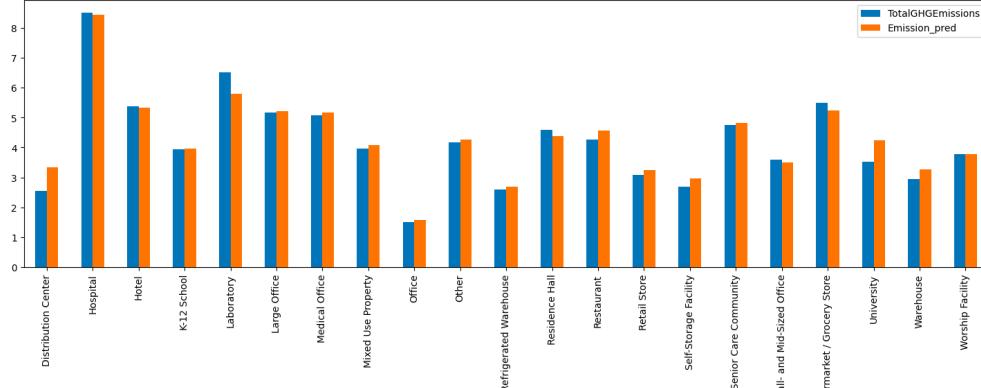
Analyse des prédictions

Emission CO₂

Ecarts de prédition sur la variable TotalGHGEmissions par nombre d'usage

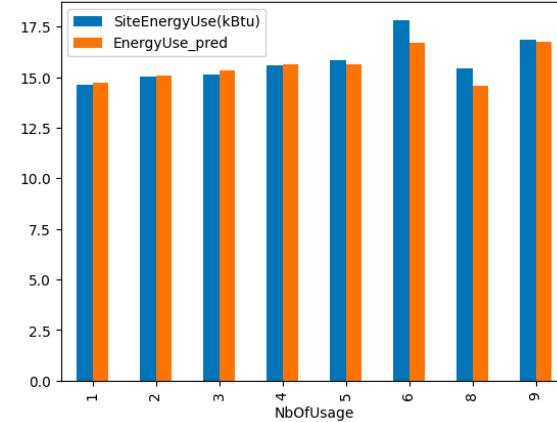


Ecarts de prédition sur la variable SiteEnergyUse par type de bâtiment

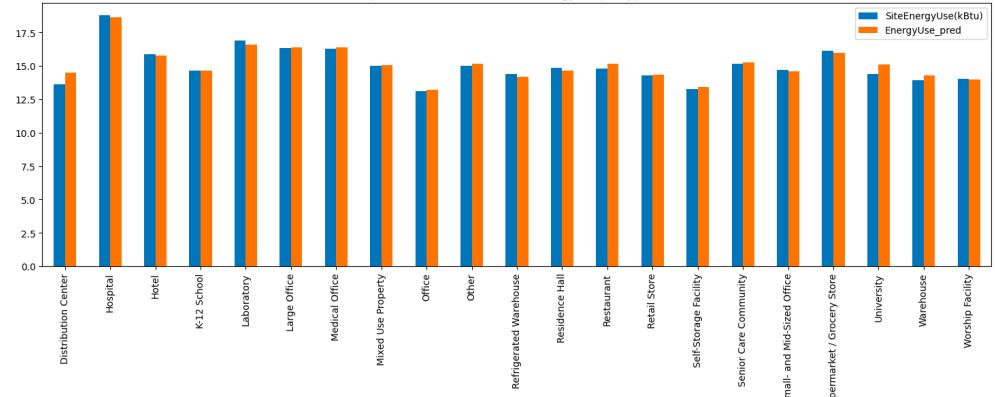


Consommation d'énergie

Ecarts de prédition sur la variable SiteEnergyUse par nombre d'usage

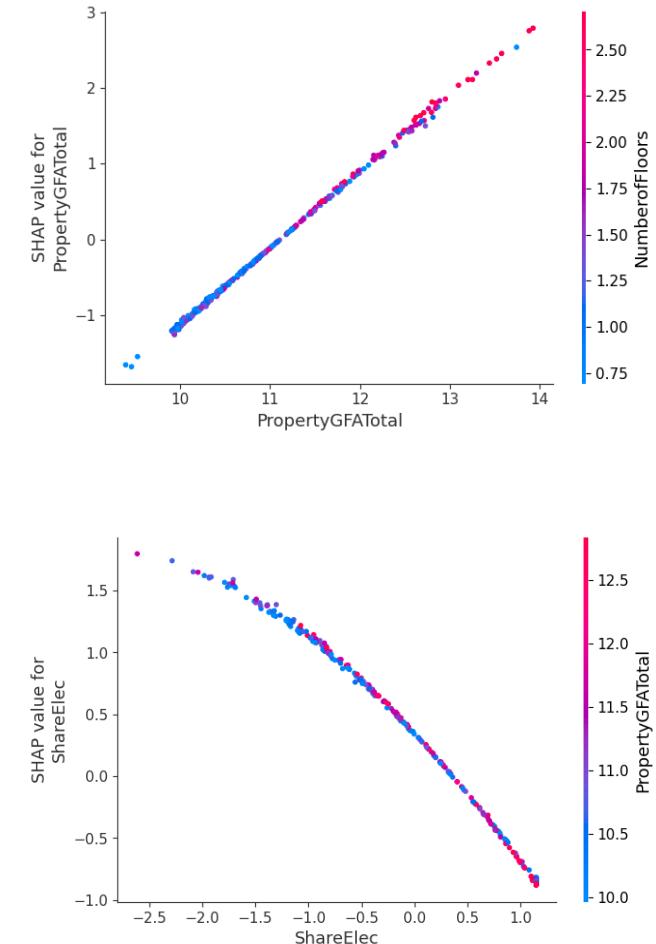
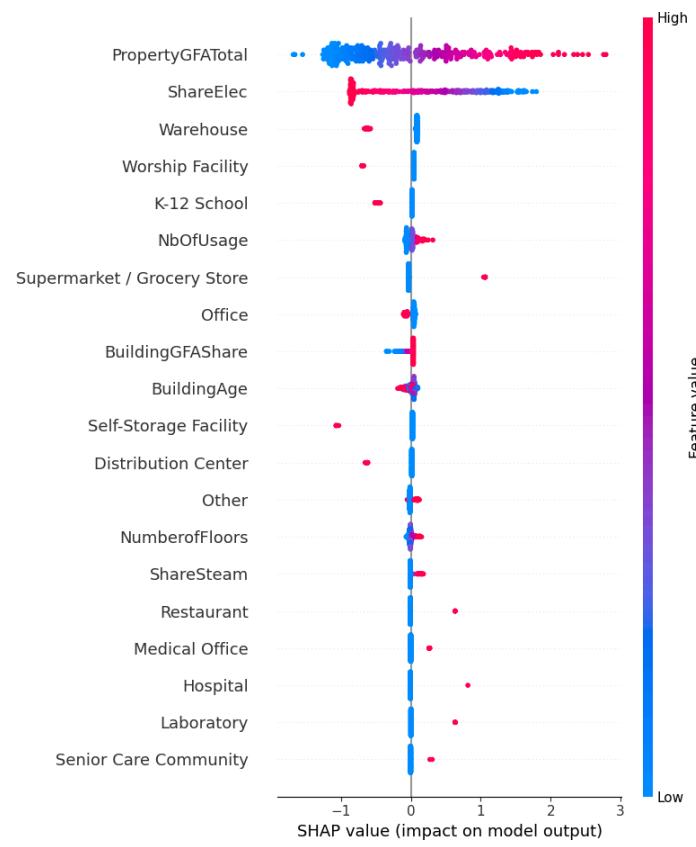
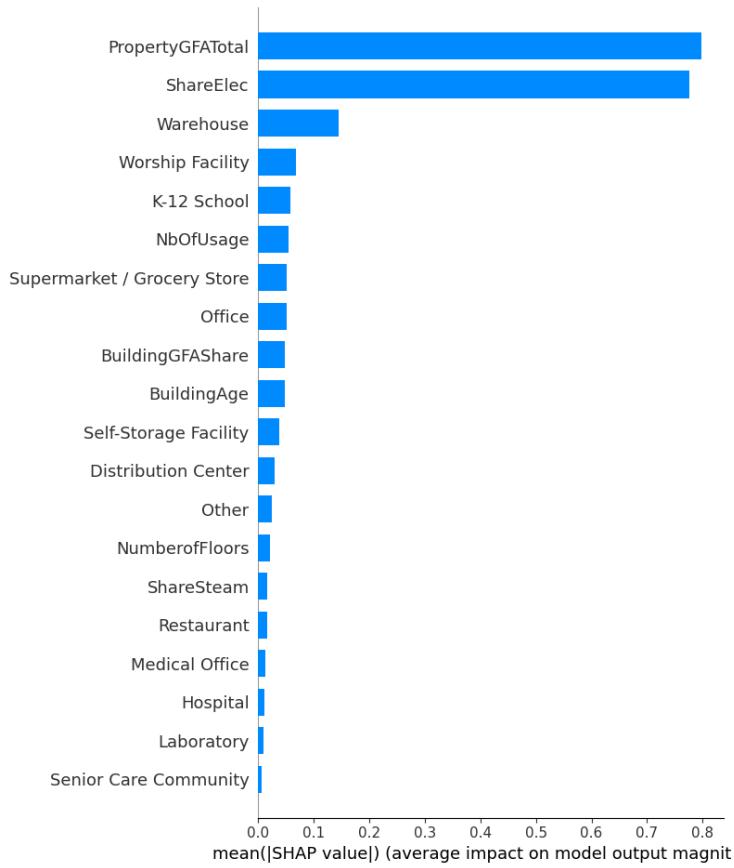


Ecarts de prédition sur la variable SiteEnergyUse par type de bâtiment



Analyse du modèle retenu

Global interpretability : Emission de CO2



Modèle retenu : SVR

Prise en compte Energy Star Score

 Sur la base des caractéristiques du bâtiment (taille, localisation, nombre d'occupants...), un algorithme estime combien d'énergie le bâtiment devrait consommer s'il s'agissait du plus performant, du moins performant, et de tous les niveaux intermédiaires. L'algorithme compare cette estimation avec la consommation effective du bâtiment pour déterminer son classement par rapport à ses pairs. Le score va de 1 à 100. 50 est la médiane : un bâtiment qui a un score inférieur à 50 performe moins bien que 50% de bâtiments comparables.

Emission CO₂

--SVR--

Le meilleur score obtenu avec la GridSearch est un r2 de 0.92 avec les paramètres {'C': 20, 'degree': 2, 'epsilon': 0.21544346900318845, 'kernel': 'rbf'}.

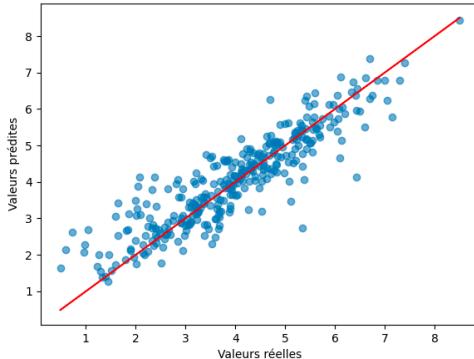
Le R2 moyen de la cross validation est de 0.92 avec un écart type de 0.01. Le RMSE moyen de la cross validation est de 0.39 avec un écart type de 0.14. Le MAE moyen de la cross validation est de 0.27 avec un écart type de 0.01.

R2
Training score:0.93
Test score:0.92

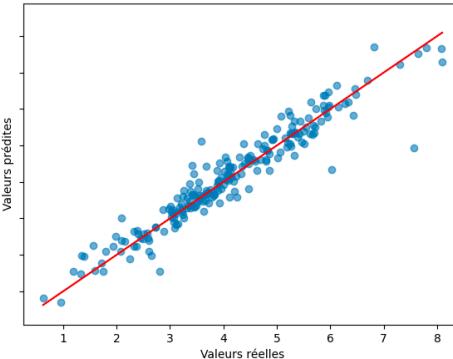
RMSE
Training score:0.37
Test score:0.40

MAE
Training score:0.25
Test score:0.27

Valeurs prédites vs valeur réelles - Emissions de CO₂ (sans Escore)



Valeurs prédites vs valeur réelles - Emissions de CO₂ (avec Escore)



Consommation d'énergie

--SVR--

Le meilleur score obtenu avec la GridSearch est un r2 de 0.90 avec les paramètres {'C': 100.0, 'degree': 2, 'epsilon': 0.21544346900318845, 'kernel': 'rbf'}.

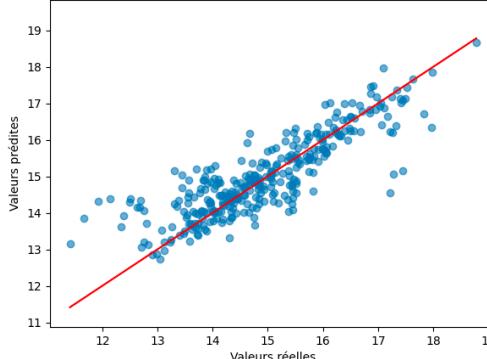
Le R2 moyen de la cross validation est de 0.90 avec un écart type de 0.02. Le RMSE moyen de la cross validation est de 0.40 avec un écart type de 0.15. Le MAE moyen de la cross validation est de 0.27 avec un écart type de 0.01.

R2
Training score:0.92
Test score:0.89

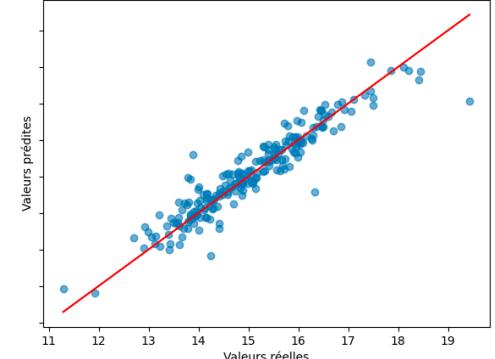
RMSE
Training score:0.36
Test score:0.42

MAE
Training score:0.24
Test score:0.28

Valeurs prédites vs valeur réelles - Conso d'énergie (sans Escore)



Valeurs prédites vs valeur réelles - Conso d'énergie (avec Escore)



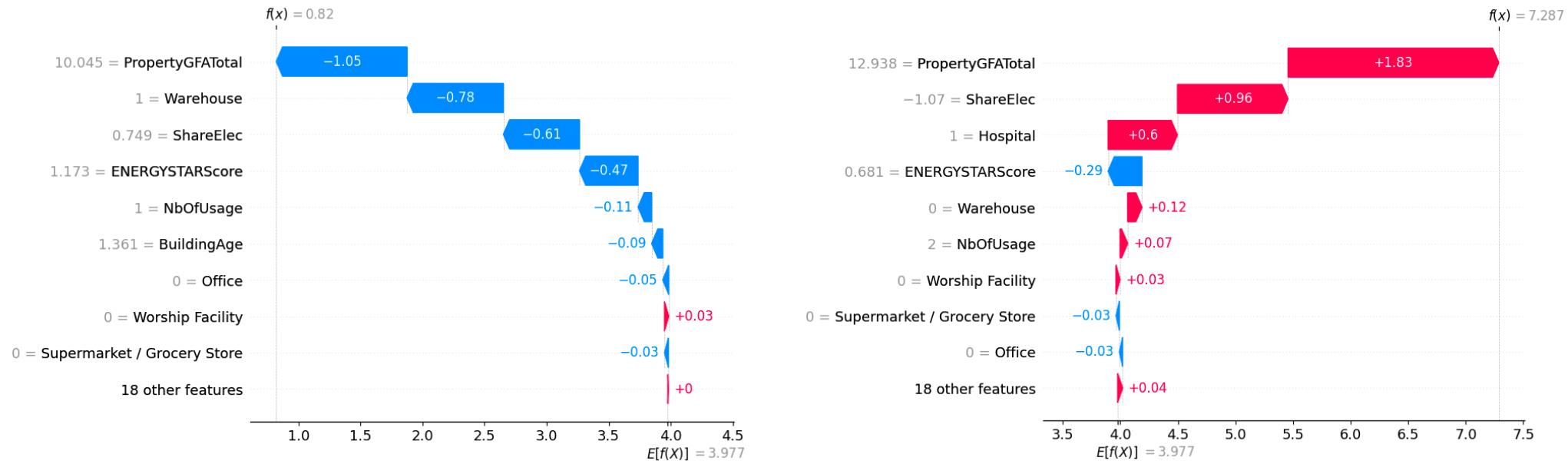
Analyse du modèle retenu

Local interpretability : Emission CO₂

Force plot d'une prédiction random



Waterfall plots pour la prédiction la plus élevée et la moins élevée



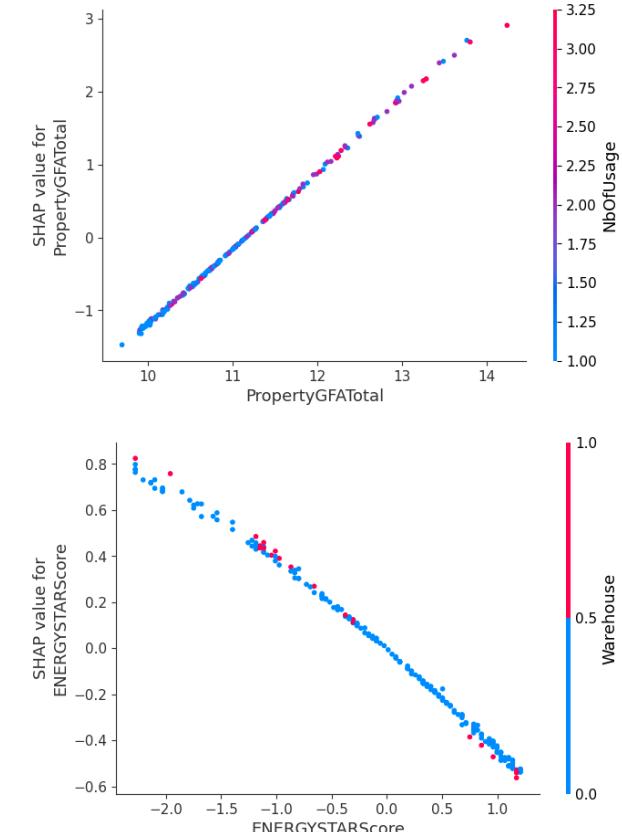
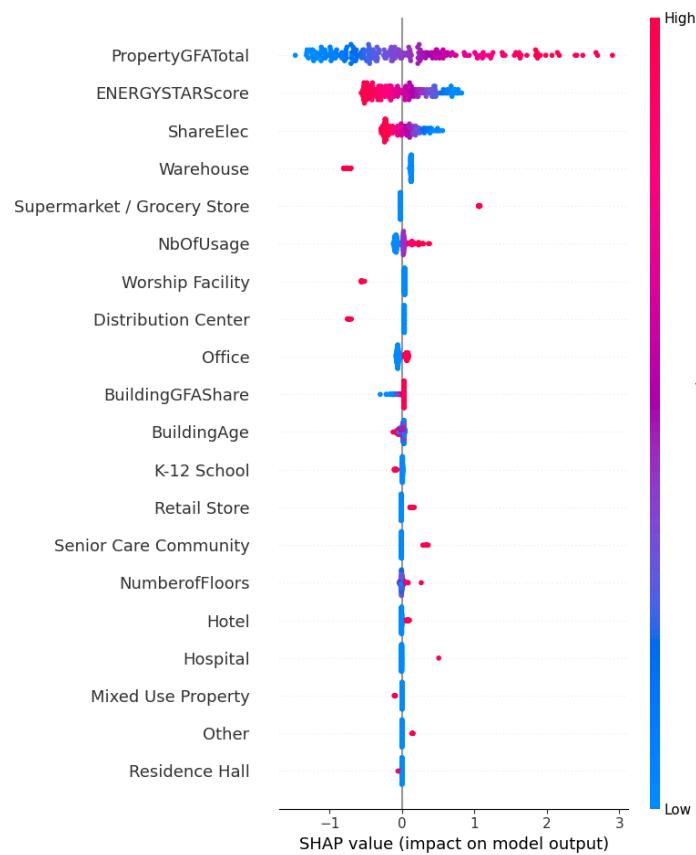
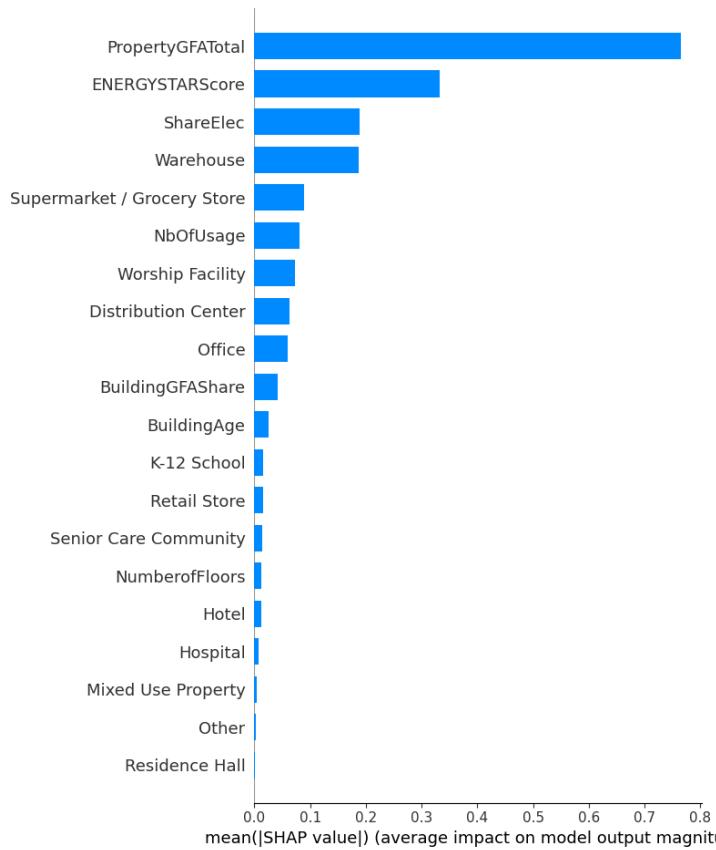
Annexes

Technologies utilisées

- Les travaux ont été réalisés sur Python (3.9.6) avec les bibliothèques ci-dessous:
 - numpy 1.23.5
 - pandas 2.0.2
 - matplotlib 3.7.1
 - seaborn 0.12.2
 - scikit-learn 1.2.2
 - shap 0.41.0
- Les modèles de machine learning ont été implémentés grâce à la librairie scikit-learn via les méthodes suivantes:
 - Régression linéaire : `sklearn.linear_model.LinearRegression`
 - Régression Ridge : `sklearn.linear_model.Ridge`
 - Régression Elastic Net : `sklearn.linear_model.ElasticNet`
 - Régression K Neighbors : `sklearn.neighbors.KNeighborsRegressor`
 - Régression Support Vector (SVR) : `sklearn.svm.SVR`
 - Régression Random Forest : `sklearn.ensemble.RandomForestRegressor`
 - Régression Gradient Boosting : `sklearn.ensemble.GradientBoostingRegressor`

Analyse du modèle retenu

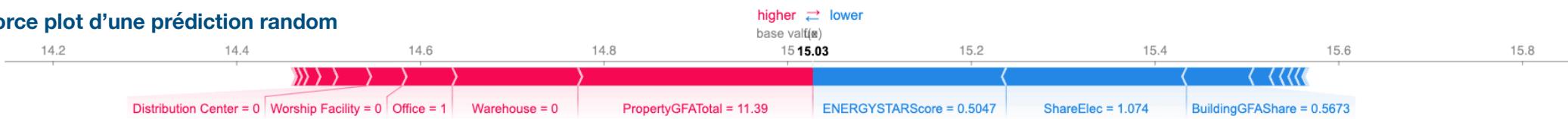
Global interpretability : consommation d'énergie avec Energy Star Score



Analyse du modèle retenu

Local interpretability : consommation d'énergie avec Energy Star Score

Force plot d'une prédiction random



Waterfall plots pour la prédiction la plus élevée et la moins élevée

