



Anticipez les besoins en consommation de bâtiments

Projet 3 du parcours Machine Learning Engineer

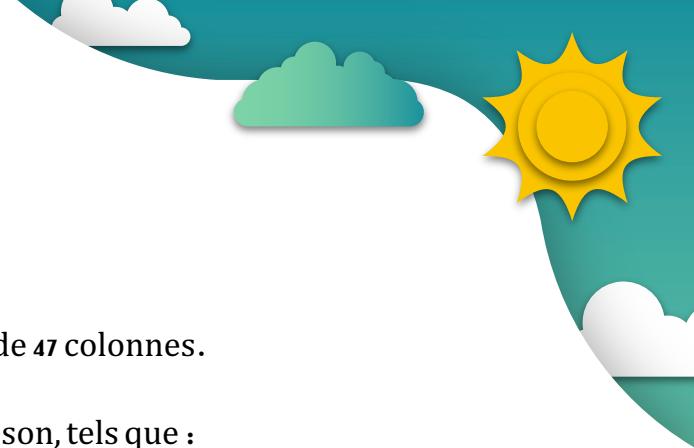
Contexte et objectif

Pour son objectif de ville neutre en **2050**. La ville de Seattle souhaite analyser et comprendre les tendances de consommation d'énergie des bâtiments.

Objectif : L'objectif principal de cette étude est d'explorer les modèles de consommation d'énergie des bâtiments de Seattle à l'aide de techniques d'apprentissage supervisé. En analysant ces modèles, nous cherchons à identifier les principaux facteurs qui influent sur la consommation d'énergie des bâtiments et à développer des modèles prédictifs précis pour estimer la consommation d'énergie future.



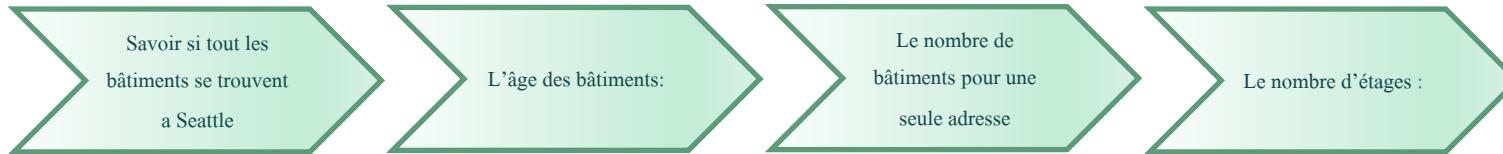
Présentation des données



Le jeu de données est fourni par la ville de Seattle.

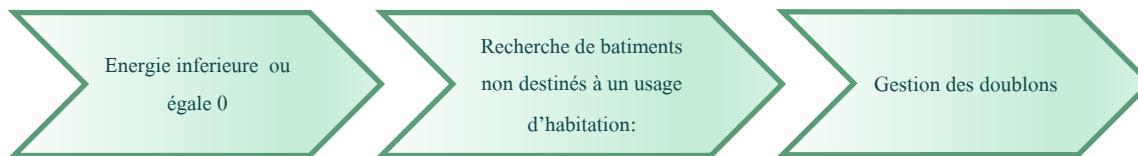
- Nombre de Colonnes : Le jeu de données comprend un total de 47 colonnes.
- Types de Données : Les données comprennent une combinaison, tels que :
 - Données catégorielles(15) : Informations sur le type de bâtiment, le nom de la propriété, l'adresse, etc.
 - Données numériques (31) : Informations sur les superficies, les années de construction, les scores ENERGY STAR, les consommations d'énergie, etc.
- Variables d'Intérêt : Certaines variables clés incluent :
 - Type de bâtiment et usage principal de la propriété.
 - Localisation géographique : Latitude, longitude, quartier, etc.
 - Caractéristiques du bâtiment : Année de construction, nombre d'étages, surface brute, etc.
 - Performances énergétiques : Scores ENERGY STAR, intensité d'utilisation de l'énergie, émissions de gaz à effet de serre, etc.

Gestion des valeurs aberrantes est des outliers



1 ligne supprimée

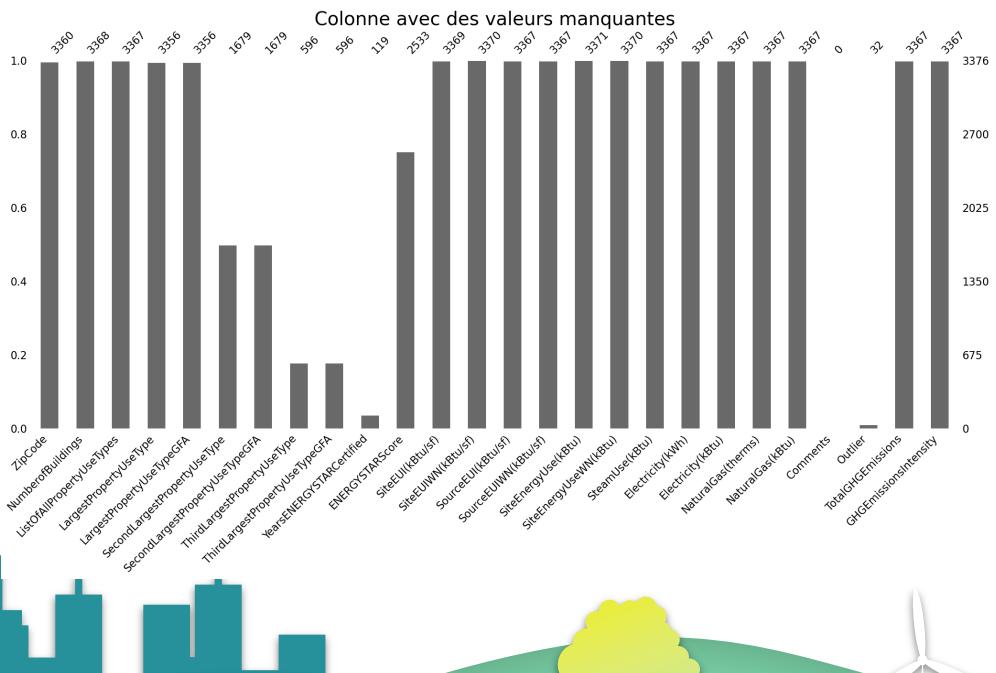
1 ligne supprimée.
99 lignes où la valeur 0 a été remplacé par 1



1 ligne où l'énergie est inférieur à 0. 14 ou elle est égale à 0. ces lignes sont supprimées

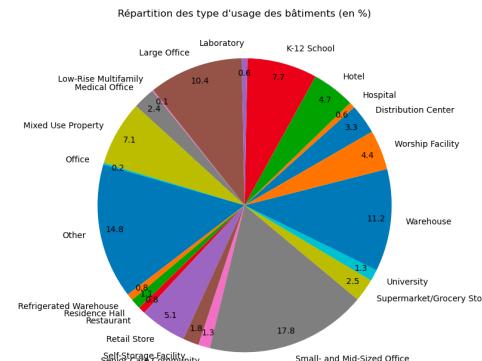
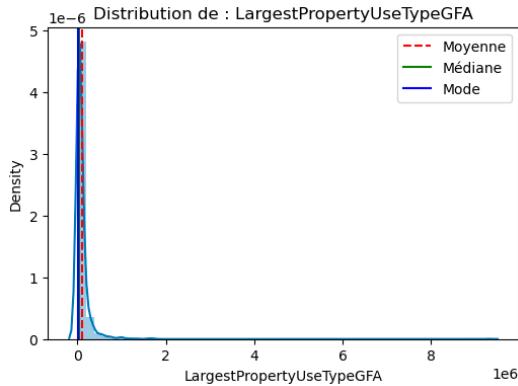
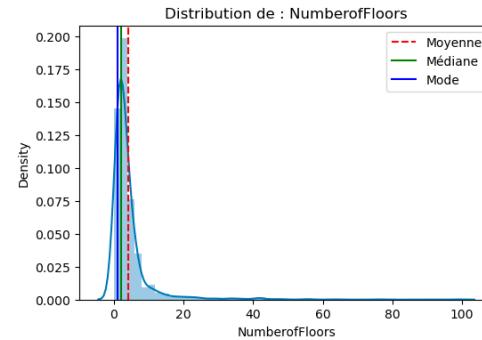
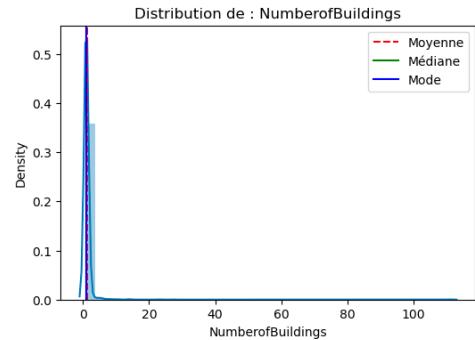
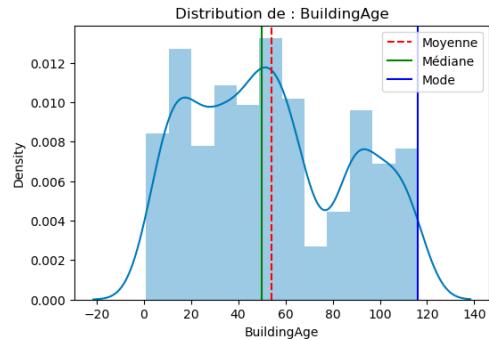
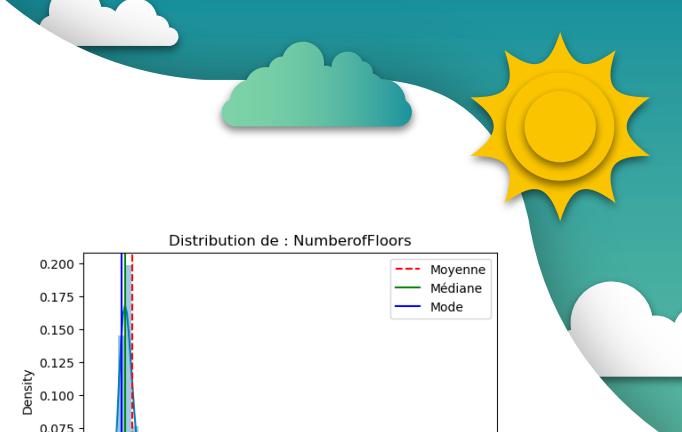
1708 lignes supprimées

Données manquantes



colonne	gestion
Colonne avec plus de 80% de val manquantes	Colonne supprimée
Variables cible	Ligne supprimée
Seconde et third property use type	'no information'
type_GFA	0
_WN, _Kwh	Colonnes non utilisées

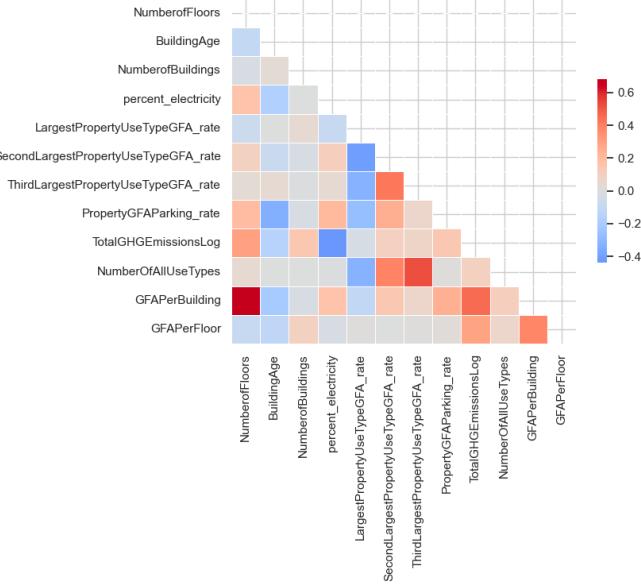
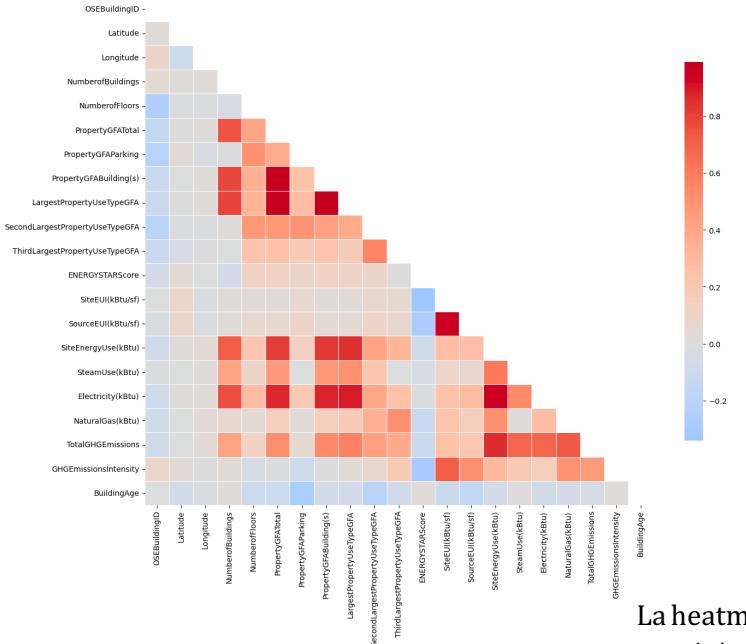
Analyse exploratoire des données



Distribution des Variables Numériques : Des histogrammes ont été utilisés pour visualiser la distribution des variables numériques telles que l'année de construction, la superficie, etc.

Visualisation des variables catégorielles : Utilisation de bar plots et pie charts pour explorer la répartition des catégories.

Corrélation entre les variables.

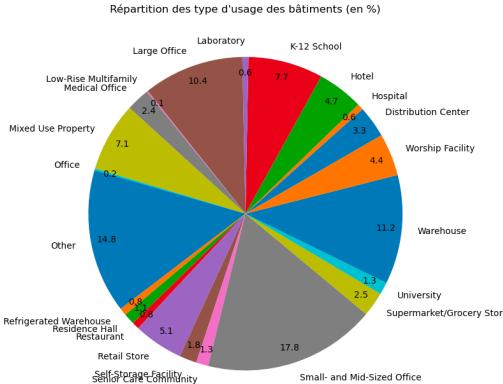


La heatmap nous permet de visualiser les liens de corrélation entre les variables. On peut constater des fortes corrélations :

- Entre les variables cibles entre elles.
- Entre les variables de surface.
- Entre les variables de surface et l'énergie utilisée.

Ces informations sont importantes à prendre en compte pour éviter les problèmes de colinéarité.

Modification et création de variables



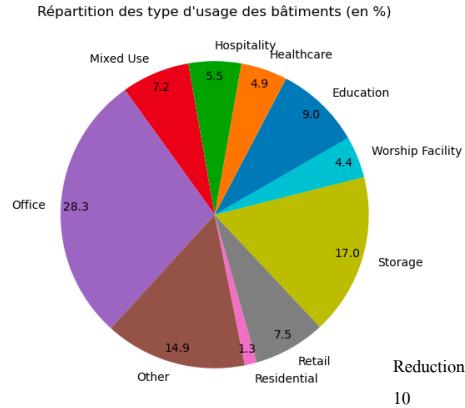
Ajout de nouvelles colonnes:

Les colonnes most_use_steam, most_use_natural et most_use_electric identifient respectivement si le chauffage: la vapeur, le gaz naturel ou l'électricité est la source d'énergie la plus utilisée pour chaque bâtiment.

Les colonnes use_stream, use_electricity et use_natural indiquent respectivement si le chauffage à la vapeur, l'électricité ou le gaz naturel est utilisé dans chaque bâtiment.

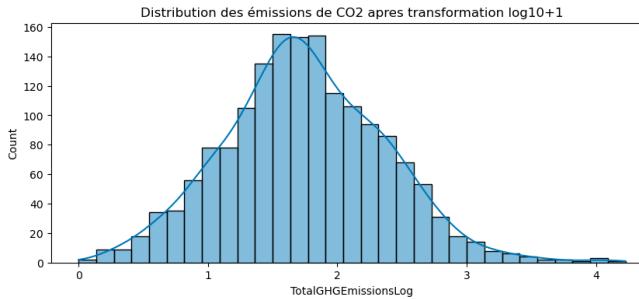
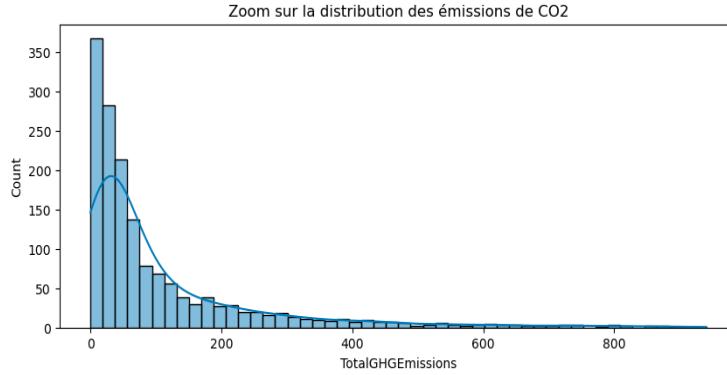
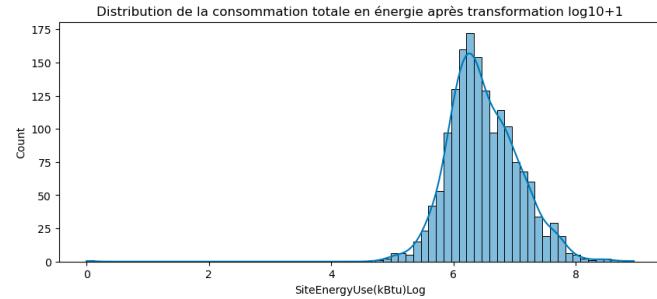
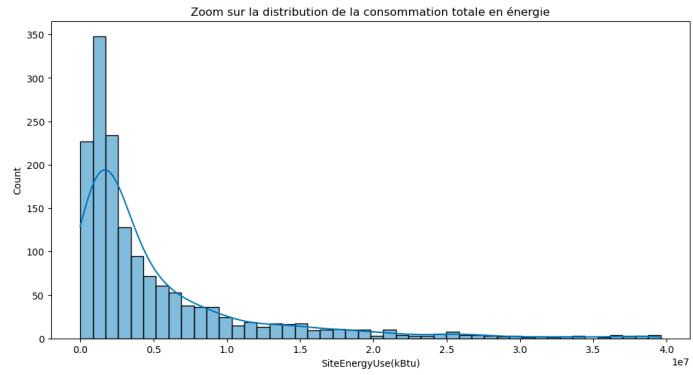
Les colonnes percent_steam, percent_natural et percent_electricity représentent le pourcentage d'utilisation de chaque source d'énergie par rapport à la consommation totale d'énergie pour chaque bâtiment.

La colonne BuildingAge représente l'âge des bâtiments, indiqué en années.



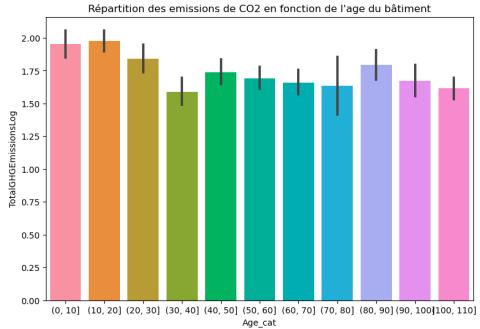
Reduction du nombre d'usage des bâtiments: de 21 à 10

Analyse des variables cibles.

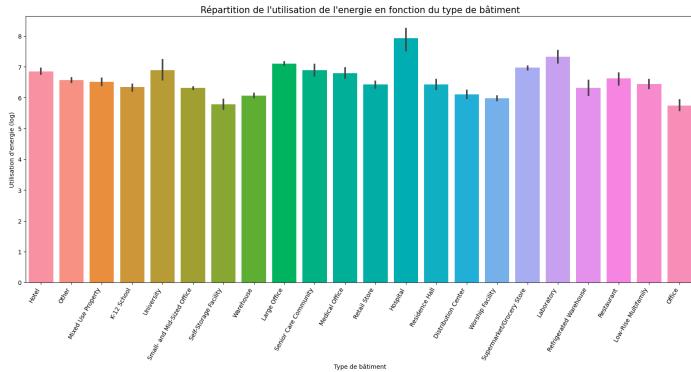
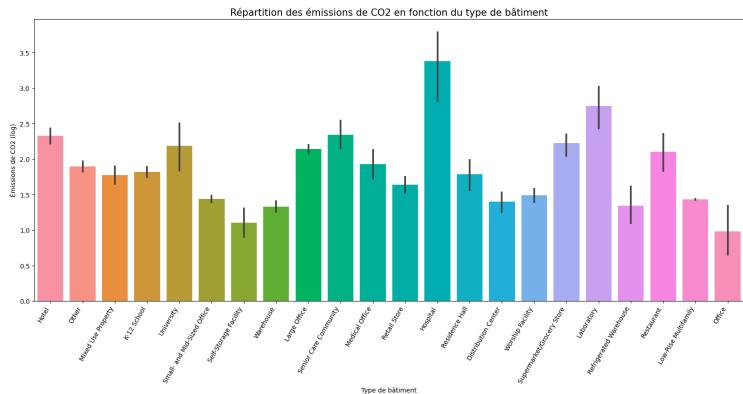
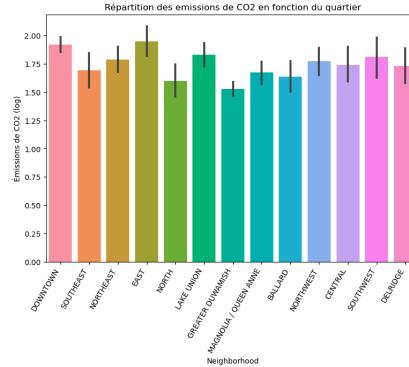


Les variables 'TotalGHGEmissions' et 'SiteEnergyUse(kBtu)' présentent des distributions fortement asymétriques à droite, avec quelques valeurs extrêmement élevées. Pour mieux modéliser ces distributions, nous avons appliqué une transformation logarithmique aux données. Cette transformation a réduit l'asymétrie et l'aplatissement des distributions.

Analyse exploratoire des variables cibles.



En observant ces exemples, on peut constater que certaines variables explicatives ont une incidence sur la variable cible.



The background features a stylized illustration of a green landscape. On the left, there's a yellow sun with rays, a white cloud, and two white wind turbines with three blades each. In the center, there are several green trees of different shapes and sizes. On the right, there are some green buildings with windows. The overall style is flat and modern.

Modélisation

Les variables explicatives

Caractéristiques du bâtiment :

-NumberofFloors, BuildingAge, NumberofBuildings, GFAPerBuilding, GFAPerFloor

Utilisation de l'énergie :

-percent_electricity, SiteEnergyUse(kBtu)Log

Type de propriété :

-LargestPropertyUseType, SecondLargestPropertyUseType, ThirdLargestPropertyUseType,

-PrimaryPropertyType

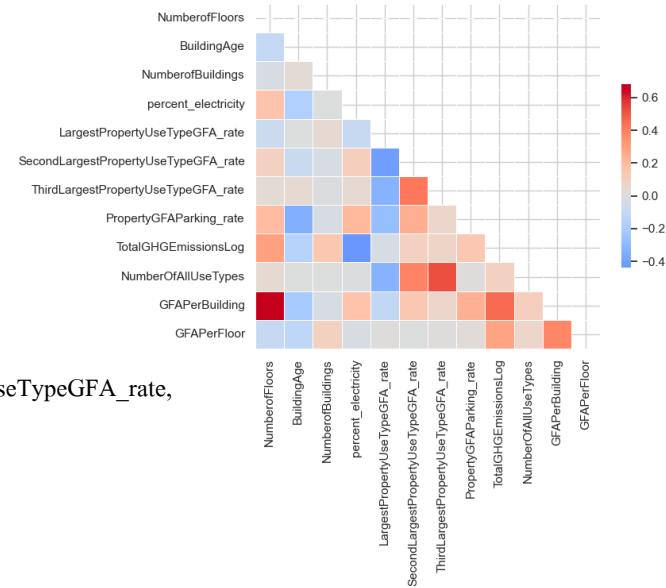
Taux de surface de la propriété :

-LargestPropertyUseTypeGFA_rate, SecondLargestPropertyUseTypeGFA_rate, ThirdLargestPropertyUseTypeGFA_rate,

PropertyGFAParking_rate

Caractéristiques environnementales :

-Neighborhood



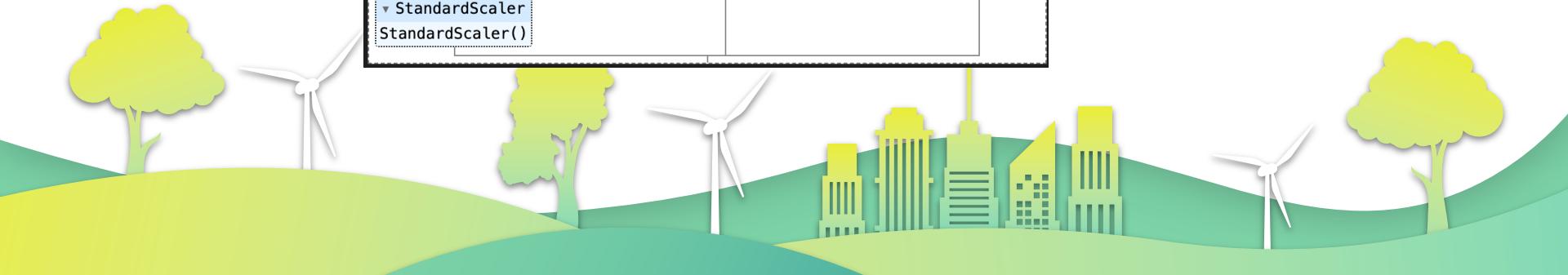
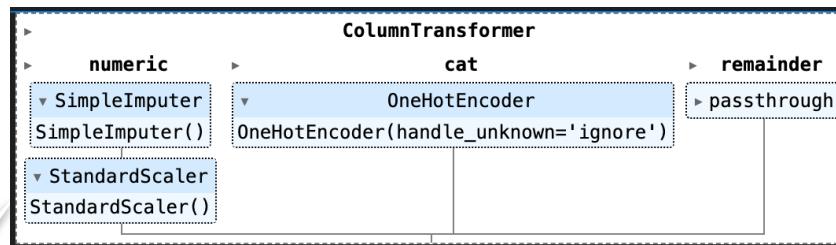
Ces variables ont été conservées car elles offrent une perspective holistique des caractéristiques physiques, énergétiques et environnementales des bâtiments, ce qui est crucial pour comprendre et prédire leur performance énergétique.

Pipelines et transformation

Pipelines: L'utilisation de pipelines simplifie et organise le processus de prétraitement des données en regroupant les différentes étapes de transformation.

Transformation des données numériques: StandardScaler standardise ces variables pour une meilleure convergence des algorithmes.

Encodage des variables catégorielles: OneHotEncoder est privilégié pour convertir les catégories en variables binaires, facilitant ainsi le traitement par les modèles.



Les algorithmes testés

Linéaire:

Regression linéaire: modélise la relation entre une variable dépendante continue et une ou plusieurs variables indépendantes.

Ensembliste:

Bagging ([RandomForest](#)): vise à réduire la variance des modèles en combinant les prédictions de plusieurs modèles de base, chacun formé sur un échantillon bootstrap différent de l'ensemble de données d'origine.

Boosting ([Adaboost](#), [LightGBM](#)): consiste à combiner plusieurs modèles de base de manière séquentielle, où chaque modèle cherche à corriger les erreurs des modèles précédents. Il met l'accent sur les échantillons mal prédits pour améliorer progressivement les performances du modèle global.

Métrique d'évaluation:

-[R2](#): Le R2 mesure la proportion de la variance de la variable dépendante qui est expliquée par le modèle.

-[RMSE](#): mesure de l'erreur moyenne entre les valeurs prédites par le modèle et les valeurs réelles.

-[MSE](#)est le carré moyen des écarts entre les valeurs prédites par le modèle et les valeurs réelles.

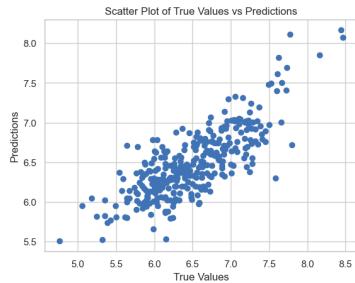
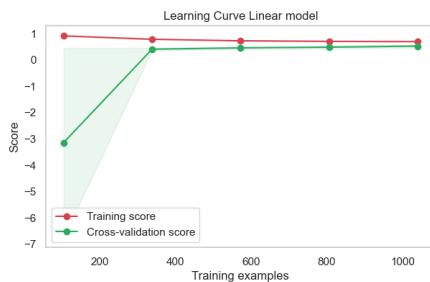
-[Fit_time](#):la durée nécessaire pour entraîner le modèle sur les données d'entraînement.



Régression linéaire

La régression linéaire est une méthode statistique visant à modéliser la relation linéaire entre une variable dépendante et une ou plusieurs variables indépendantes. L'objectif est d'ajuster un modèle qui minimise la somme des carrés des erreurs entre les valeurs prédictées et les observations réelles.

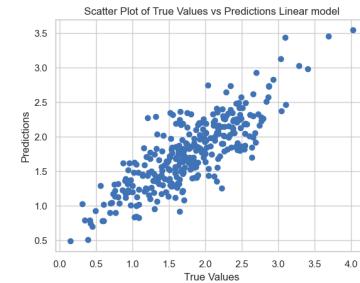
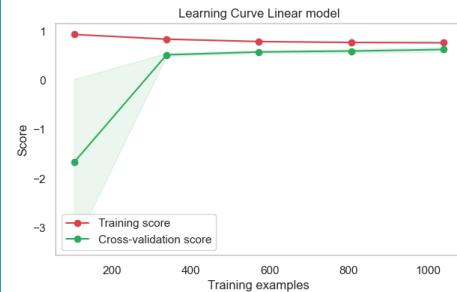
Modélisation du Site energy use



```
linear_model = Pipeline([
    ('prepa', prep),
    ('linear', linear_model.LinearRegression(fit_intercept=True))
])
```

RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test
0.311236	0.387061	0.689761	0.320466	0.292915	0.032298	0.004568

Modélisation du Co2



```
linear_model = Pipeline([
    ('prepa', prep),
    ('linear', linear_model.LinearRegression(fit_intercept=True))
])
```

RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test
0.305645	0.383392	0.756728	0.52204	0.288613	0.034284	0.004358

RandomForest

les forêts aléatoires sont une méthode robuste, flexible et facile à utiliser qui fonctionne bien dans une beaucoup de situations. Il fonctionne en construisant un ensemble de nombreux arbres de décision indépendants les uns des autres. Chaque arbre est formé sur un sous-ensemble aléatoire des données et vote pour la prédiction finale.

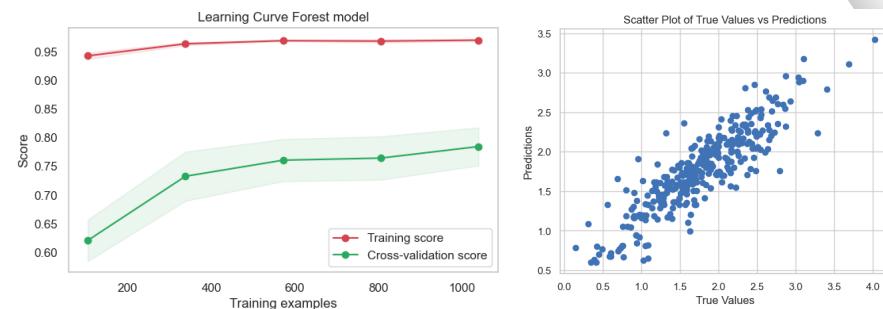
Modélisation du Site energy use



```
forest_model = Pipeline(steps=[  
    ('prepa', prepator),  
    ('forest', RandomForestRegressor(n_estimators=200, bootstrap=True, max_features=0.5,  
                                    min_samples_split=3, min_samples_leaf=1, random_state=4))
```

	RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test	
	0	0.112018	0.292641	0.959801	0.605692	0.214334	2.419845	0.02304

Modélisation du Co2



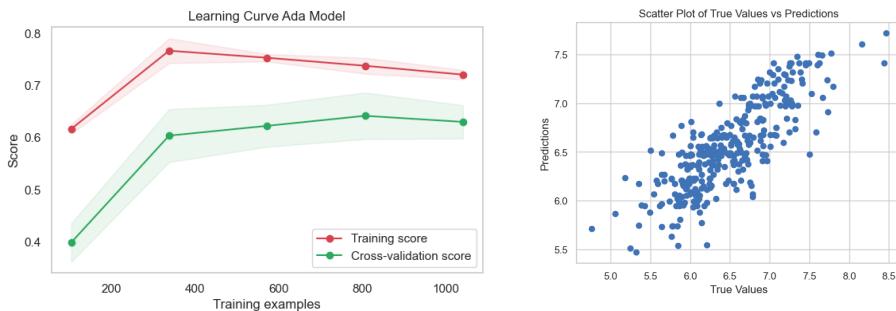
```
forest_model = Pipeline(steps=[  
    ('prepa', prepator),  
    ('forest', RandomForestRegressor(n_estimators=150, bootstrap=True, max_features=0.5,  
                                    min_samples_split=2, min_samples_leaf=1, random_state=1))
```

	RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test	
	0	0.103911	0.282254	0.971882	0.73837	0.209515	2.004015	0.01431

AdaBoost

AdaBoost fonctionne en agrégeant des classificateurs faibles. À chaque étape, il se concentre sur les exemples mal classés précédemment, ajustant ainsi le modèle pour mieux s'adapter et améliorer la prédiction globale. Cela le rend précis, adaptable et capable de bien gérer des ensembles de données complexes.

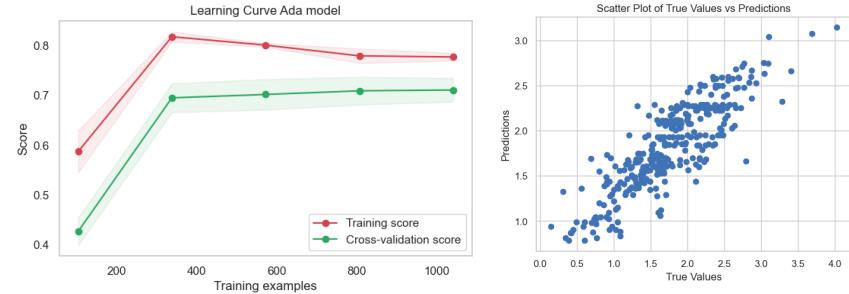
Modélisation du Site energy use



```
Ada_model = Pipeline([
    ('prepa', preprocessor),
    ('ada', AdaBoostRegressor(n_estimators=110, learning_rate=5, random_state=5, loss='exponential'))
])
```

RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test
0.253042	0.29066	0.785734	0.561991	0.225479	0.23128	0.015043

Modélisation du Co2

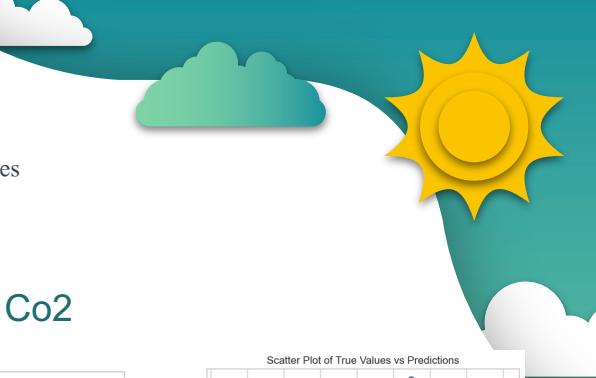


```
Ada_model = Pipeline([
    ('prepa', preprocessor),
    ('ada', AdaBoostRegressor(n_estimators=90, learning_rate=5, random_state=1, loss='exponential'))
])
```

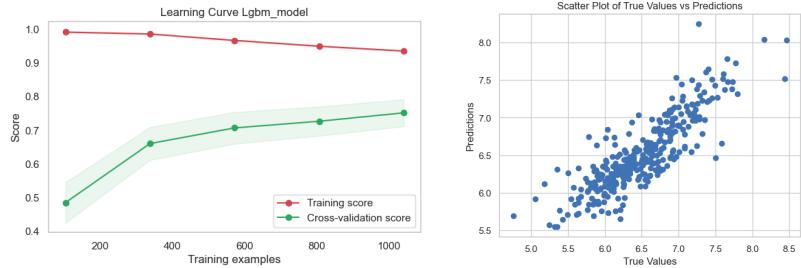
RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test
0.299337	0.335851	0.766636	0.631219	0.265693	0.266217	0.013885

Light GBM

Light gbm construit un modèle en combinant plusieurs petits arbres de décision, en se concentrant sur les caractéristiques les plus importantes. Cela le rend rapide et efficace pour résoudre des problèmes de classification et de régression



Modélisation du Site energy use



```
Lgbm_model = Pipeline([
    ('prepa', prep),
    ('Lgbm', lgb.LGBMRegressor(learning_rate=0.1, max_depth=-1,n_estimators=75,num_leaves=20 , min_child_samples=5))
])
```

	RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test
0	0.165336	0.278416	0.912386	0.63739	0.207454	0.104662	0.005586

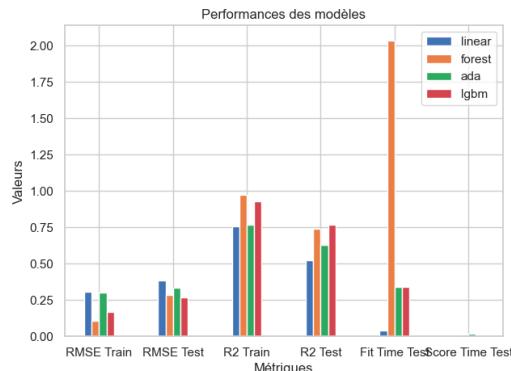
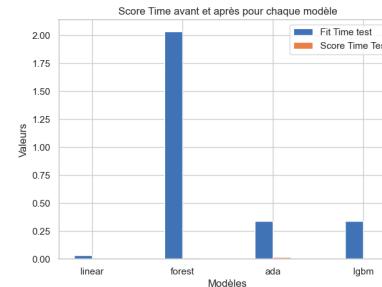
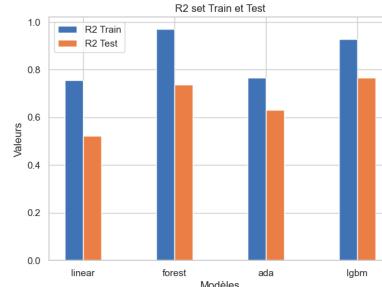
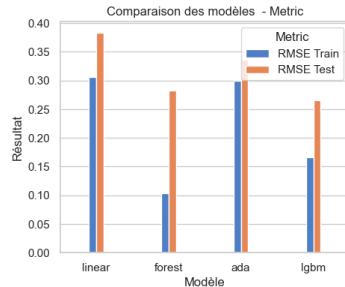
Modélisation du Co2



```
Lgbm_model = Pipeline([
    ('prepa', prep),
    ('Lgbm', lgb.LGBMRegressor(learning_rate=0.1, max_depth=-1,n_estimators=90,num_leaves=15 , min_child_samples=5))
])
```

	RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test
0	0.165784	0.265418	0.928393	0.767436	0.198189	0.344498	0.006068

Comparaison des résultats pour le Co2

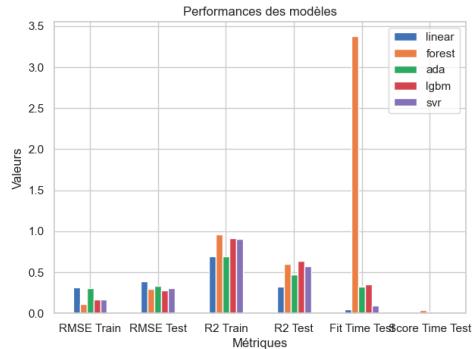
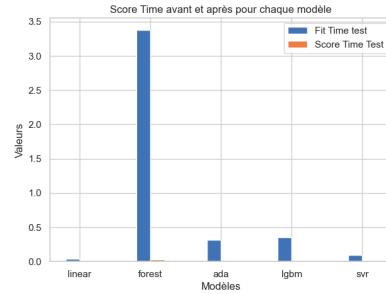
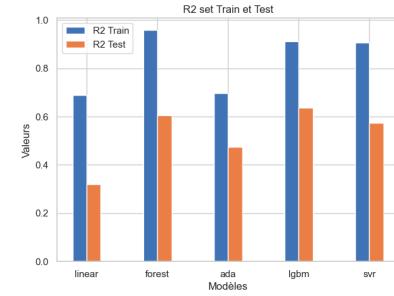
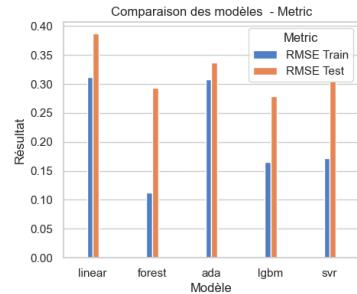


Le modèle Random Forest semble être robuste et bénéficie d'améliorations après le tuning. Il peut être un choix solide en raison de sa capacité à gérer des ensembles de données complexes par contre le temps d'ajustement est relativement long.

LightGBM maintient de bonnes performances après le tuning et offre une efficacité de calcul élevée. Contrairement au Random Forest le temps d'ajustement est beaucoup plus court.

	RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test
linear	0.305645	0.383392	0.756728	0.522040	0.288613	0.037120	0.004125
forest	0.103911	0.282254	0.971882	0.738370	0.209515	2.038013	0.014170
ada	0.299337	0.335851	0.766636	0.631219	0.265693	0.341167	0.015901
lgbm	0.165784	0.265418	0.928393	0.767436	0.198189	0.339385	0.007492

Comparaison des résultats pour Site energy use



Dans cette modélisation de la variable "site energy use", les modèles les plus robustes restent LightGBM et Random Forest. On remarque également que les performances métriques sont relativement inférieures par rapport à la modélisation du CO2.

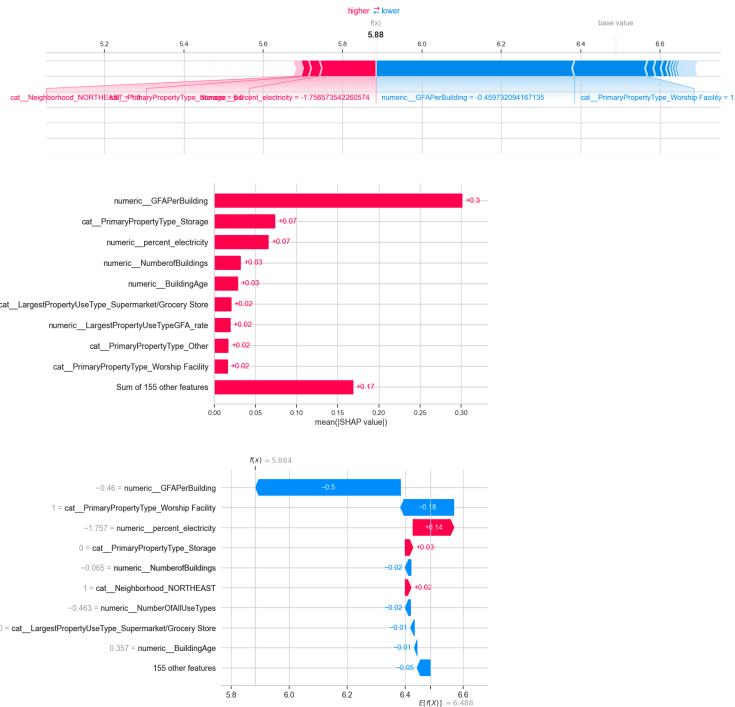
	RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test
linear	0.311236	0.387061	0.689761	0.320466	0.292915	0.043635	0.005572
forest	0.112018	0.292641	0.959801	0.605692	0.214334	3.381793	0.033884
ada	0.307935	0.337344	0.696344	0.474141	0.266999	0.320130	0.016135
lgbm	0.165336	0.278416	0.912386	0.637390	0.207454	0.350057	0.005528
svr	0.171470	0.304840	0.905776	0.573870	0.226225	0.095717	0.011258

A stylized illustration of a green landscape. It features a yellow sun with rays at the top left, a white cloud with a yellow center, and a large green tree on the right. In the foreground, there are several wind turbines with white blades and green bases, positioned between green buildings with blue windows. The background is a teal gradient.

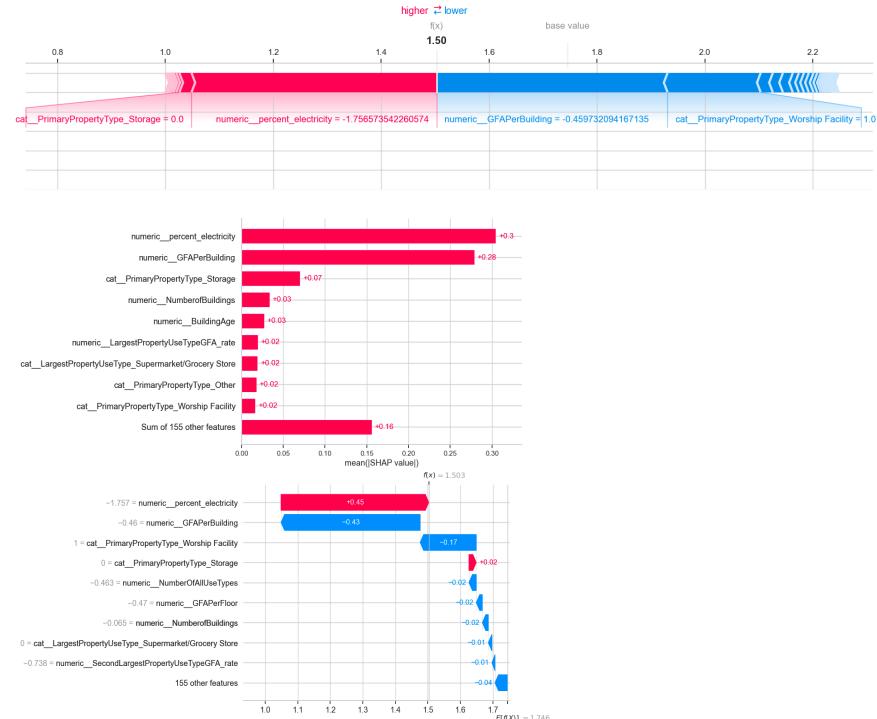
Importance des variables

Importance des caractéristiques

Modélisation Site Energy use



Modélisation du Co2





EnergyStarScore

Importance de la variable energyStars score

On teste l'importance de la variable "energyStar score" avec le modèle sélectionné. Étant donné que nous avons choisi le modèle LightGBM pour les deux modélisations, nous allons l'utiliser pour ces tests, avec les hyperparamètres sélectionnés pour chaque modèle.

Site energy use

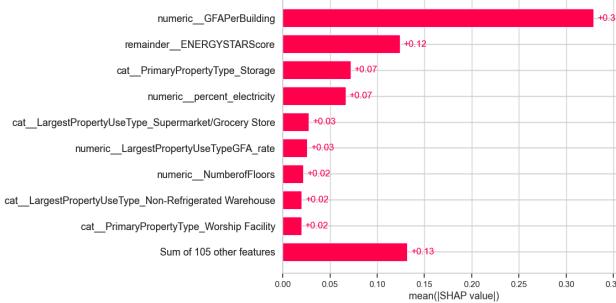
Test set avec
energyStar Score

	RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test
0	0.064685	0.195317	0.985998	0.802809	0.132484	0.318067	0.005716

Test sans
energyStar Score

	RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test
0	0.104996	0.25013	0.963043	0.671279	0.182585	0.340562	0.00511

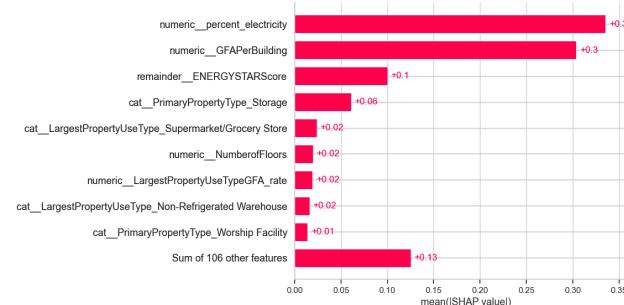
Importance des
variables



Modélisation du Co2

	RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test
0	0.062947	0.198906	0.989171	0.854824	0.140928	0.309671	0.005537

	RMSE Train	RMSE Test	R2 Train	R2 Test	MAE Test	Fit Time Test	Score Time Test
0	0.097318	0.241397	0.974067	0.78795	0.177463	0.326475	0.005408



Conclusion

Après une analyse approfondie des différents modèles et de leurs performances, nous avons opté pour l'utilisation du modèle LightGBM dans la modélisation des émissions de CO₂ et de la consommation d'énergie, en nous basant sur les critères suivants :

- Robustesse des performances.
- Efficacité de calcul.
- Capacité à gérer des ensembles de données complexes.

Le calcul de l'EnergyStar Score peut potentiellement présenter un intérêt dans la prédiction de l'utilisation total d'énergie. Cependant, en ce qui concerne la modélisation de la consommation de CO₂, l'utilité de cette variable demeure limitée.

