Prédiction de la consommation énergétique et des émissions de CO_2 des bâtiments non résidentiels à Seattle.

Bouzouita Hayette - Analyse exploratoire et Modélisation prédictive.
Février 2026

Sommaire

- Introduction
- Compréhension et préparation des données
- Analyse exploratoire des données (EDA)
- Méthodologie de modélisation
- Modélisation
- Comparaison des modèles
- Discussion & Conclusion

Introduction

Contexte énergétique et environnemental ,
Présentation du jeu de donnée (Seattle), Objectif et problématique du projet.

Contexte , Objectif & Problématique

🎯 La ville de Seattle souhaite atteindre son objectif de ville neutre en émission de carbone en 2050.

💰 Des relevés minutieux ont été effectués par les agents de la ville en 2016. Ces relevés sont coûteux à obtenir !

🤖 **J'interviens dans cette mission pour tenter de prédire les émissions de CO2 et la consommation totale d'énergie de bâtiments non destinés à l'habitation.**

🏗️ La prédiction doit se baser sur les données structurelles des bâtiments (taille et usage des bâtiments, date de construction, situation géographique, etc.)

❓ *Peut-on prédire la performance énergétique d'un bâtiment non résidentiel à partir de ses caractéristiques structurelles, et quel modèle offre les meilleures performances prédictives ?*

Jeu de données

- Jeu de données issu du site officiel de la ville de Seattle.
- Notre jeu de données contient, au départ, 3376 bâtiments décrits par 46 variables sur l'année 2016.

Variables d'identification & contexte

Variables concernant la structure du bâtiment

Variables sur l'énergie & émissions du bâtiment

Variables de localisation

Variables décrivant l'usage du bâtiment

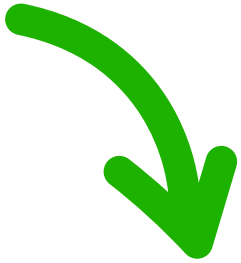
Variables concernant la qualité et la conformité des données

Compréhension et préparation des données

Description des variables, nettoyages des données, traitements des valeurs manquantes, détection et gestion des valeurs aberrantes, Feature engineering.

2 variables cibles (Targets)

L'objectif étant de prédire la consommation énergétique totale des bâtiments ainsi que leurs émissions de CO2. Les variables cibles retenus sont donc :



SiteEnergyUse (kBtu), représentant la consommation énergétique annuelle du bâtiment.

TotalGHGEmissions, représentant les émissions totales de gaz à effet de serre (CO2).

Suppression des autres colonnes énergétiques pour éviter tout Data Leakage

Définition du périmètre

Le projet concerne **uniquement les bâtiments non résidentiels** (bureaux, établissements scolaires, campus, bâtiments institutionnels, etc.)

Ainsi, les bâtiments à usages résidentiels (immeubles multi-familiaux) ont été exclus du périmètre d'étude.

BuildingType	
NonResidential	1460
Multifamily LR (1-4)	1018
Multifamily MR (5-9)	580
Multifamily HR (10+)	110
SPS-District K-12	98
Nonresidential COS	85
Campus	24
Nonresidential WA	1

Nettoyage

- **Traitements des valeurs manquantes** (Colonnes à + de 90% de valeurs manquantes on été supprimées, 2 bâtiments dont les valeurs sont manquantes pour les variables cibles ont été supprimées, Imputation etc.)
- **Suppression de colonnes non pertinentes** (*City*, *State*, *DataYear* etc.) n'apportant aucune information utile à l'analyse et la modélisation.
- **Choix éclairé entre les variables sémantiquement proches** (ex : *PrimaryPropertyType* vs *LargestPropertyUseType*)
- **Évaluation de la qualité des données**
- **Traitement des valeurs aberrantes** (incohérence physique, valeur négative, etc.)
- **Traitement des outliers** (un seuil basé sur le 99e percentile a été appliqué sur les variables cibles)

Toute suppression ou modification a été explicitement justifié dans le Notebook.

À ce stade, le jeu de données comprend 1614 bâtiments décrit par 21 variables

Création de variables

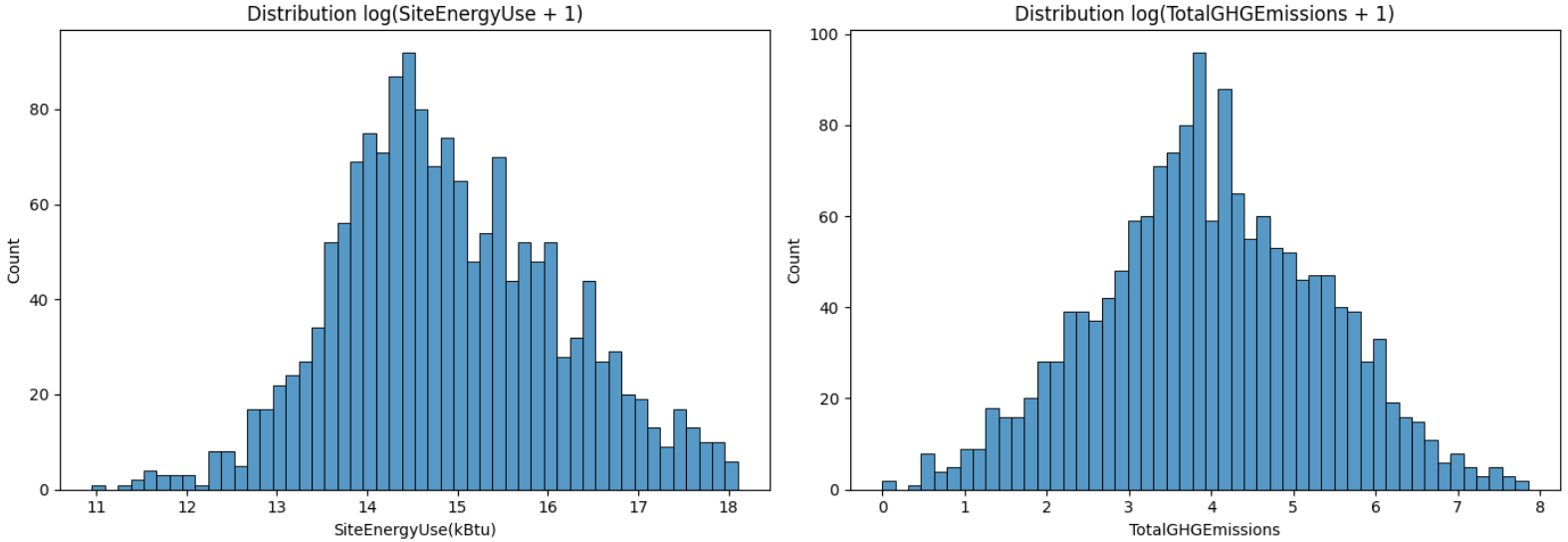
Création d'une variable binaire *has_secondary_use* indiquant si le bâtiment est mono-usage ou multi-usage.

Création d'une variable *BuildingAge* indiquant l'âge du bâtiment afin d'analyser l'influence de l'ancienneté des bâtiments.

Analyse exploratoire des données

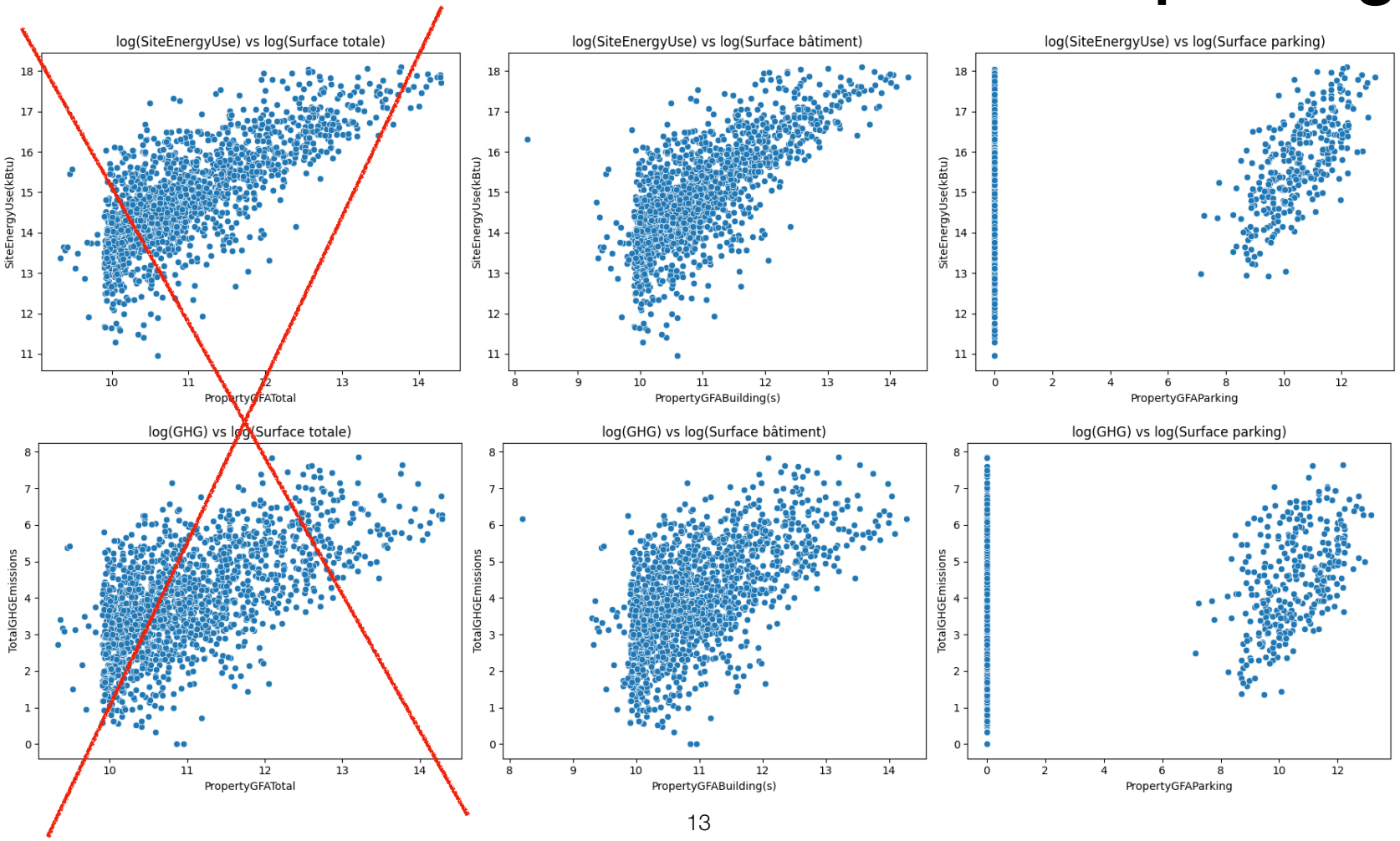
Analyse descriptive, Analyse des variables cibles, Analyse des corrélations, Analyse par type de bâtiment.

Distribution des variables cibles

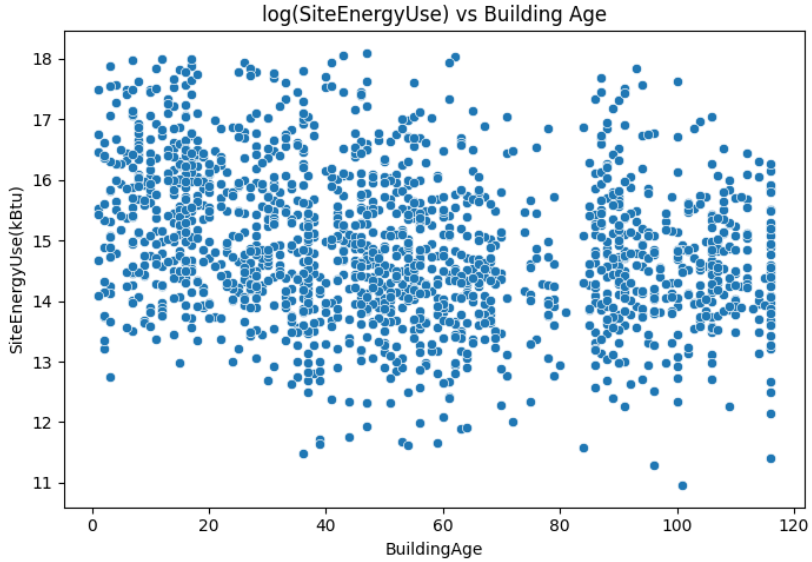


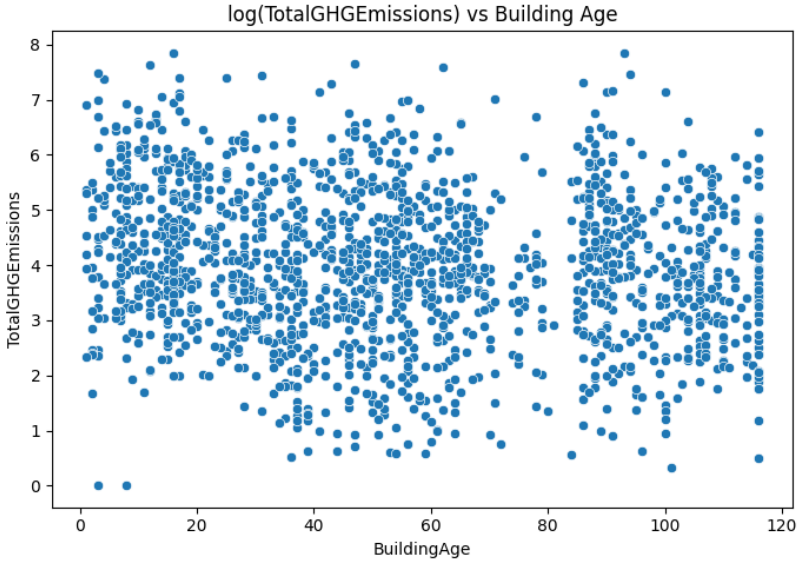
Une transformation logarithmique a été appliquée lors de la phase de modélisation, afin de réduire l'impact de l'asymétrie des variables cibles. En effet, une majorité de bâtiments présentent des niveaux de consommation et d'émissions relativement faibles tandis qu'une minorité de bâtiments est très énergivore.

Surface totale ou Surface bâtiment + parking ?

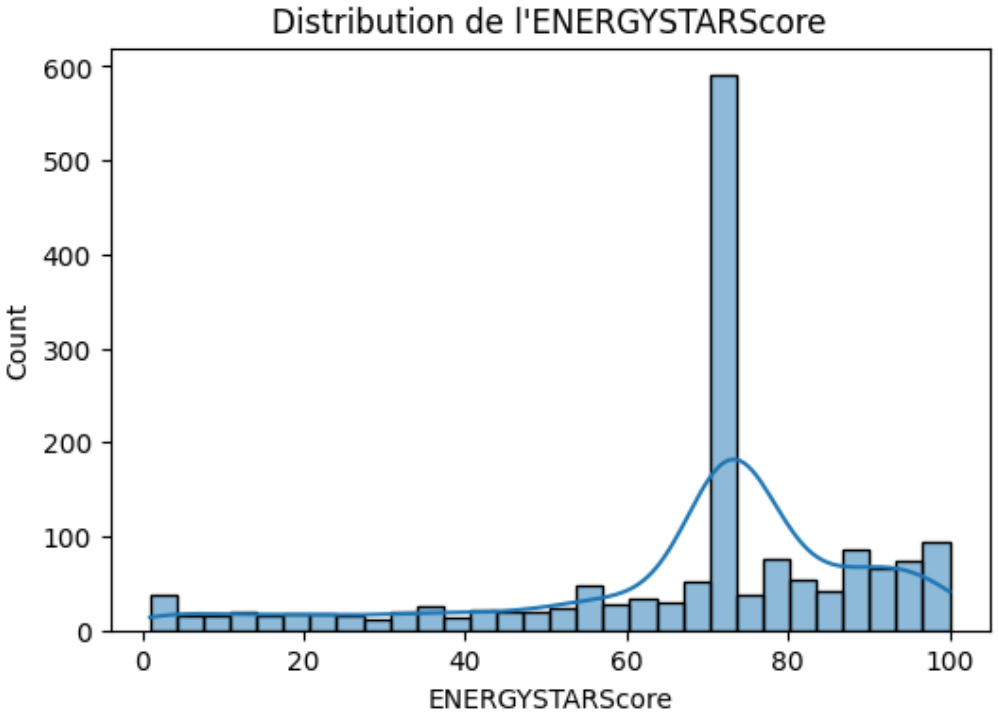


L’âge du bâtiment est-il un facteur explicatif ?





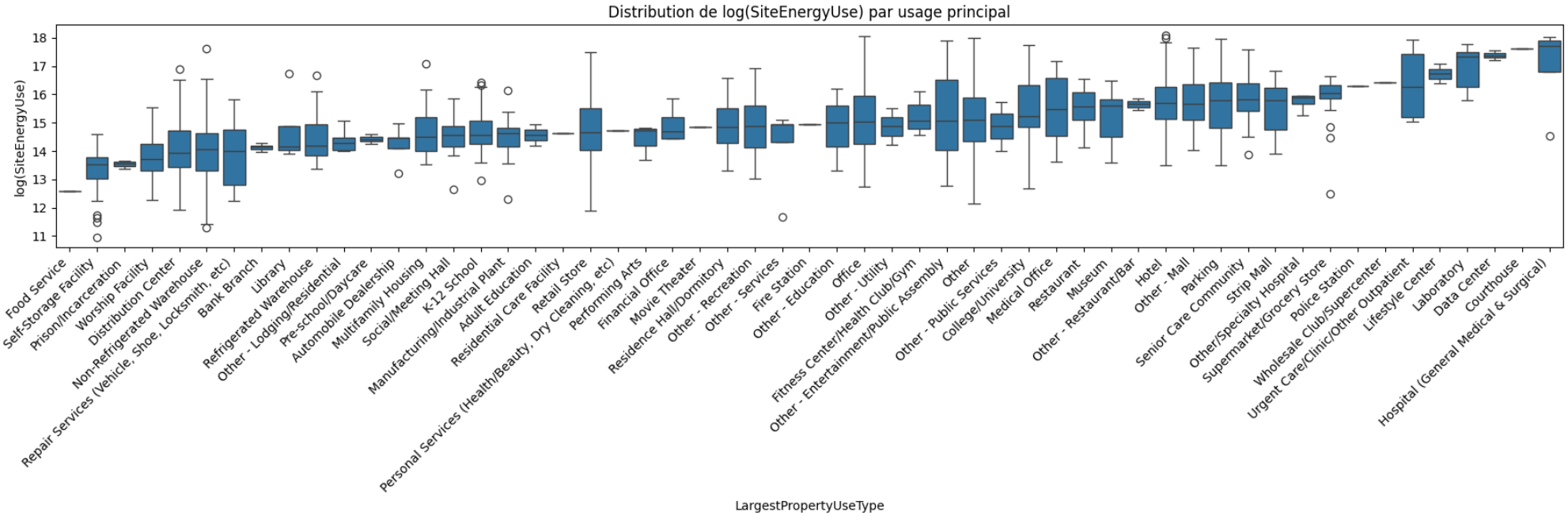
ENERGYSTARScore



⭐ La majorité des bâtiments de Seattle présentent un score surpérieur à 50, suggérant une performance énergétique globalement meilleure que la médiane nationale.

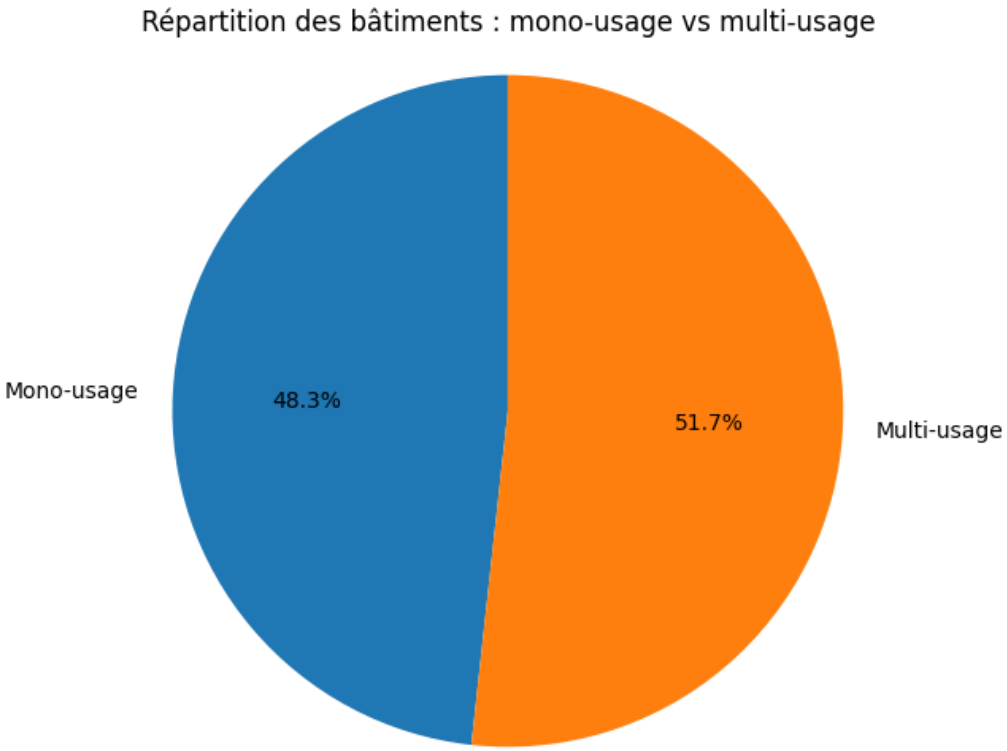
Cette variable a été utilisée uniquement pour l'EDA.

LargestPropertyUseType

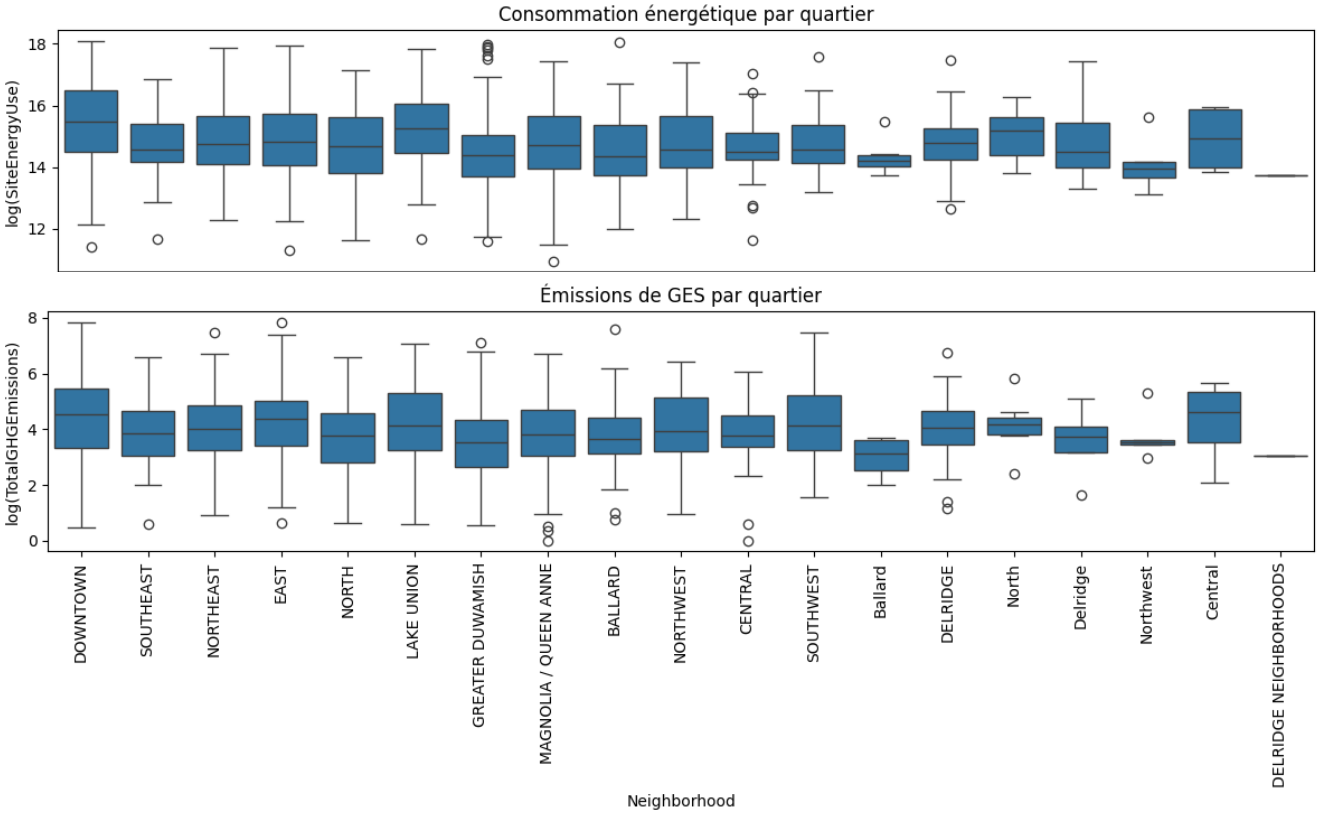


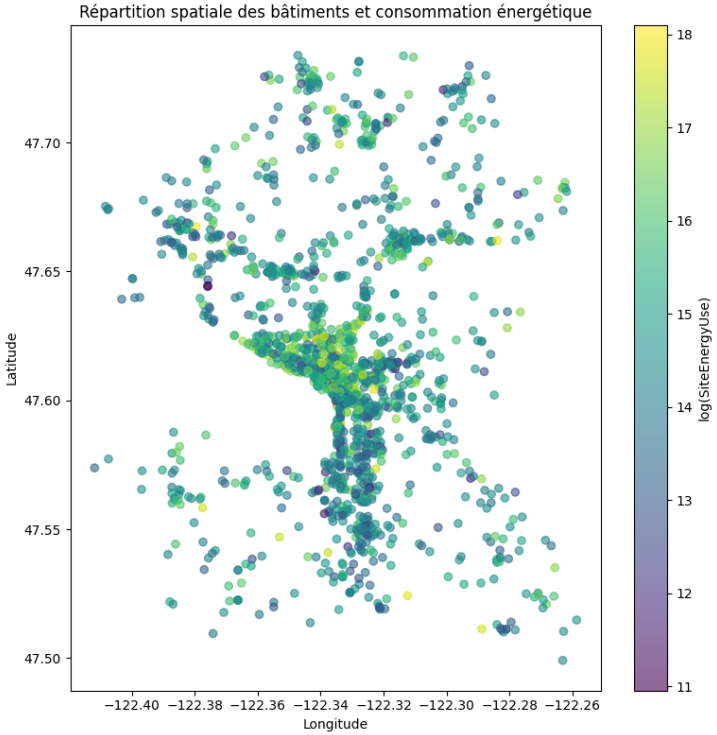
L'usage principal du bâtiment influence la distribution de la consommation énergétique. Cependant, la dispersion intra-usage reste importante. Nous tirons la même conclusion pour l'analyse des émissions de CO2.

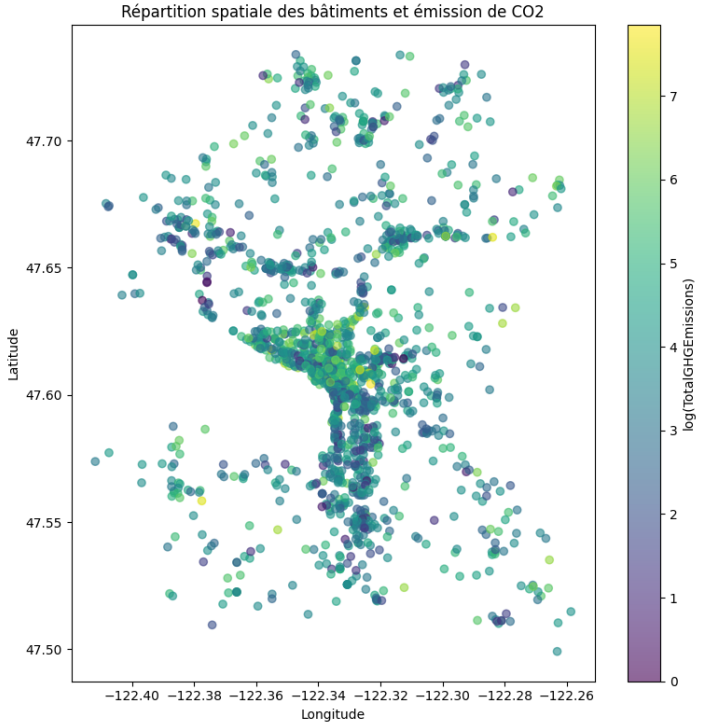
La diversité des usages au sein d’un même bâtiment est une caractéristique fréquente du parc étudié.



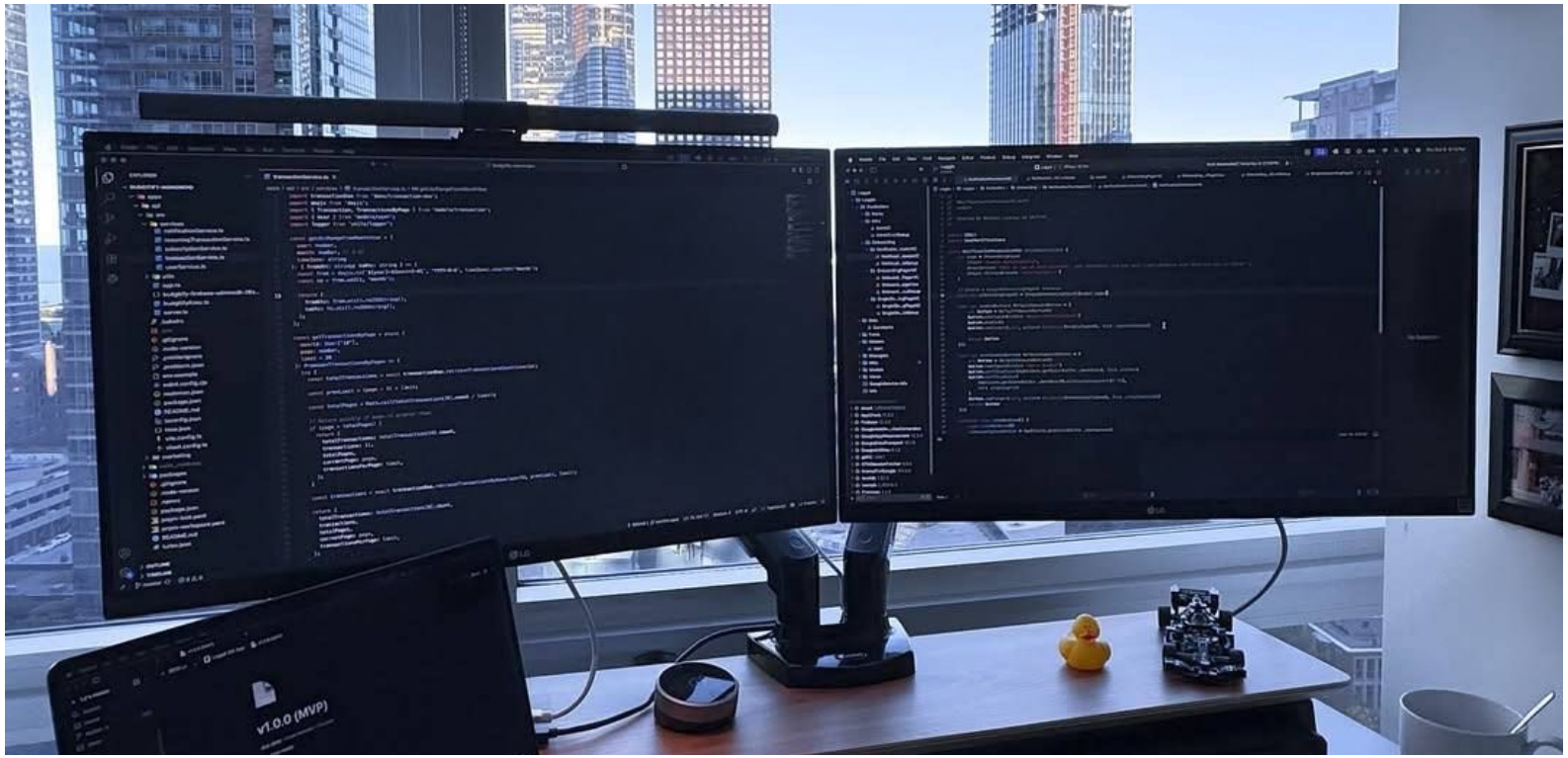
La localisation du bâtiment est-il un facteur explicatif ?







Les représentations spatiales mettent en évidence une forte concentration des bâtiments dans certaines zones, notamment au centre de la ville. Cependant pour une même zone géographique, on observe une grande variabilité des niveaux de consommation énergétique et d'émissions de CO2



Méthodologie de modélisation

Choix des variables explicatives, encodages des variables catégorielles, transformation des variables (log, normalisation), stratégie de validation croisée, métriques d’évaluation.

8 variables explicatives

```
categorical_features = ["Neighborhood", "LargestPropertyUseType"]

numeric_features = [
    "NumberofBuildings",
    "PropertyGFAParking",
    "PropertyGFABuilding(s)",
    "LargestPropertyUseTypeGFA",
    "has_secondary_use",
    "BuildingAge"
]
```

Diversité des variables (contexte, localisation, structure du bâtiment, usage du bâtiment).

Méthodologie

Encodage pour les variables catégorielles.
Normalisation des variables numériques lorsque le modèle en a besoin.

Une **fonction générique d'évaluation** a été créée. Elle repose sur une stratégie de **validation croisée** afin d'estimer mes performances de manière plus robustes qu'un simple découpage train/test.

```
preprocessor_lr = ColumnTransformer(
    transformers=[
        ("cat", OneHotEncoder(handle_unknown="ignore"), categorical_features),
        ("num", StandardScaler(), numeric_features),
    ]
)

lr_pipeline = Pipeline([
    ("preprocessor", preprocessor_lr),
    ("model", LinearRegression())
])

# Évaluation energy
lr_results_energy = evaluate_model_cv(lr_pipeline, X, y_energy)
print("Résultat pour l'énergie : ")
print(lr_results_energy)

# Evaluation CO2
lr_results_ghg = evaluate_model_cv(lr_pipeline, X, y_ghg)
print("Résultat pour le CO2 : ")
print(lr_results_ghg)
```

Pipeline : preprocessor + appel du modèle (avec ou sans réglages des hyper-paramètres)

3 métriques d’évaluation

R2

Coefficient de détermination

Mesure la proportion de variance expliquée par le modèle.

R^2 = 1 -> modèle « parfait »

R^2 = 0 -> modèle n’explique rien

MAE

Mean Absolute Error

Mesure l’erreur moyenne absolue entre les valeurs réelles et prédites.

En unité réelle.

RMSE

Root Mean Squared Error

Mesure la racine de la moyenne des erreurs au carré.

Pénalise davantage les grosses erreurs.

Modélisation

Régression Linéaire Multivariée, Random Forest, Gradient Boosting, SVM

4 modèles testés

Baseline Model : Régression Linéaire Multivariée

```
lr_pipeline = Pipeline([
    ("preprocessor", preprocessor_lr),
    ("model", LinearRegression())
])
```

Modèle Non-linéaire : Random Forest

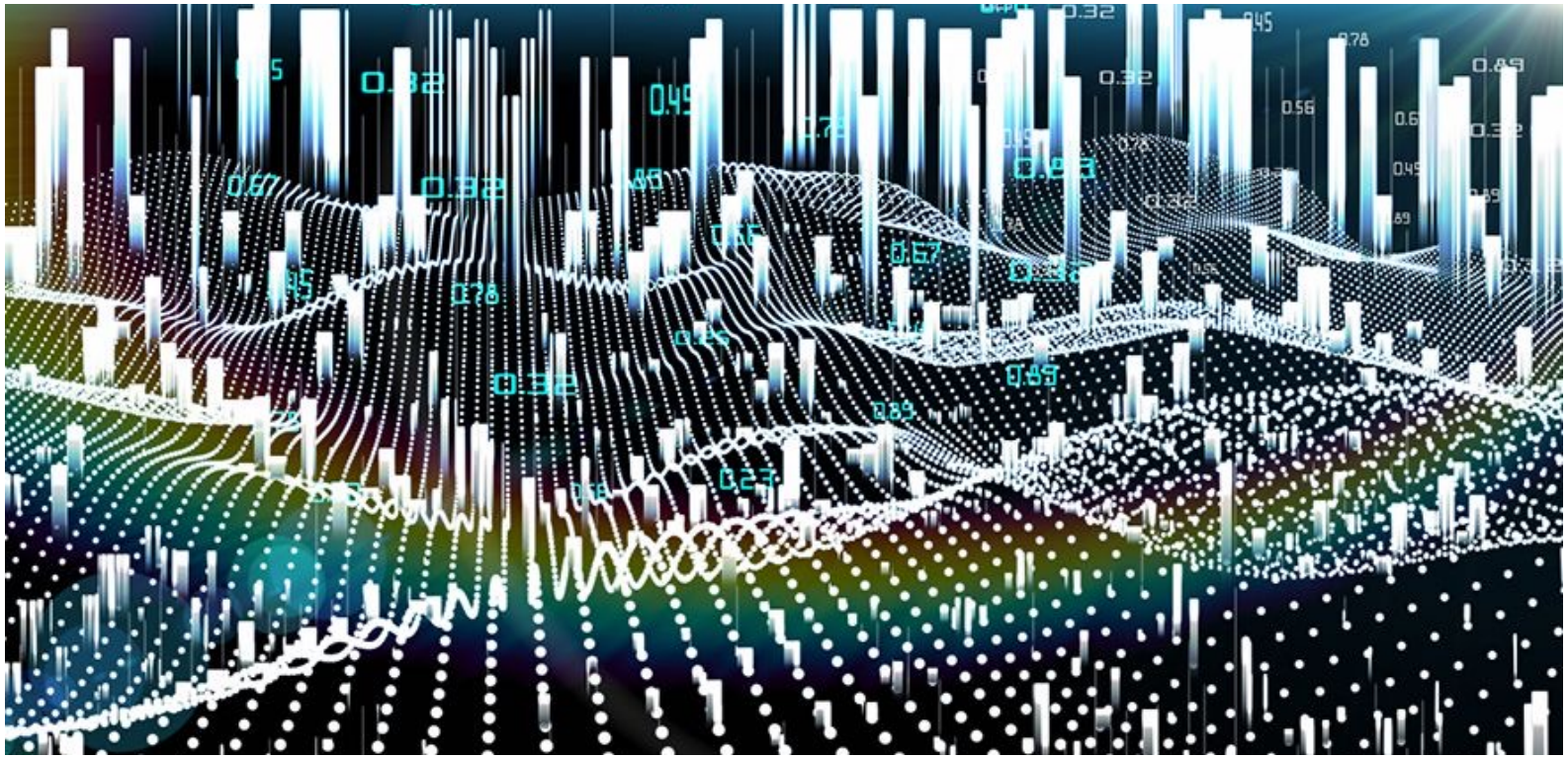
```
rf_pipeline = Pipeline([
    ("preprocessor", preprocessor),
    ("model", RandomForestRegressor(
        n_estimators=300,
        random_state=42,
        n_jobs=-1
    ))
])
```

Modèle avancé : Gradient Boosting

```
gb_pipeline = Pipeline([
    ("preprocessor", preprocessor_gb),
    ("model", GradientBoostingRegressor(
        n_estimators=300,
        learning_rate=0.05,
        max_depth=3,
        random_state=42
    ))
])
```

Modèle alternatif testé: SVM

```
svm_pipeline = Pipeline([
    ("preprocessor", preprocessor_svm),
    ("model", SVR(
        kernel="rbf",
        C=10,
        epsilon=0.1
    ))
])
```



Comparaison des modèles

Résultats des performances, Optimisation du modèle retenu

Comparaison des résultats de performance

Énergie :

	R2_log_mean	R2_log_std	MAE_real_mean	RMSE_real_mean	training_time_sec
LinearRegression	0.545241	0.039186	7.817847e+06	6.560826e+07	0.119999
RandomForest	0.686387	0.036281	2.671516e+06	5.968868e+06	5.948424
GradientBoosting	0.708018	0.033466	2.611703e+06	5.902730e+06	4.093410
SVM	0.672454	0.021603	3.017177e+06	6.827361e+06	1.000795

CO2 :

	R2_log_mean	R2_log_std	MAE_real_mean	RMSE_real_mean	training_time_sec
LinearRegression	0.388498	0.032651	137.420478	809.596554	0.143080
RandomForest	0.474291	0.044017	78.709983	183.673141	6.132335
GradientBoosting	0.507459	0.042804	76.663234	175.275049	3.994004
SVM	0.461780	0.020747	84.152476	191.297995	0.912006

Optimisation de Gradient Boosting

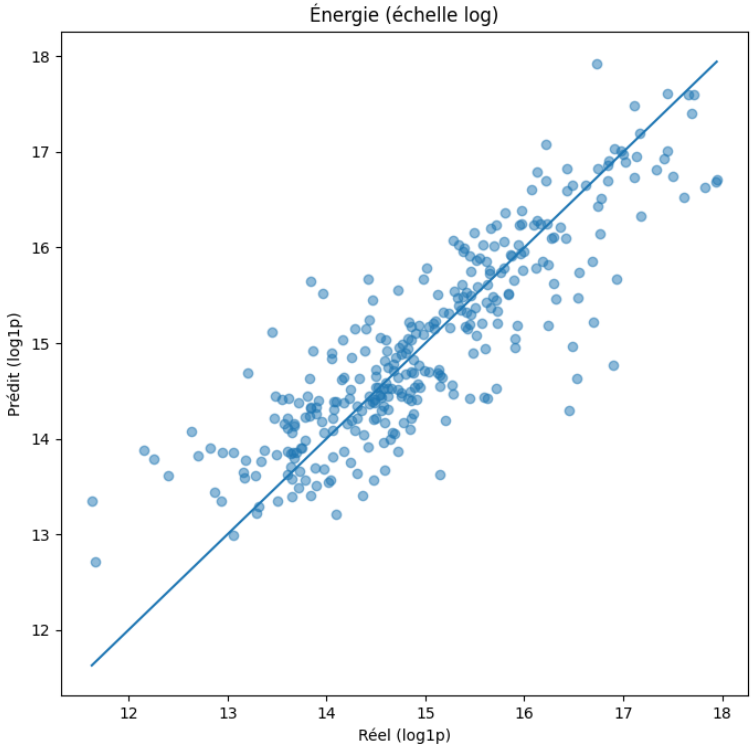
```python
# Grille d'hyperparamètres
param_grid_gb = {
    "model__n_estimators": [200, 300, 500],
    "model__learning_rate": [0.03, 0.05, 0.1],
    "model__max_depth": [1, 2, 3],
    "model__min_samples_leaf": [1, 3],
    "model__subsample": [0.7, 0.8]
}

# GridSearchCV
grid_gb = GridSearchCV(
    estimator=gb_pipeline,
    param_grid=param_grid_gb,
    scoring="r2",          # optimisation en log via y_train_log
    cv=cv,
    n_jobs=-1,
    verbose=2,
    refit=True
)

grid_gb.fit(X_train, y_train_log)
```

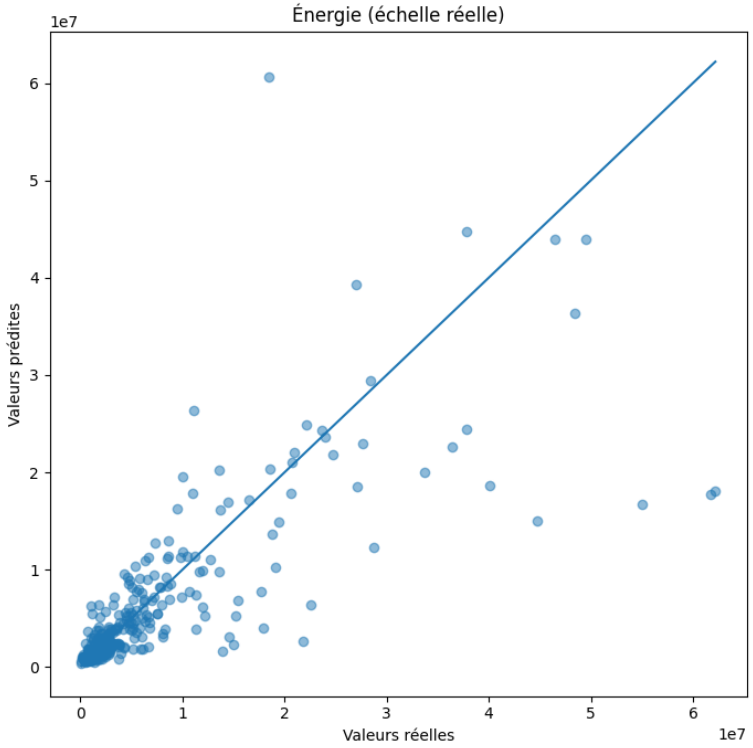
Résultats après optimisation

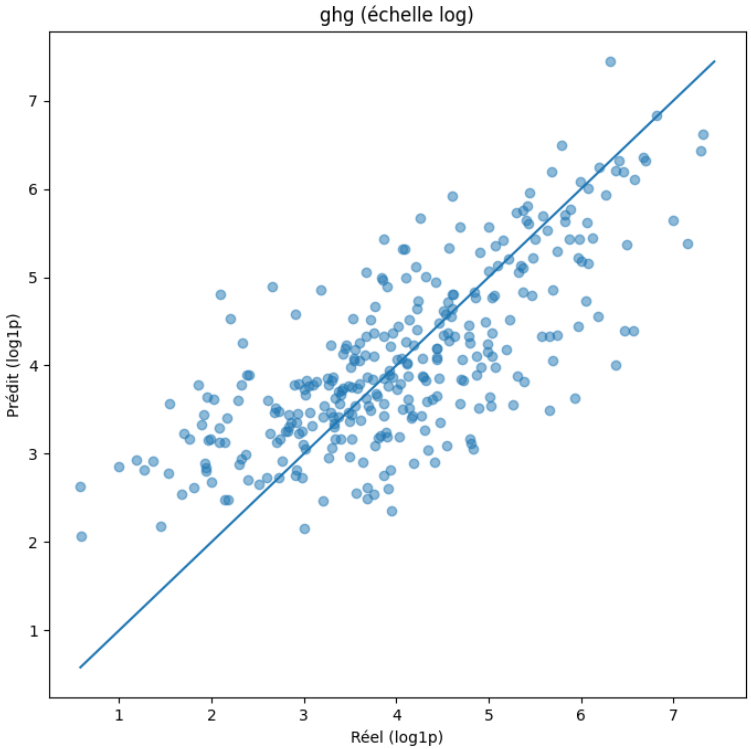
```
Fitting 5 folds for each of 108 candidates, totalling 540 fits
 Best CV R² (log): 0.7102370489427285
 Best params: {'model__learning_rate': 0.1, 'model__max_depth': 2, 'model__min_samples_leaf': 3, 'model__n_estimators': 200, 'model__subsample': 0.8}

=== Final Test Metrics (REAL scale) pour l'énergie ===
R² log  : 0.739
R² réel  : 0.587
RMSE réel: 6310232.29
MAE réel : 2671917.56
```

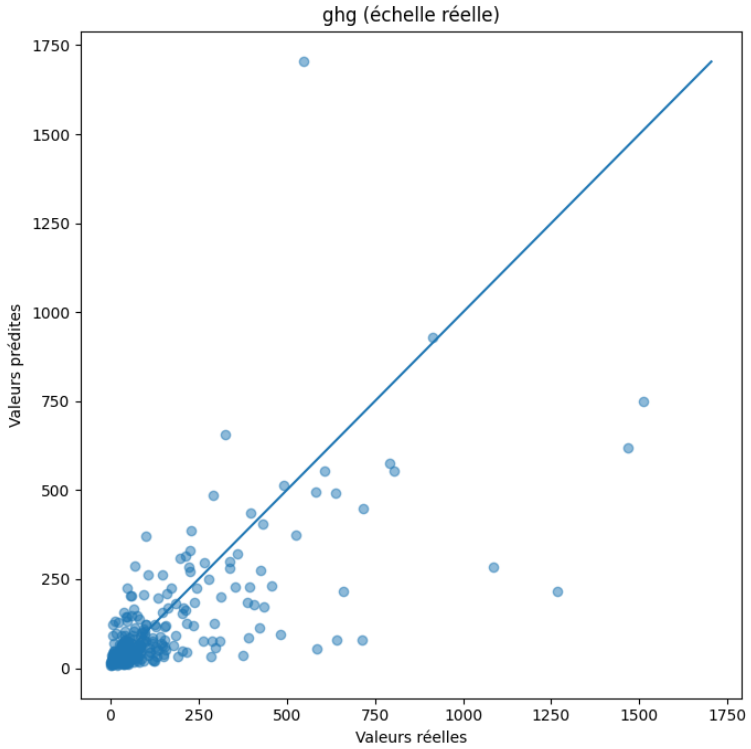
```
Fitting 5 folds for each of 108 candidates, totalling 540 fits
 Best CV R² (log): 0.5141917086637996
 Best params: {'model__learning_rate': 0.05, 'model__max_depth': 2, 'model__min_samples_leaf': 3, 'model__n_estimators': 300, 'model__subsample': 0.7}

=== Final Test Metrics (REAL scale) pour le CO2 ===
R² log  : 0.539
R² réel  : 0.428
RMSE réel: 155.83
MAE réel : 71.12
```

Valeur prédite vs Valeur réelle






ghg (échelle log)
Prédit (log1p)
Réel (log1p)
1
2
3
4
5
6
7

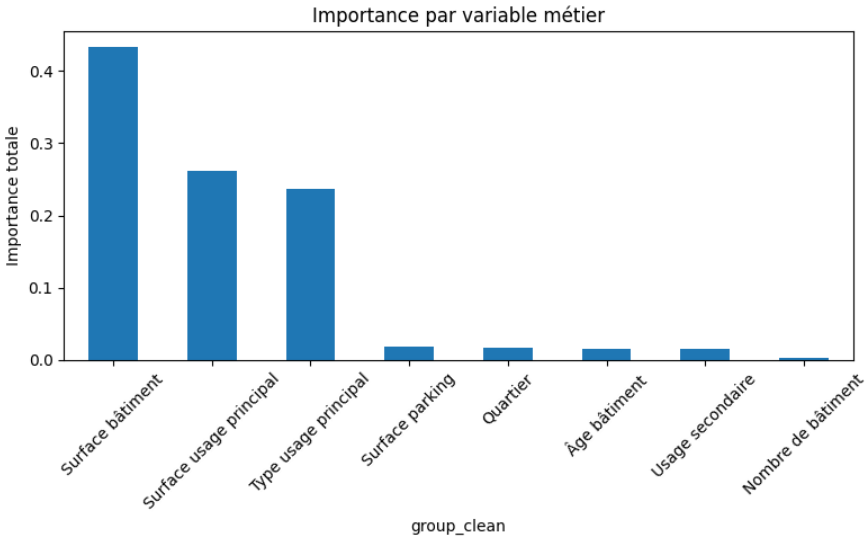

ghg (échelle réelle)
Valeurs prédites
Valeurs réelles
0
250
500
750
1000
1250
1500
1750

Discussion & Conclusion

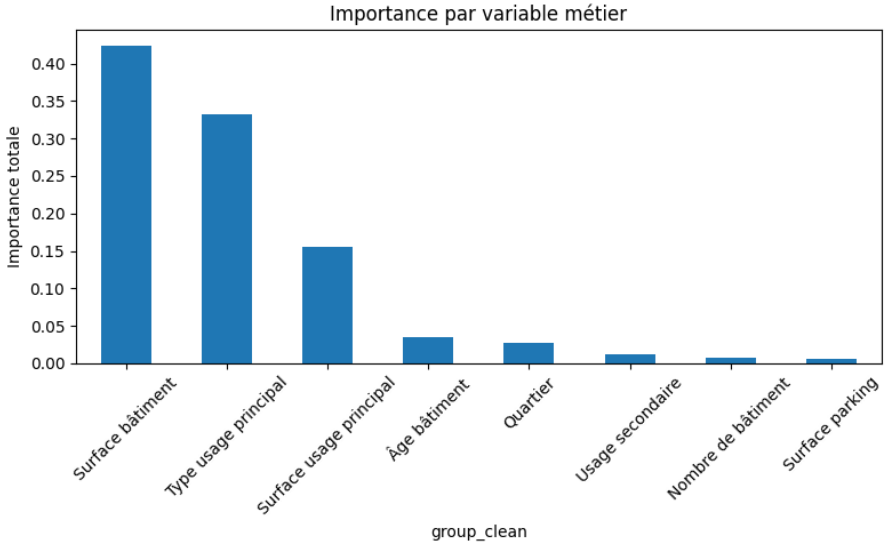
Feature importance, Limite du projet,
Pistes d’amélioration et suggestions

Feature Importance

Énergie :



CO2 :



La problématique de ce projet était de **déterminer s'il est possible de prédire la performance énergétique d'un bâtiment non résidentiel à partir de ses caractéristiques structurelles**, et d'identifier le modèle offrant les meilleures performances prédictives.

Les résultats obtenus montrent qu'**une telle prédiction est effectivement possible**. Les variables structurelles et fonctionnelles **permettent d'expliquer environ 59 % de la variance de la consommation énergétique et 43 % de celle des émissions de CO_2**, ce qui traduit une capacité prédictive significative malgré la complexité du phénomène étudié.

Parmi les modèles testés, **les approches non linéaires**, et en particulier le Gradient Boosting, ont offert les meilleures performances, confirmant l'existence de relations complexes entre les caractéristiques des bâtiments et leur performance énergétique.

Toutefois, la prédiction des émissions de CO_2 s'est révélée plus difficile, suggérant que certains facteurs déterminants, notamment liés au mix énergétique ou aux équipements techniques, ne sont pas entièrement capturés par les données structurelles disponibles.

Ainsi, ce travail démontre que **les caractéristiques d'un bâtiment constituent une base pertinente pour anticiper sa performance énergétique**, tout en soulignant l'importance d'**un enrichissement des données pour améliorer davantage la précision des modèles.**