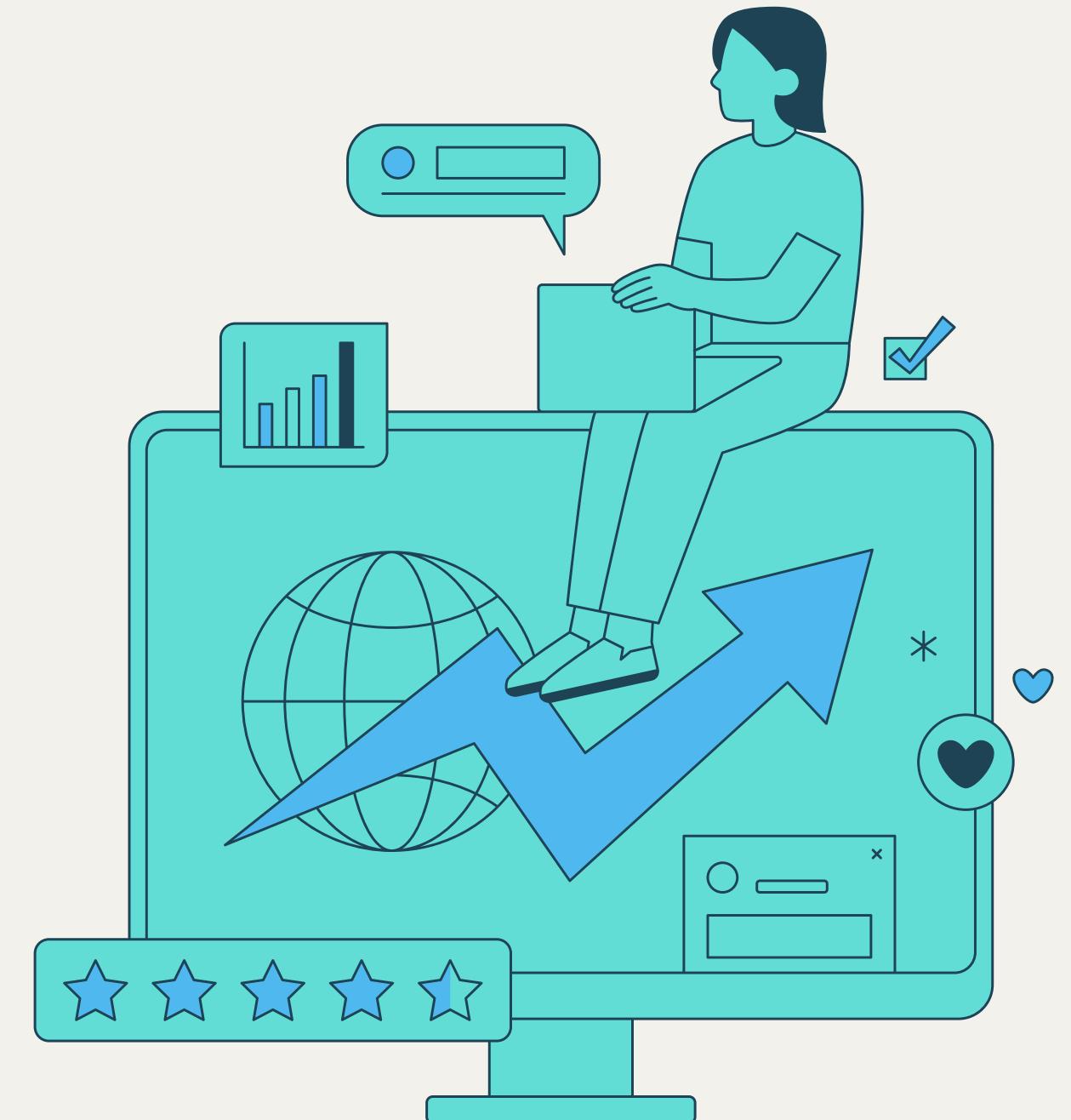


Anticipez les besoins en consommation de bâtiments

La Ville de Seattle

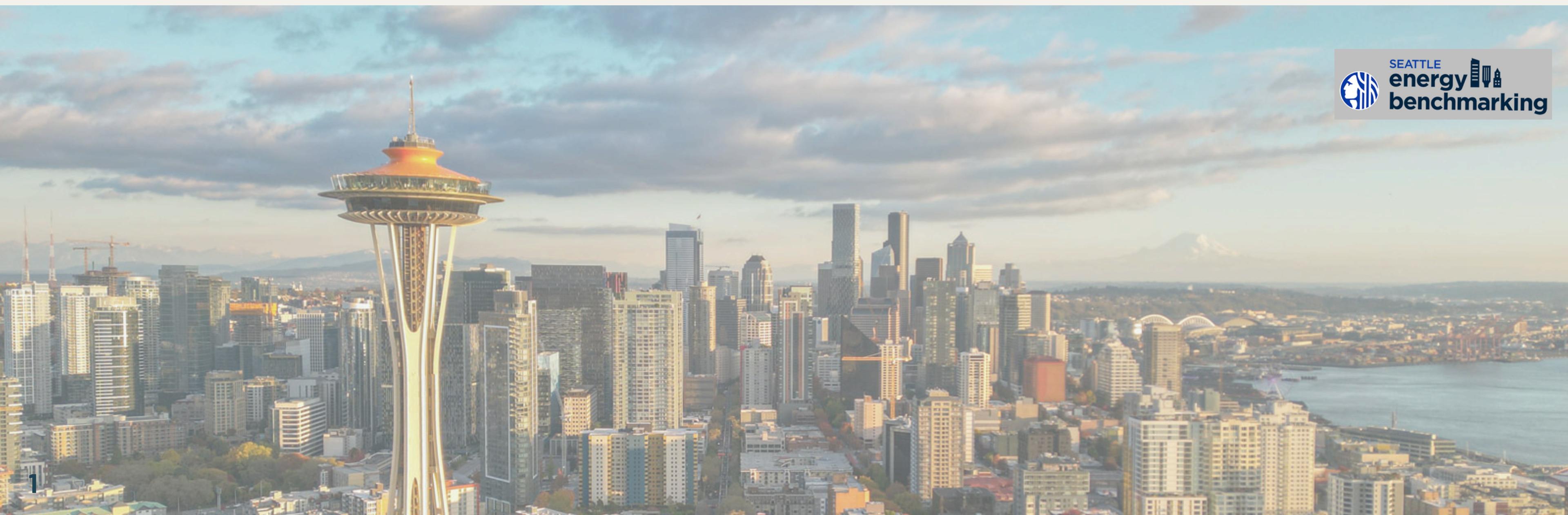
Mars 2025

Natascha Minnitt

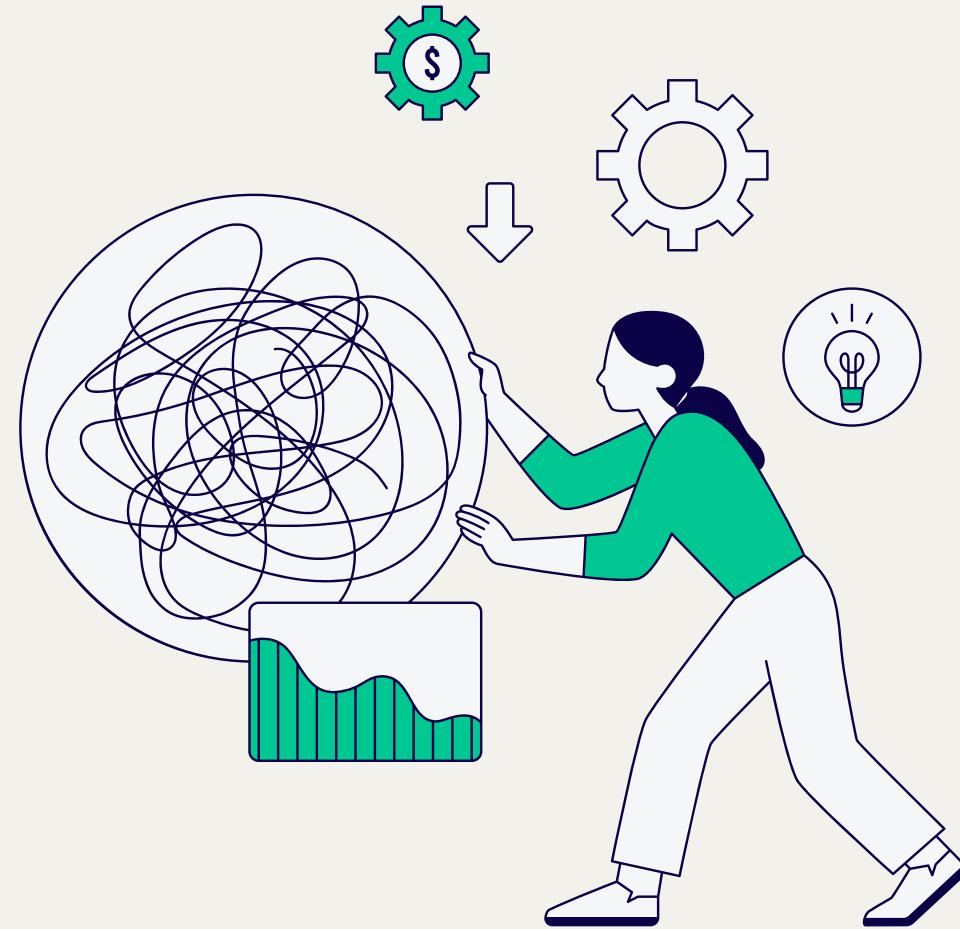


La mission

100% d'énergie propre dans nos bâtiments d'ici 2050



Traitement des données et feature engineering

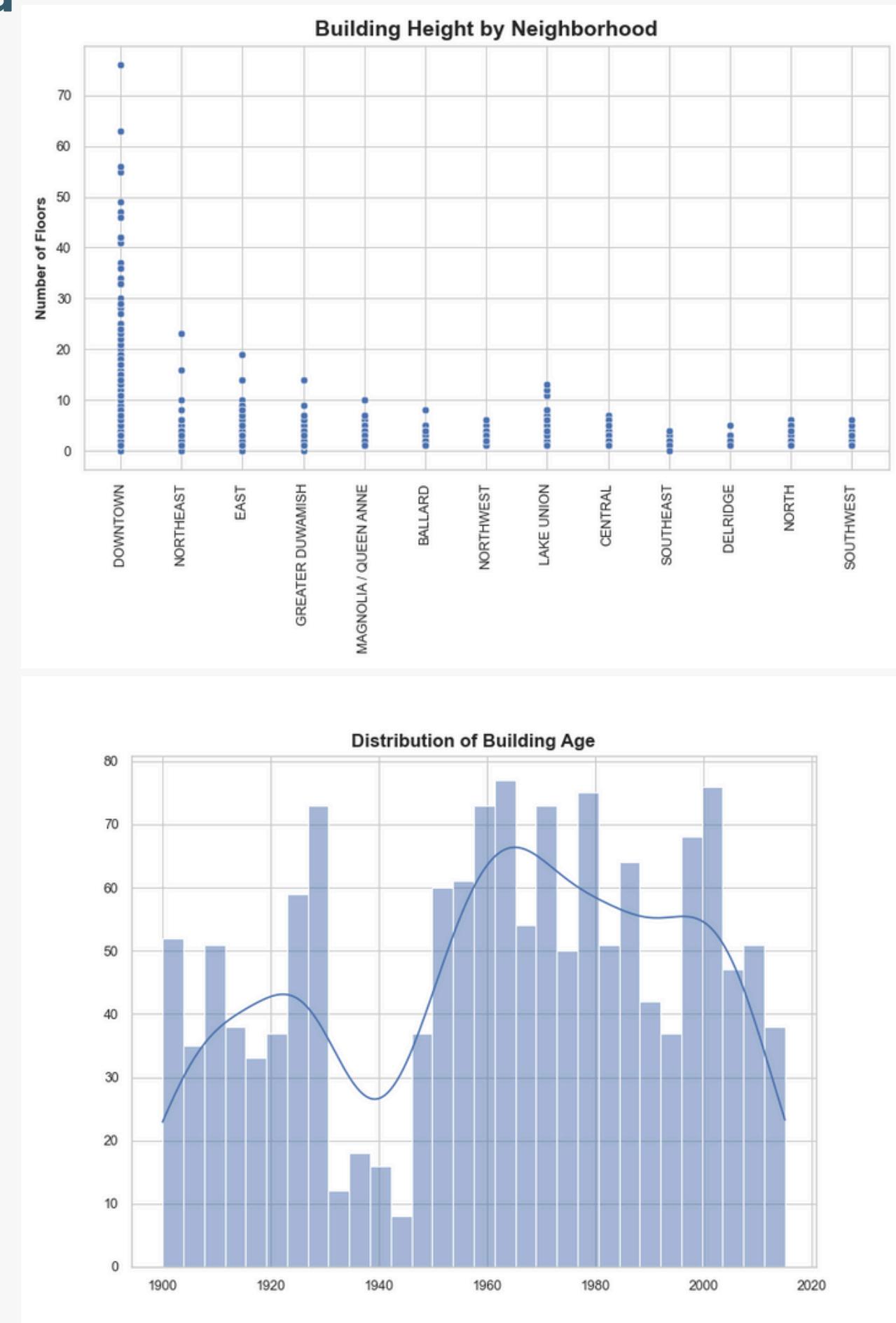
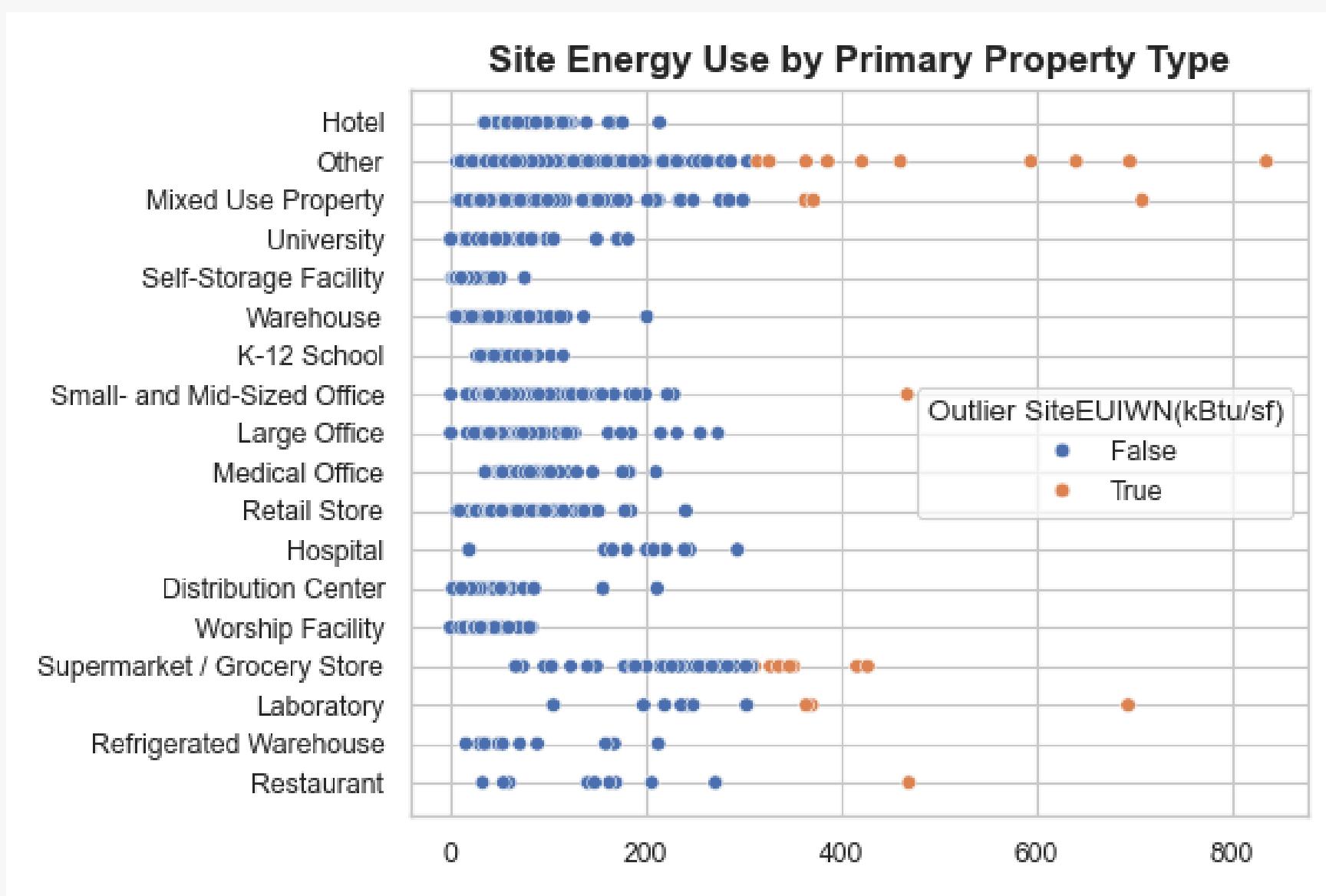


Traitement des données

Seattle's 2016 Building Energy Benchmarking Data

1. Filtrage et exploration des données

- Suppression des bâtiments résidentiels
- Suppression des outliers et bâtiments non-conformes

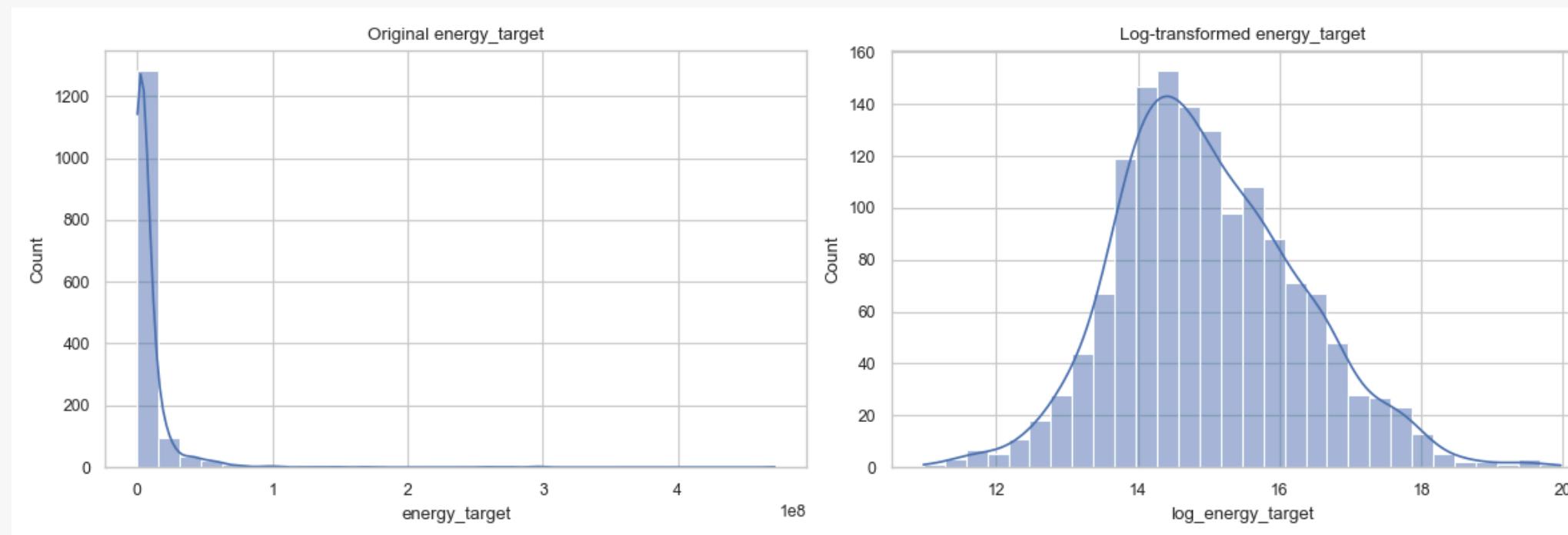


Seattle's 2016 Building Energy Benchmarking Data

2. Sélection et transformation des cibles (targets)

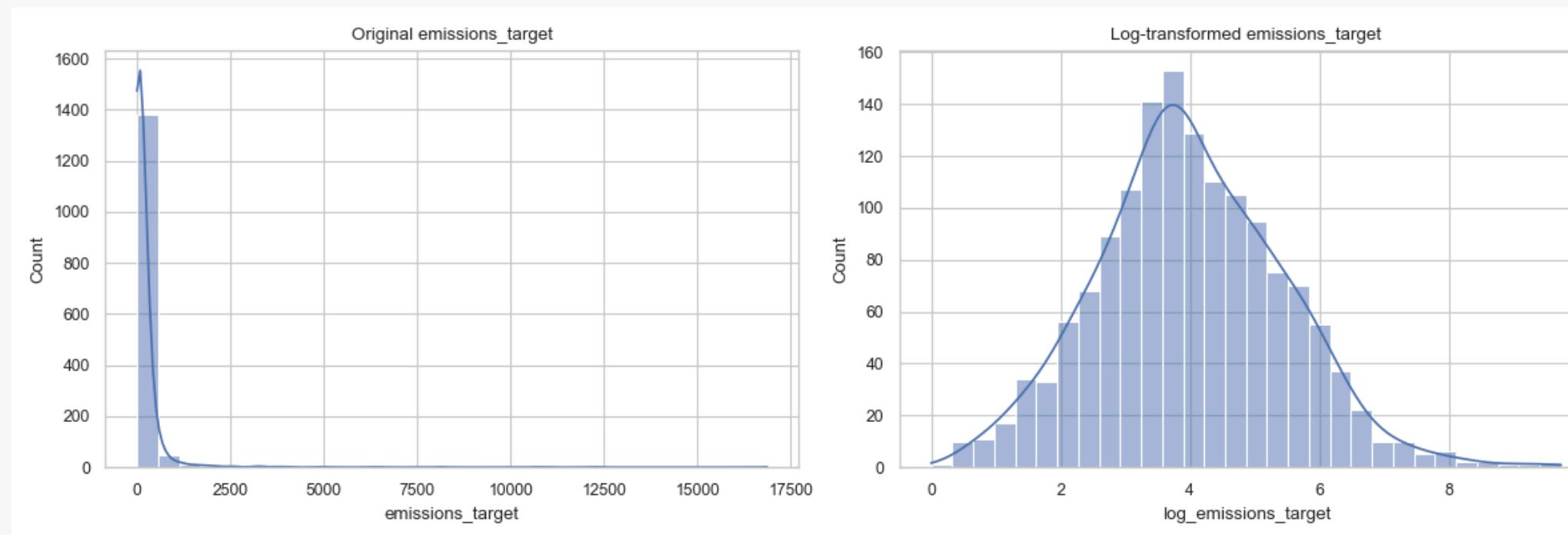
[1] SiteEnergyUseWN(kBtu) *Transformation logarithmique*

[2] TotalGHGEmissions *Transformation logarithmique*



Asymétrie

Avant	Après
10.77	0.38

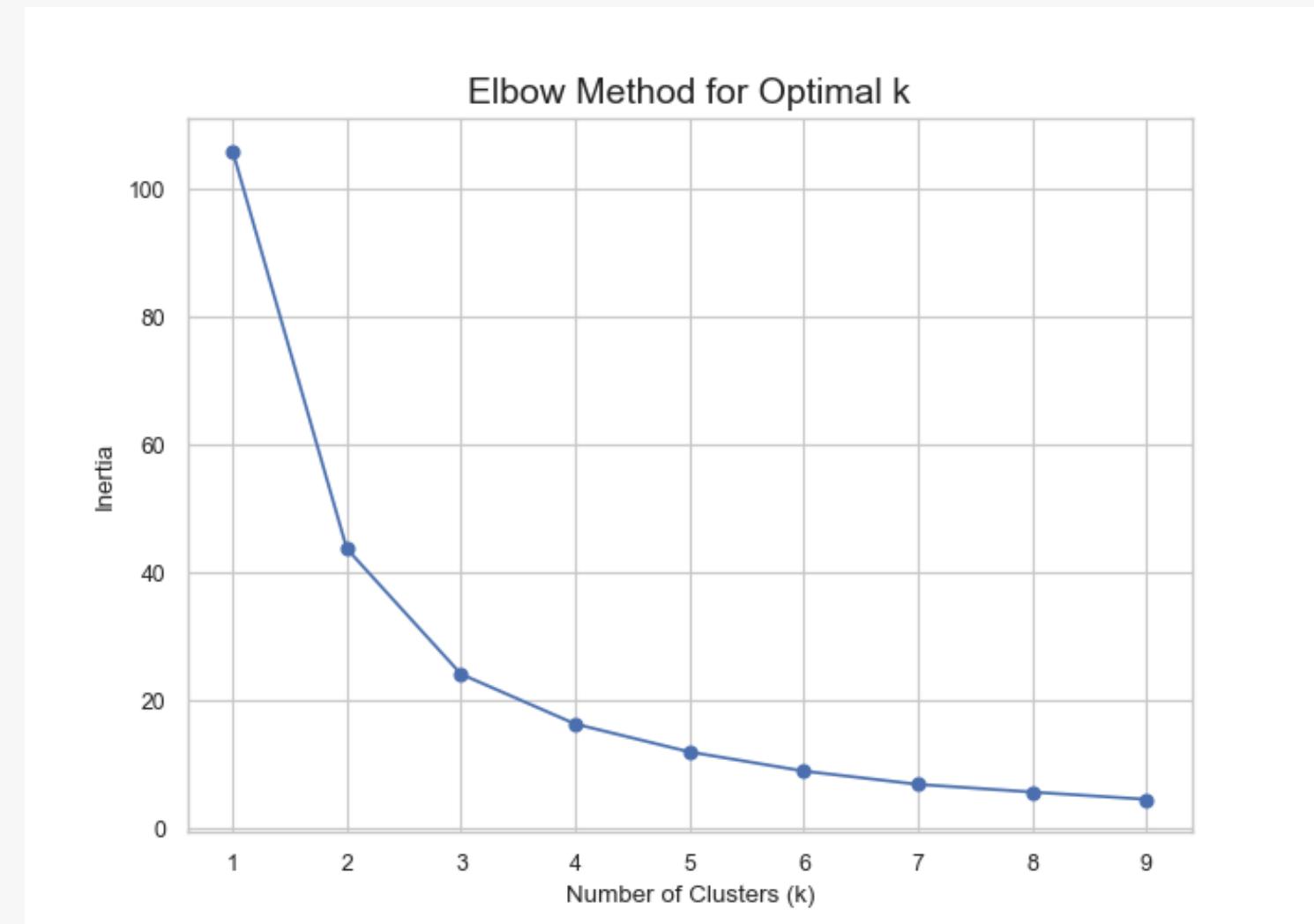
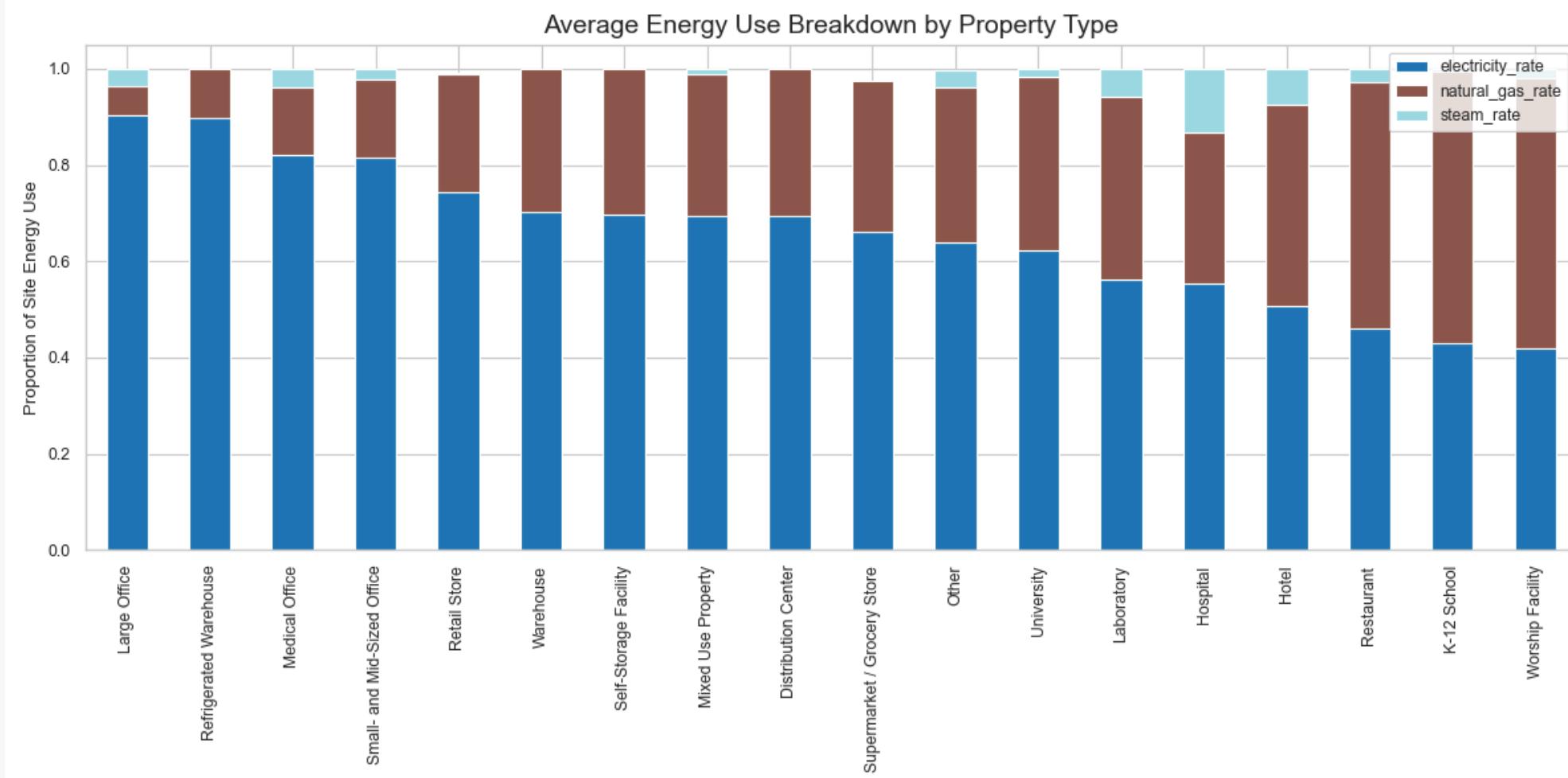


Avant	Après
14.38	0.23

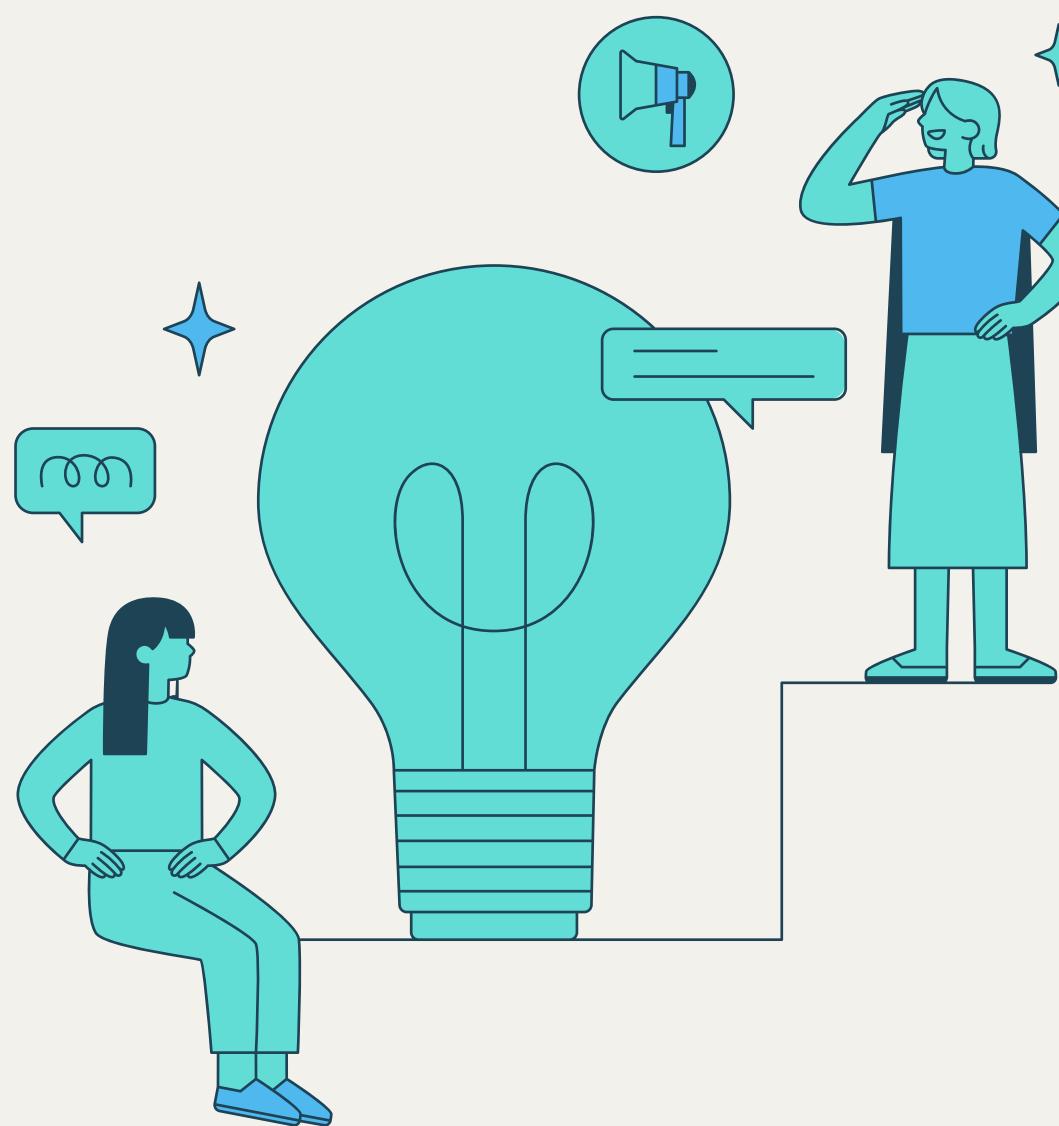
Feature engineering

3. Sélection des features

[1] YearBuilt	
[2] Number of floors	Suprême > 78
[3] Property GFA Total	Transformation logarithmique
[4] Parking Proportion	PropertyGFAParking / PropertyGFA Total
[5] Use rate	Électricité, gaz, vapeur / SiteEnergyUse(kBtu)
[6] Use counts	Compte de ListOfAllPropertyUseTypes
[7] Largest Property Use Type	K-means clustering(5 clusters)
[8] Primary Property Type	OneHotEncoder(18 catégories)
[9] Neighbourhood	OneHotEncoder(13 catégories)
[10] ENERGYSSTARScore	



L'approche de modélisation et présentation des résultats



L'approche de modélisation

Préparation des données :

- Division Train-Test : 70 % d'entraînement / 30 % de test (random_state = 42)

Construction du pipeline :

- Standard Scaler pour la normalisation des variables
- Intégration d'une grille d'hyperparamètres pour l'optimisation du modèle

Ajustement du modèle avec GridSearchCV :

- Validation croisée à 5 blocs

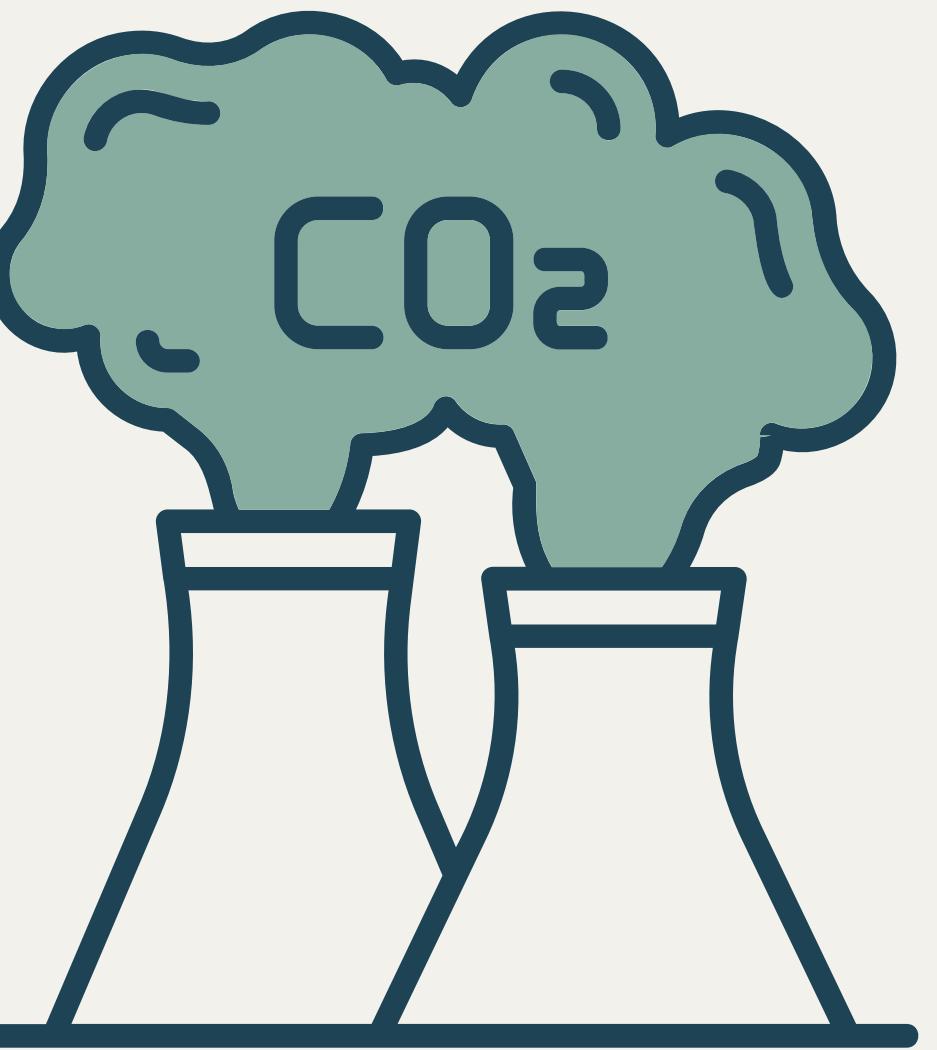
Modèles évalués :

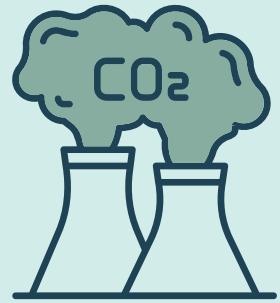
- Modèles linéaires : Régression Linéaire, Ridge, Lasso
- Modèles basés sur des arbres : Arbre de décision, Forêt aléatoire
- Régression par vecteurs de support (SVR)
- Modèles de boosting : XGBRegressor, LightGBM Regressor

Critères de sélection :

- Les meilleurs modèles sont choisis en fonction du score R² le plus élevé.

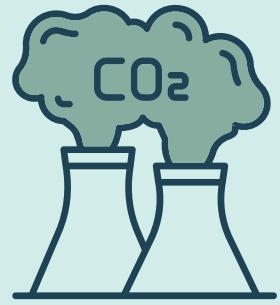
Cible Émissions





Scores de validation croisée sur l'ensemble d'entraînement

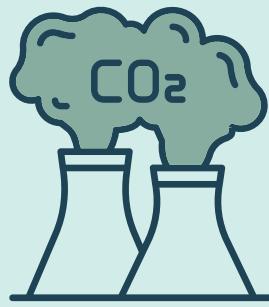
	R² (Mean ± Std)	MAE (Mean ± Std)	RMSE (Mean ± Std)
Régression Linéaire	0.779 ± 0.061	0.478 ± 0.044	0.656 ± 0.097
Ridge	0.776 ± 0.066	0.473 ± 0.039	0.660 ± 0.096
Lasso	0.786 ± 0.057	0.472 ± 0.038	0.645 ± 0.081
Arbre de décision	0.513 ± 0.074	0.758 ± 0.048	0.982 ± 0.064
Forêt aléatoire	0.662 ± 0.038	0.638 ± 0.027	0.819 ± 0.033
SVR	0.737 ± 0.132	0.476 ± 0.040	0.704 ± 0.177
XGBRegressor	0.799 ± 0.049	0.470 ± 0.031	0.628 ± 0.062
LightGBM Regressor	0.804 ± 0.047	0.461 ± 0.022	0.620 ± 0.056



Résultats de validation croisée Train-Test

	CV Train R ²	CV Test R ²	ΔR ²
Régression Linéaire	0.827	0.779	0.048
Ridge	0.829	0.776	0.053
Lasso	0.824	0.786	0.038
Arbre de décision	0.697	0.546	0.151
Forêt aléatoire	0.783	0.662	0.121
SVR	0.828	0.737	0.091
XGBRegressor	0.933	0.799	0.134
LightGBM Regressor	0.92	0.804	0.116

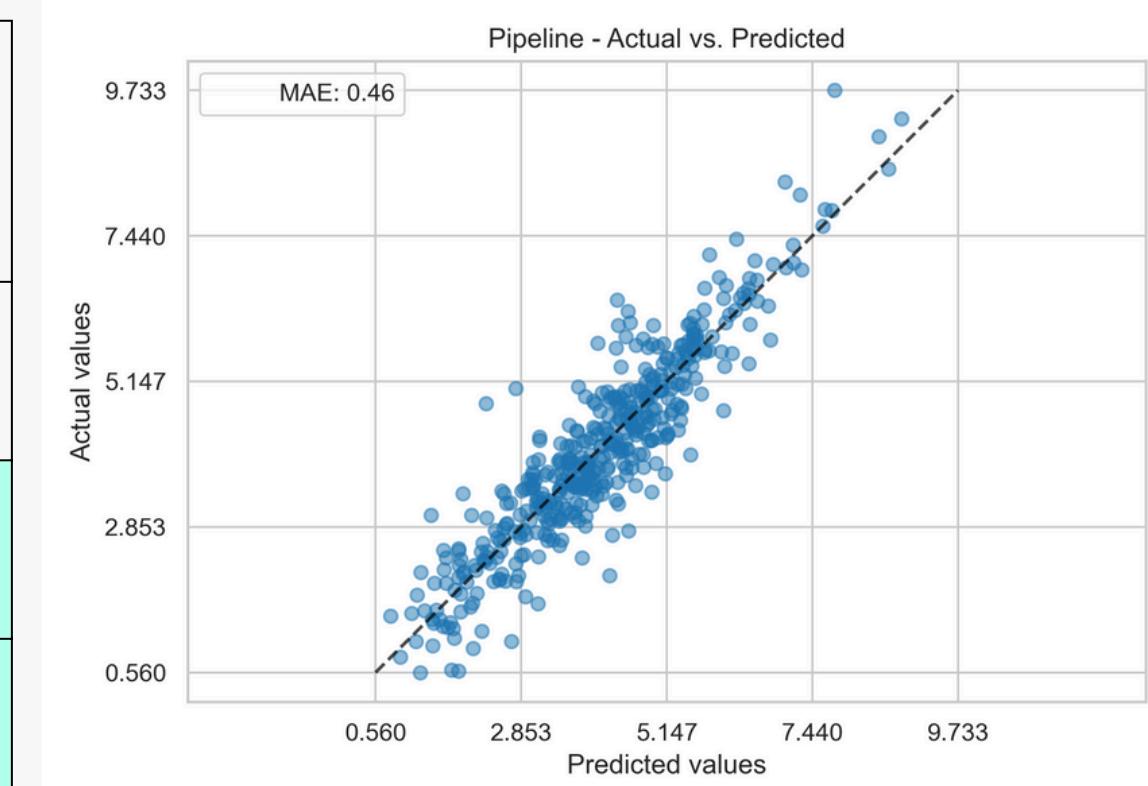
Cible Émissions



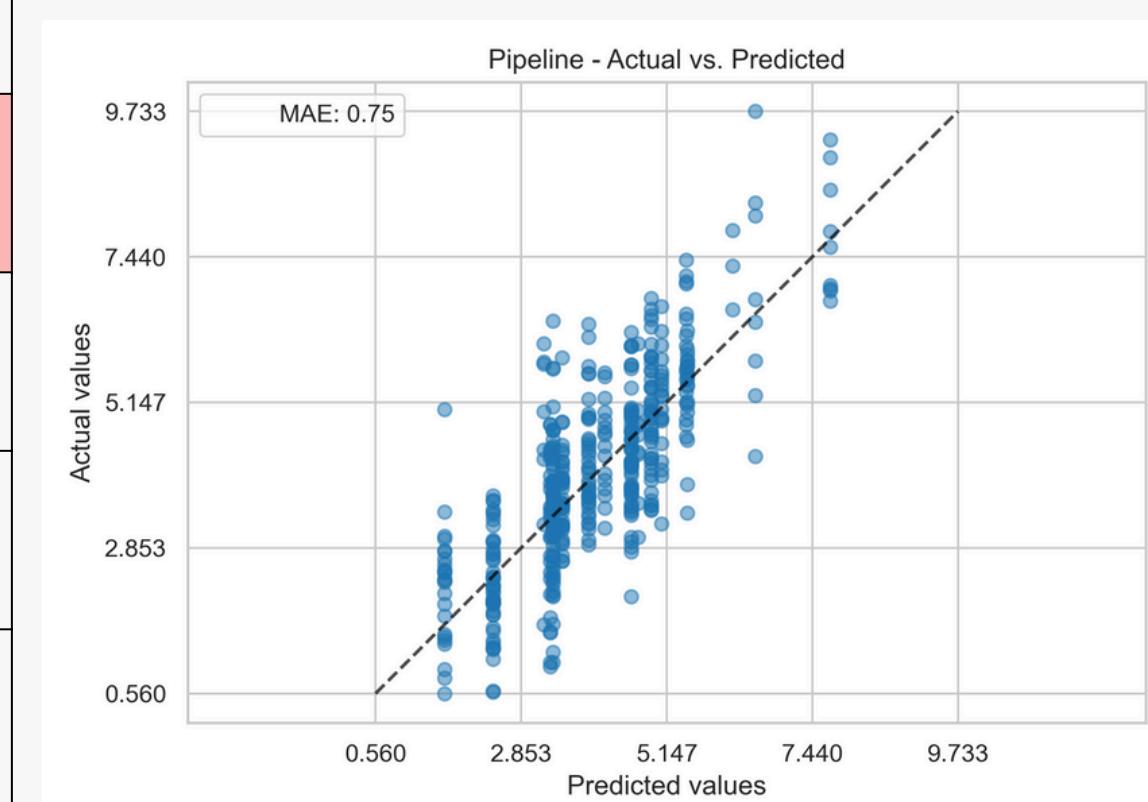
Performance sur des données test

Modèle	R ²	MAE	RMSE	Duration
Régression Linéaire	0.819	0.489	0.646	0.020
Ridge	0.820	0.488	0.644	0.003
Lasso	0.819	0.487	0.645	0.003
Arbre de décision	0.596	0.752	0.965	0.012
Forêt aléatoire	0.669	0.673	0.873	0.196
SVR	0.819	0.490	0.645	0.097
XGBRegressor	0.835	0.462	0.616	0.048
LightGBM Regressor	0.832	0.464	0.623	0.067

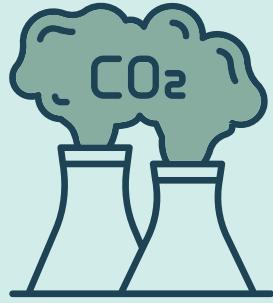
XGB Regressor



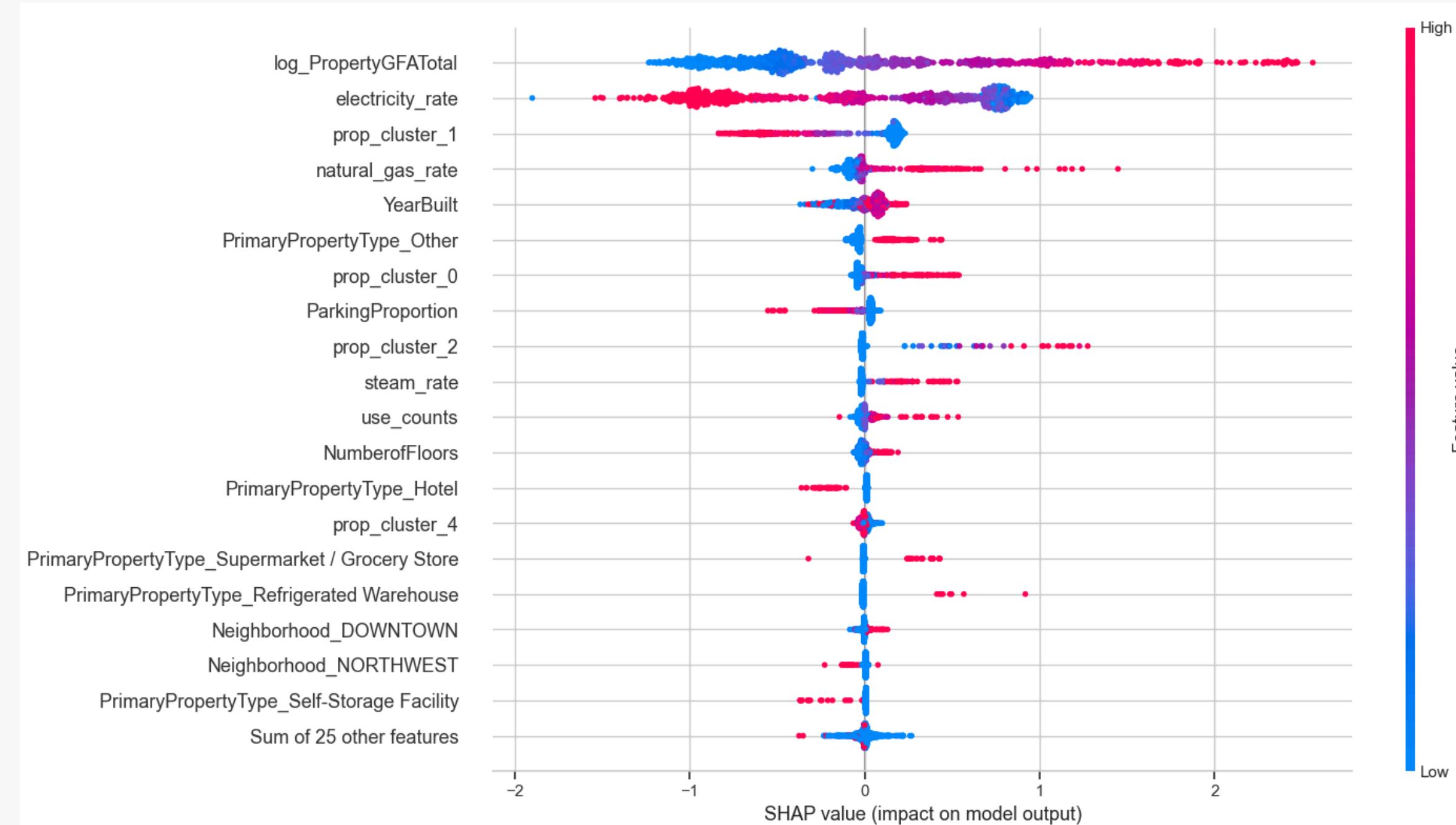
Arbre de décision



Cible Émissions

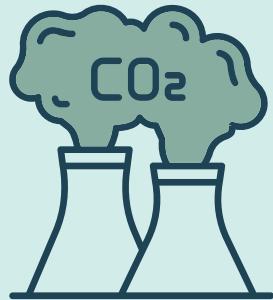


Meilleur modèle: XGB Regressor



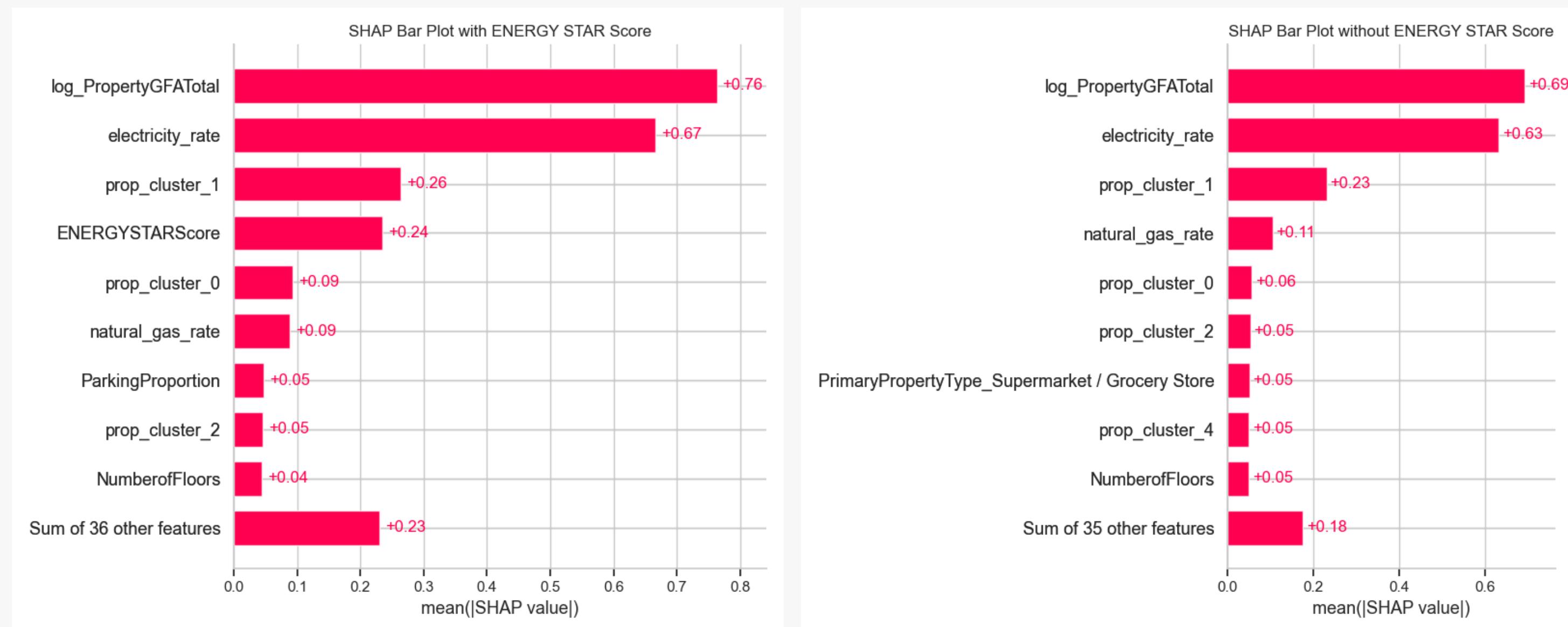
- La taille du bâtiment a la plus grande influence sur le modèle.
- Un taux d'électricité plus élevé correspond à des émissions plus faibles.
- Un taux de vapeur plus élevé correspond à des émissions plus élevées.
- Une proportion de parking plus élevée correspond à des émissions plus faibles.
- Un nombre d'usages plus élevé correspond à des émissions plus élevées.

Cible Émissions



Meilleur modèle: XGB Regressor avec/sans ENERGYSTARSCORE

Evaluation	R ²	MAE	RMSE	Durée
Avec ENERGY STAR Score	0.931	0.287	0.382	0.077
Sans ENERGY STAR Score	0.883	0.383	0.5	0.021



Cible Energie





Scores de validation croisée sur l'ensemble d'entraînement

Modèle	R ² (Mean ± Std)	MAE (Mean ± Std)	RMSE (Mean ± Std)
Régression Linéaire	0.717 ± 0.080	0.482 ± 0.039	0.668 ± 0.108
Ridge	0.743 ± 0.037	0.481 ± 0.033	0.640 ± 0.058
Lasso	0.751 ± 0.038	0.470 ± 0.034	0.629 ± 0.057
Arbre de décision	0.443 ± 0.124	0.715 ± 0.057	0.934 ± 0.074
Forêt aléatoire	0.638 ± 0.028	0.589 ± 0.035	0.759 ± 0.039
SVR	0.710 ± 0.089	0.480 ± 0.035	0.675 ± 0.116
XGBRegressor	0.760 ± 0.032	0.473 ± 0.031	0.618 ± 0.047
LightGBM Regressor	0.760 ± 0.034	0.471 ± 0.036	0.618 ± 0.055



Résultats de validation croisée Train-Test

Modèle	CV Train R^2	CV Test R^2	ΔR^2
Régression Linéaire	0.777	0.717	0.06
Ridge	0.773	0.743	0.03
Lasso	0.782	0.751	0.031
Arbre de décision	0.605	0.496	0.109
Forêt aléatoire	0.767	0.638	0.129
SVR	0.781	0.71	0.071
XGBRegressor	0.868	0.76	0.108
LightGBM Regressor	0.883	0.76	0.123

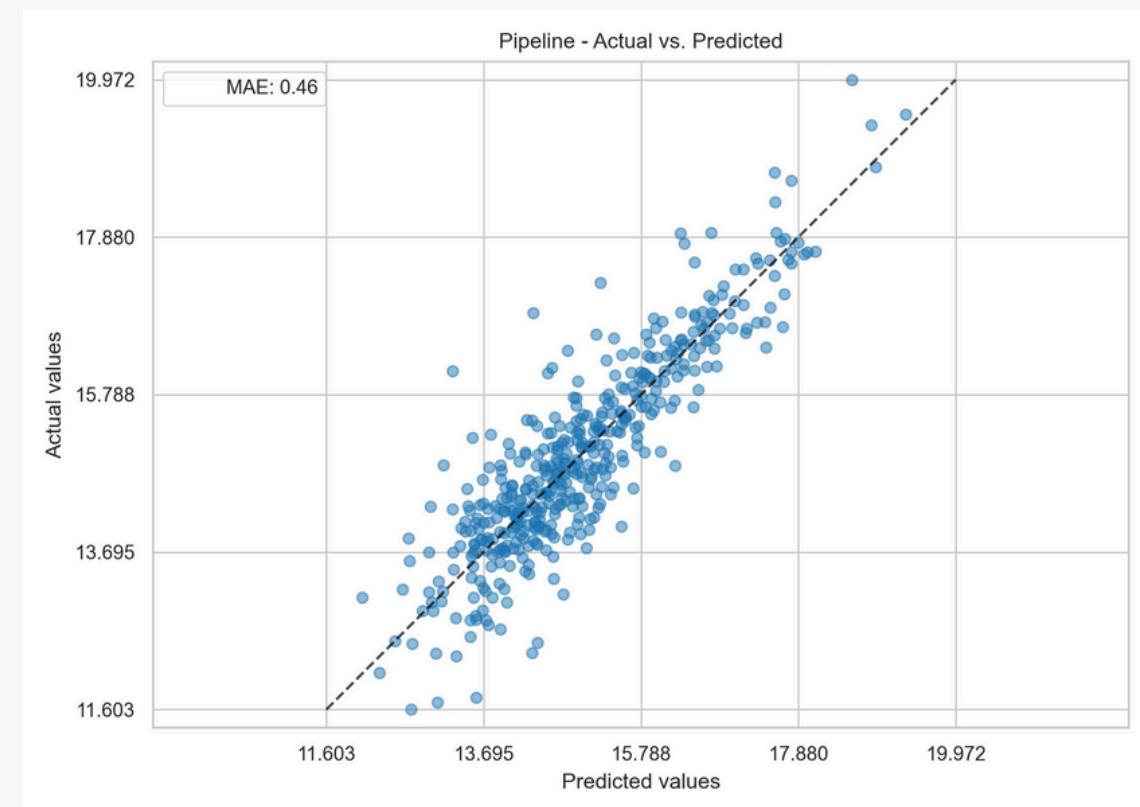
Cible Energie



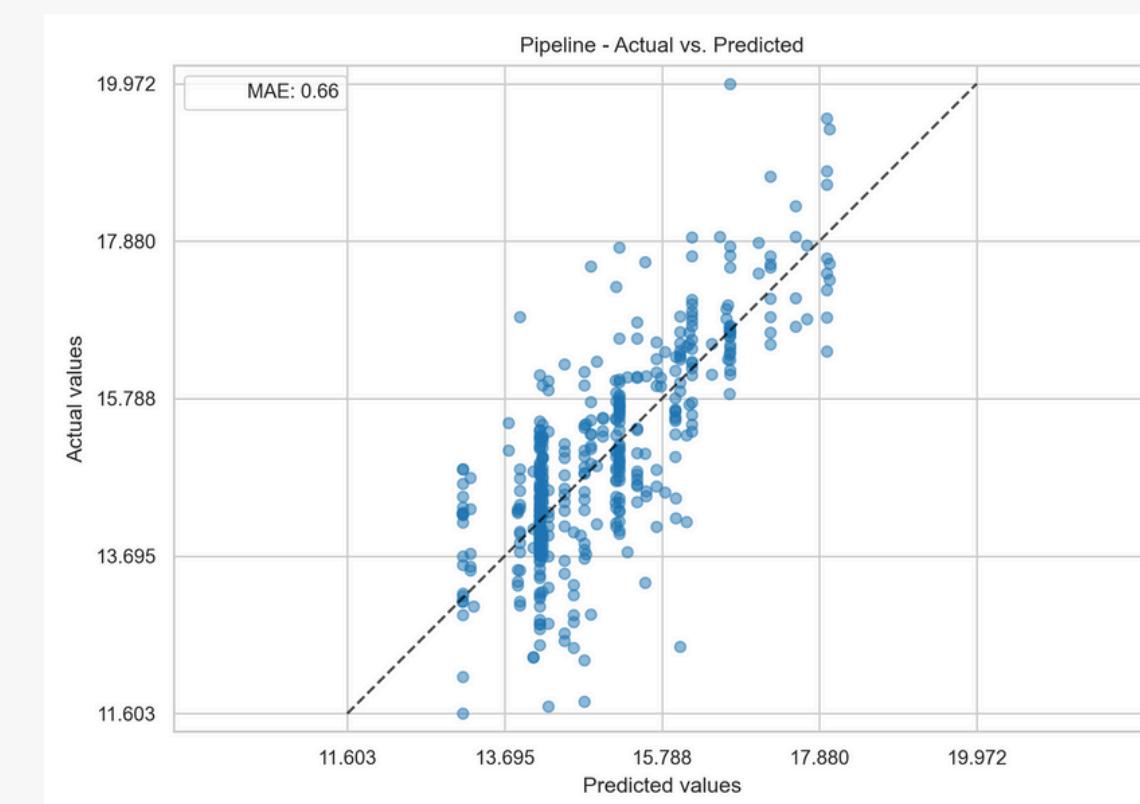
Performance sur des données test

Modèle	R ²	MAE	RMSE	Duration
Régression Linéaire	0.772	0.487	0.648	0.011
Ridge	0.760	0.499	0.665	0.005
Lasso	0.772	0.483	0.648	0.005
Arbre de décision	0.589	0.658	0.869	0.004
Forêt aléatoire	0.630	0.622	0.825	0.074
SVR	0.769	0.494	0.652	0.032
XGBRegressor	0.789	0.458	0.623	0.178
LightGBM Regressor	0.789	0.465	0.624	0.012

Light GBM Regressor



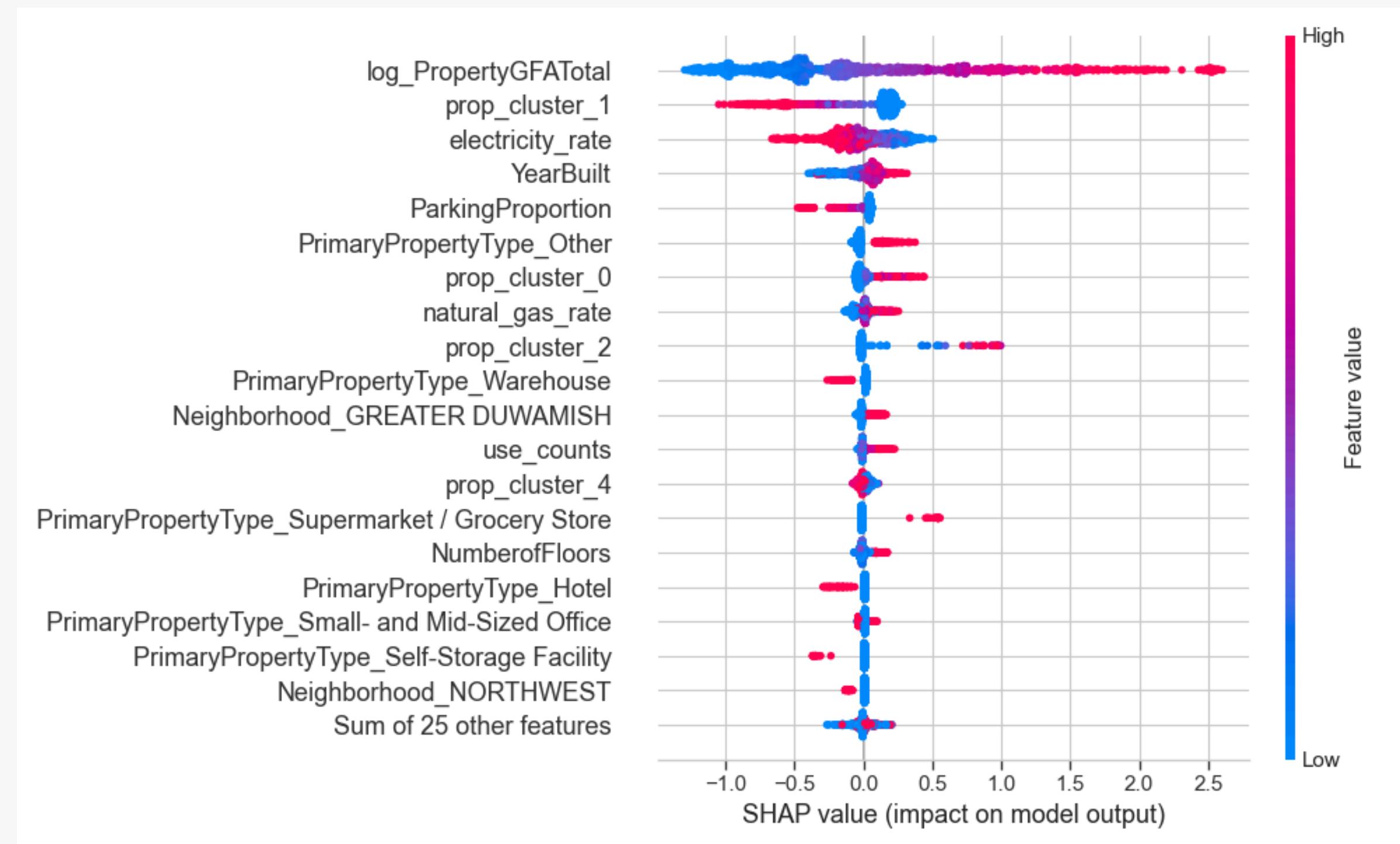
Arbre de décision



Cible Energie



Meilleur modèle: LightGBM Regressor



Prop_cluster_1

- Non-Refrigerated Warehouse,
- Worship Facility,
- Movie Theater
- Library
- Self-Storage Facility
- Distribution Center
- Bank Branch
- Automobile Dealership
- Performing Arts

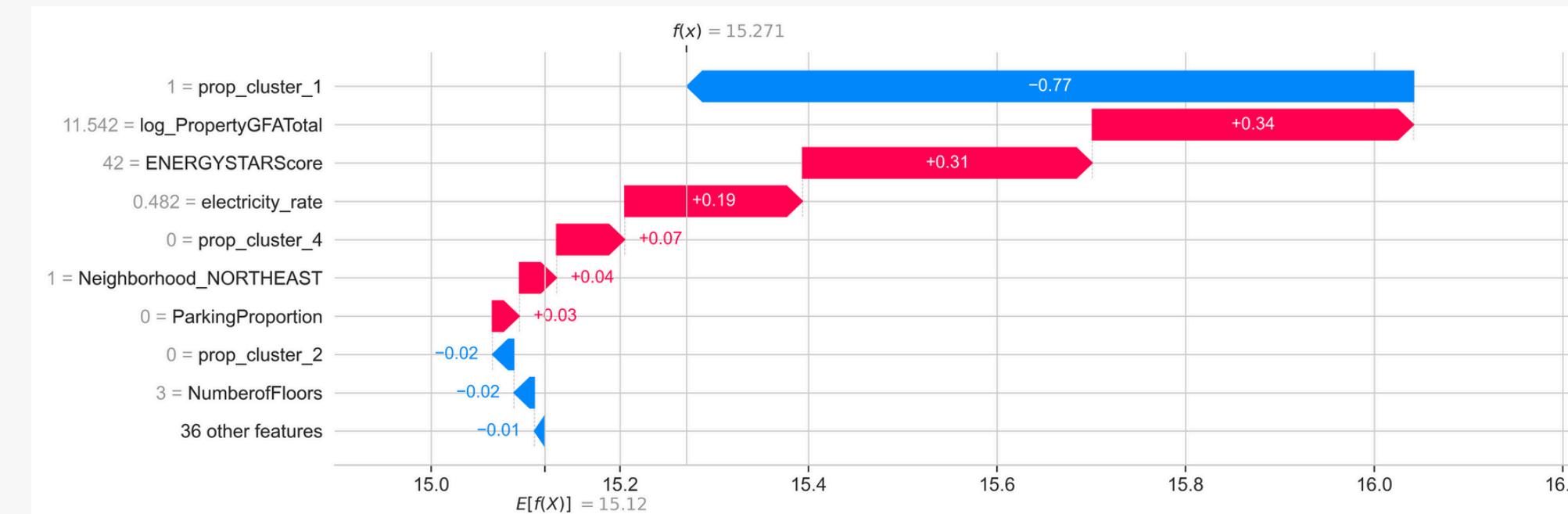
Cible Energie



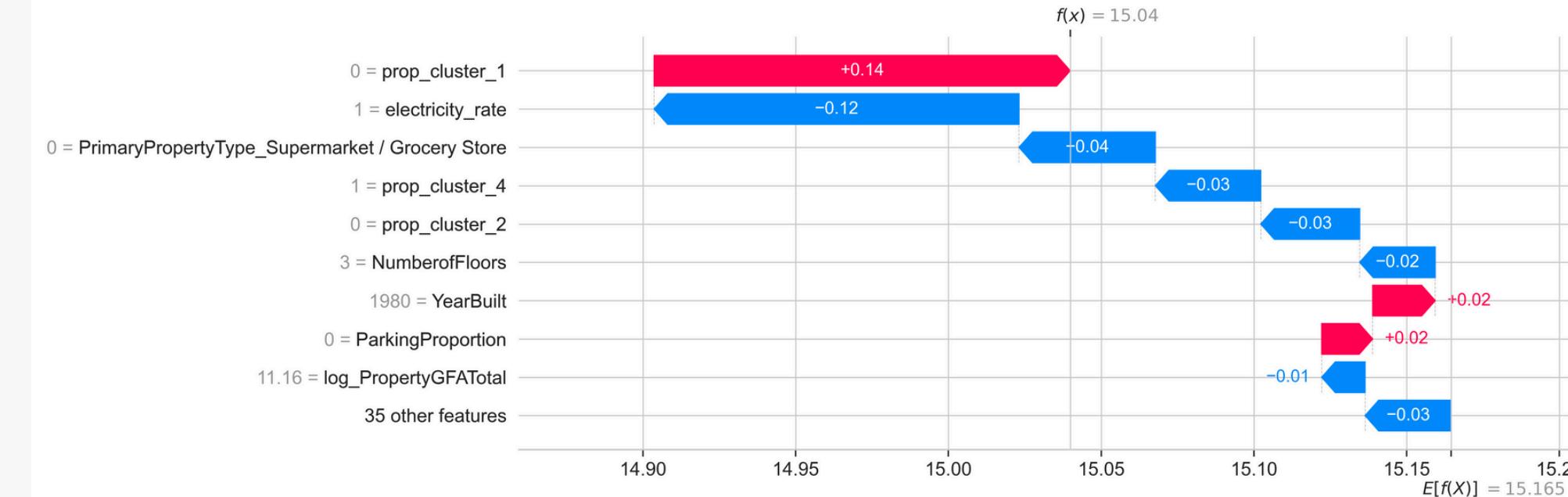
Meilleur modèle: LightGBM Regressor avec/sans ENERGYSTARSCORE

Evaluation	R ²	MAE	RMSE	Durée
Avec ENERGY STAR Score	0.904	0.304	0.407	0.100
Sans ENERGY STAR Score	0.830	0.426	0.542	0.012

Avec ENERGY STAR SCORE



Sans ENERGY STAR SCORE



Conclusion

Les modèles de boosting excellent

Les modèles de boosting offrent la meilleure précision prédictive.

L'ENERGY STAR Score est essentiel

L'inclusion de l'ENERGY STAR Score améliore significativement le R^2 et les métriques d'erreur.

Des insights exploitables

Un modèle prédictif peut être utilisé dans la stratégie de Seattle pour atteindre la neutralité carbone d'ici 2050.