



FORMATION DATA SCIENTIST : PROJET 4

Anticipez les besoins en consommation de bâtiments



Sommaire

1. Introduction
2. Données et sources
3. Analyse exploratoire et prétraitement des données
4. Méthodologie : Consommation énergétique
5. Modélisation : Consommation énergétique
6. Méthodologie : Emissions GES
7. Modélisation : Emission GES
8. Résultats et interprétation
9. Conclusion et Recommendations



Seattle

1. Introduction

- Contexte :

Vers une Seattle neutre en carbone en 2050



- Objectifs :

- Prédire les émissions de CO2
- Estimer la consommation d'énergie



- Importance de l'Energy Star Score

Evaluer l'efficacité de l'Energy Star Score pour des prédictions précises



2. Données et Sources

- Données Collectées en 2016 auprès de la ville de Seattle

Collecte exhaustive sur les bâtiments non résidentiels

- Variables Explicatives Potentielles :



Groupes	Noms des variables
GES	TotalGHGEmissions, GHGEmissionsIntensity
Energie	SiteEUI(kBtu/sf), SiteEUIWN(kBtu/sf) SourceEUI(kBtu/sf), SourceEUIWN(kBtu/sf) SiteEnergyUse(kBtu), SiteEnergyUseWN(kBtu) , SteamUse(kBtu), Electricity(kWh), Electricity(kBtu), NaturalGas(therms), NaturalGas(kBtu)
Bâtiment	YearBuilt, NumberofBuildings, NumberofFloors
Energy Star Score	ENERGYSTARScore
Localisation	Neighborhood
Type bâtiment/utilisation	BuildingType, ListOfAllPropertyUseTypes, PrimaryPropertyType, LargestPropertyUseType, SecondLargestPropertyUseType, ThirdLargestPropertyUseType
Surface	PropertyGFATotal, PropertyGFAParking, PropertyGFABuilding(s), LargestPropertyUseTypeGFA, SecondLargestPropertyUseTypeGFA, ThirdLargestPropertyUseTypeGFA
Autres	ComplianceStatus, Outlier

3. Analyse exploratoire et prétraitement des données

- Analyse Exploratoire et Prétraitement

Processus itératif pour une qualité de données optimale

- Nettoyage des Données

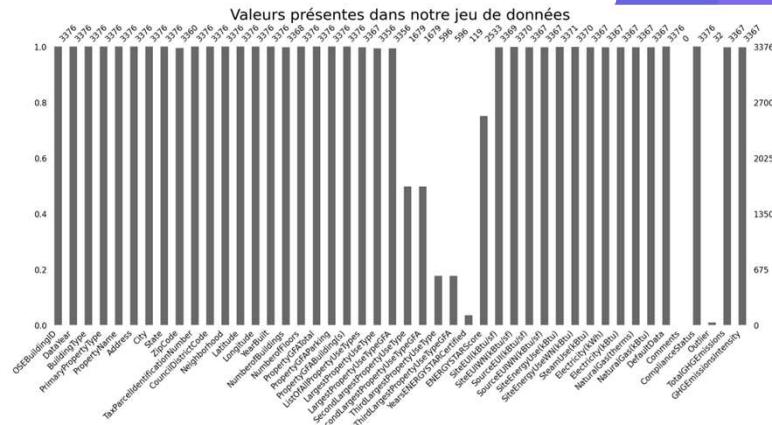
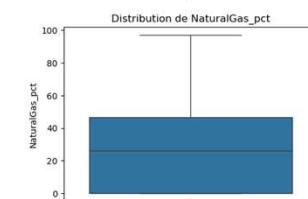
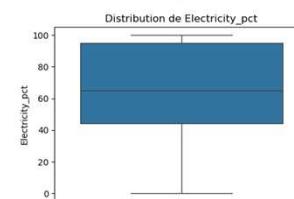
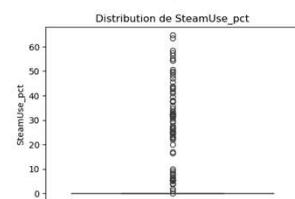
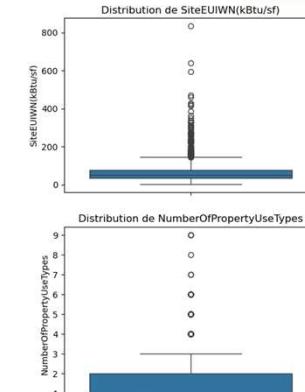
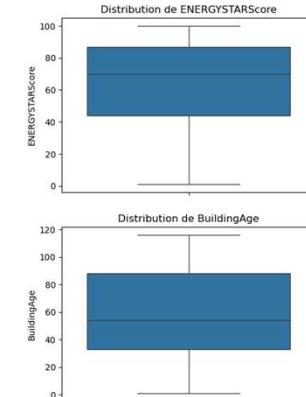
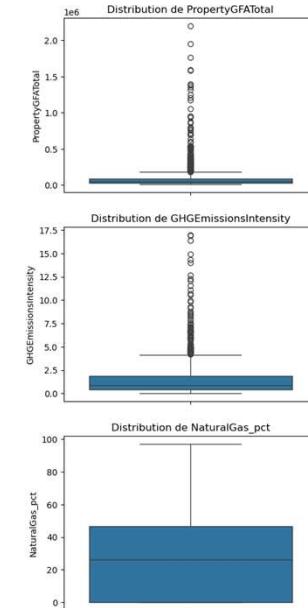
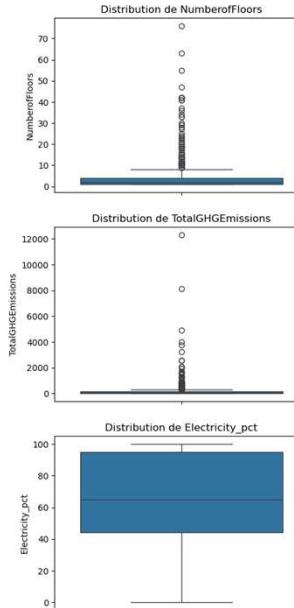
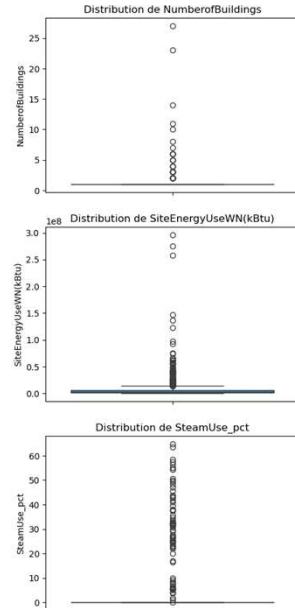
Formats, doublons, valeurs manquantes

- Déetecter et Traiter les Outliers

Vérifications manuelles sur Google Maps,
Règles métier,
IQR

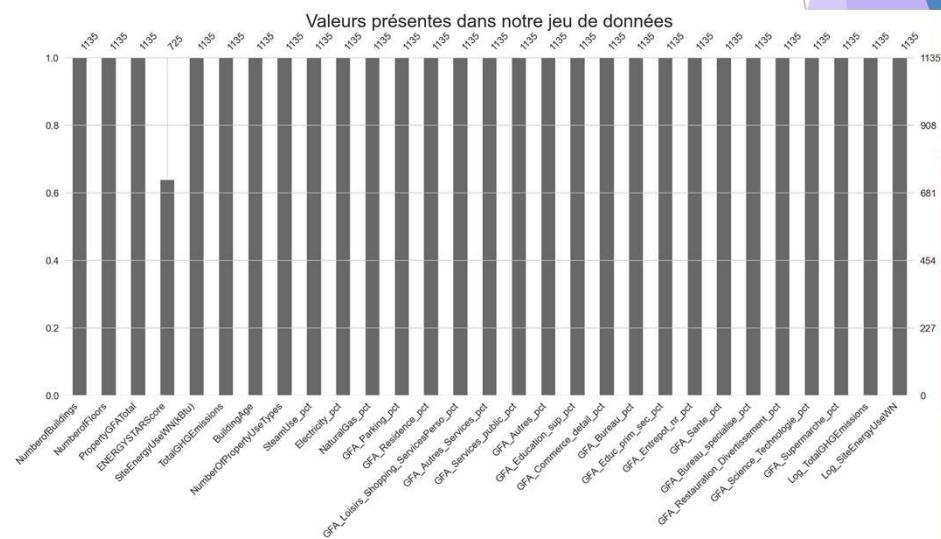
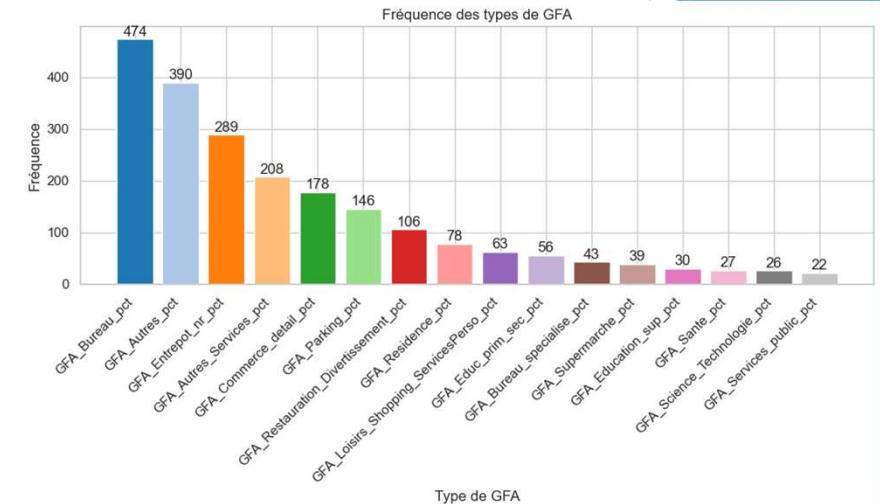
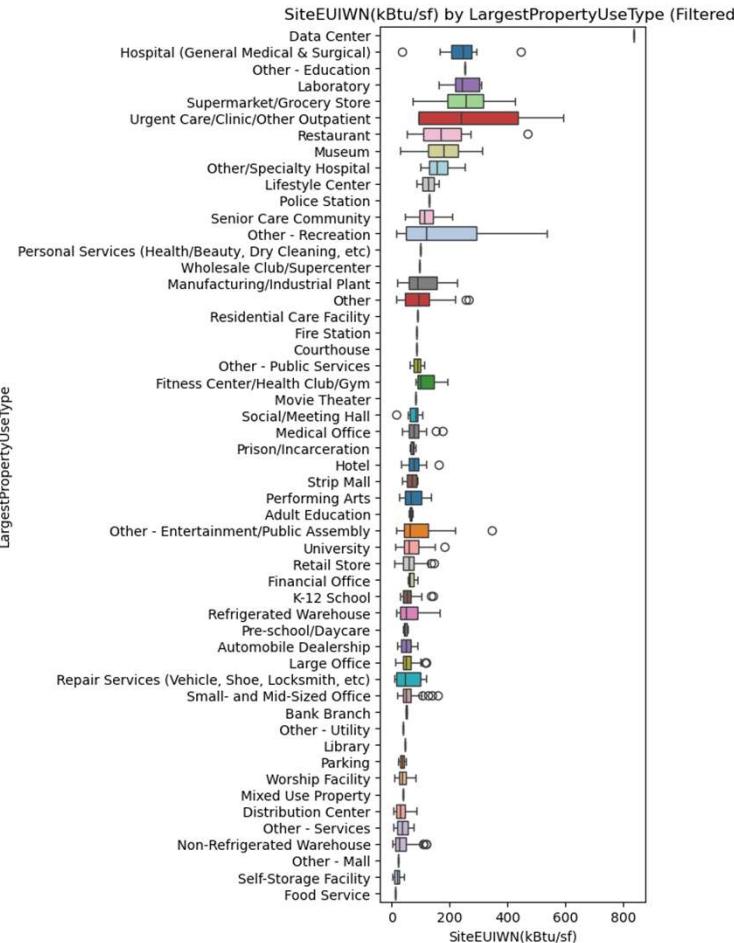
- Gestion des fuites de données

Suppressions ou transformation en %



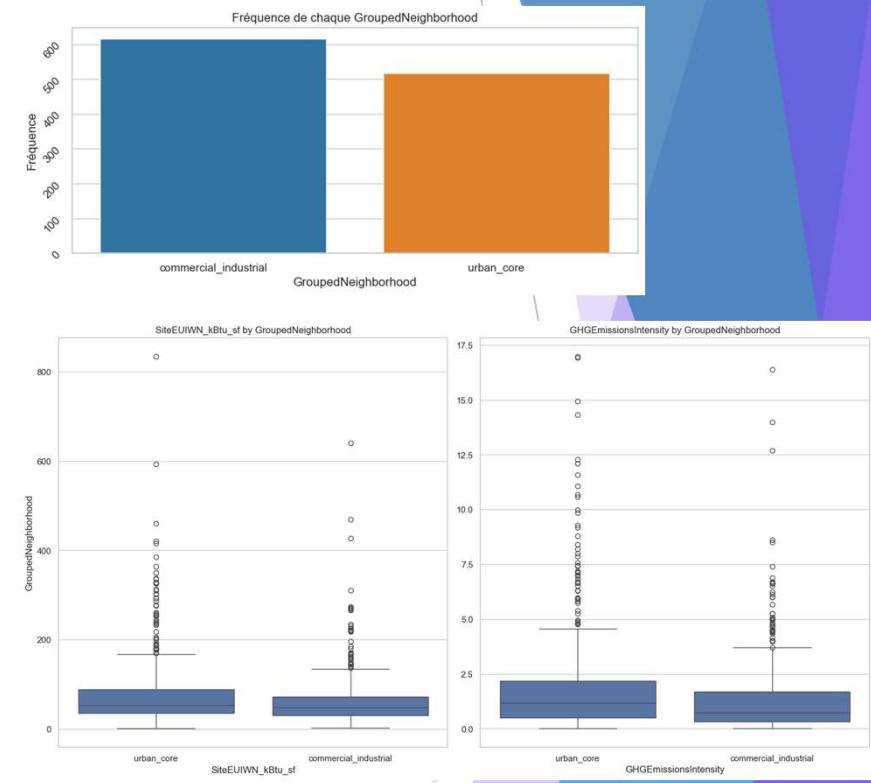
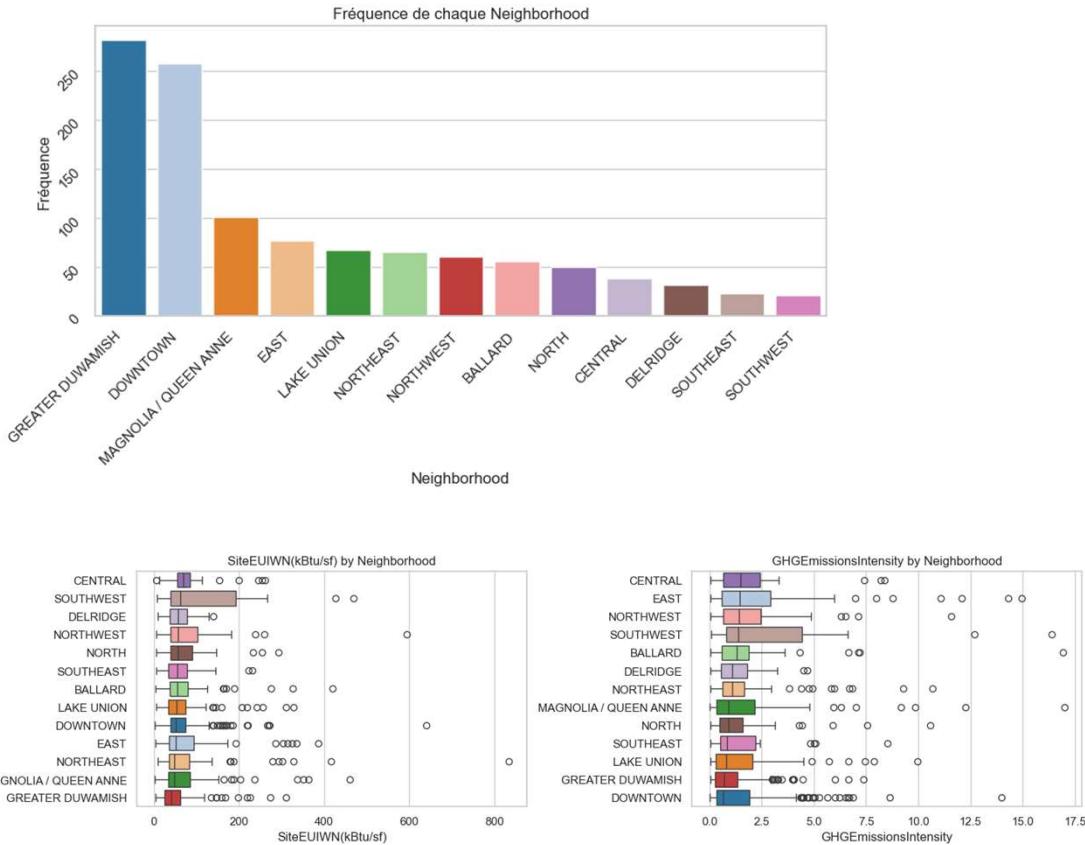
3. Analyse exploratoire et prétraitement des données

- Création de nouvelles variables



3. Analyse exploratoire et prétraitement des données

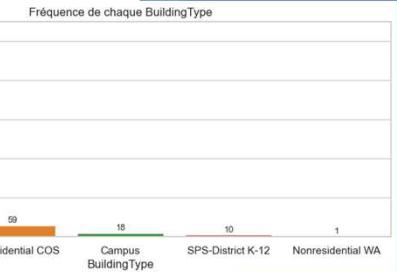
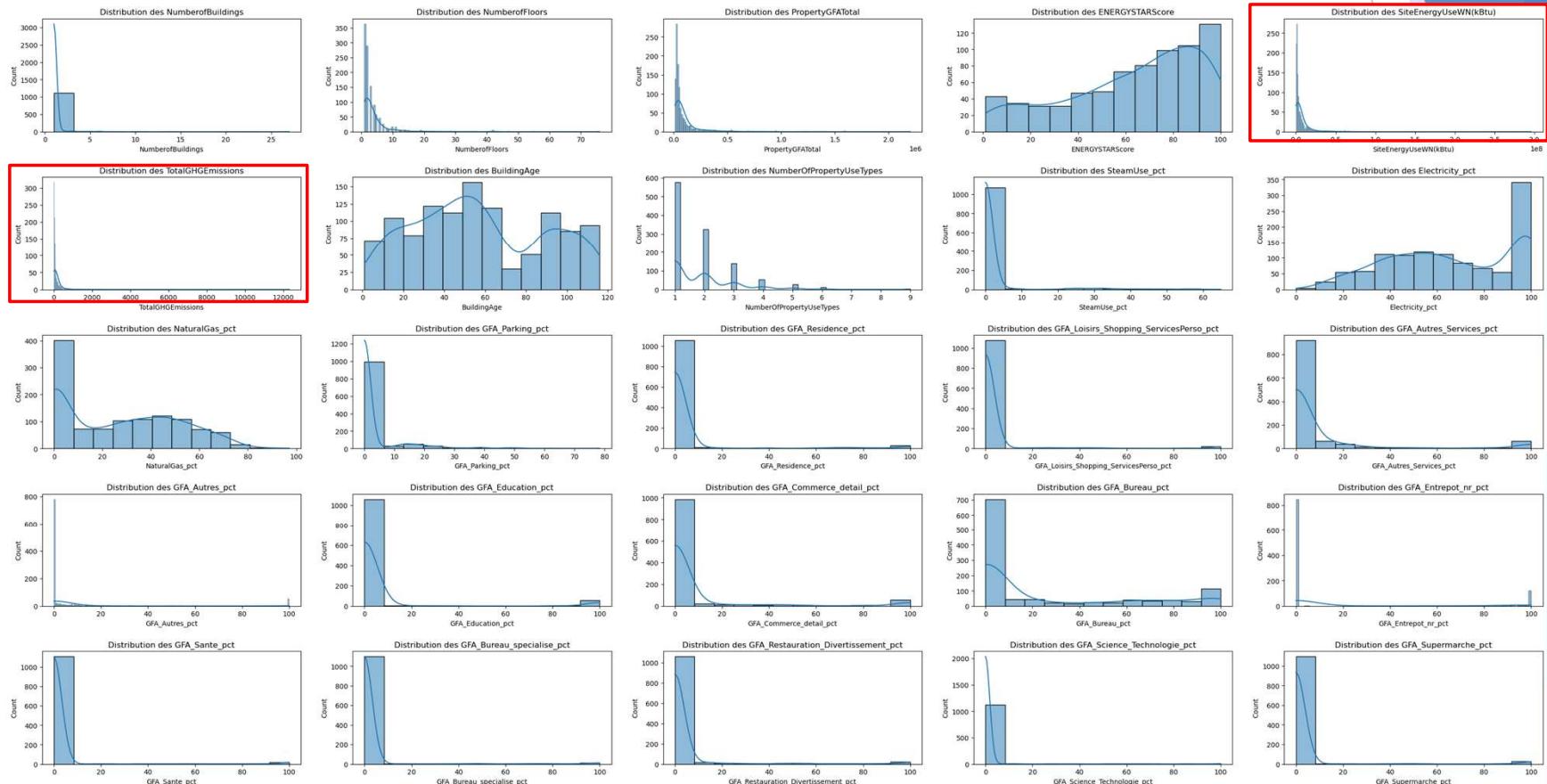
- Visualisation des Données



Chi-squared value: 935.6163943522128
P-value: 0.43305896577369646

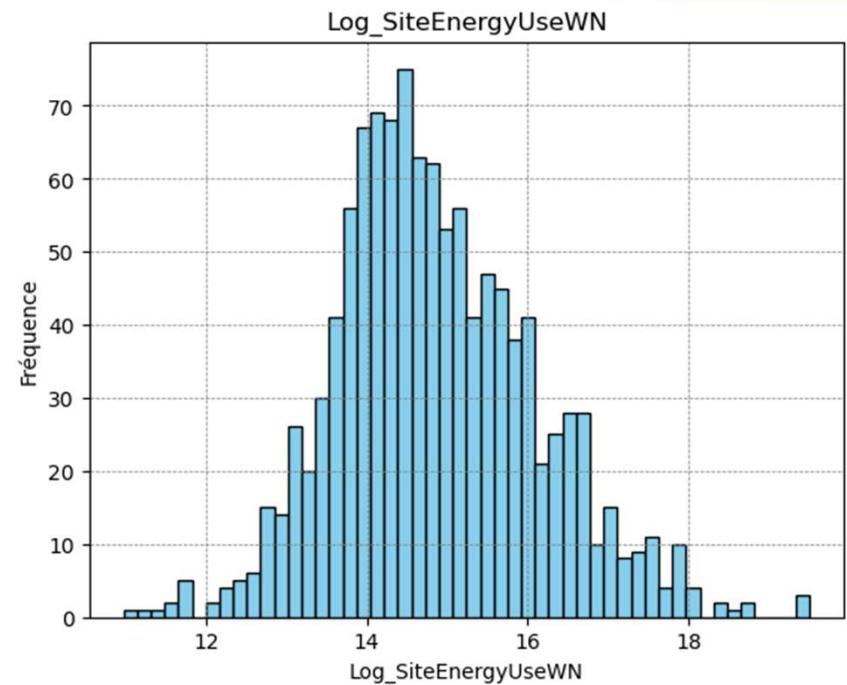
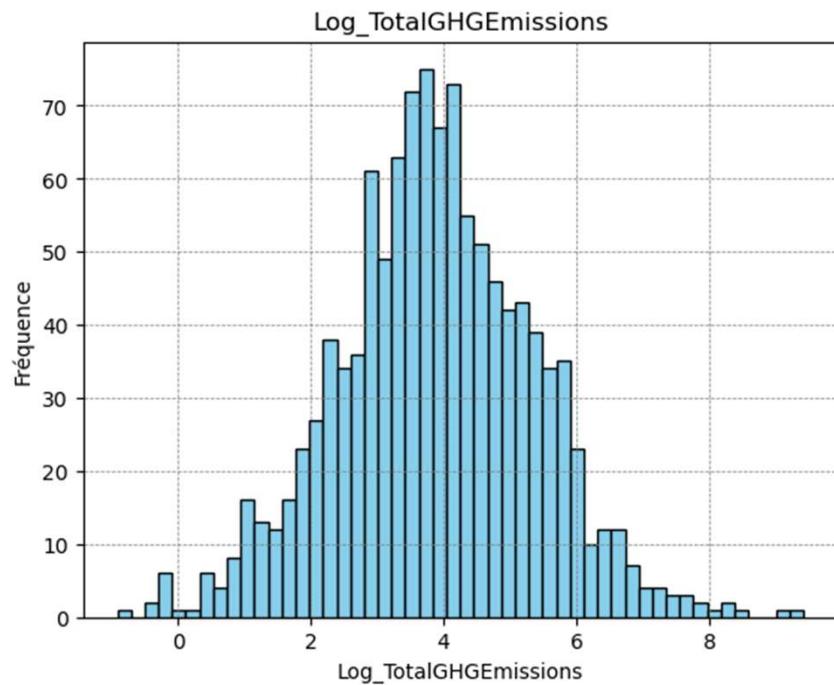
3. Analyse exploratoire et prétraitement des données

- Visualisation des Données



3. Analyse exploratoire et prétraitement des données

- Transformation des Variables :



4. Méthodologie :

Modèles de prédition de l'énergie consommée

- Approche Méthodologique

Processus structuré pour prédition précise

- Séparation des données en jeu de test et d'entraînement
- Normalisation par StandardScaler
- (OneHotEncoder)
- Test de Student

- Sélection des Modèles à Tester

Liste des modèles :

- Dummy Regressor Median
- Régression linéaire
- Ridge, Lasso
- ElasticNet
- SVR
- Gradient Boosting et XGBoost
- Random Forest
- Ada Boost

- Critères d'Évaluation et Validation Croisée

Recherche des hyperparamètres : RandomizedSearchCV + GridSearchCV

Validation croisée sur le meilleur modèle

Résultats sur jeu de test vs jeu d'entraînement

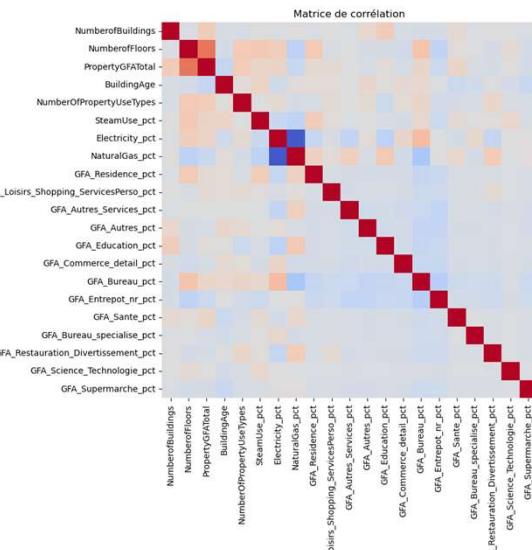
Validation Croisée RMSE: 0.5608 (± 0.0426)

Validation Croisée MAE: 0.4275 (± 0.0298)

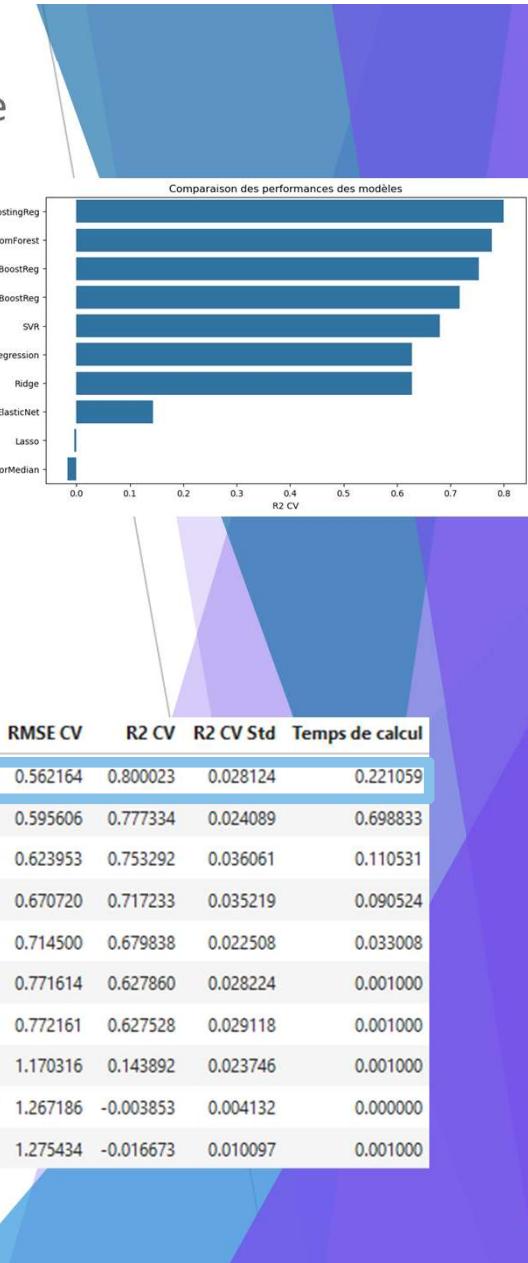
Validation Croisée R²: 0.8021 (± 0.0244)

Ensemble d'entraînement : RMSE = 0.3929, R² = 0.9043, MAE = 0.3026

Ensemble de test : RMSE = 0.5513, R² = 0.7822, MAE = 0.4263

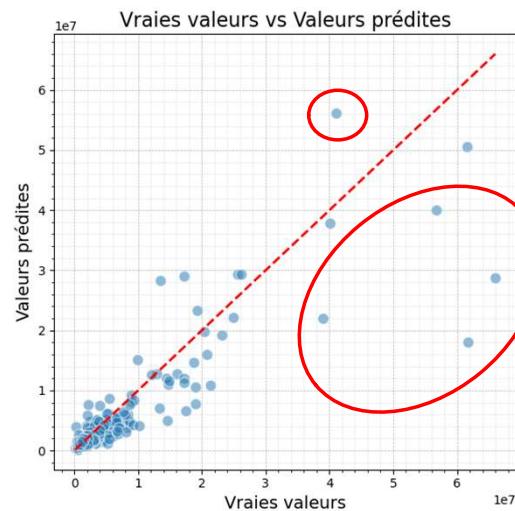
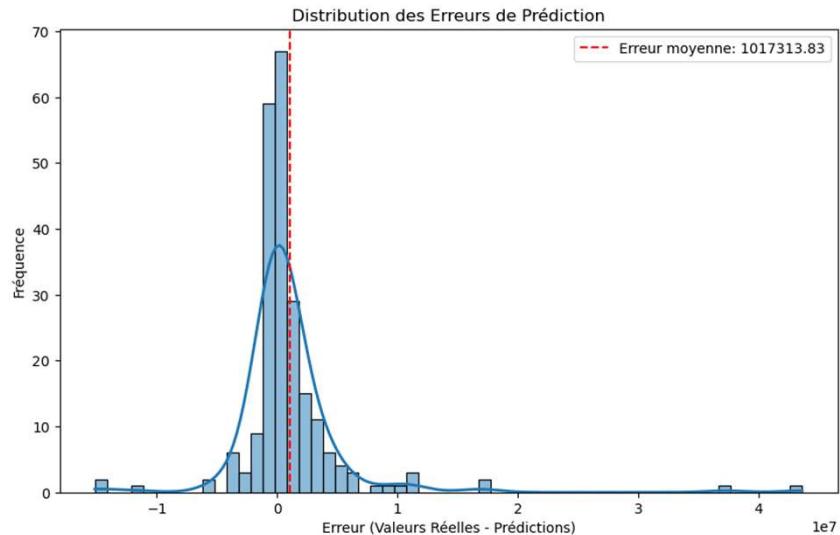


Modèle	RMSE	MAE	R2	RMSE CV	R2 CV	R2 CV Std	Temps de calcul
GradientBoostingReg	0.556543	0.432993	0.778049	0.562164	0.800023	0.028124	0.221059
RandomForest	0.565114	0.438163	0.771160	0.595606	0.777334	0.024089	0.698833
XGBoostReg	0.586868	0.453769	0.753203	0.623953	0.753292	0.036061	0.110531
AdaBoostReg	0.644207	0.502388	0.702621	0.670720	0.717233	0.035219	0.090524
SVR	0.695099	0.545066	0.653780	0.714500	0.679838	0.022508	0.033008
LinearRegression	0.792181	0.630172	0.550315	0.771614	0.627860	0.028224	0.001000
Ridge	0.793610	0.631469	0.548691	0.772161	0.627528	0.029118	0.001000
ElasticNet	1.096742	0.891666	0.138078	1.170316	0.143892	0.023746	0.001000
Lasso	1.181363	0.946430	-0.000060	1.267186	-0.003853	0.004132	0.000000
DummyRegressorMedian	1.189829	0.943303	-0.014445	1.275434	-0.016673	0.010097	0.001000

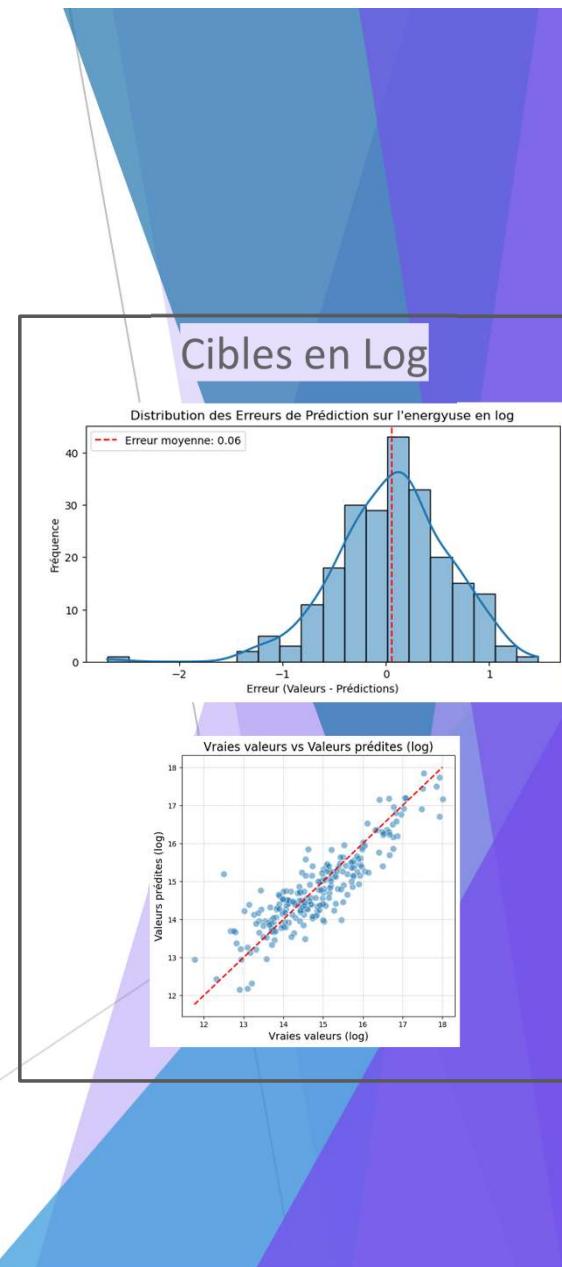


5. Modélisation : Modèles de prédition de l'énergie consommée

- Analyse des erreurs



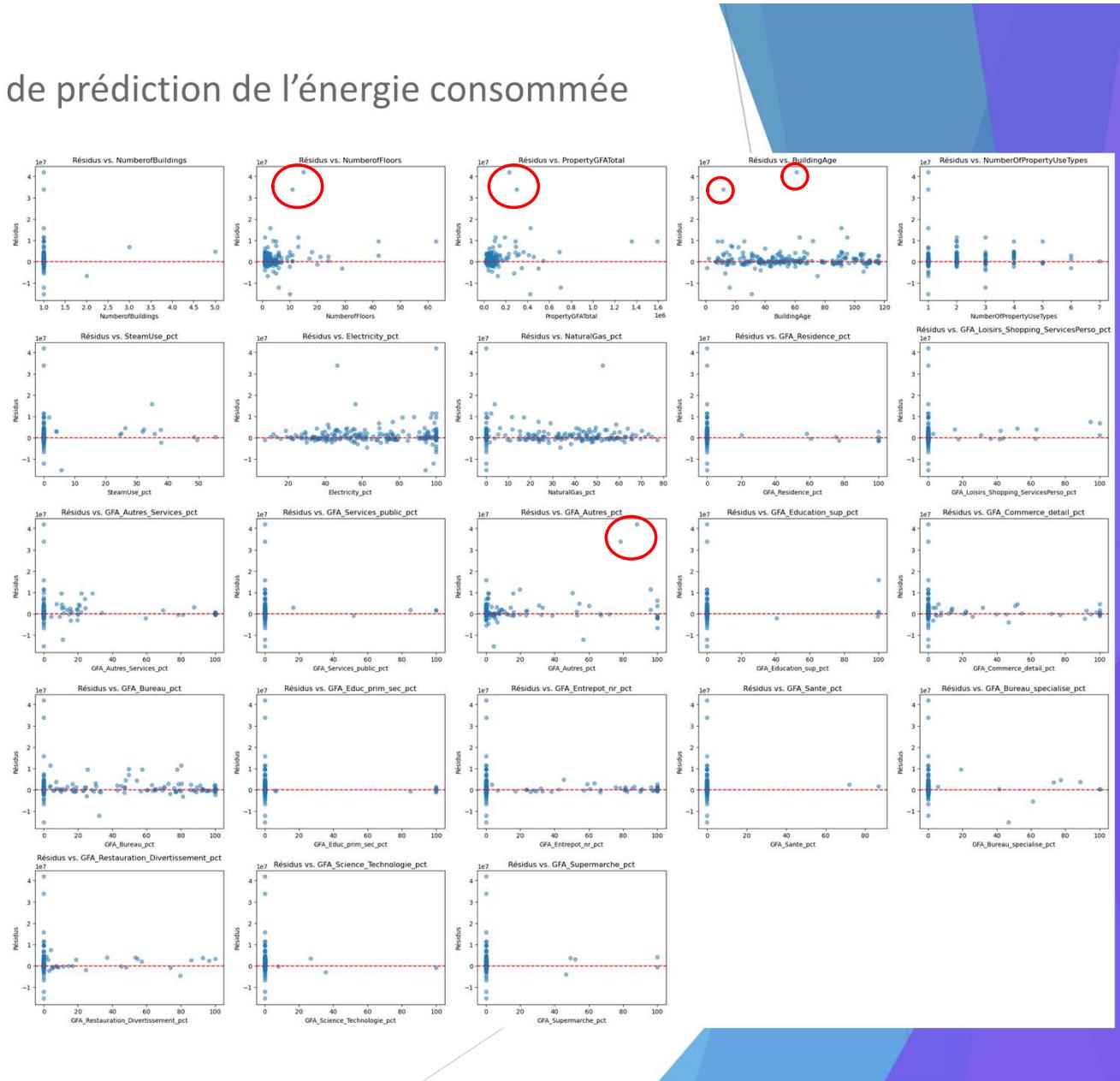
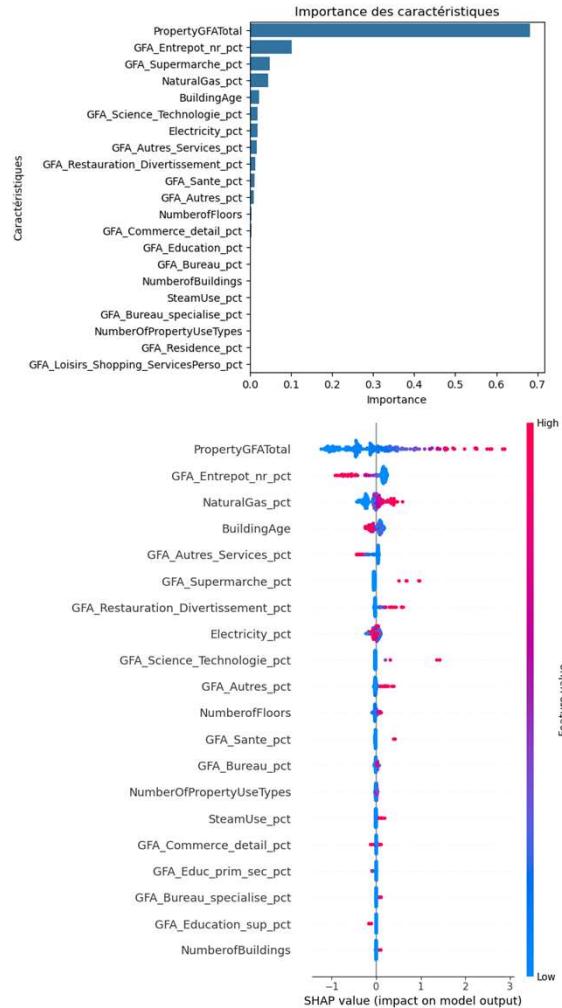
SiteEnergyUseWN(kBtu)	ListOfAllPropertyUseTypes	PrimaryPropertyType	SecondLargestPropertyUseType	ThirdLargestPropertyUseType
66000296.0	Other	Other	NaN	NaN
61674856.0	Other	Other	NaN	NaN
56785916.0	Data Center, Financial Office, Office, Parking...	Large Office	Parking	Financial Office
38977108.0	College/University	University	NaN	NaN
41078600.0	Data Center, Laboratory, Museum, Office, Other...	Mixed Use Property	Office	Parking



5. Modélisation :

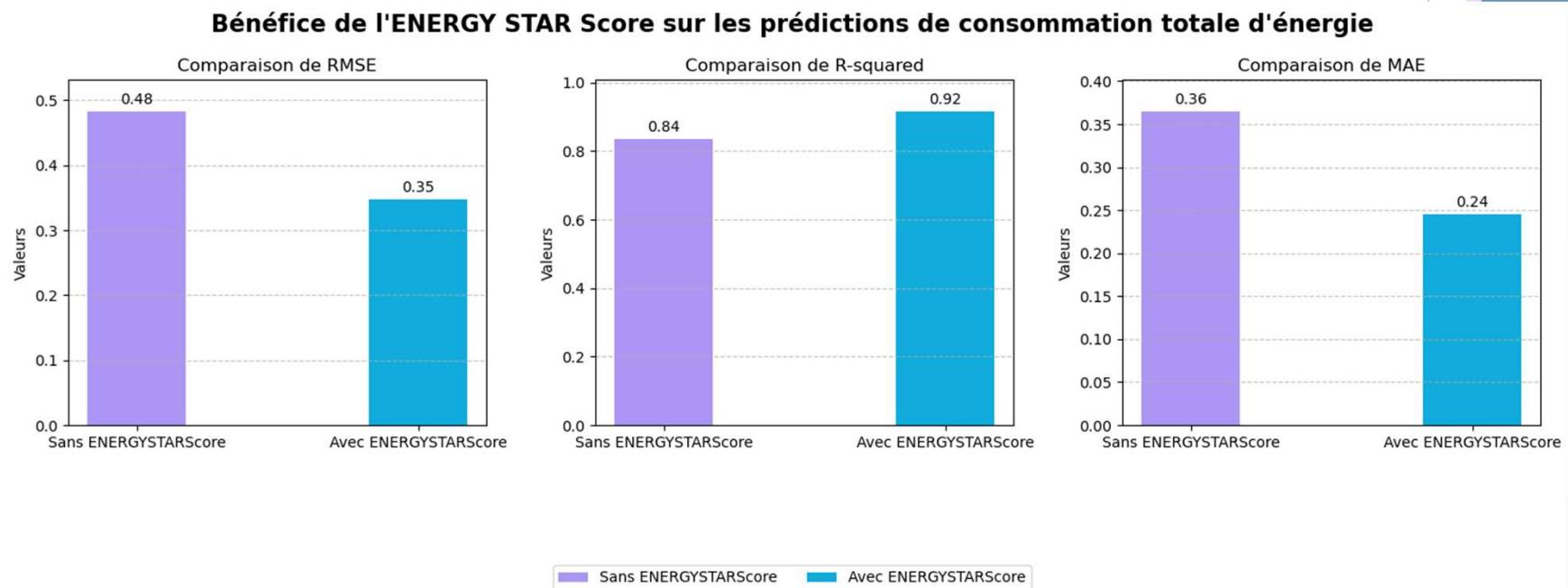
Modèles de prédition de l'énergie consommée

- Importance des variables



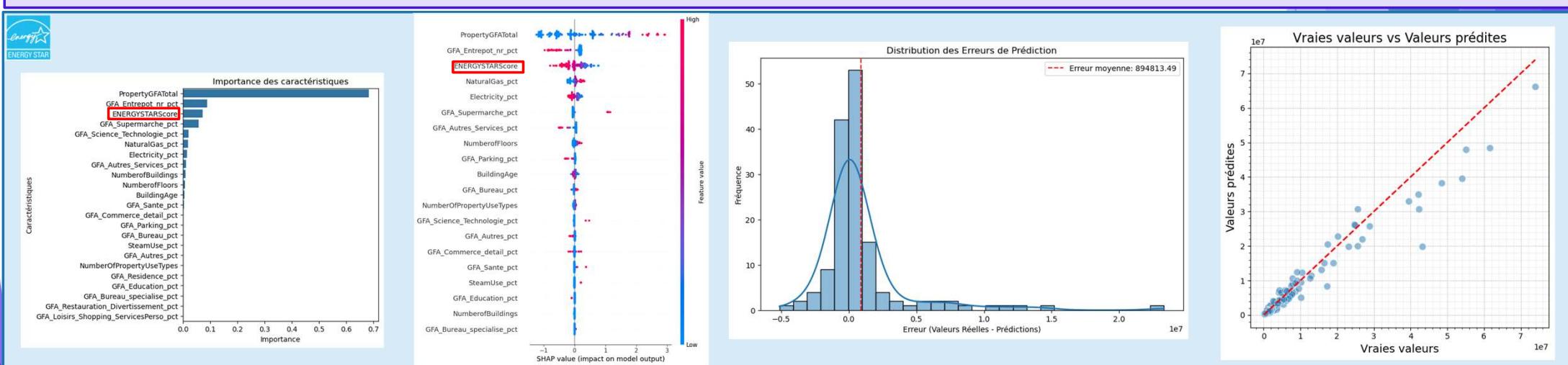
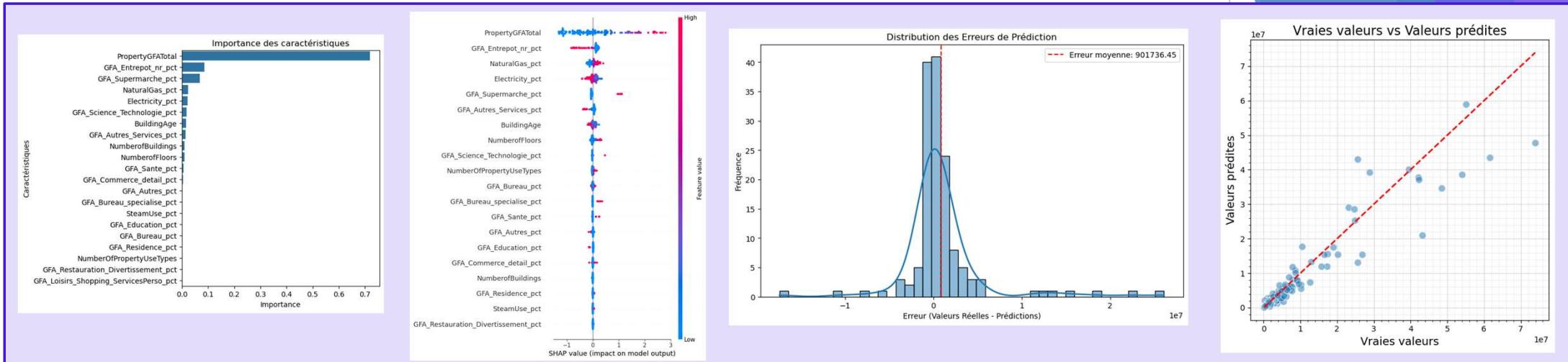
5. Modélisation : Modèles de prédiction de l'énergie consommée

- Impact de l'Energy Star Score



5. Modélisation : Modèles de prédition de l'énergie consommée

- Importance des variables



6. Méthodologie : Modèles de prédiction des émissions de GES

- Approche Méthodologique

Identique prédictions de consommation énergétique

- Sélection des Modèles à Tester

Liste des modèles :

- Dummy Regressor Median
- Régression linéaire
- Ridge, Lasso
- ElasticNet
- SVR
- Gradient Boosting et XGBoost
- Random Forest
- Ada Boost

- Critères d'Évaluation et Validation Croisée

Identiques prédictions de consommation énergétique

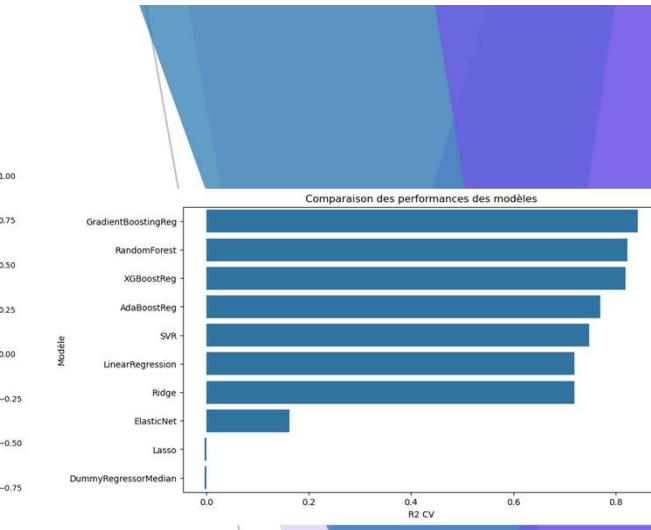
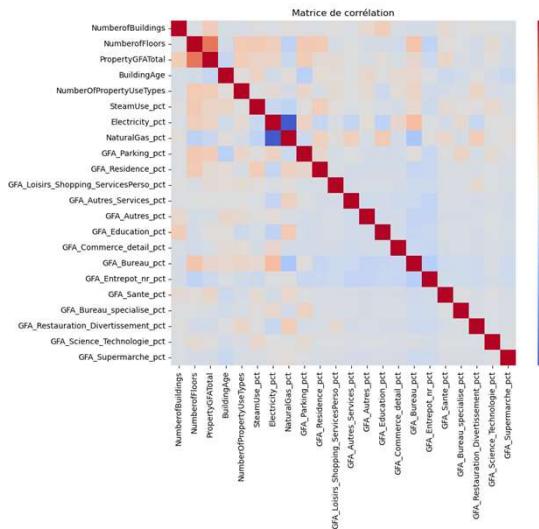
Validation Croisée RMSE: 0.5764 (± 0.0321)

Validation Croisée MAE: 0.4456 (± 0.0284)

Validation Croisée R²: 0.8523 (± 0.0143)

Ensemble d'entraînement : RMSE = 0.4172, R² = 0.9232, MAE = 0.3226

Ensemble de test : RMSE = 0.5544, R² = 0.8361, MAE = 0.4270

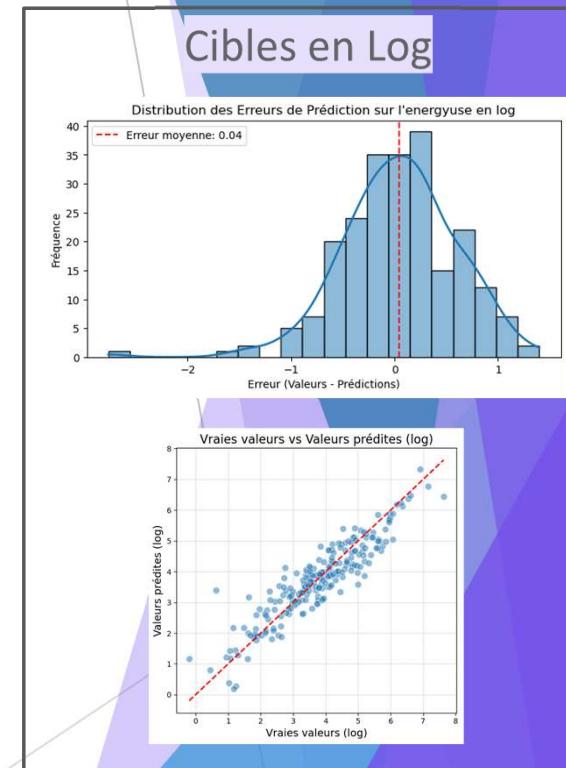
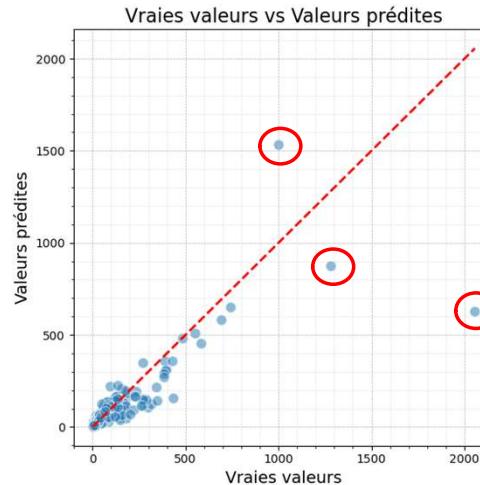
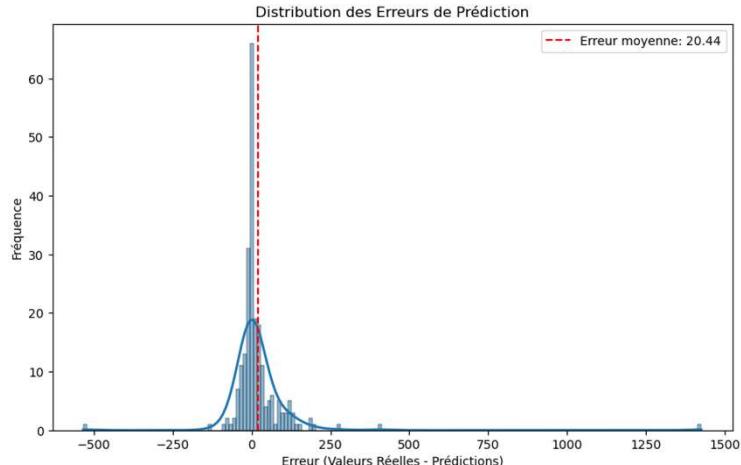


Modèle	RMSE	MAE	R2	RMSE CV	R2 CV	R2 CV Std	Temps de calcul
GradientBoostingReg	0.553152	0.434706	0.836816	0.592377	0.843686	0.017342	0.215173
RandomForest	0.580644	0.457789	0.820192	0.632745	0.822545	0.016368	0.719377
XGBoostReg	0.610420	0.476833	0.801278	0.638188	0.819246	0.020060	0.111529
AdaBoostReg	0.669892	0.532317	0.760670	0.719504	0.769668	0.030407	0.096527
SVR	0.716561	0.557932	0.726162	0.753813	0.748115	0.021868	0.029006
LinearRegression	0.819566	0.642561	0.641775	0.795698	0.719576	0.032217	0.000000
Ridge	0.817698	0.642831	0.643406	0.796339	0.719155	0.034162	0.001001
ElasticNet	1.266421	0.994872	0.144649	1.375381	0.162269	0.012501	0.000000
Lasso	1.372498	1.076277	-0.004642	1.504665	-0.002870	0.003210	0.001000
DummyRegressorMedian	1.370029	1.075025	-0.001031	1.505240	-0.003594	0.002272	0.000000

7. Modélisation :

Modèles de prédition des émissions de GES

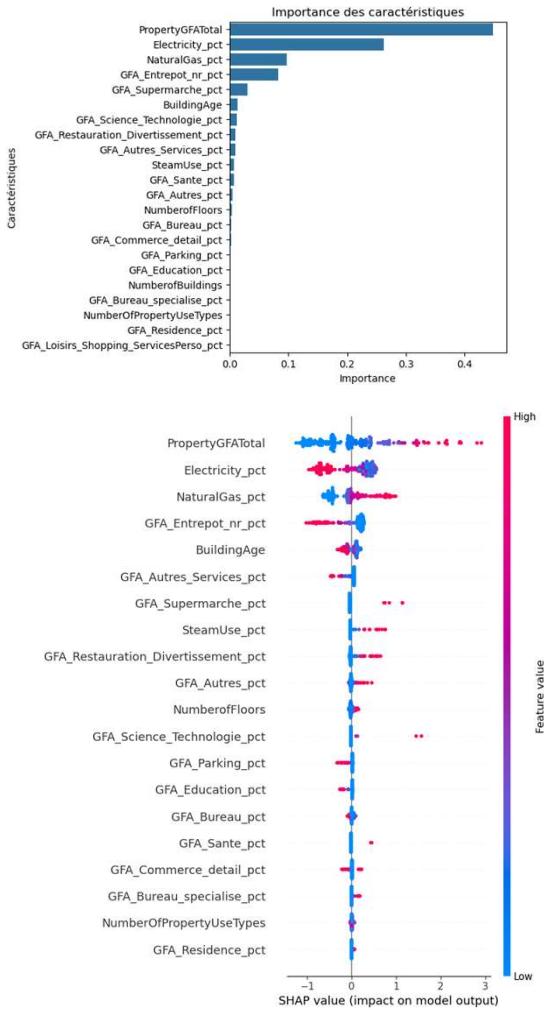
- Analyse des erreurs



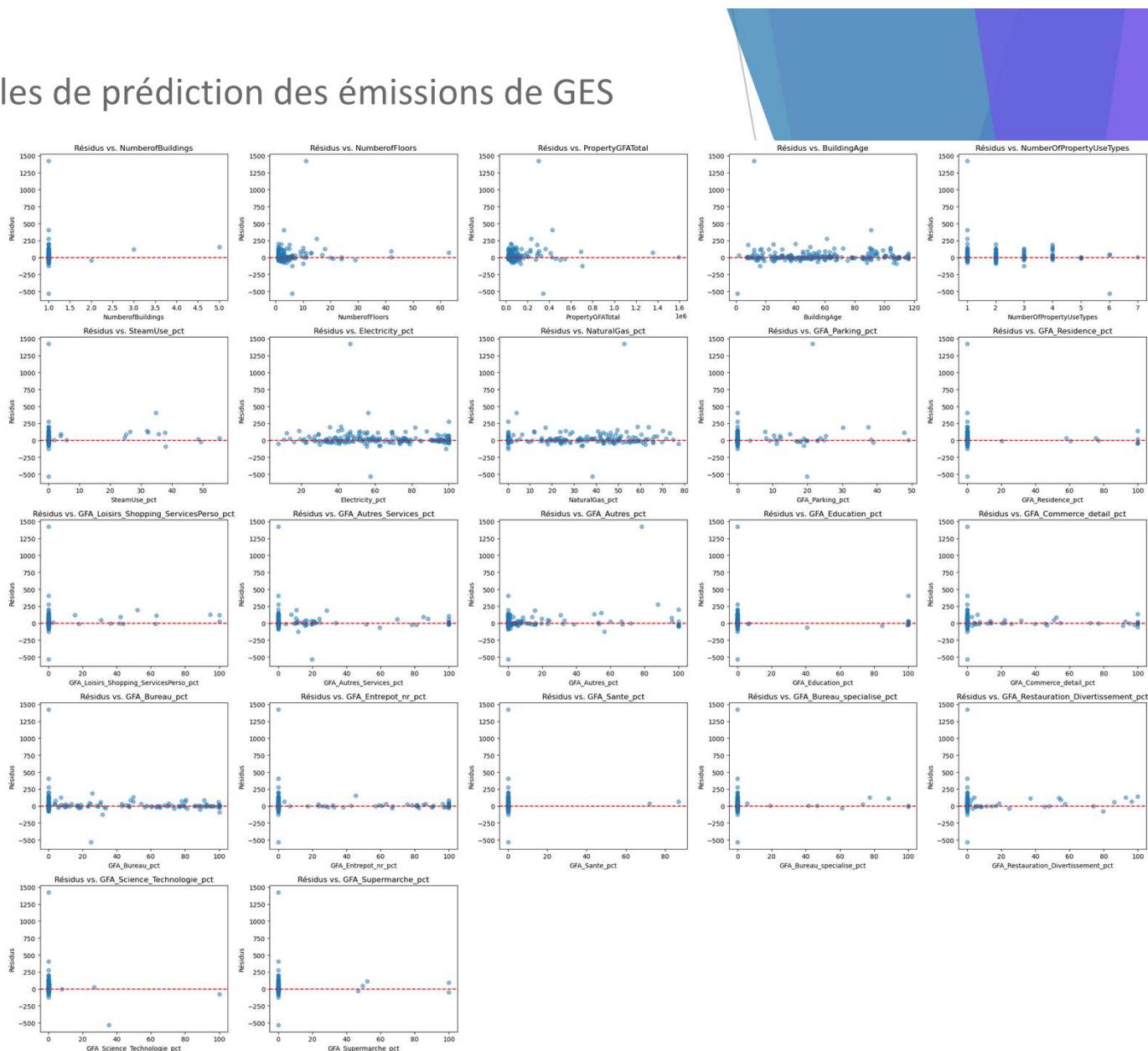
TotalGHGEmissions	ListAllPropertyUseTypes	PrimaryPropertyType	SecondLargestPropertyUseType	ThirdLargestPropertyUseType
2055.82		Other		
1280.81	College, University	University		
1000.06	Data Center, Laboratory, Museum, Office, Other...	Mixed Use Property	Office	Parking

7. Modélisation :

- Importance des variables

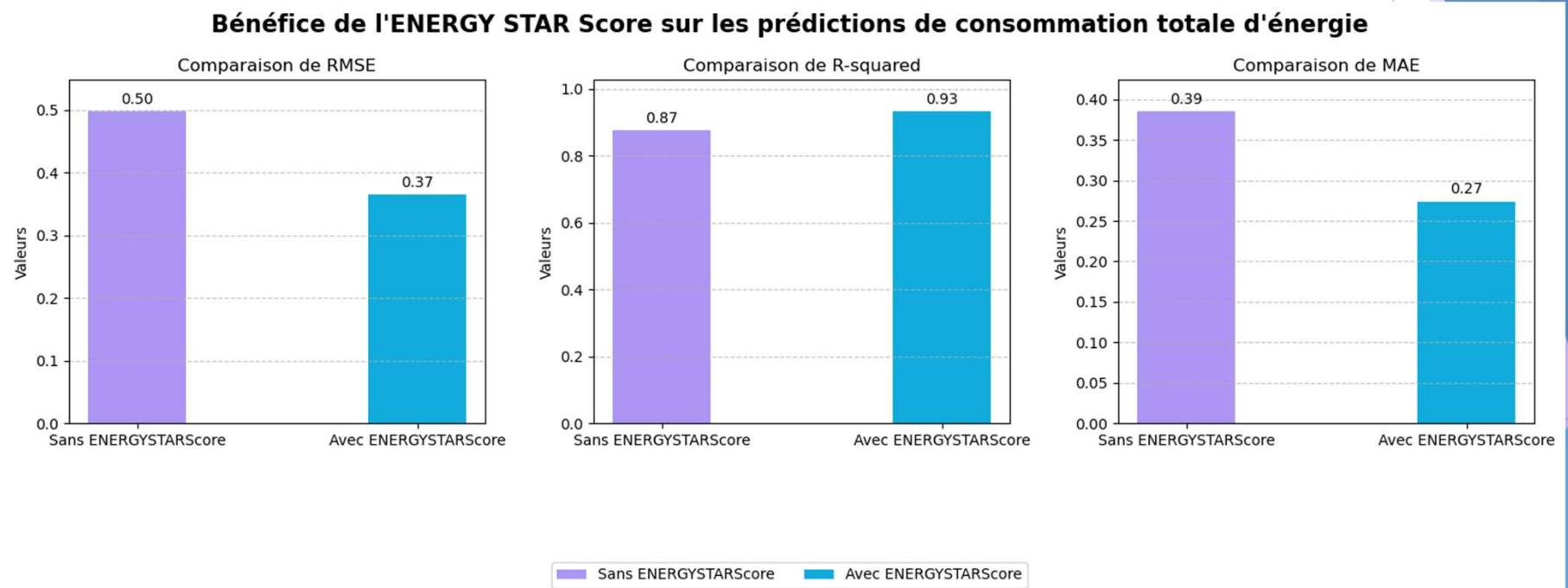


Modèles de prédition des émissions de GES



7. Modélisation : Modèles de prédiction des émissions de GES

- Impact de l'Energy Star Score



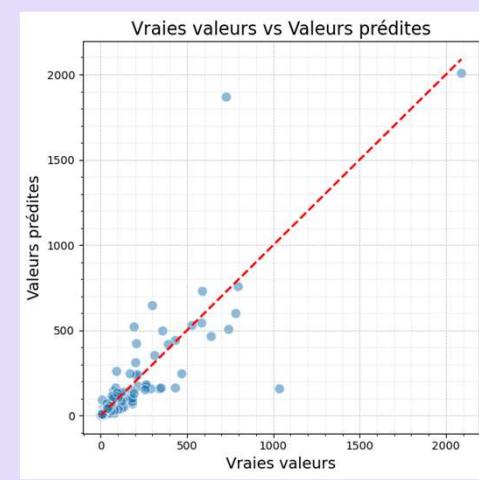
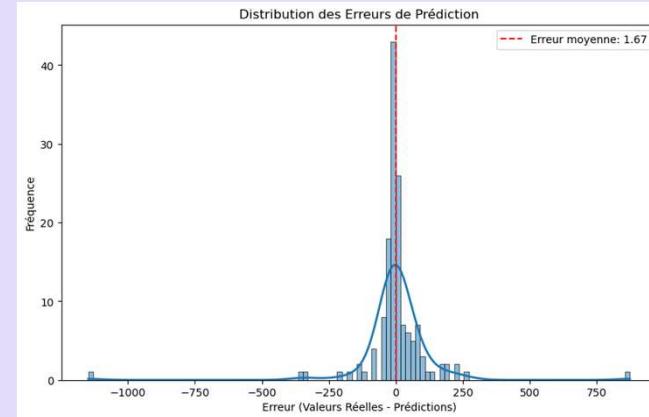
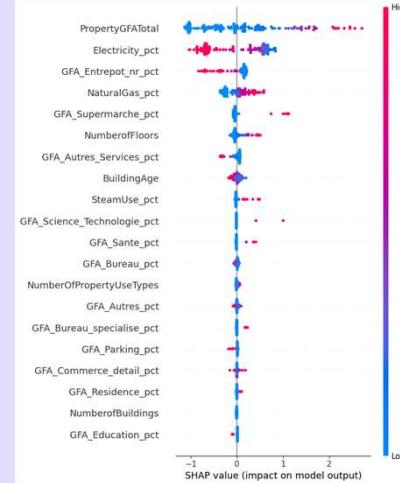
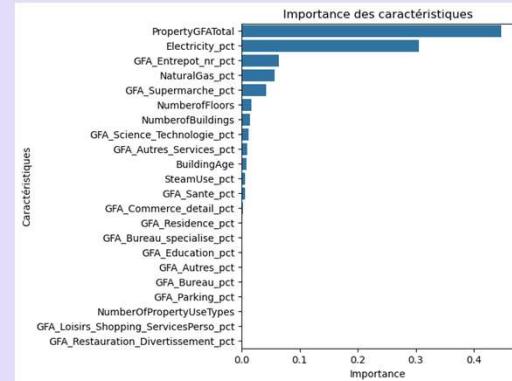
7. Modélisation :

Modèles de prédition des émissions de GES

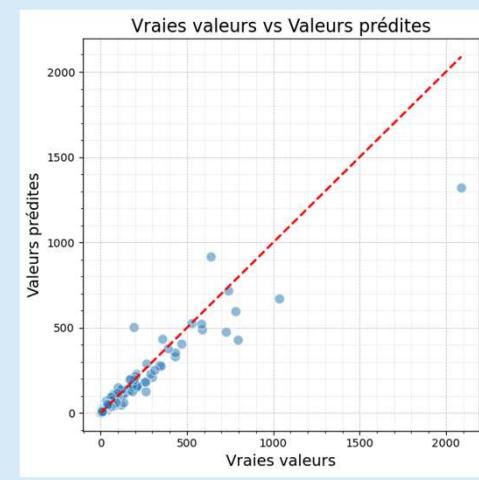
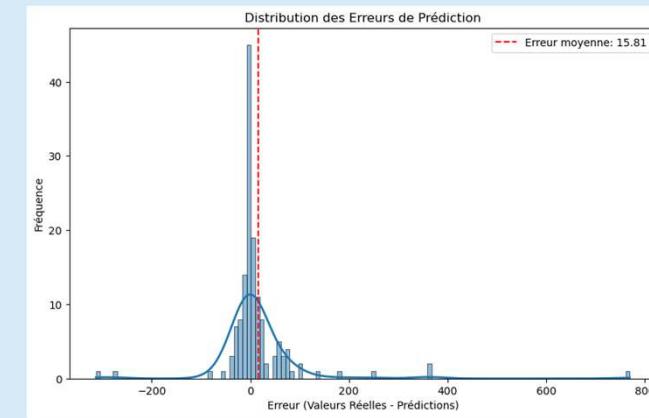
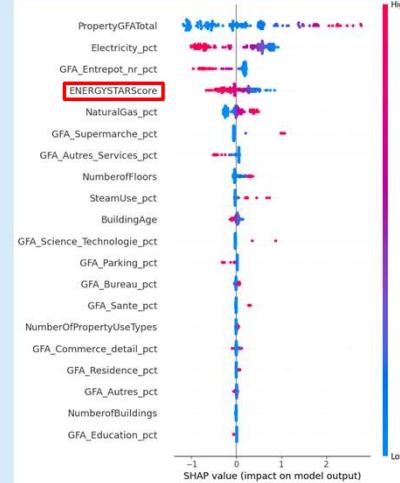
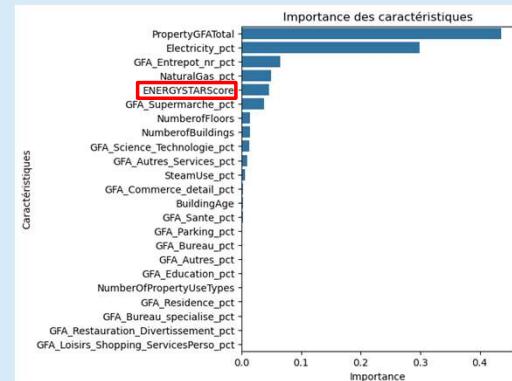
- Importance des variables

- Analyse des erreurs

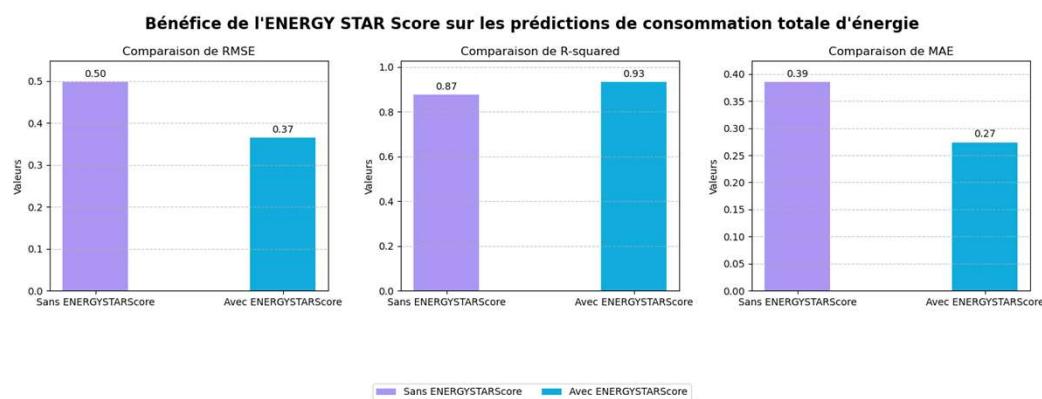
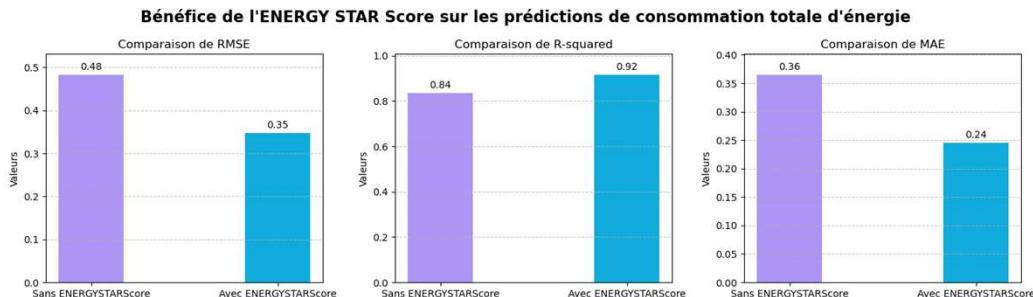
Sans l'Energy Star Score



Avec l'Energy Star Score



8. Résultats et interprétation



- Modèle de prédiction énergétique
 - RMSE : - 27%
 - R^2 : + 8%
 - MAE : - 33,3%

- Modèle de prédiction GES
 - RMSE : - 26%
 - R^2 : + 6%
 - MAE : - 30,8%

Conclusion et Recommendations

• Synthèse des Résultats

- Amélioration significative des prédictions avec l'intégration de l'ENERGY STAR Score.
- Importance confirmée de l'ENERGY STAR Score dans la performance énergétique et les émissions de GES.

• Prochaines Étapes

- Évaluer le coût-bénéfice de l'utilisation de l'ENERGY STAR Score.

• Recommandations pour Améliorations

- Intégrer des données plus récentes et étendre l'analyse à plus de bâtiments.

• Perspectives pour la ville de Seattle vers 2050

- Utiliser les modèles prédictifs pour orienter les politiques d'efficacité énergétique.
- Mettre en place un suivi continu pour évaluer les progrès vers la neutralité carbone.



Merci pour votre attention

Des questions ?

