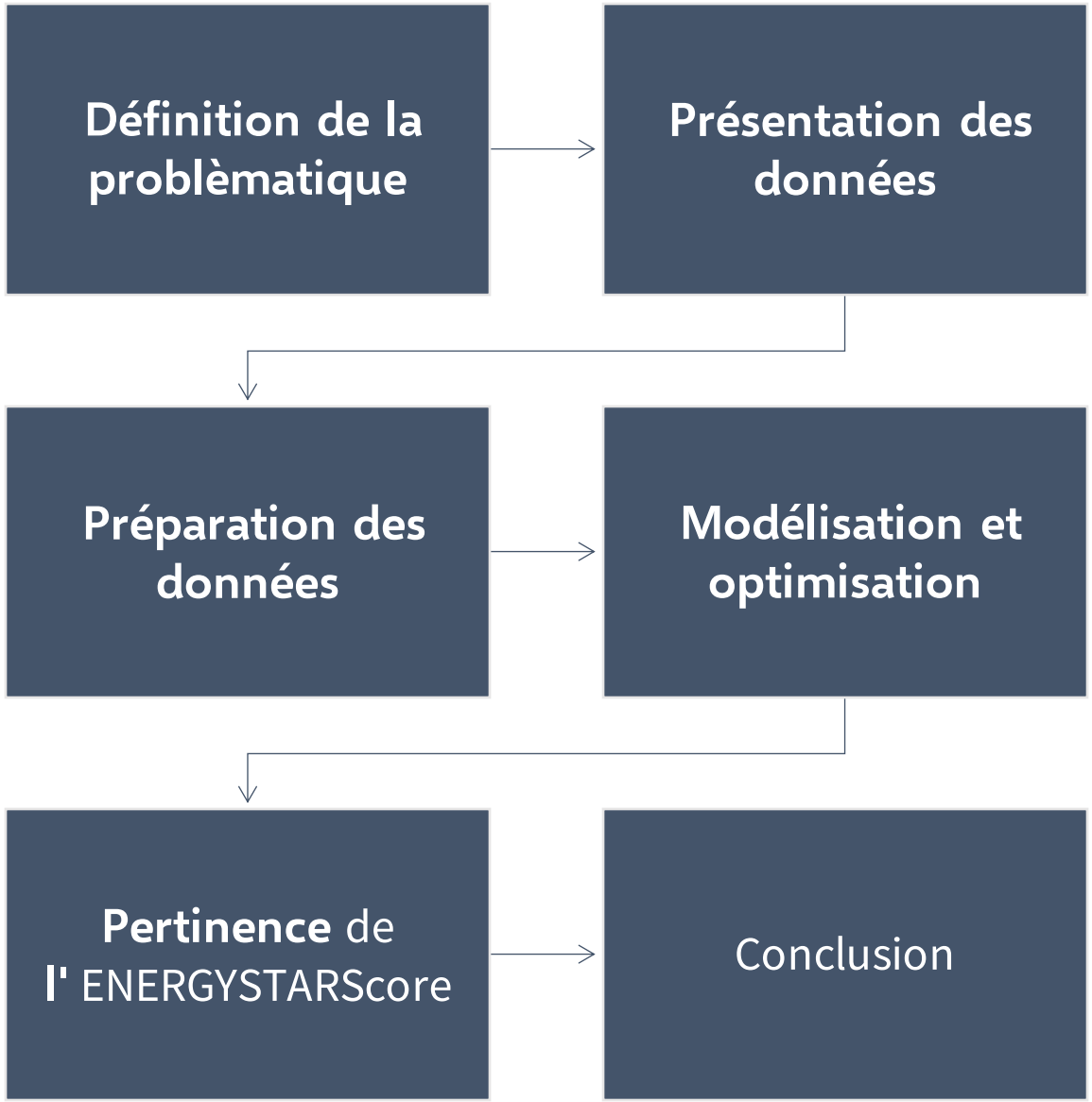
Anticipez les besoins en consommation de bâtiments

Réalisée par
Mounira Abderrahmani

Examinateur
Slim Ouerghi

Plan :


Définition de la problèmatique
Présentation des données
Préparation des données
Modélisation et optimisation
Pertinence de l' ENERGYSTARScore
Conclusion

Contexte :

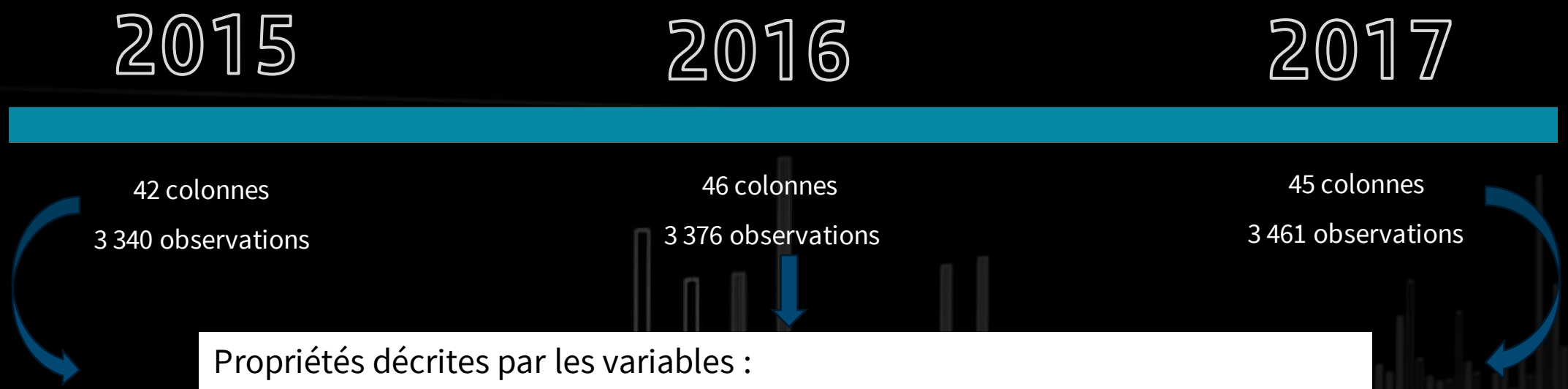
- Objectif : une ville neutre en émission de carbone en 2050
- Problématique : des relevés minutieux réalisés mais coûteux à obtenir

I. Objectif de la démarche :

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

II. Présentation des données :

Source : Seattle Seattle Open Data



Propriétés décrites par les variables :

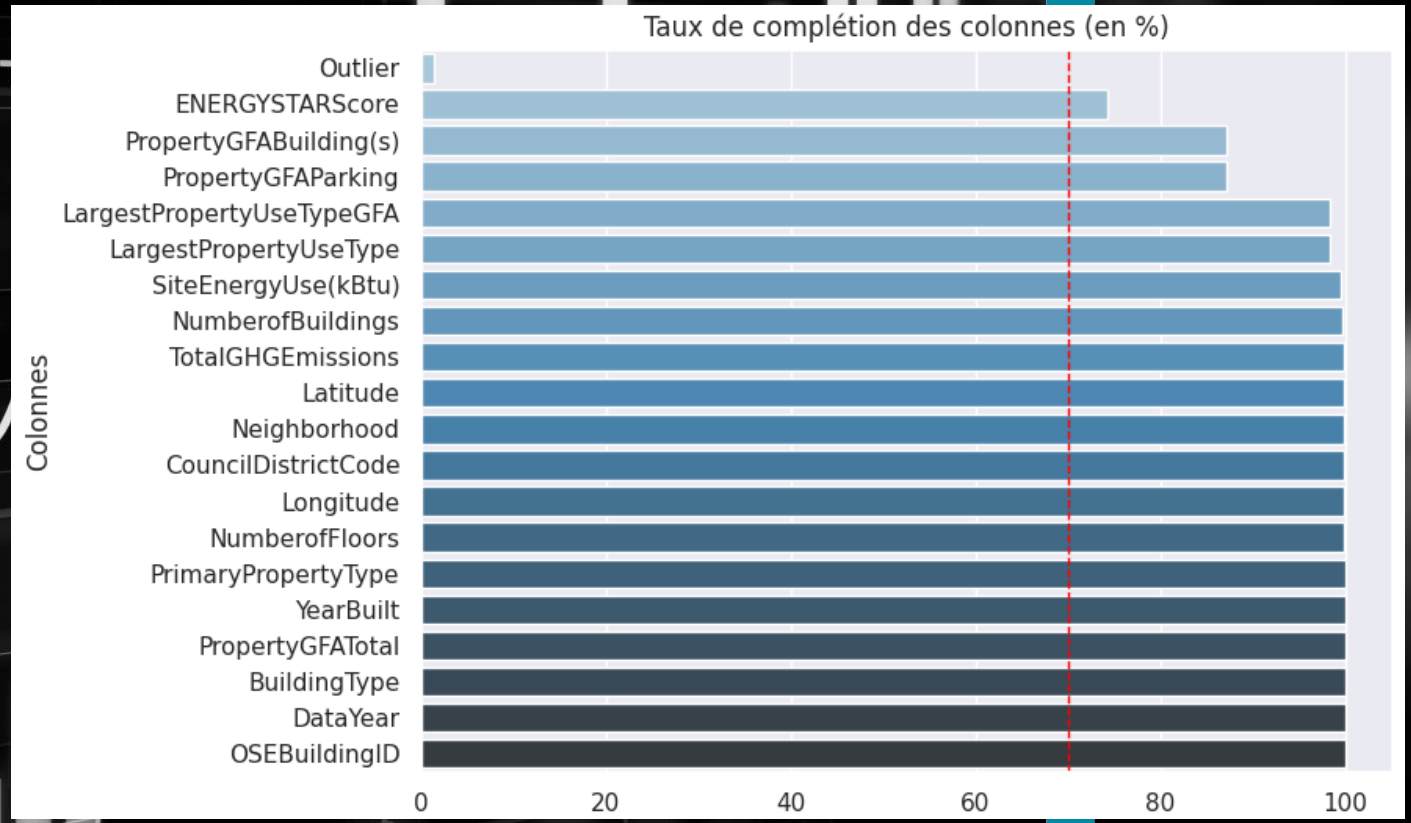
- **Géographiques** (longitude, latitude, …)
- **Architecturales** (YearBuilt, NumberofFloors, ...)
- **Usage** (BuildingType, PrimaprPrepertyType, ...)
- **Energétiques et émissions** (Electricity(kBtu), TotalGHGEmissions, …)

II. Présentation des données

1. Prétraitement :

- Uniformisation des noms des colonnes : données de 2015, 2016 et 2017
- Suppression des colonnes avec 50% de Nan
- Choix des features pertinentes : 20 variables
- Identification des variables cibles pour les prédictions : SiteEnergyUse(KBtu) et TotalGHGEmissions

10 177 observations
20 variables

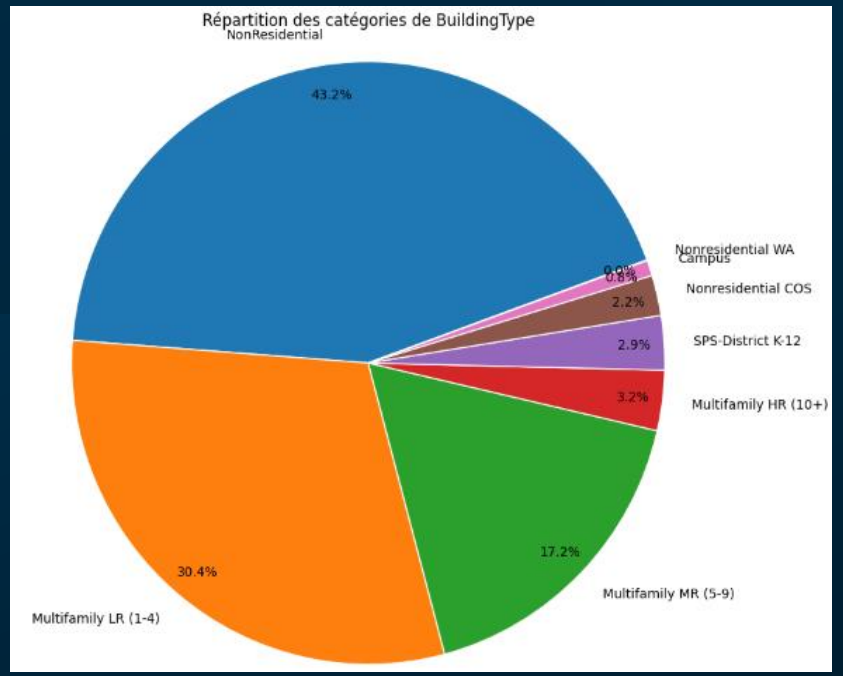


II. Présentation des données

2. Nettoyage :

Variables catégorielles

- Uniformisation des catégories
- Suppression des bâtiments résidentiels



4471 observations
20 variables

Variables numériques

- Suppression des lignes avec trop de Nan
- Suppression des observations négatives

OSEBuildingID	1.00000
DataYear	2015.00000
Latitude	47.49917
Longitude	-122.41182
YearBuilt	1900.00000
NumberofBuildings	0.00000
NumberofFloors	0.00000
PropertyGFATotal	11285.00000
PropertyGFAParking	-2.00000
PropertyGFABuilding(s)	-50550.00000
LargestPropertyUseTypeGFA	5656.00000
ENERGYSTARScore	1.00000
SiteEnergyUse(kBtu)	0.00000
TotalGHGEmissions	-0.80000
BuildingAge	1.00000

dtype: float64

III. Préparation des données

Analyse exploratoire

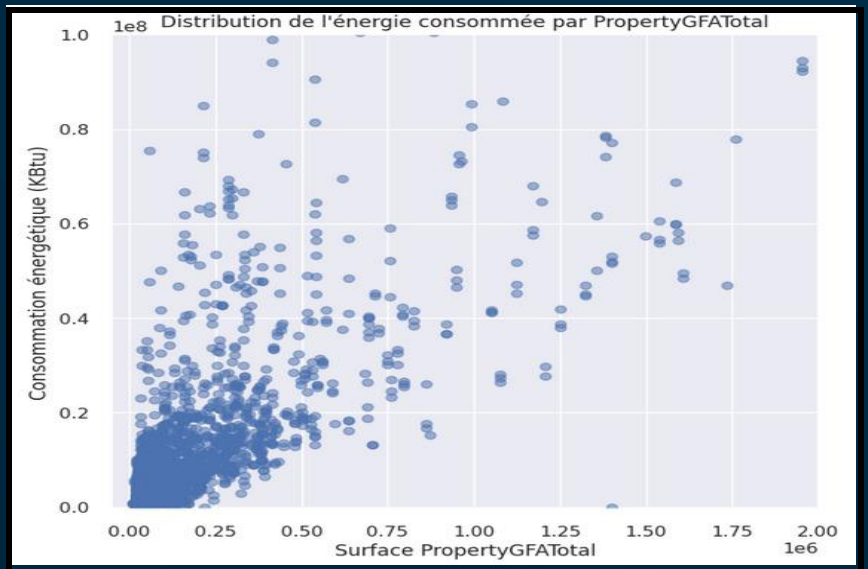
- Analyse des corrélations
- Analyse des distributions des targets

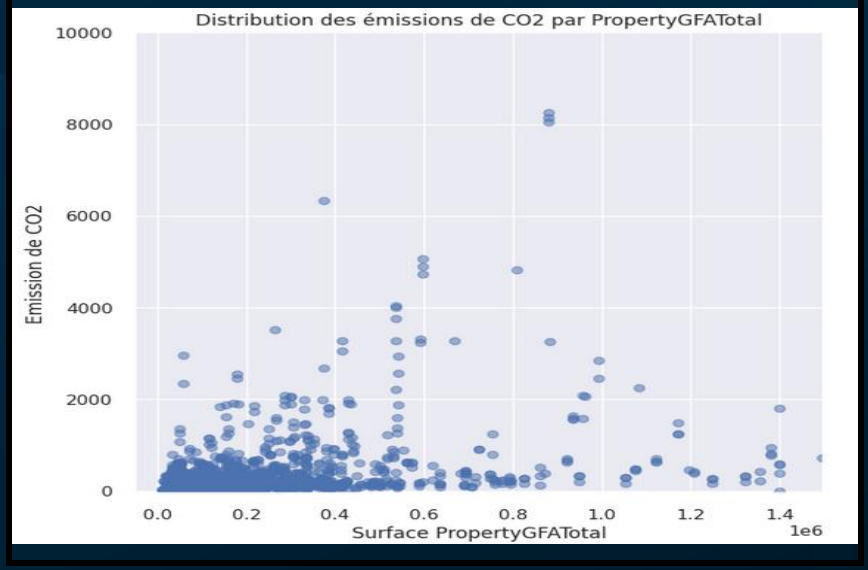
Feature engineering

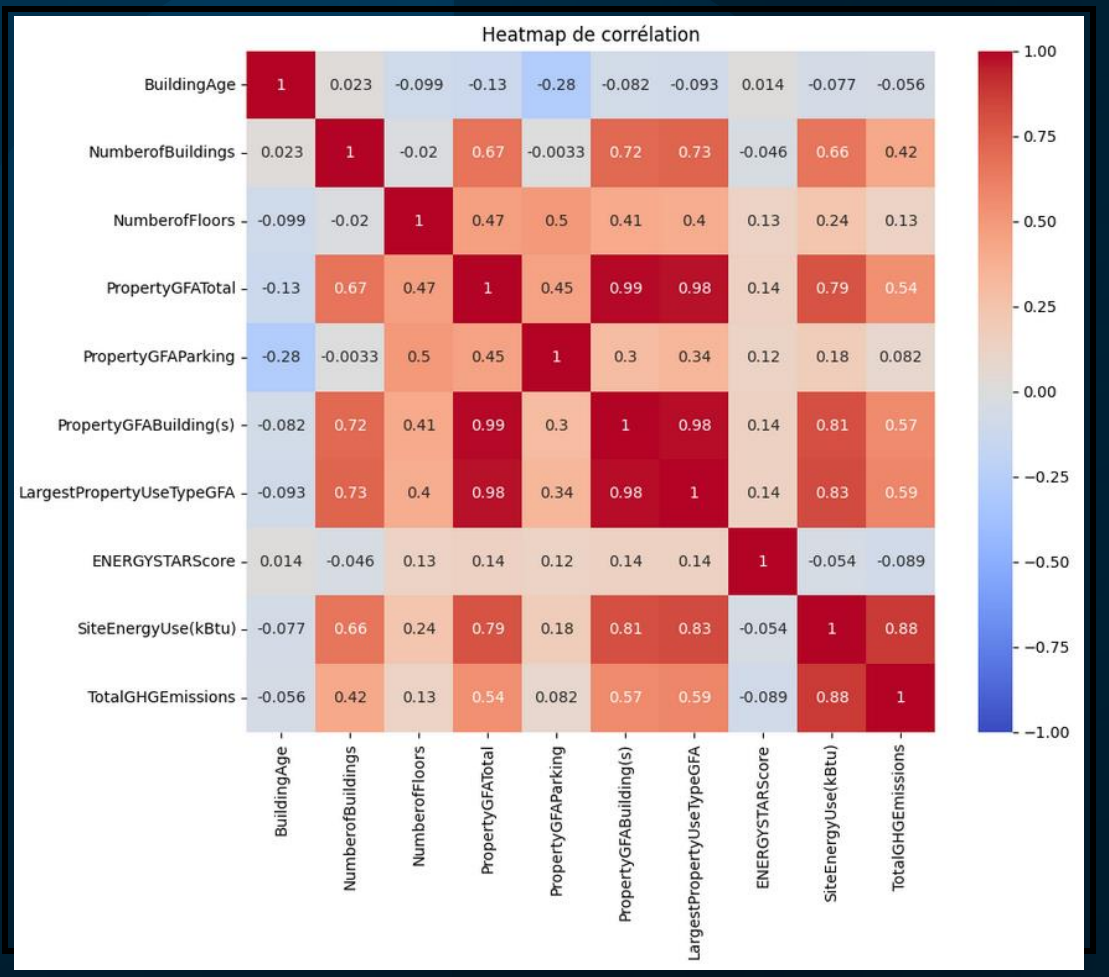
- Création de nouvelles variables
- Transformation des targets : Log

III. Préparation des données

1. Analyse exploratoire





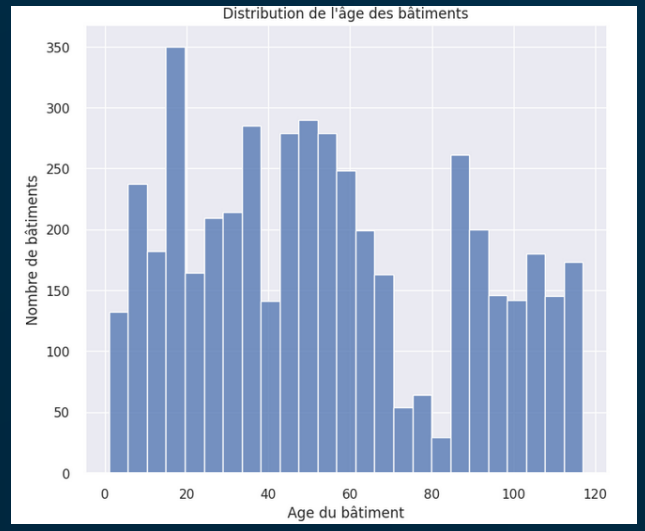


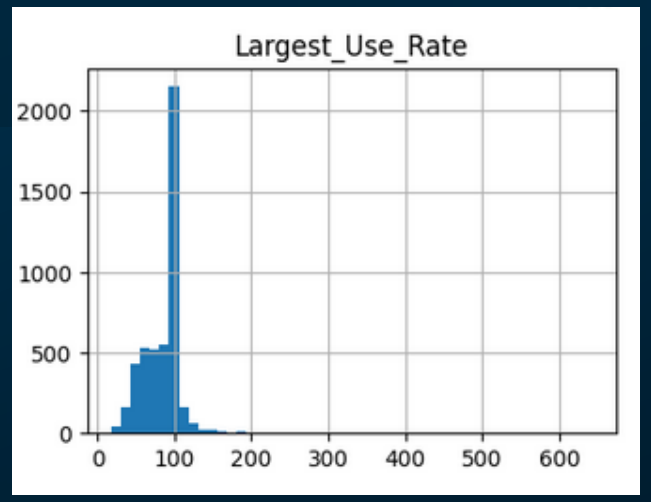
- Corrélation entre les targets et la surface totale
- Forte corrélation de l'énergie avec les émissions

III. Préparation des données

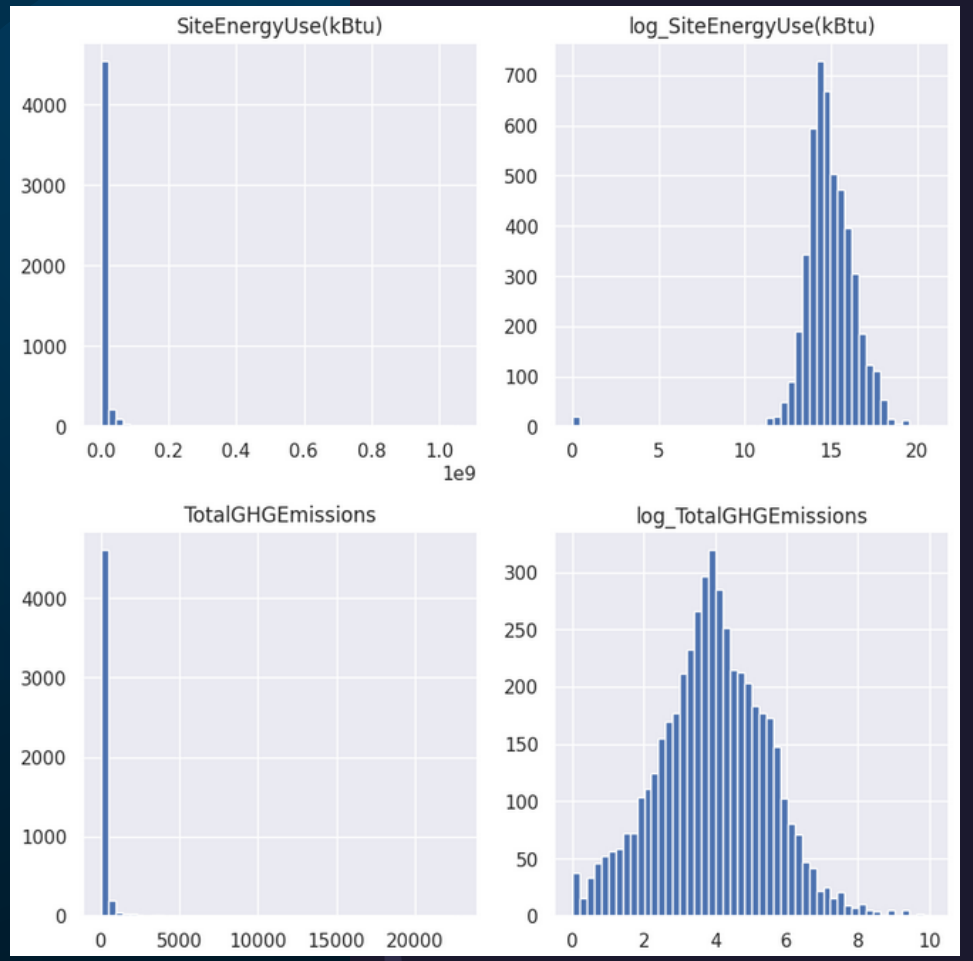
2. Feature engineering :

Building_Age -> suppression de DataYear et Year_Built





Proportion de LargestPropertyUseType



Visualisation des cibles transformées au log

IV. Modélisation et optimisation

IV. Modélisation et optimisation

1. Preprocessing

- Séparation des données : entraînement, test
- Encodage: TargetEncoder
- Feature selection : RFECV

2. Pipeline

- Scaler : RobustScaler
- Modèle : linéaire, non linéaire, ensembliste

3. Evaluation

- Validation croisée
- GridSearchCV,
- RandomizedSearchCV

IV. Modélisation et optimisation

Preprocessing

Train test split → **TargetEncoder** → **Recursive feature elimination**

```
Données d'entraînement :  (3758, 13)
Données de test :  (940, 13)
```

Train_test_plit
(test_size = 0.2)

	BuildingType_encoded	PrimaryPropertyType_encoded	LargestPropertyUseType_encoded	Neighborhood_encoded
1831	14.942503	15.211435	15.177833	14.918621
2174	14.942503	15.211435	15.177833	15.530043
1228	14.942503	15.234731	15.144186	15.005398
3411	14.942503	15.211435	15.177833	15.530043
4183	14.942503	14.892388	14.674313	14.749136

Variable cible :
- SiteEnergyUse(kBtu)

Modèle utilisé :
régression linéaire

Features Sélectionnées par la RFECV :

	Feature	Ranking	Selected
0	BuildingType	1	True
1	PrimaryPropertyType	1	True
2	LargestPropertyUseType	1	True
3	Neighborhood	1	True
4	CouncilDistrictCode	1	True
5	BuildingAge	3	False
6	Latitude	1	True
7	Longitude	1	True
8	NumberofBuildings	1	True
9	NumberofFloors	2	False
10	PropertyGFATotal	1	True
11	PropertyGFABuilding(s)	1	True
12	LargestPropertyUseTypeGFA	5	False
13	PropertyGFAParking	1	True
14	Building_Rate	1	True
15	Largest_Use_Rate	4	False
16	Parking_Rate	1	True

IV. Modélisation et optimisation

Modèles :

Linéaires

- Régression linéaire
- Ridge
- Lasso
- ElasticNet

Non linéaires

- KernelRidge
- SVR

Ensemblistes

- RandomForest
- XGBoost

Prédiction

- Consommation totale d'énergie
- Emissions de CO2

Métriques

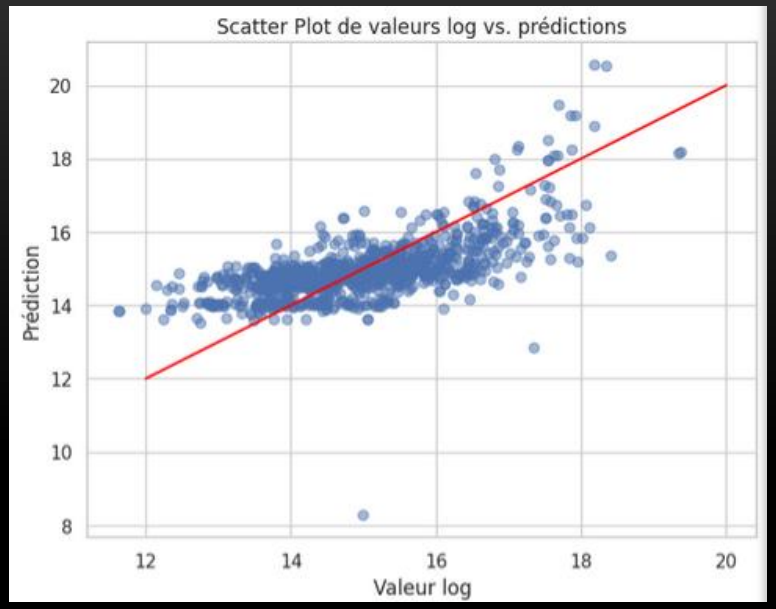
- R2 (score de détermination)
- RMSE (écart quadratique moyen) — score
- RMSE relatif
- MAE (erreur absolue moyenne)
- MeAE (erreur absolue médiane)
- Time (entraînement, prédiction)

IV. Modélisation et optimisation

1. Consommation d'énergie

Régression linéaire

Avec transformation Log



Scores avec la cible transformée en log puis exp

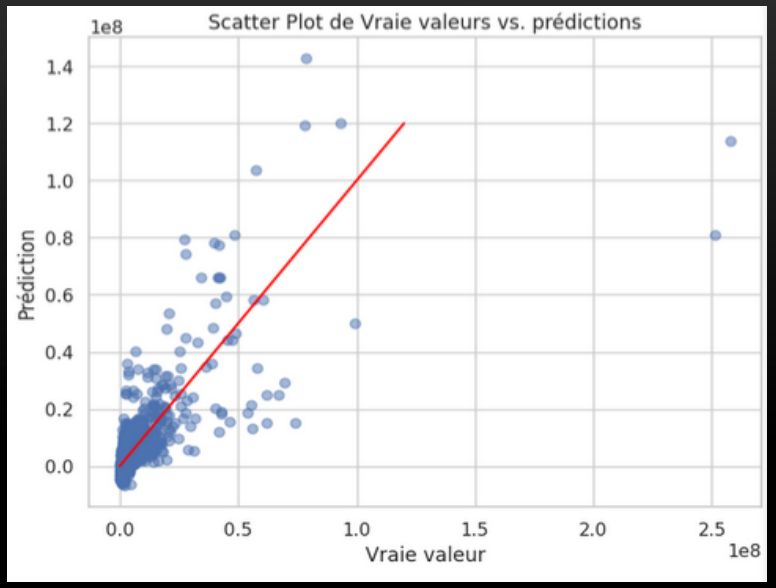
	Training scores	Test scores log	Test scores
RMSE	0.902888	0.931081	3.836384e+07
MAE	0.702853	0.728584	6.772472e+06
R2	0.486995	0.424815	-4.693841e+00
Median Abs Err	0.580556	0.619739	1.459816e+06
Time	0.017231	0.000485	4.851818e-04

Prédiction en log

Passage à exp

- RMSE, MAE augmentent
- Median augmente
- R2 négatif

Sans transformation Log



Scores avec cible non transformée

	Training scores	Test scores
R2	7.068392e-01	5.015082e-01
RMSE	1.728966e+07	1.135137e+07
Relative RMSE	1.632724e+00	4.401329e+00
MAE	5.727079e+06	5.101810e+06
Median Abs Err	2.688634e+06	2.672325e+06
Time	1.605606e-02	5.786419e-04

Prédiction valeur réelle

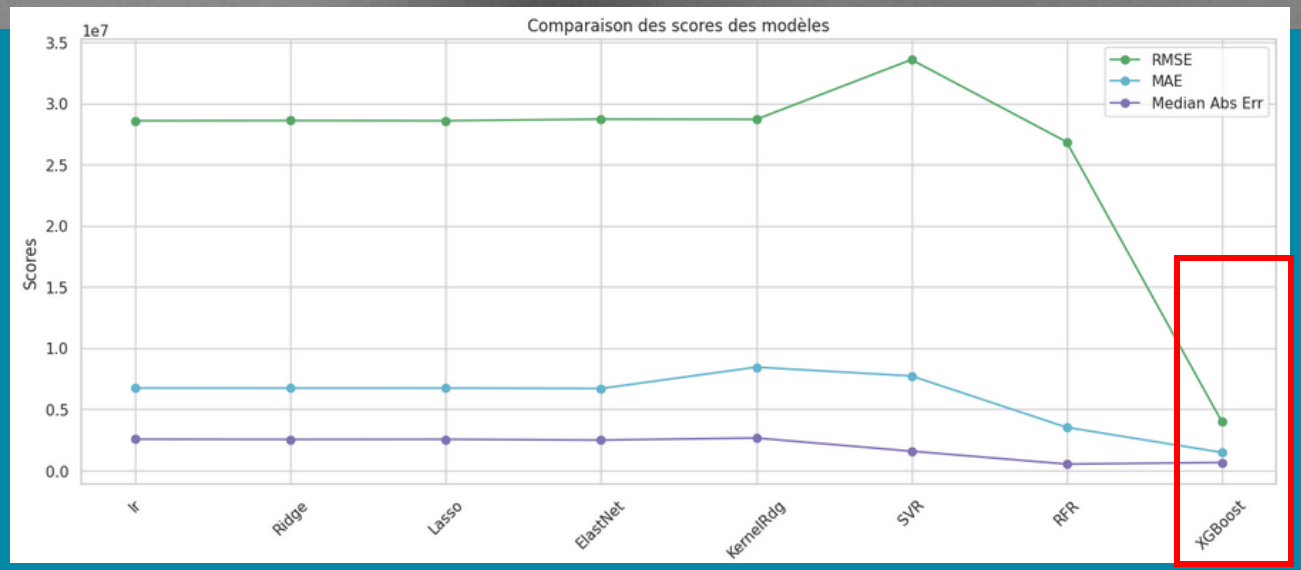
- RMSE, MAE diminuent
- Median augmente
- R2 positif

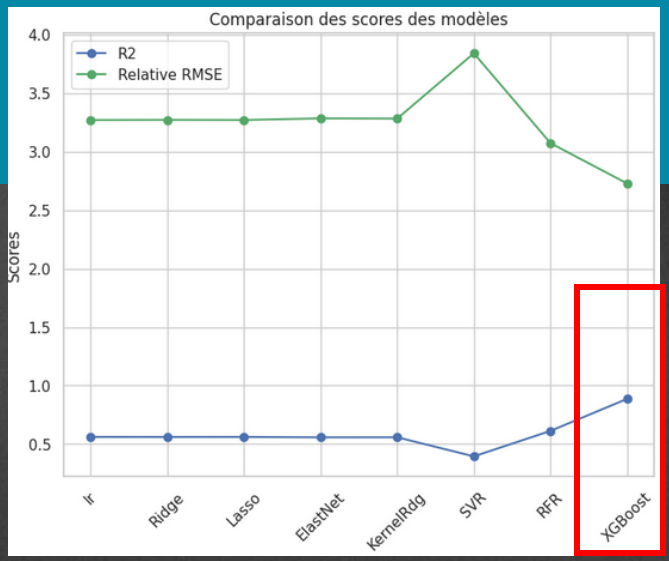
Prédire des valeurs réelles

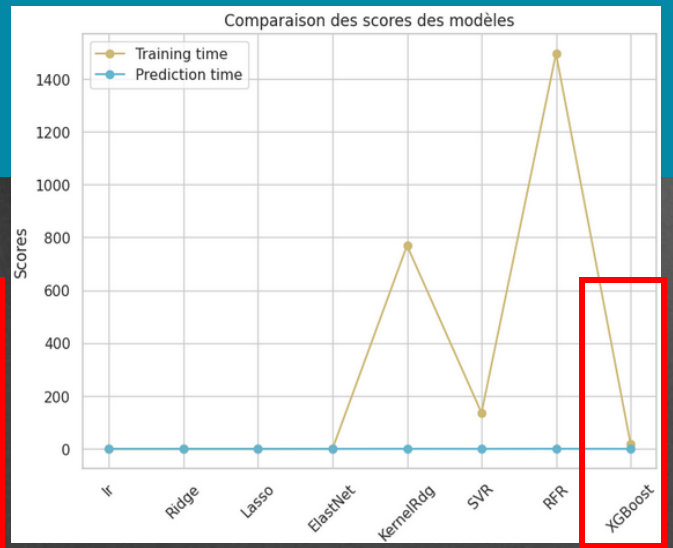
IV. Modélisation et optimisation

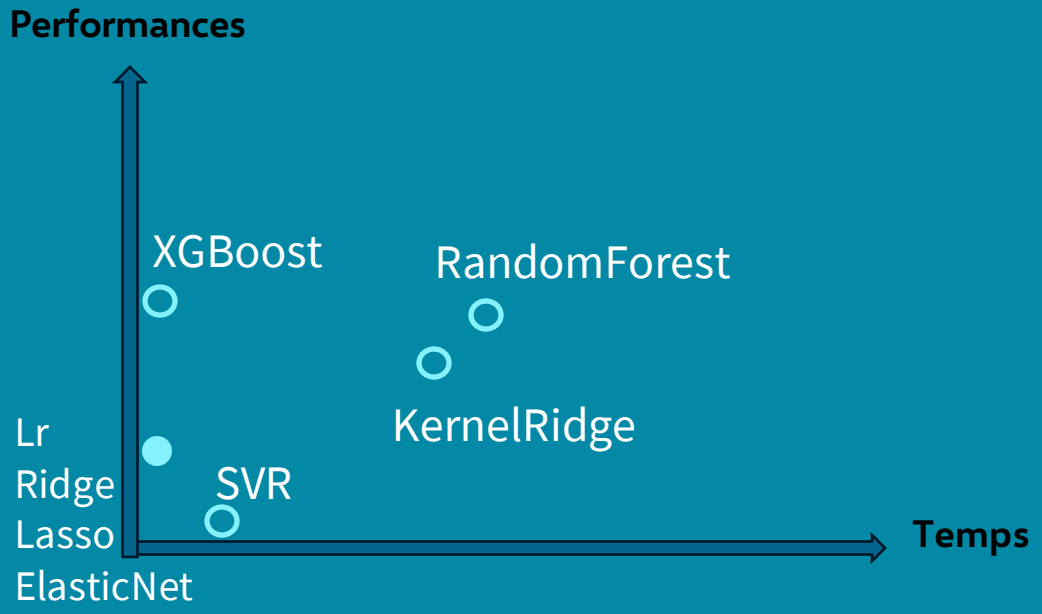
1. Consommation d'énergie

Comparaison des scores des modèles









Modèle retenu pour l'énergie : XGBoost

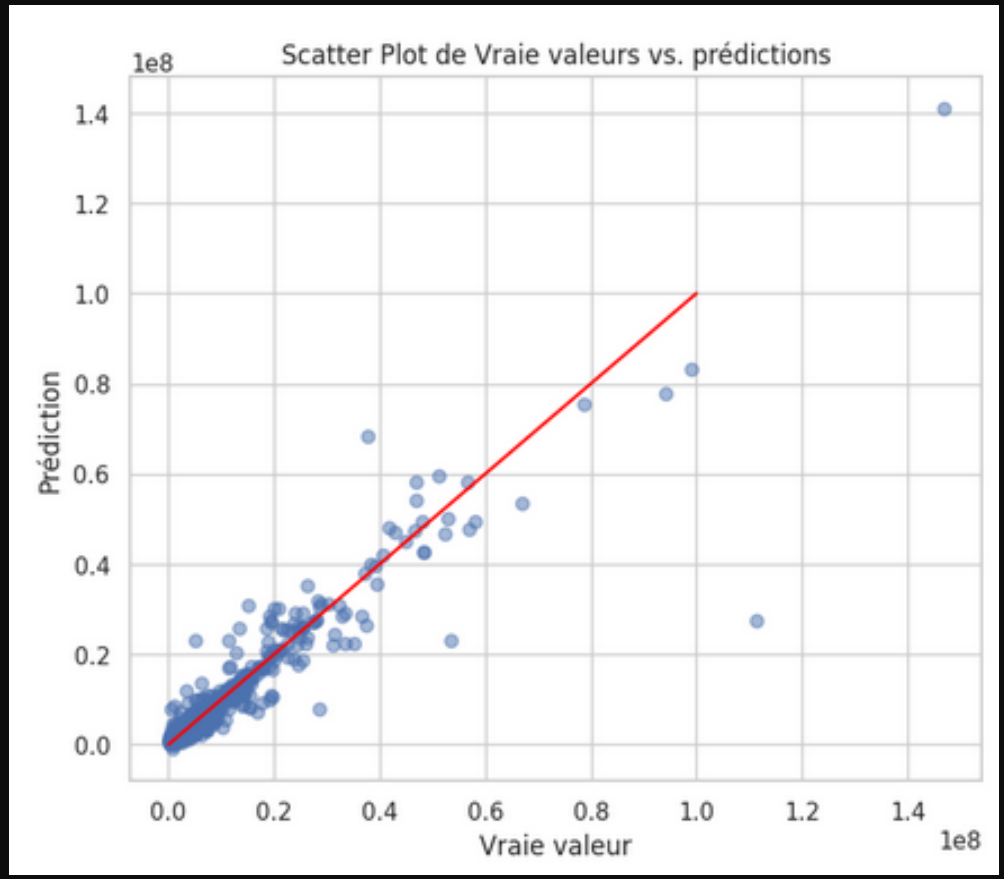
IV. Modélisation et optimisation

1. Consommation d'énergie

Modèle final : XGBRegressor

Modèle final : XGBRegressor
Meilleurs paramètres : n_estimators=300, max_depth=5, learning_rate=0.2

	Gridsearch scores	Test scores
R2	8.365303e-01	8.866144e-01
RMSE	1.218480e+07	4.002454e+06
Relative RMSE	1.150654e+00	2.727716e+00
MAE	3.514826e+06	1.476902e+06
Median Abs Err	1.666925e+06	6.522830e+05
Time	4.875211e+01	3.101587e-03

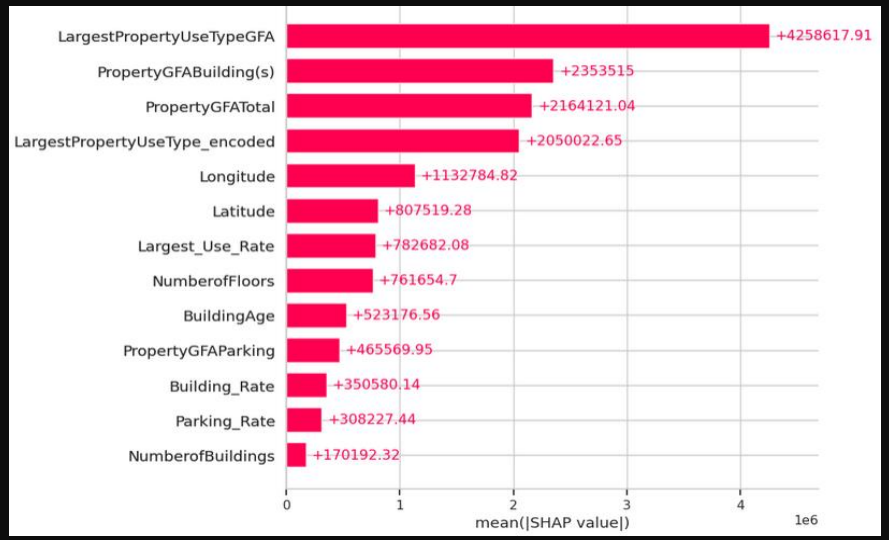


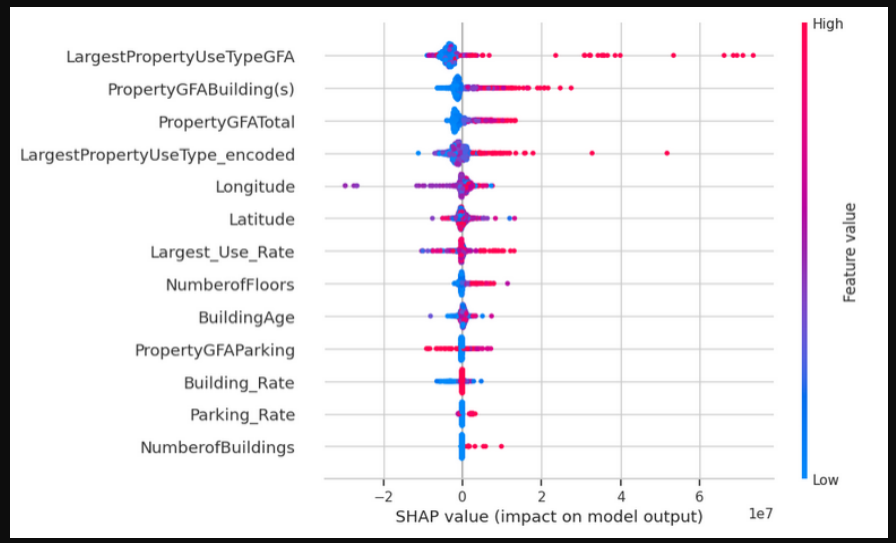
IV. Modélisation et optimisation

1. Consommation d'énergie

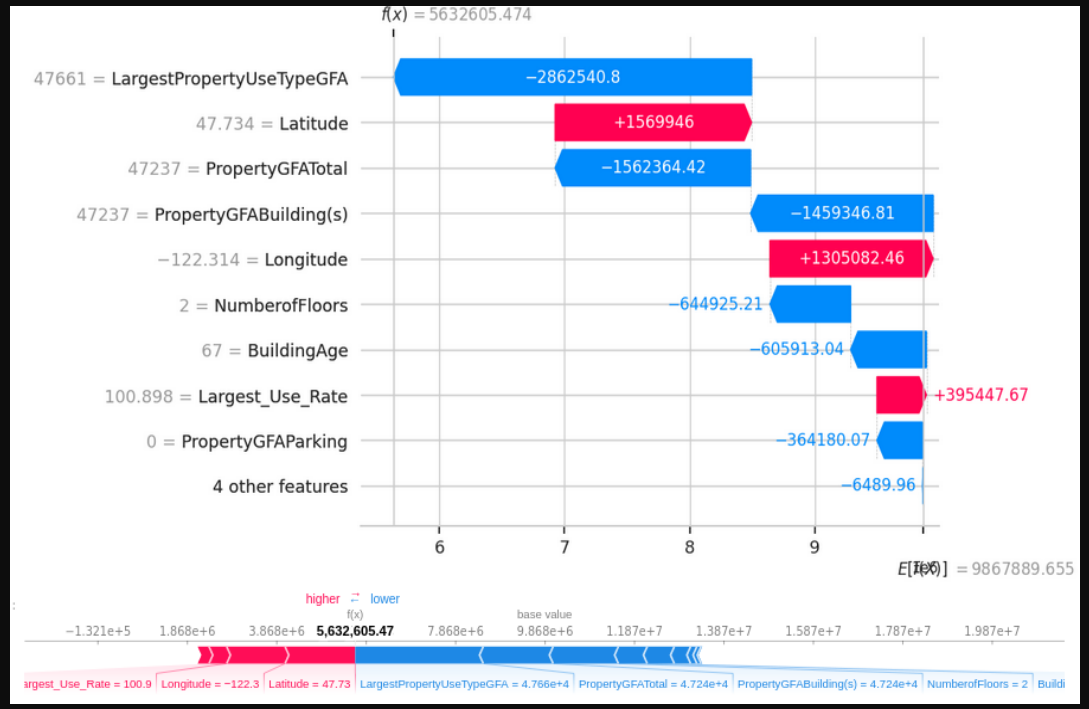
Features importances XGBRegressor

Analyse globale





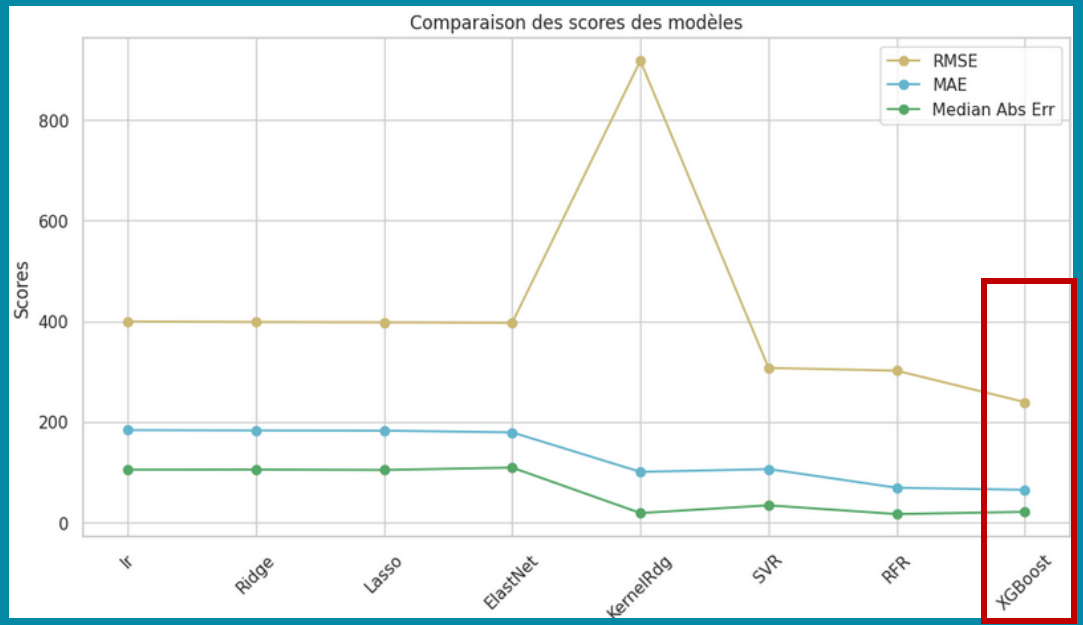
Analyse locale

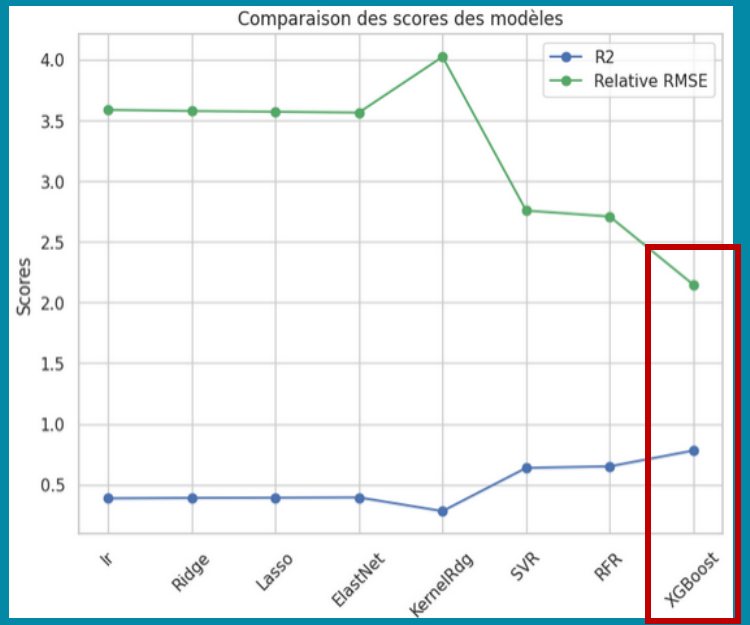


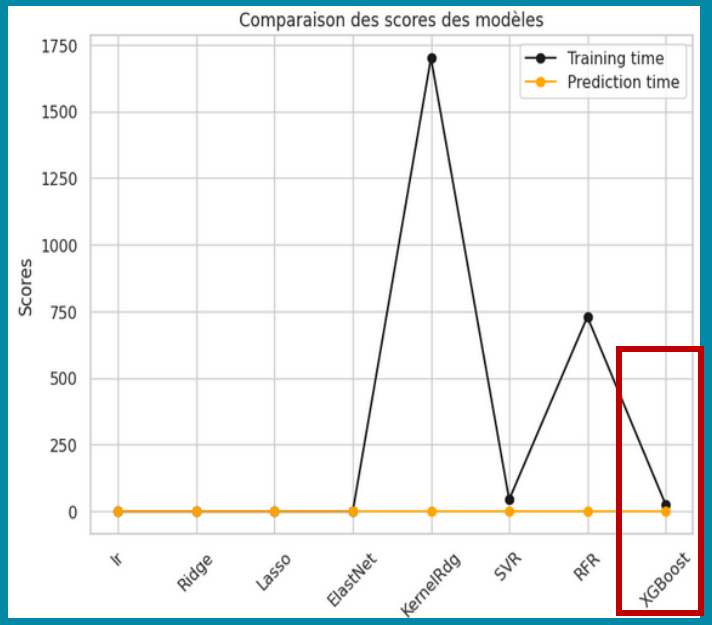
IV. Modélisation et optimisation

2. Emissions de CO2

Comparaison des scores des modèles







Modèle retenu pour les émissions de CO2 : XGBoost

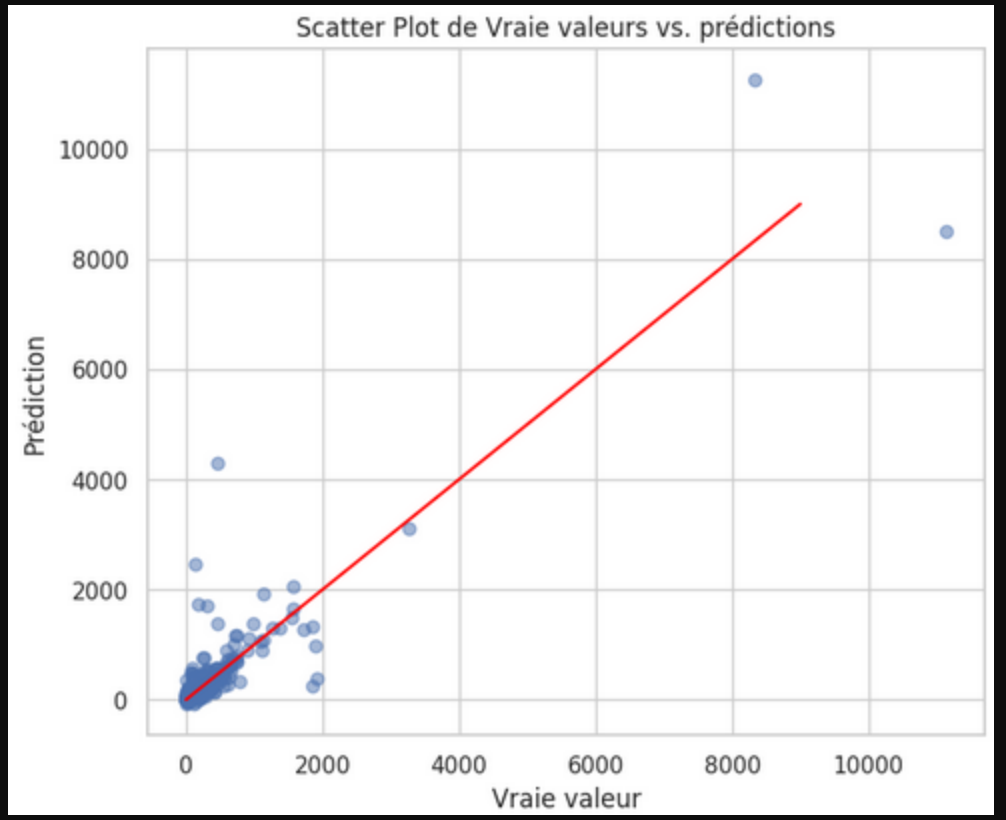
IV. Modélisation et optimisation

2. Emissions de CO2

Modèle final : XGBRegressor

Modèle final : XGBRegressor
Meilleurs paramètres : n_estimators=300, max_depth=5, learning_rate=0.2

	Gridsearch scores	Test scores
R2	0.797959	0.780268
RMSE	350.318134	239.164153
Relative RMSE	1.535638	2.146992
MAE	90.723312	64.843141
Median Abs Err	39.788963	21.126569
Time	26.165442	0.002236

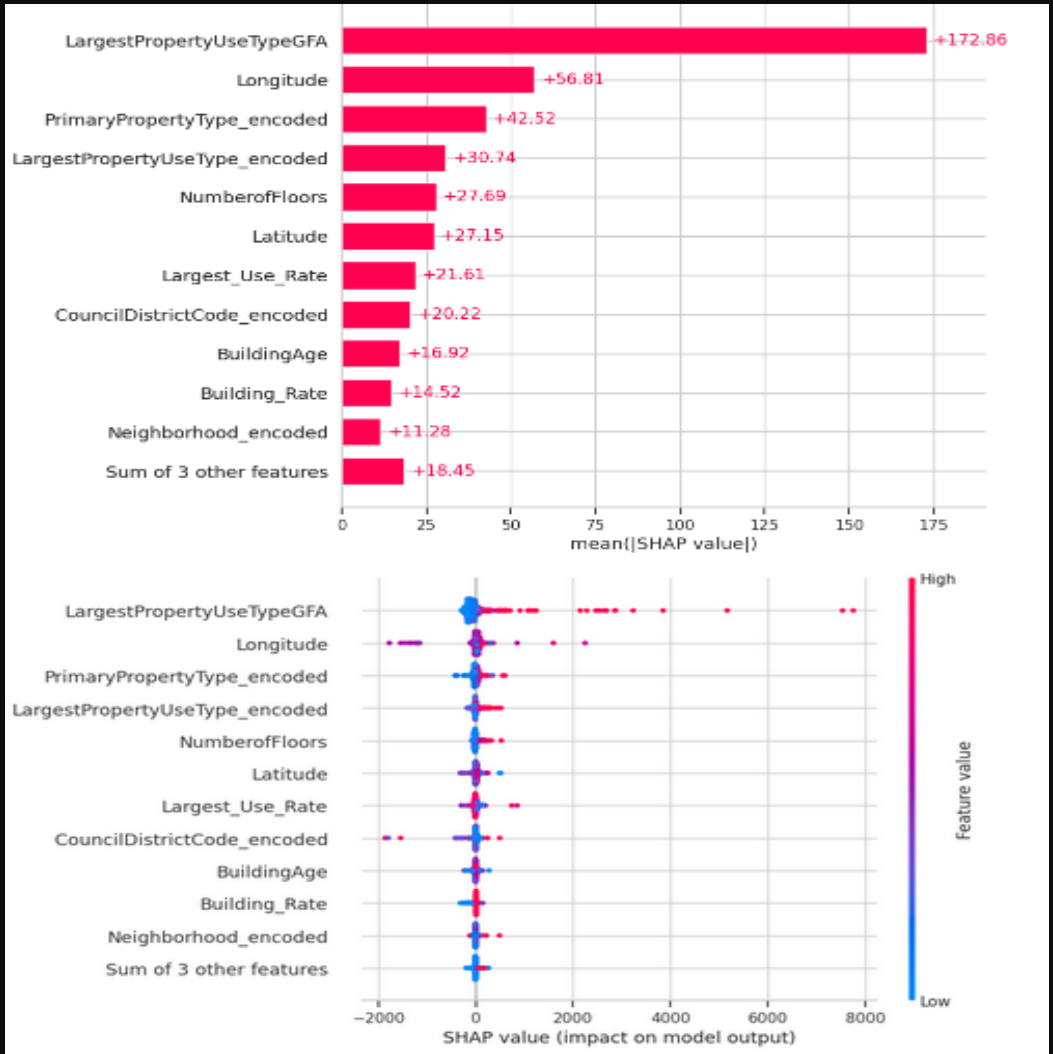


IV. Modélisation et optimisation

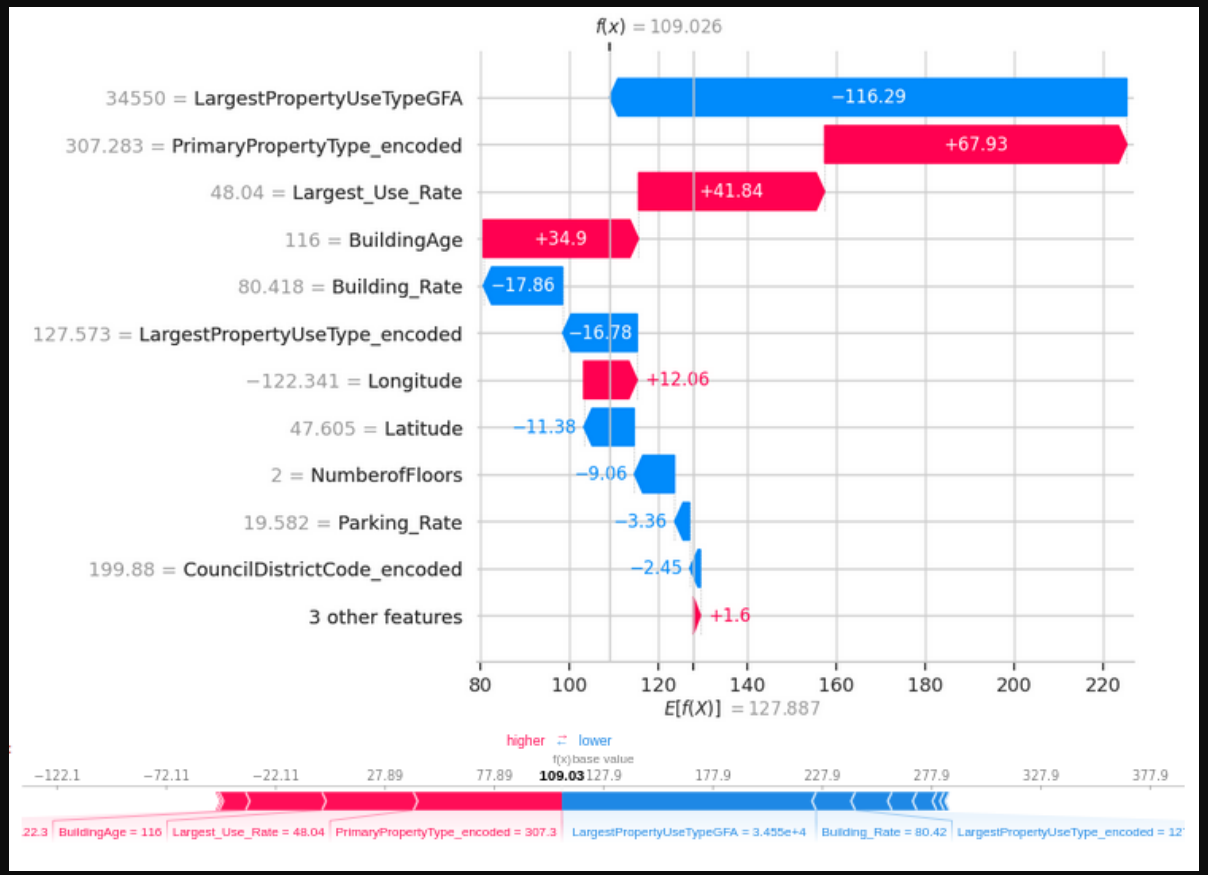
2. Emissions de CO2

Analyse globale

Features importances XGBRegressor



Analyse locale

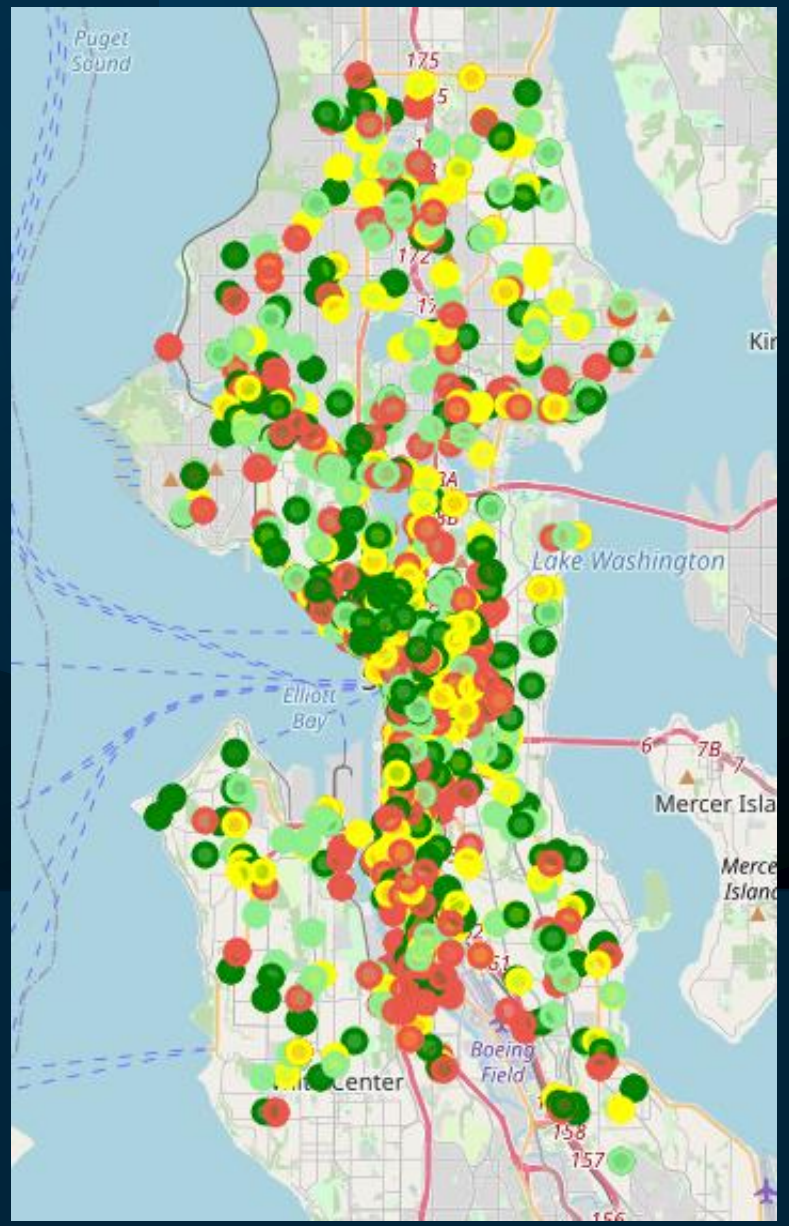


V. Pertinence de ENERGYSTARScore

ENERGYSTARScore

i. Outil d'analyse comparative.
ii. Permet d'évaluer le rendement énergétique des bâtiments commerciaux.

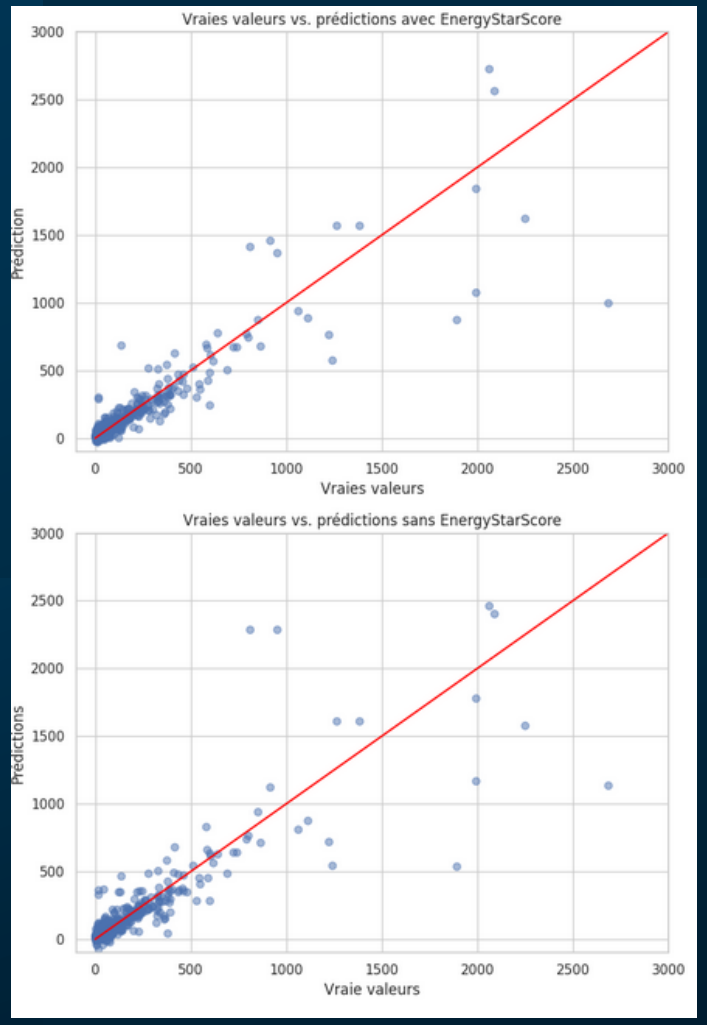
- Variable non utilisée pour la modélisation
- Taux de remplissage ~ 75%
- Réduction du nombre d'observations : 3 126

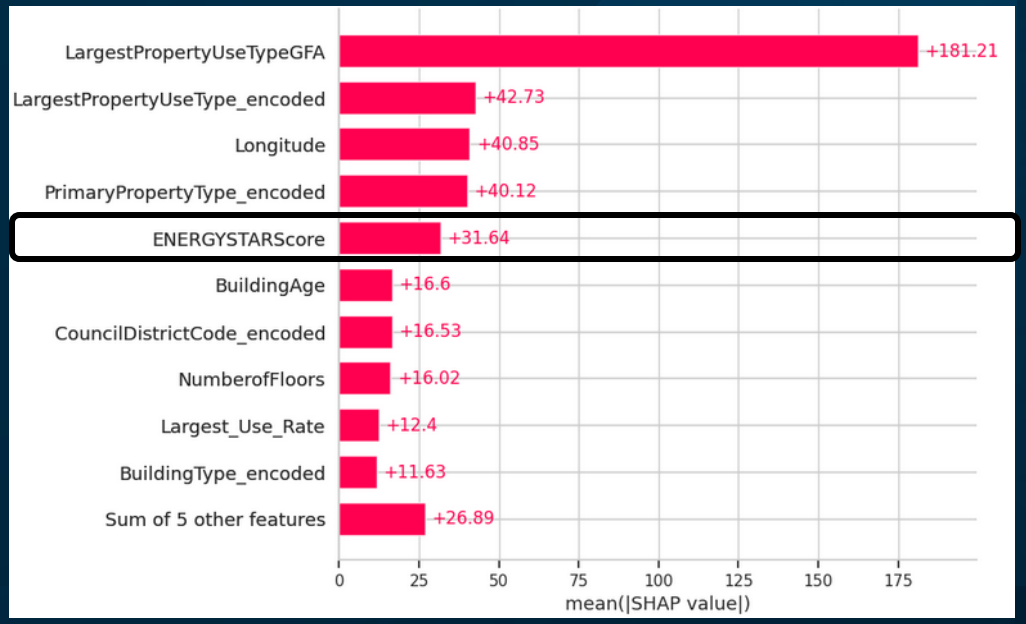


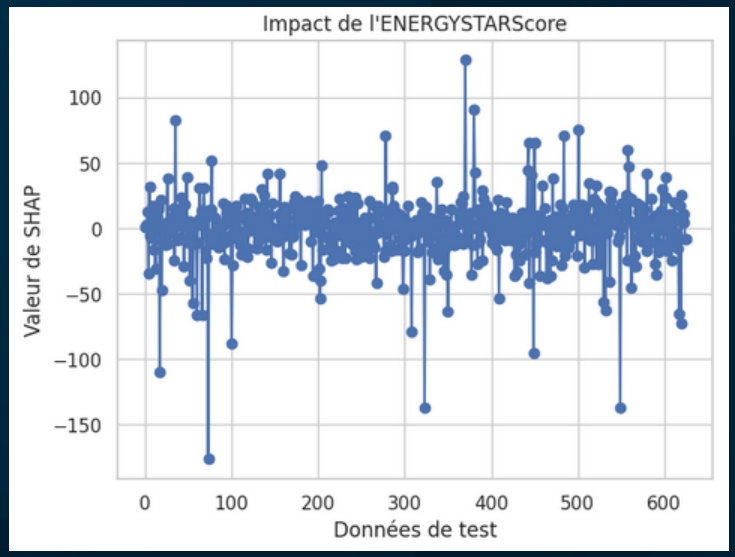
Répartition du EnergyStarScore sur la carte de Seattle

V. Pertinence de ENERGYSTARScore

Prédiction avec XGBRegressor



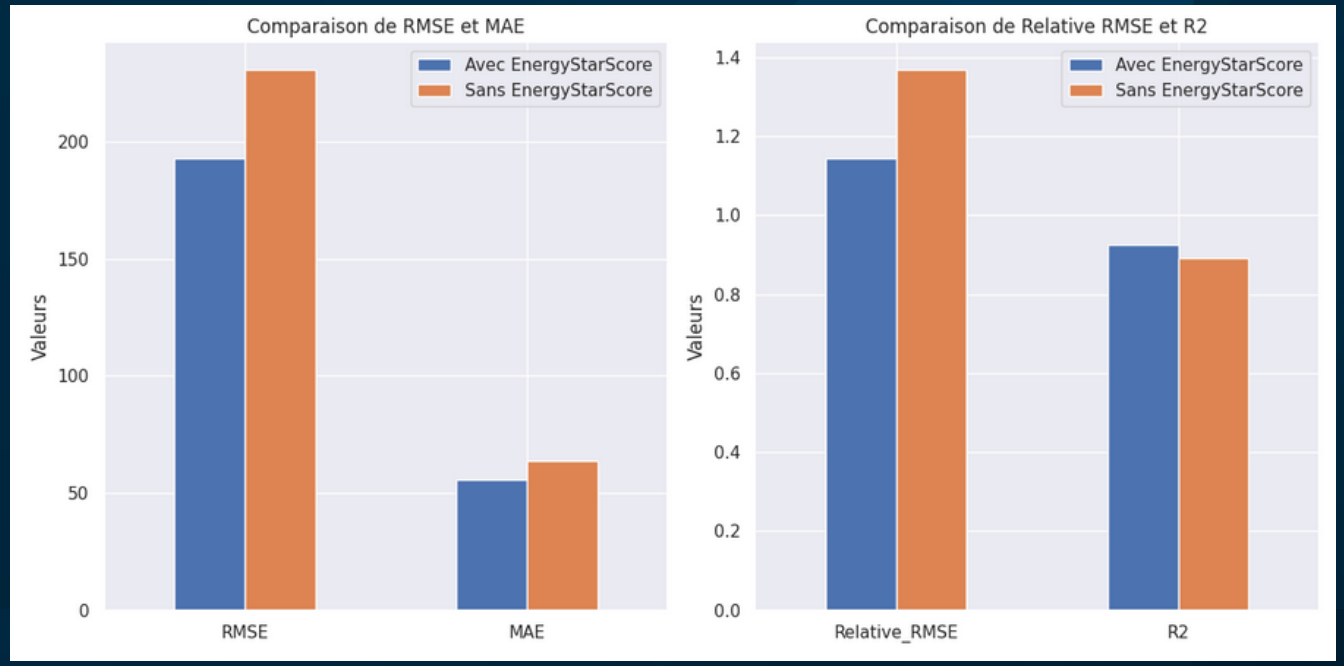




- ENERGYSTARScore importante pour le modèle
- Quel impact sur les précisions ?
- Scores ?

V. Pertinence de ENERGYSTARScore

Comparaisons des scores



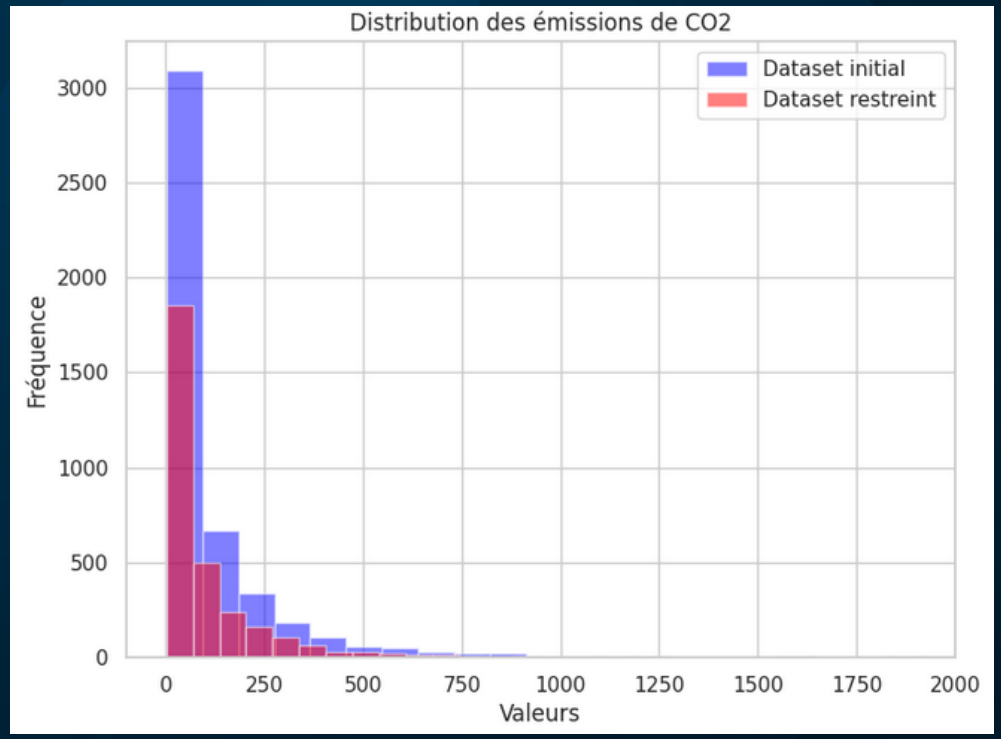
- Meilleurs scores avec la variable ENERGYSTARScore
- Entraînement du modèle sur moins d'observations
- Fiabilité ?

V. Pertinence de ENERGYSTARScore

Fiabilité des scores

Test
Comparaison des distributions
Emissions de CO2

```
Test de Kolmogorov-Smirnov :
----------------------------
Statistique KS = 0.016981663739943642
Valeur de p = 0.6427914928053903
Les distributions des émissins de CO2 sont identiques.
```



Scores fiables
ENARGYSTARScore pertiente pour prédire
les émissions de CO2

VI. Conclusion

- Modèle retenu pour les prédictions : XGBoost
- Les prédictions des **émissions de CO2 moins** précises par rapport aux prédictions de la **consommation énergétique**
- **ENERGYStarScore** semble **améliorer** les prédictions des **émissions de CO2**
- Avec
    i. Un plus grand nombre de données
    ii. Des informations supplémentaires sur les bâtiments (matérieux de construction, isolants, présence de panneaux solaires, …)

Il est possible d'améliorer les performances des modèles.