Anticipez les besoins en consommation de bâtiments

Formation Data Scientist – Projet 4

Octave POUILLOT

Mai 2023

Sommaire

- Mission
- Présentation du jeu de données
- Nettoyage & Exploration
- Modeling
- Intérêt de ENERGY STAR Score
- Conclusions

Mission

Ville de Seattle, relevés des bâtiments non destinés à l'habitation de 2016

- Prédire les émissions de CO2eq de bâtiments
- Prédire la consommation totale d'énergie de bâtiments
- Evaluer l'intérêt de l'"ENERGY STAR Score"

L'objectif étant de se passer des relevés de consommation annuels futurs.

Présentation du Jeu de données

2016 Building Energy Benchmarking

Présentation du Jeu de données

- 3 376 lignes, 46 colonnes
- Chaque ligne est un bâtiment
- Informations structurelles :
  - Années de construction
  - Adresse (ville, état, quartier, GPS, …)
  - Usage (type, principal, secondaire, …)
  - Superficie (Totale, Parking, Etages, …)
- Energétiques :
  - Energy Star Score
  - Energie utilisée (Electricité, Gaz fossile, Vapeur, totale, par surface, …)
  - Emissions (totale, intensité)
- Conformité des données (Statut, outliers, commentaires)

Cibles : SiteEnergyUse(kBtu) & TotalGHGEmissions

Nettoyage et Exploration

Prise en main du jeu de données et préparation aux modélisations

Nettoyage et Exploration – Suppression

ComplianceStatus

- Compliant 3211
- Default Data 113
- Non-Compliant 37
- Missing Data 15

Outliers

- Low Outliers 23
- High Outliers 9

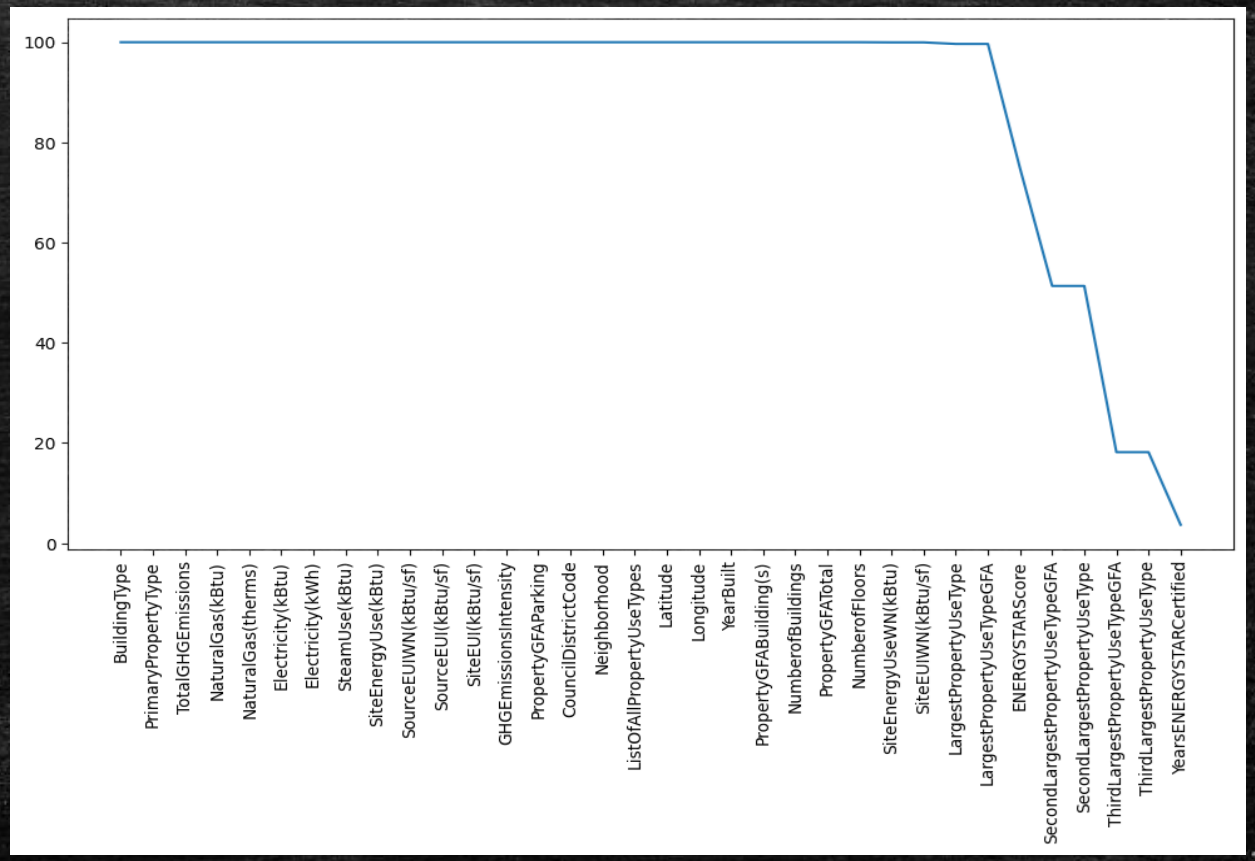
DefaultData

- False 3263
- True 113

- Vérification des doublons (aucun)
- Suppression des « ComplianceStatus » NOK
- Suppression des « Outliers » (plus d'outliers)
- Suppression des NaN sur cibles (4 lignes)

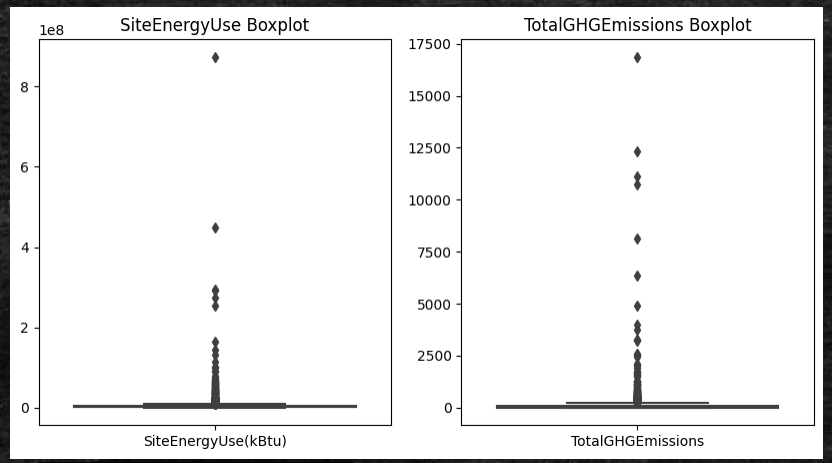
Nettoyage et Exploration – Sélection

- Pré-sélection de features
- Taux de remplissage des features
- Suppression des features < 70%
- ENERGYSTAR Score : 74.8 %



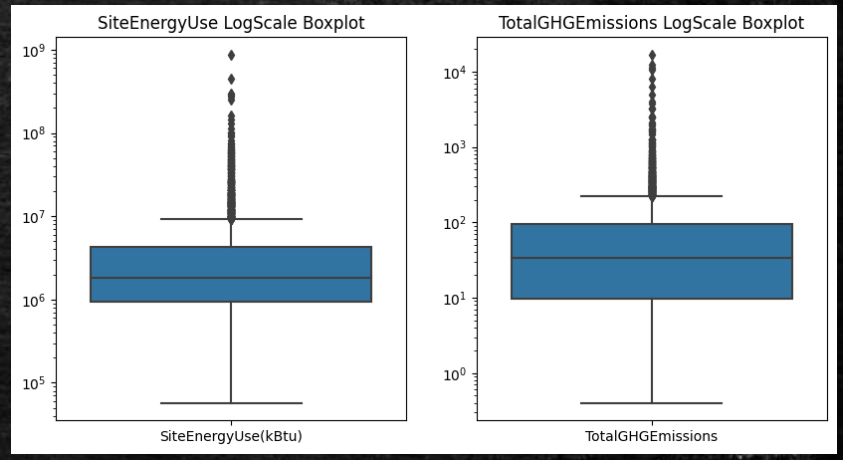
Nettoyage et Exploration – Analyse univariée

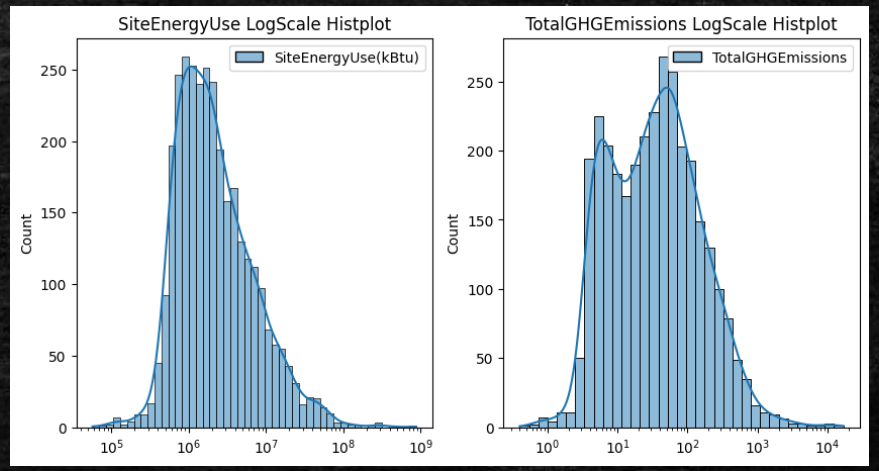
Dispersions des cibles



	TotalGHGEmissions	SiteEnergyUse(kBtu)
count	3207.000000	3.207000e+03
mean	122.211886	5.539669e+06
var	3.044616e+05	4.898833e+14
std	551.534876	2.212364e+07
min	-0.800000	5.713320e+04
25%	9.640000	9.387549e+05
50%	33.920000	1.809587e+06
75%	94.385000	4.277747e+06
max	16870.980000	8.739237e+08

Passage au log



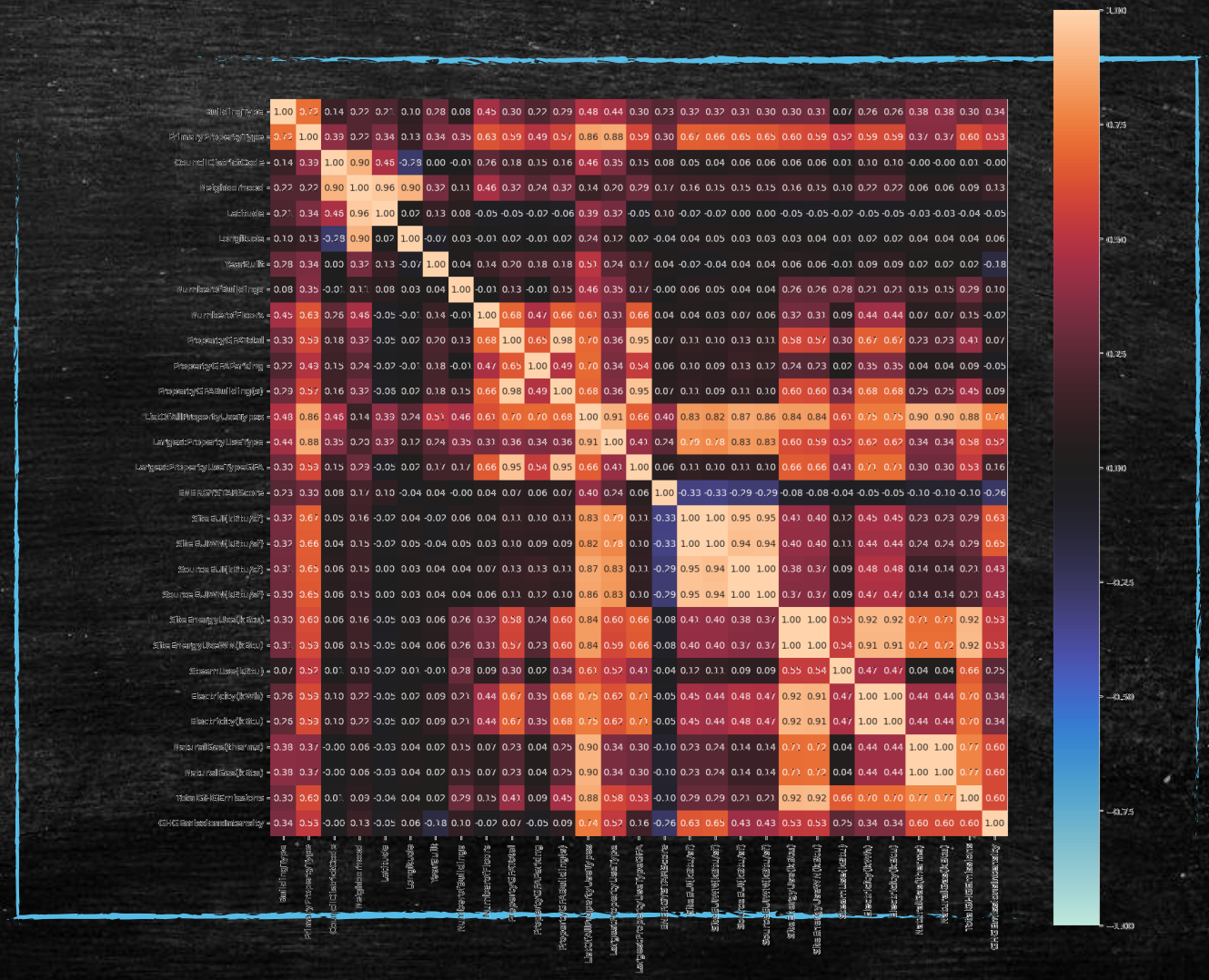


Nettoyage et Exploration – Matrice de corrélation

Quelques « clusters de corrélation »

Suppression des features fortement corrélées & non structurelles :

Suppression ElecBool (tout les bâtiments utilisent de l'électricité)



Nettoyage et Exploration – Matrice de corrélation

Quelques « clusters de corrélation »

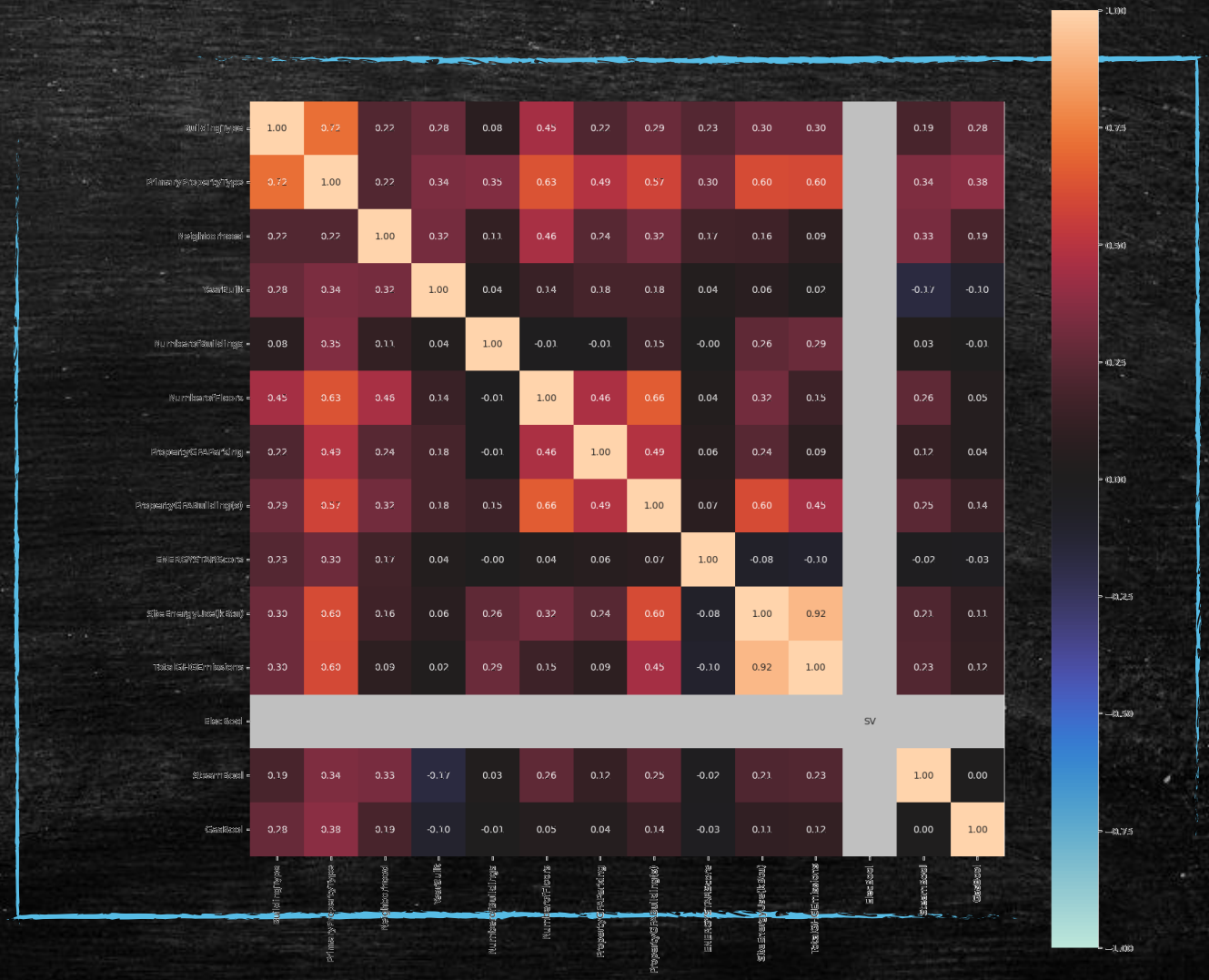
Suppression des features fortement corrélées & non structurelles :

Suppression ElecBool (tout les bâtiments utilisent de l'électricité)

Taille finale :

2758 lignes

11 features + 2 targets



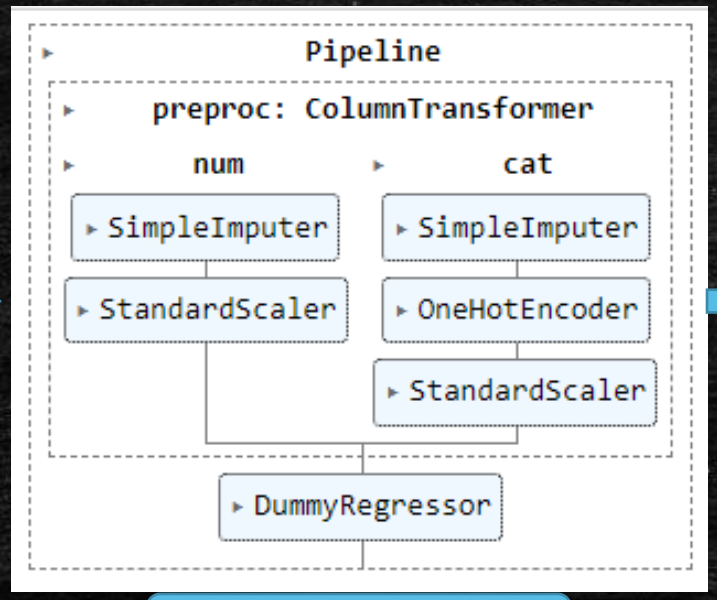
Modeling

Consommation d'énergie & Emissions de gaz à effet de serres

Modeling – Préprocessing

- train_test_split : 80% train, random_state
- pipe = model_pipe(preprocessor, regressor)

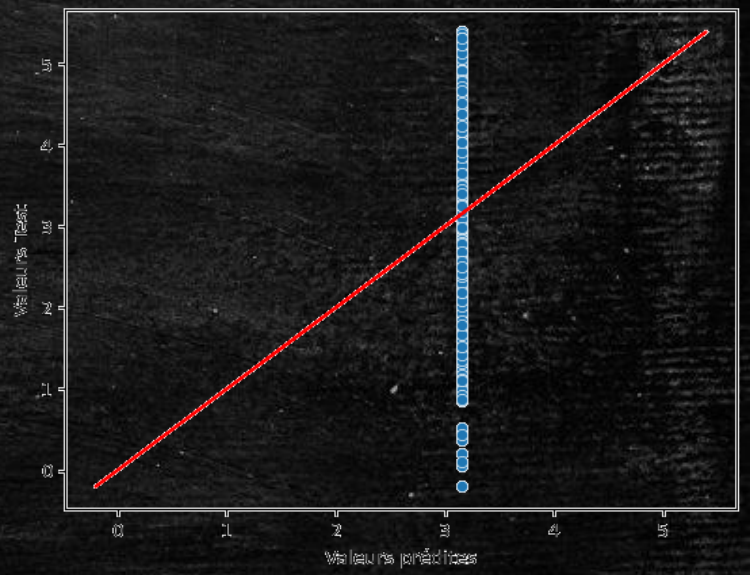
X_train, y_train



fit

Score
(X_test, y_test)

DummyRegressor :
Score: -0,005



Modeling – Modèles & GridSearch

Modèles

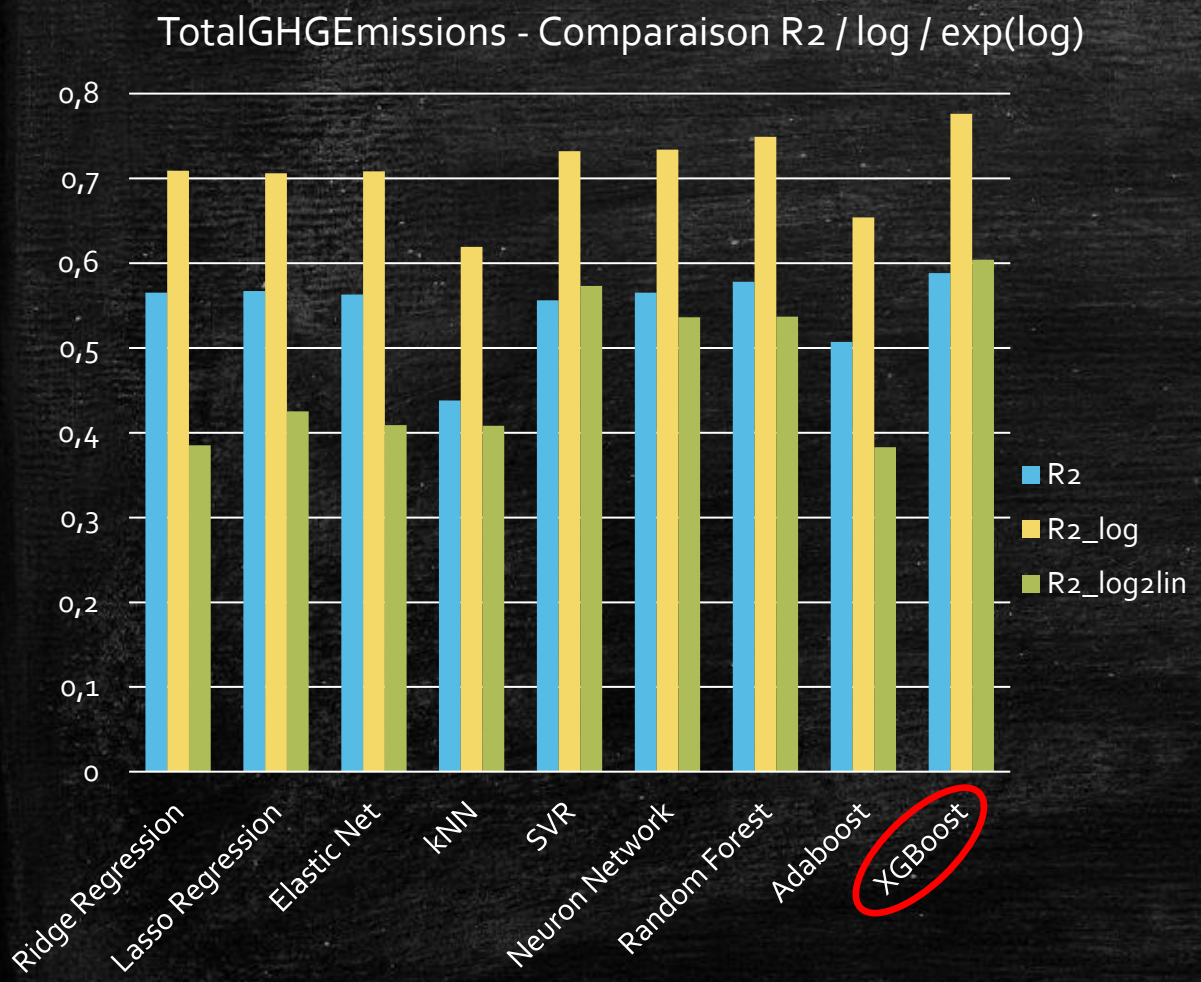
- Ridge Regression
- Lasso Regression
- Elastic Net
- kNN
- SVR
- Neuron Network
- Random Forest
- Adaboost
- XGBoost

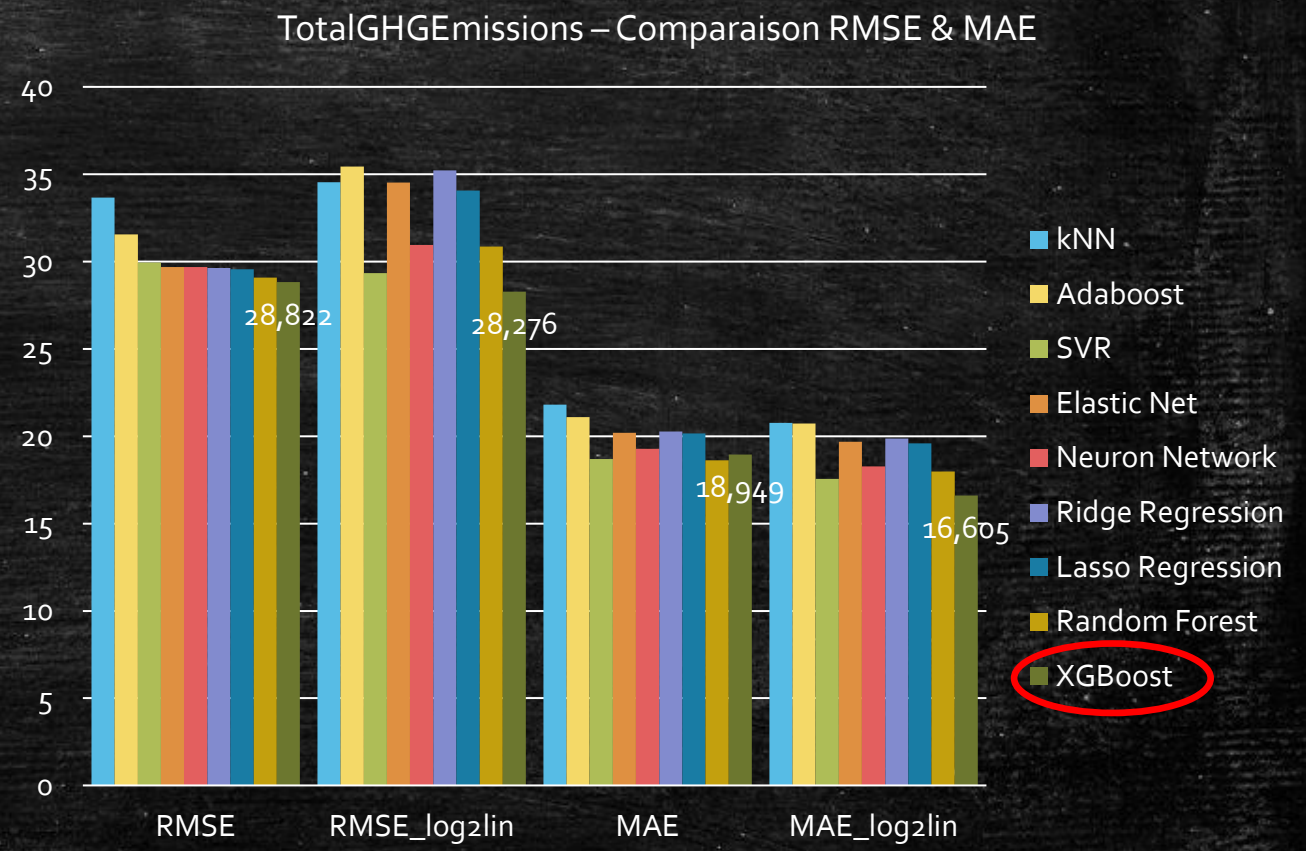
- Echelle d'origine
- Echelle logarithmique

GridSearch

- Kfold : 10 folds, shuffle, random_state
- Tableaux modèles & paramètres
- Boucle for (modèle, paramètres)
  - GridSearchCV (modèle, paramètres)
  - Stockage score r2
  - Stockage GridSearch (best_params, best_estimator, …)

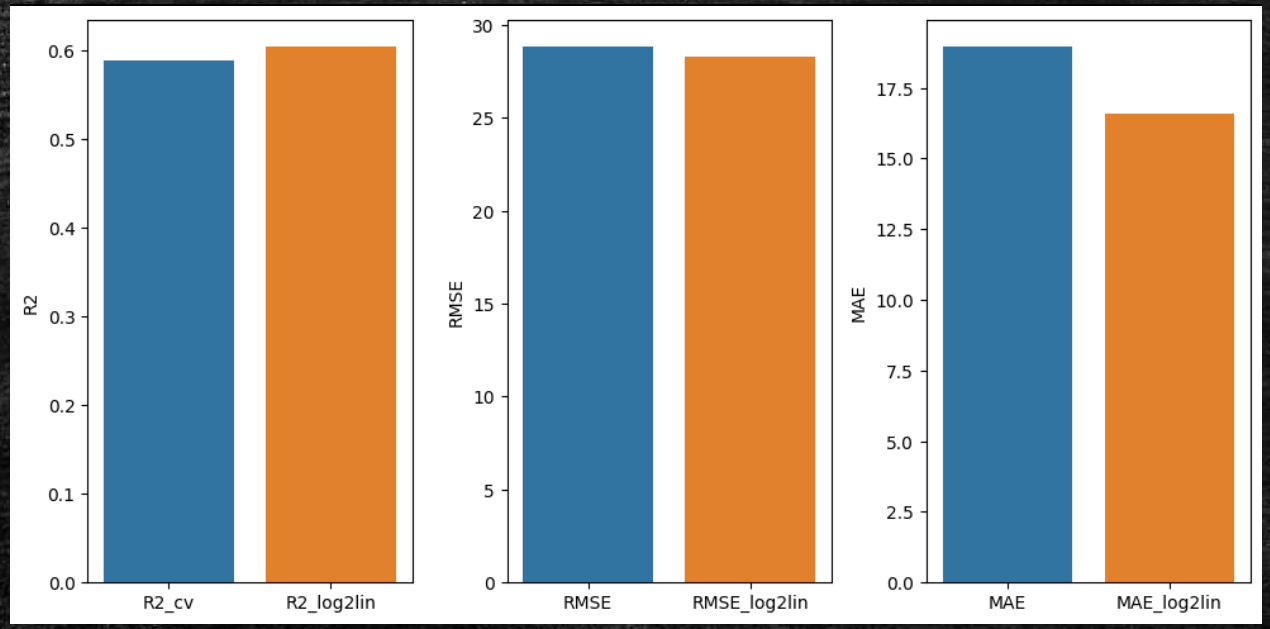
Modeling – Résultats GHG Emissions





Meilleur modèle : XGBoost

Modeling – Résultats GHG Emissions

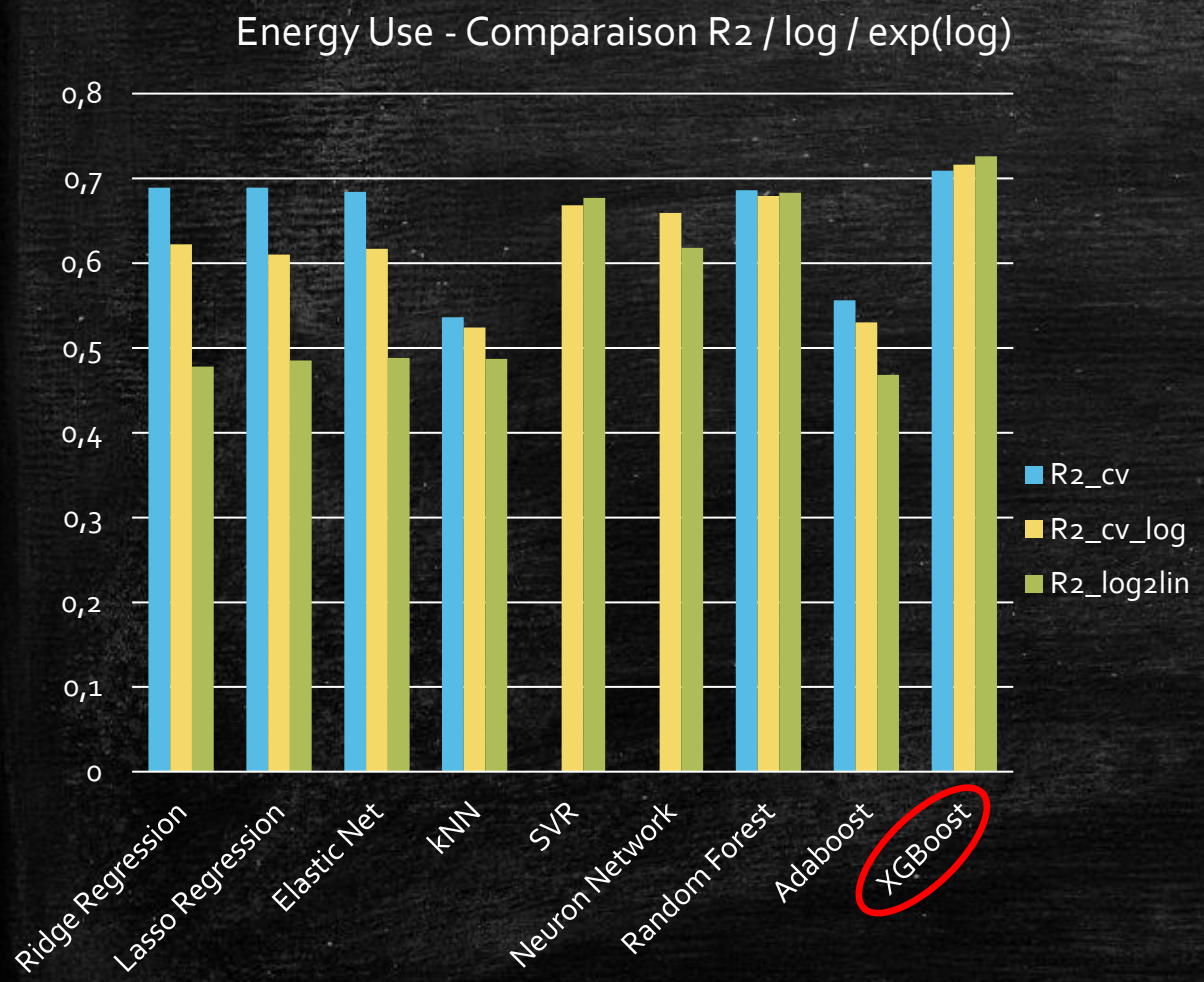


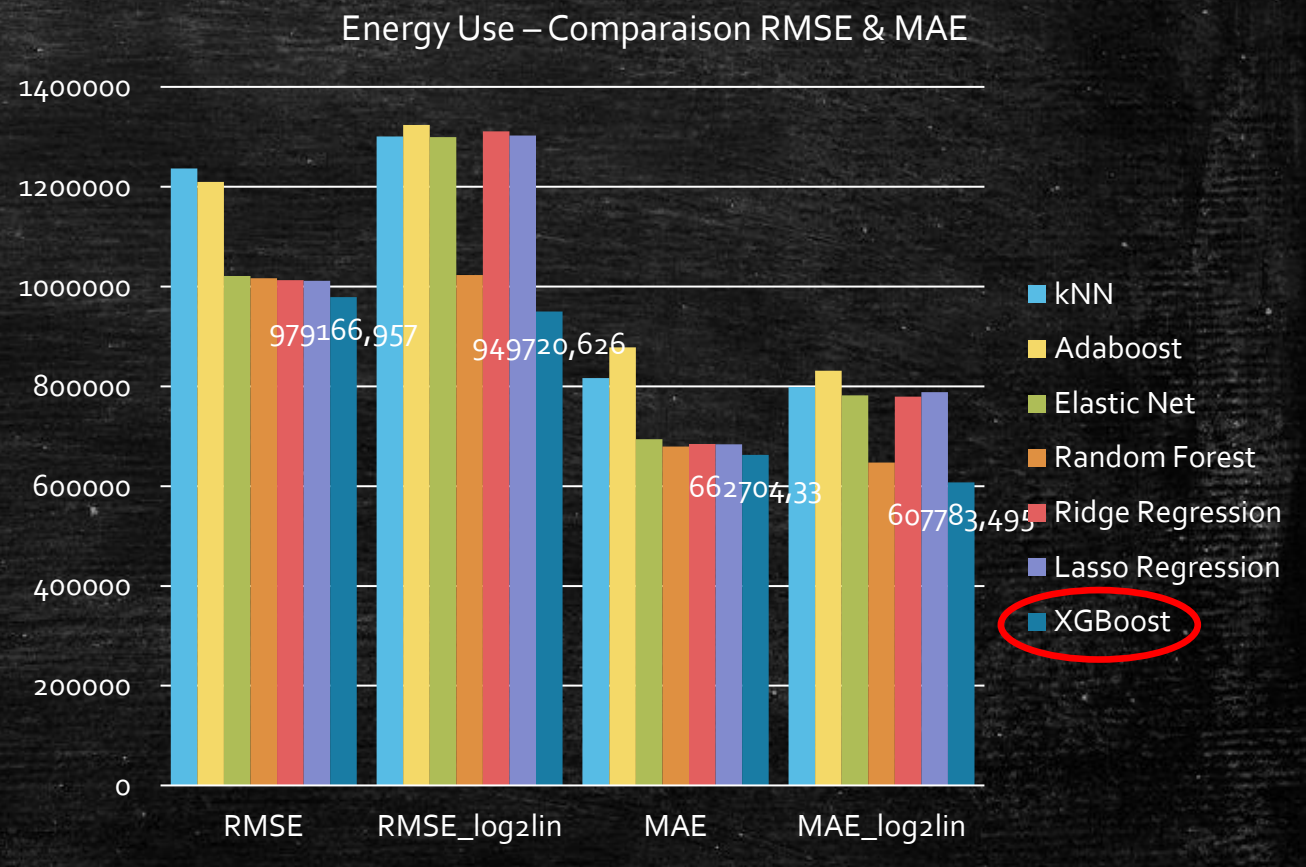
Score R2 : 0.588

Score R2_log2lin : 0.604

Meilleure échelle : Log

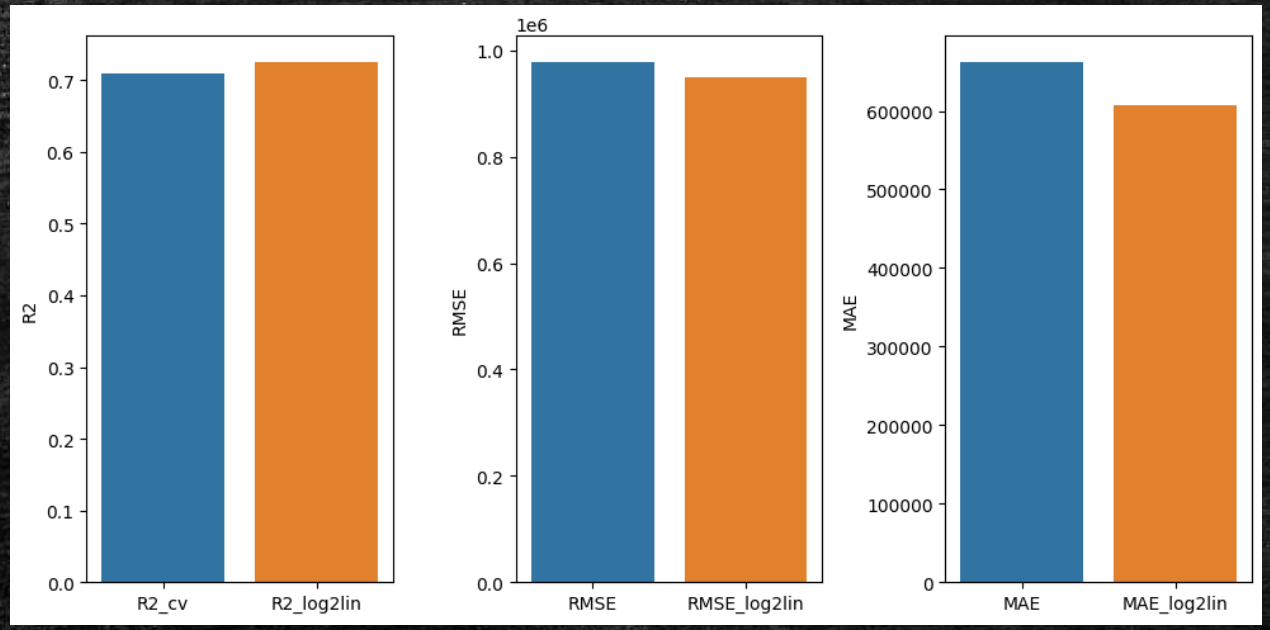
Modeling – Résultats Energy Use





Meilleur modèle : XGBoost

Modeling – Résultats Energy Use

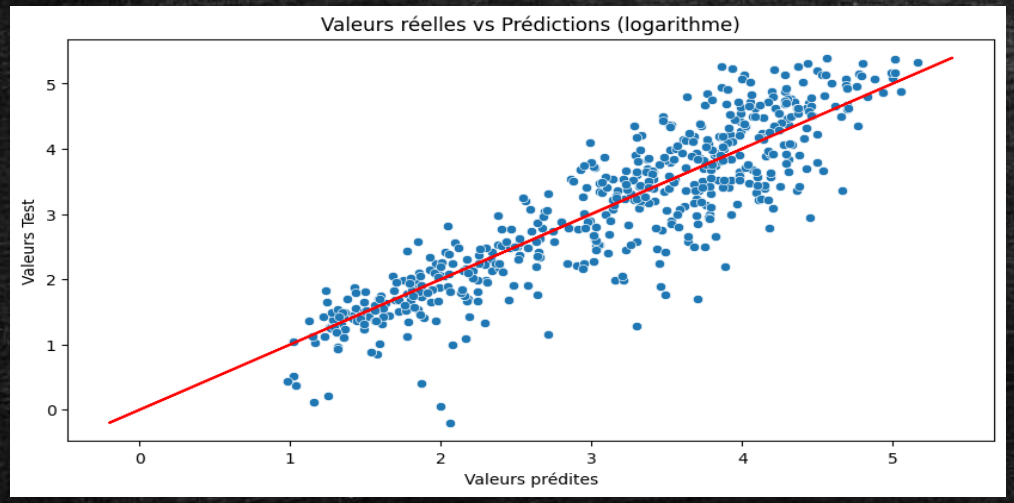


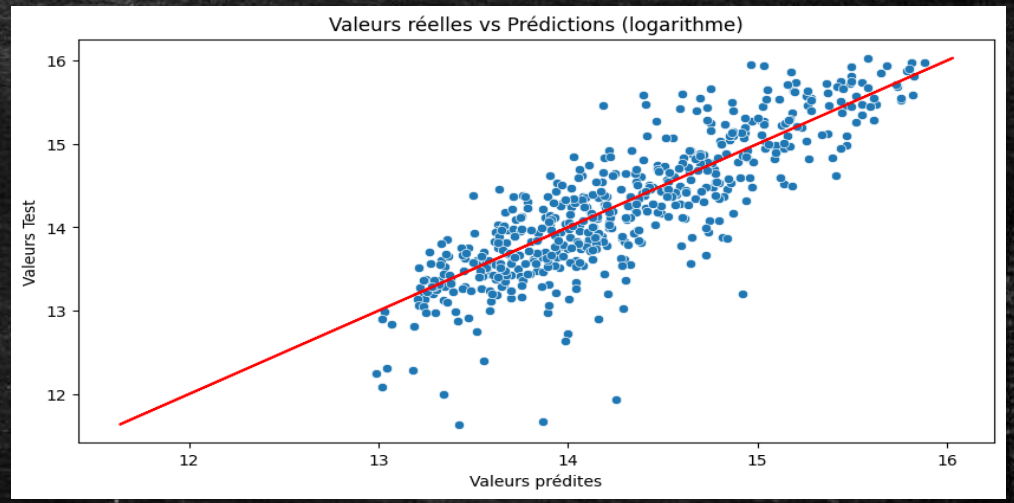
Score R2 : 0.709

Score R2_log2lin : 0.726

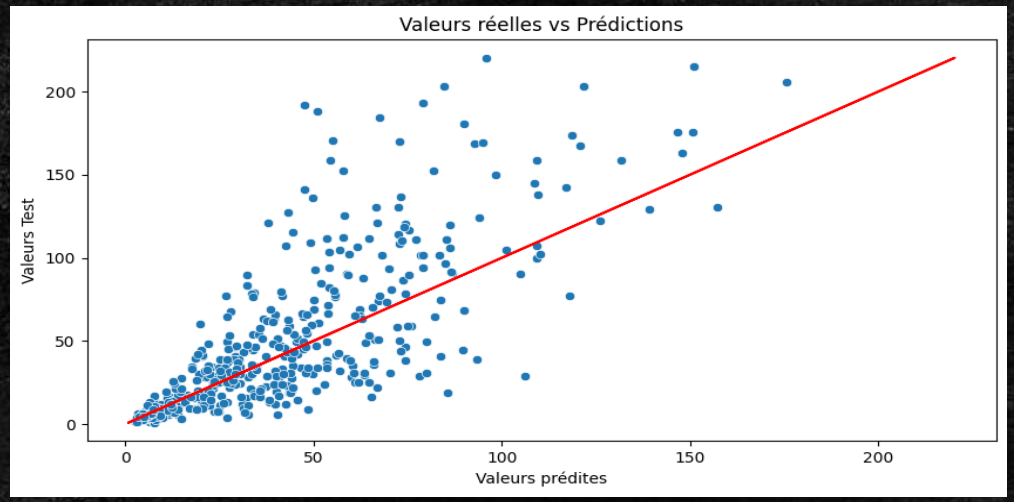
Meilleure échelle : Log

Modeling – Interprétation

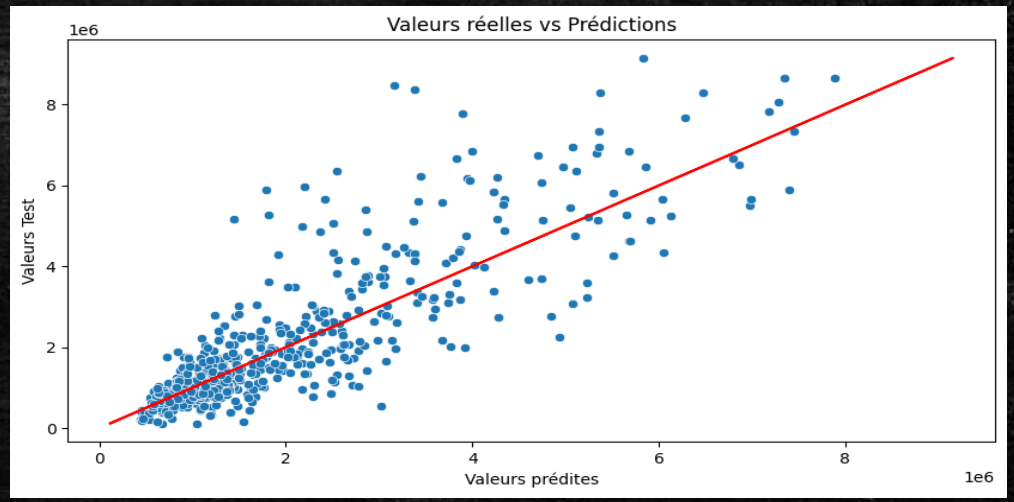




GHG Emissions



Energy Use

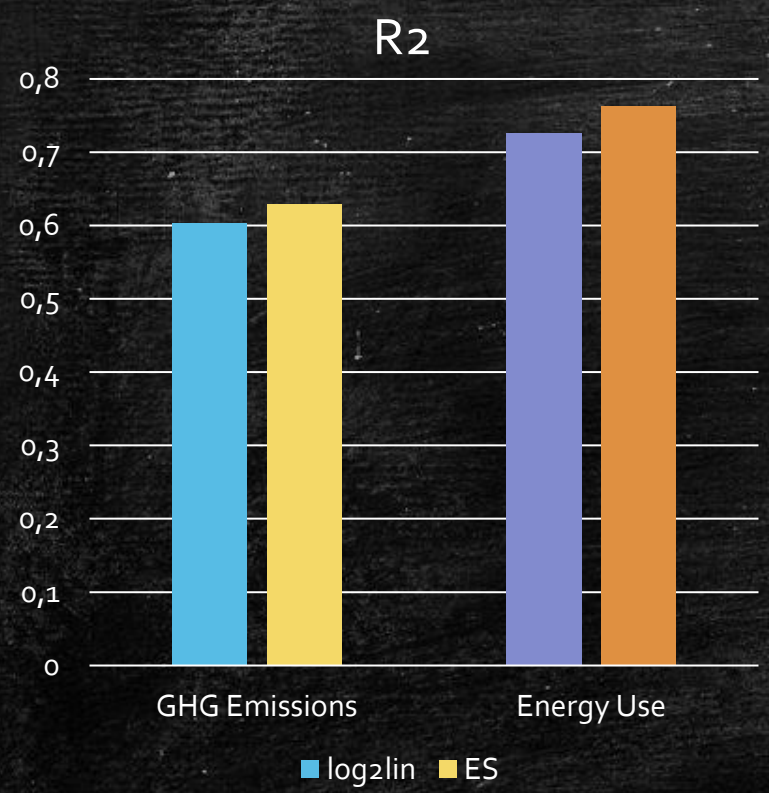


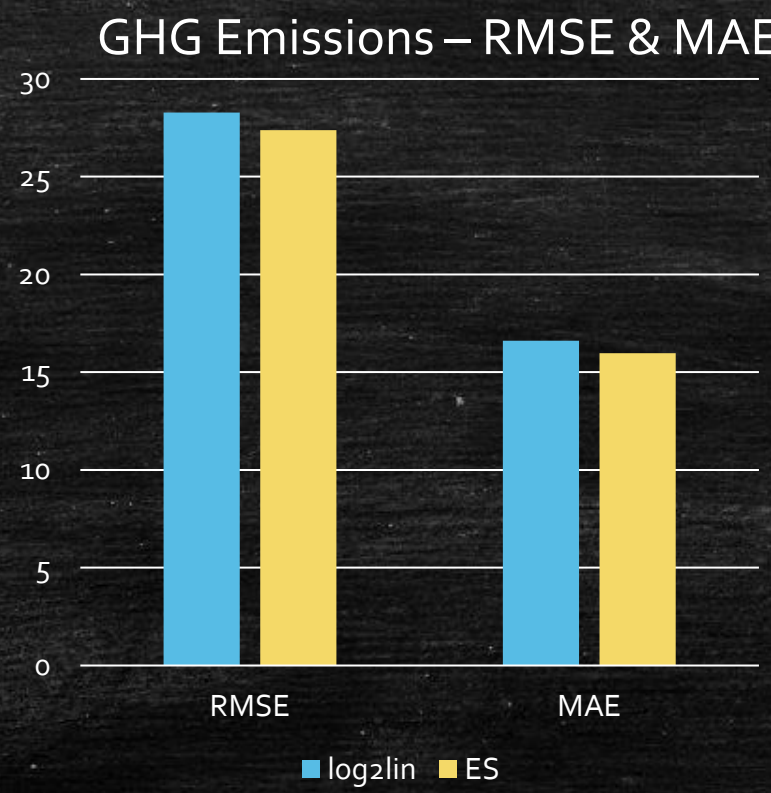
Intérêt de ENERGY STAR Score

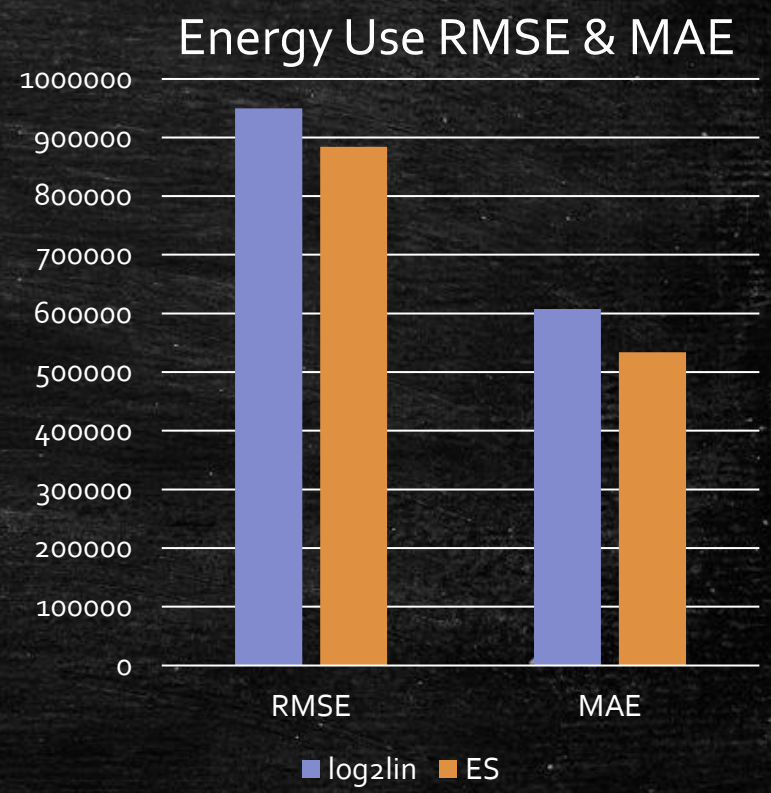
GridSearch avec ES Score

Intérêt de ENERGY STAR Score – Comparaison des scores

XGBoost – Comparaison Sans / Avec ES Score





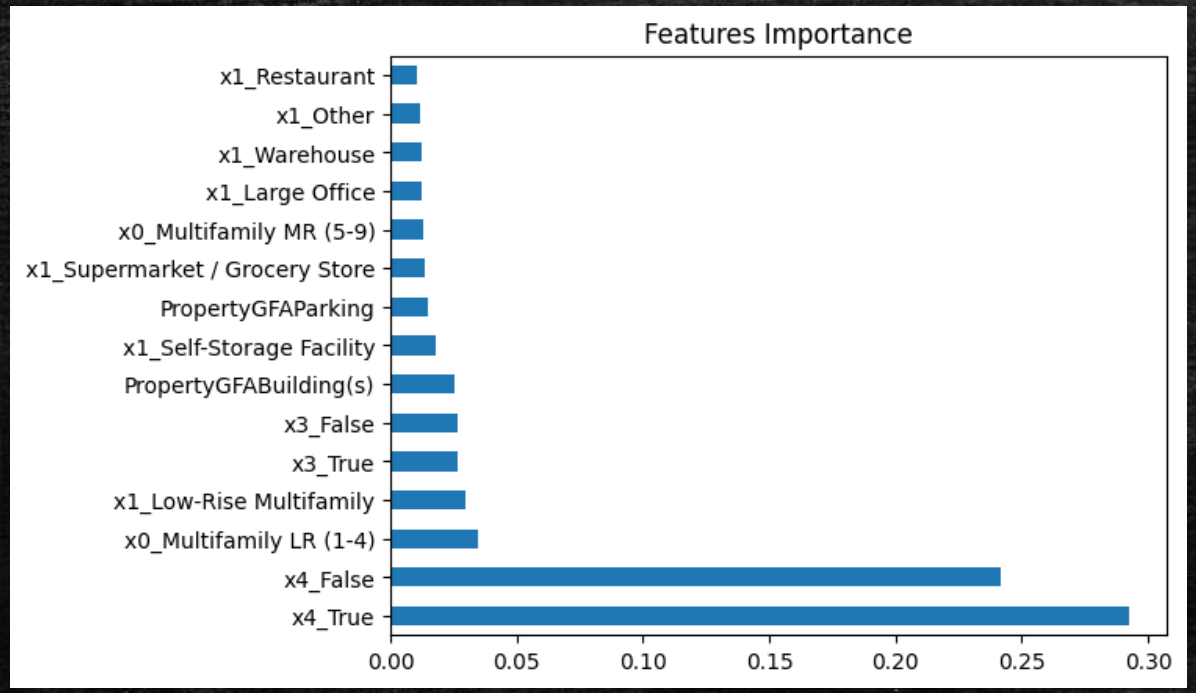


GHG Emissions : R2 : +2.7 % RMSE : -1.61 % MAE : -1.96 %

Energy Use : R2 : +6.1 % RMSE : -3.56 % MAE : -6.47 %

Intérêt de ENERGY STAR Score – Features Importance

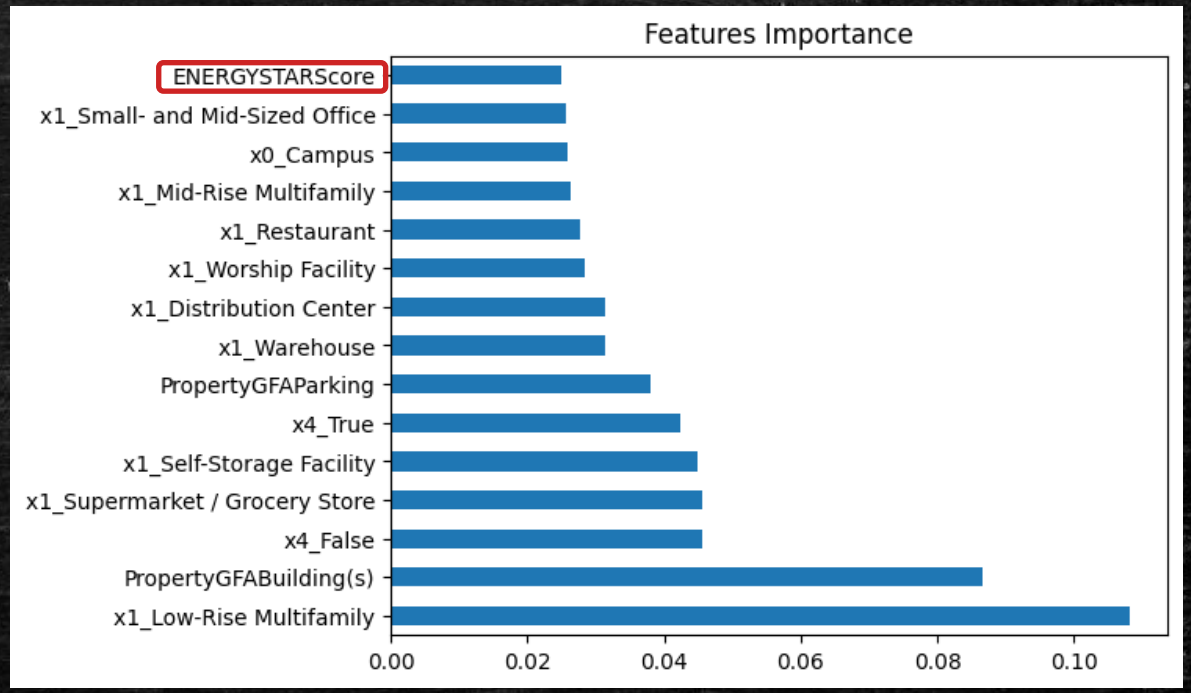
GHG Emissions



ENERGYSTARScore: 0.008

25ème / 60

Energy Use



ENERGYSTARScore: 0.025

15ème / 60

Conclusions

Conclusions

- Meilleur modèle : XGBoost
- Meilleur échelle : Log
- Axes d'amélioration :
  - Augmenter la taille d'échantillon (années, villes)
  - Ajouter des métriques (Type d'isolation, matériaux, …)
- ENERGY STAR Score
  - Importance relative
  - Gain vs Complexité ?
- Prédictions =/= Mesures