

P4

Alexandre Jacqueline

Open Class Room

2022

Sommaire :



1 Ajustement et Nettoyage des données



2 Analyse exploratoire & Feature Engineering



3 Modélisation et Machine Learning



Seattle

Contexte

La ville de Seattle étudie les émissions de GES et la consommation d'énergie des bâtiments dans un objectif de ville neutre

La ville voudrait prédire la consommation d'énergie et les émissions de GES des bâtiments manquants
La ville veut connaître l'importance de l'ENERGY STAR Score

Ajustement et Nettoyage des données



Les données:

2 Daraframes

data_2015.shape

(3340, 47)

data_2016.shape

(3376, 46)

Types de données

OSEBuildingID	int64
DataYear	int64
BuildingType	object
PrimaryPropertyType	object
PropertyName	object
TaxParcelIdentificationNumber	object
Location	object
CouncilDistrictCode	int64
Neighborhood	object
YearBuilt	int64
NumberOfBuildings	int64
NumberOfFloors	int64
PropertyGFATotal	float64
PropertyGFAParking	int64
PropertyGFABuilding(s)	int64
ListofAllPropertyUseTypes	int64
LargestPropertyUseType	object
LargestPropertyUseTypeGFA	object
SecondLargestPropertyUseType	float64
SecondLargestPropertyUseTypeGFA	object
ThirdLargestPropertyUseType	float64
ThirdLargestPropertyUseTypeGFA	object
YearsENERGYSTARCertified	float64
ENERGYSTARScore	float64
SiteEUI(kBtu/sf)	float64
SiteEUIWN(kBtu/sf)	float64
SourceEUI(kBtu/sf)	float64
SourceEUIWN(kBtu/sf)	float64
SiteEnergyUse(kBtu)	float64
SiteEnergyUseWN(kBtu)	float64
SteamUse(kBtu)	float64
Electricity(kwh)	float64
Electricity(kBtu)	float64
NaturalGas(therms)	float64
NaturalGas(kBtu)	float64
OtherFuelUse(kBtu)	float64
GHGEmissions(MetricTonsCO2e)	float64
GHGEmissionsIntensity(kgCO2e/ft2)	float64
DefaultData	object

Source :



Adjustement:

-Vérifications de la similitudes des dataset Colonne, Raws

-Ajustement des données de localisation, variable imbriquées dans la dataset2015, utilisation de la librairie AST afin d'extraire les variables du dict

-Renommage des colonnes et suppression des colonnes qui ne sont pas communes aux deux datasets

-Concaténation des deux datasets

```
df = pd.concat([data_2015[data_2016.columns],data_2016], axis = 0).sort_values(["DataYear", "OSEBuildingID"])
df.shape
```

(6716, 46)

Nettoyage :

Actions:	Ligne	Colonne
Suppression des données d'habitations	3318	46
Suppression des variables qui ne nous intéressent pas	3318	42
Suppression des variables avec un suffixe WN	3318	39
Suppression des variables redondantes	3318	37
MAJ de quelque variables uniques	3318	37
Suppression des variables négatives pour PropertyGFAParking et PropertyGFABuilding(s)	3318	37
Correction du nombre de bâtiments et étages nan to 0	3313	37

NaN:

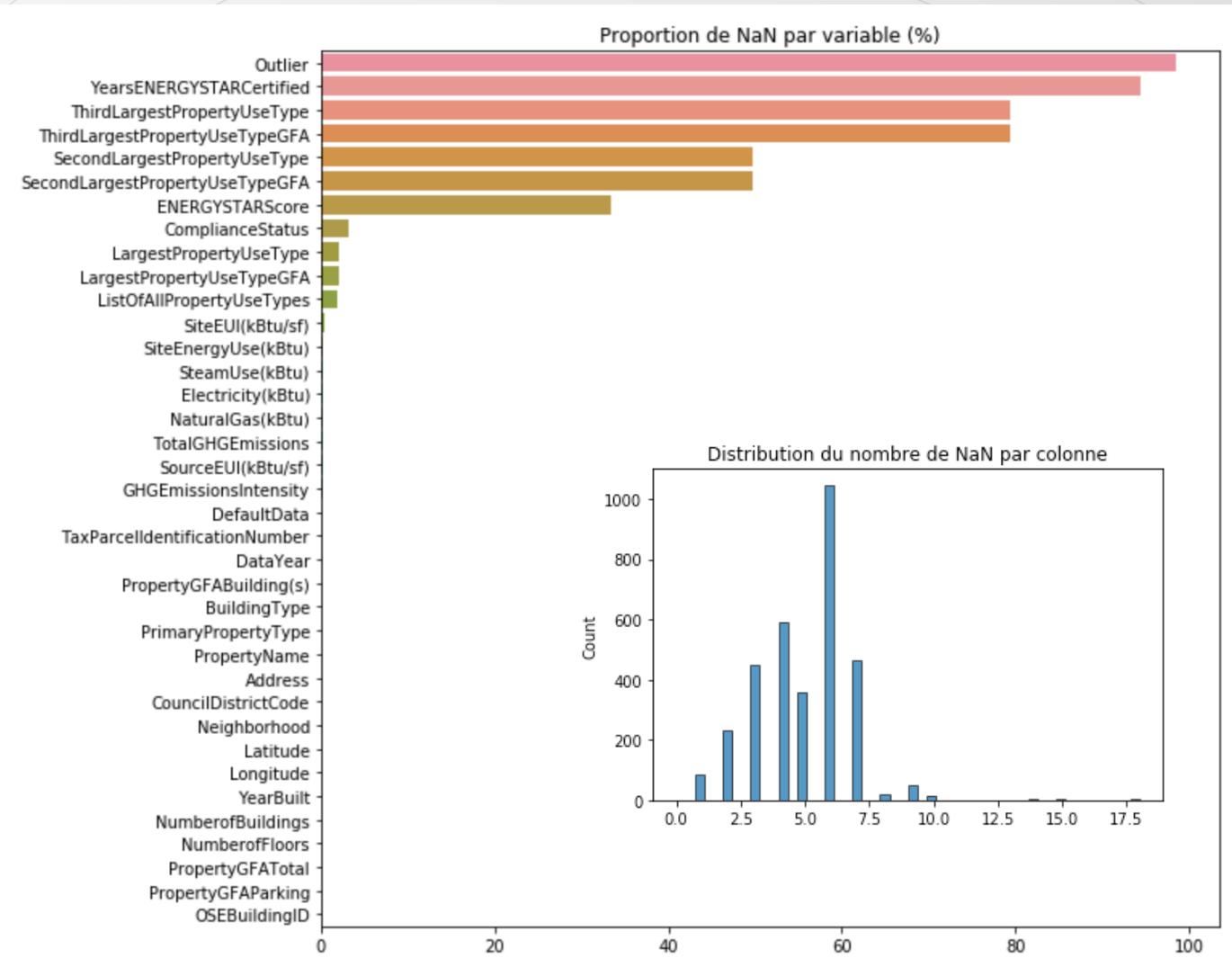
Elimination des lignes ne comportant des NaNs

Complétion des NaN de certaines variables par 0

Suppression des données ne comportant pas d'information sur la consommation énergétique

`df.shape`

(3283, 37)



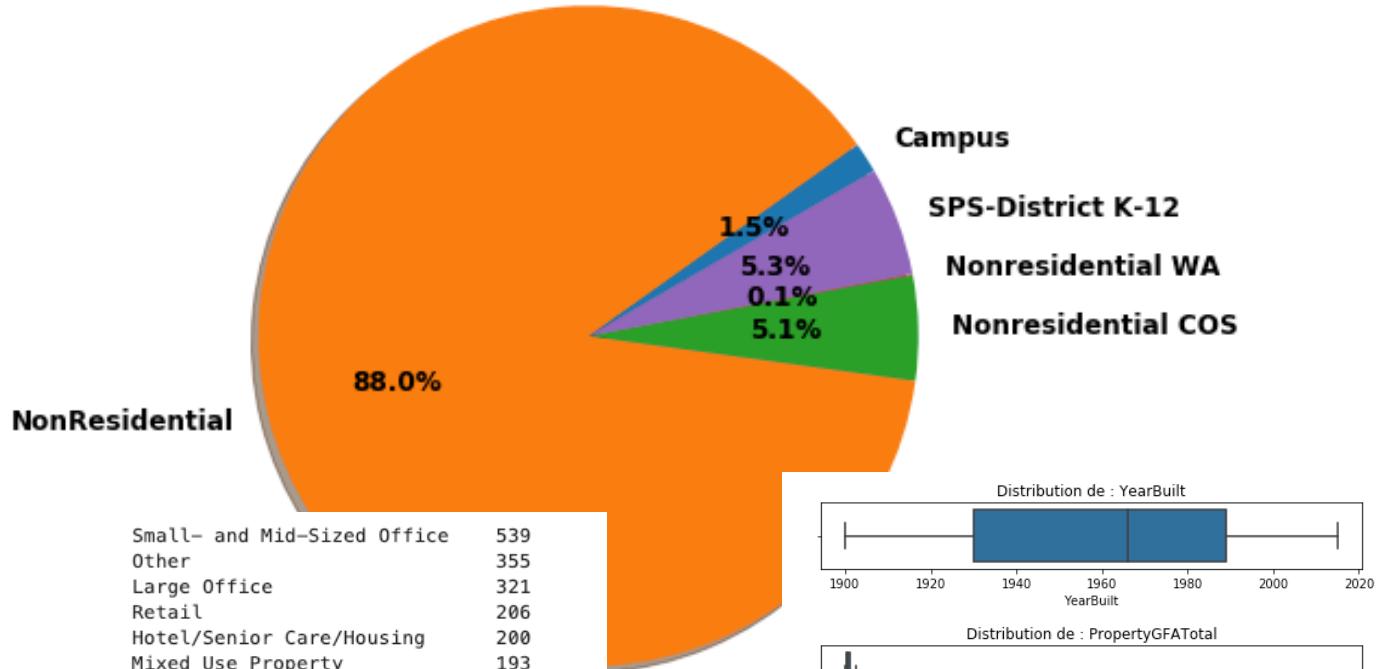


Analyse exploratoire

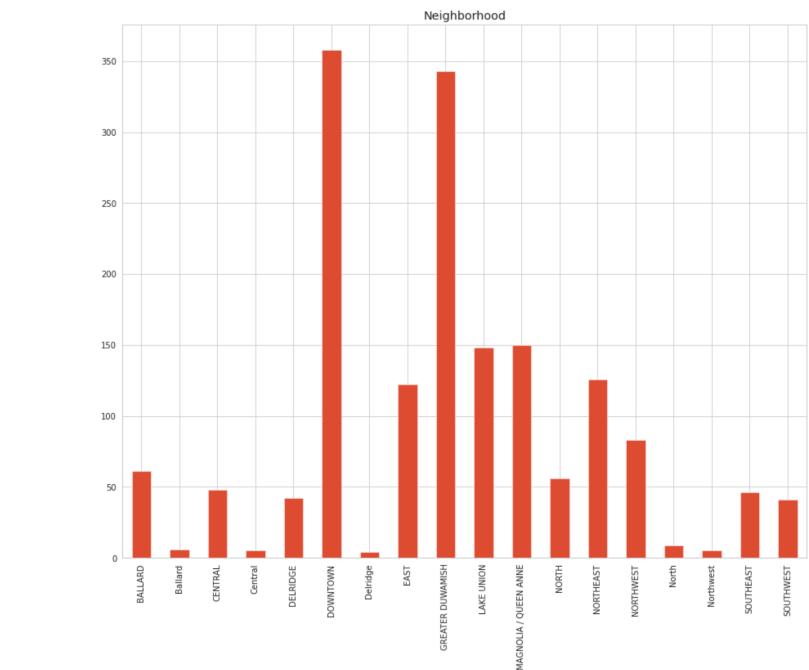
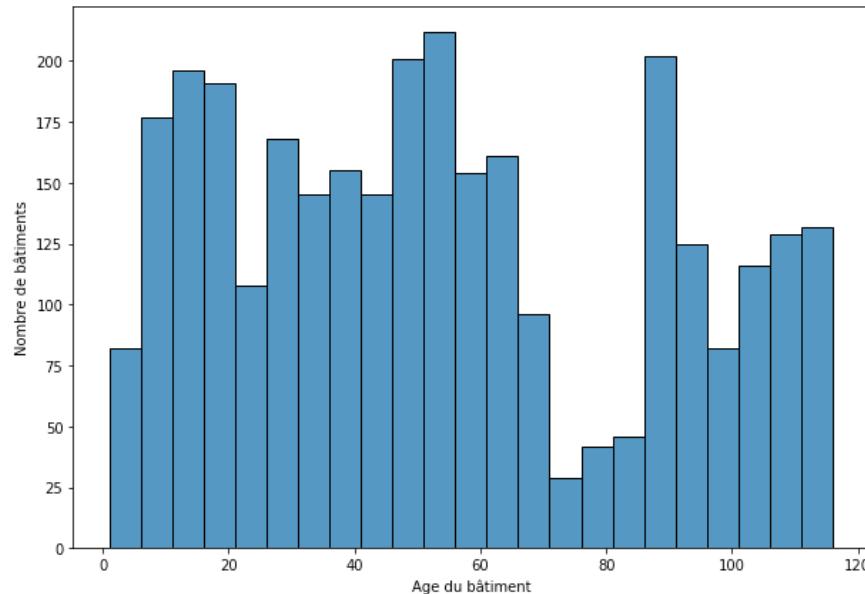
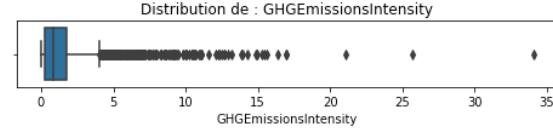
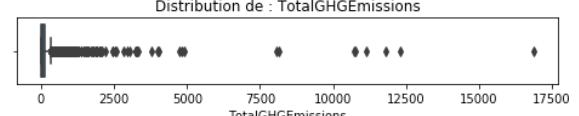
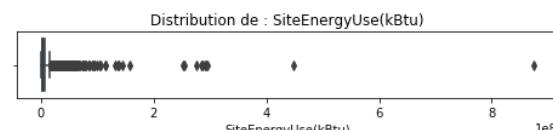
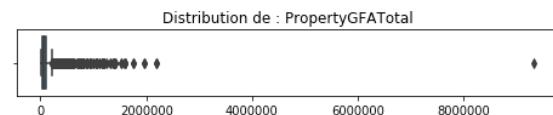
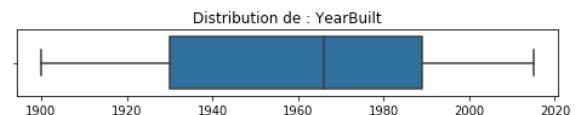


Distribution de l'âge des bâtiments

Répartition des types de bâtiments du dataset

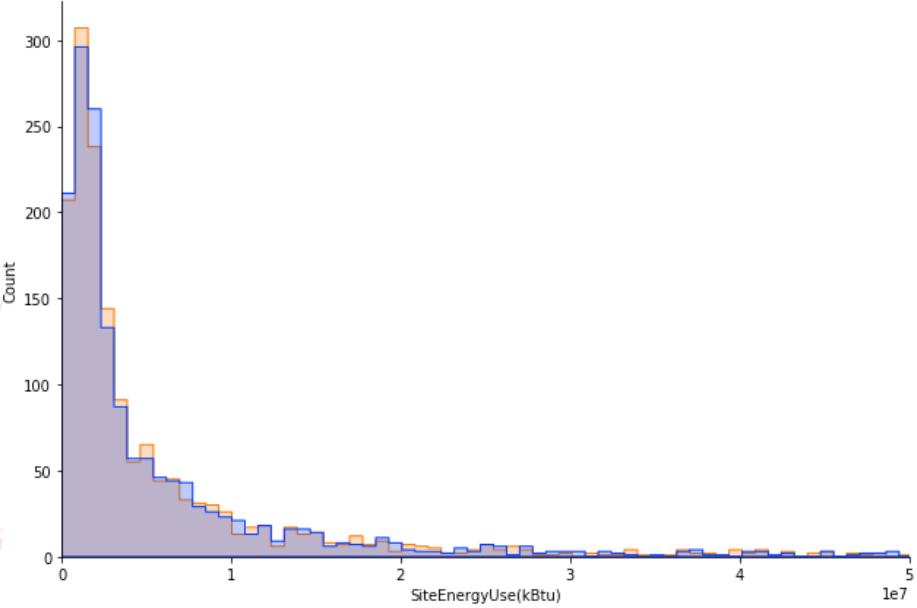


Name: PrimaryPropertyType, dtype: int64

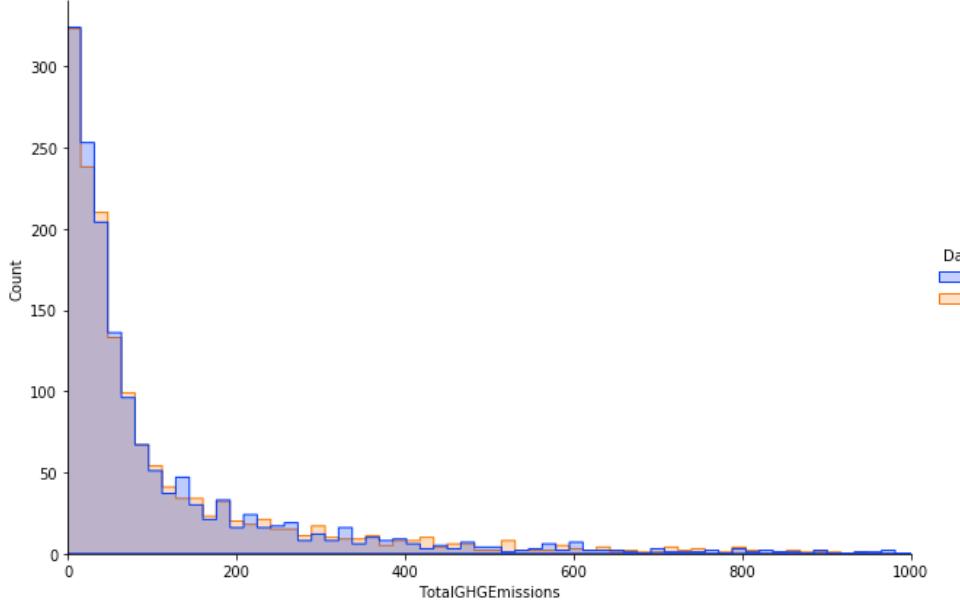


Ecarts de consommation entre les années

Distribution des écarts de consommation normalisée 2015-2016 pour des bâtiments identiques (en %)

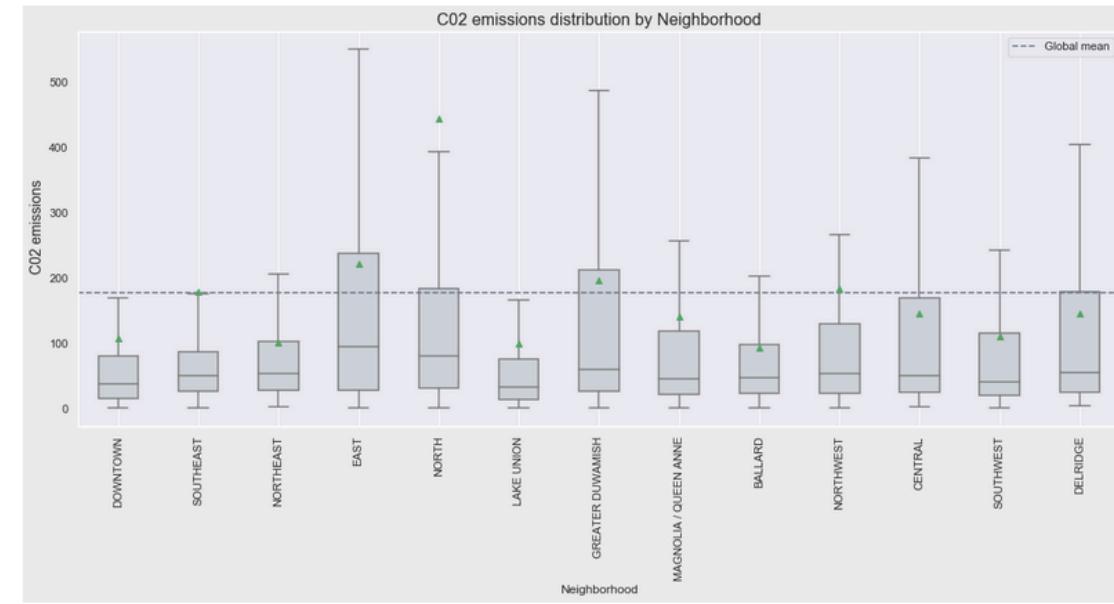
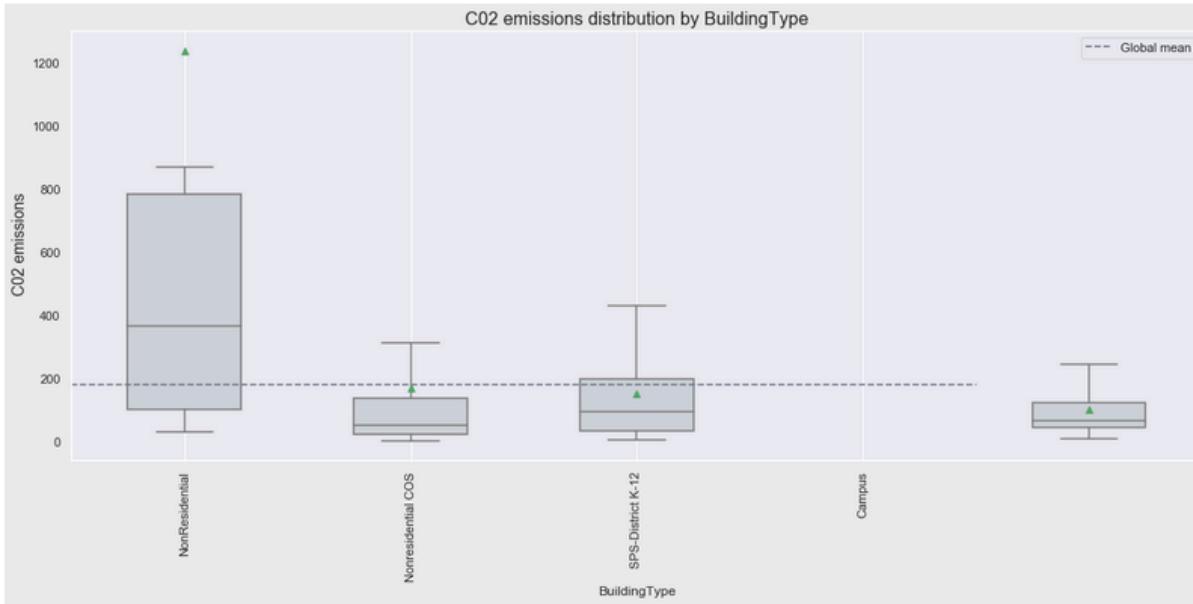


Distribution des écarts d'émissions de CO2 normalisée 2015-2016 pour des bâtiments identiques (en %)

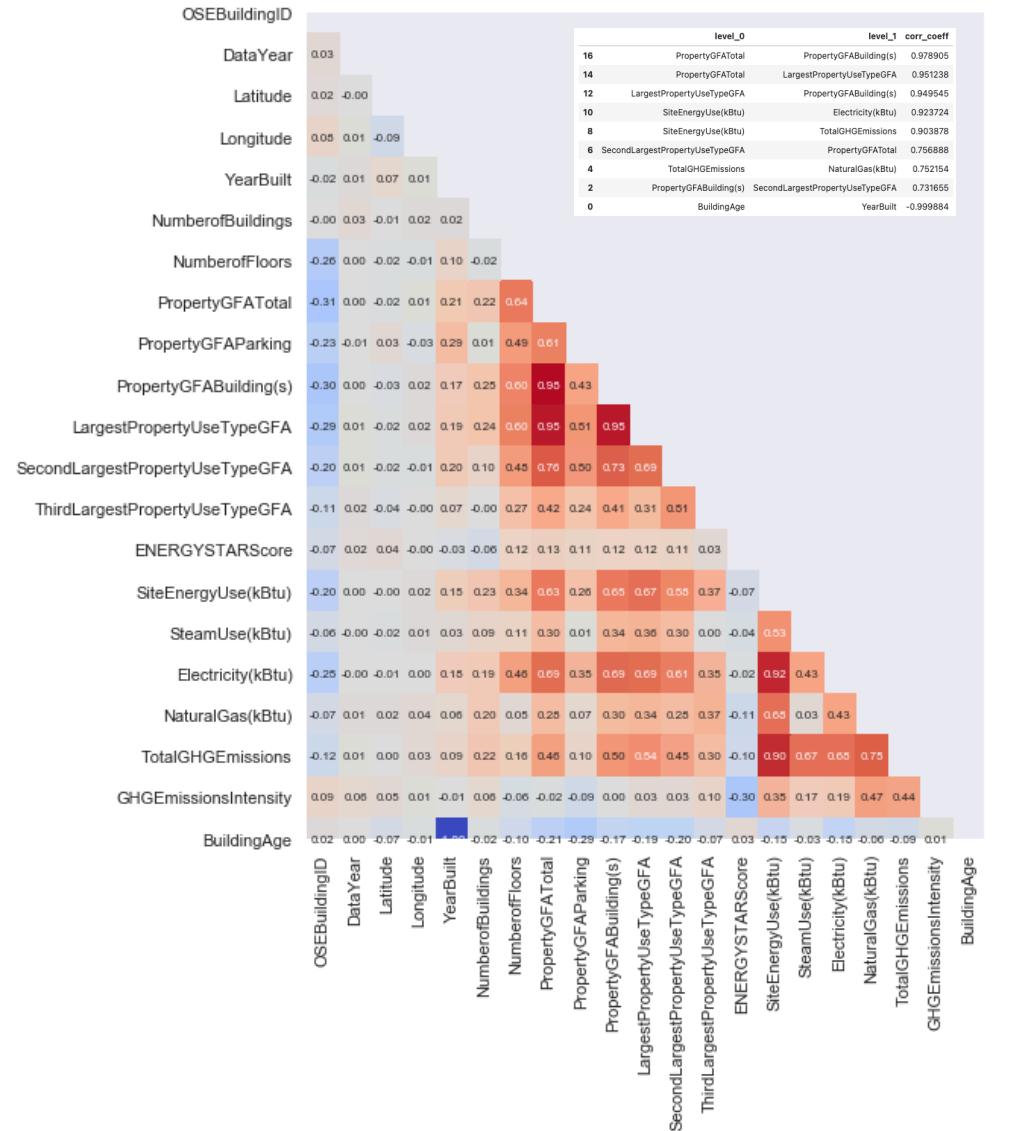


- la moyenne d'écart de consommation normalisée est de 10,4 %
- le minimum d'écart est nul
- le maximum d'écart est de 100 % (!)
- la médiane est à 5,4 %
- la déviation standard est à 15,4 %

CO2 VS Building Type / Neighborhood



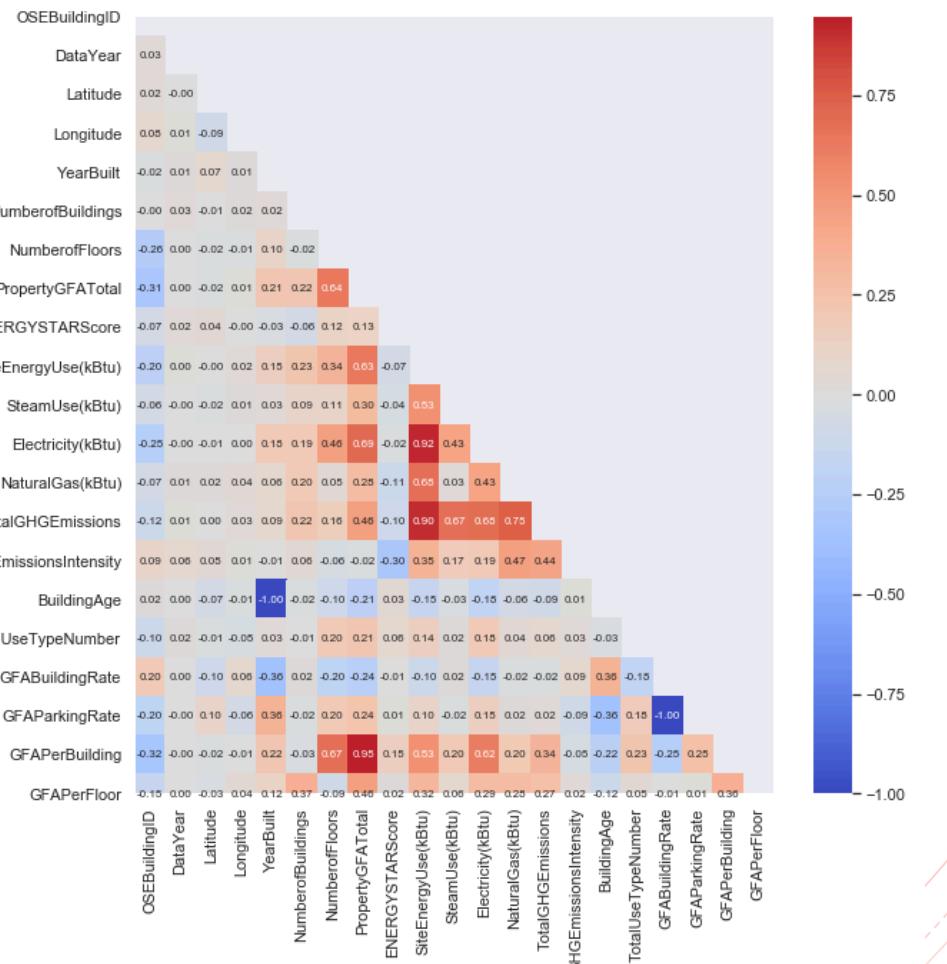
Heatmap des corrélations linéaires



df.shape

(3092, 26)

Heatmap des corrélations linéaires



Preprocessing



Encodage et standardisation



Centrage réduction

Test de differents modèles : Choix des modeles

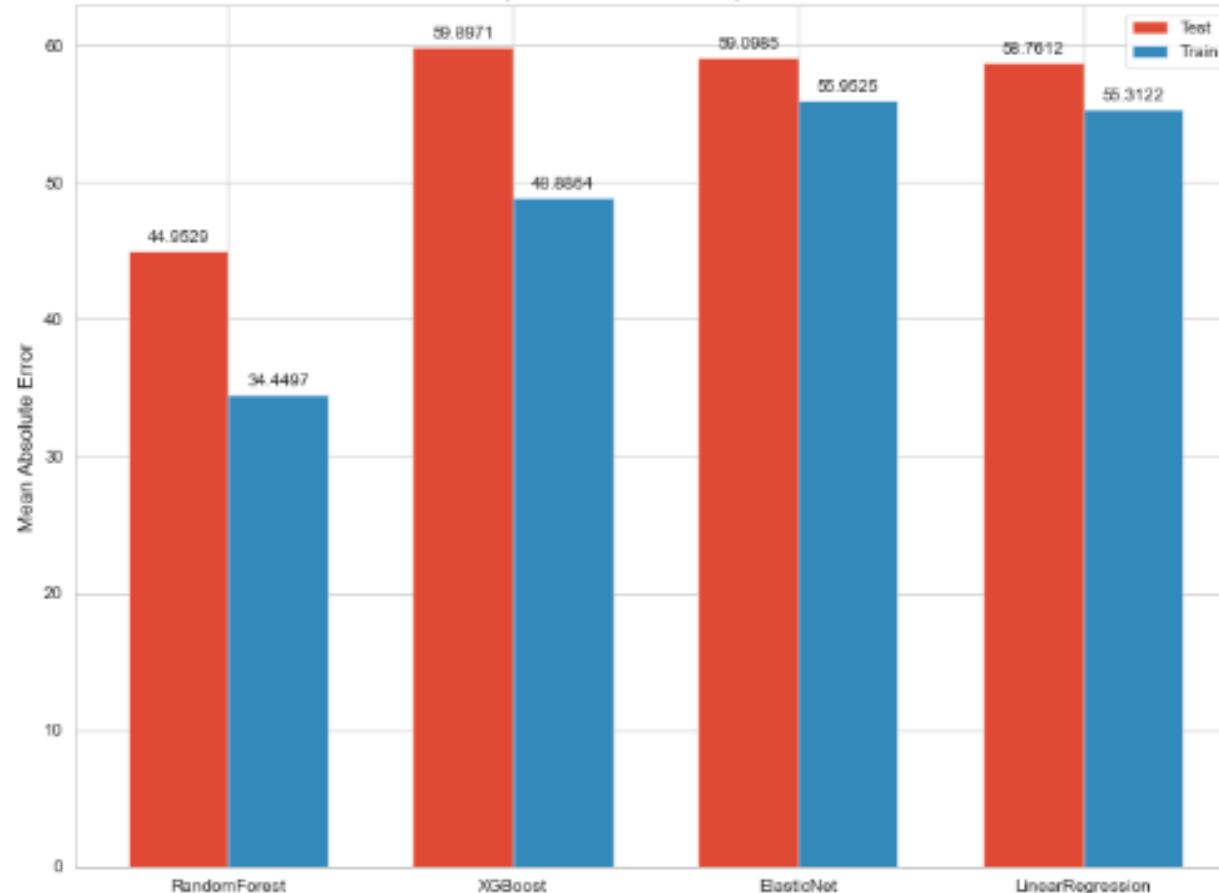
- Modèle Baseline : Régression linéaire multivariée
- Modèle linéaires : ElascticNet et SVR
- Modèle non-linéaires : XGBoost et RandomForestRegressor

	mean_fit_time	mean_score_time	mean_test_neg_mean_absolute_error	mean_train_neg_mean_absolute_error
RandomForest	0.647721	0.032200	-44.952917	-34.449667
XGBoost	0.472775	0.006023	-59.897115	-48.886410
LinearSVR	0.048192	0.001412	-108.113022	-103.791562
ElasticNet	0.005954	0.001135	-59.098471	-55.952496
LinearRegression	0.063207	0.001193	-58.761168	-55.312199

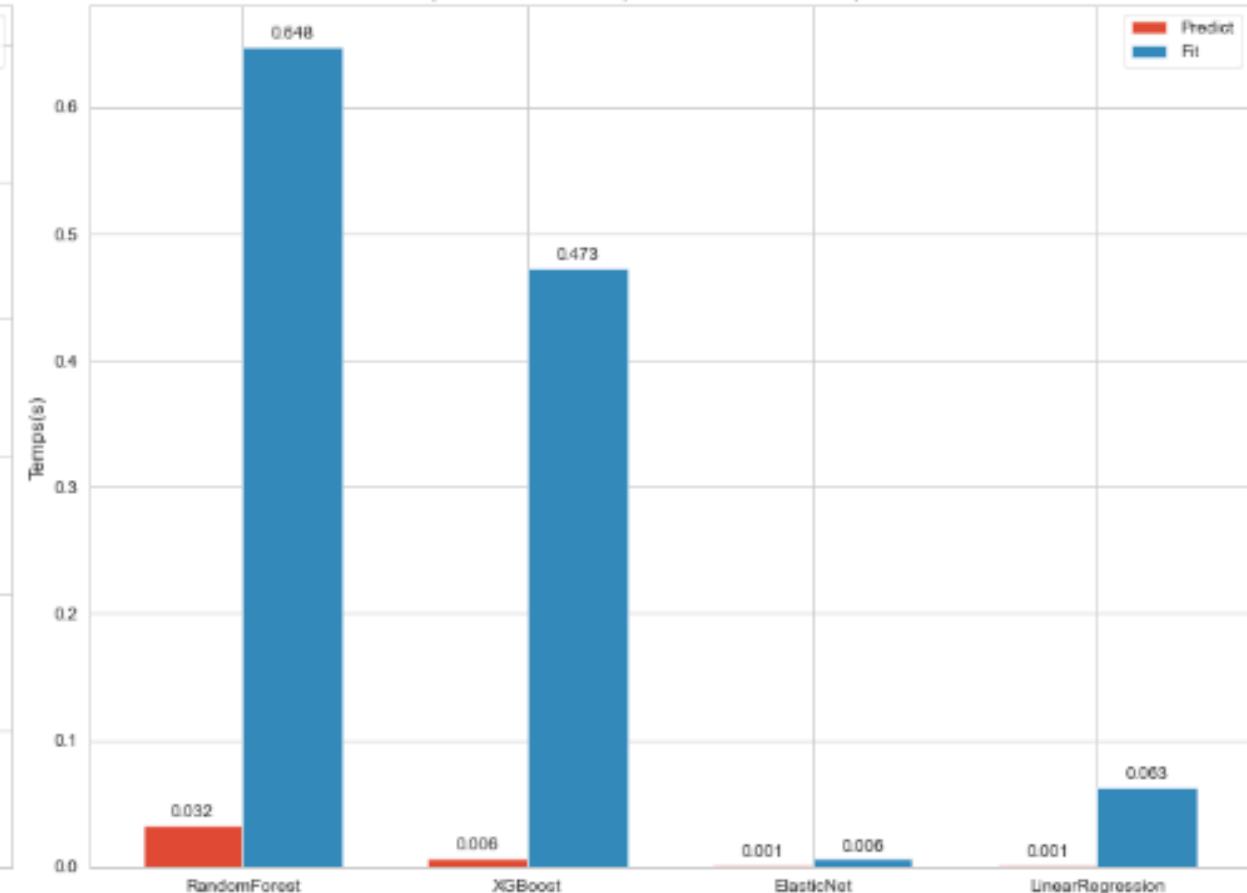
Sélection des meilleurs modèles : Emissions de CO₂

Modélisations sur la variable TotalGHGEmissions

Comparaison des scores par modèle

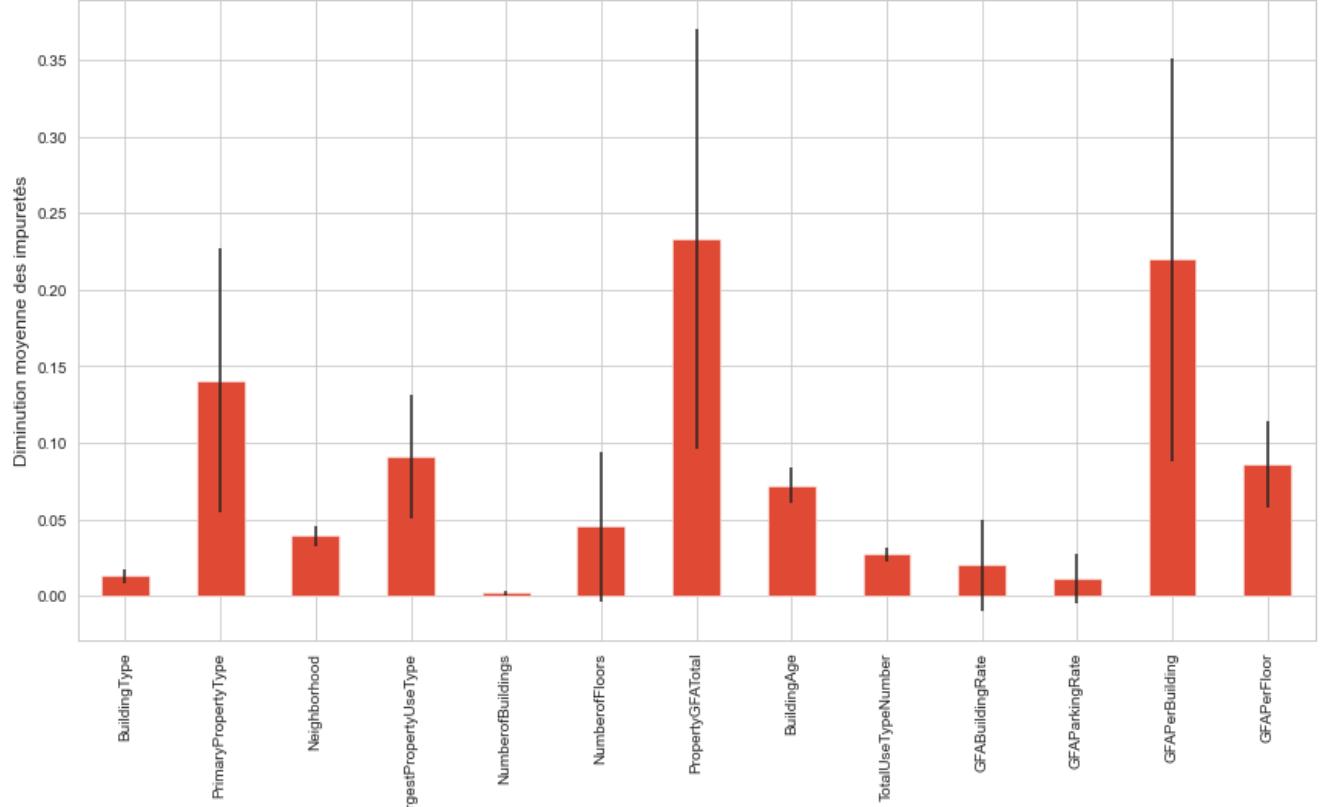


Comparaison des temps d'entraînement et prédiction

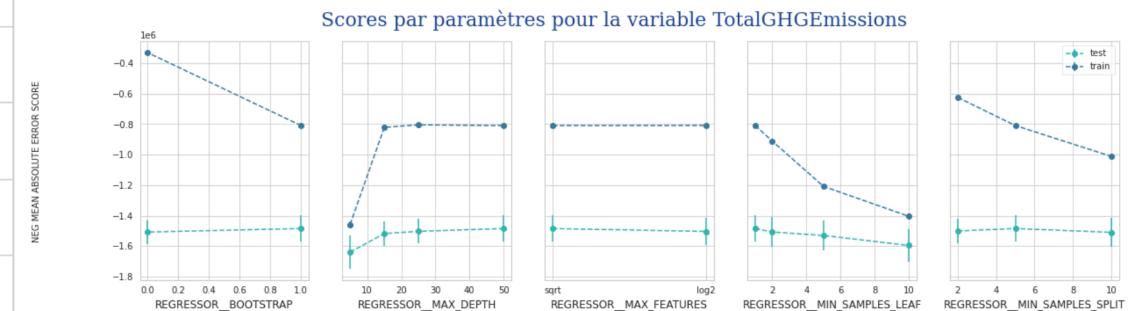


Sélection des meilleurs modèles : Emissions de CO2

Feature importances du modèle RandomForestRegressor sur les émissions de CO2



Scores par paramètres pour la variable TotalGHGEmissions

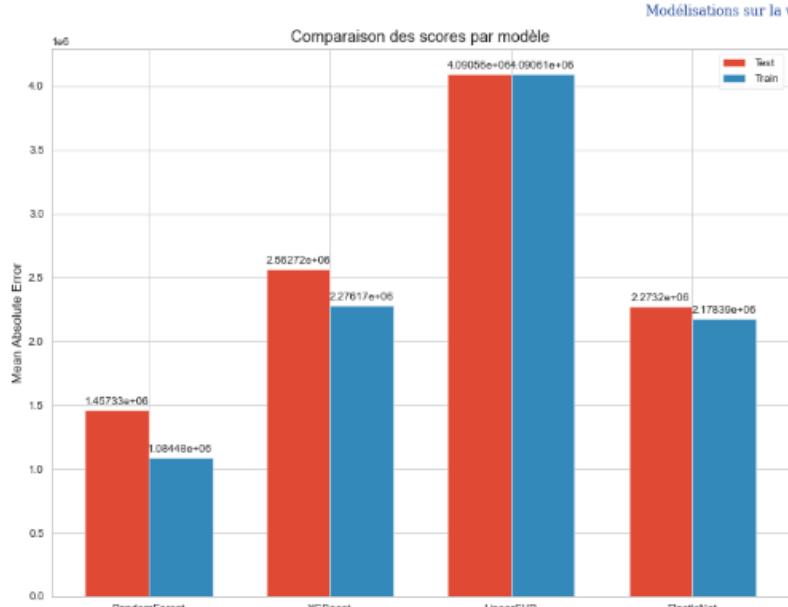


Rappel des meilleurs paramètres :

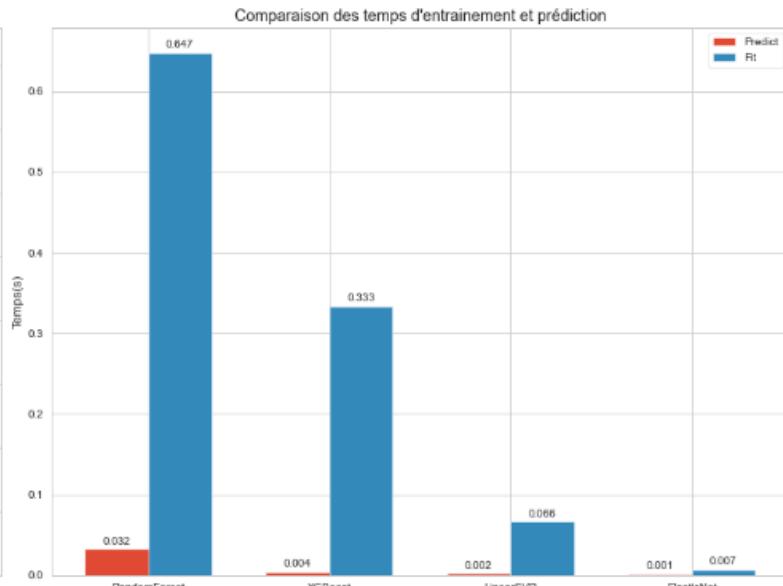
```
{'regressor__bootstrap': True, 'regressor__max_depth': 50, 'regressor__max_features': 'sqrt', 'regressor__min_samples_leaf': 1, 'regressor__min_samples_split': 5}
```

Sélection des meilleurs modèles : Consommations d'énergie

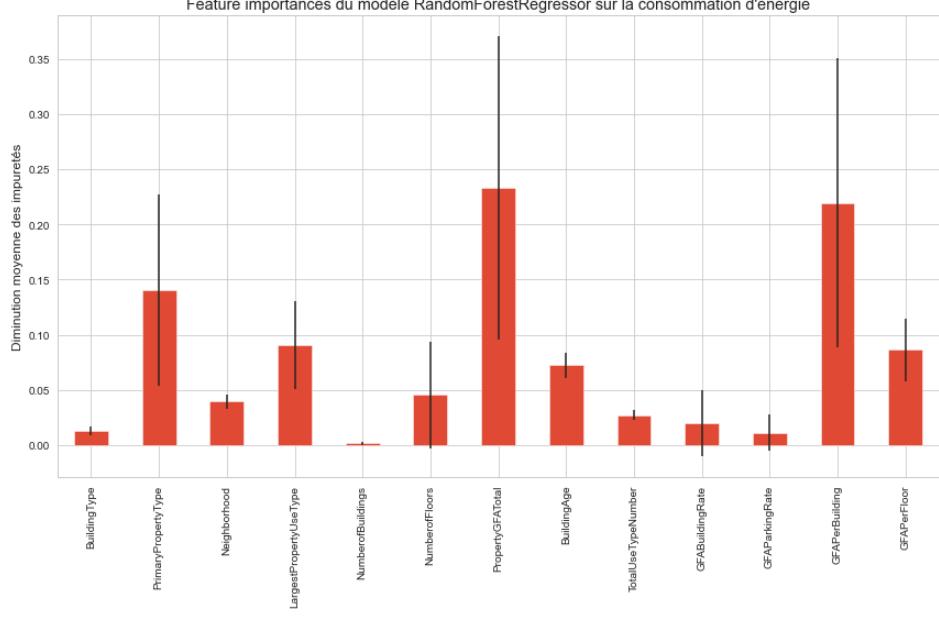
Comparaison des scores par modèle



Comparaison des temps d'entraînement et prédiction

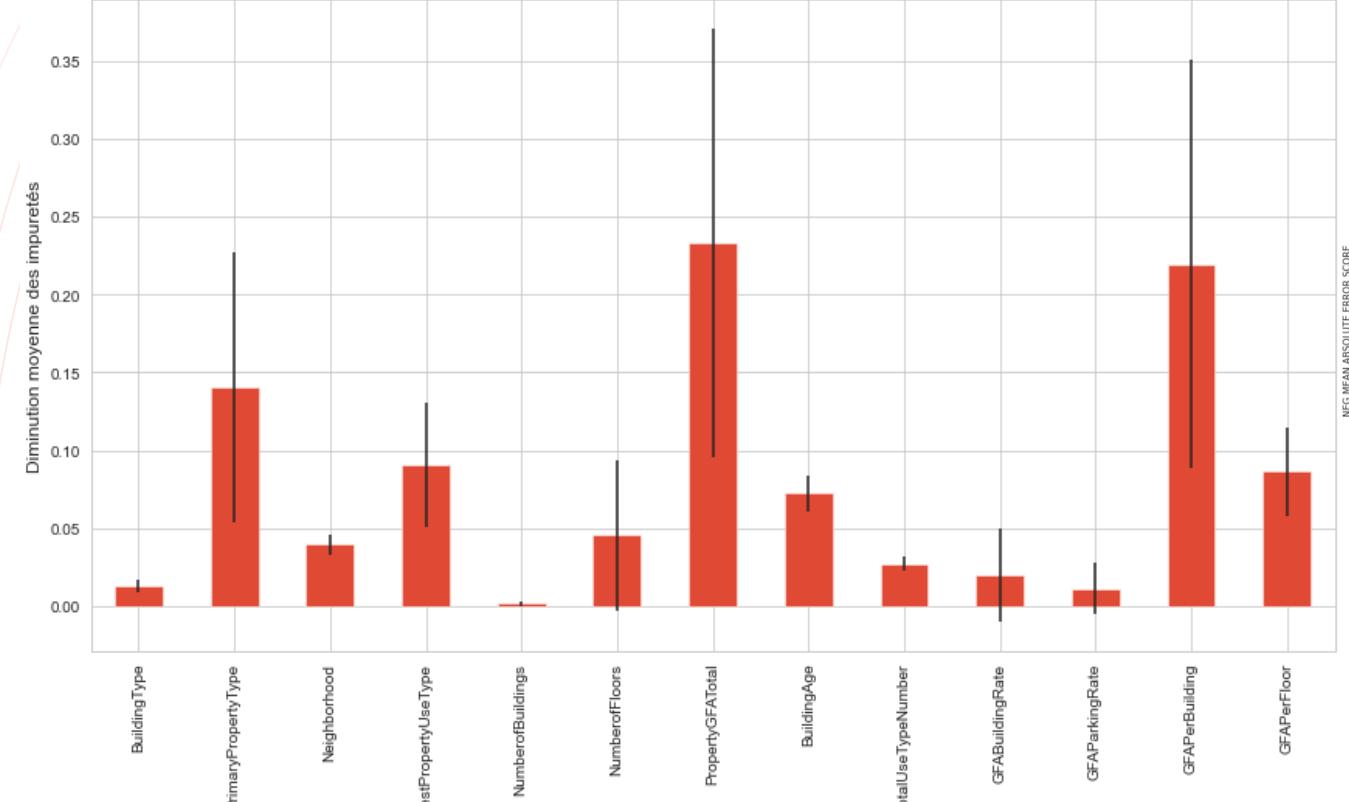


Feature importances du modèle RandomForestRegressor sur la consommation d'énergie

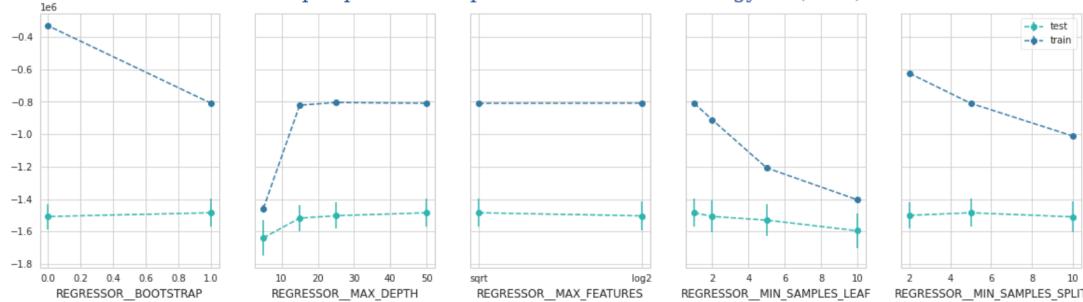


Sélection des meilleurs modèles : Consommations d'énergie

Feature importances du modèle RandomForestRegressor sur la consommation d'énergie



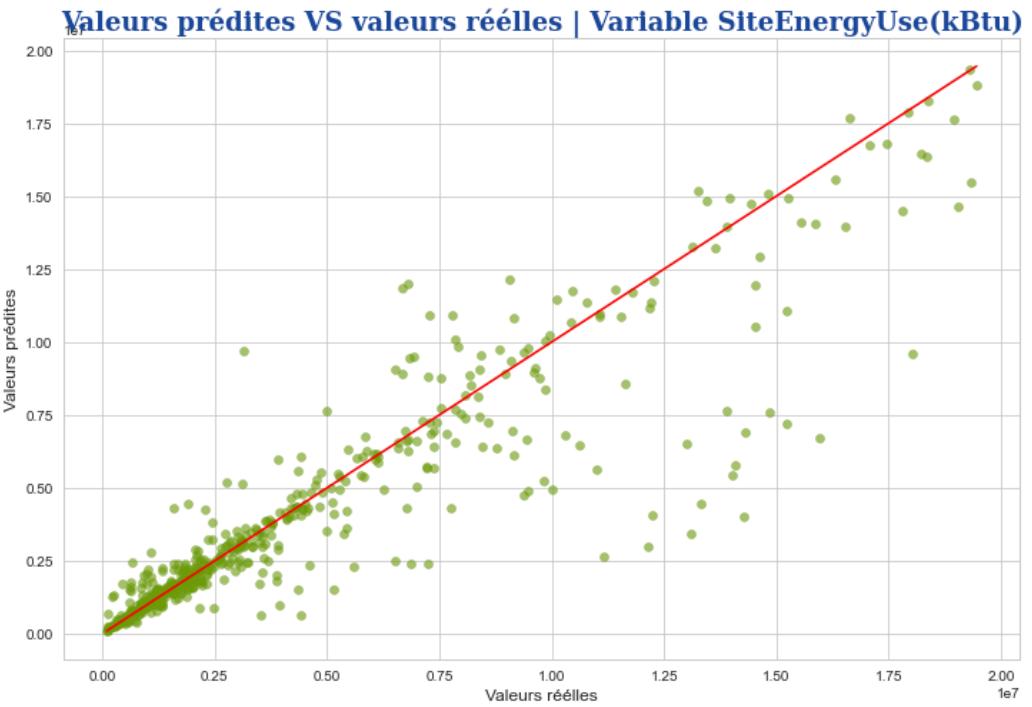
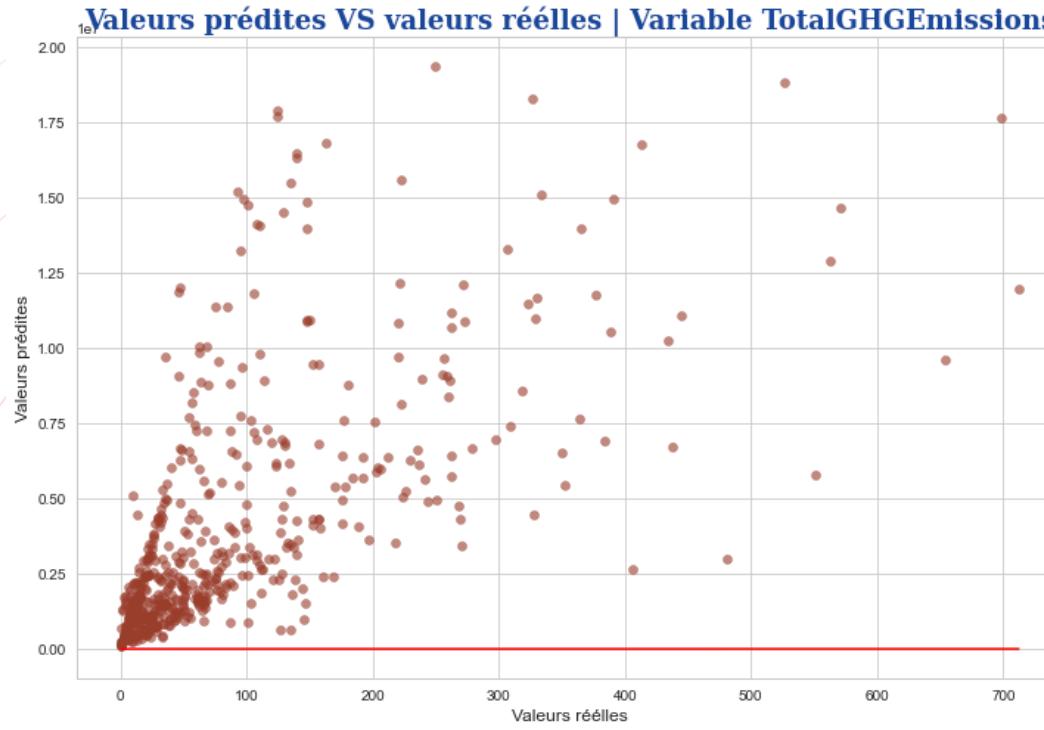
Scores par paramètres pour la variable SiteEnergyUse(kBtu)



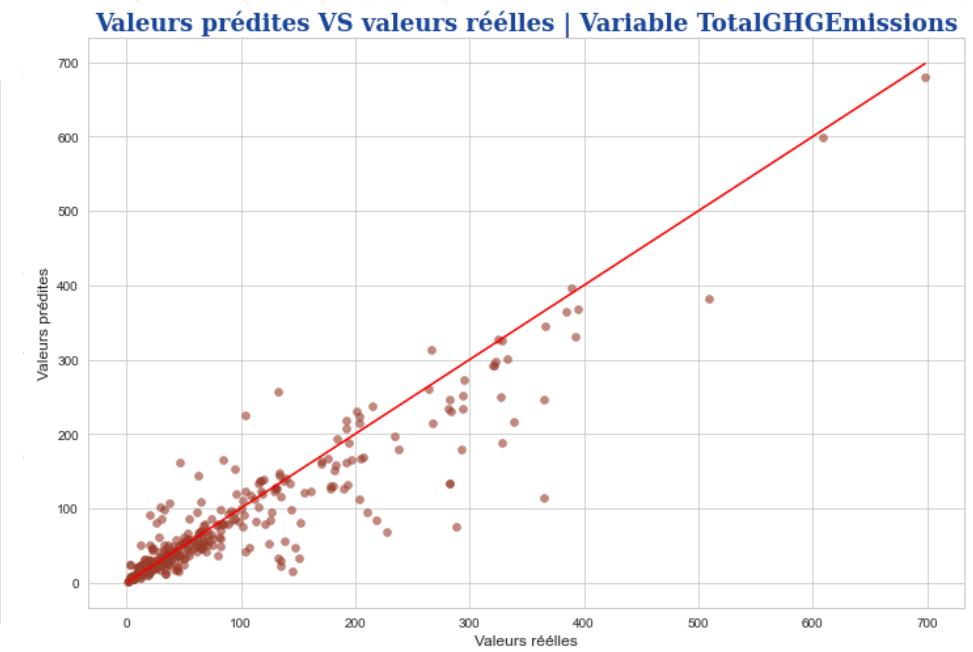
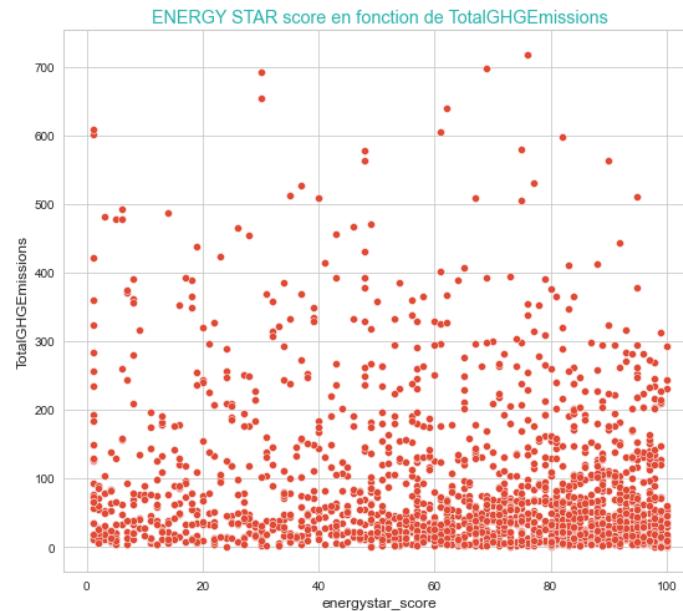
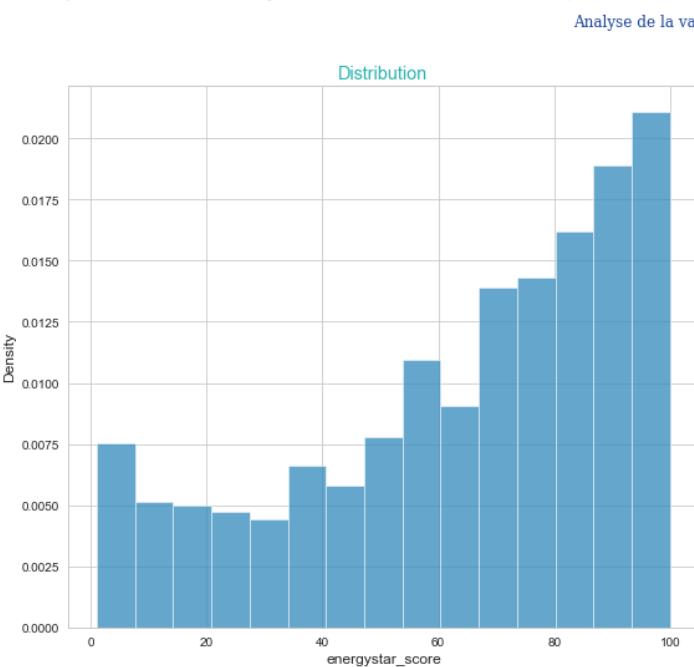
Rappel des meilleurs paramètres :

```
{'regressor__bootstrap': True, 'regressor__max_depth': 50, 'regressor__max_features': 'sqrt', 'regressor__min_samples_leaf': 1, 'regressor__min_samples_split': 5}
```

Test des modèles sélectionnés



Influence du score ENERGY STAR



Conclusions

- Base de données nettoyée : harmonisée, fusionnée, suppression des valeurs aberrantes ..
- Création de nouvelles variables
- Plusieurs modèles testés
- Random Forest permet d'obtenir les meilleures performances
- L'Energy star Score permet d'augmenter la performance, mais pas de manière significative
- Améliorations ...