Seattle

PROJET 4: ANTICIPEZ LES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS.

Sommaire

I- Présentation de la problématique.

Présentation de la problématique:

- *Données de consommation disponibles pour les bâtiments de la ville de Seattle pour les années 2015 et 2016.*
- *Coût important d'obtention des relevés/fastidieuses à collecter.*

Objectif du projet:

- Prédire les émissions de CO_2 et la consommation totale d'énergie sans les relevés annuels.
- Evaluer l'intérêt de l'ENERGY STAR Score.
- Mettre en place un modèle de prédiction réutilisable.

Seattle

Interprétation de la problématique:

1- Prévision

- Features: caractéristiques intrinsèques des bâtiments (hors consommations).
- Données à prédire
  - Consommation totale des bâtiments SiteEnergyUseWN(kBtu)
  - Emissions totales des bâtiments TotalGHGEmissions

=> 2 modèles différents

2-ENERGY STAR Score :

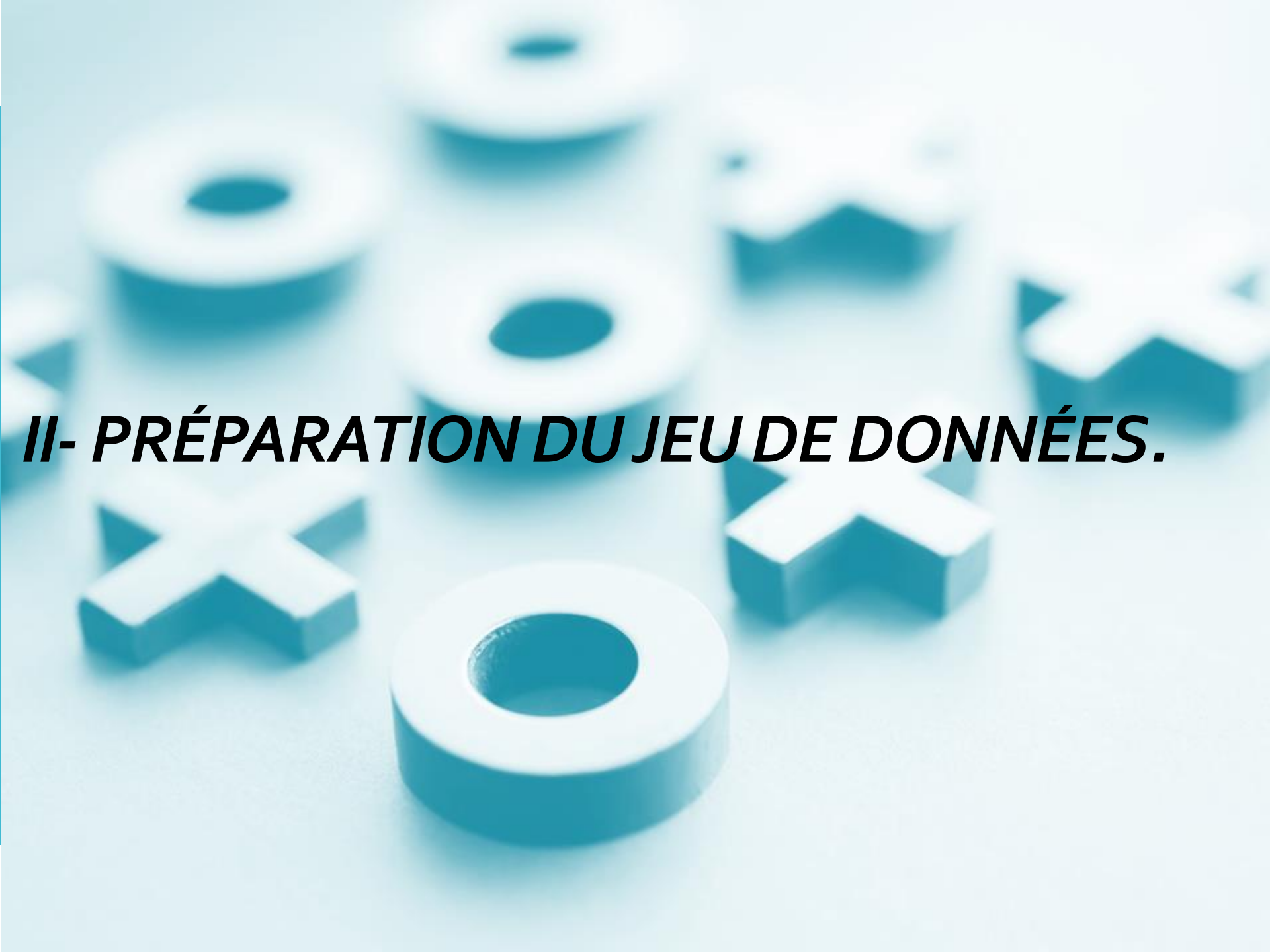
- Comparaison de son intérêt en essayant de modéliser avec et sans ce feature.

Présentation des données par années:

	OSEBuildingID	DataYear	BuildingType	PrimaryPropertyType	PropertyName	TaxParcelIdentificationNumber	Location
0	1	2015	NonResidential	Hotel	MAYFLOWER PARK HOTEL	659000030	{'human_address': '{"address":"405 OLIVE WAY","city":"SEATTLE","state":"WA","zip":"98101"}', 'latitude': '47.61219025', 'needs_recoding': False, 'longitude': '-122.33799744'}
1	2	2015	NonResidential	Hotel	PARAMOUNT HOTEL	659000220	{'human_address': '{"address":"724 PINE ST","city":"SEATTLE","state":"WA","zip":"98101"}', 'latitude': '47.61310583', 'needs_recoding': False, 'longitude': '-122.33335756'}
2	3	2015	NonResidential	Hotel	WESTIN HOTEL	659000475	{'human_address': '{"address":"1900 5TH AVE","city":"SEATTLE","state":"WA","zip":"98101"}', 'latitude': '47.61334897', 'needs_recoding': False, 'longitude': '-122.33769944'}

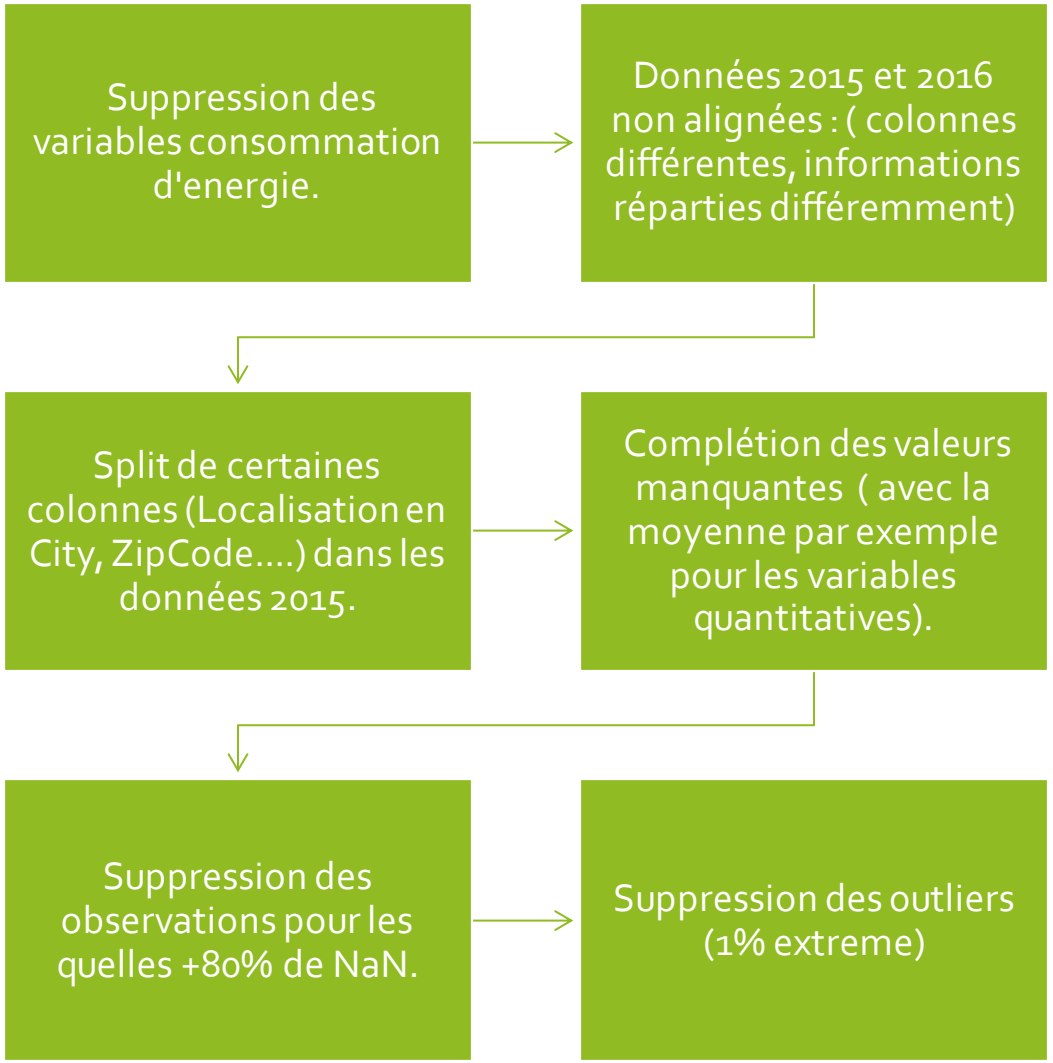
Données 2015: 3340 lignes, 42 features

Données 2016: 3376 lignes, 46 features

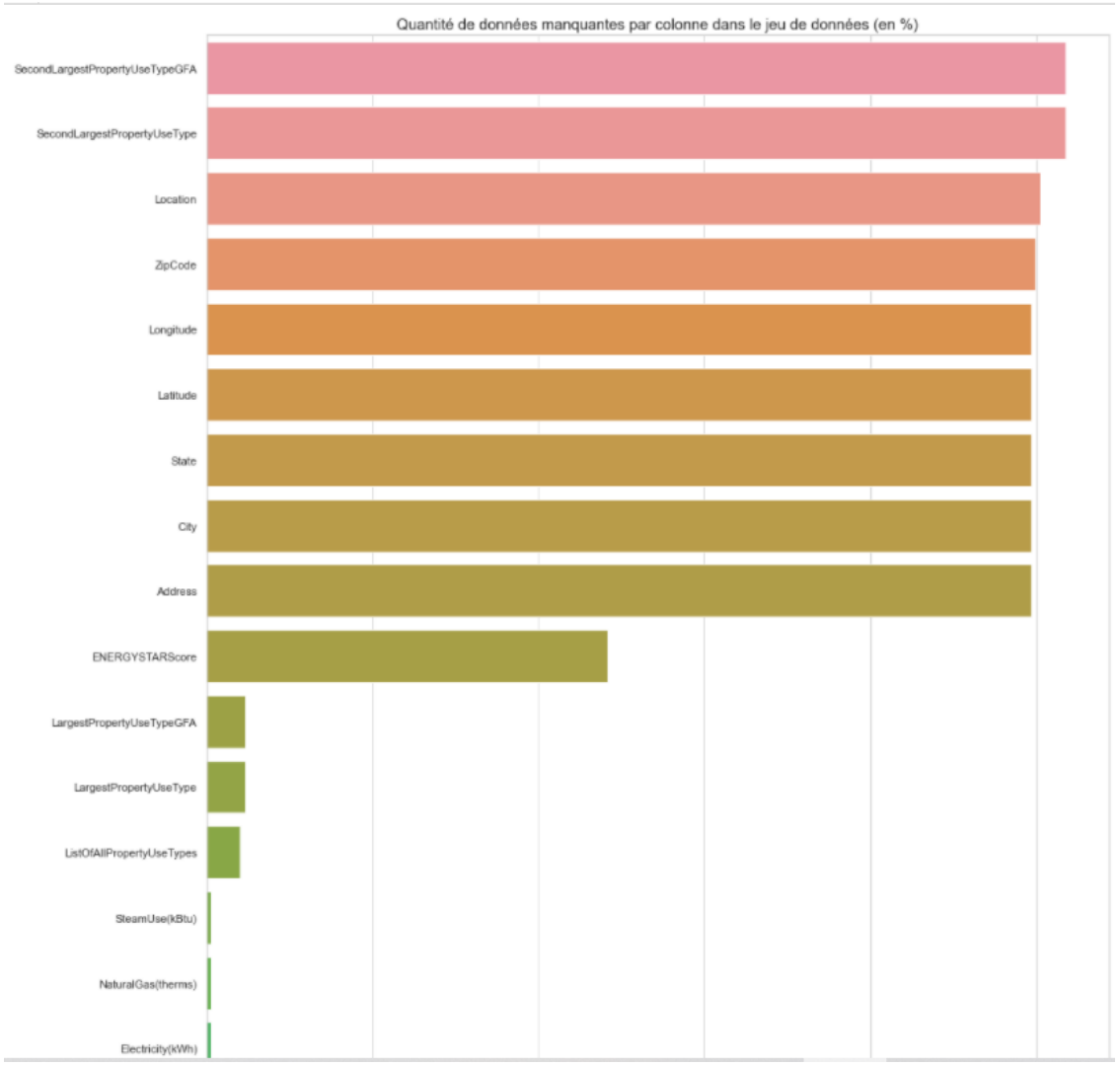


II- PRÉPARATION DU JEU DE DONNÉES.

Démarche Cleaning

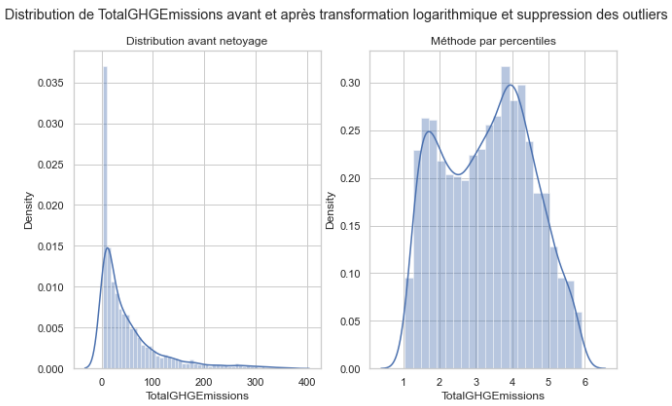


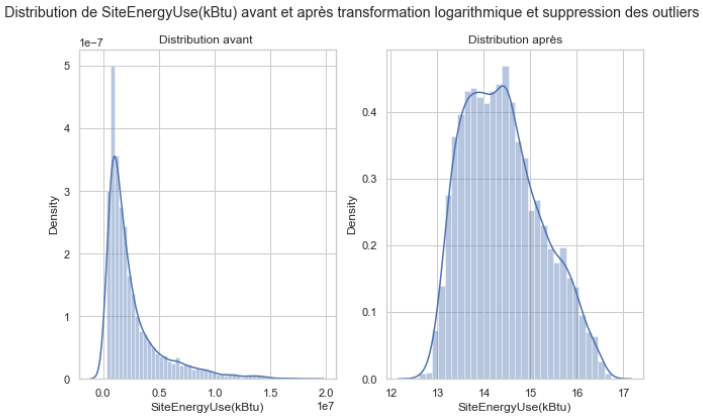
Suppression des features +80 % de NaN.



Feature engine ering

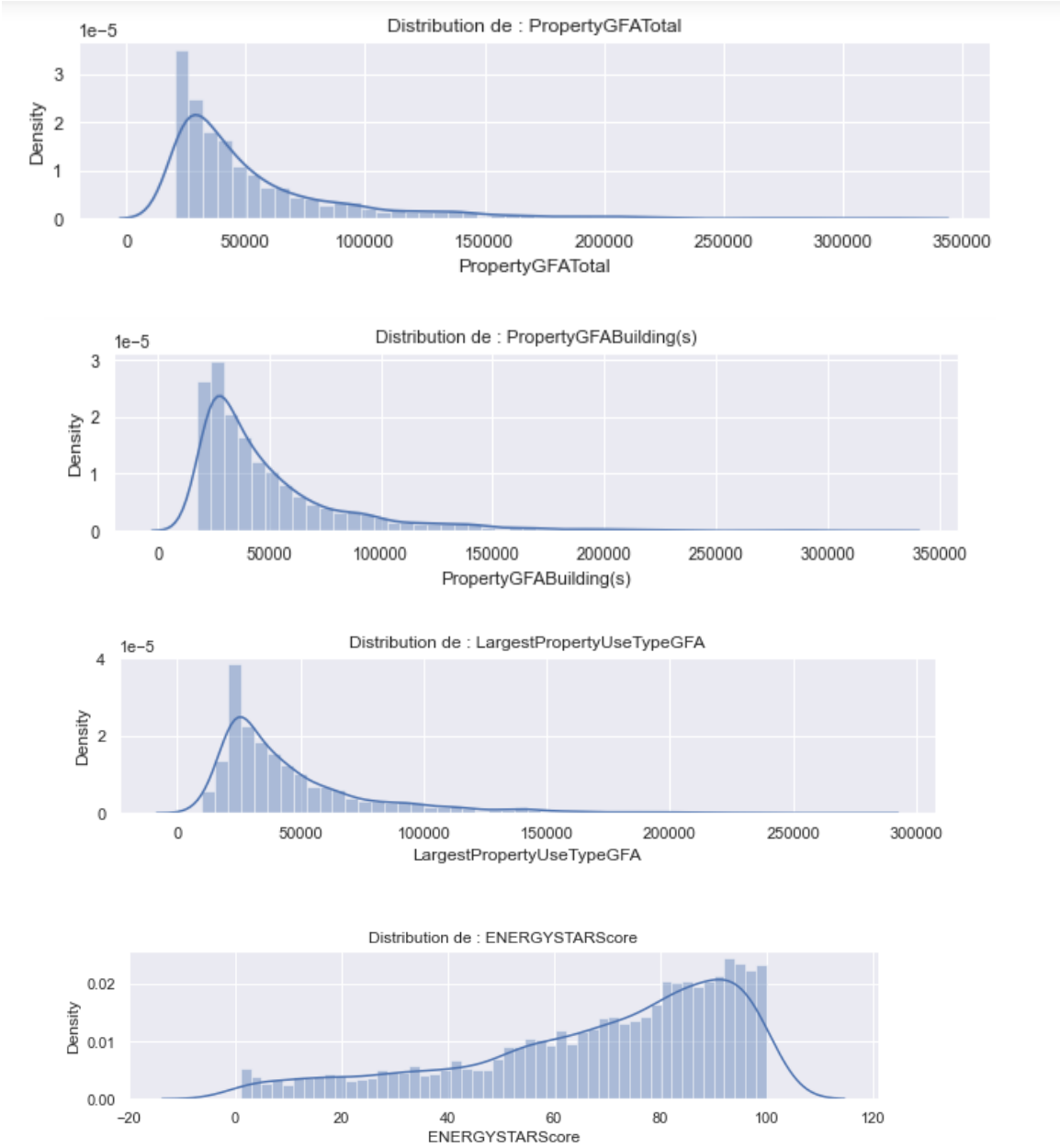
- Features liées à la proportion des sources d'énergie (coûteux à obtenir pour futures données)
- Utilisation du Energy Star score (mis de côté pour une analyse ultérieure) Idées retenues.
- Séparrer les features à prédire et les les autres features.
- Catégorisation (Encoding) des données pour certaines colonnes (I.e colonnes pertinents avec un nombre de catégorie raisonable).
- Suppression de colonnes non pertinentes pour notre modèle
  - Données sans catégorisation possible (example: Comment)
  - Données avec une unique information (exemple : State)
  - Données sans information pertinente pour le modèle
    - DefaultData : sens de la feature non expliqué + booléen avec beaucoup de NaN
    - SPD Beats : informations non utiles à la problématique + beaucoup de NaN
  - Features redondantes (address / zipcode remplacées par latitude et longitude)
- Log2-transformation variable de prédiction



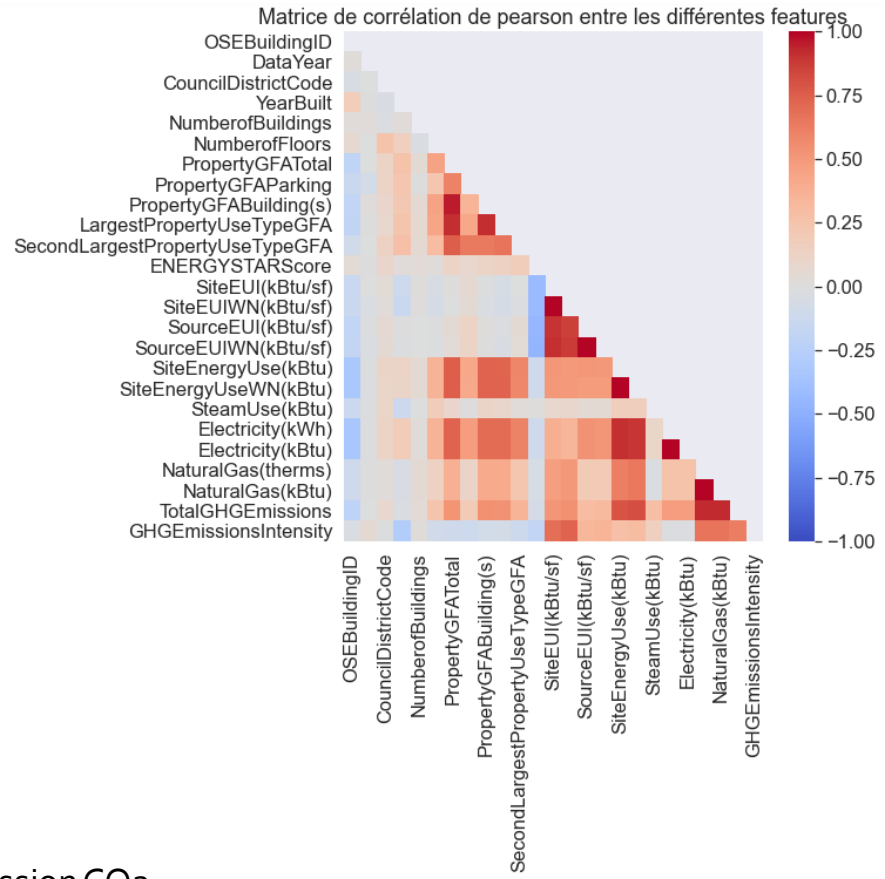


Feature engineering: Log2-transformation variable de prédiction

Analyse univariée: Distribution de quelques features.



Matrice de corrélation:



Emission CO_2:

- Une corrélation importante entre TotalGHGEmissions et PropertyGFATotal, PropertyGFABuilding, SiteEUI(kBtu/sf), SiteEnergyUse, NaturalGas(kBtu).

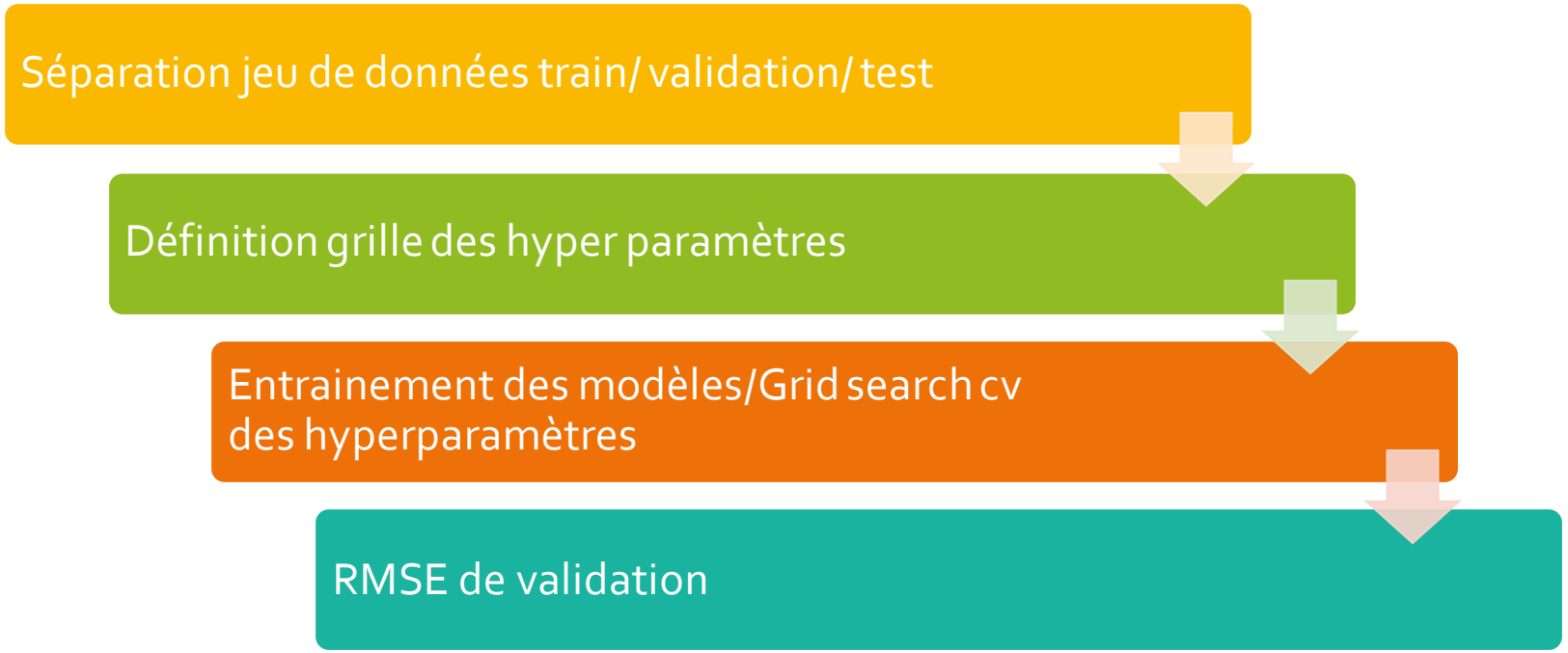
Consomation

- Une corrélation importante entre SiteEnergyUse(kBtu) et PropertyGFATotal, PropertyGFABuilding, NaturalGas(kBtu), etles autres variable de consommation(SiteEUI(kBtu/sf,..) .

Autres relation:

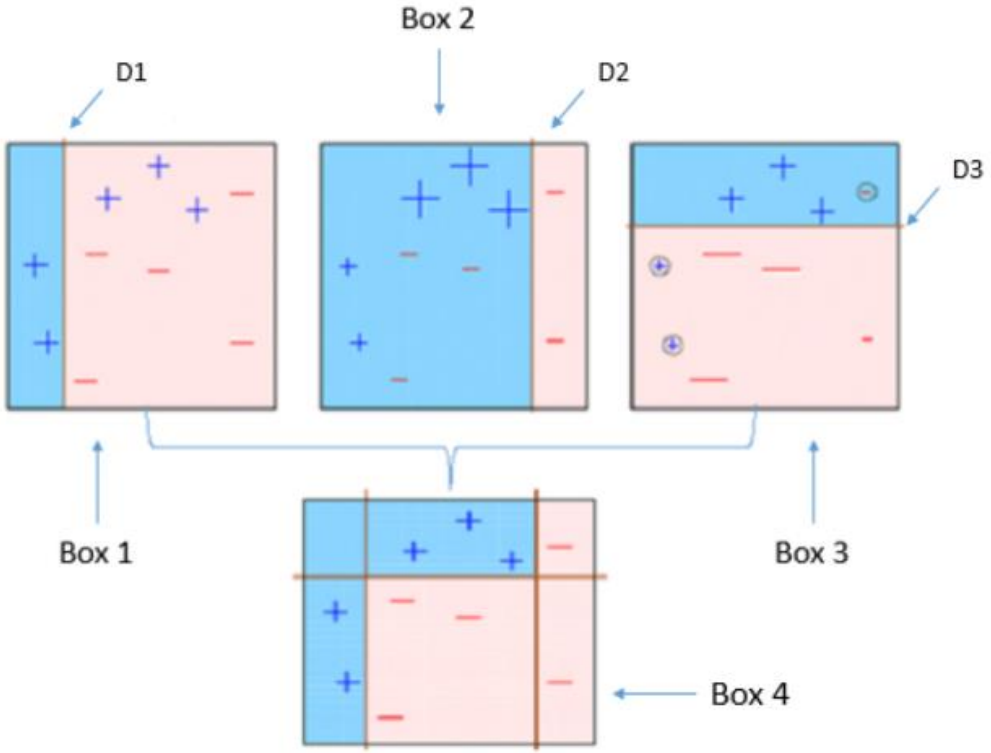
- On remarque une correlation entre les variable cibles TotalGHGEmissions et SiteEnergyUse(kBtu).
- SiteEUI(kBtu/sf) avec une forte correlation avec NaturalGas(kBtu).

III- Pistes de modélisation:

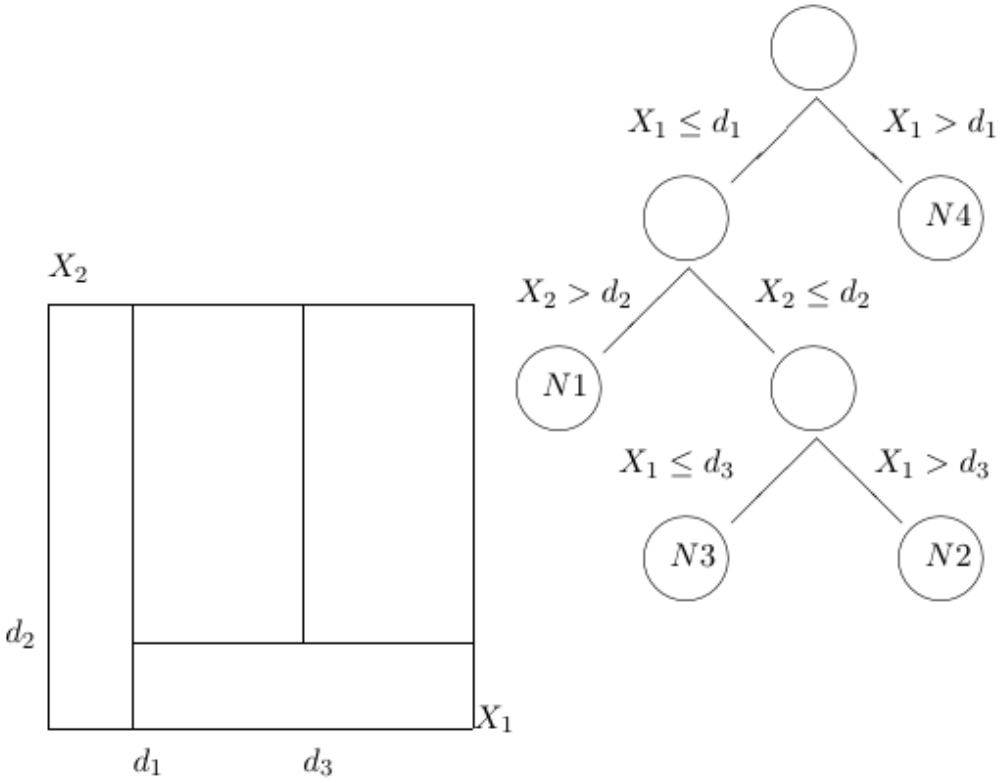


Démarche de modélisation

XGBoost/ Boosting Algorithm



Random Forest Regression/ Cart algorithm



Elastic Net	SVR	XGBoost	Random Forest Regressor
Alpha : [10^-4, 10^-3, ..., 10, 10^2]	Gamma : [10^-3,10^-2 , ... , 10^3]	N_estimators : 100	N_estimators : [10, 50, 100, 300, 500]
	C : [0.001, 0.01, 0.1, 1, 10]		Min_samples_leaf : [1, 3, 5, 10]

Elastic Net	SVR	XGBoost	Random Forest Regressor
Alpha: 0.001	Gamma : 0.01	N_estimators : 100	N_estimators: 300
1_ratio: 1.0	C : 100.0		Min_samples_leaf : 10

Définition des hyperparamètres à optimiser.

Impacte de Energy Star Score

Implémentation avec Energy Star score

```
Méthode: LinearRegression
+ utilisation de pipelines
Prédiction de  TotalGHGEmissions
score de la prédiction:  0.8273719938108743
MAE =  0.3654730349236491
RMSE =  0.46305661571507706
median abs err =  0.3200920051435352

Méthode: LinearRegression
+ utilisation de pipelines
Prédiction de  SiteEnergyUse(kBtu)
score de la prédiction:  0.8485060898652765
MAE =  0.2166901659378328
RMSE =  0.28998772909648024
median abs err =  0.16750426292299814
```

Implémentation sans Energy Star score

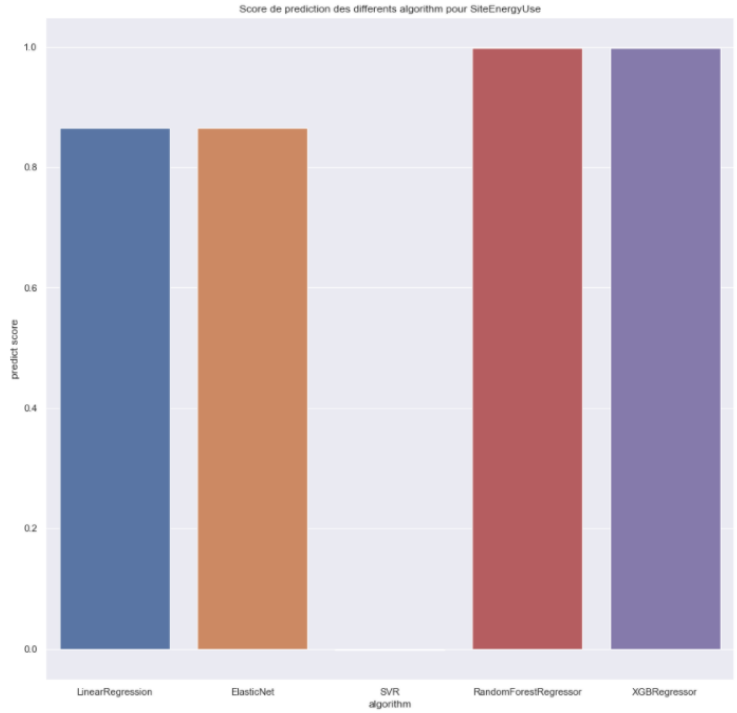
```
Méthode: LinearRegression
+ utilisation de pipelines
Prédiction de  TotalGHGEmissions
score de la prédiction:  0.8083916919344072
MAE =  0.3925692538620816
RMSE =  0.4998621974470686
median abs err =  0.3313018075356078

Méthode: LinearRegression
+ utilisation de pipelines
Prédiction de  SiteEnergyUse(kBtu)
score de la prédiction:  0.8271374436353547
MAE =  0.22519245899299464
RMSE =  0.31480854716911943
median abs err =  0.172474535399056
```

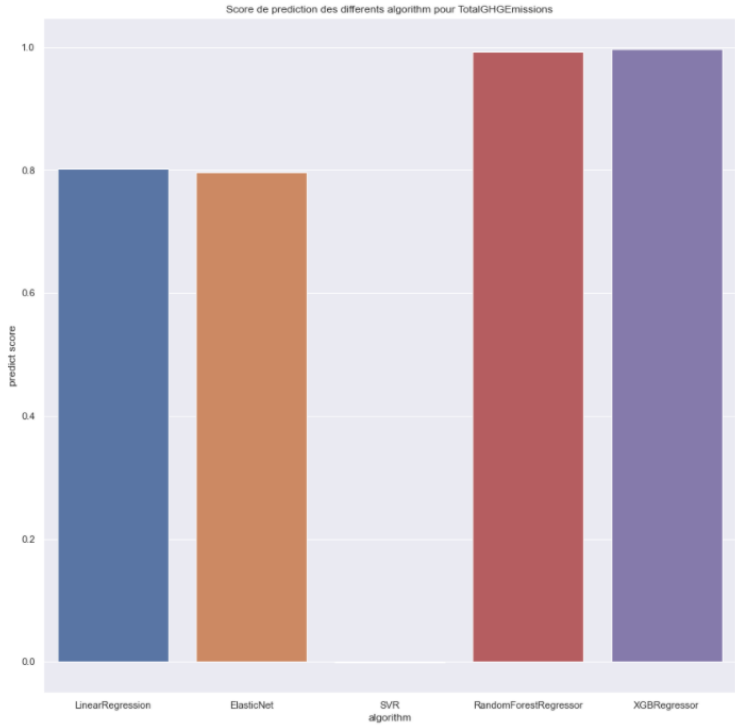
- Une lègère augmentation de la RMSE pour la prediction de la TotalGHGEmissions et SiteEnergyUse sans Energy Star Score. Energy Star Score impacte légèrement la prédiction des variables cible.

	index	algorithm	column	predict score	MAE	RMSE	median abs err
0	8	XGBRegressor	TotalGHGEmissions	0.997139	0.033594	0.059385	0.018327
1	6	RandomForestRegressor	TotalGHGEmissions	0.992897	0.040774	0.093576	0.016274
2	0	LinearRegression	TotalGHGEmissions	0.802517	0.374326	0.493412	0.326074
3	2	ElasticNet	TotalGHGEmissions	0.796517	0.379840	0.500851	0.325355
4	4	SVR	TotalGHGEmissions	-0.000349	0.851300	1.110504	0.783498
5	9	XGBRegressor	SiteEnergyUse(kBtu)	0.998457	0.019769	0.029721	0.013932
6	7	RandomForestRegressor	SiteEnergyUse(kBtu)	0.997875	0.022759	0.034883	0.015724
7	1	LinearRegression	SiteEnergyUse(kBtu)	0.865437	0.201397	0.277562	0.163357
8	3	ElasticNet	SiteEnergyUse(kBtu)	0.864789	0.201721	0.278229	0.166993
9	5	SVR	SiteEnergyUse(kBtu)	-0.001409	0.559644	0.757186	0.436308

Comparaison des differents algorithms.

Le meilleur algorithm en terme de score de prediction est XGBoost ou Boosting suivis par Random Forest suivis par LinearRegression suivis par ElasticNet et au dernier rang SVR pour TotalGHGEmissions et SIteEnergyUse.

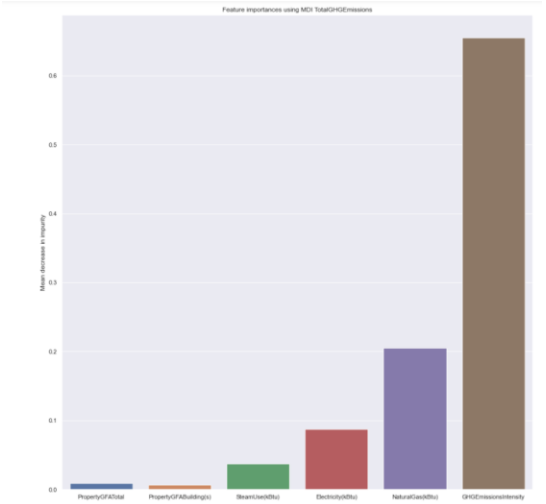


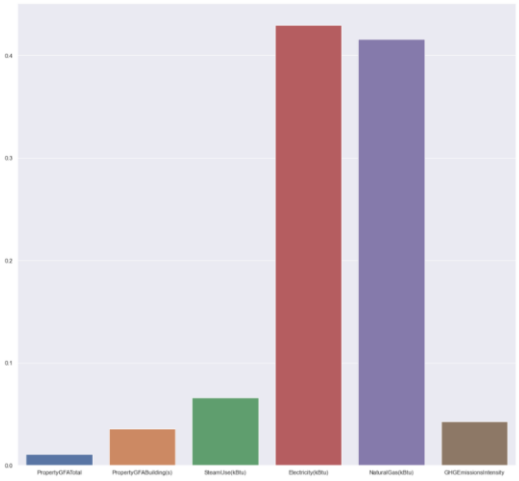


Comparaison des differents algorithms.

Importance des variables selectionnées (Gradient Boosting model)

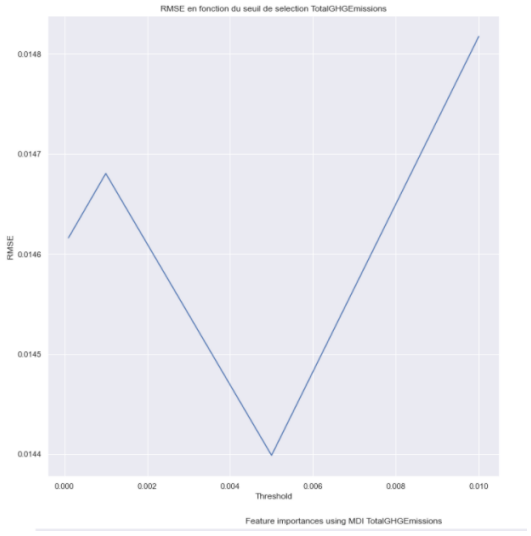
On est parvenu de limiter notre selection de variable à 5 et 6 feature tout en gardant les même performances RMSE et score

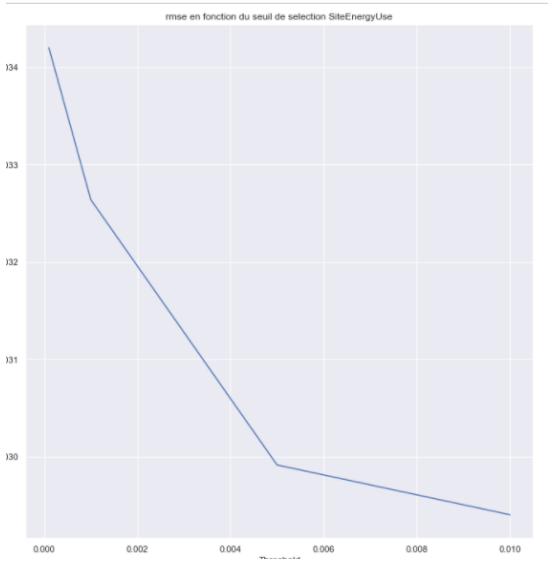




RMSE en fonction des seuils de selection (Gradient Boosting model)

Meilleur seuil de *selection* 0.001 energyUse et 0.005 pour TotalGHGEmissions






Conclusion: