Anticipez les besoins en consommation de bâtiments

Seattle

PLE Coline

PROJET 4

14/07/2023

Plan de la présentation

- Présentation de la problématique
- Présentation du jeu de données
- Feature engineering
- Modélisations: Approche et résultats
- Energy Star Score (émission gaz à effet de serre)?
- Discussion

RAPPEL DE LA PROBLEMATIQUE

Objectif de la ville de Seattle

Ville neutre en émission de carbone en 2050

Focus de l'équipe

Consommation et émission des bâtiments non destinées à l'habitation

Mission

Prédiction des émissions des gaz à effet de serre et de la consommation totale d'énergie de bâtiments **non destinés à l'habitation** pour lesquels les mesures n'ont pas encore été réalisées

Problématique

Relevés couteux

Mise à disposition des mesures déjà effectuées en 2016

\+

ENERGY STAR

Calcul fastidieux

Par l'approche utilisée dans notre équipe

=

Prédiction par ML

PRESENTATION ET NETTOYAGE DU JEU DE DONNEES

```
2 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3376 entries, 0 to 3375
Data columns (total 46 columns):
```

- Relevés énergétiques de 2016 de 3376 propriétés de Seattle
- 46 variables (Superficies, localisation, conso énergétique et émission de gaz à effet de serre)
- Source: https://data.seattle.gov/dataset/2016-Building-Energy-Benchmarking
- 0 doublon

NETTOYAGE DES DONNEES

1. Elimination des propriétés destinées à l'habitation

```
df["BuildingType"].unique()
```

```
'NonResidential'
'Nonresidential WA'
'Nonresidential COS'
'SPS-District K-12'
'Campus'
'Multifamily LR (1-4)'
'Multifamily MR (5-9)'
'Multifamily HR (10+)'
```

1708

2. Conservation des propriétés conformes

```
1 df["ComplianceStatus"].value_counts()
```

```
Compliant                      1546
Error - Correct Default Data     88
Non-Compliant                    17
Missing Data                     14
```

3. Conservation des données exprimées en kBtu pour la consommation en énergie et les émissions de gaz à effet de serre (homogénéité)

```
2 df.info()
```

```
SteamUse(kBtu)
Electricity(kWh)
Electricity(kBtu)
NaturalGas(therms)
NaturalGas(kBtu)
```

4. Choix des 2 cibles/targets

Cible Consommation en énergie

```
SiteEUI(kBtu/sf)
SiteEUIWN(kBtu/sf)
SourceEUI(kBtu/sf)
SourceEUIWN(kBtu/sf)
SiteEnergyUse(kBtu)
SiteEnergyUseWN(kBtu)
```

Consommation ramenée à la surface du bâtiment (intensité de consommation)

Consommation normalisée sur les conditions météorologiques

Cible Emission des gaz à effet de serre

```
TotalGHGEmissions
GHGEmissionsIntensity
```

Intensité de consommation (homogène par rapport au choix de la 1ère target)

Elimination des cibles/targets potentielles non retenues

PRESENTATION ET NETTOYAGE DU JEU DE DONNEES

NETTOYAGE DES DONNEES

4. Gestion des données erronées (quelques exemples)

```
# Liste du nombre de bâtiments
print(df_1["NumberofBuildings"].unique().tolist())
```

```
[1.0, 3.0, 0.0, 2.0, 4.0, 27.0, 6.0, 11.0, 14.0, 9.0, 7.0, 5.0, 8.0, 23.0, 10.0, 111.0]
```

52 propriétés = 1

NumberofFloors	SiteEUIWN(kBtu/sf)
0	0.0

```
print(df_1["NumberofFloors"].unique().tolist())
```

```
[12, 11, 41, 10, 18, 2, 8, 15, 25, 9, 33, 6, 28, 5, 19, 7, 4, 3, 24, 20, 34, 1, 0, 16, 23, 17, 36, 22, 47, 29, 14, 49, 37, 42, 63, 13, 21, 55, 46, 30, 56, 76, 27, 99]
```

= 2

```
# Description pour le nombre d'étages
culte["NumberofFloors"].describe()
```

```
count    69.000000
mean      3.347826
std      11.709506
min       1.000000
25%       1.000000
50%       2.000000
75%       2.000000
max      99.000000
```

Les superficies

Enormément d'incohérences!!!

Modifications se basant sur les variables 'ListOfAllPropertyUseTypes', 'LargestPropertyUseTypeGFA', 'SecondLargestPropertyUseTypeGFA' et 'ThirdLargestPropertyUseTypeGFA'.

PropertyGFABuilding(s)	ListOfAllPropertyUseTypes	LargestPropertyUseType	LargestPropertyUseTypeGFA	SecondLargestPropertyUseType
99005	Office	Office	79555.0	NaN

5. Gestion des données manquantes

<u>Les superficies</u>

- Remplacement par Nothing si absence d'activité secondaire ou tertiaire pour les variables PropertyUsedType
- Remplacement par 0 si absence d'activité secondaire ou tertiaire pour les variables PropertyUsedTypeGFA

<u>Les ZIP Code</u>

Recherche sur internet *via* l'adresse de la propriété

PRESENTATION ET NETTOYAGE DU JEU DE DONNEES

Seattle

GESTION DES VALEURS ABERRANTES ET OUTLIERS

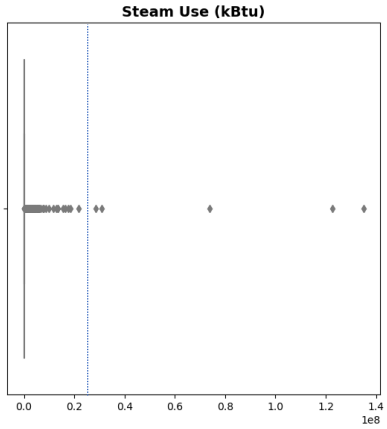
1. Les propriétés présentant des consommations énergétiques ou des émissions de gaz à effet de serre < ou = à 0 (zéro OK pour SteamUse)

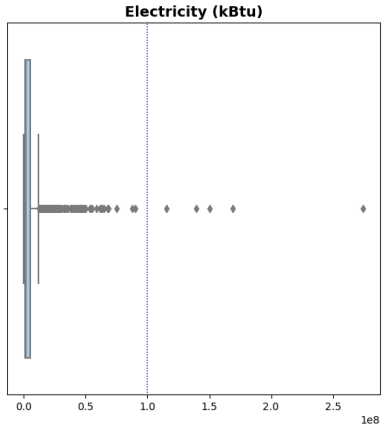
```
1 print(f'{valeur_nulle.shape[0]} propriétés présentent des valeurs aberrantes.')
```

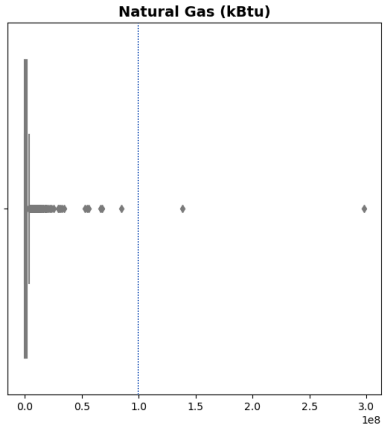
```
12 propriétés présentent des valeurs aberrantes.
```

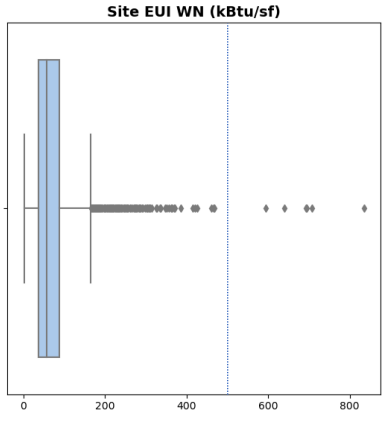
SiteEnergyUse(kBtu)	SiteEnergyUseWN(kBtu)	SteamUse(kBtu)	Electricity(kBtu)	NaturalGas(kBtu)
1.150804e+07	1.185445e+07	0.00	0.0	11508035.0
1.252517e+07	1.284386e+07	0.00	0.0	0.0

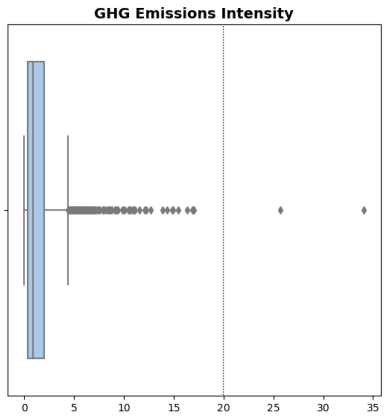
2. La gestion des outliers/valeurs atypiques









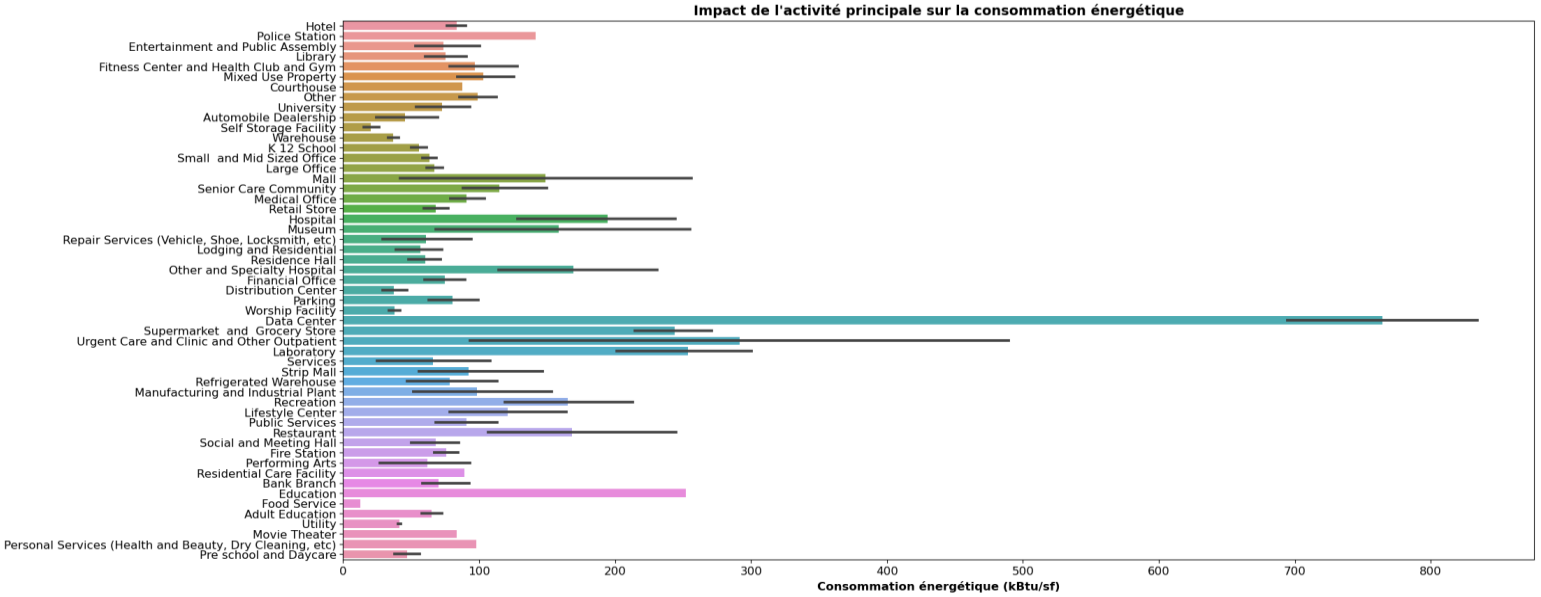


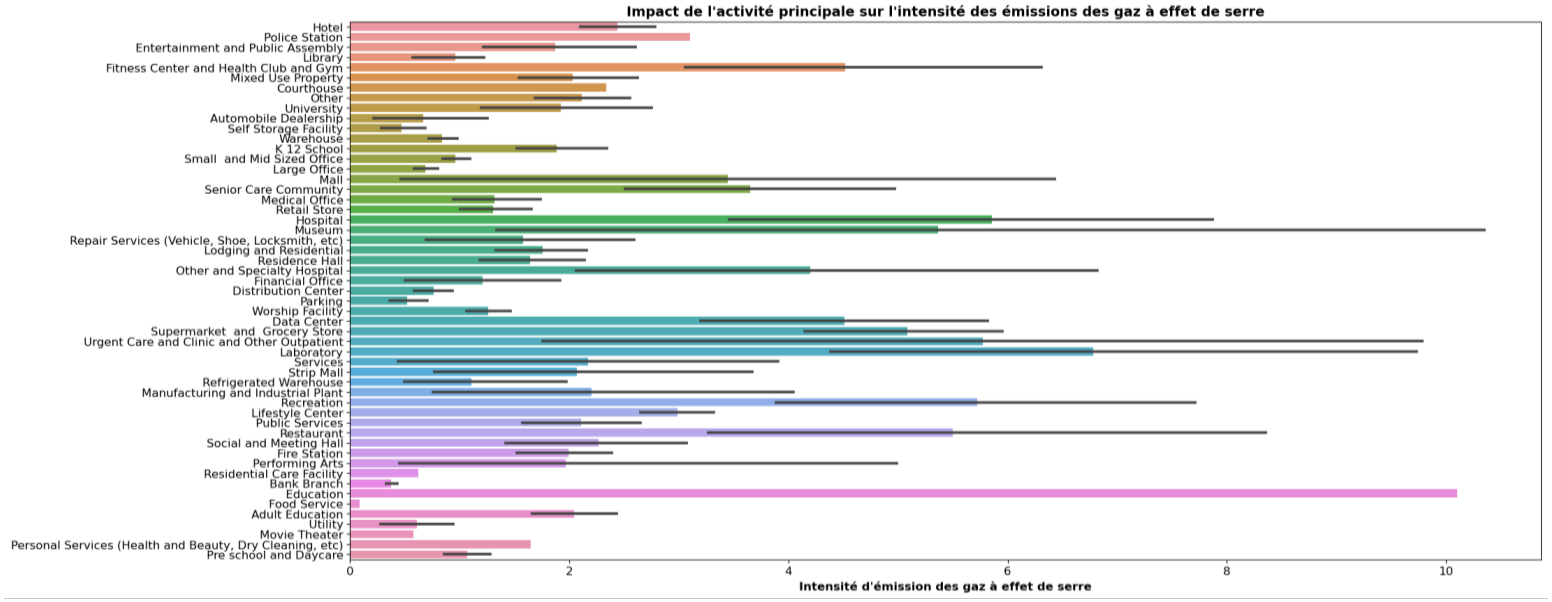
Seuil de tolérance

1483 propriétés entièrement renseignées pour toutes les variables à l'exception de l'Energy Star Score (967 données)

VISUALISATIONS PRINCIPALES DES DONNEES:
Impact de l'activité principale sur les 2 cibles

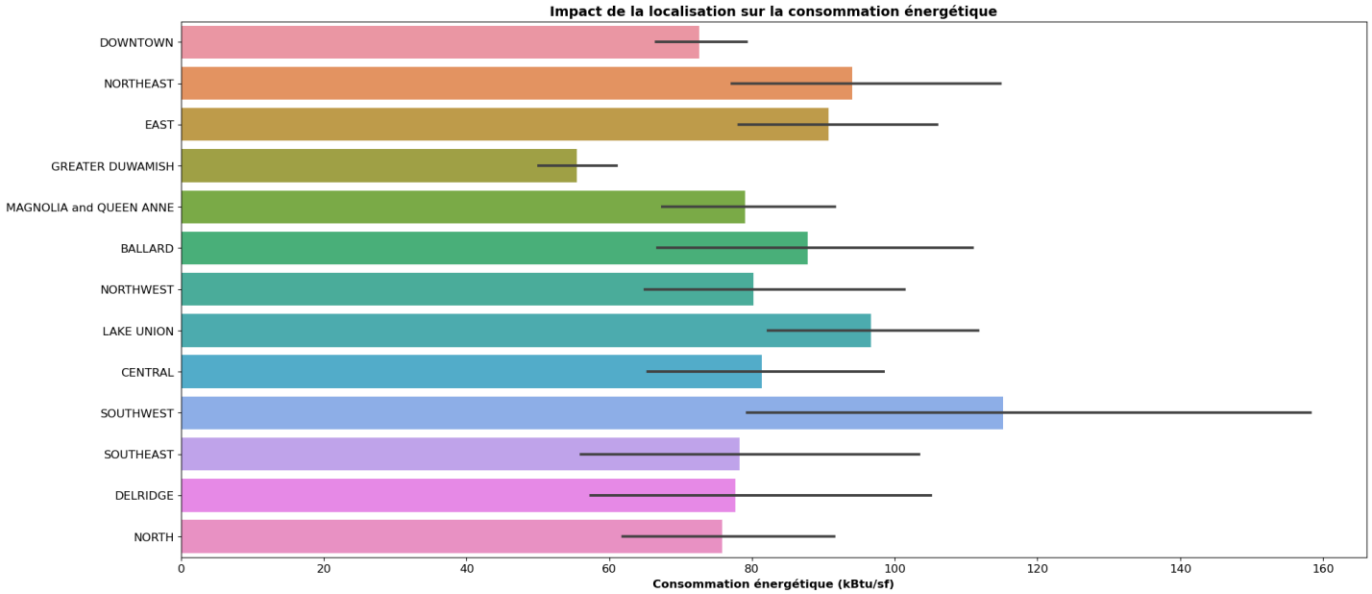
Seattle

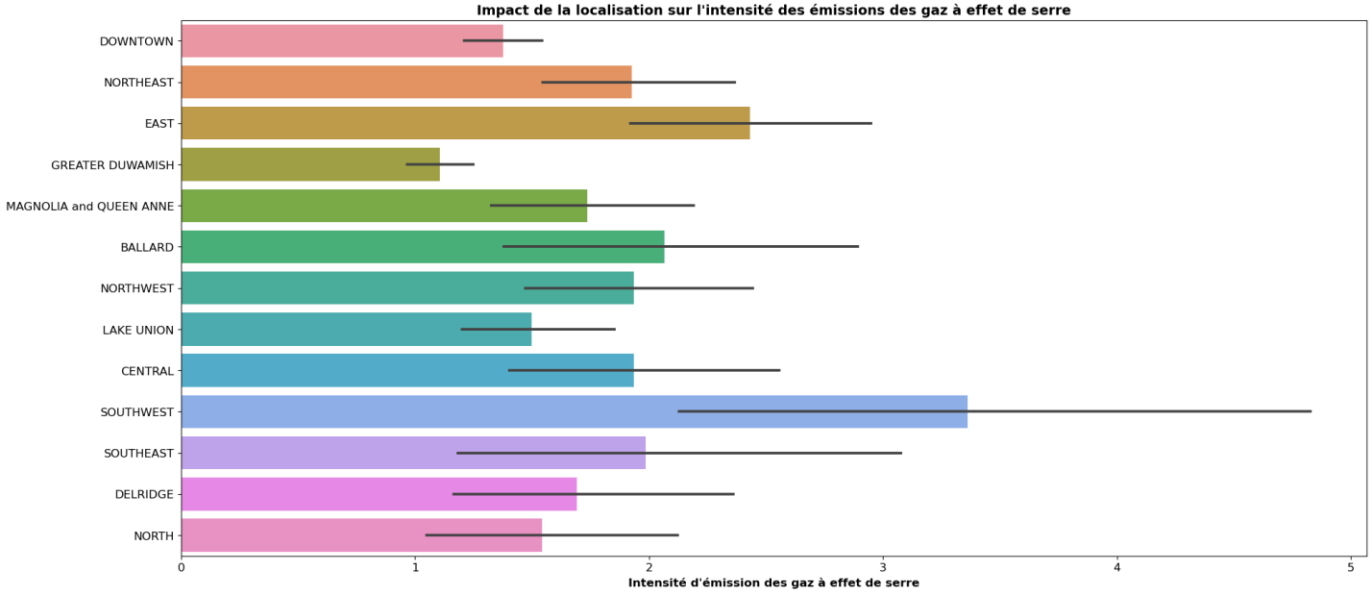




VISUALISATIONS PRINCIPALES DES DONNEES:
Impact de la localisation sur les 2 cibles

Seattle





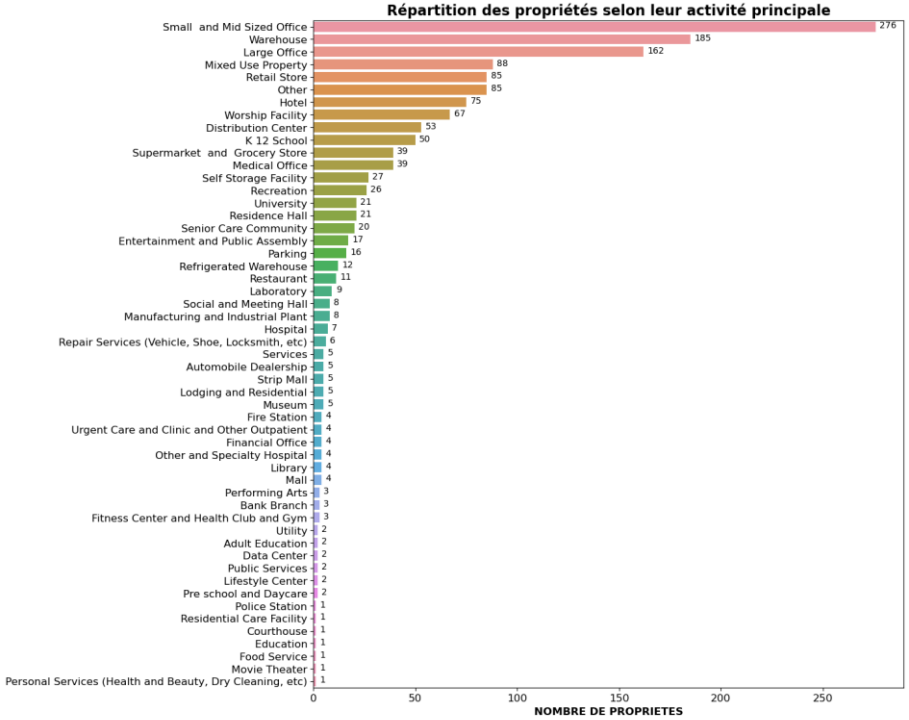
FEATURE ENGINEERING

1. Passage en pourcentage des différents types d'énergie avant leur élimination

> **Douglas** : L'objectif est de te passer des relevés de consommation annuels futurs (attention à la fuite de données). Nous ferons de toute façon pour tout nouveau bâtiment un premier relevé de référence la première année, donc rien ne t'interdit d'en déduire des variables structurelles aux bâtiments, par exemple la nature et proportions des sources d'énergie utilisées..

2. Création d'une feature Age de la propriété avant élimination de la variable 'YearBuilt'

3. Création de nouvelles variables à partir des variables existantes pour les types d'activités



- Présence d'activités très peu représentées
- Regroupement permettant à la fois un meilleur apprentissage et un encodage plus simple

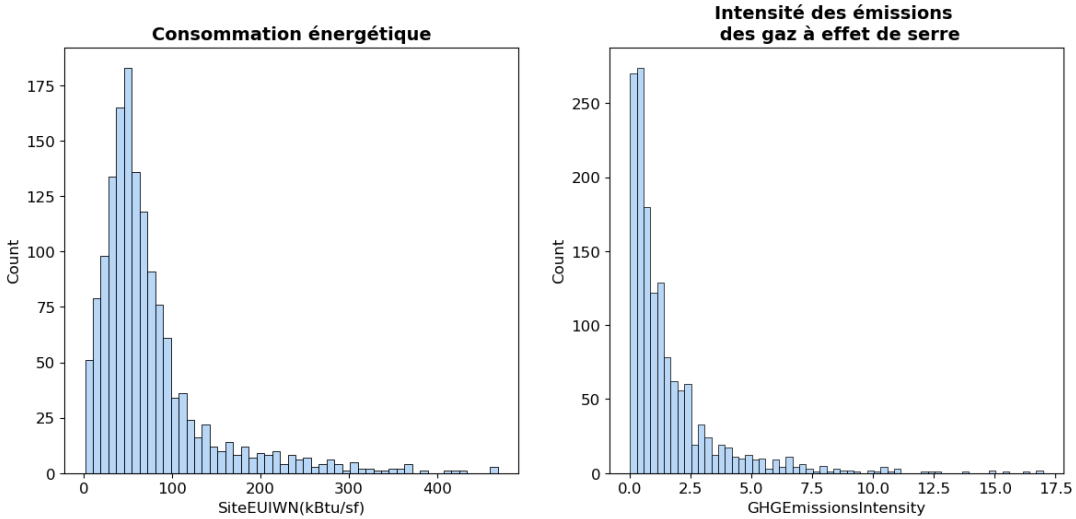
1. Hébergement
2. Entrepôts réfrigérés et Data Center
3. Stockage
4. Santé et Recherche
5. Restauration
6. Gros commerces
7. Petits commerces
8. Education
9. Bureaux
10. Autres et services
11. Les usines
12. Parking
13. Nothing

- Encodage manuel (% des activités en fonction de la superficie totale de la propriété)

FEATURE ENGINEERING

4. Passage à l'échelle logarithmique des 2 cibles présentant une distribution non normale

AVANT PASSAGE

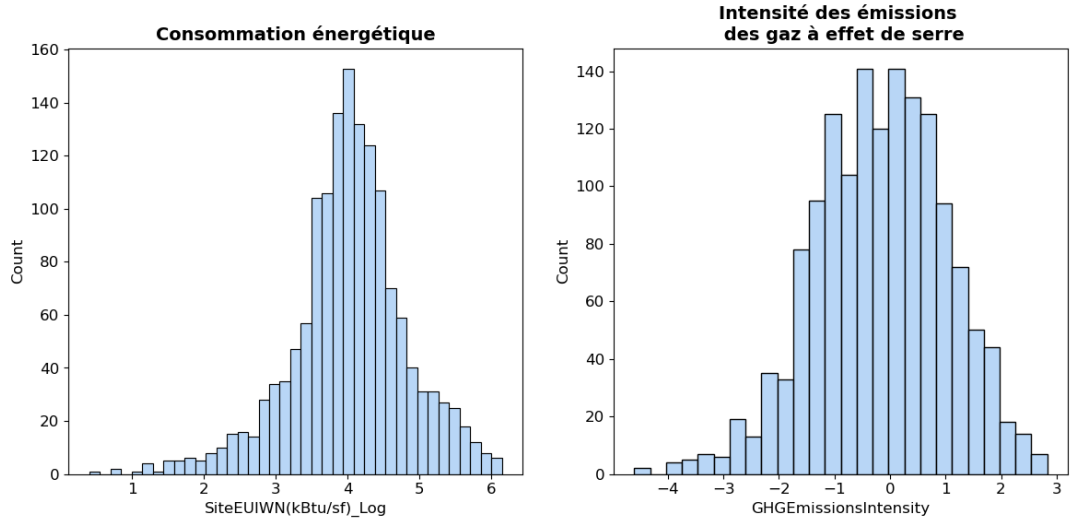


Test de Skew

```
Test sur la target énergie: 23.3889295701429

Test sur la target émission: 26.672138901291113
```

APRES PASSAGE (FunctionTransformer de Sklearn)



Les cibles seront passées au Log lors des modélisations (meilleur apprentissage)

FEATURE ENGINEERING

Seattle

5. Choix de la variable de localisation : Neighborhood

6. Création du jeu de modélisation et du jeu de validation

JEU DE MODELISATION APRES ENCODAGE

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 967 entries, 0 to 966
Data columns (total 36 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   NumberofBuildings                       967 non-null    float64
 1   NumberofFloors                          967 non-null    int64
 2   PropertyGFATotal                        967 non-null    float64
 3   ENERGYSTARScore                         967 non-null    float64
 4   SiteEUIWN(kBtu/sf)                      967 non-null    float64
 5   GHGEmissionsIntensity                   967 non-null    float64
 6   SteamUse(%)                             967 non-null    float64
 7   Electricity(%)                          967 non-null    float64
 8   NaturalGas(%)                           967 non-null    float64
 9   AgePropriete                            967 non-null    int64
 10  Usines                                  967 non-null    float64
 11  Bureaux                                 967 non-null    float64
 12  Santé et Recherche                      967 non-null    float64
 13  Autre et services                       967 non-null    float64
 14  Stockage                                967 non-null    float64
 15  Hébergement                             967 non-null    float64
 16  Petits commerces                        967 non-null    float64
 17  Entrepôts réfrigérés et Data Center     967 non-null    float64
 18  Parking                                 967 non-null    float64
 19  Education                               967 non-null    float64
 20  Activités sociales                      967 non-null    float64
 21  Restauration                            967 non-null    float64
 22  Gros commerces                          967 non-null    float64
 23  Neighborhood_ballard                    967 non-null    int64
 24  Neighborhood_central                    967 non-null    int64
 25  Neighborhood_delridge                   967 non-null    int64
 26  Neighborhood_downtown                   967 non-null    int64
 27  Neighborhood_east                       967 non-null    int64
 28  Neighborhood_greater duwamish           967 non-null    int64
 29  Neighborhood_lake union                 967 non-null    int64
 30  Neighborhood_magnolia and queen anne    967 non-null    int64
 31  Neighborhood_north                      967 non-null    int64
 32  Neighborhood_northeast                  967 non-null    int64
 33  Neighborhood_northwest                  967 non-null    int64
 34  Neighborhood_southeast                  967 non-null    int64
 35  Neighborhood_southwest                  967 non-null    int64
dtypes: float64(21), int64(15)
```

- 80% des données dans le jeu d'entraînement
- 20% des données dans le jeu test

JEU DE VALIDATION APRES ENCODAGE

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 516 entries, 0 to 515
Data columns (total 36 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   NumberofBuildings                       516 non-null    float64
 1   NumberofFloors                          516 non-null    int64
 2   PropertyGFATotal                        516 non-null    float64
 3   ENERGYSTARScore                         0 non-null      float64
 4   SiteEUIWN(kBtu/sf)                      516 non-null    float64
 5   GHGEmissionsIntensity                   516 non-null    float64
 6   SteamUse(%)                             516 non-null    float64
 7   Electricity(%)                          516 non-null    float64
 8   NaturalGas(%)                           516 non-null    float64
 9   AgePropriete                            516 non-null    int64
 10  Usines                                  516 non-null    float64
 11  Bureaux                                 516 non-null    float64
 12  Santé et Recherche                      516 non-null    float64
 13  Autre et services                       516 non-null    float64
 14  Stockage                                516 non-null    float64
 15  Hébergement                             516 non-null    float64
 16  Petits commerces                        516 non-null    float64
 17  Entrepôts réfrigérés et Data Center     516 non-null    float64
 18  Parking                                 516 non-null    float64
 19  Education                               516 non-null    float64
 20  Activités sociales                      516 non-null    float64
 21  Restauration                            516 non-null    float64
 22  Gros commerces                          516 non-null    float64
 23  Neighborhood_ballard                    516 non-null    int64
 24  Neighborhood_central                    516 non-null    int64
 25  Neighborhood_delridge                   516 non-null    int64
 26  Neighborhood_downtown                   516 non-null    int64
 27  Neighborhood_east                       516 non-null    int64
 28  Neighborhood_greater duwamish           516 non-null    int64
 29  Neighborhood_lake union                 516 non-null    int64
 30  Neighborhood_magnolia and queen anne    516 non-null    int64
 31  Neighborhood_north                      516 non-null    int64
 32  Neighborhood_northeast                  516 non-null    int64
 33  Neighborhood_northwest                  516 non-null    int64
 34  Neighborhood_southeast                  516 non-null    int64
 35  Neighborhood_southwest                  516 non-null    int64
dtypes: float64(21), int64(15)
```

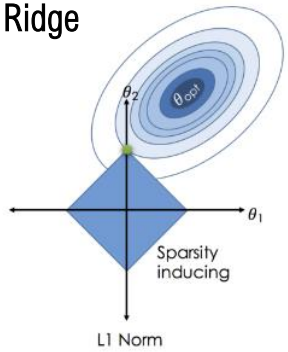
MODELISATION: Choix des modèles de régression

Seattle

Modèle servant de baseline: Linear Regression (Classique)

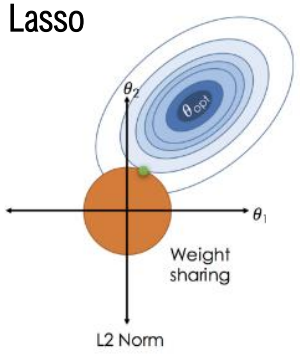
Modèles de régression linéaire suivant un chemin de régularisation

Ridge



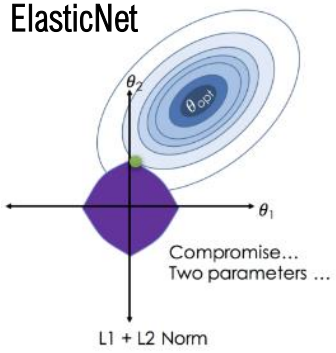
Réduction du poids de l'amplitude des variables

Lasso



Réduction de dimension supervisée

ElasticNet



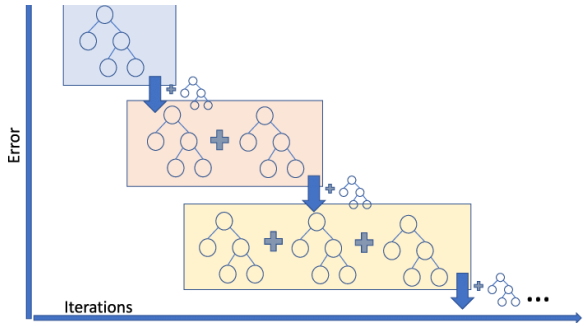
Combinaison des 2 modèles

Méthodes ensemblistes de régression non linéaires basées sur des arbres

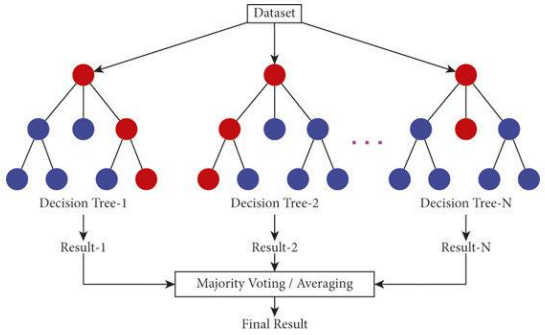
XGBoost

Version améliorée du Gradient Boosting Regressor

Utilisation de paramètres avancés de régularisation (L1 et L2)



Gradient Boosting Regressor



Random Forest Regressor

MODELISATION: Prédiction de la consommation en énergie

Seattle

MODELES LINEAIRES	METHODES ENSEMBLISTES
Ridge (random_state = 42, cv=5) **alpha** : [0.0001, 0.001, 0.01, 0.1, 1] **max_iter** : [10, 100, 200, 400, 600, 800, 1000]	**Random Forest Regressor** (random_state = 42, cv=5) **n_estimators** : [50, 100, 200, 400], **max_depth** : [5, 10, 15, 20] **max_features** : ['sqrt', 'log2'], **criterion**: ['squared_error', 'absolute_error']
Lasso (random_state = 42, cv=5) **alpha** : [0.0001, 0.001, 0.01, 0.1, 1] **max_iter** : [1000, 2000, 5000, 10000]	**Gradient Boosting Regressor** (random_state = 42, cv=5) **n_estimators** : [50, 100, 200, 400, 1000], **max_depth**: [5, 10, 15, 20, 25] **max_features** : ['sqrt', 'log2'], **criterion**: ['squared_error', 'absolute_error']
ElasticNet (random_state = 42, cv=5) **l1-ratio** : [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1] **alpha** : [10 ,1 , 0.1, 0.01, 0.001] **max_iter** : [500, 1000, 2000, 5000]	**XGBRegressor** (cv=5) **n_estimators** : [10,50,100,500,1000], **max_depth** : [2,4,8,16]

	LinearRegression	Ridge	Lasso	ElasticNet	RFR	GBR	XGBR
Grid Training Time	0.014315	0.022668	0.050628	0.129862	1.603764	1.80675	1.904382
Grid Train R^2	0.555011	0.555006	0.554992	0.553906	0.912153	0.998212	0.728402
Grid Train RMSE	39.199428	39.199647	39.200288	39.248065	17.416808	2.484572	30.624462
Grid Train MSE	1536.595151	1536.612295	1536.662603	1540.410609	303.345205	6.1731	937.857653
Grid Train MAE	25.040228	25.038394	25.030529	24.855676	10.705327	1.680279	20.032042
Grid Test R^2	0.41836	0.418983	0.420176	0.430508	0.502572	0.48517	0.496288
Grid Test RMSE	34.793804	34.775164	34.739433	34.428527	32.176591	32.734584	32.379177
Grid Test MSE	1210.608782	1209.312028	1206.828179	1185.3235	1035.332982	1071.552994	1048.411098
Grid Test MAE	24.37326	24.353228	24.342202	23.936694	22.914644	23.875604	23.201573

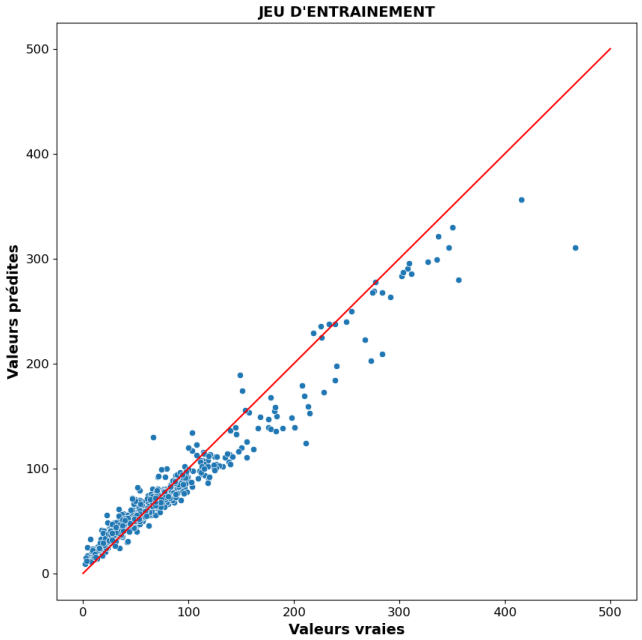
RandomForestRegressor

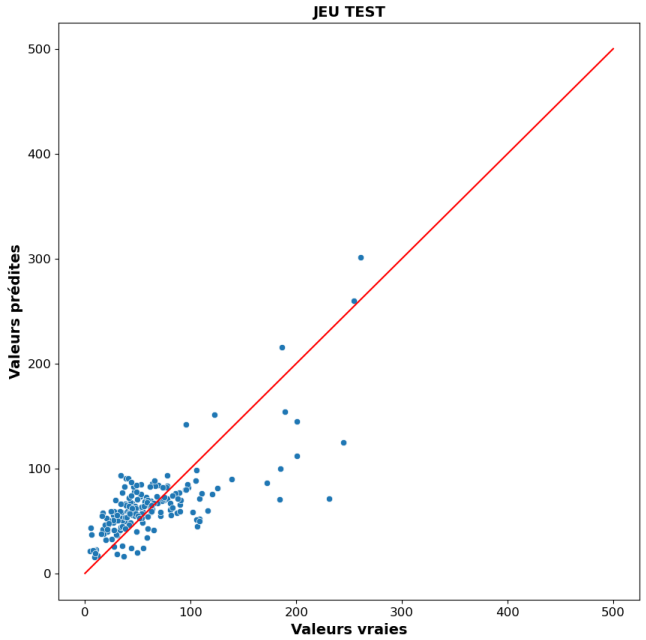
MODELISATION: Prédiction de la consommation en énergie

Seattle

RandomForestRegressor 2	RandomForestRegressor 3
n_estimators : [10, 20, 50, 100, 200, 400], max_depth : [10, 20, 25], max_features : ['sqrt', 'log2'], criterion: ['squared_error', 'absolute_error']	n_estimators : [100, 200, 400, 800], max_depth : [8, 16, 32] max_features : ['sqrt', 'log2'], criterion: ['squared_error', 'absolute_error']

	RFR 1	RFR 2	RFR 3
Grid Training Time	1.692208	15.931489	45.323845
Grid Train R^2	0.912153	0.928917	0.93302
Grid Train RMSE	17.416808	15.667131	15.208157
Grid Train MSE	303.345205	245.458997	231.288026
Grid Train MAE	10.705327	9.855735	9.544225
Grid Test R^2	0.502572	0.533306	0.538615
Grid Test RMSE	32.176591	31.166698	30.988915
Grid Test MSE	1035.332982	971.363043	960.312857
Grid Test MAE	22.914644	22.590148	22.647489

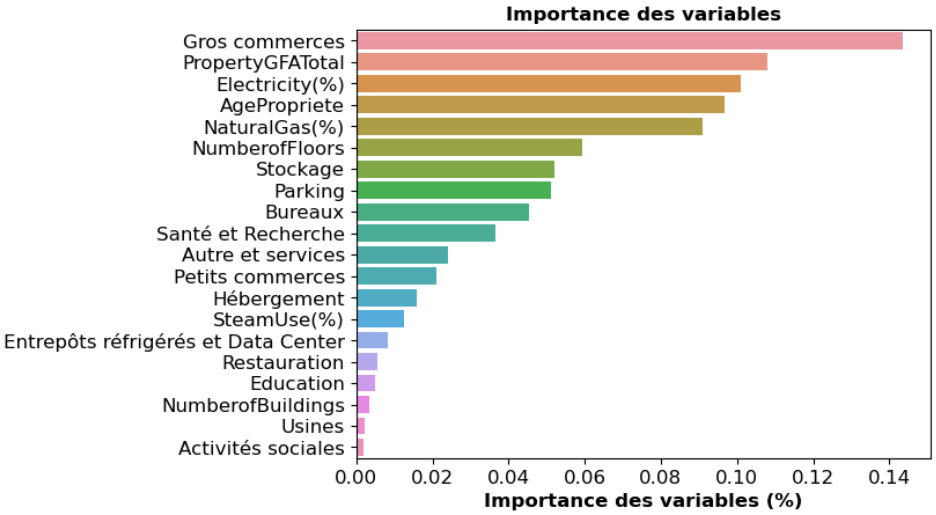




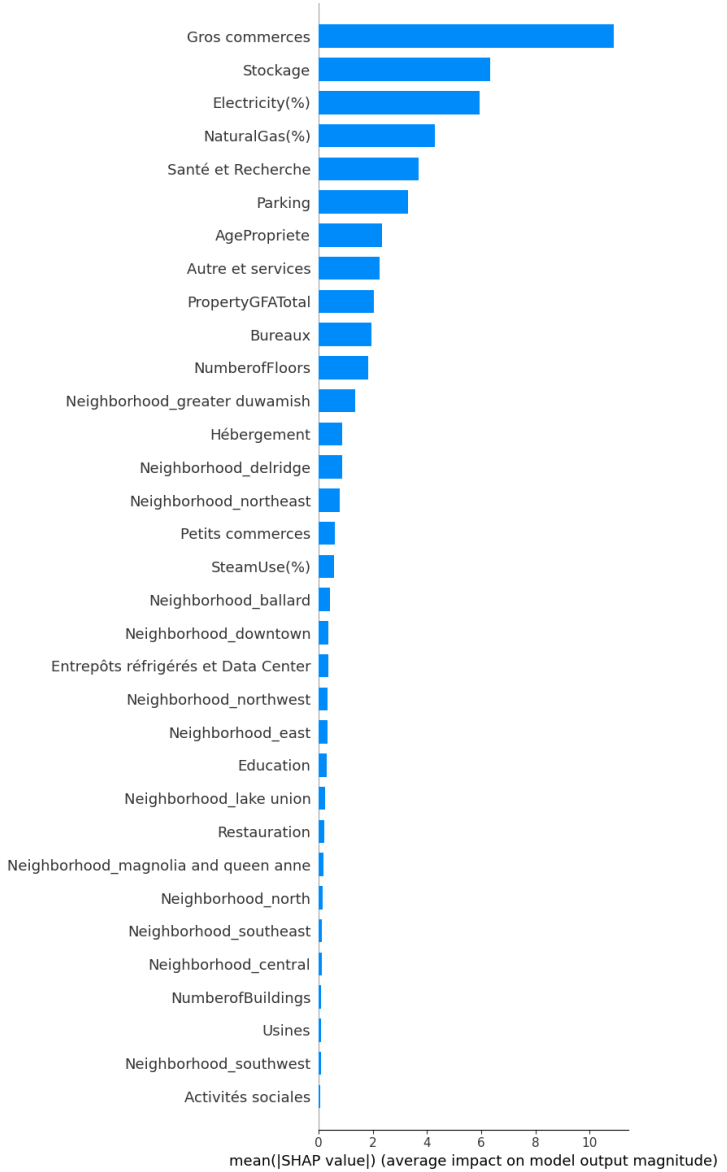
MODELISATION: Prédiction de la consommation en énergie

Seattle

Les 10 variables les plus importantes pour l'entraînement du modèle: features_importances



Importance des variables sur la prédiction du modèle: Les valeurs de Shapley (TreeExplainer)



MODELISATION: Prédiction de la consommation en énergie sur de nouvelles données

Seattle

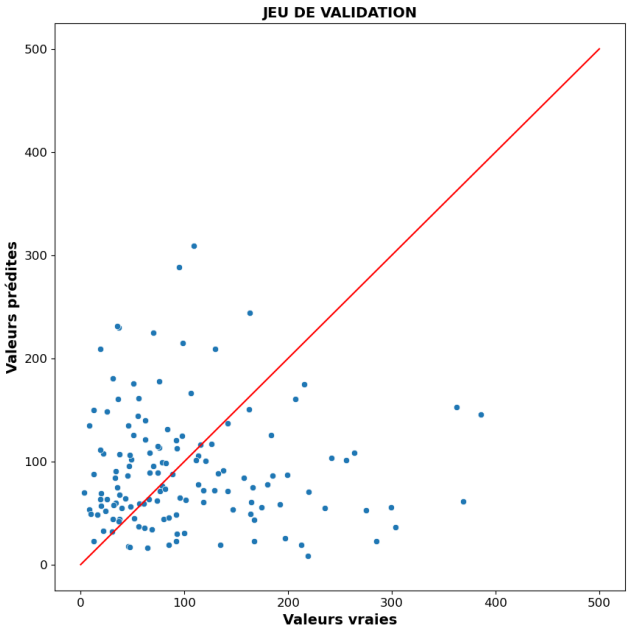
SCORE SUR LE JEU DE VALIDATION

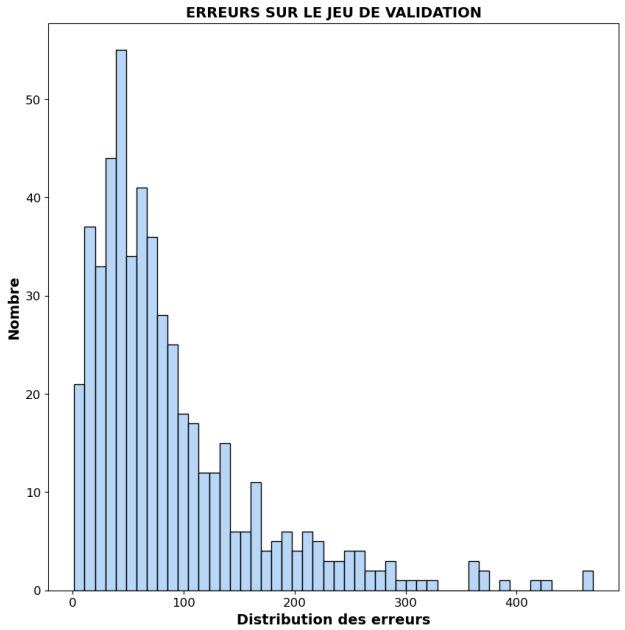
```
R²: 0.8937233636738723
RMSE: 25.3191157572625
MSE: 641.0576227296583
MAE: 17.786946830750235
```

POURCENTAGE D'ERREURS SUR LE JEU DE VALIDATION

```
3 print(1-rfr_val.score(X_val, y_val))
```

```
0.10627663632612772
```





Majorité des erreurs de faibles ampleurs

CONCLUSION POUR LA PREDICTION DE LA CONSOMMATION EN ENERGIE

Modèle utilisable mais qui serait davantage robuste par la mise à disposition de plus de données

MODELISATION: Prédiction des émissions de gaz à effet de serre

Seattle

MODELES LINEAIRES	METHODES ENSEMBLISTES
Ridge (random_state = 42, cv=5) **alpha** : [0.0001, 0.001, 0.01, 0.1, 1] **max_iter** : [10, 100, 200, 400, 600, 800, 1000]	**Random Forest Regressor** (random_state = 42, cv=5) **n_estimators** : [50, 100, 200, 400], **max_depth** : [5, 10, 15, 20] **max_features** : ['sqrt', 'log2'], **criterion**: ['squared_error', 'absolute_error']
Lasso (random_state = 42, cv=5) **alpha** : [0.0001, 0.001, 0.01, 0.1, 1] **max_iter** : [2000, 5000, 10000, 1000000]	**Gradient Boosting Regressor** (random_state = 42, cv=5) **n_estimators** : [50, 100, 200, 400, 1000], **max_depth**: [5, 10, 15, 20, 25] **max_features** : ['sqrt', 'log2'], **criterion**: ['squared_error', 'absolute_error']
ElasticNet (random_state = 42, cv=5) **l1-ratio** : [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1] **alpha** : [100, 10 ,1 , 0.1, 0.01, 0.001] **max_iter** : [1000, 2000, 5000]	**XGBRegressor** (cv=5) **n_estimators** : [10,50,100,500,1000], **max_depth** : [2,4,8,16]

	LinearRegression	Ridge	Lasso	ElasticNet	RFR	GBR	XGBR
Grid Training Time	0.008048	0.02528	0.029827	0.096135	1.274874	5.232662	5.342513
Grid Train R^2	0.637779	0.633705	0.486115	0.636281	0.939131	0.902864	0.813935
Grid Train RMSE	1.024497	1.030242	1.220273	1.026614	0.419974	0.530536	0.73427
Grid Train MSE	1.049593	1.061399	1.489066	1.053936	0.176378	0.281469	0.539152
Grid Train MAE	0.601932	0.593388	0.729805	0.598232	0.22999	0.262768	0.435181
Grid Test R^2	0.540722	0.57977	0.494271	0.546846	0.74221	0.732488	0.713737
Grid Test RMSE	0.849015	0.812122	0.890915	0.843335	0.636078	0.647961	0.670287
Grid Test MSE	0.720827	0.659542	0.79373	0.711215	0.404595	0.419854	0.449284
Grid Test MAE	0.550068	0.53361	0.630064	0.545772	0.42279	0.403539	0.444862

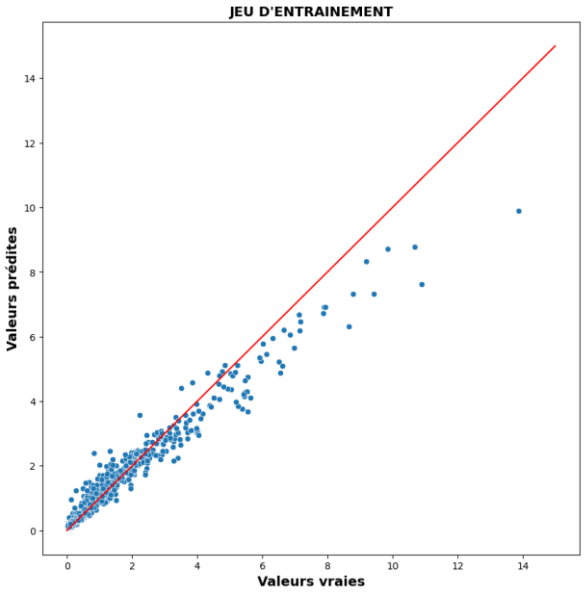
RandomForestRegressor?
GradientBoostingRegressor?

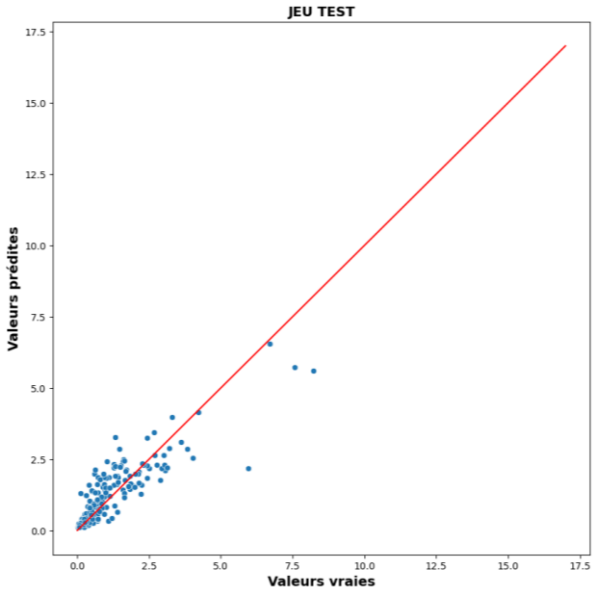
MODELISATION: Prédiction des émissions de gaz à effet de serre

Seattle

RandomForestRegressor 2	GradientBoostingRegressor 2
n_estimators : [50, 100, 200, 400], max_depth : [15, 20, 25], max_features : ['sqrt', 'log2'], criterion: ['squared_error', 'absolute_error']	n_estimators : [25, 50, 100, 200, 400], max_depth : [2, 4, 8, 16] max_features : ['sqrt', 'log2'], criterion: ['squared_error', 'absolute_error']

	RFR 1	RFR 2	GBR 1	GBR 2
Grid Training Time	3.246331	6.456486	10.141948	10.415677
Grid Train R^2	0.939131	0.942969	0.902864	0.883486
Grid Train RMSE	0.419974	0.406518	0.530536	0.58105
Grid Train MSE	0.176378	0.165257	0.281469	0.337619
Grid Train MAE	0.22999	0.21641	0.262768	0.365437
Grid Test R^2	0.74221	0.737025	0.732488	0.713738
Grid Test RMSE	0.636078	0.642444	0.647961	0.670284
Grid Test MSE	0.404595	0.412734	0.419854	0.449281
Grid Test MAE	0.42279	0.422229	0.403539	0.463936

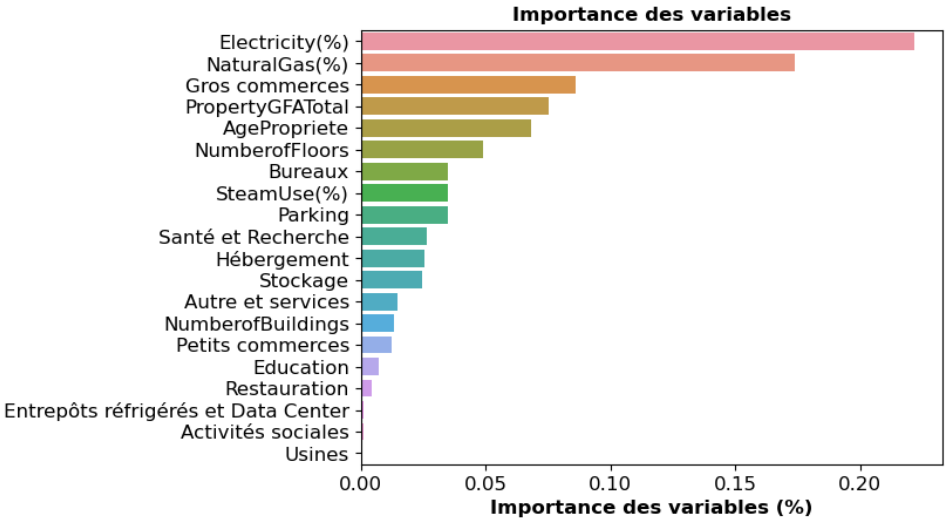




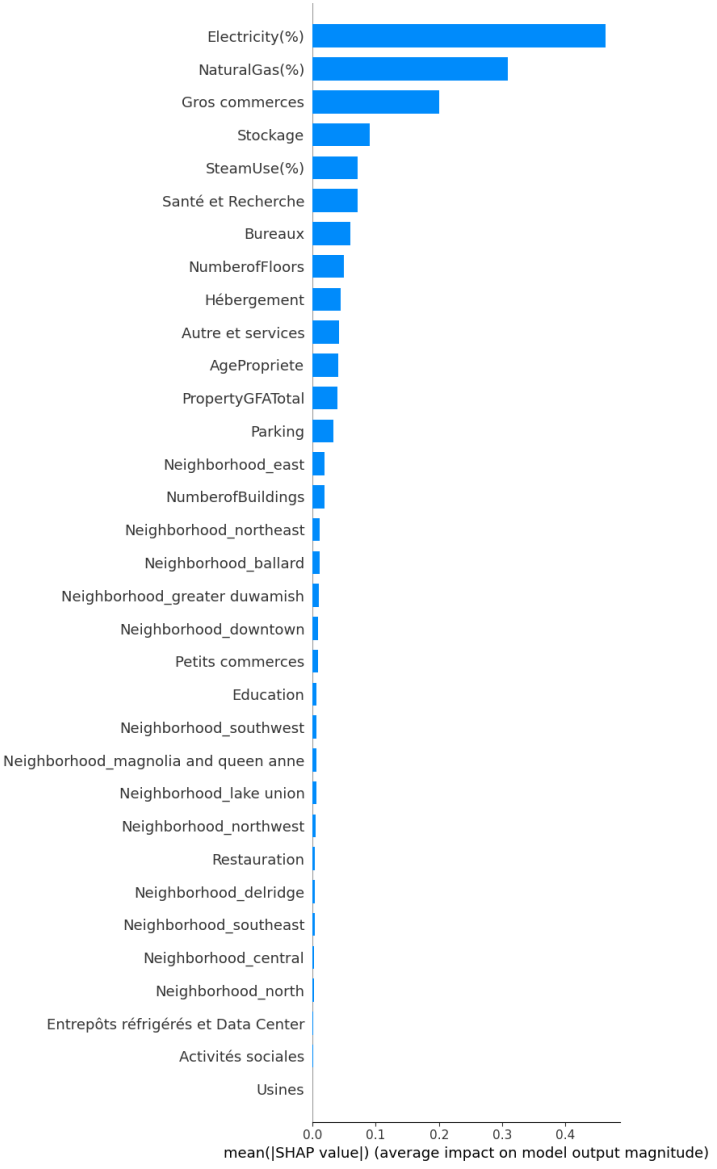
MODELISATION: Prédiction des émissions de gaz à effet de serre

Seattle

Les 10 variables les plus importantes pour l'entraînement du modèle: features_importances



Importance des variables sur la prédiction du modèle: Les valeurs de Shapley (TreeExplainer)



MODELISATION: Prédiction des émissions de gaz à effet de serre sur de nouvelles données

Seattle

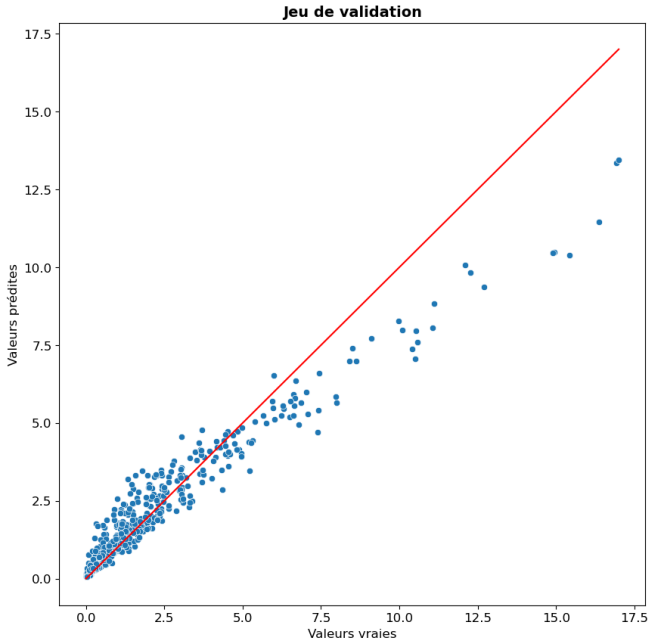
SCORE SUR LE JEU DE VALIDATION

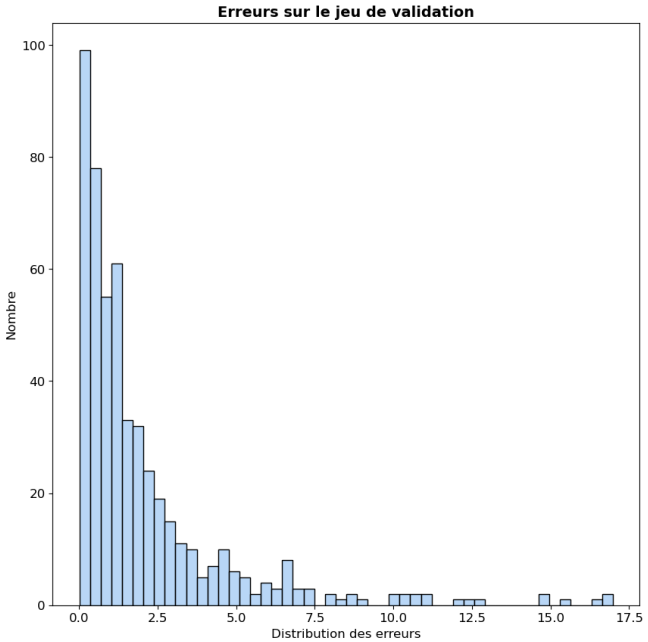
```
R²: 0.9028826989512889
RMSE: 0.8344032771331279
MSE: 0.6962288288905035
MAE: 0.49247897286821696
```

POURCENTAGE D'ERREURS SUR LE JEU DE VALIDATION

```
3 print(1-rfr_val.score(X_val, y_val))
```

```
0.09711730104871108
```





CONCLUSION POUR LA PREDICTION DES EMISSIONS DES GAZ A EFFET DE SERRE:

Modèle davantage robuste que celui obtenu pour la prédiction de la consommation en énergie
Optimal si mise à disposition de plus de données…

MODELISATION: Impact de l'Energy Star Score sur la prédiction des émissions de gaz à effet de serre

Seattle

	RFR sans EnergyStarScore	RFR avec EnergyStarScore
Train R^2	0.939131	0.948906
Train RMSE	0.419974	0.384776
Train MSE	0.176378	0.148053
Train MAE	0.22999	0.196019
Test R^2	0.74221	0.777909
Test RMSE	0.636078	0.590395
Test MSE	0.404595	0.348566
Test MAE	0.42279	0.365365

CONCLUSION DE L'IMPACT DE L'ENERGY STAR SCORE POUR LA PREDICTION DES EMISSIONS DES GAZ A EFFET DE SERRE

Impact non négligeable de l'incorporation de l'Energy Star Score

CONCLUSION GENERALE

Modèles utilisables avec le modèle de prédiction de l'émission de gaz davantage performant

- Malgré le cout, des relevés supplémentaires seraient vraiment souhaitables…
- Intérêt non négligeable de l'incorporation de l'Energy Star Score même si calcul fastidieux…

Seattle

- Débat / Réflexion -

Informations supplémentaires

Analyse des corrélations entre variables > ou = à 0,85

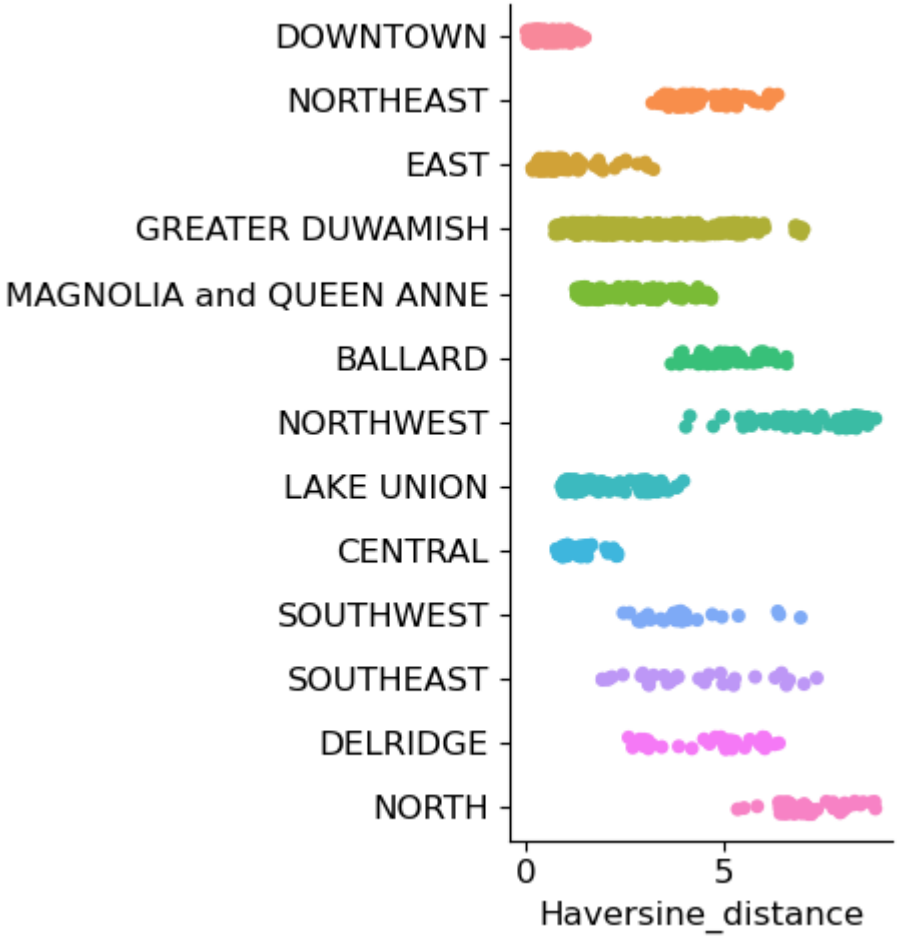
```
strong_pairs = sorted_pairs[(abs(sorted_pairs) >= 0.85) & (abs(sorted_pairs)!=1)]
print(strong_pairs)
```

```
TotalGHGEmissions                SiteEnergyUseWN(kBtu)            0.893592
SiteEnergyUseWN(kBtu)            TotalGHGEmissions                0.893592
SecondLargestPropertyUseTypeGFA  PropertyGFAParking               0.896050
PropertyGFAParking               SecondLargestPropertyUseTypeGFA  0.896050
TotalGHGEmissions                NaturalGas(kBtu)                 0.911821
NaturalGas(kBtu)                 TotalGHGEmissions                0.911821
Electricity(kBtu)                SiteEnergyUseWN(kBtu)            0.922263
SiteEnergyUseWN(kBtu)            Electricity(kBtu)                0.922263
SourceEUIWN(kBtu/sf)             SiteEUIWN(kBtu/sf)               0.945329
SiteEUIWN(kBtu/sf)               SourceEUIWN(kBtu/sf)             0.945329
LargestPropertyUseTypeGFA        PropertyGFATotal                 0.976698
PropertyGFATotal                 LargestPropertyUseTypeGFA        0.976698
                                 PropertyGFABuilding(s)           0.980842
PropertyGFABuilding(s)           PropertyGFATotal                 0.980842
                                 LargestPropertyUseTypeGFA        0.983838
LargestPropertyUseTypeGFA        PropertyGFABuilding(s)           0.983838
dtype: float64
```

Informations supplémentaires

Seattle

Transformation des Latitudes et Longitudes en distance Haversienne à partir des coordonnées de centre de Seattle?



Très mauvaise idée car disparition des quartiers!!!

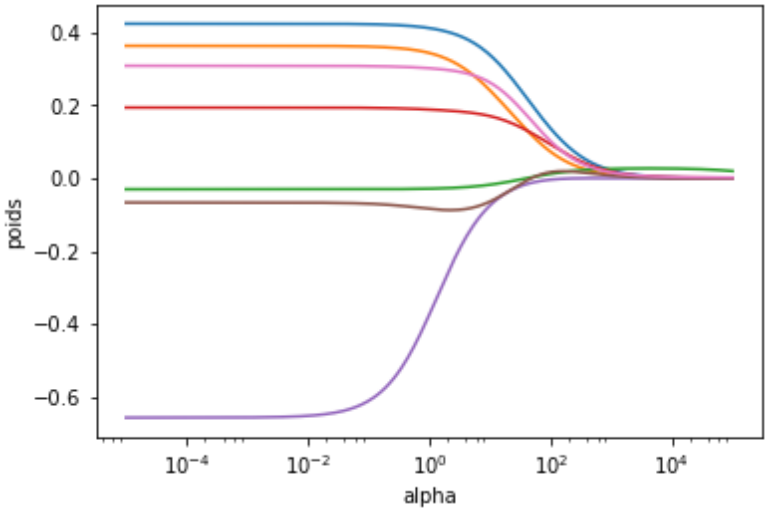
→ Conservation de la variable Neighborhood

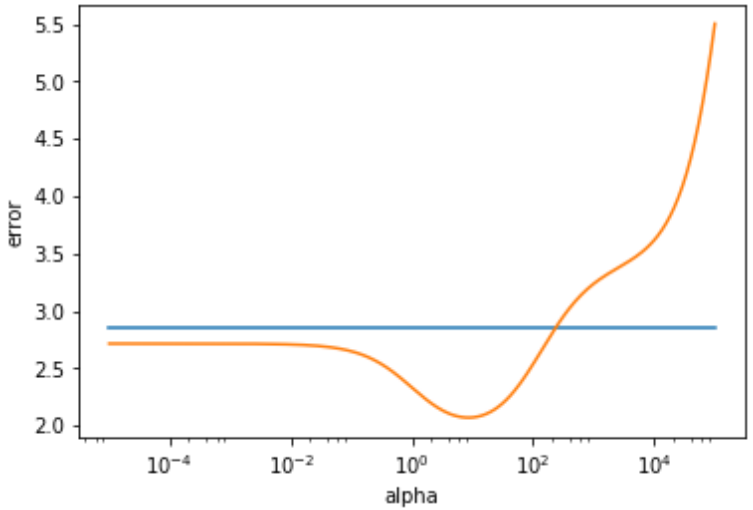
Informations supplémentaires

Seattle

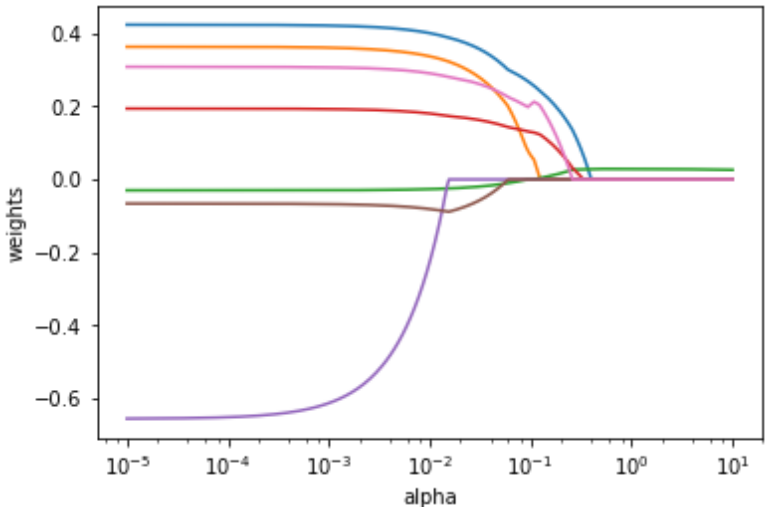
Scaling has an impact on the magnitude of coefficients but it does not have any impact on the model predictions

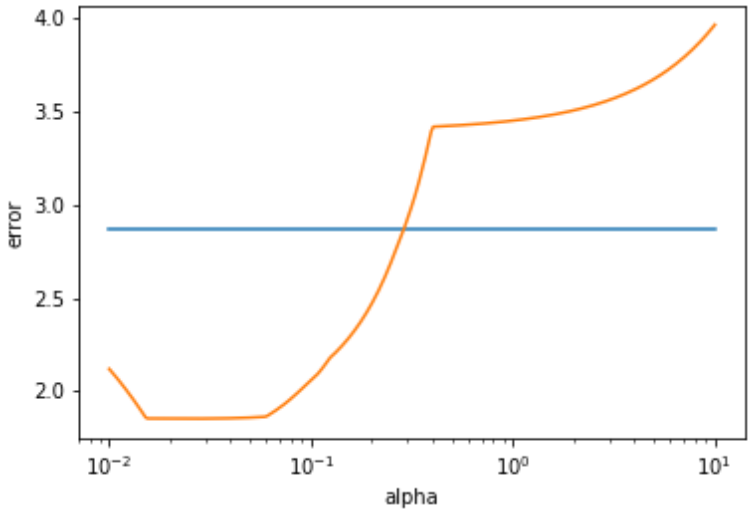
Le chemin de régulation du modèle Ridge





Le chemin de régulation du modèle Lasso





Informations supplémentaires

Seattle

PERFORMANCE POUR LA PREDICTION EN CONSOMMATION D'ENERGIE

Choix du modèle pour la prédiction de la consommation en énergie

	RFR 1	RFR 2	GBR 1	GBR 2
Grid Training Time	1.485762	14.243245	14.495047	18.521609
Grid Train R^2	0.912153	0.928917	0.998212	0.642353
Grid Train RMSE	17.416808	15.667131	2.484572	35.142501
Grid Train MSE	303.345205	245.458997	6.1731	1234.995377
Grid Train MAE	10.705327	9.855735	1.680279	19.12991
Grid Test R^2	0.502572	0.533306	0.48517	0.5181
Grid Test RMSE	32.176591	31.166698	32.734584	31.670374
Grid Test MSE	1035.332982	971.363043	1071.552994	1003.012615
Grid Test MAE	22.914644	22.590148	23.875604	20.521693

Informations supplémentaires

Seattle

PERFORMANCE BASIQUE DES DIFFERENTS MODELES

Comparaison des différents modèles pour la prédiction de la consommation en énergie

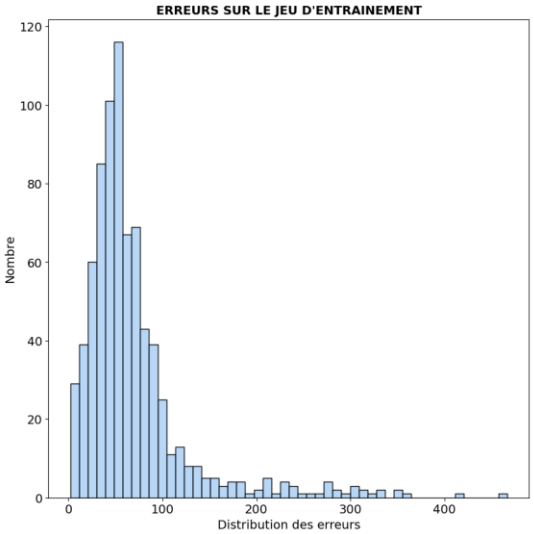
	LinearRegression	Ridge	Lasso	ElasticNet	RFR	GBR	XGBR
Training Time	0.012702	0.025505	0.049137	0.065884	1.12785	1.407364	2.007535
Train R^2	0.566386	0.566168	0.429531	0.491093	0.893437	0.73301	0.993598
Train RMSE	38.695159	38.704901	44.383445	41.920284	19.182597	30.363566	4.701817
Train MSE	1497.315325	1498.069331	1969.890171	1757.310241	367.972036	921.946123	22.107087
Train MAE	23.153982	23.140432	25.753467	24.567122	9.42841	17.61397	2.488531
Test R^2	0.527032	0.527473	0.436409	0.480473	0.50413	0.489696	0.452347
Test RMSE	31.375479	31.360862	34.249702	32.883547	32.126134	32.590371	33.761938
Test MSE	984.420695	983.503687	1173.042069	1081.327643	1032.088512	1062.132253	1139.868445
Test MAE	20.869678	20.836738	22.359822	21.207503	21.284427	21.113808	23.227955

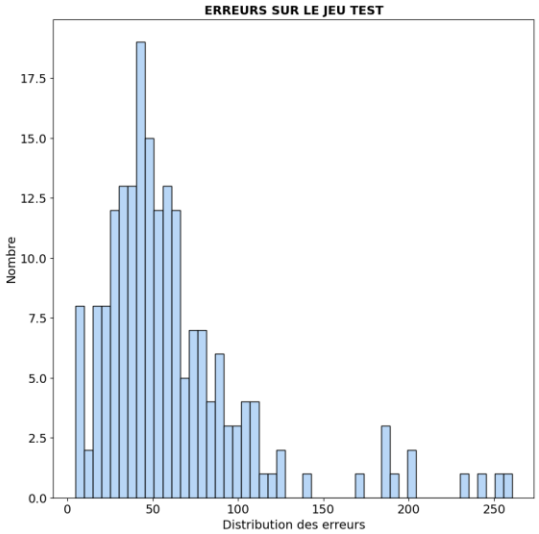
Comparaison des différents modèles pour la prédiction
de l'intensité de l'émission des gaz à effet de serre

	LinearRegression	Ridge	Lasso	ElasticNet	RFR	GBR	XGBR
Training Time	0.01301	0.016727	0.028735	0.033494	0.457	0.582772	1.205339
Train R^2	0.624847	0.625227	0.534675	0.576581	0.911985	0.82556	0.996601
Train RMSE	1.042625	1.042096	1.161187	1.107666	0.505013	0.710963	0.099248
Train MSE	1.087068	1.085965	1.348356	1.226925	0.255038	0.505468	0.00985
Train MAE	0.540101	0.539771	0.599104	0.57187	0.21878	0.368146	0.047067
Test R^2	0.477957	0.489518	0.455774	0.521177	0.741191	0.760482	0.714724
Test RMSE	0.905172	0.895092	0.924202	0.866893	0.637334	0.613121	0.66913
Test MSE	0.819336	0.80119	0.85415	0.751503	0.406195	0.375918	0.447735
Test MAE	0.448059	0.446534	0.480689	0.455982	0.377495	0.371699	0.405424

Informations supplémentaires

DISTRIBUTION DES ERREURS DU MODELE POUR PREDICTION DE LA CONSOMMATION EN ENERGIE





DISTRIBUTION DES ERREURS DU MODELE POUR PREDICTION DES EMISSIONS DES GAZ A EFFET DE SERRE