ANTICIPEZ LA CONSOMMATION EN ENERGIE ET LES EMISSIONS

SOMMAIRE

1. Objectifs
2. Présentation du jeu de données
3. Feature engineering
4. Approche de modélisation
5. Résultats – consommation d’énergie
6. Résultats – émissions CO2
7. Intérêt Energy Star Score

OBJECTIFS

A partir de relevés effectués par les agents de la ville en 2016 :

1. Prédire les émissions de CO2 et la consommation totale d'énergie des bâtiments non destinés à l'habitation.
2. Évaluer l'intérêt de l'énergie star score pour la prévision des émissions de CO2.

PRESENTATION DU JEU DE DONNEES

Relevés de consommation d'énergie des bâtiments de la ville de seattle pour l'année 2016

	Fichier de départ	Fichier nettoyé	Commentaires
Nombre de colonnes	45	37	Suppression des colonnes 'Comments','Outlier', 'YearsENERGYSTARCertified', 'DefaultData', 'ComplianceStatus' qui sont insuffisament renseignées et ne correspondent pas à notre objectif Suppression des colonnes avec des informations redondantes : - Consommation d'énergie en différentes unités - consommation par sources d'énergie en différentes unités
Nombre d'entrées	3376	1664	Suppression des bâtiments résidentiels Suppression des valeurs aberrantes

Variables retenues pour les prédictions :

- SiteEnergyUseWN(kBtu)
- TotalGHGEmissions

VARIABLES RETENUES

Variables retenues pour la modélisation :

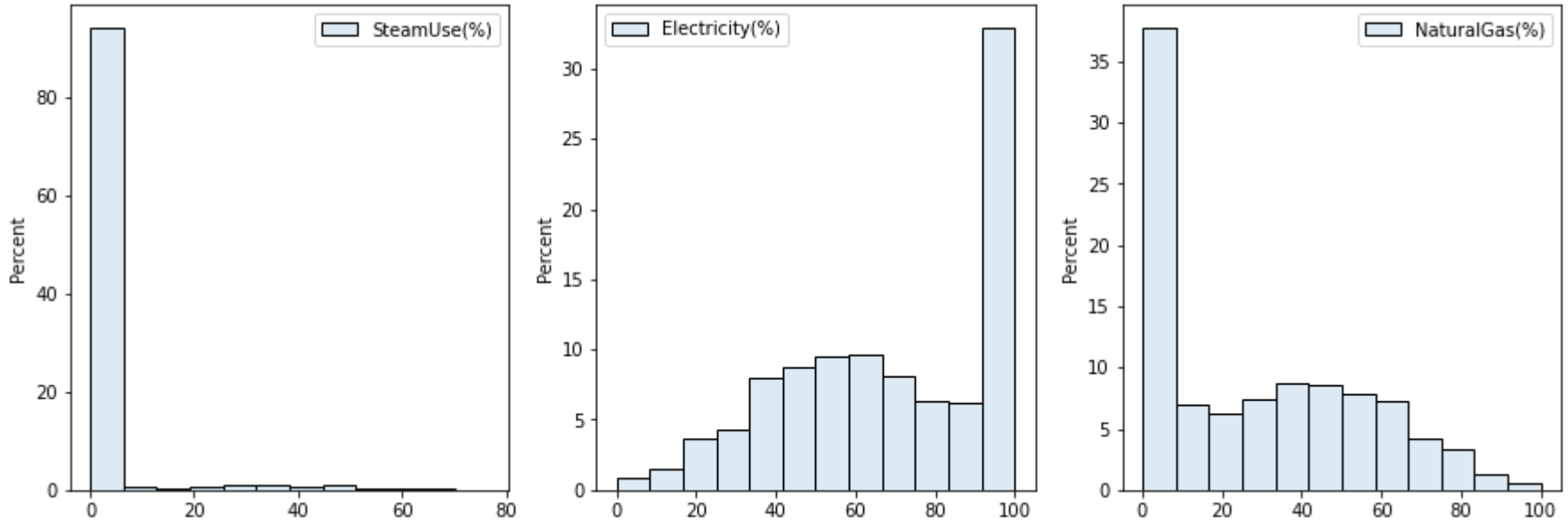
Variables catégorielles	**Variables numériques**
Neighborhood	NumberofBuildings
LargestPropertyUseType	NumberofFloors
	PropertyGFATotal
	PropertyGFABuilding(s)
	PropertyGFAParking
	LargestPropertyUseTypeGFA
	SteamUse(%)
	Electricity(%)
	NaturalGas(%)
	Age

FEATURE ENGINEERING

Réduction du nombre de catégories de 57 à 22 pour les types d'usages des immeubles (property use type)

Calcul de l'ancienneté des bâtiments à partir de l'année de construction

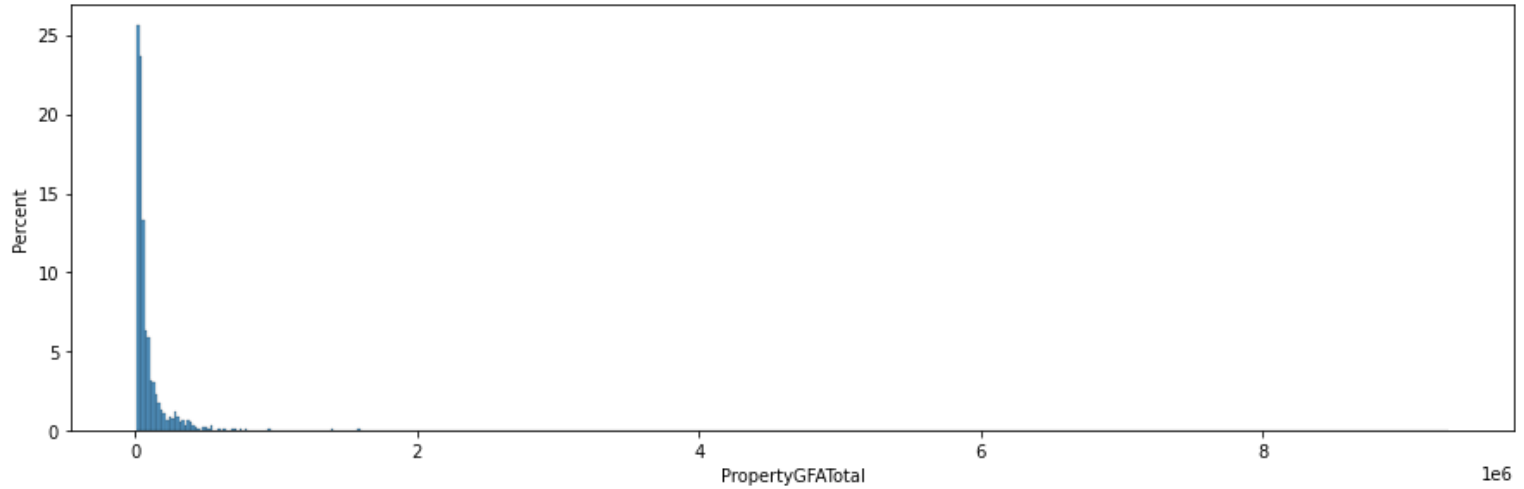
Calcul des proportions des sources d'énergie utilisées

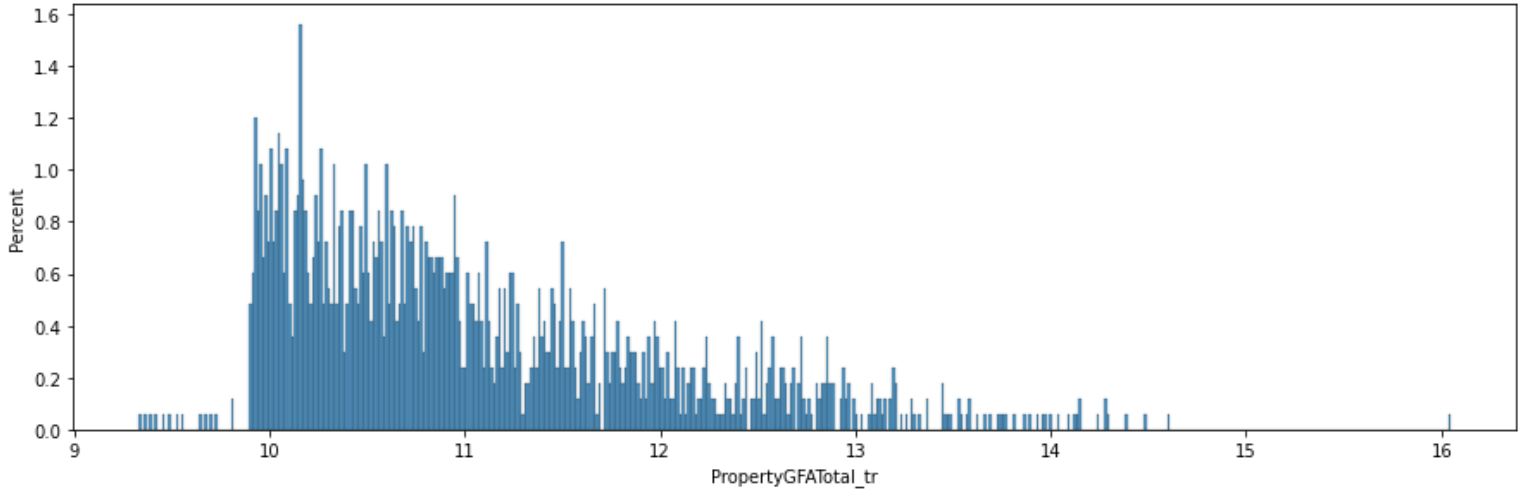


FEATURE ENGINEERING

Passage au log(x+1) des variables indiquant les surfaces (GFA) qui ont des distributions étalées à droite :

- PropertyGFATotal
- PropertyGFABuilding(s)
- PropertyGFAParking
- LargestPropertyUseTypeGFA





APPROCHE DE MODELISATION

Modèles testés
Régression Linéaire Simple
Ridge
Lasso
Kernel Regression
Arbre de decision
Gradient boosting Regression

- Séparation du jeu de données en un jeu d'entraînement (75%) et un jeu de test (25%)
- Standardisation et creation de dummy variables pour les données catégorielles
- Transformation log(x+1) pour les cibles

APPROCHE DE MODELISATION

Recherche des paramètres par validation croisée en 5 plis sur le jeu d'entraînement

Hyper-paramètres par famille de modèles:

Ridge Regression

```
n_alphas = 300
alphas = np.logspace(-5, 5, n_alphas)

param_grid = {
    'regressor__alpha': alphas
}
```

Lasso Regression

```
n_alphas = 300
alphas = np.logspace(-5, 0, n_alphas)

param_grid = {
    'regressor__alpha': alphas,
    'regressor__max_iter' :[10000]
}
```

Kernel Regression

```
alphas = np.logspace(-3, 0, 10)
gammas=np.logspace(-6, -3, 10)

param_grid = {
    'regressor__alpha': alphas,
    'regressor__degree':[2,3],
    'regressor__gamma' :gammas,
    'regressor__kernel':['rbf','polynomial']
}
```

Arbre de Décision

```
param_grid = {
    "max_depth": [4,5,6,7],
    "min_samples_split":[2,5,10,15,20],
    "min_samples_leaf":[5,10,15,20],
    }
```

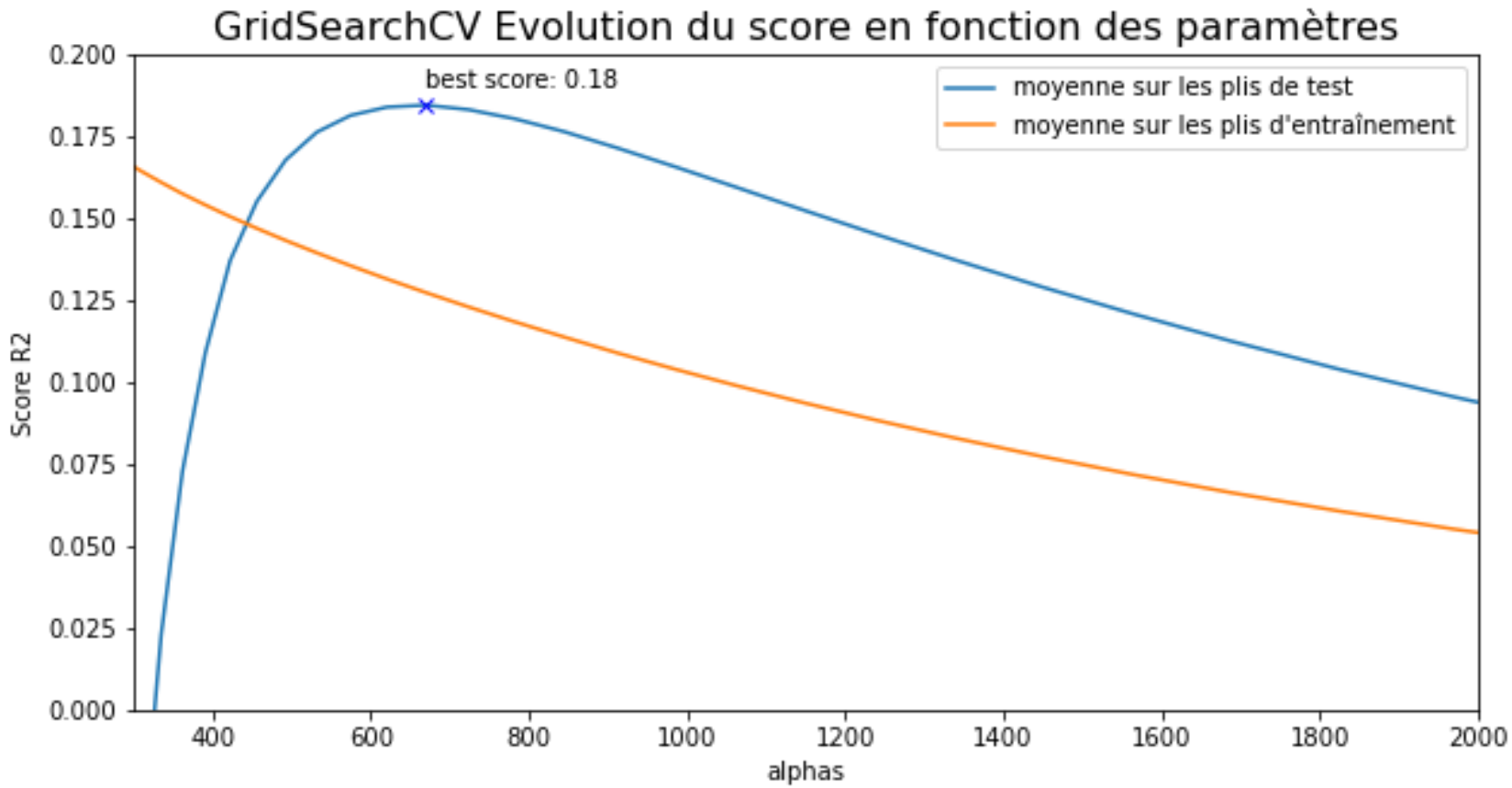
Gradient Boosting Regression

```
learning_rates = np.logspace(-3, 0, 10)

#paramètres pour la recherche
param_grid = {
    "regressor__n_estimators": [100,500],
    "regressor__max_depth": [3,4,5],
    "regressor__learning_rate": learning_rates,
}
```

APPROCHE DE MODELISATION

- Recherche des paramètres par validation croisée en 5 plis sur le jeu d'entraînement
- Nous avons testé trois méthodes de scoring:
  - R2
  - Mean squared error
  - Mean Abolute error



RESULTATS - CONSOMMATION ENERGIE

Le critères de selection des meilleurs paramètres qui a été choisi lors de la cross validation est le score R2

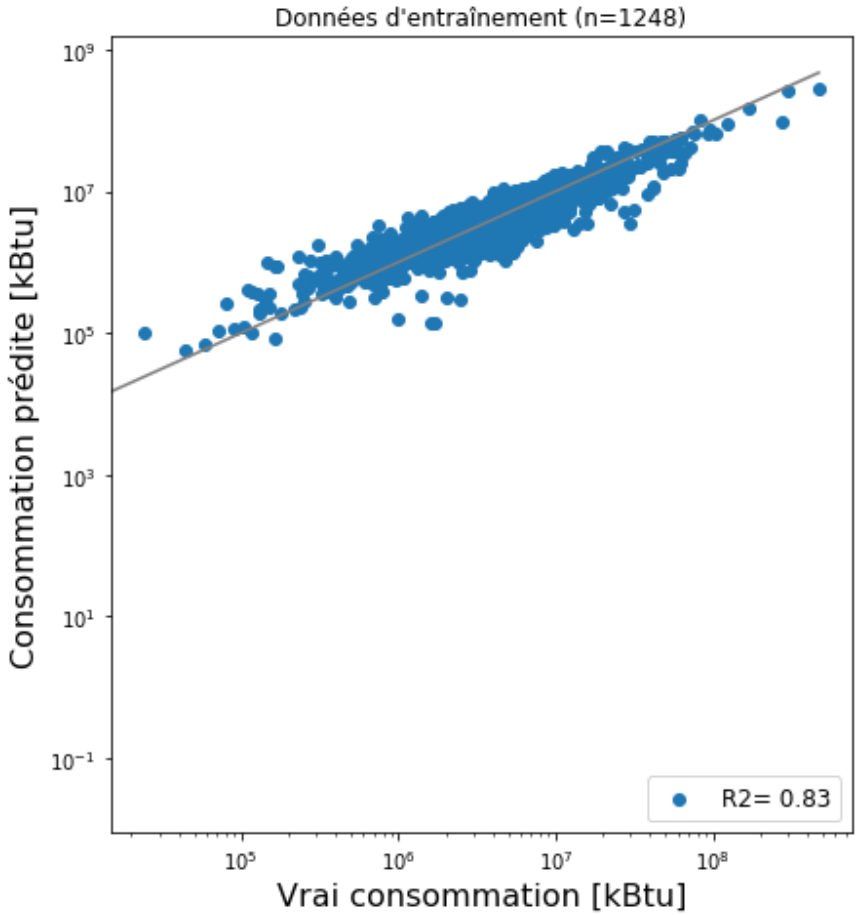
	Résultats Cross validation			
Modèle	**Mean R2**	**MSE*10^12**	**MAE*10^3**	**Mean fit Time**
Régression linéaire simple	-2.68			
Régression Ridge	0.185	405	4894	2.594 ms
Régression Lasso	0.395	323	4226	18.753 ms
Kernel Regression	0.493	276	3932	47.081 ms
Arbre de decision	0.429	298	4762	2.995 ms
Gradient boosting regression	0.566	263	3723	154.236 ms

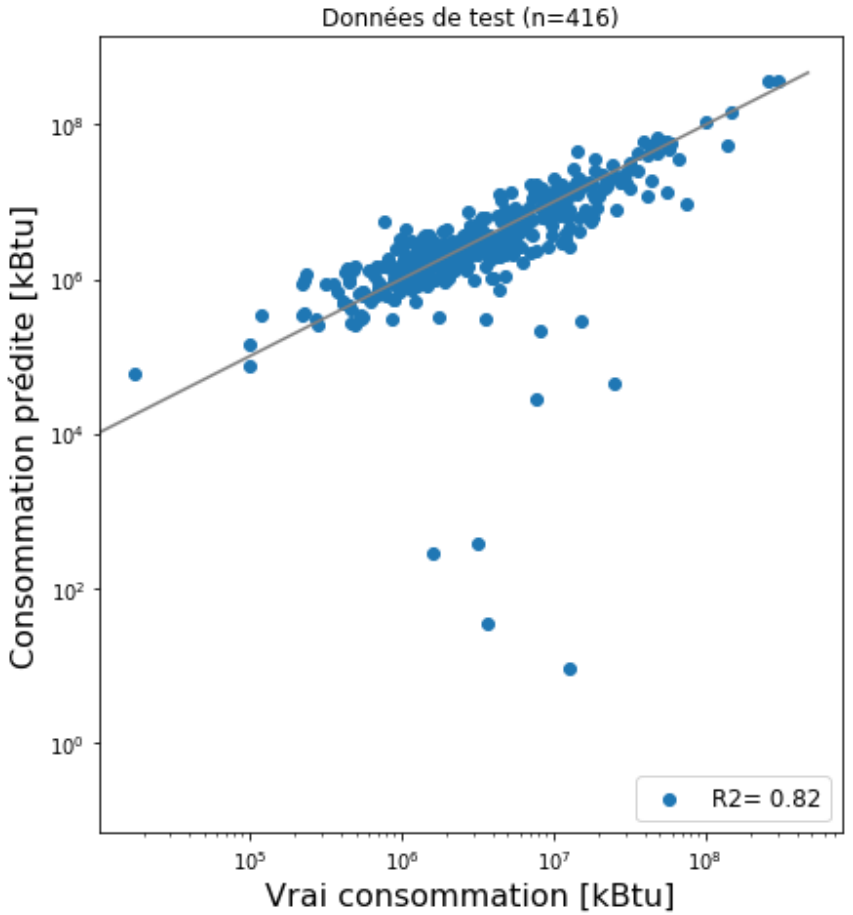
RESULTATS CONSOMMATION ENERGIE

Evaluation des modèles – performances sur de nouvelles données

	Test set		
Modèle	**RMSE/Mean**	**MAE/ Mean**	**R2**
Régression linéaire simple	1.45	0.46	0.68
Régression Ridge	2.41	0.64	0.13
Régression Lasso	2.21	0.56	0.27
Kernel Regression	1.41	0.44	0.70
Arbre de decision	1.88	0.62	0.47
Gradient boosting regression	1.10	0.38	0.83

RESULTATS GRADIENT BOOSTING REGRESSOR



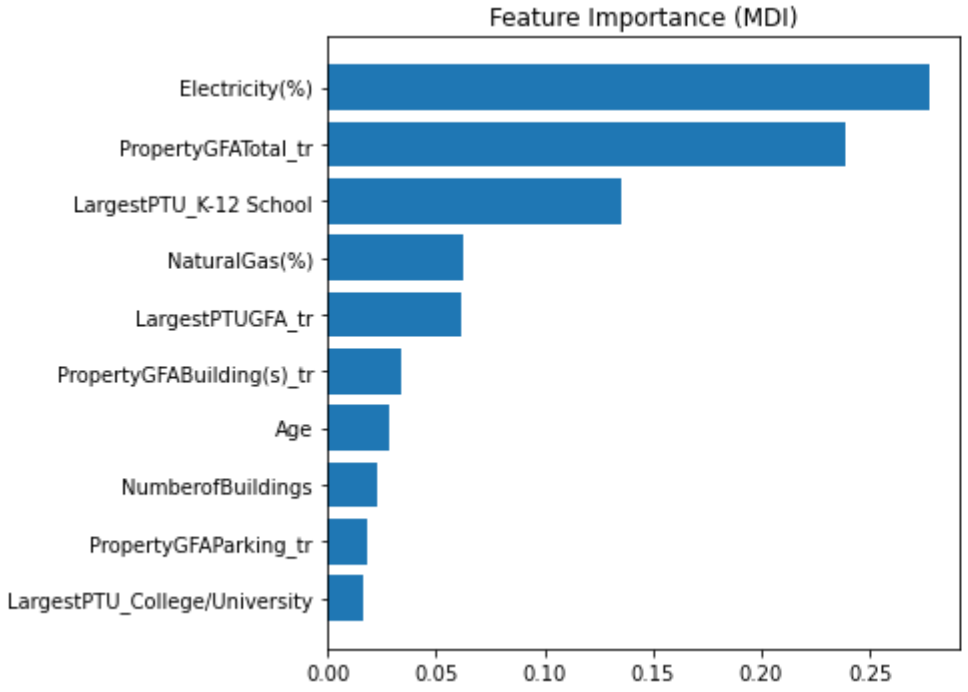


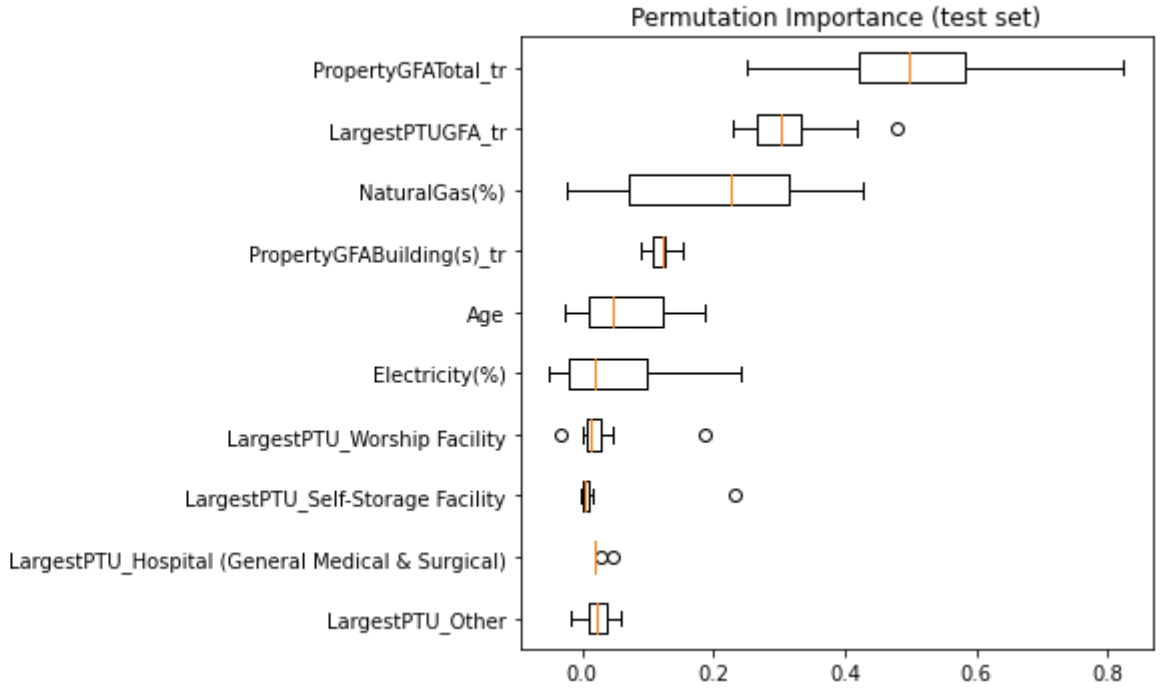
RMSE entrainement/consommation moyenne: 1.16
MAE entrainement/consommation moyenne:: 0.31

RMSE test/consommation moyenne:: 1.10
MAE test/consommation moyenne:: 0.38

IMPORTANCE DES VARIABLES

Analyse des 10 variables qui influent le plus sur la prédiction de consommation d'énergie.





RESULTATS EMISSIONS CO2

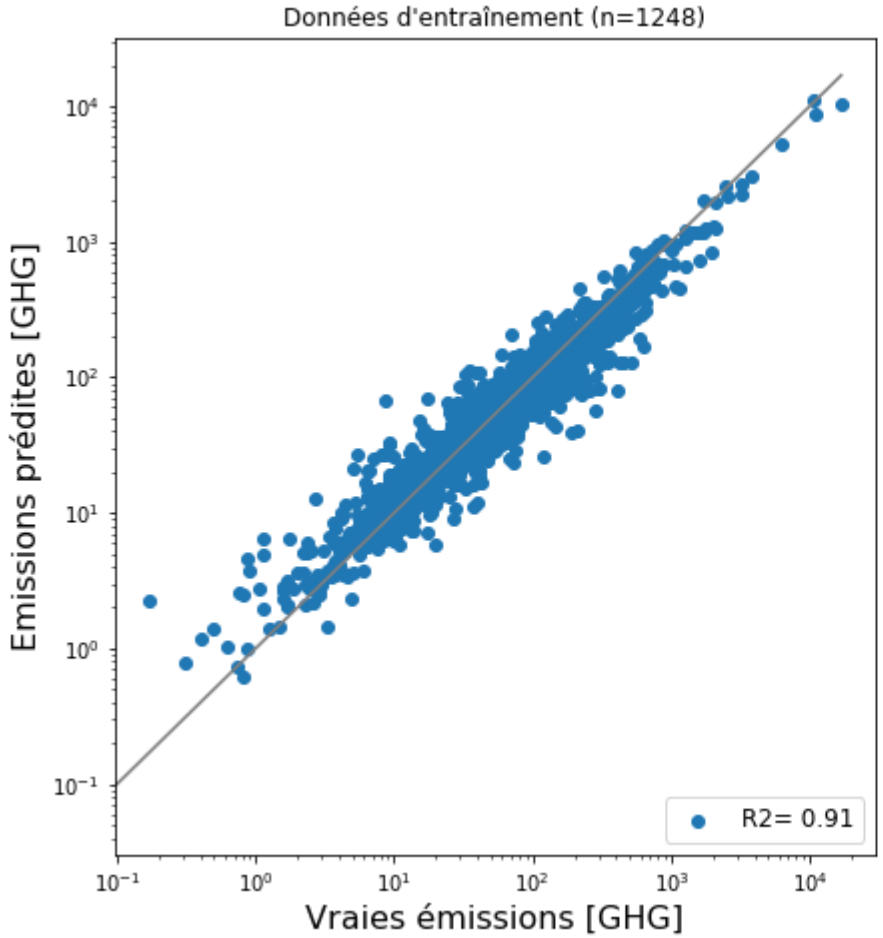
	Résultats Cross validation				
Modèle	**Mean R2**	**std_R2**	**MSE*10^3**	**MAE**	**Mean fit Time**
Régression linéaire simple	-16947				
Régression Ridge	0.48		319	98.1	6.248 ms
Régression Lasso	0.58	0.23	319	87.5	24.994 ms
Kernel Regression	0.58	0.25	279	88.9	68.734 ms
Arbre de decision	0.26	0.11	403	141	6.254 ms
Gradient boosting regression	0.50	0.23	360	93	206.261 ms

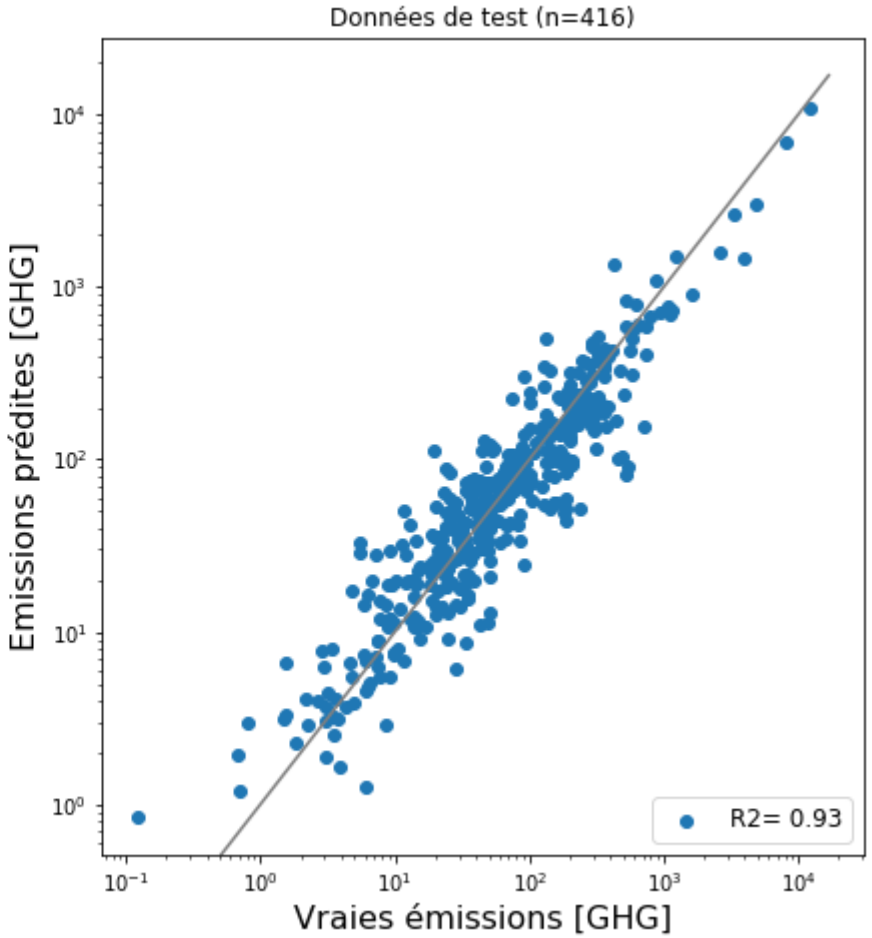
RESULTATS EMISSIONS CO2

Evaluation des modèles – performances sur de nouvelles données

	Test set		
Modèle	**RMSE/Mean**	**MAE/ Mean**	**R2**
Régression linéaire simple	1.57	0.37	0.84
Régression Ridge	2.58	0.55	0.56
Régression Lasso	2.16	0.48	0.69
Kernel Regression	1.83	0.44	0.78
Arbre de decision	3.36	0.90	0.25
Gradient boosting regression	1.02	0.33	0.93

RESULTATS GRADIENT BOOSTING REGRESSOR



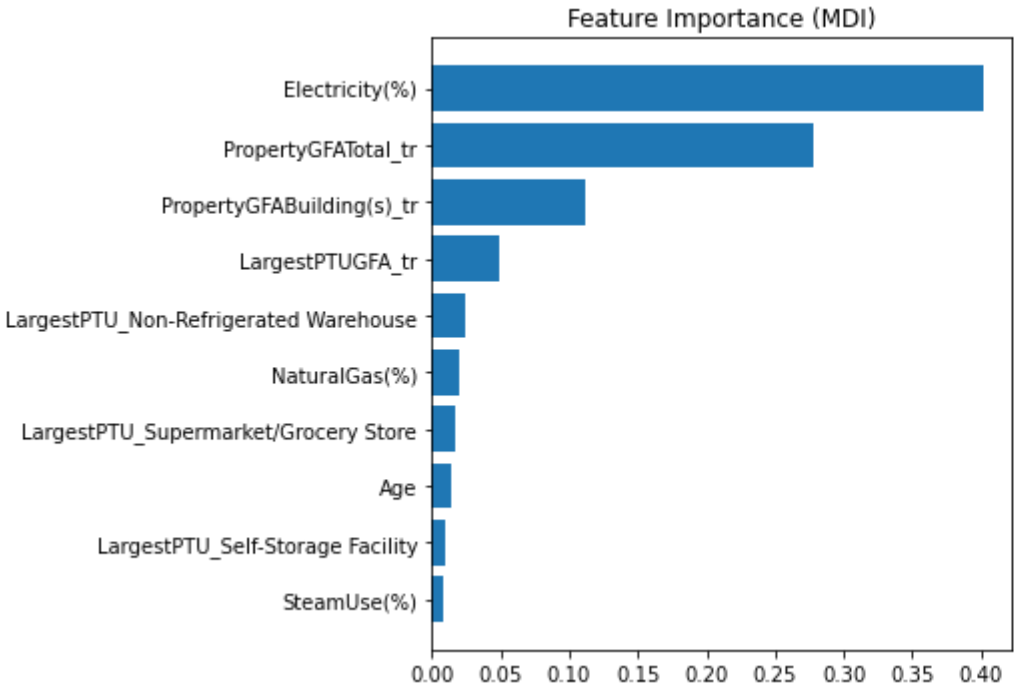


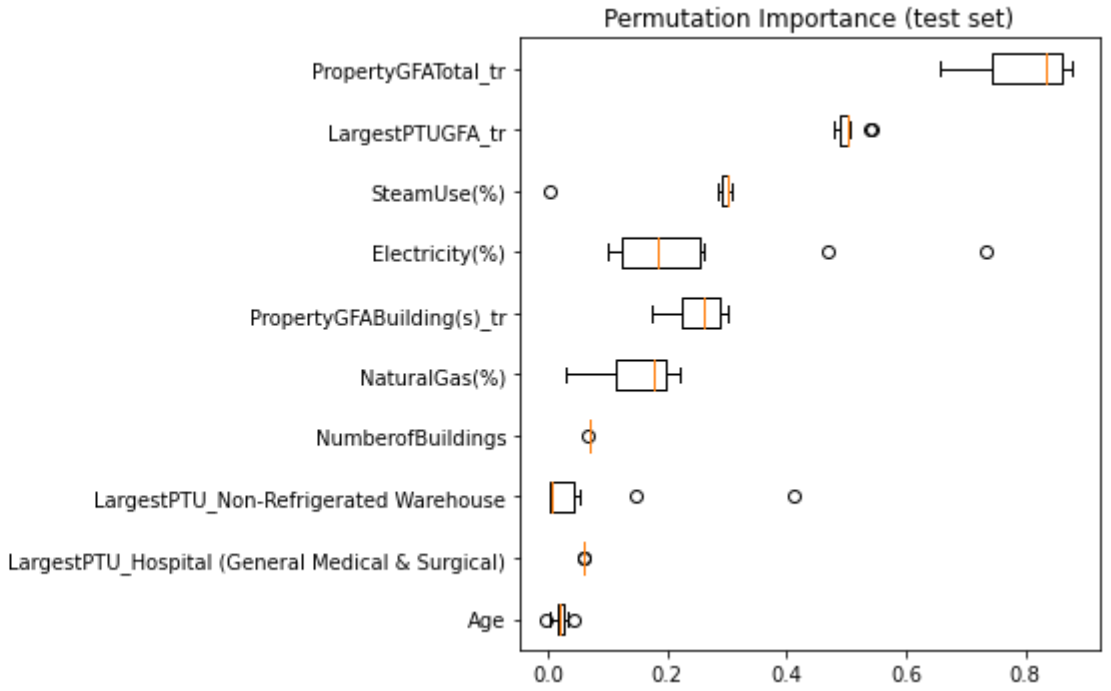
RMSE entrainement/émission moyenne: 1.25
MAE entrainement/émission moyenne:: 0.27

RMSE test/émission moyenne:: 1.02
MAE test/émission moyenne:: 0.33

IMPORTANCE DES VARIABLES

Analyse des 10 variables qui influent le plus sur la prédiction démissions de CO2





ENERGY STAR SCORE

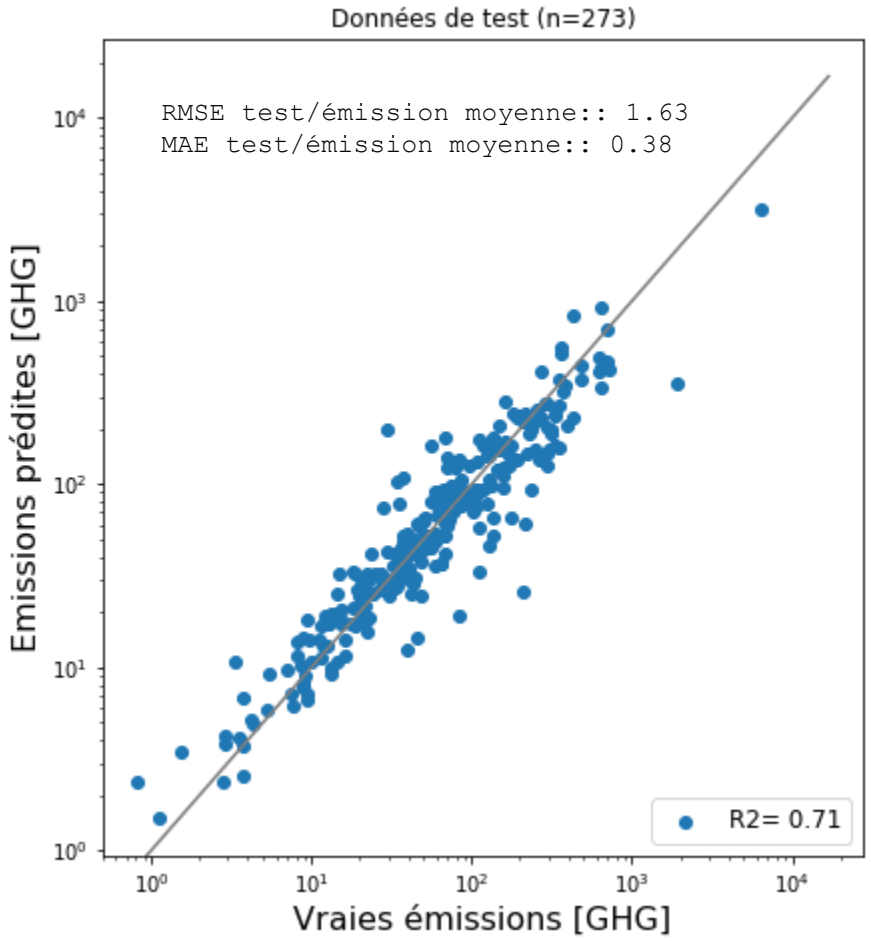
Evaluation de l'intérêt de la variable Energy Star Score pour prédire les émissions de CO2 sur les données pour lesquelles l'indicateur est non nul soit un échantillon de 1092 bâtiments

Résultat de la validation croisée:

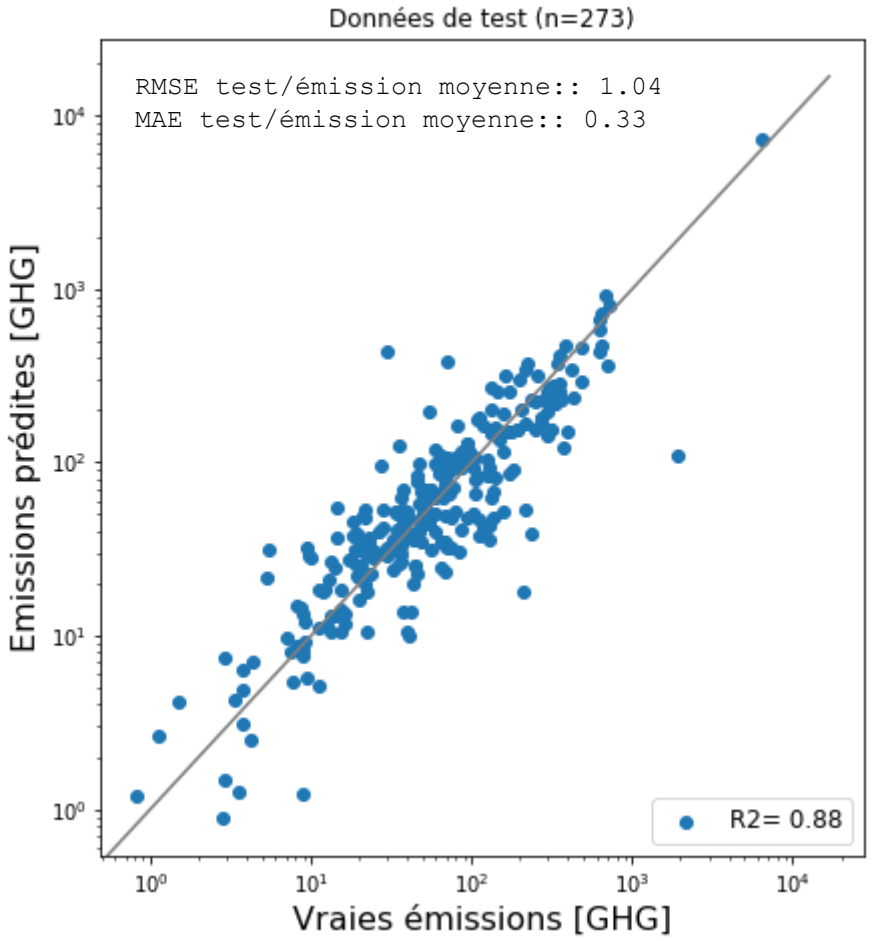
Gradient Boosting Regressor	**Sans Energy Star Score**	**Avec Energy Star Score**
Mean R2	0.60	0.62
Std_R2	0.24	0.17

ENERGY STAR SCORE

Avec



Sans



CONCLUSION

- Nous recommandons le Gradient Boosting Regressor aussi bien pour les predictions de consommations d'énergie que les émissions de CO2:
  - C'est l'algorithme qui se généralise le mieux face à de nouvelles données
  - Le taille du jeu de données permet de palier la contrainte de temps de calcul.

- Nous recommandons au vu des contrainte de calcul de l'indicateur de ne pas intégrer l'Energie Star Score au modèle :
  - L'Energie Star Score n'apporte pas de réelle valeur ajoutée dans les résultats.