Anticipez les besoins en consommation de bâtiments

Soutenance du 31 mai 2023

Yann Pham-Van

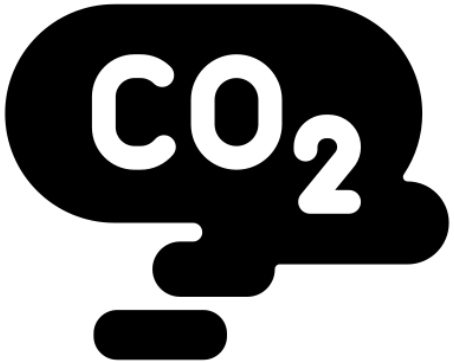
Anticipez les besoins en consommation de bâtiments

- Problématique
- Jeu de données
- Feature engineering
- Modélisation
- Résultats

Anticipez les besoins en consommation de bâtiments

- **Problématique**
- Jeu de données
- Feature engineering
- Modélisation
- Résultats

Problématique



Anticipez les besoins en consommation de bâtiments

- Problématique
- **Jeu de données**
- Feature engineering
- Modélisation
- Résultats

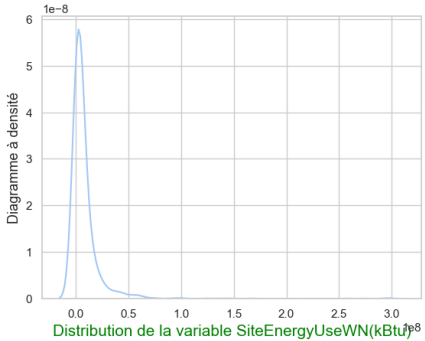
Jeu de données > inspection

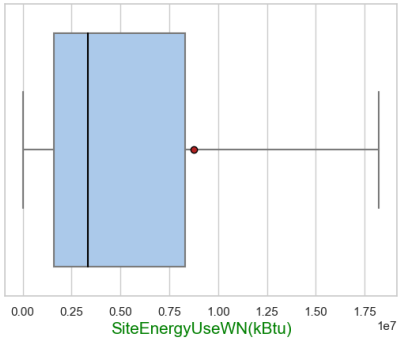
Opération	Nombre d'observations	Nombre de variables
Jeu de données initial	3376	46
Suppression habitat résidentiel	1668	46
Suppression des observations aberrantes, outliers, variables inutiles	1210	27
Export pour modélisations	1210	9

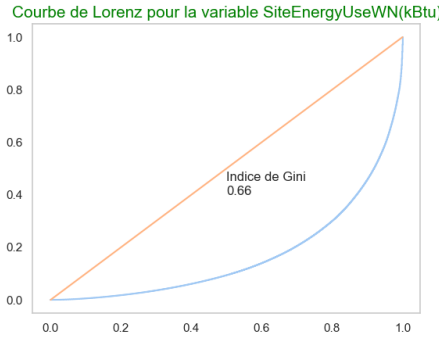
Choix des targets :

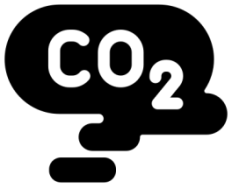
- SiteEnergyUseWN(kBtu)
- TotalGHGEmissions

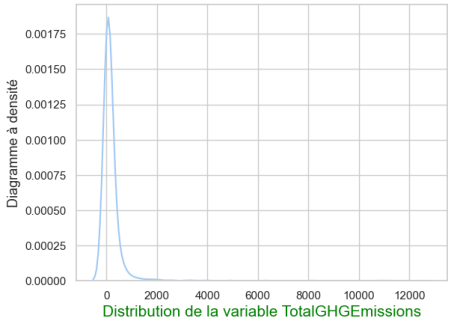
Jeu de données > targets

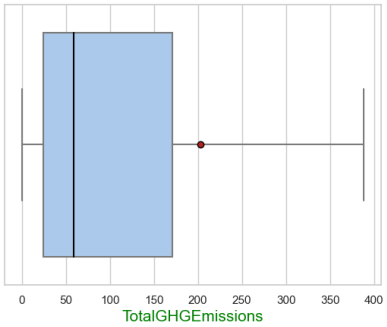


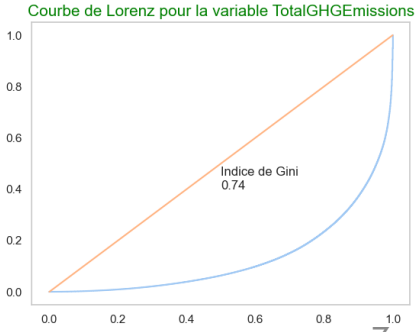












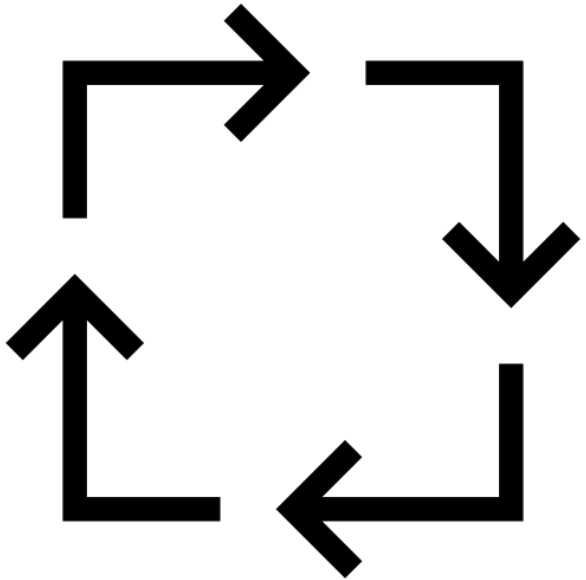
Anticipez les besoins en consommation de bâtiments

- Problématique
- Jeu de données
- **Feature engineering**
- Modélisation
- Résultats

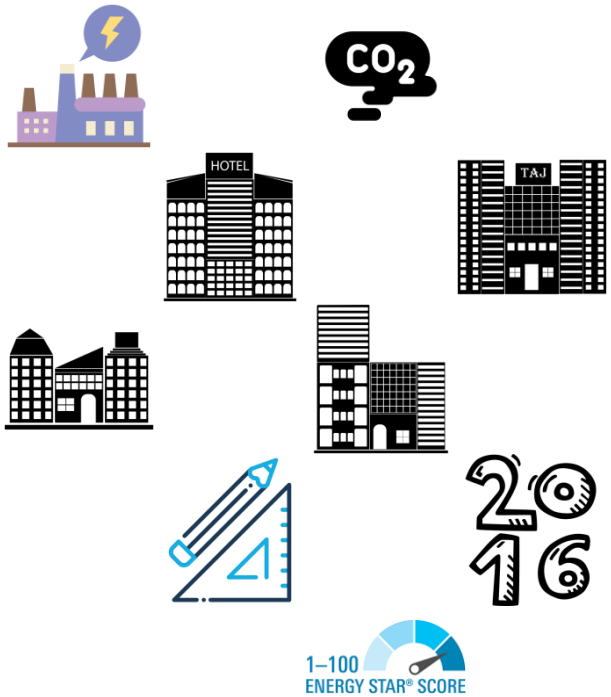
Feature engineering

Démarche itérative d'amélioration des modèles

- Recherche de variables
- Analyse des corrélations
- Transformation des variables
- Pré-processing



Feature engineering > sélection des variables



- Consommation d'énergies
- Emission de CO_2
- Type de bâtiment
- Usage principal
- Nombre de bâtiments
- Nombre d'étages
- Surface totale
- Date de construction
- Option : ENERGY STAR Score

Feature engineering > corrélations

Feature engineering > transformations

Transformation testée	Validée
Remplacement manquants sur type et surface des usages non principaux	
Date de construction → Age	
Nombre de bâtiments → variable binaire	
Passage au log	
Mise à l'échelle des variables numériques	
Encodage des variables qualitatives	

Anticipez les besoins en consommation de bâtiments

- Problématique
- Jeu de données
- Feature engineering
- **Modélisation**
- Résultats

Modélisation > approche

- Pipeline
  - Pré-processing
  - Modélisation
- Baseline, approche naïve par la médiane
- Test de différentes classes de modèles
- Evaluation automatisée
- Optimisation par validation croisée

Modélisation > modèles testés

Classe	Méthode	Modèle
	Plus proches voisins	Kneighbors
Linéaire	Régression linéaire	Ridge
Non linéaire	Support Vector Machine à noyau	SVR
	Régression ridge à noyau	kRR
Ensembliste	Parallèle	Random Forest
	Séquentielle et boosting	Adaboost
	Boosting et descente de gradient	Gradient Boosting

Anticipez les besoins en consommation de bâtiments

- Problématique
- Jeu de données
- Feature engineering
- Modélisation
- **Résultats**

Résultats > évaluation initiale

Consommation d'énergies									
Score	**Modèle**								
	Baseline	**Naïf**	**KNN**	**Ridge**	**SVM**	**kRR**	**Random Forest**	**AdaBoost**	**Gradient Boosting**
Durée			2.369s	2.243s	0.067s	0.061s	0.405s	0.177s	0.190s
R^2	0.62	-0.10	0.33	0.62	-0.10	0.62	0.80	0.77	0.89
RMSE	1.31e+07	2.23e+07	1.74e+07	1.31e+07	2.23e+07	1.32e+07	9.52e+06	1.02e+07	7.14e+06
MAE	5.93e+06	8.23e+06	5.44e+06	5.93e+06	8.23e+06	5.98e+06	4.42e+06	7.39e+06	3.69e+06
R^2 / durée			0.1	0.3	-1.4	10.1	2.0	4.4	4.7

Émission de CO_2									
Durée			2.325s	2.221s	0.069s	0.065s	0.371s	0.169s	0.202s
R^2	0.58	-0.07	0.19	0.58	-0.03	0.59	0.57	0.53	0.78
RMSE	4.10e+02	6.54e+02	5.72e+02	4.09e+02	6.42e+02	4.04e+02	4.17e+02	4.34e+02	3.00e+02
MAE	1.50e+02	1.96e+02	1.52e+02	1.50e+02	1.85e+02	1.51e+02	1.50e+02	3.37e+02	1.31e+02
R^2 / durée			0.1	0.3	-0.4	9.2	1.5	3.1	3.8

Résultats > optimisation par validation croisée

Consommation d'énergies							
Score	**Modèle**						
	KNeighbors	**Ridge**	**SVM**	**kRR**	**Random Forest**	**AdaBoost**	**Gradient Boosting**
R^2 train CV	0.48 (+/-0.08)	0.53 (+/-0.07)	0.67 (+/-0.06)	0.53 (+/-0.07)	0.93 (+/-0.01)	0.85 (+/-0.08)	0.97 (+/-0.01)
R^2 test CV	0.36 (+/-0.38)	0.01 (+/-1.34)	0.38 (+/-0.27)	0.01 (+/-1.34)	0.50 (+/-0.40)	0.36 (+/-0.46)	0.42 (+/-0.66)
R^2 train complet	0.48	0.50	0.66	0.50	0.94	0.85	0.96
R^2 jeu de test	**0.50**	**0.62**	**0.44**	**0.62**	**0.81**	**0.76**	**0.90**
Hyper paramètres	n_neighbors : 6	alphas : 0.1	C : 1e8 gamma : 0.1	alpha : 0.1 gamma : 1e-05	n_estimators : 200	loss : 'square'	learning_rate : .2 n_estimators : 72

Résultats > optimisation par validation croisée

Émission de CO_2							
Score	**Modèle**						
	KNeighbors	**Ridge**	**SVM**	**kRR**	**Random Forest**	**AdaBoost**	**Gradient Boosting**
R^2 train CV	0.44 (+/-0.05)	0.71 (+/-0.11)	0.49 (+/-0.16)	0.71 (+/-0.11)	0.93 (+/-0.02)	0.84 (+/-0.20)	0.98 (+/-0.01)
R^2 test CV	0.25 (+/-0.16)	0.52 (+/-0.47)	0.40 (+/-0.21)	0.52 (+/-0.47)	0.44 (+/-0.95)	0.31 (+/-0.84)	0.06 (+/-2.50)
R^2 train complet	0.46	0.72	0.53	0.71	0.94	0.84	0.97
R^2 jeu de test	**0.44**	**0.59**	**0.39**	**0.59**	**0.59**	**0.57**	**0.82**
Hyper paramètres	n_neighbors : 6	alphas : 1	C : 1e5 gamma : 0.001	alpha : 1 gamma : 1e-05	n_estimators : 200	loss : 'square'	learning_rate : .2 n_estimators : 50

Résultats > intérêt ENERGY STAR Score

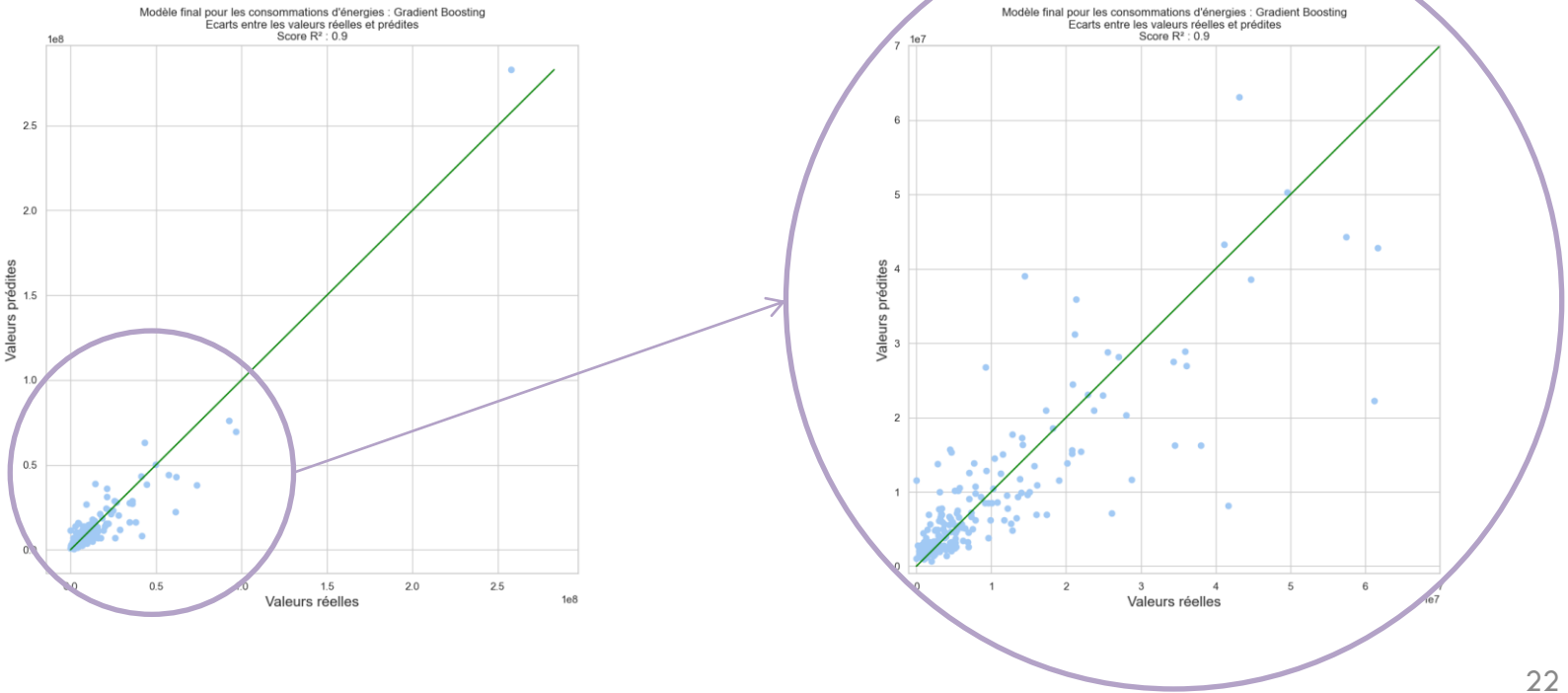
Consommation d'énergies							
Score	**Modèle avec ENERGY STAR Score**						
	KNeighbors	**Ridge**	**SVM**	**kRR**	**Random Forest**	**AdaBoost**	**Gradient Boosting**
R^2 train CV	0.53 (+/-0.13)	0.61 (+/-0.15)	0.46 (+/-0.09)	0.61 (+/-0.15)	0.93 (+/-0.03)	0.88 (+/-0.06)	1.00 (+/-0.00)
R^2 test CV	0.46 (+/-0.32)	0.39 (+/-0.45)	0.48 (+/-0.40)	0.39 (+/-0.45)	0.45 (+/-0.54)	0.28 (+/-0.79)	0.61 (+/-0.36)
R^2 train complet	0.53	0.61	0.48	0.61	0.95	0.87	1.00
R^2 jeu de test	**0.54**	**0.66**	**0.55**	**0.66**	**0.90**	**0.79**	**0.88**
Score	**Modèle sans ENERGY STAR Score**						
	KNeighbors	**Ridge**	**SVM**	**kRR**	**Random Forest**	**AdaBoost**	**Gradient Boosting**
R^2 train CV	0.52 (+/-0.11)	0.60 (+/-0.15)	0.46 (+/-0.09)	0.60 (+/-0.15)	0.93 (+/-0.03)	0.81 (+/-0.08)	0.97 (+/-0.01)
R^2 test CV	0.45 (+/-0.33)	0.38 (+/-0.45)	0.47 (+/-0.32)	0.38 (+/-0.45)	0.49 (+/-0.48)	0.20 (+/-0.91)	0.61 (+/-0.44)
R^2 train complet	0.52	0.61	0.48	0.61	0.95	0.78	0.97
R^2 jeu de test	**0.56**	**0.65**	**0.50**	**0.65**	**0.92**	**0.71**	**0.90**

Résultats > intérêt ENERGY STAR Score

Émission de CO_2							
Score	**Modèle avec ENERGY STAR Score**						
	KNeighbors	**Ridge**	**SVM**	**kRR**	**Random Forest**	**AdaBoost**	**Gradient Boosting**
R^2 train CV	0.46 (+/-0.16)	0.62 (+/-0.14)	0.92 (+/-0.07)	0.62 (+/-0.14)	0.93 (+/-0.04)	0.94 (+/-0.07)	1.00 (+/-0.00)
R^2 test CV	0.32 (+/-0.15)	-0.43 (+/-2.08)	0.62 (+/-0.31)	-0.43 (+/-2.10)	-0.07 (+/-1.67)	0.39 (+/-0.73)	0.57 (+/-0.43)
R^2 train complet	0.48	0.65	0.93	0.65	0.96	0.95	0.99
R^2 jeu de test	**0.46**	**0.55**	**0.86**	**0.54**	**0.80**	**0.72**	**0.77**
Score	**Modèle sans ENERGY STAR Score**						
	KNeighbors	**Ridge**	**SVM**	**kRR**	**Random Forest**	**AdaBoost**	**Gradient Boosting**
R^2 train CV	0.45 (+/-0.13)	0.62 (+/-0.14)	0.92 (+/-0.07)	0.62 (+/-0.14)	0.94 (+/-0.05)	0.94 (+/-0.05)	0.99 (+/-0.01)
R^2 test CV	0.34 (+/-0.13)	-0.41 (+/-2.07)	0.63 (+/-0.25)	-0.42 (+/-2.09)	0.08 (+/-1.38)	0.29 (+/-1.06)	0.51 (+/-0.48)
R^2 train complet	0.48	0.64	0.92	0.64	0.96	0.95	0.99
R^2 jeu de test	**0.49**	**0.52**	**0.81**	**0.52**	**0.85**	**0.73**	**0.74**

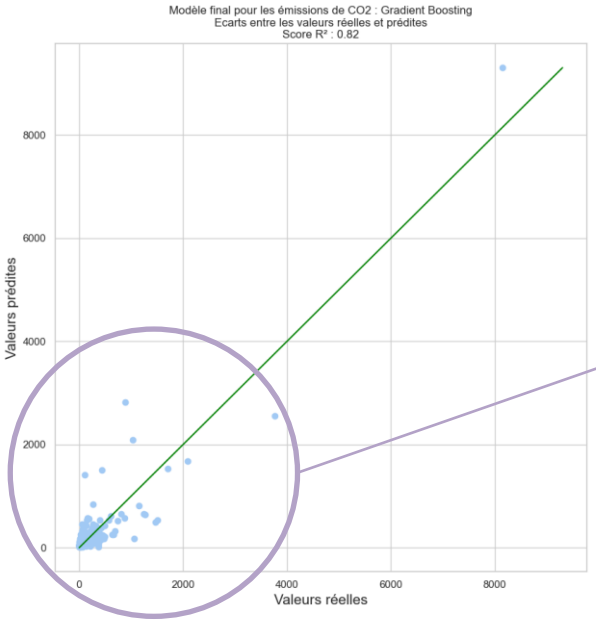
Résultats > choix final énergies

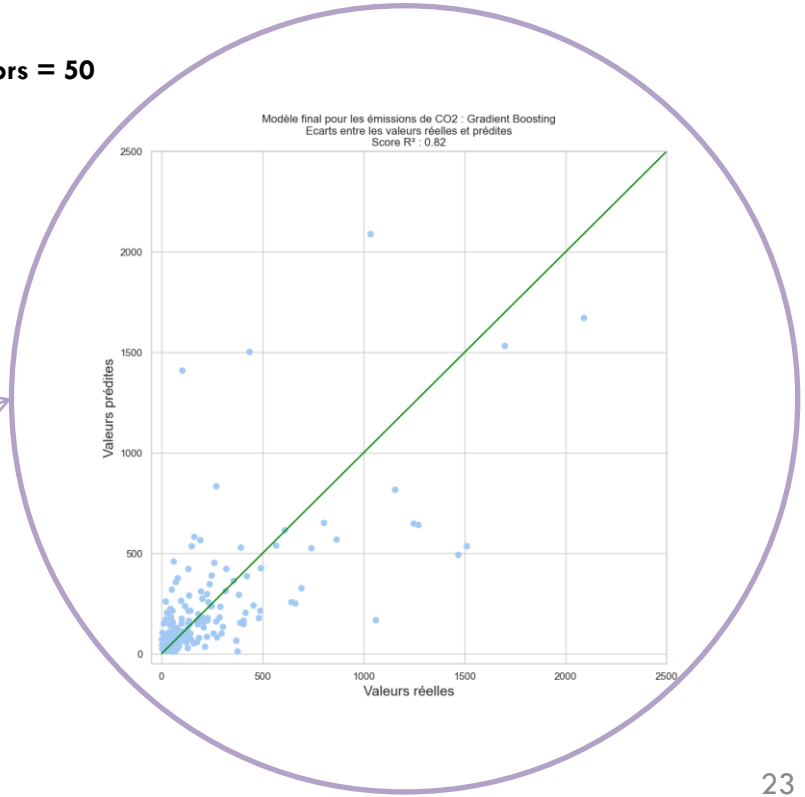
Modèle : Gradient Boosting
Hyper paramètres : learning_rate = 0.2, n_estimators = 72



Résultats > choix final CO_2

Modèle : Gradient Boosting
Hyper paramètres : learning_rate = 0.2, n_estimators = 50





Avez-vous des questions ?