


Anticipez les besoins en consommation électrique de bâtiments

Par : BAHRI
Abdelghani

SOMMAIRE


- ▶ Contexte et objectifs
 - ▶ Analyse exploratoire
 - * Nettoyage et Concaténation
 - * L'analyse univariée/ multivariée
 - ▶ Feature engineering
 - ▶ Modélisation et comparaison des modèles
 - ▶ Conclusion
- 

Contexte et objectifs

Contexte

- ▶ Atteindre l'objectif de la ville de Seattle neutre en émissions de carbone en 2050

Objectifs


- ▶ Prédire les émissions de CO₂ et la consommation totale d'énergie de bâtiments
 - ▶ Tester les différents modèles de prédiction
 - ▶ Sélectionner le meilleur modèle et effectuer un réglage hyperparamétrique pour optimiser le modèle.
- 

Seattle

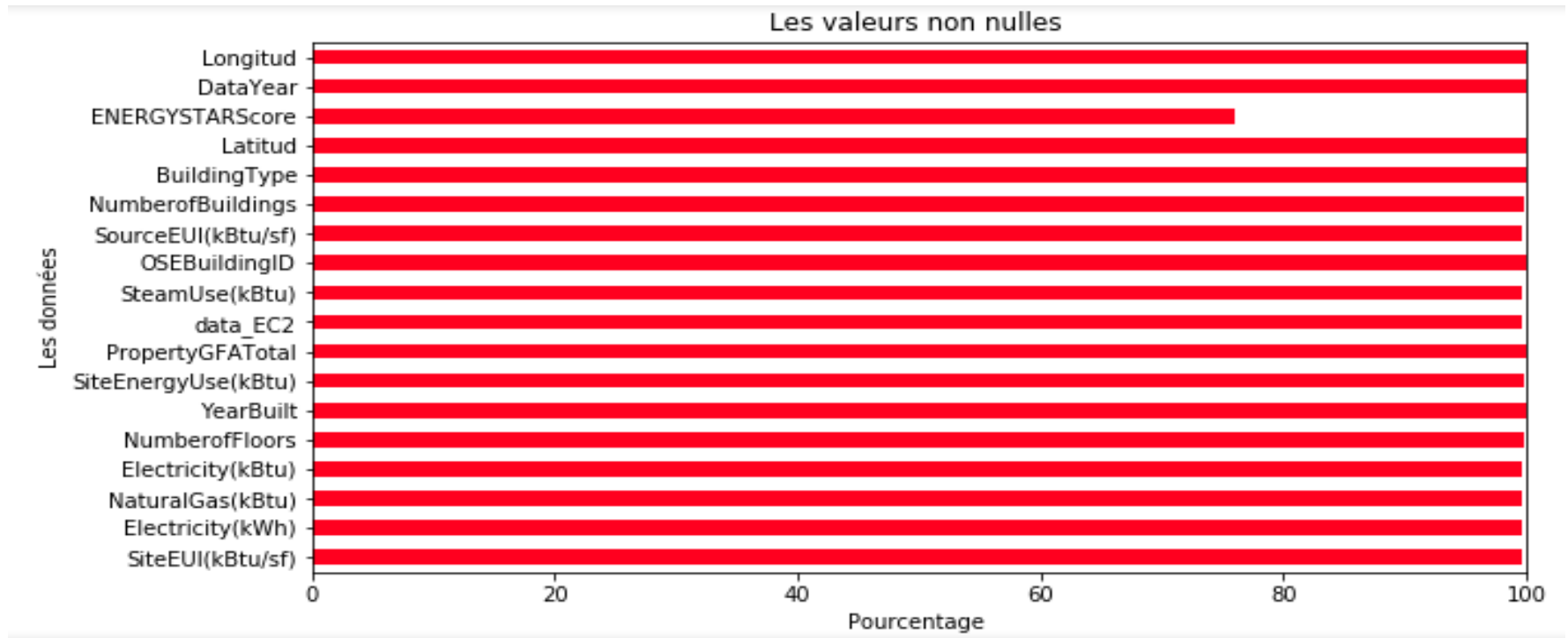
- ▶ Seattle est l'une des villes les plus importantes de la côte américaine
- ▶ La métropole de Seattle est le centre technologique commercial, culturel et avancé des États-Unis



Nettoyage et Concaténation

- ▶ Une ligne d'un fichier représente les informations pour un bâtiment:
 - informations du permis de construction
 - informations sur la consommation d'énergie
 - ▶ Merger les deux fichiers 2015 et 2016
 - ▶ Supprimer les données avec plus de 60% de valeur NaN en fonction des méthodes statistiques utilisées
 - ▶ Supprimer les champs inutiles
- 

Visualiser la consistance des données non nulles



Analyse exploratoire

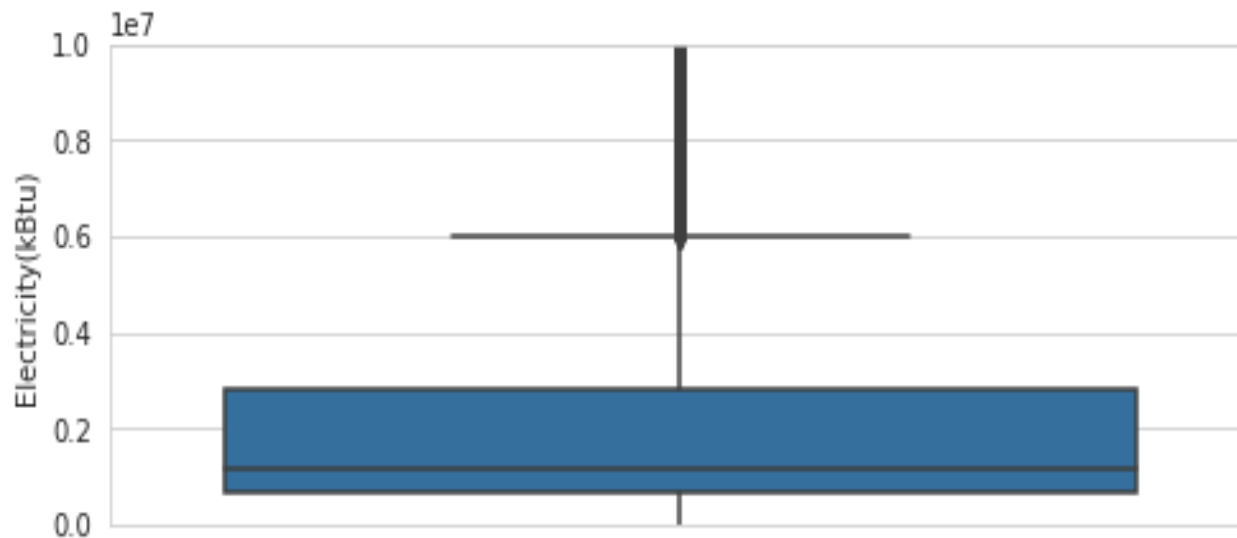
- ▶ L'analyse univariée
- ▶ L'analyse multivariée



Analyse Univariée

Electricité

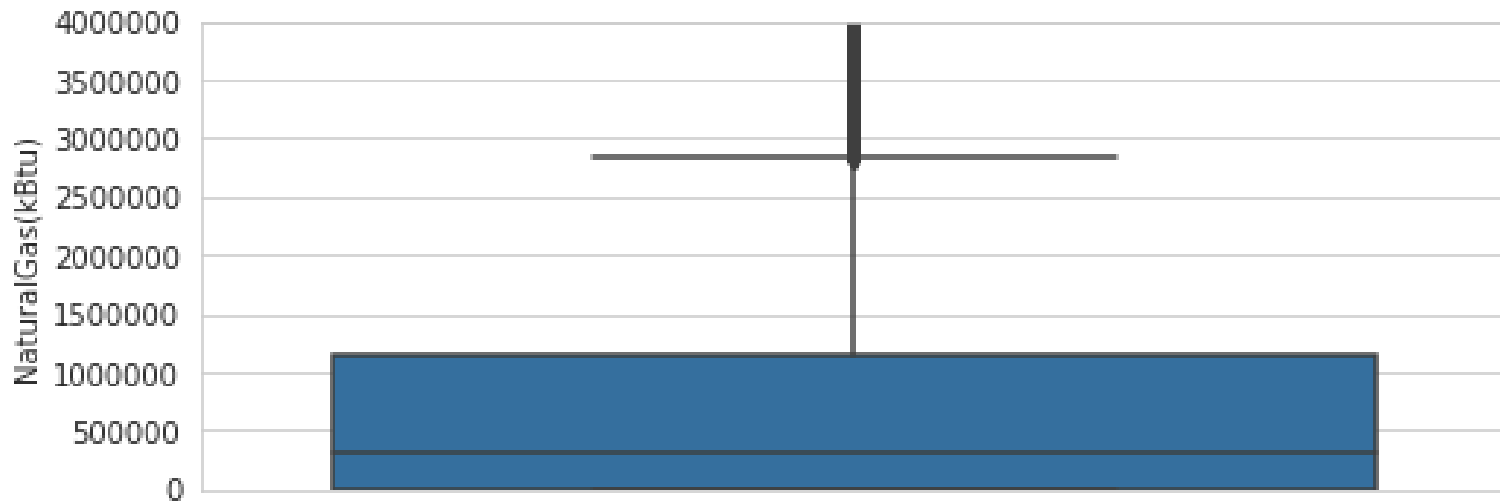
Un peu bancal



Analyse Univariée

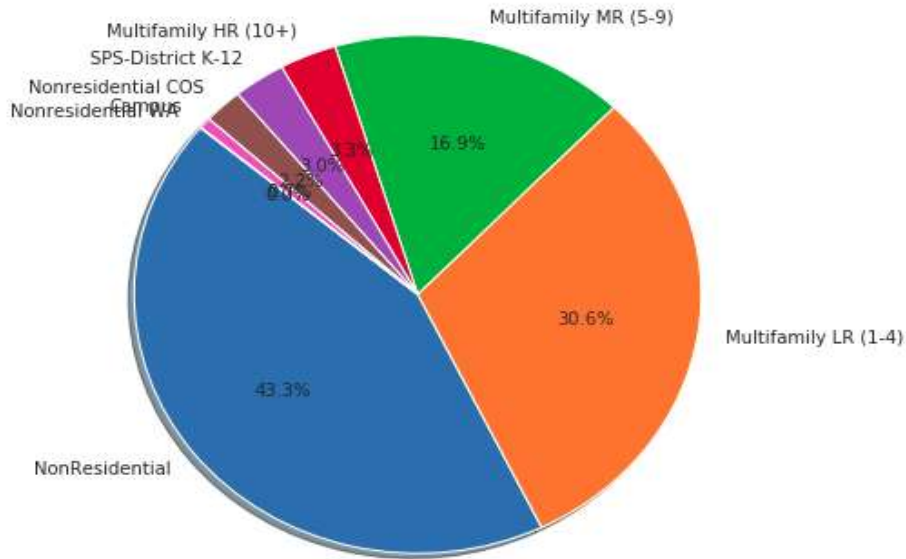
Gaz naturel

Moins écrasée



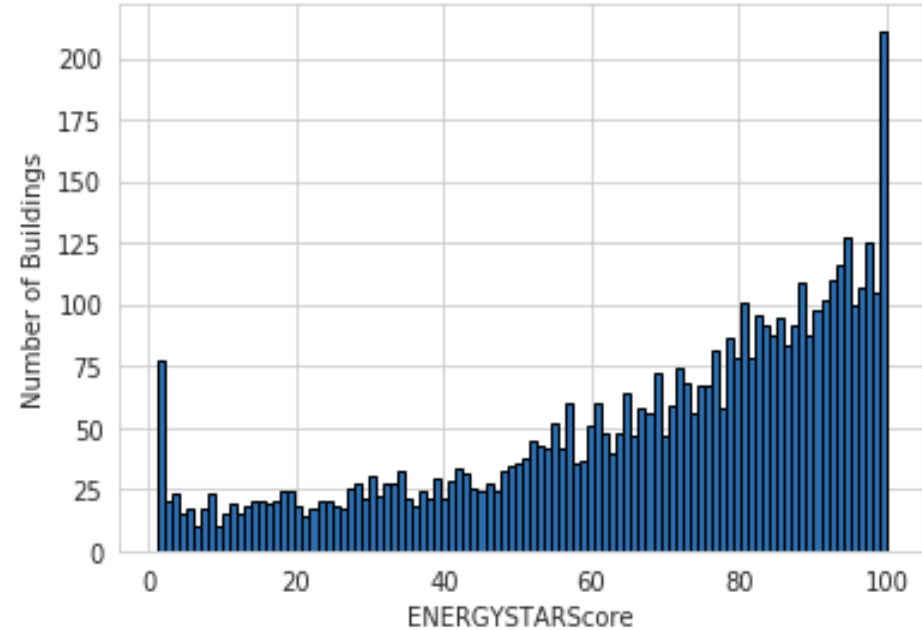
Distribution

Répartition des batiments selon leurs classement denergie



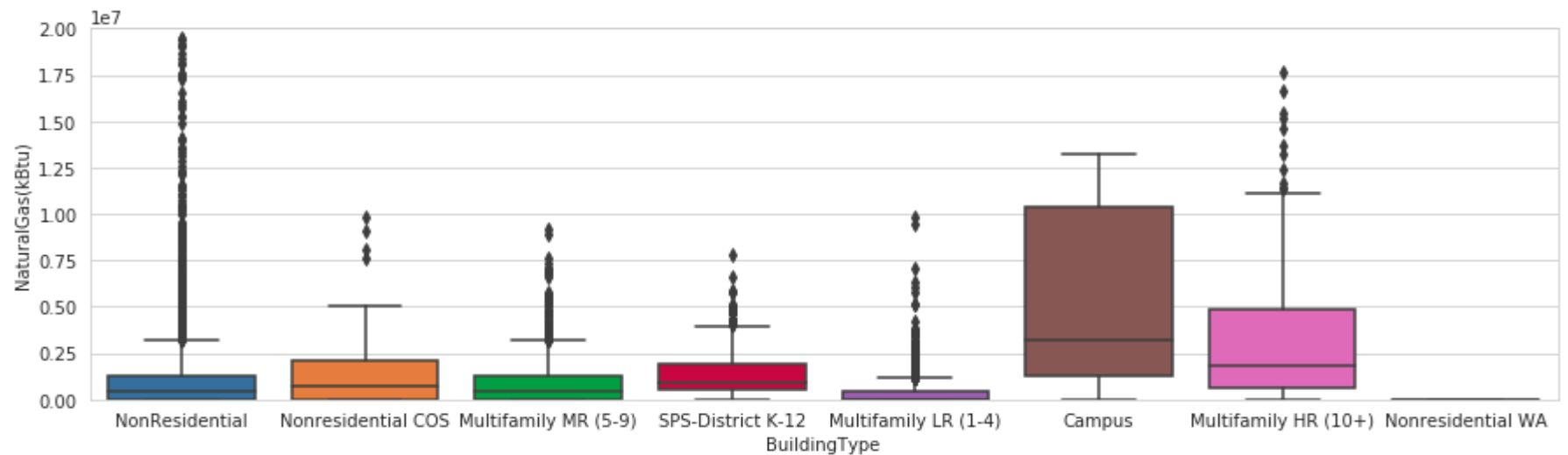
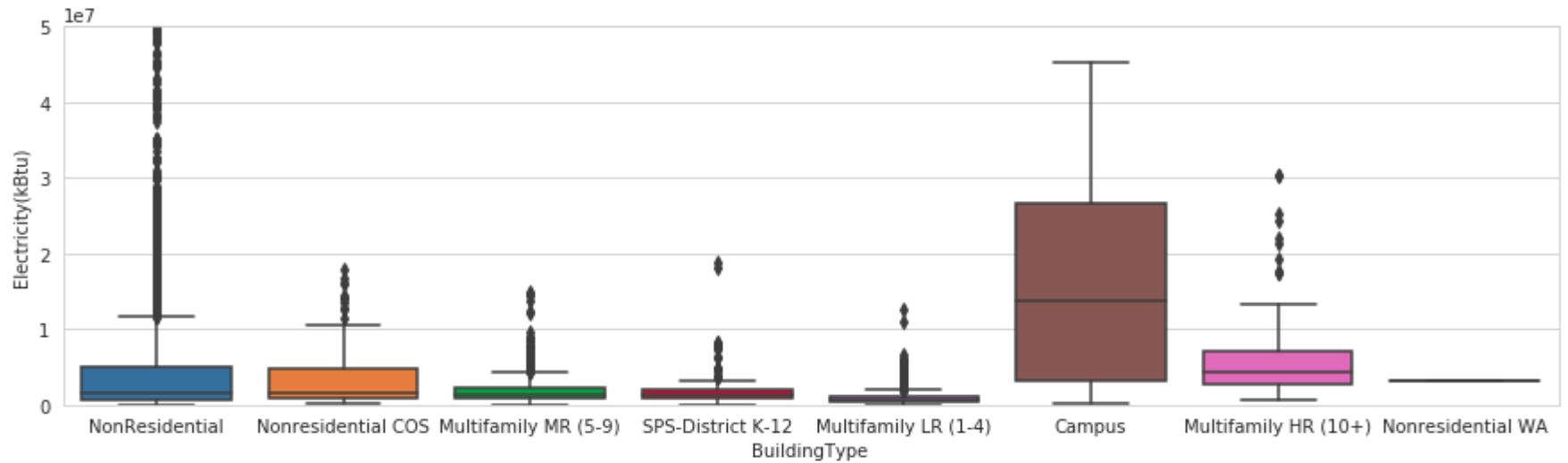
Répartition par type de batiment

Energy Star Score Distribution




Distribution par Energy Star Score

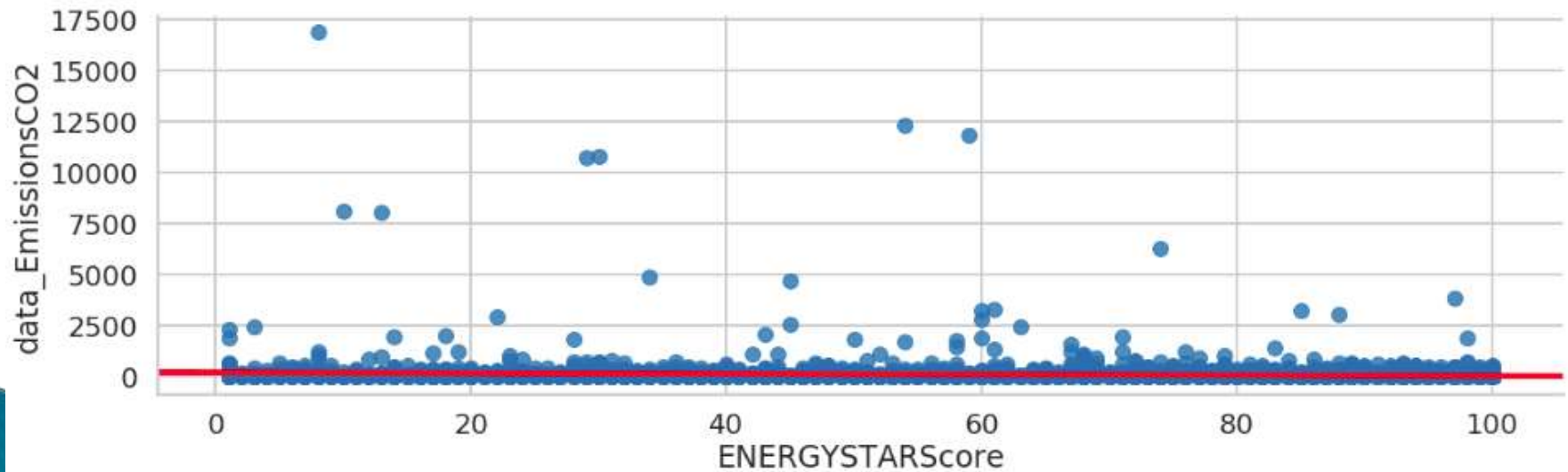
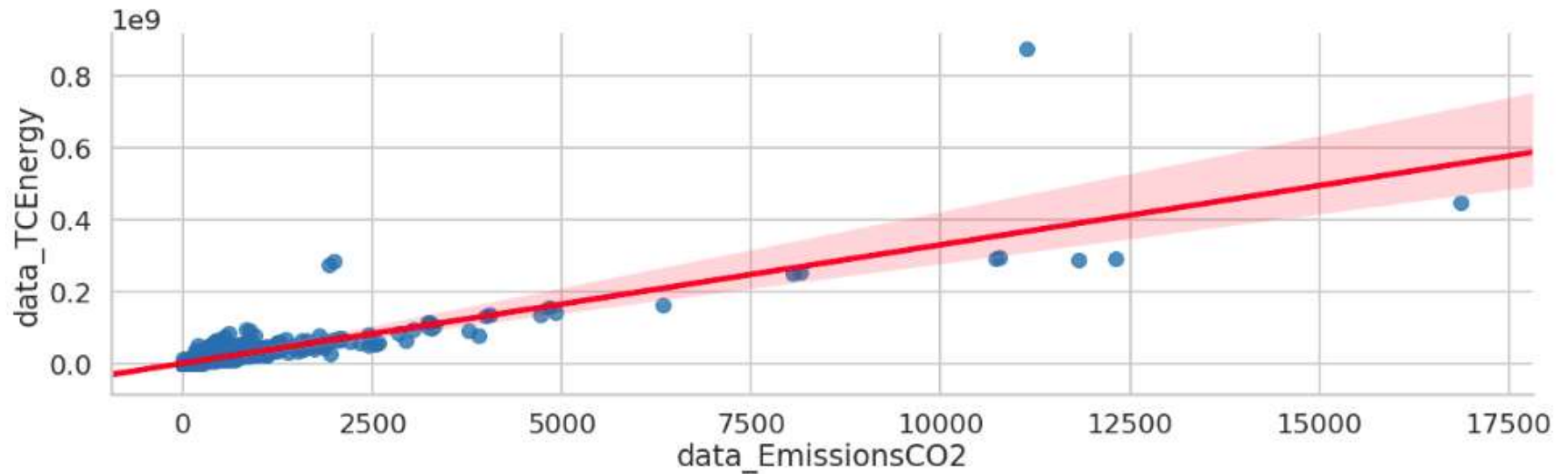
Analyse Multivariée



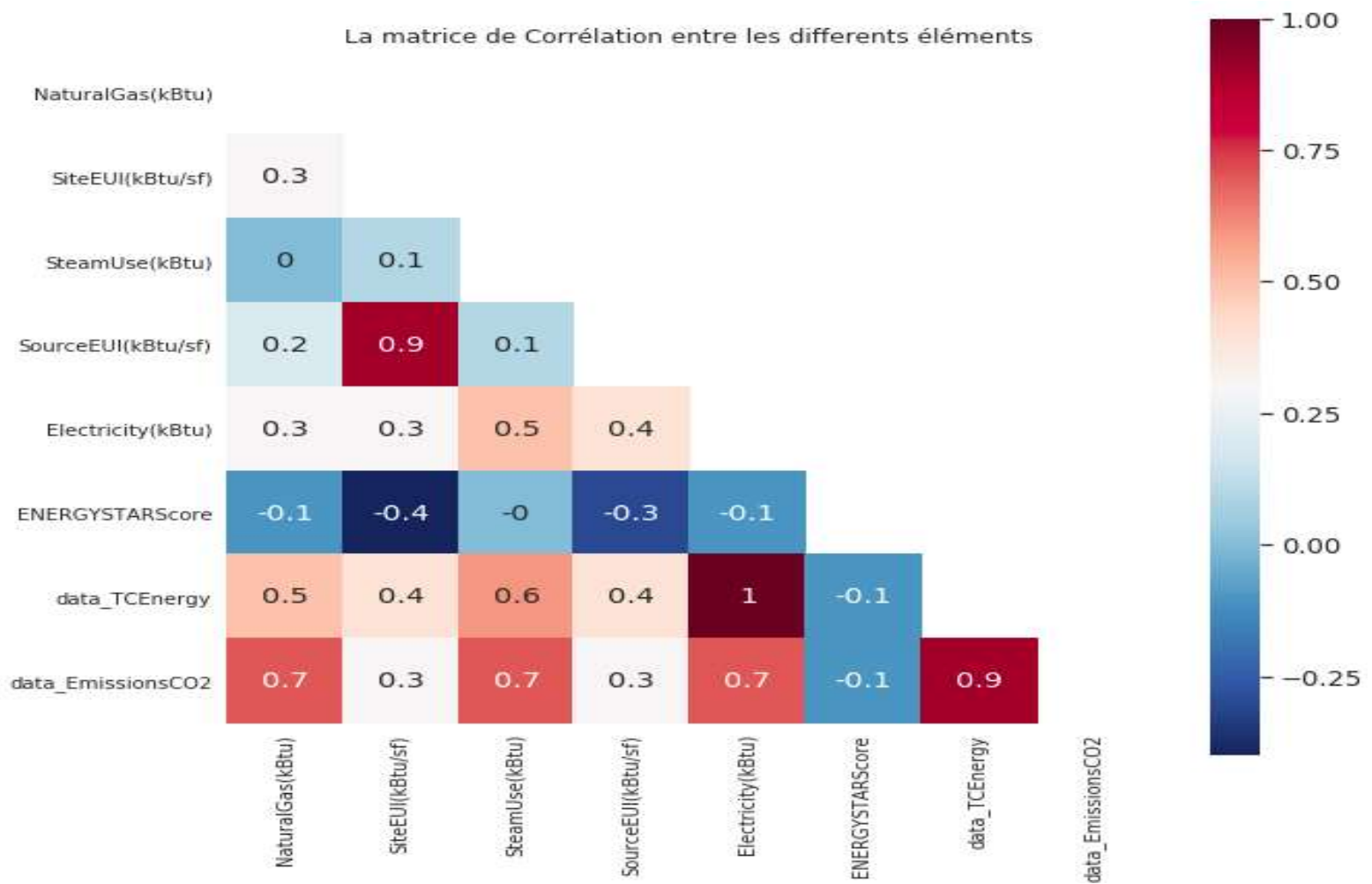
Feature engineering

- ▶ Une target qui représente la consommation totale d'énergie
 - ▶ Une autre target qui représente les émissions de CO2
 - ▶ La prédiction ce basera sur les données déclaratives du permis d'exploitation commerciale
- 

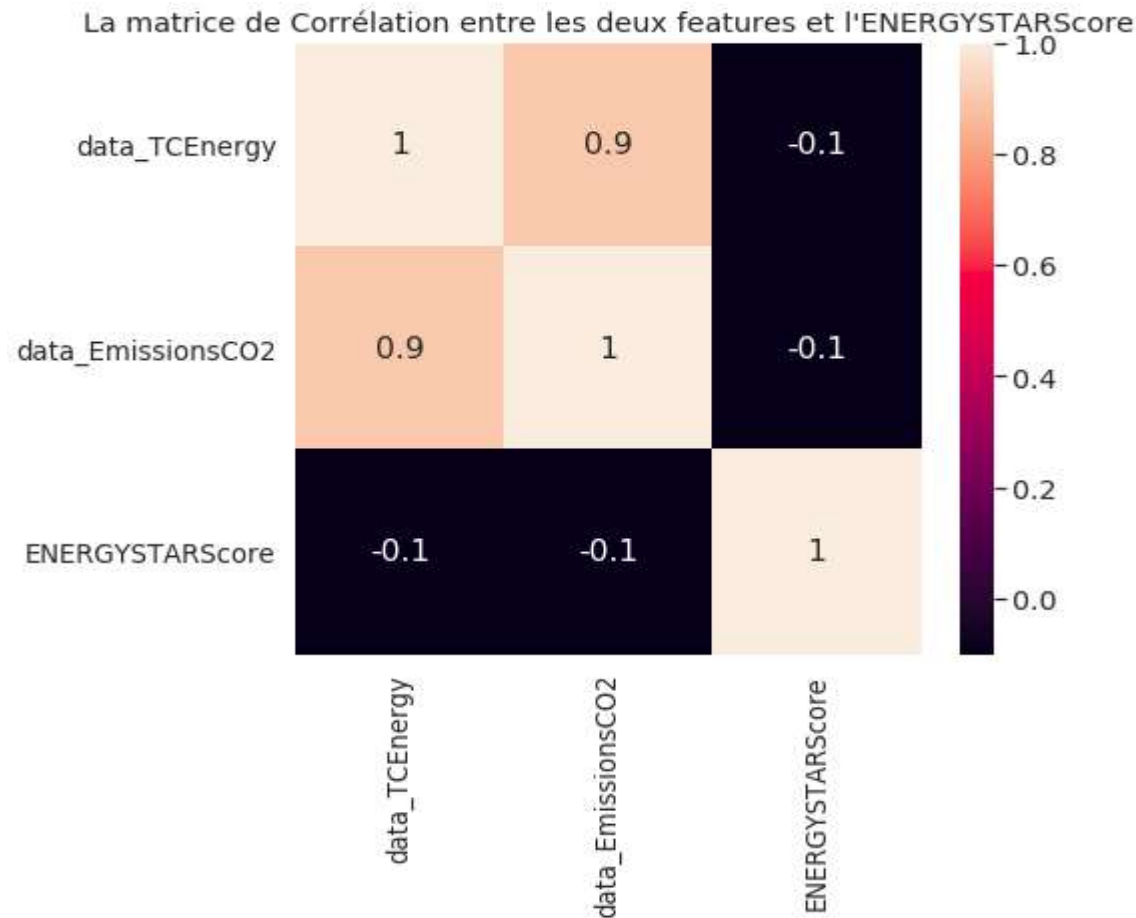
Représentation des données selon deux dimensions



Matrice de corrélation




Corrélation entre les deux features et l'EnergyStarScore



Modèles de prédiction

- ▶ Modèle de la consommation totale d'énergies (CTE)
- ▶ Modèle des émissions du CO₂ (ECO₂)

Modélisation

- ▶ Transformer les données de la colonne des types de batiments on forme binnaire
 - ▶ Choisir les composants qui correspond à chaque modèle de prédiction
 - ▶ Créer une nouvelle base de données qui correspond à nos modèles de prédiction (soit CTE et ECO2)
- 


Modélisation

Hyperparamètre

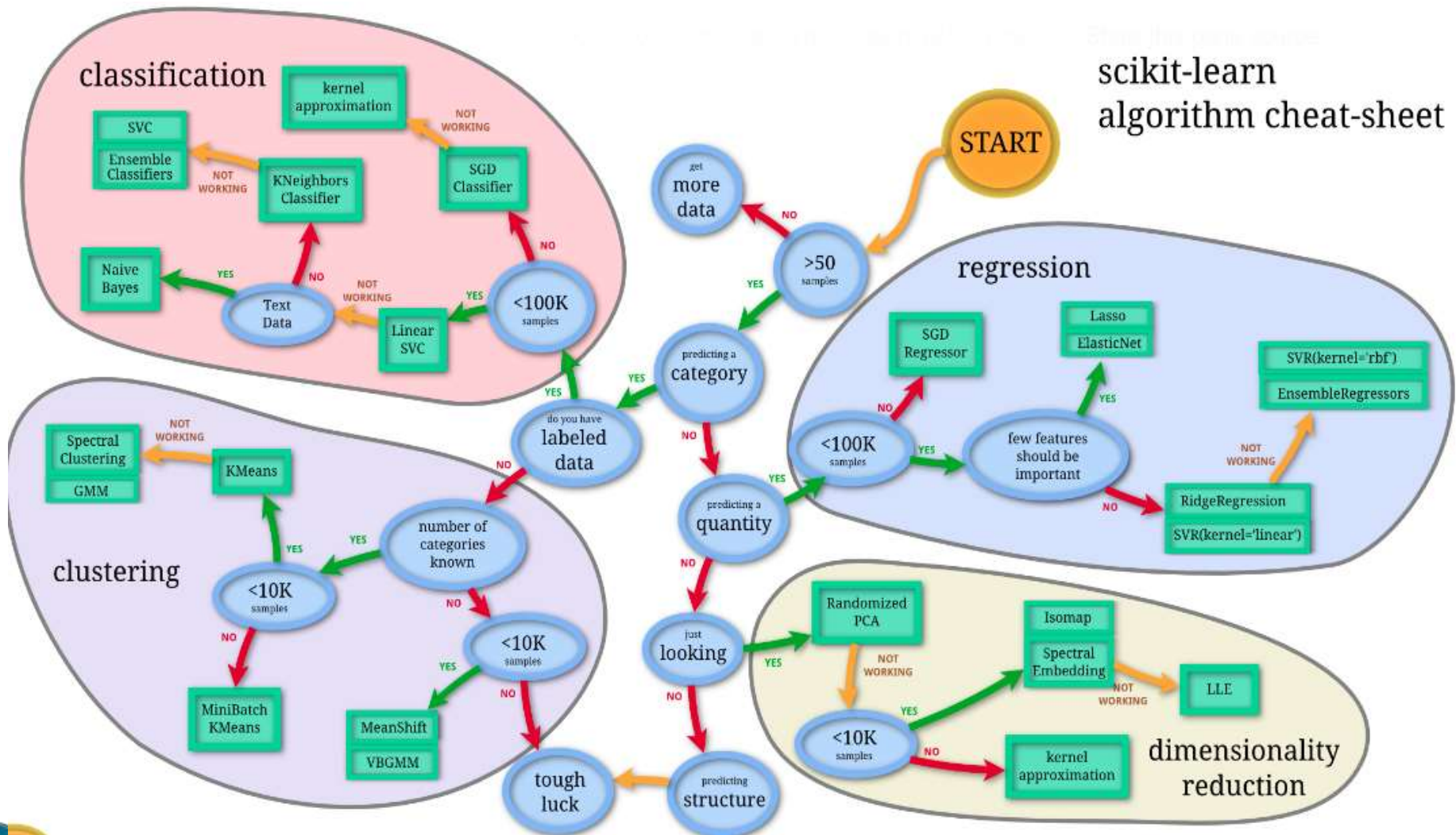
Un hyperparamètre est un paramètre dont la valeur est définie avant le début du processus d'apprentissage

Validation croisée

La validation croisée va nous permettre d'utiliser l'intégralité de notre jeu de données pour l'entraînement et pour la validation



Modèles de prédiction



Modèle de prédiction CTE et ECO2

► Utiliser 5 modèles de regression soit :

1 – Linear Regression (LR)

2 – Kernel Ridge Regression (linear & non linear avec cross-validation)

3 – SVR non linear avec cross-validation

4 – MLPRegressor

5 – Gradient Boosted Regression Trees avec cross-validation



Evaluation et modèle final de (CTE et ECO2)

- ▶ Modèle final de CTE
- ▶ Modèle final de ECO2

Evaluation et modèle final de (CTE)

Model	R2(Score)	RMSE	M. Coefficients	Temp du calcul
Linear Regression	44.11%	8049789	PropertyGFATotal	0.005s
Ridge Regression	44.11%	8049789	PropertyGFATotal	0.003s
Ridge Regression–CV	44.11%	8049789	PropertyGFATotal	0.292s
Kernel Ridge Regression–Non Linear	51.33%	7511353	/	20.400s
Kernel Ridge Regression–Non Linear –CV	7.68%	10345202	/	1455.042s
SVR non linear –CV	-8.07%	11193302	/	27.714s
MLPRegressor	32.06%	8874532	/	6.200s
Gradient Boosted Regression Trees – CV	78.24%	5022537	PropertyGFATotal	114.324s

Evaluation et modèle final de (ECO2)

Model	R2(Score)	RMSE	M. Coefficients	Temp du calcul
Linear Regression	25.89%	446.19	PropertyGFATotal	0.007s
Ridge Regression	25.88%	446.21	PropertyGFATotal	0.005s
Ridge Regression – CV	25.88%	446.21	PropertyGFATotal	0.376s
Kernel Ridge Regression – Non Linear – CV	38.11%	407.74	/	681.333s
SVR non linear – CV	11.30%	488.15	/	28.847s
MLPRegressor	12.70%	484.28	/	0.527s
Gradient Boosted Regression Trees – CV	70.19%	282.95	PropertyGFATotal	115.878s

Conclusion

- ▶ Une analyse de données exploratoire a permis de filtrer et interpréter les données importantes
 - ▶ Développer un ensemble de features que nous utiliserons pour nos modèles
 - ▶ Evaluation les différents modèles de regression sur CTE et ECO2
 - ▶ Optimiser et trouver le meilleur modèle.
- 