ANTICIPEZ LES BESOINS EN CONSOMMATION DE BÂTIMENTS

DATA SCIENCE

MÉTHODOLOGIE

I. PRESENTATION

II.DESCRIPTION DES DONNEES

III. DATA CLEANING

IV. ANALYSE EXPLORATOIRE

V. MODELISATION

VI. CONCLUSION

PRESENTATION

Contexte

- Faire de Seattle, une ville neutre en émission de carbone en 2050.
- S'intéresser à la consommation et aux émissions des bâtiments non destinés à l'habitation

Objectifs

- Trouver un meilleur modèle de prédiction des émissions de CO2 et la consommation totale d'énergie de bâtiments non destinés à l'habitation
- Evaluer l'intérêt de " l'ENERGY STAR Score " pour la prédiction d'émissions

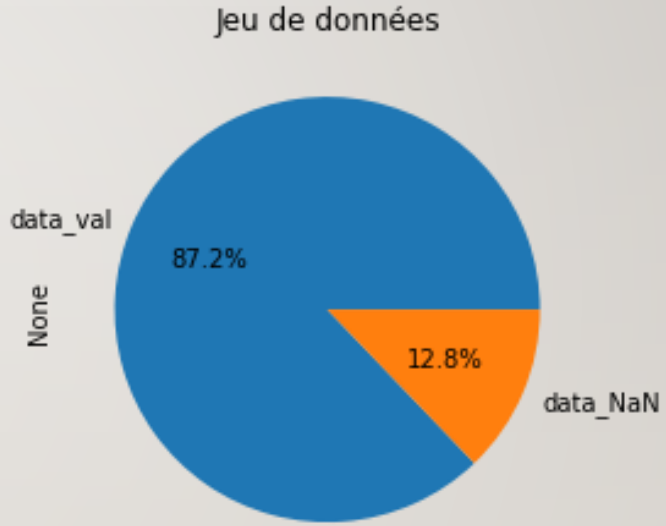
Contraintes

- Relevés de consommation de 2016
- La prédiction se basera sur les données structurelles des bâtiments (taille et utilisation des bâtiments, date de construction, situation géographique, ...)

DESCRIPTION DES DONNEES

Le jeu de données est un fichier de :

- ✓ 3.376 bâtiments (ligne)
- ✓ 46 champs ou variables (colonnes)



Aucune ligne en doublons n’a été identifiée dans le jeu de données

DATA CLEANING

Il a consisté à :

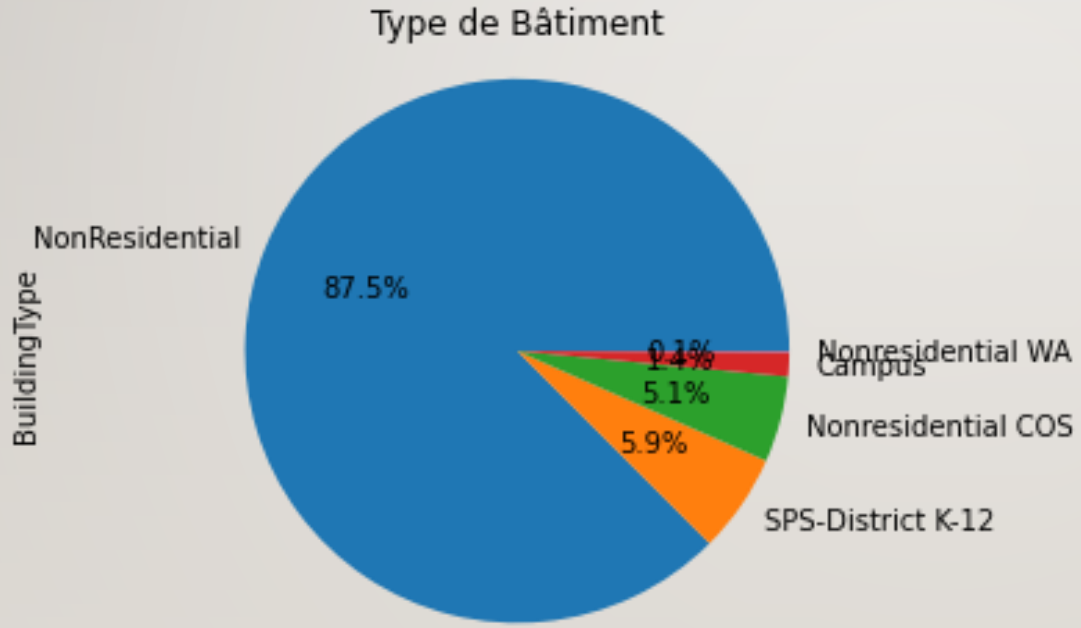
- Sélectionner les variables pertinentes
- Eliminer les valeurs manquantes
- Corriger les Outliers

	Données initiales	Données finales
Nbre de Lignes	3.376	1.663
Nbre de colonnes	46	25

ANALYSE EXPLORATOIRE

Analyse des variables des bâtiments

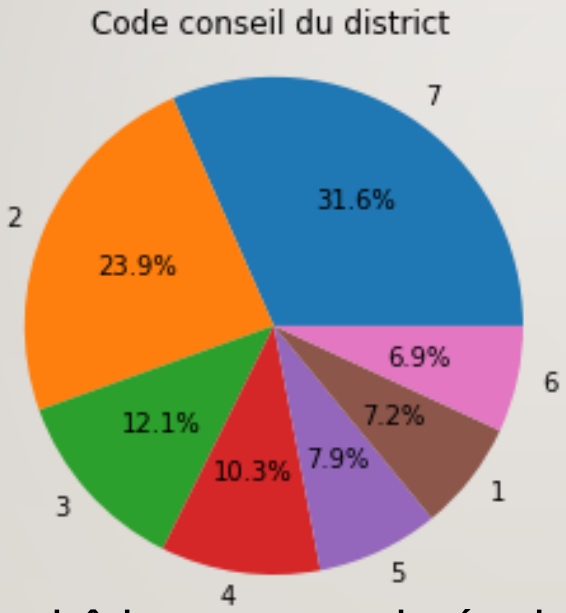
Type de bâtiments



ANALYSE EXPLORATOIRE

Analyse des variables des bâtiments

Conseil de district et Quartier



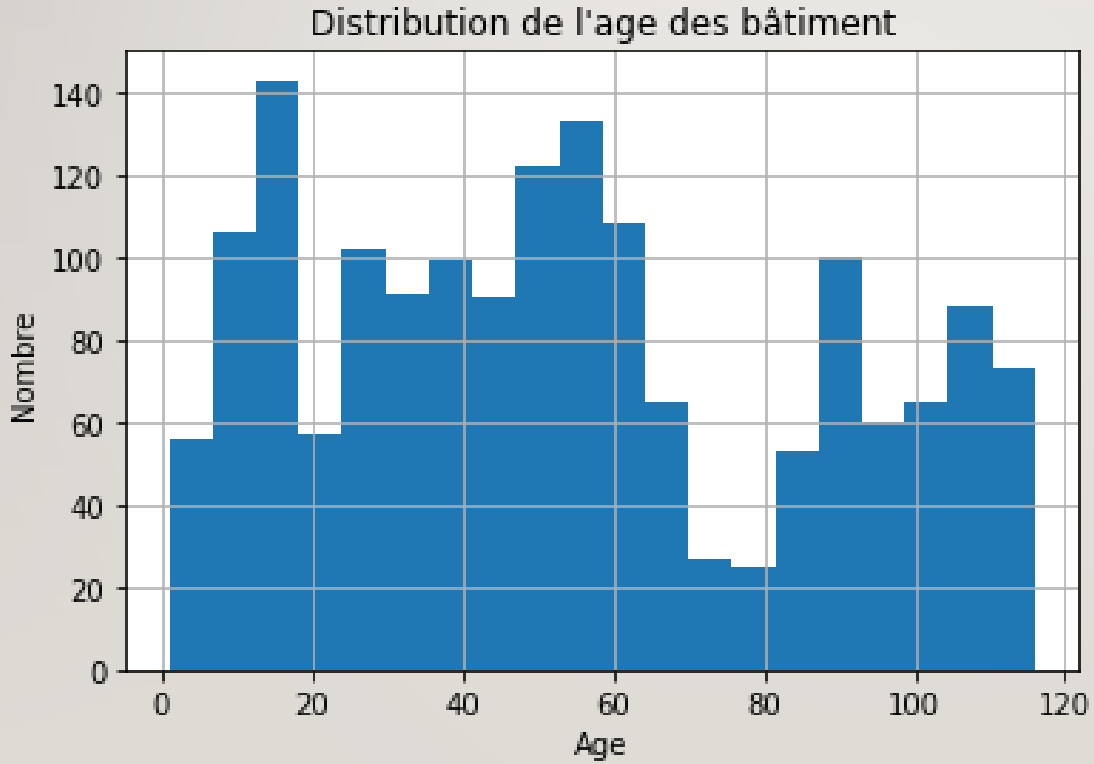
Quartier	Nbre	Pourcent (%)
DOWNTOWN	358	0.215
GREATER DUWAMISH	346	0.208
MAGNOLIA / QUEEN ANNE	151	0.091
LAKE UNION	148	0.089
NORTHEAST	127	0.07

- **60%** des bâtiment sont situés dans 4 quartiers sur les 19
- Plus de la moitié des bâtiments sont situés dans les conseils de district 7 et 2

ANALYSE EXPLORATOIRE

Analyse des variables des bâtiments

Age des bâtiments

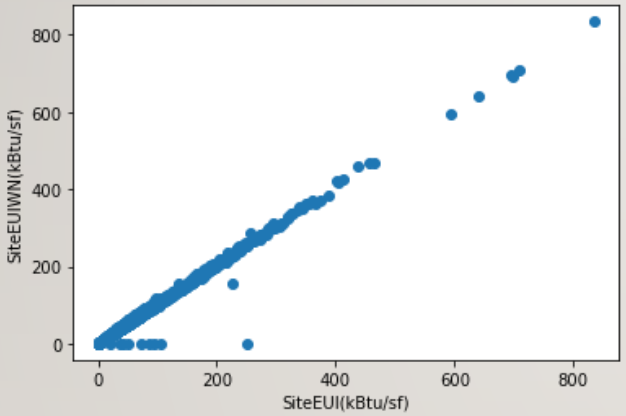


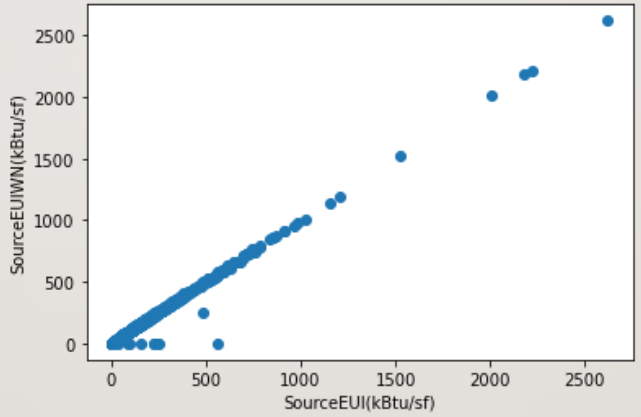
- La moyenne d'âge des bâtiment est de : **54**
- Le mode d'âge des bâtiment est de : **116**

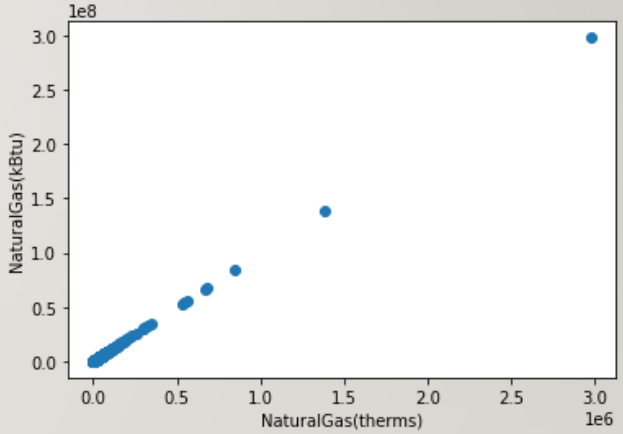
ANALYSE EXPLORATOIRE

Analyse de la corrélation

La corrélation entre certaines variable





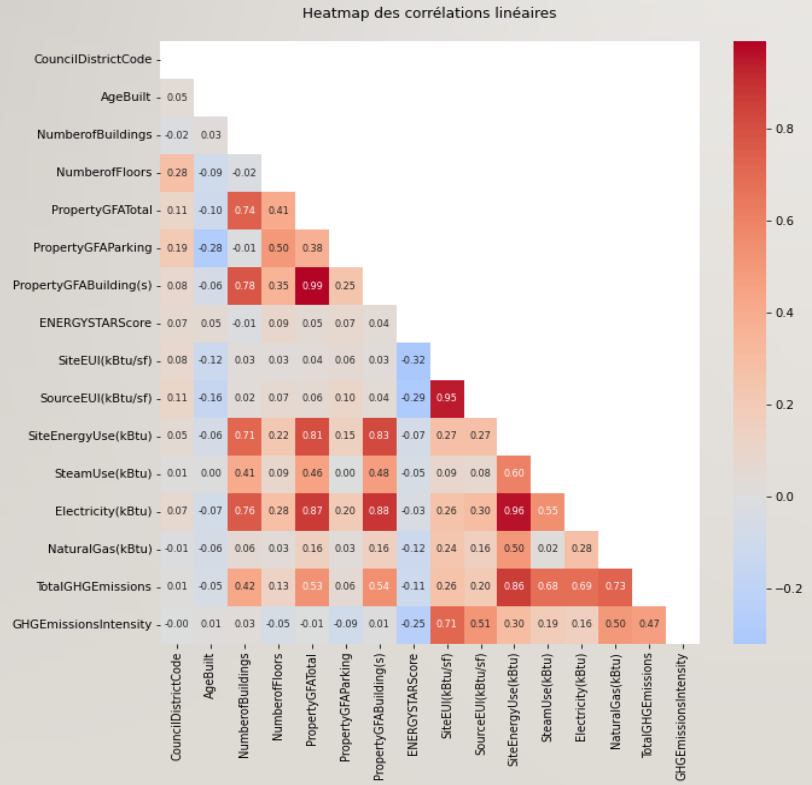


- Certaines variables sont répétées mais en des unités de mesures différentes
- Nous avons donc procéder à leur suppression .

ANALYSE EXPLORATOIRE

Analyse de la corrélation

Heatmap des Corrélations linéaires



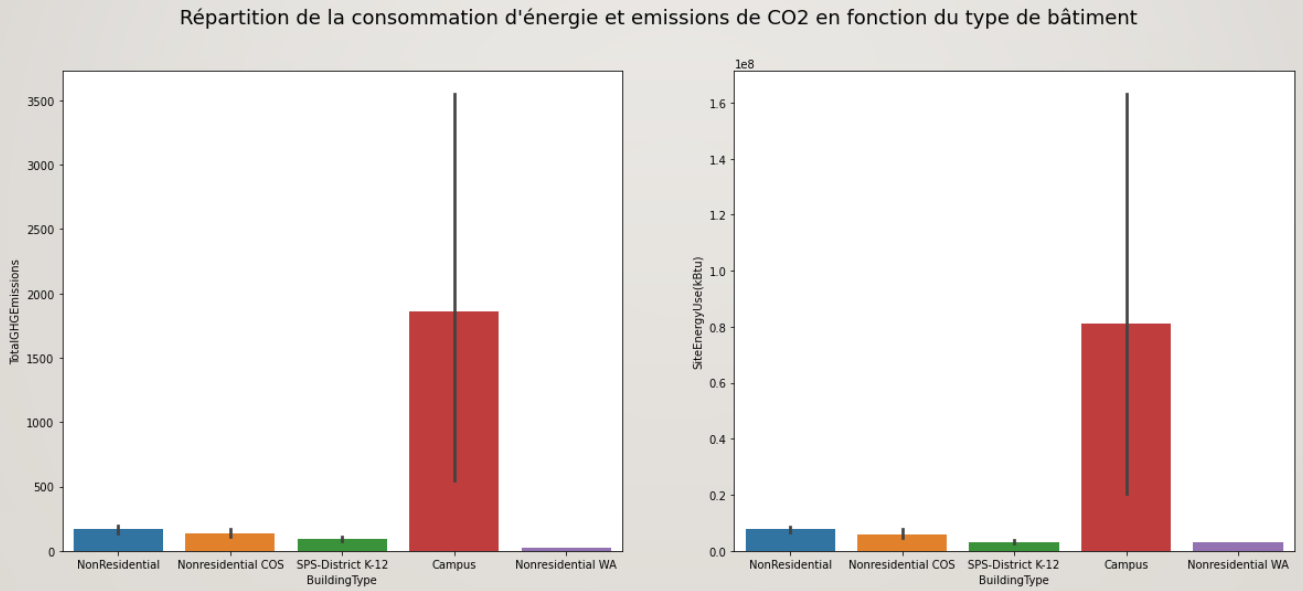
Pour les varaibles à prédire TotalGHGEmissions et SiteEnergyUse(kBtu), on remarque des corrélations linéaires avec :

- les variables de relevés (les consommations)
- les variables de nombre de bâtiments et de surfaces,

ANALYSE EXPLORATOIRE

Analyse de la corrélation

Les variables à prédire



Les campus se démarquent largement en terme de consommation et de rejets de CO2

PREDICTION

Feature Engineering

- **Transformation logarithmique**
- **Renommage de variables**
- **One Hot Encoding**

PREDICTION

Feature Engineering

Out[5]:

CouncilDistrictCode	AgeBuilt	NumberofBuildings	NumberofFloors	PropertyGFATotal	PropertyGFAParking	PropertyGFABuilding(s)	ENERGYSTARScore	SiteEUI(kE
7	89	1.0	12	11.390012	0.000000	11.390012	60.0	4.4
7	20	1.0	11	11.547964	9.620063	11.390780	61.0	4.5
7	47	1.0	41	13.770628	12.189527	13.540273	43.0	4.5
7	90	1.0	10	11.023861	0.000000	11.023861	56.0	4.7
7	36	1.0	18	12.075850	11.034890	11.640263	75.0	4.7

ws × 59 columns

Out[5]:

PropertyGFAParking	PropertyGFABuilding(s)	ENERGYSTARScore	SiteEUI(kBtu/sf)	SourceEUI(kBtu/sf)	...	Neighborhood_9	Neighborhood_10	Neighborhood_11	N
0.000000	11.390012	60.0	4.403054	5.206750	...	0.0	0.0	0.0	
9.620063	11.390780	61.0	4.551769	5.171052	...	0.0	0.0	0.0	
12.189527	13.540273	43.0	4.564348	5.488524	...	0.0	0.0	0.0	
0.000000	11.023861	56.0	4.707727	5.376204	...	0.0	0.0	0.0	
11.034890	11.640263	75.0	4.743192	5.353752	...	0.0	0.0	0.0	

MODÉLISATION

Prédiction de l'énergie

Méthodologie

- Création du modèle
- Evaluation du modèle
- Features importance

MODÉLISATION

Prédiction de l'énergie

Algorithmes

- Régression linéaire
- Forêt Aléatoire
- Gradient Boosting
- MLP
- XGBoost

- Features importance

MODÉLISATION

Prédiction de l'énergie

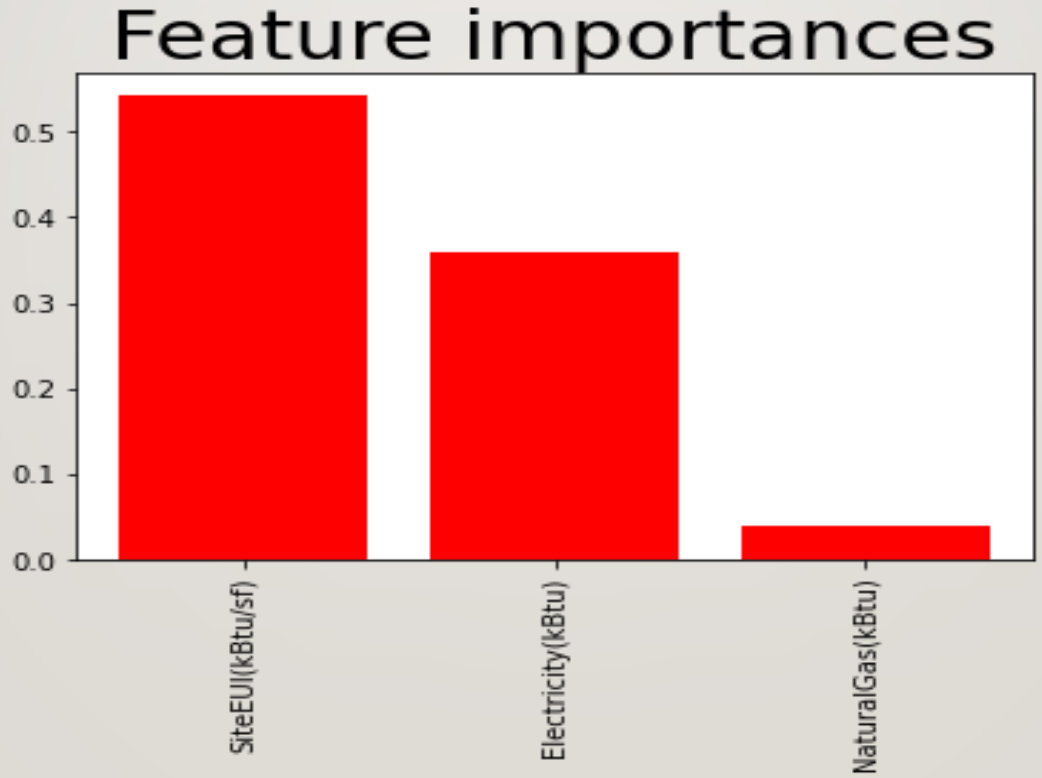
Choix du modèles

Algorithmes	R^2
Régression Linéaire	0.654792
Gradient Boosting	0.898771
Forêt Aléatoire	0.94188
MLP	0.842322
XGBoost	0.936882

MODÉLISATION

Prédiction de l'énergie

Features importance



MODÉLISATION

Prédiction de l'énergie

Tableau comparatif

Score	Régression Linéaire	Gradient Boosting	Forêt Aléatoire	MLP	XGBoost
R^2(avec ENERGYSTARScore)	0.821597	0.979401	0.954553	0.912775	0.986726
R^2(Sans ENERGYSTARScore)	0.773453	0.979553	0.975560	0.748997	0.994139

MODÉLISATION

Prédiction du CO2

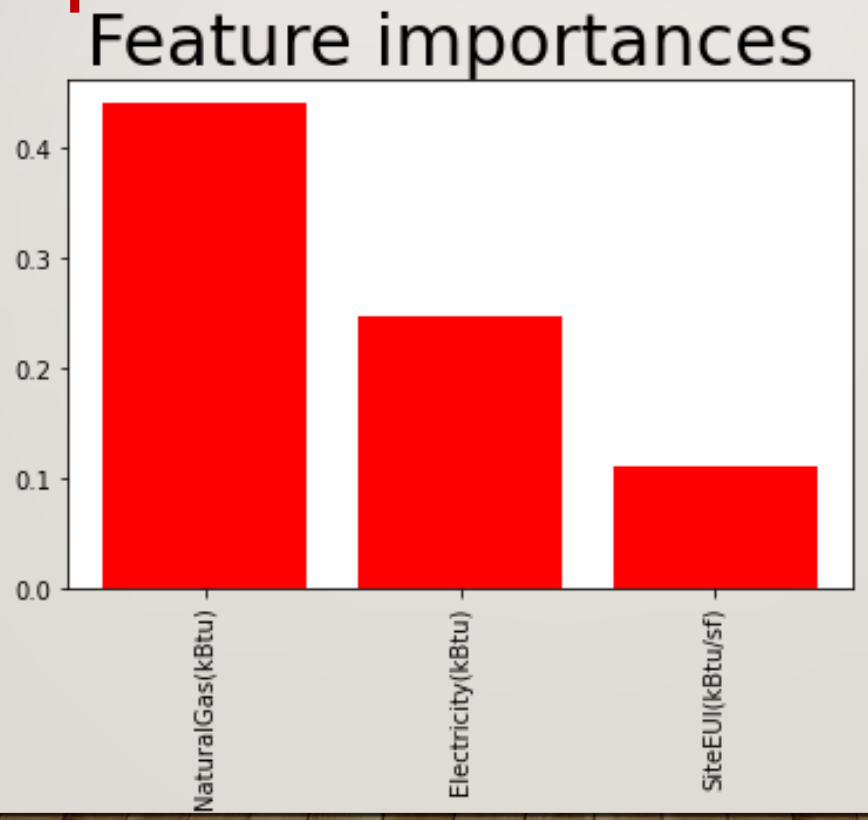
Choix du modèle

Algorithmes	R^2
Régression Linéaire	0.76759
Gradient Boosting	0.95251
Forêt Aléatoire	0.99088
MLP	0.834827
XGBoost	0.99271

MODÉLISATION

Prédiction du CO2

Features importance



MODÉLISATION

Prédiction du CO2

Tableau comparatif

Scores	Régression Linéaire	Gradient Boosting	Forêt Aléatoire	MLP	XGBoost
R^2(avec ENERGYSTARScore)	0.916986	0.994943	0.961819	0.879166	0.996297
R^2(Sans ENERGYSTARScore)	0.918099	0.988192	0.953314	0.842674	0.988086

CONCLUSION

Au vu de l'analyse de nos différents algorithmes de modélisation, il ressort que :

Le meilleur modèle pour la prédiction de la consommation de l'énergie est le **Gradiant Boosting** et de l'émission de CO_2 est l'algorithme **XGBoost.**

La suppression de la variable "ENERGYSTARScore" n'a pas d'impact majeur sur nos modèles. Aussi, Les variables ayant le plus d'importance dans la prédiction :

CONCLUSION

L'ENERGIE :

SiteEUI(kBtu/sf) (54%)

Electricity(kBtu) (36%)

CO2 :

NaturalGas(kBtu) (44%)

Electricity(kBtu) (24%).

MERCI

OPENCLASSROOM