Anticiper les besoins en consommation électrique des bâtiments

Seattle

Neutralité Carbone 2050

Sommaire

1. **Problématique**
2. Données
3. Modélisation
4. Conclusion

Problématique - Contexte

Objectif de la ville de Seattle:

Neutralité Carbone en 2050

➡ connaitre leurs **consommations** en **énergie** et **émission**

Problème:

Des relevés minutieux et couteux ont été effectués sur les années antérieures (2015 et 2016)

Missions:

Avec **uniquement** les données récoltées:

- **Prédire** la **consommation** totale d'énergie
- **Prédire** les **émissions** de Co2
- Evaluer l'intérêt de la variable **ENERGY STAR Score** pour la **prédiction** d'**émission de Co2**

Problématique – Interprétations/Indications

2 jeux de données (2015 et 2016):

Similarité, doublons, ? Grouper les données

Sélection les variables cibles :

Total des émissions, Intensité?

Site/Site WN?

Sélection des variables indépendantes:

caractéristiques propres aux bâtiments

➡ exclusion des variables d'énergie

Bâtiments non résidentiels:

Filtrer les bâtiments multi-familles ?

Modélisation:

2 variables cibles quantitatives à prédire

➡ 2 modélisations de régression

SiteEnergyUseWN(kBtu)

➡ modèle sur la consommation d'énergie

TotalGHGEmissions

modèle sur l'intensité des émissions (GES)

Intérêt de **ENERGYSTARScore**

➡ 2 modèles à comparer(avec ou sans la variable)

Problématique: Dataset

2 jeux de données (dataset benchmarking 2015 et 2016):

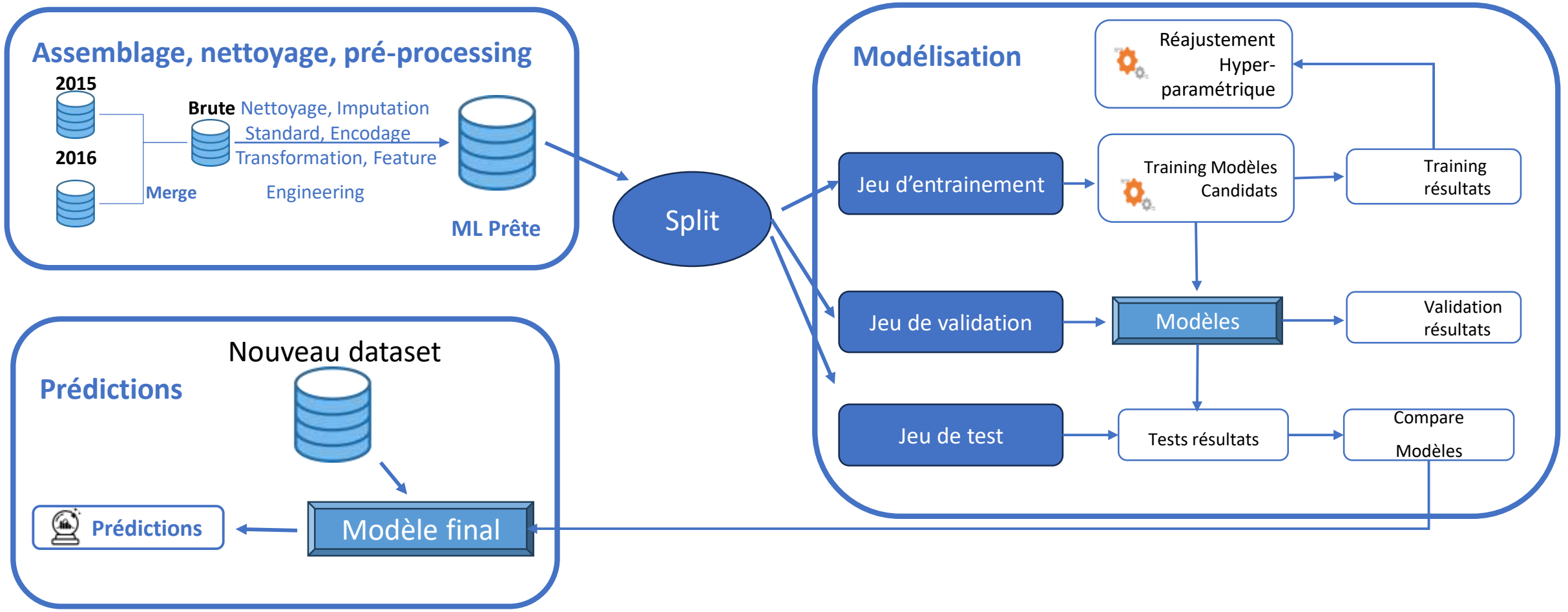
Informations: type, localisation, usage, consommation énergétique, émissions GES

		Données 2015	Données 2016
Lignes	**Taille**	**3340**	**3376**
Variables	**Nombre**	**47**	**46**
	Similaires	Location	Latitude, Longitude, Address, City, Zip Code, State
		GHGEmissions (MetricTonsCO2e)	TotalGHGEmissions
		GHGEmissionsIntensity(kgCO2e/ft2)	GHGEmissionsIntensity
		Comment	Comments
	Additionnels	OtherFuelUse(kBtu), SPD Beats, Seattle Police Department Micro Community Policing Plan Areas, City Council Districts, Zip Codes, 2010 Census Tracts	

Variables Cibles: SiteEnergyUse(kBtu), TotalGHGEmissions

EnergyStarScore: EnergyStarScore

Problématique: Cheminement



Sommaire

1. Problématique
2. **Données**
3. Modélisation
4. Conclusion

Données

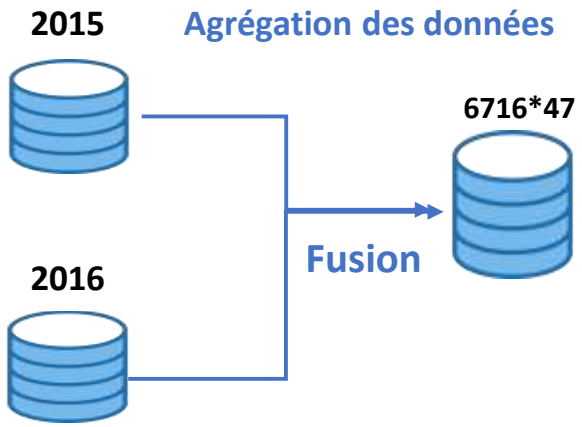
Métier	- Compréhension du métier - Groupement des données
Nettoyage	- Suppression des données inutiles, filtre - Valeurs manquantes, aberrantes - Harmonisation des caractères, doublons
Analyse	- Analyse univariée - Analyse bivariée/multivariée
Pré processing	- Feature engineering - Imputation - Types de variables, transformation variables cibles

Données - Groupement des données

Compréhension métier

3340	3376
47	**46**
Location	Latitude, Longitude, Address, City, Zip Code, State
GHGEmissions (MetricTonsCO2e)	TotalGHGEmissions
GHGEmissionsIntensity(kgCO2e/ft2)	GHGEmissionsIntensity
Comment	Comments
OtherFuelUse(kBtu), SPD Beats, Seattle Police Department Micro Community Policing Plan Areas, City Council Districts, Zip Codes, 2010 Census Tracts	

Agrégation des données



Avec 3284 doublons

Stratégie: 3511 bâtiments

+92 (2016)

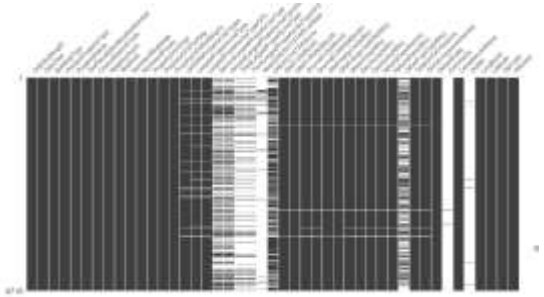
PARTIE FIXE 2016: 3284

+56 (2015)

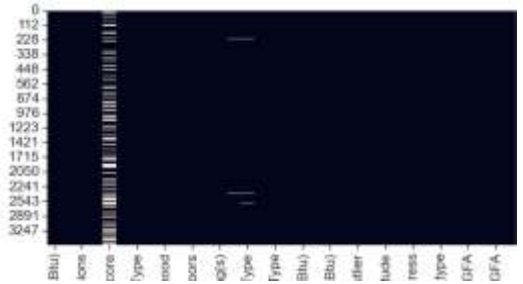
2015/2016
Var > 100%
+79

Données - Nettoyage

De 6716 lignes, 45 variables



A 1755 lignes, et 34 variables



- Fusion des données
- Suppression variables inutiles
- Homogénéisation du contenu des variables catégorielles: Minuscules, majuscules, Suppression des accents **->** Doublons
- Réduction des modalités des variables catégorielles
- Filtre des bâtiments (2016 + sup 2015 + var > 100 %)
- Filtre des bâtiments non résidentiels
- Gestion des valeurs manquantes - imputation
- Gestion des valeurs aberrantes
- Feature engineering
- Type (catégorielle en object, float/int 64 en 32)

Nb lignes	Nb Var
6716	**45**
6716	21
6716	21
6716	21
3511	21
1760	21
1755	21
1755	21
1755	**34**

Données – Nettoyage – Suppression des variables

Variables	Raison
City Council Districts, Zip Codes, 2010 Census Tracts, Seattle Police Department Micro Community Policing Plan Areas, SPD Beats	Seulement en 2015, abandonnées avant fusion
PropertyName, TaxParcelIdentificationNumber, OSEBuildingID, YearsENERGYSTARCertified, DefaultData, ComplianceStatus	Inutiles pour notre problématique ou trop de valeurs manquantes
Electricity(kWh), NaturalGas(therms)	Autres unités de mesure d'énergie
SiteEUI(kBtu/sf), SiteEUIWN(kBtu/sf), SourceEUI(kBtu/sf), SourceEUIWN(kBtu/sf)	Unités en fonction de la surface en pieds carrés
SiteEnergyUse(kBtu), GHGEmissionsIntensity	Redondantes avec les cibles
DataYear, YearBuilt, PropertyGFAParking, LargestPropertyUseTypeGFA, SecondLargestPropertyUseTypeGFA, ThirdLargestPropertyUseTypeGFA	Après Feature engineering
ListOfAllPropertyUseTypes, PropertyGFATotal	Après imputation
City (SEATTLE), State (WA)	1 seule valeur

INUTILES

DOUBLONS

PRE-PROCESSING

Données – Nettoyage - Modalités

Plusieurs modalités avec répétitions, espace en plus, minuscules, majuscules

→ groupement des modalités redondantes

	BuildingType	Neighborhood	PrimaryPropertyType	LargestPropertyUseType	SecondLargestPropertyUseType	ThirdLargestPropertyUseType
unique	6	13	28	57	50	45

Variable	Modalité	Aggrégation
BuildingType	6	6
Neighborhood	13	13
PrimaryPropertyType	28	10
LargestPropertyUseType	57	10
SecondLargestPropertyUseType	50	10
ThirdLargestPropertyUseType	45	10

Données – Nettoyage – Valeurs aberrantes

	NumberofBuildings	NumberofFloors	PropertyGFATotal	PropertyGFAParking	PropertyGFABuilding(s)	LargestPropertyUseTypeGFA	SecondLargestPropertyUseTypeGFA	ThirdLargestPropertyUseTypeGFA	TotalGHGEmissions	GHGEmissionsIntensity
type	int32	int32	int64	int64	int64	float64	float64	float64	float64	float64
nb_nan	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
%_nan	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
count	1755.0	1755.0	1755.0	1755.0	1755.0	1755.0	1755.0	1755.0	1755.0	1755.0
mean	1.193162	4.164103	120861.124217	13137.523077	107723.60114	100262.479772	19655.979087	3243.299942	186.911425	1.581242
std	2.852665	6.640411	297092.5252	43444.642055	282866.47618	273891.182304	56554.564849	18966.998388	749.272019	2.317105
min	1.0	0.0	11285.0	0.0	3636.0	5656.0	0.0	0.0	0.0	0.0
25%	1.0	1.0	29503.0	0.0	28496.0	25552.0	0.0	0.0	19.585	0.33
50%	1.0	2.0	49760.0	0.0	47673.0	44091.0	0.0	0.0	49.17	0.85
75%	1.0	4.0	107751.0	0.0	96124.5	92580.0	12770.5	0.0	141.575	1.87
max	111.0	99.0	9320156.0	512608.0	9320156.0	9320156.0	686750.0	459748.0	16870.98	34.09

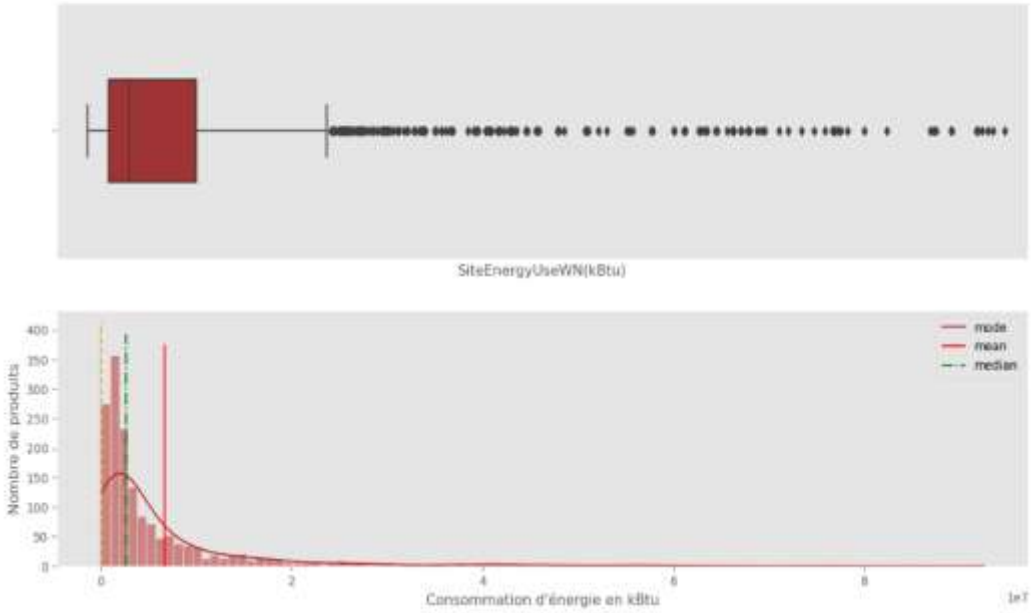
Aberrante

Outliers ?

Variable	Aberrante ?
NumberofFloors	A la main, 2 au lieu de 99 : église moderne de 2 étages maximum (google street), tour la plus haute 93 étages
TotalGHGEmissions, PropertyGFABuilding(s)	Max aberrant? Non, les bâtiments = hôpitaux, campus

Données – Analyse univariée

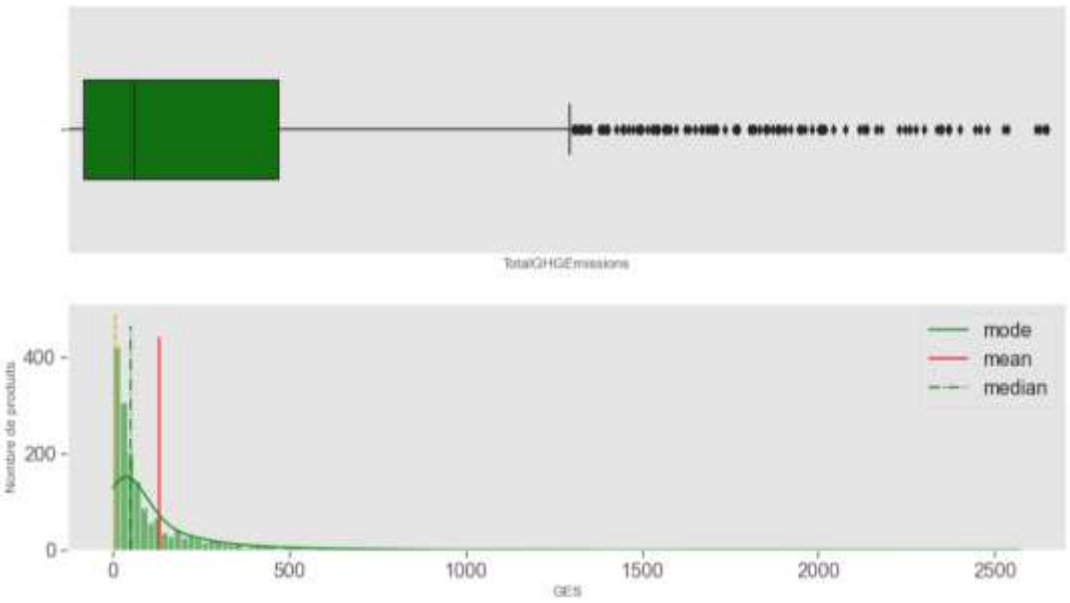
Consommation totale de l'énergie



Variable name	SiteEnergyUseWN(kBtu)
Mean	6641394.261947
Median	2652254.5
Skew	3.535695
Kurtosis	15.059948
Variance	116516924194899.578125
Stdev	10794300.54218
min	0.0
25%	1297661.6875
50%	2652254.5
75%	7020946.75
max	92537256.0

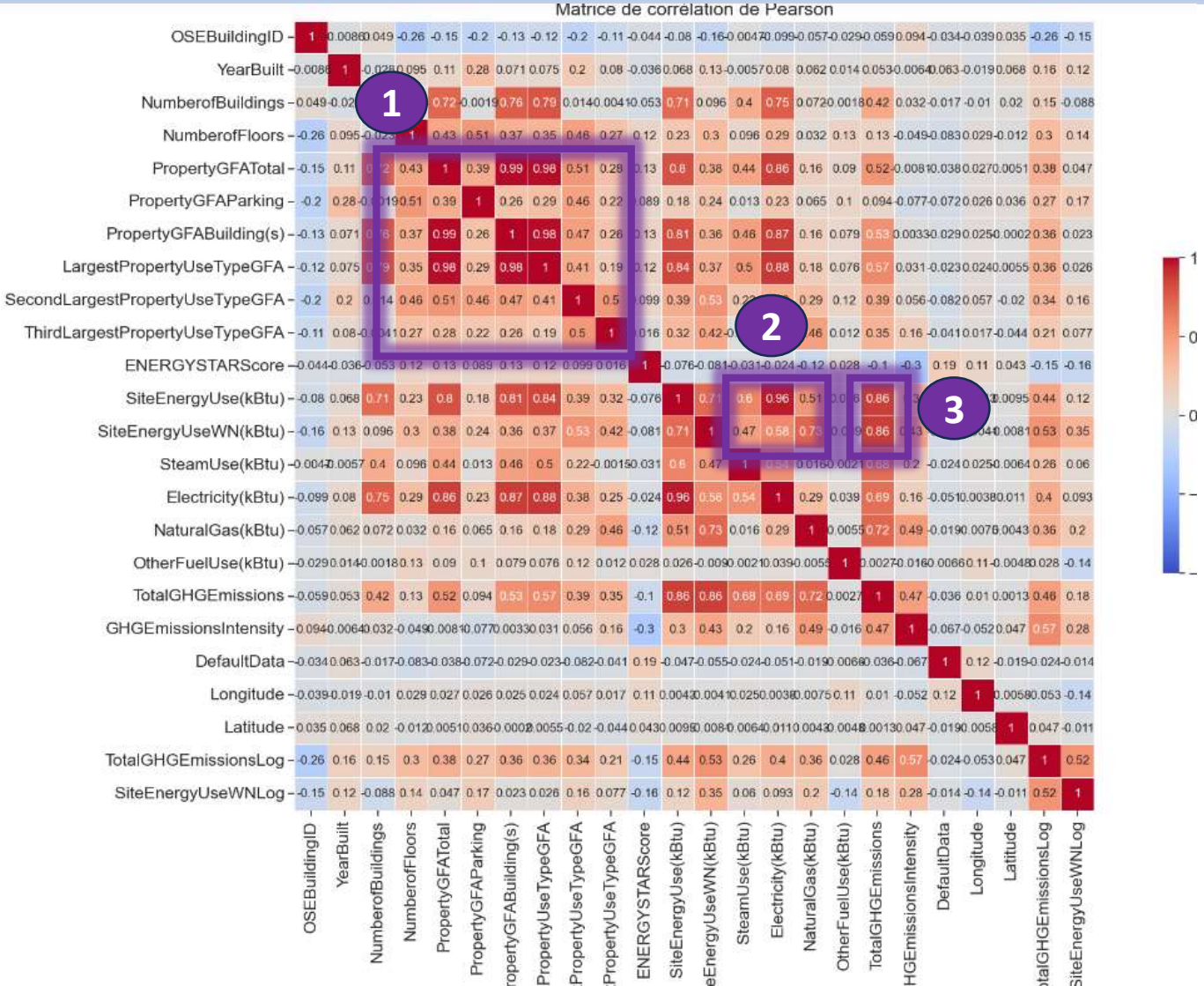
SKEWNESS > 1 ⇒ transformation logarithmique

Emission des gaz à effet de serre



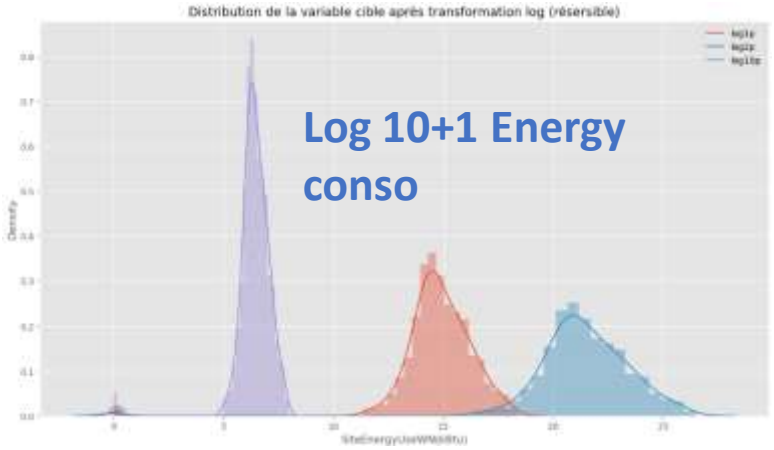
Variable name	TotalGHGEmissions
Mean	133.309405
Median	49.18
Skew	4.824349
Kurtosis	31.119229
Variance	61425.753221
Stdev	247.842194
min	0.5
25%	20.005
50%	49.18
75%	135.025
max	2573.75

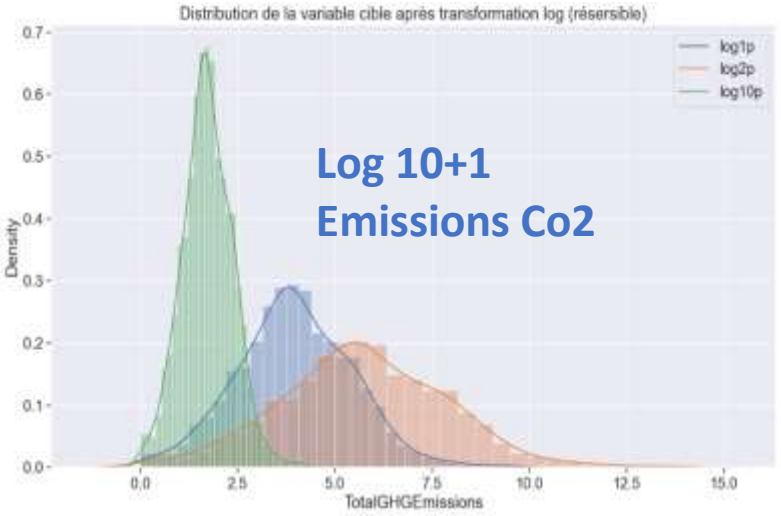
Données – Analyse multivariée



1. Features engineering(variables caractéristiques): nouvelles variables
2. Features Engineering : Seule information dans Le permis de construire : les sources d'énergie
3. Cibles fortement corrélées

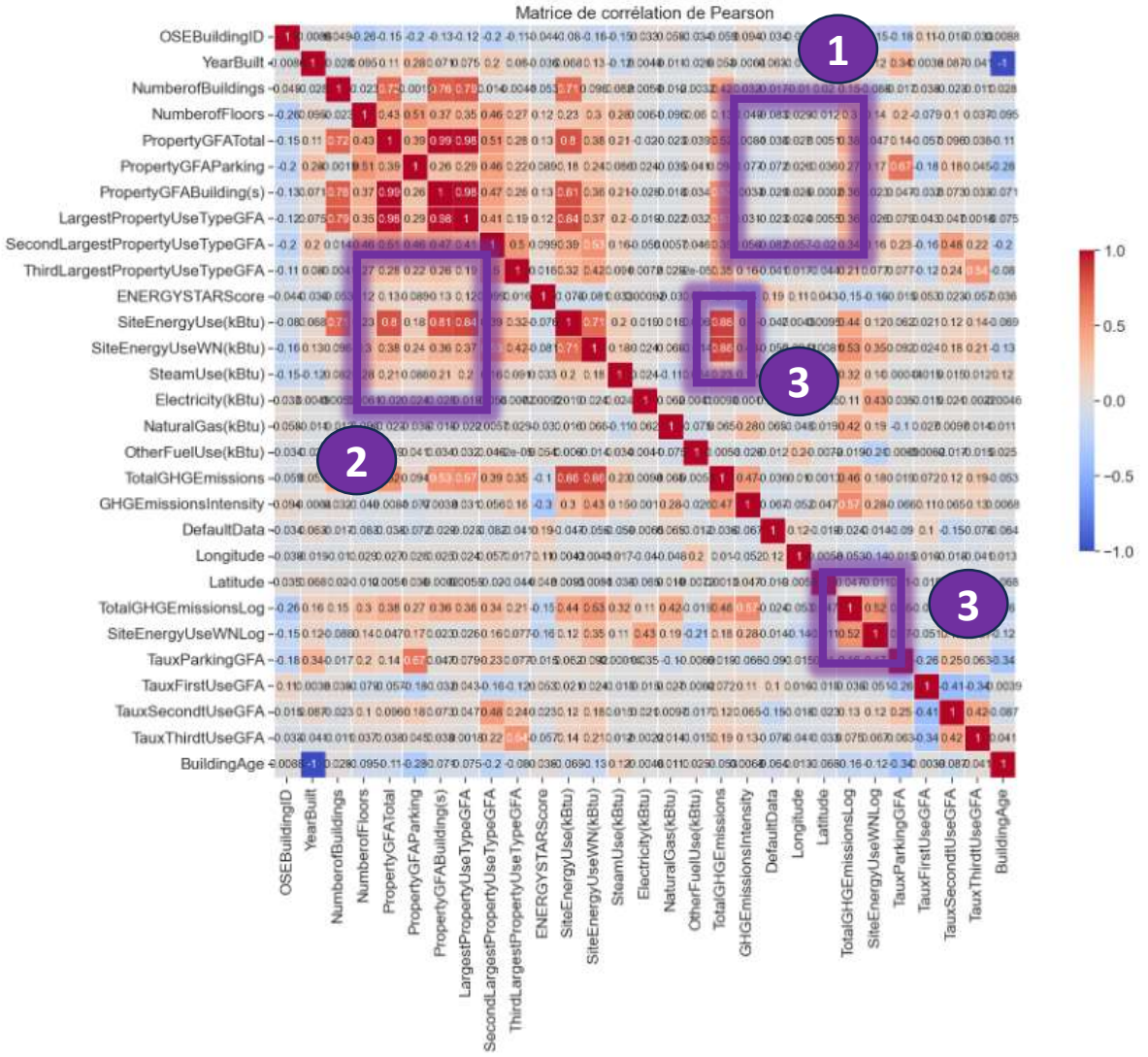
Données – Nettoyage – Feature Engineering





	Variables	Description
Localisation bâtiments	ListOfAllPropertyUseTypes	Compte le nombre de type de propriété pour chaque bâtiment
	Latitude/Longitude	Cartographie des bâtiments avec la transformation en variable binaire de la latitude et la longitude et en faisant la somme
	Address	Influence si le bâtiment est dans une rue, avenue, chemin? ➔ WAY, AVENUE ou STREET
Construction bâtiments	DataYear, YearBuilt	L'âge du bâtiment, ou rénovation
	PropertyGFAParking PropertyGFATotal LargestPropertyUseTypeGFA SecondLargestPropertyUseTypeGFA ThirdLargestPropertyUseTypeGFA	Ratio de la surface du parking sur la surface totale Ratio de la surface de la première (2nde , 3ième) sur la surface totale
Variables d'énergie, émission	SteamUse(kBtu), Electricity(kBtu), NaturalGas(kBtu), OtherFuelUse(kBtu)	0 : n'utilise pas cette énergie, 1 : utilise cette énergie.
	SiteEnergyUseWN(kBtu)	Transformation en log10 + 1
	TotalGHGEmissions	Transformation en log10 + 1

Données – Analyse multivariée



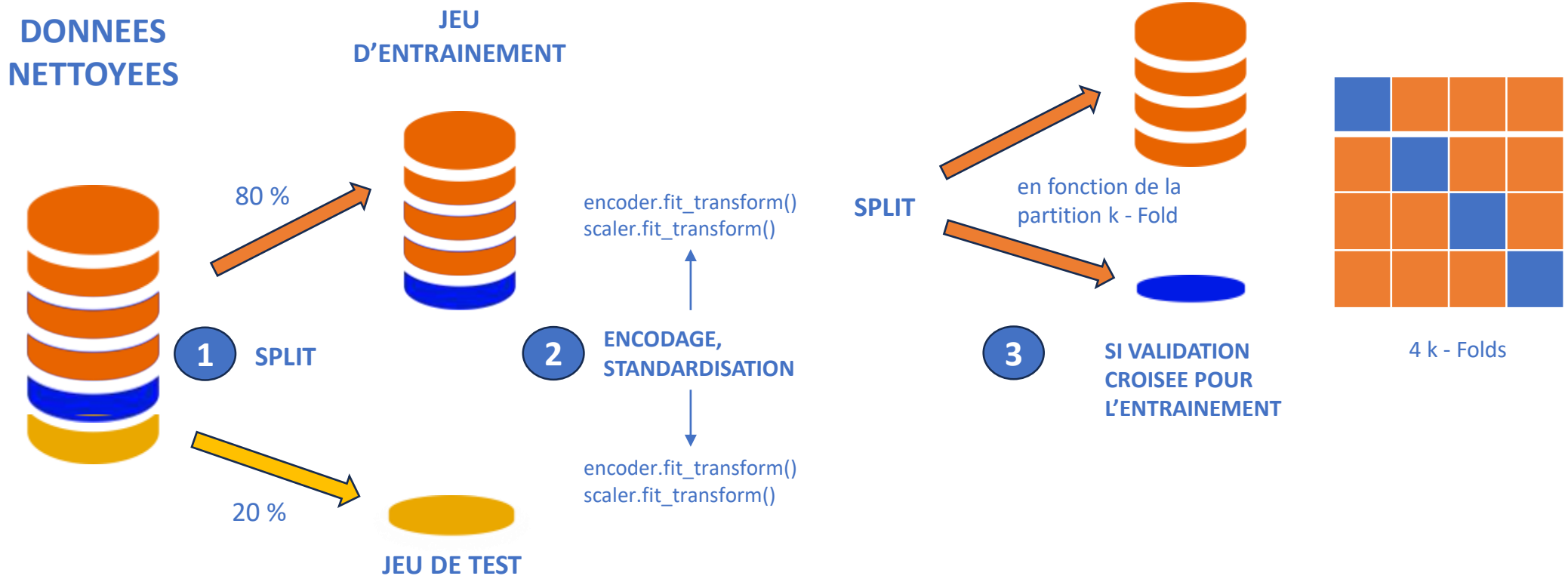
APRES FEATURE ENGINEERING

1. Corrélation fortement réduite
2. Corrélation fortement réduite
3. Variables cibles fortement corrélées

Sommaire

Modélisation – Consommation d'énergie

SPLIT - ENCODAGE/STANDARDISATION : préparation des données au Machine Learning

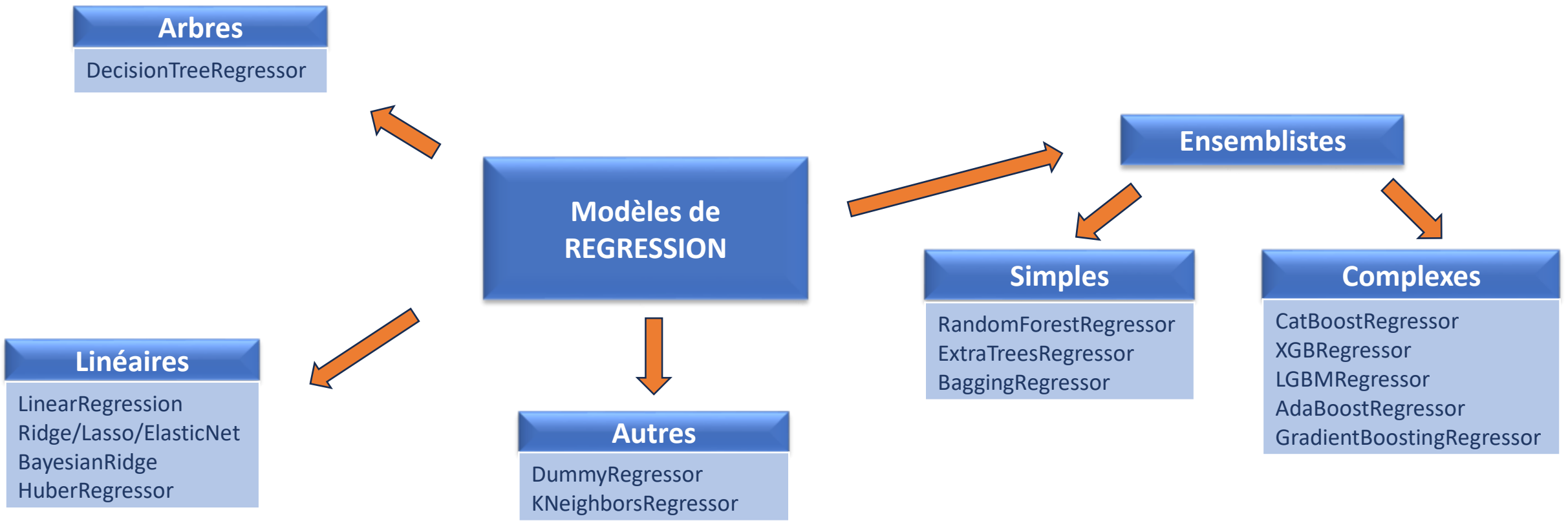


Variables catégorielles : **encodage** avec encoder = **TargetEncoder**
Variables numériques : **standardisation** avec **scaler**

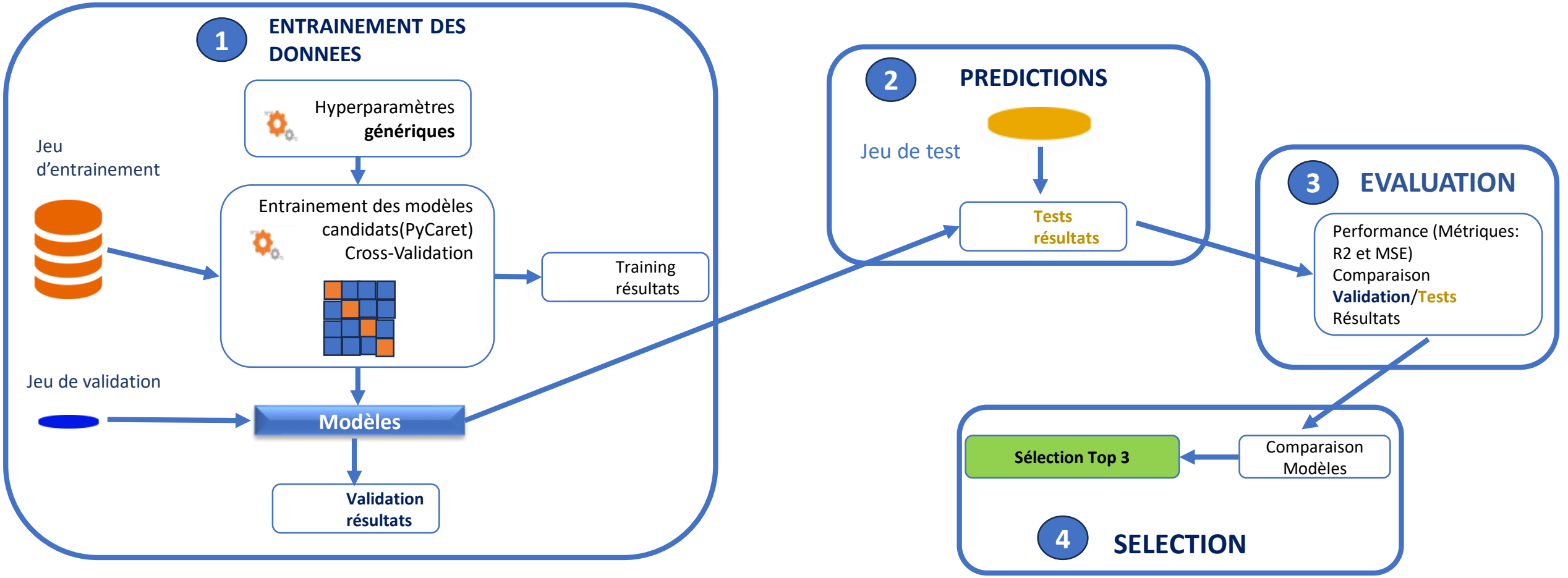
Modélisation: Consommation d'énergie

SELECTION MODELES DE BASE

Cible **SiteEnergyUseWNLog** numérique ➔ RÉGRESSION



Modélisation: Consommation d'énergie

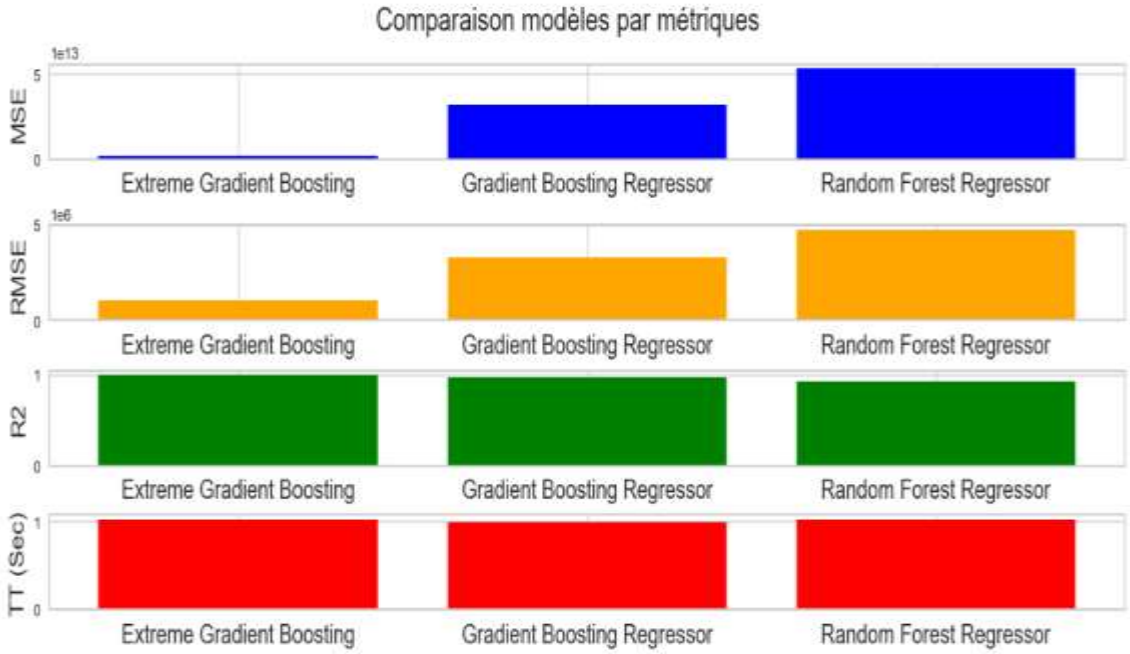


Modélisation: Consommation d'énergie

PERFORMANCES, COMPARAISON DES MODELES

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
xgboost	Extreme Gradient Boosting	220289.3820	1929822989721.6001	1034154.4109	0.9955	1.3366	0.0293	0.4410
gbr	Gradient Boosting Regressor	609367.7366	32359739096739.8203	3331225.3337	0.9644	1.4171	0.0373	0.4400
rf	Random Forest Regressor	820417.8483	53145886294504.9219	4764054.8494	0.9295	1.3441	0.0489	0.4620
dt	Decision Tree Regressor	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
ada	AdaBoost Regressor	2398919.2066	59910702950876.9688	5844971.7633	0.8950	2.1964	1.5781	0.4470
catboost	CatBoost Regressor	2218859.7665	134019141055222.4062	8492258.6326	0.7884	1.9670	0.6229	0.4330
lightgbm	Light Gradient Boosting Machine	1799968.1220	109183031268813.0312	8582860.6211	0.7637	1.8017	0.2214	0.4490
et	Extra Trees Regressor	3396315.9403	204422402961202.9688	11403368.1937	0.6021	2.0986	1.0506	0.4590
ridge	Ridge Regression	5747093.7419	288135218821448.4375	13624420.6367	0.3517	2.3914	2.8373	0.4290
lasso	Lasso Regression	5750456.8429	290540674030861.3750	13711650.2989	0.3344	2.3909	2.8016	0.4440
lr	Linear Regression	5751102.8227	290567845691725.3750	13713678.5757	0.3342	2.3907	2.8017	0.4420
llar	Lasso Least Angle Regression	5753785.1508	290664431482809.1875	13718633.1595	0.3338	2.3906	2.8055	0.4140
en	Elastic Net	6245949.5672	317950500188934.1875	14567768.0662	0.2556	2.4442	3.0668	0.4180
knn	K Neighbors Regressor	6790608.6000	299940478923571.1875	15510303.3500	0.0779	2.5045	3.0502	0.4460
dummy	Dummy Regressor	8123899.9000	377689665542553.6250	17172327.3500	-0.0537	2.6636	4.6825	0.4390
br	Bayesian Ridge	7888708.1283	426906948732541.8750	17722190.9561	-0.0683	2.5975	4.0410	0.4470
omp	Orthogonal Matching Pursuit	7699013.2870	439902019895552.3750	18023761.5261	-0.1031	2.5848	3.9237	0.4570
par	Passive Aggressive Regressor	8524914.0865	474191135432294.9375	19981970.6087	-0.6447	2.5923	3.3648	0.4210
huber	Huber Regressor	8490098.3932	563847279321870.2500	21191762.1566	-0.8906	2.5994	3.9311	0.4250

TOP 3



Modélisation: Consommation d'énergie

OPTIMISATION DES MODELES

Extreme Gradient Boosting

Gradient Boosting Regressor

Random Forest

Modèles à optimiser

OPTIMISATION

Hyperparamètres

TUNING

Sélection manuelle

Recherche Automatique

Randomized Search CV

Grid Search CV

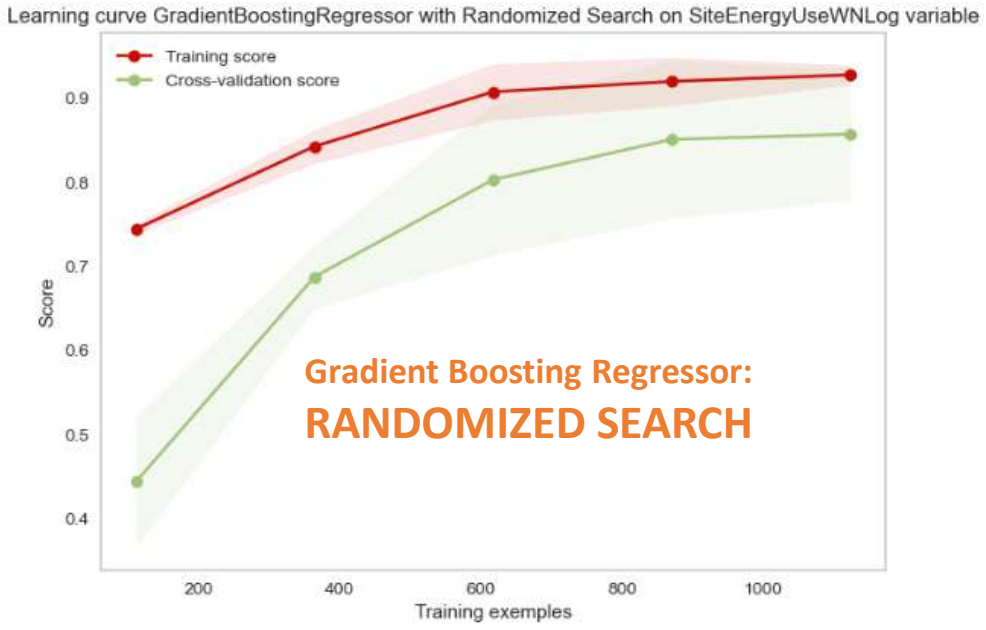
Modélisation: Consommation d'énergie

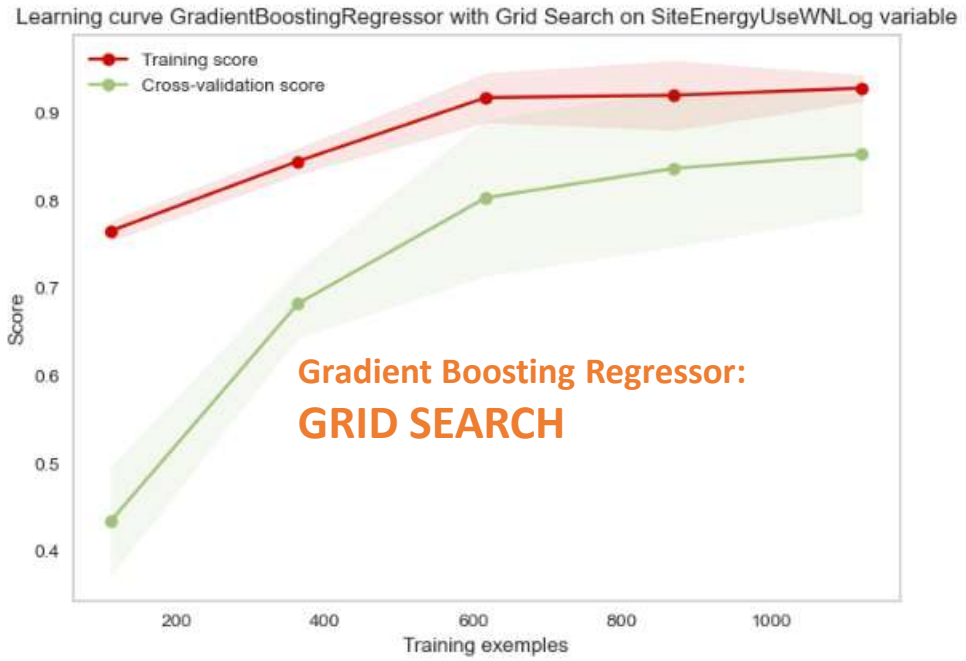
REGLAGE DES HYPERPARAMETRES

Modèle	Hyperparamètre	Défaut	Grille de recherche	Performances
Extreme Gradient Boosting	n_estimators	100	[110,130]	**130**
	max_depth	None	[0, 2, 4, 6, 8, 10, 12, 14]	**2**
	learning_rate	Auto	[0.001, 0.01, 0.03, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5]	**0.2**
Random Forest	n_estimators	100	[110,130]	**130**
	max_depth	None	[2, 4, 6, 8, 10, 12, 14]	**8**
	max_features	Auto	[2, 4, 6, 8, 10, 12]	**2**
Gradient Boosting Regressor	n_estimators	100	[110,130]	**110**
	min_samples_split	2	[0, 2, 4, 6, 8, 10, 12, 14]	**6**
	min_samples_leaf	1	[1, 2, 3, 4, 5, 6, 10]	**2**
	max_depth	3	[2, 4, 6, 8, 10, 12, 14]	**2**
	learning_rate	0,1	[0.001, 0.01, 0.03, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5]	**0.05**

Modélisation: Consommation d'énergie

PERFORMANCE – COMPARAISON LEARNING CURVE



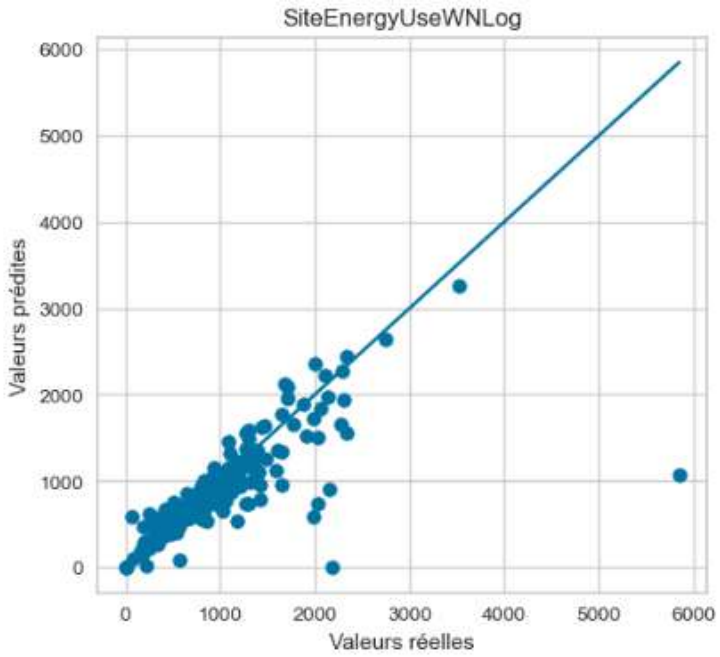


Meilleur Résultat: Méthode du Grid Search avec le GradientBoostingRegressor

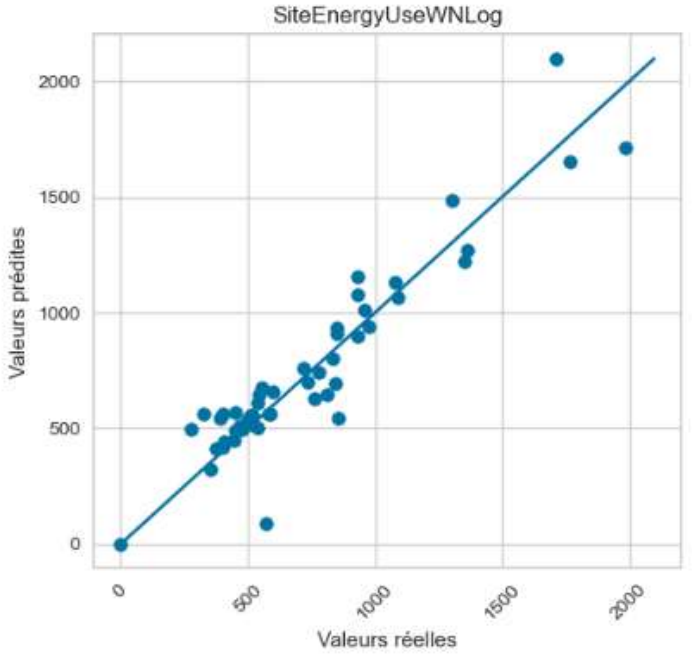
Modélisation: Consommation d’énergie

PRÉDICTIONS du modèle FINAL

Visualisation des erreurs

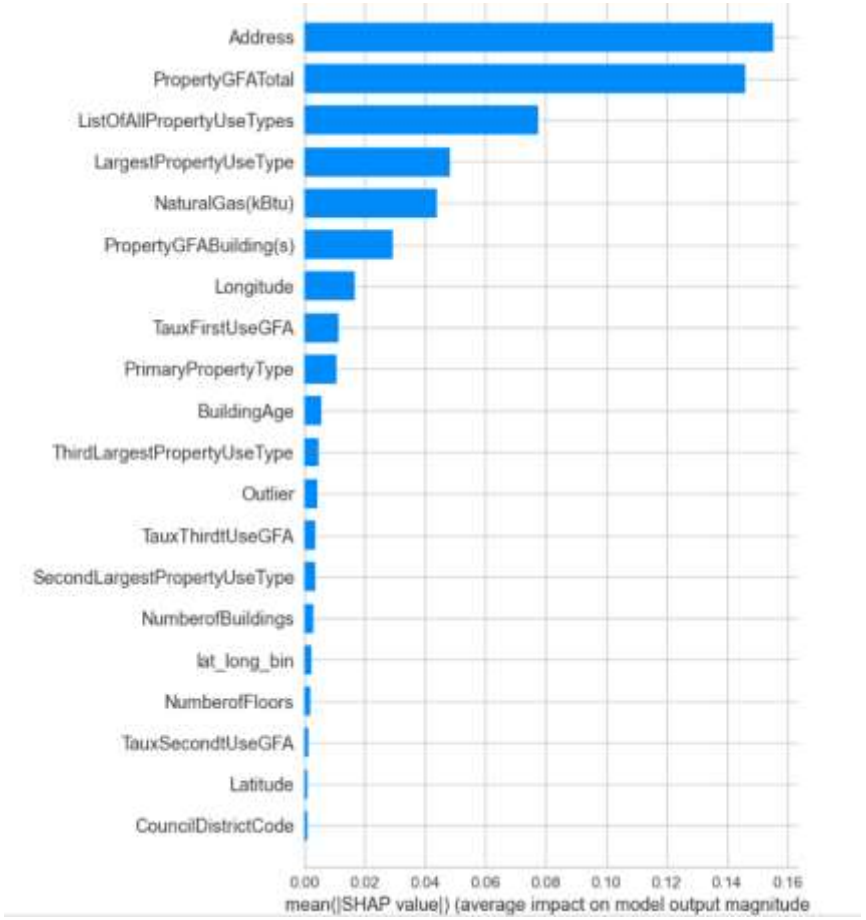


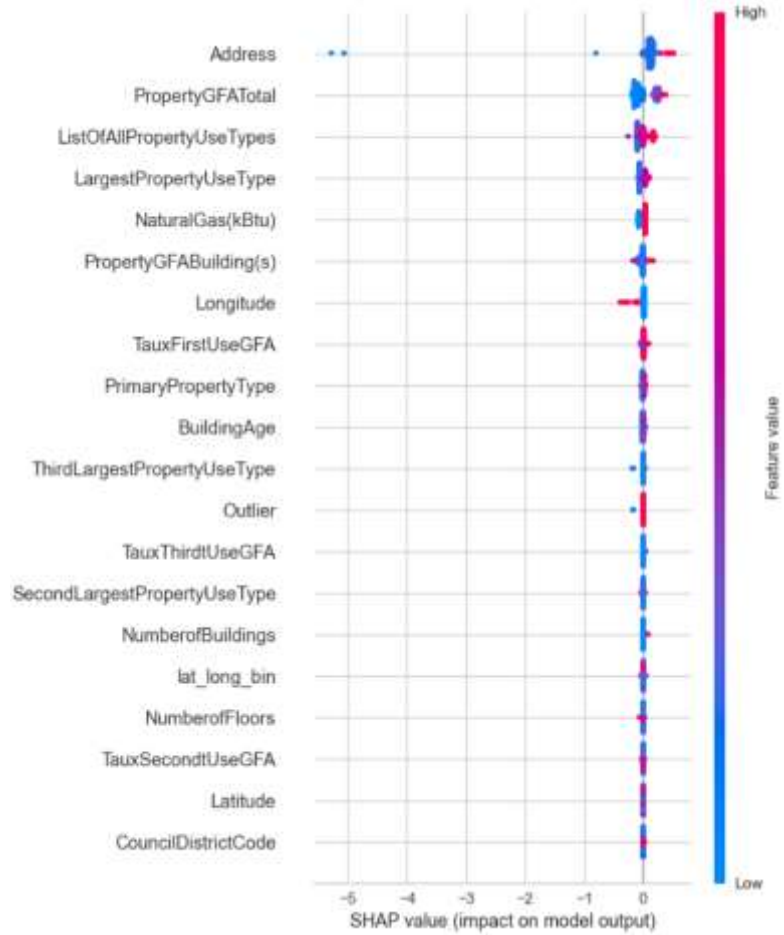
Prédiction de la variable cible



Modélisation: Consommation d’énergie

FEATURES IMPORTANCE

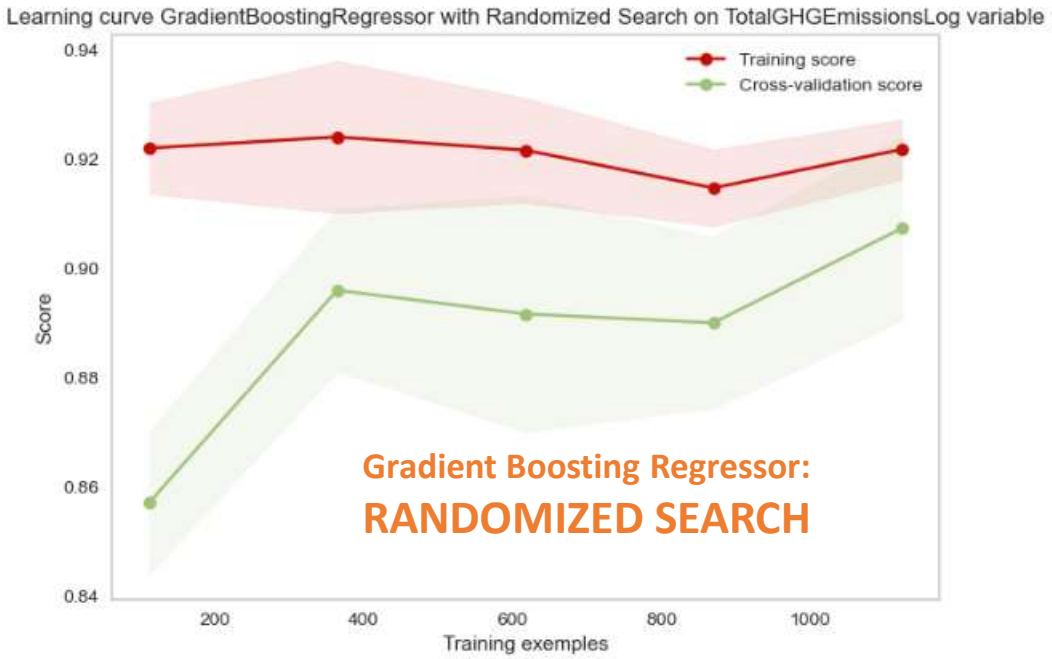


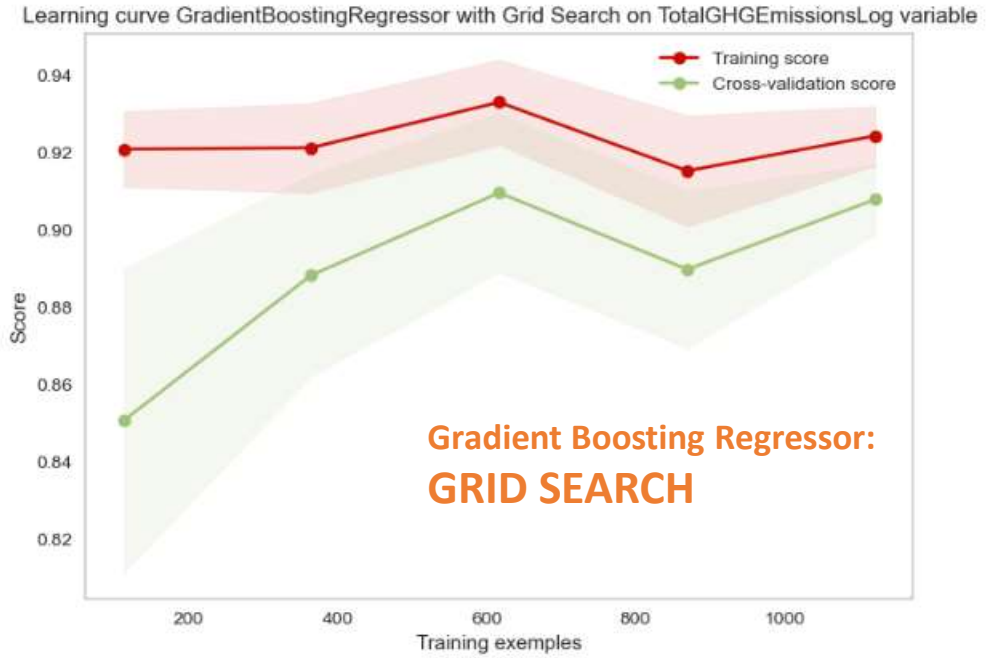


Sommaire

Modélisation – Émissions de CO2

PERFORMANCE – COMPARAISON LEARNING CURVE



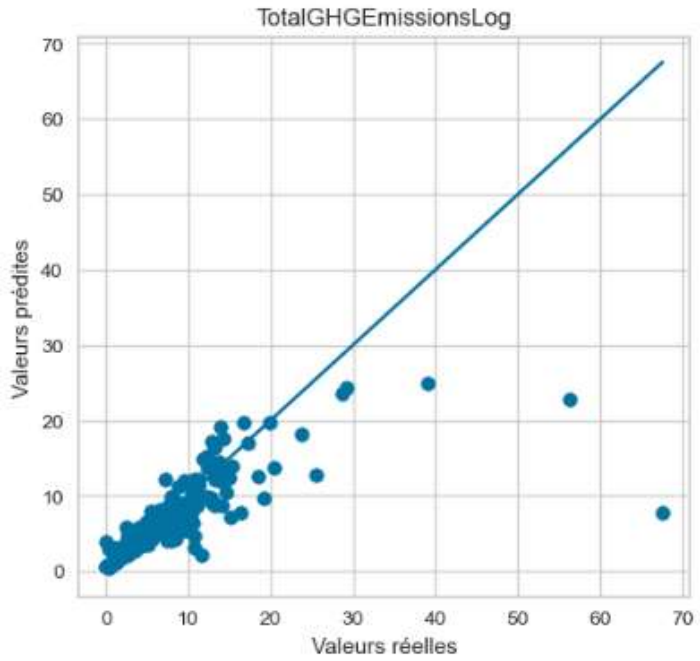


Meilleur Résultat: Méthode du Grid Search avec le GradientBoostingRegressor

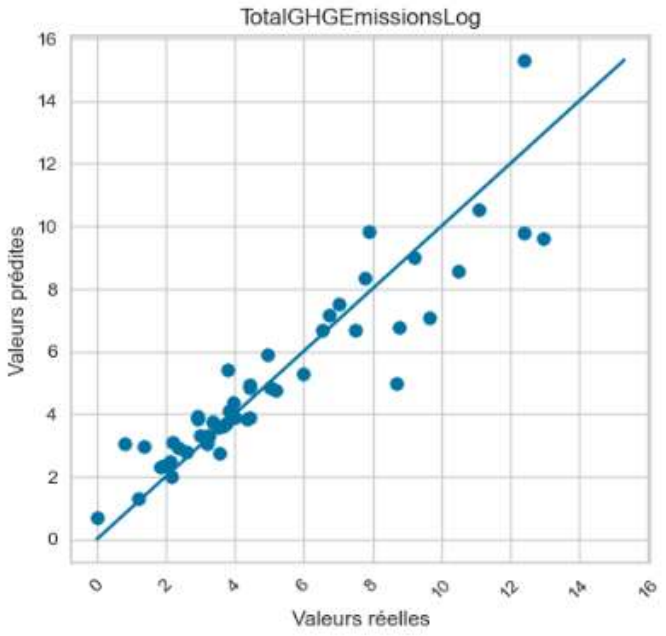
Modélisation – Émissions de CO2

PRÉDICTIONS du modèle FINAL

Visualisation des erreurs

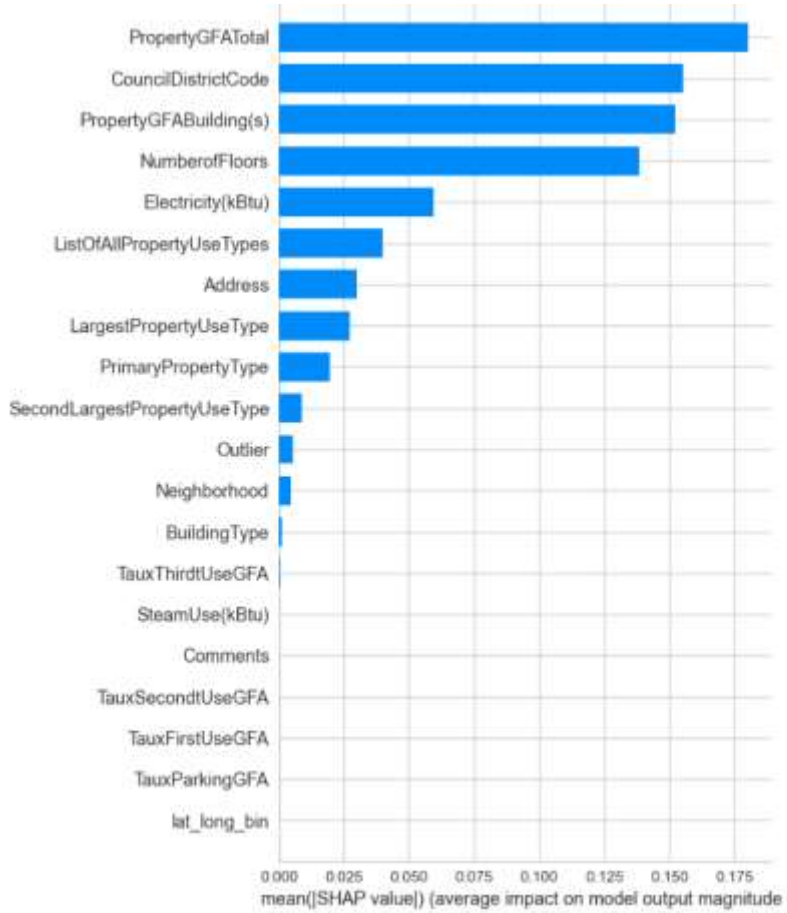


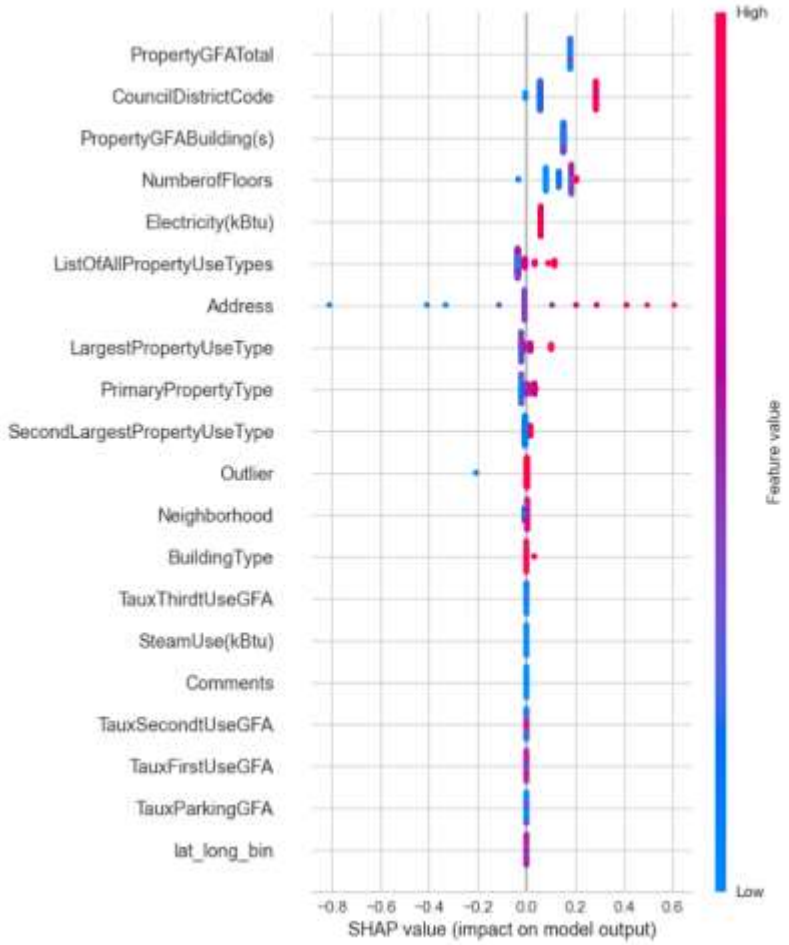
Prédiction de la variable cible



Modélisation – Émissions de CO2

FEATURES IMPORTANCE





Sommaire

1. Problématique
2. Données
3. **Modélisation** **Intérêt de l'EnergyStar sur l'émission de CO2**
4. Conclusion

Modélisation – Intérêt

COMPARAISON

Avec ou sans **ENERGYSTARScore:**
Même démarche : split, encodage, standardisation
Optimisation d'un modèle GradientBoostingRegressor sans la variable **ENERGYSTARScore** et un autre avec **ENERGYSTARScore**

Modèle	R2	MSE	RMSE	MAE	Erreur moy	Précision	Durée	Test R2 CV	Test R2 +/-	Test MSE CV	Train R2 CV
Gradient_Boosting_optimise_final_2(Randomized Search)	0.475841	0.250325	0.500325	0.381322	0.381322	-inf	0.128774	0.905710	0.019746	0.037569	0.921891
Gradient_Boosting_optimise_final_2(Grid Search)	0.412508	0.280572	0.529690	0.405647	0.405647	-inf	0.121254	0.900072	0.020114	0.039768	0.918094
Gradient_Boosting_optimise_final_2(Randomized Search)_avec_ENERGYStarScore	0.434943	0.269857	0.519478	0.396407	0.396407	-inf	0.115091	0.894939	0.021911	0.042238	0.910098
Gradient_Boosting_optimise_final_2(Grid Search)_avec_ENERGYStarScore	0.417421	0.278225	0.527471	0.402169	0.402169	-inf	0.121446	0.908187	0.011007	0.036551	0.921421

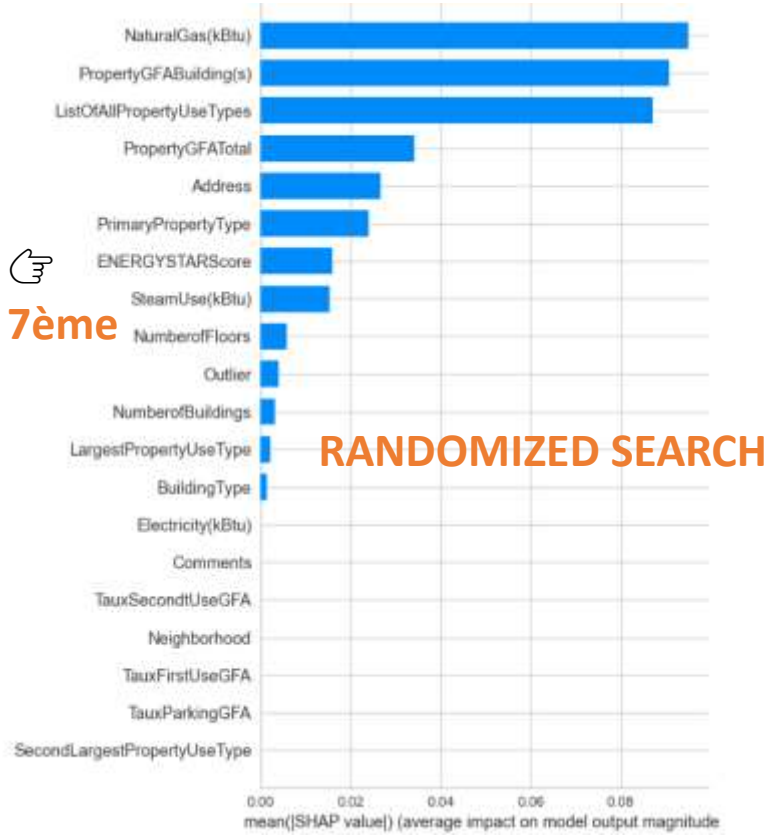
GRID Search -> Hausse: 0,0081 % R2 Test
Baisse de : 0,003 % MSE

Randomized Search -> Baisse: 0,01 % R2 Test
Hausse de : 0,019 % MSE

ENERGYSTARScore n' améliore pas significativement les 2 méthodes
Plus coûteuse en termes de temps ➔
Arbitrage à faire

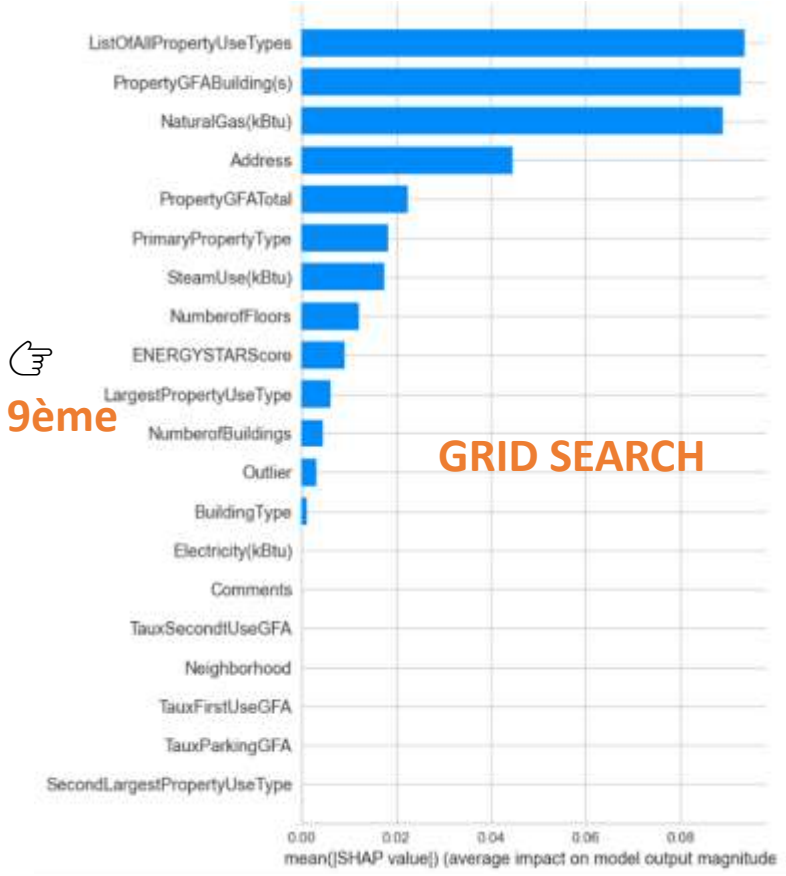
Modélisation – Intérêt

FEATURES IMPORTANCE



7ème

RANDOMIZED SEARCH



9ème

GRID SEARCH

Contribution plus importante de la variable **ENERGYSTARScore** sur la RANDOMIZED SEARCH que la GRID SEARCH

Sommaire

1. Problématique
2. Données
3. Modélisation
4. **Conclusion**

Idées d’amélioration

1. **Dataset**

   Discussion de la problématique au client: - récolte des données sur internet

   - arbitrage ‘EnergyStar score’

2. **Modélisation**

   2 modèles ?

   ACP : en utilisant moins de composantes?, Amélioration sur les features engineering, performances modèles

   Tester avec les réseaux de neurones?