Ville de Seattle

Anticipation des besoins en consommation de bâtiments

Présenté par Thiery KAPPE

Le contexte

Dans son objectif de devenir une ville neutre en émission de carbone à l'horizon 2050, la ville de Seattle souhaite à partir des ***données structurelles*** *des* ***bâtiments non destinés à l'habitation****, prédire les* ***émissions de CO2*** *et la* ***consommation totale d'énergie*** *de ces bâtiments.*

*Pour la prédiction des émissions de CO2, la ville souhaite également évaluer l'****intérêt de l'ENERGY STAR Score*** *qui est la mesure de la performance énergétique des bâtiments.*

Nos travaux consisterons à explorer et analyser les données disponibles, à identifier et tester différents modèles de prédiction afin de répondre au mieux à la problématique d'anticipation.

NB : *Les travaux tels que présentés ici sont plutôt linéaires afin de faciliter la présentation. Dans la pratique, les étapes étaient plutôt itératives avec des essais et des allers- retours.*

Cette présentation est soutenue par 3 notebooks : un notebook d'exploration et un notebook pour chacune des prédictions.

Vue globale des étapes clés

01 → 02 → 03 → 04 → 05

Etape 1

Importation des données,

Filtrage, sélection des bâtiments pertinents pour notre étude,

Exploration et visualisation des données.

Etape 2

Analyses exploratoires (valeurs aberrantes, traitement des valeurs manquantes, création de nouvelles variables, sélection des variables)

Etape 3

Analyses univariées et bivariées,
Préparation des données (encodage, suppression variables non structurelles, visualisation des distributions, analyse des corrélations).

Etape 4

Modélisation consommation
(Split train/test,
Standardisation,
Choix des modèles à tester,
Métriques d'évaluation,
Optimisation hyperparamètres,
Calcul des performances,
Sélection du modèle à retenir,
Feature importance,
Interprétabilité globale et locale.)

Etape 5

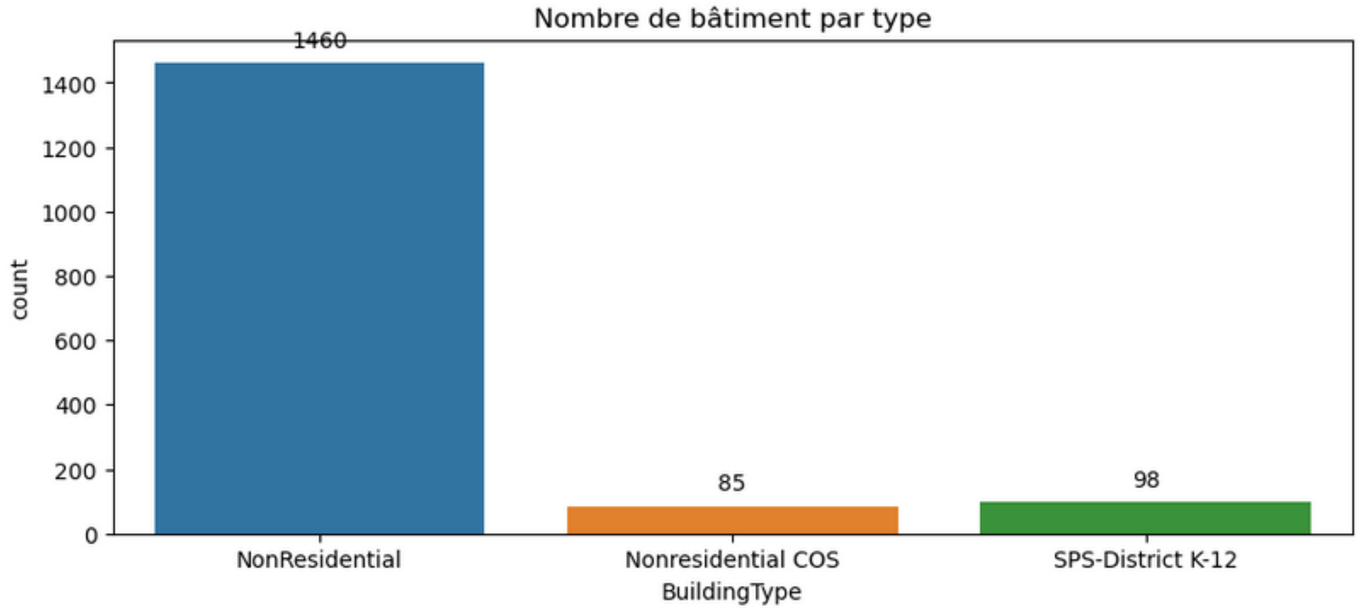
Modélisation émission de CO2
(Split train/test,
Standardisation,
Choix des modèles à tester,
Métriques d'évaluation,
Optimisation hyperparamètres,
Calcul des performances,
Sélection du modèle à retenir,
Feature importance,
Interprétabilité globale et locale.)
Pertinence de l'ENERGYSTARScore.

01

Importation des données.
Filtrage, sélection des bâtiments pertinents pour notre étude
Exploration et visualisation des données

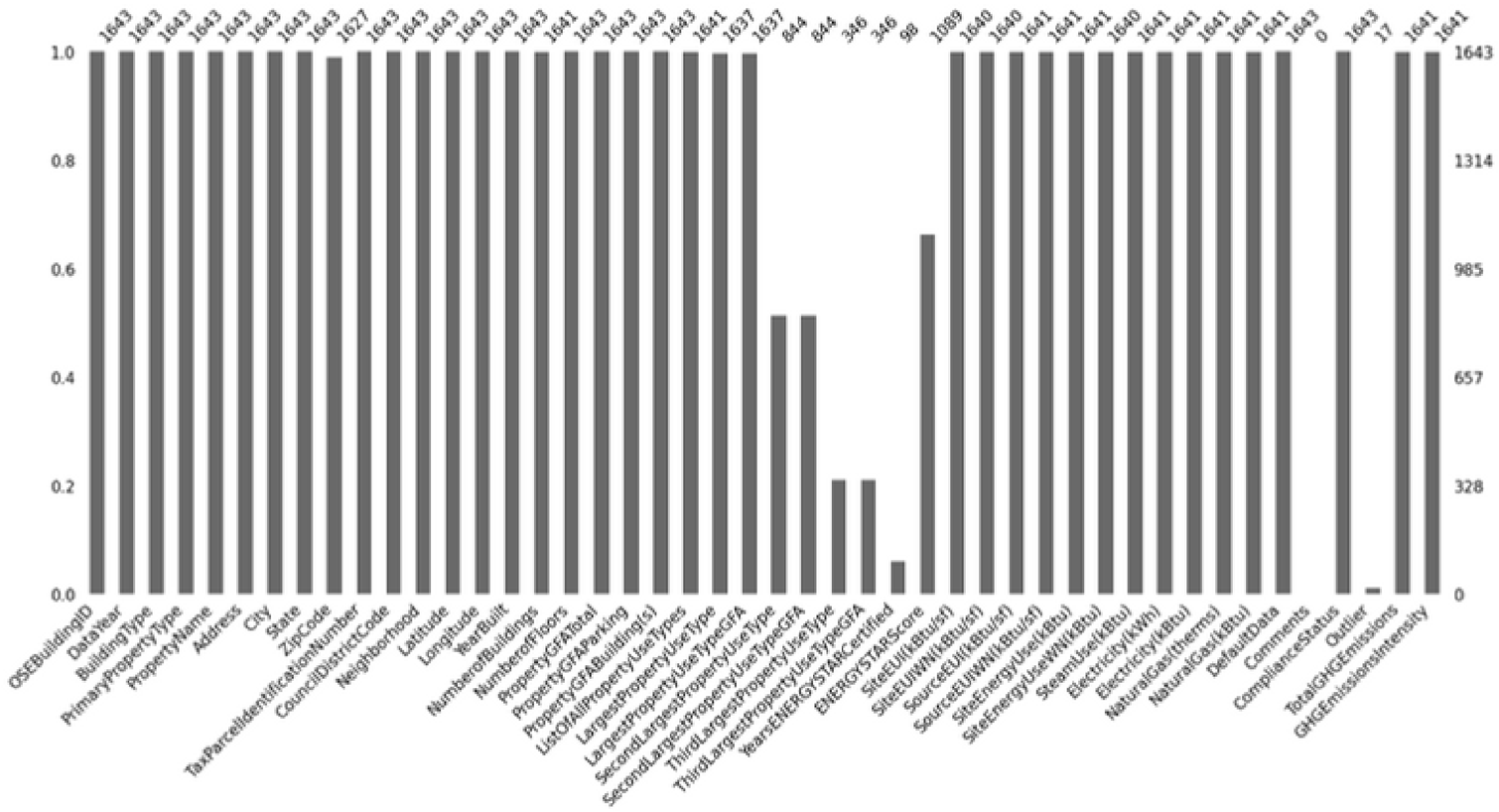
- *Après importation des données, nous avons un data frame 3376 bâtiments chacun décrit par 46 variables*
- *La variable ‘BuildingType’ nous permet de filtrer les données pour ne retenir que les bâtiments non destinés à l’habitation objet de notre étude.*
- *Les bâtiments de type ‘Campus’ présentent un caractère atypique. Ceux-ci sont en effet de regroupements de multiples bâtiments dont la modélisation ne serait pas aisée. Nous les supprimons des individus à analyser.*
- *Absence de doublons dans la base de données (aucun bâtiment n’est présent plus d’une fois)*

- ***Visualisation après filtrage***



02 *Analyses exploratoires (valeurs aberrantes, traitement des valeurs manquantes, création nouvelles variables)*

- ***Etat des valeurs manquantes par variable***



- ***Traitement des valeurs manquantes***

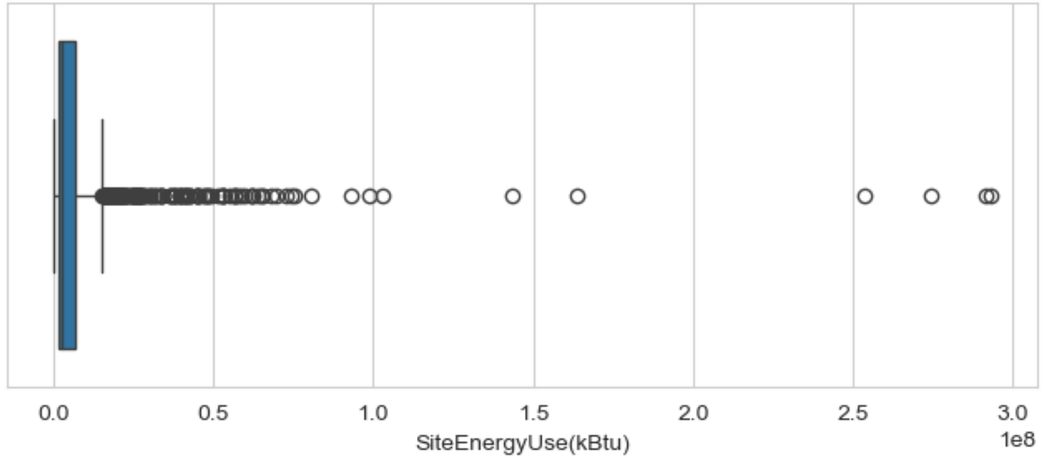
Variables	***Traitement apporté à la valeur manquante identifiée dans la variable***
NumberofBuildings	*Suppression*
LargestPropertyUseType	*Suppression*
SecondLargestPropertyUseTypeGFA	*Imputation à 0*
ThirdLargestPropertyUseTypeGFA	*Imputation à 0*
ENERGYSTARScore	*Imputation via KNN, moyenne des 20 plus proches voisins.*

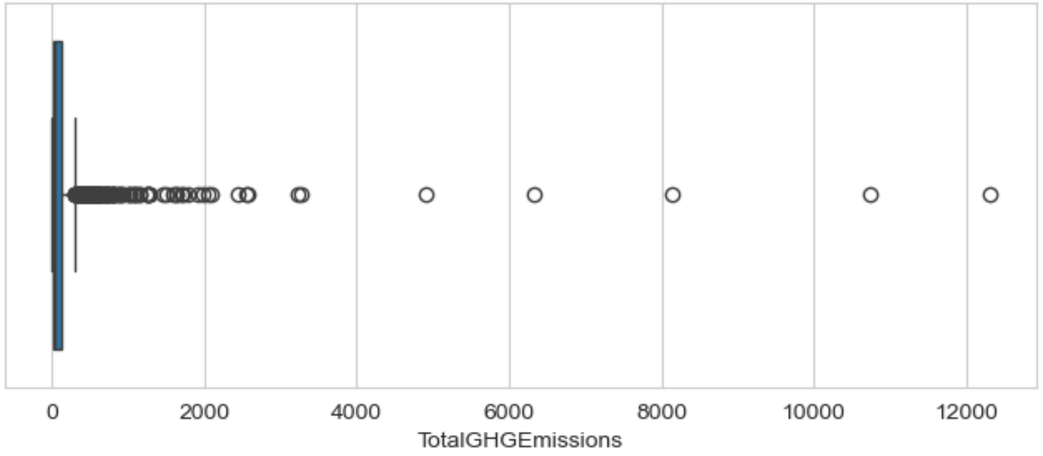
- ***Création de nouvelles variables***

Nouvelles variables créées	***Traitement effectué***
BuildingAge	*Utilisation avec la variable 'YearBuilt' et DataYear pour la création de la nouvelle variable 'BuildingAge'.*
SteamUse_ratio	*Proportion d'utilisation 'SteamUse (kBtu)' sur la quantité totale d'énergie consommée du site 'SiteEnergyUse(kBtu)'.*
Electricity_ratio	*Proportion d'utilisation 'Electricity(kBtu)' sur la quantité totale d'énergie consommée du site 'SiteEnergyUse(kBtu)'.*
NaturalGas_ratio	*Proportion d'utilisation 'NaturalGas(kBtu)' sur la quantité totale d'énergie consommée du site 'SiteEnergyUse(kBtu)'.*

02 *Analyses exploratoires (valeurs aberrantes, traitement des valeurs manquantes, création nouvelles variables)*

- ***Visualisation des outliers***





- ***Traitement des valeurs atypiques***

Outliers	***traitement effectué***
TotalGHGEmissions < 0	*Suppression des bâtiments concernés*
TotalGHGEmissions, SiteEnergyUse(kBtu)	*Suppression de l'ensemble des bâtiments dont la valeur absolue du Z_score est supérieure à 2. Ce sont des valeurs très atypiques qui peuvent pertuber la modélisation.*

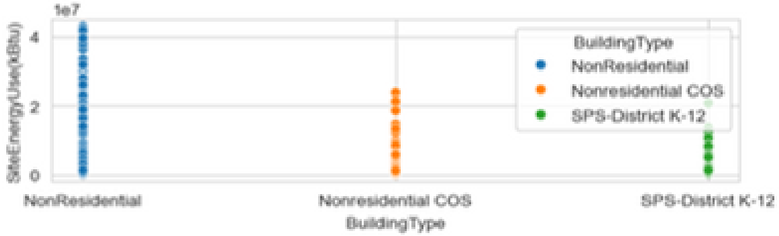
03 *Analyses univariées et bivariées, Préparation des données (encodage, suppression variables non structurelles, visualisation des distributions, analyse des corrélations).*

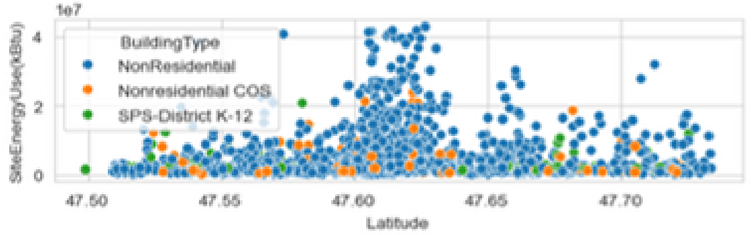
- ***Suppression de variables non structurelles***

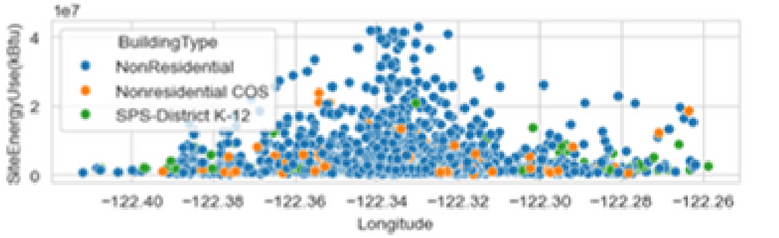
Variables non structurelles
PrimaryPropertyType
PropertyName, Address
City, State
TaxParcelIdentificationNumber
CouncilDistrictCode
Neighborhood, ZipCode
ListOfAllPropertyUseTypes
SecondLargestPropertyUseType
ThirdLargestPropertyUseType
YearsENERGYSTARCertified
Comments, SiteEUI(kBtu/sf)
SiteEUIWN(kBtu/sf)
SourceEUI(kBtu/sf)
SourceEUIWN(kBtu/sf)
SiteEnergyUseWN(kBtu)
SteamUse(kBtu)
Electricity(kWh)
Electricity(kBtu)
NaturalGas(therms)
NaturalGas(kBtu), DefaultData
ComplianceStatus, Outlier
GHGEmissionsIntensity

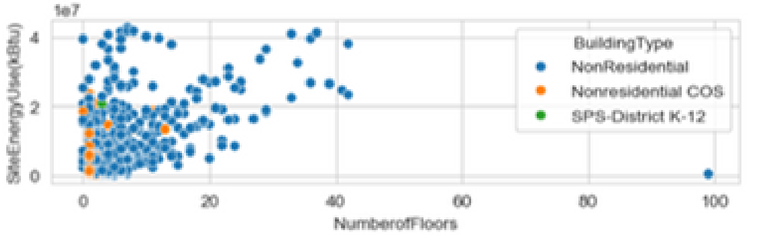
Variables non structurelles supprimées car sans intérêt pour notre analyse

- ***Analyse des relations entre la variable cible 'SiteEnergyUse (ktBu)' et les variables explicatives***





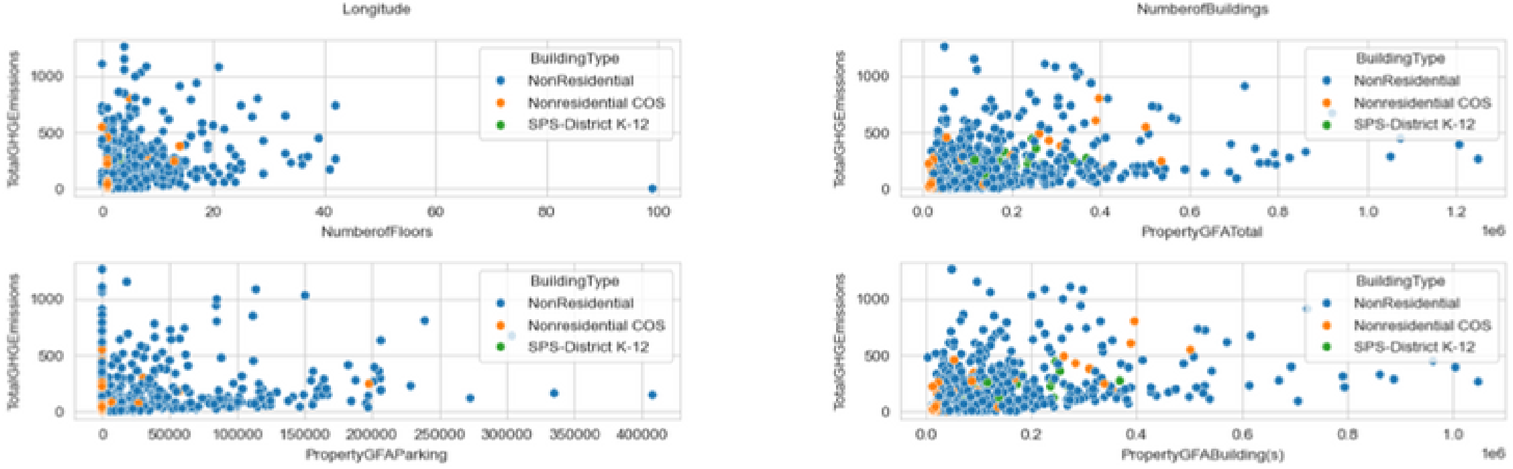




03

Analyses univariées et bivariées, Préparation des données (encodage, suppression variables non structurelles, visualisation des distributions, analyse des corrélations).

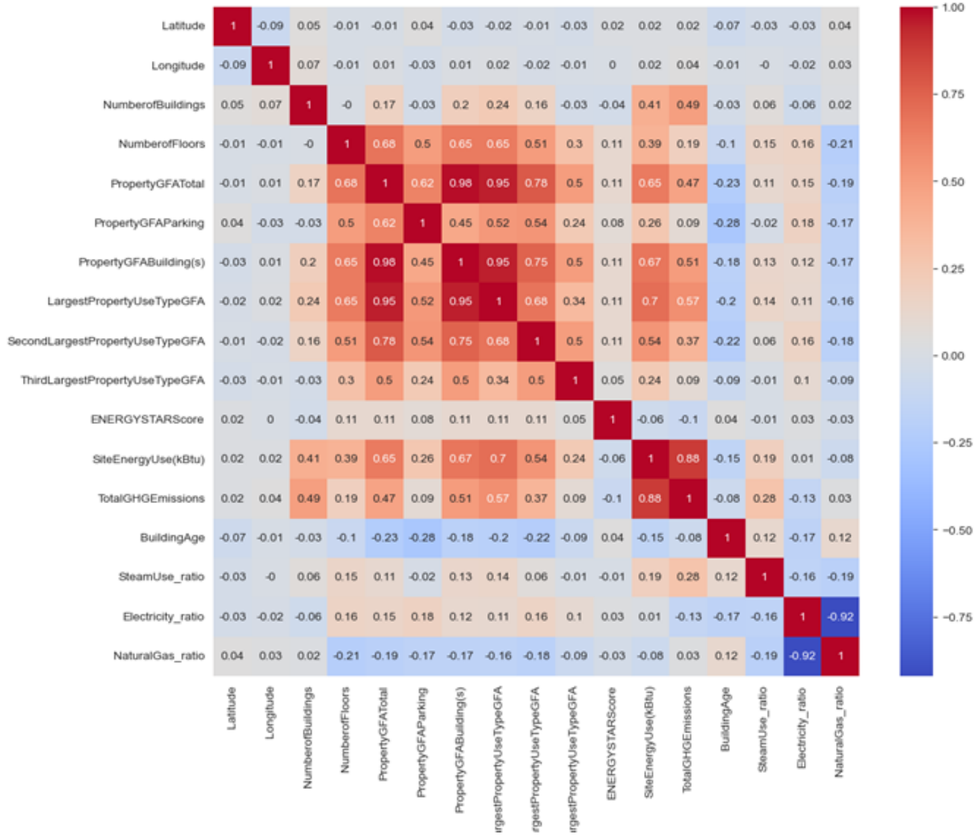
- ***Analyse des relations entre la variable cible 'TotalGHGEmissions' et les variables explicatives***



Variables catégorielles encodées	***Méthodologie utilisée***
LargestPropertyUseType	*One Hot Encoding avec suppression de la première modalité et ensuite suppression de la variable initiale.*
BuildingType	*One Hot Encoding avec suppression de la première modalité et ensuite suppression de la variable initiale.*

Variables fortement correlées entre elles	***Traitement effectué***
PropertyGFATotal, PropertyGFABuilding(s) LargestPropertyUseTypeGFA	*Existence d'une très forte corrélation positive >0.8 ==> Supression des variables PropertyGFATotal et LargestPropertyUseTypeGFA afin d'éviter toute collinéarité*
Electricity_ratio, NaturalGas_ratio	*Existence d'une très fortecorrélation négative de -0.92 ==> Suppression de la variable NaturalGas_ratio pour éviter toute collinéarité.*
TotalGHGEmissions, SiteEnergyUse(kBtu)	*Forte corrélation positive de 0.88, ce sont des variables cibles qui ne seront pas utilisées comme variables explicatives.*

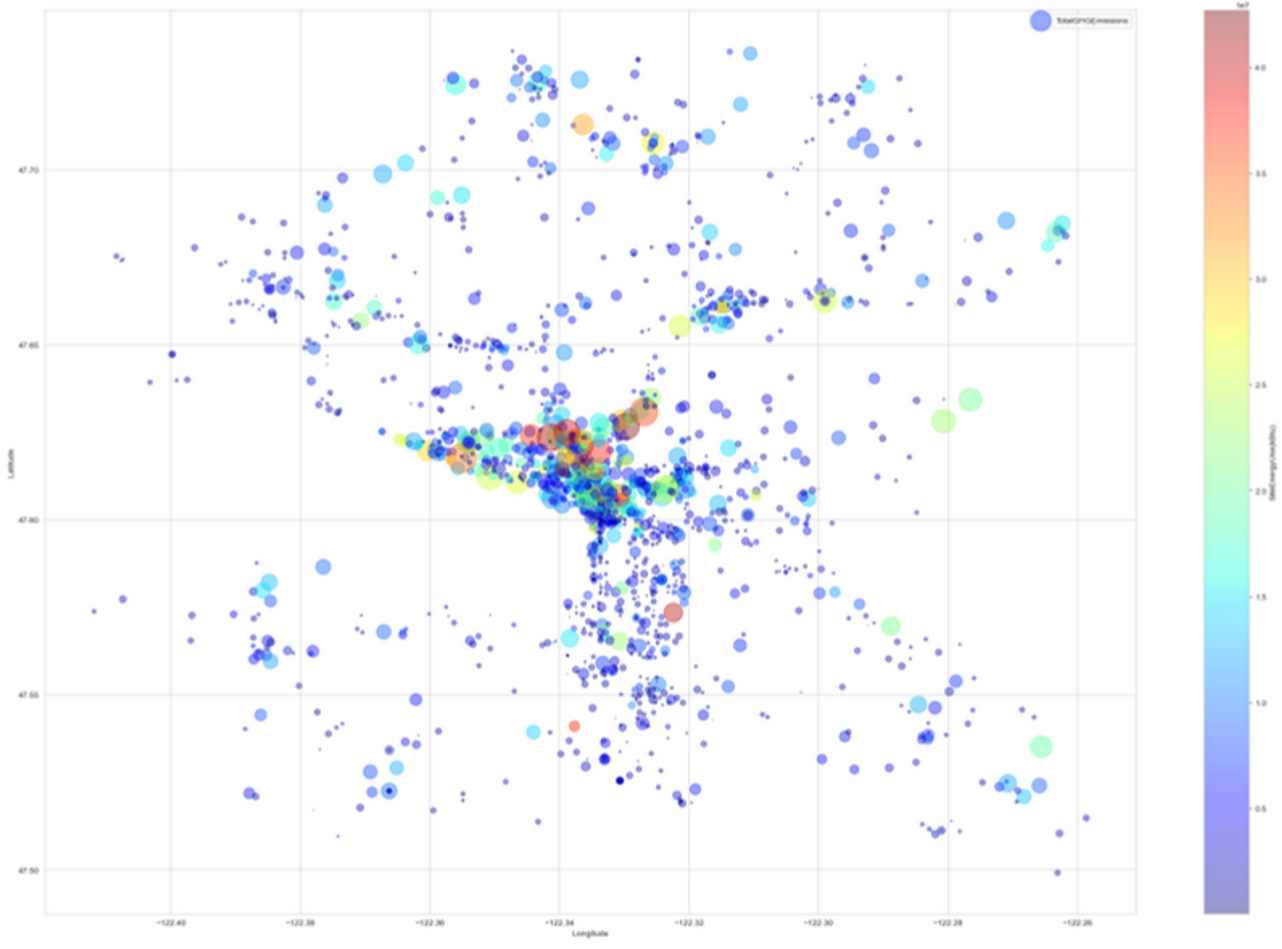
- ***Analyse des corrélations***



03

Analyses univariées et bivariées, Préparation des données (encodage, suppression variables non structurelles, visualisation des distributions, analyse des corrélations).

- ***Visualisation géographique des variables cibles***



- ***Synthèse de l'ensemble des variables***

	count	mean	std	min	25%	50%	75%	max
Latitude	1571.0	4.761609e+01	4.853839e-02	47.49917	4.758437e+01	4.761285e+01	4.765016e+01	4.773387e+01
Longitude	1571.0	-1.223330e+02	2.469276e-02	-122.41182	-1.223436e+02	-1.223329e+02	-1.223216e+02	-1.222586e+02
NumberofBuildings	1571.0	1.014004e+00	4.265787e-01	0.00000	1.000000e+00	1.000000e+00	1.000000e+00	8.000000e+00
NumberofFloors	1571.0	3.711012e+00	5.195221e+00	0.00000	1.000000e+00	2.000000e+00	4.000000e+00	9.900000e+01
PropertyGFAParking	1571.0	1.118304e+04	3.544274e+04	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	4.077950e+05
PropertyGFABuilding(s)	1571.0	8.009118e+04	1.048321e+05	3636.00000	2.781900e+04	4.500000e+04	8.687550e+04	1.047934e+06
SecondLargestPropertyUseTypeGFA	1571.0	1.460592e+04	3.700212e+04	0.00000	0.000000e+00	0.000000e+00	1.179450e+04	3.806390e+05
ThirdLargestPropertyUseTypeGFA	1571.0	2.017665e+03	8.759412e+03	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.772100e+05
ENERGYSTARScore	1571.0	6.504583e+01	2.423274e+01	1.00000	5.300000e+01	6.900000e+01	8.300000e+01	1.000000e+02
SiteEnergyUse(kBtu)	1571.0	5.210508e+06	6.903208e+06	16808.90039	1.216333e+06	2.439485e+06	6.197825e+06	4.270962e+07
TotalGHGEmissions	1571.0	1.061692e+02	1.557337e+02	0.00000	1.958500e+01	4.712000e+01	1.240550e+02	1.266060e+03
BuildingAge	1571.0	5.461235e+01	3.275836e+01	1.00000	2.800000e+01	5.100000e+01	8.600000e+01	1.160000e+02
SteamUse_ratio	1571.0	1.866505e-02	8.624389e-02	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	7.669874e-01
Electricity_ratio	1571.0	6.970903e-01	2.651376e-01	0.00000	4.858442e-01	6.982951e-01	9.999977e-01	1.000011e+00
LargestPropertyUseType_Catering	1571.0	8.911521e-03	9.400921e-02	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Culture	1571.0	1.718651e-02	1.300073e-01	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Data Center	1571.0	1.273074e-03	3.566880e-02	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Education	1571.0	9.293444e-02	2.904330e-01	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Factory	1571.0	4.455761e-03	6.662381e-02	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Healthcare	1571.0	5.092298e-02	2.199105e-01	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Hotel	1571.0	6.174411e-02	2.407668e-01	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Housing	1571.0	7.001910e-03	8.341050e-02	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Office	1571.0	3.176321e-01	4.657038e-01	0.00000	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
LargestPropertyUseType_Other	1571.0	5.983450e-02	2.372555e-01	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Parking	1571.0	1.591343e-02	1.251805e-01	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Stores	1571.0	1.005729e-01	3.008581e-01	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Warehouse	1571.0	1.858689e-01	3.891247e-01	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Well-being	1571.0	2.482495e-02	1.556409e-01	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
LargestPropertyUseType_Worship Facility	1571.0	4.519414e-02	2.077959e-01	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
BuildingType_Nonresidential COS	1571.0	5.346913e-02	2.250387e-01	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
BuildingType_SPS-District K-12	1571.0	5.346913e-02	2.250387e-01	0.00000	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00

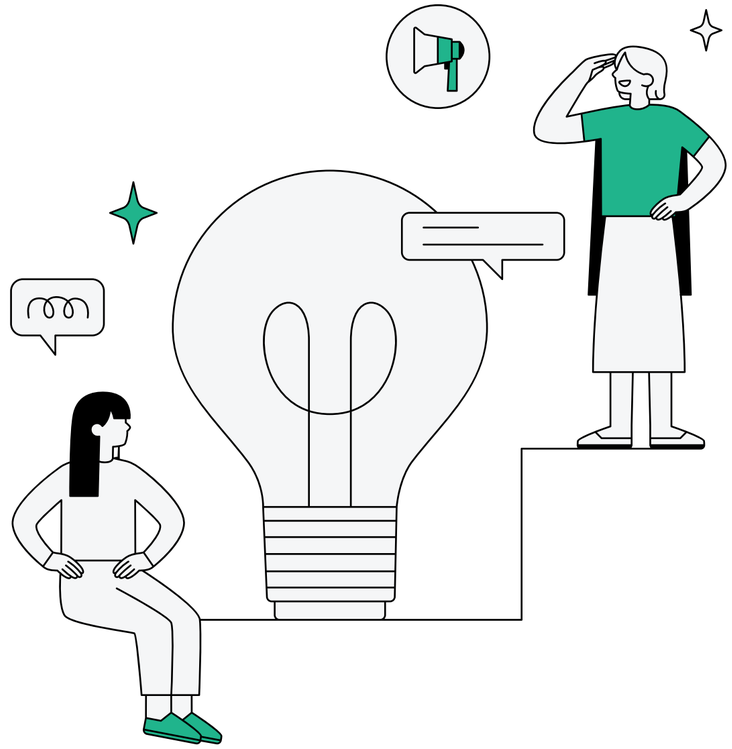
04 Modélisation de la consommation totale d'énergie

Split train/test, Standardisation, Choix des modèles à tester, Métriques d'évaluation, Optimisation hyperparamètres, Calcul des performances, Sélection du modèle à retenir, Feature importance, Interprétabilité globale et locale.

La démarche adoptée

- ***La variable cible*** *pour cette prédiction c'est la variable 'SiteEnergyUse(kBtu)'*
- ***Les variables explicatives*** *sont composées de l'ensemble des variables de l'état de synthèse (l'étape 3) sans la variable 'TotalGHGEmissions'*
- ***Nous divisons l'ensemble des données à 70% pour l'entrainement et 30% pour l'évaluation*** *afin d'identifier une éventuelle présence d'overfitting ou underfitting avec précision d'un random_state pour la reproductivité de l'échantillonnage.*
- ***Standardisation des données :*** *Les données ne sont pas à la même échelle de mesure, elles présentent de nombreuses valeurs aberrantes et ne suivent pas une distribution normale. Nous avons standardisé en utilisant RobustScaler.*

Modèles à tester	Justification du choix
DummyRegressor	Nous l'utiliserons comme base de référence simple en prédisant la moyenne
LinearRegression	Pour ajuster un modèle linéaire avec des coefficients et minimiser la somme des carrés résiduelle entre les données observées et les données prédictites.
KNeighborsRegressor	Pour prédire la valeur de la cible en fonction des K valeurs qui lui sont les plus proches.
RandomForestRegressor	Pour utiliser plusieurs regresseurs d'arbre de décision sur différents échantillons de nos données.
Ridge , ElasticNet	Pour ajouter de la pénalisation, régulariser et choisir des variables pertinentes pour ElasticNet.
XGBRegressor	Pour utiliser la force de la combinaison d'un ensemble de modèles prédictifs faibles et booster les performances.
LGBMRegressor	Pour utiliser la force de la combinaison d'un ensemble de modèles prédictifs faibles et booster les performances.



04 Modélisation de la consommation totale d'énergie

Split train/test, Standardisation, Choix des modèles à tester, Métriques d'évaluation, Optimisation hyperparamètres, Calcul des performances, Sélection du modèle à retenir, Feature importance, Interprétabilité globale et locale.

- ***Résultats des modéles non paramétriques***

DummyRegressor

Métrique	train	test
R2	0.00E+00	-3.15E-04
RMSE	7.04E+06	6.56E+06
MAE	4.75E+06	4.50E+06
EVS	0.00E+00	0.00E+00

LinearRegression

Métrique	train	test
R2	6.51E-01	6.28E-01
RMSE	4.16E+06	4.01E+06
MAE	2.46E+06	2.42E+06
EVS	6.51E-01	6.28E-01

- ***Optimisation des hyperparamètres***

==> KNeighborsRegressor : *n_neighbors :12 ; weights : uniform*

==> RandomForestRegressor :
```
Best parameters set :
{'max_depth': 25, 'n_estimators': 200}
```

==> Ridge : *Best alpha = 0.01*

==> ElasticNet : *alpha = 10000.0 ; l1_ratio = 1.0*

==> XGBRegresssor :
```
Best parameters set :
{'learning_rate': 0.01, 'max_depth': 10, 'n_estimators': 400, 'subsample': 0.3}
```

==> LGBMRegressor:
```
Best parameters set :
{'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 400}
```

- ***Calcul des performances et choix du meilleur modèle***

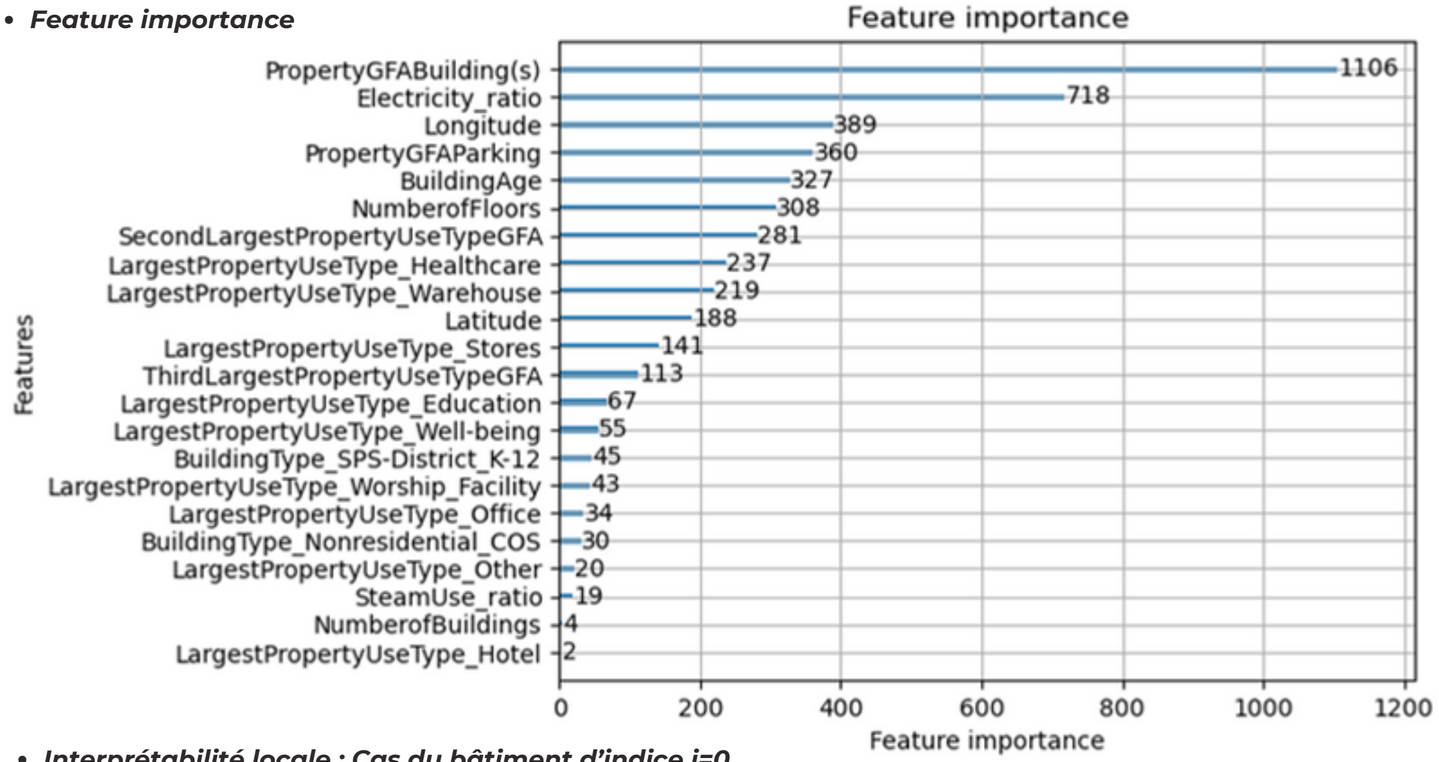
Modèles	R2_train	R2_test	RMSE_train	RMSE_test	MAE_train	MAE_test	EVS_train	EVS_test	Temps (s)
KNeighborsRegressor	4.70E-01	3.80E-01	5.12E+06	5.17E+06	2.97E+06	2.92E+06	4.74E-01	3.88E-01	24.19
RandomForestRegressor	9.48E-01	6.60E-01	1.60E+06	3.83E+06	8.91E+05	2.26E+06	9.48E-01	6.60E-01	252.51
Ridge	6.51E-01	6.28E-01	4.16E+06	4.00E+06	2.46E+06	2.42E+06	6.51E-01	6.28E-01	0.12
ElasticNet	6.47E-01	6.33E-01	4.18E+06	3.98E+06	2.47E+06	2.38E+06	6.47E-01	6.33E-01	2.54
XGBRegressor	8.91E-01	6.71E-01	2.32E+06	3.76E+06	1.31E+06	2.15E+06	8.91E-01	6.72E-01	985.19
LGBMRegressor	**7.69E-01**	**6.18E-01**	**3.38E+06**	**4.05E+06**	**1.96E+06**	**2.32E+06**	**7.69E-01**	**6.19E-01**	**226.06**

04

Modélisation de la consommation totale d'énergie

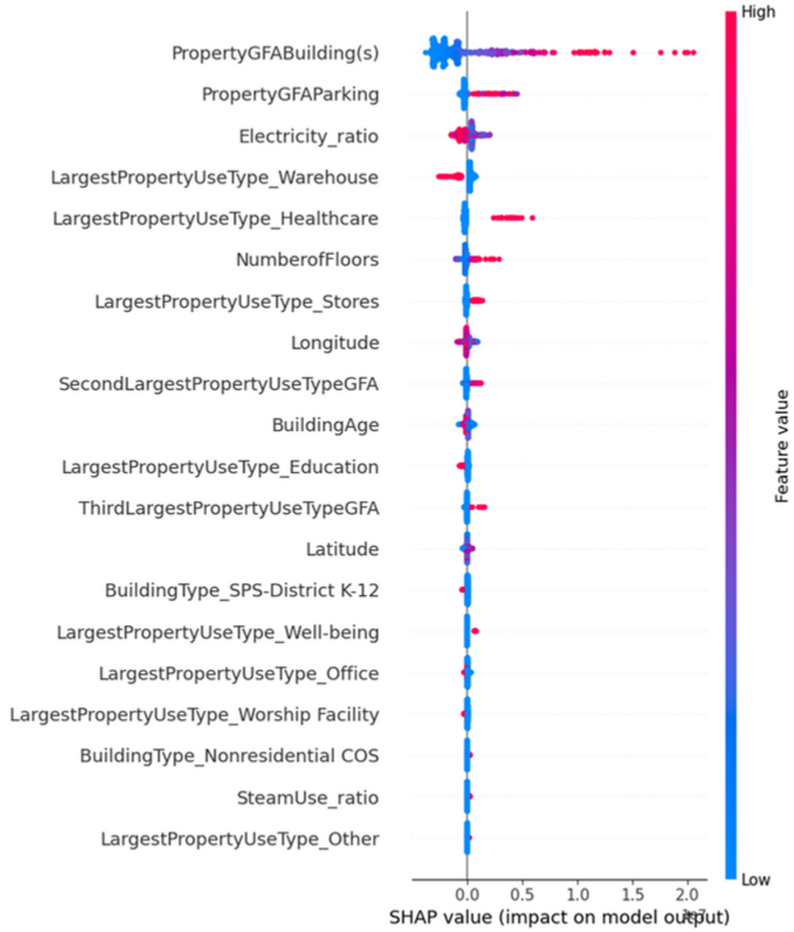
Split train/test, Standardisation, Choix des modèles à tester, Métriques d'évaluation, Optimisation hyperparamètres, Calcul des performances, Sélection du modèle à retenir, Feature importance, Interprétabilité globale et locale.

- ***Feature importance***



- ***Interprétabilité locale : Cas du bâtiment d'indice i=0***

- ***Interprétabilité globale***



05

Modélisation des émissions de CO2

Split train/test, Standardisation, Choix des modèles à tester, Métriques d'évaluation, Optimisation hyperparamètres, Calcul des performances, Sélection du modèle à retenir, Feature importance, Interprétabilité globale et locale, Pertinence ENERGYSTARScore.

Prédiction des émissions de CO2 avec l'ENERGYSTARScore comme variable explicative

- ***La variable cible*** *pour cette prédiction c'est la variable 'TotalGHGEmissions'*
- ***Création de la variable GHGE_class*** *pour la stratification lors de l'échantillonnage et permettre un bon brassage des données.*
- ***Les variables explicatives*** *sont composées de l'ensemble des variables de l'état de synthèse (l'étape 3) sans les variables 'SiteEnergyUse(kBtu)' et 'GHGE_class'.*
- ***Nous divisons l'ensemble des données à 70% pour l'entrainement et 30% pour l'évaluation*** *afin d'identifier une éventuelle présence d'overfitting ou underfitting avec précision d'un random_state pour la reproductivité de l'échantillonnage.*
- ***Standardisation des données :*** *Les données ne sont pas à la même échelle de mesure, elles présentent de nombreuses valeurs aberrantes et ne suivent pas une distribution normale. Nous avons standardisé en utilisant RobustScaler.*
- ***Performances sur les modèles choisis avec les paramètres par défaut***

	R2_train	R2_test	RMSE_train	RMSE_test	MAE_train	MAE_test	MEDAE_train	MEDAE_test	EVS_train	EVS_test
lgbm_reg	0.947368	0.658550	35.553218	91.930636	19.144300	46.573438	10.491676	19.389295	0.947368	6.585505e-01
rf_reg	0.954741	0.652359	32.969053	92.760263	17.286737	45.871453	7.007300	16.975650	0.954742	6.526263e-01
xgb_reg	0.999277	0.640517	4.168294	94.326984	2.863969	46.986600	1.915154	20.340714	0.999277	6.405196e-01
Lin_reg	0.531512	0.499463	106.072433	111.305099	64.549128	64.671606	40.071030	41.103333	0.531512	4.998652e-01
ela_reg	0.362043	0.320548	123.779632	129.680935	73.283783	72.591739	46.969106	44.351584	0.362043	3.207824e-01
knn_reg	0.438095	0.143793	116.167566	145.574775	64.109383	81.040123	27.176000	32.661000	0.439459	1.470630e-01
ada_reg	0.205707	0.054232	138.115942	152.999134	129.139236	135.992646	134.151405	136.259486	0.682073	5.094452e-01
dum_reg	0.000000	-0.000092	154.972065	157.331874	100.488322	100.471588	78.656069	77.451069	0.000000	1.110223e-16

- ***Optimisation des paramètres sur les modèles les plus performants***

==> LGBMRegressor

```
Best parameters set :
{'learning_rate': 0.01, 'max_depth': 20, 'n_estimators': 300}
```

==> RandomForestRegressor

```
Best parameters set :
{'criterion': 'absolute_error', 'max_depth': 15, 'max_features': None, 'n_estimators': 300}
```

==> XGBRegressor

```
Best parameters set :
{'learning_rate': 0.01, 'max_depth': 25, 'n_estimators': 400, 'subsample': 0.3}
```

- ***Performances des modèles optimisés et choix***

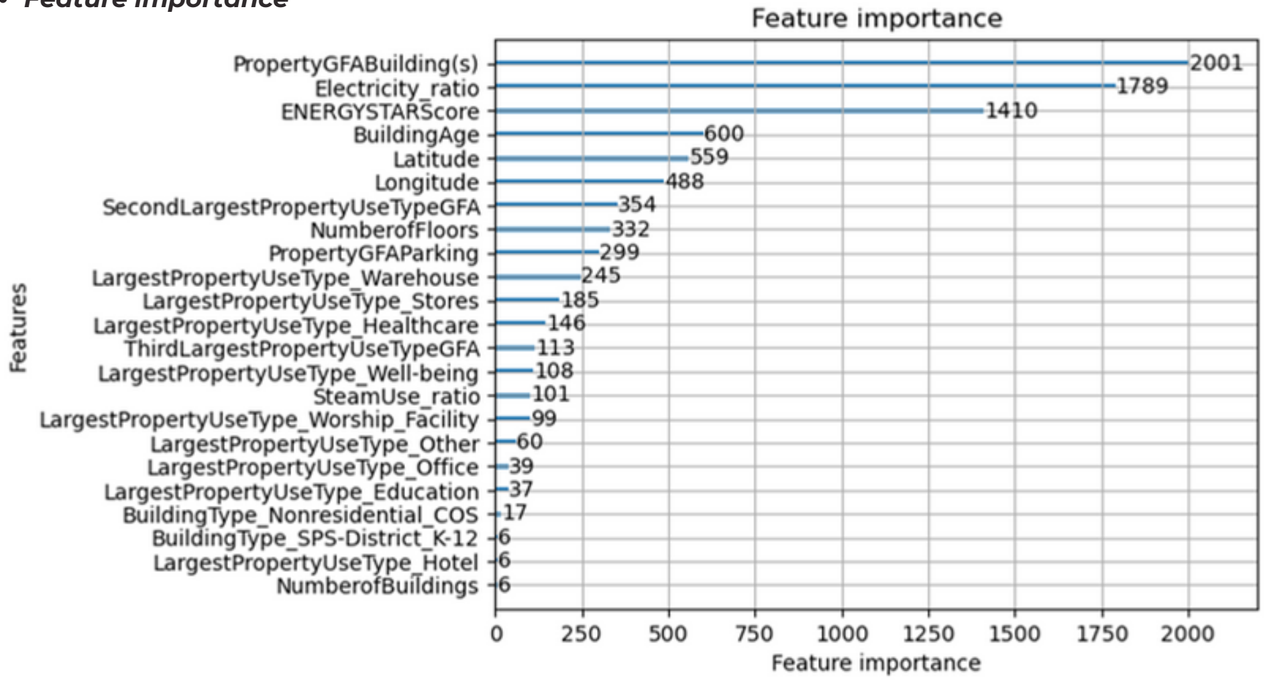
Modèles	R2_train	R2_test	RMSE_train	RMSE_test	MAE_train	MAE_test	MEDAE_train	MEDAE_test	EVS_train	EVS_test	Temps (s)
LGBMRegressor	**8.25E-01**	**6.48E-01**	**64.91**	**93.29**	**33.18**	**46.56**	**14.85**	**18.98**	**8.25E-01**	**6.48E-01**	**21.63**
RandomForestRegressor	9.51E-01	6.28E-01	34.43	95.98	18.33	46.35	8.03	16.43	9.51E-01	6.28E-01	325.26
XGBRegressor	9.20E-01	6.80E-01	43.91	88.95	20.99	43.90	9.02	19.13	9.20E-01	6.80E-01	443.81

05

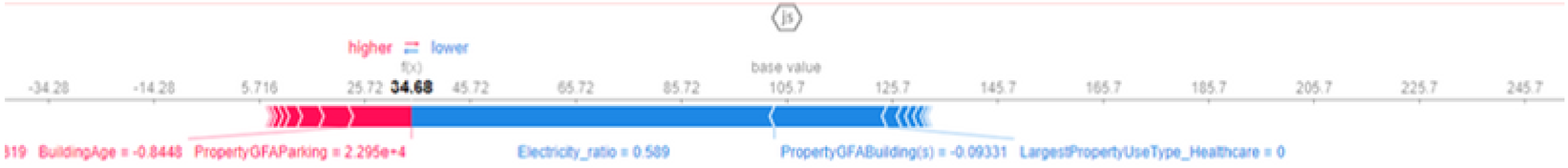
Modélisation des émissions de CO2

Split train/test, Standardisation, Choix des modèles à tester, Métriques d'évaluation, Optimisation hyperparamètres, Calcul des performances, Sélection du modèle à retenir, Feature importance, Interprétabilité globale et locale, Pertinence ENERGYSTARScore.

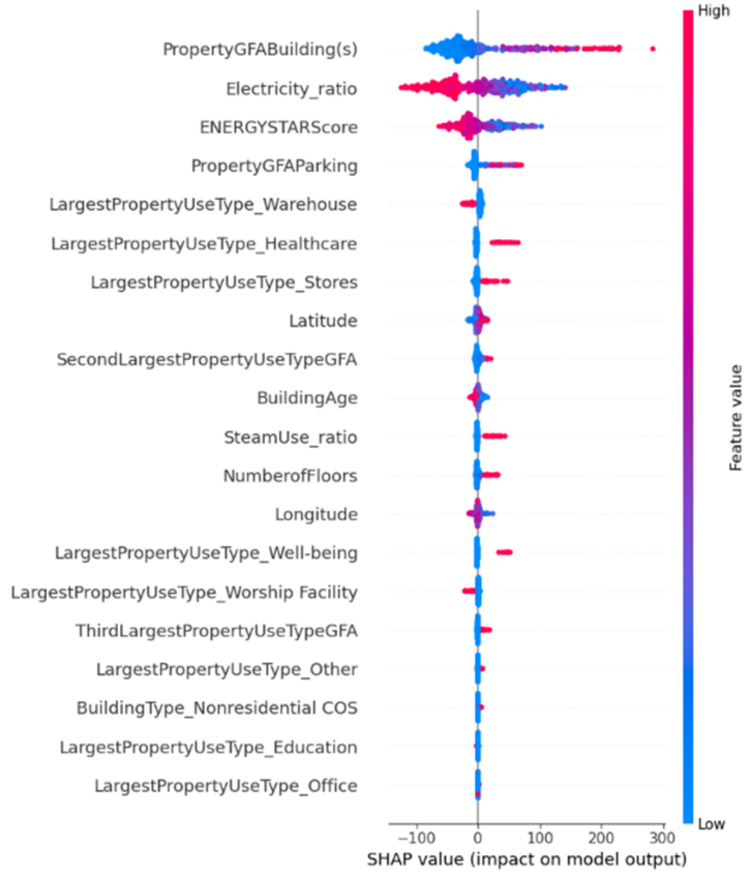
- ***Feature importance***



- ***Interprétabilité locale : Cas du bâtiment d'indice i =0***



- ***Interprétabilité globale***



05

Modélisation des émissions de CO2

Split train/test, Standardisation, Choix des modèles à tester, Métriques d'évaluation, Optimisation hyperparamètres, Calcul des performances, Sélection du modèle à retenir, Feature importance, Interprétabilité globale et locale, Pertinence ENERGYSTARScore.

Prédiction des émissions de CO2 sans l'ENERGYSTARScore comme variable explicative

- ***La variable cible*** *ne change pas 'TotalGHGEmissions'*
- *Nous appliquons toujours la stratification à l'aide de la variable GHGE_class.*
- ***Les variables explicatives*** *sont composées de l'ensemble des variables de l'état de synthèse (l'étape 3) sans les variables 'SiteEnergyUse(kBtu)' , 'ENERGYSTARScore' et 'GHGE_class'.*
- ***Nous divisons l'ensemble des données à 70% pour l'entrainement et 30% pour l'évaluation*** *afin d'identifier une éventuelle présence d'overfitting ou underfitting avec précision d'un random_state pour la reproductivité de l'échantillonnage.*
- ***Standardisation des données :*** *Les données ne sont pas à la même échelle de mesure, elles présentent de nombreuses valeurs aberrantes et ne suivent pas une distribution normale. Nous avons standardisé en utilisant RobustScaler.*

- ***Afin de juger de la pertinence uniquement de la variable ENEERGYSTARScore, nous utilisons le meilleur modèle optimisé sélectionné précédemment LGBMRegressor***

```
Best parameters set :
{'learning_rate': 0.01, 'max_depth': 20, 'n_estimators': 300}
```

- ***Performances du modèle optimisés***

Modèles	R2_train	R2_test	RMSE_train	RMSE_test	MAE_train	MAE_test	MEDAE_train	MEDAE_test	EVS_train	EVS_test	Temps (s)
LGBMRegressor	**7.89E-01**	**6.21E-01**	**71.21**	**96.80**	**38.04**	**48.88**	**17.07**	**21.28**	**7.89E-01**	**6.22E-01**	**448.04**
RandomForestRegressor	9.31E-01	6.32E-01	40.69	95.38	22.19	47.67	10.21	20.53	9.31E-01	6.32E-01	477.95
XGBRegressor	9.02E-01	6.43E-01	48.53	93.96	24.11	48.25	10.86	21.70	9.02E-01	6.44E-01	1.26

- ***Conclusion sur la pertinence de la variable ENERGYSTARScore***

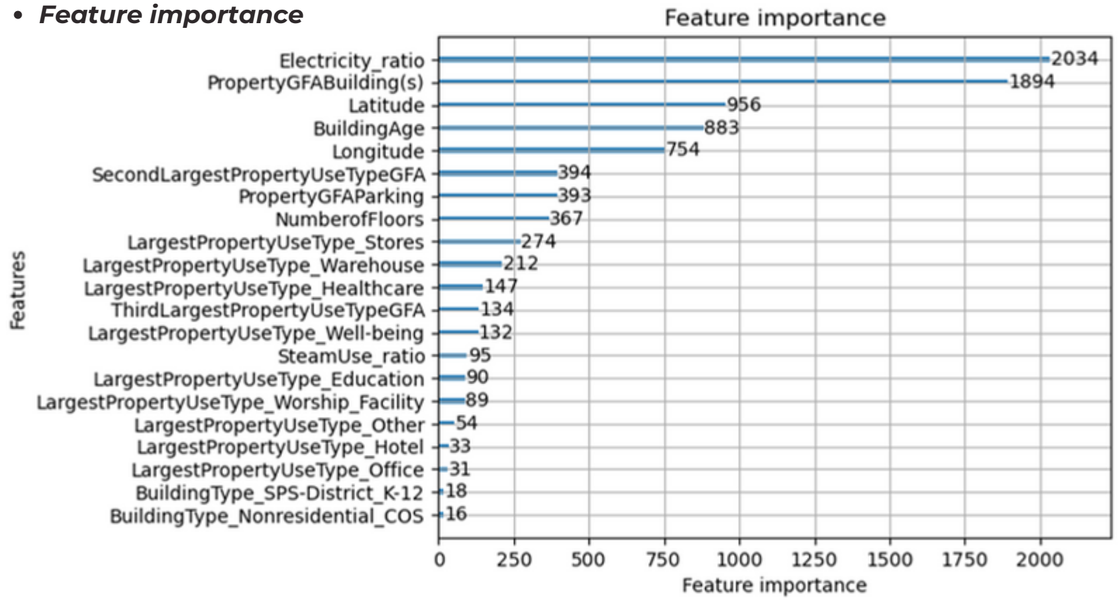
En prédisant les émissions de CO2 avec le même modèle LGBM et sans retenir l'ENERGYSTARScore comme variable explicative, les performances r2_score sur les données de test ne varient pas énormément 0.621 contre 0.648 précédemment. Sur les données d'entrainement ces performances sont de 0.789 contre 0.825 précédemment. Nous constatons que sans l'ENERGYSTARScore comme variable explicative, le modèle surapprend un peu moins. Cependant ces performances ne sont pas significativement différentes.
Compte tenu des difficultés rencontrées pour obtenir ces données (coût et calcul fastidieux), cette variable n'est pas pertinente pour la prédiction des émissions de CO2.

05

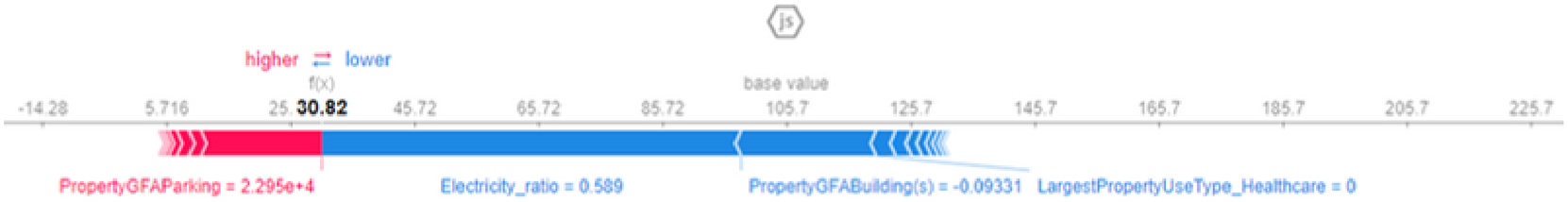
Modélisation des émissions de CO2

Split train/test, Standardisation, Choix des modèles à tester, Métriques d'évaluation, Optimisation hyperparamètres, Calcul des performances, Sélection du modèle à retenir, Feature importance, Interprétabilité globale et locale, Pertinence ENERGYSTARScore.

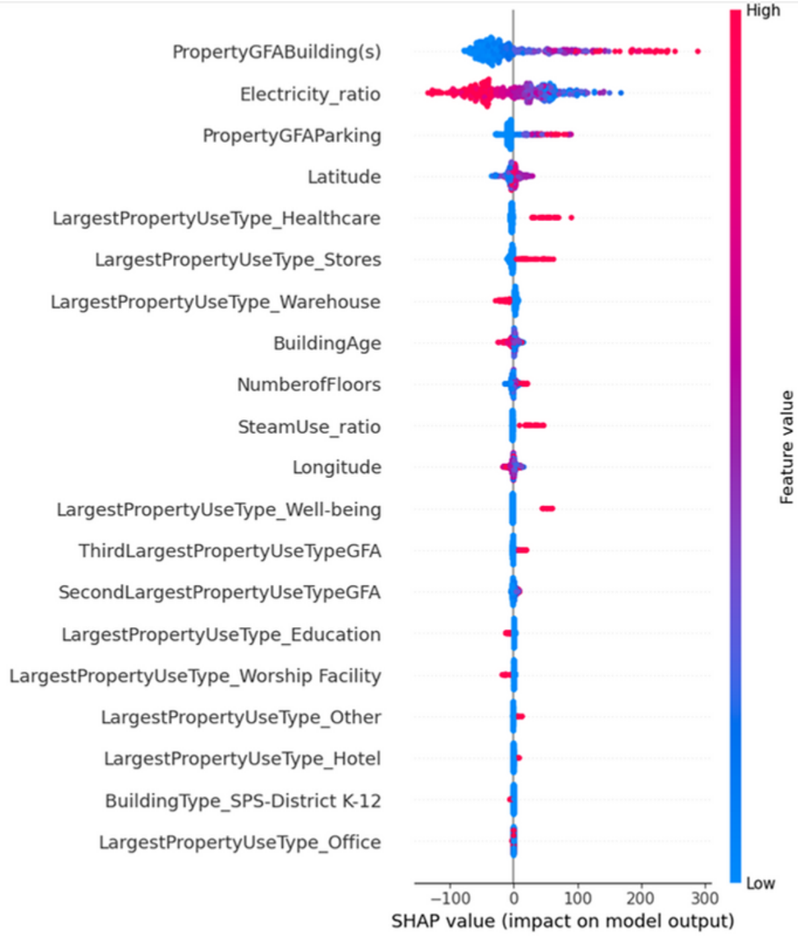
- ***Feature importance***

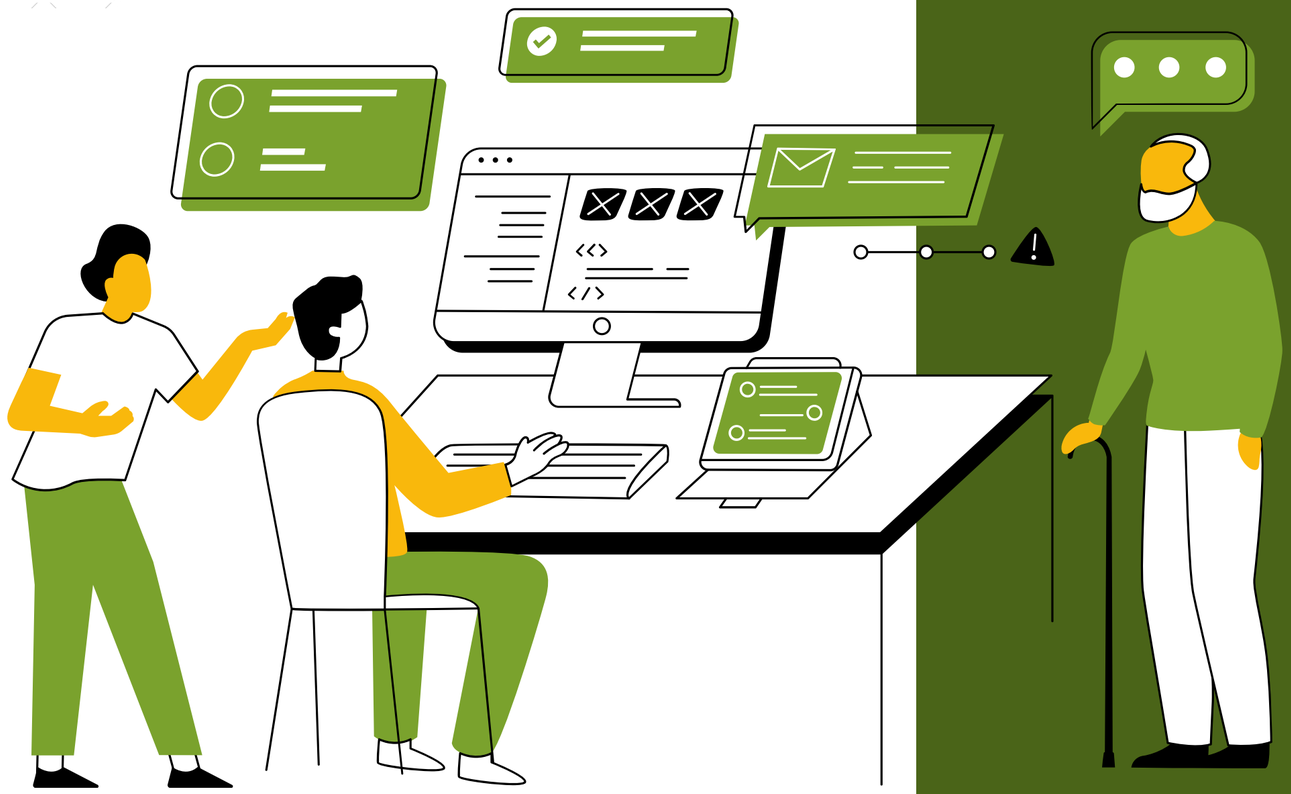


- ***Interprétabilité locale : Cas du bâtiment d'indice i=0***



- ***Interprétabilité globale***





Merci