Anticipez les besoins en consommation de bâtiments

Seattle

Objectif: Neutralité carbone 2050

Soutenance Projet 4 : Habibatou BA 29/09/2023

Sommaire

- Problématique
- Analyser les données
- Description des données
- Pré-traitement des données
- Nettoyage des données et Analyse exploratoire des données
- Modélisation
- Conclusion
- Suite du Projet

Problématique : Contexte

Objectif de la ville de Seattle :

Neutralité carbone en 2050.
33% des émissions par bâtiments non résidentiels
➔ connaître leurs **consommation** en énergie et **émission.**

Problématique : Mission

A partir des données récoltées :

- Prédire la consommation totale d’énergie.
- Prédire l’émission.
- Evaluer l’intérêt de l’ENERGY STAR Score pour la prédiction d’émission.

Analyser les données

- Réaliser une courte analyse exploratoire.
- Tester différents modèles de prédiction afin de répondre au mieux à la problématique.
- Mettre en place une évaluation rigoureuse des performances, et optimiser les hyperparamètres et le choix d'algorithmes de ML à l'aide d'une validation croisée.
- Tester au minimum 4 algorithmes de famille différente (par exemple : ElasticNet, SVM, GradientBoosting, XGBoost, RandomForest).

Descriptions des données

- Données disponibles à l'adresse : Seatlle Data
- Constitué de **3376 lignes** et **46 colonnes**
- **Filtrage sur les batiments non-residentiels**

Sélectionner les targets :

- Total/Intensité?
- Source/Site
  Site/SiteWN

Modélisation :

2 variables cibles quantitatives à prédire

SiteEnergyUseWN(kBtu)

➔ modèle consommation d'énergie

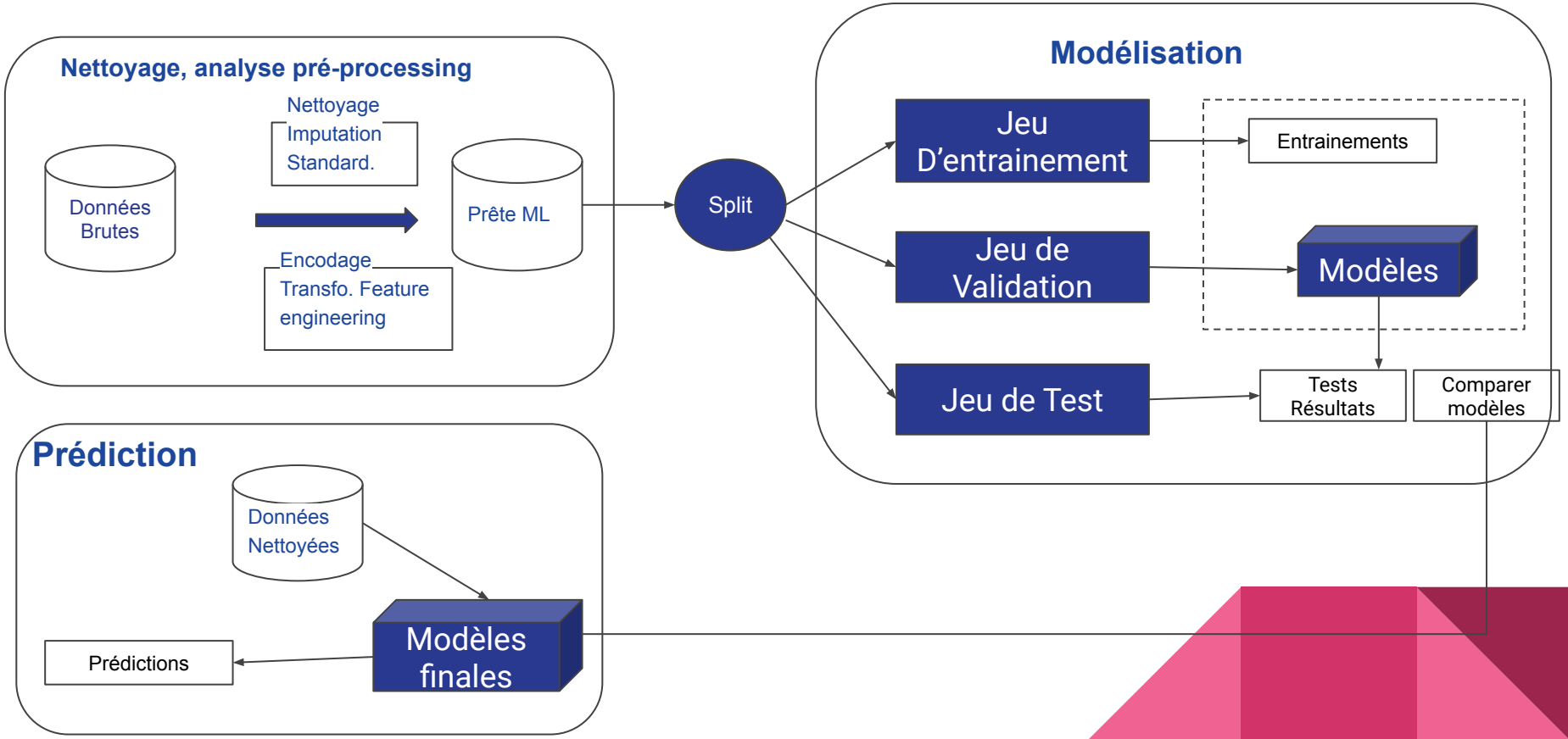
TotalGHGEmissions

➔modèle émission de GES

Intérêt de **EnergyStar Score**

➔ modèle à comparer avec EnergyStar Score

Pré-traitements des données : Démarche



Nettoyage : réductions des données

- Chargement des données => 3376 lignes et 46 colonnes
- Suppressions des colonnes vides à 80% et suppressions des variables inutiles après analyses métiers => 3376 lignes et 31 colonnes
- Conserver uniquements les bâtiments non-résidentiels => 1668 lignes et 31 colonnes
- Imputations des valeurs manquantes => 1665 lignes et 31 colonnes
- Feature engineering et filtre sur les variables utiles => 1665 lignes et 23 colonnes

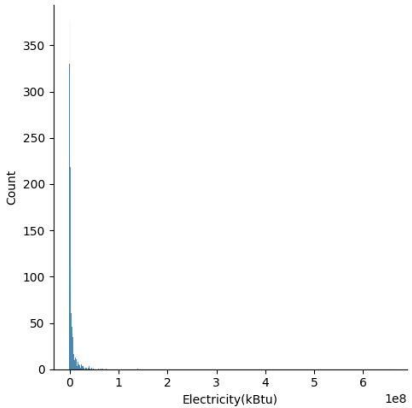
Nettoyage : valeurs manquantes : Méthodes d’Imputations

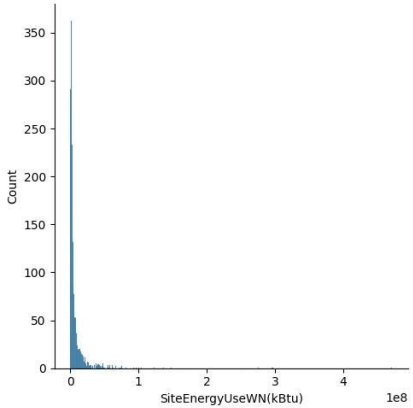
Seattle

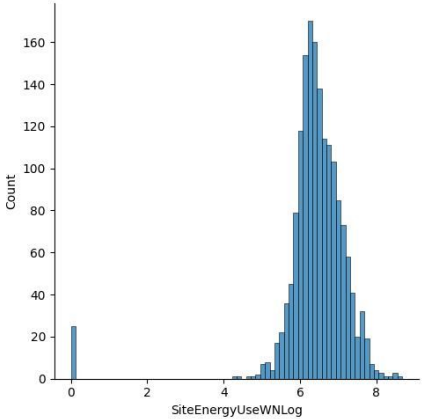
- IMputation par mise à '0' et ‘nonUtilisée’ pour certaines variables
- IMputation pour EnergyStarScore par la méthode : SimpleImputer

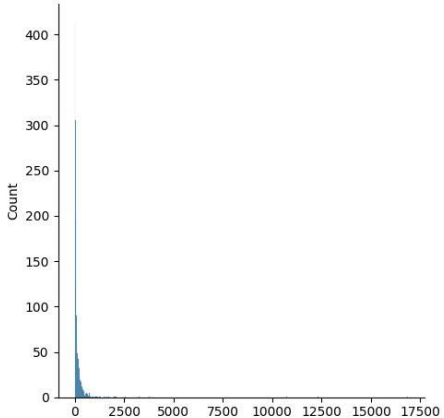
Analyse exploratoire des données

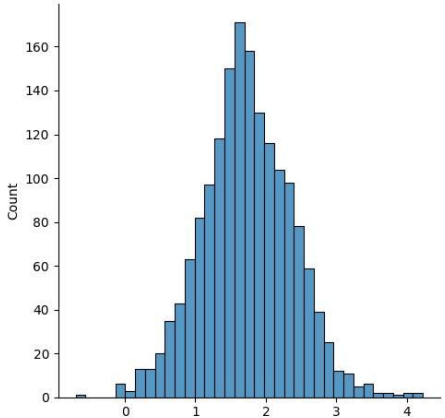
Seattle

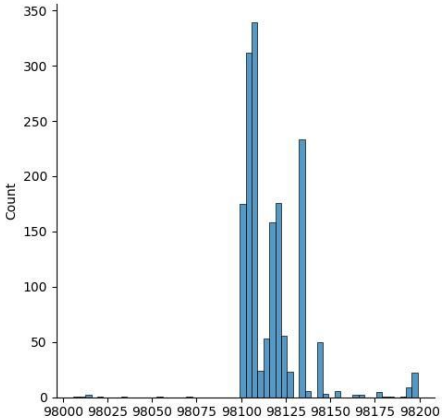






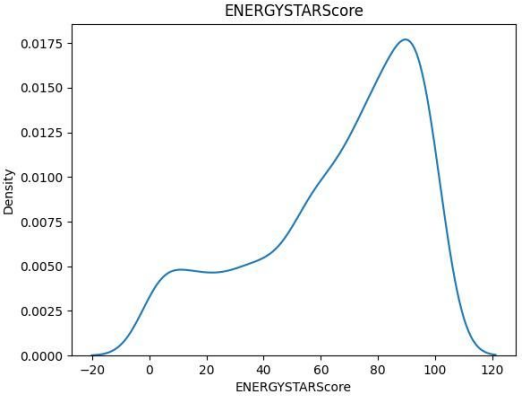


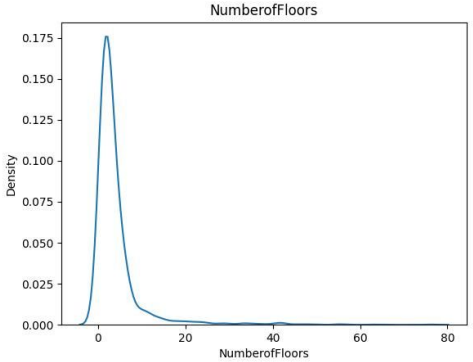


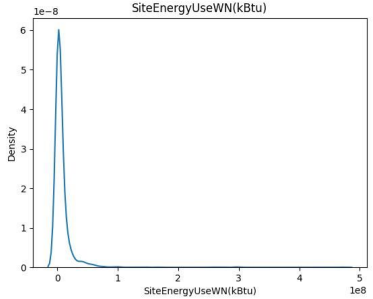


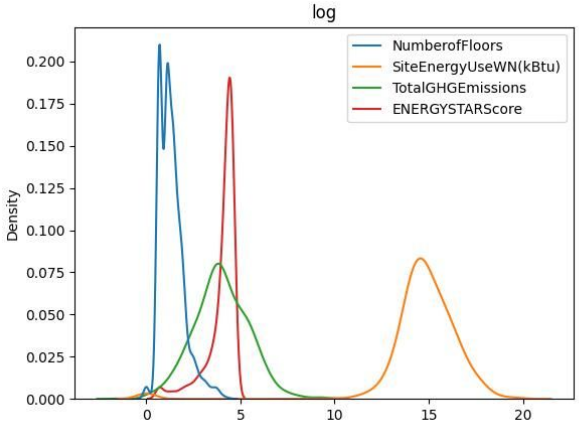
Analyse exploratoire des données

Seattle



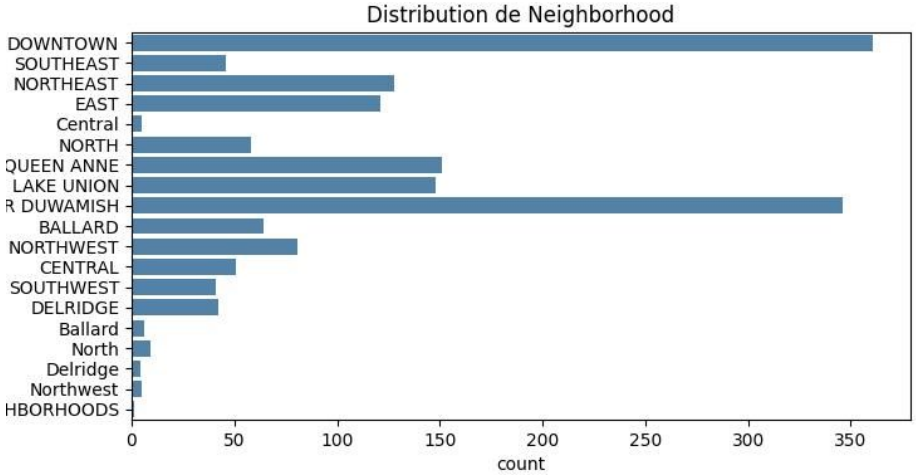


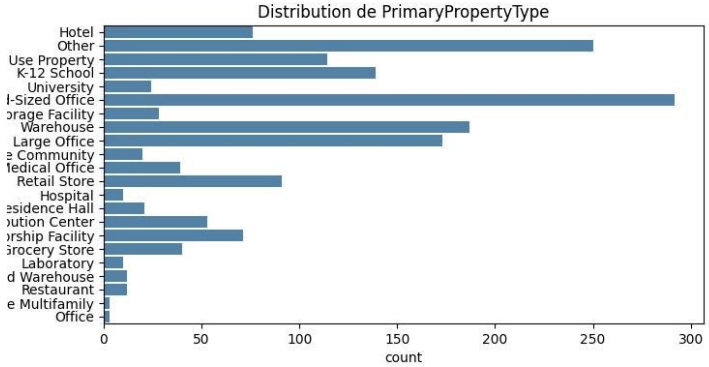


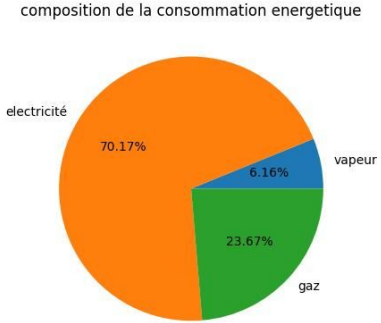


Analyse exploratoire des données

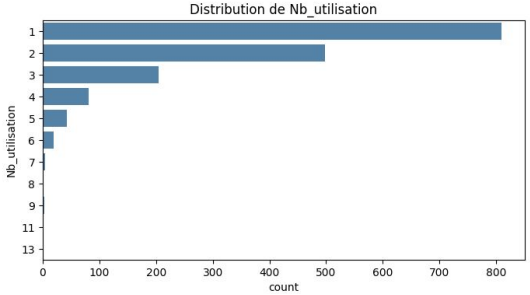
Seattle

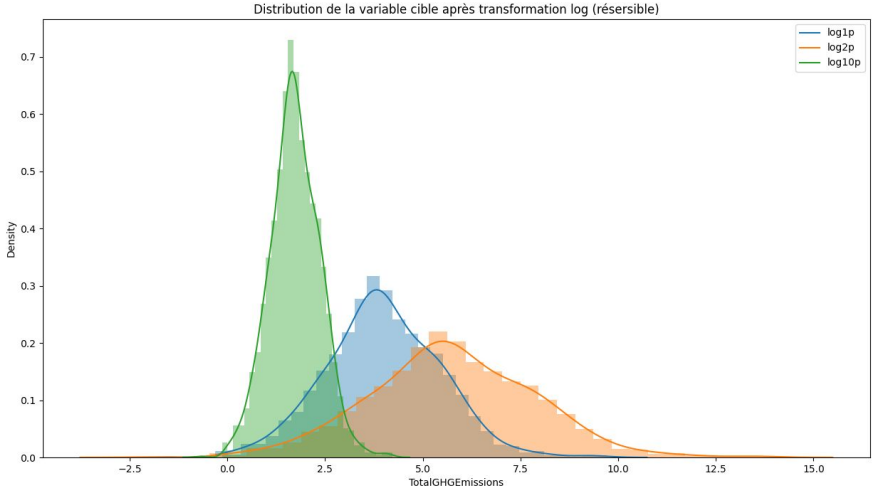


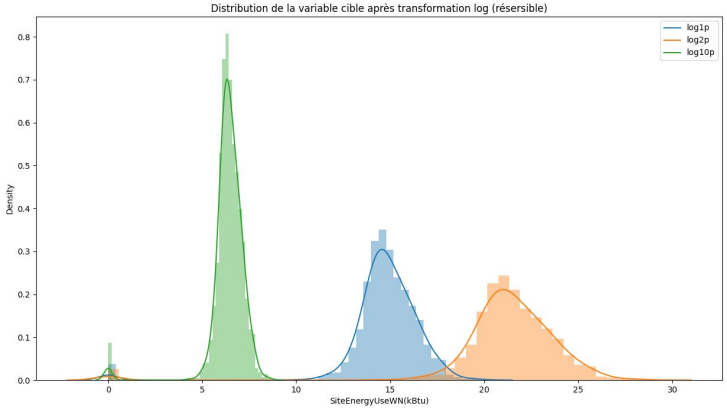




Analyse exploratoire des données: Feature Engineering



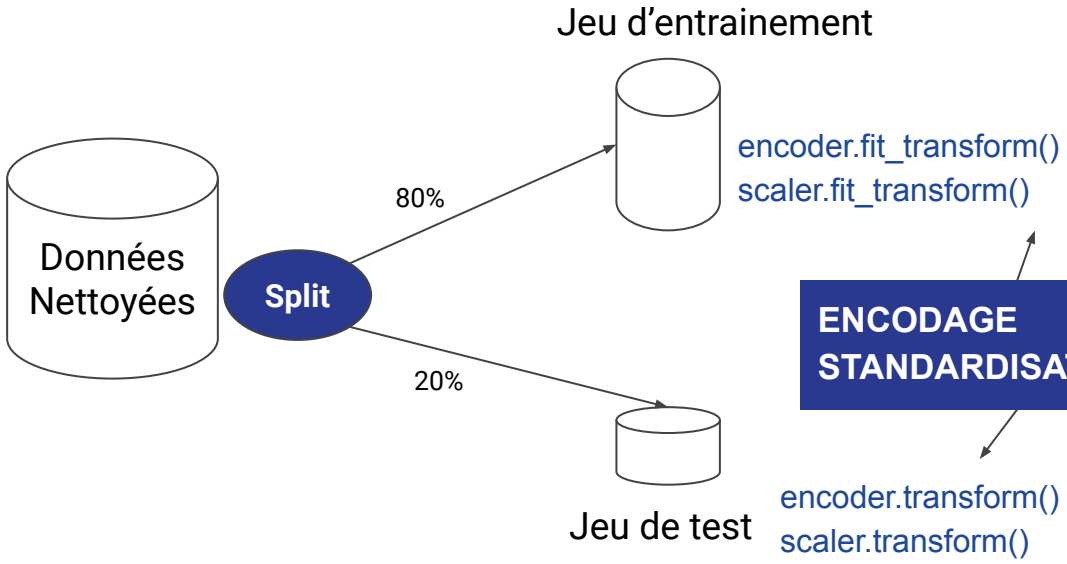




Modélisation : Prédiction de la Consommation d'Énergie

Seattle

Split - Encodage/Standardisation



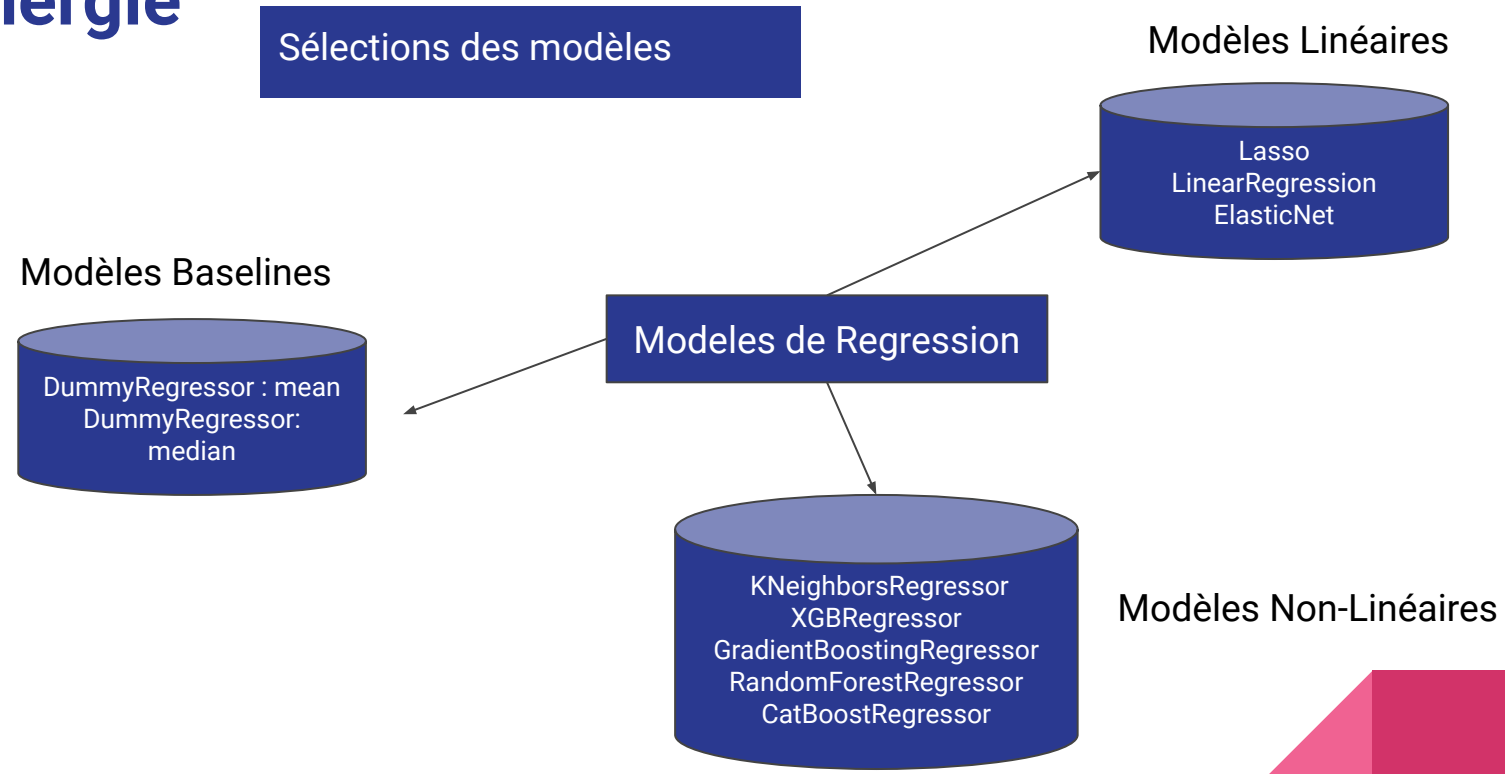
Variables catégorielles : encodage avec encoder : **TargetEncoder**

Variables numériques : standardisation avec scaler : **RobustScaler**

VALIDATION CROISÉE LORS DE L'ENTRAÎNEMENT

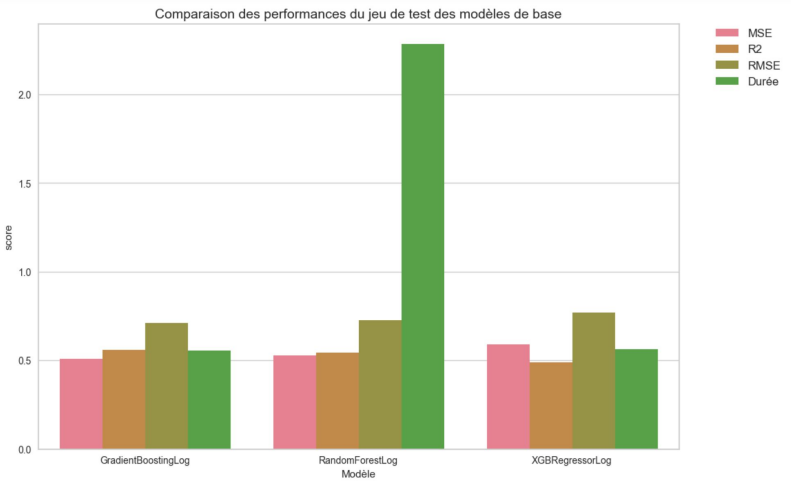
Modélisation:Prédiction de la Consommation d'Énergie

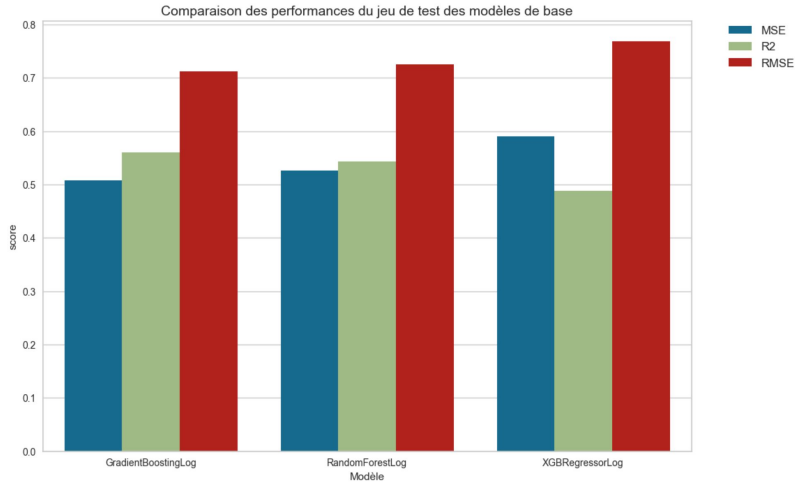
Seattle



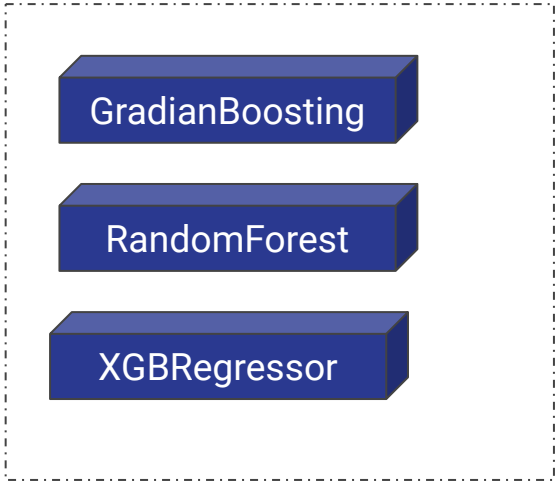
Modélisation : Performance et Co

Modèle	R2	MSE	RMSE	MAE	Erreur moy	Précision	Durée	Test R2 CV	Test R2 +/-
GradientBoostingLog	0.560088	0.507537	0.712416	0.199849	0.199849	-inf	0.555769	0.552387	0.215114
RandomForestLog	0.543401	0.526789	0.725802	0.164686	0.164686	-inf	2.284299	0.521722	0.284616
XGBRegressorLog	0.488215	0.590459	0.768413	0.196762	0.196762	-inf	0.562517	0.495372	0.317290
CatBoostRegressor_log	0.242668	0.873751	0.934747	0.242177	0.242177	-inf	2.857576	0.588850	0.290842
LinearRegressionLog	0.212338	0.908744	0.953281	0.439155	0.439155	-inf	0.025898	0.050988	0.545138
KNeighborsRegressorLog	0.124500	1.010085	1.005030	0.362802	0.362802	-inf	0.017675	0.230063	0.177624
ElasticNetLog	0.016628	1.134539	1.065148	0.531083	0.531083	-inf	0.026365	-0.028435	0.234924
Lasso_log	0.008236	1.144222	1.069683	0.538103	0.538103	-inf	0.024967	0.011320	0.044914
DummyRegressor_meanlog	-0.000844	1.154697	1.074569	0.546214	0.546214	-inf	0.001144	-0.008290	0.007637
DummyRegressor_medianlog	-0.002474	1.156578	1.075443	0.545697	0.545697	-inf	0.001602	-0.006541	0.007419





Modélisation : Optimisation du Top 3



Optimisation : Hyperparamètres

Recherche Manuelle

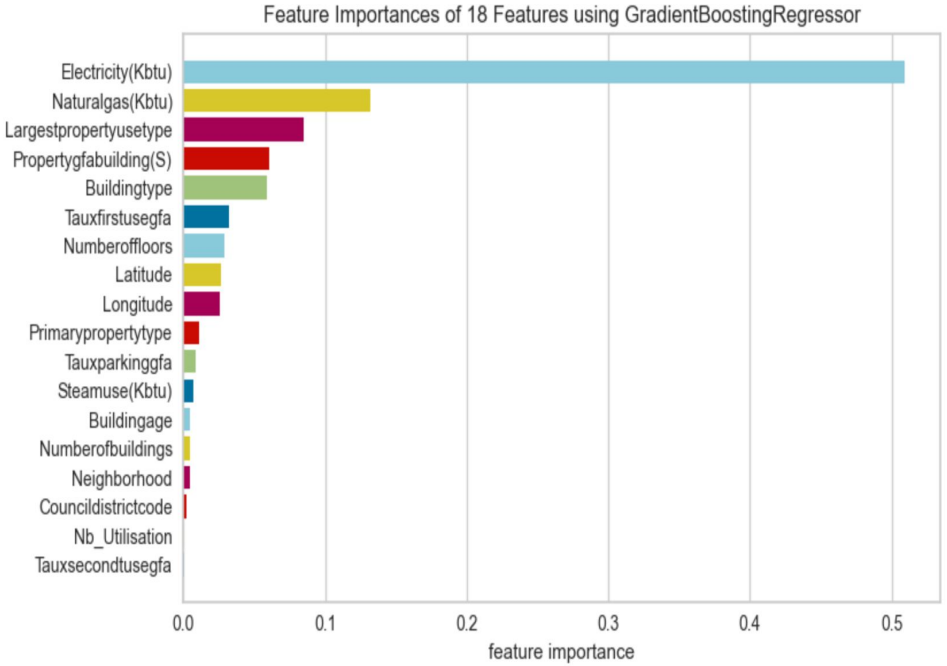
Recherche automatique

- **GridSearchCV**
- **RandomizedSearchCV**

Modélisation : Exemple Optimisation GradientBoosting

Optimisation : Hyperparamètres :
n_estimators=107
min_samples_split=2
min_samples_leaf=1
max_depth=5

Modèle	R2	MSE	RMSE	MAE	Erreur moy	Précision	Durée
GradientBoosting_Log	0.560088	0.507537	0.712416	0.199849	0.199849	-inf	0.541330
GradientBoosting_Log_optimisé	0.581736	0.482561	0.694666	0.169330	0.169330	-inf	0.974997



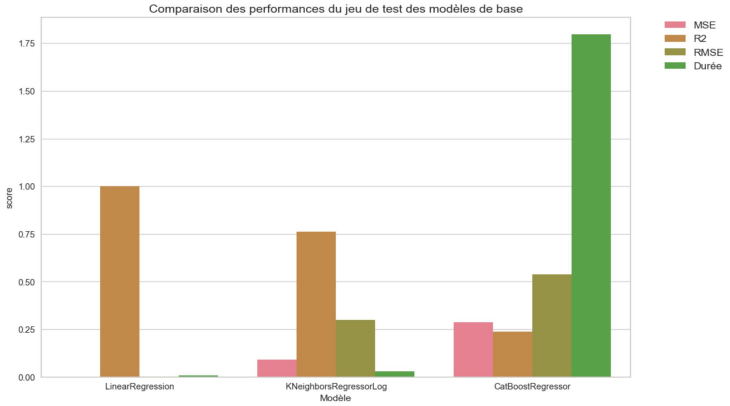
Modélisation : Prédiction de l'émission de Co2

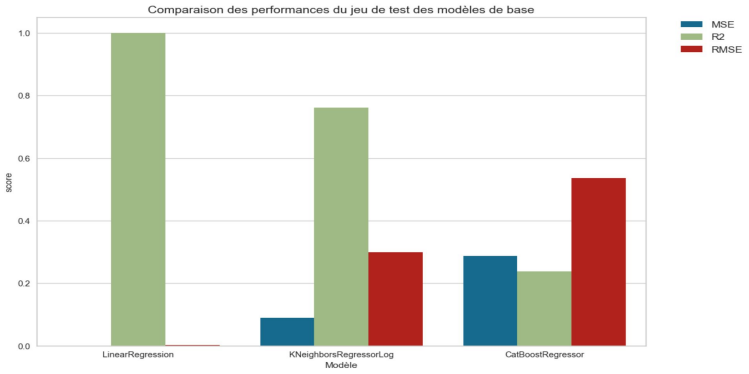
Sans Energy Star Score

Cible **TotalGHGEmissions** ➔RÉGRESSION. Même démarche : split, encodage, standardisation, modèle de base.

Sélections des modèles

Modèle	R2	MSE	RMSE	MAE	Erreur moy	Précision	Durée	Test R2 CV	Test R2 +/-	Test MSE CV
LinearRegression	1.000000	0.000009	0.003031	0.002650	0.002650	99.987976	0.007654	1.000000	0.000000	0.000009
KNeighborsRegressorLog	0.762072	0.090012	0.300019	0.208750	0.208750	83.037387	0.029567	0.755737	0.033124	0.100414
CatBoostRegressor	0.238762	0.287988	0.536645	0.414214	0.414214	67.481144	1.796563	0.993494	0.004484	0.002702
XGBRegressorLog	0.209817	0.298938	0.546752	0.414511	0.414511	65.734211	0.197850	0.994459	0.003965	0.002274
ElasticNetLog	0.149968	0.321580	0.567080	0.454982	0.454982	64.473231	0.006075	0.153862	0.046412	0.350407
GradientBoostingLog	0.112282	0.335837	0.579514	0.450769	0.450769	64.345995	0.588411	0.995638	0.003229	0.001772
Lasso_log	0.094746	0.342471	0.585210	0.466357	0.466357	63.666522	0.014447	0.130417	0.042854	0.360059
DummyRegressor_meanlog	-0.002247	0.379165	0.615764	0.485176	0.485176	62.320522	0.006345	-0.004079	0.004768	0.416932
LinearRegressionLog	-0.005842	0.380525	0.616867	0.485262	0.485262	62.475571	0.009527	0.996783	0.003033	0.001305
DummyRegressor_medianlog	-0.009316	0.381839	0.617931	0.487143	0.487143	62.866267	0.000547	-0.006184	0.009461	0.417956
RandomForestLog	-0.009414	0.381877	0.617962	0.486135	0.486135	62.468049	1.098904	0.994653	0.004324	0.002193
Lasso	0.807785	14615.970357	120.896527	78.212686	78.212686	-308.496143	0.043290	0.999373	0.000826	303.442111
CatBoostRegressor	0.538921	35060.399418	187.244224	107.599907	107.599907	-386.429285	2.106173	0.870699	0.190216	203716.982574





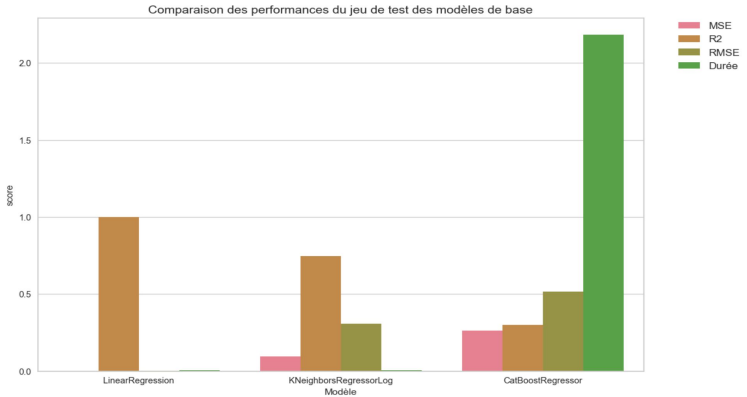
Modélisation : Prédiction de l'émission de Co2 Avec Energy Star Score

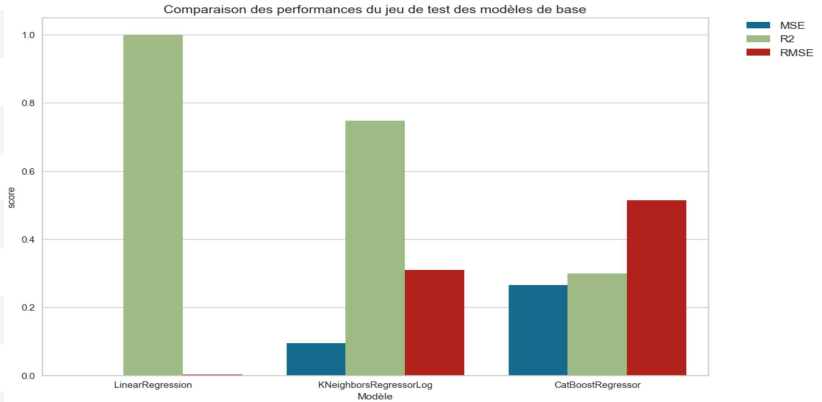
Seattle

Cible **TotalGHGEmissions** ➔RÉGRESSION. Même démarche : split, encodage, standardisation, modèle de base.

Sélections des modèles

Modèle	R2	MSE	RMSE	MAE	Erreur moy	Précision	Durée	Test R2 CV	Test R2 +/-
LinearRegression	1.000000	0.000009	0.003032	0.002653	0.002653	99.987964	0.004874	1.000000	0.000000
KNeighborsRegressorLog	0.747648	0.095468	0.308980	0.231425	0.231425	80.976519	0.007761	0.722896	0.024739
CatBoostRegressor	0.299916	0.264852	0.514638	0.389061	0.389061	68.111193	2.182935	0.993643	0.004060
XGBRegressorLog	0.231348	0.290793	0.539252	0.409883	0.409883	66.196596	0.238021	0.994181	0.003921
ElasticNetLog	0.149968	0.321580	0.567080	0.454982	0.454982	64.473231	0.007669	0.153862	0.046412
GradientBoostingLog	0.121132	0.332489	0.576619	0.449046	0.449046	64.447828	0.511827	0.995513	0.003271
Lasso_log	0.094746	0.342471	0.585210	0.466357	0.466357	63.666522	0.015866	0.130417	0.042854
DummyRegressor_meanlog	-0.002247	0.379165	0.615764	0.485176	0.485176	62.320522	0.000874	-0.004079	0.004768
LinearRegressionLog	-0.005971	0.380574	0.616907	0.485292	0.485292	62.473612	0.007251	0.996776	0.003038
DummyRegressor_medianlog	-0.009316	0.381839	0.617931	0.487143	0.487143	62.866267	0.000563	-0.006184	0.009461



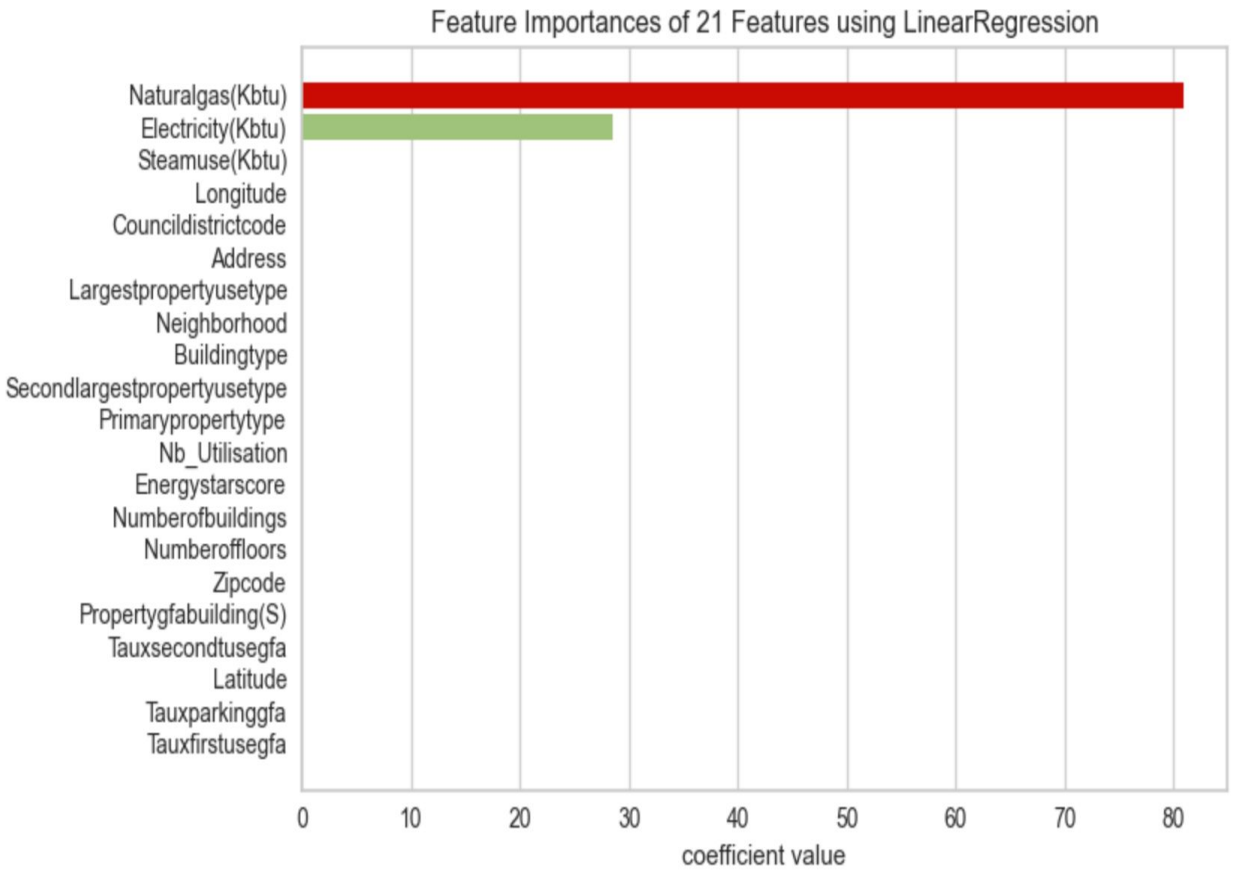


Modélisation : Prédiction de l'émission de Co2

Seattle

Features Importances

Prédiction de l'émission de CO2 AVEC ou SANS Energy Star Score est légèrement différent voire quasi identique.



Conclusion et suite du projet

- Pour la prédiction de l'émission, ils nous faut plus de features pour éviter l'overfitting sur les modèles non-linéaires et linéaires.
- Discussion avec les décideurs pour élargir la récolte et intégrer les bâtiments résidentiels pour plus de performance.
- Intégrer L'ACP en utilisant moins de composantes (5 eboulis?).
- Tester avec les réseaux de neurones.