Projet 4 : Anticipez les besoins en consommation de bâtiments

Eva Rondeau

Seattle

Présentation

<u>Objectif</u> : prédire les émissions de CO2 et la consommation totale d'énergie de bâtiments non destinés à l'habitation pour la ville de Seattle

- 3376 lignes (bâtiments)
- 46 colonnes (informations relevées en 2016)

Seattle

Informations relevés de 2016
Type de bâtiment (résidentiel, non résidentiel, campus, …)
Localisation (latitude, longitude, adresse, …)
Année de construction
Structure (nombre étages, nombre bâtiments, taille)
Types d'utilisations (primaire, secondaire, tertiaire)
ENERGY STAR Score
Sources d'énergie et émissions de gaz à effet de serre

Seattle

Présentation

ENERGY STAR :

- Programme gouvernemental
- Promeut les économies d'énergie (USA, Canada, Australie et UE)
- Environmental Protection Agency (EPA) en 1992
- Réduction des émissions de gaz à effet de serre

ENERGY STAR Score :

- Outil fondé sur données réelles mesurées
- Evaluation du rendement d'un bâtiment
- Caractéristiques : taille, emplacement géographique, isolation, …
- Score calculé entre 1 et 100

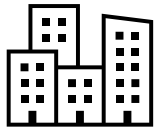
Score ≥ 75

Bâtiment admissible à la **certification**

Seattle

Analyse exploratoire

Bâtiments NON résidentiels



3376 bâtiments (résidentiels + non résidentiels)

1615 bâtiments (non résidentiels)

Traitement données non conformes

- 14 « Missing Data »
- 17 bâtiments « non conformes » dont 16 présentant des outliers
  - 2 valeurs aberrantes élevées
  - 14 valeurs aberrantes faibles
- Valeurs négatives ou nulles

→ **Suppression de ces données**

Traitement données manquantes

LargestPropertyUseType	2nd et 3rd PropertyUseType (GFA)	ENERGYSTARScore	SiteEUIWN(kBtu/sf) et SiteEnergyUseWN(kBtu/sf)	ZipCode	NumberofBuildings	YearsENERGYSTARCertified
Valeur la plus fréquente en se basant sur le type de propriété principal	« Lack Data » (0)	-	0	Utilisation de l'adresse pour obtenir le ZipCode correspondant (Nominatim: géocodage)	Suppression (non conformes)	-

Seattle

Analyse exploratoire

Feature Engineering

- Création de nouvelles variables à partir des données
- But : améliorer la performance des modèles d'apprentissage automatiques.

SURFACE TOTALE	AGE	CONSOMMATION ET ÉMISSIONS PAR SURFACE	TYPES DE BÂTIMENTS
= LargestPropertyUseTypeGFA + 2ndLargestPropertyUseTypeGFA + 3rdLargestPropertyUseTypeGFA	= Année de prise des mesures (2016) – Année de construction	= Consommation/Emissions % Superficie totale	Regroupement en plus grands ensembles de types de bâtiments

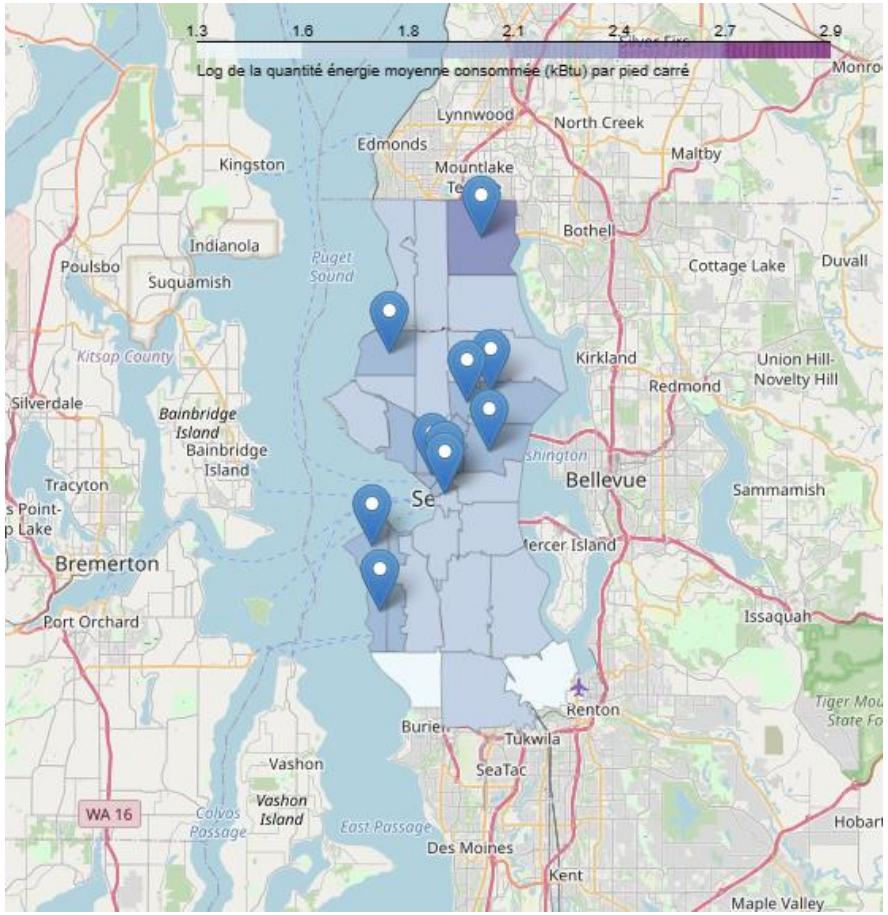
PROPORTIONS ÉNERGIES (VAPEUR, ÉLECTRICITÉ, GAZ)	SURFACE PAR ÉTAGE	SURFACE PAR BÂTIMENT	TYPES D'ÉNERGIE LA PLUS UTILISÉE
= Vapeur + Electricité + Gaz % Consommation totale *100	= Surface du bâtiment % Nombre d'étages	= Surface du bâtiment % Nombre de bâtiments	Attribution de valeurs (selon l'énergie le plus utilisée) : 1 -> Electricité 2 -> Gaz Naturel 3 -> Vapeur

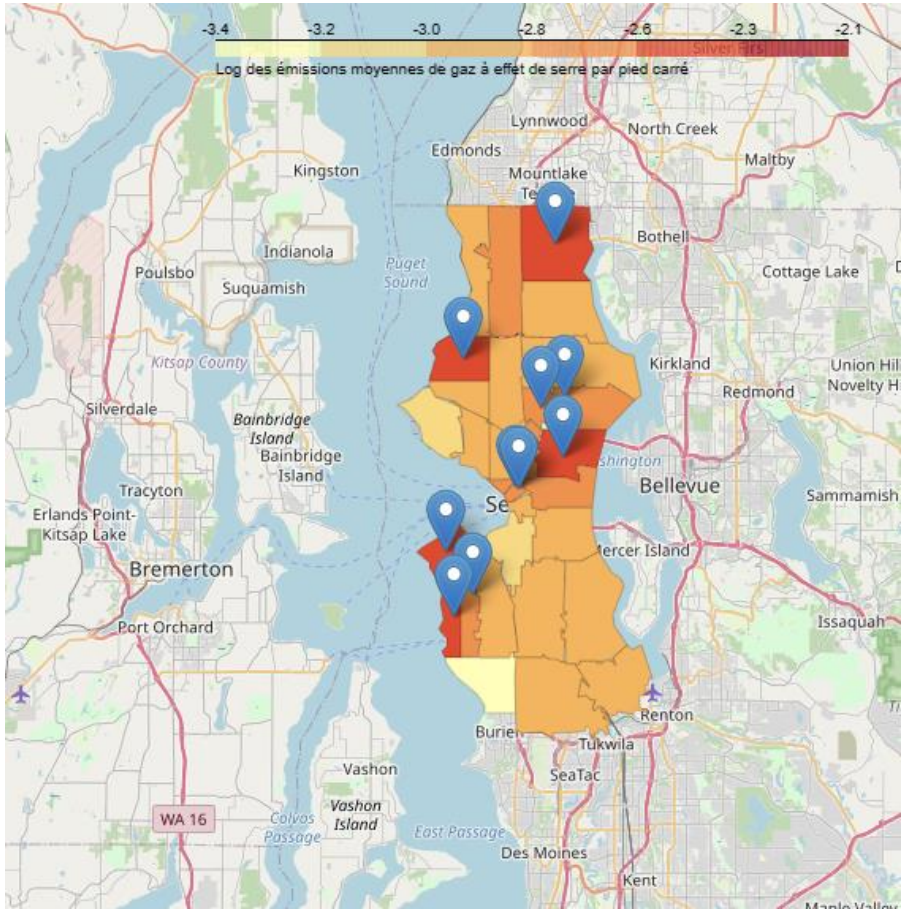
Seattle

Analyse exploratoire

Cartographies

Consommation d'énergie moyenne (à gauche) et émissions moyennes (à droite) par pied carré et affichage des 10 codes postaux les plus consommateurs d'énergie et émetteurs de CO2



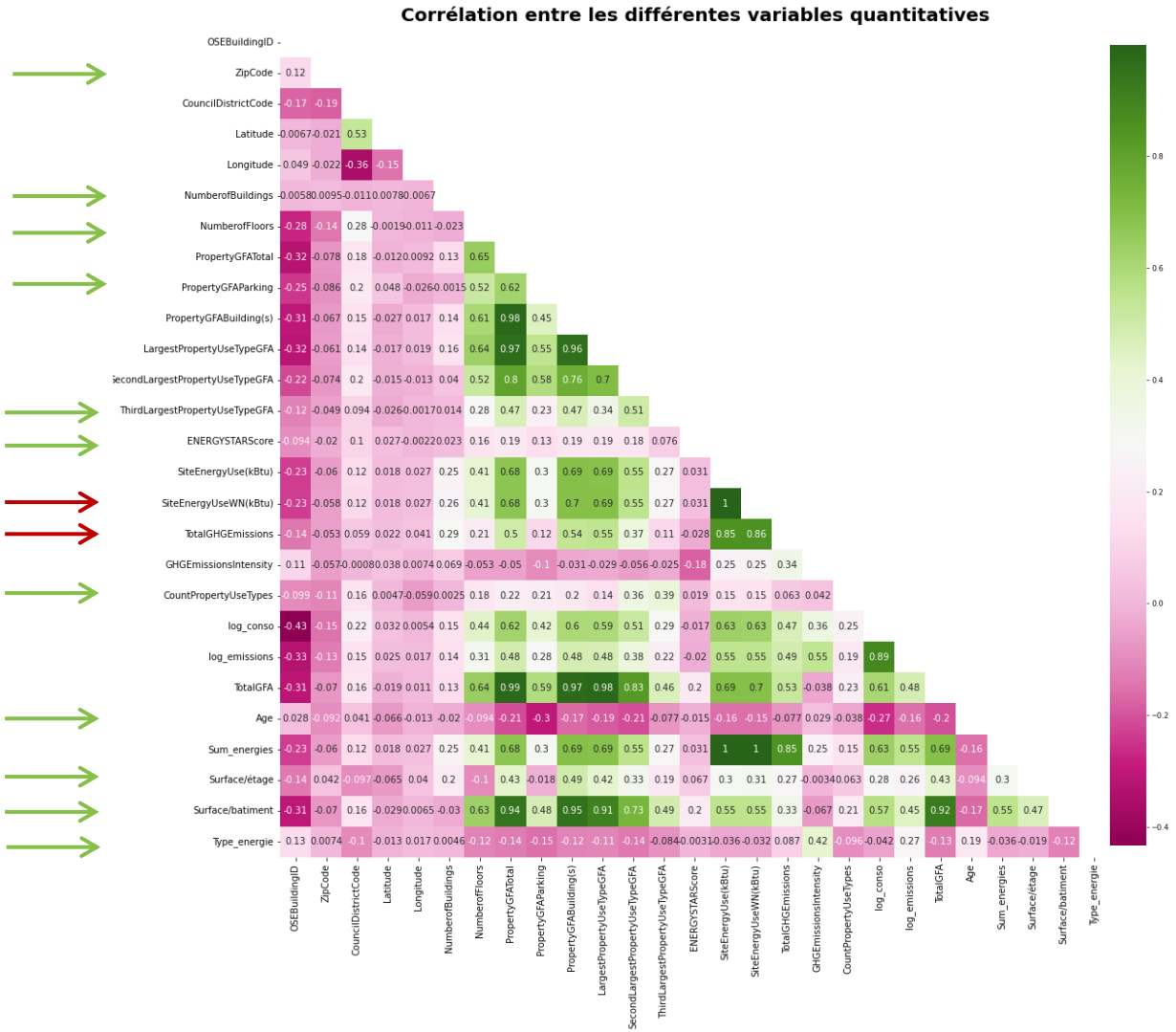


Seattle

Modélisation

1. Variables corrélées (> 0,7) non prises en compte pour la modélisation : risque de **sur-apprentissage**

2. 11 variables quantitatives + 2 variables qualitatives
   - PropertyType
   - Neighborhood



Variables cibles (targets) →
Caractéristiques (features) →

Seattle

Modélisation

Encodage variables qualitatives

Impossible de traiter directement les variables catégorielles : **encodage** nécessaire.

- Méthode **get_dummies**:

Transformation variables catégorielles en plusieurs colonnes binaires (1 colonne = catégorie unique)
Valeur 1 ou 0 : catégorie présente ou non

OSEBuildingID	Neighborhood_DOWNTOWN	Neighborhood_EAST	Neighborhood_GREATER DUWAMISH	Neighborhood_LAKE UNION	Neighborhood_MAGNOLIA / QUEEN ANNE	Neighborhood_NORTH
1	1	0	0	0	0	0
2	1	0	0	0	0	0
3	1	0	0	0	0	0
5	1	0	0	0	0	0
8	1	0	0	0	0	0

- **Retirement d'une modalité** (k-1) (Rakotomalala, 2015) :

Evite la redondance et la colinéarité (corrélation entre plusieurs variables indépendantes)

- PropertyType : 12 modalités ⟹ 11 modalités
- Neighborhood : 13 modalités ⟹ 12 modalités

Seattle

Modélisation

Sélection variables pertinentes

A partir de StatsModels :

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.6752	0.050	133.878	0.000	6.577	6.773
Age	-0.0930	0.011	-8.132	0.000	-0.115	-0.071
PropertyGFAParking	0.0504	0.013	3.972	0.000	0.026	0.075
ENERGYSTARScore	-0.0945	0.011	-8.993	0.000	-0.115	-0.074
ZipCode	-0.0229	0.011	-2.114	0.035	-0.044	-0.002
→ ThirdLargestPropertyUseTypeGFA	-0.0066	0.012	-0.543	0.587	-0.031	0.017
NumberofBuildings	0.0798	0.010	7.631	0.000	0.059	0.100
NumberofFloors	0.0419	0.017	2.496	0.013	0.009	0.075
CountPropertyUseTypes	0.0695	0.012	5.718	0.000	0.046	0.093
Surface/étage	0.0613	0.015	4.153	0.000	0.032	0.090
Surface/batiment	0.1962	0.020	9.705	0.000	0.157	0.236
Type_energie	0.0556	0.011	5.184	0.000	0.035	0.077
→ PropertyType_Food Store	0.0782	0.055	1.434	0.152	-0.029	0.185
PropertyType_Hospital_Medical Office	0.1756	0.074	2.373	0.018	0.030	0.321
→ PropertyType_Hotel	0.1015	0.063	1.621	0.105	-0.021	0.224
PropertyType_Laboratory	0.6685	0.143	4.665	0.000	0.387	0.950
PropertyType_Mixed Use Property	-0.1601	0.062	-2.584	0.010	-0.282	-0.039
→ PropertyType_Office	-0.0403	0.048	-0.845	0.398	-0.134	0.053

Sélection des coefficients les plus pertinents (p-value < 0,05)

Méthode de **sélection des variables** :

Réduction des coûts (temps de calcul, espace mémoire)

☆☆☆ Amélioration de la qualité d'apprentissage

Transformation logarithmique

Variables cibles (targets) transformées par le logarithme ⟶ **Amélioration performance du modèle**

Seattle

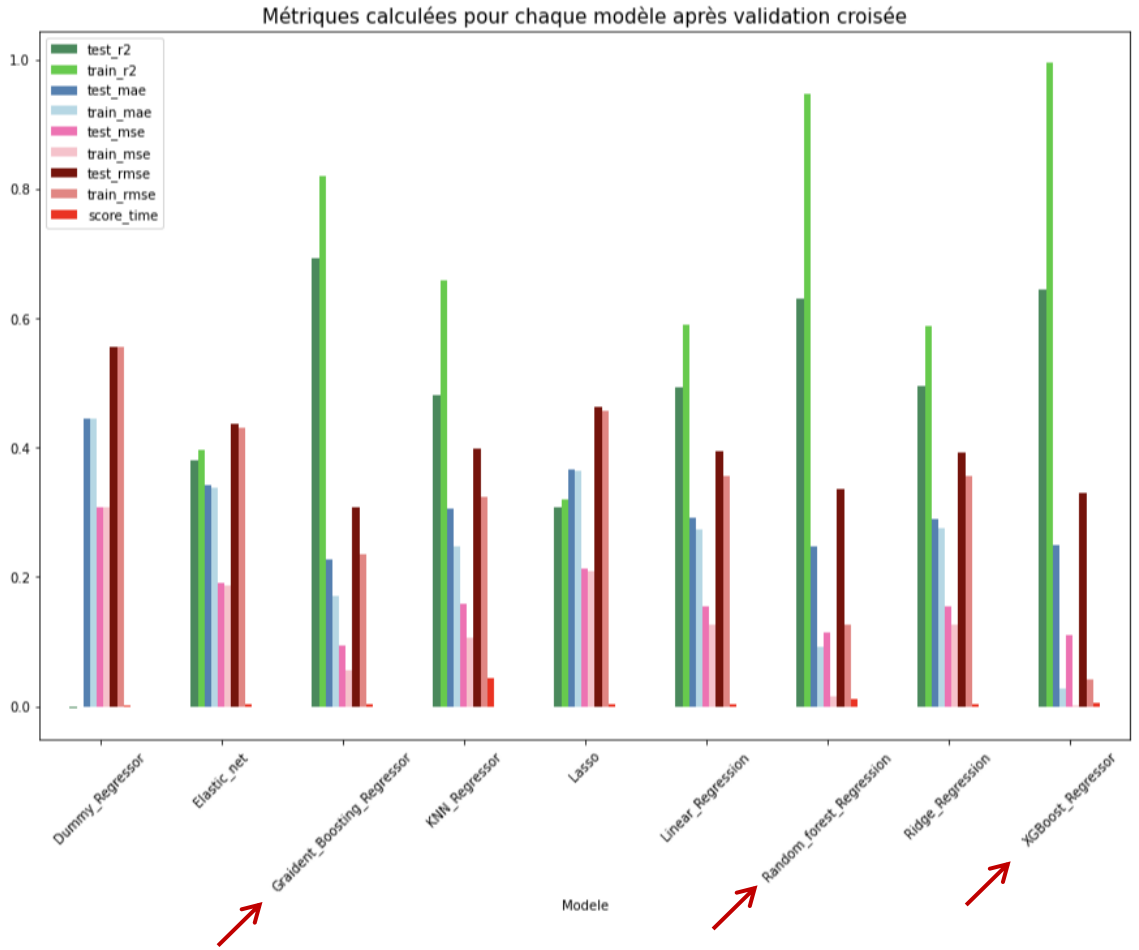
Modélisation

1. Consommation totale d'énergie

	Modele	fit_time	score_time	test_r2	train_r2	test_mae	train_mae	test_mse	train_mse	test_rmse	train_rmse
0	Dummy_Regressor	0.001258	0.002471	-0.002017	0.000000	0.444410	0.444011	0.308688	0.308425	0.555597	0.555361
1	Elastic_net	0.003014	0.003251	0.393504	0.411474	0.339307	0.335489	0.186771	0.181471	0.432170	0.425994
2	Graident_Boosting_Regressor	0.142229	0.003172	0.708473	0.838874	0.221093	0.164347	0.089439	0.049680	0.299063	0.222890
3	KNN_Regressor	0.002219	0.040531	0.527090	0.683202	0.291982	0.238879	0.145152	0.097704	0.380988	0.312577
4	Lasso	0.003497	0.004201	0.307867	0.320736	0.366403	0.363347	0.213486	0.209462	0.462046	0.457671
5	Linear_Regression	0.005002	0.004801	0.493858	0.589506	0.290946	0.274338	0.154903	0.126575	0.393577	0.355774
6	Random_forest_Regression	0.471662	0.010319	0.677307	0.955405	0.233672	0.086102	0.099244	0.013751	0.315031	0.117264
7	Ridge_Regression	0.004090	0.004800	0.495976	0.588955	0.290221	0.274435	0.154248	0.126745	0.392745	0.356012
8	XGBoost_Regressor	0.089836	0.005801	0.637600	0.995802	0.249712	0.023940	0.111170	0.001292	0.333422	0.035944

Valeurs de scores des données d'entrainement les plus élevées :

- Gradient Boosting Regressor
- Random Forest Regressor
- XGBoost Regressor

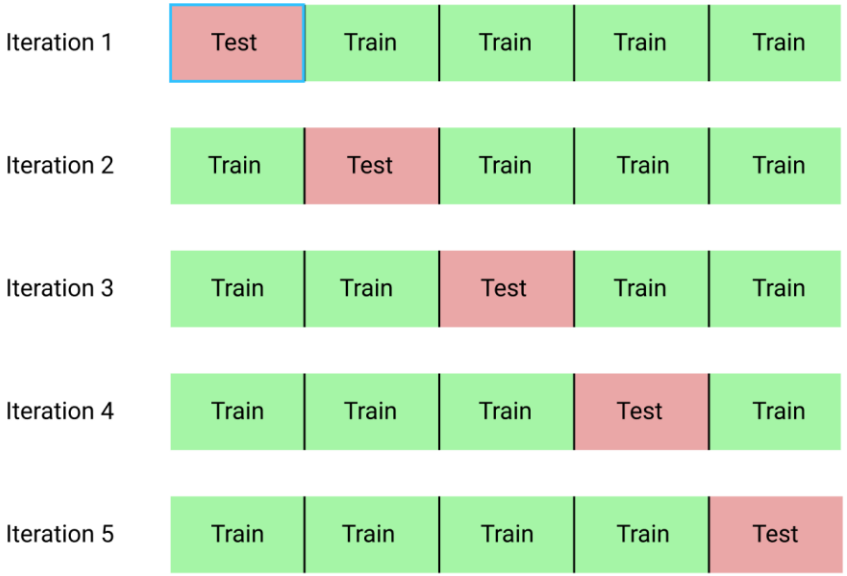


Seattle

Modélisation

Validation croisée

- Evalue les performances d'un modèle sur les nouvelles données
  - Division du jeu de données en plusieurs parties (= folds)
    - Entrainement sur une partie des données
- Scores moyennés : estimation finale de la performance du modèle

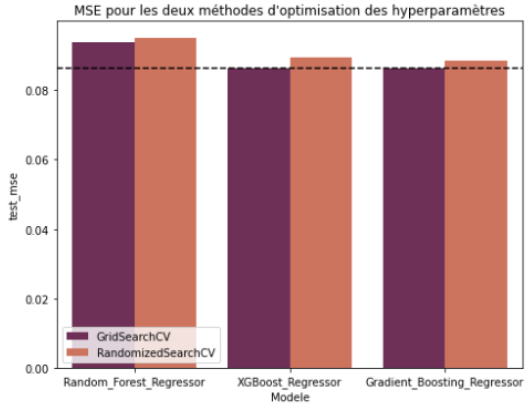


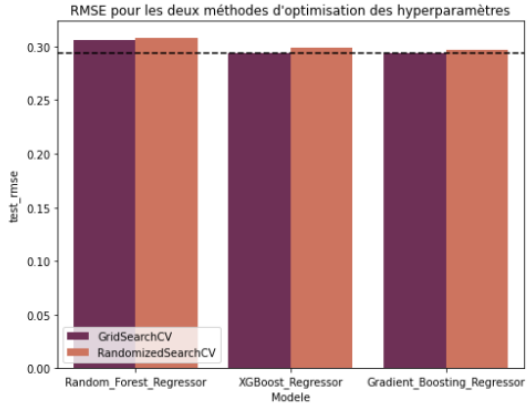
Seattle

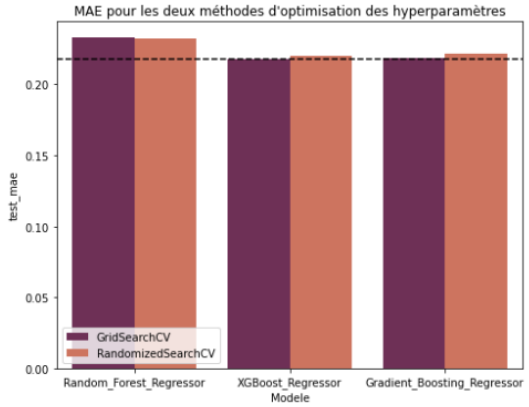
Modélisation

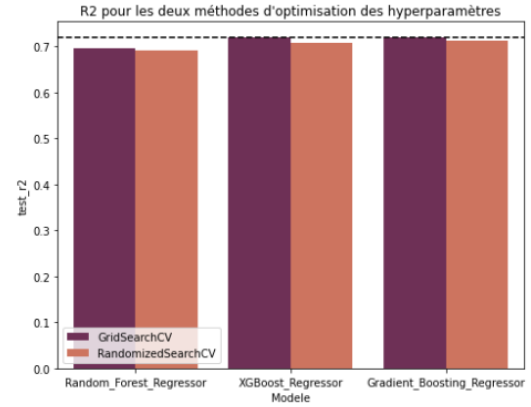
1. Consommation totale d'énergie

Optimisation hyperparamètres : GridSearchCV *vs.* RandomizedSearchCV

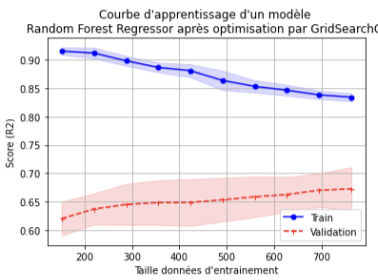


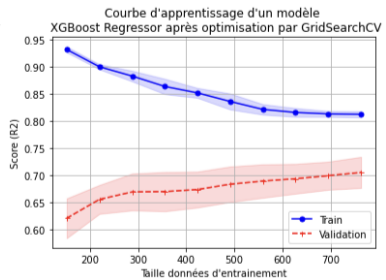


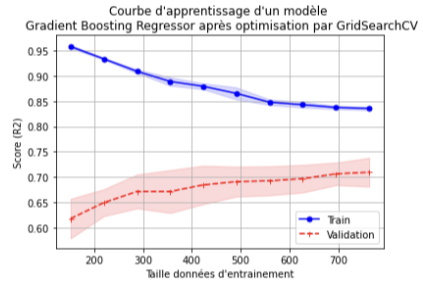




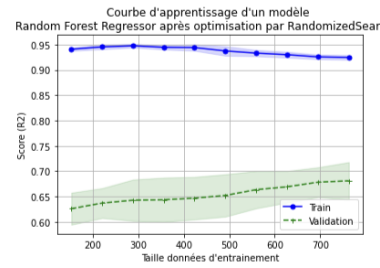
GridSearchCV

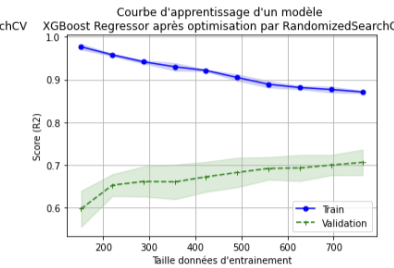


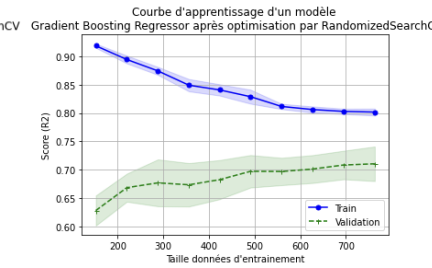




RandomizedSearchCV







⇒ Modèle choisi : **Gradient Boosting Regressor**

Seattle

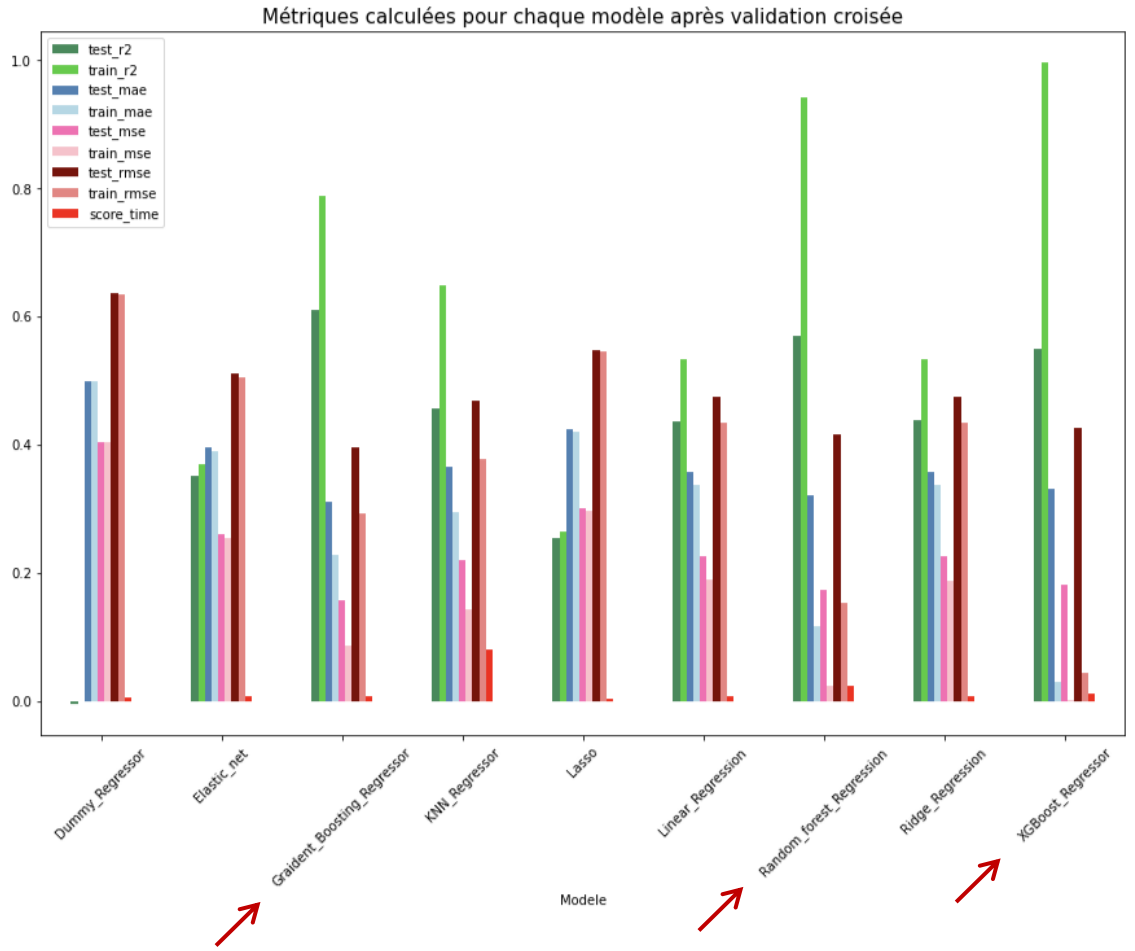
Modélisation

2. Emissions de CO2

	Modele	fit_time	score_time	test_r2	train_r2	test_mae	train_mae	test_mse	train_mse	test_rmse	train_rmse
0	Dummy_Regressor	0.001600	0.004801	-0.004242	0.000000	0.498401	0.497918	0.403787	0.403066	0.635443	0.634875
1	Elastic_net	0.006401	0.008047	0.351275	0.370219	0.394744	0.389869	0.260971	0.253785	0.510853	0.503771
2	Graident_Boosting_Regressor	0.343469	0.006445	0.609127	0.788731	0.310173	0.227567	0.156589	0.085122	0.395713	0.291757
3	KNN_Regressor	0.001662	0.080663	0.456161	0.647548	0.365510	0.294444	0.218993	0.142092	0.467967	0.376951
4	Lasso	0.006400	0.003200	0.253149	0.264246	0.422933	0.420347	0.300494	0.296529	0.548174	0.544545
5	Linear_Regression	0.008149	0.007905	0.435135	0.533009	0.356886	0.336406	0.225822	0.188182	0.475207	0.433799
6	Random_forest_Regression	1.083294	0.024044	0.570052	0.941693	0.319869	0.116572	0.172509	0.023497	0.415342	0.153287
7	Ridge_Regression	0.006400	0.008001	0.437583	0.534054	0.356267	0.335915	0.224965	0.187758	0.474305	0.433311
8	XGBoost_Regressor	0.166799	0.011398	0.548899	0.995446	0.331011	0.029161	0.181017	0.001836	0.425461	0.042853

Valeurs de scores des données d'entrainement les plus élevées :

- Gradient Boosting Regressor
- Random Forest Regressor
- XGBoost Regressor

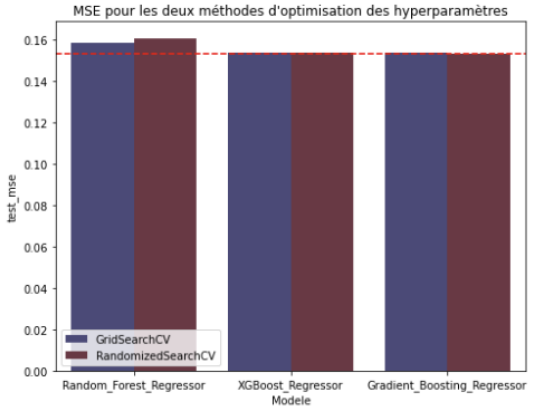


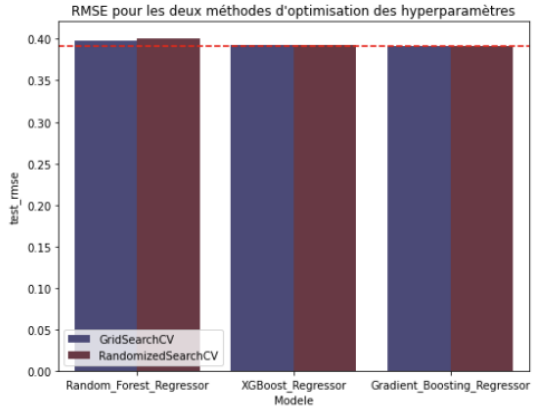
Seattle

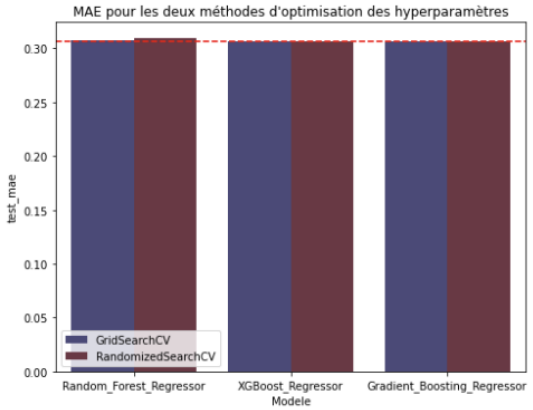
Modélisation

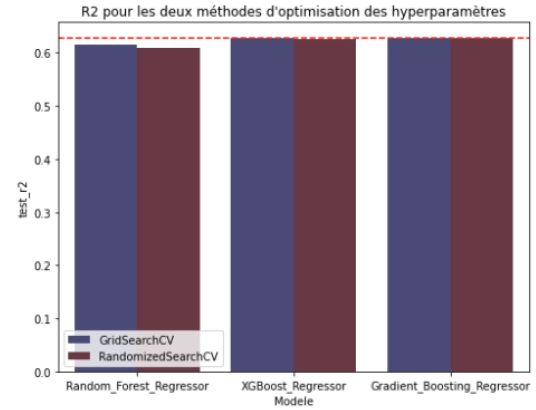
2. Emissions de CO2

Optimisation hyperparamètres : GridSearchCV *vs.* RandomizedSearchCV

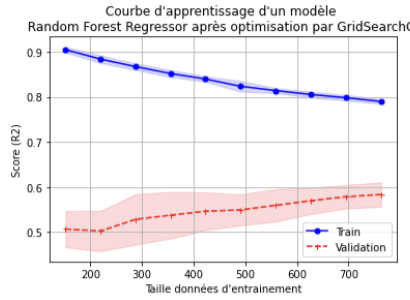


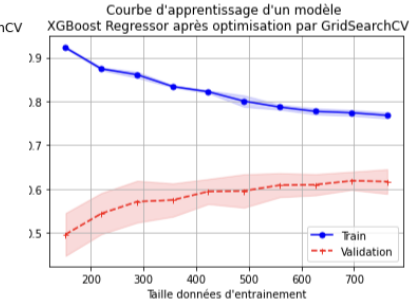


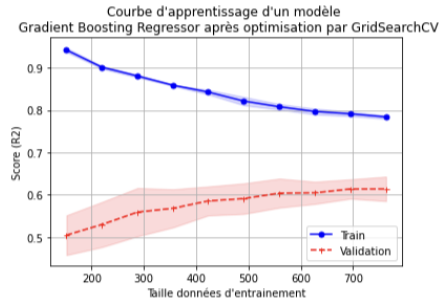




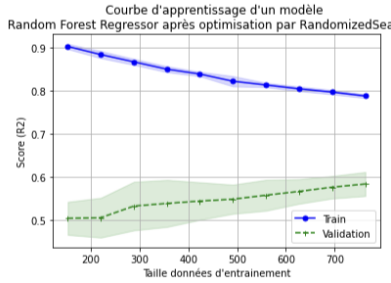
GridSearchCV

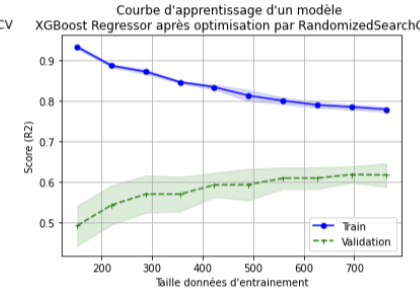


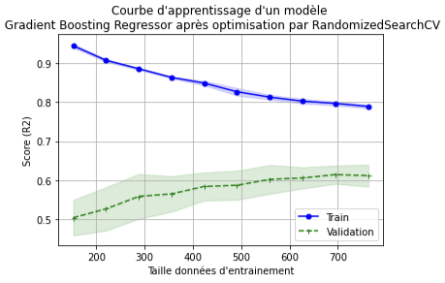




RandomizedSearchCV







⇒ Modèle choisi : **Gradient Boosting Regressor**

Seattle

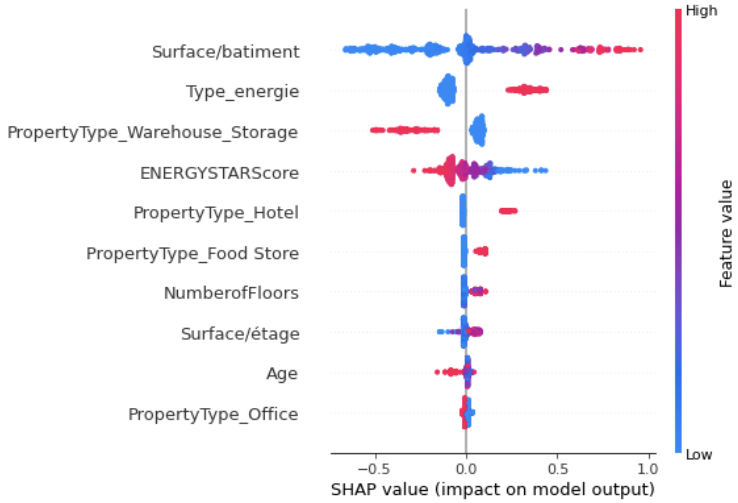
Modélisation

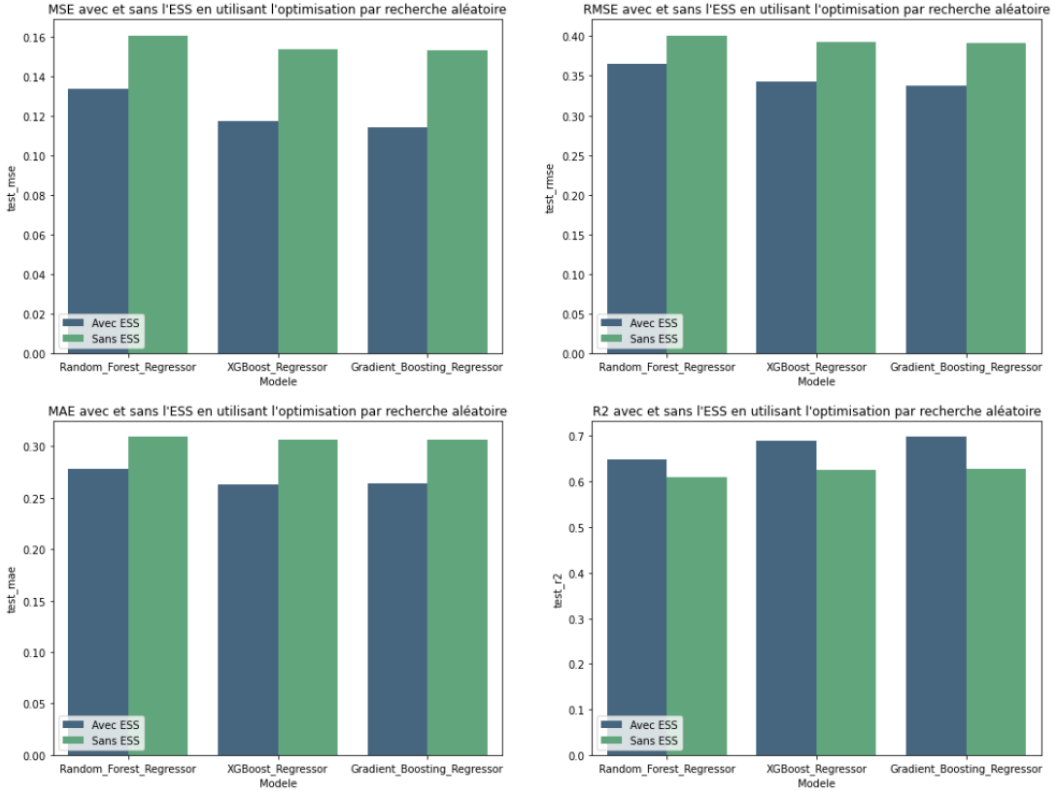
2. Emissions de CO2

Intérêt de **l'ENERGY STAR Score** pour la prédiction d'émissions

Valeurs SHAP

Valeurs de SHAP: identification et visualisation des variables ayant le plus contribué à chaque prédiction



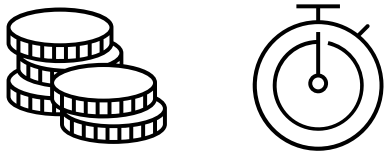


Seattle

Conclusion

- **Intérêt de l'ENERGY STAR Score** pour la prédiction des émissions de CO2.

- Possible d'effectuer des prévisions à partir de données antérieures pour anticiper le futur
- Objectif : ville neutre en carbone en 2050



- Modèles de prédiction : meilleure **efficience** des prévisions futures
- Evite déplacements et mesures **coûteuses** et **chronophages**