Neutralité carbone 2050 - Seattle

A partir de 2 campagnes de relevés (2015 et 2016) :

- ✓ Prédire les émissions de CO2 et la *consommation totale d'énergie*
- ✓ Evaluer l'intérêt de l'EnergyStar Score pour la prédiction d'émissions

Seattle
Washington, États-Unis

Seattle

Sommaire & livrables

- Problématique & pistes
- 0 • Merge
- 1 • Exploration - Résultats
- 2 • Exploration - Profils
- 3 • Préparation - Feature Engineering
- 4 • Modélisation
- Conclusions

6 livrables en plus du présent support

- 0_Merge.ipynb
- 1_Results.ipynb
- 2_Profiles.ipynb
- 3_Prepare.ipynb

Les 4 notebooks

- 4_Models.ipynb

Le Code

Pélec_03_Restultats.xlsx

Les Résultats

Problématique et pistes

Comment contribuer à la finalité de neutralité carbone?

- **Problématiques ouvertes :**
  - Choisir la bonne cible en termes de « *consommation totale d'énergie* »? En particulier :
    - **Total** *versus* **Intensité**? i.e. « rapporter » à la surface l'énergie consommée sur un site
    - **Source** *versus* **Site**? i.e. « remonter » au moyens de productions et acheminement de l'énergie sur un site
  - Déterminer les meilleures données de « profil » caractérisant les bâtiments, en ratio (Gain prédiction) / (Coût de collecte)
  - Prioriser les bâtiments ***non destinés à l'habitation :***
    - Définir à quoi correspondent ces bâtiments?
    - Ecarter une partie des données ou au contraire arbitrer les données à prendre en compte selon le gain de prédiction?

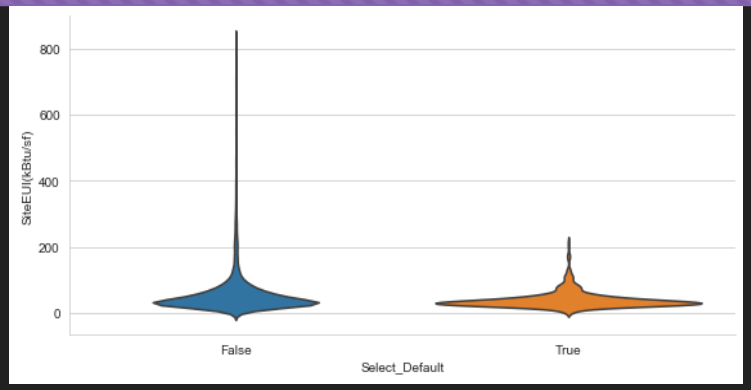
- **Piste :**
  - Apprendre l'intensité d'énergie type pour un usage donné, pour étalonner les bâtiments (principe de l'ENERGYSTARScore)
    - Conserver l'information de l'**usage de la surface** occupée.
    - Obtenir des informations de surface cohérentes (ex. « Building » versus « Parking »)

0 - Merge

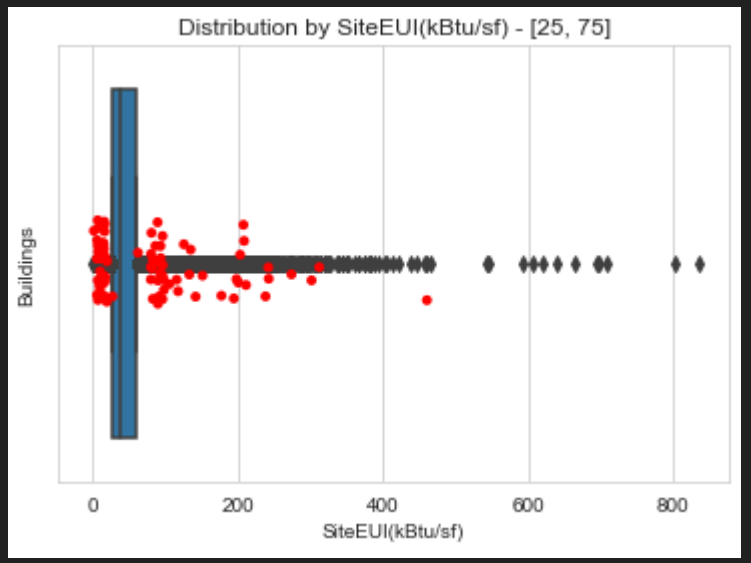
Comparaison et concaténation des relevés 2015 et 2016

- **Identifier** les features disjoints, les rendre homogènes (ex. adresse) et les rationaliser
- **Conserver** les relevés « en double ».
  - Evolutions liées à la façon de prendre en compte les parkings.
- **Restreindre** aux données valides :
  - **Garder** les données **Compliant** (parmi : 'Not Compliant', 'Error - Correct Default Data', 'Missing Data', 'Non-Compliant'],
  - **Garder** les valeurs par défaut : **True** (*nb valeurs issues du Score EnergyStar**)
  - **Conserver** le feature « **Outliers** » pour usage ultérieur : sélection avec ou sans « outliers » (*nb : quartiles*).
  - **Isoler** un jeu de données pour validation (ex. *données des relevés non recouvrées, répartition représentative au sens de la cible, etc.*).

* On pourra tester si le retrait de valeurs « default » == 'True' ne dégrade pas la prise en compte de l' *EnergyStar Score*



Les valeurs par défaut restreignent la distribution de l'énergie

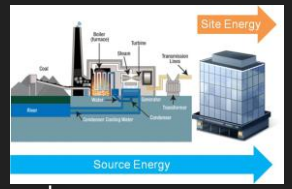


Les outliers des relevés sont 1er – dernier quartile et sont incomplets

1 – Exploration - Résultats

Résultats : mesures de l'énergie consommée et des émissions (finaux et composition)

- **Utiliser** les corrélations pour éliminer les cibles « redondantes » (WN : Normalisation Météo)
- **Pré-sélectionner** une cible entre Energie ou Emission **Totale** versus **Intensité**
  - Une **Intensité** caractérise le bâtiment indépendamment de sa taille
  - Mais la **Surface** est introduite par calcul (cf. portfolio)

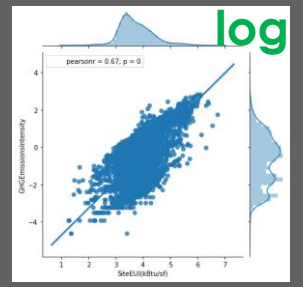


- **Choisir** entre données **Site** versus **Source**
  - Les données **Site** favorisent les caractéristiques intrinsèques de bâtiments.
- **Favoriser** les meilleures distributions et les meilleurs niveaux de corrélation
  - Limiter l'asymétrie des distributions via **passage au log**
  - Tester l'effet des **suppressions d'Outliers** au sens statistique sur ces corrélations
  - Les faibles émissions sont **sur-représentées** : *introduire des catégories de profil?*
- **Dériver** une information : hasNaturalGas et hasSteam
  - *les sites sans électricité et avec fuel étant marginaux*

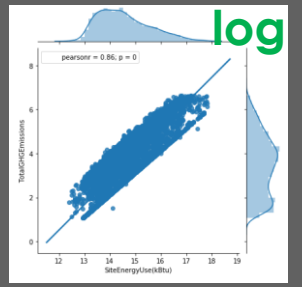
En résumé :

	SiteEnergyUse(kBtu)	TotalGHGEmissions	SiteEUI(kBtu/sf)	GHGEmissionsIntensity
SiteEnergyUse(kBtu)	1.00	0.79	0.45	0.24
TotalGHGEmissions	0.79	1.00	0.51	0.55
SiteEUI(kBtu/sf)	0.45	0.51	1.00	0.75
GHGEmissionsIntensity	0.24	0.55	0.75	1.00

Niveaux de corrélations des features résultats

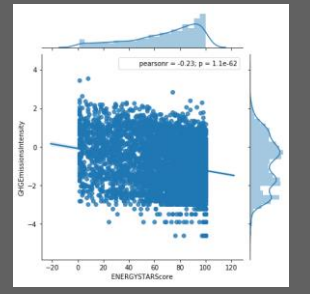


SiteEUI(kBtu/sf) x GHGEmissionsIntensity

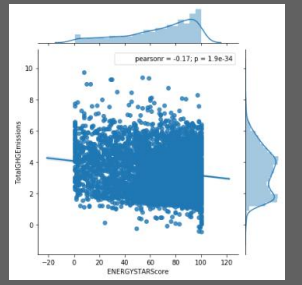


SiteEnergyUse(kBtu x TotalGHGEmissions

L'EnergyStarScore est corrélé négativement aux émissions:



GHGEmissionsIntensity



TotalGHGEmissions

2 – Exploration – Profils (1/3)

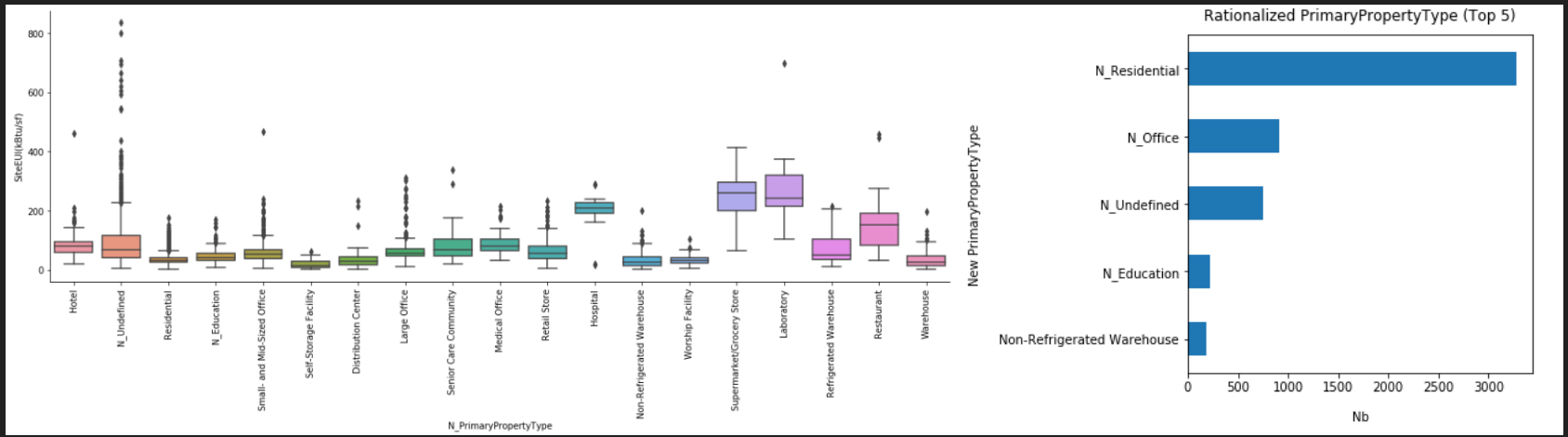
Profils : informations qui ne nécessitent pas de relevés pour être connues

2.1. Types et Usages

- **Comparer** le Type vs usage dominant :
  - On trouve ***habitation familiale*** (PrimaryPropertyType – […]Multifamily), pour type ***non résidentiel*** (BuildingType - NonResidential)
- **Observer** et **rationaliser** les usages détaillés :
  - L'usage dominant calculé (> 50% parmi ceux détaillés en 1er, 2ème et 3ème usage, par ordre d'importance en surface) est déterminant.

En résumé :

- Filtre **moins restrictif** basés sur les usages détaillés pour isoler les bâtiments « non résidentiels »
- **2 pertes d'information** (hors rationalisation) :
  - Usage Other ou Mixed-used (les + dispersés)
  - Perte d'information de surfaces (spécifiées dans les 1er, 2ème et 3ème usages).



Effet des usages dominants sur l'intensité

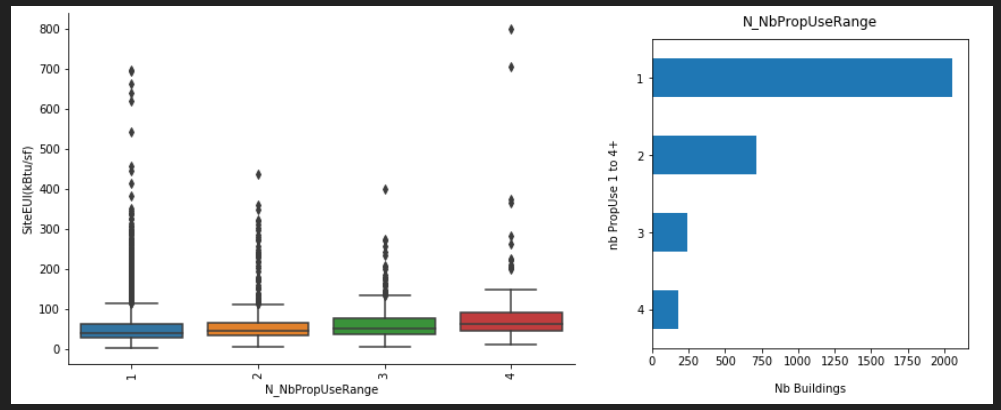
Répartition des usages dominants

2 – Exploration – Profils (2/3)

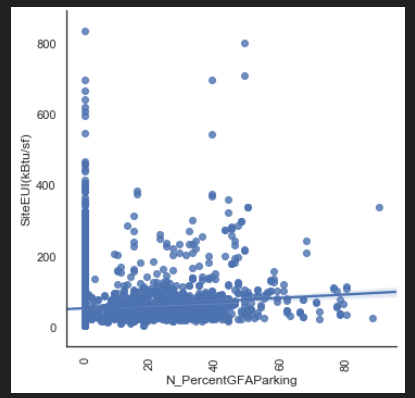
2.2. Usages combinés aux Surfaces

- Distinction Parking versus Building(s) :
  - Cruciale en cas de cible intensité d'énergie
  - **Contre-intuitif :** l'intensité d'énergie augmente avec la surface de parking
  - La surface reconstruite (3 usages) **est parfois supérieure** à la surface totale

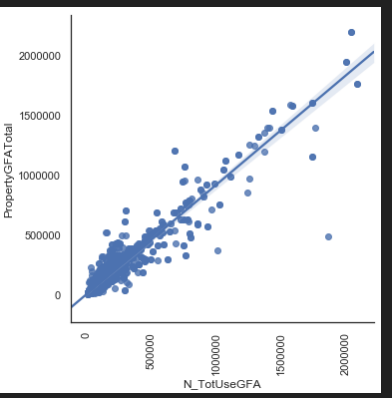
- Usages détaillés :
  - Une soixantaine d'usages connus
  - **Information dérivée**, nb d'usages : 1, 2, 3, et « 4 et + »



Effet des nb d'usages sur l'intensité | Décompte des usages



Effet sur l'intensité, du % surface parking déclarée



Ecarts surface reconstruite vs total délcaré

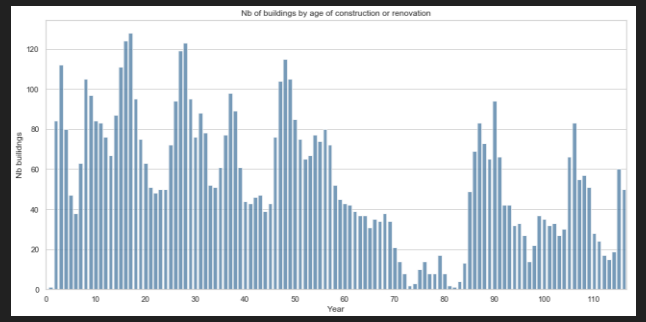
En résumé :

- Usage détaillé (hors « Other ») et nb d'usages **à valoriser**,
  - Vigilance sur les effectifs par usage et cas non spécifiés (2%)
- Conséquences négatives sur les intensités :
  - **Incohérence des surfaces**

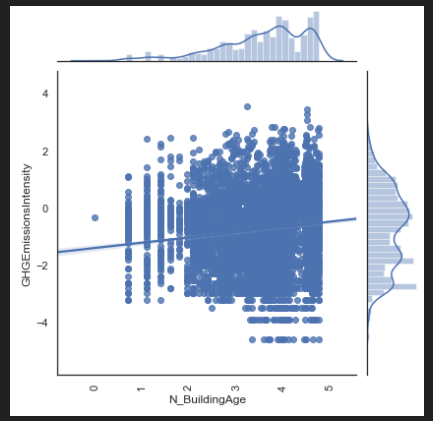
2 – Exploration – Profils (3/3)

2.3. Age, composition et emplacement

- Age de construction ou rénovation des bâtiments.

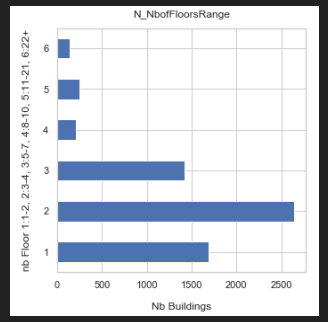


Répartition par âge du bâtiment

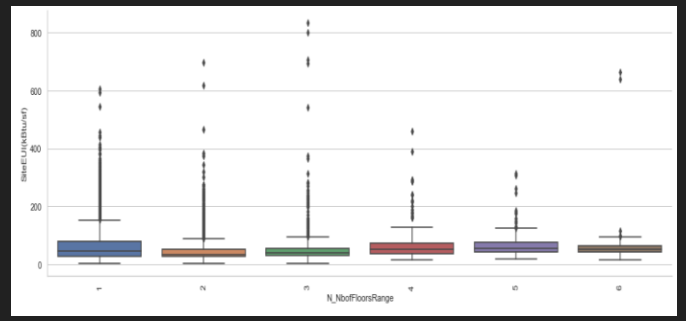


Effet de l'âge du bâtiment sur l'intensité

- Composition en nb de bâtiments et nb d'étages (et regroupements par catégories)

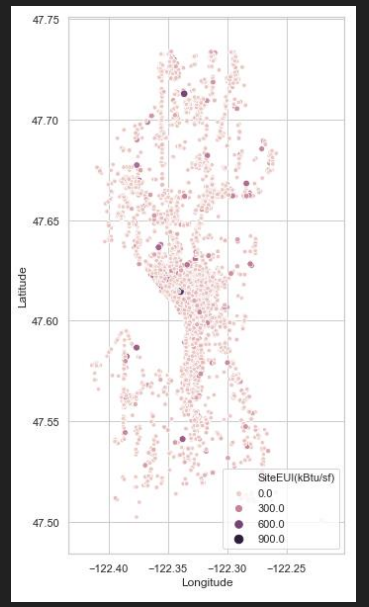


Répartition par nb d'étages

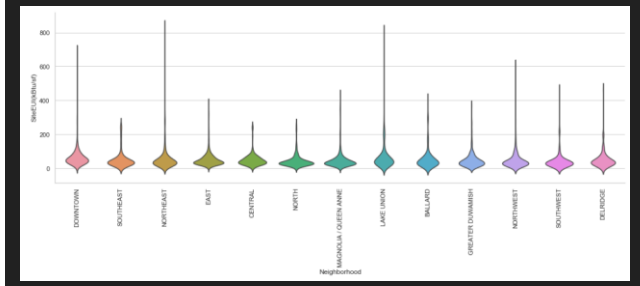


Effet du nb d'étage sur l'intensité

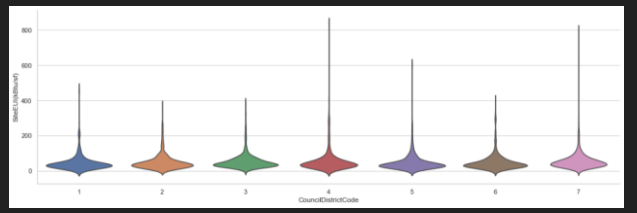
- L'emplacement:
  - Latitude et Longitude
  - Découpages territoriaux
    - Neighborhood
    - District Code



Latitude & Longitude



Neighborhoods



Disctricts

3 – Préparation

Récapitulatif du Feature Engineering :
Traitements préparatoires et validité des données

- Le feature engineering abouti à plusieurs familles de données :
  - Pertinence testée par ajouts (regression) ou retrait (arbres) :

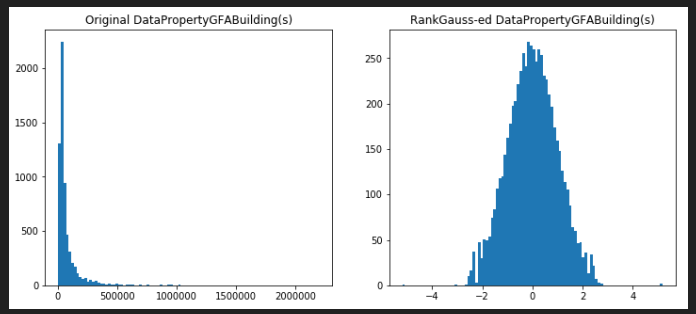
- **1.** Encoder (One Hot) les **features catégoriels**
  - BuildingType, CouncilDistrictCode, Neighborhood, N_PrimaryPropertyType (rationnalisé)

- **2.** Créer la **UseTypeList** des Usages uniques détaillés définis dans les relevés.
  - Affecter la part de surface allouée à l'usage
  - « **Usage x surfaceUsage / surfaceUtile** »

- **3.** Transformer des données **numériques** brutes : Rank Gauss

- **4.** Les features **pseudo-numériques**, dérivés et utilisables en l'état
  - Ex. nb usages, nb étages, nature énergie, etc.

- \+ Focus sur la validité des features dans le cas de régressions
  - Corrélations de Pearson ou Kendall avec les 4 cibles
  - SiteEnergyUse et TotalGHGEmissions est le binôme cible le plus judicieux



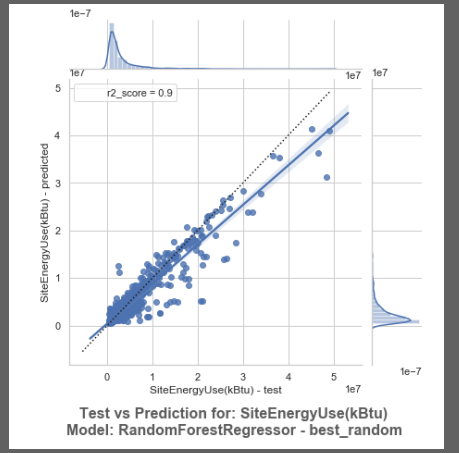
Exemple de transformation Rank Gauss

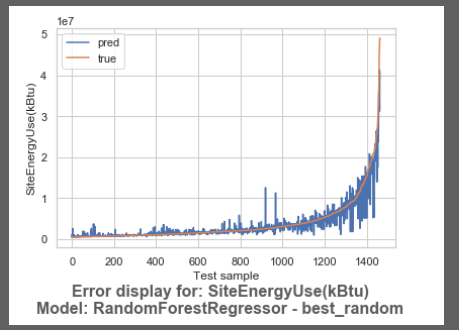
3 – Modélisation (1/3)

3.1. Méthode

- **Création d'une fonction intégrée :**
  - Entrainement,
  - Prédictions,
  - Calcul des scores : **R2**, rmse, mae
  - Affichage predict vs test
  - Features « importance » :
    - observation en direct dans le cas d'un Random Forest,
    - observation sur les coefficients en Régression
- **Stratification** des splits (train-test & folds) par Building Type
- **Sélection du scope et des cibles**
  - Tests de pertinence des cas de feature engineering

Illustration :





	feature_names	feature_importances
9	N_RG_PropertyGFATotal	0.255126
11	N_RG_PropertyGFABuilding(s)	0.192328
12	N_RG_LargestPropertyUseTypeGFA	0.188807
13	N_RG_SecondLargestPropertyUseTypeGFA	0.061784
8	N_RG_NumberofFloors	0.042114
2	N_D_BuildingType	0.041354
15	N_RG_Latitude	0.032239
5	N_D_NbofFloorsRange	0.029335
16	N_RG_Longitude	0.028907
6	N_RG_YearBuilt	0.028632
0	N_D_hasNaturalGas	0.025951
17	N_RG_N_BuildingAge	0.025812
10	N_RG_PropertyGFAParking	0.025140
14	N_RG_ThirdLargestPropertyUseTypeGFA	0.011758
3	N_D_NbPropUseRange	0.008471
1	N_D_hasSteam	0.001541
7	N_RG_NumberofBuildings	0.000476
4	N_D_NbofBuildingsRange	0.000226

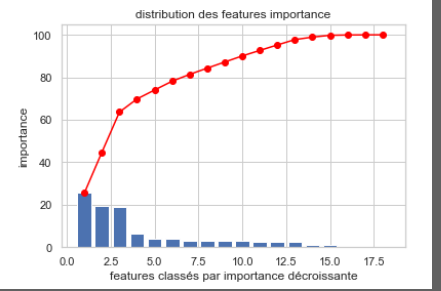


Illustration pour SiteEnergyUse – RandomForest, optimisé par Randomized Search CV

3 – Modélisation (2/3)

3.2. Approche & Principaux résultats

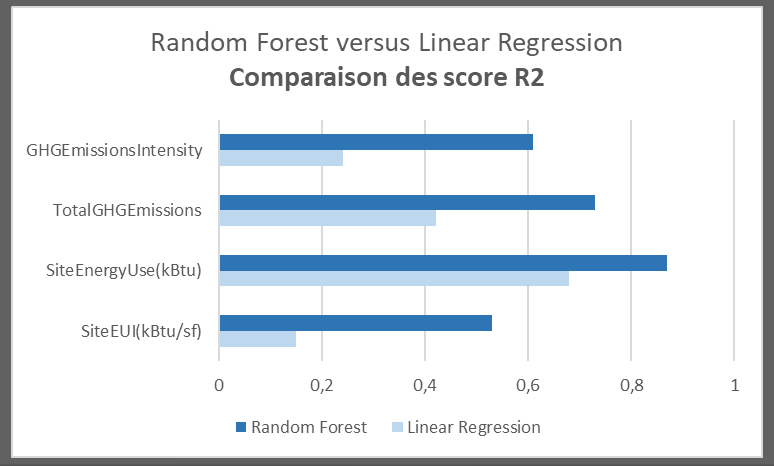
- **Balayer les cibles :**
  - Comparer les performance :

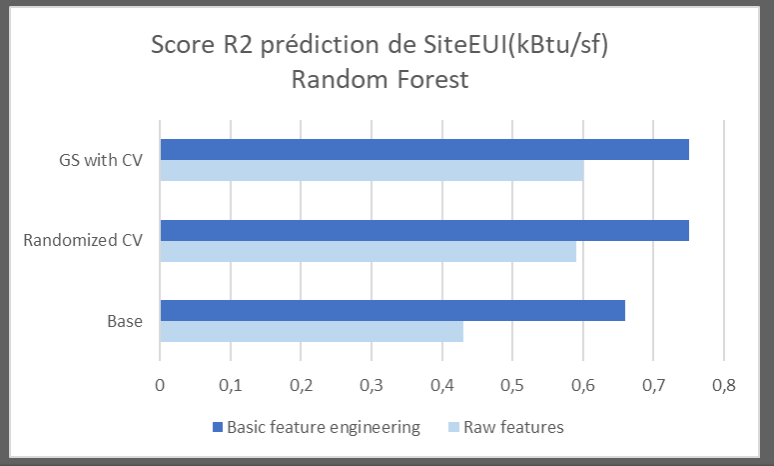
    Meilleures pour les Totaux {SiteEnergyUse, TotalGHGEmission}
- **Tester la pertinence** d'une approche **régression linéaire** :
  - Ajouter les features pertinents
  - Comparer au Random Forest : **perdant y compris en conditions favorables**
  - **Améliorer avec CV Ridge et Lasso : ne rattrape pas le Random Forest**

- Estimer l'**apport du feature engineering**
  - **C**omparer features bruts versus traitements simples:

    Invalide l'intensité par type de surface : trop dispersée
- **Optimiser le Random Forest :**
  - Ajuster les hyperparamètres Randomized Search **CV** et Grid Search **CV**
  - Evaluer les temps de calcul : **4-15 min pour 100 à 300 fits**
  - Recentrer sur les features pertinents : **alléger les données de profil**

Principaux résultats :



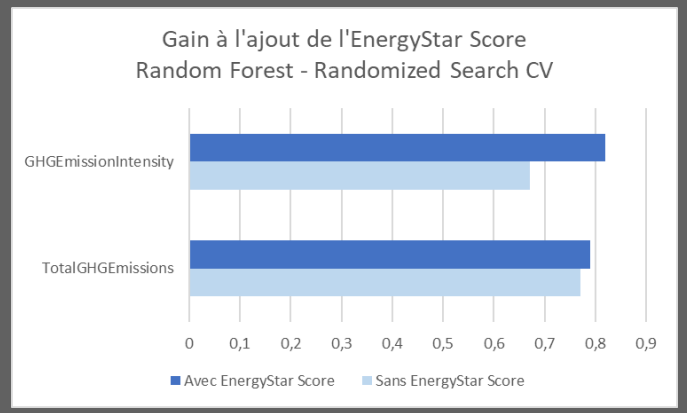


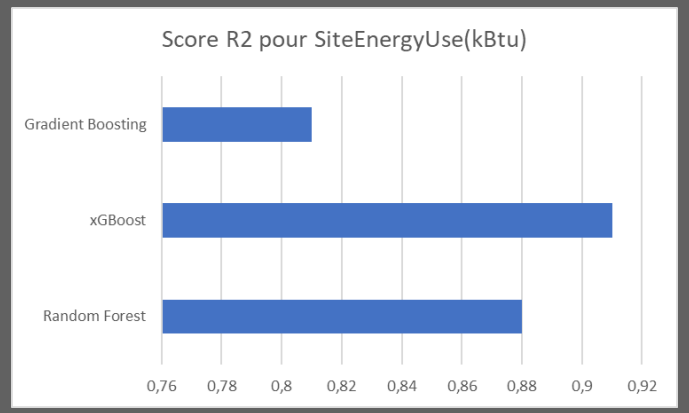
3 – Modélisation (3/3)

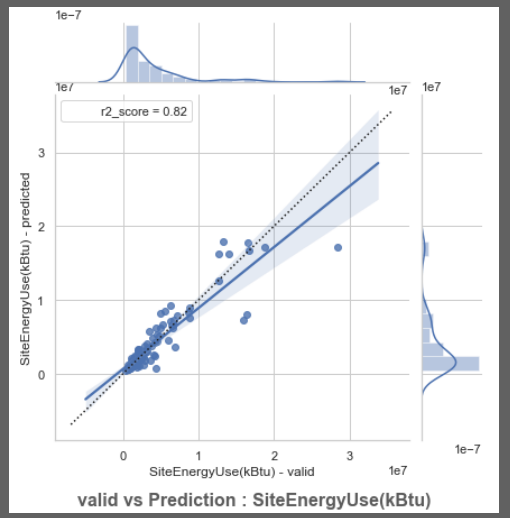
3.2. Approche & Principaux résultats

- **Tester l'apport** du gain de prise en compte **EnergyStar Score**
  - Restriction aux cas de score connus
  - **Gain initial significatif – affaibli après feature engineering**
- **Explorer GradientBoosting et xGBoost** ***(/!\ non optimisés)***
  - Comparer gains versus tps de calcul et features considérés
- **Confronter** le modèle à des données isolées
  - **Isoler** un jeu de donnée pour validation
  - **Apprendre** hors données (R2 amélioré à 0,92)
  - **Prédire** sur ces données (R2 dégradé à 0,82)
- Meilleur modèle : **RandomForestRegressor**

Principaux résultats :








```
RandomForestRegressor
{'bootstrap': False,
 'criterion': 'mse',
 'max_depth': None,
 'max_features': 'sqrt',
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 300,
 'n_jobs': None,
 'oob_score': False,
 'random_state': None,
 'verbose': 0,
 'warm_start': False}
```

Conclusions

Random Forest (with randomized & CV) offre le meilleur compromis

Modèle	Cible	Pouvoir prédictif		Feat. Engineering	Complexité		Rang
		R2	*RMSE*	*Features*	*Tuning*	*Calcul*	
Random Forest ***(best : Rand with CV)***	*Energy*	**0,9**	**1,7 E6**	Light	Medium	Medium	**1.**
	Emissions	**0,81**	**40**	Medium			
xGBoost *non optimisé*	*Energy*	+ 0,1%	- 7 %	Medium	Medium	*« Heavy »*	**2.**
Random Forest *Avant optimisation*	*Energy*	- 3 %	+ 13 %	Light	Light	Light	**3.**
	Emissions	-9,9 %	+ 35 %	Light			
Gradient Boosting *non optimisé*	*Energy*	- 10 %	+ 39 %	Medium	Medium	*« Heavy »*	**4.**
Ridge Reg. with CV	*Energy*	- 24 %	+ 80 %	Light	Light	Light	**5.**
Dummy model (median)	*Energy*	- 116 %	+ 297 %	-			**6.**
	Emissions	- 118 %	+ 177 %	-			

Conclusions

- **Perspectives :**
  - Les approchent **xGBoost** et **GradientBoosting**, prisent comme oustiders, mériteraient optimisation :
    - Le temps de calcul **n'est pas pénalisant** en contexte métier
    - **Les optimisations ont toutes les chances d'améliorer les résultats.**
  - Les modèles sur-estiment pour les faibles valeur et sous-estiment pour les plus grandes valeurs :
    - Une idée serait de combiner à un clustering judicieux des données de profil pour ajuster le modèle.
  - L'approche intensité par usage n'a pas abouti :
    - Les données d'autres sources seraient à agréger dans cette optique d'étalonnage.

- **Retour d'expérience projet :**
  - Maitrise de l'EDA orientée « finalité ML »
  - Sources documentaires pléthoriques
  - Codage d'une fonction intégrée