Parcours Data Scientist : Projet 4

Anticipez les besoins en consommation électrique de bâtiments

Problématique

- Tenter de prédire les émissions de CO2 et la consommation totale d'énergie de bâtiments.
- Evaluer l'intérêt de l'ENERGY-STAR Score pour la prédiction d'émissions

- Combiner les données 2015 et 2016 en un seul tableau
- Faire une étude des données pour sélectionner les données les plus pertinentes et recalculer/simplifier/normaliser certaines données
- Chercher un modèle de régression pour la consommation d'énergie et les émissions en CO2
- Étudier les différences des résultats des modèles avec ou sans ESS pour les émissions

Cleaning et feature engineering

Création d’un seul fichier de données

- On fait en sorte que les deux tableaux aient les mêmes variables :
  - En 2015 on divise la colonne adresse avec les dictionnaires à l’intérieur pour obtenir les coordonnées GPS.
  - En 2015 on supprime les colonnes: 2010 Census Tract, City Counsil Districts, SPD Beats, Seattle Police Dpt Micro … et ZipCodes qui n’ont pas de correspondance en 2016
  - On crée une colonne « Comment » et une colonne « OtherFuelUse » en 2016
- On uniformise les noms des variables des deux tableaux et on concatène.

Nettoyage rapide des données

- Suppressions de colonnes inutiles :
  - Données géographiques (on a la position GPS) : City, State, Address, ZipCode, District, Neigbourhood
  - Données administratives : TawParcelIdentificationNumber, YearofDeliveryESS
  - Via une HotMap on supprime les colonnes Energy(kWh) et NaturalGas(Therms) corrélés à 100 % avec leurs homologues en kBtu.
  - On supprime également SourceEUI et SourceEUI/WN qui ne nous intéressent pas pour la problématique (Simple prise en compte en plus du transport énergétique)
  - On observe d'autres corrélations (WN) mais pour le moment on va conserver le tout pour le traitement des données manquantes, des doublons et des outliers.
- On utilise la colonne Comment pour supprimer les bâtiments en construction/réhabilitation sur 2015 et 2016
- On utilise la colonne PrimaryPropertyType pour remplir les valeurs manquantes de LargestPropertyUseType

Traitement des données catégorielles

- Après analyse ANOVA des catégories de Buildings
  - On simplifie en 2 catégories :
    - Residential et NonResidential
  - Pour Largest, SecondLargest et ThirdLargest on crée 5 catégories :
    - High, Middle, Parking, Other et None

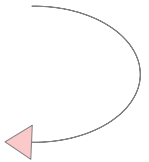
BuildingType	LargestPropType	SecondLargestPropType	ThirdLargestPropType
NonResidential	Medical Office	Laboratory	Restaurant
Multifamily LR(1-4)	Low-Rise Multifamily	Other	NaN
Campus	College/University	Data Center	Parking
Multifamily HR(10+)	Residence Hall/Dormitory	Restaurant	NaN

	BuildingType	LargestPropType	SecondLargestPropType	ThirdLargestPropType
0	NonResidential	Middle	Middle	Middle
1	Residential	Other	Other	None
2	NonResidential	Other	High	Parking
3	Residential	Other	Middle	None

Traitement des surfaces

- Parking n’a rien à voir dans les usages on compile le tout avec PropertyGFAParking :
  - Au cas où cette valeur était à 0 on la remplace par la valeur en usage et on réévalue la surface totale
  - On fait monter les usages en dessous de parking.

TotalGFA	ParkingGFA	LargestType	LargestTypeGFA	SecondUseType	SecondUseTypeGFA	ThirdUseType	ThirdUseTypeGFA
2000	100	Other	100	High	20	Parking	10
3000	0	Middle	2000	Parking	100	High	20
500	200	Parking	300	Other	140	None	0



TotalGFA	ParkingGFA	LargestType	LargestTypeGFA	SecondUseType	SecondUseTypeGFA	ThirdUseType	ThirdUseTypeGFA
2000	100	Other	100	High	20	None	0
3100	100	Middle	2000	High	20	None	0
500	200	Other	140	None	0	None	0

Traitement des doublons

- La majorité des batiments sont disponibles en 2015 et 2016 :
  - On crée 3 tableaux :
    - Les non doublons (152)
    - Les doublons 2015 (3269)
    - Les doublons 2016 (3269)
  - On joint les 2 tableaux de doublons sur la colonne ID et on va traiter les données manquantes et aberrantes sur ce tableau avant de refaire un tableau global.

Traitement des données

- Suppression des Buildings avec de trop grosses incohérences :
  - Coordonnées GPS trop éloignées d'une année à l'autre, différences dans le nombre de building et/ou étages
  - Suppression des outliers pour les données chiffrées (outliers en 2015 ou 2016)
- Traitement des valeurs manquantes
  - On prend la valeurs de l'autre année quand elle est disponible
- Imputation de la valeur finale
  - On calcule la moyenne des deux valeurs

PropertyGFAParking_15	PropertyGFAParking_16
50.0	NaN
NaN	100.0
NaN	NaN
60.0	80.0

PropertyGFAParking_15	PropertyGFAParking_16
50.0	50.0
100.0	100.0
NaN	NaN
60.0	80.0

PropertyGFAParking
50.0
100.0
NaN
70.0

Nettoyage final de la base de donnée

- On concatène le tableau des données dupliquées avec celui des non dupliquées
- Traitement des données manquantes :
  - On extrapole les données de surfaces avec les données disponibles
  - Suppression d'une ligne sans données de consommation
- Suppression de colonnes inutiles :
  - Données dont on a plus besoin pour les données manquantes: 'PropertyName','ListOfAllPropertyUseTypes','DefaultData' 'DataYear'
  - Données trop fortement corrélées avec la cible :

    'SiteEUIWN(kBtu/sf)' , 'SiteEnergyUseWN(kBtu)', 'GHGEmissionsIntensity','SiteEUI(kBtu/sf)'
- Suppression des outliers par analyse d'histogrammes et suppression des données labelisées low and high outliers (seulement 2%) et suppression de la colonne outliers

Traitement des données

- Données énergétiques :
  - Suppression des lignes dont la somme des sources s'éloigne de plus de 10 % de la quantité d'énergie totale
  - Remplacement des valeurs par la proportion de chaque énergie dans la somme de ces énergies

Electricity(kBtu)	NaturalGas(kBtu)	SteamUse(kBtu)	OtherFuelUse(kBtu)
30	20	10	0
20	10	0	5
25	15	5	0

Electricity(kBtu)	NaturalGas(kBtu)	SteamUse(kBtu)	OtherFuelUse(kBtu)
0.50	0.33	0.17	0.00
0.57	0.29	0.00	0.14
0.56	0.33	0.11	0.00

Traitement des données

- Surfaces et types de bâtiment :
  - On a plus que 3 types de batiment : High, Middle et Other
  - On remplace LargestGFA, SecondLargestGFA et ThirdLargestGFA par les surfaces High, Middle et Other de ces batiments
  - On supprime ainsi les données catégorielles

LargestType	LargestTypeGFA	SecondUseType	SecondUseTypeGFA	ThirdUseType	ThirdUseTypeGFA
Other	100	High	20	Other	10
Middle	200	Middle	100	High	20
Other	100	Other	40	None	0

LargestType	SecondLargestType	ThirdLargestType
20	0	110
20	300	0
0	0	140

Suppression des dimensions liées

- PropertyGFATotal=PropertyGFAParking+PropertyGFABuilding(s)
  - On supprime PropertyGFATotal

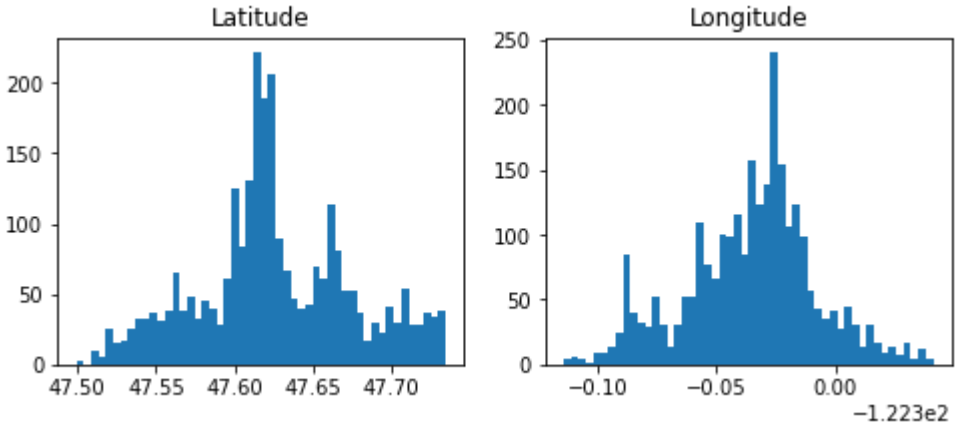
- On a Electricity+Steam+Gas+OtherFuel=1
  - On supprime Steam

Numérisation et normalisation des données

- On modifie notre variables catégorielle en {1;0}

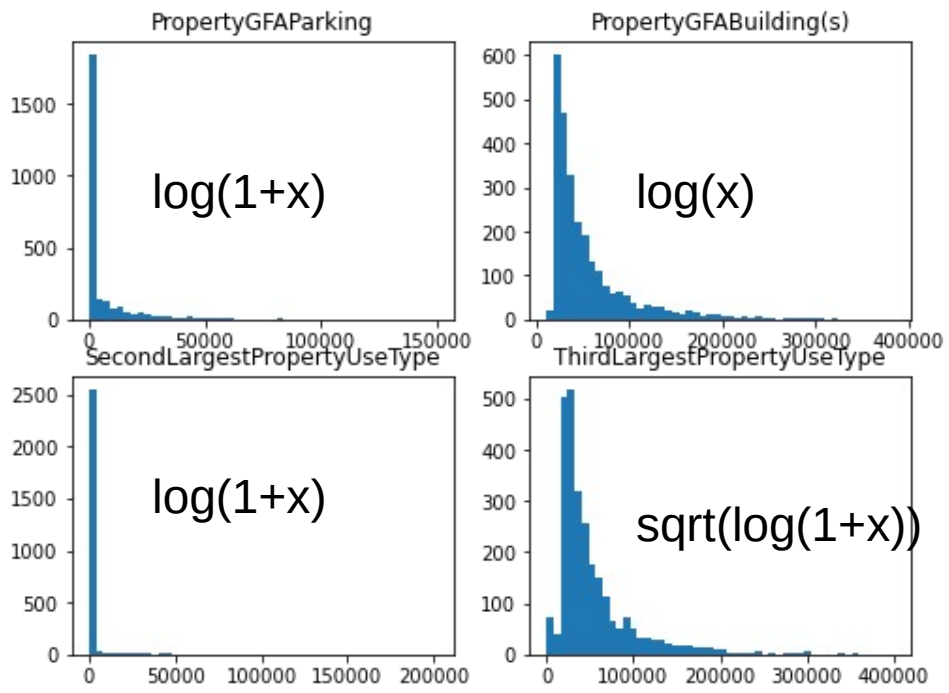
  ‘NonResidential’=1 et ‘Residential’=0

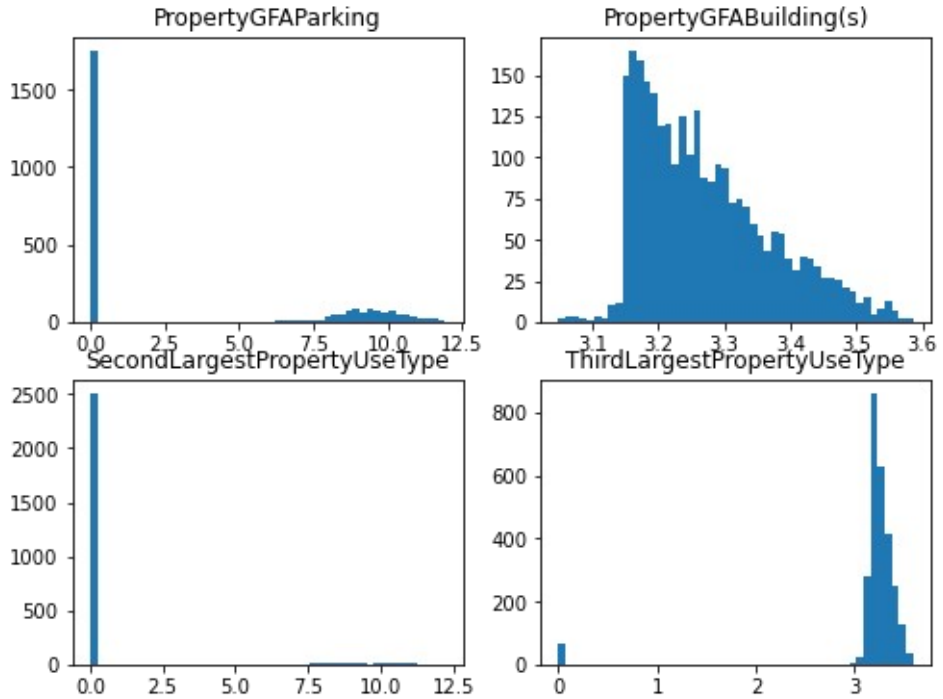
- Pour les données chiffrées on regarde les distributions



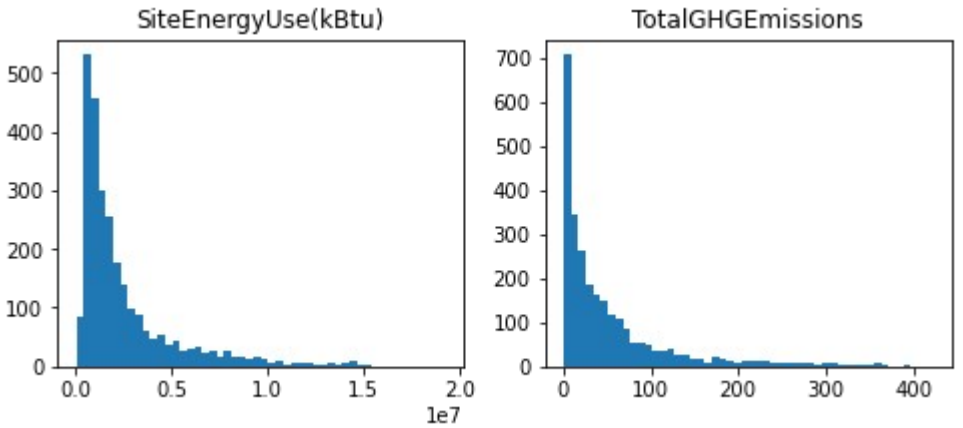
Ici on ne fait rien, la distribution est presque normale

Normalisation des données

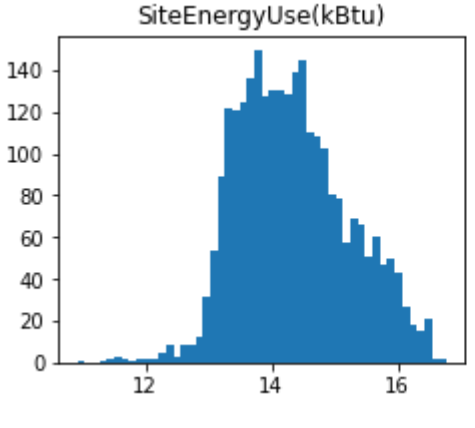


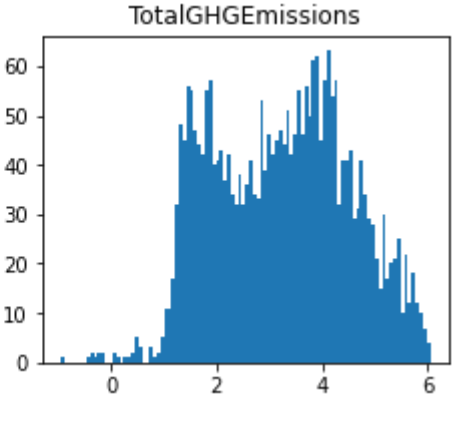


Normalisation des données

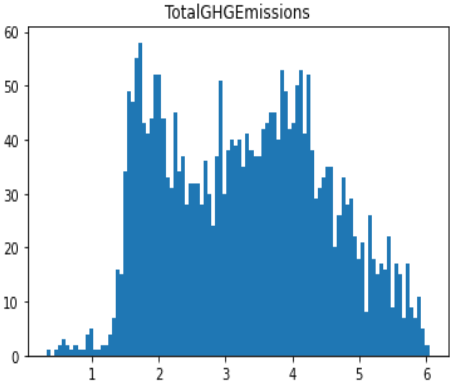


log





log(1+x) pour GHG

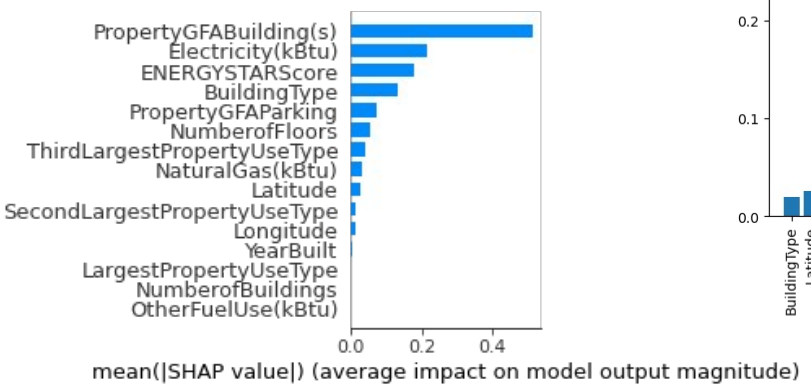


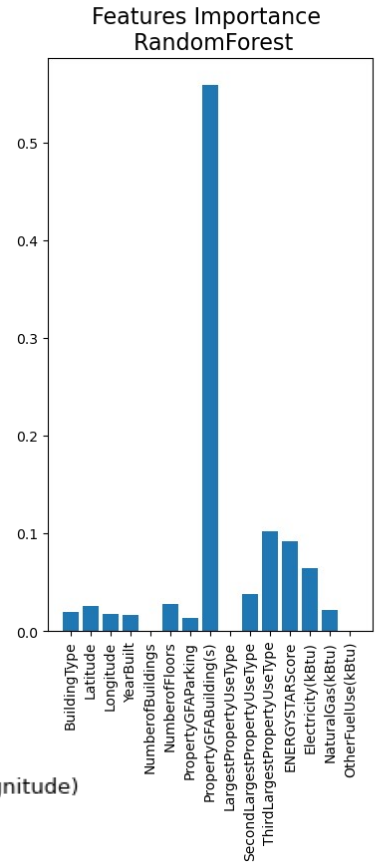
Recherche des modèles de prédiction

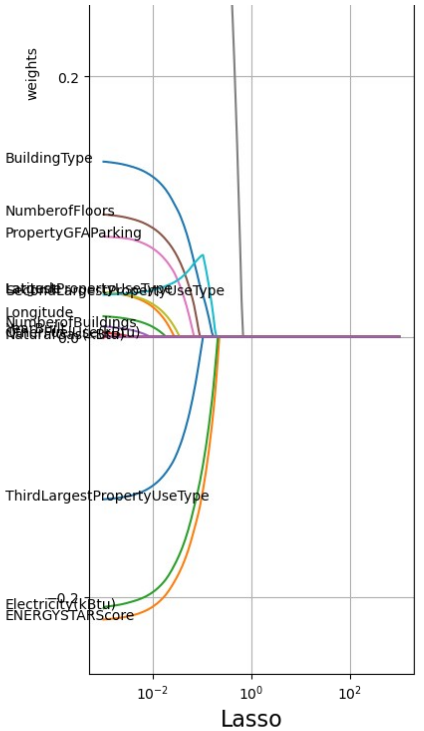
Consommation énergétique

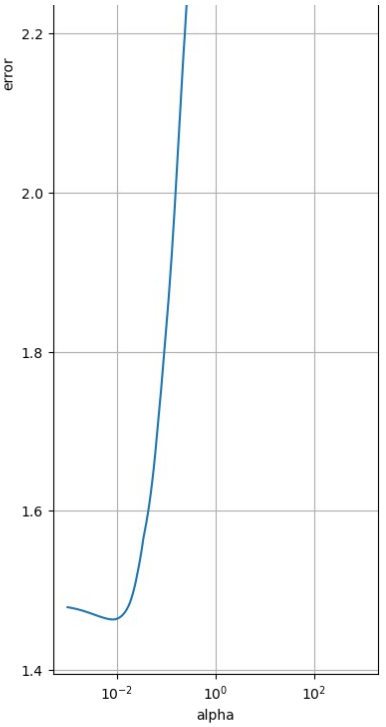
Feature Selection

On regarde l'importance des variables pour une régression linéaire simple, un Lasso et une random forest.









Features selection

- On regarde les résultats en supprimant quelques variables :

	R2	RMSE	MAE	MSLE
Dummy	-1.486846e+32	2.839051e+06	1.693402e+06	0.776284
Régression Simple	5.985233e-01	1.479445e+06	7.906859e+05	0.192749
Régression Ridge	5.810977e-01	1.485831e+06	7.912428e+05	0.193195
Régression Lasso	-5.289863e-01	1.827681e+06	9.362064e+05	0.253743
Random Forrest	6.241748e-01	1.276060e+06	6.779992e+05	0.166045

Avec tout

R2	RMSE	MAE	MSLE
-1.486846e+32	2.839051e+06	1.693402e+06	0.776284
5.985233e-01	1.479445e+06	7.906859e+05	0.192749
5.925525e-01	1.481055e+06	7.898267e+05	0.192680
-5.289863e-01	1.827681e+06	9.362064e+05	0.253743
6.263650e-01	1.272275e+06	6.790721e+05	0.165753

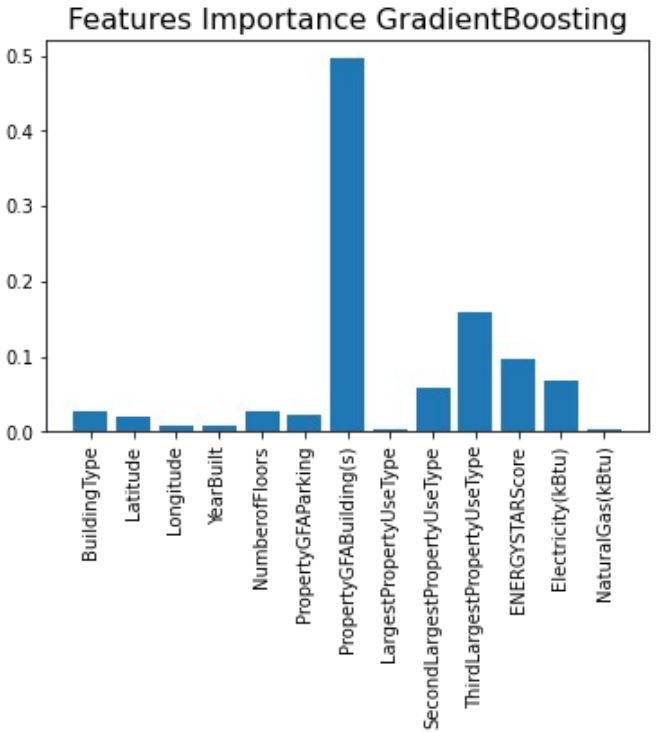
Sans OtherFuel et sans NumberofBuildings

R2	RMSE	MAE	MSLE
-1.486846e+32	2.839051e+06	1.693402e+06	0.776284
5.985233e-01	1.479445e+06	7.906859e+05	0.192749
5.342142e-01	1.498991e+06	7.938985e+05	0.198890
-5.289863e-01	1.827681e+06	9.362064e+05	0.253743
6.167457e-01	1.284532e+06	6.794196e+05	0.165254

Sans OtherFuel, NumberofBuildings et LargestPropertyUse

Features Selection

- On fait la même chose avec un GradientBoostRegression après une GridSearch sur le R2



On obtient un meilleur score en supprimant en plus:
Longitude, YearBuilt, NaturalGas et LargestPropertyUseType

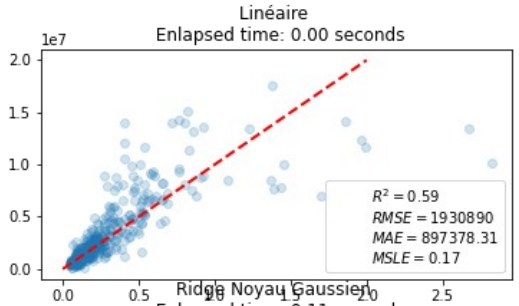
Recheche de la meilleure regression

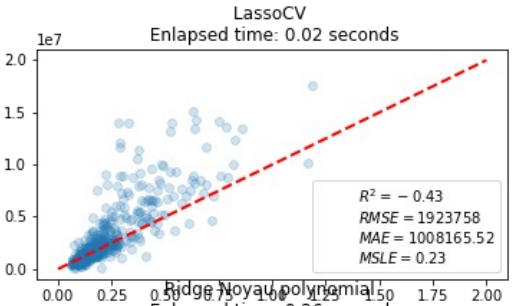
- On ne conserve plus que les autres variables et on lance plusieurs gridsearch pour chercher les meilleurs modèles. On obtient :

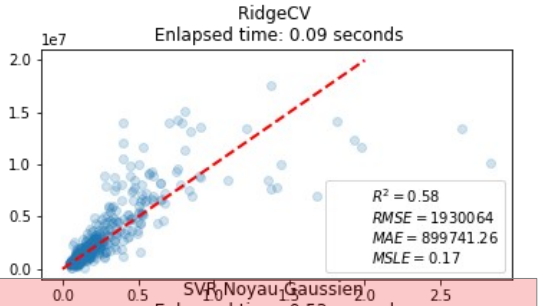
	Best Parameters	R2
KernelRidge rbf	{'alpha': 0.00013894954943731373, 'gamma': 0.0...	0.822620
KernelRidge sigmoid	{'alpha': 7.196856730011514e-06, 'gamma': 2.51...	0.786710
KernelRidge Polynomiale	{'alpha': 0.025118864315095794, 'degree': 3, '...	0.826718
SVR linear	{'C': 215.44346900318823, 'gamma': 1e-06}	0.779699
SVR rbf	{'C': 1000.0, 'gamma': 0.0021544346900318843}	0.826021
SVR sigmoid	{'C': 10.0, 'gamma': 0.0021544346900318843}	0.779433

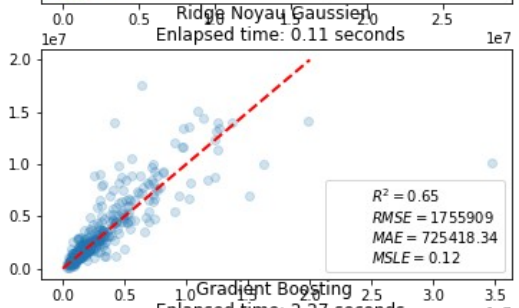
Comparaison des modèles

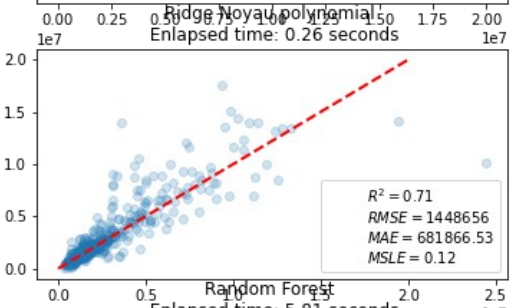
RealValue=f(PredictedValue)

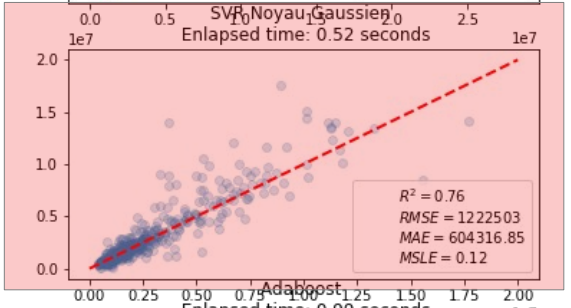


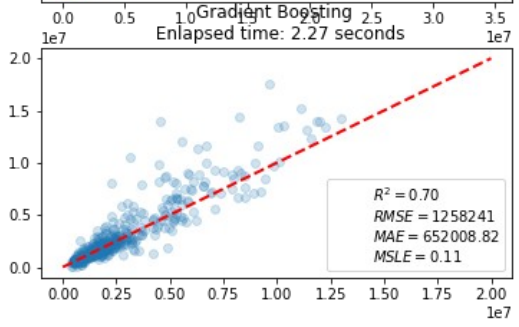


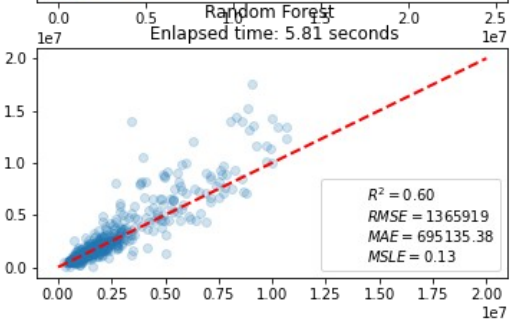


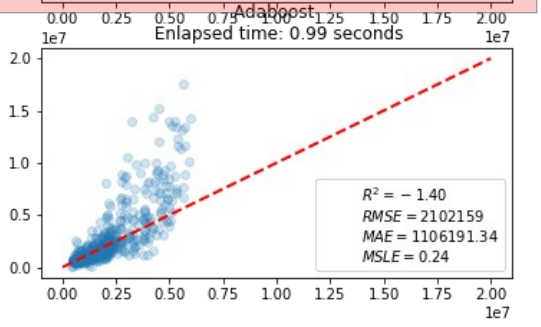






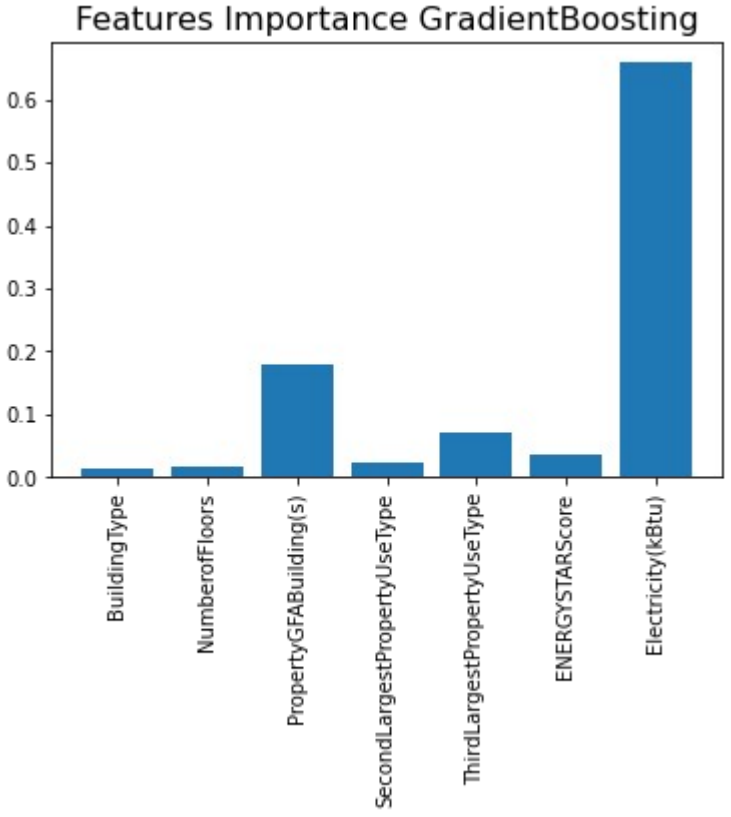






Emissions de CO2 avec ESS

- Avec la même méthode, en gardant ces variables :

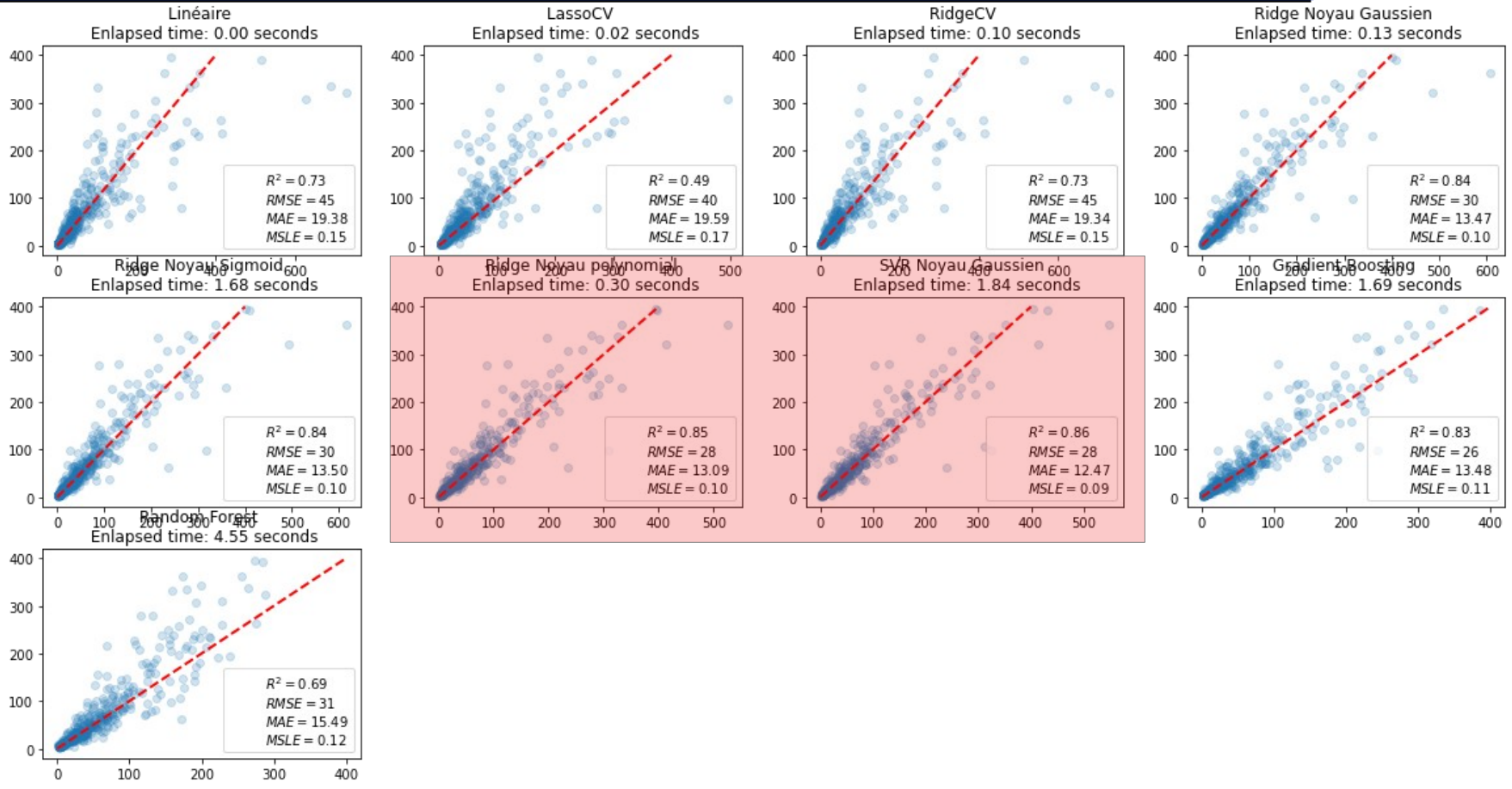


On obtient un score au R2 de 0.91 sur une GridSearch.

On baisse le score en enlevant de nouvelles variables.
On va donc faire l'étude avec ces variables.

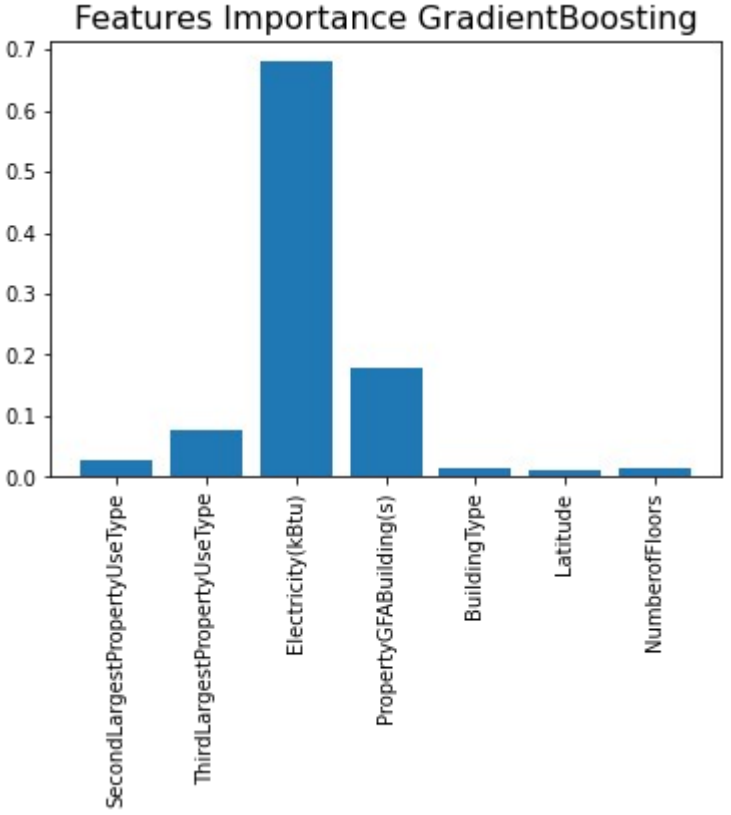
Résultats des modèles :

CO2 Avec ESS: RealValue=f(PredictedValue)



Emissions de CO2 Sans ESS

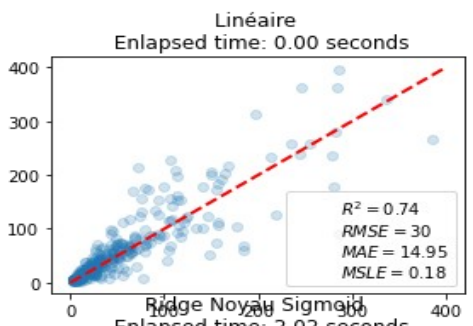
- Avec la même méthode, en gardant ces variables :

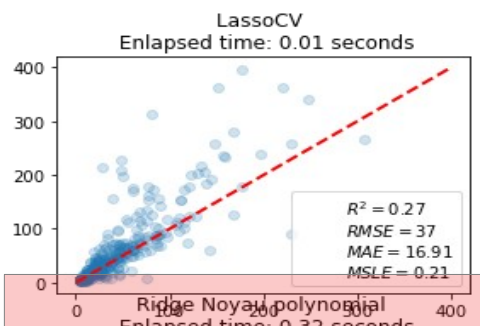


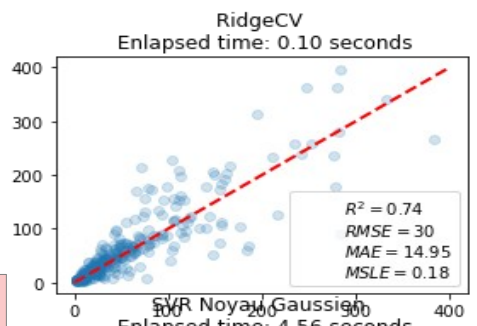
On obtient un score au R2 de 0.84 sur une GridSearch avec ces variables.

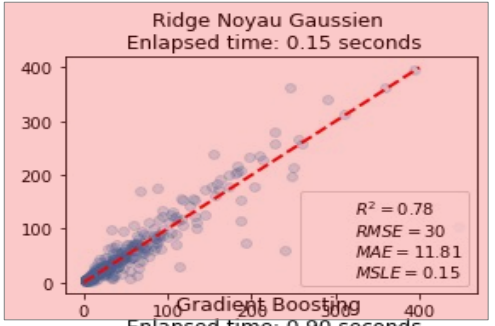
Résultats des modèles :

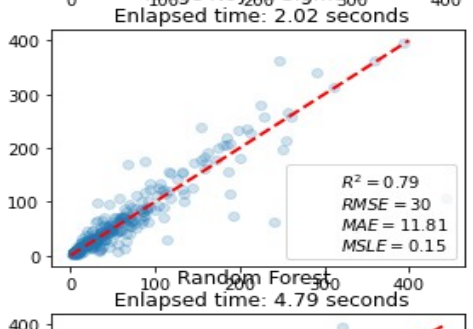
CO2 Sans ESS: RealValue=f(PredictedValue)

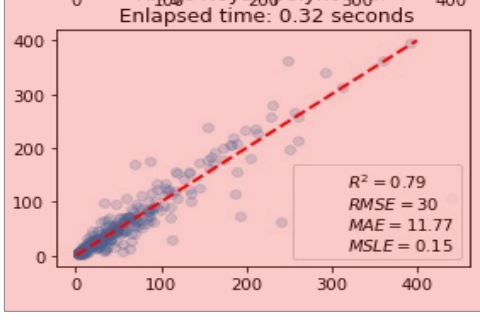


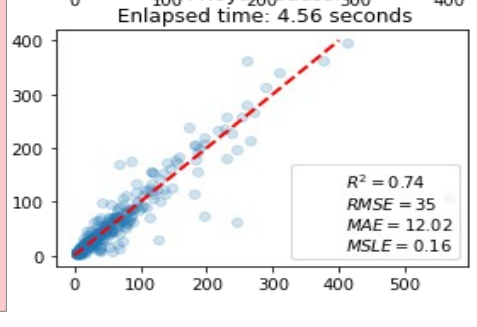


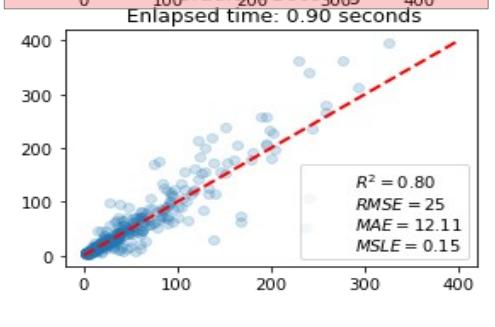


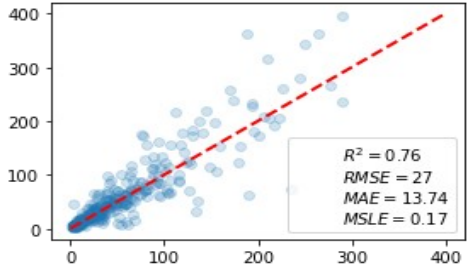












Conclusions

- Pour la prédiction de la consommation énergétique :
  - Les données importantes sont :
    - Le type de building (Résidentiel ou non) et le nombre d’étages
    - La latitude
    - La répartition des surfaces (building, parking et usages)
    - Les proportions de consommation énergétique par source
  - Le modèle à sélectionner
    - Une SVR noyau gaussien avec un gamma de 0.002 et un C de 1000

Conclusions

- Pour les prédictions d'émissions de CO2 :
  - Le modèle est souvent plus précis avec l'ESS mais la différence de précision ne justifie pas un effort laborieux de calcul de l'ESS. En effet on obtient les meilleurs résultats suivant (Avec ESS vs sans ESS):
    - R2 : 0.86 vs 0.8
    - RMSE : 26 vs 25
    - MAE : 12.47 vs 11.77
    - MSLE : 0.09 vs 0.15
  - Ainsi on conseille une régression ridge à noyau gaussien avec alpha=0.00005 et gamma=0.0006 et en conservant pour données :
    - Le type de building et son nombre d'étages
    - La part d'électricité dans la consommation énergétiques
    - La répartition des usages du bâtiment hors parking