Parcours Data Scientist

Projet N°4 :
Anticipez les besoins en consommation électrique de bâtiments

Daniel CHASTANET

Sommaire

- ❑ Rappel de la problématique
- ❑ Description de la base de données
- ❑ Nettoyage de la BDD
- ❑ Préparation de données et Résultat repère (« baseline »)
- ❑ Sélection du modèle et choix des attributs pertinents
- ❑ Améliorations et résultats
- ❑ Conclusion

Rappel de la problématique

Seattle : Objectif ville neutre en émissions de carbone en 2050

On s'intéresse ici aux émissions des bâtiments non destinés à l'habitation.

On veut prédire :

- La consomation d'énergie 'SiteEnergyUse(kBtu)'
- Les émission de CO^2 'GHGEmissions(MetricTonsCO2e)'

BDD : 2015 – 2016

- Relévés minucieux et couteux de nos 2 attributs
- → Tenter de prédire à partir des autres données

Merge des BDD

2015 / 2016 (OSE building):

« » A unique identifier assigned to each property covered by the Seattle Benchmarking Ordinance for tracking and identification purposes »

(3432, 96)

TaxParcelIdentificationNumber_x	TaxParcelIdentificationNumber_y	PropertyName_x	PropertyName_y	YearBuilt_x	YearBuilt_y	Latitude_x	Latitude_y
323049024.0	369000400.0	seattle warehouse office building	seattle 11-13	1961.0	1961.0	47.510603	47.51138
369000400.0	425049022.0	lawton elementary school (sps-district)	lawton elementary	1990.0	1990.0	47.657262	47.65671
425049022.0	467000429.0	eckstein middle (sps-district)	eckstein middle	1950.0	1950.0	47.682450	47.68252

Analyse pré exploratoires

Les données des propriétés

(51 colonnes) :

L'identification : OSEBuildingID, PropertyNamae, TaxParcelIdentificationNumber, Address, CouncilDistrictCode, Neighborhood

Le batiment en lui même : BuildingType, PrimaryPropertyType, YearBuilt, NumberofBuildings, NumberofFloors, PropertyGFATotal, PropertyGFAParking, PropertyGFABuilding(s), ListOfAllPropertyUseTypes, Largest, Second, Third, → equivalent en GFA

Consommation (kBtu) : SiteEnergyUse/WN, SiteEUI/sf/WN, SourceEUIsf/ WN, Steam, Electricity, NaturalGas, OtherFuelUse, GHGEmissions(MetricTonsCO2e)/ ft2

Extras : DefaultData, Com, ComplianceStatus, Outlier, 2010 Census Tracts, SPDMCPPA, City Council Districts, SPD Beats

Energy star : YearsENERGYSTARCertified, ENERGYSTARScore

L'Energy Star score

Outils d'évaluation crée par l'agence pour la protection environnementale (EPA)

Score de 1 à 100 : propriétés physiques du batîments
type d'utilization
comportement des occupants

Il ne permet pas en lui même d'expliquer pourquoi le batîment fonctionne d'une certaine façon ou de comment changer les performances. Il permet néanmoins d'évaluer les performances du batîment et d'identifier quel batîment offrent les meilleurs opportunities d'amelioration.

The Score Does	The Score Does Not
• Evaluate actual billed energy data • Normalize for business activity (hours, workers, climate) • Compare buildings to the national population • Indicate the level of energy performance	• Sum the energy use of each piece of equipment • Credit specific technologies • Compare buildings with others in Portfolio Manager • Explain why a building performs well or poorly

Nettoyag de la BDD

Valeurs aberrantes

(3432,

```
Multifamily LR (1-4)     1040
Multifamily MR (5-9)      584
```

(1698,

Retrait des batîment d'habitation :

		PrimaryPropertyType	PropertyName	NumberofBuildings	NumberofFloors	PropertyGFATotal	SiteEnergyUse(kBtu)
59	166	Hotel	Grand Hyatt Seattle	1.0	0	934292	6.504728e+07
72	487	Medical Office	Arnold Pavilion	1.0	0	225982	2.056062e+07
155	564	Other	Pacific Place	1.0	0	947987	4.651096e+07
191	1754	Medical Office	HART First Hill LLC	1.0	0	274568	2.531153e+07
229	1993	Other	(ID#24086)Campus1:KC Metro Transit Atlantic Central Bases	10.0	0	230971	2.102229e+07
252	3130	Warehouse	Sandpoint #5	1.0	0	384772	1.520676e+07
267	3131	Medical Office	Sandpoint #25	1.0	0	30287	2.193115e+06
269	3132	Small- and Mid-Sized Office	Sandpoint #29	1.0	0	21931	3.947209e+06
280	3168	Other	Magnuson	8.0	0	502030	1.847034e+07
347	3273	Other	Smilow Rainier Vista Boys & Girls Club	1.0	0	40265	2.159170e+06
	3274	University	University of Washington - Seattle Campus	111.0	0	9320156	8.739237e+08

Pre processing

Utilisation de pipelines

```
Pipeline(steps=[('columntransformer',
                 ColumnTransformer(transformers=[('pipeline-1',
                                                  Pipeline(steps=[('robustscaler',
                                                                   RobustScaler())]),
                                                  ['columns_1']),
                                                 ('pipeline-2',
                                                  Pipeline(steps=[('onehotencoder',
                                                                   OneHotEncoder(handle_unknown='ignore'))]),
                                                  ['columns_2']),
                ('model', Regressionmodel())])
```

RobustScaler() : median / inter-quartile
OneHotEncoder() : attributs sous forme de catégories 1-0

Machine learning baseline

LinearRegression()

```
mll = d16[['BuildingType', 'PrimaryPropertyType', 'LargestPropertyUseType', 'SecondLargestPropertyUseType',
           'ThirdLargestPropertyUseType',
           'YearBuilt',
           'CouncilDistrictCode', 'Neighborhood',
           'NumberofBuildings', 'NumberofFloors', 'PropertyGFATotal', 'PropertyGFAParking', 'PropertyGFABuilding(s)',
           'LargestPropertyUseTypeGFA', 'SecondLargestPropertyUseTypeGFA', 'ThirdLargestPropertyUseTypeGFA',
           'ENERGYSTARScore',
           'SiteEnergyUse(kBtu)']]
```

Data : Full model

Model : LinearRegression()

Results : sans Energystar

```
train : 0.8063377185863464
test : -0.07083898539640376
```

/ avec Energystar

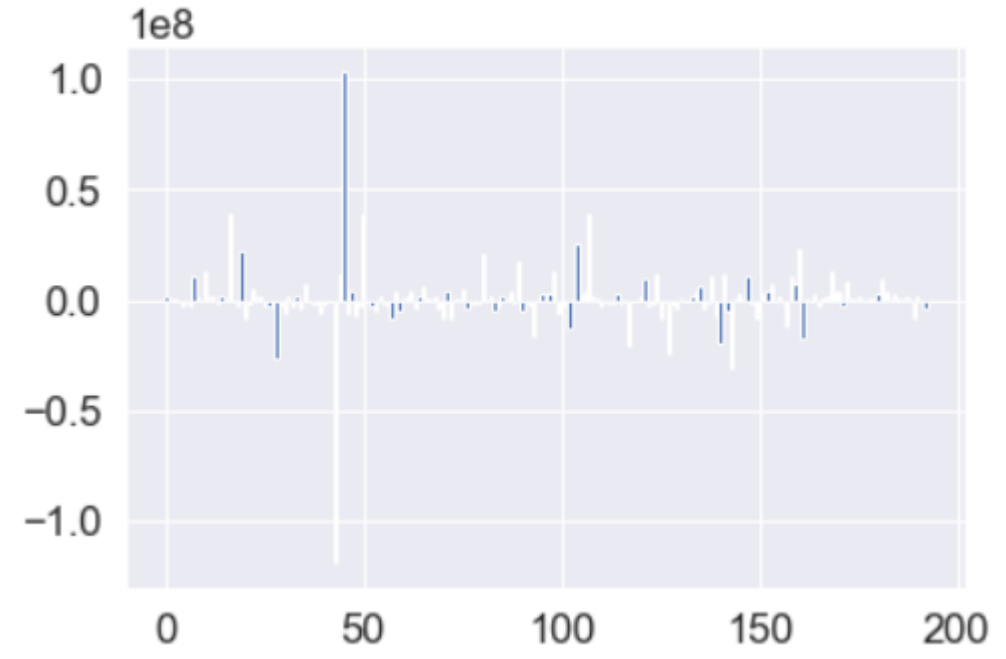
```
train : 0.8522277049761375
test : 0.5322021178728733
```

LinearRegression()

coefficients

192 features !

```
Feature: 191, Score: -839056.36893
Feature: 192, Score: -3456730.11412
```



Feature engineering

New columns and data selection

Création de nouveaux attributs :

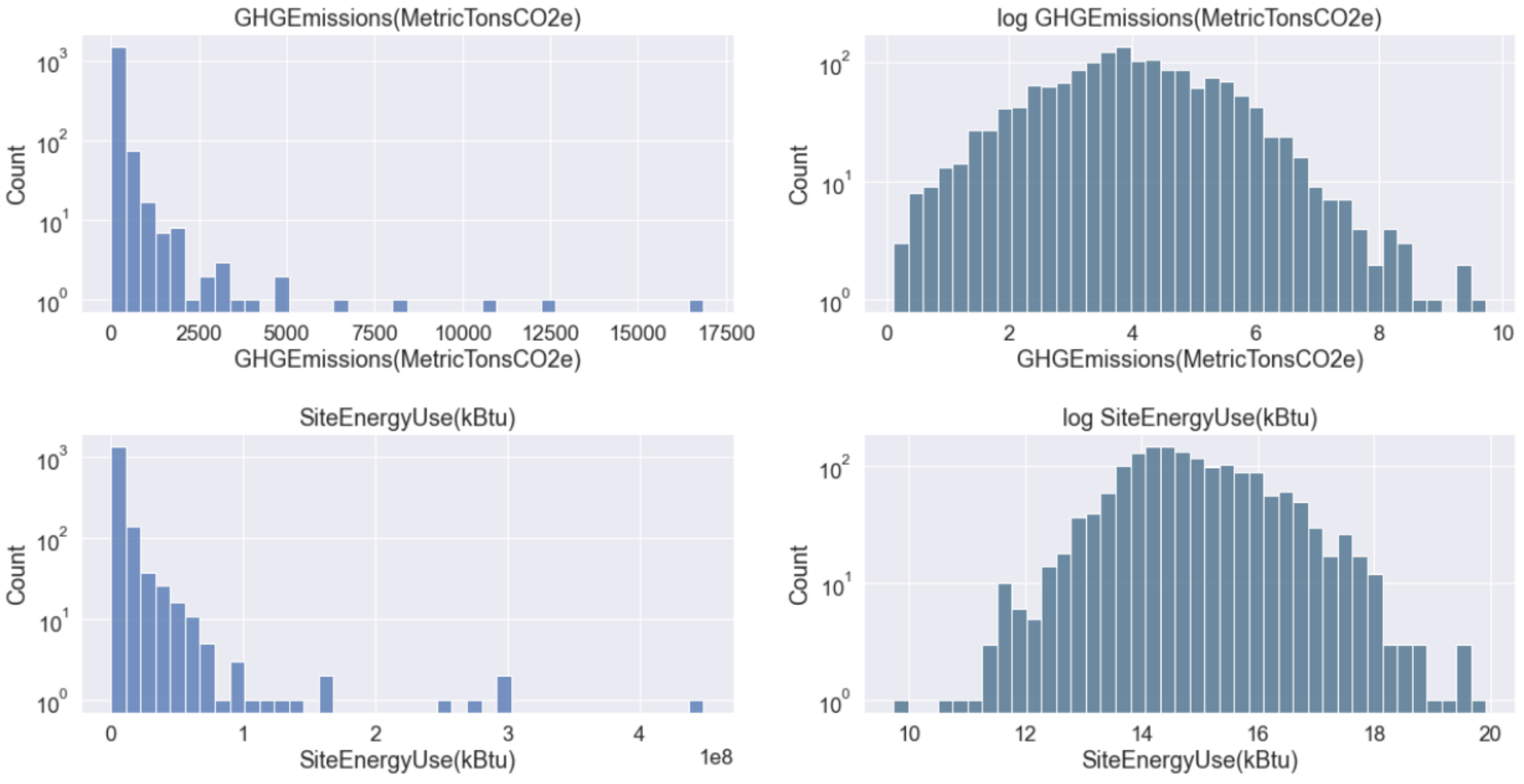
- Age : 2016 – « YearBuilt »
- Second_Type → 0 ou 1 en fonction de « SecondLargestPropertyUseType » (seconde surface d' d'exploitation principale)
- Third_Type → 0 ou 1 en fonction de « ThirdLargestPropertyUseType » (troisième surface d'exploitation principale)
- Prop_Type → Type of Use of the Building with 30+ ind (« LargestPropertyUseType améliorée »)
- % d'Energie :
    - Steam(%) = « SteamUse(kBtu) » / Somme * 100
    - Elec(%)
    - Gas(%)
    - Other(%)

Log(targets) : « SiteEnergyUse(kBtu) », « GHGEmissions(MetricTonsCO2e) »tu tr

Passage au log

Vers une distribution + « normale »

Right or positive skewed distribution



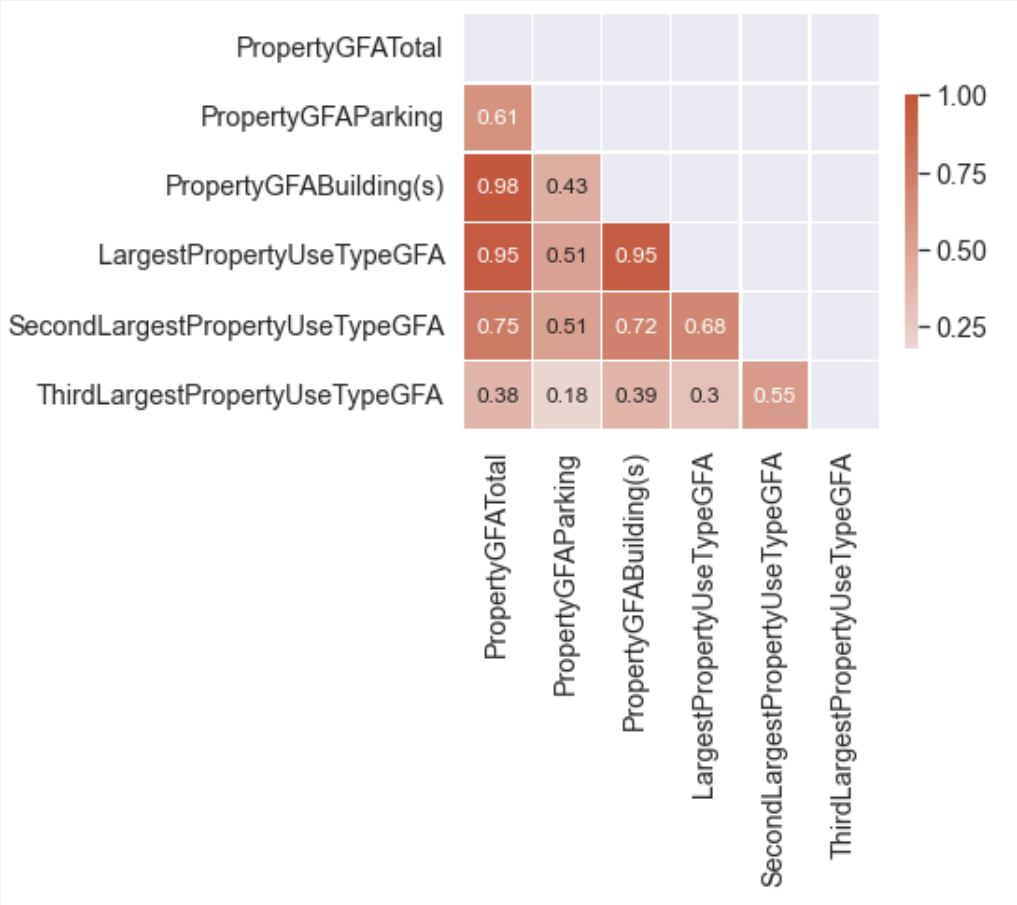
Feature engineering

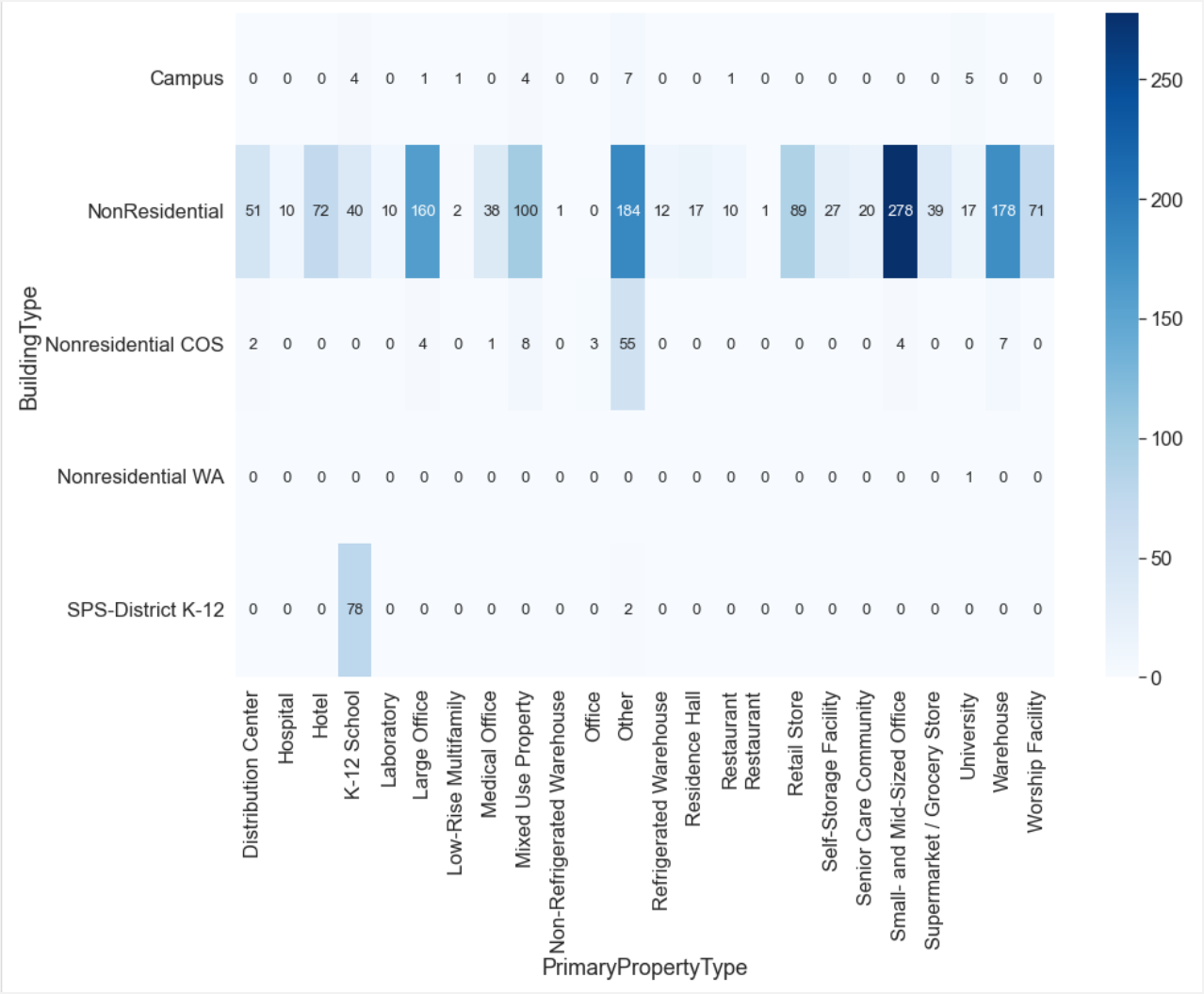
New columns and data selection

SiteEnergyUse(kBtu)

Feature engineering

data selection





New columns and data selection

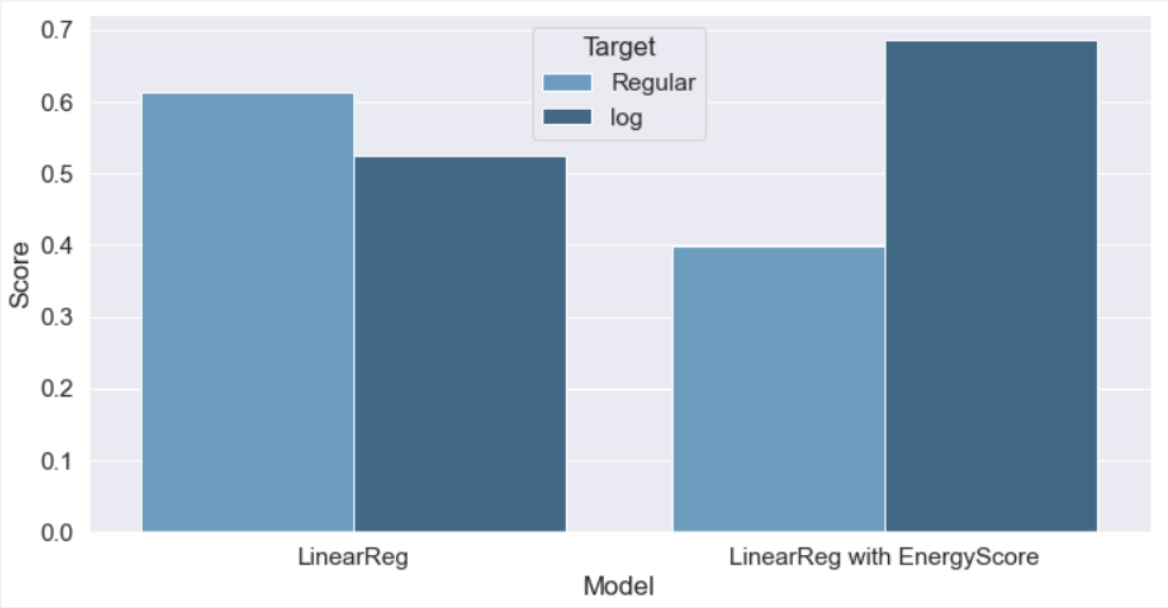
Sélection de nouvelles variables :

Vers un modèle simple : 5 attributs

- Age
- Prop_Type
- LargestPropertyUseTypeGFA
- SecondLargestPropertyUseTypeGFA
- Second_Type

Baseline : LinearRegression()

Score des « tests »



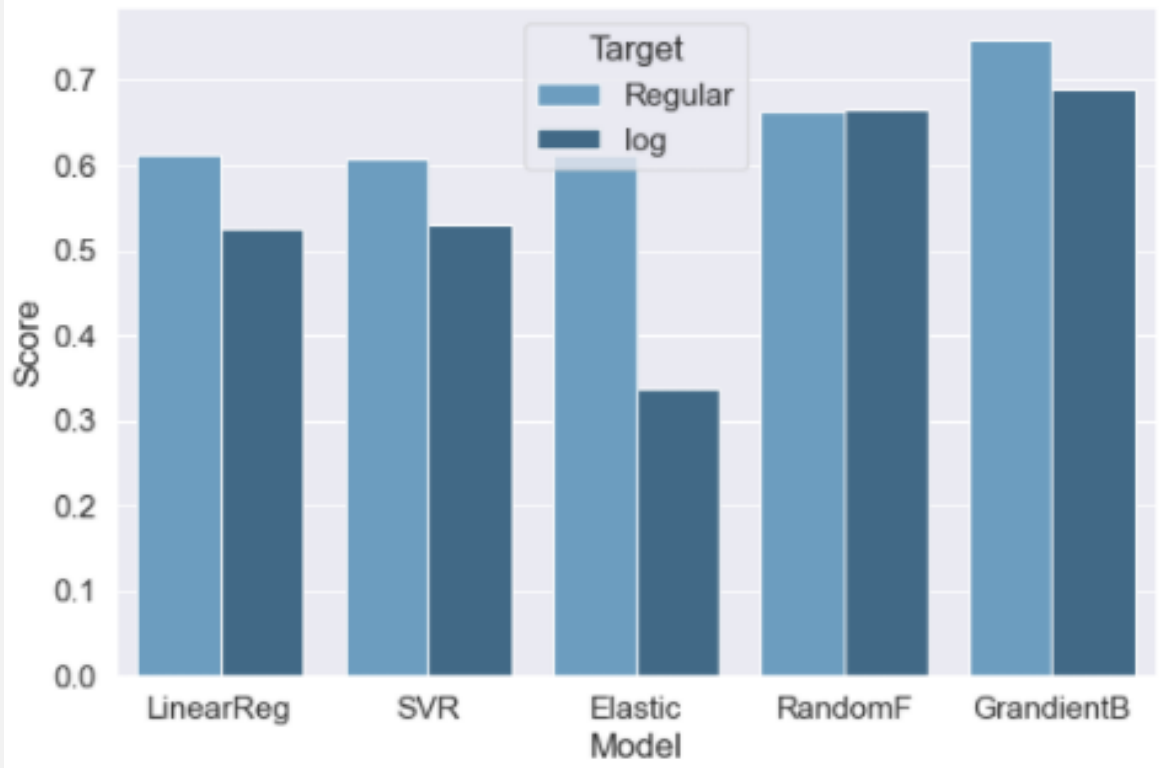
Baseline :

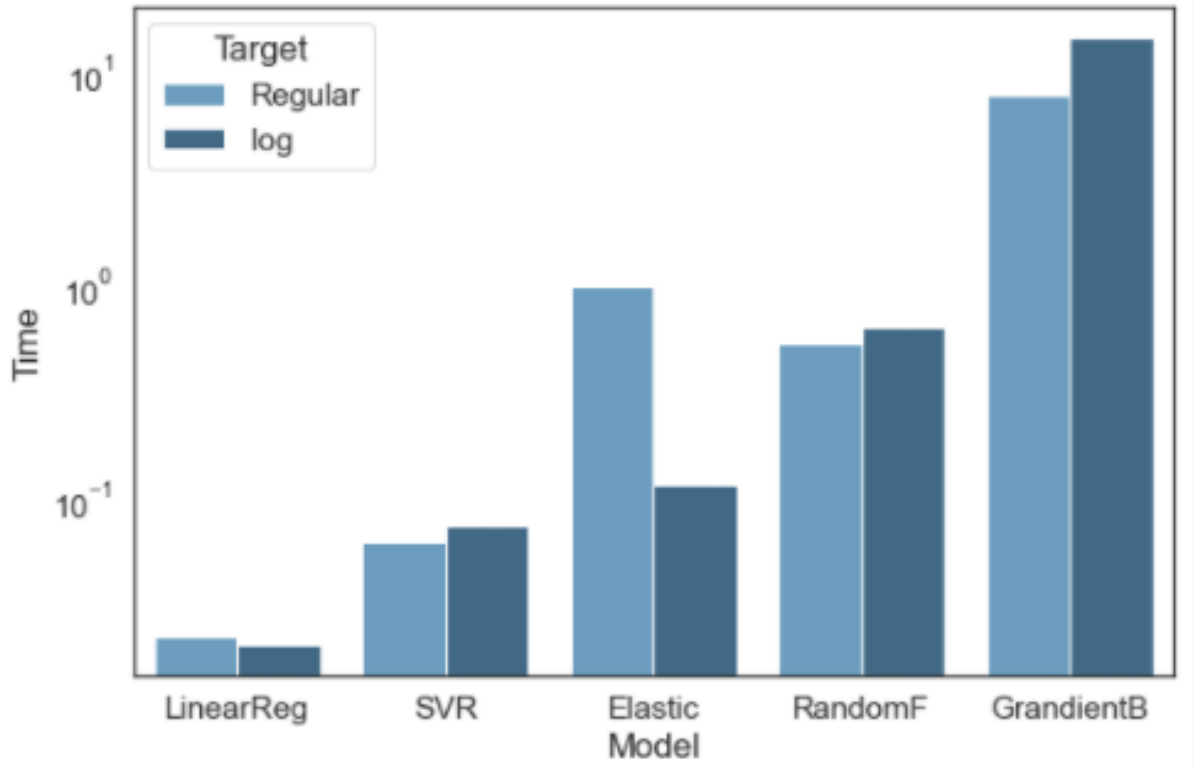
```
train : 0.8063377185863464
test : -0.07083898539640376
```

```
train : 0.8522277049761375
test : 0.5322021178728733
```

Modèle 1 : modèle simple

Score des « tests »





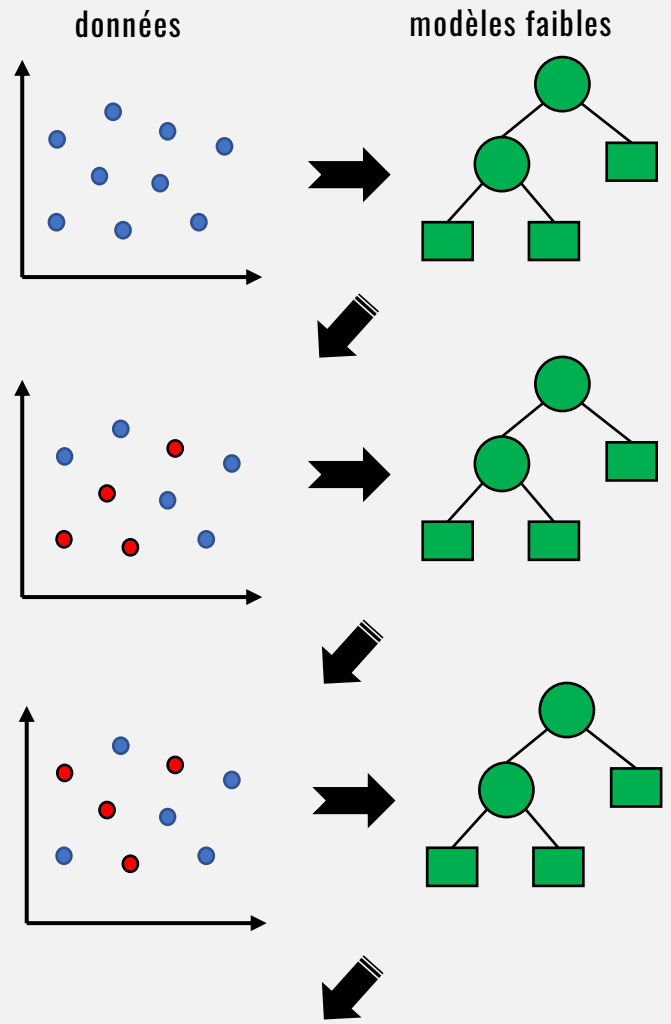
Gradient Boosting

Boosting :

Entrainement successif d'arbre
Système de poids

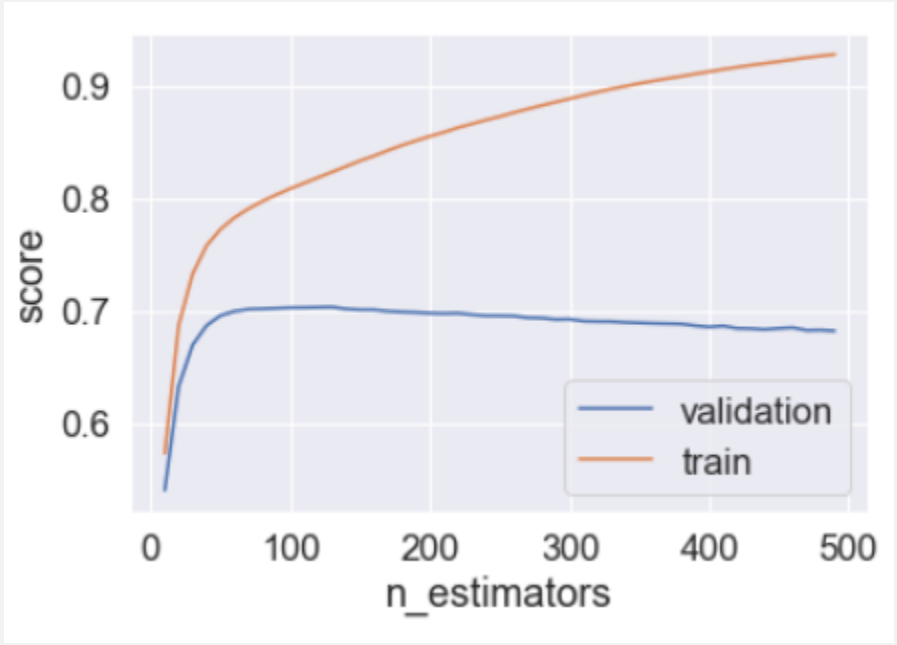
Paramètres :

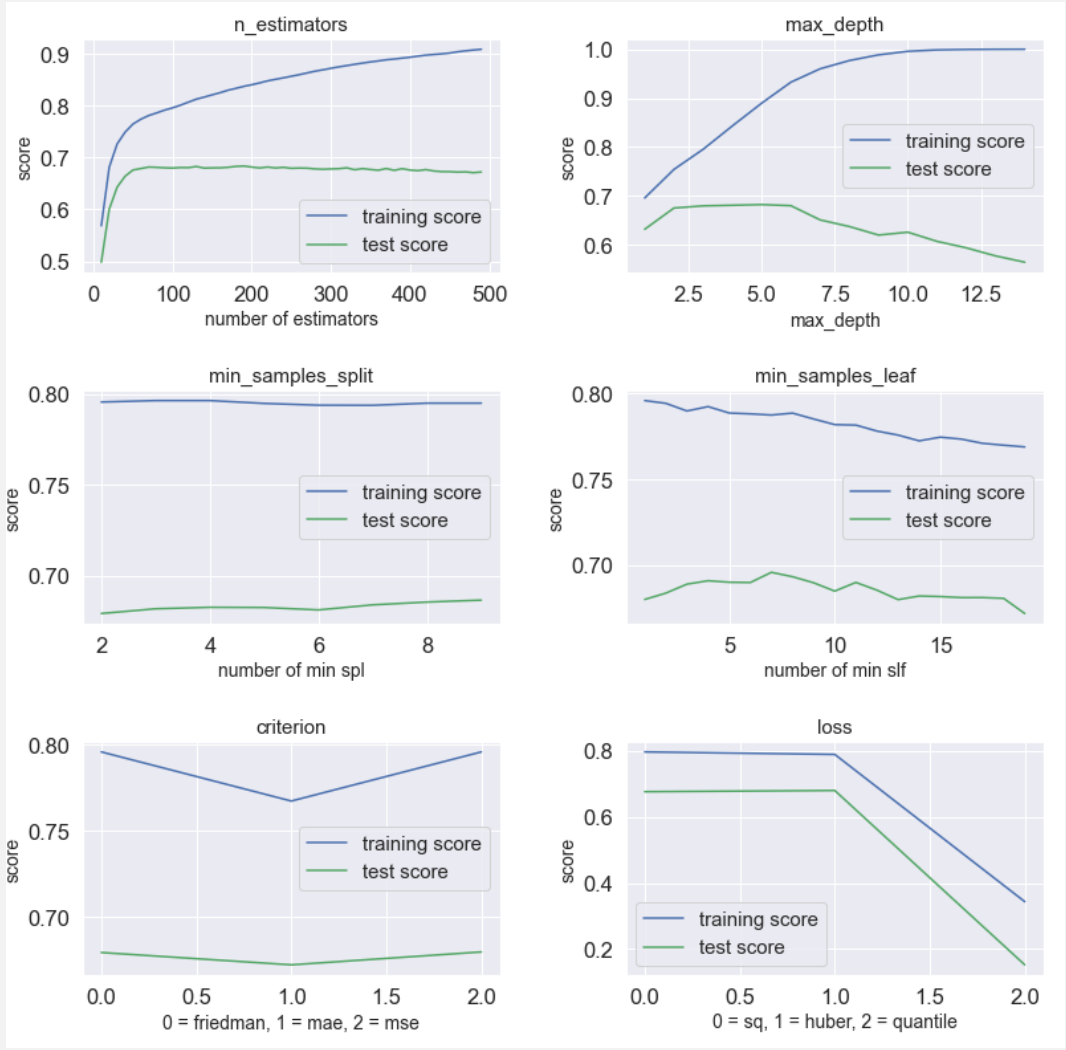
n_estimators : nombre d'arbres
max_depth : le profondeur de l'arbre (nombre max de noeuds)
min_samples_leaf : Nombre d'individus min dans chaque branche
min_samples_split : individys requis pour séparer un nœud
criterion : fonction de mesure de la qualité de la séparation
loss : fonction de perte qui doit être optimisée



Validation Curve et GridSearch

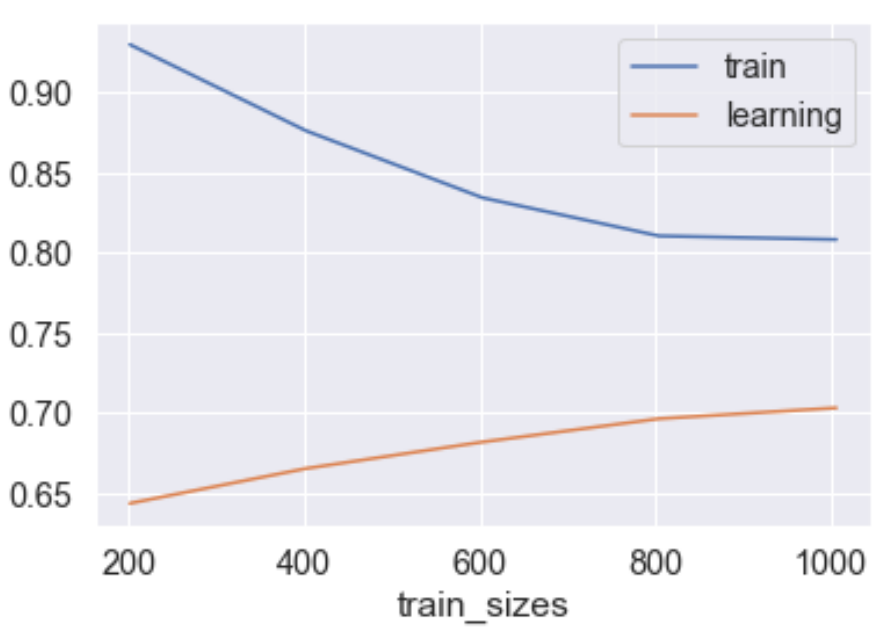
Modèle 1 : modèle simple

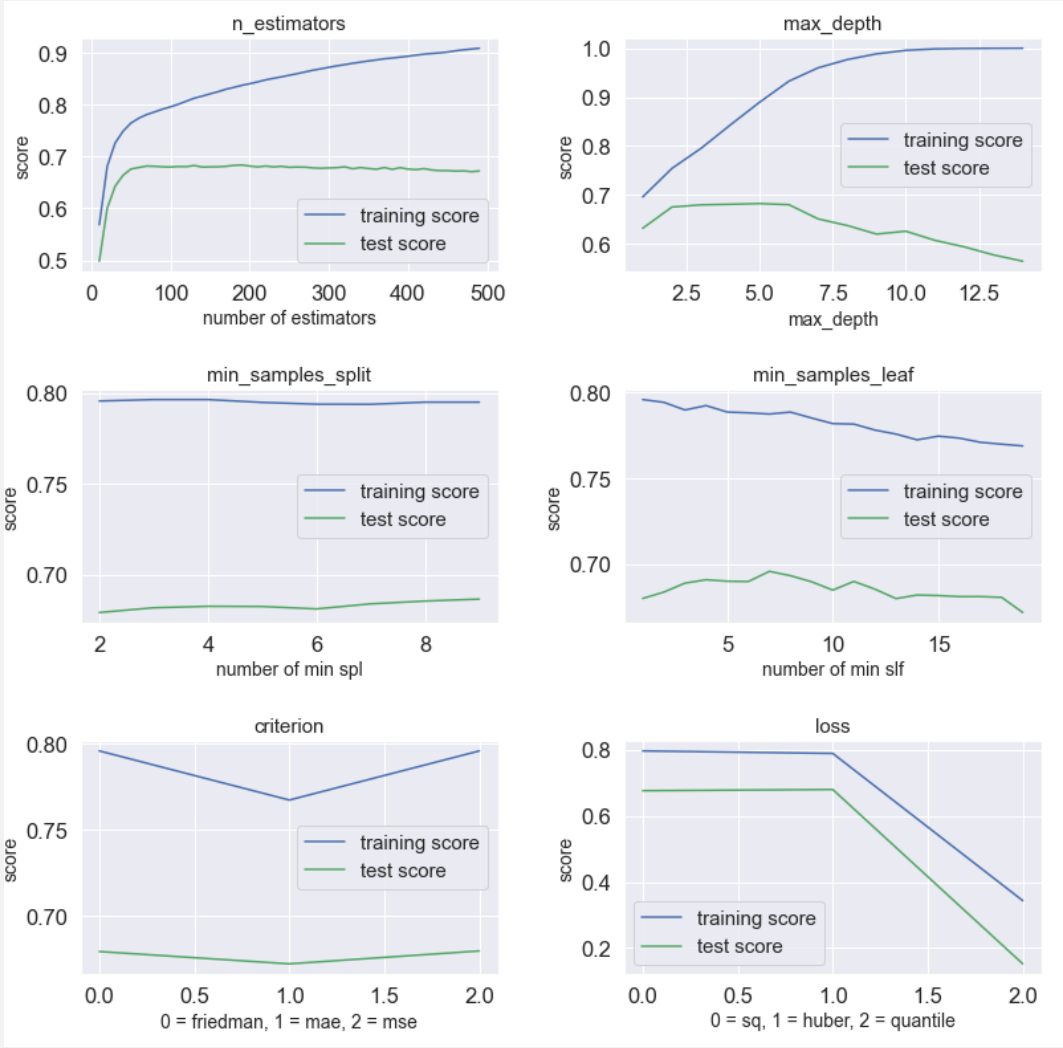




Learning Curve et GridSearch

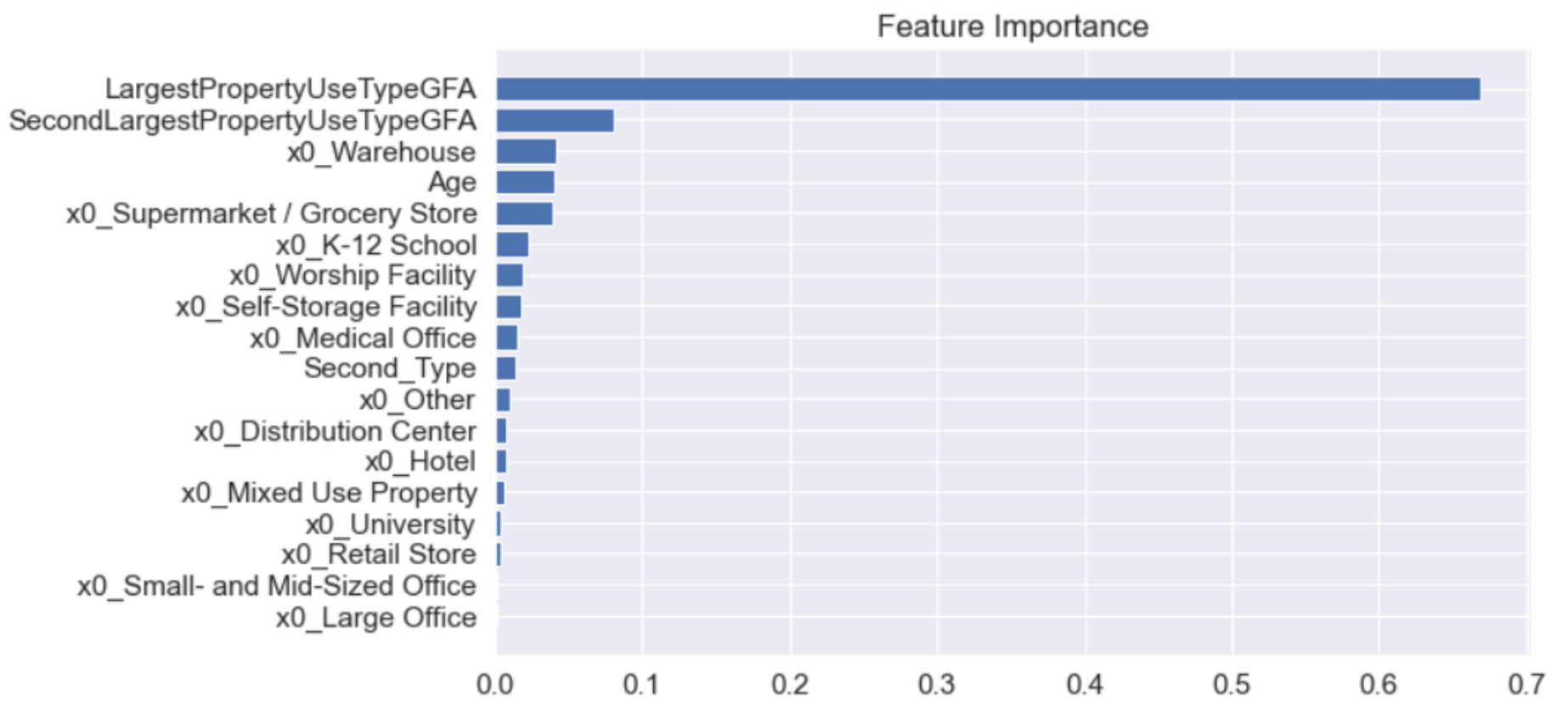
Modèle 1 : modèle simple





SiteEnergyUse Results

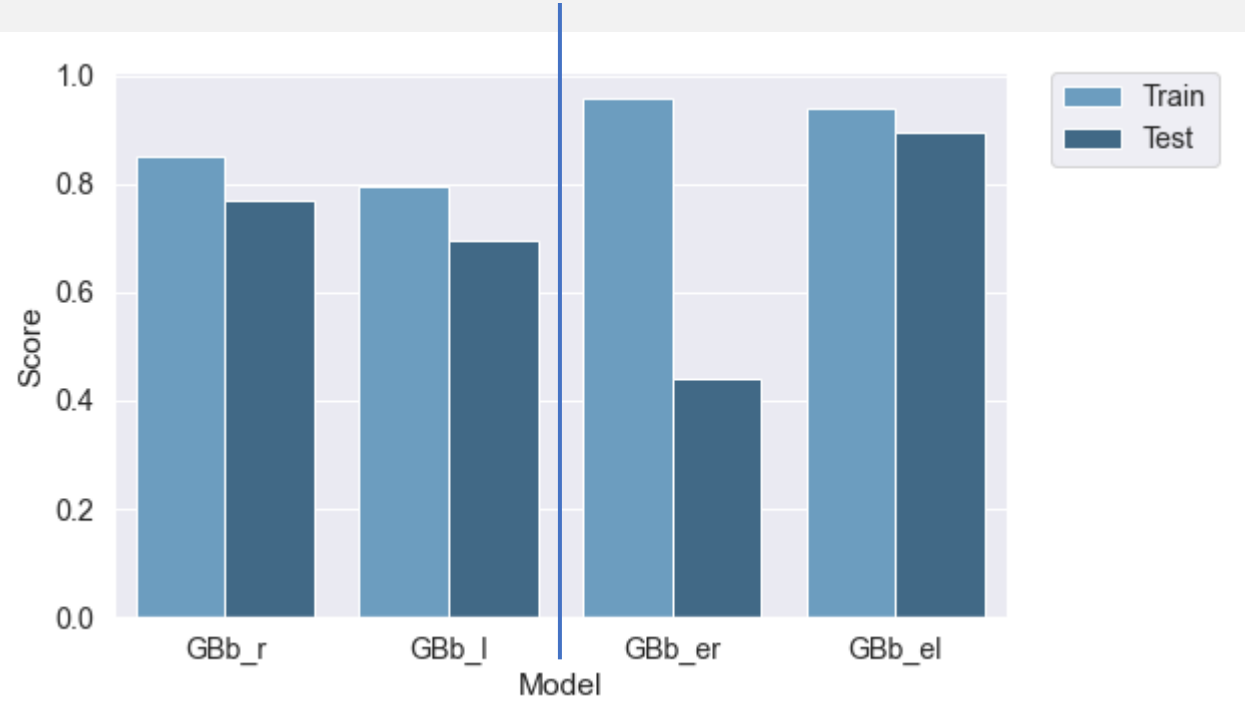
Feature Importances



SiteEnergyUse Results

Meilleurs paramètres

Meilleur modèle : Prop_Type , Age, LargestPropertyUseTypeGFA, SecondLargestPropertyUseTypeGFA, ThirdLargestPropertyUseTypeGFA, Second_Type

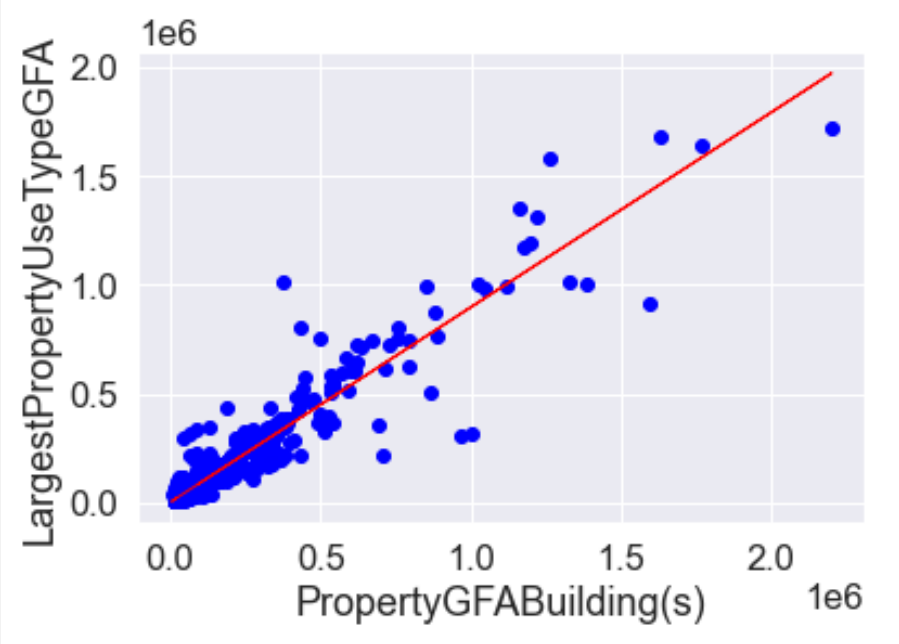


r : regular
l : log
e : EnergyStarScore

GHG Results

GraB_1 : = Model « SiteEnergyUse » //// GrabB_ 2 : Prediction de « SiteEnergyUse »

GraB_ 3 : Prediction de « SiteEnergyUse » + ratio d'Energies (%)



Conclusion

- Les 2 meilleurs modèle sont des modèles ensembliste
- Gradient Boosting

Limitation de l'étude

- Traitement des « comments »
- Traitement des « outliers »
- Peu de données
- ErnergyStarScore

Perspectives

- Retravailler les attributs « Prop_Type »
- Stabilité du modèle
- Les modèles

Fin de la présentation

Merci de m’avoir écouté !