Anticipez les besoins en consommation électrique de bâtiments

Projet n°4 – Parcours Data Scientist – Jérémy Vangansberg

Le cadre du projet

- La ville de **Seattle** souhaite atteindre son objectif de ville neutre en émission de carbone en 2050
- Elle s'intéresse donc à la performance énergétique des bâtiments **non destinés à l'habitation**
- En tant que Data Scientist, je vais utiliser des algorithmes afin de prédire les **émissions de CO2** et la **consommation totale d'énergie** en fonction de certaines variables

Table des matières

La problématique

Problématique

- La ville de Seattle souhaite créer un outil qui permet de prédire la consommation d'énergie et les émissions de CO2 d'un bâtiment à partir de ses **données déclaratives**
- En effet, la ville dispose d'un jeu de données comprenant des relevés de consommation et des données déclaratives par bâtiment
- L'objectif est de **se passer des relevés** qui sont couteux à obtenir

Objectif : prédire une variable continue

- Lorsque l’on souhaite prédire le résultat d’une **variable continue**, nous nous trouvons dans un cas de **régression**
- Je vais tester plusieurs algorithmes de régression, évaluer leurs performances et sélectionner le meilleur

L’enjeux de la fuite de données

- Le **data leakage ou leakage** est un concept de statistiques et de machine learning qui caractérise des données qui ne **devraient pas être disponible au moment de la prédiction**
- Dans notre cas, ce sont les données relatives aux **relevés** de consommation et d’émissions
- Cependant, il est possible de déduire certaines variables à partir de ces informations

L’intérêt d’Energy Star Score ?

- **L’Energy Star Score** ou ESS est une note de 1 à 100 qui permet de mesurer la performance énergétique d’un bâtiment
- Cette information est **fastidieuse** à obtenir
- J’ai donc tenté d’évaluer son intérêt dans les modélisations

Pistes

Avant de vous présenter les données, voici l'hypothèse que j'ai envisagée :

*A priori, quels seront les variables les plus **importantes** pour effectuer cette prédiction ?*

- La surface d'un bâtiment
- L'année de construction du bâtiment
- Les variables déduites à partir des relevés

La présentation des données

Les données

- Deux jeux de données :
  - 2015 : 45 colonnes et 3376 lignes
  - 2016 : 47 colonnes et 3340 lignes

Les données

- Consolider les jeux de données ? (moyenne)
- Utiliser une année pour entraîner les données et une autre pour les tester ?
  - 3284 entrées sont communes aux 2 datasets (soit 97% et 98% des jeux de données)
- L'année 2016 a été retenu car ce sont les données les plus récentes à notre disposition

Les caractéristiques du jeu de données

INFORMATIONS SUR LES VARIABLES

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3376 entries, 0 to 3375
Data columns (total 46 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   OSEBuildingID                    3376 non-null   int64
 1   DataYear                         3376 non-null   int64
 2   BuildingType                     3376 non-null   object
 3   PrimaryPropertyType              3376 non-null   object
 4   PropertyName                     3376 non-null   object
 5   Address                          3376 non-null   object
 6   City                             3376 non-null   object
 7   State                            3376 non-null   object
 8   ZipCode                          3360 non-null   float64
 9   TaxParcelIdentificationNumber    3376 non-null   object
 10  CouncilDistrictCode              3376 non-null   int64
 11  Neighborhood                     3376 non-null   object
 12  Latitude                         3376 non-null   float64
 13  Longitude                        3376 non-null   float64
 14  YearBuilt                        3376 non-null   int64
 15  NumberofBuildings                3368 non-null   float64
 16  NumberofFloors                   3376 non-null   int64
 17  PropertyGFATotal                 3376 non-null   int64
 18  PropertyGFAParking               3376 non-null   int64
 19  PropertyGFABuilding(s)           3376 non-null   int64
 20  ListOfAllPropertyUseTypes        3367 non-null   object
 21  LargestPropertyUseType           3356 non-null   object
 22  LargestPropertyUseTypeGFA        3356 non-null   float64
 23  SecondLargestPropertyUseType     1679 non-null   object
 24  SecondLargestPropertyUseTypeGFA  1679 non-null   float64
 25  ThirdLargestPropertyUseType      596 non-null    object
 26  ThirdLargestPropertyUseTypeGFA   596 non-null    float64
 27  YearsENERGYSTARCertified         119 non-null    object
 28  ENERGYSTARScore                  2533 non-null   float64
 29  SiteEUI(kBtu/sf)                 3369 non-null   float64
 30  SiteEUIWN(kBtu/sf)               3370 non-null   float64
 31  SourceEUI(kBtu/sf)               3367 non-null   float64
 32  SourceEUIWN(kBtu/sf)             3367 non-null   float64
 33  SiteEnergyUse(kBtu)              3371 non-null   float64
 34  SiteEnergyUseWN(kBtu)            3370 non-null   float64
 35  SteamUse(kBtu)                   3367 non-null   float64
 36  Electricity(kWh)                 3367 non-null   float64
 37  Electricity(kBtu)                3367 non-null   float64
 38  NaturalGas(therms)               3367 non-null   float64
 39  NaturalGas(kBtu)                 3367 non-null   float64
 40  DefaultData                      3376 non-null   bool
 41  Comments                         0 non-null      float64
 42  ComplianceStatus                 3376 non-null   object
 43  Outlier                          32 non-null     object
 44  TotalGHGEmissions                3367 non-null   float64
 45  GHGEmissionsIntensity            3367 non-null   float64
dtypes: bool(1), float64(22), int64(8), object(15)
memory usage: 1.2+ MB
```

- Index
- Données déclaratives
- ESS
- Relevé de consommation
- Autres Données déclaratives
- Relevé d'émissions

La préparation des données

Variables catégorielles : Pré-sélection

LES VARIABLES CATÉGORIELLES : TAUX DE REMPLISSAGE ET CARDINALITÉ

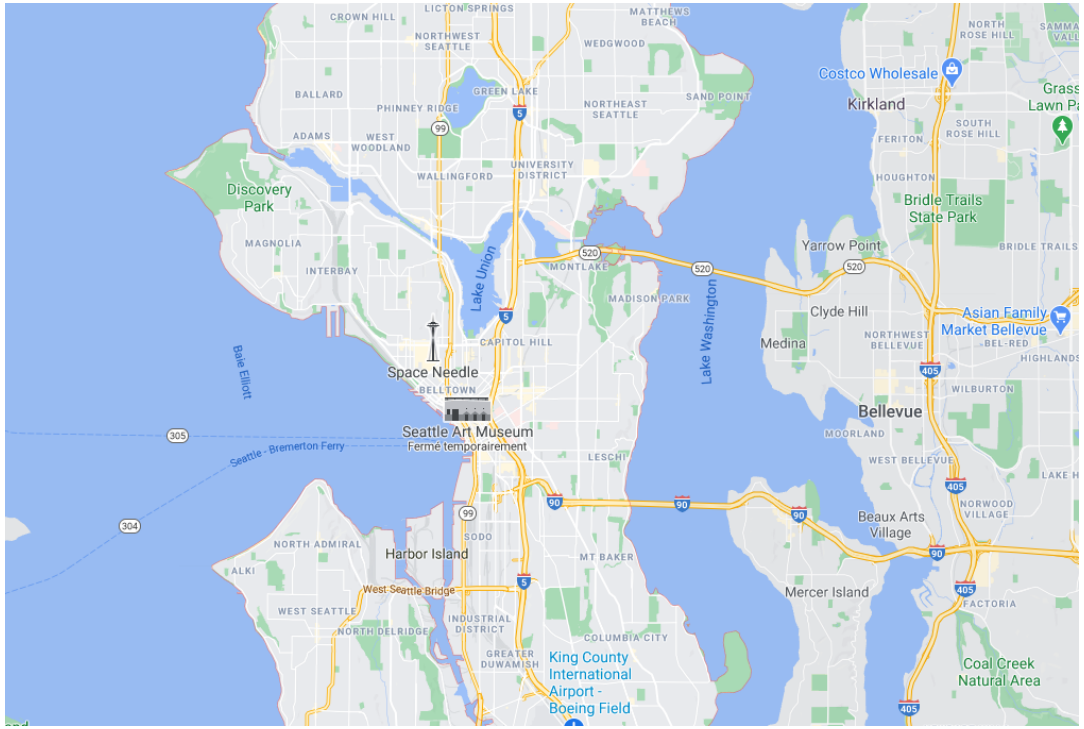
Nom de la variable	Taux de remplissage (%)	Cardinalité
BuildingType	**100**	**8**
PrimaryPropertyType	**100**	**24**
PropertyName	100	3362
Address	100	3354
City	100	1
State	100	1
TaxParcelIdentificationNumber	100	3268
Neighborhood	**100**	**19**
ListOfAllPropertyUseTypes	100	467
LargestPropertyUseType	99	57
SecondLargestPropertyUseType	50	51
ThirdLargestPropertyUseType	18	45
YearsENERGYSTARCertified	4	66
ComplianceStatus	**100**	**4**
Outlier	1	3

Leakage : Les données relatives aux relevés

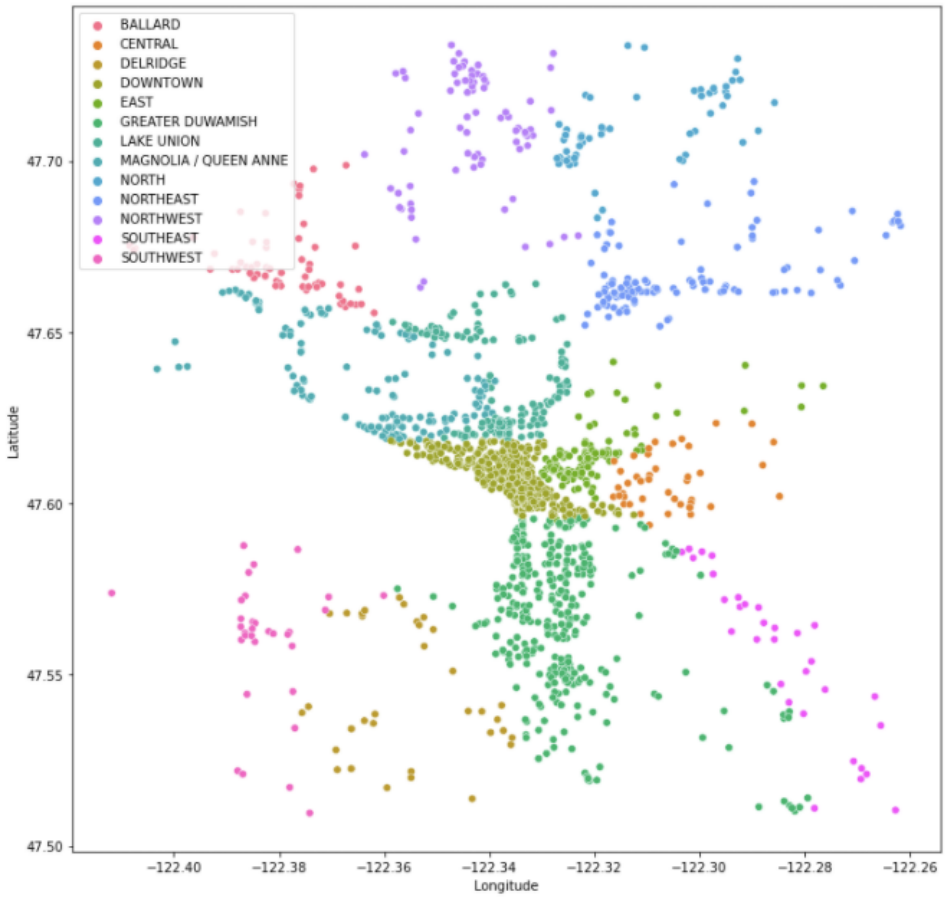
- Objectif : Prévoir la consommation et les émissions sans les informations des relevés
- Ces variables ont été **supprimées**
- Cependant de nouvelles variables ont été créées :
  - La consommation/émission moyen par type de bâtiment
- Suite à cette modification, j'estime que la variable « PrimaryPropertyType » n'est plus utile car elle est **colinéaire** à ce nouvel ensemble de variables

Les données spatiales

VUE DE SEATTLE SUR GOOGLE MAP



SCATTER PLOT : LONGITUDE ET LATITUDE

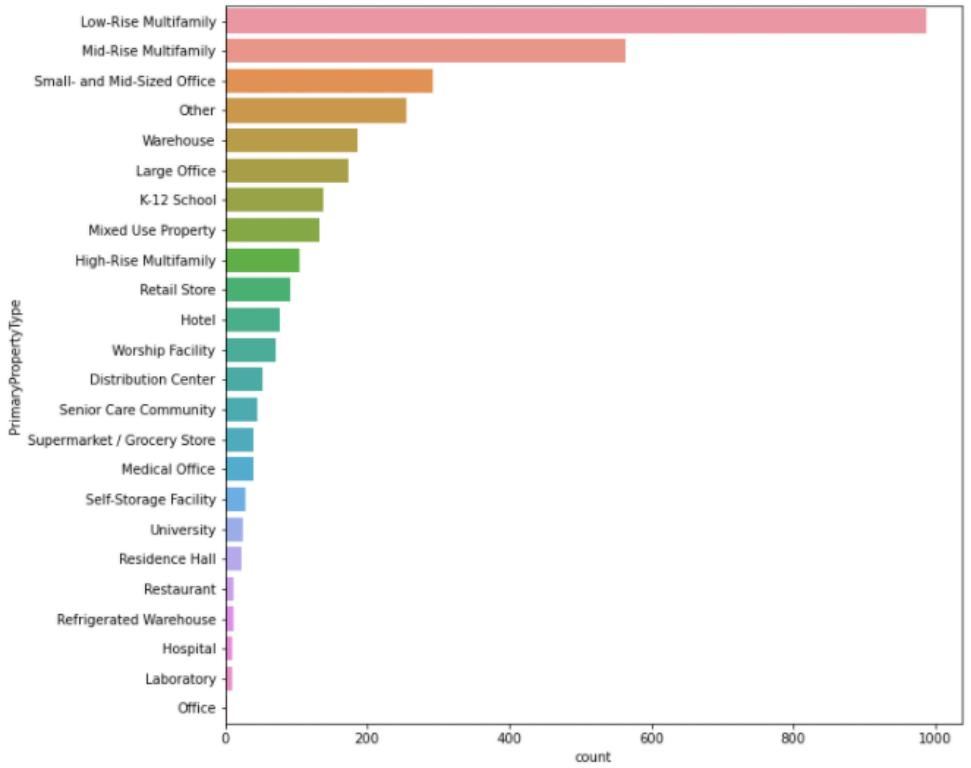


Focus sur les bâtiments non-résidentiels

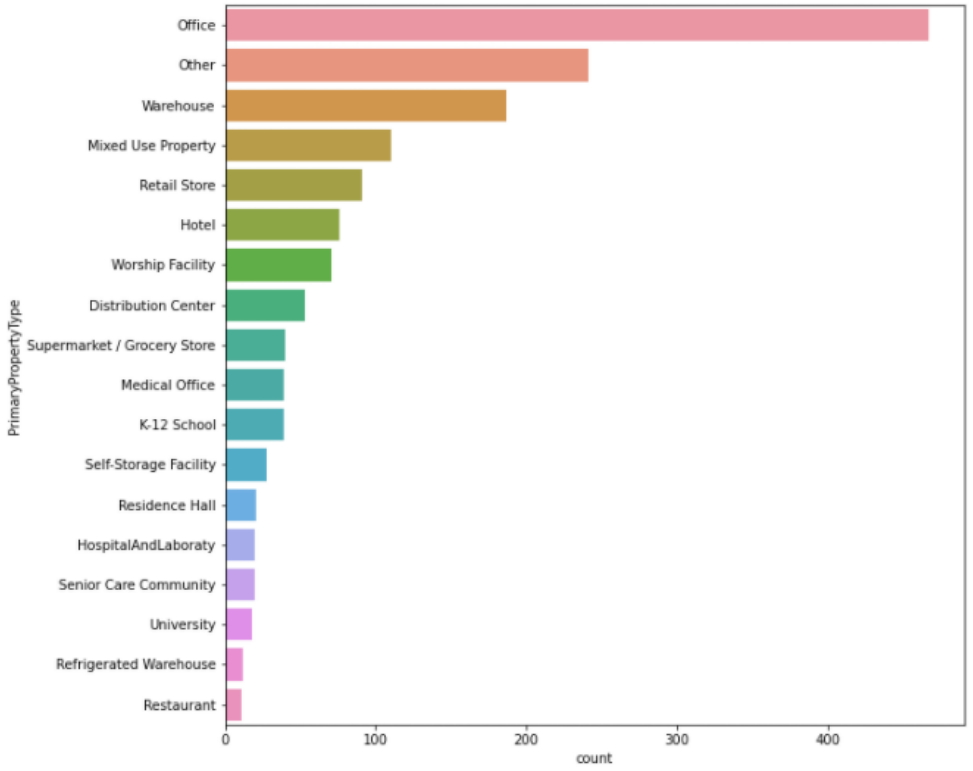
- « BuildingType » valeurs conservées :
  - 'NonResidential’
  - 'Nonresidential COS’
  - 'Nonresidential WA’
- Suite à cette opération, le jeu de données passe de 3376 entrées à 1544
- Envisager des solutions pour séparer les données avec une stratification en fonction du type de bâtiment

Retravail des catégories

RÉPARTITION AVANT NETTOYAGE



RÉPARTITION APRÈS NETTOYAGE (GROUPE MINIMUM : 12)



- 'Small- and Mid-Sized Office', 'Large Office', 'Office' ➔ 'Office'
- 'Non-Refrigerated Warehouse', 'Warehouse' ➔ 'Warehouse'
- 'Hospital', 'Laboratory' ➔ 'HospitalAndLaboraty'

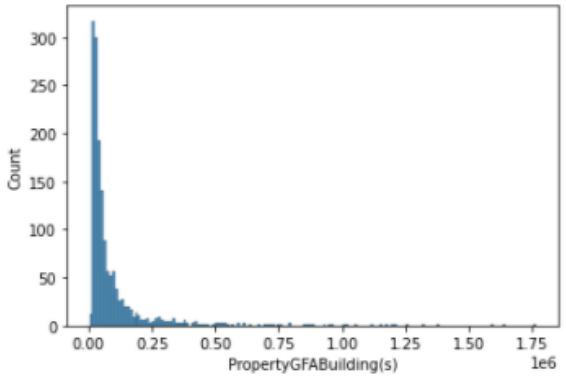
Pipeline

- Données numériques
  - Transformation en log + 1 :
    - ❑ 'PropertyGFABuilding(s)'
    - ❑ 'NumberofBuildings'
    - ❑ 'NumberofFloors'
    - ❑ 'ENERGYSTARScore'
  - Transformation avec Standard Scaler

$$z = \frac{x - \mu}{\sigma}$$

- Données catégorielles
  - Transformation avec one hot encoder

DISTRIBUTION



Les modélisations

Séparation en train/test set

- Etant donné que le jeu de données est **petit**, une séparation des données par utilisation du hasard peut conduire à un déséquilibre important
- J'ai donc utilisé une **méthode de stratification** afin d'avoir un train set représentatif du test set
- En général, la stratification est plutôt utilisé dans le cas d'une classification
- J'ai stratifié les données en fonction de « PrimaryPropertyType »

Les modèles envisagés

- Modèles **linéaires** :
  - Régression linéaire
  - Lasso
- Modèles **non-linéaires** :
  - Random Forest Regressor
  - Gradient Boosting Regressor

Les métriques et outils d’évaluation

- R^2
- RMSE
- Validation Curves
- Temps d’entraînement

Les variantes

- Les modèles ont été entraînés enfin de prédire la valeur de deux variables cibles:
  - Target 1 : Emissions de CO2
  - Target 2 : Consommation d'énergie totale

- Afin d'évaluer l'intérêt de la variable Energy Star Score, chaque modèle a une version avec et sans cette variable

Baseline : Dummy Regressor

- Stratégie utilisée : la moyenne
- Score Target 1 :
  - RMSE : 1,54
  - R^2 : - 0,0016
- Score Target 2 :
  - RMSE : 1,58
  - R^2 : - 6,51

Régression linéaire (1)

- Bref définition :

Algorithme de régression le plus répandu. Il cherche à **minimiser la somme des moindres carrés** en optimisant les coefficients associés à chaque variable

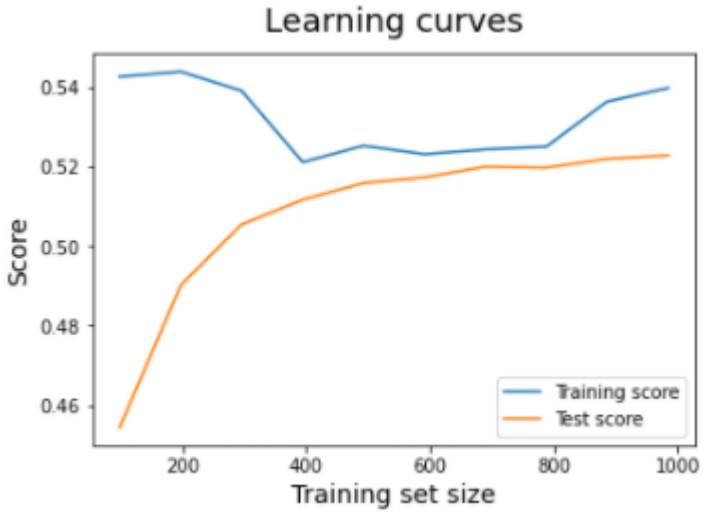
- Hyperparamètres : ∅

RÉCAPITULATIF DES SCORES

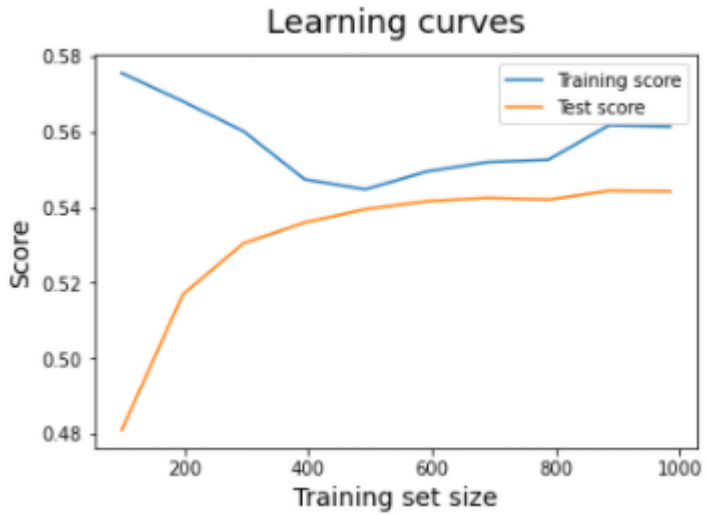
Target	R^2	RMSE
N°1	0,575	1,002
N°1 (ESS)	0,598	0,974
N°2	0,583	1,019
N°2 (ESS)	0,614	0,984

Régression linéaire (2)

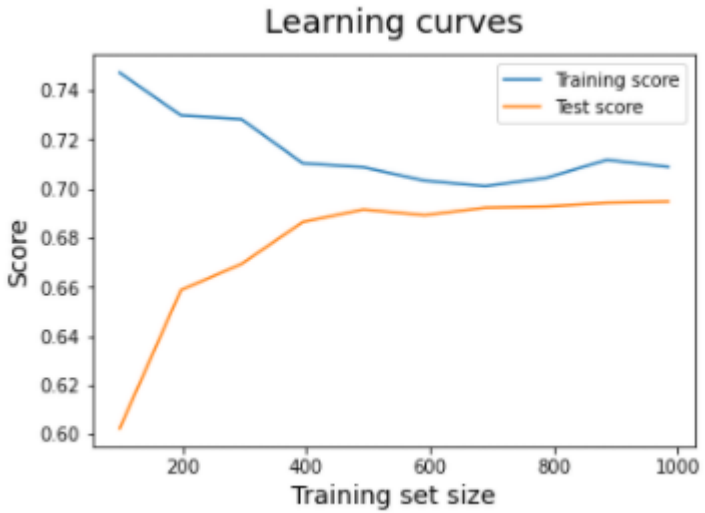
TARGET 1



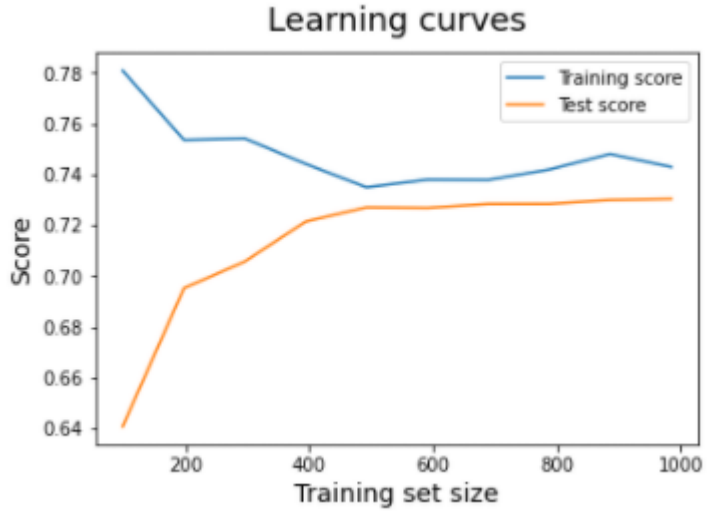
TARGET 1 ESS



TARGET 2



TARGET 2 ESS



Lasso (1)

- Bref définition :

C’est un algorithme de régression linéaire avec un terme de régularisation qui permet de réduire l’overfitting

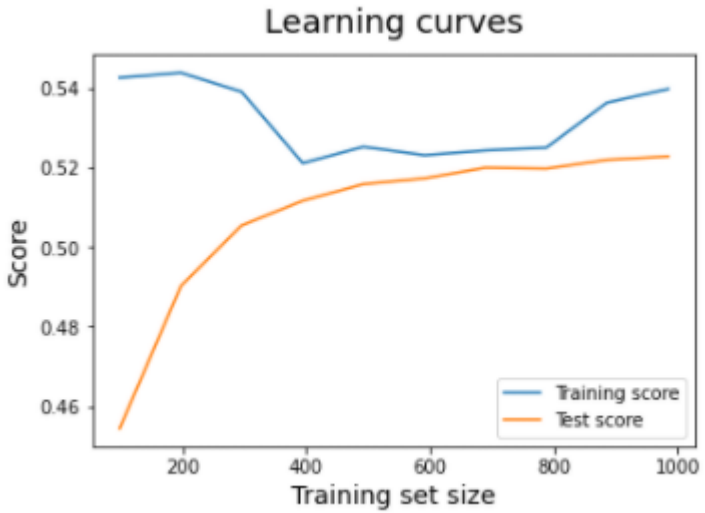
- Hyperparamètre : alpha

RÉCAPITULATIF DES SCORES

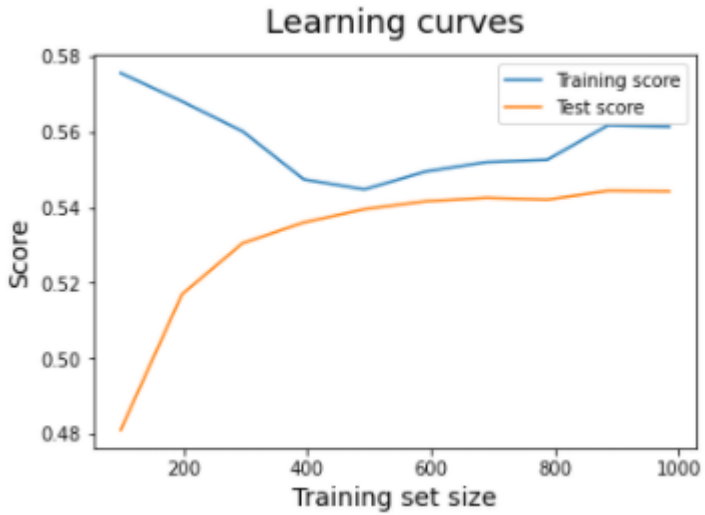
Target	R^2	RMSE
N°1	0,575	1,002
N°1 (ESS)	0,598	0,974
N°2	0,578	1,028
N°2 (ESS)	0,603	0,997

Lasso (2)

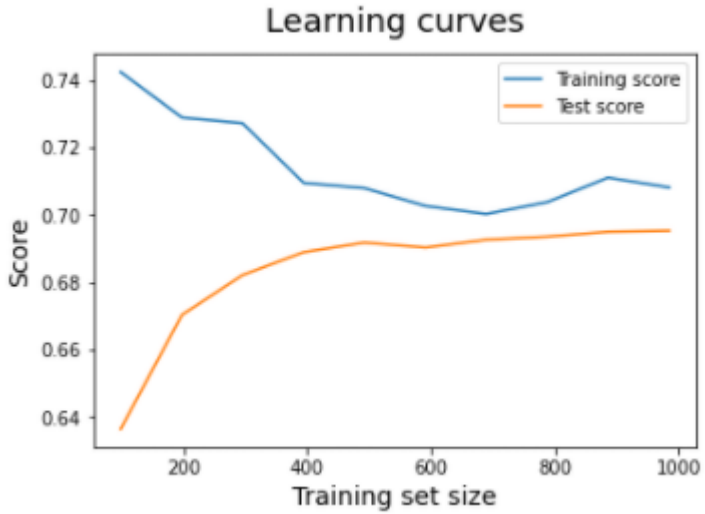
TARGET 1



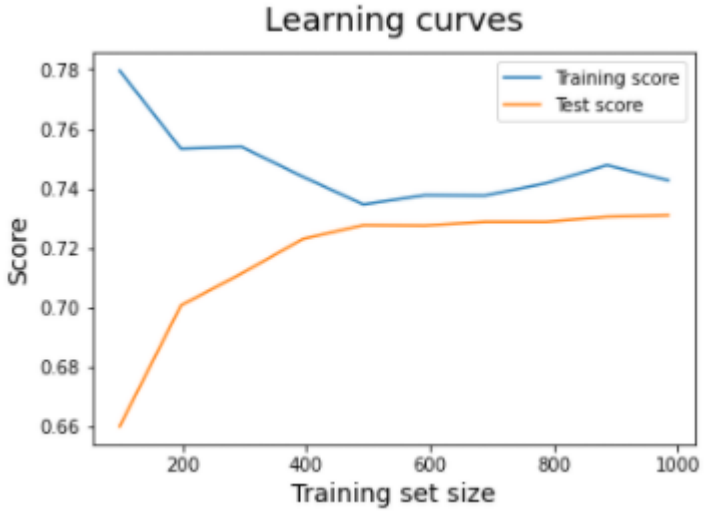
TARGET 1 ESS



TARGET 2



TARGET 2 ESS



Lasso (3)

TARGET 1 : COEFFICIENTS

	features	coefs
10	NaturalGas(kBtu)_mean_property_type	1.527322
0	YearBuilt	0.824314
3	PropertyGFABuilding(s)	0.519850
11	ComplianceStatus_Compliant	0.116515
13	ComplianceStatus_Non-Compliant	0.058373
1	NumberofBuildings	0.057067
6	SourceEUI(kBtu/sf)_mean_property_type	0.028426
8	SteamUse(kBtu)_mean_property_type	0.004421
4	Latitude	0.000000
12	ComplianceStatus_Error - Correct Default Data	-0.013927
2	NumberofFloors	-0.041677
7	SiteEUI(kBtu/sf)_mean_property_type	-0.049575
9	Electricity(kBtu)_mean_property_type	-1.098427
5	Longitude	-1.189855

TARGET 2 : COEFFICIENTS

	features	coefs
0	YearBuilt	0.850800
3	PropertyGFABuilding(s)	0.553800
10	NaturalGas(kBtu)_mean_property_type	0.502111
6	SourceEUI(kBtu/sf)_mean_property_type	0.117584
1	NumberofBuildings	0.041505
2	NumberofFloors	0.038806
11	ComplianceStatus_Compliant	0.023170
13	ComplianceStatus_Non-Compliant	0.003975
8	SteamUse(kBtu)_mean_property_type	0.000224
4	Latitude	-0.000000
9	Electricity(kBtu)_mean_property_type	0.000000
7	SiteEUI(kBtu/sf)_mean_property_type	-0.024896
12	ComplianceStatus_Error - Correct Default Data	-0.032265
5	Longitude	-0.834754

Random Forest (1)

- Bref définition :

Méthode ensembliste et parallèle. Cet algorithme combine plusieurs arbres de décision simple (apprenants faibles) afin d'effectuer une prédiction.

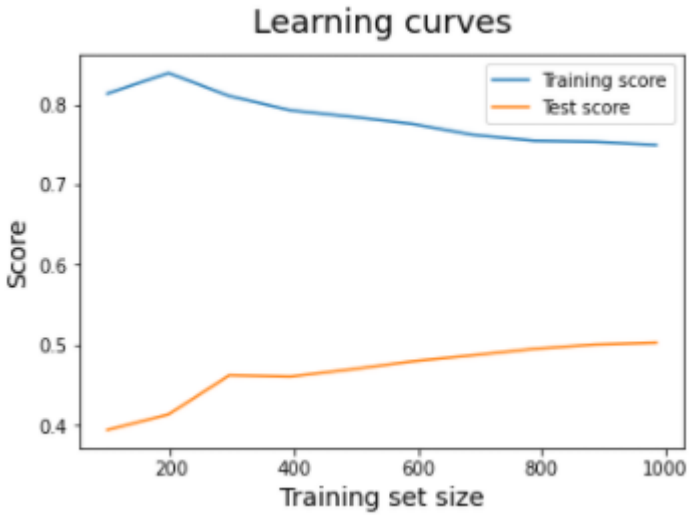
- Hyperparamètres principaux : *n_estimators, max_features, max_depth, min_samples_split, min_samples_leaf*

RÉCAPITULATIF DES SCORES

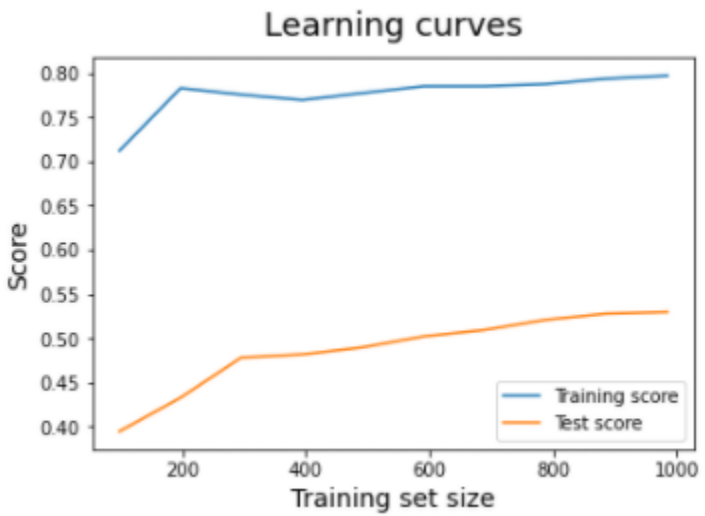
Target	R^2	RMSE
N°1	0,557	1,023
N°1 (ESS)	0,580	0,995
N°2	0,574	1,032
N°2 (ESS)	0,607	0,997

Random Forest (2)

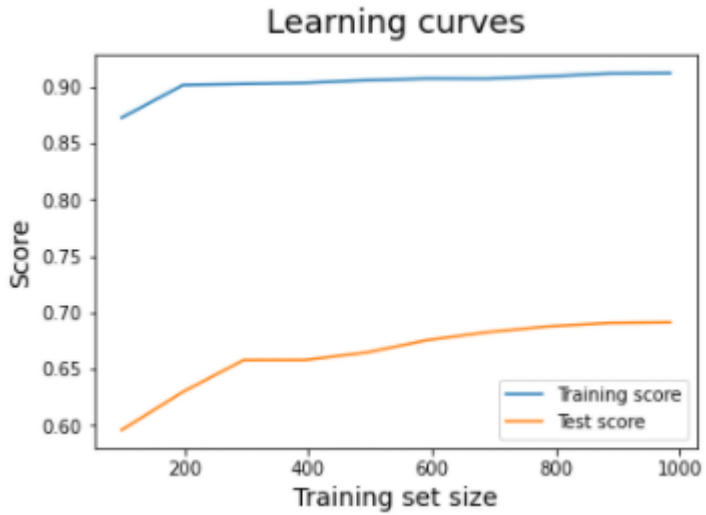
TARGET 1



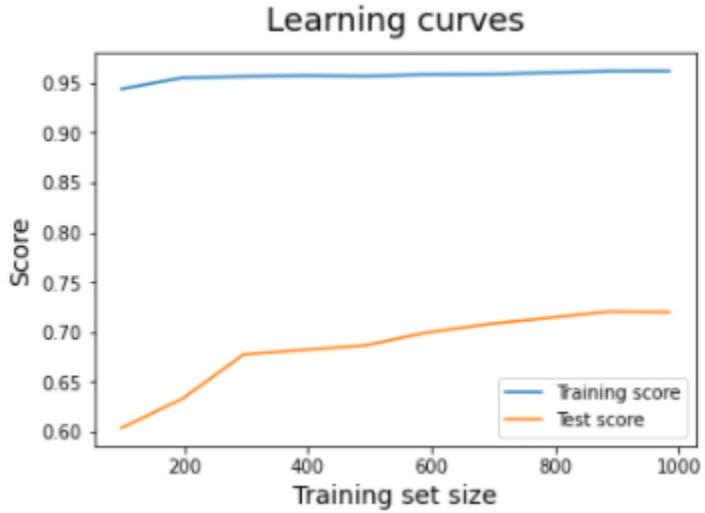
TARGET 1 ESS



TARGET 2



TARGET 2 ESS



Gradient Boosting Regressor (1)

- Définitions

Méthode ensembliste séquentielle. Cette algorithme combine plusieurs apprenants faibles. De manière successive, les arbres vont accorder plus de poids aux erreurs faites par les estimateurs précédents

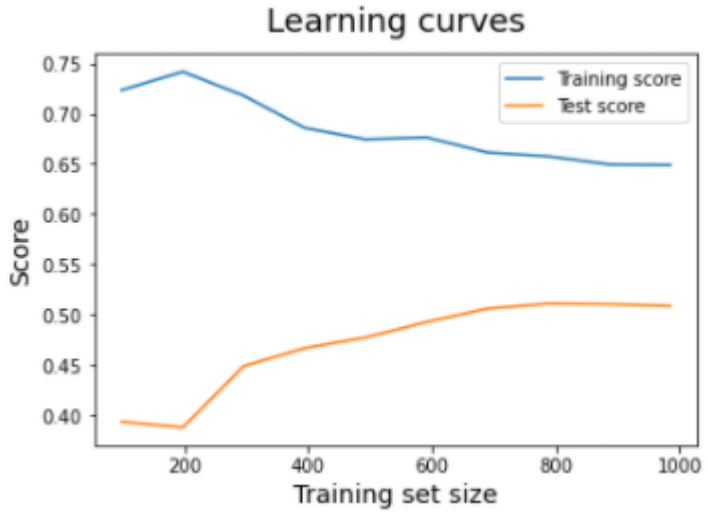
- Hyperparamètres : *n_estimators, max_features, max_depth, min_samples_split, min_samples_leaf, subsample, learning_rate*

RÉCAPITULATIF DES SCORES

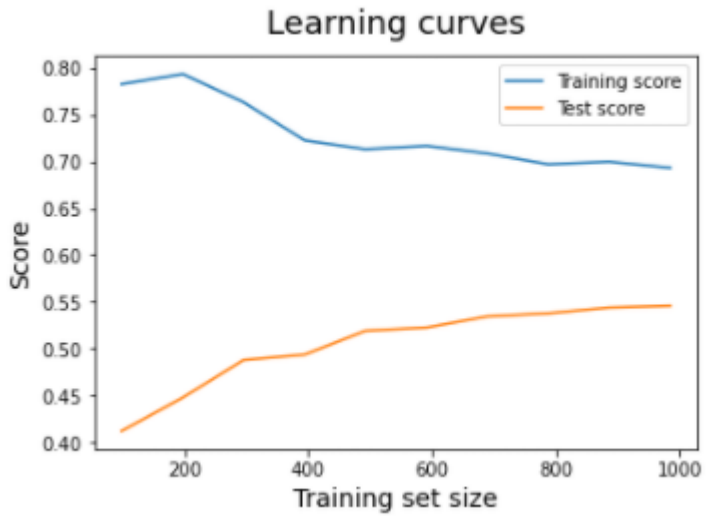
Target	R^2	RMSE
N°1	0,589	0,985
N°1 (ESS)	0,608	0,962
N°2	0,592	1,101
N°2 (ESS)	0,625	0,969

Gradient Boosting Regressor (2)

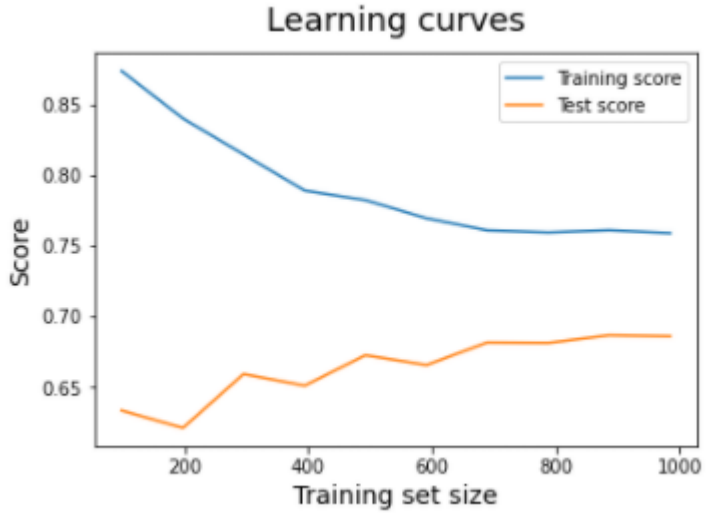
TARGET 1



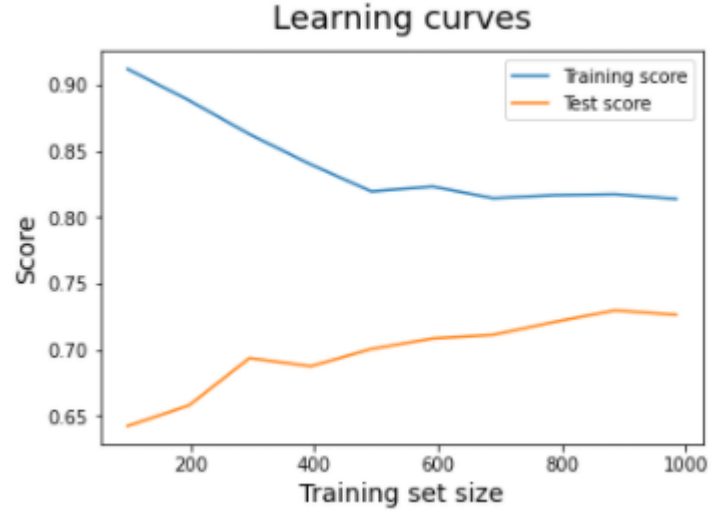
TARGET 1 ESS



TARGET 2

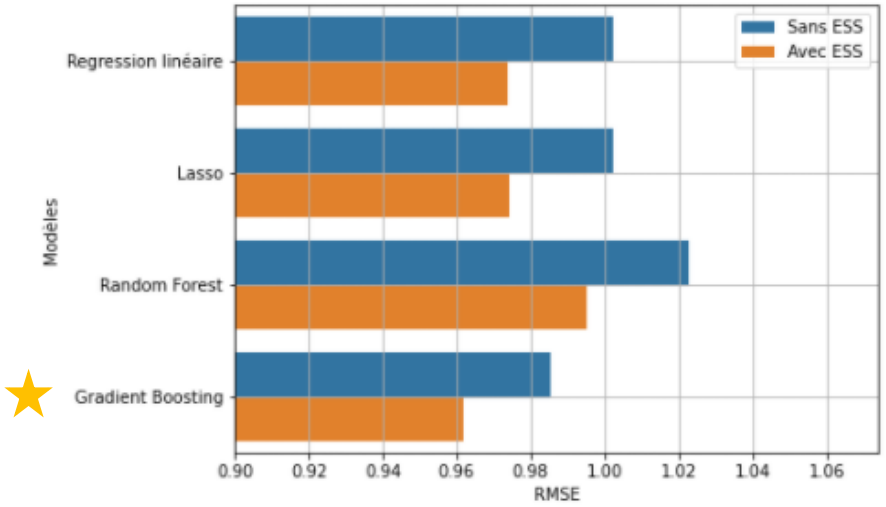


TARGET 2 ESS

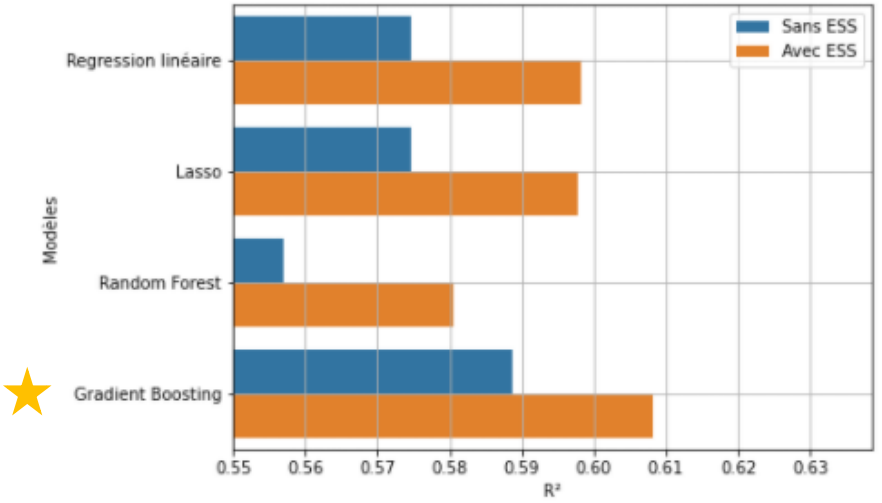


Comparatif des scores (début de l'échelle ≠0)

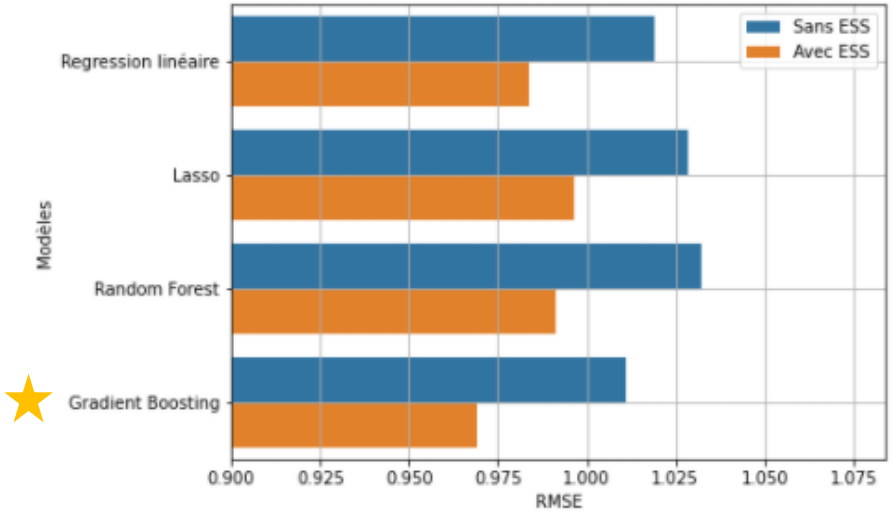
TARGET 1 : RMSE



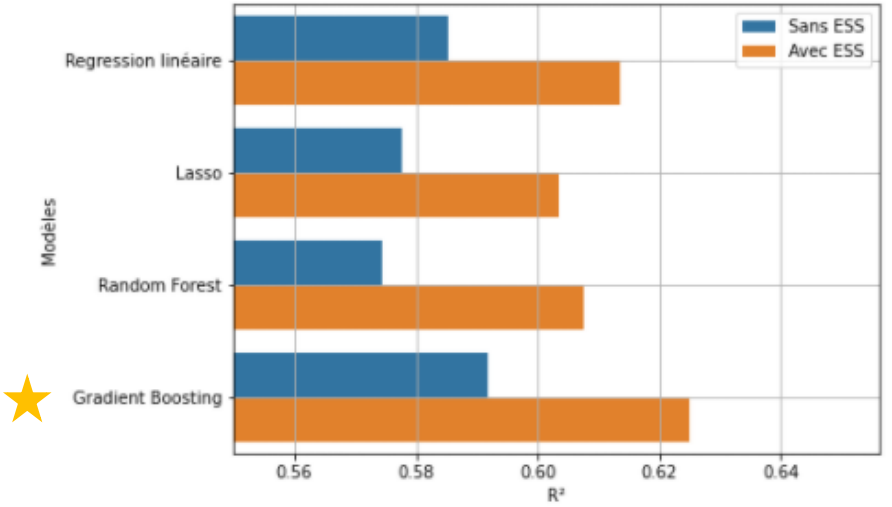
TARGET 1 : R^2



TARGET 2 : RMSE

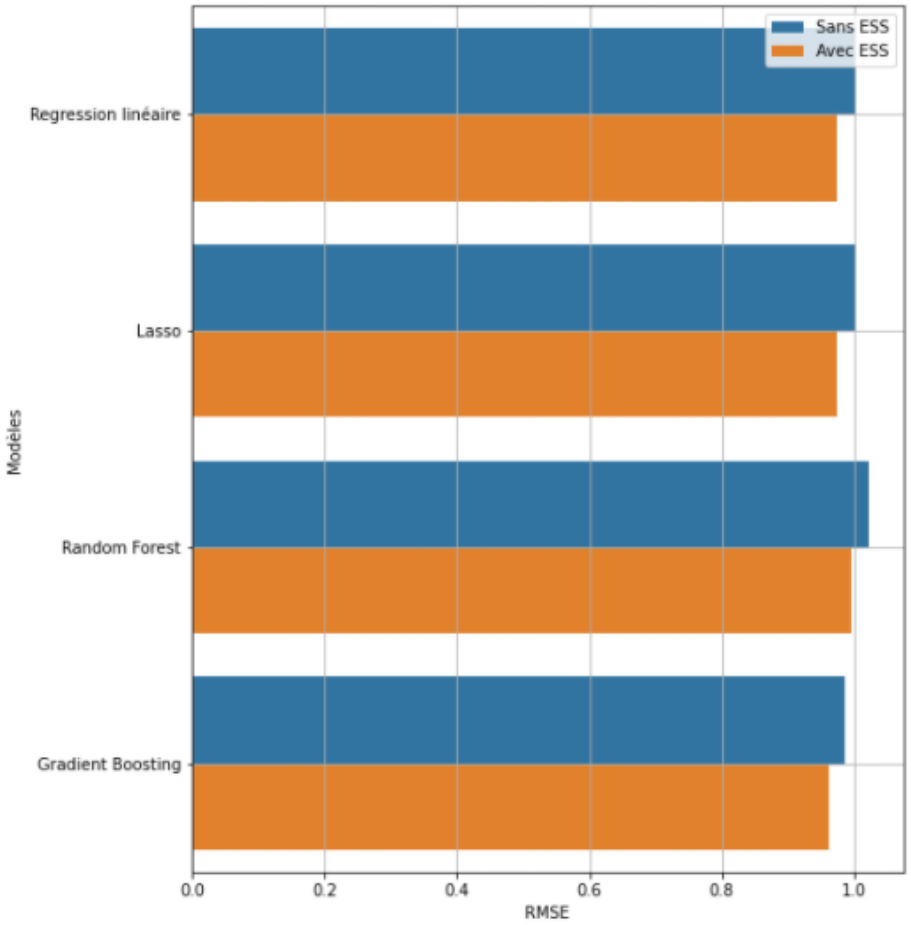


TARGET 2 : R^2



Comparatif des scores (début de l'échelle à 0)

TARGET 1 : RMSE



Bilan :

1. La différence de performance entre les algorithmes est faible
2. L'intérêt de la variable ESS est limité

Autres indicateurs de performance : temps

- Les temps d’exécution des algorithmes sont très courts (– d’une seconde). C’est donc un estimateur trivial
- Toutefois, la recherche d’hyperparamètre peut atteindre jusqu’à 5 minutes par recherche.

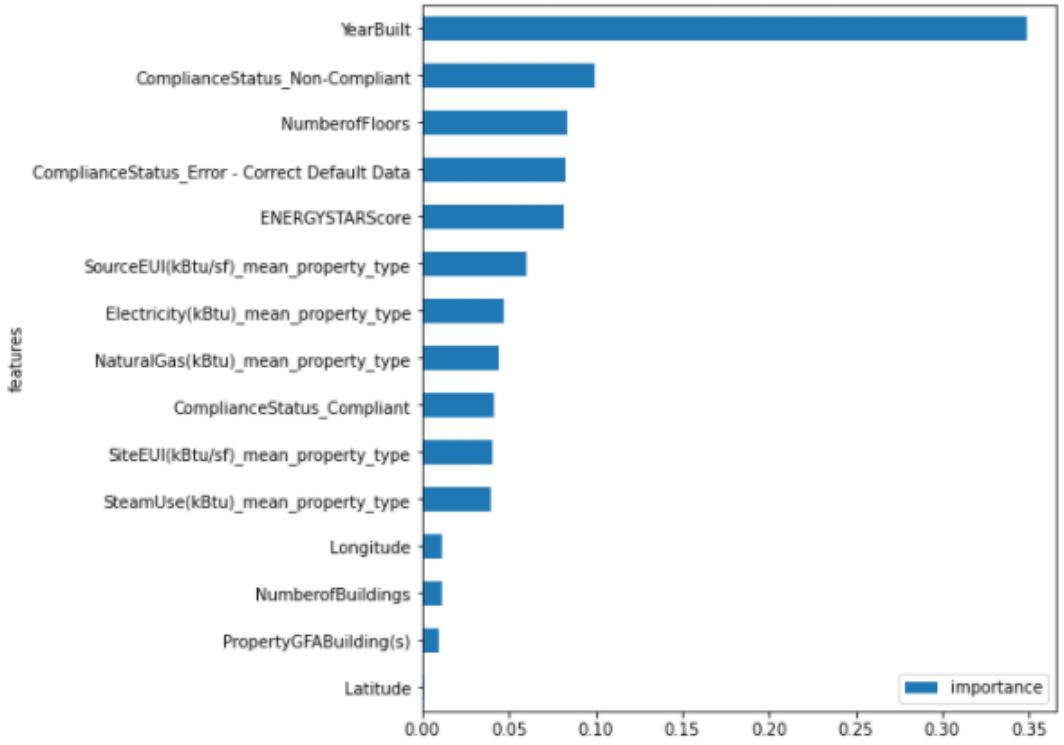
Le modèle finale

Le modèle final : Gradient Boosting

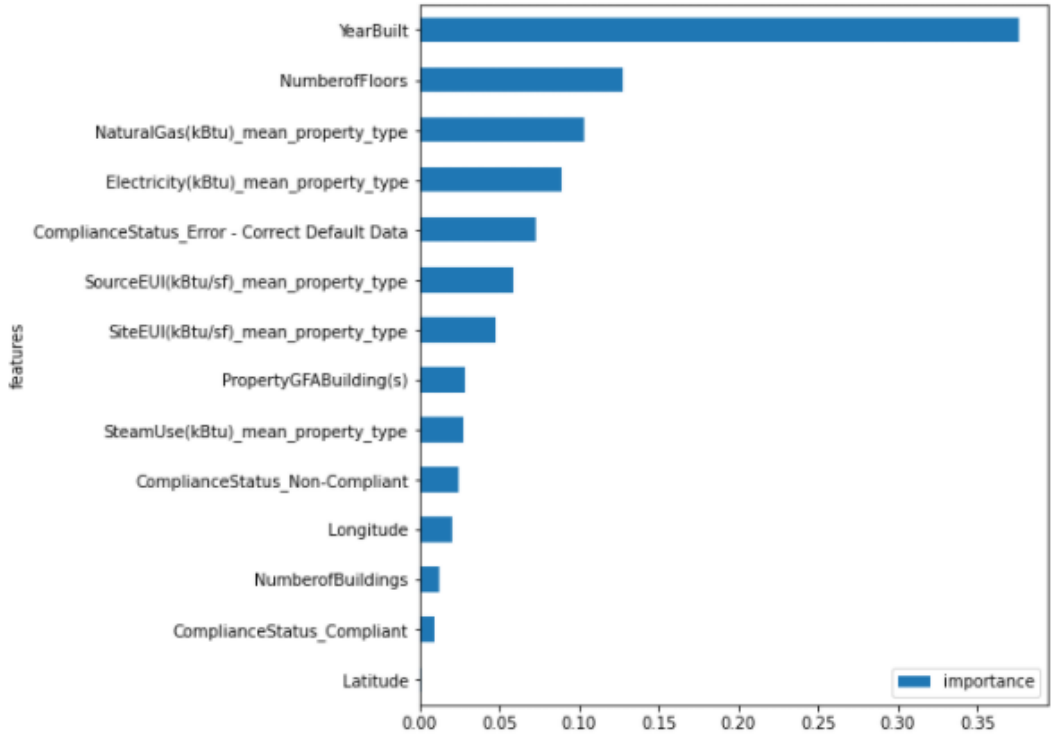
- **Stratégie d'optimisation** des paramètres en plusieurs étapes :
  1. Choix d'un learning rate assez élevé : 0,1
  2. Recherche du nombre d'estimateurs : de 10 à 150 par palier de 5
  3. Recherche des paramètres spécifiques aux arbres :
     - min_samples_split : de 2 à 40 par palier de 2
     - min_samples_leaf : de 1 à 10 par palier de 1
     - max_depth : de 2 à 8 par palier de 1
     - Subsample : [0.6,0.7,0.8,0.85,0.9]
  4. Réduction du learning rate et augmentation proportionnelle du nombre d'estimateurs : division du learning rate par deux et multiplication du nombre d'estimateurs par deux

L’importance des variables

TARGET 1



TARGET 2



Les erreurs

TARGET 1 : ERREUR PAR CATÉGORIE

PrimaryPropertyType	
K-12 School	0.462860
Medical Office	0.574584
Office	0.592915
Hotel	0.601901
Worship Facility	0.650862
Self-Storage Facility	0.684444
Mixed Use Property	0.725369
Retail Store	0.735587
Refrigerated Warehouse	0.827915
Other	0.857088
Warehouse	0.875245
HospitalAndLaboraty	0.912684
University	0.917669
Residence Hall	0.920644
Senior Care Community	0.926131
Distribution Center	0.932522
Restaurant	1.301800
Supermarket / Grocery Store	1.415109

Name: error, dtype: float64

TARGET 2 : ERREUR PAR CATÉGORIE

PrimaryPropertyType	
Supermarket / Grocery Store	0.313725
Office	0.371728
Worship Facility	0.384909
Hotel	0.386181
Residence Hall	0.396569
University	0.409546
Mixed Use Property	0.444804
Refrigerated Warehouse	0.474045
Retail Store	0.487051
K-12 School	0.494121
Medical Office	0.539435
HospitalAndLaboraty	0.588498
Senior Care Community	0.648775
Warehouse	0.652931
Distribution Center	0.755905
Self-Storage Facility	0.850614
Other	0.868855
Restaurant	1.076984

Name: error, dtype: float64

Variables modifiées

Top/Bot erreur

Conclusion

- ESS a un intérêt limité
- Le faible nombre de données d’entraînement rend les modèles instables
- Le meilleur modèle explique environ 62 % de la variance (R^2)

Pistes d’approfondissements

- Tester d’autres modèles comme XGBoost
- Calculer le taux d’évolution des variables entre les années