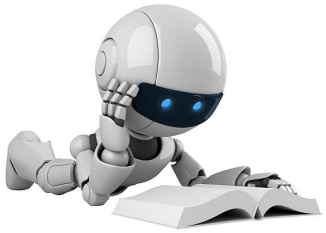


Anticipez les besoins en consommation électrique de batiments

L.khellouf[1] Mentor: B.Beaufils[2]

[12]OpenClassRooms

Projet 3, Janvier 2021

Sommaire

Présentation

La ville de Seattle veut atteindre son objectif de ville neutre en émissions de carbone en 2050 et pour cela des relevés manuels minutieux on été effectués en 2015 et 2016. Ces relevés sont très couteux et il reste encore des bâtiments à mesurer.Et évaluer l'intéret de l'ENERGY STAR Score pour la prédiction d'émissions.

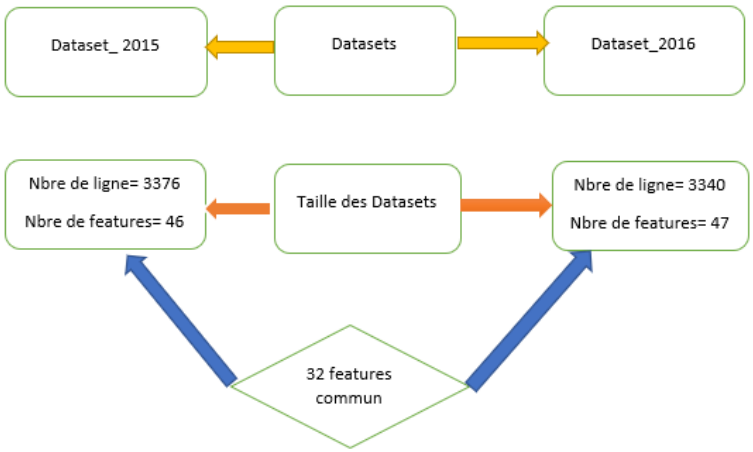
Mission:

Prédictions des émissions de CO_2 et de consommation totale d'énergie à partir des données déjà existantes.

Données:

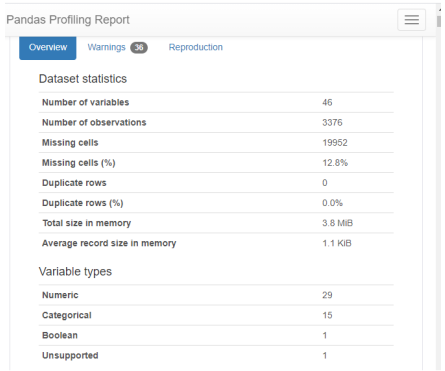
https://www.kaggle.com/city-of-seattle/sea-building-energy-benchmarking2015-building-energy-benchmarking.csv

Présentation des données



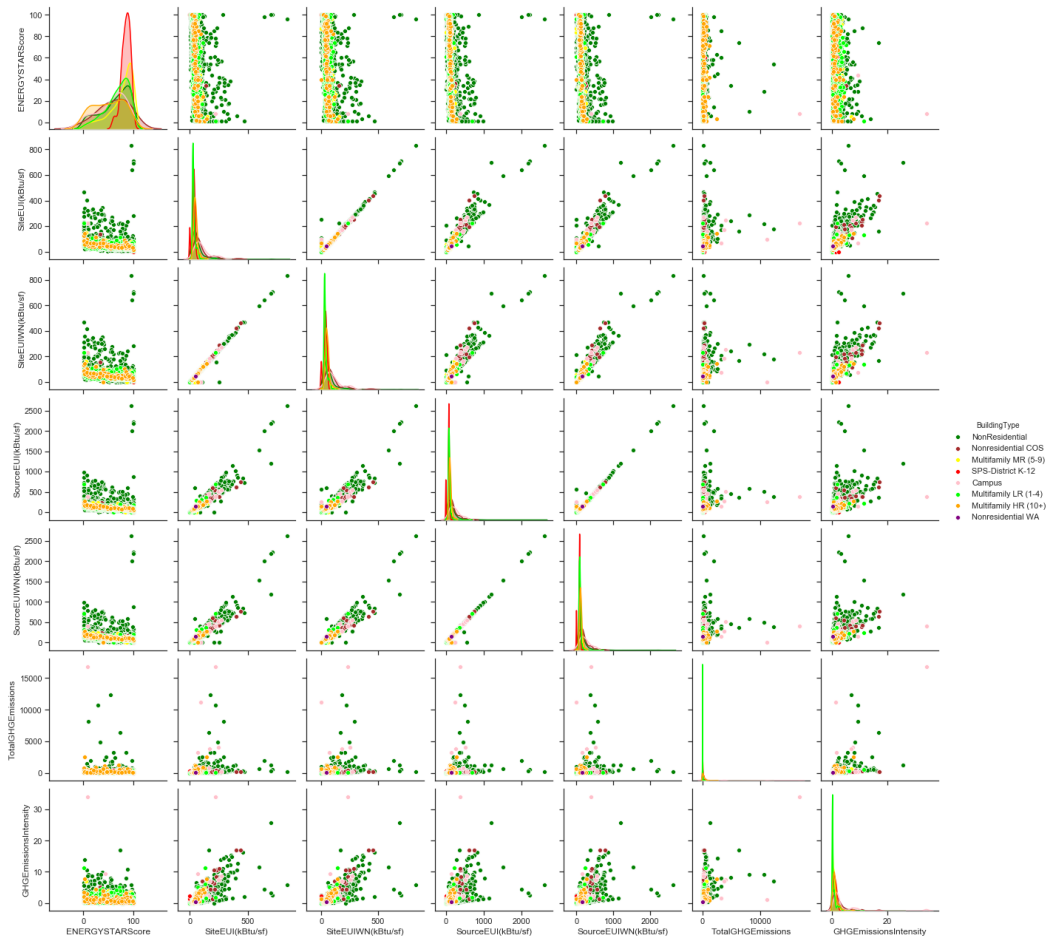
la nouvelle datasets est obtenu à partir de la concaténation de dataset_2015 et dataset_2016, avec une remise en forme des colonnes et la suppression des doublons avec moins de NAN.

Profiling pandas de la dataset_2016



Analyse Exploratoire

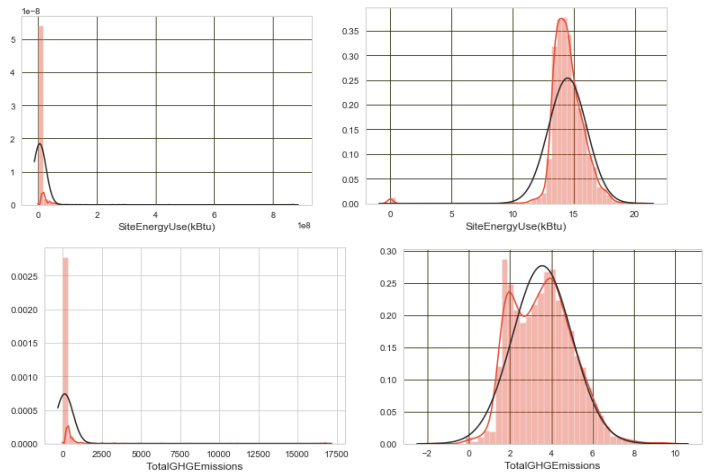
Visualisation de pair plot



Les Target:

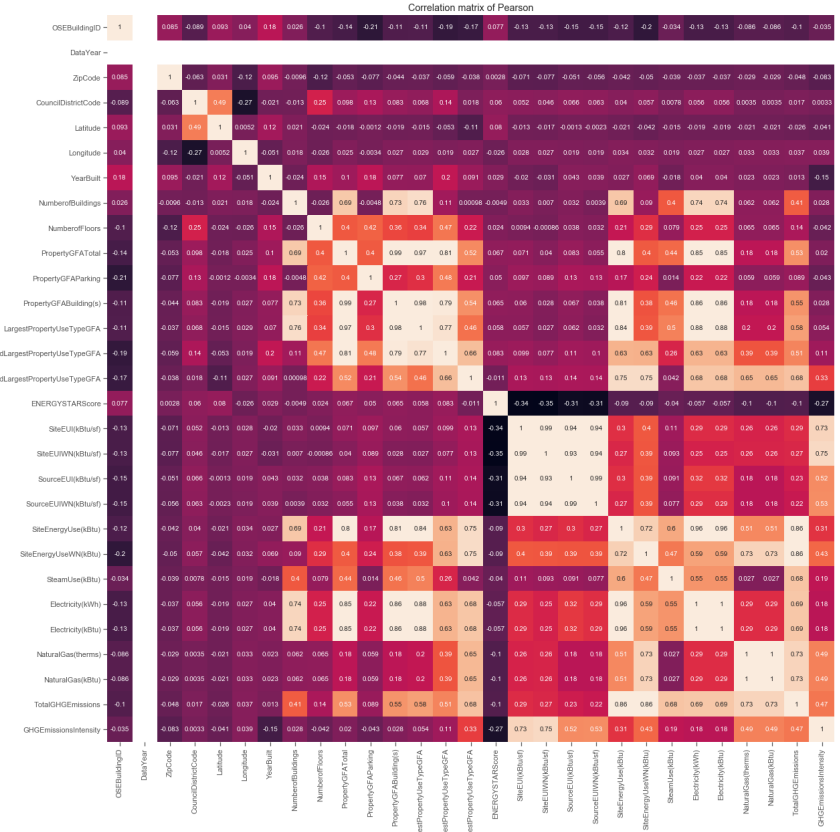
- TotalGHGEmissions: La quantité totale d'émissions de gaz.
- SiteEnergyUse(kBtu): La quantité annuelle d'énergie consommée par la propriété à partir de toutes les sources d'énergie.

Distribution des targets:

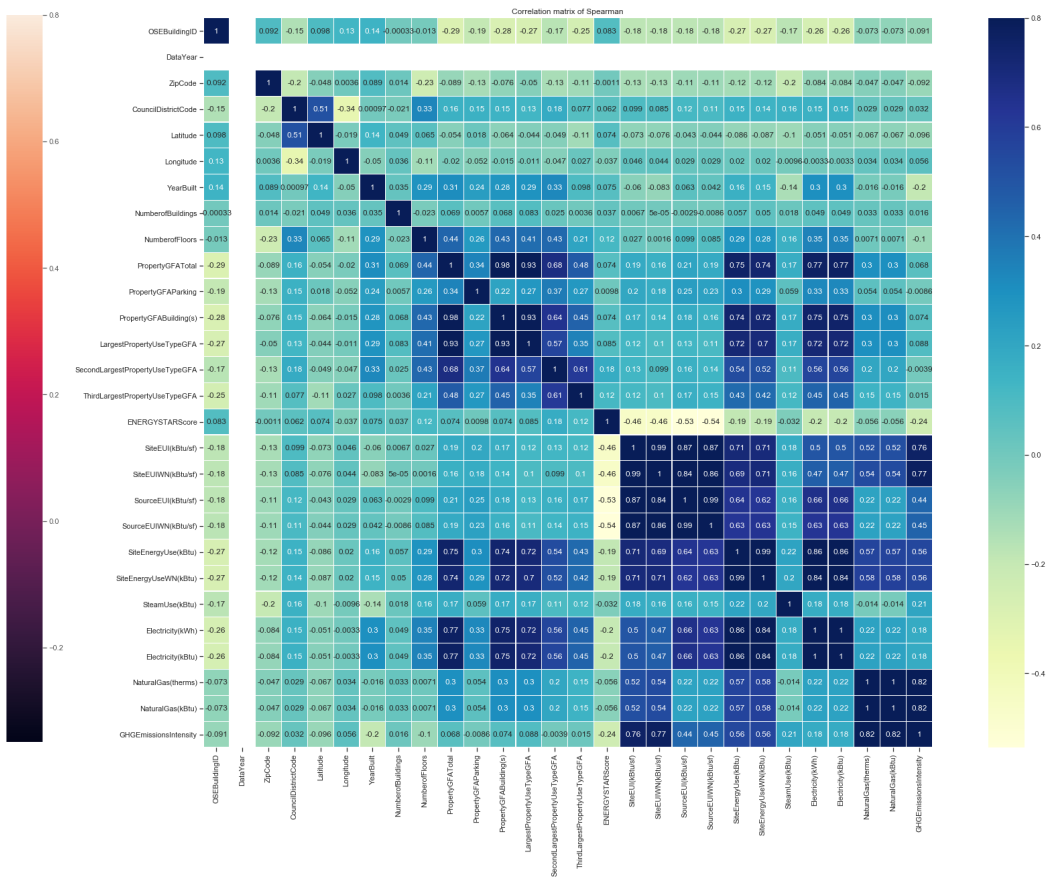


Analyse Exploratoire

Matrice de corrélation de Pearson

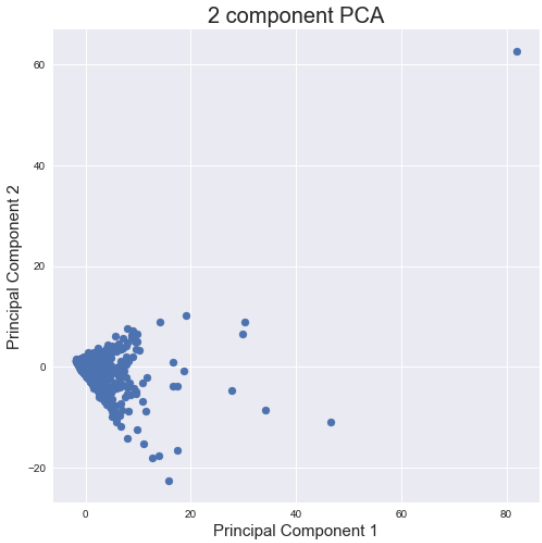


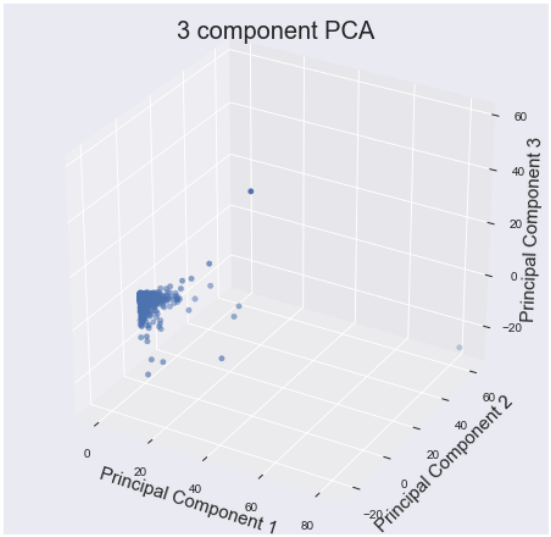
Matrice de corrélation de Spearman

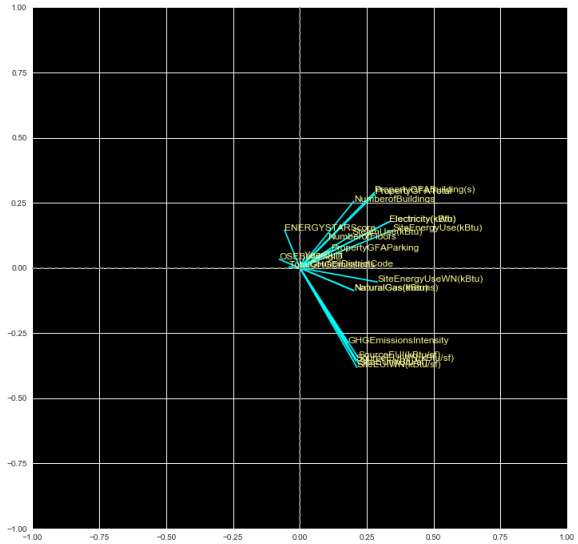


Dimensionality reduction avec la PCA

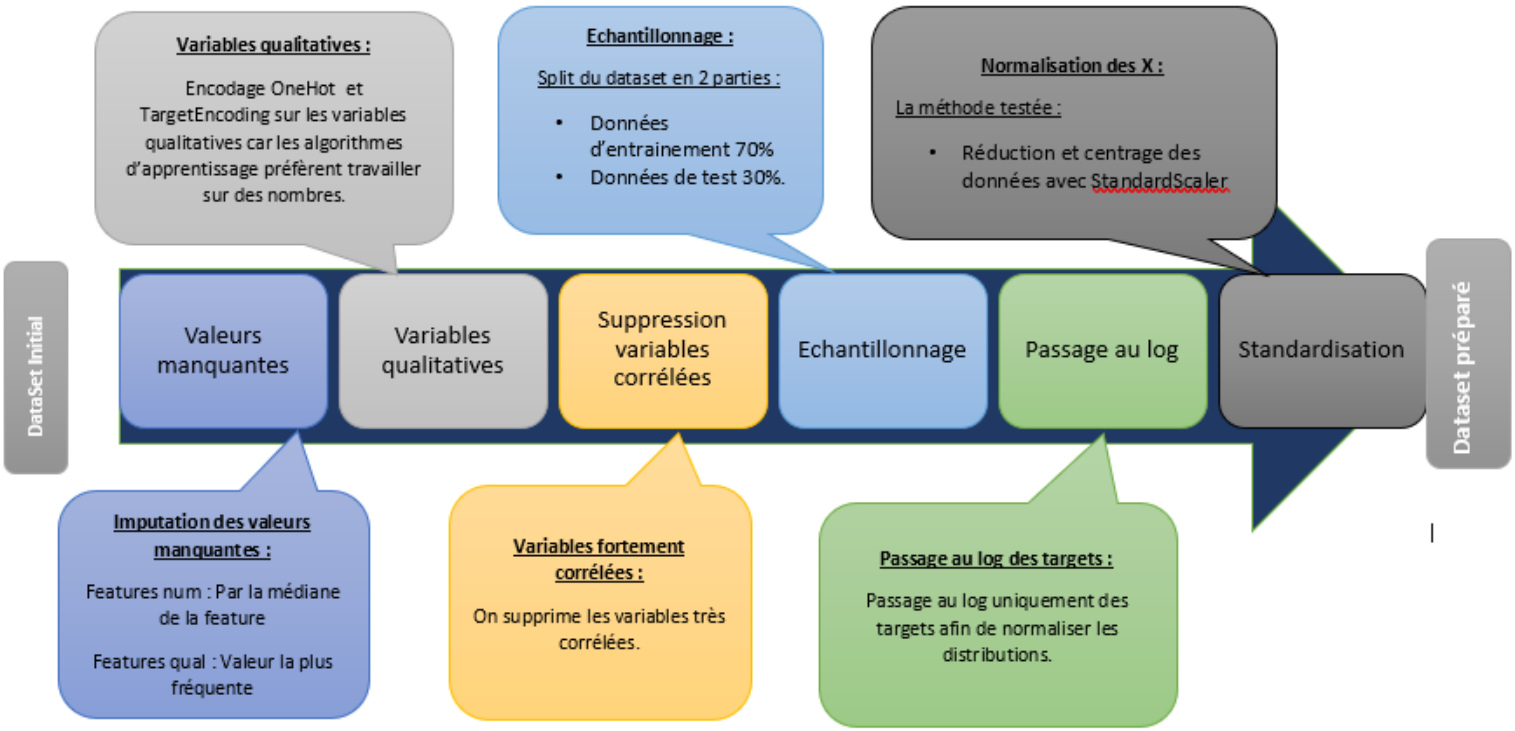
Analyse de la PCA:







Pre-Processing



Modélisation et Prédiction

les points importants pour bien cerner le problème :

- Les données d'entraînement sont étiquetés donc la tache est dite supervisée.
- Nous charchons une valeur donc c'est un problème de régression.
- Notre dataset contient plusieurs indicateurs donc c'est une régression multivariée.

GridSearchCV

GridSearchCV

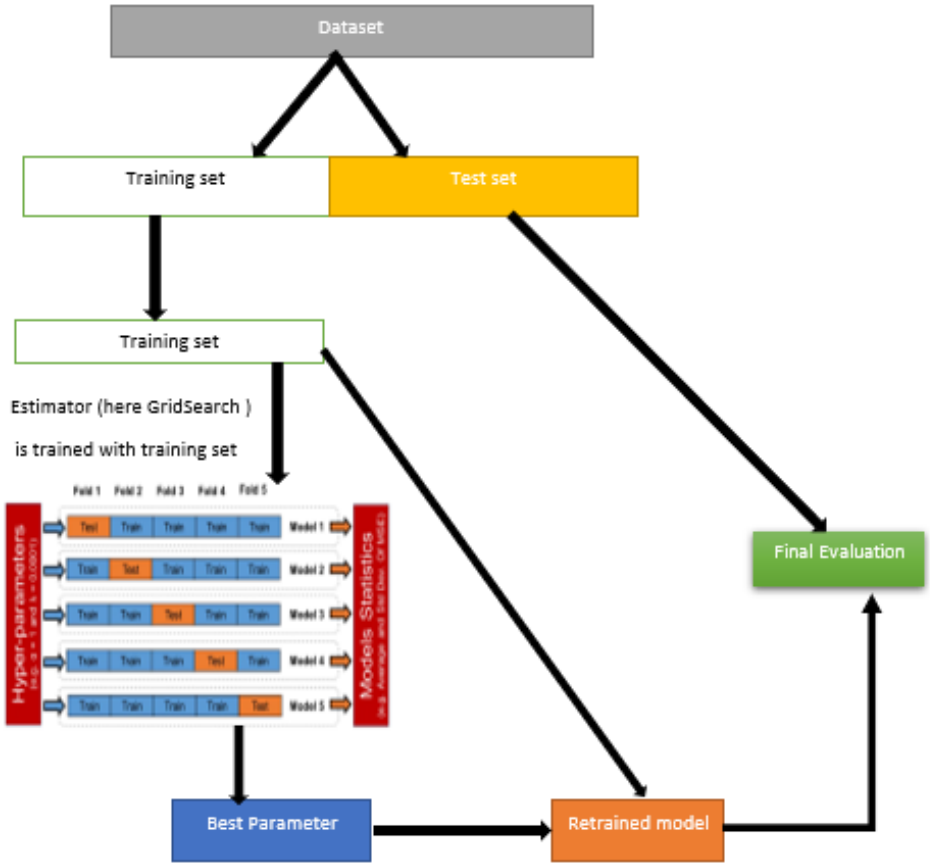
C'est une méthode d'optimisation (hyperparameter optimization) qui va nous permettre de tester une série de paramètres et de comparer les performances pour en déduire le meilleur paramétrage.

K_Fold

Elle consiste à découper le data set en k échantillons. On sélectionne x échantillons pour constituer l'échantillon d'apprentissage. Les $k - x$ échantillons restants permettront d'évaluer la performance du modèle.Pour construire le modèle suivant on sélectionne les échantillons différemment de manière à ne jamais avoir les mêmes échantillons d'apprentissage et de validation.

GridSearchCV

Procédure de la modélisation avec GridSearchCV:



Les modèles testées:

Linear Regression:

- sklearn.linear_model.Ridge
- sklearn.linear_model.Lasso
- sklearn.linear_model.ElasticNet

Support Vector Machines:

- sklearn.svm.LinearSVR
- sklearn.svm.SVR

Nearest Neighbours:

- sklearn.neighbors.KNearestNeighborsRegressor

Tree Based:

- sklearn.ensemble.RandomForestRegressor
- sklearn.ensemble.GradientBoostingRegressor
- xgboost.XGBRegressor

Neural Network

- sklearn.neural_network.MLPRegressor

Les Hyperparamètres d'un estimateur

Les hyper-paramètres sont des paramètres qui ne sont pas directement appris dans les estimateurs. Dans scikit-learn, ils sont passés comme arguments au constructeur des classes d'estimateur. Une recherche comprend:

- an estimator (regression ou classification comme sklearn.svm.SVC())
- parameter space (comme kernel, degree, gamma pour le SVM)
- a method for searching or sampling candidates
- a cross_validation
- a score function (comme score pour SVM)

LASSO

Le Lasso est un modèle linéaire qui estime les sparse coefficients . Il est utile dans certains contextes en raison de sa tendance à préférer des solutions avec moins de coefficients non nuls, ce qui réduit efficacement le nombre de caractéristiques dont dépend la solution donnée. - Ce modèle peut être très sensible aux valeurs aberrantes. L'hyper parametre utilisé:

- alpha: Constante qui multiplie le terme L1. alpha $= 0$ équivaut à un moindres carrés ordinaires, résolu par l'objet LinearRegression.

Gradient Boosting

- C'est une méthode séquentielle.
- Utilisation de plusieurs modèles
- Pondération des individu (donner un poids plus important aux individus pour lesquels la valeur a été mal prédite pour la construction du modèle suivant)

Cet algorithme utilise le gradient de la fonction de perte pour le calcul des poids des individus lors de la construction de chaque nouveau modèle.
Pour eviter le sur-apprentissage il faut :

- Limiter la taille des arbres
- Construire les modèles sur des échantillons de la population (on parle de stochastic gradient boosting)

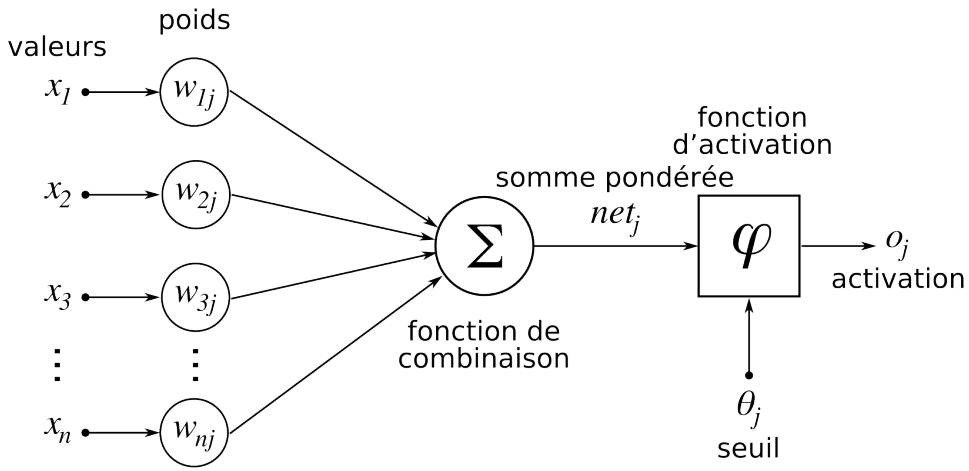
Gradient Boosting

Les hyper paramètres de Gradient Boosting:

- n_estimators : Nbr d'étape de boosting a effectuer
- max_depth : profondeur maximale des estimateurs de régression individuels
- min_samples_split : Le nombre minimum d'échantillons requis pour scinder un nœud interne
- learning_rate : regulation de la contribution de chaque arbre
- loss : fonction de perte

Réseau de neurones — Multi-layer Perceptron regressor

C'est une méthode qu'on peut utiliser pour des problématiques de prédiction et de classement en particulier pour des phénomènes complexes à modéliser et/ou non linéaires.



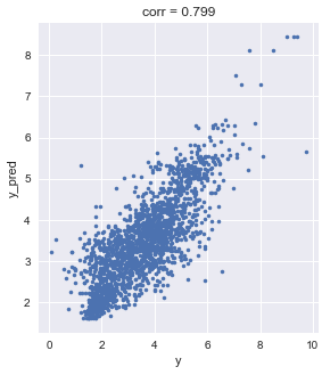
Les Hyperparamètres de cette méthodes sont:

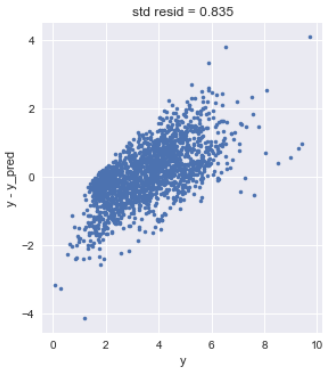
- alpha : paramètre de régulation
- Nbr de neurones dans chaque couche
- learning_rate : taux d'apprentissage pour les mises à jour de poids
- solver : la méthode d'optimisation du poids

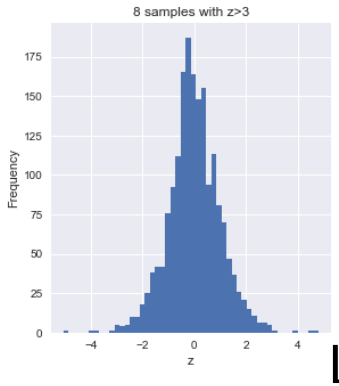
Synthèse des résultats

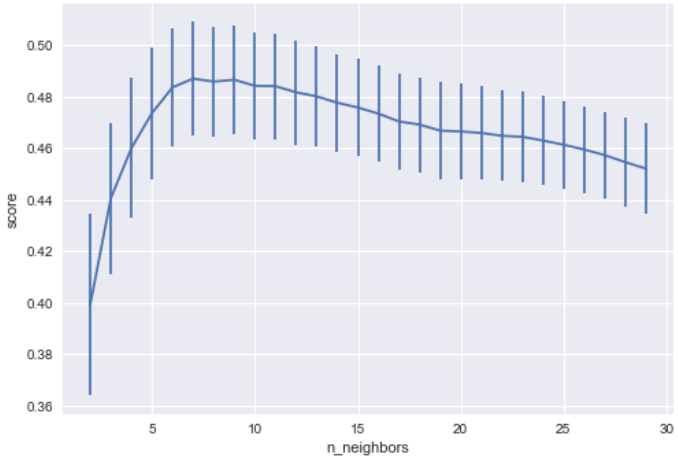
Prédiction de TotalGHGEmissions

Exemple de GridSearchCV pour le KNNs:



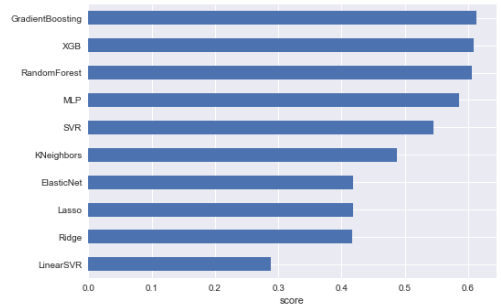




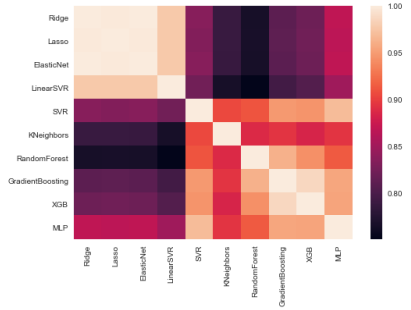


Le Score des méthodes utilisées:

	mean	std
LinearSVR	0.288665	0.105953
Ridge	0.416877	0.042132
Lasso	0.417721	0.038217
ElasticNet	0.417984	0.038091
KNeighbors	0.486929	0.049130
SVR	0.544471	0.026076
MLP	0.584738	0.034496
RandomForest	0.606227	0.030072
XGB	0.608790	0.036820
GradientBoosting	0.613326	0.031853



La matrice de corrélations des méthodes optimisées:

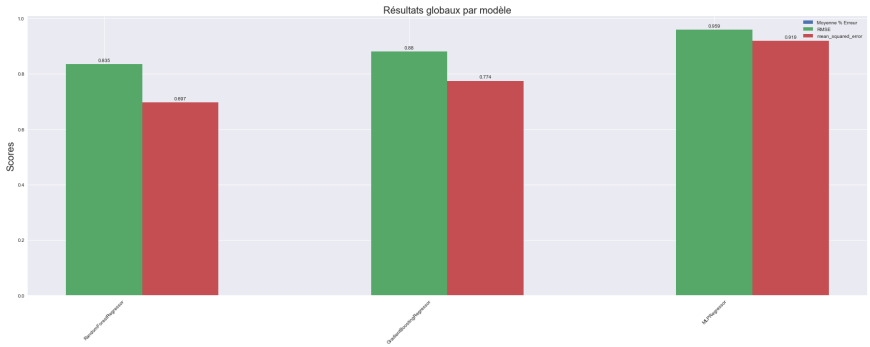


Synthèse des résultats

Prédiction de TotalGHGEmissions

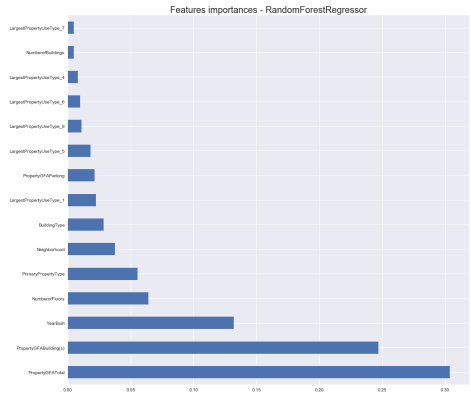
Prédiction sur x_test:

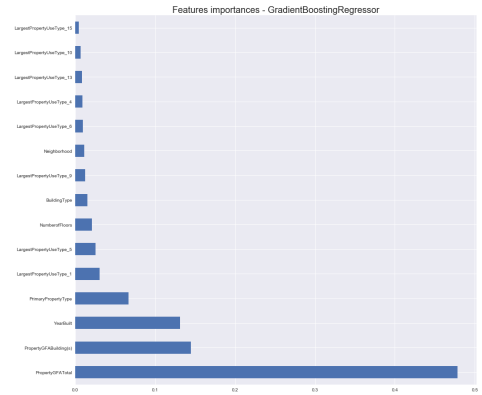
	Modeles	RMSE	mean_absolute_percentage_error	mean_squared_error	R2	Time
0	RandomForestRegressor	0.835	inf	0.697	0.653	0.878643
1	GradientBoostingRegressor	0.880	inf	0.774	0.615	3.104120
2	MLPRegressor	0.959	inf	0.919	0.543	0.177495



Feature importance:

```
Features importances - RandomForestRegressor
PropertyGFATotal            0.303868
PropertyGFABuilding(s)      0.247106
YearBuilt                   0.131906
NumberofFloors              0.063896
PrimaryPropertyType         0.055581
Neighborhood                0.037343
BuildingType                0.028438
LargestPropertyUseType_1    0.022341
PropertyGFAParking          0.021282
LargestPropertyUseType_5    0.018040
dtype: float64

Features importances - GradientBoostingRegressor
PropertyGFATotal            0.478159
PropertyGFABuilding(s)      0.144565
YearBuilt                   0.131287
PrimaryPropertyType         0.066758
LargestPropertyUseType_1    0.030758
LargestPropertyUseType_5    0.025215
NumberofFloors              0.021118
BuildingType                0.015299
LargestPropertyUseType_9    0.012304
Neighborhood                0.010977
dtype: float64
```

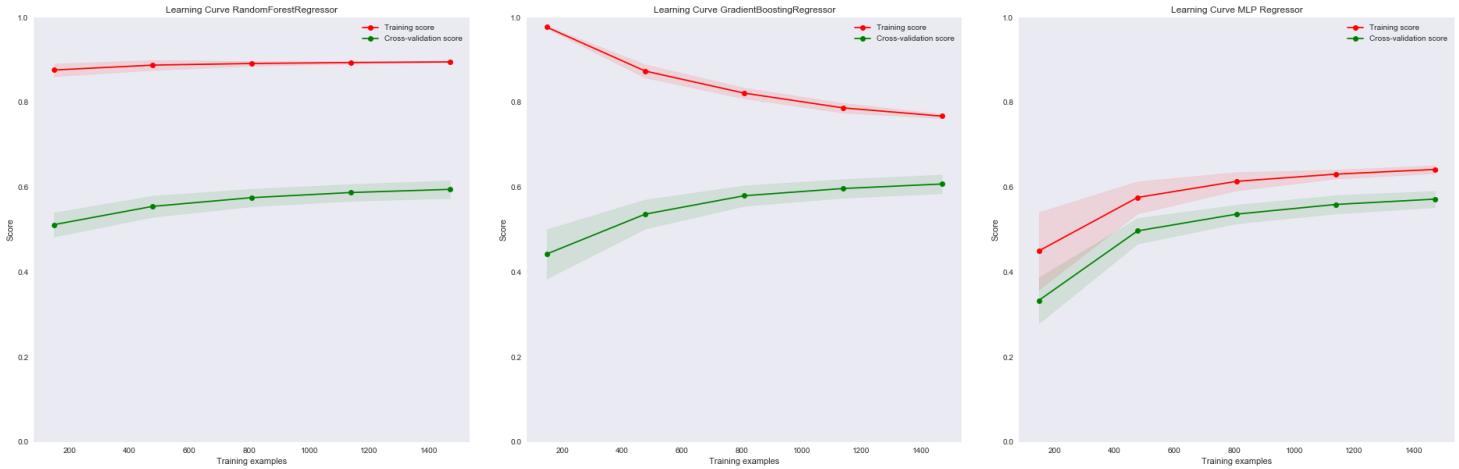




Synthèse des résultats

Prédiction de TotalGHGEmissions

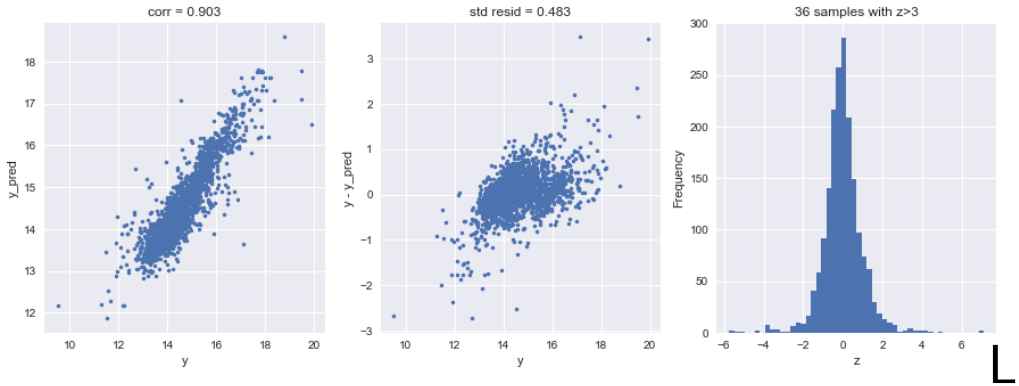
Les courbes d'apprentissage nous montre les performances des modèles par rapport à la taille des échantillons.
Les courbes ont été tracées avec les modèles optimisés par la GridSearchCV
Analyse : Les 3 modèles ont encore une marge de progression potentielle
Le Gradient Boosting Regressor est celui avec la courbe la plus intéressante
Il faudrait plus de données pour avoir de meilleures performances

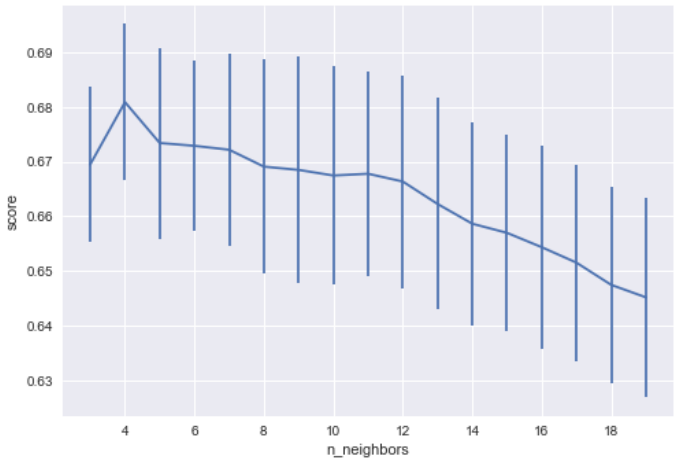


Synthèse des résultats

Prédiction de SiteEnergyUse(kBtu)

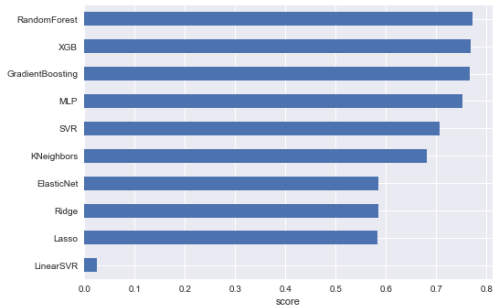
Exemple de GridSearchCV pour le KNN:



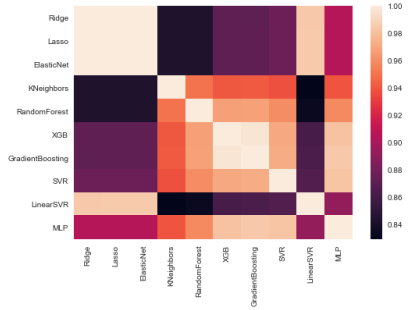


Le Score des méthodes utilisées:

	mean	std
LinearSVR	0.024235	0.484828
Lasso	0.583506	0.037287
Ridge	0.584439	0.040254
ElasticNet	0.585720	0.046894
KNeighbors	0.680854	0.031944
SVR	0.707183	0.042851
MLP	0.751853	0.045811
GradientBoosting	0.766564	0.038413
XGB	0.768077	0.028578
RandomForest	0.772983	0.031269



La matrice de corrélations des méthodes optimisées:

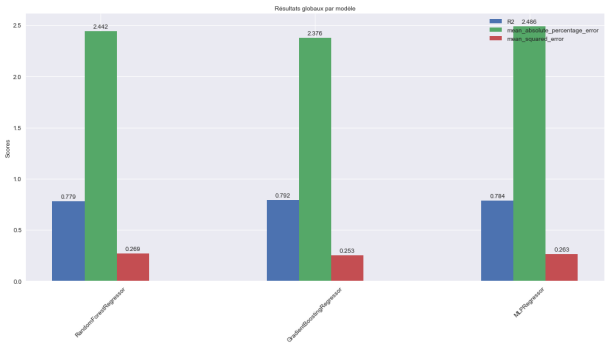


Synthèse des résultats

Prédiction de SiteEnergyUse(kBtu)

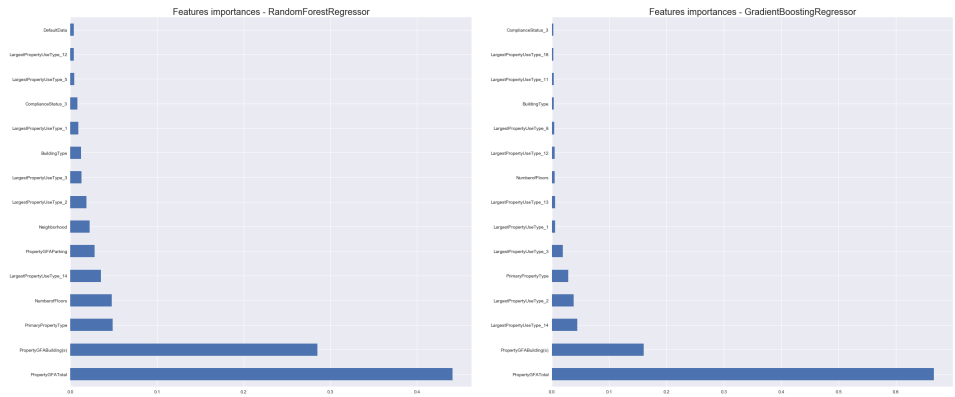
Prédiction sur x_test:

	Modeles	RMSE	mean_absolute_percentage_error	mean_squared_error	R2	Time
0	RandomForestRegressor	0.519	2.442	0.269	0.779	2.559154
1	GradientBoostingRegressor	0.503	2.376	0.253	0.792	0.317150
2	MLPRegressor	0.513	2.486	0.263	0.784	0.613435



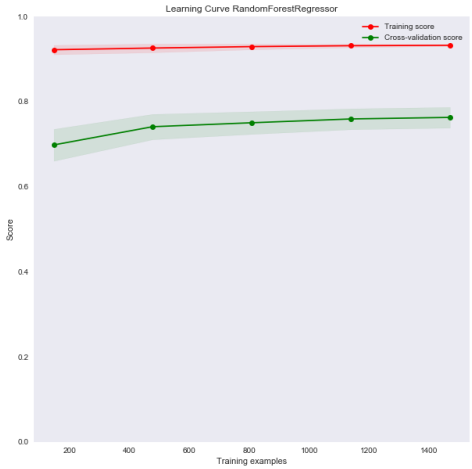
Feature importance:

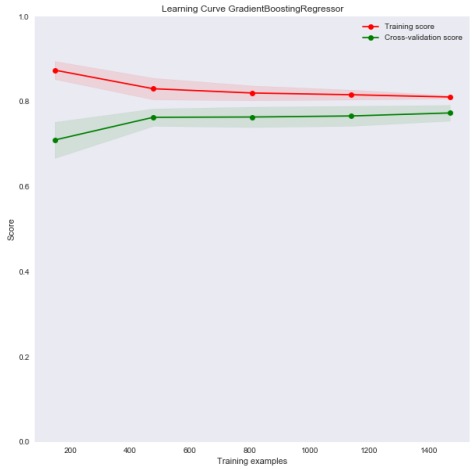
```
Features importances - RandomForestRegressor
PropertyGFATotal              0.441047
PropertyGFABuilding(s)        0.285097
PrimaryPropertyType           0.048938
NumberofFloors                0.047500
LargestPropertyUseType_14     0.035315
PropertyGFAParking            0.027857
Neighborhood                  0.022112
LargestPropertyUseType_2      0.018584
LargestPropertyUseType_3      0.012666
BuildingType                  0.012286
dtype: float64

Features importances - GradientBoostingRegressor
PropertyGFATotal              0.666338
PropertyGFABuilding(s)        0.159987
LargestPropertyUseType_14     0.044150
LargestPropertyUseType_2      0.038380
PrimaryPropertyType           0.028878
LargestPropertyUseType_3      0.019357
LargestPropertyUseType_1      0.006031
LargestPropertyUseType_13     0.005429
NumberofFloors                0.004906
LargestPropertyUseType_12     0.004895
dtype: float64
```

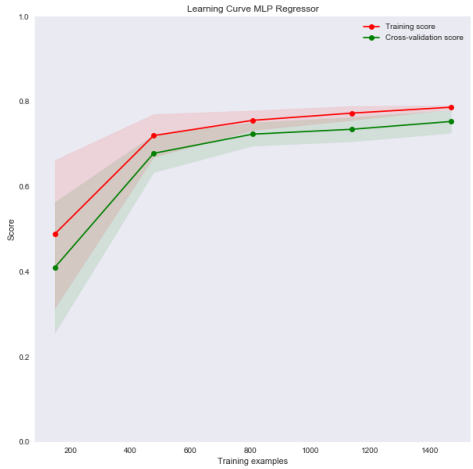


Synthèse des résultats

Prédiction de TotalGHGEmissions







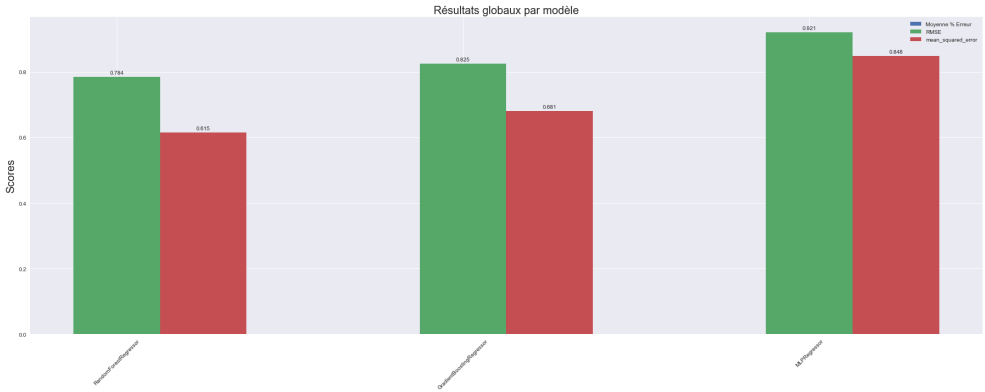
Prédiction avec L'ENERGY STAR

L'ENERGY STAR Score est un outil de dépistage aidant à évaluer les performances d'émission de GES d'une propriété par rapport à des bâtiments similaires. Cet indicateur se base sur une échelle de 0 à 100 dont la médiane est 50. Si le score est $>= 75$, le bâtiment peut être admissible à la certification ENERGY STAR.

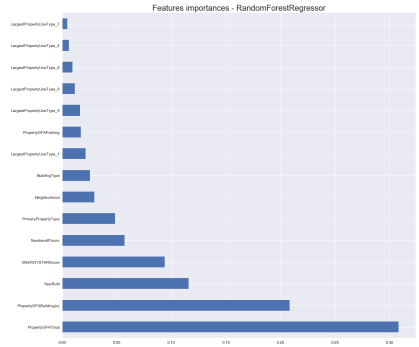
Prédiction avec L'ENERGY STAR

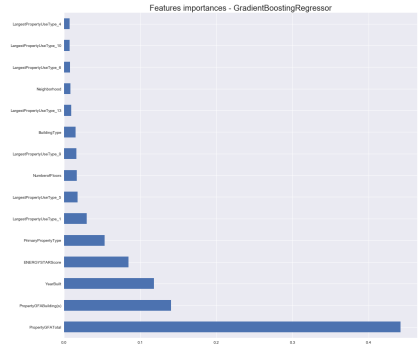
Prédiction de TotalGHGEmissions :

	Modeles	RMSE	mean_absolute_percentage_error	mean_squared_error	R2	Time
0	RandomForestRegressor	0.784	inf	0.615	0.694	0.900925
1	GradientBoostingRegressor	0.825	inf	0.681	0.662	2.940839
2	MLPRegressor	0.921	inf	0.848	0.578	0.190463



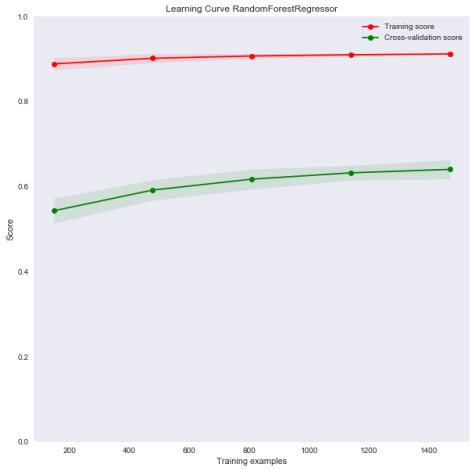

```
Features importances - RandomForestRegressor
PropertyGFATotal            0.308452
PropertyGFABuilding(s)      0.208521
YearBuilt                   0.115862
ENERGYSTARScore             0.093876
NumberofFloors              0.057018
PrimaryPropertyType         0.048162
Neighborhood                0.029070
BuildingType                0.025315
LargestPropertyUseType_1    0.021346
PropertyGFAParking          0.016923
dtype: float64

Features importances - GradientBoostingRegressor
PropertyGFATotal            0.441809
PropertyGFABuilding(s)      0.140332
YearBuilt                   0.117973
ENERGYSTARScore             0.084799
PrimaryPropertyType         0.053472
LargestPropertyUseType_1    0.029963
LargestPropertyUseType_5    0.017750
NumberofFloors              0.016500
LargestPropertyUseType_9    0.015938
BuildingType                0.015260
dtype: float64
```

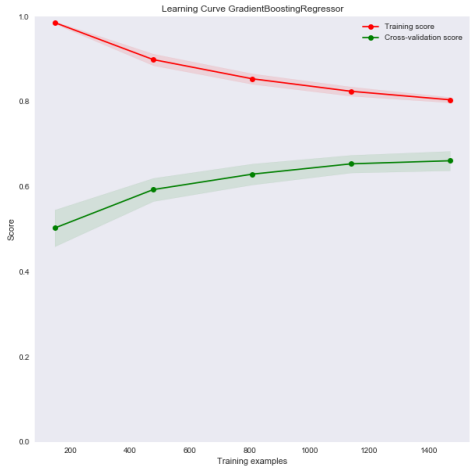


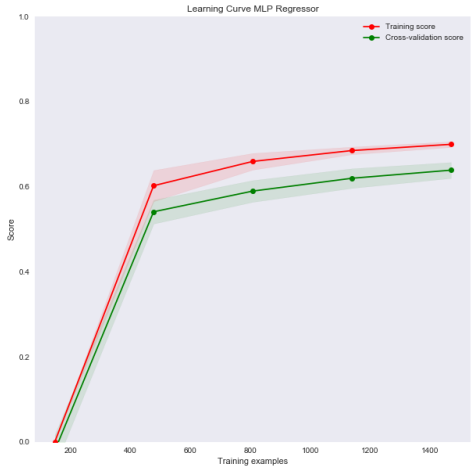


Synthèse des résultats

Prédiction de TotalGHGEmissions



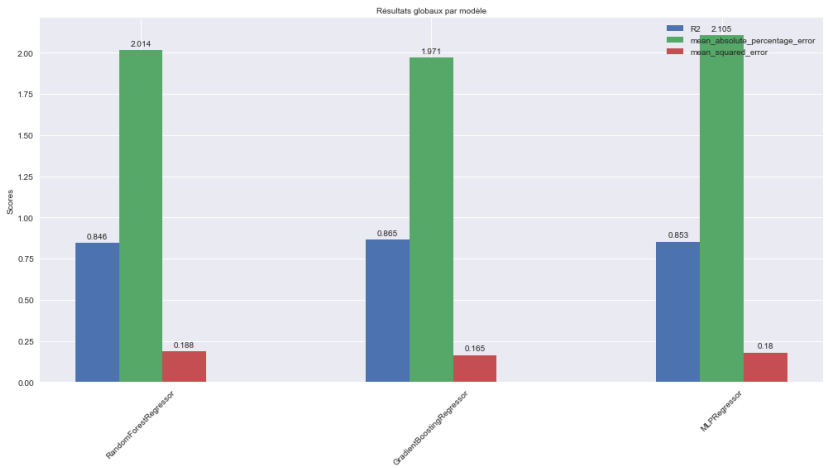




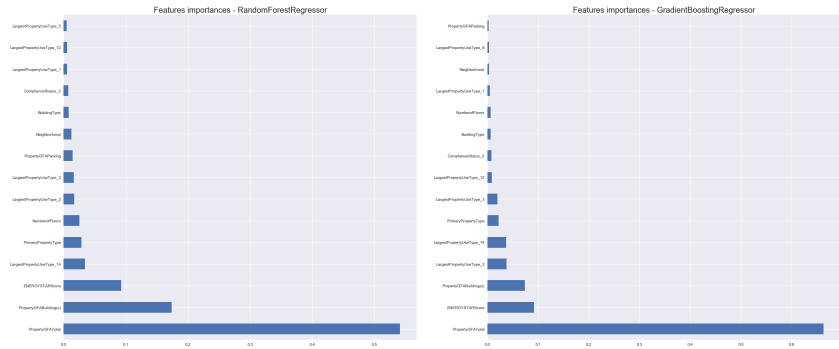
Prédiction avec L'ENERGY STAR

Prédiction de SiteEnergyUse(kBtu) :

	Modeles	RMSE	mean_absolute_percentage_error	mean_squared_error	R2	Time
0	RandomForestRegressor	0.434	2.014	0.188	0.846	3.552007
1	GradientBoostingRegressor	0.406	1.971	0.165	0.865	0.400039
2	MLPRegressor	0.424	2.105	0.180	0.853	0.933697

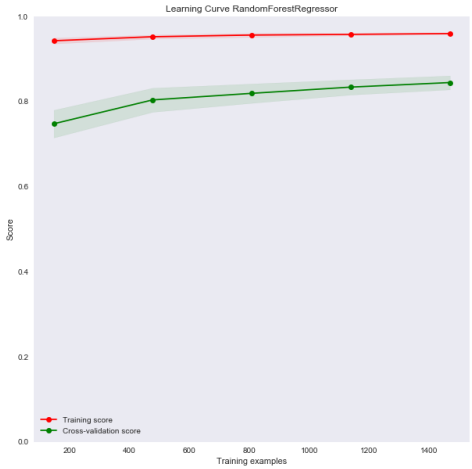


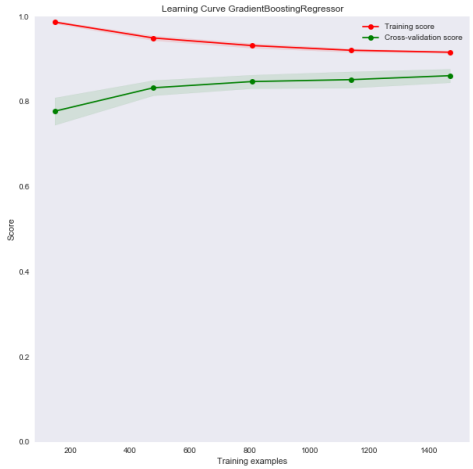

```
Features importances - RandomForestRegressor
PropertyGFATotal              0.541151
PropertyGFABuilding(s)        0.174106
ENERGYSTARScore               0.092670
LargestPropertyUseType_14     0.034525
PrimaryPropertyType           0.028663
NumberofFloors                0.025249
LargestPropertyUseType_2      0.017055
LargestPropertyUseType_3      0.016266
PropertyGFAParking            0.014116
Neighborhood                  0.012552
dtype: float64

Features importances - GradientBoostingRegressor
PropertyGFATotal              0.663014
ENERGYSTARScore               0.092153
PropertyGFABuilding(s)        0.073926
LargestPropertyUseType_2      0.037671
LargestPropertyUseType_14     0.037175
PrimaryPropertyType           0.021929
LargestPropertyUseType_3      0.019436
LargestPropertyUseType_12     0.009040
ComplianceStatus_3            0.007698
BuildingType                  0.006758
dtype: float64
```

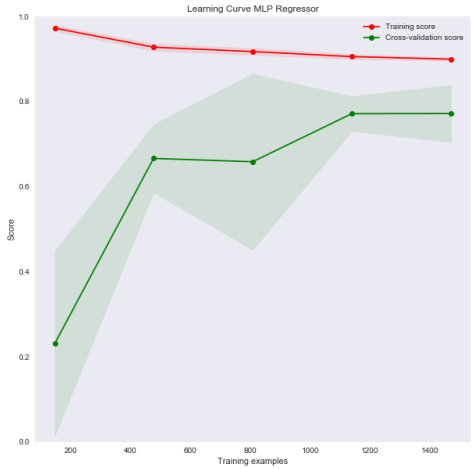


Synthèse des résultats

Prédiction de SiteEnergyUse(kBtu)







Conclusion

Les résultats sont globalement décevants est Cela dû à la taille de la dataset.
Prédictions AVEC la feature ENERGY STAR Score légèrement meilleures que les prédictions SANS cette feature La feature ne représente que peu d'intérêt

Merci pour votre attention