Projet 3

Anticiper les besoins Energétiques de Seattle

Problématique

Prédire consommation et émission

Seattle

2050

Problématique

Réduire
les coûts

relevé

modélisation

Problématique

Les données

Dataframe

47 Variables

Immeubles

3340 + 3376 = 6716

- Localisation
- Construction
- Energétique

Nettoyage

Sélection des Features

Energétique

Consommation
Emission CO^2

Construction

- Caractéristique bâtiment
- Type d'usage
- Commentaires
- Conformité données

Localisation

- District
- Quartier
- Lon/Lat
- Zip Code
- Adresse

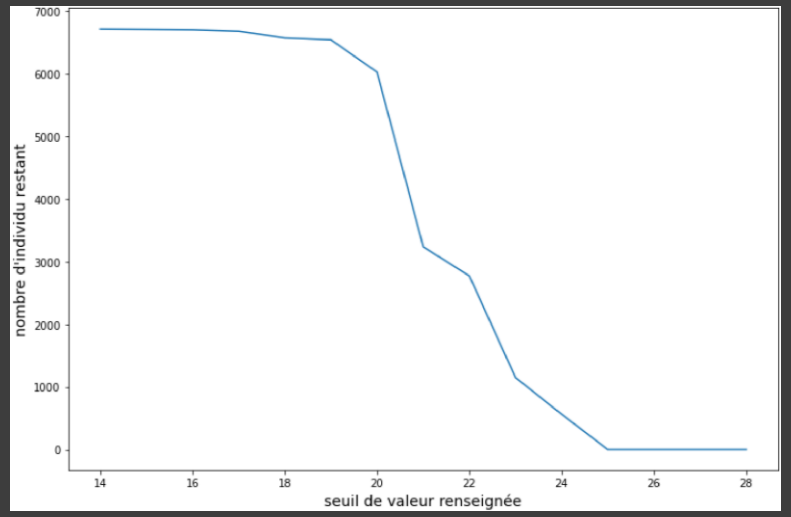
Labels

Features

Nettoyage

Filtrage des individus

filtrage des individus moins renseigné



Retire 172 individus

Suppression des doublons

Si
- Longitude
- Latitude
- TaxParcelIdentificationNumber
- DataYear

Son identique → Doublon

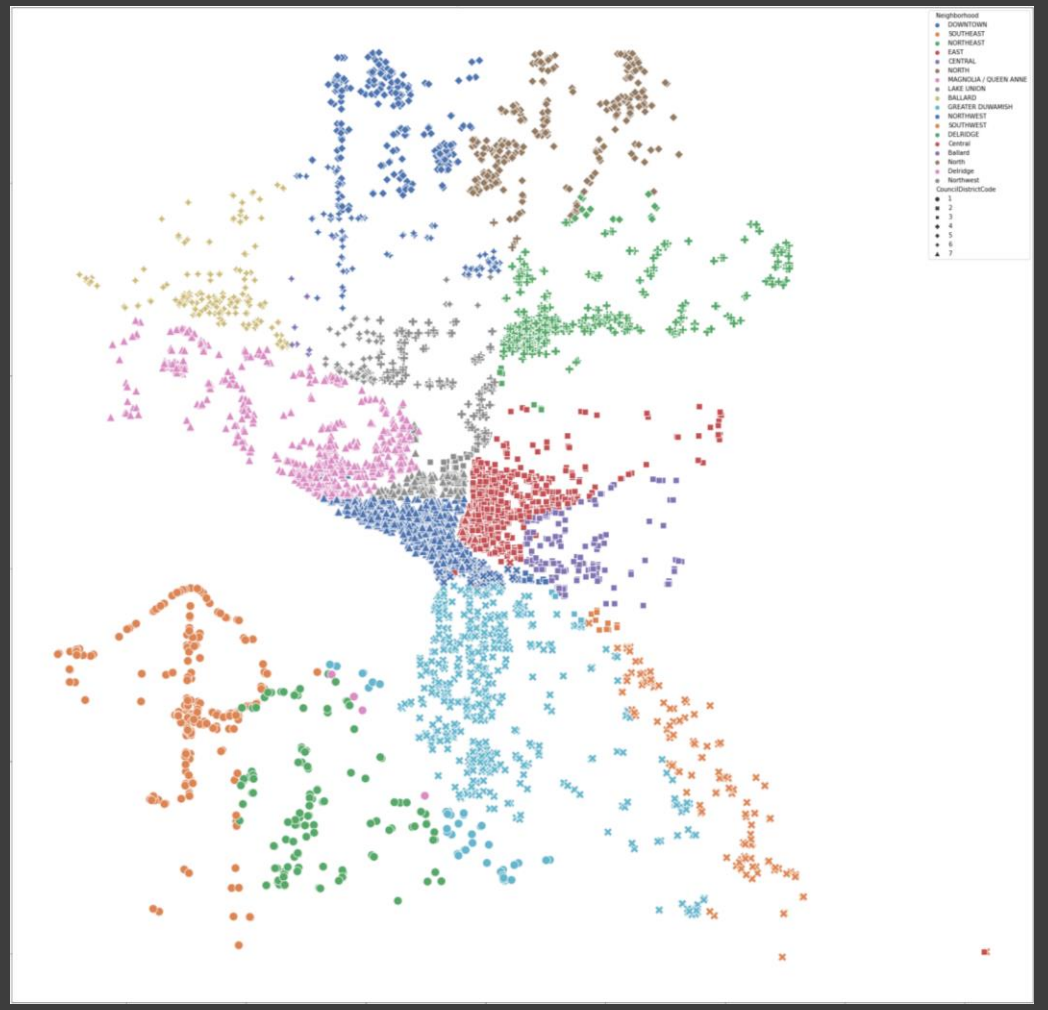
Retire 127 individus

Nettoyage

Variable de localisation

Seattle

Ses districts et quartiers



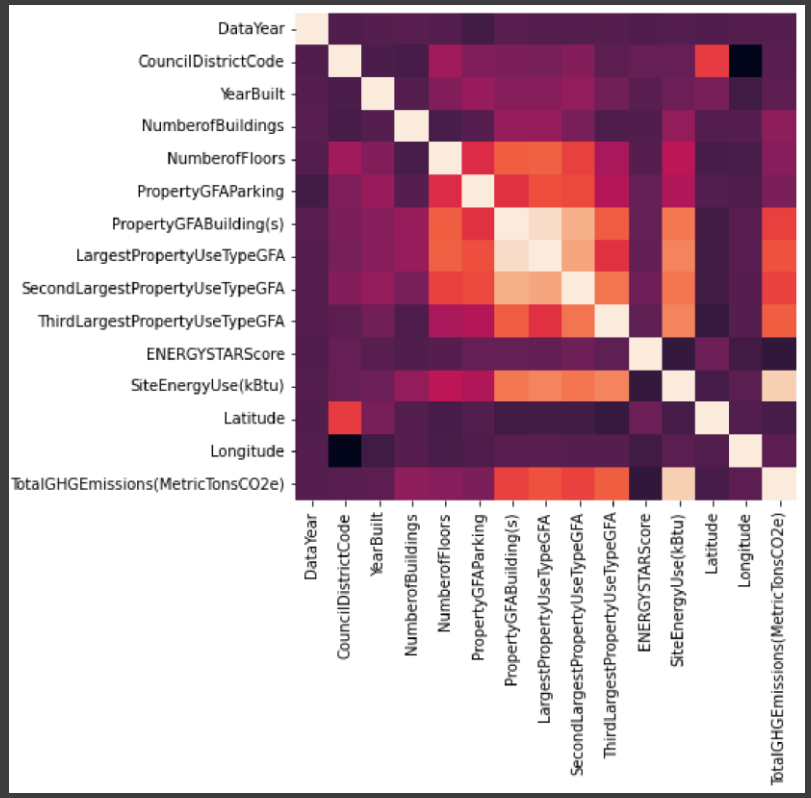
ZipCode : 83 différents

Adresse : non exploitable

Nettoyage

Corrélations

Corrélations



Nettoyage

Peu de valeurs manquantes
distribution concentrée
distribution asymétrique

Combler par la médiane

Nettoyage

Nouvelles variables

Avant

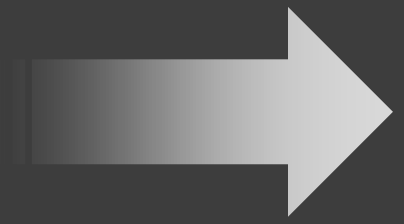
LargestPropertyUseType	LargestPropertyUseTypeGFA
Hotel	88434
Office	51029
Hotel	348329
Other	385274
Office	103501

Après

Hotel	Office	Other
88434	0	0
0	51029	0
348329	0	0
0	0	385274
0	103501	0

Nettoyage

- Liste de tout les types d’usage
- Type de bâtiment
- Quartier
- District



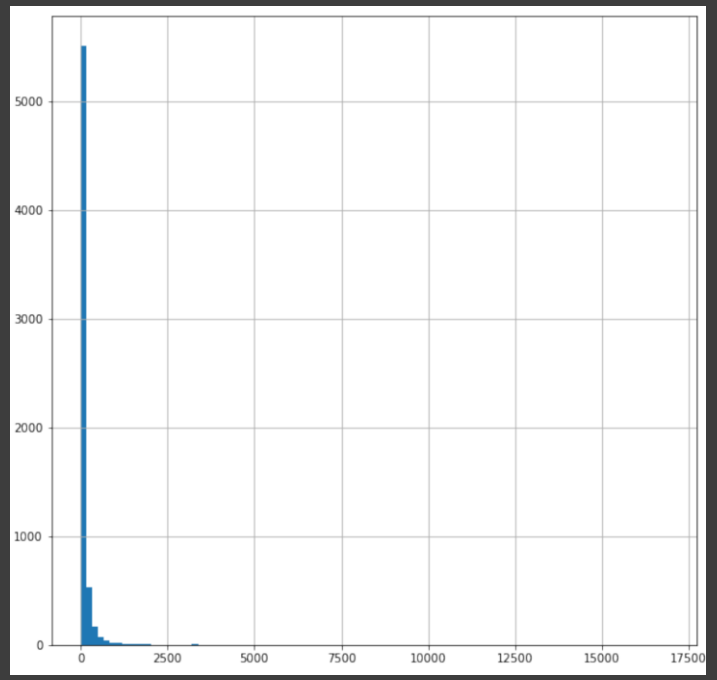
One Hot Encoding

Nettoyage

Nouvelles variables

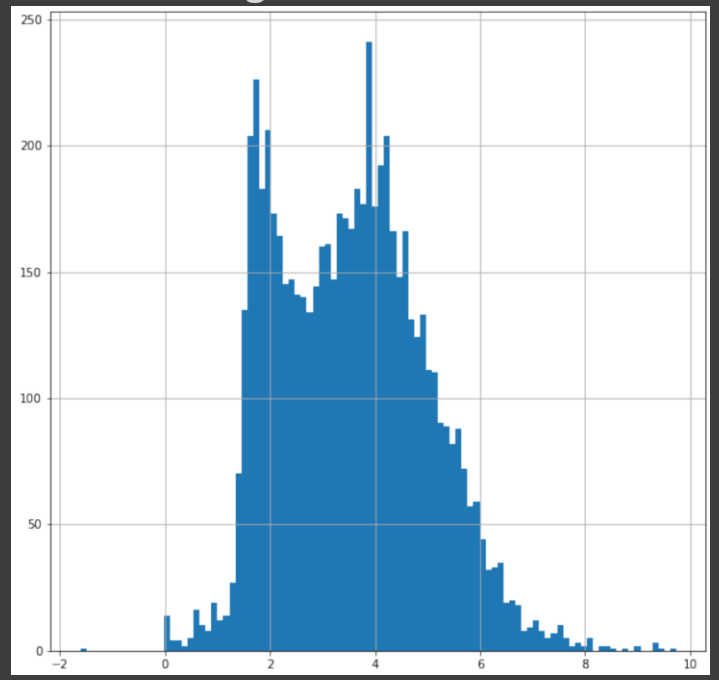
Avant

Consommation
Emission co²



Après

Log Consommation
Log Emission co²



Modélisation

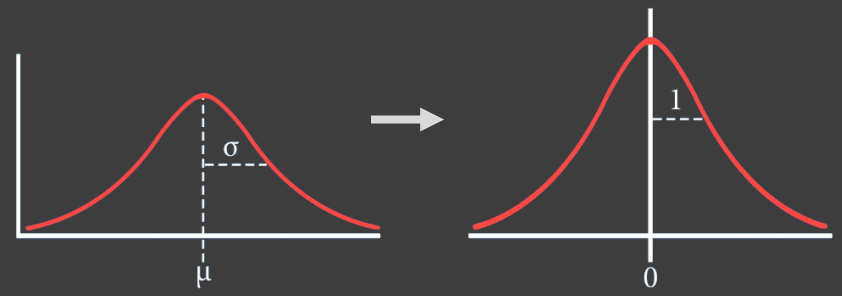
Train Set

80%

Test Set

20%

Standardise



Modélisation

Modèles Linéaire

Baseline

Régression linéaire

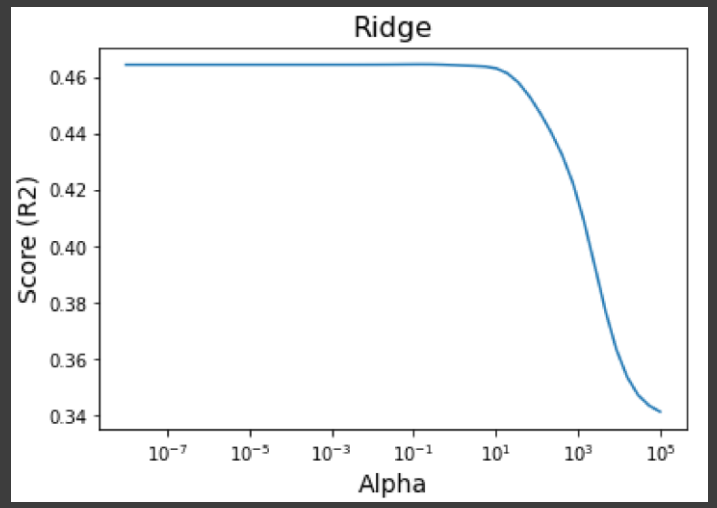
Score (R^2) : 0.481

Learning time : xxx

Régression Ridge

Score (R^2) : 0.481

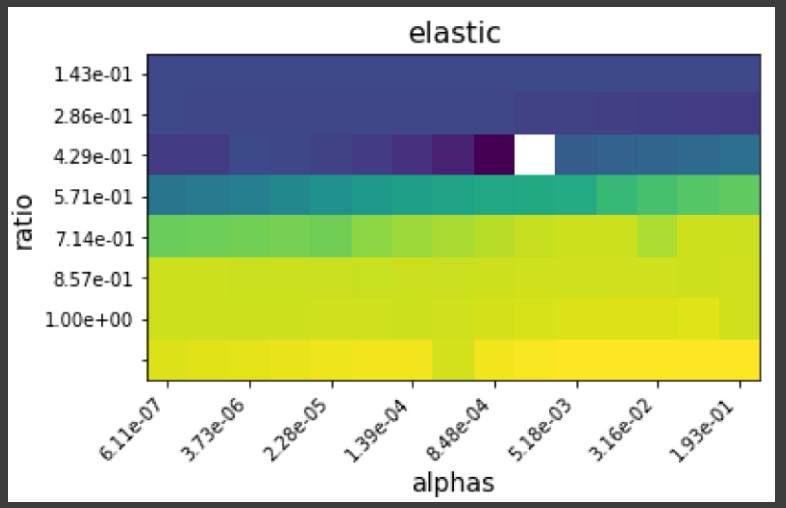
Learning time : 0.02



Régression Elastic Net

Score (R^2) : 0.481

Learning time : 0.54

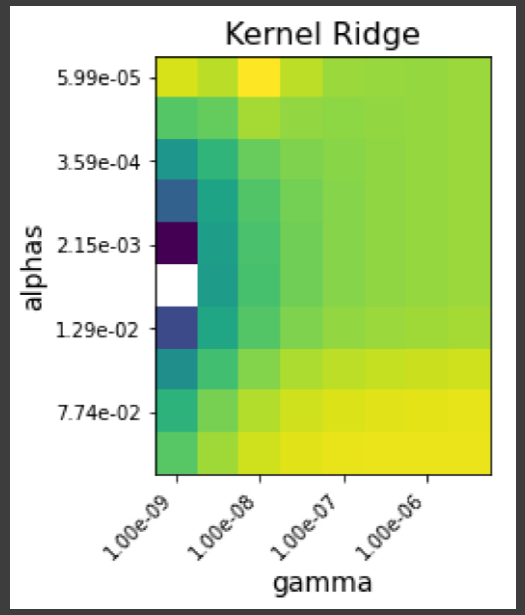


Modélisation

Modèles à Noyau

Kernel Ridge

Score (R^2) : 0.447
Learning time : 0.89

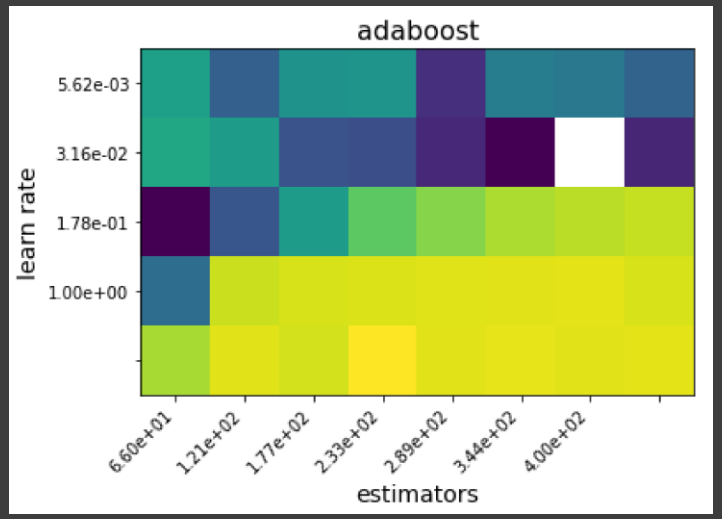


Modélisation

Modèles Ensembliste

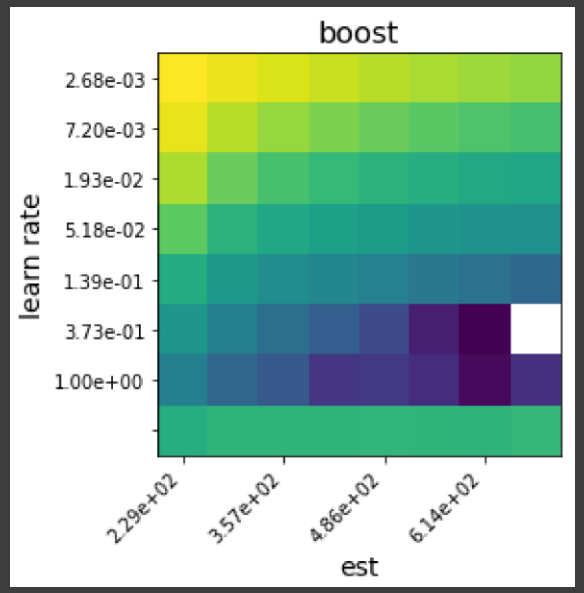
Ada Boost

Score (R^2) : 0.489
Learning time : 7.93



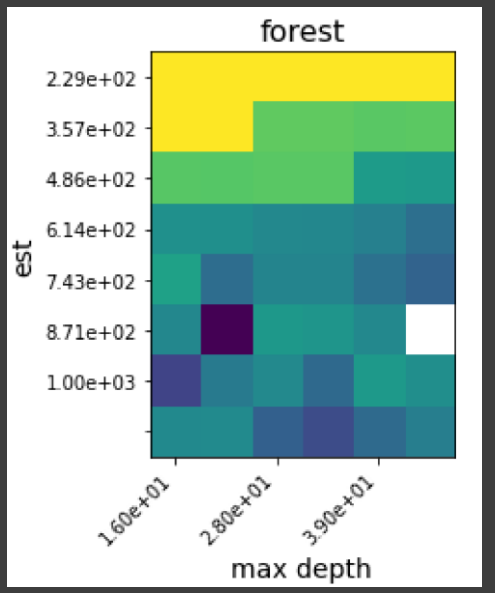
Gradiant Boosting

Score (R^2) : 0.754
Learning time : 16.2



Random Forest

Score (R^2) : 0.801
Learning time : 26.8



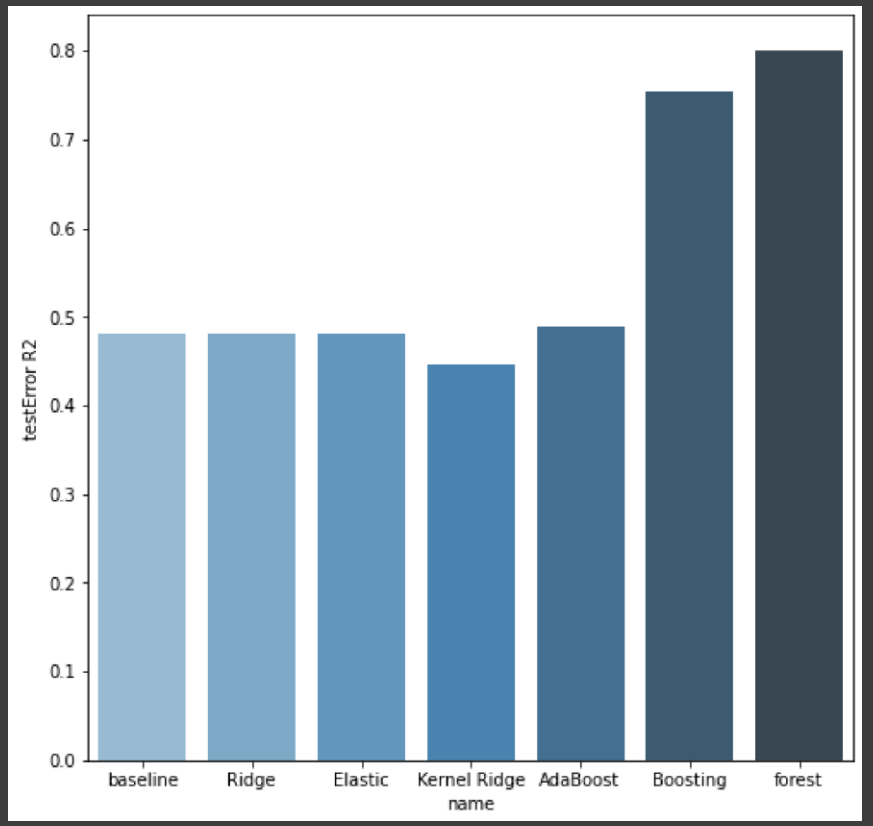
Modélisation

Résumé

Résumé des modèles

name	trainError	testError R2	testError RMSE	learningTime	predictTime
baseline	0.524629	0.481074	1.026192	NaN	0.00509963
Ridge	0.524629	0.481074	1.026192	0.017230	0.00682005
Elastic	0.524629	0.481087	1.026180	0.542988	0.00499864
Kernel Ridge	0.623419	0.446675	1.059660	2.043311	0.890072
AdaBoost	0.491015	0.489334	1.017992	7.931864	0.517571
Boosting	0.902375	0.753696	0.706987	16.209746	0.0691218
forest	0.974181	0.800998	0.635484	26.763308	0.633284

Score par modèles



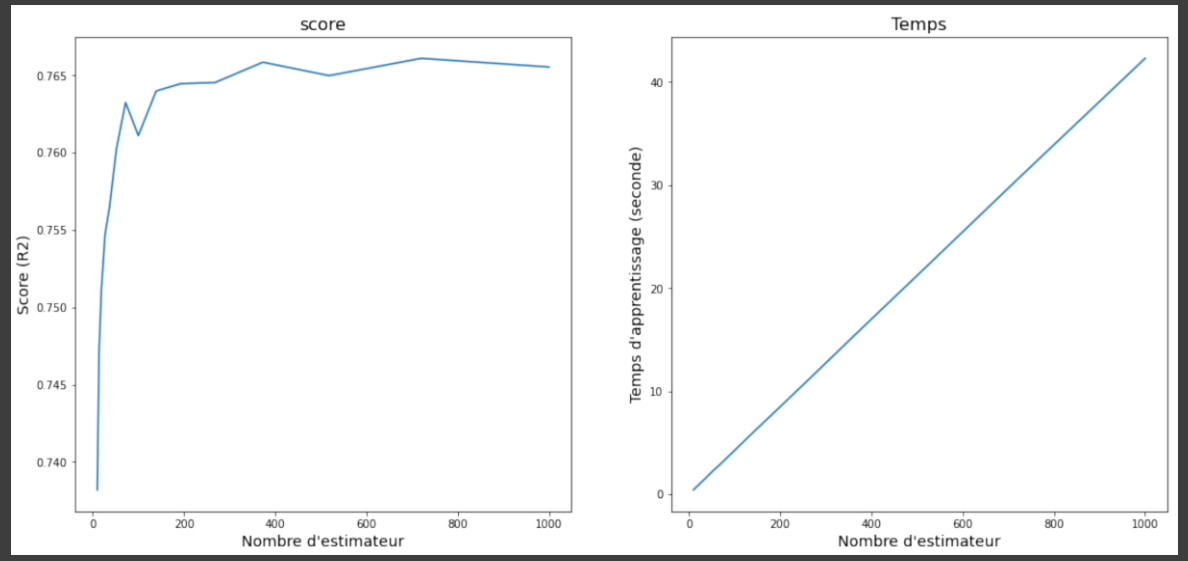
Modèle Final

Affinage des hyperparamètres

Sélection
Nombre d'estimateur

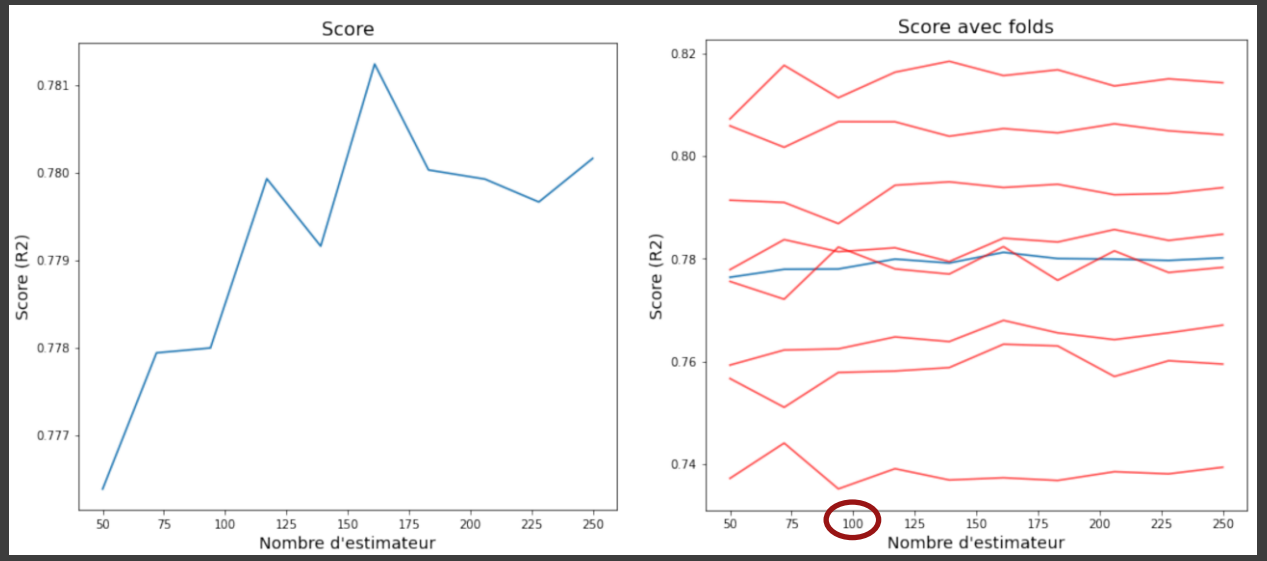
Intervalle large et peu précis

5 Folds



Intervalle étroit et plus précis

8 Folds

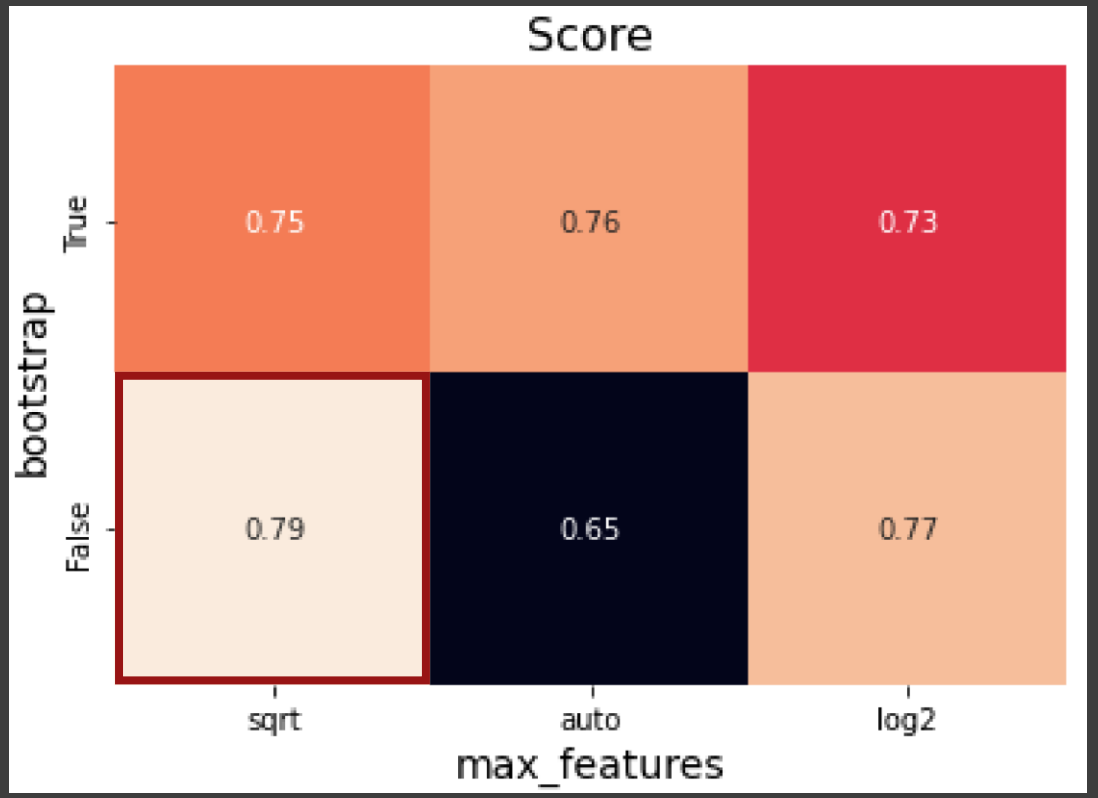


Modèle Final

Affinage des hyperparamètres

Sélection
bootstrap et max_features

5 Folds

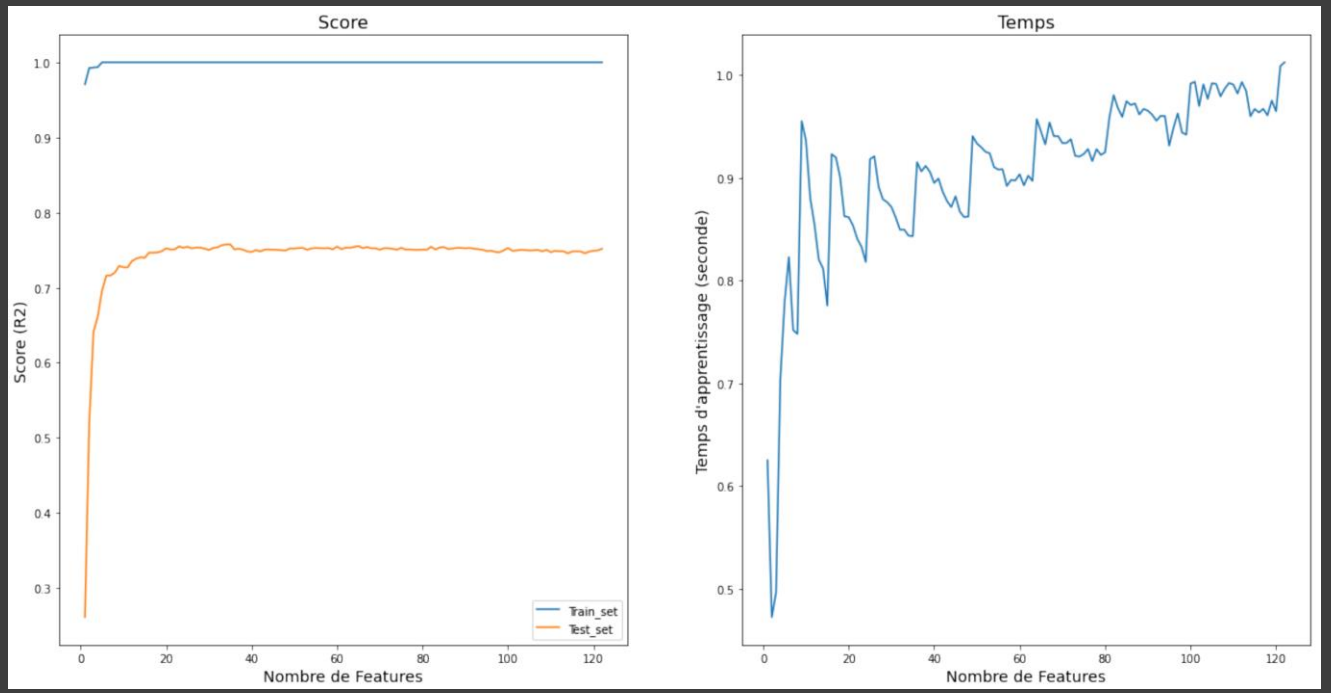


Modèle Final

Feature Selection

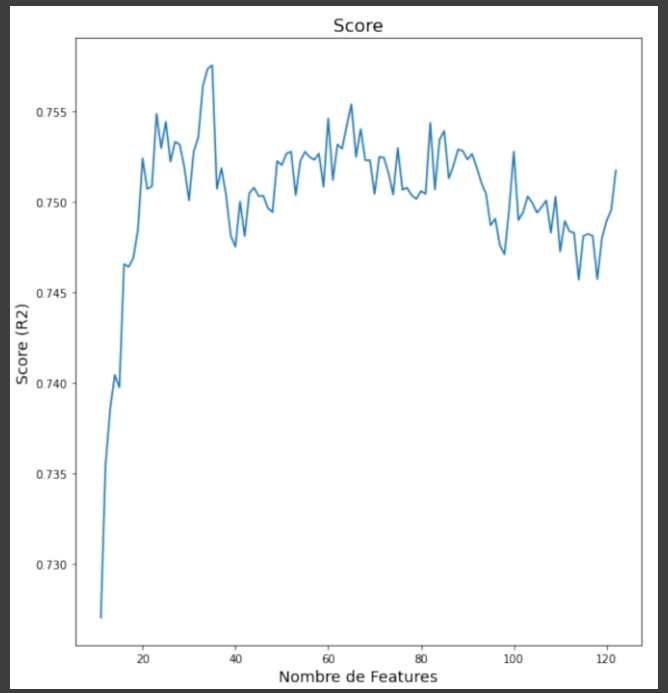
Intervalle large et peu précis

5 Folds



Intervalle de 10 à plus

5 Folds

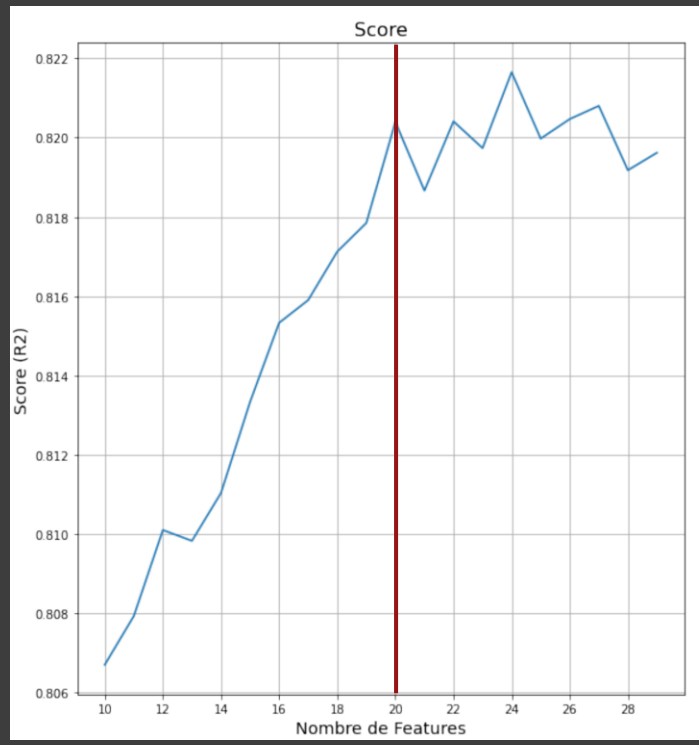


Modèle Final

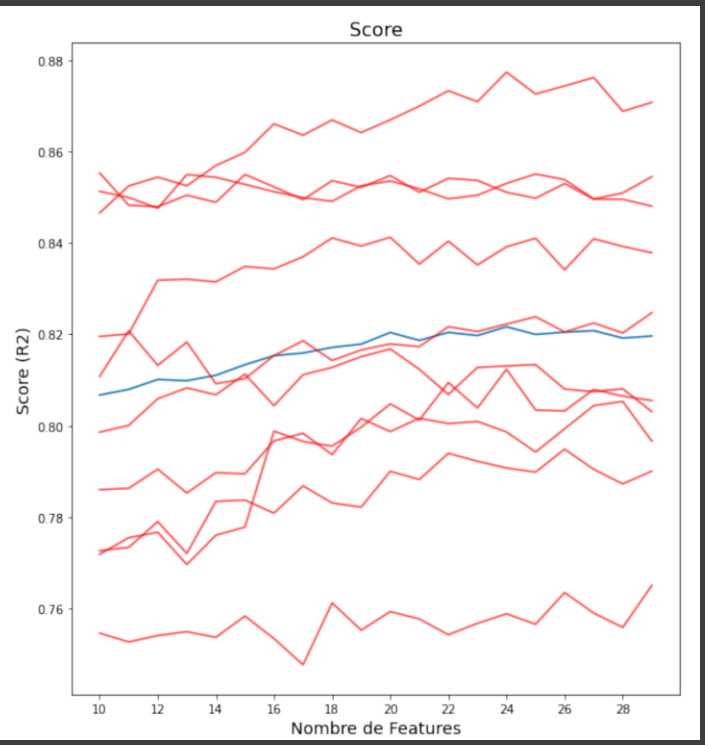
Feature Selection

Intervalle étroit et précis

10 Folds



Avec Folds



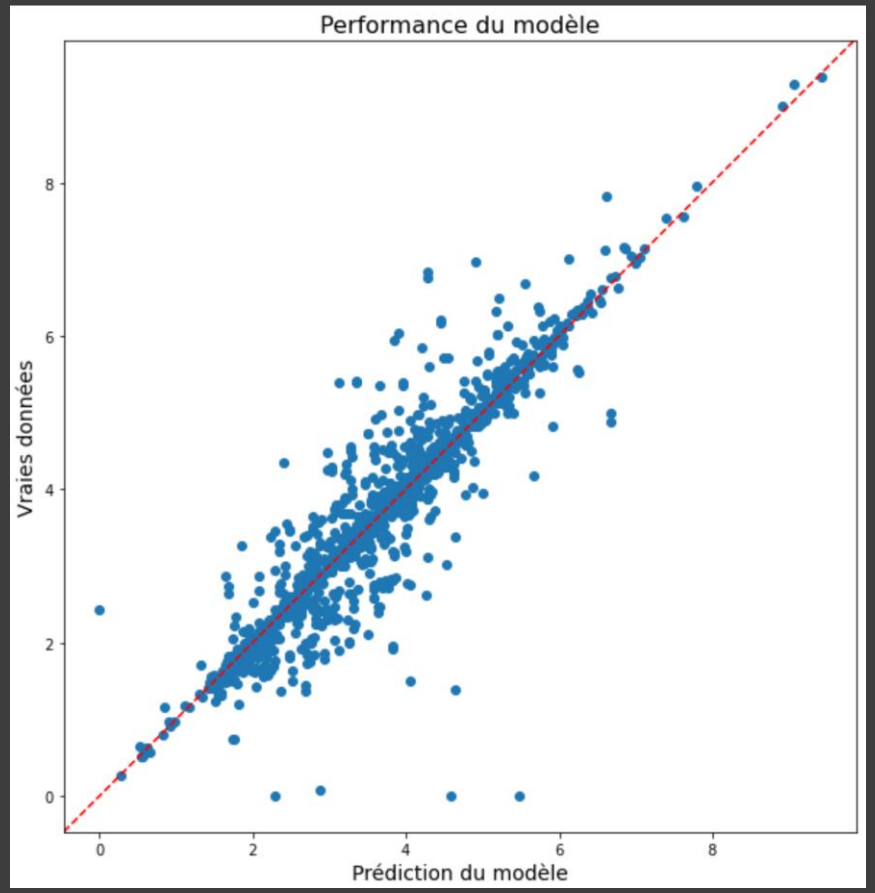
Modèle Final

Informations retenues

1. Surface du bâtiment
2. Année de fabrication
3. Surface par type d'usage pour :
   - Multifamily Housing
   - Parking
   - Office
   - Non-Refrigerated Warehouse
   - Other
   - Hotel
   - Retail Store
   - Supermarket/Grocery Store
   - Hospital (General Medical & Surgical)
4. Nombre d'étage
5. Latitude
6. Longitude
7. Type d'usage pour :
   - Multifamily LR (1-4)
   - Multifamily Housing
   - NonResidential
   - Office
8. Surface du parking
9. S'il est du centre ville (DOWNTOWN)

Modèle Final

EnergyStar Score



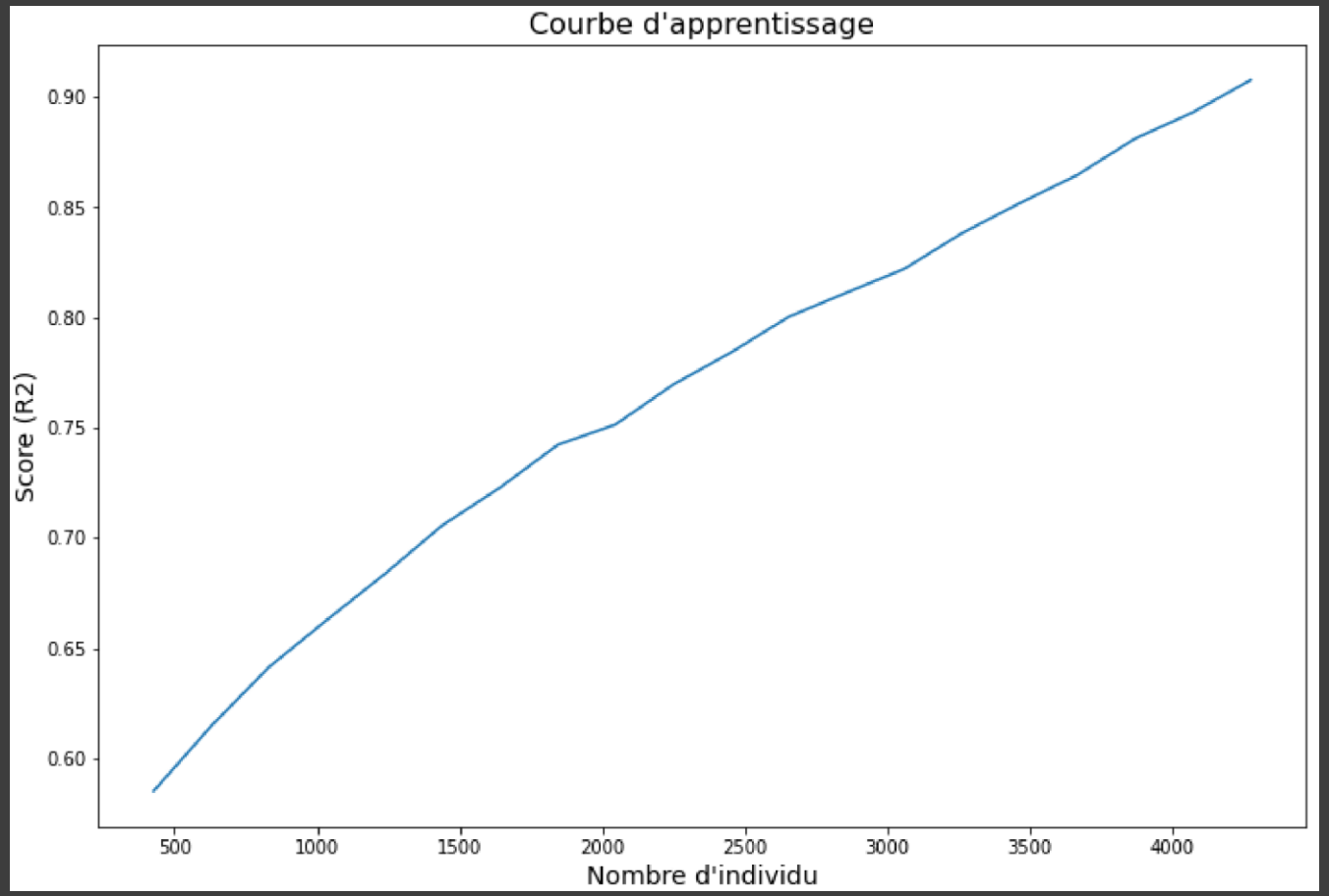
Score R^2 avec EnergyStarScore : 0.904

Score R^2 sans EnergyStarScore : 0.903

Modèle Final

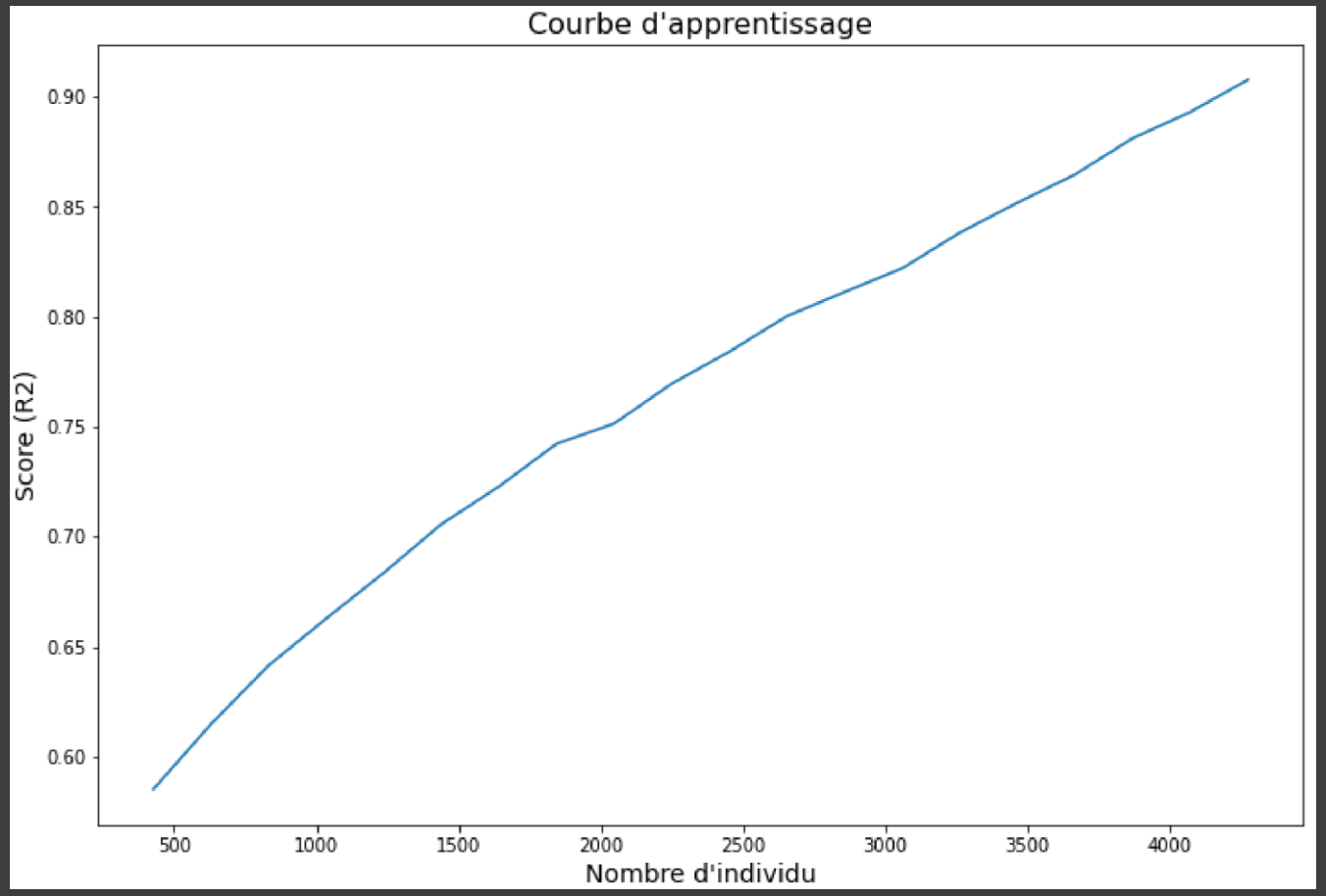
Learning Curve

8 Folds



Modèle Final

Learning Curve



Plus de données !