
**EA Stephen**
Data scientist

# Anticipez les besoins en consommation électrique de bâtiments


# SOMMAIRE

Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.

Présentation du cleaning effectué, du feature engineering et de l'exploration.

Présentation des différentes pistes de modélisation effectuées.

Présentation du modèle final sélectionné ainsi que des améliorations effectuées.

# La mission

Vous travaillez pour la **ville de Seattle**. Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, votre équipe s'intéresse de près aux émissions des bâtiments non destinés à l'habitation.

Des relevés minutieux ont été effectués par vos agents en 2015 et en 2016. Cependant, ces relevés sont coûteux à obtenir, et à partir de ceux déjà réalisés, **vous voulez tenter de prédire les émissions de CO2 et la consommation totale d'énergie** de bâtiments pour lesquels elles n'ont pas encore été mesurées.

Vous cherchez également à **évaluer l'intérêt de l'"ENERGY STAR Score" pour la prédiction d'émissions**, qui est fastidieux à calculer avec l'approche utilisée actuellement par votre équipe.

# Les variables par types

**Identification**

- **OSE Building ID**
- **PropertyName**
- **Taxe Parcel Identification Number**

**Localisation**

- **Address**
- **City**
- **State**
- **ZipCode**
- **Council District Code**
- **Neighborhood**
- **Latitude**
- **Longitude**

**Les données déclaratives du permis d'exploitation commerciale**

- **Building Type**
- **Primary Property Type**
- **YearBuilt**
- **Number of Buildings**
- **Number of Floors**
- **Property GFA Total**
- **Property GFA Parking**
- **Property GFA Building(s)**
- **List Of All Property Use Types**
- **Largest Property Type**
- **Largest Property Use Type GFA**
- **Second Largest Property Use Type**
- **Second Largest Property Use Type GFA**
- **Third Largest Property Use Type**
- **Third Largest Property Use Type GFA**

**Relevés énergétiques**

- **Site EUI(kBtu/sf)**
- **SiteEUIWN(kBtu/sf)**
- **SourceEUI(kBtu/sf)**
- **SourceEUIWN(kBtu/sf)**
- **Site Energy Use(kBtu)**
- **Site Energy Use(kBtu)**
- **Steam Use(kBtu)**
- **Electricity(kWh)**
- **Electricity(kBtu)**
- **Natural Gas(therms)**
- **Natural Gas(kBtu)**

**Performance énergétique du bâtiment**

- **Years ENERGY STAR Certified**
- **ENERGYSTAR Score**

**Calcul des émissions CO2**

- Total GHG Emissions
- GHG Emissions Intensity

**Info data**

- Data Year
- Comments
- DefaultData
- Outlier
- Compliance Status

# Objectif

**Prédire les émissions de CO2 et la consommation totale d'énergie**

- **Les targets:** Total GHG Emissions, Site Energy Use(kBtu)
- **Data cleaning** afin de modéliser le phénomène le plus représentatif et ne pas sensibiliser les modèles
- **Evaluer l'intérêt de l'"ENERGY STAR Score" pour la prédiction d'émissions**
- **Les modèles se baseront sur les fonctionnalités** Identification, Localisation, Les données déclaratives du permis d'exploitation commerciale et des variables produites à l'aide de features engineering

# Data leakage

Méthodologie suivie afin de ne pas fournir des modèles trop optimiste et invalide en production

➢ Features leakage

1. Vérifier relation avec la target une forte relation peut indiquer un problème de target leakage
2. Vérifier que le modèle ne fournit pas des performances trop optimiste cela peut etre indication.
3. Pour éviter ce type de fuite de données, toute variable mise à jour (ou créée) après la réalisation de la valeur cible doit être exclue.

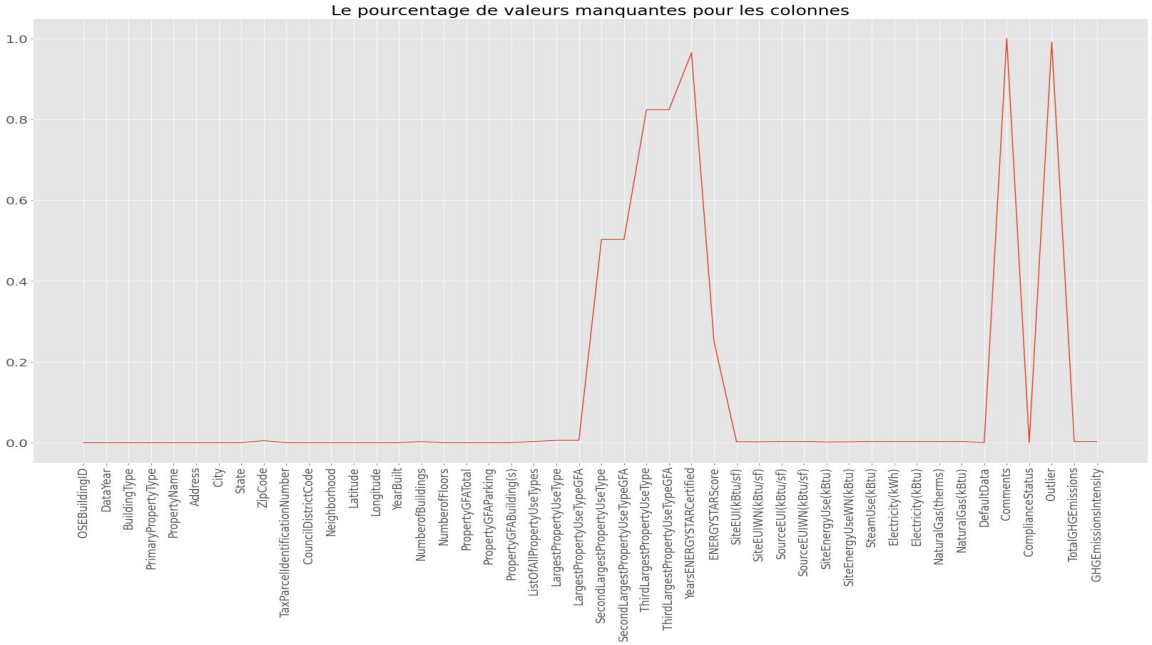
➢ Data leakage

1. Utiliser pipeline
2. Incohérence data (effectuer m transformation
3. Ne pas faire transformation sur le dataset (fit, fit_transform)
4. Faire des transformations uniquement sur le train_set.
5. Séparer le dataset en train_set et test_set avant l'exploration

# Description du dataset 2016

- Les dimensions du dataset 2016: (3376, 46)
- Le types des colonnes du dataset :
  dtypes: bool(1), float64(22), int64(8), object(15)
- Memory usage; 1.2+ MB
- 12.85 % NaN dans le dataset

  soit (19 9552 NaN)




| | Count_NaN | %_NaN_col | Total_NaN_in_dataset | %_NaN_in_dataset | Types |
|---|---|---|---|---|---|
| Comments | 3376 | 100.0 | 19952 | 12.85 | float64 |
| Outlier | 3344 | 99.05 | 19952 | 12.85 | object |
| YearsENERGYSTARCertified | 3257 | 96.48 | 19952 | 12.85 | object |
| ThirdLargestPropertyUseType | 2780 | 82.35 | 19952 | 12.85 | object |
| ThirdLargestPropertyUseTypeGFA | 2780 | 82.35 | 19952 | 12.85 | float64 |
| SecondLargestPropertyUseType | 1697 | 50.27 | 19952 | 12.85 | object |
| SecondLargestPropertyUseTypeGFA | 1697 | 50.27 | 19952 | 12.85 | float64 |
| ENERGYSTARScore | 843 | 24.97 | 19952 | 12.85 | float64 |
| LargestPropertyUseTypeGFA | 20 | 0.59 | 19952 | 12.85 | float64 |
| LargestPropertyUseType | 20 | 0.59 | 19952 | 12.85 | object |
| ZipCode | 16 | 0.47 | 19952 | 12.85 | float64 |
| ListOfAllPropertyUseTypes | 9 | 0.27 | 19952 | 12.85 | object |
| SourceEUIWN(kBtu/sf) | 9 | 0.27 | 19952 | 12.85 | float64 |
| SourceEUI(kBtu/sf) | 9 | 0.27 | 19952 | 12.85 | float64 |
| Electricity(kWh) | 9 | 0.27 | 19952 | 12.85 | float64 |
| Electricity(kBtu) | 9 | 0.27 | 19952 | 12.85 | float64 |
| NaturalGas(therms) | 9 | 0.27 | 19952 | 12.85 | float64 |
| NaturalGas(kBtu) | 9 | 0.27 | 19952 | 12.85 | float64 |
| TotalGHGEmissions | 9 | 0.27 | 19952 | 12.85 | float64 |
| SteamUse(kBtu) | 9 | 0.27 | 19952 | 12.85 | float64 |
| GHGEmissionsIntensity | 9 | 0.27 | 19952 | 12.85 | float64 |
| NumberofBuildings | 8 | 0.24 | 19952 | 12.85 | float64 |
| SiteEUI(kBtu/sf) | 7 | 0.21 | 19952 | 12.85 | float64 |
| SiteEUIWN(kBtu/sf) | 6 | 0.18 | 19952 | 12.85 | float64 |
| SiteEnergyUseWN(kBtu) | 6 | 0.18 | 19952 | 12.85 | float64 |
| SiteEnergyUse(kBtu) | 5 | 0.15 | 19952 | 12.85 | float64 |
| TaxParcelIdentificationNumber | 0 | 0.0 | 19952 | 12.85 | object |
| BuildingType | 0 | 0.0 | 19952 | 12.85 | object |
| PrimaryPropertyType | 0 | 0.0 | 19952 | 12.85 | object |
| ComplianceStatus | 0 | 0.0 | 19952 | 12.85 | object |
| PropertyName | 0 | 0.0 | 19952 | 12.85 | object |
| DefaultData | 0 | 0.0 | 19952 | 12.85 | bool |
| Address | 0 | 0.0 | 19952 | 12.85 | object |
| City | 0 | 0.0 | 19952 | 12.85 | object |
| State | 0 | 0.0 | 19952 | 12.85 | object |
| PropertyGFABuilding(s) | 0 | 0.0 | 19952 | 12.85 | int64 |
| CouncilDistrictCode | 0 | 0.0 | 19952 | 12.85 | int64 |
| PropertyGFAParking | 0 | 0.0 | 19952 | 12.85 | int64 |
| Neighborhood | 0 | 0.0 | 19952 | 12.85 | object |
| Latitude | 0 | 0.0 | 19952 | 12.85 | float64 |
| Longitude | 0 | 0.0 | 19952 | 12.85 | float64 |
| YearBuilt | 0 | 0.0 | 19952 | 12.85 | int64 |
| NumberofFloors | 0 | 0.0 | 19952 | 12.85 | int64 |
| PropertyGFATotal | 0 | 0.0 | 19952 | 12.85 | int64 |
| DataYear | 0 | 0.0 | 19952 | 12.85 | int64 |
| OSEBuildingID | 0 | 0.0 | 19952 | 12.85 | int64 |

# Data cleaning train_set (80%)

- Filtrer les bâtiments non habitations
  'NonResidential','Nonresidential COS','SPS-District K-12', 'Nonresidential WA', 'Campus'

- Filtrer les bâtiments non habitations
  Compliance Status = compliant

- Supprimer OSbuildingID 4978
  Valeur aberrante sur Total GHG Emissions (-0.8 )
  Supprimer OSbuildingID 700
  Valeurs aberrantes Total GHG Emissions(0) et SiteEnergie kBtu (12525174)

- Supprimer OSbuildingID 19445 , Supprimer OSbuildingID 21481, Supprimer OSbuildingID 25674
  Low-Rise Multifamily (PrimaryPropertyType)

- Supprimer les **Number of Buildings == 0**

- Transformation lower
  La colonne Neighborhood car mélange en majuscules et minuscules qui ne fait pas correspondre les catégories entre elles

- Fillna 'not outlier'
  Après le data clean, il n'y a plus de 'High outlier' 'Low outlier'

- Supprimer les colonnes:
  'City', 'State', 'Comments', 'Outlier', 'DefaultData', 'NaturalGas(therms)', 'Electricity(kWh)',
  'SiteEnergyUseWN(kBtu)','SourceEUIWN(kBtu/sf)','SiteEUIWN(kBtu/sf)'

# Feature engineering

Création nouvelles variables pct énergie

**% Electricity, % Steam, % Natural Gas**

Création variable booléen pour usage oui ou non de l'énergie

**Bool Electricity, Bool Steam, Bool Natural Gas**

Création variable superficie par étages

**GFAperFloor**

pct surface parking du bâtiments
pct surface du bâtiment hors parking

**% Property GFA Parking, % Property GFA Building(s)**

Création âge du bâtiment

**BuildingAge**

Création colonnes targets en Log

**Total GHG Emissions LOG, Site Energy Use kBtu LOG**

## Les variables conservées pour la modélisation

- % Electricity
- % Steam
- % Natural Gas
- BuildingAge
- Total GHG Emissions LOG,
- Site Energy Use kBtu LOG

# Sélections des variables sur le train_set (80%)

- ❑ **Non conservées dans le modèle final**
- ❑ **Redondance et non conservées dans le modèle final**
- ❑ Redondance et non conservées dans le modèle final
- ❑ **Features leakage**

## Les variables supprimées

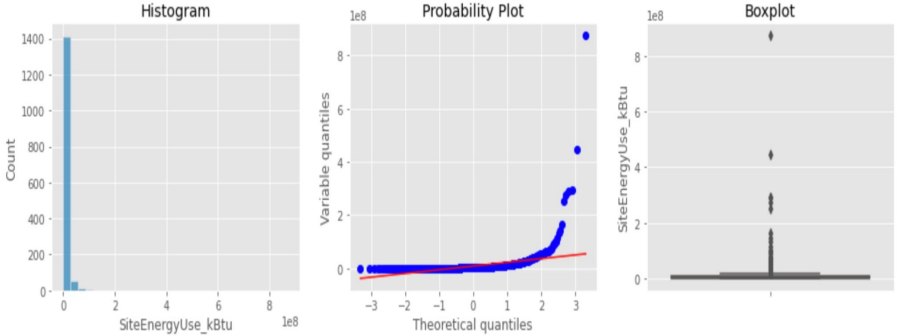
| | |
|---|---|
| DataYear | 0 |
| Address | 0 |
| ZipCode | 7 |
| TaxParcelIdentificationNumber | 0 |
| CouncilDistrictCode | 0 |
| Latitude | 0 |
| Longitude | 0 |
| YearBuilt | 0 |
| NumberofBuildings | 0 |
| NumberofFloors | 0 |
| PropertyGFAParking | 0 |
| PropertyGFABuilding(s) | 0 |
| ListOfAllPropertyUseTypes | 0 |
| LargestPropertyUseType | 3 |
| LargestPropertyUseTypeGFA | 3 |
| SecondLargestPropertyUseType | 560 |
| SecondLargestPropertyUseTypeGFA | 560 |
| ThirdLargestPropertyUseType | 937 |
| ThirdLargestPropertyUseTypeGFA | 937 |
| YearsENERGYSTARCertified | 1117 |
| SiteEUI_kBtu/sf | 0 |
| SourceEUI_kBtu/sf | 0 |
| SteamUse_kBtu | 0 |
| Electricity_kBtu | 0 |
| NaturalGas_kBtu | 0 |
| GHGEmissionsIntensity | 0 |

## Les variables conservées

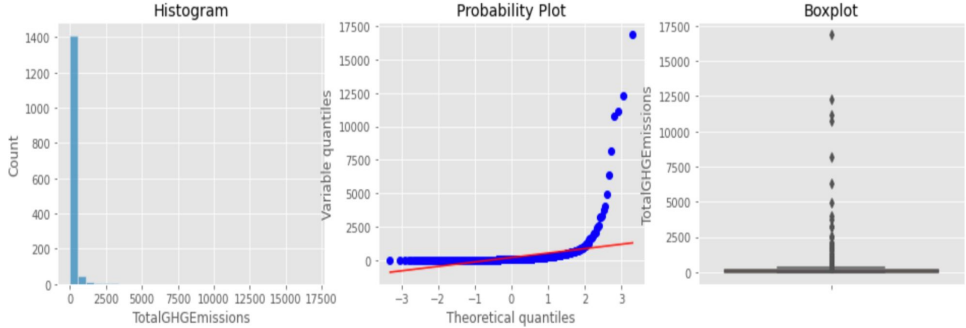
| | |
|---|---|
| PrimaryPropertyType | 0 |
| PropertyName | 0 |
| Neighborhood | 0 |
| PropertyGFATotal | 0 |
| ENERGYSTARScore | 409 |
| BuildingAge | 0 |
| TotalGHGEmissions_LOG | 0 |
| SiteEnergyUse_kBtu_LOG | 0 |
| SiteEnergyUse_kBtu | 0 |
| TotalGHGEmissions | 0 |
| %_Electricity | 0 |
| %_Steam | 0 |
| %_NaturalGas | 0 |

# Data exploration des targets

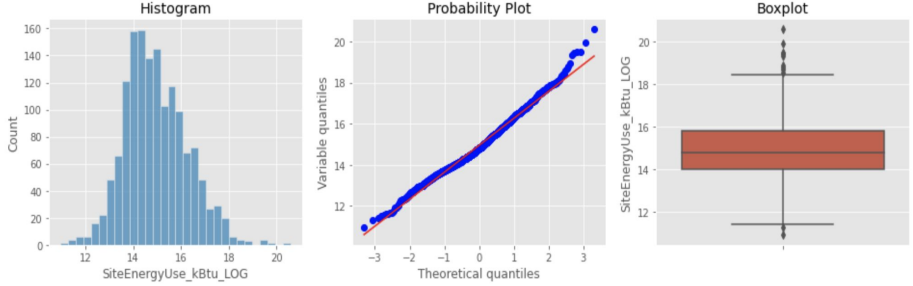
SiteEnergyUse_kBtu
Test Shapiro
stat=0.195, p=0.000
Probablement pas Gaussien
Test normaltest
stat=3137.305, p=0.000
Probablement pas Gaussien



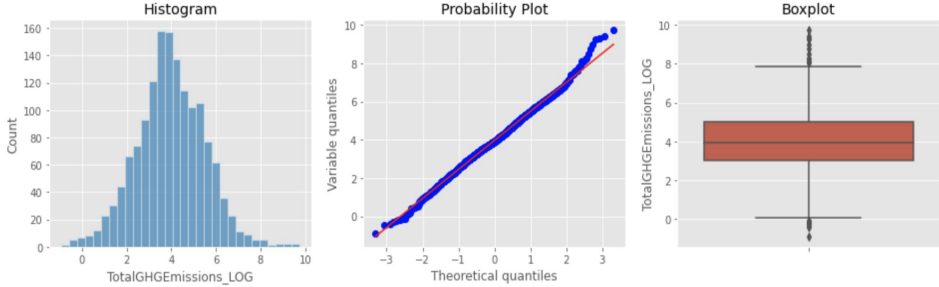

TotalGHGEmissions
Test Shapiro
stat=0.178, p=0.000
Probablement pas Gaussien
Test normaltest
stat=2782.591, p=0.000
Probablement pas Gaussien



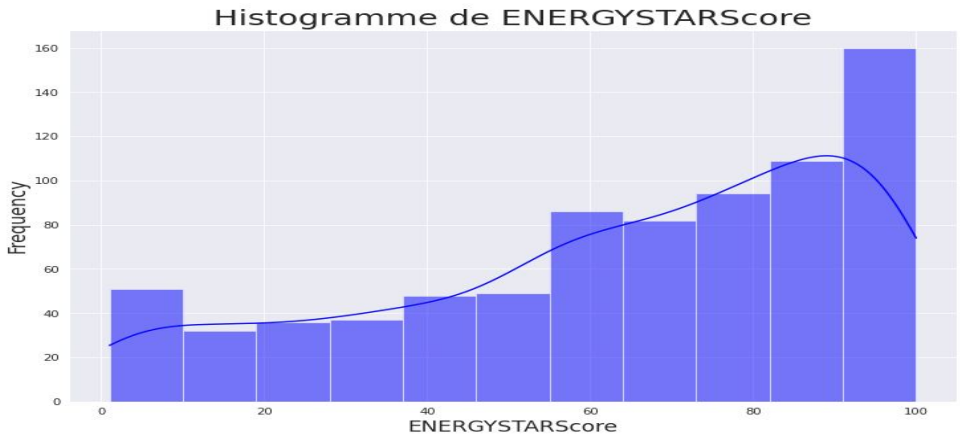

SiteEnergyUse_kBtu_LOG
Test Shapiro
stat=0.989, p=0.000
Probablement pas Gaussien
Test normaltest
stat=42.133, p=0.000
Probablement pas Gaussien

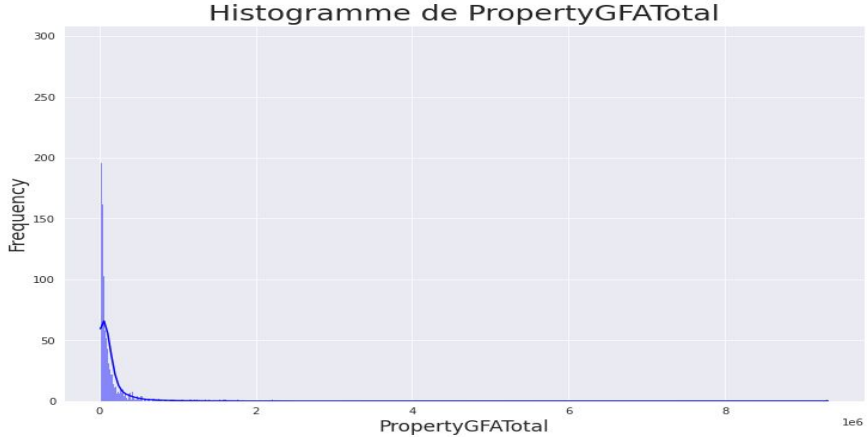


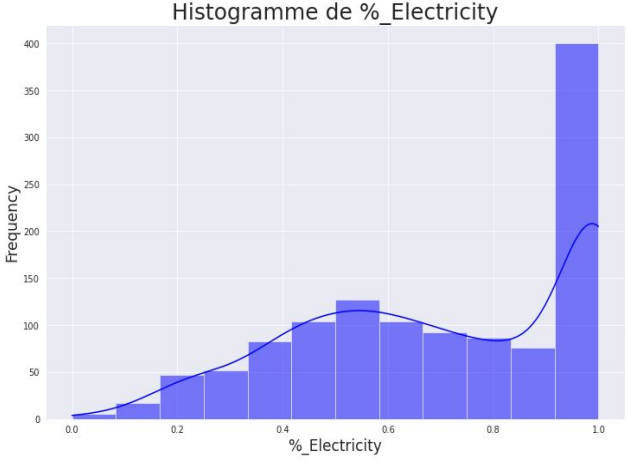

TotalGHGEmissions_LOG
Test Shapiro
stat=0.997, p=0.006
Probablement pas Gaussien
Test normaltest
stat=8.508, p=0.014
Probablement pas Gaussien

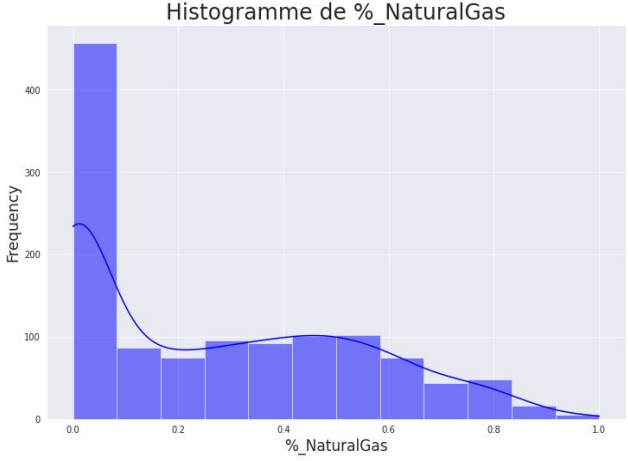


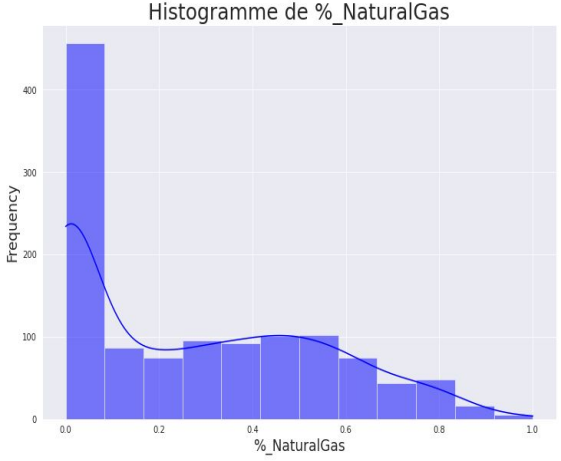
# Data exploration quantitatives



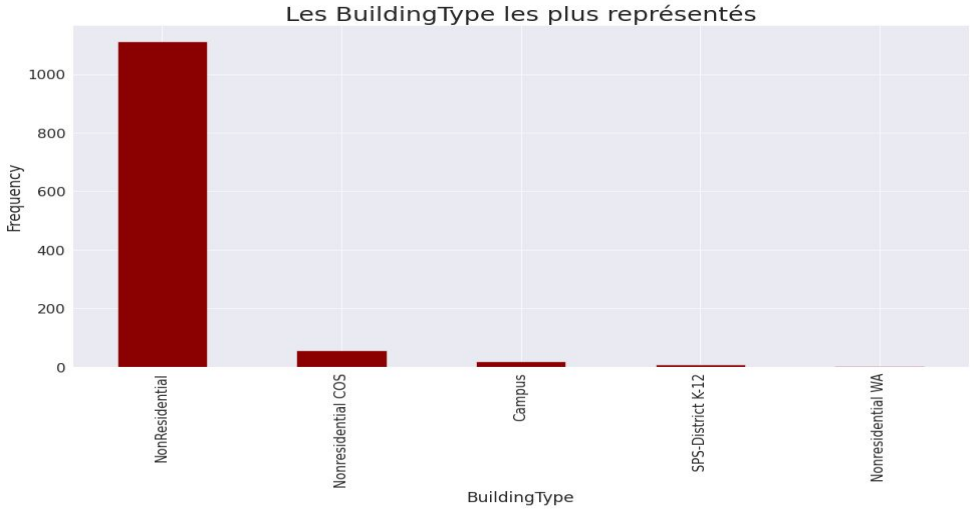



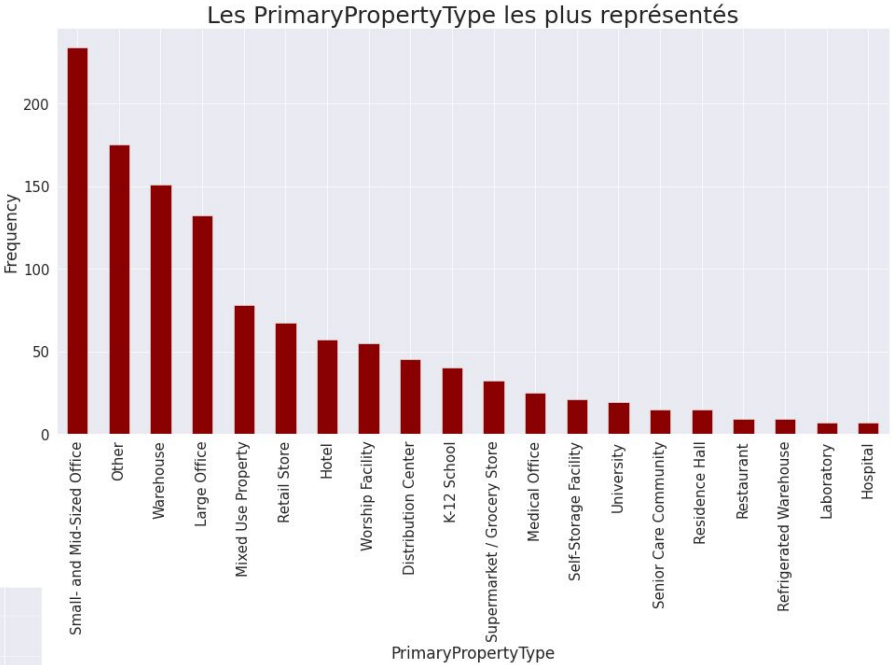



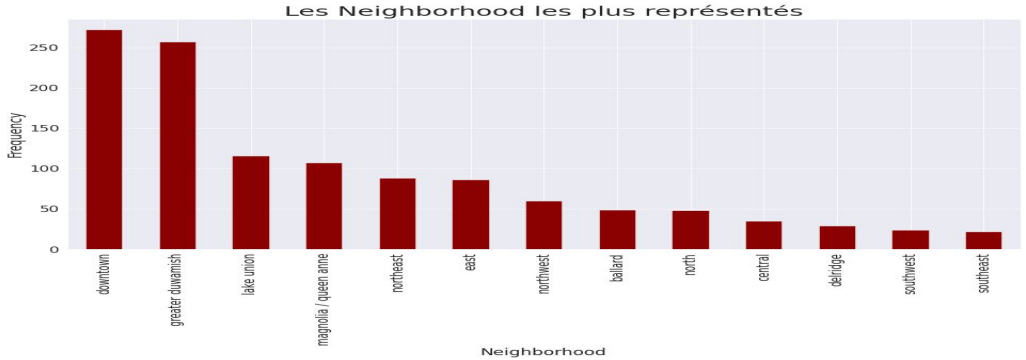



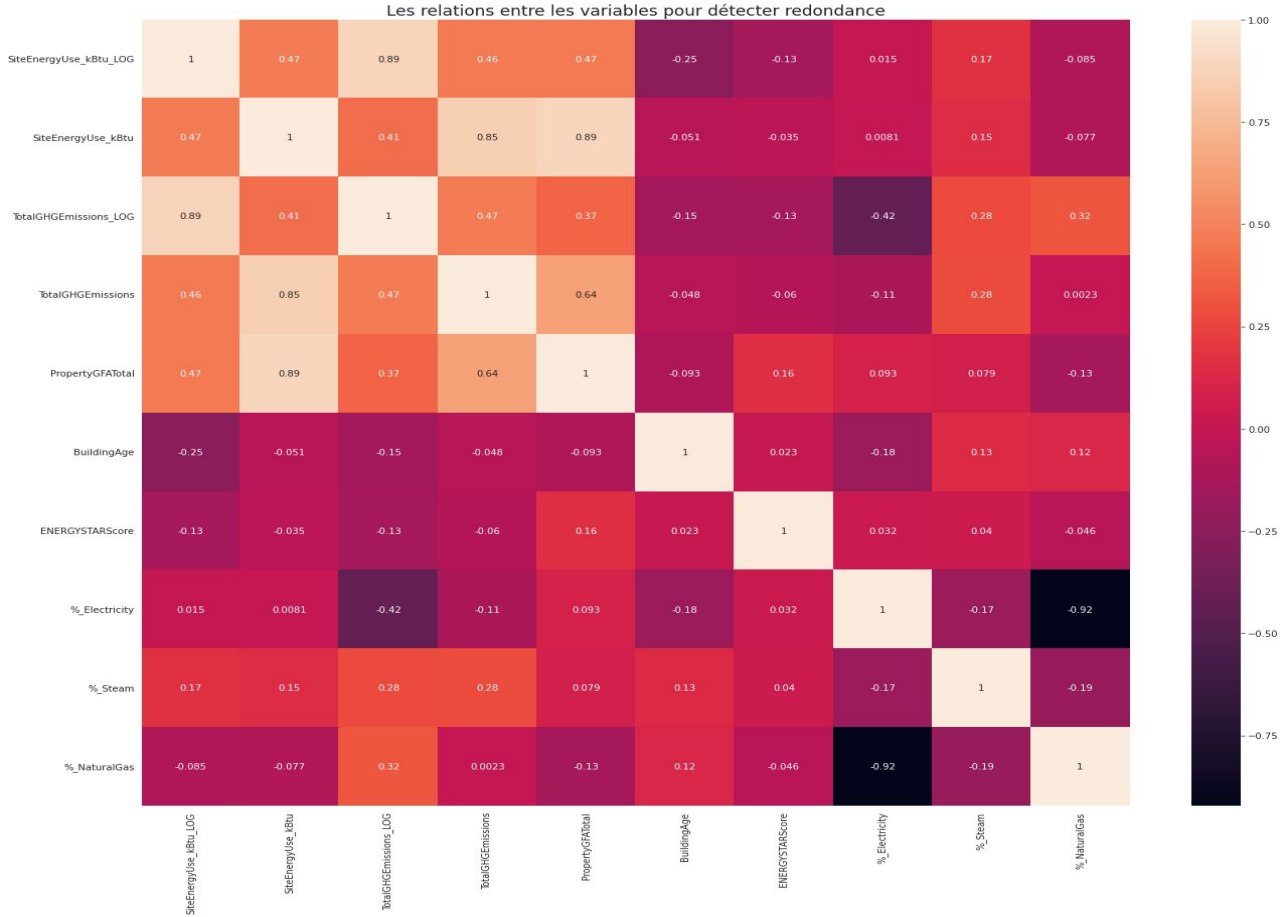
# Data exploration catégorielle



# Data exploration bivariées quantitatives



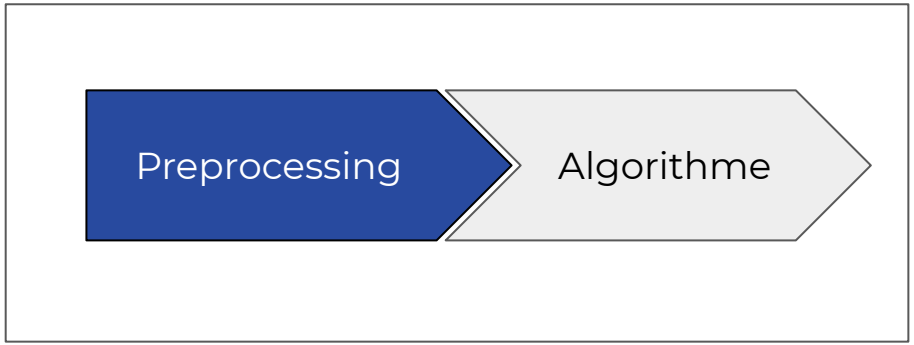

| | feature | VIF |
|---|---|---|
| 0 | PropertyGFATotal | 1.155777 |
| 1 | %_Electricity | 1.409629 |
| 2 | %_Steam | 1.038979 |
| 3 | %_NaturalGas | 1.249653 |

# Plan procédure d'évaluation de la meilleur pipeline

Outliers

X_train

Pipeline




RandomizedSearchCV et Grid Search CV

**(shuffle=True, cv=5, random_state=77)**
**Recherche des meilleurs hyperparamètres**
**grid.fit(X_train, y_train)**

Best pipeline

Best score CV
Temps prédiction

pipeline ajustée

Evaluation des pipelines paramètrer

**Scores métrics CV (R2, RMSE, MSE, MAE)**
**Courbe d'apprentissage**
**Influence des hyperparamètres**
**Importance des fonctionnalités**

# Méthodologie preprocessing pipeline

## Les variables quantitatives

### 1 Feature engineering

- avec Polynomial Features ou sans

### 2 Scaling

- StandardScaler, RobustScaler, QuantileTransformer, PowerTransformer

### 3 Feature selection ou réduction dimensions

- SelectKBest(**f_regression**, **mutual_info_regression**), VarianceThreshold, ACP, Isomap

## Les variables quantitatives NaN

### 1 Imputation

- SimpleImpute (mean, median, constant, most_fréquent)
- DropNa
- KNN imputer
- Iterative imputer

### 2 Scaling

### 3 Feature selection ou réduction dimensions

## Les variables catégorielles

### Encodage

- OneHotEncoder, OrdinalEncoder,

# Les algorithmes choisis pour les pipelines

## Baseline

DummyRegressor
(strategy='mean')

## Linéaire

Linear Regression
Lasso
Ridge
ElasticNet
SGDRegressor

## Non linéaire

Polynomial regression
SVR
KernelRidge
KNN Regressor

## Ensemble

DecisionTreeRegressor
RandomForestRegressor

## Bagging

BaggingRegressor
Base_estimator =
KernelRidge
BaggingRegressor
Base_estimator =
RandomForrest

## Boosting

GradientBoostingRegressor
AdaBoostRegressor
XGBRegressor

# Améliorer les modèles

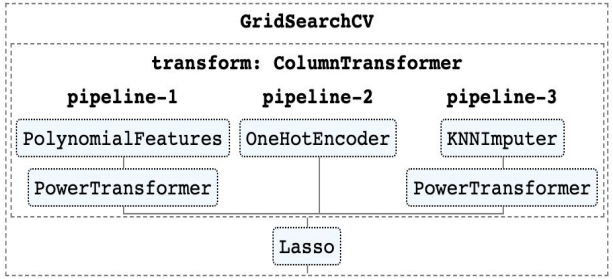
## Améliorer le modèle

1. Collecte de données : Augmenter le nombre d'exemples de formation
2. Traitement des entités : Ajouter d'autres variables et un meilleur traitement des entités
3. Vérifier l'importance des fonctionnalités afin de supprimer les variables qui n'ont pas d'importance
4. Réglage des hyperparamètres du modèle : Considérez d'autres valeurs pour les paramètres de formation utilisés par les algorithmes d'apprentissages

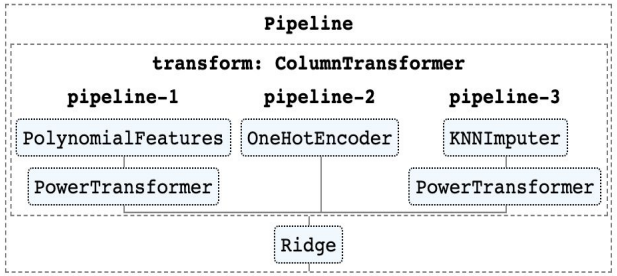
## Overfitting

1. Ajouter des données d'entraînements
2. Retirer des features (multicolinéarité , pas de variance)
3. Méthodes de régularisations
4. Réglage hyperparamètres
5. Choix algorithme faible variance

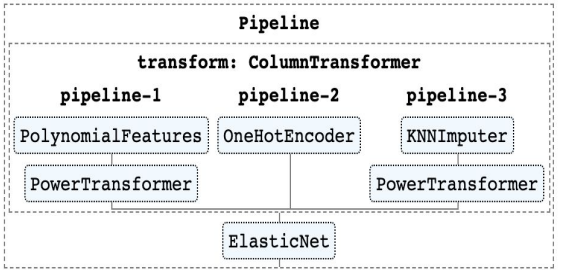
# Procédure d'évaluation consommation totale énergie



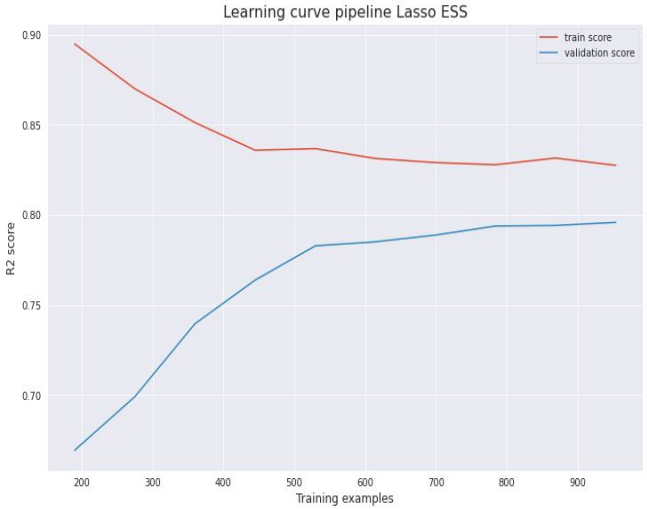

Best score R2: 0.7984

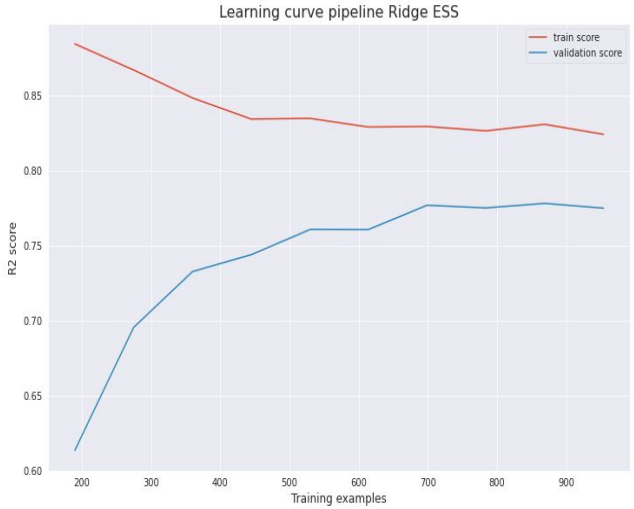


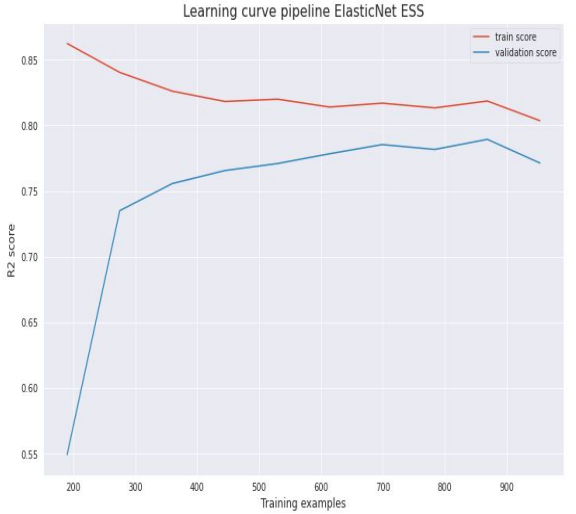

Best score R2: 0.7911



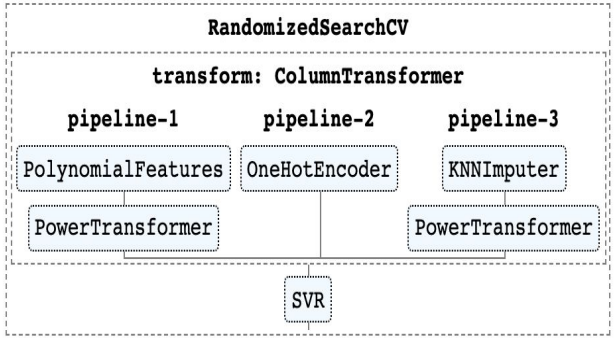

Best score R2: 0.7783



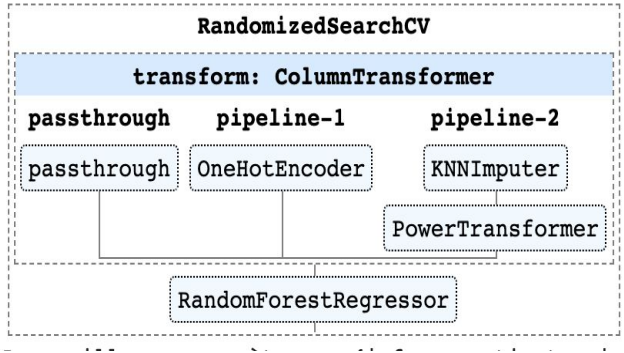



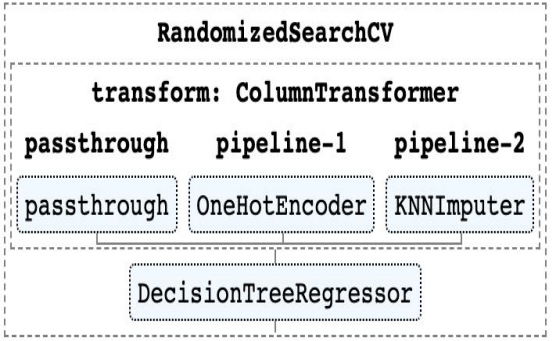
# Procédure d'évaluation consommation totale énergie



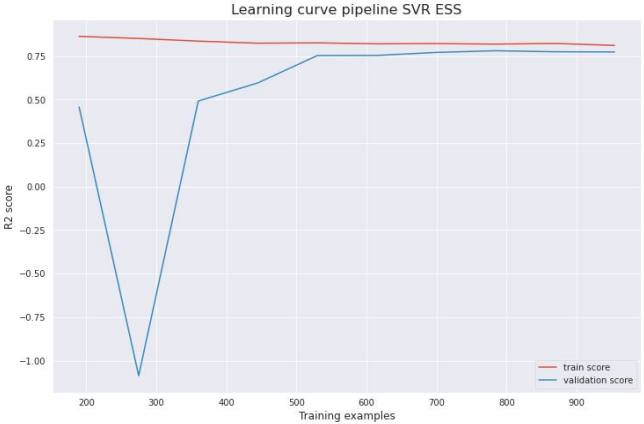

Best score R2: 0.7835

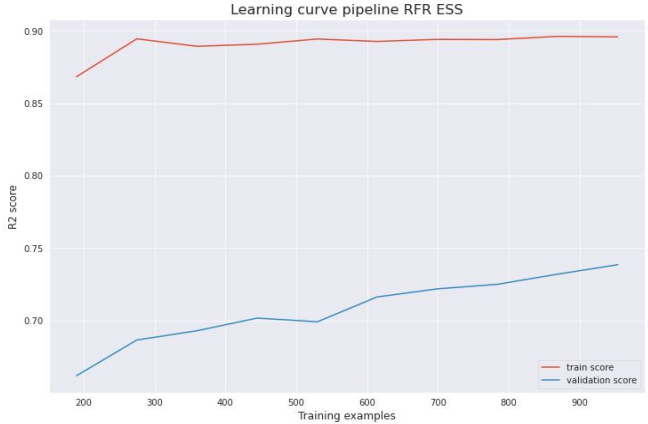


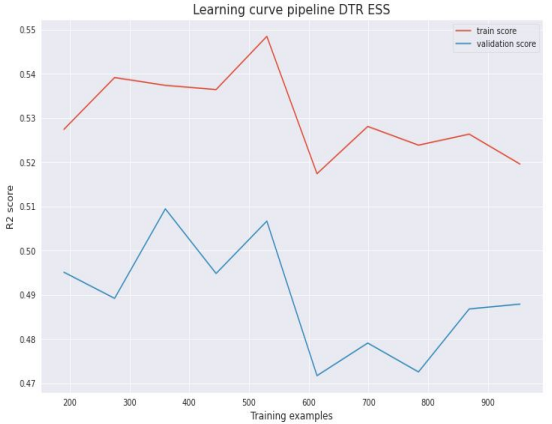

Best score R2: 0.7421



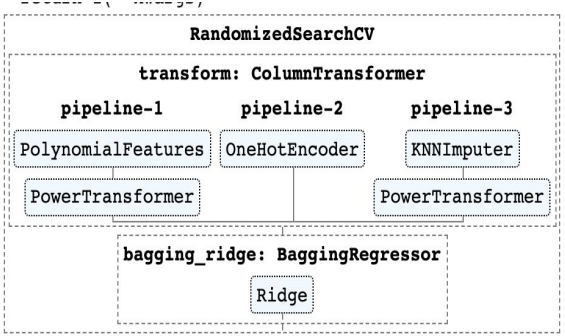

Best score R2: 0.5323



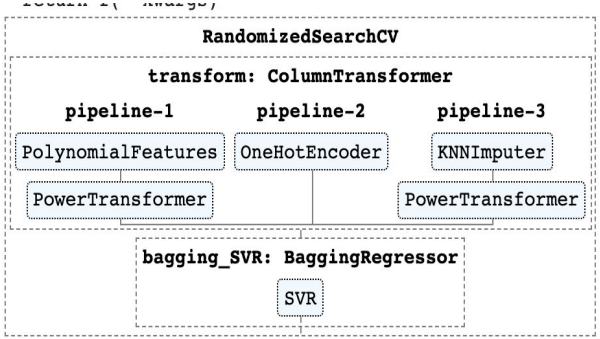



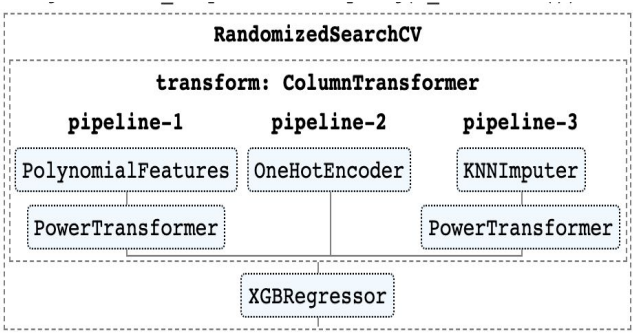
# Procédure d'évaluation consommation totale énergie



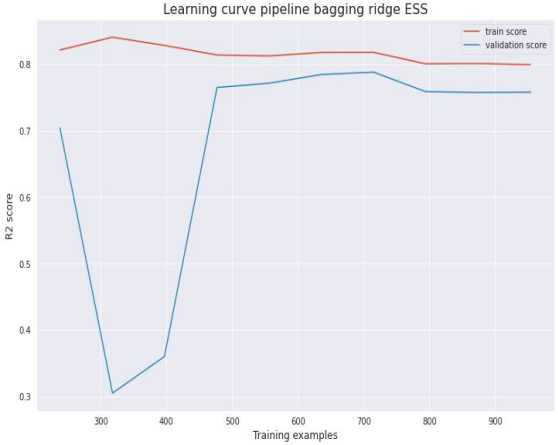

Best score R2: 0.7911

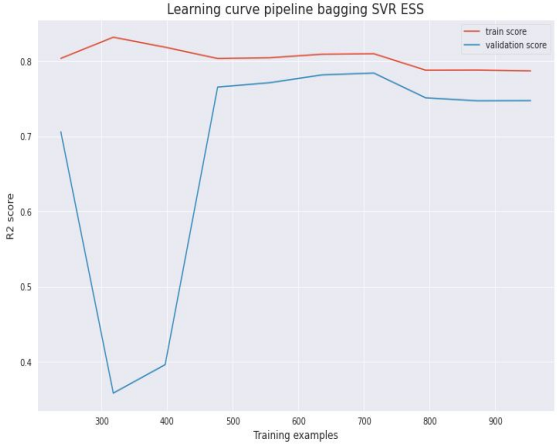


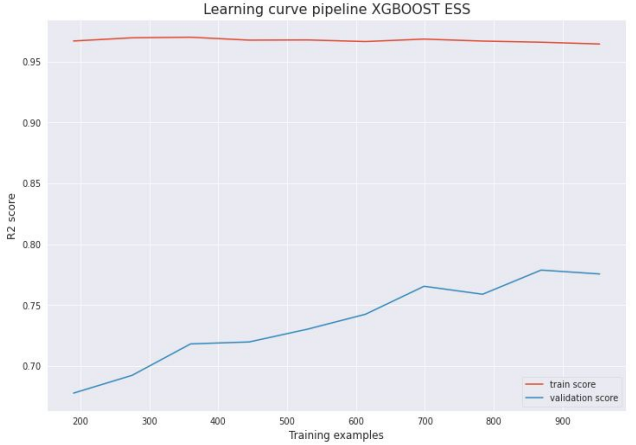

Best score R2: 0.7878



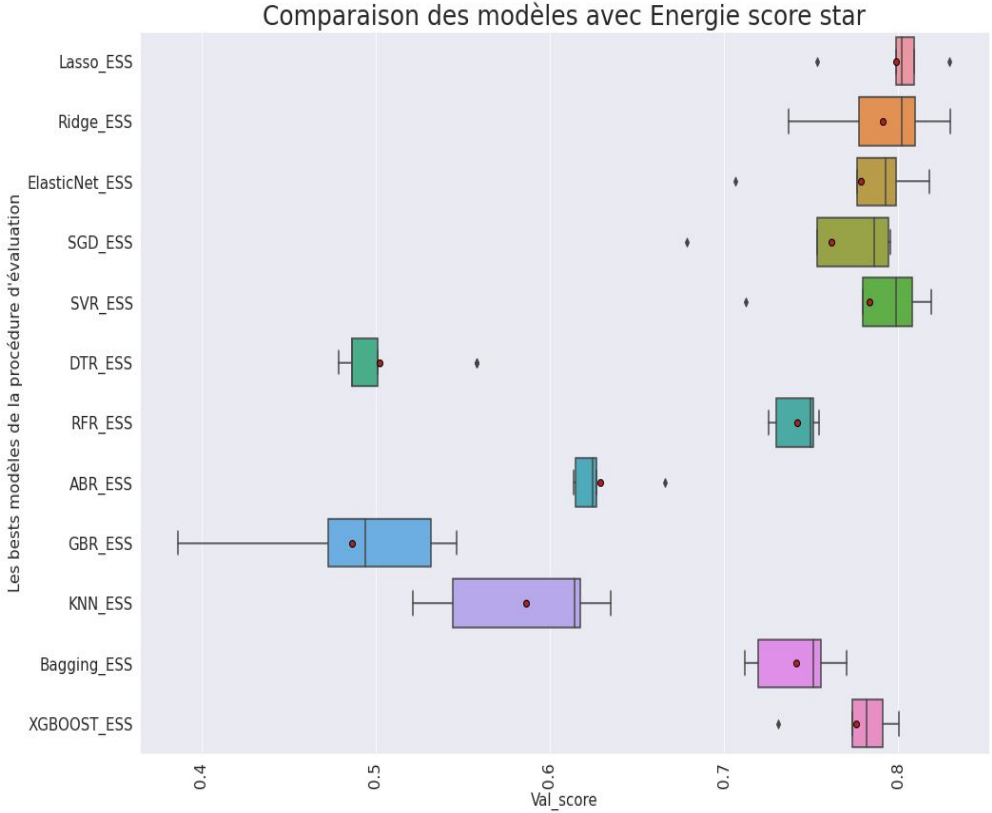

Best score R2: 0.7754

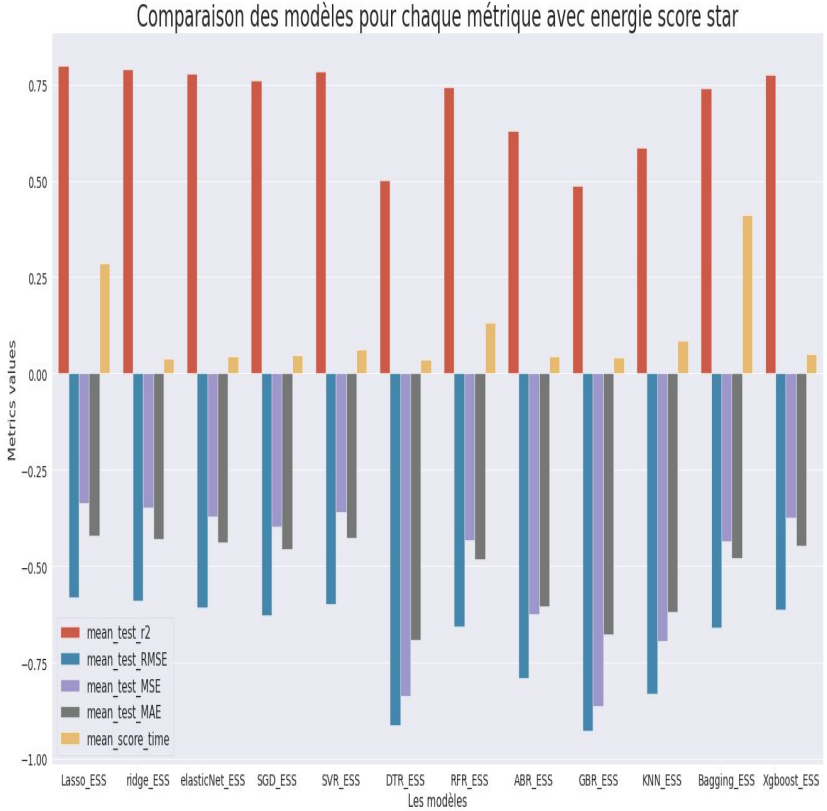


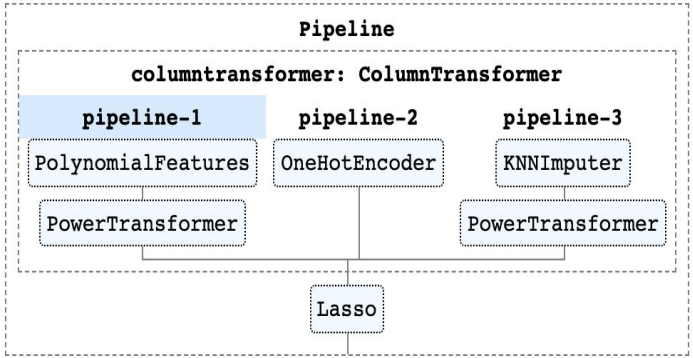
# Comparaison des pipelines energie totale

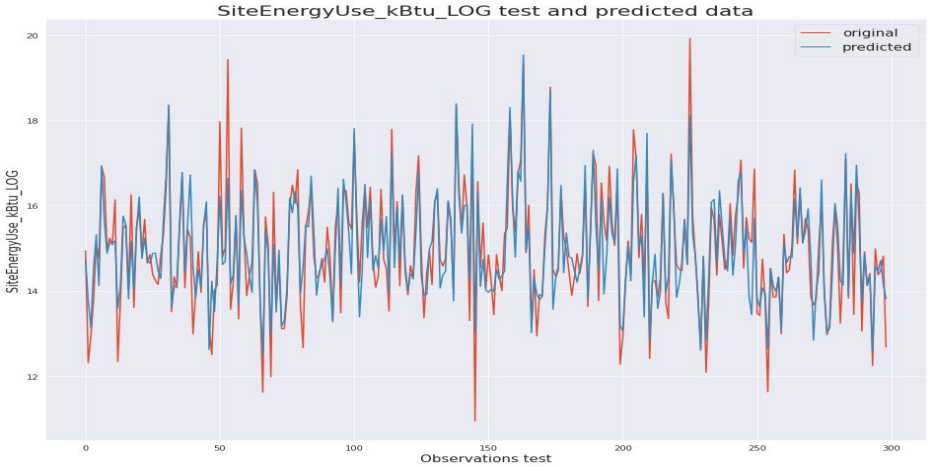


# Le modèle final sélectionné

LASSO

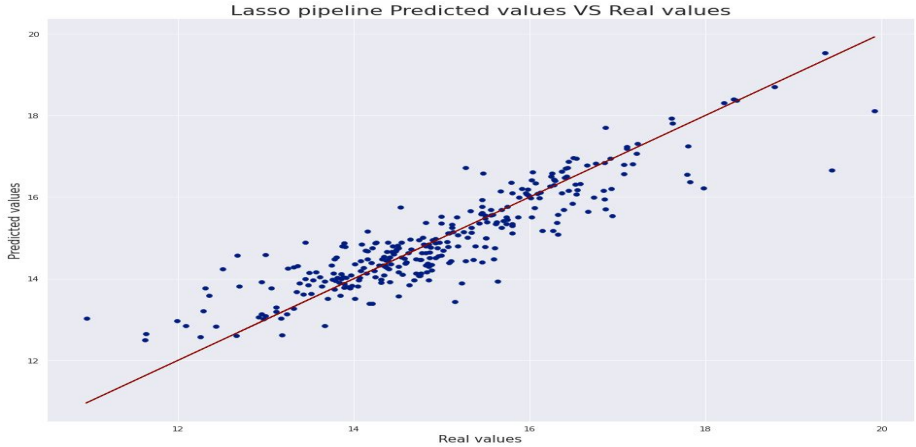


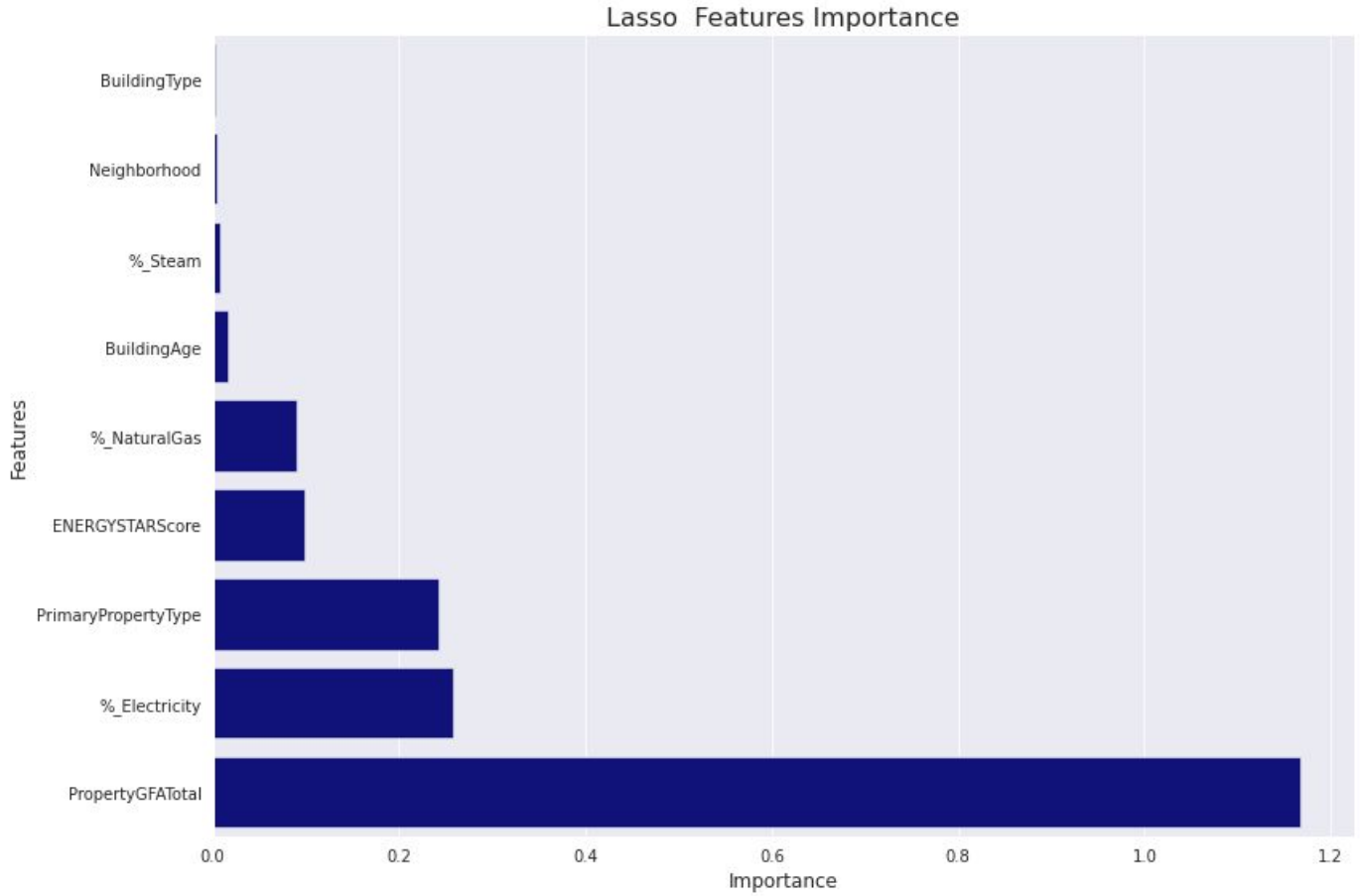

Le score R2 sur le test_set la pipeline Lasso: 0.8079505288519834

train MAE: 0.3903179126898014
train MSE: 0.29612299244936696
train RMSE: 0.5441718409191778
train R2: 0.823815795731122

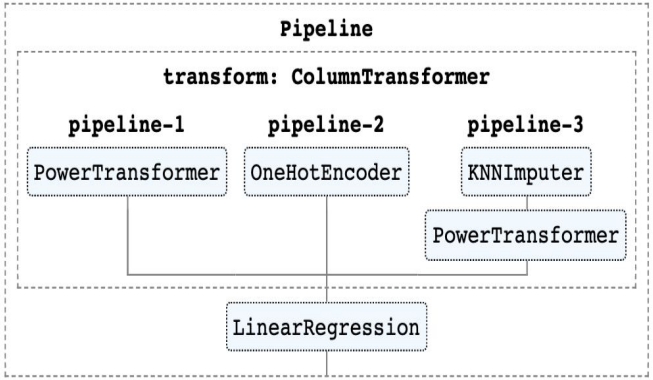
test MAE: 0.3903179126898014
test MSE: 0.36933292489162584
test RMSE: 0.6077276732975271
test R2: 0.8079505288519834



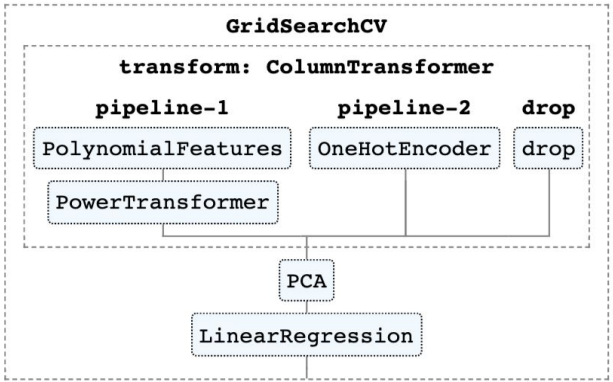
# Interprétation du modèle consommation energie total



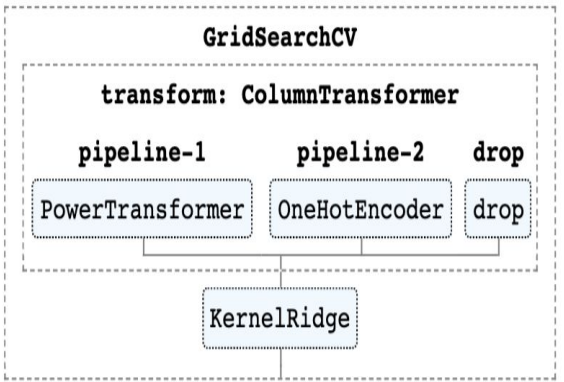
# Procédure d'évaluation émissions de CO2



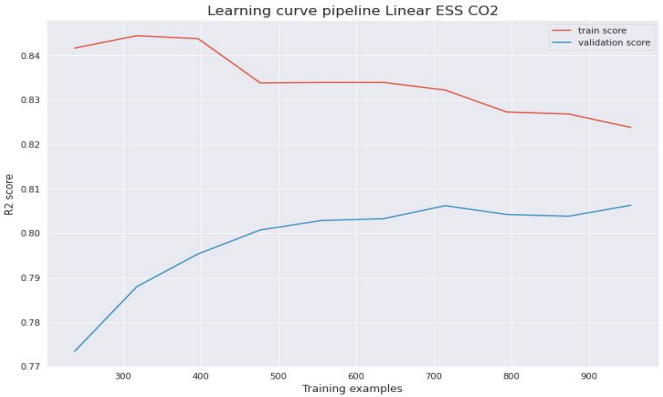

Best score R2: 0.8053

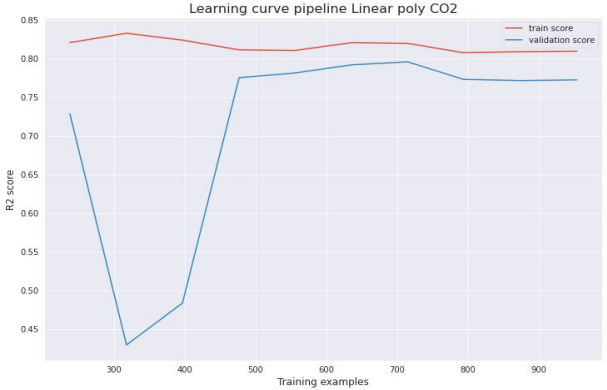


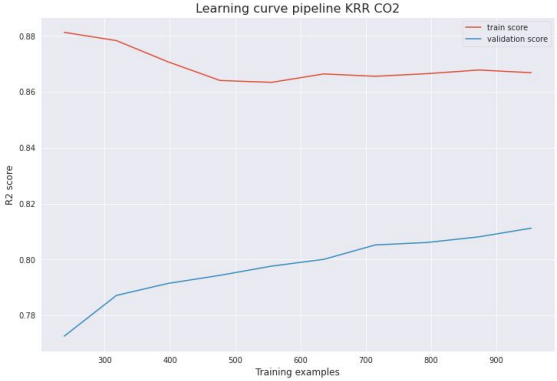

Best score R2: 0.7871



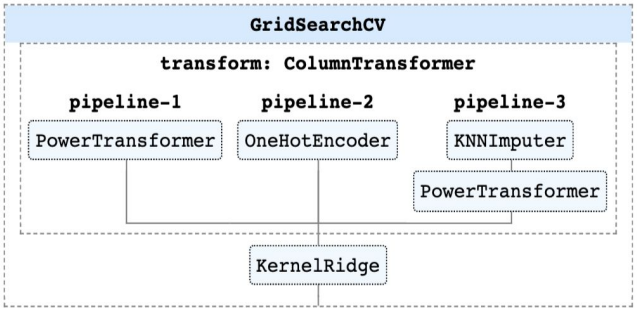

Best score R2: 0.8116



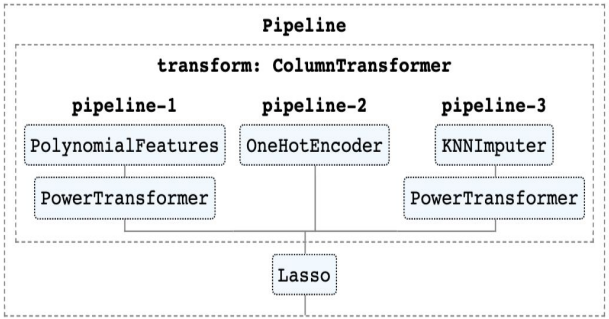



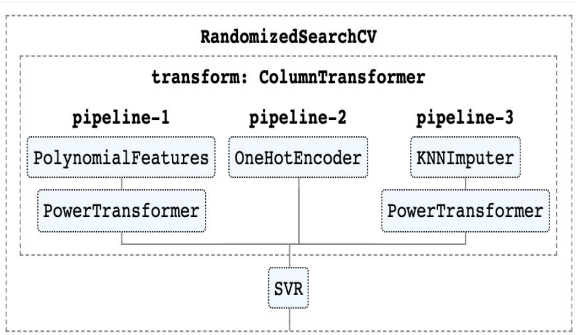
# Procédure d'évaluation émissions de CO2



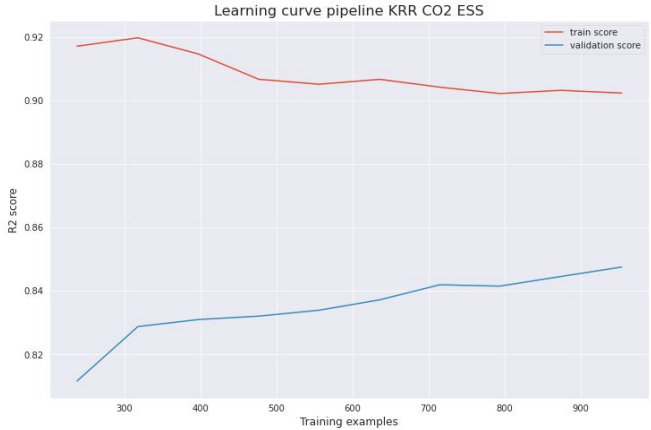

Best score Rcarré: 0.8500

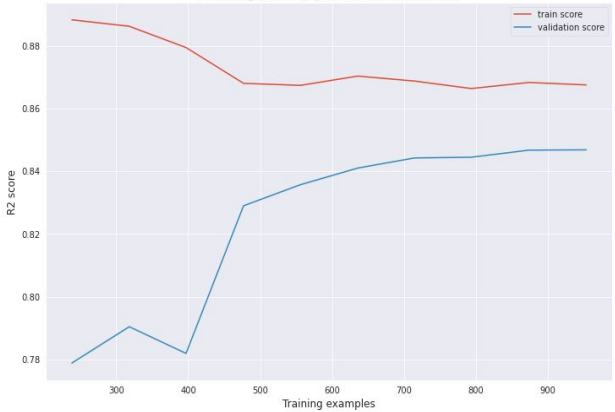


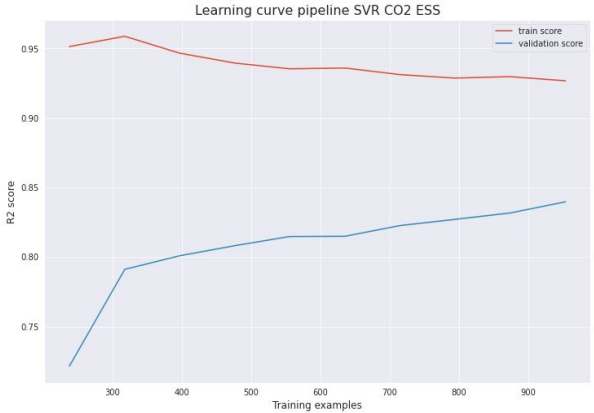

Best score Rcarré: 0.8494



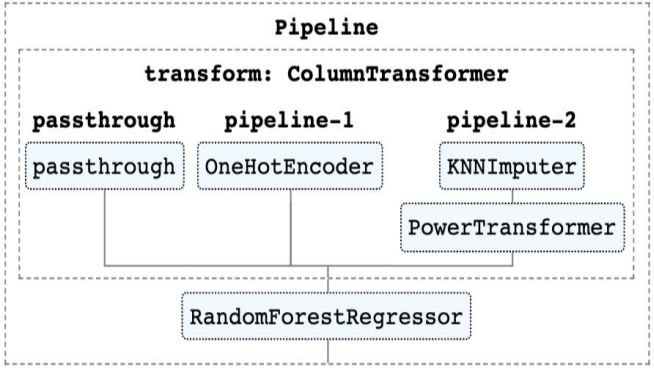

Best score Rcarré: 0.8290



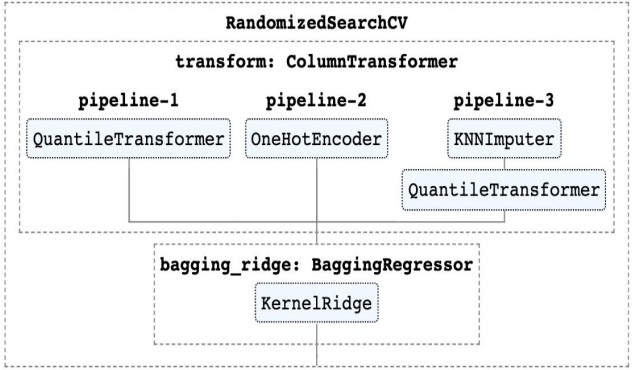



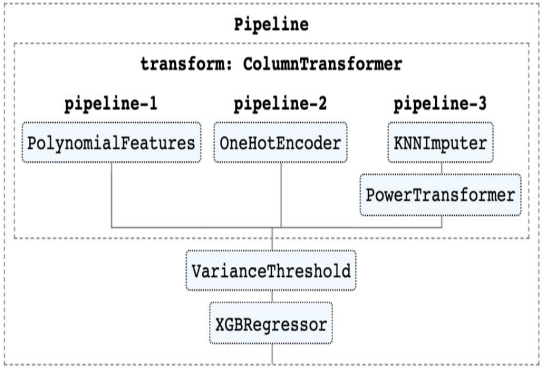
# Procédure d'évaluation émissions de CO2



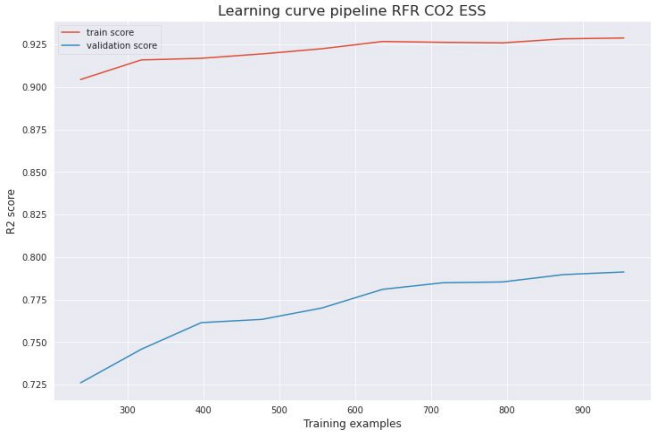

Best score Rcarré: 0.7867

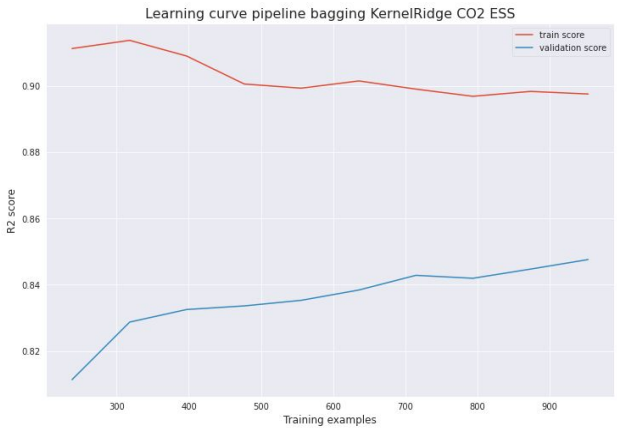


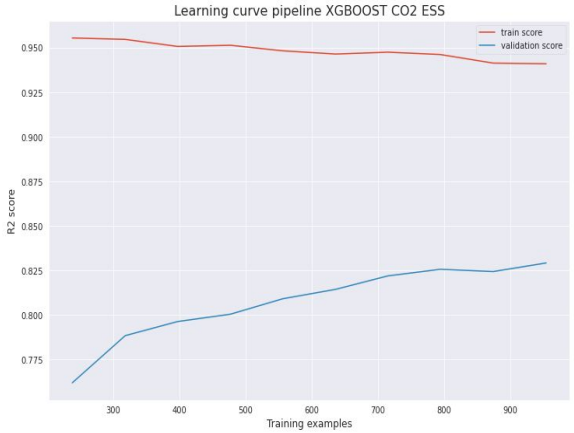

Best score Rcarré: 0.8499



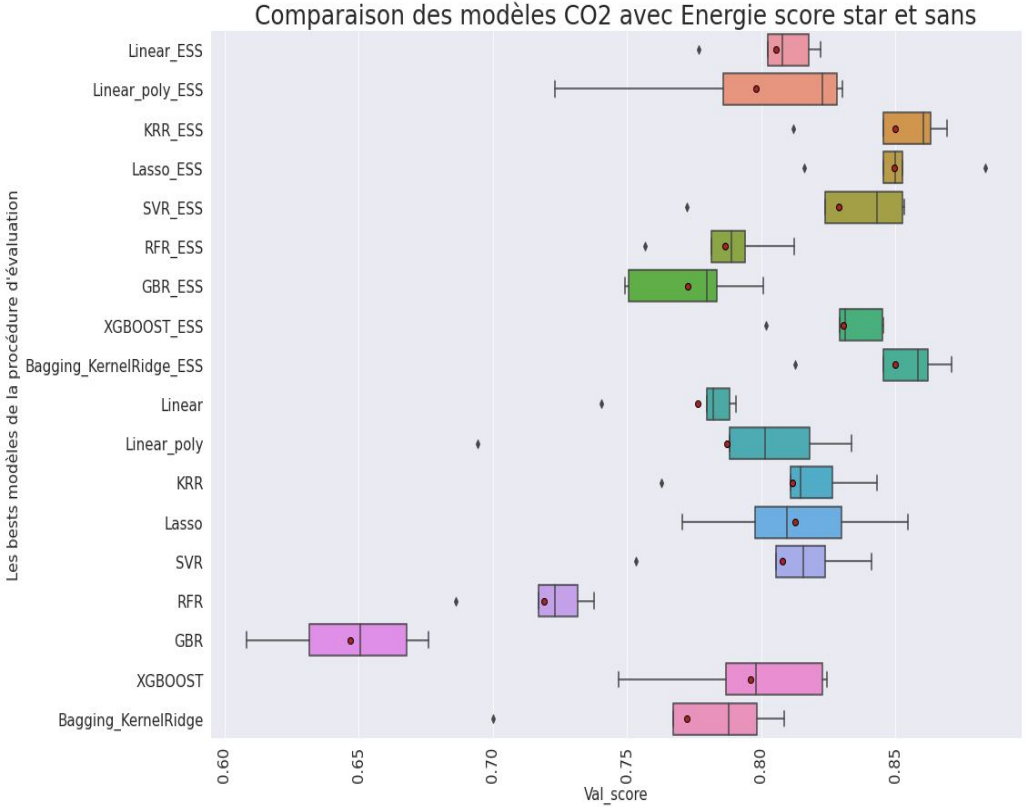

Best score Rcarré: 0.8306

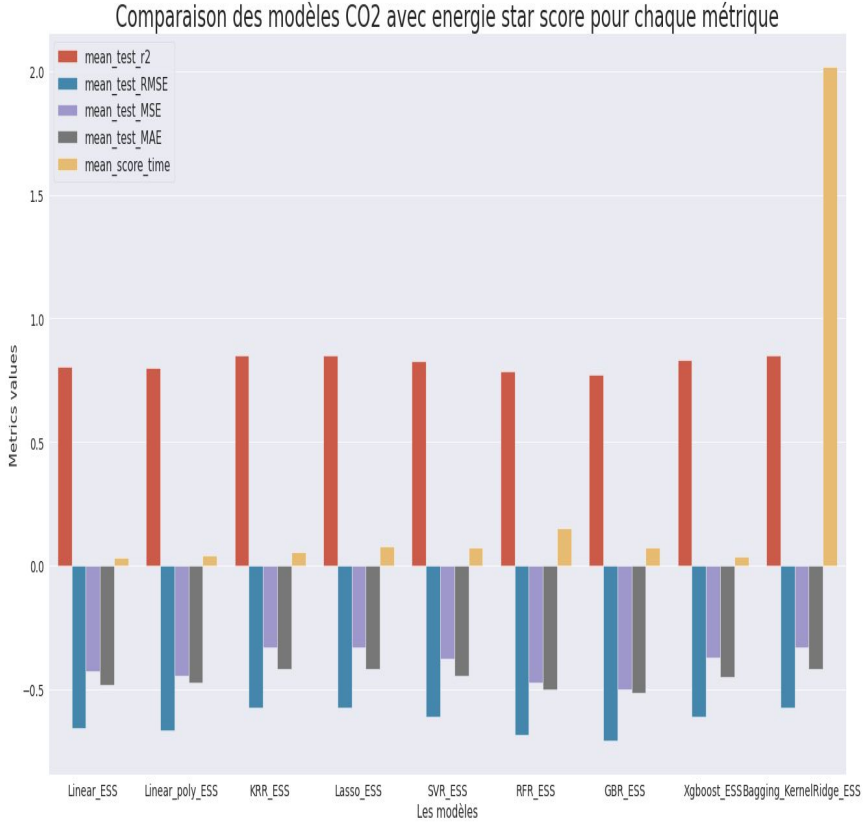


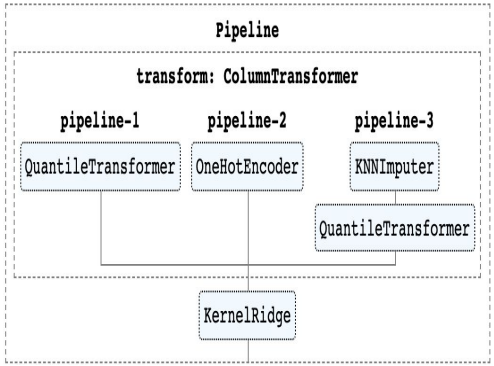
# Comparaison des pipelines avec ESS et sans

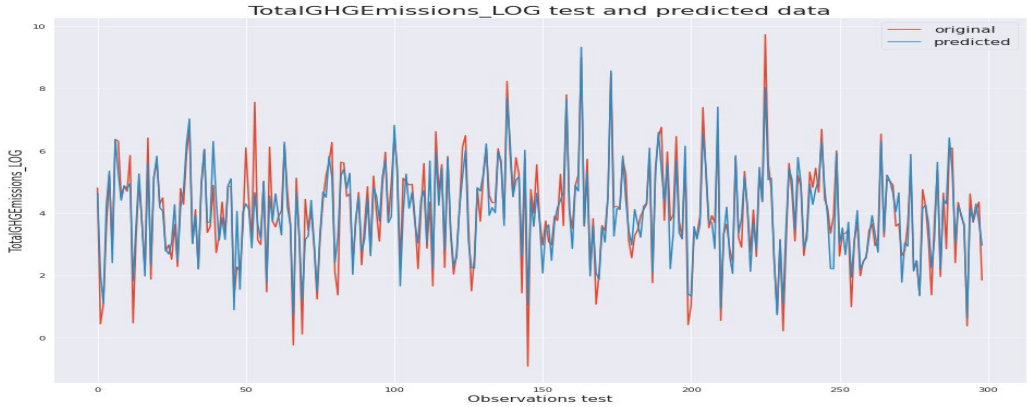


# Le modèle final sélectionné

KERNELRIDGE

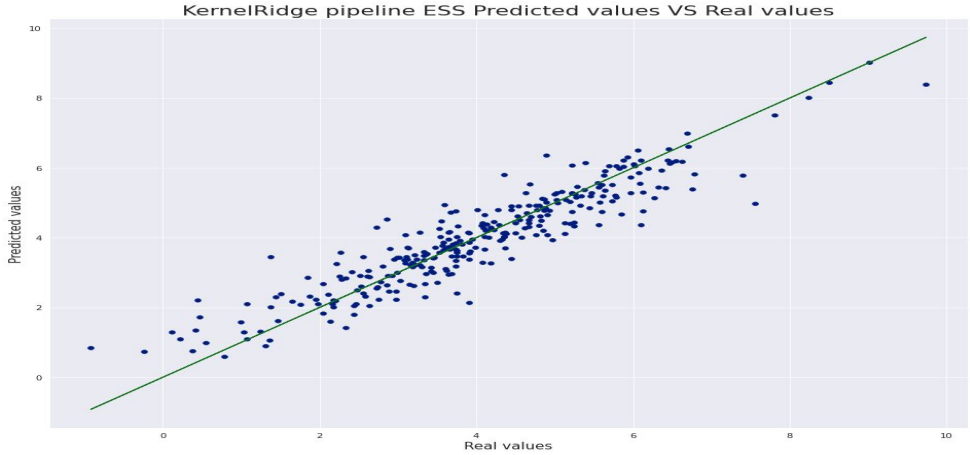


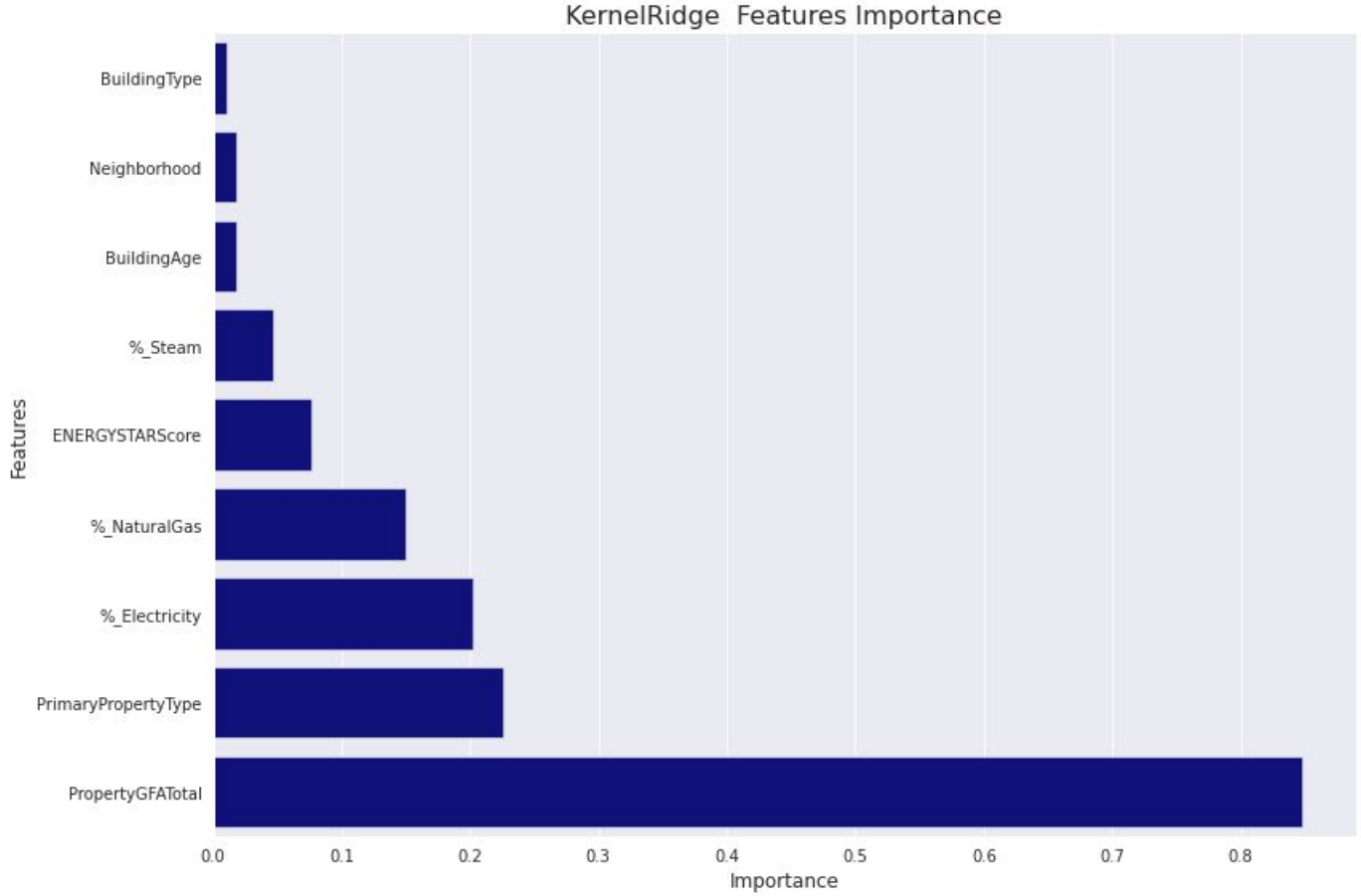

Le score R2 sur le test_set CO2 la pipeline KernelRidge avec Energie star score : 0.8579029162025882

train MAE: 0.3452452265391251
train MSE: 0.23116702729050984
train RMSE: 0.4807983228865403
train R2: 0.8957916689361856

test MAE: 0.3452452265391251
test MSE: 0.3727221389657137
test RMSE: 0.6105097369950079
test R2: 0.8579029162025882



# Interprétation du modèle CO2



# Conclusion

**Prédiction consommation totale d'énergie**

- Les courbes d'apprentissage montrent qu'une augmentation de la data améliorerait les modèles.
- Le modèle choisi  Lasso a un meilleur score sur le test_set(0.7984 (CV) = > 0.8079(Test_set)).

**Prédiction émissions de CO2**

- Les courbes d'apprentissage montrent qu'une augmentation de la data améliorerait les modèles.
- Le modèle choisi KernelRidge  a un meilleur score sur le test_set(0.8500 (CV) = > 0.8579(Test_set)).

**La localité du bâtiment**

- Oui les modèles indiquent que la localité à un pouvoir prédictif malgrés un faible poids.

**Evaluation de Energie Star score**

- La variable améliore les modèles lorsqu'elle est utilisée en input
- C'est une mesure que je recommande à récolter afin d'améliorer les modèles.