OpenClassrooms – Parcours Data Scientist

# Projet 4 :

# Anticipation des besoins en consommation électrique de bâtiments

Fanjamalala Rajaonalison

01/10/2021

# Présentation de l'étude

# Contexte

**Client :** Seattle

2050

neutral

**Mission :** étudier les émissions des bâtiments non destinés à l'habitation.

⇒ Des relevés minutieux des consommations électriques ont été ainsi effectués en 2015 et en 2016.

# Problématique

**Problématique :** Relevés coûteux à obtenir

**Objectifs :** Etude de ces relevés

- Prédiction des émissions de CO2 et de la consommation totale d'énergie de bâtiments pour lesquels elles n'ont pas encore été mesurées.

- Evaluer l'intérêt de l'ENERGY STAR Score pour la prédiction d'émissions, qui est fastidieux à calculer.

**⇒ Problème de régression**

# Méthodologie

**Traitements et Analyses des données**

- **Cleaning**
- **Exploration**
- **Feature engineering**

**Essais de différents modèles**

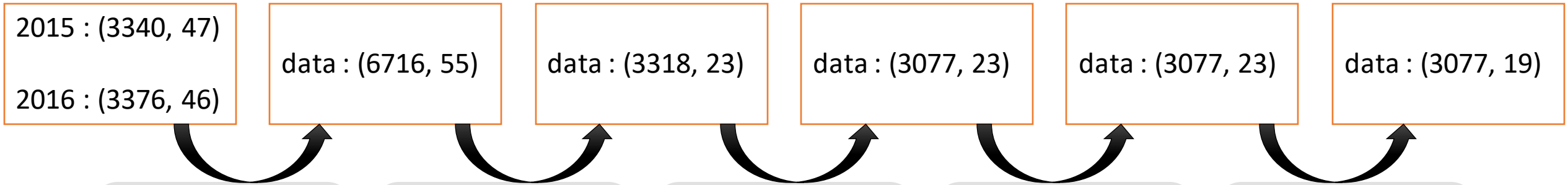
- Simple → Complexe
- Evaluation de la pertinence de chaque modèle (R2, RMSE)
- Comparaison des modèles **(R2, RMSE, temps d'apprentissage, temps de prédiction)**

**Finalisation de l'étude**

- **Sélection du modèle final**
- Prédiction des émissions de CO2 et consommation totale en énergie
- Evaluation de l'importance d'ENERGY STAR SCORE

# Traitements
# Analyses
# des données

# Cleaning




## Mise en forme des données

- Renommage des **variables** communes
- Extraction d'infos à partir de l'adresse : "nom de rue"
- Concaténation des 2 années

## 1ère Selection des données

- LIGNE : Bâtiment non résidentiel 'family' ∉ BuildingType
- COLONNE : - Selon la définition - Selon le taux de remplissage (> 50%)

## Traitement des outliers

- Détection des outliers : |Zscore| < 3
- Suppression des outliers

## Traitements des valeurs manquantes

- Variable catégorielle → "inconnu"
- Variable numérique → **Médiane de la colonne en** filtrant **les** bâtiments du **même** type

## Selection finale des features

- Corrélation avec les autres features < 0,8

# Exploration

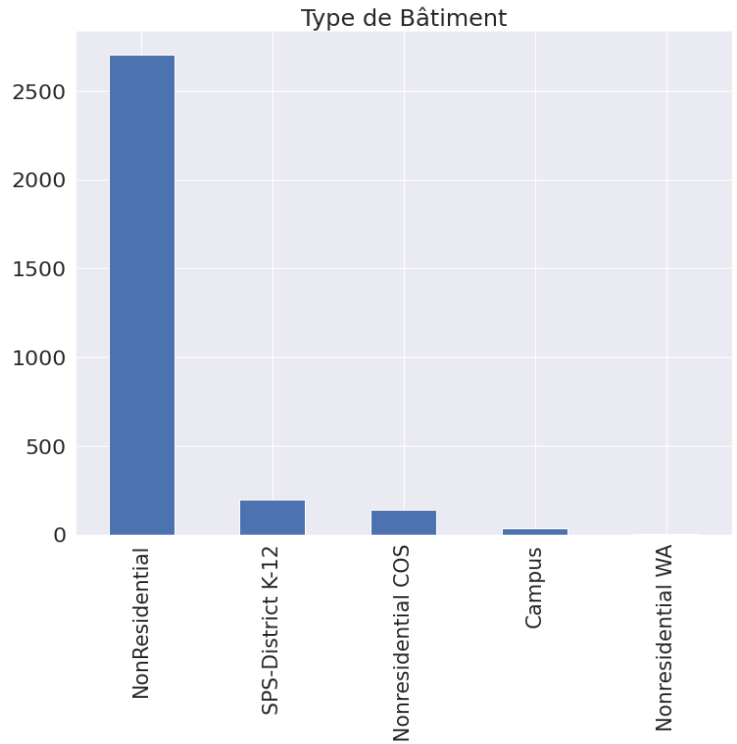
| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | OSEBuildingID | 3077 non-null | int64 |
| 1 | BuildingType | 3077 non-null | object |
| 2 | DataYear | 3077 non-null | int64 |
| 3 | Age | 3077 non-null | float64 |
| 4 | Street | 3077 non-null | object |
| 5 | ZipCode | 3064 non-null | float64 |
| 6 | PrimaryPropertyType | 3077 non-null | object |
| 7 | LargestPropertyUseType | 3077 non-null | object |
| 8 | Neighborhood | 3077 non-null | object |
| 9 | NumberofBuildings | 3077 non-null | float64 |
| 10 | NumberofFloors | 3077 non-null | float64 |
| 11 | PropertyGFAParking | 3077 non-null | int64 |
| 12 | PropertyGFABuilding(s) | 3077 non-null | int64 |
| 13 | ENERGYSTARScore | 3077 non-null | float64 |
| 14 | SiteEUI(kBtu/sf) | 3077 non-null | float64 |
| 15 | SiteEnergyUse(kBtu) | 3077 non-null | float64 |
| 16 | SteamUse(kBtu) | 3077 non-null | float64 |
| 17 | TotalGHGEmissions | 3077 non-null | float64 |
| 18 | GHGEmissionsIntensity | 3077 non-null | float64 |

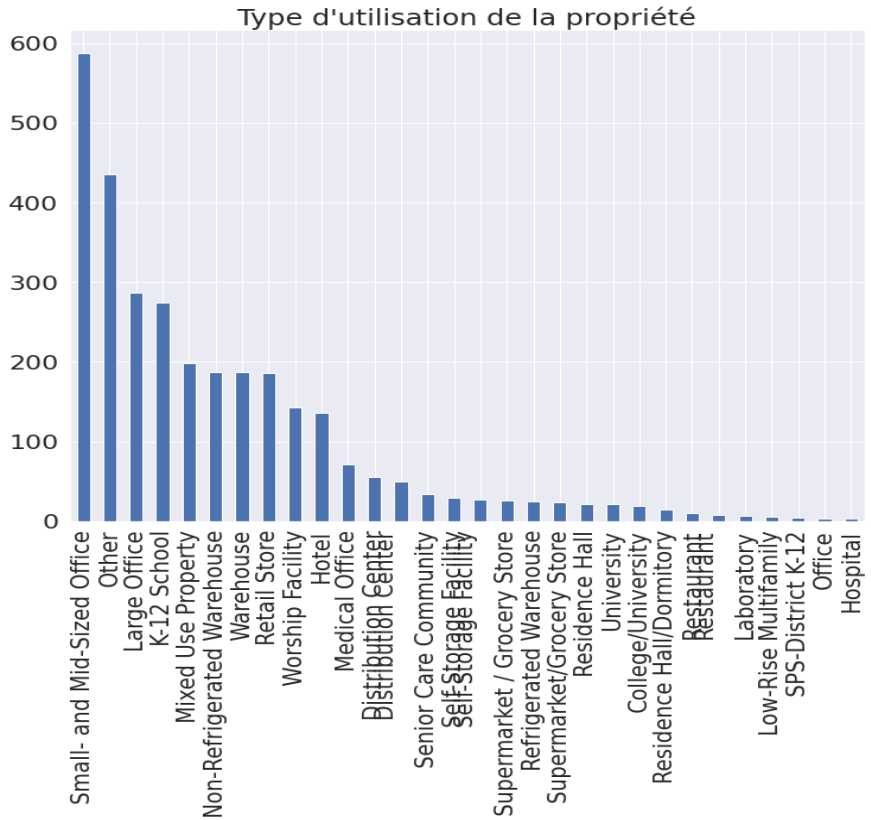
Données : (3077, 19)

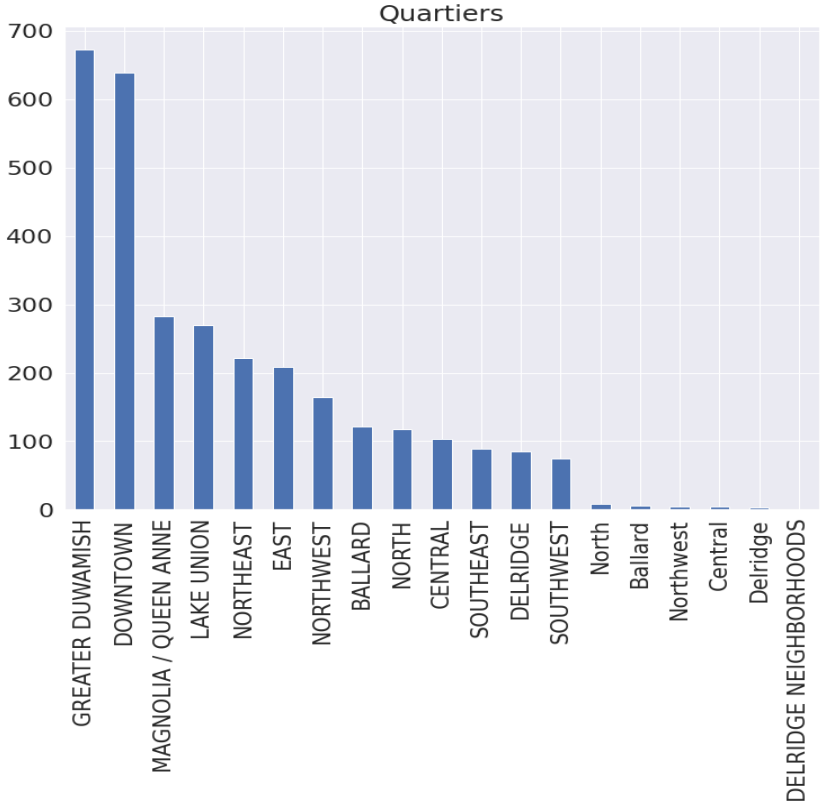
- 5 Variables catégorielles
- 15 Variables numériques
  - 2 Variables cibles

# Exploration

→ Variable catégorielle



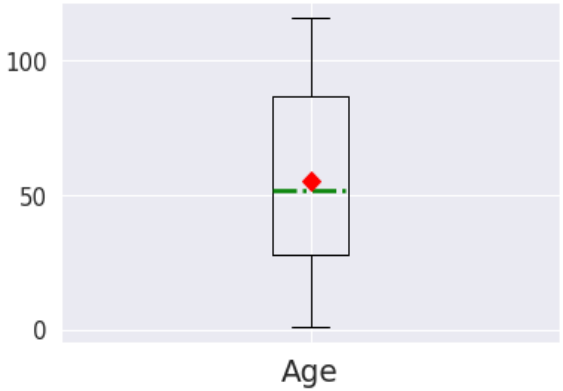

Les bâtiments sont en grand nombre des bâtiments non residentielles.

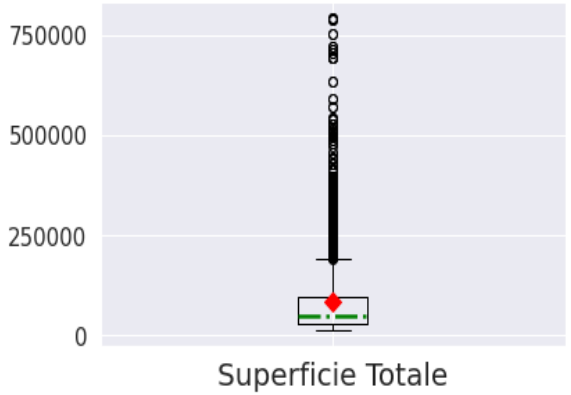
Les 5 premières utilisations sont des bureaux, des écoles et des utilisations mixtes.

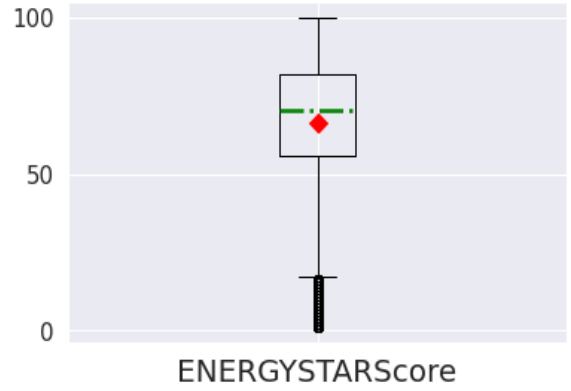
Le plus grand nombre de ces bâtiments sont localisés dans les quartiers : Greater Duwamish et Downtown.

# Exploration

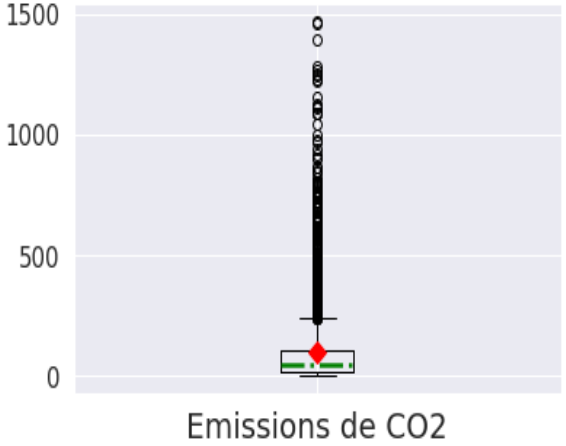
→ Variable numérique

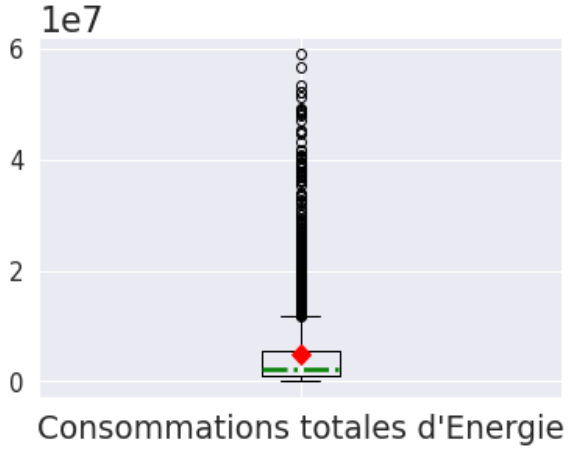



Les Bâtiments sont plutôt âgés (~ 40 ans), avec une superficie totale assez variante mais globalement inférieure à 100000, et un ENERGYSTARScore de 70 en moyenne.

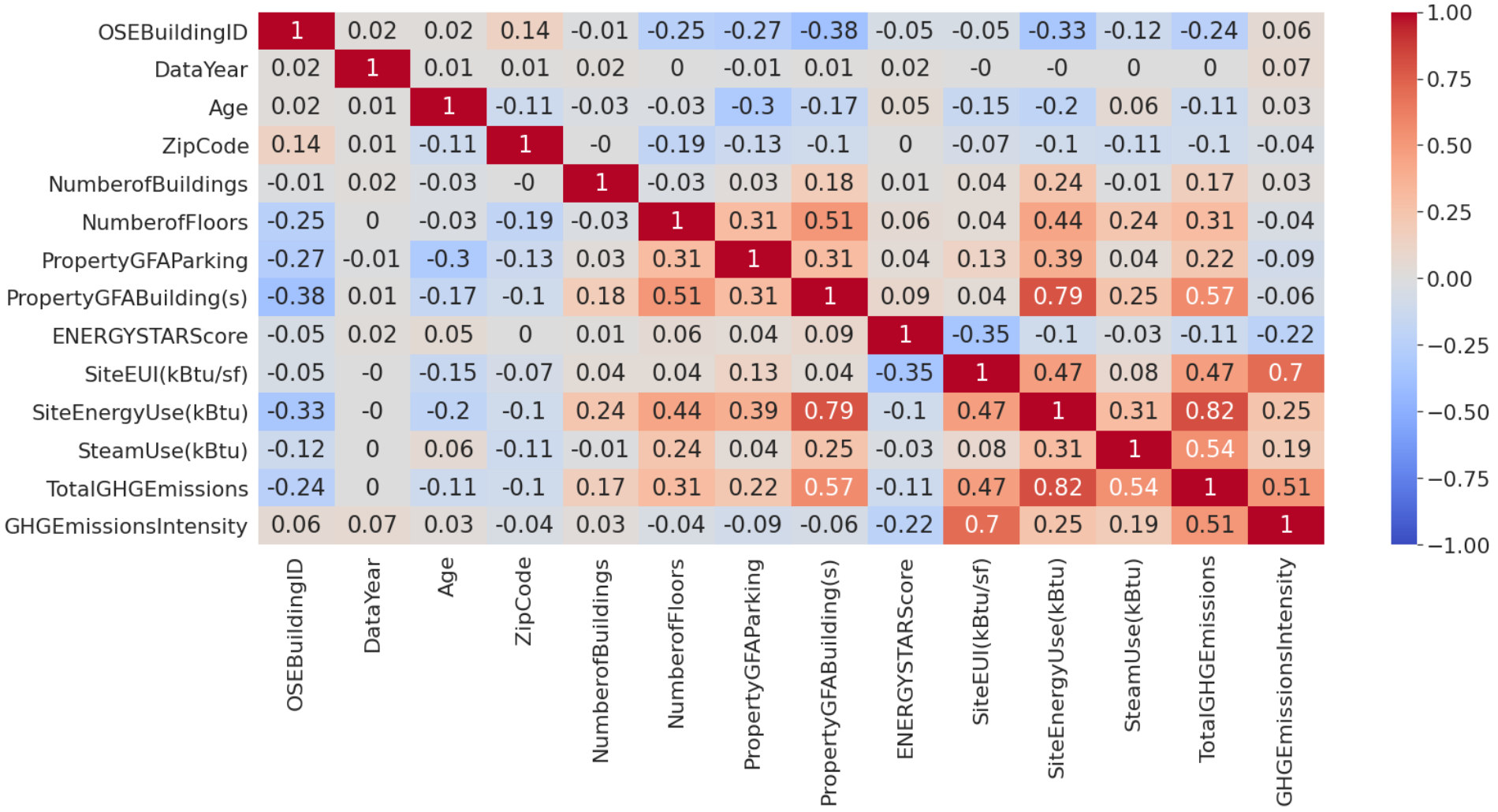



Les émissions de CO2 et consommation d'énergie sont très variants mais dont les moyennes respectives sont de 98 et 4.8M.

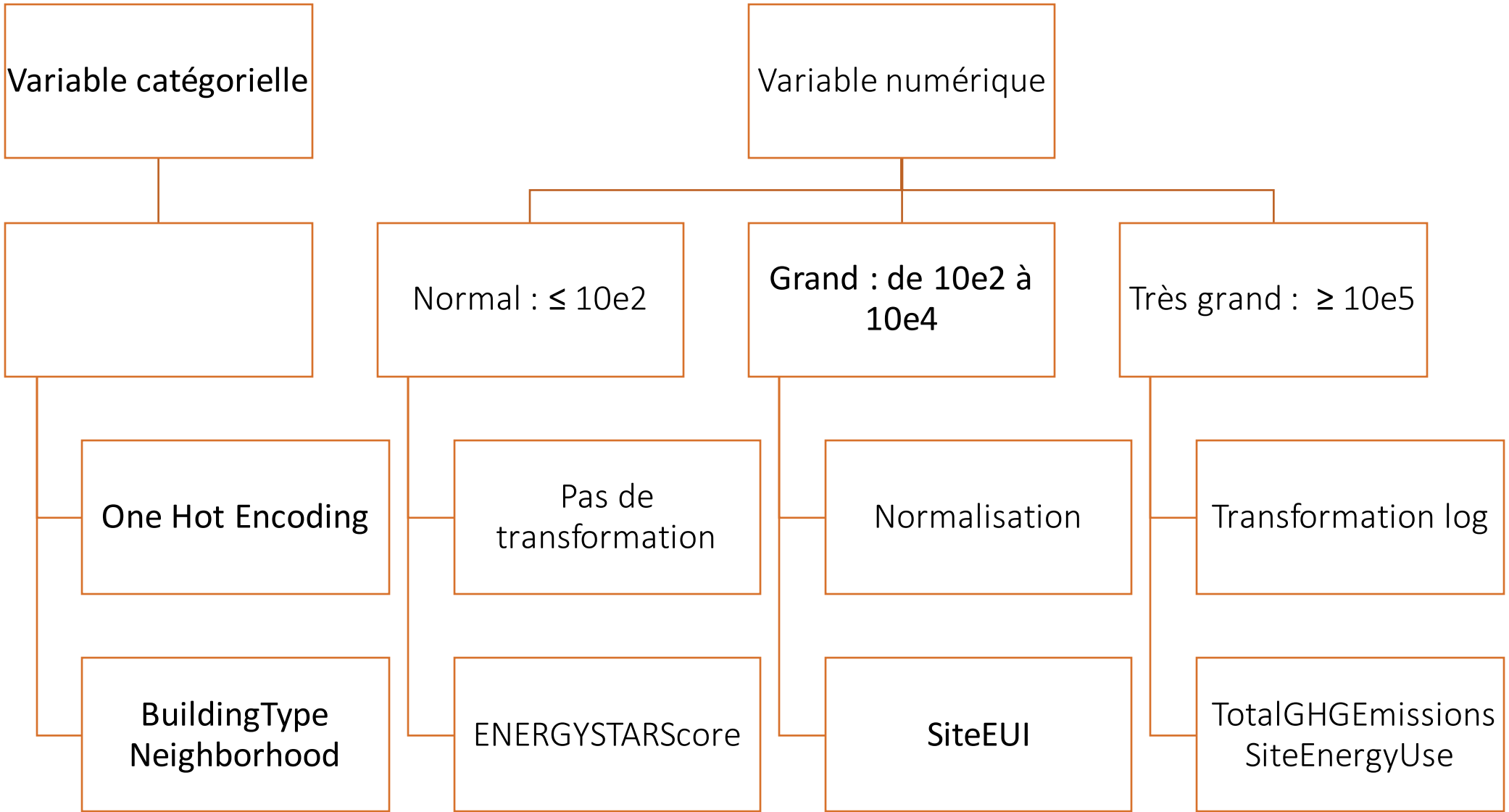
# Exploration

→ Variable numérique




Bien qu'on ait supprimé certaines variables fortement correlées, les variables numériques restantes présentent néanmoins une certaine corrélation.

# Feature Engineering



# Modélisation

# Modèles testés

| Modèles | Linéaire | Ridge | Lasso | Dummy | Random Forest | XGBoost |
|---|---|---|---|---|---|---|
| Type de prédiction | Fonction linéaire | Fonction linéaire avec régularisation de l'amplitude des poids | Fonction linéaire avec régularisation du nombre de variables utilisées | Stratégies simples | Ajustement d'ensemble d'arbres de décision | Ensemble construit à partir de modèles d'arbres de décision mais optimisé |
| Méthode | Moindres carrés | Contrainte quadratique | Contrainte linéaire | Ignore les données d'entrée | Apprentissage d'ensemble | Agrégation de modèles de manière séquentielle |
| Utilisation | Prédiction | Variables correlés | Variables correlés | Modèle de base (Comparaison avec d'autres modèles) | Precision de l'ajustement, Contrôle de l'over fitting | Ajustement et correction des erreurs |

# Paramètres d'évaluation des modèles

| R2 | Error (Mean / Std) | RMSE | Fit time | Prediction time |
|---|---|---|---|---|
| Proportion expliquée de la variance de la variable cible | Différence entre les observations et les prédictions | Erreur **quadratique moyenne entre les observations et les prédictions** | Temps que l'algorithme a mis pour l'apprentissage | **Temps que l'algorithme a mis pour** la prediction |

→ Qualité de la régression

→ Qualité de l'algorithme

# Modelisation des Emissions de CO2

→ Variable cible : **TotalGHGEmissions**

→ Comparaison des modèles testés

| model | r2 | mean error | std error | rmse | fit time | pred time | r2_cv |
|---|---|---|---|---|---|---|---|
| linear | -4.633807e+10 | -16669.205414 | 290376.413869 | 290854.472444 | 0 days 00:00:00.103882 | 0 days 00:00:00.006677 | [-29224941333747.402, 0.8060308539902012, 0.78... |
| ridge | 7.968787e-01 | 0.017684 | 0.596694 | 0.596956 | 0 days 00:00:00.014094 | 0 days 00:00:00.000512 | [0.7444219696445493, 0.6870970906237612, 0.818... |
| lasso | 2.302582e-01 | -0.064158 | 1.174432 | 1.176183 | 0 days 00:00:00.010845 | 0 days 00:00:00.000571 | [0.2899435681204498 6, 0.2323815971620799 6, 0.2... |
| dummy | -1.499404e-03 | -0.050547 | 1.305374 | 1.306352 | 0 days 00:00:00.000453 | 0 days 00:00:00.000099 | [-0.002846875977950 62, -0.002945695650296 2317... |
| randomforest | 9.806254e-01 | -0.007276 | 0.185636 | 0.185778 | 0 days 00:00:03.193553 | 0 days 00:00:00.038594 | [0.9758453137296385, 0.9843439407115906, 0.982... |
| XGBoost | 9.817005e-01 | -0.005098 | 0.180795 | 0.180867 | 0 days 00:00:22.594098 | 0 days 00:00:00.016033 | [0.9763055316694904, 0.9763437127902298, 0.960... |

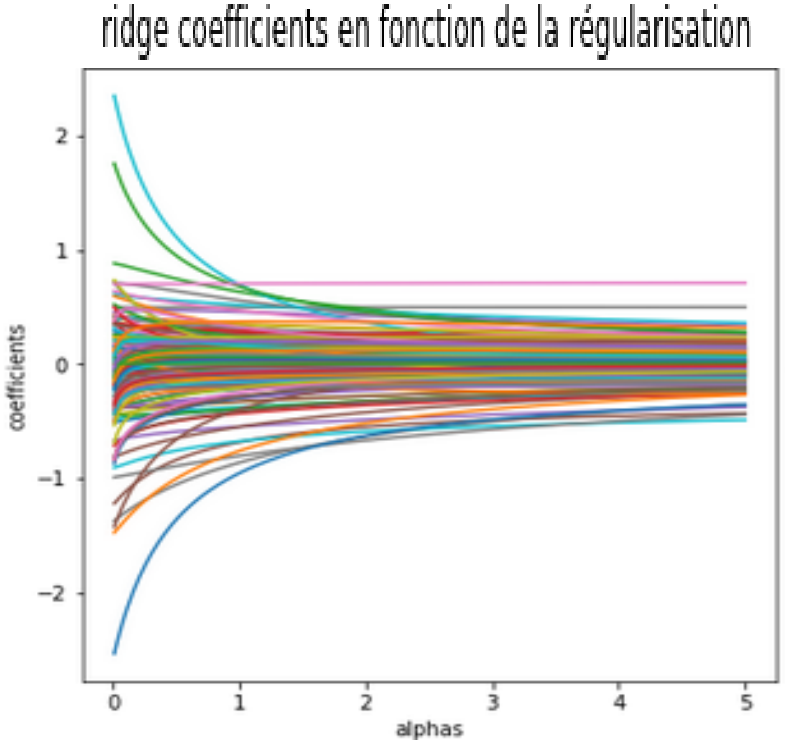
R2 non conforme

R2 faible

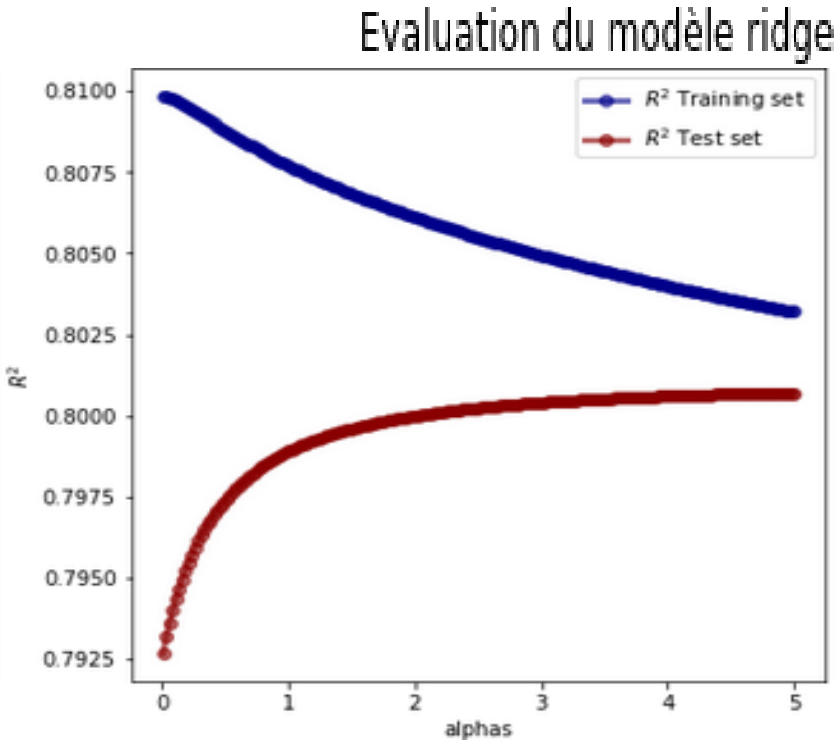
R2 très faible

# Modelisation des Emissions de CO2

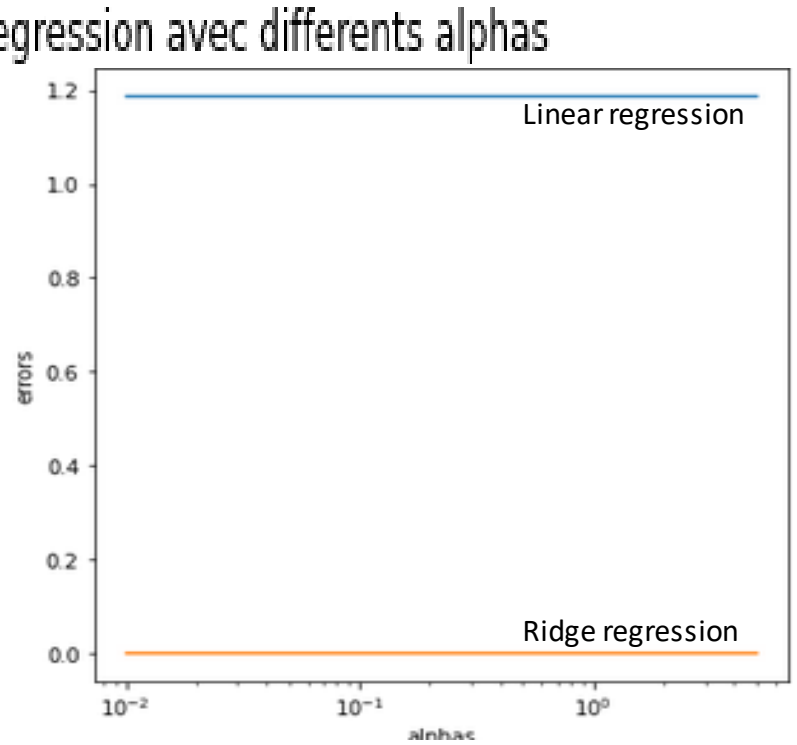
→ Régularisation du modèle Ridge




Alpha diminue les poids des paramètres de la régression jusqu'à une stabilisation




- R2 training set a une tendance décroissante
- R2 test set a une tendance croissante

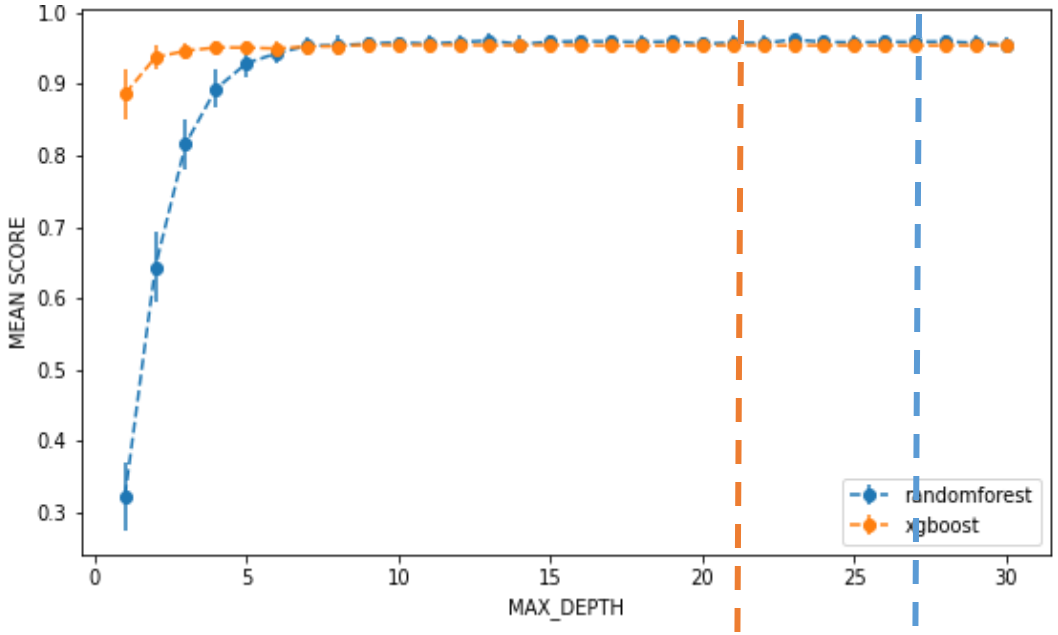


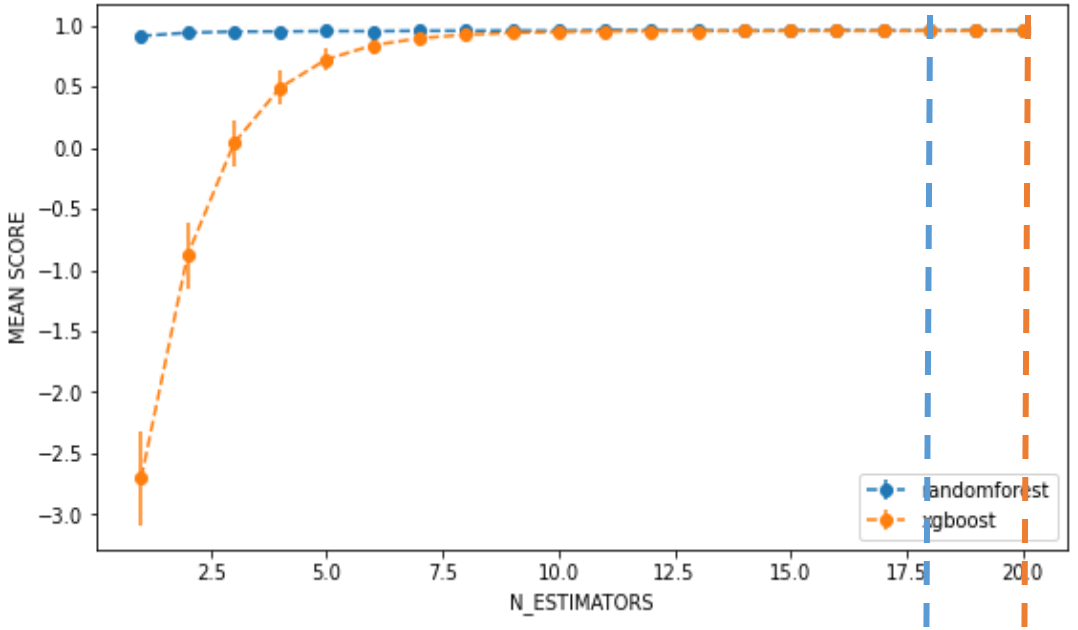

Erreur faible

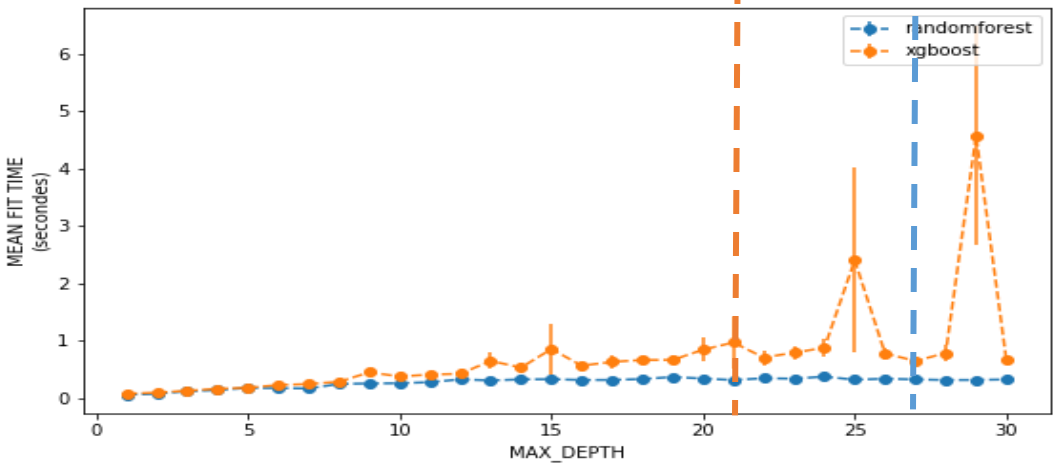
→ R2 plus faible que pour RF et XG

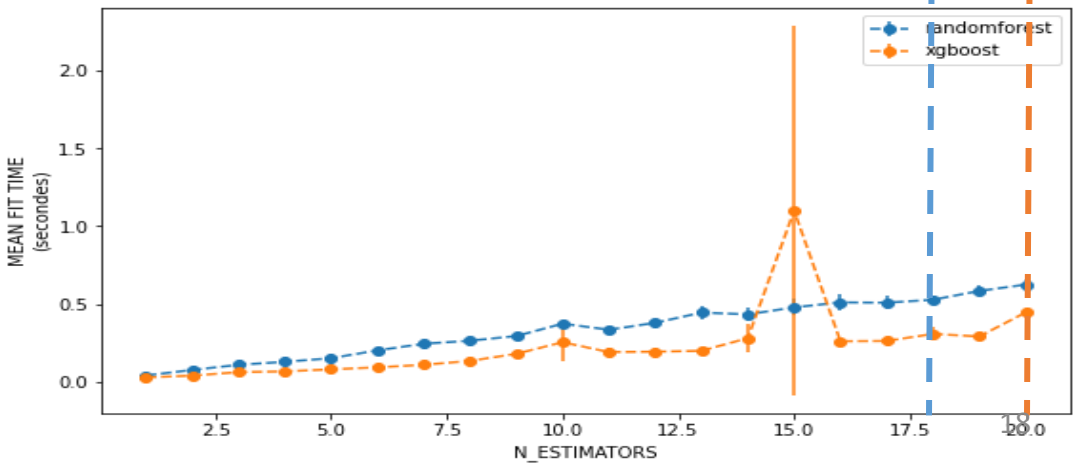
# Modelisation des Emissions de CO2

→ Optimisation de paramètres pour random forest et xgboost : **Max_depth & N_estimator**



# Modelisation des Emissions de CO2

→ Evaluation des modèles random forest et xgboost optimisés

| model | r2 | mean error | std error | rmse | fit time | pred time | r2_cv |
|---|---|---|---|---|---|---|---|
| randomforest | 0.920893 | 0.007233 | 0.362379 | 0.362452 | 0 days 00:00:01.102934 | 0 days 00:00:00.009551 | [0.9838605641225466, 0.9773075795078823, 0.976... |
| XGBoost | 0.982240 | -0.019180 | 0.160848 | 0.161987 | 0 days 00:00:00.853205 | 0 days 00:00:00.005448 | [0.9822076439393237, 0.9359328323096218, 0.981... |

⇒ **XGBoost meilleur**

# Modelisation de la Consommation d'énergie

→ Variable cible : **SiteEnergyUse**

→ Comparaison des modèles testés

| model | r2 | mean error | std error | rmse | fit time | pred time | r2_cv |
|---|---|---|---|---|---|---|---|
| linear | -1.089801e+09 | -2743.170011 | 59022.439439 | 59086.151839 | 0 days 00:00:00.043161 | 0 days 00:00:00.015698 | [0.687174024558933, -251194852581.39627, 0.43... |
| ridge | 4.777843e-01 | -0.053327 | 1.086438 | 1.087746 | 0 days 00:00:00.017793 | 0 days 00:00:00.001481 | [0.4383451446723134, 0.4098225770960781, 0.5... |
| lasso | 1.543760e-01 | 0.007837 | 1.555125 | 1.555145 | 0 days 00:00:00.015785 | 0 days 00:00:00.007080 | [0.3022614996718181, 0.040965781518480404, 0.... |
| dummy | -1.182469e-03 | -0.044402 | 1.291252 | 1.292015 | 0 days 00:00:00.000533 | 0 days 00:00:00.000103 | [-0.0023429349213355266, -0.001605872824299359... |
| randomforest | 9.898144e-01 | 0.000383 | 0.173552 | 0.173552 | 0 days 00:00:03.178424 | 0 days 00:00:00.033756 | [0.9821910088565635, 0.9822443672393822, 0.984... |
| XGBoost | 9.414576e-01 | -0.013676 | 0.350379 | 0.350646 | 0 days 00:00:01.090498 | 0 days 00:00:00.006045 | [0.9804390320201267, 0.9870484222395715, 0.987... |

R2 non conforme

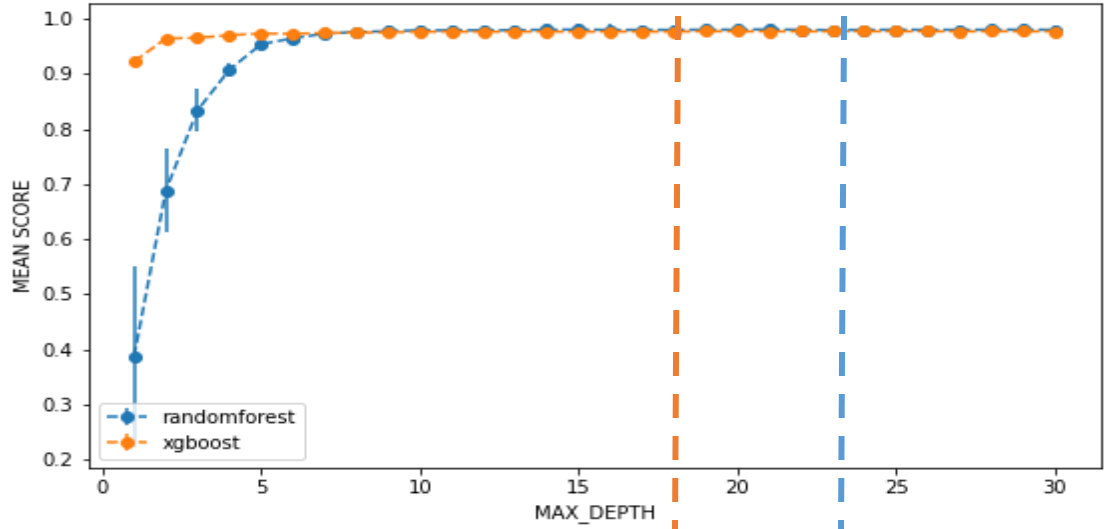
R2 moyen

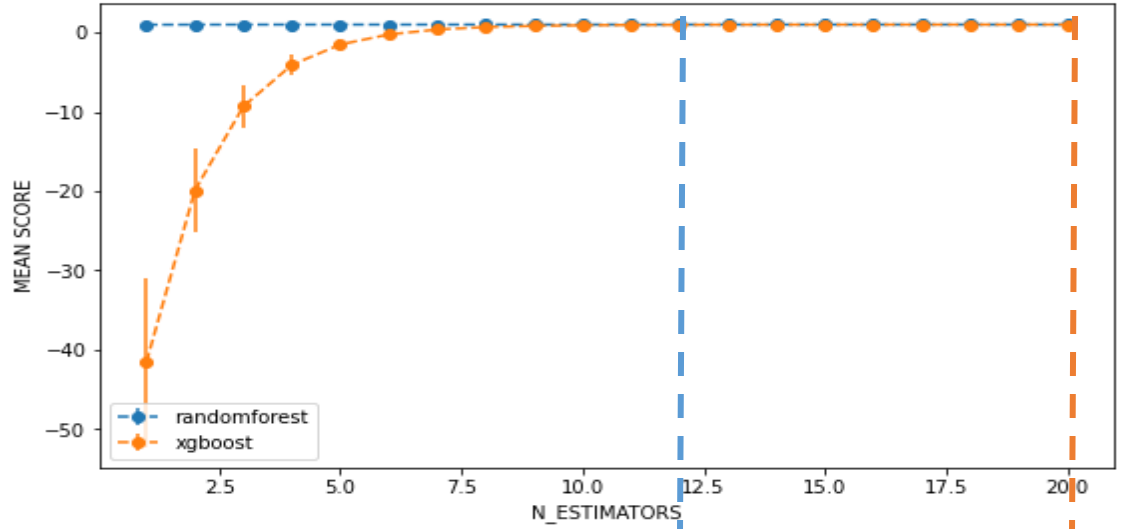
R2 très faible

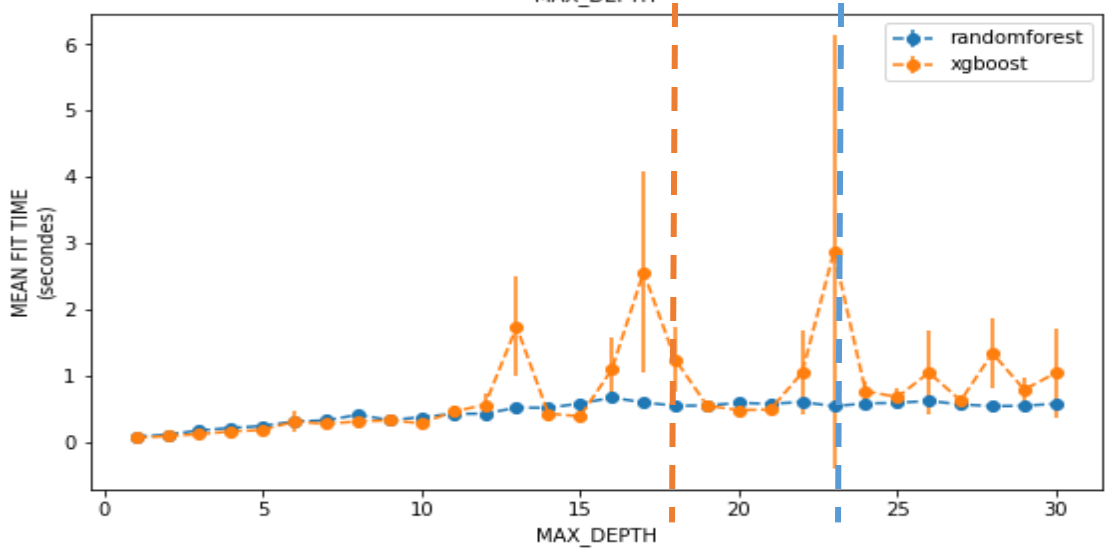
R2 très faible

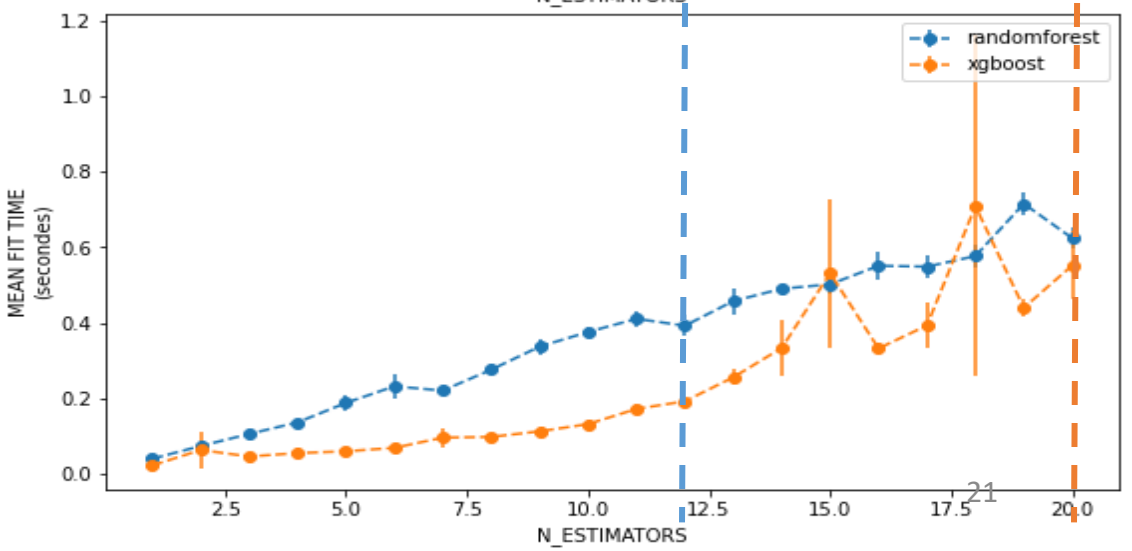
# Modelisation de la Consommation d'énergie

→ Optimisation de paramètres pour random forest et xgboost : **Max_depth** & **N_estimator**



# Modelisation dela Consommation d'énergie

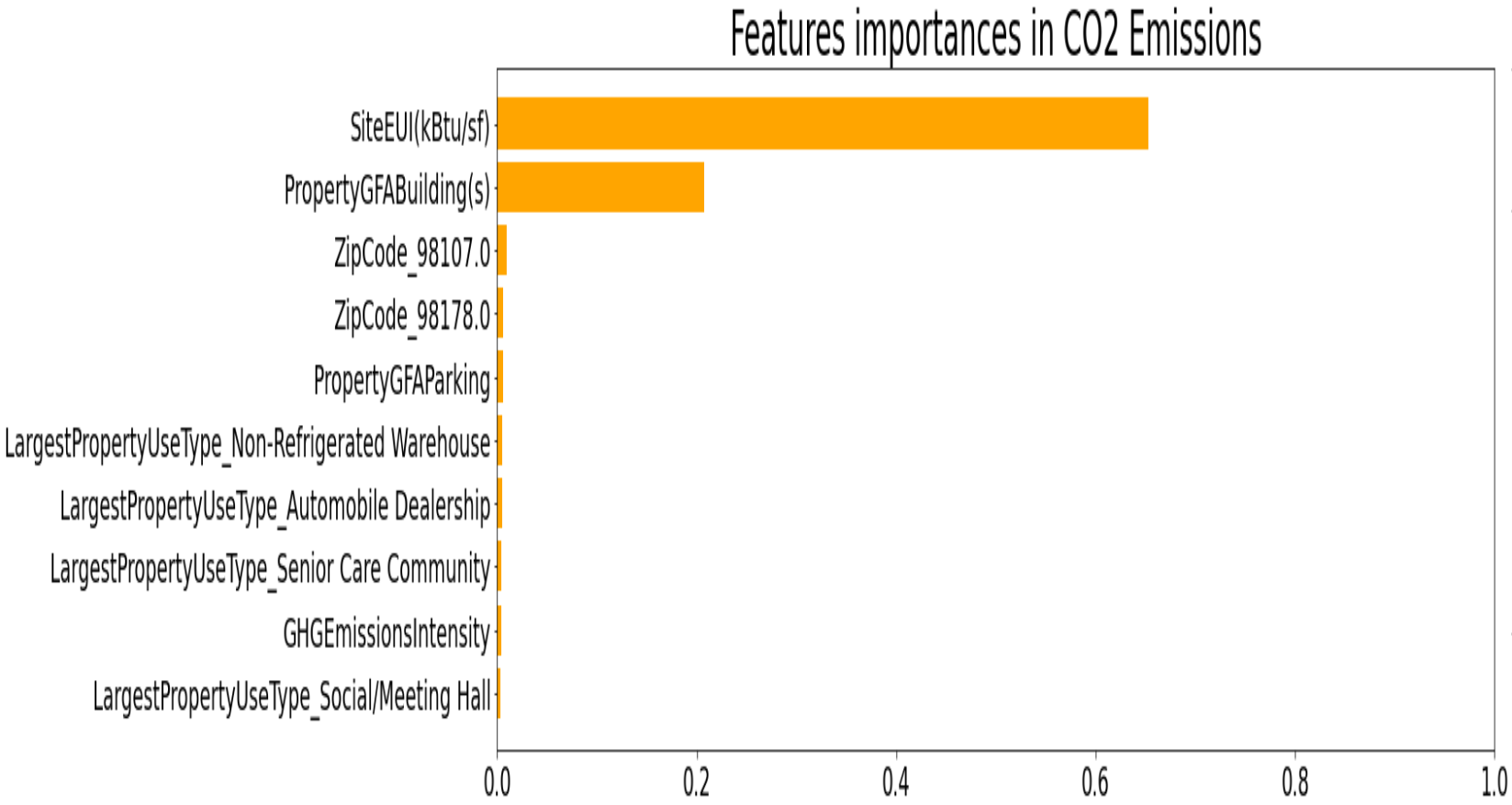
→ Optimisation de paramètres pour random forest et xgboost

| model | r2 | mean error | std error | rmse | fit time | pred time | r2_cv |
|---|---|---|---|---|---|---|---|
| randomforest | 0.973724 | -0.001537 | 0.230648 | 0.230653 | 0 days 00:00:00.902245 | 0 days 00:00:00.011565 | [0.786819083734008, 0.939314668510721, 0.97590... |
| XGBoost | 0.983897 | -0.024431 | 0.164917 | 0.166716 | 0 days 00:00:02.167755 | 0 days 00:00:00.005814 | [0.9658100342291462, 0.93614107551718, 0.97255... |

**⇒ XGBoost meilleur**

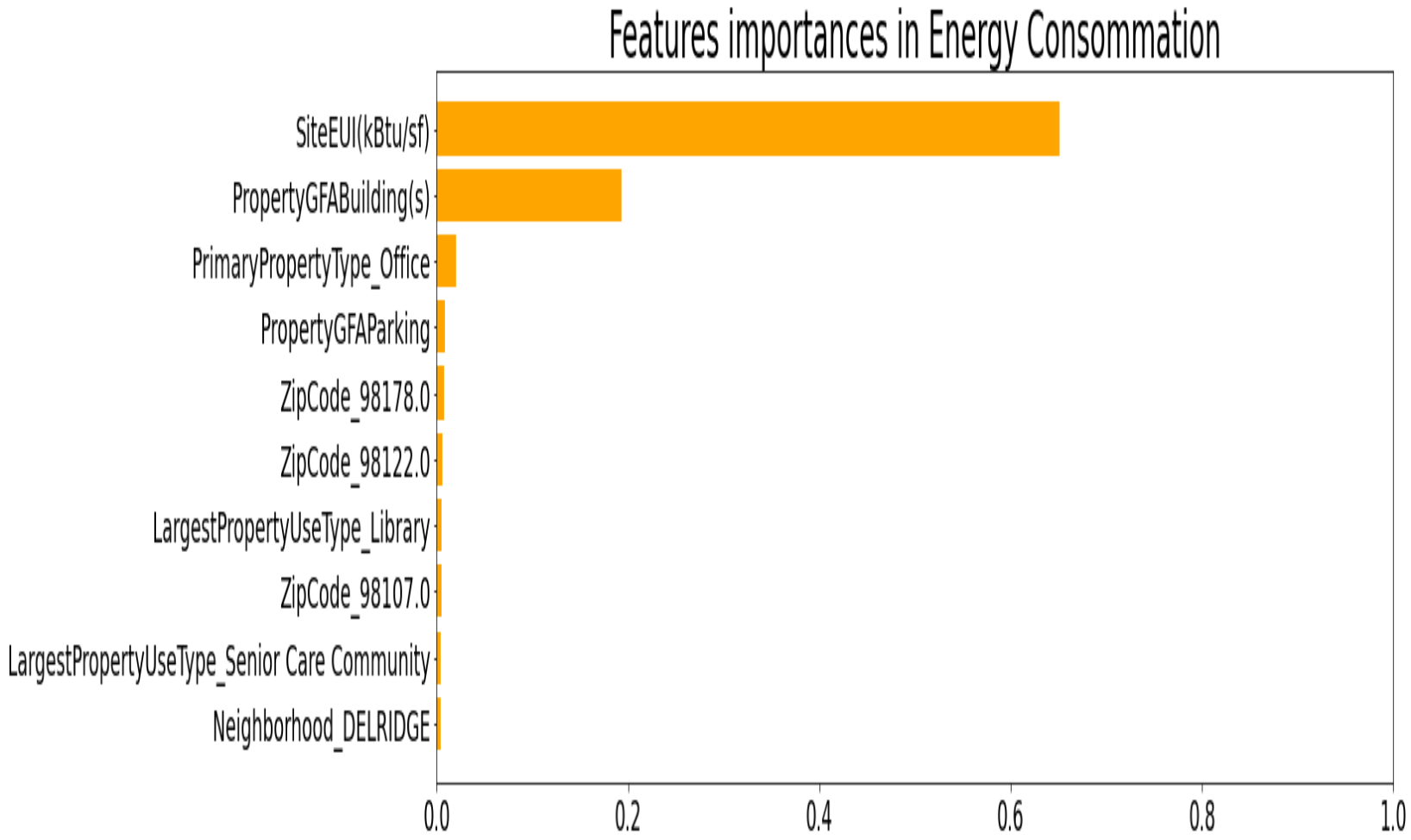
# Résultats

# Prédiction de les émissions de CO2




- Modèle XGBoost
- ~80% variance des émissions de CO2 est expliquée l'intensité d'utilisation de l'énergie du site et de la surperficie des bâtiment
- L'importance de l'ENERGYSTARScore dans la prédiction est minime : ~0.12%

# Prédiction de la consommation d'énergie




- Modèle XGBoost
- ~80% variance de la consommation totale d'énergie est expliquée par l'intensité d'utilisation de l'énergie du site et de la surperficie des bâtiment
- L'importance de l'ENERGYSTARScore dans la prédiction est minime : ~0.07%

# Améliorations

## Cleaning

- Réduire la corrélation des variables
- Ne garder qu'une donnée par bâtiment (~95% des bâtiments ont des données en 2015 et en 2016)

## Feature engineering

- Essayer de regrouper les variables catégorielles avant la transformation One Hot Encoding

## Optimisation

- Choisir des nombres d'estimateurs plus grands
- Essayer d'autres paramètres

# MERCI

Questions

Incompréhensions