Présentation projet 4: Anticiper les besoins en consommation électrique de bâtiments

Parcours Data Scientist : Bagué Pierre

Sommaire

1. Présentation des données

Provenance et description des données

- Données de la ville de Seattle
- Deux années de bilan énergétiques
- Plus de 3300 bâtiments
- Plus de 40 variables : type des bâtiments, surface, lieu, année de construction, nombre d’étages, consommation en tout genre,...

2. Nettoyage du jeu de données

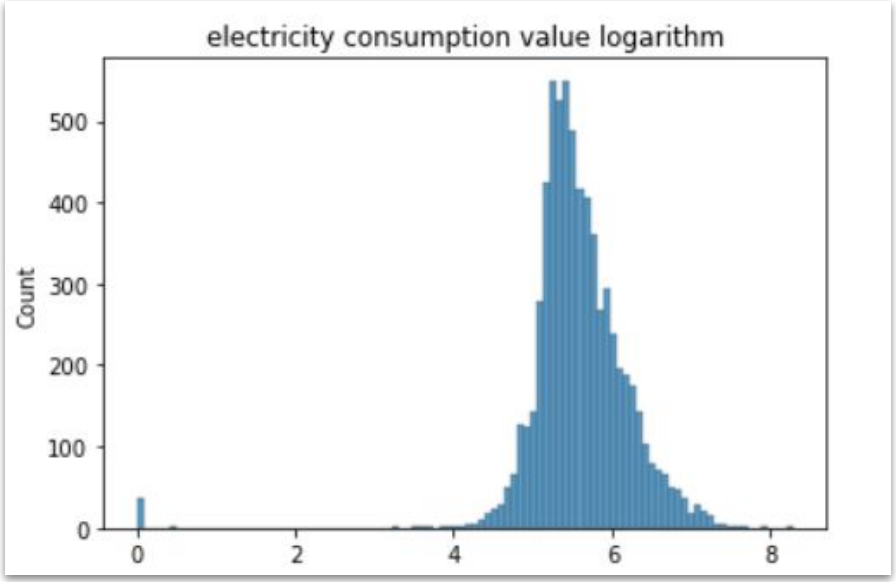
Visualisation et nettoyage des données

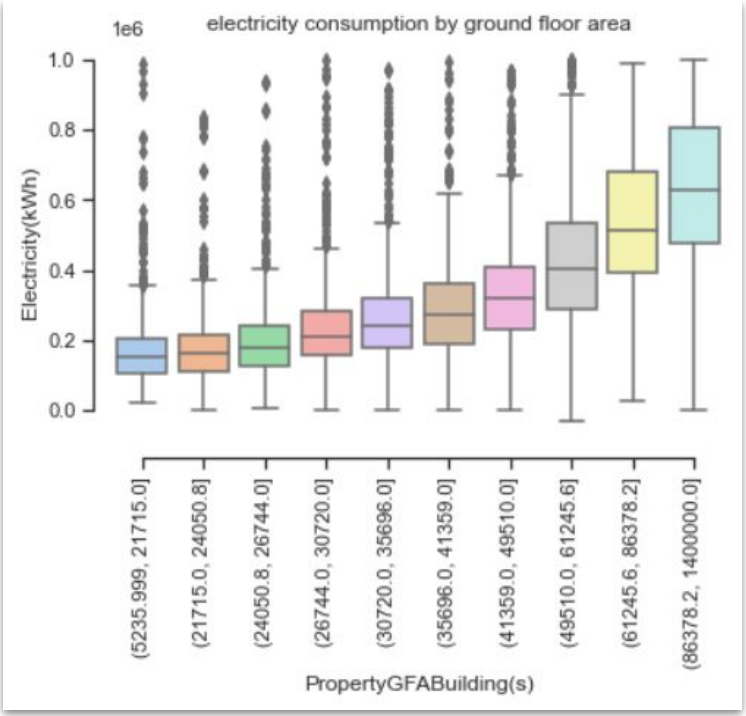
- Deux fichiers avec des différences de variables:
  - noms différents :
    - changements de noms
  - variable rassemblée ou divisée d’une année à l’autre :
    - coordonnées gps
  - variables enlevées ou ajoutées d’une année à l’autre :
    - données parcellaires
- concaténation des fichiers

3. Préparation des variables cibles pour le k-fold

Préparation des valeurs cibles

- Regard sur la répartition des valeurs :





Préparation des variables cibles

- pour l’électricité et l’émission de gaz à effet de serre:
  - mise à 0 pour les valeurs non définies (consommation électrique ne rejetant pas de gaz)
  - opposition des valeurs négatives
  - Création d’une seconde cible avec le logarithme en base 10 pour l’électricité

k-fold

- 5 folds prévus
- création de 5 index de données pour les entraînements et les tests des algorithmes
- répartition équitable par intervalle de valeur des données cibles

4. Preprocessing

4.1 Les variables numériques

- Sélection des variables numériques dans le dataset
- Corrélation de pearson entre les variables numériques
- Suppression des variables à corrélation supérieur à 99% (duplicatas d'une autre unité de mesure)
- Sélection des variables avec une bonne corrélation

	SiteEUI(kBtu/sf)	SiteEUIWN(kBtu/sf)
SiteEUIWN(kBtu/sf)	1.000000	0.994560
SourceEUI(kBtu/sf)	0.994560	1.000000

- Suppression des valeurs aberrantes
- Remplissage de variables en fonction du contexte

4.2 variables numériques + non numériques

- Sélection des variables autre que numériques dans le dataset

LargestPropertyUseType	ThirdLargestPropertyUseType
Hotel	NaN
Hotel	Restaurant

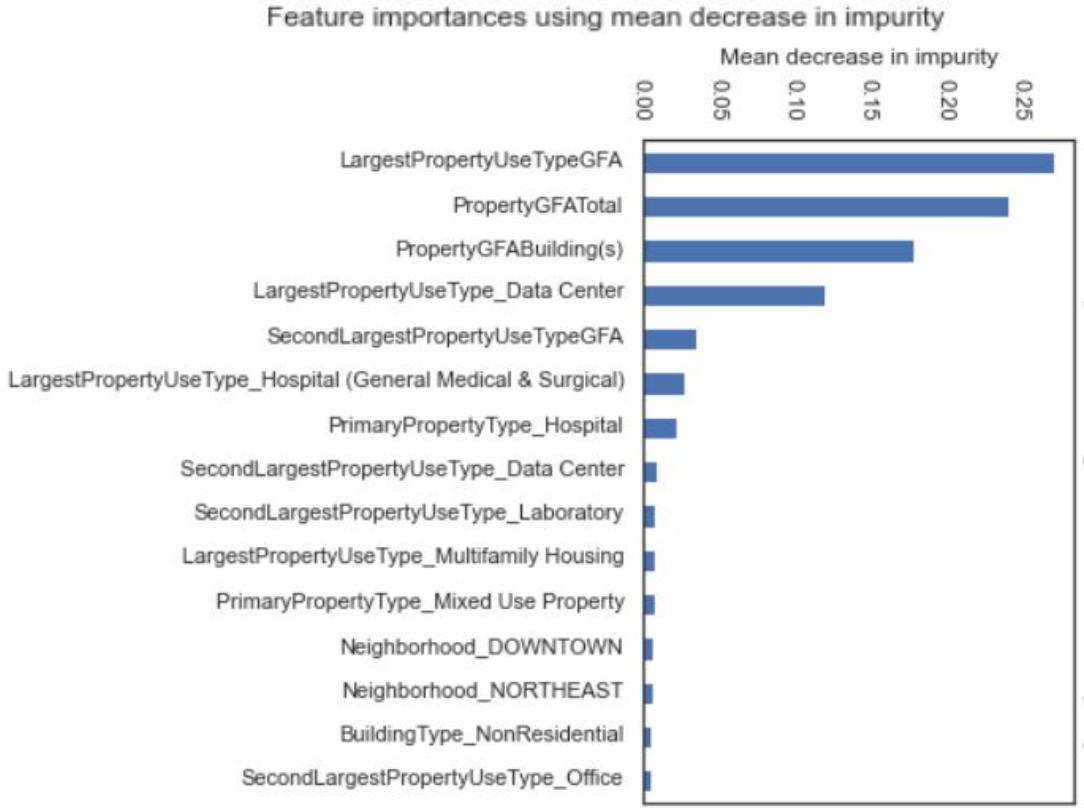
- Transformation des variables avec one hot encoding
- Concaténation avec les variables numériques

4.3 Variables numérique à forte corrélation

- Sélection des variables numériques à forte corrélation avec la cible (électricité ou émission de gaz à effet de serre):
  - surface du plus gros bâtiment
  - surface des bâtiments
  - surface de la propriété
  - surface du deuxième bâtiment le plus gros
  - surface du troisième bâtiment le plus gros

4.4 Variables numériques à forte corrélation + non numériques les plus utilisées

- Sélection des variables non numériques les plus utilisées
  - deux algorithmes utilisés
    - mesure de permutation des variables
    - diminution des impuretées
  - intersection des deux



5. Les algorithmes utilisés

5.1 Diversité du benchmark

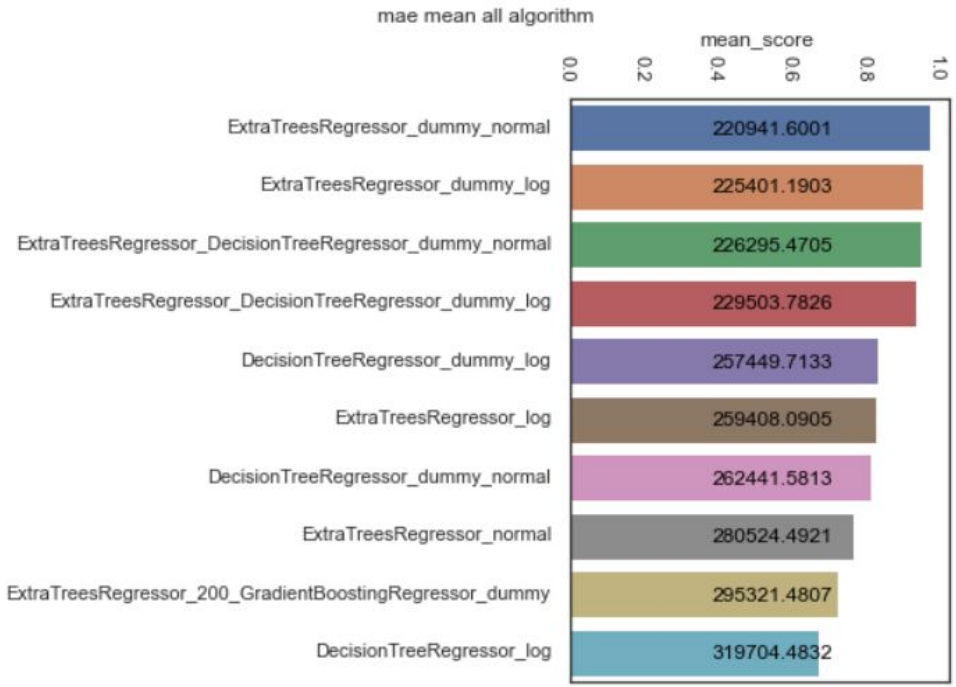
- 1 Algorithme étalon :
  - Dummy Regressor
    - moyenne
    - médiane
- 13 Algorithmes simples :
  - Linear Regressor
  - Elastic Net CV
  - Elastic Net
  - Ridge
  - Lasso
  - MLP Regressor
  - Support Vector Regression
  - Random Forest Regressor
  - K Neighbors Regressor
  - Decision Tree Regressor
  - Extratrees Regressor
  - Ada Boost Regressor
  - Gradient Boosting Regressor
- 1 Algorithme conjugué:
  - Voting Regressor
    - Extratrees
    - Decision Tree

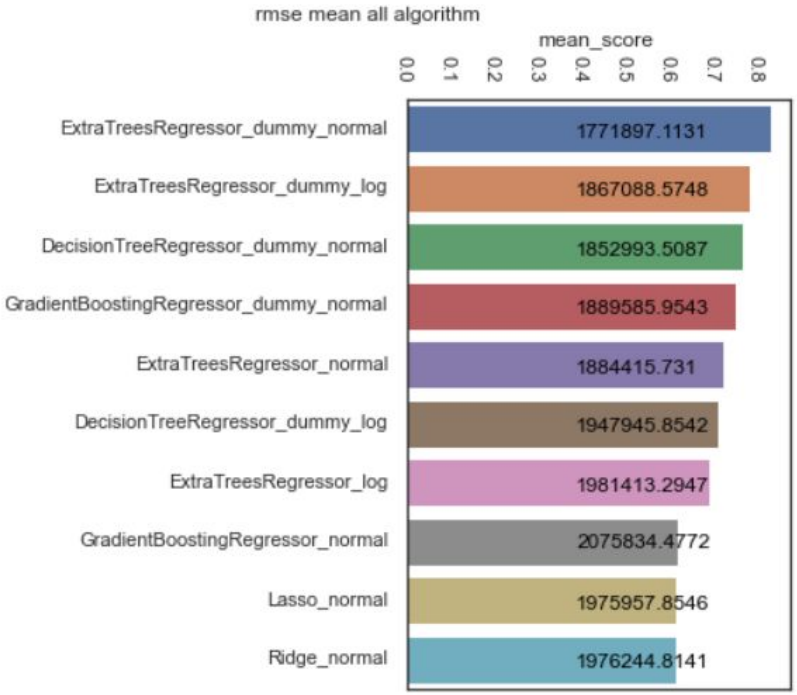
5.2 Mesure des résultats

- Interface générique
- Deux mesures utilisées:
  - mae : mean absolute error
  - rmse : root mean squared error
- pourcentage final :
  - disparité entre les folds :
    - normalisation du résultat entre 0 et 1 (1 étant le meilleur)
    - moyenne sur les folds

5.3 Résultats pour la consommation d’électricité

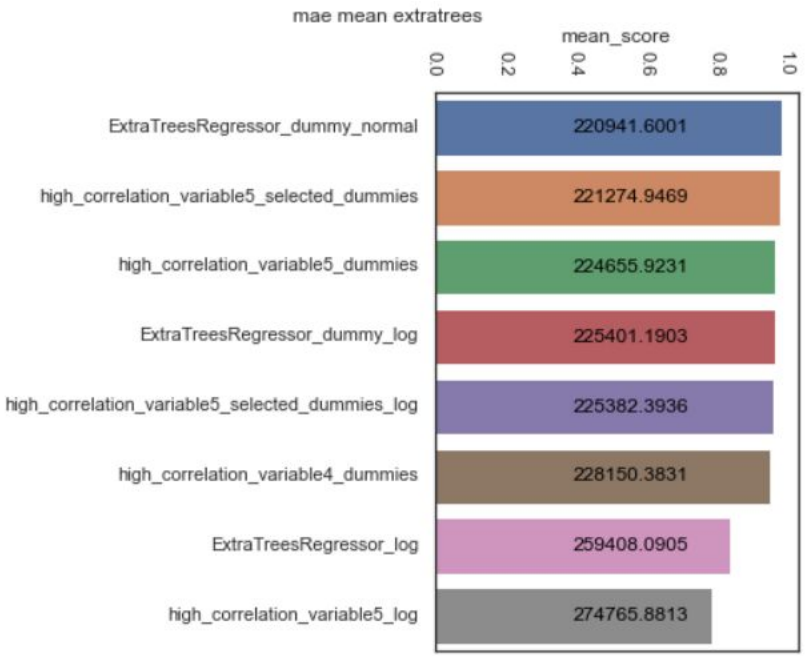
5.3.1 Résultats pour la consommation d'électricité

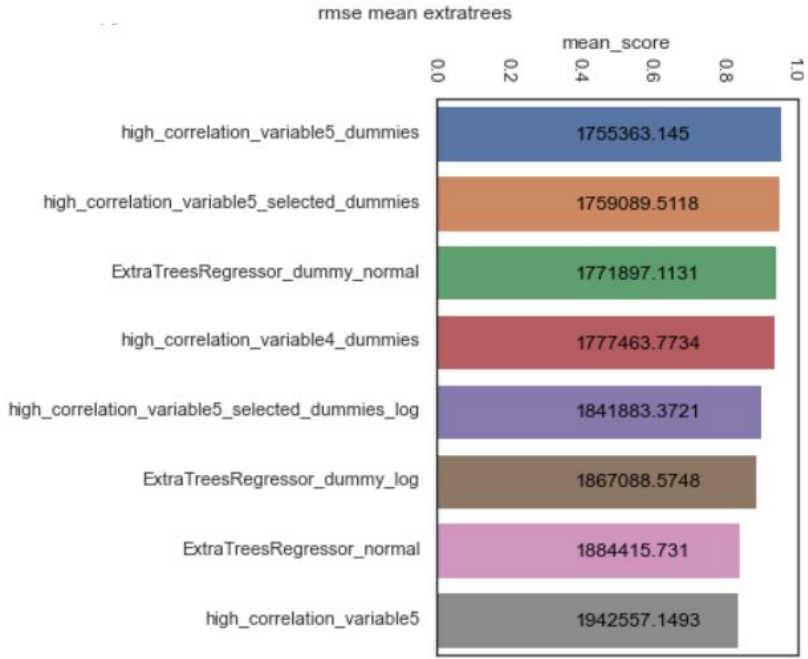




Dataset mean : 1050000, Dummy regressor : mae=850000,rmse=3439000

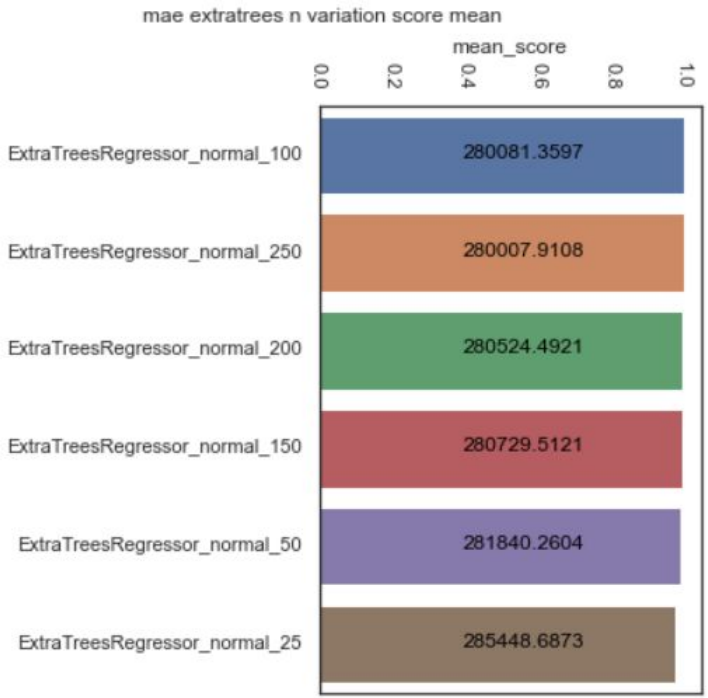
5.3.2 Affinage des résultats : variables à haute corrélation

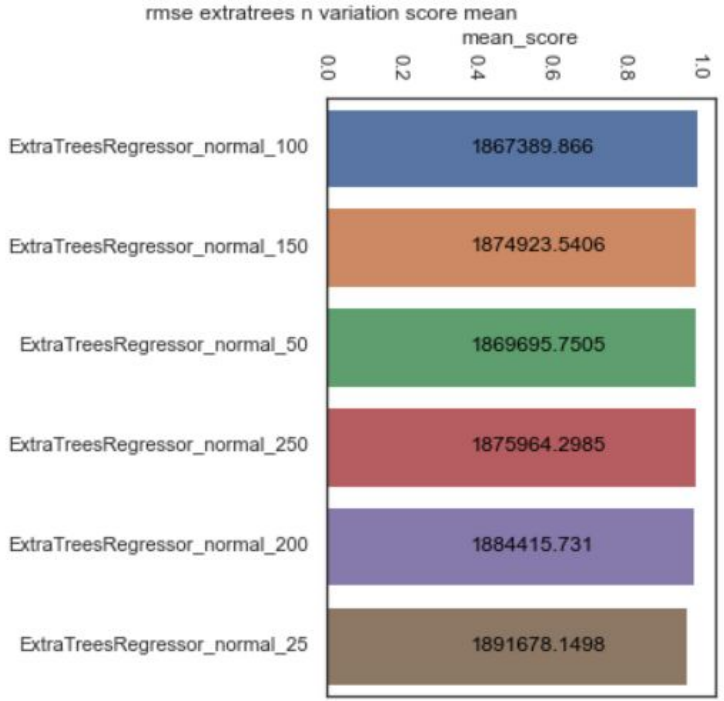




Dataset mean : 1050000, Dummy regressor : mae=850000,rmse=3439000

5.3.3 Affinage des résultats : paramètre n estimators





Dataset mean : 1050000, Dummy regressor : mae=850000,rmse=3439000

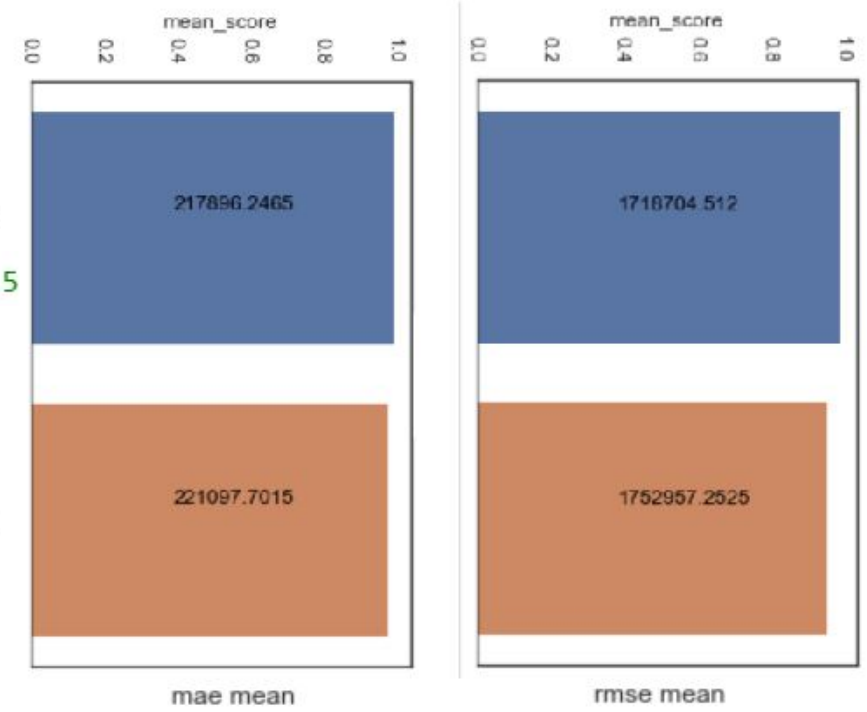
5.3.3 Affinage des résultats : gridsearchCV et extratrees

- Tester sur les variables à fortes corrélation pour un gain de temps
- Recherche paramètre par paramètre
- Recherche par liste paramètres pertinents

ExtraTreesRegressor_best_param

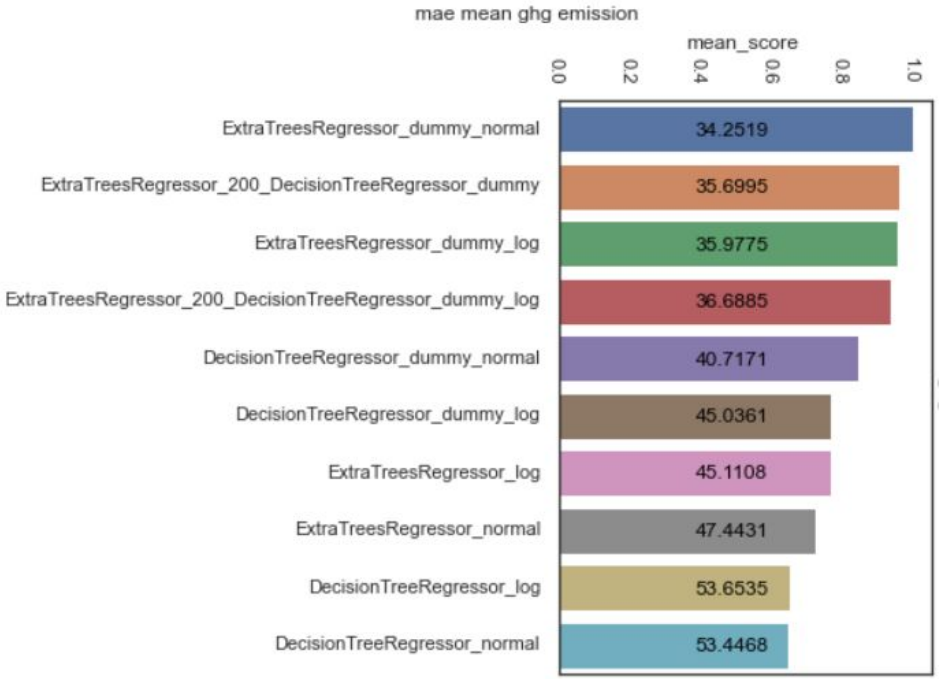
```
min_impurity_decrease=0.5
```

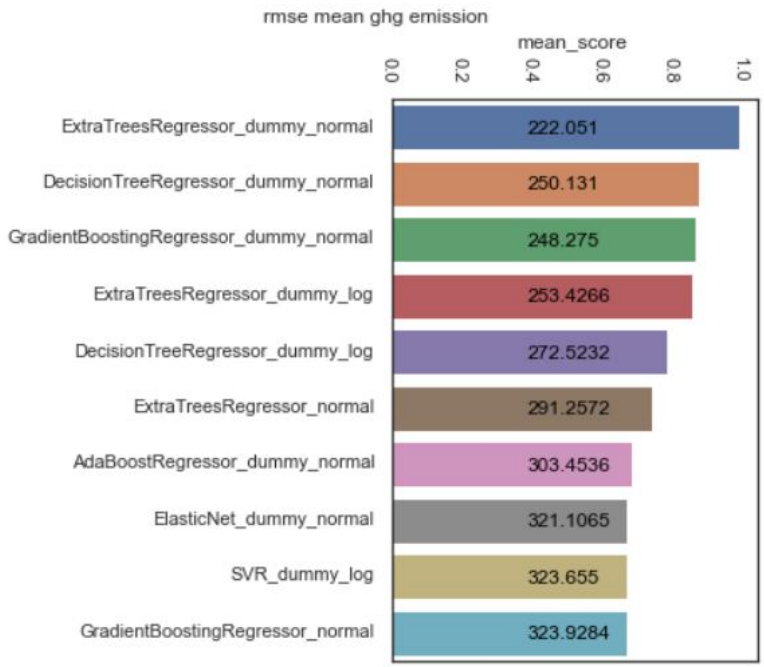
ExtraTreesRegressor_default



5.4 Résultats pour l’émission de gaz à effet de serre

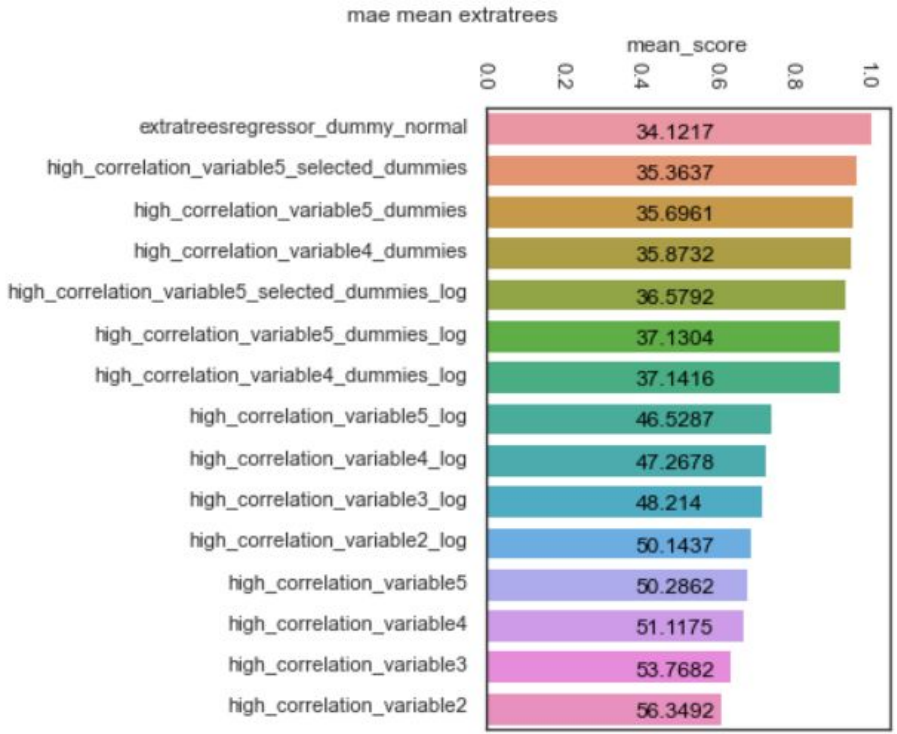
5.4.1 Résultats pour l'émission de gaz à effet de serre

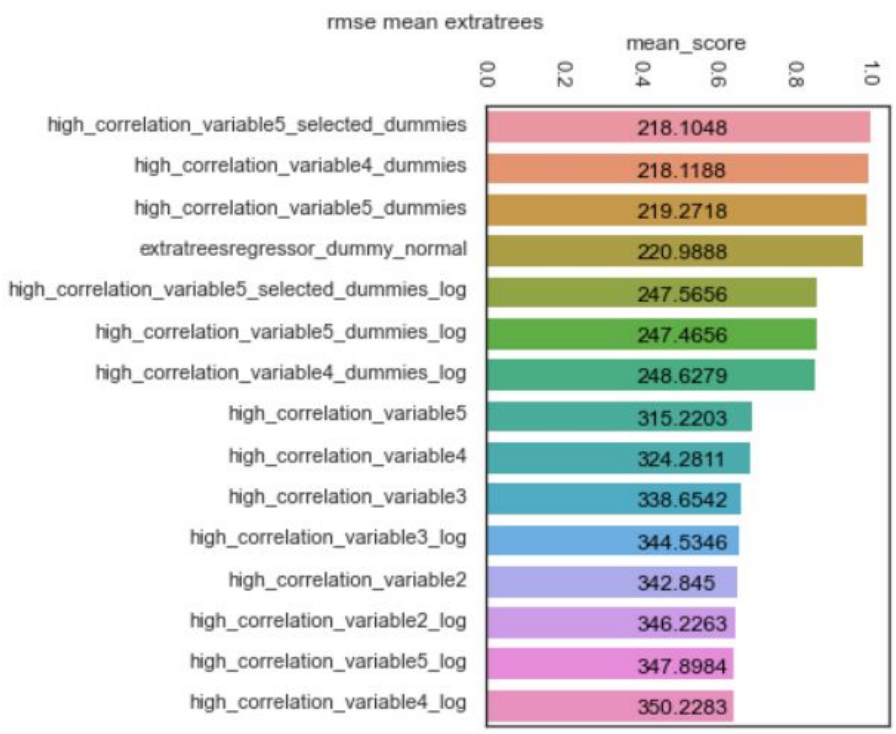




Dataset mean : 114, Dummy regressor : mae=101,rmse=456

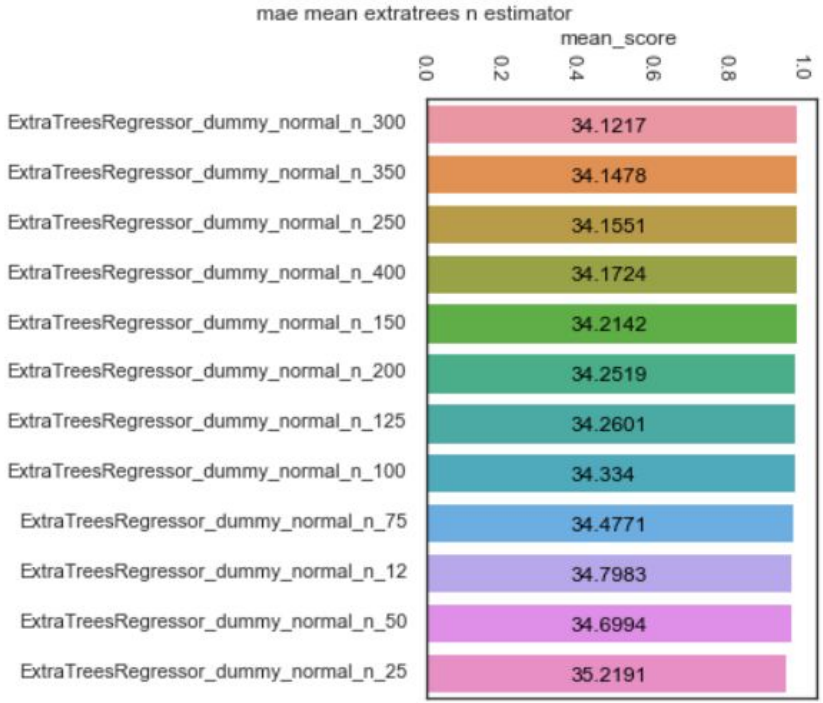
5.4.2 Affinage des résultats : variables à haute corrélation

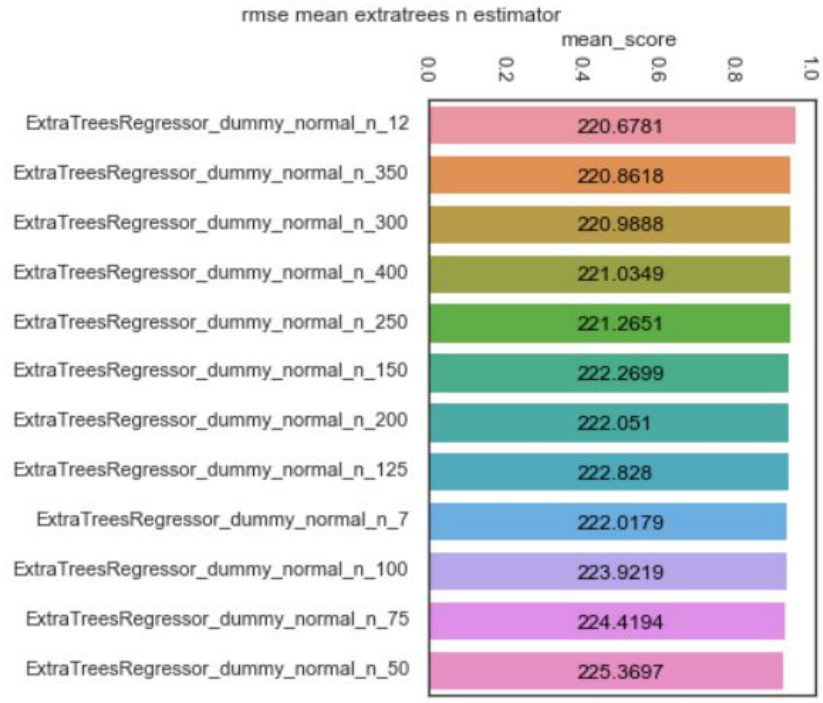




Dataset mean : 114, Dummy regressor : mae=101,rmse=456

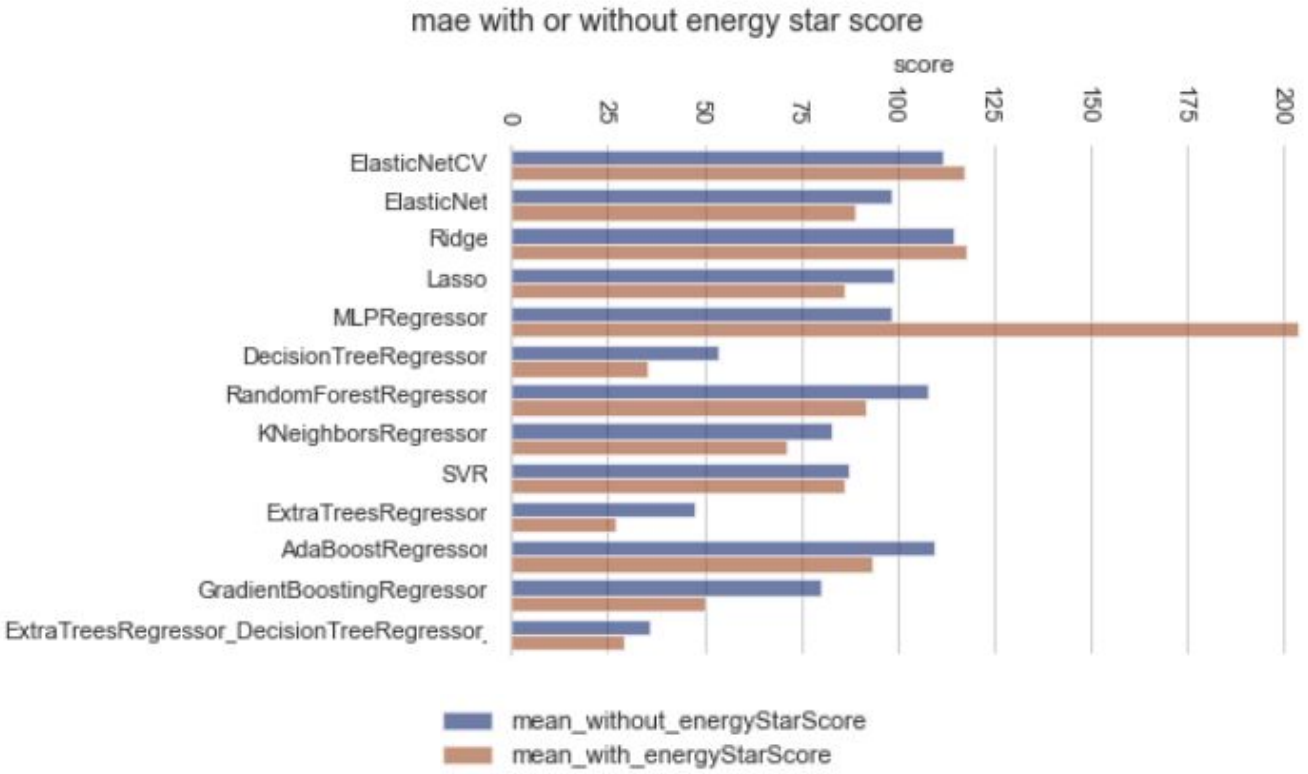
5.4.3 Affinage des résultats : paramètre n estimators



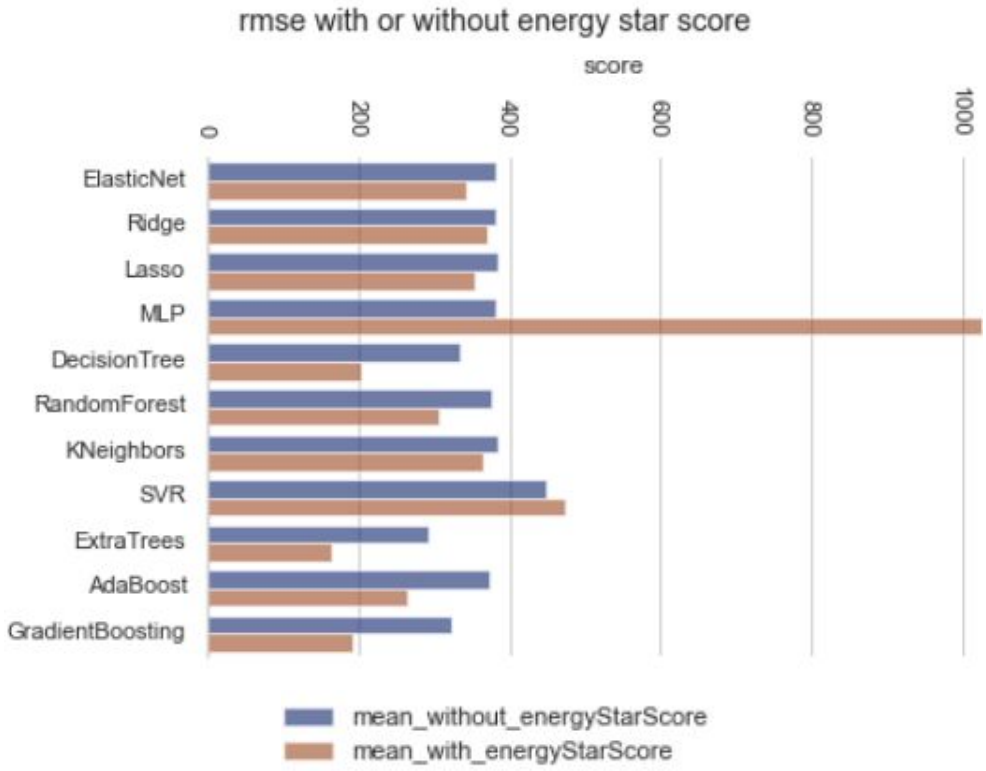


Dataset mean : 114, Dummy regressor : mae=101,rmse=456

5.4.4 Incidence de l'energy star score

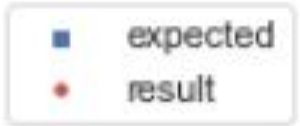


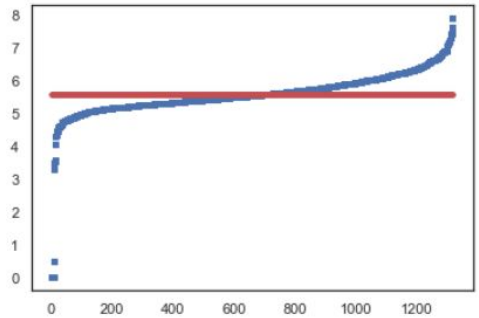
5.4.4 Incidence de l'energy star score



5.5 Axes d’amélioration

- Amélioration des variables d’entrées
  - les variables de type chaîne de caractère :
    - rassembler les valeurs de certains traits (type de bâtiments)
    - remplacer les valeurs discrètes par des nombres
  - les variables numériques :
    - trouver une conversion des données gps en valeurs
    - faire un compteur du nombre d’années energy star certified
    - trouver des valeurs pour les energy star score manquants
- Essayer du gridsearchCV sur le deuxième meilleur algorithme
- Trouver la cause des résultats sans corrélation





6. Conclusion

- **Pour l'électricité :**
  - moyenne du dataset = 1.050.000
  - Dummy regressor best score :
    - mae moyenne = 850.000
    - rmse moyenne 3.440.000
  - Meilleur algorithme pour la mae et rmse :
    - Extratrees
    - mae moyenne = 221.000
    - rmse moyenne = 1.770.000
  - variables d'entrées :
    - variables numériques à haute corrélation ou non
    - les one hot encoding totaux ou à haute utilité
  - meilleurs paramètres :
    - mae moyenne = 218.000
    - rmse moyenne = 1.718.000
- **Pour l'émission de gaz à effet de serre :**
  - moyenne du dataset = 114
  - Dummy regressor best score :
    - mae moyenne = 101
    - rmse moyenne 456
  - Meilleur algorithme pour la mae ou rmse :
    - Extratrees
    - mae moyenne = 26
    - rmse moyenne = 163
  - variables d'entrées :
    - gain (très léger) de mae ou de rmse en fonction du choix des variables d'entrées
  - incidence de l'energy star score :
    - forte incidence, amélioration très significative du score
      - mae : 34 → 26
      - rmse : 222 → 163