Projet 4 : Anticipez les besoins en consommation électrique de bâtiments

Sommaire

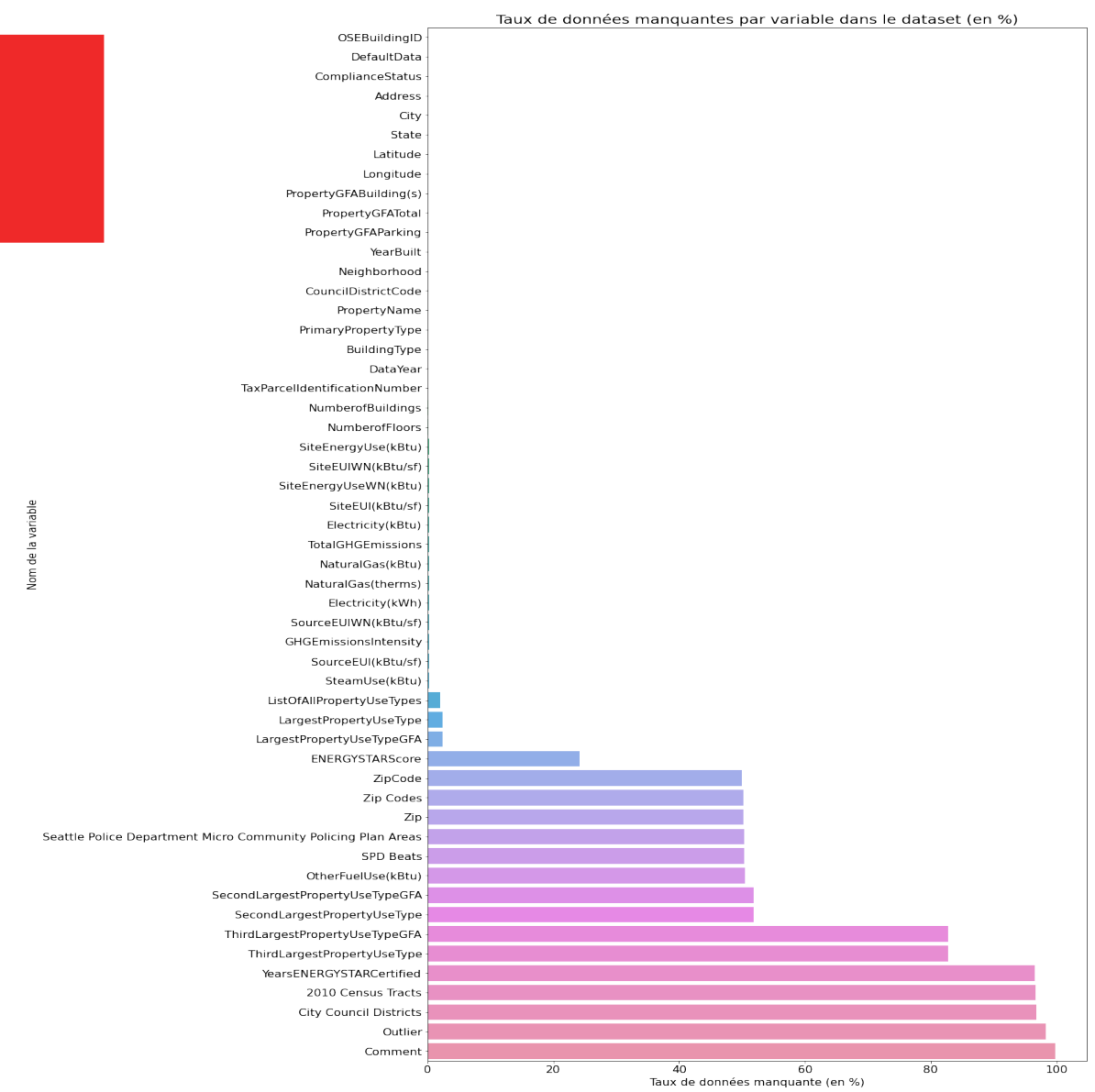
Problématique

- Des relevés de données de consommation ont été effectués pour les années 2015 et 2016 pour la ville de Seattle.
- Coût important et obtention fastidieuse

- Mission :
  - Prédire les émissions de CO2 et la consommation d'énergie
  - Évaluer l'intérêt de l'« ENERGY STAR Score » pour la prédiction d'émissions

Exploration des données

- 2 datasets pour les années 2015 et 2016 :
- 53 variables réparties comme suit :
  - 16 variables catégorielles
  - 36 variables numériques
  - 1 variable booléenne
  - Data2015 : 3340L, 47C / Data2016 : 3376 L, 46C
  - Certaines variables différentes
  - Données quasi-identiques et nom qui diffère pour certaines
  - filtration des bâtiments habitables

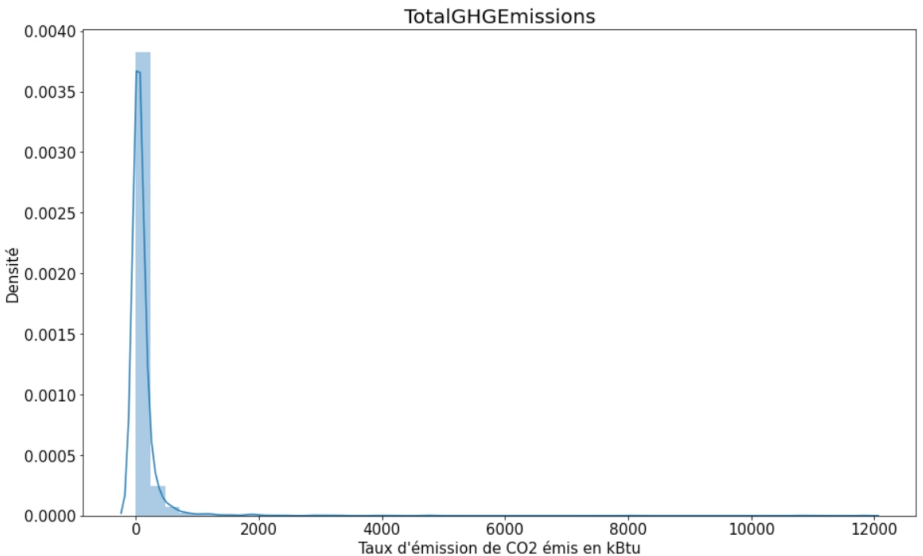


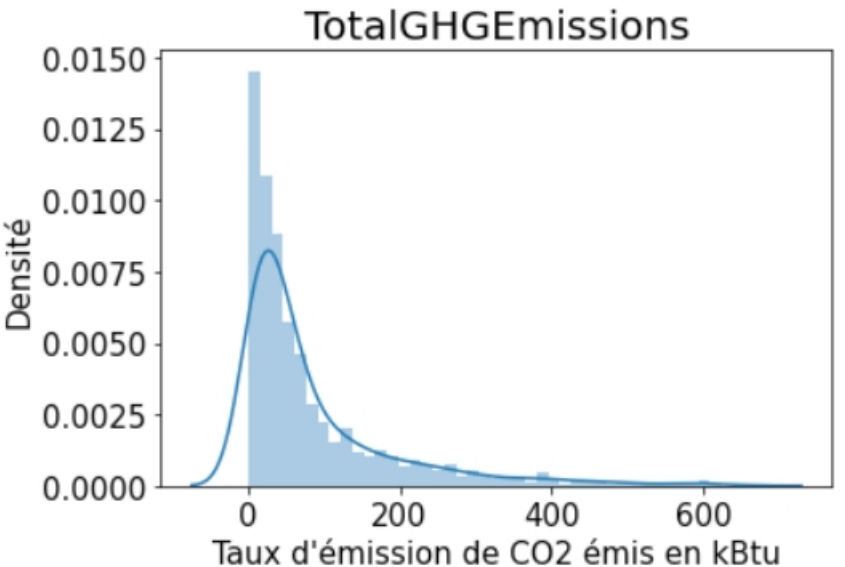
Exploration des données :

- Suppression des variables et individus ayant trop de valeurs manquantes et non intéressantes pour l'étude

Exploration des données :

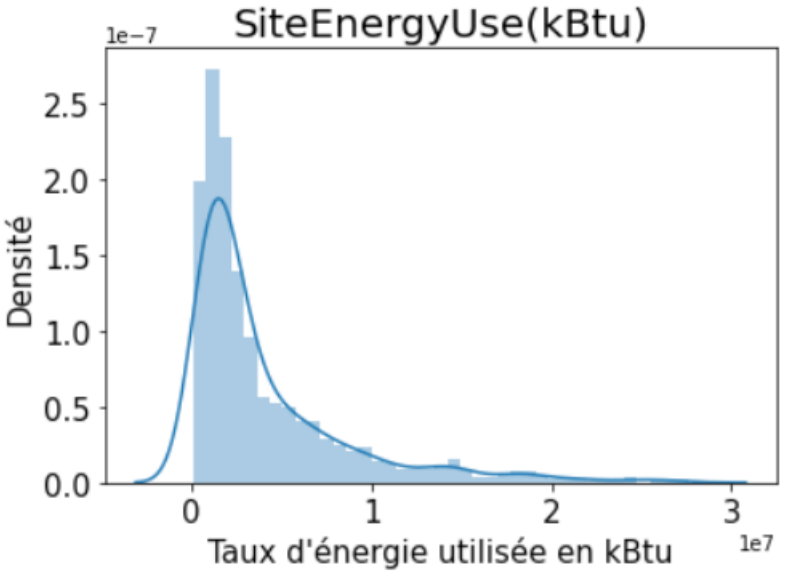
- Suppression des valeurs aberrantes
- Exemple avec la variable TotalGHGEmissions :

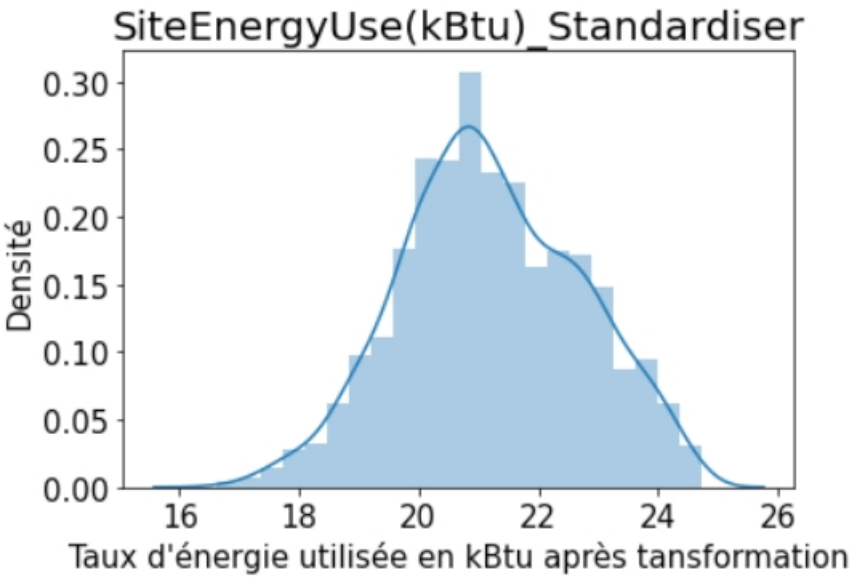




Exploration des données :

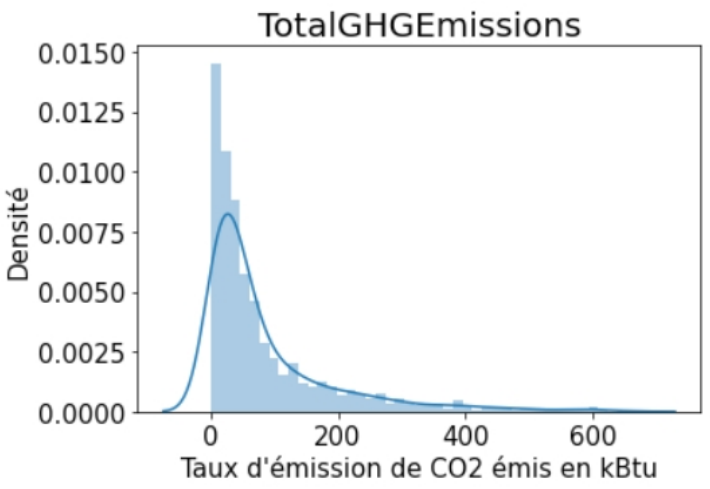
- Transformation de la distribution pour nos variables cibles :

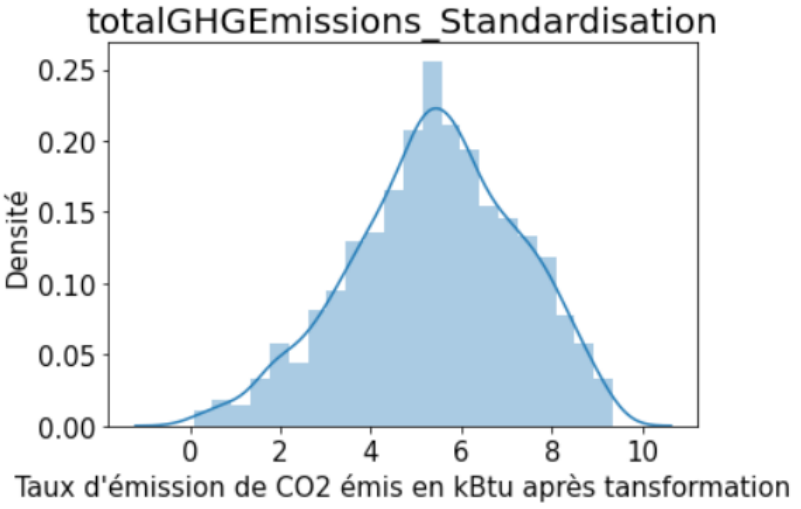




Exploration des données :

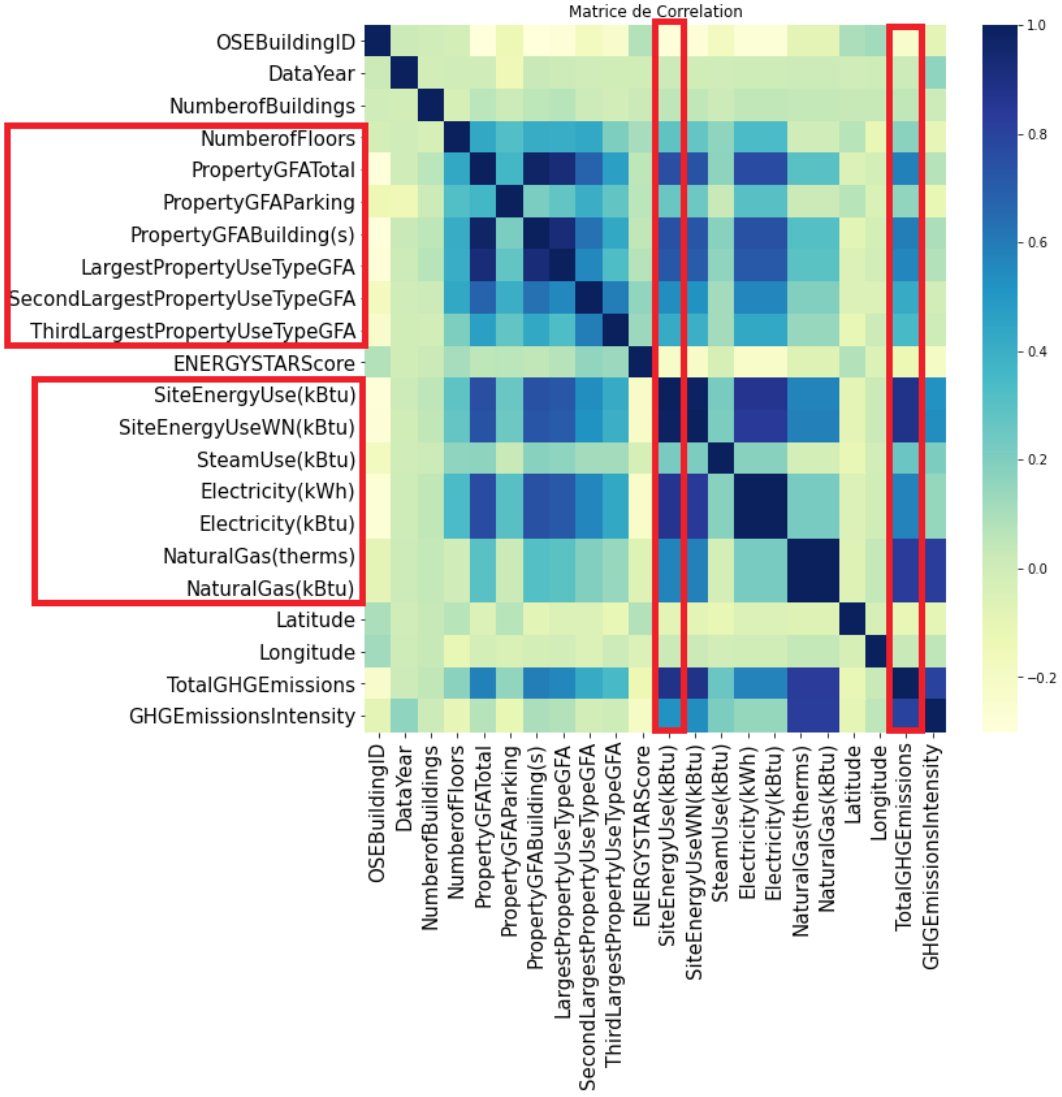
- Transformation de la distribution pour nos variables cibles :





Exploration des données :

- Matrice de Corrélation:



Modélisation

- Démarche :
  - Préparation du jeu de données
  - Construction des features
  - Entraînement des modèles
  - Optimisation des hyperparamètres
  - Évaluation du modèle
  - Sélection du meilleur modèle
  - Intérêt de la variable EnergyStarScore

Modélisation : Préparation du jeu de donnée

- Pour chaque variable cible :
  - Séparation des données en sous-ensembles d'entraînement et de test dans notre cas :
    - SiteEnergyUse ou TotalGHGEmissions
  - Nous prendrons 70 % du jeu de données pour l'entraînement et 30 % pour le test

Modélisation : Construction des features

- Passage au logarithme pour les variables cibles
- Modification des autres variables :
  - Standardisation des variables numériques
  - Encodage des variables catégorielles

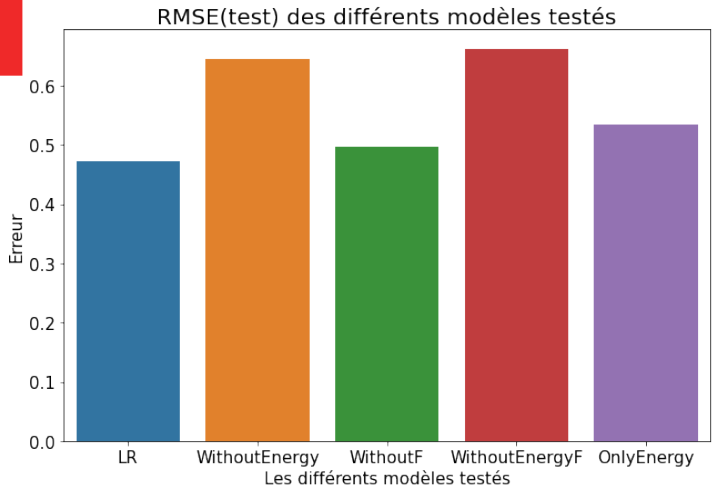
Modélisation : Les différents modèles

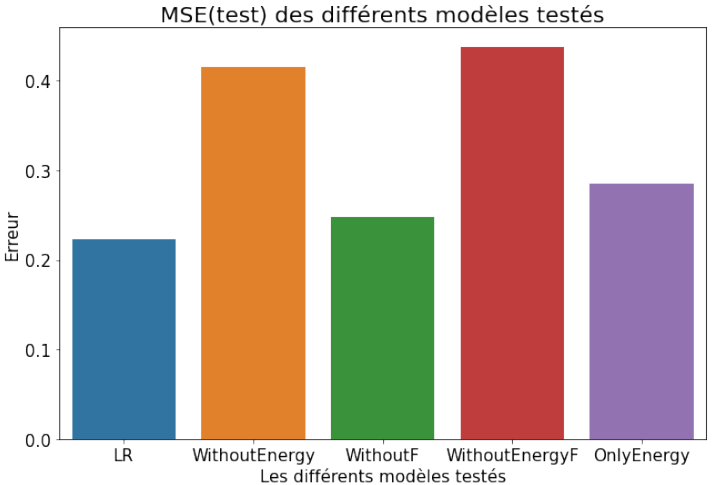
Linear Regressor	Pas d'hyperparamètre
Ridge Regressor	alpha : [0.0001, 0.001, 0.01, 0.1, 1, 10, 100]
Random Forest Regressor	n_estimators : [800, 1200] max_depth : [15, 30, None] max_features : ['auto', 'sqrt', 'log2'] min_samples_leaf : [1, 2] min_samples_split : [2, 5]

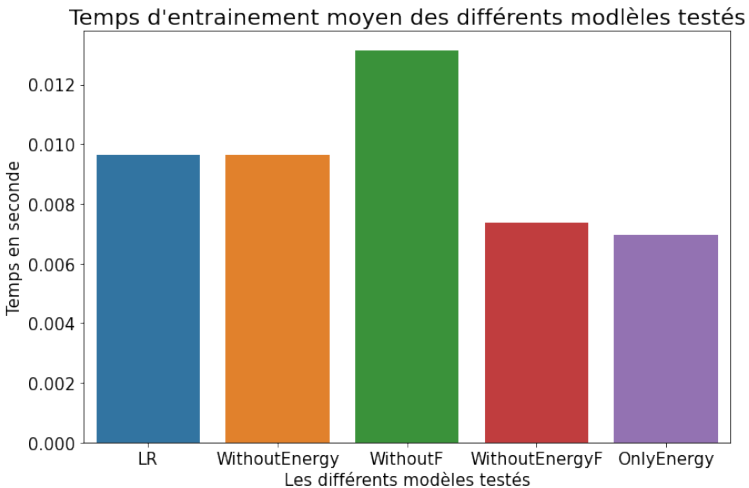
Modélisation : Optimisation des hyperparamètres

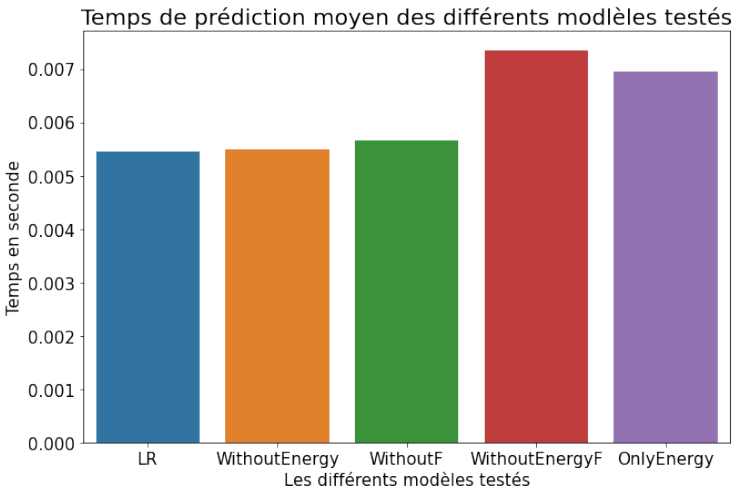
- Pour chaque modèle, après avoir séparé nos données en jeu d'entraînement et en jeu de test :
  - Recherche des meilleurs paramètres avec la méthode de recherche par grille
  - Optimisation des meilleurs paramètres en utilisant la cross-validation et en minimisant la fonction de coût (dans notre cas Erreur absolue moyenne négative).

Modélisation : Consommation d’énergie





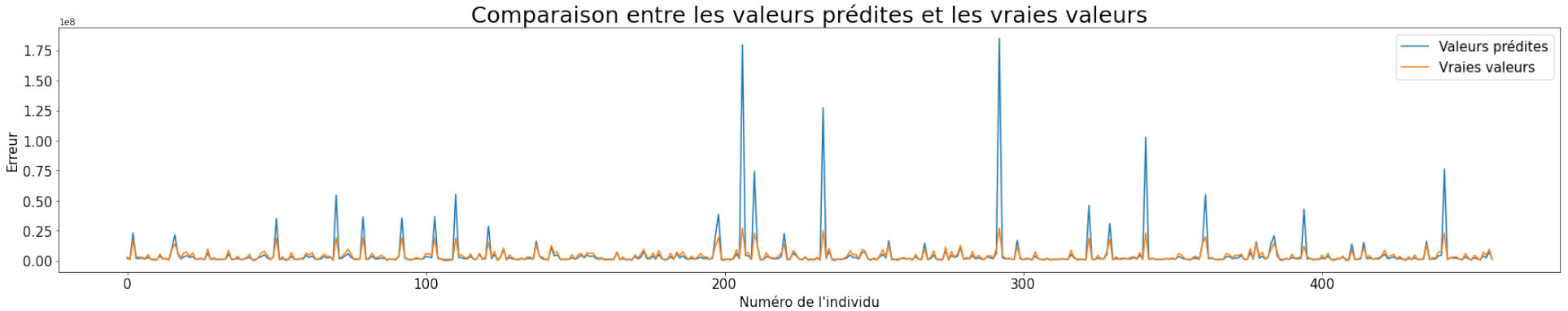




- Impact variable:

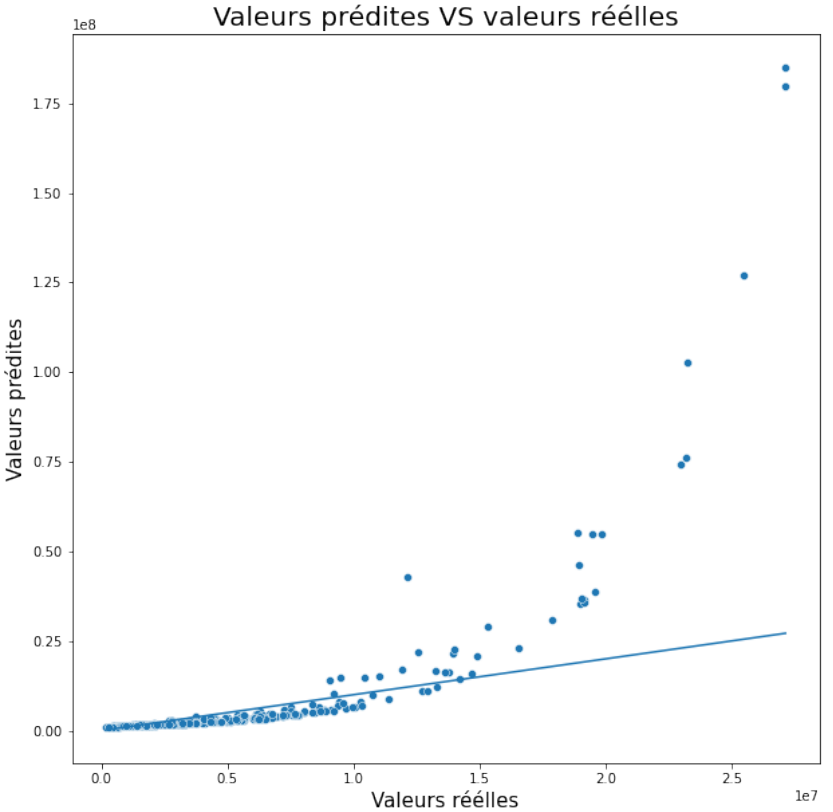
Modélisation : Consommation d’énergie

- Régression linéaire:



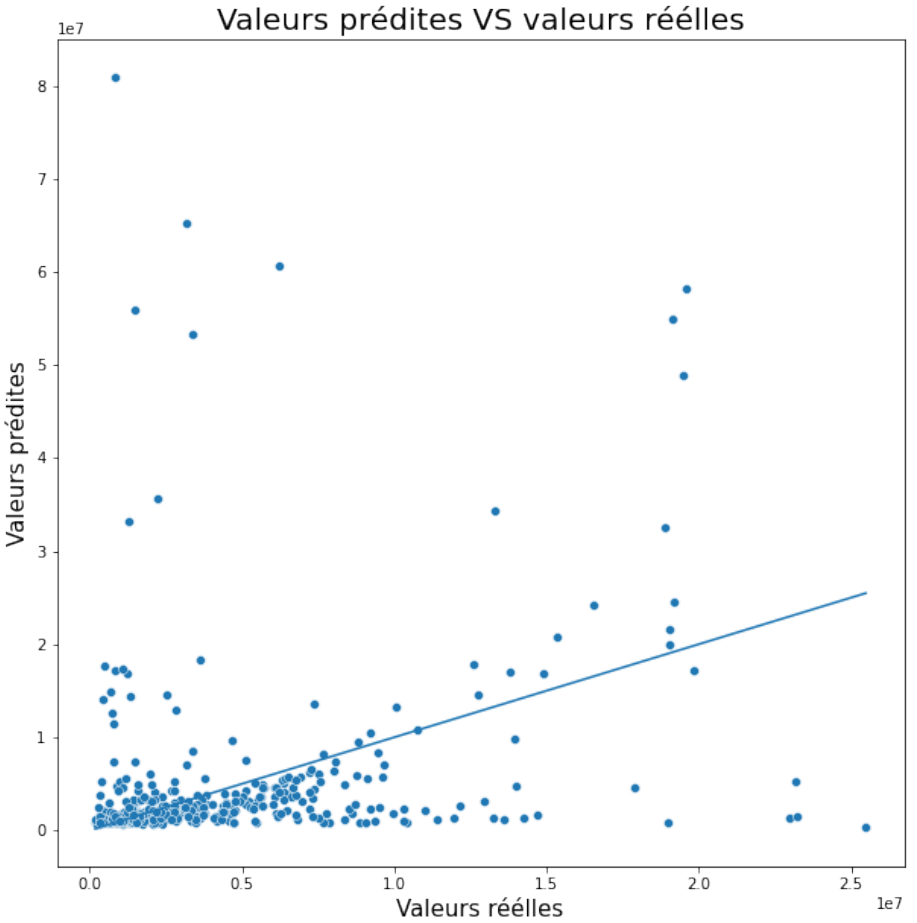
Modélisation : Consommation d’énergie

- Régression linéaire:



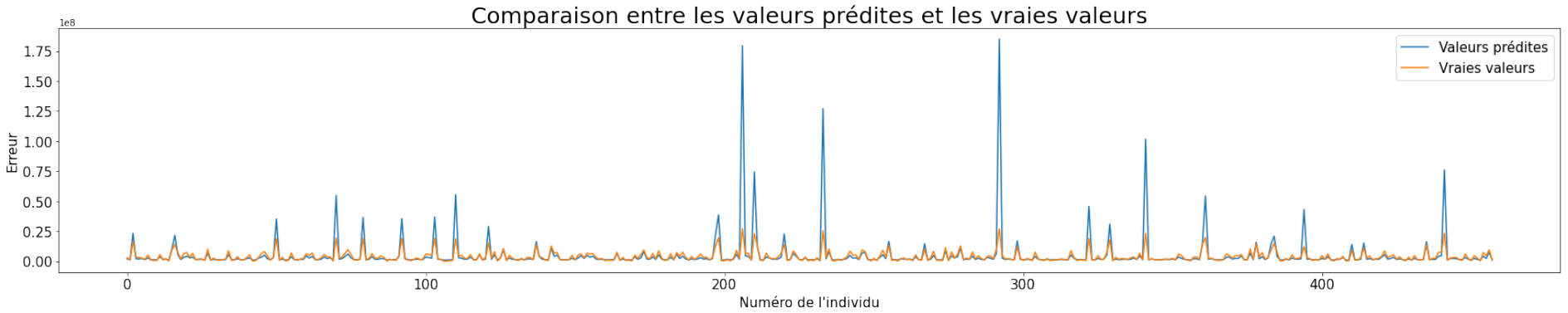
Modélisation : Consommation d’énergie

- Régression linéaire:



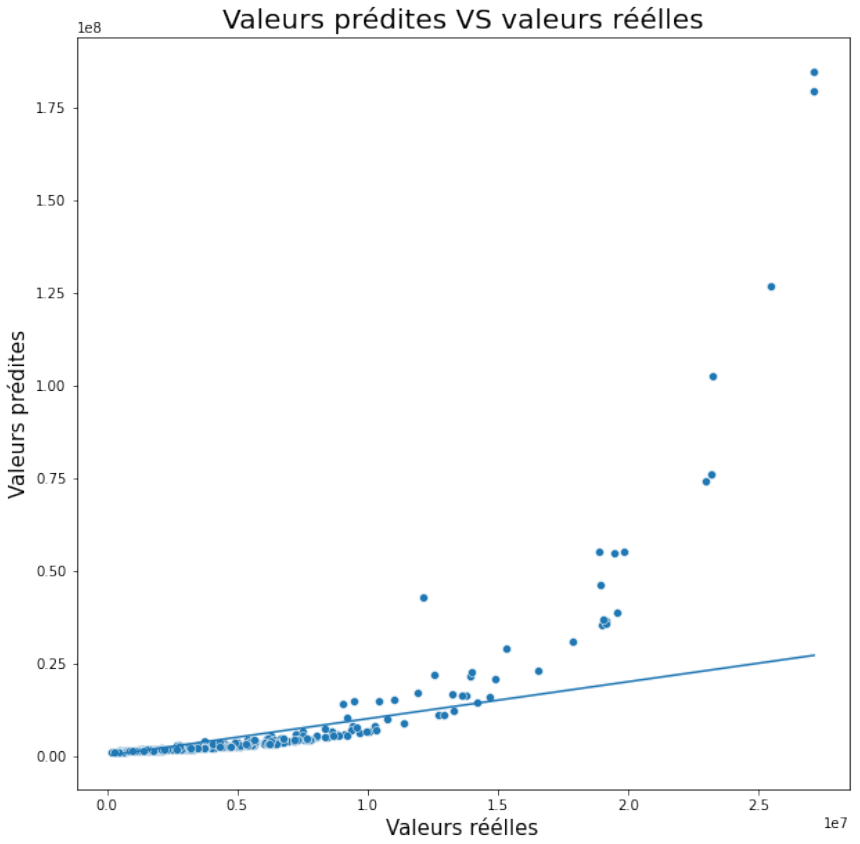
Modélisation : Consommation d’énergie

- Régression Ridge :



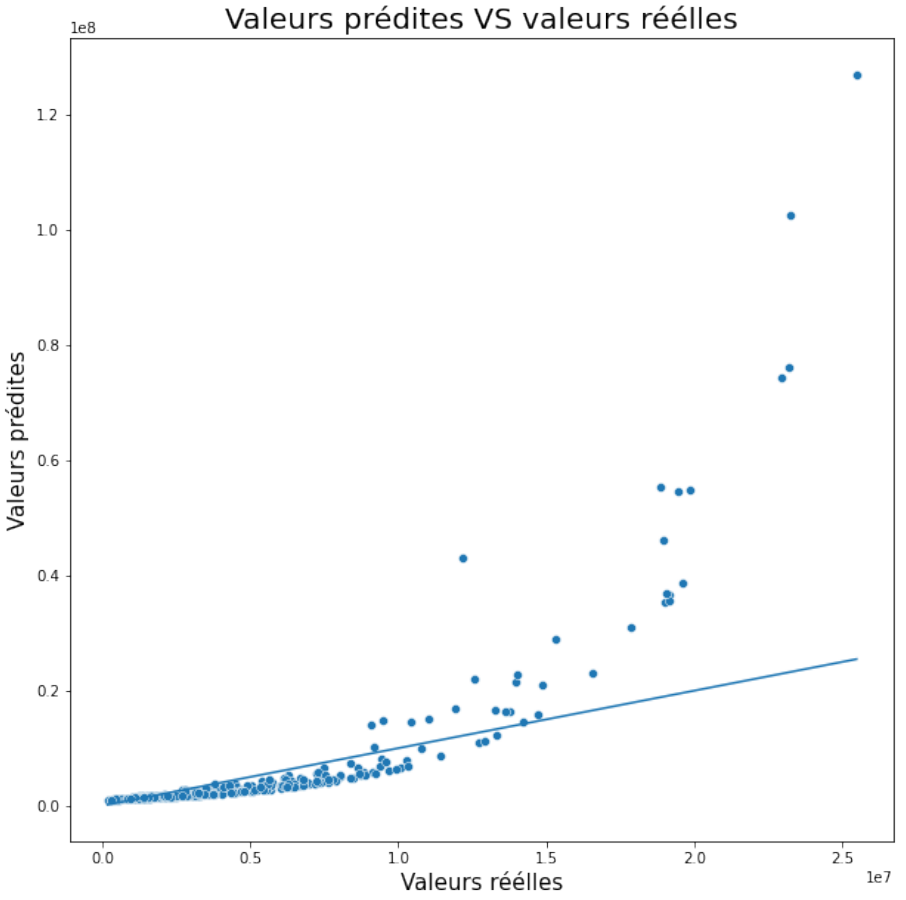
Modélisation : Consommation d’énergie

- Régression Ridge :



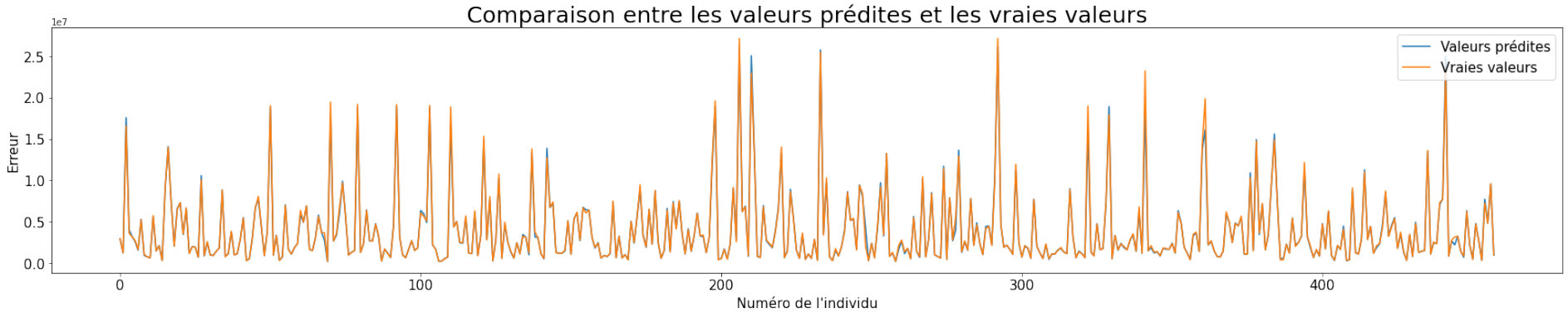
Modélisation : Consommation d’énergie

- Régression Ridge :



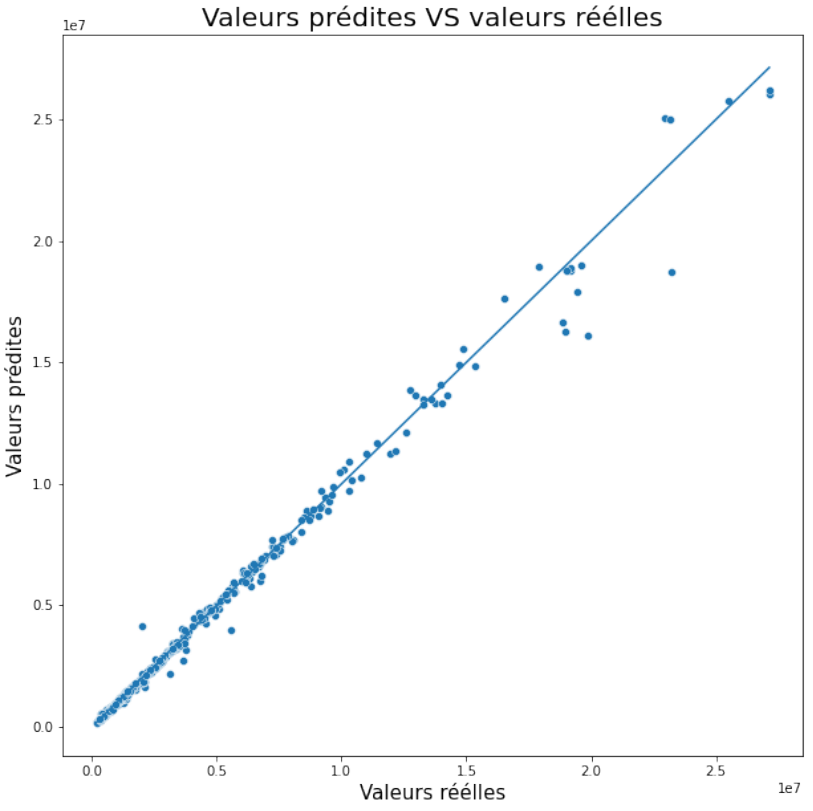
Modélisation : Consommation d’énergie

- Random Forest Regressor :



Modélisation : Consommation d’énergie

- Random Forest Regressor :



Modélisation

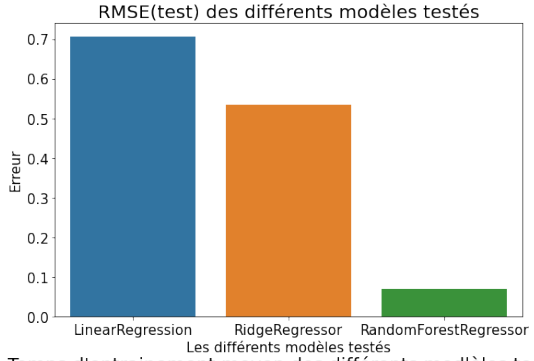
- Comparaison des performances des 3 modèles pour la prédiction de la consommation d'énergie :

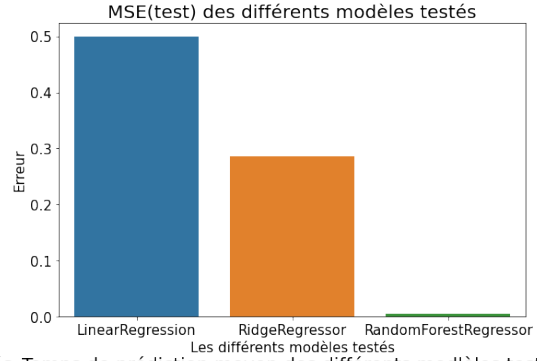
  <u>Meilleur Modèle :</u>

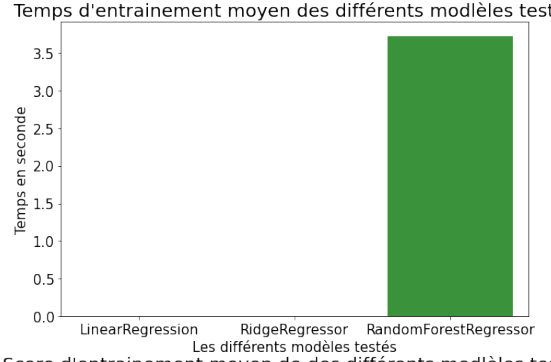
  - Random Forest Regressor
  - RMSE : 0,06971

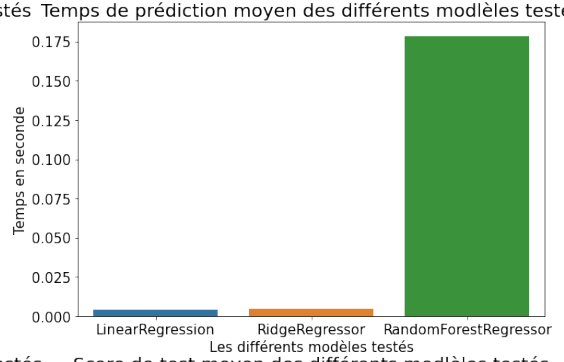
  <u>Combinaison des hyperparamètres :</u>

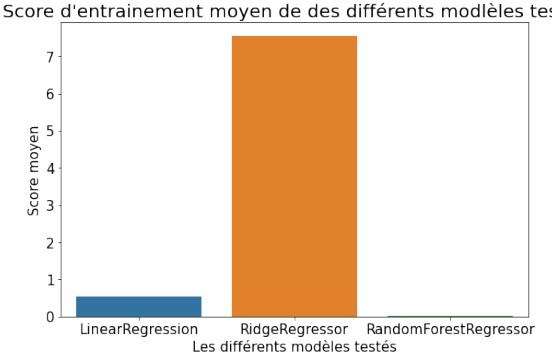
  - max_depth : 15
  - max_features : auto
  - min_samples_leaf : 1
  - min_samples_split : 2
  - n_estimators : 1200

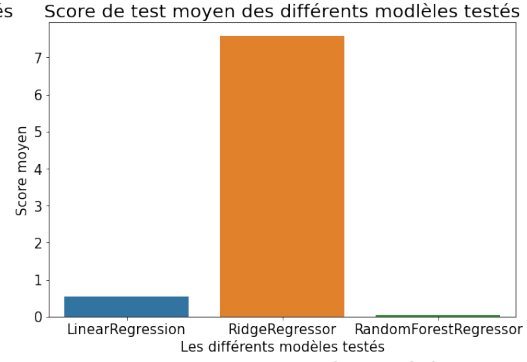




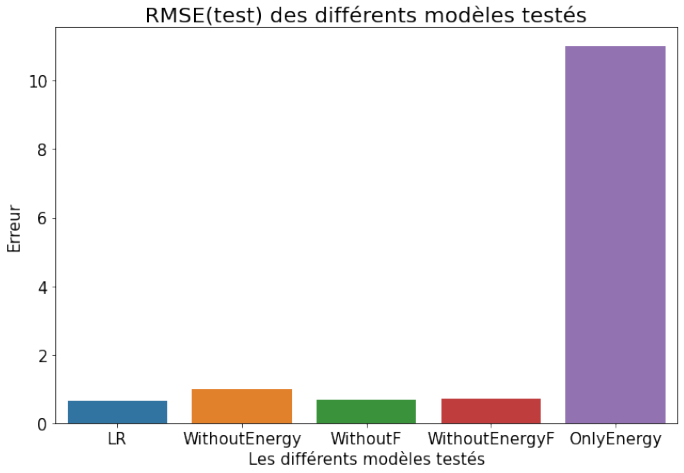


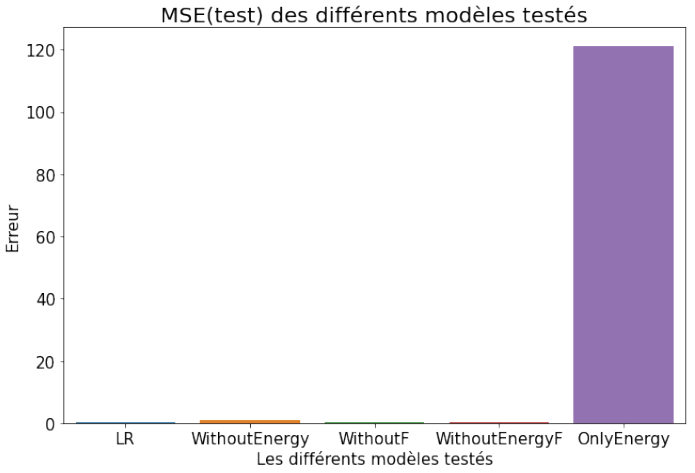


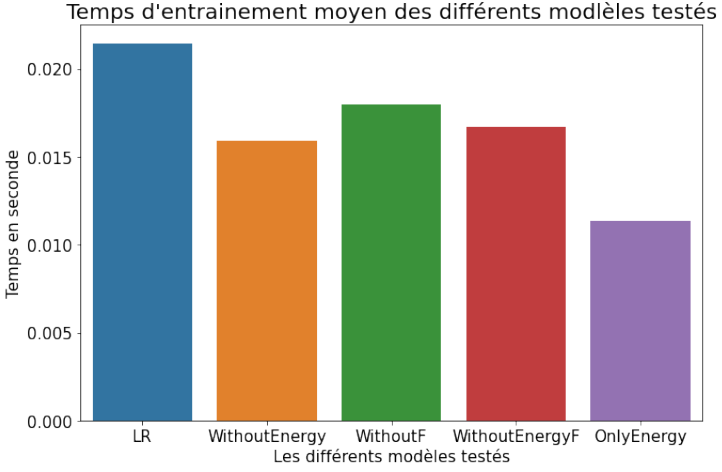


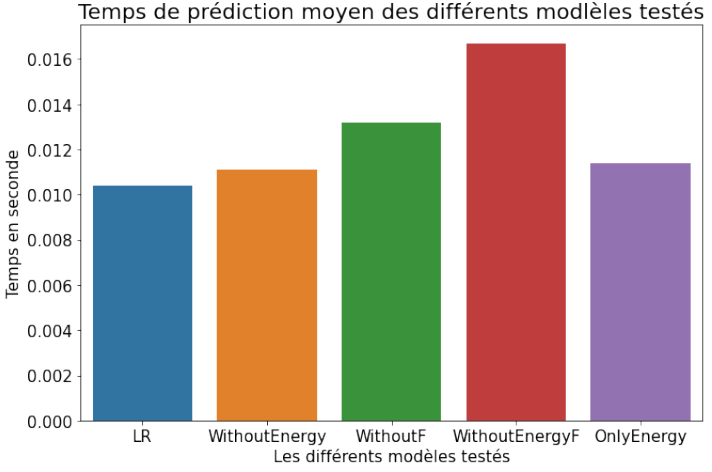


Modélisation : Émission de CO2





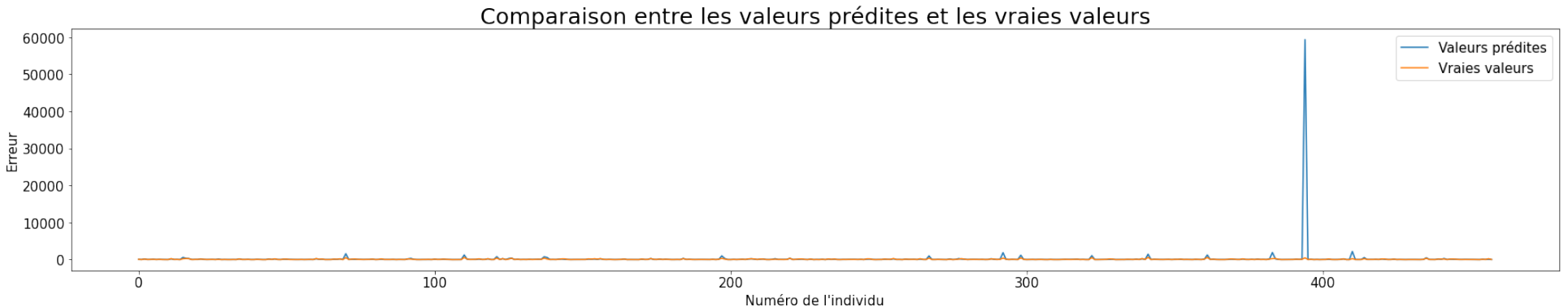




- Impact variable:

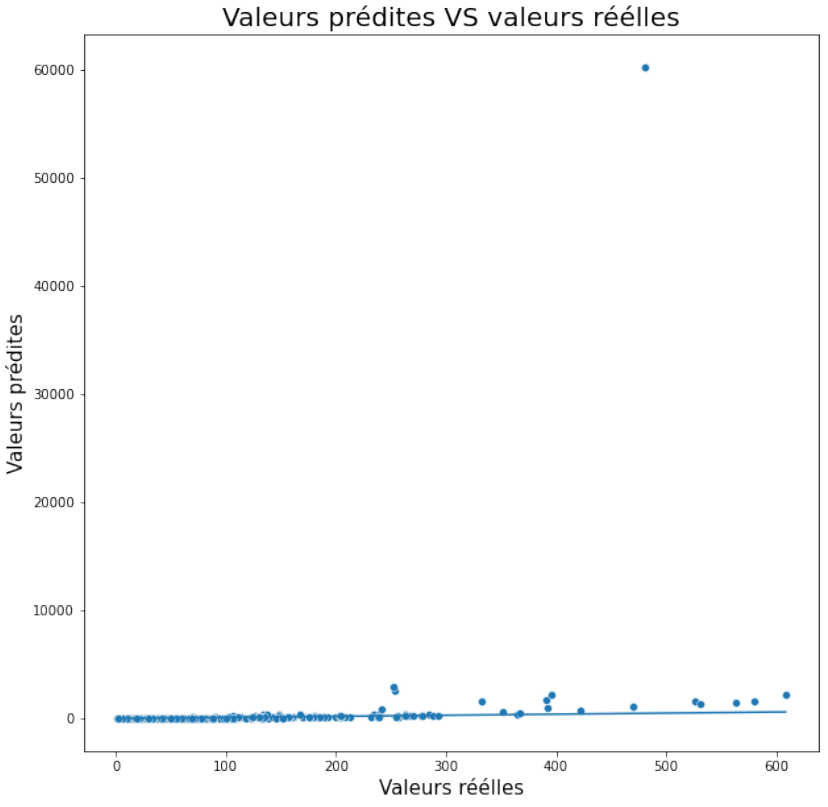
Modélisation : Émission de CO2

- Régression linéaire:



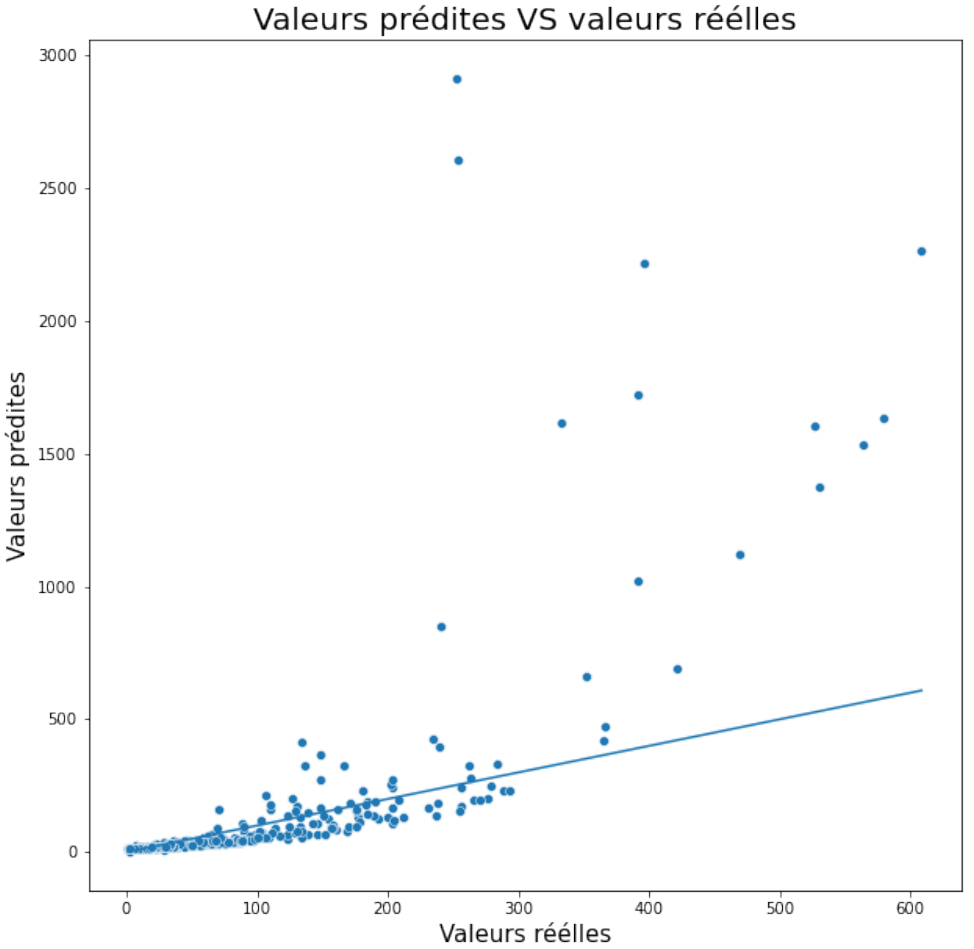
Modélisation : Émission de CO2

- Régression linéaire:



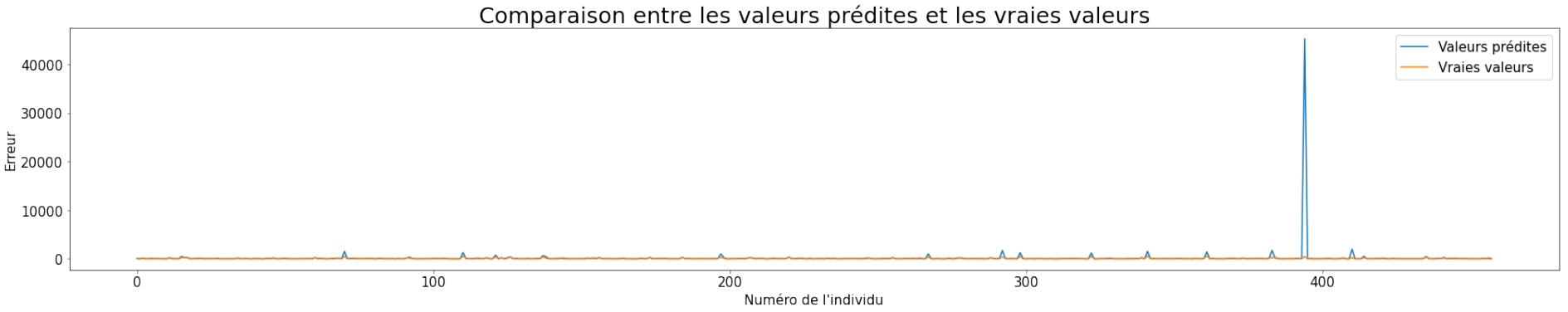
Modélisation : Émission de CO2

- Régression linéaire:



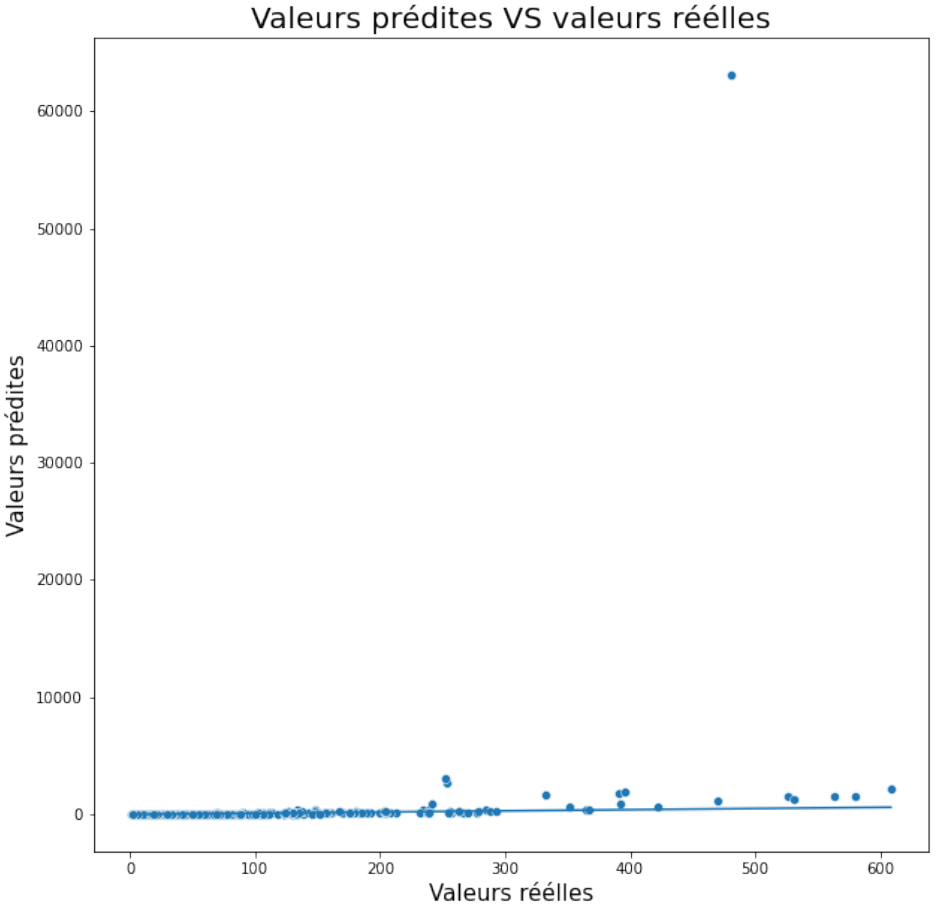
Modélisation : Émission de CO2

- Régression Ridge:



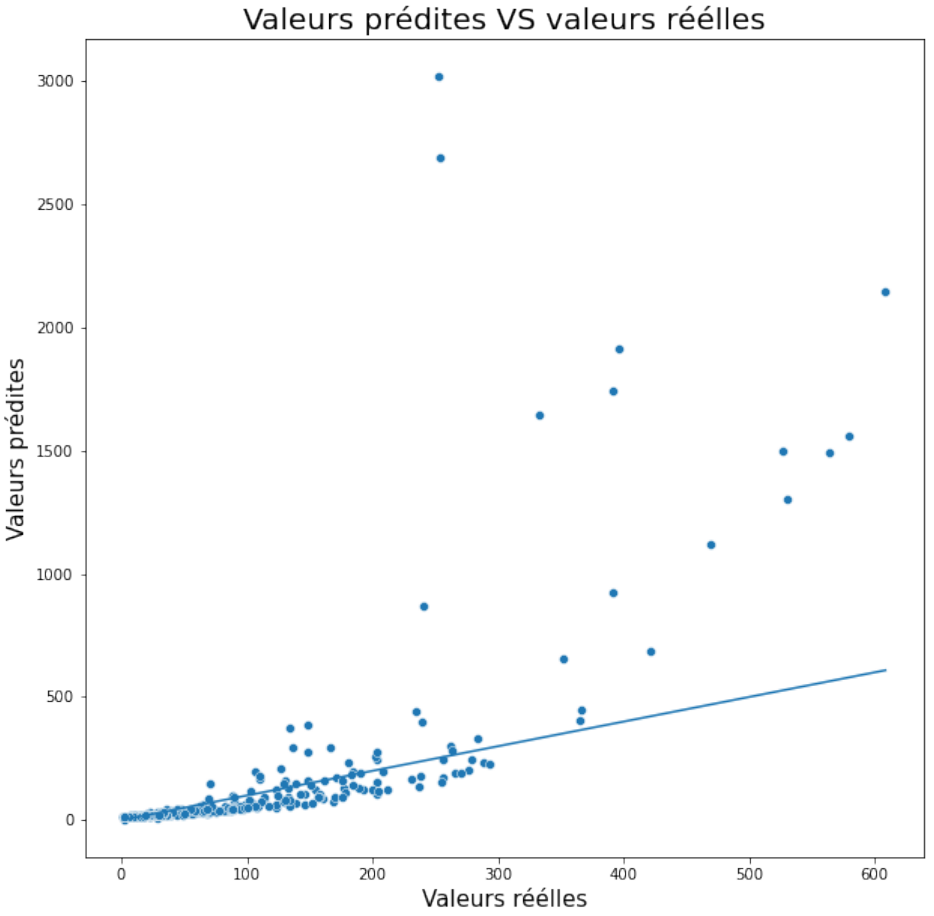
Modélisation : Émission de CO2

- Régression Ridge:



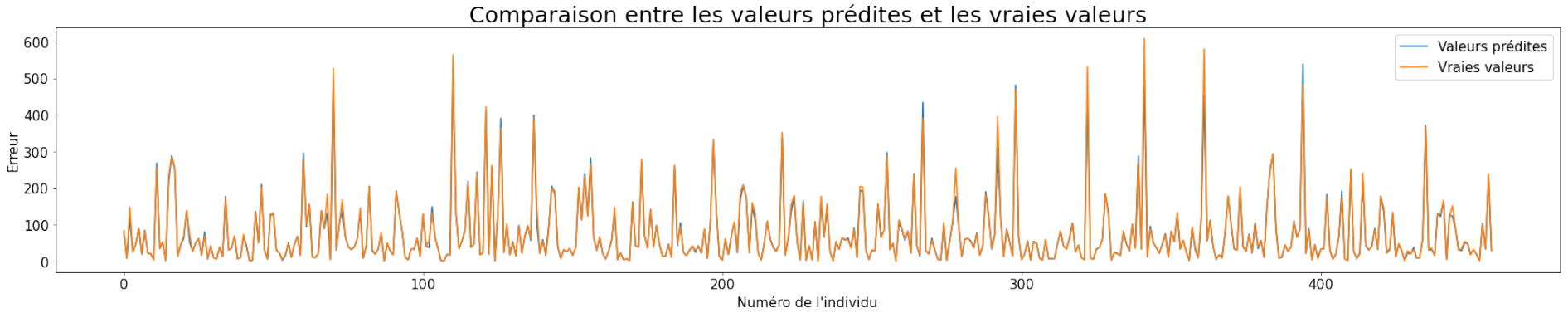
Modélisation : Émission de CO2

- Régression Ridge:



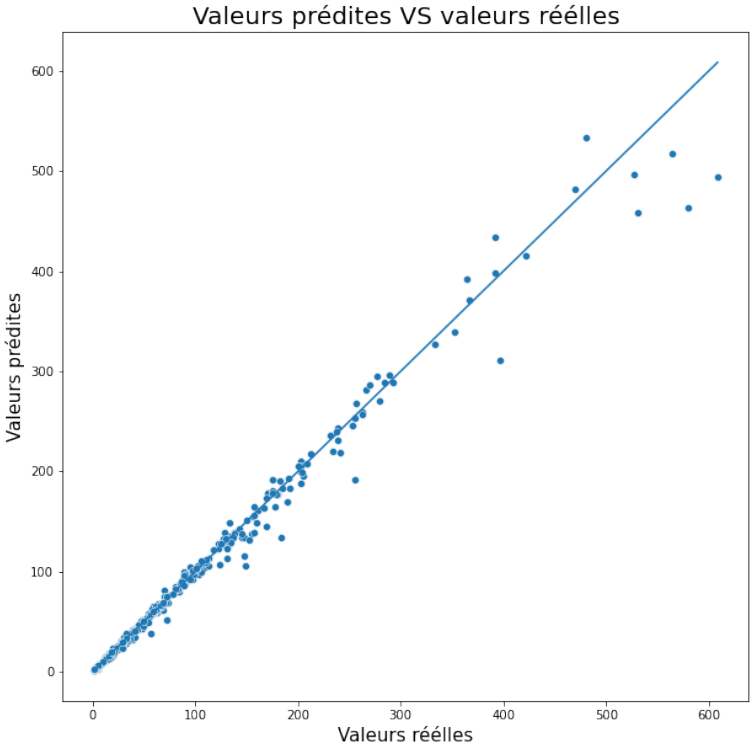
Modélisation : Émission de CO2

- Random Forest Regressor:



Modélisation : Émission de CO2

- Random Forest Regressor:



Modélisation

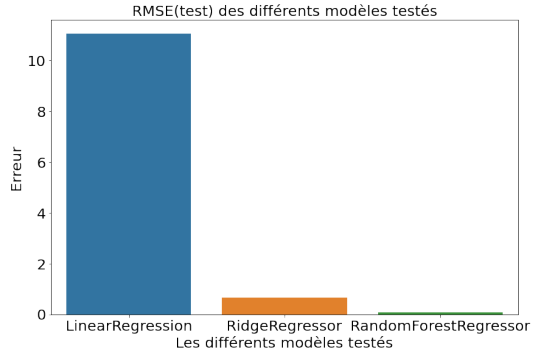
- Comparaison des performances des 3 modèles pour la prédiction d'émission de CO2 :

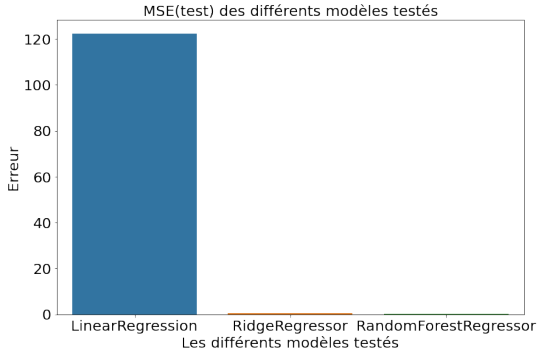
<u>Meilleur Modèle :</u>

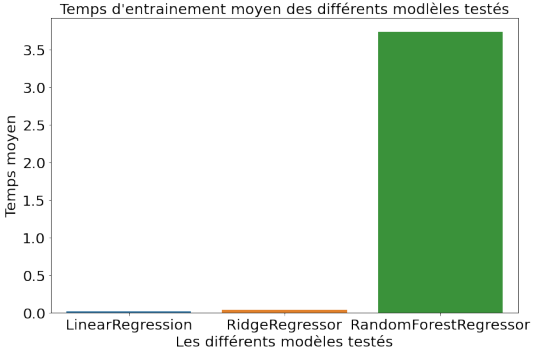
  - Random Forest Regressor
  - RMSE : 0,07320

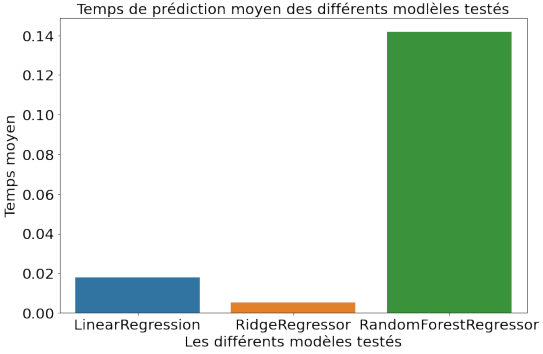
<u>Combinaison des hyperparamètres :</u>

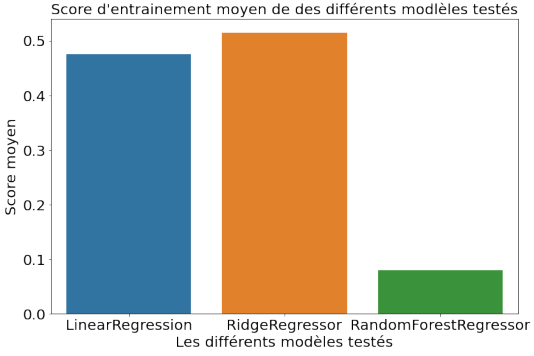
  - max_depth : None
  - max_features : auto
  - min_samples_leaf : 1
  - min_samples_split : 2
  - n_estimators : 1200

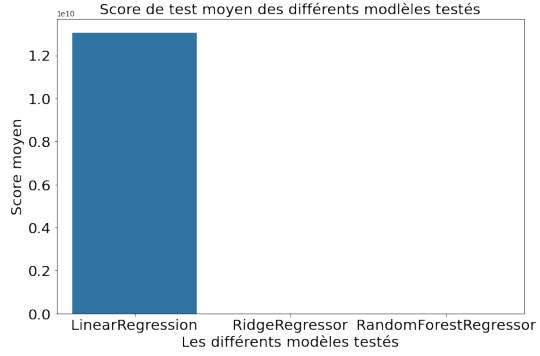






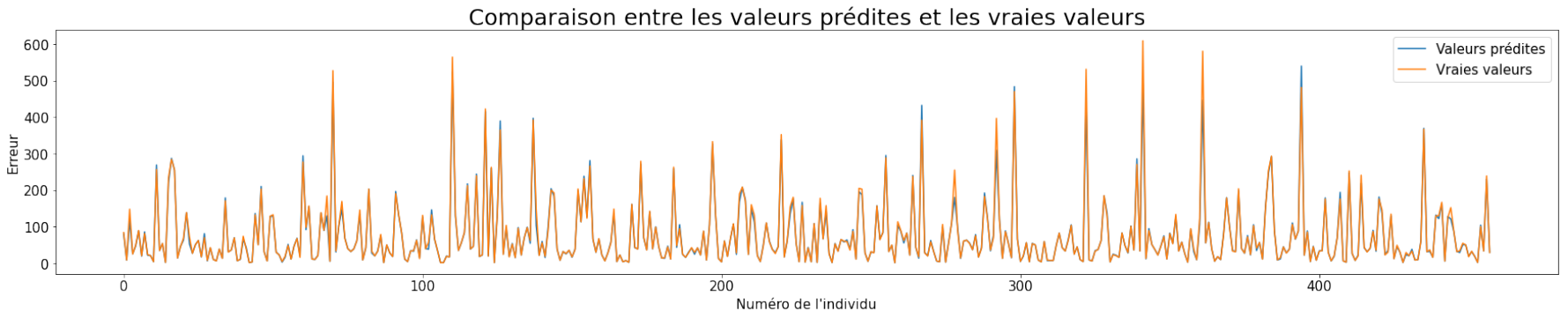






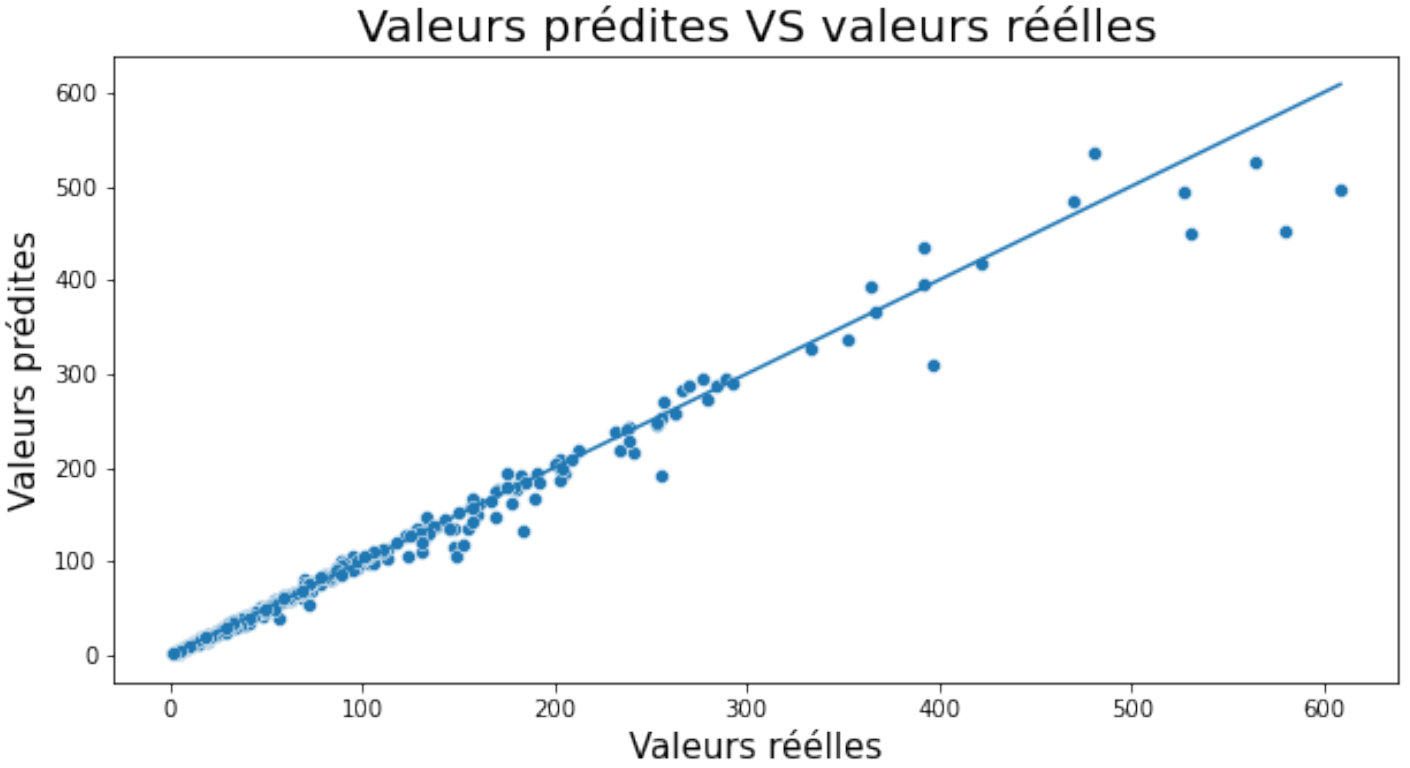
Modélisation : Intérêt EnergyStarScore

- Random Forest Regressor:



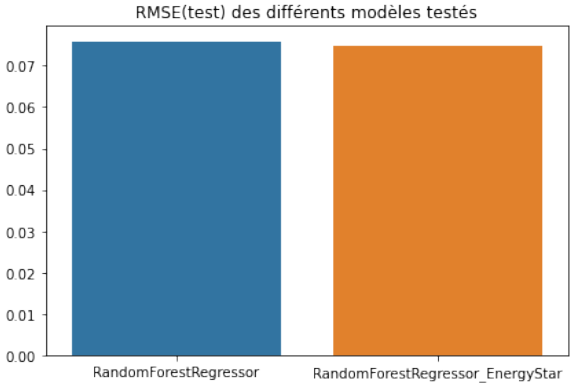
Modélisation : Intérêt EnergyStarScore

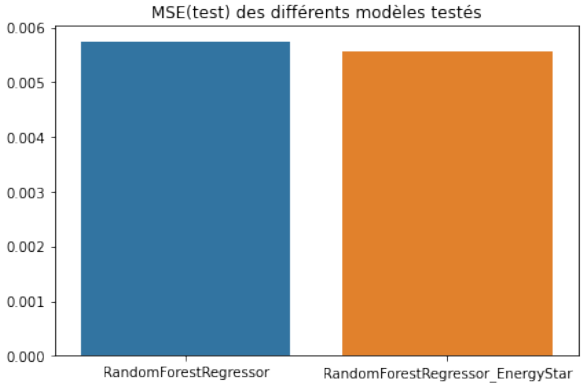
- Random Forest Regressor:

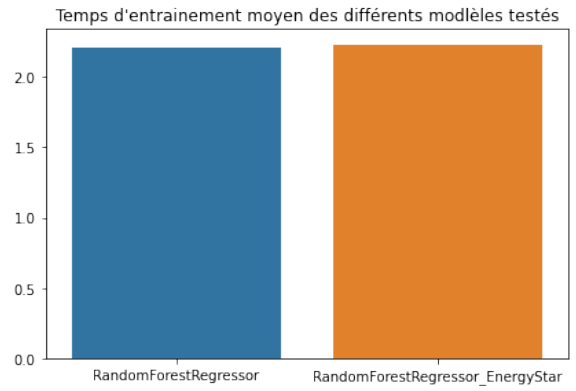


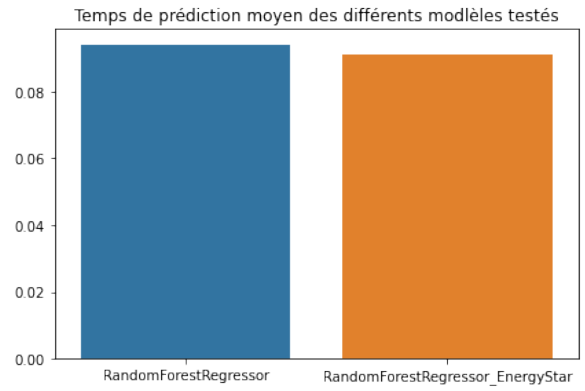
Modélisation

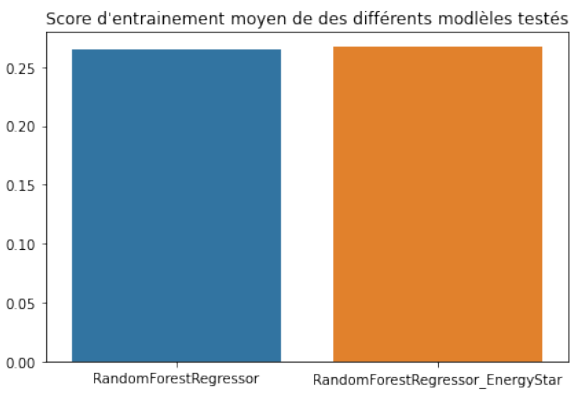
- Intérêt de la variable EnergyStarScore :

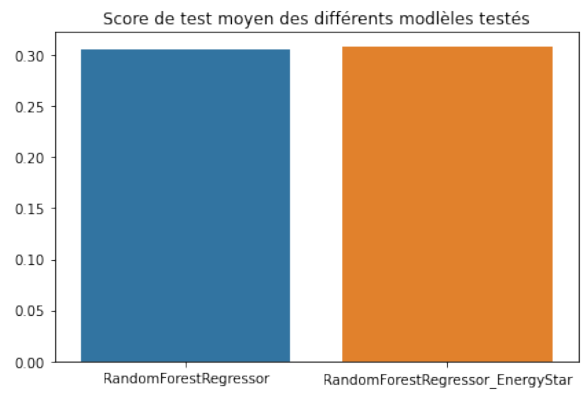












Conclusion

- Random Forest Regressor est le meilleur modèle pour notre problème
- Intérêt négligeable de EnergyStarScore

Merci de votre attention