Projet 3 - Anticipez les besoins en consommation électrique de bâtiments

Aurélien Corroyer-Dulmont, PhD

Ingénieur imagerie médicale

Centre Baclesse
L'excellence pour vaincre votre cancer

Rappel de l'appel à projet

- Problématique :
  - La ville de Seattle souhaite atteindre la neutralité carbone en 2050
  - Score pour prédire la consommation en énergie et émission en GES
    => création de l'ENERGYSTARScore
    - basé sur les relevés de consommation en énergie
    - couteux à avoir car les relevés sont compliqués à obtenir
  - Nécessité d'une solution n'utilisant pas ces relevés pour la prédiction

Centre Baclesse
L'excellence pour vaincre votre cancer

Rappel de l'appel à projet (5min)

- Interprétation de l'AAP :
  - Utilisation des données des permis d'exploitation (surface, usage, date de construction…) ?
  - Utilisation des données de 2015 et 2016 sur plus de 3000 bâtiments pour chaque année relevant plus de 45 informations différentes

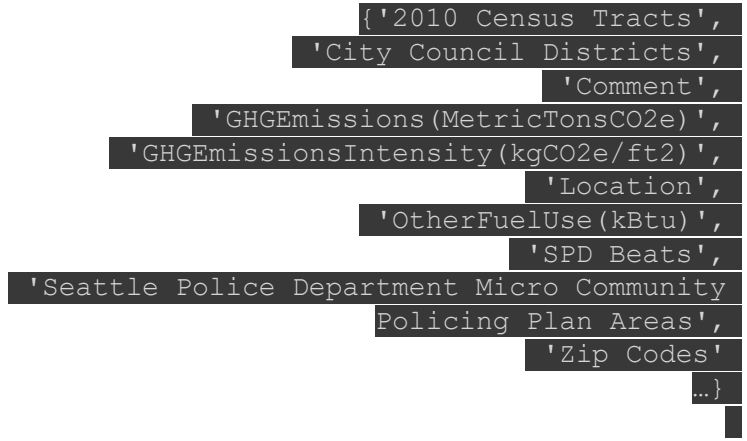
Rappel de l’appel à projet

- Pistes de recherche envisagées :
  - À partir de ces données, création d’un modèle de machine learning supervisé avec régression linéaire, non linéaire ou de type ensembliste pour une meilleure prédiction des variables d’intérêt
  - Permettant, sans les informations de relevés de compteurs, de prédire la consommation en énergie, mais aussi des émissions des GES pour les prochaines années

Nettoyage des données

- Construction d'une base de donnée 2015+2016 :
  - Certains features ne concordent pas entre les deux années
    - Suppression des features qui ne sont que pour une année

ou changement de nom



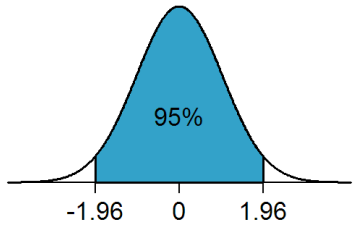
- Dans l'AAP il est mentionné *"votre équipe s'intéresse de près aux émissions des bâtiments non destinés à l'habitation"*.
  - J'ai supprimé les lignes correspondantes aux maisons résidentielles, information que l'on trouve dans le feature : *BuildingType*

    - Décompte des variables présentant un nombre de données manquantes trop important :
      - suppression de ces variables si *NaN* > 50 %

Centre Baclesse
L'excellence pour vaincre votre cancer

Nettoyage des données

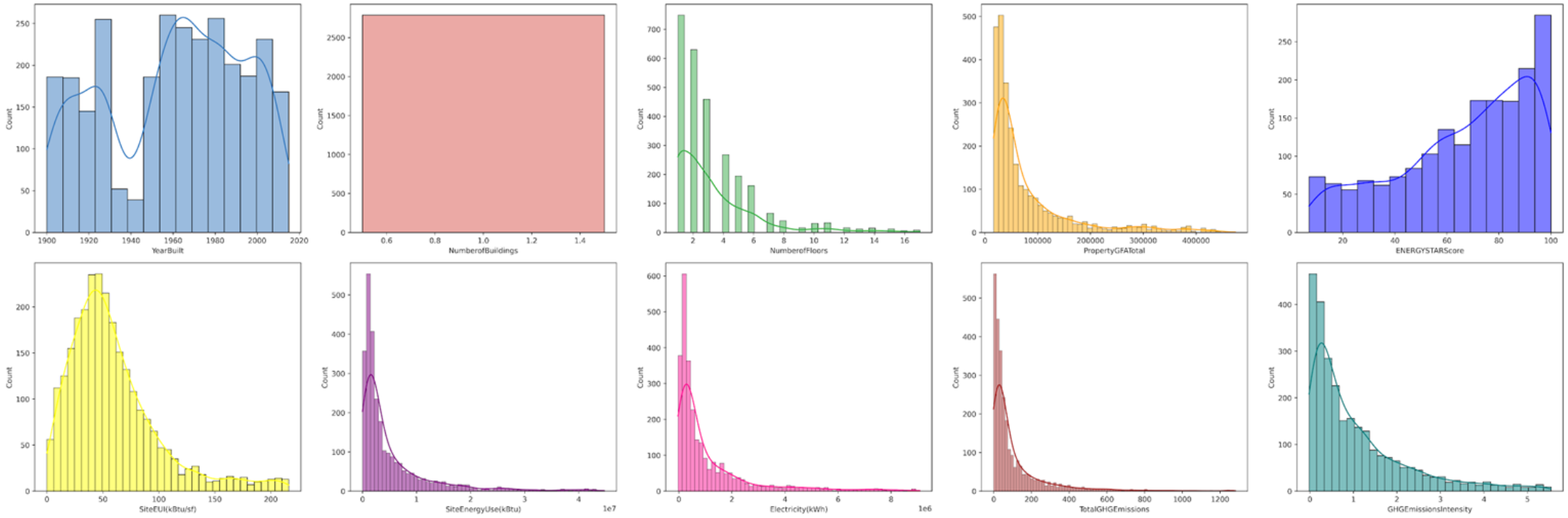
- Nettoyage des valeurs aberrantes :
  - Des valeurs négatives sont retrouvées dans certaines variables comme la surface
    - Remplacement de ces valeurs incohérentes par des *NaN*
  - Vérification des valeurs dupliquées
  - Suppression des données significativement ($p<0.05$) différentes de la valeur moyenne
    - Différence avec la moyenne > 1.96 * écart-type



Centre Baclesse
L'excellence pour vaincre votre cancer

Exploration des données

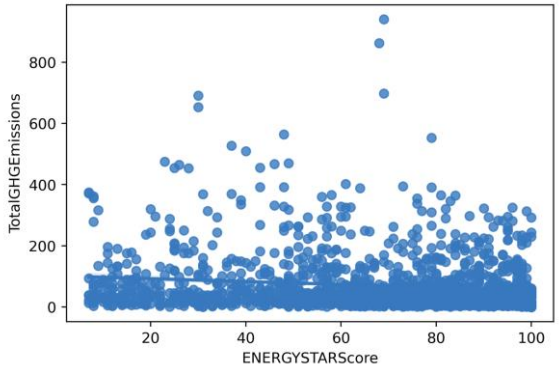
- Exploration globale des variables d'intérêt
  - On remarque qu'il y a eu beaucoup moins de construction pendant la 2nd guerre mondiale.
  - Il y a globalement plus de bâtiment avec un energystarsore important.



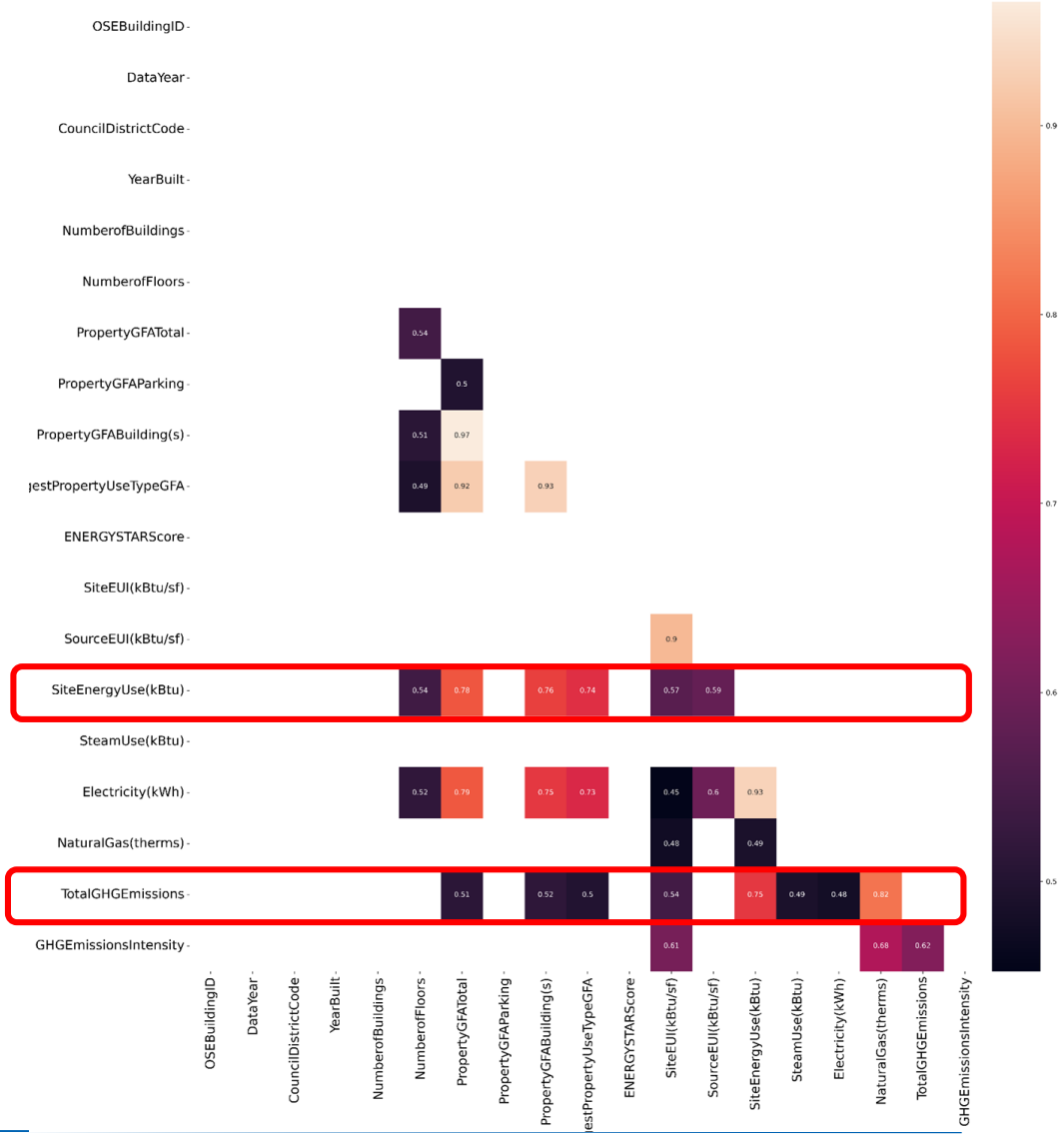
Exploration des données

- Etude des corrélations entre les variables
  - La **consommation en énergie** est principalement corrélée avec les surfaces
  - Les **émissions de GES** sont corrélées avec la consommation en énergie et en gaz

- l'**ENERGYSTARScore** n'est corrélé avec aucune variable

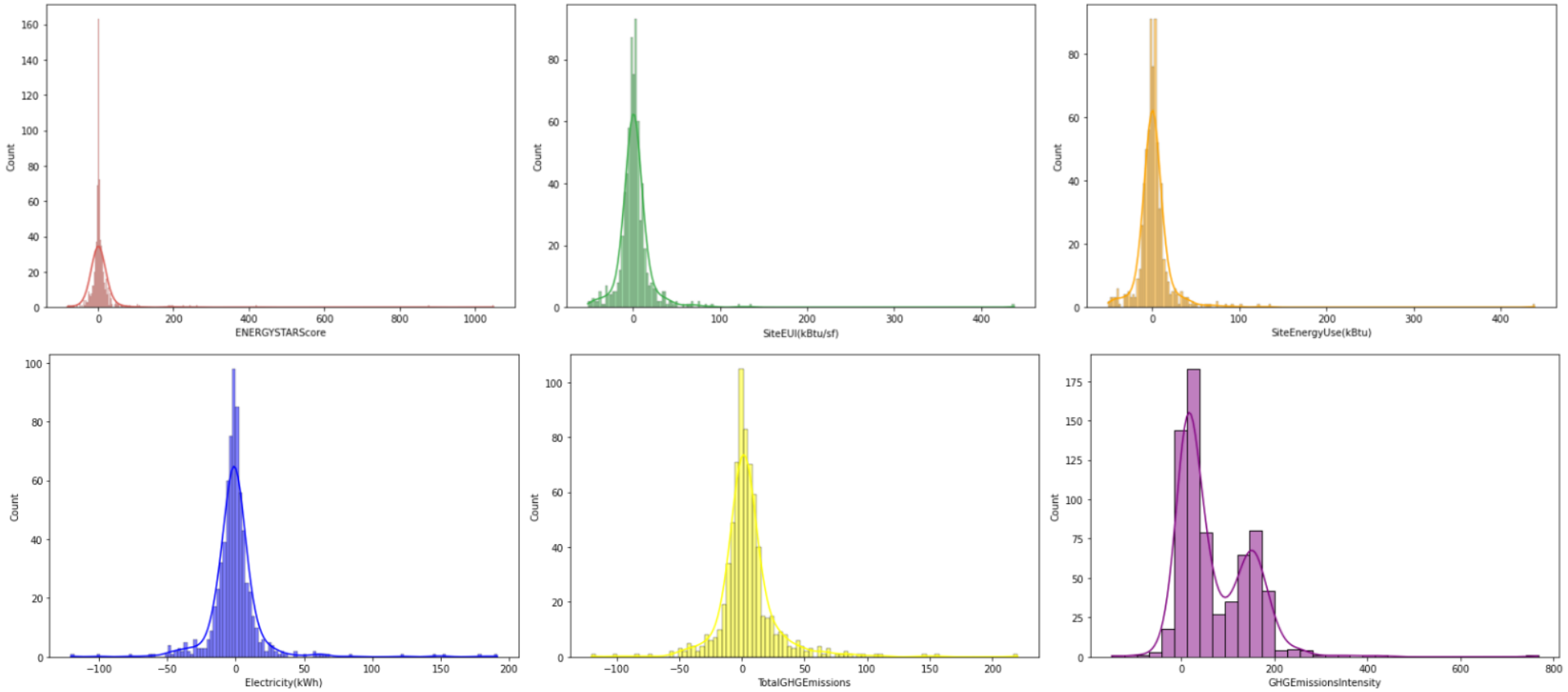


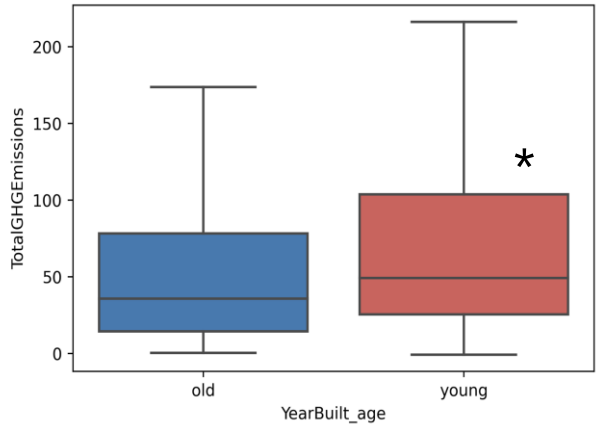
son calcul dépend de plusieurs variables et non pas d'une seule



Exploration des données

- Existe-t-il une différence entre 2015 et 2016
  - Une tendance à l'augmentation des émissions de GES
  - Les bâtiments récents émettent plus de GES





* p< 0.05

Features engineering

- Colonnes Dataframes crées :
  - *« GFA_per_floor »*
    - En plus de la surface totale et du nombre d'étage, la surface par étage pourrait être une information intéressante pour notre modèle
    - En utilisant les variables “*'PropertyGFATotal'*” et “*'NumberofFloors'*”
  - *« Building_age »*
    - Exprime l'âge du batiment directement, ce sera sûrement plus parlant que la date de construction qui ne varie que de 5% (1900 vs 2000)
    - En soustrayant la date de construction à 2022

Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation

- Features **utilisés** pour l'entrainement des modèles :

  - Features **quantitatifs** :
    - CouncilDistrictCode
    - NumberofFloors
    - PropertyGFATotal
    - PropertyGFAParking
    - PropertyGFABuilding(s)
    - LargestPropertyUseTypeGFA
    - GFA_per_floor
    - Building_age

  - Features **catégoriels** :
    - PrimaryPropertyType
    - Neighborhood
    - LargestPropertyUseType
    - YearBuilt_age

Modélisation

- Features **non utilisés** pour l'entrainement des modèles :

  - Car **non informatifs** :
    - YearBuilt
    - OSEBuildingID
    - GHGEmissionsIntensity
    - BuildingType
    - DataYear
    - ListOfAllPropertyUseTypes

  - Car **hors AAP** :
    - SiteEUI(kBtu/sf)
    - SourceEUI(kBtu/sf)
    - SteamUse(kBtu)
    - Electricity(kWh)
    - NaturalGas(therms)
    - ENERGYSTARScore

Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation

- Choix des modèles de régression étudiés :

  - Méthodes **linéaires** :
    - LinearRegression(),
    - Lasso()
    - Ridge()
    - ElasticNet()
    - TweedieRegressor()
    - HuberRegressor()
  - Méthodes **non-linéaires** :
    - SVR()
  - Méthodes **ensemblistes** :
    - XGBRegressor()
    - RandomForestRegressor()

Modélisation

- Performance des modèles - consommation en énergie

  - Méthodes **linéaires** :

LinearRegression()

Score entrainement	**0.63**
Score de prédiction	**0.70**
MAE	1448111
RMSE	2378088

Lasso()

Score entrainement	**0.63**
Score de prédiction	**0.70**
MAE	1448111
RMSE	2378088

HuberRegressor()

Score entrainement	**0.60**
Score de prédiction	**0.66**
MAE	1373031
RMSE	2421812

TweedieRegressor()

Score entrainement	**0.57**
Score de prédiction	**0.62**
MAE	1662786
RMSE	2558493

Les modèles linéaires offrent des performances similaires et moyennes

Centre Baclesse
L'excellence pour vaincre votre cancer

MAE : Mean Absolute Error / RMSE : Root Mean Square Error

Modélisation

- Performance des modèles - consommation en énergie

  - Méthodes **non linéaires** :

SVR()

Score entrainement	**-0.15**
Score de prédiction	**-0.14**
MAE	2503738
RMSE	4412233

  - Méthodes **ensemblistes** :

XGBRegressor()

Score entrainement	**0.85**
Score de prédiction	**0.74**
MAE	1285233
RMSE	2107071

RandomForestRegressor()

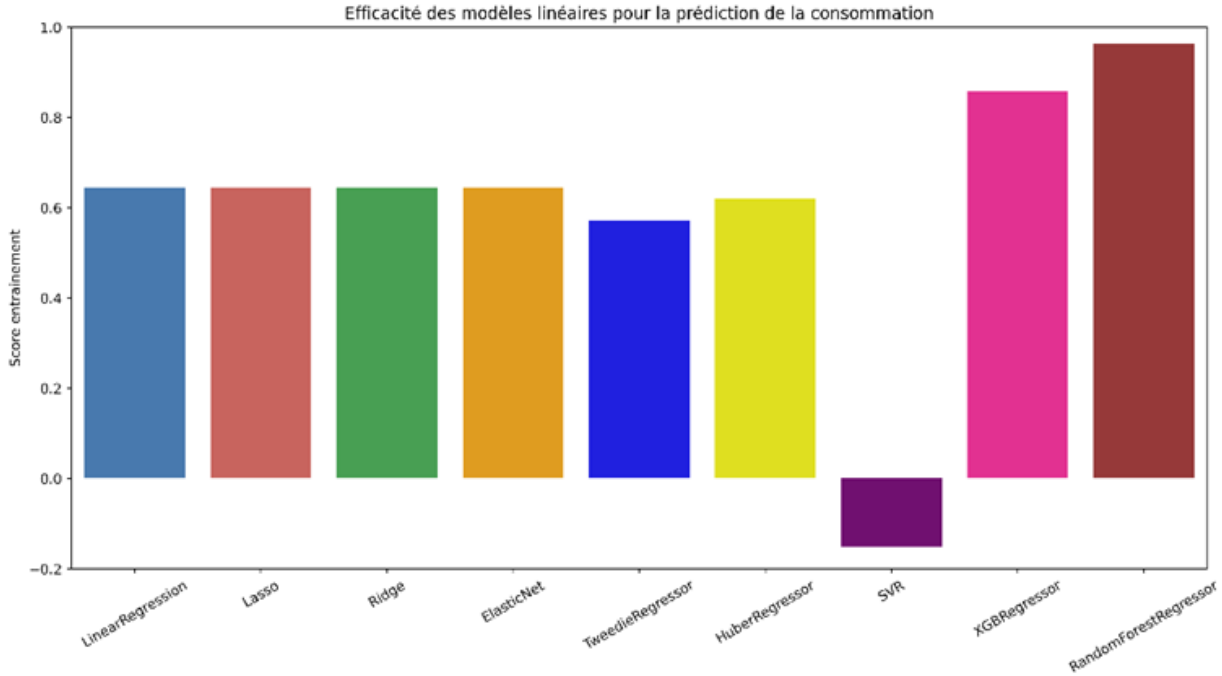
Score entrainement	**0.97**
Score de prédiction	**0.83**
MAE	9265988
RMSE	1719883

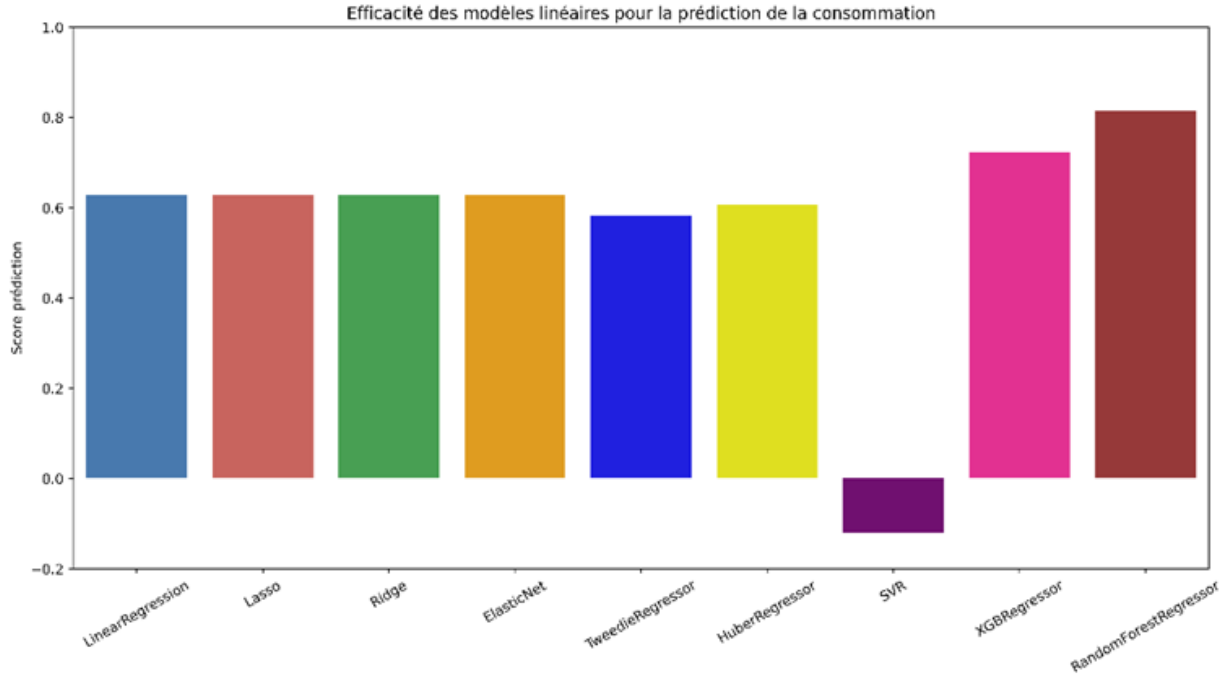
Les modèles ensemblistes sans optimisation préalable proposent des performances très intéressantes

Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation

- Performance des modèles - consommation en énergie





Pour la prédiction de la consommation en énergie, les modèles ensemblistes sont les plus performants

Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation

- Performance des modèles - émission des GES

  - Méthodes **linéaires** :

LinearRegression()

Score entrainement	**0.35**
Score de prédiction	**0.30**
MAE	50
RMSE	78

Lasso()

Score entrainement	**0.35**
Score de prédiction	**0.30**
MAE	49
RMSE	77

HuberRegressor()

Score entrainement	**0.19**
Score de prédiction	**0.20**
MAE	45
RMSE	83

TweedieRegressor()

Score entrainement	**0.27**
Score de prédiction	**0.24**
MAE	55
RMSE	81

Centre Baclesse
L'excellence pour vaincre votre cancer

De façon assez logique, les surfaces sont importantes, ensuite le nombre d'étage, l'âge du bâtiment et enfin le n° du district

Modélisation

- Performance des modèles - émission des GES

- Méthodes **non linéaires** :

SVR()

Score entrainement	**0.10**
Score de prédiction	**0.11**
MAE	40
RMSE	64

- Méthodes **ensemblistes** :

XGBRegressor()

Score entrainement	**0.71**
Score de prédiction	**0.52**
MAE	40
RMSE	64

RandomForestRegressor()

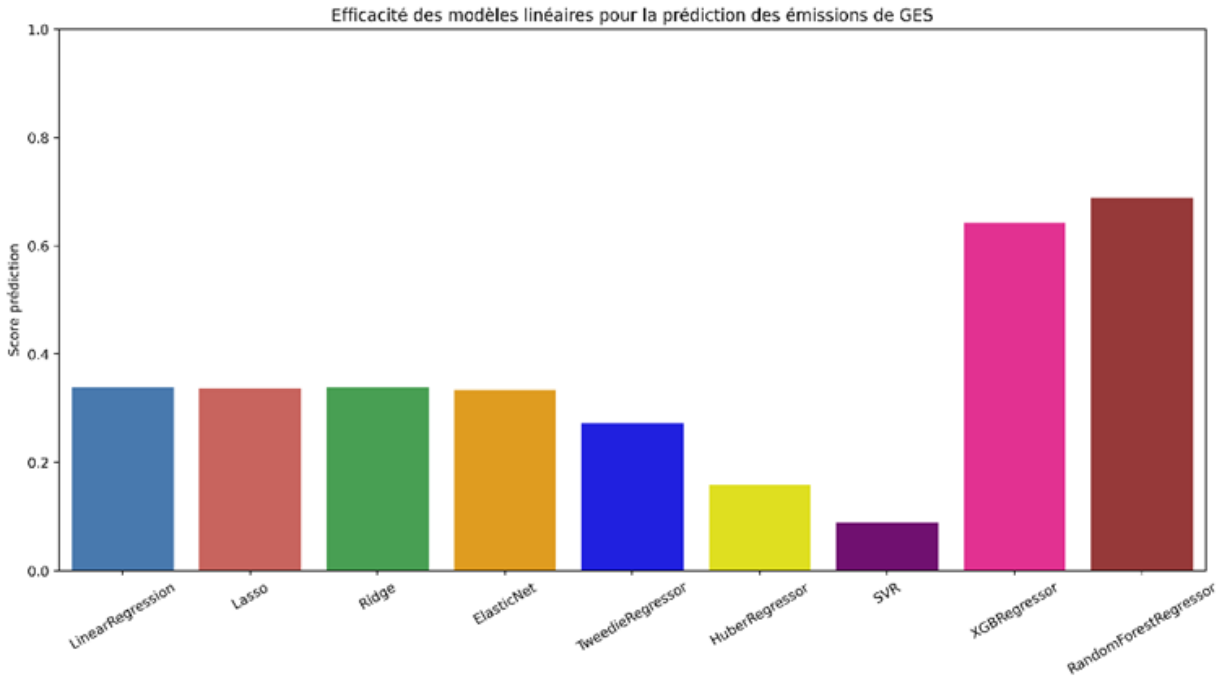
Score entrainement	**0.93**
Score de prédiction	**0.65**
MAE	30
RMSE	54

Le modèle XGB est intéressant, RandomForest est cependant le plus intéressant

Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation

- Performance des modèles - émission des GES

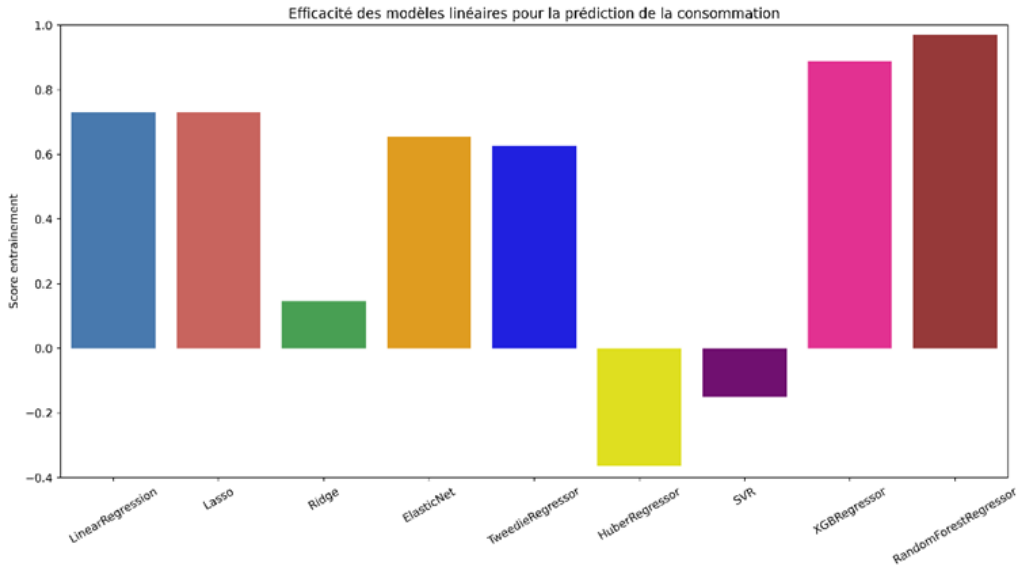


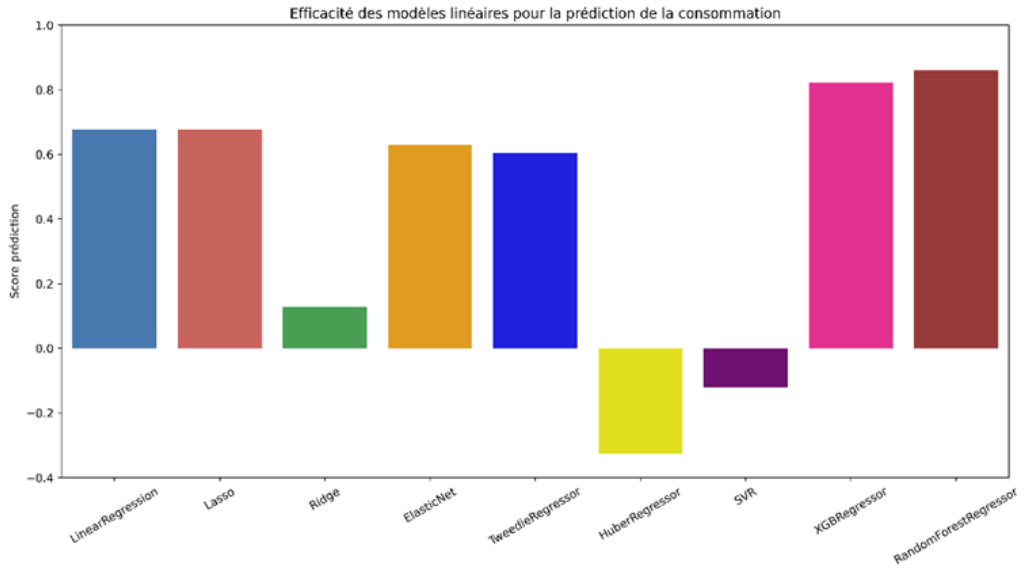
Pour la prédiction de la consommation en énergie, les modèles non linéaires et ensemblistes sont les plus performants

Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation

- Performance des modèles - intérêt des variables catégorielles ?
  - Prédiction de **la consommation** :



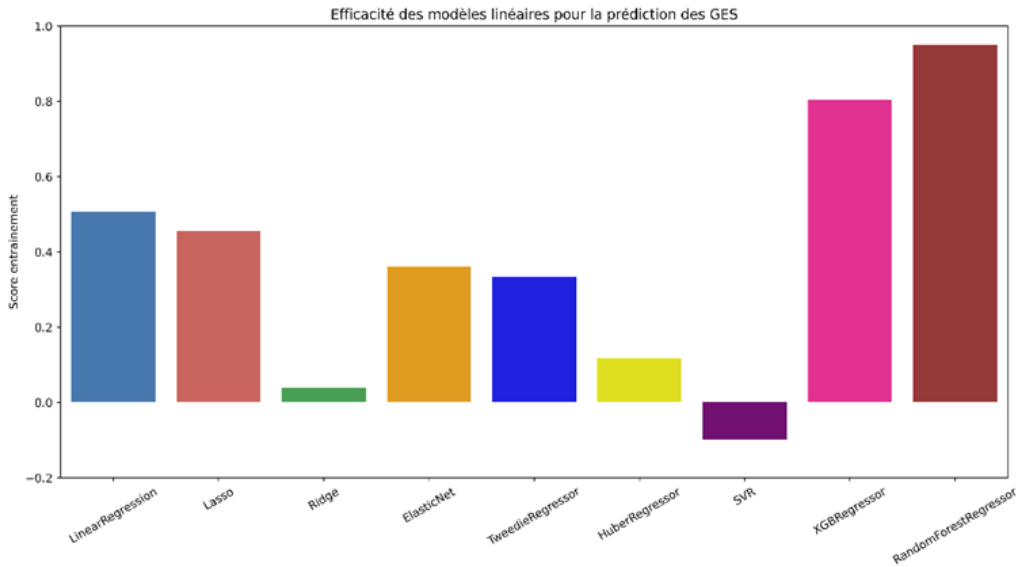


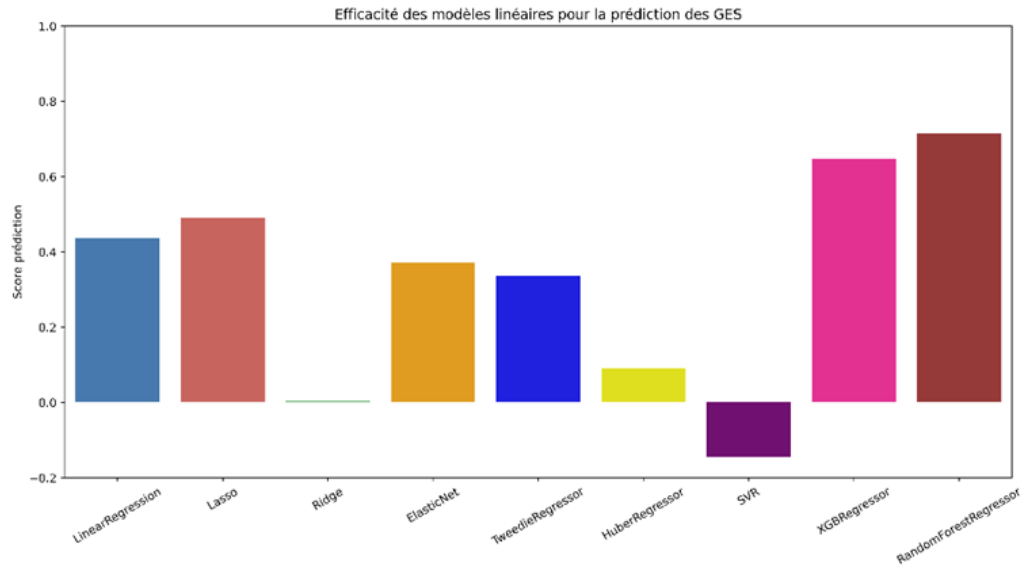
L'utilisation des variables catégorielles n'augmente pas la performance des modèles sur la prédiction de la consommation

Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation

- Performance des modèles - intérêt des variables catégorielles ?
  - Prédiction des **émissions de GES** :



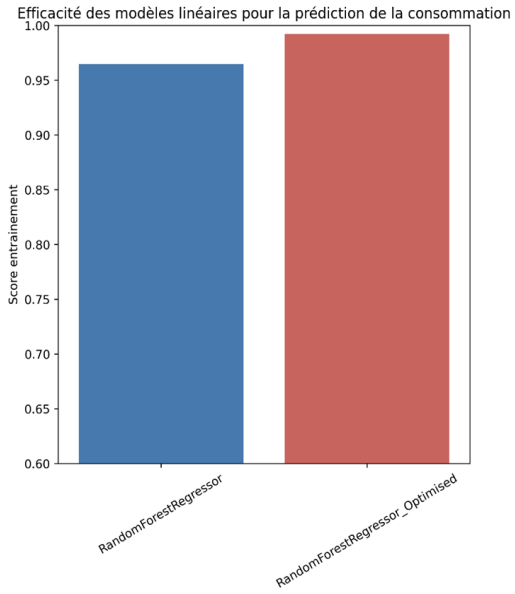


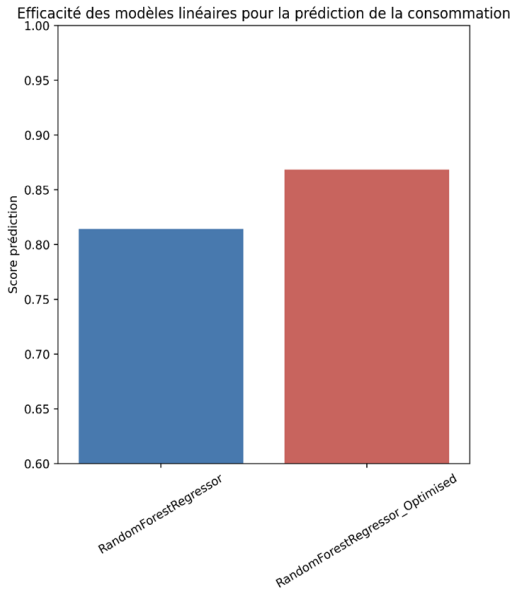
L'utilisation des variables catégorielles n'augmente pas la performance des modèles sur la prédiction des émissions des GES

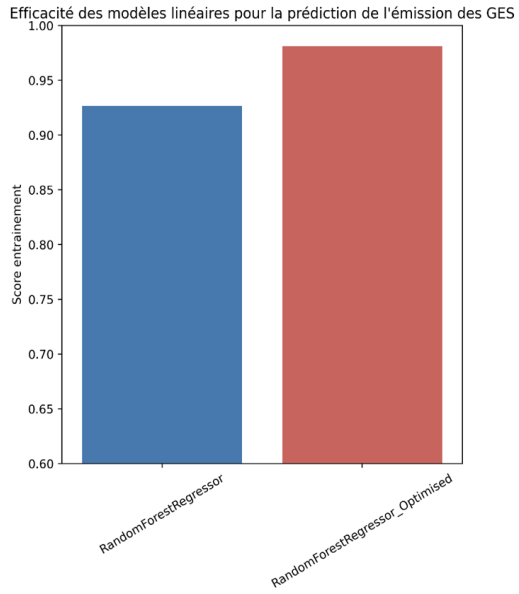
Centre Baclesse
L'excellence pour vaincre votre cancer

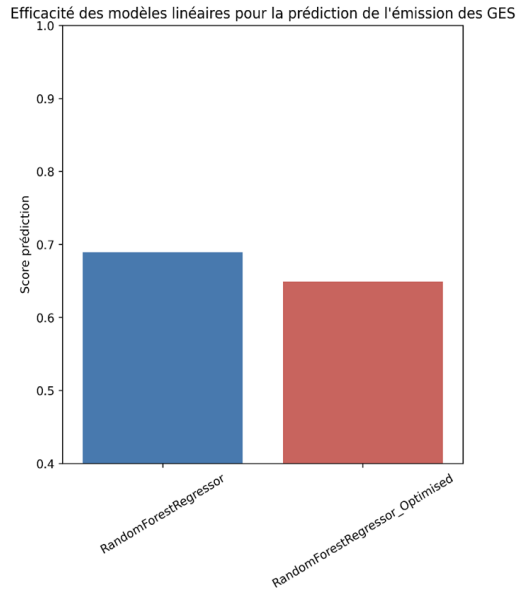
Optimisation du modèle choisi

- Optimisation de l'algorithme Random Forest (validation croisée et grid search)





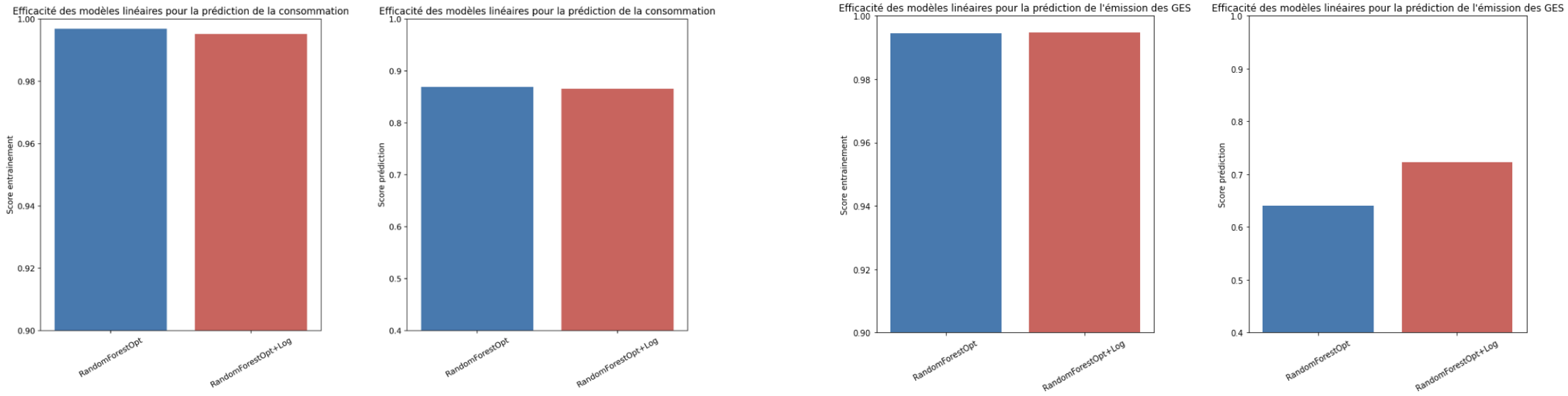




L'optimisation améliore l'algorithme pour la prédiction de la consommation.
Cependant, cette optimisation créée de l'overfitting pour la prédiction des émissions de GES et l'entrainement prend beaucoup de temps (environ 8min) pour un gain relativement faible (0.87 (optimisé) à la place de 0.81 (non-optimisé)).

Optimisation du modèle choisi

- La normalisation des données par un log améliore-t-elle les performances du modèle ?

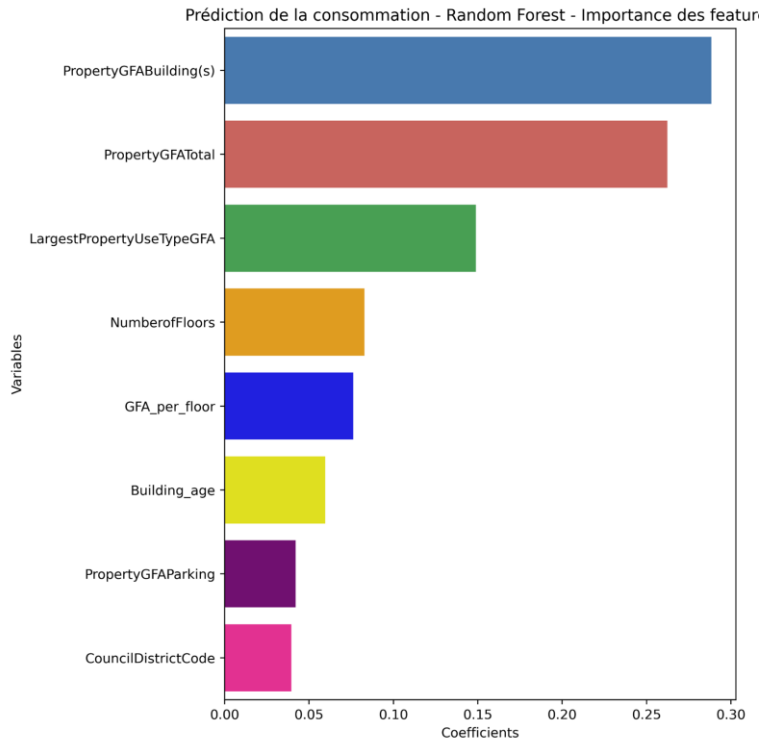


Alors qu'il ne semble pas avoir d'intérêt à normaliser les données de consommation par le log, cette opération améliore légèrement la prédiction d'émissions de GES

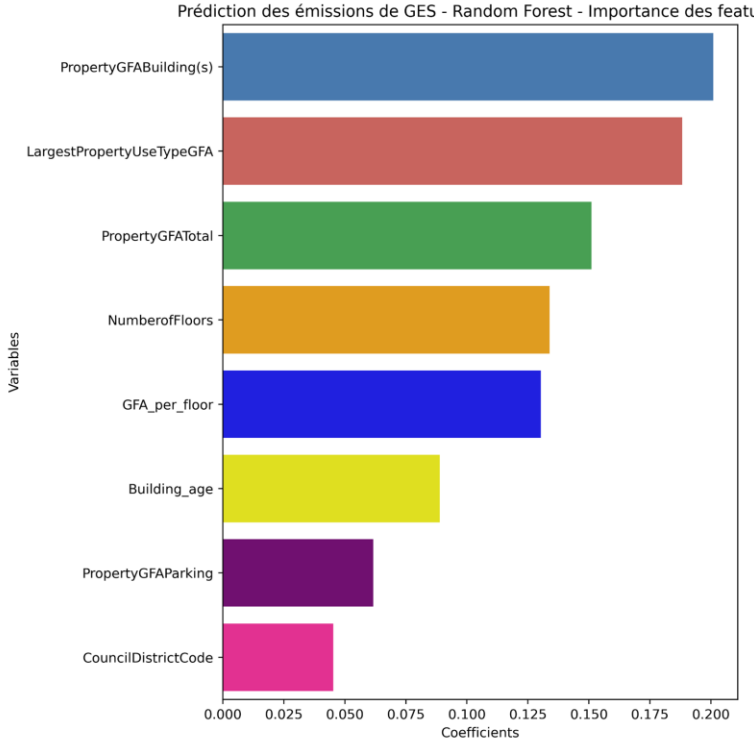
Centre Baclesse
L'excellence pour vaincre votre cancer

Importance des features dans les modèles prédictifs :

Émissions de la consommation :



Émissions de GES :



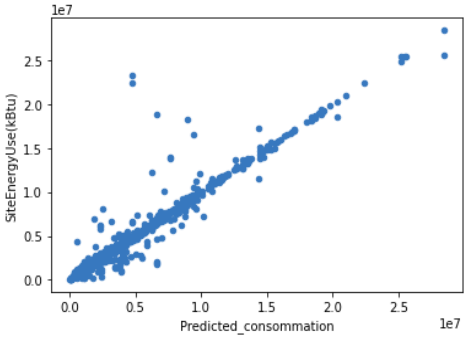
De façon assez logique, les surfaces sont importantes, ensuite le nombre d'étage, l'âge du bâtiment et enfin le n° du district

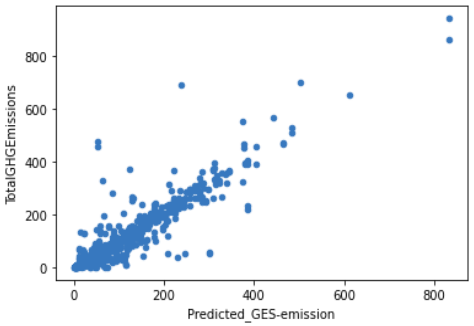
Centre Baclesse
L'excellence pour vaincre votre cancer

CONCLUSION

- Rappel de la problématique :
  - Créer un modèle capable de prédire la consommation en énergie et en émission de GES à partir d'informations structurelles (surface du logement, nombre d'étage…) plutôt que sur des données coûteuses à obtenir (relevés de compteurs…)

- Résultats :
  - Le meilleur modèle permet une bonne prédiction de ces variables
  - Le modèle choisi correspond à un modèle de type ensembliste ***RandomForest*** avec optimisation par validation croisée et recherche des meilleurs hyperparamètres
  - Une transformation log des données d'émission de GES permet une optimisation des prédictions du modèle





Centre Baclesse
L'excellence pour vaincre votre cancer

Projet 3 - Anticipez les besoins en consommation électrique de bâtiments

Aurélien Corroyer-Dulmont, PhD

Ingénieur imagerie médicale

Centre Baclesse
L'excellence pour vaincre votre cancer