Anticipez les besoins en consommation électrique des bâtiments

Parcours Data Scientist - *OPENCLASSROOMS*

Plan de la présentation

1. Exploration des données
    a. Présentation de la problématique
    b. Nettoyage des données
    c. Transformation des variables
2. Modélisation des données
    a. Démarche effectuée
    b. Les différentes modélisations
    c. Etude de la variable l'EnergySTARScore
3. Conclusion

Exploration des données

Présentation de la problématique

Présentation de la problématique

- Prédire les consommations d'énergie et d'émissions en CO2 sur les années à venir pour la ville de Seattle
- Évaluer l'importance de la feature EnergyStarScore
- Existence de deux relevés (2015 et 2016) + description des colonnes
- En 2015 on a 3340 lignes et 47 colonnes, en 2016 : 3376 lignes et 46 colonnes
- Approche choisie :
    - Avoir les mêmes colonnes en 2015 et 2016
    - Faire les même transformations sur les deux données
    - Faire la modélisation sur 2016 et tester les résultats sur 2015

Exploration des données

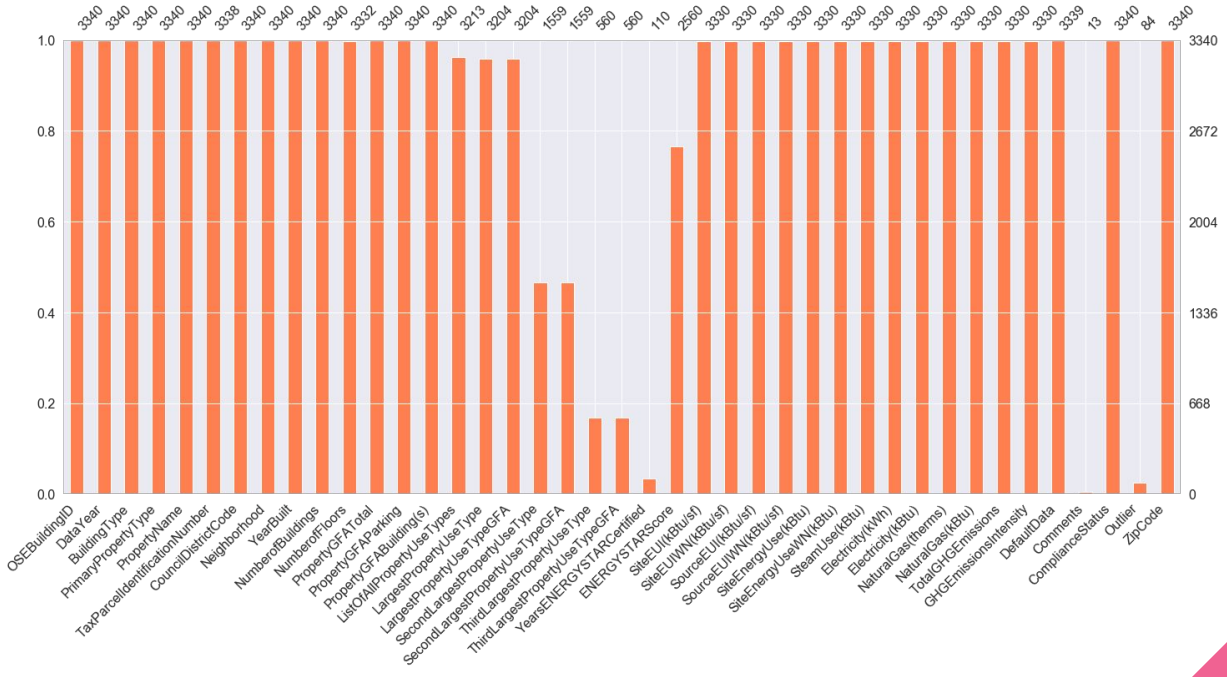
Nettoyage des données

Nettoyage des données

- Nombre de colonnes communes -> 37 :
  - Utilisation des différentes sources d'énergie
  - Différentes données sur la propriété
  - Informations sur la donnée collectée
- Les colonnes non communes:
  - 4 colonnes qui ont été renommées
  - 6 seulement en 2015: représentent des informations redondantes ou non intéressantes
  - 5 seulement en 2016: idem

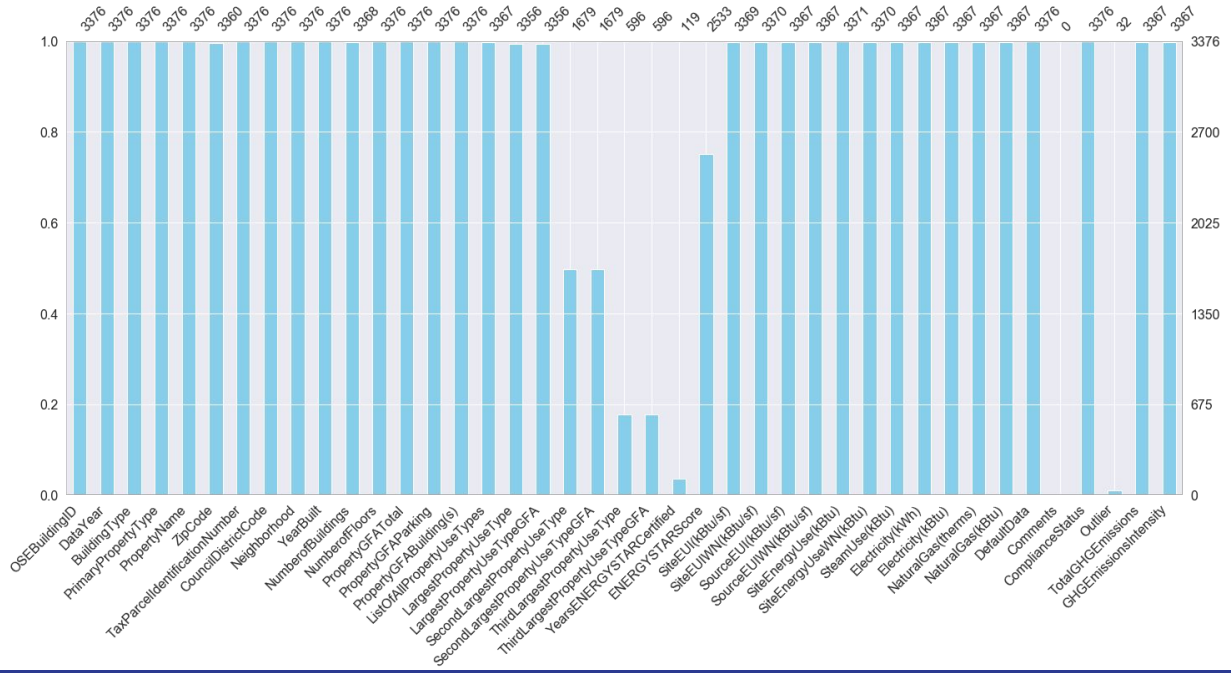
Nettoyage des données

- Données 2015 :



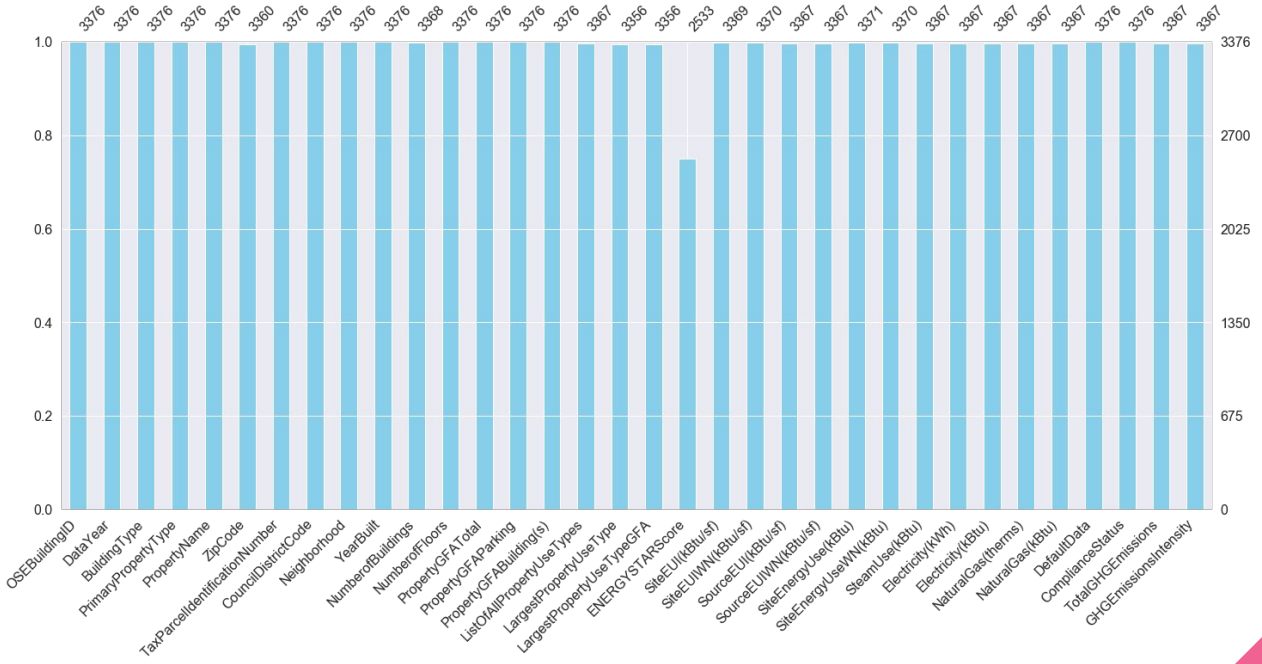
Nettoyage des données

- Données 2016 :



Nettoyage des données

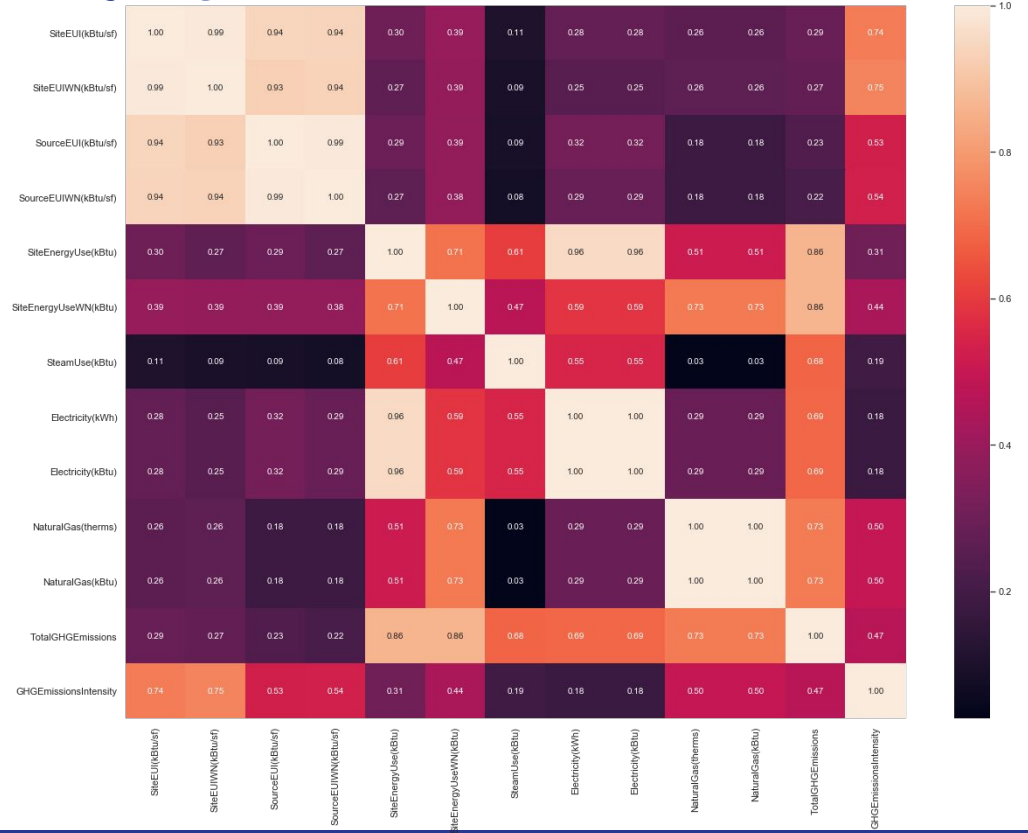
- Enlever les colonnes remplies en dessous de 50%



Nettoyage des données

- Les colonnes servant pour identifier la donnée 'OSEBuildingID', 'TaxParcelIdentificationNumber', 'PropertyName'
- Les informations redondantes 'ListOfAllPropertyUseTypes',
- Remplacer la colonne ‘YearBuilt’ par ‘Building Age’
- Enlever les données erronées (non compliant, ou compliance à Error)

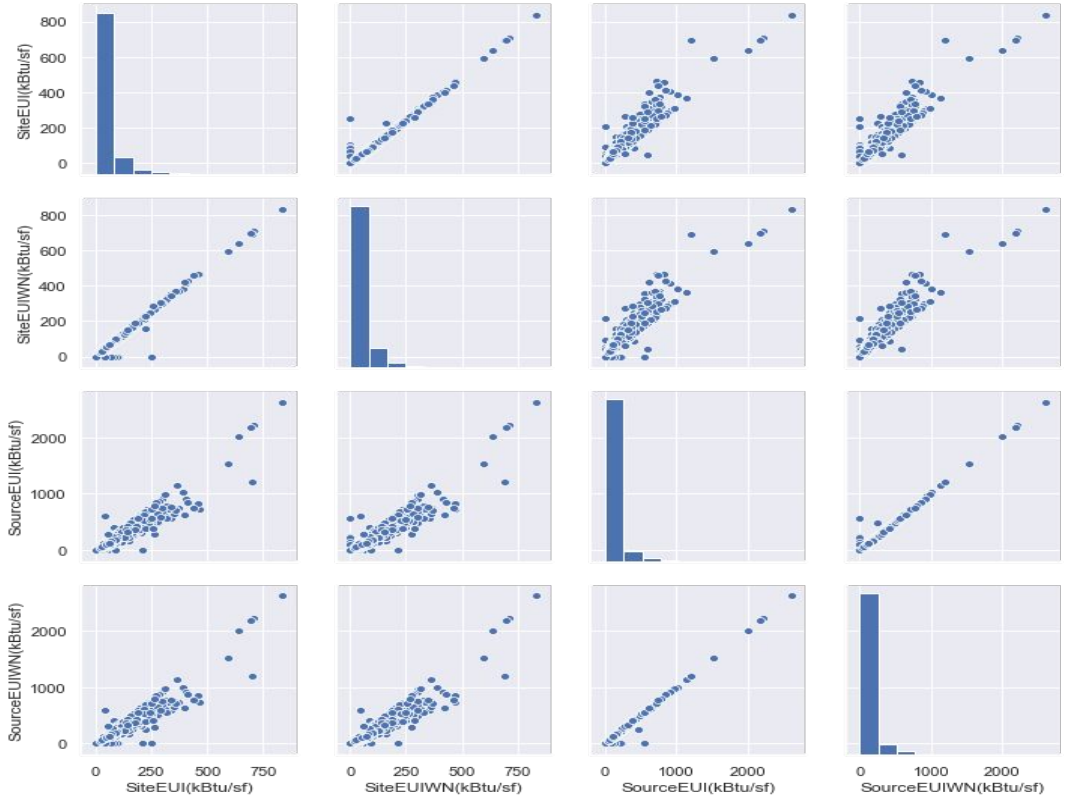
Nettoyage des données



- Certaines features en rapport avec l'énergie sont très corrélées.

- D'après la définition des colonnes ce sont bien des information redondantes

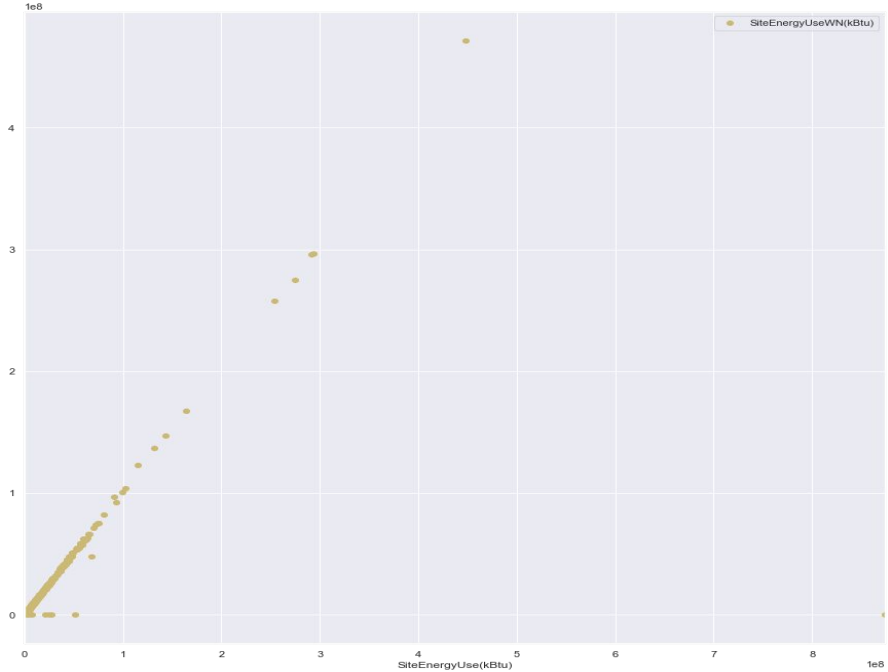
Nettoyage des données



- Le test du Chi-2 donne une p-value = 0 pour chaque couple de features d'intensité énergétique

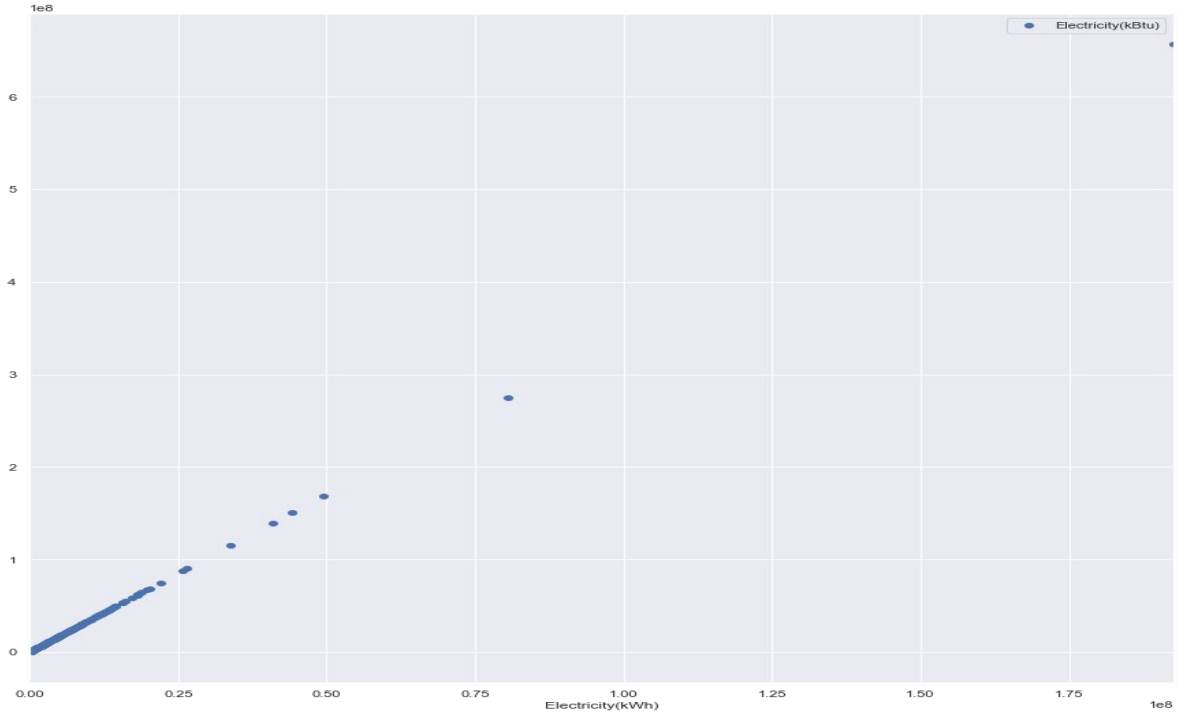
Nettoyage des données

- De même pour l’énergie



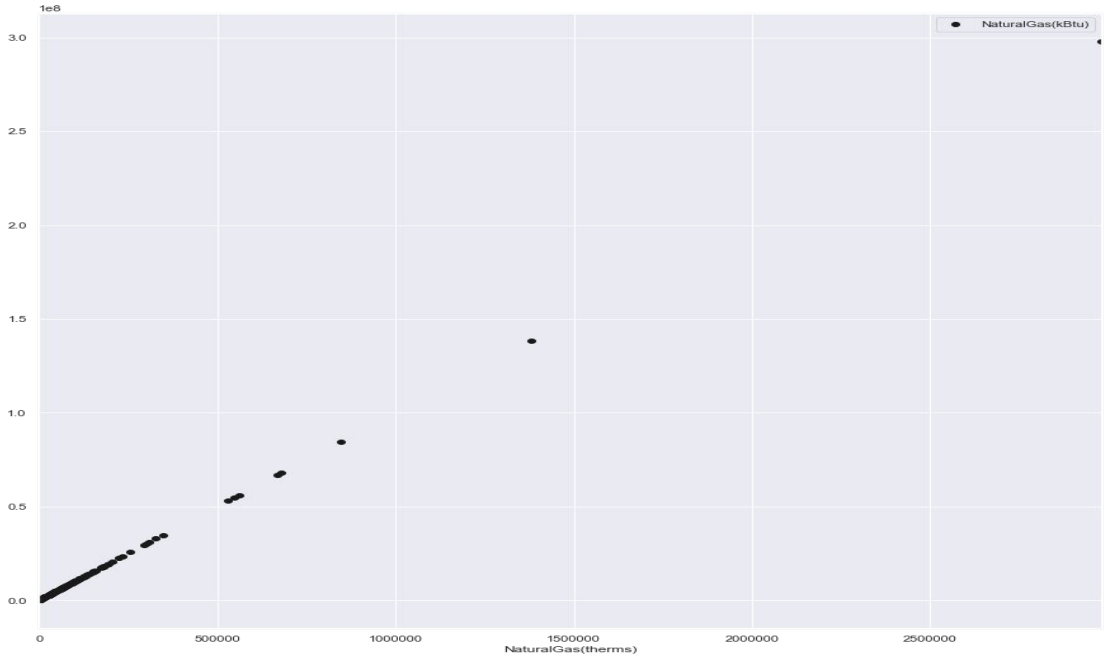
Nettoyage des données

- …. l’éléctricité

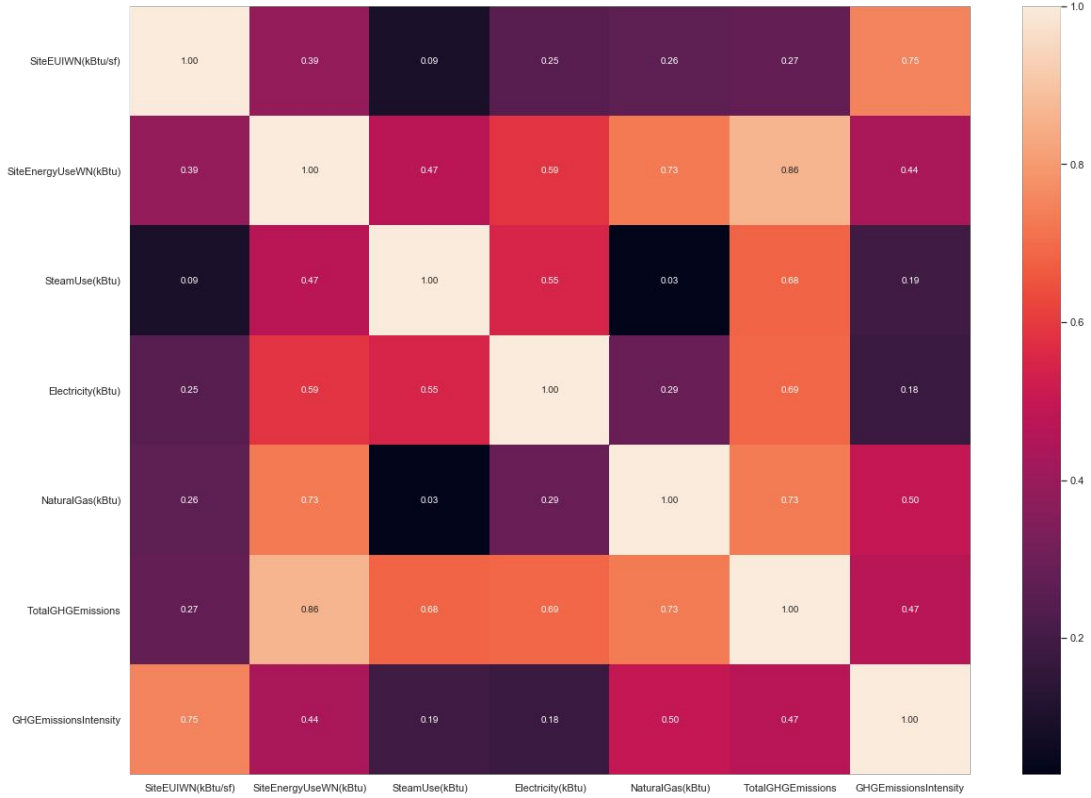


Nettoyage des données

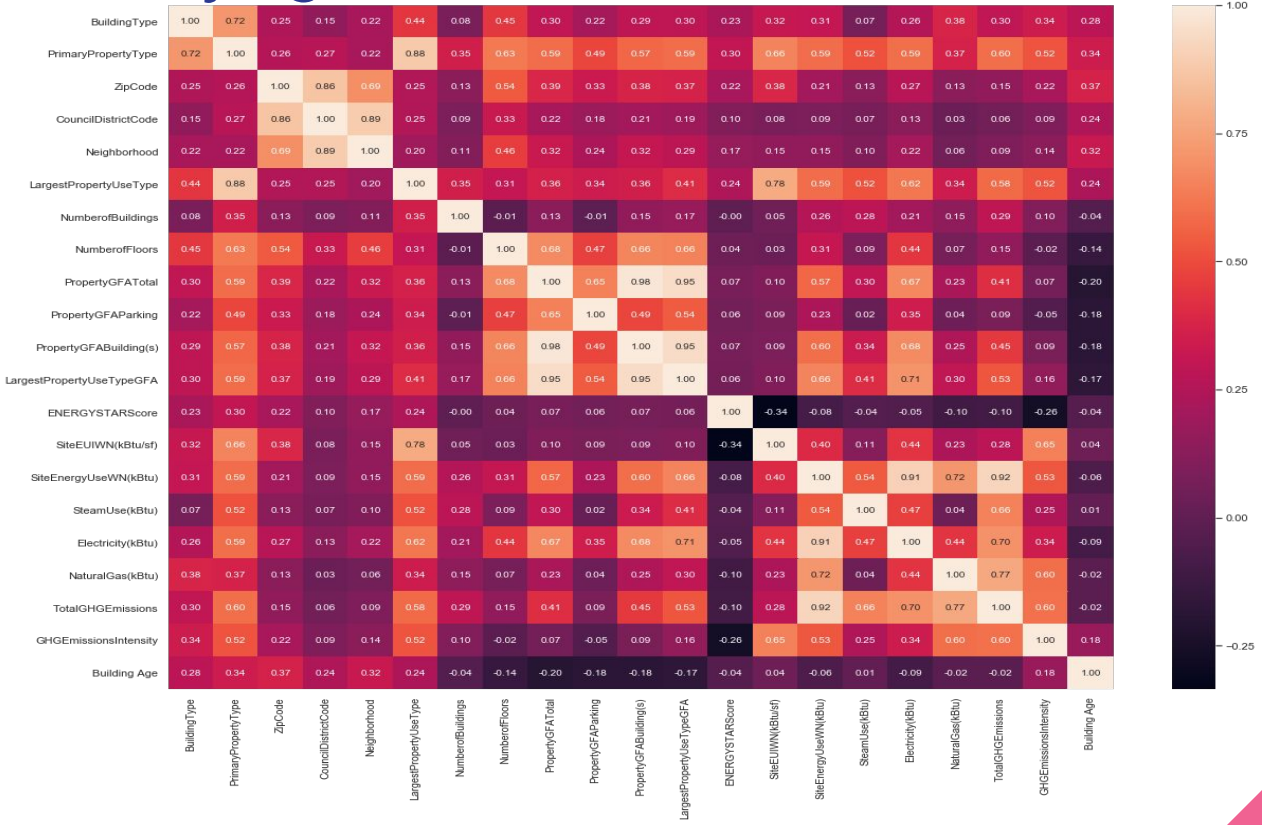
- Et le gaz naturel



Nettoyage des données



Nettoyage des données



- Certaines variables propres au bâtiment sont aussi corrélées

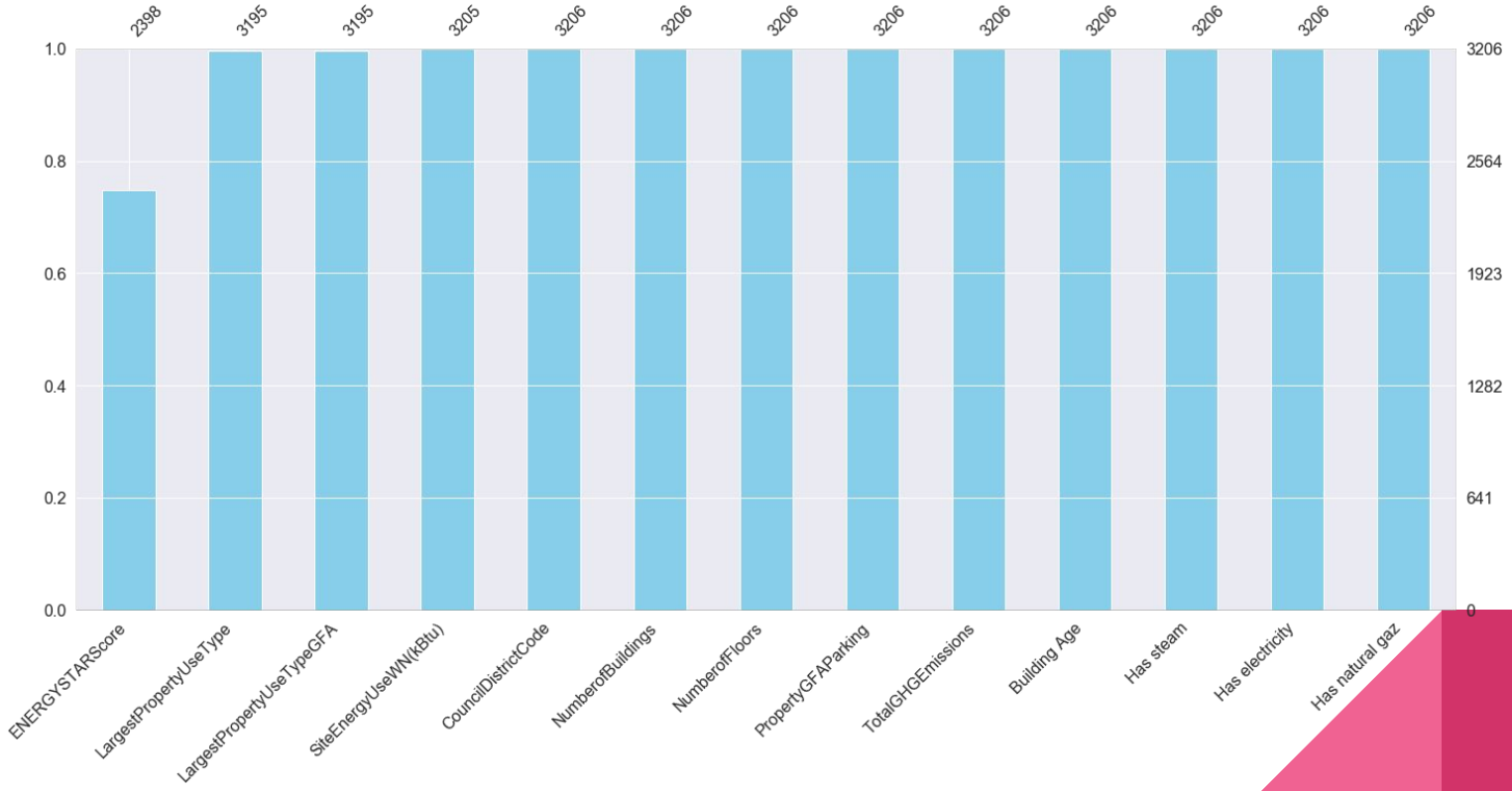
Exploration des données

Transformation des données

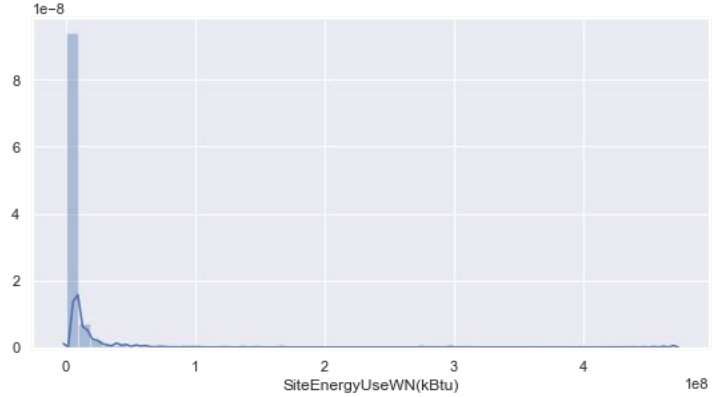
Transformation des données

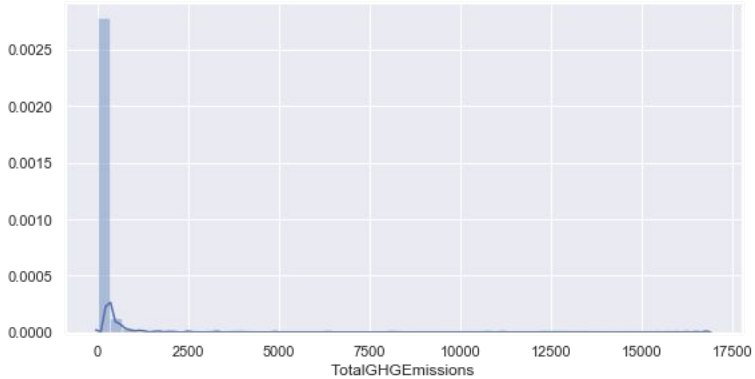
- Transformation de l’utilisation de natural gaz, d'électricité, ou de vapeur en booleans
- Transformation des colonnes objet en variables catégorielles
- Certaines données sont mal réparties, donc on les passe au log
- Et finalement on standardise nos données en utilisant un StandardScaler

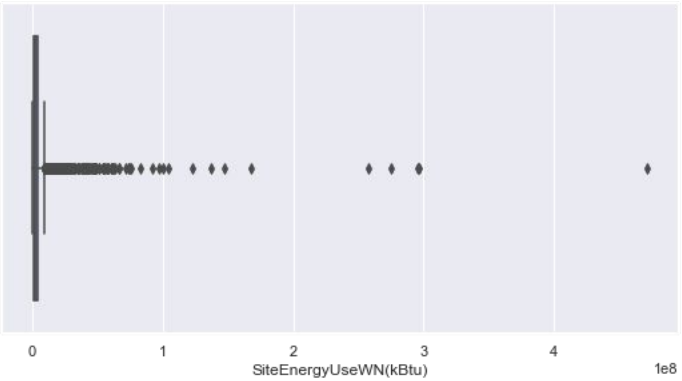
Transformation des données

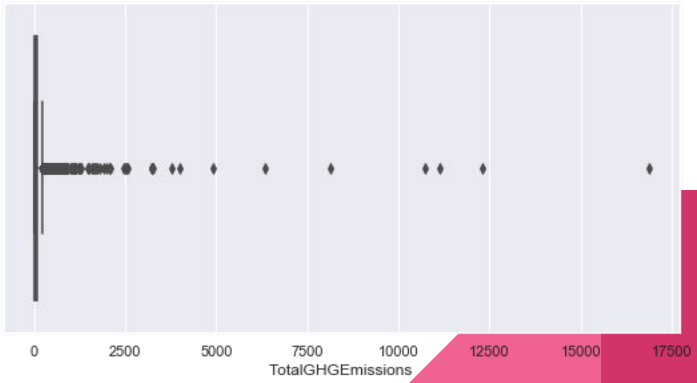


Transformation des données

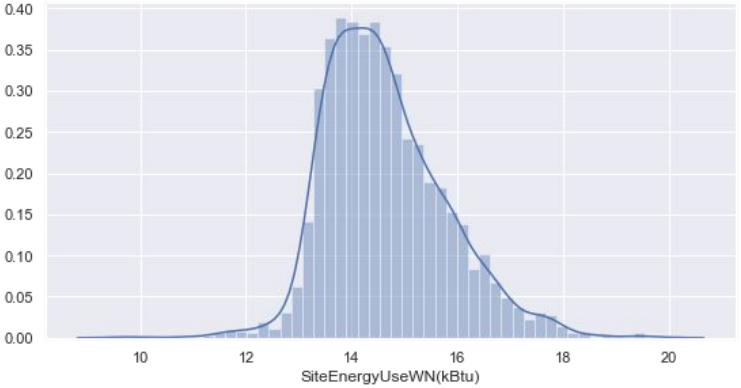


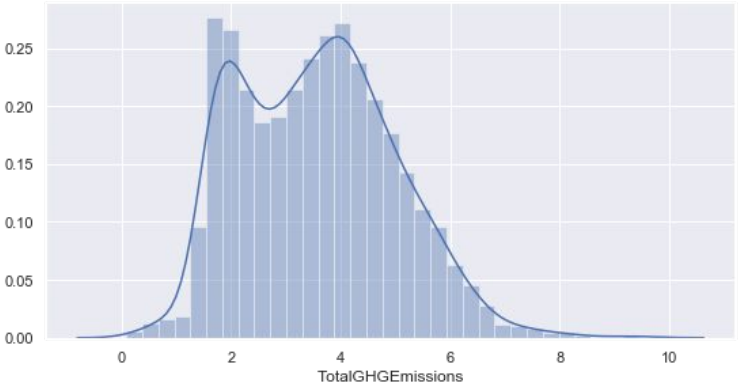


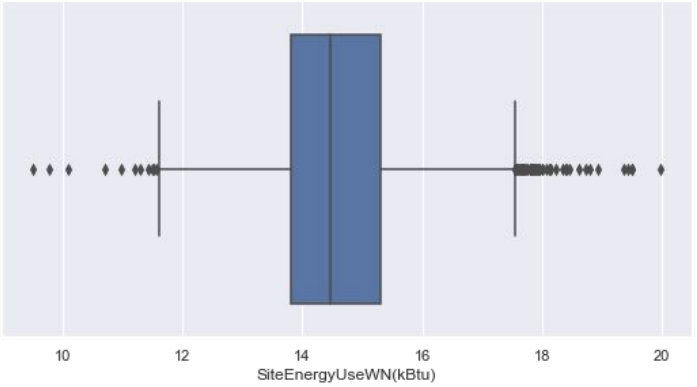


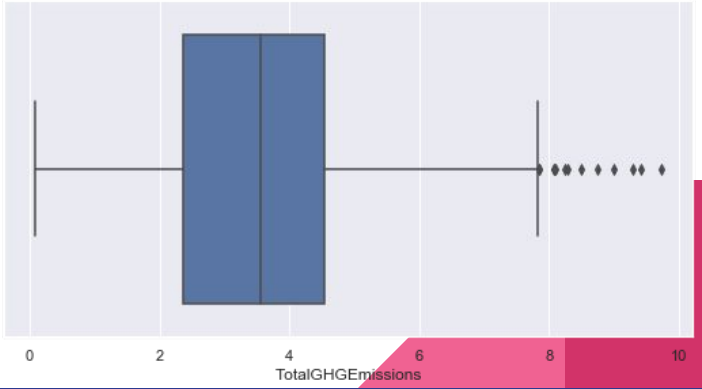


Transformation des données

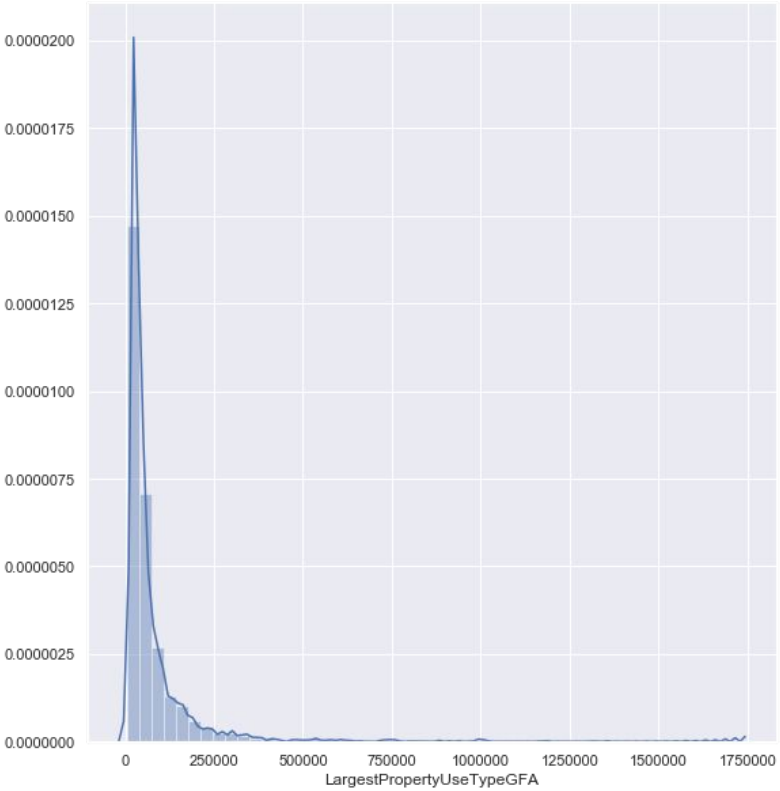




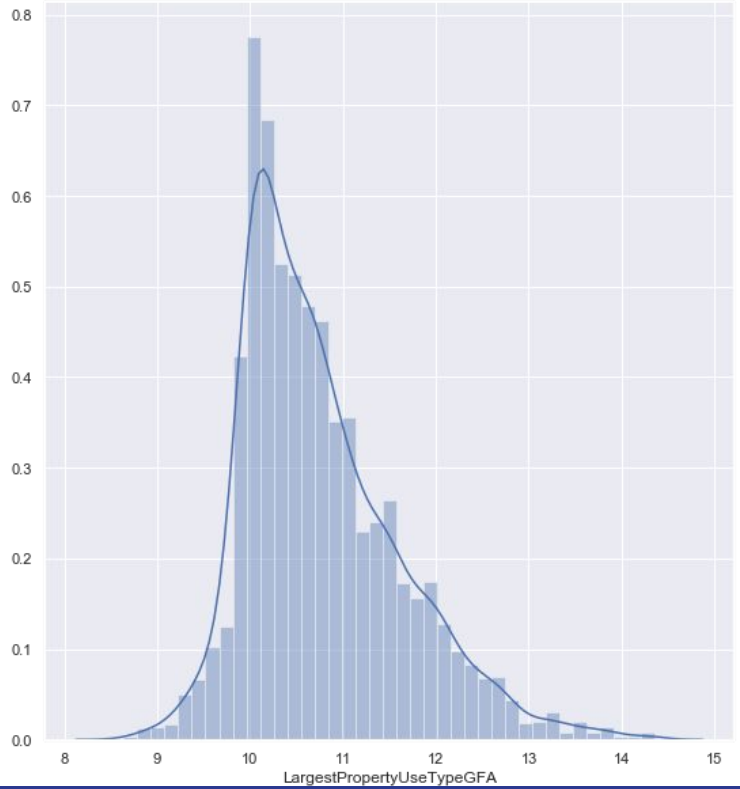




Transformation des données



Transformation des données



Modélisation des données

Démarche effectuée

Démarche effectuée

- 2 colonnes target: TotalGHGEmissions et SiteEnergyUseWN(kBtu)
- Choix d’une baseline pour chacune des valeurs target
- Test de différents modèles
- Test du modèle sur les données 2015
- intégration de la variable EnergySTARScore
- Test avec la nouvelle variable sur les données 2015

Modélisation des données

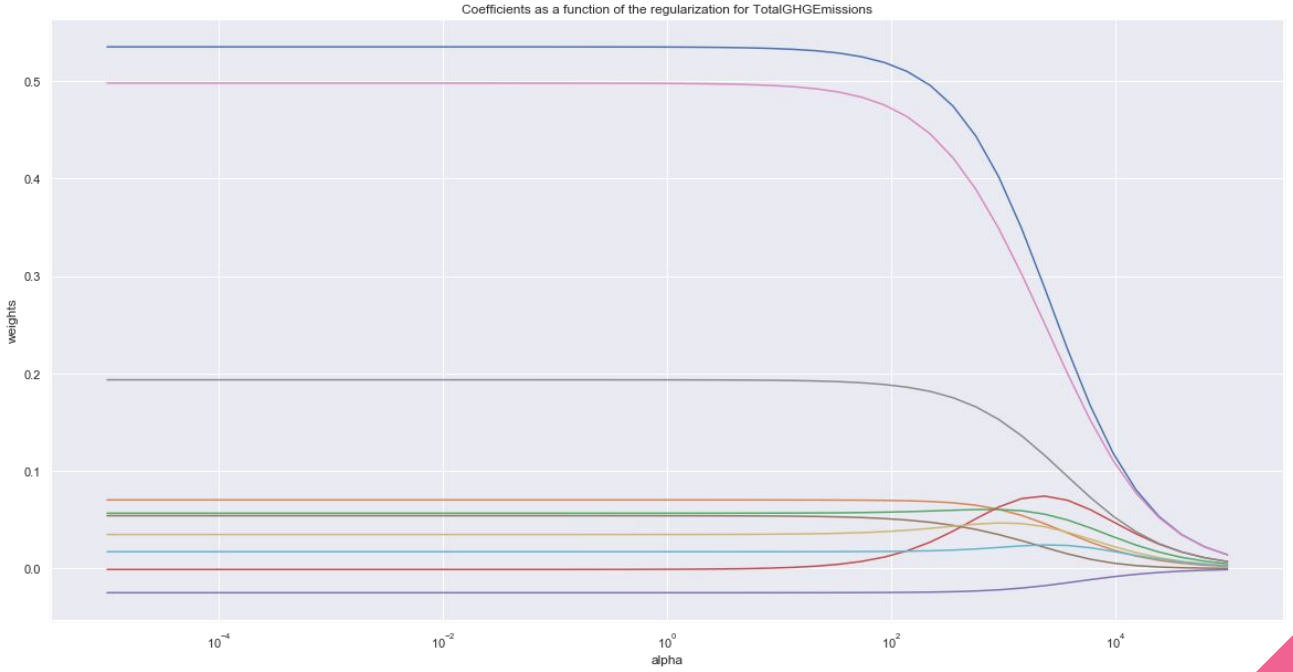
Les différentes modélisations

Les différentes modélisations :baseline

- Régression linéaire :
  - Pour les émissions de gaz :
    - MSE 0.27
    - RMSE 0.52
    - Testing Score 0.73
    - Training Score 0.71
    - Cross val RMSE: 0.54 +/- 0.02
  - Pour les consommation d’énergie :
    - MSE 0.31
    - RMSE 0.56
    - Testing Score 0.67
    - Training Score 0.67
    - Cross val RMSE: 0.58 +/- 0.02

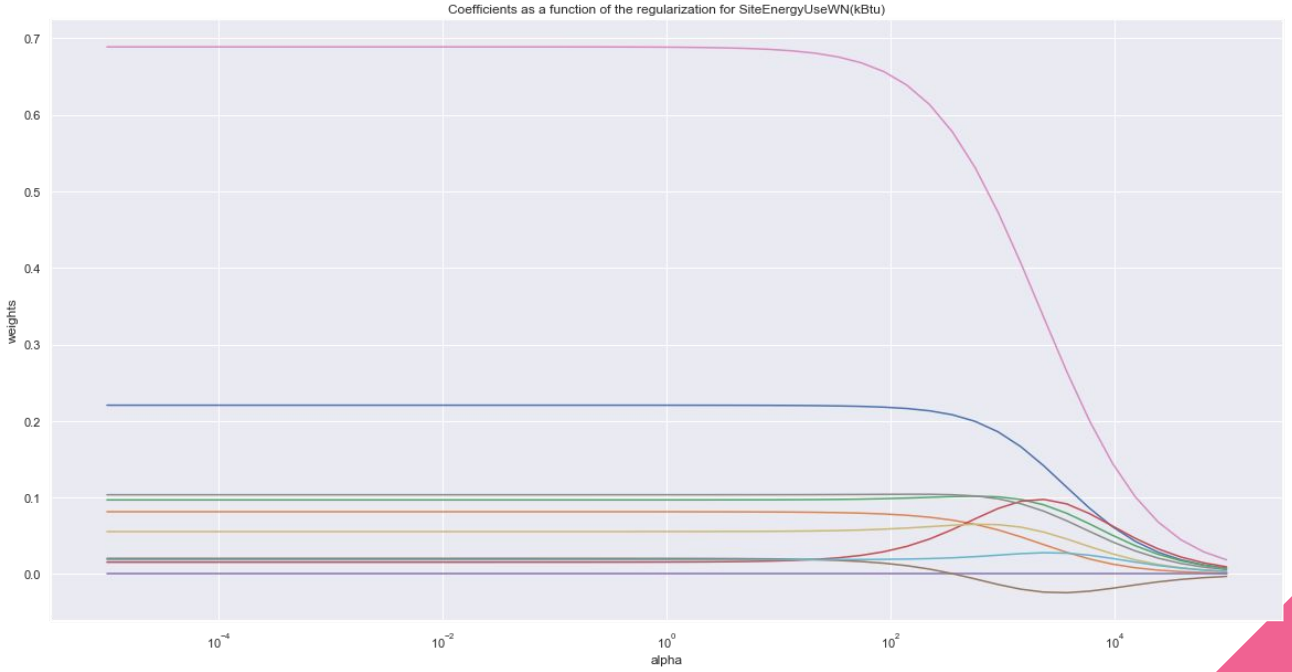
Les différentes modélisation : Ridge

- Pour les émissions de gaz, alpha = 0.52, score=0.69



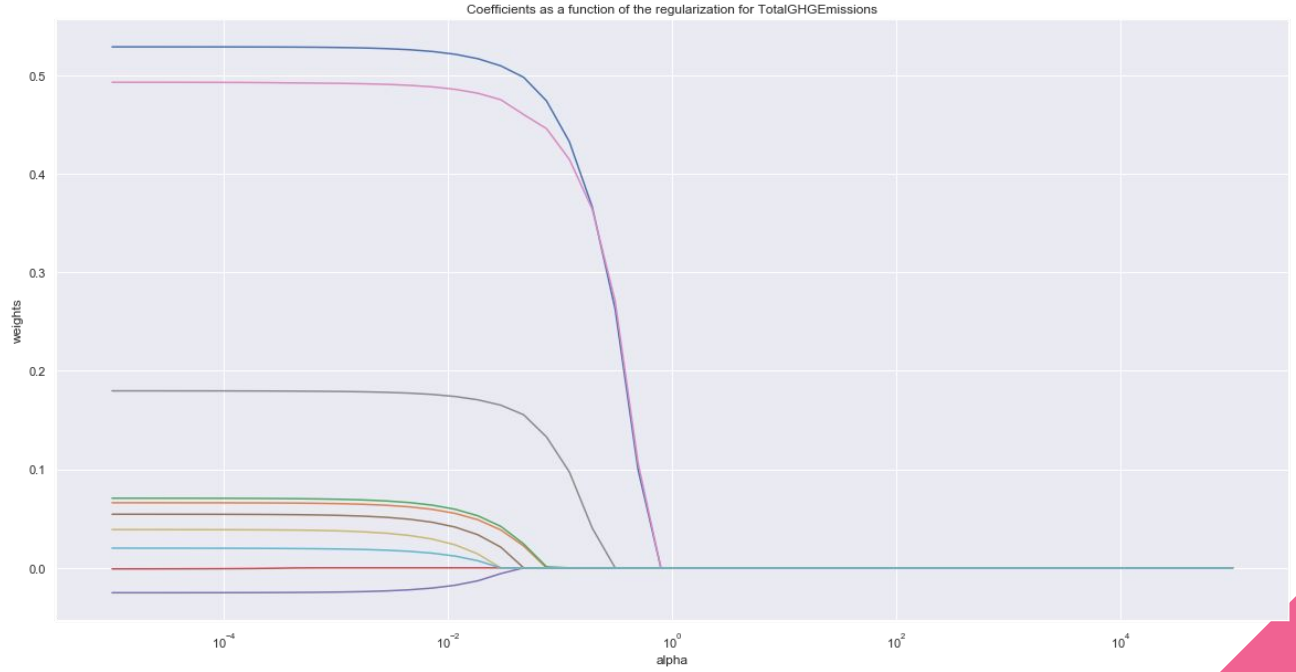
Les différentes modélisation : Ridge

- Pour la consommation d'électricité alpha = 138.95 et score = 0.65



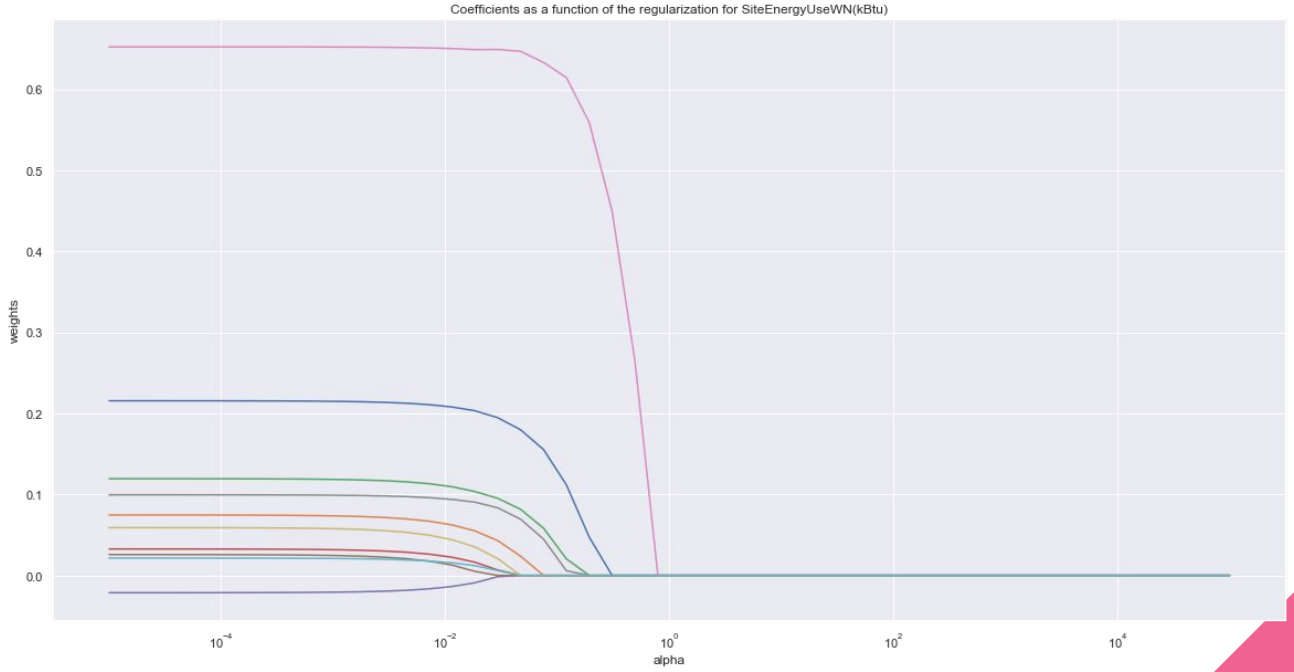
Les différentes modélisation : Lasso

- Pour les émissions de gaz alpha = 1e-5 et score = 0.70



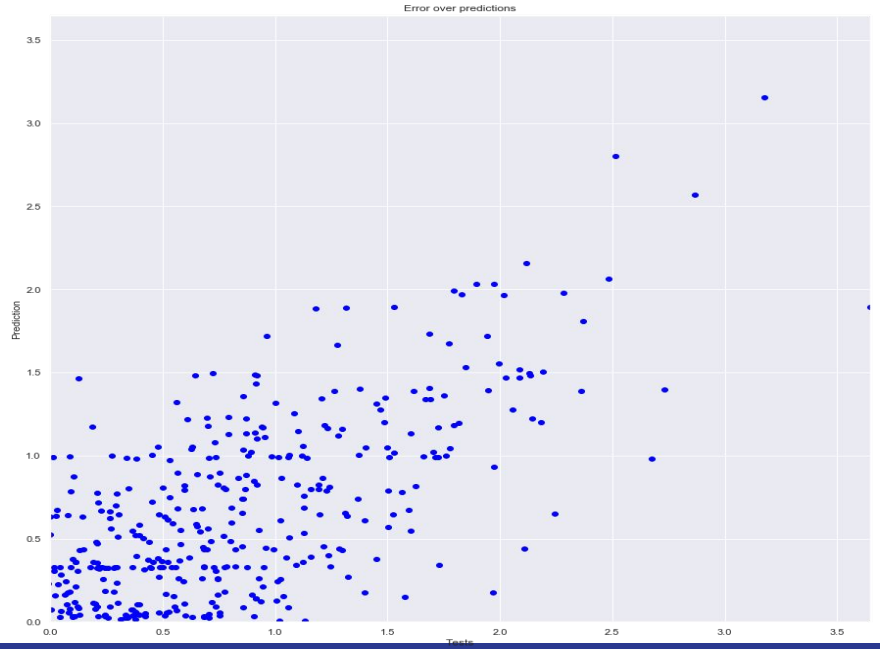
Les différentes modélisation : Lasso

- Pour la consommation d'énergie alpha = 0.005 et score = 0.65

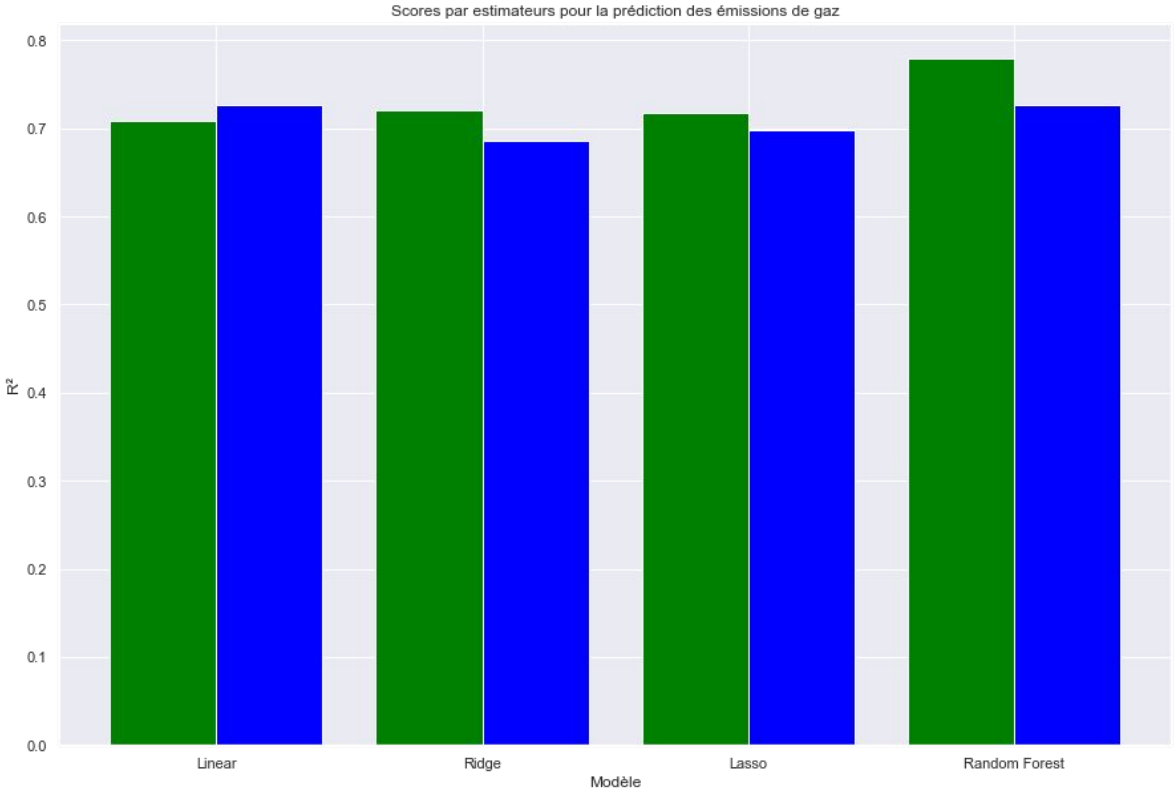


Les différentes modélisations : Random Forest

- Pour les émissions de gaz : score = 0.73, training score= 0.78
- {'bootstrap': True, 'max_depth': 5, 'min_samples_leaf': 3, 'min_samples_split': 3, 'n_estimators': 200}

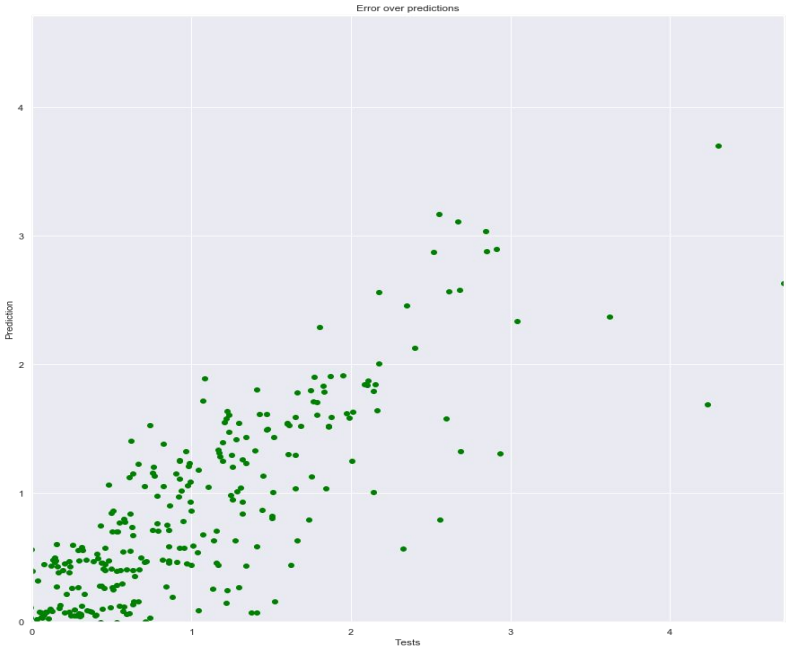


Les différentes modélisations : Random Forest

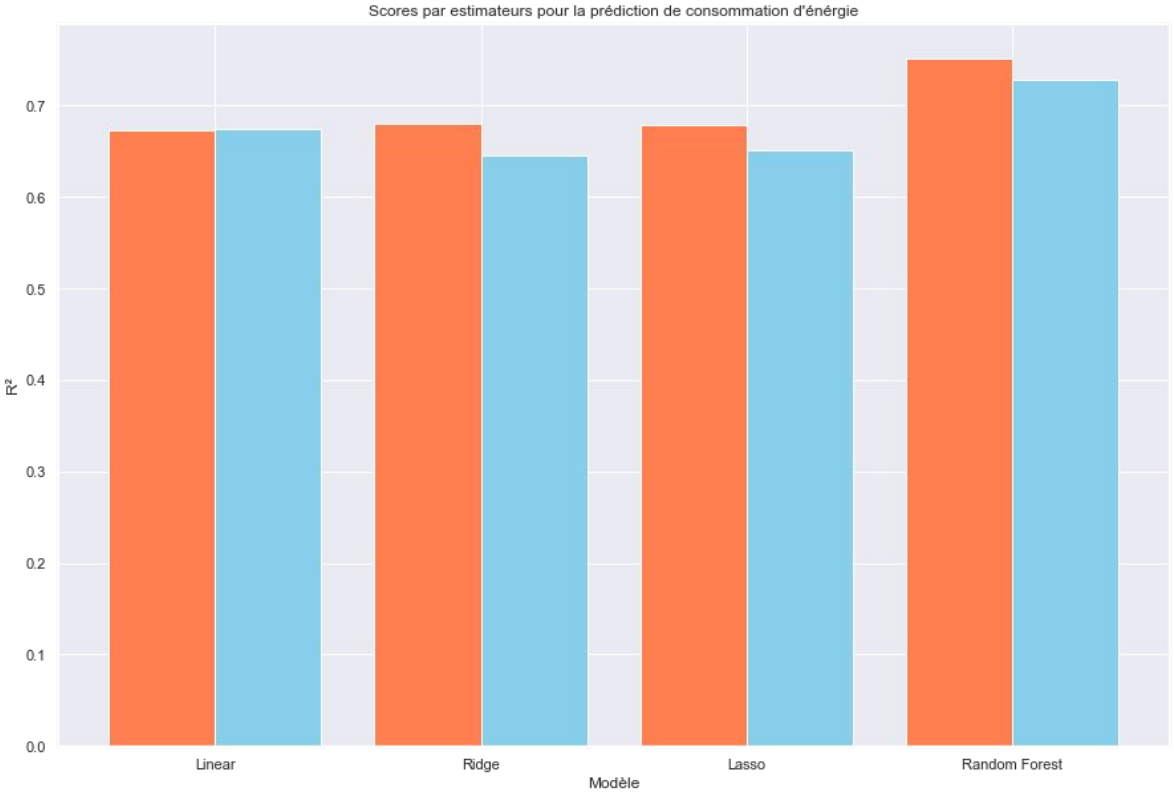


Les différentes modélisations : Random Forest

- Pour la consommation d'énergie : score = 0.73, training score=0.75
- {'bootstrap': True, 'max_depth': 5, 'min_samples_leaf': 3, 'min_samples_split': 3, 'n_estimators': 200}

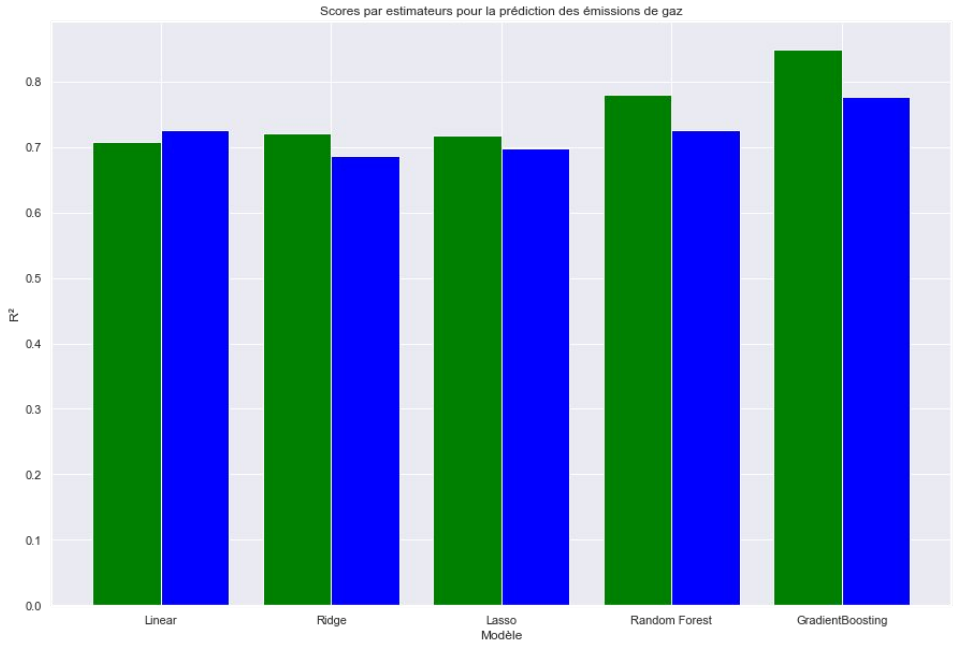


Les différentes modélisations : Random Forest



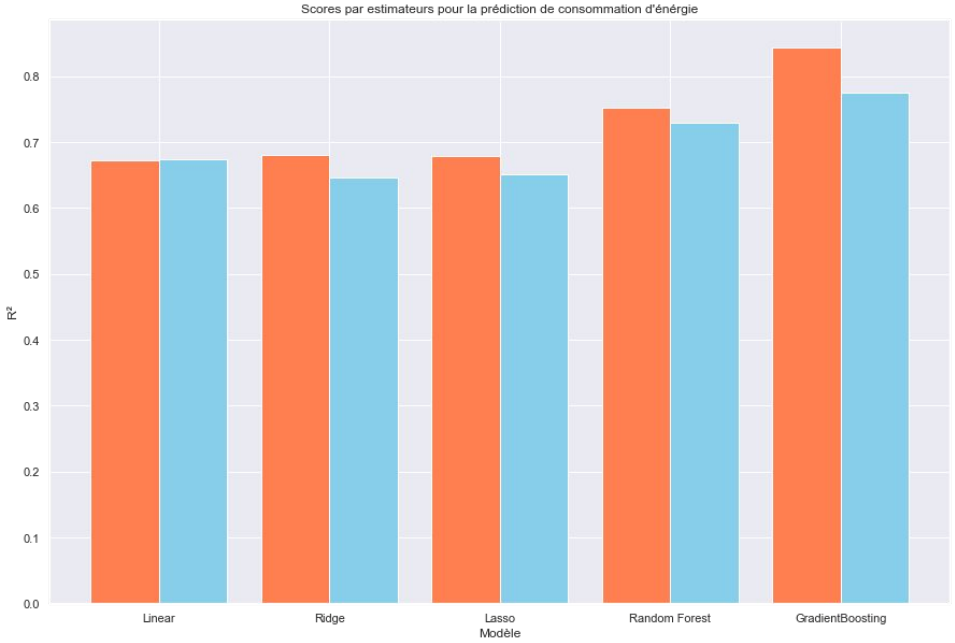
Les différentes modélisations : Gradient Boosting

- Pour les émissions de gaz : score = 0.78, training score =0.84
- {'max_depth': 3, 'min_samples_leaf': 6, 'min_samples_split': 3, 'n_estimators': 200}



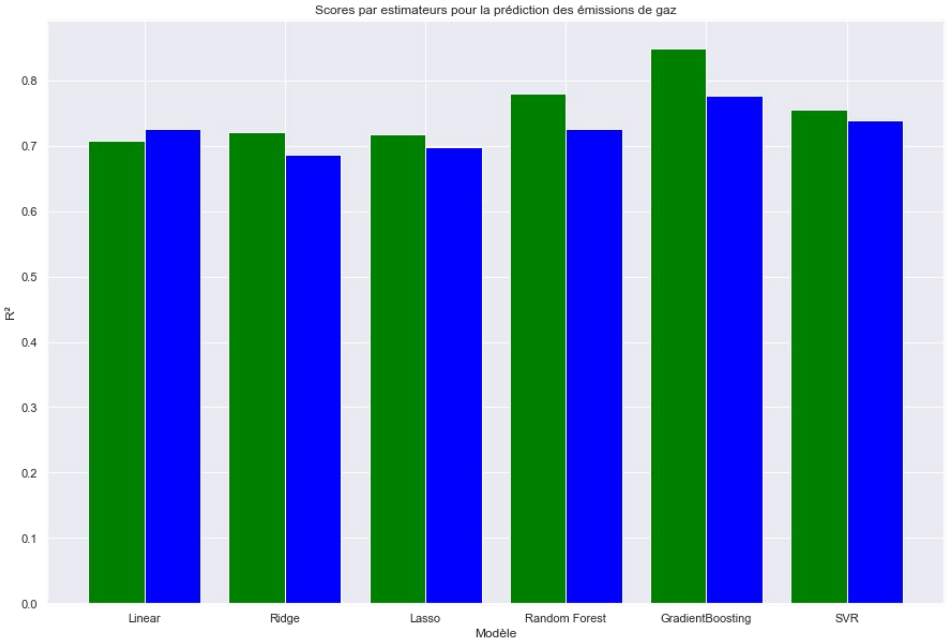
Les différentes modélisations : Gradient Boosting

- Pour la consommation en énergie: score = 0.77, training score =0.84
- {'max_depth': 3, 'min_samples_leaf': 5, 'min_samples_split': 3, 'n_estimators': 200}



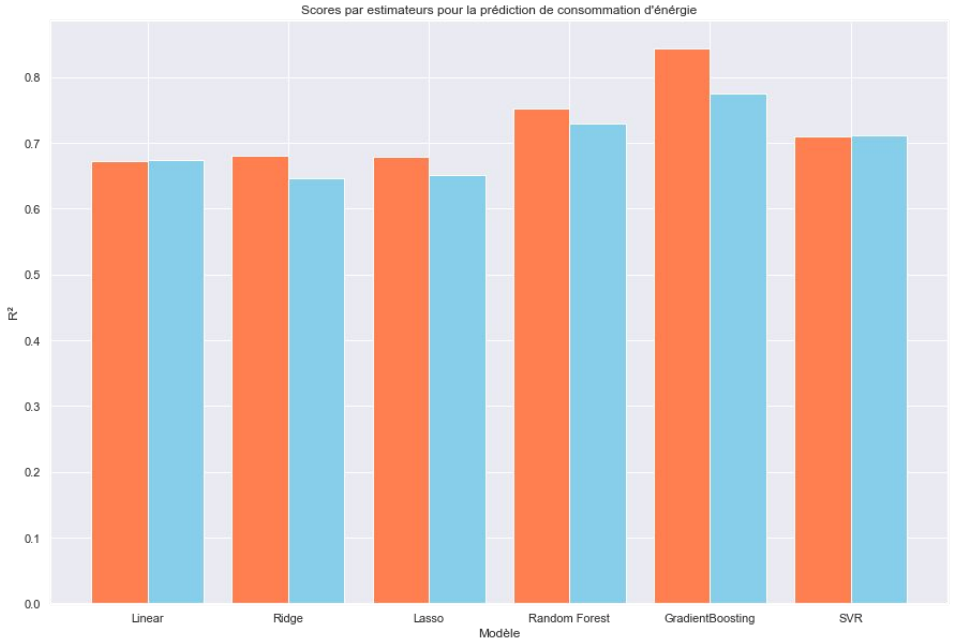
Les différentes modélisations : SVR

- Pour les émissions de gaz : score = 0.74, training score = 0.76
- {'C': 20.0, 'gamma': 0.01, 'kernel': 'rbf'}



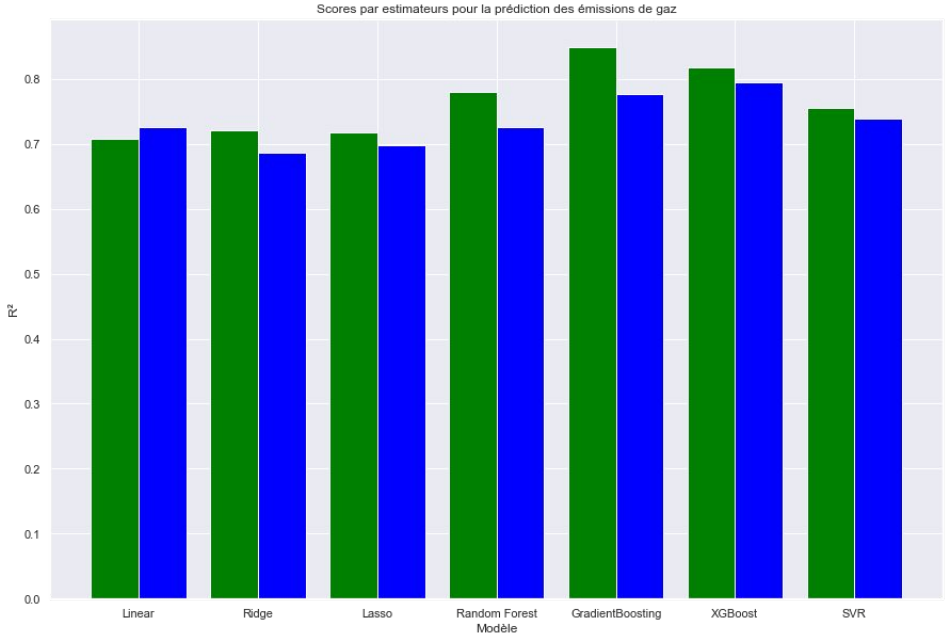
Les différentes modélisations : SVR

- Pour la consommation d'énergie : score = 0.71, training score = 0.71
- {'C': 17.33, 'gamma': 0.01, 'kernel': 'rbf'}



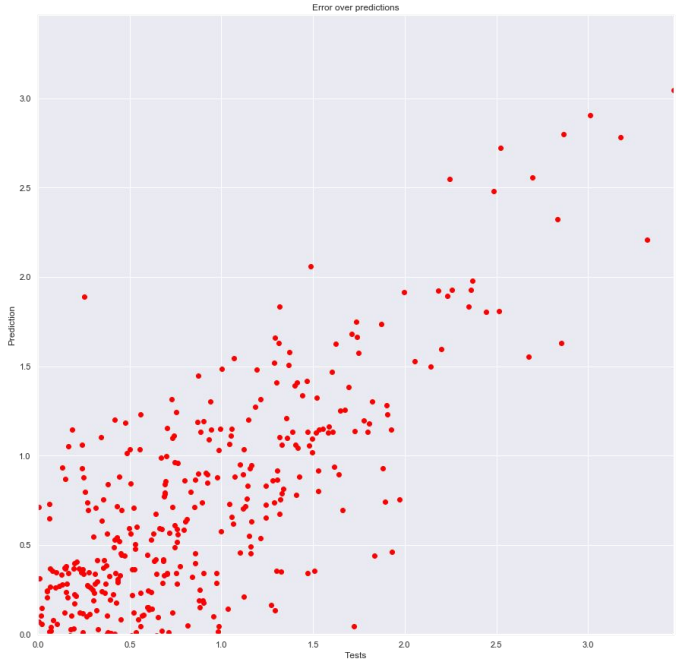
Les différentes modélisations : XGBoost

- Pour les émissions de gaz : score = 0.79, training score = 0.82
- {'colsample_bytree': 0.8, 'learning_rate': 0.1, 'max_depth': 3, 'min_child_weight': 7, 'n_estimators': 1000, 'subsample': 1.0}, early stopping 104



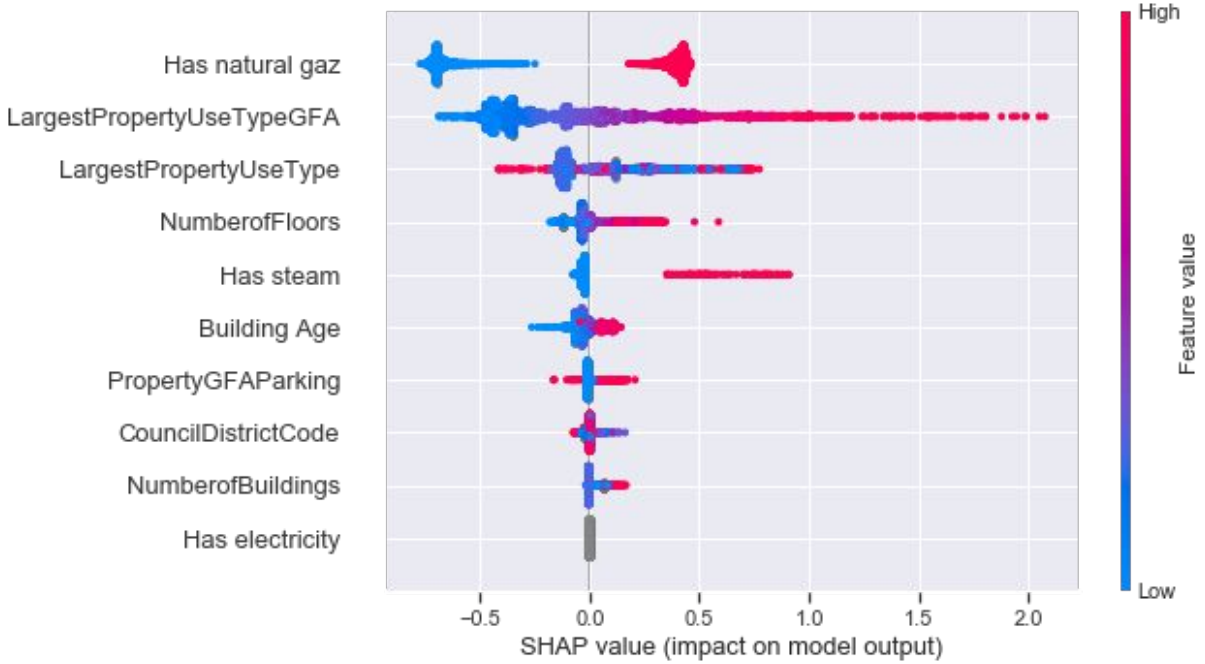
Les différentes modélisations : XGBoost

- Prédiction vs target pour les émissions de gaz



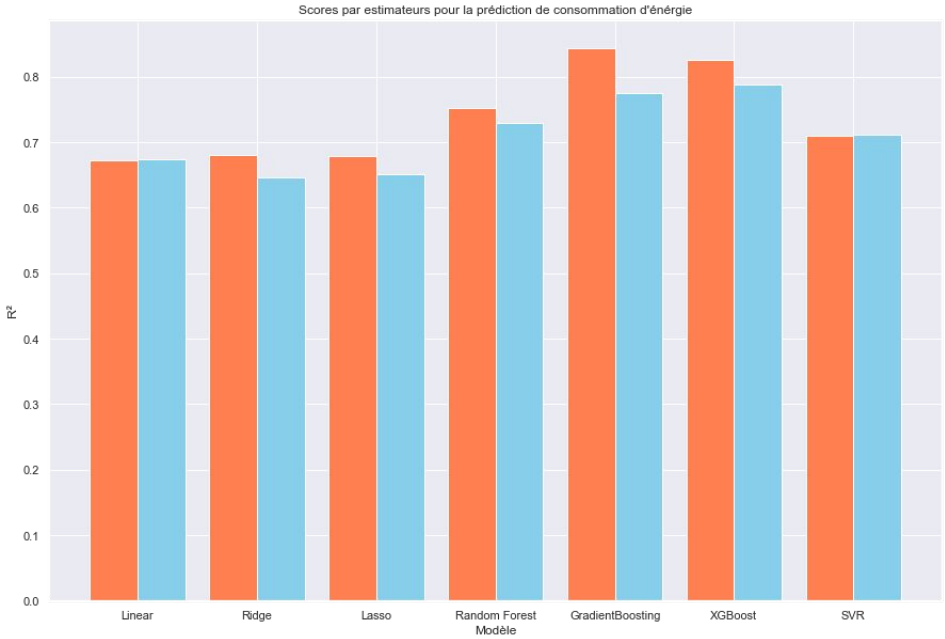
Les différentes modélisations : XGBoost

- Importance des variables pour les émissions de gaz



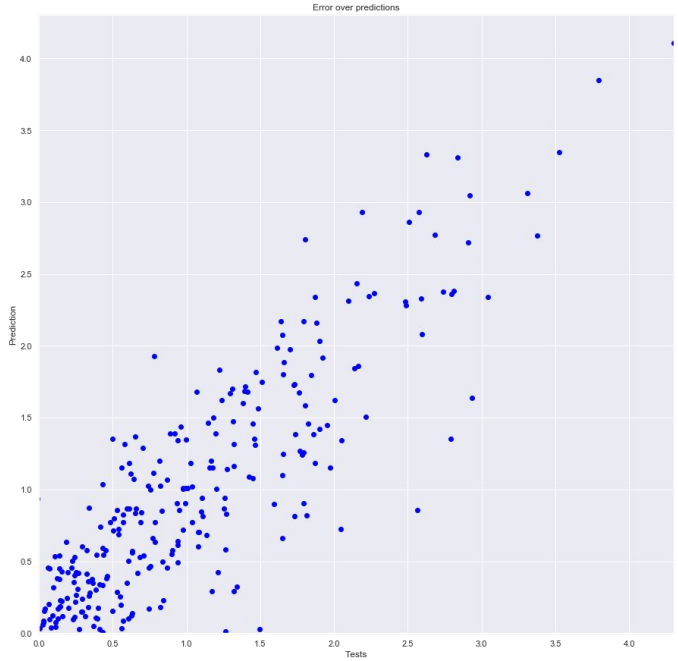
Les différentes modélisations : XGBoost

- Pour la consommation d'énergie : score = 0.79, training score = 0.82
- {'colsample_bytree': 0.6, 'learning_rate': 0.1, 'max_depth': 3, 'min_child_weight': 6, 'n_estimators': 200, 'subsample': 1.0} , Early stopping round = 161



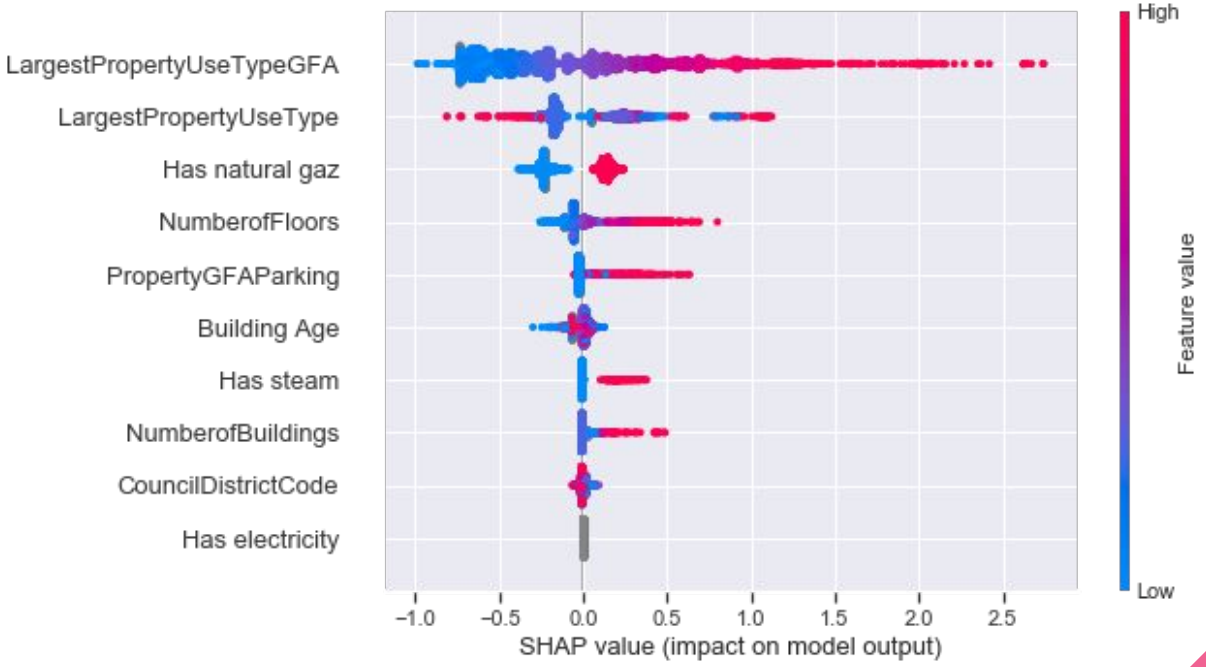
Les différentes modélisations : XGBoost

- Prédiction vs target pour la consommation en énergie



Les différentes modélisations : XGBoost

- Importance des variables pour la consommation d'énergie



Les différentes modélisations

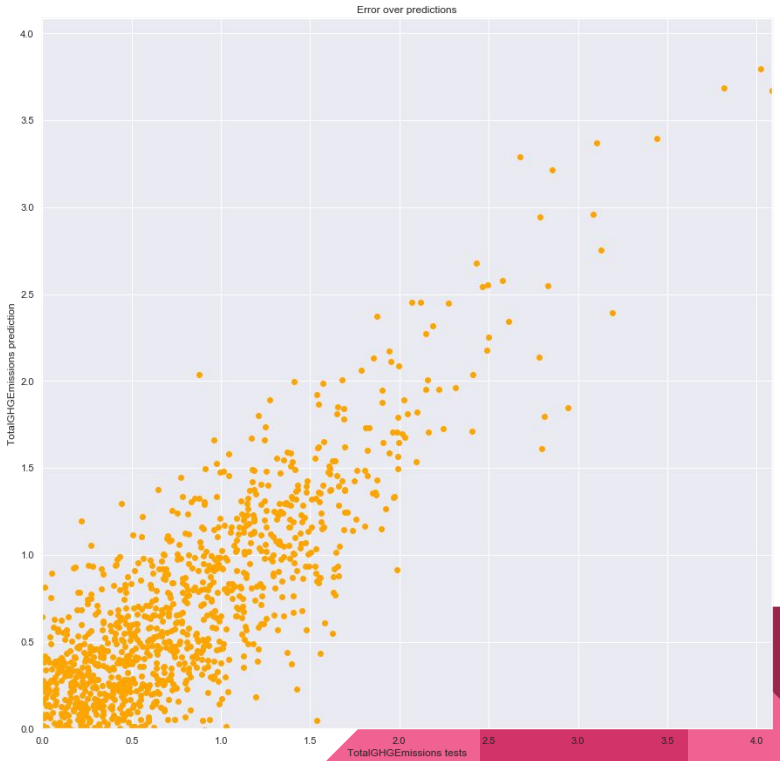
- Sur les données 2015 :
  - Score en utilisant le XGBoost pour les émissions de gaz : 0.79
  - Score en utilisant le XGboost pour la consommation d'énergie : 0.78

Modélisation des données

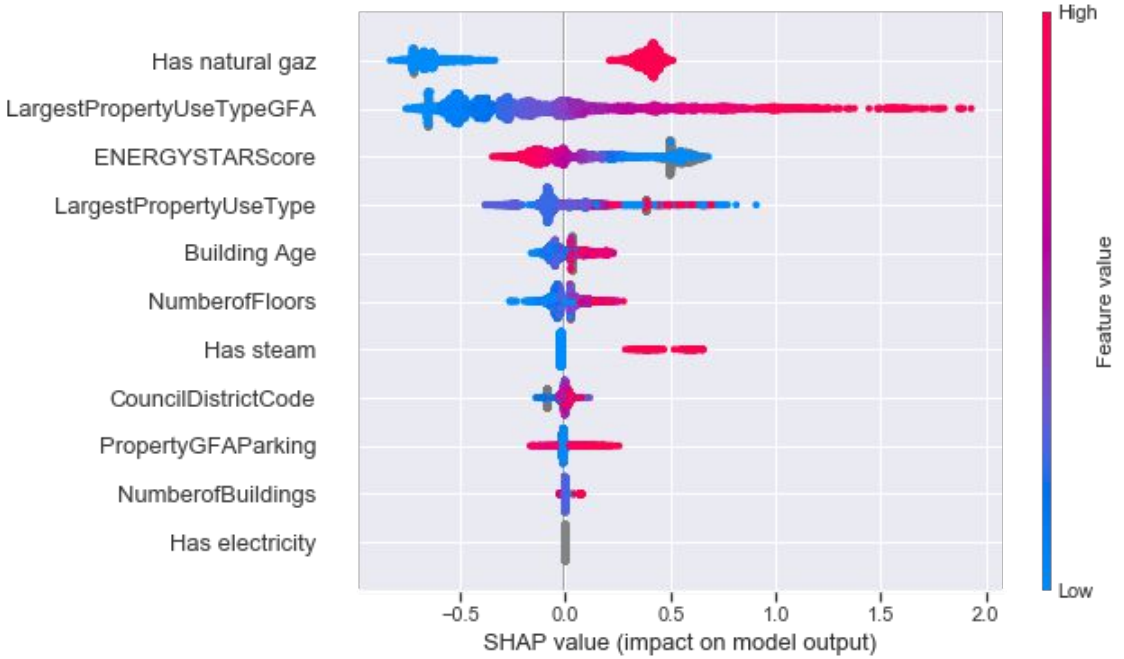
Etude de la variable EnergySTARScore

Etude de la variable EnergySTARScore

- Pour les émissions de gaz :
  - Training score= 0.89 vs 0.82 avant
  - Testing score=0.85 vs 0.79 avant
  - data 2015 score = 0.87 vs 0.79 avant

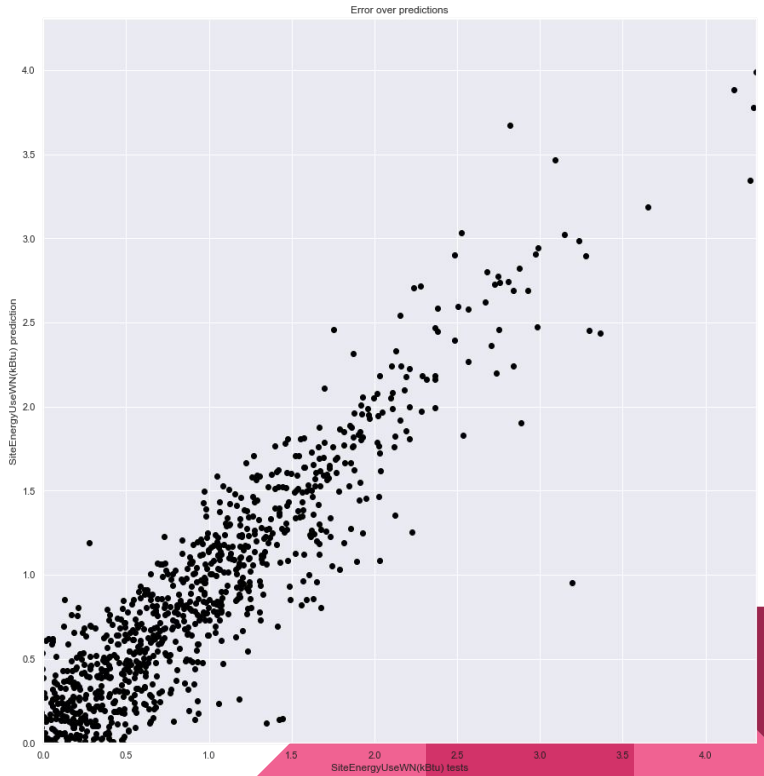


Etude de la variable EnergySTARScore

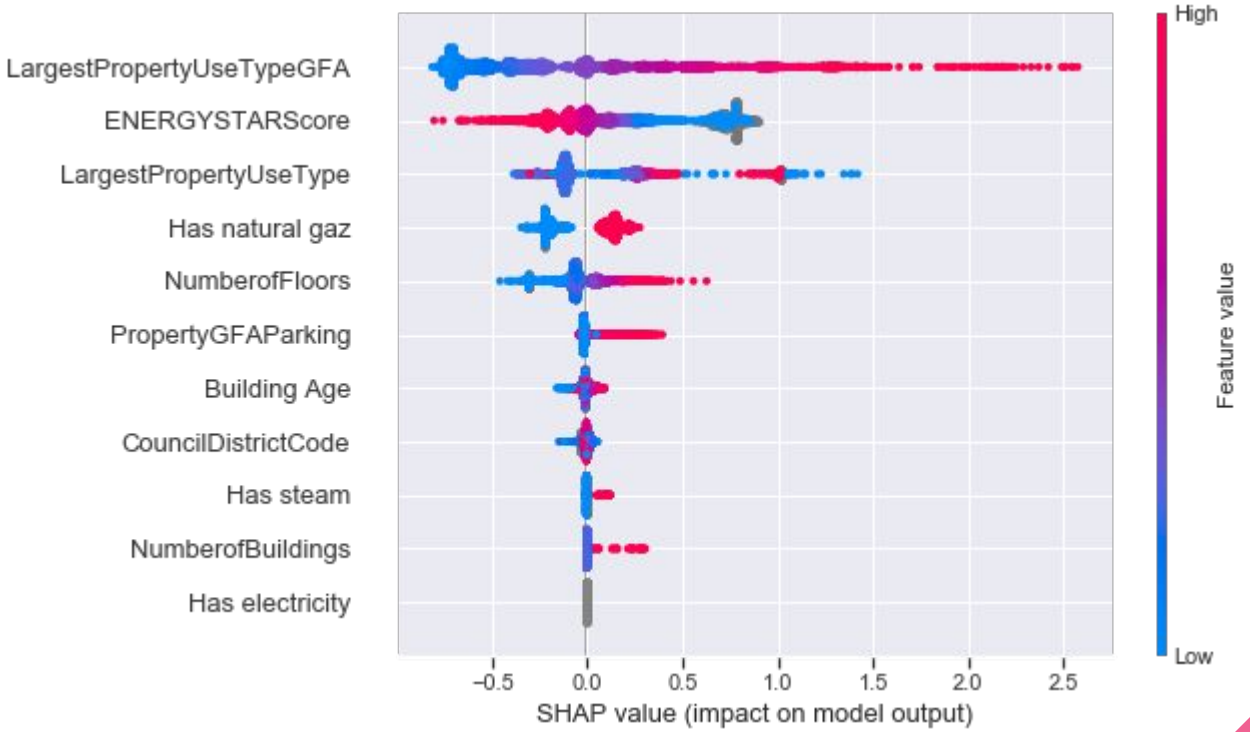


Etude de la variable EnergySTARScore

- Pour la consommation en énergie
  - Training score= 0.93 vs 0.82 avant
  - Testing score=0.88 vs 0.79 avant
  - data 2015 score = 0.90 vs 0.78 avant



Etude de la variable EnergySTARScore



Conclusion

Conclusion

- Meilleur compromis c'est l'utilisation du XGBoost
- Importance de la variable EnergyStarScore
- Pistes d'améliorations
  - Prendre en compte les utilisations secondaires de chaque propriété
  - Essayer d'input l'EnergySTARScore
  - Refaire la modélisation pour les lignes contenant uniquement l'EnergySTARScore