

Projet 3 : Anticipez les besoins en consommation d'un bâtiments

Antoine Maby - 25/08/2021

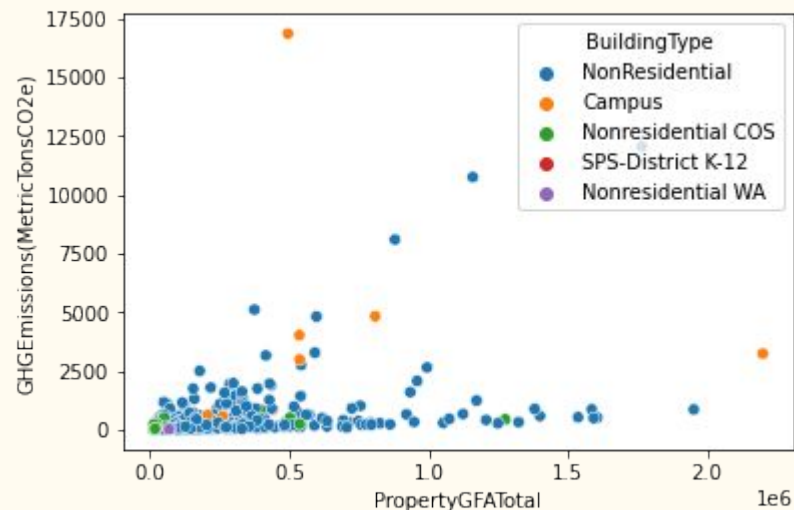
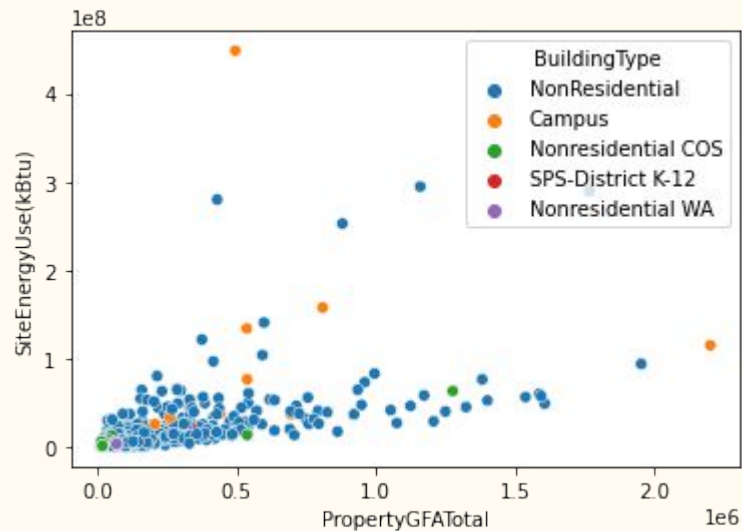
Problématique

- **Prédire les émissions de CO2 et la consommation totale d'énergies**
- **Des relevés ont été effectué en 2015-2016 mais ils sont coûteux**
- **Objectif : Prédire en fonction des données déclaratives du permis d'exploitation commerciale**
- **Evaluer l'intérêt de l'ENERGIE STAR Score**

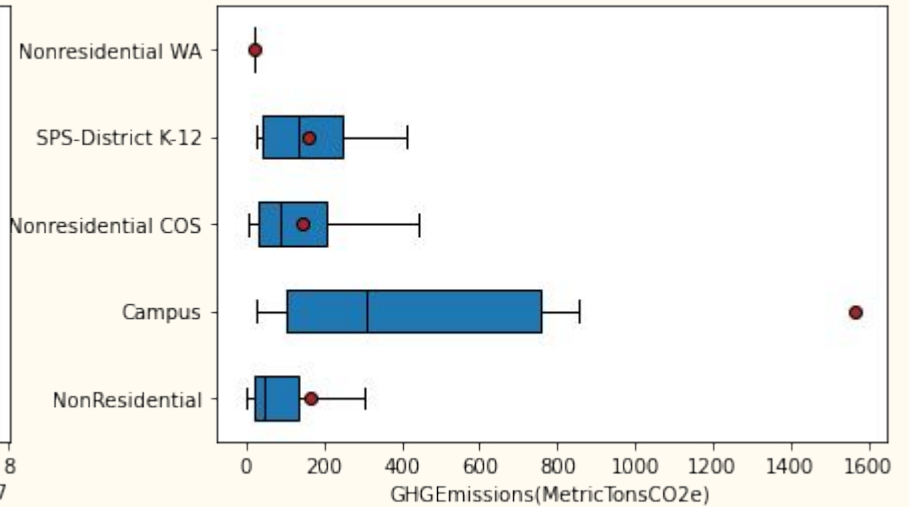
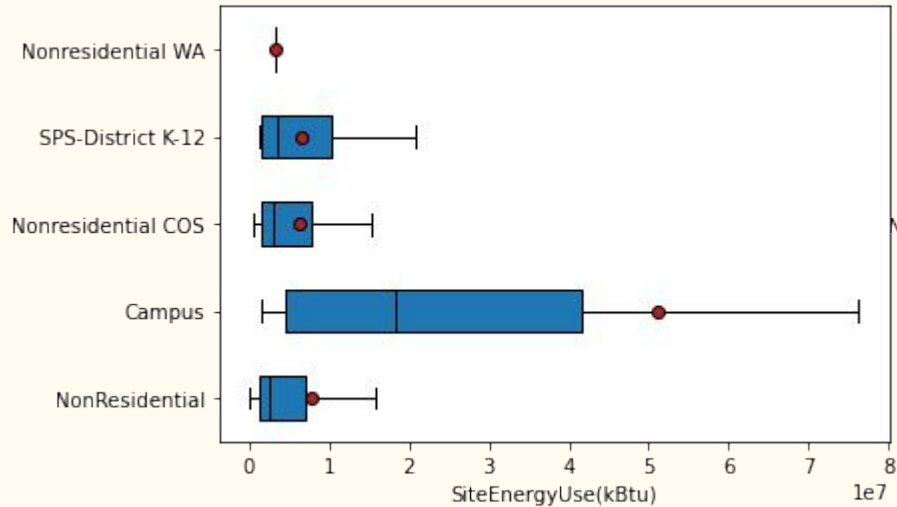
Nettoyage et Feature Engineering

- Fusion des deux databases par année et moyenne sur les bâtiments
- Suppression des outliers répertoriés
- Suppression des données moyennées sur le climat
- Création de nouvelles variables (Age bâtiment, Ratio des surfaces)

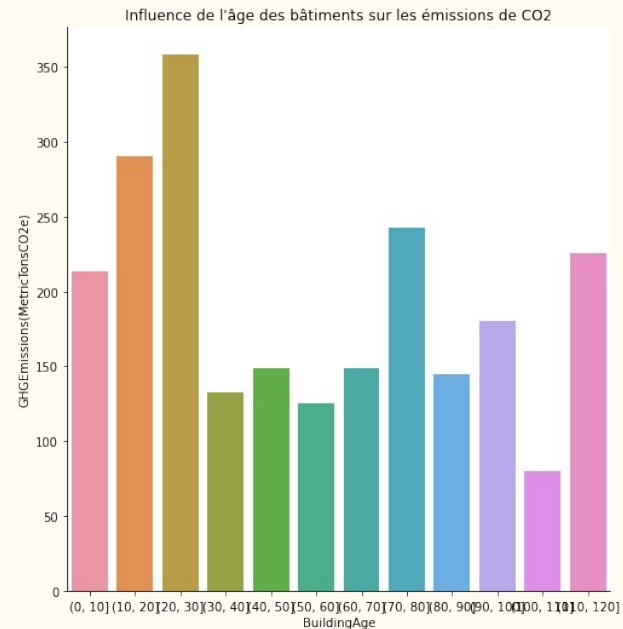
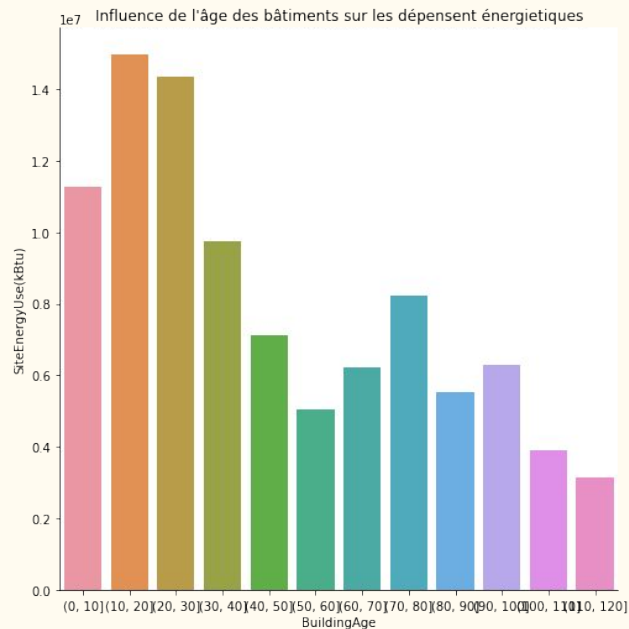
Exploration



Exploration

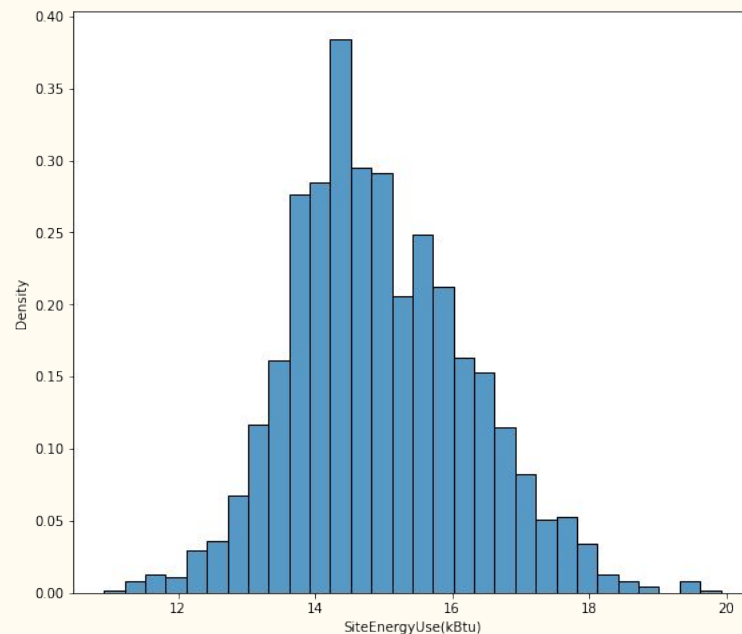
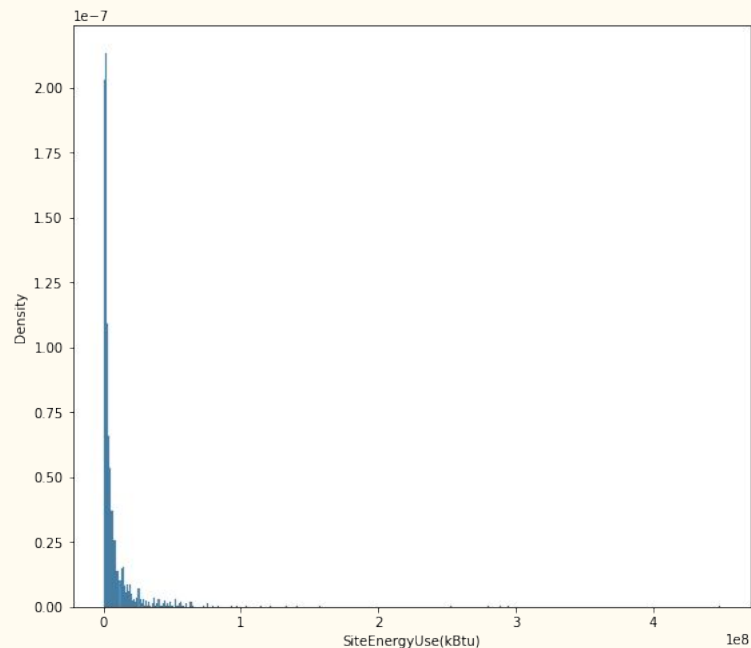


Exploration



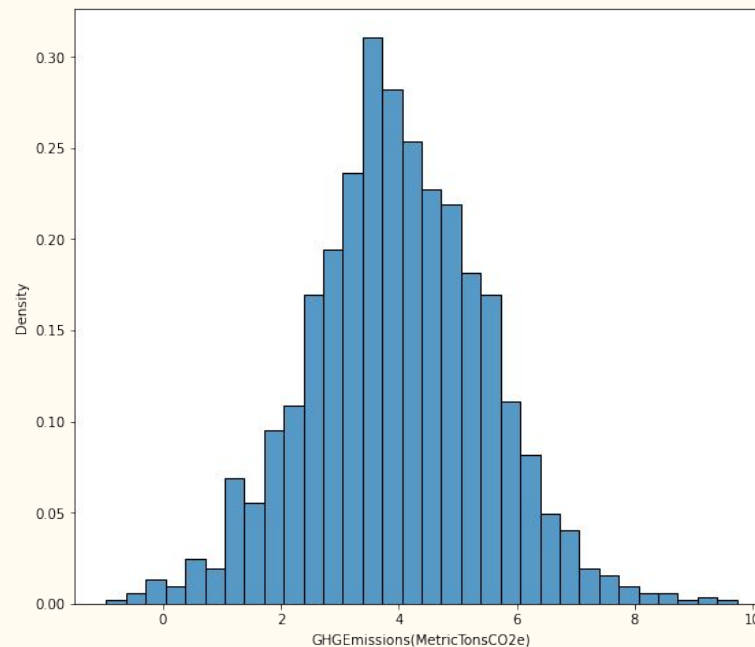
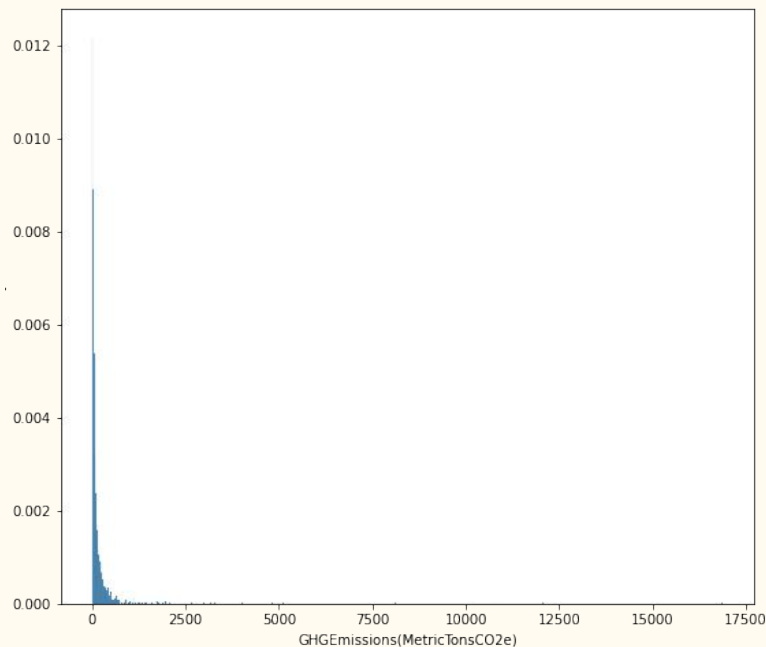
Exploration

Passage au logarithme pour obtenir distribution normale



Exploration

Passage au logarithme pour obtenir distribution normale



Modélisation

Régression Ridge

Hyperparamètres :

- α

Cross validation pour l'énergie

```
Fitting 5 folds for each of 200 candidates, totalling 1000 fits  
Best R2: [0.56523004]  
Best MAE: [-0.65049045]  
Temps moyen [0.01060333] s  
Best Params: {'Ridge__alpha': 178.34308769319094}
```

Cross validation pour le CO2

```
Fitting 5 folds for each of 200 candidates, totalling 1000 fits  
Best R2: [0.42514713]  
Best MAE: [-0.87870694]  
Temps moyen [0.00880232] s  
Best Params: {'Ridge__alpha': 246.5811075822604}
```

Modélisation

Régression Lasso

Hyperparamètres :

- alpha

Cross validation pour l'énergie

```
Fitting 5 folds for each of 200 candidates, totalling 1000 fits  
Best R2: [0.57311699]  
Best MAE: [-0.64302703]  
Temps moyen [0.01801257] s  
Best Params: {'Lasso__alpha': 0.01762914118095948}
```

Cross validation pour le CO2

```
Fitting 5 folds for each of 200 candidates, totalling 1000 fits  
Best R2: [0.43203837]  
Best MAE: [-0.87051355]  
Temps moyen [0.02160492] s  
Best Params: {'Lasso__alpha': 0.02354286414322418}
```

Modélisation

Elastic Net

Hyperparamètres :

- alpha
- l1_ratio

Cross validation pour l'énergie

```
Fitting 5 folds for each of 625 candidates, totalling 3125 fits  
Best R2: [0.57311657]  
Best MAE: [-0.64304808]  
Temps moyen [0.01540456] s  
Best Params: {'Elastic__alpha': 0.01778279410038923, 'Elastic__l1_ratio': 1.0}
```

Modélisation

Elastic Net

Hyperparamètres :

- alpha
- l1_ratio

Cross validation pour le CO2

```
Fitting 5 folds for each of 625 candidates, totalling 3125 fits
Best R2: [0.4320347]
Best MAE: [-0.87054647]
Temps moyen [0.02320571] s
Best Params: {'Elastic__alpha': 0.023713737056616554, 'Elastic__l1_ratio': 1.0}
```

Modélisation

SVR

Hyperparamètres :

- C
- Epsilon

Cross validation pour l'énergie

```
Fitting 5 folds for each of 625 candidates, totalling 3125 fits  
Best R2: [0.55278879]  
Best MAE: [-0.65138984]  
Temps moyen [0.52911696] s  
Best Params: {'svr__C': 2.8729848333536645, 'svr__epsilon': 0.3831186849557287}
```

Modélisation

SVR

Hyperparamètres :

- C
- Epsilon

Cross validation pour le CO2

```
Fitting 5 folds for each of 625 candidates, totalling 3125 fits
Best R2:  [0.40556503]
Best MAE:  [-0.88557567]
Temps moyen [0.07601595] s
Best Params:  {'svr__C': 0.11006941712522095, 'svr__epsilon': 0.5623413251903491}
```

Modélisation

Random Forest

Hyperparamètres :

- n_estimator
- max_depth
- min_samples_leaf
- min_samples_split

Cross validation pour l'énergie

```
Fitting 5 folds for each of 192 candidates, totalling 960 fits
Best R2: [0.71108208]
Best MAE: [-0.51963341]
Temps moyen [1.60067663] s
Best Params: {'rfr__max_depth': 30, 'rfr__min_samples_leaf': 1, 'rfr__min_samples_split': 20, 'rfr__n_estimators': 250}
```

Modélisation

Random Forest

Hyperparamètres :

- n_estimator
- max_depth
- min_samples_leaf
- min_samples_split

Cross validation pour le CO2

```
Fitting 5 folds for each of 144 candidates, totalling 720 fits
Best R2: [0.53101241]
Best MAE: [-0.79977567]
Temps moyen [1.64377007] s
Best Params: {'rfr__max_depth': 30, 'rfr__min_samples_leaf': 1, 'rfr__min_samples_split': 15, 'rfr__n_estimators': 250}
```


Modélisation

Gradient Boosting

Hyperparamètres :

- n_estimator
- max_depth
- min_samples_leaf
- min_samples_split
- Learning_rate

Cross validation pour l'énergie

```
Fitting 5 folds for each of 432 candidates, totalling 2160 fits
Best R2: [0.72203752]
Best MAE: [-0.51290138]
Temps moyen [0.71427555] s
Best Params: {'gbr__learning_rate': 0.1, 'gbr__max_depth': 1, 'gbr__min_samples_leaf': 1, 'gbr__min_samples_split': 6, 'gbr__n_estimators': 500}
```

Modélisation

Gradient Boosting

Hyperparamètres :

- n_estimator
- max_depth
- min_samples_leaf
- min_samples_split
- Learning_rate

Cross validation pour le CO2

```
Fitting 5 folds for each of 576 candidates, totalling 2880 fits
```

```
Best R2: [0.5426816]
```

```
Best MAE: [-0.79631872]
```

```
Temps moyen [0.79732223] s
```

```
Best Params: {'gbr__learning_rate': 0.2, 'gbr__max_depth': 1, 'gbr__min_samples_leaf': 5, 'gbr__min_samples_split': 2, 'gbr__n_estimators': 500}
```

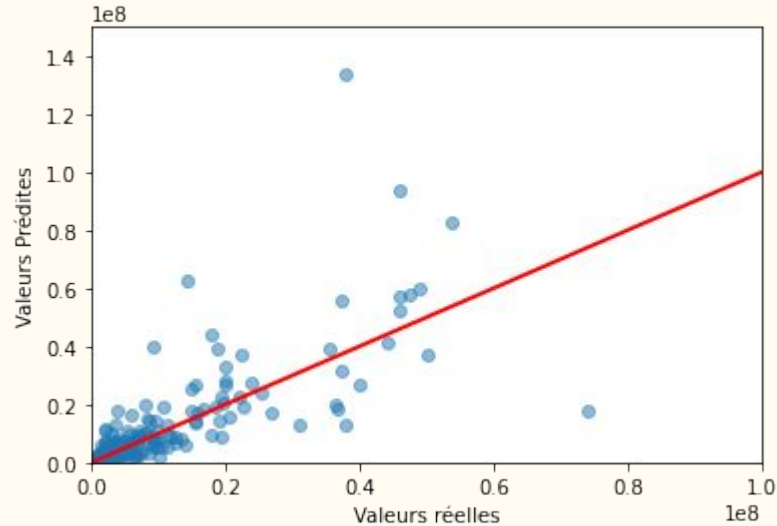
Sélection du modèle final

Modèle pour la consommation énergétique

level_0	mean_fit_time	mean_score_time	mean_test_neg_mean_absolute_error	mean_train_neg_mean_absolute_error	mean_test_r2	mean_train_r2
RandomForest	1.600677	0.026206	-0.519633	-0.368704	0.711082	0.860362
XGBoost	0.714276	0.001601	-0.512901	-0.460724	0.722038	0.777682
LinearSVR	0.529117	0.002402	-0.651390	-0.591132	0.552789	0.631904
ElasticNet	0.015405	0.001801	-0.643048	-0.599357	0.573117	0.636991
Ridge	0.010603	0.002000	-0.650490	-0.594795	0.565230	0.638732
Lasso	0.018013	0.001600	-0.643027	-0.599161	0.573117	0.637166

Sélection du modèle final

Modèle pour la consommation énergétique



Sans Transformation inverse `r2 test : 0.7131305666343533`

Avec Transformation inverse `r2 test : 0.70739557891258`

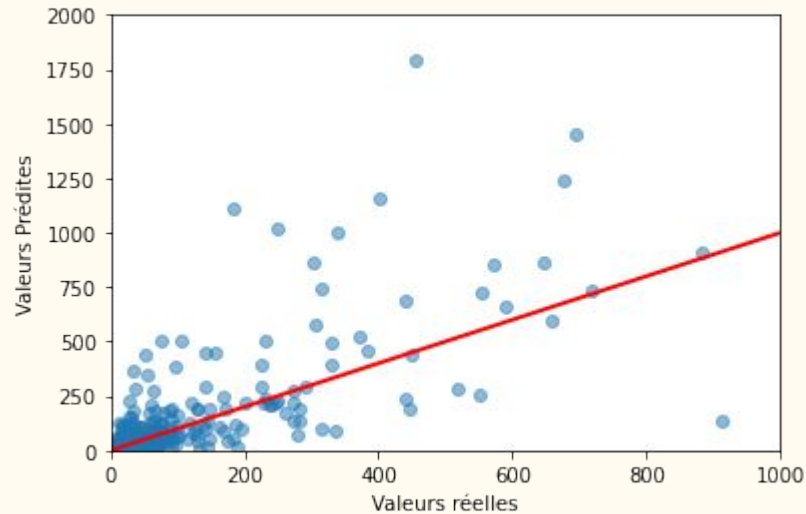
Sélection du modèle final

Modèle pour les émissions de CO2

level_0	mean_fit_time	mean_score_time	mean_test_neg_mean_absolute_error	mean_train_neg_mean_absolute_error	mean_test_r2	mean_train_r2
RandomForest	1.643770	0.027607	-0.799776	-0.536476	0.531012	0.790141
XGBoost	0.797322	0.002801	-0.796319	-0.713765	0.542682	0.635263
LinearSVR	0.076016	0.001600	-0.885576	-0.808275	0.405565	0.514011
ElasticNet	0.023206	0.001799	-0.870546	-0.818243	0.432035	0.505321
Ridge	0.008802	0.001601	-0.878707	-0.811984	0.425147	0.508202
Lasso	0.021605	0.001801	-0.870514	-0.818058	0.432038	0.505520

Sélection du modèle final

Modèle pour les émissions de CO2



Sans Transformation inverse

`r2 test : 0.5221212849918768`

Avec Transformation inverse

`r2 test : 0.52988631091429`

Améliorations du modèle

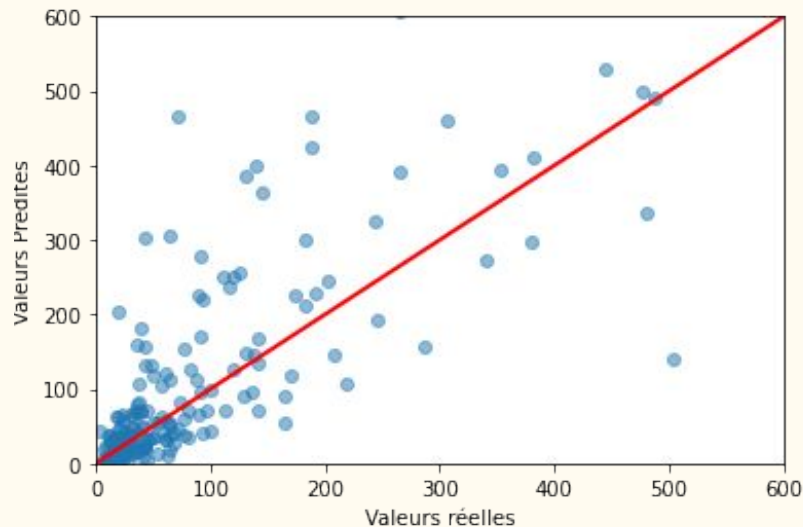
Ajout de la variable Energy Star Score

Ce score permet de réaliser plusieurs actions

- Évaluer les données énergétiques réelles facturées
- Normaliser pour l'activité commerciale
- Comparer les bâtiments à la population
- Indiquer le niveau de performance énergétique

Améliorations du modèle

Modèle pour la consommation énergétique

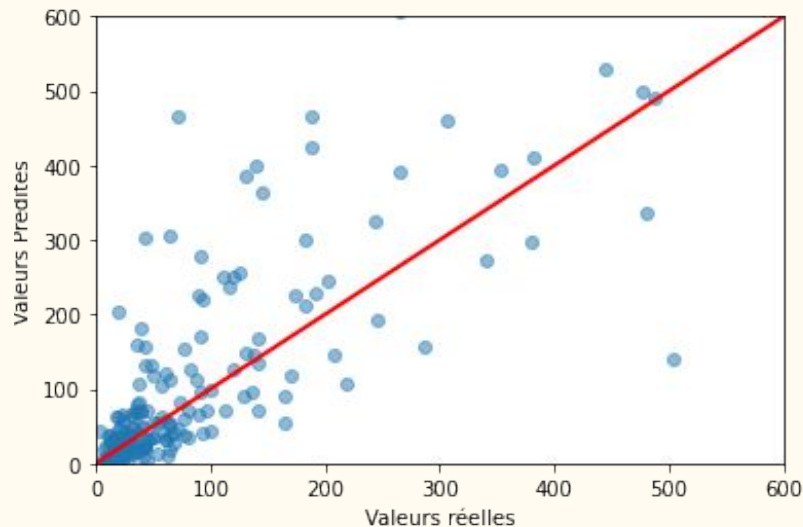


Sans Transformation inverse `r2 test : 0.8672808422775804`

Avec Transformation inverse `r2 test : 0.8970965554970651`

Améliorations du modèle

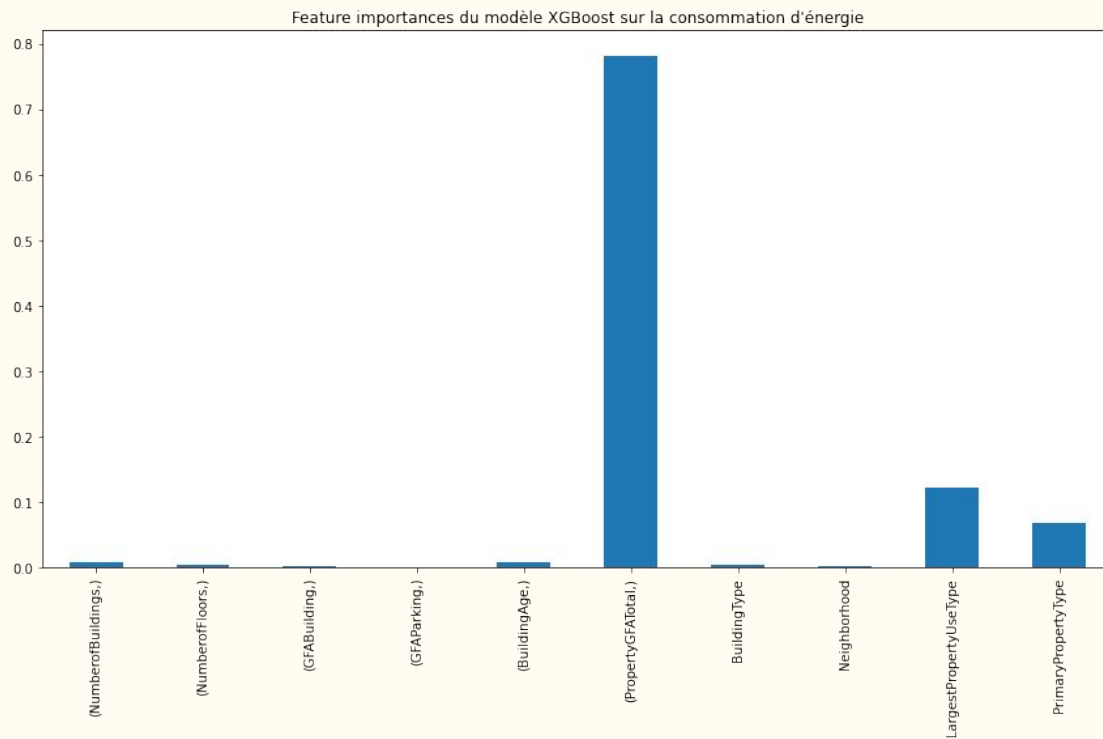
Modèle pour les émission de CO2



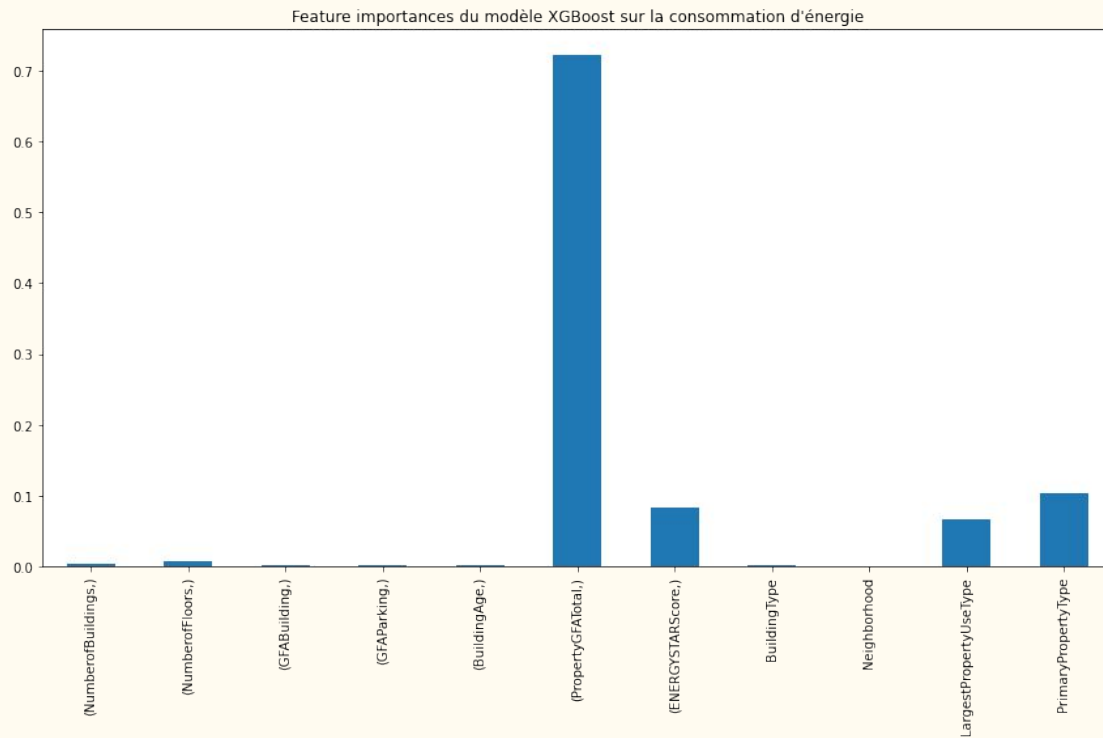
Sans Transformation inverse `r2 test : 0.6888315999991013`

Avec Transformation inverse `r2 test : 0.6960562323176634`

Améliorations du modèle



Améliorations du modèle



Améliorations du modèle

- Tests des modèles avec outliers : pas d'améliorations
- Création d'un pickle avec tous les modèles
- Création d'une classe pour prédire les émissions pour un nouveau bâtiment

Merci pour votre attention