

*For a carbon neutral city*



**Seattle**



Anticipez les besoins en consommation  
électrique de bâtiments

---

Data & Analytics

Eric Blanvillain - 16-11-2021

# Problématique



## Mon rôle :

- Je travaille pour la ville de Seattle. Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, mon équipe s'intéresse de près aux émissions des bâtiments non destinés à l'habitation.
- Des relevés minutieux ont été effectués par des agents en 2015 et en 2016. **Cependant, ces relevés sont coûteux à obtenir.**
- A partir de ceux déjà réalisés, je veux tenter de prédire les émissions de CO<sub>2</sub> et la consommation totale d'énergie de bâtiments pour lesquels elles n'ont pas encore été mesurées.

## Les points à aborder :

- Présentation de la donnée de la ville de Seattle, quelles sont les données utiles ?
- Comment traiter la donnée (cleaning/feature engineering) -> création du dataset prédition
- Présentation des indicateurs clefs (TotalGHGEmissions, SiteEnergyUse(kBtu))
- Etude des modèles de prédition “potentiels”
- Evaluation de l'utilité du EnergyStar Score

# Présentation du Dataset (1/2) “pre cleaning”

## Dataset statistics

Number of variables	47
Number of observations	3340
Missing cells	26512
Missing cells (%)	16.9%
Total size in memory	1.2 MiB
Average record size in memory	376.0 B

## Dataset statistics

Number of variables	46
Number of observations	3376
Missing cells	19952
Missing cells (%)	12.8%
Total size in memory	1.2 MiB
Average record size in memory	361.0 B

## Variable types

Numeric	32
Categorical	15

2015

- 1 - Adapter les colonnes non-conformes
- 2 - Remplir la donnée manquante
- 3 - Supprimer/modifier certaines colonnes

## Variable types

Numeric	29
Categorical	15
Boolean	1
Unsupported	1

2016

# Présentation du Dataset (2/2) “post cleaning”

## Dataset statistics

Number of variables	24
Number of observations	3093
Missing cells	1019
Missing cells (%)	1.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	517.9 KiB
Average record size in memory	171.5 B

→ Non Residential Buildings

## Variable types

Numeric	19
Categorical	5

→ (EnergySTAR Score) = 33% vide

Et maintenant ?

- > Etude de la donnée/indicateurs
- > Choix du modèle
- > Modélisation

# Explication des variables

Les variables à prédire sont :

- TotalGHGEmissions - "Emission C02"
- SiteEnergyUse(kBtu) - "Consommation d'énergie"

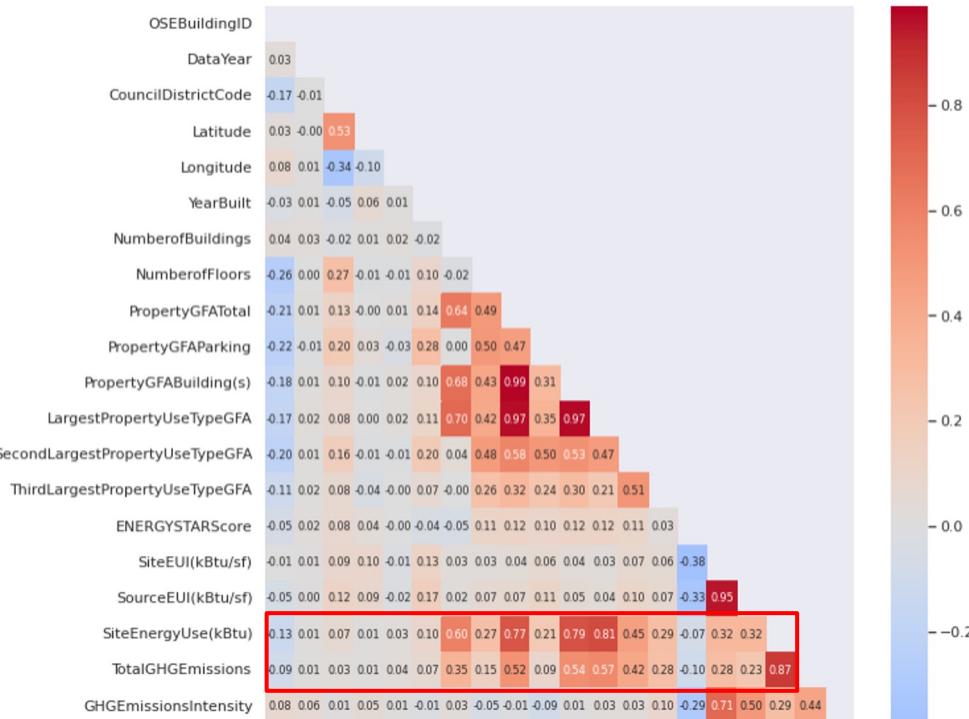
Plusieurs variables concernent les surfaces (GFA = Gross floor area)



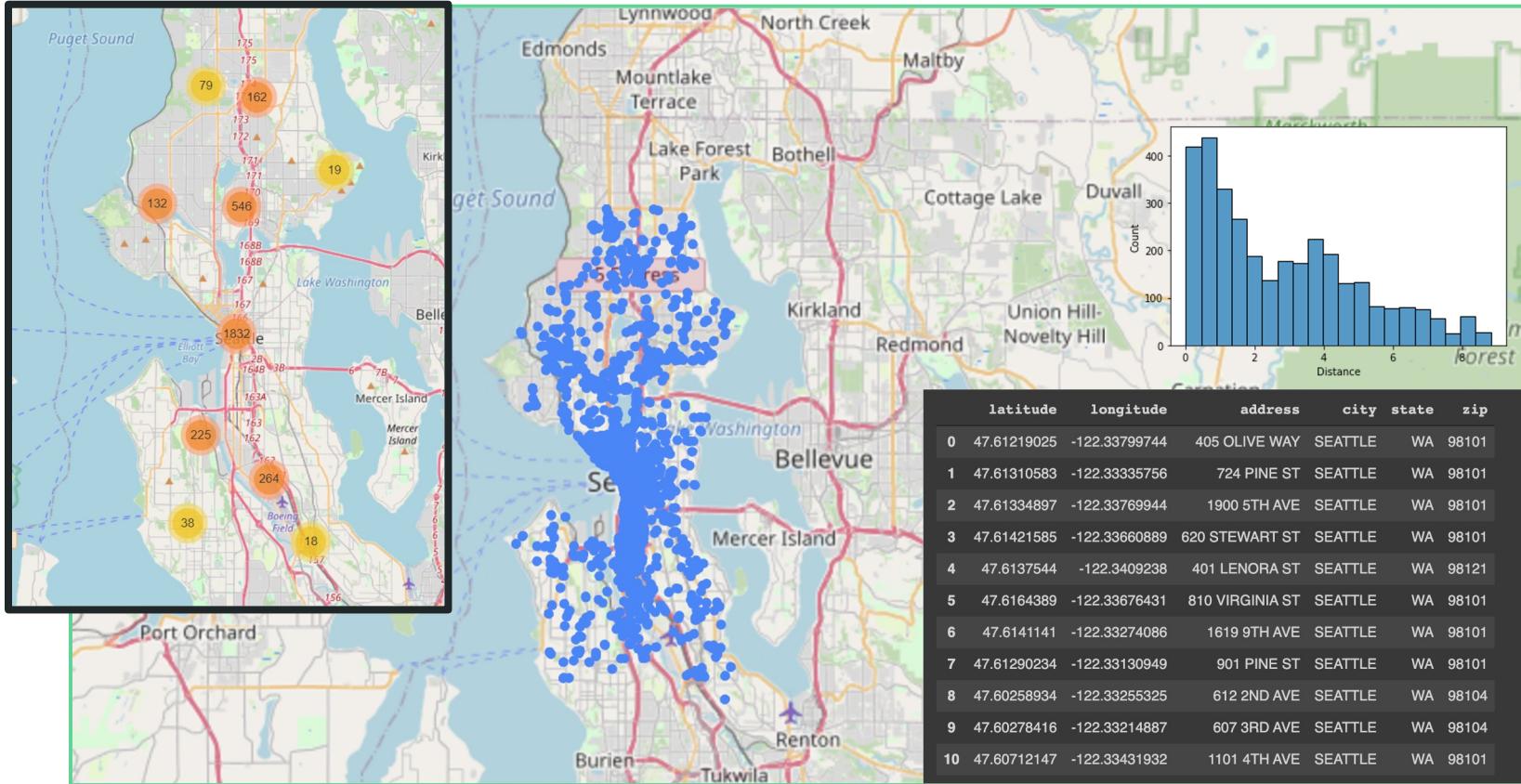
En prenant en compte toutes les consommations d'énergie, le score "SiteEnergyUse(kBtu)" fournit une évaluation complète de l'efficacité énergétique d'un bâtiment.



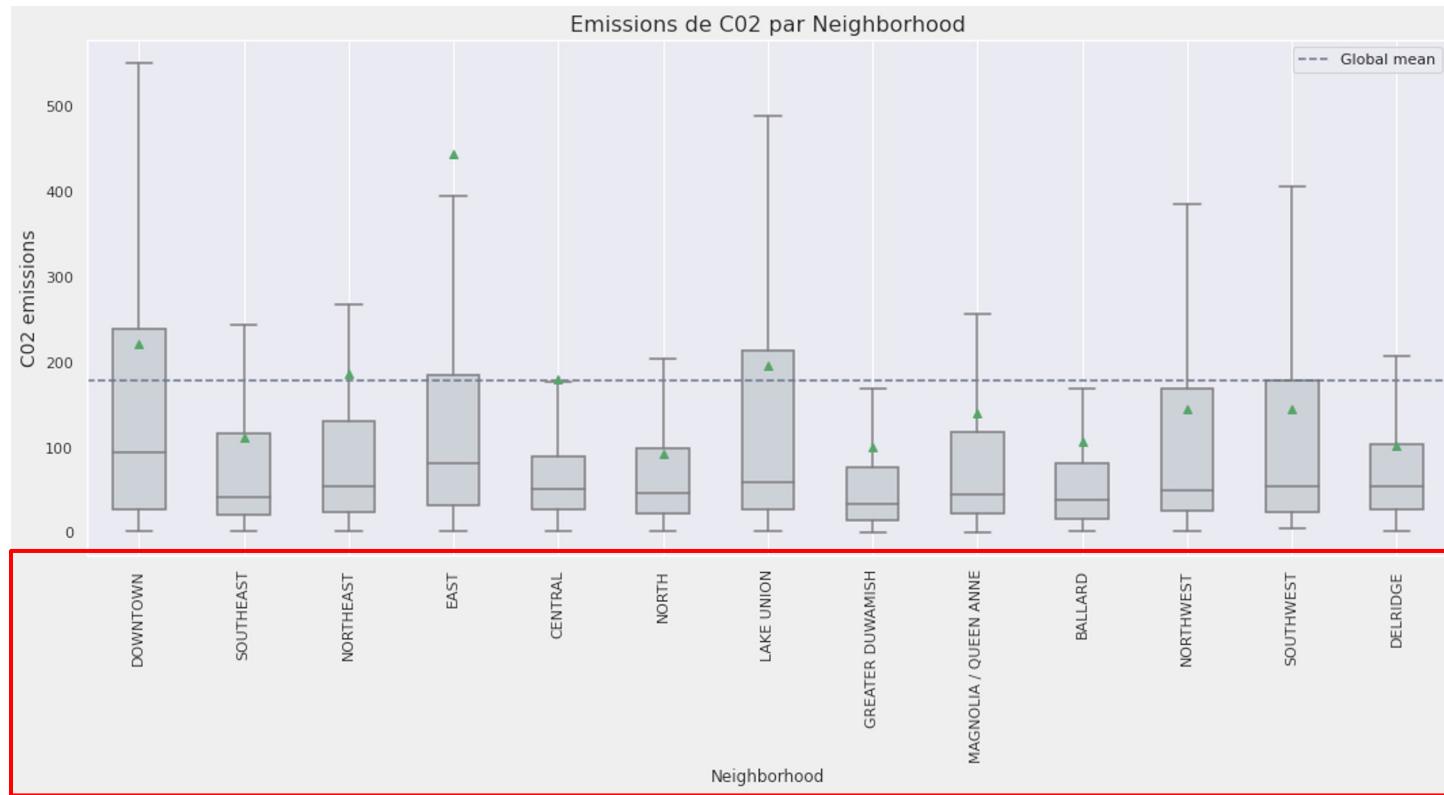
Heatmap des corrélations linéaires



# Etude des bâtiments et features engineering (1/7) - localisation

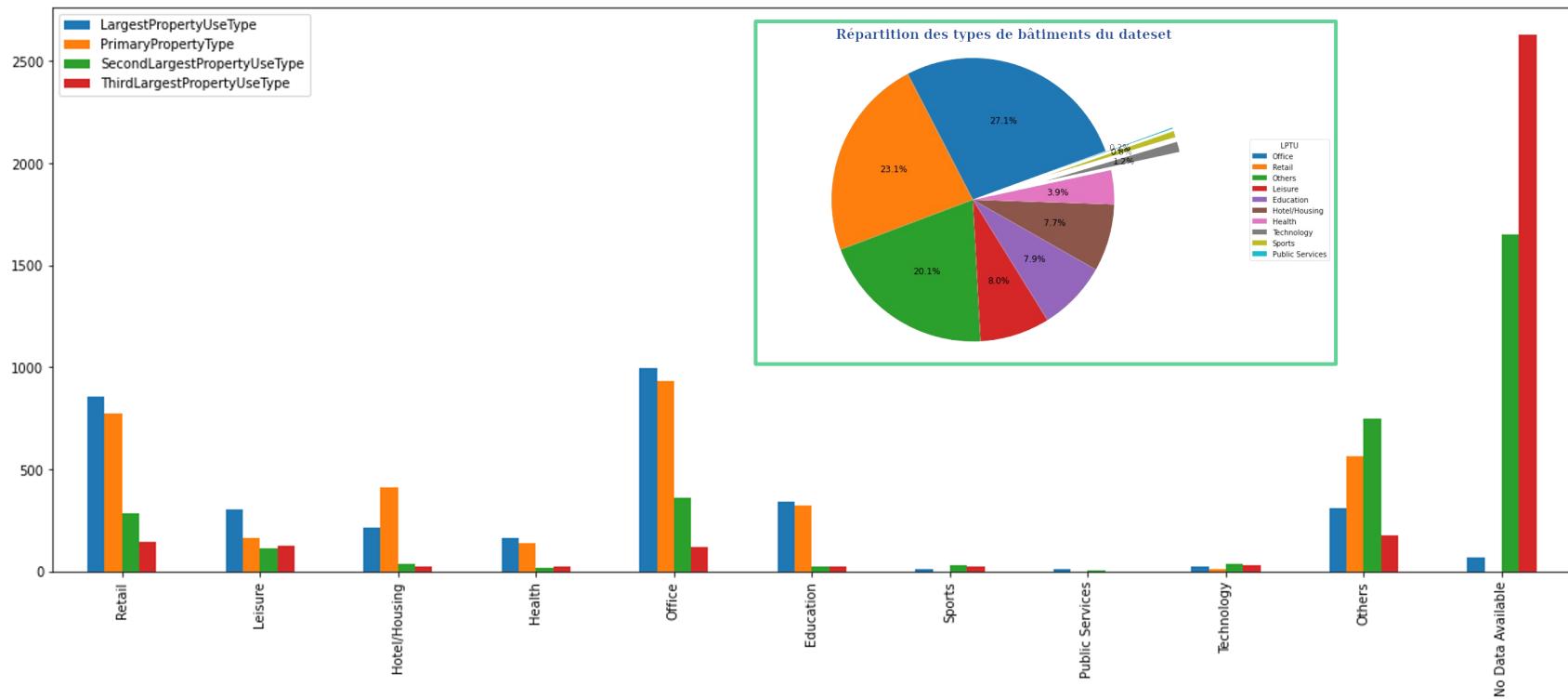


# Etude des bâtiments et features engineering (2/7) - quartier

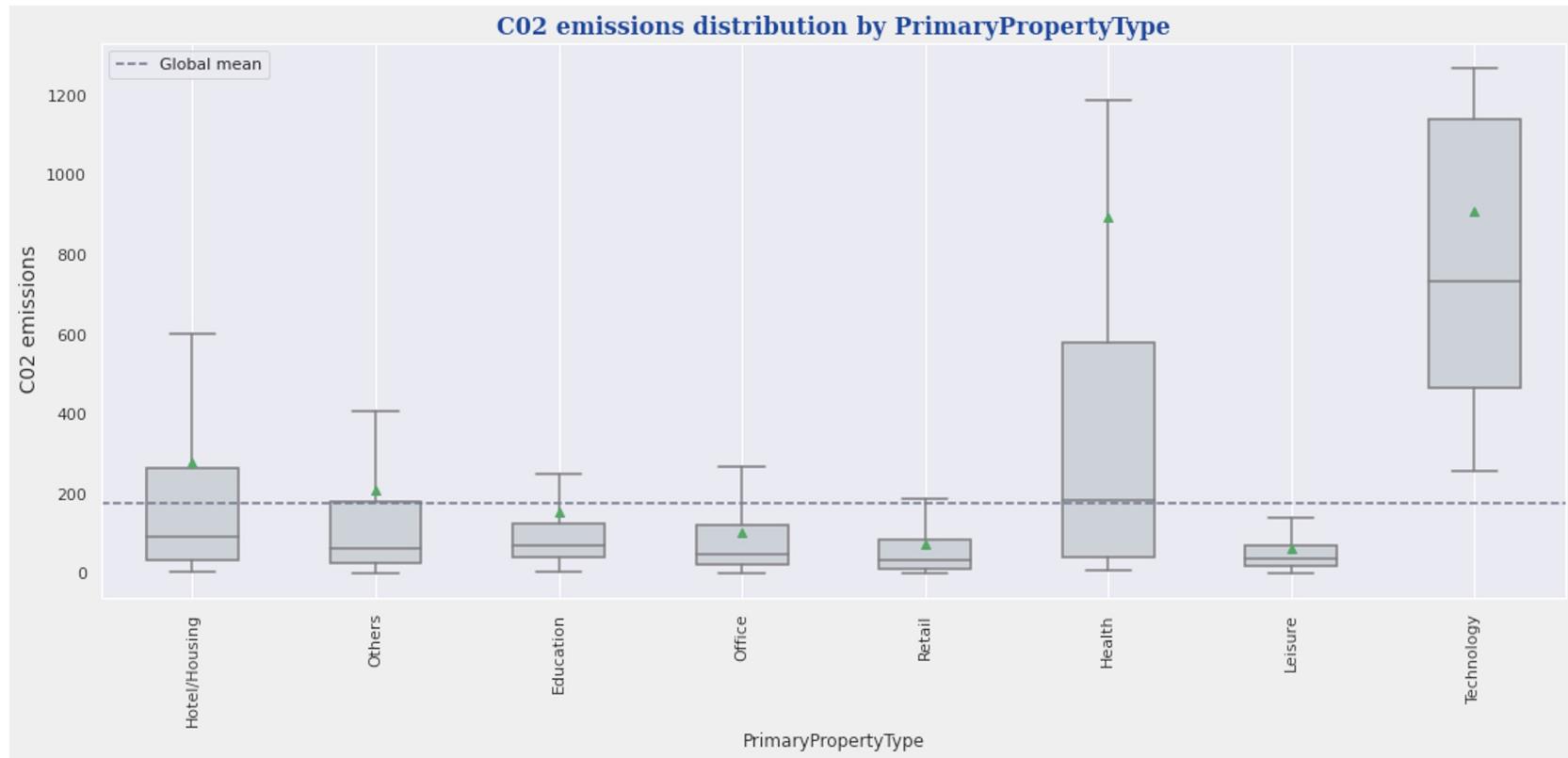


# Etude des bâtiments et features engineering (3/7) - type

## Les nouvelles catégories de building les plus représentées

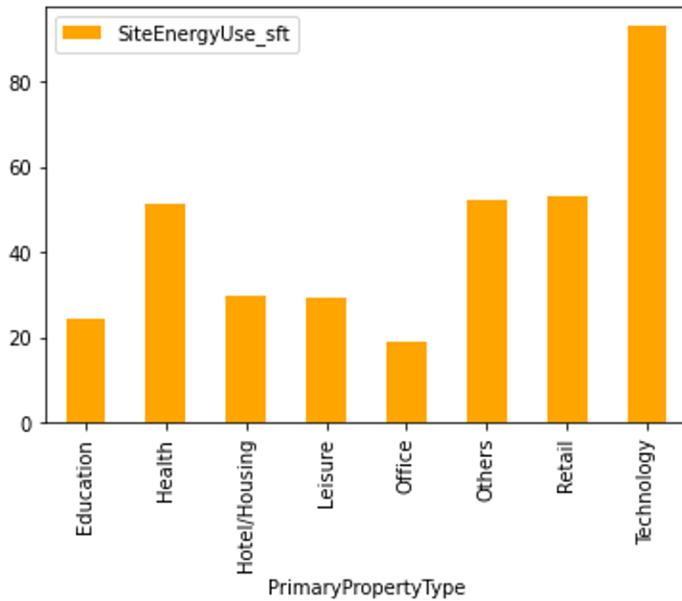
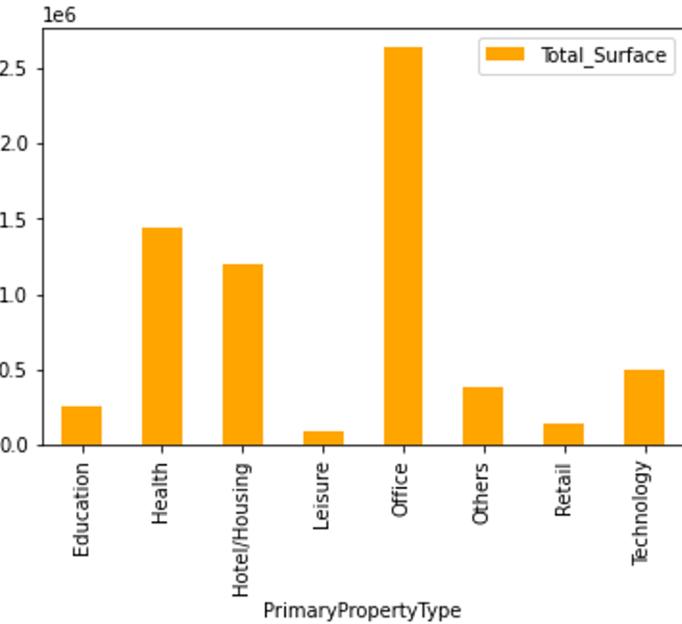


# Etude des bâtiments et features engineering (4/7) - type



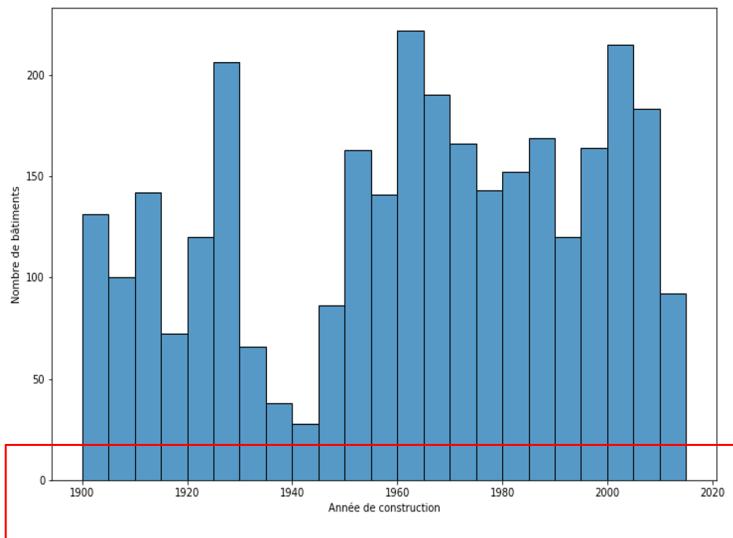
# Etude des bâtiments et features engineering (5/7) - surface

Surface totale des batiments par categories   Consommation énergétique des batiments par sft

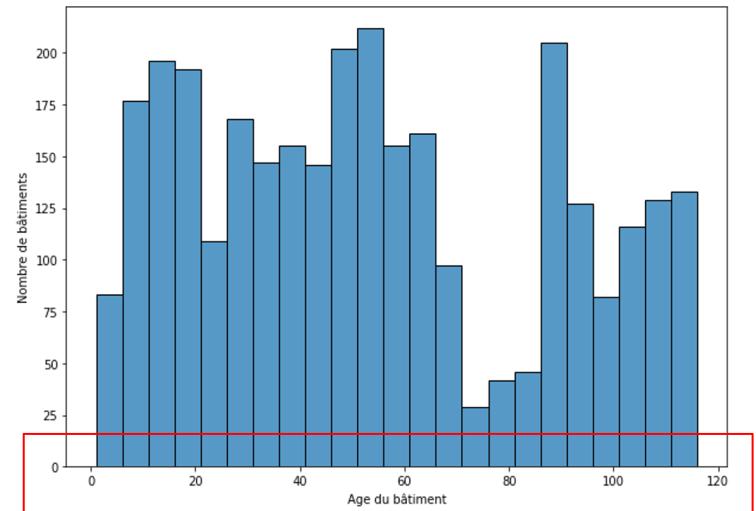


# Etude des bâtiments et features engineering (6/7) - age

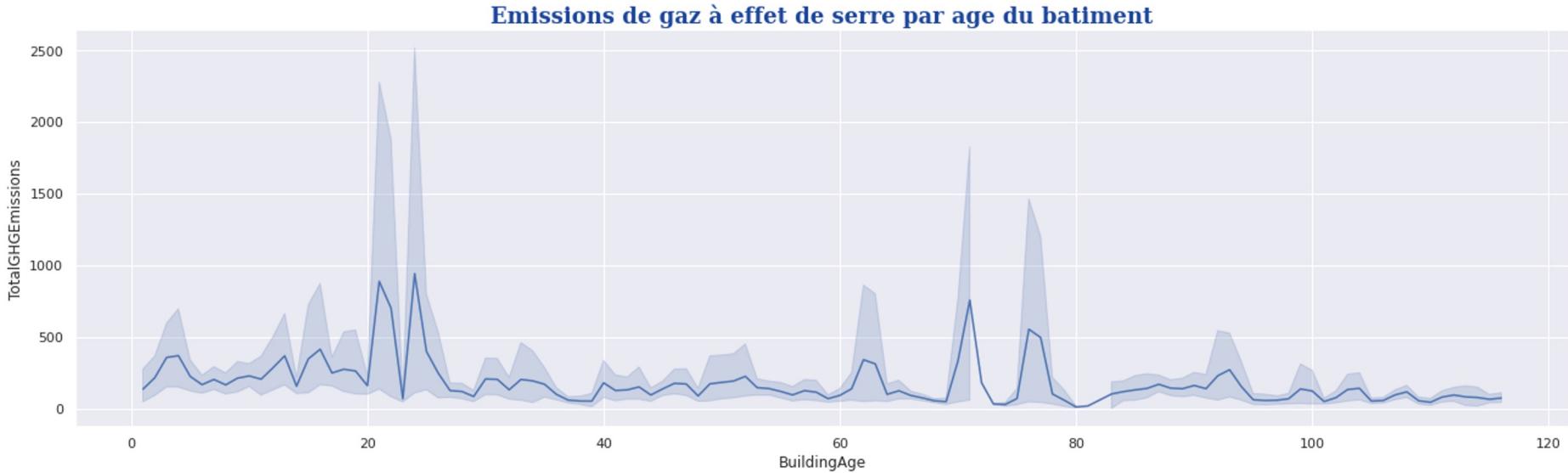
Distribution des années de construction des bâtiments



Distribution de l'âge des bâtiments



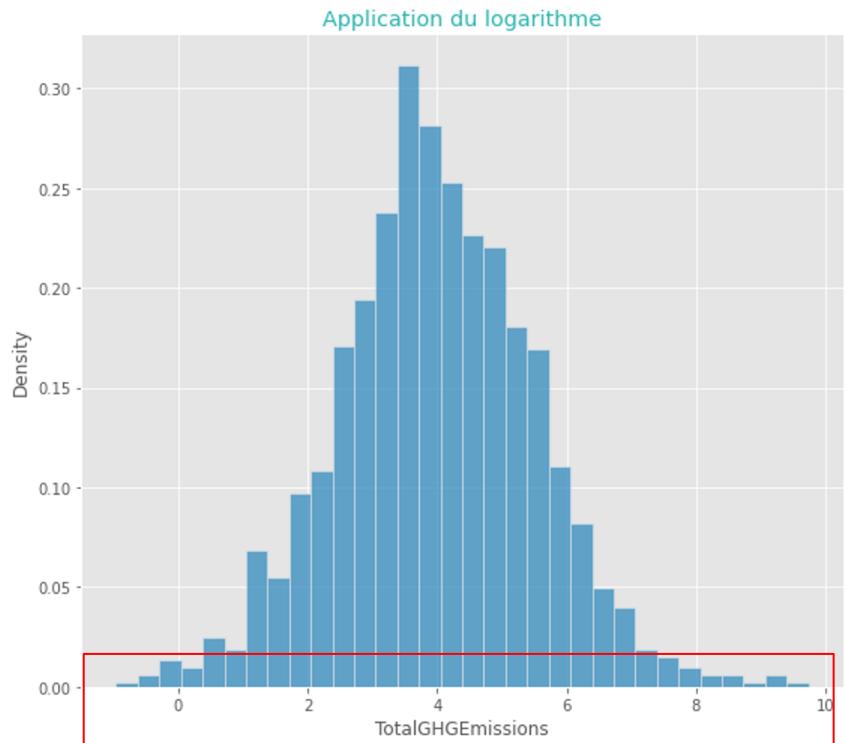
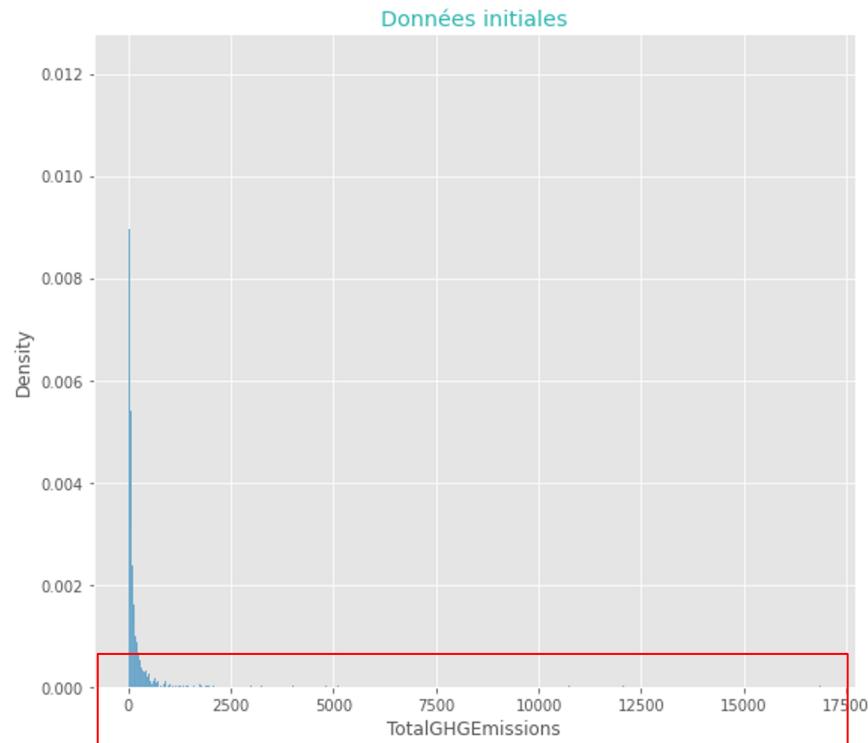
# Etude des bâtiments et features engineering (7/7) - age



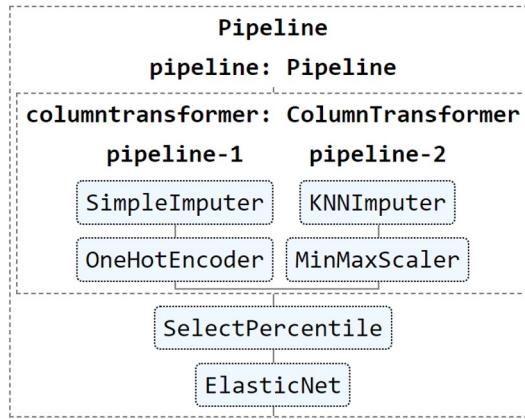
Et maintenant ?  
-> Choix du modèle  
-> Modélisation

# Modélisation - “Passage au log” de Y

Distribution des émissions de CO<sub>2</sub> avec changement d'échelle



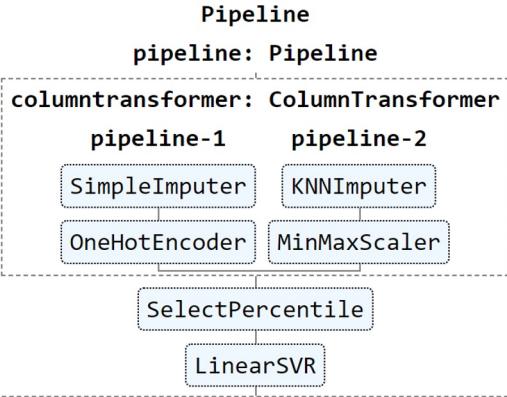
# Choix des modèles à tester (1/2) - Modèles linéaires



régression linéaire →



ElasticNet



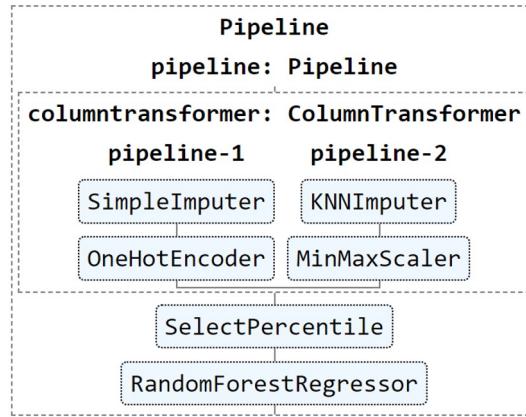
régression linéaire →



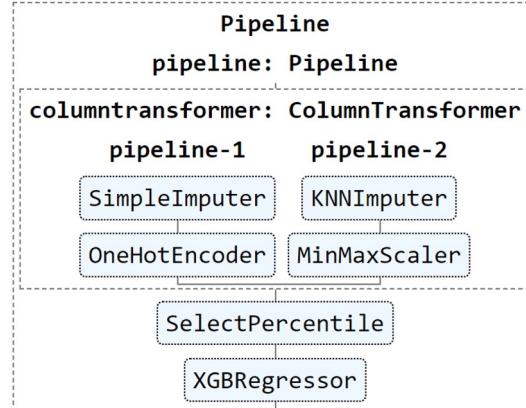
Linear Support Vector Regression

nb max of features  
= sqrt of 3093  
= 55

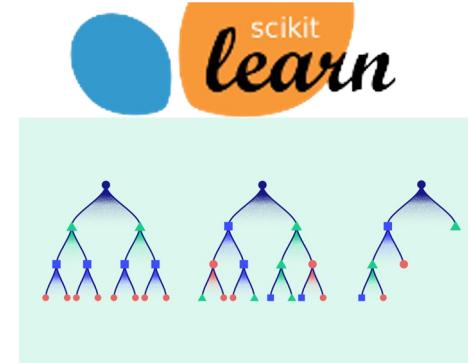
## Choix des modèles à tester (2/2) - Modèle non-linéaires



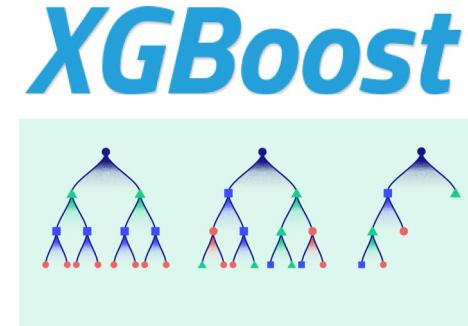
arbres de décision boosté



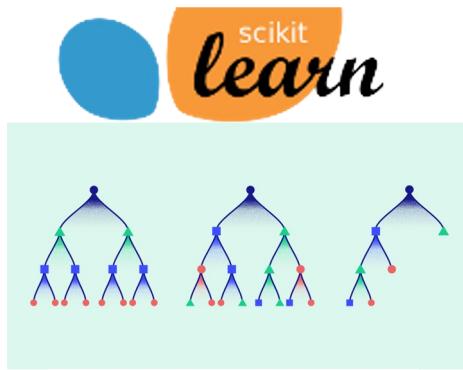
arbres de décision boosté



Random Forest Regressor

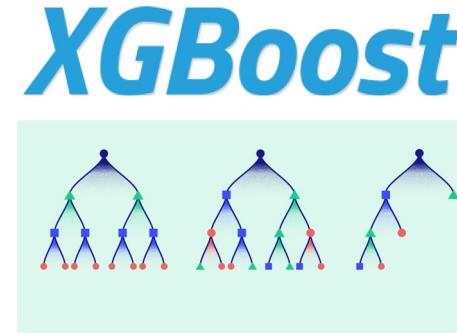


# Hyperparamètrage des modèles (GridSearch)



Random Forest Regressor

```
max_depth = 25  
min_samples_split = 2  
min_samples_leaf = 1  
bootstrap = False  
max_features = sqrt  
test_size = 0.25
```



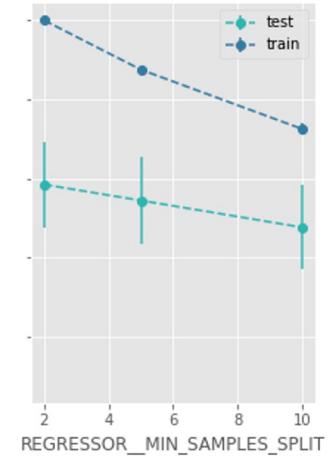
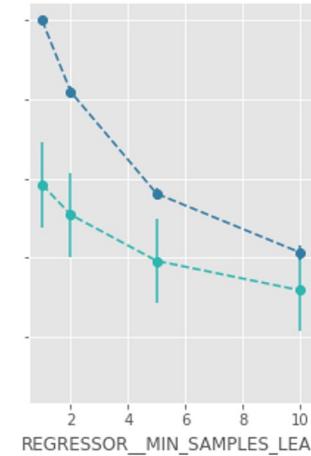
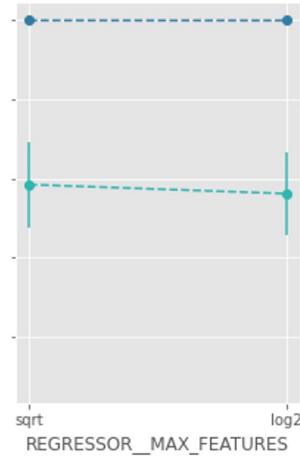
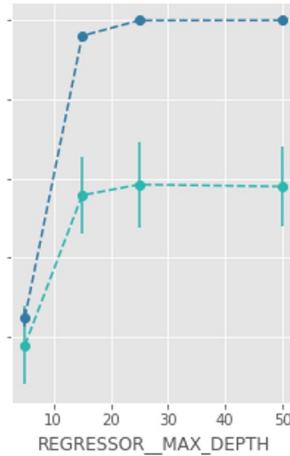
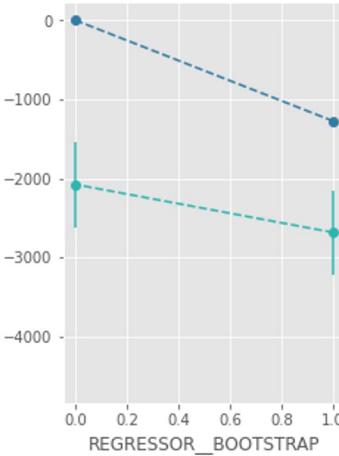
```
max_depth = 15  
n_estimators = 500  
learning_rate = 0.1  
min_child_weight = 10  
gamma = 0.25  
test_size = 0.25
```

# Hyperparamètrage des modèles (GridSearch)

- Random Forest Regressor

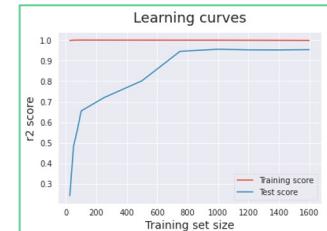
Scores par paramètres pour la variable TotalGHGEmissions

NEG MEAN ABSOLUTE ERROR SCORE



Et maintenant ?

- > Choix du modèle
- > Etude du modèle sélectionné

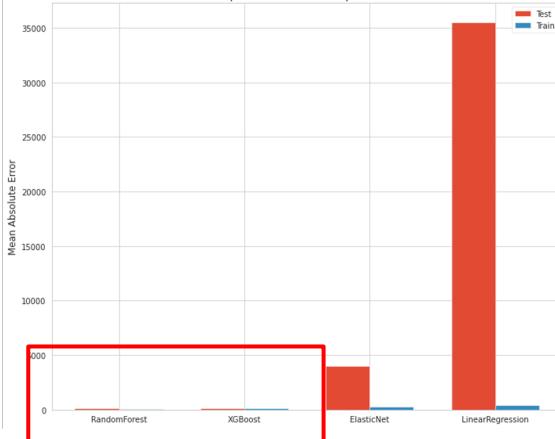


# Evaluation des modèles (Comparaison)

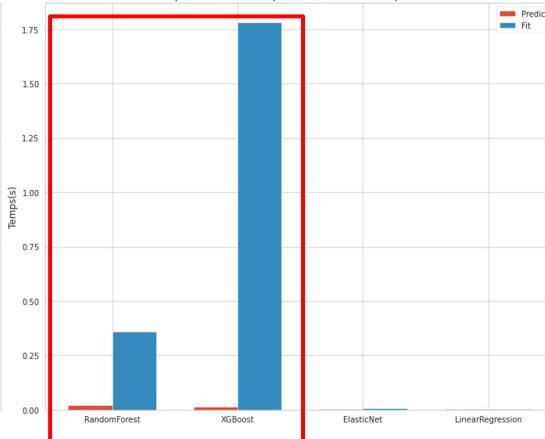
	fit_time	score_time	test_neg_mean_absolute_error	train_neg_mean_absolute_error
--	----------	------------	------------------------------	-------------------------------

RandomForest	0.539960	0.023493	-1.044920e+02	-82.659144
XGBoost	4.584033	0.015586	-1.267571e+02	-112.167805
LinearSVR	0.011123	0.002070	-1.292696e+03	-251.629375
ElasticNet	0.025989	0.002080	-2.834776e+05	-1053.521678
LinearRegression	0.007582	0.001832	-3.197074e+06	-6806.847949

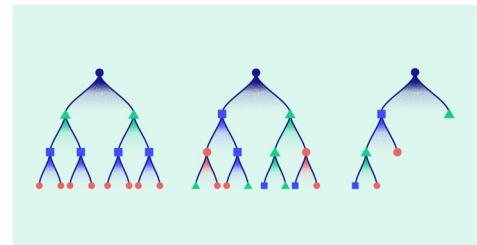
Comparaison des scores par modèle



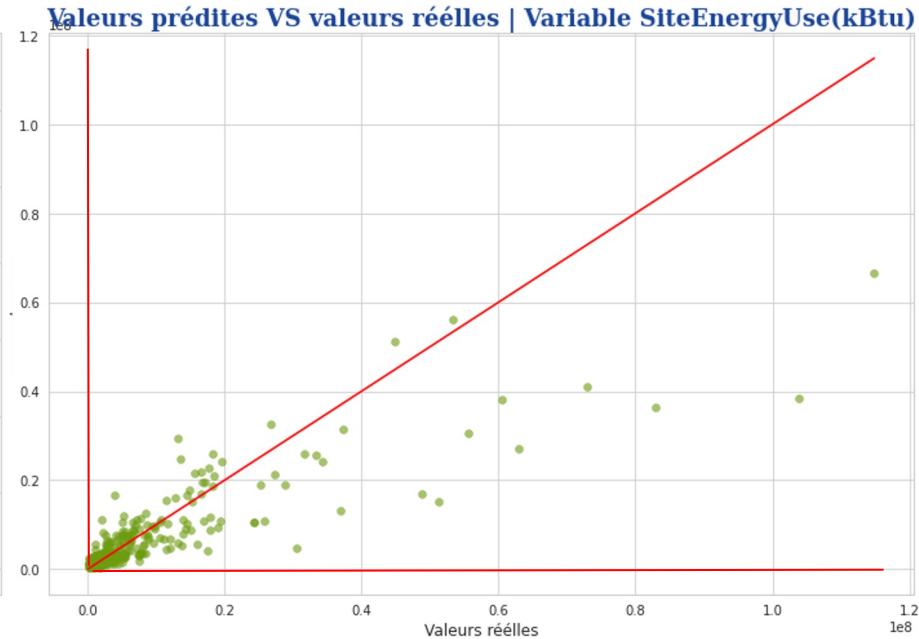
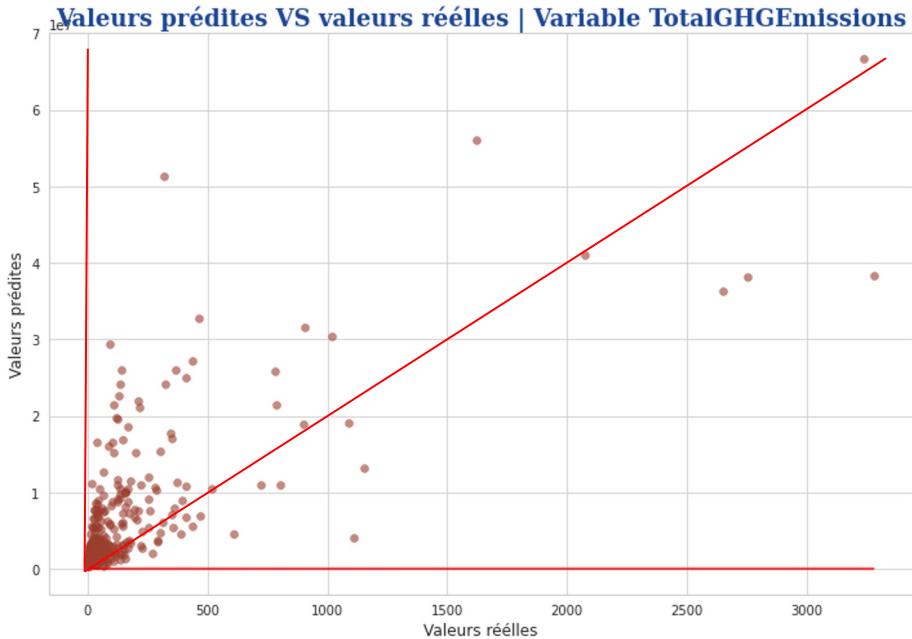
Comparaison des temps d'entraînement et prédiction



## RandomForest



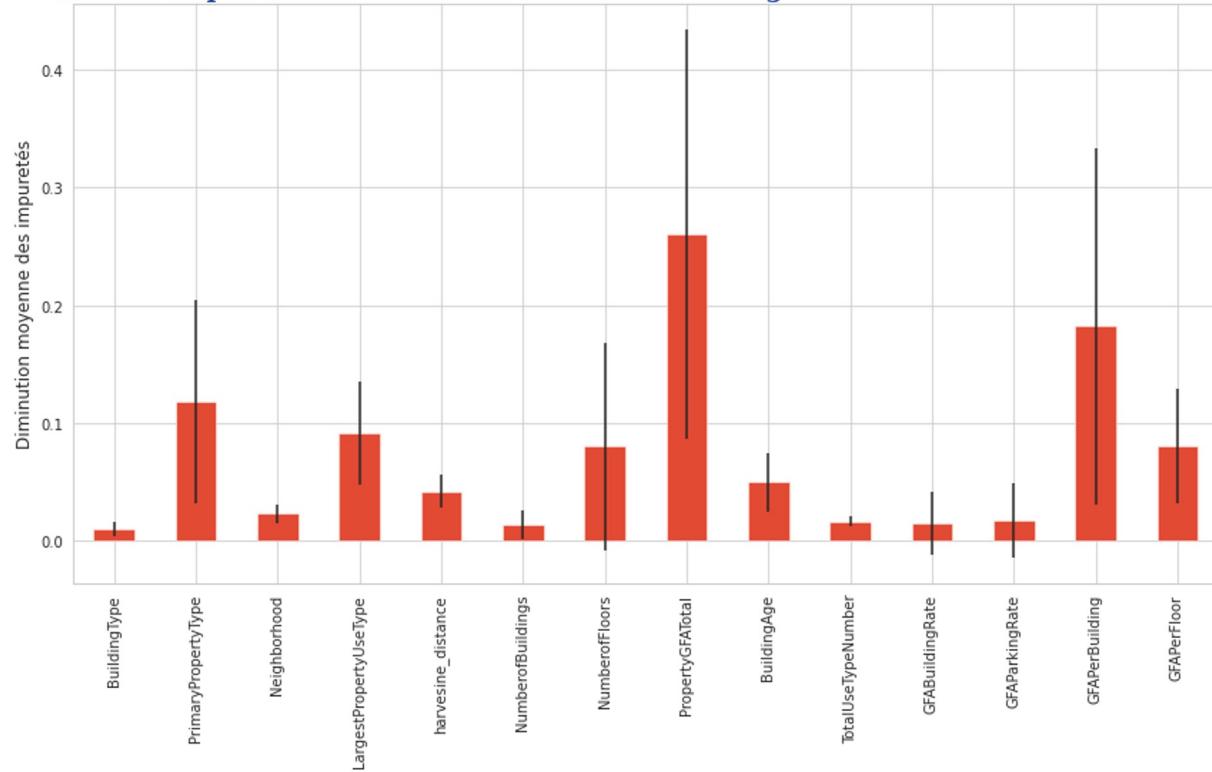
# Etude du modèle sélectionné (RandomForest) - (1/4)



-> SANS ENERGYSTAR SCORE

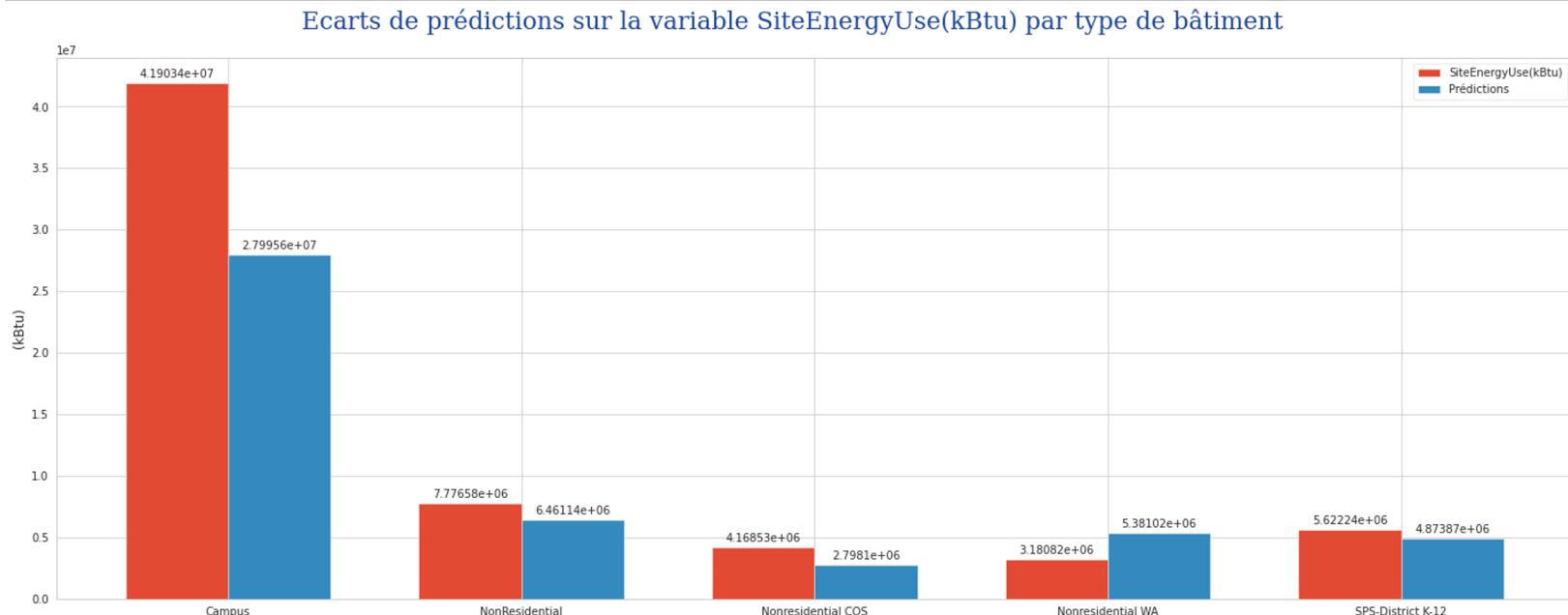
# Etude du modèle sélectionné (RandomForest) - (2/4)

Feature importances du modèle RandomForestRegressor sur les émissions de CO<sub>2</sub>



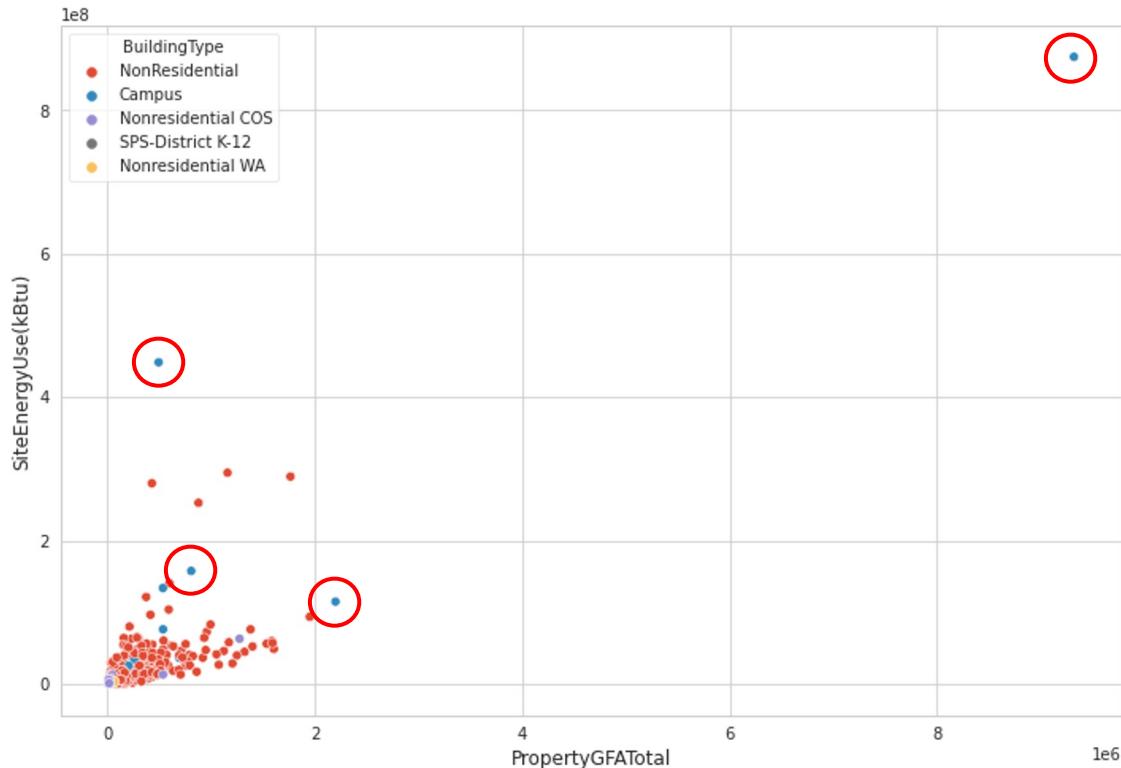
- Most Important Features :
- PropertyGFATotal
  - GFAPerBuilding
  - PrimaryPropertyType
  - LargestPropertyUseType
  - NumberofFloors
  - GFAPerFloor

# Etude du modèle sélectionné (RandomForest) - (3/4)



# Etude du modèle sélectionné (RandomForest) - (4/4)

Consommations d'énergie par surface totale au sol et par type de bâtiment



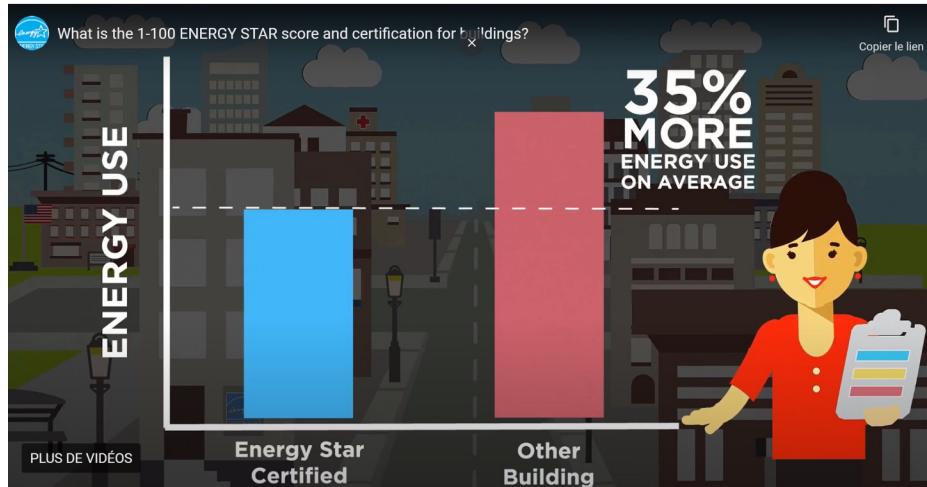
- Campus

Les campus sont distribués de manières moins prédictibles par notre algorithmes

Et maintenant ?

- > Impact de EnergyStar Score
- > Conclusion

# Qu'est ce que l'ENERGY STAR SCORE



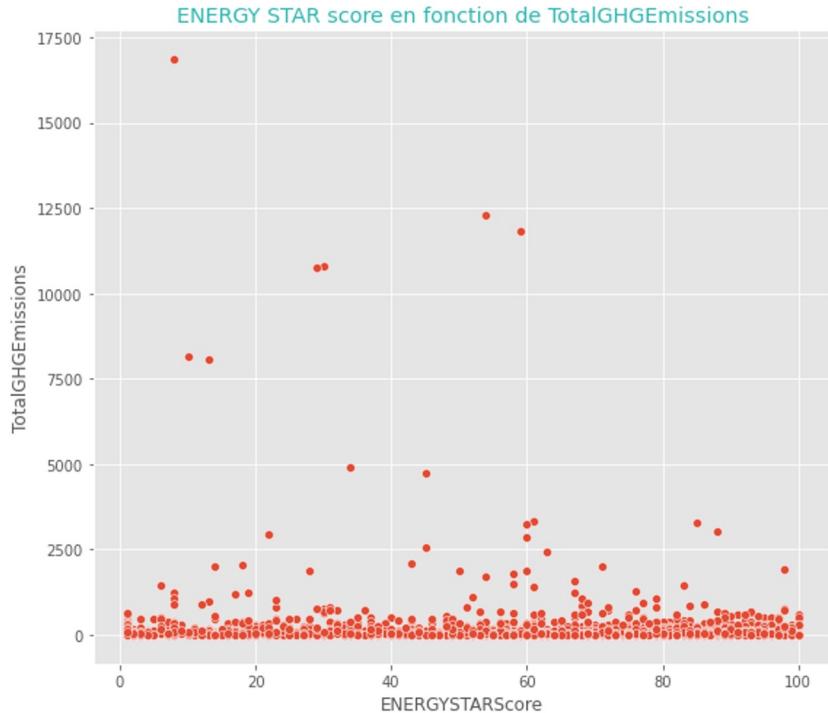
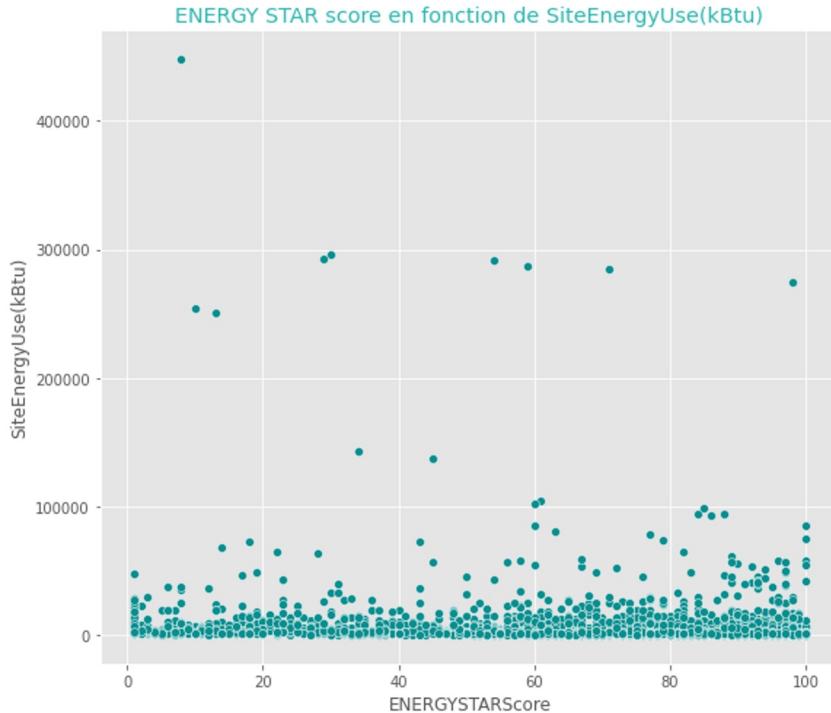
## ENERGY STAR Score Details for Buildings in the United States

- Data center
- Hospital (general medical and surgical)
- Hotel
- K-12 school
- Medical office
- Multifamily housing
- Office (covers office, bank branch, financial office, and courthouse)
- Parking
- Residence hall/ dormitory
- Retail store (covers retail and wholesale club/supercenter)
- Senior living community
- Supermarket/grocery store
- Swimming pool
- Warehouse (covers distribution center, non-refrigerated warehouse, and refrigerated warehouse)
- Wastewater treatment plant
- Worship facility

# Analyse de l'Energy Star Score (1/4)

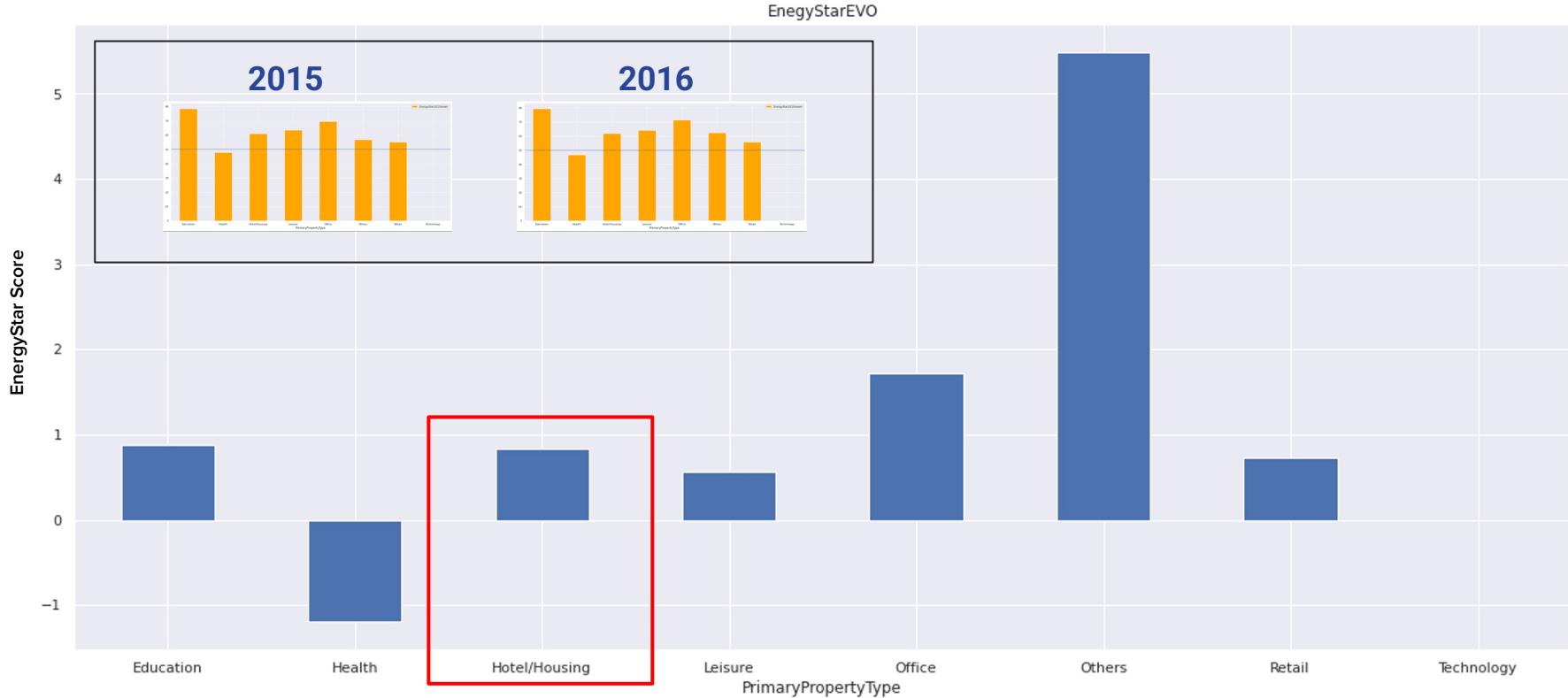


## Analyse de la variable ENERGY STAR Score



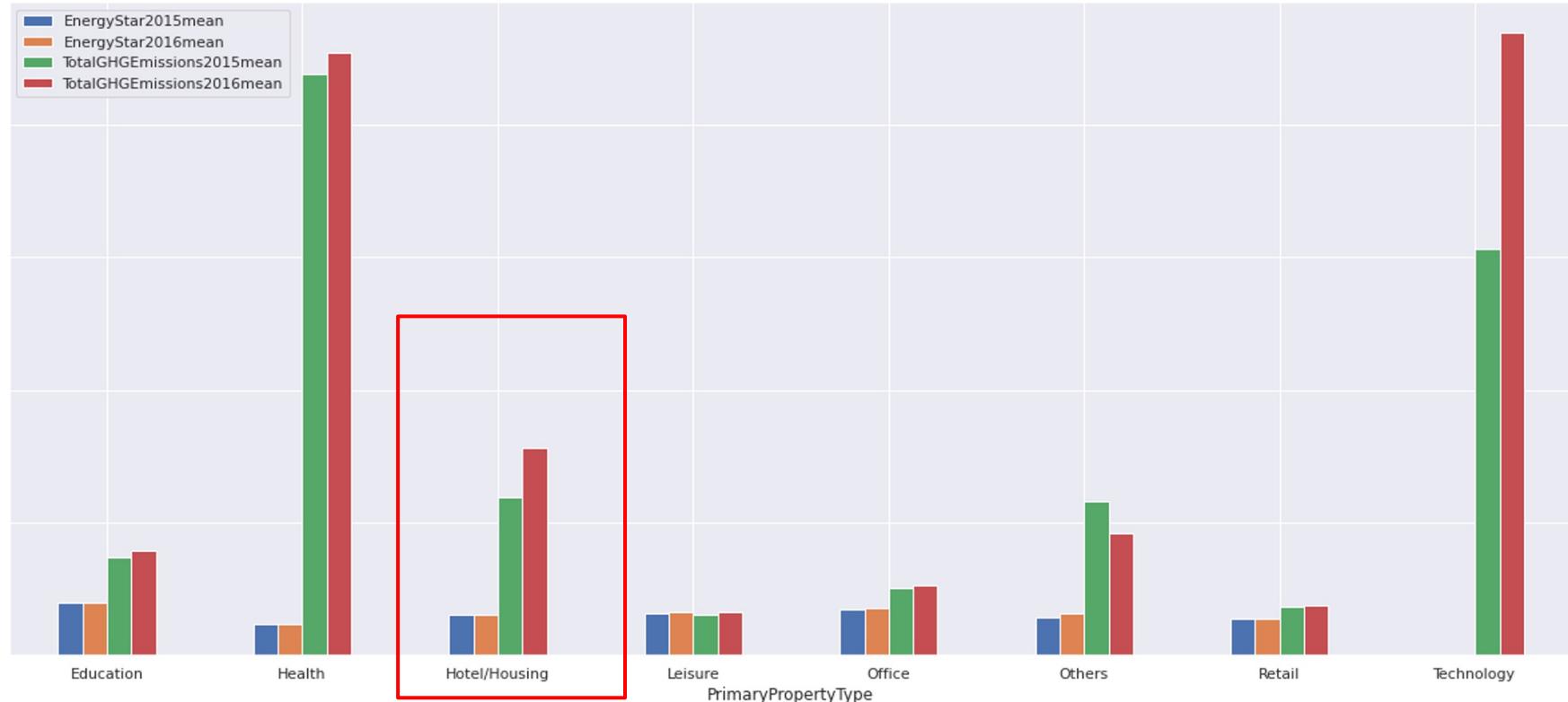


# Analyse de l'Energy Star Score (2/4)



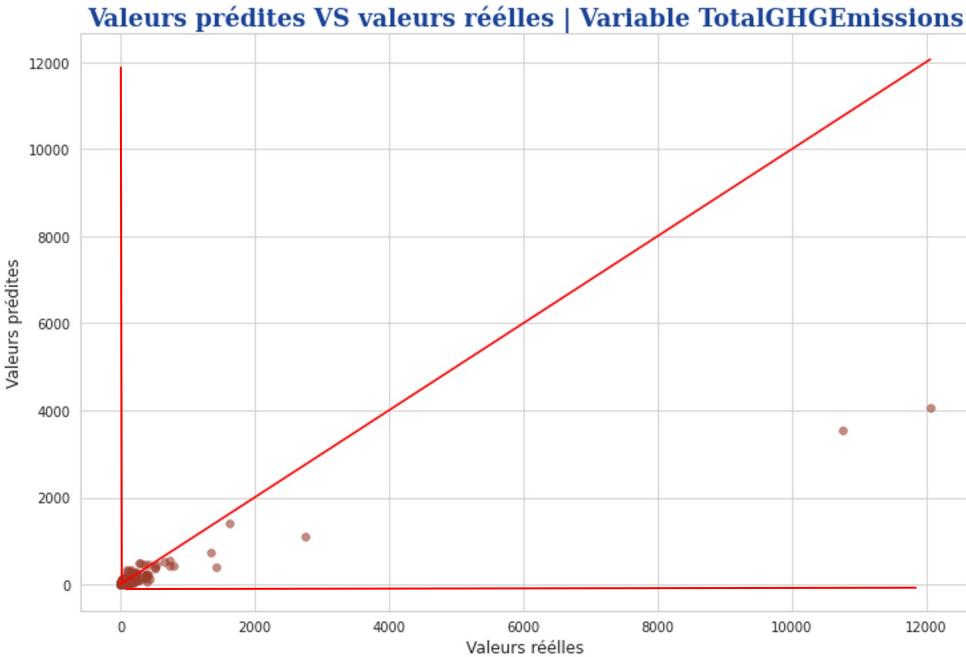


# Analyse de l'Energy Star Score (3/4)





# Analyse de l'Energy Star Score (4/4)



## Améliorations:

- les métriques se sont améliorées grâce à la prise en compte de l'ENERGY STAR Score
- Les valeurs prédictes sont plus resserrées sur la première bissectrice

## Faiblesses:

- Cette variable est peu renseignée et le jeu de données comporte peu d'entrées. Il est donc difficile de savoir si cette amélioration est réellement significative
- Il faut également prendre en compte le bénéfice vis à vis du coût de cet ENERGY STAR Score

-> AVEC ENERGYSTAR SCORE

# Conclusions

1- Peut-on bien modéliser la consommation des bâtiments de SEATTLE sans appels aux agents tiers :

-> La consommation d'un bâtiments peut-être bien modélisée à l'aide d'informations fiables telles que la surface totale du bâtiment ou son type d'utilisation principal

2 - L'Energy Star Score est-elle indispensable à la modélisation de nos consommations :

-> L'Energy Star Score est un score relatif et non absolu, il nous faut trouver un meilleur référentiel de scoring

3 - La ville de SEATTLE peut-elle atteindre son objectif de neutralité carbone d'ici 2050 :

-> Cette question requiert d'obtenir des relevés de consommation sur un plus grand nombre d'années afin de modéliser un comportement global, en revanche les économies réalisées par notre algorithme pourraient permettre de mettre en place des plans de politique de réduction de consommation par la ville