

# Anticipation de la consommation d'énergie et des émissions de CO<sub>2</sub>

Cécile Guillot, Ingénieure Machine Learning





01

...  
INTRODUCTION

02

...  
ANALYSES  
DESCRIPTIVES

03

...  
MACHINE LEARNING  
ET MODELISATION

04

...  
CONCLUSION

# 01

...

## INTRODUCTION

Problématique, hypothèses et pistes  
d'exploration



# Contexte de l'étude

**Objectif** : Être une ville neutre en émissions de gaz à effets de serre pour 2050

*Données* : Deux jeux de données sur les bâtiments non-résidentiels (2015 & 2016)



- Prédire la consommation d'énergie
- Prédire les émissions de CO2
- Intérêt de l'utilisation du score ENERGY STAR

# Contexte de l'étude (2)

- Utilisation des données liées aux informations inscrites sur le permis de construire :
  - Usage de la propriété
  - Date de construction (ou de gros travaux)
  - Nombre de bâtiments et d'étages
  - Superficie de la propriété (bâtiments, parking et autres)
  - Localisation (quartier, adresse et géolocalisation)
- Pistes d'explorations :
  - Prédire la consommation d'énergie à l'aide des informations disponibles
  - Prédire les émissions de CO2 avec les informations des permis de construire
    - Avec prise en compte du ENERGY STAR Score
    - Sans prise en compte du ENERGY STAR Score



# Déroulement de l'étude

## Nettoyage des données

- Réorganisation des données
- Retrait des valeurs aberrantes

## Analyses exploratoires

- Analyses univariées, multivariées
- Permet de dégager des insights

## Modélisation avec algorithme de régression

- Essais de plusieurs modèles
- Choix et optimisation d'un modèle



# 02



## ANALYSES EXPLORATOIRES

Analyses statistiques descriptives  
univariées et bivariées



# Traitement et nettoyage des données

- Rassemblement des données de 2015 et 2016 :
  - Harmonisation des chaînes de caractères dans les données de 2015
  - Suppression des colonnes avec plus de 95% de valeurs manquantes
- Harmonisation des chaînes de caractères (présence de retour chariot, etc.)
- Retrait des valeurs aberrantes
  - Retrait des bâtiments avec des consommations négatives
  - Cas du *Bullitt Center* (<https://bullittcenter.org/>)
  - Retrait de l'université de Washington
  - Retrait du campus Boeing



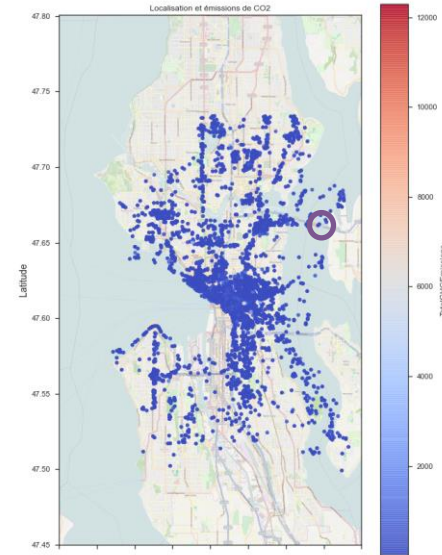
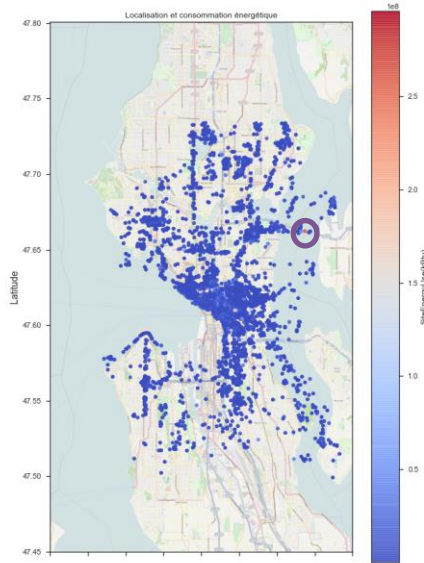


# Feature Engineering

- Simplification de variables
  - Rassemblement de catégories de bâtiments
  - Office, Warehouse, K-12 School, Supermarket
- Création de nouvelles variables :
  - HasParking (Yes/No)
  - Clusters (0 => 4)
  - Age (Année – Année de construction)
  - Température (Degree Day)
    - Equivalent des DJU en France
    - Une valeur : Chaud

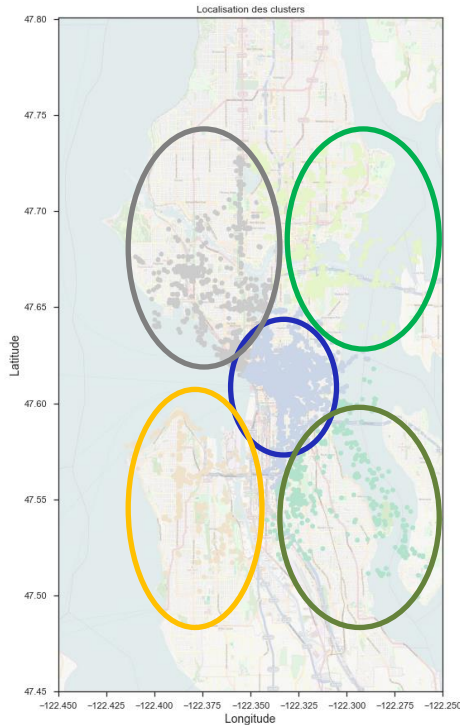


# Données géographiques



- Grande concentration de bâtiments non-résidentiels dans le centre-ville (Downtown)
- Bâtiments les plus polluants : Eloignées des autres

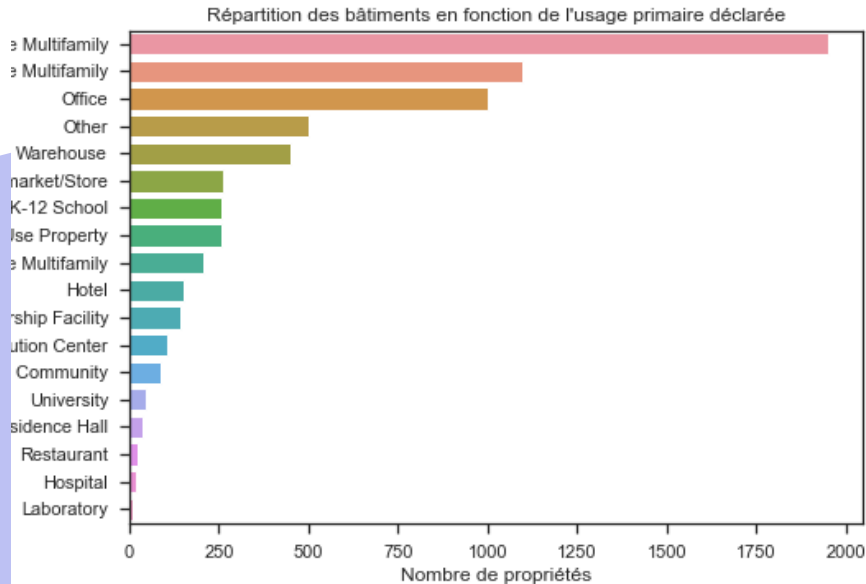
## Données géographiques (2)



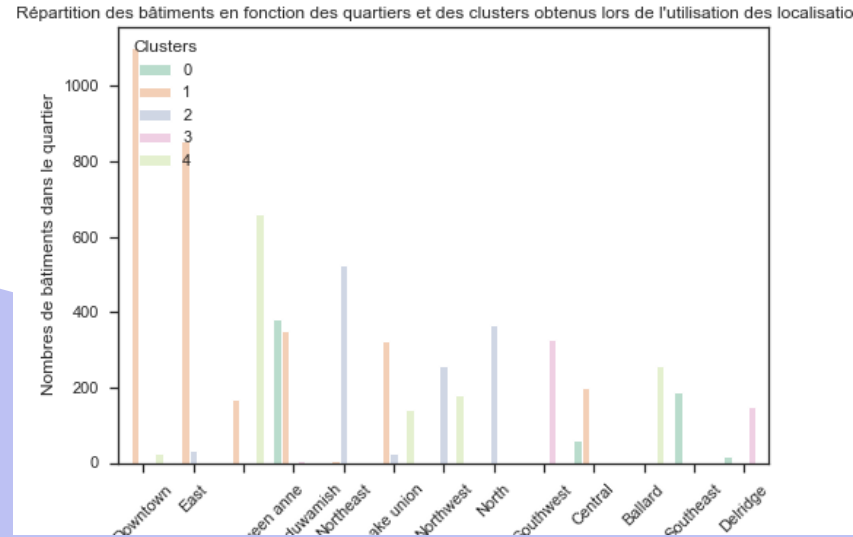
- Découpage en 5 clusters : Centre, Nord-Est, Nord-Ouest, Sud-Est, Sud-Ouest
- Détermination du nombre de clusters avec méthode du coude

# Analyses univariées

- Répartition des bâtiments

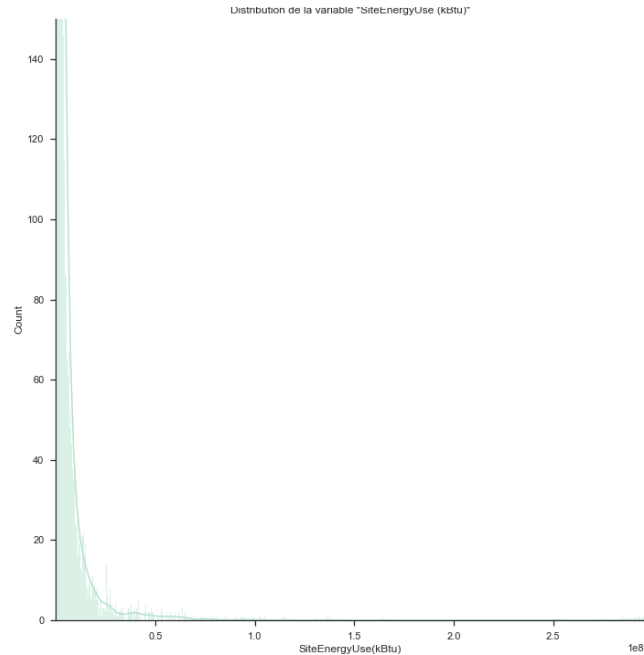


- Répartition en fonction des clusters

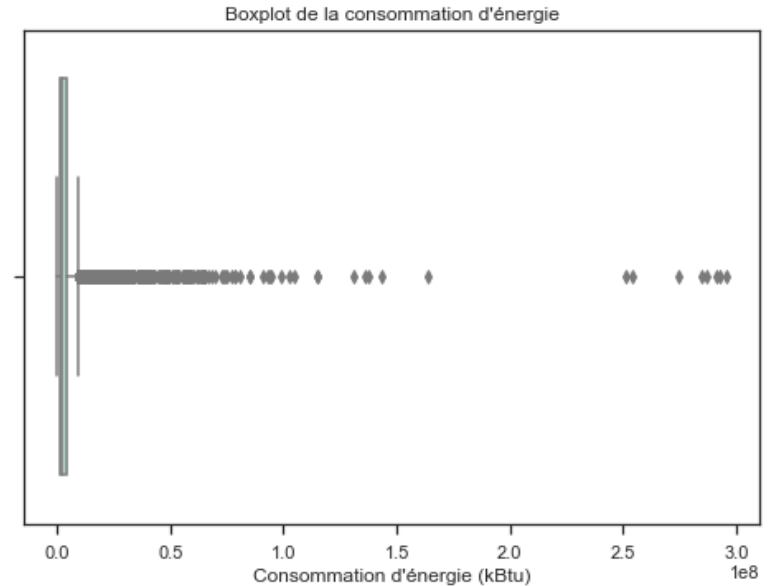


# Analyses univariées (2)

- Distribution de la variable « Site Energy Use »

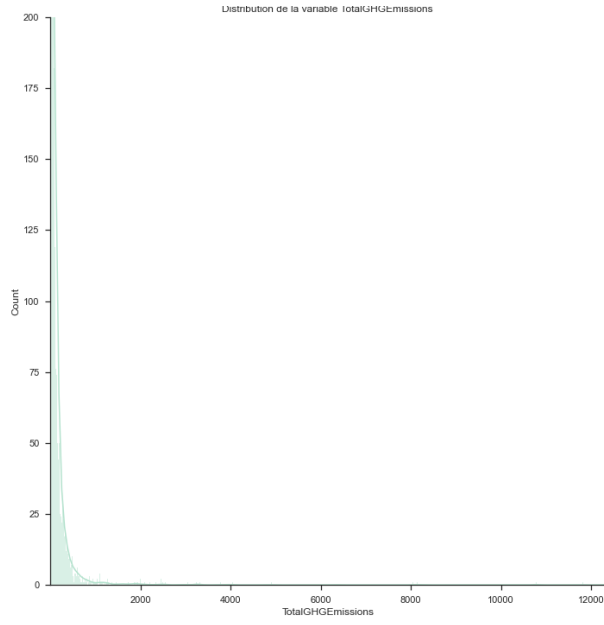


- Boîte à moustache de la variable « Site Energy Use »

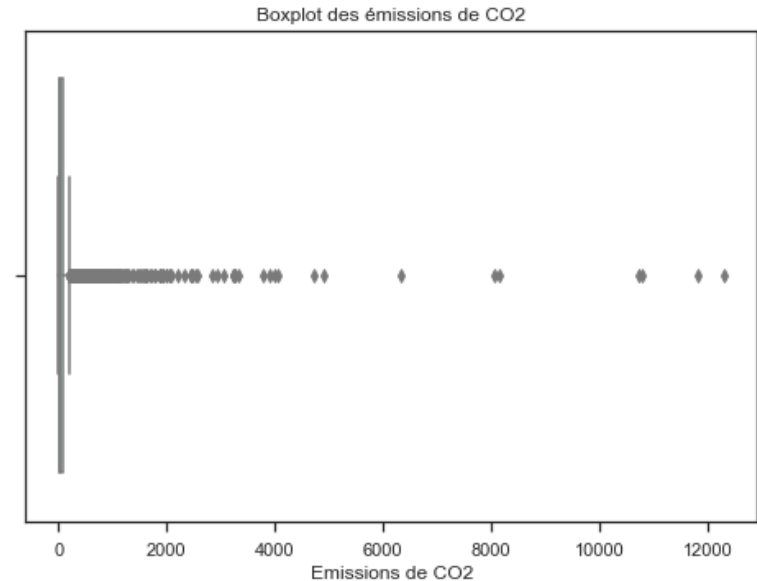


# Analyses univariées (3)

- Distribution de la variable « Total GHG Emissions »

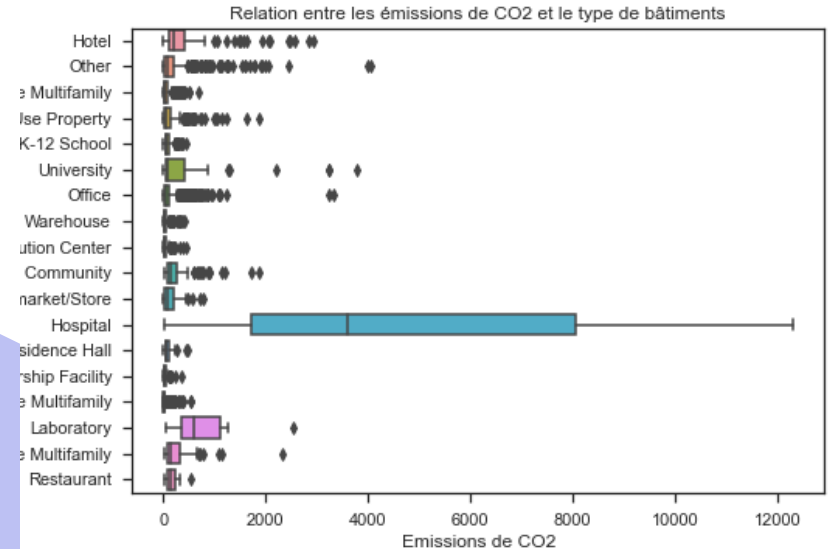
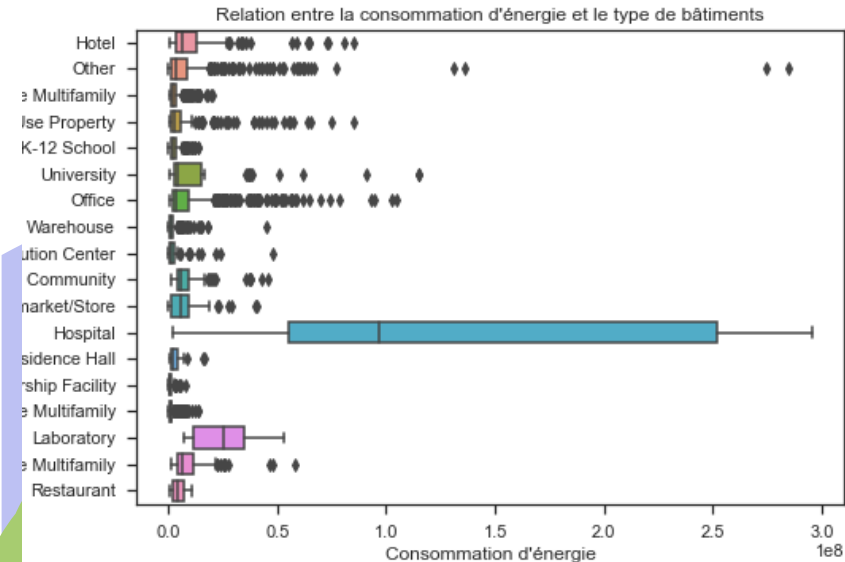


- Boîte à moustache de la variable « Total GHG Emissions »



# Analyses bivariées

- Les hôpitaux sont les bâtiments les plus polluants...
- ...Suivi par les laboratoires



# Analyses bivariées (2)

- Consommation d'énergie

Var.	Coef. de corrélation
Emissions de CO2	0.88
PropertyGFATotal	0.66
SecondLargestPropertyUseGFA	0.54
NumberofFloors	0.34
ThirdLargestPropertyUseGFA	0.24

- Emissions de CO<sub>2</sub>

Var.	Coef. de corrélation
Conso. d'énergie	0.88
PropertyGFATotal	0.50
SecondLargestPropertyUseGFA	0.40
NumberofFloors	0.20
NumberofBuildings	0.19





# 03

...

## MODELISATION

Recherche d'un algorithme de  
régression



# Méthodologie

- Pré-traitement des données :
  - Création de deux jeux de données stratifiés sur le type de bâtiments (30% de bureaux)
  - Séparation des variables catégorielles et numériques
    - Variables catégorielles : *Imputer* → Mode - *OneHotEncoding*
    - Variables numériques : *Imputer* → Médiane - *StandardScaler*
  - Création d'un pipeline de pré-traitement
- Tests de plusieurs modèles de machine learning :
  - Linéaire (Régression linéaire, Lasso, Ridge, SVM), non-linéaire (Ridge Kernel) et ensembliste (Random Forest, XGBoost, Adaboost, etc.)
- Recherche des hyperparamètres les plus optimisés :
  - Grille de recherche + Validation croisée
- Evaluation du modèle :
  - MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), RMPSE (% de RMSE), MAPE (Mean Absolute Percentage Error) &  $R^2$  (Coefficient de détermination)
- Enregistrement dans un pipeline :
  - Pré-traitement + modèle avec hyperparamètres

# Prédiction de la consommation d'énergie



# Test de plusieurs modèles

- Test de plusieurs modèles
- Aucun paramètre fixé (sauf *random\_state*)
- Utilisation de **MAE**, **MSE**, **RMSE** et **R<sup>2</sup>** pour choisir un modèle

	Dummy Regressor	Linear Regression	Ridge	Lasso	DecisionTree	SVM	Ridge Kernel	AdaBoost	Bagging	GradientBoosting	Random Forest	XGBoost
MAE	0.895	4.189360e+07	0.543	0.903	0.244	0.401	0.545	0.519	0.281	0.396	0.267	0.325
MSE	1.306	7.734025e+17	0.495	1.270	0.208	0.305	0.496	0.454	0.177	0.295	0.158	0.202
RMSE	1.143	8.794331e+08	0.703	1.127	0.456	0.553	0.704	0.673	0.420	0.543	0.398	0.449
R <sup>2</sup>	-0.028	-6.090599e+17	0.610	-0.001	0.836	0.759	0.609	0.643	0.861	0.768	0.876	0.841


- Problème de régression non linéaire
- Choix d'un modèle ensembliste
- **RandomForest**

# Hyperparamètres du modèle

- Random Forest :

- Méthode ensembliste
- « Wisdom of crowd »
- Entraînement de plusieurs arbres de décision
- Calcul de la prédiction moyenne de chaque arbre

- Hyperparamètres

- Nombre d'arbres (n\_estimators)
  - Nombre de variables (max\_features)
  - Profondeur des arbres (max\_depth)
  - Nombre de valeurs minimales pour un nœud (min\_samples\_split)
  - Nombre de valeurs minimales pour une feuille (min\_samples\_leaf)
  - Méthode de bootstrapping (bootstrap)
- 

# Algorithme final : RandomForest

RandomForestRegressor

```
RandomForestRegressor(max_depth=100, min_samples_leaf=2, min_samples_split=3,  
                      n_estimators=550, n_jobs=-1, random_state=42)
```

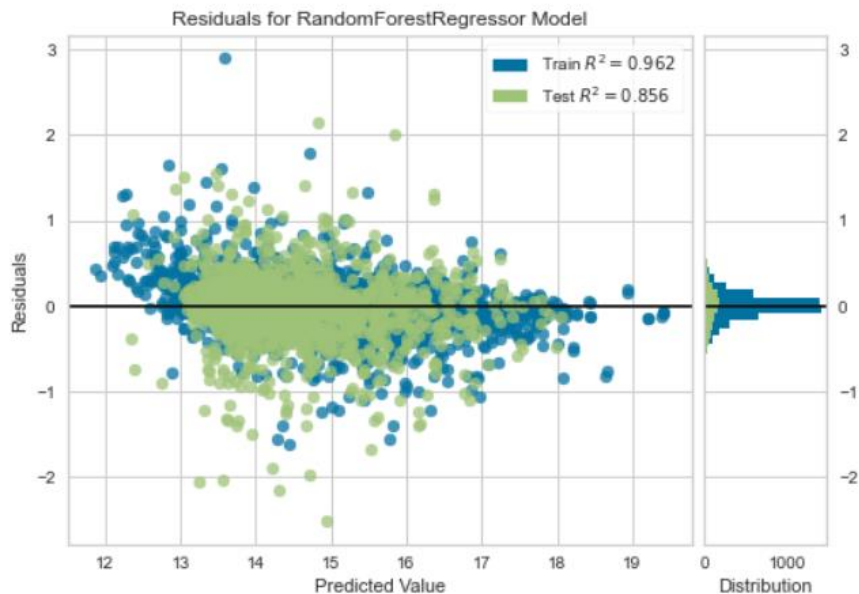


## Feature Importances

Features	Score
Supermarket/Store	0.72
Vocational School	0.06
No (HasParking)	0.022
ThirdLargestPropertyUseGFA	0.022
Office	0.021

# Evaluation du modèle

## *Analyse des résidus*



## *Evaluation des performances*

- **MAE** : 0.29
- **MSE** : 0.18 ( $\log^2$ )
- **RMSE** : 0.42 (log)
- **RMPSE** : 0.10%
- **MAPE** : 0.02%
- **$R^2$**  : 0.86



Prédiction des émissions de CO<sub>2</sub> :

- 1) Sans score Energy Star
- 2) Avec score Energy Star



# Test de plusieurs modèles

- Test de plusieurs modèles
- Aucun paramètre fixé (sauf *random\_state*)
- Utilisation de **MAE**, **MSE**, **RMSE** et **R<sup>2</sup>** pour choisir un modèle

	Dummy Regressor	Linear Regression	Ridge	Lasso	DecisionTree	SVM	Ridge Kernel	AdaBoost	Bagging	GradientBoosting	Random Forest	XGBoost
MAE	1.219	1.538226e+08	0.894	1.221	0.426	0.711	0.894	0.920	0.487	0.722	0.474	0.554
MSE	2.223	6.256071e+18	1.258	2.228	0.669	0.909	1.261	1.226	0.496	0.853	0.452	0.563
RMSE	1.491	2.501214e+09	1.122	1.493	0.818	0.953	1.123	1.107	0.704	0.923	0.673	0.751
R <sup>2</sup>	-0.007	2.834794e+18	0.430	-0.010	0.697	0.588	0.429	0.444	0.775	0.614	0.795	0.745

- Problème de régression non linéaire
- Choix d'un modèle ensembliste
- **RandomForest**

# Algorithme final : Sans score Energy Star

- Hyperparamètres de l'algorithme:

Hyperparamètres	Valeurs
Bootstrap	True
max_depth	80
max_features	auto
min_samples_leaf	2
min_samples_split	3
n_estimators	650

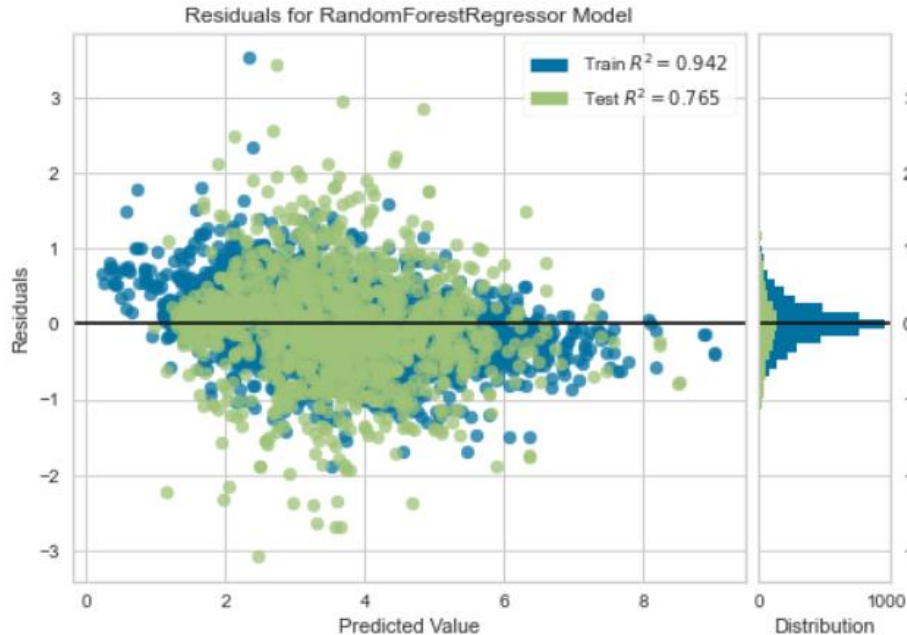


## Feature Importances

Features	Score
Strip Mall	0.52
Swimming Pool	0.15
ThirdLargestPropertyUseGFA	0.06
No (HasParking)	0.036
Worship Facility	0.035

# Evaluation du modèle : Sans score Energy Star

## Analyse des résidus



## Evaluation des performances

- **MAE** : 0.52
- **MSE** :  $0.51(\log^2)$
- **RMSE** : 0.72 (log)
- **RMPSE** : 2.6%
- **MAPE** : 0.25%
- **$R^2$**  : 0.77

# Test de plusieurs modèles

- Test de plusieurs modèles
- Aucun paramètre fixé (sauf *random\_state*)
- Utilisation de **MAE**, **MSE**, **RMSE** et **R<sup>2</sup>** pour choisir un modèle

	Dummy Regressor	Linear Regression	Ridge	Lasso	DecisionTree	SVM	Ridge Kernel	AdaBoost	Bagging	GradientBoosting	Random Forest	XGBoost
MAE	1.219	8.030599e+08	0.880	1.221	0.468	0.684	0.881	0.941	0.498	0.702	0.489	0.544
MSE	2.223	1.705130e+20	1.204	2.228	0.714	0.841	1.207	1.243	0.497	0.789	0.461	0.539
RMSE	1.491	1.305806e+10	1.097	1.493	0.845	0.917	1.099	1.115	0.705	0.888	0.679	0.734
R <sup>2</sup>	-0.007	-7.726400e+19	0.455	-0.010	0.676	0.619	0.453	0.437	0.775	0.642	0.791	0.756

- Problème de régression non linéaire
- Choix d'un modèle ensembliste
- **RandomForest**

# Algorithme final : Avec score Energy Star

- Hyperparamètres de l'algorithme:

Hyperparamètres	Valeurs
Bootstrap	True
max_depth	80
max_features	Auto
min_samples_leaf	2
min_samples_split	2
n_estimators	650



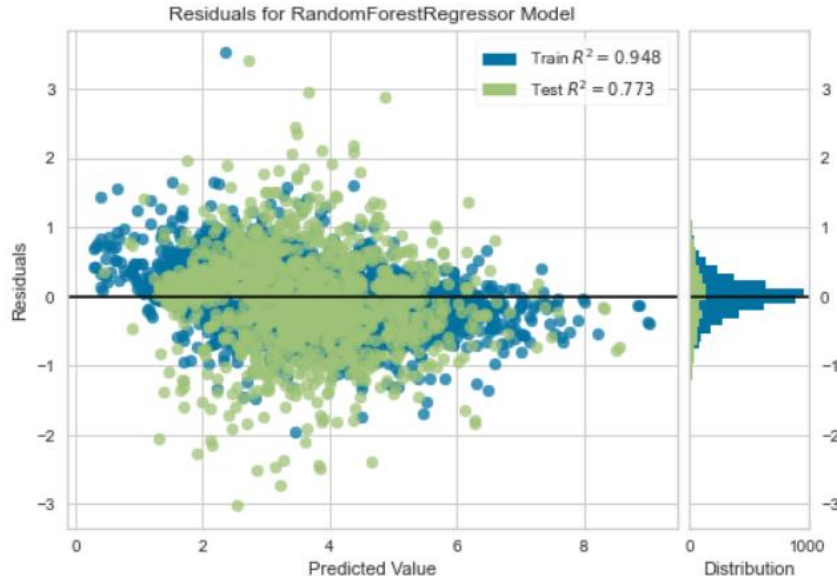
## Feature Importances

Features	Score
Residence Hall	0.49
Social/Meeting Hall	0.14
Worship Facility	0.07
ThirdLargestPropertyUseGFA	0.06
Supermarket/Store	0.03



# Evaluation du modèle : Avec score Energy Star

## *Analyse des résidus*



## *Evaluation des performances*

- **MAE** : 0.52
- **MSE** : 0.50 ( $\log^2$ )
- **RMSE** : 0.70 (log)
- **RMPSE** : 2.6%
- **MAPE** : 0.21%
- **$R^2$**  : 0.77

# 04

...

## CONCLUSION



# Choix du modèle à déployer

## Concernant la consommation d'énergie :

- Informations suffisantes pour la mise en place d'une prédiction
- Modèle avec de bonnes performances (80% de variance expliquée, 0.1% d'erreur dans les prédictions)

## Concernant les émissions de CO2 :

- Ajout du score Energy Star : Amélioration de certaines performances
- Beaucoup de valeurs manquantes → Imputation de la médiane
- Peu d'intérêt d'utiliser le score Energy Star
- Performances moins bonnes que pour la consommation d'énergie
- Manque d'informations pour construire ce modèle ?

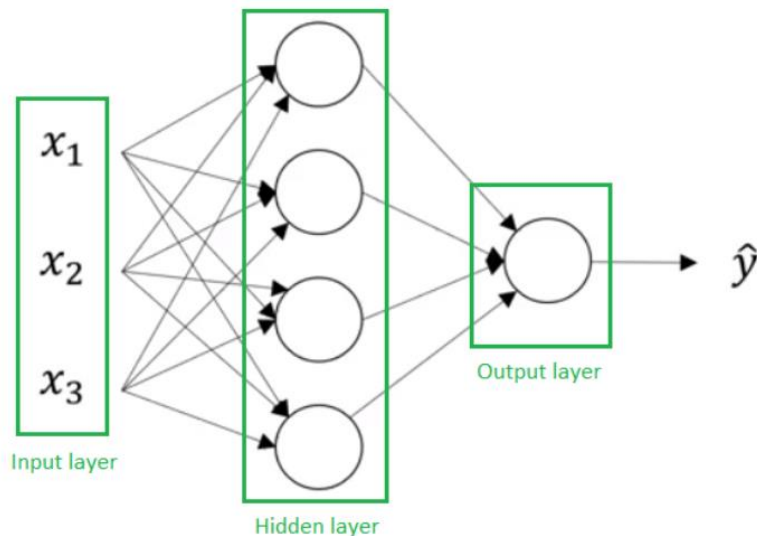


# Conclusion

- Possibilité de prédire la consommation d'énergie de manière fiable
- Possibilité de prédire les émissions de CO2 mais légèrement moins fiable
  - Score Energy Star : inutile à la prédiction
- Est-ce suffisant pour arriver à un objectif d'une ville neutre en carbone pour 2050 ?
- Hôpitaux et laboratoires : Bâtiments les plus « polluants »
  - Rénover ces bâtiments sur le modèle des bâtiments à énergie positive
- Downtown (centre-ville) : Beaucoup de bureaux
  - Imposer des normes pour réduire la consommation d'énergie (exemple du *Bullitt Center*)
- Mise à jour du modèle en intégrant les bâtiments à basse consommation

# Bonus : Prédiction des émissions de CO2 via un réseau de neurones

- Modèle séquentiel avec deux couches cachées
- Fonction d'activation : « relu »



## *Evaluation des performances*

- **MAE** : 0.66
- **MSE** : 0.91 ( $\log^2$ )
- **RMSE** : 0.95 (log)
- **RMPSE** : 1.16%
- **MAPE** : 0.28%
- **R<sup>2</sup>** : 0.58

# Références

- Différence entre source et site ([https://www.energystar.gov/buildings/benchmark/understand\\_metrics/source\\_site\\_difference](https://www.energystar.gov/buildings/benchmark/understand_metrics/source_site_difference))
- Définition du score ENERGY STAR ([https://www.energystar.gov/buildings/benchmark/analyze\\_benchmarking\\_results](https://www.energystar.gov/buildings/benchmark/analyze_benchmarking_results))
- Définition des types de bâtiments (<https://portfoliomanager.energystar.gov/pm/glossary>)
- Site de la ville de Seattle (<https://data.seattle.gov/dataset/2015-Building-Energy-Benchmarking/h7rm-fz6m>)
- Fuite de données (<https://machinelearningmastery.com/data-leakage-machine-learning/>)





CentraleSupélec

Merci pour votre attention !