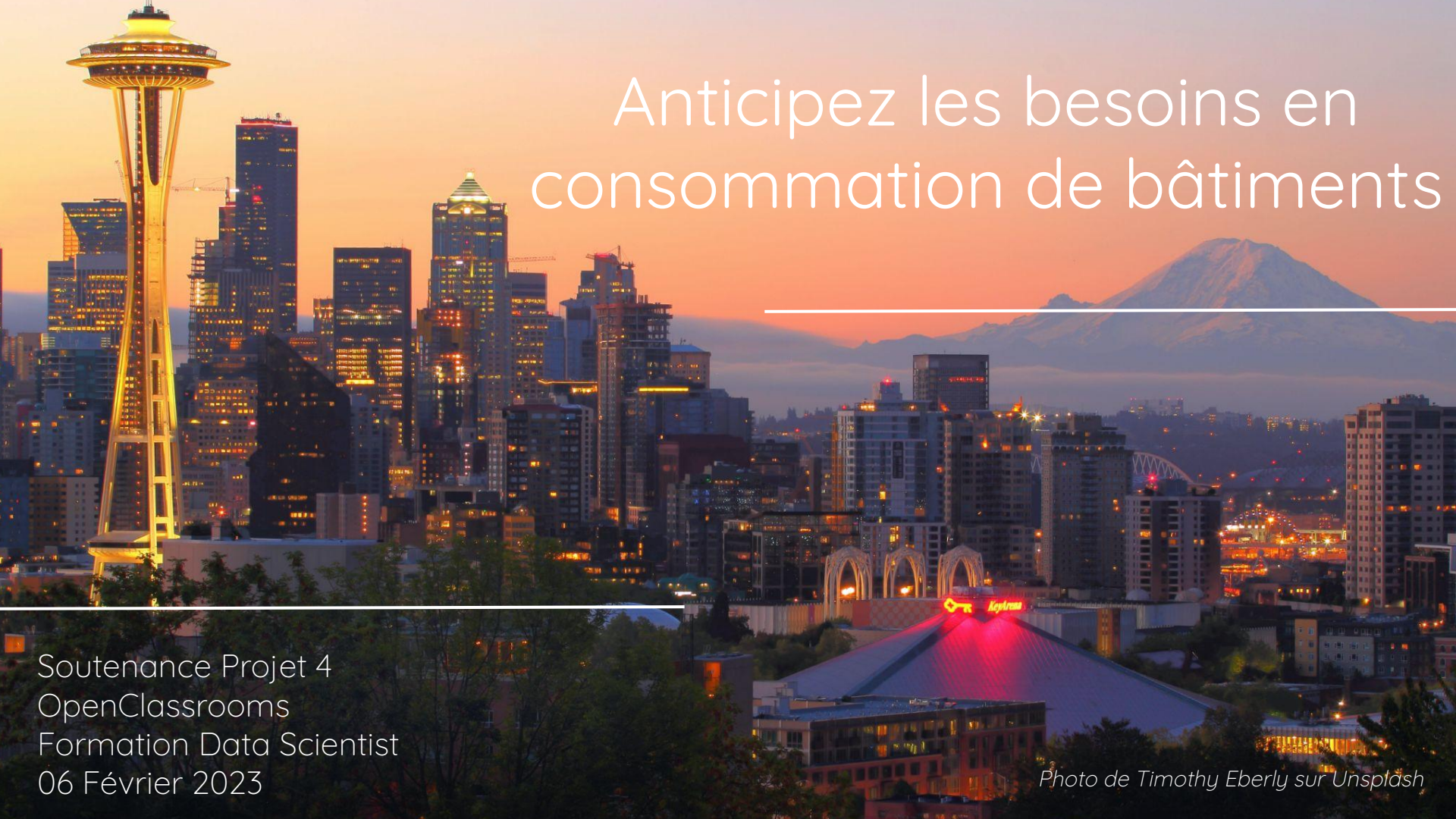
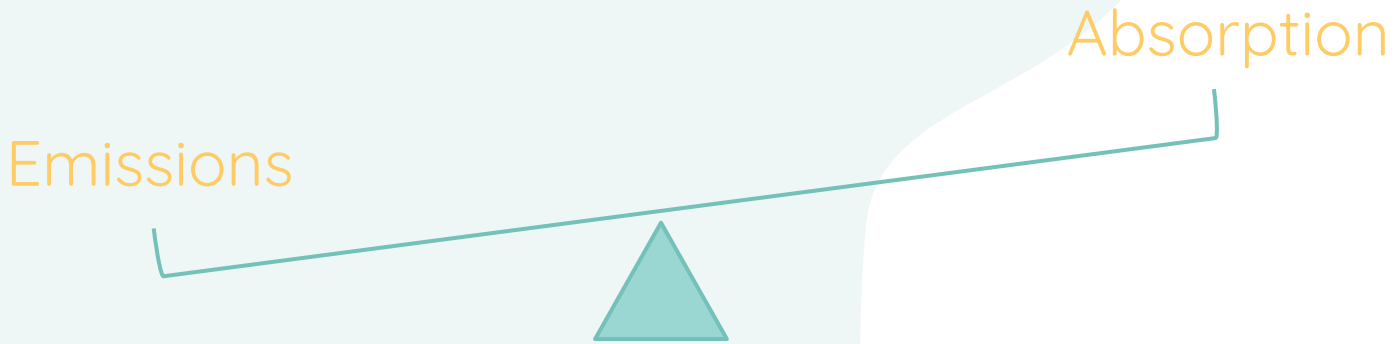
# Anticipez les besoins en consommation de bâtiments



Soutenance Projet 4
OpenClassrooms
Formation Data Scientist
06 Février 2023

*Photo de Timothy Eberly sur Unsplash*

# Neutralité carbone 2050




Bâtiments non résidentiels :

- ↘ Consommation énergie
- ↘ Émissions CO2

# Problématique 1 :

Prédire

- Consommation d'énergie
- Emissions de CO2

# *Une ligne = Un bien immobilier

## Colonnes

### STRUCTURE

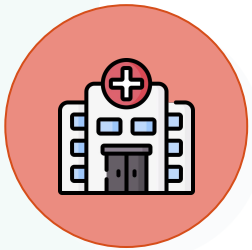
- ✓ Surface totale
- ✓ Nb étages
- ✓ Nb bâtiments
- ✓ Année construction

...

### LOCALISATION

- ✓ Quartier
- ✓ Adresse
- ✓ District

...

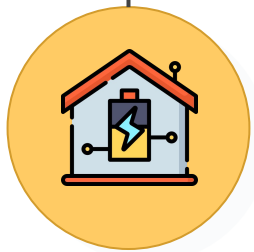


### FONCTIONS


### EMISSIONS

- ✓ Totale
- ✓ Par sq.ft.

...



### CONSO

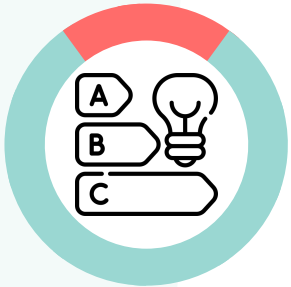
- ✓ Mix énergétique
- ✓ Primaire
- ✓ Normalisée par météo

...

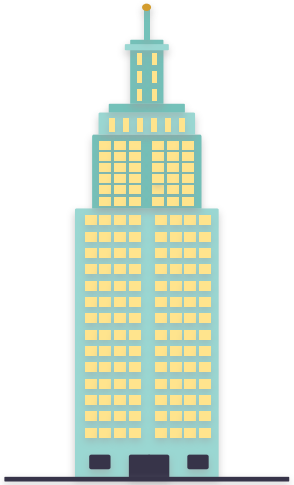
# Performance énergétique

100

1



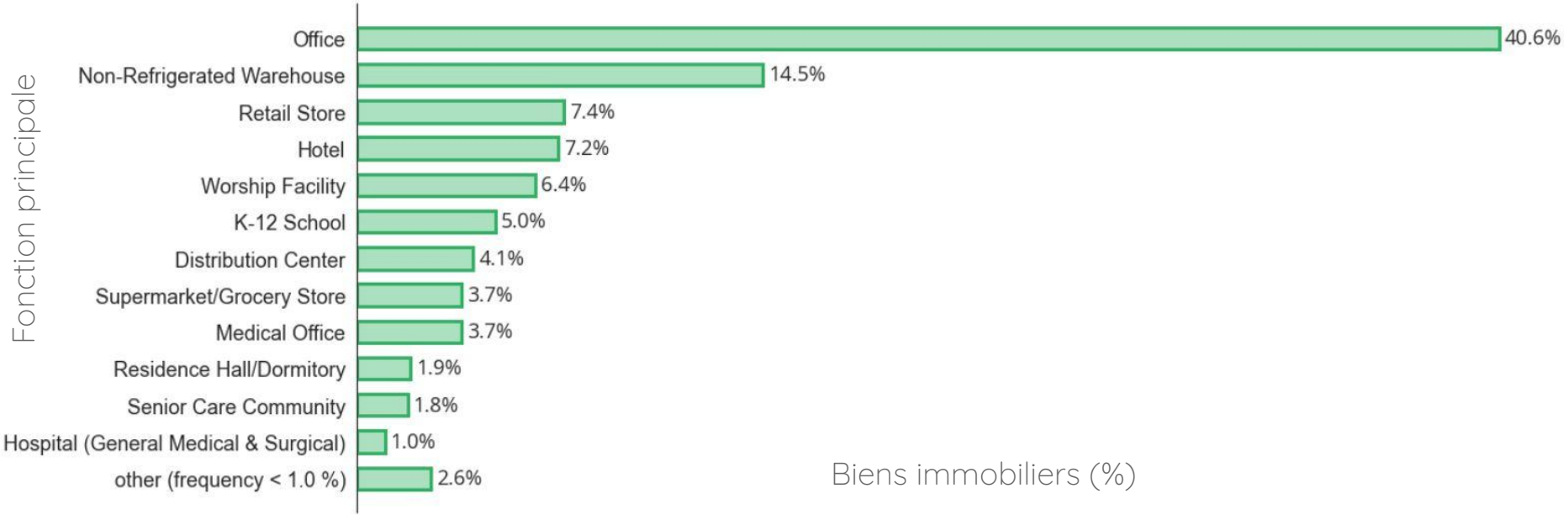

Caractéristiques du bâtiment



# Problématique 2 :

- Quel est l'intérêt de l'energy star score pour la prédiction ?

# Biens immobiliers par fonction principale



# Plan Soutenance

- Feature Engineering
- Modélisation & Optimisation
- Conclusion

# Feature Engineering

*Photo de Nitish Meena sur Unsplash*

# Variables Qualitatives

## Ordinales

- Taille du bien immobilier

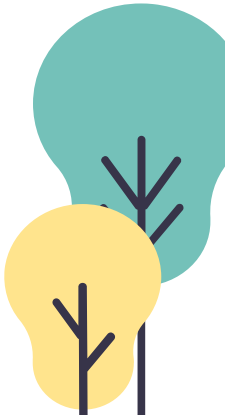
**Ordinal Encoder*

Small -> 0
Mid -> 1
Large -> 2

## Nominales

- Quartier
- Fonction principale

**Target Encoder*



# Target Encoder

$$\text{poids} \times \bar{y}_{\text{modalité}} + (1-\text{poids}) \times \bar{y}_{\text{générale}}$$

- Fuite données test vers train -> pipeline
- Surapprentissage -> hyperparamètres du poids

$$\text{*Poids} = \frac{1}{1 + \exp[(-n + \text{min\_leaf\_sample})/\text{smoothing})]}$$

# Variables Quantitatives

## Discrètes

- Age du bâtiment

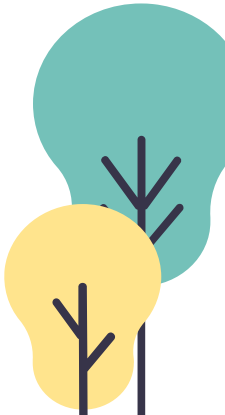
**Robust Scaler*

$$X\_scaled = \frac{X - médiane}{Q3 - Q1}$$

## Continues

- Surfaces par fonction
- Proportions sources d'énergie

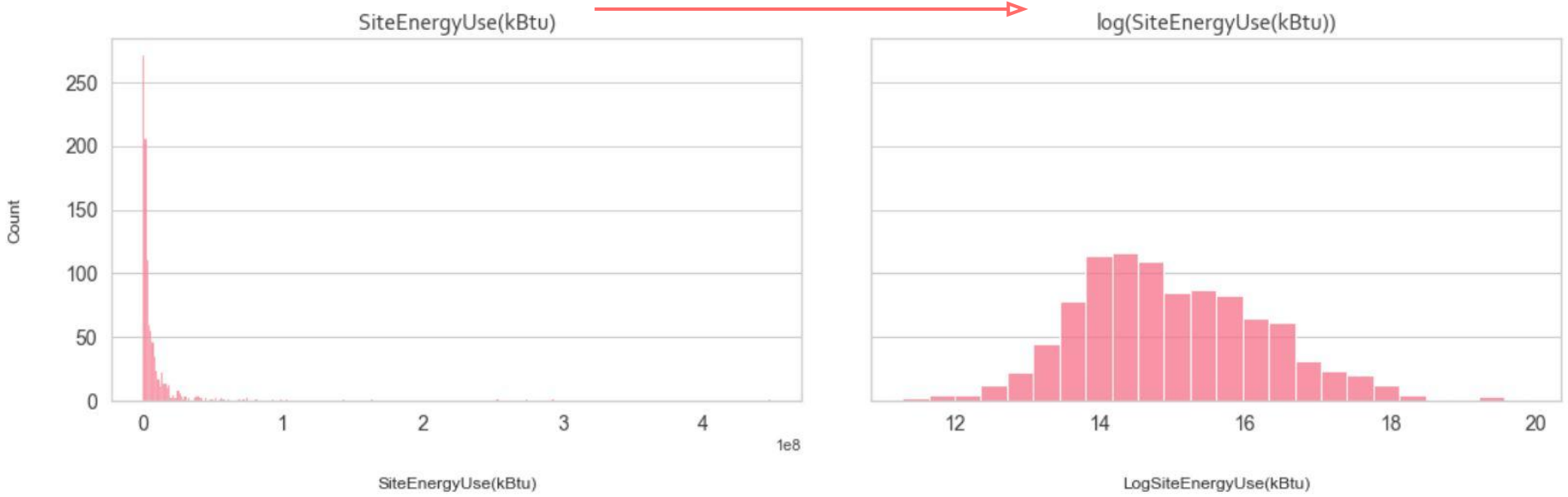
**Passage au log*



# Asymétrie positive

## Cible 1 : Consommation

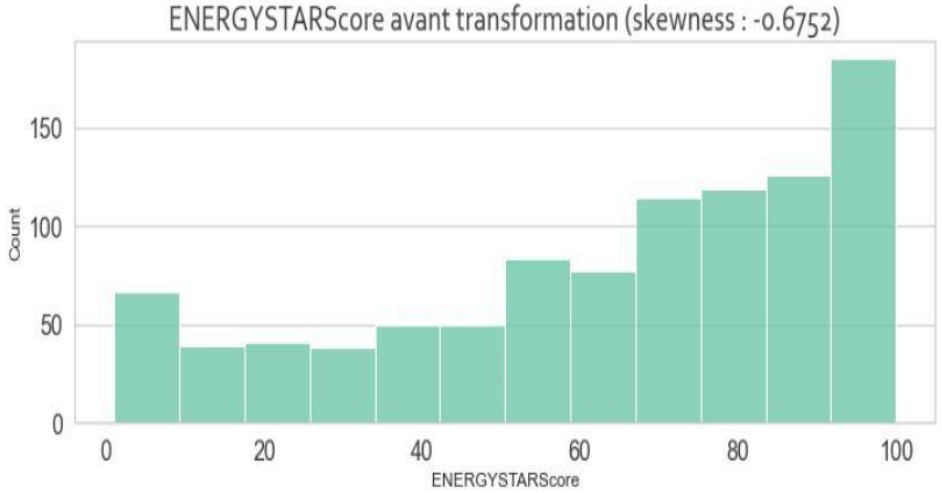
*Passage au log*

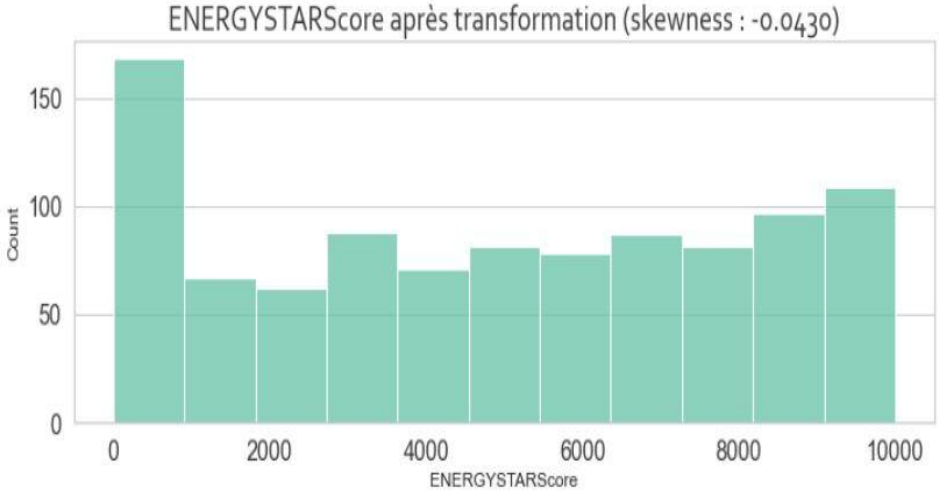


# Asymétrie négative

## Energy star score

*Passage au carré*



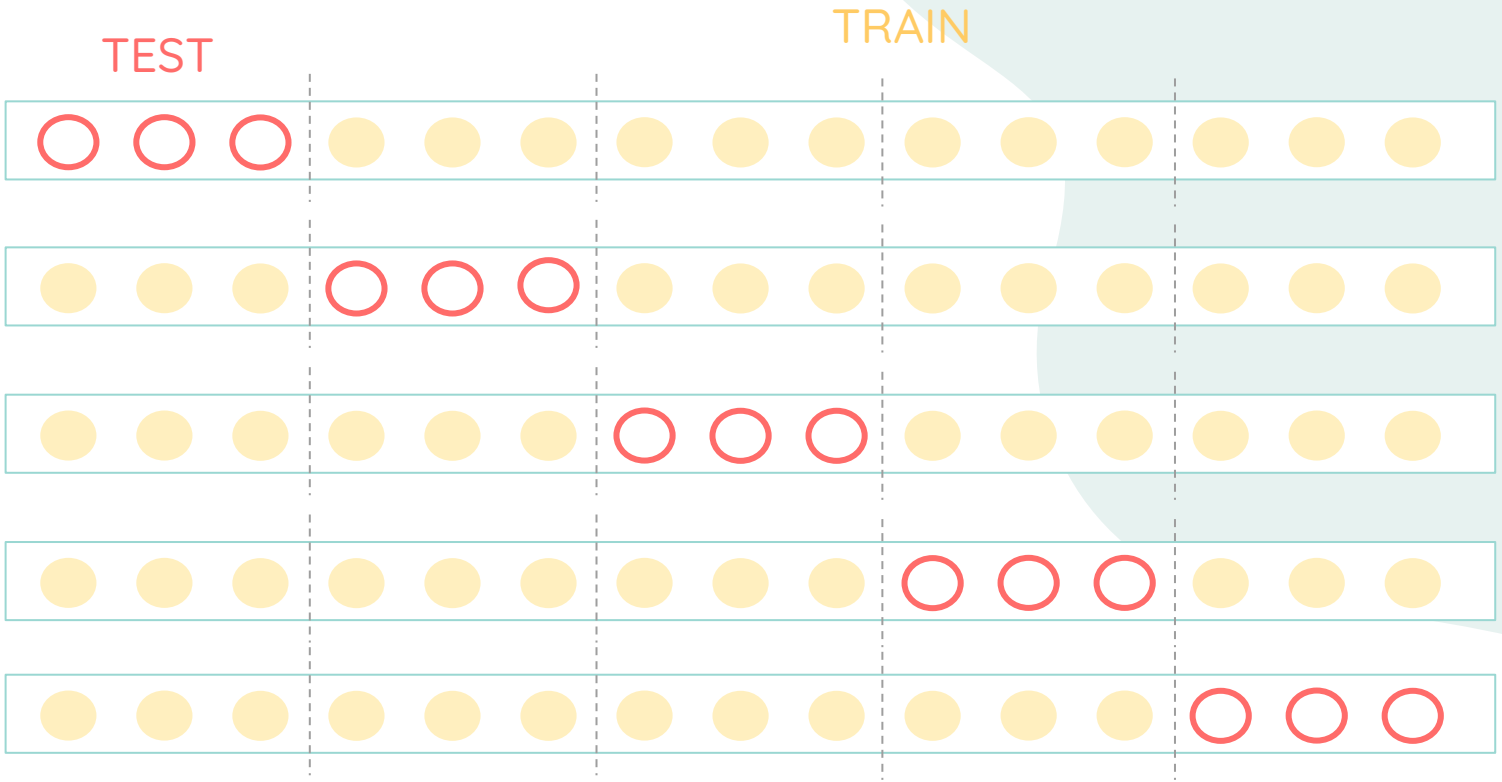
# Modélisation & Optimisation

*Photo de Nitish Meena sur Unsplash*

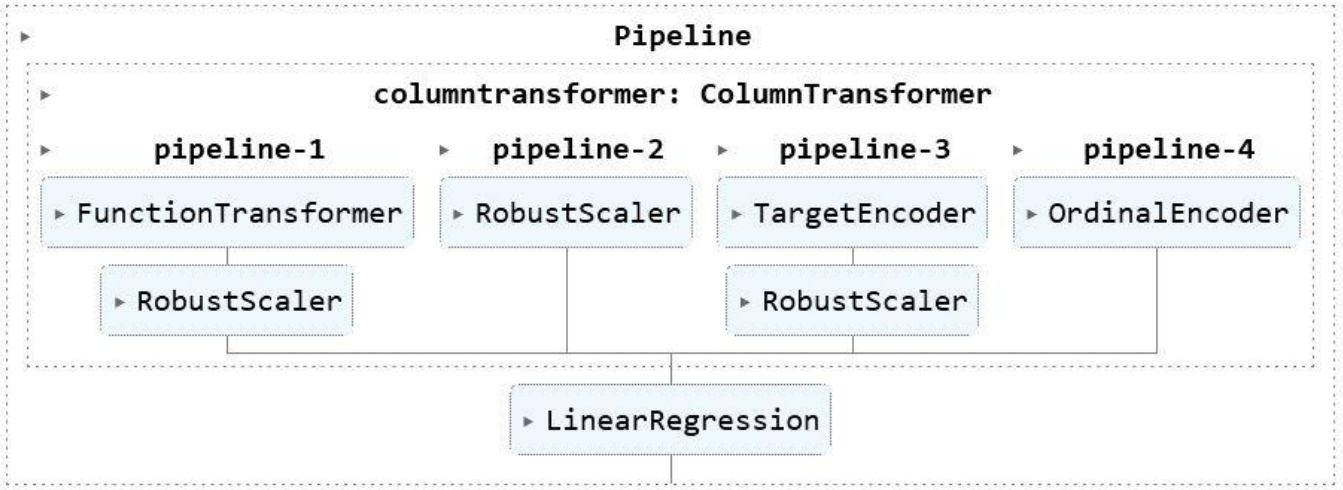
# Modélisation

- Tester sans energy star score
- Optimiser les meilleurs
- Recommencer
  avec energy star score

# Validation croisée



# Pipeline




Entraîne avec
TRAIN SET

Modifie
TEST SET

Robust Scaler :

$$X\_test\_scaled = \frac{X\_test - médiane(X\_train)}{Q3(X\_train) - Q1(X\_train)}$$

# Mesures de performance

## Distance

- Erreur quadratique moyenne

$$MSE = 1/n \Sigma (prédite - vraie)^2$$

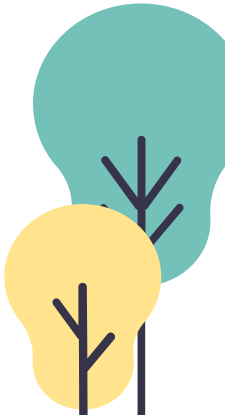
- Erreur absolue moyenne

$$MAE = 1/n \Sigma |prédite - vraie|$$

## Corrélation

- Coefficient de détermination

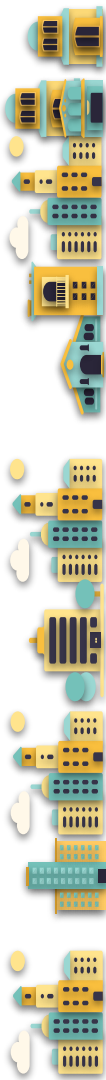
$$R2 = 1 - \underbrace{\frac{\Sigma(vraie\text{-}prédite)^2}{\Sigma(vraie\text{-}moyenne)^2}}_{RSE}$$



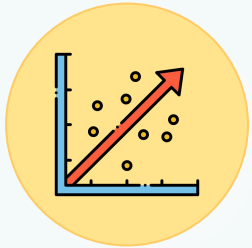
# Approche naïve

*Médiane du jeu d'entraînement

| R2 | RMSE | MAE | Temps d'entraînement |
|---|---|---|---|
| -0.07 | 23 millions | 7 millions | 0.2 secondes |



# Modèles supervisés testés

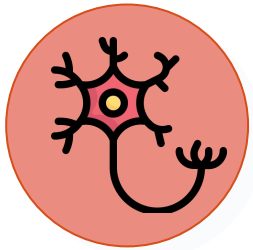


## LINEAIRES

- ✓ Régression linéaire
- ✓ Régression ridge
- ✓ Lasso
- ✓ Elastic Net
- ✓ SVM

## A NOYAU

- ✓ Régression Ridge à noyau
- ✓ SVM à noyau

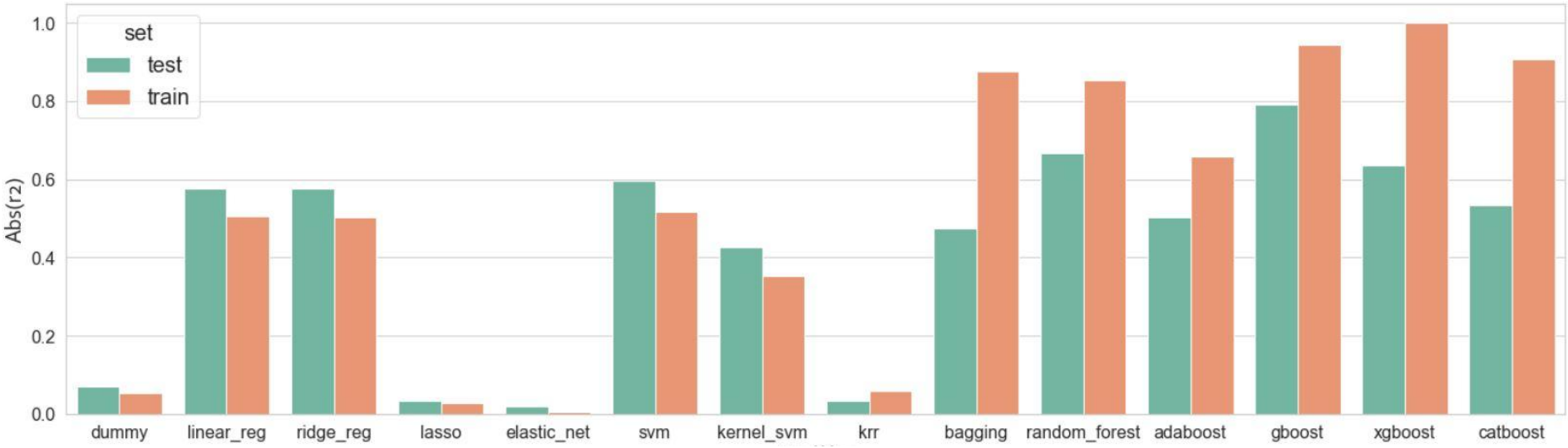


## PERCEPTRON MULTI-COUCHES

## ENSEMBLISTES

- ✓ Bagging
- ✓ Forêt aléatoire
- ✓ Adaboost
- ✓ Gradient Boost
- ✓ XGBoost
- ✓ CatBoost

# R2 moyen après validation croisée (valeur absolue)



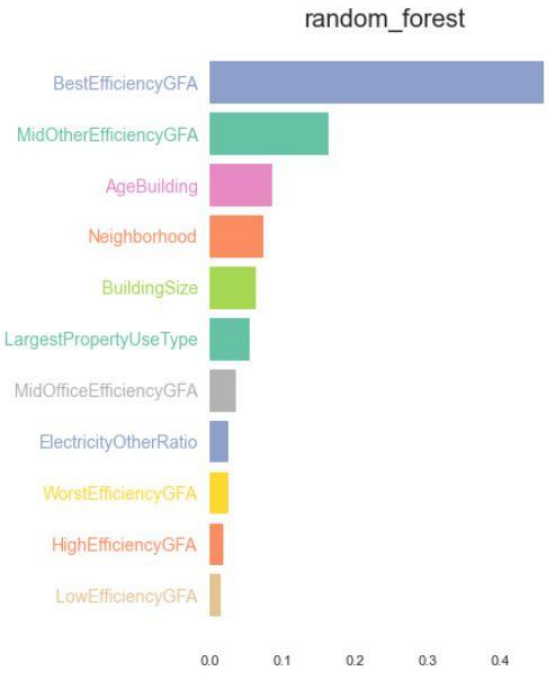

Modèles (consommation)

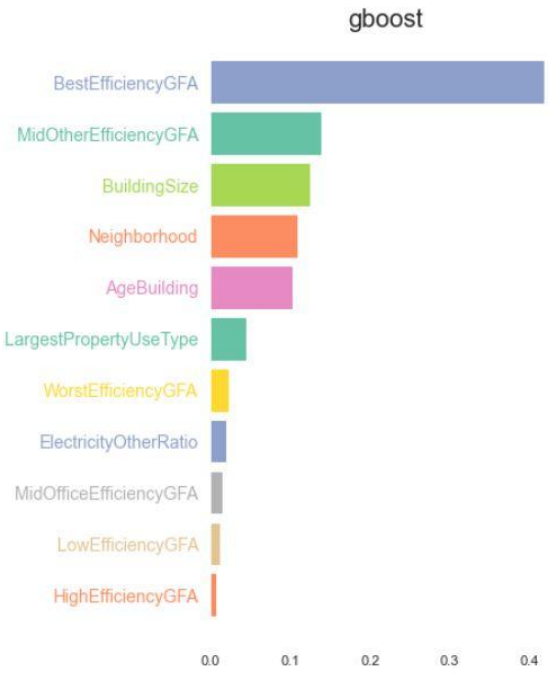
# Top 3

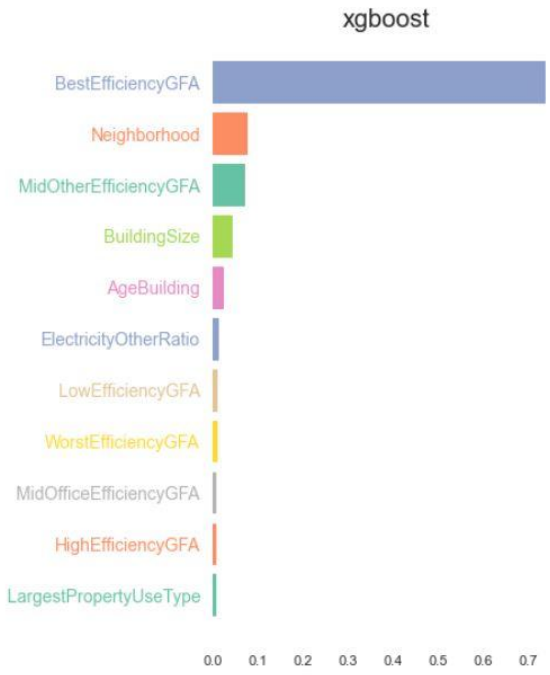
| Modèle | R2 | RMSE (millions) | MAE (millions) | Temps d'entraînement (secondes) |
|---|---|---|---|---|
| GBoost | 0.79* | 10,5 | 3,6 | 0.48 |
| Forêt Aléatoire | 0.67* | 13,8 | 3,9 | 1.26 |
| XGBoost | 0.64* | 13 | 4 | 0.37 |
| Dummy | -0.07 | 23 | 7 | 0.2 |

*écart-type R2 : 0.2

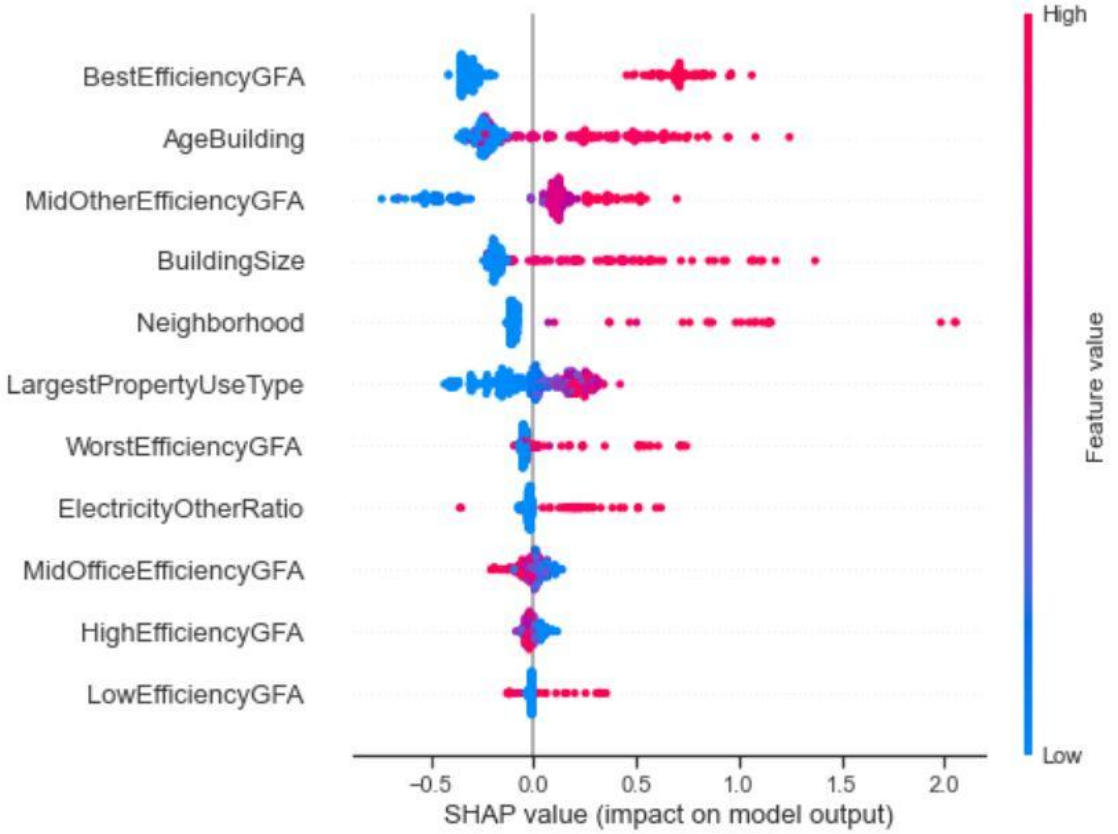
# Importance des Features (consommation)



# Summary plot (GBoost)




# Optimisation

*Photo de* Meriç Dağlı *sur Unsplash*

# Algorithmes d'optimisation

## Random Search

n combinaisons

au hasard

Baseline

## Grid Search

toutes combinaisons possibles

exhaustif

Affinage

## Optimisation Bayesienne

n combinaisons

Utilise les itérations précédentes pour déterminer les prochaines combinaisons à tester

Proba conditionnelles

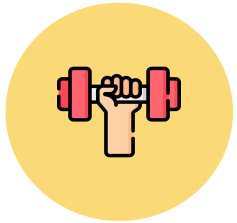
Sklearn

HyperOpt

# Optimisation


70%

30%



TRAIN

TEST

Validation croisée : n itérations

80%

20%

Train
80% * 70% = 56% données

Validation : -R2
20% * 70% = 14% données

minimise score moyen

Recherche hyperparamètres

Mesure performance : R2

Avec meilleurs hyperparamètres

# Top 3 après optimisation

| Modèle | R2 (TEST) | R2 (TRAIN) | Ecart type R2 Test | Temps d'entraînement (secondes) |
|---|---|---|---|---|
| GBoost | 0.89 | 0.90 | 0.19 | 0.13 |
| Forêt Aléatoire | 0.81 | 0.82 | 0.23 | 9 |
| XGBoost | 0.88 | 0.95 | 0.10 | 0.8 |
| Dummy | -0.07 | -0.05 | 0.08 | 0.2 |

# Energy Star Score

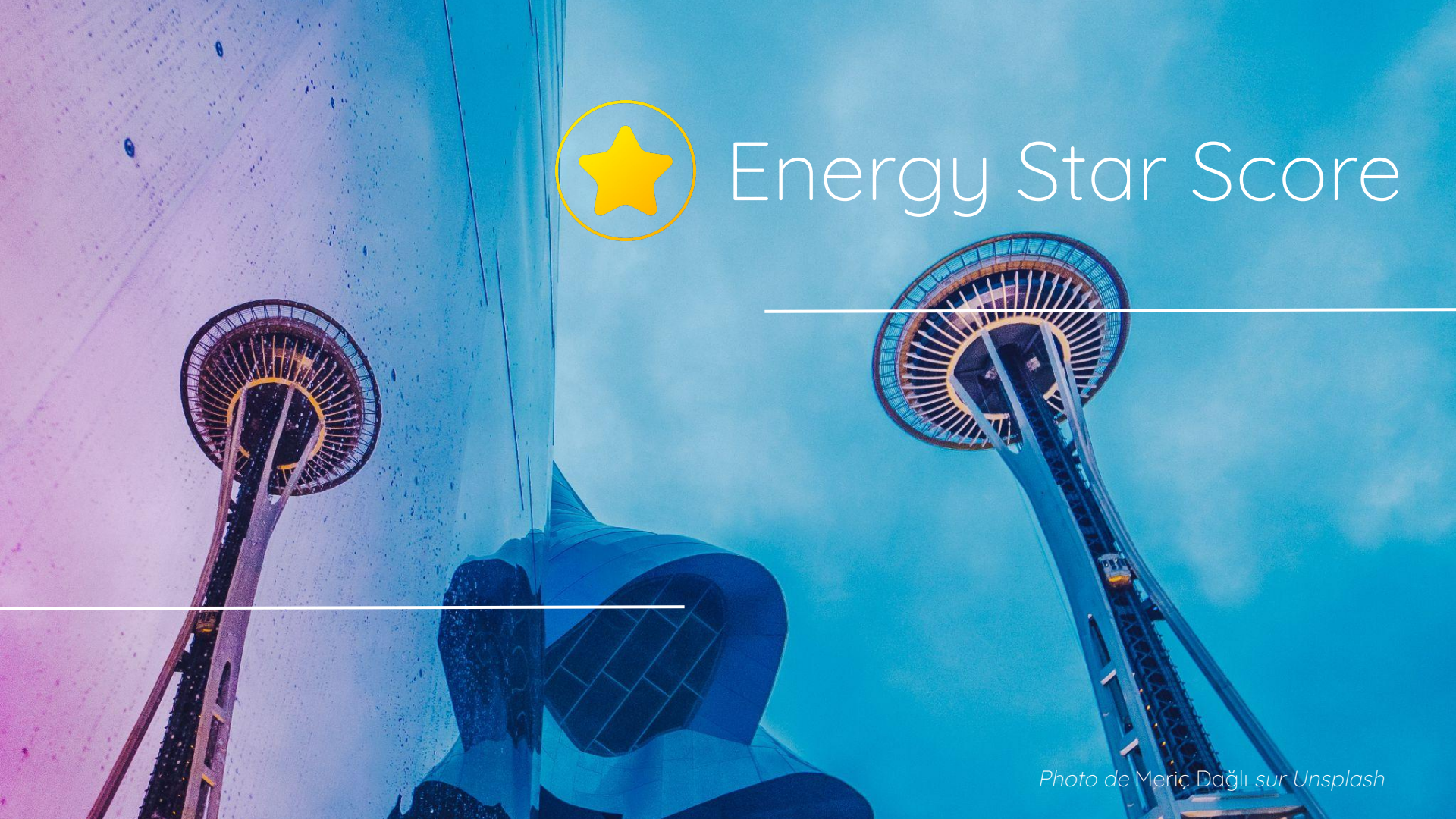
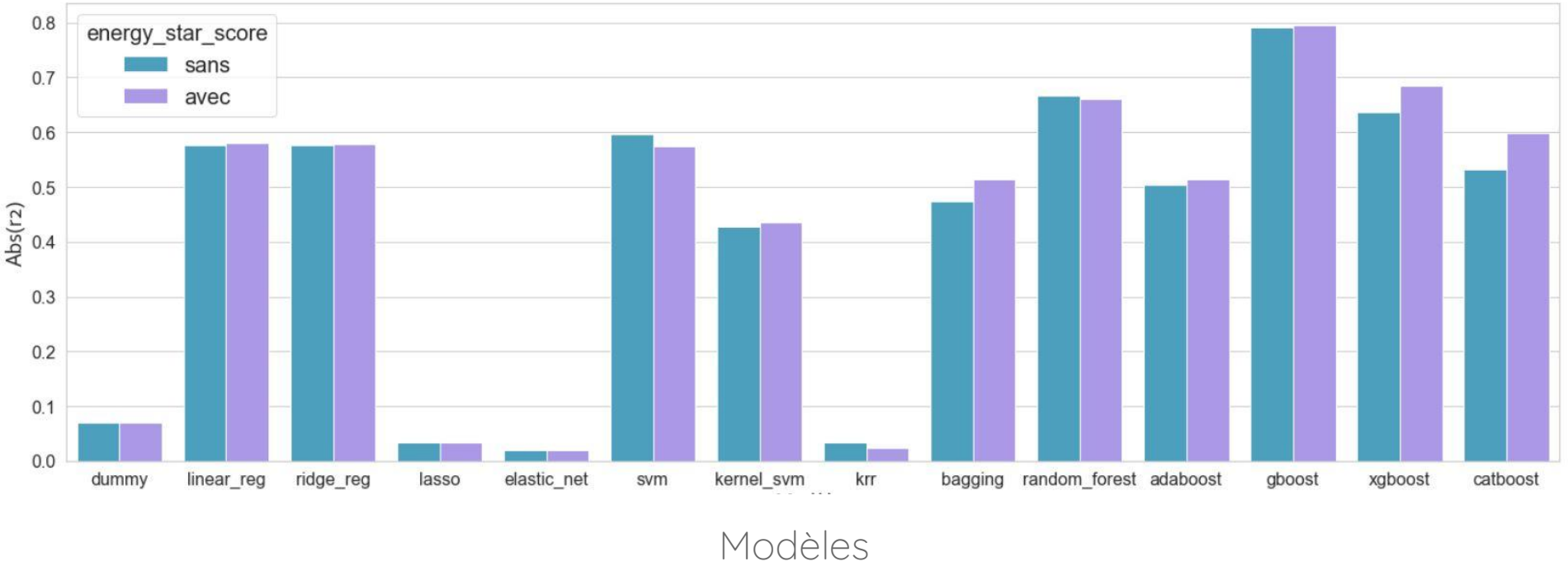


*Photo de* Meriç Dağlı *sur Unsplash*

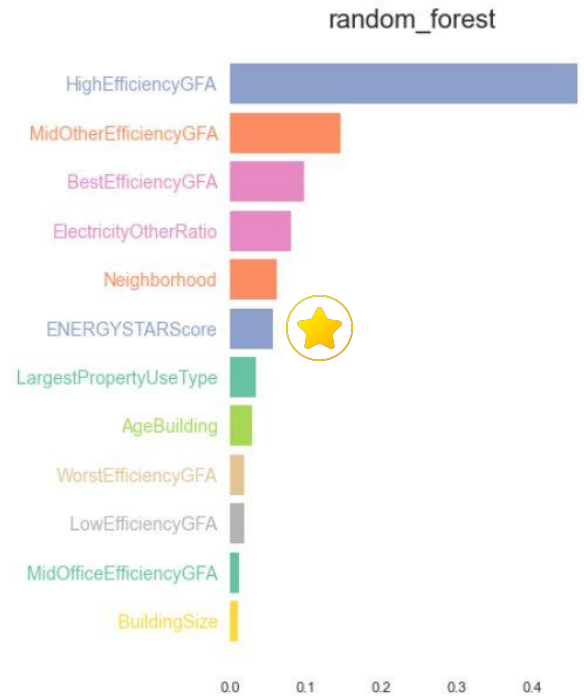
# Cible : Consommation

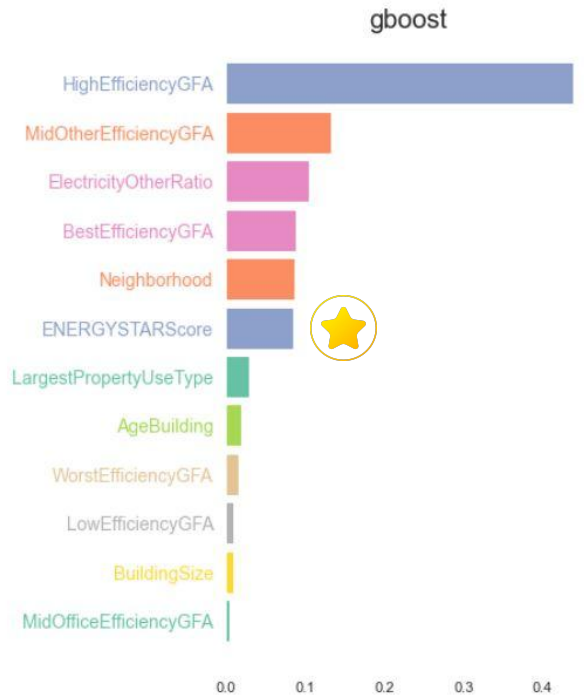
## R2 moyen après validation croisée (valeur absolue) - TEST -

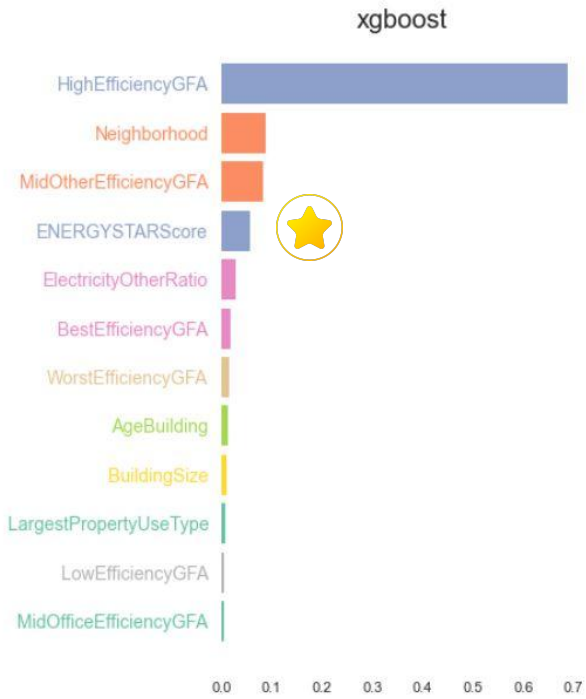


# Cible : Consommation

## Importance des Features



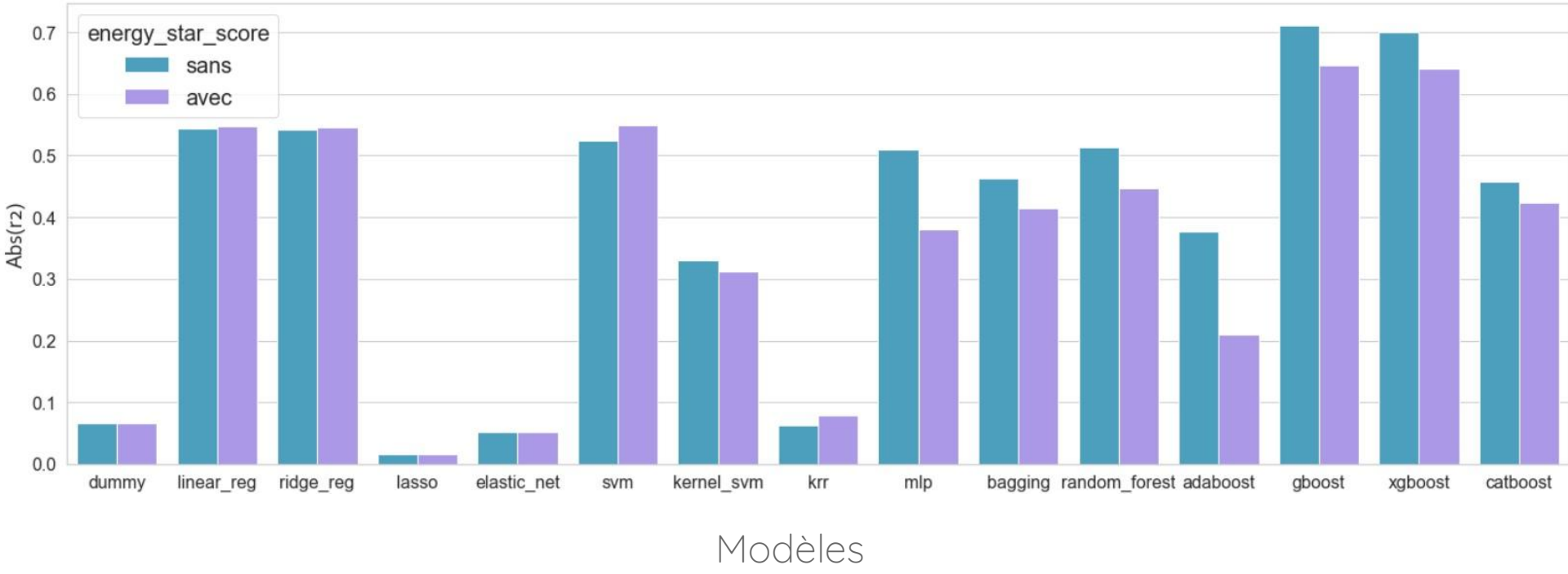



# Top optimisé (cible conso)

| Modèle | R2 avec energy star (TEST) | R2 sans (TEST) |
|---|---|---|
| GBoost | 0.91 (+/- 0.19) | 0.89 (+/-0.19) |
| XGBoost | 0.93 (+/- 0.06) | 0.88 (+/-0.10) |
| Dummy | -0.07 | -0.05 |

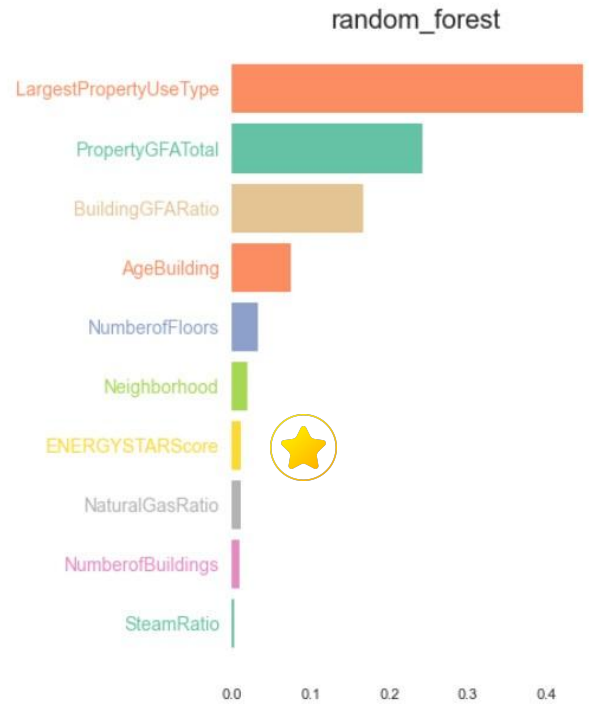
# Cible : Emissions

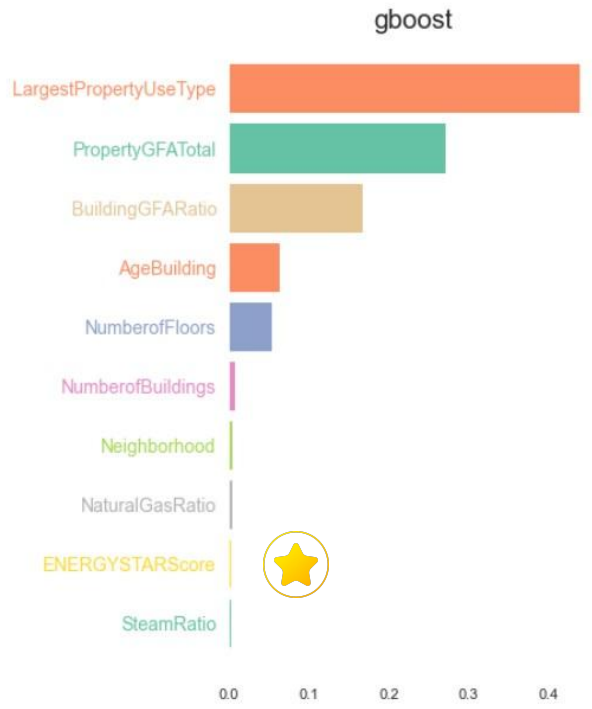
## R2 moyen après validation croisée (valeur absolue) - TEST -

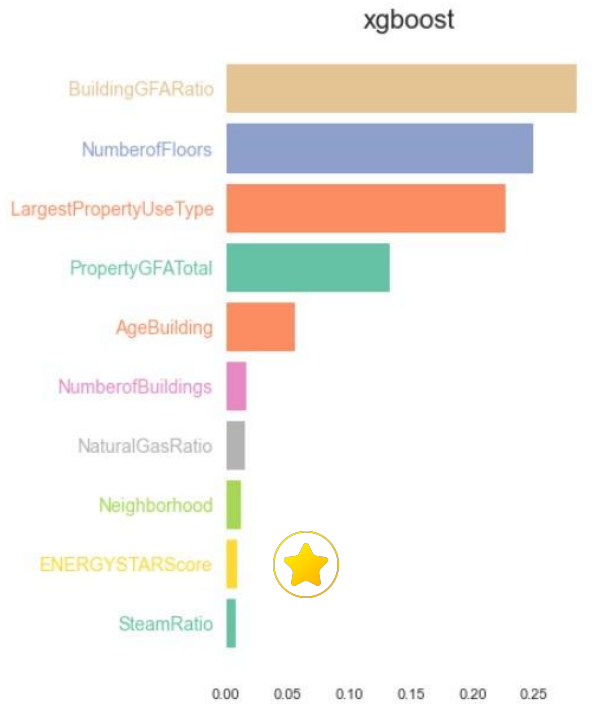


# Cible : Emissions

## Importance des Features




# Top optimisé (cible émissions)

| Modèle | R2 avec energy star (TEST) | R2 sans (TEST) |
|---|---|---|
| GBoost | 0.79 (+/- 0.3) | 0.86 (+/-0.3) |
| XGBoost | 0.86 (+/- 0.3) | 0.91 (+/-0.3) |
| Dummy | -0.07 | -0.05 |

# Conclusion

Qualité du modèle :

- XGBoost
- Construit avec peu de données
- Peu robuste
- Peu stable

Utilité Energy Star Score :

- Pas impact significatif
- Perte observations/info

# CREDITS

This presentation template was created by **Slidego**, including icons by **Flaticon**, and infographics & images by **Freepik**.

**Please keep this slide for attribution.**