# Anticipez les Besoins en Consommation de Bâtiments

OPENCLASSROOMS

Seattle

# ENJEUX

**Seattle : 2050 → Neutralité carbone**

**Prédire les émissions de CO2 & la consommation totale d'énergie de bâtiments non destinés à l'habitation**

# SOMMAIRE

# EXPLORATION DES DONNÉES & 1ER FEATURE ENGINEERING

## BUILDING ENERGY BENCHMARKING

- **3376 propriétés × 46 features**

→ 1668 non-résidentielles

- **Compliance Status =** Compliant
- **DefaultData** = False
- **Outlier** = Null
- **PropertyGFATotal =** PropertyGFAParking + PropertyGFABuilding(s)

## FEATURE ENGINEERING:

- **Variables Cibles =**
  - **TotalGHGEmissions**
  - **SiteEnergyUseWN(kBtu)**
- **Création de nouvelles variables :**
  - **Electricity-NaturalGas-SteamUse(%) :** Electricity-NaturalGas-SteamUse(kBtu) /SiteEnergyUse
  - **Age :** DataYear - Yearbuilt

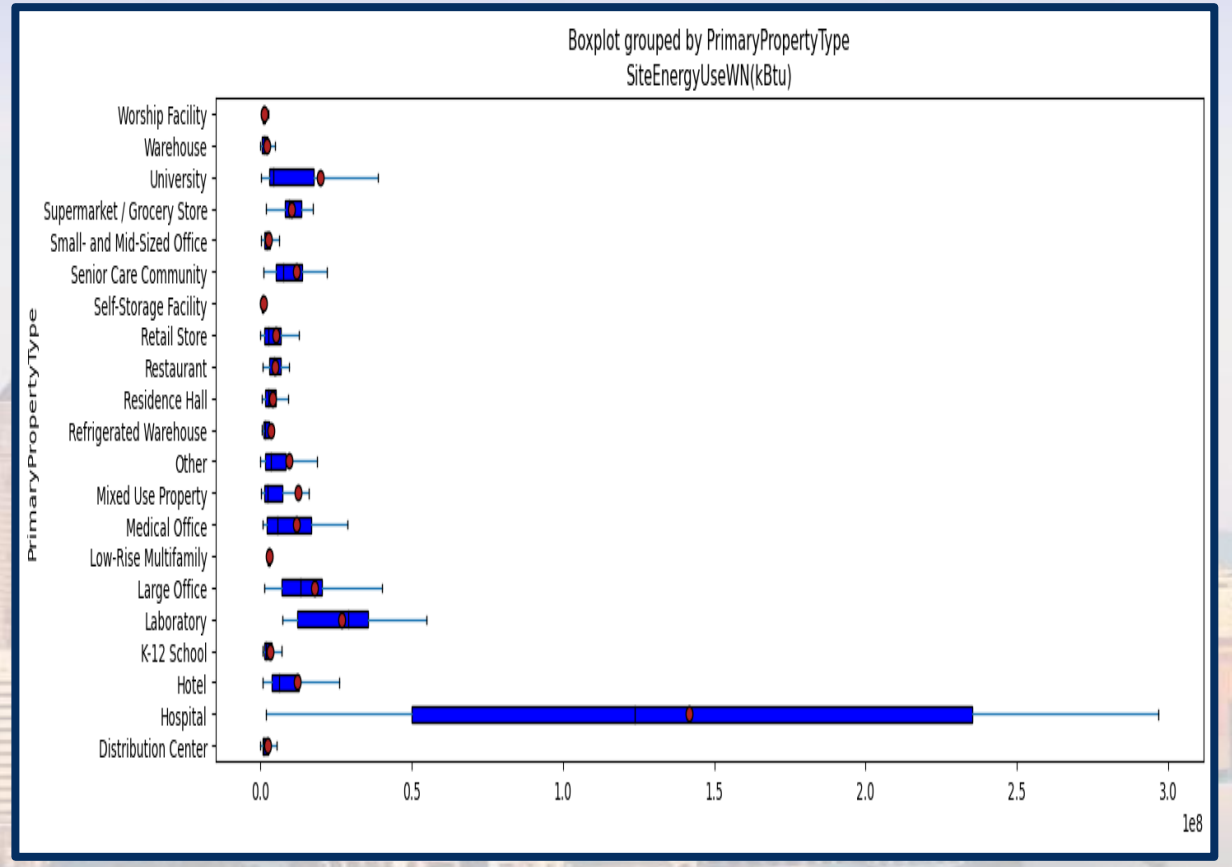
# SÉLECTION DES FEATURES

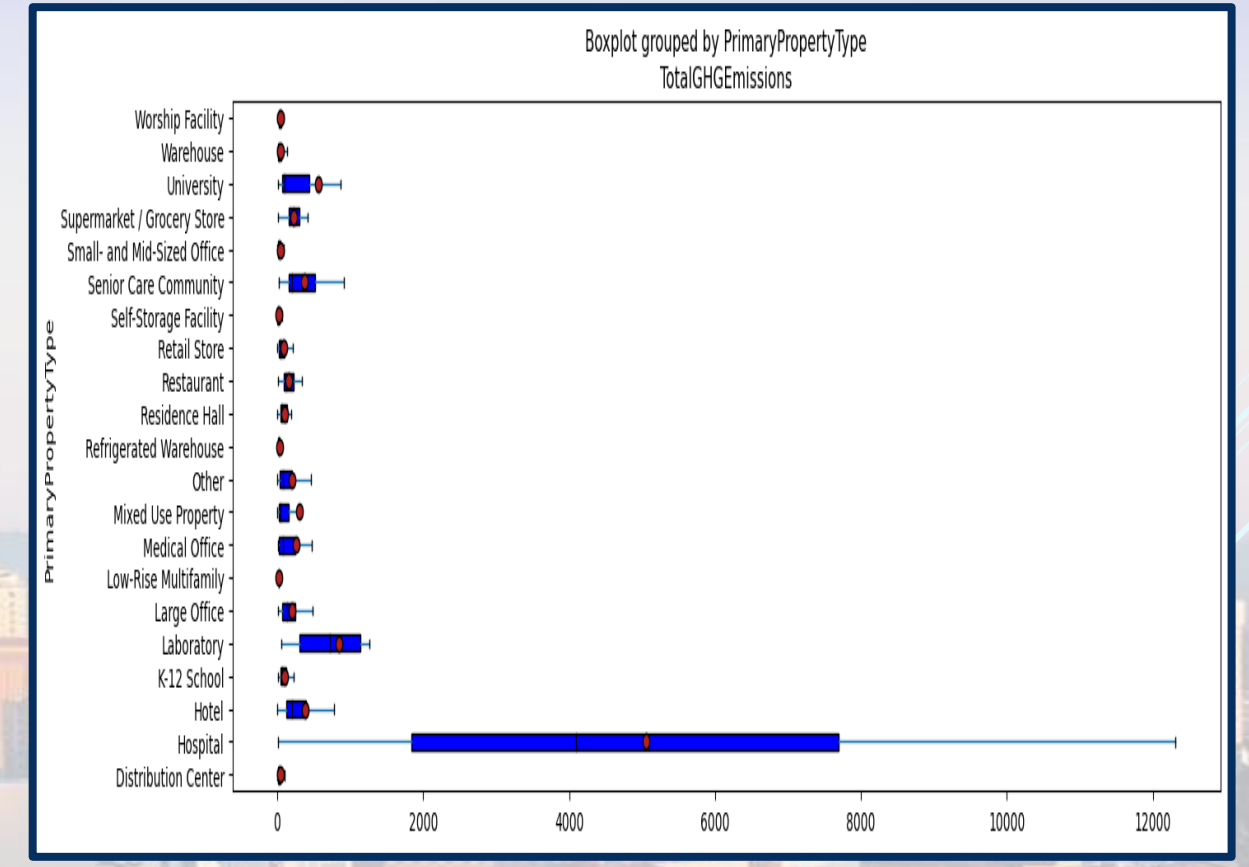
- **Caractéristiques Géographiques :** Latitude & Longitude
- **Caractéristiques Fonctionnelles =** PrimaryPropertyType
- **Caractéristiques Structurelles =** NumberofBuildings, NumberofFloors, PropertyGFABuilding(s), PropertyGFAParking, Age
- **Caractéristiques Energétiques :** Electricity(%), NaturalGas(%), SteamUse(%)

**Data Leakage :**

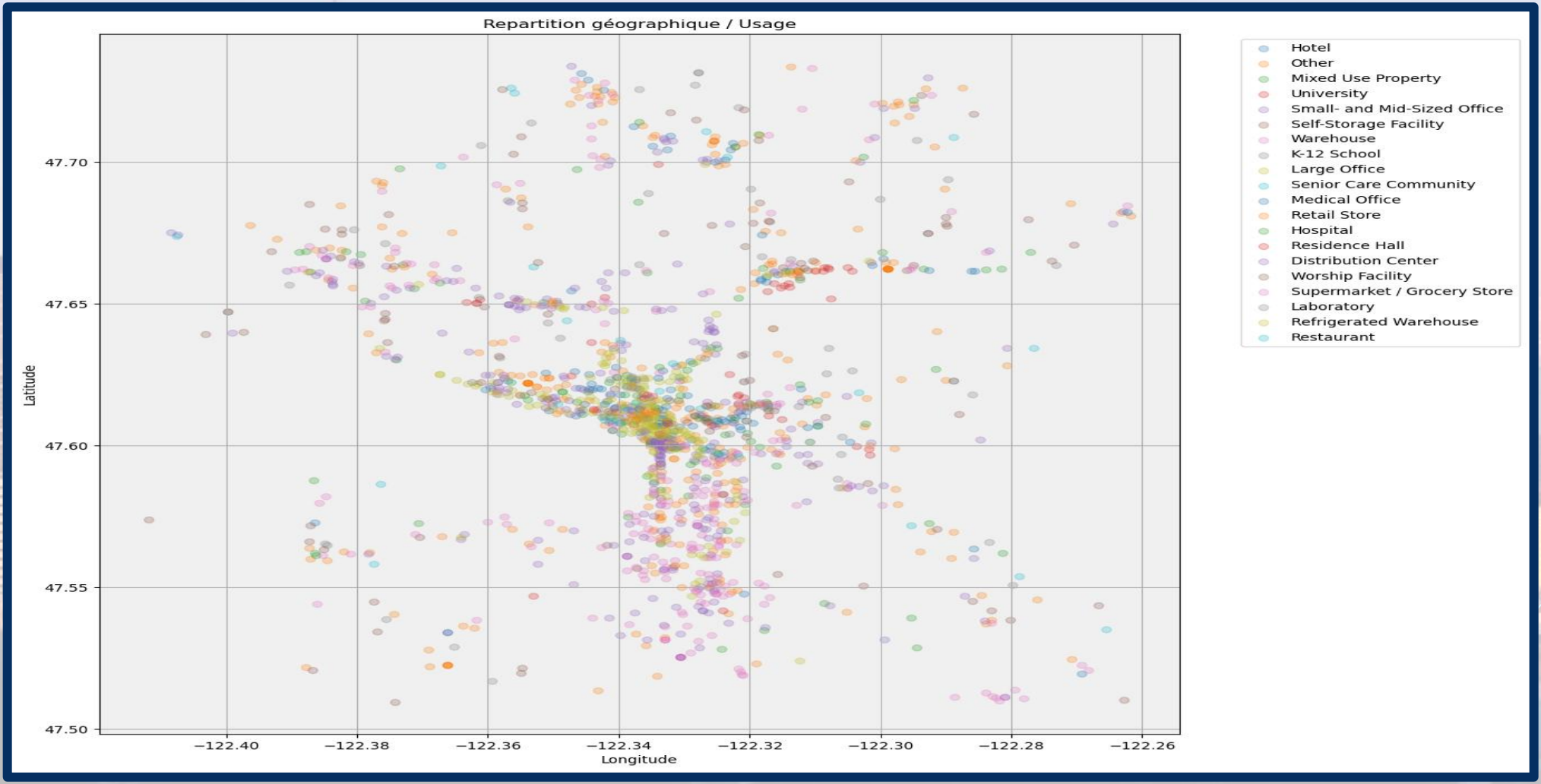
**~~GHGEmissionsIntensity~~, ~~SiteEUI(kBtu/sf)~~, etc.**

# ANALYSE EXPLORATOIRE : PROPERTY TYPE

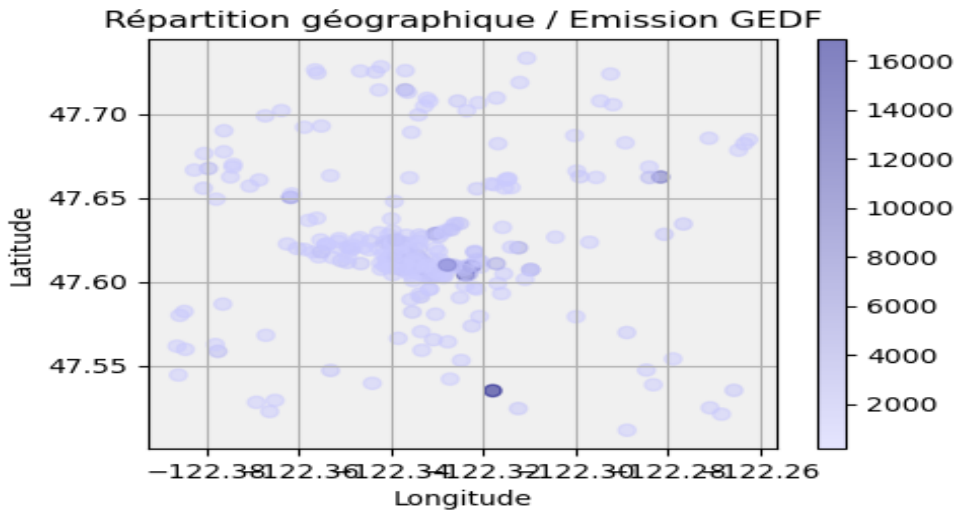


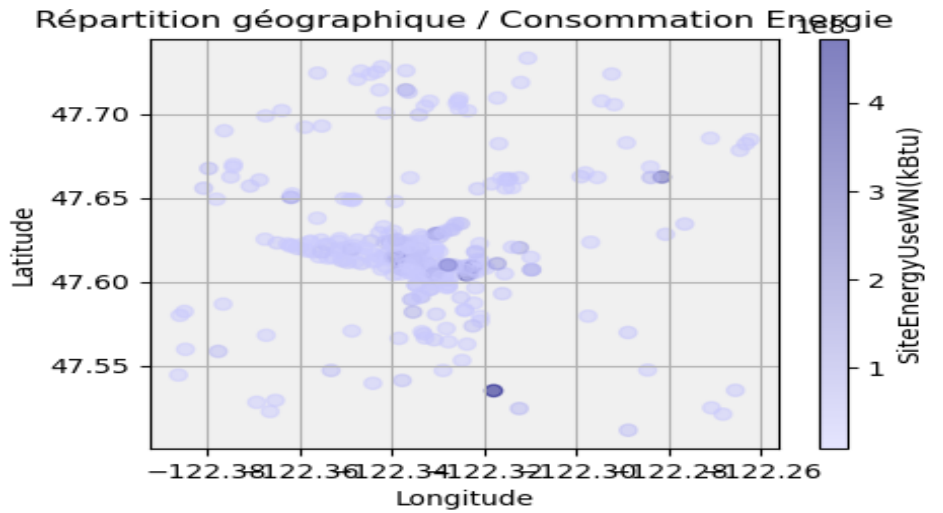



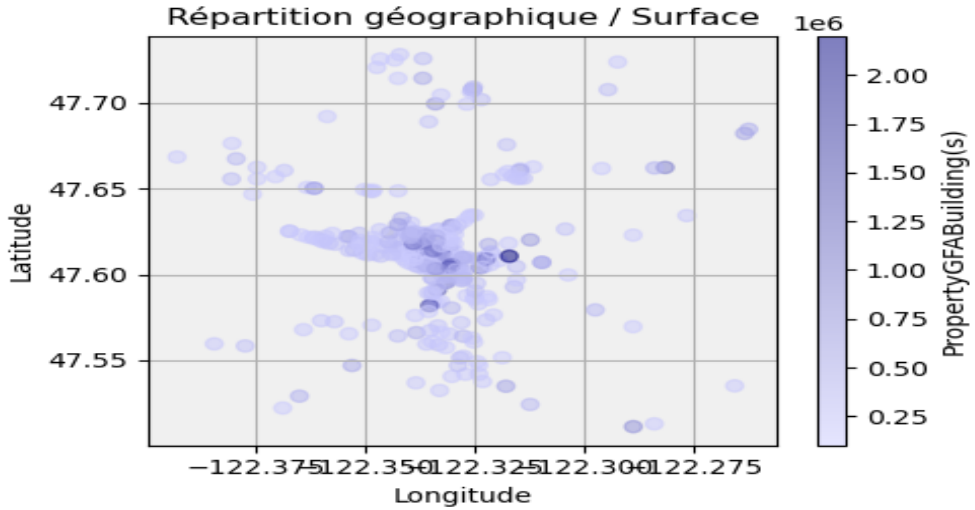
# ANALYSE EXPLORATOIRE : GEOGRAPHIE

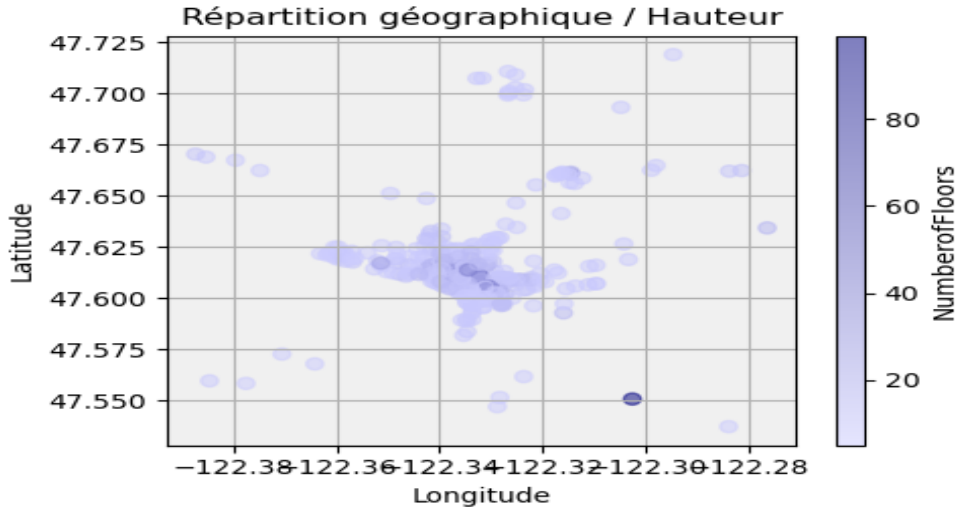


# ANALYSE EXPLORATOIRE : GEOGRAPHIE

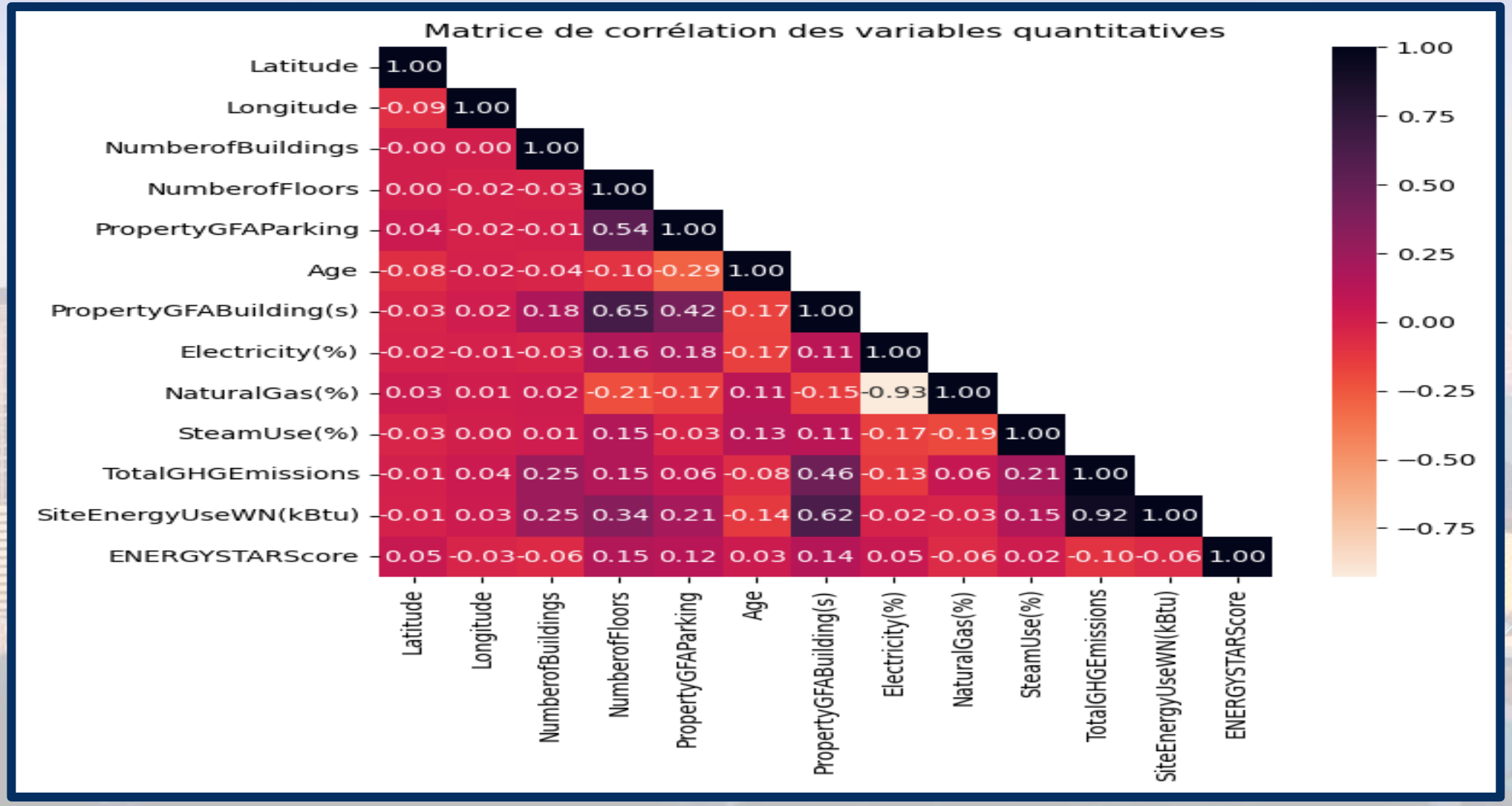


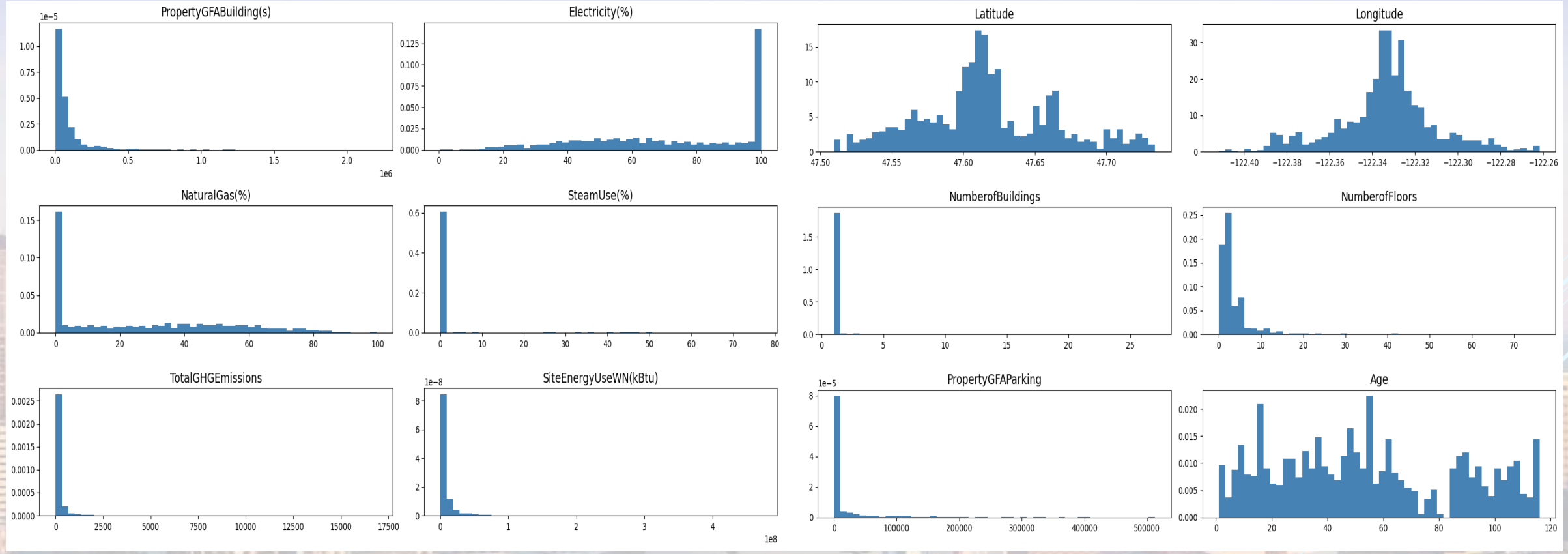



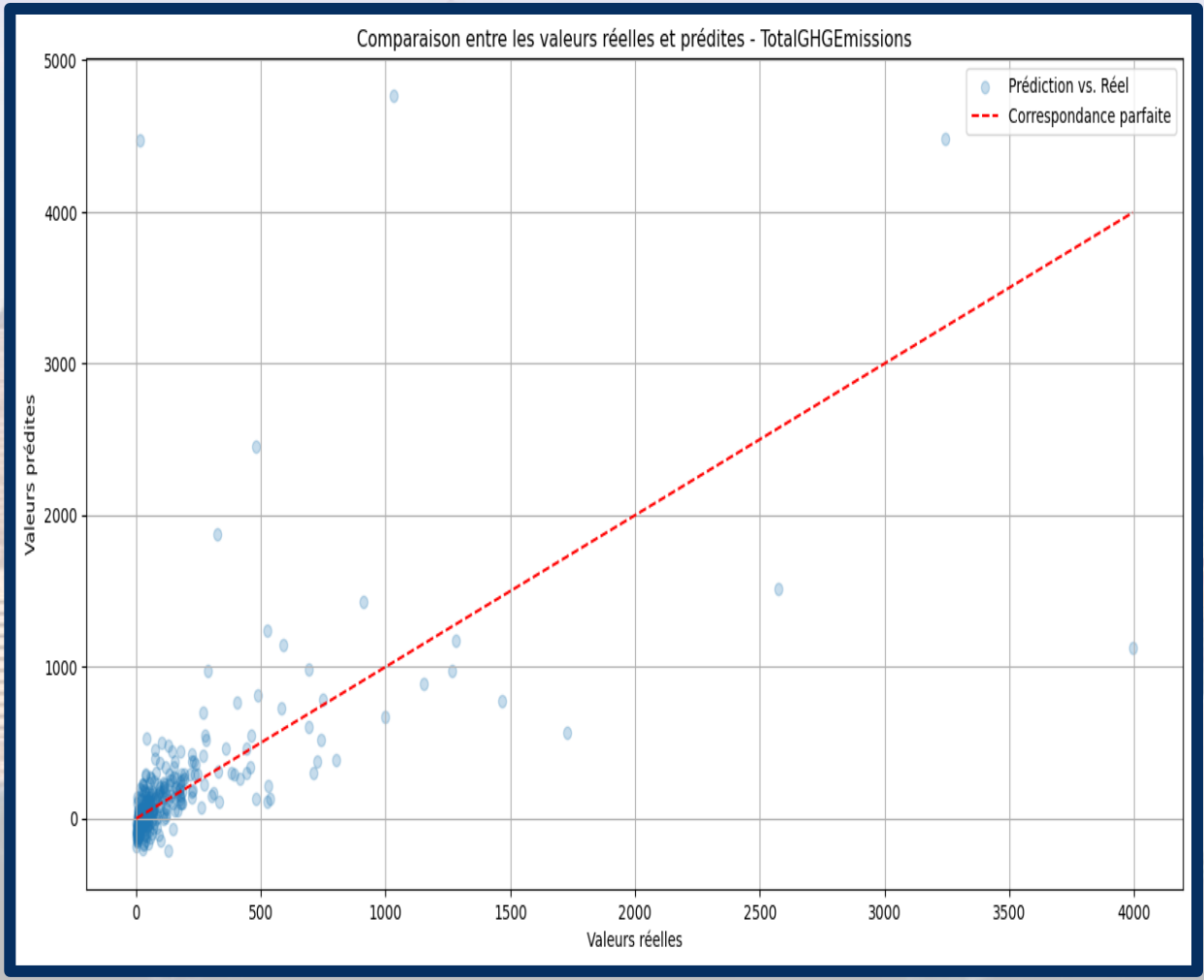
# ANALYSE EXPLORATOIRE : VARIABLES CIBLES



# ANALYSE EXPLORATOIRE : REPARTITION GLOBALE



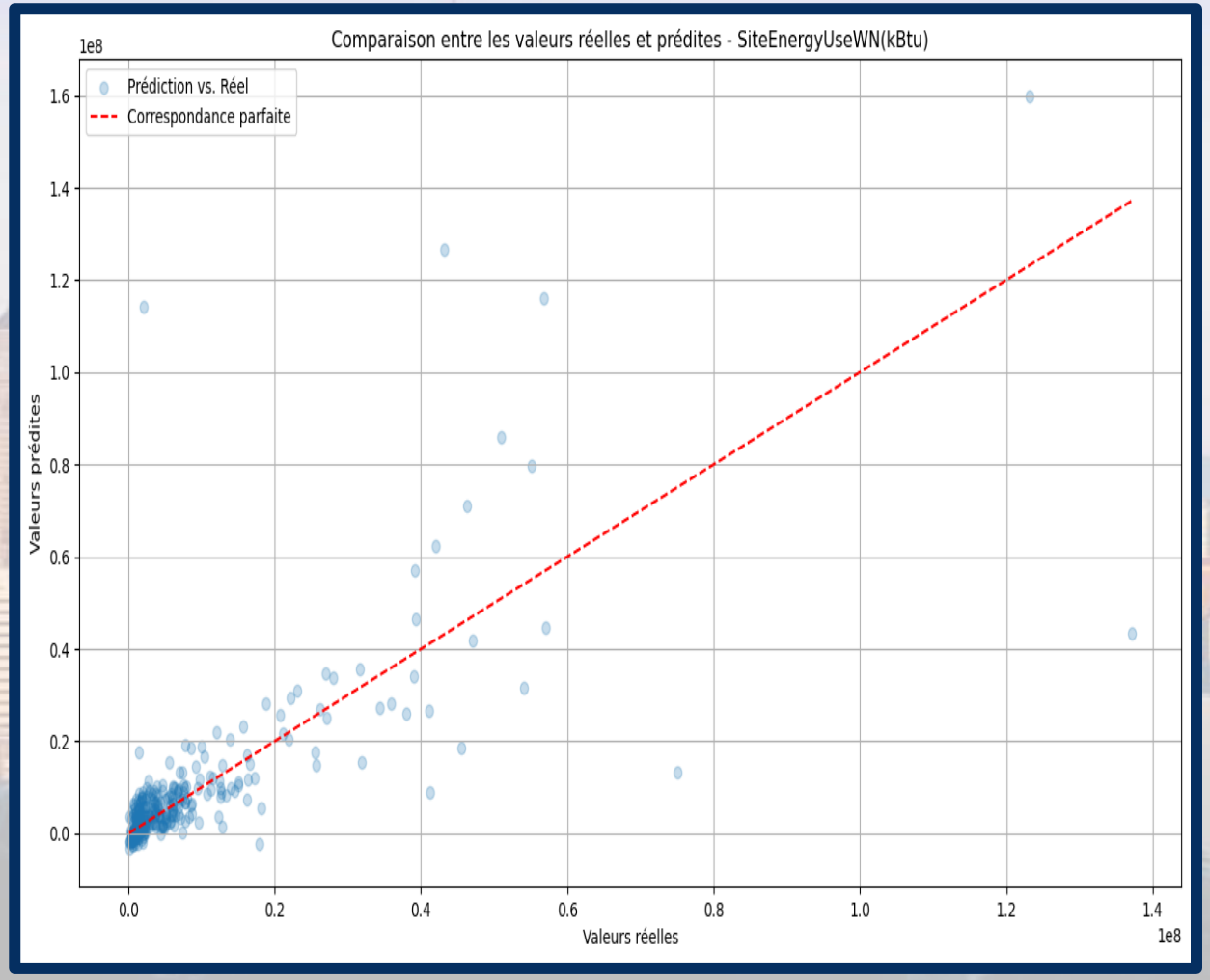
# DÉVELOPPEMENT & SIMULATION D'UN PREMIER MODÈLE : TOTALGHGEMISSIONS




## Modèle de Départ: Régression Linéaire

- TotalGHGEmissions :
  - $R^2$ = -0.28
  - MAE = 170.46
  - RMSE = 442.90

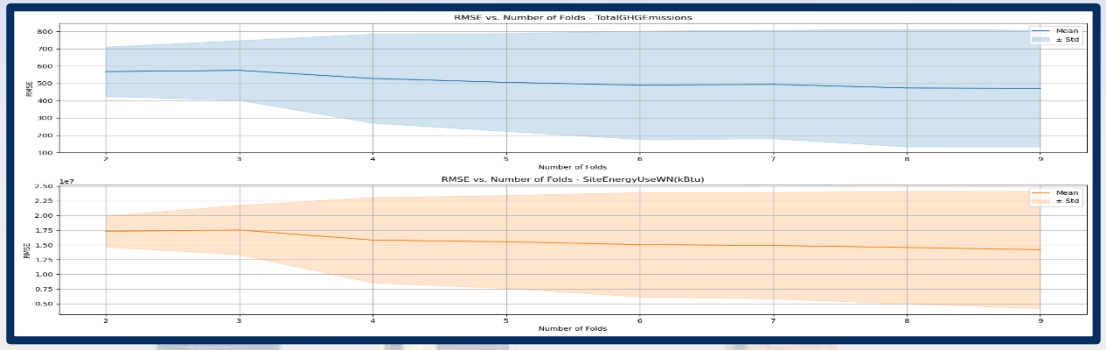
# DÉVELOPPEMENT & SIMULATION D'UN PREMIER MODÈLE : SITEENERGYUSEWN(KBTU)

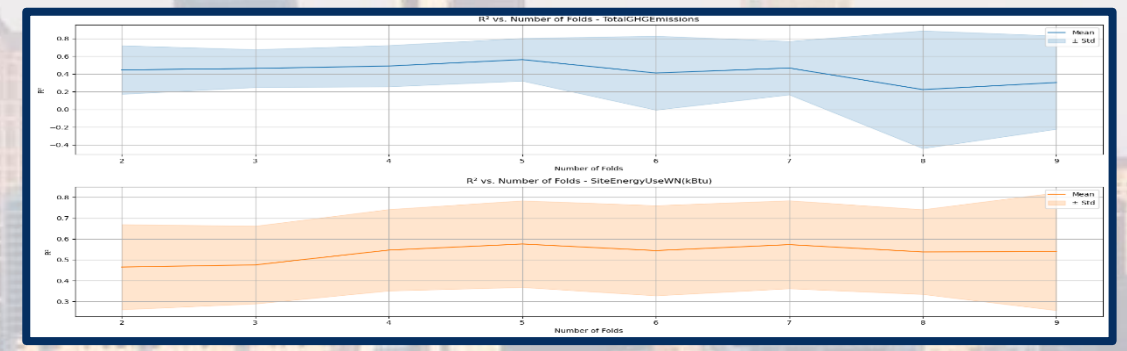


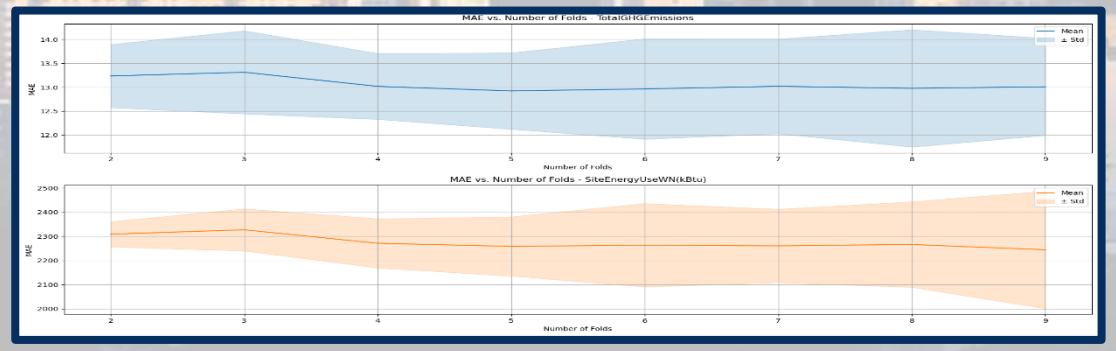

## Modèle de Départ: Régression Linéaire

- SiteEnergyUseWN(kBtu) :
  - $R^2 = 0.34$
  - MAE = 5110950.26
  - RMSE = 12500303.1

# DÉVELOPPEMENT & SIMULATION D'UN PREMIER MODÈLE



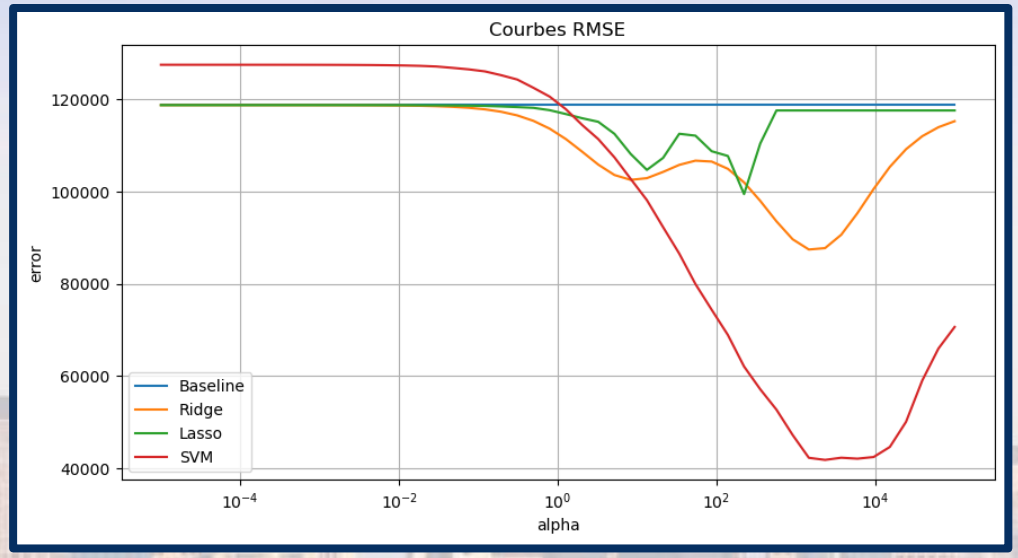



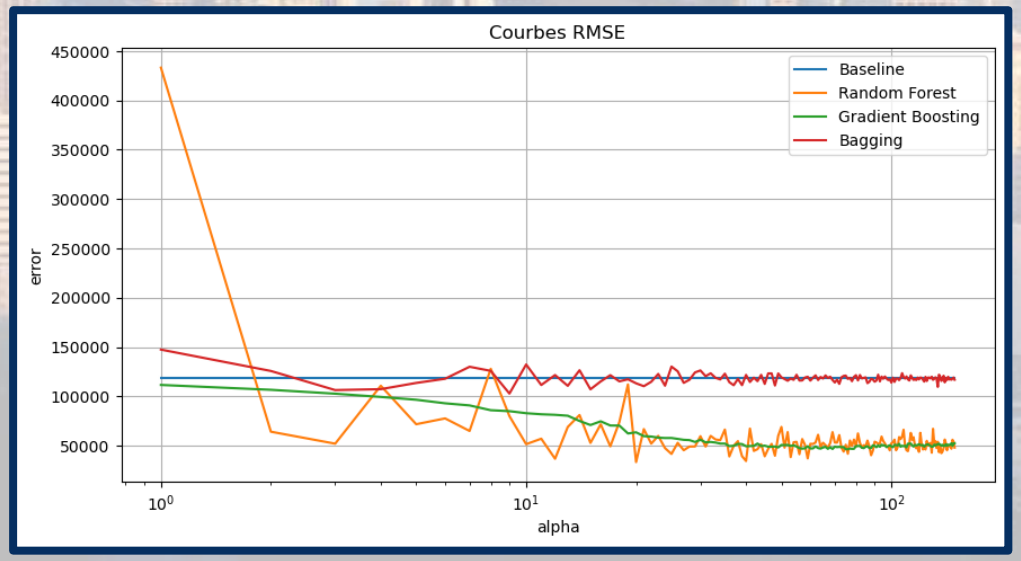




## Méthode de Validation Croisée

- **TotalGHGEmissions :**
  - **$R^2$ = 0.49**
  - **MAE = 13.02**
  - **RMSE = 527.84**
- **SiteEnergyUseWN(kBtu) :**
  - **$R^2$ = 0.55**
  - **MAE = 2271.22**
  - **RMSE = 15823938.70**

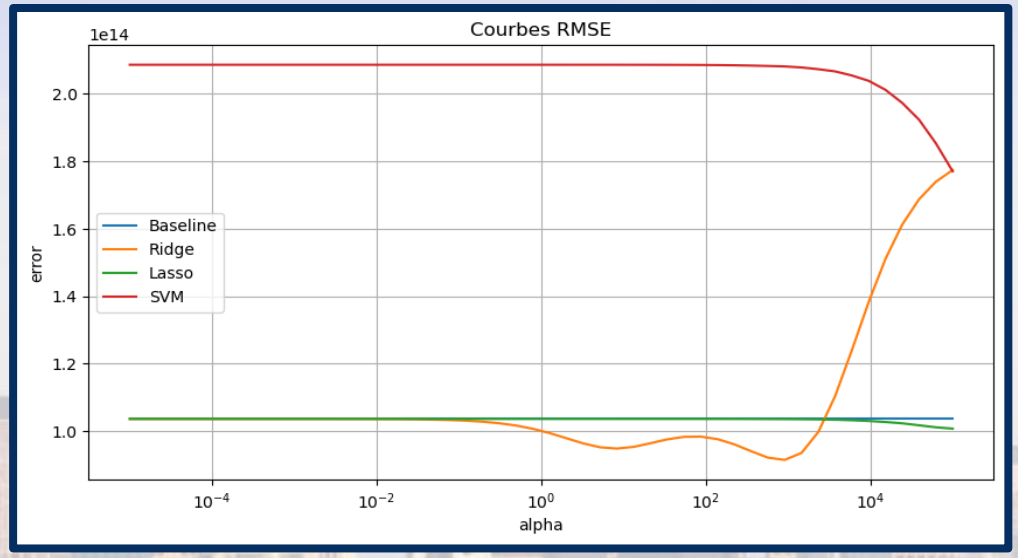
# SIMULATION D'AUTRES MODÈLES & CHOIX D'UN MODÈLE FINAL : TOTALGHGEMISSIONS

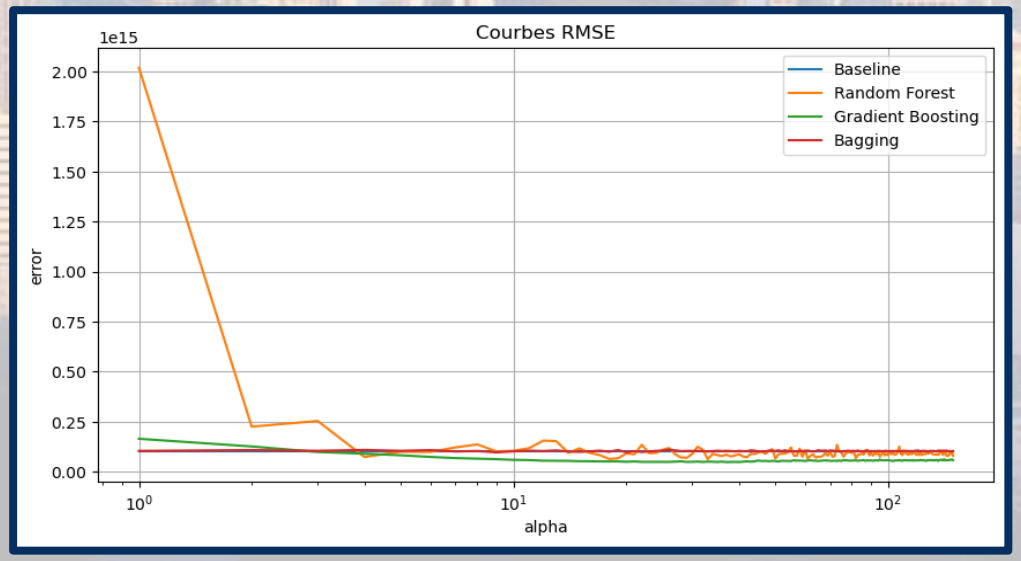



| Model | Best RMSE Parameters | Training RMSE | Test RMSE |
|---|---|---|---|
| Linear Regression | {'model__fit_intercept': False} | 589.47 | 344.58 |
| Lasso Regression | {'model__alpha': 0.1} | 741.47 | 315.35 |
| Ridge Regression | {'model__alpha': 0.1} | 713.05 | 295.68 |
| Random Forest Regression | {'model__n_estimators': 100} | 600.44 | 218.69 |
| Gradient Boosting Regression | {'model__n_estimators': 200} | 595.40 | 217.31 |
| Bagging Regression | {'model__n_estimators': 100} | 593.33 | 324.47 |
| Support Vector Machine | {'model__C': 10} | 699.39 | 204.46 |
| Transformed Target Regression | | 12838.33 | 545.88 |
| Stacking Regression | | 598.43 | 321.09 |

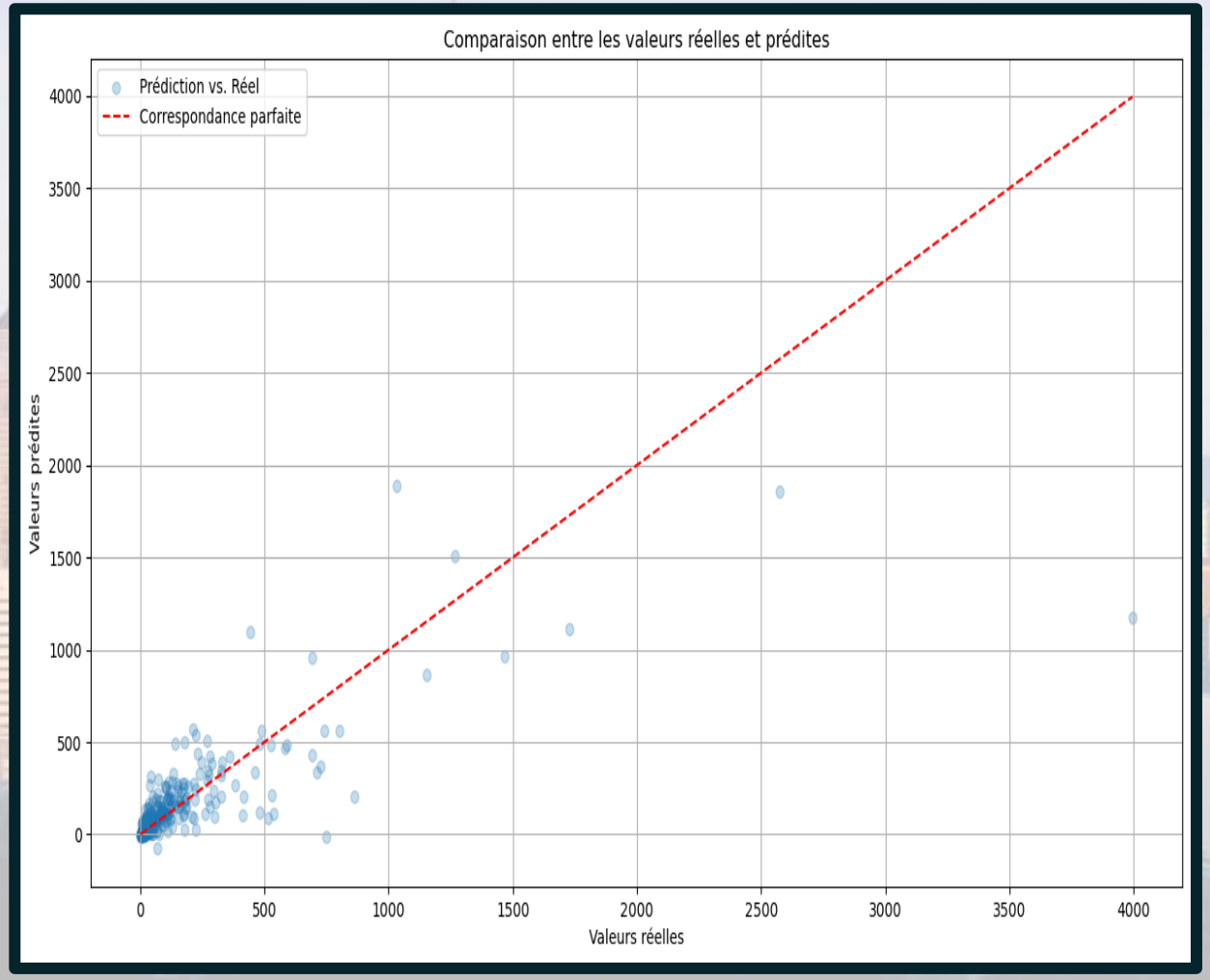
# SIMULATION D'AUTRES MODÈLES & CHOIX D'UN MODÈLE FINAL : SITEENERGYUSEWN(KBTU)




| Model | Best RMSE Parameters | Training RMSE | Test RMSE |
|---|---|---|---|
| Linear Regression | {'model__fit_intercept': False} | 1.776894e+07 | 1.776894e+07 |
| Lasso Regression | {'model__alpha': 0.1} | 1.776834e+07 | 1.018118e+07 |
| Ridge Regression | {'model__alpha': 0.1} | 1.776708e+07 | 1.015967e+07 |
| Random Forest Regression | {'model__n_estimators': 100} | 1.843309e+07 | 9.102588e+06 |
| Gradient Boosting Regression | {'model__n_estimators': 200} | 1.932904e+07 | 7.760410e+06 |
| Bagging Regression | {'model__n_estimators': 100} | 1.777025e+07 | 1.026720e+07 |
| Support Vector Machine | {'model__C': 10} | 2.533614e+07 | 1.444466e+07 |
| Transformed Target Regression | | 3.600371e+08 | 5.410965e+07 |
| Stacking Regression | | 2.412944e+10 | 1.946931e+10 |

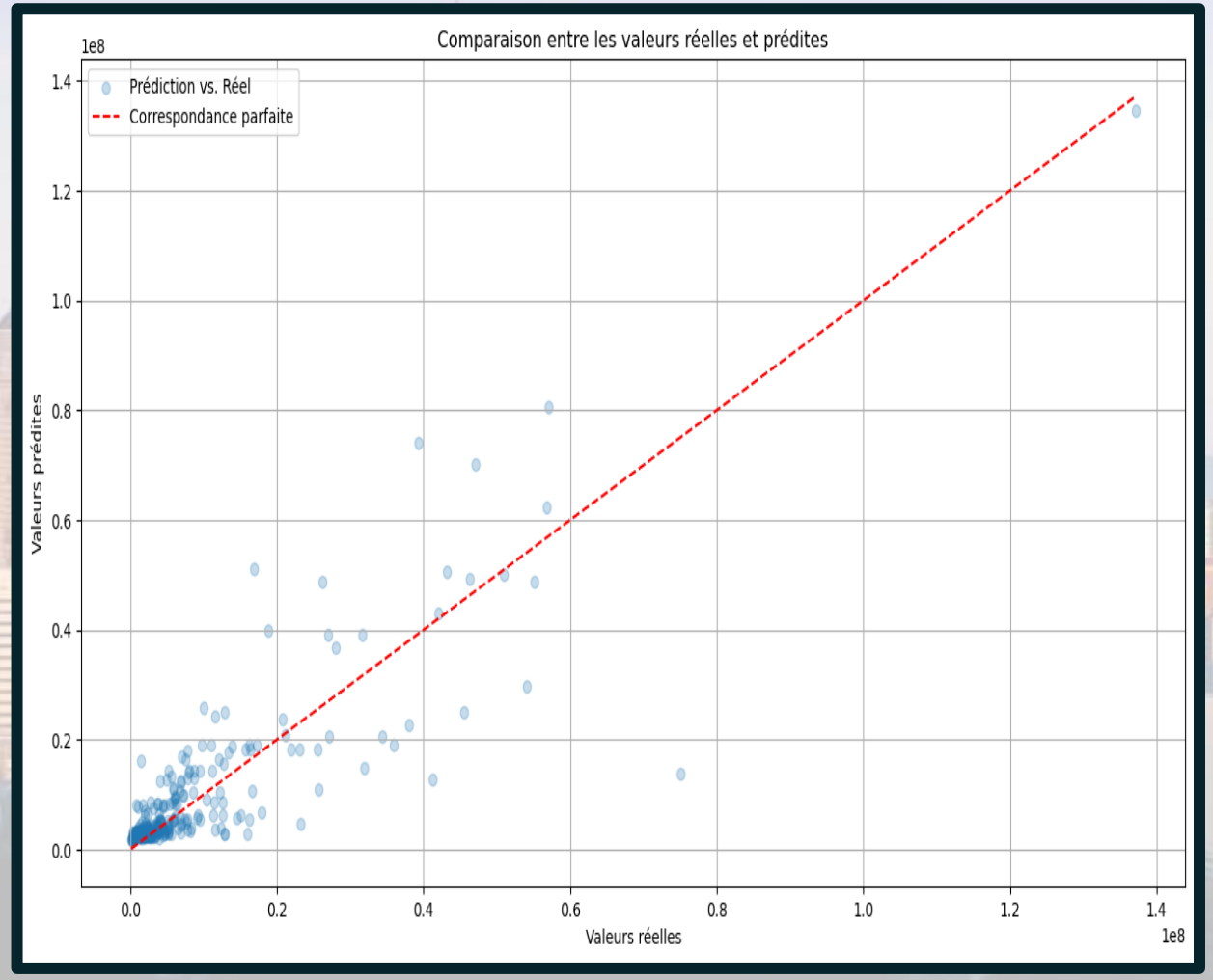
# SIMULATION D'AUTRES MODÈLES & CHOIX D'UN MODÈLE FINAL : TOTALGHGEMISSIONS




## Modèle Final: Gradient Boosting Regression

- N° of estimators = 70
  - $R^2$ = 0.6
  - MAE = 86.54
  - RMSE = 216.81

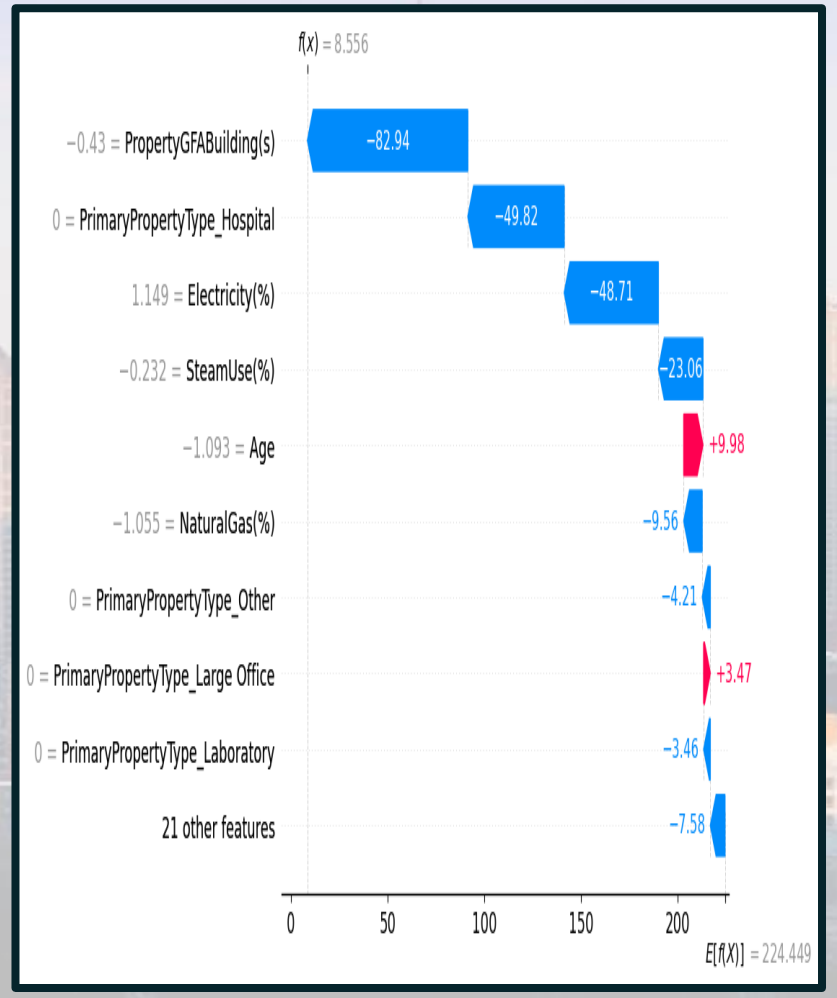
# SIMULATION D'AUTRES MODÈLES & CHOIX D'UN MODÈLE FINAL : SITEENERGYUSEWN(KBTU)

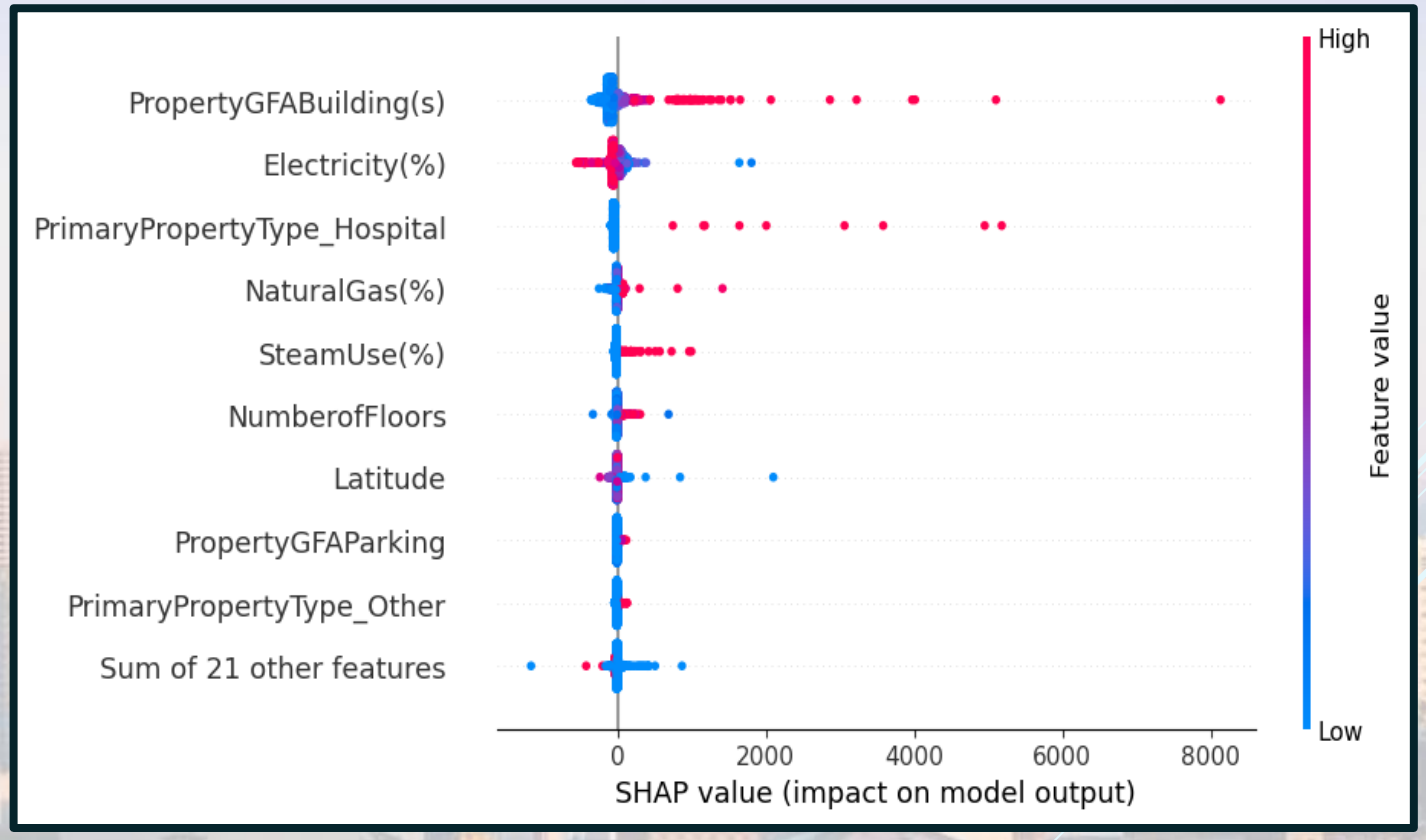



## Modèle Final: Gradient Boosting Regression

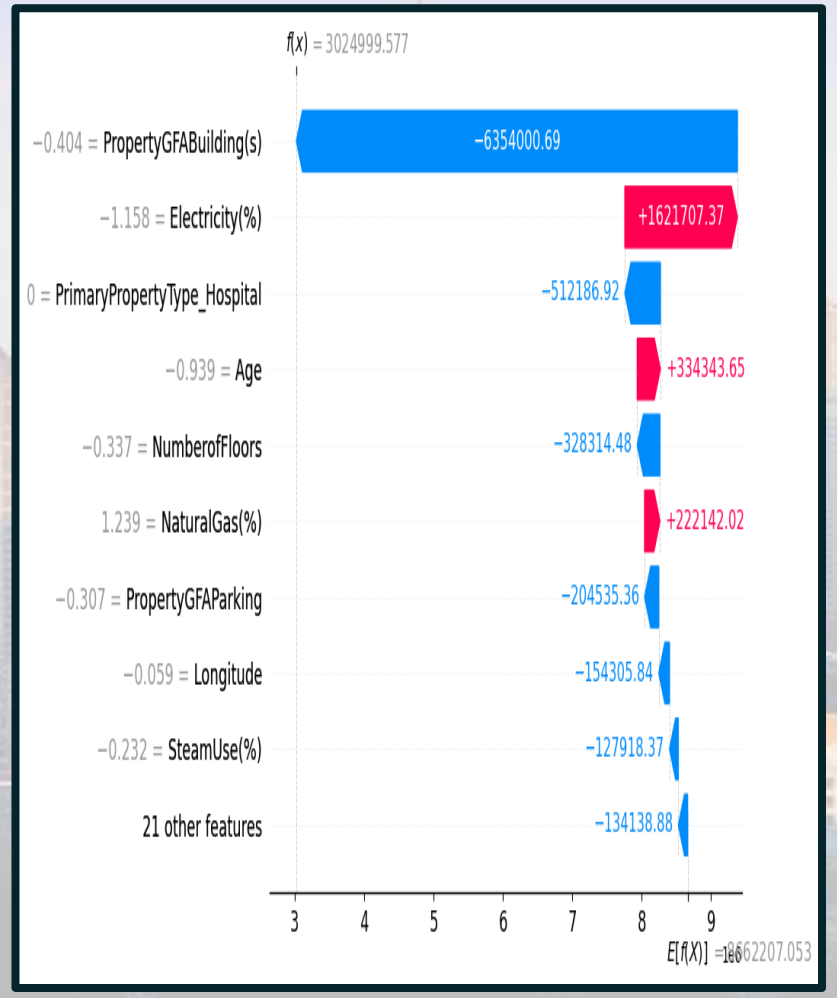
- N° of estimators = 40
  - $R^2$ = 0.71
  - MAE = 3858793.03
  - RMSE = 7224883.46

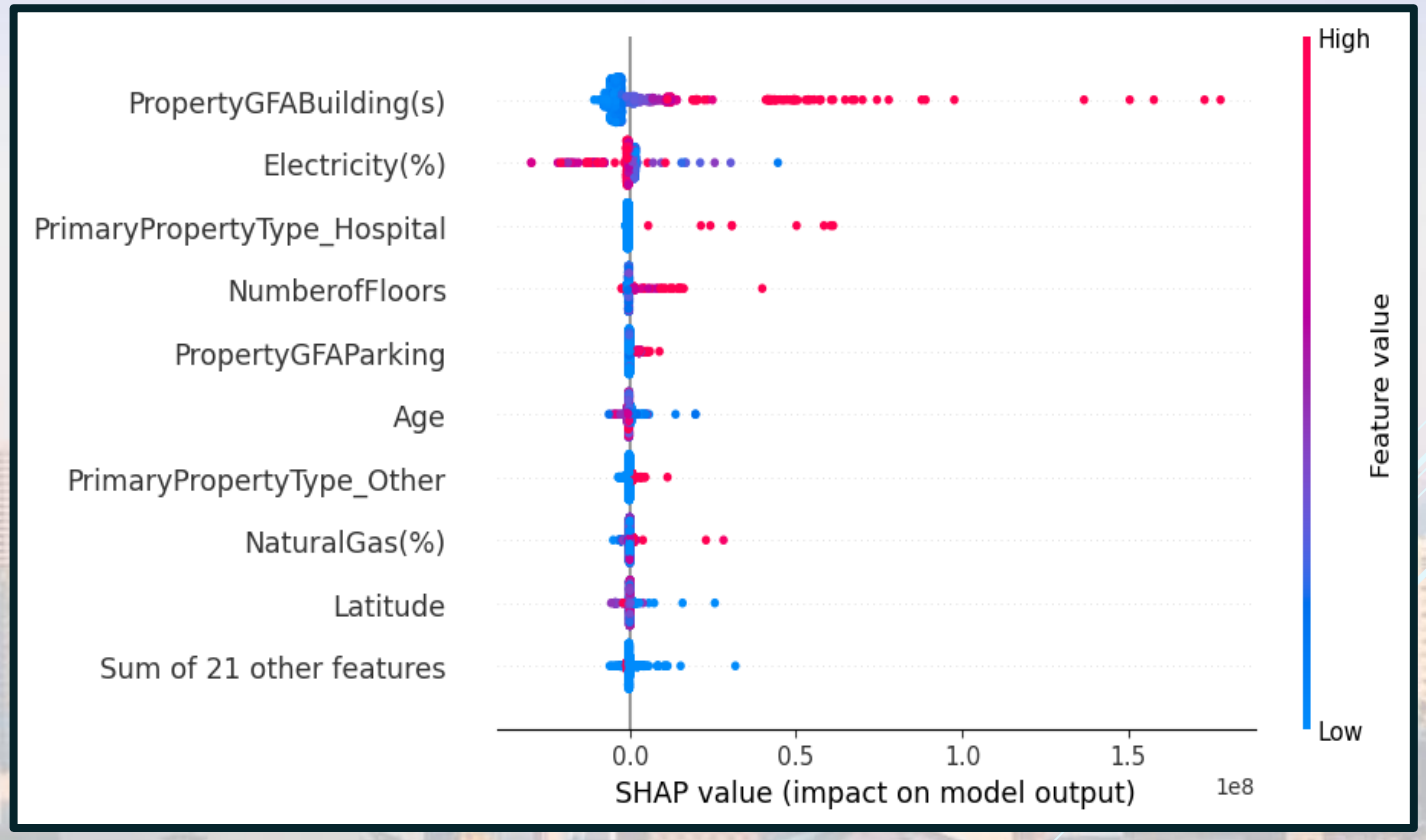
# ANALYSE DE LA "FEATURE IMPORTANCE" GLOBALE & LOCALE : TOTALGHGEMISSIONS



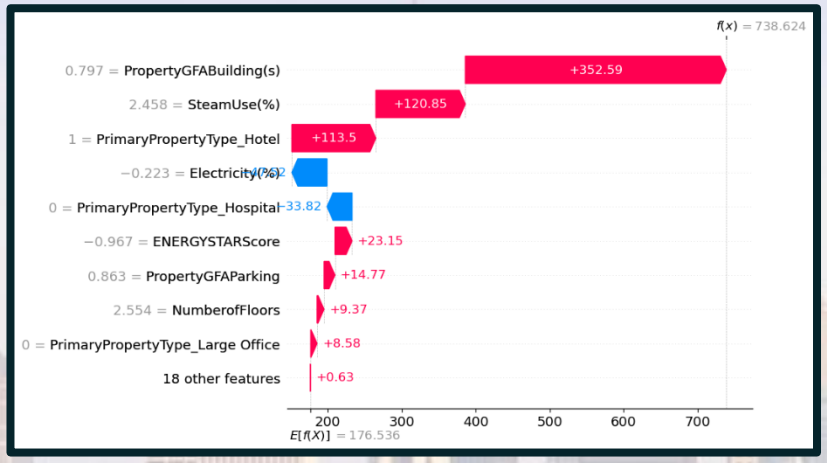



# ANALYSE DE LA "FEATURE IMPORTANCE" GLOBALE & LOCALE : SITEENERGYUSEWN(KBTU)

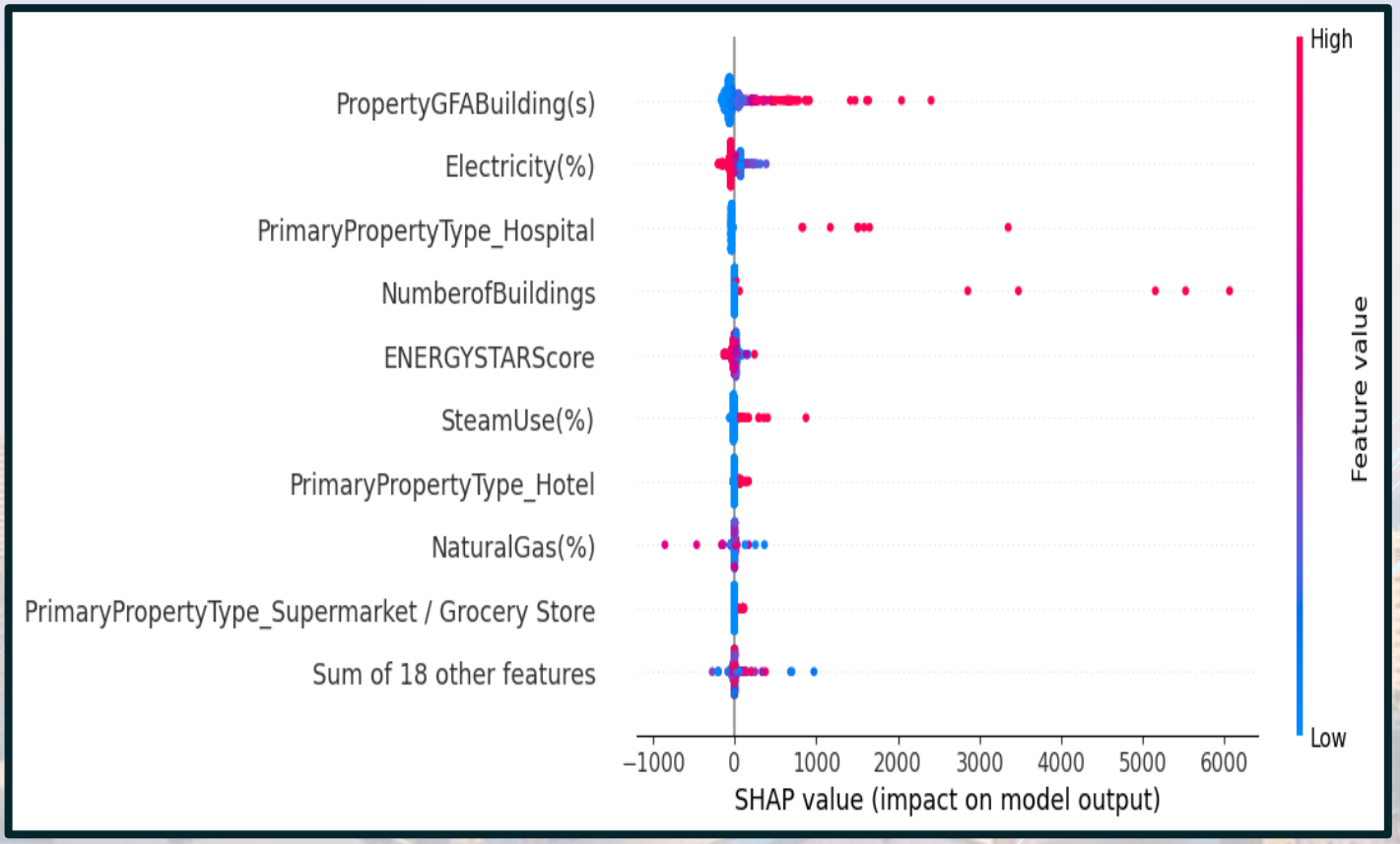


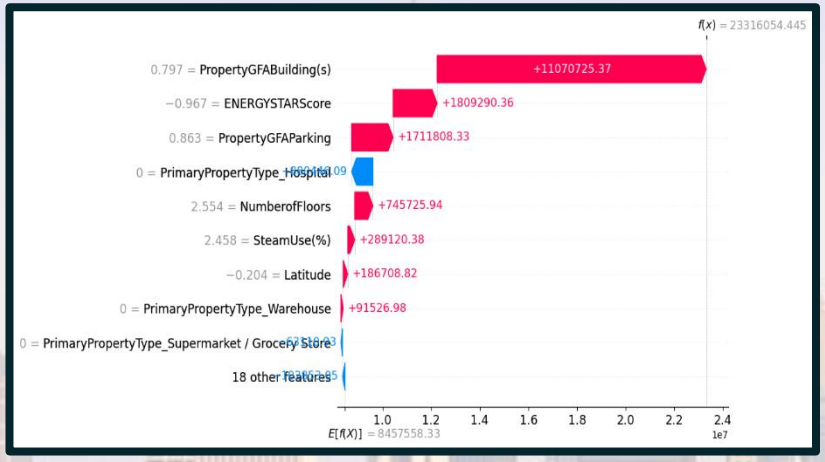
# ANALYSE DE L'INFLUENCE DE L'ENERGYSTARSCORE : TOTALGHGEMISSIONS



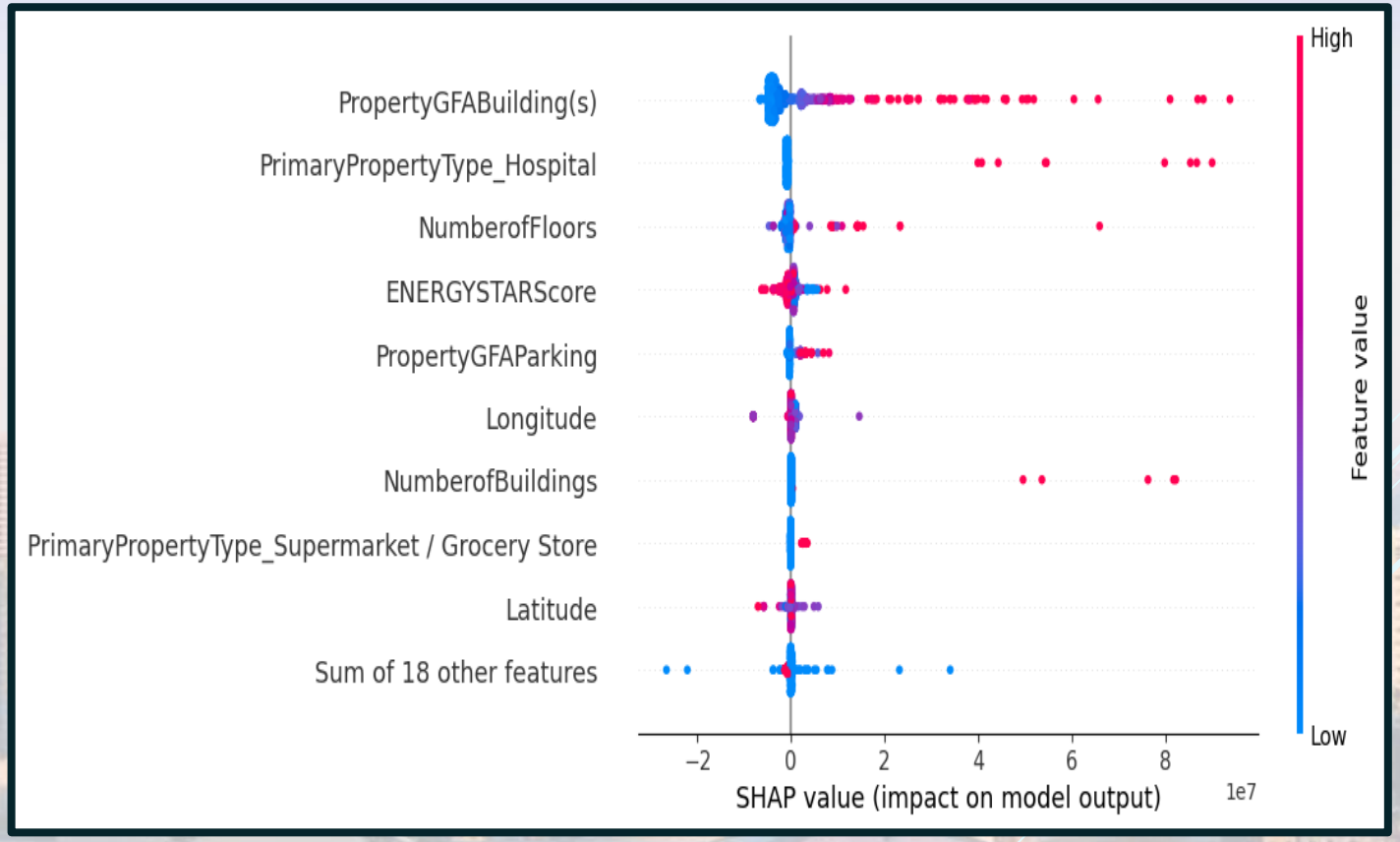

| Métrique | Avec EnergyStarScore | Sans EnergyStarScore |
|---|---|---|
| $R^2$ | 0.12 | 0.11 |
| MAE | 145.58 | 143.95 |
| RMSE | 1134.74 | 1144.52 |



# ANALYSE DE L'INFLUENCE DE L'ENERGYSTARSCORE : SITEENERGYUSEWN(KBTU)




| Métrique | Avec EnergyStarScore | Sans EnergyStarScore |
|---|---|---|
| $R^2$ | 0.14 | 0.12 |
| MAE | 5131972.86 | 5384023.76 |
| RMSE | 31868482.04 | 32132763.43 |



# CONCLUSION

**Modèle & Paramétrages Final – TotalGHGEmissions :**

**Gradient Boosting Regression (learning_rate=0.1, n_estimators= 70)**

**Modèle & Paramétrages Final – TotalGHGEmissions :**

**Gradient Boosting Regression (learning_rate=0.1, n_estimators= 40)**

**Variances particulièrement élevées**

**Généraliser Energy Star Score à tout le dataset = Performance++**

**Notifier Variables Cibles aux Valeurs Aberrantes**

# Merci pour votre attention