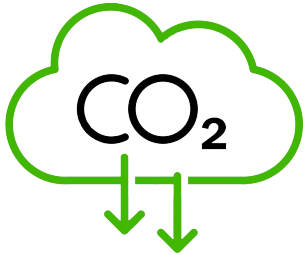
Prédiction des émissions et de la consommation d'énergie de bâtiments

15/12/2023

Seattle

Introduction

- Seattle: objectif 2050, neutralité carbone

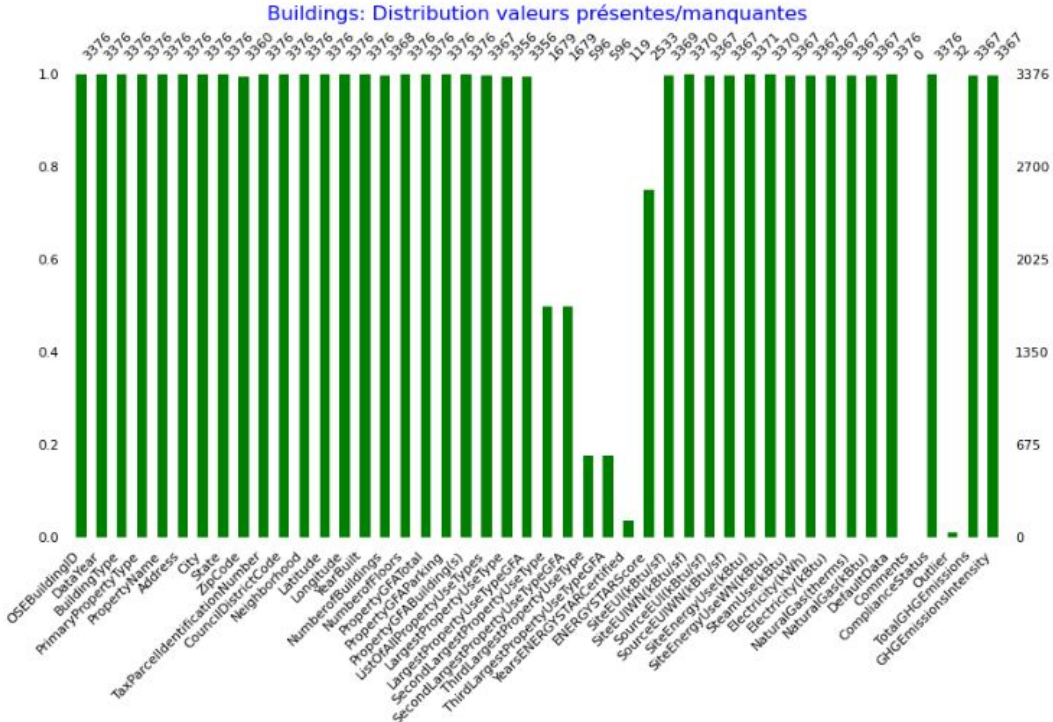


- Contexte:
  - Etude des consommations et des émissions de bâtiments
  - Relevés minutieux effectués en 2016 - coûteux à obtenir

- But de cette étude:
  - Prédire les consommations et émissions (sans les relevés)
  - Identifier le meilleur algorithme de prédiction
  - Evaluer l'intérêt de l'Energy Star Score dans la prédiction d'émissions

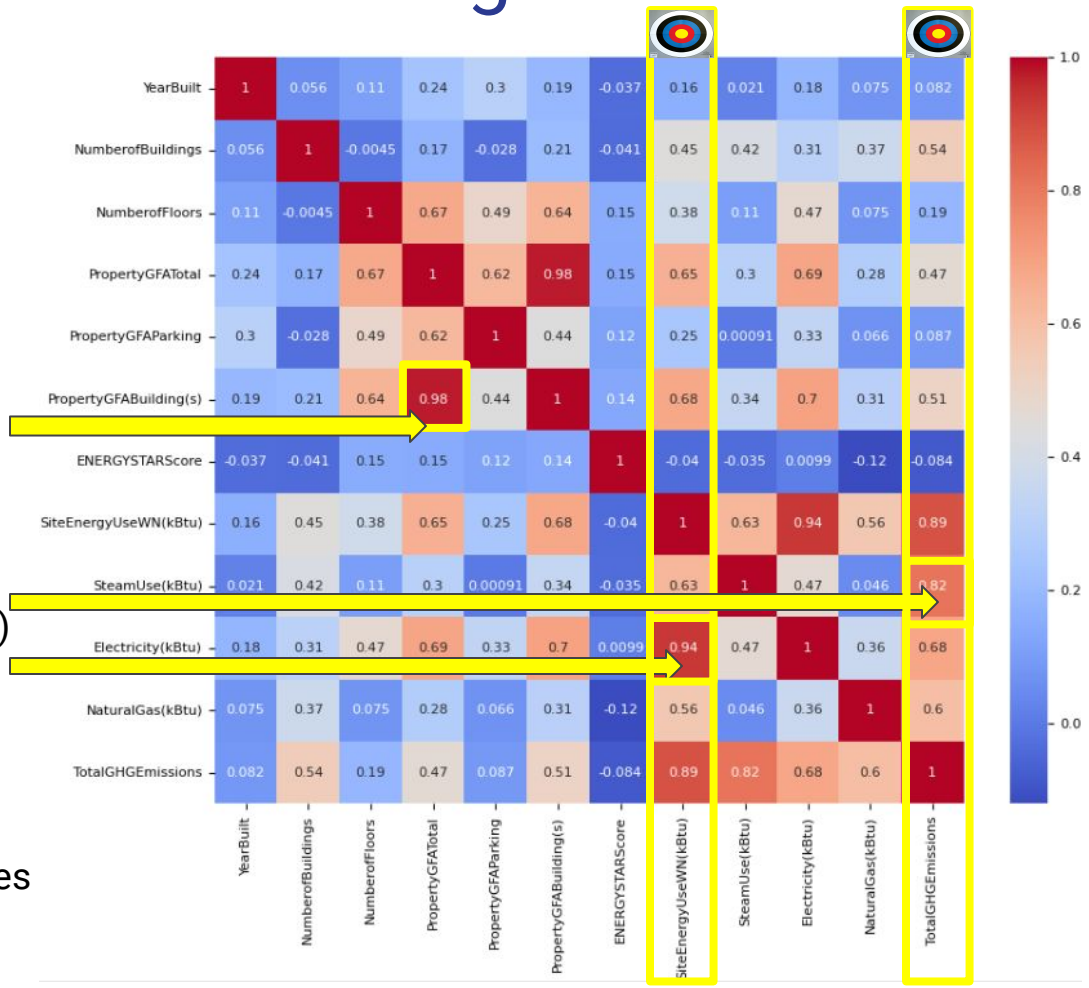
Description du jeu de données

- Source: Ville de Seattle - développement durable et environnement
- Contenu: relevés de 2016 - bâtiments non résidentiels
- Variables principales: localisation, surfaces, types, consommation par type …
- 1 fichier de 3376 lignes et 46 colonnes
- Remplissage: 87,15%
- Nettoyage: Doublons, outliers, métier …



Corrélation des features et des targets

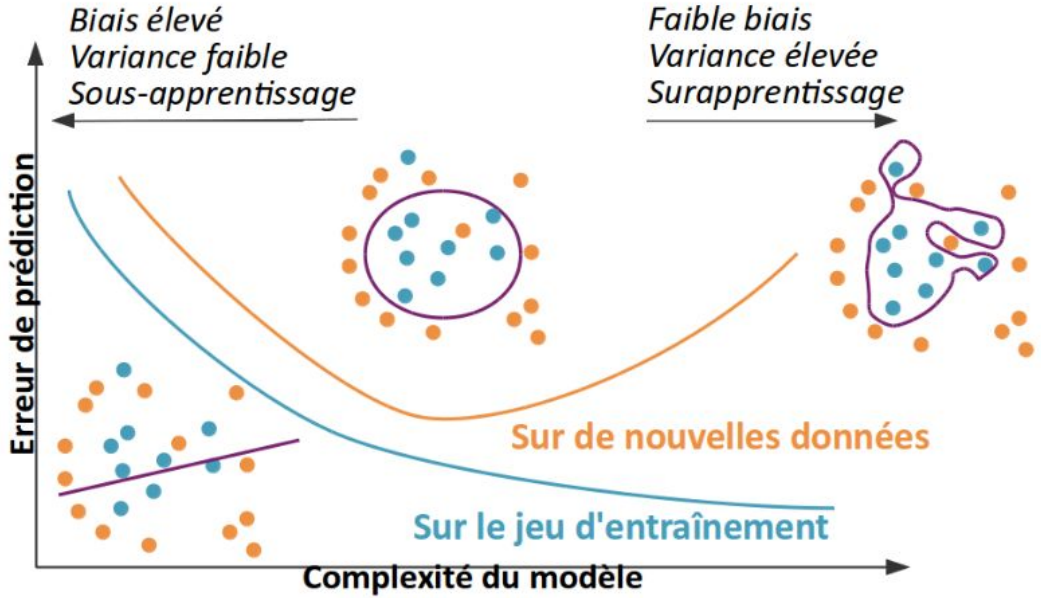
- Variables trop corrélées entre elles
  (ex: surface building et surface totale)
- Fuite de données (features <-> targets)
- Solution: transformer certaines features



Partie 1: Feature engineering

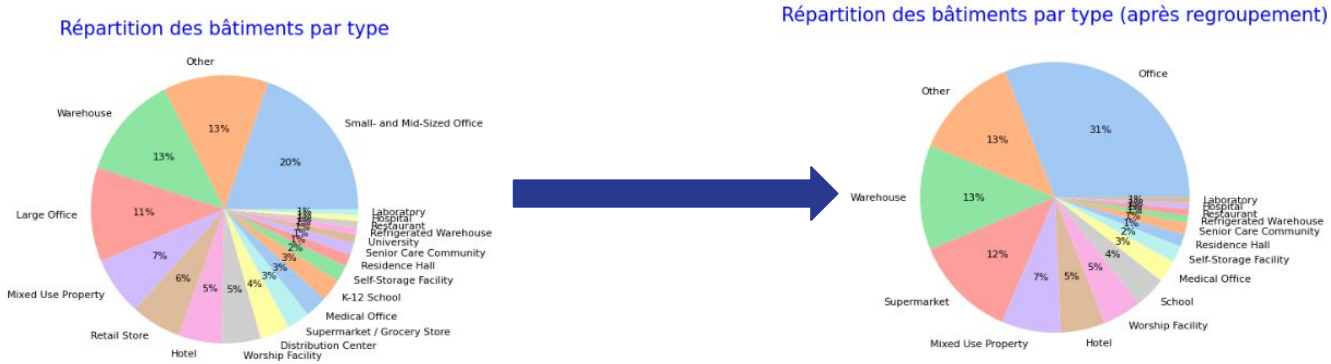
Feature engineering: vue d’ensemble

- Rendre les données compatibles aux modèles (valeurs numériques)
- Améliorer les performances (ajuster complexité modèle)
- Techniques: mise en intervalle, encodage, passage au log …



Transformation des variables catégorielles

- Regroupement des catégories



- Transformer des catégories: n catégories -> n variables binaires

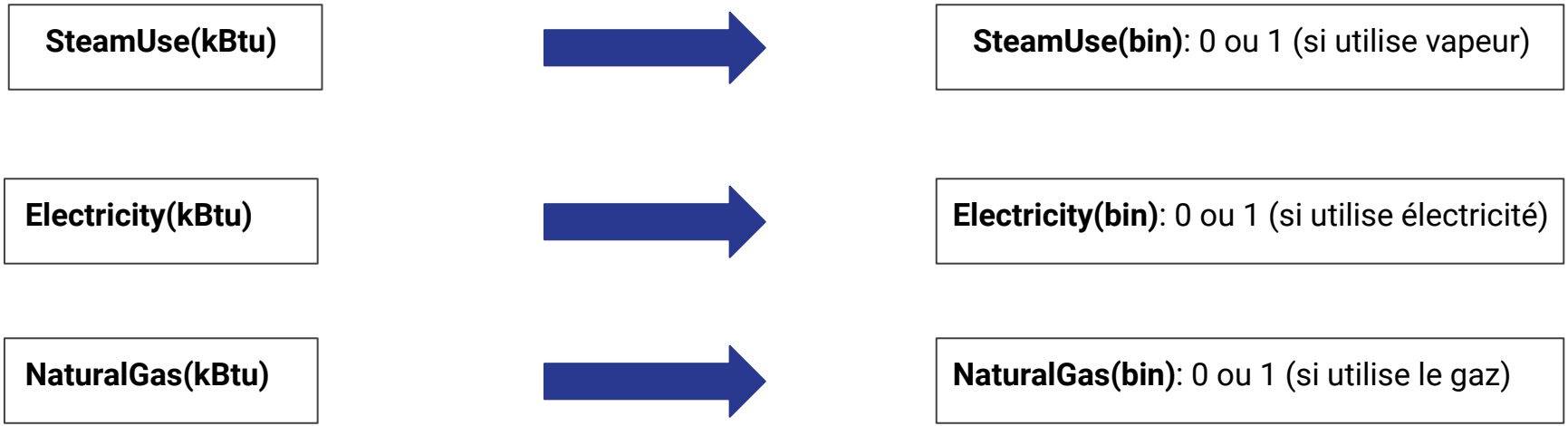
```
DOWNTOWN
GREATER DUWAMISH
LAKE UNION
MAGNOLIA / QUEEN ANNE
EAST
NORTHEAST
NORTHWEST
BALLARD
NORTH
CENTRAL
SOUTHWEST
DELRIDGE
SOUTHEAST
North
Delridge
Ballard
Central
Northwest
```

Encodage OneHot

```
Neighborhood_BALLARD
Neighborhood_CENTRAL
Neighborhood_DELRIDGE
Neighborhood_DOWNTOWN
Neighborhood_EAST
Neighborhood_GREATER DUWAMISH
Neighborhood_LAKE UNION
Neighborhood_MAGNOLIA / QUEEN ANNE
Neighborhood_NORTH
Neighborhood_NORTHEAST
Neighborhood_NORTHWEST
Neighborhood_SOUTHEAST
Neighborhood_SOUTHWEST
```

Transformation des variables quantitatives en binaire

Objectif: se passer des relevés de consommation pour prédire

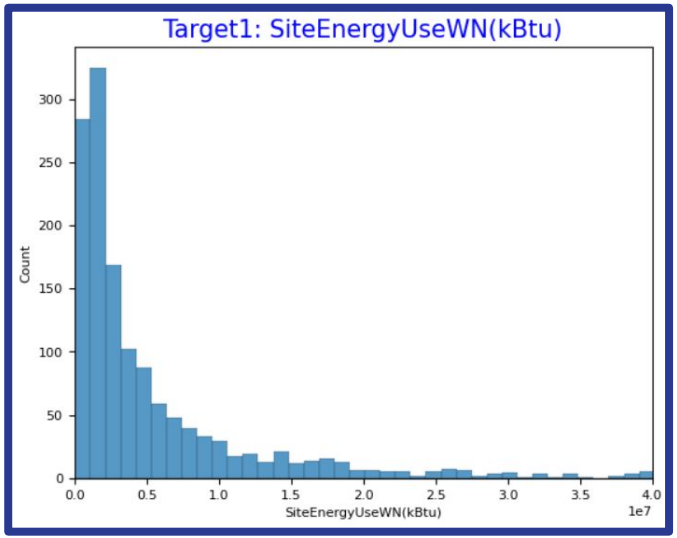


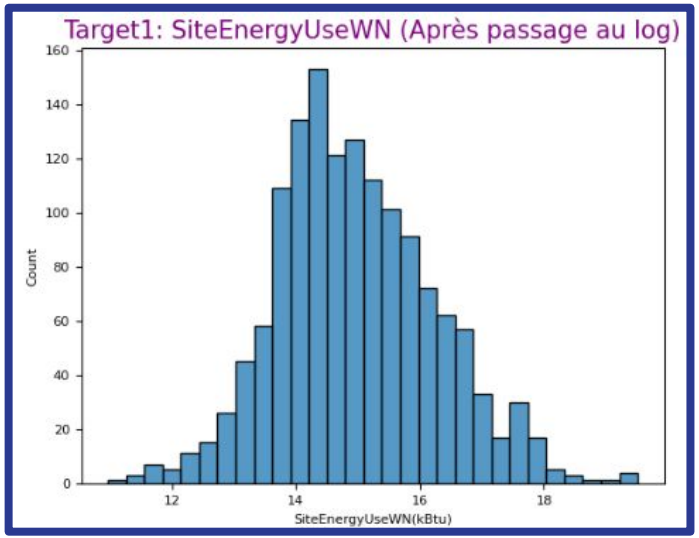
Transformation de variable quantitative en ratio

- Variables concernées: PropertyGFATotal et PropertyGFAParking
- Objectifs:
  - réduire le nombre de variables
  - avoir une valeur de ratio entre 0 et 1
- Principe: calcul du ratio PropertyGFAParking/PropertyGFATotal

Passage au log

- Réduire l'amplitude de variables sans perte d'information
- Réduire l'influence des valeurs atypiques
- Targets + PropertyGFABuilding(s)





Partie 2: Modélisation

Etapes de prétraitement

- **Séparation des features et des targets**
  - Une matrice X et un vecteur Y
  - Target 1: consommation totale d'énergie
  - Target 2: émission de CO2

- **Séparation des données de train et de test**
  - Fonction train_test_split de Scikit-learn
  - 30% des données dans le jeu de test
  - Attribut random_state pour avoir toujours les mêmes jeux de train et de test

- **Standardisation des données** => moyenne nulle et un écart type de 1

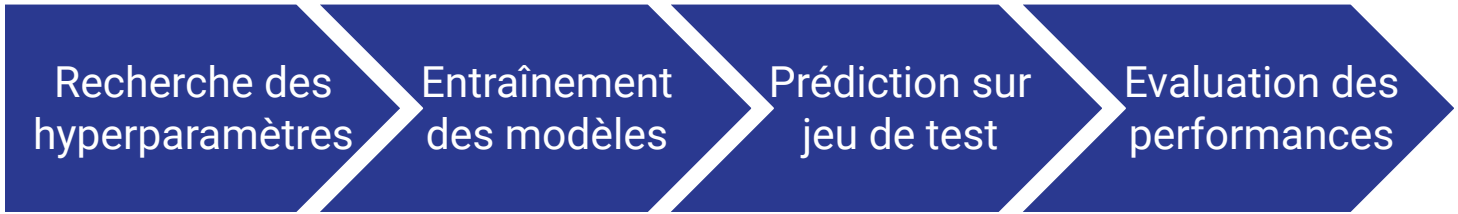
Modèles sélectionnés

- **Baseline**: Régresseur naïf

- **Modèles linéaires**
  - Régression linéaire
  - Régression Ridge
  - SVR linéaire

- **Modèles non linéaires**
  - SVR non linéaire
  - ElasticNet
  - GradientBoosting
  - RandomForest

Etapes de la modélisation



Comparaison des performances

Target1: Consommation

Algorithme	RMSE	R2	MAE	Train time	Test Time
Dummy	1.293862	-0.000468	1.025597	0.000346	0.000079
LinearRegression	0.717609	0.692247	0.548646	0.002200	0.000361
RidgeRegression	0.717726	0.692146	0.548623	0.001182	0.000323
ElasticNet	0.715433	0.685638	0.542945	0.001163	0.000273
LinearSVR	0.725273	0.685638	0.547614	0.021239	0.000502
NonLinearSVR	0.788401	0.628531	0.613518	0.050658	0.027183
GradientBoosting	0.693427	0.712638	0.522880	5.351501	0.013460
RandomForest	0.720334	0.689905	0.547403	0.730587	0.020331

Target2: Emissions

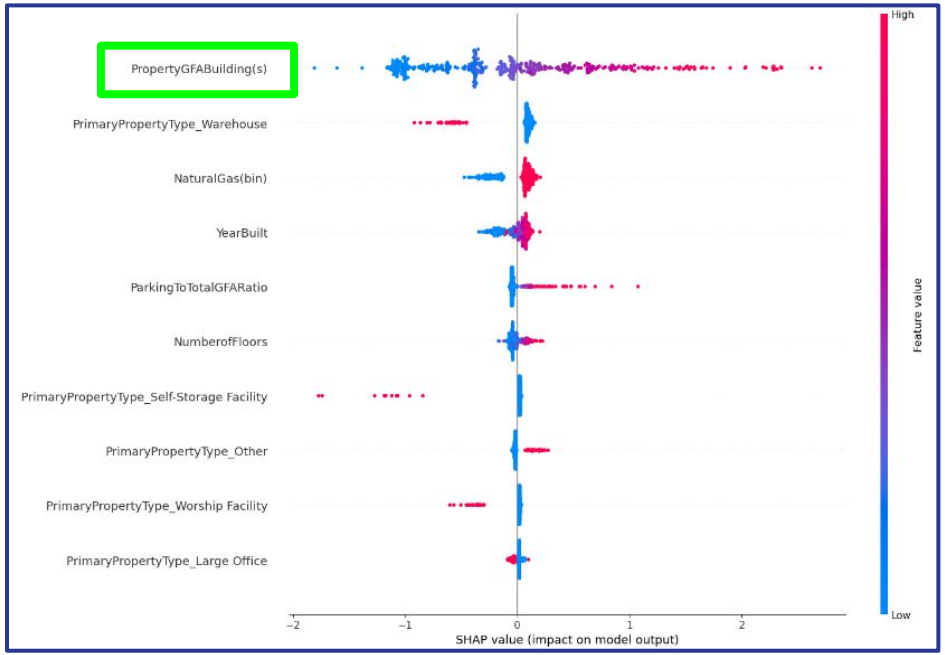
Algorithme	RMSE	R2	MAE	Train time	Test Time
Dummy	1.444219	-0.000203	1.133790	0.000551	0.000214
LinearRegression	0.836582	0.664386	0.647929	0.001771	0.000285
RidgeRegression	0.837484	0.663663	0.648851	0.001512	0.000327
ElasticNet	0.831768	0.668238	0.645645	0.001469	0.000325
LinearSVR	0.838664	0.662714	0.651735	0.017630	0.000423
NonLinearSVR	0.920211	0.593934	0.690176	0.044847	0.027910
GradientBoosting	0.775989	0.711243	0.597467	6.817825	0.012745
RandomForest	0.830978	0.668868	0.646089	28.837639	0.554988

- **Le GradientBoosting a les meilleurs performances** (RMSE, R2, MAE)
- Train time et test time: pas des critères déterminants de choix

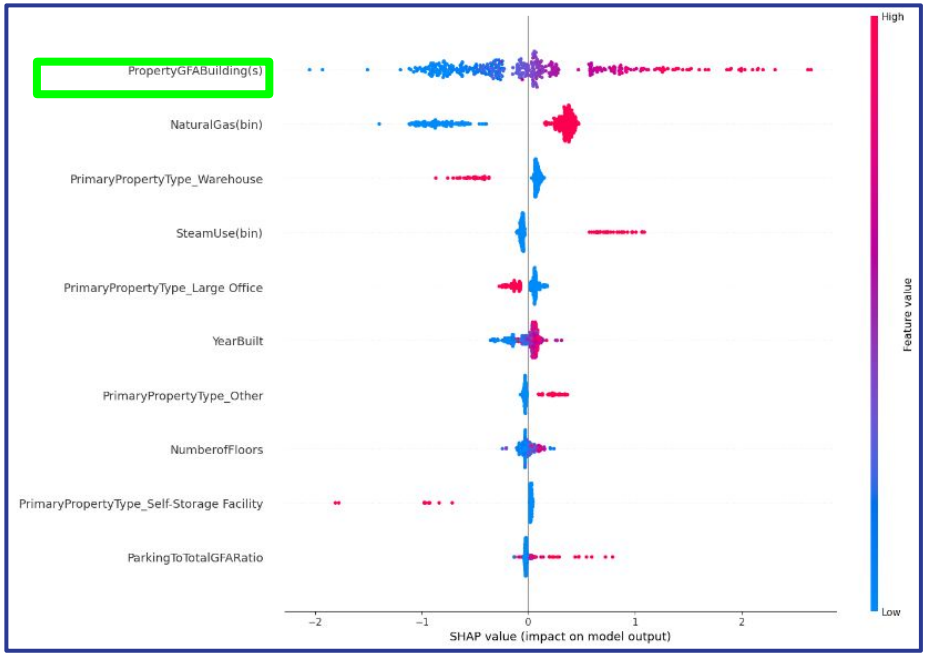
Analyse de l'importance des variables avec SHAP

- **SHAP**: Contribution de variables à la différence entre la valeur prédite par le modèle et la moyenne des prédictions
- La surface des bâtiments est la variable la plus importante

Target1: Consommation



Target2: Emissions



Intégration de l'Energy Star Score (½)

Target1: Consommation

	Algorithme	RMSE	R2	MAE
	GradientBoosting	0.693672	0.712435	0.523052
	GradientBoosting with EnergyStarScore	0.545813	0.834670	0.359449

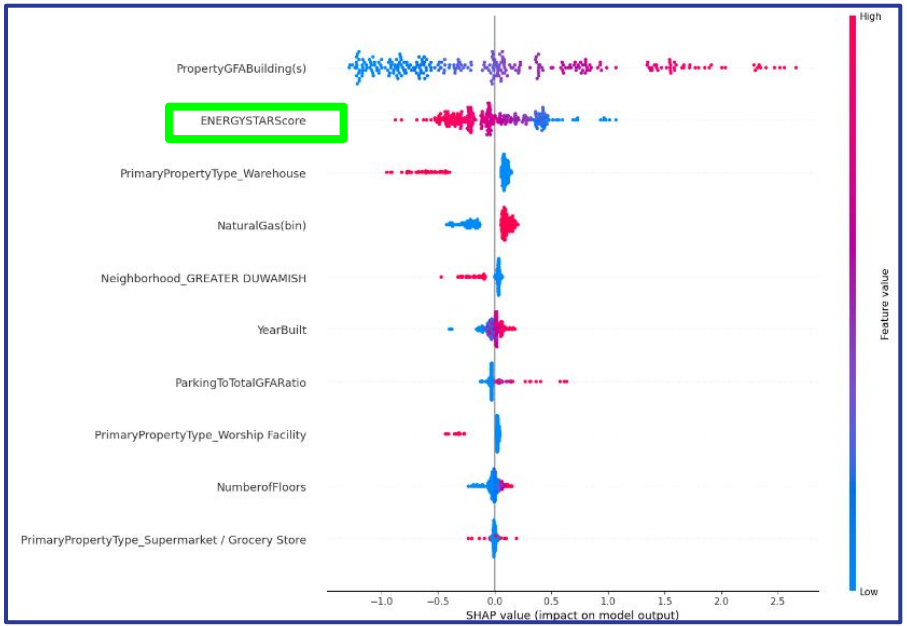
Target2: Emissions

	Algorithme	RMSE	R2	MAE
	GradientBoosting	0.775989	0.711243	0.597467
	GradientBoosting with EnergyStarScore	0.672708	0.805607	0.512721

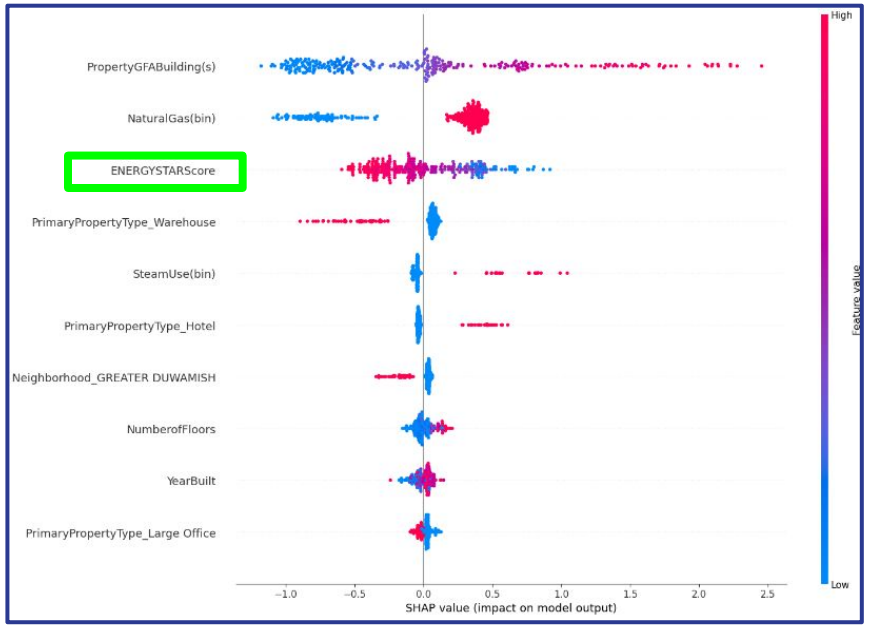
- L'intégration de l'**Energy Star Score** au modèle GradientBoosting **améliore nettement les performances** (RMSE, R2, MAE)

Intégration de l'Energy Star Score (2/2)

Target1: Consommation



Target2: Emissions



- L'**Energy Star Score** est la 2ème variable la plus importante pour prédire la consommation et la 3ème pour prédire les émissions

Conclusion

- Le **GradientBoosting** présente les meilleurs performances

- L’**Energy Star Score** bien que fastidieux à calculer, **améliore nettement les performances** de notre meilleur modèle de prédiction d’émissions