Anticipez les besoins en consommation de bâtiments

Seattle

21/05/2023

Khoty WOLIE

SOMMAIRE

1. Contexte et problématique
2. Présentation des données et feature engineering
3. Méthodologie de modélisation
4. Résultats pour la Target N°1 : SiteEnergyUse(kBtu)
5. Résultats pour la Target N°2 : TotalGHGEmissions
6. Conclusions

1.Contexte et problématique

Approche :
Analyser la consommation énergétique et les émissions de CO2 des bâtiments à usage non résidentiel.

Collectes :
Des relevés minutieux ont été effectués par les agents de la ville en 2016.

Données :
Les données sont disponibles sur le site : [2016 Building Energy Benchmarking](https://data.seattle.gov/Permitting/2016-Building-Energy-Benchmarking/2bpz-gwpy/about_data)

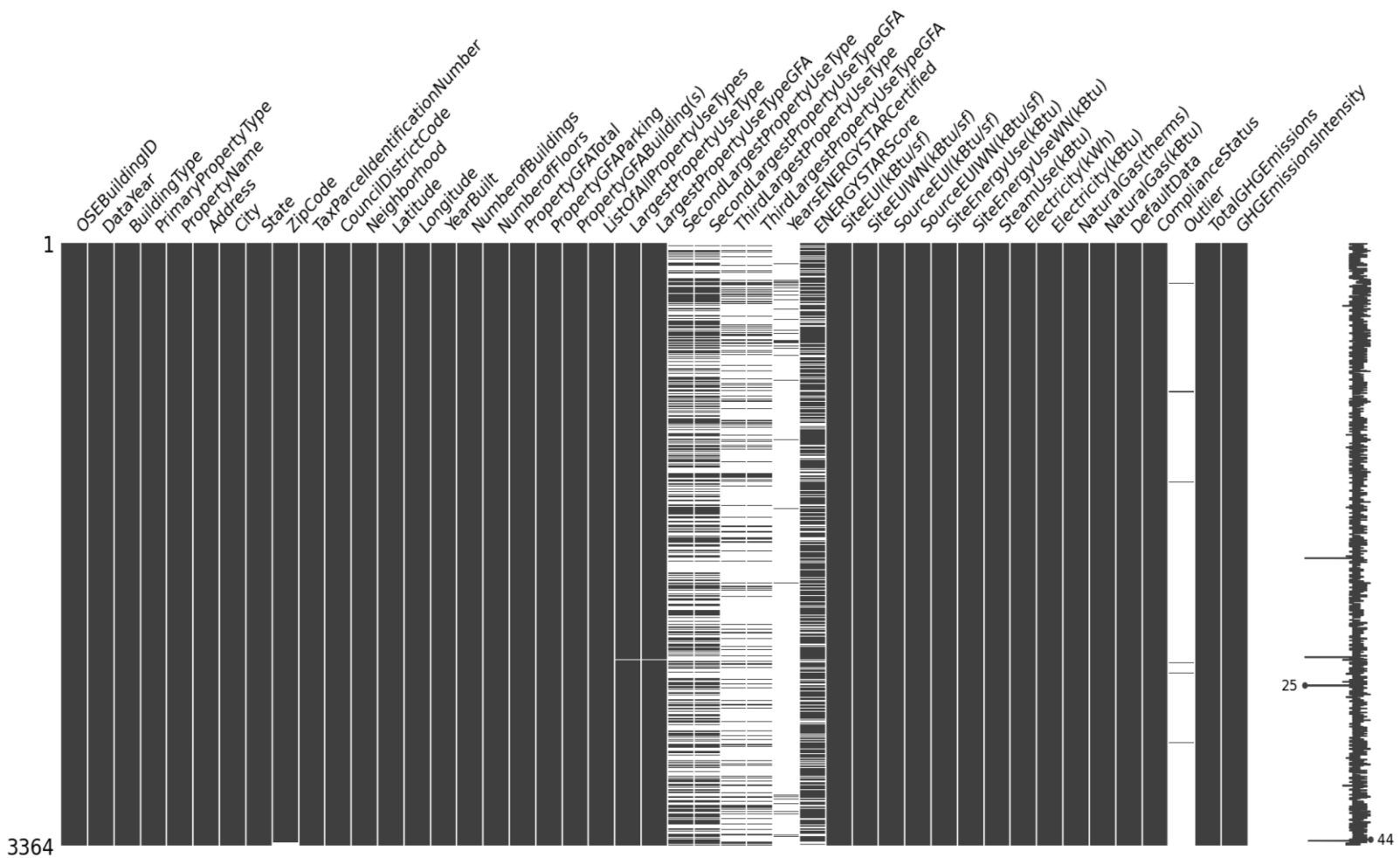
Problématique :
La collecte de ces relevés est coûteuse.

Objectif :
Atteindre la neutralité carbone pour la ville d'ici 2050.
Prédire les émissions de CO2 et la consommation énergétique totale des bâtiments non encore mesurés en utilisant les données existantes et les premiers relevés (**électricité, gaz, vapeur**).

2.Présentation des données et feature engineering

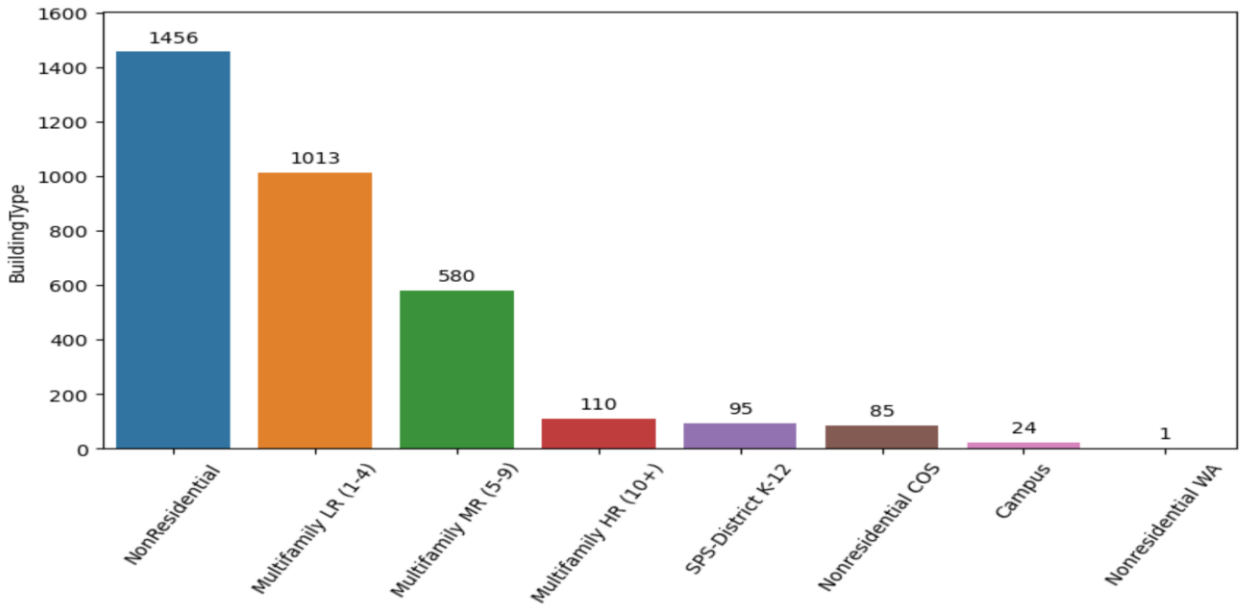
- Nombres de variables: 46
- Nombre de doublon: 0
- Nombres de % NAN: 13%
- Nombres de construction: 3376

- Nous avons 15 variables qualitatives.
- Nous avons 30 variables numériques

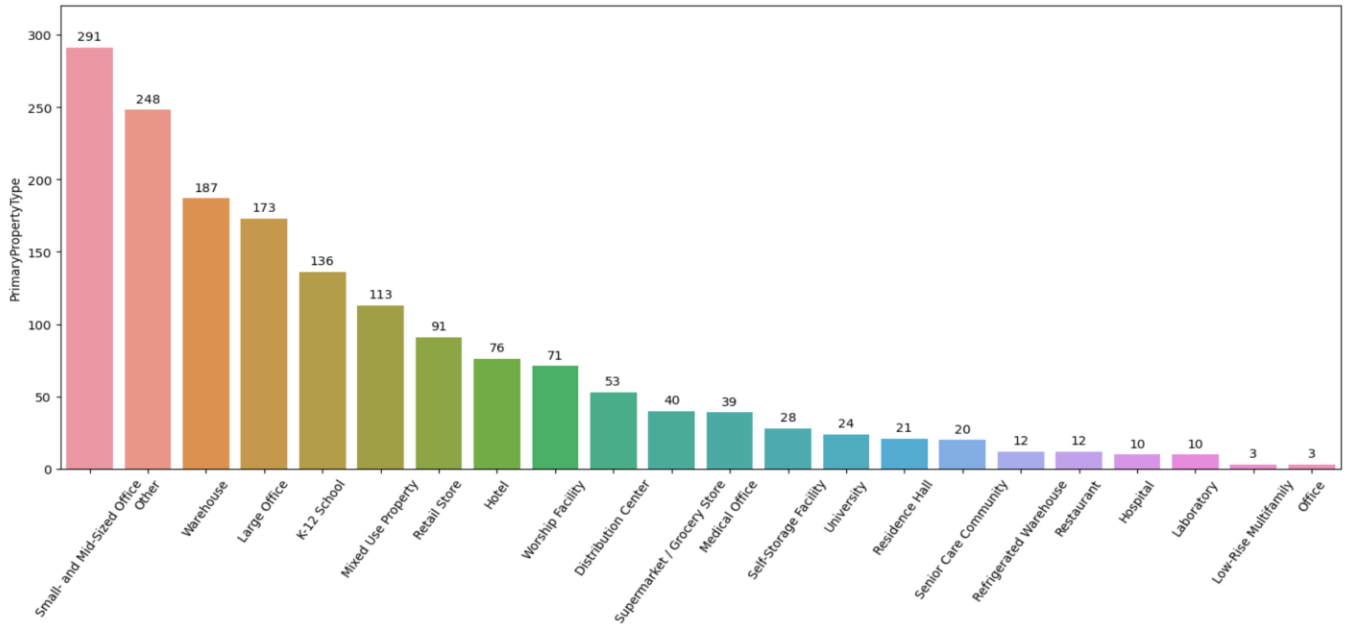


Types de variables des bâtiments

Types de constructions



Types de bâtiments principaux

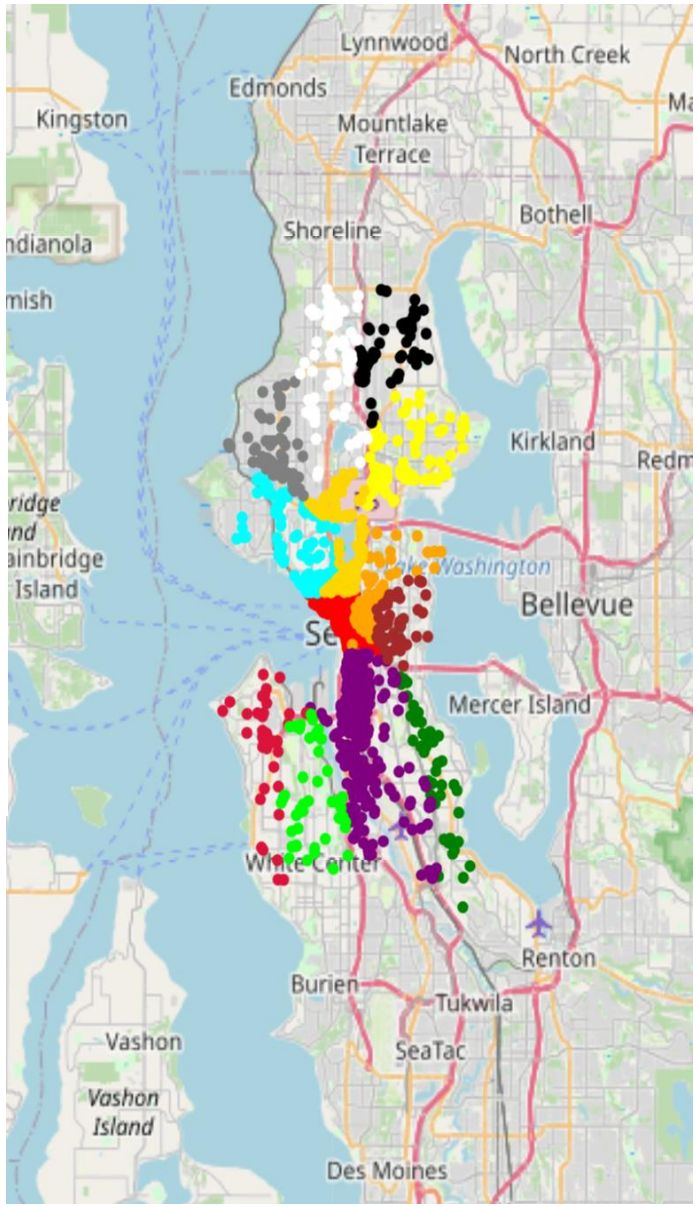


Type de variables de localisation des bâtiments

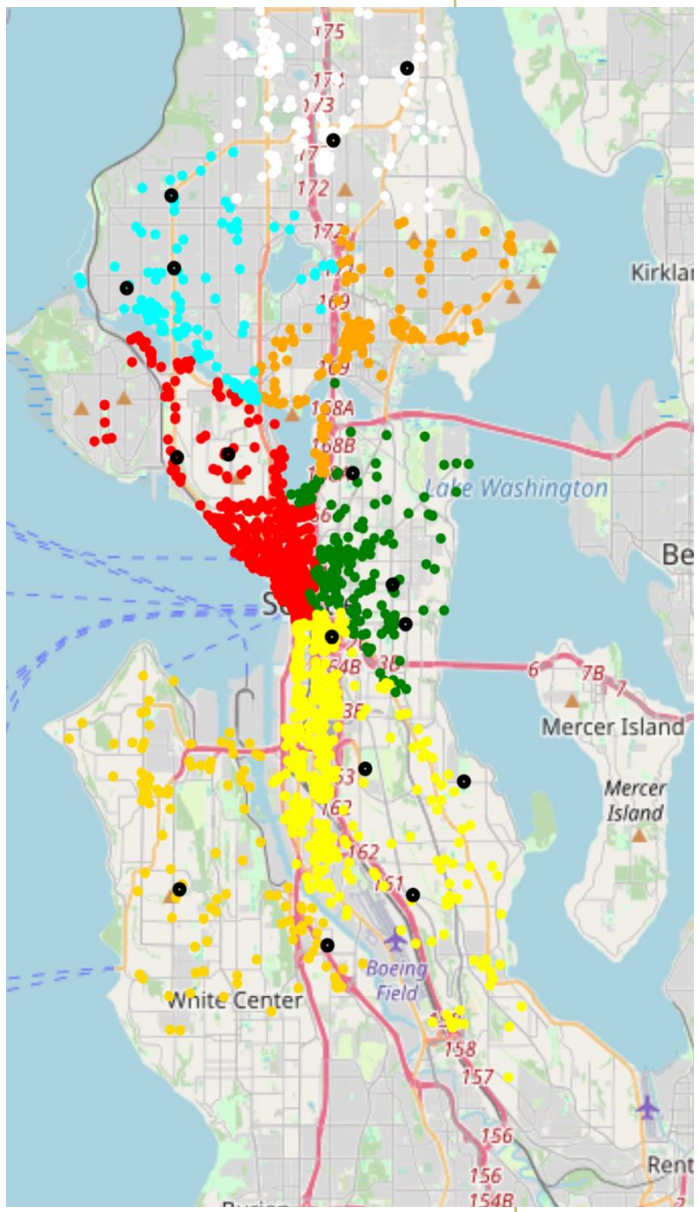
- Quartier
- District
- Code postal & Adresse
- Latitude & longitude

Observation:
Les bâtiments sans code postal (en noir) sont répartis dans divers districts.

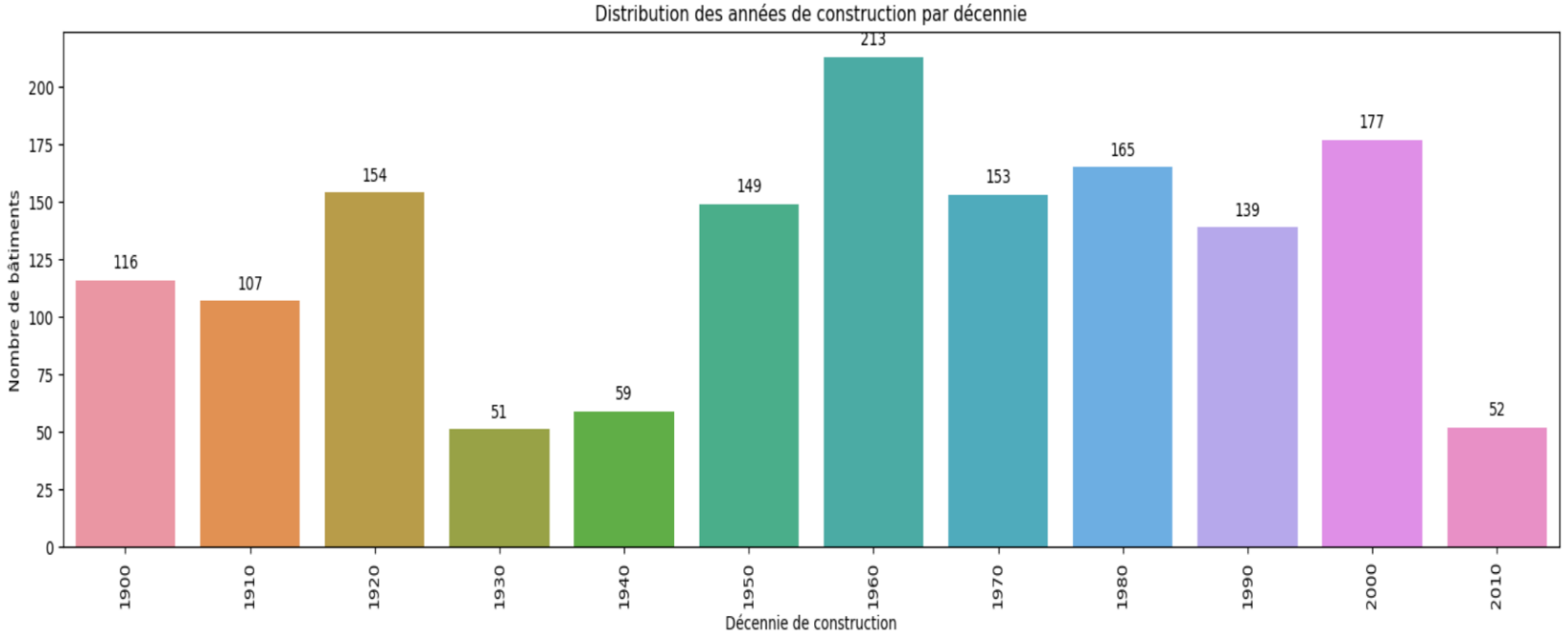
quartier

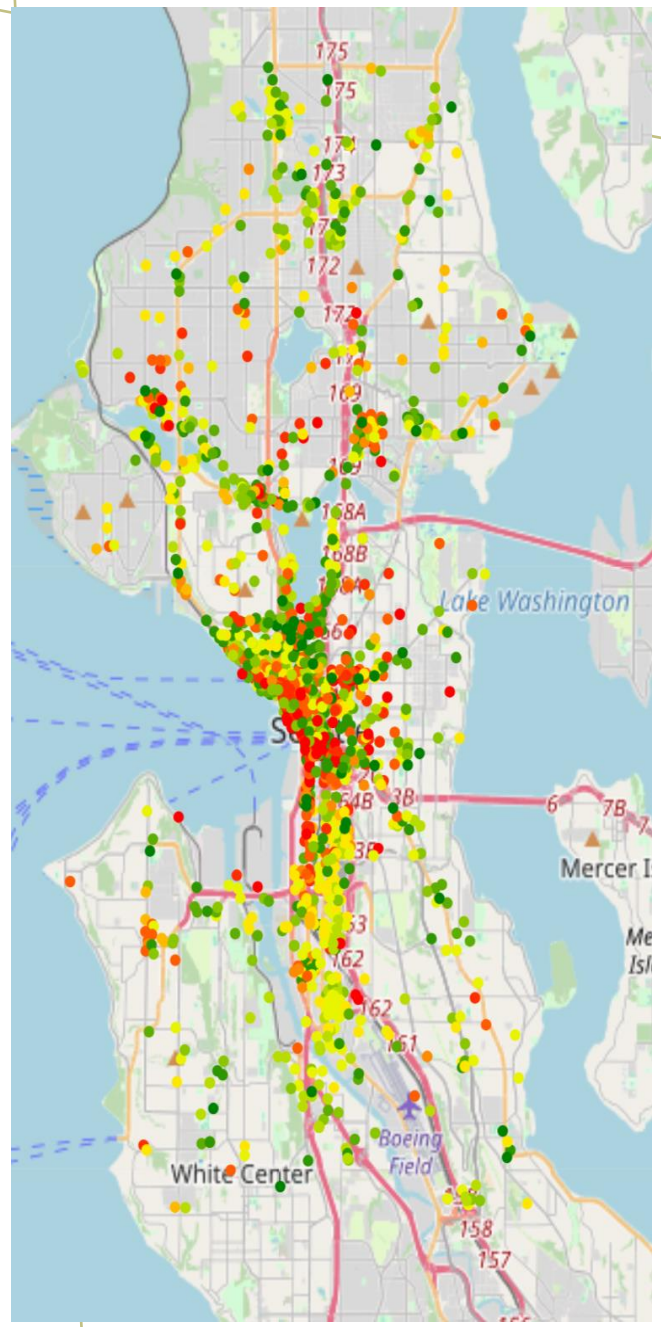


District



Type de variables générales concernant toujours les bâtiments

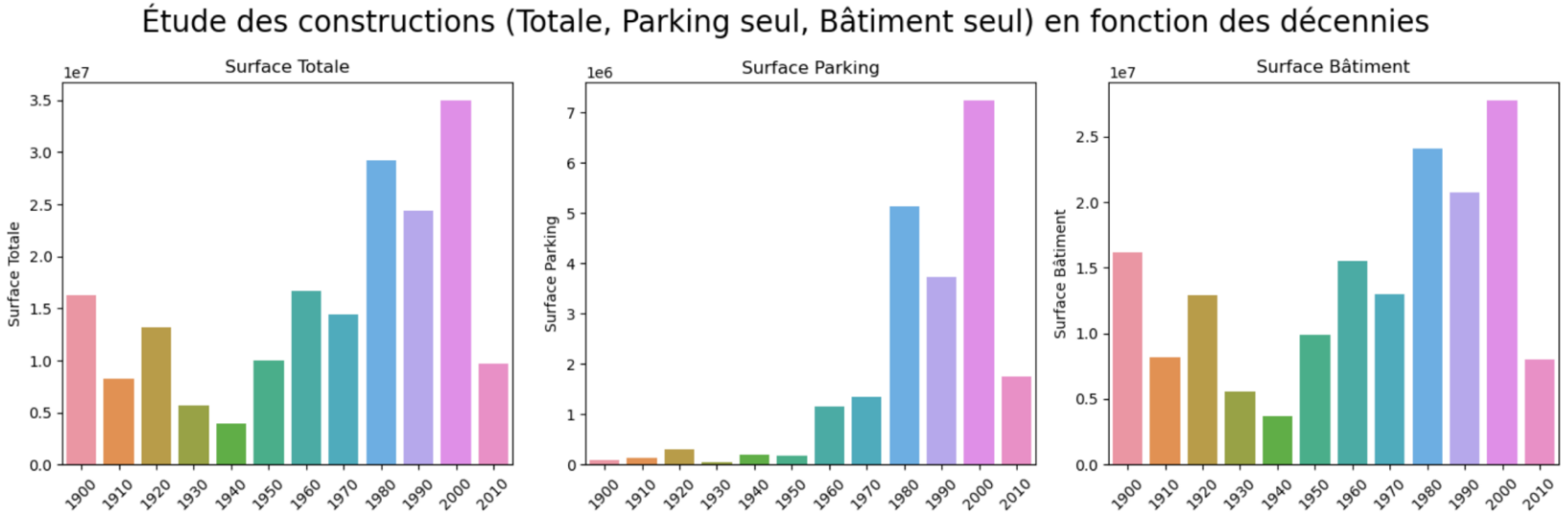




Observation:

- La construction des bâtiments ne semble pas évoluer au fil des décennies.
- La plupart des bâtiments les plus anciens sont principalement construits dans le centre-ville.

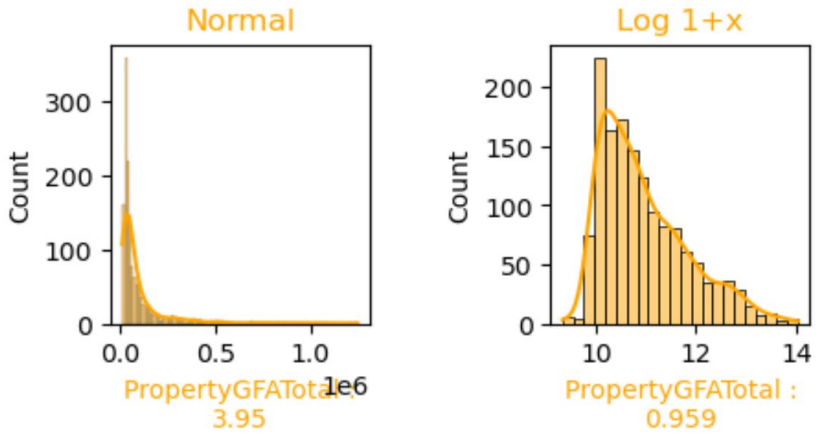
Type de surfaces des bâtiments

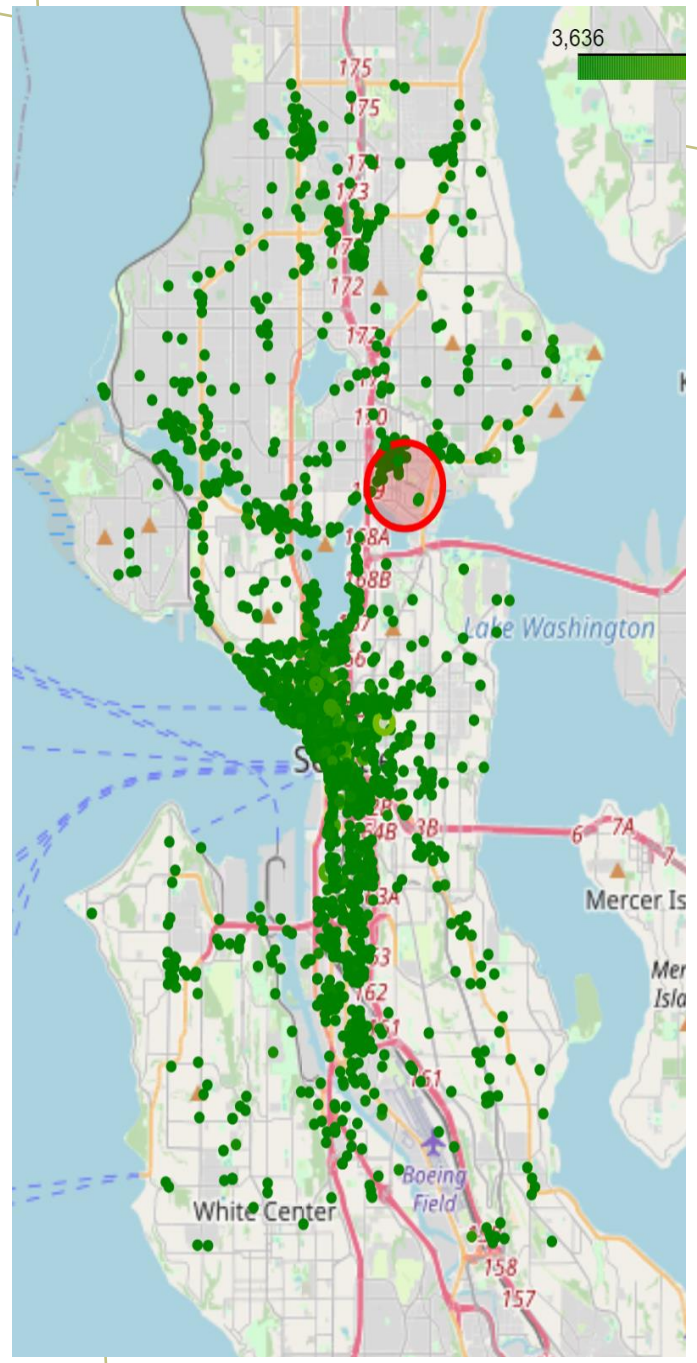


Observation:

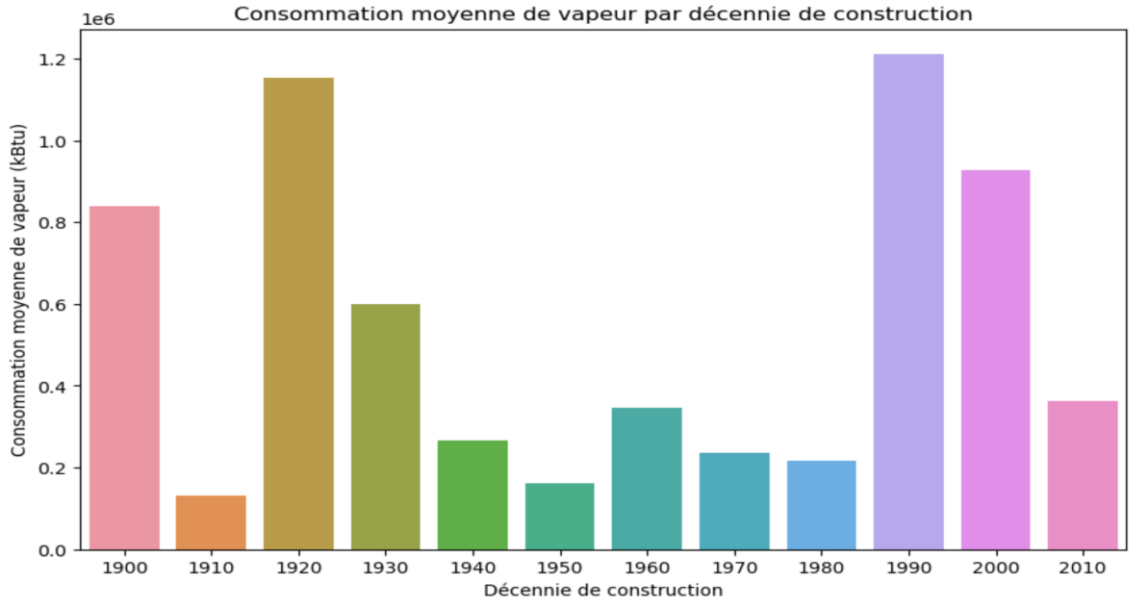
- Les variables les plus pertinentes pour la consommation d'énergie sont principalement celles liées aux surfaces totales des bâtiments.
- Une augmentation notable est observée au cours des 40 dernières années.
- D'importantes surfaces de bâtiments sont observées dans le centre-ville.

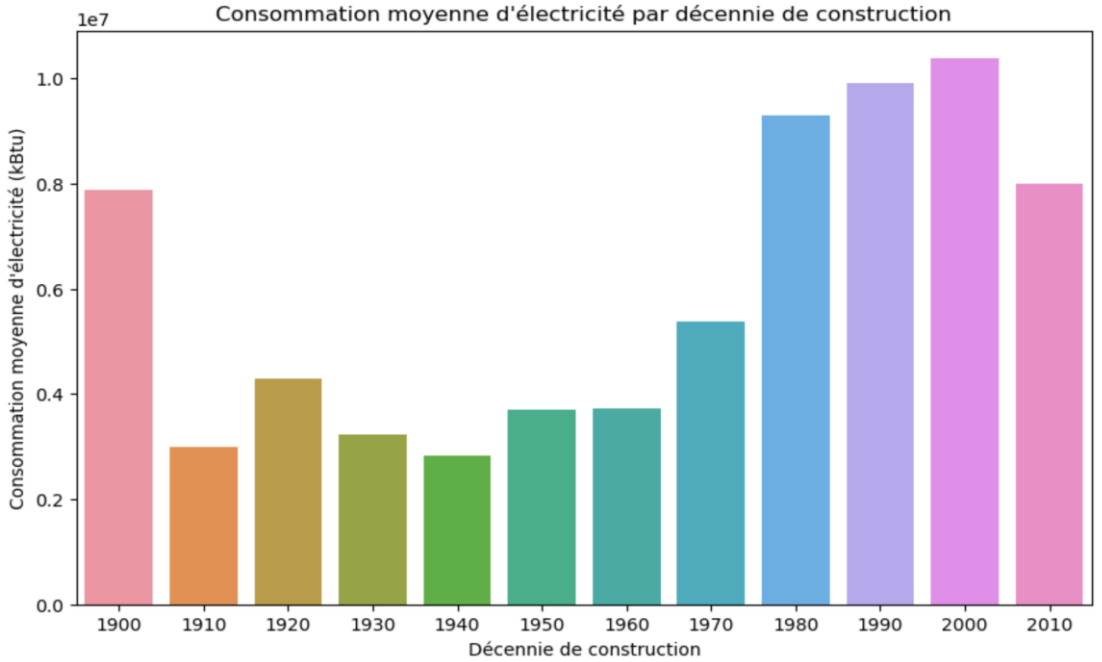
Transformation logarithmique de la surface globale

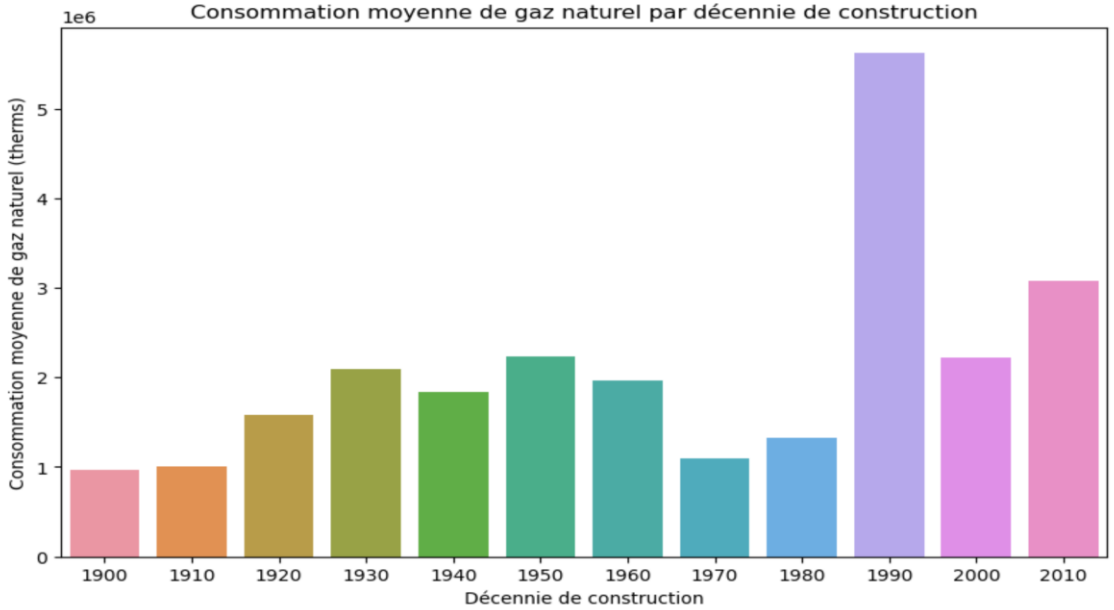




Variables de consommation énergétique SteamUse(kBtu), Electricity(kBtu), et NaturalGas(kBtu)







Observations:

- Il y a eu une forte hausse de consommation de **vapeur** dans les années 20, mais la tendance a été relativement stable depuis, à l'exception des années 1950.
- Nous observons une consommation d'**électricité** relativement élevée au cours des années 1900, suivie d'une utilisation stable au fil des années, à l'exception d'une augmentation notable entre 1980 et 2010.
- La consommation de **gaz naturel** a connu une tendance globalement stable avant 1980, à l'exception d'une augmentation marquée durant les années 1990.

Processus de transformation des variables qualitatives importantes

Transformation de deux variables qualitatives:

- **BuildingType**
- **PrimaryPropertyType**

Utilisation de **OneHotEncoder** pour transformer **BuildingType** et **PrimaryPropertyType** en une matrice de variables binaires.
Ce processus permet de convertir les variables qualitatives en une forme numérique adaptée aux algorithmes de machine learning.

BuildingType

```
['nonresidential',
 'nonresidential cos',
 'sps-district k-12',
 'campus',
 'nonresidential wa']
```

PrimaryPropertyType

```
['hotel',
 'other',
 'mixed use property',
 'university',
 'small- and mid-sized office',
 'self-storage facility',
 'k-12 school',
 'large office',
 'senior care community',
 'medical office',
 'retail store',
 'warehouse',
 'distribution center',
 'worship facility',
 'supermarket / grocery store',
 'laboratory',
 'refrigerated warehouse',
 'restaurant',
 'hospital']
```

Processus de transformation des variables quantitatives importantes

Transformation de trois variables quantitatives:

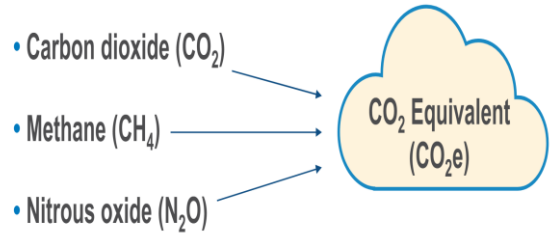
- **SteamUse(kBtu)**
- **Electricity(kBtu)**
- **NaturalGas(kBtu)**

Utilisation d'une fonction **transform_to_binary** qui convertit une valeur en 1 si elle est strictement supérieure à 0 (indiquant une consommation d'énergie) et en 0 sinon (indiquant aucune consommation d'énergie).
Nous appliquons la fonction aux colonnes **SteamUse(kBtu), Electricity(kBtu)**, et **NaturalGas(kBtu)** pour créer de nouvelles colonnes binaires correspondantes.
Ce processus permet de convertir les valeurs énergétiques en variables binaires, ce qui facilite leur utilisation dans des analyses et des modèles de machine learning.

Présentation des Targets

- **SiteEnergyUse(kBtu) :** fait référence la consommation totale d'énergie d'un site, mesurée en kilo-British thermal units (kBtu). Cela inclut toutes les formes d'énergie utilisées sur le site, telles que l'électricité, le gaz naturel, la vapeur, et autres, converties en une unité commune, le kBtu. Cette mesure permet d'évaluer et de comparer l'efficacité énergétique de différents sites ou bâtiments.

- **TotalGHGEmissions :** désigne les émissions totales de gaz à effet de serre (GES) d'un site, mesurées en termes de dioxyde de carbone équivalent (CO2e) dans la ville de Seattle. Cela inclut tous les types de GES produits par les activités du site, tels que le dioxyde de carbone (CO2), le méthane (CH4), et le protoxyde d'azote (N2O), parmi d'autres. La mesure en CO2e permet de comparer et d'évaluer l'impact environnemental des émissions de GES en utilisant une unité standardisée.



- **ENERGYSTARScore :** Le score énergétique actuel est complexe à calculer selon la méthode actuellement employée. Qui est une évaluation comparative de l'efficacité énergétique d'un bâtiment, notée sur une échelle de 1 à 100. Cette évaluation est fournie par le programme ENERGY STAR de l'Environmental Protection Agency (EPA) des États-Unis. Un score de 50 indique que le bâtiment a une performance énergétique médiane par rapport à des bâtiments similaires. Un score de 75 ou plus indique que le bâtiment se situe dans les 25% les plus performants et peut être éligible à la certification ENERGY STAR, signalant une efficacité énergétique supérieure.

3.Méthodologie de modélisation

- **Etape 1:** Utilisation de **LazyRegressor** pour explorer rapidement différents modèles de régression sans avoir à ajuster manuellement chaque algorithme.

```
100%|████████| 42/42 [00:17<00:00,  2.47it/s]
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.000764 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 558
[LightGBM] [Info] Number of data points in the train set: 1016, number of used features: 20
[LightGBM] [Info] Start training from score 14.840675
```

Model	Adjusted R-Squared	R-Squared	RMSE
GradientBoostingRegressor	0.61	0.64	0.71
ExtraTreesRegressor	0.59	0.61	0.73
SVR	0.58	0.61	0.73
RandomForestRegressor	0.56	0.59	0.75
NuSVR	0.56	0.59	0.75
LGBMRegressor	0.55	0.58	0.76
HistGradientBoostingRegressor	0.55	0.58	0.76
XGBRegressor	0.55	0.58	0.76
BaggingRegressor	0.53	0.56	0.77
HuberRegressor	0.50	0.54	0.79
LinearSVR	0.50	0.54	0.79
LinearRegression	0.50	0.53	0.80
TransformedTargetRegressor	0.50	0.53	0.80
RidgeCV	0.50	0.53	0.80
Ridge	0.50	0.53	0.80
LassoCV	0.50	0.53	0.80
ElasticNetCV	0.50	0.53	0.80
BayesianRidge	0.50	0.53	0.80
SGDRegressor	0.49	0.53	0.80
Lars	0.49	0.52	0.80
LassoLarsCV	0.49	0.52	0.80
LarsCV	0.49	0.52	0.81
LassoLarsIC	0.49	0.52	0.81
KNeighborsRegressor	0.48	0.52	0.81
AdaBoostRegressor	0.48	0.52	0.81
PoissonRegressor	0.47	0.51	0.82
RANSACRegressor	0.46	0.50	0.82
OrthogonalMatchingPursuitCV	0.35	0.40	0.90
TweedieRegressor	0.35	0.40	0.91
GammaRegressor	0.35	0.40	0.91
OrthogonalMatchingPursuit	0.33	0.38	0.92
ExtraTreeRegressor	0.29	0.34	0.95
DecisionTreeRegressor	0.25	0.30	0.98
ElasticNet	0.04	0.11	1.10
QuantileRegressor	-0.07	-0.00	1.17
DummyRegressor	-0.08	-0.00	1.17
Lasso	-0.08	-0.00	1.17
LassoLars	-0.08	-0.00	1.17
PassiveAggressiveRegressor	-0.29	-0.20	1.28
MLPRegressor	-1.15	-1.00	1.65
KernelRidge	-171.82	-159.91	14.81
GaussianProcessRegressor	-31121.05	-28974.70	198.72

- **Etape 2 :** Modélisation avec des modèles naïfs, tels que Random et Dummy.
- **Etape 3 :** Modélisation avec divers modèles de régression linéaire avec régularisation (Lasso, Ridge et Elastic Net).
- **Etape 4 :** Modélisation avec des modèles de régression plus complexes (KNN et arbre de décision)
- **Etape 5 :** Modélisation avec des modèles de machine learning plus avancés (Random Forest, XGBoost et SVM).
- **Etape 6 :** Optimisation des hyperparamètres des modèles complexes (Max_depth, n_estimators, n_neighbors, weights).
- **Etape 7 :** Sélection des trois modèles ayant le meilleur coefficient de détermination **R^2** (Random Forest, XGBoost et SVM).
- **Etape 8 :** Réexécution des trois modèles sélectionnés en incluant la variable ENERGYSTARScore.

Trois premiers modèles

	RMSE	R^2
Support Vector Machine (SVM)	0.70	0.64
XGBoost	0.70	0.64
Random Forest	0.74	0.60
K Neighbors	0.79	0.55
Ridge	0.80	0.53
Lasso	0.80	0.53
ElasticNet	0.80	0.53
Linear Regression	0.80	0.53
Decision Tree	0.83	0.49
Dummy	1.17	-0.00
Random Test	2.79	-4.73
Random Train	2.81	-4.86

4.Résultats pour la Target N°1 : SiteEnergyUse(kBtu)

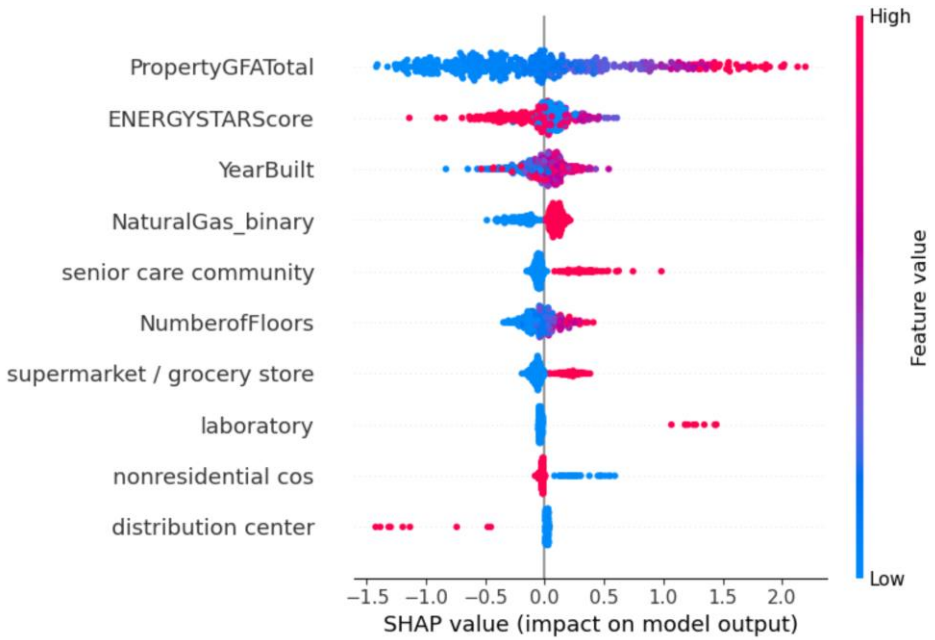
Sans la variable **ENERGYSTARScore**

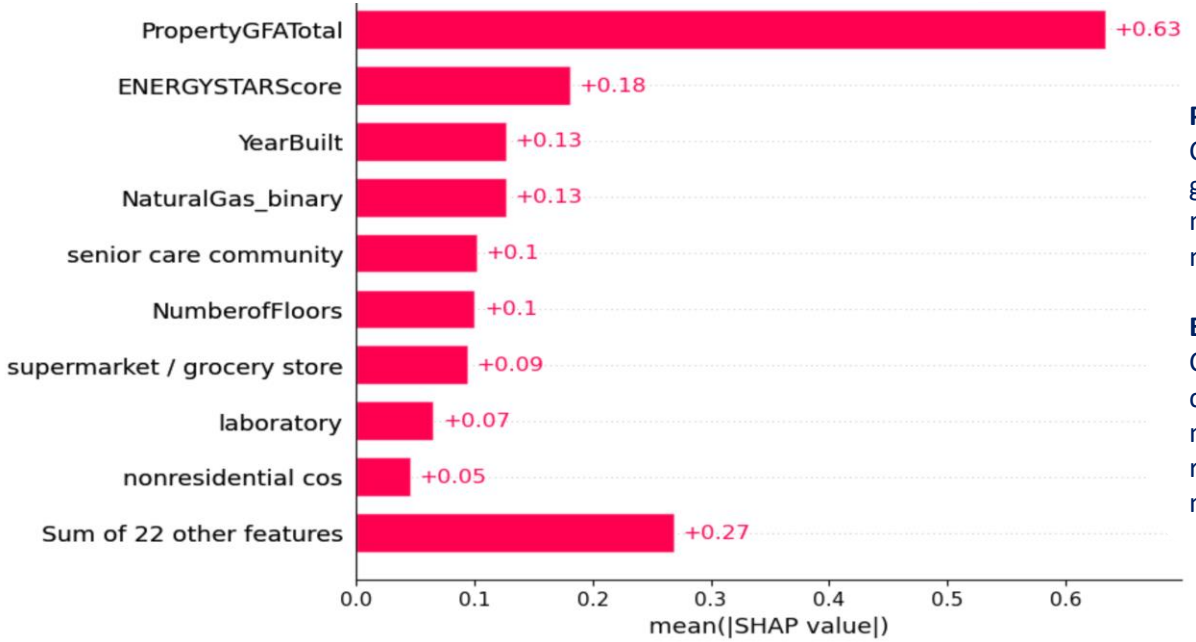
	RMSE	R^2
Support Vector Machine (SVM)	0.70	0.64
XGBoost	0.70	0.64
Random Forest	0.74	0.60

Avec la variable **ENERGYSTARScore**

	RMSE	R^2
Support Vector Machine (SVM)	0.67	0.67
XGBoost	0.63	0.71
Random Forest	0.69	0.65

Interprétation théorique des modèles avec SHAP





PropertyGFATotal (+0.63) : Cette variable a la plus grande contribution positive moyenne aux prédictions du modèle.

Energystarscore (+0.18) : Cette variable a une contribution positive moyenne plus faible par rapport à PropertyGFATotal, mais elle reste importante

5.Résultats pour la Target N°2 : TotalGHGEmissions

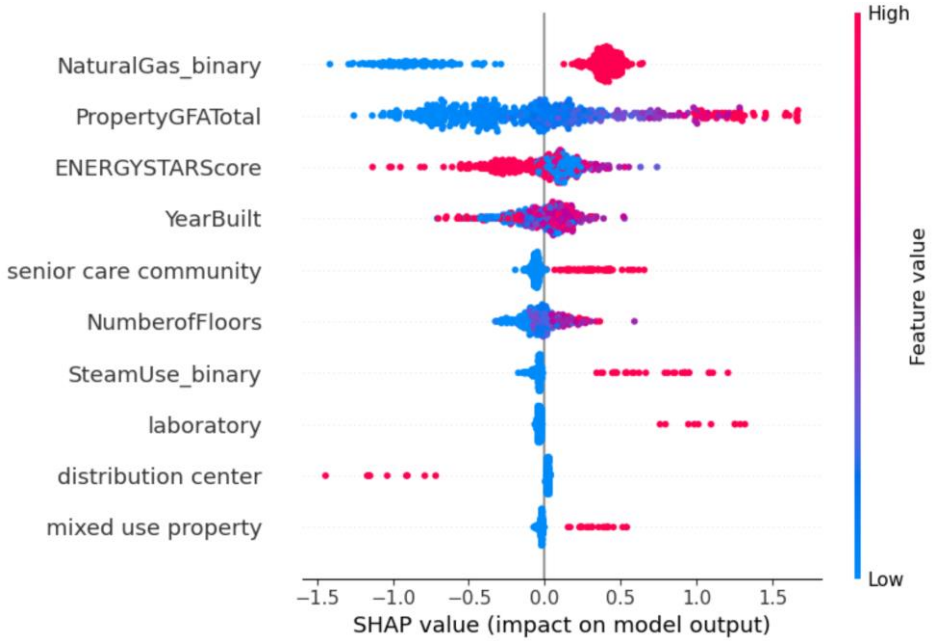
Sans la variable **ENERGYSTARScore**

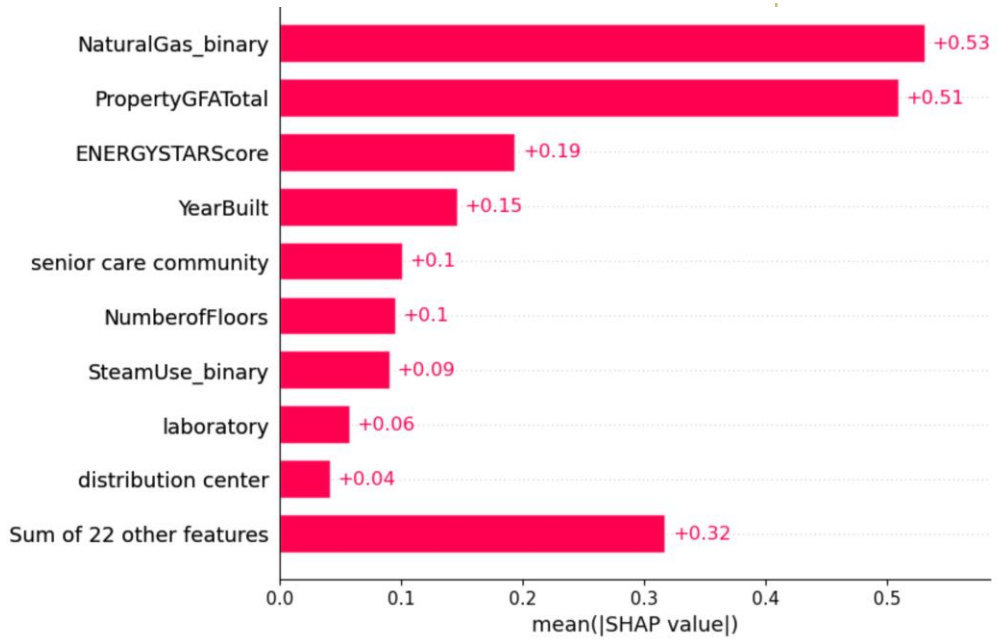
	RMSE	R²
Support Vector Machine (SVM)	0.79	0.61
XGBoost	0.80	0.61
Random Forest	0.80	0.61

Avec la variable **ENERGYSTARScore**

	RMSE	R²
Support Vector Machine (SVM)	0.77	0.64
XGBoost	0.75	0.65
Random Forest	0.79	0.62

Interprétation théorique des modèles avec SHAP





NaturalGAs (+0.53) : Cette variable a la plus grande contribution positive moyenne aux prédictions du modèle.

PropertyGFATotal (+0.51) : Cette variable a une contribution positive moyenne légèrement faible par rapport à NaturalGAs, mais elle reste importante.

6.Conclusions

- L'intégration de la variable **ENERGYSTARScore** dans nos modèles de machine learning a significativement amélioré la performance de nos prédictions.

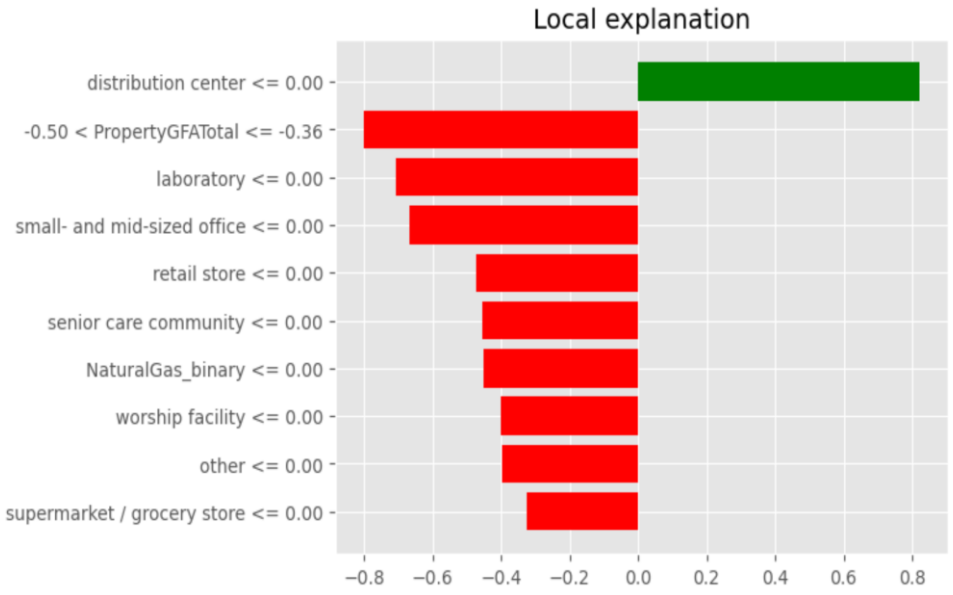
- Les modèles ont tous montré des réductions notables de l'erreur quadratique moyenne (RMSE) et des augmentations du coefficient de détermination (R^2)

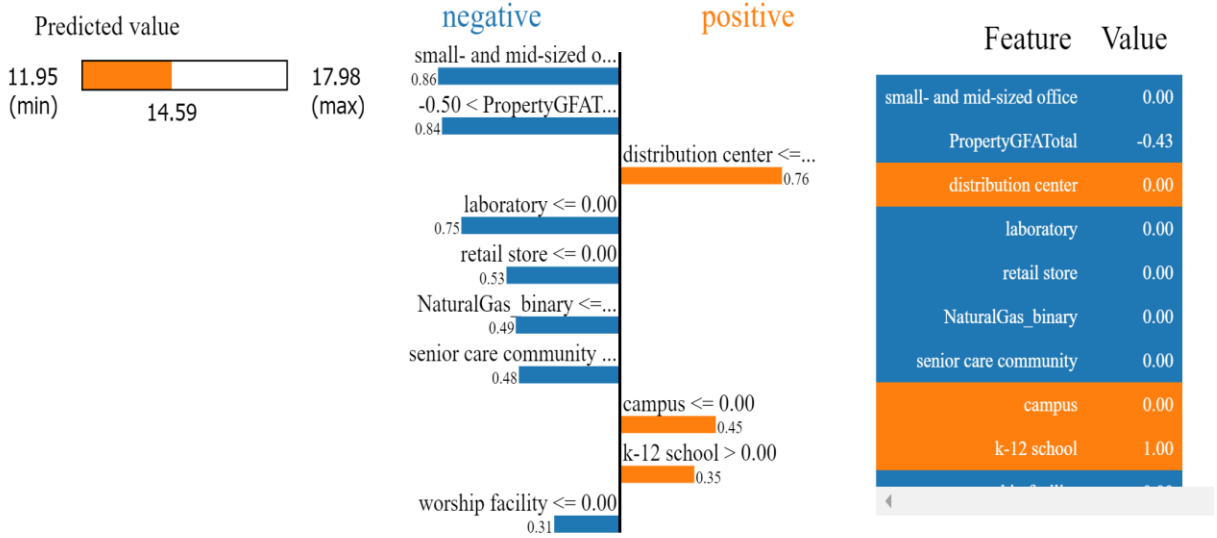
- Cette variable devrait donc être considérée comme une composante essentielle dans les futurs modèles de prédiction pour optimiser la précision des résultats.

- Cependant, l'ENERGY STAR nécessite des calculs préalables coûteux et n'est pas disponible pour tous les permis d'exploitation, ce qui limite son intérêt étant donné l'amélioration relativement faible observée.

ANNEXE: Avec LIME de manière plus locale

SiteEnergyUse(kBtu)





TotalGHGEmissions