Seattle

Anticipez les besoins en consommation de bâtiments

Synthèse

Sommaire

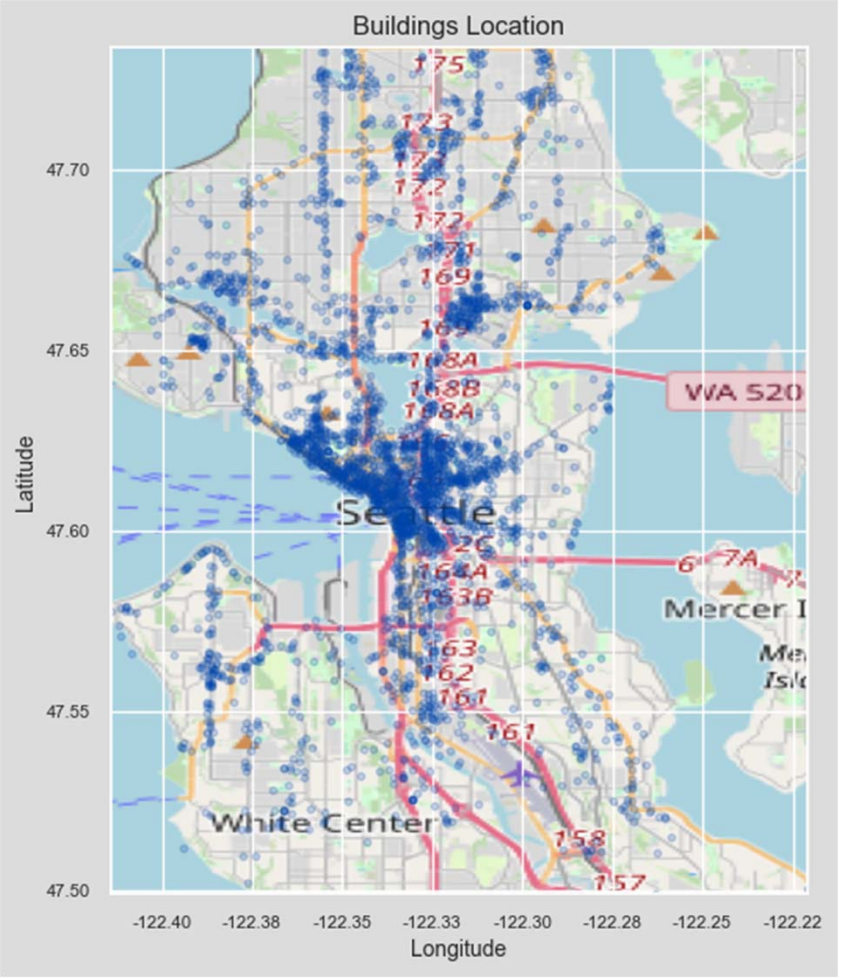
Contexte & Objectifs

PROJET

Sur la base des relevés de l'année 2016, prédire la consommation énergétique totale et les émissions de gaz à effet de serre des immeubles à usage non-résidentiel de la ville de Seattle, et évaluer l'utilité de l'Energy Star Score dans ces prédictions.

DÉMARCHE

Après une analyse exploratoire des données, utiliser plusieurs familles de modèles et les évaluer selon plusieurs critères afin d'en déduire les meilleures prédictions et le meilleur jeu de données en entrée.



1 – Analyse Exploratoire des Données

1.1 – Nettoyage & Feature Engineering

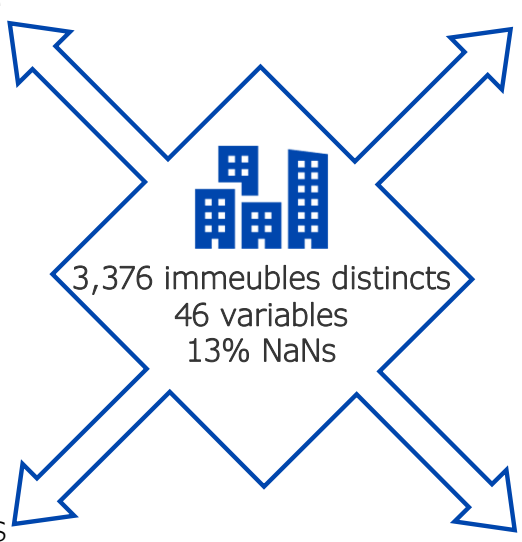
JEU DE DONNEES INITIAL* :

Données géographiques :
- Latitude & longitude
- Ville / État
- Code postal etc…

➢ 9 variables

Données énergétiques & environnementales :
- Consommation
- Types d'énergie
- Émissions de gaz à effet de serre etc…

➢ 15 variables



Données structurelles, réglementaires & fiscales :
- Conformité des rapports
- Source des données
- Age des immeubles
- Nombre de bâtiments etc…

➢ 12 variables

Données d'utilisation principale & secondaires :
- Habitation
- Commerces
- Parkings
- Superficies etc…

➢ 10 variables

Seattle

* Source : https://data.seattle.gov/Built-Environment/2016-Building-Energy-Benchmarking/2bpz-gwpy/about_data

1 – Analyse Exploratoire des Données

1.1 – Nettoyage

SUPPRESSIONS – 87 immeubles sur 1,470 non-résidentiels

- Données non pertinentes:
  - immeubles à usage résidentiel (d'apres BuildingType, PrimaryPropertyType, LargestPropertyUseType, SecondLargestPropertyUseType et ThirdLargestPropertyUseType)
  - données purement administratives (Zip code, information fiscale, commentaires etc…)
  - données non-differenciées (1 seule valeur) comme State ou City

- Données non qualitatives:
  - immeubles utilisant des DefaultData
  - immeubles pour lesquels la consommation totale des 3 types d'énergie dévie de plus de 5% en valeur absolue du total rapporté
  - immeubles pour lesquels le LargestPropertyUseTypeGFA est manquant
  - immeubles pour lesquels le NumberofBuildings est nul

- Outliers

CORRECTION

- des labels dans la colonne Neighborhood

1 – Analyse Exploratoire des Données

1.1 – Feature Engineering

- Remplacement de l'année de construction par l'âge de l'immeuble

- Conversion des usages primaires, secondaires & tertiaires en colonnes de % de la surface totale & groupement des valeurs les moins fréquentes (<1%) en « Other_aggregated_usage »

- Discrétisation des variables-cibles en 3 quantiles (low / medium / high) en vue du stratified k-fold

- Addition d'une colonne binaire pour l'utilisation de chaque type d'énergie (électricité, gaz & vapeur)

- Ajout d'une colonne par type de consommation (électricité, gaz & vapeur) en % du total

1 – Analyse Exploratoire des Données

1.2 – Sélection des Features & Variables-Cibles

JEU DE DONNÉES NETTOYÉ :

35 FEATURES NUMÉRIQUES
+ ENERGYSTARScore (33% NaNs)

1,383 IMMEUBLES
Non-Résidentiels

2 VARIABLES-CIBLES :

Consommation d'énergie
SiteEnergyUseWN(kBtu)
Unité : milliers de British Thermal Units*

Émissions de gaz à effet de serre
TotalGHGEmissions
Unité : tonnes métriques d'équivalent CO_2

Seattle

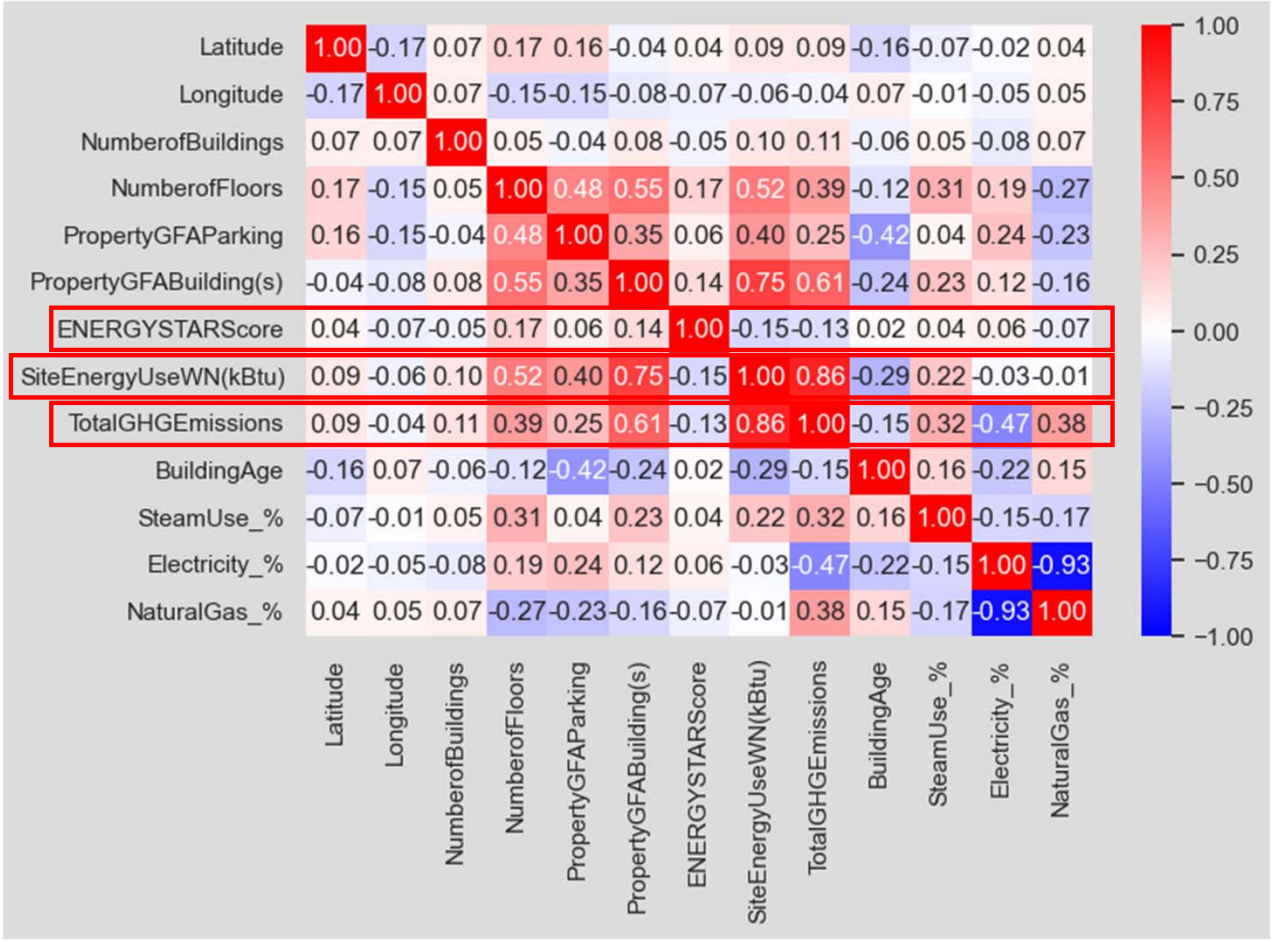
* Defined as the amount of heat needed to raise 1 lb of water at maximum density through one degree Fahrenheit, equivalent to 1.055 × 103 joules or 0.2931 W·h (watt hours). Source : https://www.eia.gov/energyexplained/units-and-calculators/british-thermal-units.php

1 – Analyse Exploratoire des Données

1.3 – Statistiques Descriptives

CORRÉLATIONS:

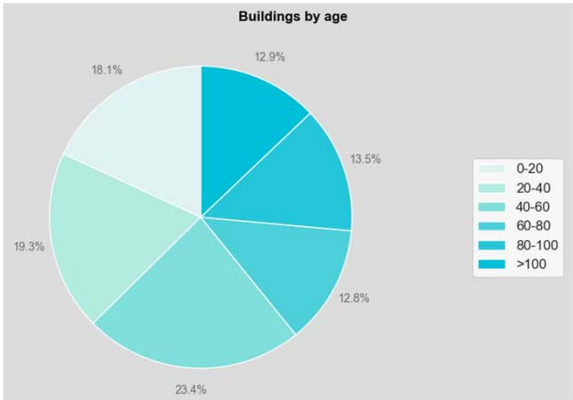
- Skew
- Variables non-gaussiennes
  - Coefficient de Spearman

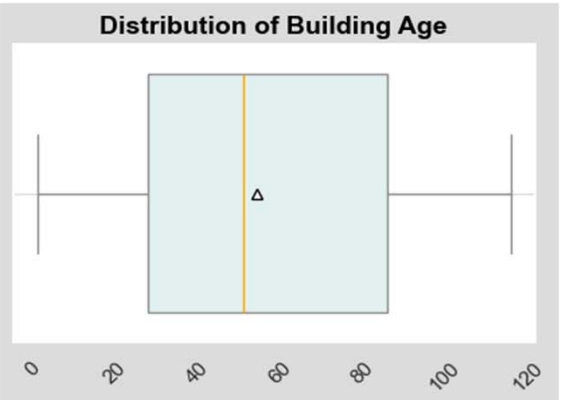


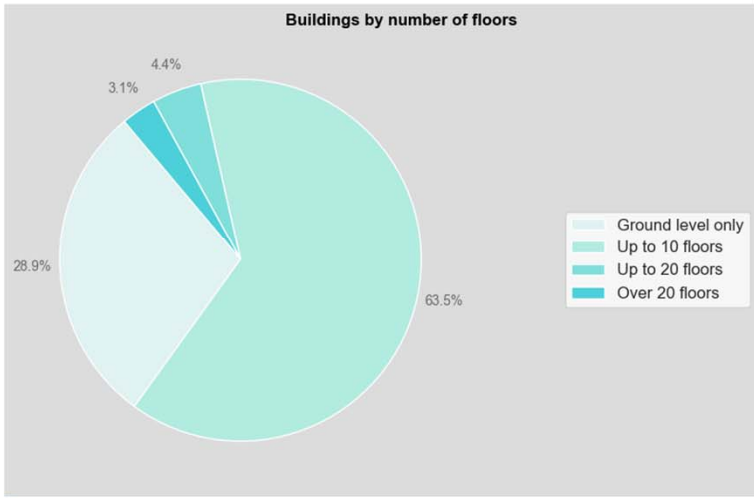
1 – Analyse Exploratoire des Données

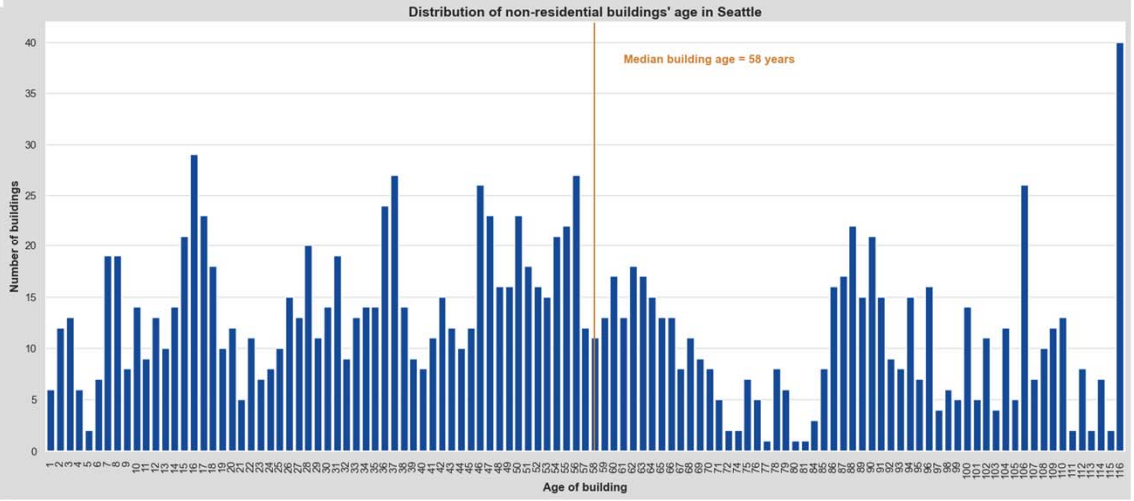
1.3 – Statistiques Descriptives

Caractéristiques du Parc Immobilier Non-Résidentiel de Seattle
–
Âge & Hauteur des Immeubles





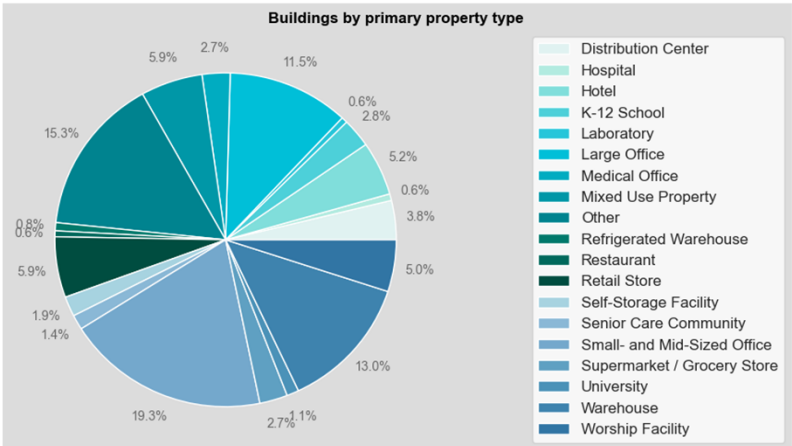


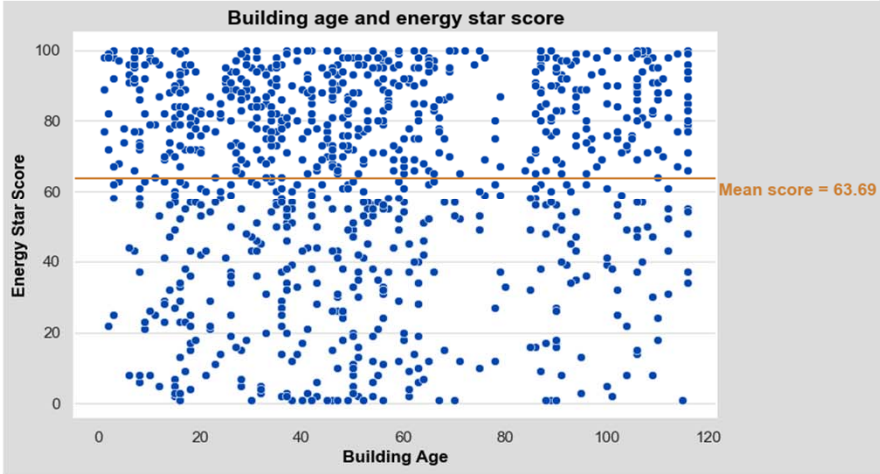


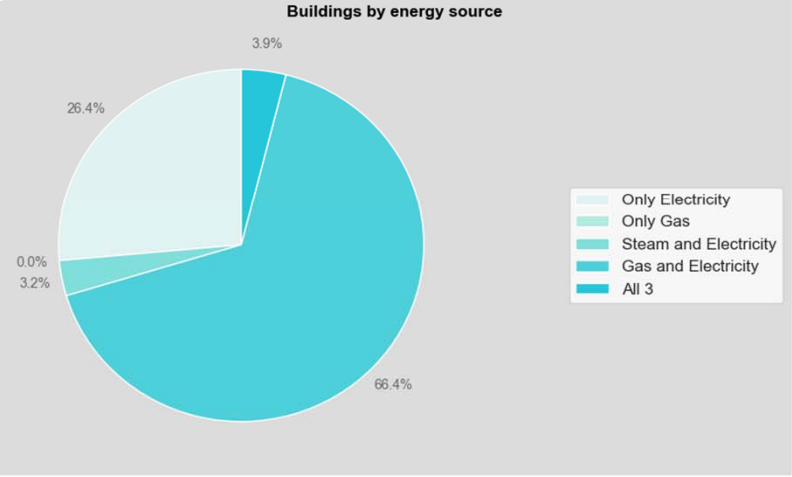
1 – Analyse Exploratoire des Données

1.3 – Statistiques Descriptives

Caractéristiques du Parc Immobilier Non-Résidentiel de Seattle – Types d'Usage & d'Énergie



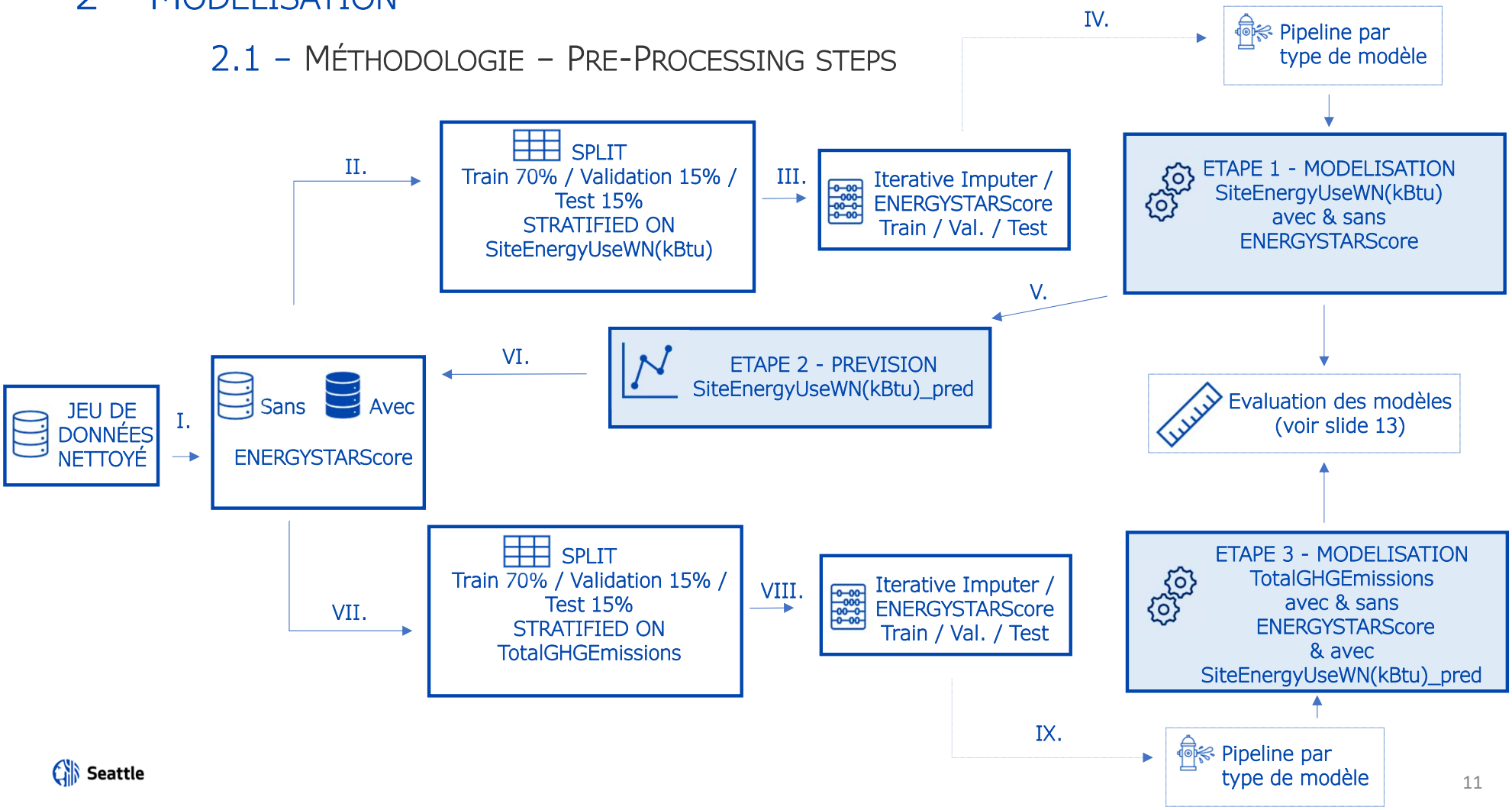




Seattle

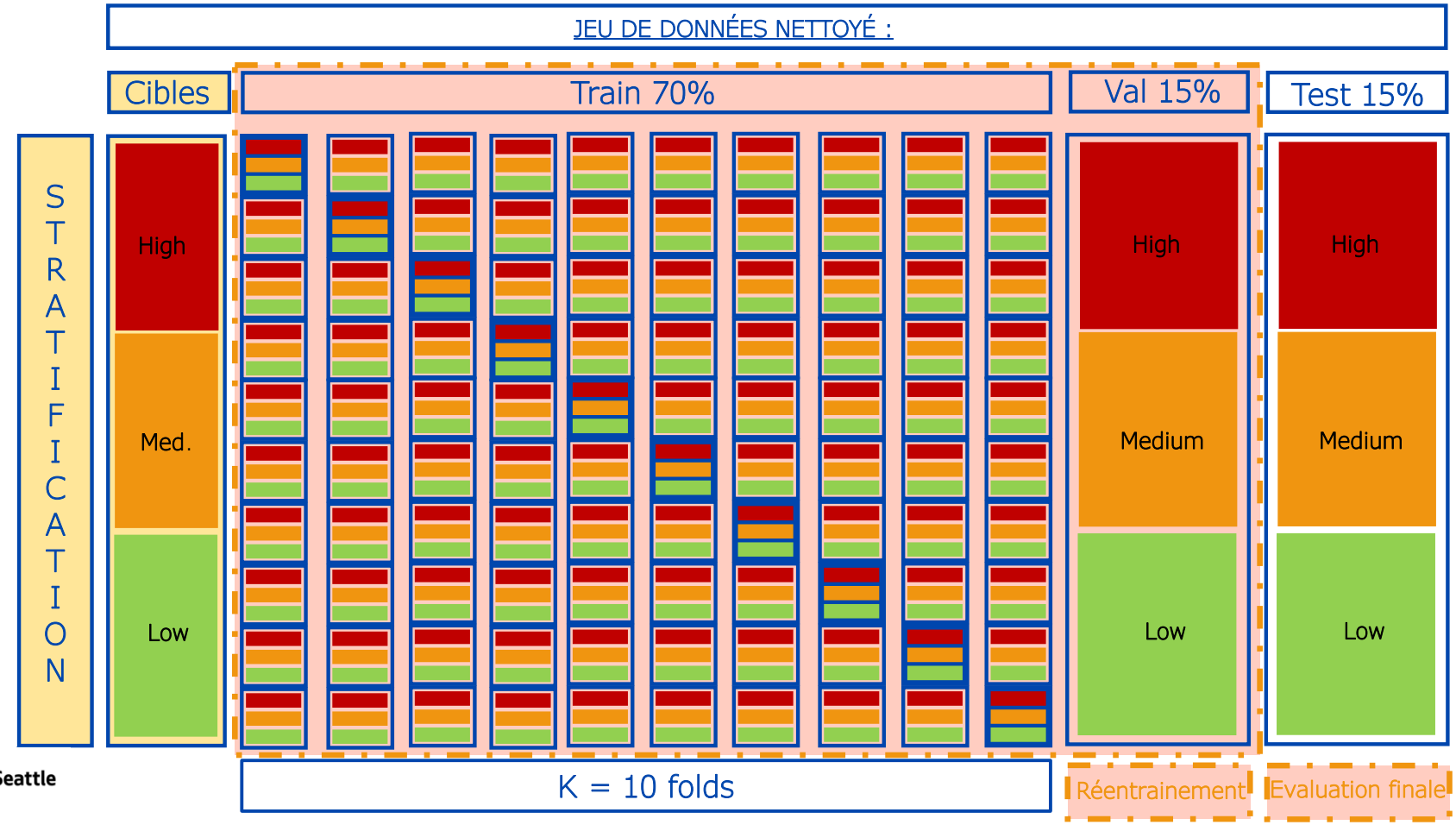
2 – MODÉLISATION

2.1 – MÉTHODOLOGIE – PRE-PROCESSING STEPS



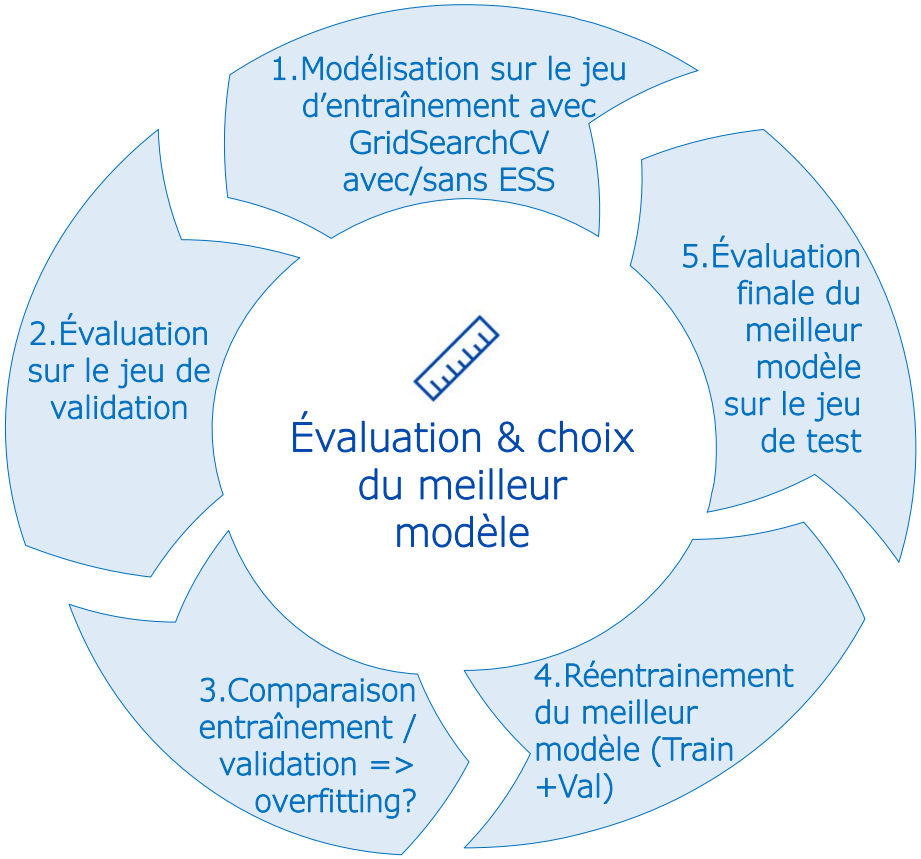
2 – Modélisation

2.1 – Méthodologie – Stratified K-Fold



2 – Modélisation

2.1 – Évaluation des modèles



Critères d'Évaluation :

- R^2
- MAE
- Fit time

2 – Modélisation

2.2 – Choix des Modèles & Features

❖ Modèles testés :

- Baseline = régression linéaire
- ElasticNet
- K-Nearest Neighbors
- Support Vector Regression
- Random Forest Regression
- Gradient Boosting Regression

❖ Combinaisons de features testées :

- Latitude & Longitude / neighborhood
- Catégories d'énergie utilisées / % du total
- ENERGYSTARScore exclus ou inclus
- Et, dans le cas des Émissions de Gas à effet de serre, prédiction de la consommation totale d'énergie exclue ou incluse

2 – Modélisation

2.2 – Optimisation des hyper-paramètres

	Valeur-cible	
Modèle	**SiteEnergyUseWN(kBtu)**	**TotalGHGEmissions**
Linear Regression	fit_intercept: [True, False], model__copy_X : [True]	fit_intercept: [True, False], model__copy_X : [True]
ElasticNet	alpha : [0.001, 0.01, 0.1, 1, 10], random_state: [rs], model__copy_X : [True]	alpha : [0.01, 0.1, 1], random_state: [rs], model__copy_X : [True], selection : ['random'], tol : [0.001], l1_ratio : [0.1, 0.2, 0.5], max_iter': [2000]
KNN	n_neighbors: [5, 10, 15], weights : ['uniform', 'distance']	n_neighbors: [10, 15, 20], weights : ['distance'], algorithm : ['ball_tree', 'kd_tree'], leaf_size : [15, 30, 45], p : [1, 2], n_jobs : [-1]
SVR	kernel : ['poly'], degree : [2,3], gamma : [1, 10], C : [0.1, 1], epsilon : [0.1, 0.5]	kernel : ['rbf'], gamma : [0.01, 0.1, 1], C : [0.01, 0.1, 1, 10, 100, 1000], epsilon : [0.05, 0.01, 0.1], max_iter: [2000], tol' : [0.001]
Random Forest	n_estimators: [100, 200], random_state : [rs], max_depth : [None]	n_estimators: [100, 200], random_state' : [rs], max_depth : [5, 10], bootstrap': [True, False], min_samples_leaf:[2, 3], min_samples_split':[2, 5], max_features : ['sqrt', 'log2'], max_leaf_nodes': [20, 50], n_jobs': [-1]
Gradient Boosting	learning_rate : [0.5, 0.25, 0.1], n_estimators' : [100, 200], random_state' : [rs]	learning_rate : [0.5, 0.25, 0.1], n_estimators : [200, 400], random_state' : [rs], loss : ['squared_error'], max_features : ['sqrt', 'log2'], max_depth': [5, 10], subsample': [0.5, 0.75], min_samples_split': [2, 5]

2 – Modélisation

2.4 – Évaluation sur les Jeux d'Entraînement & de Validation

SiteEnergyUseWN(kBtu)

❖ Sans ENERGYSTARScore :

Model	Train R2	Train fit time	Val R2	Val MAE
Linear Regression	0.1572	5.12	0.2297	8,343,228.5918
ElasticNet	0.2876	6.49	0.2214	6,682,449.4851
KNN	0.2846	10.84	0.5081	4,764,916.8436
SVR	0.3919	29.85	0.1994	5,558,838.0088
Random Forest Regression	0.5845	37.22	0.7721	3,686,620.3270
Gradient Boosting Regression	0.5339	32.62	0.7380	4,054,347.4131

Paramètres :

➢ {'max_depth': None, 'n_estimators': 200, 'random_state': 42}

❖ Avec ENERGYSTARScore :

Model	Train R2	Train fit time	Val R2	Val MAE
Linear Regression	0.1673	2.21	0.2424	8,283,946.3256
ElasticNet	0.2944	5.54	0.2296	6,636,985.4928
KNN	0.2783	6.86	0.4584	4,748,229.2635
SVR	0.3941	28.00	0.0651	5,809,382.7329
Random Forest Regression	0.6002	40.65	0.7993	3,443,620.8714
Gradient Boosting Regression	0.5751	35.07	0.7735	3,590,709.1203

➢ {'max_depth': None, 'n_estimators': 200, 'random_state': 42}

2 – Modélisation

2.4 – Évaluation sur les Jeux d'Entraînement & de Validation GHGEmissions

❖ Sans ENERGYSTARScore :

Model	Train R2	Train fit time	Val R2	Val MAE
Linear Regression	- 0.6320	2.46	- 0.1465	183.6420
ElasticNet	0.0416	9.90	0.1768	142.2493
KNN	0.0980	47.32	0.3224	88.8791
SVR	0.3290	88.00	0.4430	80.9461
Random Forest Regression	0.3269	293.99	0.6732	80.1427
Gradient Boosting Regression	0.3307	390.90	0.6113	88.7281

❖ Avec ENERGYSTARScore :

Model	Train R2	Train fit time	Val R2	Val MAE
Linear Regression	- 0.6320	2.16	- 0.1465	183.6420
ElasticNet	0.0416	9.05	0.1768	142.2502
KNN	0.2285	46.66	0.3161	88.9497
SVR	0.3337	91.51	0.4301	82.1739
Random Forest Regression	0.3615	293.76	0.6579	80.0998
Gradient Boosting Regression	0.3047	418.69	0.5850	84.3626

Paramètres :

➢ {'bootstrap': True,
'max_depth': 10,
'max_features': 'sqrt',
'max_leaf_nodes': 50,
'min_samples_leaf': 2,
'min_samples_split': 2,
'n_estimators': 200,
'n_jobs': -1,
'random_state': 42}

➢ {'bootstrap': True,
'max_depth': 10,
'max_features': 'sqrt',
'max_leaf_nodes': 50,
'min_samples_leaf': 2,
'min_samples_split': 2,
'n_estimators': 200,
'n_jobs': -1,
'random_state': 42}

3 – Évaluation du Modèle Final

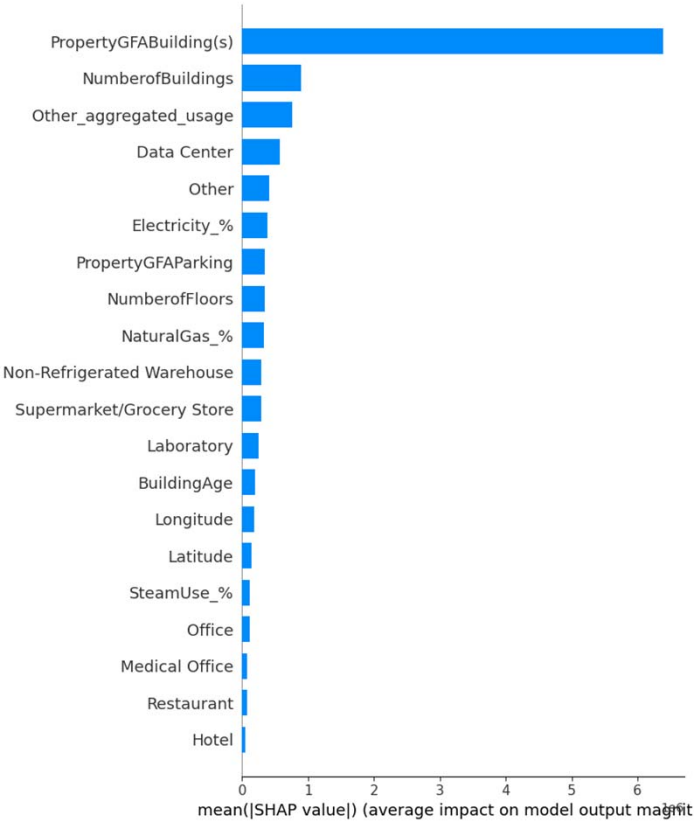
3.1 – Sur le jeu de Test

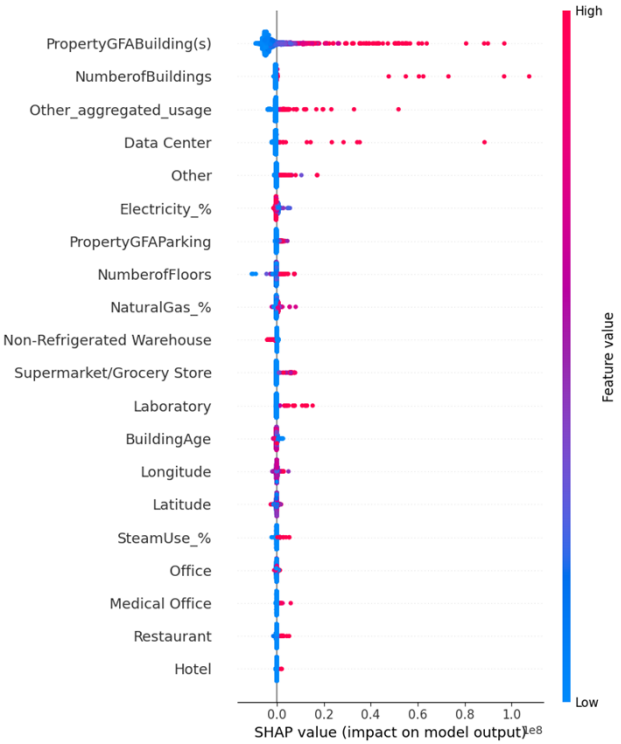
❖ Résultats sur le jeu de Test après réentrainement sur jeu (Train + Validation) :

Modele	Variable-Cible	Features	Test R2	Test MAE
Random Forest Regression	SiteEnergyUseWN(kBtu)	Sans ENERGYSTARScore	0.6890	3,084,918.7219
		Avec ENERGYSTARScore	0.6992	2,934,380.6128
	GHGEmissions	Sans ENERGYSTARScore	0.6166	80.6622
		Avec ENERGYSTARScore	0.5641	85.8598

3 – Évaluation du Modèle Final

3.2 – Feature Importance Globale SiteEnergyUseWN(kBtu)

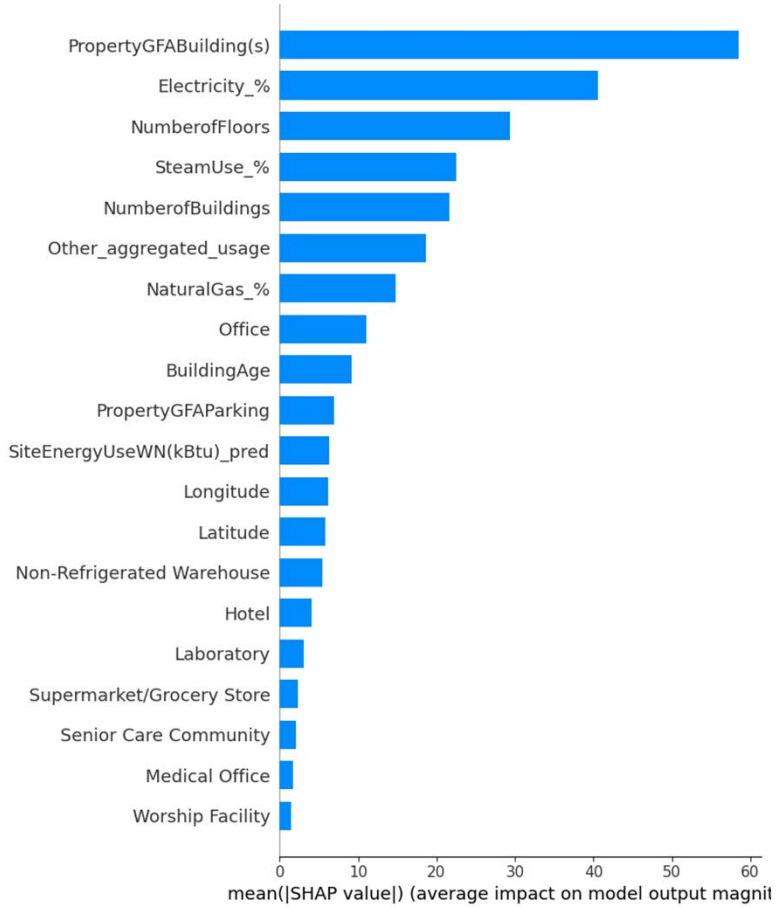


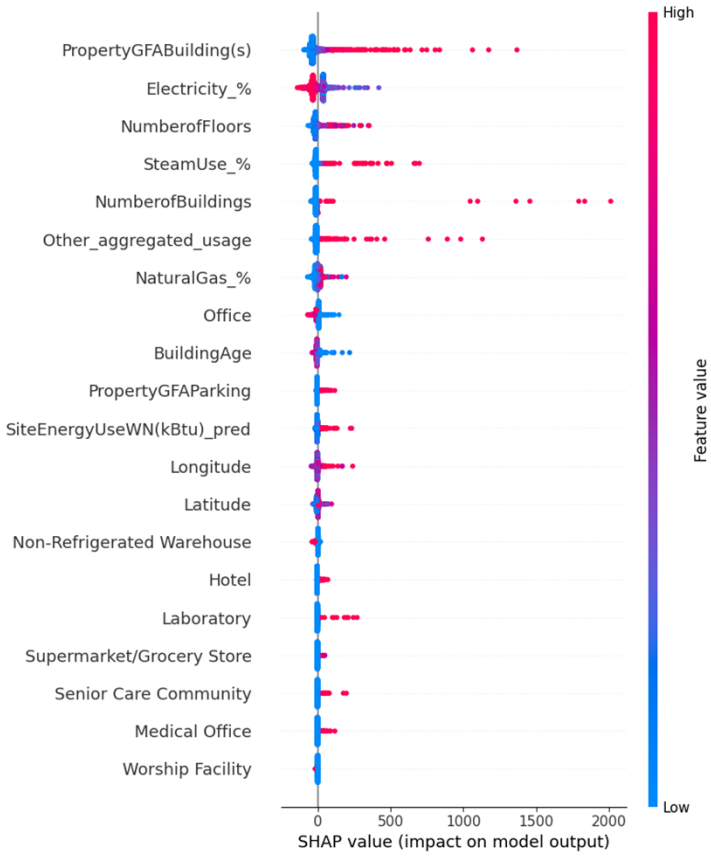


Seattle

3 – Évaluation du Modèle Final

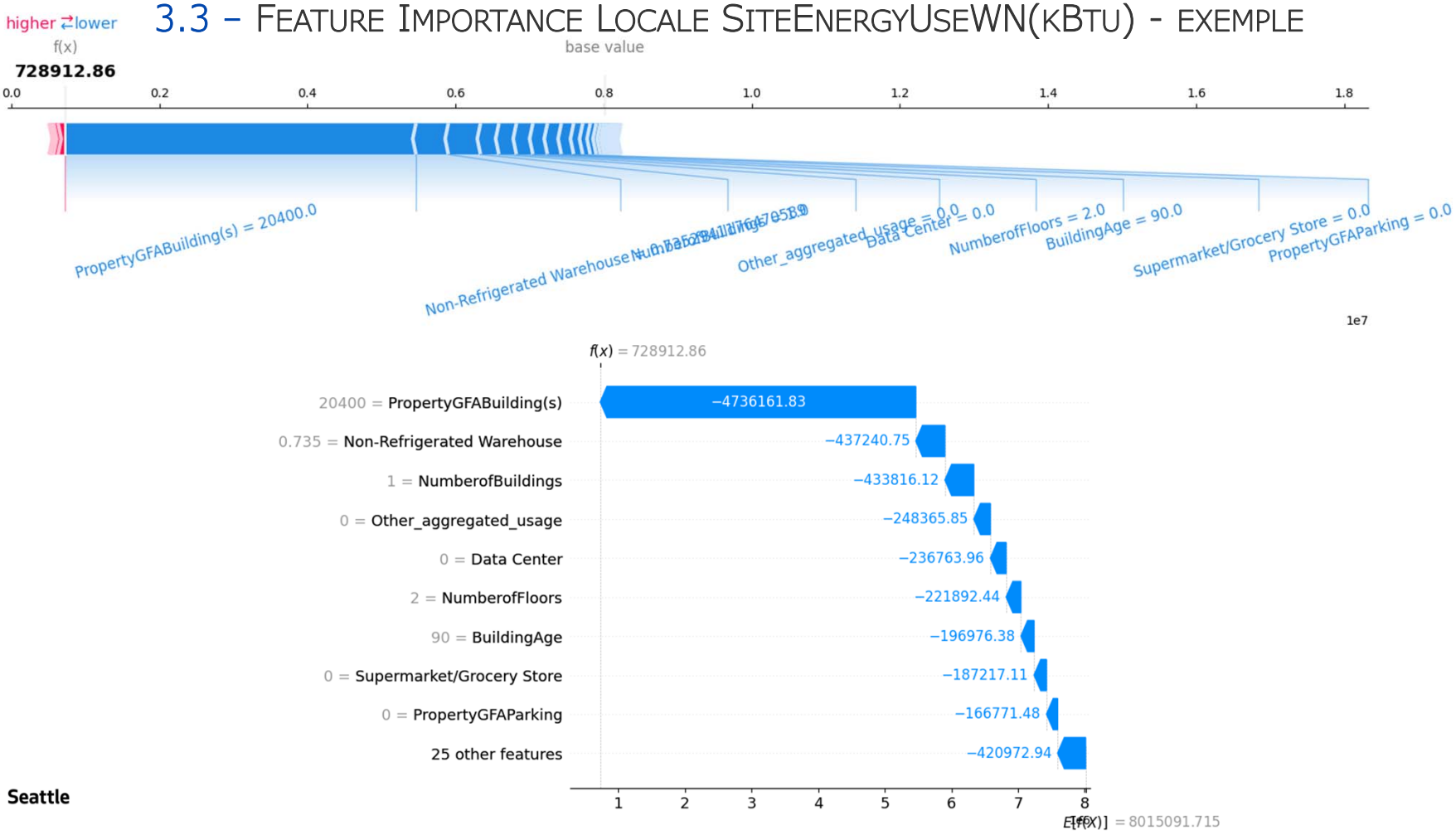
3.2 – Feature Importance Globale GHGEmissions





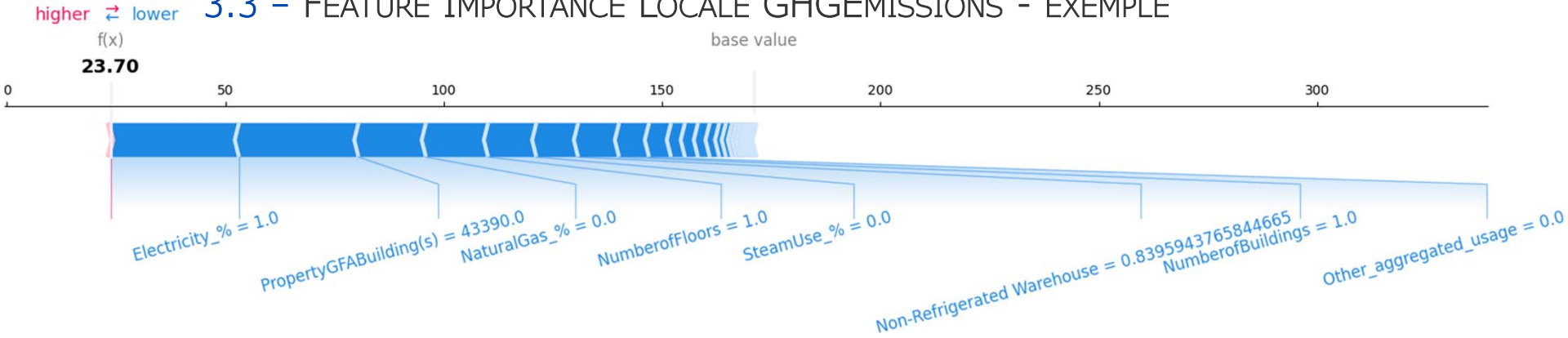
3 – Évaluation du Modèle Final

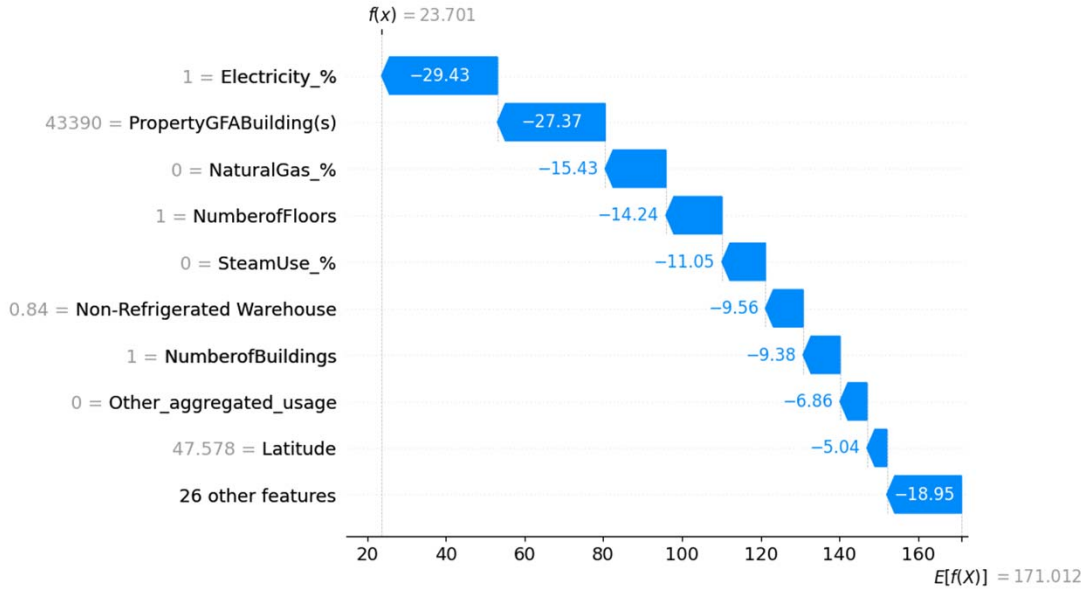
3.3 – Feature Importance Locale SiteEnergyUseWN(kBtu) - exemple



3 – Évaluation du Modèle Final

3.3 – Feature Importance Locale GHGEmissions - exemple





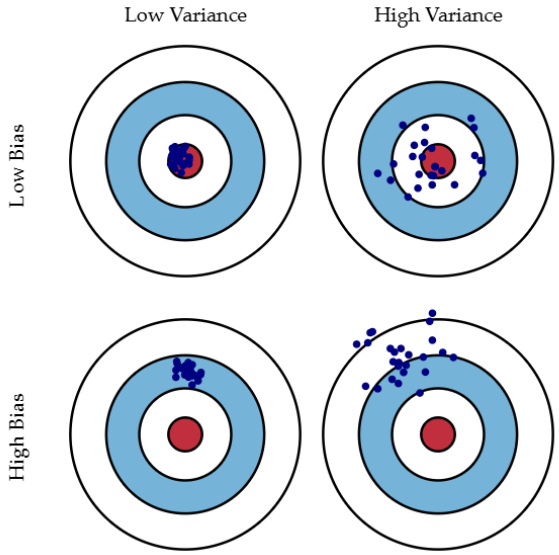
Seattle

Conclusion & perspectives

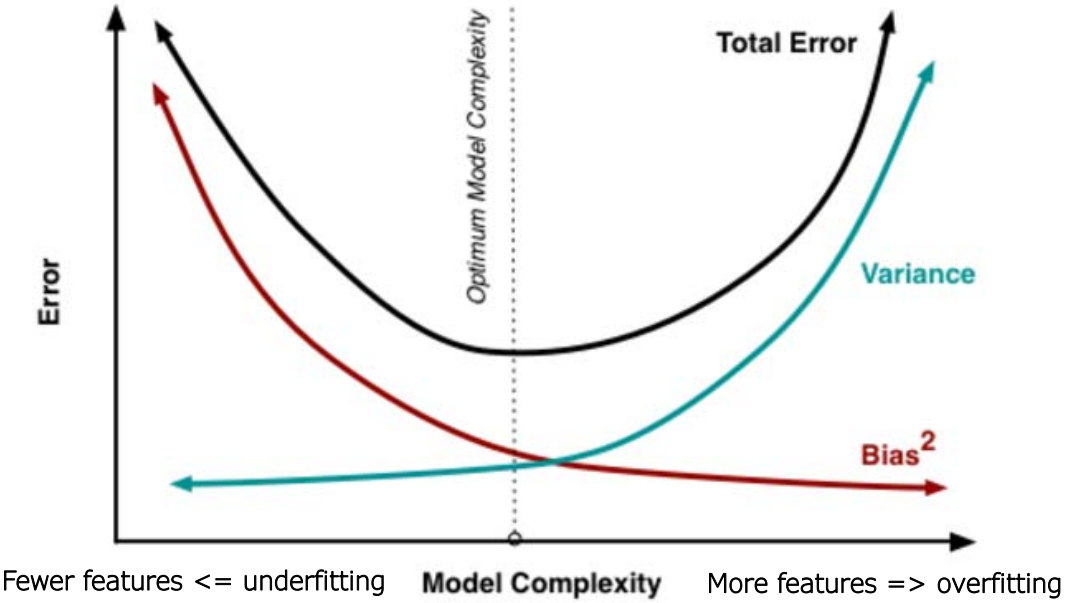
- Ajout de données supplémentaires en entrée ne conduit pas nécessairement à une amélioration des résultats d'un modèle
  - Minimiser l'erreur totale = optimiser le compromis biais / variance[#]
- Point de vigilance, particulièrement lorsque ces données sont complexes à calculer[*] & coûteuses à acquérir
  - Energy Star Score peu utile & pas toujours disponible
- Les résultats présentés ne sont que des exemples
  - Pas de « meilleure modélisation » absolue
  - Autres modèles, paramètres, hypothèses[+] & groupes de données à tester

[#] voir Annexe 1
[*] Source : https://www.energystar.gov/buildings/benchmark/understand-metrics/score-details
[+] voir Annexe 2

Annexe 1 – The Bias / Variance Trade-Off



Graphical illustration of bias & variance



Bias & variance vs. model total error

Bias is the tendency of a model to provide inaccurate predictions.
Variance is the tendency of a model to provide very different predictions when trained with different sets of features.

Source : https://scott.fortmann-roe.com/docs/BiasVariance.html

ANNEXE 2 – IMPACT DE L'ENERGYSTARSCORE SUR LES PRÉDICTIONS

Modèle = Random Forest Regressor