Anticipation des besoins en consommation des bâtiments de Seattle

OC Data Scientist
Projet 4

Clément Dubois
18/10/2022

Contexte générale

Contexte :

La ville de Seattle souhaite devenir une ville neutre en émission de carbone en 2050.

Mission :

- Analyser la consommation en énergie et les émissions de G.E.S. des **bâtiments non destinés à l'habitation**.
- Tenter de prédire ces deux valeurs à partir des propriétés géographiques, architecturales et d'usage des bâtiments.
- Tester la pertinence de l'indicateur ENERGYSTARScore, difficile et couteux à acquérir.

Données :

Relevé des différents indicateurs pour les bâtiments de Seattle pour l'année 2016.

Sommaire

Partie I : **Présentation et préparation des données**

Présentation des données et des variables d'intérêts

- **Données fournies le 15 mars 2018 par Seattle sur des relevés de 2016.**
- **3,376 propriétés**
- **46 variables décrivant les propriétés (géographiques, architecturales, usage, consommations, émissions)**

GEOGRAPHIQUES	ARCHITECTURAUX	USAGE	EMISSIONS/CONSO
• Longitude • Latitude • Code de District (District Code) • Quartier (Neighborhood)	• Nombre de bâtiments • Nombre d'étages • Année de construction • Surface totale (propertyGFATotal)	• Type de propriété principale	• TotalGHGEmmissions • SiteEnergyUse

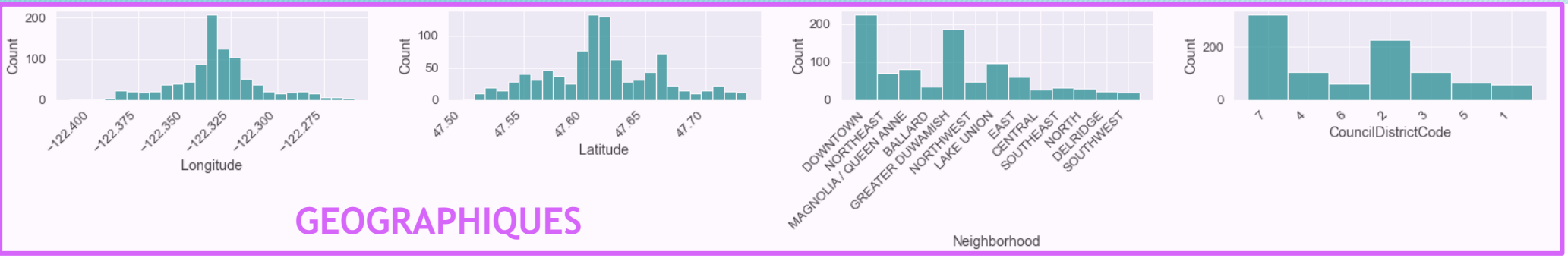
Modification et création de variable :

Type de propriété principale : 'Office', 'Small- and Mid-Sized Office', 'Large Office' = Office
Parking_per : Surface parking/Surface totale
Largest_UseType_per = Surface d'utilisation principale / Surface totale

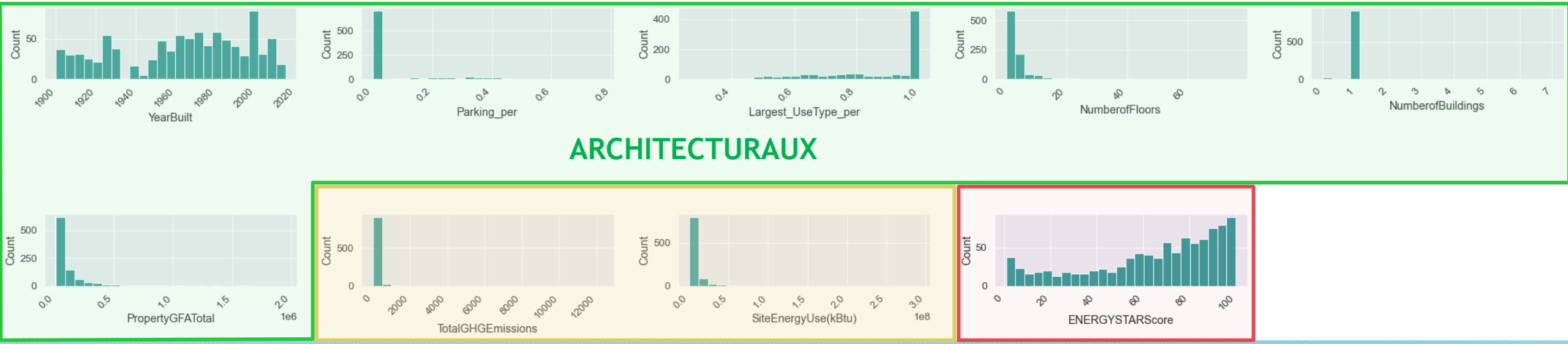
Répartition empirique et modification des données

USAGE

GEOGRAPHIQUES



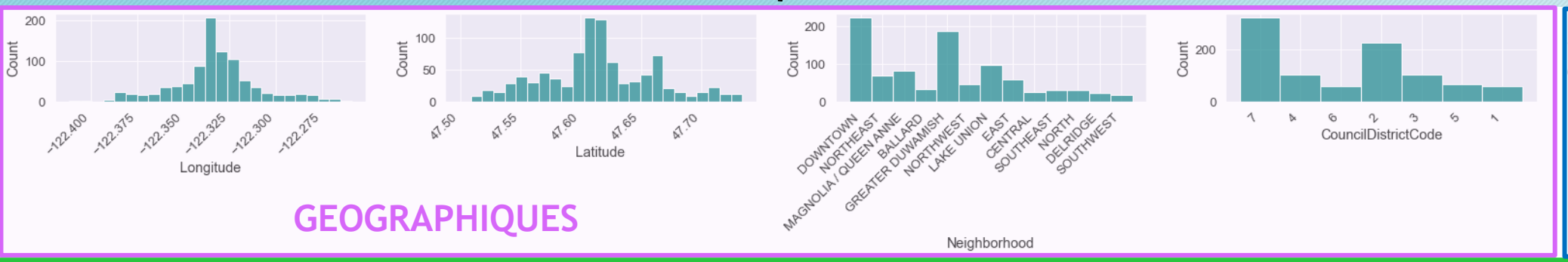
ARCHITECTURAUX

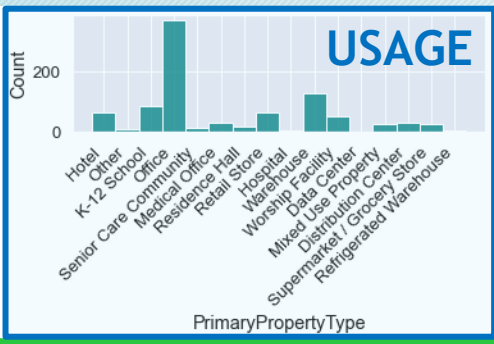


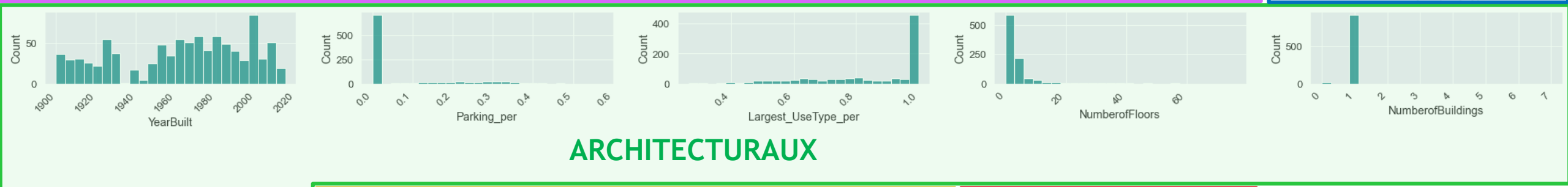
EMISSION + CONSOMMATION

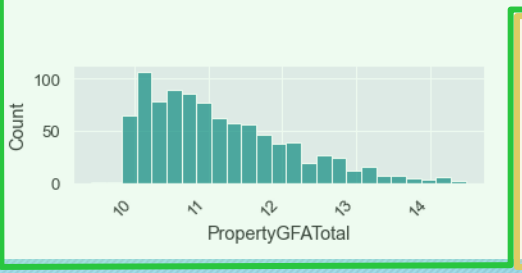
Répartition empirique et modification des données

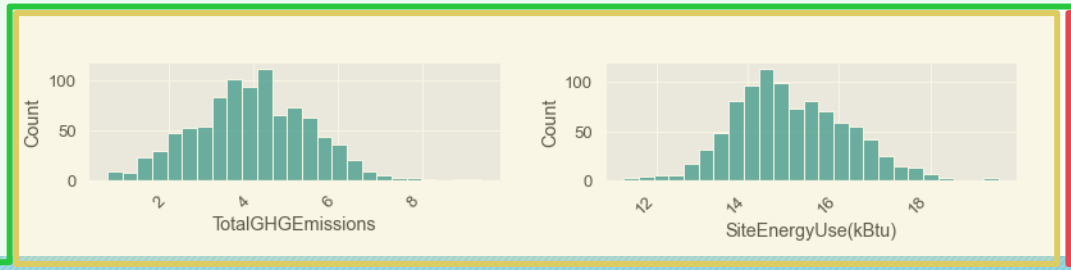
Encodage des données catégorielles (OneHotEncoding)

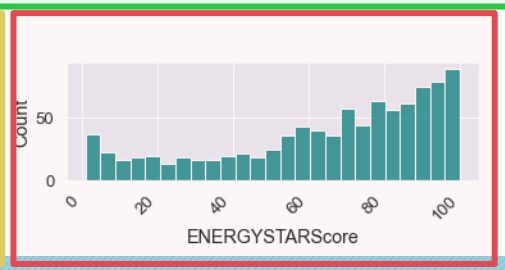










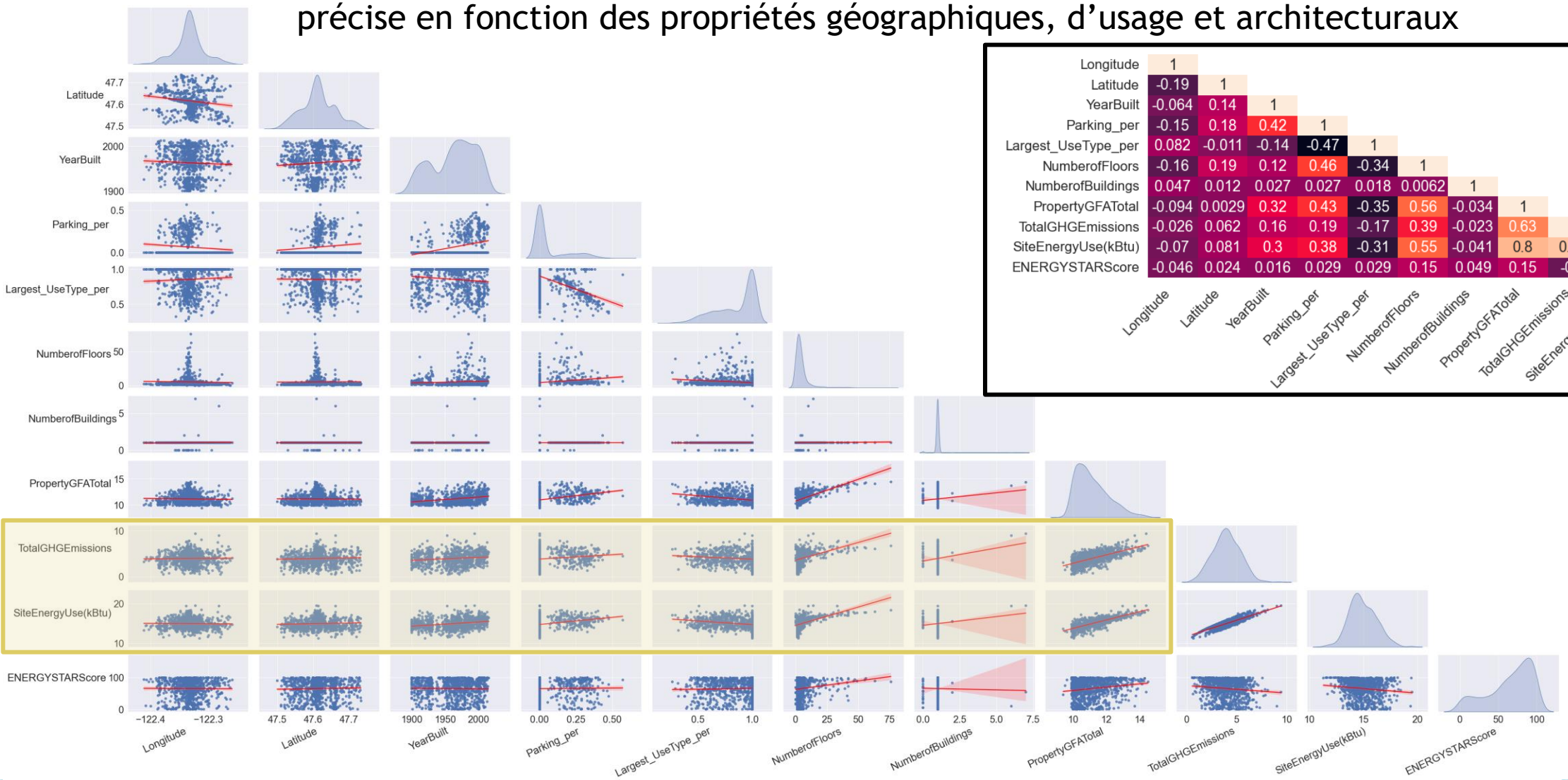


EMISSION + CONSOMMATION

Transformation en logarithme naturel (+1)

Brève analyse de la relation entre les variables

Objectif : Prédire la consommation en énergie et les émissions de GES de façon précise en fonction des propriétés géographiques, d'usage et architecturaux



Partie II : **Méthodes de prédiction**

Présentation de la démarche de modélisation

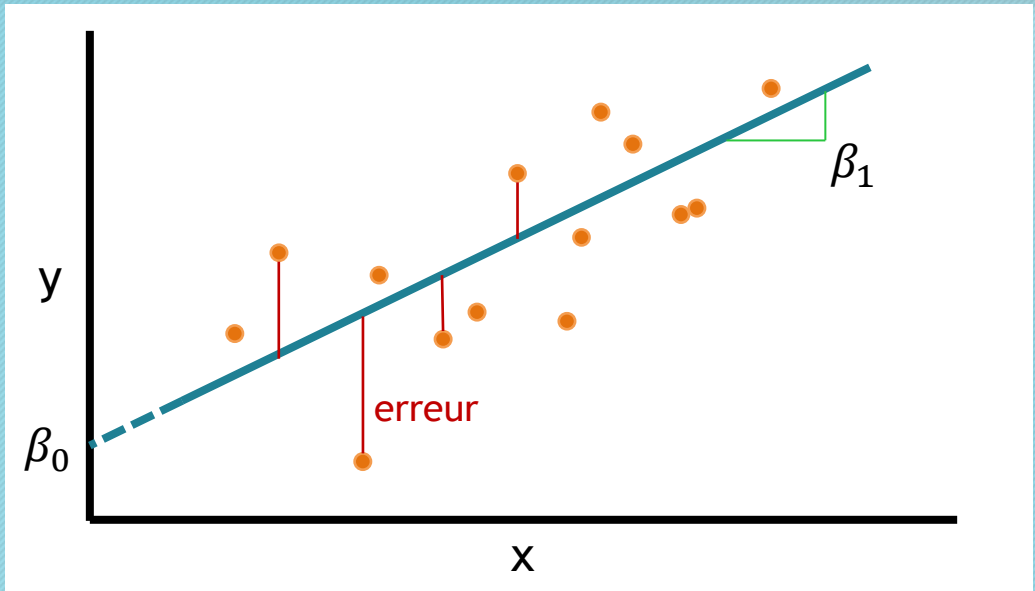
1) Prédictions des valeurs de Consommation et des Emissions de GES

- Comparaison : 10 modèles de Machine Learning succinctement optimisés par validation croisée
- 4 critères (métriques)
- Données standardisées ou non
- Différents échantillons de départ (Random state de l'échantillonnage (split) des données d'entraînement/test)

2) Optimisation plus poussée du meilleur modèle + Courbe d'apprentissage

3) Importance des variables (dont ENERGYSTARScore)

Les modèles utilisés : Modèles linéaires



$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i + \varepsilon$$

Objectif : Minimiser la somme des carrés des erreurs (ou résidus) = Méthode des Moindres Carrés (OLS)

Régression linéaire : Ordinary Least Squares

$$L_{OLS}(\hat{\beta}) = \sum_{i-1}^{n} (y_i - x_i' \hat{\beta})^2$$

Régression Ridge : OLS + L1 penalty

- Le poids des variables

$$L_{lasso}(\hat{\beta}) = \sum_{i-1}^{n} (y_i - x_i' \hat{\beta})^2 + \lambda \sum_{j-1}^{m} |\hat{\beta}_j|.$$

Régression Lasso : OLS + L2 penalty

- Le nombre de variables

$$L_{ridge}(\hat{\beta}) = \sum_{i-1}^{n} (y_i - x_i' \hat{\beta})^2 + \lambda \sum_{j-1}^{m} \hat{\beta}_j^2$$

Régression ElasticNet : OLS + L1 + L2 penalty

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i-1}^{n} (y_i - x_i' \hat{\beta})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j-1}^{m} \hat{\beta}_j^2 + \alpha \sum_{j-1}^{m} |\hat{\beta}_j| \right),$$

Les modèles utilisés : Modèles ensemblistes

- Méthodes parallèles

Bagging (tree) + Random Forest

- Méthodes séquentielles

Adaboost + (X)Gradient Boosting

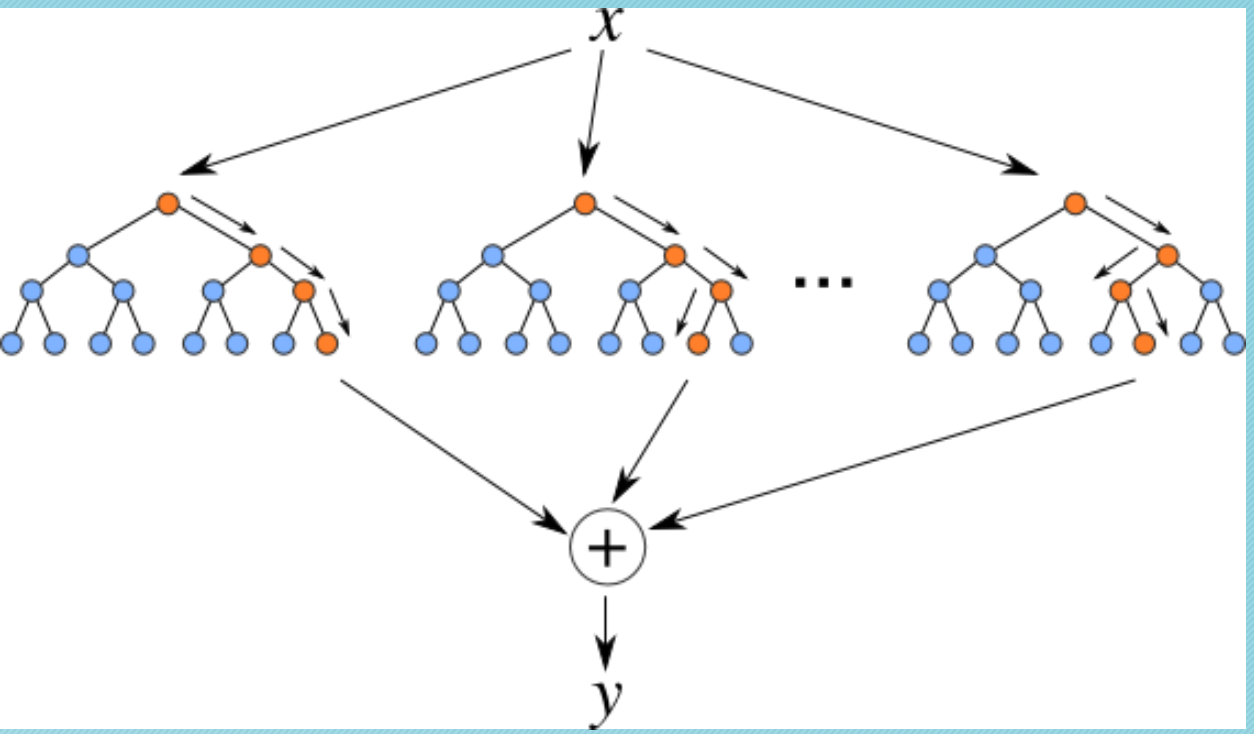


Image : Bakshi, Random Forest Regression. Medium, 2020

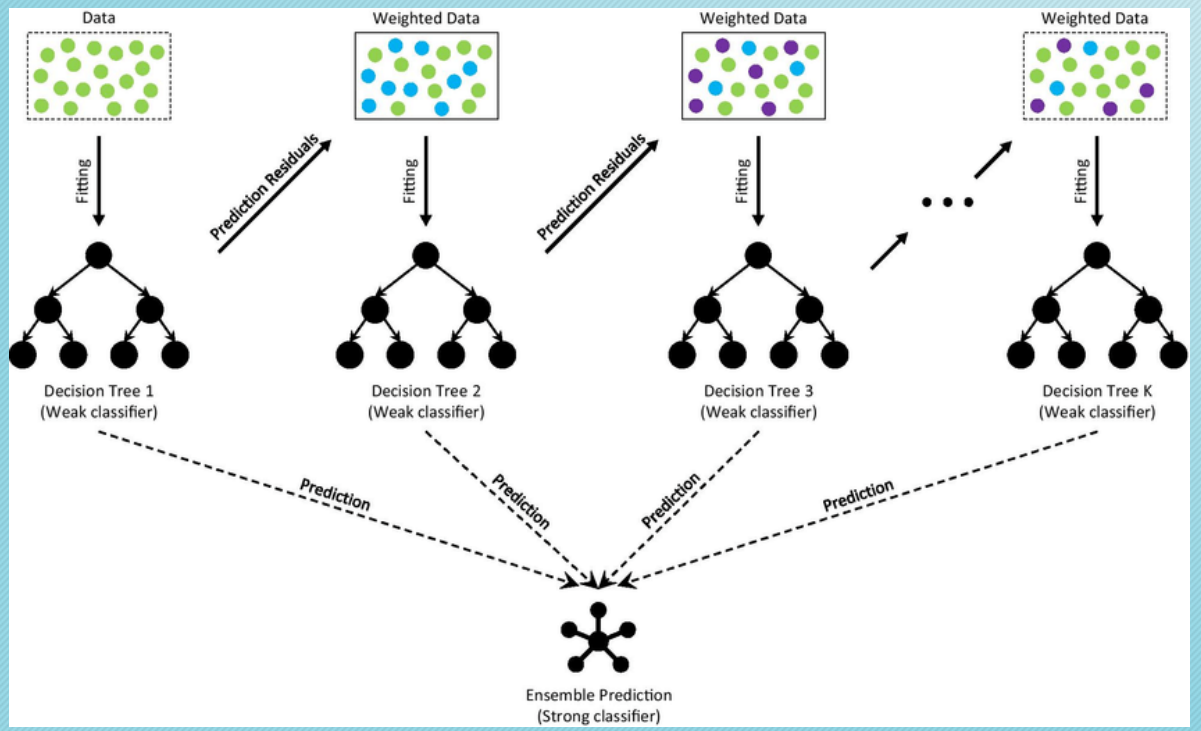


Image : Deng et al. BMC Medical Informatics and Decision Making 2021

Les modèles utilisés : Autres modèles

- Méthode de support vector machine
- Méthode des K plus proches voisins
- Méthode de Perceptron multicouches
- Modèle nul (moyenne)

Critères de comparaison : les différentes métriques

- Erreur quadratique moyenne (Mean Squared error MSE) $$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
- Erreur quadratique moyenne (Mean Squared error) $$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$
- Coefficient de détermination R^2 $$R^2 = 1 - \frac{\overbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}^{\text{erreur quadratique}}}{\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{variance}}}$$
- Erreur Moyenne Absolue (Mean Absolute Error MAE) $$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

- y_i valeur vraie
- $\hat{y}_i$ valeur prédite
- $\bar{y}$ valeur moyenne

Importance des Variables : Valeurs de Shapley

Lundberg & Lee 2017 :

Algorithme basé sur la théorie des jeux coopératifs *(Lloyd S. Shapley)*
La prédiction d'un individu = somme de contributions de chacune des variables.

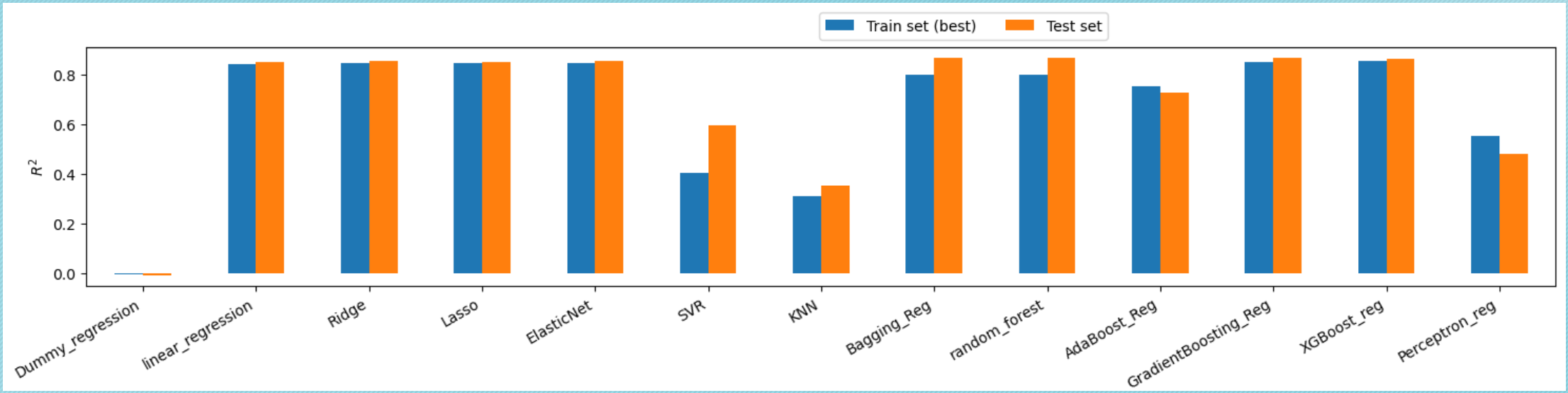
Partie III : **Résultats**

Résultats 1 : Prédictions de la consommation énergétique

Résultats 2 : Prédictions des émissions de G.E.S.

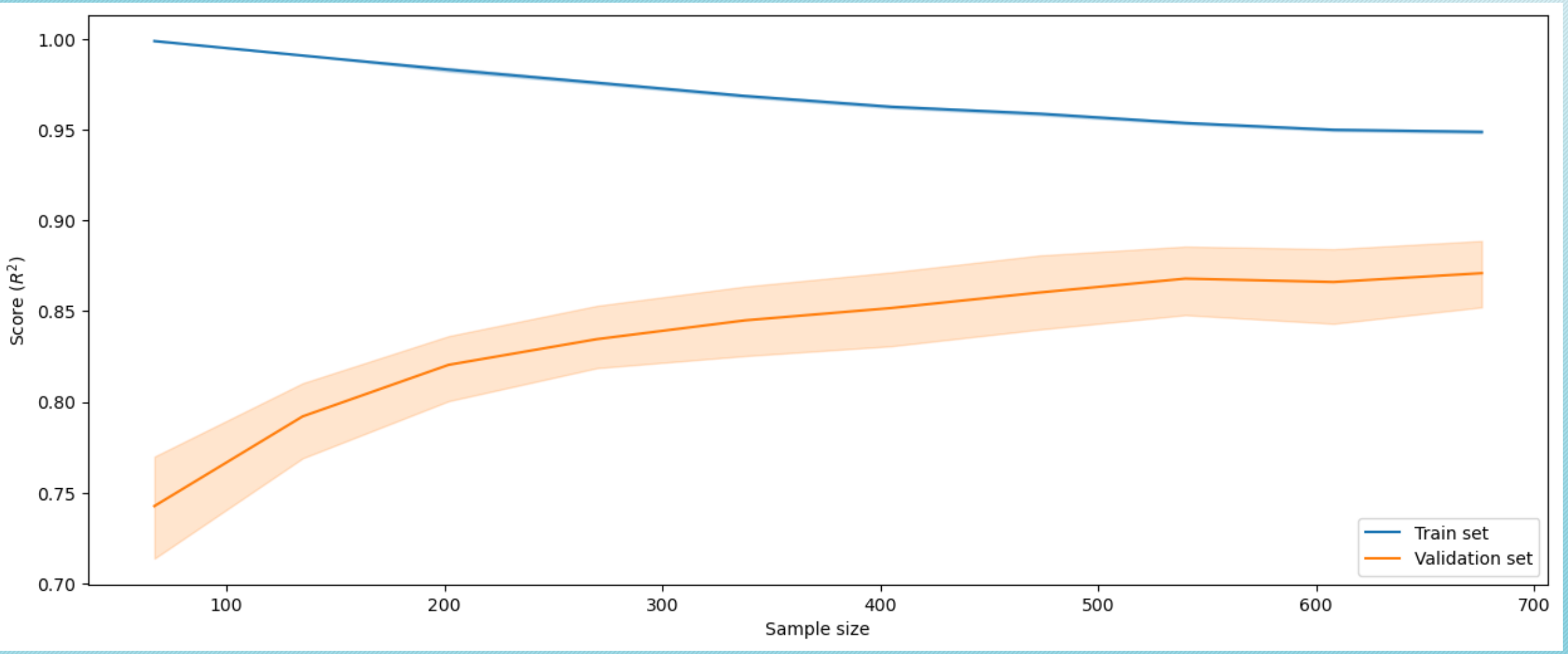
Résultats 1 : Prédictions de la consommation énergétique

Modèle	R^2 : entrainement (meilleurs param)	R^2 : test	MSE : test	RMSE : test	MAE : test
Dummy_regression	-0.006	-0.000	1.178	1.387	0.957
linear_regression	0.857	0.871	0.423	0.179	0.331
Ridge	0.858	0.872	0.422	0.178	0.330
Lasso	0.864	0.871	0.423	0.179	0.329
ElasticNet	0.861	0.872	0.421	0.178	0.328
SVR	0.834	0.868	0.429	0.184	0.319
KNN	0.607	0.658	0.689	0.474	0.533
Bagging_Reg	0.809	0.847	0.460	0.212	0.344
random_forest	0.808	0.841	0.469	0.220	0.350
AdaBoost_Reg	0.740	0.776	0.557	0.310	0.458
GradientBoosting_Reg	0.848	0.884	0.401	0.161	0.300
XGBoost_reg	0.843	0.874	0.419	0.175	0.304
Perceptron_reg	0.713	0.872	0.421	0.177	0.321

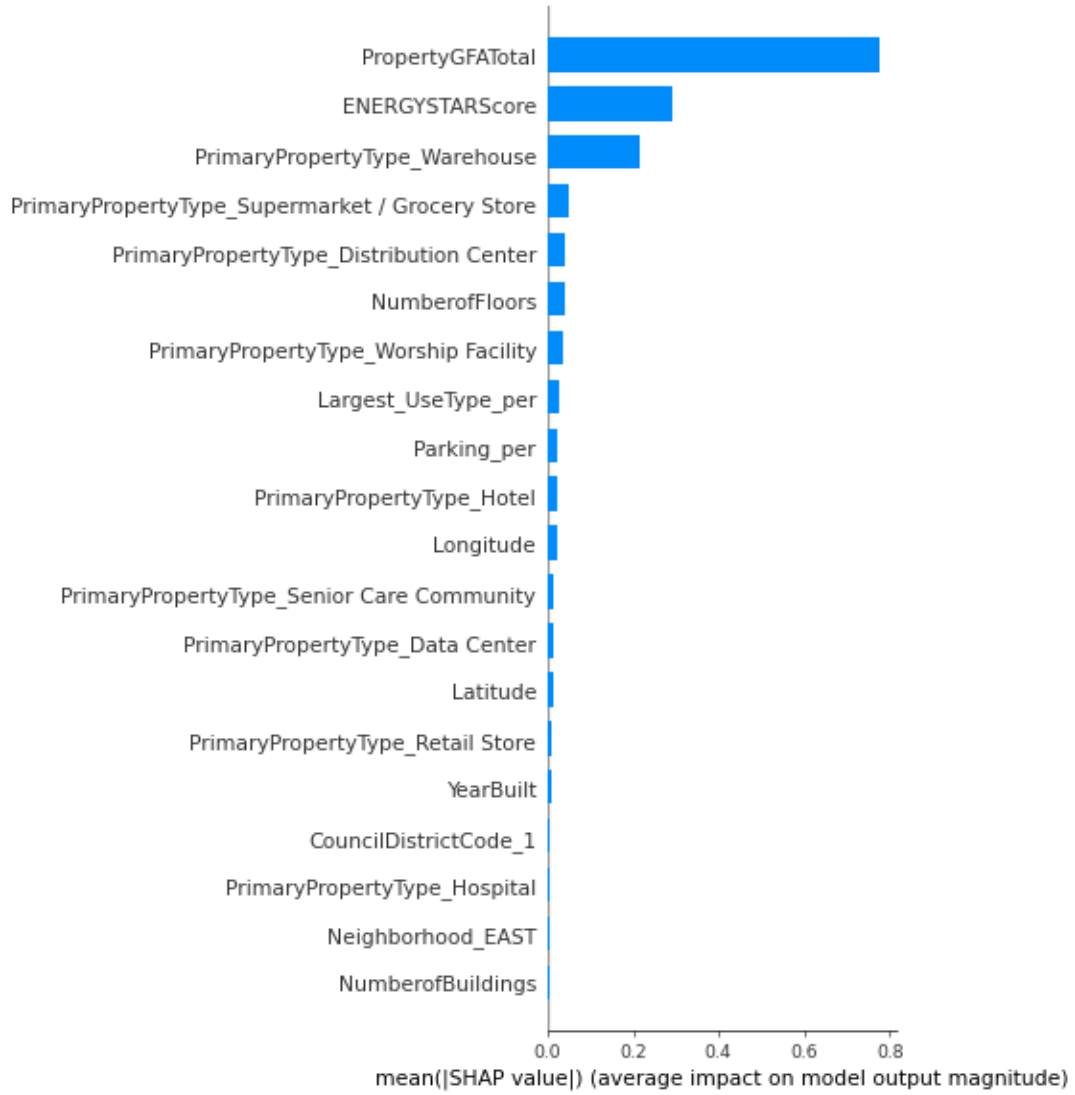


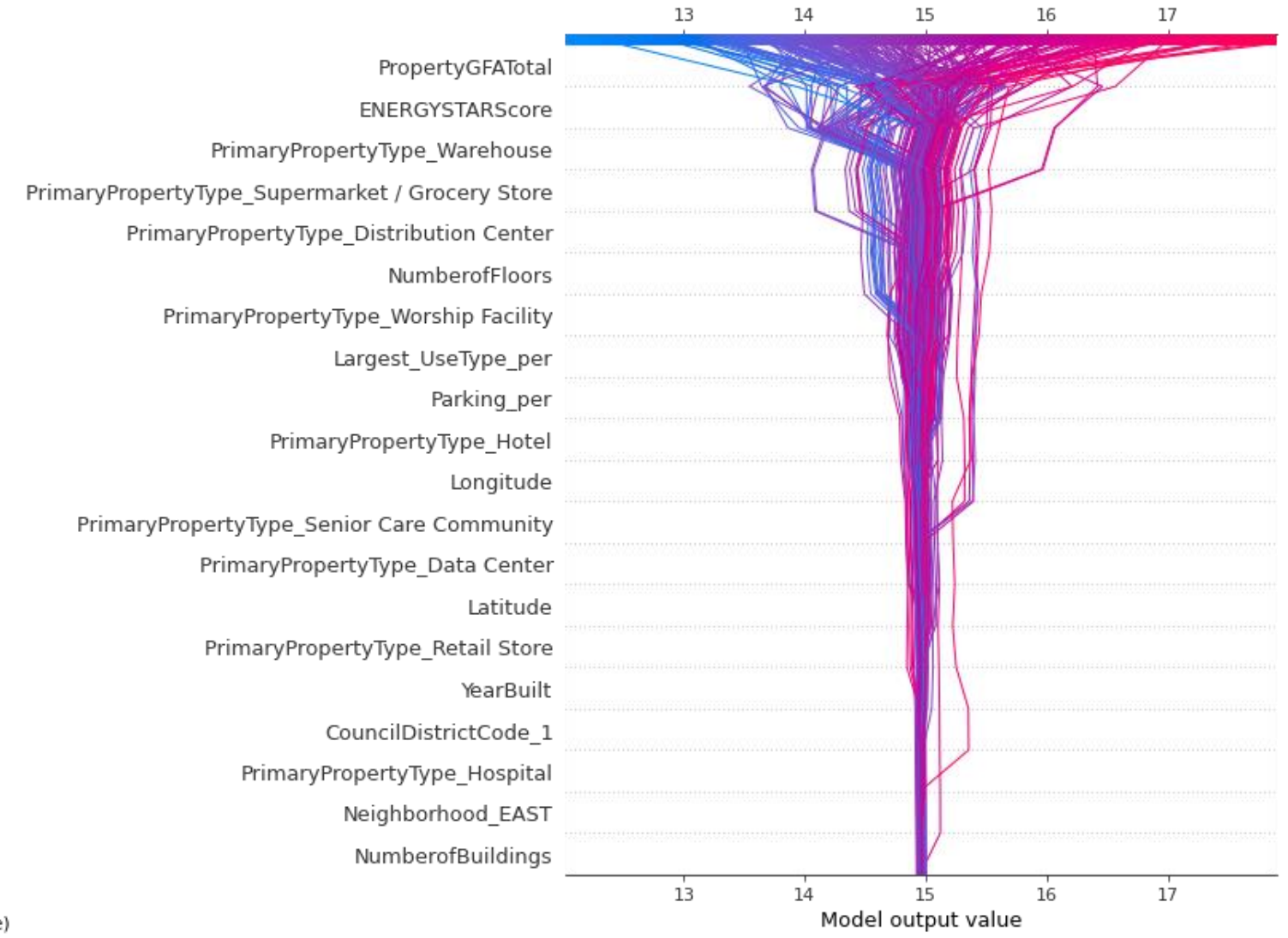
Résultats 1 : Optimisation des meilleurs modèles

Données numériques	R^2 : test	MSE : test	RMSE : test	MAE : test
Non Standardisées	0.888	0.395	0.156	0.303

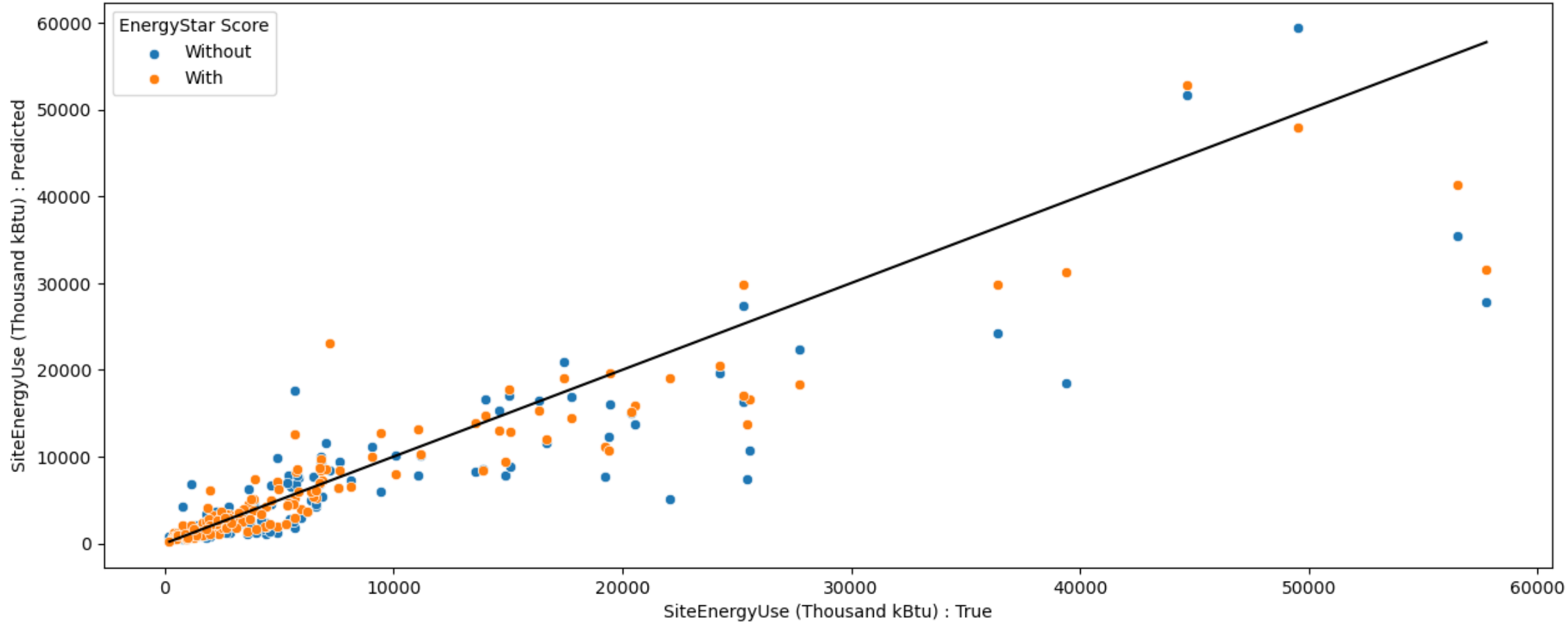


Résultats 1 : Importances des variables





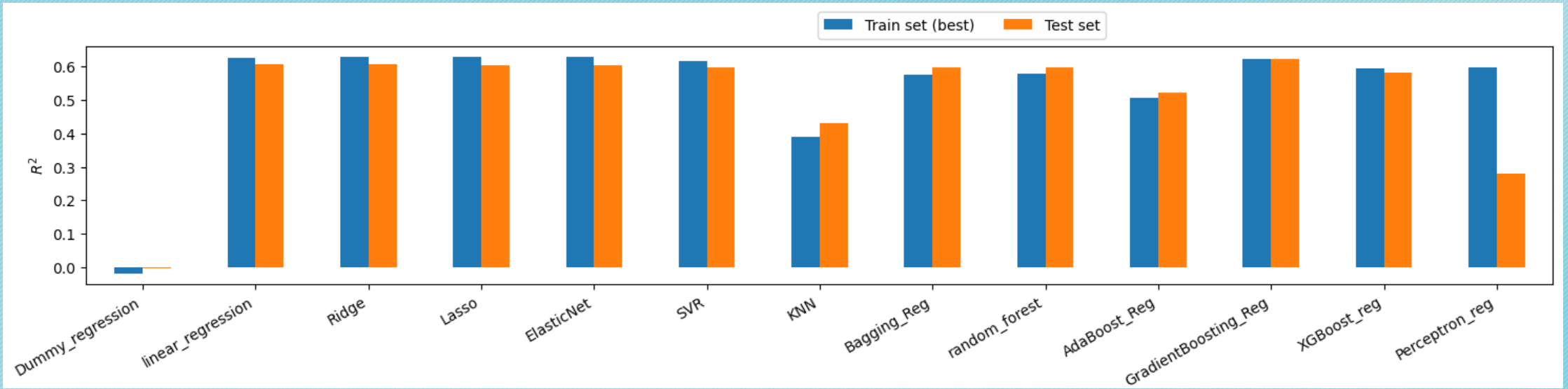
Résultats 1 : Comparaison des valeurs prédites/vraies



ENERGYSTARScore	R^2 : test	MSE : test	RMSE : test	MAE : test
Avec	0.888	0.395	0.156	0.303
Sans	0.793	0.536	0.287	0.406
Ecart	**-10,7 %**	**+35,7 %**	**+84,0 %**	**+34,0 %**

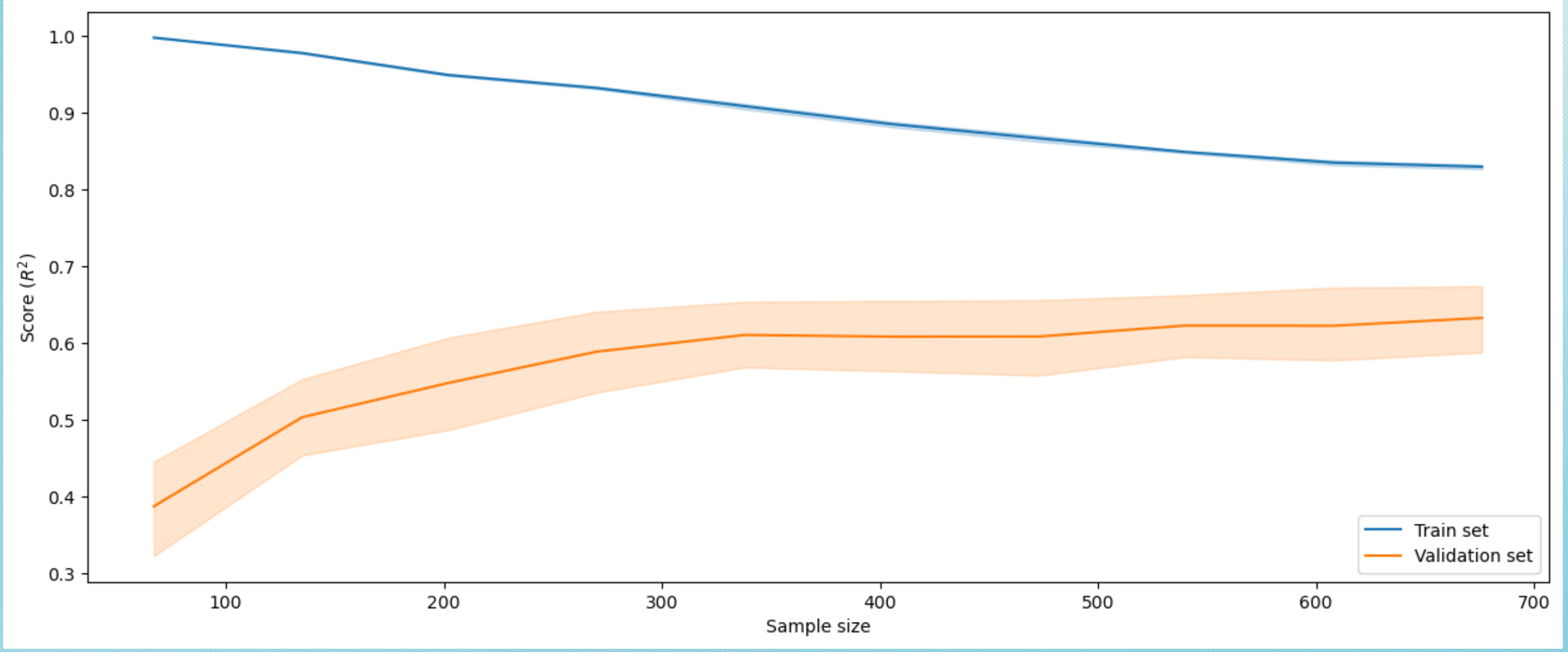
Résultats 2 : Prédictions des émissions de G.E.S.

Modèle	R^2 : entrainement (meilleurs param)	R^2 : test	MSE : test	RMSE : test	MAE : test
Dummy_regression	-0.018	-0.003	1.339	1.792	1.062
linear_regression	0.626	0.607	0.839	0.703	0.685
Ridge	0.630	0.606	0.839	0.704	0.685
Lasso	0.630	0.603	0.843	0.710	0.686
ElasticNet	0.630	0.604	0.841	0.707	0.685
SVR	0.616	0.597	0.849	0.720	0.666
KNN	0.390	0.432	1.008	1.015	0.801
Bagging_Reg	0.577	0.598	0.848	0.718	0.676
random_forest	0.578	0.597	0.849	0.721	0.677
AdaBoost_Reg	0.508	0.523	0.923	0.853	0.778
GradientBoosting_Reg	**0.622**	**0.623**	**0.821**	**0.674**	**0.653**
XGBoost_reg	0.594	0.582	0.864	0.747	0.683
Perceptron_reg	0.597	0.281	1.134	1.285	0.901

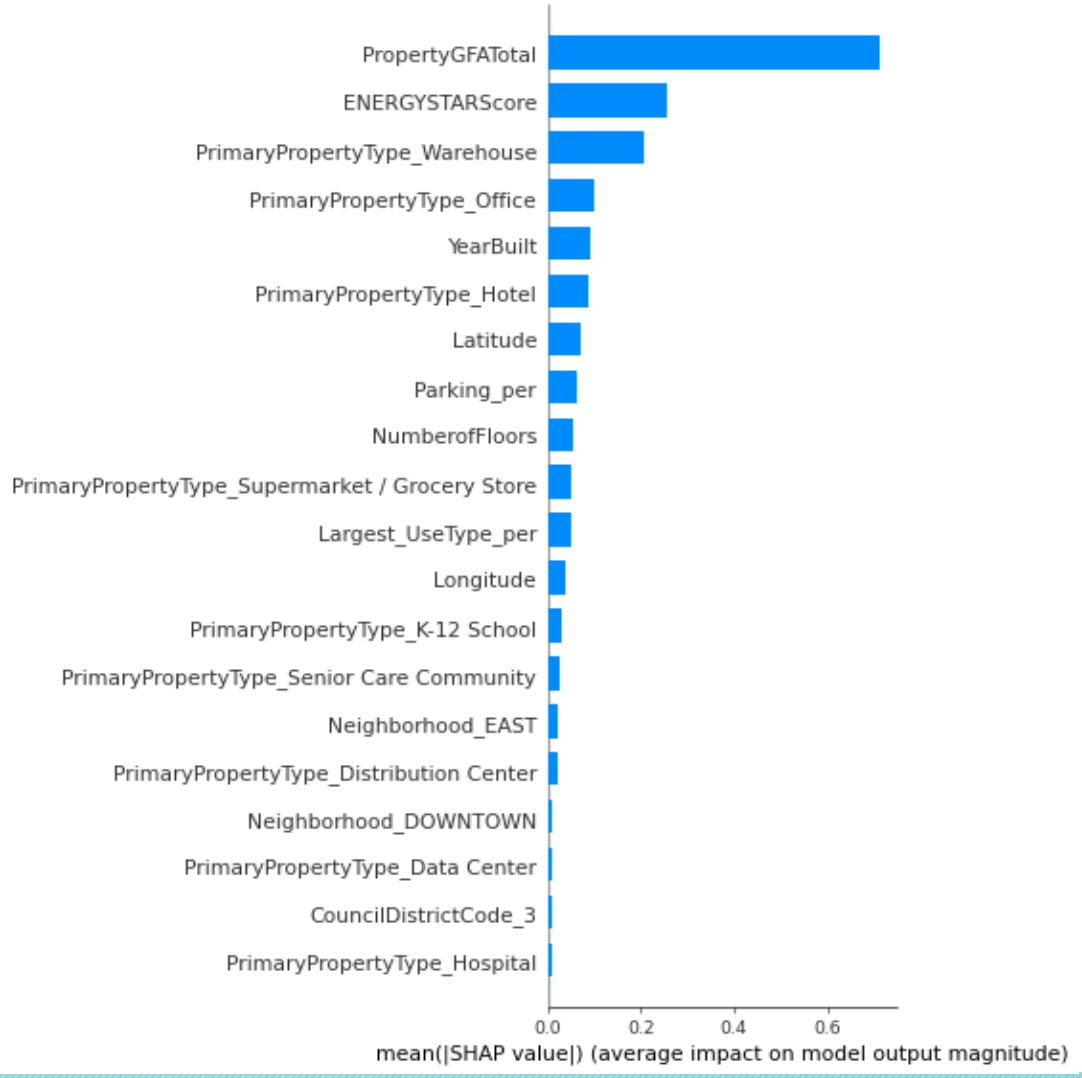


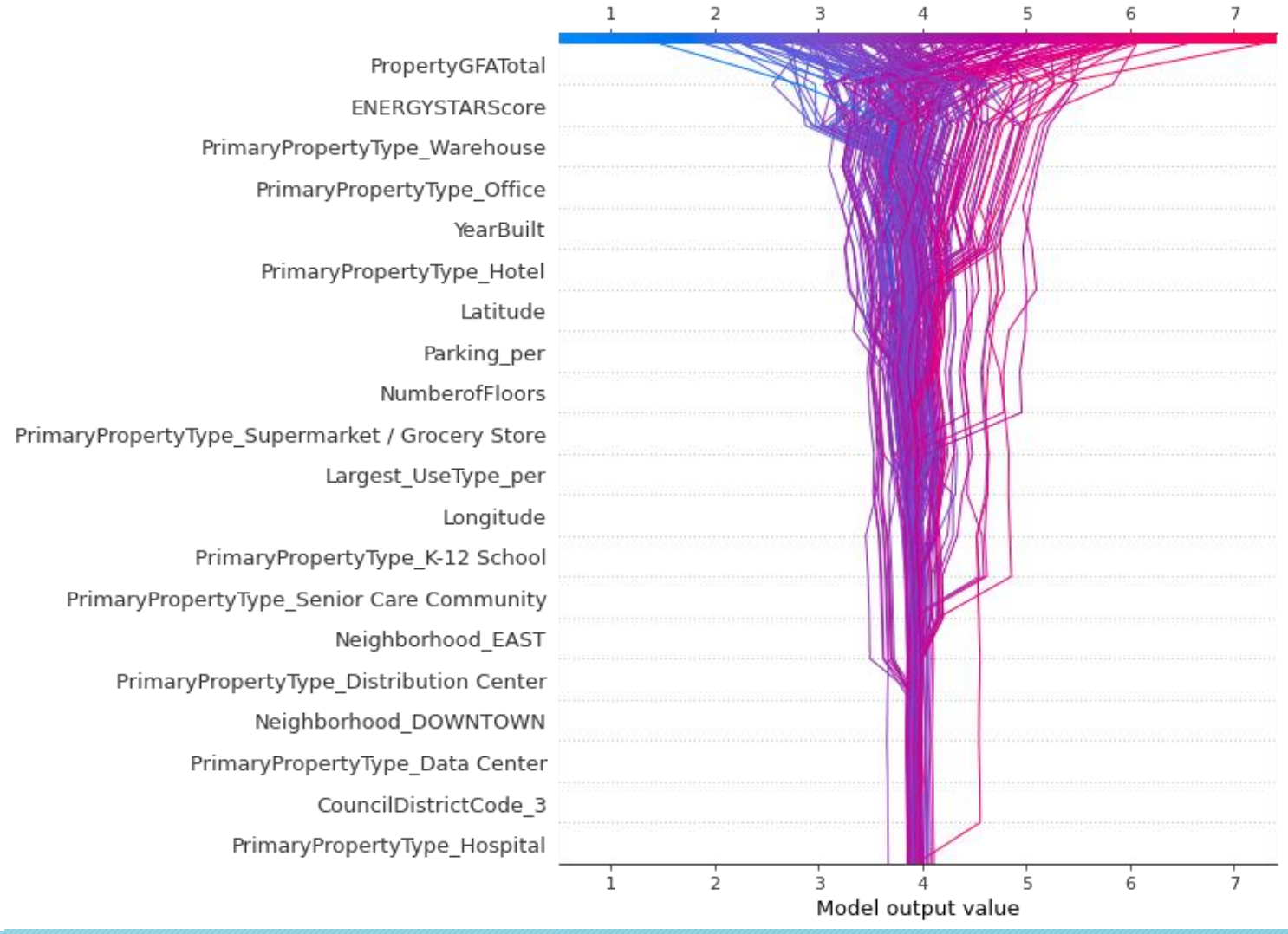
Résultats 2 : Optimisation des meilleurs modèles

Données numériques	R^2 : test	MSE : test	RMSE : test	MAE : test
Non standardisées	0.636	0.806	0.65	0.652

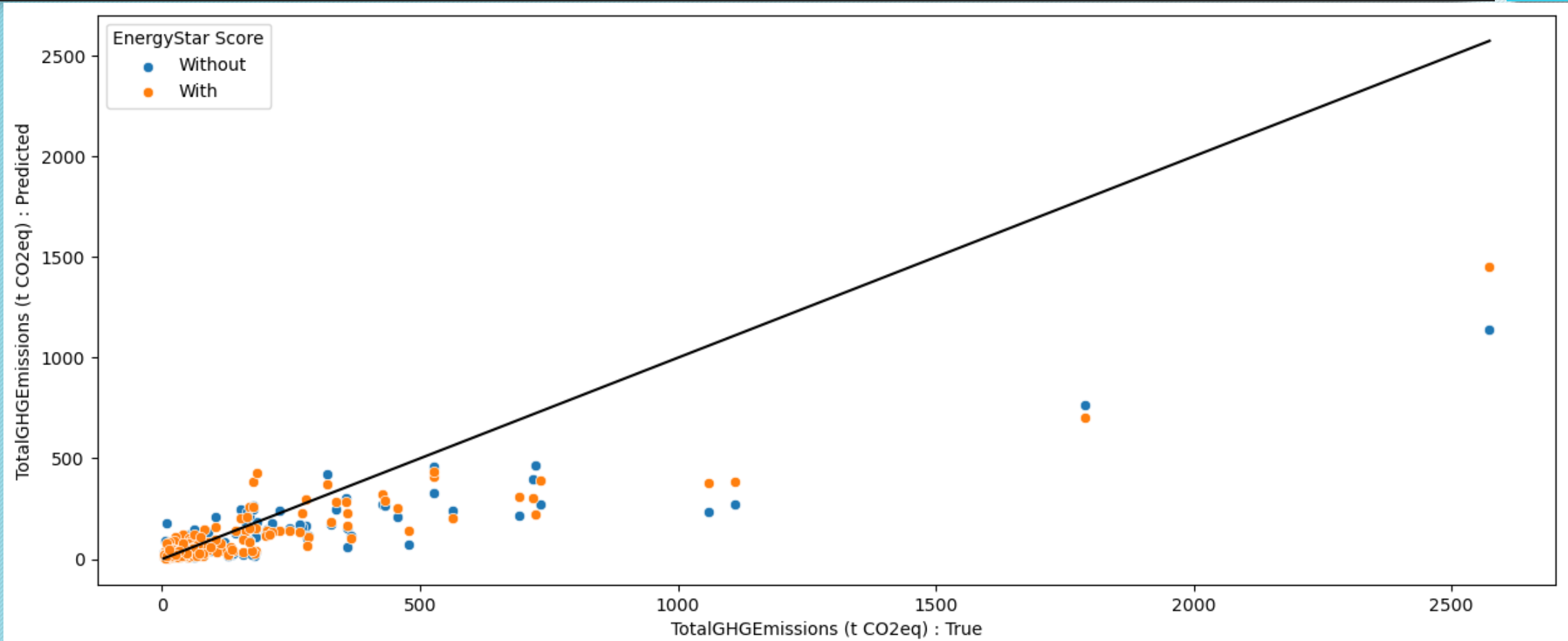


Résultats 2 : Importances des variables





Résultats 2 : Comparaison des valeurs prédites/vraies



ENERGYSTARScore	R^2 : test	MSE : test	RMSE : test	MAE : test
Avec	0.636	0.806	0.65	0.652
Sans	0.594	0.852	0.726	0.67
Ecart	-6,6 %	+5,7 %	+11,7 %	+2,8 %

Conclusion

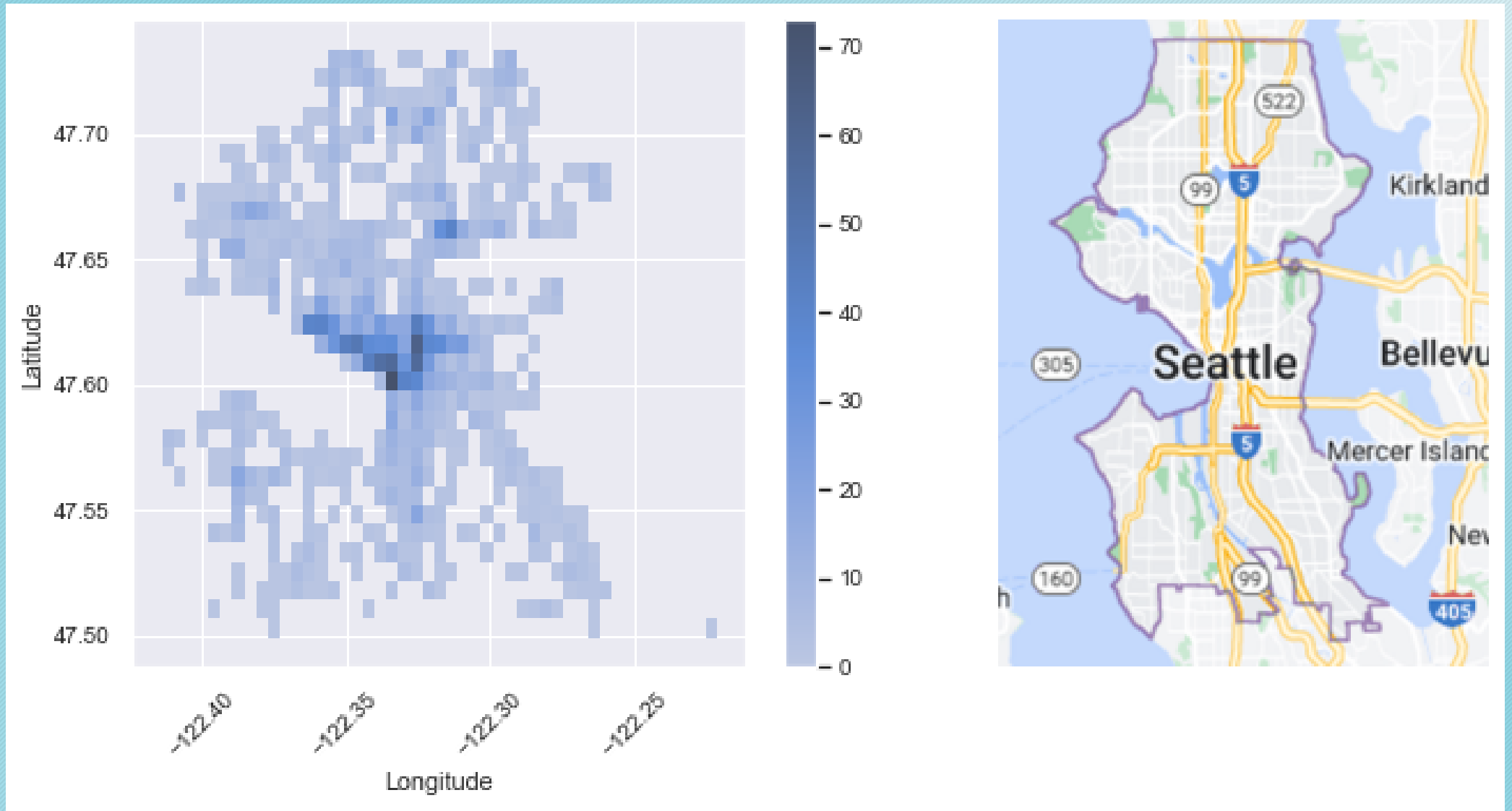
1) Le modèle prédit avec précision la consommation énergétique des bâtiments non destinées à l'habitation.

2) Les prédictions des émissions de G.E.S. sont moins précises.

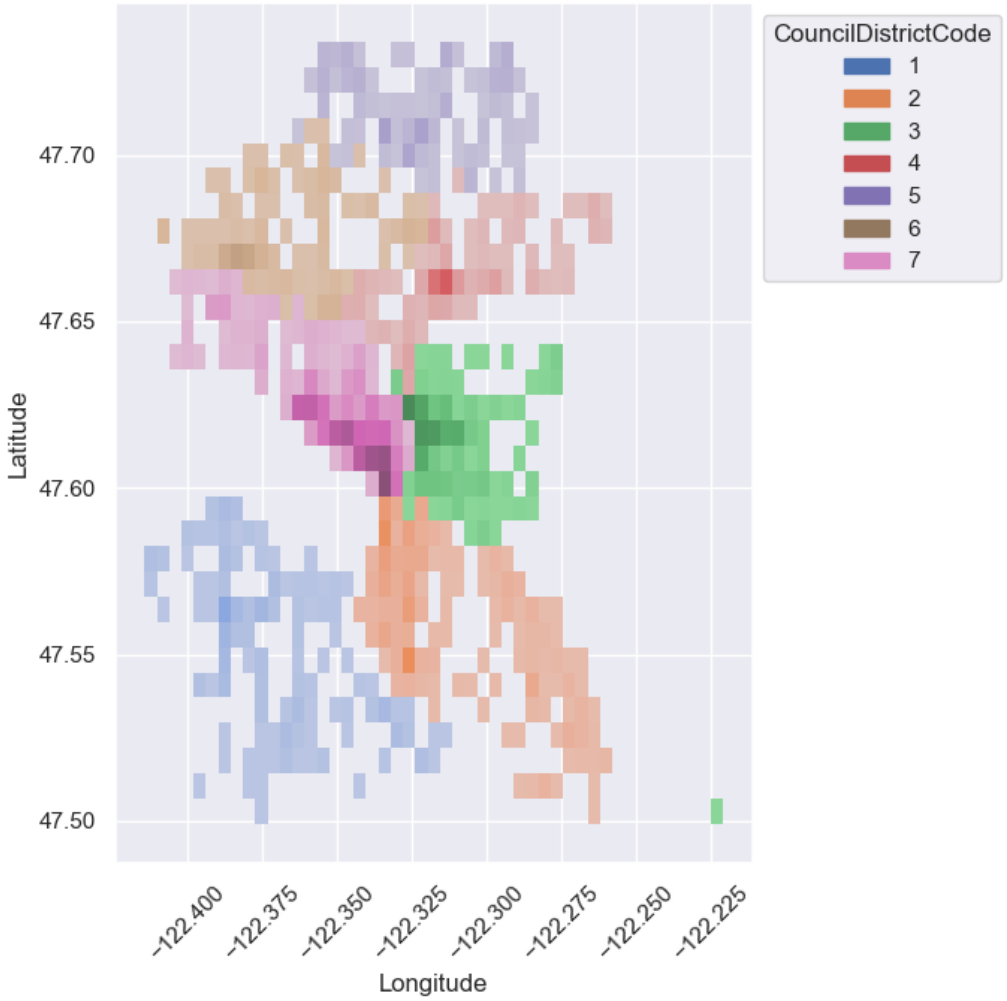
3) Dans les deux cas, ENERGYSTARScore augmentent la précision des prédictions.

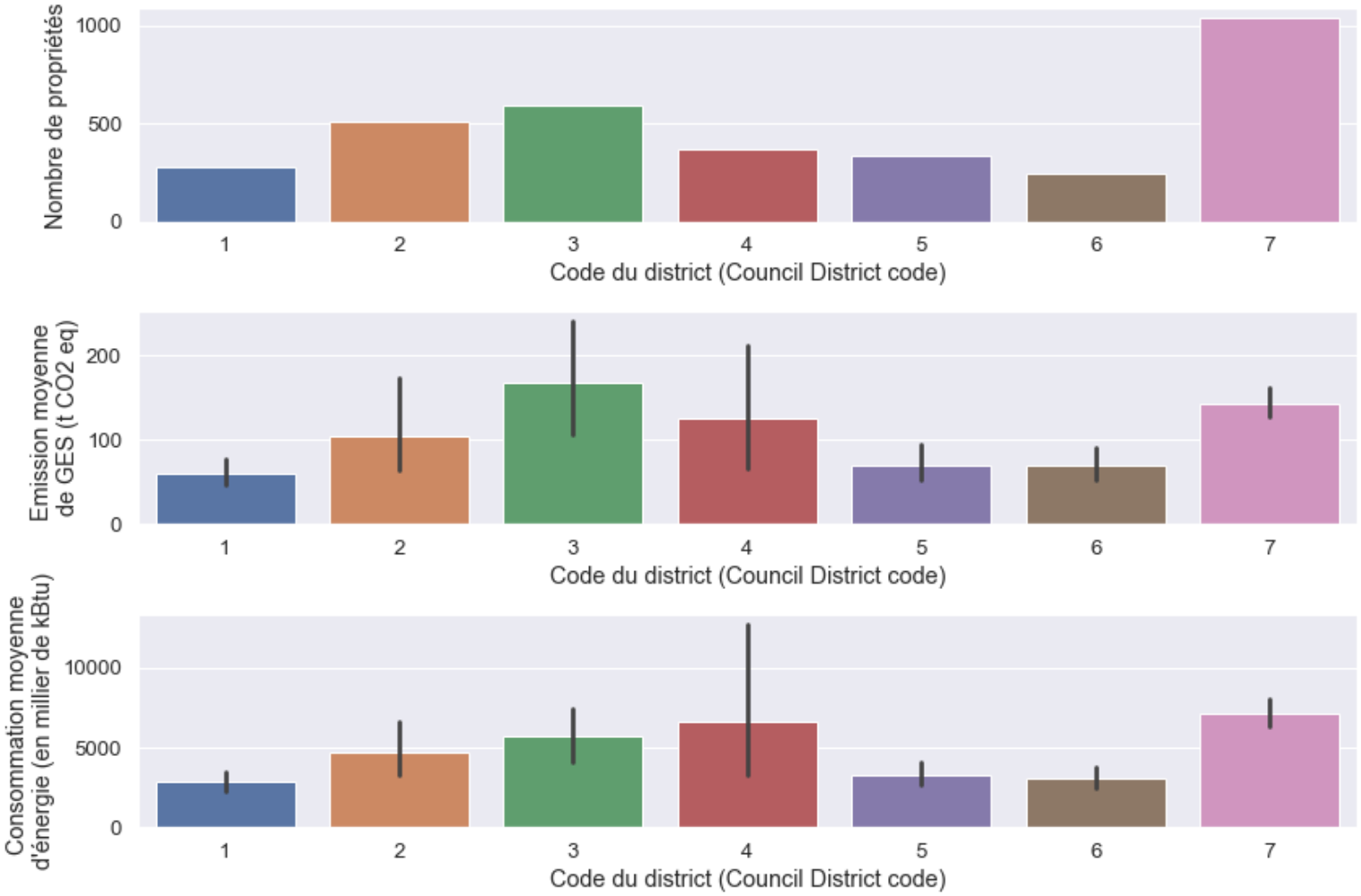
Annexe : **Analyse exploratoire des indicateurs**

Indicateurs géographiques : Longitude latitude

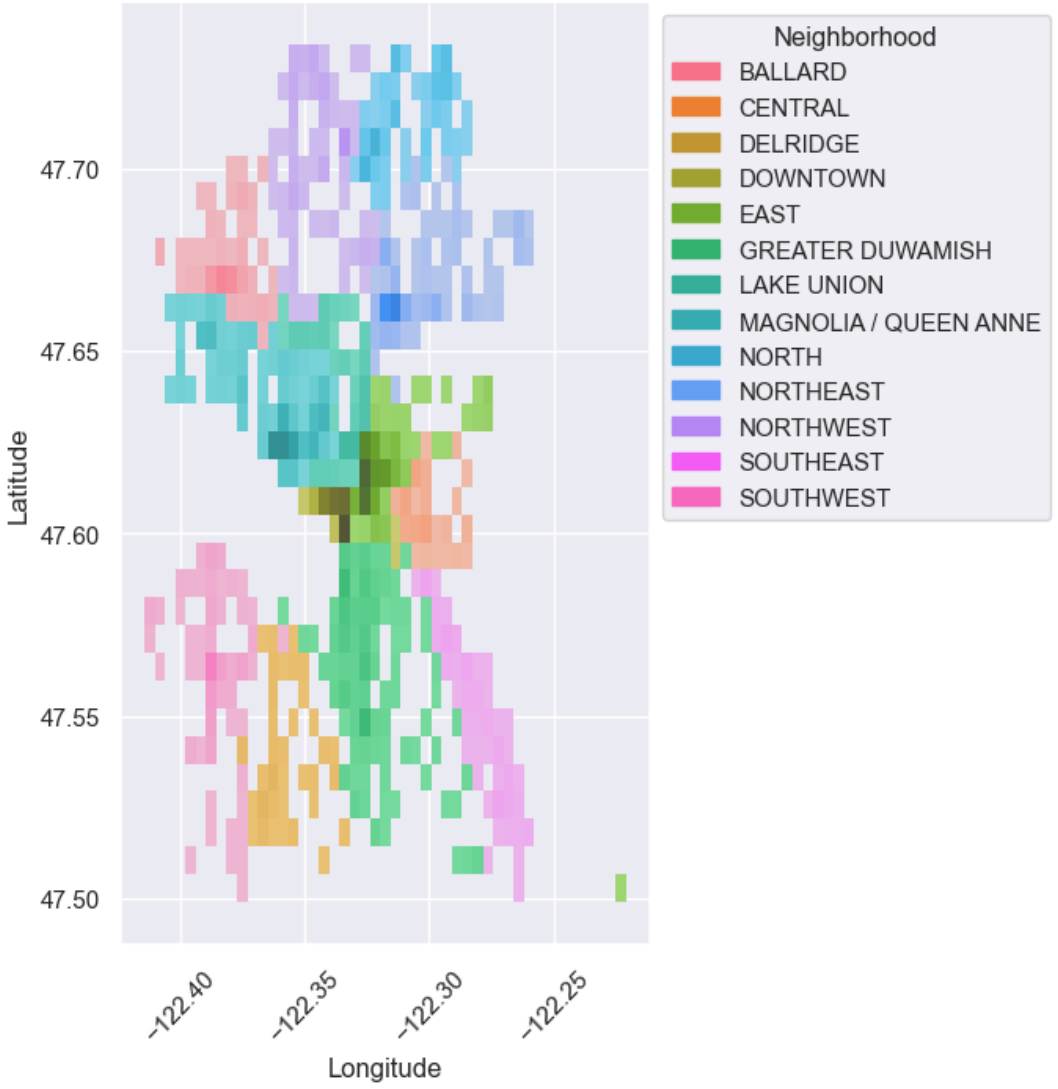


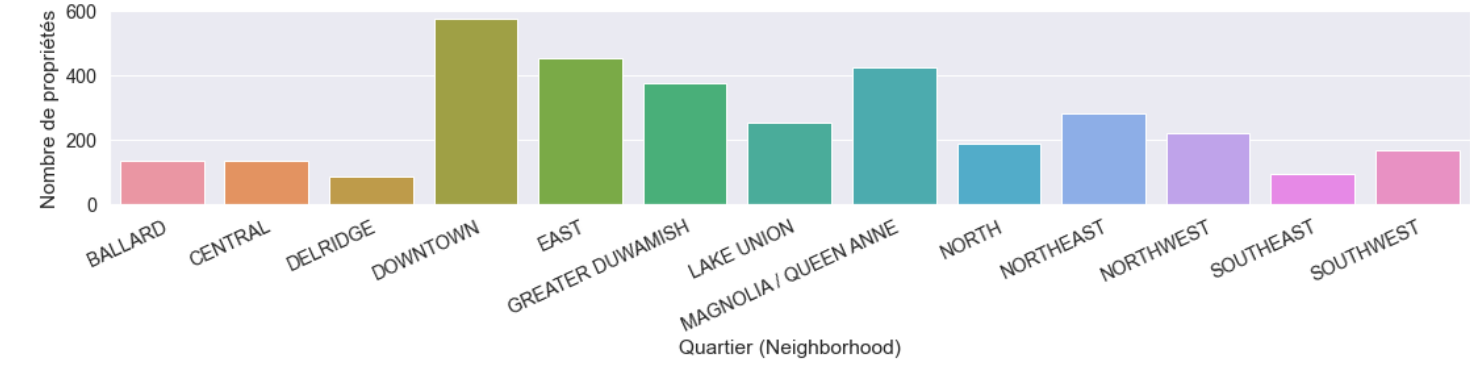
Indicateurs géographiques : District

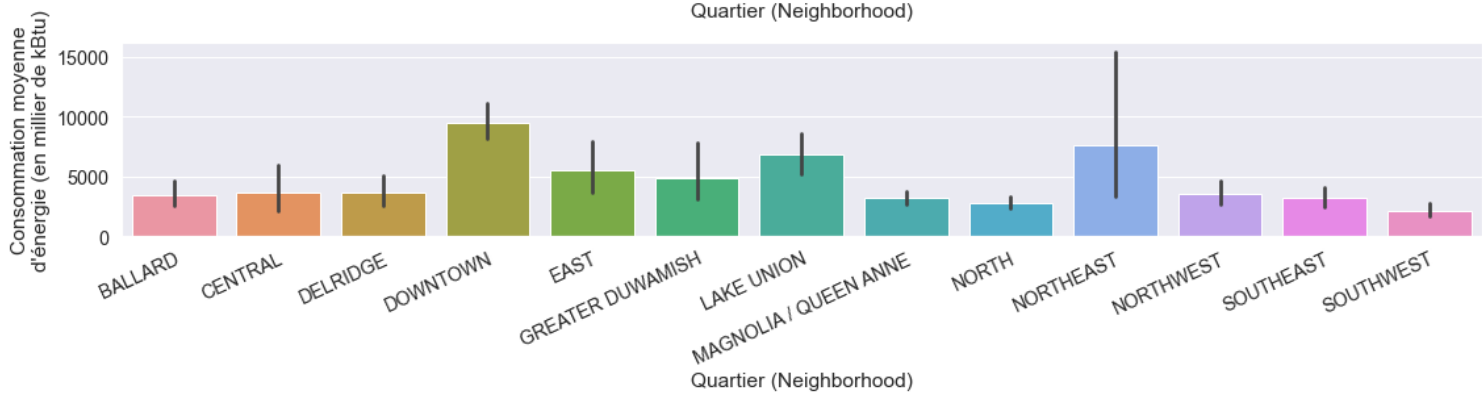




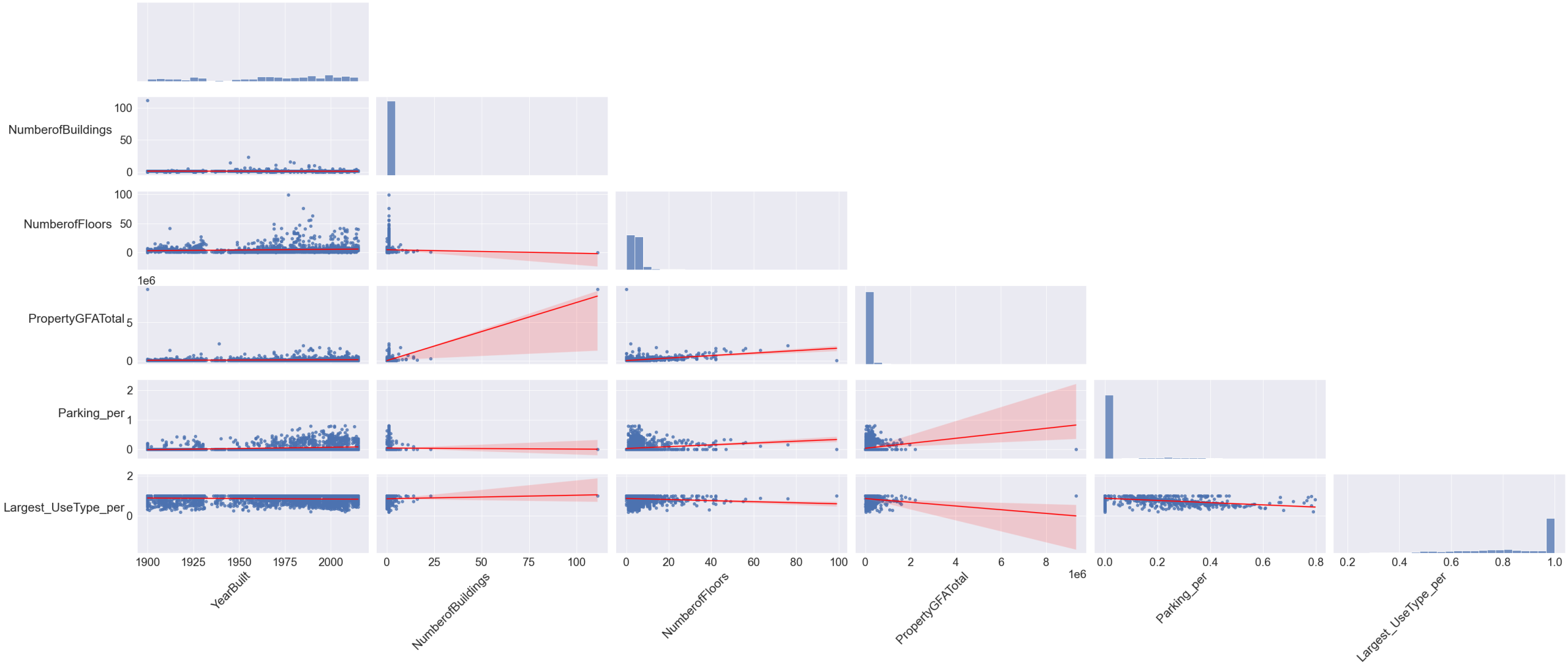
Indicateurs géographiques : Quartier



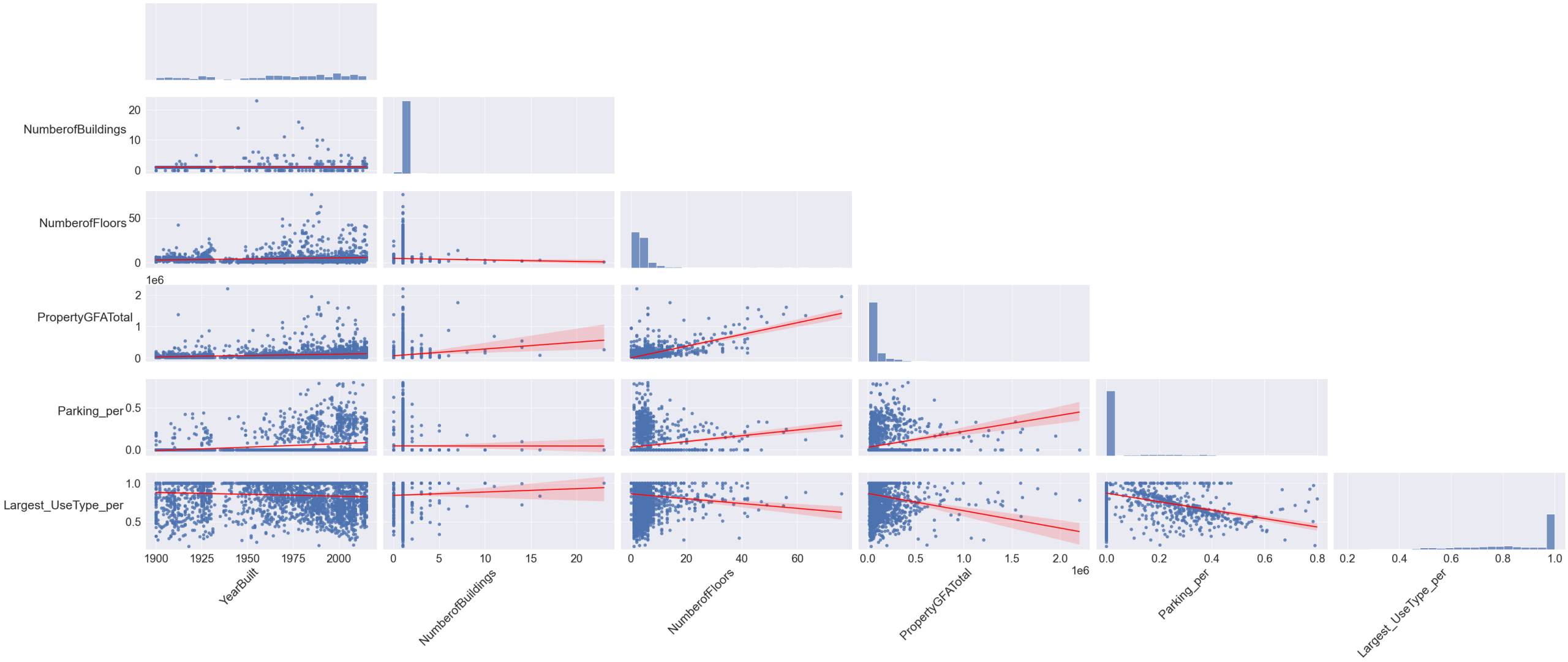




Indicateurs architecturaux :



Indicateurs architecturaux :



Indicateurs d'usages : Type de propriété principale