Seattle

Anticipez les besoins en consommation de bâtiments

Charlotte DUBUS

Soutenance du 12 février 2023 - Parcours Data Scientist - Projet 4

MISSION

+ ***Objectif de Seattle : Neutralité carbone en 2050***

+ **<u>Notre action :</u>**
    - prédictions des émissions de CO^2 et d'énergie sur les bâtiments non résidentiels
    - évaluation de l'impact de l'ENERGY STAR Score

+ Sources : données collectées en 2016 par la ville

SOMMAIRE

COMPREHENSION DU JEU DE DONNEES

3376 enregistrements & 46 variables

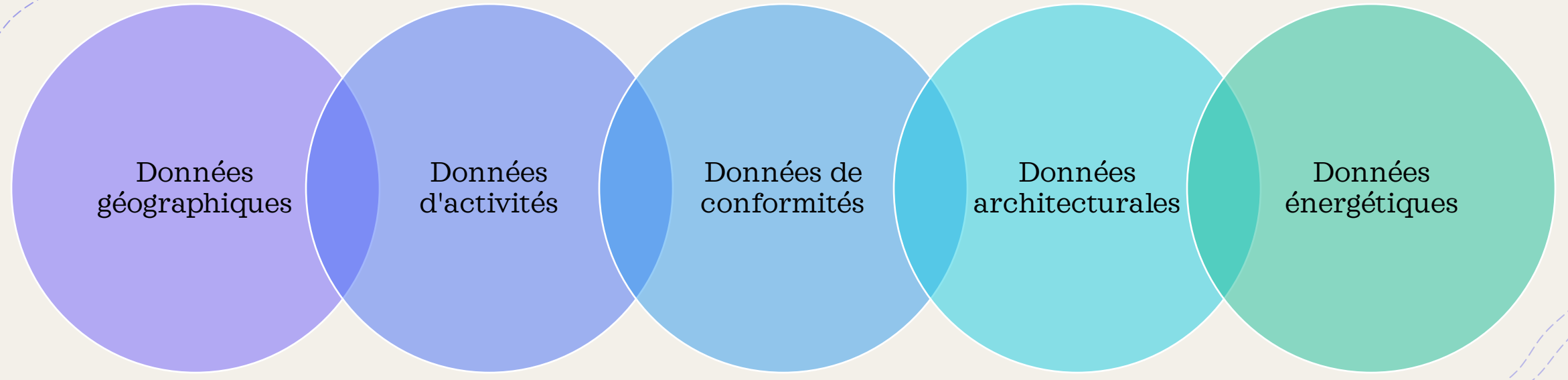
Données numériques ,catégorielles & 1 booléenne

12.8% d'informations manquantes

- Nettoyage de données standard
- Gestion approfondie des absences
  - Colonne ENERGYSTAR Score
- Suppression colonne vide

Base solide

EXPLICATION VARIABLES



ANALYSE EXPLORATOIRE

Localisation

- Nettoyage des ZipCode
- Correction erreurs typographiques
- Rapport Quartiers/Concils

Catégorisation des types de bâtiments

- Filtrage : "Non résidentiels"
- Regroupement
- Graphiques des tendances liés aux targets

Conformité

- Examen de l'influence d' 'Outlier' et 'ComplianceStatus'
- Test Chi² et Heatmap
- Filtrage des bâtiments

Données constructives

- Traitement des valeurs aberrantes
- Visualisation des targets et de l'âge des bâtiments

Approche sélective > Filtrage > Conservation des variables les plus informatives

ANALYSE EXPLORATOIRE

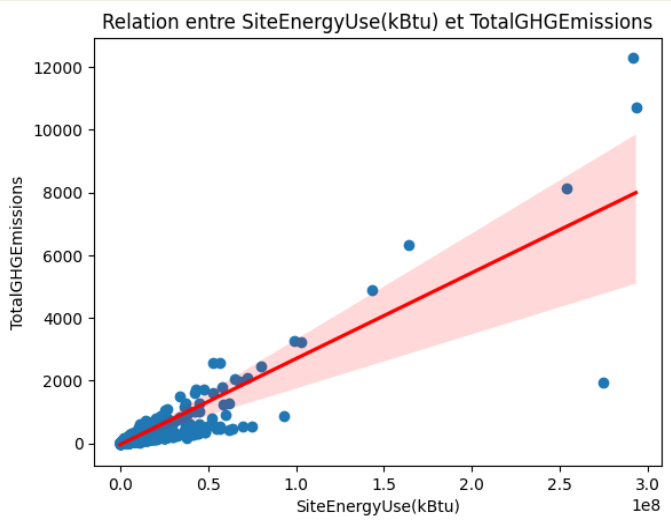
Création de nouvelles caractéristiques

Application d'une transformation logarithmique sur 8 de nos variables

One Hot Encoding sur les données catégorielles

Création d'un Pipeline de pré-traitement

Jeu de donnée prêt pour la modélisation : 1328 enregistrements & 36 features



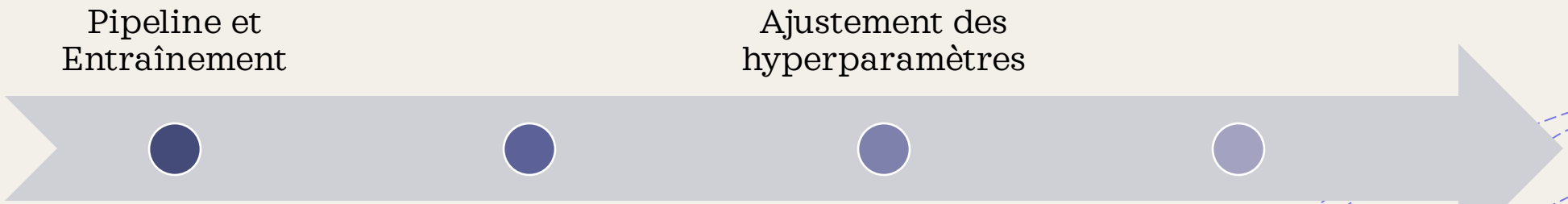
Column	Skewness
Latitude	0.279626
Longitude	-0.181425
NumberofBuildings	9.801974
PropertyGFATotal	4.822461
ENERGYSTARScore	-0.686352
SiteEnergyUse(kBtu)	9.358619
TotalGHGEmissions	13.422347
BuildingAge	0.275485
NumberOfActivities	1.962929
ResidentialSection	1.939626
AverageFloorArea	6.118060
Parking_ratio	2.269647
LargestPropertyUseTypeGFA_ratio	6.205282

MODELISATION

Objectif :

- construire des modèles prédictifs capables d'estimer avec précision la consommation énergétique des bâtiments et leurs émissions de gaz à effet de serre

Test de 9 modèles de régression	3 métriques d'évaluation
• Régression linéaire, Ridge, Lasso, ElasticNet, SVR, Bagging, Boosting et Stacking	• RMSE, MAE et R^2



Modèles linéaires

Régression linéaire

- établit une relation linéaire entre des variables indépendantes et une variable dépendante.

Ridge

- méthode de régression linéaire régularisée qui inclut une pénalité L2 pour réduire la complexité du modèle et prévenir le surajustement.

Lasso

- technique de régression linéaire qui utilise une pénalité L1 pour encourager la parcimonie des coefficients, permettant ainsi la sélection de caractéristiques.

ElasticNet

- combine les pénalités L1 et L2 pour bénéficier à la fois de la sélection de caractéristiques du Lasso et de la régularisation de Ridge.

Modèles ensemblistes

Bagging

Random Forest Regressor :
algorithme d'ensemble qui construit de nombreux arbres de décision indépendants en échantillonnant aléatoirement des sous-ensembles de données et en moyennant leurs prédictions pour une meilleure robustesse et précision.

Boosting

AdaBoost :
ajuste les poids des instances de données pour se concentrer sur les prédictions difficiles, améliorant ainsi la performance du modèle.

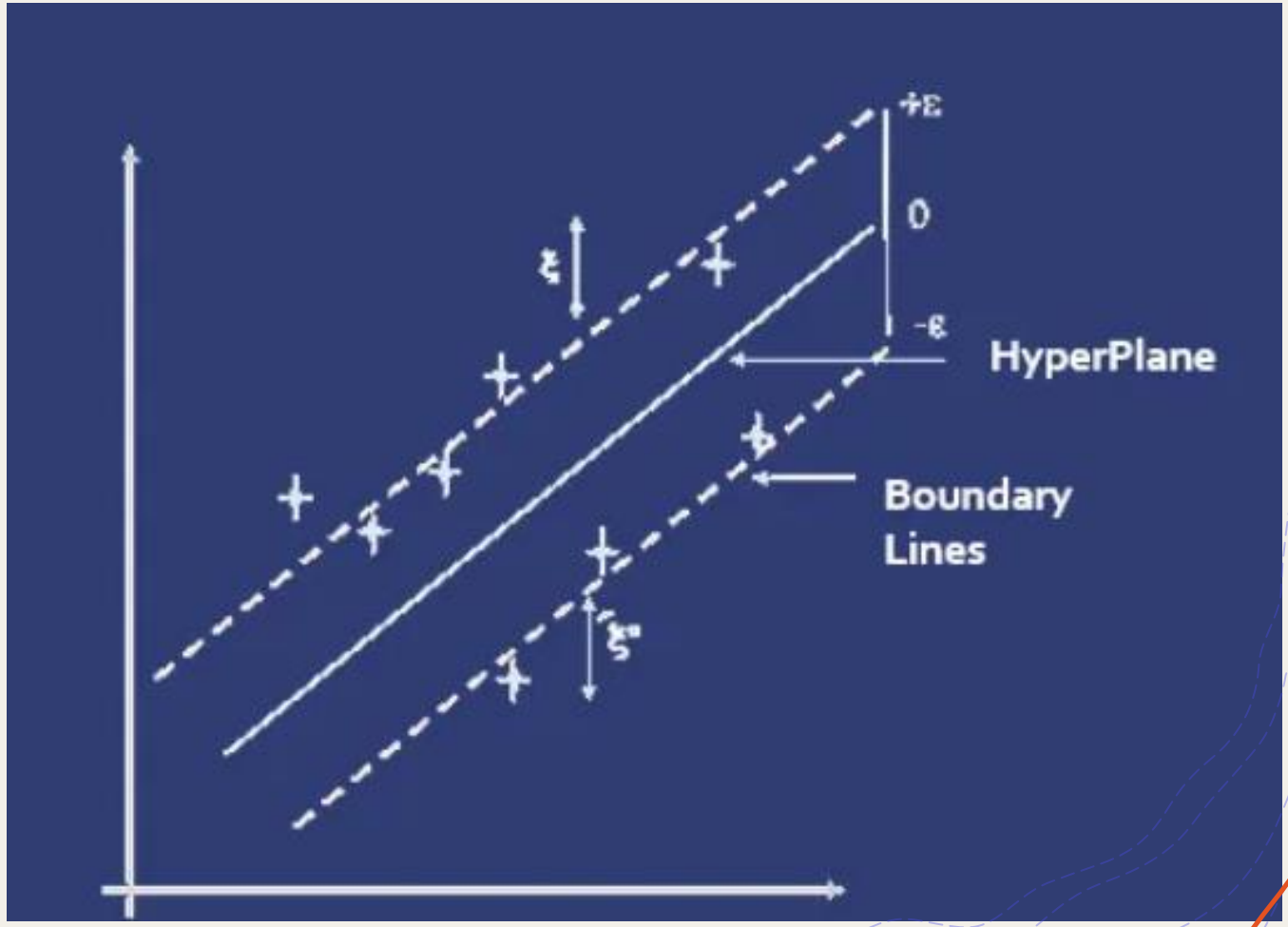
XgBoost :
algorithme d'optimisation de gradient boosting efficace et évolutif qui utilise des arbres de décision avancés et des techniques de régularisation pour des performances de prédiction supérieures.

les machines à vecteurs de support

+ modélise des relations complexes entre les variables en trouvant l'hyperplan qui s'adapte le mieux aux données

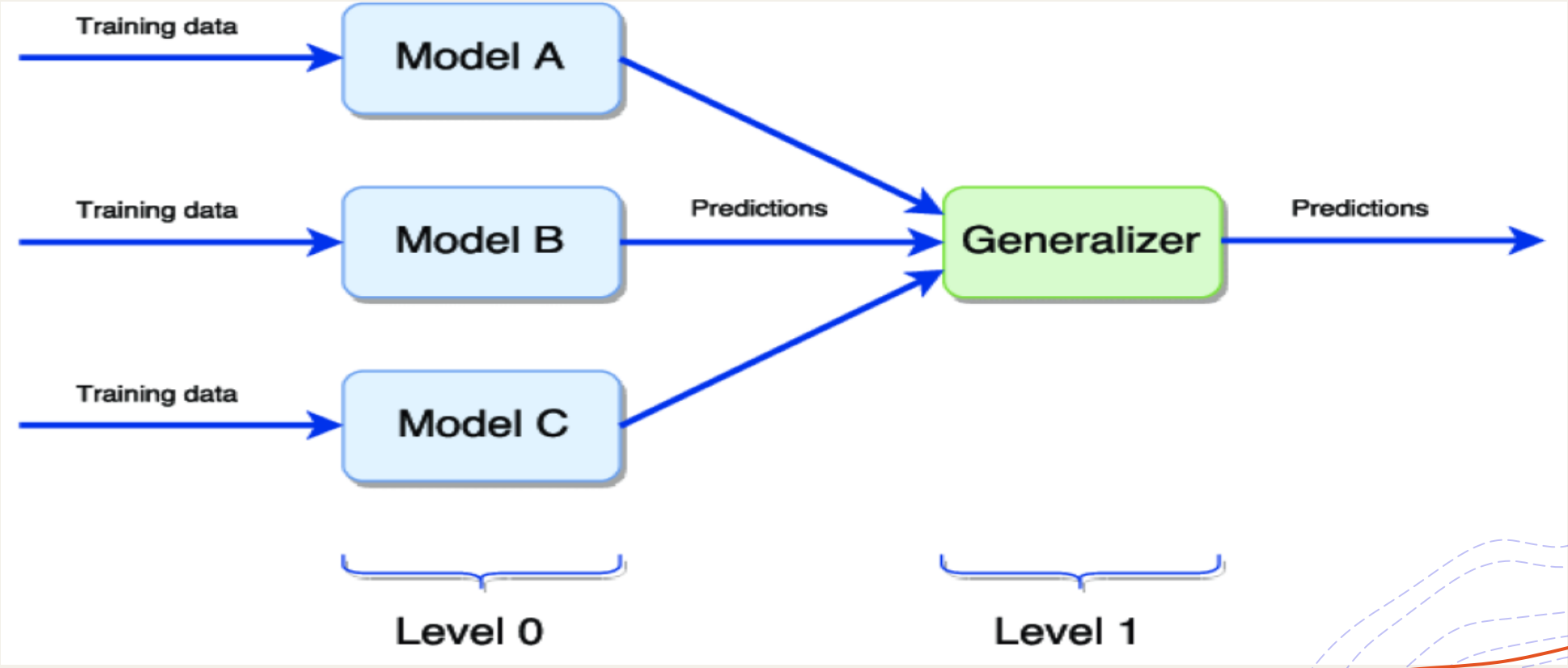
+ en optimisant l'équilibre entre la maximisation de la marge entre les différentes catégories et la minimisation de l'erreur de prédiction,

+ souvent avec l'utilisation de noyaux pour gérer les espaces non linéaires.



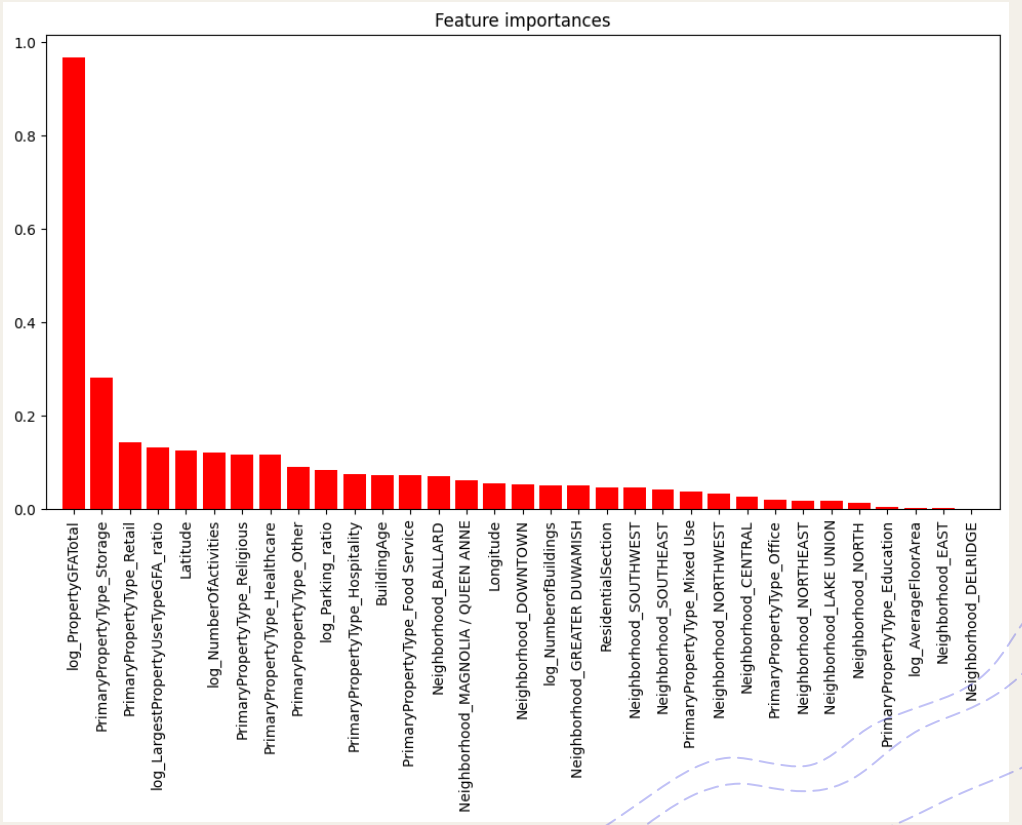
Modèle ensembliste : stacking

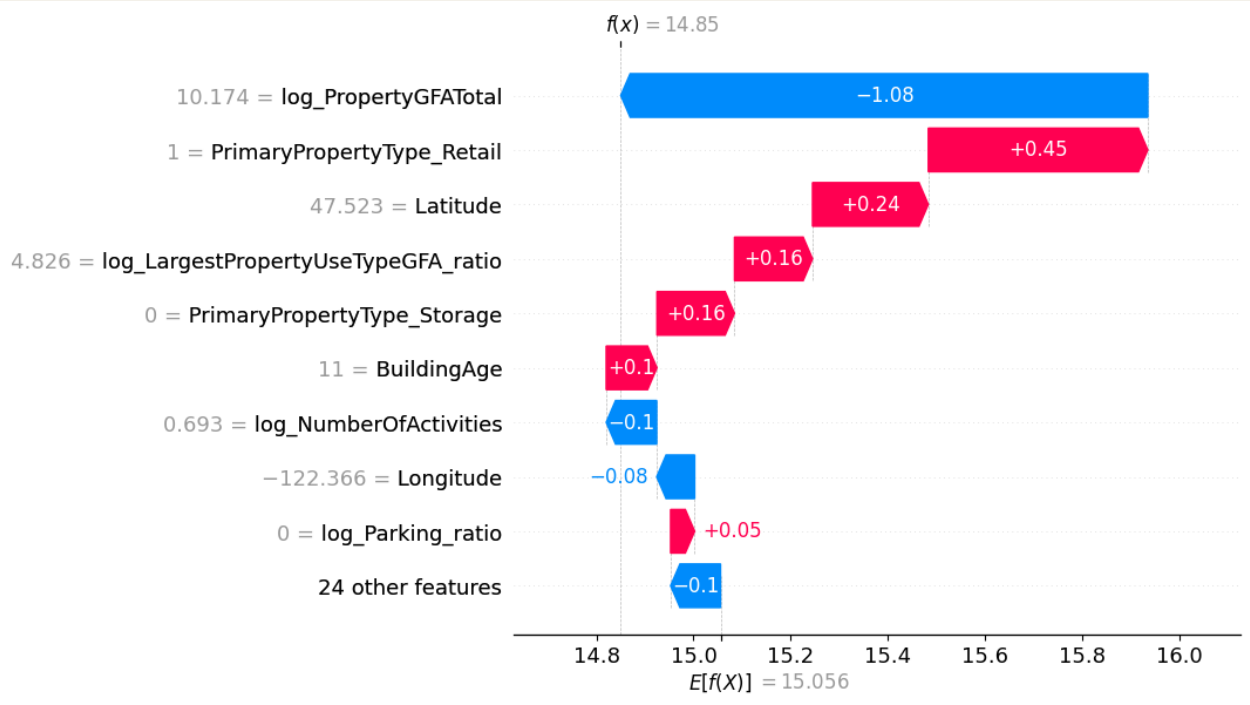
+ technique d'ensemble qui combine les prédictions de plusieurs modèles d'apprentissage machine pour produire une prédiction finale plus précise en utilisant un méta-modèle.



Résultat : prédiction de l'énergie

	Model	MAE	RMSE	R2	Execution Time (s)
0	Linear Regression	0.563448	0.761764	0.671344	0.014793
1	Linear Regression avec StandardScaler	0.562861	0.761374	0.671681	0.039084
2	Lasso	0.632512	0.856731	0.584292	0.032191
3	Lasso avec StandardScaler	1.042843	1.330878	-0.003174	0.050717
4	Ridge	0.567981	0.771881	0.662557	0.025717
5	Ridge avec StandardScaler	0.561220	0.759781	0.673053	0.036335
6	Elastic Net	0.595286	0.810281	0.628147	0.020607
7	Elastic Net avec StandardScaler	0.559302	0.764567	0.668922	0.014118
8	Bagging RandomForest	0.581304	0.796060	0.641085	5.564466
9	Boosting AdaBoost	0.593726	0.800912	0.636697	0.470575
10	Boosting XGBoost	0.576893	0.792084	0.644661	0.813713
11	SVR	0.616461	0.833584	0.606451	0.075619
12	SVR avec StandardScaler	0.561890	0.768238	0.665734	0.066088
13	Stacking Regressor	0.566901	0.766499	0.667247	0.861994





Conclusion :

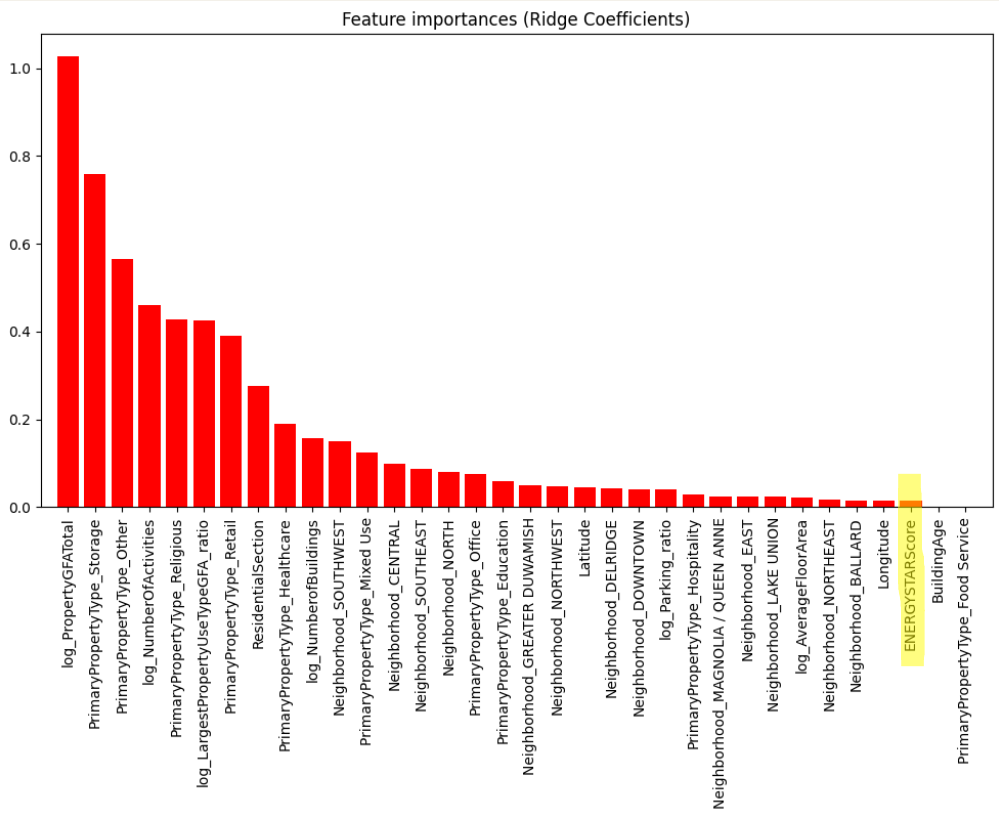
Choix du Modèle Optimal: le modèle Ridge avec StandardScaler

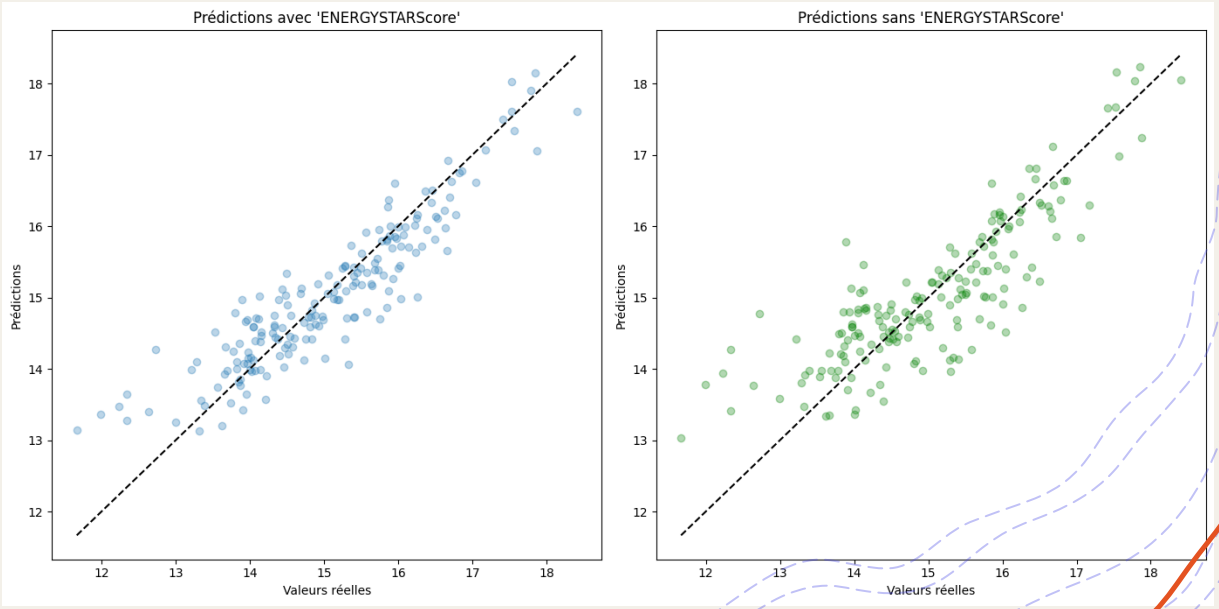
Importance des Caractéristiques: La surface totale de la propriété est la caractéristique la plus influente, suivie par le type de propriété et la latitude

Analyse SHAP : L'outil SHAP révèle que la surface totale et le type 'Retail' sont les principaux facteurs influençant la consommation d'énergie, avec des impacts respectivement négatif et positif

ENERGYSTAR SCORE

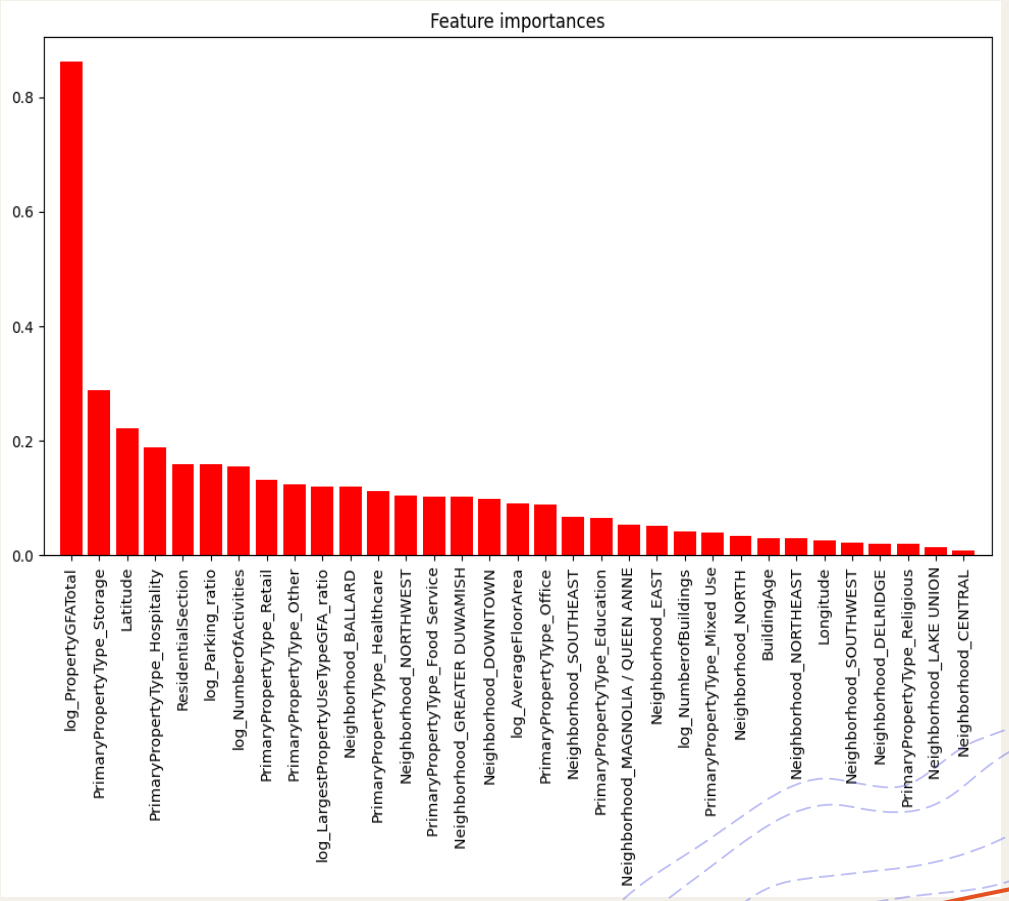
- Amélioration des métriques : poids sur le modèle
- Facteur non dominant dans les feature importances

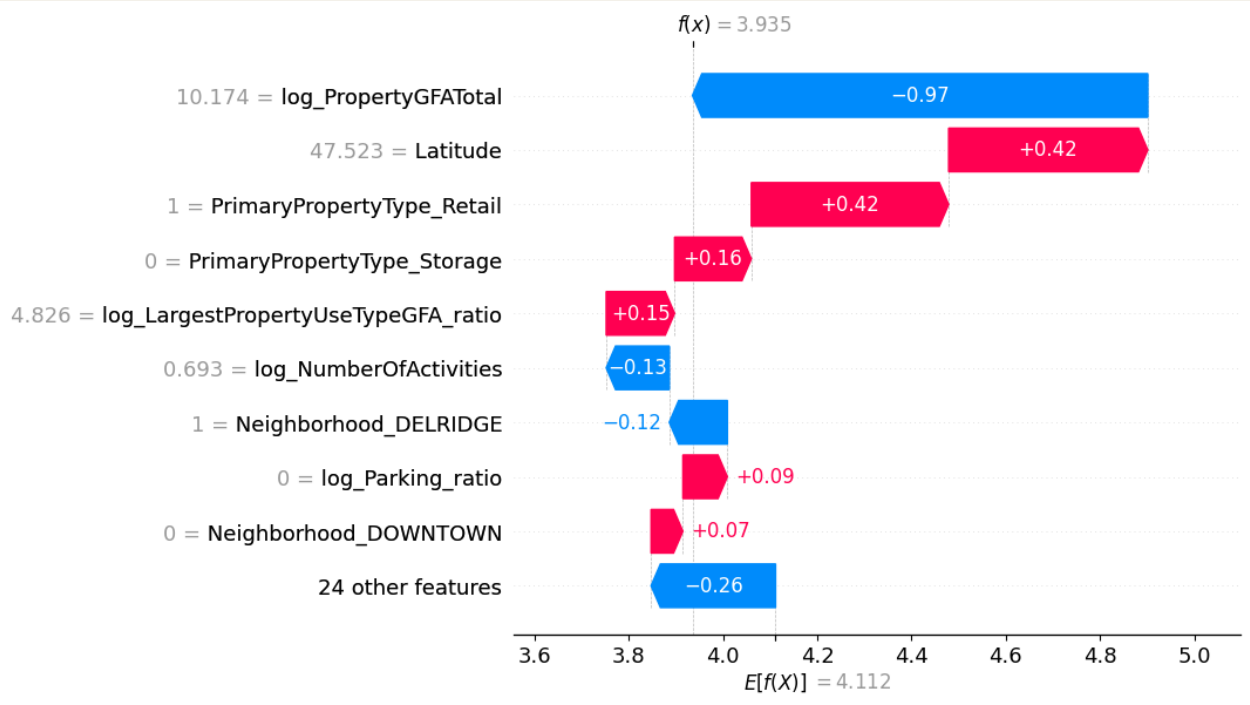




Résultat : prédiction de l'émission CO^2

	Model	MAE	RMSE	R2	Execution Time (s)
0	Linear Regression	0.785890	0.996729	0.525216	0.078123
1	Linear Regression avec StandardScaler	0.789977	0.999048	0.523004	0.042225
2	Lasso	0.892204	1.111732	0.409334	0.035588
3	Lasso avec StandardScaler	1.113083	1.446718	-0.000253	0.014810
4	Ridge	0.794958	1.005764	0.516569	0.038192
5	Ridge avec StandardScaler	0.785249	0.995562	0.526327	0.019953
6	Elastic Net	0.837673	1.052042	0.471058	0.010024
7	Elastic Net avec StandardScaler	0.797518	1.005856	0.516480	0.034443
8	Bagging RandomForest	0.798185	0.997736	0.524256	5.451461
9	Boosting AdaBoost	0.815678	1.008776	0.513669	0.653112
10	Boosting XGBoost	0.795544	0.998225	0.523789	0.234867
11	SVR	0.838406	1.070150	0.452692	0.053003
12	SVR avec StandardScaler	0.787222	1.020864	0.501944	0.062347
13	Stacking Regressor	0.798404	1.001630	0.520535	0.405630





Conclusion :

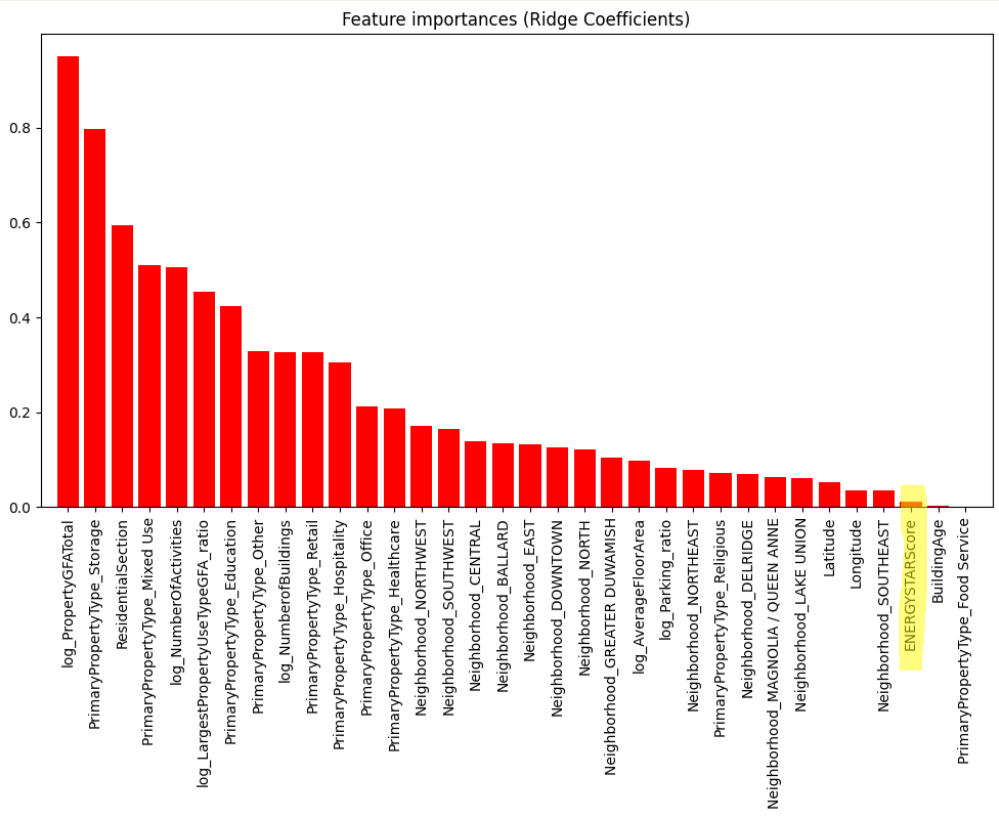
Choix du Modèle Optimal: le modèle Ridge avec StandardScaler

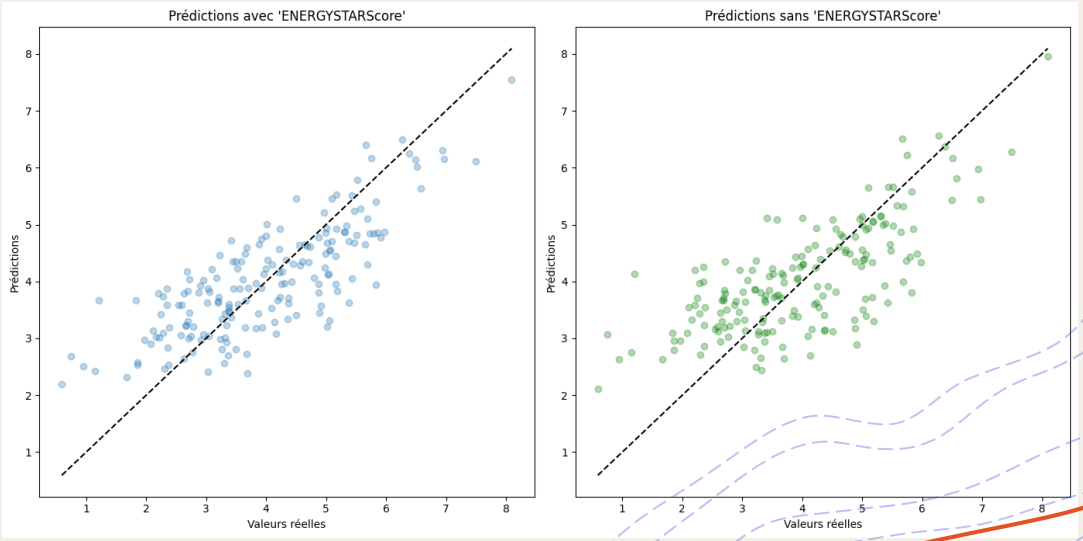
Importance des Caractéristiques: La surface totale de la propriété est la caractéristique la plus influente, suivie par le type de propriété et le nombre d'activités

Analyse SHAP : L'outil SHAP révèle que la surface totale, la latitude et le Type 'Retail' sont les principaux facteurs influençant la consommation d'énergie, avec des impacts respectivement négatif et positif

ENERGYSTAR SCORE

- Amélioration des métriques : poids sur le modèle
- Facteur non dominant dans les featureq importances





CONCLUSION

Prédiction:

- modèle unique pour les deux targets à prédire
- Avantages : cohérence, efficacité et simplification de la mise en production.

ENERGYSTARScore :

- améliore la précision des prédictions.
- Facteur non déterminant seul

Optimisation Énergétique :

- Potentiel d'utilisation du modèle
- Recommandations d'actions ciblées sur l'énergie

Seattle

Merci pour votre attention !