Projet 3 en Ingénieur Machine learning

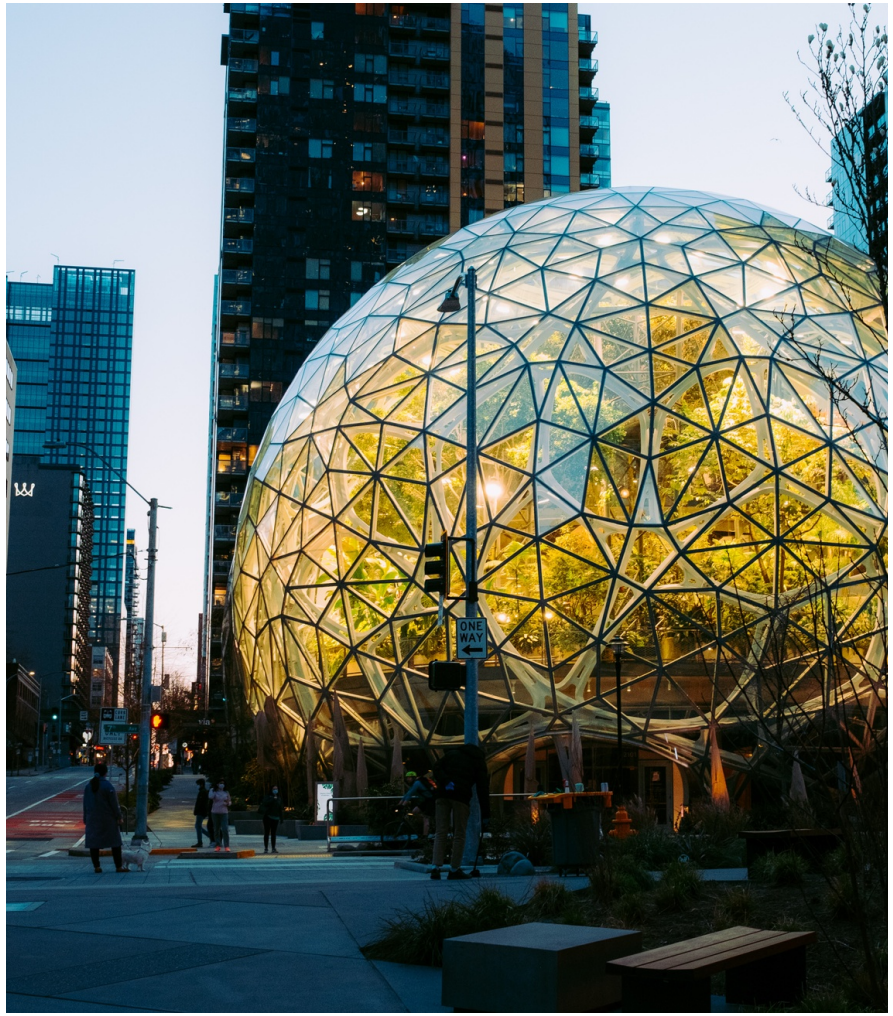
Anticipez les besoins en consommation de bâtiments

Yassine MASSOUATH

Mai 2023

Sommaire

1.Problématique et Découvertes des données



Notre projet vise à prédire les émissions de CO2 et la consommation d'énergie des bâtiments non résidentiels à Seattle, dans le cadre de l'objectif de la ville de devenir neutre en émissions de carbone d'ici 2050

1.Problématique et Découvertes des données

Pour ce faire, nous avons utilisé des données minutieuses collectées par les agents de la ville en 2016. Ces données incluent des informations sur la taille et l'usage des bâtiments, la date de construction, la situation géographique, et d'autres caractéristiques structurelles. Nous avons également examiné l'Energy Star Score, une mesure de la performance énergétique des bâtiments, pour voir comment il pourrait nous aider dans notre prédiction.

2.Préparation des données

Une fois que nous avons compris nos données, la prochaine étape consiste à les préparer pour la modélisation. Nous avons commencé par nettoyer nos données, ce qui impliquait de traiter les valeurs manquantes et aberrantes. En outre, nous avons créé de nouvelles variables à partir des données existantes pour extraire des informations supplémentaires

les différentes étapes du nettoyage ?

On changer le nombres de bâtiments 0 par 1

On supprime les bâtiments avec
High outlier et Low outlier

On Supprime les bâtiments destinés à l'habitation

les différentes étapes du nettoyage ?

On supprime les bâtiments ayant des valeurs incohérentes pour les features "SiteEnergyUse(kBtu)" et "TotalGHGEmissions"

On Supprime des variables à faible cardinalité et /ou inutiles

les différentes étapes du nettoyage ?

On conserve les variable avec kBtu : British thermal unit

On supprime les variables avec plus de 90% valeurs manquantes

Traitement d'autres valeurs manquantes

Features engineering

1. ## Création de la variable BuildingAge

La création de la variable "BuildingAge" peut permettre de capturer l'impact des variations d'efficacité énergétique et des normes de construction sur la consommation d'énergie et les émissions de CO2 en fonction de l'âge du bâtiment.

2. ## Calcule de pourcentage de chaque variable de surface par rapport au surface total.

Nous avons remarqué une colinéarité entre les différentes variables de surface de nos bâtiments. Pour résoudre ce problème, nous avons transformé ces variables en calculant le pourcentage de chaque type de surface par rapport à la surface totale du bâtiment. Cette approche réduit la colinéarité en standardisant les mesures de surface par rapport à la taille du bâtiment.

3. ## Calcule la distance de Haversine par rapport au centre de Seattle.

Le calcul de la distance Haversine par rapport au centre de Seattle peut servir à introduire une variable géographique dans votre modèle de prédiction. La distance au centre de la ville peut être liée aux émissions de CO2 et à la consommation

4. ## Passage au Log10+1 des variables cibles

La transformation log10+1 de la variable cible aide à réduire l'asymétrie, à minimiser l'effet des valeurs extrêmes, et à stabiliser la variance, ce qui peut améliorer les performances des modèles de régression.

Analyse exploratoire

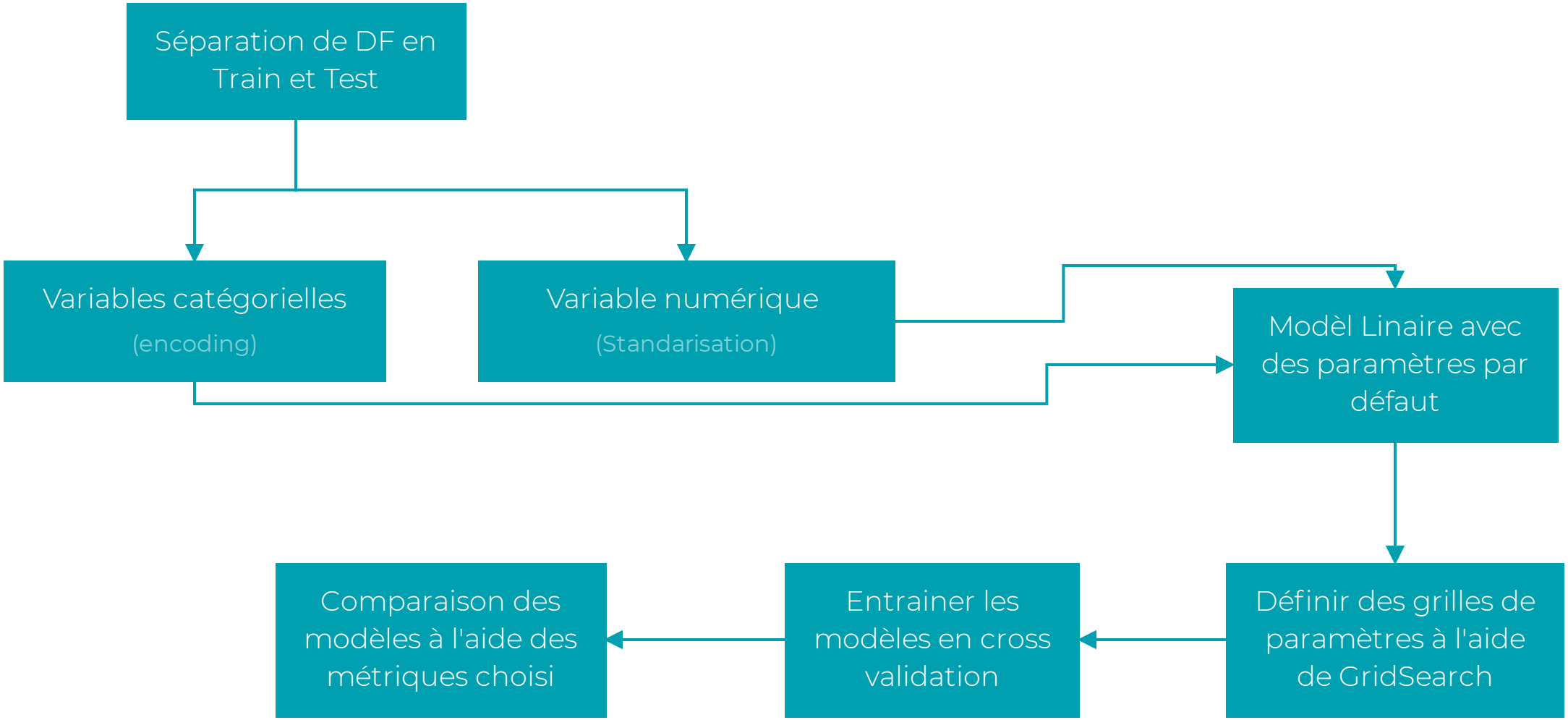
Passage au Log10+1 des variables cibles

La transformation log10+1 de la variable cible aide à réduire l'asymétrie, à minimiser l'effet des valeurs extrêmes, et à stabiliser la variance, ce qui peut améliorer les performances des modèles de régression.

3.Modélisation :

Passons maintenant à l'étape de la modélisation. Au cours de cette phase, nous avons d'abord sélectionné le modèle le plus approprié en fonction de nos données et de notre problème. Ensuite, nous avons entraîné et optimisé ce modèle sur notre ensemble de données d'entraînement, en ajustant les hyperparamètres pour améliorer ses performances. Enfin, nous avons évalué notre modèle sur l'ensemble de données de validation pour vérifier sa capacité à prédire les émissions de CO2 et la consommation d'énergie des bâtiments de manière précise et fiable.

Les étapes pour modélisation



Consomation d'energie

Cross validation

Baseline

Model	Duration	Test R2	Train R2	Test RMSE	Test MAE	Train RMSE	Train MAE
linear_regression	0.002775	0.476230	0.602240	0.356256	0.274844	0.356256	0.274844
elastic net	0.002165	0.197920	0.211854	0.501462	0.396544	0.501462	0.396544
random_forest	0.010702	0.718584	0.961364	0.111027	0.080693	0.111027	0.080693
svr	0.012171	0.666475	0.691416	0.313790	0.233596	0.313790	0.233596
XGBM	0.002828	0.701536	0.995074	0.039643	0.026562	0.039643	0.026562

Consomation d'energie

Cross validation

Régression Avec GridSearch

	modèle	Fit time	Score time	Mean train score RMSE	Mean test score RMSE	Best estimator	RMSE best estimator
0	XGBoost	0.096630	0.002572	0.926111	1.007365	{'colsample_bytree': 0.8, 'gamma': 0.2, 'learn...	1.820903e-01
1	RandomForest	0.256313	0.009872	0.205162	0.315922	{'bootstrap': False, 'max_depth': 50, 'max_fea...	7.973377e-15
2	SVR	0.012852	0.001244	0.516799	0.519855	{'C': 0.1, 'epsilon': 0, 'loss': 'squared_epsi...	3.425855e-01
3	ElasticNet	0.002880	0.001135	0.524220	0.524436	{'alpha': 0.001, 'l1_ratio': 0.8}	3.419092e-01

Consomation d'energie

Feature importance

	feature	importance
12	RatePerBuildings	0.319617
0	PrimaryPropertyType	0.144638
2	LargestPropertyUseType	0.098125
11	RatePerFloors	0.097512
8	BuildingAge	0.051914
17	haversine_distance	0.049061
13	RateLargestPropertyUseType	0.034936
3	SecondLargestPropertyUseType	0.034856
16	NumberOfAllUseTypes	0.032597
1	Neighborhood	0.028571
7	NaturalGas(kBtu)	0.022208
14	RateSecondLargestPropertyUseType	0.021866
10	RateBuilding	0.019809
9	RateParking	0.017430
4	ThirdLargestPropertyUseType	0.012011
15	RateThirdLargestPropertyUseType	0.008769
5	SteamUse(kBtu)	0.005700
6	Electricity(kBtu)	0.000380

Emission CO2

Cross validation

Basline

modèle	Fit time	Durée	Test R2	Train R2	Test RMSE	Test MAE	Train RMSE	Train MAE
linear_regression	0.014563	0.003509	0.604240	0.626138	0.375038	0.293345	0.375038	0.293345
elastic net	0.003441	0.002478	0.081483	0.087483	0.585940	0.459832	0.585940	0.459832
random_forest	0.868794	0.016102	0.705598	0.958391	0.125116	0.094720	0.125116	0.094720
svr	0.097889	0.021666	0.675576	0.725972	0.321079	0.240807	0.321079	0.240807
XGBM	0.508061	0.003400	0.681995	0.996139	0.038116	0.025848	0.038116	0.025848

Emission CO2

Cross validation

Aved GridSearch

	modèle	Fit time	Score time	Mean train score RMSE	Mean test score RMSE	Best estimator	RMSE best estimator
0	XGBoost	0.131348	0.002822	0.344157	0.447732	{'colsample_bytree': 1.0, 'gamma': 0.2, 'learn...	0.204049
1	RandomForest	0.368106	0.014467	0.234047	0.360692	{'bootstrap': False, 'max_depth': 50, 'max_fea...	0.104410
2	SVR	0.015681	0.002194	0.581311	0.584992	{'C': 0.1, 'epsilon': 0, 'loss': 'squared_epsi...	0.377069
3	ElasticNet	0.002982	0.001160	0.524220	0.524436	{'alpha': 0.001, 'l1_ratio': 0.8}	1.232268

Emission CO2

Feature Importance

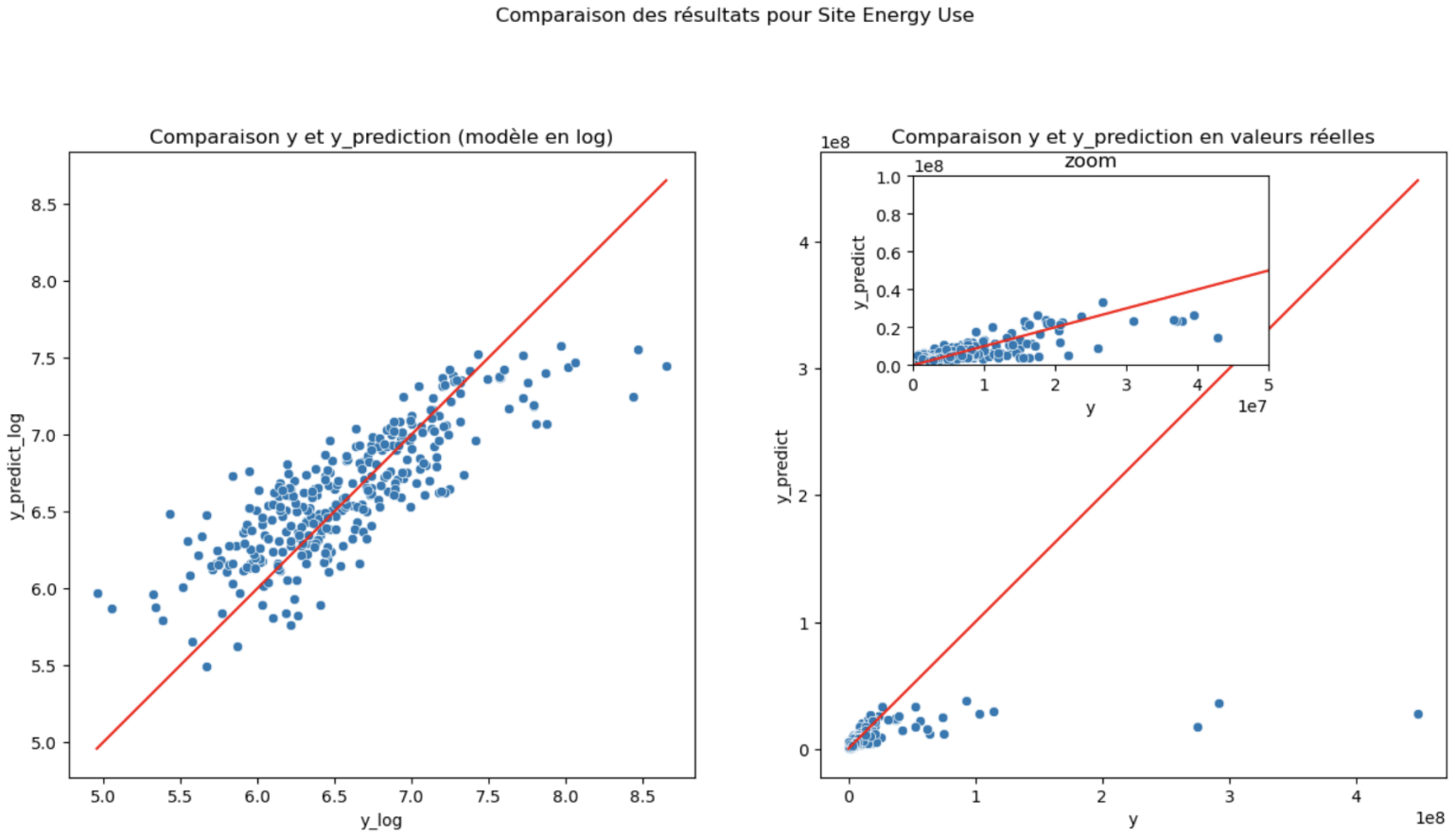
	feature	importance
12	RatePerBuildings	0.221483
7	NaturalGas(kBtu)	0.165469
0	PrimaryPropertyType	0.145644
2	LargestPropertyUseType	0.106941
11	RatePerFloors	0.087114
5	SteamUse(kBtu)	0.040249
17	haversine_distance	0.039354
8	BuildingAge	0.032212
13	RateLargestPropertyUseType	0.027803
16	NumberOfAllUseTypes	0.026267
1	Neighborhood	0.024099
3	SecondLargestPropertyUseType	0.022460
14	RateSecondLargestPropertyUseType	0.019650
10	RateBuilding	0.012704
9	RateParking	0.010664
4	ThirdLargestPropertyUseType	0.010165
15	RateThirdLargestPropertyUseType	0.007723
6	Electricity(kBtu)	0.000000

Choix de modèl

D'après les résultats partagés, le modèle RandomForest se révèle être le plus performant parmi les différents modèles testés. En particulier, il obtient le score RMSE (Root Mean Squared Error) le plus faible sur les ensembles d'entraînement et de test, ce qui indique une précision de prédiction supérieure. De plus, l'écart relativement faible entre les scores d'entraînement et de test suggère que le modèle n'est pas en surapprentissage, c'est-à-dire qu'il n'est pas excessivement adapté aux données d'entraînement au détriment de sa capacité à généraliser à de nouvelles données. Bien que le temps d'apprentissage pour RandomForest soit légèrement plus long que pour les autres modèles, la différence est compensée par une meilleure performance globale.

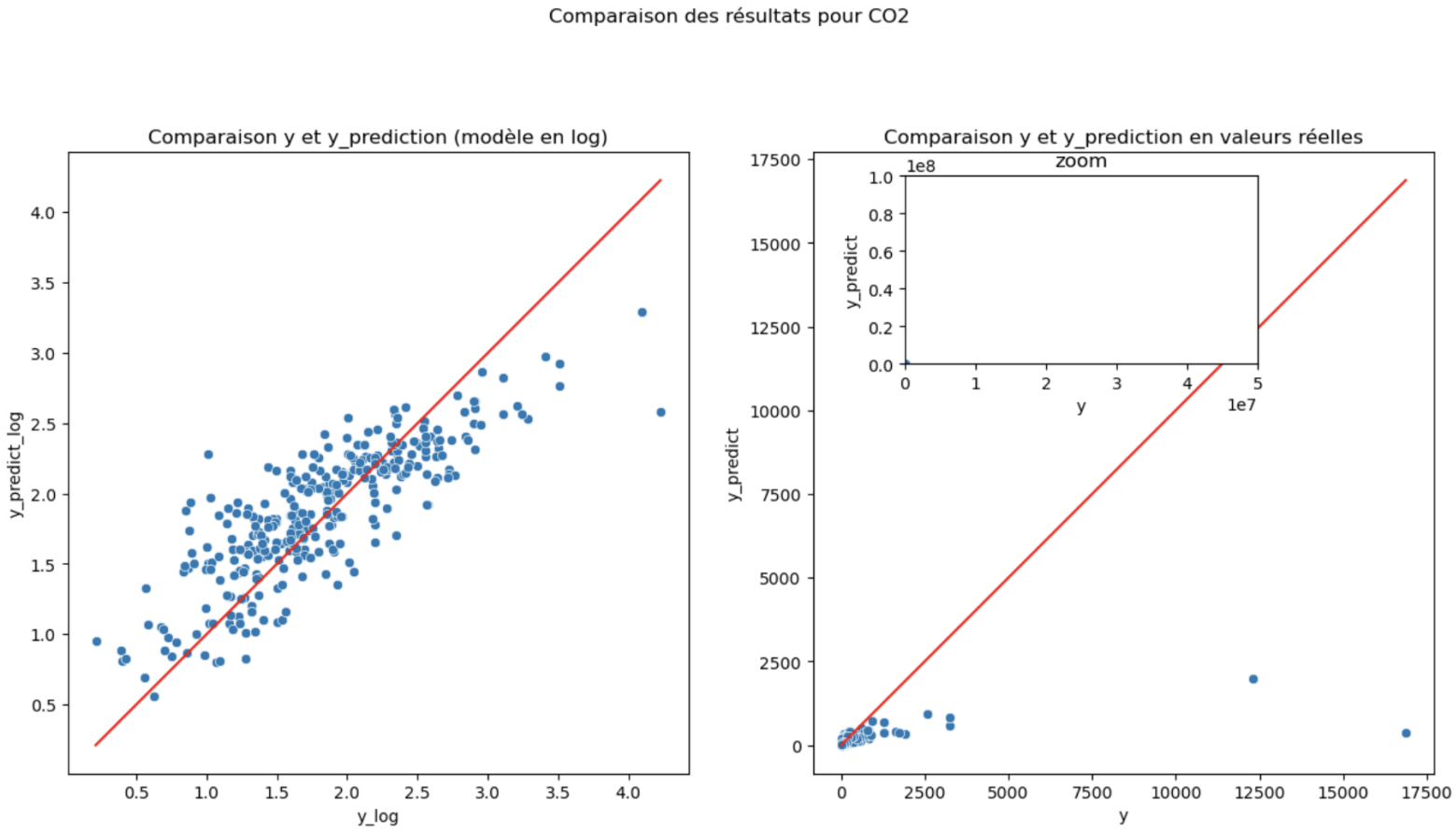
Consomation d'energie

Analyse des prédictions



Emission CO2

Analyse des prédictions



3.Evaluation l'intérêt de EnergyScore :

Enfin, nous avons évalué l'utilité du score EnergyStar dans notre modèle. Le score EnergyStar est un indicateur de l'efficacité énergétique d'un bâtiment, mais il est coûteux et fastidieux à calculer. Nous avons donc voulu savoir si l'inclure dans notre modèle améliorait réellement les prédictions. Pour cela, nous avons comparé la performance de notre modèle avec et sans le score EnergyStar

Emission CO2

Intérêt de EnergyScore

Comparaison des métriques d'evaluation RMSE

	modèle	Fit time	Score time	Mean train score RMSE	Mean test score RMSE	Best estimator	RMSE best estimator
0	RandomForest	0.380921	0.009713	0.247432	0.3629	{'bootstrap': True, 'max_depth': 10, 'max_feat...	0.228975

	modèle	Fit time	Score time	Mean train score RMSE	Mean test score RMSE	Best estimator	RMSE best estimator
0	RF_ES_Co2	0.40741	0.009819	0.239156	0.350603	{'bootstrap': True, 'max_depth': 20, 'max_feat...	0.15271

Consommation d'energie

Intérêt de EnergyScore

Comparaison des métriques d'evaluation RMSE

	modèle	Fit time	Score time	Mean train score RMSE	Mean test score RMSE	Best estimator	RMSE best estimator
0	RandomForest	0.380921	0.009713	0.247432	0.3629	{'bootstrap': True, 'max_depth': 10, 'max_feat...	0.228975

	modèle	Fit time	Score time	Mean train score RMSE	Mean test score RMSE	Best estimator	RMSE best estimator
0	RF_ES_Co2	0.40741	0.009819	0.239156	0.350603	{'bootstrap': True, 'max_depth': 20, 'max_feat...	0.15271

Conclusion

Pour conclure, notre modèle basé sur la méthode Random Forest a réussi à prédire de manière relativement précise les émissions de CO2 et la consommation d'énergie des bâtiments de Seattle. L'inclusion du score EnergyStar, bien qu'il soit coûteux et fastidieux à calculer, a légèrement amélioré la performance de notre modèle. Cependant, il convient de peser cette amélioration par rapport au coût de calcul du score EnergyStar.

Pour améliorer davantage notre modèle à l'avenir, nous pourrions envisager plusieurs pistes :

Exploration de modèles plus sophistiqués : Bien que le modèle Random Forest ait donné de bons résultats, il existe d'autres modèles, comme les réseaux de neurones, qui pourraient potentiellement offrir une meilleure performance.

Enrichissement des données : Si possible, nous pourrions chercher à obtenir plus de données ou à intégrer d'autres types de données, comme des informations plus détaillées sur l'utilisation des bâtiments.

Amélioration de feature ingeniring : Nous pourrions explorer d'autres transformations de nos caractéristiques existantes, ou créer de nouvelles caractéristiques à partir des données existantes.

Optimisation des hyper paramètres : Nous pourrions consacrer plus de temps à l'optimisation des hyper paramètres de notre modèle pour améliorer ses performances.