Anticipez les besoins en consommation de bâtiments

Projet 4 du parcour
Data Scientist

Dernière MàJ
12 février 2022

Raquel Sanchez Pellicer

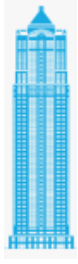
OPENCLASSROOMS

Problématique

Données

Démarche :

- EDA, Nettoyage et Feature engineering
- Modélisation Consommation E. et Émissions CO_2
- Évaluation intérêt ‘Energy Star score’



Conclusion

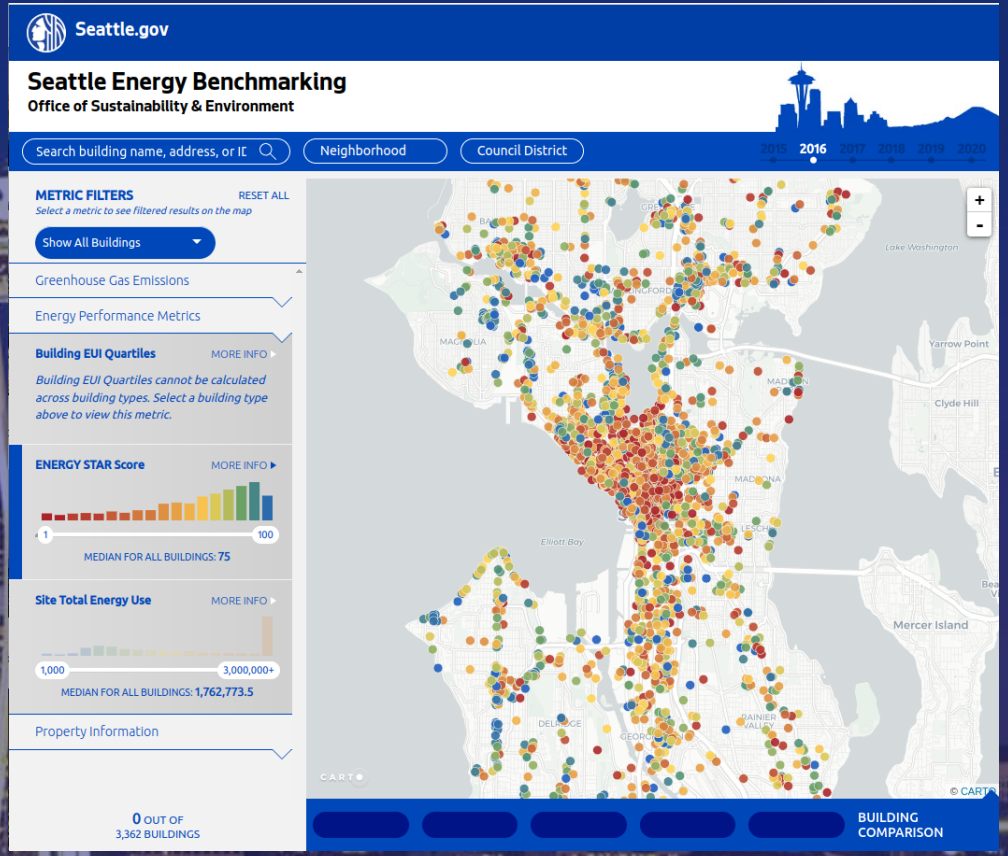
Problématique : **Contexte**

La ville de **Seattle** a adopté un plan d'action climatique visant à atteindre des **émissions nettes** de gaz à effet de serre (GES) **nulles d'ici 2050**.

Le programme exige aux propriétaires de **bâtiments non résidentiels et multifamiliaux** de faire le **suivi** de leurs **performances énergétiques et** de **présenter un rapport annuel** à la ville de Seattle.

Des **relevés** minutieux ont été effectués par les agents de la ville dès **2015**.

Ces relevés sont **coûteux** à obtenir.



https://www.seattle.gov/energybenchmarkingmap/#seattle/2020?layer=total_ghg_emissions&sort=total_ghg_emissions&order=desc&lat=47.61&lng=-122.33&zoom=14

Problématique : **Objectif**

City of Seattle

Trouver une solution pour éviter les relevés annuels de consommation
évaluer l'intérêt de l'"ENERGY STAR Score"

1. **Prédire** la **consommation totale d'énergie** de bâtiments non destinés à l'habitation

2. **Prédire** les **émissions de CO_2** des bâtiments non destinés à l'habitation

3. **Évaluer l'intérêt de l'"ENERGY STAR Score"** pour la prédiction d'émissions

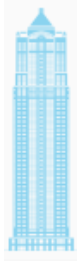
ENERGY STAR

Problématique

Données

Démarche :

- EDA, Nettoyage et Feature engineering
- Modélisation Consommation E. et Émissions CO_2
- Évaluation intérêt ‘Energy Star score’



Conclusion

Données

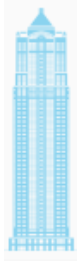
Données 2016	
N° colonnes : 46	N° lignes : 3376
Information général sur les propriétés: 14	Nom ; Adresse ; lat / long ; ID ; Année de construction ; Quartier ...
Nombre de bâtiments et nombre d'étages : 2	Si la propriété est déclarée comme campus ; plusieurs bâtiments peuvent être inclus dans un rapport
Surface des propriétés: 3	Surface totale ; Surface Parking ; Surface batîments
Utilisation des propriétés: 8	Listes d'utilisations ; détails sur les trois utilisations principales et la surface dédiée à ces utilisations
Consommation énergie : 11	Consommation annuelle de différents types d'énergie ; Énergie annuelle totale ; Énergie annuelle totale normalisée en fonction des conditions météorologiques
Émissions gaz à effet serre : 2	La quantité totale d'émissions GES ; Émissions en fonction de la superficie totale
Energy Star Score : 2	Score année 2016 ; Années où la propriété a reçu la certification
Autres variables : 4	Commentaires ; Outlier ; Default data ; Compliance status

Problématique

Données

Démarche :

- **EDA, Nettoyage et Feature engineering**
- **Modélisation Consommation E. et Émissions CO_2**
- **Évaluation intérêt ‘Energy Star score’**



Conclusion

Démarche :

City of Seattle

EDA et Nettoyage
Feature engineering

Data-set brut

Sélection des variables
Analyse exploratoire
Transformation variables

Data-set pour ML

Info . pertinente
0 données manquants
Corrélation réduite

Prédiction (Évaluation performance)

New dataset

MODÈLE FINAL

Prédictions

Modélisation
(Émissions, Consommation ; Évaluation 'Energy Star score')

Pre-processing
Split

Train-set

Test-set

Sélection de Modèles à tester

Résultats 1er entraînement

Validation

Sélection des modèles les plus performants

Optimisation

Entraînement modèles optimisés

Comparaison et sélection

Validation modèles optimisés

Démarche : **Analyse Exploratoire de Données, Nettoyage et Feature Engineering**

City of Seattle

Analyse métier

Misse en évidence des données pertinents et relevants pour la modélisation des émissions de GES et de la consommation d'énergie.

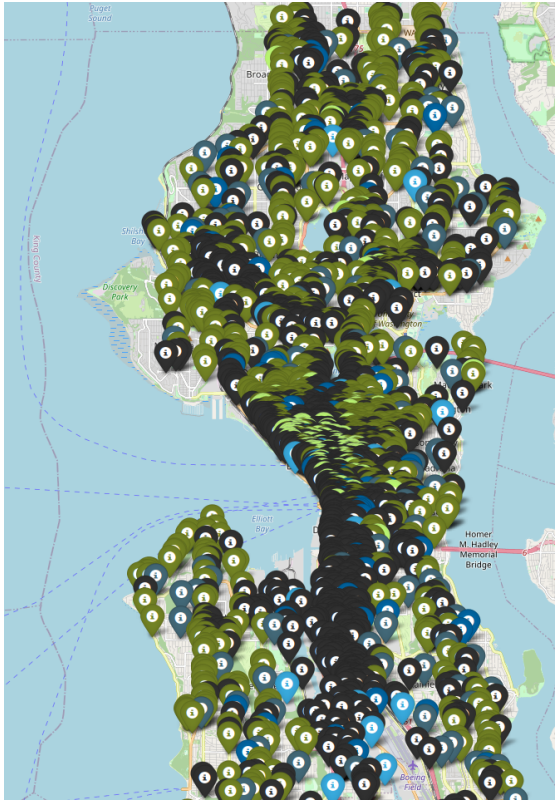
Compréhension des variables.

Identification des variables à modéliser:

- **TotalGHGEmissions :** Émissions de gaz à effet de serre
- **SiteEnergyUseWN(kBtu) :** Consommation annuelle énergie normalisée en fonction des conditions météorologiques (WN)

Identification des variables non relevantes / pertinentes :

- Variables non relevantes : ex. année de régistre des données, ville, état
- Variables non pertinentes : ex. ID évaluation fiscale
- Variables vides : ex. comments



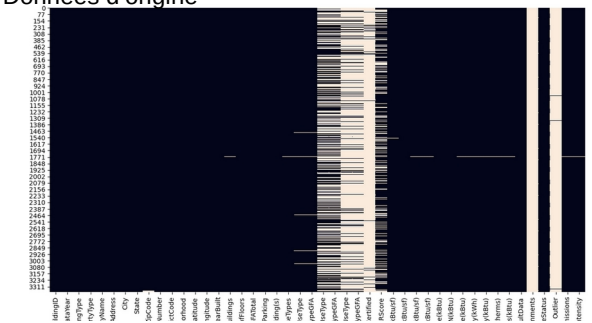
Carte de la distribution géographique des biens considérés par la base de données

Démarche : **Analyse Exploratoire de Données, Nettoyage et Feature Engineering**

City of Seattle

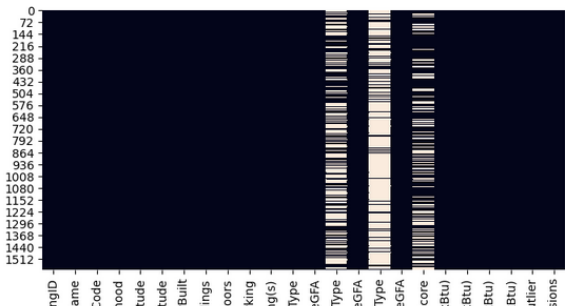
Nettoyage

Données d'origine



Building type = 'Multifamily'
&
List of all use types contains 'Multifamily'

Données après nettoyage



Actions	Dimensions dataset
Dataset source : 12,85 % données manquants	Dimensions : 3376 lignes , 46 colonnes
Filtre des bâtiments non résidentiels	Dimensions : 1701 lignes , 46 colonnes
Harmonisation (orthographe) :	Neigborhood
Vérification doublons	
Gestion des NaN	Attribution valeur après recherche internet Attribution valeur déduit à partir du dataset
Gestion valeurs aberrantes	Vérification recherche internet
Vérification variables corrélées	Sup. Totale = Parking + Buildings Utilisation 1ère, 2nde, 3rs / surfaces
Suppression des lignes ne satisfaisant pas les exigences du benchmarking et les lignes sans information pour les variables cible	Dimensions : 1579 lignes , 46 colonnes
Suppression des variables non relevantes / pertinentes	Dimensions : 1579 lignes, 24 colonnes

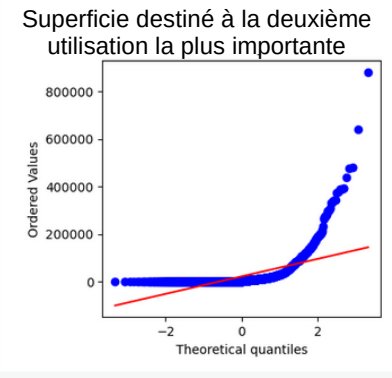
Démarche : **Analyse Exploratoire de Données, Nettoyage et Feature Engineering**

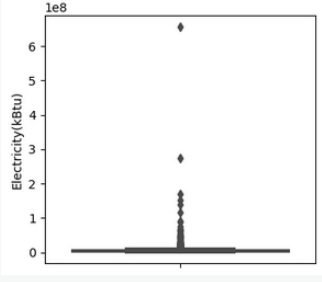
City of Seattle

Exploration : Analyses univariées

Variables quantitatives

- D'après le test D'Agostino et l'analyse graphique ; ces variables ne présentent pas une distribution normale
- Le unités et amplitudes des distributions sont très différents; le variables doivent être standardisées
- Certaines variables présentent des valeurs aberrants pouvant rendre la modélisation difficile, un filtre sera appliqué

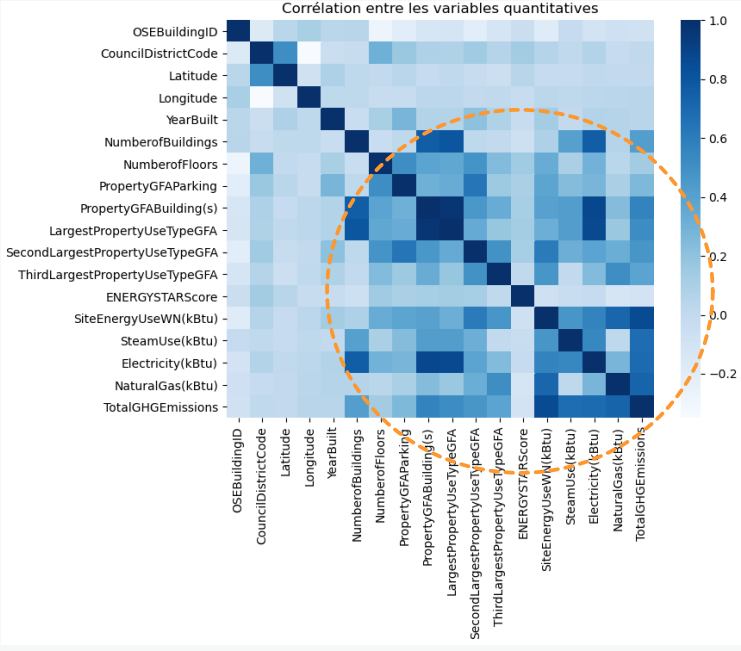




Exploration : Analyses multivariées

Corrélation entre variables quantitatives

- Forte corrélation entre certaines variables



Démarche : **Analyse Exploratoire de Données, Nettoyage et Feature Engineering**

City of Seattle

Feature Engineering

Dimensions dataset avant feature Engineering : 1579 lignes, 24 colonnes

Variable initial		Variable final
SiteEnergyUseWN(kBtu) *TotalGHGEmissions*	Suppression des valeurs aberrants, sélection d'un seuil	*SiteEnergyUseWN(kBtu)* *TotalGHGEmissions*
SteamUse(kBtu) *Electricity(kBtu)* *NaturalGas(kBtu)*	Transformation en pourcentage sur le total de la consommation	*SteamUse%* *Electricity%* *NaturalGas%*
YearBuilt	Calcul de l'âge de la propriété l'année de collecte des données	*PropertyAge*
CouncilDistrict	Création de variables dummy	*CouncilDistrict_1* à *CouncilDistrict_7*
LargestPropertyUseType *SecondLargestPropertyUseType* *ThirdLargestPropertyUseType* *LargestPropertyUseTypeGFA* *SecondLargestPropertyUseTypeGFA* *ThirdLargestPropertyUseTypeGFA*	Homogénéisation et réduction des termes. Création d'une colonne par utilisation. Les surfaces dédiés à l'utilisation principal, secondaire et tertiaire dévienent les valeurs dans les colonnes par utilisation.	35 colonnes indicant des utilisations differentes et la surface dédié à l'utilisation dans la propriété

Dimensions dataset après feature Engineering : 1572 lignes, 56 colonnes

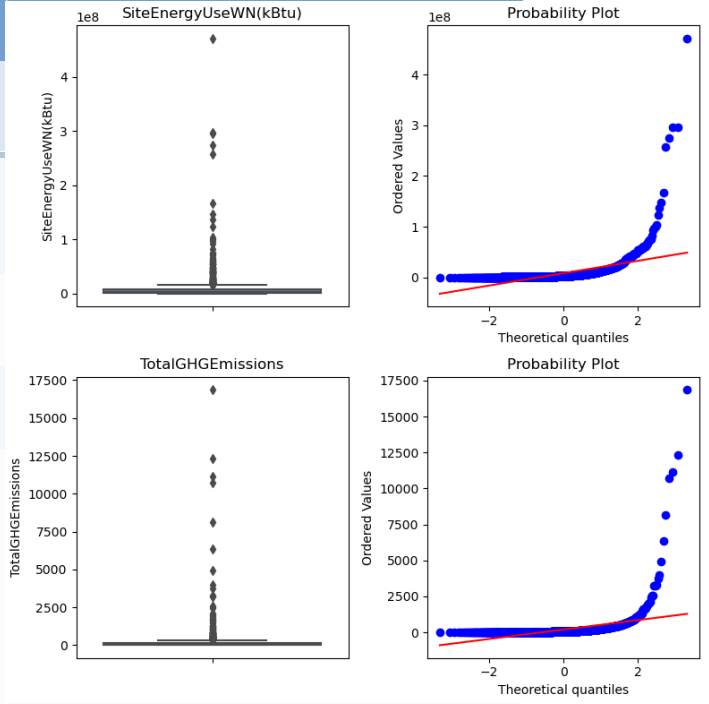
Démarche : **Analyse Exploratoire de Données, Nettoyage et Feature Engineering**

City of Seattle

Feature Engineering

Dimensions dataset avant feature Engineering : 1579 lignes, 24 colonnes

Variable initial	
SiteEnergyUseWN(kBtu) *TotalGHGEmissions*	Suppression des valeurs aberrants, sélection d'un seuil
SteamUse(kBtu) *Electricity(kBtu)* *NaturalGas(kBtu)*	Transformation en pourcentage sur le total de la consommation
YearBuilt	Calcul de l'âge de la propriété l'année de collecte des données
CouncilDistrict	Création de variables dummy
LargestPropertyUseType *SecondLargestPropertyUseType* *ThirdLargestPropertyUseType* *LargestPropertyUseTypeGFA* *SecondLargestPropertyUseTypeGFA* *ThirdLargestPropertyUseTypeGFA*	Homogénéisation et réduction des termes. Création d'une colonne par utilisation. Les surfaces dédiés à l'utilisation principal, secondaire et tertiaire dévienent les valeurs dans les colonnes par utilisation.



Dimensions dataset après feature Engineering : 1572 lignes, 56 colonnes

Démarche : **Analyse Exploratoire de Données, Nettoyage et Feature Engineering**

City of Seattle

Feature Engineering

Dimensions dataset avant feature Engineering : 1579 lignes, 24 colonnes

Variable initial		Variable final
SiteEnergyUseWN(kBtu) *TotalGHGEmissions*	Suppression des valeurs aberrants, sélection d'un seuil	*SiteEnergyUseWN(kBtu)* *TotalGHGEmissions*
SteamUse(kBtu) *Electricity(kBtu)* *NaturalGas(kBtu)*	Transformation en pourcentage sur le total de la consommation	*SteamUse%* *Electricity%* *NaturalGas%*
YearBuilt	Calcul de l'âge de la propriété l'année de collecte des données	*PropertyAge*
CouncilDistrict	Création de variables dummy	*CouncilDistrict_1* à *CouncilDistrict_7*
LargestPropertyUseType *SecondLargestPropertyUseType* *ThirdLargestPropertyUseType* *LargestPropertyUseTypeGFA* *SecondLargestPropertyUseTypeGFA* *ThirdLargestPropertyUseTypeGFA*	Homogénéisation et réduction des termes. Création d'une colonne par utilisation. Les surfaces dédiés à l'utilisation principal, secondaire et tertiaire dévienent les valeurs dans les colonnes par utilisation.	35 colonnes indicant des utilisations differentes et la surface dédié à l'utilisation dans la propriété

Dimensions dataset après feature Engineering : 1572 lignes, 56 colonnes

Démarche :

City of Seattle

EDA et Nettoyage
Feature engineering

Data-set brut

Sélection des variables
Analyse exploratoire
Transformation variables

Data-set pour ML

Info . pertinente
0 données manquants
Corrélation réduite

Prédiction (Évaluation performance)

New dataset

MODÈLE FINAL

Prédictions

Modélisation
(Émissions, Consommation ; Évaluation 'Energy Star score')

Pre-processing
Split

Train-set

Test-set

Apprentissage
Modèles à tester

Résultats 1er entraînement

Sélection des modèles les plus performants

Apprentissage
modèles
optimisés

Comparaison et sélection

Optimisation

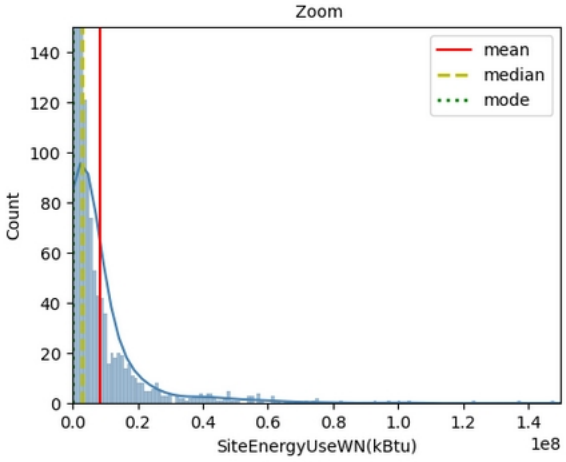
Validation

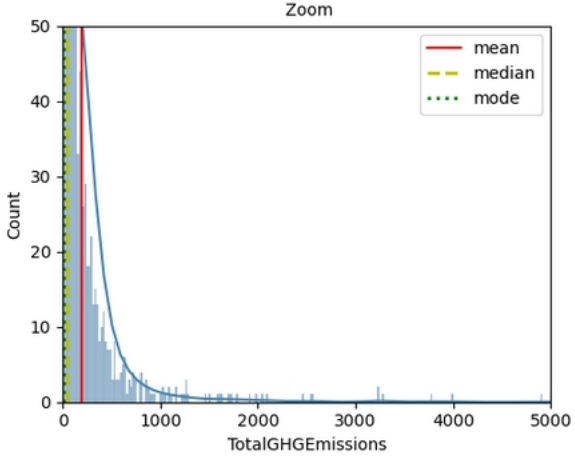
Validation modèles optimisés

Démarche : **Modélisation Émissions et Consommation**

City of Seattle

Pre-processing : Stratification





Les variables cibles présentent un biais important vers des valeurs élevées (skewness élevée).
Afin de favoriser la distribution homogène des valeurs, nous créons une nouvelle variable catégorielle indiquant à quel quartile appartient l'échantillon en question.

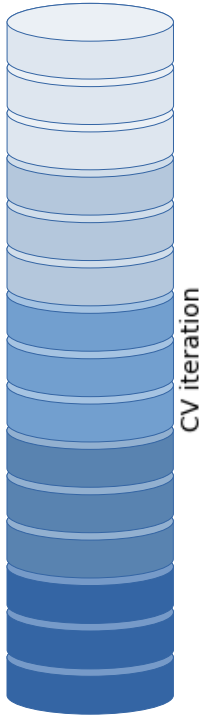
L'échantillonnage pour la séparation en set de entraînement et de test est effectué en fonction de la nouvelle variable.

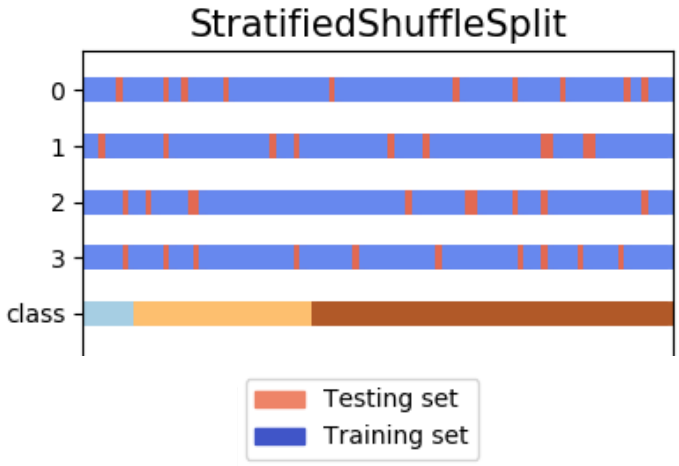
Démarche : **Modélisation Émissions et Consommation**

City of Seattle

Pre-processing

Données nettoyés





STRATIFICATION

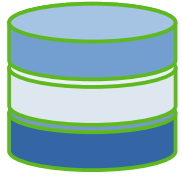
SPLIT

Training set

Encodage Et mise à l'échelle

Données servant à l'entraînement des modèles

Test set



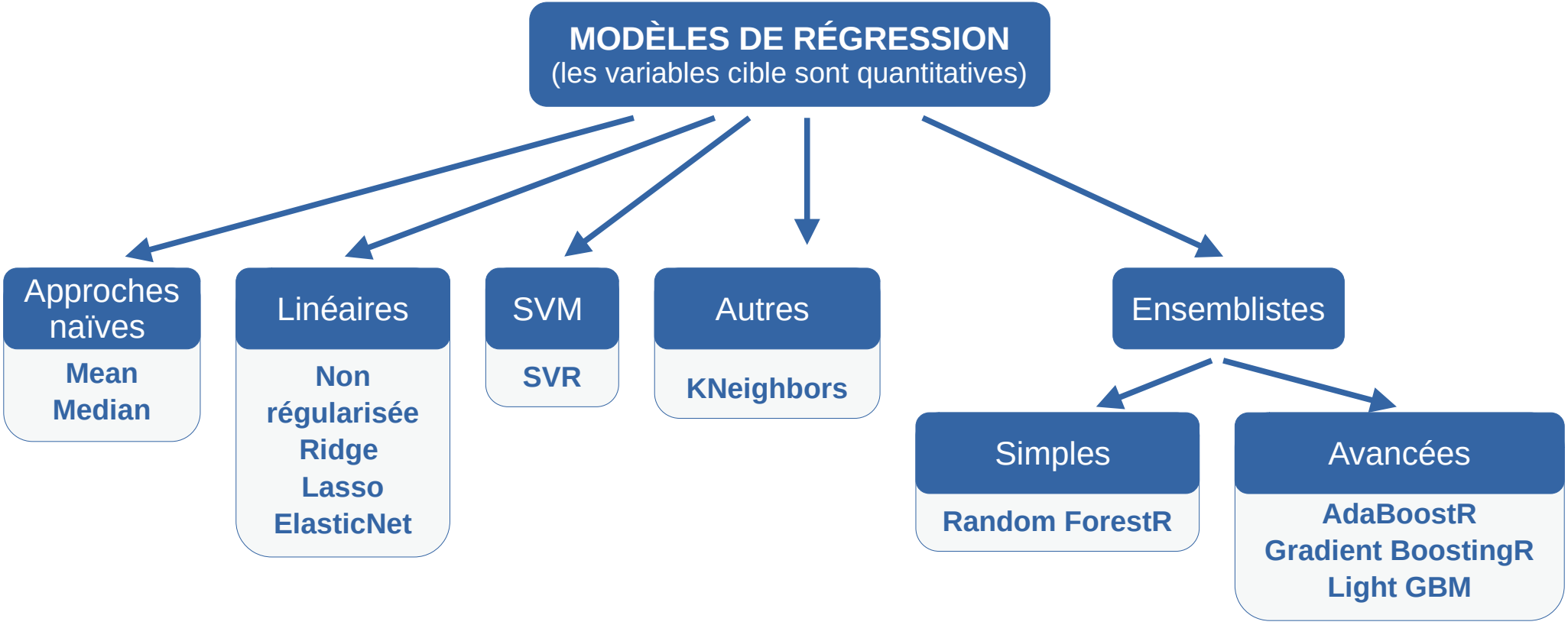
Encodage Et mise à l'échelle

Données servant à valider et classer les modèles en fonction de leurs performances

Démarche : **Modélisation Émissions et Consommation**

City of Seattle

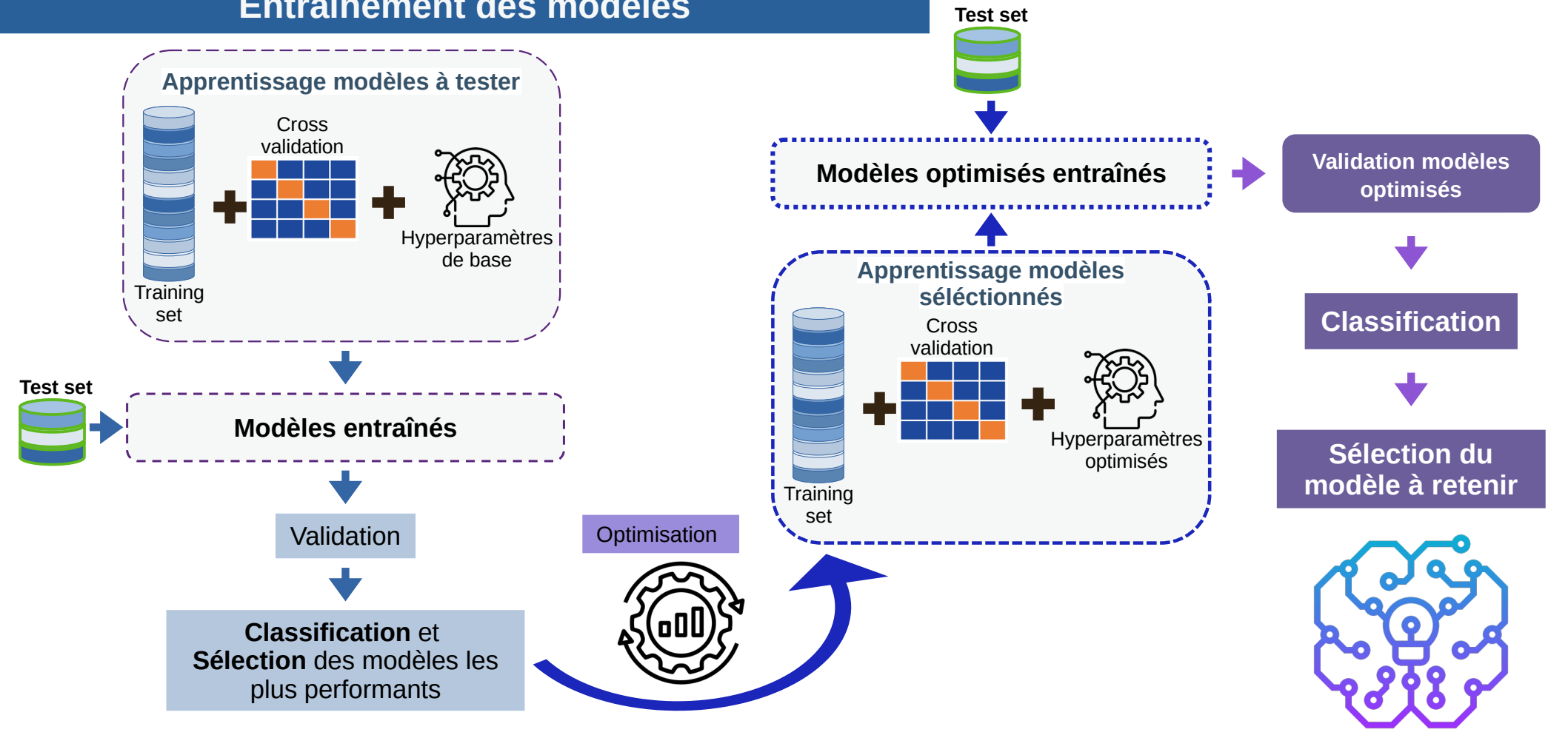
Sélection de modèles à tester



Démarche : **Modélisation Émissions et Consommation**

City of Seattle

Entraînement des modèles

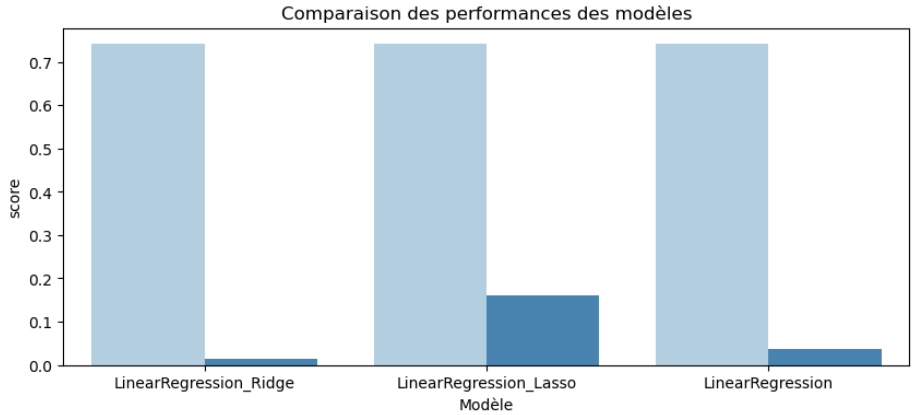


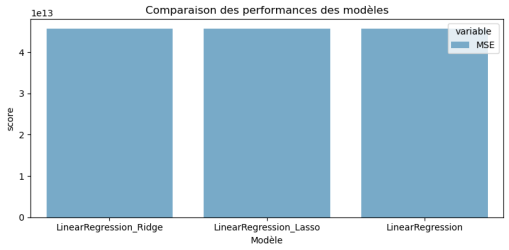
Démarche : **Modélisation Consommation**

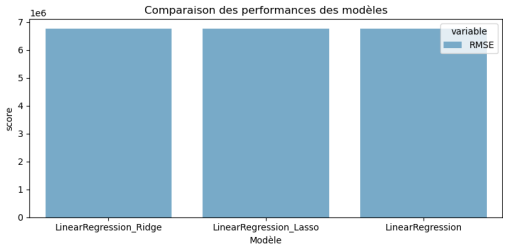
City of Seattle

Entraînement des modèles

Modèle	Scaler	Durée	R2	MAE	MSE	RMSE
LinearRegression_Ridge	StandardScaler	0.013480	0.741440	3224572.000000	45683655213998.796875	6758968.500000
LinearRegression	StandardScaler	0.036870	0.741380	3225734.000000	45694136027506.101562	6759743.800000
LinearRegression_Ridge	RobustScaler	0.023530	0.741380	3225003.000000	45694856266605.703125	6759797.100000
LinearRegression_Lasso	MinMaxScaler	0.161250	0.741160	3225204.000000	45732930565711.601562	6762612.700000
LinearRegression_Lasso	StandardScaler	0.135460	0.741160	3225226.000000	45734117133172.398438	6762700.400000
LinearRegression_Lasso	RobustScaler	0.256970	0.741160	3225225.000000	45734149302616.703125	6762702.800000
LinearRegression	RobustScaler	0.037080	0.741160	3225228.000000	45734207735758.898438	6762707.100000
LinearRegression	MinMaxScaler	0.058660	0.741150	3225336.000000	45734955217963.101562	6762762.400000
LinearRegression_ElasticNet	RobustScaler	0.066250	0.730930	3198545.000000	47541544472467.398438	6895037.700000
Random_Forest	StandardScaler	1.074620	0.727770	3074734.000000	48098747105876.796875	6935326.000000
Random_Forest	MinMaxScaler	1.085450	0.727720	3076721.000000	48107315092538.203125	6935943.700000
Random_Forest	RobustScaler	1.086030	0.727720	3076009.000000	48108881766710.703125	6936056.600000
Gradient_Boosting	StandardScaler	0.291040	0.717840	3083081.000000	49853533622882.398438	7060703.500000
Gradient_Boosting	RobustScaler	0.303390	0.717840	3083081.000000	49853533622882.398438	7060703.500000
Gradient_Boosting	MinMaxScaler	0.353310	0.717840	3083081.000000	49853533622882.398438	7060703.500000
LinearRegression_ElasticNet	StandardScaler	0.044780	0.711060	3421440.000000	51051105323911.398438	7145005.600000
LinearRegression_Ridge	MinMaxScaler	0.006230	0.710110	3370592.000000	51219018145522.898438	7156746.300000
Light_GBM	RobustScaler	0.368570	0.566620	3695073.000000	76572089553341.406250	8750548.000000
Light_GBM	MinMaxScaler	0.204390	0.564220	3696596.000000	76996811015937.500000	8774782.700000
K_NeighborsRegression	StandardScaler	0.100670	0.561300	3749359.000000	77511976937114.093750	8804088.600000
Light_GBM	StandardScaler	0.156830	0.551680	3904437.000000	79211106226102.703125	8900062.100000
K_NeighborsRegression	RobustScaler	0.020170	0.498870	4093689.000000	88543289362673.796875	9409744.400000
K_NeighborsRegression	MinMaxScaler	0.241230	0.347670	4372238.000000	115257554425566.406250	10735807.100000
Ada_Boost	MinMaxScaler	0.137390	0.237150	9615986.000000	134784655721532.593750	11609679.400000







Démarche : **Modélisation Consommation**

City of Seattle

Optimisation modèles plus performants

Régression linéaire Ridge

'alpha' : [.0001, 0.001,0.01, 0.01,1],
'tol' : [0.001,0.0001,0.00001],
'solver' : ['auto', 'svd', 'cholesky', 'lsqr',
'sparse_cg', 'sag', 'saga'],
'max_iter' : [None, 1000,15000]

```
Amélioration du coeff R2 : 0.01%.
Amélioration de la durée : -77.14%.
```

Régression linéaire Lasso

'alpha' : [.0001, 0.001,0.01, 0.01,1],
'tol' : [0.001,0.0001,0.00001],
'max_iter' : [1000,10000, 20000],
'selection' : ['cyclic','random']

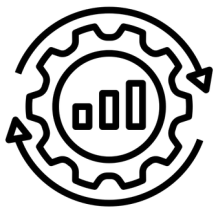
```
Amélioration du coeff R2 : 0.0%.
Amélioration de la durée : 17.23%.
```

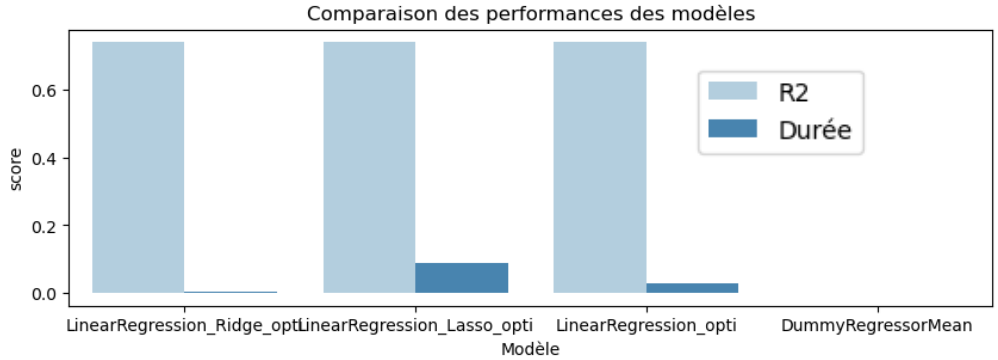
Régression Linéaire non régularisée

'n_jobs' : [None, 1, -1]

```
Amélioration du coeff R2 : 0.0%.
Amélioration de la durée : 1.7%.
```

Modèle	Scaler	Durée	R2	MAE	MSE	RMSE
LinearRegression_Ridge_opti	RobustScaler	0.005040	0.741440	3224572.000000	45683655213998.796875	6758968.500000
LinearRegression_Lasso_opti	MinMaxScaler	0.240870	0.741160	3225204.000000	45732930565711.601562	6762612.700000
LinearRegression_opti	RobustScaler	0.020130	0.741160	3225228.000000	45734207735758.898438	6762707.100000





Démarche : **Modélisation Consommation**

City of Seattle

Prédictions modèle final : Régression Linéaire Ridge, Robust Scaler

Valeurs de Shapley

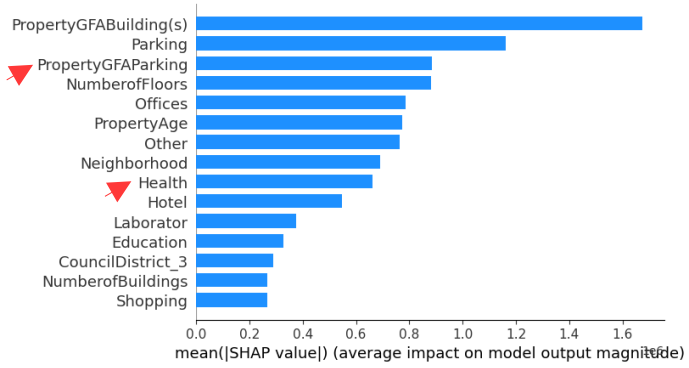
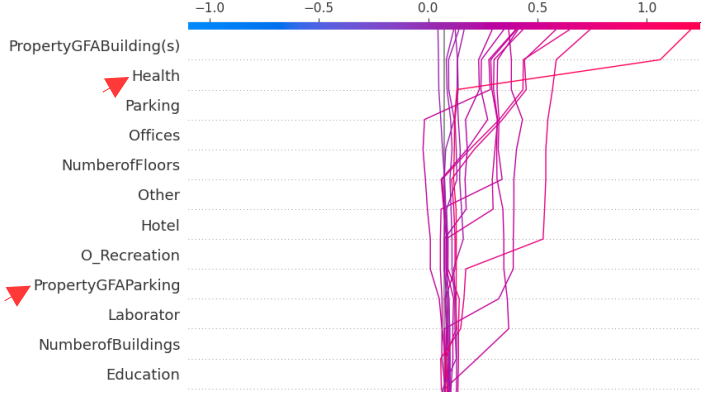
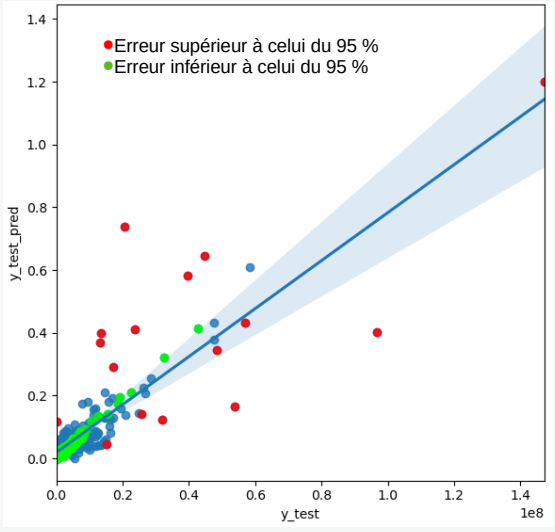
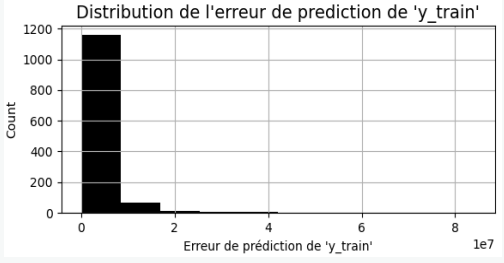


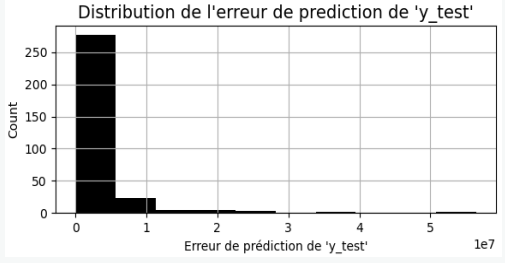
Diagramme de décision des items avec un erreur associé supérieur à celui du 95% des prédictions restantes



Erreurs de prédiction





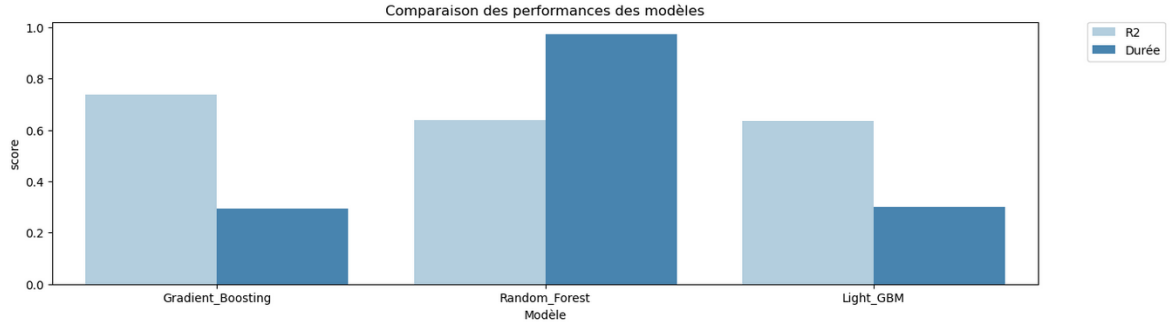


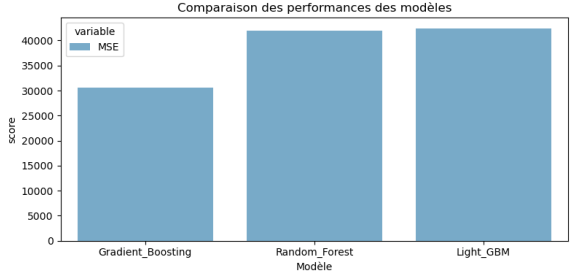
Démarche : **Modélisation Émissions de CO_2**

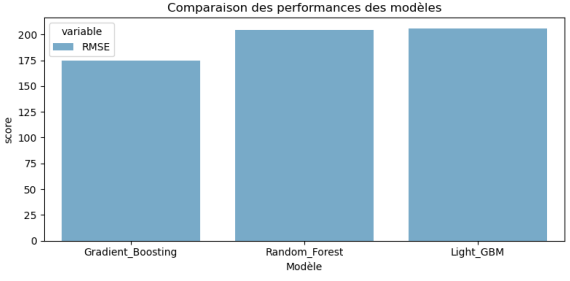
City of Seattle

Entraînement des modèles

Modèle	Scaler	Durée	R2	MAE	MSE	RMSE
Gradient_Boosting	MinMaxScaler	0.281800	0.736110	73.000000	30562.000000	174.800000
Gradient_Boosting	RobustScaler	0.291950	0.736110	73.000000	30562.000000	174.800000
Gradient_Boosting	StandardScaler	0.283700	0.736110	73.000000	30562.600000	174.800000
Random_Forest	StandardScaler	0.972070	0.638250	76.000000	41896.500000	204.700000
Random_Forest	MinMaxScaler	1.472540	0.638240	76.000000	41896.900000	204.700000
Random_Forest	RobustScaler	0.997600	0.638150	76.000000	41907.700000	204.700000
Light_GBM	RobustScaler	0.299860	0.633690	83.000000	42423.900000	206.000000
Light_GBM	StandardScaler	0.109480	0.633000	81.000000	42504.100000	206.200000
Light_GBM	MinMaxScaler	0.093790	0.632010	82.000000	42619.200000	206.400000
LinearRegression_ElasticNet	StandardScaler	0.014470	0.407830	103.000000	68582.500000	261.900000
LinearRegression_Ridge	MinMaxScaler	0.014060	0.372780	111.000000	72641.300000	269.500000
LinearRegression_Lasso	MinMaxScaler	0.042150	0.355920	110.000000	74594.200000	273.100000
LinearRegression_Lasso	StandardScaler	0.034760	0.338960	112.000000	76558.700000	276.700000
LinearRegression	StandardScaler	0.065810	0.327200	114.000000	77920.700000	279.100000
LinearRegression_Ridge	StandardScaler	0.008770	0.326350	114.000000	78018.800000	279.300000
LinearRegression	RobustScaler	0.128820	0.326270	114.000000	78027.800000	279.300000
LinearRegression_Ridge	RobustScaler	0.009220	0.326250	114.000000	78030.000000	279.300000
LinearRegression	MinMaxScaler	0.080880	0.326230	114.000000	78033.300000	279.300000
LinearRegression_Lasso	RobustScaler	0.127890	0.324240	113.000000	78263.000000	279.800000
LinearRegression_ElasticNet	RobustScaler	0.047160	0.297880	107.000000	81316.000000	285.200000
K_NeighborsRegression	RobustScaler	0.226570	0.248350	110.000000	87052.100000	295.000000
K_NeighborsRegression	StandardScaler	0.243140	0.221180	95.000000	90199.200000	300.300000
K_NeighborsRegression	MinMaxScaler	0.579930	0.170220	105.000000	96101.500000	310.000000
LinearRegression_ElasticNet	MinMaxScaler	0.033040	0.049720	154.000000	110057.200000	331.700000







Démarche : **Modélisation Émissions de CO_2**

City of Seattle

Optimisation modèles plus performants

Gradient Boosting Regressor

'n_estimators' : [50, 100, **150**],
'learning_rate' : [0.05,**0.1**, 0.2],
'max_depth' : [**3**, 7, None],
'min_samples_split' : [2, **4**]}

```
Amélioration du coeff R2 : 1.65%.
Amélioration de la durée : 49.89%.
```

Random Forest Regressor

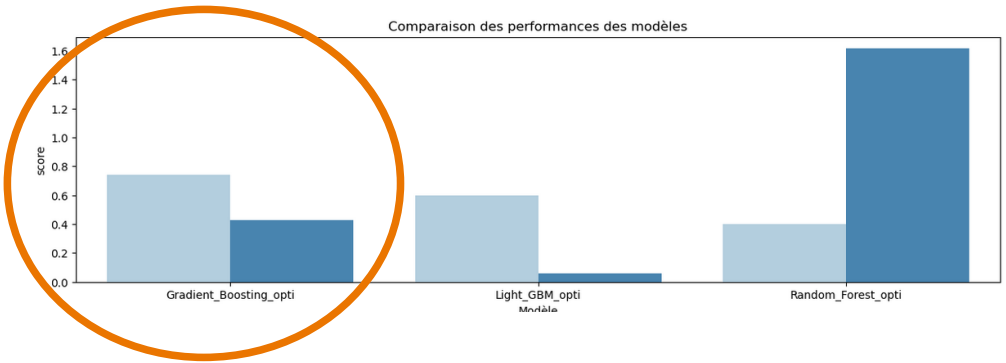
'n_estimators' : [**50**, 100, 150],
'max_depth': [**None**, 5,10],
'min_samples_split': [**2**,3,5,8,15]

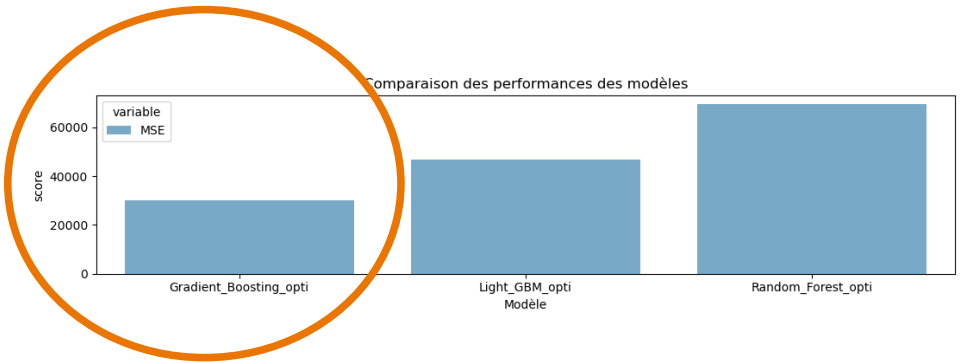
```
Amélioration du coeff R2 : -38.13%.
Amélioration de la durée : 72.77%.
```

Light Gradient Boosting Machine

'n_estimators' : [50, **100**, 150],
'boosting_type': ['gbdt',**'dart'**],
'colsample_bytree': [**0.75**,1.0],
'learning_rate' : [0.05,0.1, **0.2**],
'max_depth' : [-1, 3, **7**],
'num_leaves': [15,**31**,50]

```
Amélioration du coeff R2 : -6.11%.
Amélioration de la durée : -39.04%.
```





Démarche : **Modélisation Émissions de CO_2**

City of Seattle

Prédictions modèle final : Gradien Boosting Regressor, Robust Scaler

Valeurs de Shapley

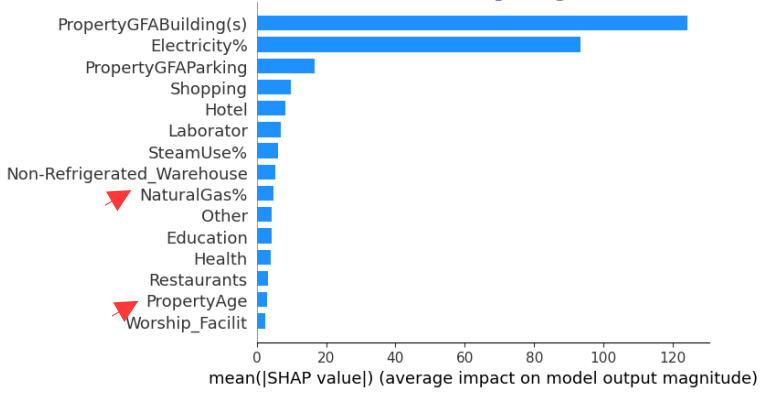
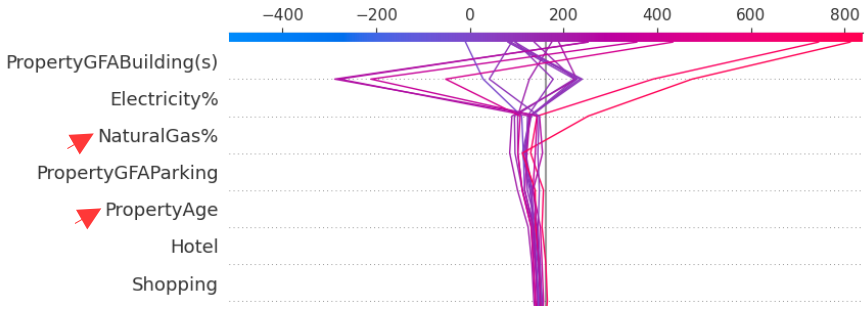
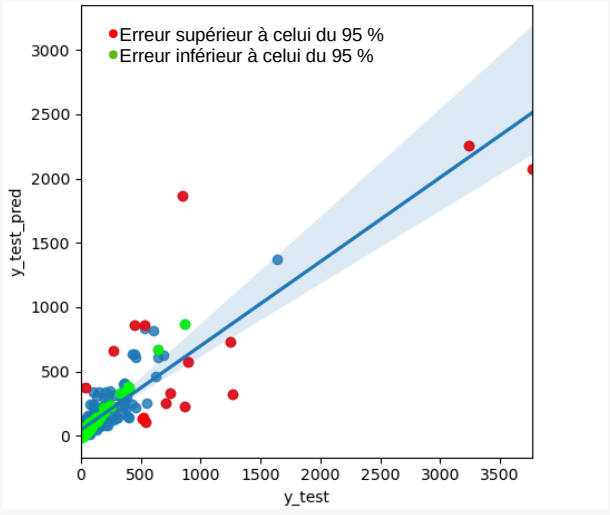
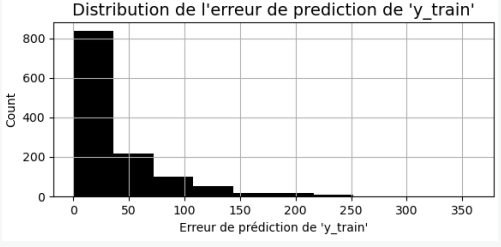


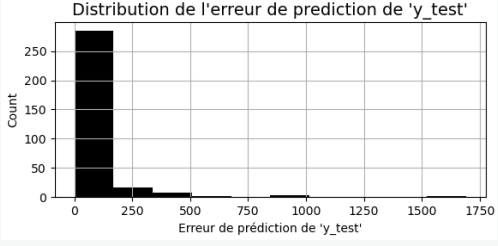
Diagramme de décision des items avec un erreur associé supérieur à celui du 95% des prédictions restantes



Erreurs de prédiction





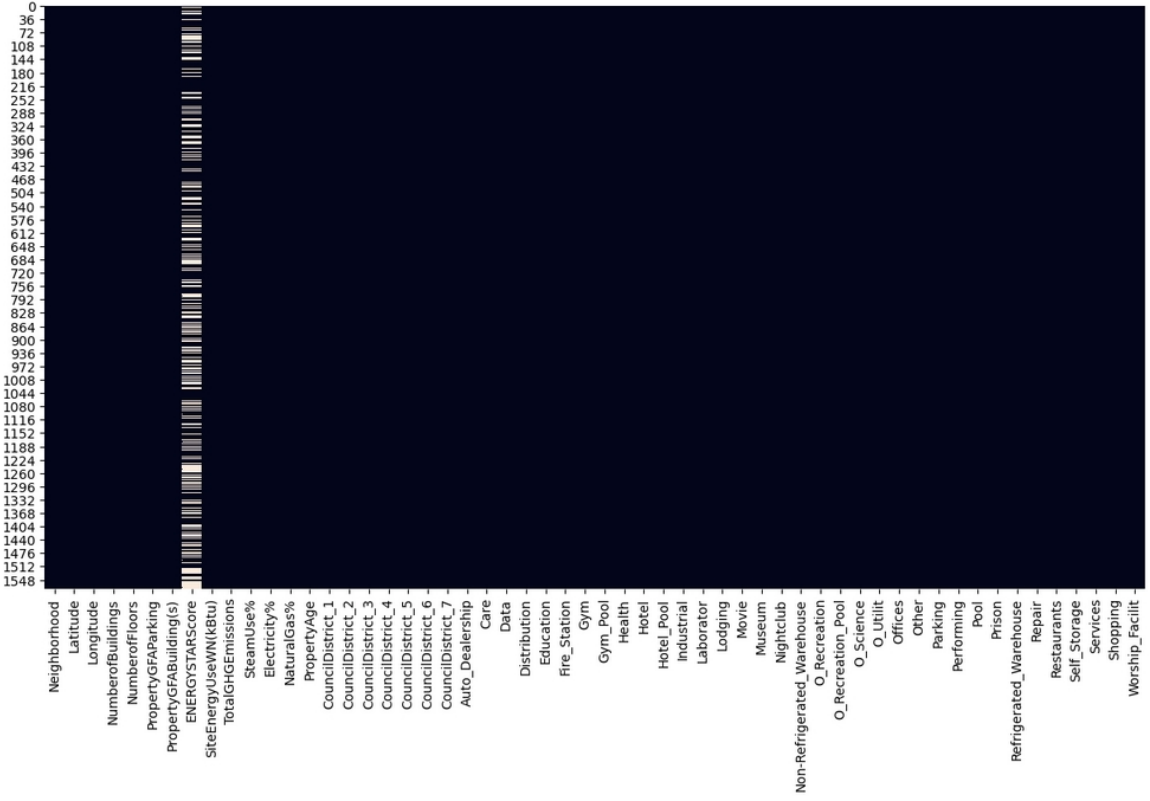


Démarche : **Évaluation intérêt ‘Energy Star score’ pred CO_2**

City of Seattle

(1016 lignes) **Entraînements des modèles :**

Sans ‘Energy Star Score’ **Avec ‘Energy Star Score’**



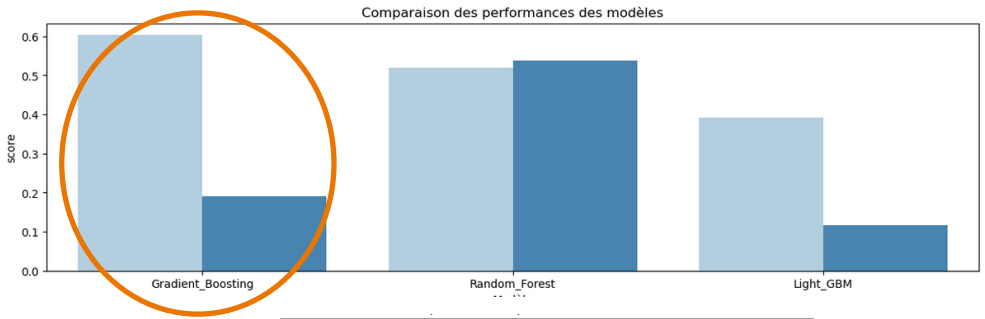
Démarche : **Évaluation intérêt 'Energy Star score' pred CO_2**

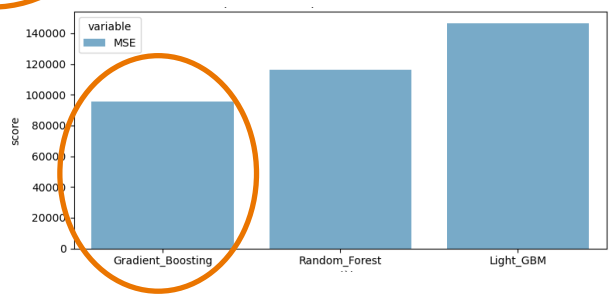
City of Seattle

(1016 lignes) Entraînements des modèles :

Sans 'Energy Star Score'

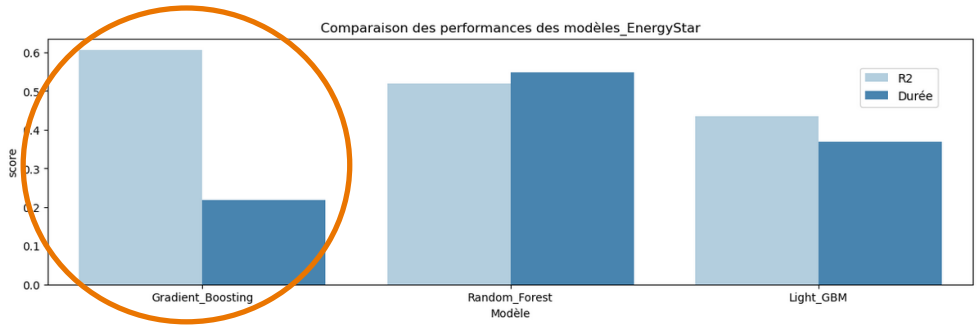
Modèle	Scaler	Durée	R2	MAE	MSE	RMSE
Gradient_Boosting	RobustScaler	0.190160	0.602880	88.000000	95775.300000	309.500000
Random_Forest	RobustScaler	0.536780	0.518540	89.000000	116114.300000	340.800000
Light_GBM	RobustScaler	0.115820	0.392350	102.000000	146547.300000	382.800000

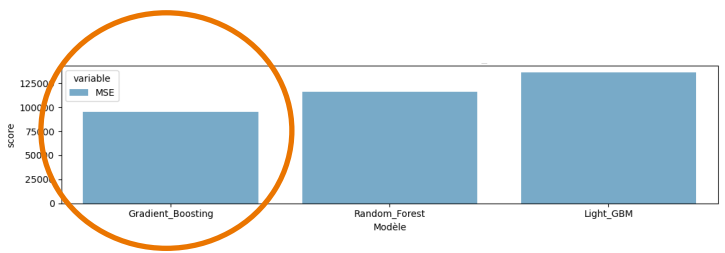




Avec 'Energy Star Score'

Modèle	Scaler	Durée	R2	MAE	MSE	RMSE
Gradient_Boosting	RobustScaler	0.218760	0.605750	80.000000	95083.100000	308.400000
Random_Forest	RobustScaler	0.547650	0.518640	83.000000	116091.600000	340.700000
Light_GBM	RobustScaler	0.368030	0.434920	91.000000	136280.600000	369.200000





Démarche : **Évaluation intérêt 'Energy Star score' pred CO_2**

City of Seattle

Optimisation Gradient Boosting Regressor :

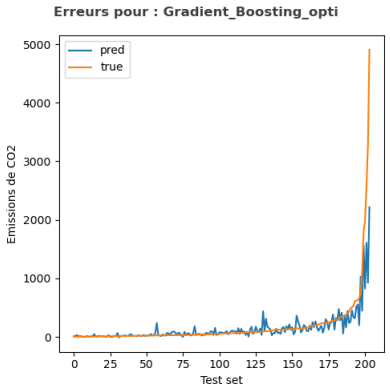
Sans 'Energy Star Score'

Modèle	Scaler	Durée	R2	MAE	MSE	RMSE
Gradient_Boosting_opti	RobustScaler	0.289030	0.656720	84.000000	82790.500000	287.700000

Gradient Boosting Regressor

'n_estimators' : [50, 100, **150**],
'learning_rate' : [0.05,0.1, **0.2**],
'max_depth' : [**3**, 7, None],
'min_samples_split' : [**2**, 4]

```
Amélioration du coeff R2 : 8.93%.
Amélioration de la durée : 51.99%.
```



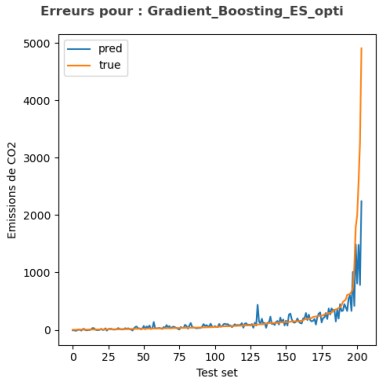
Avec 'Energy Star Score'

Modèle	Scaler	Durée	R2	MAE	MSE	RMSE
Gradient_Boosting_opti	RobustScaler	0.304270	0.648200	76.000000	84844.300000	291.300000

Gradient Boosting Regressor

'n_estimators' : [50, 100, **150**],
'learning_rate' : [0.05,0.1, **0.2**],
'max_depth' : [**3**, 7, None],
'min_samples_split' : [**2**, 4]

```
Amélioration du coeff R2 : 7.01%.
Amélioration de la durée : 51.65%.
```



Démarche : **Évaluation intérêt 'Energy Star score' pred CO_2**

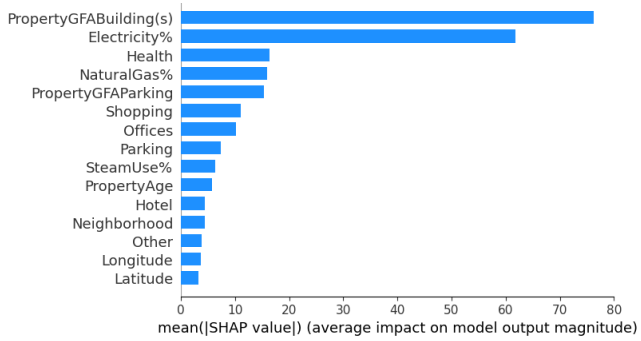
City of Seattle

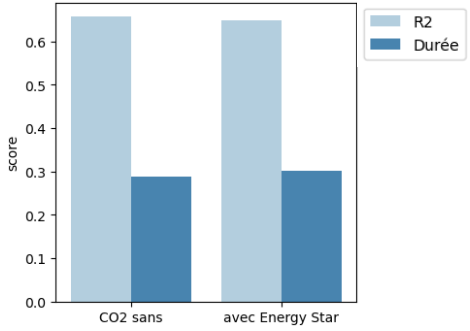
Comparaison Prédictions :

Sans 'Energy Star Score'

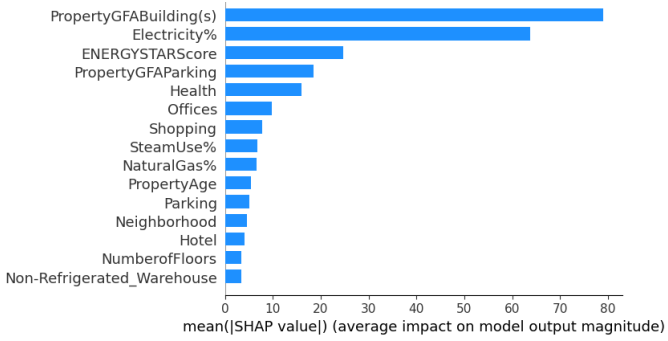
Avec 'Energy Star Score'

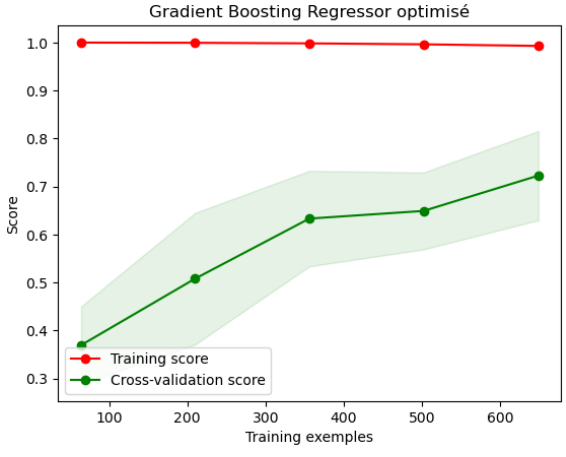
Valeurs de Shapley

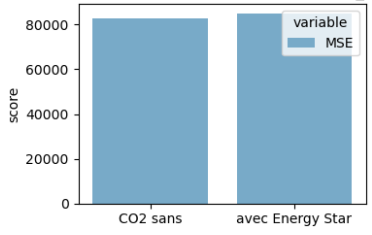


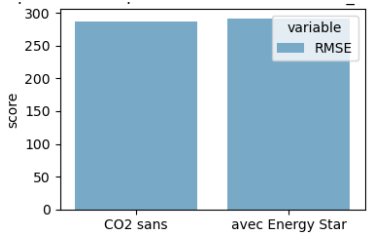


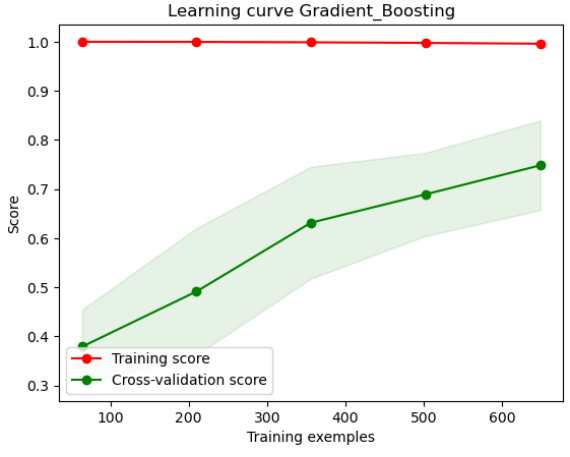
Valeurs de Shapley









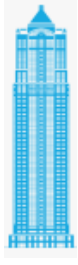


Problématique

Données

Démarche :

- EDA, Nettoyage et Feature engineering
- Modélisation Consommation E. et Émissions CO_2
- Évaluation intérêt ‘Energy Star score’



Conclusion

Conclusion

City of Seattle

- **Modèle le plus performant pour la prédiction de la consomation totale d'énergie est la *Régression Linéaire Ridge* sur des données mis à l'échelle avec Robust Scaler**
- **L'algorithme le plus performant pour la prédiction des émmisions de CO_2 est 'Gradient Boosting Regressor sur des données mis à l'échelle avec Robust Scaler**
- **Prendre en compte le score 'Energy Star' n'améliore pas les performances de l'algorithme de prédiction des émmisions de CO_2**

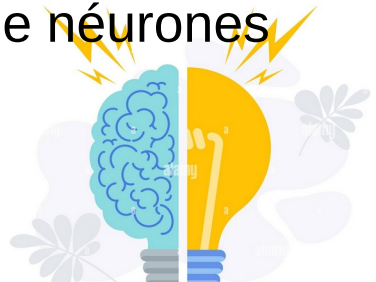
Perspectives

- Augmenter la taille du data-set

- Essayer différentes stratégies de « features engineering »

- Améliorer l’optimisation des hyperparamètres des modèles de régression
  - Plupart des cas amélioration uniquement de la durée
  - Random Forest dégradation très importante de la performance

- Tester les réseaux de néurones

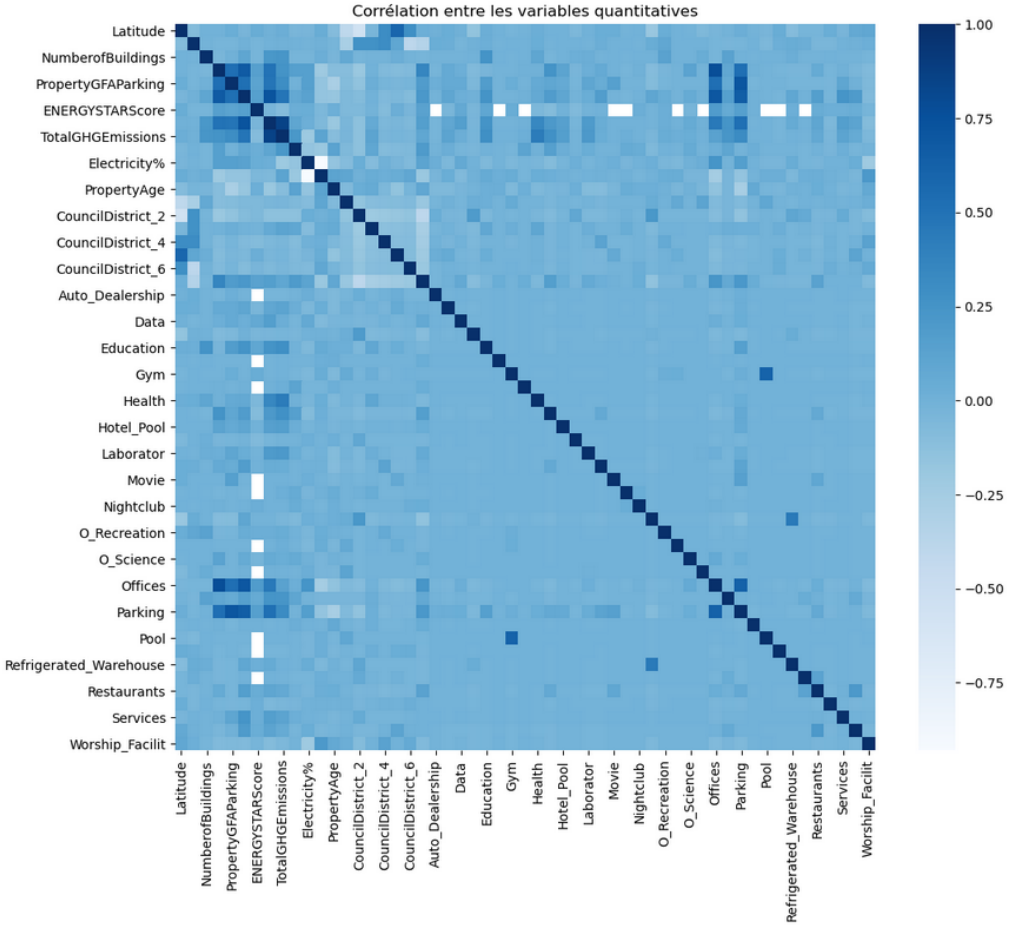


Boîte à outils

Démarche : **Analyse Exploratoire de Données, Nettoyage et Feature Engineering**

City of Seattle

Feature Engineering

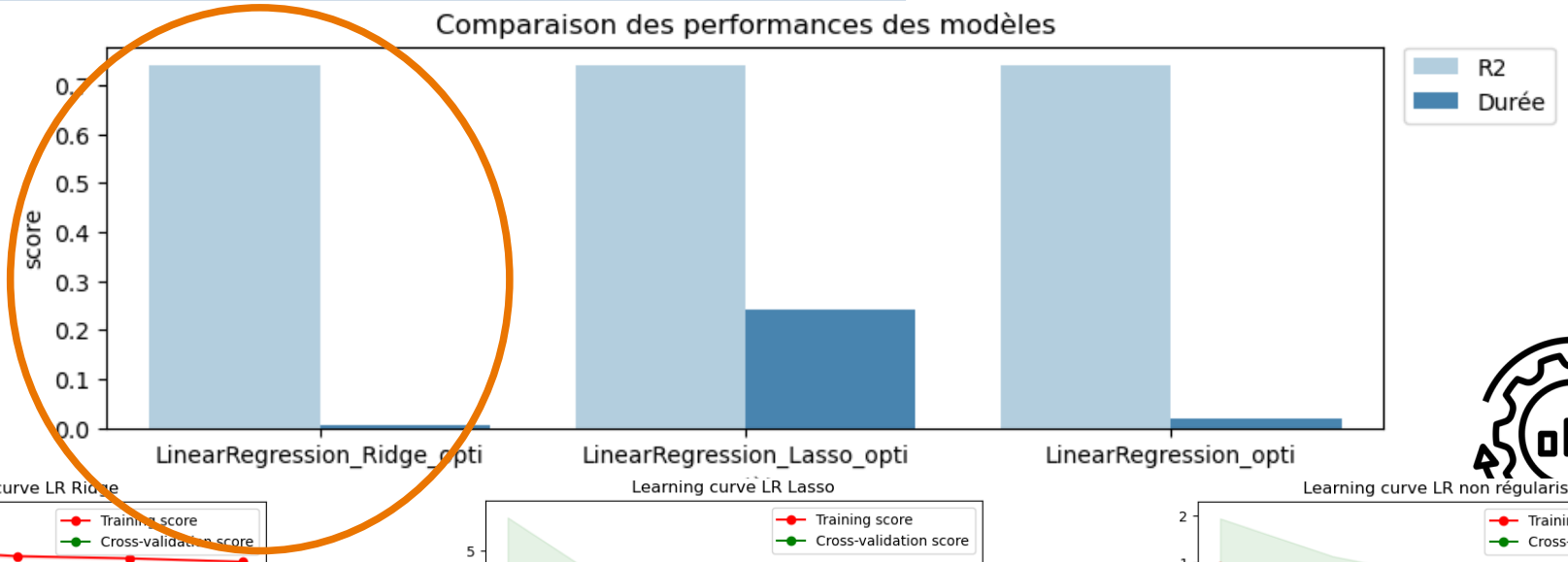


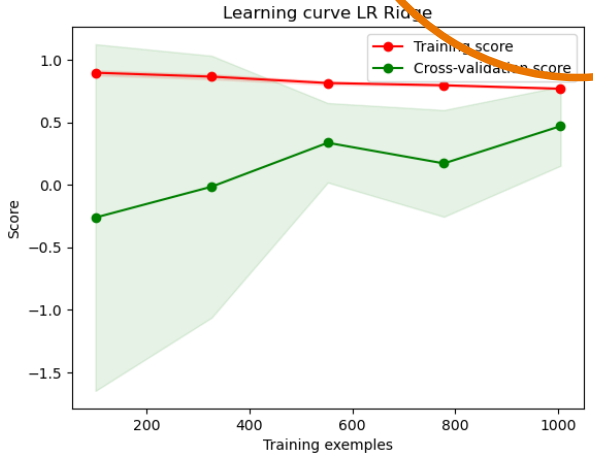
Résultat : diminution de la corrélation entre les variables

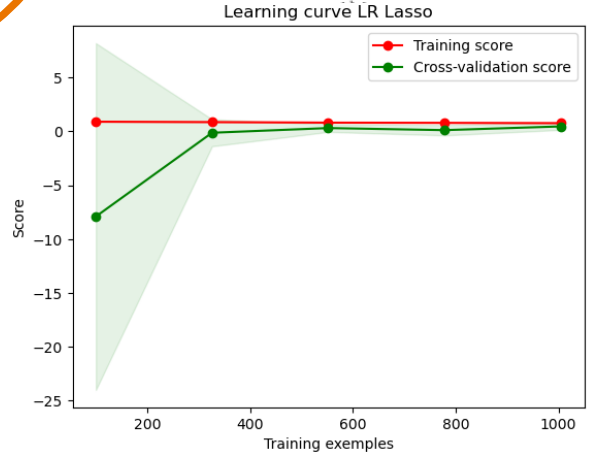
Démarche : **Modélisation Consommation**

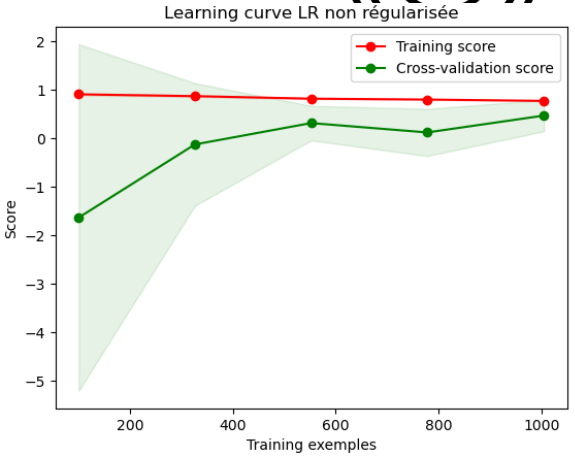
City of Seattle

Optimisation modèles plus performants









J'ai considéré la possibilité de calculer la précision (100- MAPE, mean absolute percentage error) mais comme il y a y_test = 0 ce n'était pas pertinente