



ANTICIPEZ LES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS DE LA VILLE DE SEATTLE

Victoire MOHEBI

Février 2022



OPENCLASSROOMS

SOMMAIRE

- Rappel du problématique
- Mission
- Source
- Présentation de jeu de données
- Nettoyages effectués
- Analyse exploratoire
- Modélisation
- Synthèse





PROBLÉMATIQUE

MISSION

SOURCE



City of Seattle

PROBLÉMATIQUE

- ❖ Prédiction de la consommation énergétique et de l'émission de CO2 des bâtiments non résidentiels de la ville de Seattle
- ❖ Evaluer l'intérêt de « EnergyStarScore » en essayant de modéliser sans et avec

MISSION

- ❖ Réaliser une analyse exploratoire
- ❖ Tester différents modèles de prédiction
- ❖ Evaluer l'intérêt de l'*EnergyStarScore* pour la prédiction de l'émission de CO2

SOURCE

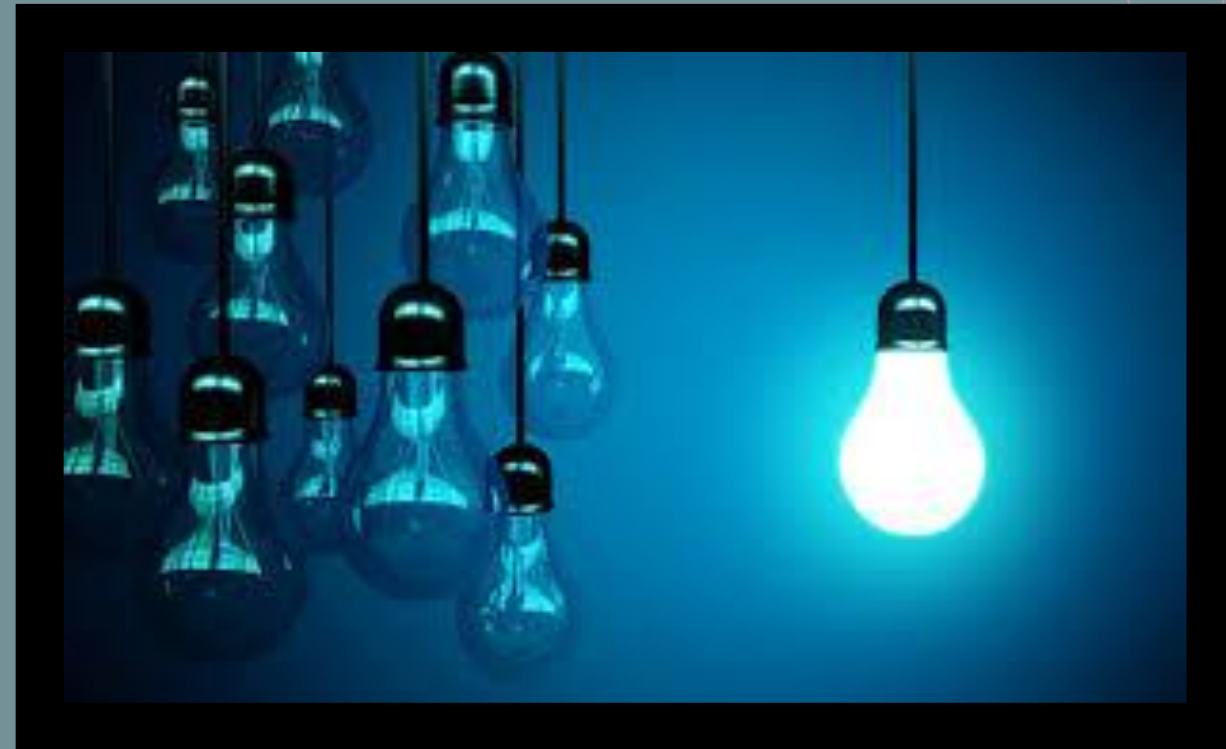
Des données détaillées sur des relevés minutieux de 2015 et 2016 et les caractéristiques des bâtiments disponible via le portail de la [ville de Seattle](#)



PRÉSENTATION DU JEU DE DONNÉES

Dimension et les variables

- Dataframe de 2015 a 3340 lignes et 47 colonnes
- Dataframe de 2016 a 3376 lignes et 46 colonnes
- Les deux dataframes contiennent les informations sur :
 - Les relèves minutieux des établissements non résidentiels
 - Superficie (en square feet)
 - Nombre des étages
 - Année de construction
 - Type et usage de bâtiments
 - Coordonnées GPS, latitude et longitude
 - Adresses





NETTOYAGE DES DONNÉES

**Extraire les donnée de
localisation**

Typage de données

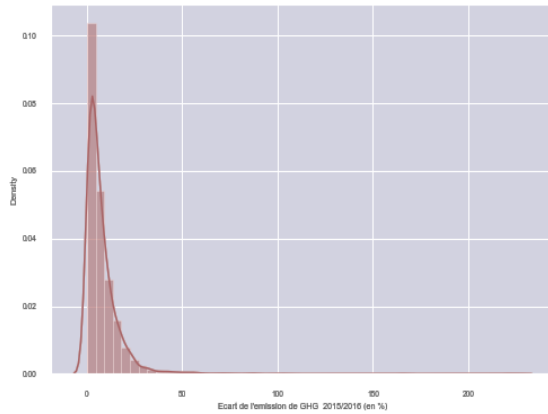
Valeurs aberrantes

**Anomalie des variables cibles
pour les bâtiments identiques**

Valeurs manquantes

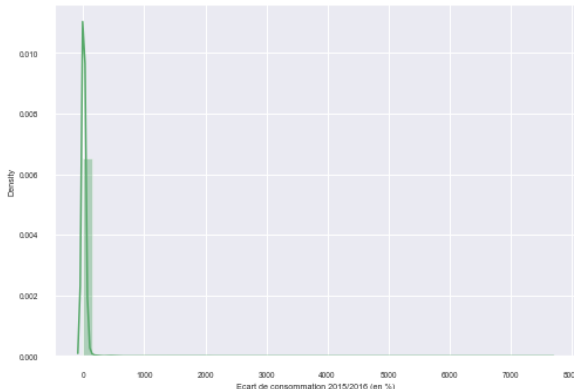
Anomalie des variables cibles pour les bâtiments identiques

Distribution des écarts de émission de GHG 2015-2016 pour des bâtiments identiques (en %)



- Identifier et supprimer les bâtiments dont l'écart entre les relevé de la consommation énergétique de 2015 et 2016 est supérieur à 20%

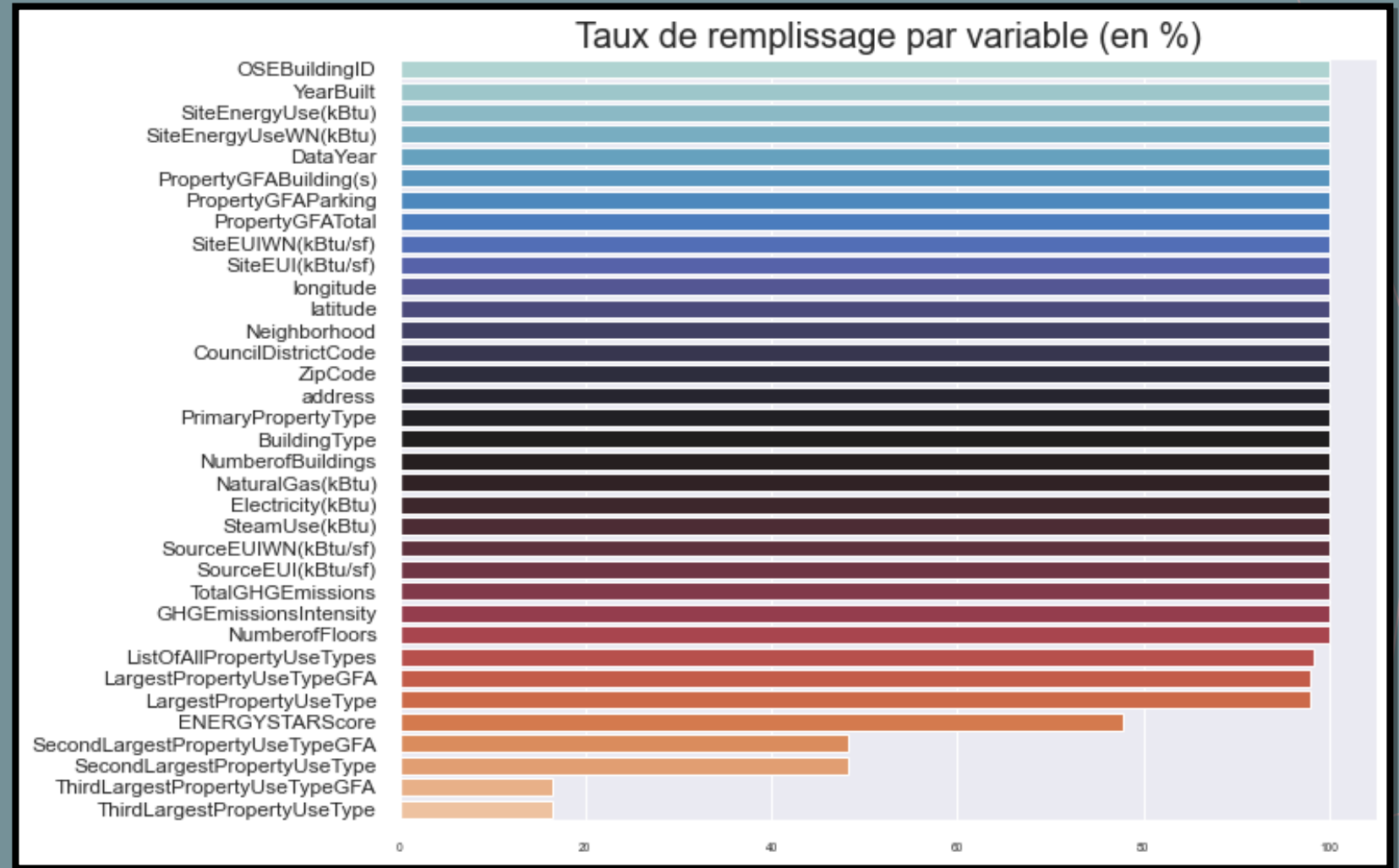
Distribution des écarts de consommation énergétique 2015-2016 pour des bâtiments identiques (en %)



- Identifier et supprimer les bâtiments dont l'écart entre les relevé de l'émission de CO2 de 2015 et 2016 est supérieur à 20%

Taux de remplissage de jeu de données

- Les variables sont majoritairement bien renseigné
- Les variables peu renseignées sont celle de « Third & Second Property Type » et « EnergyStarScore »



Type de « *missingness* » des valeurs manquantes

- Une relation systématique entre les valeurs manquantes de second & third usage.
- Pour les autres variable les valeurs manquantes sont aléatoires et leur nombre est faible
- Supprimer les variables avec plus de 80% de valeurs manquantes.

Matrix de Missingo

OSEBuildingID
DataYear
BuildingType
PrimaryPropertyType
address
ZipCode
CouncilDistrictCode
Neighborhood
latitude
longitude
YearBuilt
NumberofBuildings
NumberofFloors
PropertyGFATotal
PropertyGFAParking
PropertyGFABuilding(s)
ListOfAllPropertyUseTypes
LargestPropertyUseType
LargestPropertyUseTypeGFA
SecondLargestPropertyUseType
SecondLargestPropertyUseTypeGFA
ThirdLargestPropertyUseType
ThirdLargestPropertyUseTypeGFA
ENERGYSTARScore
SiteEUI(kBtu/sf)
SiteEUIWN(kBtu/sf)
SourceEUI(kBtu/sf)
SourceEUIWN(kBtu/sf)
SiteEnergyUse(kBtu)
SiteEnergyUseWN(kBtu)
SteamUse(kBtu)
Electricity(kBtu)
NaturalGas(kBtu)
TotalGHGEmissions
GHGEmissionsIntensity



REGROUPEMENT DE TYPE D'USAGE DES ÉTABLISSEMENTS

8 catégories unique de types des bâtiment

91 sous catégories d'usage des bâtiments

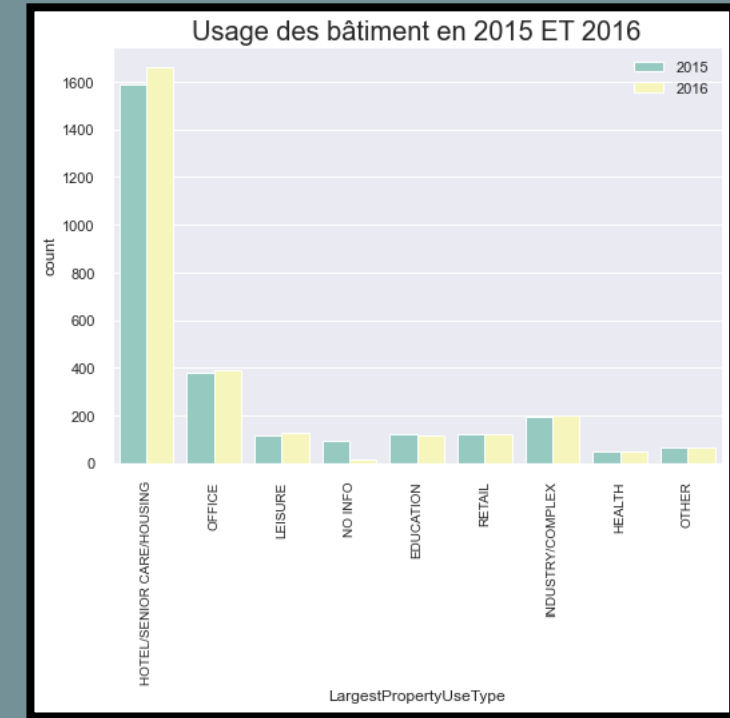
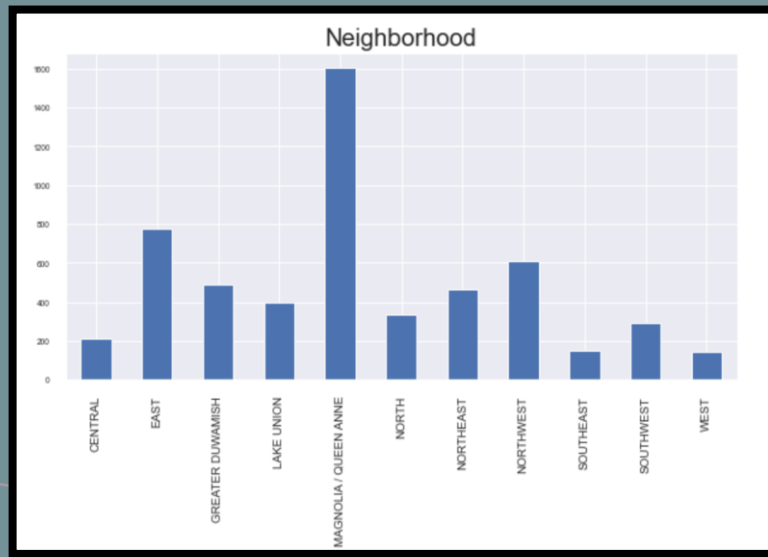
Type d'usages des bâtiments (PropertyType) ne sont pas sous-catégories uniques des types de bâtiments(BuildingType)

Regrouper la modalité d'usage du bâtiments en 8 groups

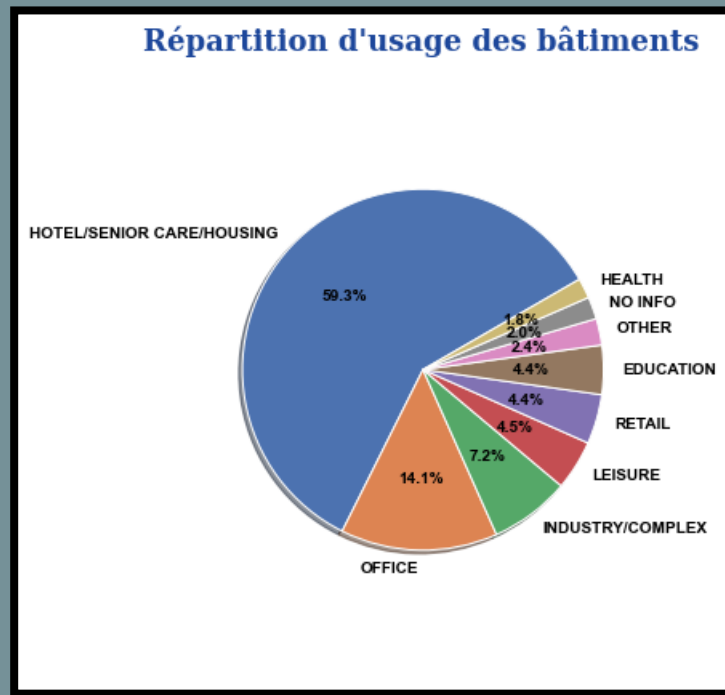


ANALYSE EXPLOIRATOIRE

- Les type d'usage de bâtiments sous 2 ans sont presque également représentés.
- Certains quartiers sont très peu représentés.

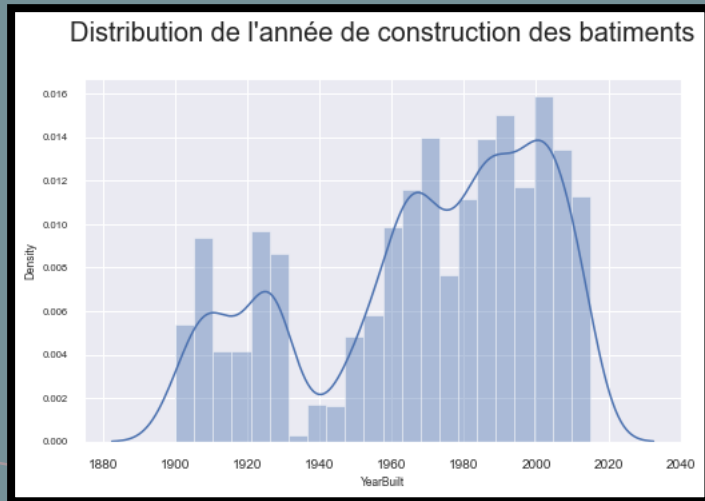
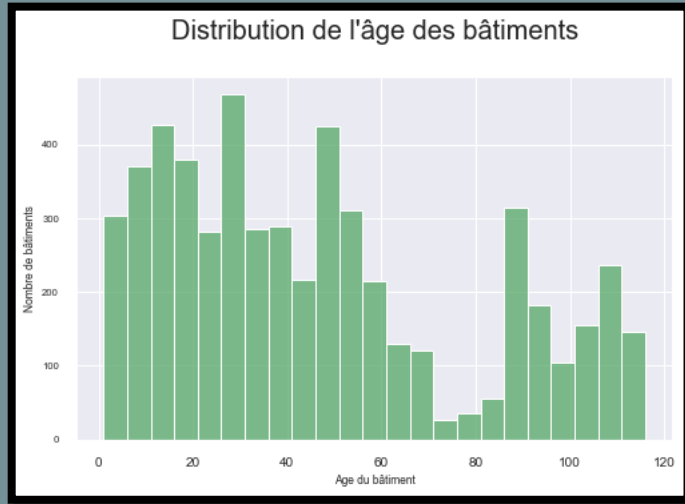


Le type d'usage des établissements



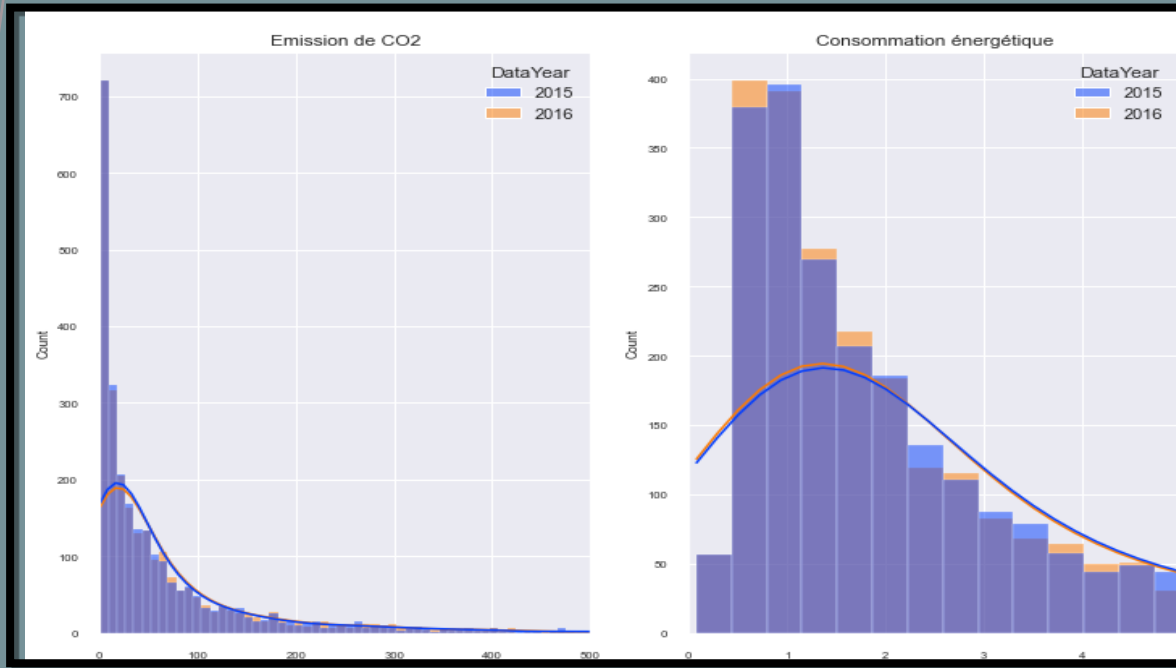
- La modalité d'usage des bâtiments ne sont pas représenté également.
- Les « Hotels » est le plus nombreux dans le jeu de données

L'âges des établissement



- On a les données sur les bâtiments depuis 1900
- Très peu de construction bâtie aux années 40 ce qui est expliqué par la crise de la 2^{de} guerre mondiale
- Peu de construction récentes

Distribution des variables cibles sous deux ans



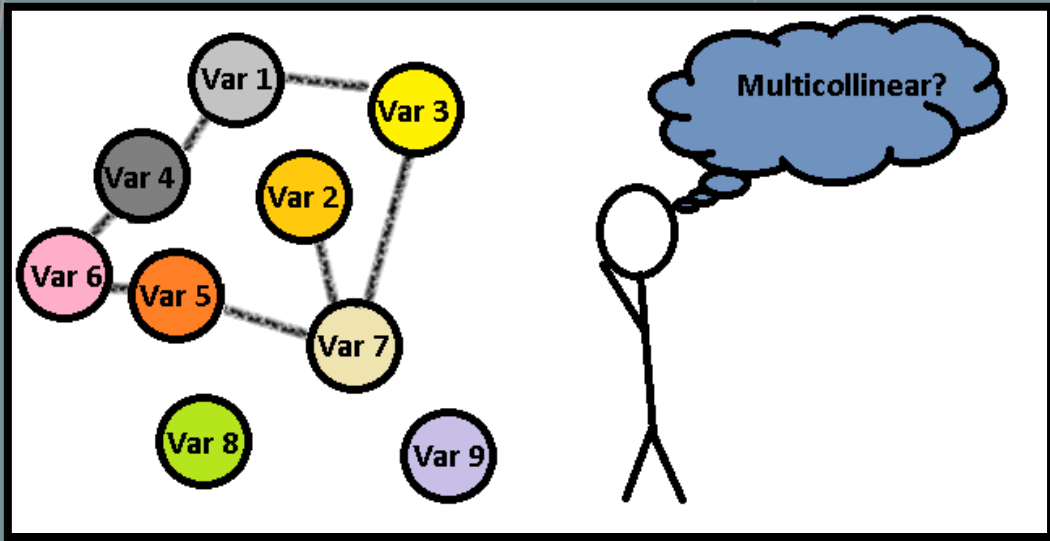
- La distribution de 2 variables cibles est asymétrique vers le gauche
- Les deux variables ont quasiment la même distribution sous deux ans

Analyse bivariée

- L'analyse bivariée montre qu'il y a des variables indépendantes fortement corrélées entre elles:
 - Les variables de la consommation d'énergie normalisées (avec suffix WN) et non normalisées.
 - Les variables avec suffix GFA (superficie)
- D'où la nécessité d'analyse de multicollinéarité avant la modélisations



Multicolinéarité ?!



- La multicolinéarité se produit lorsque deux ou plusieurs variables indépendantes sont fortement corrélées entre elles dans un modèle de régression.
- Cela signifie qu'une variable indépendante peut être prédite à partir d'une autre variable indépendante dans un modèle de régression.

Vérification de multicolinéarité avec « variance inflation factor »

- Créer de nouvelles variables en calculant Building surface/total
- Garder les deux variables cibles « SiteEnergyUse » et « Total GHGEmission »

	feature	VIF
0	<u>LargestPropertyUseTypeGFA</u>	12.254500
1	<u>PropertyGFABuilding(s)</u>	28.008235
2	<u>PropertyGFATotal</u>	28.507590

	feature	VIF
0	LargestPropertyUseTypeGFA	13.374800
1	SiteEnergyUseWN(kBtu)	1384.329087
2	SiteEnergyUse(kBtu)	1297.137668
3	SourceEUI(kBtu/sf)	3518.256379
4	PropertyGFABuilding(s)	29.964655
5	GHGEmissionsIntensity	18.574783
7	PropertyGFATotal	38.992097
8	SiteEUIWN(kBtu/sf)	2277.212088
9	TotalGHGEmissions	9.759358
10	SourceEUIWN(kBtu/sf)	3286.008892
11	SiteEUI(kBtu/sf)	2610.098251



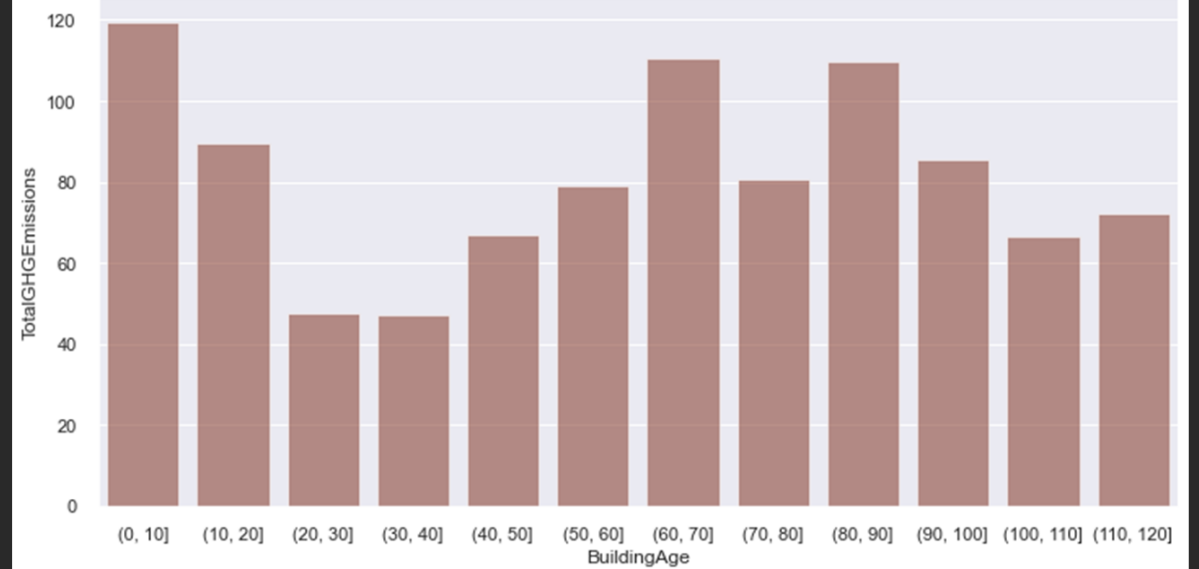
ANALYSE DES VARIABLES À PRÉDIRE

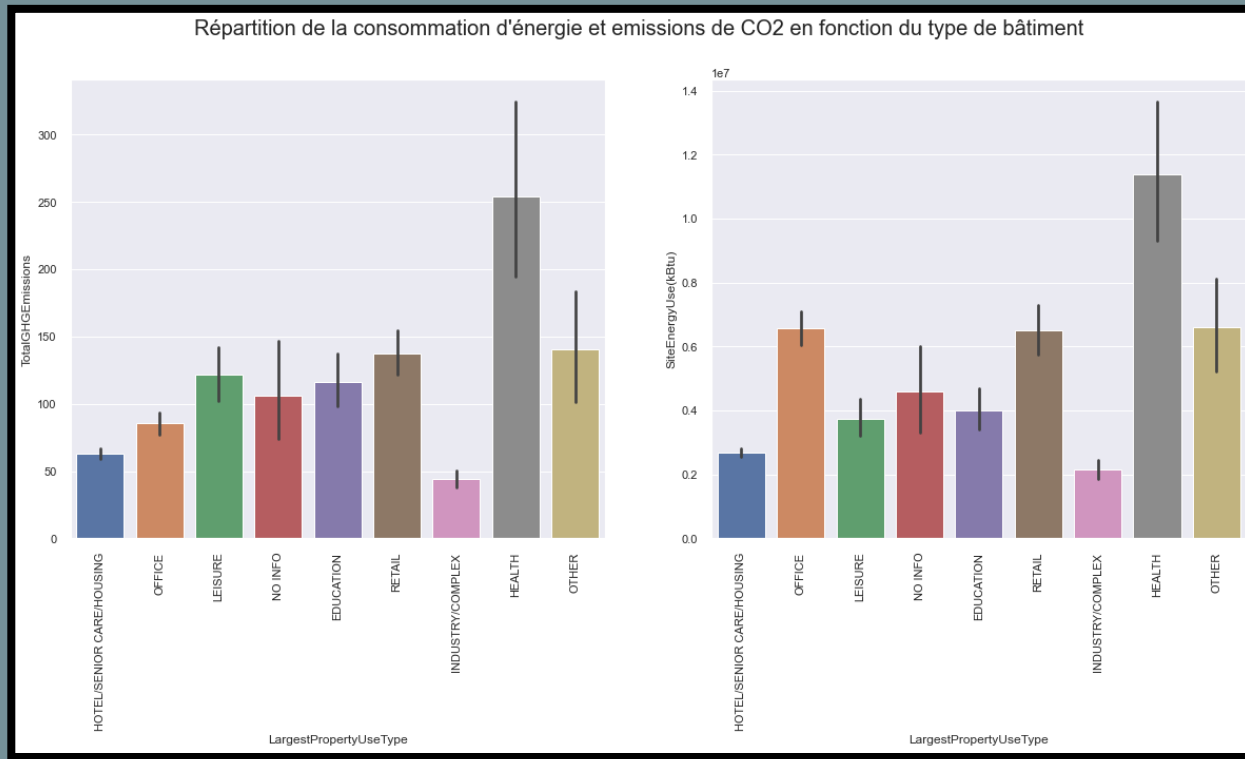


Consommations d'énergie vs. emission de GHG par l'usage des bâtiment



Influence de l'âge des bâtiments sur les émissions de CO2

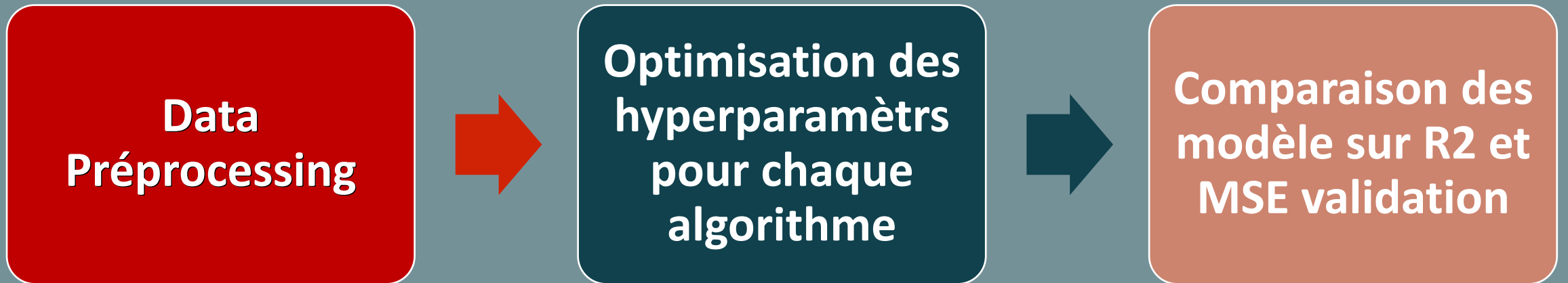




- Les établissements avec type d'usage "Health" et les "Office" consomment plus que les autres de l'énergie.
- - Plus de gaz à effet de serre est émis par les établissements du type d'usage "Health" et "Retail"

MODÉLISATION

Démarche de modélisation



Processus : data preprocessing

01

One-Hot Encoding des données quantitatives

02

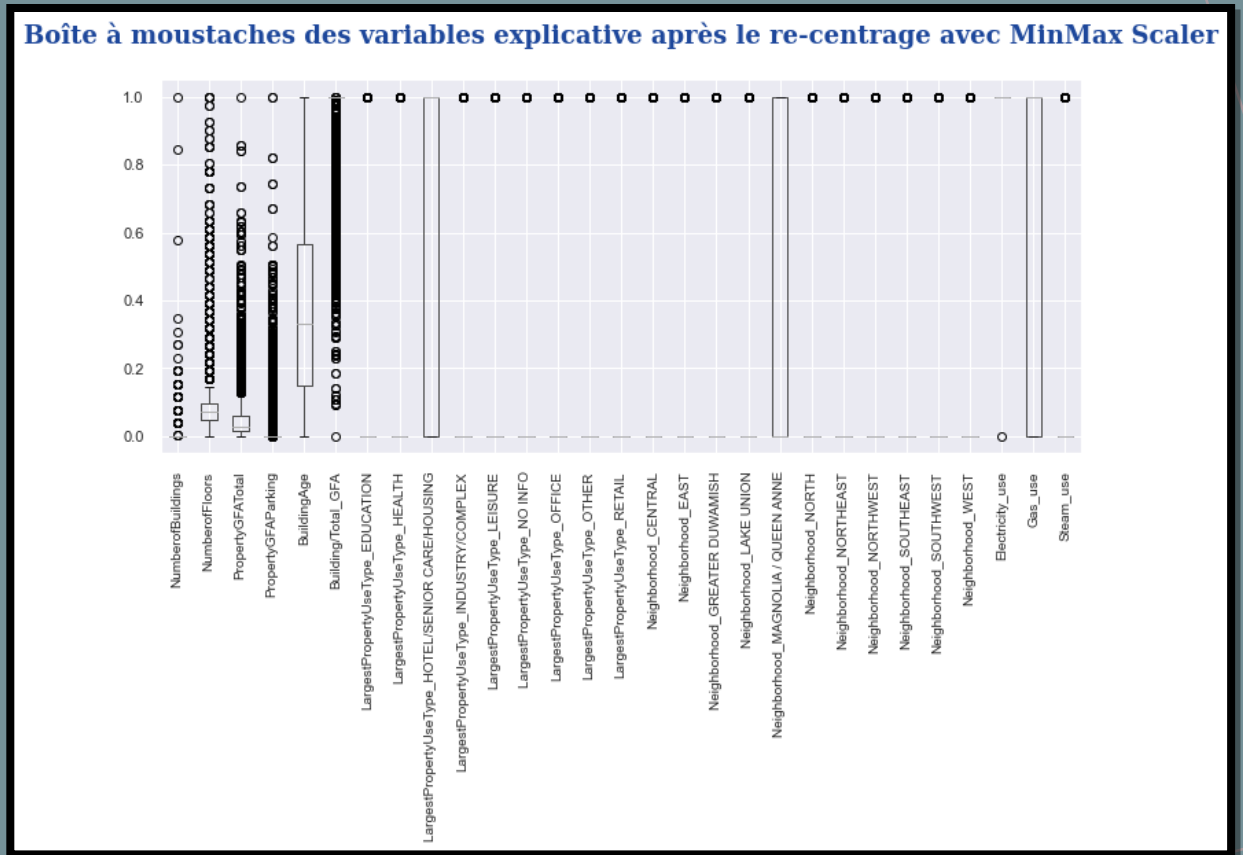
Data Transformation (Features scaling)

03

Séparation des donnée en Train/Test

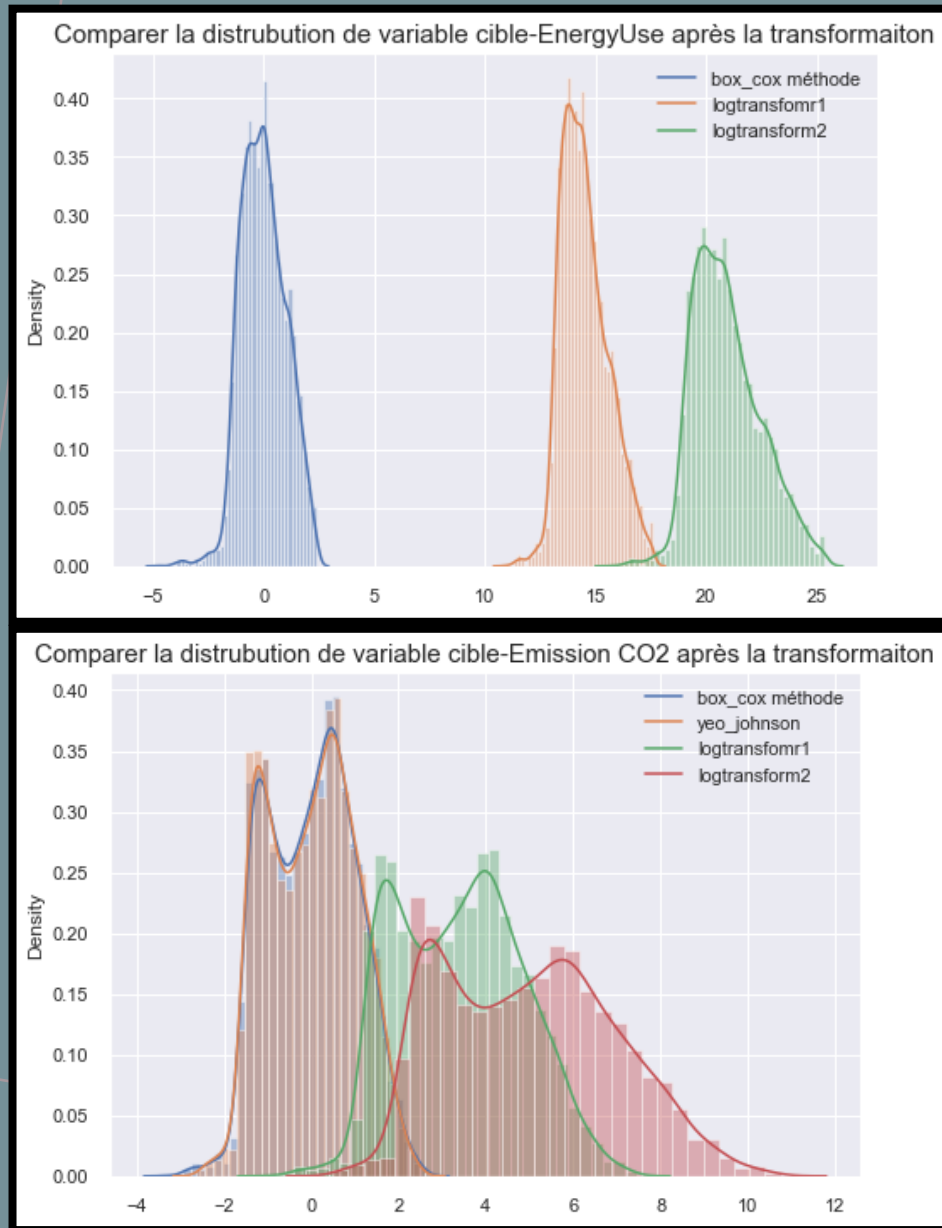
Data transformation : recentrage des variables explicatives

- Le dataframe final pour la modélisation a 5477 lignes et 13 colonnes
- Transformation des variables qualitatives de « Neighbourhood », « Electricity », « Gaz » et « Steam » en variables booléenne 0/1



Data transformation : recentrage des variables cibles

- La transformation logarithmique est la plus pertinente pour les deux variables cibles.
- Ce transformateur a normalisé au mieux la distribution de la variable cible.



MODÈLES DE RÉGRESSION À TESTER

Regréssion linéaire multivariée

Plusieurs
variables
explicative pour
la prediction
d'une variable
continue

ElasticNet

Régression
régularisée

Random Forest Regressor

Régression non-
linéaire,
méthode
ensembliste

Xgboost

Algorithme de
gradient
boosting basée
sur des arbres
de décision

LightGBM

Basée sur le
Gradient
Boosting
Machine, Moins
d'utilisation de
la mémoire



Ajustement des hyperparamètres du modèle

Définition de
la grille de
recherche

Entraînement
du modèle

Evaluer sa
performance

Hyperparametres tuning

Random search

Random Search configure une grille de valeurs d'hyperparamètres et sélectionne des combinaisons aléatoires pour former le modèle et le score

RandomSearchCV

Adjustment des hyperparamètre

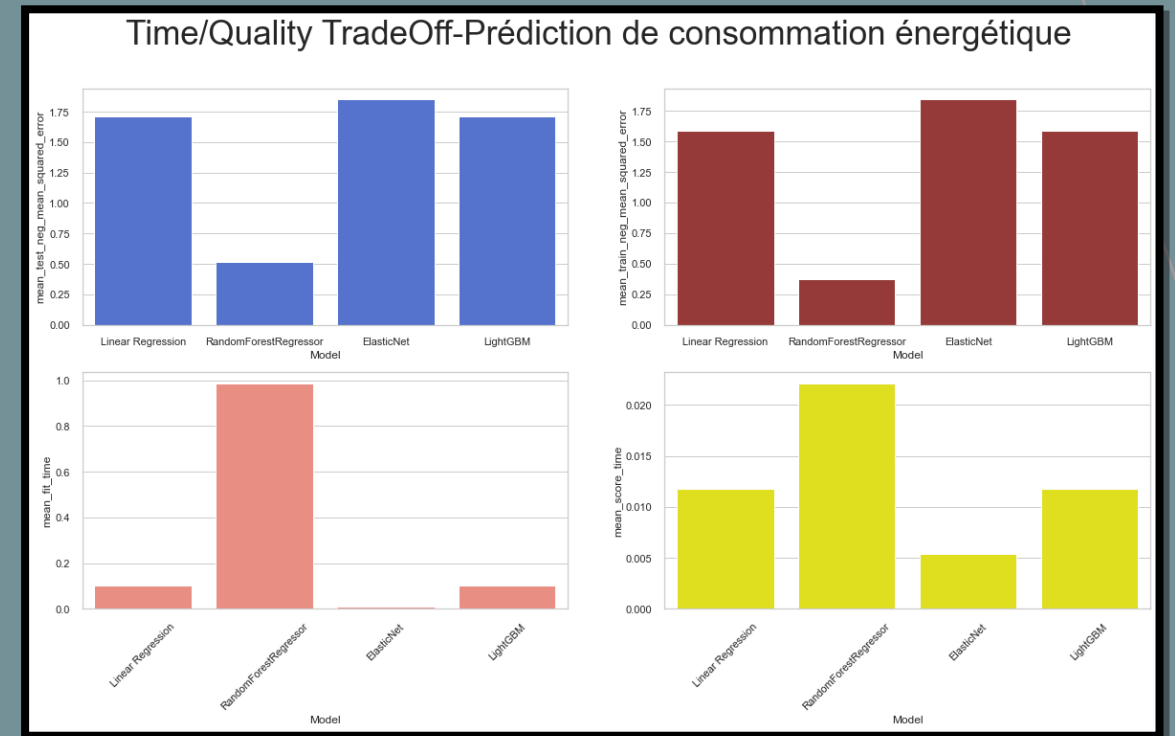
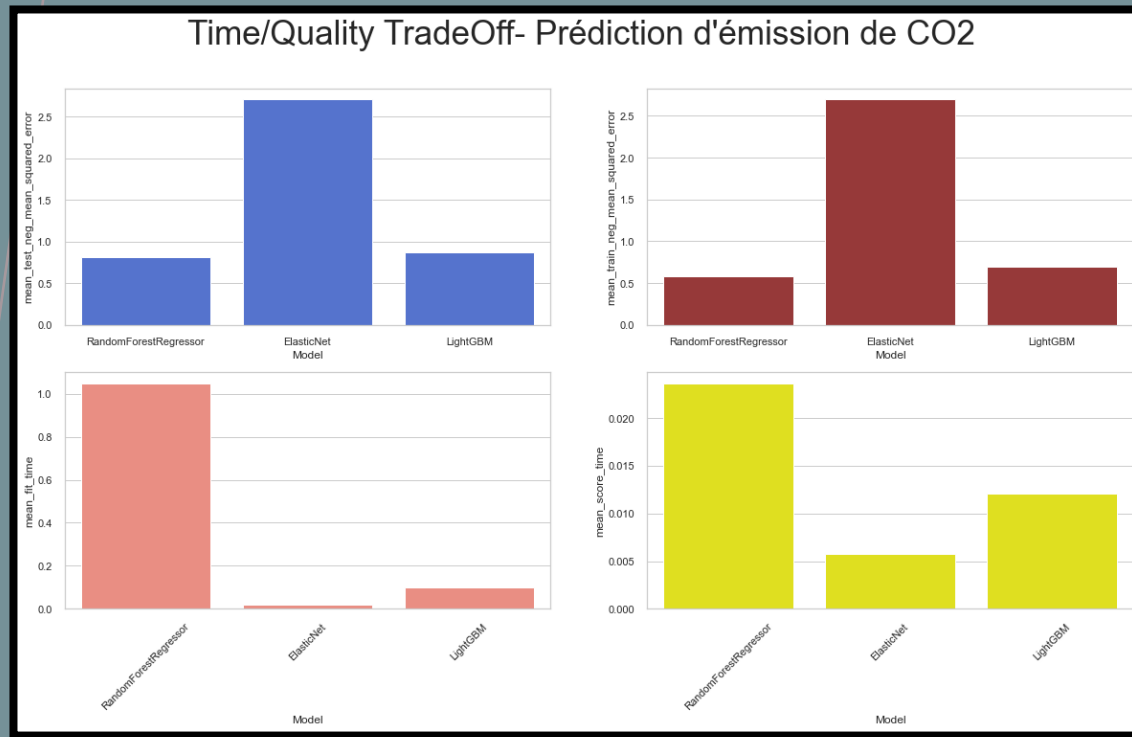
Tester différentes configurations des hyperparamètres lors de l'apprentissage du modèle, et de retenir celle qui minimise le taux d'erreur

Grid Search

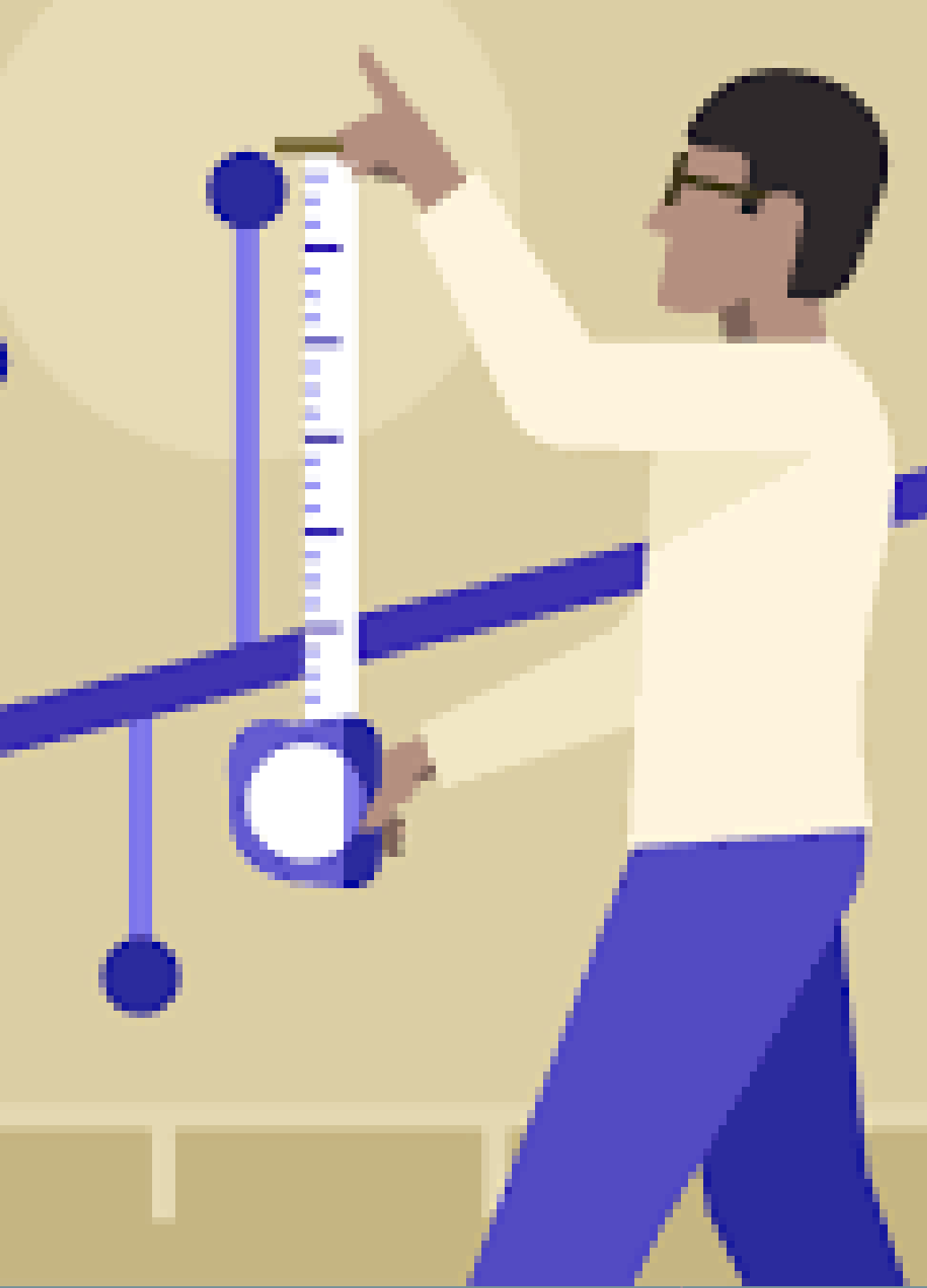
Grid Search configure une grille de valeurs des hyperparamètres et sélectionne leur combinaison pour créer un meilleur modèle

GridSearchCV

Comparaison des modèles



La projection graphique montre que le modèle LightGBM offre le meilleur compromis score / temps



ÉVALUER LA PERFORMANCE DU MODÈLE CHOISI

Mesures de la performance algorithme régression



Coefficient de détermination



Mean Squared Error



Mean Absolute Error



Racine carré de MSE

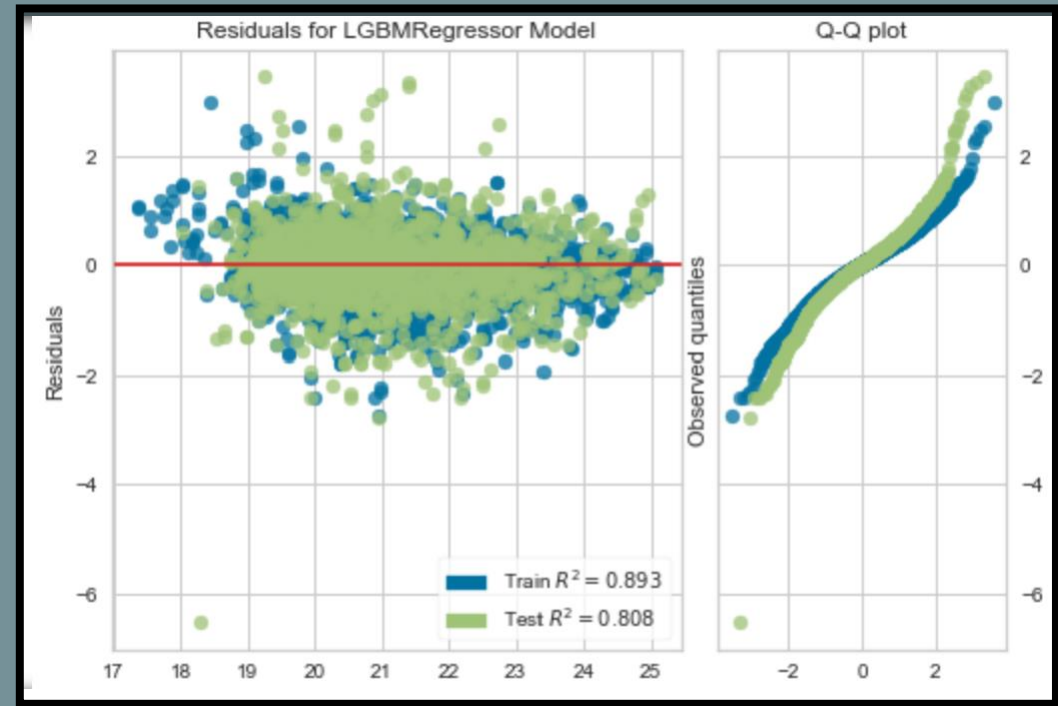
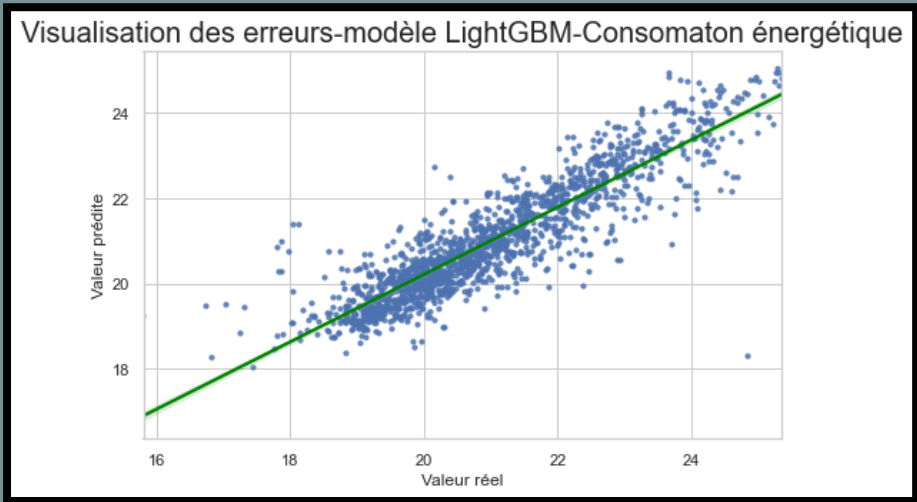


Root Mean Squared
Log Error

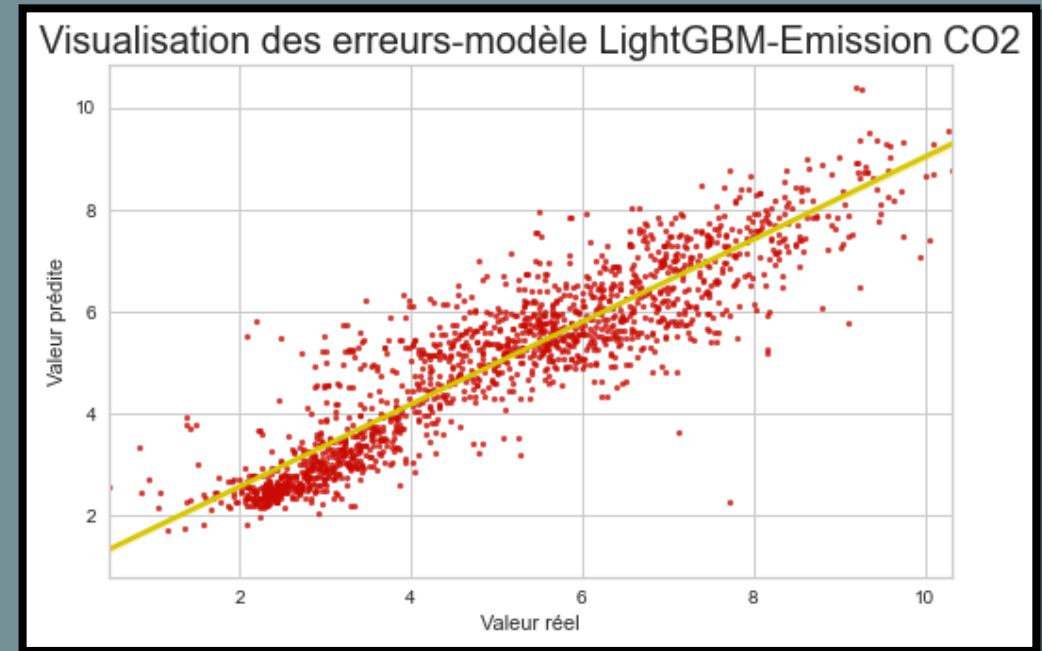
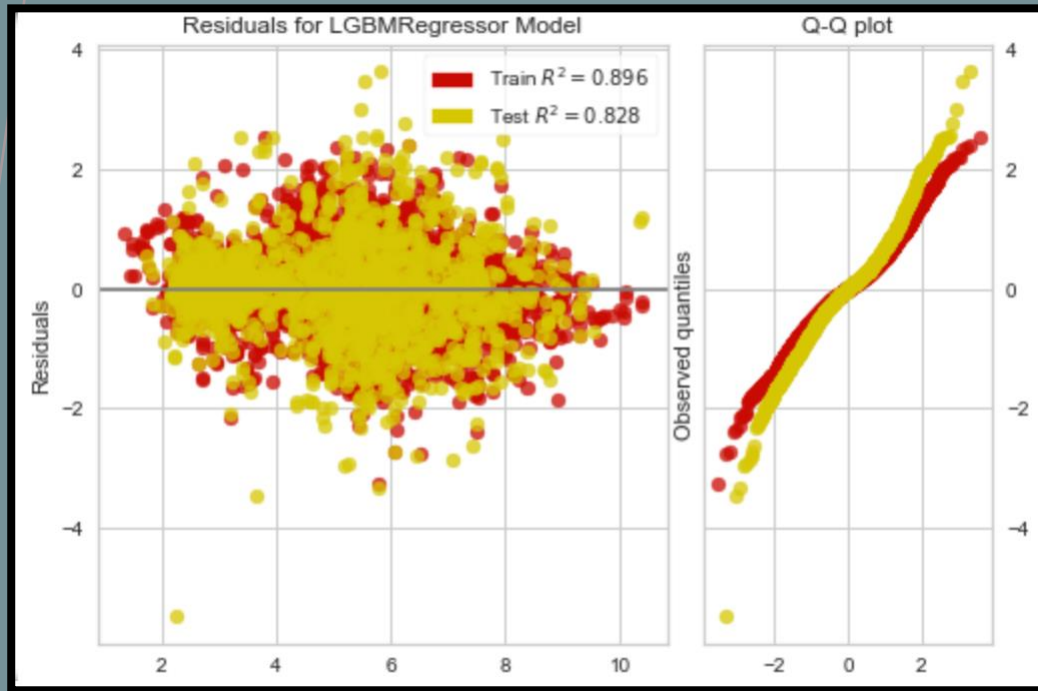


Mean Absolute
Percentage Error

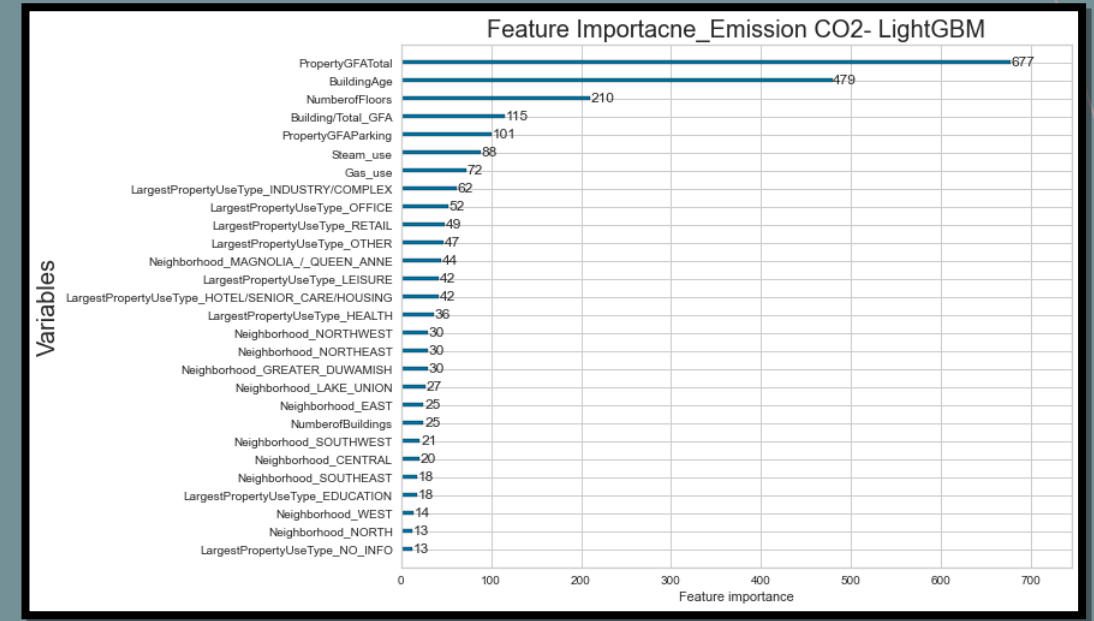
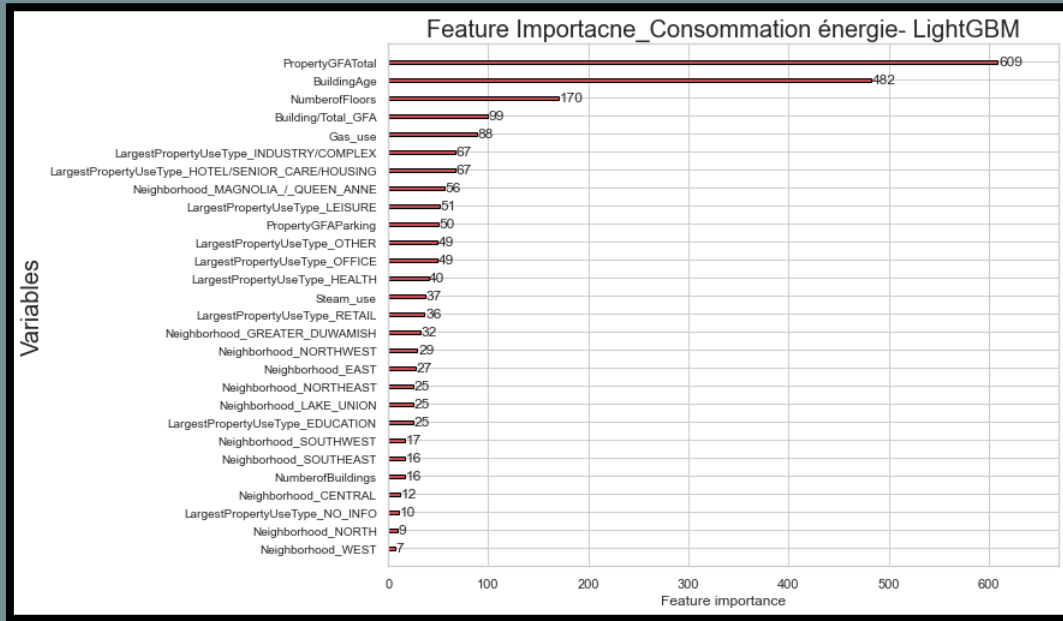
Visualisation de la performance du modèle de la prédiction de la Consommation énergétique



Visualisation de la performance du modèle de la prédiction d'émission de CO2



Importance des variables pour les deux modèles



- Le résultat de la prédiction dépend principalement de la superficie, et de l'année de construction des établissements
- Le quartier et le type d'usage des établissements ont des impacts modestes sur le modèle



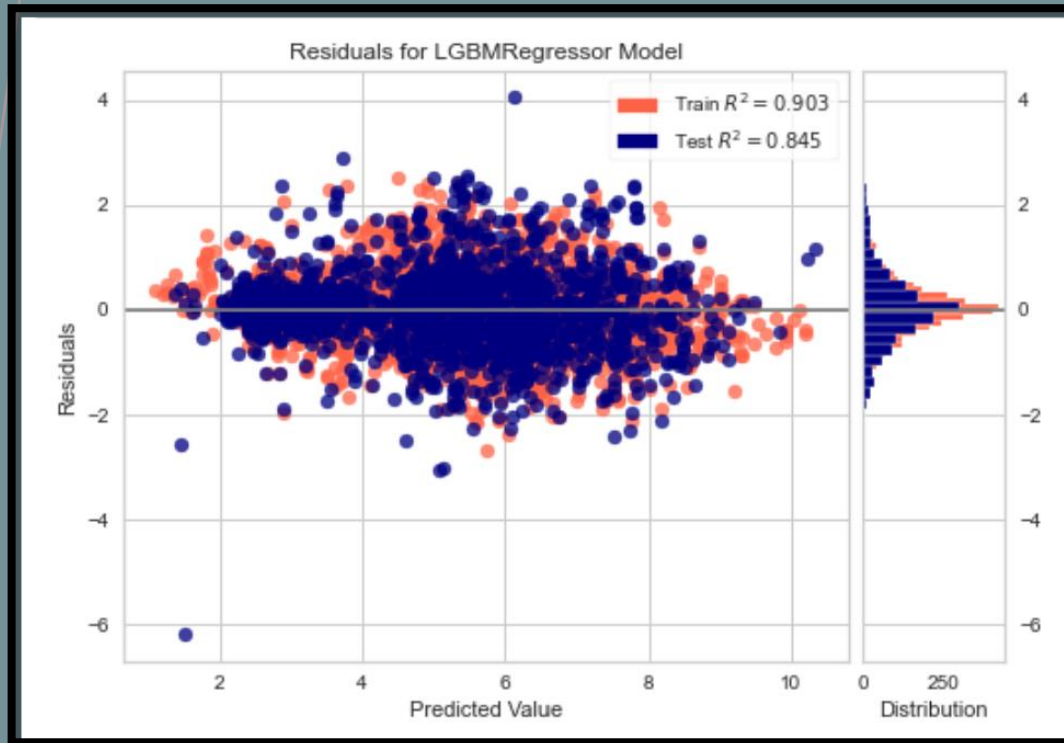
INTÉRÊT DE
« ENERGY STAR
SCORE » POUR
LE MODÈLE

La description de “EnergyStarScore

- Mesure de la performance énergétique des bâtiments
- Basée sur un échelle de 1 à 100.
- Plus le score est élevé, meilleure est la performance énergétique du bâtiment



Visualisation de la performance du modèle de la prédiction d'émission de CO2



```
print("Meilleurs modèle pour l'émission de CO2' :\n", '\n', lgb_gaz_s.best_estimator_, "\n")  
lgb_gaz_pred_s = lgb_gaz_s.best_estimator_.predict(xtest_s)  
print("Score du model final pour l'émission de CO2_avec EnergyStarScore : ")  
evaluate(ytest_gaz_s['TotalGHGEmissions'].values, lgb_gaz_pred_s)
```

Meilleurs modèle pour l'émission de CO2' :

LGBMRegressor(max_depth=8, min_child_samples=2, num_leaves=21)

Score du model final pour l'émission de CO2_avec EnergyStarScore :

Coefficient de détermination (R^2) = 0.845

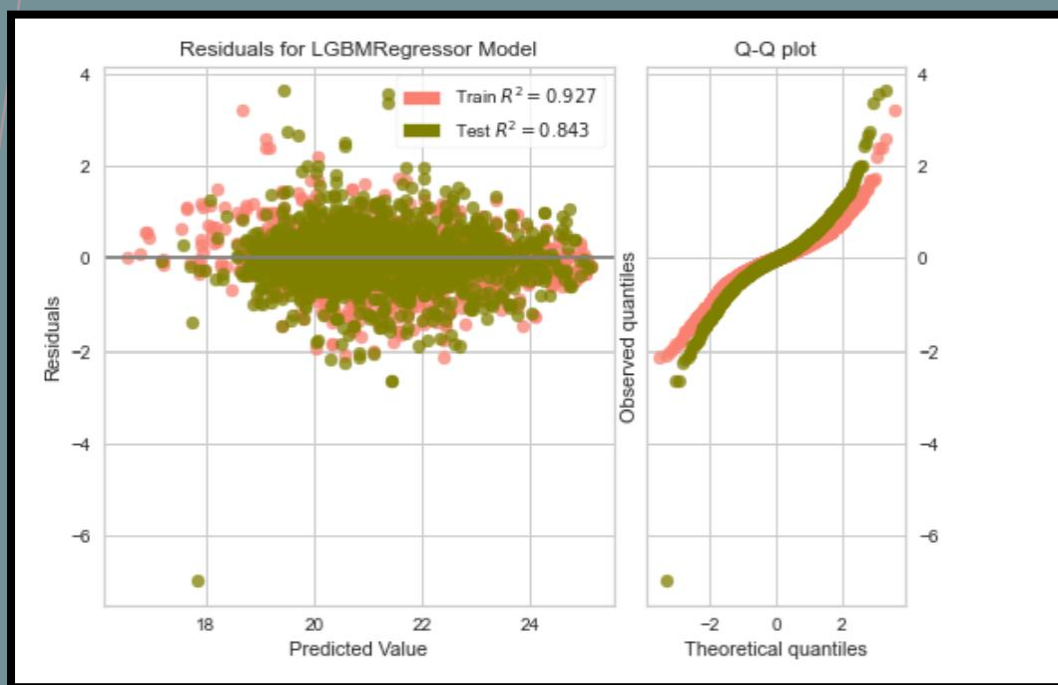
Le carré moyen des erreurs (MSE) = 0.625

Racine carré de MSE (RMSE) = 0.791

Mean absolute Error (MAE) = 0.561

	Métrique	Résultats
0	MSE	0.625163
1	R2	0.844942
2	RMSE	0.790672
3	MAE	0.561402

Visualisation de la performance du modèle de prédiction de consommation énergétique



```
print("Meilleurs modèle pour la consommation de l'energy :\n", '\n', lgb_energy_s.best_estimator_, "\n")
lgb_energy_pred_s = lgb_energy_s.best_estimator_.predict(xtest_s)

print("Score du model final pour la consommation énergétique_avec EnergyStarScore : ")
evaluate(ytest_energy_s['SiteEnergyUse(kBtu)'].values, lgb_energy_pred_s)
```

Meilleurs modèle pour la consommation de l'energy :

LGBMRegressor(max_depth=9, min_child_samples=6, num_leaves=29)

Score du model final pour la consommation énergétique_avec EnergyStarScore :

Coefficient de détermination (R^2) = 0.843

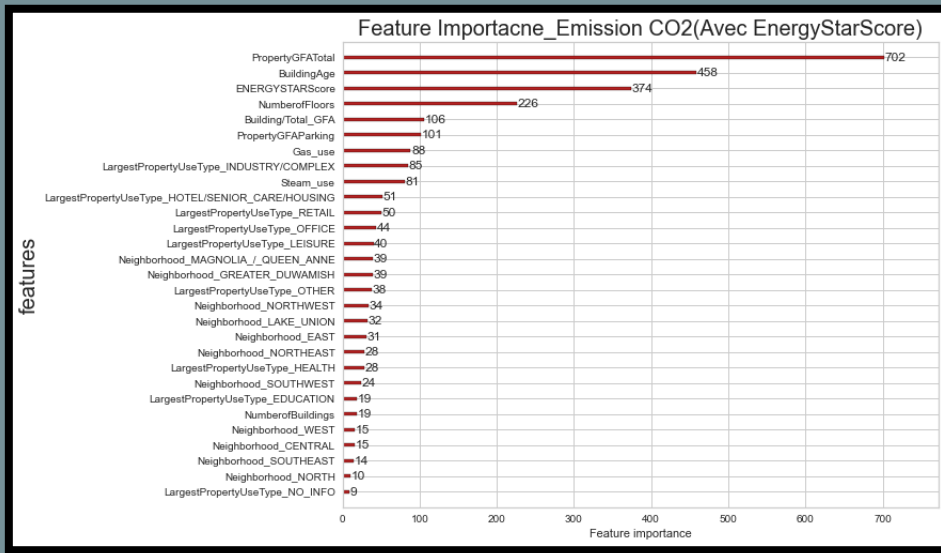
Le carré moyen des erreurs (MSE) = 0.376

Racine carré de MSE (RMSE) = 0.614

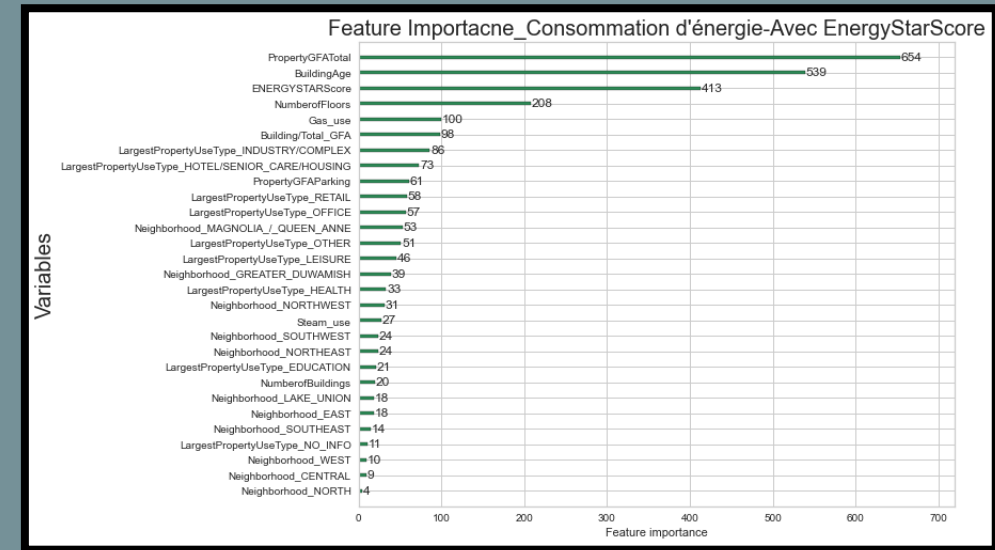
Mean absolute Error (MAE) = 0.411

	Métrique	Résultats
0	MSE	0.376383
1	R2	0.842522
2	RMSE	0.613500
3	MAE	0.411012

Evaluation de l'importance de chaque variable pour le modèle



Le résultat de la modélisation avec *EnergyStarScore* montre que ce dernier a des impacts importants sur le modèle.



RÉCAPITULATIF

1. Préparation des données

2. Analyse exploratoire et descriptive

3. Modélisation

4. Data preprocessing

5. Tester 5 algorithmes de régression

6. Optimiser les hyperparamètres

7. Sélectionner le modèle performant : LighGBM

8. Entraînement du modèle sélectionné


9. R² de plus de 80% pour les deux modèles

10. Evaluer l'intérêt de la variable "*Energy score*" pour le modèle

SYNTHÈSE ET CONCLUSION



« EnergySTARScore » améliore la performance de 2 modèles



« EnergySTARScore » est fatidieux à calucler et la performance du modèle est faiblement améliorée





**MERCI DE VOTRE
ATTENTION**