



Seattle

PROJET 4 - “ Anticipez les besoin en consommation de bâtiments”

Soutenance de projet Juillet 2023

Mauyves NKONDO
Data Scientist

- ▶ **Problématique**
- ▶ **Préparation du jeu de données**
- ▶ **Pistes de modélisations**
- ▶ **Présentation du modèle final**



Seattle



PROBLÉMATIQUE

- Problématique
- Présentation de la problématique
- Interprétation de la problématique




Seattle



Seattle

Afin d'atteindre l'objectif de devenir une ville neutre en émissions de carbone d'ici 2050, la ville de Seattle accorde une grande importance à l'évaluation des émissions provenant des bâtiments non destinés à l'habitation.

Des relevés minutieux ont été effectués en 2016. Cependant, ces relevés sont coûteux à réaliser, et compte tenu de ceux déjà effectués, nous devons maintenant chercher à prédire les émissions de CO₂ et la consommation totale d'énergie des bâtiments pour lesquels ces données n'ont pas encore été mesurées.



- Données de consommation disponibles pour les bâtiments de la ville de Seattle pour l'année 2016.

- Coût important d'obtention des relevés / fastidieuses à collecter.
- Nous avons un jeu de données contenant 3376 lignes , 46 colonnes et aucune observation dupliquée.

La Mission :

- Prédire les émissions de CO2 et la consommation totale d'énergie sans les relevés annuels
- Evaluer l'intérêt de l'ENERGY STAR Score
- Mettre en place un modele de prediction performants réutilisable



Seattle

- Préviation :

- Features : Caractéristiques intrinsèques des bâtiments (hors consommation)
- Données à prédire :
 - Consommation totale des bâtiments [SiteEnergyUseWN\(kBu\)](#)
 - Emissions totales des batiments [TotalGHGEEmissions](#)
-  [2 modèles différents](#)
- Type de Bâtiments : Non résidentiel
- ENERGY STAR Score :
 - Comparaison de son intérêt en essayant de modéliser avec et sans




PRÉPARATION DU JEU DE DONNÉES

- Nettoyage
- Feature Engineering
- Exploration

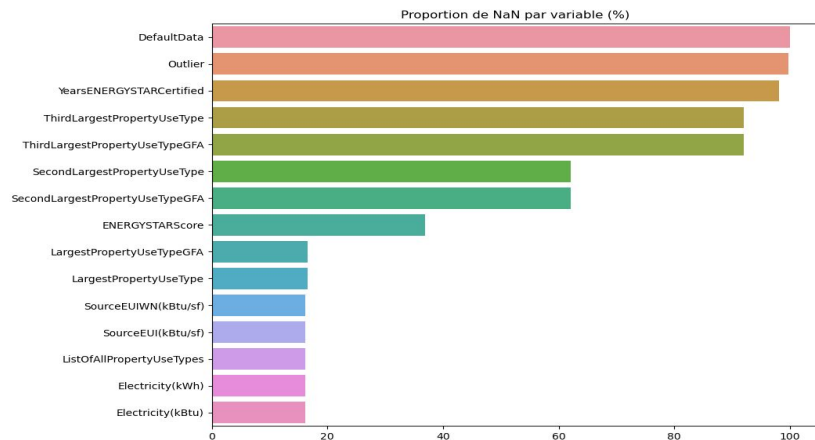


Seattle

- **Valeur manquantes (NaN) :**

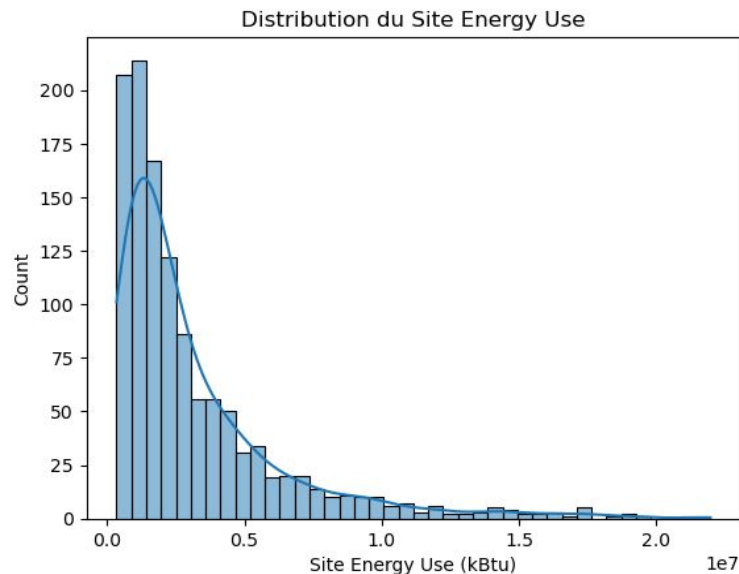
- Complétion des valeurs manquantes.

- Suppression des observations pour lesquelles on a beaucoup de NaN pour conserver un maximum de features.

Représentons la proportion de valeurs manquantes (NaN) pour chaque variable d'un jeu de données data sous forme de pourcentage.



- **Suppression des Outliers :**

- Outliers univariés
- Outliers multivariées (distance aux 5 plus proches voisins)



Feature Engineering

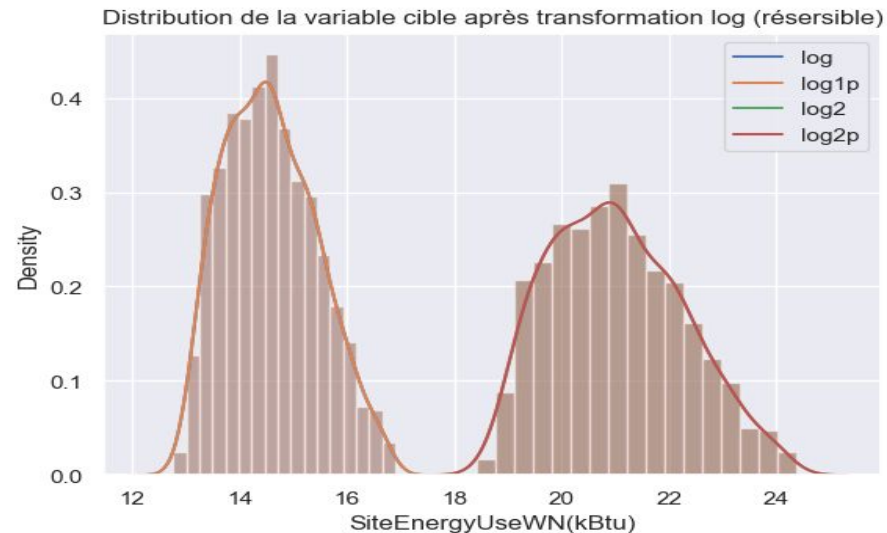
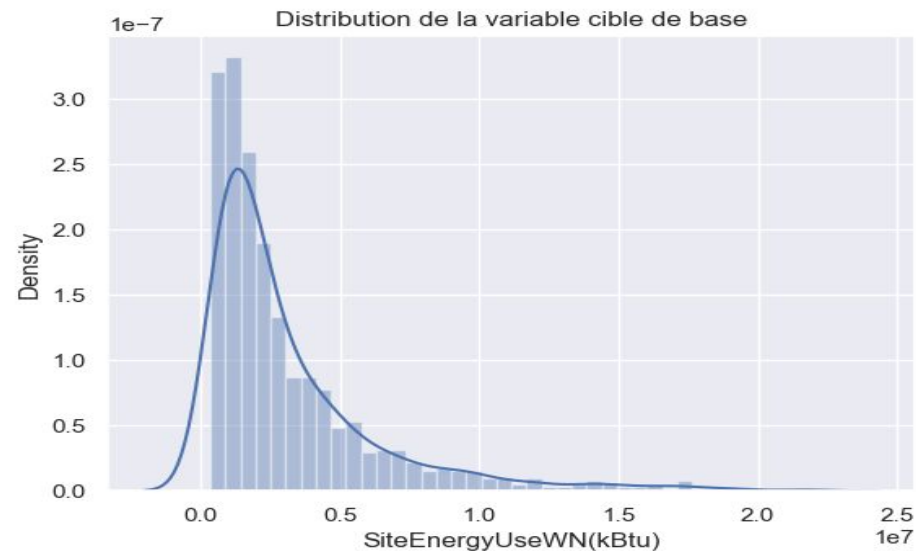
✓ Idées écartées :

- Features liées aux compteurs (coûteux à obtenir pour les futures données)
- Energy Star score est mis de côté pour une analyse ultérieure.

✓ Idées retenues :

- Suppression des features de consommation à l'exception des 2 features que nous cherchons à prédire
- Catégorisation des données pour certaines colonnes (Usage)
- One Hot Encoding : Transformation d'une feature avec n catégories en n features booléennes.
- Suppression de colonnes non pertinentes pour notre modèle
 - Données sans catégorisation possible (Comment)
 - Données avec une unique information (exemple : State)
 - Données sans information pertinente pour le modèle (voir exemple)
 - Default : sens de la feature non expliqué + booléen avec beaucoup de NaN

Log2 - transformation variable de prédiction

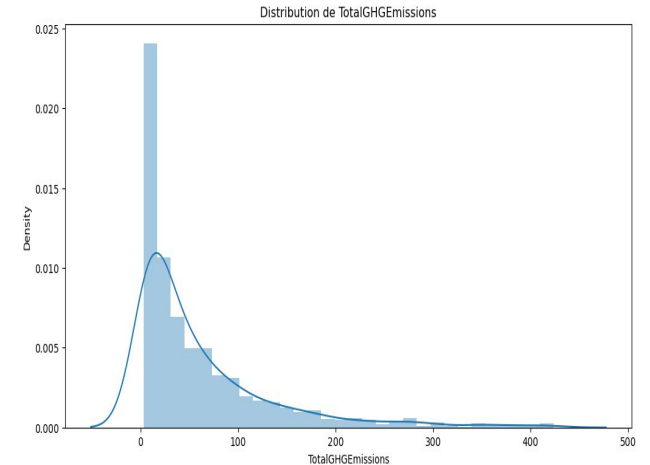
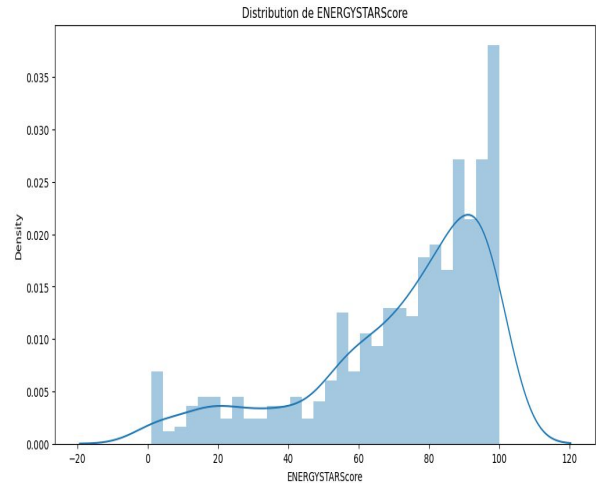
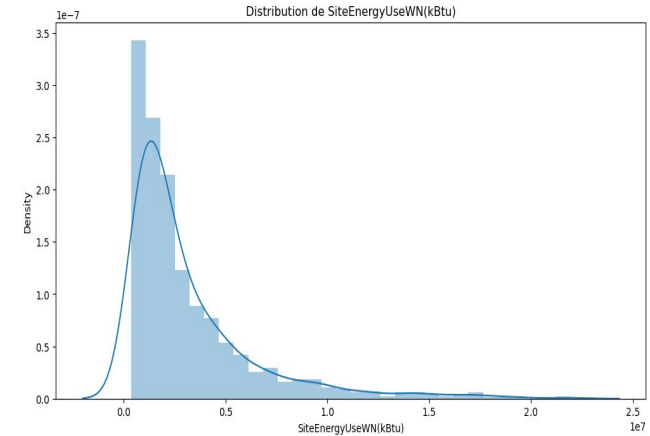
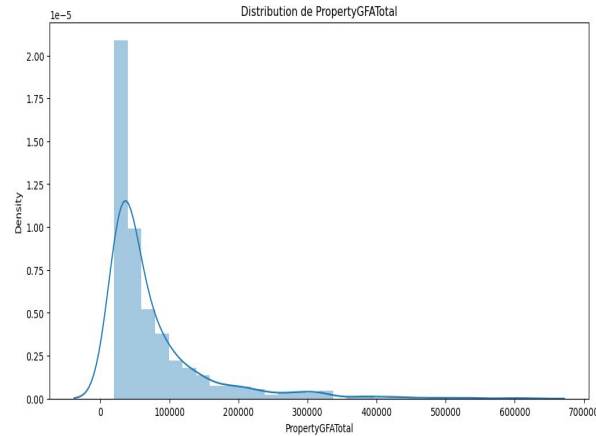


- La variable **SiteEnergyUseWN(kBtu)** de prédiction dans un modèle de machine learning a été transformée à l'aide du logarithme base 2. Cette transformation est utile notamment pour réduire l'effet d'échelle de la variable, ou pour rendre les données plus linéaires si elles présentent une tendance exponentielle.

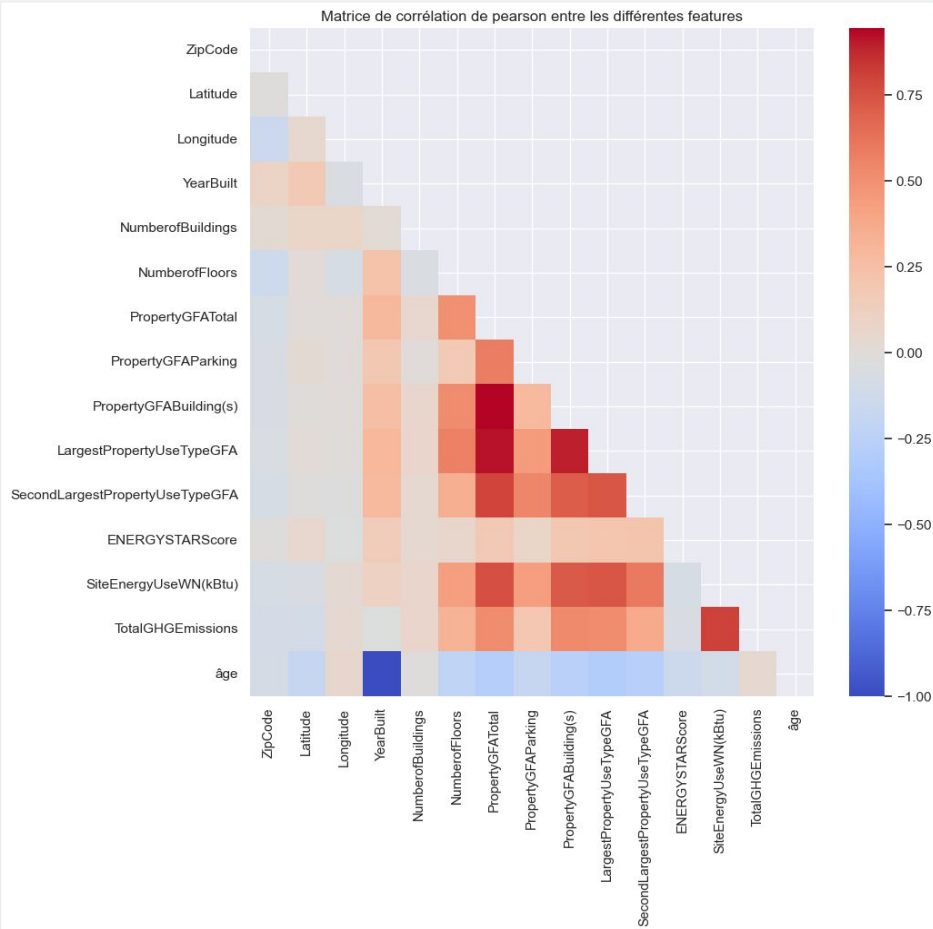
Analyse Exploratoire



On observe la distribution de différentes variables liées aux bâtiments : **PropertyGFATotal**, **SiteEnergyWN(Kbtu)**, **EnergyStarScore**, et **TotalGHGEmissions**.



Exploration : Corrélations



Point Majeurs :

Consommation : Corrélation importante avec :

- PropertyGFATotal,
- PropertyGFABuilding,
- LargestPropertyUseTypeFGA

Emissions : mêmes corrélation (dans moindre mesure) + corrélation importante avec la corrélation.

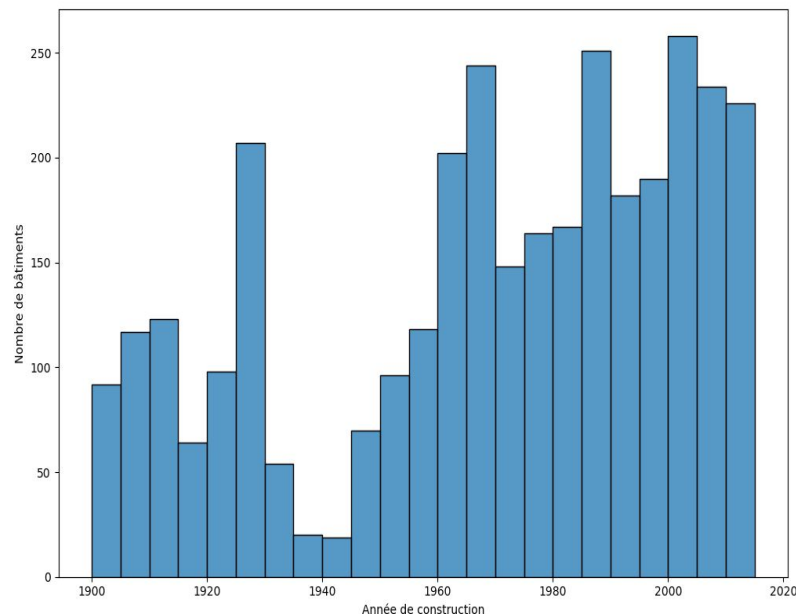
Autres points notables :

- Corrélation importante entre
 - PropertyGFATotal et PropertyGFABuilding
 - ,PropertyGFATotal et LargestPropertyUseTypeFGA
 - LargestPropertyUseTypeFGA et PropertyGFABuilding
- Energy Star Score : pas de corrélation notable

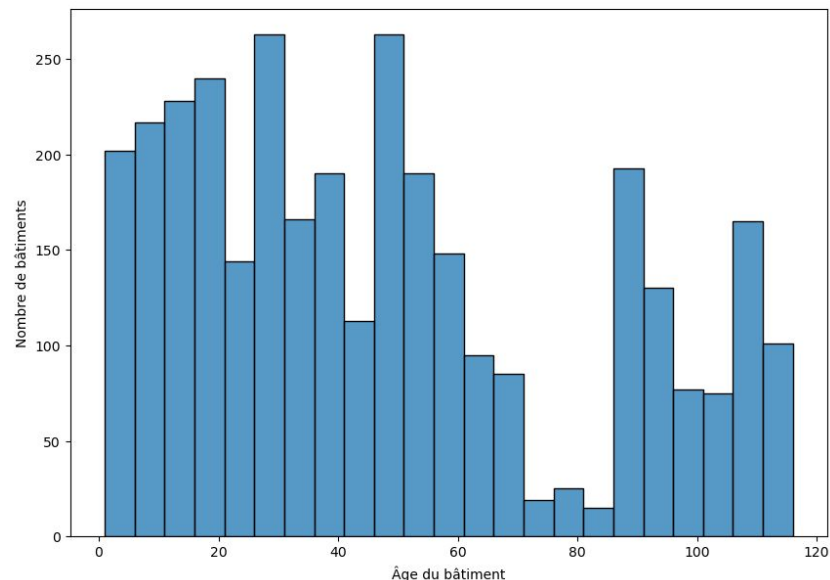


Distribution des années de construction des bâtiments permet d'analyser comment les bâtiments ont été construits au fil du temps dans cette ville de Seattle.

Distribution des années de construction des bâtiments



Distribution de l'âge des bâtiments



Distribution de l'âge des bâtiments permet de visualiser comment les bâtiments sont répartis en fonction de leur âge, ce qui peut être utile pour comprendre l'évolution.



PISTE DE MODÉLISATIONS

- Découpage de l'échantillon
- Modèles entraînés
- Modèle de consommation : Démarche
- Modèle de consommation : Hyperparamètres
- Modèle Émission : Démarche



Seattle

Données d'entraînement vs. données de test

Utilisation de l'ensemble des données d'entraînement et de test par **la Validation croisée**

Le but est de conserver une certaine cohérence et surtout une représentativité

Fonction `train_test_split()` :

- 20% attribués au test

Validation Croisée :

- plis-folds fixé à 5
- éviter le biais potentiel
- pas d'évaluation unique



Seattle

Modèles entraînés

Choix de modèles

Plusieurs algorithmes sont utilisées afin de pouvoir comparer leurs performances.

Critères de performances traités :

R² - RMSE

L'erreur des modèles les plus performants sera visualisée graphiquement

Modèle simple :

- Ridge Regressor

Méthodes ensemblistes :

- Random Forest Regressor
- Gradient Boosting Regressor
- XGBoost



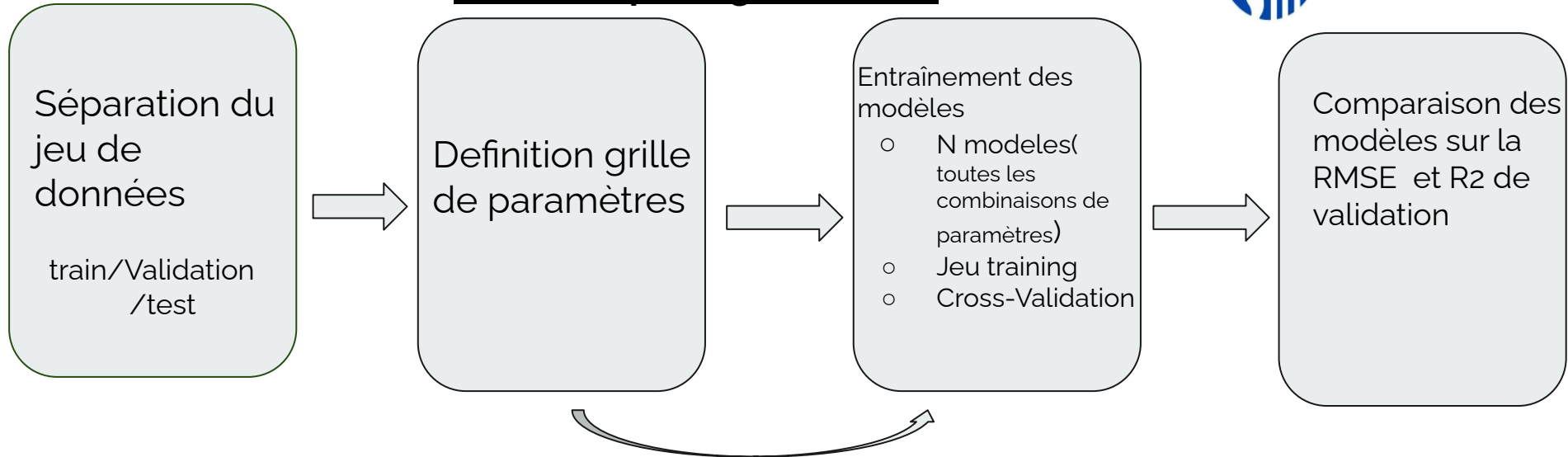
Seattle

Modèle de Consommation : Démarche



Seattle

Pour Chaque Algorithme (*)



L'affinage des paramètres

- l'affinage des paramètres est un processus itératif d'ajustement de ces algorithmes dans le but d'optimiser la performance globale du modèle. Cela peut impliquer de faire des essais et des erreurs, de comparer différents algorithmes, de faire usage de techniques d'optimisation, ou d'utiliser des approches plus sophistiquées telles que la descente de gradient, le réglage des hyperparamètres, l'évaluation croisée, etc.

Modèle de Consommation : Hyperparamètres

Meilleurs hyperparamètres en vert

Hyperparamètres

Après l'ajustement des caractéristiques métier, optimisation des principaux hyperparamètres : régresseur (alpha), arbres de décision (n_estimators)...

- Une méthode est appliquée :

GridSearch

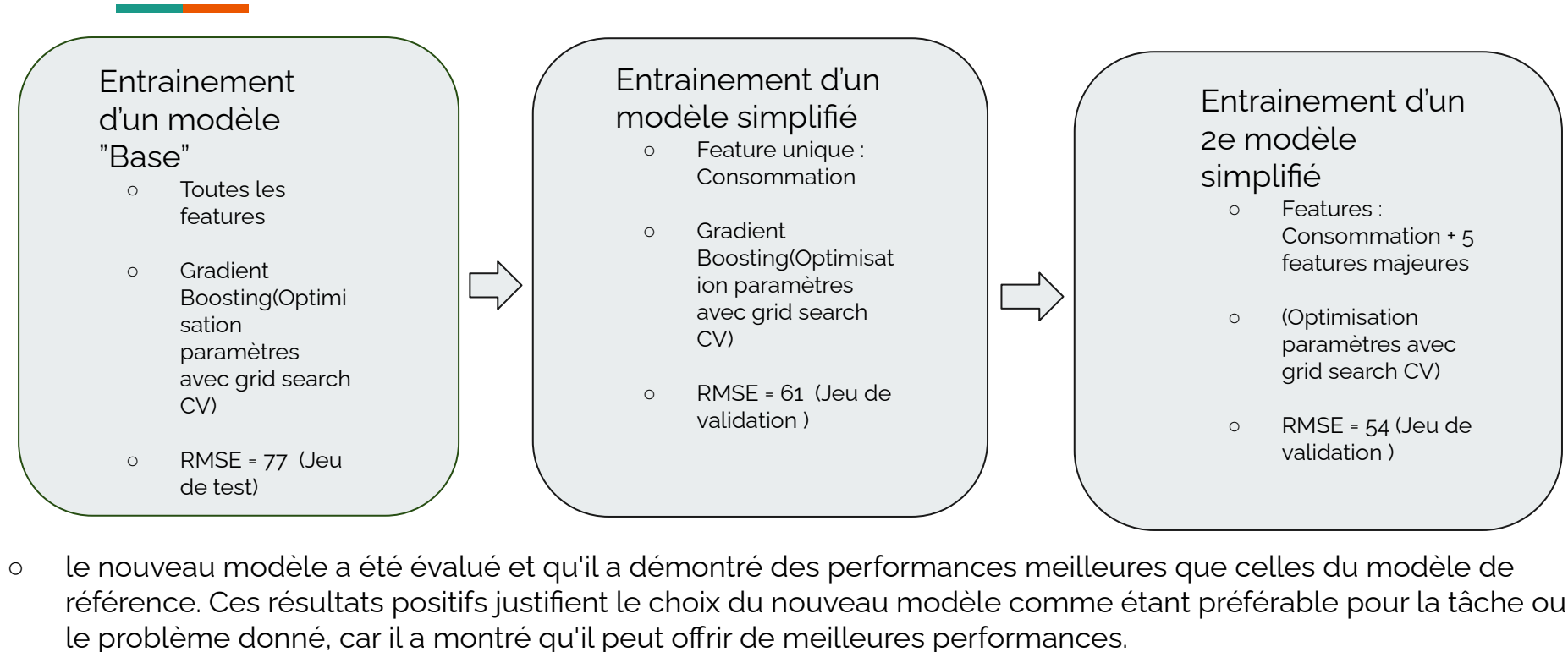
pour couvrir l'espace des valeurs pertinentes des hyperparamètres

- "**Combinaison optimale des hyperparamètres**" représente le processus de recherche et de sélection des meilleurs paramètres pour obtenir les performances les plus élevées dans un système ou un modèle donné.

<u>Ridge Regressor</u>	<u>Random Forest Regressor</u>	<u>Gradient Boosting Regressor</u>	<u>XGBoost</u>
{ 'Alpha ' : 4.89 }	'n_estimators' : [10, 50, 100, 300, 500]	'n_estimators' : [50, 100, 150 , 200]	'n_estimators' : [100 , 500, 1000, 2000]
	'min_samples_leaf' : [1 , 3, 5, 10]	'learning_rate' : [0.01, 0.1 , 0.5]	
	'max_features' : ['auto', ' sqrt ']	'max_depth' : [3 , 5, 7]	

Modèle Émissions : Démarche

Idée : Faire un modèle simplifié à partir de la prédiction de Consommation.





Seattle

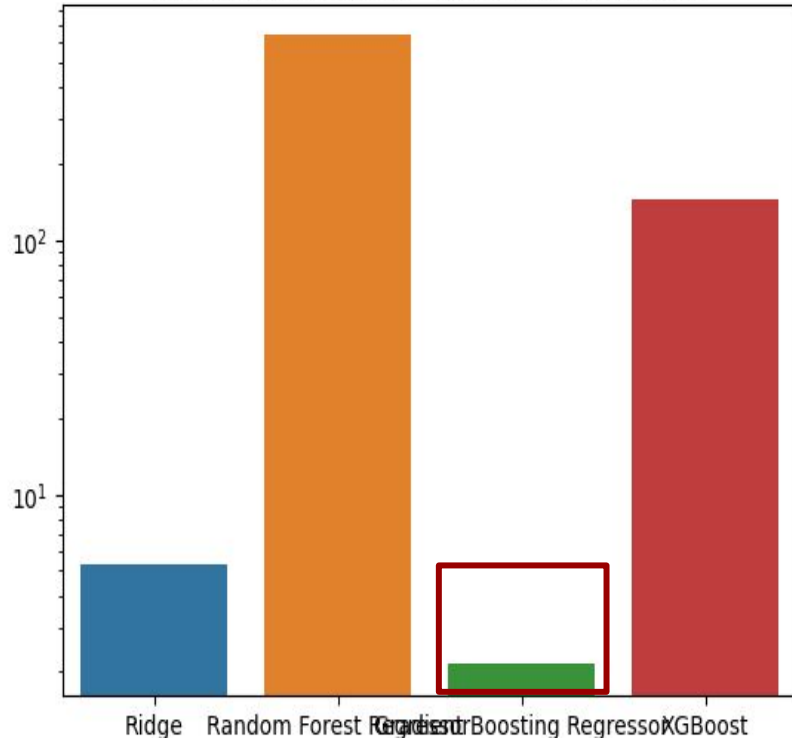
PRÉSENTATION DU MODÈLE FINAL

-
- Modèle Obtenus (Consommation)
 - Modèle Final
 - Intérêt de l'energy star score

Modèle Obtenus (Consommation)

"Comparaison sur le jeu de validation"

Temps d'exécution des algorithmes pour la prédiction
(jeu d'entraînement) - échelle logarithmique



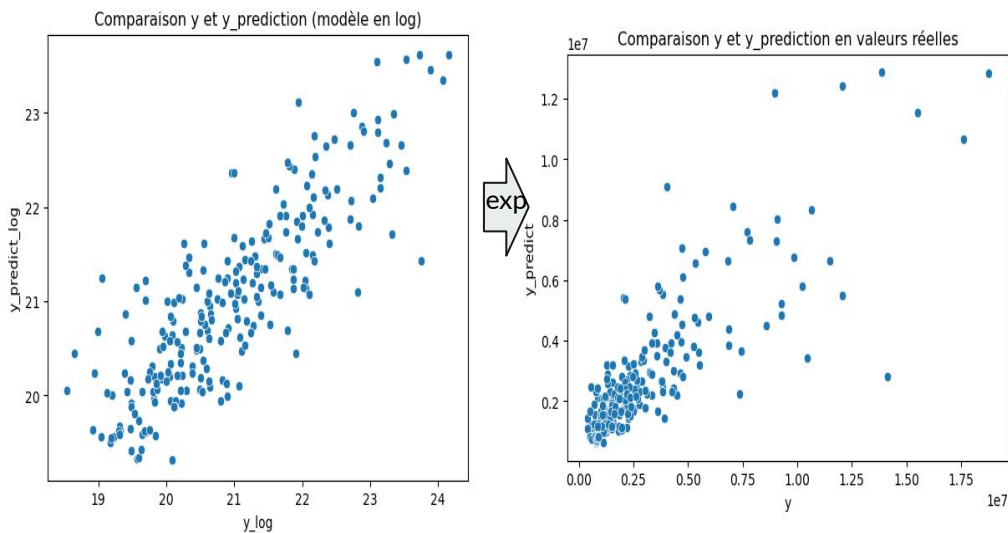
4 Algorithmes testés et comparés.
Sans et avec validation croisée

La tendance générale est meilleure avec GridSearch.

	RMSE	R ²	RMSE_relative
Ridge	0.750561	0.606933	1.0
Random Forest Regressor	0.651582	0.703767	0.868128
Gradient Boosting Regressor	0.639401	0.714739	0.851898
XGBoost	0.673558	0.683448	0.897406

- Le Gradient Boosting Regressor semble plus rapide à entraîner.

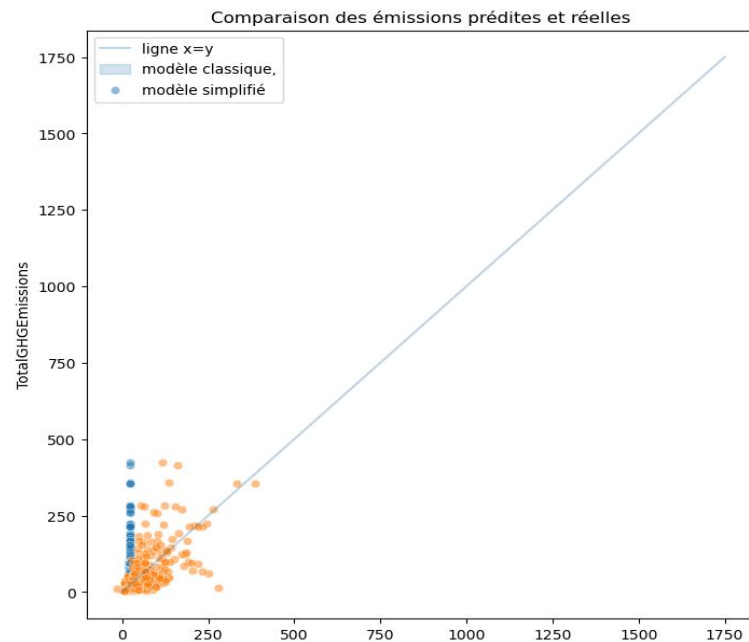
Une méthode ensembliste permet de gagner en précision avec un temps de calcul acceptable pour le Gradient Boosting Regressor.



Les échantillons présentent certaines valeurs extrêmes...

La décision de maintenir ces valeurs a été prise en réponse au contexte métier..

Les algorithmes robustes permettront de modéliser sans éliminer ces valeurs.



Des valeurs atypiques sont clairement observables.

Visuellement, ces distributions présentent des queues longues.



- La feature a été traitée séparément du modèle initial en raison d'une quantité réduite de données disponibles.
- L'entraînement d'un modèle Gradient Boosting Regressor avec une recherche de grille (Grid search CV) sans energy star score a abouti à un best score de 0.79, ce qui est légèrement meilleur que la valeur avec energy star score de 0.85. Cela a conduit à une légère amélioration des performances du modèle.
- Le RMSE obtenu sur le jeu de test $0.44 < 0.58$ améliore légèrement la performance de l'Energy star score.



Problématique de régression identifiée et traitée.

4 algorithmes traités et optimisés en validation croisée nous donne une meilleure performance en méthode ensembliste par le choix d'un modèle en Gradient Boosting Regressor.

Le résultat obtenu avec l'Energy Star Score est relativement proche de celui du modèle initial, mais son impact semble aléatoire; par conséquent, il n'est pas impératif de le conserver.



Seattle

MERCI DE VOTRE ATTENTION

○ Fin.