Questions :

- Vous avez justifié le choix final de l'algorithme et des hyperparamètres.  → Je dois justifier la selection de chaque hyperparameter pour chaque model ? Sinon le choix c’est simplement le meilleur en fonction du scoring.

- diff entre feature engineering et l’approche de modélisation ? Mon approche comporte également les feature engineering.

- I input -1 for missing Energyscore and cannot remove the empty lines. Should I use here the mean/median or something else instead ? I’m afraid of inputing something that would skewer the results but -1 already does.

Tu prends le meilleur modèle, ajoute la colonne d’energy, supprime les valeurs nulles. Puis tu refais 2 pass de gridsearch avec et sans cette colonne.

- et locale sur chaque individu du jeu de données. → Shap et Lime sur chaque individu ? Ou juste en prendre 2-3 pour l’analyse locale suffirait ? 2-3 oui.

- Why use SiteEnergyUse and not the source qui inclus toute la chaine de l’énergie ? Why not EUI qui normalize par surface ? https://www.opendatanetwork.com/dataset/data.seattle.gov/h7rm-fz6m

Prend la non normalisée niveau emission aussi.

Inclure dans la prez pour chaque modèle les hyperparamètres (mostly data) du meilleur score ?

Justifie dans prez que ton target encoding ninduit pas de leakage parce que ton fit/transform only uses the training set. Vu que jutilise une pipeline cest bon de toute facon.

One really important effect is the Target Leakage. By using the probability of the target to encode the features we are feeding them with information of the very variable we are trying to model. This is like “cheating” since the model will learn from a variable that contains the target in itself.

Vous avez vérifié qu'il n’y a pas de problème de data leakage (c'est-à-dire, des variables trop corrélées à la variable cible et inconnues a priori dans les données en entrée du modèle).

Prez :

Trop de variance vu le petit echantillon de donnees.

##### Sélectionner, transformer et créer les variables pertinentes d'un modèle d'apprentissage supervisé ou non supervisé (feature engineering) afin de réaliser un apprentissage optimal.

CE1 Vous avez identifié les variables catégorielles (qualitatives).

CE2 Les variables catégorielles identifiées ont été transformées en fonction du besoin (par exemple via OneHotEncoder ou TargetEncoder).

CE3 Vous avez créé de nouvelles variables à partir de variables existantes.

CE4 Vous avez réalisé des transformations mathématiques lorsque c'est requis pour transformer les distributions de variables.

CE5 Vous avez normalisé les variables lorsque c'est requis.

##### Sélectionner, entraîner et évaluer des modèles d’apprentissage supervisé, afin de choisir le modèle le plus performant pour la problématique métier.

CE1 Vous avez choisi la ou les variables cibles pertinentes. Cela signifie dans le cadre de ce projet :

* Vous avez choisi la variable cible pour la consommation totale d’énergie,
* Vous avez choisi la variable cible pour la consommation de CO2.

CE2 Vous avez vérifié qu'il n’y a pas de problème de data leakage (c'est-à-dire, des variables trop corrélées à la variable cible et inconnues a priori dans les données en entrée du modèle).

CE3 Vous avez choisi des métriques adaptées pour évaluer la performance d'un algorithme (par exemple : R2, RMSE, MAE en régression.)

CE4 Vous avez exploré d'autres indicateurs de performance que le score pour comprendre les résultats (par exemple : coefficients des variables en fonction de la pénalisation, visualisation des erreurs en fonction des variables du modèle, temps de calcul...).

CE5 Vous avez séparé les données en train/test pour les évaluer de façon pertinente et détecter l'overfitting.

CE6 Vous avez mis en place une validation croisée (via GridsearchCV, RandomizedSearchCV ou équivalent) afin de réaliser une Cross Validation sur le set de train et d’optimiser les hyperparamètres.

CE7 Vous avez testé au minimum 4 algorithmes de famille différente (par exemple : ElasticNet, SVM, GradientBoosting, RandomForest). Vous avez présenté une synthèse comparative de l'ensemble des résultats. Vous avez justifié le choix final de l'algorithme et des hyperparamètres.

CE8 Vous avez réalisé l’analyse de l’importance des variables (feature importance) globale sur l’ensemble du jeu de données et locale sur chaque individu du jeu de données.