PROJET 3 : ANTICIPEZ LES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BATIMENTS

SOUTENANCE OPENCLASSROOMS, LE XX/04/2022

ERWAN CHESNEAU

Seattle

PLAN :

I. Contexte
II. Nettoyage de la base de données
III. Tests de prédictions
IV. Interprétation du modèle sélectionné
V. Conclusions
VI. Améliorations à envisager

I. CONTEXTE:

- La ville de Seattle étudie les émissions de GES et la consommation d'énergie des batiments dans un objectif de ville neutre en émissions carbone.
- Un relevé a déjà été réalisé en 2015 et 2016 pour mesurer ces valeurs.
- Cout important de ces relevés

➔ La ville voudrait prédire la consommation d'énergie et les émissions de GES des batiments manquants

➔ La ville veut connaitre l'importance de l'ENERGY STAR Score

- Démarches :
  - Extraire les variables d'intérets dans le jeu de données
  - Calculer de nouvelles variables
  - Tester différents modèles pour évaluer leurs performances
  - Extraire et optimiser le meilleur modèle
  - Conclure sur la faisabilité et la performance escomptée

II. NETTOYAGE : HARMONISATION DES JEUX DE DONNÉES

- 2 jeux de données disponibles : 2015 et 2016
  - Vérifications que les variables des 2 années sont égales
- Liste des variables différentes et actions effectuées :

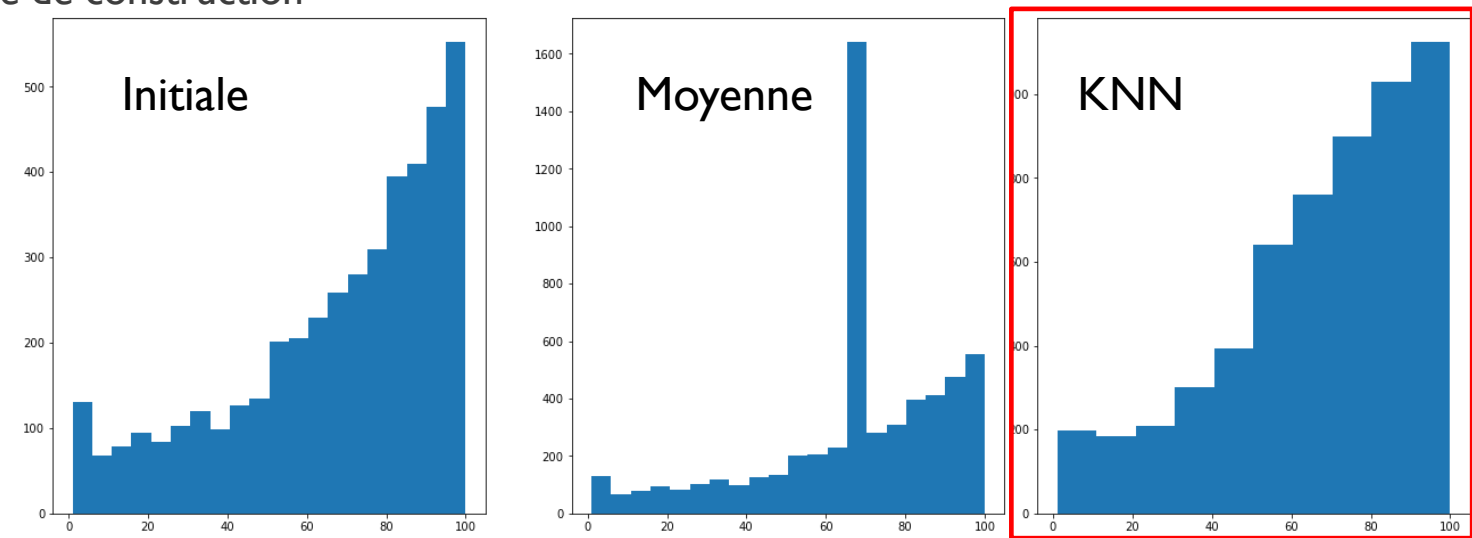
Variables	Actions
Comment	Equivalent à Comments
GHGEmissionsIntensity	Equivalent à GHGEmissionsIntensity(kgCO2e/ft2)
TotalGHGEmissions	Équivalent à GHGEmissions(MetricTonsCO2e)
Location	Contient les variable : City, State ZipCode, Lattitude Longitude
2010 Census Tracts	Supprimée
Seattle Police Department Micro Community Policing Plan Areas	Supprimée
SPD Beats	Supprimée
City Council Districts	Supprimée

II. NETTOYAGE : CHOIX DES VARIABLES ET NETTOYAGE

- Sélections des variables utiles pour l'étude :
  - Pour caractériser le batiment : BuildingType, PrimaryPropertyType, YearBuilt, NumberofFloors, NumberofBuildings, PropertyGFATotal, PropertyGFAParking, PropertyGFABuilding(s), LargestPropertyUseType, LargestPropertyUseTypeGFA, LargestPropertyUseType, SecondLargestPropertyUseTypeGFA : ThirdLargestPropertyUseType, ThirdLargestPropertyUseTypeGFA
  - Pour caractériser la consommation électrique : Electricity(kWh), Electricity(kBtu), NaturalGas(kBtu), SteamUse(kBtu), OtherFuelUse(kBtu)
  - Autres : ENERGYSTARScore, Outlier, DataYear
- Vérifications du type des valeurs
- Cohérence entre les variables :
  - Entre les différents types d'énergie et la consommation totale : si différence > 1% ➔ supprimée
  - Entre la surface totale et la surface de l'immeuble + parking : si différence inférieure à 0 ➔ supprimée

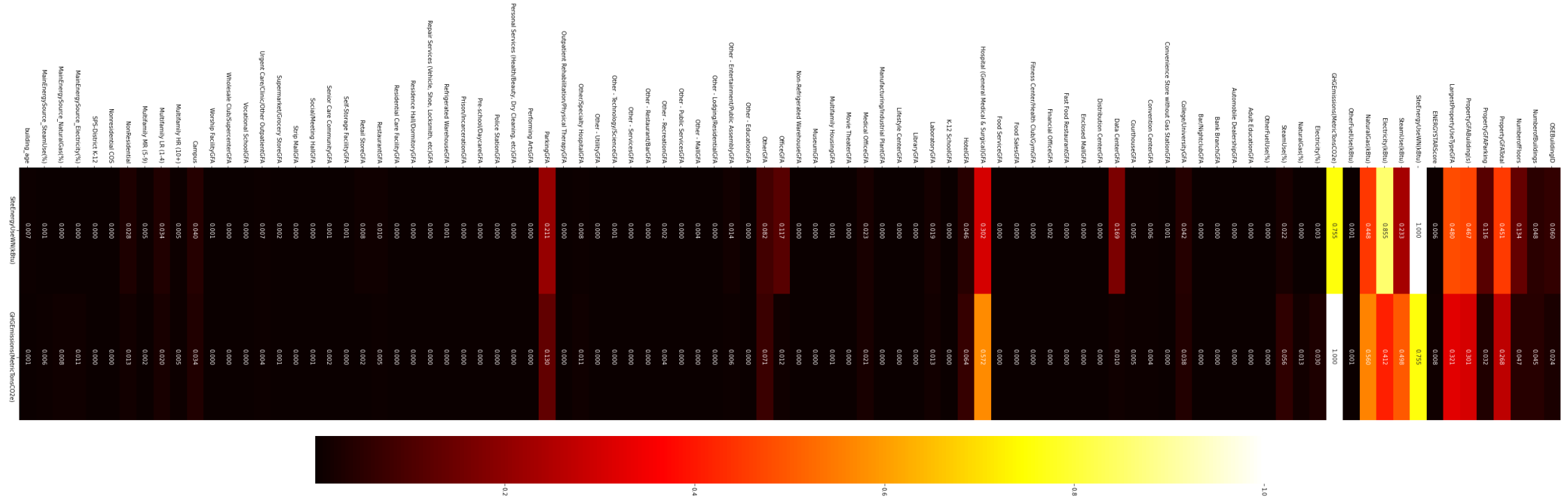
II. NETTOYAGE : FEATURES ENGINEERING

- Création de nouvelles variables à partir des données disponibles
  - Mix énergétique ➔ %électricité, %gaz, %vapeur, %autres
  - Source principale d'énergie ➔ OneHotEncoder sur le plus gros % du mix énergétique
  - Surface occupée par chaque activité ➔ regroupement des variables LargestPropertyUseType et LargestPropertyUseTypeGFA
  - Niveau de outliers ➔ 0 si non outlier, 1 faible outlier, 2 haut outlier
  - Type d'immeuble ➔ OneHotEncoder
  - Age de l'immeuble ➔ année du relevée – année de construction
- Imputation de l'ENERGY STAR Score



II. NETTOYAGE : EXPLORATION

- Taux de complétion : visualisation des variables les mieux renseignées
- Distribution : caractériser la distribution de chaque variable
- Corrélation : identification des variables les plus corrélées aux variables à prédire



III. TEST DE DIFFÉRENTS MODÈLES : CHOIX DES VARIABLES

- Séparation des variables en 4 niveaux :
  - Niveau 1:
    - YearBuild
    - MainEnergySource_Electricity(%)', 'MainEnergySource_NaturalGas(%)', 'MainEnergySource_SteamUse(%)'
    - 'NumberofBuildings', 'NumberofFloors', 'PropertyGFATotal','PropertyGFAParking', 'PropertyGFABuilding(s)'
  - Niveau 2 :
    - level 1
    - 'Campus', 'Multifamily HR (10+)', 'Multifamily LR (1-4)', 'Multifamily MR (5-9)', 'NonResidential', 'Nonresidential COS', 'SPS-District K-12'
    - 'Data CenterGFA' 'Hospital (General Medical & Surgical)GFA', 'OfficeGFA', 'College/UniversityGFA' 'HotelGFA'
  - Niveau 3 :
    - level 2
    - Electricity(%)', 'NaturalGas(%)', 'SteamUse(%)','OtherFuelUse(%)'
  - Niveau « all » : toutes les variables

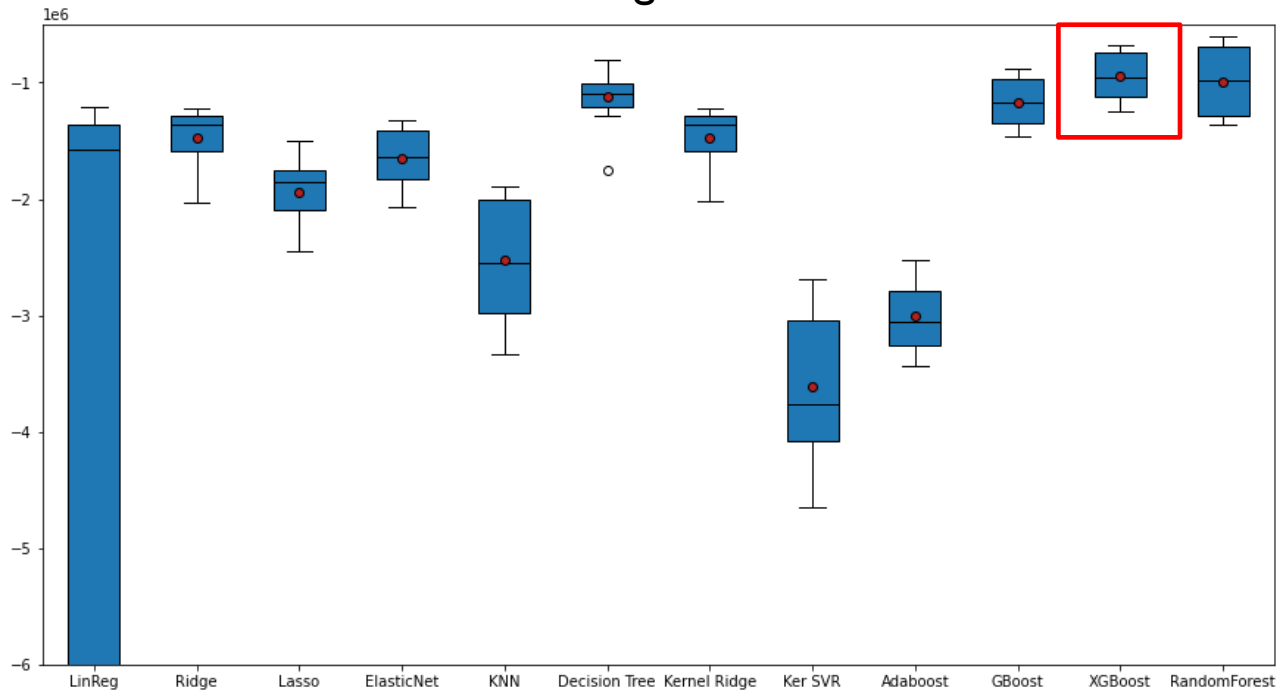
III. TEST DE DIFFÉRENTS MODÈLES : CHOIX DES MODELES

- modèles linéaire :
  - LinearRegression
  - Ridge Regression
  - Lasso Regression
  - ElasticNet Regression
  - Support vector machine Regression
- KNN
- arbres de décision
- modèles non linéaires :
- Kernel Ridge Regression
- Kernel Support Vector machine
- méthodes ensemblistes :
  - AdaBoost
  - GBoost
  - XGBoost
  - random forest

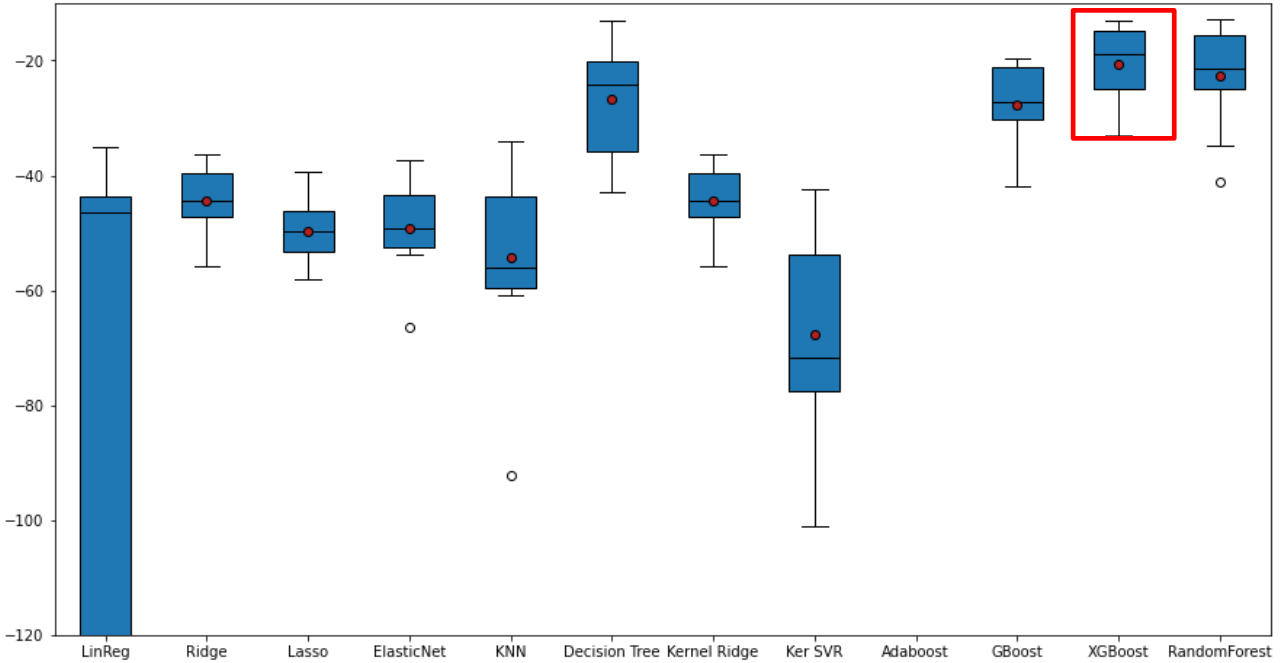
III. TEST DE DIFFÉRENTS MODÈLES : RESULTATS

- Paramètres :
  - 10 Kfolds, random_state : 42, scoring : neg_mean_absolute_error, différentes normalisations

Energie



GES

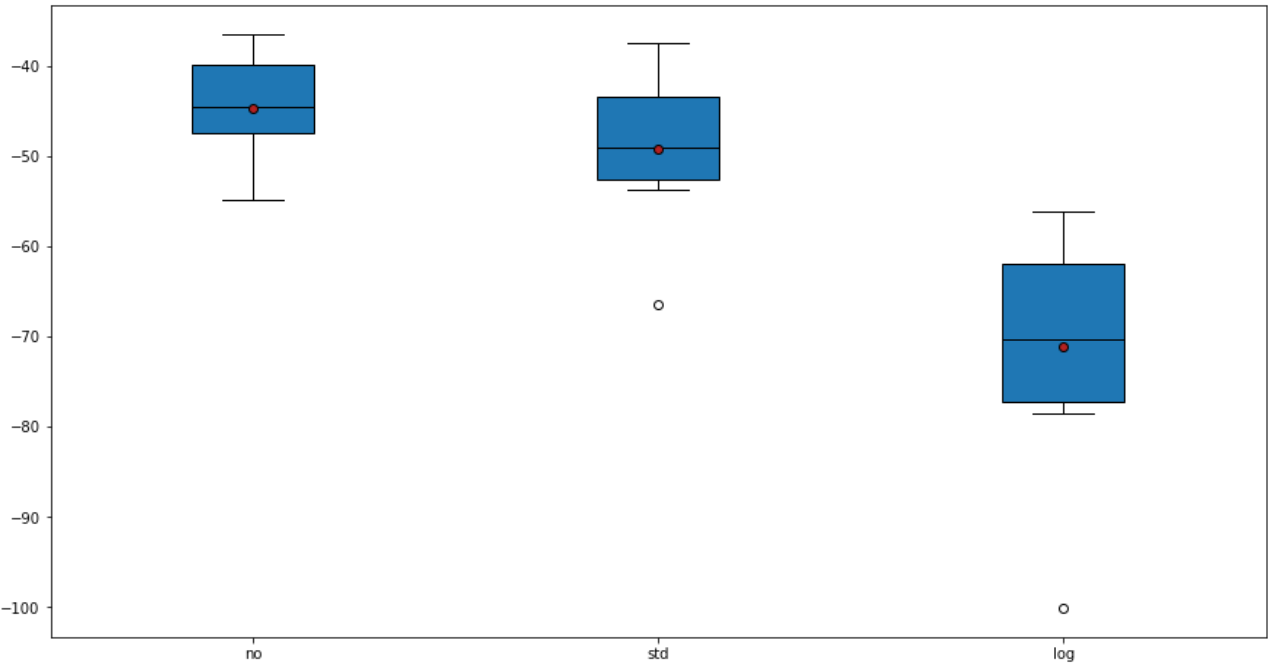


- XGBoost donne les meilleures performances

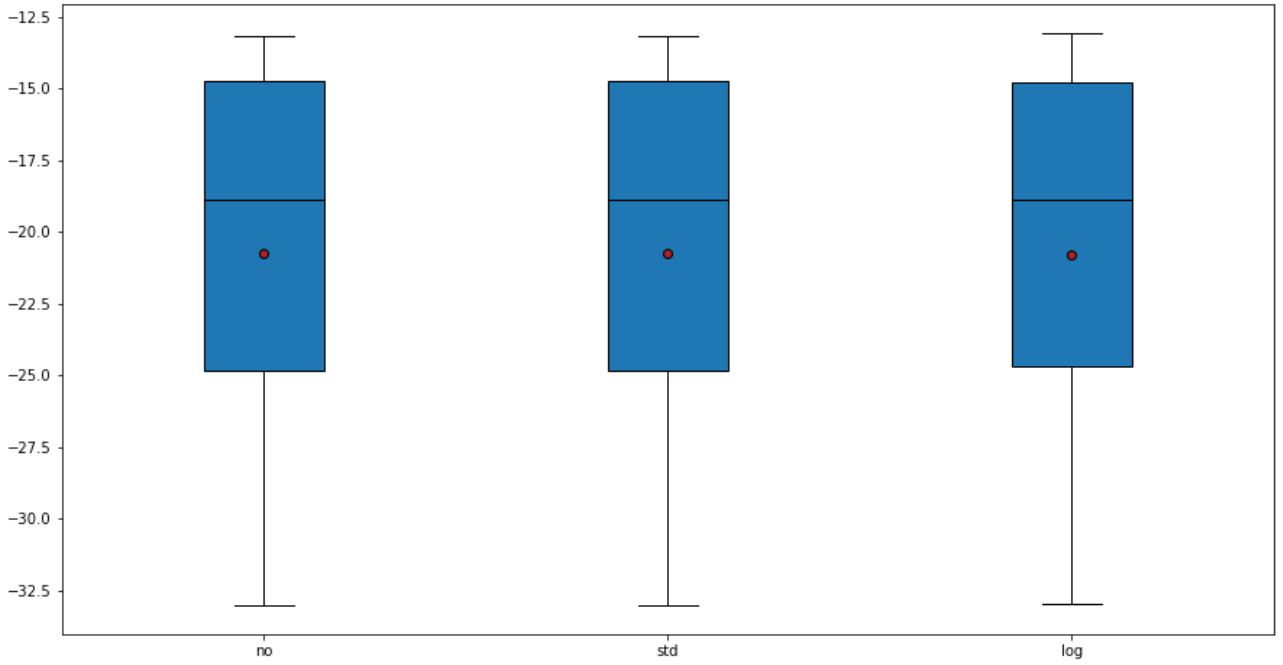
III. TEST DE DIFFÉRENTS MODÈLES : RESULTATS

- Effet de la normalisation :

ElasticNet



XGBoost



- Normalisation inutile pour le modèle XGBoost

IV. INTERPRÉTATIONS DU MODÈLE OPTIMAL

- Optimisation des hyperparamètres

	XGBoost Energie	XGBoost GES
N_estimators	400	300
Max_depth	16	21
Subsample	0.6	0.8
Colsample_bytree	0.8	0.8
Learning_rate	0.1	0.1

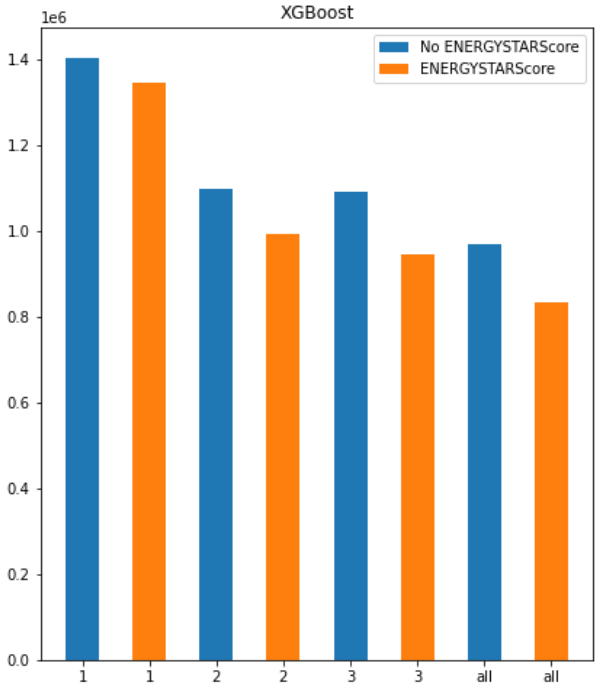
Modele	Score Energy (MAE) (level 2)	Score GES (MAE) (level 2)
Baseline	$4.577*10^6$	98.87
XGBoost non tunné	$1.31141*10^6$	35.831
XGBoost tunné	$1.0967*10^6$	31.49

- Le modèle optimisé conduit à de meilleurs performances
  - Gain important par rapport à la baseline

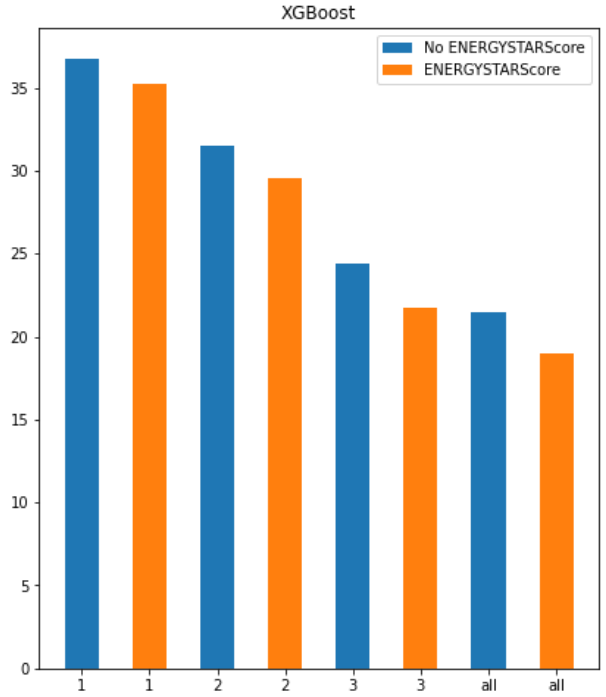
IV. INTERPRÉTATIONS DU MODÈLE OPTIMAL

- Effet de la variable Energy Star Score

Energie



GES



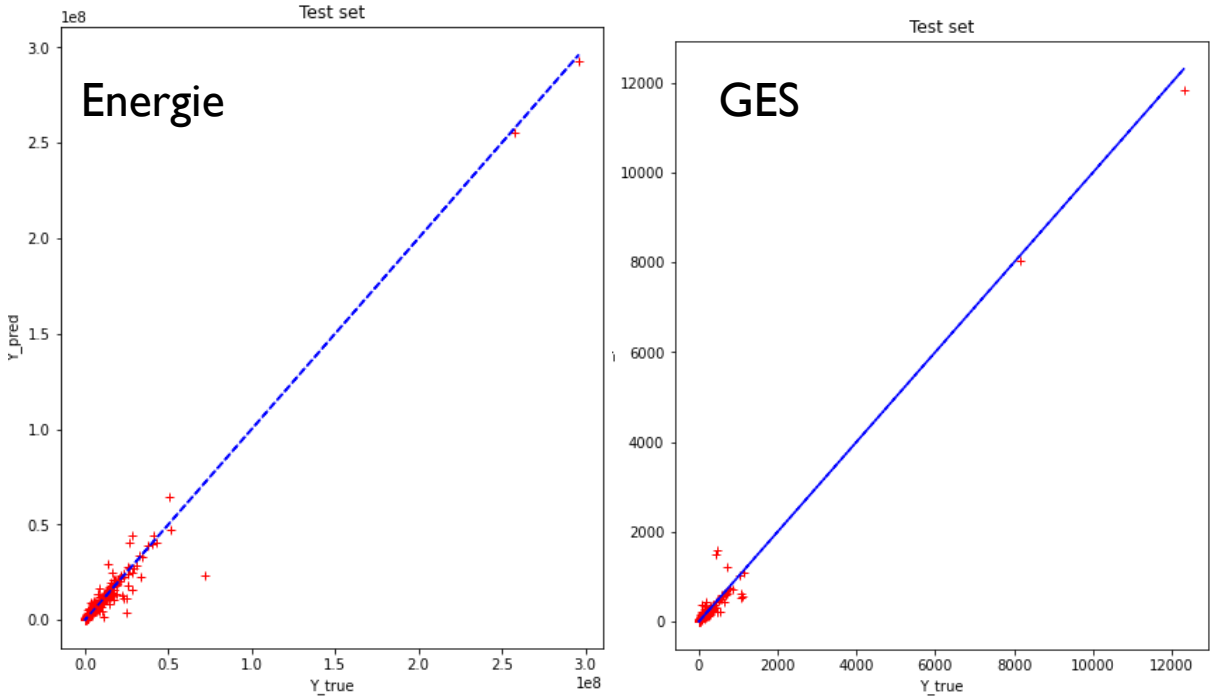
- Permet d'augmenter légèrement la performance du modèle.
- Le gain est similaire à celui obtenu en ajoutant plus de variables

  ➔ Pas essentiel

IV. INTERPRÉTATIONS DU MODÈLE OPTIMAL

- Performance sur le jeu de test

	XGBoost Energie	XGBoost GES
Baseline	$4.58*10^6$	98.87
Cross validation	$8.31*10^5$	19.00
Test set	$7.05*10^5$	17.47

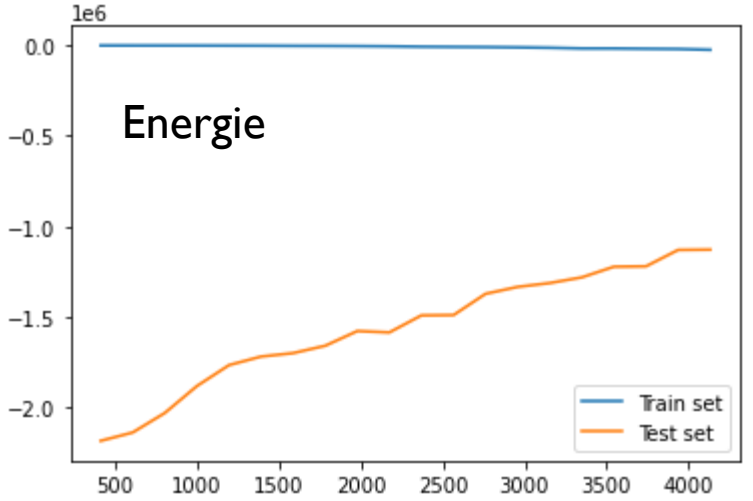


- L'erreur obtenue est du même ordre de grandeur que lors de la cross validation
- Un facteur 5 est obtenue par rapport à la baseline

IV. INTERPRÉTATIONS DU MODÈLE OPTIMAL

- Ajouter plus d'observations?

	XGBoost Energie	XGBoost GES
Petit Dataset	$8.31*10^5$	19.00
Gros Dataset	$1.21*10^6$	25.33



- Etrangement ajouter plus d'observations dégrade la performance…
  - Il faut peut être optimiser à nouveau les hyperparamètres
- La courbe d'apprentissage du jeu de test semble être encore croissante

➔ La performance peut être améliorée en augmentant la taille de la base de données

IV. INTERPRÉTATIONS DU MODÈLE OPTIMAL

- Effet de l'imputation

	XGBoost Energie	XGBoost GES
Sans ENERGYSTARScore	$1.21*10^6$	25.33
Avec ENERGYSTARScore	$1.30*10^6$	27.03

- Les performances obtenues avec la variable ENERGYSCORStar avec imputation dégrade les performances
  - L'imputation par la méthode des KNN n'est pas adaptée

V. CONCLUSIONS

- Base de données nettoyée : harmonisée, fusionnée, suppression des valeurs aberrantes…
- Création de nouvelles variables
- Plusieurs modèles testés
  - XGBoost permet d'obtenir les meilleurs performances
- Plusieurs traitements de données testés :
  - Pas d'effet sur le modèle XGBoost
- L'Energy Star Score permet d'augmenter la performance, mais pas de manière significative
- La prédiction de l'Energy Star Score par la méthode KNN est contre productive

V. AMÉLIORATIONS

- Augmenter la taille de la base de données afin d'obtenir des modèles plus performant
- Identifier et analyser les outliers sur le jeu d'apprentissage
- Aller plus loin de le feature engineering
  - Demander l'expertise des experts du métier