SOUTENANCE DU PROJET 4: ANTICIPATION DES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS

Présenté par Check KOUTAME

PLAN

I/ Mission & Description du projet

- Exploration des données de consommation/Emission de la ville de Seattle & explication des prédictions attendues
- Observations des données: Formes et qualités

II/ Nettoyages des données & analyse exploratoire

- Les différentes étapes de nettoyages
- Analyse exploratoire

III/ Modélisation et optimisation

- Mise en place de plusieurs modèles
- Optimisation de 4 modèles + la baseline
- Evaluation des performances de nos modèles et choix final

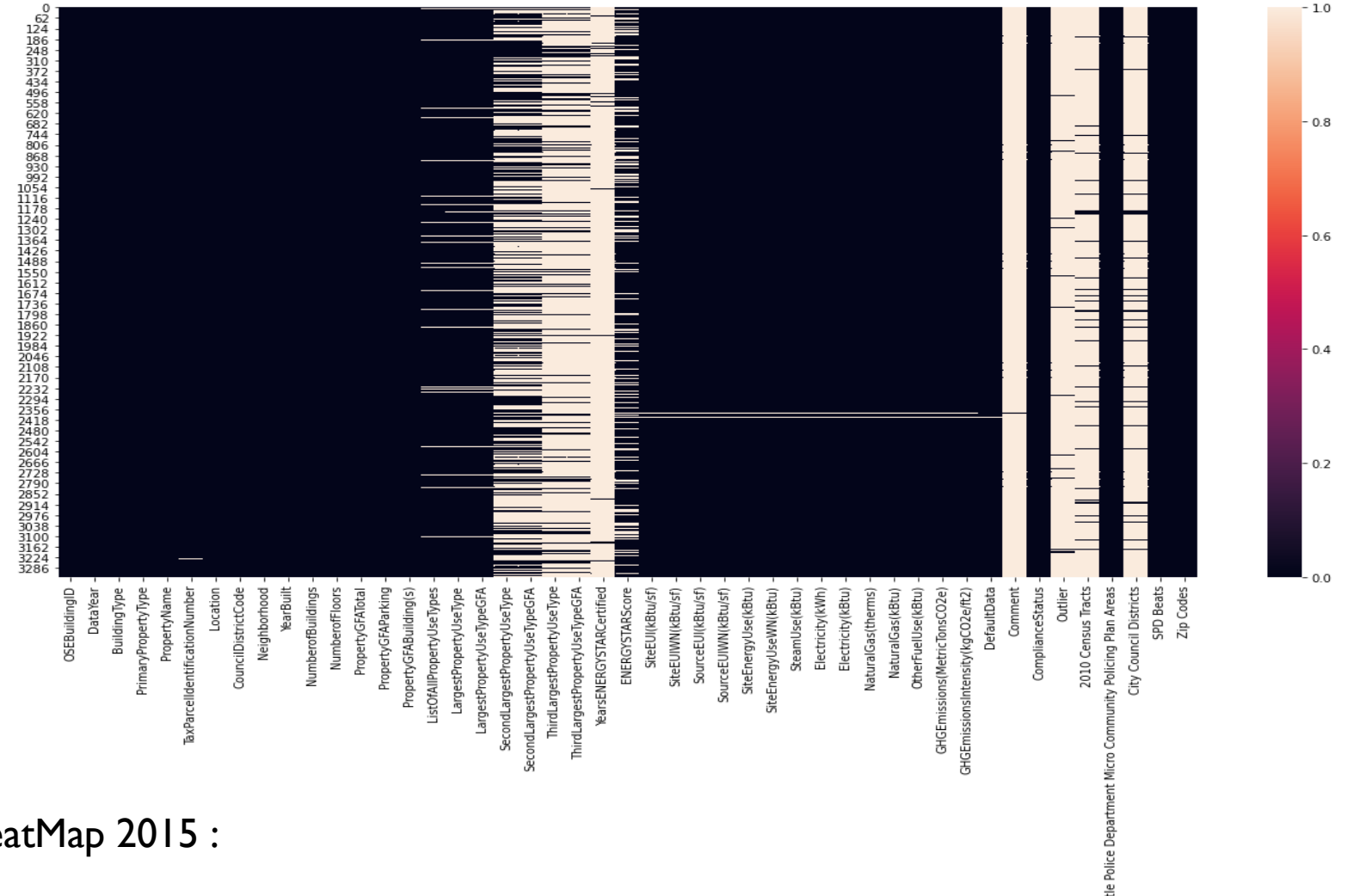
IV/ Intérêt d'utiliser l'Energystarscore

Conclusion

I. MISSION & DESCRIPTION DU PROJET:

- L'objectif de la ville de Seattle : Ville neutre en émissions de carbone en 2050
- **Problématique :**
  - Relevés manuels minutieux effectués en 2015 et 2016.
  - Ces relevés sont très couteux et il reste encore des bâtiments à mesurer.
  - Intérêt de l'indicateur *Energy Star Score* pour les prédictions de GES.
- **Mission :**
  - Prédictions des émissions de CO_2 et de consommation totale d'énergie à partir des données déjà existantes.
    - Les données existantes sont toutes les mesures qui ont été effectuées en 2015 et en 2016
  - Evaluer la performances de nos modèles en utilisant un certain nombres de métriques

I. MISSION & DESCRIPTION DU PROJET: NOS DONNÉES: OBSERVATIONS

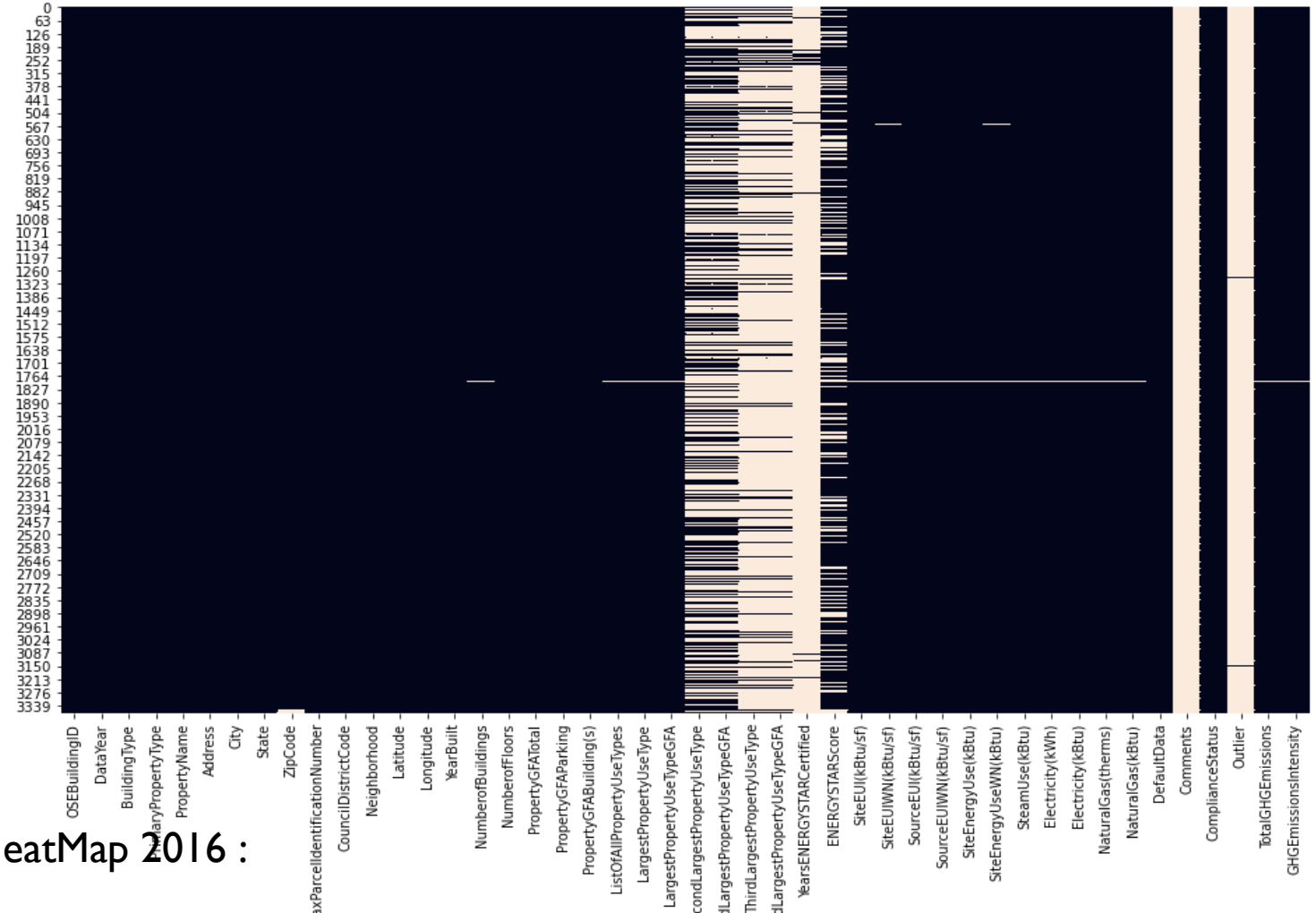


Informations sur la data 2015

```
Donnée : ['data_brute_2015']
Nombre de variable : 47
Nomres de types de variables : float64    23
object     15
int64       9
dtype: int64
Nombre observation : 3340
Nombre de cellules manquantes : 26512
% de cellules manquantes : 16.89%
Nombre de lignes dupliquées : 0
% de lignes dupliquées : 0.00%
```

HeatMap 2015 :

I. MISSION & DESCRIPTION DU PROJET: NOS DONNÉES:OBSERVATIONS



HeatMap 2016 :

Informations sur la data 2016

```
Donnée : ['data_brute_2016']
Nombre de variable : 46
Nomres de types de variables : float64    22
object     15
int64       8
bool        1
dtype: int64
Nombre observation : 3376
Nombre de cellules manquantes : 19952
% de cellules manquantes : 12.85%
Nombre de lignes dupliquées : 0
% de lignes dupliquées : 0.00%
```

I. MISSION & DESCRIPTION DU PROJET: NOS DONNÉES:

- 2 data séparés

Années	Observations	Variables
2015	3340	47
2016	3376	46

- 6 catégories de données

Catégories
Identifications
Infos liées aux données
Localisation
Usages et construction (variables qualitatives)
Usages et constructions (variables quantitatives)
Relevés énergétiques et calcul des émissions

OSEBuildingID
PropertyName
TaxParcelIdentificationNumber

DataYear
DefaultData
Comments
ComplianceStatus
Outlier

CouncilDistrictCode
Neighborhood
ZipCode
Latitude
Longitude
Address

BuildingType
PrimaryPropertyType
YearBuilt
ListOfAllPropertyUseTypes
LargestPropertyUseType
SecondLargestPropertyUseType
ThirdLargestPropertyUseType

SiteEUI(kBtu/sf)
SiteEUIWN(kBtu/sf)
SourceEUI(kBtu/sf)
SourceEUIWN(kBtu/sf)
SiteEnergyUse(kBtu)
SiteEnergyUseWN(kBtu)
SteamUse(kBtu)
Electricity(kWh)
Electricity(kBtu)
NaturalGas(therms)
NaturalGas(kBtu)
OtherFuelUse(kBtu)
TotalGHGEmissions
GHGEmissionsIntensity

YearsENERGYSTARCertified
ENERGYSTARScore

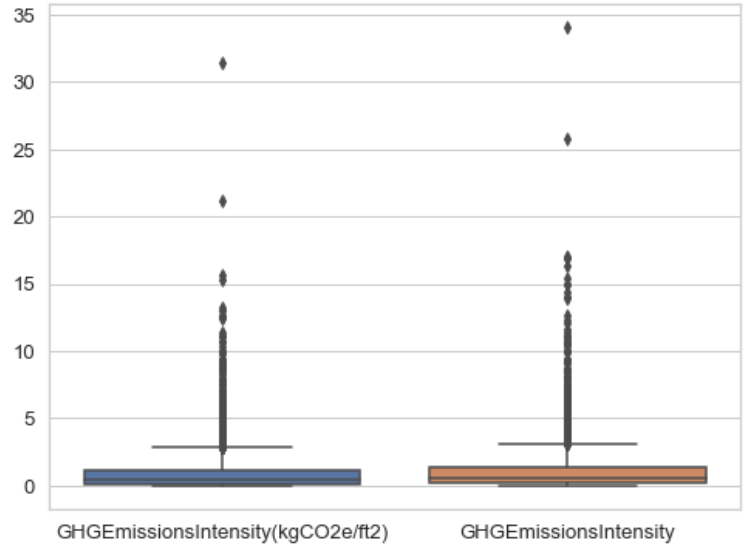
NumberofBuildings
NumberofFloors
PropertyGFATotal
PropertyGFAParking
PropertyGFABuilding(s)
LargestPropertyUseTypeGFA
SecondLargestPropertyUseTypeGFA
ThirdLargestPropertyUseTypeGFA

I. MISSION & DESCRIPTION DU PROJET: NOS DONNÉES: ETAPES AVANT CONCATÉNATION

- Sur les deux data set: plusieurs variables non communes
  - Chercher les variables différentes
    - Supprimer celles qui ne sont que dans une dataSet

•Regrouper les variables qui diffèrent par leur noms

- GHGEmissionsIntensity(kgCO2e/ft2) et GHGEmissionsIntensity
- GHGEmissions(MetricTonsCO2e) et TotalGHGEmissions



- Sur la variable Location de 2015 par exemple
  - Decomposer le disctionaire qu'il constitue pour en faire d'autres varaibles (Longitude, Latitude, ZipCode, et…)

II. NETTOYAGE DES DONNÉES:

- Suppressions des données dupliquées
- Suppressions des variables sans intérêts pour nos prédiction
- Traitement des valeurs NaN
- Traitements des bâtiments non résidentiels
- Suppression des variables corrélées
- Imputations des valeurs manquantes sur ENERGYSTARScore
- Traitement des valeurs catégorielles

II. NETTOYAGE DES DONNÉES: A/VALEURS DUPLIQUÉES

- Suppressions des valeurs dupliquées en se basant sur le l'indicateur OSEBuilding
  - 6716 rows × 46 columns → 3432 rows × 46 columns
  - Il s'agit de supprimer toutes les observations qui sont faites sur les mêmes bâtiments.

II. NETTOYAGE DES DONNÉES: A/SUPPRESSIONS DES VARIABLES SANS INTÉRÊT

Rappel des variables à prédire: SiteEnergyUseWN(kBtu), TotalGHGEmissions, et importance de la variable energieStarScore dans cette prédiction...

Non exploitable telles quelles

DefaultData, Comments, ComplianceStatus, Comments, YearsENERGYSTARCertified, OSEBuildingID, DataYear, PropertyName, TaxParcelIdentificationNumber, YearBuilt, ListOfAllPropertyUseTypes, Latitude, Longitude, Address,

Trop directement liées aux variables cibles

ENERGYSTARScore, *Certified*, SiteEUI(kBtu/sf), SiteEUIWN(kBtu/sf), SourceEUI(kBtu/sf), SourceEUIWN(kBtu/sf), SiteEnergyUse(kBtu), SiteEnergyUseWN(kBtu), SteamUse(kBtu), Electricity(kWh), Electricity(kBtu), NaturalGas(therms), NaturalGas(kBtu), OtherFuelUse(kBtu), TotalGHGEmissions, GHGEmissionsIntensity

Quantitatives

- **Usages des bâtiments**

LargestPropertyUseTypeGFA, SecondLargestPropertyUseTypeGFA, ThirdLargestPropertyUseTypeGFA,

- **Surfaces et état du bâtiment**

BuildingAge, NumberofBuildings, NumberofFloors, PropertyGFATotal, PropertyGFAParking, PropertyGFABuilding(s), *ExtsurfVolRatio*, *MeanGFAperFloor*, *ParkingGFARatio*,

Catégorielles

- **Liées au profil énergétique**

CertifiedPreviousYear, *NbYearsCertified*, *EnergyProfile*, *MainEnergy*, Outlier

- **Liées aux usages des bâtiments**

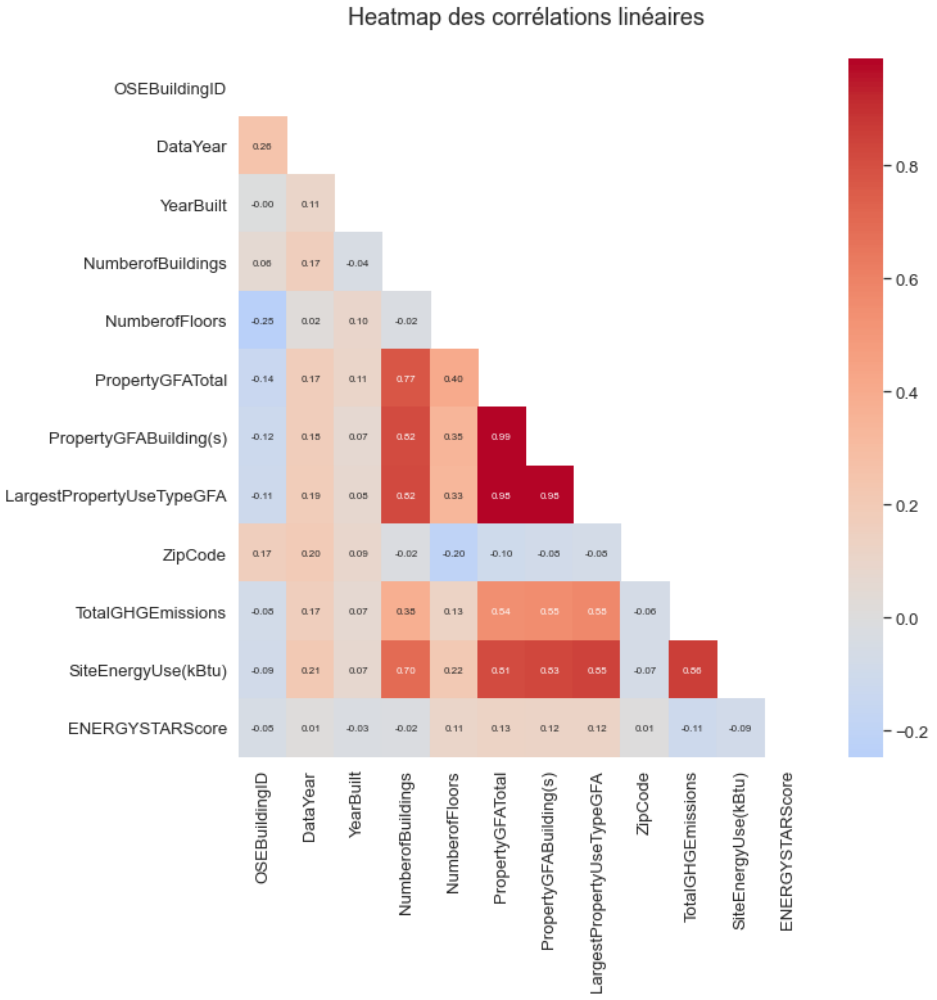
BuildingType, PrimaryPropertyType, LargestPropertyUseType, SecondLargestPropertyUseType, ThirdLargestPropertyUseType,

- **Liées à l'emplacement des bâtiments**

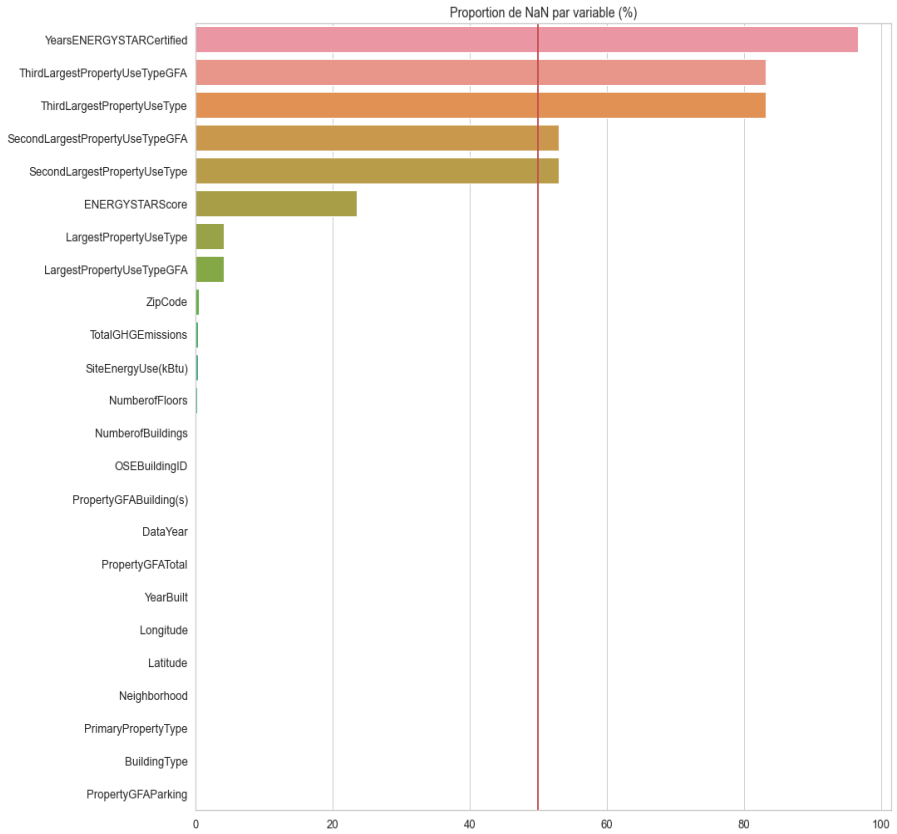
CouncilDistrictCode, Neighborhood, ZipCode

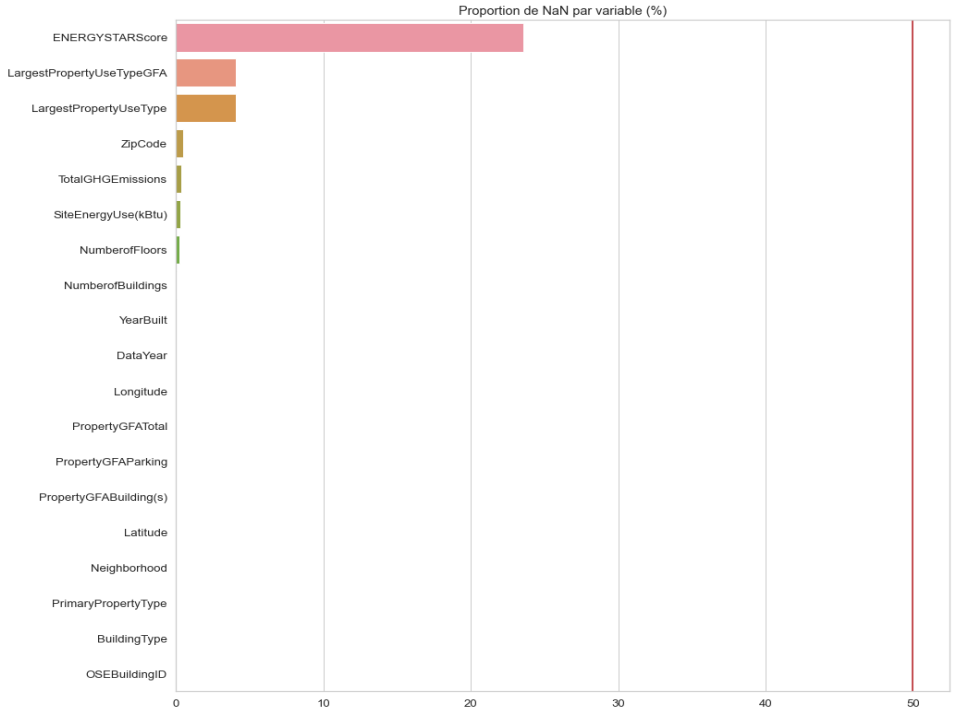
II. NETTOYAGE DES DONNÉES: A/SUPPRESSIONS DES VARIABLES SUIVANTES

- **Suppression des valeurs bâtiments non résidentiels**
  - Toutes les valeurs correspondant à des habitations sont supprimées dans la variable Building Type
- **Suppression des variables corrélées**
  - Ex: PropertyGFATotal, PropertyGFABuilding(s), LargestPropertyUseTypeGFA sont très corrélés donc seulement la variable PropertyGFATotalva être retenu pour notre modèle.



II. NETTOYAGE DES DONNÉES: A/TRAITEMENT DES NAN

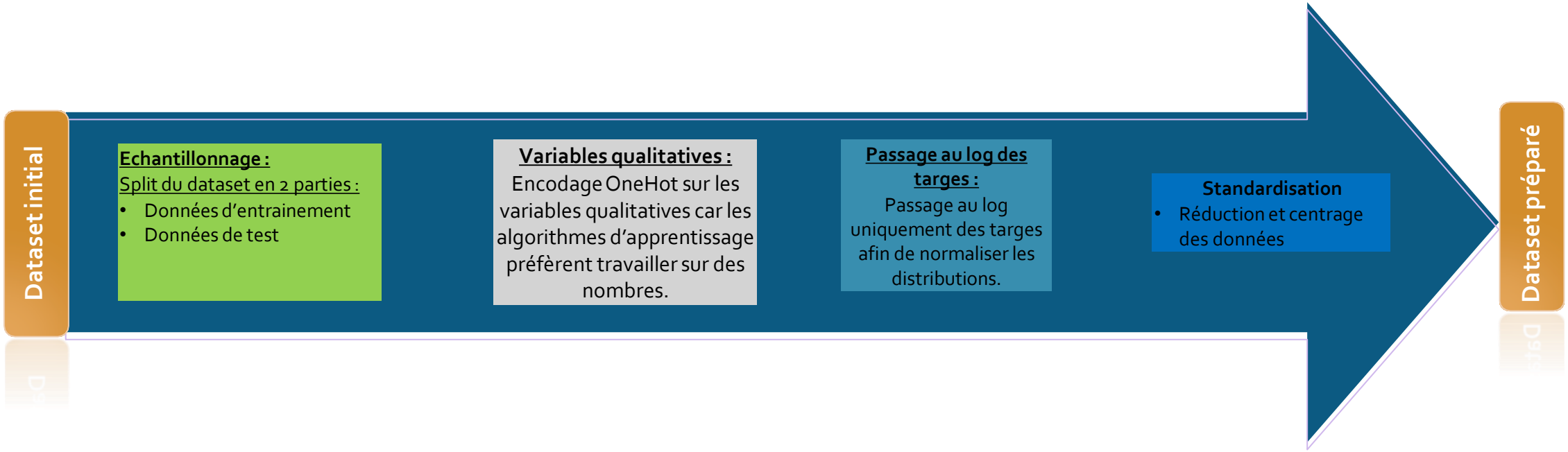




- Suppressions des lignes et variables vides
- Suppressions des variables ayant plus de 50% de NaN

- Certaines de ces variables présentes un taux très faible de données manquantes < 2.5% (exemple : **LargestPropertyUseType** et ***ZipCode.*** Les quelques observations manquantes vont être supprimés pour ces variables.
- La variable ENERGYSTARScore (34%): imputées avec l'algorithme KNN.

III. MODÉLISATION
A/ PREPROCESSING



III. MODÉLISATION
A/ PLAN SUIVI

Utilisation à l'aveugle de plusieurs modèles de régression

Identification de quelques modèles potentiellement prometteur.

Choix de quelques modèles en plus de la Baseline

Optimisation de ces modèles (Elastic Net – SVM – Random Forest – Gradient boosting)
- Hyperparamètres
- Sélection de variables

Evaluation des modèles (entraînement)
- Courbe d'apprentissage
- Scores en cross-validation
- Influence des hyperparamètres
- Importance des variables

Evaluation finale (test)
- Scores
- Distribution des erreurs
- Temps de prédiction

Importance de l'ENERGYSTAR Score

III. MODÉLISATION
A/ MÉTHODES D'ÉVALUATION

b/Résultats de test

- **MAE** : intuitif, importance proportionnelle à la valeur des erreurs
- **RMSE** : donne une évaluation proportionnelle des erreurs, mais peut être perturbé par les erreurs sur les plus petites valeurs.
- R^2 : évalue la proportion de variance expliquée par le modèle

➢ Toutes les modélisations sont faites sur nos deux targes: TotalGHGEmission et SiteEnergieUse

III. MODÉLISATION
A/ UTILISATION À L'AVEUGLE DE PLUSIEURS MODÈLES DE RÉGRESSION

b/Modélisation

Liste des modèles testés :

1 Linear Regression
2 Regression Ridge
3 Regresion Lasso
4 ElasticNet
5 kNN
6 SVR
7 DecisionTree Regressor

Modèles simples

8 Random Forest Regressor
9 AdaBoostRegressor
10 Gradient Boosting Regressor

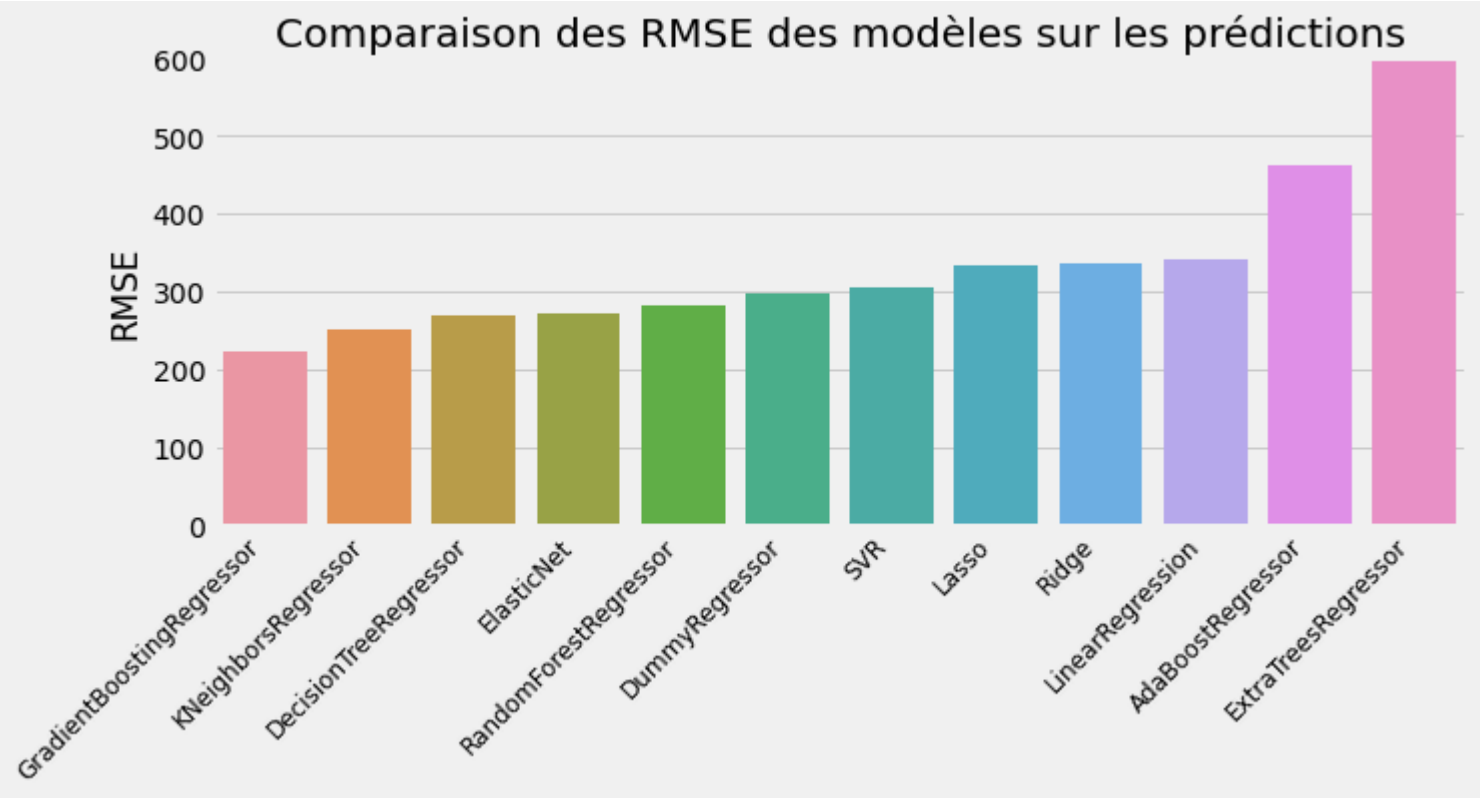
Méthodes ensemblistes

Idée:

- L'idée dans cette étape de notre étude est de voir avec une seule Target, quels sont les modèles potentiellement prometteurs qu'on peut optimiser par la suite. Pour cela, nous allons:
  - Entrainement et faire des prédiction en utilisant plusieurs modèles non optimisés
  - On utilisera, le métriques RMSE pour la validation croisée sur la target **TotalGHGEmissions**

III. MODÉLISATION
A/ UTILISATION À L'AVEUGLE DE PLUSIEURS MODÈLES DE RÉGRESSION

Modèles potentiellement prometteurs



	Modeles	RMSE
10	GradientBoostingRegressor	222.156
6	KNeighborsRegressor	249.962
2	DecisionTreeRegressor	269.740
9	ElasticNet	271.477
0	RandomForestRegressor	281.661
8	DummyRegressor	297.128
7	SVR	305.655
5	Lasso	333.414
11	Ridge	335.812
1	LinearRegression	340.683
3	AdaBoostRegressor	462.709
4	ExtraTreesRegressor	705.929

Première conclusions: Au vu de ces résultats, on peut voir que les modèles assemblistes non hyperparametrés ne permettent pas d'avoir de meilleurs résultats par rapport aux modèles linéaires. Cependant, ce graphique peut nous servir de support afin de suivre l'évolution de la performance de nos modèles paramétrés.

III. MODÉLISATION
B/ OPTIMISATION DE QUELQUES MODÈLES

a/Modèles potentiellement prometteurs

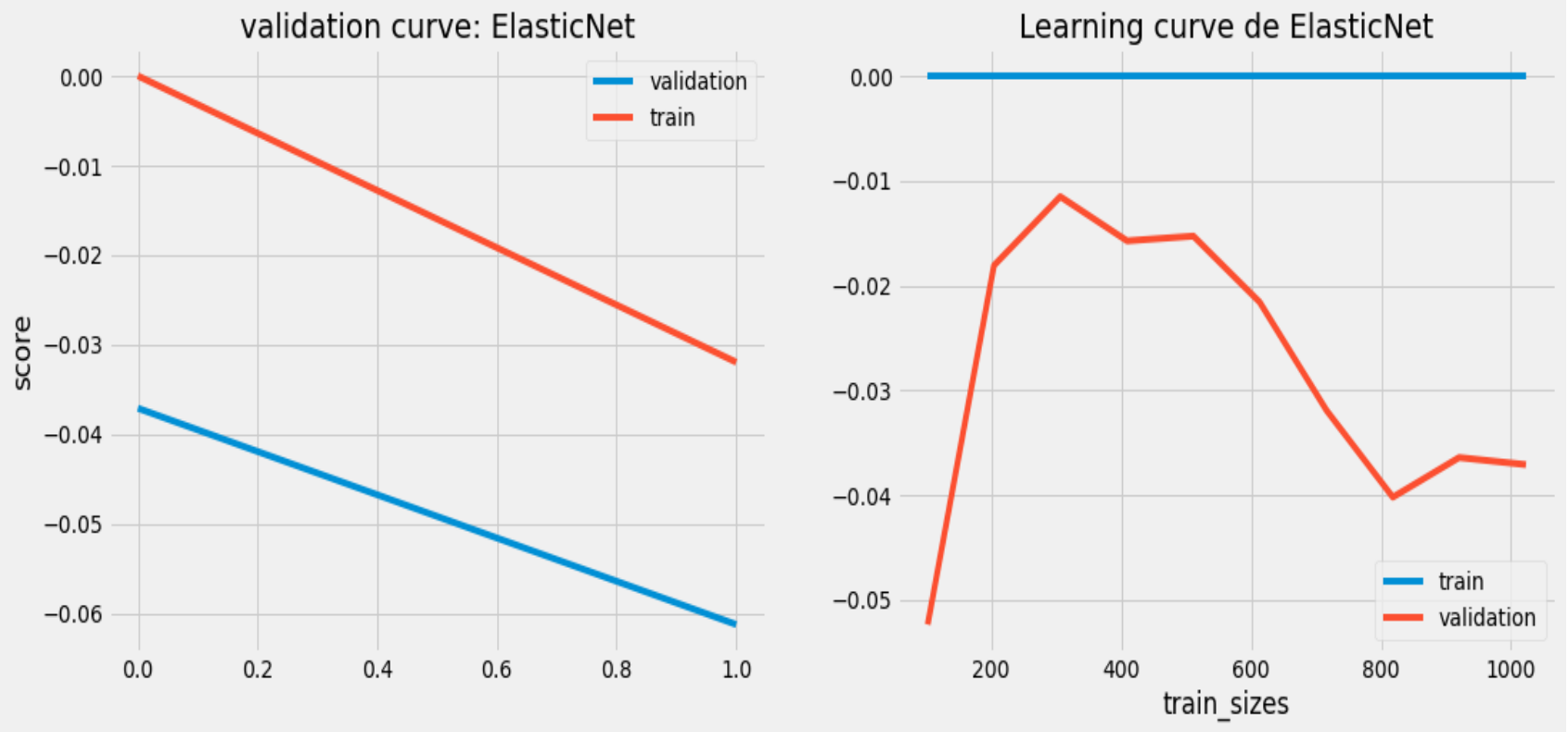
- **Pour l'entrainement et l'optimisation de nos modèles, nous choisirons donc les modèles suivants:**
  - Baseline: Dummy
  - Régressions linéaires: Lasso, Ridge, Elactic Net, SVM
  - Assemblistes: RandomForest et gradient boost

III. MODÉLISATION
B/ OPTIMISATION DE QUELQUES MODÈLES

a/Entrainement de nos modèles, Target TotalGHGEmission

Optimisation par recherche par quadrillage en validation croisée (5 passes

- **Dummy**

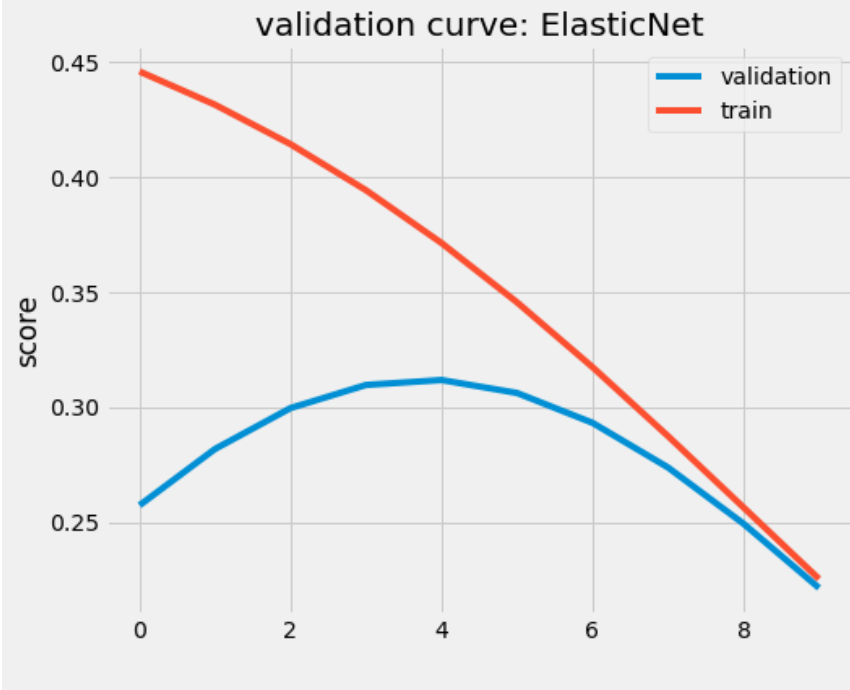


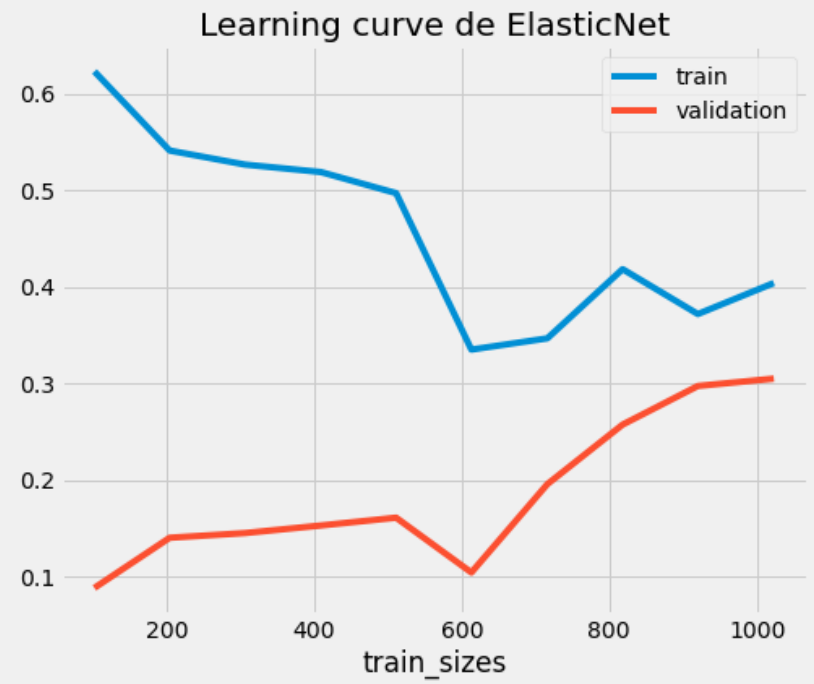
III. MODÉLISATION
B/ OPTIMISATION DE QUELQUES MODÈLES

a/Entrainement de nos modèles, Target TotalGHGEmission

Optimisation par recherche par quadrillage en validation croisée (5 passes

- **Le ElasticNet:**
  - alpha : alpha, coef qui multiplie le terme de pénalité
  - L1: =1 équivaut à un Lasso, 0 à un Ridge



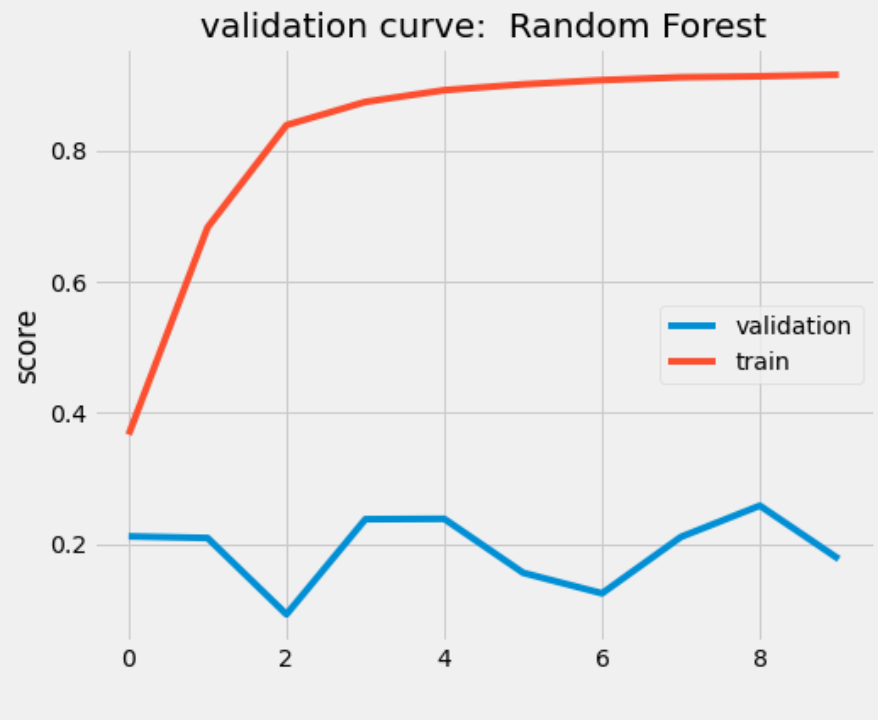


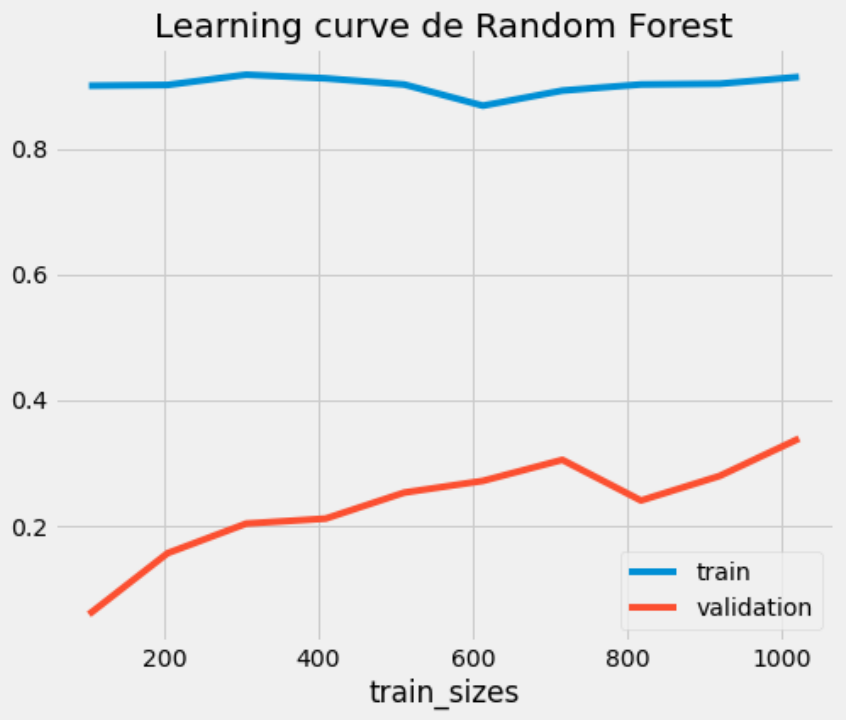
III. MODÉLISATION
B/ OPTIMISATION DE QUELQUES MODÈLES

a/Entrainement de nos modèles, Target TotalGHGEmission

Optimisation par recherche par quadrillage en validation croisée (5 passes

- **La forêt aléatoire :**
  - Nombre d'arbres qui composent la forêt
  - Nombre de variables à considérer
  - Profondeur de l'arbre
  - min_samples_split : Le nombre minimum d'échantillons requis pour scinder un nœud interne



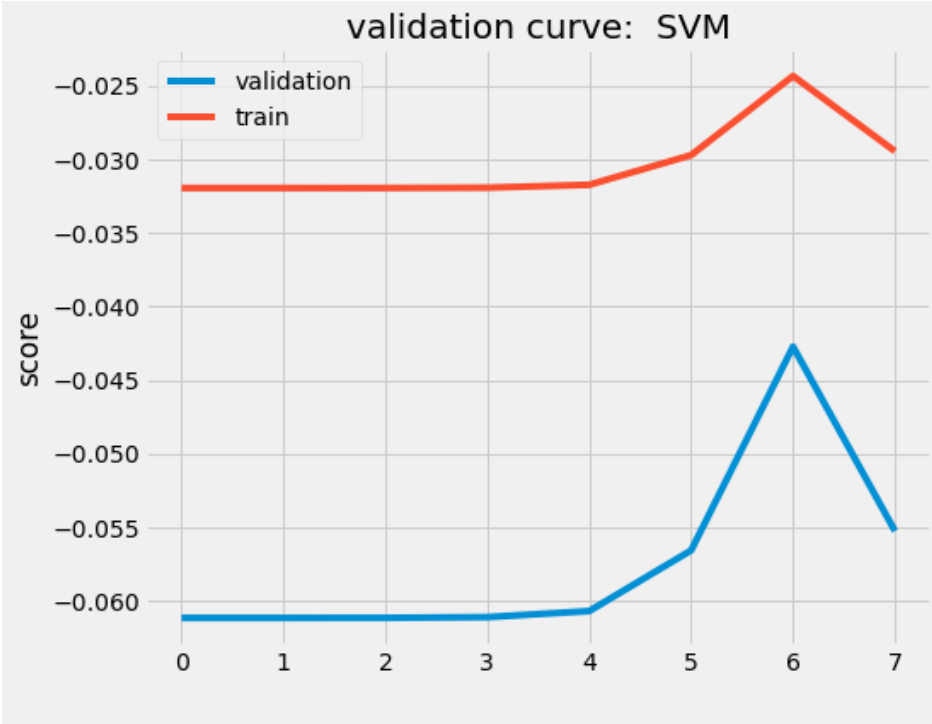


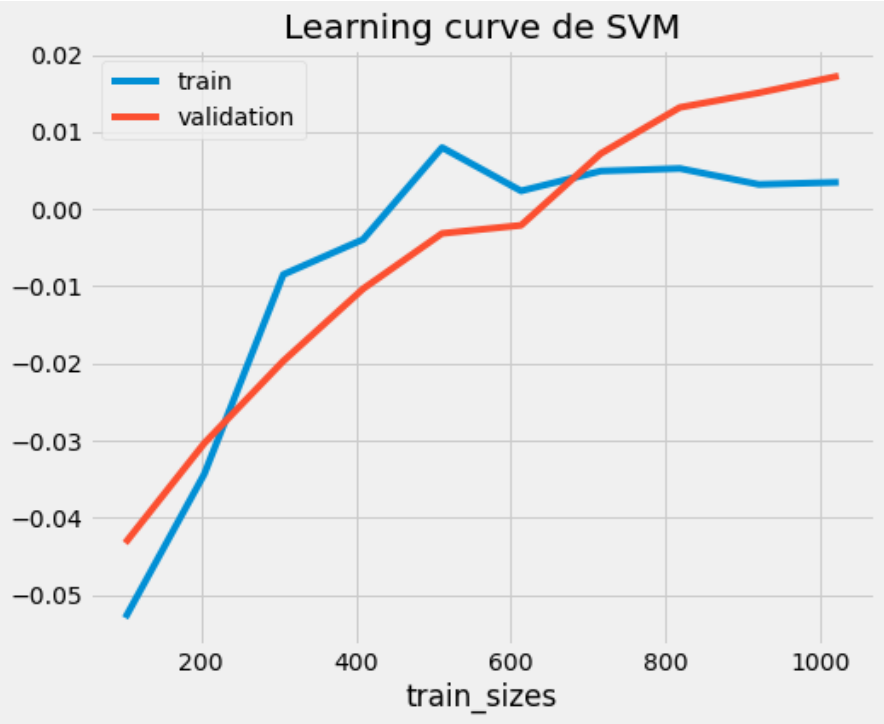
III. MODÉLISATION
B/ OPTIMISATION DE QUELQUES MODÈLES

a/Entrainement de nos modèles, Target TotalGHGEmission

Optimisation par recherche par quadrillage en validation croisée (5 passes

- **Le SVM:**
  - gamma : #kernel coefficient [ici kernel = Radial Basis Functionepsilon
  - Epsilon: rate erreur tolérée par l'algorithme
  - C :paramètre de régularisation



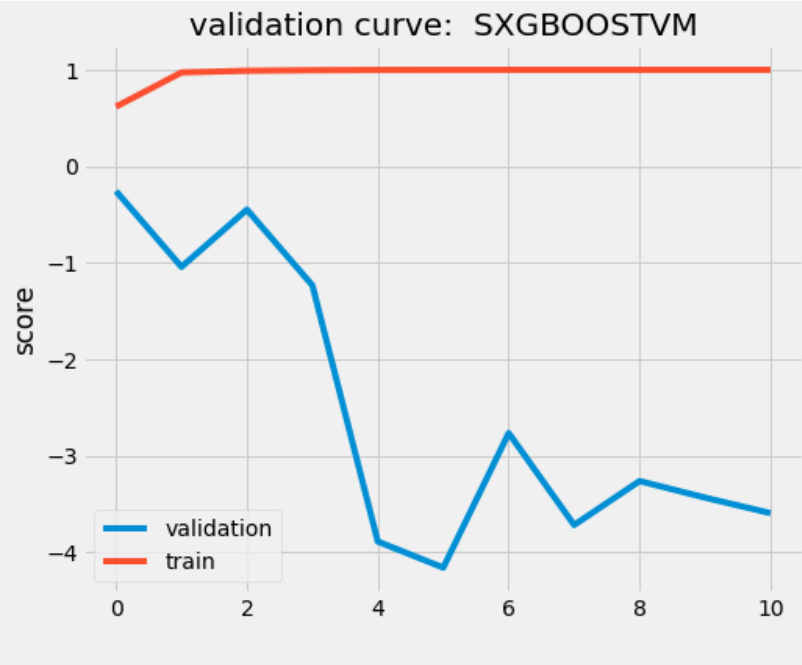


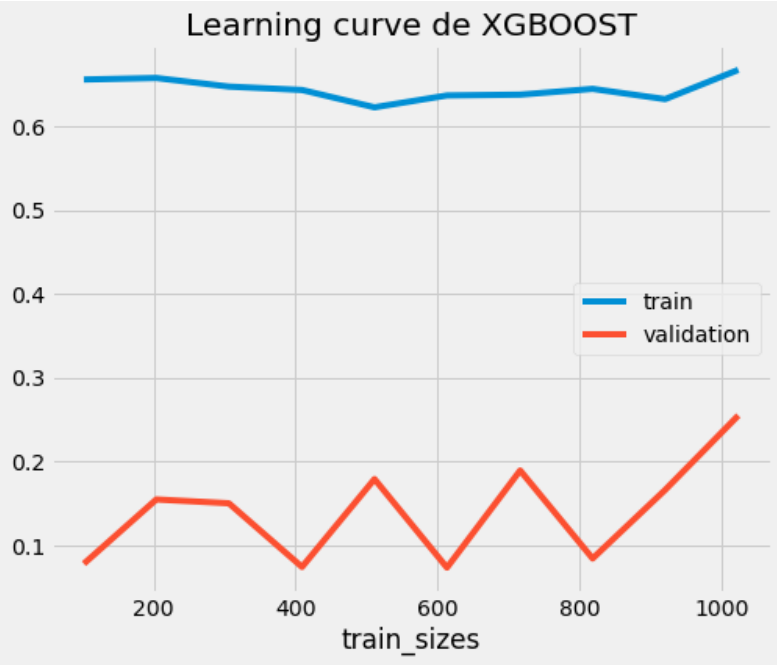
III. MODÉLISATION
B/ OPTIMISATION DE QUELQUES MODÈLES

a/Entrainement de nos modèles, Target TotalGHGEmission

Optimisation par recherche par quadrillage en validation croisée (5 passes)

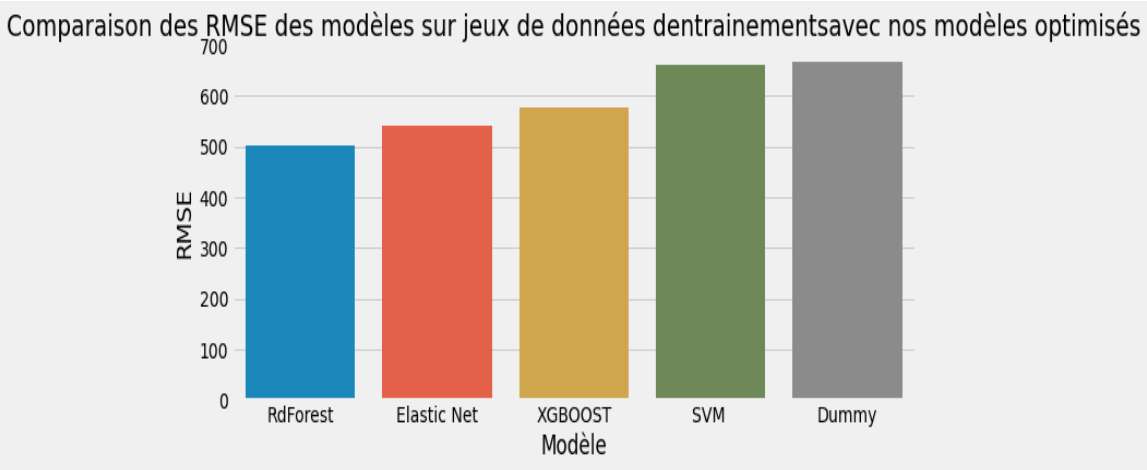
- **Le Gradient Boosting Regressor :**
  - n_estimators : Nbr d'étape de boosting à effectuer
  - max_depth : profondeur maximale des estimateurs de régression individuels
  - min_samples_split : Le nombre minimum d'échantillons requis pour scinder un nœud interne
  - learning_rate : régulation de la contribution de chaque arbre
  - loss : fonction de perte

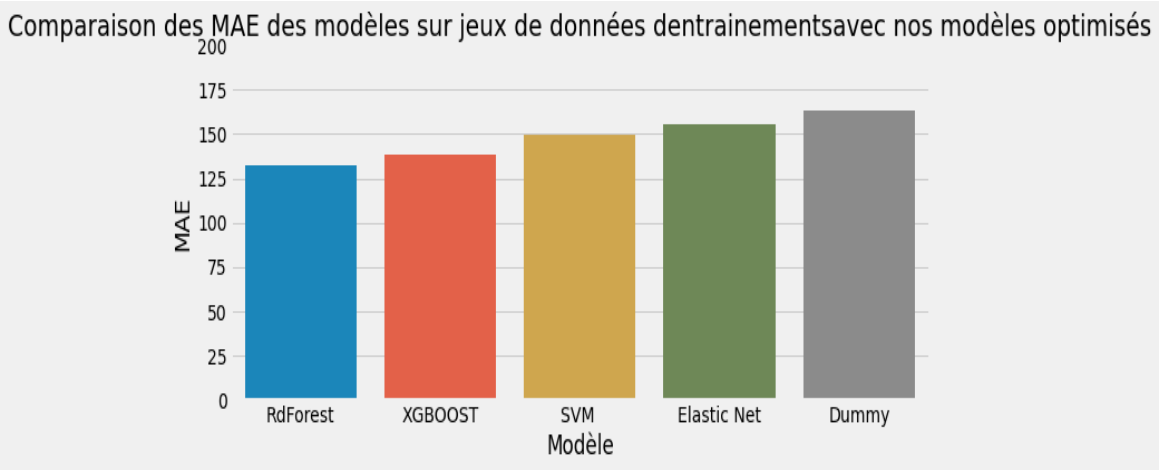


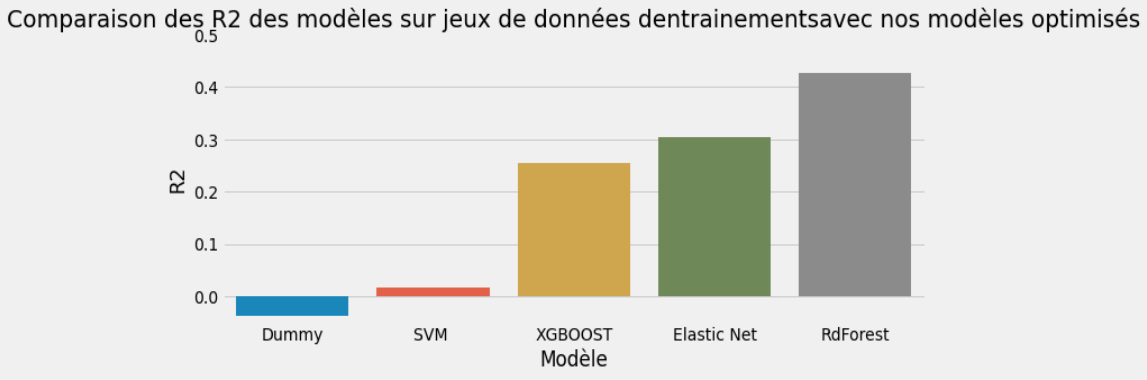


III. MODÉLISATION
B/ OPTIMISATION DE QUELQUES MODÈLES

b/Evaluation des modèles sur les jeux d'entrainement, Target TotalGHGEmission

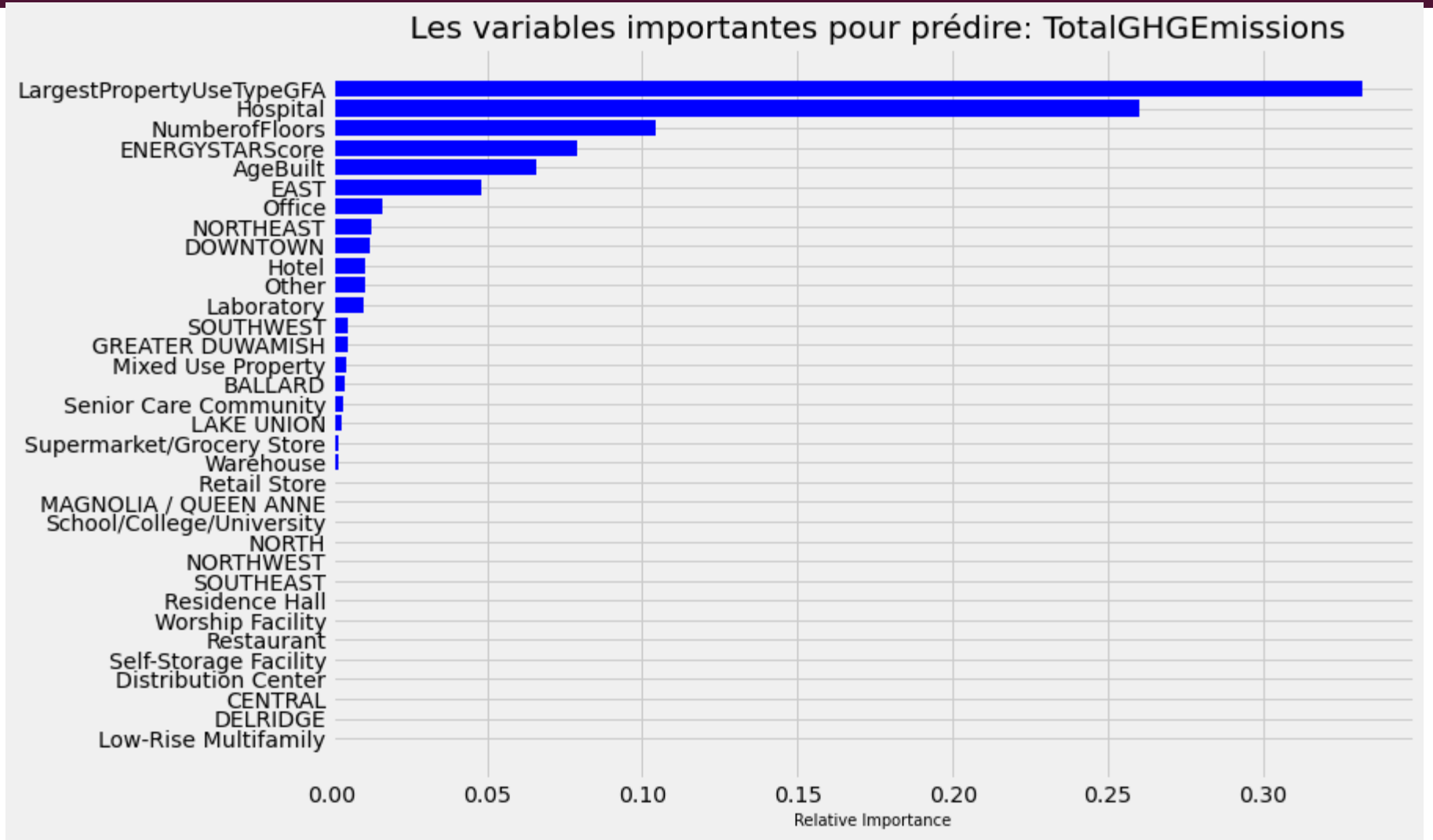






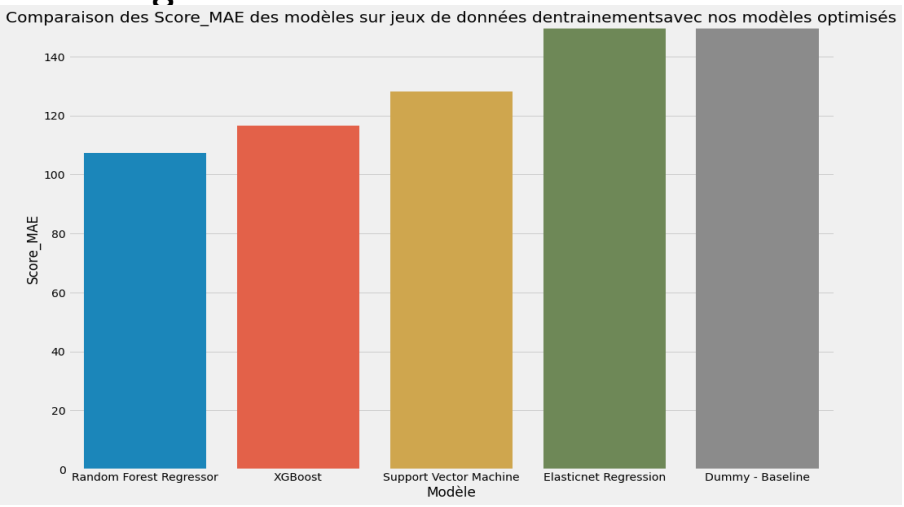
	Modèle	Best Param	RMSE	R2	MAE
0	Dummy	{'strategy': 'mean'}	2.119695e+07	-0.004658	6.868329e+06
1	Elastic Net	{'alpha': 1, 'l1_ratio': 0.7000000000000001, '...	1.407553e+07	0.581688	4.807309e+06
2	RdForest	{'max_features': 'auto', 'min_samples_leaf': 1...	1.245056e+07	0.682193	3.709011e+06
3	SVM	{'C': 10, 'epsilon': 0.001, 'gamma': 0.01}	2.191065e+07	-0.083715	6.868253e+06
4	XGBOOST	{'learning_rate': 0.1, 'max_depth': 4, 'n_esti...	1.390253e+07	0.571854	3.766517e+06

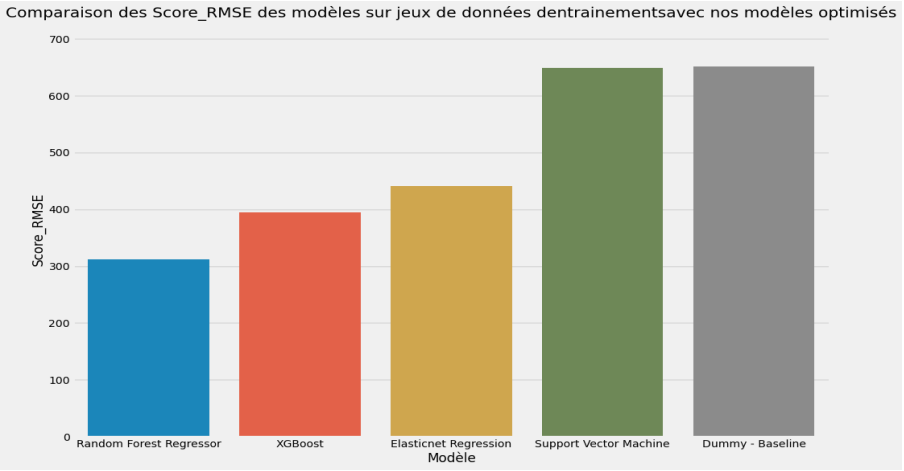
III. MODÉLISATION
C/ ETAPE 3: FEATURES IMPORTANTES

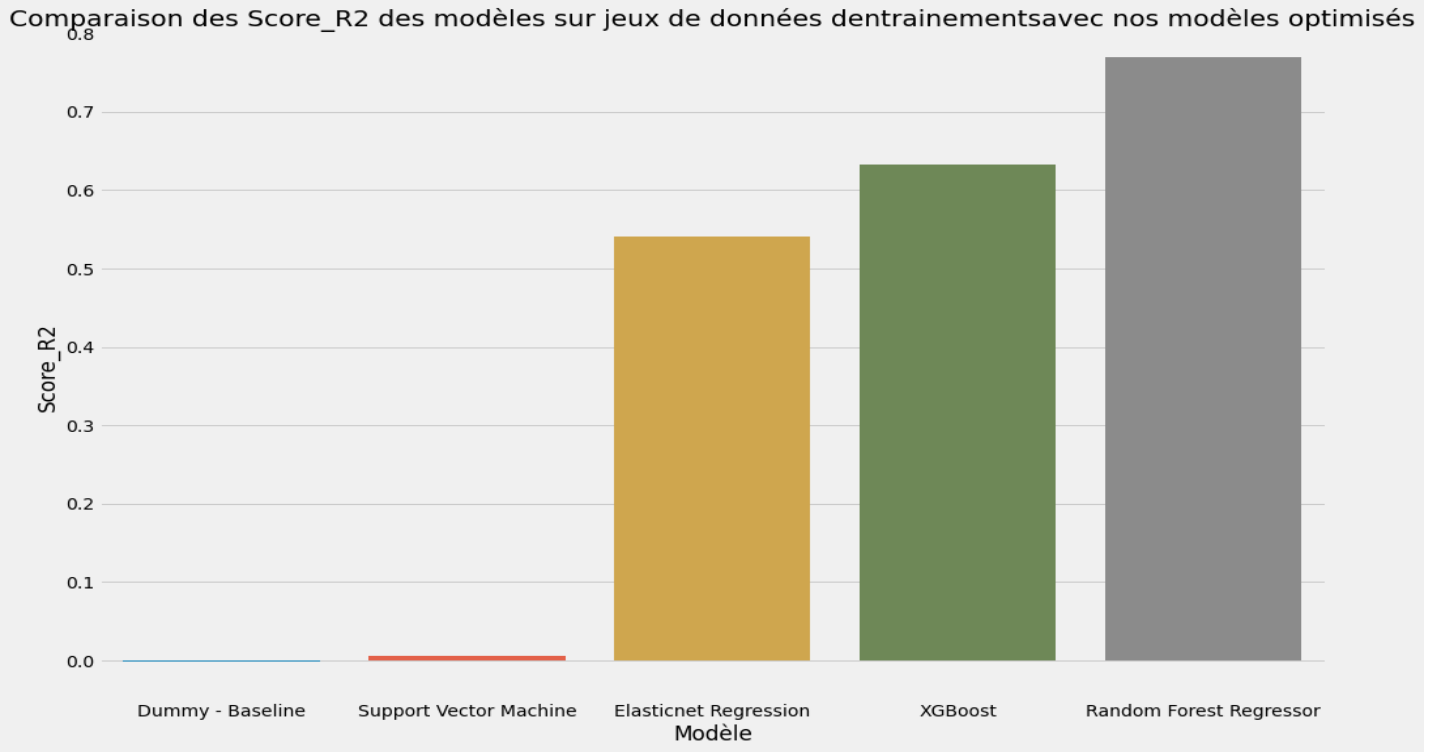


III. MODÉLISATION
D/ **PRÉDICTION SUR LES DONNÉES DE TEST ET ÉVALUATION**

Target: TotalGHGEmission



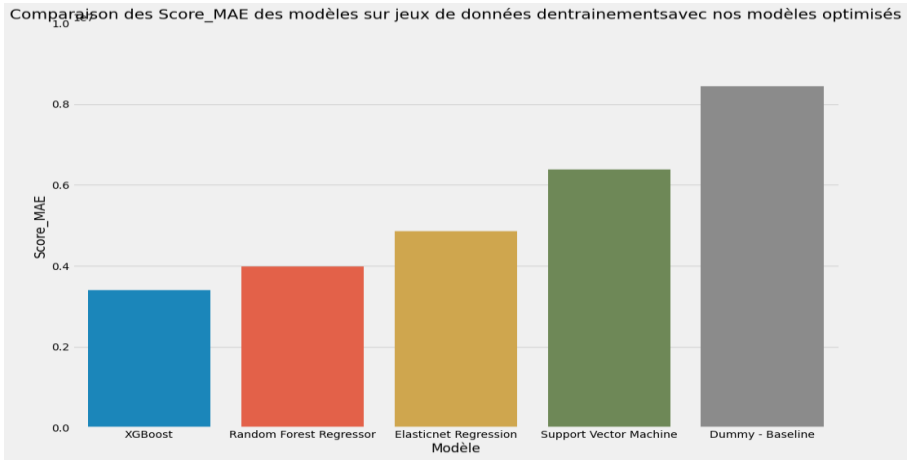


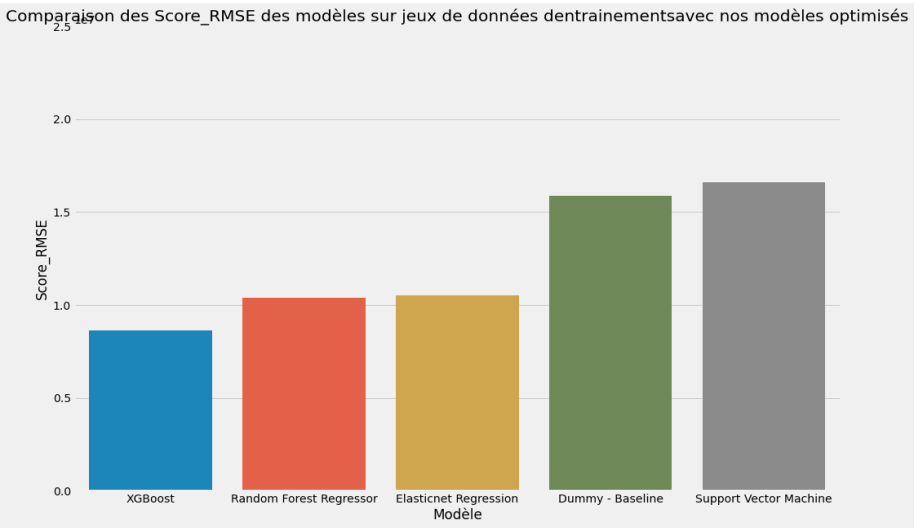


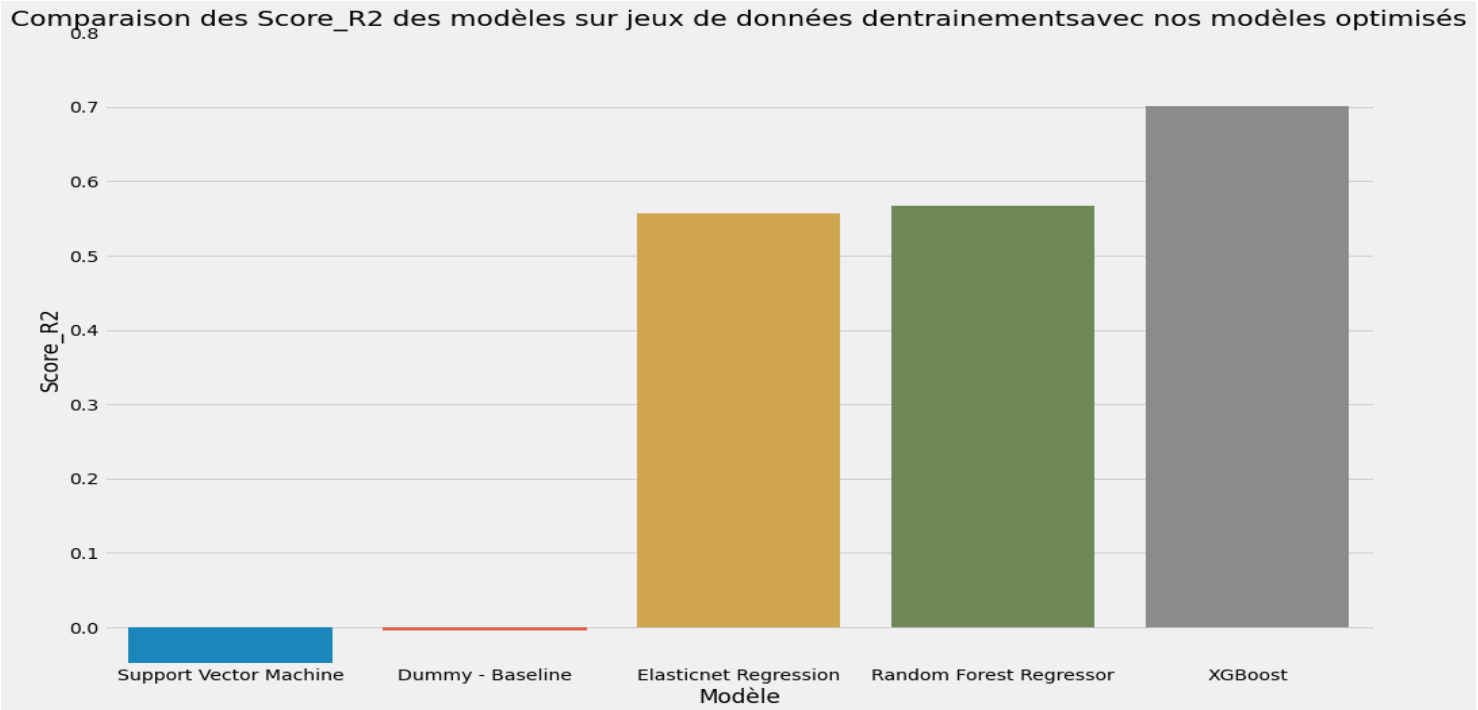
	Modèle	Best Param	Score_RMSE	Score_R2	Score_MAE
0	Dummy - Baseline	{'strategy': 'mean'}	651.116025	-0.001223	196.793647
1	Elasticnet Regression	{'alpha': 1, 'l1_ratio': 0.9, 'tol': 0.0001}	440.650750	0.541433	153.617349
2	Random Forest Regressor	{'max_features': 'auto', 'min_samples_leaf': 1...	312.604333	0.769217	107.263866
3	Support Vector Machine	{'C': 10, 'epsilon': 1, 'gamma': 0.01}	648.706828	0.006172	128.170537
4	XGBoost	{'learning_rate': 0.1, 'max_depth': 7, 'n_esti...	394.173801	0.633064	116.702478

III. MODÉLISATION
D/ **PRÉDICTION SUR LES DONNÉES DE TEST ET ÉVALUATION**

Target: EnergySiteUse

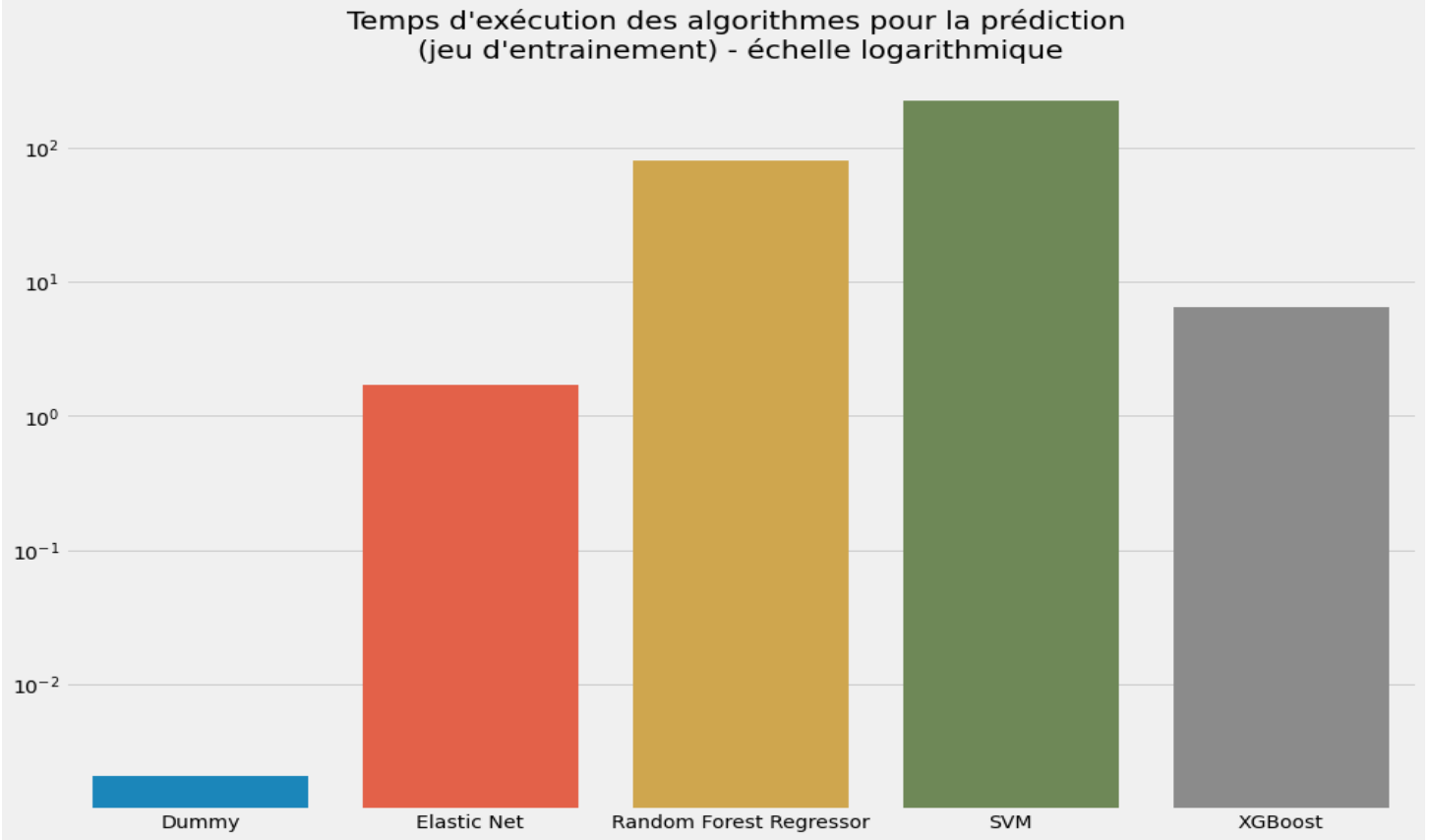






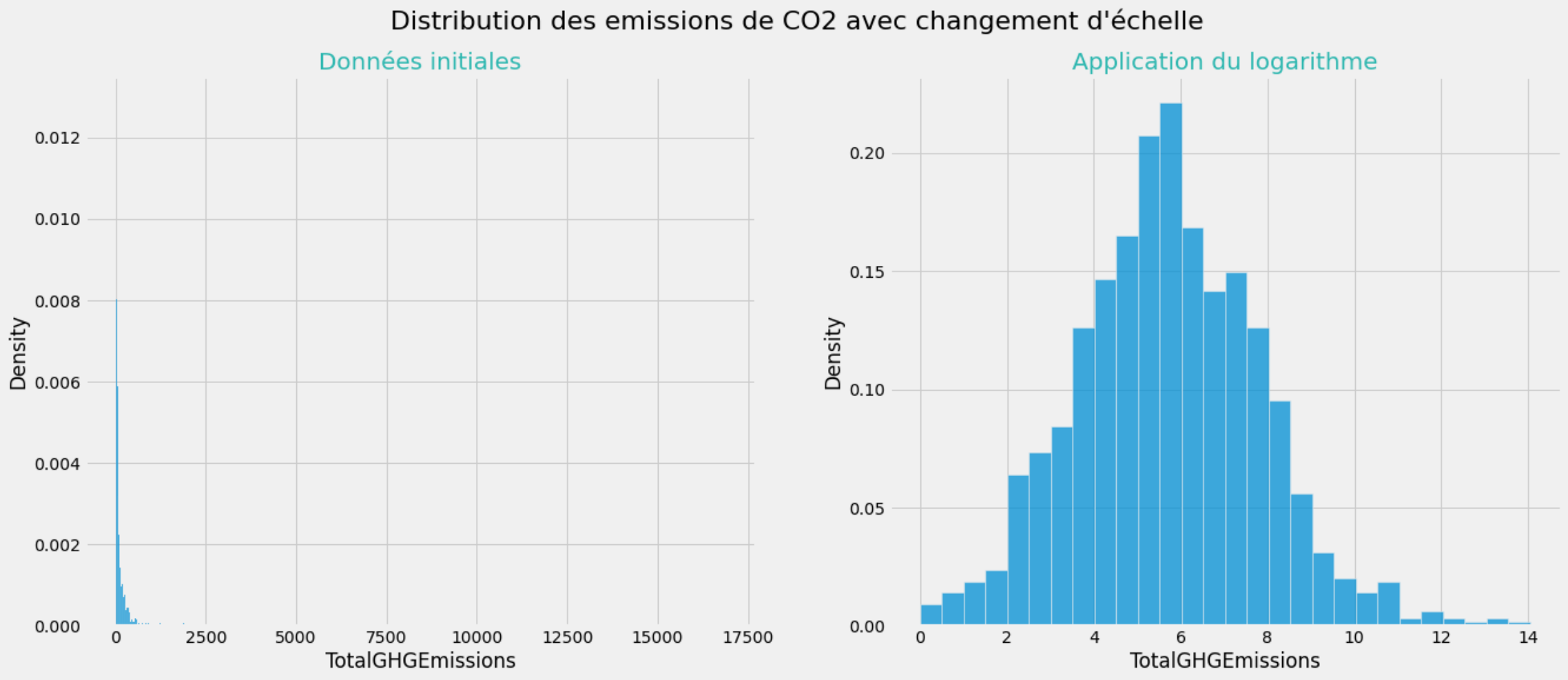
	Modèle	Best Param	Score_RMSE	Score_R2	Score_MAE
0	Dummy - Baseline	{'strategy': 'mean'}	1.585495e+07	-0.005105	8.443624e+06
1	Elasticnet Regression	{'alpha': 0.1, 'l1_ratio': 0.4, 'tol': 0.0001}	1.053596e+07	0.556156	4.857446e+06
2	Random Forest Regressor	{'max_features': 'sqrt', 'min_samples_leaf': 1...	1.040778e+07	0.566890	3.981983e+06
3	Support Vector Machine	{'C': 10, 'epsilon': 0.001, 'gamma': 0.01}	1.661759e+07	-0.104124	6.388047e+06
4	XGBoost	{'learning_rate': 0.01, 'max_depth': 8, 'n_est...	8.643736e+06	0.701266	3.391709e+06

III. MODÉLISATION
D/ **PRÉDICTION SUR LES DONNÉES DE TEST ET ÉVALUATION**

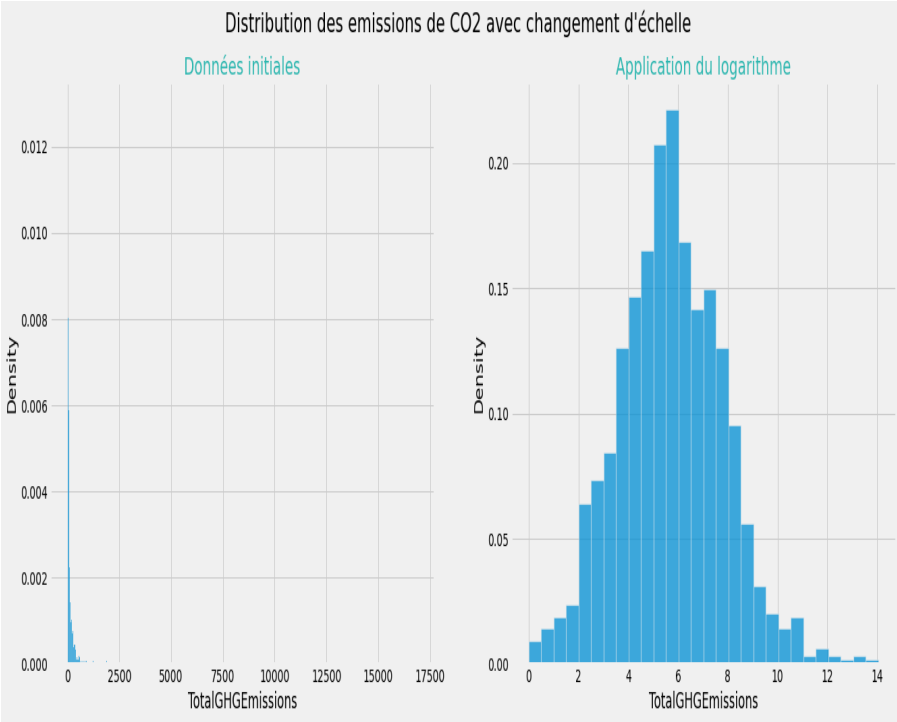


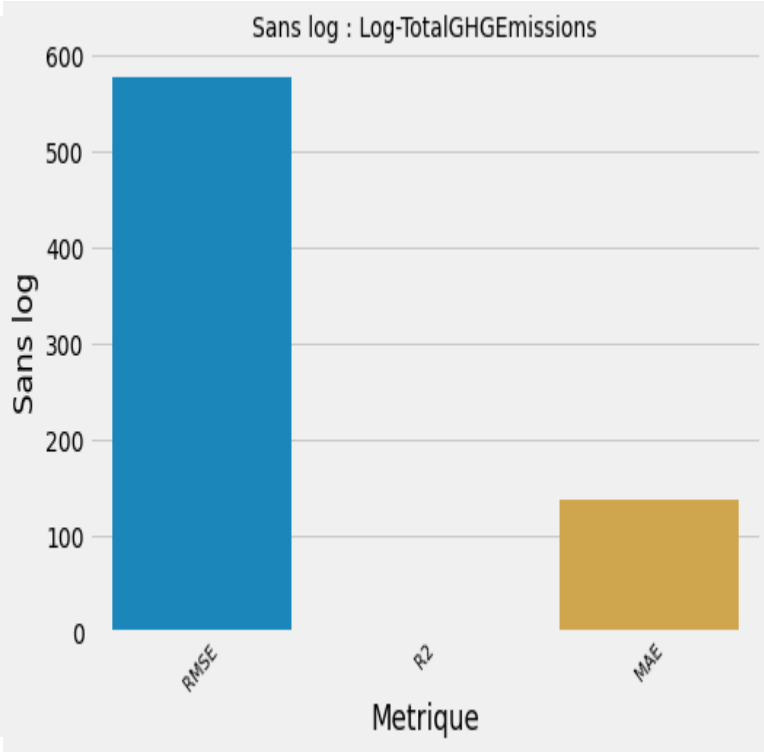
Pourquoi nous avons choisi le XGBoost comme modèle final

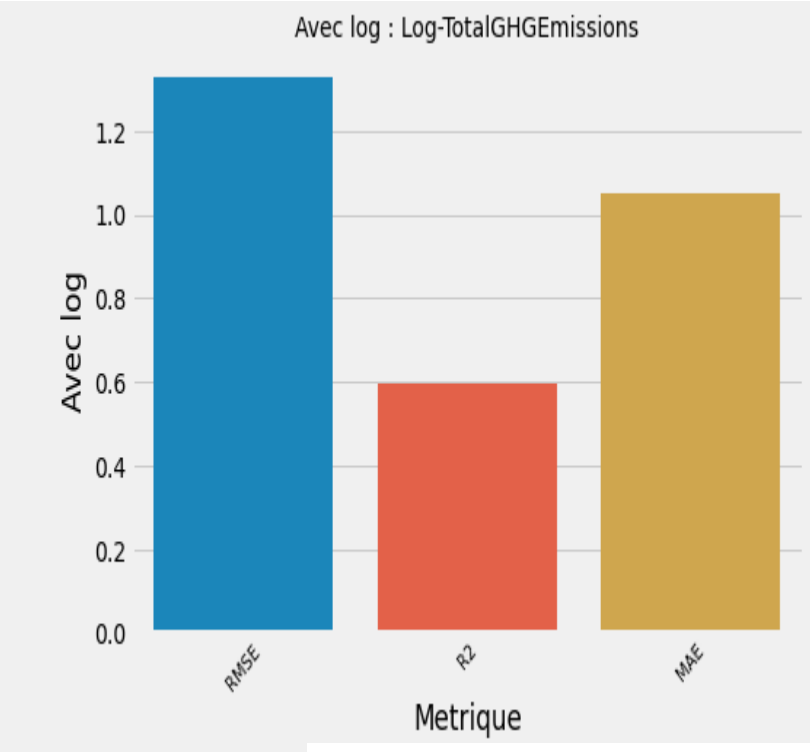
III. MODÉLISATION
D/ **PRÉDICTION SUR LES DONNÉES AVEC CHANGEMENT D'ECHELLE**



III. MODÉLISATION
D/ **PRÉDICTION SUR LES DONNÉES AVEC CHANGEMENT D'ECHELLE**







	Metrique	Sans log	Avec log
0	RMSE	1.390253e+07	1.181014
1	R2	5.718538e-01	0.673413
2	MAE	3.766517e+06	0.667571

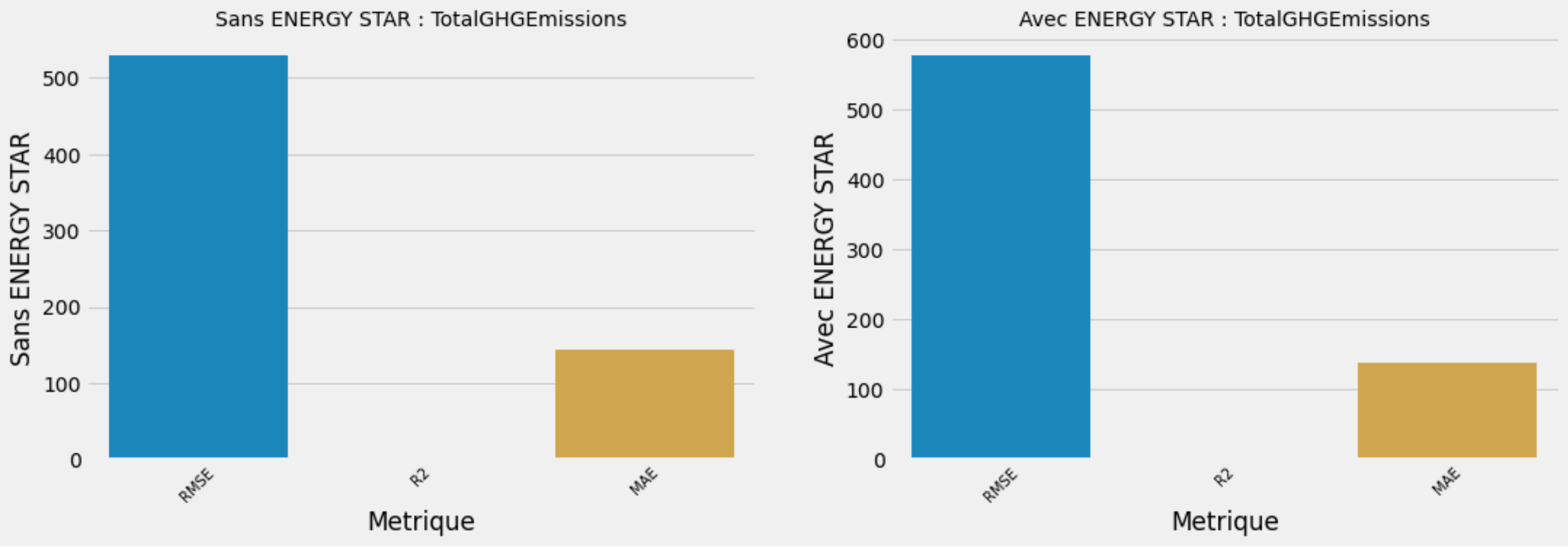
IV. ENERGYSTARSCORE
A/ **INTERÊT**

- L'ENERGY STAR Score est un outil de dépistage aidant à évaluer les performances d'émission de GES d'une propriété par rapport à des bâtiments similaires.
- Cet indicateur se base sur une échelle de 0 à 100 dont la médiane est 50.
  - Si le score est >= 75, le bâtiment peut être admissible à la certification ENERGY STAR.

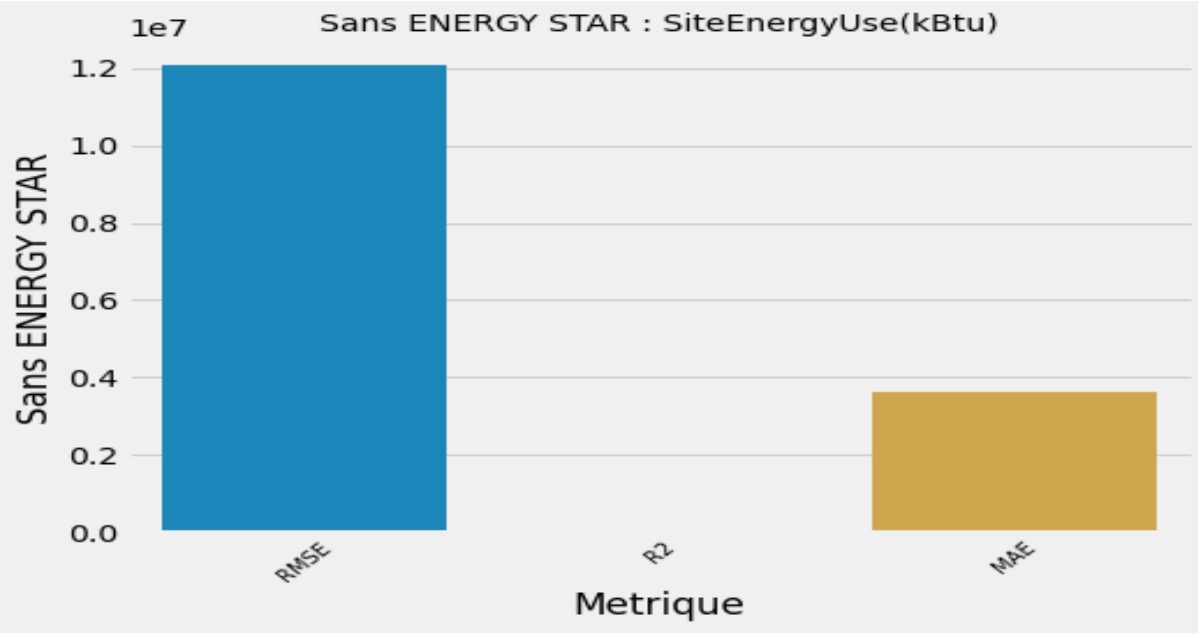
- A partir du meilleur résultat après optimisation, nous allons reconstruire un modèle avec les meilleurs paramètres obtenus, mais en enlevant de la dataframe la variable ENERGYSTARScore
- Ensuite nous allons analyser quelles sont les variables les plus importantes pour ce modèle.

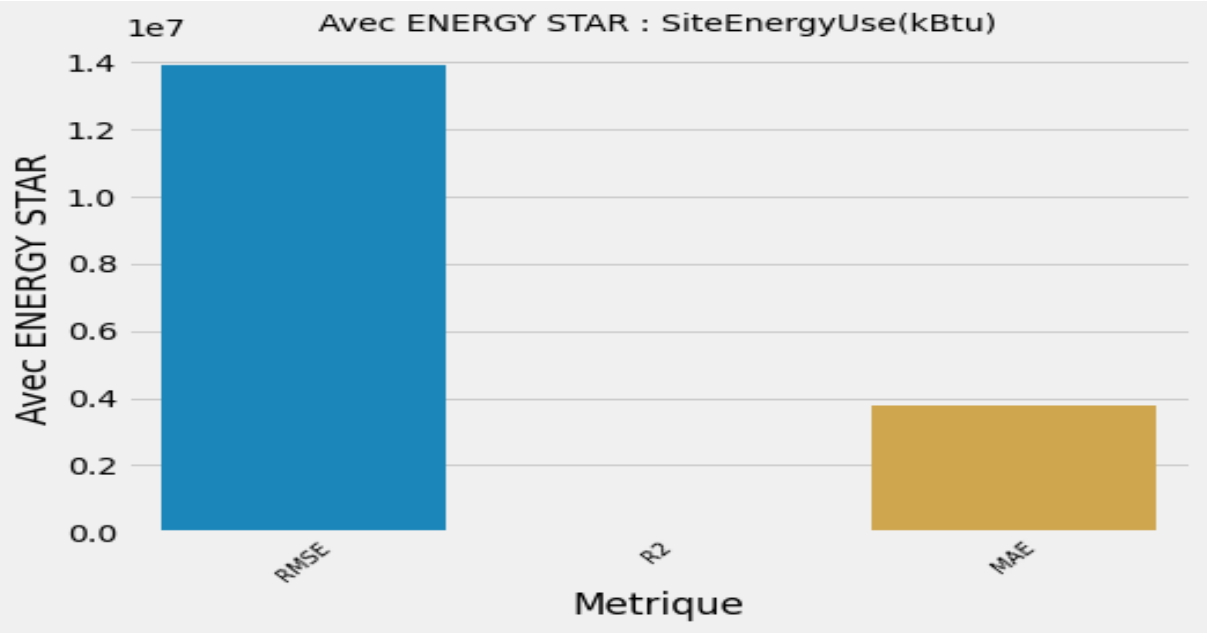
Afin de comparer le fonctionnement de modèles différents, nous allons travailler avec le RandomForestRegressor et le GradientBoostingRegressor.

IV. ENERGYSTARSCORE
A/ **INTERÊT**



IV. ENERGYSTARSCORE
A/ **INTERÊT**





L'ENERGY STAR Score est un outil de dépistage aidant à évaluer les performances d'émission de GES d'une propriété par rapport à des bâtiments similaires.

- Cet indicateur se base sur une échelle de 0 à 100 dont la médiane est 50.
- Si le score est >= 75, le bâtiment peut être admissible à la certification ENERGY STAR.

CONCLUSION

- Les résultats sont globalement décevants. Cela est en partie dû aux données dont nous disposons en entrée.
- Il serait bien d'avoir quelques informations techniques du type :
  - Travaux de rénovation récents
  - Type d'isolation, d'éclairage (LED…)
  - Type de chauffage
- Une base de données avec plus d'observations serait un plus. Nous avons pu constater qu'une marge d'amélioration est possible de ce côté avec la learning curve.

- **Intérêt de l'ENERGY STAR Score :**
  - Prédictions GES AVEC la feature légèrement meilleures que les prédictions GES SANS la feature
  - La feature ne représente que peu d'intérêt