Soutenance Projet OC P4 DS: Anticipez les besoins en consommation électrique des bâtiments

Candidat: David CAPELLE
Mentor: Nicolas MICHEL

Evaluateur: Rick BOURGET
Formation 100% Pôle Emploi

16/09/2021

Plan de la soutenance

- Problématique et présentation de la démarche
- Présentation du nettoyage des données, du feature engineering et de l'analyse exploratoire des données
- Présentation des pistes de modélisation testées
  - Synthèse sur les métriques
  - Synthèse sur l'analyse des performances des modèles
- Conclusion sur le choix du modèle retenu et améliorations

Problématique du projet

- Prédire la consommation d'énergie et les émissions de CO2 des bâtiments de Seattle avec du machine learning, via un apprentissage sur des modèles supervisés.
- Ne pas utiliser les données relevés de 2015/2016 car les ressources sont trop coûteuses pour la ville.
- Création de nouvelles variables pertinentes pour les prédictions.
- Testes plusieurs modèles en optimisant leurs hyper-paramètres.
- Choix du modèle final en se basant sur plusieurs hypothèses en amont dans le traitement des données.
- Analyse de l'impact du score EnergyStar sur les émissions de CO2.

Démarche pour le choix du modèle

- Itérations de la boucle suivante sur les 4 hypothèses sur le traitement des données (pistes de modélisation) :
  - Réglage des hyper-paramètres sur une sélection de modèles (GridSearchCV, CV=10).
  - Entraînement de modèles et prédictions à partir des meilleurs paramètres.
  - Comparaison des performances des modèles par rapport à la métrique de référence (RMSE)
  - Détermination du meilleur modèle suivant l'hypothèse de traitement des données
- Choix final du meilleur modèle et améliorations.

Hypothèses testées pour pré-traitement des données

- **Hypothèse 1**: modélisation simple à partir de 11 variables, standardisation des données en entrée, valeurs réelles sur variables cible.

- **Hypothèse 2**: modélisation à partir de 11 variables, standardisation des données en entrée, **variables cible à l'échelle logarithmique**.

- **Hypothèse 3**: modélisation à partir de 11 variables, standardisation des données en entrée, variables cible à l'échelle logarithmique, **suppression des variables atypiques sur les variables cible (méthode inter-quartiles)**.

- **Hypothèse 4**: **modélisation à partir de 29 variables (ajout variables catégorielles avec encodage OneHot)**, standardisation des données en entrée, variables cible à l'échelle logarithmique, suppression des variables atypiques sur les variables cibles (méthode inter-quartile).

Sélection de modèles, baseline et métriques

- **Sélection de modèles pour régression linéaire :**
  - Modèles linéaires: LinearRegression, ElasticNet.
  - Modèles non linéaires: SVR, Random Forest Regressor, Gradient Boosting Regressor, Extreme Gradient Boosting.

- **Baseline**: modèle LinearRegression, modèle simple, sans réglage d'hyper-paramètre.

- **Métriques retenues**:
  - Métrique principale (RMSE), sensible aux valeurs atypiques.
  - 2 métriques secondaires : MAPE, plus compréhensible par les utilisateurs (taux d'erreur moyen sur un jeu de données) et R^2 pour évaluer le over/underfiting sur les jeux de données d'apprentissage.

Présentation du jeu de données

- **2 types de données dans les dataset 2015 et 2016 :**
  - Données relatives au permis d'exploitation commerciale (emplacement géographique, nombre de bâtiments et d'étages, types d'utilisation,…)
  - Relevés de consommation et d'émission 2015 et 2016 de plusieurs indices: électricité, gaz, …
  - Score Energy Star

- Le dataset correspondant aux données 2015 compte 3340 bâtiments et 47 variables.

- Le dataset correspondant aux données 2016 compte 3376 bâtiments et 46 variables.

Constitution du jeu de données

- Concaténation des datasets 2015 et 2016.
- Suppression des doublons de bâtiments sur les 2 années.
- Suppression des valeurs redondante sur les relevés.
- Suppression de variables normalisées par rapport aux conditions météorologiques
- Suppression de variables quantitatives et qualitatives inutiles à la modélisation.

=> Taille du jeu de données : 3432 lignes, 34 variables

Proportion de NaN : 14.43 %

Feature engineering : création de nouvelles variables

- **Création des variables suivantes (phase préparation données) :**
  - Variable « TotalUseTypeCounts » (nombre total d'usage du bâtiment).
  - Variable « AgeOfBuilding » (age du bâtiments).
  - Variable « Distance_Harversine » à partir de la latitude/longitude.

- **Création des variables suivantes (phase analyse exploratoire) :**
  - Variable « AreaBuildingsMean » (surface réelle moyenne par bâtiment).
  - Variable « AreaParkingMean » (surface moyenne de parking par bâtiment).

NB : les variables à l'origine de ces nouvelles données ont été supprimées.

Nettoyage des données - Traitement valeurs manquantes

- **Traitement de valeurs quantitatives :**
  - Suppression des lignes NaN pour les relevés
  - Imputation de valeurs NaN pour le variables exprimées en GFA en cohérence avec le nombre d'usage des bâtiments.
  - Imputation valeur moyenne pour la variable "ENERGYSTARScore".

  => Taille du jeu de données : (3428 lignes, 32 variables)
  Proportion de NaN : 10.4 %

- **Traitement de valeurs qualitatives :**
  - Imputation valeur 'Other' si présence de valeurs NaN pour variables « xxxPropertyUseType ».
  - Suppression des variables « Comments » et « Outliers ».

  => Taille du jeu de données : (3262 lignes, 29 variables)
  Proportion de NaN : 0 %

Nettoyage des données - Traitement des outliers

- **Traitement des outliers - valeurs aberrantes :**
  - Suppression des valeurs négatives pour la variable "PropertyGFAParking"
  - Suppression des valeurs négatives pour la variable "PropertyGFABuilding(s)".

- **Traitement des outliers - analyse valeurs atypiques :**
  - Pour les variables à prédire
  - On décide de garder les valeurs atypiques correspondant à des campus universitaires avec des grandes surfaces et des nombreux usages des bâtiments.

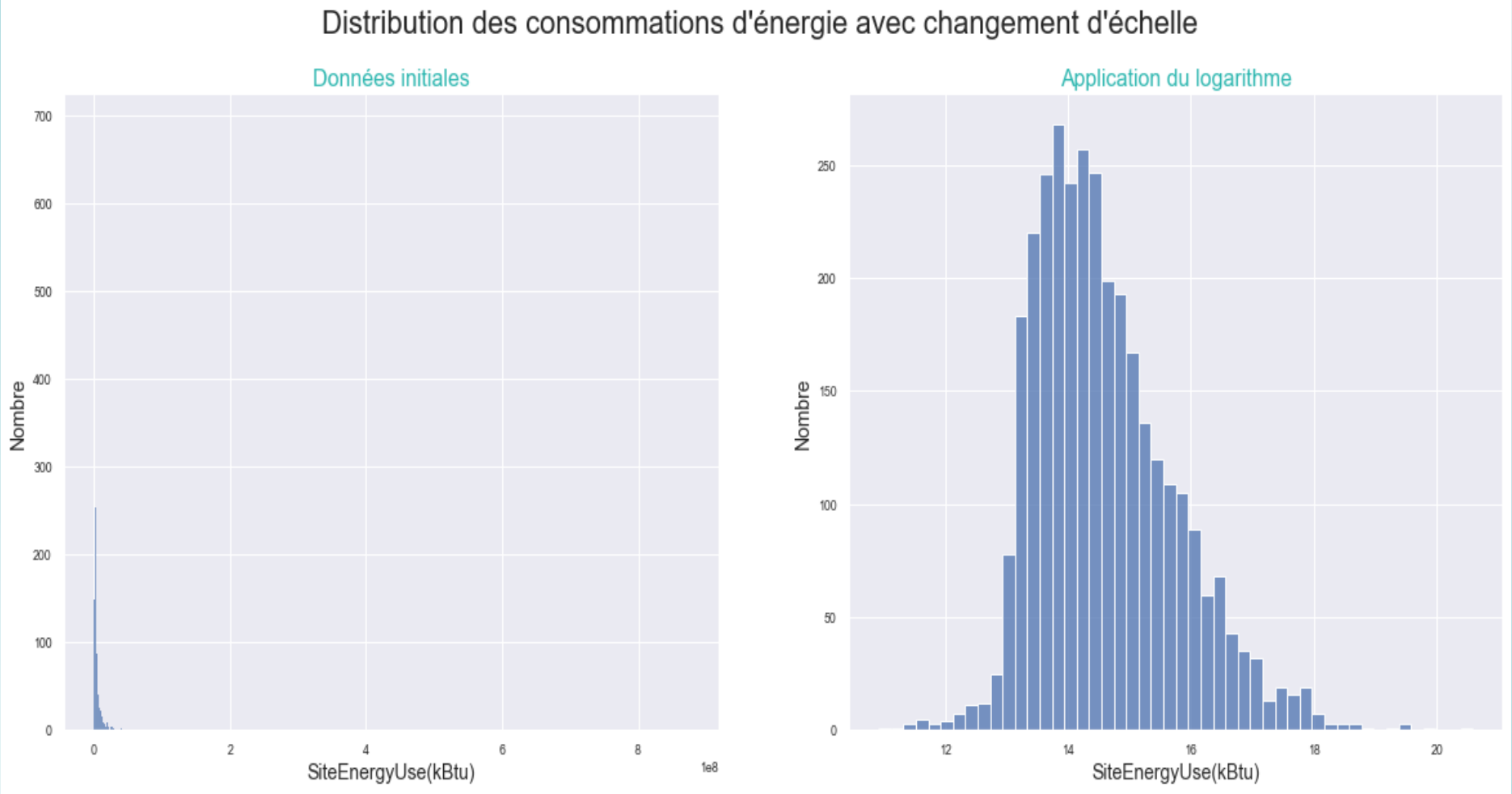
=> Taille du jeu de données : (3259 lignes, 29 variables)
Proportion de NaN : 0 %

Analyse exploratoire des données
Analyse univariée

- **Variables à prédire:**
  - Distribution très asymétrique pour les 2 variables
  - Passage à l'échelle logarithmique pour obtenir une distribution normale.

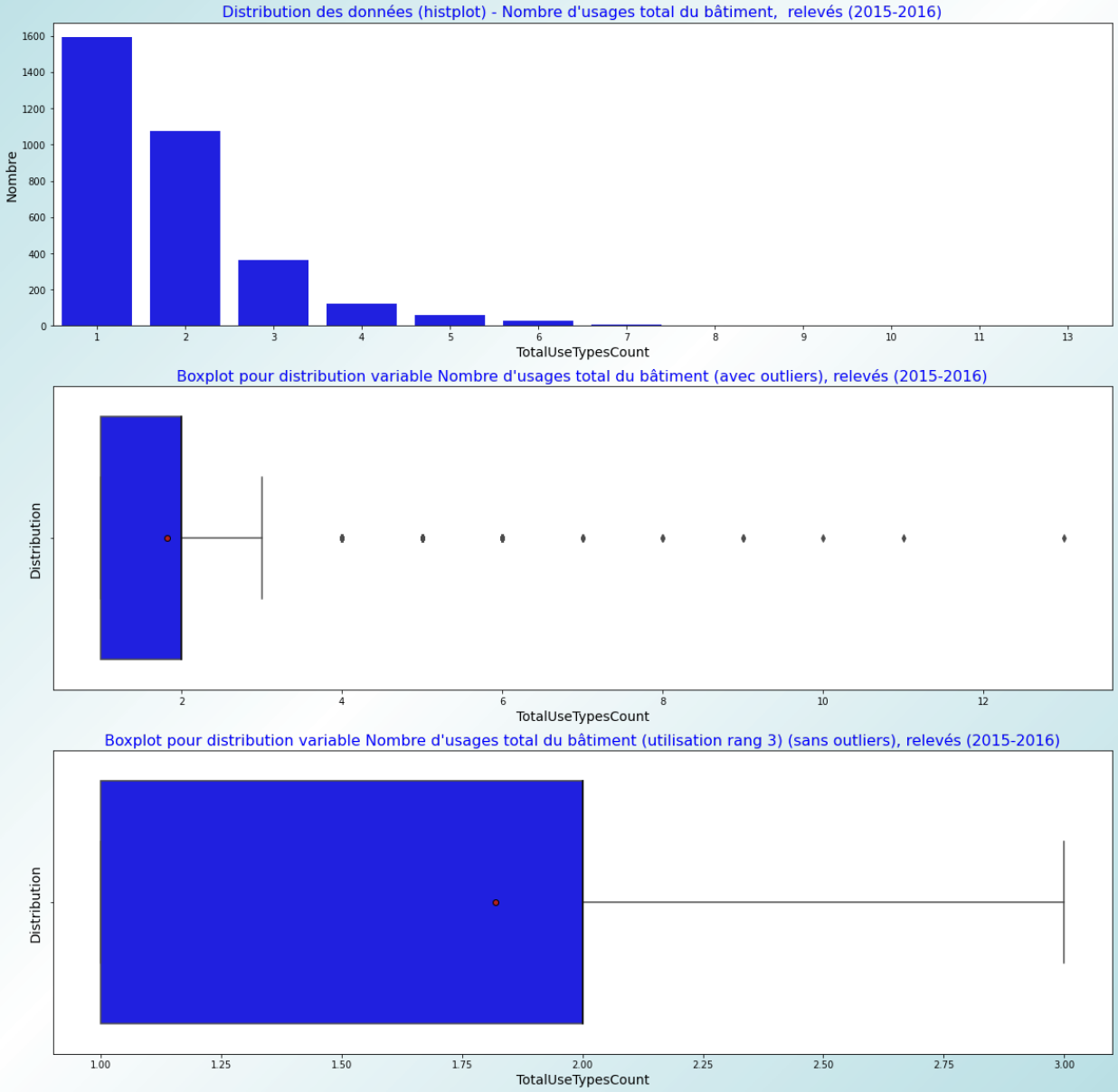
- **Autres variables :**
  - Distribution asymétrique des données pour les surfaces
  - La standardisation des données dans le pré-traitement permet d'atteindre une distribution quasi normale.

Analyse exploratoire des données
Analyse univariée – Passage au log

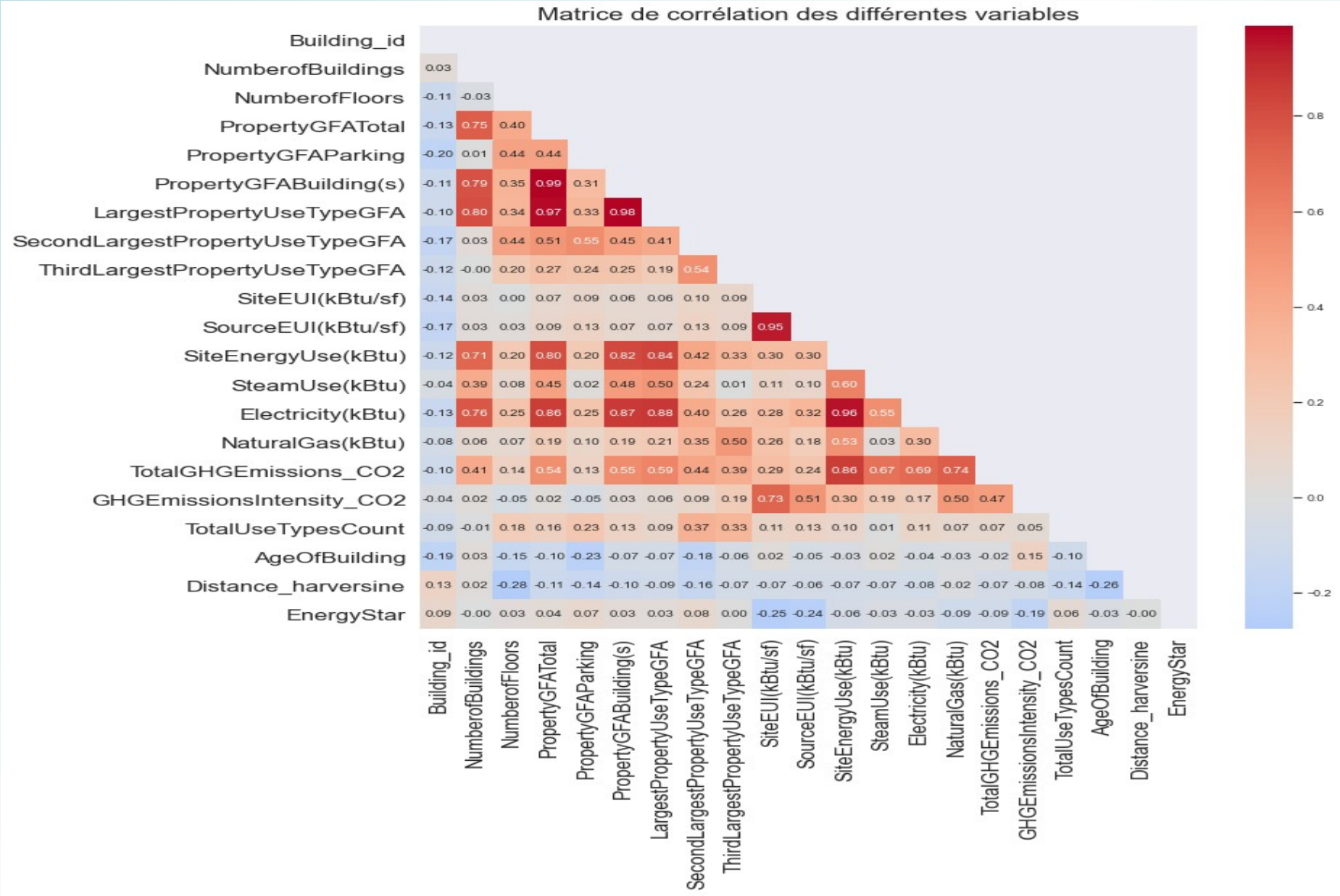


Analyse exploratoire des données

Analyse univariée – Distribution variable « Nbre total usage bâtiment »



Analyse exploratoire des données
Analyse multivariée – Matrice des corrélations



Piste de modélisation – Hypothèse 1 traitement données

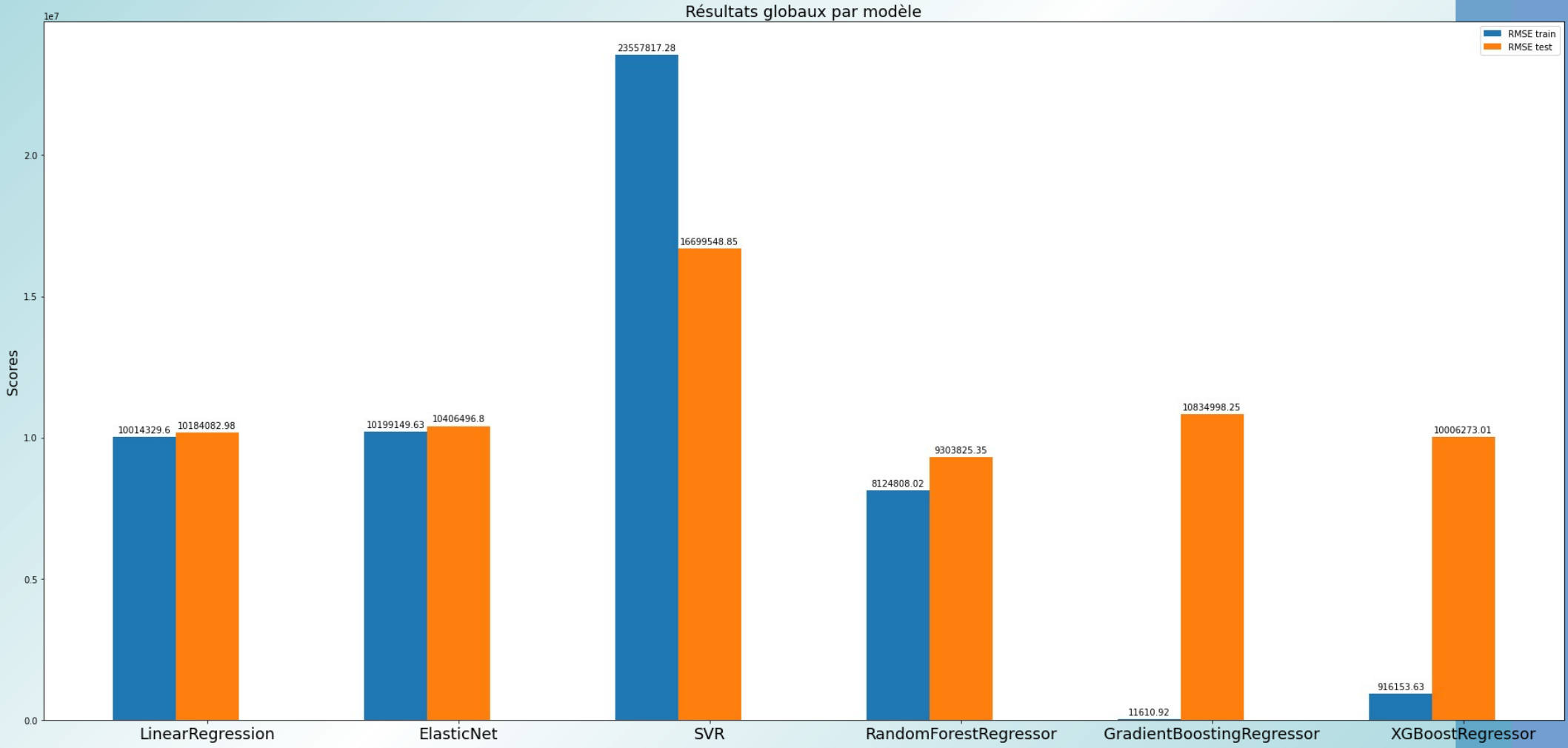
Principes

- Modélisation à partir de peu de variables en entrée (11 variables).
- Prédictions sur des variables en valeurs réelles
- Standardisation des données en entrée par StandardScaler()
- Conservation des valeurs atypiques dans les variables à prédire.
- Non prise en compte de la variable EnergyStar score

=> hypothèse 1 = modélisation simple, sans grande transformation des données.

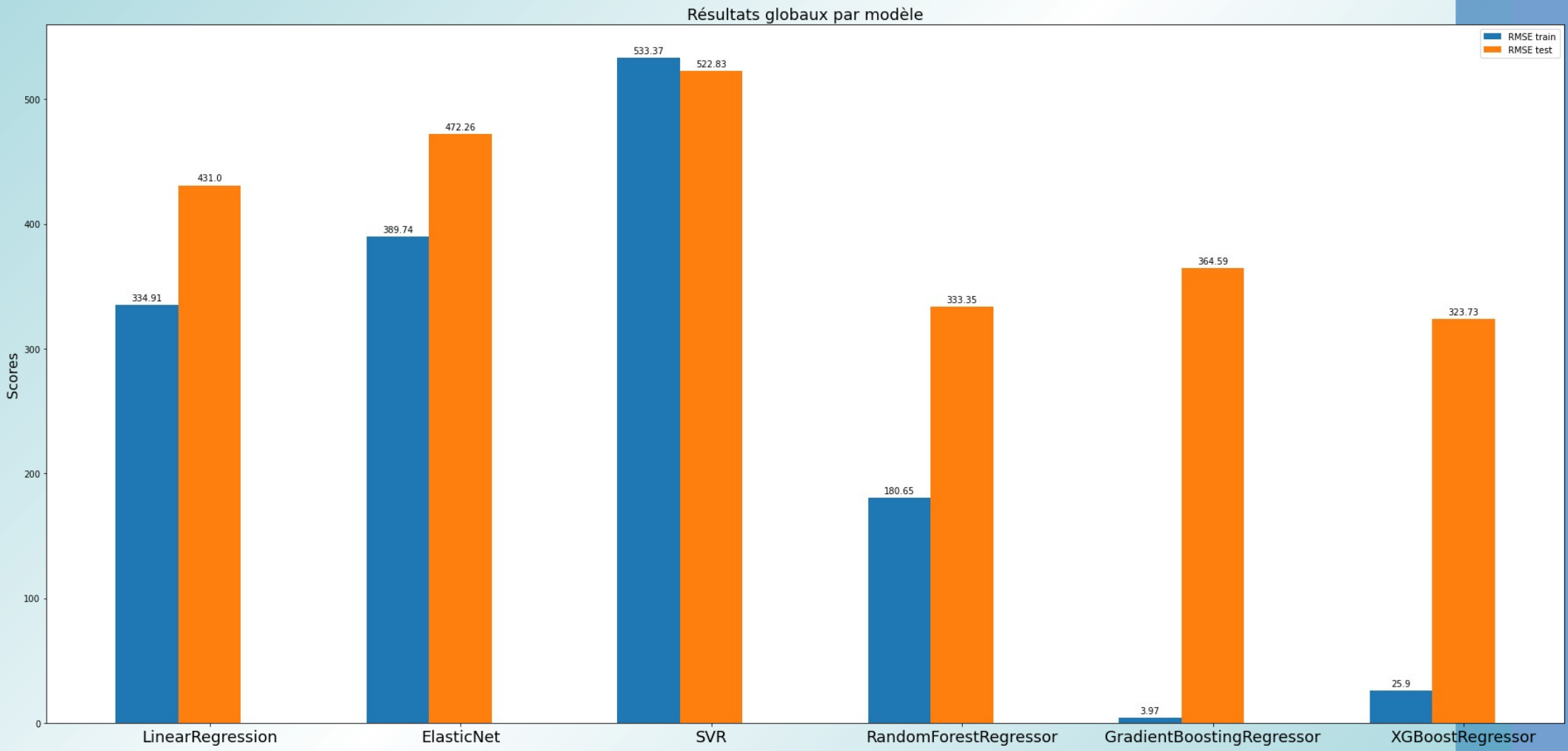
Piste de modélisation – Hypothèse 1

Métrique RMSE – Consommation d'énergie



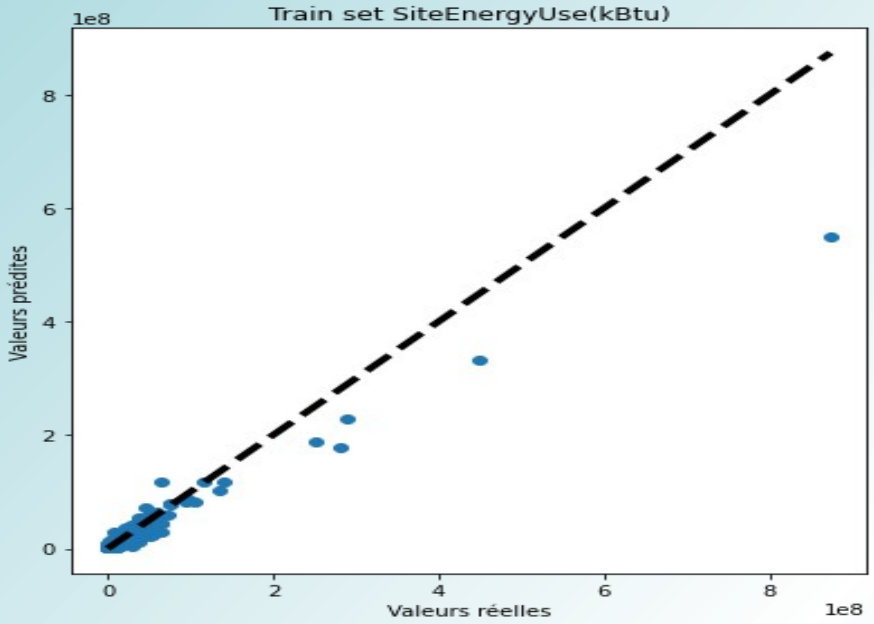
Piste de modélisation – Hypothèse 1

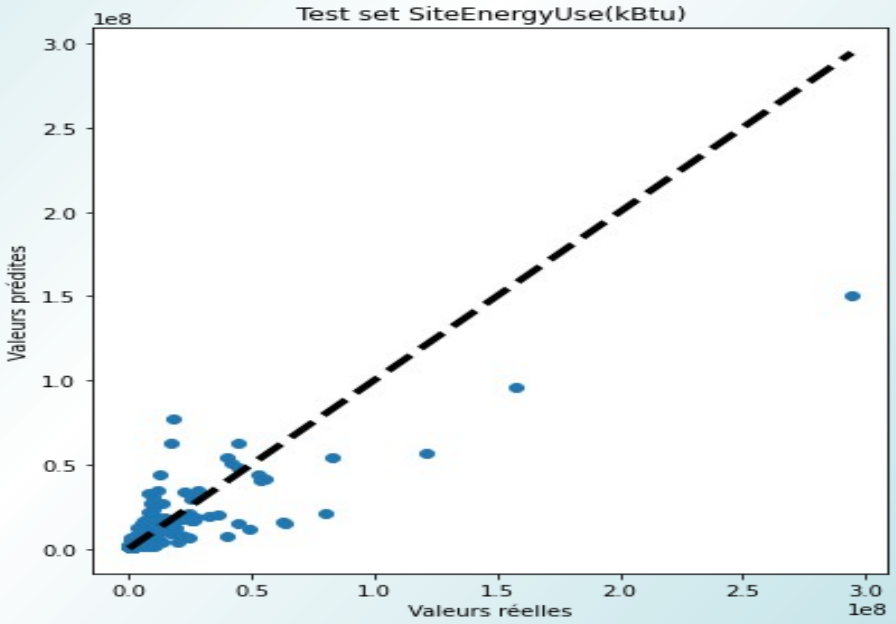
Métrique RMSE – Émissions de CO2

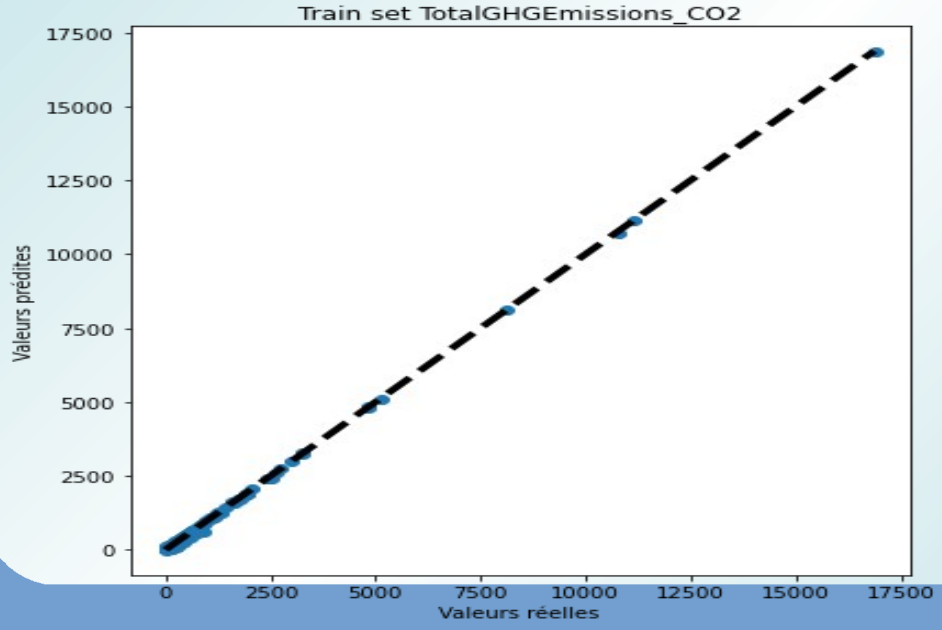


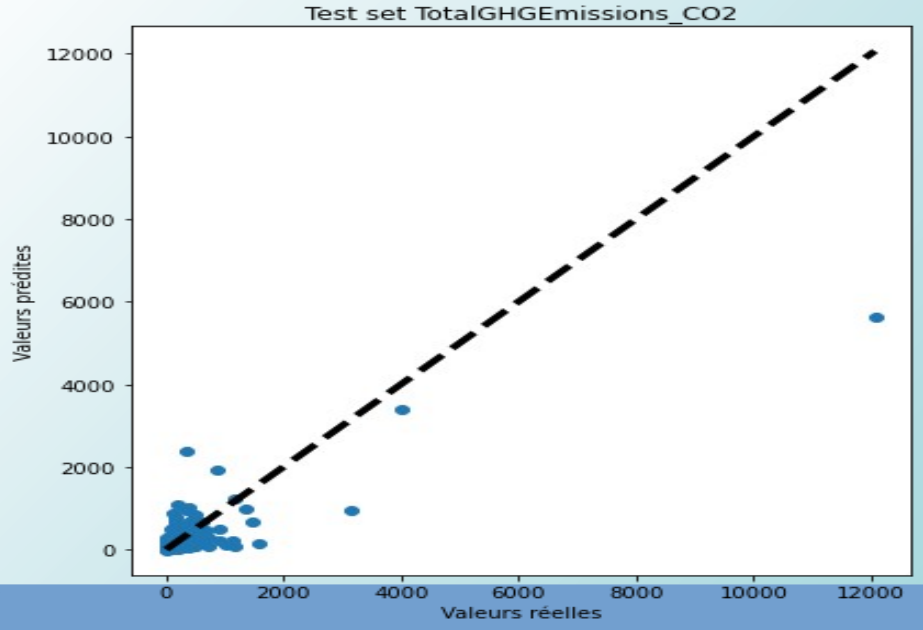
Piste de modélisation – Hypothèse 1

Graphe valeurs cible réelles / valeurs cible prédites









Piste de modélisation – Hypothèse 1

Interprétation des résultats et métriques secondaires

- Variable « Consommation d'énergie » :
  - **Meilleur modèle dans ce cas : Random Forest Regressor (RFR).**

Métriques	RMSE	MAPE	R^2
Baseline : LinearRegression	10 184 082	94 %	0,61
Meilleur modèle : RFR	9 303 825	77 %	0,67

- Variable « Emissions de CO2 » :
  - **Meilleur modèle dans ce cas : Extreme Gradient Boosting (XGR).**

Métriques	RMSE	MAPE	R^2
Baseline : LinearRegression	431	364 %	0,36
Meilleur modèle : XGR	323,73	207 %	0,64

=> Sur-apprentissage de l'ensemble des estimateurs sur le jeu de train.

=> Influence néfaste de quelques valeurs atypiques.

=> Variance du modèle retenu (RMSE relative) = 167 %(Energie) / 256 % (CO2) de la moyenne des observations (jeu de test), mauvaise qualité des prédictions.

r

Piste de modélisation – Hypothèse 2 traitement données

Principes

- Point d'entrée : hypothèse 1.

- \+ Passage à l'échelle logarithmique des variables à prédire.

=> Résultats hypothèse 2 :

- Résultat pire que celui de l'hypothèse 1.

- Passage à l'échelle logarithmique de la variable cible permet de réduire le sur-apprentissage sur le jeu train

- Mais la conversion des cette variable aux valeurs réelles dégrade la RMSE.

Piste de modélisation – Hypothèse 2

Interprétation des résultats et métriques secondaires

- Variable « Consommation d'énergie » :
  - **Meilleur modèle dans ce cas : Gradient Boosting Regressor (GBR).**

Métriques	RMSE	MAPE	R^2
Baseline : LinearRegression	626 533 900	90 %	0,48
Meilleur modèle : GBR	17 176 150	56 %	0,71

- Variable « Emissions de CO2 » :
  - **Meilleur modèle dans ce cas : Gradient Boosting Regressor (GBR).**

Métriques	RMSE	MAPE	R^2
Baseline : LinearRegression	1,395696e+26	4.906367e+20	-1,73
Meilleur modèle : GBR	380	120 %	0,59

=> Les modèles linéaires (dont la baseline) ne semblent pas être adaptés pour cette hypothèse.

=> Variance du modèle retenu (RMSE relative) = 274 % (Energie) / 301 % (CO2) de la moyenne des observations (jeu de test), très mauvaise qualité des prédictions.

Piste de modélisation – Hypothèse 3 traitement données

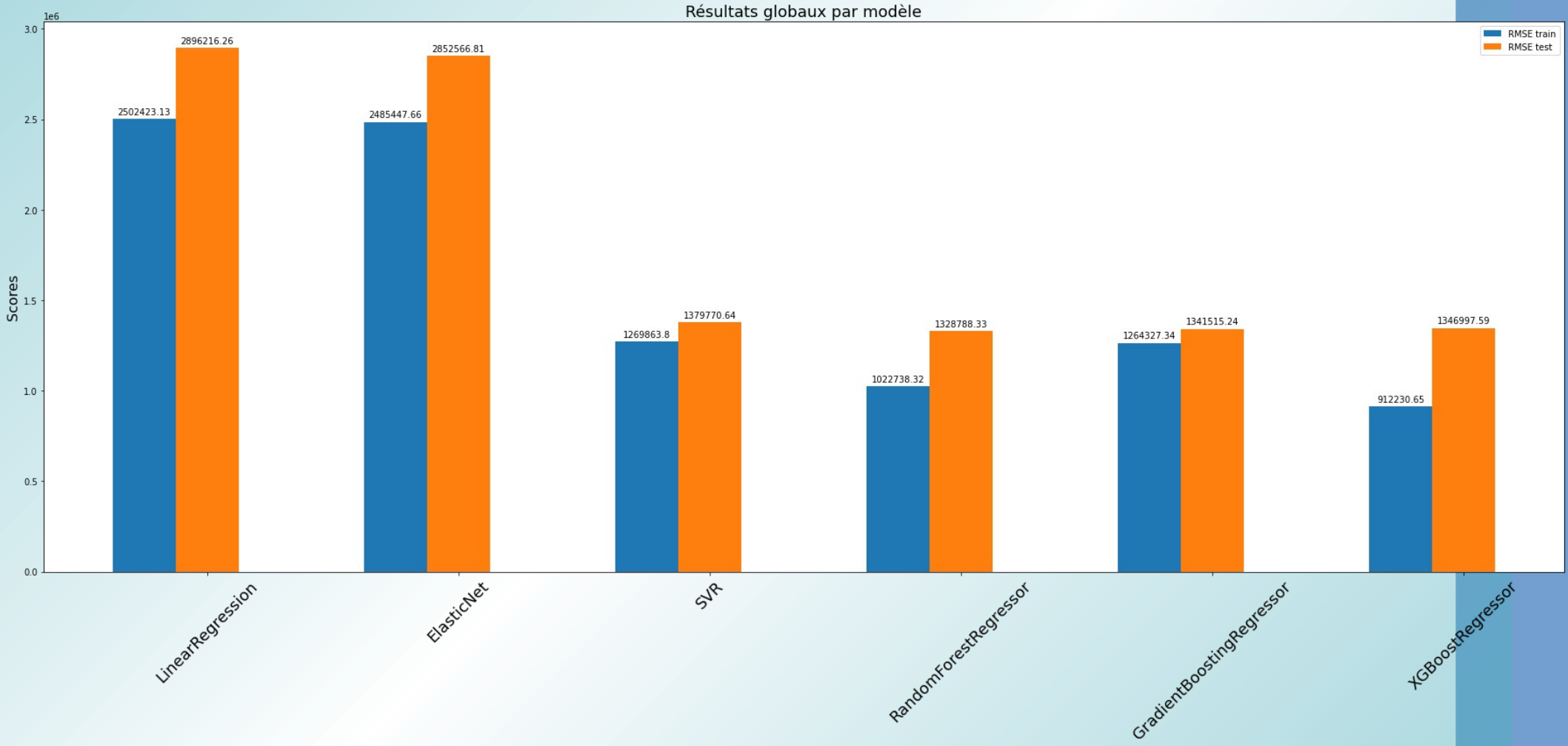
Principes

- Point d'entrée : hypothèse 2.
- + Suppression des valeurs atypiques par méthode inter-quartiles sur les variables à prédire, suite passage à l'échelle logarithmique de ces variables.
- Suppression de 11 % des valeurs du jeu de données pour les 2 variables .

=> Cette piste de modélisation proposera des prédictions peu pertinentes si un jeu de données inédit contient des valeurs atypiques.

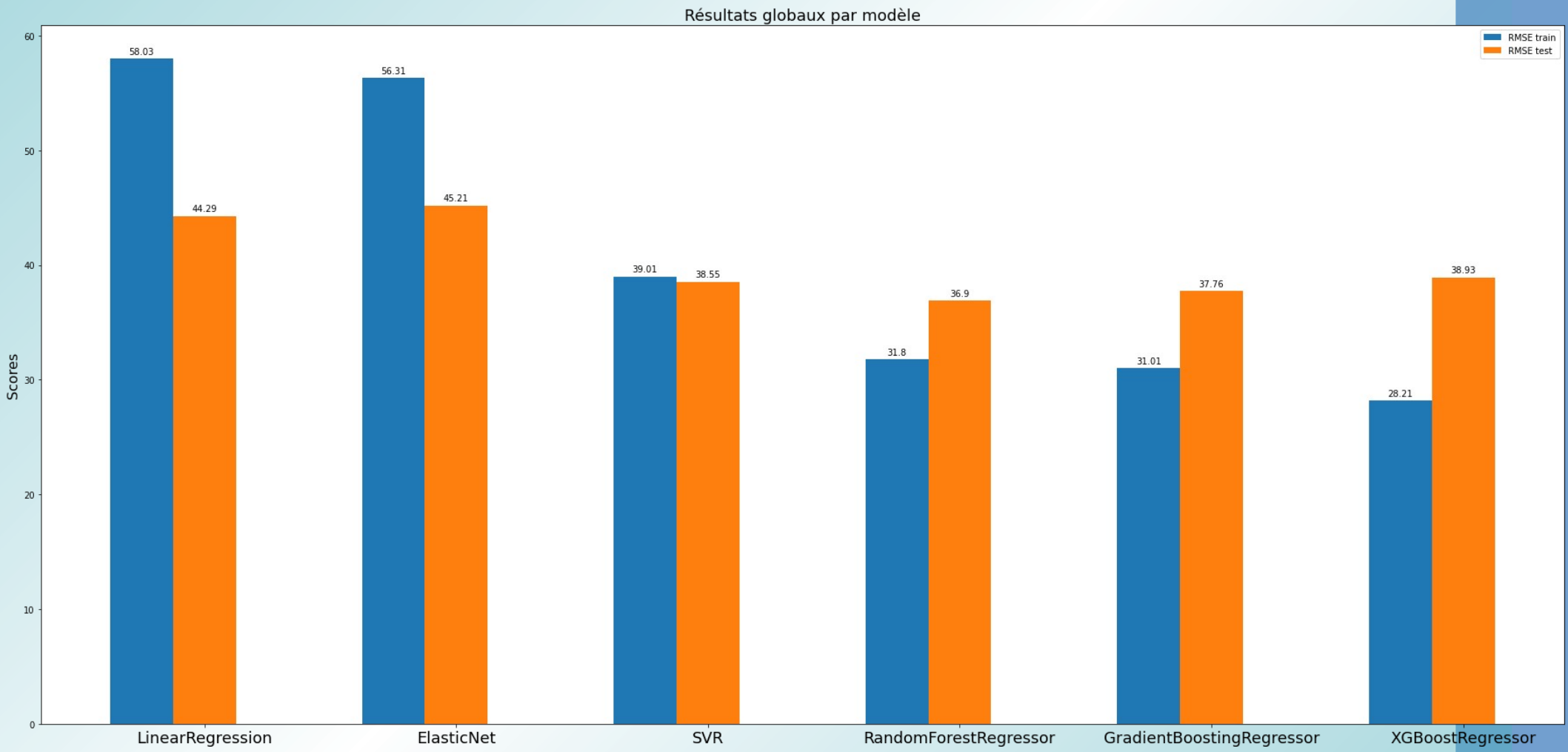
Piste de modélisation – Hypothèse 3

Métrique RMSE – Consommation d'énergie



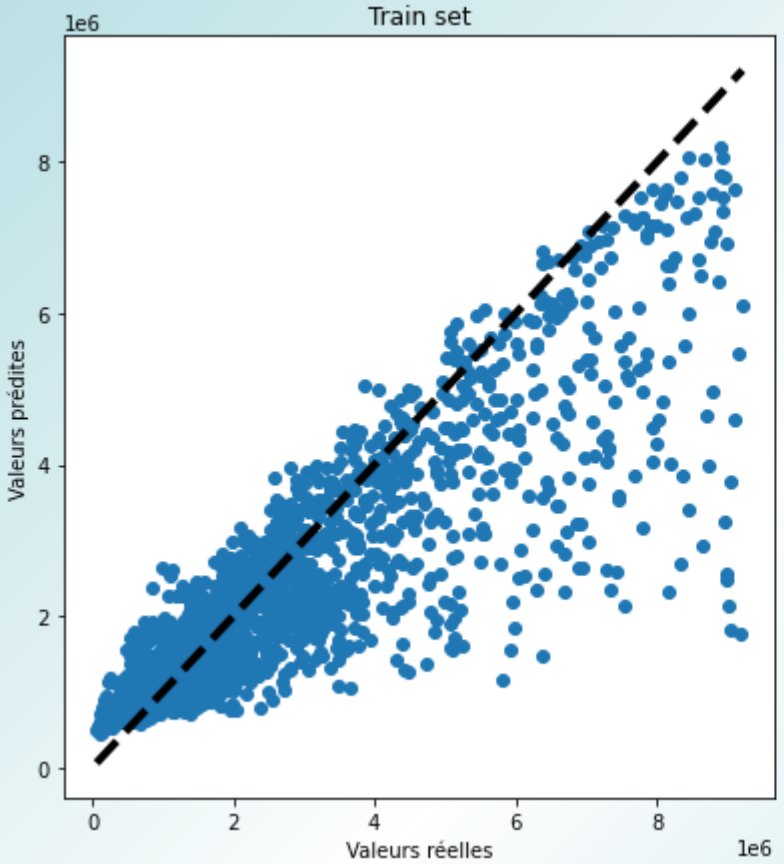
Piste de modélisation – Hypothèse 3

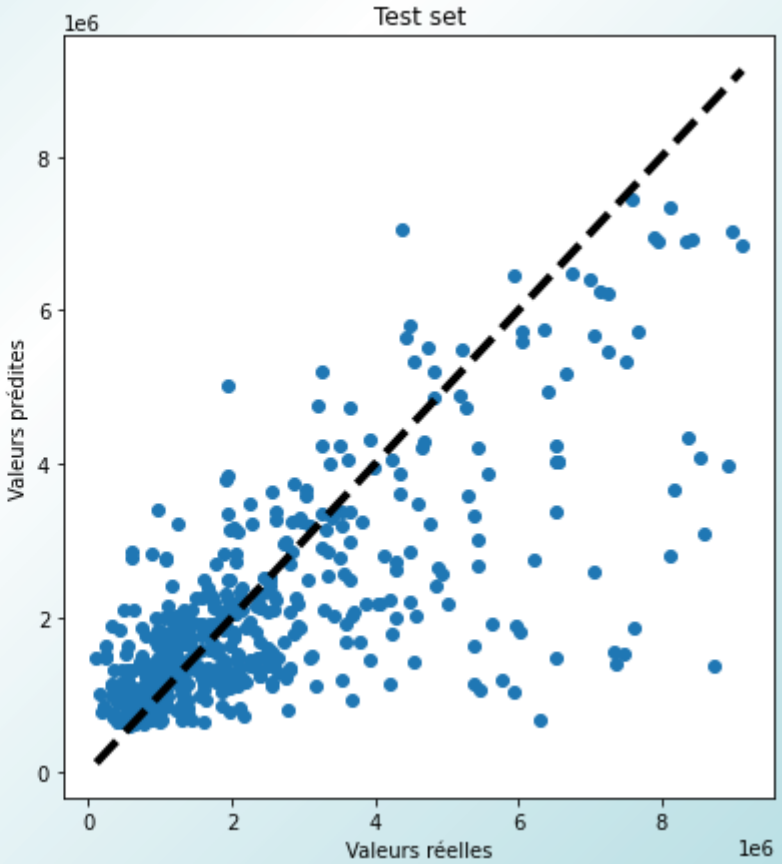
Métrique RMSE – Émissions de CO2



Piste de modélisation – Hypothèse 3

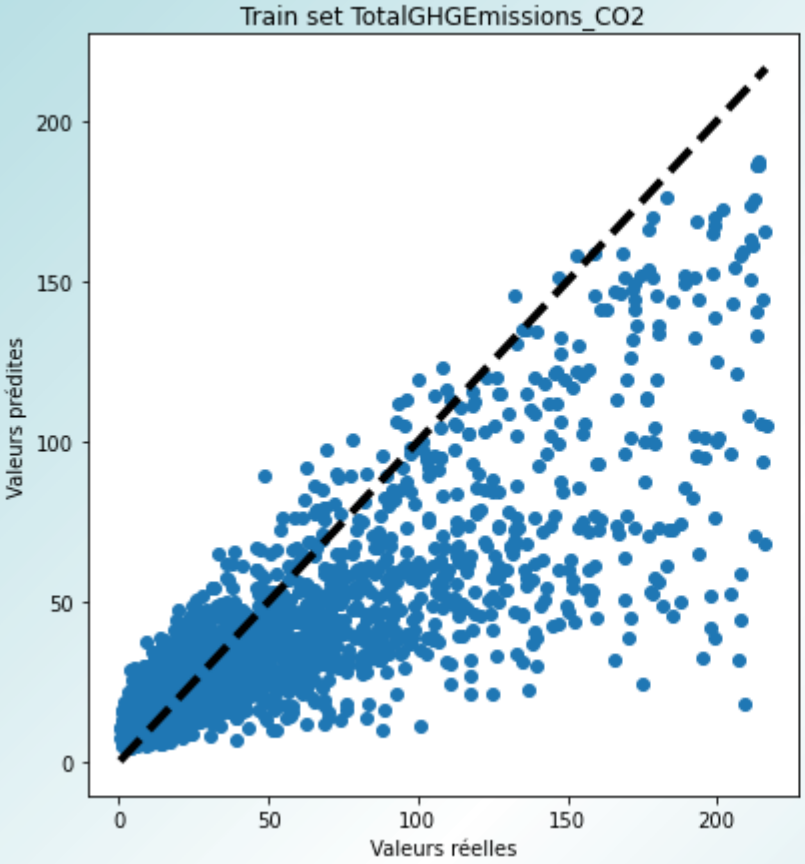
Graphe valeurs cible réelles / valeurs cible prédites Energie

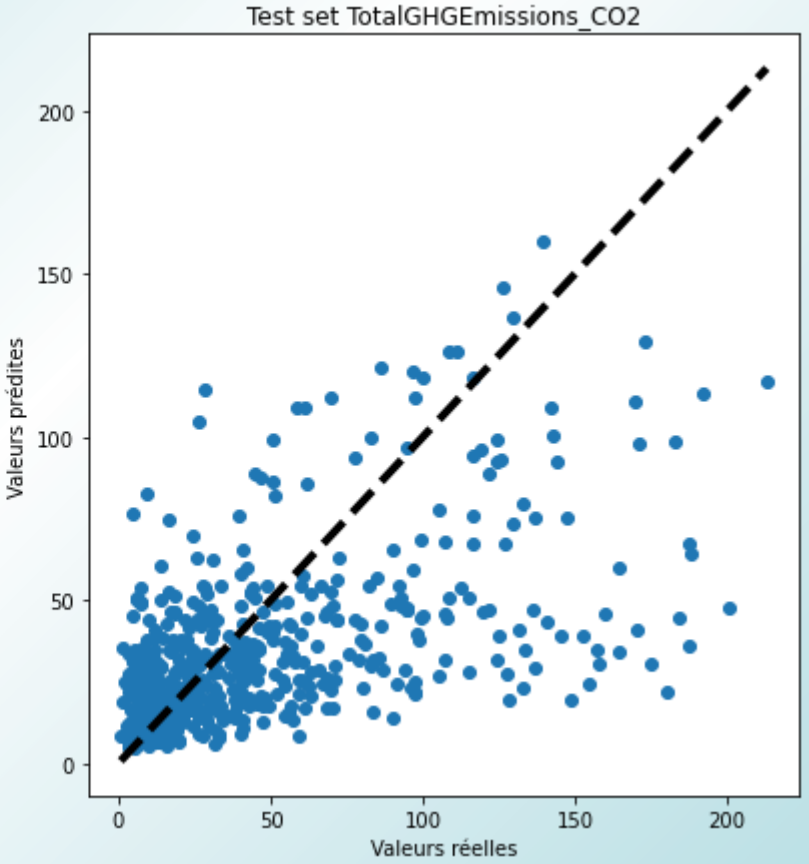




Piste de modélisation – Hypothèse 3

Graphe valeurs cible réelles / valeurs cible prédites Emissions CO2





Piste de modélisation – Hypothèse 3

Interprétation des résultats et métriques secondaires

- Variable « Consommation d'énergie » :
  - **Meilleur modèle dans ce cas : Random Forest Regressor (RFR).**

Métriques	RMSE	MAPE	R^2	RMSE relative
Baseline : LinearRegression	2 896 216	61 %	0,38	N/A
Meilleur modèle : RFR	1 328 788	48 %	0,53	60 % moy obs.

- Variable « Emissions de CO2 » :
  - **Meilleur modèle dans ce cas : Random Forest Regressor (RFR).**

Métriques	RMSE	MAPE	R^2	RMSE relative
Baseline : LinearRegression	44,29	121 %	0,25	N/A
Meilleur modèle : RFR	36,9	101 %	0,44	87 % moy obs.

=> Amélioration sensible de la RMSE, mais la conversion aux valeurs réelles montrent des valeurs sous-estimées.

=> Variance du modèle retenu (RMSE relative) = qualité médiocre des prédictions.

=> Les valeurs atypiques représentent environ 86% (Energie) et 88 % (CO2) de la RMSE totale

r

Piste de modélisation – Hypothèse 4 traitement données

Principes

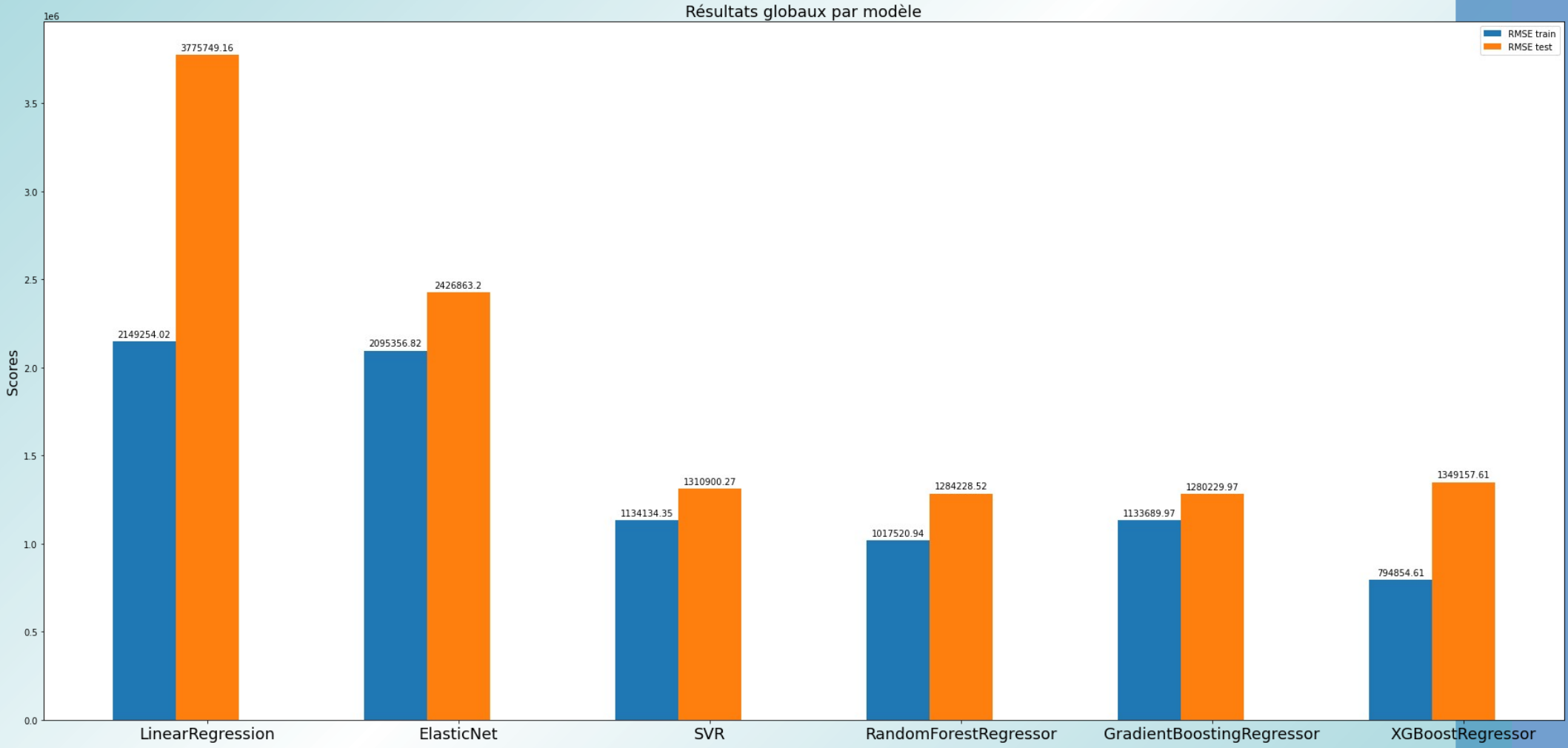
- Point d'entrée : hypothèse 3.
- + Encodage OneHot sur des variables catégorielles sur les types et les utilisations des bâtiments
- Création de 18 variables supplémentaires

=> hypothèse 4 = plus grande complexité du modèle.

=> Cette hypothèse permet un amélioration de la RMSE pour la variable Energie, mais pas pour la variable Emissions de CO2.

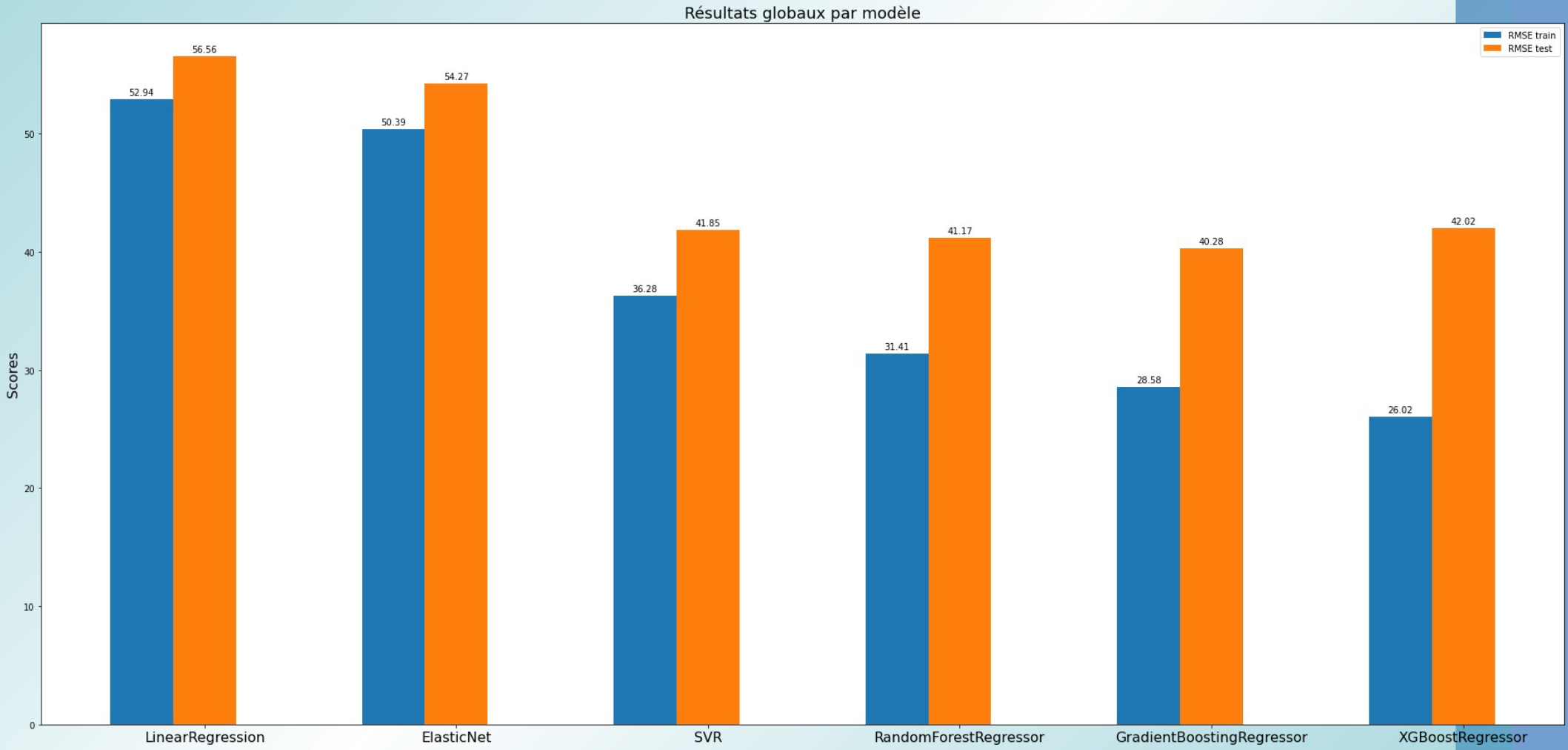
Piste de modélisation – Hypothèse 4

Métrique RMSE – Consommation d'énergie



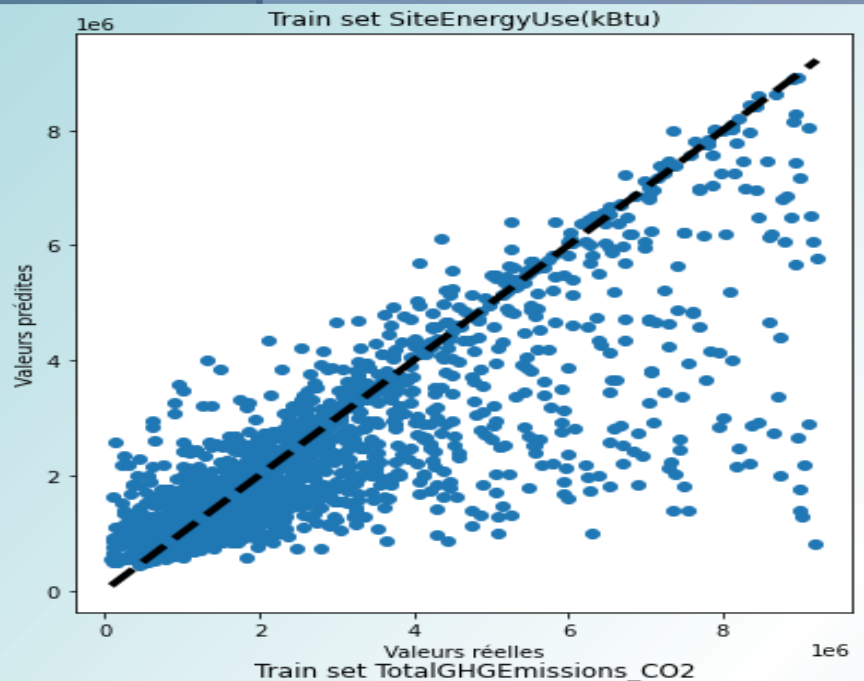
Piste de modélisation – Hypothèse 4

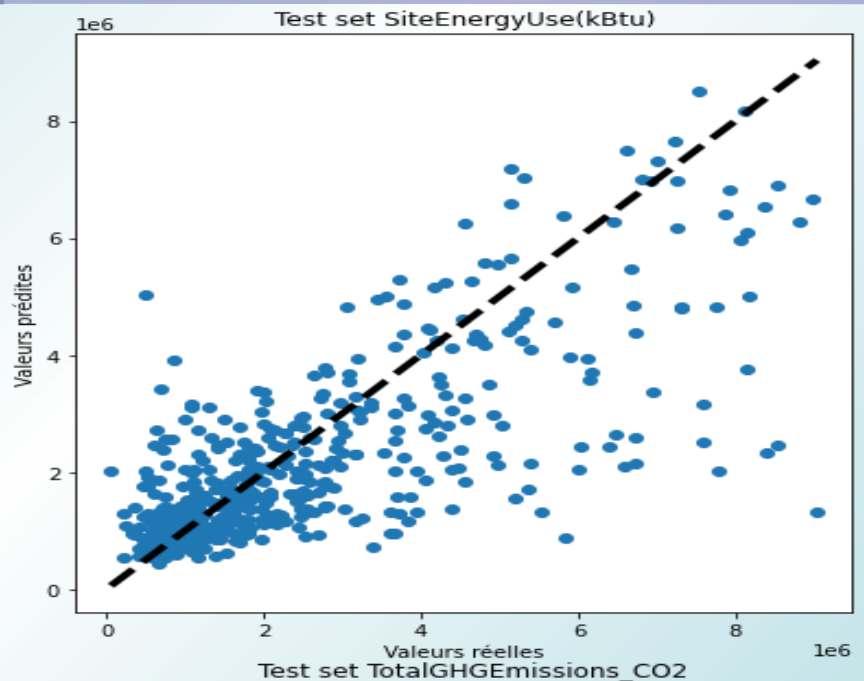
Métrique RMSE – Émissions de CO2

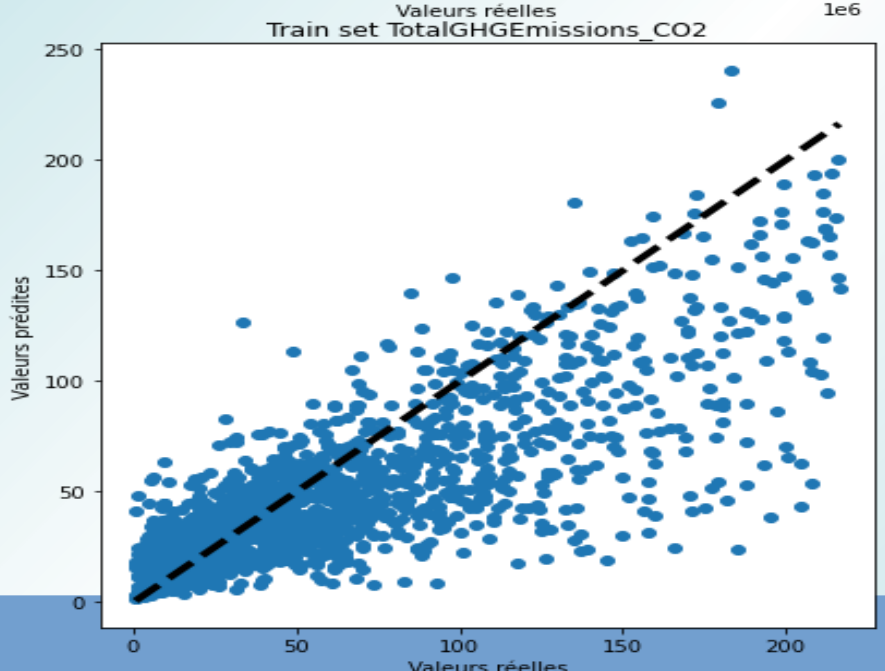


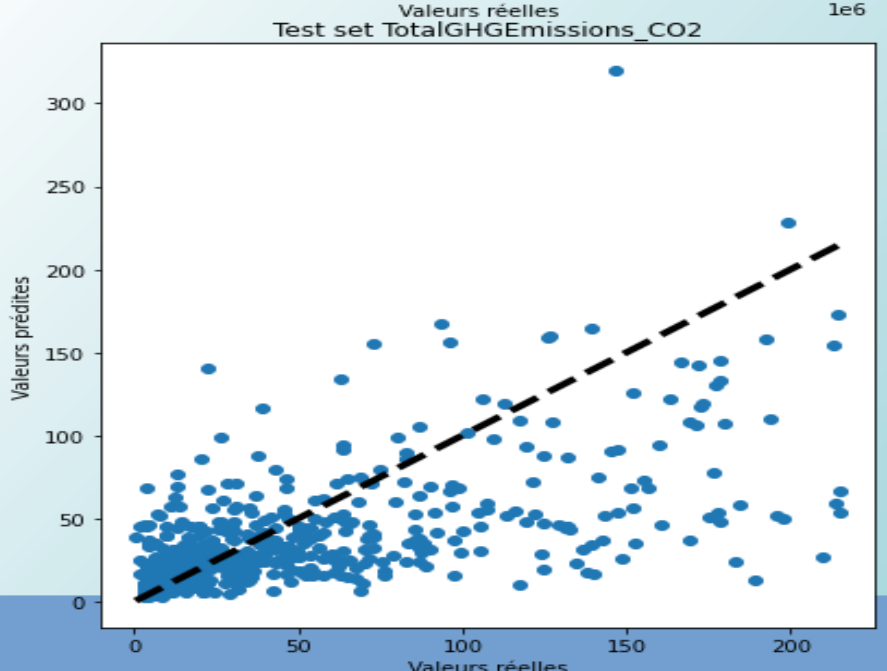
Piste de modélisation – Hypothèse 4

Graphe valeurs cible réelles / valeurs cible prédites









Piste de modélisation – Hypothèse 4

Interprétation des résultats et métriques secondaires

- Variable « Consommation d'énergie » :
  - **Meilleur modèle dans ce cas : Gradient Boosting Regressor (GBR).**

Métriques	RMSE	MAPE	R^2	RMSE relative
Baseline : LinearRegression	3 775 749	61 %	0,38	N/A
Meilleur modèle : GBR	1 280 229	48 %	0,56	56 % moy obs.

- Variable « Emissions de CO2 » :
  - **Meilleur modèle dans ce cas : Gradient Boosting Regressor (GBR).**

Métriques	RMSE	MAPE	R^2	RMSE relative
Baseline : LinearRegression	56,56	126 %	0,38	N/A
Meilleur modèle : GBR	40,28	109 %	0,51	89 % moy obs.

=> Sur-apprentissage plus faible par rapport au jeu de train.

=> L'encodage OneHot des variables catégorielles améliore nettement toutes les métriques pour la variable Energie, pas pour la variable CO2

=> Variance du modèle retenu (RMSE relative) = qualité prédiction moyenne (Energie) et médiocre (CO2).

r

Récapitulatif des pistes modélisation

Choix du modèle – Variable Consommations d'énergie

Hypothèses	Hypothèse 1	Hypothèse 2	Hypothèse 3	Hypothèse 4	Amélioration Hyp 4 AME1	Amélioration Hyp 4 REL1
11 variables dont 4 nouvelles	X	X	X	X	X	X
Variables réelles	X	X	X	X	X	X
Echelle log variables cible		X	X	X	X	X
Suppression valeurs atypique			X	X	X	X
Encodage OneHot variables catégorielles (18 variables)				X	X	X
Permutation features (GBR): suppression 10 features suite hyp 4					X	
Ajout 3 variables relevés – encodage binaire						X
Resultats:	**Energie**	**Energie**	**Energie**	**Energie**	**N/A**	**N/A**
Baseline	**LR**	**LR**	LR	LR	N/A	N/A
RMSE	10 184 082	626 533 900	2 896 216	3 775 749	N/A	N/A
MAPE	0.94	0.90	0.61	0.94	N/A	N/A
R^2	0.61	0.48	0.38	0.38	N/A	N/A
Meilleur modèle	**RFR**	**GBR**	**RFR**	**GBR**	**GBR**	**GBR**
RMSE	9 303 825	17 176 150	1 328 788	1 280 229	1 486 558	1 157 556
RMSE_rel	167% val moy	274% val moy	60% val moy	56% val moy	61% val moy	53% val moy
MAPE	0.77	0.56	0.48	0.48	0.43	0.49
R^2	0.67	0.71	0.53	0.56	0.58	0.59

r

Récapitulatif des pistes de modélisation
Choix du modèle – Variable Émissions CO2

Hypothèses	Hypothèse 1	Hypothèse 2	Hypothèse 3	Hypothèse 4	Amélioration Hyp 4 AME2	Amélioration Hyp 3 REL1	Impact EnergyStar
11 variables	X	X	X	X	X	X	X
Variables réelles	X	X	X	X	X	X	X
Echelle log variables cible		X	X	X	X	X	X
Suppression valeurs atypique			X	X	X	X	X
Encodage OneHot variables catégorielles				X	X		
Permutation features (GBR): suppression 9 features suite hyp 4					X		
Ajout 3 variables relevés – encodage binaire						X	X
Ajout EnegyStar							X
Resultats:	**CO2**	**CO2**	**CO2**	**CO2**	**CO2**	**CO2**	**CO2**
Baseline	**LR**	LR	LR	LR	**N/A**	**N/A**	**N/A**
RMSE	431	1.395696e+26	44.29	56.56	**N/A**	**N/A**	**N/A**
MAPE	3.64	4.906367e+20	1.21	1.26	**N/A**	**N/A**	**N/A**
R^2	0.36	-1.73	0.25	0.38	**N/A**	**N/A**	**N/A**
Meilleur modèle	**XGR**	**GBR**	**RFR**	**GBR**	**GBR**	**RFR**	**RFR**
RMSE	323	380	36.9	40.28	39.73	36.72	33.88
RMSE_rel	256% val moy	301% val moy	87% val moy	89% val moy	90% val moy	80% val moy	72% val moy
MAPE	2.07	1.2	1.01	1.09	1.17	0.65	0.67
R^2	0.64	0.59	0.44	0.51	0.41	0.68	0.73

r

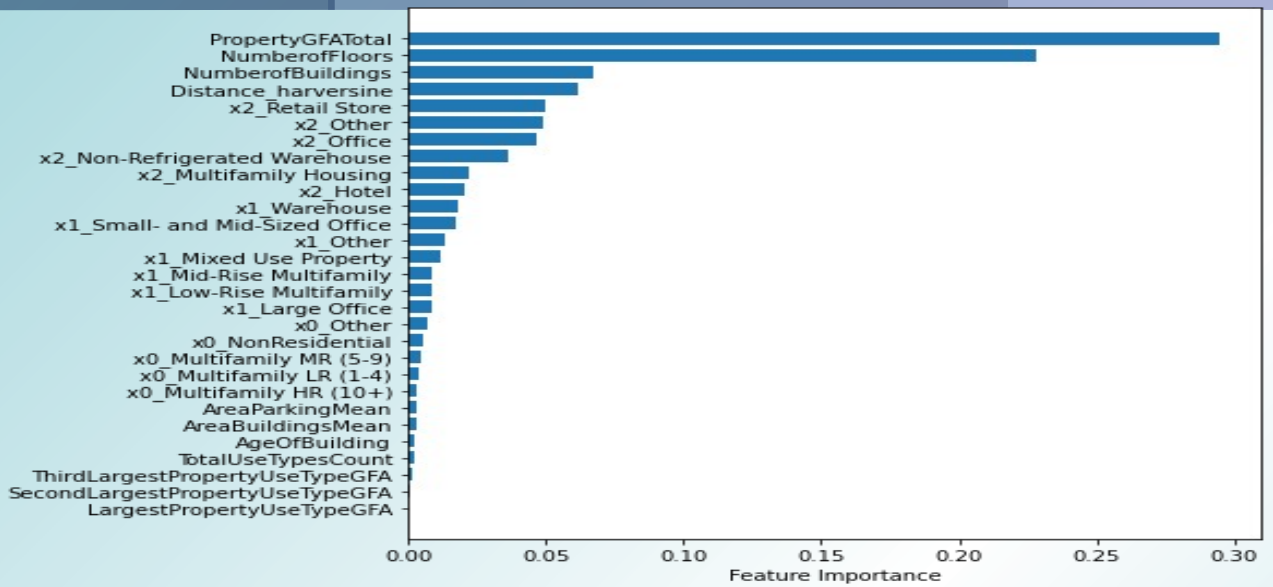
Choix et amélioration du modèle final

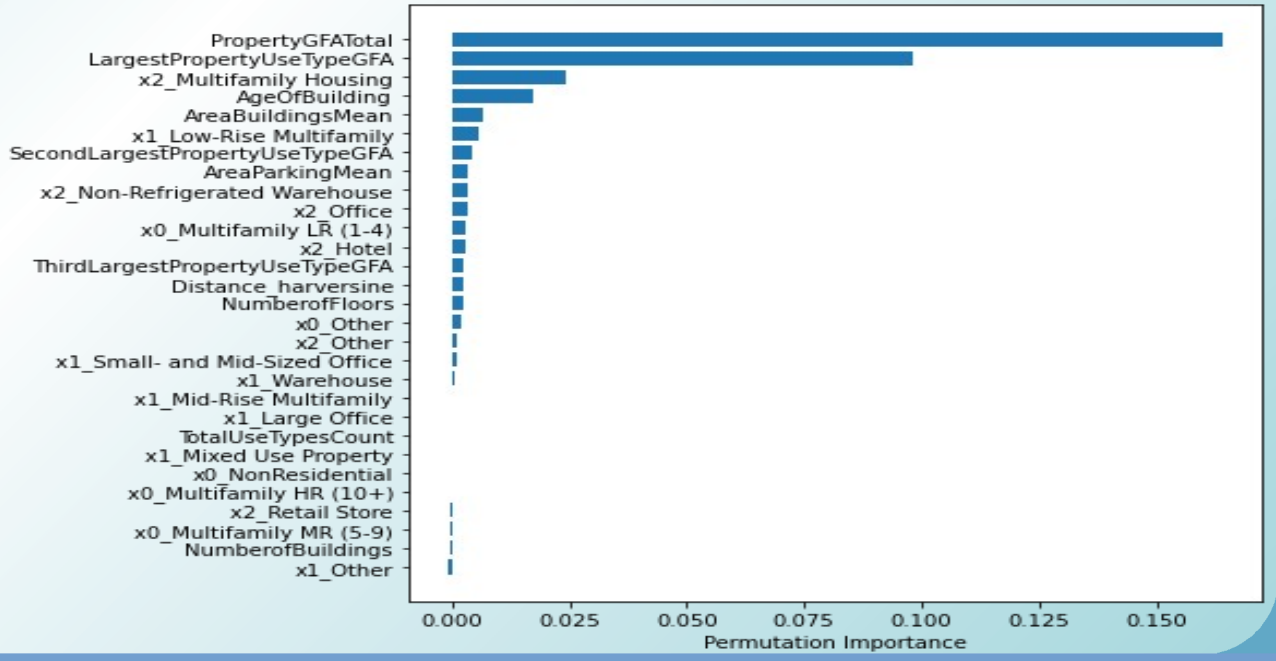
- Suite aux différentes piste de modélisation, 2 modèles sont retenus :
  - **Modèle Gradient Boosting Regressor (GBR) avec 29 variables en entrée pour la variable Consommation d'énergie (hypothèse 4)**
  - **Modèle Random Forest Regressor (RFR) avec 11 variables en entrée pour la variable Emissions de CO2 (hypothèse 3) .**

- Améliorations des modèles finaux :
  - Suppression de features suite permutation de l'importance des features sur modèle GBR de l'hypothèse 4 pour les variables Energie et CO2.
  - Ajout de 3 variables indiquant l'utilisation ou pas d'un type d'énergie par bâtiment pour modèle GBR de l'hypothèse 4 (Energie), pour modèle RFR de l'hypothèse 3 (CO2).

Features / Permutation importance

Modèle GBR Hypothèse 4 – Variable « Consommation d'énergie »

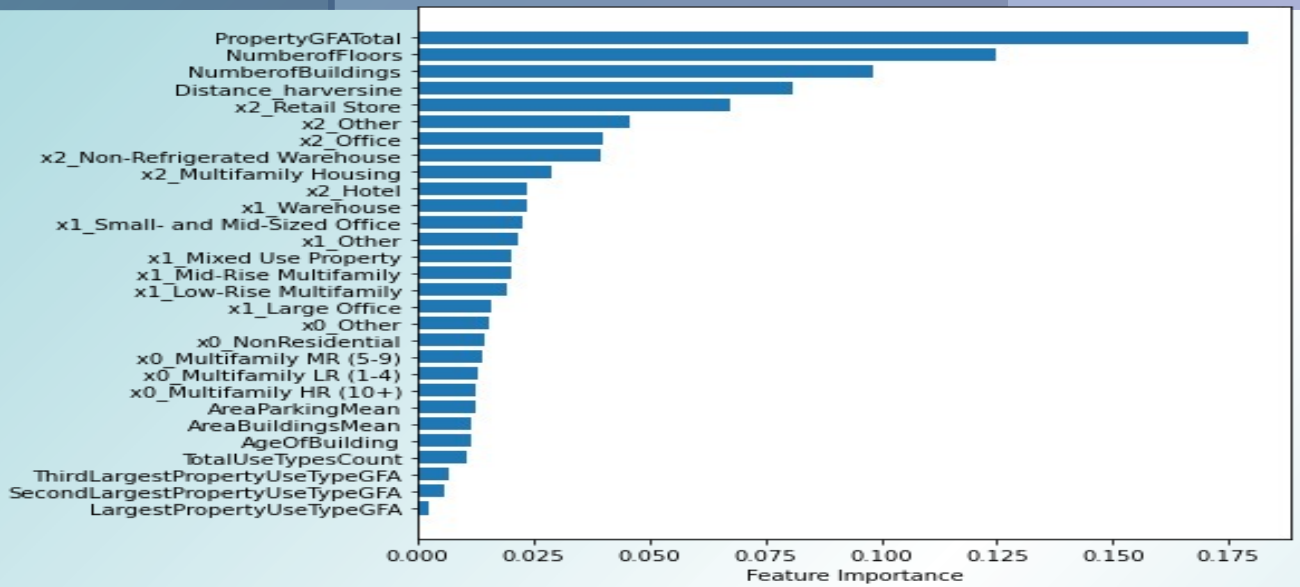


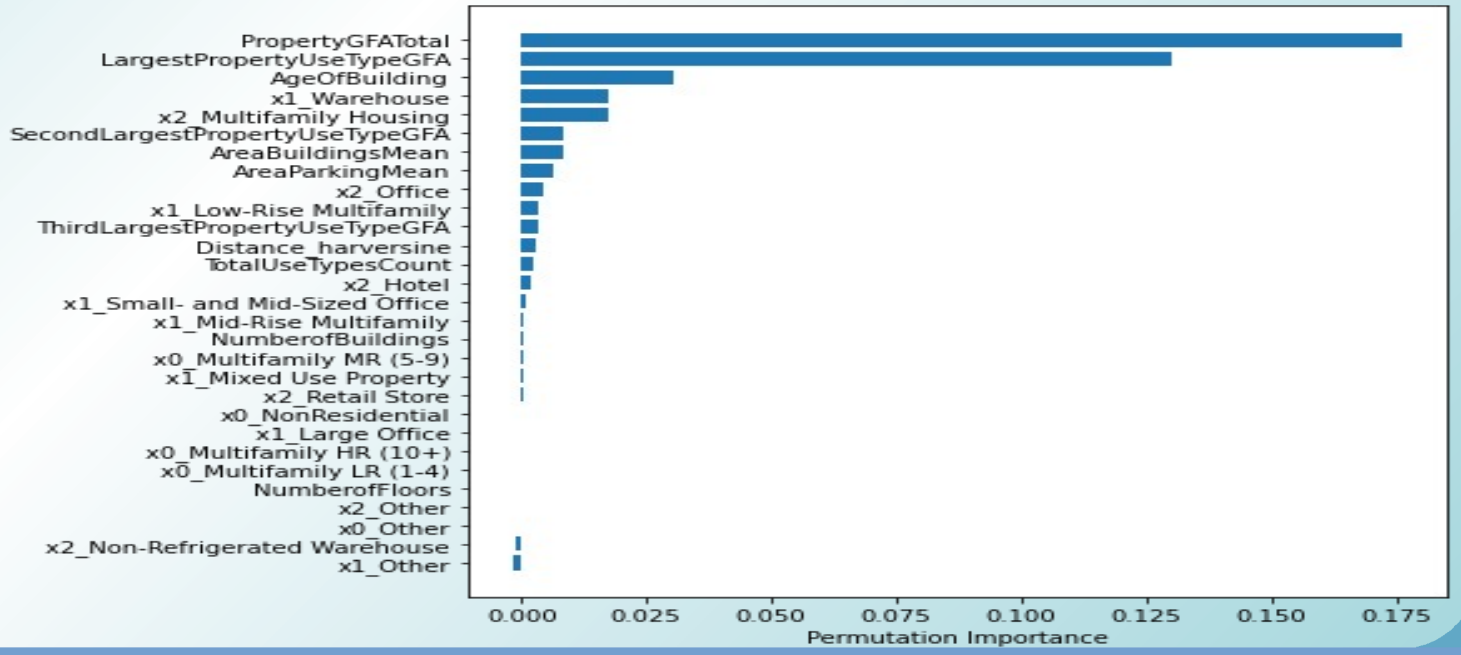


r

Features / Permutation importance

Modèle GBR Hypothèse 4 – Variable « Emissions CO2 »





r

Résultats de la suppression des features

- Suppression des features avec un score négatif ou égal à 0.

- Variable « Consommation d'énergie » :
  - Suppression de 10 variables
  - RMSE (1 486 558) dégradé par rapport au modèle GBR de l'hypothèse 4, mais R^2 (0,58) et MAPE (43%) améliorés.

  => Amélioration non retenue.

- Variable « Emissions CO2 » :
  - Suppression de 9 variables
  - RMSE (39,73) un peu amélioré par rapport au modèle GBR de l'hypothèse 4, mais R^2 (0,41) et MAPE (117 %) dégradés.

  => Amélioration non retenue.

Résultats ajout 3 variables relevés par encodage binaire

- Encodage binaire (0=absence, 1=présence) pour indiquer l'utilisation ou pas du type d'énergie par la bâtiment (variables « Electricity », « NaturalGas », « SteamUse »).

- Variable « Consommation d'énergie » :
  - Modèle avec 32 variables
  - RMSE (1 157 556) sensiblement amélioré par rapport au modèle GBR de l'hypothèse 4, idem pour R^2 (0,59) et RMSE relative (53%) améliorés.

  => Amélioration retenue, meilleur modèle pour la variable Energie.

- Variable « Emissions CO2 » :
  - Modèle avec 14 variables
  - RMSE (36,72) un peu amélioré par rapport au modèle RFR de l'hypothèse 3, R^2 (0,68) et MAPE (65 %) fortement améliorés.

  => Amélioration retenue ,meilleur modèle pour la variable CO2

Impact EnergyStar score sur émissions CO2

- Ajout de la variable « EnergyStar » à partir d'un modèle basé sur Random Forest Regressor (11 variables + 3 variables sur présence type énergie), soit 15 variables.

- Imputation des valeurs manquantes par la moyenne des valeurs de l'EnergyStar.

- Toutes les métriques sont améliorées permettant de fournir un meilleur modèle:
  - **RMSE = 33,38, RMSE relative = 72 %**
  - **MAPE = 67 % (taux moyen d'erreur sur le jeu de test).**
  - **R^2 = 0,73**

- Performance médiocres sur modèles sur les prédictions avec une RMSE normalisée correspondant à à 72% de la moyenne des observations sur le jeu de test.

Conclusion sur les modèles finaux

- L'analyse des features montre l'importance des variables surface (PropertyGFATotal) et de localisation (Distance_harversine) pour développer des modèles performants.
- L'impact du score EnergyStar améliore les performances du modèle sur la variable Emissions de CO2 sur toutes les métriques.
- Qualité moyenne (RMSE relative) pour les prédictions Consommations d'énergie.
- Qualité médiocre (RMSE relative) pour les prédictions Emissions CO2.
- Difficulté pour évaluer les performances avec un volume de données trop faible, pour généraliser un modèle.

Annexes

Annexe 1 : Réglage hyper-paramètres ElasticNet

Annexe 2 : Réglage hyper-paramètres SVR

Annexe 3 : réglage hyper-paramètres RFR

Annexe 4 : réglages hyper-paramètres GBR

Annexe 5 : réglages hyper-paramètres XGBoost

Annexe 1 : Réglage hyper-paramètres ElasticNet

- Liste des valeurs des hyper-paramètres à régler :
  - 'alpha': [0.0001, 0.001, 0.01, 0.1, 1, 10, 100]
  - 'l1_ratio': np.arange(0.0, 1.0, 0.1)
  - 'tol': [0.1, 0.01, 0.001, 0.0001]

- Signification des hyper-paramètres :
  - alpha : multiplication du terme de pénalité
  - l1_ratio : ratio entre pénalité L1 et L2
  - tol : tolérance pour l'optimisation

- ElasticNet incorpore les 2 pénalités (régularisation L2 du modèle Ridge et régularisation L2 du modèle Lasso).

Annexe 2 : Réglage hyper-paramètres SVR

- Liste des valeurs des hyper-paramètres à régler :
  - 'C': [0.001, 0.01, 0.1, 1, 10]
  - 'epsilon': [0.001, 0.01, 0.1, 1]
  - 'gamma': [1e-8, 1e-7, 1e-6, 1e-5,1e-4, 1e-3, 1e-2, 1e-1]

- Signification des hyper-paramètres :
  - C : paramètre de régularisation (importance relative du terme d'erreur et du terme de marge)
  - epsilon : largeur de la zone d'indécision
  - gamma : coefficient du noyau pour rbf et poly

Annexe 3 : Réglage hyper-paramètres RFR Random Forest Regressor

- Liste des valeurs des hyper-paramètres à régler :
  - 'bootstrap' : [True]
  - 'max_depth' : [4, 6, 8, 10]
  - 'max_features' : ['log2', 'sqrt', 'auto']
  - 'n_estimators' : [500]

- Signification des hyper-paramètres :
  - bootstrap : indique si on prend l'intégralité de données ou non
  - max_depth : profondeur maximale de l'arbre de décision
  - max_features : nombre de variables à prendre en compte pour le feature sampling.
  - n_estimators : nombre d'arbres dans la fôret.

Annexe 4 : Réglage hyper-paramètres GBR Gradient Boosting Regressor

- Liste des valeurs des hyper-paramètres à régler :
  - 'learning_rate': [0.2, 0.4, 0.7]
  - 'max_depth' : [4, 6, 8, 10]
  - 'loss' : ['ls','lad','huber']

- Signification des hyper-paramètres :
  - learning_rate : taux d'apprentissage indiquant la contribution à chaque arbre.
  - max_depth : profondeur maximale des estimateurs de régression individuels
  - loss : fonction de perte à optimiser.

Annexe 5 : Réglage hyper-paramètres XGR XGBoost

- Liste des valeurs des hyper-paramètres à régler :
  - 'n_estimators' : [100, 500, 1000, 2000]
  - 'max_depth'   : [4, 6, 10]

- Signification des hyper-paramètres :
  - n_estimators : nombre d'arbres séquentiels pour corriger les arbres précédents.
  - max_depth : détermine à quelle profondeur chaque arbre est autorisé à pousser pendant n'importe quel tour de boost.