PROJET 4 – « ANTICIPEZ LES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS »

Soutenance de projet

Décembre 2019

Sommaire

I - PROBLÉMATIQUE

Rappel de la problématique

Interprétation

Pistes de recherche envisagées

Présentation de la problématique

Seattle

- Données de consommation disponibles pour les bâtiments de la ville de Seattle pour les années 2015 et 2016
- Coût important d'obtention des relevés / fastidieuses à collecter

- 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 modèle de prédiction réutilisable

Interprétation de la problématique

Seattle

- Prévision
  - Features: caractéristiques intrinsèques des bâtiments (hors consommations)
  - Données à prédire
    - Consommation totale des bâtiments *SiteEnergyUseWN(kBtu)*
    - Emissions totales des bâtiments *TotalGHGEmissions*

      => 2 modèles différents

- ENERGY STAR Score :
  - Comparaison de son intérêt en essayant de modéliser avec et sans

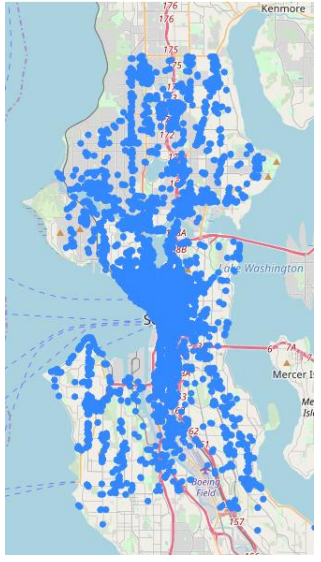
II – PRÉPARATION DU JEU DE DONNÉES

Cleaning

Feature engineering

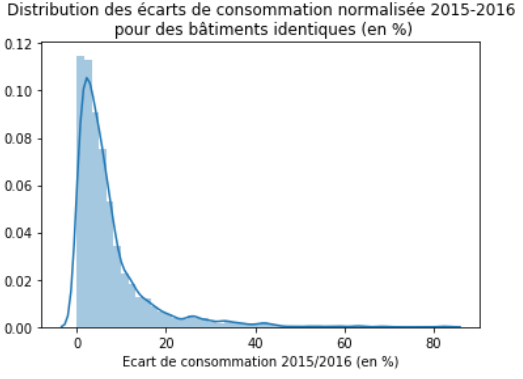
Exploration

Cleaning



Défauts » du jeu de données initial:

- Données 2015 et 2016 non alignées : (colonnes différentes, informations réparties différemment)
- Casse de certaines colonnes / informations quasi-identiques
- NaN :
  - complétion des valeurs manquantes quand applicable (e.g. catégories « unknown »)
  - Suppression des observations pour lesquelles on a beaucoup de NaN pour conserver un maximum de features
- Suppression des outliers :
  - Outliers univariés (1% extreme)
  - Outliers multivariés (distance aux 5 plus proches voisins / 1 % extreme)
  - Ecart de consommation entre 2015 et 2016 pour les mêmes bâtiments : $\mu + 3\sigma$



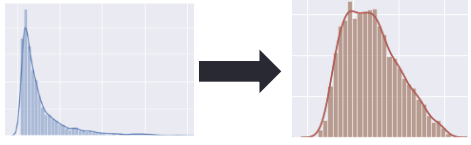
Feature engineering

Idées écartées

- Features liées à la proportion des sources d'énergie (coûteux à obtenir pour futures données)
- Utilisation du Energy Star score (mis de côté pour analyse ultérieure)

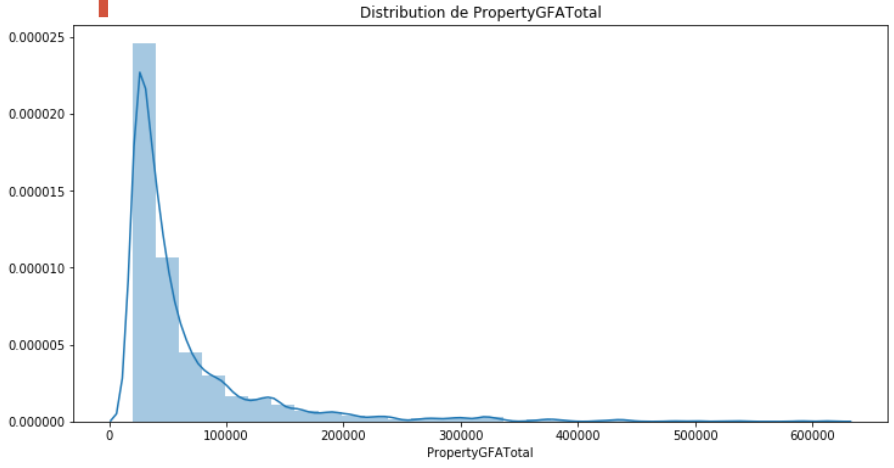
Idées retenues

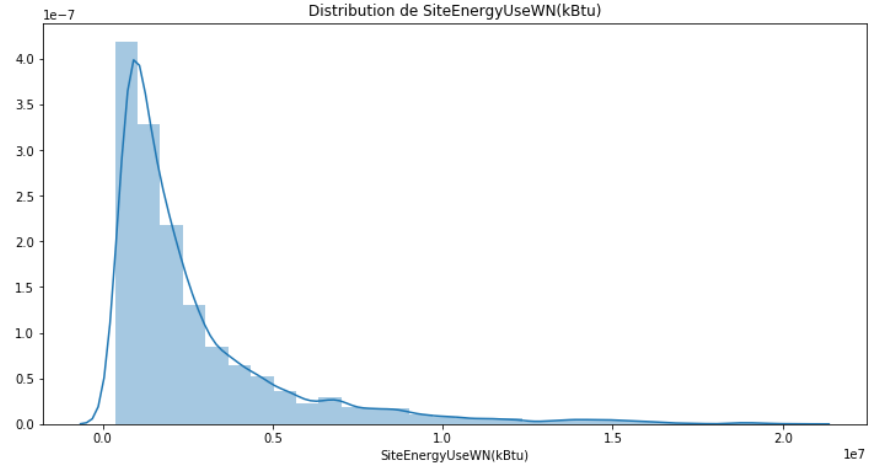
- Suppression des features de consommation (ormis les 2 features qu'on cherche à prédire)
- Catégorisation des données pour certaines colonnes (usage)
- One Hot Encoding : Transformation d'une feature avec n catégories en n features booléennes.
- Suppression de colonnes non pertinentes pour notre modèle
  - Données sans catégorisation possible (Comment)
  - Données avec une unique information (exemple : State)
  - Données sans information pertinente pour le modèle (voir exemples)
    - DefaultData : sens de la feature non expliqué + booléen avec beaucoup de NaN
    - SPD Beats : informations non utiles à la problématique + beaucoup de NaN
    - Features redondantes (address / zipcode remplacées par latitude et longitude)
- Log2-transformation variable de prédiction

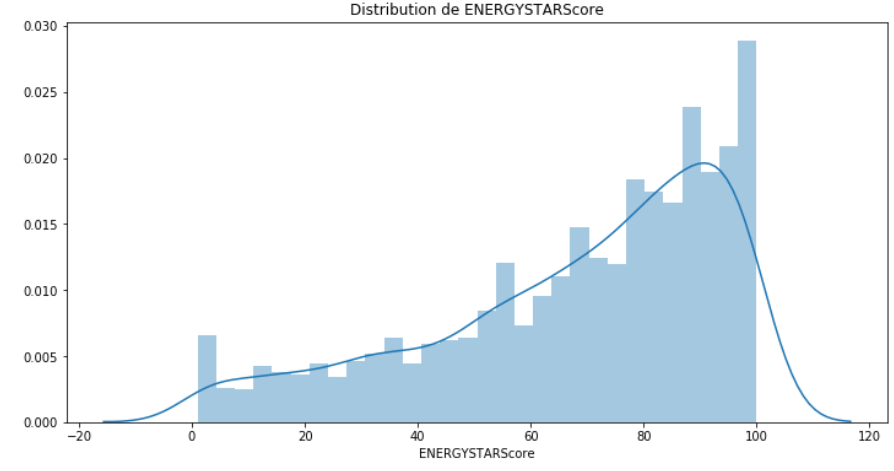


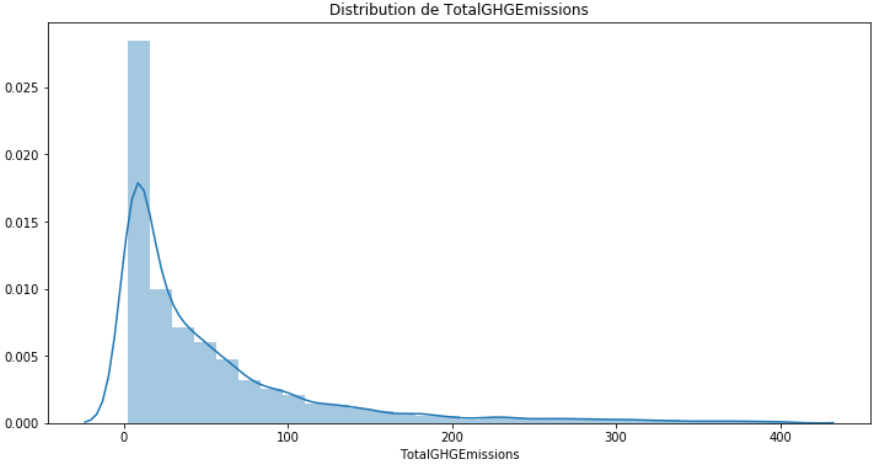
Jeu final *sans Energy Star Score*	• 5748 lignes • 24 colonnes (3 variables de prédiction)
Jeu final *avec Energy Star Score*	• 4289 lignes • 25 colonnes (3 variables de prédiction)

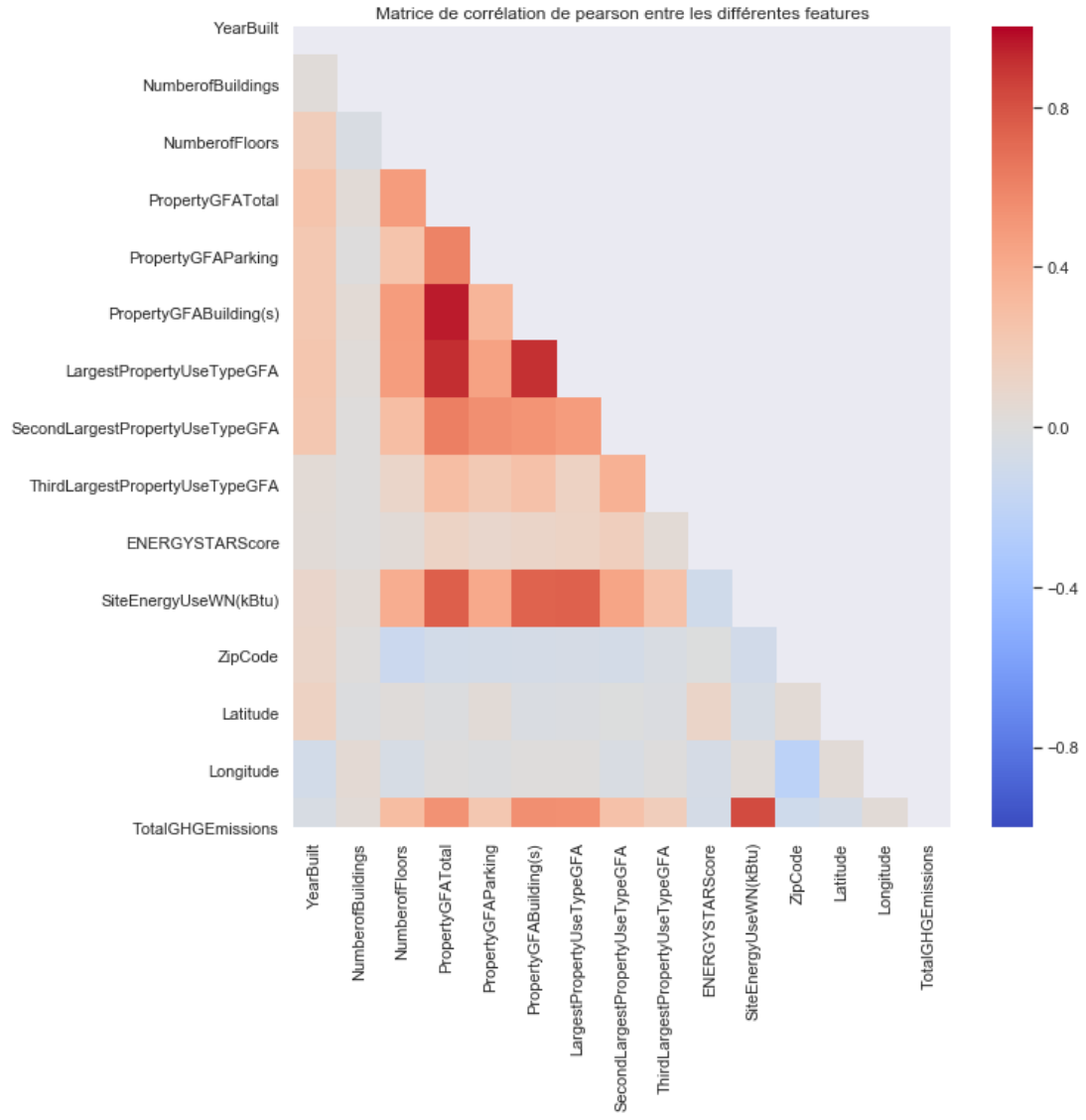
Exploration











Exploration : Corrélations

Points Majeurs:

<u>Consommation:</u> Corrélation importante de la avec:

- PropertyGFATotal,
- PropertyGFABuilding,
- LargestPropertyUseTypeFGA

<u>Emissions:</u> Mêmes corrélations (dans moindre mesure) + corrélation importante avec la consommation

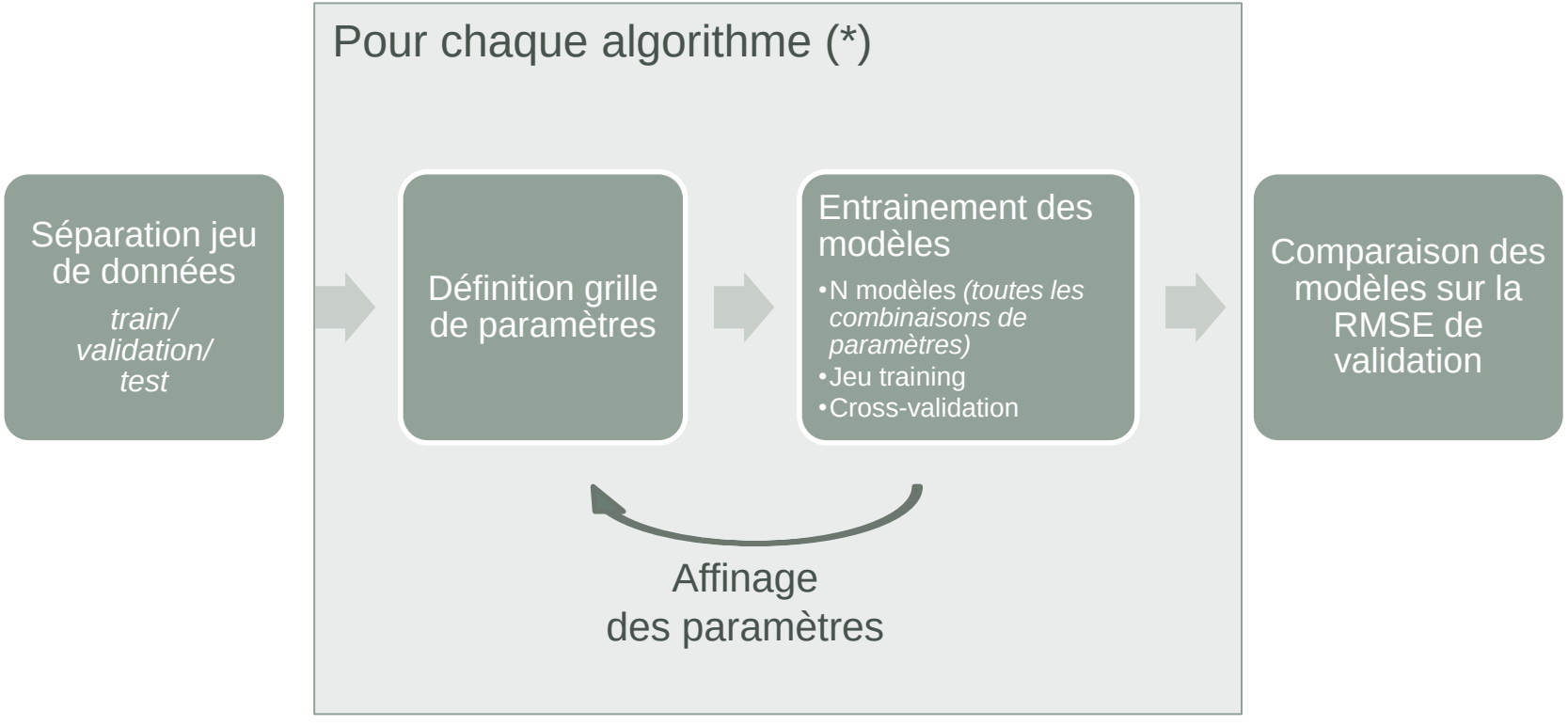
Autres points notables:

- Corrélation importante entre
  - PropertyGFATotal et PropertyGFABuildings
  - PropertyFGATotal et LargestPropertyUseTypeGFA
  - LargestPropertyUseTypeGFA et PropertyFGABuilding(s)

- Energy Star Score : pas de corrélation notable

III – PISTES DE MODÉLISATIONS

Modèle consommation : démarche



(*) Modèles entraînés : Elastic Net / SVR / Random Forest Regressor / XGBoost

Modèle consommation : paramètres

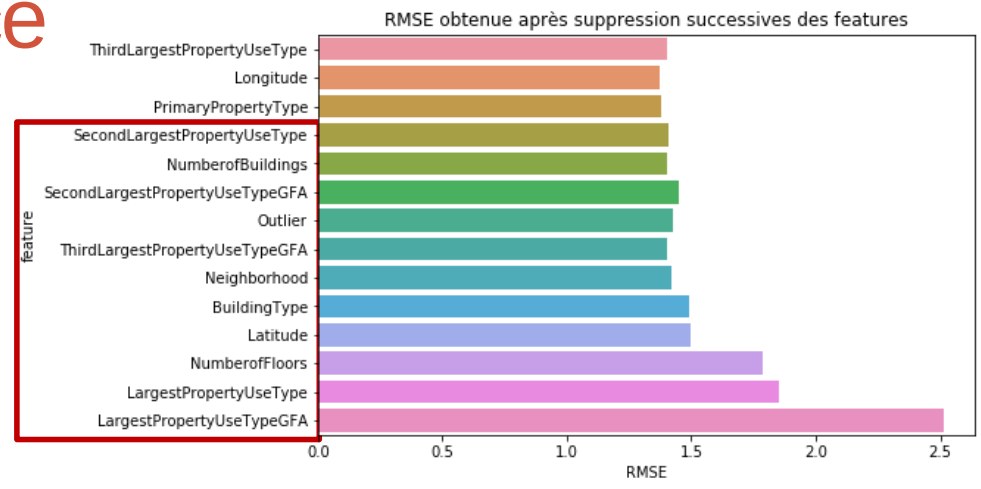
Elastic Net	SVR	XGBoost	Random Forest Regressor
Alpha : [10^{-4}, 10^{-3}, ..., 10, 10^{2}]	Gamma : 10^{-8},10^{-7} , ... , 10^{-1}	N_estimators : [100, 500, 1000, 2000]	N_estimators : [10, 50, 100, 300, 500]
L1_ratio : [0.1, 0.2, 0.3 ...0.6.... 0.9	Epsilon : [0.001, 0.01, 0.1, 1]		Min_samples_leaf : [1, 3, 5, 10]
Tol : [0.1,0.01,0.001,0.0001]	C : [0.001, 0.01, 0.1, 1, 10]		Max_features : [auto, sqrt]

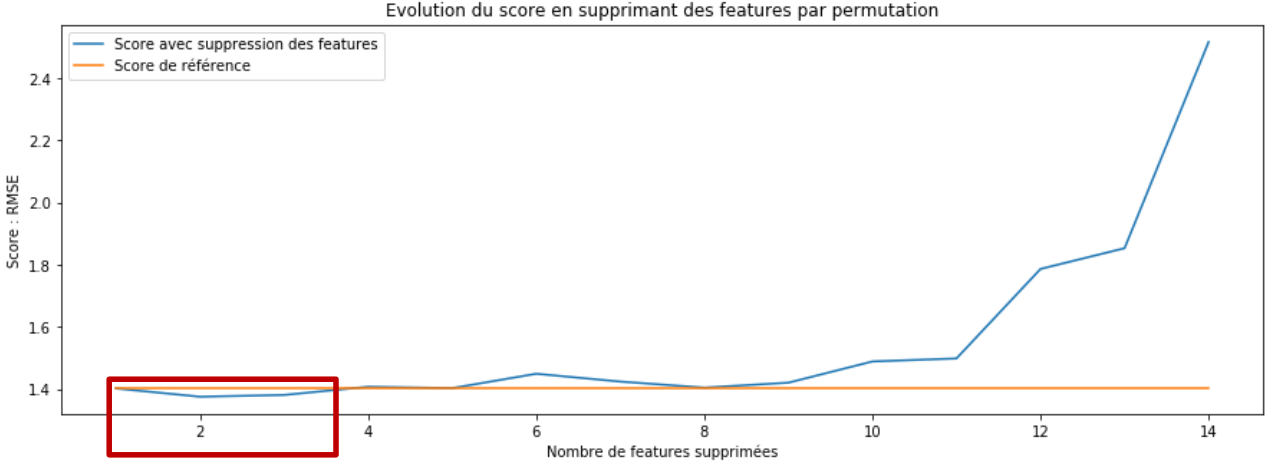
Combinaison optimale des paramètres

Complément : Pertinence des variables par permutation

Boucle :

- Fit modèle avec ensemble des features
- Permutation aléatoire d'une feature (ou bloc de feature pré OHE)
- Calcul du score
- Suppression de la feature qui dégrade le moins le score
- Calcul score





Modèle émissions : démarche

Idée: Faire un modèle simplifié à partir de la prédiction de consommation

Entraînement d'un modèle « base »

- Toutes les features
- Random Forest (optimisation paramètres avec grid search CV)
- **RMSE = 103** (jeu de test)

Entraînement d'un modèle simplifié:

- Feature unique : consommation prédite
- Random Forest (optimisation paramètres avec grid search CV)
- **RMSE : 103** (jeu de validation)

Entraînement d'un 2e modèle simplfié:

- Features:consommation prédite + 5 features majeures
- (optimisation paramètres avec grid search CV)
- **RMSE : 84** (jeu de validation)

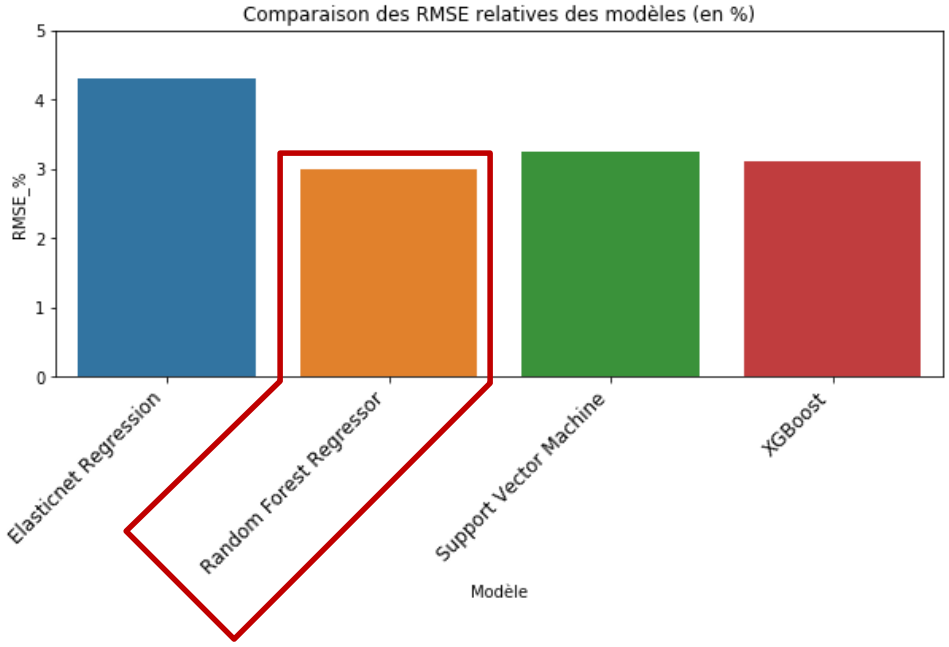
- Le modèle obtenu est encore plus performant que le modèle étalon et peut être retenu.

IV –PRÉSENTATION DU MODÈLE FINAL

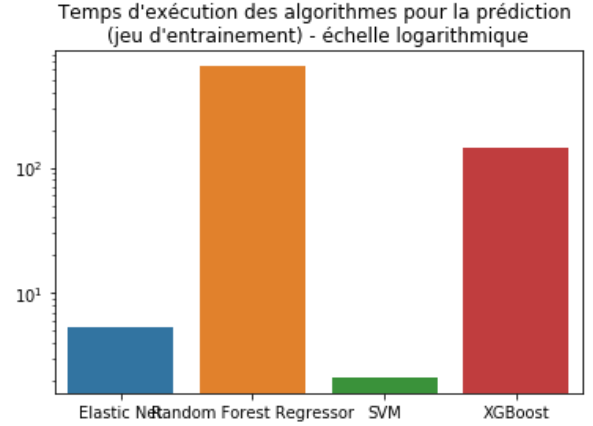
ainsi que des améliorations effectuées.

Modèles obtenus (consommation)

Comparaison sur jeu de test



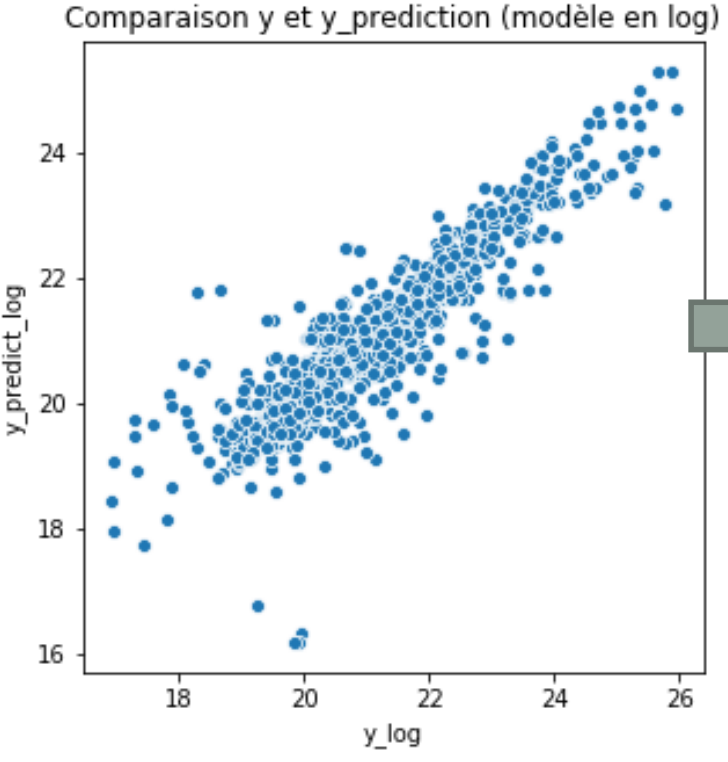
	Modèle	Score_RMSE	RMSE_%
0	Elasticnet Regression	0.906957	0.043171
1	Random Forest Regressor	0.630444	0.030009
2	Support Vector Machine	0.685175	0.032614
3	XGBoost	0.653062	0.031086



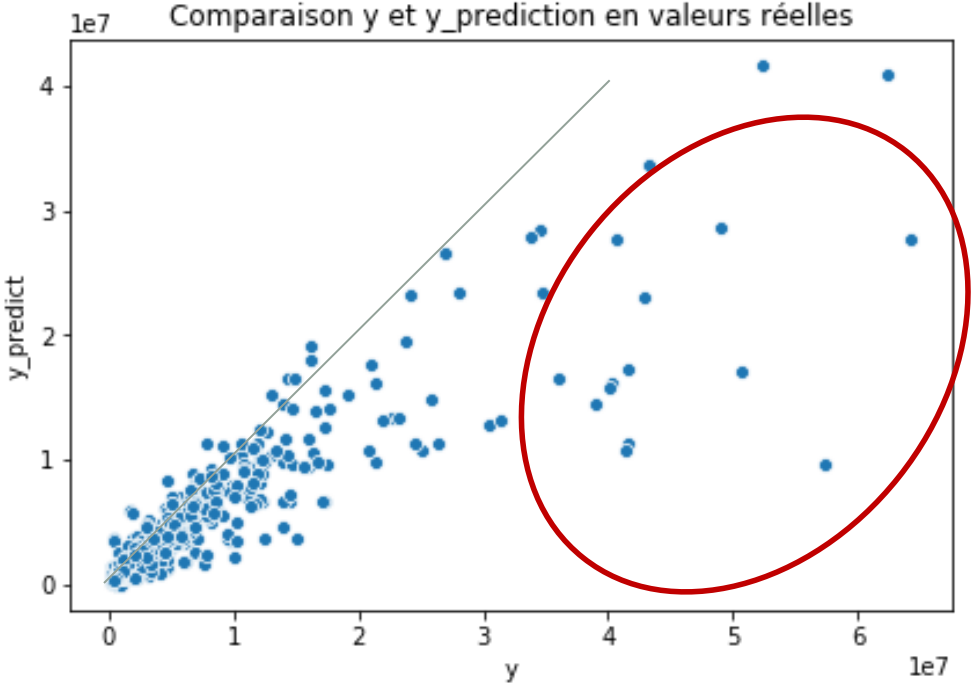
Pour comparaison : un estimateur donnant comme prédiction la moyenne donne une RMSE de 1,50

Modèle final :

- Prédiction et limites



exp



Intérêt du ENERGY STAR Score

- Feature traitée à part du modèle initial (moins de données disponibles)
- Entraînement d'un modèle Random Forest Regressor (grid search CV)
- RMSE obtenue sur jeu de test : 1,14 < 1,26 Améliore très légèrement la performance du modèle

- Arbitrage à réaliser :
  - fastidieux à calculer / complexité
  - améliore la performance faiblement

Complément : Modèle d'ensemble

- Entrainement d'un modèle Ridge

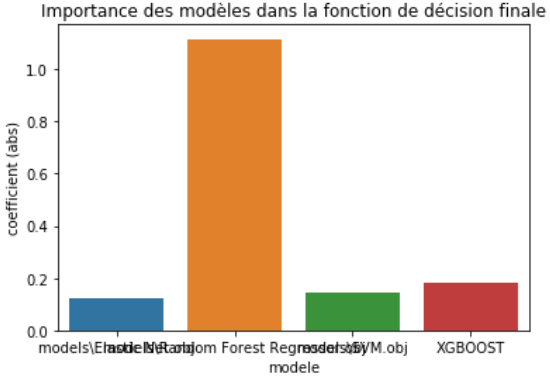
Modèle Elastic Net

Modèle SVR

Modèle Random Forest Regressor

Modèle XGBOOST

Ridge Regression



Prédiction finale

RMSE_test= 0,6300
(< 0,6304)

MERCI DE VOTRE ATTENTION