Projet 4 OC :

"Anticipez les besoins en consommation électrique de bâtiments"

- Problématique
- Données
- Modélisation
- Conclusion

Problématique

Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, la ville de Seattle s'intéresse de près aux émissions des bâtiments non destinés à l'habitation.

- Prédire les émissions de CO2 et la consommation totale d'énergie sans nouveaux relevés annuels (onéreux)
- Evaluer l'intérêt de l'ENERGY STAR Score
- Mettre en place un modèle de prédiction réutilisable

Données du dataset

Identification	Infos liées aux données	Localisation	Usage et construction (catégorielles)	Usage et construction (quantitatives)	Relevés énergétiques émissions
OSEBuildingID PropertyName TaxParcelIdentificationNumber	DataYear DefaultData Comments ComplianceStatus Outlier	CouncilDistrictCode Neighborhood ZipCode Latitude Longitude Address	BuildingType PrimaryPropertyType YearBuilt ListOfAllPropertyUseTypes LargestPropertyUseType SecondLargestPropertyUseType ThirdLargestPropertyUseType	NumberofBuildings NumberofFloors PropertyGFATotal PropertyGFAParking PropertyGFABuilding(s) LargestPropertyUseTypeGFA SecondLargestPropertyUseTypeGFA ThirdLargestPropertyUseTypeGFA	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

Données brute du dataset

- Données de la ville de Seattle
- Deux dataset : pour les années 2015 et 2016
- Environ 3 400 bâtiments dans chaque table avec 46 variables
- Les fichiers .csv pèsent 1,5 Mo + 1,2 Mo
- Données 2015 et 2016 non alignées :
- Les deux datasets identifient par ligne un batiment unique avec leurs caractéristiques techniques

Colonnes différentes, informations réparties différemment, NaN, donnée dupliqué

Données nettoyé du dataset

Renommage des variables communes et mal nommées

Supprimer les variables non communes aux 2 datasets

Aplatissement des variables de localisation pour une concatenation parfaite

Suppression des variables inutiles (id, info, unique et localisation)

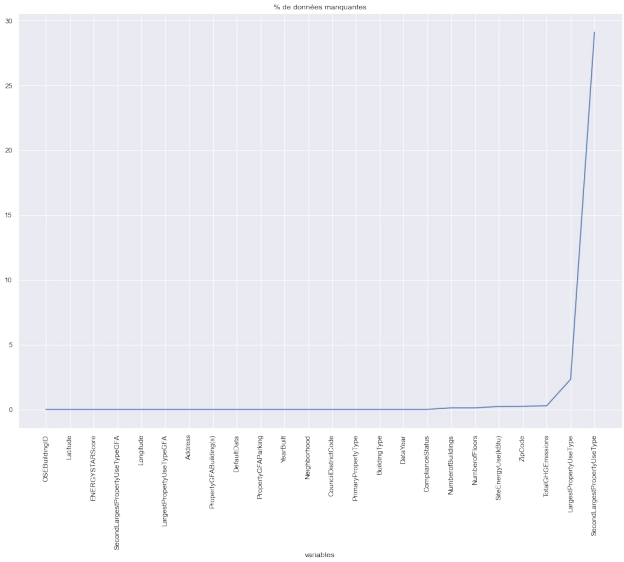
Remplissage des valeurs manquantes en utilisant une année différente ou au cas par cas (en fonction des distributions et boxplots)

Valeurs aberrantes pour chaque variables

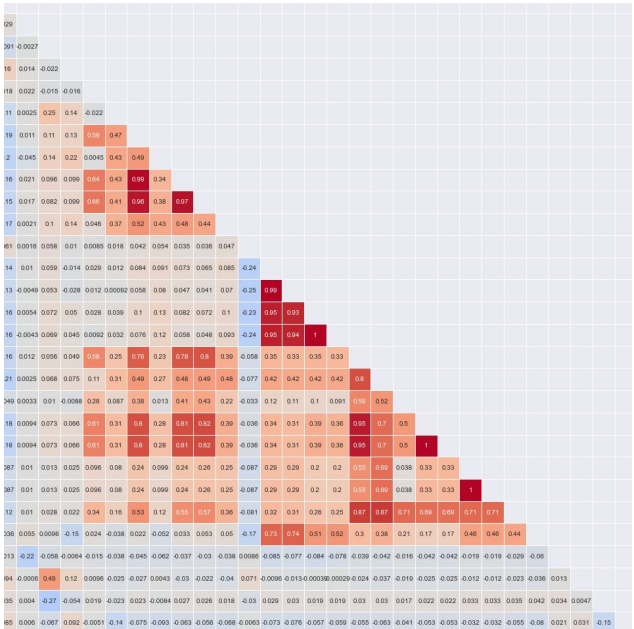
Concaténation finale

Suppression des colonnes 100% NaN et > 80% ou imputation médiane / mode

Suppression outliers (IQR * 1.5) et valeurs négatives pour les targets



Feature enginerring



Features corrélées à supprimer

"SiteEUIWN(kBtu/sf)"
"SiteEUI(kBtu/sf)"
"SourceEUI(kBtu/sf)"
"SourceEUIWN(kBtu/sf)"
"PropertyGFATotal"
"SiteEnergyUseWN(kBtu)"
"Electricity(kWh)"
"NaturalGas(therms)"
"GHGEmissionsIntensity"
"Electricity(kBtu)"
"SteamUse(kBtu)"
"NaturalGas(kBtu)"

Se résume avec des variables synthétiques

“SiteEnergyUse(kBtu)”
“TotalGHGEmissions”
“PropertyGFABuilding(s)”
“LargestPropertyUseTypeGFA”
“LargestPropertyUseType”

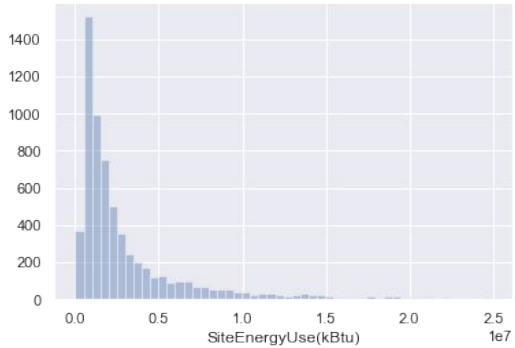
Feature enginerring et variables cibles

Suppression des variables de source d'énergie secondaire et de sous-multiple d'unité de mesure

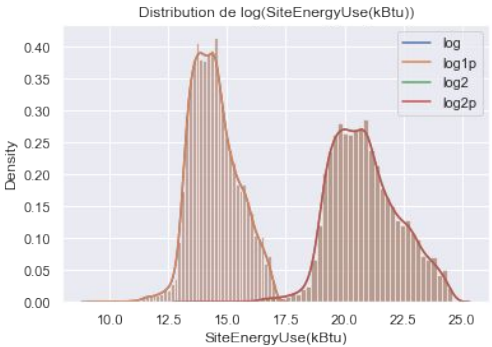
2 variables qui ne dépendent pas des caractéristiques intrinsèques du bâtiment. Les distributions des variables numériques sont globalement aplaties

SiteEnergyUse(kBtu) : "The annual amount of energy consumed by the property from all sources of energy"

TotalGHGEmissions : "The total amount of greenhouse gas emissions, including carbon dioxide, methane etc."



Log



Amélioration du biais

Modélisation

Tester et à évaluer plusieurs régresseurs de prédiction des variables SiteEnergyUse(kBtu) et TotalGHGEmissions

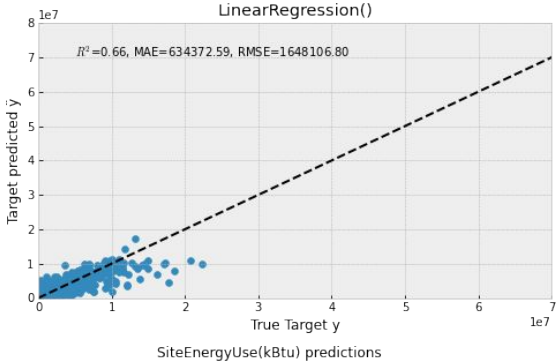
- Définition des variables explicatives et à prédire.
- Séparation variable numérique et catégorielle.
- Définition du pré traitement différencié selon les colonnes
- Split en jeu d'entraînement et de test
- Début du Pipeline des données :
- Imputation
- Target encoding
- Standardisation
- Définition des régresseurs

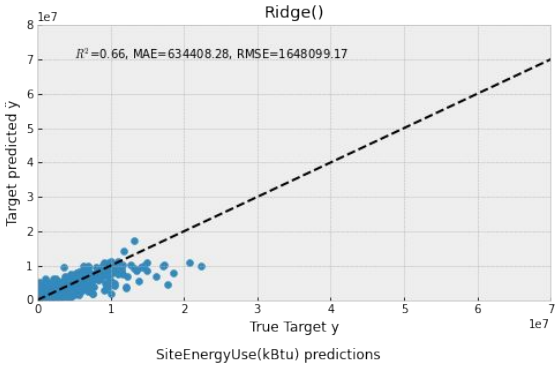
Régresseurs

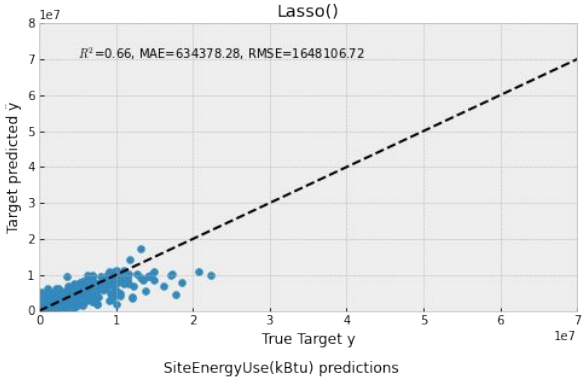
```
'Dummy': DummyRegressor(),
'Linear': LinearRegression(),
'Ridge' : Ridge(),
'Lasso' : Lasso(),
'Elastic Net': ElasticNet(),
'SVR': SVR(kernel="rbf", C=300, gamma=1),
'Random Forest': RandomForestRegressor(n_estimators=100, n_jobs=-1),
'Extra Tree': ExtraTreesRegressor(n_estimators=100, n_jobs=-1),
'Gradient Boosting': GradientBoostingRegressor(n_estimators=100, random_state = 42),
```

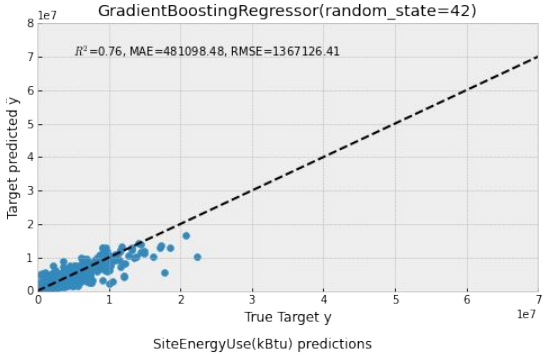
Baselines : Recherche du meilleur régresseur

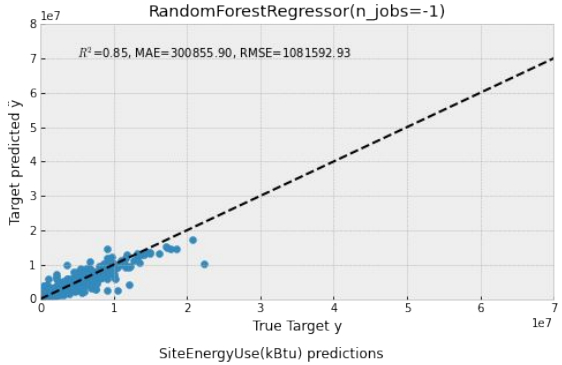
Courbe prédiction/réelle Y et utilisation le coefficient de détermination R^2 et du RMSE

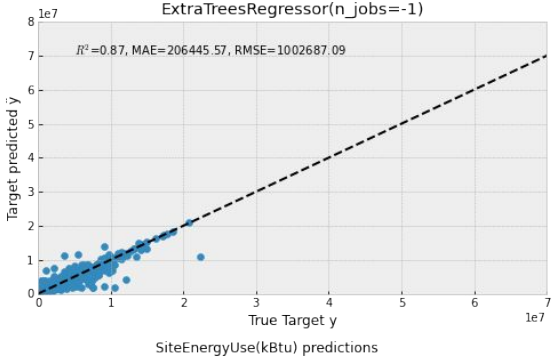








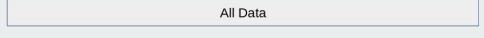


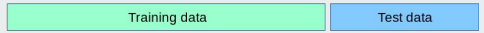


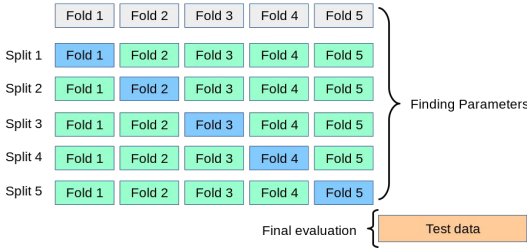
Cross-validation

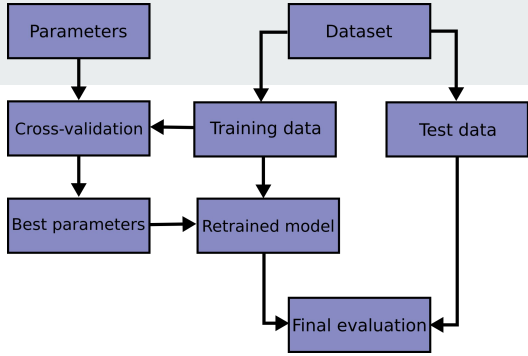
9.27e-01 9.27e-01

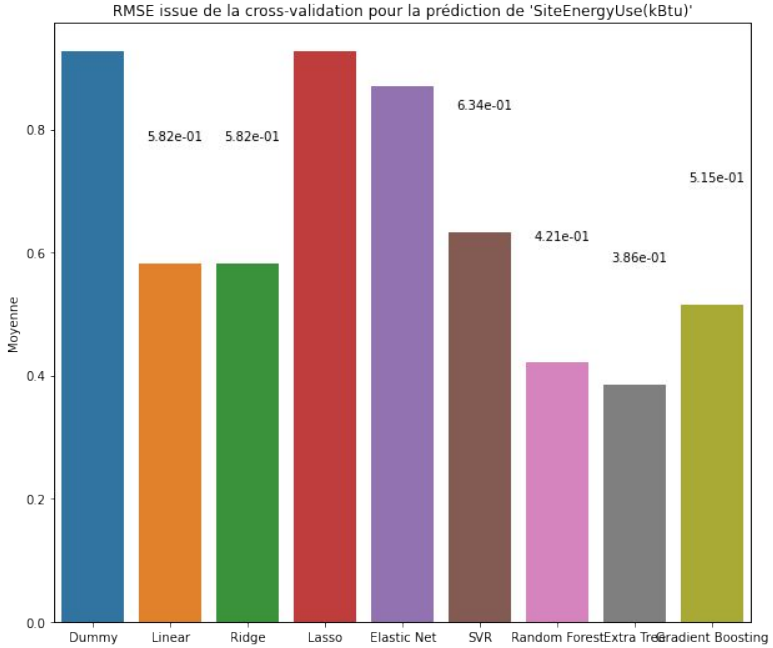
8.70e-01

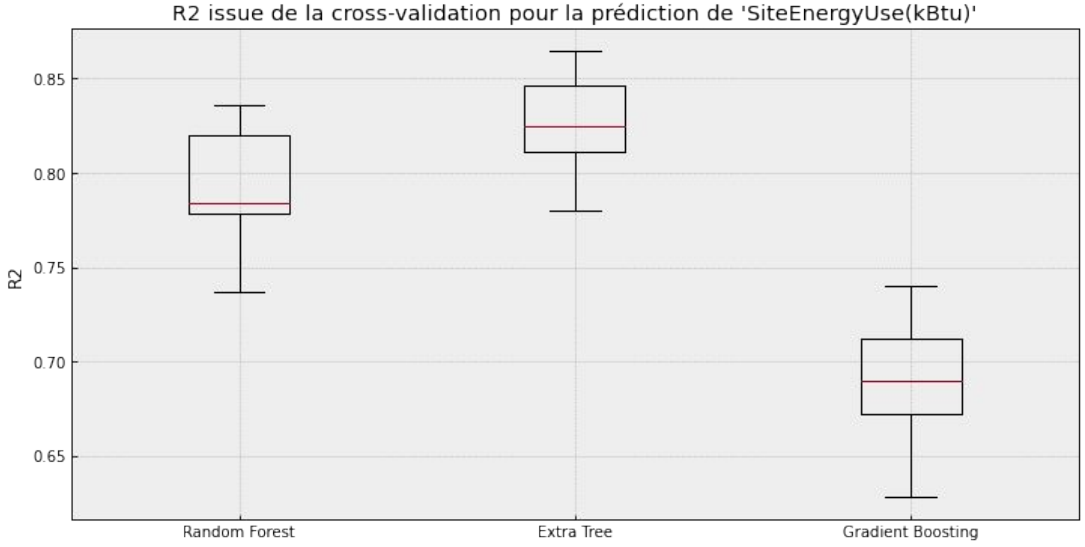












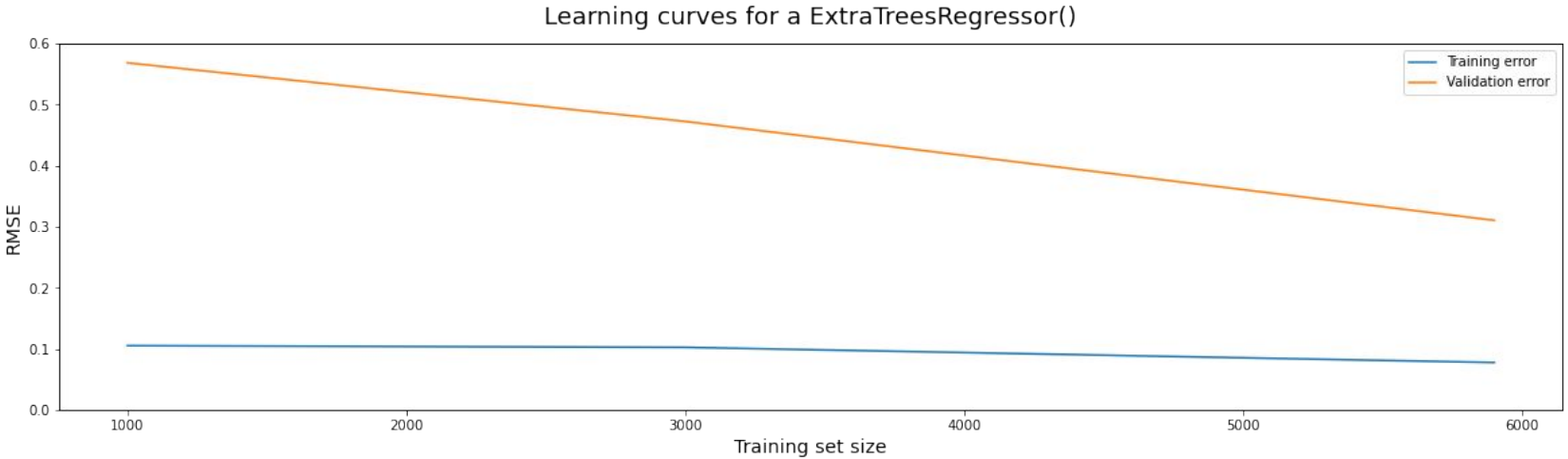
Hyperparamètres et grid search pour Extra tree

Hyperparamètres du meilleur modèle : {'preprocessor__cat__imputer__strategy': 'mean', 'preprocessor__num__imputer__strategy': 'most_frequent', 'regressor__max_depth': None, 'regressor__max_features': 'auto', 'regressor__min_samples_leaf': 1, 'regressor__min_samples_split': 4, 'regressor__n_estimators': 500}

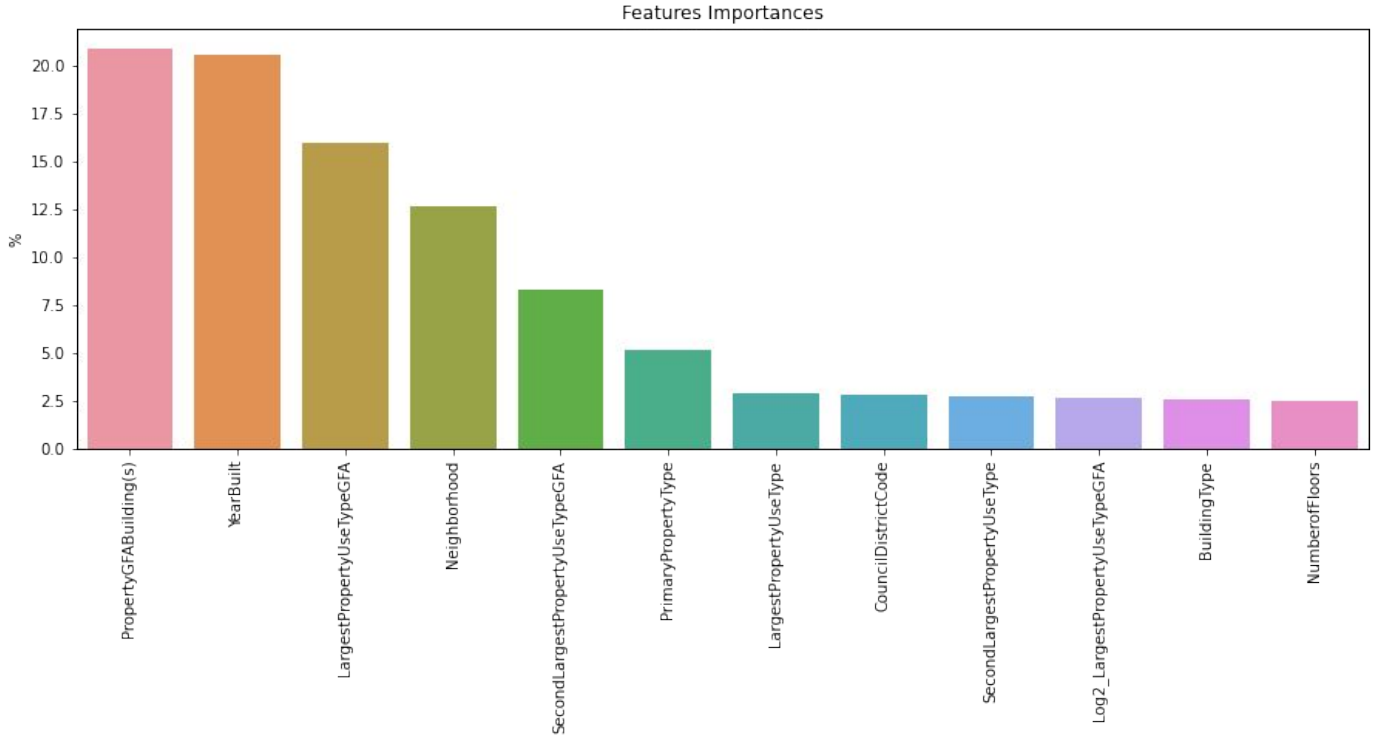
___ SCORES :___

R2 score sur tous le jeux d'entrainement: 0.9932067364390967

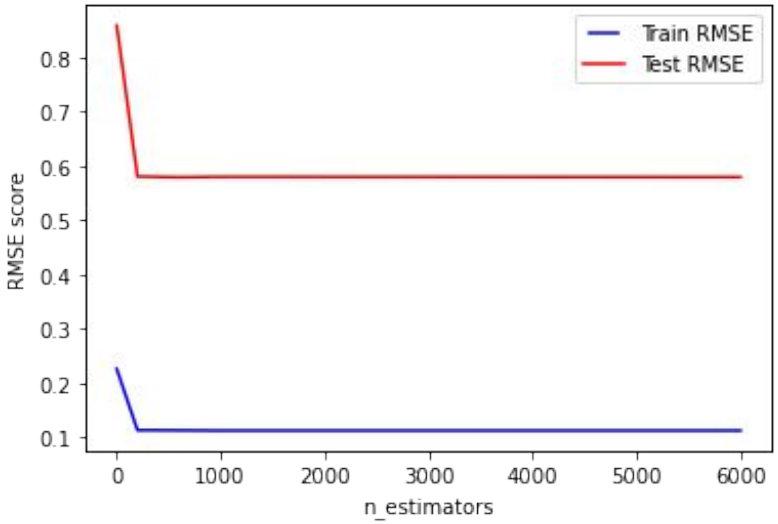
Moyenne cross-validée de la RMSE du meilleur modèle sur le jeu d'entrainement: 0.3939622763329374

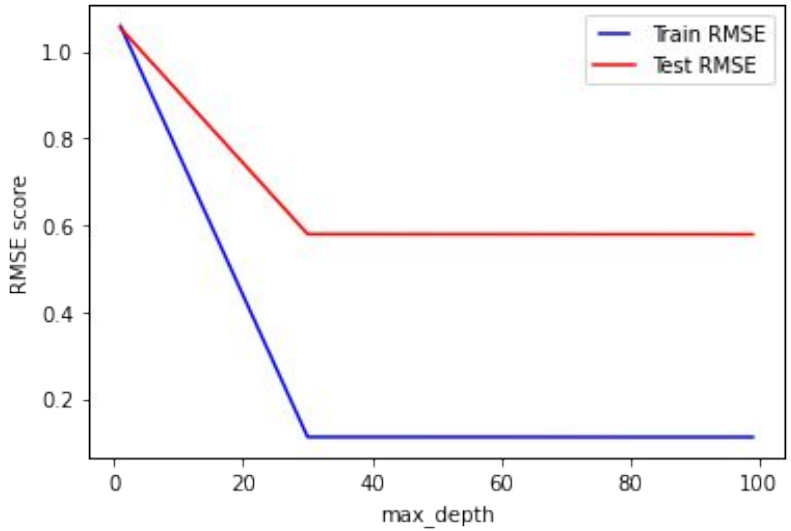


Importance des variables pour Extra Tree



Evaluation de l'influence des hyperparamètres sur le RMSE





Conclusion

- Amélioration en utilisant les variables cibles en log
- Pas d'amélioration en utilisant un jeu de variable restreint (% importance des features)
- Amélioration en utilisant RandomizedSearchCV sur Extra Tree
- Amélioration en supprimant la variable EnergyStars et DataYear
- ENERGYSTARScore n'est pas une variable pertinente pour la prédiction (faible % importance des features)
- La taille du jeu de données d'entraînement détermine la qualité de prédiction de notre modèle
- Amélioration du temps de calcul avec l'évaluation de l'influence des hyperparamètres