OPENCLASSROOM

Parcours data scientist en alternance

13 mai 2022

Maxime Dupouy

Livrable 4

Seattle

Missions	Objectifs
Des relevés coûteux ont été effectués entre en 2015 et en 2016 à Seattle. A partir de ces relevés, **prédire les émissions de CO2 et la consommation totale d'énergie** des bâtiments non résidentiels pour lesquels elles n'ont pas encore été mesurées. Evaluer l'intérêt de l'ENERGYStarScore.	- Réaliser une courte analyse exploratoire. - Tester différents modèles de prédiction Attention : - se passer des relevés annuels - optimiser les performances avec des transformations simples aux variables - évaluation rigoureuse des performances de la régression - optimiser les hyperparamètres et le choix d'algorithme de ML à l'aide d'une validation croisée.

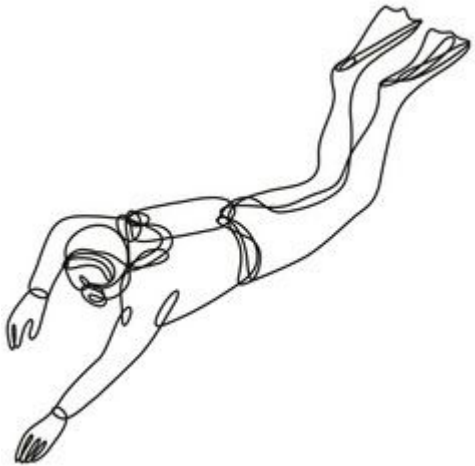
La source

Dataset provenant de la ville de Seattle et hébergée sur **Kaggle.**
La donnée est mise à jour régulièrement.

Data en 2 datasets : 2015 et 2016

Data Cleaning 🧹

Prérequis



Réunir les deux datasets :

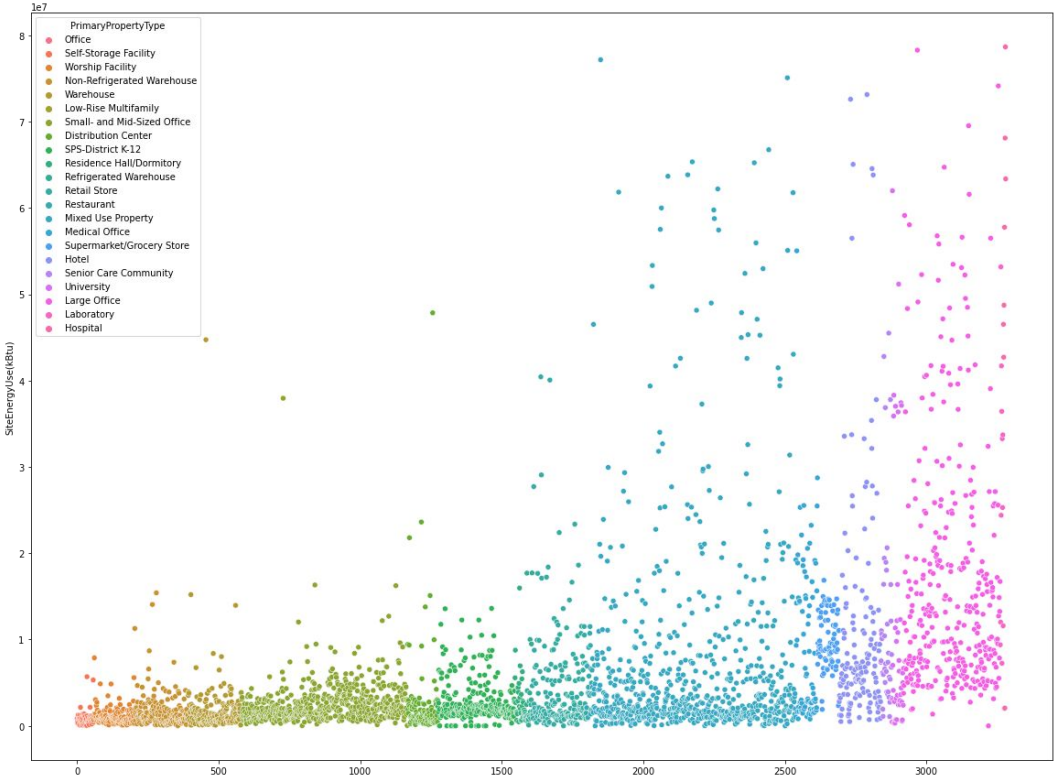
- **homogénéiser les colonnes**
- **concatener les datasets**

Déterminer les cibles :

- **SiteEnergyUse(kBtu),**
- **TotalGHGEmissions,**
- **(ENERGYStarScore)**

Variables catégorielles (1)

- **BuildingType : filtre les bâtiments non résidentiels**
- **Suppression des features sans intérêt (exemple TaxParcelIdentificationNumber)**



- Utilisation **test chi 2** pour choisir certaines features (PrimaryPropertyType, LargestPropertyType)
- PrimaryPropertyType : réduire le nombre de type possible (passage de 30 à 17)

Variables catégorielles (2)

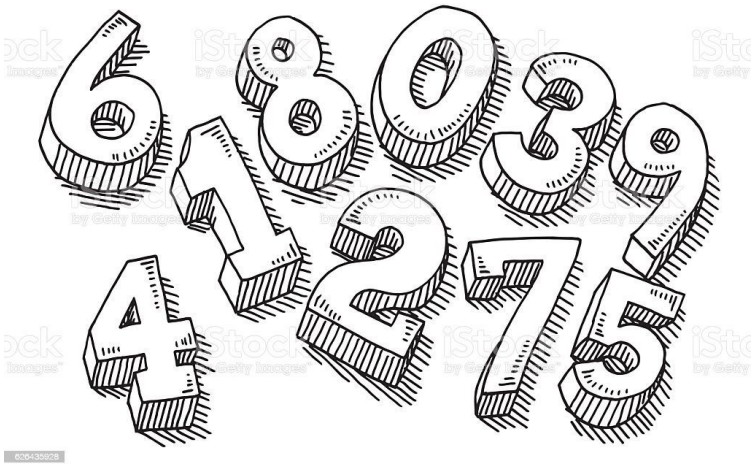
Variables catégorielles (3)

Variables sélectionnées :

- **PrimaryPropertyType**
- **Neighborhood**
- **BuildingType**

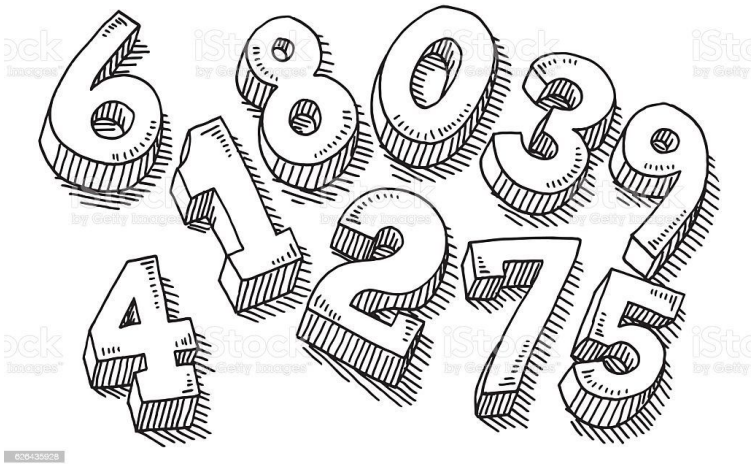
-> One hot encoding

Variables quantitatives (1)



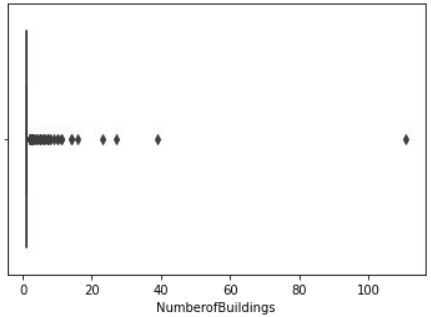
- **Suppression de toutes variables provenant des relevés**
- **Data-engineering : calcul age du bâtiment plutôt que l'année, ratio parking/surfacetotal, buidling/surfacetotal**
- **Analyse corrélation**

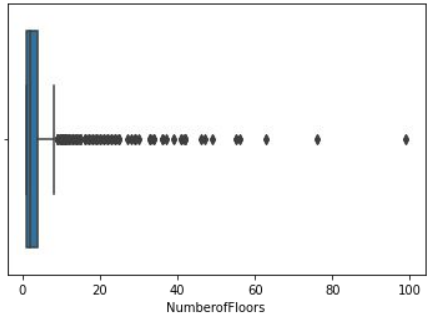
Variables quantitatives (2)



- **Vérification des outliers (surfaces / nombre étage négatives)**
- **Remplir par la valeur de l'autre année si possible**

Variables quantitatives (3)





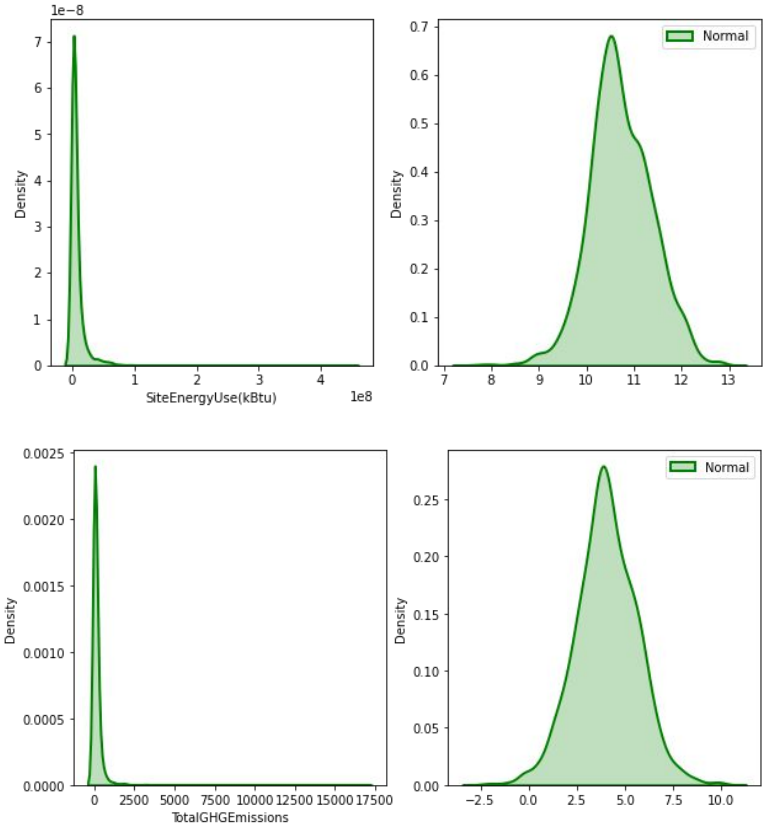
Variables sélectionnées :

- Numberofbuildings
- Numberoffloors
- parking ratio
- building ratio
- DataYear (encodée)
- PropertyGFABuilding(s)
- BuildingAge

-> Normalisation par **RobustScaler** due à la distribution présentant des outliers

$$X_{\text{scale}} = \frac{x_i - x_{\text{med}}}{x_{75} - x_{25}}$$

Cibles



Application de la transformée de Box-Cox sur les targets

$$y(\lambda) = \begin{cases} (y^{\lambda} - 1)/\lambda & if\ \lambda \neq 0; \\ \log y & if\ \lambda = 0. \end{cases}$$

Normalise + retour donnée originale possible

Modèles Sélection 🧐

Baseline

DummyRegressor sur médiane :

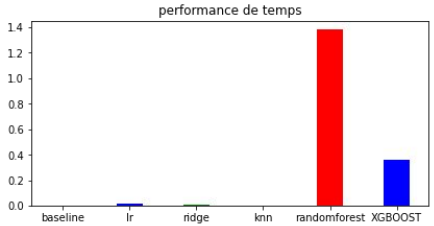
EnergySiteUse (médiane = 2 500 701 kBtu) :

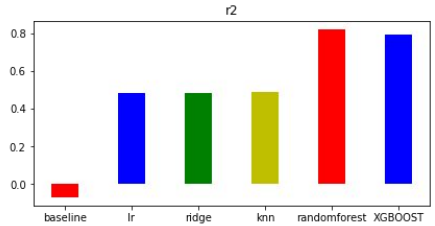
- **MAE :** 1 663 251
- **RMSE :** 19 070 933
- **r2: -**0.07
- **MDAPE** : 71

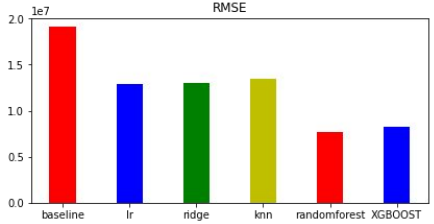
Emission (médiane=49):

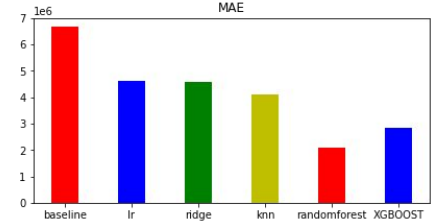
- **MAE**: 36
- **RMSE** : 607
- **r2**: -0.04
- **MDAPE :** 75

Premiers modèles









Gridsearching pour finetuner + **cross validation** (k=5) sur **régression linéaire, ridge, lasso, knn, randomforest et XGBoost**

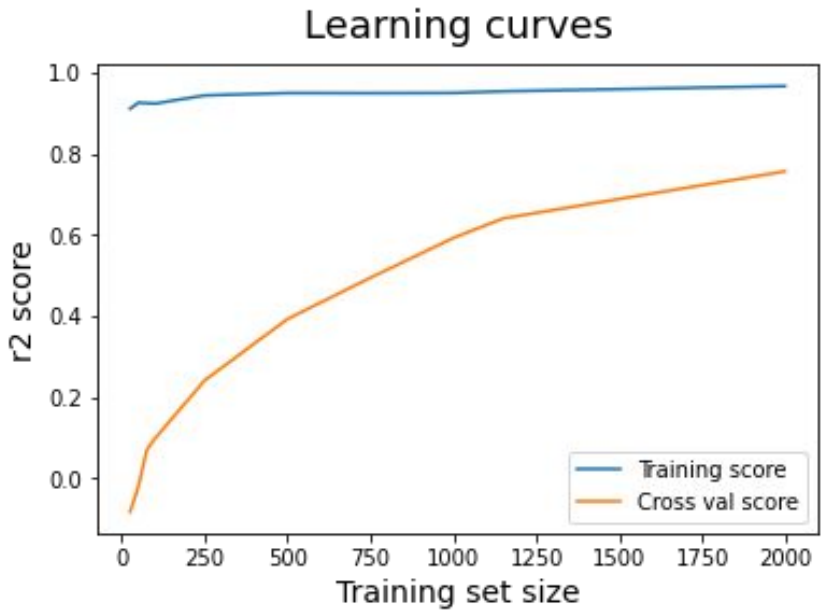
Résultats similaires sur la prédiction des émissions

Evaluation sur :

- MAE,
- RMSE,
- R2
- (MDAPE)

Meilleure modèle : **Random forest**

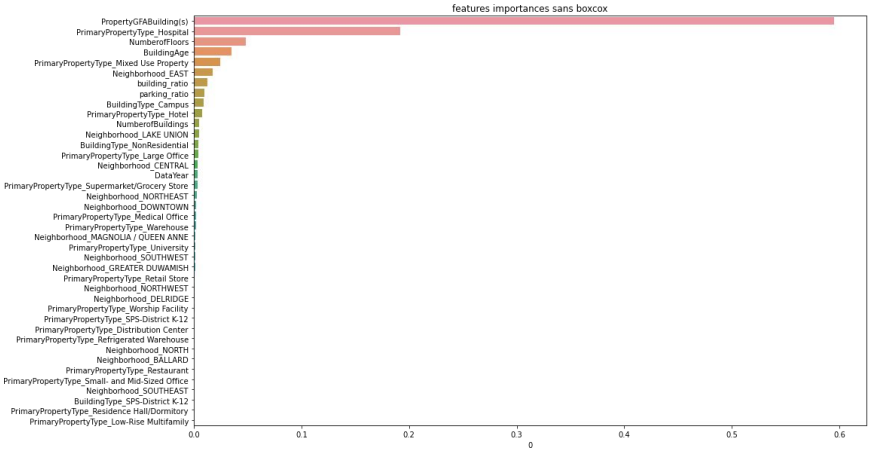
Premiers modèles

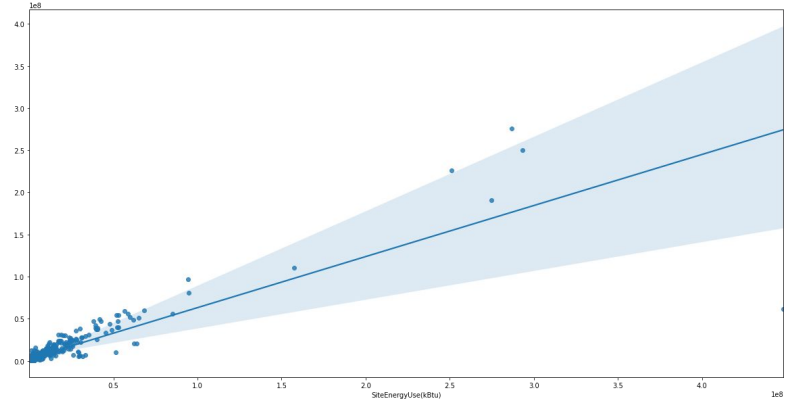


Vérification tradeoff overfitting/underfitting par l**earning curves**

-> Tradeoff ok, mais semble “manquer” de donnée pour atteindre un plateau

Random Forest





- Features importances sur la donnée box cox
- Outliers semblent perturber le modèle
- BoxCox transformation améliore les performances
- Résultats similaires sur la target émission

EnergieStarScore

	metrics	Avec EStarScore	Sans EStarScore
CONSOM-MATION	**r2**	0.9	0.78
	r2 ajusté	0.89	0.77
EMISSION	**r2**	0.82	0.81
	r2 ajusté	0.80	0.8

Sélection donnée avec EnergieStarScore

(perte environ 1000 observations/3000)

Réentrainement modèle avec et sans energie star score

-> impact positif sur la prédiction de la consommation,

-> pas d'impact sur prédiction d' émission.

Entrainements 

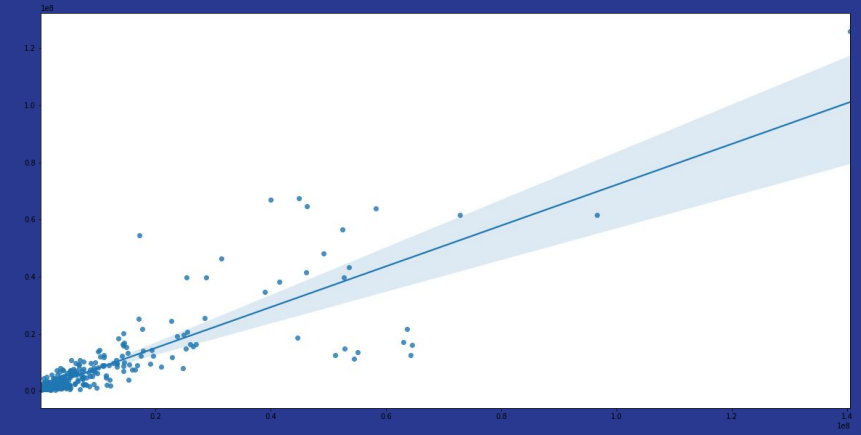
Modèle Final

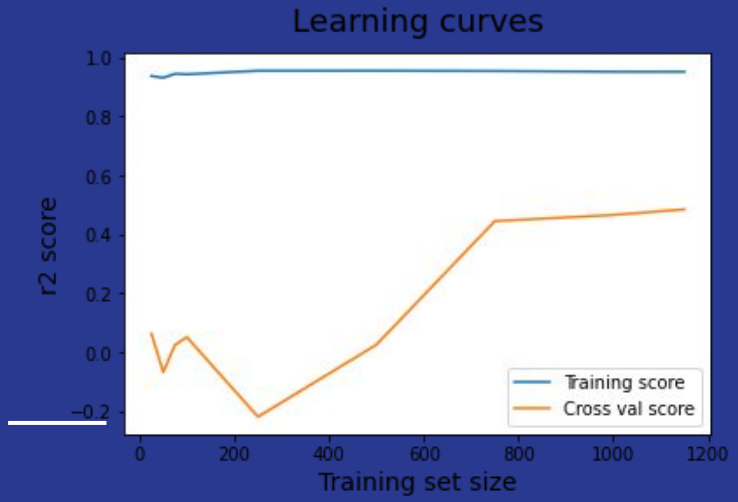
Random Forest :

- sans outliers,
- sans EnergieStarScore (enlève trop de donnée)
- Gridsearch pour finetuning hyperparams

Modèle Final Consommation

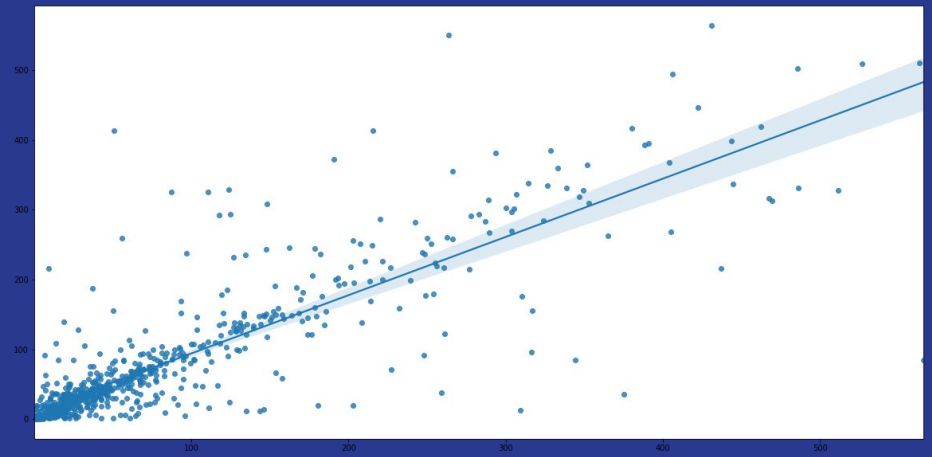
	BASELINE	Résultats
MAE	~~1 663 251~~ 1 802 401	417 621 979 627
RMSE	~~19 070 933~~ 17 310 936	1 917 678 7733288
R2	~~0.07~~ 0.06	0.8 0.71
mdape	~~71~~ 72	21 35





Modèle Final Emission

	BASELINE	Résultats
MAE	~~36~~ 39	7.5
RMSE	~~607~~ 405	53
R2	~~-0.04~~ -0.08	0.71
mdape	~~75~~ 78	17



#60648792

Modèle limité pour répondre à nos objectifs :

- Baseline largement améliorée
- R2 0.8 et 0.7
- mais MAE/RMSE hautes

EnergyStarScore semble très pertinent sur la consommation

Pistes d'améliorations :

- avoir plus de données
- travailler colonne LargestPropertyUses ?

Conclusions personnelles :

- Difficulté pour cerner la consigne
- Prédiction de plusieurs cibles + évaluation d'une feature -> beaucoup de pistes

Je vous remercie pour votre attention