Projet 3

Anticipez les besoins en consommation électrique de bâtiments

Adrian Rodriguez - Ingénieur Machine Learning

Seattle

OPENCLASSROOMS

CentraleSupélec

Problématique

Emissions carbone et consommation d’énergie

1. Orientation / lumière naturelle
2. Isolation
3. Etanchéité à l’air
4. Performance des équipements
5. Mode de chauffage

Présentation du jeu de données

- 3340 observations
- 47 variables

2015

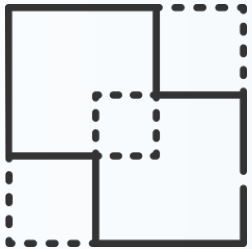
2016

- 3376 observations
- 46 variables

- Traitement des bâtiments en double
- Exclusion des bâtiments résidentiels

- 1727 observations
- 46 variables

Consistance des données



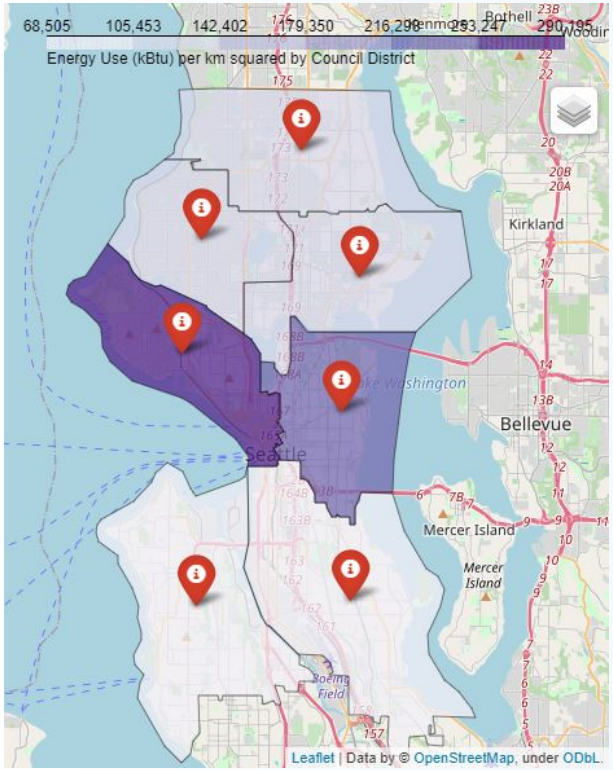
Surface GFA

Tower

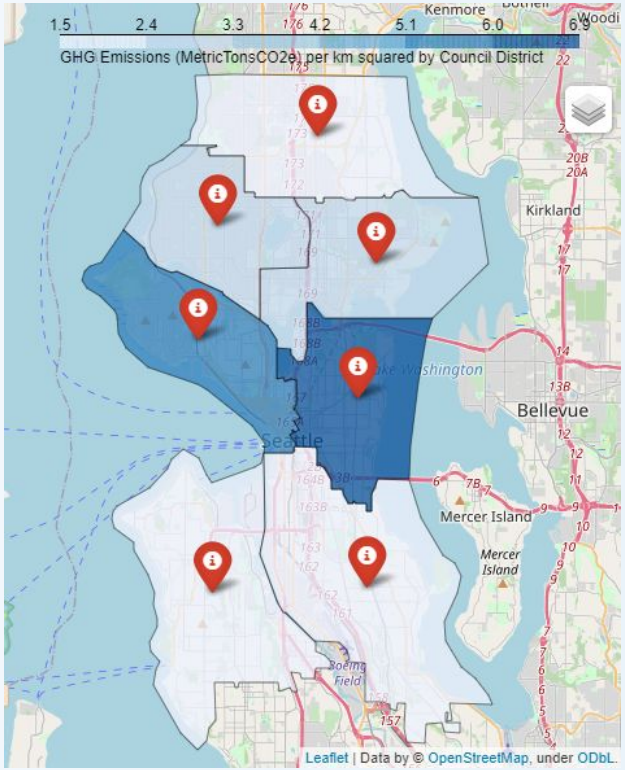
Energy use

Feature engineering - Localisation

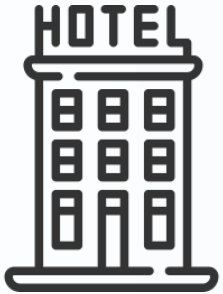
Site Energy Use (kBtu)



GHG Emissions (TonsCO2)

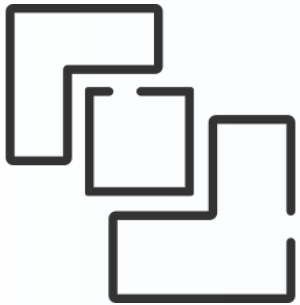


Features engineering - Usage/Equipement



Features engineering - Bâtiment

Ratio des surfaces

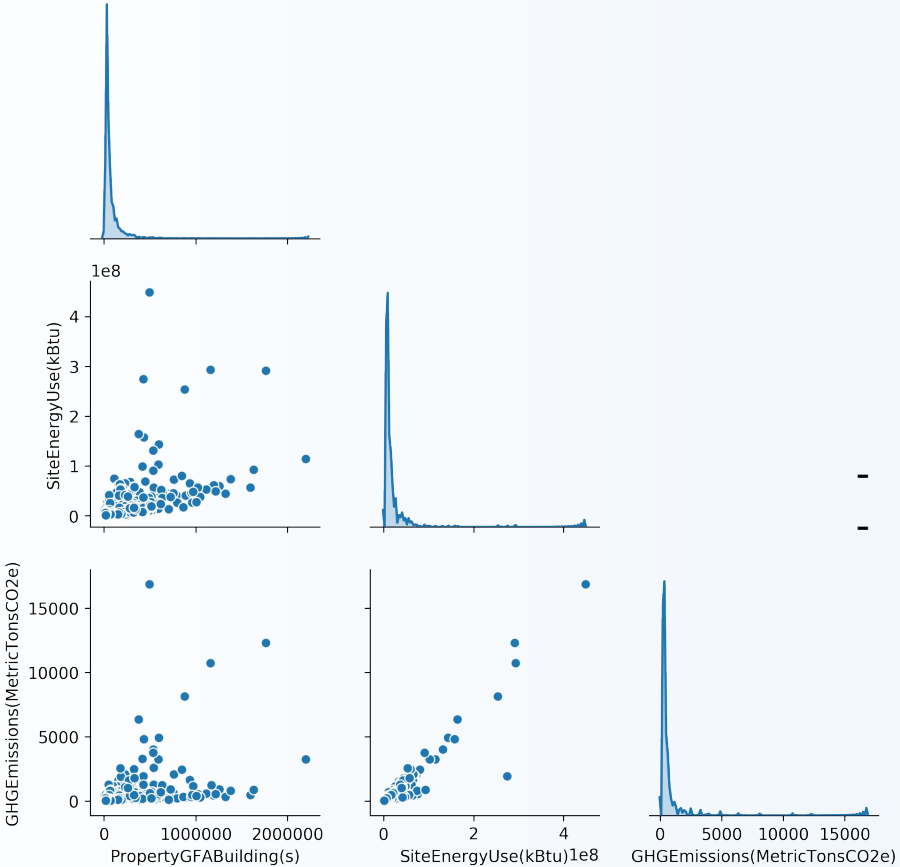


Hauteur des bâtiments

Période de construction

Score > 75 = Top performer

Exploration des variables quantitatives

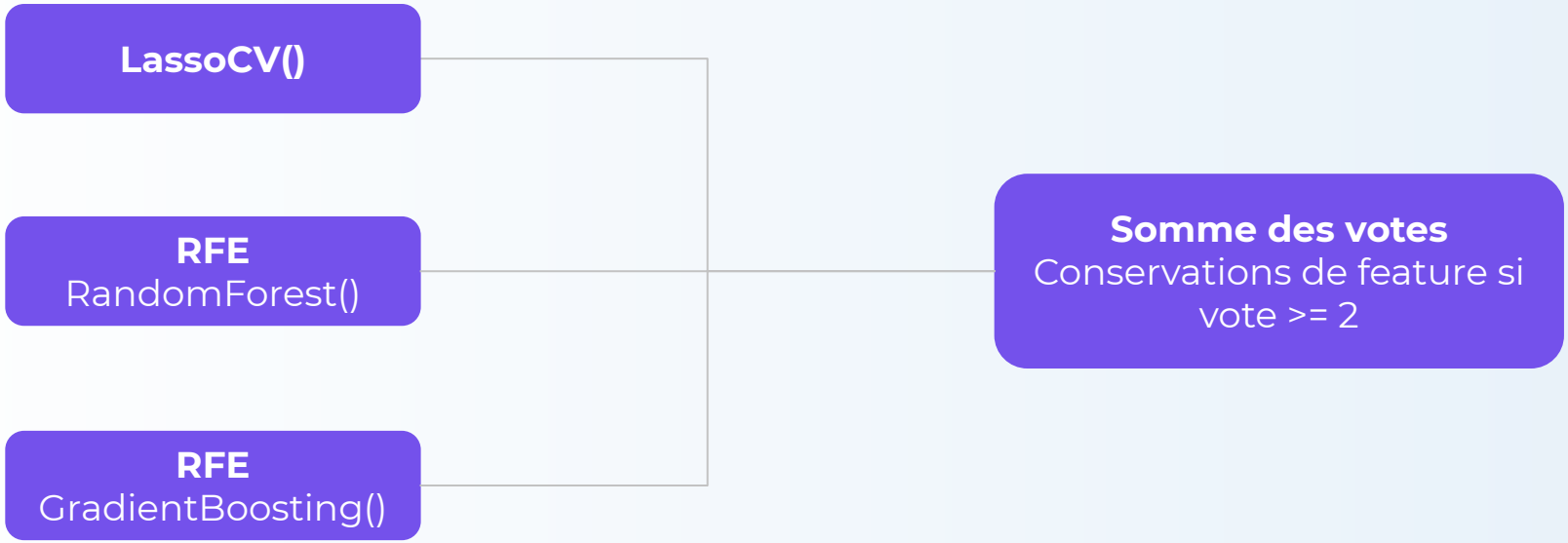


- Distributions non gaussienne
- Variances élevées

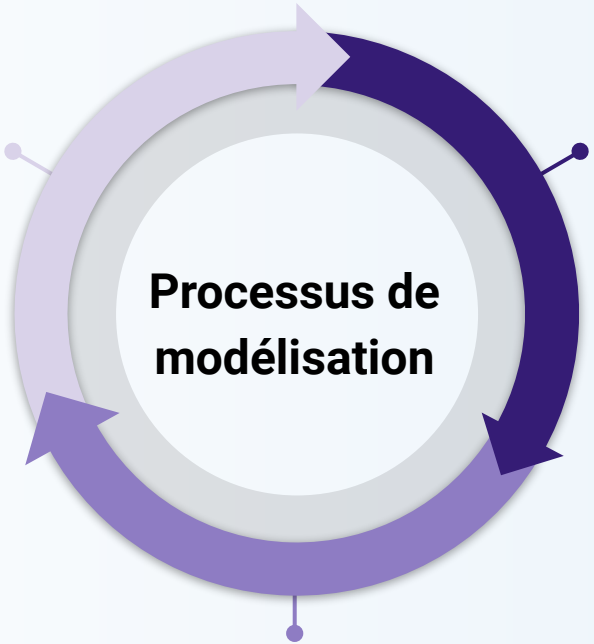
Choix des modèles

Méthodes linéaires	Arbre de décision	Méthode ensembliste
- Ridge - Lasso	- DecisionTreeRegressor	- RandomForestRegressor - GradientBoostingRegressor
Hyperparamètres : - alpha	Hyperparamètres : - max_depth - min_samples_split	Hyperparamètres : - max_depth - min_samples_split - n_estimators

Features selection



Modélisation



Choix du modèle

Réglage des plages des hyperparamètres

Analyse des résultats sur données de test

Récupération des RMSE et r2

Ajustement du modèle sur données d'entraînement

Cross-Validation 5 fois / Métrique RMSE / Récupération des meilleurs hyperparamètres

Validation des modèles

Consommation d'énergie

EnergyUse	Test : RMSE	Test : R2	mean_test_score	std_test_score	alpha	max_depth	min_samples_split	n_estimators
Ridge	-0.4367	0.8084	-0.4831	0.0479	7.5753	nan	nan	nan
Lasso	-0.4363	0.8088	-0.4813	0.0485	0.0116	nan	nan	nan
Decision-Tree	-0.4972	0.7516	-0.5292	0.0477	nan	8.0	20.0	nan
RandomForest	-0.4326	0.812	-0.4908	0.0483	nan	8.0	12.0	1000.0
GradientBoosting	-0.44	0.8055	-0.4788	0.0424	nan	4.0	500.0	100.0

Emissions carbone

GHGE	Test : RMSE	Test : R2	mean_test_score	std_test_score	alpha	max_depth	min_samples_split	n_estimators
Ridge	-0.6403	0.5806	-0.6687	0.0465	10.7189	nan	nan	nan
Lasso	-0.6386	0.5827	-0.6687	0.046	0.0041	nan	nan	nan
Decision-Tree	-0.655	0.561	-0.7347	0.0564	nan	6.0	50.0	nan
RandomForest	-0.6348	0.5876	-0.6796	0.0503	nan	8.0	20.0	100.0
GradientBoosting	-0.6371	0.5846	-0.6677	0.0447	nan	4.0	200.0	50.0

Rôle de l'Energy Star Score

Émissions carbone **avec** Energy Star Score

GHGE	Test : RMSE	Test : R2	mean_test_score	std_test_score	alpha	max_depth	min_samples_split	n_estimators
Ridge	-0.6403	0.5806	-0.6687	0.0465	10.7189	nan	nan	nan
Lasso	-0.6386	0.5827	-0.6687	0.046	0.0041	nan	nan	nan
Decision-Tree	-0.655	0.561	-0.7347	0.0564	nan	6.0	50.0	nan
RandomForest	-0.6348	0.5876	-0.6796	0.0503	nan	8.0	20.0	100.0
GradientBoosting	-0.6371	0.5846	-0.6677	0.0447	nan	4.0	200.0	50.0

Émissions carbone **sans** Energy Star Score

GHGE_sans_ESS	Test : RMSE	Test : R2	mean_test_score	std_test_score	alpha	max_depth	min_samples_split	n_estimators
Ridge	-0.6946	0.5064	-0.7123	0.049	9.5477	nan	nan	nan
Lasso	-0.6935	0.5079	-0.7114	0.0493	0.0065	nan	nan	nan
Decision-Tree	-0.6954	0.5052	-0.7449	0.0659	nan	6.0	50.0	nan
RandomForest	-0.6801	0.5268	-0.7072	0.0542	nan	6.0	20.0	50.0
GradientBoosting	-0.6854	0.5194	-0.702	0.0437	nan	4.0	200.0	50.0

Modèle final

	Consommation d'énergie	Emissions carbone
Modèle	Random Forest	Random Forest
RMSE	0.44	0.63
r2	0.81	0.58