Anticipez les besoins en consommation électrique de bâtiments

Seattle

Soutenance de projet

Nicolas FAUCONNIER
Parcours Ingénieur ML
31/01/2022

Plan

1. Problématique
Projet et missions

2. Data Cleaning
Nettoyage et pré-sélection des données

3. Analyse exploratoire
Analyse univariée et multivariée

4. Feature Engineering
Création de nouvelles variables et pipeline de pré-processing

5. Modélisation
Entrainement et évaluation des modèles

1.

Problématique

Projet et missions

1. Problématique

Contexte

La ville de Seattle mène différentes études afin d'atteindre son objectif de ville neutre en émissions de carbone en 2050. La prédiction des émissions de gaz à effet de serre et de la consommation des bâtiments de la ville seraient utiles pour atteindre cet objectif. Cela permettrait également de diminuer les coûts liés aux relevés de consommation.

Missions

- Réaliser une courte analyse exploratoire des données
- Tester différents modèles de prédiction, sélectionner et optimiser les plus performants
- Juger de l'intérêt de l'ENERGYSTARScore comme variable explicative des émissions de gaz à effet de serre

2.

Data Cleaning

Nettoyage et présélection des variables explicatives pertinentes pour la modélisation

Dataset

Caractéristiques techniques de bâtiments de la ville de Seattle, pour les années 2015 et 2016. Données issues du permis d'exploitation commerciale.

Le dataset spécifiquement utilisé ici est hébergé sur Kaggle au lien suivant: https://www.kaggle.com/city-of-seattle/

Aussi disponible sur la plateforme opendata de Seattle.

- 2 .csv, **pour un total de 3Mb**
- **~3300 lignes** et **~47 colonnes par année**

Outils utilisés: colab jupyter

Données

- Informations sur les consommations d’énergie
- Nature du bâtiment: différents types d’usages
- Surfaces: total, par type d’usage, parking
- Données de localisation: quartier, Zipcode, longitude et latitude
- Autres: commentaires, est un outlier, année de construction, année du relevé

Data Cleaning

- Suppression des colonnes non exploitables ou inutiles: commentaires, adresses, données inconnues sans descriptions ou trop éparses

- Suppression des colonnes redondantes: différentes unités d’une même mesure, données normalisées (selon la surface et les variations de températures annuelles)

- Harmonisation des colonnes entre 2015 et 2016: noms, dtypes, extraction des longitudes et latitudes d’une colonne en .json dans le dataset de 2015

Data Cleaning

- Harmonisation des catégories
- Suppression des outliers indiqués dans le dataset
- Valeurs incohérentes: valeurs négatives, incohérence des surfaces

Data Cleaning

Surfaces par type d'usage:

- Fill par 0 des usages 2 et 3
- création de la catégorie « None » quand XLargestUsage = NaN

Suppression de l'année 2015:

- Valeurs quasi-identiques avec 2016, et constitue presque la totalité des observations
- Pas de déséquilibre dans le training pour les quelques observations uniques

→ 1622 observations

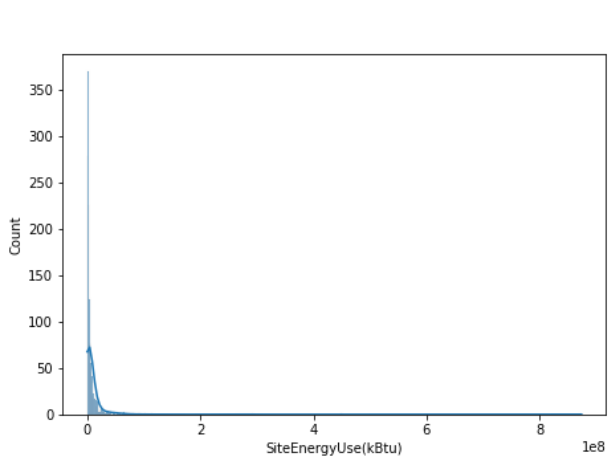
(*Nb. Pas de doublons sur la même année*)

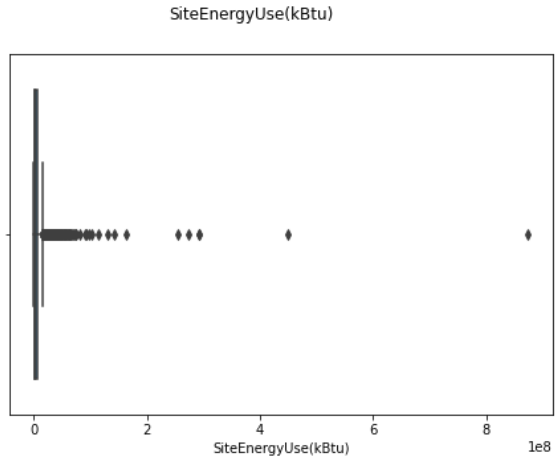
3.
Analyse exploratoire

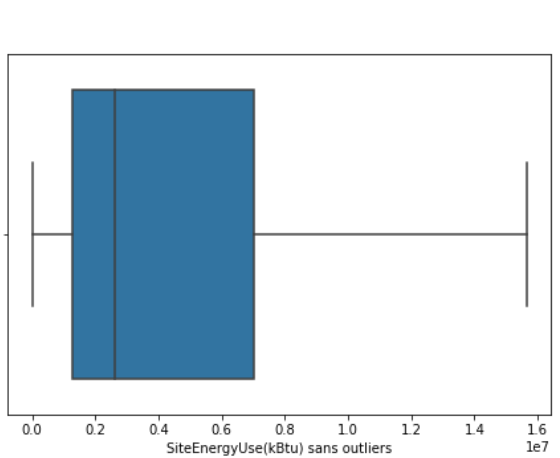
Analyse univariée et multivariée

Analyse univariée

Beaucoup de variables quantitatives ont une majorité de valeur proche de zéro:







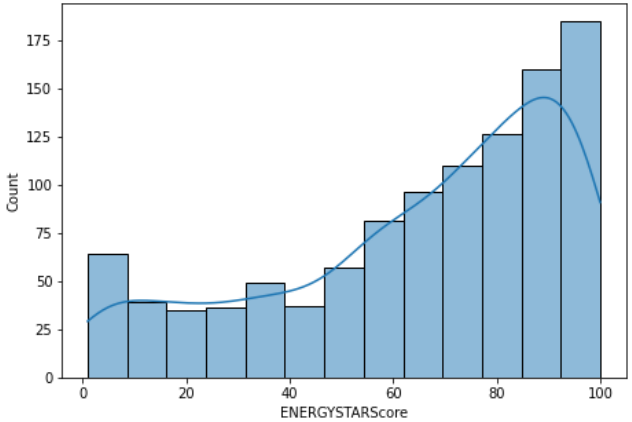
Notamment les variables à prédire:

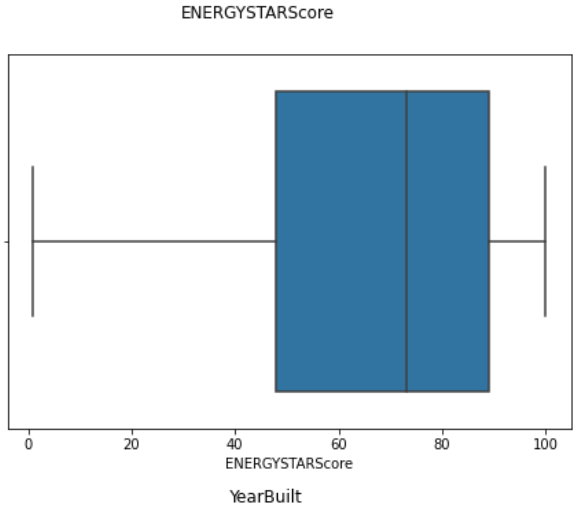
Considération de la log-transformation

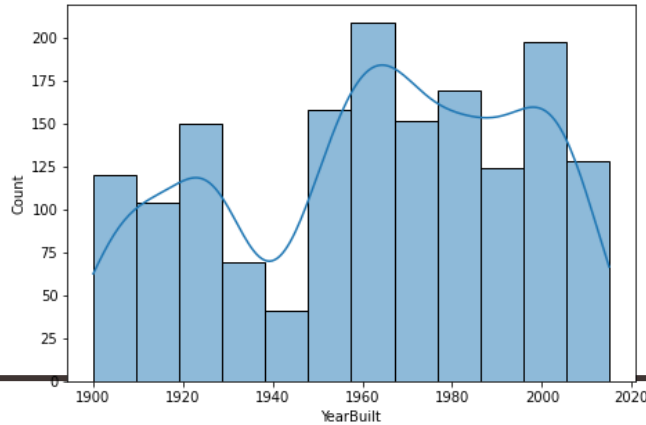
Analyse univariée

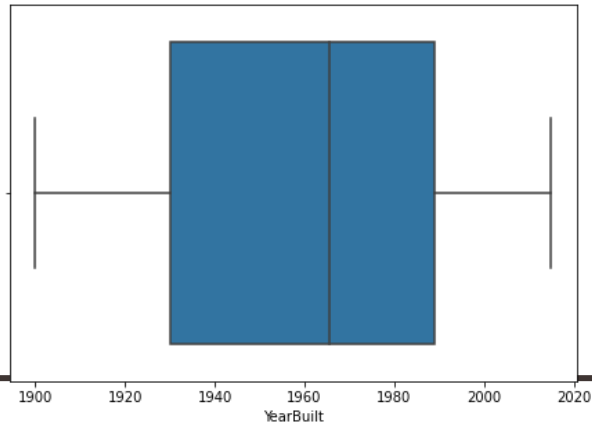
Mais ce n’est pas le cas pour toutes les variables:

(Nb. Aucune variable quantitative n’est normalement distribuée)



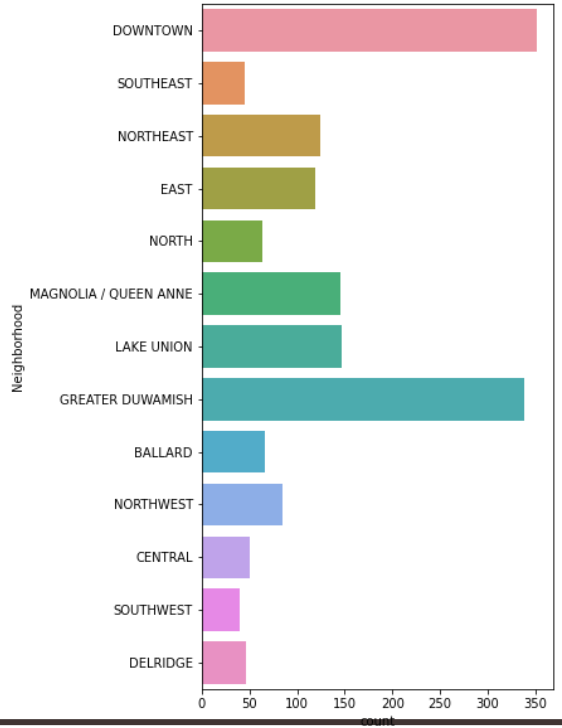


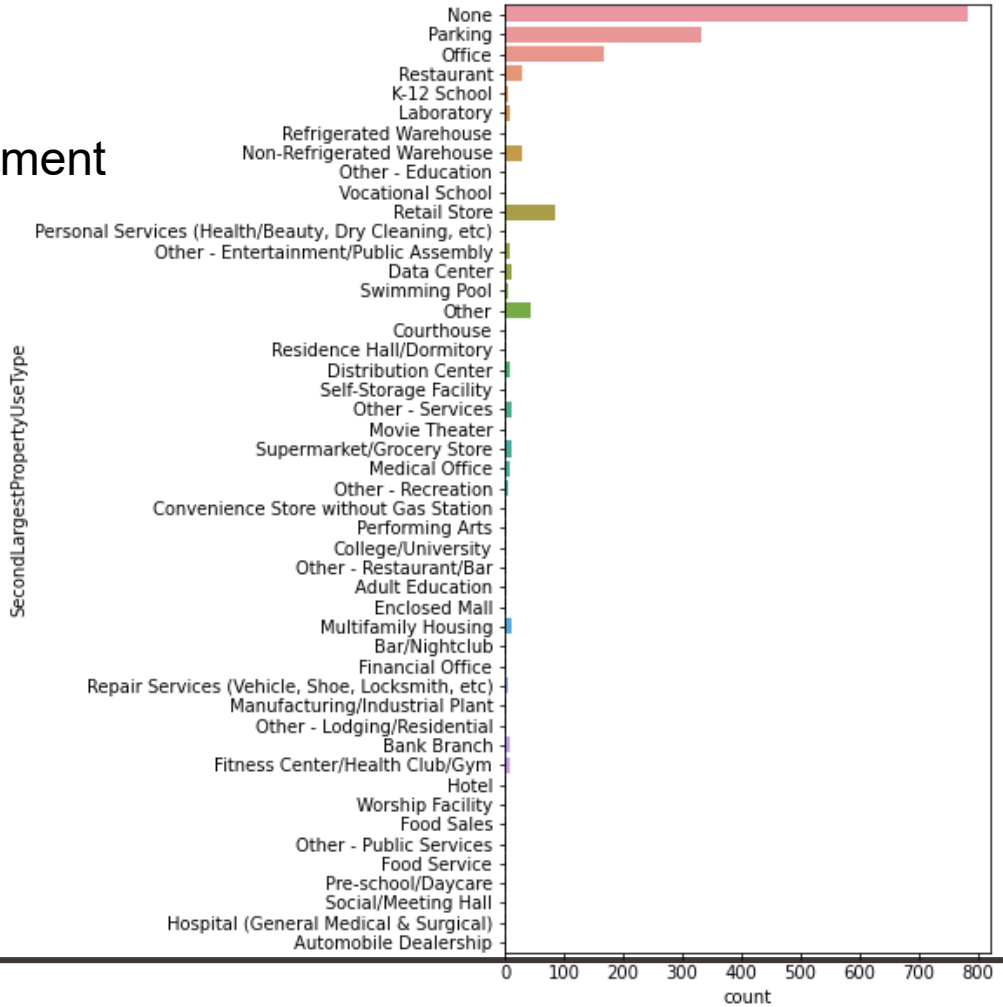




Analyse univariée

Aucune variable catégorielle n'est également distribuée

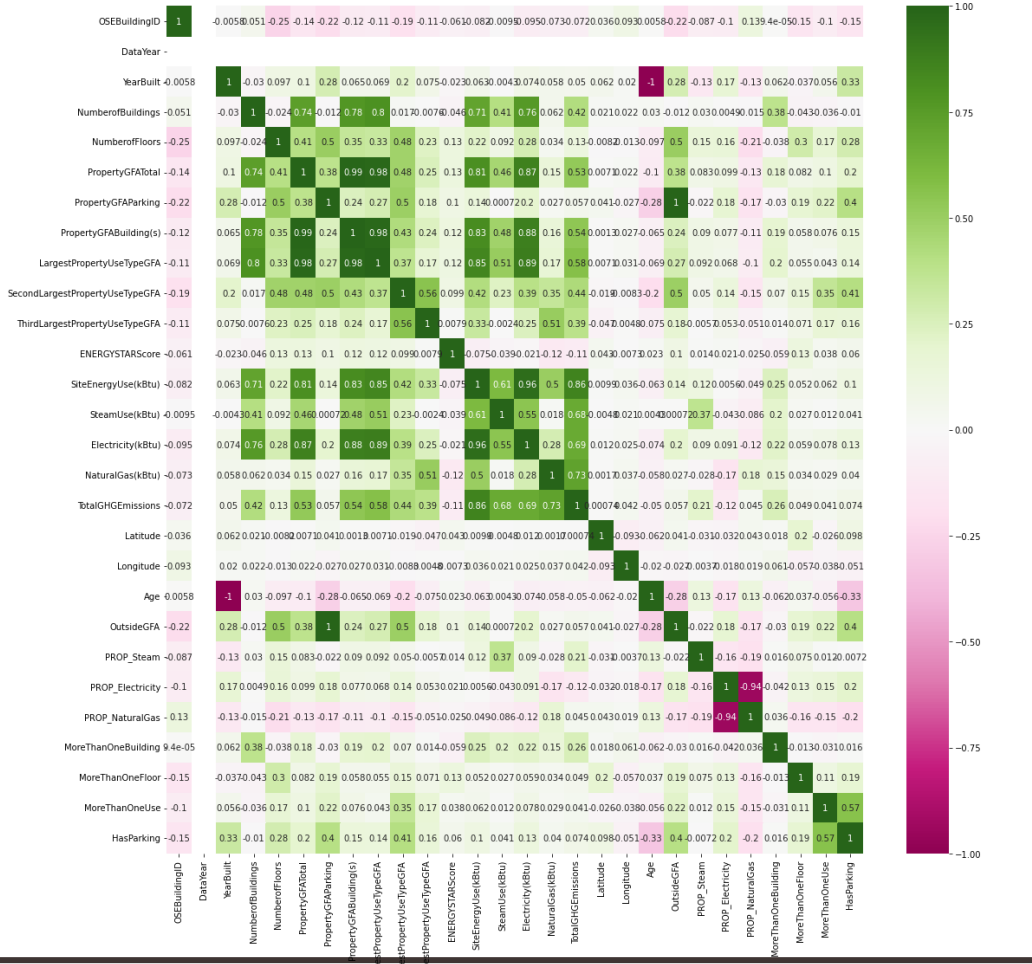




Analyse bivariée

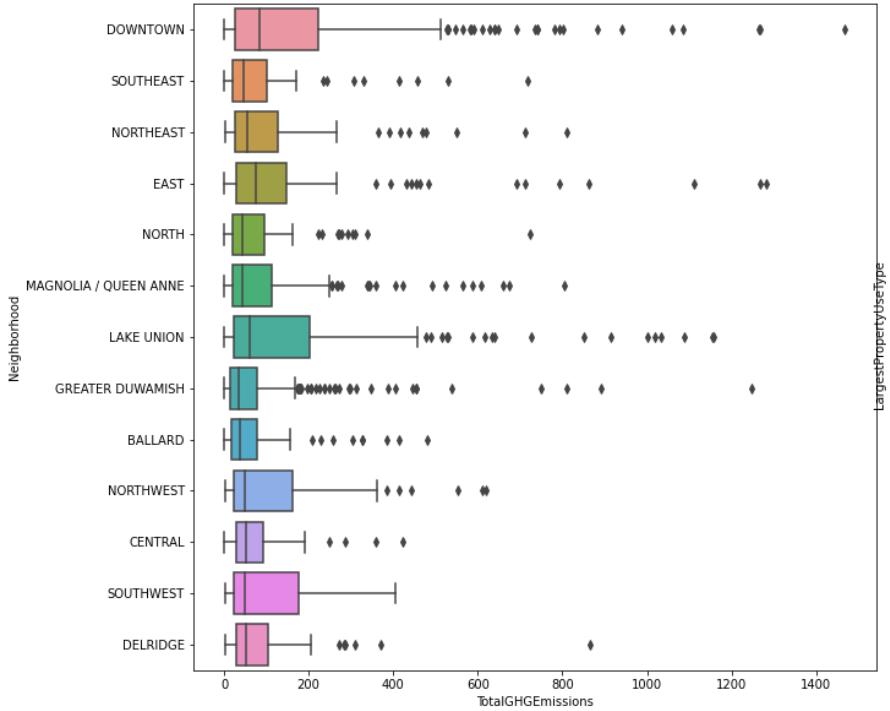
Les deux variables à expliquer sont corrélées avec de nombreuses features.

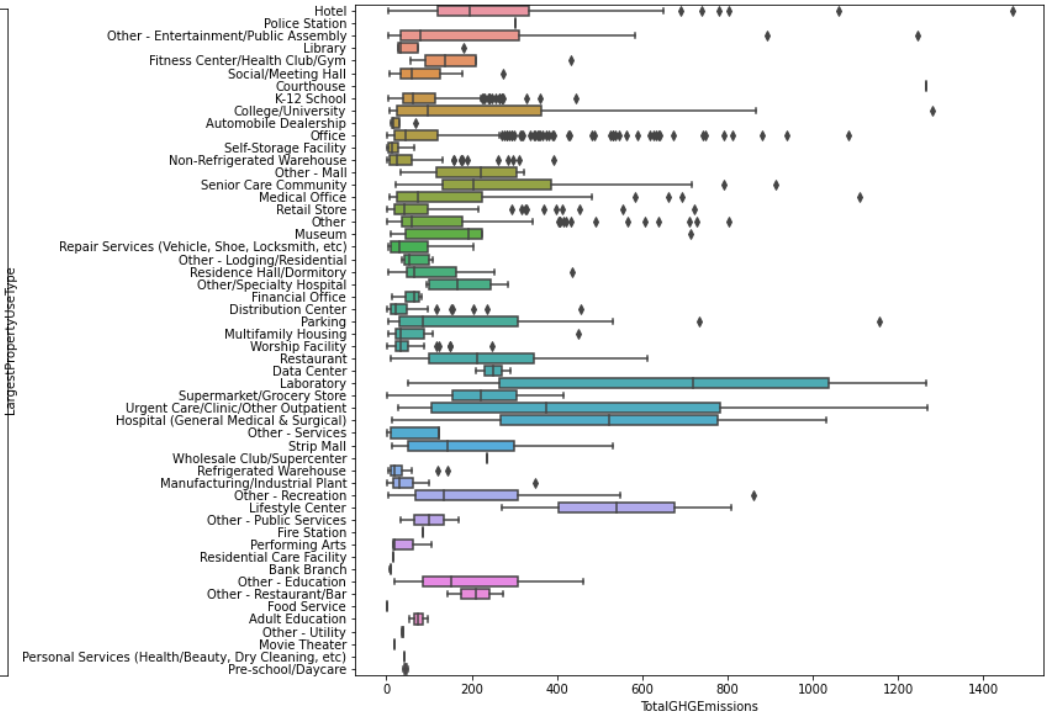
La corrélation négative et forte entre la consommation/proportion de gaz et d'électricité suggère qu'un type d'énergie se soustrait à l'autre dans la plupart des cas.



Analyse bivariée

Le type d'usage semble avoir un lien avec les variables à expliquer, pas la localisation:





4.

Feature Engineering

Création de nouvelles variables et pipeline de pré-processing

Création de nouvelles variables

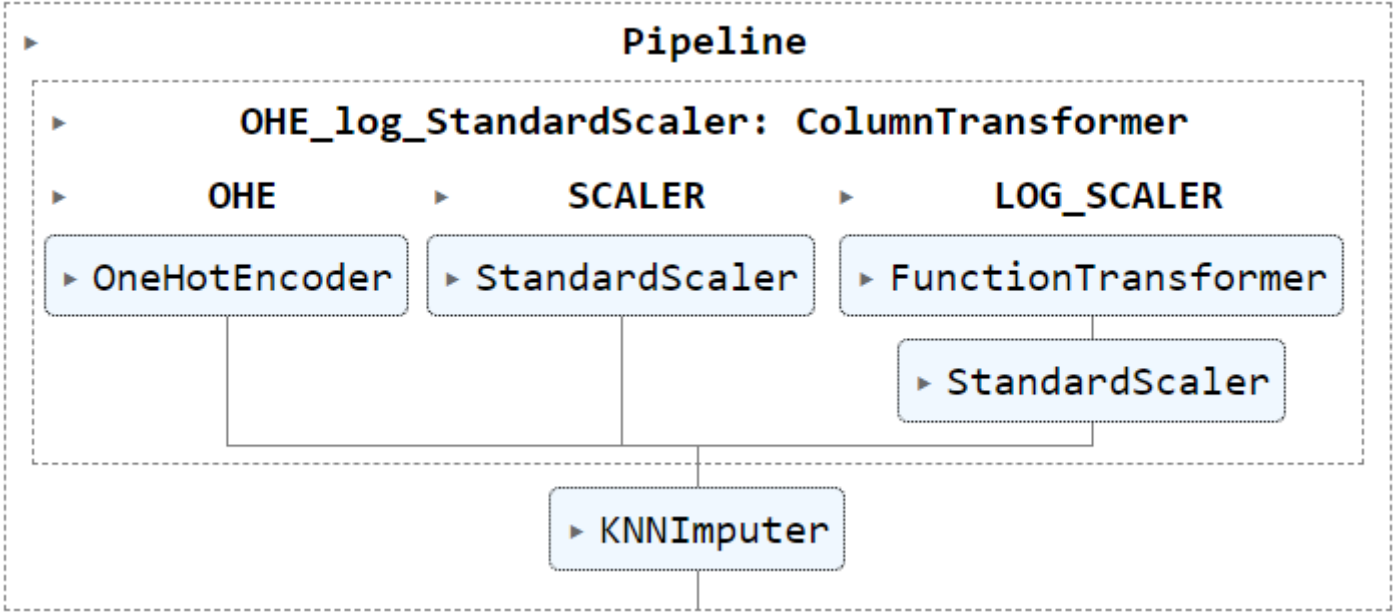
Pivot des types d’usages (et déduction du 4ème usage le plus important quand c’est possible)

Age de la propriété

Variables Dummies:

- Plus d’un bâtiment
- Plus d’un étage
- Plus d’un usage
- Possède un parking

Pipeline de pré-processing



5.
Modélisation

Entrainement et évaluation des modèles

Features et transformations

Features en commun: Surface intérieure, âge de la propriété

Combinaisons de features testées:

- Nombre de bâtiments, d'étages et surface du parking VS. leurs dummies
- Longitude et latitude VS. Neighborhood
- LargestUsage, sa surface et dummy plus d'un usage VS. surfaces pivotées
- ENERGYSTARScore; avec VS. Sans

Transformations:

- OneHotEncoding
- StandardScaler
- Log-transformation puis StandardScaler (testé avec ou sans)

Modèles

- Regression linéaire → **Baseline**
- Regression Ridge
- Regression Lasso
- Regression ElasticNet
- SVM
- RandomForestRegressor
- XGBRegressor

GHGEmissions: Recherche du meilleur modèle

Fonction testant toutes les combinaisons de Feature x Transformation x Modèle sur le training test (20% du dataset) par Cross-Validation (k=5)

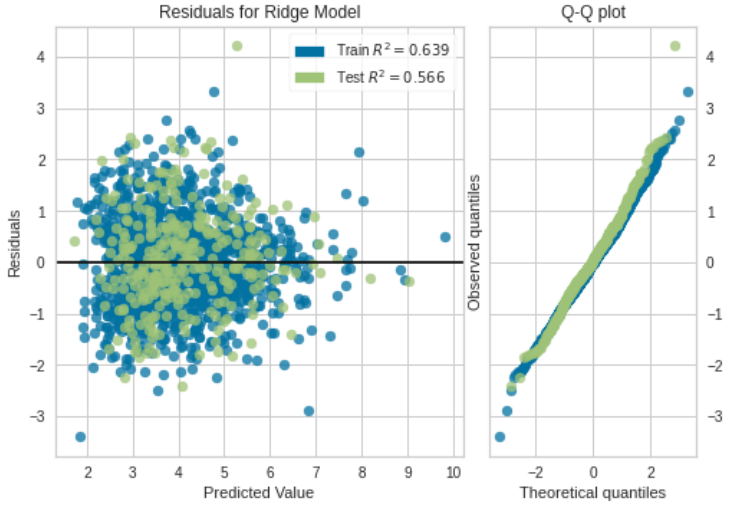
	Model	label	Log Transformation	neg_mean_absolute_error	fit_time
60	Ridge(random_state=1337)	Dummies_Lon_lat_LargestUsage_ENERGYSTARScore	1.0	-7.124000e-01	0.0030
52	Ridge(random_state=1337)	Dummies_Neighborhood_LargestUsage_ENERGYSTARScore	1.0	-7.134000e-01	0.0092
36	Ridge(random_state=1337)	Continues_Neighborhood_LargestUsage_ENERGYSTAR...	1.0	-7.184000e-01	0.0112
44	Ridge(random_state=1337)	Continues_Lon_lat_LargestUsage_ENERGYSTARScore	1.0	-7.184000e-01	0.0040
165	XGBRegressor(random_state=1337)	Continues_Neighborhood_LargestUsage_ENERGYSTAR...	0.0	-7.188000e-01	0.2082
180	XGBRegressor(random_state=1337)	Dummies_Neighborhood_LargestUsage_ENERGYSTARScore	1.0	-7.232000e-01	0.2056
181	XGBRegressor(random_state=1337)	Dummies_Neighborhood_LargestUsage_ENERGYSTARScore	0.0	-7.246000e-01	0.2046
		...			
26	LinearRegression()	Dummies_Lon_lat_UGFA_No_ENERGYSTARScore	1.0	-3.464497e+11	0.0222
27	LinearRegression()	Dummies_Lon_lat_UGFA_No_ENERGYSTARScore	0.0	-3.896742e+11	0.0122
10	LinearRegression()	Continues_Lon_lat_UGFA_No_ENERGYSTARScore	1.0	-1.155827e+12	0.0522
24	LinearRegression()	Dummies_Lon_lat_UGFA_ENERGYSTARScore	1.0	-2.480234e+12	0.0214
25	LinearRegression()	Dummies_Lon_lat_UGFA_ENERGYSTARScore	0.0	-2.601300e+12	0.0228

GHGEmissions: Régression Ridge

Recherche de la meilleure valeur d'alpha par GridSearch et CV, puis évaluation :

```
['PropertyGFATotal',
 'PropertyGFABuilding(s)',
 'Age',
 'MoreThanOneBuilding',
 'MoreThanOneFloor',
 'HasParking',
 'Longitude',
 'Latitude',
 'LargestPropertyUseType',
 'LargestPropertyUseTypeGFA',
 'MoreThanOneUse',
 'ENERGYSTARScore']
```

	param_alpha	mean_test_score	mean_score_time
5	1.25	-0.712286	0.000713
6	1.4	-0.712333	0.000725
4	1.1	-0.712342	0.000748
3	1	-0.712390	0.000729
7	1.5	-0.712443	0.000775
2	0.9	-0.712499	0.000769
1	0.75	-0.712751	0.000897
8	2	-0.713186	0.000982
0	0.5	-0.713468	0.000882
9	10	-0.727478	0.000689
10	100	-0.791326	0.000712
11	1000	-0.876194	0.000680



MAE = 0.727

échelle log, sur set de validation

MAE à l'échelle = 1.069

Tonnes de C02

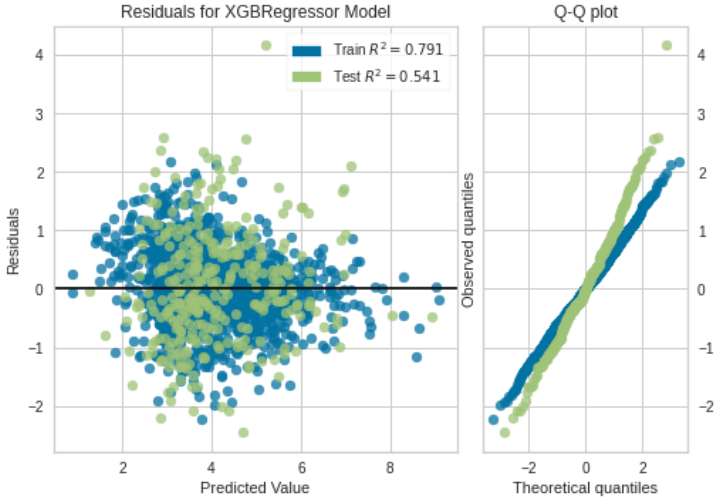
GHGEmissions: XGBoost

Recherche des meilleurs hyperparamètres par GridSearch et CV, puis évaluation :

	param_learning_rate	param_max_depth	param_n_estimators	param_reg_alpha	param_reg_lambda	mean_test_score	mean_fit_time
40	0.1	4	150	0	1.3	-0.706054	0.618665
46	0.1	4	200	0	1.3	-0.706674	0.801353
111	0.15	3	150	0	1.1	-0.708083	0.473094
127	0.15	4	100	0.5	1.3	-0.708127	0.422218
19	0.1	3	200	0.5	1.3	-0.708484	0.655359

```
['PropertyGFATotal',
 'PropertyGFABuilding(s)',
 'Age',
 'NumberofBuildings',
 'NumberofFloors',
 'PropertyGFAParking',
 'Neighborhood',
 'LargestPropertyUseType',
 'LargestPropertyUseTypeGFA',
 'MoreThanOneUse',
 'ENERGYSTARScore']
```

```
parameters = {
    'n_estimators': [50, 100, 150, 200],
    'learning_rate': [0.1, 0.15, 0.3],
    'max_depth': [3, 4, 5, 6],
    'reg_alpha': [0.5, 0],
    'reg_lambda': [1.1, 1.3, 1.5]
}
```



MAE = 0.760 (*vs. 0.727)*

échelle log, sur set de validation

MAE à l'échelle = 1.138

Tonnes de C02

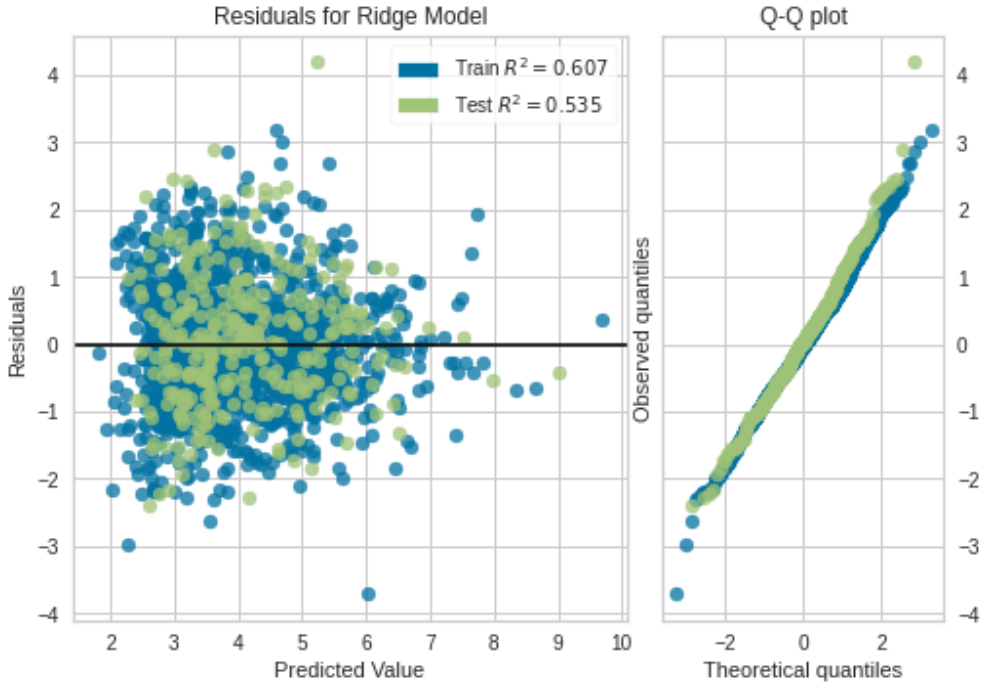
GHGEmissions: Intérêt de l'ENERGYSTARScore

Entrainement d'une régressions Ridge avec le même alpha (1.25) et les mêmes features, mais SANS l'ENERGYSTARScore:

MAE avec ENERGYSTARScore = 0.727

MAE sans ENERGYSTARScore = 0.761

→ l'ENERGYSTARSCore est une feature utile !



SiteEnergyUse: Recherche du meilleur modèle

Même démarche: test de toutes les combinaisons de Feature x Transformation x Modèle sur le training test par Cross-Validation, **SANS l'ENERGYSTARScore** (data leakage)

	Features	Log Transformation	Model	R²	fit_time	label	neg_mean_absolute_error	neg_mean_absolute_percentage_error	neg_mean_squared_error
30	[PropertyGFATotal, PropertyGFABuilding(s), Age...	1.0	Ridge(random_state=1337)	4.856000e-01	0.0080	Dummies_Lon_lat_LargestUsage	-6.202000e-01	-3.450100e+14	-1.598400e+00
22	[PropertyGFATotal, PropertyGFABuilding(s), Age...	1.0	Ridge(random_state=1337)	4.848000e-01	0.0030	Continues_Lon_lat_LargestUsage	-6.204000e-01	-3.446441e+14	-1.598400e+00
95	[PropertyGFATotal, PropertyGFABuilding(s), Age...	0.0	XGBRegressor(random_state=1337)	4.400000e-01	0.1852	Dummies_Lon_lat_LargestUsage	-6.274000e-01	-3.514656e+14	-1.731000e+00
94	[PropertyGFATotal, PropertyGFABuilding(s), Age...	1.0	XGBRegressor(random_state=1337)	4.400000e-01	0.1830	Dummies_Lon_lat_LargestUsage	-6.274000e-01	-3.514656e+14	-1.731000e+00
83	[PropertyGFATotal, PropertyGFABuilding(s), Age...	0.0	XGBRegressor(random_state=1337)	4.532000e-01	0.1956	Continues_Neighborhood_LargestUsage	-6.286000e-01	-3.501257e+14	-1.685200e+00
	...								
1	[PropertyGFATotal, PropertyGFABuilding(s), Age...	0.0	LinearRegression()	-8.025323e+24	0.0168	Continues_Neighborhood_UGFA	-2.798725e+11	-3.616580e+14	-1.617929e+25
5	[PropertyGFATotal, PropertyGFABuilding(s), Age...	0.0	LinearRegression()	-1.816020e+25	0.0144	Continues_Lon_lat_UGFA	-4.160329e+11	-3.572319e+14	-4.396692e+25
13	[PropertyGFATotal, PropertyGFABuilding(s), Age...	0.0	LinearRegression()	-1.249242e+26	0.0082	Dummies_Lon_lat_UGFA	-5.379766e+11	-3.549054e+14	-2.106209e+26
4	[PropertyGFATotal, PropertyGFABuilding(s), Age...	1.0	LinearRegression()	-3.508557e+25	0.0090	Continues_Lon_lat_UGFA	-7.274076e+11	-3.600080e+14	-1.057333e+26

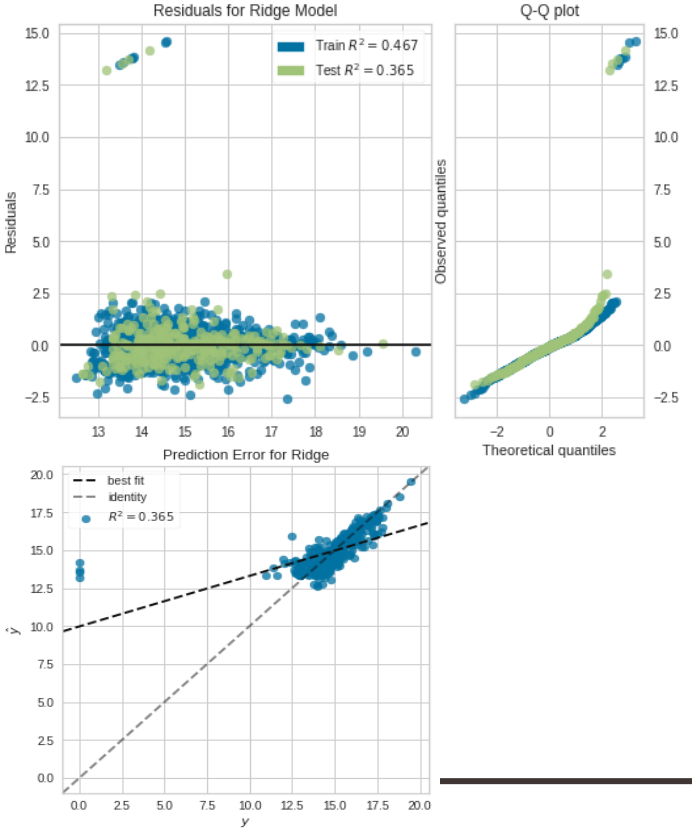
→ La régression Ridge avec log-transformation surperforme des modèles non linéaires

SiteEnergyUse : Régression Ridge

Recherche de la meilleure valeur d'alpha par GridSearch et CV, puis évaluation :

```
['PropertyGFATotal',
 'PropertyGFABuilding(s)',
 'Age',
 'MoreThanOneBuilding',
 'MoreThanOneFloor',
 'HasParking',
 'Longitude',
 'Latitude',
 'LargestPropertyUseType',
 'LargestPropertyUseTypeGFA',
 'MoreThanOneUse']
```

	param_alpha	mean_test_score	mean_score_time
9	2.5	-0.618393	0.000754
8	2	-0.618708	0.000865
7	1.5	-0.619337	0.000716
6	1.4	-0.619505	0.000671
5	1.25	-0.619785	0.000684
4	1.1	-0.620085	0.000662
3	1	-0.620313	0.000741
2	0.9	-0.620575	0.001112
1	0.75	-0.621070	0.000998
0	0.5	-0.622088	0.001228



MAE = 0.713

R^2 très faible (0.365)

Pistes d’amélioration

- Obtenir plus d’observations
- Obtenir plus de features
- Méthode de sélection des features plus avancée (LOFO)
- Enjeux de qualité des données

Merci

Avez-vous des questions?