PROJET 4 – ANTICIPEZ LES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS

Sommaire

Présentation de la problématique :

- Les données de consommation pour les bâtiments de la ville de Seattle pour l’année 2016 est disponible.

- L'obtention de relevés est coûteuse et fastidieuse.

La mission :

- L'objectif est de prédire les émissions de CO2 et la consommation totale d'énergie sans avoir recours aux relevés annuels.
- Évaluer l'intérêt du score ENERGY STAR.
- Mettre en place un modèle de prédiction réutilisable.

Interprétation de la problématique

- Données à prédire

- Consommation totale des bâtiments
SiteEnergyUseWN(kBtu)

- Emissions totales des bâtiments
TotalGHGEmissions

- **Score ENERGY STAR :**
  Comparaison de l'intérêt de la modélisation avec et sans.

Datas

```
print ("Le dataset compte {} lignes et {} variables".format(datas.shape[0], datas.shape[1]))
```

```
Le dataset compte 3376 lignes et 46 variables
```

```
#Nombre de valeurs manquantes totales
nb_nan_tot     = datas.isna().sum().sum()
nb_donnees_tot = np.product(datas.shape)
pourc_nan_tot  = round((nb_nan_tot/nb_donnees_tot)*100,2)

print(f'Valeurs manquantes : {nb_nan_tot} NaN pour {nb_donnees_tot} données ({pourc_nan_tot} %)'
```

```
Valeurs manquantes : 19952 NaN pour 155296 données (12.85 %)
```

2 – PRÉPARATION DU JEU DE DONNÉES

- Cleaning
- Exploration

Cleaning : Casse de certaines colonnes / informations quasi-identiques

	LargestPropertyUseType	SecondLargestPropertyUseType	ListOfAllPropertyUseTypes
1709	Multifamily Housing	NaN	Multifamily Housing
833	Multifamily Housing	Office	Multifamily Housing, Office
3244	Multifamily Housing	Parking	Multifamily Housing, Parking
2078	Medical Office	Office	Medical Office, Office

ListOfAllPropertyUseTypes = LargestPropertyUseType + SecondLargestPropertyUseType

	SiteEUI(kBtu/sf)	SiteEUIWN(kBtu/sf)	SourceEUI(kBtu/sf)	SourceEUIWN(kBtu/sf)	SiteEnergyUse(kBtu)	SiteEnergyUseWN(kBtu)
1043	52.4	62.9	93.00000	104.10000	1.25703e+06	1.51011e+06
1545	32.0	33.3	100.40000	104.50000	1.44554e+06	1.50492e+06
149	100.5	106.1	262.70001	272.20001	9.47236e+06	1.00055e+07
3245	50.1	52.4	109.80000	115.70000	3.34483e+06	3.49665e+06

Le projet stipule que seuls les édifices qui ne sont pas destinés à l'habitation seront examinés. Ainsi, nous allons éliminer toutes les lignes relatives aux logements en utilisant la variable BuildingType comme référence.

```
NonResidential              : 1460
Multifamily LR (1-4)        : 1018
Multifamily MR (5-9)        : 580
Multifamily HR (10+)        : 110
SPS-District K-12           : 98
Nonresidential COS          : 85
Campus                      : 24
Nonresidential WA           : 1
Total:                        3376
```

- Suppression des outliers

Le résultat final est un nouveau DataFrame avec 1668 lignes et 37 colonnes, dans lequel les lignes considérées comme des valeurs aberrantes ont été supprimées.

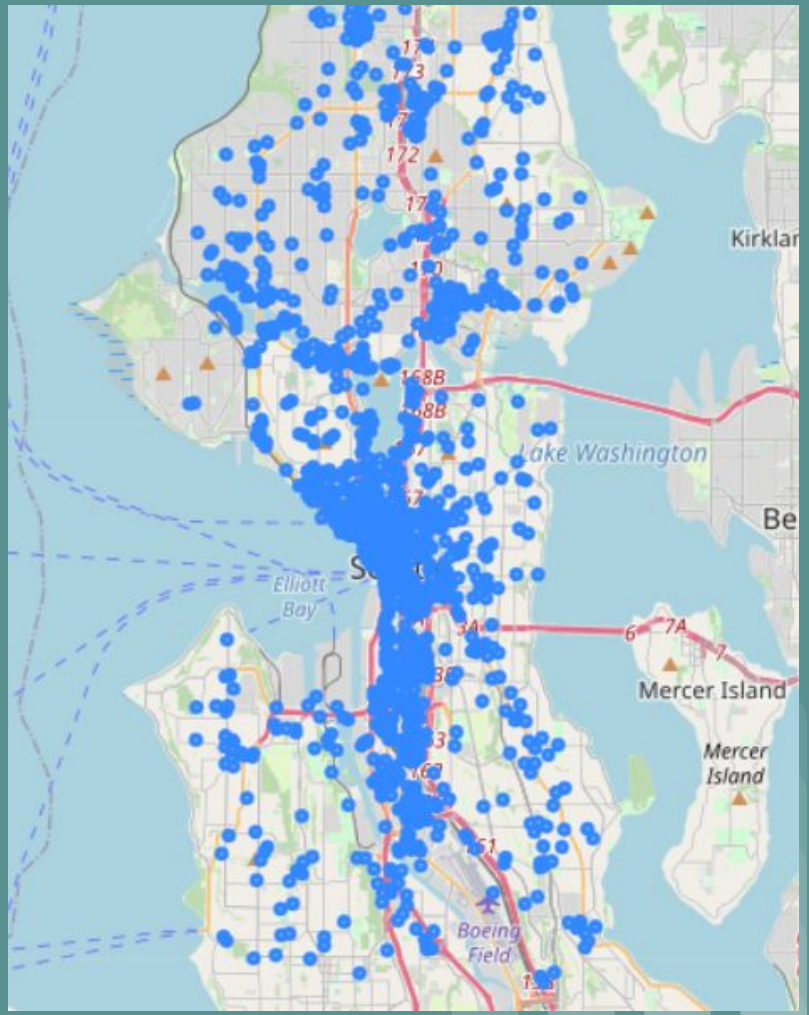
- Correction du nombre de bâtiments et d'étages (ne peut être nul ou NaN)

- Suppression des valeurs aberrantes

TotalGHGEmissions -0.80000
GHGEmissionsIntensity -0.02000

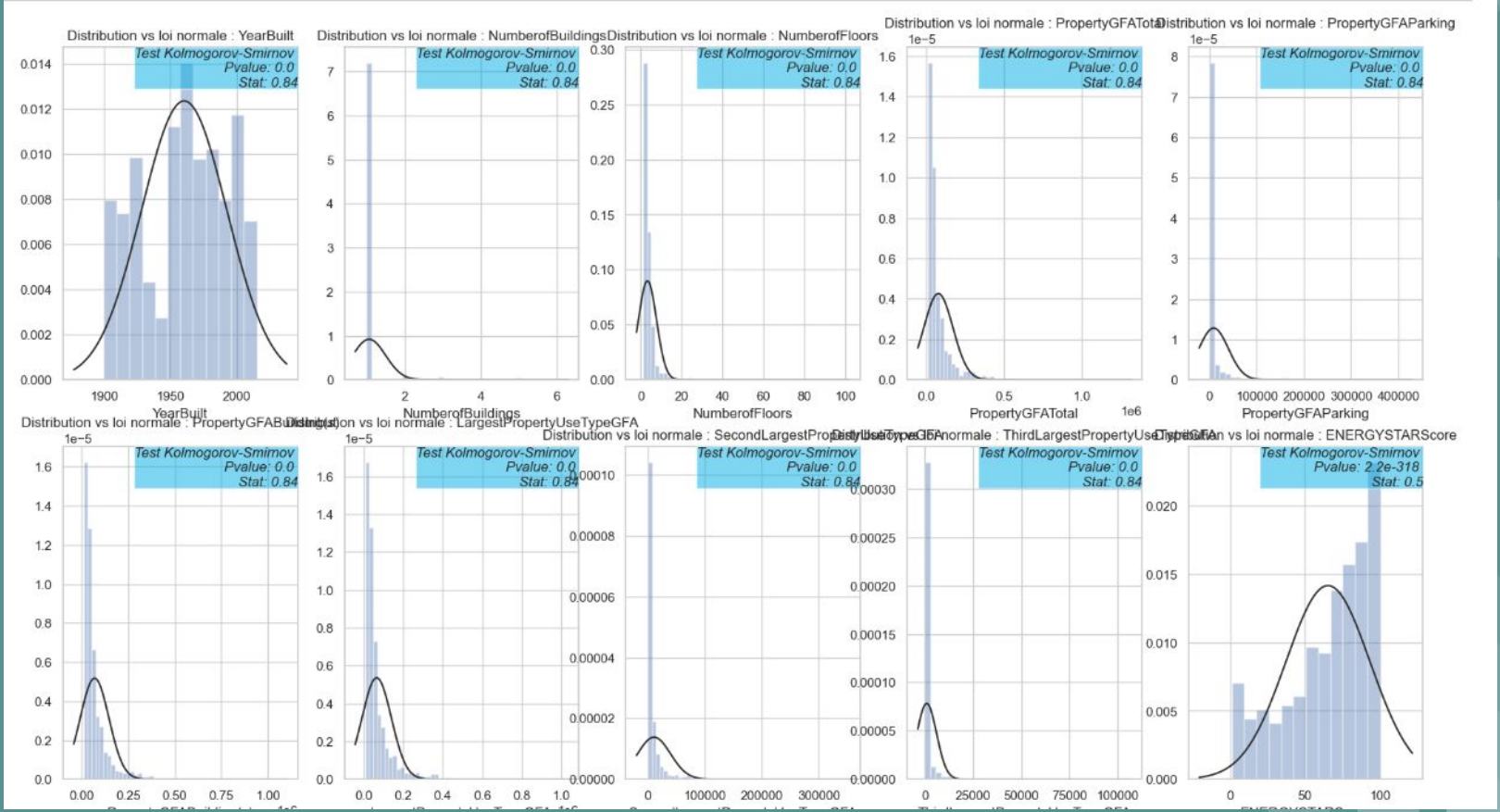
- NaN :
- complétion des valeurs manquantes quand applicable (e.g. catégories « No information»)
- Suppression des observations pour lesquelles on a beaucoup de NaN pour conserver un maximum de feature

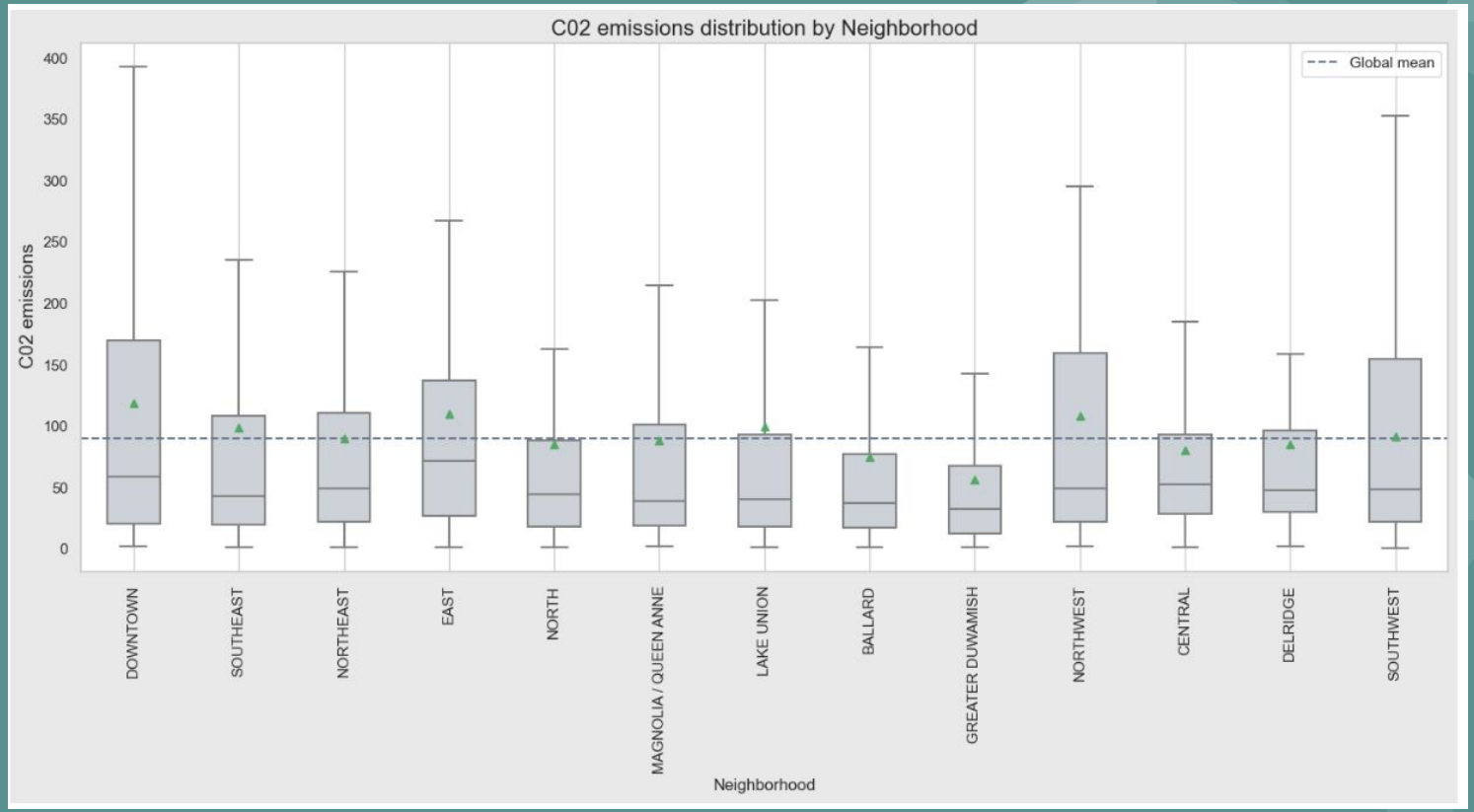
Les bâtiments sont localisés dans Seattle.



Exploration

Analisis univariée



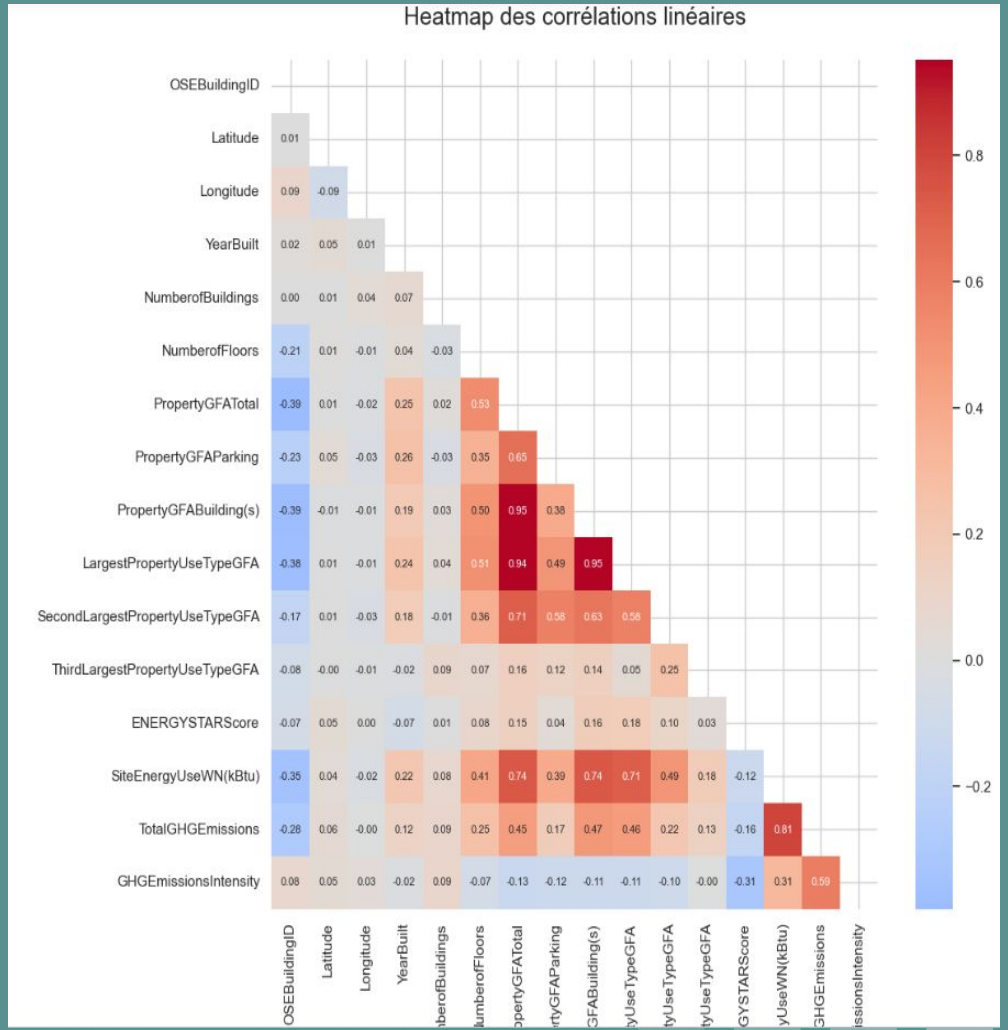

C02 emissions distribution by Neighborhood
Global mean
C02 emissions
400
350
300
250
200
150
100
50
0
DOWNTOWN
SOUTHEAST
NORTHEAST
EAST
NORTH
MAGNOLIA / QUEEN ANNE
LAKE UNION
BALLARD
GREATER DUWAMISH
NORTHWEST
CENTRAL
DELRIDGE
SOUTHWEST
Neighborhood

Observations

- Corrélation importante entre : PropertyGFATotal et PropertyGFABuilding(s) / LargestPropertyUseTypeGFA

- Corrélation importante entre PropertyGBABuilding(s) et LargestPropertyUseTypeGFA

- Corrélation importante entre TotalGHEmissions et SiteEnergyUseWN(kBtu) (on notera que ce sont les deux variables qu'on cherche à prédire, ce qui laisse présager



3. PISTES DE MODÉLISATIONS

Modèle consommation : démarche

Modèles entraînés : Elastic Net / SVR / Random Forest Regressor / XGBoost

1. Séparation jeu de données train/ validation/ test
2. Définition grille de paramètres
3. Entrainement des modèles
   - N modèles (toutes les combinaisons de paramètres)
   - Jeu training
   - Cross-validation
4. Comparaison des modèles sur la RMSE de validation

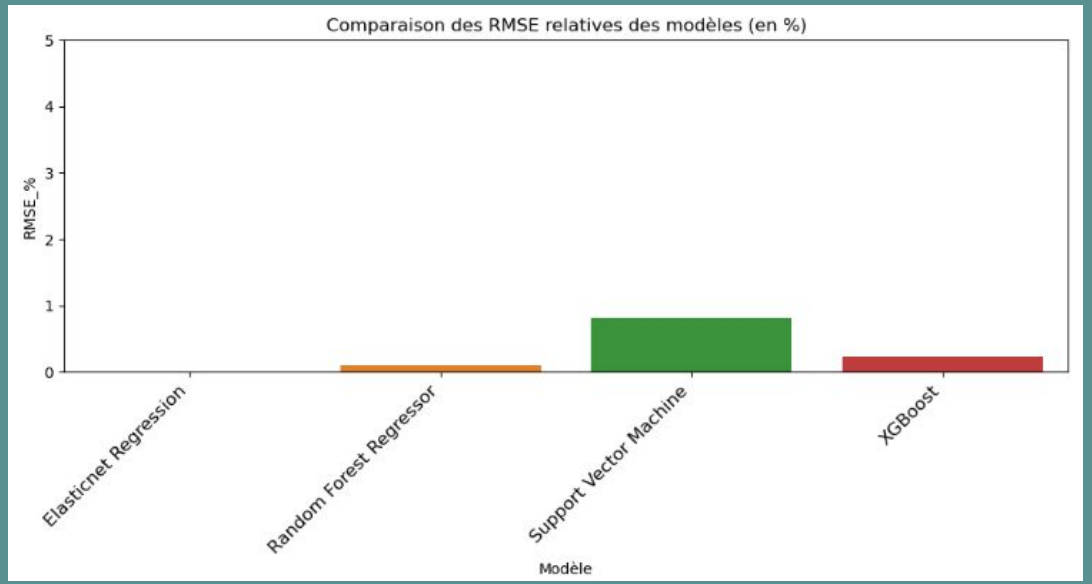
Modèle consommation : best_params_

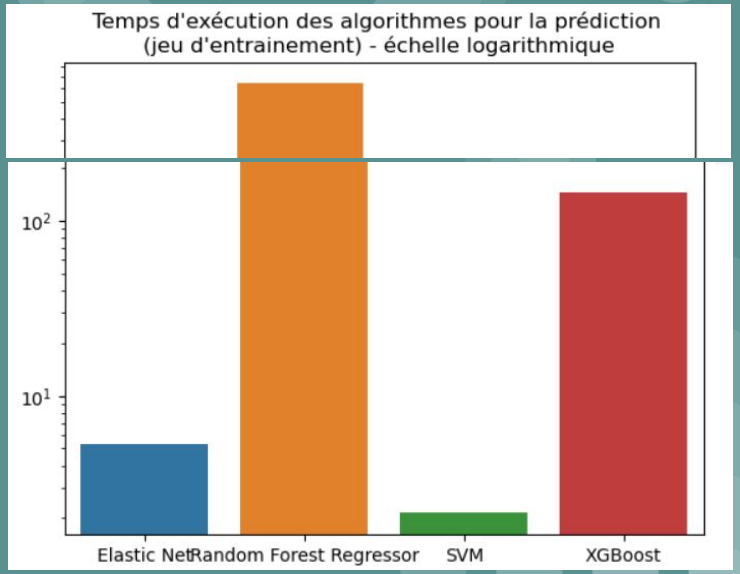
Elastic Net	SVR	XGBoost	Random Forest Regresso
'alpha': 0.0001	'gamma': 0.001	'n_estimators': [2000]	max_features': 'auto'
l1_ratio': 0.9	epsilon': [0.001]		min_samples_leaf': 1
tol': 0.0001	'C': [10]		n_estimators': 10

04. PRÉSENTATION DU MODÈLE FINAL

	Modèle	Score_RMSE	RMSE_%
0	Elasticnet Regression	0.00023	0.00001
1	Random Forest Regressor	0.02002	0.00097
2	Support Vector Machine	0.16984	0.00823
3	XGBoost	0.04843	0.00235

Le temps d'exécution du modèle **XGBoost** est significativement plus faible que celui du modèle **Random Forest** (environ 4 fois plus rapide). Cette différence de temps d'exécution peut être considérée comme un critère important lors du choix entre XGBoost et Random Forest Regressor, même si cela implique une légère dégradation des performances.





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): **0.1258**
- Entrainons un autre modèle avec les mêmes données **sans** le Energy Star Score: **0.11897**

MERCI DE VOTRE ATTENTION