Anticipez les besoins en consommation de bâtiments



Entreprise: RES Group

Ecole: OPENCLASSROOM

Divine Tulomba

Année: 2022/2023

SOMMAIRE

- · Présentation de la problématique
- Préparation du jeu de données
- Pistes de modélisations
- Présentation des résultats

Rappel de la problématique

Interprétation

Pistes de recherche envisagées

Problématique

Présentation de la problématique

- Données de consommation disponibles pour les bâtiments de la ville de Seattle pour l'année 2016
- Coût important d'obtention des relevés / Fastidieuses à collecter

La mission:

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

Présentation de la problématique

Prévision:

- > Features : Caractéristiques intrinsèques des bâtiments (hors consommations)
- > Données à prédire à partir de deux modèles différents:
 - Consommation totale des bâtiments SiteEnergyUseWN(kBtu)
 - Emissions totales des bâtiments TotalGHEmissions
- Energy Star Score :
 - Comparaison de son intérêt en essayant de modéliser avec et sans pour le meilleur modèle simplement

II - Préparation du jeu de données



Nettoyage de données



Feature Engineering



Exploration de données

Etude de la dataframe

Etude Générale de notre dataframe :

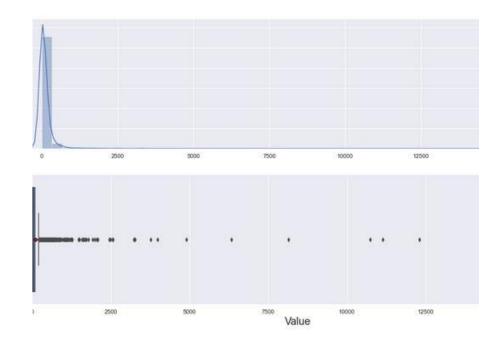
- Le dataframe a 3376 lignes et 46 colonnes
- Le dataframe a un taux de remplissage de 87.15%
- Il y a 19 952 valeurs manquantes sur 115 296

Nettoyage de données

Suppression de colonnes non pertinentes pour notre modèle, tells que :

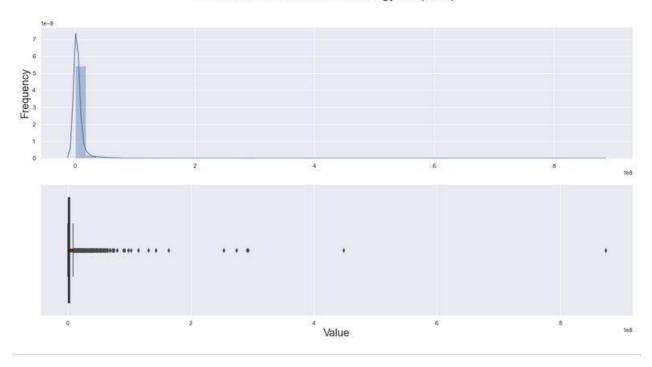
- Données sans catégorisation possible
- Données avec une unique information (exemple : State)
- Suppression de données non destinés à l'habitation
- Suppression de données issues des relevés de consommation annuels
- Élimination des colonnes trop peu remplies
- Données sans information pertinente pour le modèle :
 - DefaultDate : booléen avec beaucoup de NaN
 - SPD Beats : Informations non utiles à la problématique + beaucoup de NaN
 - Features redondantes (address/zipcode remplacées par latitude et longitude)

Statistical distribution: TotalGHGEmissions



Distributions empiriques des variables cibles

Statistical distribution: SiteEnergyUse(kBtu)



Distributions empiriques des variables cibles

Feature engineering

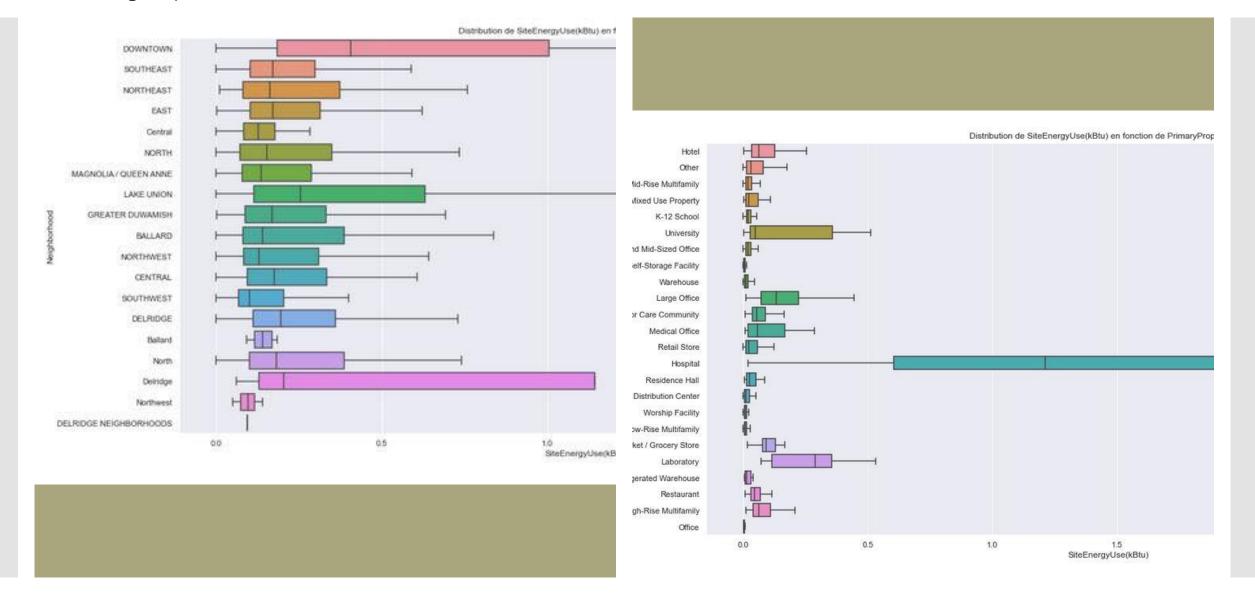
• Prévision :

- Idées écartées
 - Features liées à la proportion des sources d'énergie (coûteux à obtenir pour futures données)

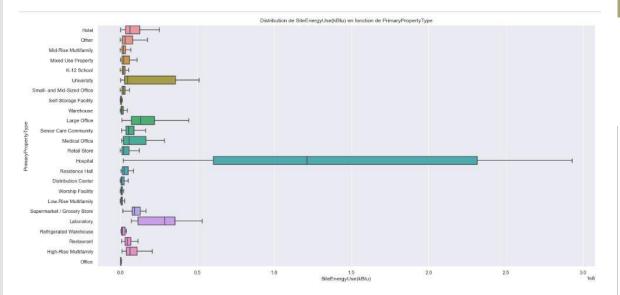
Idées retenues :

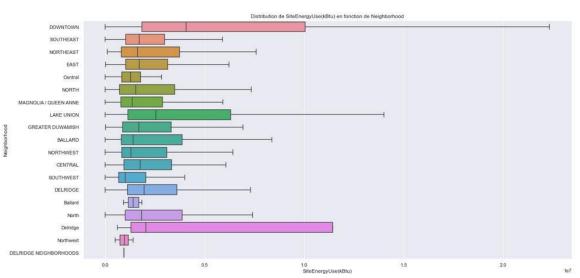
- Suppression des features de consommation (ormis les 2 features qu'on cherche à prédire)
- Utilisation du Energy Star score (mis de côté pour analyse ultérieure)
- Création de nouvelles features (nb_use_types, distance_center, BuildingAge)
- One hot encoding : Transformation d'une feature avec n catégories en n features booléennes.
- Log2-transformation variable de prédiction

Avant regroupement ...



Après regroupement ...





Exploration de données

Points Majeurs:

Points Majeurs

Pour des variables de consommation : Corrélation importante entre :

- PropertyGFATotal et PropertyGFABuildings
- PropertyGFATotal et LargestPropertyUseTypeGFA
- LargestPropertyUseTypeFGA et PropertyGFABuildings

Energy Star Score : pas de corrélation notable

Triangle Correlation Heatmap

NumberofBuildings	1	-0.026	0.69	-0.0047	-0.0049	0.69	-0.01	0.41	0.024	0.031	-0.014	0.0033	0.082	-0.0086
NumberofFloors	-0.026		0.4	0.42	0.023	0.21	-0.074	0.14	-0.15	-0.26	0.19	0.03	0.22	0.019
PropertyGFATotal	0.69	0.4	1	0.4	0.067	0.8	-0.019	0.53	-0.1	-0.098	0.16	-0.021	0.2	0.06
PropertyGFAParking	-0.0047	0.42	0.4	1	0.049	0.17	-0.043	0.089	-0.18	-0.14	0.21	0.015	0.082	0.037
ENERGYSTARScore	-0.0049	0.023	0.067	0.049	1	-0.09	0.1	-0.1	-0.029	-0.034	0.068	0.005	-0.03	-0.019
SiteEnergyUse(kBtu)	0.69	0.21	0.8	0.17	-0.09	1	-0.023	0.86	-0.027	-0.052	0.1	0.011	0.2	0.064
DefaultData	-0.01	-0.074	-0.019	-0.043	0.1	-0.023	1	-0.017	-0.0017	0.17	-0.11	-0.014	-0.029	0.042
TotalGHGEmissions	0.41	0.14	0.53	0.089	-0.1	0.86	-0.017	1	-0.013	-0.033	0.075	0.003	0.23	0.11
BuildingAge	0.024	-0.15	-0.1	-0.18	-0.029	-0.027	-0.0017	-0.013	1	-0.14	-0.094	-0.0082	0.15	0.096
distance_center	0.031	-0.26	-0.098	-0.14	-0.034	-0.052	0.17	-0.033	-0.14	1	-0.15	-0.02	-0.16	-0.087
nb_use_types	-0.014	0.19	0.16	0.21	0.068	0.1	-0.11	0.075	-0.094	-0.15	1	0.022	0.052	0.081
Electricity	0.0033	0.03	-0.021	0.015	0.005	0.011	-0.014	0.003	-0.0082	-0.02	0.022	1	0.013	0.036
SteamUse	0.082	0.22	0.2	0.082	-0.03	0.2	-0.029	0.23	0.15	-0.16	0.052	0.013	1	-0.02
NaturalGas	-0.0086	0.019	0.06	0.037	-0.019	0.064	0.042	0.11	0.096	-0.087	0.081	0.036	-0.02	1
	NumberofBuildings	NumberofFloors	PropertyGFATotal	PropertyGFAParking	ENERGYSTARScore	Site Energy Use (kBtu)	DefaultData	TotalGHGEmissions	BuildingAge	distance_center	nb_use_types	Bectricity	SteamUse	NaturalGas

III - Modélisations

Modèle de consommation : démarche

Séparation du jeu de données (train/validation et test)

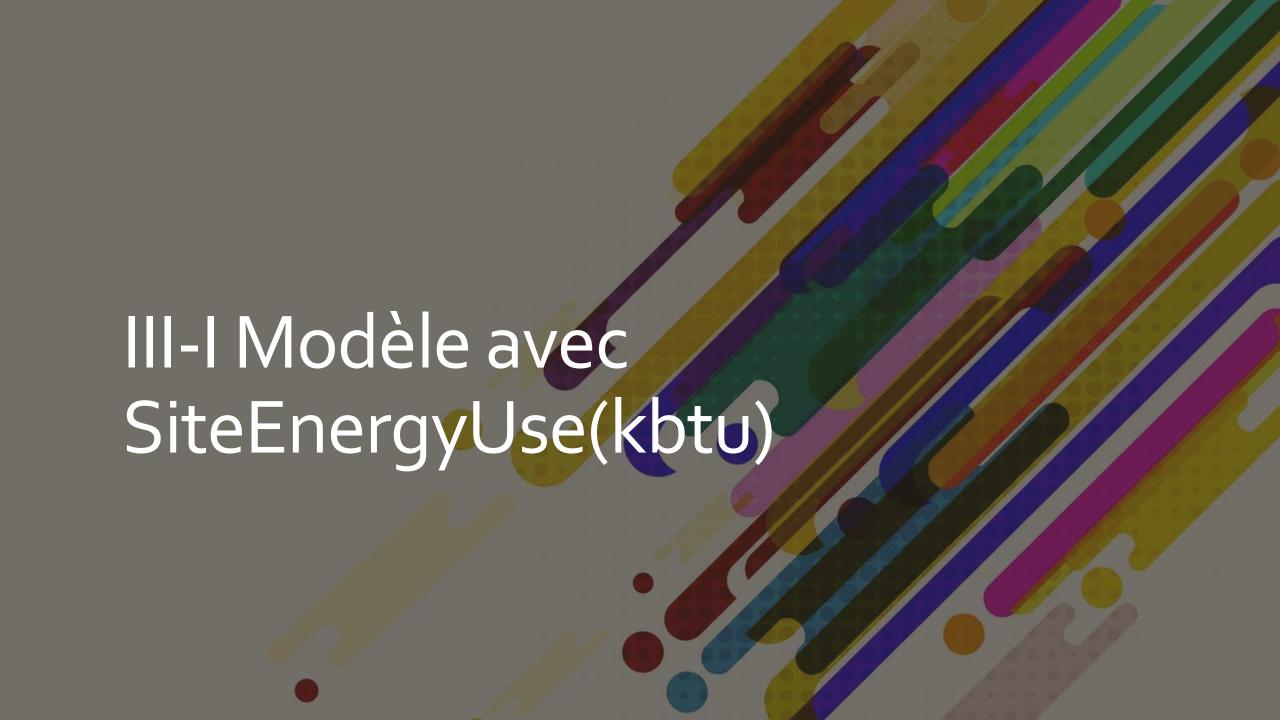


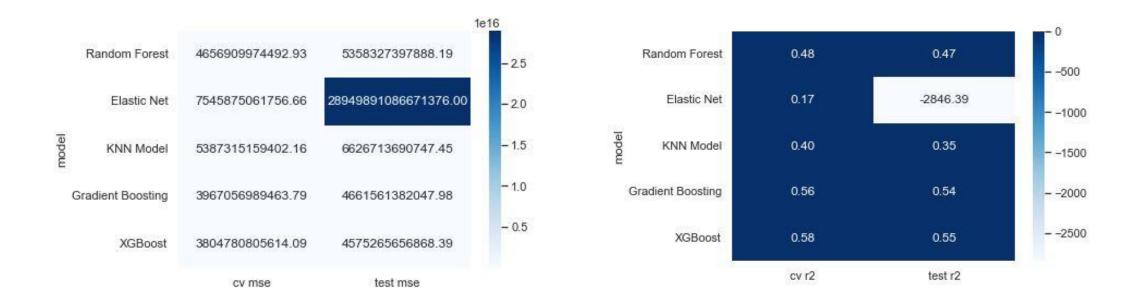
Pour chaque algorithme

- - Définition grille de paramètres
- Entrainement des modèles (N modèles, Jeu training, Crossvalidation)

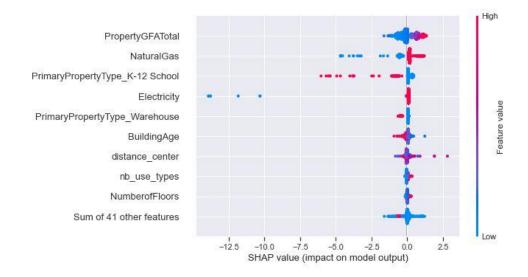


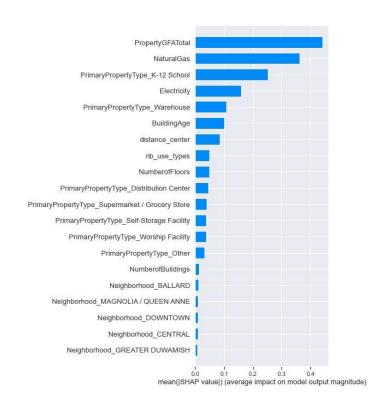
Comparaison des modèles sur la RMSE et le R2 de validation



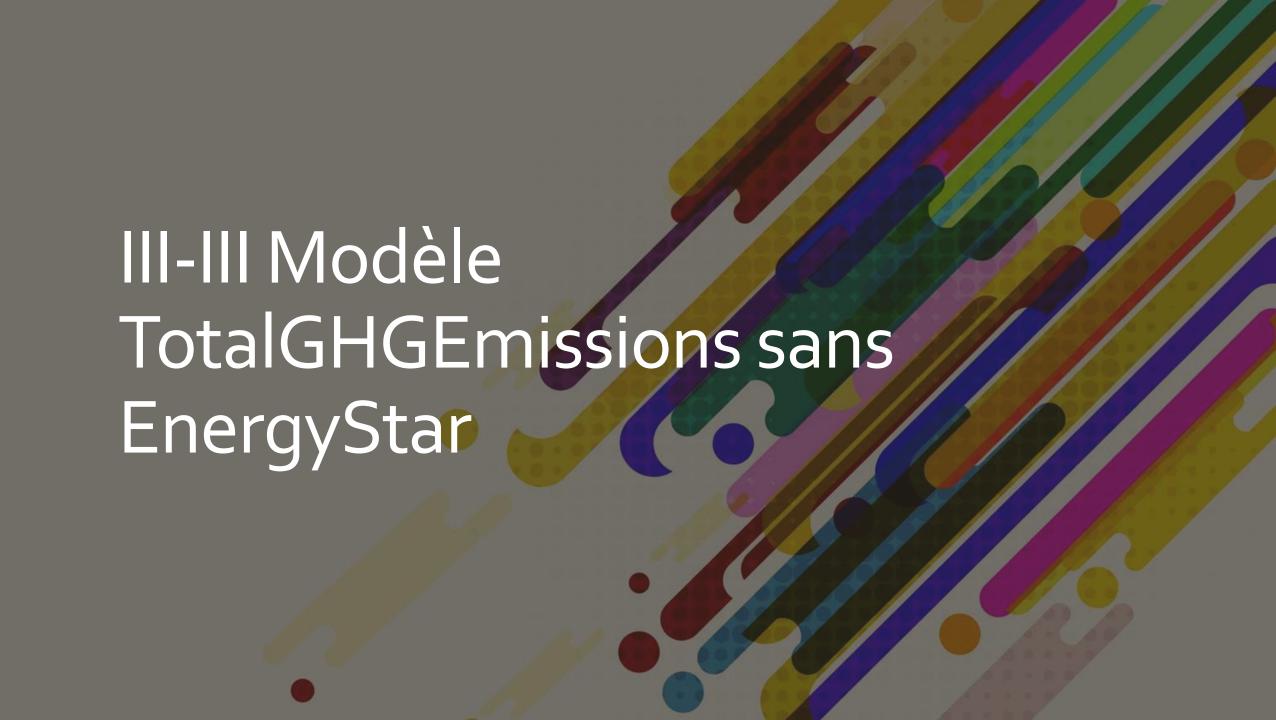


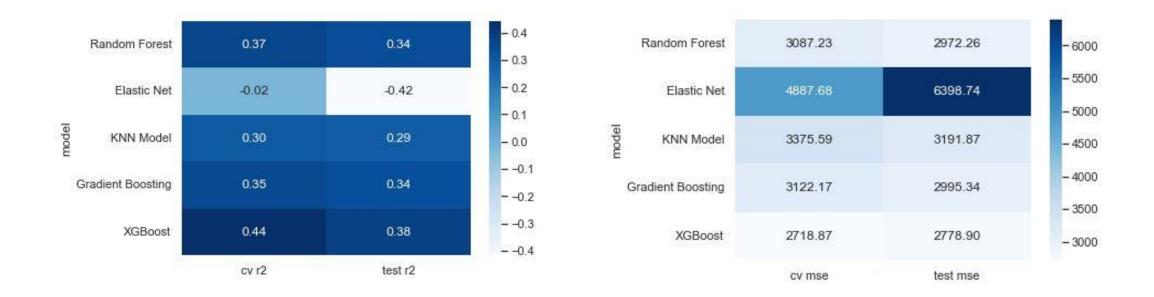
Modèle avec : SiteEnergyUse(kBtu)



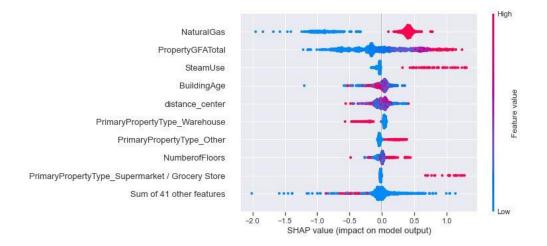


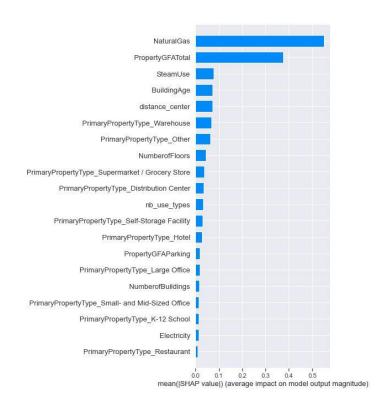
Interprétation globale du modèle avec SiteEnergyUse(kBtu)





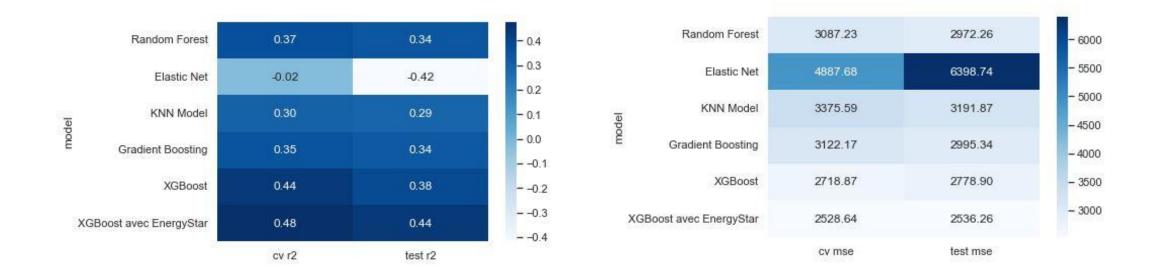
Modèle: SiteEnergyUse(kBtu)



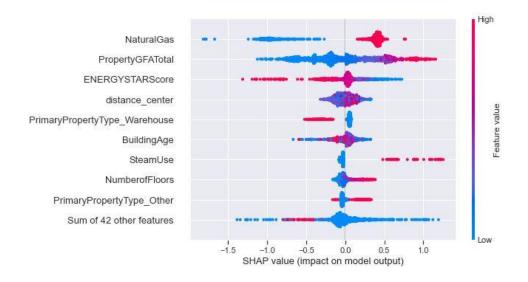


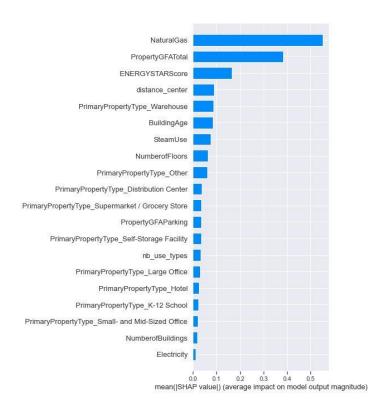
Interprétation globale du modèle avec SiteEnergyUse(kBtu)





Modèle: TotalGHGEmissions





Interprétation globale du modèle avec TotalGHGEmissions

Conclusion

Intérêt du Energy star Score

Le meilleur modèle

MERCIPOUR VOTRE ATTENTION