

Projet 4: Anticipez les besoins en consommation de bâtiments

Maodo FALL

OpenClassrooms

Soutenance du projet

19 juillet 2024

Sommaire

- 1 Analyse exploratoire des données
- 2 Prédiction de la consommation totale d'énergie
- 3 Prédiction de l'émission de gaz à effet de serre
- 4 Conclusion

Sommaire

- 1 Analyse exploratoire des données
- 2 Prédiction de la consommation totale d'énergie
- 3 Prédiction de l'émission de gaz à effet de serre
- 4 Conclusion

Problématique et jeu de données

Problématique et objectifs de la ville de Seattle :

- Ville neutre en émissions de carbone en 2050.

But :

- Prédire les émissions de CO₂ et la consommation totale d'énergie de bâtiments non destinés à l'habitation.

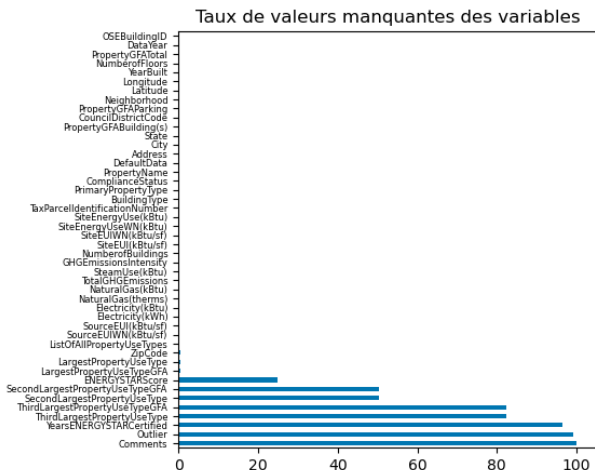
Le jeu de données et sa documentation sont disponibles sur le [site officiel](#) de la ville de Seattle.

Ville de Seattle



Seattle

Traitement des valeurs manquantes



- Nous avons supprimé les variables non renseignées à plus de 80%(sauf la variable Outlier).

Traitement des valeurs aberrantes

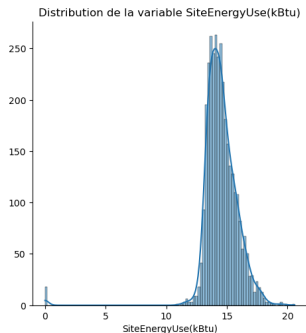
- La variable Outlier nous dit si une propriété est une valeur aberrante élevée ou faible (O/N). Nous les avons enlevées.

	Outlier	count	percent	cumulative_count	cumulative_percent
0	Low outlier	23	71.88%	23	71.88%
1	High outlier	9	28.12%	32	100.00%

Analyse univariée - un exemple

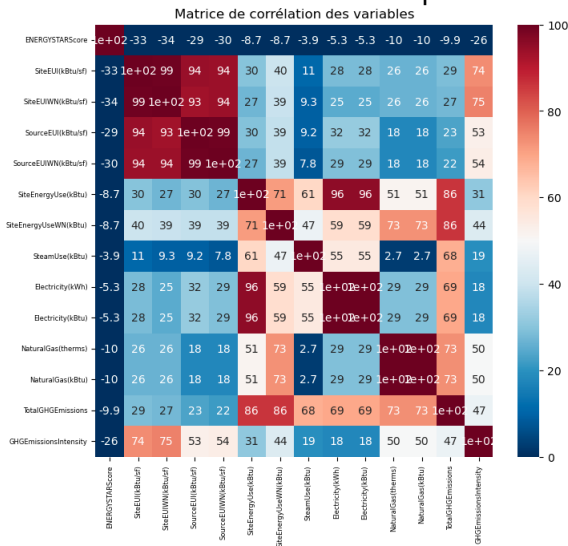
- Voici la description statistique et la distribution du log de la translation par 1 de la variable d'énergie SiteEnergyUse(kBtu) qu'on souhaite prédire.

```
mean : 5414823.429  
std : 21696835.466  
median : 1809101.5  
quartile q1: 935188.781  
quartile q3 : 4218830.5  
min : 0.0  
max : 873923712.0
```



Analyse bivariée

● Matrice de corrélation des variables de prédiction.



Sélection des variables de prédiction

- Bâtiments non destinés à l'habitation.
- Sélection des variables.
- Imputation des valeurs manquantes par k-plus proches voisins.
- Création de deux nouveaux fichiers à partir des données sélectionnées. L'un avec la variable ENERGYSTARScore et l'autre sans.
- Mise à l'échelle des variables.

Modèles de prédictions testés

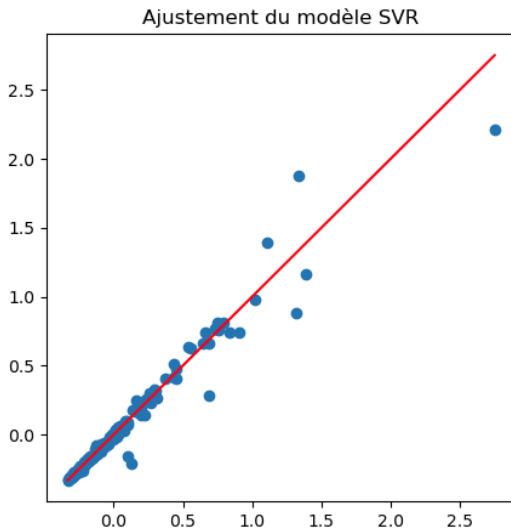
- Transformation de la variable d'énergie.
- Nous avons testé 4 modèles différents :
 - Régression linéaire.
 - Forêt aléatoire avec une grille de recherche des hyperparamètres.
 - Machine à vecteur de support avec une grille de recherche des hyperparamètres.
 - Renforcement extrême du gradient(Xgboost) avec une grille de recherche de ses hyperparamètres.

Récapitulation des résultats de prédiction

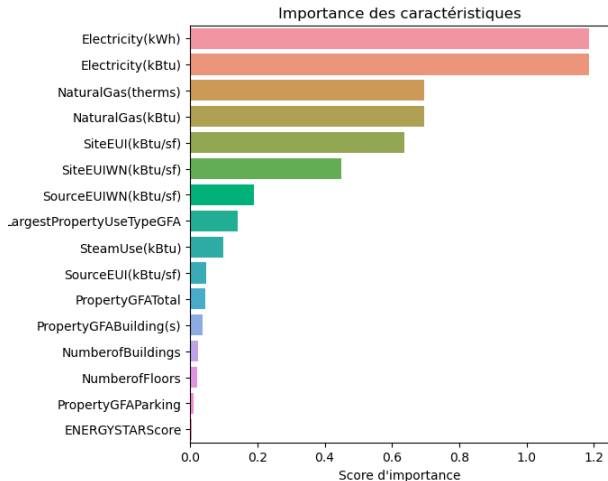
	R2 Scores	RMSE	MAPE
SVR	0.884755	0.109318	0.133832
XGBRegressor	0.960634	0.063891	0.150343
RandomForestRegressor	0.958231	0.065812	0.191058
LinearRegression	0.556621	0.214420	1.139679

- La plupart de ces modèles choisis prédisent bien la consommation totale d'énergie.
- Le modèle de machine à vecteur de support est néanmoins le plus performant pour la prédiction de la consommation totale d'énergie car sa MAPE est la plus petite.

Ajustement du modèle SVR



Importance des variables



Sommaire

- 1 Analyse exploratoire des données
- 2 Prédiction de la consommation totale d'énergie
- 3 Prédiction de l'émission de gaz à effet de serre
- 4 Conclusion

Prédiction avec présence de la variable ENERGYSTARScore

Nous avons effectué les mêmes étapes que précédemment :

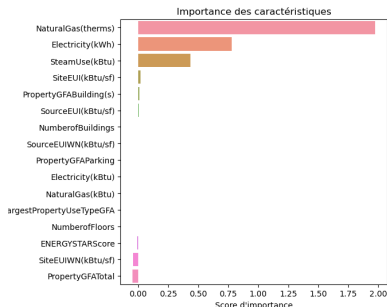
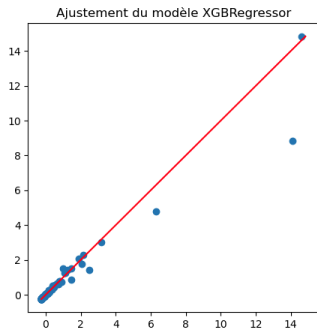
- On modélise le log de la translation par 1 de la variable d'émission de CO₂.
- Nous avons également essayé les mêmes modèles que ceux précédents.

Récapitulation

	R2 Scores	RMSE	MAPE
XGBRegressor	0.922182	0.098859	0.171940
SVR	0.778422	0.166818	0.182453
RandomForestRegressor	0.915122	0.103247	0.306277
LinearRegression	0.426781	0.268311	1.176550

- La plupart des modèles choisis prédisent assez fidèlement l'émission de CO₂, excepté le modèle de régression linéaire.
- Le modèle de régression par gradient boosting est le plus performant pour la prédiction de l'émission de gaz à effet de serre car sa MAPE est la plus petite.

Ajustement du modèle Xgboost - Importance des variables

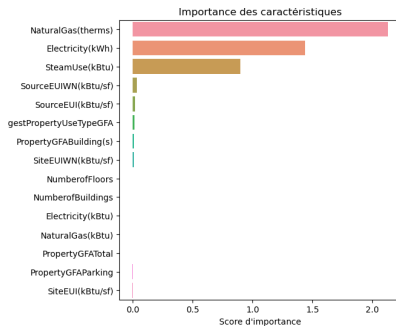
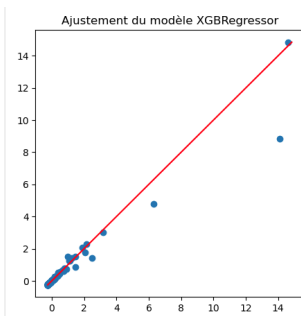


Prédiction sans la présence de la variable ENERGYSTARScore

	R2 Scores	RMSE	MAPE
XGBRegressor	0.936875	0.30659	0.116705

- En l'absence de la variable ENERGYSTARScore, la prédiction de l'émission de CO2 par le meilleur modèle est légèrement plus performante.

Ajustement



Sommaire

- 1 Analyse exploratoire des données
- 2 Prédiction de la consommation totale d'énergie
- 3 Prédiction de l'émission de gaz à effet de serre
- 4 Conclusion

Conclusion

- On arrive à bien prédire la consommation totale d'énergie des bâtiments non habitables.
- La prédiction de l'émission de gaz à effet de serre est également assez performante.
- Prédire l'émission de CO₂ avec ou sans la présence de ENERGYSTARScore n'a pas une grande différence de résultat.

MERCI BEAUCOUP !