



Projet 4

Anticipez les besoins en consommation de bâtiments

Zeynep Erdem
01-06-2023



Sommaire

- Rappel de la Problématique et Environnement
- Présentation et nettoyage des données
- Analyses Exploratoires des Données
- Présentation de démarche de modélisation
- Première itération de modélisation
- Feature Engineering
- Modélisation et évaluations des modèles
- Evaluation l'intérêt de l'Energy Star Score et feature engineering
- Conclusion et Recommandations

Rappel de la Problématique



- ❖ Atteindre la neutralité carbone à Seattle d'ici 2050.
- ❖ La consommation et l'émission des bâtiments non destinés à l'habitation.
- ❖ Obtenir des relevés est coûteux et chronophage pour chaque bâtiment.
- ❖ Utiliser des données de l'année 2016 pour prédire les émissions de CO2 et la consommation d'énergie de bâtiments non mesurés.
- ❖ Intégrer l'"ENERGY STAR Score" pour évaluer sa pertinence.

Environnement

- Python: 3.8.16
- Pandas: 1.5.3
- Numpy: 1.23.5
- Seaborn: 0.12.2
- Matplotlib: 3.7.1
- Missingno: 0.5.2
- Sklearn: 1.2.1

Présentation du Jeu de Données



Informations générales

Données structurelles

Relevés de
consommations et
d'émissions

Energy Star Score

- Il y a environ 3400 lignes et 46 colonnes dans notre dataframe
- On a 13 % de valeurs null dans notre dataframe
- On a constaté certaines incohérences

Les étapes du nettoyage



- Elimination des colonnes trop de valeurs manquantes
- SiteEnergyUseWN(kBtu)
- TotalGHGEmissions

- Suppression si moins de 5 %

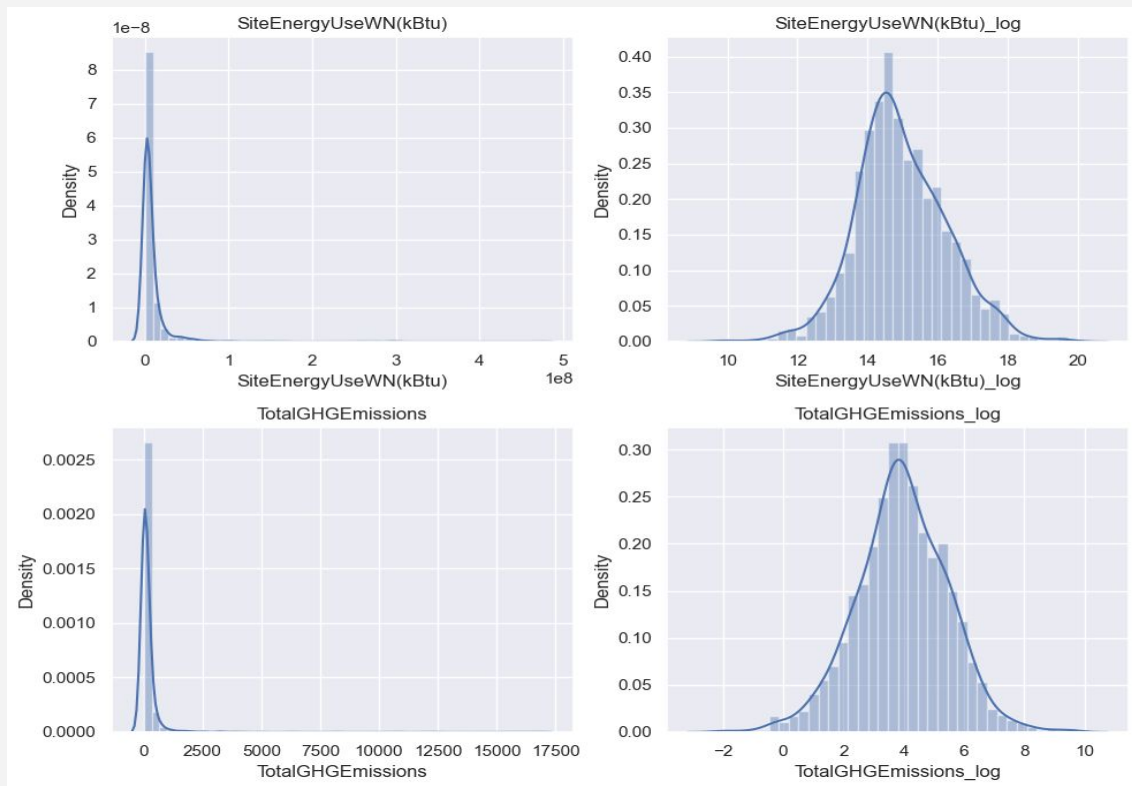
- Les consommation d'énergie négatives
- Nombre de bâtiments et étages zéro



Remplacé par nan et imputées avec médian de première use type

- Building age
- N of building(uni/multi)
- N of floors(high/low)

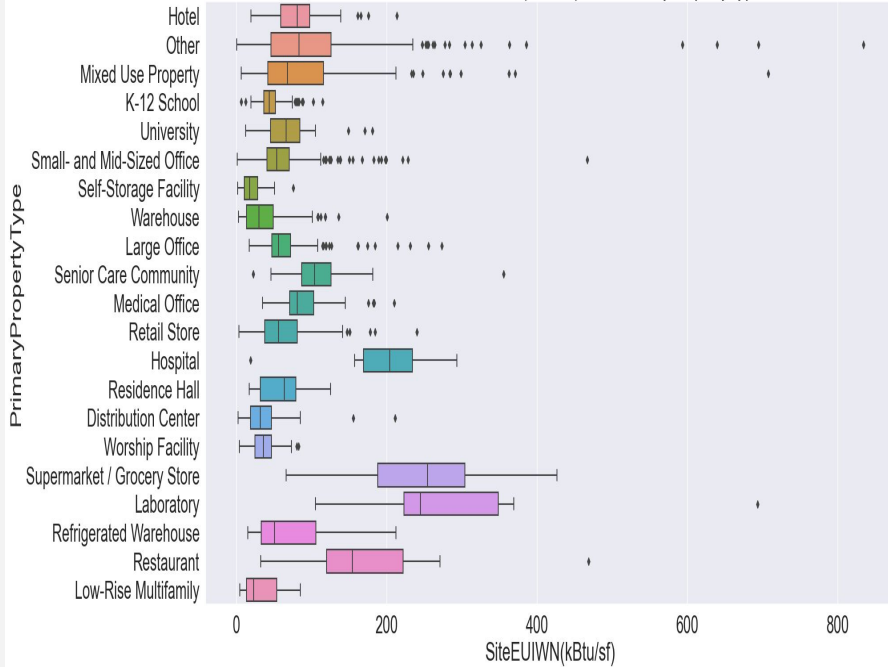
Analyses Exploratoires des Données



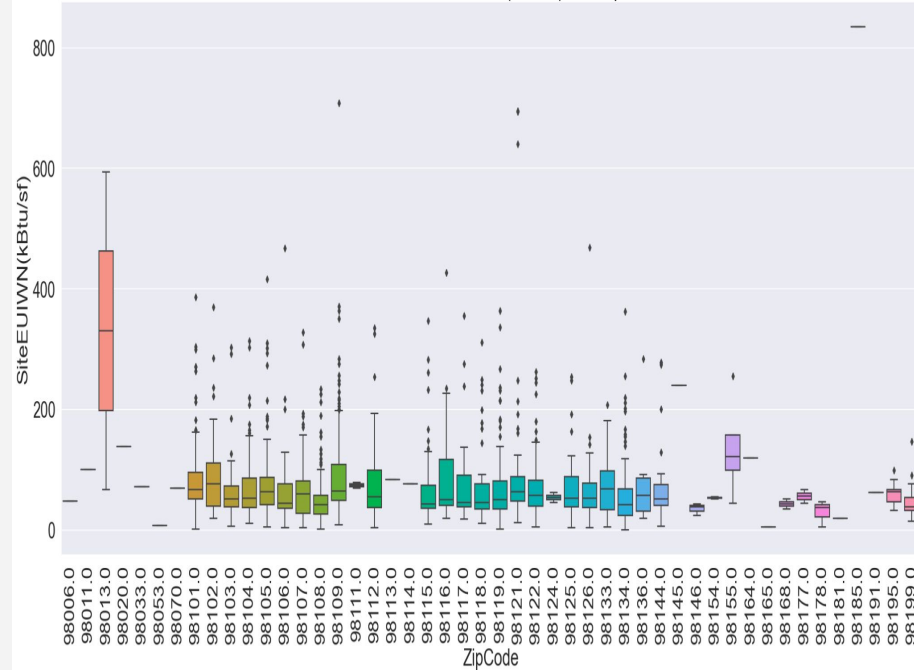
La transformation logarithmique des cibles permet d'obtenir une distribution plus proche de la normale.

Analyses Exploratoires des Données

Le lien entre le SiteEUIWN(kBtu/sf) et le PrimaryPropertyType

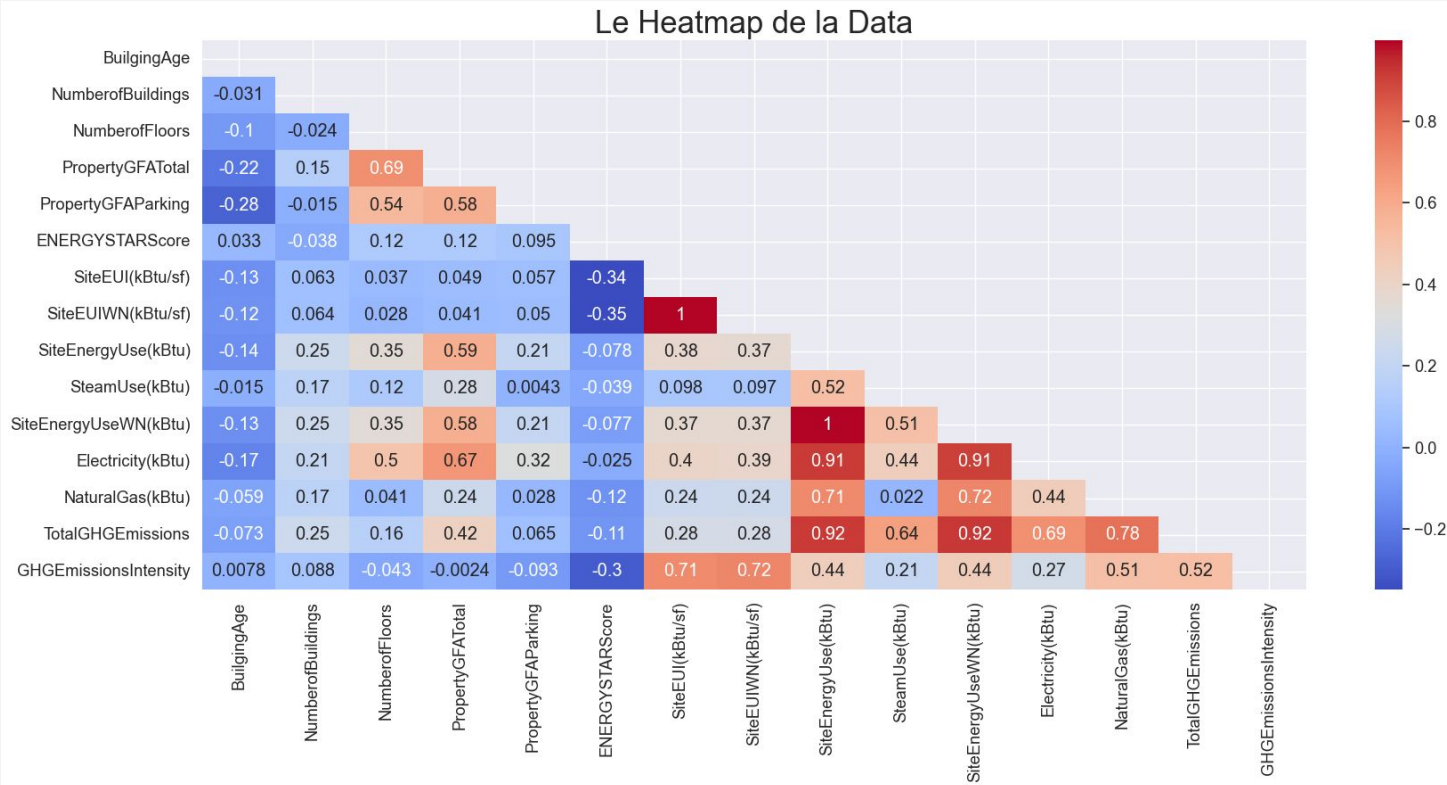


Le lien entre le SiteEUIWN(kBtu/sf) et le ZipCode



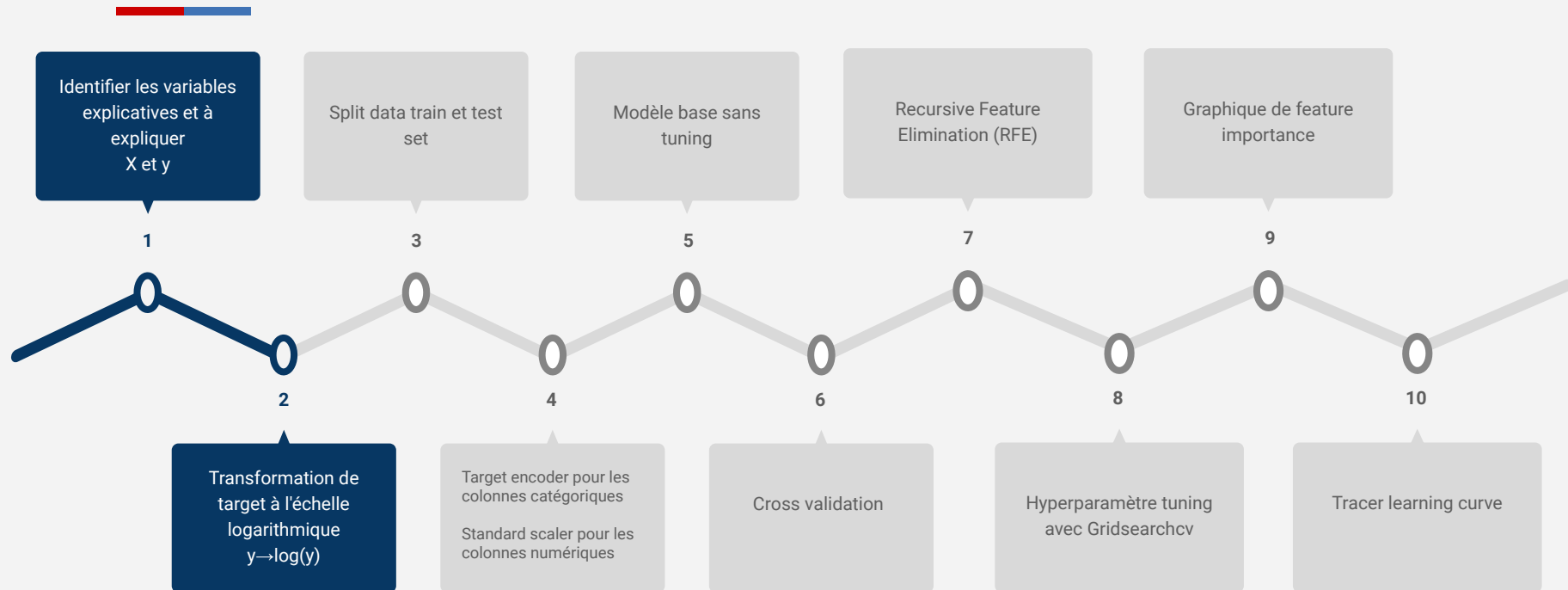
Les variables “Primary Use Type” et “Zipcode” ont un impact sur la consommation d'énergie.

Analyses Exploratoires des Données



- Attention aux variables fortement corrélées entre elles.
- Attention au risque de "data leakage" et aux variables endogènes.

La démarche de modélisation



Première itération de modélisation

Consommation d'énergie

model	r2_train	cv_mean	r2_test	mean_absolute_error	mean_absolute_percentage_error	RMSE	training_time
Random Forest Tuning	0.75	0.67	0.67	0.53	0.04	0.72	0.078022
Adaboost Tuning	0.66	0.63	0.63	0.55	0.04	0.76	0.070014
SVR Tuning	0.69	0.61	0.60	0.58	0.04	0.79	0.094484
LinearRegression RFE	0.53	0.52	0.51	0.65	0.04	0.87	0.002565
LassoCV	0.53	0.52	0.51	0.65	0.04	0.87	0.023320

Emission de gaz

model	r2_train	cv_mean	r2_test	mean_absolute_error	mean_absolute_percentage_error	RMSE	training_time
Random Forest Tuning	0.62	0.48	0.49	0.83	0.41	1.04	0.050393
Adaboost Tuning	0.50	0.43	0.41	0.92	0.41	1.12	0.299950
SVR Tuning	0.50	0.39	0.44	0.85	0.45	1.09	0.091453
LinearRegression RFE	0.39	0.37	0.38	0.90	0.44	1.15	0.002128
LassoCV	0.38	0.37	0.38	0.89	0.44	1.15	0.023930

Les variables

1. BuildingType
2. PrimaryProperty Type
3. ZipCode
4. Latitude
5. Longitude
6. BuilgingAge
7. PropertyGFATotal
8. NofBuildings
9. NofFloors

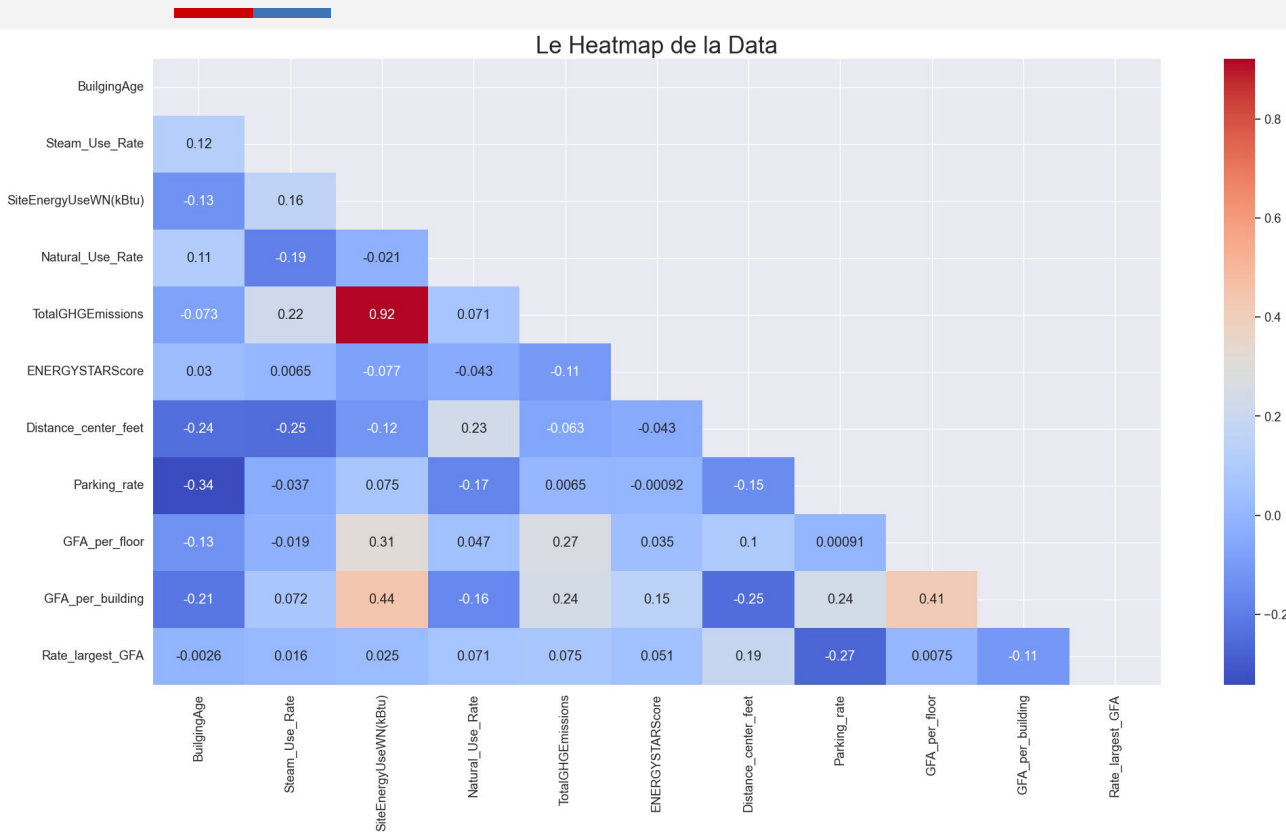


Feature Engineering

“ Même si tu as de meilleures machines, si tu n'as pas de bons ingrédients de qualité, tu ne peux pas faire un bon gâteau. C'est le même principe en machine learning. ”

Mon mentor :)

Feature Engineering



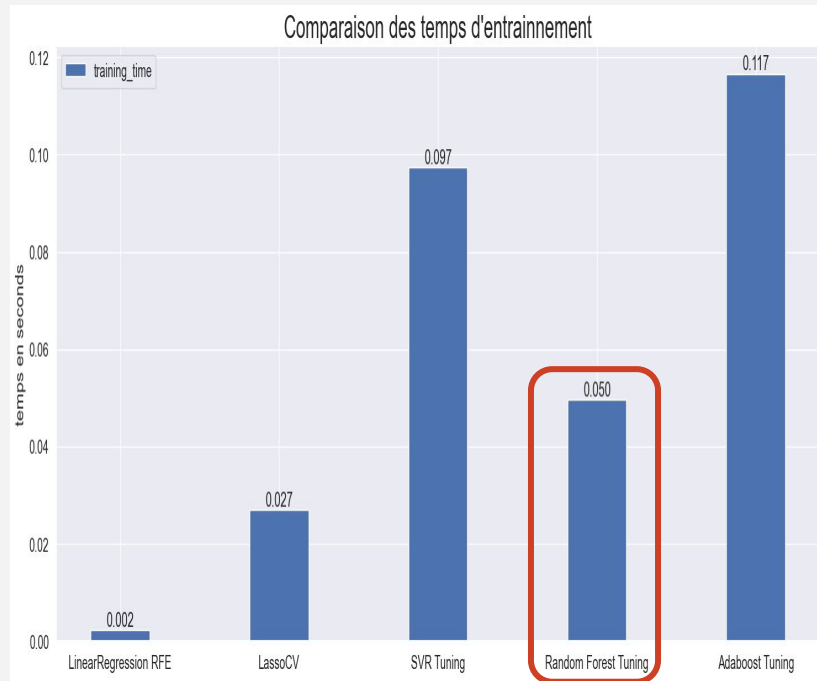
Les variables

1. BuildingType
2. PrimaryProperty Type
3. BuildingAge
4. ZipCode
5. LargestPropertyUseType

Feature Engineering

6. Distance_center_feet
7. Parking_rate
8. Rate_largest_GFA
9. GFA_per_floor
10. GFA_per_building
11. Steam_Use_Rate
12. Natural_Use_Rate

Les résultats pour la consommation d'énergie



On a choisi le modèle Random Forest en fonction de sa performance R2 et de son temps d'entraînement.

Présentation de la démarche de modélisation avec Random Forest



Cross val R2 values for model RandomForestRegressor : [0.71 0.7 0.67 0.7 0.69]

	model	r2_train	cv_mean	r2_test	mean_absolute_error	mean_absolute_percentage_error	RMSE	training_time
0	Random Forest sans tuning	0.96	0.69	0.68	0.51	0.03	0.69	0.501772

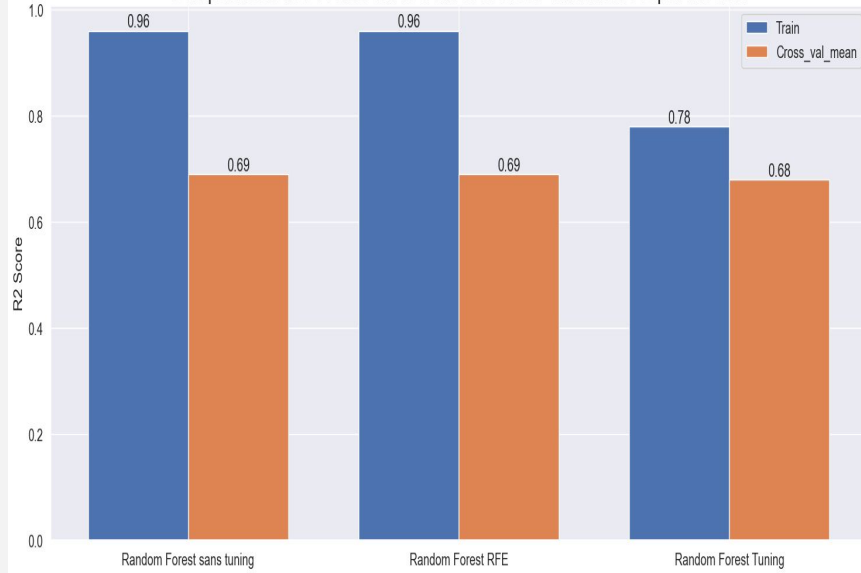


- Les résultats des validations croisées sont proches et cela est rassurant.
- À partir de 8 variables, les résultats ne changent pas beaucoup.

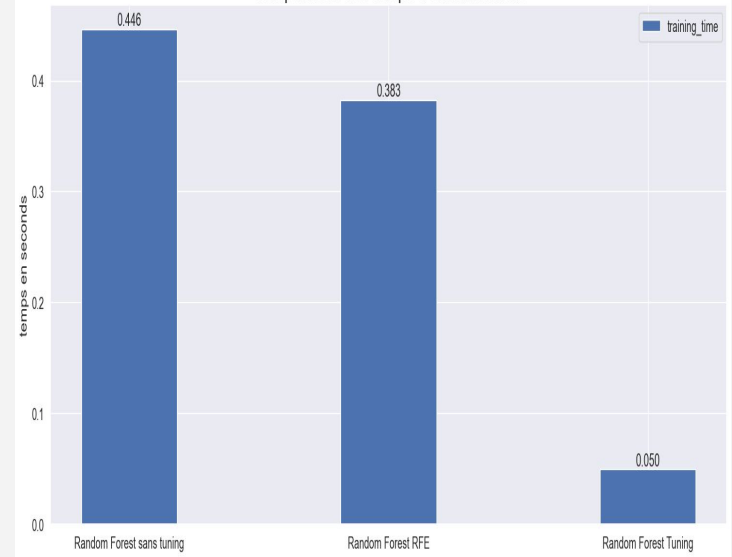
Présentation de la démarche de modélisation avec Random Forest



Comparaison des scores sur le train et le cross validation set par modèle

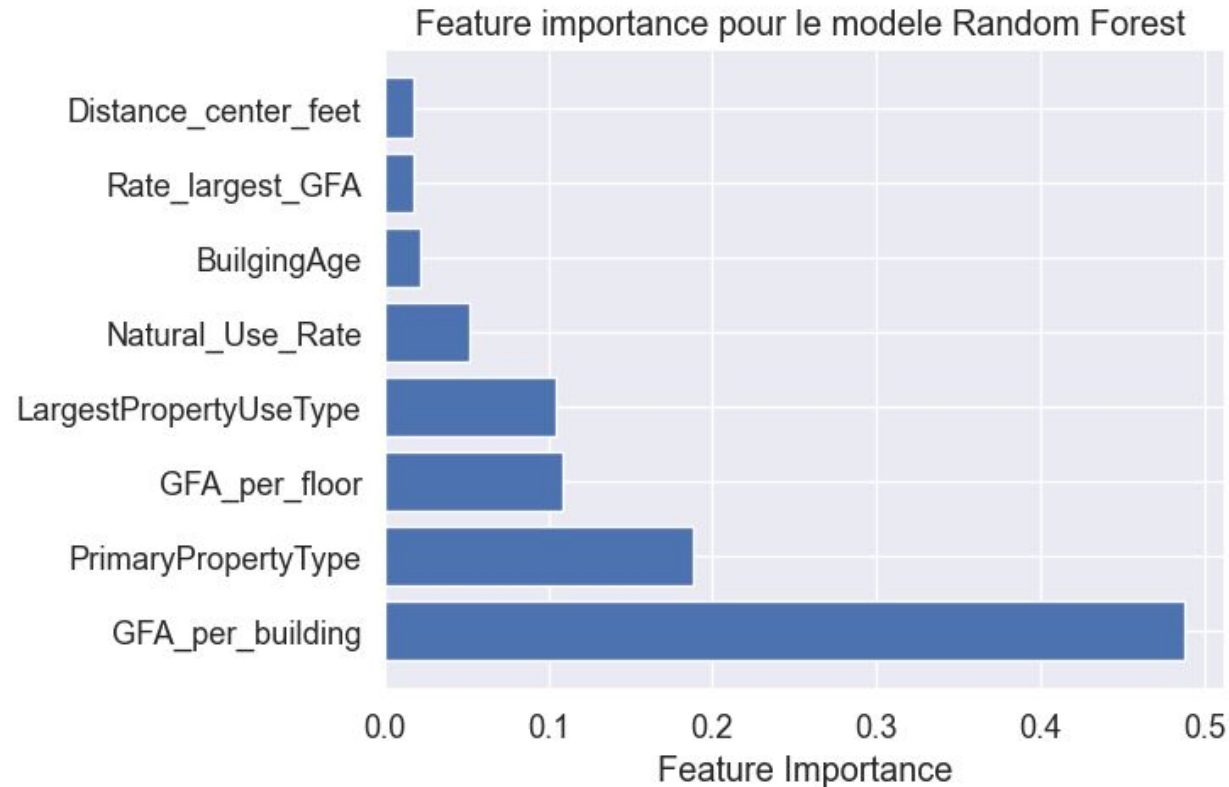


Comparaison des temps d'entraînement



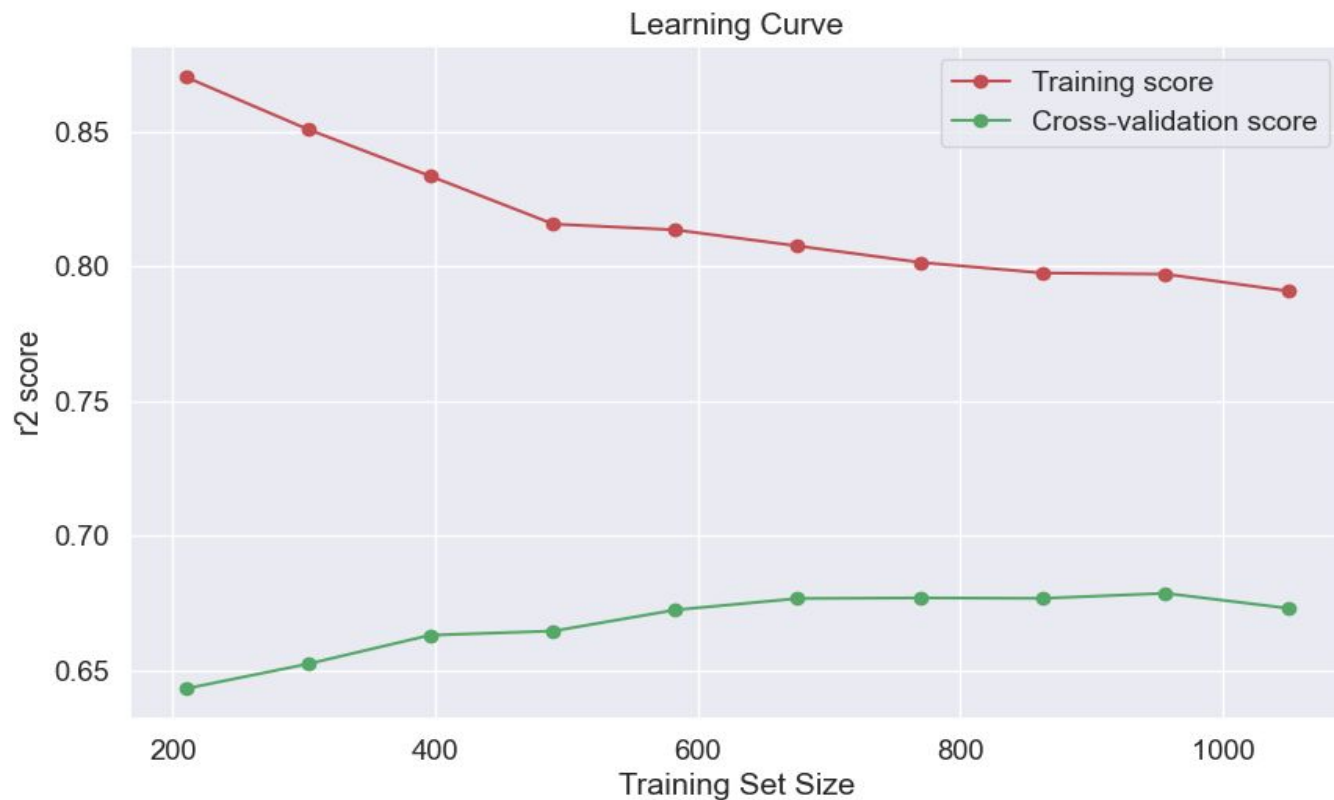
Les démarches RFE (Recursive Feature Elimination) et GridSearch améliorent la performance du modèle en réduisant l'overfitting et le temps d'entraînement.

Présentation de la démarche de modélisation avec Random Forest



- Les variables les plus importantes sont la surface par bâtiment, le type d'usage de bâtiment et la surface par étage.
- La distance au centre-ville n'est pas très significative.

Présentation de la démarche de modélisation avec Random Forest







Il y a un léger overfitting.

Best hyperparameters

- bootstrap: True
- max_depth: 6
- max_features: 0.5
- min_samples_leaf: 2
- n_estimators: 50

R2 test score:
0.68

Evaluation l'intérêt de l'Energy Star Score pour consommation d'énergie

	model	r2_train	cv_mean	r2_test	mean_absolute_error	mean_absolute_percentage_error	RMSE	training_time
	LinearRegression RFE No Star	0.68	0.67	0.63	0.57	0.04	0.73	0.002161
	LinearRegression RFE Star	0.75	0.73	0.73	 0.47	0.03	0.62	0.002140
	Random Forest No Star Tuning	0.87	0.77	0.76	0.44	0.03	0.59	0.097126
	Random Forest Star Tuning	0.90	0.82	0.82	 0.38	0.03	0.51	0.097549

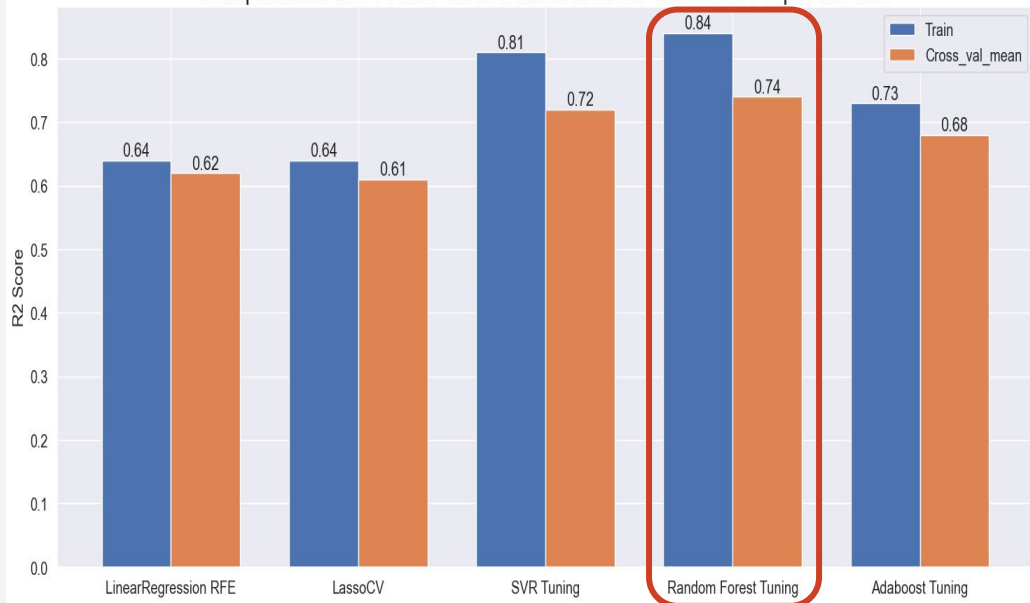
- On a supprimé les valeurs manquantes pour Energy Star Score.
- On a recalculé les résultats pour ce nouveau dataframe.
- On a entraîné deux modèles, la régression linéaire et le Random Forest.
- On a constaté une (légère) amélioration des scores en incluant la variable Energy Star Score dans nos modèles.

Les résultats de l' émission de gaz

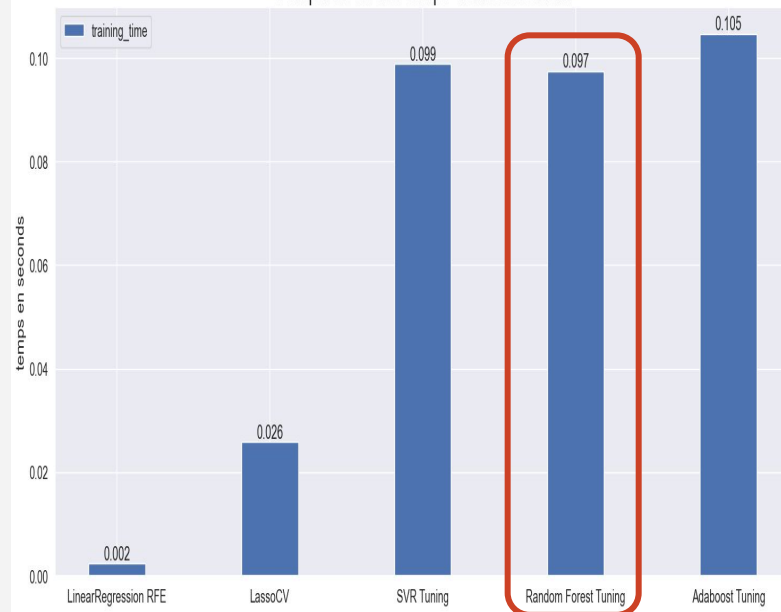


Les résultats pour l'émission de gaz

Comparaison des scores sur le train et le cross validation set par modèle

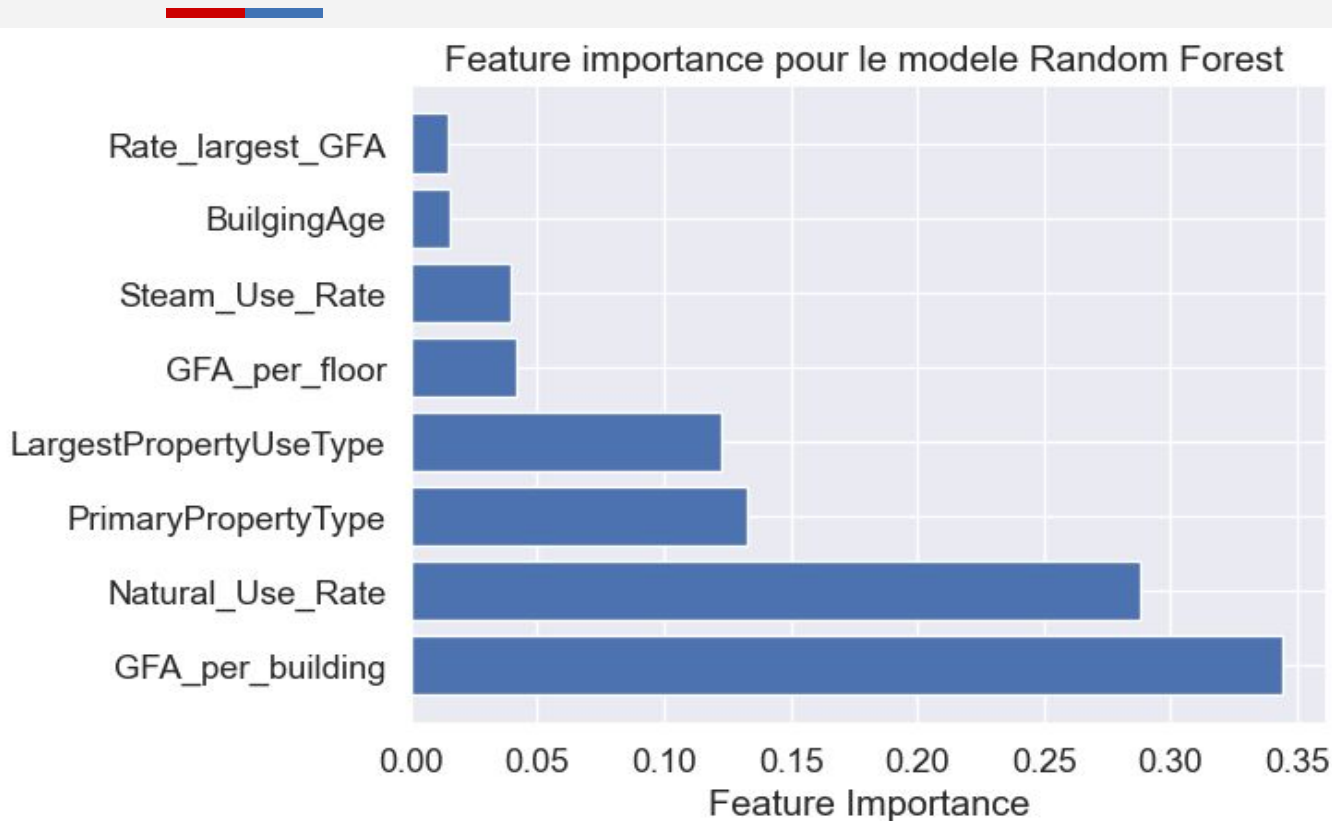


Comparaison des temps d'entraînement



On a sélectionné le modèle Random Forest en fonction de sa performance R2 et de son temps d'entraînement.

Feature Importance et Best Paramètres de Random Forest







Les variables les plus importantes sont la surface par bâtiment et le taux d'utilisation gaz naturel.

Best hyperparameters

- bootstrap: True
- max_depth: 6
- max_features: 0.8
- min_samples_leaf: 2
- n_estimators: 150

R2 test score:
0.68

Evaluation l'intérêt de l'Energy Star Score pour émission de gaz

	model	r2_train	cv_mean	r2_test	mean_absolute_error	mean_absolute_percentage_error	RMSE	training_time
	LinearRegression RFE No Star	0.72	0.71	0.72	0.63	0.25	0.79	0.002329
	LinearRegression RFE Star	0.75	0.73	0.73	 0.47	0.03	0.63	0.002486
	Random Forest No Star Tuning	0.88	0.79	0.78	0.52	0.18	0.69	0.051344
	Random Forest Star Tuning	0.90	0.82	0.82	 0.38	0.03	0.51	0.096982

- On a supprimé les valeurs manquantes pour Energy Star Score.
- On a recalculé les résultats pour ce nouveau dataframe.
- On a entraîné deux modèles, la régression linéaire et le Random Forest.
- Les scores obtenus avec Energy Star Score sont presque les mêmes que ceux obtenus sans cette variable.

Evaluation l'intérêt de feature engineering

Consommation d'énergie

Avant

model	r2_train	cv_mean	r2_test
Random Forest Tuning	0.75	0.67	0.67
Adaboost Tuning	0.66	0.63	0.63
SVR Tuning	0.69	0.61	0.60
LinearRegression RFE	0.53	0.52	0.51
LassoCV	0.53	0.52	0.51

Après

model	r2_train	cv_mean	r2_test
Random Forest Tuning	0.78	0.68	0.68
SVR Tuning	0.76	0.66	0.65
Adaboost Tuning	0.68	0.62	0.51
LinearRegression RFE	0.58	0.56	0.54
LassoCV	0.58	0.55	0.54

Emission de gaz

Avant

model	r2_train	cv_mean	r2_test
Random Forest Tuning	0.62	0.48	0.49
Adaboost Tuning	0.50	0.43	0.41
SVR Tuning	0.50	0.39	0.44
LinearRegression RFE	0.39	0.37	0.38
LassoCV	0.38	0.37	0.38

Après

model	r2_train	cv_mean	r2_test
Random Forest Tuning	0.84	0.74	0.68
SVR Tuning	0.81	0.72	0.72
Adaboost Tuning	0.73	0.68	0.58
LinearRegression RFE	0.64	0.62	0.61
LassoCV	0.64	0.61	0.61

Le feature engineering n'a pas eu un impact majeur sur la prédiction de la consommation, mais il a considérablement amélioré la performance pour la prédiction des émissions de gaz.

Conclusion et Recommandations



- On a choisi le modèle Random Forest pour effectuer deux prédictions avec différents hyperparamètres.
 - On a obtenu des scores moyens pour les deux modèles avec un léger overfitting (surajustement).
 - On a observé que le tuning des hyperparamètres améliore les résultats et le temps d'entraînement des modèles.
 - Le feature engineering a un impact sur la prédiction de l'émission de gaz.
 - La variable "Energy Star Score" améliore les scores pour la consommation d'énergie, mais elle a peu d'impact sur les émissions de gaz, ce qui la rend non indispensable.
-
- On peut améliorer nos résultats en utilisant des techniques de feature engineering plus sophistiquées, telles que l'utilisation de bibliothèques comme featuretools.
 - On peut également améliorer le réglage de nos modèles.
 - On peut essayer de collecter des données supplémentaires.

Merci

