Seattle

Anticipez les besoins en consommation de bâtiments

Soutenance Projet 4 - OPENCLASSROOMS
Samira MAHJOUB
27/04/2024

SOMMAIRE

- **Problématique**
- **Données**
- **Modélisation**
- **Conclusion**

Problématique - Contexte

Moyens mis à disposition:

Les relevés des bâtiments effectués en 2016

Objectif de ville Seattle :

Neutralité Carbonne en 2050.

Problématique:

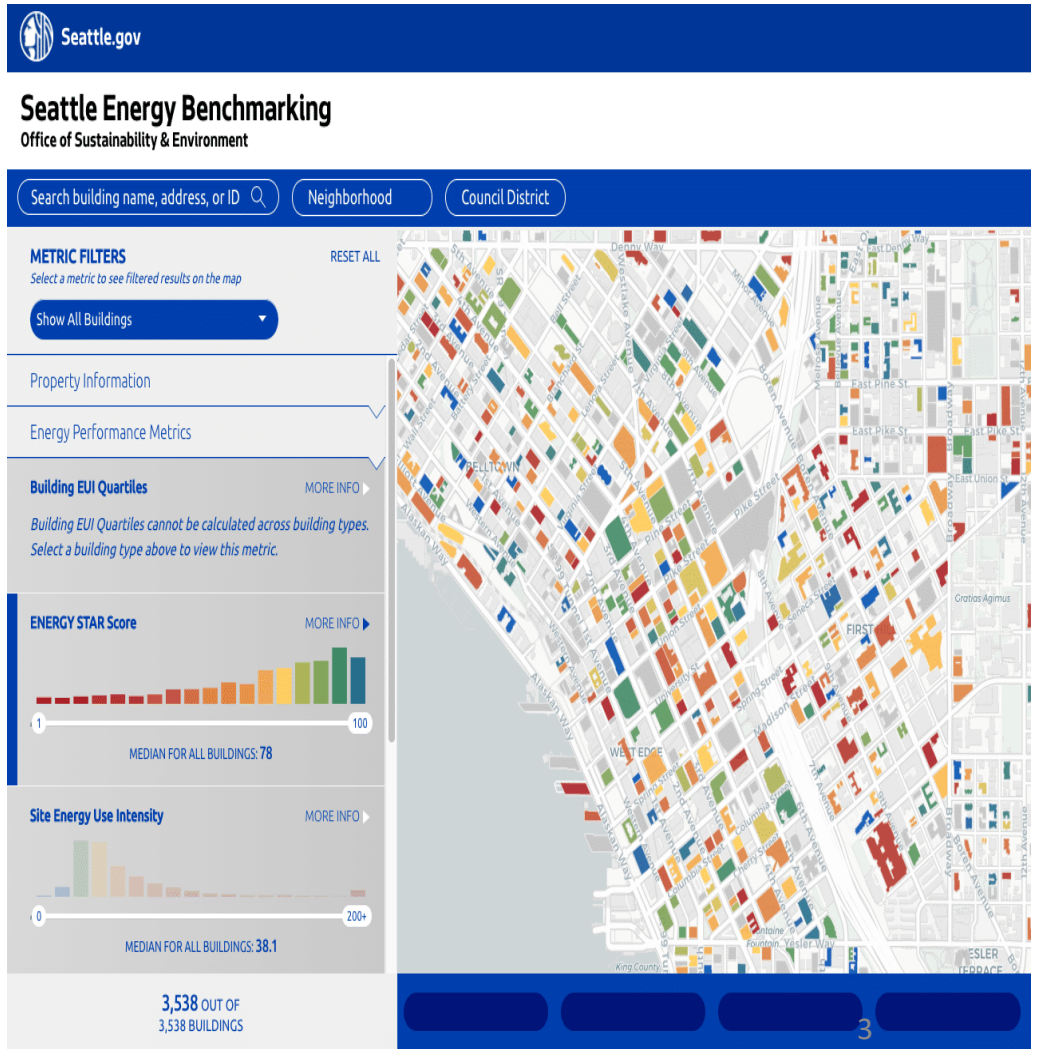
Comment prédire la consommation d'énergie et les émissions de CO2 des bâtiments non résidentiels de Seattle de manière économique, en limitant la dépendance aux relevés coûteux ?

Missions:

A partir des données récoltées :

- **Développer un modèle prédictif** pour estimer la **consommation d'énergie** et les **émissions de CO2** des **bâtiments non résidentiels** de Seattle, afin de contribuer aux objectifs de durabilité de la ville.

- Evaluer l'intérêt de **l'ENERGY STAR Score** pour la **prédiction d'émission**.



Données

Relevés des bâtiments effectués en 2016

Informations	Données 2016
Lignes	3376
Taille	46
Données Structurelles	nombre de bâtiments, nombre d' étages, année de construction, type d'utilisation, surfaces,…
Données de localisation	Adresse, quartier, coordonnées Lat./Long, numéro de District,...
Données énergétiques	Consommation énergétique totale , consommation de gaz ,consommation de vapeur ,consommation électrique, émissions de CO2,…

Données

Étapes Clés du Traitement des Données

Métier:

- ✓ Compréhension du métier :
- ✓ Assemblage des jeux de données

Nettoyage:

- ✓ Suppression des données inutiles, filtre
- ✓ Valeurs manquantes, aberrantes
- ✓ Gestion des doublons

Analyse :

- ✓ Analyse univariée
- ✓ Analyse bivariée / multi-variée

Pré processing :

- ✓ Features engineering
- ✓ Imputation
- ✓ Type des variables, transformation cible

Processus de Sélection et de Nettoyage des Données

Flux de Préparation des Données

Sélection des Variables Cibles (Targets)

- TotalGHGEmissions
- SiteEnergyUse(kBtu)

Score Energy Star

Score élevé = une meilleure efficacité énergétique

Traitement des Données

- Nettoyage des données
- Gestion des valeurs extremes/manquantes

Feature Engineering

Encodages/Transformations

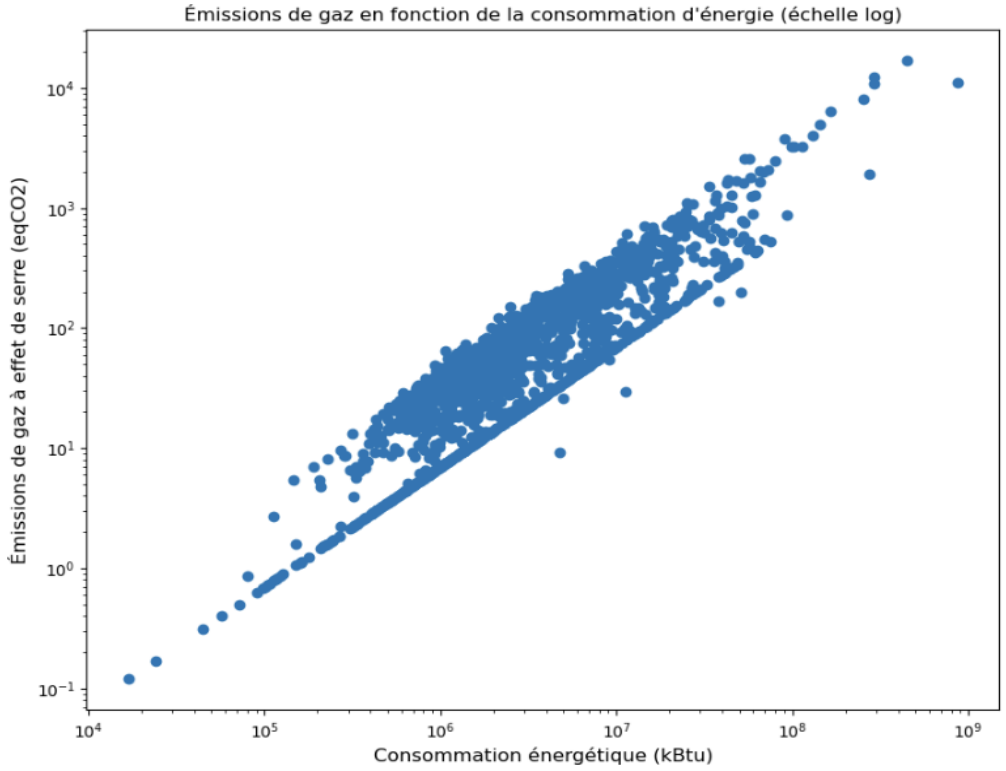
Vérification du data leakage

Processus de Sélection et de Nettoyage des Données

Étape	Description	Méthodes Utilisées	Variables Exemples
Identification des Variables	Analyse initiale pour déterminer les variables à traiter.	Statistiques descriptives	Toutes les variables initiales
Suppression pour Multicollinéarité	Élimination des variables redondantes pour éviter les problèmes de multicollinéarité.	Analyse de corrélation	LargestPropertyUseTypeGFA, Electricity(kBtu), SiteEUI(kBtu)
Conservation des Informations Uniques	Conservation des variables qui apportent des informations uniques.	Analyse de contribution unique	SecondLargestPropertyUseTypeGFA, ThirdLargestPropertyUseTypeGFA
Traitement des Valeurs Extrêmes et Manquantes	Correction des valeurs aberrantes et imputation des valeurs manquantes.	Imputation médiane, suppression des outliers	NaturalGas(kBtu)
Binarisation et Encodage	Transformation de variables continues en catégorielles et encodage des variables catégorielles.	Binarisation, encodage OneHot	NaturalGas(kBtu), BuildingType, PrimaryPropertyType

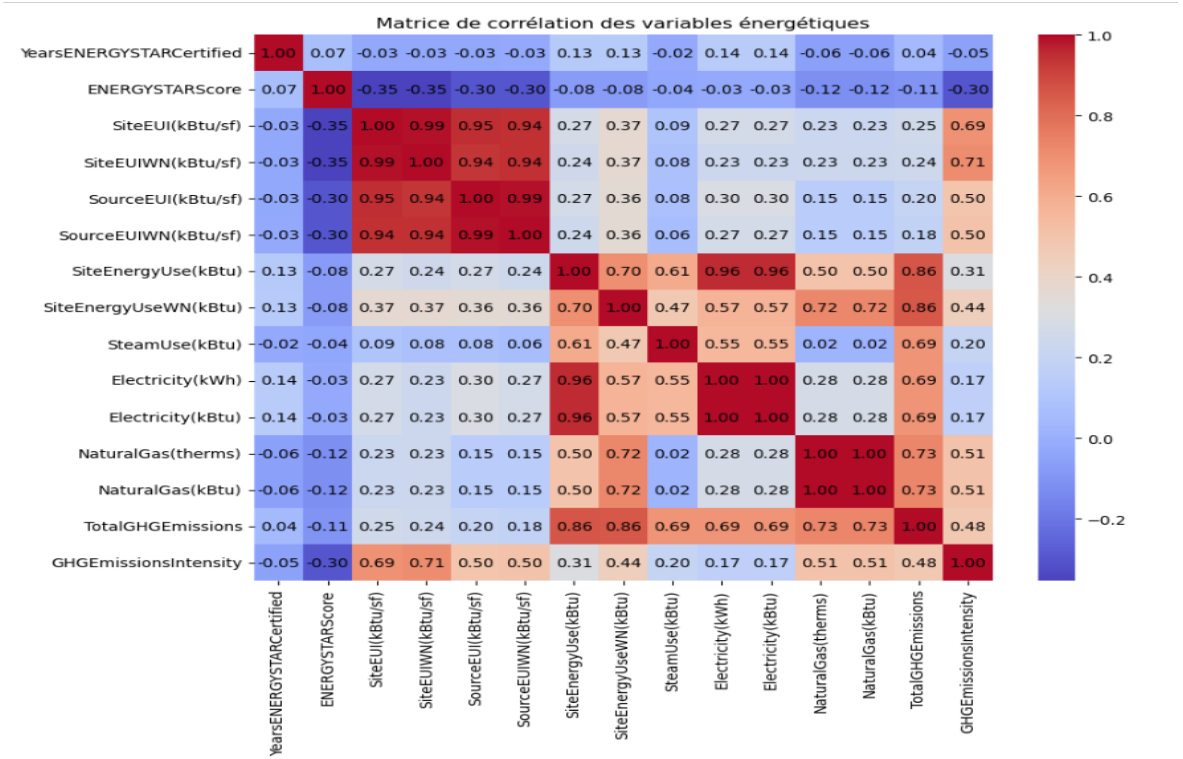
Chaque étape du nettoyage est justifiée par la nécessité de réduire les erreurs de modèle et d'améliorer la précision des prédictions.

Processus de Sélection et de Nettoyage des Données



Coefficient de Pearson entre les émissions de gaz et la consommation d'énergie : 0.8599129061367168

- Le nuage de points met en évidence une corrélation positive forte entre la consommation énergétique et les émissions de gaz à effet de serre, suggérant que l'efficacité énergétique est un levier clé pour la réduction des émissions.

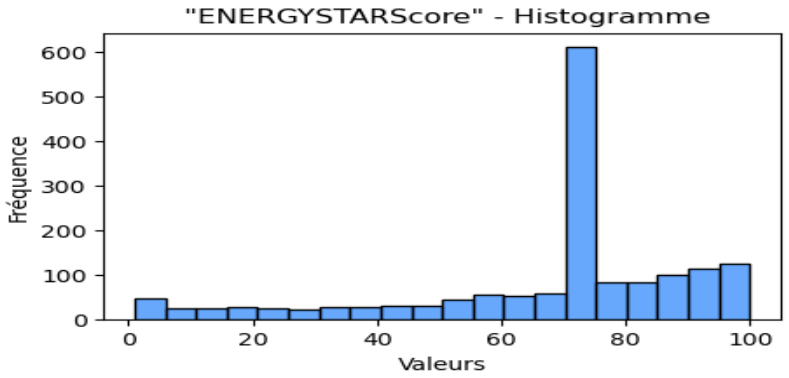


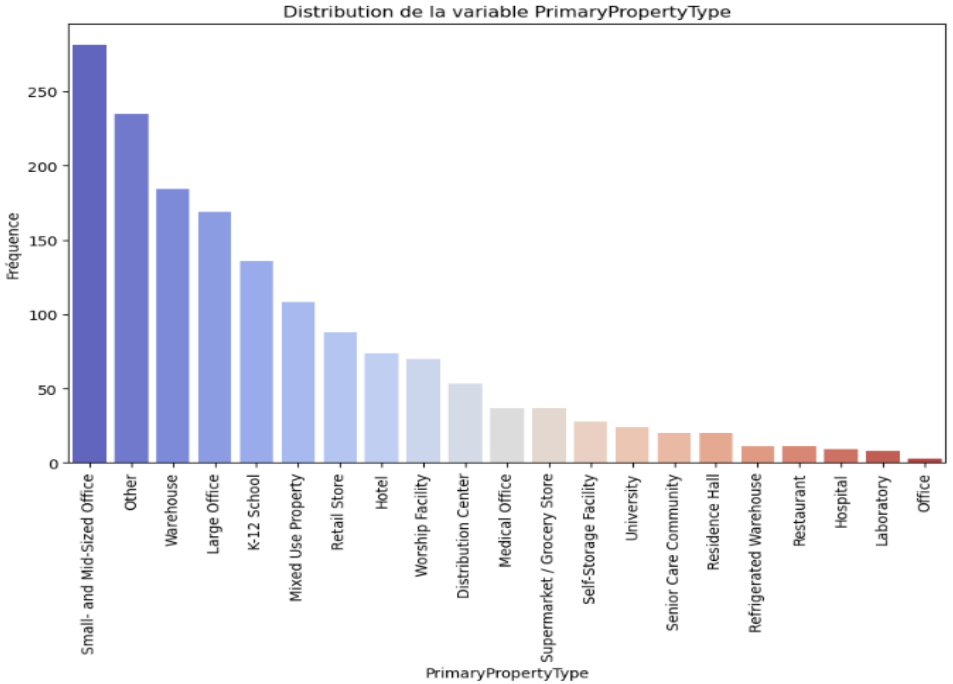
- Des corrélations élevées indiquent la nécessité de sélectionner une seule variable représentative par groupe de corrélations élevées pour éviter la redondance dans le modèle.

Analyse Exploratoire Des Données

Analyse Univariée

```
Statistiques pour 'ENERGYSTARScore':
Moyenne: 68.28
Médiane: 73.50
Écart Type: 23.48
Variance: 551.12
Minimum: 1.00
Maximum: 100.00
Quartile 25%: 63.00
Quartile 75%: 82.00
Skewness empirique: -1.27
Kurtosis empirique: 1.12
```

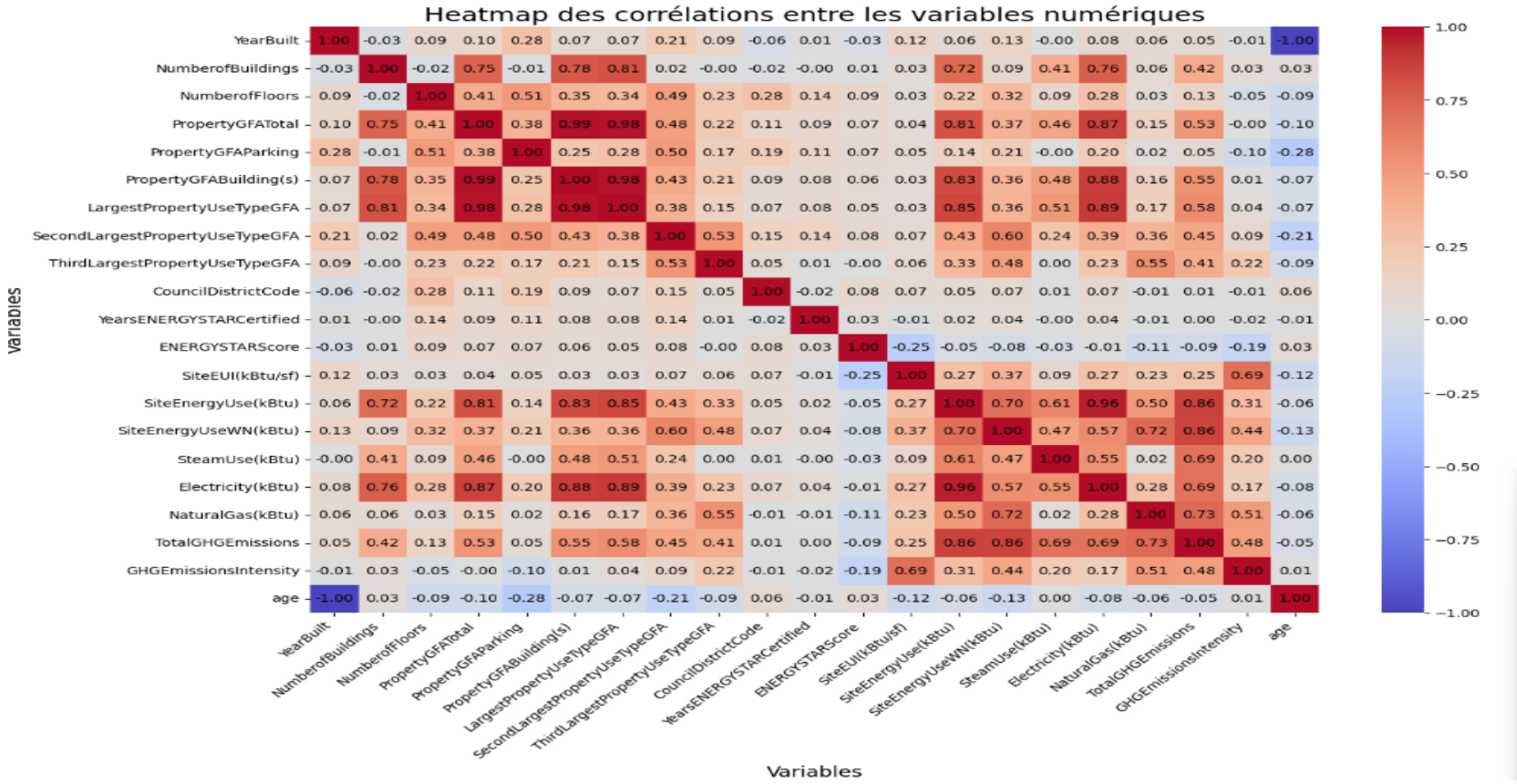




- L'histogramme montre une distribution avec une concentration élevée de scores ENERGYSTAR autour de 70-80, suggérant que la majorité des bâtiments ont une performance énergétique supérieure à la moyenne.
- Le graphique à barres révèle une prédominance des bureaux parmi les types de propriété principaux.

Analyse Exploratoire Des Données

Analyse Multi-variée



- Cette version actualisée de la heatmap révèle des interdépendances subtiles entre les variables, éclairant le chemin vers une sélection stratégique de variables pour affiner le modèle prédictif.

Analyse Exploratoire Des Données

Étapes Clés du Prétraitement des Données pour la Modélisation

Étape	Objectif	Méthode
Binarisation	Simplifier l'analyse en transformant une variable quantitative en binaire.	Codification des valeurs supérieures à un seuil prédéfini (500 kBtu) comme 1 (haute consommation) et les autres comme 0.
Encodage One-Hot	Préparer les variables catégorielles pour les algorithmes d'apprentissage automatique.	Création de colonnes binaires pour chaque catégorie avec OneHotEncoder, sans ordre hiérarchique implicite.
Standardisation	Normaliser les variables numériques pour faciliter la comparaison.	Ajustement des variables numériques à une moyenne de 0 et un écart-type de 1 avec StandardScaler.
Imputation	Traiter les valeurs manquantes pour permettre l'analyse.	Remplacement des valeurs manquantes par la moyenne pour les variables numériques et la catégorie la plus fréquente pour les variables catégorielles avec SimpleImputer.

Modélisation

Émissions de CO2

Consommation d’énergie

Modèles selectionnés

Linéaires	Ensemblistes	Basé sur la proximité:
Régression classique Ridge Lasso Elastic Net	Gradient Boosting RandomForest Xgboost	K-NN

- **GridSearchCV** est utilisé sur chaque modèle : un outil de sélection de modèle fourni par la bibliothèque scikit-learn, qui permet de trouver les meilleurs hyperparamètres pour un modèle d'apprentissage automatique.

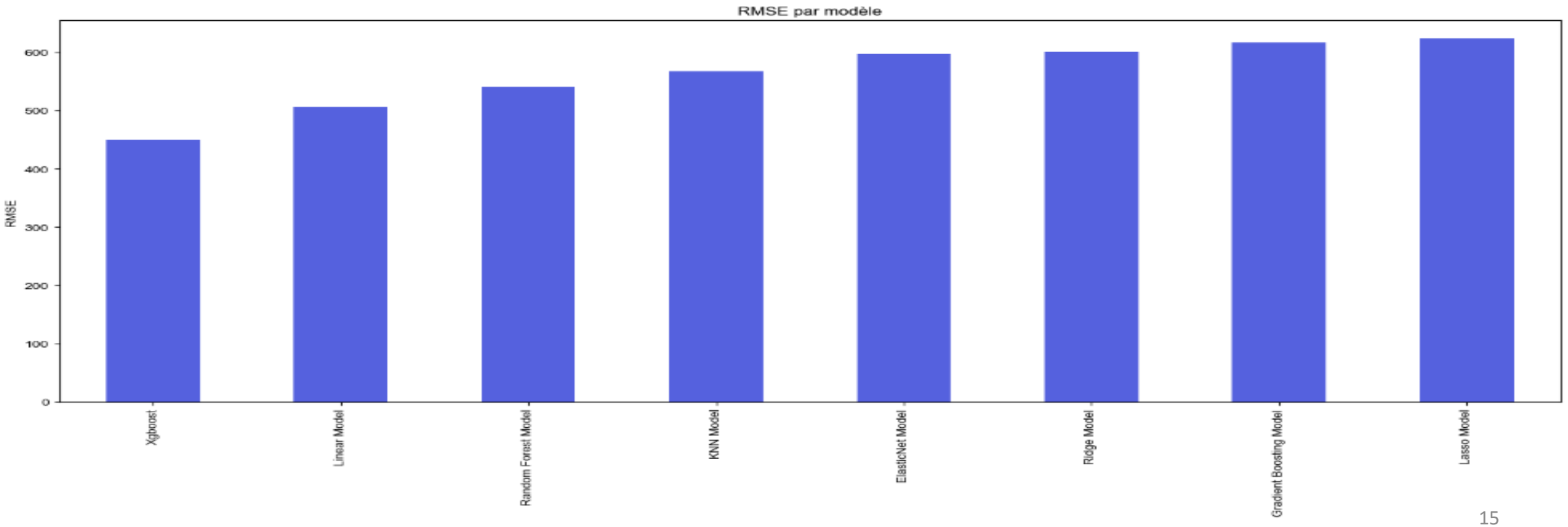
Synthèse des Métriques de Performance Modèle

MÉTRIQUES	DÉFINITION
MSE(Mean Squared Error)	**Erreurs quadratiques:** mesure de la moyenne des carrés des erreurs Meilleur = le score le plus bas
RMSE (Root Mean Squared Error)	**Erreurs quadratiques racinées :** Racine carrée de la MSE pour des unités de mesure cohérentes. Meilleur = le score le plus bas
MAE (Mean Absolute Error)	**Erreurs absolues:** mesure la moyenne des valeurs absolues des erreurs. Meilleur = le score le plus bas
R^2 (Coefficient de Détermination)	**Variance expliquée :** évalue la proportion de la variance expliquée par le modèle. Varie de 0 (aucune explication) à 1 (explication parfaite).

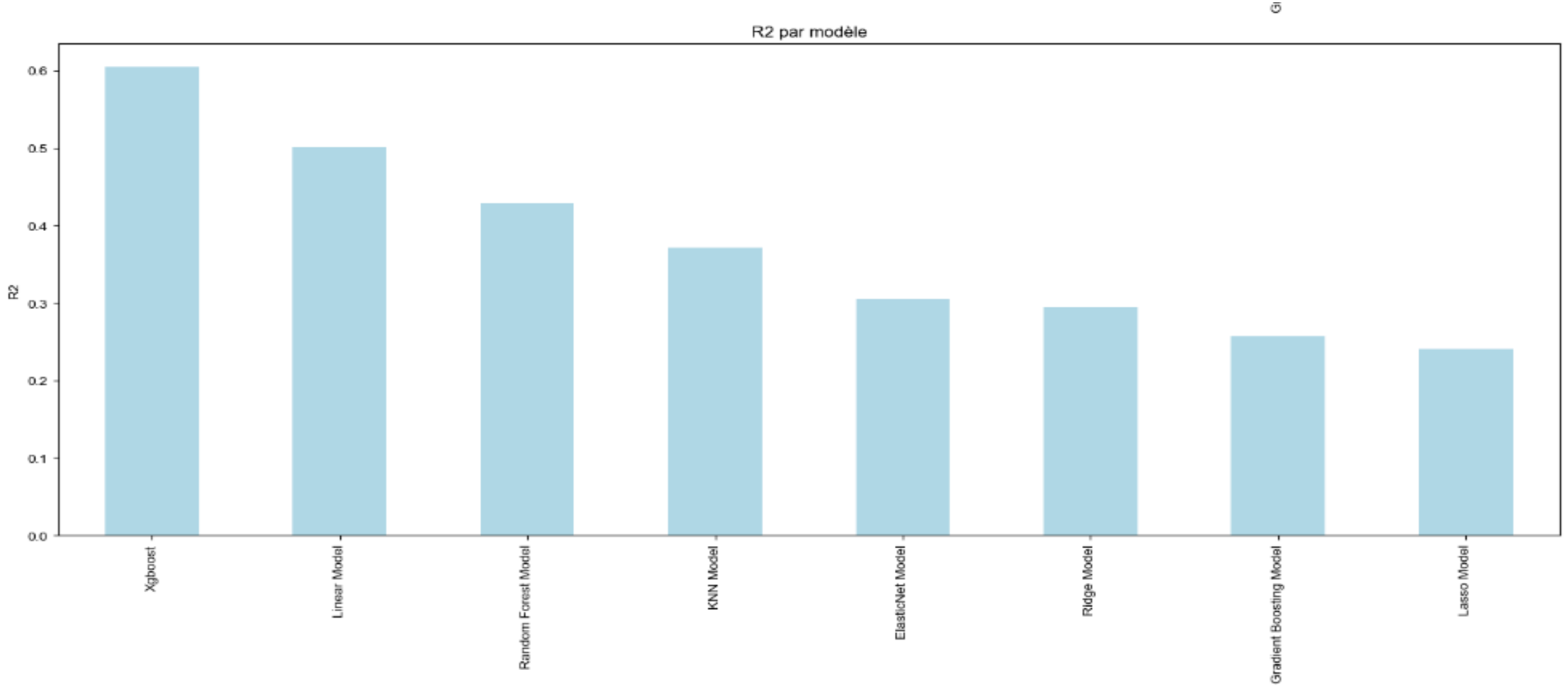
Modélisation

Emission de CO_2

- **Cible** :TotalGHGEmissions
- **Démarche** : split, encodage, standardisation, modèle de base



Modélisation Emission de CO2

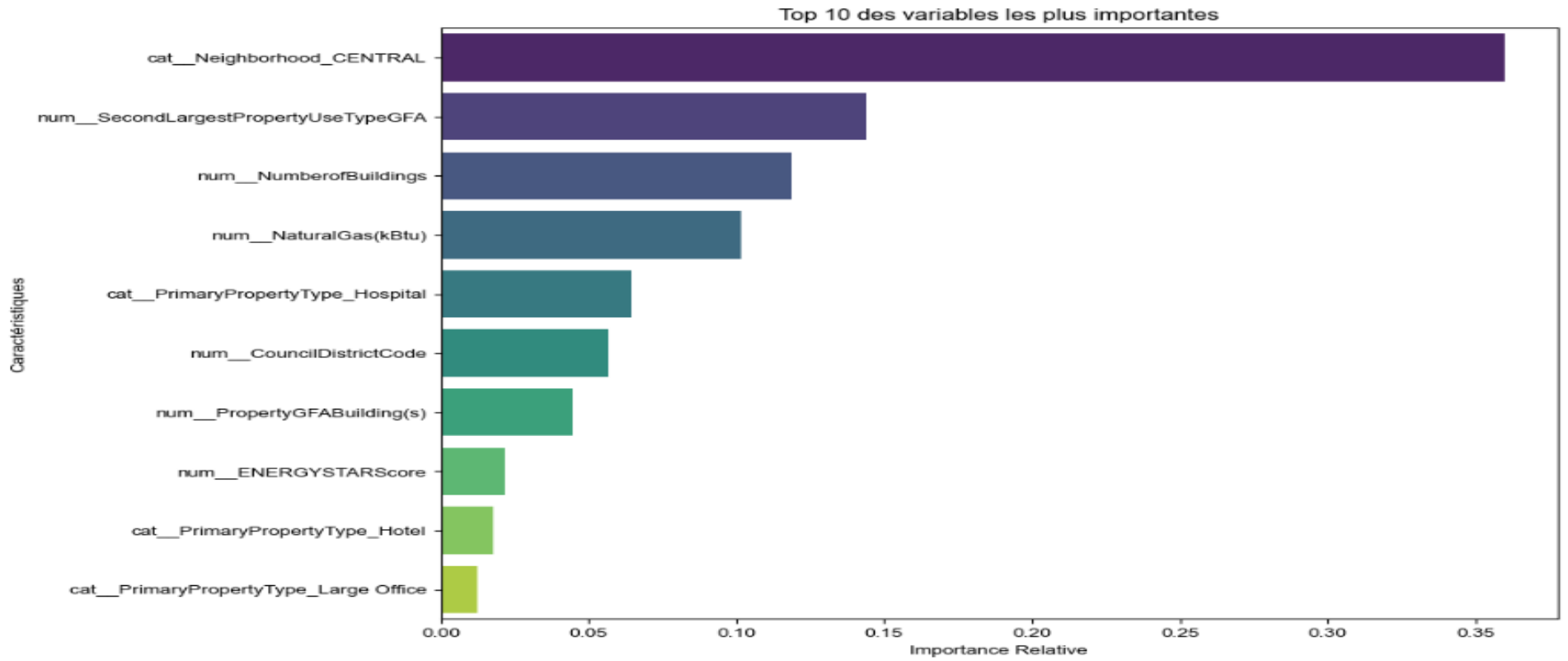


Modèle à OPTIMISER : Xgboost (eXtreme Gradient Boosting)

Sélection du modèle final

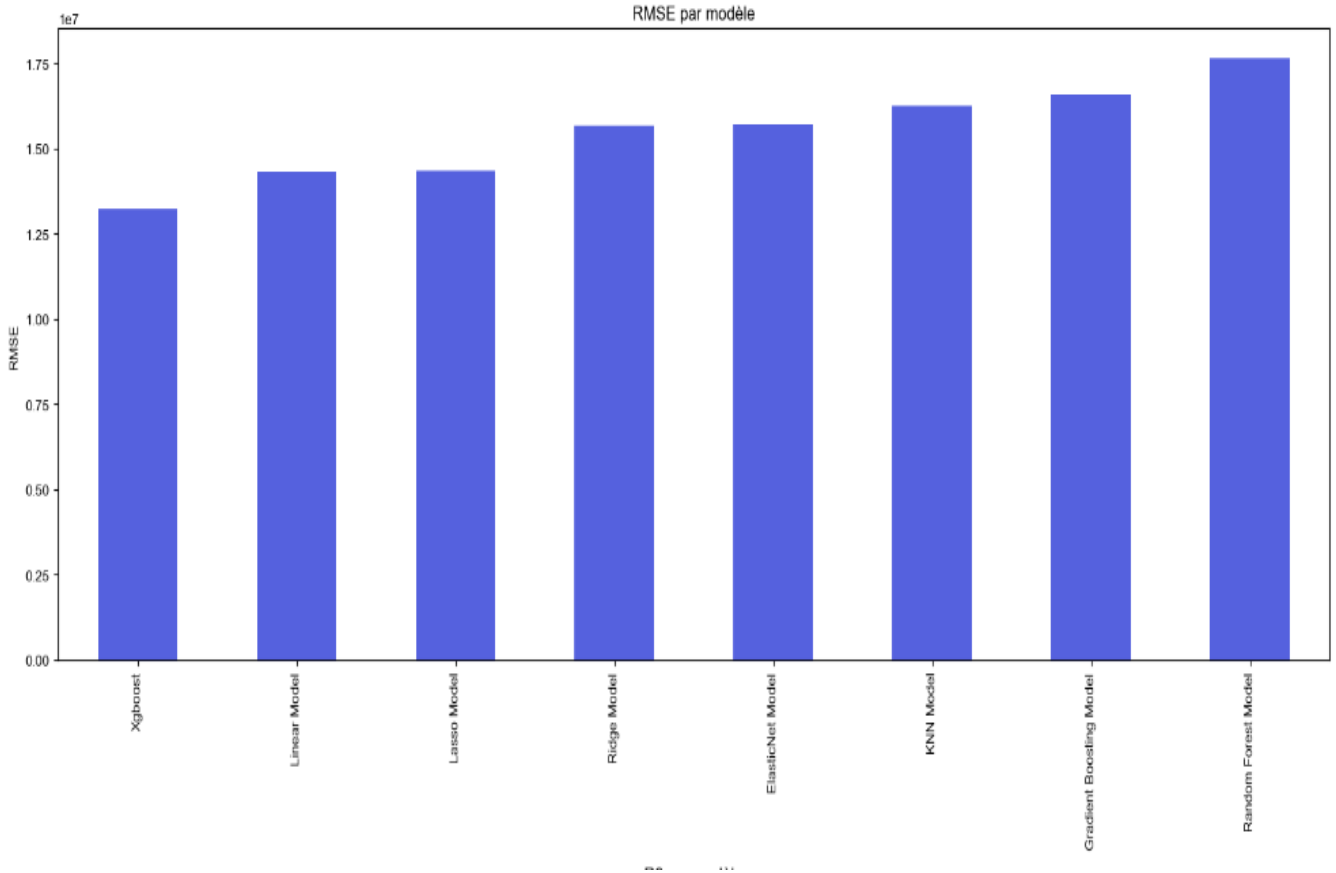
FEATURES IMPORTANCE

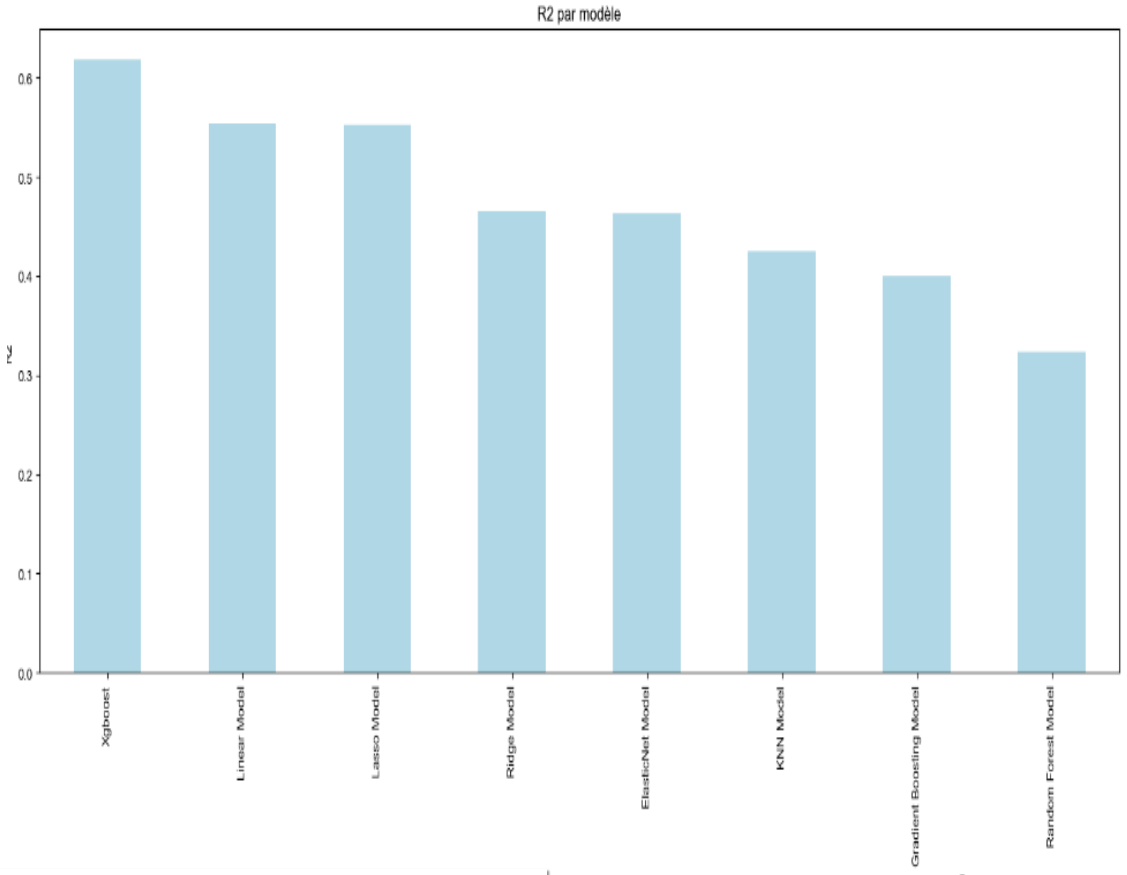
Xgboost



Modélisation
Consommation d'énergie

- **Cible** : SiteEnergyUse(Kbtu)
- **Démarche** : split, encodage, standardisation, modèle de base



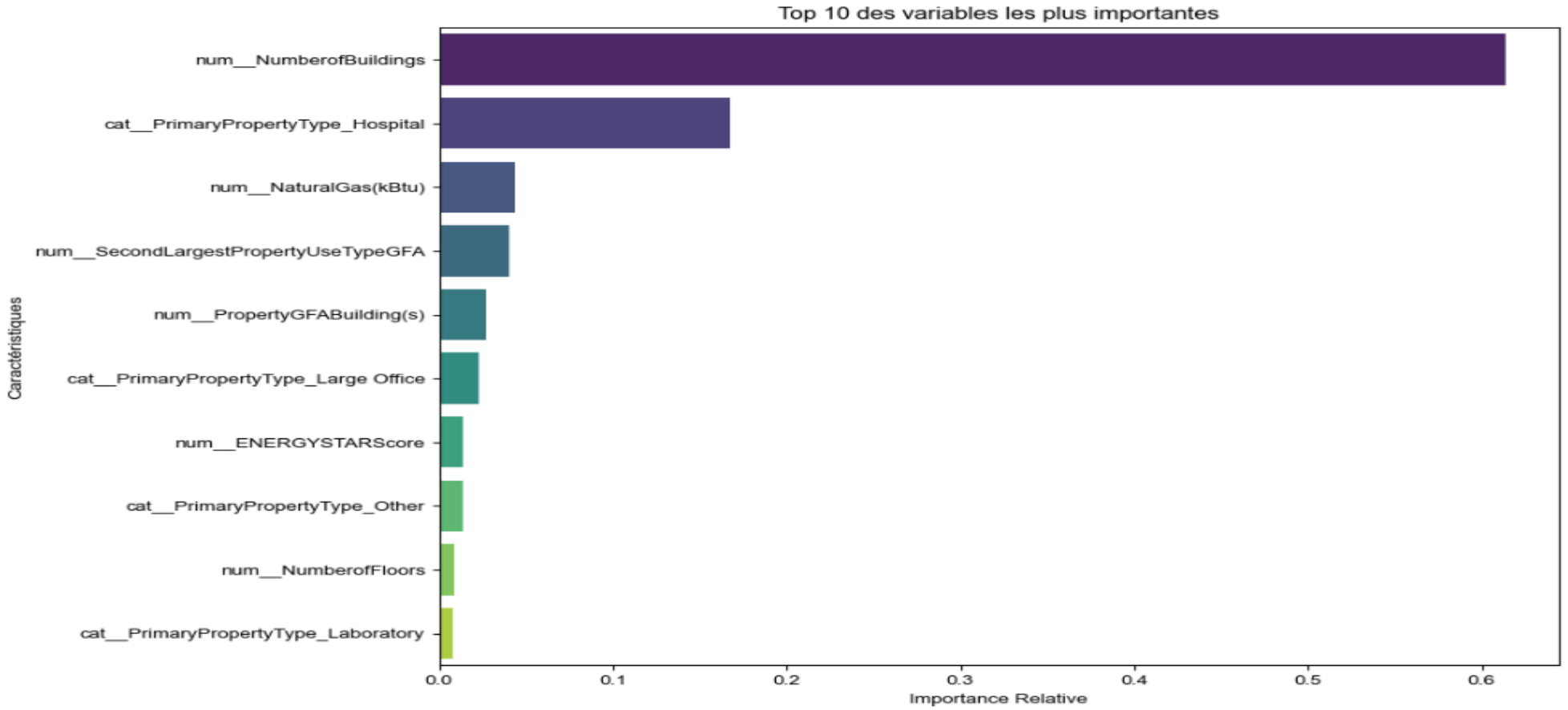


Modèle à OPTIMISER : Xgboost (eXtreme Gradient Boosting)

Sélection du modèle final

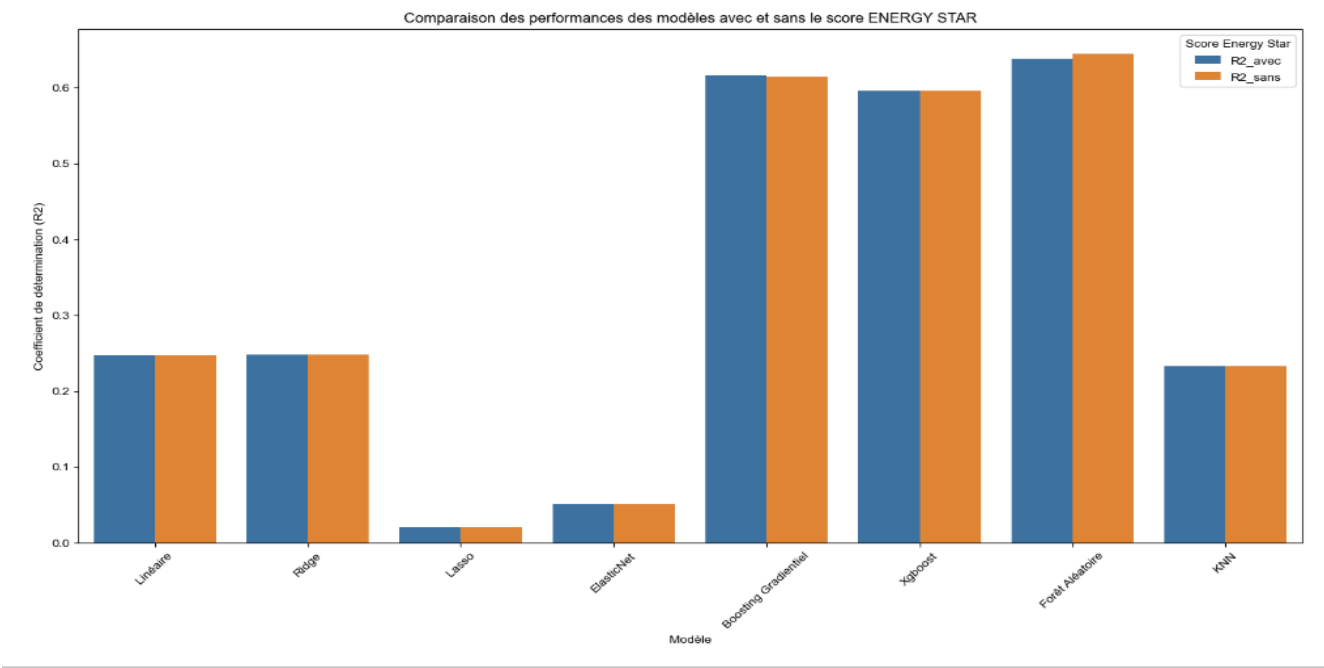
FEATURES IMPORTANCE

Xgboost



Intérêt de l'Energy Star sur l'émission de CO2

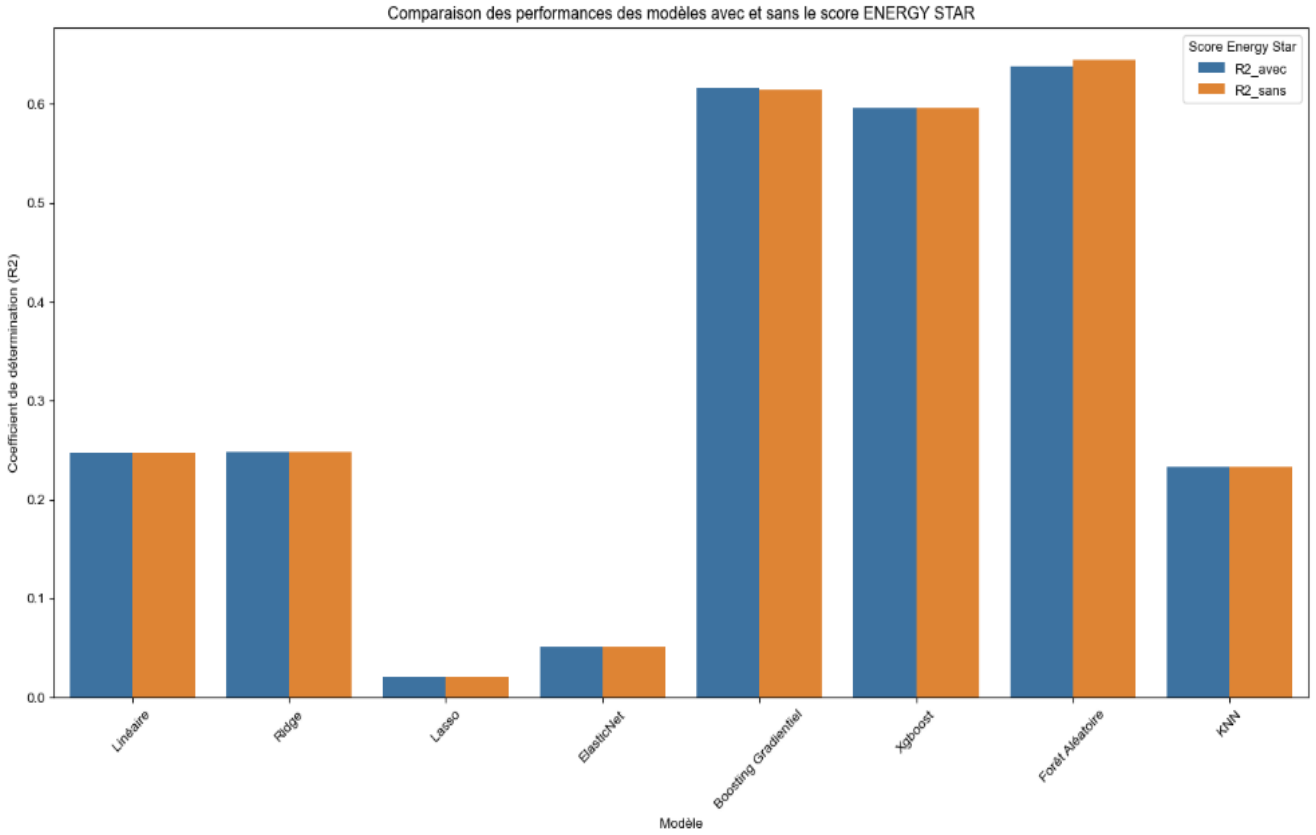
Comparaison avec ou sans ENERGYSTARScore



ENERGYSTARScore améliore légèrement la précision de certains modèles.

Intérêt de l’Energy Star sur la consummation d’énergie

Comparaison avec ou sans ENERGYSTARScore



Conclusion

•**Performance des Modèles :** Utilisation efficace de modèles de machine learning, particulièrement XGBoost et Random Forest, pour prédire la consommation énergétique et les émissions de gaz à effet de serre des bâtiments.
•L'intégration de l'Energy Star Score a significativement amélioré la précision des prédictions.

Axes d'Amélioration :
•**Exploration de modèles avancés :** Implémenter des architectures de réseaux de neurones comme les réseaux neuronaux profonds (Deep Neural Networks, DNN) ou les réseaux neuronaux convolutifs (CNN) adaptés à l'analyse temporelle, pour améliorer la précision des prédictions sur les données complexes et volumineuses.

•**Amélioration de la performance :** Les réseaux de neurones pourraient offrir de meilleures performances en termes de précision des prédictions grâce à leur capacité à apprendre des niveaux élevés d'interactions non linéaires entre les caractéristiques.

MERCI !

Des questions ?