PROJET N°3: ANTICIPEZ LES BESOINS EN CONSOMMATION ÉLECTRIQUE DE BÂTIMENTS

BOURBON VICENTE

SOMMAIRE

- Problématique
- Nettoyage et Feature Engineering
- Exploration des données
- Modèles de prédiction
- Choix du modèle final

PROBLÉMATIQUE

- Bâtiments non résidentiels de Seattle
- Prédiction émissions CO2 et consommation d'énergie
- Données de 2015 et 2016
- Variables à prédire:
  - GHGEmissions(MetricTonsCO2e)
  - SiteEnergyUse(kBtu)

NETTOYAGE ET FEATURE ENGINEERING
1. UNION DES DEUX JEUX DE DONNÉES

- Comparaison des structures et des variables
- Décomposition de la variable Location
- Harmonisation des noms de variables
- Suppression des variables uniques

NETTOYAGE ET FEATURE ENGINEERING
2. VALEURS MANQUANTES

- Relevés onéreux: suppression minimale d'observations
- Implémentation par la valeur moyenne ou la valeur la plus fréquente
- Utilisation des autres variables pour l'évaluation

NETTOYAGE ET FEATURE ENGINEERING
3. DOUBLONS

- Lignes identiques
- Bâtiments n'apparaissant pas deux fois dans les relevés d'une même année

NETTOYAGE ET FEATURE ENGINEERING
4. OUTLIERS

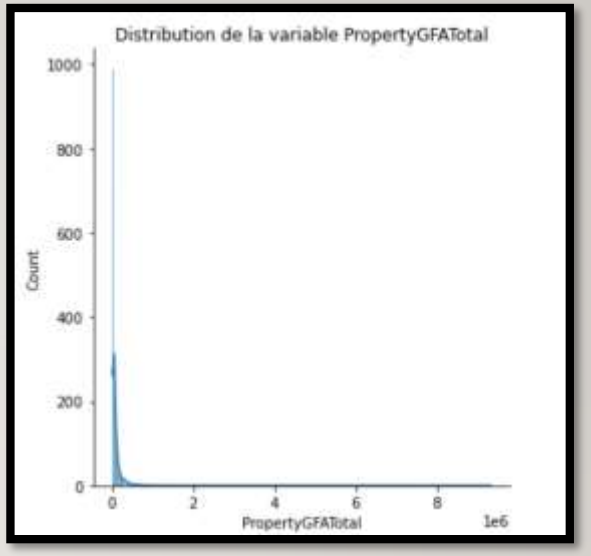
- Utilisation de la variable Outlier
- Pas d'informations dans la documentation sur les critères d'outlier
- Suppression des « high outlier »

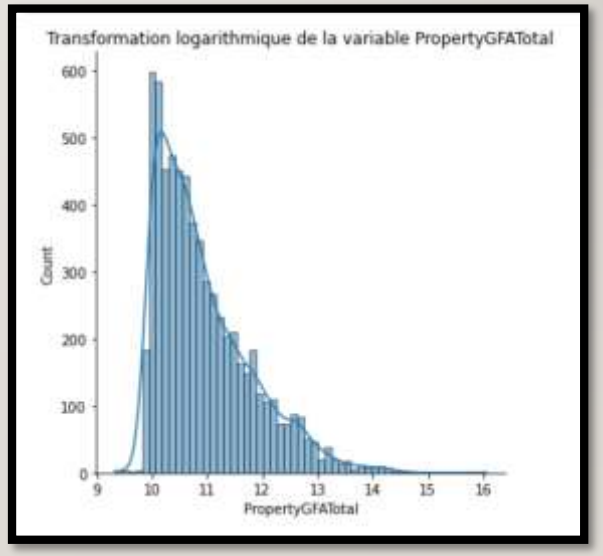
NETTOYAGE ET FEATURE ENGINEERING
5. AGE ET TYPE DES BÂTIMENTS

- Pas de bâtiment destiné à l'habitation individuelle
- Calcul de l'âge des bâtiments

NETTOYAGE ET FEATURE ENGINEERING
6. TRANSFORMATION DE VARIABLES

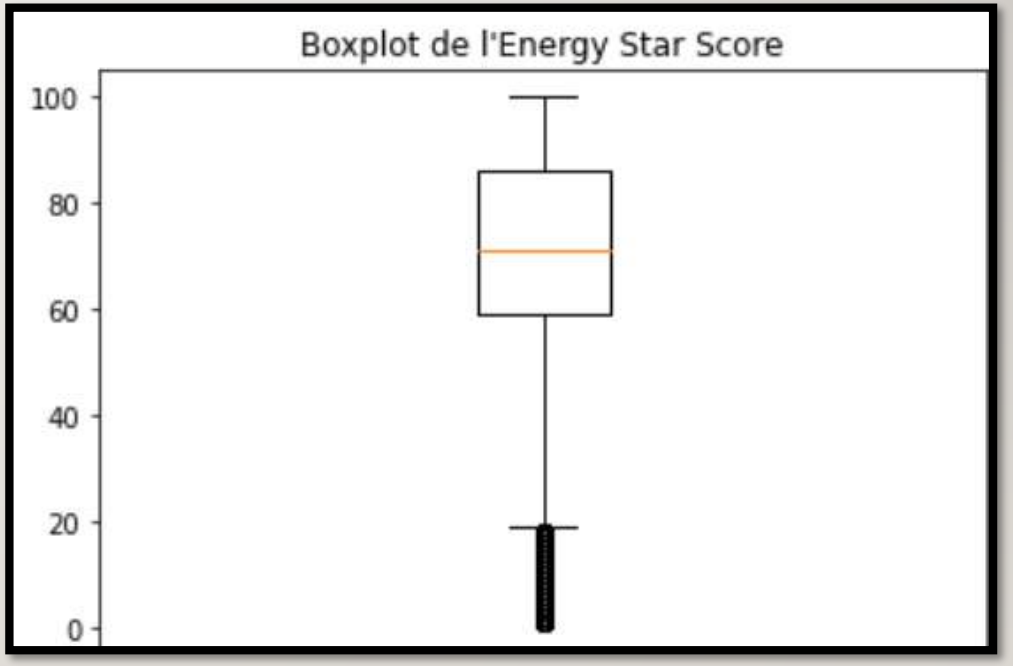
- Discrétisation des variables catégorielles nominales
- Passage au log des variables numériques continues
- Standardisation des données





EXPLORATION DES DONNÉES

- Majorité d'immeubles
- Majorité de bâtiments ayant un très bon Energy Star Score
- Principales corrélations entre consommation d'énergie, surface et émissions de CO2



MODÈLES DE PRÉDICTION
1. PRÉPARATION

- Différents jeux de données pour permettre la comparaison
- Choix des targets et des features
- Séparation jeu d'entraînement / jeu de test

MODÈLES DE PRÉDICTION
2. DÉMARCHE DE TEST D'UN MODÈLE

- Recherche sur grille des hyperparamètres
- Validation croisée sur le jeu d'entraînement
- Entraînement du modèle sur le jeu d'entraînement
- Evaluation du modèle sur le jeu de test
- Scores du modèles

MODÈLES DE PRÉDICTION
3. MODÈLES TESTÉS

- Complexité croissante
- Régressions linéaires
- Méthodes ensemblistes
- Méthodes non linéaires

CHOIX DU MODÈLE FINAL
1. ÉMISSIONS DE CO2

- Modèles plus robustes avec les transformations logarithmiques et l'Energy Star Score
- Conserver l'Energy Star Score
- Choix du Gradient Boosting pour la prédiction des émissions de CO2
- Possibilité d'optimisation des hyperparamètres

	Bagging	Random Forest	Gradient Boosting	Ridge à noyau
R2 train	0.988381	0.988336	0.922928	0.958989
R2 test	0.903300	0.904800	0.880700	0.703200
Temps calcul (s)	11.559200	6.938200	2.647100	4.547200

CHOIX DU MODÈLE FINAL
2. CONSOMMATION TOTALE D'ÉNERGIE

- Utilisation de l'Energy Star Score et de la quantité d'émissions de CO2
- Pas de gain avec les transformations logarithmiques
- Gradient Boosting pour la prédiction de la consommation totale d'énergie mais un peu d'instabilité
- Possibilité d'optimisation des hyperparamètres

	Bagging	Random Forest	Gradient Boosting
R2 train	0.9825816	0.9265382	0.9930764
R2 test	0.8683000	0.8687000	0.8750000
Temps calcul (s)	1.157800	6.033500	2.946200