Prédiction des émissions de CO2 et de la consommation énergétique des bâtiments non résidentiels à Seattle

Prédiction des émissions de CO2 et de la consommation énergétique des bâtiments non résidentiels à Seattle

Contexte du projet :
• Objectif de rendre Seattle neutre en carbone d'ici 2050.
• Focalisation sur la consommation énergétique et les émissions des bâtiments non résidentiels.

Problématique :
• Les relevés de consommation annuels sont coûteux et limités.
• Prédire les émissions de CO2 et la consommation énergétique pour les bâtiments non mesurés à partir de données structurelles (taille, usage, année de construction, localisation...).

Objectifs :
• Développer un modèle prédictif basé sur les caractéristiques des bâtiments.
• Intégrer l'ENERGY STAR Score pour améliorer les prédictions.

Plan

- Analyse Exploratoire des Données
  - Structure des données
  - Sélection des bâtiments conformes
  - Gestion des valeurs aberrantes et manquantes
- Préparation des données
  - Sélection des variables
  - Création de variables
- Analyse des relations et corrélations
- Transformations des variables
- Modélisation des prédictions
  - Performance des modèles pour la prédiction des émissions de CO_2
  - Validation du modèle (émissions de CO_2)
  - Importance des variables dans le modèle (émissions de CO_2)
  - Analyse de l'impact du score d'énergie (émissions de CO_2)
- Modélisation pour la consommation d'énergie
  - Performance des modèles pour la prédiction de la consommation d'énergie
  - Validation du modèle (consommation d'énergie)
  - Importance des variables dans le modèle (consommation d'énergie)
  - Analyse de l'impact du score d'énergie (consommation d'énergie)

Structure des Données

46 colonnes, dont 30 numériques et 15 textuelles, couvrant divers aspects des bâtiments.
3376 propriétés.

Catégorie	Détails	Exemples de Variables
Identification et Localisation	Informations permettant d'identifier et localiser le bâtiment.	OSEBuildingID, TaxParcelIdentificationNumber, Adresse, Ville, État, ZipCode, Latitude, Longitude, CouncilDistrictCode, Neighborhood
Caractéristiques Structurelles	Données relatives aux caractéristiques physiques et structurelles du bâtiment.	BuildingType, PrimaryPropertyType, PropertyName, YearBuilt, NumberofBuildings, NumberofFloors, PropertyGFATotal, PropertyGFAParking, PropertyGFABuilding(s)
Usage et Surface	Informations sur l'usage des bâtiments et les surfaces associées.	ListOfAllPropertyUseTypes, LargestPropertyUseType, LargestPropertyUseTypeGFA, etc
Performance Énergétique	Données sur la performance énergétique du bâtiment.	ENERGYSTARScore, YearsENERGYSTARCertified, SiteEUI(kBtu/sf), SiteEUIWN(kBtu/sf), SourceEUI(kBtu/sf), SourceEUIWN(kBtu/sf), SiteEnergyUse(kBtu), SiteEnergyUseWN(kBtu), SteamUse(kBtu), Electricity(kWh), Electricity(kBtu), NaturalGas(therms), NaturalGas(kBtu)
Émissions de Gaz à Effet de Serre	Données sur les émissions de gaz à effet de serre et leur intensité.	TotalGHGEmissions, GHGEmissionsIntensity
Autres Informations	Autres données relatives à la conformité, aux anomalies et aux commentaires.	DefaultData, Comments, ComplianceStatus, Outlier

Sélection des Bâtiments Conformes

Nombre initial de propriétés : 3 376.

Filtrage des bâtiments non résidentiels : 1 578 bâtiments non résidentiels.

Analyse de la colonne ComplianceStatus : Suppression des propriétés avec un ComplianceStatus autre que "Compliant" : 1 459 propriété restantes.

Analyse de la colonne DefaultData :

- Résultat : Toutes les propriétés ont DefaultData = False, signifiant qu'aucune donnée par défaut n'a été utilisée dans les propriétés restantes.

Analyse de la colonne Outlier :

- Résultat : La colonne est vide, il n'y a donc aucune valeur aberrante les propriétés renstantes.

Suppression des colonnes avec plus de 50% de valeurs manquantes :

- Avant nettoyage : 46 colonnes.
- Après nettoyage : 41 colonnes, après suppression de celles avec plus de 50% de valeurs manquantes.

Gestion des valeurs aberrantes et manquantes

Recherche des valeurs aberrantes

- Émissions de CO_2 (tonnes métriques):

Vérifications des émissions ≤ 0 ou > 2200 → suppression de 3 bâtiments

- Consommation d'énergie :

Vérification des consommations > 77 000 000 kBtu → aucune anomalie détectée

- NumberOfBuildings = 0 :

21 remplacements après validation manuelle

- NumberOfFloors = 0 ou > 26 :

8 remplacements après recherche sur Internet

- PropertyGFABuilding(s) (pieds carrés) :

Vérification des tailles> 1 500 000 → aucune anomalie détectée

Valeurs manquantes

- LargestPropertyUseType :

Imputation de 2 valeurs manquantes en se basant sur PrimaryPropertyType

- ENERGYSTARScore :

Suppression des lignes avec 35 % de valeurs manquantes pour garantir la fiabilité des analyses

Sélection des variables

Critères de sélection

- Pertinence métier : Garder uniquement les variables qui ont du sens dans le contexte du projet.
- Éviter la redondance : Ne garder qu'une seule version des variables qui sont trop similaires ou qui apportent la même information, pour éviter des doublons inutiles.
- Eviter le Data Leakage : Exclure les variables qui ne seraient pas accessibles au moment de la prédiction (comme les relevés de consommation d'énergie) ou celles qui sont directement corrélées à la cible, pour éviter d'introduire une fuite d'information.

Variables sélectionnées

- Usage principal de la propriété
- District de la propriété
- Nombre d'étages
- Nombre de bâtiments
- Taille de la propriété
- Taille de l'usage principal de la propriété

Création de variables

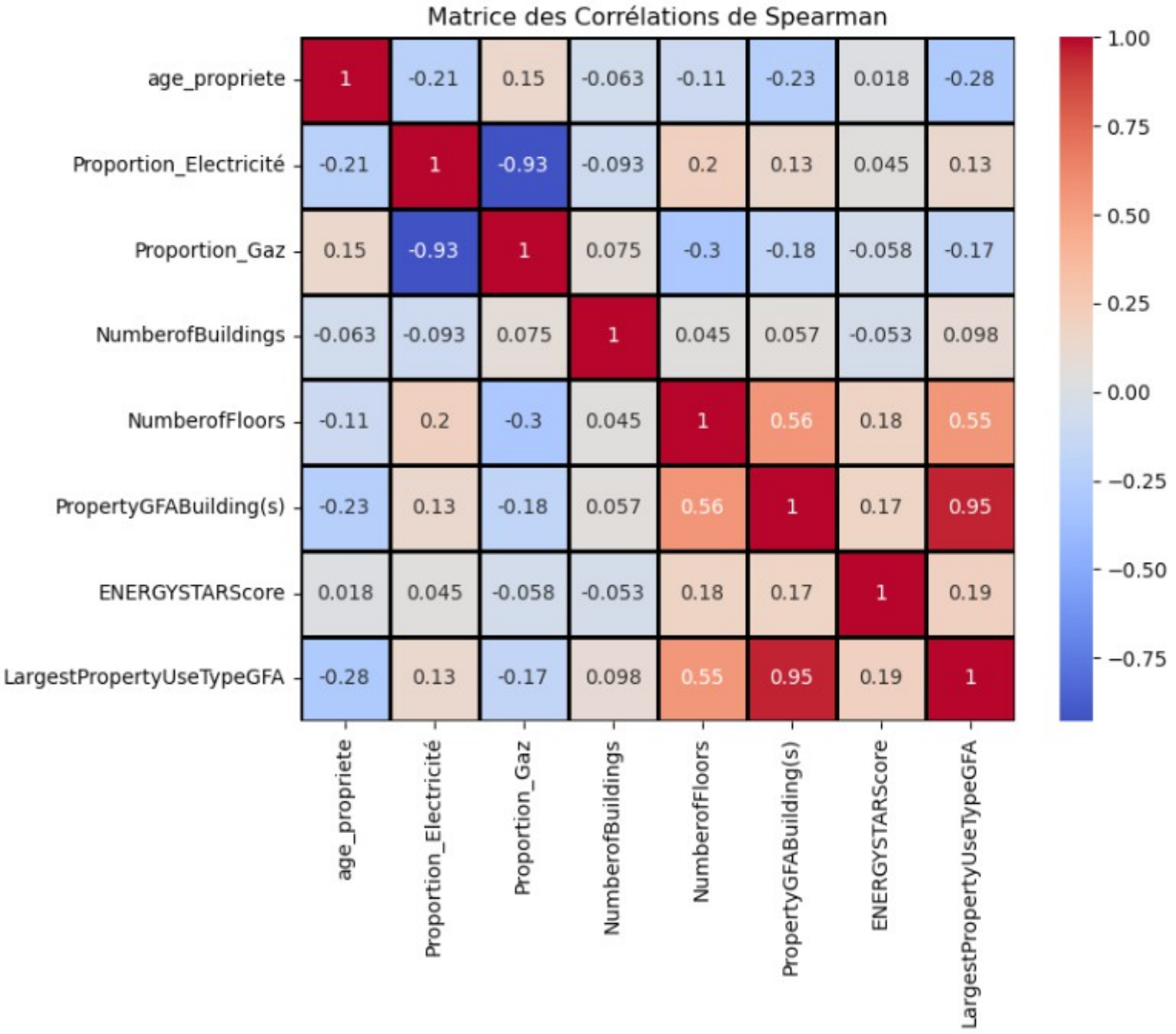
• **Âge de la propriété** : Différence entre l'année actuelle et l'année de construction/rénovation
→ Impact du vieillissement sur les performances énergétiques.

• **Proportion d'électricité et de gaz** : Répartition des sources d'énergie
→ Compréhension de l'impact des sources d'énergie sur la consommation et les émissions de CO_2.

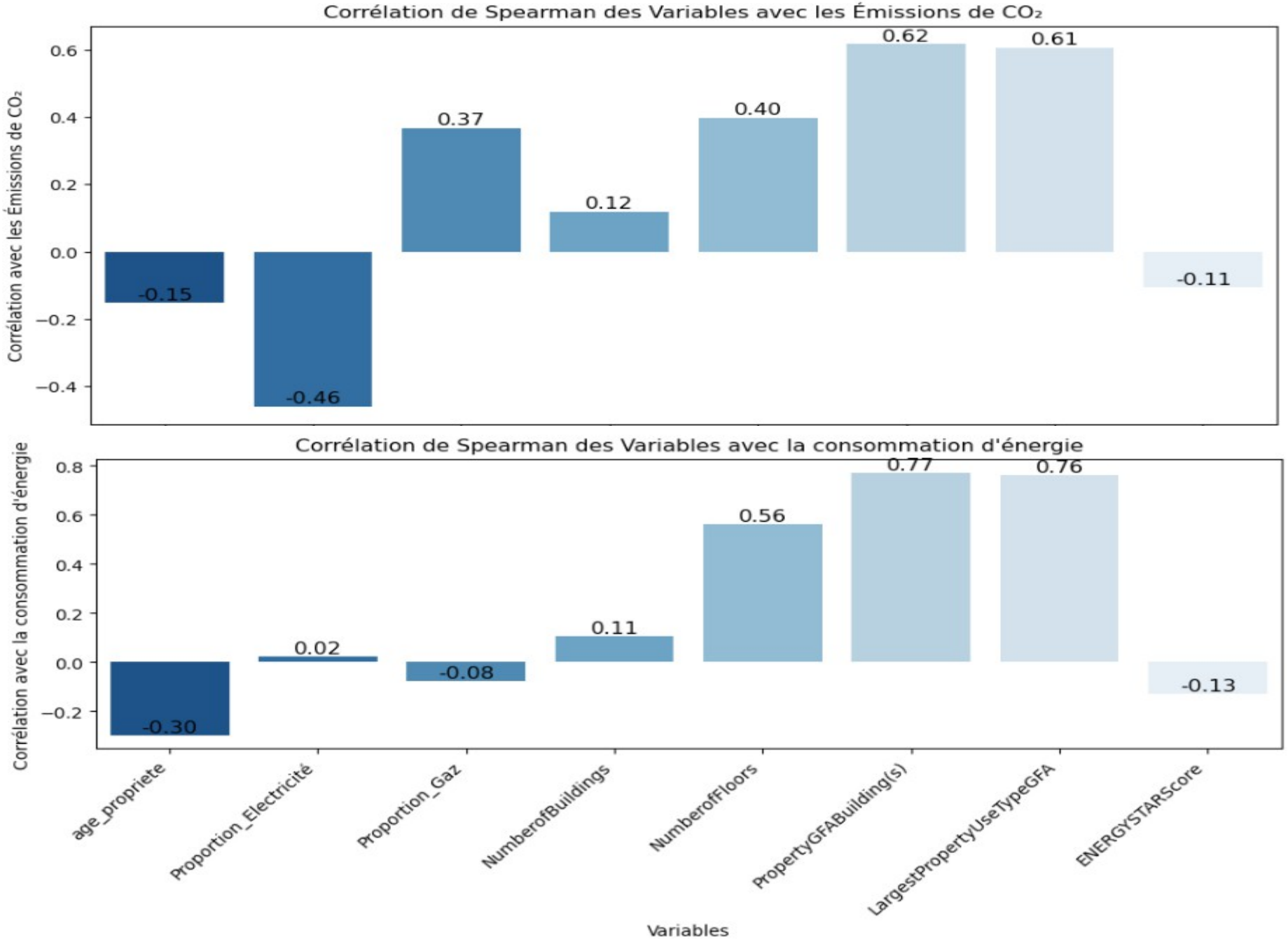
• **Combinaison Taille/Usage** : Interaction entre taille du bâtiment et type d'usage
→ Permet de capturer l'impact combiné sur la consommation d'énergie et les émissions de CO_2.

Analyse des relations entre les variables

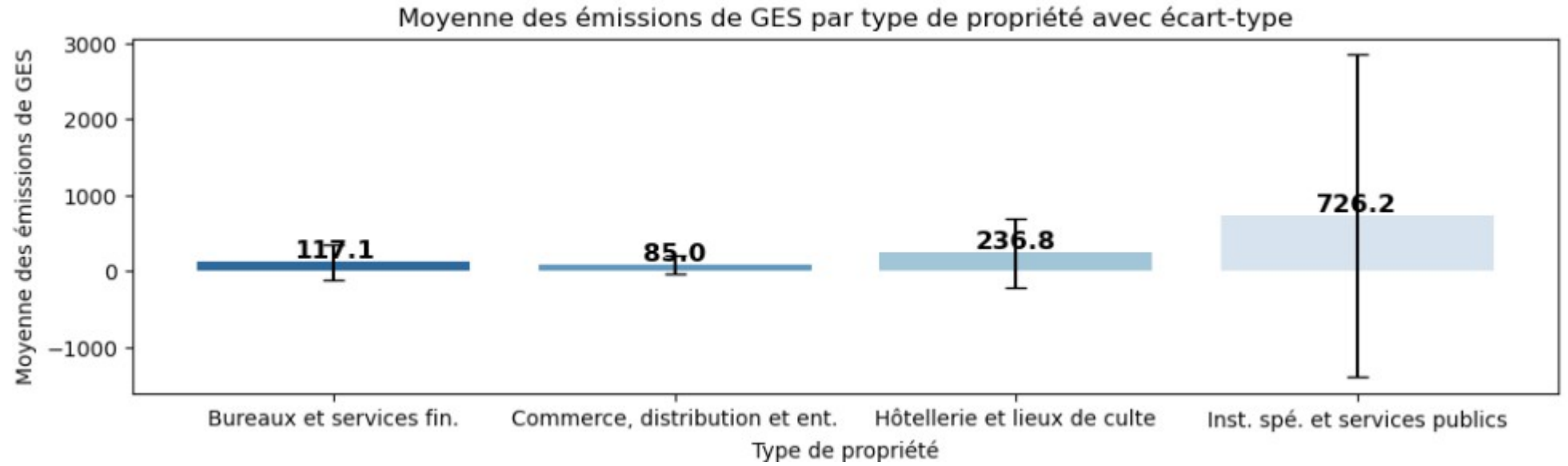
Corrélation entre les variables explicatives

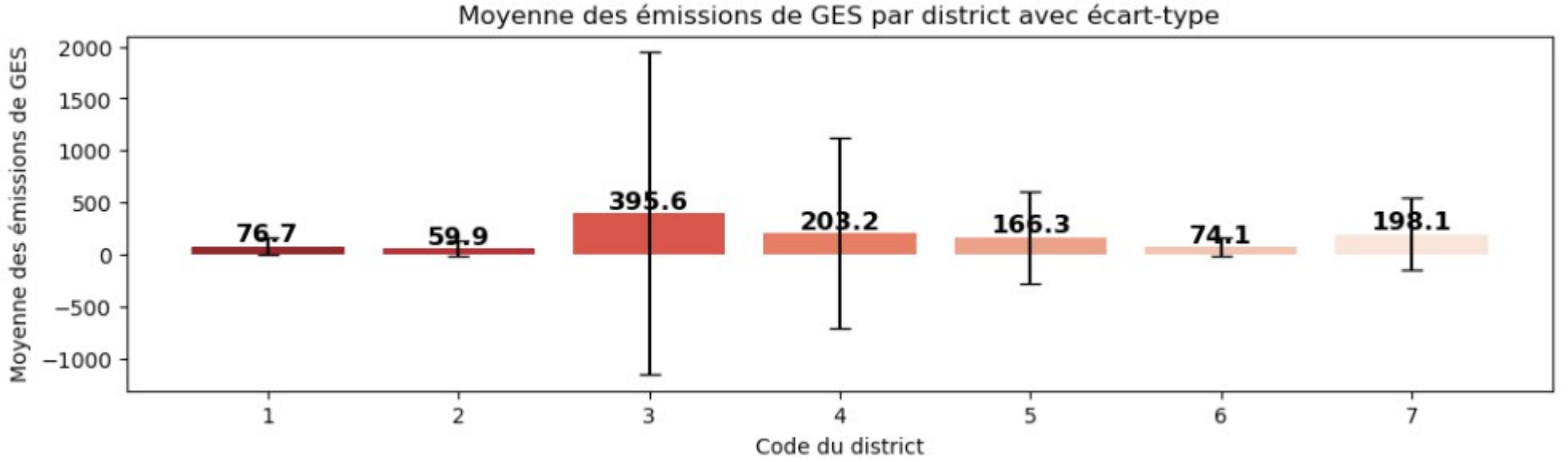


Corrélations des variables avec nos cibles

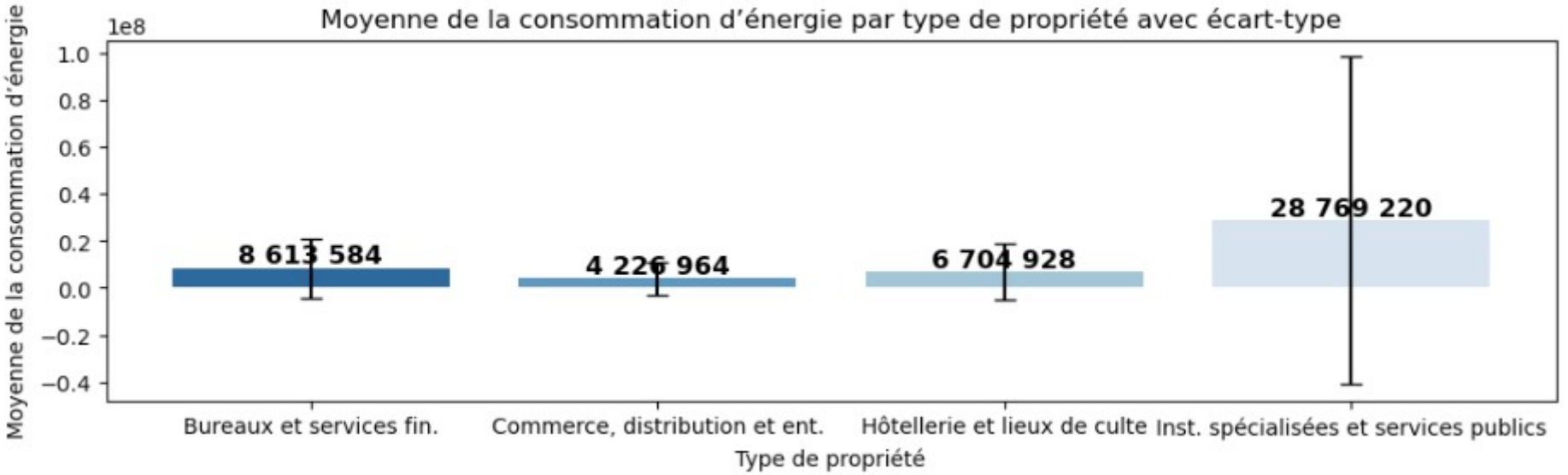


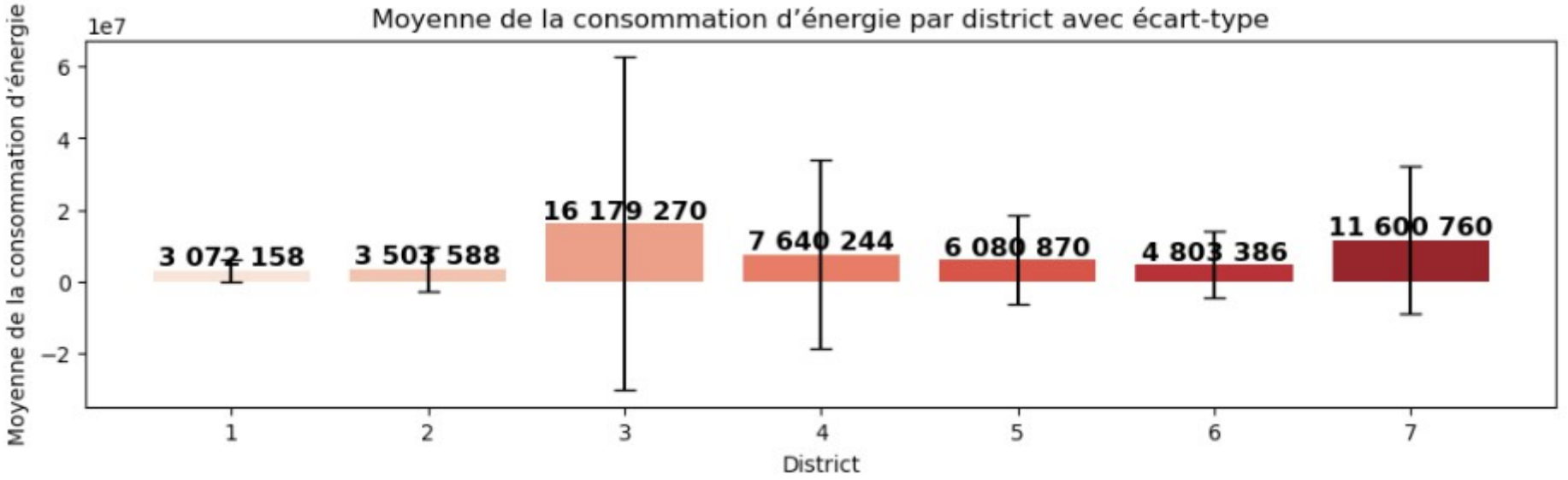
Émissions de GES par Type de Bâtiment et District





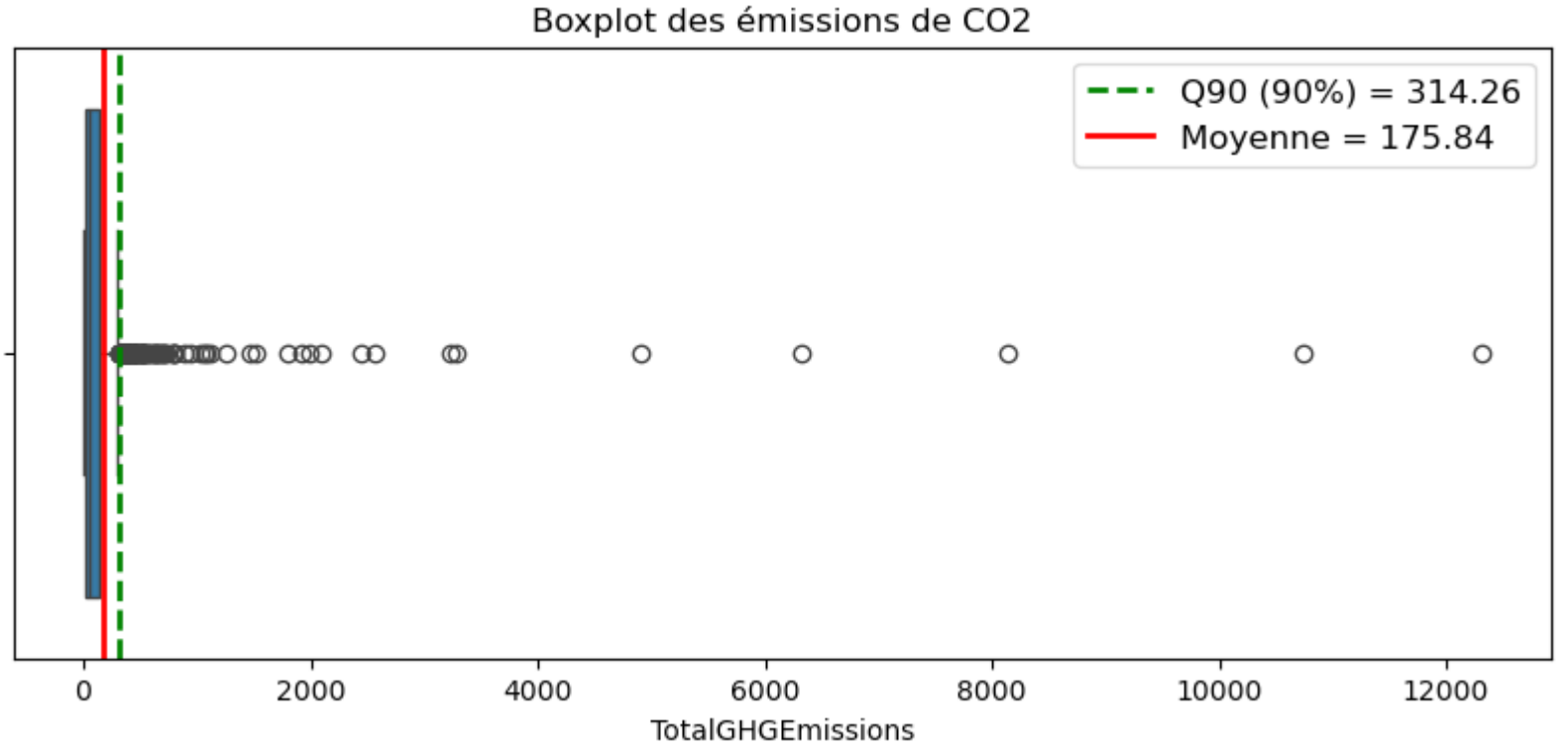
Consommation d'énergie par Type de Bâtiment et District

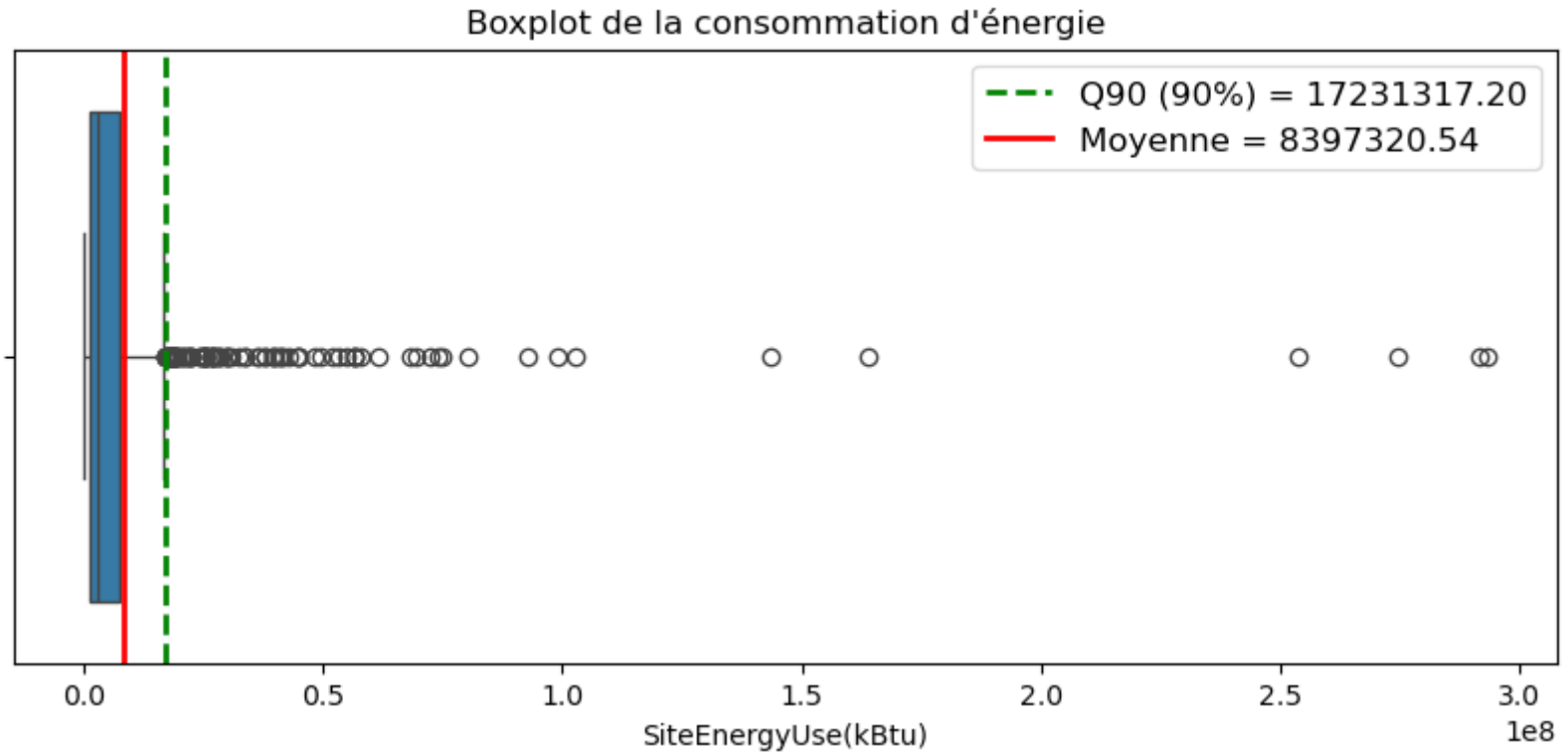




Tranformations des variables

Variables	Transformations
Usage de la propriété	One-hot encoding (avec regroupement pour catégories peu représentées)
District	One-hot encoding
Combinaison Usage/ taille	One-hot encoding
Nombre d'étage	→ **Division en 3 groupes** (1-2 / 3-4 / 5+ étages) → **MinMaxScaler** (pour la normalisation des variables numériques) → **RobustScaler** (pour traiter les outliers) → **Log** (pour rendre les relations plus linéaires)
Nombre de bâtiments	→ **Binarisation** (1 bâtiment / plusieurs bâtiments) → **MinMaxScaler** (pour la normalisation des variables numériques) → **RobustScaler** (pour traiter les outliers) → **Log** (pour rendre les relations plus linéaires)
Taille de la propriété	→ **Discrétisation en 5 groupes** via KBinsDiscretizer (quantiles) → **MinMaxScaler** (pour la normalisation des variables numériques) → **RobustScaler** (pour traiter les outliers) → **Log** (pour rendre les relations plus linéaires)
Taille de l'usage principal	→ **Discrétisation en 5 groupes** via KBinsDiscretizer (quantiles) → **MinMaxScaler** (pour la normalisation des variables numériques) → **RobustScaler** (pour traiter les outliers)
Age de la propriété	→ **MinMaxScaler** (pour la normalisation des variables numériques) → **Sqrt**...
Proportion de gaz	→ **Log +1**


Boxplot des émissions de CO2
Q90 (90%) = 314.26
Moyenne = 175.84
0
2000
4000
6000
8000
10000
12000
TotalGHGEmissions


Boxplot de la consommation d'énergie
Q90 (90%) = 17231317.20
Moyenne = 8397320.54
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e8
SiteEnergyUse(kBtu)

Comparaison des performances des modèles testés

- Objectif : Comparer les performances des modèles pour prédire les émissions de CO_2 et la consommation d'énergie.

- Séparation des données : Les données ont été stratifiées pour garantir une répartition correcte, en particulier pour les valeurs extrêmes.

- Validation croisée stratifiée : Utilisation de la validation croisée pour évaluer les modèles de manière fiable, en garantissant que les sous-ensembles reflètent bien la répartition des données.

Performances des Modèles Testés pour les émissions de co2

Modèle	R2	MAE	RMSE	MAPE	Temps d'entrainement
Régression linéaire	0,53	94,7	321,31	0,84	0,03s
Random Forest	0,65	80,52	329,21	0,68	0,75s
Extreme Gradient Boosting	0,73	75,65	264,65	0,74	0,18s
Support Vector Regression	0,55	93	383,89	0,74	0,08s

Transformations appliquées au meilleur modèle (XGBoost):

- Regroupement des catégories pour la variable Usage
- Encodage One-Hot pour Council District et Usage
- Transformation logarithmique sur la variable cible (Émissions de CO_2)
- Suppression des variables proportion_Gaz et taille de l'usage principal

Optimisation des paramètres

Les scores avant optimisation :

Jeu de donnée	R2	MAE	RMSE	MAPE
Entrainement (train)	1	3,95	11,92	0,04
Test (données jamais vues)	0,57	109,64	602,09	0,61

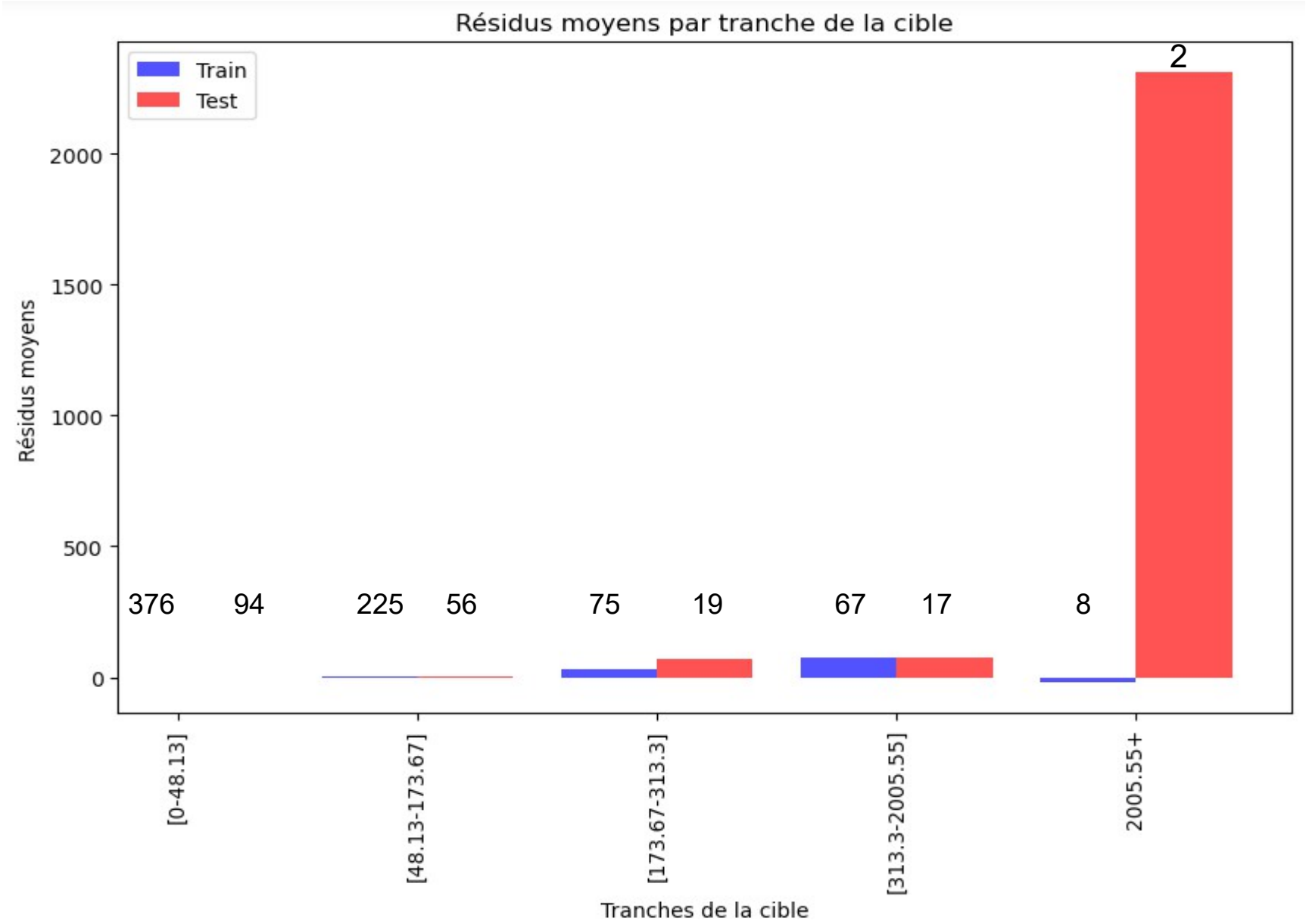
Les scores après optimisation

Jeu de donnée	R2	MAE	RMSE	MAPE
Entrainement (train)	0,96	43,65	233,14	0,33
Test (données jamais vues)	0,84	77,93	338,82	0,62

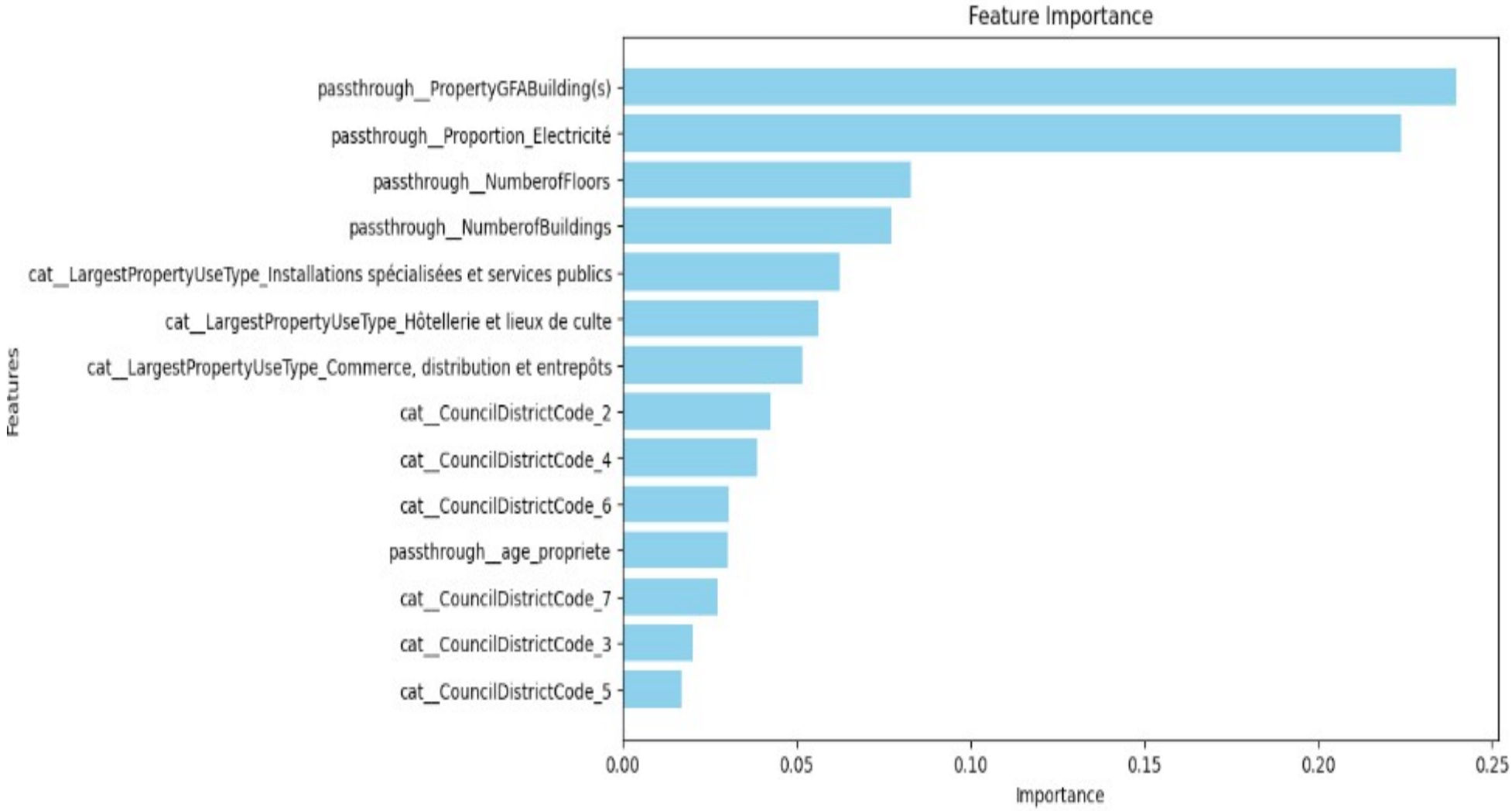
Principaux réglages :

- max_depth = 3 : Limite la profondeur des arbres pour éviter le surapprentissage.
- n_estimators = 100 : Nombre modéré d'arbres, bon compromis entre précision et temps de calcul.
- learning_rate = 0.3 : Apprentissage rapide, bien équilibré avec les autres paramètres.
- min_child_weight = 3 : Évite les divisions inutiles des nœuds, limite la complexité.
- reg_alpha = 0.5, reg_lambda = 0.1 : Régularisation pour contrôler le surajustement.
- subsample = 0.7, colsample_bytree = 0.8 : Randomisation pour améliorer la robustesse du modèle.

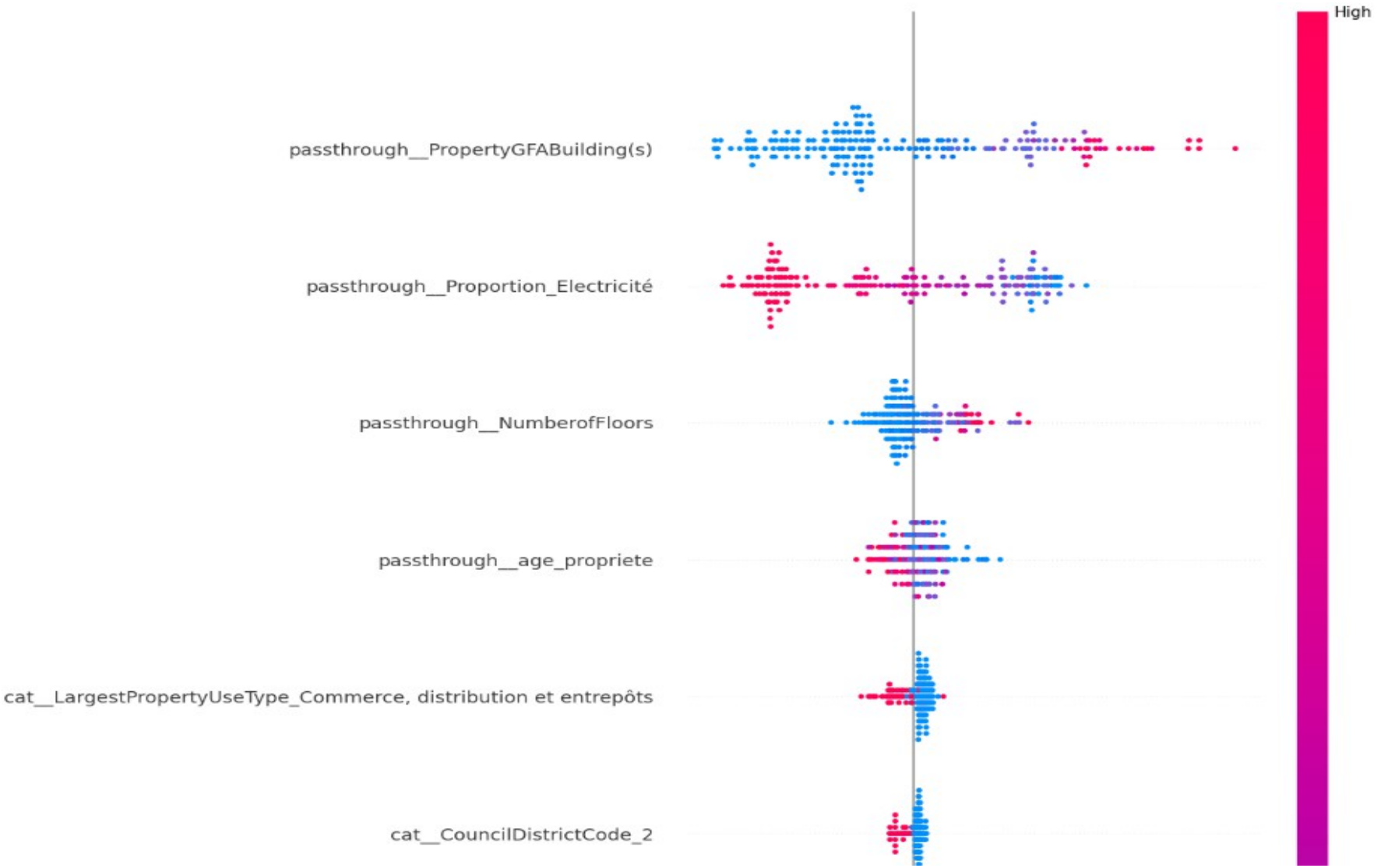
Analyse des résidus – Validation du modèle



Importance des Variables dans le Modèle

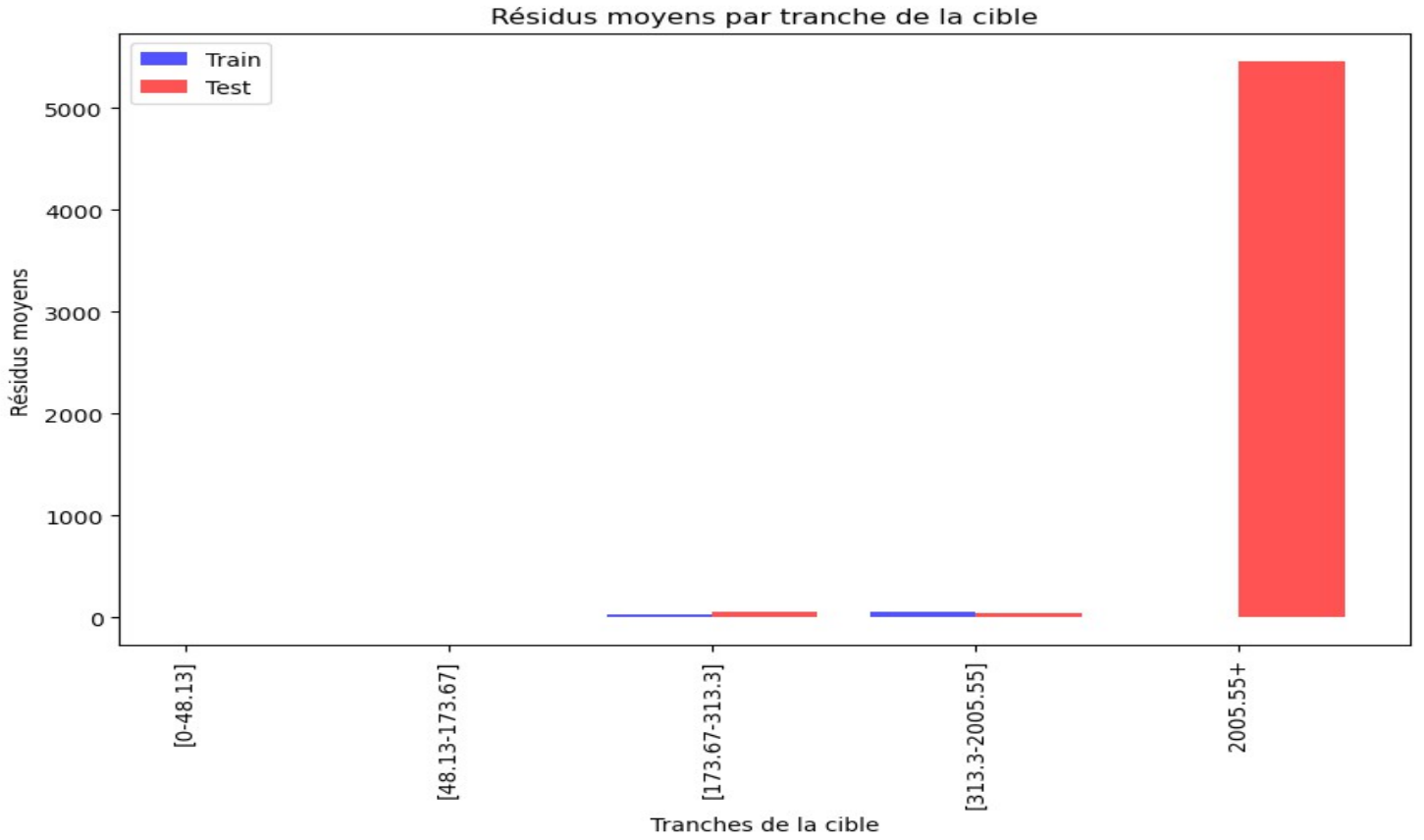


Analyse SHAP – Comprendre l'impact des variables sur les prédictions

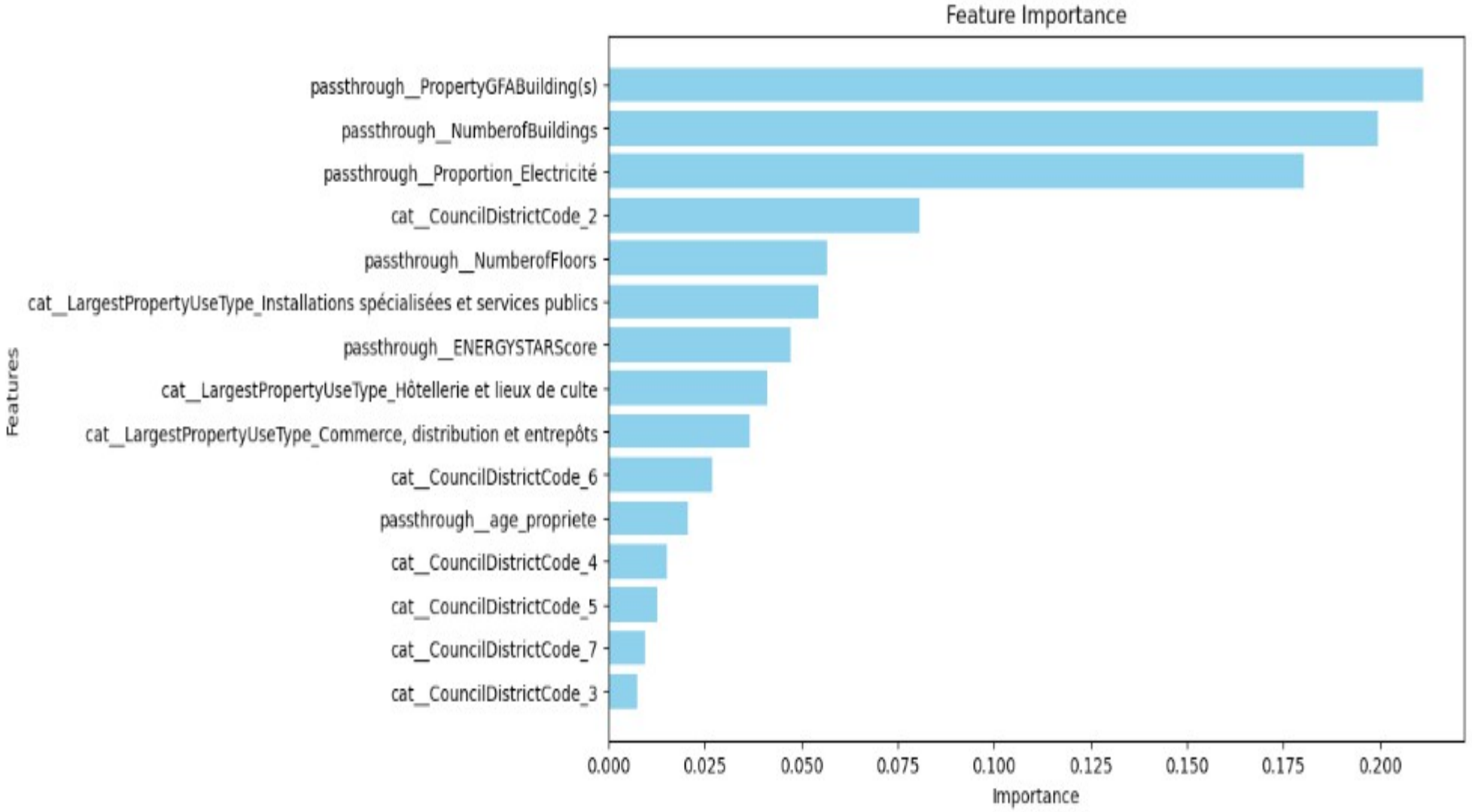


Test de l'impact de la variable 'ENERGYSTARScore' sur les prédictions

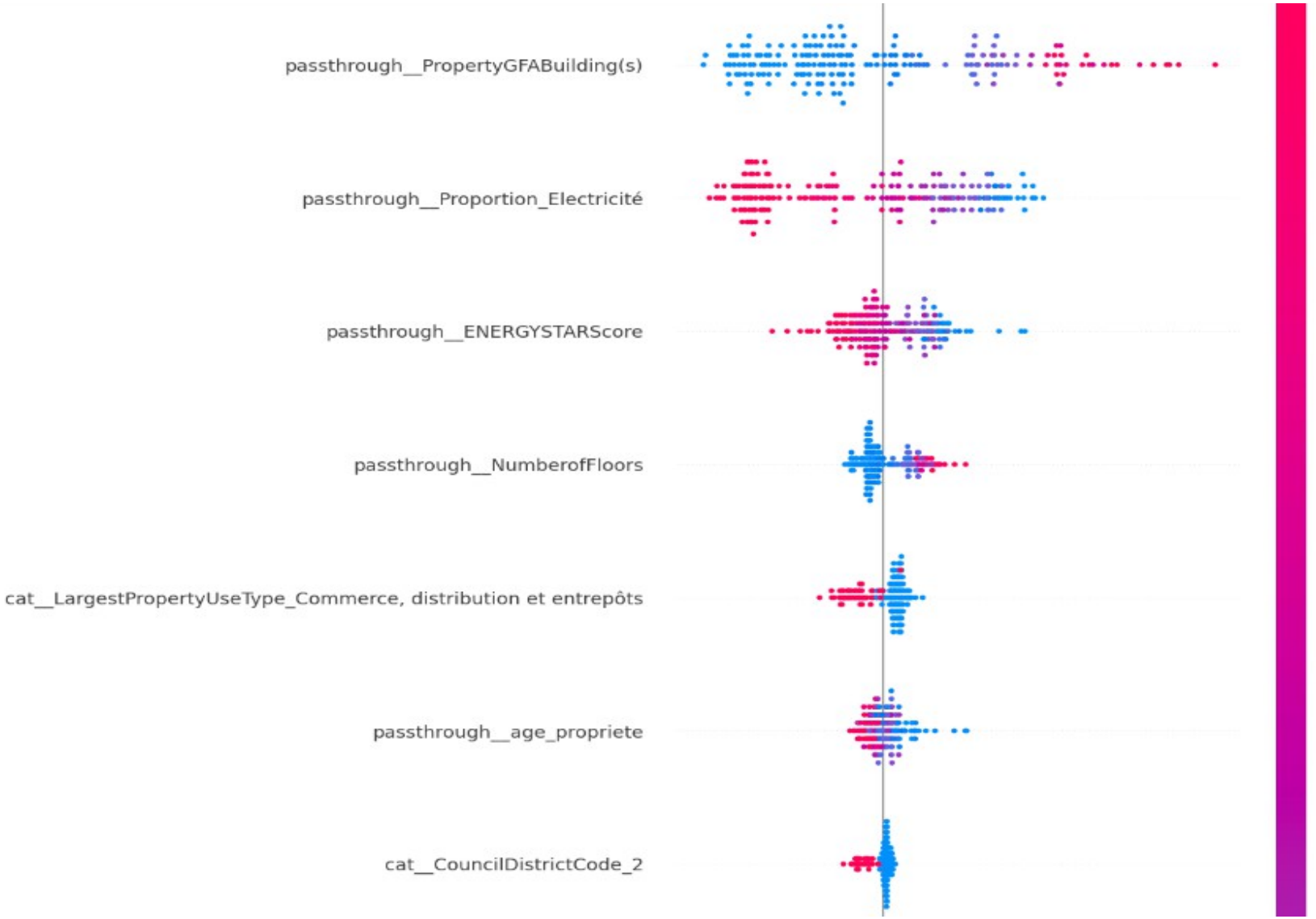
	Jeu de donnée	R2	MAE	RMSE	MAPE
Sans EnergyStarScore	Entrainement (train)	0,96	43,65	233,14	0,33
	Test (données jamais vues)	0,84	77,93	338,82	0,62
Avec EnergyStarScore	Entrainement (train)	0,97	29,46	104,85	0,24
	Test (données jamais vues)	0,60	107,91	580,67	0,49



Importance des Variables dans le Modèle après l'ajout "d'ENERGYSTARScore"



Analyse SHAP – Comprendre l'impact des variables sur les prédictions



Conclusion sur l'ajout de la variable EnergyStarScore

Précision améliorée pour 99% des observations

- Meilleures prédictions pour les bâtiments avec émissions moyennes à élevées

Impact de la nouvelle variable sur l'importance des features

- EnergyStarScore devient une variable significative du modèle

Limites et risques identifiés

- Dégradation sur les valeurs extrêmes
- Impact négatif sur 1% des observations

Performances des Modèles Testés pour la consommation d'energie

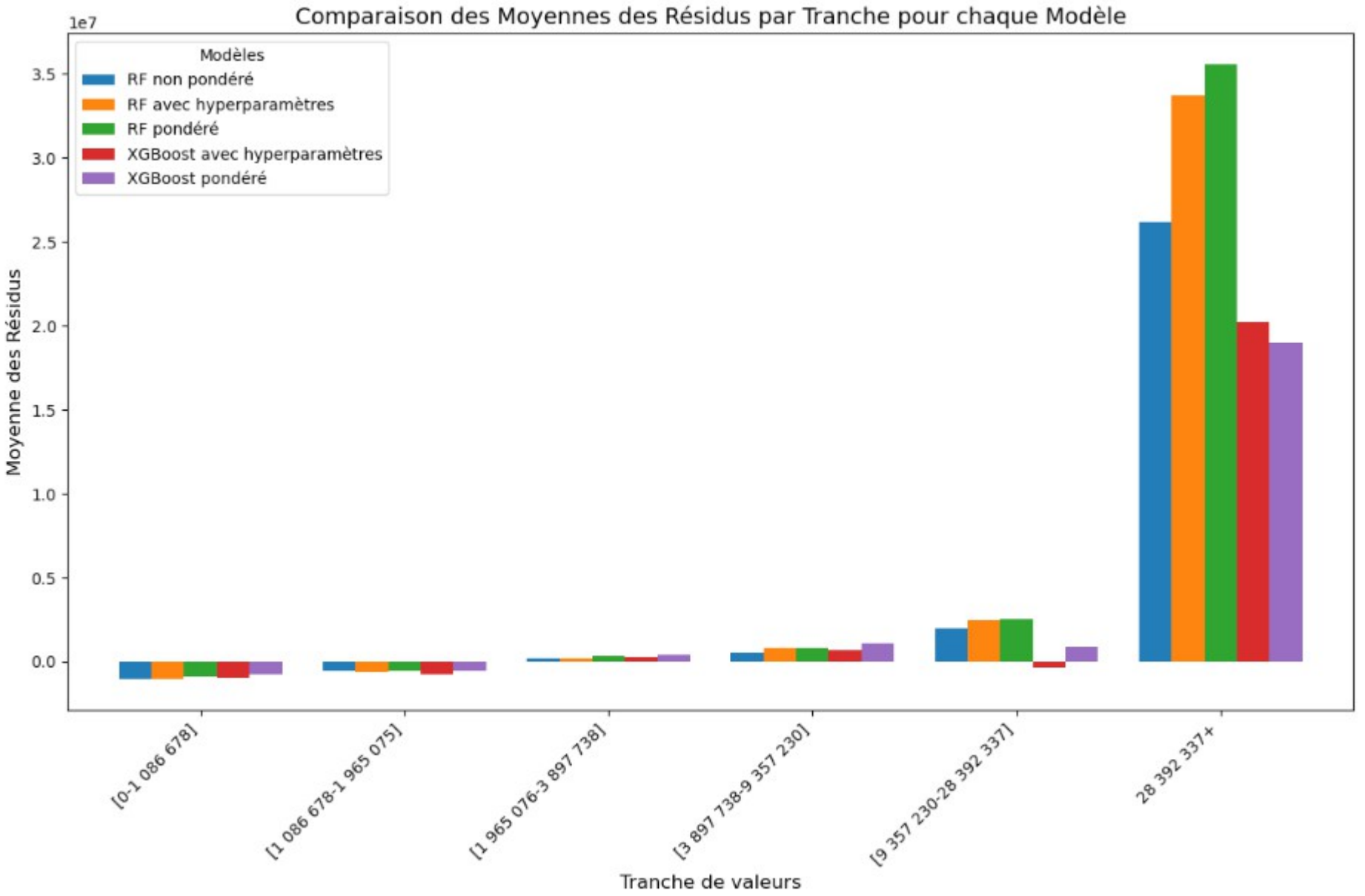
Modèle	R2	MAE	RMSE	MAPE
Régression linéaire	0,60	3 752 061	11 916 381	0,62
Random Forest	0,65	3 590 607	12 062 271	0,64
Extreme Gradient Boosting	0,63	4 065 932	12 587 009	0,66
Support Vector Regression	0,43	4 210 315	15 996 570	0,79

Transformations appliquées au meilleur modèle (XGBoost):

- Regroupement des catégories pour la variable Usage
- Discrétisation du nombre d'étages
- Encodage One-Hot pour Council District et Usage
- Transformation logarithmique sur la variable cible (Conso d'energie)
- Sans taille de l'usage et proportion d'électricité

Transformations appliquées au meilleur modèle (RF):

- Regroupement des catégories pour la variable Usage
- MinMax sur certaines variables
- Combinaison taille/usage
- Transformation logarithmique sur la variable cible (Conso d'energie)
- Sans le district et le nombre de bâtiments


Comparaison des Moyennes des Résidus par Tranche pour chaque Modèle
1e7
Modèles
RF non pondéré
RF avec hyperparamètres
RF pondéré
XGBoost avec hyperparamètres
XGBoost pondéré
Moyenne des Résidus
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
[0-1 086 678]
[1 086 678-1 965 075]
[1 965 076-3 897 738]
[3 897 738-9 357 230]
[9 357 230-28 392 337]
28 392 337+
Tranche de valeurs

Résulats du modèle XGBoost pondéré

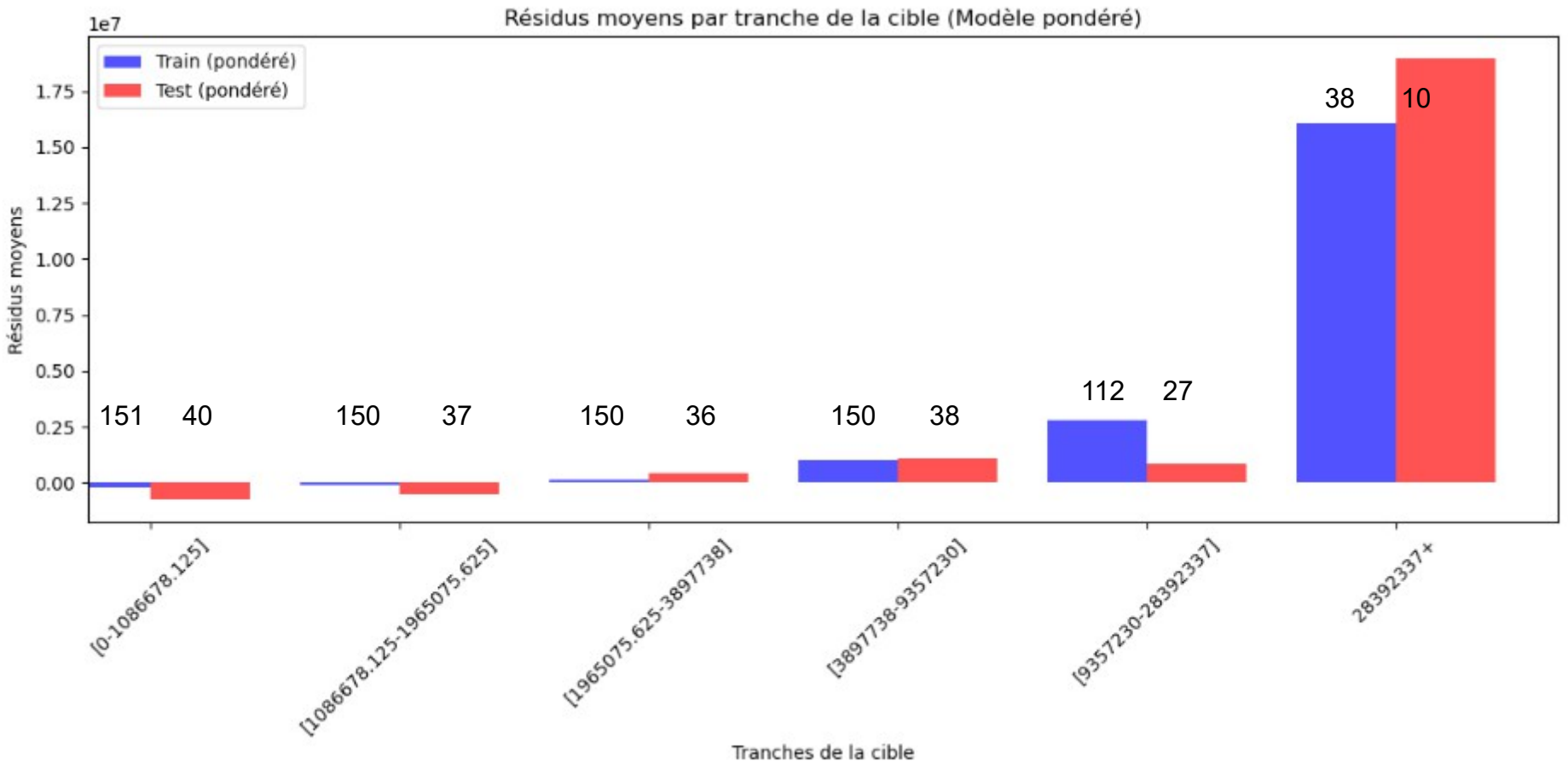
Jeu de donnée	R2	MAE	RMSE	MAPE
Entrainement (train)	0,83	2 424 071	8 927 634	0,29
Test (données jamais vues)	0,88	3 378 029	8 388 802	0,74

Tranche de consommation (%)	Poids
0-20	4,2
21-40	4
41-60	2
61-90	1,5
91-95	1,2

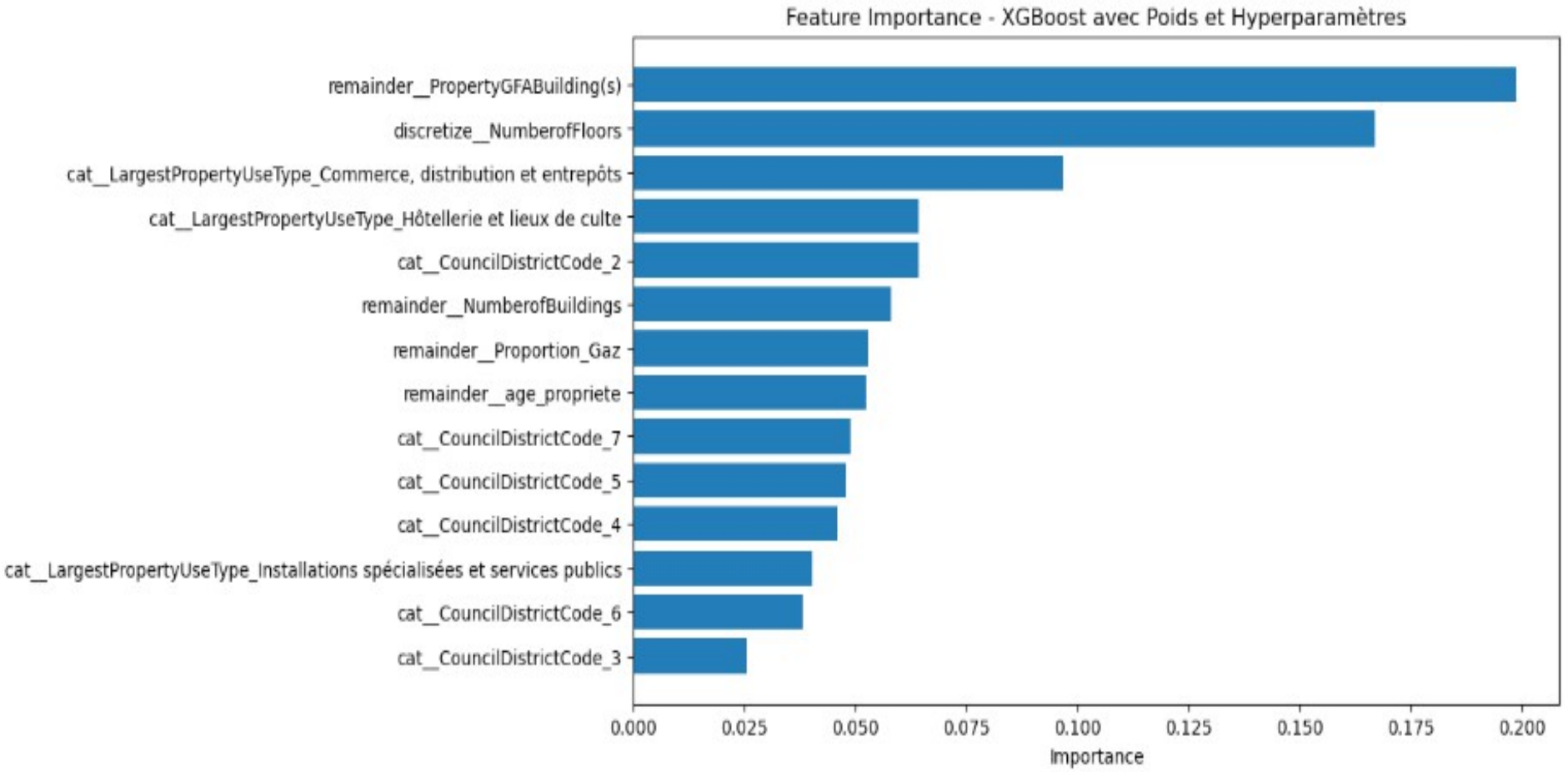
Principaux réglages :

- max_depth = 3 : Limite la profondeur des arbres pour éviter le surapprentissage.
- n_estimators = 300
- learning_rate = 0.08
- min_child_weight = 3 : Évite les divisions inutiles des nœuds, limite la complexité.
- reg_alpha = 0.2, reg_lambda = 0.2 : Régularisation pour contrôler le surajustement.
- subsample = 0.7, colsample_bytree = 0.7 : Randomisation pour améliorer la robustesse du modèle.

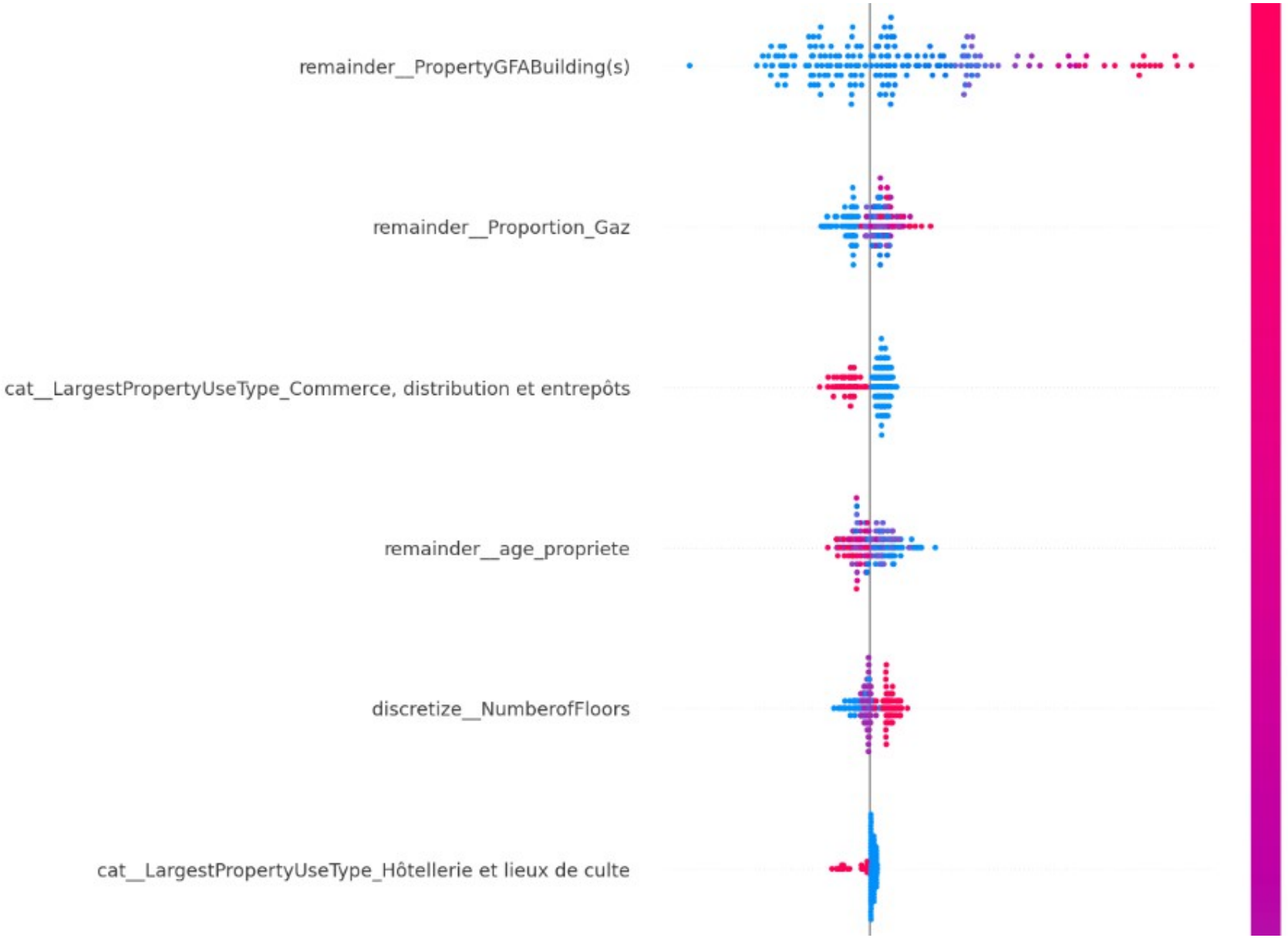
Analyse des résidus



Importance des Variables dans le Modèle



Analyse SHAP – Comprendre l'impact des variables sur les prédictions



Test de l'impact de la variable 'ENERGYSTARScore' sur les prédictions

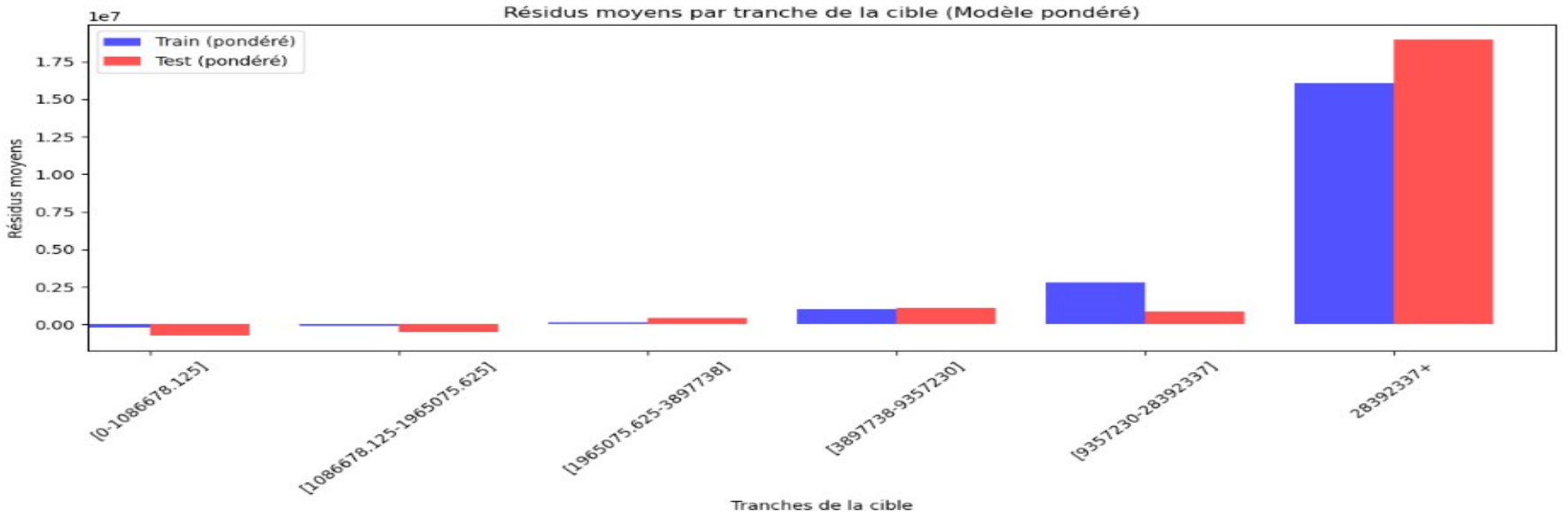
	Jeu de donnée	R2	MAE	RMSE	MAPE
Sans EnergyStarScore	Entrainement (train)	0,83	2 424 071	8 927 634	0,29
	Test (données jamais vues)	0,88	3 378 029	8 388 802	0,74
Avec EnergyStarScore	Entrainement (train)	0,82	1 984 574	9 154 627	0,21
	Test (données jamais vues)	0,87	3 127 149	8 775 521	0,59

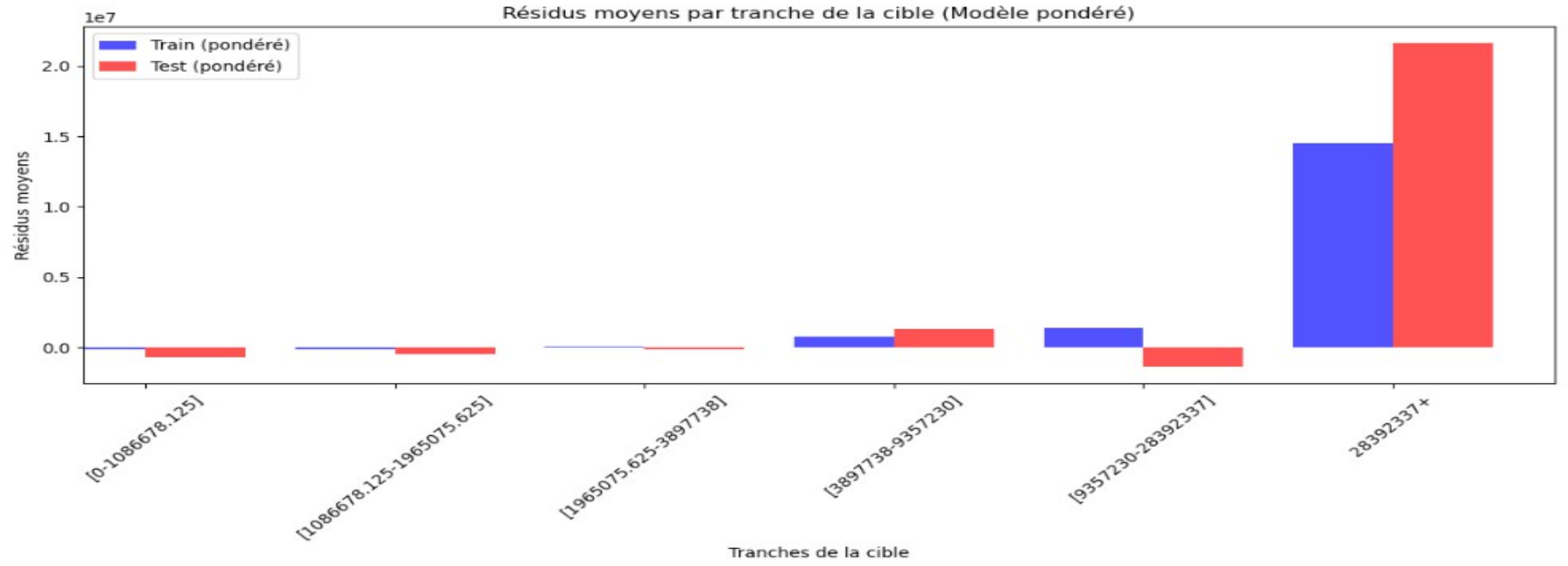
• Légère baisse du R^2 (de 0.88 à 0.87 sur le test) → le modèle explique légèrement moins bien la variabilité des consommations.

• Réduction des erreurs absolues (baisse du MAE et du MAPE) → meilleure précision sur la majorité des prédictions.

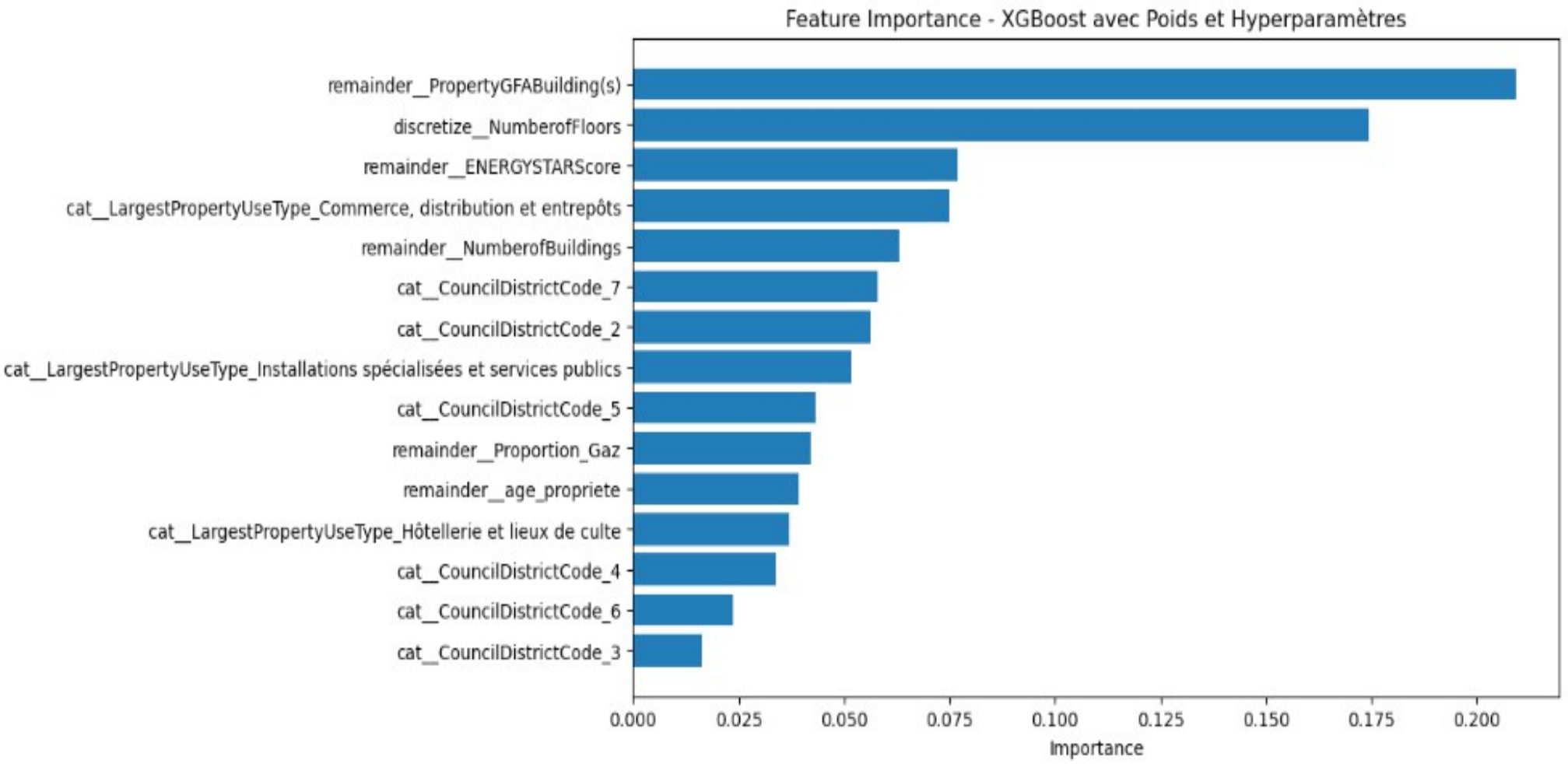
• Légère hausse du RMSE → erreurs plus importantes sur certaines valeurs extrêmes.

Analyse des résidus avant/après

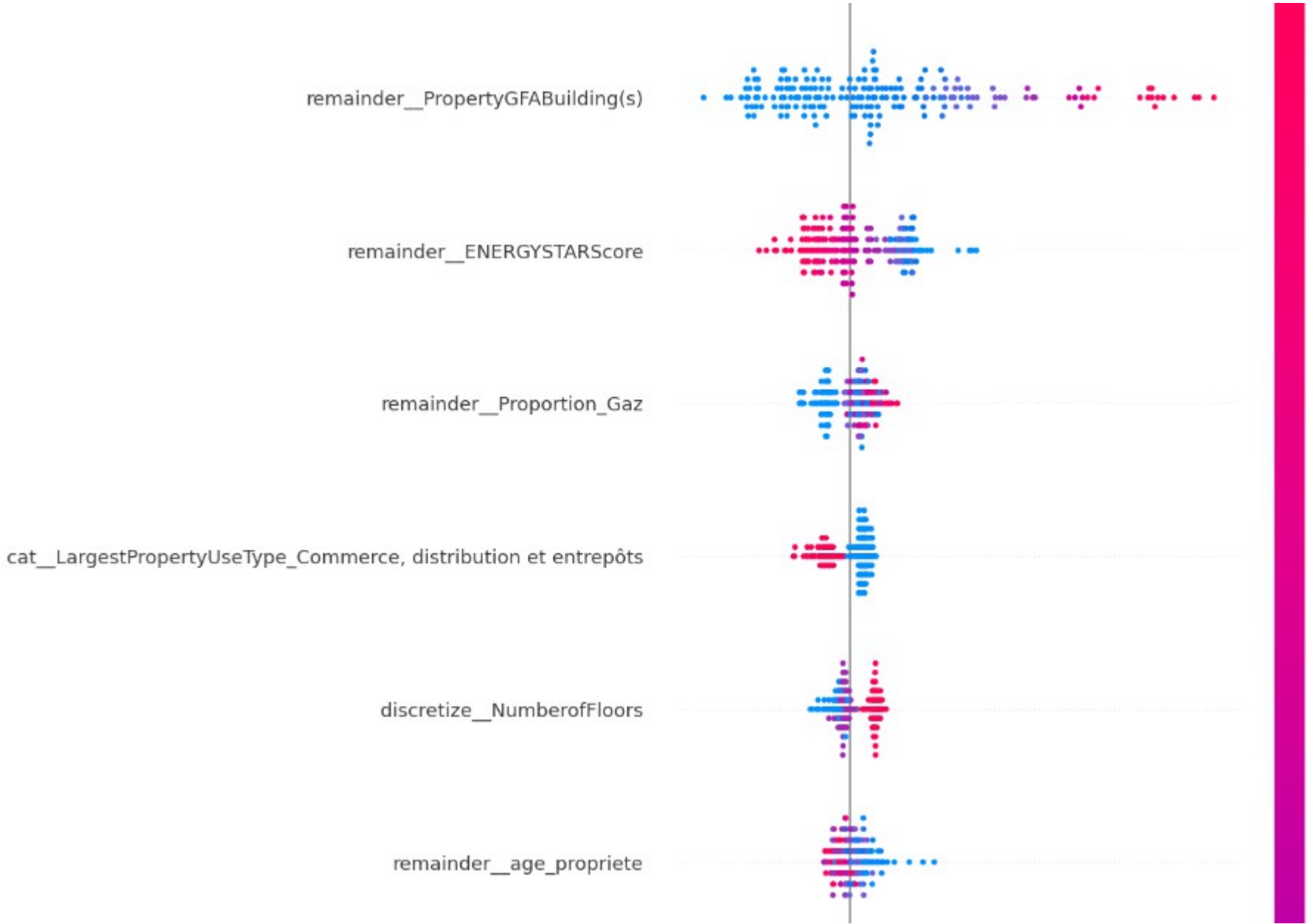




Importance des Variables dans le Modèle après l'ajout "d'ENERGYSTARScore"



Analyse SHAP – Comprendre l'impact des variables sur les prédictions



Conclusion générale sur l'ajout de la variable EnergyStarScore

Impact sur la précision du modèle :

- MAE & MAPE : Amélioration (prédictions plus précises)
- RMSE : Augmentation (erreurs extrêmes plus marquées)

Amélioration par tranche de consommation :

- Prédictions améliorées pour consommations faibles et modérées
- Prédictions dégradées pour consommations élevées

Changements dans l'importance des variables :

- ENERGYSTARScore : Devient un facteur clé

Impact global :

- Amélioration des prédictions pour cas courants
- Nouvelle dimension sans perturbation majeure du modèle

Conclusion du projet

Objectif : Étudier les facteurs influençant les émissions de CO_2 et la consommation d'énergie des bâtiments pour mieux les prédire.

Analyse exploratoire : Sélection des variables pertinentes, identification des facteurs clés (taille, source d'énergie des bâtiments) et absence de fuite de données.

Modélisation : Test et optimisation de plusieurs algorithmes. XGBoost est le plus performant.

Défis : Gestion des valeurs extrêmes et sous-représentation des données affectant les erreurs de prédiction.