Prédiction du prix de vente des biens immobiliers dans les Deux-Sèvres (2023-2024)

Introduction

Ce projet vise à développer un modèle prédictif des prix de vente immobiliers dans le département des Deux-Sèvres, en utilisant les ventes réalisées en 2023 et début 2024.

Deux fichiers sont à disposition :

- train.csv : Contient les données de biens immobiliers vendus, avec leurs prix.
- test.csv : Contient des biens sans indication de prix, nécessitant une prédiction.

L'objectif final est d'estimer les valeurs manquantes de la colonne « valeur_fonciere » du fichier test.csv, en s'appuyant sur un modèle linéaire.

Analyse exploratoire (cf-> graphiques)

1. Avant la modélisation, une analyse préliminaire des données a été menée afin d'identifier les variables explicatives pertinentes.

   Variable cible : valeur_fonciere
   Cette variable présente une dispersion importante. La majorité des transactions tournent autour de 100 000 €, mais certains biens atteignent des valeurs bien supérieures, rendant une prédiction directe complexe.

2. Transformation logarithmique de la surface
   Afin de stabiliser la variance et d'établir une relation plus linéaire entre la surface bâtie et le prix, une transformation logarithmique de la surface bâtie réelle a été appliquée.

3. Variables influentes :

Plusieurs variables corrélées significativement au prix ont été identifiées :

- Surface bâtie : directement liée à la valeur.
- Type de bien (Maison ou Appartement) : ces types de biens présentent des dynamiques de prix très distinctes.
- Code postal : influence marquée de la localisation géographique sur les prix (zones rurales, Niort, etc.).

Ces observations justifient une approche segmentée pour la modélisation.

Modèle retenu : Régression log-log par secteur

Formulation du modèle ->

Un modèle de régression linéaire log-log a été retenu :

Ce modèle implique que le prix évolue comme une puissance de la surface. Les coefficients a et b sont estimés par calcul direct utilisant la covariance et la variance, respectant ainsi la contrainte d'absence de librairie externe.

La prédiction finale est obtenue en revenant à l'échelle réelle :

- Segmentation selon le type de bien

Les maisons et appartements ont été modélisés séparément, considérant leurs différences notables en termes de prix au m^2, distributions de surface et localisation.

- Modèle par code postal

Compte tenu des variations importantes de prix entre secteurs, des modèles distincts ont été établis pour chaque code postal comportant au moins cinq ventes dans train.csv. Cette segmentation améliore la précision tout en évitant l'overfitting.

- Modèle global de secours

Si un code postal présent dans test.csv n'existe pas ou est insuffisamment représenté dans train.csv, un modèle global basé sur l'ensemble des ventes du département pour chaque type de bien est utilisé.

- Évaluation du modèle

Le modèle a été évalué sur le fichier train.csv en simulant des prédictions. Plusieurs métriques standards ont été calculées :

Résultat

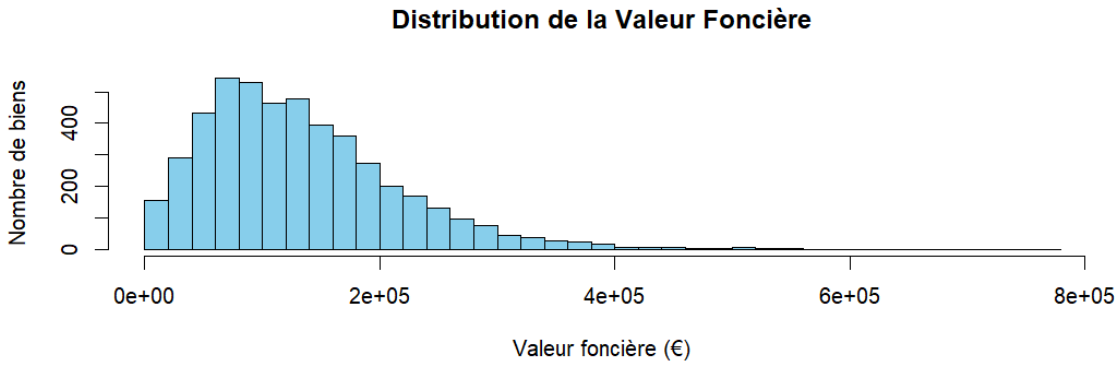
- SCR (Erreur quadratique) :17 474 976 843 111
- R (Corrélation) : 0,7132
- R^2 (Coefficient de détermination) : 0,5087
- MAE (Erreur absolue moyenne) : 44 668 €

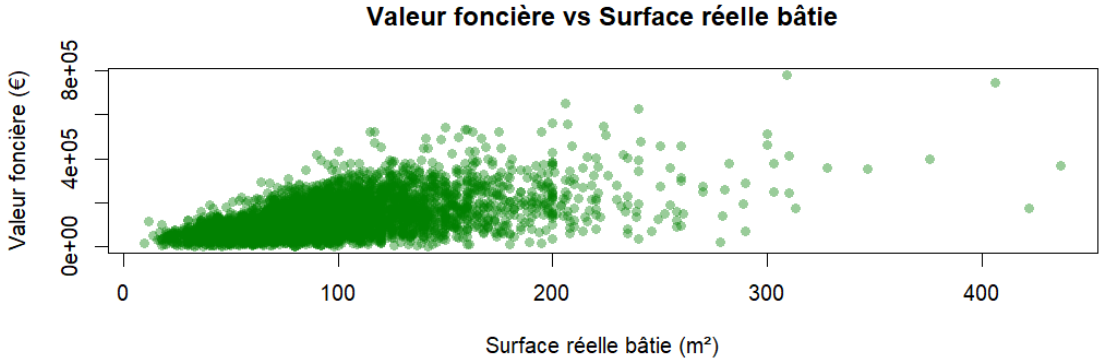
Conclusion

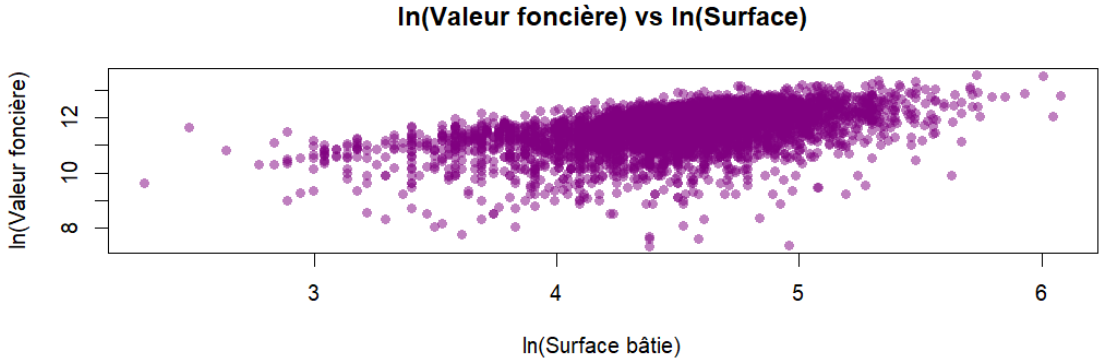
Ce projet a permis de créer un modèle fiable et adapté aux contraintes. En séparant maisons et appartements, et en tenant compte du code postal, on a bien capté les différences de prix dans les Deux-Sèvres avec un modèle log-log.

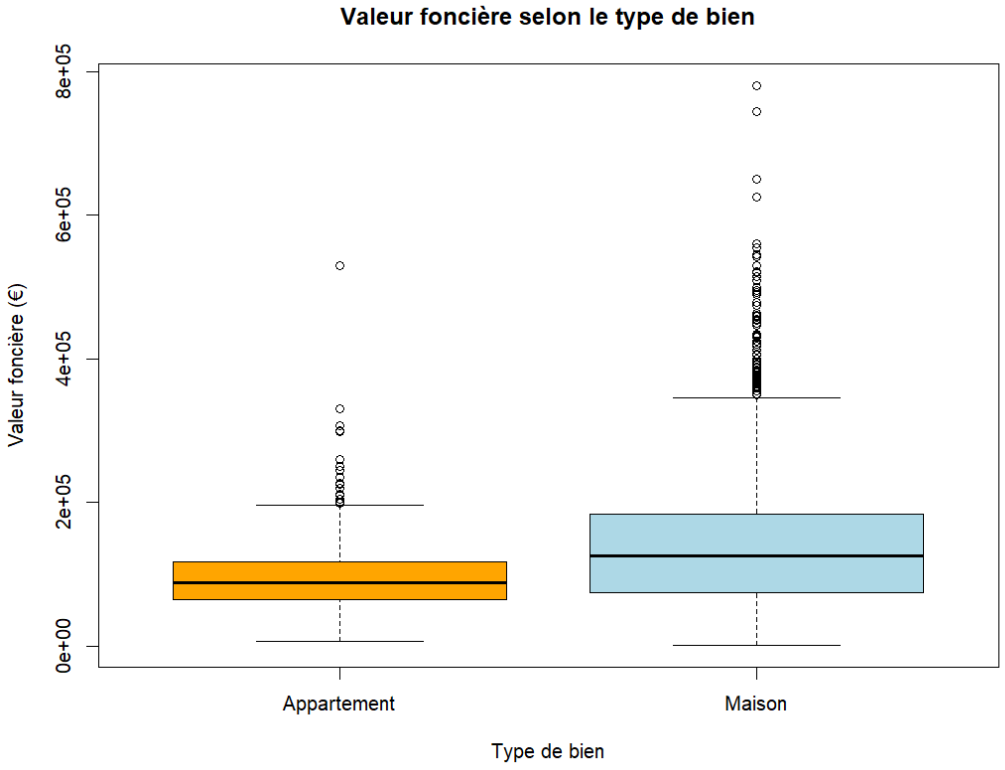
Des améliorations possibles seraient :

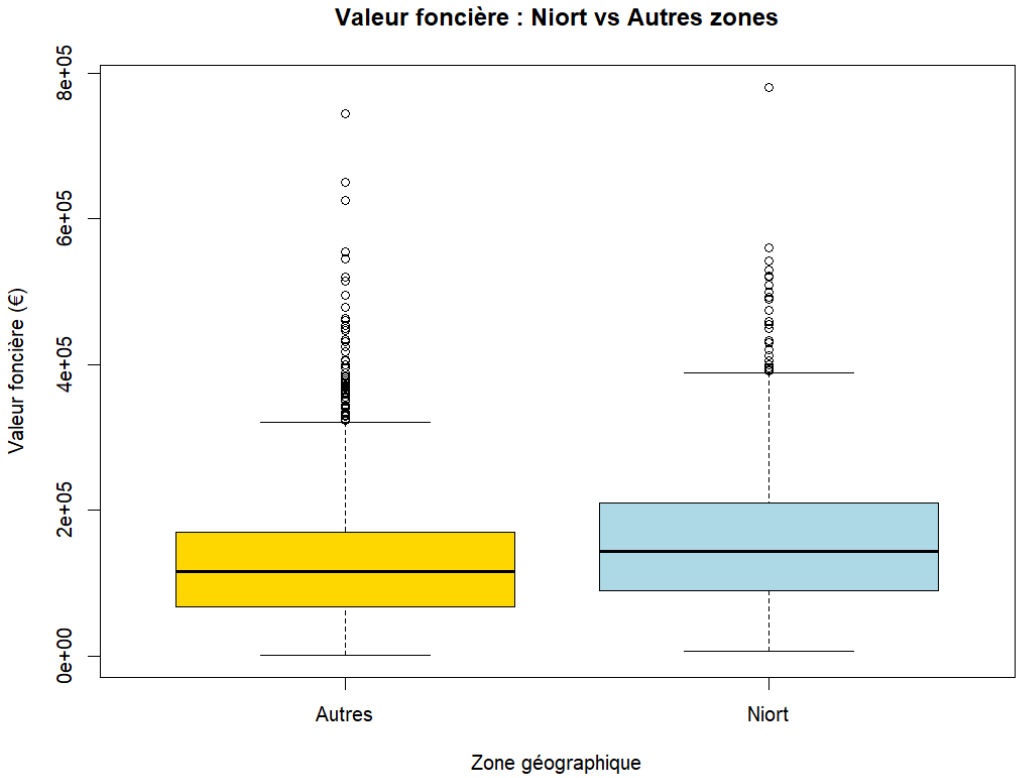
- Ajouter la surface du terrain et le nombre de pièces pour affiner le prix.
- Utiliser l'année de construction pour tenir compte de l'état du bien.
- Mieux localiser les logements, par exemple par quartier, et pas seulement par code postal.


Distribution de la Valeur Foncière
Nombre de biens
0
200
400
0e+00
2e+05
4e+05
6e+05
8e+05
Valeur foncière (€)


Valeur foncière vs Surface réelle bâtie
Valeur foncière (€)
0e+00
4e+05
8e+05
0
100
200
300
400
Surface réelle bâtie (m²)


ln(Valeur foncière) vs ln(Surface)
ln(Valeur foncière)
8
10
12
3
4
5
6
ln(Surface bâtie)


Valeur foncière selon le type de bien
Valeur foncière (€)
0e+00
2e+05
4e+05
6e+05
8e+05
Appartement
Maison
Type de bien


Valeur foncière : Niort vs Autres zones
Valeur foncière (€)
0e+00
2e+05
4e+05
6e+05
8e+05
Autres
Niort
Zone géographique