Régréssion sur des données réelles

Théo PETIT – Jérémy PITON

04/04/25

Introduction

Ce projet de régression sur des données réelles a pour but de prédire le prix de l'immobilier dans les Deux-Sèvres. L'objectif est d'établir un modèle qui correspond au mieux à la valeur foncière des biens.

Nous avions un fichier CSV, nommé *Train*, qui correspond aux ventes immobilières de 2023 en Deux-Sèvres et qui contient de nombreuses informations telles que la localisation du bien, le nombre de pièces, la superficie… et la valeur foncière. Il nous servait pour nos recherches et nos tests de prédiction.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id		Date.mutati	Nature.muta	Code.postal	Commune	Code.depart	Type.local	Surface.reell	Nombre.piec	Surface.terra	Valeur.fonciere		SURFACE CC	Surface Terrain
2		1	06/01/2023	Vente	79100	PLAINE-ET-VA	79	Maison	60	1	948	122000		60	948
3		2	09/01/2023	Vente	79000	NIORT	79	Appartement	43	2	NA	73500		0	0
4		3	06/01/2023	Vente	79110	CHEF-BOUTO	79	Maison	92	4	244	116000		92	244
5		5	09/01/2023	Vente	79390	DOUX	79	Maison	114	4	1603	147000		114	0
6		6	06/01/2023	Vente	79000	NIORT	79	Appartement	28	2	NA	53000		0	0
7		7	04/01/2023	Vente	79100	THOUARS	79	Maison	89	4	79	80581		89	0
8		8	10/01/2023	Vente	79800	LA MOTHE-SA	79	Maison	163	5	275	170000		0	275
9		9	02/01/2023	Vente	79100	THOUARS	79	Maison	80	6	707	150000		80	707
10		10	09/01/2023	Vente	79000	NIORT	79	Appartement	37	1	NA	95000		0	0
11		11	03/01/2023	Vente	79100	THOUARS	79	Maison	49	3	312	69000		0	312
12		12	05/01/2023	Vente	79100	THOUARS	79	Maison	99	4	1046	122800		99	1046
13		13	10/01/2023	Vente	79100	PAS-DE-JEU	79	Maison	50	0	466	6500		0	466
14		14	03/01/2023	Vente	79700	MAULEON	79	Maison	72	4	551	160000		72	551
15		15	03/01/2023	Vente	79350	CHICHE	79	Maison	81	3	298	65000		81	298
16		16	10/01/2023	Vente	79100	LOUZY	79	Maison	167	4	1610	255000		0	0
17		17	10/01/2023	Vente	79220	CHAMPDENI	79	Maison	97	3	577	43500		97	577
18		18	02/01/2023	Vente	79120	LEZAY	79	Maison	72	3	495	27500		72	495
19		19	09/01/2023	Vente	79370	BEAUSSAIS-V	79	Maison	113	3	928	115000		113	928
20		21	02/01/2023	Vente	79800	SOUDAN	79	Maison	51	2	172	63000		0	172
21		22	02/01/2023	Vente	79310	SAINT-PARD	79	Maison	119	4	678	155000		119	678
22		23	04/01/2023	Vente	79320	MONCOUTAI	79	Maison	90	4	84	66254		90	0

Nous avions également un autre fichier, nommé *Test*, qui contient les mêmes variables que *Train*, mais sans le prix de vente. C'est sur celui-ci qu'il faut prédire les valeurs foncières. Notre objectif est de rendre un tableau avec l'identifiant des logements concernés et leur valeur foncière prédite pour vérifier qu'elle soit le plus juste possible.

Les recherches

Notre objectif était d'étudier les communes en fonction du nombre d'habitants afin de réaliser une analyse pertinente. Pour ce faire, nous avons cherché à regrouper les communes en différentes catégories selon leur population.

Nous avons commencé par rechercher un fichier de l'INSEE contenant les codes postaux ainsi que le nombre d'habitants par commune. Cependant, un problème s'est posé : l'INSEE n'utilise pas les codes postaux mais les codes INSEE des communes. Nous avons donc dû identifier un fichier permettant de faire la correspondance entre ces codes INSEE et les codes postaux.

Une difficulté supplémentaire est apparue : certains codes postaux correspondent à plusieurs communes. Pour pallier ce problème, nous avons décidé d'utiliser les noms de communes comme identifiants uniques afin d'associer correctement un nombre d'habitants à chaque localité.

Une fois les données consolidées, nous avons établi quatre grands groupes en fonction du nombre d'habitants :

- Moins de 500 habitants
- Entre 500 et 1 000 habitants
- Entre 1 000 et 10 000 habitants

• Plus de 10 000 habitants

Ce choix s'est imposé car il nous semblait pertinent d'analyser les modèles en fonction de la taille des communes, qui peut influencer les tendances observées.

Pour assurer la qualité et la cohérence de notre jeu de données, nous avons réalisé plusieurs traitements de mise en forme à l'aide d'intelligence artificielle. Parmi ces ajustements, nous avons :

• Uniformisé les accents et apostrophes

• Mis en majuscule les noms de communes

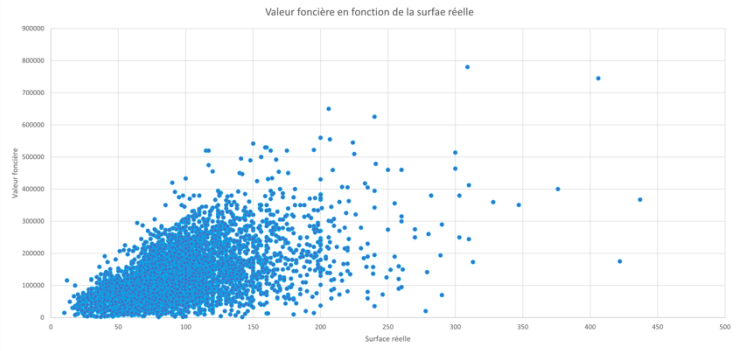
• Supprimé les caractères spéciaux

• Standardisé les abréviations (par exemple, conversion de « ST » en « Saint » et inversement)

Ces corrections ont pris du temps mais étaient essentielles pour garantir la fiabilité des données utilisées.

Afin d'améliorer la précision de notre modèle, nous avons choisi d'éliminer les valeurs extrêmes en supprimant les 10 % des valeurs les plus basses et les 10 % des valeurs les plus élevées des valeurs foncières. Ce nettoyage nous a permis de réduire notre base de données de **4 785 à 3 835 enregistrements**.

Enfin, nous avons décidé de croiser plusieurs variables afin de mieux comprendre les tendances : la surface réelle du bâti, le nombre de pièces et la commune d'appartenance (classée dans nos quatre grands groupes). Ce croisement nous a permis d'affiner notre analyse et d'identifier des corrélations pertinentes entre ces différentes variables.



Grâce à ces étapes rigoureuses de collecte, de nettoyage et d'analyse des données, nous avons pu constituer un jeu de données fiable et représentatif. Ce travail nous permet d'obtenir des modèles plus précis et adaptés aux spécificités des différentes communes en fonction de leur taille et de leur population.

Choix du modèle

Après avoir établi nos catégories, nous avons testé quatre modèles afin de déterminer lequel était le plus adapté à notre répartition des données.

Pour cela, nous avons filtré les données selon deux critères :

• Le groupe d'habitants, pour les maisons

• Le nombre de pièces, pour les appartements

Cette distinction nous a semblé pertinente, car les appartements représentaient une part bien plus faible du jeu de données par rapport aux maisons (environ **10 %**).

Nous avons ensuite évalué chaque modèle en calculant **la somme des résidus au carré**, notre objectif étant d'obtenir la valeur la plus faible possible pour optimiser la précision de la prédiction. Nous avons testé les modèles **exponentiel, puissance, linéaire et logarithmique**.

À l'issue de ces analyses, le **modèle logarithmique s'est révélé être le plus adapté** à notre jeu de données. De manière surprenante, il s'est avéré optimal pour toutes les catégories que nous avons définies.

Application du modèle

Sur R, nous avons commencé par importer les deux fichiers nécessaires à notre analyse : les fichiers de données d'entraînement (*Train*) et de test (*Test*). Une fois ces fichiers chargés, nous avons appliqué un premier traitement en supprimant, comme mentionné précédemment, les valeurs aberrantes du fichier *Train*.

Ensuite, nous avons segmenté notre jeu de données en différentes catégories afin d'appliquer notre modèle de manière plus précise. Les catégories définies sont les suivantes :

• Maison - moins de 500 habitants

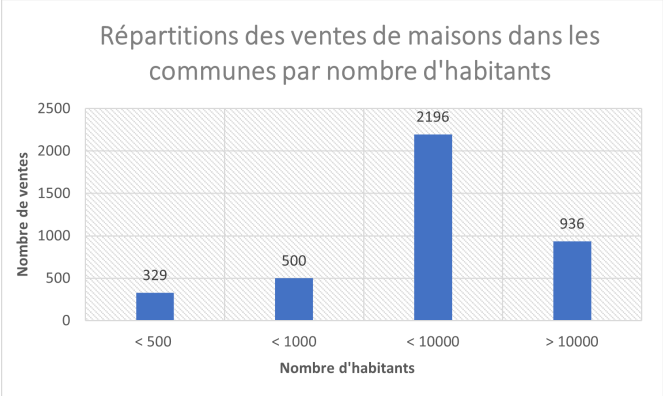
• Maison - entre 500 et 1 000 habitants

• Maison - entre 1 000 et 10 000 habitants

• Maison - plus de 10 000 habitants

• Appartement - moins de 2 pièces

• Appartement - 2 pièces ou plus

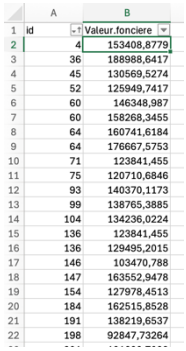


Une fois cette segmentation réalisée, nous avons procédé à l'estimation des coefficients A et B afin de retrouver l'équation du modèle pour chaque catégorie. Cela nous a permis d'estimer les valeurs foncières prédictives et de les comparer aux valeurs du fichier *Train*.

Après avoir appliqué ces modèles, nous avons regroupé les logements en fonction des filtres préalablement définis dans le fichier *Train*. Chaque catégorie a été traitée indépendamment, générant plusieurs DataFrames contenant les valeurs prédites.

Afin d'obtenir un fichier consolidé, nous avons rassemblé ces différentes tables en un seul DataFrame contenant l'identifiant des logements (ID) et leur valeur foncière prédite (Valeur.fonciere).

Pour finaliser le traitement, nous avons utilisé la fonction rbind() afin d'assembler l'ensemble des prédictions en un seul tableau. Enfin, nous avons exporté ce fichier final au format CSV2 pour une exploitation ultérieure.



	A	B
1	id	Valeur.fonciere
2	4	153408,8779
3	36	188988,6417
4	45	130569,5274
5	52	125949,7417
6	60	146348,987
7	60	158268,3455
8	64	160741,6184
9	64	176667,5753
10	71	123841,455
11	75	120710,6846
12	93	140370,1173
13	99	138765,3885
14	104	134236,0224
15	136	123841,455
16	136	129495,2015
17	146	103470,788
18	147	163552,9478
19	154	127978,4513
20	184	162515,8528
21	191	138219,6537
22	198	92847,73264

Conclusion

Pour conclure, nous avons choisi d'utiliser **le modèle logarithmique** pour prédire les valeurs foncières en fonction des différentes caractéristiques immobilières. Ce modèle nous a permis d'intégrer des variables clés telles que **la surface, le type de logement, la localisation et le nombre de pièces**, afin d'optimiser la précision des prédictions.

Grâce à cette approche, nous avons pu obtenir des estimations cohérentes et adaptées aux différentes catégories de biens immobiliers analysées.

Nous avons trouvé ce projet particulièrement intéressant, notamment parce qu'il repose sur des données réelles. De plus, il nous a permis d'explorer la manière d'expliquer une variable à partir d'une ou plusieurs autres. Le fait de partir d'un fichier brut pour en extraire des conclusions pertinentes et produire un fichier de sortie structuré a également été un exercice enrichissant.

Enfin, pour faire le bilan, ce projet nous a permis d'explorer en profondeur **R**, en manipulant des propriétés et des formules complexes, tout en intégrant les mathématiques pour la recherche du modèle optimal. De plus, il nous a conduits à suivre les étapes essentielles à la réalisation d'un compte rendu structuré :

• **Acquérir** les données,

• **Filtrer** les informations pertinentes,

• **Analyser** les résultats obtenus,

• **Tirer** des conclusions pertinentes,

• **Présenter** un rapport clair et concis.