Basa Benjamin

Introduction au domaine de recherche

Octobre 2022

–

Master deux de mathématiques, spécialité statistiques - Sorbonne Université (obtenu en septembre 2021)
Département de mathématiques et applications - Ecole normale supérieure

Estimation du patrimoine immobilier français par des méthodes de machine learning

Institut national de la statistique et des études économiques (Insee) – Département des Etudes Economiques (D2E)

Encadré par Mathias André et Olivier Meslin (D2E)

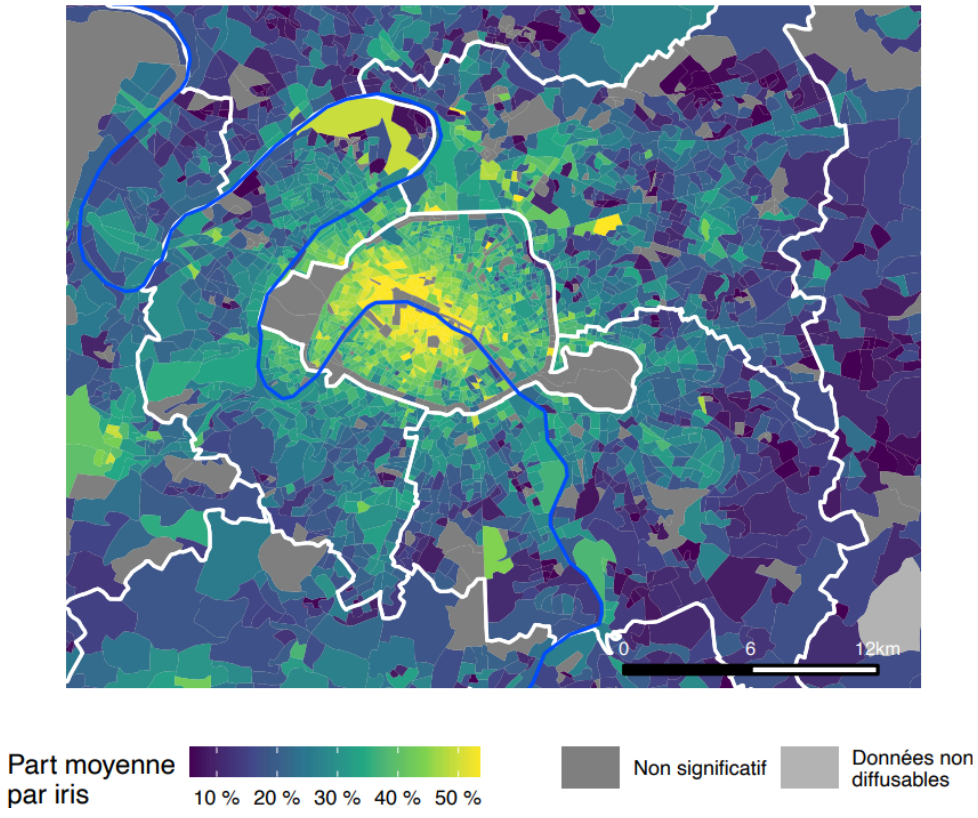

FIGURE 1 – Part des logements possédés par des ménages possédant 5 logements ou plus en région parisienne

Résumé

Dans ce rapport on s'intéresse à une estimation financière, au prix de marché, des logements (maisons et appartements) situés en France métropolitaine, afin d'en tirer des analyses sur le patrimoine immobilier détenu par les Français. Ce travail se distingue d'autres recherches similaires avant tout par son amplitude (il prend en compte l'intégralité des logements situés en France métropolitaine) et par son double objectif méthodologique et économique.

Nous commençons par estimer le prix de chaque logement situé en France à l'aide de modèles de machine-learning, puis nous utilisons ces estimations pour étudier économiquement le patrimoine immobilier des Français.

Au stade actuel, nous pouvons conclure à des résultats utilisables sur des études macro-économiques fiables, mais à améliorer en vue d'analyses économiques plus fines. Nous avons notamment mis en avant le poids prépondérant des ménages propriétaires de plusieurs logements, ainsi que la concentration marquée du patrimoine immobilier Français.

Première partie

Introduction au domaine de recherche

Table des matières

1 Introduction

1.1 Présentation sommaire du machine learning

Dans un contexte actuel de collecte et d'exploitation de données de plus en plus massives, les algorithmes dits de machine learning prennent une place de plus en plus importante dans notre société. De tels modèles utilisent une base de données (les *features*, ou *variables explicatives*) pour prédire une valeur (la *cible*, la *target* ou la *variable expliquée*), ou pour répartir les données en sous-ensemble similaires.

Notre objectif, dans la suite de ce rapport, est la prédiction du prix de chaque bien immobilier situé en France. Le prix étant une variable continue, nous nous intéressons aux modèles de machine-learning dits *de régression*, dont l'objectif est justement de prédire une variable continue.

De tels modèles sont utilisés suivant deux étapes distinctes : La première, appelée *entraînement*, ou *training*, permet au modèle d'apprendre comment, à partir du vecteur des *features*, prédire la valeur *target* la plus proche de la valeur réelle. Cette phase nécessite qu'on fournisse au modèle un jeu de données contenant des *targets* observées et réelles. Puis la seconde, appelée *prédiction*, permet au modèle de prédire une valeur *target* en fonction d'un vecteur des *features* qu'on lui fournit.

Ainsi, les modèles de machine-learning dont nous discuterons ici seront des modèles de régressions, qui apprendront, à partir d'un vecteur contenant l'ensemble des informations caractérisant un bien immobilier (taille, localisation, étages, ...) appelé vecteur des *features*, à prédire une valeur continue (le prix du bien) la plus proche de la valeur réelle si elle était observée. Dans un premier temps, il faudra lui fournir un jeu de données contenant des biens dont nous connaissons déjà le prix, pour qu'il apprenne à prédire une estimation correcte, puis nous pourrons lui faire prédire les prix de tous les logements français, simplement lui donnant les caractéristiques de ceux-ci.

1.2 Travaux précédents

Dans leurs précédent travaux ([1] et [2]), Mathias André et Olivier Meslin se sont intéressés à la répartition du patrimoine immobilier des français. En travaillant sur des donnes exhaustives, ils ont pu arriver à des conclusions purement descriptives, comme par exemple le fait qu'un quart des ménages sont multipropriétaires en France, que ces ménages détiennent les deux tiers du parc de logement, ou encore que les ménages qui ont au moins cinq logements représentent 3.5% des ménages mais détiennent 50% des appartements en locations détenus par des particuliers.

Une seconde étape naturelle à ces travaux est d'estimer le prix des logements du parc immobilier français, afin de faire des études plus précises sur le patrimoines des ménages. Les conclusions tirées ne seraient donc plus en termes de *nombre de logements*, mais en *patrimoine immobilier*. C'est dans ce contexte que des modèles de machine learning ont toute leur place.

1.3 Enjeux pour l'Insee

Un tel travail présente plusieurs intérêts pour l'Insee. Il y a tout d'abord un enjeu méthodologique : l'utilisation et la maîtrises de modèles de machine learning complexes seront des enjeux méthodologiques majeurs pour l'Insee dans les années à venir, la majorité des cadres de cet institut ayant reçu une formation principalement économique.

Le second enjeu majeur pour l'Insee est l'utilisation de base administratives et exhaustives pour étudier économiquement et socialement un aspect de la société française. En effet, la principale méthode actuelle pour estimer et étudier le patrimoine immobilier français consiste à sonder un petit échantillon de ménages [1], à qui l'enquêtrice ou l'enquêteur de l'Insee demande une estimation financière rapide de leurs biens. Cette méthode présente le double inconvénient d'être incomplet (car on ne peut

1. Dans le cas de l'étude associée, l'échantillon comporte quelques milliers de ménages.

sonder que quelques milliers de ménages tout au plus), et d'être biaisé (car il est difficile d'estimer précisément le prix de son ou ses logement(s)).

1.4 État de l'art

L'utilisation de modèles de machine learning pour estimer des prix de maisons ou d'appartement à une telle échelle est une chose étonnamment rare. Il apparaît en effet que la seule étude similaire d'estimation de la valeur des logements sur données massives est Norvégienne ([3]). Dans cet article, les auteurs ont un double objectif : Estimer le prix des logements en Norvège entre 1993 et 2015, et comparer un modèle hédonique de régression linéaire et divers modèles de machine-learning. Nous nous sommes inspiré de leur méthode pour notre étude, à savoir entraîner des modèles de machine-learning pour prédire le prix des logements.

Concernant l'étude du patrimoine immobilier des français, le travail le plus aboutit est sans doute l'enquête *Histoire de Vie et Patrimoine* (*HVP*) [4] . Cette enquête, réalisée par l'Insee tous les trois ans, présente une partie sur le patrimoine immobilier détenu par les français. Elle sert de référence pour évaluer la qualité de nos résultats, mais il faut néanmoins noter quelques différences avec le travail effectué et présenté ici : Les définitions juridiques de la propriété ne sont pas exactement les mêmes que celles utilisées lors de la construction de notre base de données, et les répondants à *HVP* peuvent, en toute bonne foi, commettre des erreurs en répondant aux question de l'enquêteur : Il a ainsi été remarqué en comparant aux données administratives que les multipropriétaires pouvaient sous-estimer le nombre de logements dont ils sont propriétaires.

1.5 Objectifs et cahier des charges

Nos objectifs sont la mise au point d'une méthode de valorisation au prix de marché de l'ensemble des logements en France métropolitaine, pour ensuite, en tirer des analyses économiques sur le patrimoine immobilier des Français. Cela passe par l'utilisation de modèles de machine-learning, car ce sont eux qui vont établir la valorisation proprement dite des biens immobiliers. Mais dans la mesure où notre second objectif est l'étude économique et sociale de la répartition du patrimoine immobilier, il faut que nous définissions un un cahier des charges pour notre modèle définitif, qui nous permette de juger de la qualité de ses prédictions au regard de la cohérence économique souhaitée.

Nous souhaitons principalement que nos modèles :

— limitent le biais par rapport aux variables explicatives (les *features*), mais aussi par rapport à d'autres variables d'intérêt qui ne seront pas des variables explicatives (niveau de vie du propriétaire notamment) ;

— et soient les plus précis possible dans la prédiction *au bien près*.

Ceci implique notamment que ces derniers soient centrés géographiquement, c'est-à-dire sans biais local, et qu'ils reflètent l'hétérogénéité du marché local à chaque endroit du territoire français.

2 Récapitulatif des étapes d'une étude sur le patrimoine immobilier

Une formalisation du fonctionnement des modèles de machine-learning qui seront utilisés est proposée en annexe 10.

Les travaux précédents de Mathias André et Olivier Meslin ont permis de produire une sous-base de données comportant les appartements et les maisons qui ont été vendus en France entre 2015 et 2019, appelée table des *mutations*. Il est donc est maintenant possible d'entraîner un modèle de*machine-learning* sur cette table des *mutations*, puis de lui faire prédire une estimation du prix des maisons et des appartements qui n'ont pas été vendus sur cette période, et donc sur lesquels il est nécessaire d'estimer un prix. Ces biens se trouvent dans une table appelée la table des *descriptions*.

Cette figure récapitule les différentes étapes de notre processus. Elle peut également s'appliquer à une majorité de problèmes d'analyse et de prédiction liés à du *machine-learning*.

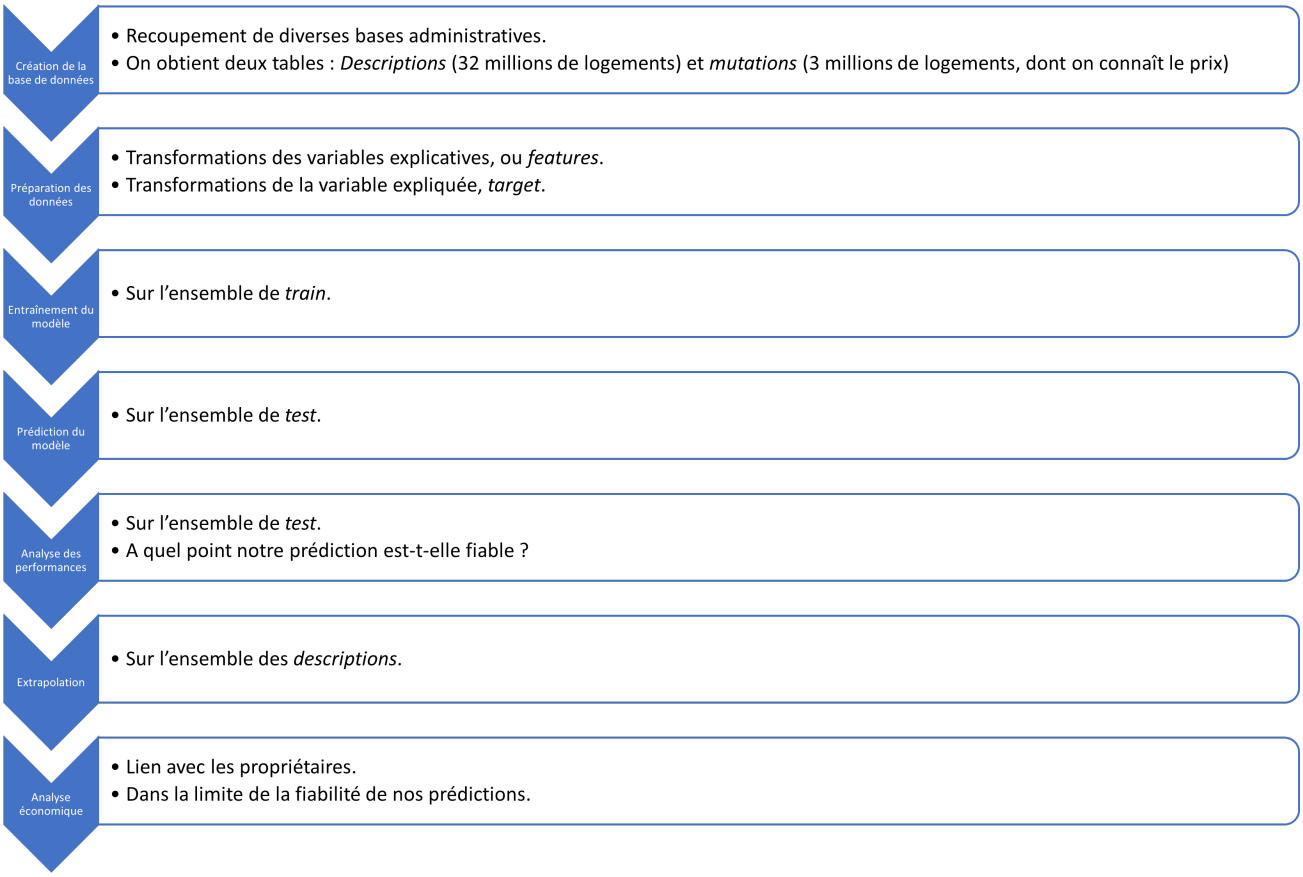

FIGURE 2 – Schéma récapitulatif de l'étude.

On peut notamment noter que la phase liée au modèle proprement dite semble finalement assez réduite, elle n'est qu'une étape parmi la dizaine nécessaires à une telle étude. C'est notamment ce constat qui pousse à penser (voir plus bas 7.3) que l'amélioration des performances obtenues passe avant tout par un travail sur les données.

3 Étapes de création, transformation et transformation des données

Les premières étapes d'une étude statistique et de machine-learning sont liées aux transformations préalables des données qui seront mises en entrée du modèle. Ces étapes sont primordiales pour permettre des résultats robustes et de bonne qualité.

Des détails concernant la création des tables des *descriptions* et des *mutations* sont disponibles en annexes 12.

3.1 Bases utilisables

Les deux bases, complètement parallèles, dont nous disposons pour évaluer le patrimoine immobilier français, sont :

— La table des *mutations* (3 millions de logements) : On connaît le prix de ces appartements et maisons, puisqu'ils ont été vendus au moins une fois entre 2015 et 2019. C'est cette table qui servira à entraîner le modèle.

— La table des *descriptions* (32 millions de logements) : On ne connaît pas le prix de ces appartements et maisons, et on cherche à l'estimer pour ensuite faire des études économiques sur cette prédiction.

3.2 Préparation des données

Malgré le travail de préparation en amont considérable, des manipulations sur les données sont encore nécessaires afin de leur permettre d'être située en entrée d'un modèle de machine-learning.

3.2.1 Sortie du modèle

Les différents modèles ne vont pas *stricto-sensus* prédire le prix d'un logement, mais une transformée de celui-ci. En effet, la variance étant trop importante, et la distribution des prix dans la base des *mutations* étant proche plus proche d'une loi log-normale que d'une loi normale, il était nécessaire de faire une transformation préalable du prix.

Ainsi, la variable prédite par nos modèles est l'écart à la moyenne locale du log du prix au mètre carré d'un logement. C'est-à-dire pour un logement au prix P et à la surface S :

$$y := \log\left(\frac{P}{S}\right) - \overline{\log\left(\frac{P}{S}\right)}^{local}.$$

3.2.2 Préparation des variables d'entrées du modèle

Préparation générale La préparation des variables explicatives qui seront utilisées en entrée des modèles hédoniques et de *machine-learning* est longue, mais en soit assez classique :

1. Ajout d'indicateurs locaux (prix moyen local, etc...).
2. Suppression des erreurs manifestes [2].
3. Normalisation, mise sous forme de *dummies* des variables catégorielles, suppression des lignes manquantes.

Construction des indicateurs locaux Afin de construire des indicateurs de prix locaux, nous avons construit des indicateurs qui reflètent les prix du marché local. Pour cela nous avions, pour chaque logement, accès aux cinq zones géographiques auxquelles il appartient :

- *Ilots Regroupés pour l'Information Statistique* (*IRIS*) : Maille élémentaire d'environ 2 000 habitants. Les communes d'au moins 10 000 habitants et une forte proportion des communes de 5 000 à 10 000 habitants sont découpées en IRIS. Ce découpage constitue une partition de leur territoire. La France compte environ 16 100 IRIS dont 650 dans les DOM.
- Commune.
- *Etablissements publics de coopération intercommunale* (*EPCI*) : Ce sont des regroupements de communes ayant pour objet l'élaboration de projets communs de développement au sein de périmètres de solidarité.
- Département.
- Région.

Pour associer à chaque bien un indicateur de prix local, le principe que nous avons utilisé est le suivant : On calcule la moyenne des prix locaux (par mètre carré) sur la zone (parmi les zones administratives ci-dessus) la plus fine qui regroupe assez de transactions pour être robuste.

Ceci est possible car les cinq zones ci-dessus forment chacune une partition de l'espace français, et sont emboîtées les unes dans les autres. Ainsi, si on se donne un nombre minimal de transactions n_t (6 dans notre cas après une recherche du paramètre optimal), on associe à chaque bien la plus petite échelle parmi les cinq qui comprenant au moins n_t transactions. Une fois cette

2. Appartements vendus à 1000 € dans Paris, maisons de 200 000m^2 à Marseille, etc...

zone locale assignée à un bien, on peut ensuite calculer divers indicateurs pour chacun des biens, tels que le prix moyen par mètre carré aux alentours par exemple.

4 Différents modèles

Après divers essais, nous avons fait le choix de nous concentrer sur trois modèles pour prédire y sur la base de *test* à partir d'un entraînement sur le *train*.

Moyenne locale Le premier modèle, prédit toujours $y = 0$. Dans la mesure où on modélisait $y = \log\left(\frac{P}{S}\right) - \overline{\log\left(\frac{P}{S}\right)}^{local}$, cela signifie que nous prédisons à chaque appartement ou maison son prix moyen local au mètre carré. En multipliant par la surface du bien, nous obtenons une estimation de son prix de vente.

Régression linéaire Le second modèle est une régression linéaire qui prédit y. Puis la transformation inverse à celle visant à obtenir y est réalisée pour obtenir une estimation du prix de chaque logement.

XGBoost Enfin, le dernier modèle étudié est le modèle *XGBoost* pour prédire y. Là encore, les transformations inverses sont ensuite appliquées pour obtenir une estimation du prix de chaque logement.

5 Performances des modèles

Une fois ces trois modèles entraînés, il est possible de leur faire prédire des prix de biens appartenant à l'ensemble de *test*. Ainsi, une première étude méthodologique permet de mettre en lumière leurs différentes performances.

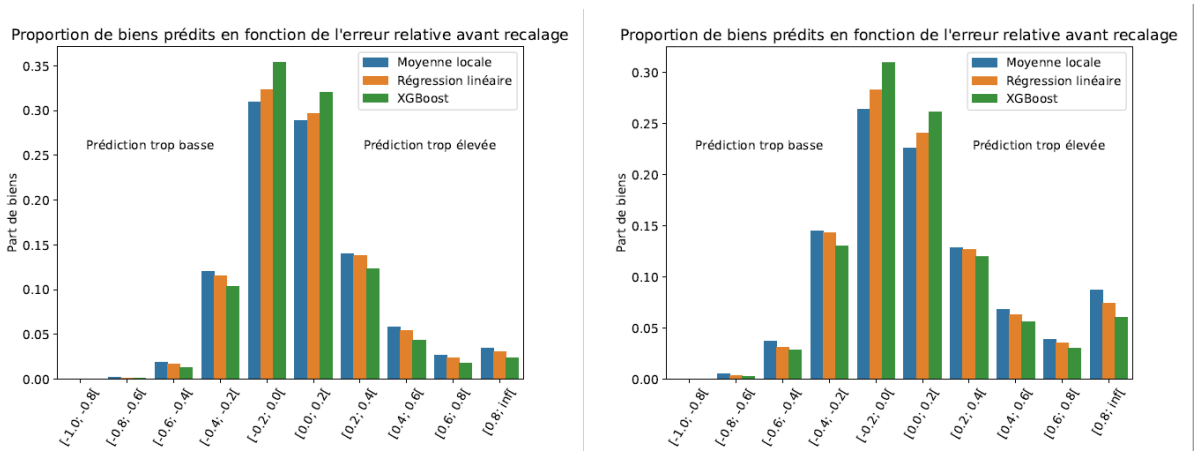

FIGURE 3 – Taux d'erreurs relatives sur l'ensemble de test. sur les appartements (gauche) et sur les maisons (droite). *Lecture : sur l'ensemble des appartements situés dans le jeu de test, le modèle XGBoost prédit, sur 13% des biens, un prix de transaction plus élevé d'entre 20% et 40% que celui observé.*

Il apparaît tout d'abord une hiérarchie stricte en terme de performance entre nos trois modèles : Le modèle de la moyenne locale présente des résultats satisfaisants, mais qui sont inférieurs à ceux de la régression linéaire, qui sont eux-mêmes inférieurs à ceux produits par *XGBoost*.

Ensuite, on peut noter une performance nettement meilleure des modèles sur les appartements. Ceci semble s'expliquer par un marché immobilier plus homogène sur les appartements, et en particulier leurs locations : La variance de la variable de sortie

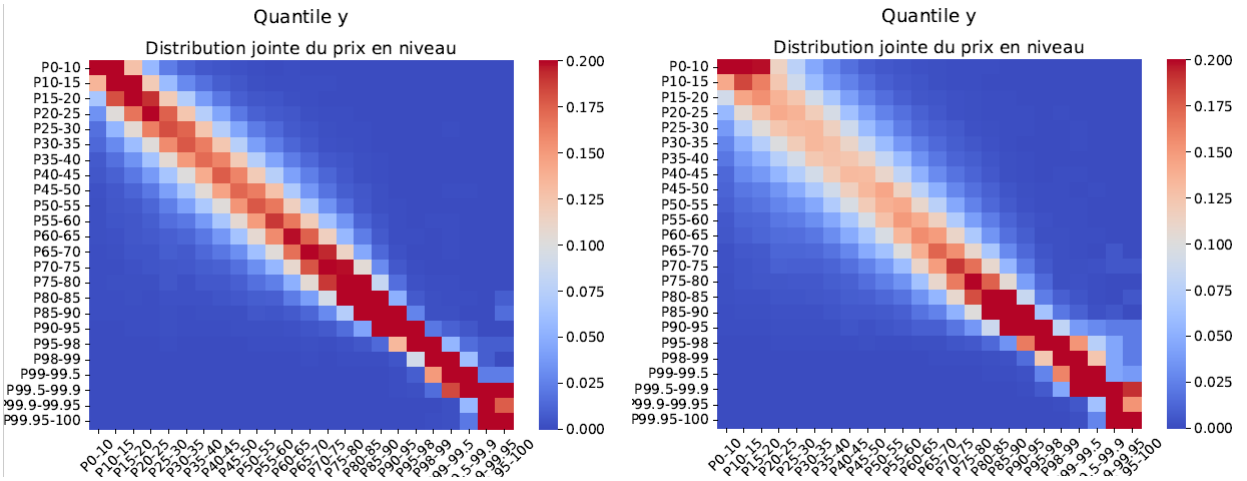

FIGURE 4 – Distributions jointes de probabilité sur les appartements (gauche) et sur les maisons (droite). Pour plus de lisibilité, tous les taux supérieurs à 0.2 = 20% ont été ramenés visuellement à exactement 0.2 = 20%. *Lecture : 18% des appartements situés dans le quantième 45-50 de la distribution des prix observés des appartements se situent dans le quantième P40-45 de la distribution des prix prédits par le modèle XGBoost des appartements.*

$y = \log\left(\frac{P}{S}\right) - \overline{\log\left(\frac{P}{S}\right)}^{local}$ est nettement plus faible pour les appartements que pour les maisons, et il semble que le nombre de variables explicatives réellement corrélées au prix du bien soit plus réduit pour les appartements que pour les maisons.

6 Analyse économique des résultats obtenus

Une fois les performances des modèles analysées, nous pouvons nous intéresser aux questions économiques qui découlent de notre étude.

Pour cela, la première étape est la prédiction, par nos modèles, de l'ensemble des prix des maisons et appartements en France au premier janvier 2017, puis l'appariement de ces estimations de prix aux ménages propriétaires. Ceci permet d'obtenir, par ménage Français, une estimation du patrimoine immobilier détenu.

Sauf mention contraire, toute l'étude économique est faite sur les prédictions du modèle le plus performant *XGBoost*.

6.1 Étude du marché de l'immobilier Français

Afin d'étudier économiquement le marché de l'immobilier, on peut par exemple s'intéresser à l'évolution du prix moyen d'un logement avec le niveau de vie du ménage propriétaire [3]

Tout d'abord, l'importante régularité des courbes obtenus est cohérente économiquement : A niveau de vie proche, le prix moyen d'un logement est proche. Ensuite, le prix moyen d'un logement augmente avec le niveau de vie. Mais d'autres tracés permettent en réalité de mettre en avant deux phénomènes distincts :

— La croissance du prix moyen d'une *maison* avec le niveau de vie est avant tout due à la croissance des *surface* des maisons avec le niveau de vie.

— La croissance du prix moyen d'un *appartement* avec le niveau de vie est avant tout due à la croissance du *prix au mètre carré* des appartements avec le niveau de vie. Dans la mesure où la précédente étude réalisée ([1] et [2]) avait mis en avant

3. Définition du niveau de vie selon l'Insee : Le niveau de vie est égal au revenu disponible du ménage divisé par le nombre d'unités de consommation (UC). Le niveau de vie est donc le même pour tous les individus d'un même ménage.

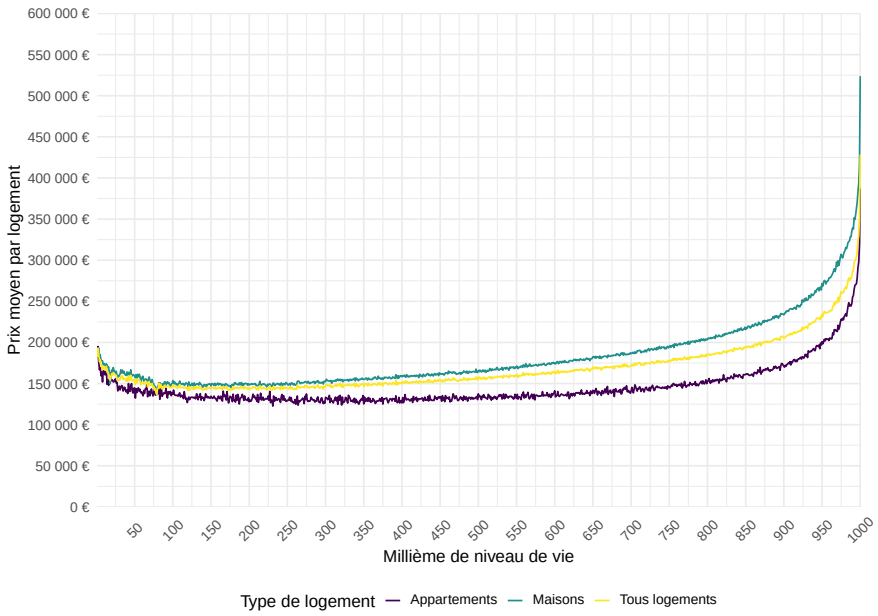

FIGURE 5 – Prix moyen d'un logement en fonction du niveau de vie du ménage propriétaire. *Lecture : Les maisons détenues par les ménages situés dans le* 650^{eme} *millième de niveau de vie coûtent en moyenne 180 000 €.*

la possession, par les multipropriétaires [4] aisés, d'appartements de taille moyenne mais situés dans les centre-villes, on peut logiquement conclure que ces multipropriétaires aisés possèdent des appartements chers du fait de leur localisation.

6.2 Étude du patrimoine immobilier

Afin d'étudier le patrimoine immobilier, on ne s'intéresse plus aux moyennes des caractéristiques des logements (prix, prix au mètre carré, surface, ...), mais on somme les prix estimés des logements détenus par un ménage. Ceci permet d'obtenir une estimation du patrimoine immobilier du ménage.

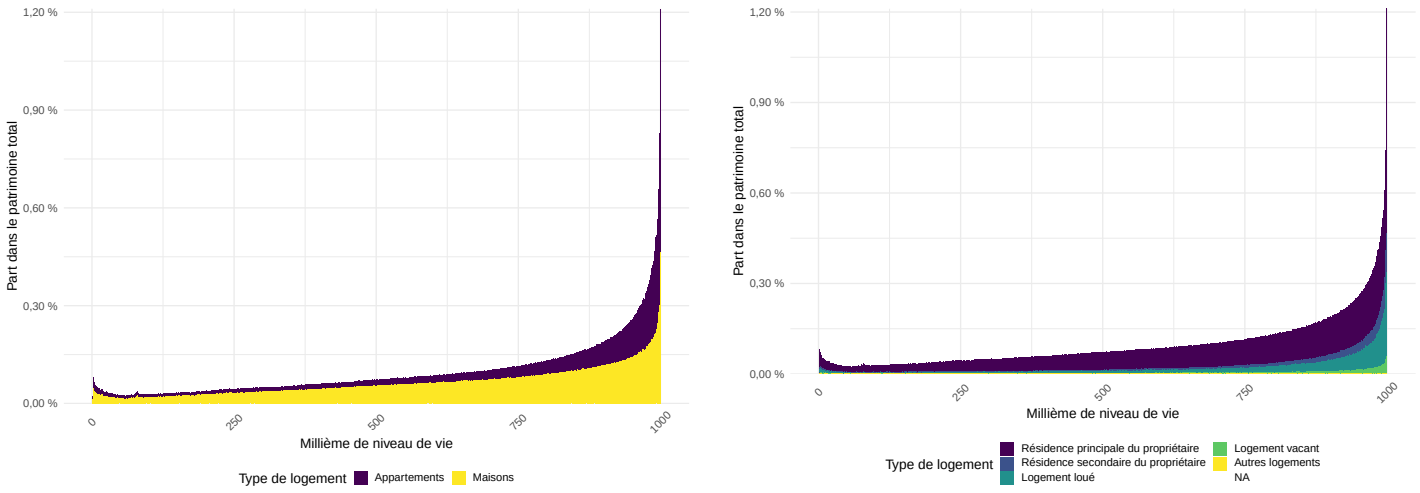

FIGURE 6 – Part du patrimoine immobilier détenu dans le patrimoine immobilier total, en fonction du niveau de vie du ménage propriétaire. *Lecture : Les logements mis en location et possédés par les ménages appartement au* 900^{eme} *millième de niveau de vie représentent* 0.05% *de la valeur du patrimoine immobilier total.*

En s'intéressant au patrimoine immobilier en fonction du niveau de vie du ménage propriétaire, on peut tout d'abord noter

4. Ménages possédant plusieurs logements.

que le patrimoine immobilier est plus concentré pour les appartements que pour les maisons (gauche). En liant cette observation aux conclusions de ([1] et [2]) sur les multipropriétaires qui possèdent des logements dans les centre-villes, mais qui sont mis en location, on peut conjecturer que cette plus grande concentration du patrimoine immobilier des appartement est principalement due aux logements en location. Cette conjecture peut être validée en décomposant la concentration du patrimoine par usage qui est fait du logement (droite).

Enfin, on peut, comparer la concentration du patrimoine immobilier, à savoir, la fonction de répartition de celui-ci, en fonction du modèle développé.

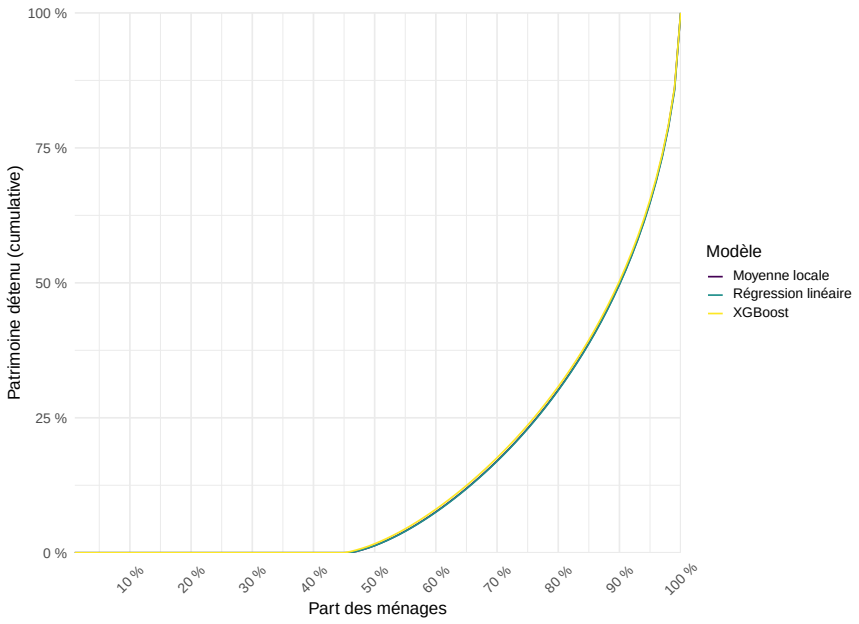

FIGURE 7 – Répartition du patrimoine immobilier (fonction de répartition), en fonction du modèle développé.

Ce graphique permet d'observer la relative proximité de nos trois modèles. Ainsi, les graphiques similaires aux précédents, tracés avec la prédiction de la régression linéaire et du prix local moyen, sont relativement similaires.

7 Perspectives d'améliorations

Il nous semble que la première amélioration possible concerne les prédictions de prix sur les maisons peu chères ; en effet, c'est ce type de biens qui présente l'erreur relative la plus importante. Et de manière générale, nos modèles semblent avoir des difficultés sur les biens extrêmes : ils prédisent un prix trop bas aux biens très chers, mais des prix trop hauts aux biens peu chères. Ainsi, c'est sans doute sur la *variance* de la *target* qu'il nous faut travailler. Un tableau illustrant ce constat est présent en annexes 13.

7.1 Ajout d'un recalage des prix après prédiction

La première piste explorée consiste à augmenter manuellement la variance de la distribution de sortie des modèles, on peut ensuite juger de la qualité de ce recalage de prix sur l'ensemble de test.

En travaillant sur la distribution de sortie du modèle, on aimerait donc que chaque quantième de la distribution prédite soit rapproché de celui de la distribution observée. Cela permettrait au bien moyen prédit situé à une position donnée dans la distribution prédite d'être égal au prix moyen observé pour les biens situés à la même position dans la distribution observés.

Afin de bien séparer les ensemble de *train* et de *test*, la méthode appliquée est la suivante :

— *Train* : On divise la distribution des prix observés et celle des prix prédits en Q quantièmes.

— *Train* : Pour chaque quantième, on définit le ratio r_q le ratio des moyennes des prix dans le quantième q.

— *Test* : On divise la distribution des prix prédits en Q sous-ensemble suivant les valeurs des quantième du *train*.

— *Test* : On applique à chaque bien dans le sous-ensemble q le ratio r_q.

Les quantiles choisis ont été : 0 - 10 - 15 - 20 - ... - 90 - 95 - 98 - 99 - 99.5 - 99.9 - 99.95 - 100

Par souci de simplification, nous ne prenons ici pas en compte d'éventuelles transformations post-prédiction (log, division par surface, soustraction de la moyenne locale), et nous séparons strictement les prix observés des prix prédits dans la notation. Une formalisation est proposée en annexes 14.1.

Les résultats issus d'un tel recalage semblent plutôt encourageants. En effet, si, sur l'ensemble de *test*, nous nous intéressons à l'erreur relative obtenue avec, et sans recalage, en fonction du niveau de vie du ménage propriétaire, nous obtenons les graphiques suivants :

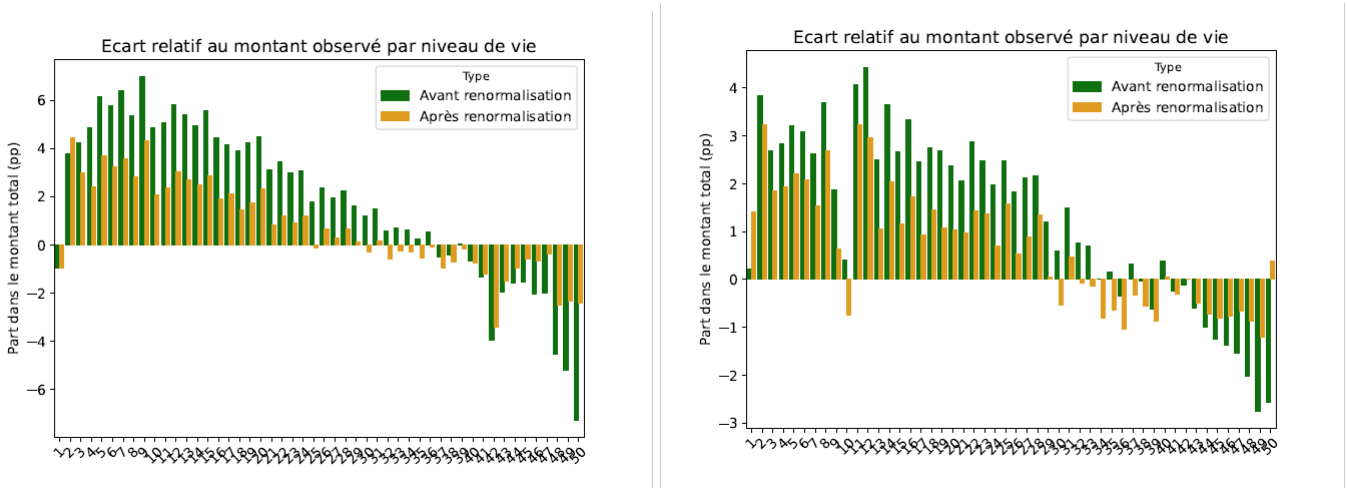

FIGURE 8 – A cinquantième q fixé : Erreur absolue entre la part dans le patrimoine total (observé) détenu par les propriétaires appartenant au cinquantième q observé, et la la part dans le patrimoine total (prédit) détenu par les propriétaires appartenant au cinquantième q prédit. *Lecture : Si les propriétaires situés dans le* 47^{eme} *cinquantième de niveau de vie possèdent* $x\%$ *du patrimoine total des appartements, alors notre modèle XGBoost prévoit que les propriétaires situés dans le* 47^{eme} *cinquantième de niveau de vie possèdent* $x-2\%$ *du patrimoine total des appartements avant renormalisation, et* $x-0.3\%$ *du patrimoine total des appartements après renormalisation.*

Il semble donc bien que le recalage réduise, parfois fortement, l'erreur relative obtenue sur des biens par notre modèle *XGBoost*.

7.2 Utilisation des méta-modèles

Une seconde piste d'amélioration de nos résultats consiste à l'utilisation de modèles d'agrégations. Sur le principe d'une forêt aléatoire, qui utilise une multitude d'arbres de qualité médiocre pour parvenir à un résultat de bonne qualité, nous pouvons utiliser une multitude de modèles *XGBoost*, chacun très performant sur un type de bien en particulier, mais médiocre sur les autres, pour parvenir à un résultat final satisfaisant.

Il est possible de configurer *XGBoost* pour l'obliger à concentrer son apprentissage sur tel ou tel type de bien. Nous avons donc pu entraîner une dizaine de modèles *XGBoost* spécialisés (appelés *weak-learners*), pour ensuite utiliser un dernier modèle *XGBoost* (appelé *strong-learner*, ou *méta-modèle*) chargé de sélectionner, pour chaque bien, le *weak-learner* le plus adapté.

Pour formaliser cela on reprend les notations vectorielles : On se donne un vecteur de *features* $\mathbb{X} = \{X_1, \cdots, X_n\}$, associé à un vecteur de prix observés $\mathbb{Y} = (Y_1, \cdots, Y_n)$.

On entraîne des *weak-learners*[5] qui prédisent après entraînement des prix $\hat{Y}_i^m = f^m(X_i)$. Notre objectif est ensuite de construire un *méta-modèle* (ou *strong-learner*) qui fait une prédiction de prix de la forme $\hat{Y}_i^{meta} = g\left((f^m(X_i))_m, ^*\right)$. Pour cela, divers choix sont possibles pour les arguments * : Ne prendre aucun argument supplémentaire, ou alors tout ou partie des features X_i, éventuellement transformées.

L'intérêt d'utiliser le modèle *XGBoost* est la possibilité d'ajouter facilement des poids pour forcer un meilleur apprentissage sur des biens ayant certaines caractéristiques. Nous avons donc principalement essayé de prendre des poids liés à la position des biens au sein de la distribution prédite. En effet, comme l'indiquent les graphiques des distributions jointes (), notre modèle actuel *XGBoost*, s'il ne prédit pas toujours correctement les prix des maisons, parvient néanmoins à positionner les biens *dans l'ordre* des prix croissants. Intuitivement, travailler de cette façon pourrait permettre de faciliter la prédiction finale du *strong-learner*. On peut ensuite utiliser divers modèles pour *strong-learners* : un modèle associant un bien, en fonction de ses caractéristiques, au *weak-learner* le plus adapté, ou alors un modèle utilisant les caractéristiques d'un bien, ainsi que les différentes prédictions des *weak-learners* pour prédire un prix comme moyenne pondéré des prédictions des *weak-learners* par exemple.

Les poids choisis pour l'entraînement des *weak-learners* ont été de la forme $\exp\left(-\beta(q - q_m)^2\right)$, avec q la position dans la distribution du bien, et q_m une position associée à un sous-modèle donné. Ainsi, chaque *weak-learner* est un modèle *XGBoost* qui se concentre, lors de l'apprentissage sur des biens situés aux alentours de tel ou tel quantile de prix prédit.

Malheureusement, de tels modèles ne semblent pas apporter des résultats bien plus satisfaisants qu'un simple modèle *XGBoost* comme présenté plus haut. Nous pensons que le problème des *méta-modèles* vient de la position de la prédiction des *weak-learners*, utilisé ensuite par le *strong-learner* pour positionner un bien : les logements bien positionnés sont très bien prédits par le *strong-learner*, et ont des performances améliorés par rapport à un simple modèle *XGBoost*, mais les quelques bien mals positionnés pénalisent fortement la prédiction et dégradent significativement les performances globalement du modèle.

7.3 Travail sur les données

Si de telles améliorations sur le modèle de prédiction sont, en soit, intéressantes, et améliorent potentiellement nos résultats, il nous apparaît comme primordial de travailler d'abord et avant tout sur les données, et notamment sur notre indicateurs de prix local moyen au mètre carré décrit en 3.2.2.

En effet, les maisons mal prédites sont majoritairement dans des zones urbaines peu denses. De plus, au sein d'une zone urbaine, le modèle semble avoir des difficultés à estimer des prix corrects principalement dans les communes de la couronne d'une grande ville.

Ainsi, c'est dans les zones où l'hétérogénéité est importante, et où la densité est faible (et donc que notre indicateur de prix moyen se calcule sur des zones grandes) que les résultats sont les plus médiocres.

Nous pensons que dans ce genre ce cas, à un bien dans une zone peu dense proche d'une ville (donc de prix plutôt bas), l'indicateur de prix moyen au mètre carré donne un prix moyen de référence qui inclus les prix des biens situés dans la ville,

5. Ici nous nous sommes concentrés sur *XGBoost*, mais d'autres modèles sont utilisables.

qui sont donc plutôt hauts. Ainsi, l'indicateur est biaisé à la hausse.

8 Conclusion

8.1 Comparaison à l'enquête Histoire de Vie et Patrimoine (*HVP*)

La dernière étape du travail réalisé consiste à une comparaison à la littérature existante la plus proche, à savoir l'enquête *HVP* 1.4. Dans cette étude, une courbe de répartition du patrimoine similaire à celle tracée dans 7 est proposée. C'est cette courbe que nous avons choisie pour comparer nos résultats.

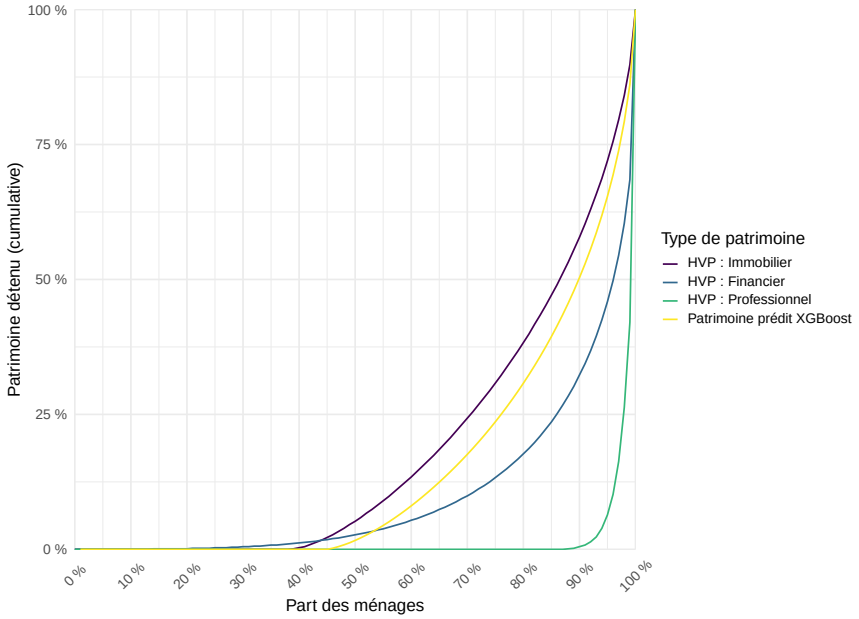

FIGURE 9 – Répartition du patrimoine immobilier (fonction de répartition), en fonction du type de patrimoine.

Il apparaît ainsi que la concentration importante du patrimoine immobilier à laquelle nous aboutissons est relativement proche de celle issue de l'enquête *HVP* ; le patrimoine immobilier est notamment, dans les deux cas, moins concentré que les patrimoines financier et professionnel.

Nous aboutissons néanmoins avec nos travaux à une plus grande concentration du patrimoine immobilier que dans l'enquête *HVP*. Arriver à une concentration plus importante est cohérent, en raison des limites intrinsèques à l'enquête *HVP* présentée dans 1.4.

8.2 Bilan méthodologique

Nous avons pour l'instant pu construire un traitement des données comprenant un modèle de *machine-learning* permettant d'estimer des prix de biens immobiliers situés en France métropolitaine. Les résultats sont utilisables en l'état pour des conclusions macro-économiques, mais sont à affiner pour des conclusions de micro-économie fines. La prédiction du prix *au bien près* est notamment à améliorer.

Nous avons également pu conclure à une stricte hiérarchie en termes de qualité des modèles : un simple modèle de moyenne locale est déjà tout à fait satisfaisant, car la surface et la géographie expliquent déjà 53% (maisons) et 85% (appartements) de la variance du prix, mais un modèle de régression linéaire fait mieux, et un modèle *XGBoost* fait encore mieux.

8.3 Bilan économique

Économiquement, il apparaît que les gros multipropriétaires ont des appartements en location de taille moyenne ($80m^2$) et situés en centre-ville, et donc chers. Notre répartition du patrimoine immobilier est plus inégale que l'*Enquête patrimoine*, mais va dans le même sens ; la différence de concentration peut sans doute s'expliquer par des erreurs de nos modèles, mais aussi par des biais intrinsèques à l'enquête *HVP* que nous pensons avoir cernés.

Enfin, la concentration du patrimoine est fortement marquée(le top 10% des ménages propriétaires possédant près de la moitié du patrimoine immobilier total), et présente des différences entre les maisons et les appartements, notamment du fait de la mise en location des appartements en centre-villes.

8.4 En perspective : Tout d'abord un travail sur les données

La continuation du travail entamé, notamment en vue d'une thèse, passe selon nous tout d'abord par un travail sur les données. Les pistes explorées passant par une amélioration plus ou moins directe des modèles de machine-learning ont en effet apporté des résultats mitigés. Il apparaît parallèlement que certains types de biens, telles que les maisons peu chères situées dans des zones peu denses en périphérie de villes moyennes, nécessitent sans doute un travail spécifique. Nous pensons donc qu'un travail sur l'indicateur des prix moyens au mètre carré dans la zone est nécessaire.

Ensuite, une amélioration en amont de la qualité des données pourrait également apporter quelques pistes intéressantes : Nettoyage plus fin pour éviter les données erronées, ajout de données liées aux zones peu denses (temps de transports, proximité des services publics, âge de la population locale, qualité de l'environnement, ...), mais aussi de données sur chaque bien (état du bien, nécessité de travaux, diagnostique de performance énergétique, ...).

De manière générale, c'est ce type de travail au plus prêt des données qui nous semble le plus pertinent dans un premier temps, pour améliorer significativement la qualité des résultats présentés.

Deuxième partie

Annexes

Table des matières

9 Liste complète des variables explicatives disponibles

9.1 Caractéristiques du bien

— Type de bien
— Eau, électricité, gaz, ascenseur, escalier de service, chauffage central, vide ordure, tout à l'égout
— Etage
— Période de construction
— Quartier prioritaire de la ville
— HLM
— Pièces, chambres, salle de bain, ...
— Surface agricole, au sol, bois, lac
— Cave, grenier, piscine, garage, autre dépendance
— log(Surface), au carré, au cube

9.2 Situation géographique

— Indicateurs locaux
— Type de zone locale (*IRIS*, commune,...)
— Littoral
— Type de station touristique
— Distance à la plus proche de ville de 50 000, 100 000, 200 000 et 500 000 habitants
— Taille de l'aire d'attraction des villes ($TAAV$)

9.3 Propriétaire(s)

— Niveau de vie
— Revenu disponible
— Part de propriétaires dans l'*IRIS* ayant un niveau de vie dans le top ou dans le bottom 1%, 5%, 10% et 20 %
— Idem pour le revenu disponible
— Nombre de propriétaires du bien

9.4 Transaction

— Année et trimestre de vente

10 Formalisation du fonctionnement d'un modèle de régression

Dans toute cette section, on considère un échantillon de couple (x, y). On note par des exposant $x^j; y^k$ le j^{eme} vecteur x et le k^{eme} vecteur y de celui-ci. Et on note par des indices x_i la i^{eme} coordonnée dans l'espace d'un point x.

10.1 Principe

On se donne un couple de variables aléatoires $z := (x; y) \in \mathbb{R}^d \times \mathbb{R}$, de loi jointe inconnue ρ. L'objectif du modèle est de prédire y en observant uniquement x, et nous avons à dispositions m observations distinctes $(x^i; y^i) \overset{iid}{\sim} \rho$. Dans notre cas, le

principale problème est que la variable de sortie y ne dépend pas de façon déterministe de x, notamment parce qu'il manque des paramètres auxquels nous n'avons pas accès (l'état du bien, les éventuelles nuisances sonores à proximité, l'orientation nord-sud, ...). Autrement dit, il n'existe pas de fonction u telle que $y = u(x)\ \forall x \in \mathbb{R}^d$.

On ne peut donc que chercher à construire une fonction u qui explique au mieux y sachant x. Pour cela, il est nécessaire de se donner une fonction de perte, ou fonction de risque, pour quantifier l'erreur faite par notre modèle u. Par exemple, le risque quadratique moyen d'un modèle est :

$$R : v \mapsto \mathbb{E}(|v(x) - y|^2)$$

Comme nous n'avons accès qu'à m observations, nous n'avons accès qu'à un risque dit *empirique*. Par exemple dans le cas d'un risque quadratique moyen, le risque empirique associé est :

$$\tilde{R}_m : v \mapsto \frac{1}{m}\sum_{i=1}^{m} |v(x^i) - y^i|^2.$$

Un problème de machine learning peut donc se résumer par la recherche de u telle que :

$$u \in \underset{v}{\text{Argmin}} \tilde{R}_m(v)$$

Bien entendu, il existe d'autres fonctions de risques R, et de risques empiriques associés $\tilde{R}$. Mais dans le cas du risque quadratique, on dit que la fonction u solution est *l'estimateur des moindres carrés*.

10.2 Régression linéaire

Le modèle le plus simple en régression est celui associé à une fonction recherche de fonction u linéaire. On parle alors de régression linéaire. On cherche u de la forme :

$$u : x \in \mathbb{R}^d \mapsto \sum_{k=1}^{d} \beta_k x_k$$

Avec x_k et β_k les $k - èmes$ coordonnées de vecteurs x et β.

Ainsi, en reprenant notre échantillon $(x^i; y^i)$, on peut écrire le problème :

$$\forall i \in \{1, \cdots, m\};\ y^i = u(x^i) + \varepsilon^i = \beta_0 + \sum_{k=1}^{d} \beta_k x_k^i + \varepsilon^i$$

On peut ensuite mettre le problème sous forme vectorielle et matricielle :

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon$$

Avec $\mathbb{Y} = (y^1; \cdots; y^m) \in \mathbb{R}^m$, $\mathbb{X} = \begin{pmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_d^1 \\ 1 & x_1^2 & x_2^2 & \cdots & x_d^2 \\ . & . & . & . & . \\ 1 & x_1^m & x_2^m & \cdots & x_d^m \end{pmatrix} \in \mathcal{M}_{m\times d+1}$ et $\beta = (\beta_0; \cdots; \beta^d) \in \mathbb{R}^{d+1}$. Le vecteur β est le vecteur associé à l'application u. Enfin, on dit que le vecteur $\varepsilon = (\varepsilon^1; \cdots; \varepsilon^m)$ est le vecteur de bruit.

L'objectif de l'entraînement d'une régression linéaire est donc de trouver β qui minimise $\varepsilon = \mathbb{Y} - \mathbb{X}\beta$.

Dans le cas où le risque considéré est le risque quadratique, u est l'estimateur des moindres carrés et le vecteur β qui minimise $\tilde{R}_m(u) = \frac{1}{m}\sum_{i=1}^{m} |v(x^i) - y^i|^2 = \frac{1}{m}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2$ est le vecteur :

$$\beta = \left({}^t\mathbb{X}\mathbb{X}\right)^{-1} {}^t\mathbb{X}Y$$

L'existence d'un tel β optimal nécessite d'avoir ${}^t\mathbb{X}\mathbb{X}$ inversible, et donc $m \geq d + 1$, et $\mathbb{X}$ de rang plein.

10.3 Forêts aléatoires et boosting

Un second modèle très utilisé en régression est appelé *XGBoost*, pour *eXtreme Gradient BOOSTing*. Il est considéré généralement comme l'un des modèles les plus performants actuellement.

Une forêt aléatoire est composée de plusieurs sous-modèles appelés des arbres aléatoires. Plus de détails sont trouvables dans [5].

10.3.1 Arbre aléatoire – Présentation

Un arbre aléatoire m_j est un modèle de régression qui prend en entrée un sous-échantillon aléatoire des m observations $((x^i; y^i))_i$ noté $\mathcal{D}^*(X^j)$. Son fonctionnement est basé sur une séparation de l'espace $\mathbb{R}^d$ en une multitude de cellules les plus homogènes possibles. Plus précisément, un arbre alterne les étapes :

— Trouver la cellule avec la variance la plus importante ;

— Déterminer la direction de l'espace $\mathbb{R}^d$ dans cette cellule avec la plus grosse variance ;

— Séparer la cellule suivant cette direction, à l'endroit de l'espace où la variance intra-cellulaire sera la plus faible ;

— On obtient alors deux sous-cellules ;

Une fois l'espace séparé en sous-cellules (qui sont donc les feuilles de l'arbre final), la sortie du modèle en un point x est donnée par :

$$m_j : x \in \mathbb{R}^d \mapsto \frac{1}{N_n(x, x_j)} \sum_{i \in \mathcal{D}_n^*(x^j)} \mathbb{1}_{x_i \in A(x, x_j)} Y^i$$

Avec $A(x, \mathcal{D}_n)$ la feuille (ou cellule) contenant x, et $N_n(x, x_j)$ le nombre de points du sous-échantillon $\mathcal{D}^*(x^j)$ qui appartiennent à la feuille $A(x, x_j)$.

Autrement dit, un arbre aléatoire associe, à un point x, une prédiction qui est la prédiction moyenne sur les points d'un sous-échantillon proche (au sens de la variance locale) de x.

10.3.2 Arbre aléatoire – Formalisation

Généralement, le critère d'optimisation généralement utilisé est appelé la *CART-criterion*.

A chaque étape de l'arbre, on sélectionne aléatoirement selon une loi uniforme $mtry \geq d$ directions de l'espace [6], et on note $\mathcal{M}_{try} \subset \{1; \cdots; d\}$ le sous-ensemble des coordonnées sélectionnées. Le CART-criterion sera optimisé sur ces coordonnées.

On considère ensuite une cellule quelconque A, et note toujours $N_n(A)$ le nombre de points tombant dans A. On appelle découpage dans A une paire (j, z) telle qu'on coupe au niveau z de la j^{th} variable dans la cellule A.

Si on note $\mathcal{C}_A$ l'ensemble des découpages possibles dans A, et que le vecteur $x^i = (x_1^i, \cdots, x_d^i)$ est dans A, alors pour tout $(j, z) \in \mathcal{C}_A$ [7] on pose :

$$L_{reg(j,z)} := \frac{1}{N_n(A)} \sum_{i=1}^{n} (y^i - \bar{y})^2 \mathbb{1}_{x^i \in A} - \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(y^i - \bar{y}_{A_L} \mathbb{1}_{x_j^i < z} - \bar{y}_{A_R} \mathbb{1}_{x_j^i \geq z} \right)^2 \mathbb{1}_{x^i \in A}$$

Avec :

— $A_L := \{x \in A, x^{(j)} < z\}$ la première sous-cellule séparée par z.

— $A_R := \{x \in A, x^{(j)} \geq z\}$ la seconde sous-cellule séparée par z.

— $\bar{y}$ est la moyenne des y^i tels que x^i est dans l'ensemble associé. On prend 0 si l'ensemble est vide.

6. on prend en général $mtr = int(d/3)$

7. Pour s'éviter les problèmes, on coupe toujours au milieu de deux points consécutifs, il n'y a donc qu'un nombre fini de découpages possibles.

Ainsi, le découpage optimal est :

$$(j^*, z^*) :\in \underset{\substack{j \in \mathcal{M}_{try} \\ (j,z) \in \mathcal{C}_A}}{\text{Argmax}} L_{reg}(j, z).$$

10.3.3 Forêt aléatoire

Une fois une multitude d'arbres construits, on peut combiner tout ces sous-modèles pour former un estimateur final appelé forêt aléatoire. Ainsi, la forêt aléatoire comprenant M arbres aura pour fonction :

$$m^M : x \mapsto \frac{1}{M} \sum_{j=1}^{M} m_j(x)$$

Ainsi, le j^{eme} arbre aléatoire est noté par un indice m_j, mais la forêt aléatoire contenant M arbres est notée par un exposant m^M.

Un algorithme complet d'entraînement des forêts aléatoires est présent ci-dessous 11.

10.3.4 Boosting

La dernière étape pour transformer une forêt aléatoire en *XGBoost* est d'entraîner la forêt non pas comme un ensemble d'arbres indépendants les uns des autres, mais en les entraînant successivement ; autrement dit, chaque arbre dépend des performances des arbres entraînés précédemment.

Dans *XGBoost*, un arbre apprend à prédire l'erreur faite par le modèle. Ceci lui permet de se concentrer sur les points les plus difficiles à prédire, car ce sont eux qui ont conduits à des erreurs importantes.

Formellement, la forêt à l'étape j est construite à partir de la même forêt à l'étape précédente, et d'un nouvel arbre de régression : $m^j = m^{j-1} + m_j$.

Ainsi, la prédiction finale du modèle en x n'est pas $m^M(x) = \frac{1}{M} \sum_{j=1}^{M} m_j(x)$ comme dans une simple forêt aléatoire, mais $m^M(x) = m^{M-1}(x) + m_M(x) = \sum_{j=1}^{M} m_j(x)$.

10.4 Overfitting

La principale difficulté vis-à-vis de la recherche d'une telle fonction u est liée à l'ensemble de recherche de u. Il est en effet facile de tomber dans un cas où la fonction u reproduit très bien l'échantillon $(x^i; y^i)$, mais se généralise très peu à un nouveau couple $(x; y) \sim \rho$.

Si on prend l'exemple d'un arbre aléatoire, il peut être tentant d'entraîner un arbre aléatoire n'ayant pas de limite inférieure de taille pour les feuilles, c'est-à-dire sans limite supérieure de profondeur. Mais dans ce cas, l'arbre entraîné découpera l'espace de façon à obtenir un point du sous-échantillon dans chaque cellule. Il sera ainsi incapable de se généraliser à un autre sous-échantillon, ou à un nouveau point x inconnu. C'est ce qu'on appelle le *sur-apprentissage*, ou *overfitting*.

Diverses techniques existent pour s'en protéger, telle que limiter la profondeur des arbres, séparer l'échantillon de travail en plusieurs sous-échantillons pour conserver des points connus qui ne sont pas utilisés à l'entraînement par exemple.

Pour se protéger de l'*overfitting*, on découpe un ensemble en deux sous-ensembles :

— L'ensemble d'entraînement, ou de *train*, qui sert à entraîner un modèle ;
— L'ensemble de validation, ou de *test*, qui sert à quantifier et juger les performances du modèle.

Seul un modèle obtenant de bonnes performances sur le *test* pourra être considéré comme performant et être validé.

11 Algorithme complet de forêt aléatoire

Algorithm 1: Random Forest selon Breinman

Input: $a_m \in \{1, \cdots, m\}$ le nombre d'échantillon qui forment $\mathcal{D}^*$. $metry \in \{1, \cdots, d\}$ le nombre de directions possible pour découper un arbre à chaque étape. $nodesize \in \{1, \cdots, a_m\}$ le nombre de points minimal par cellule (ou feuille), en desous duquel on ne peut pas faire de sous-cellule.

Output: $m^M(x)$ la prédiction de la random forest en x.

Data: $((x^1; y^1); (x^2; y^2); \cdots; (x^m; y^m))$

for $j \in \{1, \cdots, M\}$ **do**
  Former $\mathcal{D}^*$ en sélectionnant a_m points (avec ou sans remise) de manière uniforme.
  $\mathcal{P} := (\mathbb{R}^d)$ la liste contenant les cellules associées à la racine de l'arbre.
  $\mathcal{P}_{final} := \emptyset$ la liste vide.
  while $\mathcal{P} \neq \emptyset$ **do**
    $A :=$ la premier élément de $\mathcal{P}$.
    if $\#A <$ *nodesize ou si tous les* $x_i \in A$ *sont égaux* **then**
      Supprimer A de la liste $\mathcal{P}$.
      $\mathcal{P}_{final} \leftarrow Concatenate(\mathcal{P}_{final}, A)$.
    else
      Selectionner uniformément sans remise un sous-ensemble $\mathcal{M}_{try} \subset \{1, \cdots, p\}$ de cardinal $mtry$.
      Selectionner le meilleur découpage dans A d'après *CART-criterion* selon les coordonnées dans $\mathcal{M}_{try}$.
      A_L, $A_R :=$ les deux sous-cellules.
      Supprimer A de la liste $\mathcal{P}$.
      $\mathcal{P} \leftarrow Concatenate(\mathcal{P}_{final}, A_L, A_R)$.
  Calculer la valeur $m_j(x) = \frac{1}{N_n(x, x_j)} \sum_{i \in \mathcal{D}_n^*(x^j)} \mathbb{1}_{x_i \in A(x, x_j)} Y^i$ en x comme la moyenne des valeurs des Y^i tombant dans la cellule de x dans la partition $\mathcal{P}_{final}$.
Calculer et retourner $m^M(x) = \frac{1}{M} \sum_{j=1}^{M} m_j(x)$ la moyenne des résultats des arbres.

12 Création des tables des *descriptions* et des *mutations*

12.1 Base *Fidelimo*

La base de données utilisée lors de ce stage était principalement la même que celle utilisée pour les travaux de [1]. Elle a été construite en approchant trois sources administratives de 2017 :

— *Fidéli* : Construite par l'Insee. Cette table recense les données fiscales des individus et des logements (état civil, composition des ménages, revenus fiscaux, etc...).

— Fichiers *Majic* : Construite par la direction générale des Finances publiques. Elle décrit l'intégralité des propriétés bâties (maison, appartements, bureaux, garages, etc...) et non bâties (jardin, champs, lacs, etc...). Elle contient les données du cadastre, et l'identité des propriétaires (état civil, adresse, nature du droit de propriété, etc...).

— Le *registre du commerce et des sociétés* (RCS) : Construite par la greffe des tribunaux de commerce contenant les information sur les société (dénomination, forme juridique, adresse du siège, etc...) et les personnes physiques qui en sont représentantes (gérants, actionnaires, associés, etc...). Cette table permet d'identifier les individus propriétaires par l'intermédiaire d'une *Société Civile Immobilière (SCI)*.

Cette table ainsi construite présente une double innovation : Elle rapproche les informations sur les propriétaires (revenus, caractéristiques sociales, etc...) et les logements dont ils sont propriétaires. Enfin, elle intègre le parc des résidences secondaires et principales, mais aussi les biens immobiliers à usage professionnel et locatif. Elle comporte 52.4 millions de biens immobiliers, dont :

— 37.1 millions de logements (maisons ou appartements). Parmi eux, 30.3 millions sont détenus par des particuliers.

— 11.9 millions de dépendances (garages, parkings, ...).

— 3.4 millions de locaux industriels et commerciaux.

Nous nous sommes actuellement concentré uniquement sur les logements possédés par des particuliers résidant en France : 28.4 millions de logements. La base ainsi constituée s'appelle *Fidelimo*.

La base *Fidelimo* contient l'ensemble des logements, soit environ 30 millions de maisons et appartements, situés en France métropolitaine, ainsi que toutes leurs caractéristiques cadastrales, géographiques et liées à leurs propriétaires.

12.2 Ajout des données de transaction

Une dernière base a été ajoutée à *Fidelimo* : La base *Demandes de Valeurs Foncières* (*DVF*) qui comprend les transactions immobilières entre 2015 et 2019.

La base réunissant *Fidelimo* et *DVF* était déjà construite lors de mon arrivée à l'Insee.

13 Nécessité de travail supplémentaire sur les biens extrêmes

Ce tableau présente les erreurs réalisées par nos modèles sur les maisons très chères et très peu chères. La part dans le total des transactions pouvant varier significativement de la part prédite de ces maisons aux prix extrêmes justifie selon nous que des travaux futurs s'intéressent particulièrement à ce types de biens.

Prix des maisons	Fréquence observée	Fréquence prédite (régression linéaire)	Part dans le total des transactions
$>$ 5 000 000 €	0.015 %	0.005 %	0.7 %
$>$ 2 000 000 €	0.12 %	0.06 %	2.1 %
$>$ 1 000 000 €	0.70 %	0.55 %	5.6 %
$<$ 75 000 €	11.7 %	5.7 %	2.9 %
$<$ 50 000 €	4.4 %	0.82 %	0.78 %
$<$ 35 000 €	1.5 %	0.08 %	0.2 %

14 Formalisation des perspectives

14.1 Recalage de la distribution de sortie

Voici une première formalisation du rercalage par quantile de la distribution de sortie du modèle. La méthode est rappelée :

— *Train* : On divise la distribution des prix observés et celle des prix prédits en Q quantièmes.

— *Train* : Pour chaque quantième, on définit le ratio r_q le ratio des moyennes des prix dans le quantième q.

— *Test* : On divise la distribution des prix prédits en Q sous-ensemble suivant les valeurs des quantième du *train*.

— *Test* : On applique à chaque bien dans le sous-ensemble q le ratio r_q.

Les quantiles choisis ont été : 0 - 10 - 15 - 20 - ... - 90 - 95 - 98 - 99 - 99.5 - 99.9 - 99.95 - 100

Pour les prix observés : une feature est notée $x \in \mathbb{X}$, un prix prédit $y \sim Y$, la distribution observée des prix est $\mathbb{P}_Y$, et la fonction de répartition associée est $\mathbb{F}_Y : y \mapsto \mathbb{P}_Y(Y \leq y)$.

Pour les prix prédits : une feature est notée $x \in \mathbb{X}$, un prix prédit $\hat{y} \sim \hat{Y}$, la distribution prédite des prix est $\mathbb{P}_{\hat{Y}}$, et la fonction de répartition associée est $\mathbb{F}_{\hat{Y}} : \hat{y} \mapsto \mathbb{P}_{\hat{Y}}(\hat{Y} \leq \hat{y})$.

Un quantile est défini par $q_a, q_b \in [0, 1]$ tels que $q_a < q_b$. Ainsi, les prix définissant les intervalles sont :

— Observé : $a = \mathbb{F}_Y(q_a)$ et $b = \mathbb{F}_Y(q_b)$

— Prédit : $\hat{a} = \mathbb{F}_{\hat{Y}}(q_a)$ et $\hat{b} = \mathbb{F}_{\hat{Y}}(q_b)$

Ainsi, le recalage exploré est défini par :

$$\alpha(q_a, q_b, Y, \hat{Y}) := \frac{\int_{\mathbb{F}_Y^{-1}(q_a)}^{\mathbb{F}_Y^{-1}(q_b)} y \mathrm{d}\mathbb{P}_Y(y)}{\int_{\mathbb{F}_{\hat{Y}}^{-1}(q_a)}^{\mathbb{F}_{\hat{Y}}^{-1}(q_b)} \hat{y} \mathrm{d}\mathbb{P}_{\hat{Y}}(\hat{y})}$$

Troisième partie

Bibliographie

Références

[1] Mathias André et Olivier Meslin. “Et pour quelques appartements de plus : Étude de la propriété immobilière des ménages et du profil redistributif de la taxe foncière”. In : (2021). URL : https://www.insee.fr/fr/statistiques/5893223.

[2] Mathias André et Olivier Meslin. “24 % des ménages détiennent 68 % des logements possédés par des particuliers”. In : (2021). URL : https://www.insee.fr/fr/statistiques/5432517?sommaire=5435421.

[3] Torstensen Kjersti Naess Fagereng Andreas Holm Martin Blomhoff. “Housing wealth in Norway, 1993–2015”. In : (Janv 2021). URL : https://content.iospress.com/articles/journal-of-economic-and-social-measurement/jem200471.

[4] Insee. “Histoire de vie et Patrimoine, une enquête de l’Insee”. In : (2021). URL : https://www.insee.fr/fr/information/2964509.

[5] Gérard Biau & Erwan Scornet. “A Random Forest Guided Tour”. In : (18 Nov 2015). URL : https://www.researchgate.net/publication/284219299_A_Random_Forest_Guided_Tour.