Prédiction des Prix Immobiliers en Île-de-France : Une Approche par Machine Learning

Stephen Cohen Kien PHAM TRUNG Antoine Rech–Tronville

2024-2025

Table des matières

1 Introduction

Dans un contexte où le marché immobilier francilien connaît des évolutions complexes et multifactorielles, notre projet vise à développer un modèle prédictif des prix immobiliers. Cette étude s'appuie sur des techniques de machine learning pour analyser et prédire le prix d'un appartement, en intégrant à la fois les caractéristiques propres aux biens et des variables macroéconomiques susceptibles d'influencer le prix.

2 Méthodologie et Données

2.1 Sources et Structure des Données

Notre analyse repose sur un jeu de données issu de l'aggrégation de différentes bases (présentées si dessous avec leurs variables) :

- **Données immobilières** :
  - Dates des transactions
  - Typologie des biens
  - Surfaces habitables
  - Prix de vente
  - Codes INSEE
- **Indicateurs macroéconomiques** :
  - Indice de confiance des ménages
  - Taux directeur de la BCE
  - Cours du pétrole Brent

Ces données ont été récupérés sur Cerema, site de l'Etat qui regroupe les transactions immobilères.
Le Volume de données est très important avec plus de 19,6 millions d'observations.

```
Aperçu des données :

  year_month l_codinsee       type_bien  surface       prix_m2
departement  \
4    2015-04      75120  UN APPARTEMENT     61.0   4836.065574
75
5    2015-12      75111  UN APPARTEMENT     50.0   5840.000000
75
6    2024-01      75118  UN APPARTEMENT     22.0  11272.727273
75
7    2024-01      75118  UN APPARTEMENT     22.0  11272.727273
75
8    2024-01      75118  UN APPARTEMENT     22.0  11272.727273
75

   valeurfonc       date  confiance_menages  taux_directeur
cours_petrole
4    295000.0        NaT                NaN             NaN
NaN
5    292000.0        NaT                NaN             NaN
NaN
6    248000.0 2024-01-01          94.446667           3.894
82.49
7    248000.0 2024-01-02          94.573333           3.906
82.49
8    248000.0 2024-01-03          94.700000           3.904
82.49
```

FIGURE 1 – Aperçu des bases de données

2.2 Préparation des Données

3 Analyse Exploratoire

3.1 Statistiques Descriptives

Notre jeu de données comprend plus de 19,6 millions de transactions, caractérisées par :

— Un prix moyen au m² de 4 547€ avec un écart-type de 3 818€, ce qui indique une forte dispersion.
— Une surface moyenne de 619m² (écart-type : 14 011m²)
— Une période d'observation s'étendant de 2014 à 2024 donnant une bonne perspective historique

On peut déjà observer dans un premier temps qu'il existe une forte dispersion au niveau du prix moyen ainsi que des surfaces vendues. La distribution des prix est asymétrique, il y a beaucoup de biens à prix modéré et quelques biens très chers tirant la moyenne vers le haut. Les fortes de variations de prix et de distribution sont probablement dû aux différences entre départements et aux types de bien étudiés (pré ou appartement

```
Statistiques descriptives :

            surface       prix_m2    valeurfonc  \
count  1.966462e+07  1.966462e+07  1.966462e+07
mean   6.187738e+02  4.547486e+03  4.318239e+05
min    1.000000e+00  8.400000e-01  1.000000e+00
25%    4.500000e+01  1.250000e+03  1.800000e+05
50%    7.200000e+01  3.642857e+03  2.800000e+05
75%    2.580000e+02  6.891304e+03  4.400000e+05
max    6.098051e+06  1.769964e+04  3.411290e+08
std    1.401090e+04  3.817738e+03  1.435034e+06

                                date  confiance_menages
taux_directeur  \
count                       18854925       1.885492e+07
```

FIGURE 2 – Stats descriptives des prix au m²

3.2 Analyse des Corrélations

Matrice de corrélation :

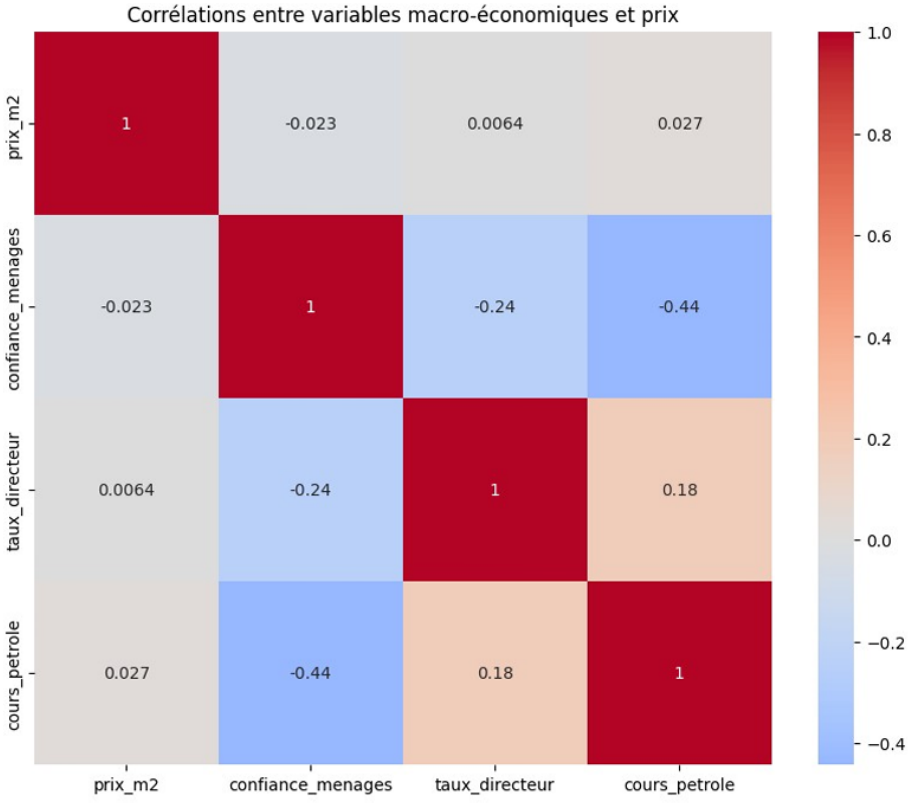

FIGURE 3 – Matrice de corrélation des variables

Les trois variables macroéconomiques que nous avons décidé de rajouter à notre jeu de données semblent a priori très peu corrélées avec le prix au mètre carré. Nous allons vérifier cela avec l'importance que les modèles prédictifs leur donnent. Cependant, cela nous semble important car pour prédire l'avenir, il faut tenir compte des tendances macroéconomiques.

3.3 Analyse graphique

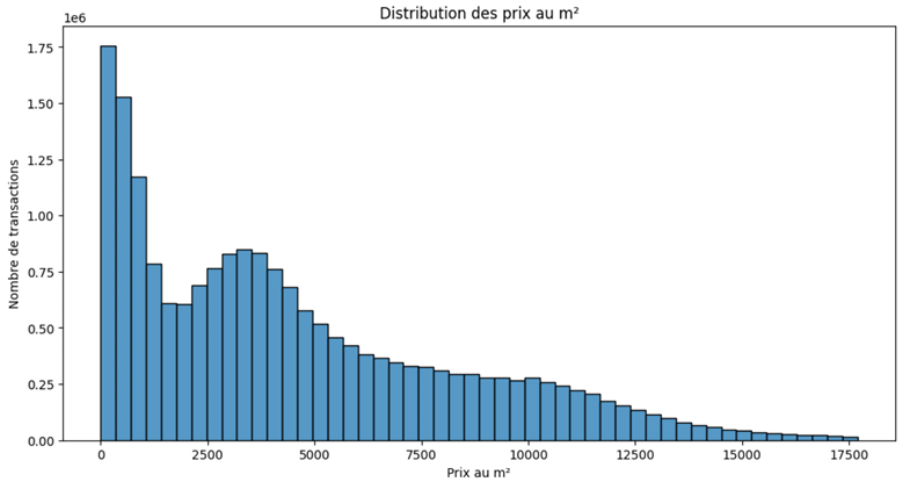

FIGURE 4 – Distribution prix au m²

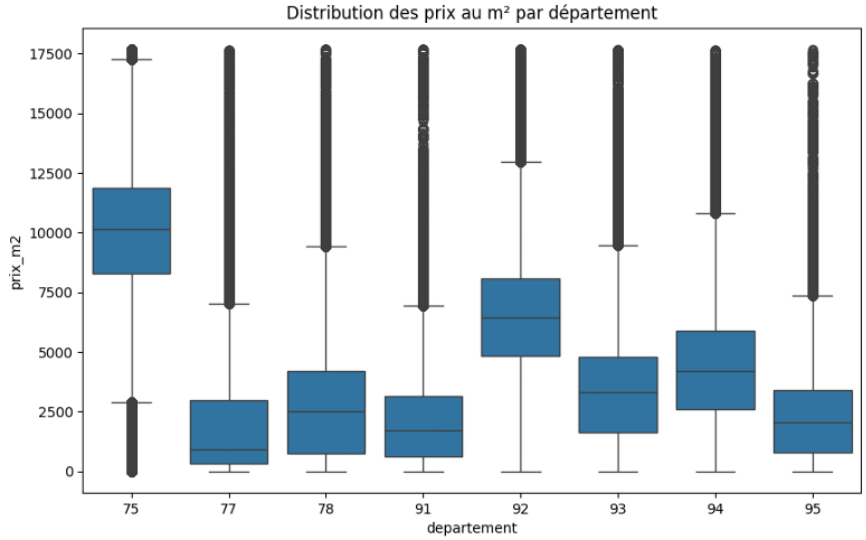

FIGURE 5 – Distribution prix au m² par département

Paris (75) se démarque nettement :
- Médiane autour de 10,000€/m²
- Forte dispersion
- Nombreux outliers vers le haut

Autres départements d'Île-de-France :
- Prix médians entre 2,500€ et 5,000€/m²
- Dispersion plus modérée
- Gradient de prix clair selon la distance à Paris Cela laisse paraître que le département est une variable clé

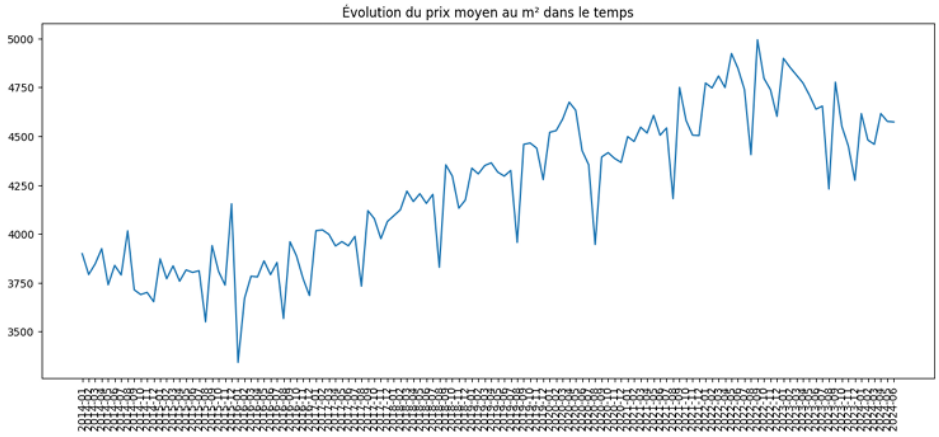

FIGURE 6 – Evolution temporelle du prix moyen au m^2

4 Modélisation

4.1 Sélection des Modèles

Nous avons retenu trois modèles différents pour la prédiction des prix :

1. **Régression Linéaire**
   — Points forts : simplicité, interprétabilité, rapidité
   — Limitations : hypothèses forte de linéarité, sensibilité aux valeurs extrêmes
2. **Random Forest**
   — Points forts : gestion des non-linéarités, robustesse des outliers, peu de paramètres à régler
   — Limitations : interprétabilité réduite, temps de calcul
3. **XGBoost**
   — Points forts : performances généralement bonnes, gestion fine des non-linéarités
   — Limitations : paramétrage complexe à paramètrer, risque de surapprentissage

4.2 Implémentation

Préparation des données :
On s'asssure du bon encodage des variables numériquess et catégoriellles. On vérifie également que notre jeu de données ne présente pas de valeurs manquantes ce qui est bien le cas.

```
Nombre de valeurs manquantes par colonne :
surface              0
confiance_menages    0
taux_directeur       0
cours_petrole        0
type_bien            0
departement          0
l_codinsee           0
dtype: int64
```

FIGURE 7 – Nombre de valeurs manquantes par variable

On divise ensuite notre jeu de données en un train set et un test set. le train set représente 80% du jeu de données tandis que le test set en représente 20%.

5 Résultats et Évaluation

5.1 Courbe d'apprentissage

On peut constater ci-dessous que les courbes d'apprentissages convergent pour le test set et le train set pour nos deux modèles.

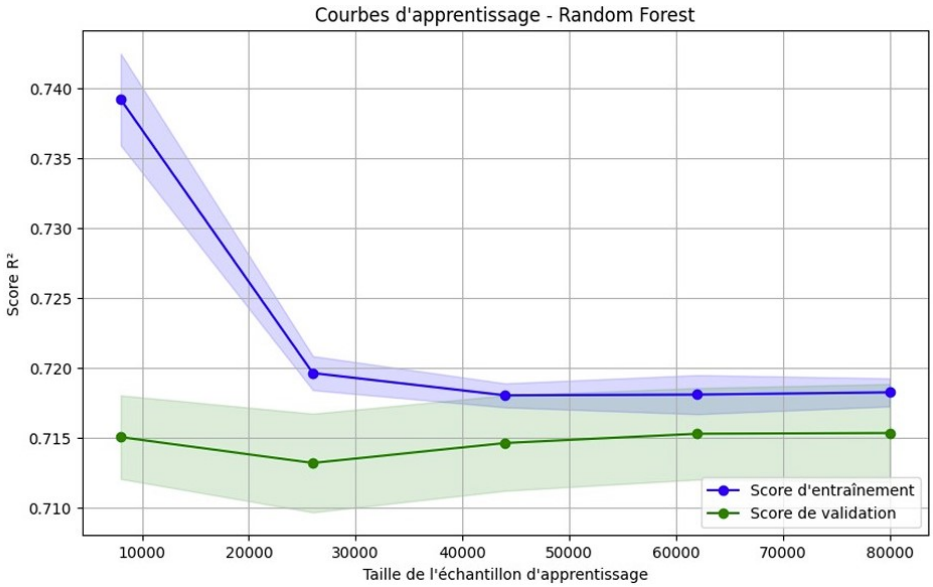

FIGURE 8 – Courbes d'apprentissage pour Random Forest

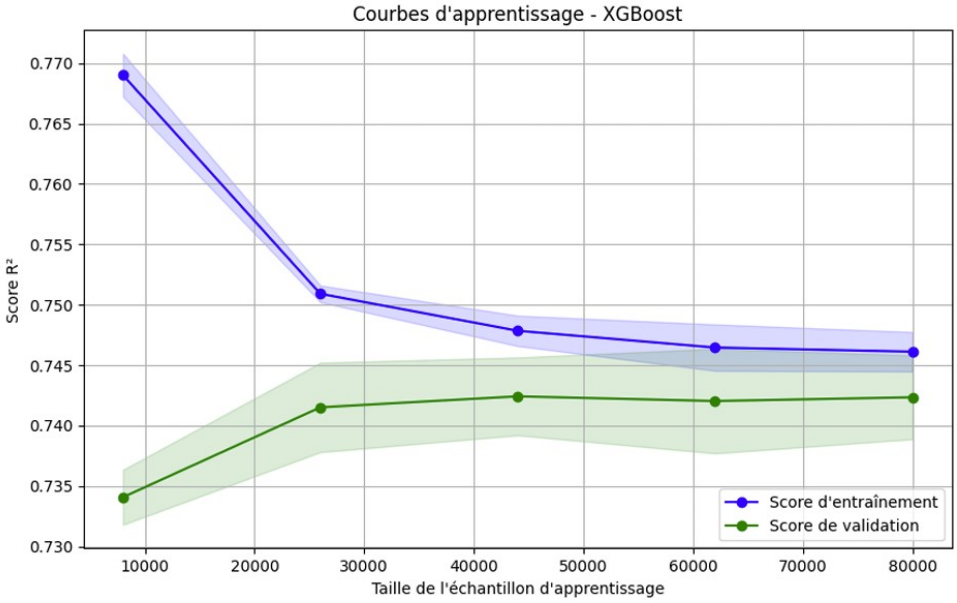

FIGURE 9 – Courbes d'apprentissage pour XG Boost

5.2 Comparaison des Performances

Les performances des modèles se distinguent significativement :

Modèle	RMSE (€/m²)	R^2	CV R^2 moyen
XGBoost	1 930,87	0,744	-0,083
Random Forest	2 043,33	0,714	-0,122
Régression linéaire	3 485,45	0,167	-2,666

TABLE 1 – Comparaison des performances des modèles

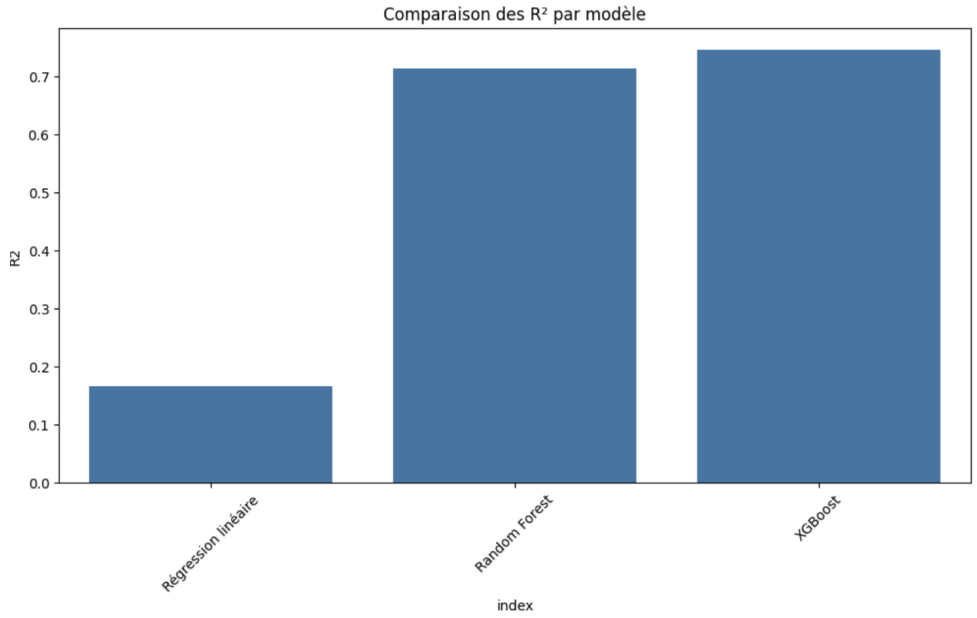

FIGURE 10 – Comparaison graphique des performances des modèles

5.3 Importance des Variables

L'analyse de l'importance relative des variables révèle :

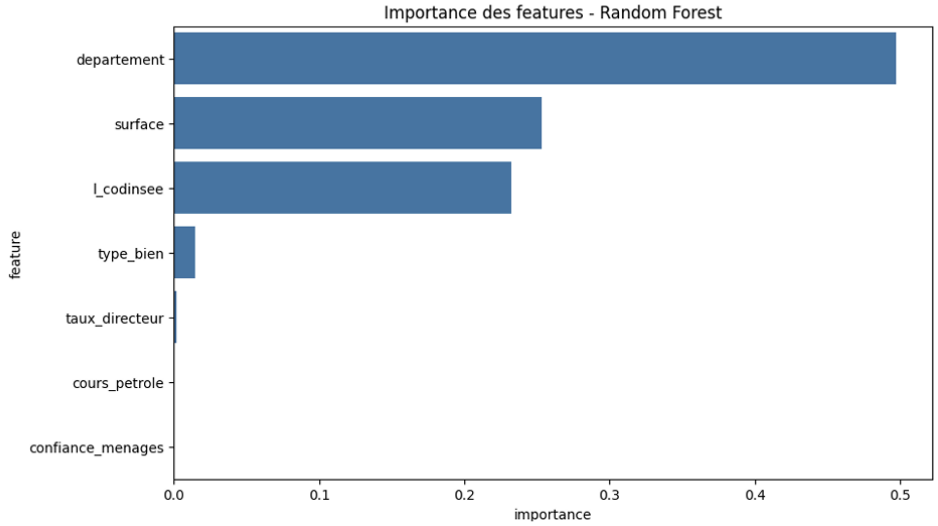

FIGURE 11 – Importance relative des variables explicatives pour Random Forest

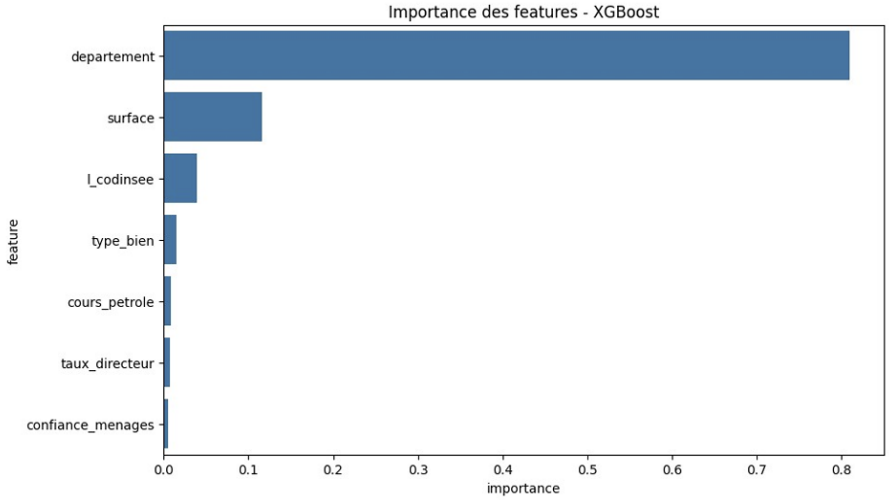

FIGURE 12 – Importance relative des variables explicatives pour XG Boost

Que la variable l plus explicative du prix selon les deux modèles est la variable departement avec 80% d'importance pour XG Boost et 50% pour Random Forest.
Elle est suivie par la surface qui a 11% d'importance pour XG Boost et 25% pour Random Forest.
On constate que le code INSEE a une importance de 24% pour Random Forest mais de seulement 5% pour XG Boost.
Quant aux variables macroéconomiques elles ont toutes moins de 5% d'importance pour les deux modèles.

6 Discussion

6.1 Analyse des Performances

Le modèle XGBoost se distingue par :
— Un coefficient R^2 de 0,744, démontrant une capacité prédictive robuste
— Une RMSE de 1 930€/m^2, permettant des estimations précises
— Une meilleure stabilité en validation croisée
Pour le modèle Random Forest on a :
— Un coefficient R^2 de 0,714, affichant une performance proche de celle de XG Boost
— Une RMSE de 2 043€/m^2, permettant également des estimations précises
— Une bonne robustesse
Quant à la régression linéaire :
— Un coefficient R^2 de 0,167, affichant une performance décevante
— Une RMSE de 3 485€/m^2, qui montre donc une précision bien inférieure aux deux autres modèles
— Cross Validation très instable

6.2 Perspectives d'Amélioration

Plusieurs axes d'amélioration ont été identifiés :
— Un rééchantillonnage pour une meilleure représentation des prix élevés
— L'application d'une transformation logarithmique sur les prix
— L'intégration de données locales complémentaires
— Une segmentation plus fine du marché

7 Conclusion

Notre étude met en évidence la prédominance des facteurs géographiques sur les variables macroéconomiques dans la formation des prix immobiliers en Île-de-France. Le modèle XGBoost développé démontre une capacité prédictive robuste, avec un R^2 de 0,744, tout en soulignant la forte segmentation géographique du marché.

L'évolution des prix sur la période 2014-2024 (+29%), particulièrement marquée après 2020, témoigne d'un marché dynamique et résilient. L'intégration future de données locales plus granulaires pourrait encore améliorer la précision des prédictions.