Compte rendu projet machine learning

HIMPE Arthur

Arts et Métiers Sciences et Technologies

Table des matières

Introduction :

L'objectif de ce projet est de calculer le prix d'un appartement parisien uniquement à partir de sa superficie et de sa localisation. Pour cela, nous allons créer quatre algorithmes qui vont nous permettre de répondre à ce problème. Une fois que nous auront répondu à la cette première question, nous essaierons d'étoffer notre modèle afin d'en améliorer la précision.

Pour entrainer notre algorithme, nous nous baserons sur le fichier de valeurs foncières géolocalisées sur Paris, disponible sur le site suivant : Page générale du dataset

Une fois les données récupérées, nous les avons triées. D'abord, nous avons limité notre document aux appartements, et avons donc supprimé les doublons ainsi que les dépendances, immeubles entiers, etc.

Ensuite, nous nous sommes aperçus que le tri des données n'était pas suffisant car il restait beaucoup de valeurs aberrantes, dues à des cessions symboliques ou à des appartements d'exception. Nous avons donc décidé de supprimer de notre étude les appartements dessous de 5 000€/m² et au-dessus de 25 000€/m².

Enfin, nous avons limité la surface car le prix du m^2 variait beaucoup sur de petits appartements ou sur des hôtels particuliers. Nous avons donc limité notre recherche aux appartements entre 40 et 200 m^2.

Après élagage des colonnes inutiles de notre dataset, il est resté 6 données réellement utiles : La longitude de l'appartement, sa latitude, le nom de sa rue, son prix, son nombre de pièces et son arrondissement.

Nous avons donc utilisé un fichier contenant un total de 5671 transactions qui correspondaient à nos critères. Ces dernières dont visibles en figure 1.

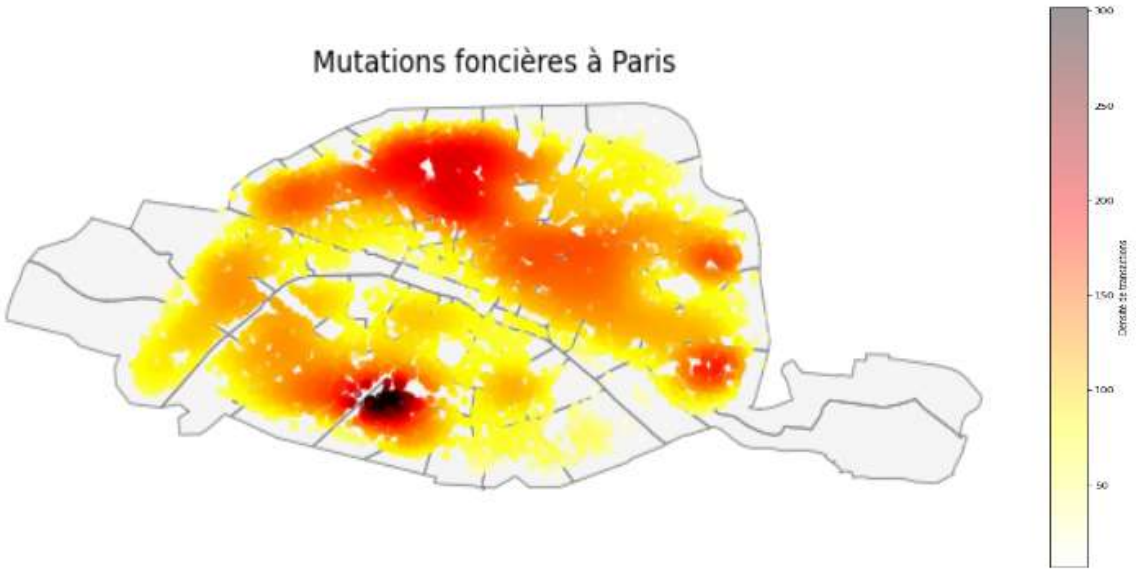

Figure 1: Mutations foncières à Paris en 2025

On peut déjà voir en figure 1 que les appartements qui valent de loin le plus cher (vue Louvre, Seine, Saint Michel) n'ont pas été très vendus, ce qui renforce notre confiance en ces données.

Partie 1) Détermination du prix d'un appartement simplement en fonction de sa position

Pour estimer le prix au m², nous avons mis en œuvre plusieurs approches distinctes : une méthode géométrique basée sur le voisinage (KNN) et une méthode ensembliste basée sur des arbres de décision (Random Forest). Pour chaque méthode, nous utiliserons 80% des données à notre disposition pour l'entrainement de l'algorithme, et 20% pour son test.

Les méthodes sont comparées sur les bases de l'indicateur R^2 et du MAE.

Le coefficient de détermination R^2 mesure la capacité du modèle à expliquer les variations de prix observées. Il compare les erreurs de nos prédictions à celles d'une simple moyenne : un R^2 de 0,45 signifie que la localisation et la surface expliquent 45 % de la valeur du bien, tandis que les 55% restants dépendent de facteurs non saisis par l'algorithme, comme l'étage ou le cachet. C'est l'indicateur clé de la pertinence de nos variables.

Le MAE (Mean Absolute Error), ou erreur absolue moyenne, traduit la performance du modèle en une valeur concrète et facile à interpréter : l'écart

moyen en euros par mètre carré. Contrairement au qui est un pourcentage abstrait, le MAE nous indique de combien nos prédictions se trompent en moyenne.

Cette métrique est essentielle pour juger de l'utilité réelle du projet, car elle permet de confronter l'erreur de l'algorithme à la réalité du marché.

I) La méthode KNN

L'algorithme des K-Plus Proches Voisins (KNN) repose sur une approche géométrique et intuitive du marché immobilier. Son principe fondamental est que des biens situés dans un périmètre géographique restreint et présentant des surfaces similaires doivent logiquement afficher des prix de vente comparables.

Pour estimer le prix d'un nouvel appartement, le modèle identifie les K transactions les plus proches au sein de notre base de données en calculant la distance euclidienne entre leurs coordonnées et leur superficie. La prédiction finale est alors obtenue en effectuant la moyenne des prix de ces voisins. Bien que cette méthode soit efficace pour capter les micro-marchés locaux, elle reste très sensible à la densité des données et peut manquer de précision dans les zones où les transactions sont moins nombreuses.

Les résultats de la figure KNN sont visibles en figure 2 :

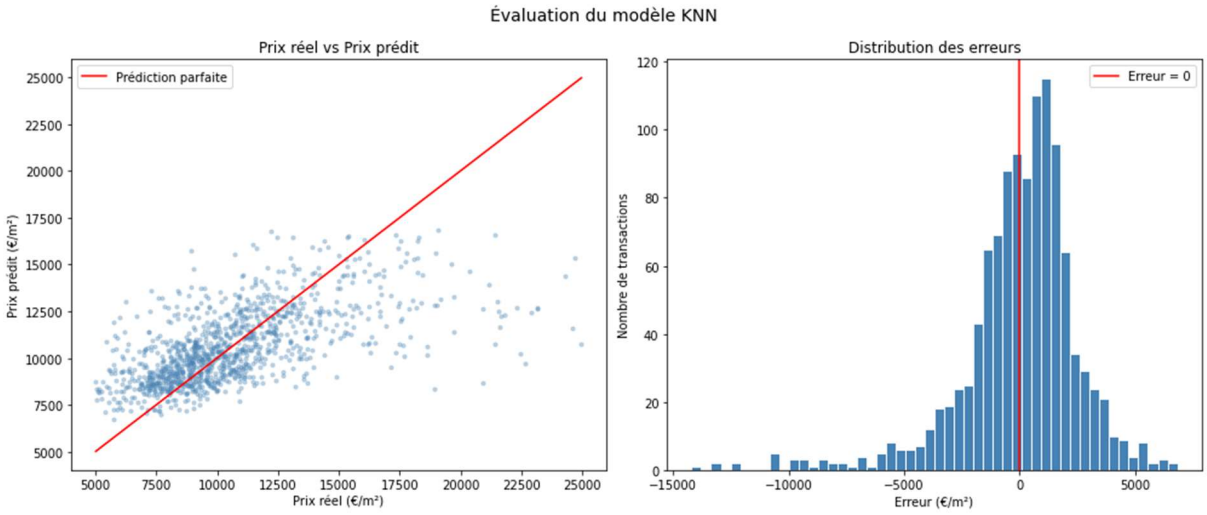

Figure 2: Tracés des écarts entre les valeurs prédites par l'algorithme KNN et les valeurs test réelles

Le tracé de la figure 2 a été réalisé pour un **K=33**, qui a permis d'optimiser la corrélation entre les données prédites et réelles (R^2).

On obtient R^2_{Train} : 0.404 et R^2_{Test} : 0.380, ce qui indique une différence de moins de 2% entre les résultats donnés et les résultats de l'algorithme.

Ce modèle se trompe en moyenne de 1799 € dans le prix du m² d'un appartement.

Néanmoins, même si la différence entre R^2_{Train} et R^2_{Test} est faible, le dernier coefficient reste plutôt faible, car il explique que 38% du prix d'un appartement est du à sa localisation, sans expliquer quels sont les facteurs manquants.

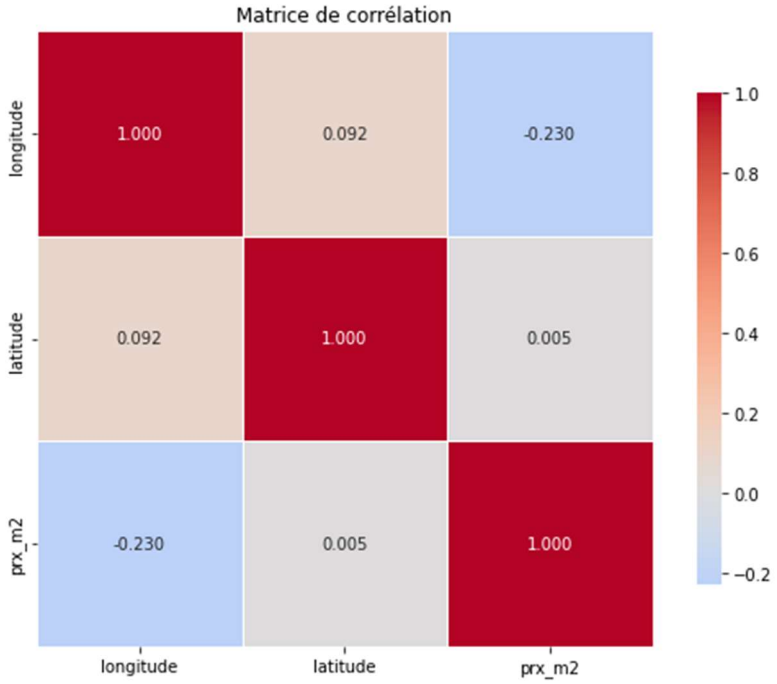

Figure 3: Matrice de correlation pour la methode KNN

La matrice de corrélation visible en figure 3 révèle une structure complexe où aucune variable n'exerce une influence linéaire dominante sur le prix. La corrélation négative de -0,23 entre la longitude et le prix au m² traduit statistiquement la baisse progressive des prix lors du passage de l'Ouest vers l'Est parisien.

En revanche, la corrélation quasi nulle de la latitude (0,005) indique qu'il n'existe pas de tendance Nord-Sud simpliste. Ces faibles coefficients confirment que la localisation ne peut être traitée de manière isolée et justifient pleinement l'usage d'algorithmes non-linéaires, comme le Random Forest, pour capturer les interactions géographiques subtiles que cette matrice ne peut révéler seule.

II) La méthode Random Forest

Pour tenter de résoudre le problème de précision de la méthode KNN, nous allons utiliser la méthode Random Forest.

Cet algorithme repose sur la combinaison de multiples arbres de décision pour gagner en précision et en stabilité. Plutôt que de s'appuyer sur un seul modèle de prédiction, il crée une "forêt" d'arbres indépendants, chacun étant entraîné sur un échantillon aléatoire de transactions et de variables. Pour estimer le prix d'un appartement, chaque arbre propose sa propre réponse, et le modèle final retient la moyenne de l'ensemble de ces résultats.

L'avantage majeur de cette approche réside dans sa robustesse et sa capacité à traiter des relations non-linéaires complexes, comme l'impact combiné de la surface et du quartier sur le prix au m². En multipliant les points de vue, le Random Forest réduit considérablement le risque de surapprentissage par rapport à un arbre de décision isolé. Dans notre étude, il permet de hiérarchiser l'importance de variables comme la localisation géographique et la superficie, offrant ainsi une vision nuancée du marché immobilier parisien malgré l'absence de données descriptives sur l'état des biens. Les résultats qu'il nous renvoient sont visibles en figure 4 :

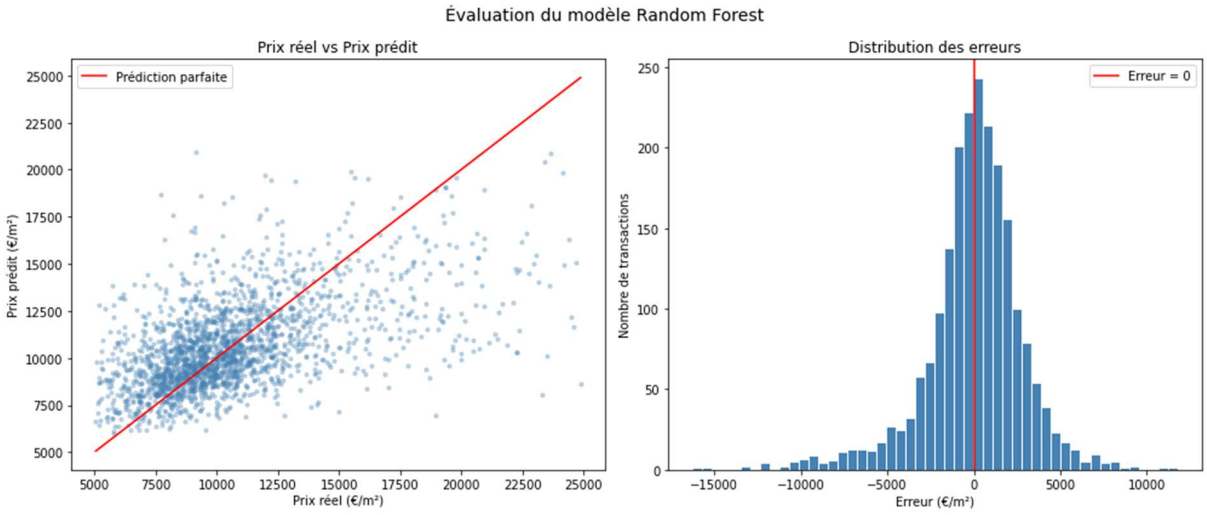

Figure 4: Résultats de la méthode Random Forest

On obtient R^2_{Train} : 0.813 et R^2_{Test} : 0.273. L'écart de 0.54 points est énorme et montre que le modèle a "mémorisé" les données d'entraînement au lieu d'apprendre des patterns généralisables. C'est le phénomène d'underfitting, qui est surement du au fait que l'algorithme ne prends en entrée que 3 données, ce qui empêche une réflexion profonde.

Ce modèle se trompe en moyenne de 1910 € dans le prix du m^2 d'un appartement.

Un R^2_{Test} : 0.273 n'est pas acceptable. Il devra être corrigé dans la partie suivante.

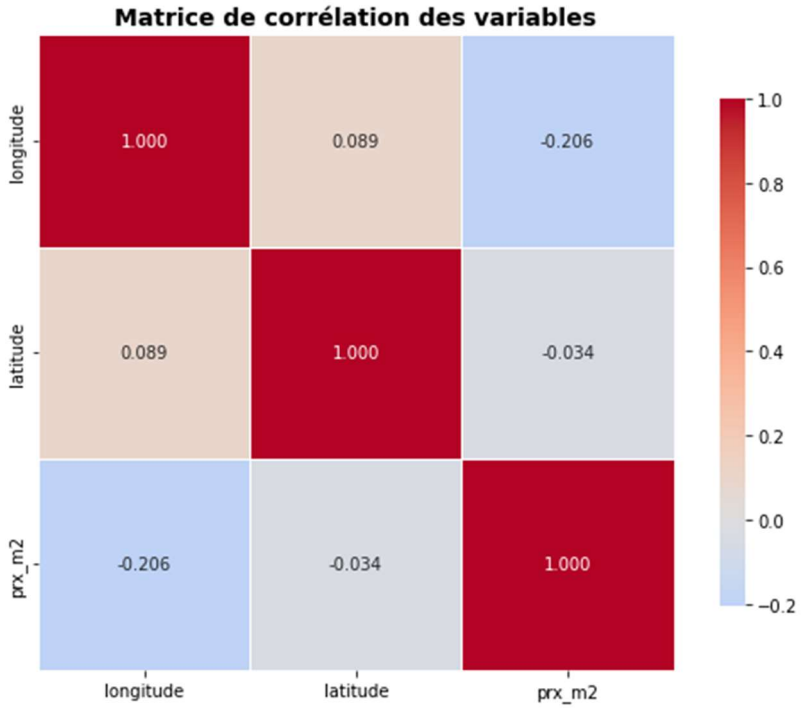

Figure 5: Matrice de correlation de l'algorithme Random Forest

Néanmoins, la matrice de corrélation visible en figure 5 montre que les liens entre les variables trouvés par l'algorithme Random Forest ne sont pas très différents de ceux trouvés par l'algorithme KNN.

Partie 2) Détermination du prix d'un appartement en fonction de sa position, de son nombre de pièces et de sa superficie

Les résultats obtenus dans la partie 1 sont partiellement satisfaisants. En effet, on en a déduit que la position géographique d'un appartement dans Paris influe sur 40% du prix de l'appartement.

Cela veut dire que les 60% restants du prix de l'appartement viennent d'autres facteurs. Pour essayer d'expliquer ces 60%, nous allons complexifier nos modèles en rajoutant des données obtenables dans le fichier CSV.

I) La méthode KNN

L'algorithme des K-Plus Proches Voisins multi-paramètres affine la logique du premier algorithme en intégrant une vision multidimensionnelle du patrimoine immobilier. Son principe fondamental reste l'analogie, mais il ne se limite plus à la seule proximité physique ou à la surface : il considère désormais le bien comme un vecteur de caractéristiques techniques (nombre de pièces, superficie, arrondissement).

Pour estimer la valeur d'un actif, le modèle projette chaque transaction dans un espace mathématique à n dimensions, où chaque paramètre représente un axe spécifique. La distance entre le bien cible et ses voisins n'est plus seulement kilométrique, mais reflète une similitude globale calculée par une distance euclidienne pondérée.

Les résultats de la figure KNN multi-paramètres sont visibles en figure 6 :

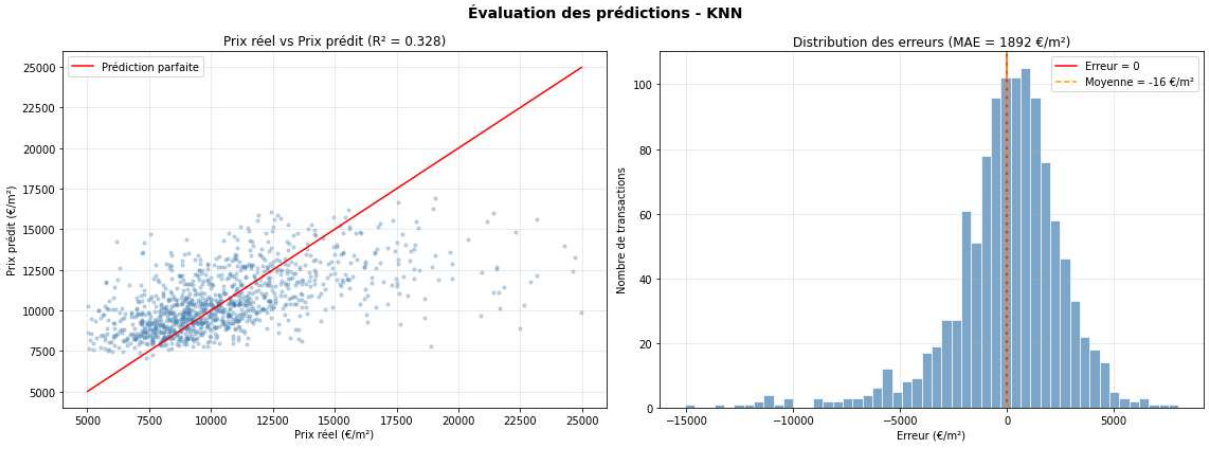

Figure 6 : Resultats de l'algoritme KNN multi parametres

Le tracé de la figure 2 a été réalisé pour un **K=27**, qui a permis d'optimiser la corrélation entre les données prédites et réelles (R^2).

On obtient R^2_{Train} : 0.393 et R^2_{Test} : 0.328. Ces résultats sont décevants car en intégrant moins de paramètres, on obtenait de meilleurs résultats (R^2_{Train} : 0.404 et R^2_{Test} : 0.380 pour le premier algo KNN)

Ce modèle se trompe en moyenne de 1891 € dans le prix du m^2 d'un appartement.

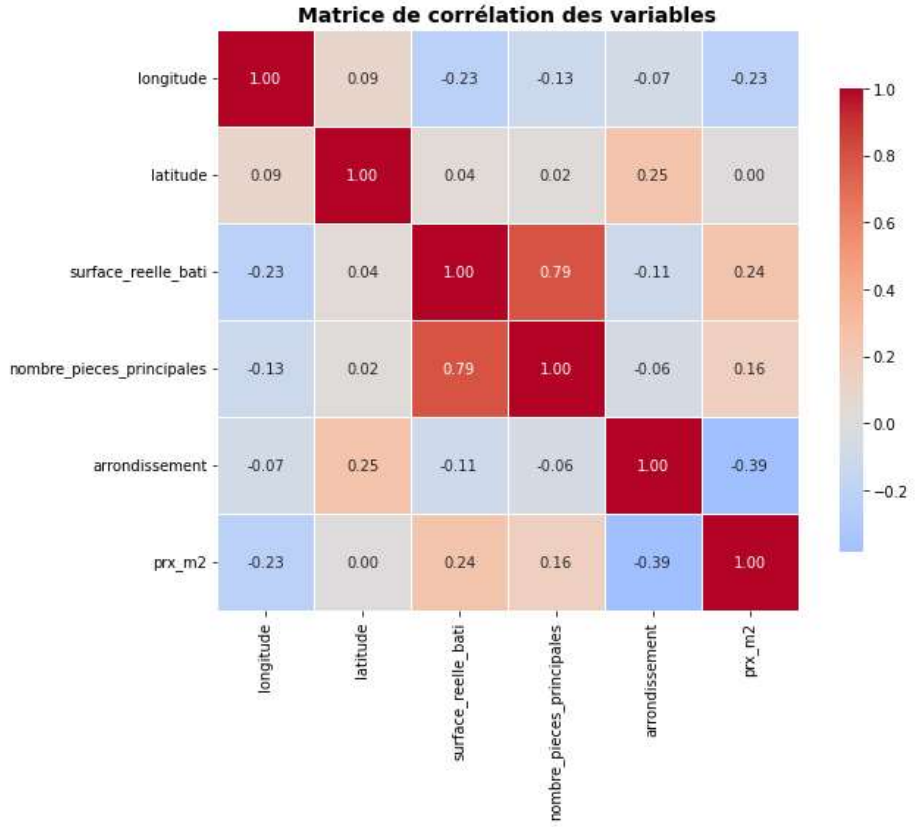

Figure 7 : Matrice de correlation de l'algorithme KNN multi-parametres

La matrice de correlation visible en figure 7 révèle une corrélation majeure (0,79) entre la surface réelle et le nombre de pièces, confirmant une forte redondance entre ces deux indicateurs de volume. Pour l'algorithme KNN, cette proximité statistique suggère qu'une seule de ces variables pourrait suffire à définir la "taille" d'un voisin, évitant ainsi de surpondérer involontairement le facteur de superficie dans le calcul des distances.

De plus, même si ce n'est pas directement visible sur cette matrice, il y a une redondance entre le numéro de l'arrondissement et la localisation. On surpondère donc surement l'emplacement des appartements par rapport aux autres paramètres.

L'analyse souligne l'influence prédominante de l'arrondissement sur le prix au m², avec une corrélation négative de -0,39. Ce résultat indique que la valeur immobilière décroît à mesure que le numéro de l'arrondissement augmente, marquant une segmentation géographique claire du marché. À l'inverse, la latitude présente une corrélation nulle (0,00), prouvant que la position Nord-Sud n'exerce aucune influence directe sur la valorisation des biens dans ce jeu de données.

Enfin, le lien modéré entre la surface et le prix au m² (0,24) démontre que la valeur unitaire n'est pas strictement proportionnelle au volume du bien. Cette décorrélation relative justifie l'utilisation d'une approche multi-paramètres : pour obtenir une prédiction fiable, le modèle ne peut se contenter de la taille du

logement, mais doit intégrer la combinaison spécifique de l'emplacement et des caractéristiques techniques pour identifier les voisins les plus comparables.

II) La méthode Random Forest

L'intégration de paramètres supplémentaires tels que l'arrondissement, le nombre de pièces et la surface permet au Random Forest d'affiner sa compréhension des disparités du marché parisien. En croisant ces variables, l'algorithme ne se contente plus d'une approche géographique brute, mais parvient à modéliser des effets d'interaction complexes.

Grâce à cette structure multi-paramètres, le modèle renvoie une estimation bien plus robuste en captant la "signature" spécifique de chaque type de bien. Là où un arbre isolé pourrait surinterpréter une transaction atypique, la forêt pondère l'influence de la surface par la rareté induite par le nombre de pièces au sein d'un quartier donné. Les résultats de ce nouvel algorithme sont visibles en figure 8 :

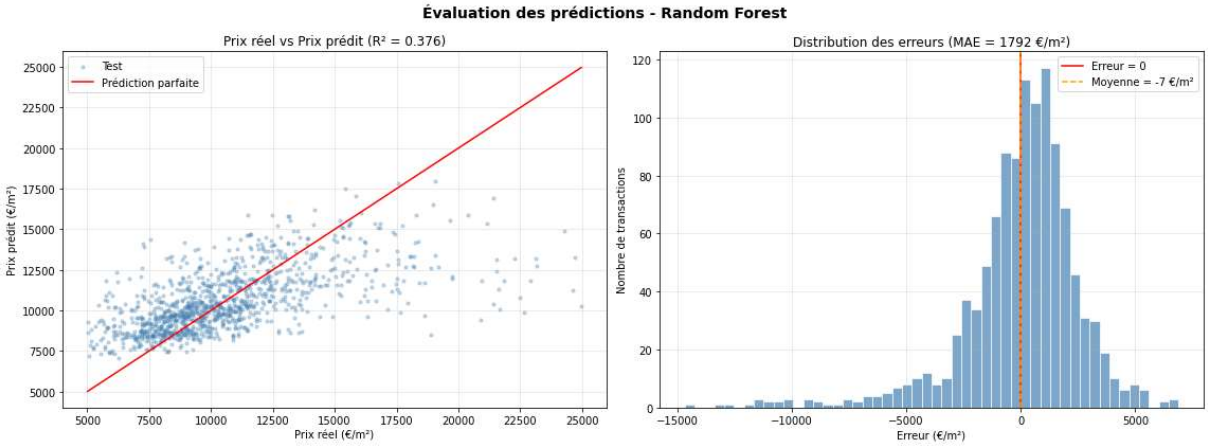

Figure 8 : résultats de l'algoritme random forest multi parametres

On obtient R^2_{Train} : 0.464 et R^2_{Test} : 0.373. Les résultats sont donc bien meilleurs que dans la partie 1, où nous avions obtenus R^2_{Train} : 0.813 et R^2_{Test} : 0.273.

Ce modèle se trompe en moyenne de 1794 € dans le prix du m^2 d'un appartement.

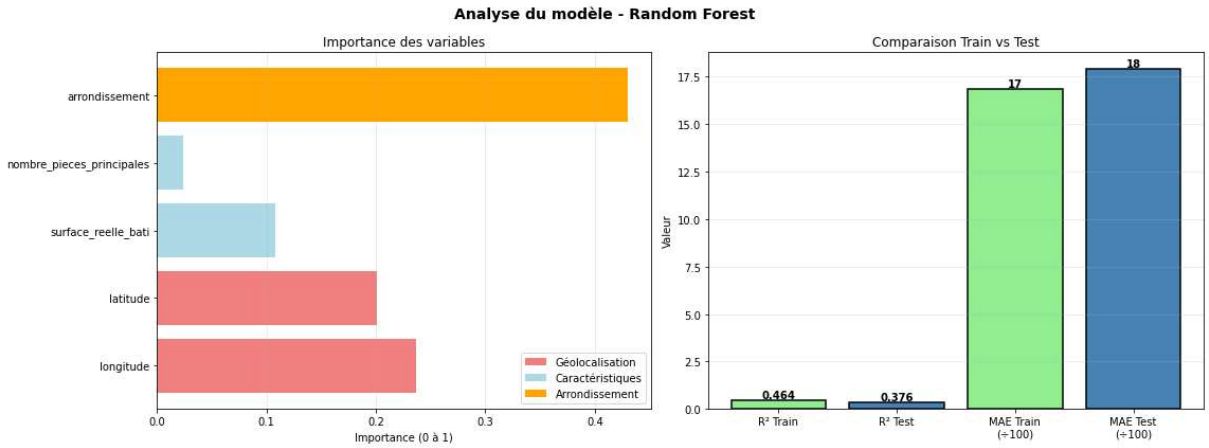

Figure 9 : résulats et importance des variables dans la methode random forest multi paramètres

L'analyse de l'importance des variables visible en figure 9 confirme que la localisation reste le moteur prédominant de la valeur immobilière, suivi de près par les coordonnées géographiques (longitude et latitude). Cette hiérarchie démontre que le Random Forest accorde plus de poids aux "micro-marché" qu'aux caractéristiques intrinsèques du bien, comme la surface ou le nombre de pièces, pour expliquer les variations de prix. En somme, l'adresse précise du logement est le signal que l'algorithme utilise prioritairement pour réduire l'erreur de prédiction.

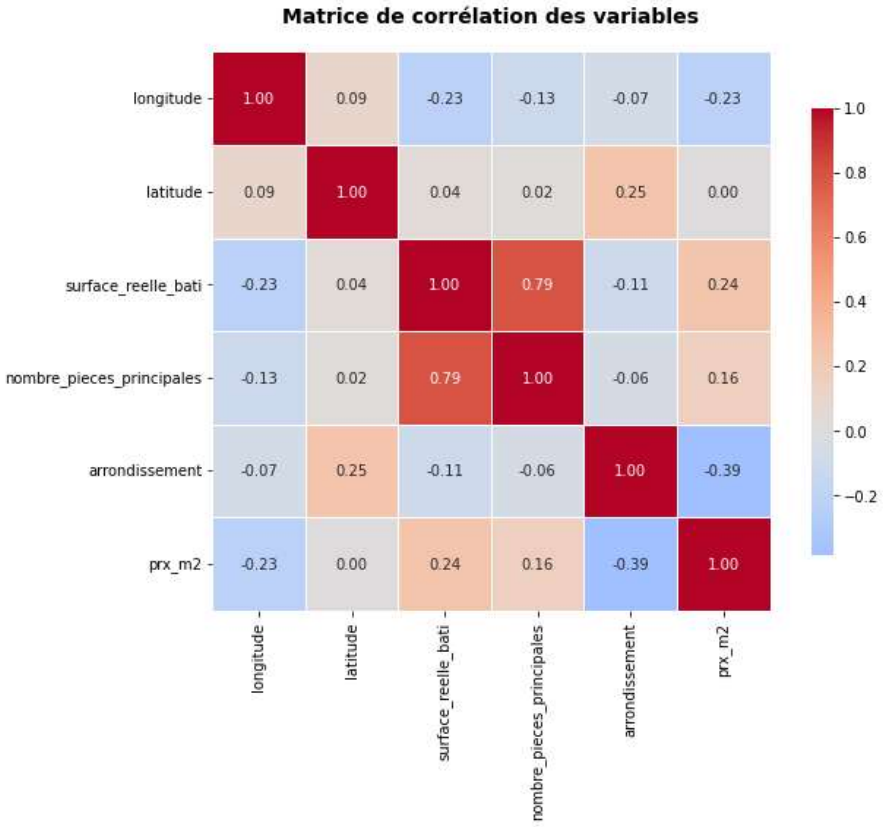

Figure 10 : matrice de correlation des variables de la methode random forest multi parametres

La matrice visible en figure 10 est exactement la même que celle obtenue pour la méthode KNN multi paramètres. On en tire donc les mêmes conclusions.

Explication des écarts et conclusion :

A travers les différentes méthodes mises en place dans notre raisonnement, nous avons su obtenir plusieurs algorithmes nous donnant des résultats intéressants.

Les algorithmes KNN et Random Forest multi-paramètres nous ont respectivement donné des R^2_{test} de 0.380 et de 0.373. Cela signifie que **les données qu'utilisent ces algorithmes expliquent ~38% de la variance du prix d'un appartement**, avec une **erreur moyenne d'environ 1800€ par m²**.

Si ces performances sont cohérentes entre les deux algorithmes, elles restent néanmoins modestes pour une application concrète d'estimation immobilière. Pour un appartement moyen de 60m², cette marge d'erreur de ±1800€/m² représente une incertitude de ±108 000€ sur le prix total, ce qui constitue une fourchette trop large pour une prise de décision fiable en matière d'achat ou de vente.

Les 62% de variance non expliquée s'expliquent principalement par l'absence de variables qualitatives essentielles dans notre jeu de données. En effet, le prix d'un bien immobilier parisien dépend fortement de facteurs que nous n'avons pas pu intégrer : l'état général du bien (rénové, à rafraîchir, vétuste), l'étage et la présence d'un ascenseur, l'exposition et la luminosité, la vue, les équipements (balcon, cave, parking), l'année de construction, ou le DPE. De plus, nos variables de localisation (longitude/latitude) ne capturent que la position géographique générale, sans tenir compte des micro-variations au sein d'un même quartier : proximité immédiate du métro, calme de la rue, ou prestige particulier de certaines adresses. Ces éléments, difficilement quantifiables mais déterminants pour le marché immobilier parisien, constituent le principal frein à l'amélioration de nos modèles.

Pour contextualiser ces résultats, il est utile de les comparer aux performances généralement observées dans le domaine de la prédiction immobilière. **Les modèles d'estimation développés par les professionnels du secteur (plateformes immobilières, banques, agences) atteignent typiquement des R² compris entre 0.65 et 0.85, avec des erreurs moyennes de 8 à 12% du prix total**. Ces performances supérieures s'expliquent par l'accès à des bases de données beaucoup plus riches, intégrant des dizaines

de variables qualitatives collectées par des experts : diagnostics de performance énergétique (DPE), descriptions détaillées de l'état intérieur, photographies analysées par vision par ordinateur, et même des données de sentiment issues d'annonces textuelles. Dans le contexte académique et de recherche, les études publiées sur l'estimation immobilière rapportent des R^2 variant de 0.50 à 0.75 selon les villes et les données disponibles. Notre résultat de 0.38, bien qu'en deçà de ces standards, reste cohérent avec un jeu de données limité à quatre variables basiques, et se situe dans la fourchette basse mais réaliste pour un modèle exploratoire fondé uniquement sur des données publiques DVF.

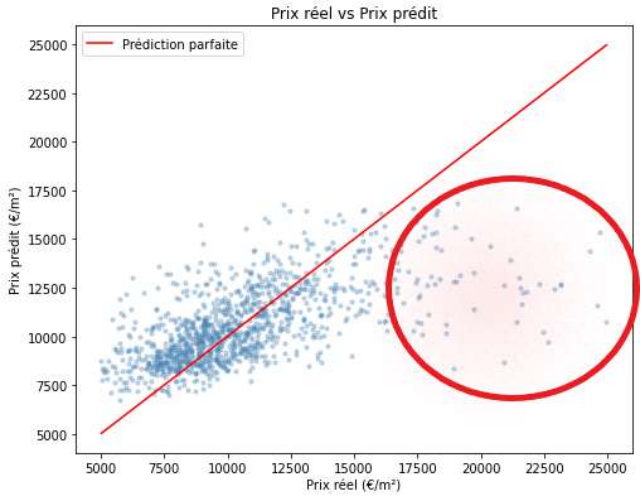

Figure 11 : Mise en avant des appartements d'exception

De plus, comme visible dans le cercle rouge de la figure 11, certains appartements ont un prix du m^2 très élevé par rapport à la prédiction, alors qu'il n'existe pas de biens dans le cas inverse. Ces appartements d'exception limitent dans tous les cas la performance de l'algorithme.

Dans leur état actuel, nos modèles restent utiles pour certaines applications à faible exigence de précision : détection d'anomalies grossières dans les prix affichés, analyse de tendances générales par quartier, ou première fourchette indicative pour un bien. En revanche, ils ne sont pas suffisamment fiables pour une estimation individuelle précise destinée à une transaction immobilière réelle. **Avec un enrichissement ciblé des données, notamment l'intégration de l'état du bien et de l'étage, il serait réaliste de viser un R^2 de 0.65 à 0.75 et une MAE de 1000 à 1200€/m²**, seuils à partir desquels le modèle deviendrait véritablement exploitable pour des estimations individuelles. Ce travail démontre néanmoins l'intérêt de l'apprentissage automatique pour l'estimation immobilière, tout en soulignant l'importance cruciale de la qualité et de la richesse des données pour obtenir des prédictions fiables.

Arts et Métiers Sciences et Technologies

Contact

—

Mail

Téléphone

Adresse