

Louvain School of Management

Estimation de la valeur de performance d'un joueur de football professionnel à partir de ses statistiques avancées : une application aux cinq grands championnats européens

Mémoire de Master

Auteur : NDIKUMANA Lukman
Promotrice : Prof. SATINET Chloé
Année académique 2025-2026
Mémoire présenté en vue de l'obtention du grade de Master en
Ingénieur de Gestion

DECLARATION

The declaration below is **mandatory and must appear on the first page (after the front page) of your master's thesis.**

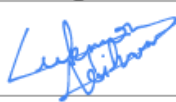
STATEMENTS ON YOUR USE OF GENERATIVE AI

Confirm the following statements:

Statements on responsible use of generative AI	Your answer
The content of my/our master's thesis is my/our original output, even if I/we used generative AI to help prepare it.	☒ Yes ☐ No
I/we master the theories, tools and methods that I/we use for the thesis.	☒ Yes ☐ No
I/we verified that the output presented in the thesis is accurate and valid, and that it contains no hallucination, no false reference, etc.	☒ Yes ☐ No
I/we acknowledge that I/we master the final output and that I am/we are able to explain it and answer questions about it.	☒ Yes ☐ No
I/we made sure not to provide generative AI tools with access to confidential data or information	☒ Yes ☐ No

SIGNATURE

Sign your declaration:

	Author last name, first name(s)	Date	Signature
1	NDIKUMANA Lukman	05/08/2026	
2			

NOTE

For this document, DeepL has been used to propose formulations, but the text is original, except the extracts explicitly mentioned as coming from other sources, which are mentioned in footnotes.

Résumé

Le marché des transferts de football alimente depuis plusieurs années l'idée d'une bulle spéculative mais qualifier une bulle suppose de disposer d'une valeur fondamentale de référence dont le prix observé s'écarte (grandeur qui n'existe nulle part dans la littérature sur les transferts). Ce mémoire propose d'en construire un proxy de la valeur de performance : à partir des seules statistiques avancées observables avant un transfert, les modèles prédisent l'indemnité que la performance sportive d'un joueur justifierait. L'écart entre ce proxy et l'indemnité réellement observée constitue l'indicateur de sur- ou sous-évaluation central de ce travail.

L'étude porte sur 1 172 transferts à indemnité réelle connue, issus des cinq grands championnats européens sur la période 2018-2024, à partir des bases FBref et Transfermarkt. La démarche combine trois étapes : une analyse en composantes principales réalisée séparément pour chacun des sept postes réduit 63 statistiques avancées à des facteurs synthétiques représentatifs des dimensions de la performance. Une décomposition Shapley-LMG assortie d'intervalles de confiance bootstrap identifie les facteurs dominants pour chaque poste. Enfin, trois modèles supervisés sont comparés en validation croisée et le Random Forest (retenu sur six postes sur sept) produit les estimations constituant le proxy de la valeur de performance.

La performance observable explique 7 % à 22 % de la variance des indemnités selon les postes, démontrant que la majorité du prix relève de déterminants extra-sportifs. La structuration des écarts confirme cette lecture : les plus fortes sur-évaluations concernent systématiquement des transferts très médiatisés (Bellingham, Rice, Kane), tandis que les sous-évaluations touchent des profils discrets ou en mobilité descendante. La convergence entre la décomposition LMG et les valeurs SHAP est confirmée sur cinq postes sur sept, renforçant la robustesse des résultats.

Ce mémoire est le premier à construire explicitement un proxy de la valeur fondamentale d'un joueur à partir de données réelles de transfert, dans un pipeline analytique entièrement documenté et reproductible.

Remerciements

Avant tout, je souhaite exprimer ma profonde gratitude à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce mémoire.

Je tiens à remercier tout particulièrement Madame Chloé Satinet, ma promotrice. Son encadrement rigoureux, sa disponibilité et la qualité de ses retours ont été d'une aide précieuse tout au long de la préparation de mon travail de fin d'études.

Je tiens également à remercier chaleureusement Souleymane Thiandoum et Abidli Imane, dont les regards extérieurs et les échanges enrichissants, nourris par leurs propres horizons scientifiques, m'ont permis d'envisager ma problématique sous des angles nouveaux et d'en affiner la réflexion.

Enfin, je souhaite également remercier ma famille et mes amis, pour leur soutien tout au long de mes études. Merci d'avoir été là et de m'avoir toujours poussée à donner le meilleur de moi-même.

Table des matières

Liste des tableaux	6
Liste des graphiques et illustrations	7
1. Introduction	8
1.1. Contexte	8
1.2. Question de recherche	9
2. Revue de littérature et cadre conceptuel	10
2.2. Statistiques avancées dans le football	14
2.2.1. Typologie et calcul des statistiques avancées mobilisées dans l'étude.....	16
2.3. Modèles quantitatifs d'estimation des prix de transfert.....	16
3. Méthodologie	18
3.1. Collecte des données	18
3.1.1. Nature et sources des données	18
3.1.2. Périmètre géographique et temporel.....	19
3.1.3. Choix de la variable cible : indemnité réelle plutôt que valeur marchande ..	20
3.1.4. Jonction des sources et périmètre des transactions retenues.....	20
3.1.5. Échantillon final	21
3.2. Nettoyage et préparation des données	22
3.2.1. Traitement des prix nuls et des prix manquants	22
3.2.2. Neutralisation des statistiques sur temps de jeu très faible	22
3.2.3. Standardisation	23
3.2.4. Vérification statistique du biais d'exclusion	23
3.2.5. Sélection des variables candidates et traitement des redondances	25
3.3. Analyse en composantes principales (ACP)	25
3.3.1. Principe de la méthode	25
3.3.2. Pertinence de l'ACP	26
3.3.3. Conditions d'utilisation et validation.....	26
3.3.4. Choix du nombre de composantes	27
3.3.5. Résultats de l'ACP	27
3.4. Décomposition de l'importance des facteurs : la méthode Shapley-LMG.....	28
3.4.1. Le problème à résoudre : mesurer la contribution de chaque facteur	28
3.4.2. Principe de la méthode LMG	29

3.4.3. Précision sur le rôle de la méthode dans ce travail.....	29
3.4.4. Conditions d'utilisation et diagnostics	30
3.4.5. Validation par bootstrap : quantifier l'incertitude	31
3.4.6. Résultats de la décomposition LMG	31
3.5. Construction du proxy de la valeur de performance par apprentissage automatique supervisé.....	32
3.5.1. Objectif : estimer une valeur qui n'est observable nulle part	32
3.5.2. L'apprentissage supervisé.....	33
3.5.3. Les trois modèles comparés	34
3.5.4. Conditions d'utilisation et précautions	35
3.5.5. Un cas d'école : la sensibilité de la régression aux valeurs extrêmes	36
3.5.6. Comparaison des modèles : métriques et tests statistiques	36
3.5.7. Sélection finale et calcul de la valeur de performance	37
3.5.8. Interprétabilité : les valeurs SHAP.....	38
4. Résultats.....	38
4.1. Capacité explicative de la performance sportive	39
4.1.1. Analyse des graphiques : les prédictions face aux prix réels	40
4.2. Hiérarchie des dimensions de la performance et validation croisée	42
4.3. Écarts entre prix réel et valeur de performance	43
5. Discussion et perspectives	44
5.1. Retour sur la question de recherche.....	44
5.2. Apports méthodologiques.....	45
5.3. Limites	45
5.4. Perspectives	46
6. Conclusion.....	47
7. Bibliographie	49
9. Annexes	51
Annexe 1 : Typologie des données mobilisées	51
Annexe 2 : Pool initial de 63 variables candidates (issu de FBref)	51
Annexe 3 : Correspondance entre F (SHAP), RC (ACP) et nom de dimension.....	56
Annexe 5 : Graphiques des valeurs de SHAP par poste.....	60
Annexe 6 : Graphiques de prédiction par poste.....	64

Annexe 7 : Les méthodes rencontrées	67
Annexe 8 : Les écarts positifs les plus importants.....	68
Annexe 9 : Test du χ^2 sur les transferts exclus	69
Annexe 10 : Facteur dominant par poste selon la décomposition LMG.....	70
Annexe 11 : Comparaison du facteur dominant identifié par le LMG.....	70
Annexe 12 : RMSE en validation croisée	71
Annexe 13 : Synthèse de l'ACP finale par poste.....	71

Liste des tableaux

Tableau 1: Répartition de l'échantillon final par poste (cas complets mobilisés dans l'ACP)	21
Tableau 2: Test du χ^2 sur les transferts exclus pour conflit de correspondance entre sources	69
Tableau 3 : Synthèse de l'ACP finale par poste	71
Tableau 4 : Facteur dominant par poste selon la décomposition LMG, avec intervalles de confiance bootstrap (1 000 répétitions).....	70
Tableau 5 : RMSE en validation croisée (échelle logarithmique du prix), par poste et par modèle	71
Tableau 6 : R^2 en validation croisée, par poste et par modèle	39
Tableau 7 : Comparaison du facteur dominant identifié par le LMG (régression) et par les valeurs SHAP (modèle d'arbres)	70
Tableau 8 : Les quinze transferts présentant les écarts positifs les plus importants (sur-évaluation par rapport au profil statistique	68
Tableau 9: Typologie des données mobilisées.....	51
Tableau 10 : Tableau correspondance spécifique au poste de défenseur central	56
Tableau 11 : Tableau correspondance spécifique au poste de milieu latéral.....	57
Tableau 12 : Tableau correspondance spécifique au poste de milieu défensif	58
Tableau 13 : Tableau correspondance spécifique au poste d'ailier	58
Tableau 14 : Tableau correspondance spécifique au poste de buteur	59
Tableau 15 : Tableau correspondance spécifique au poste de milieu offensif.....	59
Tableau 16 : Tableau correspondance spécifique au poste de gardien	59

Liste des graphiques et illustrations

Figure 1 : Facteurs influençant l'indemnité de transfert d'un joueur de football (Poli et al., s. d.)	11
Figure 2 : RMSE (échelle logarithmique) par poste et par modèle	39
Figure 3 : Prédictions contre prix réels, poste de buteur (n = 227)	41
Figure 4 : Prédictions contre prix réels, poste de milieu offensif (n = 72)	41
Figure 5 : Valeurs SHAP du modèle Random Forest, poste de défenseur central.....	60
Figure 6 : Valeurs SHAP du modèle Random Forest, poste d'ailier	61
Figure 7 : Valeurs SHAP du modèle Random Forest, poste de buteur	61
Figure 8 : Valeurs SHAP du modèle Random Forest, poste de gardien.....	62
Figure 9 : Valeurs SHAP du modèle Random Forest, poste de latéral	62
Figure 10 : Valeurs SHAP du modèle Random Forest, poste de milieu défensif	63
Figure 11 : Valeurs SHAP du modèle Random Forest, poste de milieu offensif.....	63
Figure 12 : Prédictions contre prix réels, poste de latéral (n = 161)	64
Figure 13 : Prédictions contre prix réels, poste de défenseur central (n = 203)	64
Figure 14 : Prédictions contre prix réels, poste de milieu défensif (n = 275)	65
Figure 15 : Prédictions contre prix réels, poste d'ailier (n = 166)	65
Figure 16 : Prédictions contre prix réels, poste de gardien (n = 68)	66
Figure 17 : Les méthodes rencontrées.....	Erreur ! Signet non défini.

1. Introduction

1.1. Contexte

Au fil des dernières décennies, le marché des transferts¹ est devenu l'un des domaines les plus exposés et les plus stratégiques de l'économie du football professionnel. L'augmentation constante des droits télévisés, l'arrivée d'investisseurs disposant de moyens considérables et la médiatisation croissante des recrutements ont profondément modifié les dynamiques du secteur. Dans ce contexte, les indemnités de transfert ont atteint des montants sans précédent, faisant du recrutement un enjeu à la fois sportif, économique et stratégique pour les clubs. Cette inflation continue alimente depuis plusieurs années l'idée selon laquelle le football professionnel serait entré dans une phase de « *bulle spéculative* ». Les records successifs observés sur le marché des transferts, culminant avec certains méga-transferts, ont renforcé cette lecture (Bastien Drut, 2019).

En théorie, parler de bulle spéculative suppose qu'il soit possible d'identifier un écart durable entre le prix observé d'un actif et sa valeur fondamentale. Autrement dit, une telle qualification nécessite l'existence d'un modèle permettant d'estimer ce prix fondamental. Or, appliquée au marché des transferts, cette démarche soulève immédiatement une difficulté majeure : comment déterminer la valeur fondamentale d'un joueur ? Dès lors, l'identification d'un prix "juste" constitue déjà un premier obstacle à la démonstration d'une bulle (Bastien Drut, 2019).

Parallèlement, le football professionnel est entré dans l'ère du big data : des volumes croissants de données sont désormais disponibles, les clubs s'appuient de plus en plus sur des entreprises spécialisées comme Amisco² et Opta³ pour analyser les matchs, les joueurs et leurs déplacements (Simon Kuper & Stefan Szymanski, 2016). La littérature récente confirme d'ailleurs que l'augmentation de la disponibilité des données et la montée en puissance du football analytics⁴ reconfigurent les pratiques de recrutement et d'évaluation des joueurs

¹ Le marché des transferts, ou mercato, désigne la période et le mécanisme par lesquels les clubs professionnels (notamment de football) échangent, vendent ou prêtent des joueurs. Il s'agit d'un marché saisonnier structuré, où s'échangent des talents contre des indemnités et de nouveaux contrats.

² Amisco est un acteur/système de tracking et d'analyse de la performance sportive (Amisco, s. d.).

³ Opta est une marque de données sportives, elle fournit des données collectées en temps réel par des analystes humains, enrichies par la vision par ordinateur et l'intelligence artificielle, afin de produire des statistiques, métriques et données de performance (« Opta Data from Stats Perform », s. d.).

⁴ Le football analytics correspond à l'usage de méthodes statistiques et d'outils numériques pour collecter,

(Franceschi et al., 2024). C'est dans ce cadre que s'inscrit la présente recherche. Face à la difficulté de déterminer la valeur fondamentale d'un joueur, l'objectif est d'étudier dans quelle mesure l'usage des données de performance peut contribuer à en améliorer l'estimation.

1.2. Question de recherche

Dans le prolongement de la problématique exposée ci-dessus, ce mémoire s'articule autour de la question de recherche suivante : « *Dans quelle mesure les statistiques avancées observables avant un transfert permettent-elles d'estimer la valeur de performance d'un joueur évoluant dans les championnats européens disposant de données de performance avancées, et d'en déduire l'écart avec le prix de transfert effectivement observé ?* »

Cette question suppose tout d'abord de préciser ce que l'on entend par performances observables et la manière dont celles-ci peuvent être mesurées dans le football professionnel. L'enjeu n'est pas de se limiter aux statistiques les plus visibles, comme les buts ou les passes décisives, mais d'intégrer des indicateurs plus riches susceptibles de mieux refléter la contribution réelle du joueur dans les différentes phases du jeu. Dans cette perspective, une attention particulière sera accordée aux statistiques avancées, entendues comme des mesures permettant d'appréhender non seulement la production offensive directe, mais aussi la participation à la création d'occasions ou à la réduction des opportunités adverses.

La question posée implique également de justifier plusieurs choix de cadrage. D'une part, l'analyse se concentre sur les performances observées avant le transfert, afin de ne retenir que les informations disponibles au moment où le prix est fixé. D'autre part, elle suppose de distinguer clairement trois grandeurs dont la nature diffère sensiblement. Le prix de transfert correspond à un montant effectivement observé lors d'une transaction. La valeur marchande relève quant à elle d'une estimation produite par des plateformes spécialisées, davantage proche d'un jugement collectif que d'une mesure objective. Une troisième grandeur, centrale dans ce travail, doit enfin être introduite : la valeur de performance. Rappelons que qualifier le marché des transferts de bulle spéculative suppose, comme évoqué précédemment, de pouvoir identifier un écart durable entre le prix observé d'un joueur et sa valeur fondamentale. Ce qui implique nécessairement l'existence de cette dernière (Bastien Drut, 2019). C'est cette valeur fondamentale que ce travail désigne sous le terme de valeur de performance, entendue

analyser et interpréter des données liées au football (PFSA, 2025).

comme la traduction monétaire de la contribution sportive d'un joueur au cours d'une saison, telle qu'elle peut être appréhendée à partir de ses statistiques avancées. Contrairement aux deux grandeurs précédentes, cette valeur n'est observable nulle part : aucune transaction ni aucune plateforme ne la rapporte directement. Dans la mesure où le prix effectivement payé pour un joueur intègre également des déterminants étrangers à sa seule performance sportive.

Ce mémoire ne prétend donc pas mesurer cette valeur de performance de façon exacte mais chercher à s'en approcher au mieux, à partir des informations disponibles. L'indemnité de transfert réellement observée d'une part, les statistiques avancées de performance d'autre part. Enfin, le choix de circonscrire l'analyse aux cinq championnats européens majeurs (Premier League, La Liga, Bundesliga, Serie A et Ligue 1) s'explique par une double exigence : ces championnats sont les seuls à disposer d'une couverture complète en statistiques avancées, condition indispensable à la mesure de la performance telle que définie ci-dessus, et le volume de transferts qu'ils réunissent conjointement est nécessaire à la robustesse des analyses statistiques mobilisées dans ce travail.

2. Revue de littérature et cadre conceptuel

Dans le monde du football, les indemnités de transfert (prix de transfert) sont définies comme la somme d'argent versée par un club acquéreur à un club vendeur pour racheter le contrat d'un joueur professionnel avant son terme (l'indemnité de transfert inclut tous les paiements liés à la transaction : montants fixes, variables ou conditionnels) (Infantino, s. d.). De plus le système des transferts a connu une révolution majeure avec l'arrêt Bosman⁵, une décision importante du droit européen du sport qui a fluidifié le marché des transferts. En effet, il a permis de supprimer les quotas de joueurs européens ainsi que les indemnités de transfert pour les joueurs en fin de contrat⁶.

⁵ Décision de la Cour de justice des Communautés européennes du 15 décembre 1995 ayant jugé contraires à la libre circulation des travailleurs certaines règles imposant le paiement d'une indemnité à l'ancien club lorsqu'un joueur ressortissant de l'Union européenne arrivait en fin de contrat et voulait rejoindre un club d'un autre État membre (Luca, s. d.).

⁶ Joueur qui n'est plus lié contractuellement à son ancien club.

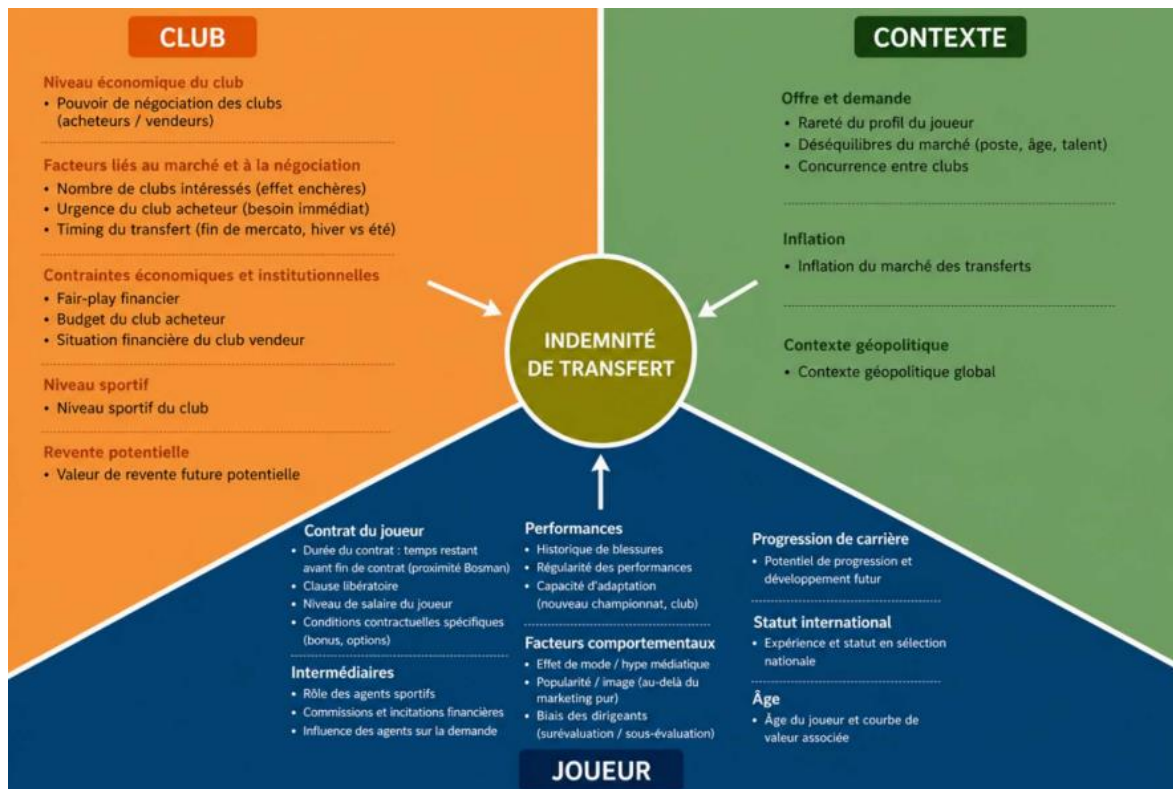


Figure 1 : Facteurs influençant l'indemnité de transfert d'un joueur de football (Poli et al., s. d.)

Ensuite, il convient de souligner que l'indemnité de transfert résulte d'un ensemble complexe de déterminants. La figure ci-dessus (Poli et al., s. d.), élaborée principalement à partir du rapport n°53 de l'Observatoire du football du CIES et complétée par la littérature académique existante, propose une synthèse structurée de ces facteurs. Elle met en évidence que l'indemnité de transfert peut être appréhendée comme une fonction de trois grandes catégories de variables : le contexte, le club et le joueur, chacune regroupant plusieurs dimensions spécifiques (Bastien Drut, 2019; Poli et al., s. d.).

Ainsi, on peut formuler de manière simplifiée :

$$\text{Indemnité de transfert} = f(x_{\text{Joueur}}, y_{\text{Club}}, z_{\text{Contexte}})$$

1. Club :

La catégorie « club » regroupe l'ensemble des variables liées aux caractéristiques économiques, financières et sportives des clubs impliqués dans la transaction. On retrouve dans cette catégorie :

- Le niveau économique du club, qui influence directement sa capacité de négociation et son pouvoir d'achat

- Les facteurs liés au marché et à la négociation, tels que le nombre de clubs intéressés (effet d'enchère), l'urgence du club acheteur ou encore le timing du transfert
- Les contraintes économiques et institutionnelles, incluant le fair-play financier⁷, le budget disponible ou encore la situation financière du club vendeur
- Le niveau sportif du club, qui peut influencer l'attractivité du projet proposé au joueur
- La valeur de revente potentielle, intégrant une logique d'investissement et d'anticipation

2. Contexte :

La catégorie « contexte » regroupe les variables macroéconomiques⁸ et structurelles⁹ qui influencent le marché des transferts dans son ensemble. Cette catégorie comprend :

- Les mécanismes d'offre et de demande, notamment la rareté du profil du joueur ou les déséquilibres structurels du marché¹⁰
- La concurrence entre clubs, qui peut générer des effets inflationnistes
- L'inflation du marché des transferts, liée à l'augmentation globale des revenus du football
- Le contexte géopolitique et économique global, susceptible d'affecter les flux financiers et les stratégies d'investissement

Ces deux premières catégories traduisent le fait que l'indemnité de transfert ne repose pas uniquement sur la valeur intrinsèque du joueur mais également sur la capacité et les contraintes des clubs à intervenir sur le marché.

⁷ Le fair-play financier désigne l'ensemble des règles mises en place par l'UEFA afin d'encadrer la gestion financière des clubs, de limiter les dépenses excessives et de favoriser leur viabilité économique.

⁸ Les variables macroéconomiques sont des indicateurs qui décrivent l'environnement économique global : revenus du secteur, inflation, taux de change, croissance économique, conditions de financement ou évolution des droits télévisés.

⁹ Les variables structurelles désignent des facteurs durables qui caractérisent l'organisation profonde du marché : répartition inégale des revenus entre clubs, concentration du pouvoir d'achat, rareté de certains profils de joueurs, poids des championnats dominants ou règles institutionnelles.

¹⁰ Déséquilibres durables qui affectent le fonctionnement d'un marché : concentration des ressources financières, inégalités entre acteurs, rareté de certains profils, asymétrie du pouvoir de négociation ou écarts persistants entre l'offre et la demande.

3. Joueur :

La catégorie « joueur » regroupe les caractéristiques individuelles qui déterminent la valeur intrinsèque du joueur. Elle inclut :

- Les performances sportives, mesurées à travers des indicateurs statistiques et analytiques.
- Le potentiel de progression et de développement futur, élément clé dans une logique d'investissement.
- Le statut international, reflétant la reconnaissance au plus haut niveau
- L'âge influence à la fois la performance actuelle et la valeur future.
- Le niveau sportif du club actuel et du championnat, servant de mesure du niveau de compétition
- Les caractéristiques contractuelles, notamment la durée restante du contrat

Ces variables sont au cœur de l'évaluation économique du joueur, puisqu'elles traduisent sa contribution sportive actuelle et future. D'ailleurs, la littérature montre que la performance et les compétences des joueurs constituent un déterminant central de leur valeur marchande, tandis que les modèles empiriques confirment l'importance de variables telles que l'âge, les performances ou l'expérience dans la formation des prix (Carmichael et al., 2011; Franceschi et al., 2024).

Il convient à présent de préciser l'ancrage théorique des grandeurs distinguées précédemment (voir chapitre 1.2. Question de recherche). En particulier celui de la valeur de performance désignée ci-après valeur intrinsèque dont la définition formelle est développée dans les paragraphes qui suivent.

La valeur marchande, pour sa part, correspond à une estimation monétaire de la valeur d'un joueur sur le marché des transferts, généralement produite par des acteurs externes (experts, plateformes spécialisées ou médias) et vise à refléter une valeur théorique dans un marché libre sans nécessairement coïncider avec l'indemnité de transfert effectivement observée (Franceschi et al., 2024).

Par analogie avec la théorie financière, nous avons la valeur fondamentale d'un actif, laquelle correspond à sa capacité à générer des performances futures. Dans le cas des joueurs de

football, cette capacité s'exprime à travers leur capital humain, incluant leurs performances sportives, leur potentiel de progression et leurs compétences individuelles. Dans le cadre de ce travail, la notion de valeur fondamentale est ainsi assimilée à une valeur intrinsèque, c'est-à-dire une valeur fondée exclusivement sur les caractéristiques propres du joueur, indépendamment des mécanismes de marché tels que l'offre et la demande ou les dynamiques de négociation. Cette approche vise à isoler la contribution réelle du joueur à sa valeur, en se concentrant sur des déterminants observables et mesurables tels que ses performances sportives, son âge ou encore son potentiel de progression (Carmichael et al., 2011).

Il convient néanmoins de noter que cette approximation reste imparfaite, dans la mesure où certaines dimensions de la valeur d'un joueur, notamment liées à son image ou à son attractivité commerciale, peuvent également contribuer à sa valorisation économique. Bien que ces éléments soient en partie liés aux caractéristiques propres du joueur, ils dépendent également de facteurs externes tels que le contexte médiatique ou le championnat dans lequel il évolue, ce qui justifie leur exclusion du périmètre d'analyse retenu dans ce travail. C'est en raison de tout cela que ma valeur de performance est assimilée à la valeur fondamentale d'un actif (non égale à celle-ci). En effet, comme indiqué précédemment, la valeur fondamentale d'un actif constitue une représentation des performances actuelles du capital humain ainsi que des revenus futurs et des perspectives de croissance probables. À l'inverse, ma valeur fondamentale ne constitue qu'une représentation monétaire des performances effectivement réalisées, sans reposer sur une spéculation quant aux performances futures.

2.2. Statistiques avancées dans le football

Les statistiques avancées désignent des indicateurs quantitatifs construits à partir d'une analyse approfondie des données, qui utilisent des méthodes statistiques, informatiques ou probabilistes pour évaluer une performance de manière plus précise, contextualisée et interprétable que les statistiques descriptives classiques. Cette idée rejoint la définition de l'analyse sportive proposée par Baumer, Matthews et Nguyen, selon laquelle l'analyse des données dans le sport vise à améliorer la compréhension et la performance athlétique en combinant données, modèles, outils statistiques et connaissance du sport (B. S. Baumer et al., 2023).

Dans cette logique, les statistiques avancées ne se limitent pas à compter ce qui s'est passé. Elles cherchent plutôt à évaluer la qualité, le contexte et la probabilité d'un événement sportif. Autrement dit, elles ne décrivent pas seulement le volume d'actions réalisées, mais tentent d'en mesurer la valeur réelle en fonction des conditions dans lesquelles elles se produisent. En analyse sportive, ces modèles servent justement à construire des indicateurs capables de capturer les compétences, les capacités et les tendances des joueurs ou des équipes, afin de produire une interprétation plus riche que les simples statistiques de comptage (Jesse et al., 2024).

Dans le monde du football, par exemple, il a été démontré que les modèles d'expected goals offrent une lecture plus détaillée de la performance que le simple score final, notamment parce qu'ils permettent d'évaluer la qualité des occasions plutôt que de se limiter au résultat brut (Borghesi et al., 2023).

Cependant, ces statistiques présentent aussi des limites. Leur fiabilité dépend de la qualité des données collectées, du choix du modèle utilisé et de la manière dont les indicateurs sont interprétés. De plus les données sportives peuvent contenir des biais, des erreurs ou des variations de définition selon les fournisseurs de données, ce qui peut influencer les résultats produits par les modèles. Ensuite, certains facteurs humains comme la fatigue, l'état physique ou la variabilité individuelle restent difficiles à intégrer parfaitement dans un indicateur chiffré. Halson montre notamment que la fatigue est un phénomène complexe qui nécessite un suivi spécifique pour être correctement comprise dans le contexte de la performance sportive (Halsen, 2014; Jesse et al., 2024).

Ces limites peuvent être regroupées en 4 catégories :

- Limites liées aux données : erreurs de collecte, définitions différentes selon les fournisseurs.
- Limites liées au contexte : rôle tactique, style de jeu, niveau de l'équipe, championnat.
- Limites liées au modèle : choix des variables, pondération, risque de surinterprétation.
- Limites liées à l'humain : fatigue, blessures, adaptation, mental, confiance.

2.2.1. Typologie et calcul des statistiques avancées mobilisées dans l'étude

Cette section vise à présenter les statistiques avancées susceptibles d'être mobilisées dans l'étude, en les classant selon leur typologie. Chaque indicateur sera rattaché à une dimension spécifique de la performance sportive : production offensive, création, progression du ballon, défense, duels, sécurité technique, pressing ou performance des gardiens.

Les tableaux en annexe synthétisent ces différentes familles d'indicateurs, précise la signification des principales statistiques retenues et met en évidence leur intérêt pour l'analyse de la performance individuelle. Cette étape constitue un préalable nécessaire à la construction d'une valeur de performance, dans la mesure où elle permet d'identifier les variables sportives qui pourront ensuite être utilisées dans l'analyse statistique (voir Annexe 2 : Pool initial de 63 variables candidates (issu de FBref)).

2.3. Modèles quantitatifs d'estimation des prix de transfert

La littérature récente consacrée à la valorisation économique des footballeurs professionnels s'est progressivement structurée autour de modèles quantitatifs visant à expliquer ou à prédire les montants associés aux joueurs. Les premiers travaux mobilisaient principalement des approches économétriques classiques, notamment des régressions linéaires, dans lesquelles le prix de transfert était expliqué à partir de variables telles que l'âge, le poste, l'expérience, le temps de jeu, le statut international, la durée contractuelle restante ou encore le niveau du club vendeur et acheteur. Plus récemment, ces approches ont été complétées par des méthodes d'apprentissage supervisé¹¹, telles le Random Forest¹², les modèles de gradient boosting¹³ et bien d'autres. Cette évolution méthodologique traduit la volonté de mieux capter la complexité d'un marché dans lequel les relations entre les caractéristiques d'un joueur et son prix ne sont pas nécessairement linéaires. McHale et Holmes soulignent d'ailleurs que les modèles de machine learning permettent d'améliorer la précision prédictive par rapport aux régressions linéaires traditionnelles.

¹¹ Définition présente au chapitre 3.5.2. L'apprentissage supervisé

¹² Définition présente au chapitre 3.5.3. Les trois modèles comparés

3.5.2. L'apprentissage supervisé

¹³ Méthode d'ensemble qui construit les modèles de manière séquentielle : chaque nouvel arbre cherche à corriger les erreurs des arbres précédents.

Ces modèles ne reposent toutefois pas tous sur la même conception de la performance. Certains travaux se limitent à des indicateurs relativement simples, tels que les buts, les passes décisives, les minutes jouées ou les apparitions internationales. C'est notamment le cas de modèles fondés sur la régression linéaire multiple, la Ridge Regression et le Support Vector Regression (Tamim et al., 2025). D'autres approches mobilisent également des données de type SoFIFA¹⁴ pour estimer la valeur marchande des joueurs, par exemple à travers une méthode en deux phases combinant un clustering automatique et une SVR optimisée par Particle Swarm Optimization (Behravan & Razavi, 2021) (voir Annexe 6 : Les méthodes rencontrées).

À l'inverse, certains travaux plus récents cherchent à intégrer des métriques de performance plus avancées afin de mieux capter la contribution sportive réelle du joueur. McHale et Holmes proposent ainsi un modèle d'estimation des indemnités de transfert reposant sur des métriques avancées de performance et des algorithmes de machine learning, notamment XGBoost, en montrant que ces variables améliorent la précision prédictive par rapport à une régression linéaire classique (McHale & Holmes, 2023).

D'autres intègrent des métriques plus avancées, notamment les expected goals, les expected assists ou des mesures plus globales de contribution à la performance collective. Les indicateurs classiques peuvent conduire à une vision partielle de la valeur sportive d'un joueur, particulièrement pour les postes dont la contribution ne se traduit pas directement par des buts ou des passes décisives. McHale et Holmes insistent précisément sur cette limite : une partie importante de la littérature utilise encore des statistiques rudimentaires ou des notations subjectives, alors que les métriques avancées permettent de mieux appréhender la qualité réelle du joueur et d'améliorer la prédiction des indemnités de transfert.

Cette distinction est importante, car les modèles fondés sur des indicateurs généraux ou agrégés risquent de mesurer davantage la valeur perçue d'un joueur que sa contribution sportive spécifique. Franceschi et al. rappellent d'ailleurs que la littérature distingue deux objets empiriques : les indemnités de transfert effectivement payées et les valeurs de marché

¹⁴ Base de données en ligne qui recense les joueurs, équipes et attributs issus des jeux EA Sports FC/FIFA. Elle fournit notamment des informations sur l'âge, le poste, le club, la note générale (overall rating), le potentiel, les caractéristiques techniques, physiques et mentales des joueurs.

proposées par des plateformes, qui reposent davantage sur une estimation collective ou médiatique (Franceschi et al., 2024).

Toutefois, ces modèles restent majoritairement orientés vers la prédiction d'un prix ou d'une valeur de marché. Leur objectif est donc de reproduire les mécanismes de valorisation observés sur le marché, et non d'isoler strictement la contribution sportive du joueur. C'est précisément sur ce point que se situe la contribution de la présente recherche : construire une valeur de performance fondée sur les statistiques avancées observables, en limitant autant que possible l'influence des facteurs extra-sportifs tels que la médiatisation, la popularité, le pouvoir de négociation du club acheteur ou la dimension contractuelle. Une telle approche ne prétend donc pas remplacer les modèles d'estimation du marché mais les compléter en proposant un indicateur de référence permettant de comparer la valeur sportive estimée d'un joueur à sa valeur marchande ou au prix effectivement payé.

3. Méthodologie

Ce chapitre présente la démarche méthodologique complète mise en œuvre pour répondre à la question de recherche. Il suit l'ordre chronologique du travail empirique : la collecte des données, leur nettoyage et leur préparation, la réduction de dimension par analyse en composantes principales, la décomposition de l'importance des facteurs par la méthode Shapley-LMG, puis la construction de la valeur de performance par apprentissage supervisé.

3.1. Collecte des données

La constitution de la base de données constitue la première étape empirique de ce travail. Elle vise à réunir, pour un ensemble de transferts de joueurs professionnels, deux types d'informations complémentaires : d'une part, les statistiques de performance sportive observées avant le transfert et d'autre part, le prix effectivement payé lors de la transaction.

3.1.1. Nature et sources des données

Les données mobilisées dans ce travail sont exclusivement des données quantitatives secondaires¹⁵. Deux catégories de données ont été réunies, provenant de deux sources

¹⁵ Des données déjà collectées et mises à disposition par des organismes tiers, par opposition à des données primaires qui auraient été collectées spécifiquement pour cette recherche.

distinctes, aucune plateforme ne réunissant à elle seule l'ensemble des informations nécessaires.

Les statistiques avancées de performance proviennent de *FBref*, une plateforme de football analytics qui met à disposition des indicateurs détaillés couvrant les différentes dimensions de la performance : production offensive, création d'occasions, progression du ballon, défense, duels, sécurité technique, pressing et performance des gardiens (voir Annexe 1 : Typologie des données mobilisées). Ces statistiques sont mesurées sur la saison précédant le transfert¹⁶, conformément à la question de recherche qui porte sur les performances *observables avant* la transaction : seules les informations disponibles au moment où le prix est fixé sont retenues. Quant aux indemnités de transfert, elles sont issues de *Transfermarkt*¹⁷.

3.1.2. Périmètre géographique et temporel

Le périmètre retenu couvre les cinq championnats majeurs européens (Premier League, La Liga, Bundesliga, Serie A, Ligue 1), généralement désignés sous l'appellation « Big 5 ». Ce choix répond à une double exigence. D'une part, ces championnats sont les seuls à disposer d'une couverture complète en statistiques avancées sur *FBref*, condition indispensable à la mesure de la performance. D'autre part, le volume d'observations nécessaire à la validité de l'analyse en composantes principales, en particulier le ratio requis entre le nombre d'observations et le nombre de variables candidates (voir 3.3.3. Conditions d'utilisation et validation). Un championnat isolé n'aurait pas fourni un nombre suffisant de transferts disposant simultanément d'une indemnité réellement payée et connue et de statistiques avancées complètes. Sur le plan temporel, la couverture *FBref* en statistiques avancées ne débutant véritablement qu'à partir de la saison 2018-2019, l'échantillon couvre les saisons 2018-2019 à 2023-2024.

¹⁶ Les statistiques mobilisées portent exclusivement sur la saison précédant le transfert. Ce choix répond à un principe de cohérence informationnelle : seules les données disponibles au moment où le prix est fixé sont retenues, conformément à la démarche des études empiriques du champ (Franceschi et al., 2024). Les performances de la saison écoulée constituent en outre le signal d'information le plus récent et le plus saillant pour les clubs acquéreurs

¹⁷ Transfermarkt est la base de données de référence sur le marché des transferts de football : elle recense, pour l'ensemble des championnats professionnels, transferts définitifs, prêts, arrivées libres, les montants des transactions rapportés par la presse et les communications officielles des clubs.

3.1.3. Choix de la variable cible : indemnité réelle plutôt que valeur marchande

Le choix de l'indemnité de transfert effectivement payée comme variable cible découle directement du constat posé précédemment : la valeur de performance (grandeur centrale de ce travail) n'est ni observable ni documentée. Cette situation impose le recours à un proxy¹⁸. Dans le cas présent, la seule grandeur monétaire objectivement observée qui reflète la valeur accordée à un joueur est l'indemnité de transfert effectivement payée : c'est donc elle qui est retenue comme variable cible, en tant que proxy de la valeur du joueur sur laquelle calibrer la relation entre performance et prix.

Une précision s'impose toutefois, déjà établie dans la revue de littérature : l'indemnité de transfert incorpore de nombreux éléments subjectifs et contextuels (pouvoir de négociation, effet d'enchère, médiatisation, situation contractuelle). C'est la raison pour laquelle l'estimation produite par ce travail doit elle-même être lue comme un proxy de la valeur de performance, et non comme une mesure exacte.

3.1.4. Jonction des sources et périmètre des transactions retenues

FBref et Transfermarkt reposant sur des identifiants de joueurs différents, une table de correspondance a été construite pour associer les deux sources, principalement sur la base du nom du joueur, du club et de la saison. Cette jonction s'est heurtée à des divergences d'orthographe entre les deux plateformes (accents, caractères spéciaux, transcriptions différentes selon les conventions linguistiques). Les transferts pour lesquels la correspondance n'a pu être établie avec certitude ont été exclus plutôt que résolus arbitrairement, afin de ne pas introduire d'appariements erronés ; l'impact de ces exclusions est vérifié statistiquement au chapitre 3.2.4. Vérification statistique du biais d'exclusion.

Par ailleurs, les transferts sous forme de prêt ont été exclus du périmètre, dans la mesure où ils ne constituent pas des transactions définitives comparables, sur le plan de la valorisation économique, à un transfert à titre définitif.

¹⁸ Un proxy (ou variable de substitution) désigne une variable observable utilisée en remplacement d'une grandeur théorique qui ne peut être mesurée directement, à condition d'entretenir avec celle-ci une relation suffisamment étroite pour en constituer une approximation acceptable.

3.1.5. Échantillon final

Au terme de cette étape, l'échantillon réunit 1 286 transferts pour lesquels l'indemnité réellement payée est connue et une couverture en statistiques avancées est disponible. Parmi eux, 1 172 constituent des cas complets, c'est-à-dire des observations disposant d'une valeur pour l'ensemble des variables candidates de leur poste, condition requise par l'analyse en composantes principales. Les 114 transferts restants présentent au moins une valeur manquante et en sont exclus. Ces valeurs manquantes proviennent de trois mécanismes principaux :

- Les statistiques exprimées par 90 minutes sont neutralisées lorsque le temps de jeu du joueur est inférieur au seuil minimal (voir 3.2.2. Neutralisation des statistiques sur temps de jeu très faible) afin d'éviter les extrapolations aberrantes
- Certains pourcentages sont indéfinis faute d'événement à mesurer
- Un gardien n'ayant affronté aucun penalty sur la saison n'a pas de taux d'arrêt de penalty, de même qu'un joueur sans aucun dribble tenté n'a pas de taux de réussite au dribble.

La classification¹⁹ par poste retenue distingue sept rôles : gardien, défenseur central, latéral, milieu défensif, milieu offensif, ailier et buteur. Le Tableau 1 présente la répartition des cas complets par poste, telle qu'utilisée dans les analyses.

Poste	Nombre d'observations (n)
Milieu défensif	275
Buteur	227
Défenseur central	203
Ailier	166
Latéral	161
Milieu offensif	72
Gardien	68
Effectif total	1172

Tableau 1: Répartition de l'échantillon final par poste (cas complets mobilisés dans l'ACP)

¹⁹ Classification par défaut de FBref

Les effectifs disponibles sont hétérogènes : les postes de gardien et de milieu offensif disposent des échantillons les plus réduits (respectivement 68 et 72 observations). Cette hétérogénéité est inhérente à la structure même d'un effectif de football, une équipe n'aligne qu'un seul gardien titulaire contre dix joueurs de champ et cela a des implications directes sur la validité de l'analyse en composantes principales pour ces deux postes, traitées à la section 3.3.3. Conditions d'utilisation et validation.

3.2. Nettoyage et préparation des données

Les données brutes issues de la collecte ne peuvent être utilisées telles quelles : elles contiennent des incohérences, des redondances et des cas particuliers susceptibles de biaiser les analyses ultérieures s'ils ne sont pas traités explicitement. Cette section documente les principales opérations de nettoyage effectuées, les problèmes rencontrés et la manière dont ils ont été résolus.

3.2.1. Traitement des prix nuls et des prix manquants

Une distinction essentielle a été opérée entre deux situations que les données brutes ne différencient pas toujours : les transferts à prix nul pour raison structurelle et les transferts à prix manquant. Un joueur arrivé en fin de contrat peut être transféré sans indemnité²⁰ : il s'agit d'un vrai zéro économique, reflet d'une réalité juridique. À l'inverse, un montant absent peut simplement signifier que l'indemnité n'a pas été divulguée publiquement : il s'agit alors d'une donnée manquante, sans signification économique. Traiter ces deux cas de manière identique aurait biaisé l'échantillon, seuls les transferts à indemnité strictement positive non nulle et effectivement rapportée ont été conservés.

3.2.2. Neutralisation des statistiques sur temps de jeu très faible

La majorité des statistiques avancées sont exprimées en taux par 90 minutes (per-90), afin de rendre comparables des joueurs aux temps de jeu différents. Cette normalisation présente toutefois un effet négatif bien identifié : pour un joueur n'ayant disputé que quelques minutes, quelques actions isolées s'extrapolent en taux par 90 minutes aberrants. Un seuil minimal²¹ de temps de jeu a donc été appliqué pour neutraliser ces cas. Ce traitement s'est révélé décisif : un cas limite (un joueur ayant disputé 92 minutes sur une saison, juste au-dessus du seuil) a

²⁰ Transfert « libre », conséquence de l'arrêt Bosman évoqué dans la revue de littérature

²¹ Le seuil de 90 minutes correspond à la durée d'un match complet, soit l'unité de normalisation sur laquelle reposent les statistiques per-90 de FBref.

produit une valeur extrême sur un facteur de l'ACP (8,8 écarts-types au-dessus de la moyenne du poste), illustrant la sensibilité des statistiques per-90 aux temps de jeu très faibles. Ce cas particulier et ses conséquences sur la modélisation sont discutés en 3.5.5.

3.2.3. Standardisation

Les statistiques brutes ne sont pas directement comparables d'un championnat à l'autre ni d'une saison à l'autre : les styles de jeu, les niveaux de compétition et les conditions de mesure varient. Chaque variable a donc été standardisée en score z^{22} (z-score) au sein de chaque combinaison entre championnat et saison. Cette standardisation locale permet de comparer la performance relative de joueurs évoluant dans des contextes différents, condition nécessaire dans notre cas.

La validité de cette standardisation suppose un nombre suffisant de joueurs dans chaque cellule (championnat $\times$ saison) pour que la moyenne et l'écart-type soient estimés de manière stable. Cette condition a été vérifiée avant l'application du traitement.

3.2.4. Vérification statistique du biais d'exclusion

Le processus de jonction entre sources (voir 3.1.4. Jonction des sources et périmètre des transactions retenues) a conduit à l'exclusion de 167 transferts uniques, soit un taux d'exclusion global de 4,9 % sur l'ensemble des transferts candidats (environ 3 400 transactions). Cette exclusion n'étant pas un choix mais la conséquence subie d'un problème d'appariement entre sources, il importe de vérifier qu'elle n'a pas introduit de biais de sélection dans l'échantillon final : si les conflits de correspondance touchaient systématiquement un certain profil de joueurs (un championnat, un poste, une nationalité), l'échantillon final s'en trouverait déformé sans que ce déséquilibre ait été choisi ni même perçu.

L'outil mobilisé pour cette vérification est le test du χ^2 , méthode statistique bivariée destinée à étudier la relation entre deux variables catégorielles (Fouss & Satinet, s. d.). Le statut d'exclusion (exclu / conservé) et chacune des trois variables considérées (poste, championnat,

²² Pour chaque joueur, on soustrait la moyenne de la variable calculée sur son championnat et sa saison, puis on divise par l'écart-type correspondant. Un score z de 0 signifie ainsi que le joueur se situe exactement dans la moyenne de son championnat cette saison-là. Un score de +1 qu'il se situe un écart-type au-dessus.

nationalité) étant toutes deux catégorielles au sens de cette typologie, le χ^2 constitue l'outil approprié.

Le test repose sur la construction d'un tableau croisant les deux variables catégorielles (par exemple, championnat en lignes, statut d'exclusion en colonnes), puis sur la comparaison, case par case, entre les effectifs observés et les effectifs qui seraient attendus si les deux variables étaient indépendantes, ces derniers se déduisant simplement du taux d'exclusion global appliqué uniformément à chaque catégorie. Pour chaque case, l'écart entre effectif observé et effectif attendu est élevé au carré puis rapporté à l'effectif attendu, ce qui neutralise le signe de l'écart et pondère les écarts relatifs plutôt qu'absolus. La somme de ces termes sur l'ensemble du tableau constitue la statistique du test. Plus cette statistique est élevée, plus la répartition observée s'écarte de ce que produirait le hasard sous l'hypothèse d'indépendance.

Pour chacune des trois dimensions²³, les hypothèses suivantes ont été posées :

- Hypothèse nulle (H0) : le statut d'exclusion est indépendant de la dimension considérée, les conflits de correspondance frappent chaque catégorie dans les mêmes proportions, aux fluctuations d'échantillonnage près.
- Hypothèse alternative (H1) : le statut d'exclusion dépend de la dimension considérée, certaines catégories sont significativement plus ou moins touchées que d'autres.

La décision de rejeter ou non H0 repose sur la p-valeur associée à la statistique du test, c'est-à-dire la probabilité d'observer. Conformément au seuil conventionnel de 5% (Ducarroz & Sinigaglia, s. d.), H0 est rejetée au profit de H1 lorsque la p-valeur est inférieure à 0,05. Au-delà de ce seuil, l'hypothèse d'indépendance n'est pas rejetée, sans pour autant être positivement démontrée. Les résultats obtenus pour les trois dimensions sont présentés dans le Tableau 13.

La lecture combinée de ces trois résultats ne permet pas de conclure à une exclusion strictement aléatoire, le test sur le poste rejette formellement l'indépendance mais elle écarte l'hypothèse d'un biais de sélection majeur : l'unique effet clairement significatif reste d'ampleur modérée, l'effet sur le championnat demeure à la limite du seuil conventionnel, et l'effet apparent sur la nationalité ne résiste pas au test une fois la variable correctement isolée.

²³ Poste, championnat et nationalité

L'exclusion s'explique ainsi essentiellement par un mécanisme technique. Les conventions orthographiques propres à certaines langues plutôt que par une caractéristique sportive ou économique des joueurs concernés.

3.2.5. Sélection des variables candidates et traitement des redondances

L'ensemble initial de variables candidates pour l'analyse factorielle comptait une soixantaine de statistiques avancées par poste. Deux traitements complémentaires ont été appliqués pour aboutir à la sélection finale.

Élimination des redondances :

Certaines variables mesurent quasiment le même phénomène sous des libellés différents (par exemple, deux décomptes voisins de passes progressives). Conserver ces paires quasi identiques gonfle artificiellement le nombre de dimensions détectées par l'analyse factorielle, sans apport informatif réel. Un premier filtrage au seuil de corrélation de 0,95 s'est révélé insuffisant : un contrôle ultérieur a identifié plusieurs dizaines de paires de variables encore corrélées au-delà de 0,90. Le seuil a donc été resserré à 0,90 : pour chaque paire dépassant ce seuil, une seule variable a été conservée, la plus interprétable ou la plus utilisée dans la littérature.

Exclusion des variables de contexte :

Certaines variables candidates, bien que disponibles dans les données de performance, reflètent davantage des décisions d'entraîneur ou un contexte d'équipe que la performance intrinsèque du joueur : c'est le cas du nombre de matchs joués (qui dépend de la politique de rotation du club, de la confiance de l'entraîneur et des blessures (voir Annexe 2 : Pool initial de 63 variables candidates (issu de FBref))). Ces variables ont été exclues du périmètre analysé, selon le même principe que l'exclusion de l'âge justifiée dans la revue de littérature : une caractéristique peut être rattachée au joueur sans pour autant constituer une mesure de sa performance observable sur le terrain.

3.3. Analyse en composantes principales (ACP)

3.3.1. Principe de la méthode

L'analyse en composantes principales est une méthode statistique descriptive de réduction de dimension. Son objectif est de résumer un grand nombre de variables quantitatives corrélées

entre elles en un petit nombre de variables synthétiques (les composantes principales) qui conservent l'essentiel de l'information contenue dans les données d'origine. Chaque composante est une combinaison linéaire des variables d'origine, construite de manière à capter la plus grande part possible de la variance totale des données (Ducarroz & Sinigaglia, s. d.).

3.3.2. Pertinence de l'ACP

Le recours à l'ACP répond à deux nécessités dans ce travail. La première est la réduction de dimension proprement dite : avec une soixantaine de statistiques avancées candidates par poste, une modélisation directe serait à la fois instable (trop de variables pour le nombre d'observations disponibles) et illisible (impossibilité d'interpréter la contribution individuelle de dizaines de variables corrélées). La seconde est le traitement de la colinéarité : les statistiques de performance sont fortement corrélées entre elles par construction (un joueur qui touche beaucoup de ballons cumule mécaniquement passes, courses et duels), ce qui rendrait les coefficients d'une régression directe ininterprétables.

L'ACP a été réalisée séparément pour chacun des sept postes. Ce choix s'impose par la nature même des données : les dimensions pertinentes de la performance diffèrent fondamentalement selon les rôles (la « finition » n'a pas de sens pour un gardien ; les statistiques d'arrêts n'en ont pas pour un buteur) et une ACP unique sur l'ensemble des joueurs aurait mélangé des structures de corrélation hétérogènes (cette segmentation par poste implique que les facteurs obtenus (leur nombre, leur composition et leur signification) sont propres à chaque rôle.

3.3.3. Conditions d'utilisation et validation

L'application de l'ACP requiert la vérification de plusieurs conditions préalables.

Nature des variables : Les variables doivent être quantitatives, mesurées sur une échelle d'intervalle ou de ratio. Cette condition est satisfaite : toutes les statistiques mobilisées sont des mesures continues, standardisées en scores z (voir 3.2.3. Standardisation), ce qui garantit également qu'aucune variable ne domine l'analyse du seul fait de son échelle de mesure.

Taille de l'échantillon : Le nombre d'observations (n) doit être suffisant au regard du nombre de variables candidates (k) ; la règle usuelle recommande un ratio n/k de 3 à 5. Cette condition a joué un rôle structurant dans ce travail : un premier passage de l'ACP a révélé que deux

postes la violaient nettement (milieu offensif : $n = 72$ pour $k = 48$ variables, soit un ratio de 1,5 et gardien : $n = 68$ pour $k = 51$, soit un ratio de 1,3). Pour ces deux postes, la sélection de variables candidates a été resserrée selon une procédure en trois étapes : exclusion des variables mal représentées, déduplication des paires très corrélées, puis priorisation théorique de 2 à 3 variables par dimension du rôle. Combiné au resserrement général décrit en 3.2.5 (seuil de corrélation à 0,90 et exclusion des variables de contexte, appliqués à l'ensemble des postes), ce traitement aboutit à des ratios compris entre 4,0 et 7,1 pour les sept postes, tous conformes à la règle.

3.3.4. Choix du nombre de composantes

Une fois l'ACP réalisée, il faut déterminer combien de composantes conserver. Le critère principal retenu est celui de Kaiser : ne conserver que les composantes dont la valeur propre (eigenvalue) est supérieure à 1. La valeur propre d'une composante mesure la quantité de variance qu'elle capte, exprimée en « équivalents-variables » : une valeur propre de 1 signifie que la composante capte autant d'information qu'une seule variable d'origine. En conserver une dont la valeur propre est inférieure à 1 reviendrait donc à retenir une dimension qui résume moins d'information qu'une variable brute.

3.3.5. Résultats de l'ACP

Le Tableau 17 synthétise les résultats de l'ACP finale pour les sept postes : le nombre d'observations, le nombre de variables candidates, le ratio n/k , l'indice KMO et le nombre de composantes retenues selon le critère de Kaiser.

Trois enseignements ressortent de ce tableau. Premièrement, tous les postes respectent désormais le ratio n/k recommandé, la plupart le dépassant largement. Deuxièmement, les indices KMO se situent entre 0,73 et 0,80 pour les cinq postes aux effectifs les plus fournis (un niveau qualifié de moyen à bon) mais restent proches du seuil minimal pour le milieu offensif (0,536) et le gardien (0,625). Troisièmement, le nombre de composantes retenues varie de 5 à 10 selon les postes.

Cette variation du nombre de composantes entre postes appelle une précision méthodologique. Elle résulte de deux mécanismes distincts : d'une part, la structure de corrélation propre à chaque rôle (les dimensions de la performance d'un buteur ne se recoupent pas de la même manière que celles d'un défenseur) et d'autre part, le resserrement plus poussé du nombre de variables candidates opéré pour le milieu offensif et le gardien afin de respecter le ratio n/k . Il ne s'agit donc pas d'une incohérence, mais de la conséquence

assumée d'un traitement différencié, appliqué uniquement là où les critères de validité l'exigeaient. Les cinq autres postes respectant ces critères avec leur sélection initiale, aucun resserrement supplémentaire n'y était justifié.

Chaque composante retenue a ensuite été nommée manuellement à partir de l'examen de ses saturations (loadings), les corrélations entre la composante et les variables d'origine, un seuil de saturation de 0,3 étant retenu pour l'interprétation conformément aux recommandations du cours (Ducarroz & Sinigaglia, s. d.). À titre d'exemple, l'ACP du milieu offensif dégage cinq dimensions interprétables : finition, création, progression, pressing et engagement défensif, sécurité technique (voir Annexe 3 : Correspondance entre F (SHAP), RC (ACP) et nom de dimension). L'ensemble des composantes des sept postes a donc été nommé manuellement à partir des saturations réelles.

Enfin, les scores factoriels ²⁴constituent les variables d'entrée de toutes les analyses ultérieures : la décomposition de l'importance des facteurs (voir 3.4. Décomposition de l'importance des facteurs : la méthode Shapley-LMG) comme les modèles prédictifs (voir 3.5. Construction du proxy de la valeur de performance par apprentissage automatique supervisé).

3.4. Décomposition de l'importance des facteurs : la méthode Shapley-LMG

3.4.1. Le problème à résoudre : mesurer la contribution de chaque facteur

Une fois les dimensions de la performance identifiées par l'ACP, une question naturelle se pose : parmi ces dimensions, lesquelles contribuent le plus à expliquer le prix de transfert ? Pour y répondre, on estime d'abord une régression linéaire multiple du prix (en logarithme) sur les scores factoriels du poste. La régression linéaire multiple modélise la relation entre une variable à expliquer et plusieurs variables explicatives. La qualité globale de cette régression se mesure par le coefficient de détermination R^2 , qui exprime la proportion de la variance de la variable à expliquer que le modèle parvient à restituer.

Le R^2 étant une mesure globale, il ne dit pas comment cette variance expliquée se répartit entre les différents facteurs. Or cette répartition est utile dans le cadre de ce travail : savoir si

²⁴ La valeur de chaque joueur sur chacune des composantes retenues de son poste.

pour un buteur, c'est la finition, la création ou la progression qui « pèse » le plus dans la relation entre performance et prix. C'est le rôle de la décomposition LMG.

3.4.2. Principe de la méthode LMG

La méthode LMG répartit le R^2 total d'une régression entre les variables explicatives, en attribuant à chacune une part non négative dont la somme reconstitue exactement le R^2 (Wabgaonkar, 2022). Sa difficulté fondatrice est la suivante : lorsque l'on ajoute les variables une à une dans une régression, la contribution apparente de chacune dépend de l'ordre dans lequel on les introduit (une variable ajoutée en premier « capte » de la variance qu'une variable corrélée ajoutée ensuite ne pourra plus revendiquer). Toute décomposition fondée sur un ordre unique serait donc arbitraire.

La solution de LMG consiste à calculer la contribution de chaque variable dans tous les ordres d'introduction possibles, puis à en faire la moyenne. Ce principe est emprunté à la théorie des jeux coopératifs²⁵, où il est connu sous le nom de valeur de Shapley. Appliqué à la régression, chaque variable explicative est un « joueur », le R^2 est le « gain » collectif, et la part LMG de chaque variable est sa contribution moyenne au R^2 , calculée sur les 2^k combinaisons possibles de variables (k étant le nombre de variables). Le résultat s'exprime en pourcentage du R^2 : une part LMG de 40 % pour un facteur signifie que ce facteur explique, en moyenne sur tous les ordres d'introduction possibles, 40 % de la variance que le modèle parvient à restituer.

3.4.3. Précision sur le rôle de la méthode dans ce travail

Il convient de clarifier le rôle exact du LMG dans la démarche d'ensemble, car une approche initialement envisagée a été écartée en cours de travail. L'idée première était de dériver directement la « valeur de performance » d'un joueur à partir de la décomposition LMG. Cette approche s'est révélée inappropriée : le LMG décompose la variance déjà expliquée par un modèle donc il répond à la question « quelle part de l'explication revient à chaque facteur ? » mais ne produit aucune prédiction de prix pour un joueur donné. C'est un outil d'explication, non de prédiction. La construction de la valeur de performance a donc été confiée à des modèles prédictifs distincts (voir 3.5. Construction du proxy de la valeur de performance par

²⁵ Pour répartir équitablement le gain d'une coalition entre ses membres, on attribue à chacun sa contribution marginale moyenne sur toutes les manières possibles de former la coalition.

apprentissage automatique supervisé), le LMG conservant un rôle explicatif : identifier, pour chaque poste, les dimensions de la performance les plus associées au prix.

3.4.4. Conditions d'utilisation et diagnostics

La décomposition LMG s'applique à une régression linéaire estimée par moindres carrés ordinaires et hérite donc de ses hypothèses. Celles-ci ont été vérifiées pour chaque poste au moyen des diagnostics usuels.

- **Absence de multicollinéarité** : le facteur d'inflation de la variance (VIF) mesure dans quelle proportion la variance d'un coefficient est gonflée par la corrélation entre variables explicatives. Les scores factoriels étant orthogonaux par construction de l'ACP, le VIF vaut exactement 1 pour tous les facteurs de tous les postes (la valeur idéale, confirmant empiriquement l'un des bénéfices attendus du passage par l'ACP).
- **Normalité des résidus** : testée par le test de Shapiro-Wilk, elle est rejetée pour trois postes (défenseur central, ailier, buteur). Cette violation n'invalide pas la décomposition LMG elle-même, qui est une décomposition de variance et non un test. Elle fragilise en revanche l'inférence paramétrique classique. C'est l'une des raisons du recours au bootstrap ²⁶présenté ci-dessous.
- **Spécification du modèle** : le test RESET de Ramsey, qui détecte une éventuelle mauvaise spécification (relation non linéaire non prise en compte, variable omise), est significatif pour le seul poste de gardien ($p = 0,007$) (le plus petit échantillon de l'étude signalant que la relation linéaire y est possiblement incomplète). Ce signal converge avec l'ensemble des indicateurs de fragilité de ce poste (Ambler, s. d.).
- **Indépendance des résidus** : le test de Durbin-Watson, conçu pour détecter l'autocorrélation dans des données ordonnées dans le temps, ressort significatif pour deux postes. Ce résultat n'est toutefois pas interprétable ici, les observations (des transferts) n'ayant pas d'ordre temporel réel au sein de l'échantillon.

²⁶ Le bootstrap est une technique de simulation qui sert à mesurer la fiabilité d'un résultat statistique à savoir à quel point il varierait si on refaisait l'étude avec un échantillon légèrement différent.

3.4.5. Validation par bootstrap : quantifier l'incertitude

Les parts LMG obtenues sur l'échantillon sont des estimations : si l'on disposait d'un autre échantillon de transferts, les pourcentages obtenus seraient différents. La question cruciale est de savoir de combien ils différeraient. Autrement dit, quelle confiance accorder au classement des facteurs. Le bootstrap répond à cette question par simulation : on tire au hasard, avec remise, un grand nombre d'échantillons alternatifs à partir des données disponibles (un même joueur pouvant être tiré plusieurs fois, d'autres pas du tout), on recalcule la décomposition LMG sur chacun, et l'on observe la variabilité des résultats.

De cette distribution de résultats se déduit un intervalle de confiance à 95 % : la plage de valeurs contenant 95 % des estimations obtenues sur les 1 000 échantillons simulés. Un intervalle étroit signale un résultat stable, peu dépendant des joueurs précis composant l'échantillon. Un intervalle large signale au contraire un résultat fragile, susceptible de varier fortement d'un échantillon à l'autre. Ce recours au bootstrap présente en outre l'avantage de ne reposer sur aucune hypothèse de normalité, ce qui répond directement aux violations relevées en 3.4.4. Conditions d'utilisation et diagnostics.

3.4.6. Résultats de la décomposition LMG

Le Tableau 14 présente le facteur dominant identifié par la décomposition LMG (pour chaque poste), c'est-à-dire celui dont la part du R^2 est la plus élevée, sa part estimée, et l'intervalle de confiance bootstrap à 95 % associé.

La lecture conjointe des estimations et de leurs intervalles conduit à distinguer trois niveaux de robustesse. Pour trois postes (buteur, latéral et défenseur central), la borne inférieure de l'intervalle demeure nettement au-dessus de zéro : même dans le scénario le plus défavorable, le facteur identifié conserve une contribution substantielle et son statut de dimension dominante peut être considéré comme robuste. La cohérence footballistique de ces résultats mérite d'être notée : la capacité de finition domine chez les buteurs, la progression balle au pied chez les latéraux.

Pour deux postes (ailier et milieu défensif), les bornes inférieures (10,4 % et 5,8 %) restent proches de zéro : le facteur dominant identifié est plausible mais son avance sur les autres facteurs ne peut être affirmée avec certitude. Enfin, pour les deux postes aux effectifs les plus réduits (gardien et milieu offensif), les intervalles sont si larges (77 et 75) que la valeur

ponctuelle estimée ne saurait être interprétée isolément : le résultat spectaculaire du milieu offensif (le pressing capterait 76 % du R^2) doit être traité comme un signal instable, vraisemblablement dû au bruit d'échantillonnage, et non comme une conclusion. Un point contre-intuitif mérite d'être relevé : ces deux postes sont précisément ceux dont la sélection de variables avait été la plus resserrée ; leurs intervalles larges confirment que la limite est bien la taille d'échantillon elle-même (68 et 72 observations), et non le nombre de variables ou de facteurs.

Plus généralement, aucun poste ne présente d'intervalle véritablement étroit dans l'absolu, même le plus resserré (milieu défensif, 38,2 points) demeure large. Ce constat n'est pas un artefact méthodologique mais une caractéristique du phénomène étudié : les R^2 des régressions sous-jacentes étant modestes (voir 4. Résultats), la variance à répartir entre facteurs est faible, et sa répartition est mécaniquement sensible à la composition de l'échantillon. Cette imprécision structurelle, loin d'affaiblir le propos, anticipe le résultat central de ce travail : la performance sportive observable ne détermine qu'une part limitée du prix de transfert. Conformément à cette lecture, l'ensemble des parts LMG est systématiquement présenté avec son intervalle de confiance dans la suite de ce travail, et jamais comme une valeur ponctuelle isolée.

3.5. Construction du proxy de la valeur de performance par apprentissage automatique supervisé

3.5.1. Objectif : estimer une valeur qui n'est observable nulle part

L'objectif de cette étape n'est pas de prédire le prix de transfert (à proprement parler) celui-ci est connu pour chaque observation de l'échantillon. Il est de construire une grandeur qui n'est observable nulle part : la valeur de performance d'un joueur, c'est-à-dire le prix que ses seules statistiques sportives justifieraient, indépendamment de tout facteur de marché (médiatisation, pouvoir de négociation, contexte contractuel). Cette valeur ne pouvant être observée directement, elle est définie de manière opérationnelle : c'est la prédiction d'un modèle entraîné à relier les facteurs de performance au prix réel, mais auquel on ne donne accès qu'aux facteurs de performance. Le modèle produit ainsi la meilleure estimation du prix que le sportif seul permet de justifier et l'écart entre le prix réellement payé et cette estimation devient l'indicateur pourrait représenter tous les facteurs non liés à la performance.

Cette logique s'inscrit dans la tradition de la sabermétrie évoquée en revue de littérature : de même que l'indicateur WAR du baseball est confronté aux salaires réels pour identifier les joueurs sous-payés, la valeur de performance construite ici est confrontée aux indemnités réelles pour identifier les transferts dont le prix s'écarte le plus de ce que la performance justifie. Une différence doit toutefois être assumée : contrairement au WAR, dérivé de la mécanique interne du jeu indépendamment des salaires, la valeur de performance construite ici est calibrée sur les prix réels, faute d'une théorie qui permettrait de convertir directement des statistiques de football en euros (c'est la raison pour laquelle, on trouvera un proxy de la valeur de performance). Ses implications sont discutées au chapitre des limites (voir 5.3. Limites) (B. Baumer & Zimbalist, 2014).

3.5.2. L'apprentissage supervisé

Les modèles mobilisés relèvent de l'apprentissage automatique supervisé : l'algorithme apprend à partir d'exemples pour lesquels la « bonne réponse » est connue. Ici, chaque exemple est un transfert, décrit par ses facteurs de performance (les entrées) et son indemnité réelle (la cible). L'apprentissage est dit supervisé car la cible est observée pour toutes les observations de l'échantillon.

Le risque central de tout apprentissage supervisé est le surapprentissage (overfitting) : un modèle trop complexe peut mémoriser les particularités de l'échantillon d'entraînement (y compris son bruit) au lieu d'en extraire la relation générale et produire alors d'excellentes performances apparentes sur les données d'entraînement mais de mauvaises prédictions sur des données nouvelles. Le phénomène inverse, le sous-apprentissage (underfitting), désigne un modèle trop simple pour capter la relation réelle. La parade standard consiste à diviser les données en un jeu d'entraînement et un jeu de test, afin d'évaluer les modèles exclusivement sur des données qu'ils n'ont pas vues lors de l'entraînement : c'est le principe de la validation croisée, appliquée ici par poste (l'échantillon est découpé en sous-ensembles (folds)), chaque modèle étant entraîné sur une partie des données et évalué sur la partie restante, en faisant tourner les rôles. Le même découpage a été imposé aux trois modèles comparés, condition indispensable pour que la comparaison soit équitable (Fouss & Satinet, s. d.).

3.5.3. Les trois modèles comparés

Trois modèles de complexité croissante ont été comparés, conformément à la démarche recommandée consistant à confronter un modèle de référence simple à des modèles plus flexibles. Ce choix précis répond à trois besoins propres à ce travail. La régression linéaire est retenue comme modèle de référence parce qu'elle est aussi le modèle sur lequel repose la décomposition LMG (voir 3.4. Décomposition de l'importance des facteurs : la méthode Shapley-LMG) : la conserver dans la comparaison garantit une cohérence entre la partie explicative et la partie prédictive du travail, et permet de vérifier si les modèles plus complexes apportent un gain réel par rapport à ce socle commun.

Le Random Forest est retenu pour sa robustesse reconnue sur des échantillons de taille modérée (Breiman, 2001), une propriété directement pertinente ici, les effectifs par poste s'échelonnant de 68 à 275 observations, un ordre de grandeur bien inférieur à celui généralement requis par d'autres.

XGBoost, enfin, est retenu parce qu'il constitue le modèle utilisé par McHale et Holmes (McHale & Holmes, 2023), l'étude de la littérature la plus proche de ce travail sur la valorisation des transferts : son inclusion permet une comparaison directe avec cette référence, tout en assumant dès le départ le risque que sa sensibilité au surapprentissage sur petits échantillons (voir 3.5.4. Conditions d'utilisation et précautions) puisse le désavantager sur certains postes mais reste un résultat informatif en soi, non un obstacle à éviter.

La régression linéaire multiple constitue le modèle de référence. Elle suppose que le prix (en logarithme²⁷) est une combinaison linéaire des facteurs de performance : chaque facteur contribue proportionnellement à sa valeur, indépendamment des autres. Sa simplicité est sa force (interprétabilité directe, stabilité sur petits échantillons) comme sa limite (incapacité à capter des relations non linéaires ou des effets d'interaction).

²⁷ Les prix de transfert présentent une distribution fortement asymétrique (la majorité des transactions se situe entre 1 et 20 M€, quelques-unes dépassant 100 M€), ce qui viole l'hypothèse de résidus approximativement normaux requise par la régression linéaire. La transformation logarithmique compresse cette échelle et ramène la distribution vers une forme plus symétrique : elle traite un écart proportionnel (par exemple, un facteur 10 entre 1 M€ et 10 M€) de la même manière qu'un écart proportionnel équivaut à une échelle plus élevée (entre 10 M€ et 100 M€), alors qu'en valeur brute ces deux écarts diffèrent d'un facteur dix en euros. Cette transformation limite également le poids disproportionné qu'auraient les quelques transferts records dans l'estimation par moindres carrés.

Le Random Forest (forêt aléatoire) agrège les prédictions d'un grand nombre d'arbres de décision. Un arbre de décision découpe l'espace des variables en régions successives par une série de questions et attribue une prédiction à chaque région. Il capte naturellement les non-linéarités et les interactions, mais un arbre isolé est très instable (de petites variations des données peuvent changer radicalement sa structure). La forêt aléatoire remédie à cette instabilité par le bagging (contraction de « bootstrap aggregating ²⁸») : chaque arbre est entraîné sur un échantillon bootstrap différent des données, avec un sous-ensemble aléatoire de variables à chaque nœud et la prédiction finale est la moyenne des arbres. Cette agrégation réduit fortement la variance des prédictions, ce qui rend la méthode particulièrement adaptée aux échantillons de taille modérée. Ces techniques de bagging et de boosting n'ont d'intérêt que pour des modèles instables et non linéaires comme les arbres, leur application à une régression linéaire n'aurait guère de sens. Ce qui justifie la comparaison de familles de modèles fondamentalement différentes plutôt que de variantes d'un même principe.

XGBoost (gradient boosting) repose également sur des arbres de décision, mais selon une logique séquentielle et non parallèle : chaque nouvel arbre est construit pour corriger les erreurs des arbres précédents, le modèle s'améliorant itérativement. Développé par Chen et Guestrin (2016), cet algorithme atteint fréquemment les meilleures performances prédictives sur données tabulaires. C'est ce modèle mis en avant par McHale et Holmes (2023) dans la littérature la plus proche de ce travail mais il est aussi plus sensible au surapprentissage, en particulier sur de petits échantillons.

3.5.4. Conditions d'utilisation et précautions

Les conditions d'application diffèrent sensiblement entre les trois modèles. La régression linéaire suppose la linéarité de la relation, l'absence de multicollinéarité forte (garantie ici par l'orthogonalité des facteurs) et des résidus bien comportés hypothèses examinées en 3.4.4. Conditions d'utilisation et diagnostics.

Les méthodes d'arbres (Random Forest, XGBoost) sont beaucoup moins exigeantes sur le plan distributionnel : elles ne supposent ni linéarité ni normalité et sont robustes aux valeurs extrêmes des variables d'entrée (Tibshirani & Friedman, s. d.). Leur exigence porte ailleurs :

²⁸ Une fois qu'on a créé plusieurs échantillons bootstrap différents et entraîné un arbre de décision sur chacun, on agrège (combine) leurs prédictions pour une régression, en faisant simplement la moyenne des prédictions de tous les arbres.

un volume de données suffisant pour que l'agrégation soit efficace et une validation rigoureuse pour contenir le surapprentissage, deux points sensibles dans ce travail compte tenu des effectifs réduits de certains postes. Pour ces postes, le risque que XGBoost surapprenne est réel et l'éventualité qu'un modèle simple y surpasse les modèles complexes a été considérée dès le départ comme un résultat possible et informatif.

3.5.5. Un cas d'école : la sensibilité de la régression aux valeurs extrêmes

L'examen des graphiques de prédiction a révélé un cas illustratif des limites respectives des modèles. Pour le poste de défenseur central, la régression linéaire produit une prédiction aberrante d'environ 370 millions d'euros pour un joueur dont l'indemnité réelle était de 2,5 millions (voir Annexe 5 : Graphiques de prédiction par poste). L'analyse de ce cas a identifié la cause : ce joueur n'avait disputé que 92 minutes sur la saison (juste au-dessus du seuil de neutralisation des statistiques per-90) de sorte que quelques actions isolées se sont extrapolées en taux par 90 minutes extrêmes (8,8 écarts-types au-dessus de la moyenne du poste sur un facteur). La régression linéaire, qui extrapole sans borne au-delà du domaine des données observées, a amplifié cette valeur en prédiction aberrante.

Les modèles d'arbres (dont les prédictions sont bornées par construction aux valeurs observées lors de l'entraînement) sont restés dans des ordres de grandeur raisonnables. Ce cas, documenté comme artefact numérique à effet de levier plutôt que comme profil de joueur atypique, illustre concrètement pourquoi la comparaison de familles de modèles différentes est précieuse et pourquoi les prédictions finales retenues sont celles de modèles d'arbres.

3.5.6. Comparaison des modèles : métriques et tests statistiques

La performance prédictive de chaque modèle a été mesurée par le RMSE²⁹ : pour chaque joueur, on calcule l'écart entre le prix prédit et le prix réel, on élève ces écarts au carré, ce qui neutralise les signes et pénalise davantage les grandes erreurs, on en fait la moyenne, puis on en prend la racine pour revenir à l'échelle d'origine. Un RMSE plus faible signale des prédictions plus proches de la réalité. Le RMSE est calculé sur l'échelle logarithmique du prix³⁰. Le RMSE est par ailleurs calculé séparément par poste, chaque poste disposant de son propre

²⁹ Root Mean Square Error, racine de l'erreur quadratique moyenne

³⁰ Une erreur absolue de même montant n'a pas la même signification sur un transfert modeste et sur un transfert record (l'échelle logarithmique ramène les erreurs à des proportions relatives comparables).

modèle et de ses propres facteurs, un RMSE global n'aurait pas de sens et systématiquement en validation croisée, jamais sur les données d'entraînement.

Le Tableau 16 présente les RMSE obtenus. Le Random Forest obtient le RMSE le plus faible sur six postes sur sept, XGBoost ne l'emportant que pour le latéral, de justesse.

Pour déterminer si ces différences sont statistiquement significatives (ou attribuables au hasard), une procédure de comparaison en trois temps a été appliquée, suivant la démarche décrite par Trawinski et al. (2012) pour la comparaison non paramétrique de plusieurs algorithmes de régression : un test global (Friedman), suivi de procédures post-hoc adaptées selon qu'il s'agit d'une comparaison de toutes les paires (Nemenyi) ou d'une comparaison ciblée entre deux modèles (Wilcoxon signé, avec correction pour comparaisons multiples). Le test de Friedman compare d'abord les rangs des trois modèles (premier, deuxième, troisième selon le RMSE) à travers les sept postes : il conclut à une différence globale significative ($p = 0,0119$), signifiant qu'il y a moins de 1,2 % de chances d'observer un tel écart de classements si les trois modèles étaient en réalité équivalents (Trawiński et al., 2012).

Deux procédures post-hoc ont ensuite cherché à identifier quelles paires de modèles diffèrent : le test de Nemenyi conclut à une supériorité significative du Random Forest sur XGBoost ($p = 0,0092$). Les tests de Wilcoxon appariés indiquent qu'avant correction, que le Random Forest surpasse à la fois XGBoost ($p = 0,0346$) et la régression linéaire ($p = 0,0225$). Après application de la correction de Holm, qui ajuste les seuils pour tenir compte du fait que plusieurs comparaisons sont menées simultanément (réduisant le risque de conclusions faussement positives) aucune paire ne demeure significative ($p = 0,069$ et $0,068$ respectivement).

Cette divergence entre procédures n'a pas été tranchée artificiellement : elle est rapportée telle quelle et s'interprète comme une conséquence de la puissance statistique limitée qu'autorise une comparaison sur sept unités seulement (les postes). La conclusion retenue est donc nuancée : le Random Forest domine systématiquement les classements et surpasse les deux autres modèles dans les comparaisons brutes, mais cette supériorité ne peut être affirmée au seuil conventionnel après correction pour comparaisons multiples, vraisemblablement faute de puissance, et non par absence de différence réelle.

3.5.7. Sélection finale et calcul de la valeur de performance

Le meilleur modèle a été retenu (poste par poste) :

- Le Random Forest pour six postes.
- XGBoost pour le latéral.

Plutôt qu'un modèle unique imposé à tous. La valeur de performance de chaque joueur est alors la prédiction de ce modèle, calculée en prédictions out-of-fold³¹. La valeur de performance étant l'exponentielle d'une prédiction en logarithme, elle est mathématiquement positive par construction (propriété vérifiée sur l'ensemble des 1 172 transferts finaux (valeurs comprises entre 1,04 et 31,8 millions d'euros)). L'écart entre l'indemnité réellement payée et cette valeur de performance constitue l'indicateur final de sur- ou sous-évaluation, dont les résultats sont analysés au chapitre 4 (voir 4. Résultats).

3.5.8. Interprétabilité : les valeurs SHAP

Les modèles d'arbres étant moins directement interprétables qu'une régression, l'importance des facteurs dans le modèle retenu a été examinée au moyen des valeurs SHAP (SHapley Additive exPlanations). Reposant sur le même principe de valeur de Shapley que la décomposition LMG, mais appliqué aux prédictions individuelles d'un modèle quelconque plutôt qu'au R^2 d'une régression, les valeurs SHAP quantifient la contribution de chaque facteur à la prédiction de chaque joueur. Elles offrent ainsi un point de comparaison indépendant avec les résultats du LMG : si deux méthodes reposant sur des modèles différents (régression linéaire d'un côté, forêt aléatoire de l'autre) identifient le même facteur dominant pour un poste, la conclusion s'en trouve renforcée. Cette comparaison croisée, présentée au chapitre 4 (voir 4. Résultats), confirme la convergence pour cinq des sept post.

4. Résultats

Ce chapitre présente les résultats de la démarche empirique en trois temps : la capacité explicative globale des modèles la hiérarchie des dimensions de la performance et sa validation croisée, puis l'analyse des écarts entre prix réel et valeur de performance, cœur de la réponse à la question de recherche.

³¹ Chaque joueur est prédit par une version du modèle entraînée sans lui, ce qui évite l'optimisme systématique qu'induirait une prédiction sur données d'entraînement

4.1. Capacité explicative de la performance sportive

Le premier résultat de ce travail réside dans la capacité explicative globale des modèles, mesurée par le coefficient de détermination R^2 . Rappelons sa lecture : le R^2 exprime la proportion des différences de prix entre joueurs que le modèle parvient à restituer à partir des seuls facteurs de performance. Le Tableau 2 présente les R^2 obtenus en validation croisée pour les trois modèles et les sept postes, et la Figure 2 en offre le pendant visuel pour le RMSE.

Poste	Régression linéaire	Random Forest	XGBoost
Défenseur central	0,140	0,138	0,097
Latéral	0,127	0,140	0,154
Milieu défensif	0,143	0,143	0,099
Ailier	0,132	0,172	0,143
Buteur	0,121	0,203	0,093
Milieu offensif	0,174	0,222	0,161
Gardien	0,153	0,107	0,067

Tableau 2 : R^2 en validation croisée, par poste et par modèle

RMSE (échelle logarithmique) par poste et modèle

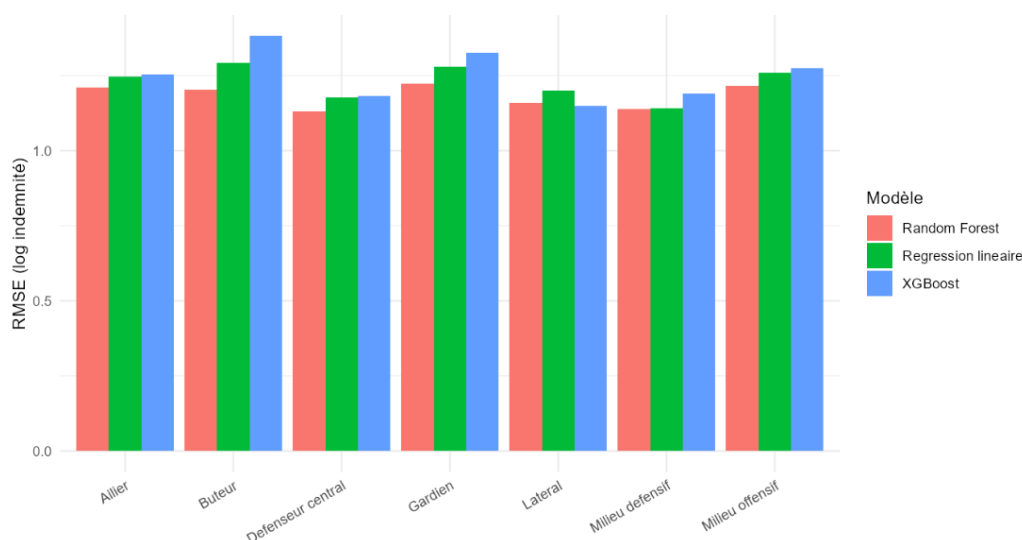


Figure 2 : RMSE (échelle logarithmique) par poste et par modèle³²

³² Pour chaque poste, la barre la plus basse désigne le modèle aux prédictions les plus proches des prix réels : le Random Forest (en rouge) domine sur six postes sur sept, XGBoost (en bleu) ne l'emportant que pour le latéral, de justesse.

Les R^2 s'échelonnent de 0,07 à 0,22 selon les postes et les modèles : la performance sportive observable, mesurée par les statistiques avancées, explique donc entre 7 % et 22 % de la variance des indemnités de transfert. Ce niveau, modeste en apparence, constitue en réalité le résultat central de ce travail au regard de la question de recherche : il quantifie précisément la mesure dans laquelle les performances observables expliquent le prix et démontre, en creux, que 78 % à 93 % de la variance des prix relève de déterminants extérieurs à la performance de terrain : les catégories « club » et « contexte » du cadre conceptuel présenté en revue de littérature (pouvoir de négociation, effet d'enchère, médiatisation, situation contractuelle), ainsi que les composantes non performatives des caractéristiques du joueur (âge, potentiel perçu, notoriété).

Un détail du tableau mérite attention : pour le gardien, la hiérarchie des modèles diffère selon la métrique retenue. Le Random Forest obtenait le meilleur RMSE (voir Tableau 16) mais c'est la régression linéaire qui atteint le meilleur R^2 (0,153 contre 0,107). RMSE et R^2 mesurant des aspects légèrement différents de la qualité prédictive (erreur typique pour l'un, proportion de variance restituée pour l'autre), leur divergence sur un même poste est le symptôme d'une instabilité d'estimation, cohérente avec l'ensemble des signaux de fragilité accumulés sur ce poste au plus petit effectif ($n = 68$) : intervalle bootstrap le plus large, test RESET significatif, indice KMO le plus faible après le milieu offensif.

4.1.1. Analyse des graphiques : les prédictions face aux prix réels

Les graphiques « Prédit vs Réel » offrent une lecture directe de ces chiffres. Leur principe est simple : chaque point représente un joueur, positionné selon son indemnité réelle (axe horizontal) et l'indemnité prédite par le modèle (axe vertical), toutes deux en échelle logarithmique. La diagonale rouge en pointillés matérialise la prédiction parfaite : un point posé exactement sur cette ligne correspond à un joueur dont le modèle a prédit le prix exact. Un point au-dessus de la ligne signale un joueur surestimé par le modèle et un point en dessous un joueur sous-estimé. Plus le nuage de points est resserré autour de la diagonale, meilleure est la qualité prédictive du modèle (toujours en gardant en tête que notre objectif n'est pas d'avoir une correspondance entre prix prédit et prix réel).

La Figure 3 présente ce graphique pour le poste de buteur, l'un des postes les plus robustes de l'étude, pour les trois modèles comparés.

Prix prédit vs prix réel : Buteur

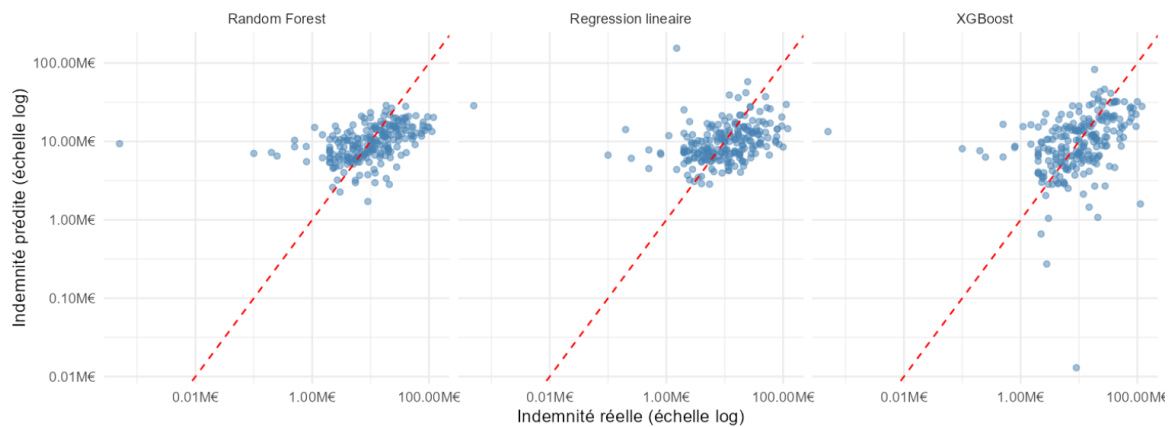


Figure 3 : Prédictions contre prix réels, poste de buteur³³ (n = 227)

La Figure 4 présente le même graphique pour le milieu offensif, le poste le plus fragile de l'étude avec le gardien et le contraste est immédiat.

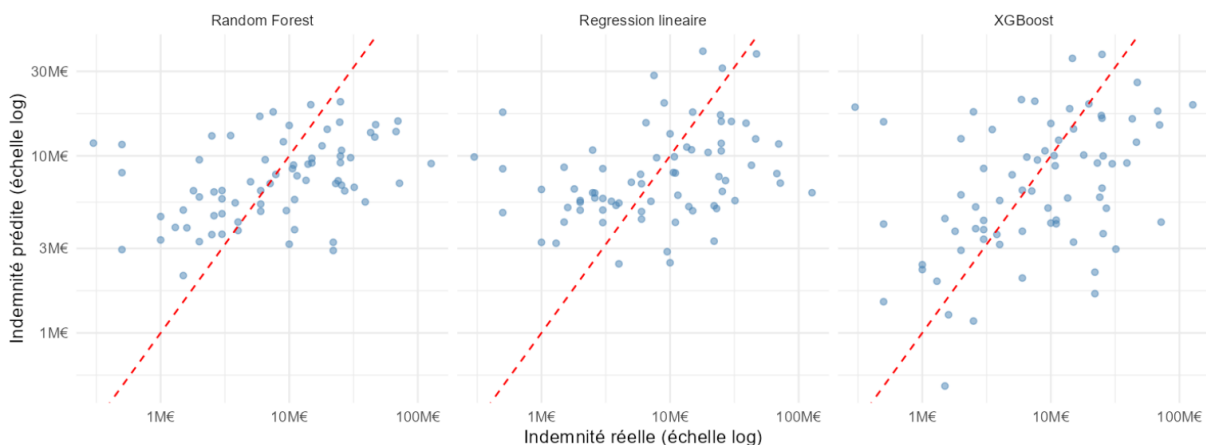


Figure 4 : Prédictions contre prix réels, poste de milieu offensif³⁴ (n = 72)

Cette lecture visuelle recoupe fidèlement la hiérarchie de robustesse établie par les intervalles bootstrap de la décomposition LMG : les postes aux intervalles resserrés (buteur, latéral, défenseur central) présentent des nuages proches de la diagonale, tandis que les postes aux intervalles les plus larges (milieu offensif, gardien) présentent les nuages les plus dispersés ou

³³ Le nuage de points suit la diagonale sur l'ensemble de la gamme de prix : lorsque le prix réel augmente, la prédiction augmente proportionnellement, signe que les modèles captent un signal réel. Le Random Forest (à gauche) présente le nuage le plus compact, cohérent avec son RMSE le plus faible.

³⁴ Les prédictions s'écrasent dans une bande étroite (environ 3 à 15 M€) quel que soit le prix réel, qui s'étend pourtant de moins de 1 M€ à plus de 50 M€ : les transferts bon marché sont systématiquement surestimés (points au-dessus de la diagonale, à gauche) et les transferts chers sous-estimés (points en dessous, à droite). Cette « compression vers la moyenne » est la traduction visuelle du faible pouvoir discriminant du modèle sur ce petit échantillon.

compressés, deux approches indépendantes convergeant vers le même diagnostic. Les graphiques des cinq autres postes, qui s'échelonnent entre ces deux situations extrêmes, figurent en Annexe 5 : Graphiques de prédiction par poste.

Un phénomène transversal mérite d'être souligné : sur l'ensemble des postes, y compris les plus robustes, la compression s'accroît aux extrémités de la gamme de prix. Les transferts les moins chers sont surestimés, les plus chers sous-estimés. Cette difficulté systématique des modèles sur les transferts extrêmes est cohérente avec la thèse de ce travail : c'est précisément sur ces transferts records médiatisés d'un côté, fins de carrière discrètes de l'autre que la composante non sportive du prix (enchères entre clubs, notoriété, âge) pèse le plus lourd, et qu'un modèle limité à la performance échoue par construction à suivre le marché.

4.2. Hiérarchie des dimensions de la performance et validation croisée

Le second résultat concerne l'identification, pour chaque poste, des dimensions de la performance les plus associées au prix. Deux méthodes indépendantes ont été mobilisées à cette fin : la décomposition LMG sur régression linéaire (voir 3.4.6. Résultats de la décomposition LMG) et les valeurs SHAP calculées sur le modèle d'arbres retenu (voir 3.5.8. Interprétabilité : les valeurs SHAP).

La Figure 5 illustre la lecture d'un graphique SHAP, sur l'exemple du défenseur central. Chaque ligne correspond à un facteur de performance, classé du plus influent (en haut) au moins influent (en bas). Chaque point représente un joueur : sa position horizontale indique l'effet du facteur sur la prédiction de prix de ce joueur. Vers la droite, le facteur pousse la prédiction vers le haut. Vers la gauche, il la tire vers le bas et sa couleur indique la valeur du facteur pour ce joueur (orange pour une valeur élevée, violet pour une valeur faible). La lecture croisée des deux informations révèle le sens de la relation : lorsque les points orange se concentrent à droite et les points violets à gauche, une valeur élevée du facteur augmente le prix prédit, le facteur influence positivement le prix. La configuration inverse signalerait une influence négative.

Le Tableau 15 systématise cette comparaison pour l'ensemble des postes, en confrontant le facteur dominant identifié par chacune des deux méthodes.

La convergence est observée pour cinq des sept postes (deux méthodes reposant sur des familles de modèles fondamentalement différentes) : l'une linéaire, l'autre non linéaire,

identifient la même dimension dominante, ce qui renforce substantiellement la crédibilité de ces résultats. Les deux divergences s'interprètent différemment. Pour le gardien, elle s'ajoute aux multiples signaux d'instabilité de ce poste et n'appelle pas d'interprétation substantielle : aucune des deux méthodes n'y produit de résultat statistiquement stable. Pour le milieu défensif, la divergence est plus intéressante : le LMG privilégie la dimension défensive traditionnelle (tacles et récupération) tandis que le SHAP met en avant la dimension de relance et de création. Deux lectures compatibles avec le profil du « récupérateur-relanceur » qui caractérise l'évolution moderne du poste. Cette interprétation demeure toutefois une piste, les deux méthodes ne convergeant pas.

4.3. Écarts entre prix réel et valeur de performance

Le troisième résultat constitue l'aboutissement de la démarche : la confrontation, pour chacun des 1 172 transferts à données complètes, entre l'indemnité réellement payée et la valeur de performance estimée. Le Tableau 12 présente les quinze écarts positifs les plus importants, les transferts dont le prix réel excède le plus ce que la performance seule justifierait.

La composition de ce classement est frappante par son homogénéité : il réunit exclusivement des transferts parmi les plus médiatisés de la période étudiée. Des joueurs à très forte notoriété, transactions record largement commentées, clubs acquéreurs au pouvoir d'achat exceptionnel. Cette concentration est précisément ce que le cadre conceptuel laissait attendre : les facteurs que le modèle exclut par construction, réputation, marketabilité, effet d'enchère entre clubs, dimension contractuelle. Ces facteurs sont exactement ceux qui caractérisent ces transferts. On observera également que les valeurs de performance estimées restent dans des ordres de grandeur cohérents (de 6,7 à 24,3 millions d'euros pour ce classement) : ce sont les prix réels qui sont extrêmes.

À l'autre extrémité du classement, les transferts les plus « sous-évalués » dont l'indemnité réelle est inférieure à la valeur de performance estimée présentent un profil symétrique : joueurs peu médiatisés (souvent en fin de carrière ou cantonnés à des rôles de rotation) et fréquemment engagés dans une mobilité descendante entre championnats (d'un championnat majeur vers un championnat moins exposé). Cette composition suggère que la sous-évaluation ne signale pas nécessairement des « talents cachés », mais reflète-t-elle aussi

le poids des facteurs de marché : l'âge, la notoriété déclinante et la visibilité réduite pèsent sur le prix indépendamment du niveau statistique de la saison précédente.

Deux précautions d'interprétation s'imposent avant de tirer des conclusions de ces écarts. D'abord, les modèles utilisés n'expliquent qu'une partie limitée du prix (R^2 modestes, entre 7 % et 22 % selon les postes). Cela veut dire que la marge d'erreur sur chaque écart individuel reste importante : on peut affirmer avec confiance quels transferts s'écartent le plus de leur profil statistique (le classement relatif) mais pas la valeur exacte de cet écart pour un joueur précis. Ces chiffres ne permettent donc pas d'affirmer, pour tel joueur en particulier, qu'il "vaut" réellement tel montant.

Ensuite, un écart important entre prix réel et valeur de performance ne signifie pas que le club a fait un mauvais choix. Un prix plus élevé que la valeur de performance peut très bien se justifier par des bénéfices réels mais non sportifs (recettes de merchandising, image de marque, valeur de revente anticipée) que ce travail ne mesure pas.

5. Discussion et perspectives

5.1. Retour sur la question de recherche

La question de recherche demandait dans quelle mesure les performances observables avant un transfert, mesurées à l'aide de statistiques avancées, permettent d'expliquer le prix de transfert. La réponse apportée par ce travail est double. Quantitativement, cette mesure est réelle mais limitée : de 7 % à 22 % de la variance des indemnités selon les postes. Qualitativement, la structure des écarts entre prix réel et valeur de performance est fortement cohérente avec le cadre conceptuel : les transferts qui s'écartent le plus de leur profil statistique, dans un sens comme dans l'autre, sont ceux où les facteurs extra-sportifs (médiatisation, notoriété, âge, trajectoire de carrière) sont les plus saillants. La performance observable constitue donc une composante mesurable mais minoritaire de la formation des prix, ce qui valide empiriquement l'intuition, présente dans la littérature, selon laquelle l'indemnité de transfert est le produit d'un système de détermination bien plus large que la seule contribution sportive.

5.2. Apports méthodologiques

Au-delà de ses résultats, ce travail présente trois caractéristiques ne sont pas réunies dans la littérature existante sur la valorisation des transferts (méthodologiques à ma connaissance).

Premièrement, la combinaison d'une décomposition explicative (LMG) et d'une modélisation prédictive (par apprentissage supervisé) sur les mêmes facteurs avec validation croisée de leurs conclusions (par les valeurs SHAP). La convergence observée sur cinq postes sur sept constituant un résultat de robustesse en soi.

Deuxièmement, la quantification systématique de l'incertitude : chaque part de contribution est accompagnée de son intervalle de confiance bootstrap, et la comparaison des modèles s'appuie sur des tests statistiques formels dont les divergences sont rapportées plutôt que dissimulées.

Troisièmement, la traçabilité intégrale des décisions : chaque exclusion de variable, chaque arbitrage de seuil et chaque correction d'anomalie ont été consignés.

5.3. Limites

Les limites de ce travail doivent être énoncées avec la même clarté que ses apports.

- La taille des échantillons par poste constitue la contrainte la plus structurante. Pour le gardien ($n = 68$) et le milieu offensif ($n = 72$), l'ensemble des indicateurs converge vers une instabilité d'estimation que ni le resserrement des variables ni le choix des modèles n'ont pu compenser : cette limite tient à la rareté intrinsèque des transferts observables sur ces postes, non à la méthode.
- Les statistiques soient standardisées par championnat et par saison, un biais de contexte d'équipe subsiste partiellement : un joueur évoluant dans une équipe dominante bénéficiera mécaniquement d'un volume de touches, de passes progressives et d'occasions créées supérieur à celui d'un joueur de niveau comparable évoluant dans une équipe défensive. La normalisation per-90 atténue ce biais sans l'éliminer entièrement.
- Le périmètre est restreint aux cinq championnats majeurs et aux transferts à indemnité divulguée : les résultats ne sont pas généralisables aux championnats secondaires, aux transferts libres, ni aux prêts, exclus par construction.

- Les intervalles de confiance demeurent larges dans l'absolu, y compris pour les postes les plus robustes : les hiérarchies de facteurs identifiées doivent être lues comme des tendances étayées, non comme des paramètres précisément estimés.
- Enfin, la couverture temporelle (2018-2024) inclut la période de la pandémie de COVID-19, dont l'effet perturbateur sur le marché des transferts n'a pas été modélisé spécifiquement.

5.4. Perspectives

La perspective la plus directe concerne l'accumulation des données. La couverture des statistiques avancées (inexistante avant 2018) s'étend chaque saison à de nouveaux championnats et s'enrichit de nouveaux indicateurs (données de tracking physique, métriques de valorisation des actions). Chaque saison écoulée accroît mécaniquement les effectifs disponibles par poste (desserrant la contrainte qui pèse aujourd'hui sur les gardiens et les milieux offensifs) et l'extension aux championnats de second rang, impossible aujourd'hui faute de couverture complète, deviendra progressivement réalisable. Les biais et instabilités documentés dans ce travail ont ainsi vocation à s'atténuer d'eux-mêmes : la méthodologie développée pourrait être ré-exécutée à l'identique sur ces données enrichies, avec une précision croissante.

Deux autres prolongements méritent mention. D'une part, l'intégration explicite de variables de contexte (notoriété mesurée par l'empreinte médiatique ou les réseaux sociaux, situation contractuelle, caractéristiques du club acquéreur) permettrait de passer d'une mesure en creux des facteurs extra-sportifs (ce que la performance n'explique pas) à leur mesure directe et de décomposer l'écart aujourd'hui résiduel.

D'autre part, une validation prospective de la valeur de performance. C'est-à-dire vérifier si les joueurs identifiés comme sous-évalués ont effectivement produit (dans les saisons suivant leur transfert) une contribution sportive supérieure à leur prix. Ce qui constituerait un test plus exigeant de la pertinence de l'indicateur construit ici, et rapprocherait ce travail de l'usage opérationnel qu'en font les départements de recrutement.

En définitive, si ce travail ne prétend pas avoir isolé la « vraie » valeur d'un joueur de football (dont l'existence même est discutable) il établit qu'une part identifiable et croissante du prix des transferts peut être rattachée à la performance observable et fournit un cadre

méthodologique reproductible pour mesurer, saison après saison, la frontière mouvante entre ce que le terrain justifie et ce que le marché décide.

6. Conclusion

Ce travail partait d'un constat posé dans la littérature : qualifier le marché des transferts de football de bulle spéculative suppose de pouvoir identifier un écart entre le prix observé d'un joueur et sa valeur fondamentale, laquelle reste jusqu'ici non définie et non mesurée. L'objectif de ce mémoire était précisément de combler ce manque, en proposant une première opérationnalisation de cette grandeur, désignée valeur de performance et définie comme la traduction monétaire des seules statistiques avancées observables avant un transfert.

Pour y répondre, une démarche en trois temps a été mise en œuvre sur un échantillon de 1 172 transferts issus des cinq grands championnats européens (2018-2024). Une analyse en composantes principales réalisée séparément pour chacun des sept postes a d'abord permis de structurer 63 statistiques avancées en dimensions synthétiques interprétables. Une décomposition Shapley-LMG a ensuite identifié, poste par poste, les dimensions les plus associées au prix. Enfin, trois modèles d'apprentissage supervisé ont été comparés en validation croisée et le Random Forest (sélectionné sur six postes sur des sept postes) a servi à produire les estimations out-of-fold constituant un proxy valeur de performance de chaque joueur.

Les résultats apportent une réponse chiffrée à la question posée : la performance sportive observable explique entre 7 % et 22 % de la variance des indemnités selon les postes — quantifiant précisément la part que les statistiques justifient et démontrant en creux que la majorité du prix relève de déterminants extra-sportifs. La structuration des écarts confirme cette lecture : les plus fortes sur-évaluations concernent systématiquement des transferts très médiatisés, tandis que les sous-évaluations touchent des profils discrets ou en mobilité descendante.

Par rapport aux travaux existants les plus proches, notamment McHale & Holmes (2023), ce mémoire apporte deux contributions distinctes. D'une part, il ne se contente pas de prédire le prix de transfert, mais construit explicitement un proxy de la valeur de performance comme grandeur économique interprétable, ancrée dans la théorie de la valeur fondamentale. D'autre

part, il combine une décomposition explicative (LMG avec intervalles de confiance bootstrap) et une approche prédictive, une double validation méthodologique absente des études antérieures.

Ces résultats s'entendent dans les limites inhérentes au travail : la marge d'incertitude sur chaque estimation individuelle reste substantielle, les postes aux effectifs les plus réduits produisent des résultats structurellement moins stables et les statistiques disponibles ne capturent pas l'ensemble des dimensions de la performance. Ces limites définissent en retour les prolongements naturels : enrichissement annuel de l'échantillon, intégration de variables de contexte pour quantifier directement les déterminants extra-sportifs, et validation prospective des estimations sur les saisons suivant les transferts.

Ce travail ne prétend pas avoir mesuré la vraie valeur d'un joueur de football. Il démontre qu'une part de cette valeur est mesurable, propose un cadre méthodologique reproductible pour la quantifier et fournit un premier outil empirique pour distinguer ce que le terrain justifie de ce que le marché décide.

7. Bibliographie

- Ambler, S. (s. d.). *ECO 4272 : Introduction à l'Économétrie Tests diagnostics*.
- Amisco. (s. d.). M Capital. Consulté 17 mai 2026, à l'adresse <https://www.mcapital.fr/portfolio/amisco/>
- Bastien Drut. (2019). *Mercato : L'économie du football au XXIe siècle* (2e édition). Bréal.
- Baumer, B. S., Matthews, G. J., & Nguyen, Q. (2023). *Big Ideas in Sports Analytics and Statistical Tools for their Investigation* (arXiv:2301.04001). arXiv. <https://doi.org/10.48550/arXiv.2301.04001>
- Baumer, B., & Zimbalist, A. (2014). The sabermetric revolution : Assessing the growth of analytics in baseball. *The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball*, 1-189.
- Behravan, I., & Razavi, S. (2021). A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing*, 1-13. <https://doi.org/10.1007/s00500-020-05319-3>
- Borghesi, M., Costa, L. D., Morra, L., & Lamberti, F. (2023). Using Temporal Convolutional Networks to estimate ball possession in soccer games. *Expert Systems with Applications*, 223, 119780. <https://doi.org/10.1016/j.eswa.2023.119780>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Carmichael, F., McHale, I., & Thomas, D. (2011). MAINTAINING MARKET POSITION : TEAM PERFORMANCE, REVENUE AND WAGE EXPENDITURE IN THE ENGLISH PREMIER LEAGUE. *Bulletin of Economic Research*, 63(4), 464-497. <https://doi.org/10.1111/j.1467-8586.2009.00340.x>
- Chen, T., & Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>
- Ducarroz, C., & Sinigaglia, N. (s. d.). *Cours de Méthodes et Modèles en Marketing 2026*.
- Fouss, F., & Satinet, C. (s. d.). *Cours d'analyse des données 2026*.
- Franceschi, M., Brocard, J.-F., Follert, F., & Gouguet, J.-J. (2024). Determinants of football players' valuation : A systematic review. *Journal of Economic Surveys*, 38(3), 577-600. <https://doi.org/10.1111/joes.12552>
- Halson, S. L. (2014). Monitoring training load to understand fatigue in athletes. *Sports Medicine*, 44 Suppl 2(Suppl 2), S139-147. <https://doi.org/10.1007/s40279-014-0253-z>

Infantino, G. (s. d.). *Fédération Internationale de Football Association*.

Jesse, D., Lotte, B., Laurens, D., Arne, J., Wannes, M., Pieter, R., Jan Van, H., & Maaike, V. R. (2024). Methodology and evaluation in sports analytics : Challenges, approaches, and lessons learned. *Machine Learning*, 113(9), 6977-7010.

Luca, B. (s. d.). *La Cour de justice acteur politique de l'intégration européenne : À propos de l'arrêt Bosman*.

McHale, I. G., & Holmes, B. (2023). Estimating transfer fees of professional footballers using advanced performance metrics and machine learning. *European Journal of Operational Research*, 306(1), 389-399. <https://doi.org/10.1016/j.ejor.2022.06.033>

Opta data from Stats Perform. (s. d.). *Stats Perform*. Consulté 17 mai 2026, à l'adresse <https://www.statsperform.com/opta/>

PFSA, T. (2025, octobre 1). Comprendre l'analyse des données dans le football : Une révolution moderne. *The PFSA*. <https://thepfsa.co.uk/fr/understanding-data-analytics-in-football/>

Poli, D. R., Ravenel, L., & Besson, R. (s. d.). *Scientific evaluation of the transfer value of football players*.

Schaden, E. (2026, février 13). *Particle swarm optimization*. <http://old.birthealarm.com>

Simon Kuper & Stefan Szymanski. (2016). *Les attaquants les plus chers ne sont pas ceux qui marquent le plus et autres mystères du football décryptés* (2e édition). De Boeck Supérieur s.a.

<https://www.furet.com/media/pdf/feuilleteage/9/7/8/2/8/0/7/3/9782807300064.pdf?srltid=AfmBOor5GZptbwbSZKA9kPpj6RAyE-G8JfjQFqpCQbjNlXW8FUd0b4J9>

Tibshirani, S., & Friedman, H. (s. d.). *The Elements of Statistical Learning*.

Trawiński, B., Smętek, M., Telec, Z., & Lasota, T. (2012). Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *International Journal of Applied Mathematics and Computer Science*, 22(4), 867-881. <https://doi.org/10.2478/v10006-012-0064-z>

Wabgaonkar, H. (2022, mai 28). Relative Importance of Variables for Explainable AI : Shapley Values and Linear Regression. *Medium*. https://medium.com/@baw_H1/relative-importance-of-variables-shapley-values-and-linear-regression-c828cbbc7e0d

9. Annexes

Annexe 1 : Typologie des données mobilisées

Variable	Quantitative / Qualitative	Sous-type	Format
Indemnité de transfert	Quantitative	Continue (montant en €)	Numérique
Statistiques avancées per-90 (xG, passes progressives, tacles...)	Quantitative	Continue	Numérique
Statistiques de comptage brutes (buts, arrêts...)	Quantitative	Discrete (avant normalisation per-90)	Numérique
Scores factoriels (issus de l'ACP)	Quantitative	Continue	Numérique
Poste (7 rôles)	Qualitative (catégorielle)	Nominale (sans ordre)	Textuel
Championnat (Big 5)	Qualitative (catégorielle)	Nominale	Textuel
Saison (2018-19 à 2023-24)	Qualitative (catégorielle)	Ordinale (ordre chronologique)	Date / textuel
Nationalité	Qualitative (catégorielle)	Nominale	Textuel
Nom du joueur / club	Qualitative (catégorielle)	Nominale (identifiant)	Textuel
Statut d'exclusion (exclu / conservé)	Qualitative (catégorielle)	Nominale binaire	Booléen

Tableau 3: Typologie des données mobilisées

Annexe 2 : Pool initial de 63 variables candidates (issu de FBref)

Liste complète des 63 statistiques avancées collectées sur FBref, point de départ commun avant toute adaptation par poste. Chaque poste n'a ensuite retenu qu'un sous-

ensemble pertinent de ce pool (46 à 51 variables selon le poste, voir 3.2.5. Sélection des variables candidates et traitement des redondances) avant le resserrement final.

Volume de jeu et disponibilité

Variable	Explication
mp_playing	Nombre de matchs joués sur la saison
starts_playing	Nombre de matchs joués comme titulaire
min_playing	Nombre total de minutes jouées

Production offensive (buts, passes décisives, xG/xA)

Variable	Explication
gls_per	Buts marqués par 90 minutes
ast_per	Passes décisives par 90 minutes
g_a_per	Buts + passes décisives par 90 minutes
g_minus_pk_per	Buts hors penalty par 90 minutes
g_a_minus_pk_per	Buts + passes décisives hors penalty par 90 minutes
x_g_per	Buts attendus (xG) par 90 minutes (qualité des occasions obtenues)
x_ag_per	Passes décisives attendues (xA) par 90 minutes (qualité des occasions créées)
x_g_x_ag_per	xG + xA combinés par 90 minutes
npx_g_per	xG hors penalty par 90 minutes
npx_g_x_ag_per	xG hors penalty + xA par 90 minutes
so_t_percent_standard	Pourcentage de tirs cadrés sur le total de tirs tentés
ga90	Buts encaissés par 90 minutes (gardien)

Passes et construction

Variable	Explication
----------	-------------

cmp_percent_total	Pourcentage de passes réussies, toutes distances confondues
cmp_percent_short	Pourcentage de passes courtes réussies
cmp_percent_medium	Pourcentage de passes moyennes réussies
cmp_percent_long	Pourcentage de passes longues réussies
per90_pas_tot_dist_total	Distance totale parcourue par les passes, par 90 minutes
per90_pas_prg_dist_total	Distance parcourue par les passes progressives (vers le but adverse), par 90 min
per90_kp	Passes clés (menant directement à un tir) par 90 minutes
cmp_percent_launched	Pourcentage de passes longues relancées (>40m) réussies (gardien)
launch_percent_passes	Pourcentage de passes du gardien jouées en relance longue
launch_percent_goal	Pourcentage de dégagements au but joués en relance longue (gardien)

Défense et duels

Variable	Explication
tkl_percent_challenges	Pourcentage de duels au sol (tacles) remportés
won_percent_aerial	Pourcentage de duels aériens remportés
per90_tkl_tackles	Tacles tentés par 90 minutes
per90_tkl_w_tackles	Tacles réussis par 90 minutes
per90_tkl_challenges	Duels défensifs tentés par 90 minutes
per90_blocks_blocks	Actions défensives bloquées (tirs et passes) par 90 minutes
per90_sh_blocks	Tirs adverses contrés par 90 minutes
per90_pass_blocks	Passes adverses contrées/interceptées par 90 minutes

per90_int	Interceptions par 90 minutes
per90_clr	Dégagements défensifs par 90 minutes

Dribbles et pertes de balle

Variable	Explication
succ_percent_take	Pourcentage de dribbles tentés réussis
tkld_percent_take	Pourcentage de dribbles tentés où le joueur a été stoppé par un tackle
per90_mis_carries	Pertes de balle par contrôle raté, par 90 minutes
per90_dis_carries	Pertes de balle par dépossession (tackle adverse en conduite) par 90 minutes

Touches de balle par zone

Variable	Explication
per90_touches_touches	Nombre total de touches de balle par 90 minutes
per90_def_pen_touches	Touches de balle dans sa propre surface de réparation, par 90 minutes
per90_def_3rd_touches	Touches de balle dans son propre tiers de terrain, par 90 minutes
per90_mid_3rd_touches	Touches de balle dans le tiers central, par 90 minutes
per90_att_3rd_touches	Touches de balle dans le tiers adverse, par 90 minutes
per90_att_pen_touches	Touches de balle dans la surface adverse, par 90 minutes

Conduite de balle (carries)

Variable	Explication
per90_carries_carries	Nombre de conduites de balle par 90 minutes
per90_tot_dist_carries	Distance totale parcourue balle au pied, par 90 minutes

per90_prg_dist_carries	Distance parcourue balle au pied vers le but adverse, par 90 minutes
per90_prg_c_carries	Nombre de conduites de balle progressives, par 90 minutes
per90_final_third_carries	Conduites de balle amenant le ballon dans le tiers adverse, par 90 minutes
per90_cpa_carries	Conduites de balle amenant le ballon dans la surface adverse, par 90 minutes

Réception de passes

Variable	Explication
per90_rec_receiving	Passes reçues par 90 minutes
per90_prg_r_receiving	Passes progressives reçues par 90 minutes

Discipline et pressing

Variable	Explication
per90_fls	Fautes commises par 90 minutes
per90_fld	Fautes subies par 90 minutes
per90_off	Hors-jeux signalés par 90 minutes
per90_recov	Ballons récupérés par 90 minutes
crd_y	Nombre de cartons jaunes reçus sur la saison
crd_r	Nombre de cartons rouges reçus sur la saison

Statistiques spécifiques gardien

Variable	Explication
save_percent	Pourcentage de tirs cadrés arrêtés
cs_percent	Pourcentage de matchs joués sans encaisser de but (clean sheet)
save_percent_penalty	Pourcentage de penalties arrêtés
stp_percent_crosses	Pourcentage de centres adverses interceptés/stoppés

--	--

Total : 63 variables. Les statistiques exprimées en "per90" sont normalisées par 90 minutes de jeu. Les pourcentages ("percent") sont bornés entre 0 et 100 avant standardisation en score z. Les variables de discipline (crd_y, crd_r) et de volume de jeu (mp_playing, starts_playing) ont été exclues du pool final lors du resserrement, en tant que proxys de contexte plutôt que de performance pure (voir 3.2.5. Sélection des variables candidates et traitement des redondances).

Annexe 3 : Correspondance entre F (SHAP), RC (ACP) et nom de dimension

Ce document rassemble, pour chacun des sept postes, la correspondance entre les labels (F1, F2...) utilisés dans les graphiques SHAP (l'ordre d'importance dans le modèle Random Forest retenu) et les composantes RC (Rotated Component) de l'analyse en composantes principales, avec le nom de dimension attribué et la variable dont la saturation (loading) est la plus forte sur cette composante.

Défenseur central :

F	RC	Dimension	Variable dominante (loading)
F1	RC1	Percussion offensive (dernier tiers)	Touches dernier tiers (0.86)
F2	RC2	Volume de jeu et progression	Distance passes progressives (0.83)
F3	RC3	Finition / efficacité offensive	Buts (0.87)
F4	RC9	Précision de passe (sécurité technique)	% passes réussies (0.85)
F5	RC4	Blocs et dégagements défensifs	Tirs contrés (0.83)
F6	RC8	Duels aériens et solidité défensive	% duels aériens gagnés (0.63)
F7	RC5	Interceptions et récupérations	Interceptions (0.75)
F8	RC6	Création / passes décisives	Passes décisives (0.86)
F9	RC7	Tacles (volume et réussite)	Tacles tentés (0.82)

Tableau 4 : Tableau correspondance spécifique au poste de défenseur central

Latéral :

F	RC	Dimension	Variable dominante (loading)
F1	RC2	Construction et précision de passe	% passes réussies (0.90)
F2	RC1	Création / passes décisives	xA (0.82)
F3	RC3	Progression balle au pied	Courses progressives (0.87)
F4	RC4	Présence défensive dans sa surface	Touches surface déf. (0.83)
F5	RC6	Finition / efficacité offensive	xG (0.91)
F6	RC5	Tacles (volume et réussite)	Tacles tentés (0.87)
F7	RC7	Discipline / engagement	Fautes subies (0.79)
F8	RC10	Interceptions et récupérations	Interceptions (0.78)
F9	RC9	Prises de risque offensives (hors-jeu)	Hors-jeu (0.74)
F10	RC8	Duels aériens	% duels aériens gagnés (0.76)

Tableau 5 : Tableau correspondance spécifique au poste de milieu latéral

Milieu défensif :

F	RC	Dimension	Variable dominante (loading)
F1	RC2	Précision de passe (sécurité technique)	% passes réussies (0.94)
F2	RC3	Progression balle au pied	Courses progressives (0.82)
F3	RC1	Création / passes décisives	xA (0.83)
F4	RC8	Finition / efficacité offensive	Buts (0.89)
F5	RC6	Présence défensive dans sa surface	Touches surface déf. (0.81)
F6	RC4	Tacles (volume et réussite)	Tacles tentés (0.84)
F7	RC5	Récupération et sobriété offensive	Récupérations (0.57)
F8	RC7	Blocs défensifs (tirs)	Tirs bloqués (0.85)

F9	RC9	Facteur diffus (précision de tir/duels mêlés)	% tirs cadrés (-0.75)
----	-----	---	-----------------------

Tableau 6 : Tableau correspondance spécifique au poste de milieu défensif

Ailier :

F	RC	Dimension	Variable dominante (loading)
F1	RC2	Progression balle au pied (percussion offensive)	Courses progressives (0.86)
F2	RC1	Solidité défensive et sobriété technique	Touches surface déf. (0.79)
F3	RC3	Tacles et récupérations	Tacles tentés (0.82)
F4	RC8	Précision de passe (sécurité technique)	% passes réussies (0.95)
F5	RC4	Création / passes décisives	Passes décisives (0.83)
F6	RC5	Finition / efficacité offensive	Buts (0.86)
F7	RC6	Duels aériens et engagement physique	% duels aériens gagnés (0.62)
F8	RC7	Réussite au dribble	% dribbles réussis (0.61)

Tableau 7 : Tableau correspondance spécifique au poste d'ailier

Buteur :

F	RC	Dimension	Variable dominante (loading)
F1	RC1	Percussion et présence en surface adverse	Touches surface adv. (0.83)
F2	RC5	Pressing et récupération haute	Tacles réussis (0.77)
F3	RC2	Finition / efficacité offensive	Buts (0.92)
F4	RC7	Volume de jeu et implication en construction	Distance totale passes (0.71)
F5	RC6	Précision de passe (sécurité technique)	% passes réussies (0.92)
F6	RC4	Duels aériens et présence défensive occasionnelle	Touches surface déf. (0.79)

F7	RC3	Pertes de balle et fautes	Pertes de balle en conduite (0.75)
F8	RC8	Création / passes décisives	Passes décisives (0.77)
F9	RC9	Efficacité au tackle	% tackles réussis (0.74)
F10	RC10	Réussite au dribble	% dribbles réussis (0.84)

Tableau 8 : Tableau correspondance spécifique au poste de buteur

Milieu offensif :

F	RC	Dimension	Variable dominante (loading)
F1	RC2	Création / passes décisives	xA (0.93)
F2	RC1	Finition / efficacité offensive	Buts (0.81)
F3	RC5	Sécurité technique et technique individuelle	% passes longues réussies (0.87)
F4	RC3	Progression balle au pied	Courses progressives (0.89)
F5	RC4	Pressing / engagement défensif	Fautes commises (0.88)

Tableau 9 : Tableau correspondance spécifique au poste de milieu offensif³⁵

Gardien :

F	RC	Dimension	Variable dominante (loading)
F1	RC1	Style de distribution (jeu court vs jeu long)	% passes réussies (0.90)
F2	RC2	Arrêts / solidité	% clean sheets (0.86)
F3	RC3	Sweeping / jeu hors surface	Distance moy. sweeper-keeper (0.87)
F4	RC5	Sécurité sur les centres	% centres stoppés (0.74)
F5	RC4	Duels aériens	% duels aériens gagnés (0.92)

Tableau 10 : Tableau correspondance spécifique au poste de gardien

³⁵ Le facteur F9 du Milieu défensif (« facteur diffus ») ne correspond pas à une dimension clairement interprétable, sa variable dominante présente une saturation négatif modéré, signe d'une composante résiduelle plutôt que d'une vraie dimension de

Annexe 4 : Graphiques des valeurs de SHAP par poste

SHAP : Défenseur central

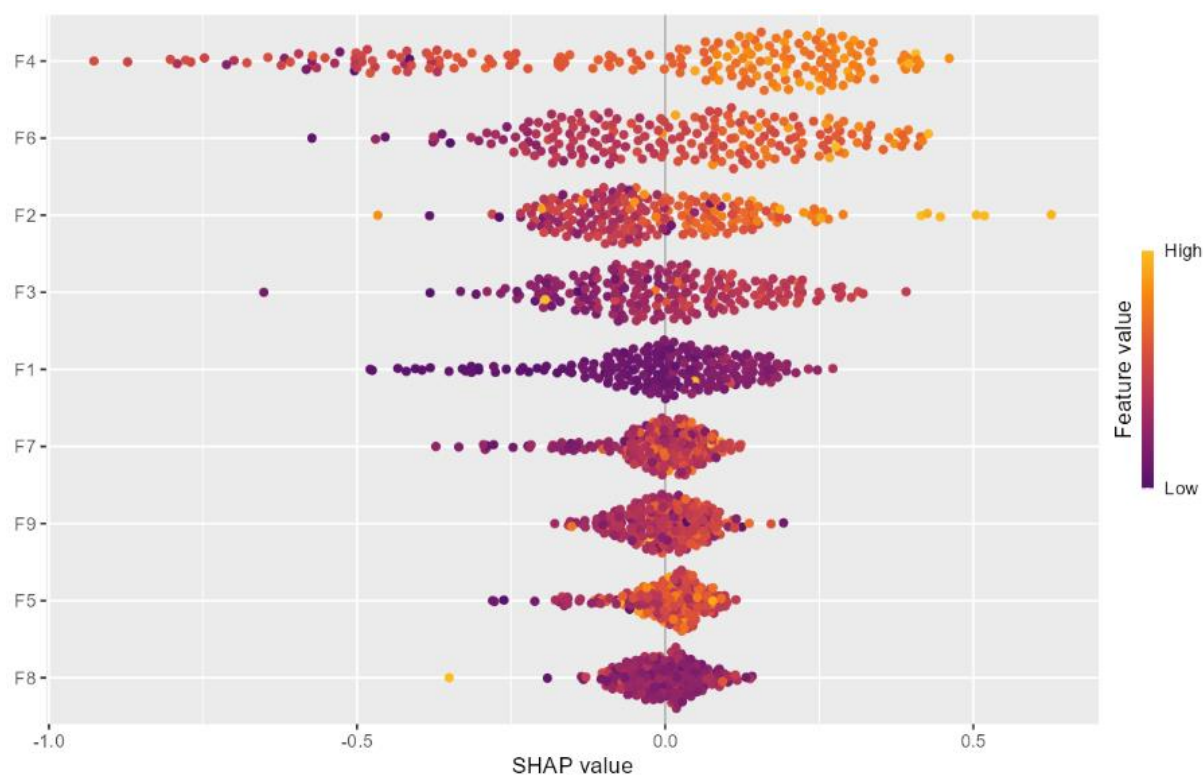


Figure 5 : Valeurs SHAP du modèle Random Forest, poste de défenseur central³⁶

³⁶ Le facteur dominant (F4, en haut) présente la séparation la plus nette : les joueurs à valeur élevée sur ce facteur (points orange, à droite) voient leur prix prédit augmenter, ceux à valeur faible (points violets, à gauche) le voient diminuer. Une influence positive, forte et cohérente. L'examen des saturations identifie ce facteur comme la précision de passe, confirmant le résultat de la décomposition LMG pour ce poste. Les graphiques SHAP des six autres postes figurent en Annexe 4 : Graphiques des valeurs de SHAP par poste



Figure 6 : Valeurs SHAP du modèle Random Forest, poste d'ailier

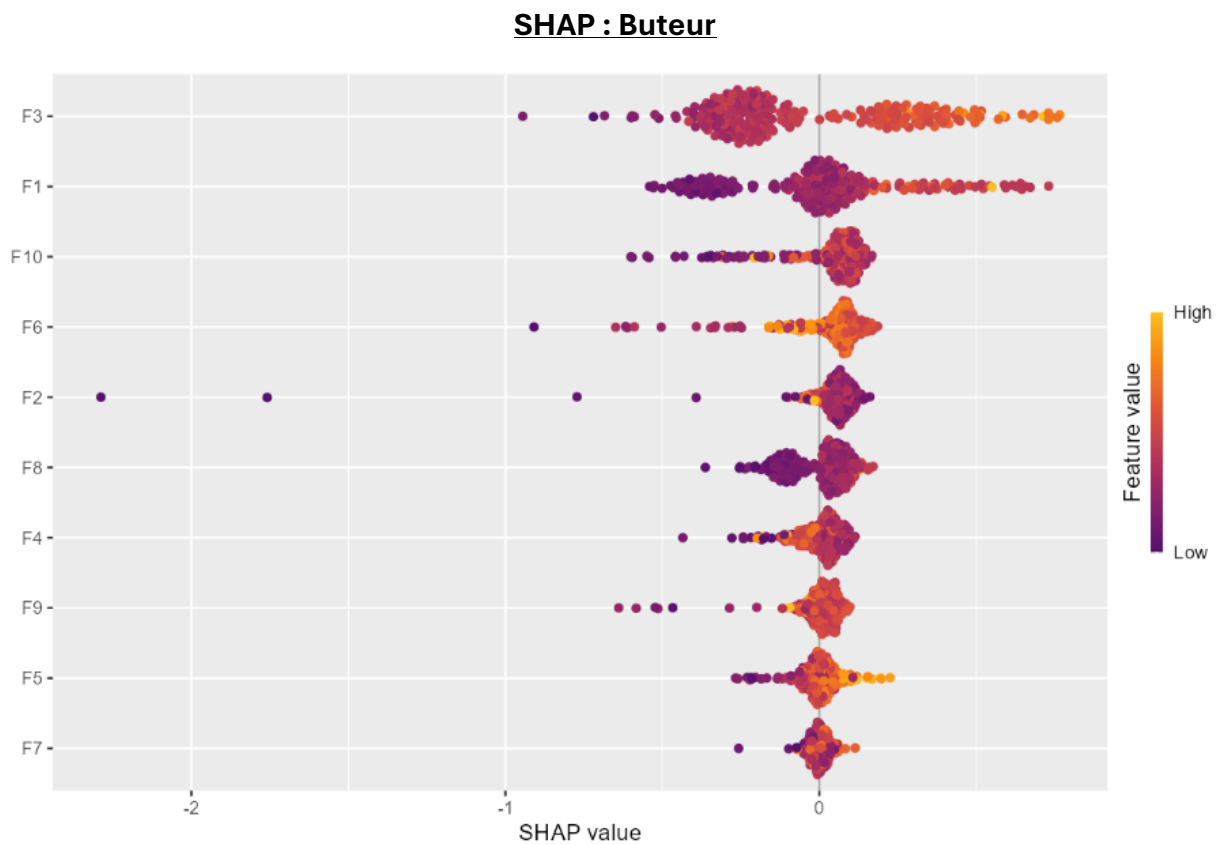


Figure 7 : Valeurs SHAP du modèle Random Forest, poste de buteur

SHAP : Gardien

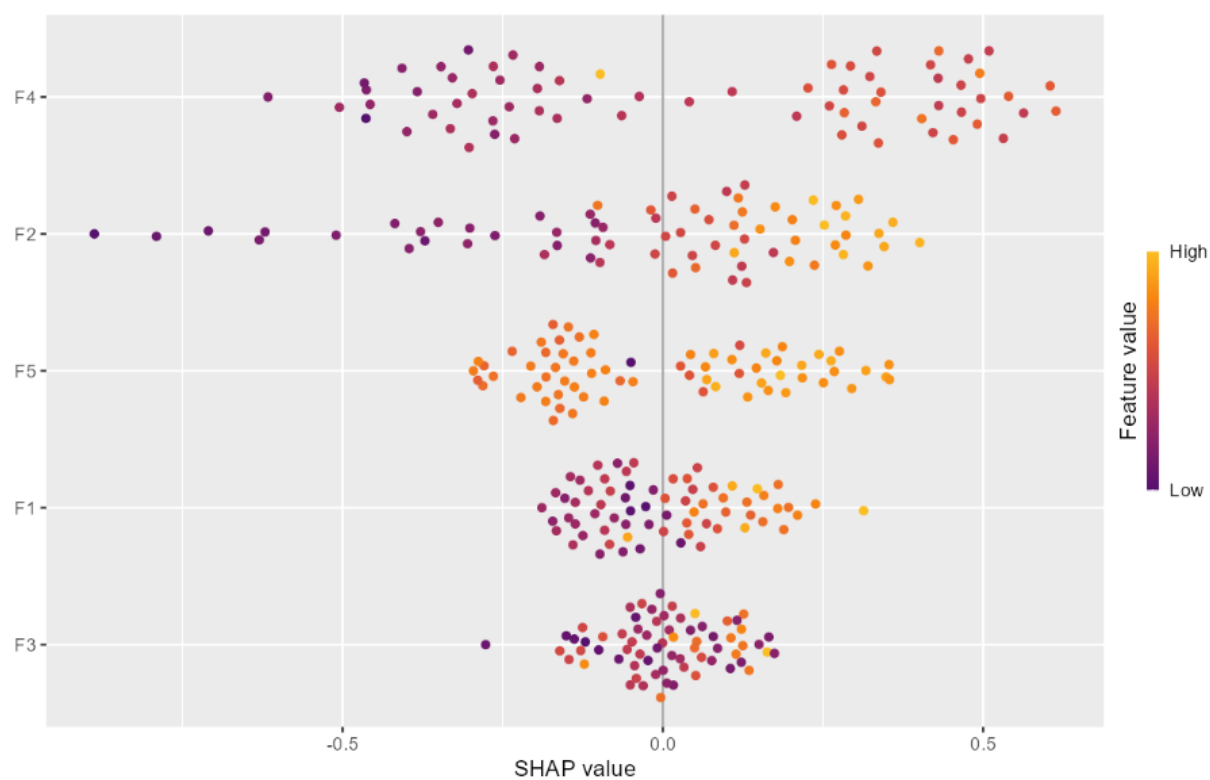


Figure 8 : Valeurs SHAP du modèle Random Forest, poste de gardien

SHAP : Latéral

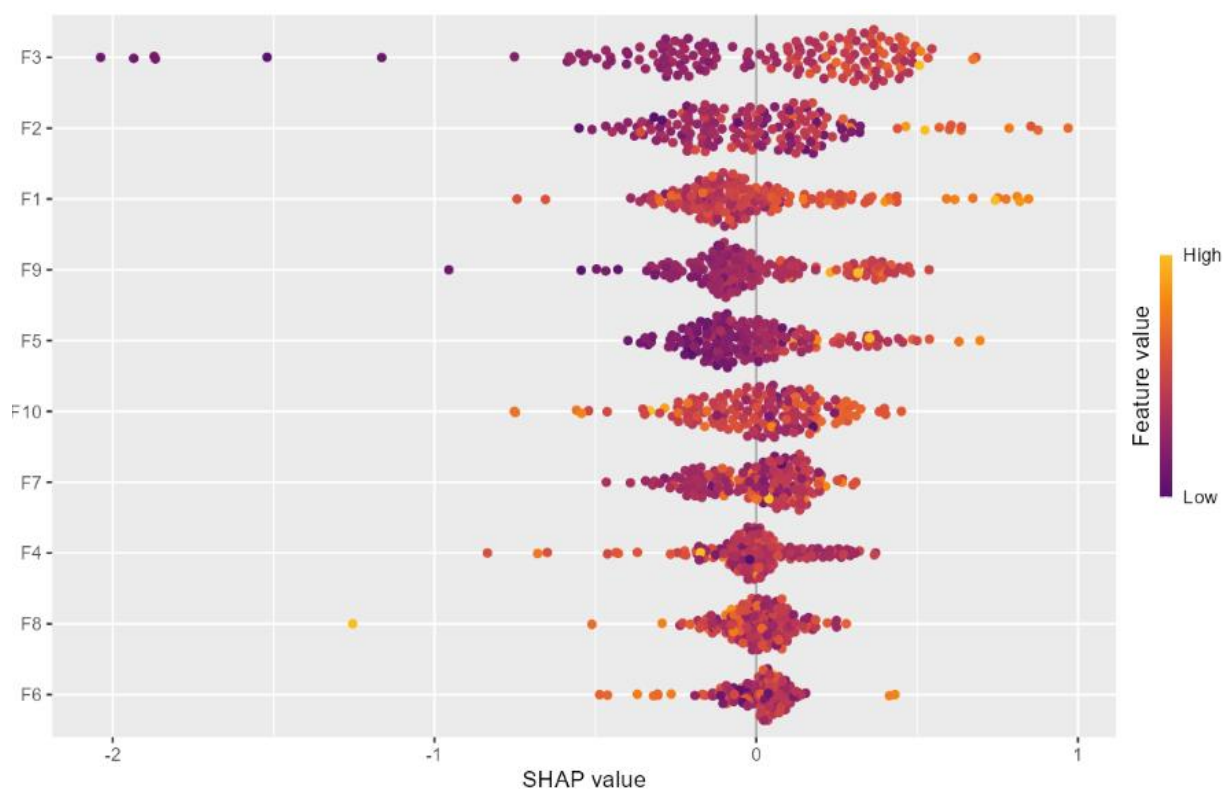


Figure 9 : Valeurs SHAP du modèle Random Forest, poste de latéral

SHAP : Milieu défensif

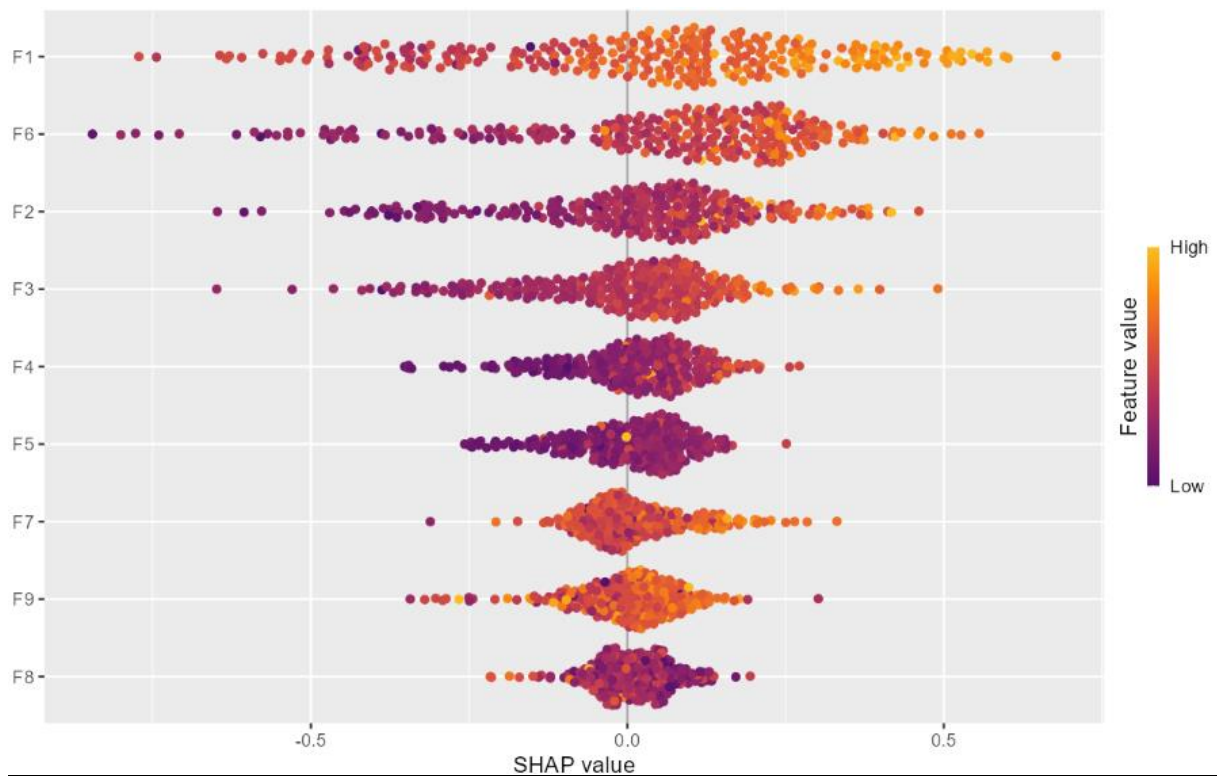


Figure 10 : Valeurs SHAP du modèle Random Forest, poste de milieu défensif

SHAP : Milieu offensif

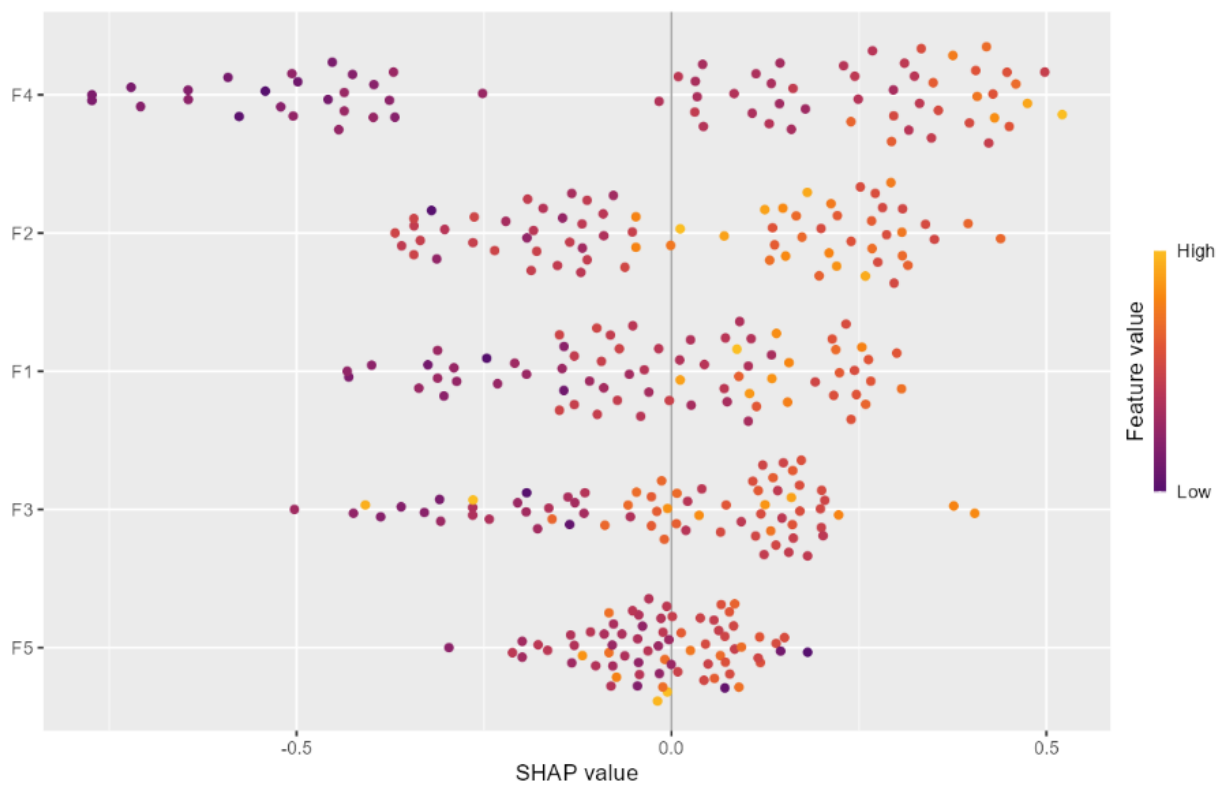


Figure 11 : Valeurs SHAP du modèle Random Forest, poste de milieu offensif

Annexe 5 : Graphiques de prédiction par poste

Cette annexe rassemble les graphiques « prédit contre réel » des cinq postes non présentés dans le corps du texte (les postes de buteur et de milieu offensif, illustrant respectivement les situations les plus et les moins favorables, figurent aux Figure 3 et Figure 4). La lecture est identique : chaque point est un joueur, la diagonale rouge matérialise la prédiction parfaite, et la proximité du nuage à cette diagonale mesure la qualité prédictive.

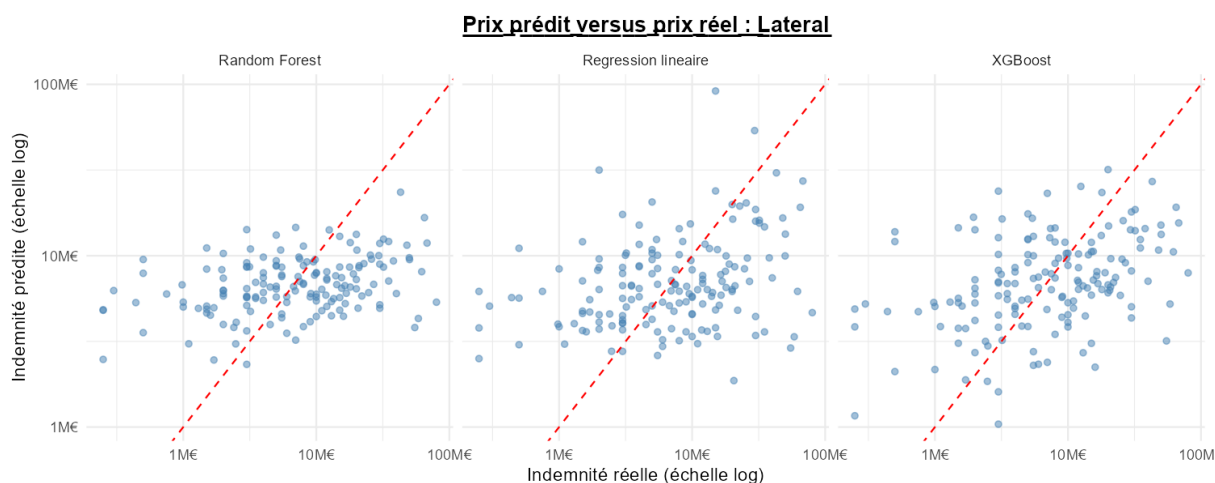


Figure 12 : Prédications contre prix réels, poste de latéral ($n = 161$)³⁷

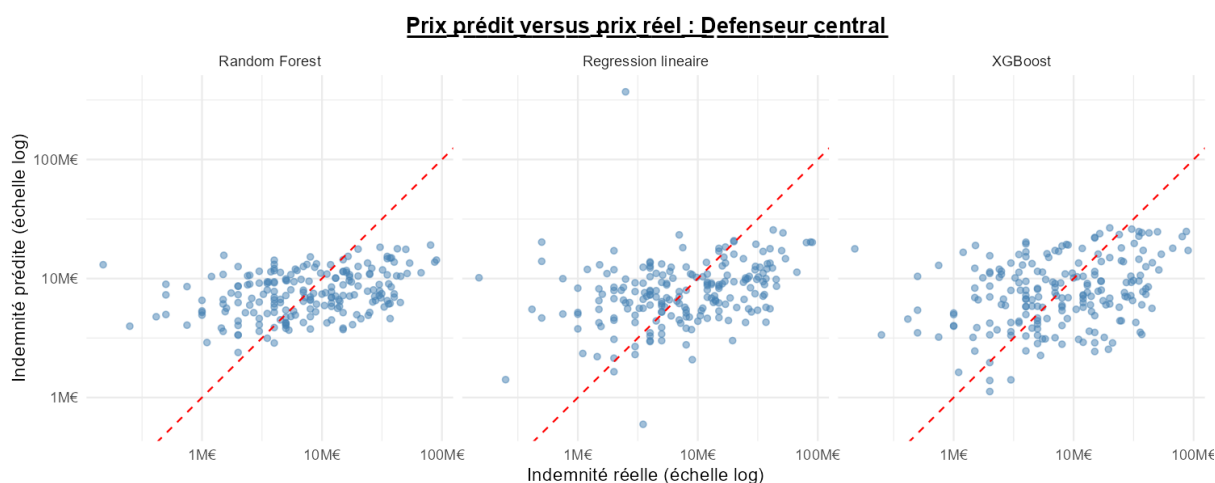


Figure 13 : Prédications contre prix réels, poste de défenseur central ($n = 203$)³⁸

³⁷ Nuage intermédiaire, cohérent avec le statut robuste de ce poste. XGBoost y obtient son unique meilleure performance.

³⁸ Nuage compact dans la zone dense ; la prédiction extrême de la régression linéaire (en haut du panneau central) correspond au cas d'école analysé en 3.5.5. Un cas d'école : la sensibilité de la régression aux valeurs extrêmes.

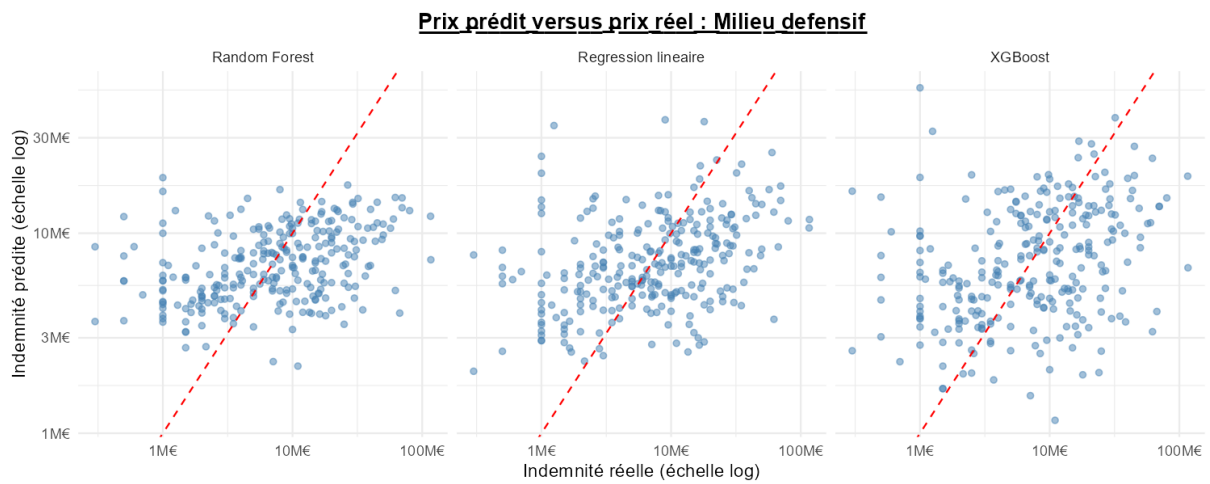


Figure 14 : Prédictions contre prix réels, poste de milieu défensif (n = 275)³⁹

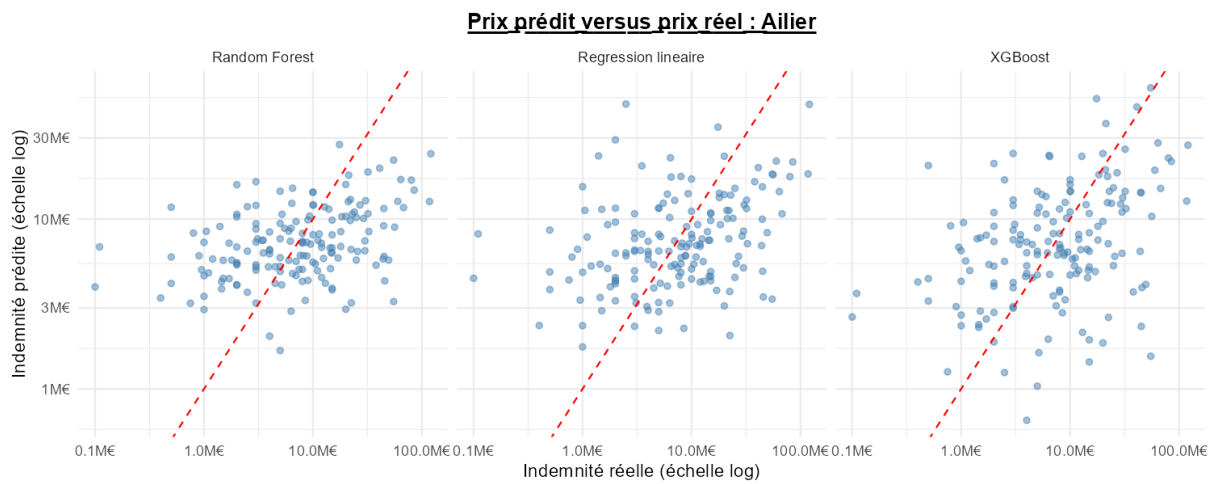


Figure 15 : Prédictions contre prix réels, poste d'ailier (n = 166)⁴⁰

³⁹ Le plus grand échantillon de l'étude ; compression modérée aux extrémités de la gamme de prix.

⁴⁰ Compression marquée pour les prix réels les plus élevés.

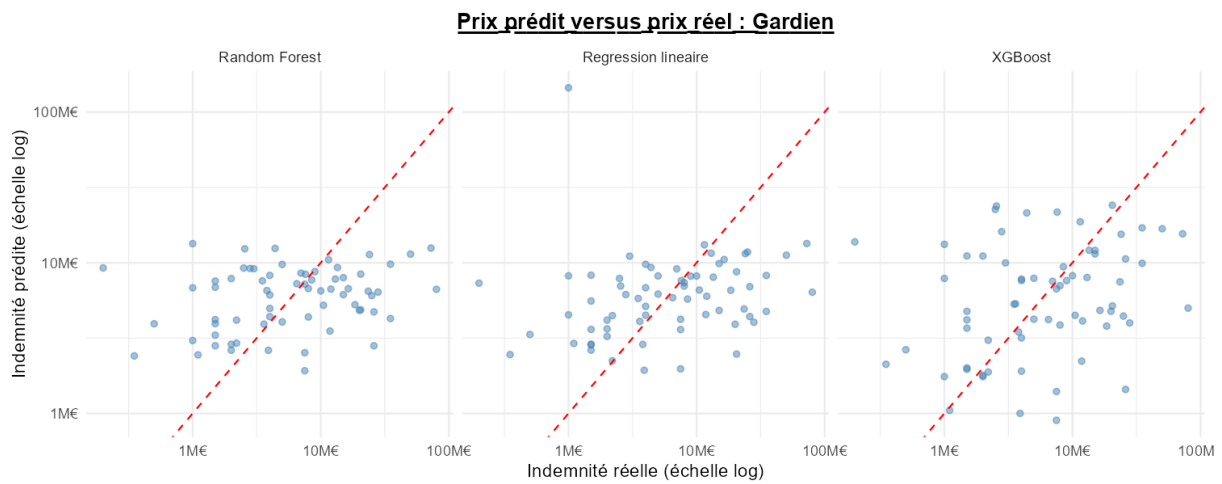


Figure 16 : Prédictions contre prix réels, poste de gardien (n = 68)⁴¹

⁴¹ Le plus petit échantillon de l'étude : nuage dispersé et fortement compressé, convergent avec l'ensemble des signaux de fragilité relevés pour ce poste.

Annexe 6 : Les méthodes rencontrées

Méthode	Fonctionnement
Régression linéaire multiple	Modélise la relation entre une variable cible et plusieurs variables explicatives en supposant une relation linéaire. Les coefficients sont estimés par les moindres carrés ordinaires. Sa limite principale est l'hypothèse de linéarité, qui peut ne pas refléter les relations réelles entre variables (Tibshirani & Friedman, s. d.).
Ridge Regression	Variante régularisée de la régression linéaire, qui ajoute une pénalité proportionnelle à la somme des carrés des coefficients à la fonction de perte minimisée. Cette régularisation réduit le risque de surapprentissage en contraignant les coefficients à rester proches de zéro, sans les annuler entièrement (Tibshirani & Friedman, s. d.).
Random Forest	Méthode d'agrégation (bagging) qui construit un grand nombre d'arbres de décision, chacun entraîné sur un échantillon bootstrap différent des données et avec un sous-ensemble aléatoire de variables à chaque nœud. La prédiction finale est la moyenne des prédictions individuelles. Cette agrégation réduit la variance et améliore la stabilité par rapport à un arbre isolé, sans supposer de forme fonctionnelle particulière pour la relation entre variables (Breiman, 2001).
Support Vector Regression (SVR)	Étend le principe des machines à vecteurs de support à la régression. Elle cherche une fonction qui s'écarte au maximum d'un seuil ε des valeurs cibles, tout en restant aussi plate que possible. Seules les observations situées au-delà de ce seuil (les vecteurs de support) contribuent à la définition de la frontière de décision (Tibshirani & Friedman, s. d.).
SVR optimisé par PSO (Particle Swarm Optimization)	Variante du SVR dans laquelle les hyperparamètres clés sont sélectionnés automatiquement par un algorithme d'optimisation (Schaden, 2026).
Gradient Boosting (XGBoost)	Méthode d'agrégation séquentielle qui construit des arbres de décision de manière itérative : chaque nouvel arbre est entraîné pour corriger les erreurs (résidus) du modèle précédent. L'optimisation repose sur une approximation au second ordre de la fonction de perte (Chen & Guestrin, 2016).

Tableau 11 : Les méthodes rencontrées

Annexe 7 : Les écarts positifs les plus importants

Joueur	Poste	Prix réel (M€)	Valeur de performance (M€)	Écart (M€)
Jude Bellingham	Milieu offensif	127,0	9,0	+118,0
Declan Rice	Milieu défensif	116,6	7,4	+109,2
Jack Grealish	Ailier	117,5	12,7	+104,8
Moisés Caicedo	Milieu défensif	116,0	12,2	+103,8
Antoine Griezmann	Buteur	120,0	20,3	+99,7
Romelu Lukaku	Buteur	113,0	13,4	+99,6
Eden Hazard	Ailier	120,8	24,3	+96,5
Kai Havertz	Buteur	100,0	15,2	+84,8
Harry Kane	Buteur	95,0	15,4	+79,6
Josko Gvardiol	Défenseur central	90,0	14,3	+75,7
Randal Kolo Muani	Buteur	95,0	20,6	+74,4
Kepa Arrizabalaga	Gardien	80,0	6,7	+73,3
Harry Maguire	Défenseur central	87,0	13,7	+73,3
Lucas Hernández	Latéral	80,0	7,9	+72,1
Dušan Vlahović	Buteur	85,4	13,8	+71,6

Tableau 12 : Les quinze transferts présentant les écarts positifs les plus importants (sur-évaluation par rapport au profil statistique)

Annexe 8 : Test du χ^2 sur les transferts exclus

Dimension	p-valeur	Décision (seuil de 5 %)	Interprétation
Poste	0,0094	H0 rejetée	Effet significatif mais modéré : taux d'exclusion de 0,9 % à 7,2 % selon le poste, les gardiens étant nettement moins touchés.
Championnat	0,056	H0 non rejetée	La Serie A ($\approx 7,0$ %) est environ deux fois plus touchée que la Premier League et la Bundesliga (3,1-3,2 %), vraisemblablement en raison des conventions de nommage (accents, caractères diacritiques) ; l'écart reste toutefois trop proche du seuil pour conclure formellement à une dépendance.
Nationalité	0,791	H0 non rejetée	Aucun écart notable par rapport à l'indépendance : l'association apparente observée dans un premier examen descriptif s'explique par la corrélation entre nationalité et championnat.

Tableau 13: Test du χ^2 sur les transferts exclus pour conflit de correspondance entre sources

Annexe 9 : Facteur dominant par poste selon la décomposition LMG

Poste	Facteur dominant	Part du R ²	IC 95 % (bootstrap)	Largeur de l'IC
Buteur	Finition / efficacité offensive	67,7 %	[27,6 ; 75,4]	47,8
Latéral	Progression balle au pied	55,2 %	[21,3 ; 68,2]	46,9
Défenseur central	Précision de passe	43,1 %	[16,8 ; 61,9]	45,1
Ailier	Précision de passe	41,0 %	[10,4 ; 61,7]	51,3
Milieu défensif	Tacles et récupération	25,3 %	[5,8 ; 44,0]	38,2
Gardien	Arrêts / solidité	44,5 %	[2,3 ; 79,3]	77,0
Milieu offensif	Pressing	76,0 %	[16,0 ; 91,0]	75,0

Tableau 14 : Facteur dominant par poste selon la décomposition LMG, avec intervalles de confiance bootstrap (1 000 répétitions)

Annexe 10 : Comparaison du facteur dominant identifié par le LMG

Poste	Facteur dominant (LMG)	Facteur dominant (SHAP)	Convergence
Buteur	Finition / efficacité offensive	Finition / efficacité offensive	Oui
Latéral	Progression balle au pied	Progression balle au pied	Oui
Défenseur central	Précision de passe	Précision de passe	Oui
Ailier	Précision de passe	Précision de passe	Oui
Milieu offensif	Pressing	Pressing	Oui
Milieu défensif	Tacles et récupération	Création / passes	Non
Gardien	Arrêts / solidité	Duels aériens	Non

Tableau 15 : Comparaison du facteur dominant identifié par le LMG (régression) et par les valeurs SHAP (modèle d'arbres)

Annexe 11 : RMSE en validation croisée

Poste	Régression linéaire	Random Forest	XGBoost	Meilleur modèle
Défenseur central	1,177	1,132	1,182	Random Forest
Latéral	1,201	1,159	1,148	XGBoost
Milieu défensif	1,141	1,138	1,189	Random Forest
Ailier	1,247	1,210	1,254	Random Forest
Buteur	1,293	1,201	1,383	Random Forest
Milieu offensif	1,259	1,215	1,274	Random Forest
Gardien	1,280	1,224	1,326	Random Forest

Tableau 16 : RMSE en validation croisée (échelle logarithmique du prix), par poste et par modèle

Annexe 12 : Synthèse de l'ACP finale par poste

Poste	n	k (variables)	Ratio n/k	KMO	Composantes retenues
Défenseur central	203	38	5,34	0,734	9
Latéral	161	40	4,03	0,741	10
Milieu défensif	275	39	7,05	0,797	9
Ailier	166	40	4,15	0,746	8
Buteur	227	39	5,82	0,731	10
Milieu offensif	72	15	4,80	0,536	5
Gardien	68	12	5,67	0,625	5

Tableau 17 : Synthèse de l'ACP finale par poste