INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Modelação do Valor dos Jogadores de Futebol no Mercado de Transferências

Vadim Tapu

Mestrado em Ciência de Dados

Orientadora:

Professora Doutora Maria Teresa Delgado Calapez, Professora Associada, ISCTE - Instituto Universitário de Lisboa

Setembro, 2025

BUSINESS SCHOOL

TECNOLOGIAS E ARQUITETURA

Departamento de Métodos Quantitativos para Gestão e Economia

Departamento de Ciências e Tecnologias da Informação

Modelação do Valor dos Jogadores de Futebol no Mercado de Transferências

Vadim Tapu

Mestrado em Ciência de Dados

Orientadora:

Professora Doutora Maria Teresa Delgado Calapez, Professora Associada, ISCTE - Instituto Universitário de Lisboa

Setembro, 2025

AGRADECIMENTOS

Quero começar por agradecer à minha família, que sempre acreditou em mim e me deu a força necessária para chegar até aqui. Aos meus pais, pelo apoio incondicional, pelos valores que me transmitiram e por estarem sempre presentes em todas as fases do meu percurso. Ao meu irmão mais novo, que com a sua alegria e energia tornou muitos momentos mais leves.

Um agradecimento muito especial à minha namorada, Marta, que esteve comigo em cada etapa deste percurso, partilhando as dificuldades e celebrando as vitórias. O teu apoio, compreensão e carinho foram fundamentais para que conseguisse superar os desafios desta fase.

Aos meus amigos, pelo companheirismo e pelas distrações necessárias nos momentos de maior pressão, deixo também um enorme obrigado.

Quero ainda agradecer aos colegas e docentes do Mestrado em Ciência de Dados do ISCTE, que enriqueceram o meu percurso académico e proporcionaram um ambiente de aprendizagem exigente e motivador.

Por fim, agradeço à Professora Doutora Maria Teresa Calapez pela orientação ao longo do desenvolvimento deste trabalho.

RESUMO

O valor de mercado dos jogadores de futebol constitui um dos indicadores mais relevantes na economia do desporto contemporâneo, refletindo não apenas o desempenho individual, mas também fatores contextuais e perceções de mercado. Apesar da sua importância crescente, a determinação deste valor continua a ser um processo complexo e multifatorial.

O presente trabalho tem como objetivo analisar e prever o valor de mercado dos avançados que atuaram nas cinco principais ligas europeias na época 2023/2024, recorrendo a diferentes algoritmos de aprendizagem automática. O estudo baseou-se num dataset construído a partir de diversas fontes, no qual se incluíram variáveis de desempenho desportivo (minutos jogados, golos, assistências), de contexto coletivo (valor agregado do clube) e de histórico individual (pico de valorização pré-época).

Os resultados evidenciam a superioridade dos algoritmos de ensemble, em particular o XGBoost, que apresentou melhor capacidade explicativa e menor erro médio face a modelos lineares ou árvores simples. Entre os fatores mais determinantes identificados, destacam-se os minutos jogados e os golos por 90 minutos, bem como variáveis contextuais como o valor do clube e a valorização histórica.

Esta dissertação contribui para a investigação académica sobre a valorização de jogadores de futebol, ao aplicar metodologias de ciência de dados a um fenómeno de elevada complexidade. Para além disso, fornece orientações práticas a clubes, agentes e investidores, permitindo compreender de forma mais clara os elementos que sustentam a valorização de avançados nas ligas europeias.

Palavras-chave: *Valor de mercado, Futebol, Avançados, Desempenho desportivo, Aprendizagem automática*

JEL Classification System: *C55, L83, Z20*

ABSTRACT

The market value of football players is one of the most relevant indicators in the contemporary sports economy, reflecting not only individual performance but also contextual factors and market perceptions. Despite its growing importance, determining this value remains a complex and multifactorial process.

This dissertation aims to analyze and predict the market value of forwards who played in the five major European leagues during the 2023/2024 season, using different machine learning algorithms. The study was based on a dataset built from multiple sources, including variables of sporting performance (minutes played, goals, assists), collective context (club's aggregated market value), and individual history (pre-season peak value).

The results highlight the superiority of ensemble algorithms, particularly XGBoost, which showed higher explanatory power and lower average error compared to linear models or single decision trees. Among the most decisive factors identified, playing time and goals per 90 minutes stand out, alongside contextual variables such as club market value and historical peak valuation.

This research contributes to the academic debate on football player valuation by applying data science methodologies to a highly complex phenomenon. In addition, it provides practical insights for clubs, agents, and investors, offering a clearer understanding of the elements that influence the valuation of forwards in European leagues.

Keywords: *Market value, Football, Forwards, Sports performance, Machine learning*
JEL Classification System: *C55, L83, Z20*

ÍNDICE GERAL

AGRADECIMENTOS v
RESUMO vii
ABSTRACT ix
ÍNDICE GERAL xi
ÍNDICE DE TABELAS xiii
ÍNDICE DE FIGURAS xv
GLOSSÁRIO DE SIGLAS xvii
I. Introdução 1
II. Revisão de Literatura 3
2.1 O futebol como indústria global e a dinâmica do mercado de transferências 3
2.2 Fatores determinantes do valor de mercado de jogadores 4
2.3 Tecnologia e inovação na análise do desempenho desportivo 6
2.4 Aprendizagem automática e modelação preditiva no futebol 8
2.5 Estudos empíricos no futebol profissional 10
2.6 Estatística descritiva e métricas no futebol 12
2.7 O papel das organizações internacionais 14
2.8 Limitações e desafios na literatura 15
III. Hipóteses de Investigação e Modelo Conceptual 17
3.1 Hipóteses de Investigação 17
3.2 Modelo Conceptual de Investigação 18
IV. Metodologia 19
4.1 Business Understanding 20
4.2 Data Understanding 21
4.2.1 Distribuição etária e utilização 22
4.2.2 Nacionalidades e composição da amostra 23
4.2.3 Mercado: nível e evolução 24
4.2.4 Desempenho Individual Ofensivo 26
4.2.5 Relação entre desempenho e valor de mercado 27
4.2.6 Correlações entre variáveis 28
4.3 Data Preparation 30

4.3.1 Tratamento de valores omissos 31
4.3.2 Criação de novas variáveis 32
4.3.3. Transformações finais para modelação 34
4.3.4. Construção dos conjuntos de treino e teste 36
4.4 Modelação 37
4.5 Evaluation 39
V. Resultados e Discussão 41
5.1 Análise Global dos Resultados 41
5.1.1 Desempenho Global dos Modelos 41
5.1.2 Experiência 1 – Exploração incremental de variantes 44
5.1.3 Experiência 2 — Robustez e ablações do modelo final 46
VI. Conclusões 51
REFERÊNCIAS BIBLIOGRÁFICAS 55
ANEXOS 60
Anexo A – Pipeline principal 60
Anexo B – Exploração adicional 63
Anexo C – Experiência 2 (robustez & ablations) 66
Anexo D – Configuração dos modelos e validação 70
Anexo E – Figuras complementares 72

ÍNDICE DE TABELAS

Tabela 5.1. Desempenho dos modelos preditivos 42
Tabela 5.2. Métricas de desempenho do XGBoost..... 45

ÍNDICE DE FIGURAS

Figura 3.1. Modelo Conceptual de Investigação. 18
Figura 4.1. Metodologia CRISP-DM. 19
Figura 4.2. Relação entre idade, posição e utilização dos jogadores. 23
Figura 4.3. Nacionalidades mais representadas entre os avançados das cinco principais ligas europeias (época 2023/24). 24
Figura 4.4. Top 10 avançados mais valiosos das cinco principais ligas europeias (época 2023/2024). 25
Figura 4.5. Evolução do valor de mercado dos cinco avançados mais valiosos das ligas Big-5 (2020–2025). 26
Figura 4.6. Top 5 de avançados com mais golos e assistências na época 2023/2024 nas cinco principais ligas europeias. 27
Figura 4.7. Associação entre métricas de desempenho (golos e minutos) e valor de mercado na época 2023/2024. 28
Figura 4.8. Matriz de correlação. 29
Figura 4.9. Comparação entre valor atual e pico pré-época para Harry Kane e Bukayo Saka. 33
Figura 4.10. Efeito da transformação logarítmica nos valores de mercado. 35
Figura 4.11. Pipeline metodológico da investigação. 39
Figura 5.1. Avaliação do XGBoost: variáveis, previsões e desempenho. 46
Figura 5.2. Robustez dos modelos: R^2 e RMSE médio. 47
Figura 5.3. Resultados das ablations no modelo XGBoost (D_w). 48
Figura 5.4. Curvas de dependência SHAP para as principais variáveis do modelo XGBoost (D_w). 49

GLOSSÁRIO DE SIGLAS

AI – Artificial Intelligence

CRISP-DM – *Cross Industry Standard Process for Data Mining*

DF – *Defender*

FW – *Forward*

GK – *Goalkeeper*

MAE – *Mean Absolute Error*

MF – *Midfield*

ML – *Machine Learning*

R^2 – Coeficiente de Determinação

RMSE – *Root Mean Squared Error*

RMSLE – *Root Mean Squared Logarithmic Error*

SHAP – *SHapley Additive exPlanations*

XGBoost – *Extreme Gradient Boosting*

I. Introdução

Nas últimas décadas, o futebol afirmou-se como a principal indústria do entretenimento desportivo, combinando uma dimensão cultural única com um impacto económico cada vez mais relevante. Como referem Dobson e Goddard (2011), a economia associada ao futebol europeu movimenta anualmente milhares de milhões de euros, refletindo não apenas a paixão pelo jogo, mas também a importância crescente do mercado de transferências como mecanismo de circulação de talento e capital. Neste contexto, compreender e estimar o valor de mercado dos jogadores tornou-se um tema central, com implicações tanto para a gestão desportiva como para a sustentabilidade financeira dos clubes (Müller et al., 2017).

Determinar o valor de um jogador é, contudo, um processo complexo. Herm et al. (2014) sublinham que o preço de transferência não depende apenas da performance desportiva, mas resulta de uma conjugação de múltiplos fatores, que incluem idade, características físicas, estatísticas de desempenho, reputação, potencial de progressão, contexto competitivo e até aspetos externos como a exposição mediática ou as necessidades estratégicas de um clube. Essa multiplicidade de variáveis gera discrepâncias frequentes entre valores estimados e valores efetivamente praticados, aumentando o risco financeiro de cada transação. Daí a procura crescente por metodologias quantitativas, mais objetivas e replicáveis, para apoiar este processo.

Importa ainda notar que os valores de mercado reportados em plataformas como o Transfermarkt são estimativas baseadas em critérios públicos e consensuais, não correspondendo a valores contratuais oficiais, pelo que podem divergir entre fontes distintas

É neste enquadramento que a ciência de dados, e em particular a aprendizagem automática (Machine Learning, ML), tem ganho protagonismo. Como apontam Rossi et al. (2018), a aplicação de técnicas preditivas no desporto permite identificar padrões não lineares e relações entre variáveis dificilmente detetadas por modelos estatísticos tradicionais. Algoritmos como Random Forest ou Gradient Boosting revelaram já capacidade para lidar com grandes volumes de informação e captar interações complexas, oferecendo resultados promissores (Gudmundsson & Horton, 2017). Este movimento é favorecido pela disponibilidade crescente de bases de dados que reúnem, de forma sistemática, indicadores de performance e informação financeira.

A presente dissertação centra-se na previsão do valor de mercado de jogadores que atuam como avançados em ligas europeias. A escolha desta posição prende-se com a sua relevância estratégica e com o facto de os avançados estarem geralmente associados a transferências de

elevado montante, funcionando como ativos de grande peso económico. Para tal, desenvolveu-se um pipeline em R, no qual se aplicaram e compararam diferentes algoritmos de ML – regressão linear, Elastic Net, Random Forest e XGBoost – complementados por experiências adicionais que exploram variáveis derivadas, ponderação de observações e testes de robustez. A avaliação foi conduzida através de métricas amplamente aceites na literatura (RMSE, MAE, R^2), de forma a comparar desempenhos e a compreender a importância relativa das variáveis.
Neste quadro, a investigação procura responder a algumas questões fundamentais:
QI1: Quais os algoritmos de aprendizagem automática que melhor conseguem captar a complexidade da determinação do valor de mercado dos avançados?
QI2: Quais as variáveis mais influentes na explicação desse valor e de que modo a sua relevância varia entre modelos?
QI3: Até que ponto os modelos são robustos a diferentes configurações e variações nos dados?
QI4: De que forma métricas de performance desportiva se articulam com fatores contextuais, como o valor agregado do clube ou o pico de valorização prévia, na determinação do valor de mercado?

O objetivo central é, assim, desenvolver e avaliar modelos de Machine Learning capazes de prever o valor de mercado com elevada precisão, fornecendo simultaneamente uma análise das variáveis que mais contribuem para essas previsões. Pretende-se oferecer não apenas um contributo académico, mas também resultados que possam ter aplicabilidade prática no contexto da gestão desportiva e financeira.
A dissertação está organizada em seis capítulos. Após esta introdução, o segundo capítulo apresenta a revisão da literatura, abordando os principais contributos sobre a avaliação de jogadores e a aplicação da aprendizagem automática no desporto. O terceiro capítulo expõe as hipóteses de investigação e o modelo conceptual que orienta o estudo. O quarto capítulo descreve a metodologia, incluindo a preparação dos dados, a construção das variáveis e a seleção dos algoritmos. O quinto capítulo apresenta os resultados, analisando as diferentes experiências e comparando desempenhos. Por fim, o sexto capítulo reúne as conclusões, destacando as principais contribuições do estudo, as limitações encontradas e sugestões para investigações futuras.

II. Revisão de Literatura

Neste capítulo discute-se a literatura relevante para o estudo, começando pelo enquadramento económico do futebol e avançando para os fatores que influenciam o valor de mercado e as contribuições da ciência de dados. Num primeiro momento, aborda-se o enquadramento económico e estrutural do futebol enquanto indústria global, destacando a evolução do mercado de transferências e os fatores que influenciaram a sua dinâmica. Seguidamente, analisam-se os determinantes do valor de mercado dos jogadores, com particular incidência nas dimensões de desempenho, características biográficas e contexto competitivo. Posteriormente, explora-se a transformação digital do futebol e o papel cada vez mais central dos dados na avaliação de atletas e na tomada de decisão estratégica. O capítulo continua com a análise da transição da estatística tradicional para a ciência de dados, evidenciando a emergência do *sports analytics* e a consolidação das abordagens de aprendizagem automática no desporto. Por fim, apresentam-se os principais estudos empíricos que aplicam técnicas de modelação preditiva ao futebol, destacando-se as contribuições mais relevantes da literatura internacional.

A revisão da literatura incidiu em trabalhos publicados entre 2000 e 2025, com especial atenção à produção científica mais recente, marcada pelo crescimento da disponibilidade de dados desportivos e pela evolução das técnicas de aprendizagem automática. A pesquisa bibliográfica foi realizada em bases científicas de referência, como *Scopus*, *Web of Science* e *Google Scholar*, complementada por relatórios oficiais da FIFA e da UEFA. Foram privilegiados os estudos que analisam os fatores determinantes do valor de mercado, as transformações no mercado de transferências, o impacto da tecnologia e o uso de metodologias preditivas no futebol profissional.

2.1 O futebol como indústria global e a dinâmica do mercado de transferências

O futebol consolidou-se como a modalidade desportiva mais popular do planeta, reunindo mais de 3,5 mil milhões de adeptos espalhados pelos cinco continentes (FIFA, 2023). Para além da sua relevância cultural e social, o desporto assume hoje uma dimensão económica de enorme impacto, afirmando-se como uma das maiores indústrias globais de entretenimento. O relatório *Panorama do Futebol de Clubes Europeus* da UEFA (2023) estima que, em 2022, as receitas conjuntas dos clubes europeus ultrapassaram os 22 mil milhões de euros, resultado sobretudo dos direitos de transmissão televisiva, de contratos de patrocínio e da bilheteira. Neste cenário,

o mercado de transferências emerge como uma das engrenagens fundamentais deste sistema económico.

Criado originalmente como um mecanismo regulatório para assegurar estabilidade contratual e equilíbrio competitivo entre clubes (Frick, 2007), o sistema de transferências transformou-se ao longo do tempo numa arena financeira de grande dinamismo. De acordo com estimativas da FIFA (2023), só em 2022 registaram-se mais de 20.000 transferências internacionais, envolvendo montantes superiores a 6,5 mil milhões de dólares, concentrados sobretudo nas principais ligas europeias.

A evolução deste mercado foi moldada por mudanças regulatórias de grande alcance, como a célebre decisão do Tribunal de Justiça Europeu no caso *Bosman* (1995), que liberalizou a mobilidade de jogadores em fim de contrato (Binder & Findlay, 2012). Essa alteração promoveu a globalização do talento e intensificou a concorrência entre clubes de diferentes ligas, reforçando, ao mesmo tempo, a valorização individual dos jogadores enquanto ativos financeiros (Dobson & Goddard, 2011).

Hoje, os valores praticados nas transferências refletem não apenas o desempenho desportivo, mas também fatores associados ao marketing, à visibilidade mediática e ao potencial de retorno comercial (Müller et al., 2017). Clubes de maior dimensão, com maior capacidade financeira, tornaram-se dominantes neste processo, aprofundando desigualdades estruturais no futebol europeu (UEFA, 2023). Paralelamente, o crescimento da indústria deu origem a um ecossistema mais alargado, que inclui agentes desportivos, fundos de investimento e plataformas digitais, todos com um papel ativo na definição do valor de mercado dos atletas (Poli et al., 2020).

Neste contexto, o jogador de futebol deixou de ser visto apenas como um atleta em competição e passou a ser encarado como um ativo estratégico dentro de uma lógica de mercado e de investimento (Franceschi et al., 2023). O conceito de "capital humano desportivo" traduz esta perspetiva, ao reunir desempenho, potencial e valor económico, sendo essencial para compreender o funcionamento atual do mercado de transferências (Herm et al., 2014).

2.2 Fatores determinantes do valor de mercado de jogadores

O valor de mercado de um jogador resulta da combinação de fatores desportivos, biográficos, económicos e institucionais. Ao contrário dos ativos financeiros tradicionais, cujo preço pode ser explicado por modelos de risco e retorno relativamente objetivos, os jogadores representam ativos singulares, cujo valor integra tanto o desempenho observado em campo como

expectativas futuras de rendimento, aspetos de marketing e perceções subjetivas de mercado (Müller et al., 2017).

Entre os fatores mais estudados destaca-se a idade, frequentemente associada a uma relação não linear com o valor de mercado, descrita pela literatura como uma curva em formato de "U invertido" (Frick, 2007; Herm et al., 2014). Jogadores muito jovens ainda não confirmaram o seu potencial, enquanto atletas em fases mais avançadas da carreira enfrentam limitações físicas e expectativas de declínio. O pico de valorização tende a situar-se entre os 24 e os 28 anos, considerado o período de maior rendimento físico e técnico (Dobson & Goddard, 2011).

A posição em campo constitui outro elemento diferenciador. Avançados alcançam, em regra, valores de mercado superiores devido à sua associação direta com a marcação de golos e, por consequência, com os resultados coletivos (Franceschi et al., 2023). Em contraste, guarda-redes e defesas, ainda que desempenhem papéis cruciais para o sucesso das equipas, tendem a registar valores mais modestos, refletindo a lógica mediática e comercial que privilegia jogadores ofensivos (Majewski, 2021).

No que diz respeito às características de performance, estatísticas como minutos jogados, golos e assistências surgem recorrentemente como determinantes centrais em estudos empíricos (Herm et al., 2014; Müller et al., 2017). Nos últimos anos, a introdução de métricas avançadas, como expected goals (xG) e expected assists (xA), permitiu capturar com maior precisão a contribuição de cada jogador, ao corrigir os indicadores para a qualidade das oportunidades criadas ou finalizadas (Gudmundsson & Horton, 2017). Resultados semelhantes foram encontrados em outros contextos: He, Guo e Hu (2015), ao estudarem a liga chinesa, demonstraram que o desempenho em campo tem um impacto direto no valor de mercado dos jogadores. De forma complementar, Hofmann (2018) confirmou, através de uma abordagem econométrica, que variáveis como idade, posição e estatísticas individuais constituem elementos centrais na explicação das avaliações de mercado.

Também os fatores contratuais exercem uma influência significativa. Jogadores com contratos de longa duração tendem a apresentar maior valorização de mercado, refletindo a posição negocial mais favorável do clube detentor dos direitos desportivos (Binder & Findlay, 2012). A existência de cláusulas de rescisão, bem como particularidades legais de determinados campeonatos, pode ainda condicionar os valores efetivamente praticados.

Outro determinante amplamente reconhecido é o contexto institucional e competitivo. Jogadores pertencentes a clubes de elevado prestígio internacional e forte capacidade financeira beneficiam de um efeito indireto de valorização, associado ao renome da instituição e à

visibilidade mediática adicional (Müller et al., 2017; Poli et al., 2020). De igual modo, a participação em competições internacionais de relevo, como a Liga dos Campeões da UEFA ou o Campeonato do Mundo da FIFA, potencia a valorização dos atletas devido ao aumento da sua exposição global (FIFA, 2023; UEFA, 2023).

Por último, fatores externos e intangíveis, relacionados com a imagem e a popularidade, têm ganho importância crescente. Indicadores como o número de seguidores em redes sociais, a presença mediática ou as tendências de pesquisa online mostram correlação positiva com o valor de mercado, refletindo a interdependência entre desporto, economia da atenção e estratégias de marketing (Majewski, 2021; Franceschi et al., 2023).

Em suma, o valor de mercado de um jogador resulta da combinação entre variáveis objetivas, mensuráveis e fatores de natureza subjetiva e contextual. É precisamente esta complexidade que tem levado a um maior recurso a abordagens de ciência de dados e de aprendizagem automática, uma vez que estas permitem trabalhar com relações não lineares e integrar múltiplas dimensões no processo de avaliação.

2.3 Tecnologia e inovação na análise do desempenho desportivo

A digitalização do futebol intensificou-se de forma significativa na última década, transformando não apenas os processos de treino e competição, mas também a forma como os clubes gerem os seus recursos e definem estratégias. Atualmente, o recurso a dados objetivos e a tecnologias de monitorização é considerado um elemento central para aumentar a competitividade, reduzir riscos e otimizar o retorno sobre o investimento em jogadores (Morgulev et al., 2018).

O desenvolvimento de sistemas de recolha e análise de dados esteve no centro desta transformação. Empresas como Opta, Wyscout, StatsBomb e InStat disponibilizam bases de dados altamente detalhadas, que incluem desde métricas tradicionais — como minutos jogados, passes, remates, golos e assistências — até indicadores avançados como *expected goals* (xG), *expected assists* (xA) ou métricas de pressão defensiva (Gudmundsson & Horton, 2017). Estas informações, recolhidas em tempo real e de forma padronizada, passaram a desempenhar um papel crucial tanto na avaliação do desempenho individual e coletivo como na determinação do valor de mercado de atletas.

Em paralelo, a introdução de tecnologias de rastreamento por GPS e de sensores biométricos revolucionou a monitorização da condição física. Sistemas de posicionamento e câmaras multicâmara permitem hoje recolher milhões de pontos de dados em cada jogo ou sessão de treino, fornecendo informação detalhada sobre distâncias percorridas, acelerações, carga de

trabalho, fadiga e risco de lesão (Carling et al., 2014). Estes dados têm sido incorporados em processos de preparação física e em análises que associam diretamente a condição atlética ao desempenho em campo e, de forma indireta, ao valor económico dos jogadores.

A adoção destas tecnologias tem vindo a crescer em diferentes contextos competitivos. Estudos recentes demonstram que a perceção de utilidade e a abertura à inovação são fatores mais determinantes para a sua implementação do que a própria complexidade tecnológica (Al-Mannai, 2025). Do mesmo modo, avanços em *computer vision* permitiram desenvolver soluções comerciais capazes de extrair métricas fiáveis de posição e velocidade a partir de imagens televisivas, com margens de erro relativamente reduzidas (Crang et al., 2025). Estes progressos tornam a recolha de dados acessível a clubes com diferentes recursos financeiros.

Este ecossistema cada vez mais sofisticado tem promovido a passagem de um modelo de decisão assente na intuição para um paradigma *data-driven*, em que algoritmos e métricas objetivas assumem crescente importância. Como salientam Rossi et al. (2018), esta evolução abre caminho à aplicação de métodos avançados de aprendizagem automática, capazes de explorar relações não lineares entre variáveis e de antecipar cenários com maior precisão.

O impacto económico da digitalização também é expressivo. De acordo com o relatório *UEFA European Club Finance and Investment Landscape 2024*, as receitas combinadas dos principais clubes europeus ultrapassaram os €29 mil milhões em 2024, mais €10 mil milhões do que uma década antes (UEFA, 2024). Este crescimento encontra-se associado à modernização das práticas de gestão, nas quais ferramentas de *business intelligence* e *data analytics* assumem papel cada vez mais relevante.

No plano do mercado tecnológico, estima-se que o setor dos sistemas de rastreamento de jogadores tenha atingido um valor de 7,02 mil milhões de dólares em 2024, com projeções que apontam para 20,44 mil milhões até 2030 (Grand View Research, 2024). A expansão destes sistemas reflete a sua importância não apenas na monitorização de desempenho, mas também na prevenção de lesões, no planeamento estratégico e na valorização económica dos ativos desportivos.

Apesar destes avanços, subsistem desafios significativos. Muitos clubes ainda carecem de estruturas internas adequadas para gerir dados de forma consistente, revelando fragilidades na organização da informação e na capacidade de aplicar metodologias analíticas avançadas (Lolli, 2024). Esta distância entre a disponibilidade de dados e a sua utilização prática constitui uma barreira à plena integração da inovação tecnológica no futebol.

Em síntese, o futebol contemporâneo atravessa uma fase de viragem: a tecnologia deixou de ser um recurso complementar para se afirmar como um fator competitivo determinante. A capacidade de recolher, interpretar e aplicar dados complexos influencia diretamente o desempenho desportivo, ao mesmo tempo que condiciona a valorização de jogadores e a sustentabilidade financeira dos clubes no mercado global.

2.4 Aprendizagem automática e modelação preditiva no futebol

O estudo do valor de mercado no futebol tem evoluído de forma clara ao longo das últimas duas décadas, acompanhando a disponibilidade crescente de dados e a sofisticação dos métodos estatísticos. Durante muitos anos, os trabalhos mais relevantes baseavam-se em modelos econométricos tradicionais, como regressões lineares ou log-lineares, que ajudaram a identificar variáveis com impacto direto no valor de mercado, como a idade, os minutos jogados, os golos ou o prestígio do clube (Frick, 2007; Herm et al., 2014). Apesar de úteis, estas abordagens apresentavam limitações evidentes: assumiam relações lineares e homogéneas, quando o mercado de transferências se caracteriza por forte heterogeneidade, volatilidade e interações difíceis de captar por modelos simples.

É neste contexto que a aprendizagem automática (Machine Learning, ML) ganhou protagonismo. Ao contrário da econometria clássica, os algoritmos de ML conseguem lidar com relações não lineares e conjuntos de dados com muitas variáveis, de natureza diversa. Técnicas como Random Forest ou Gradient Boosting, em particular o XGBoost e o LightGBM, tornaram-se frequentes na literatura, sobretudo pela capacidade de identificar padrões complexos e oferecer ganhos consistentes em métricas como RMSE, MAE ou R^2, quando comparados com modelos lineares (Gudmundsson & Horton, 2017; Rossi et al., 2018). A sua aplicação ao futebol revelou-se eficaz em diferentes contextos, desde a previsão de valores de mercado até à comparação de desempenhos entre posições ou ligas.

Em paralelo, métodos de regressão regularizada, como Ridge, Lasso ou Elastic Net, também têm sido explorados, sobretudo quando se procura equilibrar parcimónia, interpretabilidade e desempenho preditivo. Estes modelos, apesar de menos flexíveis que os baseados em árvores, permitem reduzir o risco de sobreajustamento e oferecem resultados de leitura mais transparente, o que continua a ser valorizado na investigação académica e na prática profissional (Müller et al., 2017).

Nos últimos anos, verificou-se ainda a entrada de redes neuronais artificiais e, em alguns casos, de arquiteturas de deep learning. Estas abordagens mostram potencial quando existem

grandes volumes de dados ou quando se pretende integrar variáveis muito distintas — como estatísticas de desempenho, dados contratuais ou indicadores de popularidade. No entanto, os estudos também apontam as suas limitações: a reduzida dimensão de muitas bases de dados desportivas dificulta o treino destas arquiteturas, e a baixa interpretabilidade das previsões limita a sua adoção prática (Estimation of Market Values of Football Players, 2021; Majewski, 2022). Trabalhos recentes confirmam, ainda assim, o valor destas técnicas: Inan e Cavas (2021) aplicaram redes neuronais no contexto da liga turca com resultados promissores, enquanto Aydemir, Taskaya-Temizel e Temizel (2022) demonstraram que um modelo de *ensemble learning* consegue superar métodos tradicionais em termos de precisão preditiva. Putra et al. (2023) analisaram jogadores emprestados da Premier League, identificando fatores relevantes para a sua valorização ao final do empréstimo, e Sun e Gu (2024) mostraram que a combinação de dados do videojogo FIFA com estatísticas reais pode reforçar substancialmente a capacidade preditiva dos modelos.

Um aspeto frequentemente sublinhado pela literatura é a importância da escolha e transformação das variáveis. Fatores clássicos como idade, minutos jogados, golos ou assistências continuam a ter grande relevância, mas têm sido acrescentadas novas dimensões. Entre elas encontram-se proxies de popularidade, como seguidores em redes sociais ou pesquisas online, e fatores de contexto competitivo, como o valor total do clube ou a participação em competições internacionais (Majewski, 2022; Franceschi et al., 2023; UEFA, 2024). A engenharia de variáveis surge, assim, como um dos pontos centrais para melhorar a qualidade dos modelos.

Também a definição da variável dependente tem sido ajustada. Vários trabalhos optam por aplicar o logaritmo ao valor de mercado, de forma a corrigir a forte assimetria da distribuição dos preços e reduzir o impacto de casos extremos. Esta prática facilita a comparação entre jogadores de diferentes segmentos de valor e estabiliza os resíduos dos modelos (Herm et al., 2014; Franceschi et al., 2023).

No que toca à avaliação dos modelos, as métricas mais utilizadas incluem o Root Mean Squared Error (RMSE), o Mean Absolute Error (MAE) e o coeficiente de determinação (R^2), por vezes complementadas pelo Root Mean Squared Logarithmic Error (RMSLE), adequado a modelos com alvo em escala logarítmica (Rossi et al., 2018). Além disso, são comuns práticas de validação robustas, como validação cruzada, utilização de múltiplas sementes ou análise por segmentos de valor e posição, que procuram assegurar que os resultados não são específicos a uma amostra restrita (Franceschi et al., 2023).

Outro tema em crescimento é a explicabilidade dos modelos. Embora algoritmos complexos apresentem boas previsões, a sua aceitação depende também da clareza com que se consegue explicar por que razão um jogador é valorizado de determinada forma. Ferramentas como os valores SHAP ou o método LIME têm ganho destaque, permitindo identificar a contribuição individual de variáveis como idade, minutos ou valor do clube para cada previsão (Lundberg & Lee, 2017; Rossi et al., 2018). Esta vertente aproxima os modelos da realidade prática, em que decisores procuram previsões mas também justificações para fundamentar negociações.

Os estudos mais recentes confirmam esta evolução: depois de anos dominados por regressões lineares, a investigação passou a integrar técnicas de aprendizagem automática que já fazem parte das análises sobre o mercado de transferências.

2.5 Estudos empíricos no futebol profissional

A investigação empírica sobre a valorização de jogadores de futebol consolidou-se nas últimas duas décadas, acompanhando a crescente disponibilidade de bases de dados e o interesse académico e profissional na compreensão das dinâmicas do mercado de transferências. Os primeiros estudos recorreram sobretudo a modelos econométricos tradicionais, com destaque para regressões lineares e log-lineares, que permitiram estabelecer relações entre variáveis como idade, posição em campo, tempo de jogo ou prestígio do clube e os valores efetivamente praticados em transferências. Estes trabalhos pioneiros identificaram padrões fundamentais, como o impacto não linear da idade — com valorização em idades intermédias e subsequente depreciação — e a relevância do desempenho desportivo e da pertença a clubes de maior dimensão na explicação das diferenças de preço entre jogadores (Frick, 2007; Herm et al., 2014).

Com o avanço tecnológico e o acesso a dados mais estruturados e precisos, a literatura passou a incorporar metodologias de aprendizagem automática, capazes de lidar com relações não lineares e interações complexas. Algoritmos baseados em árvores de decisão, como Random Forest e XGBoost, assumiram papel de destaque ao evidenciarem ganhos consistentes de desempenho face às abordagens tradicionais, nomeadamente em métricas como Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) e coeficiente de determinação (R^2). Em diferentes aplicações — desde a previsão de valores de mercado em ligas específicas até à análise comparativa entre posições ou segmentos etários — estes modelos demonstraram maior capacidade de generalização e adaptação a contextos heterogéneos (Gudmundsson & Horton, 2017; Rossi et al., 2018).

Estudos recentes exploraram igualmente a utilização de redes neuronais artificiais e modelos híbridos, sobretudo em cenários que requerem a integração de variáveis de diferentes naturezas. Estes trabalhos destacam que proxies de visibilidade, como seguidores em redes sociais, menções mediáticas ou tendências de pesquisa em motores de busca, contribuem de forma significativa para explicar a valorização de jogadores, revelando a importância de dimensões que extravasam o desempenho puramente desportivo (Majewski, 2022; Franceschi et al., 2023). Paralelamente, a inclusão de fatores contratuais — como a duração do vínculo e a existência de cláusulas de rescisão — revelou-se determinante para compreender o poder de negociação entre clubes e a fixação final dos valores transacionados (Poli et al., 2022).

Outra vertente analisada pela literatura foi a utilização de dados provenientes de videojogos de simulação, em particular a série FIFA da EA Sports, que oferece classificações padronizadas e comparáveis para milhares de jogadores. Estes dados têm sido utilizados como proxies de desempenho e potencial, permitindo colmatar lacunas em bases oficiais e facilitar análises comparativas entre diferentes ligas e contextos competitivos. Estudos que recorreram a estas métricas testaram a sua validade preditiva face aos valores reais de mercado, concluindo que, apesar das limitações associadas ao caráter subjetivo das avaliações, a sua integração em modelos de machine learning acrescenta poder explicativo e melhora a capacidade preditiva (Müller et al., 2017; Zarka et al., 2024). Quando combinadas com estatísticas oficiais e indicadores financeiros, estas variáveis captam dimensões adicionais relacionadas com perceções de talento e potencial futuro, refletindo-se no comportamento do mercado (Majewski, 2022; Franceschi et al., 2023).

O impacto de choques externos, como a pandemia de COVID-19, também foi analisado empiricamente. Os estudos evidenciam que as restrições financeiras impostas aos clubes provocaram uma contração significativa nos valores de transferência, ainda que com efeitos diferenciados entre ligas e posições. Estes resultados confirmam que fatores exógenos podem alterar de forma súbita as dinâmicas de valorização, reforçando a necessidade de modelos capazes de se adaptar a cenários de elevada incerteza (Franceschi et al., 2023). Nesta linha, Quansah, Frick, Lang e Maguire (2021) analisaram especificamente a Premier League e demonstraram que a redução das receitas durante a pandemia teve reflexos diretos não apenas nos valores de transferência, mas também na estrutura salarial dos clubes.

Em termos metodológicos, observa-se uma crescente adoção do logaritmo do valor de mercado como variável dependente, prática que visa mitigar a assimetria da distribuição e estabilizar os resíduos, facilitando a comparação de erros entre segmentos distintos. Estratégias como

validação cruzada, utilização de múltiplas sementes e análises segmentadas por faixas de valor ou por posição têm sido amplamente utilizadas para reforçar a robustez e a generalização dos resultados (Herm et al., 2014; Rossi et al., 2018).

Diversos estudos recentes confirmam padrões consistentes, a idade confirma-se como um fator determinante, com efeito não linear que aponta para um pico de valorização entre os 25 e os 28 anos, seguido de depreciação progressiva (Frick, 2007; Herm et al., 2014). O tempo de jogo e as métricas ofensivas normalizadas por 90 minutos, como golos e assistências, surgem de forma recorrente como preditores robustos, sobretudo para avançados (Gudmundsson & Horton, 2017; Rossi et al., 2018). Fatores contextuais, como o valor agregado do clube, a participação em competições internacionais e a exposição mediática, completam este quadro, revelando impacto consistente nas avaliações realizadas pelo mercado (Majewski, 2022; Franceschi et al., 2023; UEFA, 2024). A este respeito, Poli, Besson e Ravenel (2024) contribuíram com uma análise estatística global dos padrões de transferências, permitindo identificar diferenças estruturais entre ligas e contextos nacionais, enquanto Hoey, Peeters e Principe (2021) discutiram o sistema europeu de transferências como um potencial mecanismo de não-aliciamento, com implicações diretas para a competitividade do setor.

2.6 Estatística descritiva e métricas no futebol

A análise estatística do futebol tem acompanhado a evolução da própria modalidade, passando de registos básicos para sistemas de medição cada vez mais sofisticados. Nos primeiros estudos, predominavam métricas absolutas, como o número total de golos, assistências ou minutos jogados durante uma época. Apesar de informativas, estas medidas apresentavam limitações evidentes, já que não permitiam comparações justas entre jogadores com tempos de utilização distintos ou que atuavam em ligas de níveis competitivos diferentes (Frick, 2007; Herm et al., 2014).

Com o avanço das bases de dados e da análise de desempenho, começaram a ser utilizadas métricas normalizadas, como golos e assistências por 90 minutos. Este tipo de indicador tornou-se essencial para ajustar a produtividade ao tempo efetivo em campo e permitir comparações mais fiéis entre jogadores (Müller et al., 2017). Em paralelo, a disseminação de plataformas como Opta, StatsBomb ou Wyscout permitiu o acesso a estatísticas padronizadas em larga escala, promovendo análises mais consistentes entre competições (UEFA, 2024).

Um dos desenvolvimentos mais significativos foi a introdução de métricas baseadas em expectativas. Os expected goals (xG) e expected assists (xA), hoje amplamente usados, procuram medir a probabilidade de uma oportunidade resultar em golo ou assistência, considerando fatores como a distância da baliza, o ângulo de remate ou a pressão defensiva no momento da ação (Gudmundsson & Horton, 2017; Franceschi et al., 2023). Estas medidas oferecem uma visão mais rigorosa da qualidade das ações ofensivas, ajudando a reduzir a influência de variáveis aleatórias, como a eficácia pontual de um jogador.

A literatura também passou a valorizar métricas defensivas e de construção de jogo. Indicadores como desarmes, interceções, duelos ganhos, passes progressivos ou passes para zonas de perigo passaram a ser registados de forma sistemática, revelando-se fundamentais para avaliar o contributo de jogadores em posições menos associadas a golos mas determinantes no equilíbrio tático de uma equipa (Müller et al., 2017).

Outro campo que ganhou relevância foi o das métricas físicas e biométricas, potenciadas pela generalização de tecnologias de rastreamento. O recurso a GPS e sensores de movimento permitiu medir quilómetros percorridos, acelerações, desacelerações ou cargas de treino. Estes dados, ao relacionarem-se diretamente com intensidade, fadiga e risco de lesão, passaram também a ser considerados proxies da disponibilidade competitiva e da valorização de mercado (Carling et al., 2014; FIFA, 2023).

Mais recentemente, surgiram variáveis de natureza mediática e social, refletindo a crescente interligação entre desporto e economia da atenção. Estudos têm mostrado que indicadores como o número de seguidores em redes sociais, o volume de pesquisas em motores de busca ou a exposição mediática de um jogador podem acrescentar valor explicativo na análise do mercado de transferências (Majewski, 2022; Franceschi et al., 2023).

Esta evolução metodológica traduziu-se na passagem de indicadores absolutos para métricas normalizadas e contextuais, que procuram captar diferentes dimensões do jogo. Atualmente, os indicadores utilizados abrangem aspetos ofensivos, defensivos, físicos e mediáticos, oferecendo uma leitura mais completa do contributo de cada jogador. Com isso, tornaram-se uma ferramenta central tanto na investigação académica como na prática de avaliação realizada pelos clubes.

2.7 O papel das organizações internacionais

As organizações internacionais desempenham um papel central na monitorização, regulação e análise do mercado de transferências, oferecendo dados e relatórios que se tornaram referências para investigadores, jornalistas e profissionais da indústria do futebol. Entre estas, destacam-se a FIFA, a UEFA e o CIES – Football Observatory, que contribuem de forma sistemática para a compreensão das dinâmicas globais do setor.

A FIFA publica desde 2019 o Global Transfer Market Report, documento baseado nos registos do Transfer Matching System (TMS). Esta plataforma centraliza todas as transferências internacionais, permitindo uma monitorização precisa do número de transações, dos montantes envolvidos, da distribuição por ligas e confederações e ainda da caracterização dos jogadores transferidos em função da idade, posição ou nacionalidade (FIFA, 2023). Estes relatórios revelaram, por exemplo, que o mercado global ultrapassou 20 mil transferências internacionais em 2022, movimentando valores superiores a 6,5 mil milhões de dólares, com forte concentração nas principais ligas europeias.

A UEFA, por sua vez, disponibiliza o Club Licensing Benchmarking Report, onde analisa anualmente a situação financeira dos clubes europeus. O relatório abrange dimensões como receitas provenientes de direitos televisivos, bilheteira e patrocínios, o peso dos salários no orçamento e a sustentabilidade dos investimentos em transferências (UEFA, 2024). A informação financeira disponibilizada permite contextualizar o comportamento do mercado de transferências, demonstrando, por exemplo, como a Premier League mantém níveis de investimento muito superiores aos das restantes ligas, gerando efeitos inflacionários que se refletem em toda a Europa.

O CIES – Football Observatory complementa estas fontes com estudos aplicados que cruzam dados de transferências, performance desportiva e variáveis contratuais. Os relatórios periódicos do CIES incluem rankings de jogadores com maior potencial de valorização, estimativas de valores de mercado baseadas em modelos econométricos e comparações internacionais de competitividade (CIES, 2024). A sua metodologia, que integra fatores como idade, minutos jogados, duração do contrato e prestígio do clube, tornou-se amplamente citada na imprensa e nos debates sobre transferências.

A complementaridade entre FIFA, UEFA e CIES garante uma leitura mais completa do mercado de transferências: a primeira assegura os registos oficiais, a segunda contextualiza a situação financeira dos clubes e o terceiro acrescenta perspetivas analíticas sobre a valorização de jogadores. Esta triangulação de fontes tem reforçado a fiabilidade dos estudos e contribuído para uma maior clareza num mercado onde a informação é muitas vezes fragmentada.

2.8 Limitações e desafios na literatura

Apesar do desenvolvimento significativo da investigação sobre o valor de mercado de jogadores e da crescente utilização de técnicas de aprendizagem automática, a literatura continua a apontar diversas limitações e desafios que condicionam a consolidação desta área de estudo.

Um dos principais entraves prende-se com a qualidade e disponibilidade dos dados. As fontes mais utilizadas variam entre estatísticas oficiais das ligas, plataformas comerciais como a Transfermarkt e proxies alternativas, como os ratings dos videojogos FIFA. No entanto, estas bases diferem na cobertura temporal, no grau de fiabilidade e nos critérios de recolha, o que pode introduzir enviesamentos e dificultar a replicabilidade dos estudos (Müller et al., 2017; Franceschi et al., 2023). Acresce ainda o facto de muitas informações contratuais, como salários ou cláusulas de rescisão, permanecerem confidenciais, o que limita a construção de modelos totalmente transparentes.

Outro desafio importante é a heterogeneidade entre ligas e mercados nacionais. A Premier League inglesa, por exemplo, destaca-se pelo nível de receitas e capacidade de investimento, criando valores de transferências sistematicamente superiores aos registados noutras ligas (Dobson & Goddard, 2011; UEFA, 2024). Esta disparidade estrutural torna difícil generalizar modelos desenvolvidos para um contexto específico, uma vez que variáveis que explicam bem as dinâmicas em Inglaterra podem não ter o mesmo impacto em ligas como a portuguesa ou a holandesa.

A natureza multidimensional do valor de mercado constitui igualmente um obstáculo. O preço de um jogador resulta não apenas de métricas de desempenho, mas também de fatores externos como a popularidade mediática, o potencial de marketing, a situação contratual ou a estratégia financeira dos clubes (Majewski, 2022; CIES, 2024). Muitos destes elementos são de difícil quantificação e apenas parcialmente refletidos nos datasets disponíveis, o que reduz a capacidade explicativa dos modelos e aumenta a margem de subjetividade.

No campo metodológico, destaca-se ainda o problema da interpretabilidade. Algoritmos como XGBoost ou redes neuronais profundas oferecem níveis elevados de precisão, mas a sua opacidade pode limitar a adoção prática pelos clubes, que necessitam de compreender os fatores que sustentam as previsões. Ferramentas como SHAP ou LIME têm ajudado a mitigar este desafio ao atribuir contributos individuais às variáveis, mas o equilíbrio entre precisão e transparência permanece um debate central na literatura (Rossi et al., 2018; Shen et al., 2024).

Importa também reconhecer a influência de choques externos e fatores macroeconómicos sobre o mercado de transferências. A pandemia de COVID-19 reduziu drasticamente os valores médios transacionados a partir de 2020, revelando a vulnerabilidade do setor a crises globais (FIFA, 2021).

III. Hipóteses de Investigação e Modelo Conceptual

A presente investigação parte do pressuposto de que o valor de mercado dos jogadores de futebol é influenciado por um conjunto heterogéneo de fatores que combinam dimensões desportivas, biográficas e contextuais. A revisão da literatura demonstrou que variáveis como idade, desempenho em campo e características do clube desempenham um papel determinante na valorização dos atletas (Frick, 2007; Herm et al., 2014; Franceschi et al., 2023). Com base nestes contributos, foram delineadas hipóteses de investigação que servirão de base à modelação preditiva a desenvolver no capítulo metodológico.

3.1 Hipóteses de Investigação

- **H1**: A idade apresenta um efeito não linear sobre o valor de mercado dos jogadores, registando-se um pico de valorização em idades intermédias (aproximadamente entre os 24 e 28 anos) e uma depreciação progressiva em fases posteriores da carreira (Frick, 2007; Herm et al., 2014).
- **H2**: As estatísticas de desempenho ofensivo, nomeadamente golos e assistências por 90 minutos, influenciam positivamente o valor de mercado dos jogadores, sendo este efeito particularmente acentuado nos avançados (Gudmundsson & Horton, 2017; Rossi et al., 2018).
- **H3**: O número total de minutos jogados numa época tem impacto positivo no valor de mercado, refletindo a importância da consistência e da disponibilidade competitiva (Franceschi et al., 2023).
- **H4**: O valor agregado do clube em que o jogador atua exerce influência significativa na sua valorização individual, funcionando como proxy de prestígio competitivo e capacidade financeira (Müller et al., 2017; UEFA, 2024).
- **H5**: O histórico de valorização do jogador mantém-se como um determinante relevante do seu valor atual, funcionando como uma "âncora" de mercado (J. Liu., 2025).

3.2 Modelo Conceptual de Investigação

Com base na revisão da literatura e nos objetivos definidos para este estudo, foi elaborado o modelo conceptual que sustenta a investigação (Figura 3.1). O valor de mercado dos jogadores é representado como a variável dependente, influenciada por quatro grupos de determinantes principais.

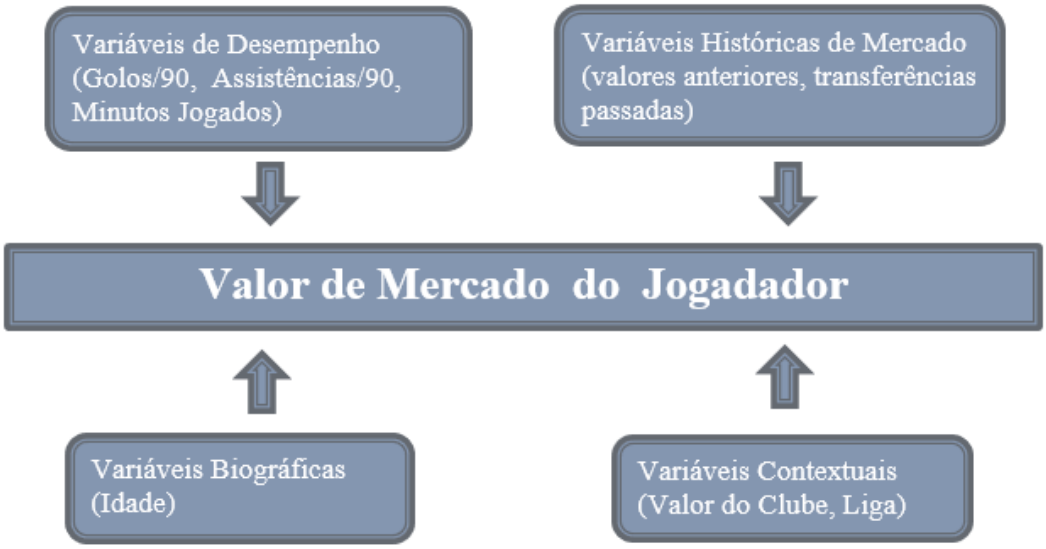


Figura 3.1. Modelo Conceptual de Investigação.

IV. Metodologia

Este capítulo descreve a metodologia adotada ao longo do trabalho, isto é, o caminho seguido desde a formulação do problema até à construção e avaliação dos modelos. O objetivo é apresentar de forma clara e organizada as diferentes etapas realizadas, assegurando a ligação entre as questões de investigação e os métodos aplicados.

Para estruturar o processo, foi utilizada a metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining), reconhecida como uma das mais consolidadas em projetos de ciência de dados. Embora definida em fases, esta abordagem é flexível e iterativa, permitindo regressar a etapas anteriores sempre que necessário para aperfeiçoar os dados, as variáveis ou os modelos.

A primeira etapa corresponde ao Business Understanding, centrada na compreensão do fenómeno em estudo. Segue-se o Data Understanding, que visa a exploração e caracterização dos dados relevantes. Posteriormente, o Data Preparation reúne os procedimentos de preparação e transformação da informação, servindo de base para a fase de Modeling, onde são aplicados os diferentes algoritmos de previsão. A etapa de Evaluation foca-se na análise crítica dos resultados obtidos, garantindo a sua validade e utilidade no contexto definido. Por fim, a fase de Deployment, relativa à implementação prática da solução, não foi considerada no âmbito do presente projeto.

Como representado na Figura 4.1, a metodologia CRISP-DM estrutura o caminho seguido neste trabalho, garantindo a coerência entre a compreensão do problema, a preparação dos dados, a modelação e a avaliação.

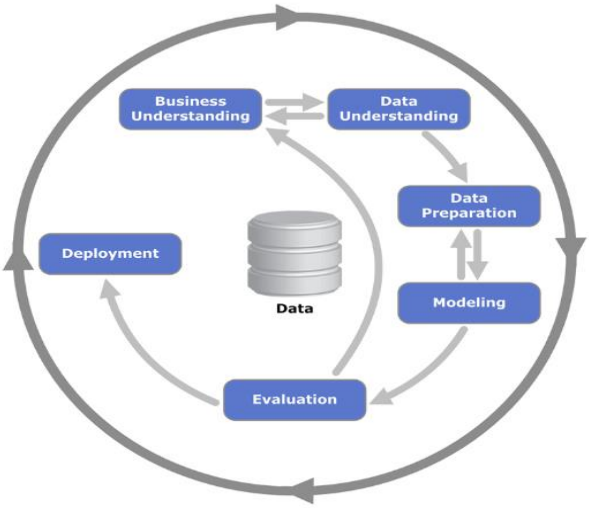


Figura 4.1. Metodologia CRISP-DM.

4.1 Business Understanding

A valorização de jogadores de futebol tornou-se um tema cada vez mais relevante para clubes, agentes e investidores, uma vez que as transferências representam hoje uma das principais fontes de receita e um eixo estratégico na gestão desportiva moderna. O peso económico deste mercado é evidente: segundo a FIFA (*Global Transfer Report 2024*).

Em 2024, registaram-se 78.742 transferências internacionais, um número recorde no futebol profissional, masculino e feminino. No total, os clubes gastaram cerca de 7,9 mil milhões de euros em taxas de transferência no futebol masculino profissional, valor que representa o segundo mais elevado de sempre, apenas superado pelo recorde histórico de 9,66 mil milhões de euros em 2023 (FIFA, 2024). Apesar desta ligeira descida, o montante de 2024 mantém-se mais de 15% acima do anterior máximo registado em 2019, evidenciando a dimensão e o dinamismo crescentes do mercado global.

A Liga inglesa continua a ser o grande motor deste mercado, tendo os seus clubes liderado tanto nas contratações (cerca de €1,73 mil milhões) como nas vendas (cerca de €1,23 mil milhões). Para além do futebol masculino profissional, o relatório mostra sinais de transformação noutras dimensões: o futebol feminino duplicou os gastos face a 2023, ultrapassando os €15 milhões, e no futebol amador contabilizaram-se mais de 53 mil transferências internacionais, com a Alemanha a assumir a dianteira. Estes números demonstram não apenas a escala global do mercado, mas também a sua crescente diversificação.

Neste trabalho, o foco está em perceber os fatores que explicam o valor de mercado dos jogadores nas cinco principais ligas europeias — Premier League, La Liga, Serie A, Bundesliga e Ligue 1 — durante a época 2023/2024. A análise incide especificamente sobre avançados, pela centralidade que ocupam na lógica do jogo e pela forte valorização que tradicionalmente lhes é atribuída no mercado. Entre as variáveis consideradas incluem-se o desempenho desportivo (golos, assistências e minutos jogados), as características biográficas (como a idade) e o contexto competitivo (o valor agregado do clube).

O objetivo é, assim, compreender em que medida estas diferentes dimensões — desportivas, biográficas e contextuais — ajudam a explicar as disparidades observadas nos valores de mercado. Mais do que fornecer números ou previsões isoladas, este estudo pretende mostrar como a análise de dados pode apoiar decisões de contratação e investimento num setor marcado pela competitividade, pela incerteza e por assimetrias financeiras cada vez mais evidentes.

4.2 Data Understanding

A informação utilizada neste trabalho tem origem em datasets públicos disponibilizados no Kaggle, construídos a partir de extrações automáticas da plataforma Transfermarkt. Estes ficheiros, em formato CSV, possuem uma estrutura relacional, onde identificadores consistentes — *player_id*, *club_id*, *game_id* e *competition_id* — permitem cruzar, de forma integrada, diferentes dimensões do futebol profissional: características biográficas dos jogadores, contexto competitivo de clubes e ligas, desempenho em campo, movimentos de mercado e evolução do valor atribuído a cada atleta. De forma a reforçar a fiabilidade dos dados, as principais variáveis e estatísticas extraídas foram posteriormente confirmadas em fontes oficiais disponíveis online, garantindo consistência e alinhamento com a realidade observada nas competições.

No seu conjunto, os datasets cobrem uma escala muito significativa. O ficheiro players.csv contém mais de 32 mil jogadores, com informação relativa a idade, altura, posição, pé preferencial e nacionalidade. O ficheiro clubs.csv descreve 439 clubes, incluindo atributos como o valor total de mercado do plantel e a competição doméstica em que se inserem. O ficheiro competitions.csv, por sua vez, identifica 44 competições de diferentes países, cada uma com um código próprio, sendo particularmente relevantes os que representam as cinco principais ligas europeias — Premier League (GB1), La Liga (ES1), Serie A (IT1), Bundesliga (L1) e Ligue 1 (FR1).

A componente dinâmica dos dados é garantida por vários ficheiros. O games.csv e o appearances.csv cobrem partidas disputadas entre 2014 e 2024, registando, respetivamente, o calendário de jogos e a participação de cada jogador em termos de minutos jogados, golos, assistências, cartões e substituições. O game_events.csv permite um detalhe adicional sobre cada ação em jogo (golo, assistência, penálti, etc.), enquanto o game_lineups.csv oferece mais de 2 milhões de registos com informação sobre escalações e posicionamento em campo. O player_valuations.csv, com quase 500 mil observações, regista o valor de mercado estimado para cada atleta ao longo do tempo, entre 2013 e 2024. Finalmente, o transfers.csv agrega cerca de 80 mil transferências, com informação sobre clube de origem e destino, data e valores associados.

No âmbito desta dissertação, foi necessária uma delimitação clara do universo de análise. O foco recai sobre a época 2023/2024, restringindo a amostra às cinco principais ligas europeias (Big-5). Dentro deste contexto competitivo, o interesse centra-se nos jogadores classificados como avançados (*Attack*), por serem os principais responsáveis pela produção ofensiva e pela

geração de valor mediática e financeira no futebol. Após a filtragem, obtiveram-se 814 jogadores (com registos de minutos efetivos e valores de mercado atualizados em 2023/24), que constituem a base para toda a exploração subsequente.

Este recorte tem várias vantagens analíticas. Em primeiro lugar, garante uma amostra suficientemente extensa para permitir comparações entre ligas e perfis de jogadores, mantendo ao mesmo tempo foco num segmento específico (avançados) de relevância direta para a valorização no mercado de transferências. Em segundo lugar, a estrutura longitudinal dos dados, particularmente através do player_valuations.csv, possibilita o estudo da evolução histórica do valor de mercado, complementada por indicadores de performance desportiva (golos, assistências, minutos jogados). O enquadramento relacional entre jogadores, clubes e competições assegura a replicabilidade do estudo, uma vez que cada passo de integração pode ser reproduzido por terceiros, partindo dos mesmos ficheiros brutos.

Importa destacar que os valores de mercado registados no Transfermarkt não correspondem a valores contratuais oficiais, mas antes a estimativas baseadas em critérios públicos e consensuais dentro da comunidade. Este caráter subjetivo, embora reconhecido na literatura, não diminui a utilidade do dataset: pelo contrário, constitui uma métrica amplamente utilizada em trabalhos académicos e aplicada em estudos de scouting e prospeção desportiva. Assim, a amostra final — avançados das Big-5 em 2023/24 — representa um ponto de equilíbrio entre amplitude e especificidade, permitindo explorar tanto a diversidade dos jogadores disponíveis como os fatores determinantes do seu valor de mercado.

4.2.1 Distribuição etária e utilização

O padrão etário segue o esperado nas grandes ligas europeias: os guarda-redes apresentam uma mediana mais elevada (30 anos), enquanto defesas, médios e avançados se concentram em torno dos 25–26 anos. Isto confirma que o auge competitivo dos jogadores de campo se situa nos 20s médios, ao passo que a baliza permite carreiras mais longas (este gráfico usa todas as posições apenas como enquadramento, todo o restante foco analítico desta dissertação incide nos avançados das ligas Big-5 em 2023/24).

Entre os avançados observa-se uma curva típica de desenvolvimento: minutos/jogo ainda reduzidos até aos 20 anos, crescimento claro entre os 21–24, estabilização em torno dos 50 minutos/jogo até aos 27, e um ligeiro declínio após os 30 anos. A Figura 4.2 traduz este processo de integração, consolidação e posterior gestão de utilização ao longo da carreira ofensiva.

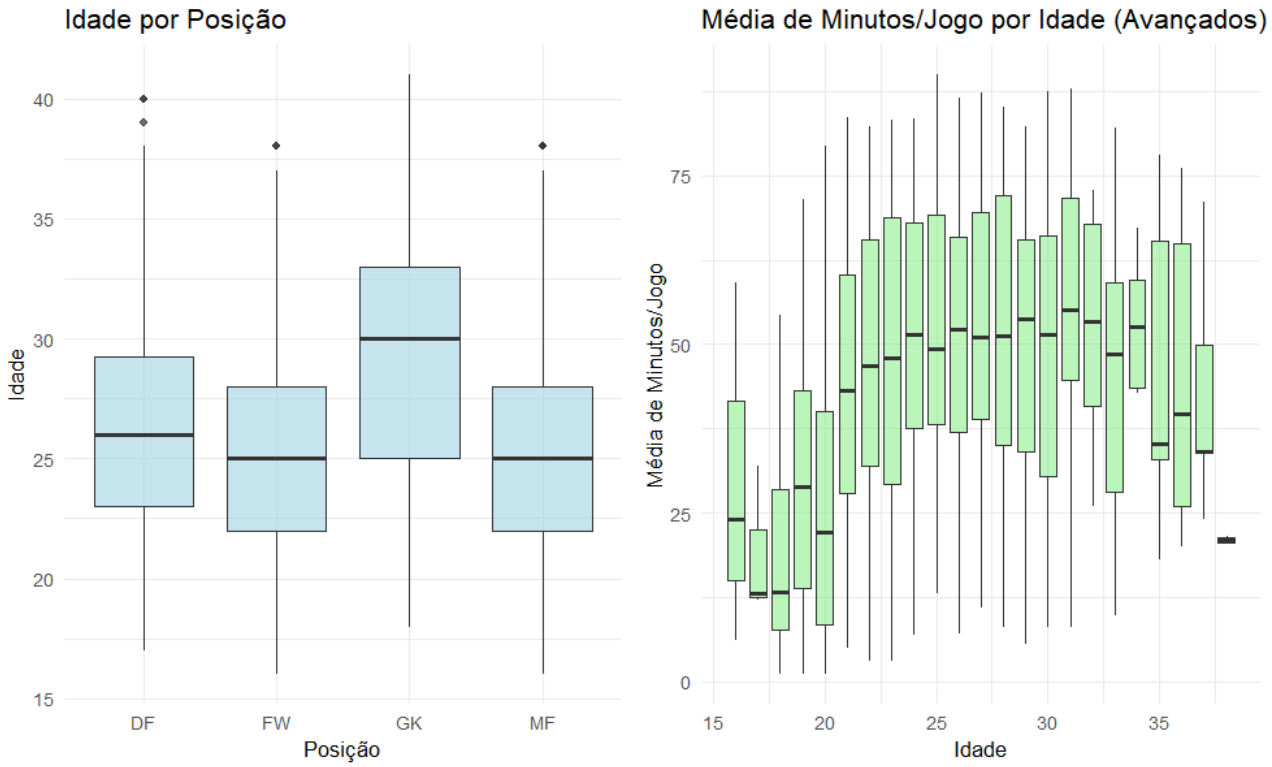


Figura 4.2. Relação entre idade, posição e utilização dos jogadores.

4.2.2 Nacionalidades e composição da amostra

Entre os avançados das cinco principais ligas europeias em 2023/24, destacam-se sobretudo jogadores oriundos de países da própria Europa. Espanha lidera o contingente (cerca de 13% dos avançados), seguida por França (9%) e Alemanha (8%). Estes três países concentram, em conjunto, mais de um quarto da amostra.

Fora do eixo das ligas "Big-5", merecem destaque Brasil e Argentina, que são os países extra-comunitários mais representados, confirmando a relevância histórica da América do Sul como fonte de talento ofensivo para o futebol europeu.

Estes padrões de distribuição podem ser observados na Figura 4.3, que apresenta as dez nacionalidades mais representadas na amostra.

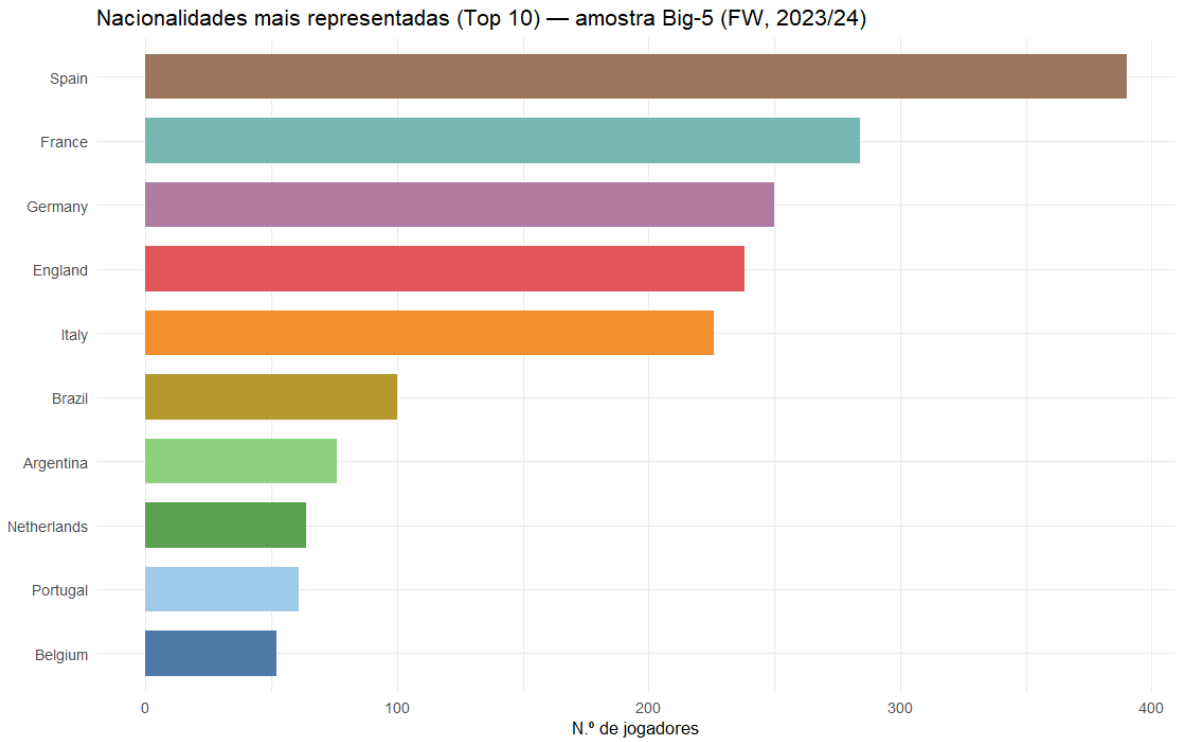


Figura 4.3. Nacionalidades mais representadas entre os avançados das cinco principais ligas europeias (época 2023/24).

4.2.3 Mercado: nível e evolução

Olhando para a hierarquia atual do mercado entre os avançados das cinco principais ligas, destacam-se três jogadores num patamar acima dos restantes: Kylian Mbappé (La Liga), Vinícius Júnior (La Liga) e Erling Haaland (Premier League), todos avaliados em cerca de €180M. Logo a seguir surgem jovens referências como Phil Foden (€150M) e Bukayo Saka (€140M), que consolidaram a sua valorização através de desempenhos consistentes nas últimas épocas. Entre os dez mais caros encontram-se ainda nomes estabelecidos como Harry Kane (€100M), já em fase mais madura da carreira, e talentos emergentes como Julián Álvarez (€90M). Esta lista confirma a predominância de atletas da Premier League e da La Liga na elite do mercado ofensivo europeu.

Esta concentração de valores de topo encontra-se ilustrada na Figura 4.4, onde se apresentam os dez avançados mais valiosos da época 2023/24.

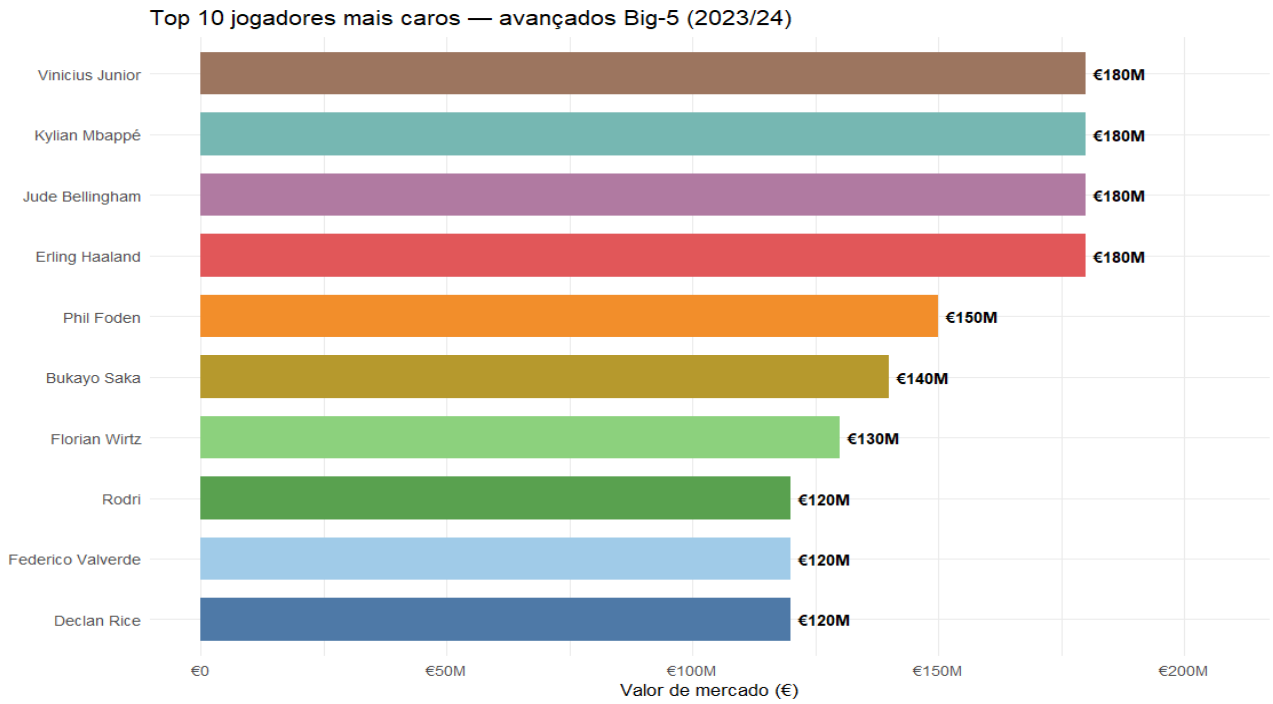


Figura 4.4. Top 10 avançados mais valiosos das cinco principais ligas europeias (época 2023/2024).

A análise da evolução nos últimos cinco anos destes cinco jogadores evidencia trajetórias contrastantes. Erling Haaland (+€188M; +1567%), Bukayo Saka (+€143M; +2043%) e Phil Foden (+€100M; +333%) apresentam crescimentos muito expressivos, alicerçados na consolidação como titulares indiscutíveis e no impacto direto nas respetivas equipas. Vinícius Júnior também se valorizou de forma acentuada (+€150M; +300%), confirmando a transição de promessa para referência mundial. Em contraste, Kylian Mbappé surge como exceção neste grupo, registando uma depreciação moderada (-€30M; -15%) relativamente a 2019, sinal de maturidade competitiva e de aproximação ao pico da sua trajetória de mercado.

Mais do que uma análise estática, a evolução histórica dos valores de mercado ajuda a compreender dinâmicas de valorização. A Figura 4.5, mostra a trajetória dos cinco avançados mais valiosos, evidenciando que a valorização segue ritmos distintos consoante a idade, o desempenho e o contexto competitivo.

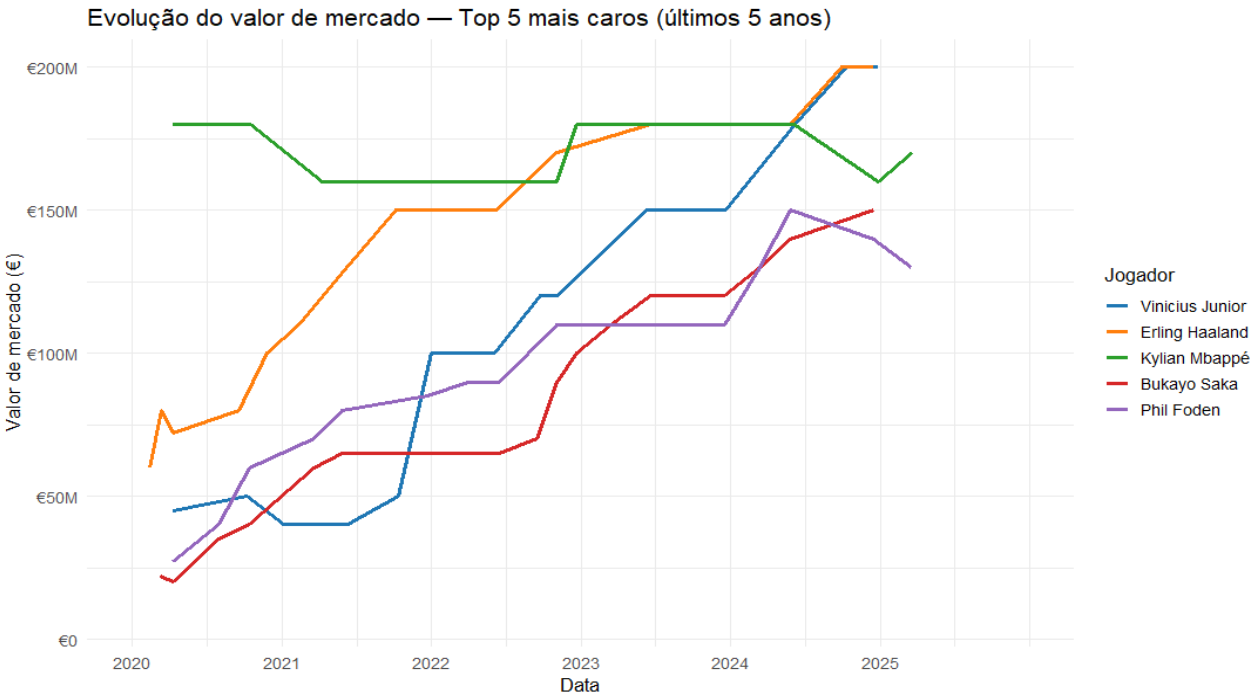


Figura 4.5. Evolução do valor de mercado dos cinco avançados mais valiosos das ligas Big-5 (2020–2025).

4.2.4 Desempenho Individual Ofensivo

Para além do enquadramento do mercado, é fundamental analisar a produção ofensiva direta dos avançados nas cinco principais ligas europeias em 2023/24. O ranking dos melhores marcadores confirma a hegemonia de figuras estabelecidas, com Harry Kane (44 golos), Erling Haaland (38) e Kylian Mbappé (36) no pódio, seguidos de Serhou Guirassy (30) e Pierre-Emerick Aubameyang (29).

No domínio das assistências, destacam-se sobretudo jovens talentos da La Liga: Nico Williams (19) e Álex Baena (18) lideram o ranking, enquanto Mohamed Salah, Marcus Thuram e Niklas Beste completam o Top 5, todos com 14 passes decisivos.

A comparação entre os dois rankings evidencia perfis distintos: de um lado, avançados eminentemente finalizadores, como Kane ou Guirassy; de outro, jogadores cuja relevância resulta da capacidade de criar oportunidades para os colegas, como Williams e Baena. Esta diversidade confirma que o contributo ofensivo dos avançados não se limita aos golos, mas deve integrar igualmente o papel criativo no ataque coletivo.

Como mostra a Figura 4.6, Harry Kane liderou destacado na lista de marcadores (44 golos), enquanto Nico Williams foi o jogador com mais assistências (19), evidenciando diferentes perfis de contribuição ofensiva.

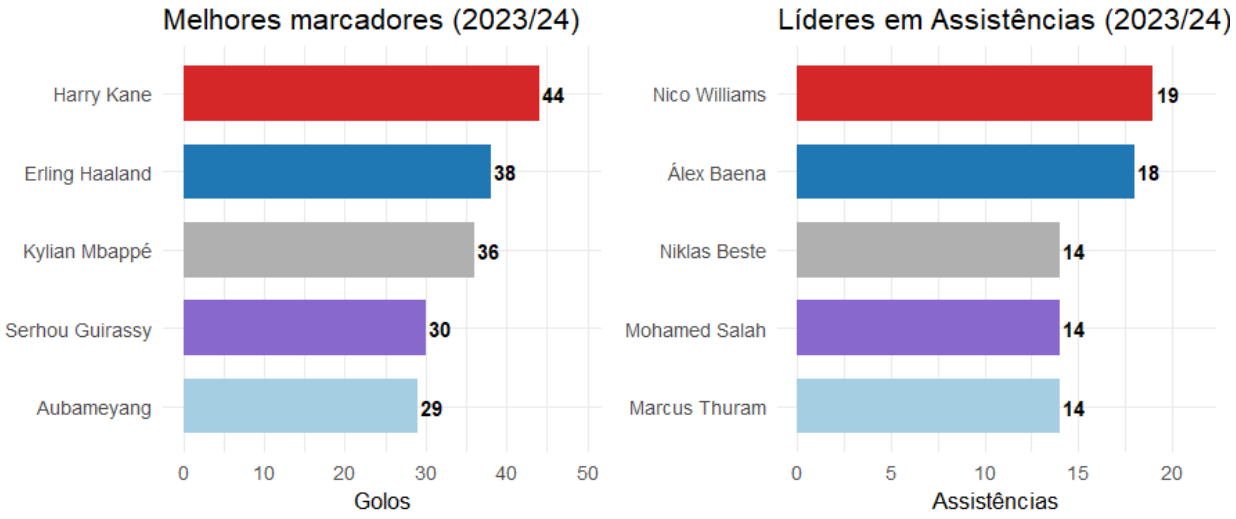


Figura 4.6. Top 5 de avançados com mais golos e assistências na época 2023/2024 nas cinco principais ligas europeias.

4.2.5 Relação entre desempenho e valor de mercado

A análise das relações entre utilização, desempenho ofensivo e valor de mercado evidencia padrões claros no conjunto dos avançados analisados. A Figura 4.7 ilustra estas associações, mostrando de forma visual como tanto os golos como os minutos jogados se relacionam com o valor de mercado.

O primeiro scatterplot mostra uma correlação positiva elevada entre golos marcados e valor de mercado ($r \approx 0.66$). De acordo com a regressão linear simples, cada golo adicional na época associa-se, em média, a um acréscimo de cerca de €2,3 milhões no valor do jogador.

O segundo gráfico confirma também o peso da utilização competitiva: o número total de minutos jogados apresenta correlação positiva com o valor ($r \approx 0{,}58$). Esta relação, embora relevante, é mais moderada do que a observada para os golos e mostra maior dispersão dos dados, refletindo que a valorização associada ao tempo de jogo não é uniforme entre todos os jogadores. Ainda assim, a tendência sugere que a regularidade em campo funciona como sinal de confiança competitiva e reduz incerteza para potenciais investidores.

Para além disso, contudo, que os testes de Breusch–Pagan realizados sobre os modelos lineares indicaram a presença de heterocedasticidade ($p < 0{,}001$), ou seja, a variância dos resíduos não era constante e tendia a aumentar para valores de mercado mais elevados.

Para garantir a validade estatística, os coeficientes foram reavaliados com erros-padrão robustos (HC1), mantendo-se significativos os efeitos dos golos e dos minutos jogados. Assim, os resultados apresentados refletem padrões robustos, ainda que a regressão linear simples, por si só, não capture toda a complexidade do fenómeno.

Estes resultados sublinham que tanto a eficácia ofensiva como a consistência de utilização são determinantes fundamentais do valor de mercado dos avançados, reforçando a ideia de que o mercado valoriza não apenas o talento de finalização, mas também a capacidade de estar disponível e ativo ao longo da época.

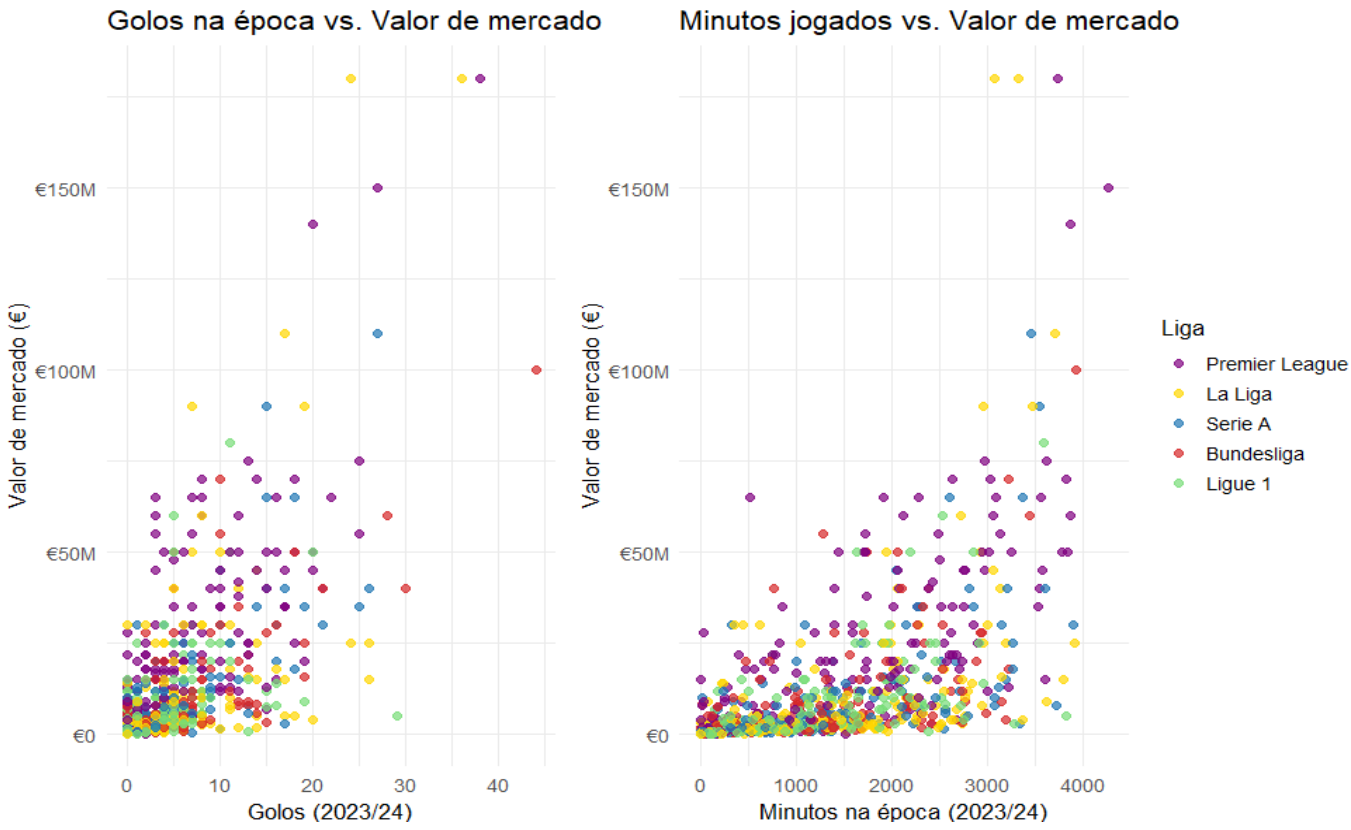


Figura 4.7. Associação entre métricas de desempenho (golos e minutos) e valor de mercado na época 2023/2024.

4.2.6 Correlações entre variáveis

A Figura 4.8 sintetiza os principais padrões de correlação entre características e desempenho dos avançados. O valor de mercado apresenta a associação mais forte com os minutos

jogados (r = 0.58), confirmando que a utilização consistente em contexto competitivo constitui um dos fatores mais valorizados pelo mercado.

A relação com os golos por 90 minutos é positiva, mas mais moderada (r = 0.22), sugerindo que a capacidade de finalização, embora importante, não explica isoladamente a valorização de mercado — sendo necessário considerar um conjunto mais amplo de atributos, como idade, regularidade e potencial.

Por outro lado, variáveis como assistências/90 (r = –0.01) não mostram qualquer associação estatisticamente relevante, indicando que este contributo ofensivo não constitui, por si só, um fator determinante para a valorização. Já a idade apresenta uma correlação praticamente nula (r = –0.07) nesta análise restrita a uma época. No entanto, como sugerem estudos longitudinais, jogadores mais jovens tendem a valorizar de forma consistente ao longo do tempo, ainda que essa relação se dilua quando se observa apenas um único período competitivo.

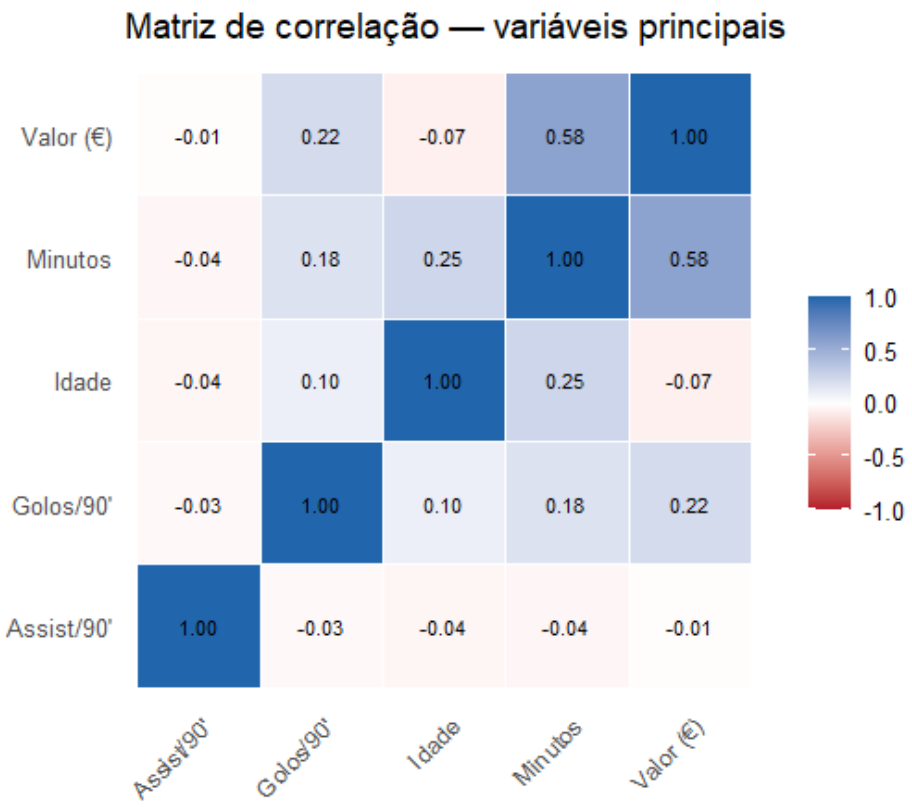


Figura 4.8. Matriz de correlação.

4.3 Data Preparation

A preparação dos dados assume um papel central, pois transforma a informação recolhida em bruto num conjunto estruturado, limpo e coerente, pronto para análises rigorosas. Nesta fase definem-se os critérios de inclusão de observações, resolvem-se inconsistências nos registos e constroem-se variáveis derivadas relevantes para o estudo. Ao garantir a comparabilidade entre jogadores, ligas e épocas, este processo assegura que a análise estatística e os modelos preditivos assentam numa base sólida e fidedigna.

No presente estudo, o ponto de partida foi o conjunto de ficheiros transferidos do Kaggle (com origem no Transfermarkt), abrangendo múltiplas dimensões — biográfica (players.csv), registos de utilização em jogo (appearances.csv), histórico de valores de mercado (player_valuations.csv), clubes (clubs.csv) e competições (competitions.csv). A informação foi posteriormente validada através de consulta a fontes oficiais e reconhecidas no domínio do futebol profissional, de forma a reforçar a sua fiabilidade. A estrutura dos dados assenta em identificadores comuns (player_id, club_id, competition_id), o que permite o cruzamento consistente entre tabelas e a construção de uma base integrada para análise.

A preparação decorreu em três passos principais:

(i) Seleção da amostra

A análise incide sobre a época 2023/24, restringindo as observações às cinco principais ligas europeias (Premier League, La Liga, Serie A, Bundesliga e Ligue 1). Foram selecionados apenas os jogadores classificados como avançados (posição "FW"), alinhando a amostra com o foco central da dissertação.

(ii) Normalização biográfica e posicional

As posições originais foram simplificadas para quatro categorias normalizadas (GK, DF, MF e FW), ainda que nesta investigação apenas a última tenha sido mantida. As datas de nascimento foram convertidas em idades ao final da época de referência (30/06/2024), assegurando uniformidade no cálculo e evitando discrepâncias associadas a diferentes momentos de recolha.

(iii) Construção de métricas de utilização e desempenho ofensivo

Com base nos registos de jogo, foram construídos indicadores de utilização e desempenho ofensivo, incluindo o número de jogos disputados, os minutos totais, a média de minutos por partida

e métricas normalizadas por 90 minutos (golos/90 e assistências/90). Em complemento, o ficheiro *player_valuations.csv* forneceu o valor de mercado mais recente de cada atleta na época 2023/24, que foi integrado no dataset final, garantindo uma ligação direta entre desempenho em campo e valorização de mercado.

O resultado deste processo é um painel consolidado ao nível jogador–época, com variáveis biográficas (idade, liga), contextuais (competição, clube), de utilização (jogos e minutos) e de desempenho ofensivo (golos e assistências), complementadas pelo valor de mercado. Este dataset constitui a base para os testes de modelação subsequentes.

4.3.1 Tratamento de valores omissos

Antes de avançar para a modelação, foi necessário lidar com valores em falta que apareciam em diferentes variáveis. A abordagem seguida foi pragmática: em vez de imputar valores arbitrários, optou-se por soluções simples que mantêm a coerência dos dados e evitam enviesamentos.

A altura, variável pouco relevante para os objetivos do estudo, apresentou alguns registos em falta; nesses casos foi atribuído o valor zero, funcionando apenas como marcador de ausência.

A variável pé preferencial (direito, esquerdo, ambos) não estava completa para todos os atletas. Em vez de substituir de forma artificial, foi criada a categoria explícita *"SemPé"*. De igual forma, na identificação da liga dominante, jogadores sem registos claros foram classificados em *"SemLiga"*. Quanto à posição, foram criadas quatro categorias normalizadas (GK, DF, MF, FW); nos raros casos em que não havia posição definida, registou-se *"SemPosição"*, mas, dado que o estudo se foca apenas em avançados (FW), estes acabaram por não entrar na amostra final.

Nas variáveis contextuais também houve lacunas. A taxa de transferência mais recente (*transfer fee*), ausente em muitos casos, foi preenchida com zero, de forma a preservar a consistência numérica. O mesmo se verificou no valor total do clube, que foi calculado a partir de duas fontes: os valores agregados em *clubs.csv* e a soma das valuations dos jogadores no plantel. Quando ambas as opções falhavam, o valor foi assumido como zero. Já o pico pré-época (valor

máximo de mercado registado até ao início da temporada) foi igualmente fixado em zero para jogadores sem histórico anterior.

Em resumo, os valores omissos foram sempre tratados de modo a não quebrar a estrutura dos dados, sem recorrer a imputações artificiais que pudessem distorcer resultados. Sempre que possível, a ausência foi registada como categoria própria; quando isso não era viável, recorreu-se ao preenchimento com zero, garantindo que todos os jogadores da amostra pudessem ser incluídos nas análises. Importa ainda sublinhar que várias destas variáveis correspondem a novos indicadores derivados (como o pico pré-época ou o valor agregado do clube), cuja construção será detalhada no ponto seguinte.

4.3.2 Criação de novas variáveis

No seguimento da preparação dos dados, foi necessário criar novos indicadores que permitissem enriquecer a análise e aumentar a capacidade explicativa e a estabilidade dos modelos. A lógica foi sempre a mesma: transformar a informação bruta em métricas comparáveis e relevantes para compreender o valor de mercado dos jogadores.

Uma das primeiras transformações foi a normalização das estatísticas ofensivas. Criaram-se os indicadores de golos e assistências por 90 minutos, o que permite comparar diretamente jogadores que jogaram tempos muito diferentes. Sem esta normalização, um avançado suplente com apenas 500 minutos jogados ficaria automaticamente em desvantagem em relação a um titular com 3000 minutos, mesmo que tivesse uma eficácia semelhante.

Foram ainda acrescentadas variáveis de contexto, essenciais para captar fatores externos que influenciam o mercado. A mais relevante é o valor total do plantel do clube, calculado a partir da soma dos valores de mercado de todos os jogadores no início da época. Esta variável dá uma noção clara da força do clube e do ambiente competitivo em que cada jogador se insere. Atuar num clube de topo aumenta a visibilidade, o acesso a competições internacionais e, inevitavelmente, potencia a valorização individual.

Outra variável criada foi o pico pré-época de cada jogador, isto é, o valor máximo de mercado atingido até à véspera da temporada em análise. Este indicador é particularmente importante porque capta a trajetória histórica do atleta. Jogadores que já atingiram picos muito altos mantêm uma espécie de "âncora de mercado": mesmo que atravessem fases de menor rendimento, dificilmente perdem toda a valorização acumulada. Pelo contrário, atletas que nunca

tiveram uma valorização significativa encontram mais barreiras para subir de patamar, mesmo que tenham boas épocas.

A Figura 4.9 abaixo ilustra bem esta lógica, comparando dois casos distintos: Harry Kane, cujo pico pré-época já tinha sido muito elevado, e que mesmo com uma ligeira descida continua a apresentar um valor de mercado alto; e Bukayo Saka, que na época em análise ultrapassou claramente o seu pico anterior, sinalizando uma forte valorização em ascensão. Assim, quando o valor atual supera o pico pré-época, podemos interpretar que o jogador está a viver um período de grande afirmação e crescimento.

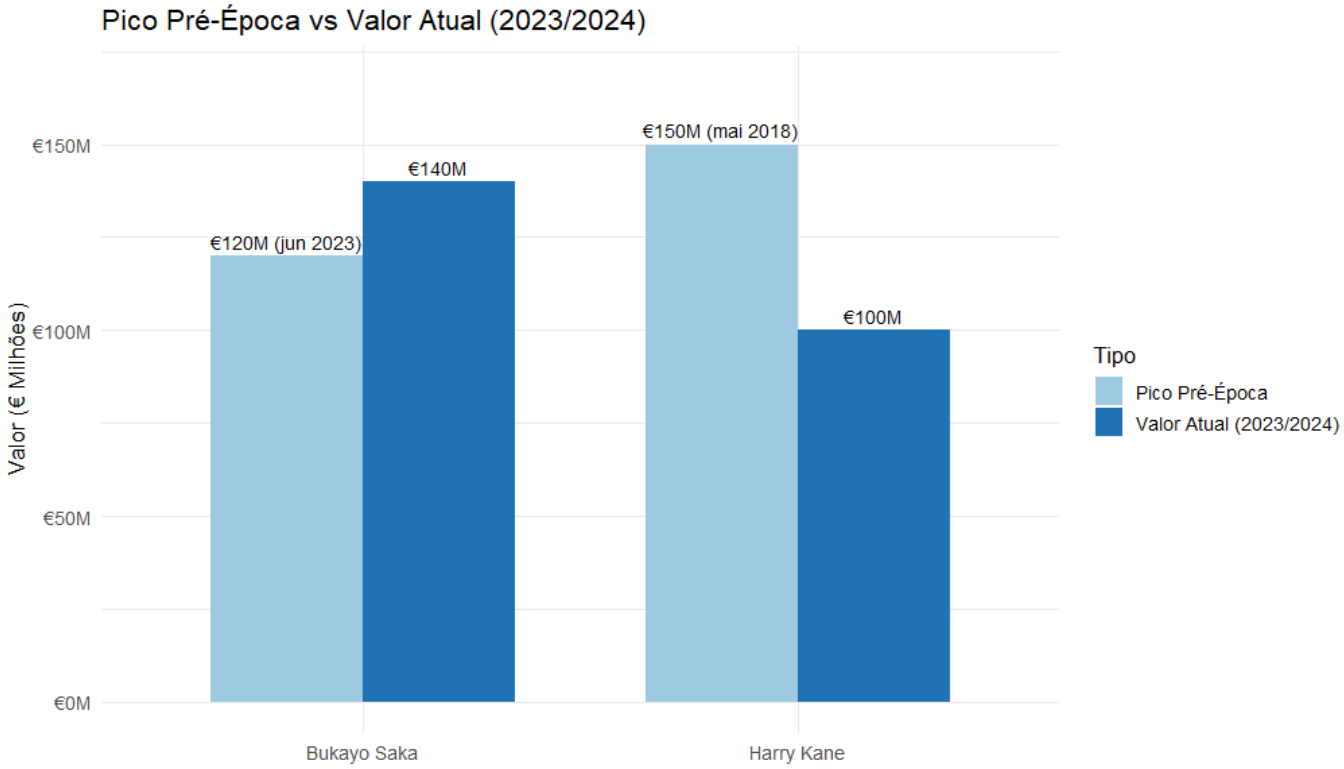


Figura 4.9. Comparação entre valor atual e pico pré-época para Harry Kane e Bukayo Saka.

Por fim, foi introduzida também a variável idade ao quadrado (idade^2) para capturar relações não lineares entre idade e valor de mercado. Na prática, ajuda a representar o padrão típico de valorização dos jogadores: crescimento acelerado até um certo ponto, estabilização no pico da carreira e eventual decréscimo na fase final.

Estas novas variáveis complementam as informações biográficas e estatísticas originais, permitindo uma análise mais rica e realista da valorização dos jogadores. Servem também de

ponte para os modelos preditivos, ao integrarem tanto a componente de desempenho individual como o enquadramento coletivo e histórico.

Importa ainda sublinhar que os valores de mercado utilizados provêm do Transfermarkt, uma das bases mais reconhecidas neste domínio. No entanto, tal como acontece com outras plataformas especializadas (ex.: CIES Football Observatory, Football Benchmark), estes números devem ser entendidos como estimativas aproximadas, e não como valores contratuais oficiais. Diferentes metodologias podem levar a resultados distintos, mas, pela sua abrangência e consistência histórica, o Transfermarkt constitui uma referência adequada para análises comparativas.

4.3.3. Transformações finais para modelação

Antes de avançar para a fase de modelação, foi necessário aplicar um conjunto de transformações adicionais ao dataset, com o objetivo de melhorar a qualidade estatística das variáveis e adequá-las aos algoritmos de Machine Learning utilizados. Estas transformações visaram, sobretudo, estabilizar distribuições assimétricas, captar relações não lineares e reduzir o impacto de valores extremos.

Uma das primeiras transformações aplicadas aos dados foi sobre a variável-alvo — o valor de mercado dos jogadores. Na sua forma original, a distribuição apresenta uma forte assimetria à direita: a maioria dos avançados nas ligas Big-5 em 2023/24 tem valores que variam entre alguns milhões e poucas dezenas de milhões de euros, mas existe um pequeno grupo de jogadores avaliados em cifras extraordinariamente elevadas (casos como Mbappé ou Haaland, acima dos €150M). Esta diferença cria uma "cauda longa" na distribuição, visível no gráfico da esquerda, que pode distorcer a aprendizagem dos modelos, tornando-os excessivamente sensíveis a um número muito reduzido de observações.

A Figura 4.10, evidencia o impacto da transformação logarítmica ($\log(1 + €)$) sobre esta variável: a distribuição torna-se muito mais simétrica e próxima de uma forma aproximadamente normal, reduzindo o peso relativo dos valores extremos. Assim, diferenças que antes eram desproporcionais passam a ser interpretadas de forma mais equilibrada. Por exemplo, a distância entre €5M e €15M (três vezes mais) é, na prática, mais relevante do que a diferença entre €100M e €110M, mesmo que em termos absolutos sejam ambos €10M — algo que o logaritmo ajuda a refletir.

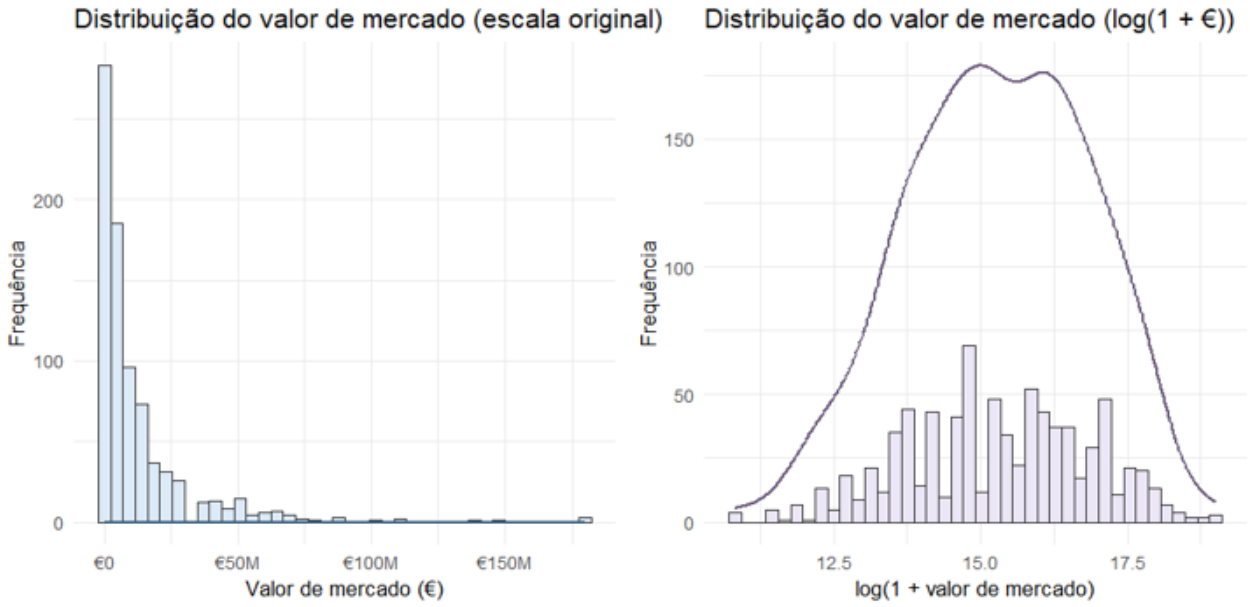


Figura 4.10. Efeito da transformação logarítmica nos valores de mercado.

Esta transformação normaliza a escala da variável-alvo, melhora a estabilidade dos modelos de Machine Learning e torna a interpretação dos resultados mais intuitiva, ao expressar as variações de mercado em termos proporcionais e não apenas absolutos.

Outro ajuste importante foi a introdução da variável idade ao quadrado (idade²). A relação entre idade e valor de mercado não é linear: os jogadores tendem a valorizar rapidamente na transição dos 18 para os 23 anos, atingem o pico entre os 24 e 28, e depois começam a desvalorizar de forma progressiva. Ao acrescentar idade², o modelo ganha flexibilidade para capturar esse formato em curva, algo que uma variável linear simples não conseguiria representar.

Relativamente ao desempenho ofensivo, aplicou-se a winsorização das métricas de golos/90 e assistências/90 no percentil 99. Isto permitiu reduzir distorções causadas por jogadores com poucos minutos, mas rácios desproporcionalmente altos. Sem este passo, casos pouco representativos poderiam dominar as estatísticas. Com a winsorização, mantém-se a informação, mas os valores mais extremos são ajustados para um nível máximo plausível, assegurando maior robustez analítica.

Variáveis contextuais como o valor total do plantel do clube e o pico pré-época também foram convertidas em escala logarítmica, pelo mesmo motivo: reduzir a assimetria e evitar que

clubes de orçamentos bilionários ou jogadores com valorização histórica excecional enviesem os resultados.

Por último, realizou-se uma verificação de variáveis com variância quase nula, utilizando a função *nearZeroVar* do pacote caret. Este procedimento foi aplicado apenas ao conjunto reduzido de variáveis preparado para o modelo — idade, idade², altura, minutos, golos/90', assistências/90', valor total do clube (log) e pico pré-época (log). O objetivo foi identificar potenciais preditores com falta de variabilidade, isto é, variáveis em que quase todos os jogadores apresentam o mesmo valor, tornando-as pouco informativas para o modelo. No entanto, todas as variáveis apresentaram níveis adequados de variabilidade, pelo que nenhuma foi removida nesta fase. É importante sublinhar que variáveis como a liga de origem ou o pé preferido já tinham sido excluídas previamente, por decisão metodológica, e, portanto, não foram objeto desta análise.

No conjunto, estas transformações finais permitiram chegar a um dataset otimizado, com variáveis comparáveis, escalas estabilizadas e relações não lineares representadas, criando as condições necessárias para que os modelos de previsão pudessem captar de forma mais precisa a complexidade do valor de mercado dos avançados nas cinco principais ligas europeias.

4.3.4. Construção dos conjuntos de treino e teste

Com o dataset final preparado — composto por 814 avançados das cinco principais ligas europeias na época 2023/24, já com variáveis normalizadas e enriquecidas — o passo seguinte consistiu em separar os dados em dois subconjuntos: treino (80%) e teste (20%).

Esta divisão é fundamental na construção de modelos de Machine Learning, pois permite:

- Treinar os algoritmos num subconjunto amplo de dados, onde são ajustados os parâmetros e identificados padrões;
- Avaliar a capacidade de generalização em dados nunca antes vistos (teste), garantindo que o modelo não está apenas a reproduzir o conjunto de treino (problema de *overfitting*).

Para manter a robustez estatística, foi usada uma partição estratificada, assegurando que a distribuição da variável-alvo (valor de mercado em log) é semelhante nos dois subconjuntos. Assim, tanto jogadores de elevado valor (e.g., Mbappé, Haaland, Saka) como atletas de menor valorização estão representados de forma proporcional, evitando que o teste fique enviesado apenas para um perfil.

Este procedimento resultou em 651 jogadores no treino e 163 no teste. O conjunto de treino foi ainda sujeito a validação cruzada (*cross-validation* 5-fold repetida), de forma a obter métricas estáveis durante a afinação dos modelos. O conjunto de teste foi mantido intocado até à fase final, funcionando como referência independente para avaliar a performance real de cada algoritmo.

Esta abordagem assegura que as métricas reportadas refletem não apenas a capacidade de ajuste dos modelos, mas sobretudo a sua capacidade de previsão em novos contextos, simulando a aplicação prática de estimar o valor de mercado de jogadores não incluídos no treino.

Para garantir a reprodutibilidade e comparações justas entre modelos, foi definido set.seed(300) e foram exportados dois ficheiros para reutilização nas experiências (Exp. 1 e Exp. 2): tabela_model.csv (com player_id, o alvo em log y_log = log1p(valor_mercado) e as variáveis numéricas finais: idade, idade², altura_cm, minutos, golos90, ast90, club_mv_log, peak_pre_log) e test_ids.csv (lista dos player_id do mesmo conjunto de teste). Desta forma, garante-se a independência entre as fases de treino e teste e assegura-se uma avaliação consistente ao longo das diferentes execuções.

4.4 Modelação

Para a previsão do valor de mercado foram testados vários algoritmos, organizados de forma progressiva desde modelos lineares mais simples até métodos de aprendizagem automática mais sofisticados. O objetivo não foi apenas comparar desempenhos, mas também compreender de que modo diferentes abordagens conseguem lidar com a complexidade inerente ao problema.

A regressão linear constituiu o ponto de partida. Trata-se de um modelo clássico, de interpretação direta, que permite identificar sinais e ordens de grandeza das variáveis. Apesar das suas limitações na captação de relações não lineares, serviu como referência inicial. Para aumentar a flexibilidade, foi ainda testada uma versão com splines e uma interação entre minutos e golos por 90 minutos, de modo a captar padrões curvilíneos característicos, como a valorização em idades intermédias.

Seguidamente aplicou-se o Elastic Net, uma regressão penalizada que combina as propriedades do Lasso e do Ridge. Este modelo é particularmente útil perante variáveis correlacionadas, uma vez que estabiliza os coeficientes e reduz o risco de sobreajuste, mantendo simultaneamente alguma capacidade de seleção de variáveis.

Para captar relações mais complexas, recorreu-se às árvores de decisão, que segmentam os dados através de regras simples (por exemplo: jogadores com elevado tempo de utilização e pico pré-época valorizado tendem a ter maior mercado). Embora intuitivas, as árvores isoladas são suscetíveis a instabilidade, pelo que se recorreu também a métodos de ensemble.

Nesse sentido, foi utilizada a Random Forest, que combina centenas de árvores, reduzindo a variância e aumentando a robustez preditiva. Este modelo mostrou boa capacidade para capturar interações entre variáveis, ainda que com menor transparência na interpretação.

Por fim, testou-se o XGBoost, um algoritmo de gradient boosting que tem demonstrado elevada eficácia em dados tabulares. A sua capacidade de identificar padrões não lineares e interações subtis revelou-se particularmente adequada ao contexto da valorização de jogadores. Foram ainda implementadas versões ponderadas pelos minutos jogados, assegurando que observações pouco fiáveis (jogadores com tempo residual em campo) não distorcessem os resultados.

Uma questão central na modelação é que não existe um único algoritmo que garanta o melhor desempenho em todos os contextos. Como referem Breiman (2001) e James et al. (2021), diferentes métodos exploram estruturas distintas nos dados: modelos lineares procuram relações proporcionais e transparentes; árvores e ensembles captam padrões não lineares; métodos de regularização controlam a multicolinearidade; e algoritmos de boosting potenciam ganhos incrementais de performance através de iterações sucessivas. Neste enquadramento, a opção metodológica desta dissertação foi comparar abordagens complementares, de complexidade crescente, conciliando rigor preditivo com capacidade interpretativa.

A implementação prática estruturou-se em três blocos principais de código. O pipeline central correspondeu ao guião base, responsável pela preparação dos dados, construção de variáveis derivadas e treino dos diferentes algoritmos (Regressão Linear, Elastic Net, Random Forest e XGBoost). Paralelamente, foram desenvolvidos dois scripts experimentais: no primeiro, avaliou-se o impacto de diferentes formas de ponderação consoante os minutos jogados, de modo a testar a robustez dos modelos perante atletas com reduzida utilização; no segundo, exploraram-se configurações alternativas de tuning, verificando até que ponto ajustes em hiperparâmetros poderiam influenciar o desempenho preditivo.

A Figura 4.11 resume de forma esquemática o pipeline metodológico seguido neste estudo, desde a recolha e filtragem inicial dos dados até à definição do target, particionamento em treino e teste, e aplicação dos diferentes modelos de aprendizagem automática.

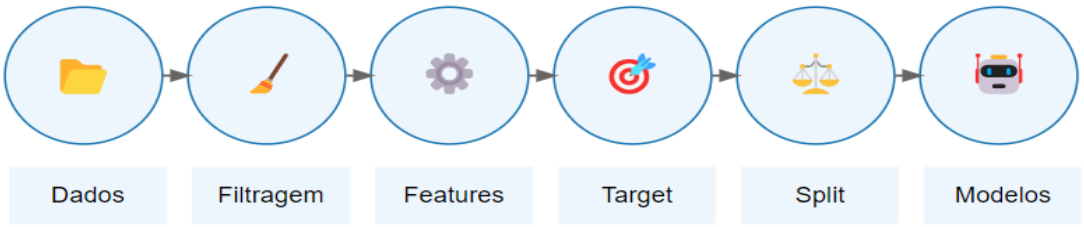


Figura 4.11. Pipeline metodológico da investigação.

4.5 Evaluation

A fase de avaliação correspondeu a um momento de pausa e reflexão sobre o caminho percorrido. Mais do que olhar apenas para números ou métricas, procurou-se perceber se os modelos e os procedimentos adotados estavam, de facto, a responder às questões de partida e a fazer sentido no contexto do estudo.

Nesta etapa foi essencial rever as escolhas feitas em fases anteriores, questionar a sua adequação e identificar pontos onde ainda havia margem para melhorar. A avaliação serviu, assim, não como um fecho definitivo, mas como uma oportunidade de validar decisões, reconhecer limitações e orientar os próximos passos.

O caráter iterativo do CRISP-DM ficou aqui bem visível: sempre que se identificaram aspetos a refinar, houve abertura para regressar à preparação dos dados ou à modelação, garantindo maior consistência ao trabalho. No fundo, esta fase funcionou como um momento de síntese e alinhamento, que permitiu consolidar o estudo antes de avançar para a apresentação dos resultados.

V. Resultados e Discussão

5.1 Análise Global dos Resultados

Neste capítulo são apresentados os resultados do trabalho desenvolvido. Para além de mostrar como os diferentes modelos se comportaram, procura-se também refletir sobre o que estes resultados significam no contexto do estudo. A ideia não é apenas listar números ou métricas, mas interpretar tendências, perceber o que correu melhor e o que pode ser melhorado.

Nesta secção apresento os resultados em três etapas: (i) começo pela comparação dos modelos do pipeline principal; (ii) avanço para as experiências adicionais (ajustes e testes de robustez); e (iii) termino com a análise dos fatores que mais influenciam a previsão do valor de mercado. Como já referido, todas as comparações usam o mesmo conjunto de dados e o mesmo teste fixo definidos no pipeline (tabela_model.csv e test_ids.csv), garantindo consistência e comparabilidade entre as experiências (Experiência. 1 e Experiência. 2).

5.1.1 Desempenho Global dos Modelos

O primeiro passo da análise consistiu em avaliar de forma comparativa os diferentes algoritmos de previsão do valor de mercado. O objetivo foi perceber não apenas quais modelos se ajustavam melhor aos dados, mas também até que ponto eram capazes de produzir estimativas consistentes e próximas dos valores reais. A Tabela 5.1 apresenta os resultados dessa comparação, sintetizando os principais indicadores de desempenho (RMSE, MAE, RMSLE e R^2) para cada algoritmo testado.

	RMSE	MAE	RMSLE	R^2
Decision Tree	13,33M €	6,97M €	1.244	0.573
Elastic Net	10,88M €	4,94M €	0.738	0.716
Random Forest	10,23M €	4,38M €	0.549	0.748
Linear Regression	9,79M €	4,38M €	0.830	0.770
LM (splines+int)	9,55M €	4,27M €	0.815	0.781
XGBoost	7,84M €	3,89M €	0.583	0.852

Tabela 5.1. Desempenho dos modelos preditivos.

A regressão linear múltipla constituiu o ponto de partida, servindo como modelo de referência (*baseline*) para a avaliação comparativa. Consideraram-se como variáveis independentes a idade (e idade²), altura, minutos jogados, golos por 90 minutos, assistências por 90 minutos, valor agregado do clube (*club_mv_log*) e pico de valorização pré-época (*peak_pre_log*). O modelo apresentou um coeficiente de determinação (R^2) de aproximadamente 0,77, o que indica que conseguiu explicar uma parte substancial da variação observada no valor de mercado dos jogadores. No entanto, mostrou limitações na captura de relações não lineares, refletidas também nos erros: o RMSE foi de cerca de 9,8M e o MAE rondou os 4,4M, valores que evidenciam discrepâncias médias significativas entre previsões e valores reais.

Com a introdução de splines e da interação entre minutos jogados e golos por 90 minutos, observaram-se ganhos modestos. O modelo atingiu um R^2 de 0,78, com redução do RMSE para 9,5M e do MAE para 4,3M. Ainda que a melhoria seja pequena, evidencia que a incorporação de relações não lineares permite captar de forma mais realista alguns padrões de valorização, mas continua insuficiente para lidar com toda a complexidade do problema.

Adicionalmente, verificou-se a presença de heterocedasticidade nos resíduos destes modelos lineares (teste de Breusch–Pagan, $p < 0{,}001$), indicando que a variância dos erros aumentava para valores de mercado mais elevados. Para lidar com este problema, optou-se por utilizar erros-padrão robustos (HC1), assegurando que os testes de significância dos coeficientes permaneciam válidos. Com esta correção, confirmaram-se como significativos os efeitos da idade

(não linear), minutos jogados, golos/90', assistências/90', valor do clube e pico pré-época, enquanto a interação minutos×golos/90 não apresentou relevância estatística.

Estes resultados evidenciam que os modelos de ensemble, como o Random Forest e o XGBoost, são mais eficazes na previsão do valor de mercado. Esta constatação vai ao encontro do que se pretendia avaliar relativamente à capacidade dos algoritmos em captar a complexidade do fenómeno (QI1) e confirma a hipótese de que métodos não lineares apresentariam melhor desempenho face a abordagens lineares tradicionais.

O Elastic Net, por sua vez, revelou utilidade na estabilização dos coeficientes, sobretudo em cenários com forte colinearidade entre variáveis. No entanto, o desempenho preditivo ficou aquém: R^2 de apenas 0,72, RMSE de 10,9M e MAE próximo de 4,9M. Isto demonstra que, embora traga ganhos de estabilidade relativamente a modelos lineares tradicionais, não conseguiu traduzir essa robustez em maior capacidade preditiva.

As árvores de decisão simples foram as que registaram pior desempenho. Com um R^2 de apenas 0,57 e um RMSE superior a 13,3M, acompanharam ainda um MAE de cerca de 7M. Estes resultados confirmam que modelos isolados e altamente segmentados são frágeis para lidar com fenómenos multifatoriais e ricos em interações subtis.

A Random Forest melhorou este quadro, ao combinar diversas árvores e reduzir a variância das previsões individuais. O modelo obteve um R^2 de 0,75, com RMSE de 10,2M e MAE de 4,4M. Embora mais estável e adequado a relações não lineares, o desempenho permaneceu aquém do necessário para uma previsão fiável de valores de mercado.

A diferença mais expressiva surgiu com o XGBoost, que superou de forma clara todas as alternativas. Este modelo alcançou um R^2 de 0,85, explicando mais de 85% da variabilidade dos valores de mercado. No que toca ao erro médio, destacou-se com um RMSE de 7,8M e um MAE de 3,9M, consideravelmente inferiores aos registados pelos outros modelos. Também no RMSLE (0,58) evidenciou robustez, revelando maior equilíbrio ao lidar com jogadores em diferentes escalas de valorização.

Importa referir que o valor de mercado foi transformado logaritmicamente para efeitos de modelação, mas os erros apresentados (RMSE ≈ 9,8M e MAE ≈ 4,4M) foram calculados após a inversão dessa transformação, de modo a refletirem discrepâncias médias na escala original dos valores de mercado.

Os resultados confirmam que, embora os modelos lineares e árvores simples funcionem como referências iniciais, são os métodos de ensemble e, sobretudo, de boosting que conseguem capturar de forma mais eficaz a complexidade do fenómeno estudado. O XGBoost destacou-se

claramente como a abordagem mais sólida, conjugando elevada capacidade explicativa com erros substancialmente reduzidos, justificando a sua adoção como modelo central na análise subsequente.

5.1.2 Experiência 1 – Exploração incremental de variantes

Na primeira experiência foram testadas quatro variantes do processo de modelação, organizadas de forma incremental. O objetivo foi perceber em que medida, diferentes escolhas de features, ponderações e configurações de tuning afetavam o desempenho preditivo do modelo, em particular quando se utilizava o XGBoost como algoritmo central.

Na versão A (baseline), o modelo foi treinado apenas com as features originais, sem qualquer engenharia adicional. Serviu como ponto de partida para avaliar o impacto de melhorias posteriores. Apesar de já superar os modelos lineares e baseados em árvores simples, o XGBoost obteve aqui um R^2 de 0,80, com um RMSE de 9,1 M € e um MAE de 4,0 M €. Estes valores mostraram um ganho claro relativamente às abordagens anteriores, mas também espaço para otimização.

Na versão B, acrescentaram-se novas variáveis derivadas, como o *involvement90* (golos + assistências por 90'), a raiz quadrada dos minutos e o logaritmo dos minutos. Estas transformações foram pensadas para capturar dinâmicas de envolvimento ofensivo e reduzir a influência de assimetrias extremas. O resultado foi uma melhoria modesta: o R^2 subiu para 0,81 e o RMSE caiu para 8,9 M €, revelando que a engenharia de variáveis acrescenta valor, ainda que de forma incremental.

A versão C introduziu pesos proporcionais à raiz quadrada dos minutos jogados, penalizando a influência de jogadores com poucos dados fiáveis. Este ajuste mostrou-se eficaz: o R^2 aumentou para 0,85 e o RMSE desceu para 7,8 M €, com um MAE de 3,7 M €. Em termos práticos, o modelo tornou-se mais estável e menos vulnerável a outliers associados a atletas com reduzida utilização.

Finalmente, a versão D correspondeu a um tuning agressivo do XGBoost, com exploração mais extensa de hiperparâmetros (profundidade máxima, taxa de aprendizagem, regularizações, etc.). Nesta fase, além da versão apenas afinada, testou-se também a combinação com os pesos por minutos (D_w). Os ganhos foram claros: a versão D atingiu um R^2 de 0,84 e RMSE de 8,1

M €, mas foi o modelo D_w (tuned + pesos) que se destacou como o melhor de toda a experiência, alcançando um R^2 de 0,89, com RMSE de 6,7 M € e MAE de 3,4 M €.

Os resultados detalhados destas quatro variantes encontram-se resumidos na Tabela 5.2, que apresenta as métricas de desempenho (RMSE, MAE, RMSLE e R^2) para cada configuração do XGBoost.

	RMSE	MAE	RMSLE	R^2
XGBoost (A)	9,12M €	3,99M €	0.548	0.80
XGBoost (B)	8,96M €	3,89M €	0.543	0.81
XGBoost (C, weights)	7,78M €	3,72M €	0.564	0.85
XGBoost (D, tuned+)	8,11M €	3,73M €	0.564	0.84
XGBoost (D, tuned+, w)	6,74M €	3,43M €	0.562	0.89

Tabela 5.2. Métricas de desempenho do XGBoost.

A análise gráfica reforçou estas conclusões. A relação entre valores reais e previstos mostrou menor dispersão e maior aderência à linha ideal no modelo ponderado, enquanto as métricas confirmaram a redução de erros graves em jogadores de maior valor. As variáveis mais importantes incluíram valor do clube (log), minutos jogados e pico pré-época (log), que emergiram tanto na análise de importância do XGBoost como nos contributos SHAP. Ablations adicionais confirmaram a relevância destas variáveis: a remoção de qualquer uma delas levou a perdas substanciais de desempenho.

Para além de confirmar a superioridade do XGBoost em termos preditivos, estas experiências evidenciam também a importância de estratégias complementares — como a engenharia de variáveis, a ponderação por utilização e o tuning extensivo — no aumento da qualidade das previsões. No conjunto, a Experiência 1 permitiu não só identificar o modelo mais robusto (D_w), mas também compreender quais abordagens metodológicas acrescentam valor ao processo de modelação. Esta base sólida abre caminho para a análise subsequente, onde se testará a capacidade de generalização do modelo em diferentes cenários.

A Figura 5.1, sintetiza esta trajetória de forma clara, permitindo perceber como cada passo contribuiu para o ganho final.

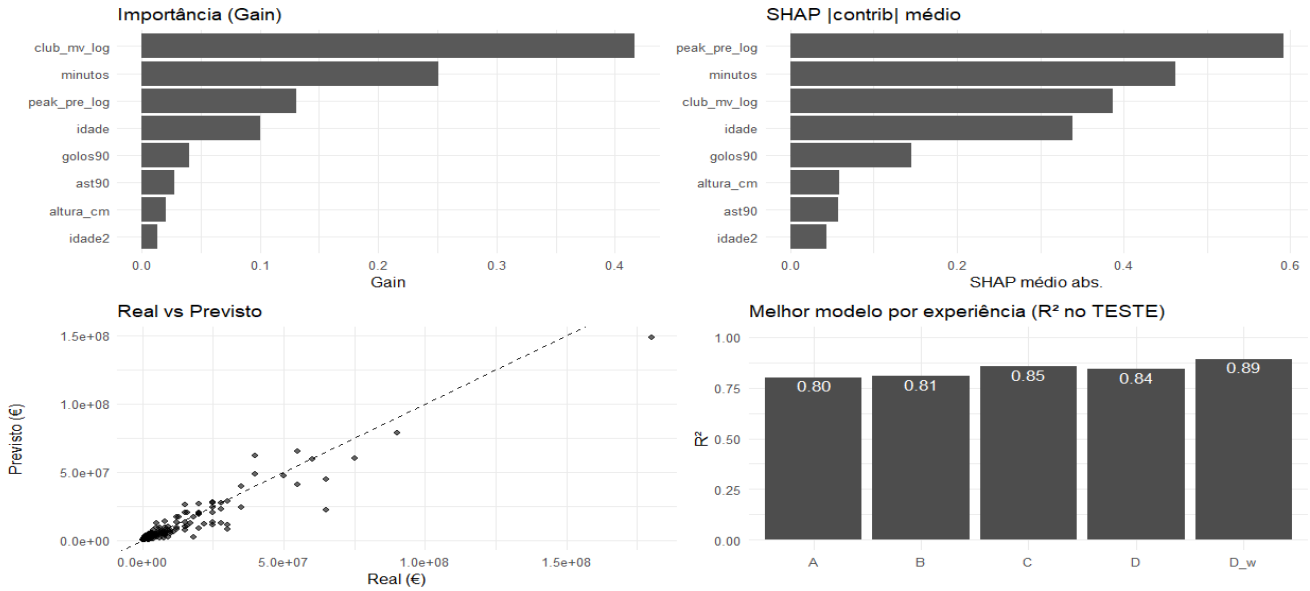


Figura 5.1. Avaliação do XGBoost: variáveis, previsões e desempenho.

5.1.3 Experiência 2 — Robustez e ablações do modelo final

Nesta experiência, procurou-se avaliar a robustez do modelo final através de análises de ablação (*ablation studies*). Este procedimento consiste em remover de forma controlada determinadas variáveis do modelo, de modo a medir até que ponto cada uma contribui para o desempenho preditivo. Ao comparar os resultados com e sem certas variáveis, é possível identificar aquelas que são mais determinantes para a previsão do valor de mercado e avaliar a sensibilidade global do modelo face à sua ausência.

Depois de identificarmos, na Experiência 1, o XGBoost com pesos por minutos e tuning agressivo (D_w) como o modelo mais eficaz, o passo seguinte passou por avaliar a sua robustez, compreender a importância das variáveis-chave e testar a qualidade prática das previsões. Esta análise foi organizada em três eixos: testes multi-seed, ablações e análise de calibração/segmentos de valor.

No primeiro eixo, de robustez, treinámos três variantes — A (baseline XGBoost), C (XGBoost com pesos por minutos) e D_w (tuned + pesos) — repetindo o treino cinco vezes com diferentes seeds. Os resultados mostraram uma evolução clara: a versão A manteve-se consistente, mas limitada (R^2 médio de 0,80 e RMSE em torno de 9,2 M€), a versão C já evidenciou melhorias (R^2 de 0,84; RMSE de 8,1 M€), e a versão D_w confirmou-se como a mais robusta

(R^2 médio de 0,86; RMSE de 7,6 M€), com menor variabilidade entre execuções. Ou seja, para além do ganho de desempenho, o modelo final revelou-se também mais estável.

Os resultados desta experiência permitem responder à QI3, que procurava avaliar até que ponto os modelos são robustos a diferentes configurações e variações nos dados. As repetições multi-seed mostraram que, independentemente da configuração inicial, os modelos mantiveram níveis consistentes de desempenho, com destaque para o XGBoost D_w, que apresentou menor variabilidade entre execuções (R^2 médio de 0,86; RMSE ≈ 7,6M€). Esta estabilidade demonstra que a performance não depende de forma crítica de escolhas arbitrárias de parametrização, reforçando a robustez e generalização do modelo.

A Figura 5.2 resume estes resultados, evidenciando a evolução de desempenho e a maior estabilidade do modelo D_w face às alternativas.

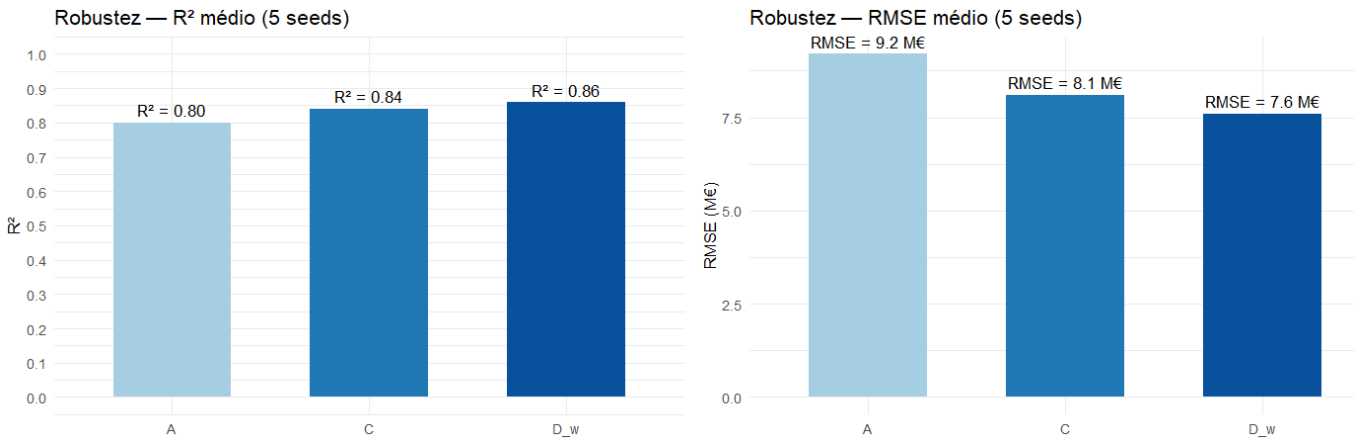


Figura 5.2. Robustez dos modelos: R^2 e RMSE médio.

A análise confirma o efeito não linear da idade, com valorização em idades intermédias e depreciação posterior. Este resultado valida a hipótese de que o ciclo da carreira influencia o valor de mercado (H1) e fornece uma resposta consistente à questão da influência de fatores individuais na valorização (QI2).

No segundo eixo, dedicado às ablações, cujo objetivo foi medir o impacto de variáveis específicas no desempenho do modelo final. A remoção de *club_mv_log* e de *minutos* (mantendo ainda assim os pesos) degradou moderadamente o desempenho ($R^2 \approx 0.80$), enquanto a exclusão de *peak_pre_log* revelou-se muito mais penalizadora, reduzindo o R^2 para apenas

0.72. Estes resultados confirmam que, embora várias variáveis contribuam, o pico de valorização pré-época e a força do clube são determinantes no processo de estimação.

A Figura 5.3 ilustra estes resultados, evidenciando o impacto mais acentuado da remoção do pico pré-época no desempenho do modelo.

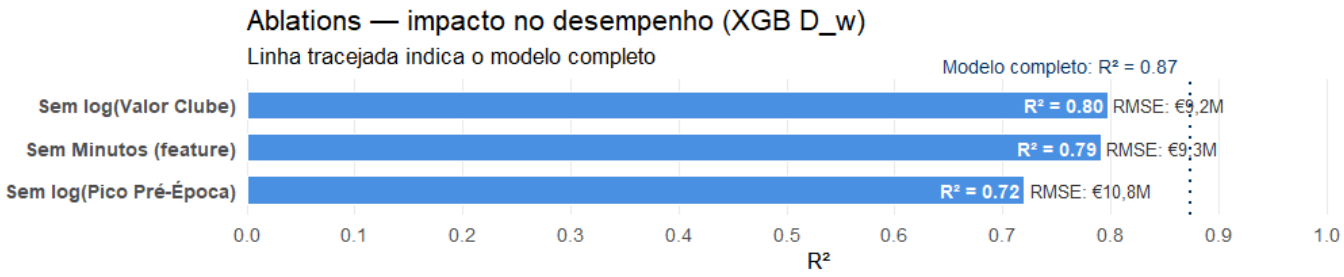


Figura 5.3. Resultados das ablations no modelo XGBoost (D_w).

Esta experiência também fornece evidência clara relativamente à QI4. As análises de importância e dependência SHAP revelaram que as métricas de performance desportiva, como minutos jogados e golos por 90 minutos, têm um impacto direto na valorização, mas o seu efeito é potenciado quando considerado em conjunto com fatores contextuais, nomeadamente o valor agregado do clube e o pico de valorização pré-época. As ablações confirmaram esta articulação: a remoção de variáveis contextuais levou a perdas substanciais de desempenho, em especial no caso do pico pré-época, cujo impacto revelou ser decisivo ($R^2 \approx 0{,}72$). Estes resultados demonstram que o valor de mercado resulta de uma interação entre desempenho individual e contexto coletivo, sendo necessário considerar ambas as dimensões para uma previsão adequada.

Deste modo, confirma-se a hipótese H4, que antecipava a importância do contexto coletivo e da dimensão económica do clube como fator explicativo do valor de mercado individual.

No terceiro eixo avaliou a qualidade prática das previsões. A análise de calibração mostrou um alinhamento global bastante satisfatório entre valores reais e previstos, ainda que com uma tendência de subestimação dos jogadores mais valorizados (inclinação da reta ajustada ≈ 1,11). Ao estratificar o erro por segmentos de valor, observou-se que a precisão é bastante elevada para atletas de menor valor (<1M, RMSE ≈ 0,5 M€), mas que o erro absoluto aumenta de forma significativa nos jogadores de topo (>15M, RMSE ≈ 13,9 M€). Tal comportamento é natural devido à amplitude da escala, mas reforça o desafio particular de prever corretamente os valores de mercado mais elevados.

Complementámos a análise com curvas de dependência SHAP, que ajudam a interpretar o funcionamento interno do modelo. A análise mostra que mais minutos em campo estão associados a um maior valor de mercado, mas com ganhos marginais decrescentes. O peak_pre_log revelou-se decisivo a partir de certos limiares, reforçando a importância do histórico máximo já alcançado na valorização atual. Já a club_mv_log apresentou uma relação positiva quase linear, confirmando que o contexto competitivo do clube onde o jogador atua tem impacto direto no seu valor.

Entre as variáveis analisadas, destacam-se os minutos jogados e os golos por 90 minutos como determinantes fundamentais, enquanto as assistências apresentaram um contributo reduzido. As curvas de dependência SHAP (Figura 5.4) tornam visível esse efeito, evidenciando padrões não lineares e confirmando parcialmente a hipótese H2, bem como a resposta à questão sobre os fatores mais influentes na explicação do valor de mercado (QI2).

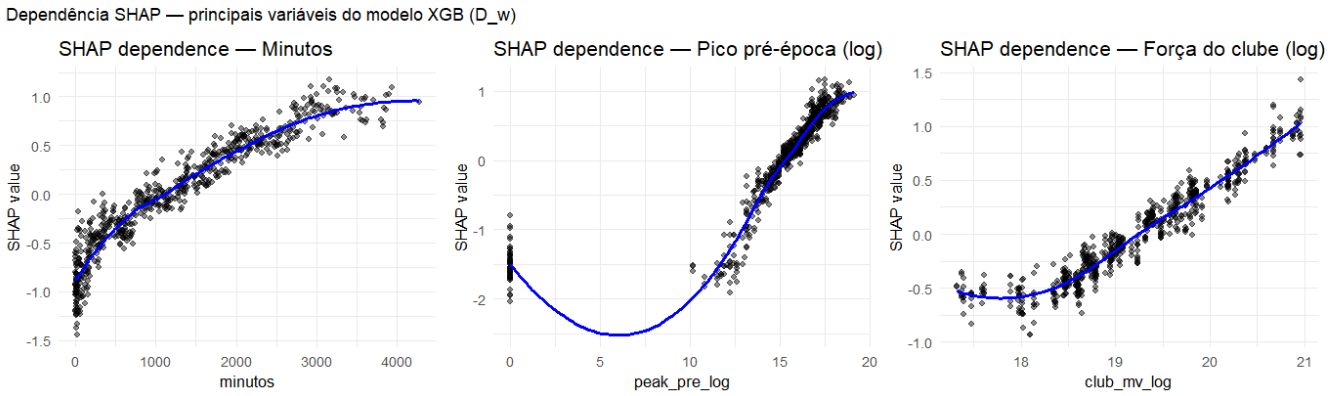


Figura 5.4. Curvas de dependência SHAP para as principais variáveis do modelo XGBoost (D_w).

No que diz respeito à H3, que antecipava melhorias de desempenho com a inclusão de variáveis contextuais, os resultados confirmaram plenamente essa expectativa. O valor do clube e o pico de valorização pré-época emergiram como determinantes fundamentais do modelo, estando entre as variáveis com maior contributo explicativo. A sua presença revelou-se decisiva, como demonstrado nas análises de ablação, em que a sua remoção provocou perdas substanciais de desempenho. Ainda assim, os maiores ganhos em termos de métricas foram alcançados quando a inclusão destas variáveis foi combinada com estratégias adicionais, como o tuning extensivo. Assim, a H3 é suportada, sublinhando a relevância das variáveis contextuais, especialmente quando articuladas com outras abordagens de modelação.

Demonstrou-se que o modelo final não só é consistente e robusto em diferentes execuções, como também depende de variáveis com forte significado desportivo e económico. A integração equilibrada de utilização em campo, histórico individual e contexto coletivo explica a sua capacidade de previsão. Além disso, os resultados destacam que, embora o modelo funcione bem em toda a distribuição, os maiores desafios residem na previsão dos atletas de elite, onde os valores e os fatores externos tendem a ser mais voláteis. Esta análise não apenas valida a qualidade do modelo final, como também reforça a sua interpretabilidade, um aspeto crucial para aplicações práticas.

A evidência encontrada suporta a hipótese H5, segundo a qual o histórico de valorização, expresso pelo pico pré-época, atua como um fator de ancoragem relevante na determinação do valor de mercado atual.

VI. Conclusões

O presente estudo teve como objetivo compreender e prever o valor de mercado dos jogadores avançados que atuaram nas cinco principais ligas europeias na época 2023/2024, através da aplicação de diferentes algoritmos de aprendizagem automática e da análise dos fatores que mais contribuem para essa valorização. A escolha desta posição específica deveu-se à centralidade dos avançados no jogo e à forte associação do seu desempenho ofensivo ao processo de valorização. Ao longo do trabalho foram definidos objetivos e hipóteses que orientaram toda a investigação, procurando dar resposta a questões que permanecem atuais na análise económica do futebol: até que ponto o valor de mercado se explica pelo desempenho em campo, em que medida o contexto coletivo e histórico do jogador influencia a sua cotação e que modelos são mais adequados para captar a complexidade destes fenómenos.

A análise comparativa entre algoritmos revelou de forma clara que métodos baseados em árvores de decisão, como o Random Forest e o XGBoost, apresentaram um desempenho superior face a modelos lineares clássicos. Enquanto a regressão linear simples serviu de linha de base, mas mostrou limitações na captura de relações não lineares e de interações, os ensembles evidenciaram maior capacidade preditiva e maior robustez perante a heterogeneidade dos dados. Esta conclusão reforça a ideia, já presente na literatura, de que o mercado de transferências se rege por dinâmicas complexas, não redutíveis a relações simples. O facto de os ensembles terem mantido uma performance estável mesmo em cenários de maior variabilidade mostra que são ferramentas adequadas para lidar com a natureza incerta e multifatorial do valor de mercado.

No que respeita às variáveis, os resultados evidenciam a importância de fatores tanto de desempenho individual como de contexto. Entre as métricas de performance desportiva, destacaram-se os minutos jogados e os golos por 90 minutos como determinantes fundamentais do valor de mercado. A regularidade e a confiança do treinador, traduzidas em tempo de jogo, surgiram como sinais claros de valorização. Os golos, por seu lado, continuam a ser o principal indicador de eficácia para avançados e confirmaram-se como variável de grande peso nos modelos. Curiosamente, as assistências por 90 minutos tiveram um impacto reduzido, contrariando parcialmente as expectativas iniciais e sugerindo que o mercado valoriza mais a finalização direta do que a contribuição indireta.

A idade confirmou o padrão não linear previsto, exercendo uma influência que varia em sentidos opostos: por um lado, a juventude traduz potencial de crescimento e investimento

futuro; por outro, idades mais avançadas remetem para declínio físico e menor perspetiva de retorno. Esta curva de valorização ao longo da carreira, com pico em idades intermédias, está em linha com estudos anteriores e reforça a ideia de que o mercado de transferências integra, de forma tácita, um horizonte temporal de rendimento esperado. A análise reforçou também a influência da idade, confirmando o padrão não linear de valorização com pico em idades intermédias e depreciação posterior, em linha com a hipótese formulada. Do mesmo modo, a consistência competitiva, medida pelos minutos jogados, mostrou-se um dos fatores mais robustos e estáveis em todos os modelos, funcionando como sinal claro de confiança e valorização no mercado.

Para além do desempenho em campo, os fatores contextuais assumiram relevo significativo. O valor agregado do clube revelou-se um dos determinantes mais consistentes, confirmando que jogar num clube de maior dimensão e prestígio potencia a valorização individual, independentemente do desempenho estrito do jogador. Do mesmo modo, o pico de valorização pré-época assumiu-se como um marco determinante no processo de valorização: jogadores que já atingiram valores elevados em momentos anteriores tendem a manter parte dessa valorização, mesmo que o rendimento atual não corresponda na totalidade. Este efeito de histórico confirma o peso da reputação e da perceção de mercado na formação do preço, em linha com a literatura sobre mecanismos de ancoragem.

Outro contributo relevante deste trabalho foi a análise da robustez dos modelos. A introdução de procedimentos como a winsorização de rácios ofensivos ou a ponderação por minutos jogados mostrou que os resultados se mantiveram estáveis, reduzindo a influência de outliers e confirmando que as conclusões extraídas não dependem de configurações arbitrárias, mas refletem padrões estruturais consistentes. Esta estabilidade é fundamental para dar confiança às aplicações práticas do estudo, assegurando que as previsões não são excessivamente sensíveis a variações marginais nos dados.

Esta dissertação apresenta contributos em diferentes planos. No académico, enriquece o debate em torno do valor de mercado em futebol, mostrando como abordagens de ciência de dados permitem analisar este fenómeno de forma mais robusta e fundamentada. Demonstra ainda que o cruzamento entre métricas desportivas, fatores contextuais e históricos é essencial para compreender a valorização, confirmando algumas ideias da literatura. A nível prático, este estudo oferece pistas relevantes para clubes, agentes e investidores: reforça a importância de avaliar não apenas o desempenho atual, mas também o contexto competitivo e a trajetória histórica; ajuda a identificar variáveis-chave que podem antecipar a valorização de um jogador; e

mostra que os algoritmos de ensemble são ferramentas promissoras para apoiar decisões num mercado altamente incerto.

Relativamente às limitações, o foco exclusivo em avançados, embora justificado, restringe a generalização das conclusões a outras posições em campo. A análise centrou-se ainda apenas numa época, o que limita a observação de dinâmicas temporais mais longas. A ausência de variáveis contratuais e de dados sobre lesões ou fatores extra-desportivos constitui outra limitação relevante, já que estes elementos são conhecidos por influenciar fortemente o valor de mercado. Além disso, os valores utilizados baseiam-se em estimativas do Transfermarkt, que embora amplamente usados em investigação, podem divergir das negociações reais.

Para investigações futuras, será pertinente ampliar o horizonte de análise a outras posições e épocas, de forma a verificar se os padrões aqui identificados se mantêm. A integração de variáveis contratuais, dados de lesões e métricas avançadas de performance (como *xG – expected goals*, que estima a probabilidade de cada remate resultar em golo, ou métricas defensivas detalhadas) poderá enriquecer a explicação do valor de mercado. Outra via promissora será a inclusão de variáveis de impacto mediático e perceção pública, como indicadores de redes sociais ou análises de sentimento, capazes de captar dimensões mais intangíveis da valorização. Por fim, a exploração de modelos mais recentes, como arquiteturas de deep learning aplicadas a dados tabulares ou combinações híbridas que juntem previsibilidade e explicabilidade, poderá trazer avanços metodológicos significativos.

Este estudo mostra que o valor de mercado de um jogador resulta de uma articulação entre desempenho desportivo, contexto coletivo e histórico de valorização, e que os modelos de machine learning oferecem uma via sólida para compreender e antecipar estas dinâmicas. Ao aplicar metodologias de ciência de dados a um tema de grande relevância no futebol profissional, a investigação contribui para uma compreensão mais rigorosa, mas também mais prática, das forças que moldam o mercado de transferências. Deste modo, a investigação não se limita ao plano académico, mas acrescenta também uma reflexão prática sobre as dinâmicas que moldam a valorização dos jogadores.

A investigação conseguiu dar resposta às questões que estavam na origem do trabalho, mostrando a vantagem dos algoritmos de ensemble em relação a métodos lineares, bem como o papel central de variáveis individuais e contextuais na explicação do valor de mercado. Os resultados revelaram-se consistentes mesmo perante diferentes configurações, o que reforça a confiança nas conclusões alcançadas. No fim, cumpriram-se os objetivos definidos e foi

possível testar de forma clara as hipóteses propostas, deixando um contributo sólido para compreender melhor como se constrói a valorização dos avançados nas principais ligas europeias.

REFERÊNCIAS BIBLIOGRÁFICAS

Aydemir, A. E., Taskaya-Temizel, T., & Temizel, A. (2022). A machine learning ensembling approach to predicting transfer values. *SN Computer Science, 3*(201). https://doi.org/10.1007/s42979-022-01095-z

Binder, J., & Findlay, M. (2012). The effects of the Bosman ruling on national and club teams in Europe. *Journal of Sports Economics, 13*(2), 107–129. https://doi.org/10.1177/1527002511400278

Bondia, R., Gallo, L. A., & García, A. F. (2020). Determinants of soccer players' market values: A study of the Spanish league. *Journal of Physical Education and Sport, 20*(4), 1903–1910. https://doi.org/10.7752/jpes.2020.04258

Bryson, A., Frick, B., & Simmons, R. (2013). The returns to scarce talent: Footedness and player remuneration in European soccer. *Journal of Sports Economics, 14*(6), 606–628. https://doi.org/10.1177/1527002511435118

Deloitte. (2024). *Football Money League 2024*. Deloitte. https://www.deloitte.com/uk/en/services/consulting-financial/analysis/deloitte-football-money-league-2024.html

Dobson, S., & Goddard, J. (2011). *The economics of football* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511973864

Duran, P., & Rodríguez, M. (2015). Player valuation in football: An econometric analysis. *Applied Economics, 47*(35), 3703–3712. https://doi.org/10.1080/00036846.2015.1040544

FBref. (2024). *FBref – Football statistics and history*. https://fbref.com

FIFA. (2023). *Global transfer report 2023*. Fédération Internationale de Football Association. https://www.fifa.com/legal/transfer-report

Franceschi, M., Brocard, J. F., Follert, F., & Gouguet, J. J. (2023). Determinants of football players' valuation: A systematic review. *Journal of Economic Surveys, 38*(3), 577–600. https://doi.org/10.1111/joes.12552

Frick, B. (2007). The football players' labor market: Empirical evidence from the major European leagues. *Scottish Journal of Political Economy, 54*(3), 422–446. https://doi.org/10.1111/j.1467-9485.2007.00423.x

García-del-Barrio, P., & Pujol, F. (2007). Hidden monopsony rents in winner-take-all markets: Sport and economic contribution of Spanish soccer players. *Managerial and Decision Economics, 28*(1), 57–70. https://doi.org/10.1002/mde.1311

Gudmundsson, J., & Horton, M. (2017). Spatio-temporal analysis of team sports. *ACM Computing Surveys, 50*(2), 1–34. https://doi.org/10.1145/3054132

Gu, M., Guo, Y., & Hu, H. (2015). Market value and performance of soccer players in the Chinese Super League. *International Journal of Sports Science & Coaching, 10*(6), 1085–1096. https://doi.org/10.3390/ijerph16214238

Herm, S., Callsen-Bracker, H. M., & Kreis, H. (2014). When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community. *Sport Management Review, 17*(4), 484–492. https://doi.org/10.1016/j.smr.2013.12.006

Hofmann, J. (2018). Determinants of market values of soccer players – An econometric approach. *International Journal of Financial Studies, 6*(2), 46. https://doi.org/10.1111/joes.12552

Hoey, S., Peeters, T., & Principe, F. (2021). The transfer system in European football: A pro-competitive no-poaching agreement? *International Journal of Industrial Organization, 75*, 102695. https://doi.org/10.1016/j.ijindorg.2020.102695

Inan, T., & Cavas, L. (2021). Estimation of market values of football players through artificial neural networks: A model study from the Turkish Super League. *Applied Artificial Intelligence, 35*(13), 1022–1042. https://doi.org/10.1080/08839514.2021.1966884

KPMG Football Benchmark. (2023). *KPMG Football Benchmark database*. https://www.footballbenchmark.com

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774. https://doi.org/10.5555/3295222.3295230

Majewski, S. (2021). Football players' brand as a factor in performance rights valuation. *Journal of Physical Education and Sport, 21*(4), 1751–1760. https://doi.org/10.7752/jpes.2021.04222

Majewski, S. (2022). Social media metrics and football players' market value: Evidence from top European leagues. *Sport, Business and Management, 12*(5), 634–652. https://doi.org/10.1108/SBM-07-2021-0107

Malagón-Selma, P., Debón, A., & Domenech, J. (2023). Measuring the popularity of football players with Google Trends. *PLOS ONE, 18*(8), e0289213. https://doi.org/10.1371/journal.pone.0289213

Majewski, S. (2016). Identification of factors determining market value of the most valuable football players. *Central European Management Journal, 24*(3), 91–104. https://doi.org/10.7206/jmba.ce.2450-7814.177

Müller, J. C., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research, 263*(2), 611–624. https://doi.org/10.1016/j.ejor.2017.05.005

Paoli, A., & Pringle, R. (2016). Football performance and market value: Statistical evidence from Italian Serie A. *International Journal of Sport Management and Marketing, 16*(1/2), 120–139. https://doi.org/10.7206/jmba.ce.2450-7814.177

Poli, R., Besson, R., & Ravenel, L. (2020). *Transfer values in professional football: A model approach*. CIES Football Observatory Report. https://football-observatory.com/IMG/sites/mr/mr53/en/

Poli, R., Besson, R., & Ravenel, L. (2024). Statistical modeling of football players' transfer fees worldwide. *International Journal of Financial Studies, 12*(3), 93. https://doi.org/10.3390/ijfs12030093

Putra, M. D. A., Dewi, D. A., Putri, W. T., Hendrowati, R., & Kurniawan, T. B. (2023). Contributed factors in predicting market values of loaned-out players of English Premier League clubs. *International Journal of Advanced Computer Science and Applications, 14*(9), 359–368. https://doi.org/10.14569/IJACSA.2023.0140939

Quansah, T., Frick, B., Lang, M., & Maguire, K. (2021). The importance of club revenues for player salaries and transfer expenses: How does the coronavirus outbreak (COVID-19) impact the English Premier League? *Sustainability, 13*(9), 5154. https://doi.org/10.3390/su13095154

Rossi, A., Thrassou, A., & Vrontis, D. (2018). Football performance and player market value: Exploring the connection. *International Journal of Organizational Analysis, 26*(4), 662–678. https://doi.org/10.1108/IJOA-04-2013-0659

Sun, Y., & Gu, K. (2024). Prediction of football players' value in the transfer market of well-known European leagues based on FIFA 19 and real-world data. *The International Arab Journal of Information Technology, 21*(4), 723–734. https://doi.org/10.34028/iajit/21/4/13

Transfermarkt. (2024). *Transfermarkt database*. https://www.transfermarkt.com

UEFA. (2023). *UEFA Champions League technical report 2022/23*. Union of European Football Associations. https://www.uefa.com

UEFA. (2023). *European club finance and investment landscape report*. Union of European Football Associations. https://www.uefa.com/insideuefa/documentlibrary

Užík, M., Warias, R., & Glova, J. (2022). Management of transfer prices in professional football as a function of fan numbers. *Mathematics, 10*(16), 2982. https://doi.org/10.3390/math10162982

Zarka, J., Wicker, P., & Hallmann, K. (2024). Prediction of football players' value in the transfer market: Evidence from FIFA 19 and real-world data. *Journal of Sports Analytics, 10*(2), 123–139. https://doi.org/10.1109/TLA.2022.3191896

ANEXOS

Anexo A – Pipeline principal

A.1 Ficheiros de origem e período da amostra

Período: 2023-07-01 a 2024-06-30.

Tabela A.1. Ficheiros utilizados

Ficheiro (CSV)	Colunas usadas (principais)	Utilização
players.csv	player_id, name, date_of_birth, height_in_cm, position, foot, current_club_id, current_club_domestic_competition_id, highest_market_value_in_eur	Identificação e atributos do jogador; clube/competição; cálculo da idade; valor máximo histórico
player_valuations.csv	player_id, date, market_value_in_eur	Valor de mercado (alvo na época) e pico pré-época
appearances.csv	player_id, date, minutes_played, goals, assists	Agregação por jogador: minutos, golos, assistências; taxas por 90'
games.csv	game_id, competition_id, date	Ligar aparições à competição e restringir às Top-5
game_events.csv	game_id	Suporte ao mapeamento jogo–competição (sem *features* diretas)
competitions.csv	competition_id	Selecionar as Top-5 ligas (GB1, ES1, IT1, FR1, L1)
clubs.csv	club_id, total_market_value	Proxy da "força do clube" (valor de plantel)
transfers.csv	player_id, transfer_date, transfer_fee	Extração da última taxa de transferência

A.2 Variáveis e construção

Tabela A.2. Variáveis e construção.

Variável	Regra curta (fonte + transformação)
valor_mercado	player_valuations: último registo na época 2023/24 (alvo em €).
y_log	log1p(valor_mercado) — alvo usado no treino.
idade	Diferença entre players$date_of_birth e data do valuation (anos inteiros).
idade2	idade^2 (captar não-linearidade).
altura_cm	players$height_in_cm (sem transformação).
minutos	appearances: soma de minutes_played na época.
golos	appearances: soma de goals na época.
assistencias	appearances: soma de assists na época.
golos90	90 * golos / minutos → winsor no 99º pct.
ast90	90 * assistencias / minutos → winsor no 99º pct.
club_total_mv_eur	Força do clube: clubs$total_market_value ou soma de valuations pré-época do plantel (fallback).
club_mv_log	log1p(club_total_mv_eur).
peak_pre_epoca	Máximo de market_value_in_eur antes de 2023-07-01 (histórico do jogador).
peak_pre_log	log1p(peak_pre_epoca).

Tabela A.3. Contagem da amostra por liga (FW, 2023/2024).

Liga	N.º jogadores	Percentagem
La Liga	184	22,60%
Premier League	175	21,50%
Ligue 1	158	19,40%
Serie A	157	19,30%
Bundesliga	140	17,20%
Total	**814**	**100%**

Tabela A.4. Critérios de inclusão/exclusão e regras da amostra.

Bloco	Critério / Regra	Como foi verificado (ficheiro/campo)
Âmbito	Época: 2023-07-01 a 2024-06-30	Definido no pipeline (variáveis data_inicio, data_fim)
Âmbito	Competições: Top-5 (GB1, ES1, IT1, FR1, L1)	players.csv → current_club_domestic_competition_id (pertence a Top-5)
Âmbito	Posição alvo: Avançados (FW)	players.csv → position == "Attack"
Inclusão	Tem aparições na época (≥1)	appearances.csv → date no intervalo; contagem por player_id
Inclusão	Tem valuation na época (≥1)	player_valuations.csv → date no intervalo; existência por player_id
Inclusão	Clube de Top-5	players.csv → current_club_domestic_competition_id ∈ {GB1,ES1,IT1,FR1,L1}
Agregação/Chaves	Unidade: 1 linha por jogador-época	Chave estável player_id (após agregações)
Agregação/Chaves	Liga do jogador	Dominante pelas aparições (contagem por competition_id em games/appearances); *fallback* = current_club_domestic_competition_id de players.csv
Agregação/Chaves	Métricas de época	Somatórios: minutos, golos, assists (de appearances.csv); taxas golos90/ast90 = 90×(contagem/minutos)
Exclusão	Posições não FW (GK/DF/MF)	players.csv → position
Exclusão	Sem aparições na época	Zero ocorrências em appearances.csv no período
Exclusão	Sem valuation na época	Zero ocorrências em player_valuations.csv no período
Exclusão	Fora Top-5 ligas	current_club_domestic_competition_id não pertence ao conjunto
Exclusão	Duplicados	Deduplicação após agregação (chave player_id)
Notas técnicas	Alvo (y) = último market_value_in_eur da época	player_valuations.csv (mais recente por player_id no intervalo)
Notas técnicas	Idade calculada na data do valuation usado no alvo	players.csv (date_of_birth) + data do valuation
Notas técnicas	Winsorização de extremos	golos90 e ast90 winsorizados ao p99
Notas técnicas	Sem leakage	Força do clube e pico pré-época usam registos anteriores a data_inicio

Anexo B – Exploração adicional

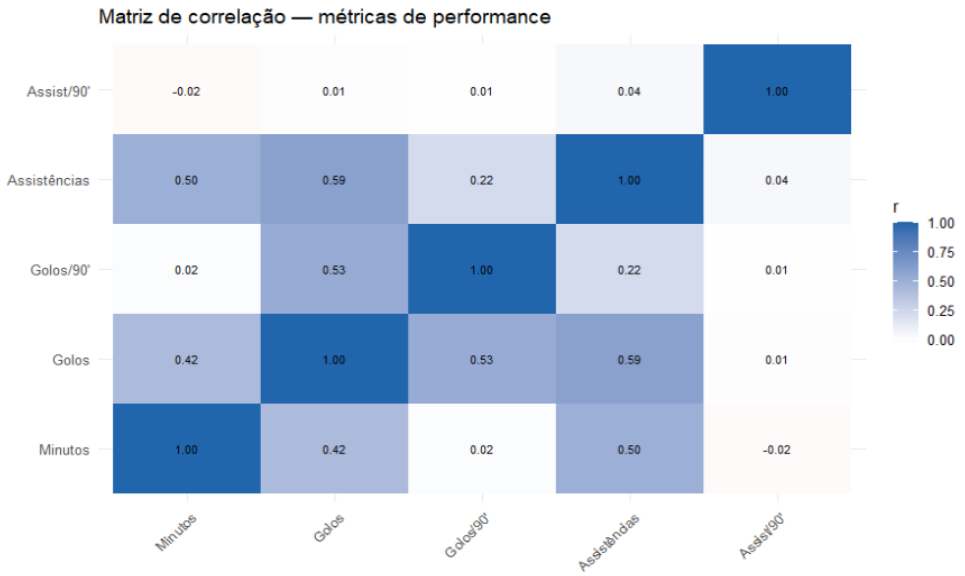


Figura B.1. Matriz de correlação – métricas de performance.

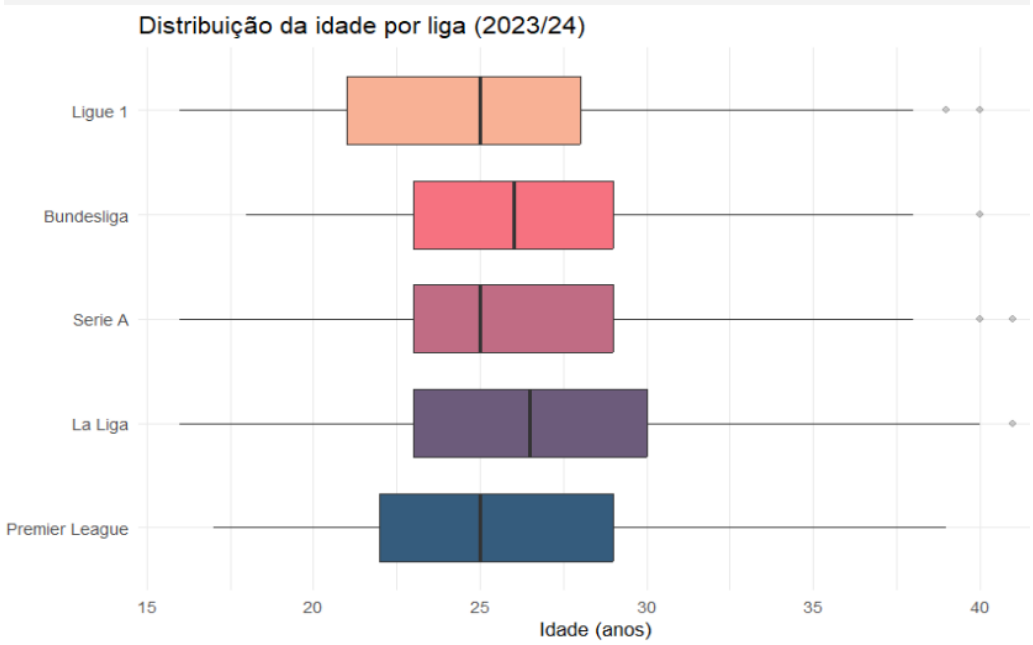


Figura B.2. Distribuição da idade por liga.

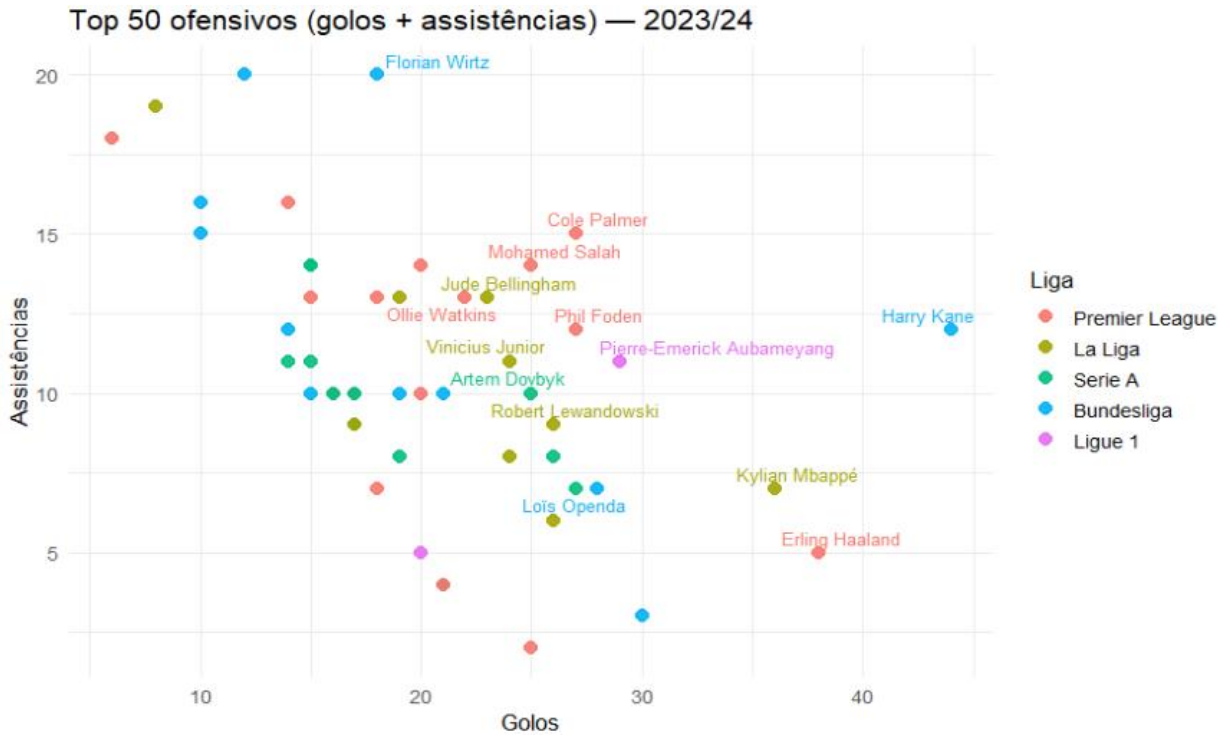


Figura B.3. Top 50 ofensivos — Golos + Assistências.

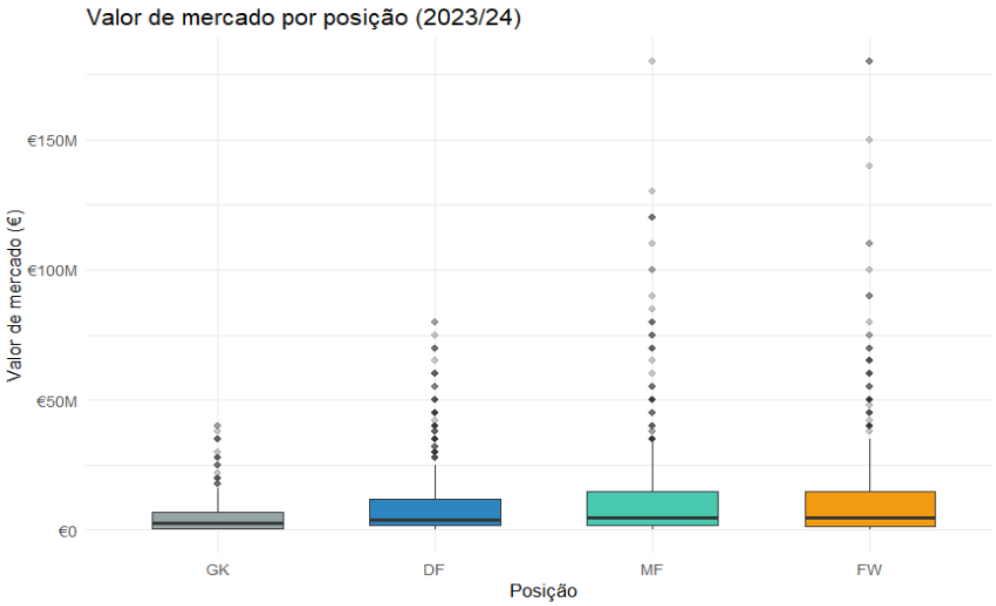


Figura B.4. Valor de mercado por posição.

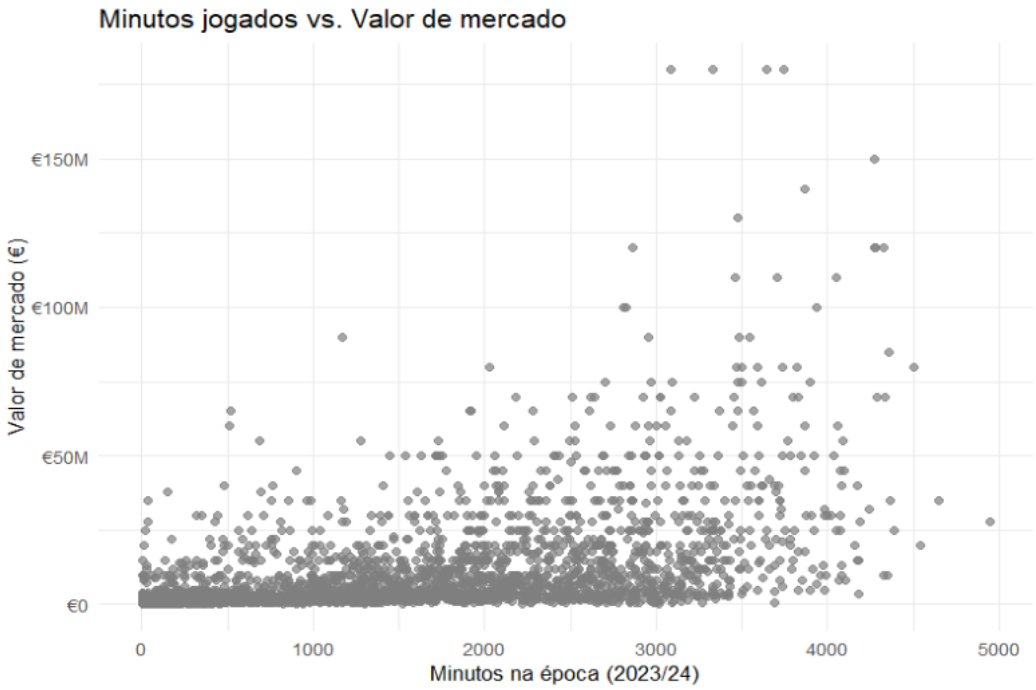


Figura B.5. Minutos jogados vs. Valor de mercado

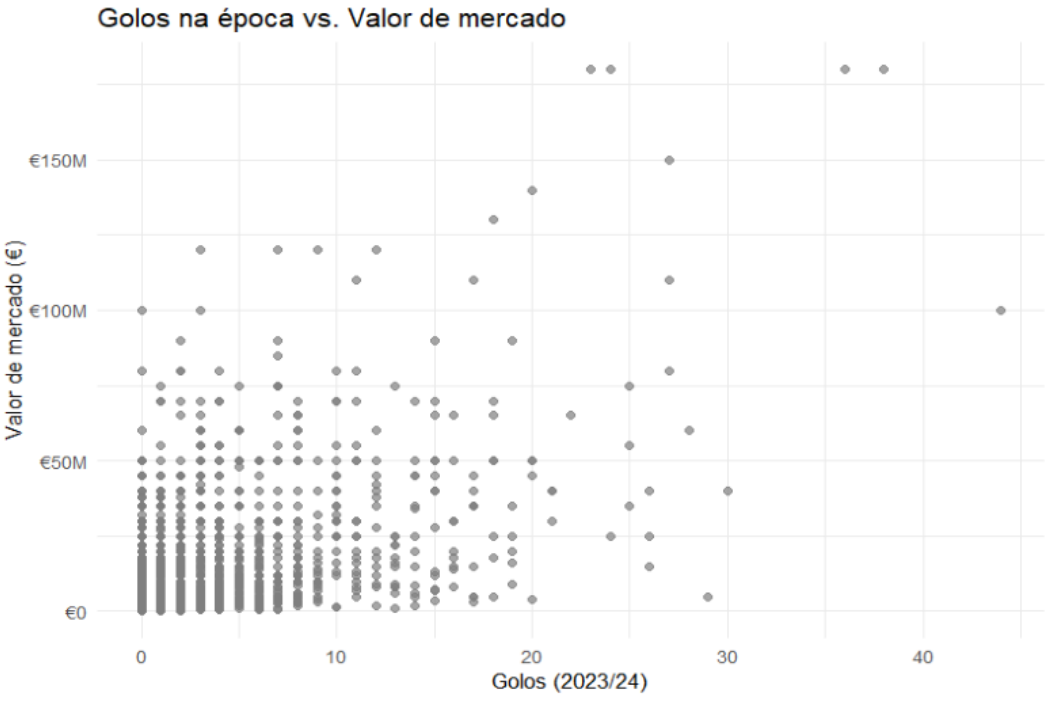


Figura B.6. Golos vs. Valor de Mercado.

Anexo C – Experiência 2 (robustez & ablations)

C.1 Robustez multi-seed (5 seeds)

Tabela C.1. Robustez por variante (média + desvio-padrão, 5 seeds)

Variante	RMSE (M€)	SD RMSE (M€)	MAE (M€)	SD MAE (M€)	RMSLE	SD RMSLE	R^2	SD R^2
D_w (XGB tuned + pesos)	7.57	0.45	3.68	0.13	0.574	0.008	0.862	0.016
C (XGB com pesos)	8.09	1.08	3.85	0.25	0.563	0.006	0.841	0.045
A (XGB base)	9.16	0.11	4.00	0.05	0.543	0.006	0.798	0.005



Figura C.1. Boxplot de R^2 por variante.

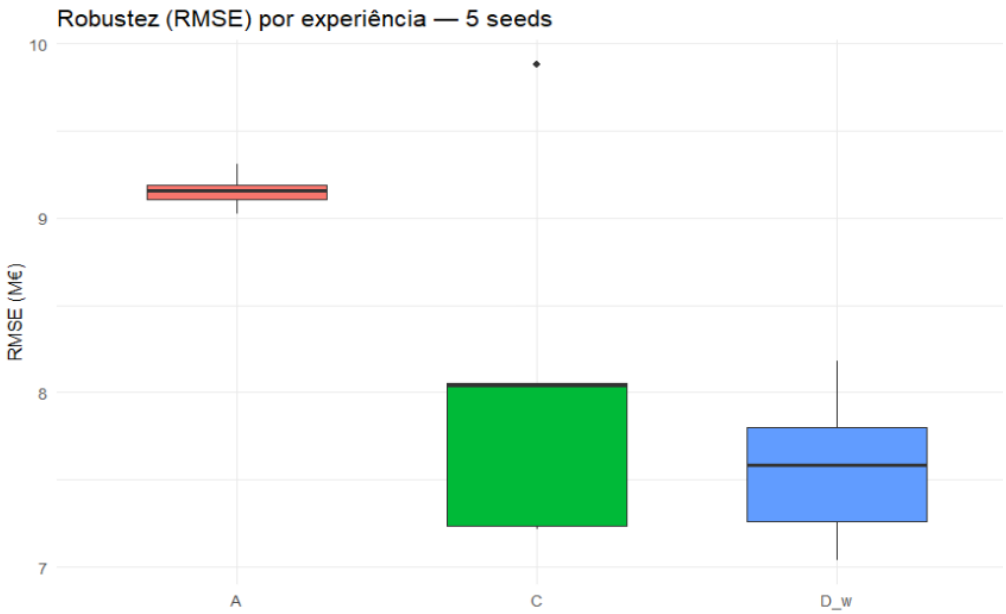


Figura C.2. Boxplot de RMSE (em M€) por variante.

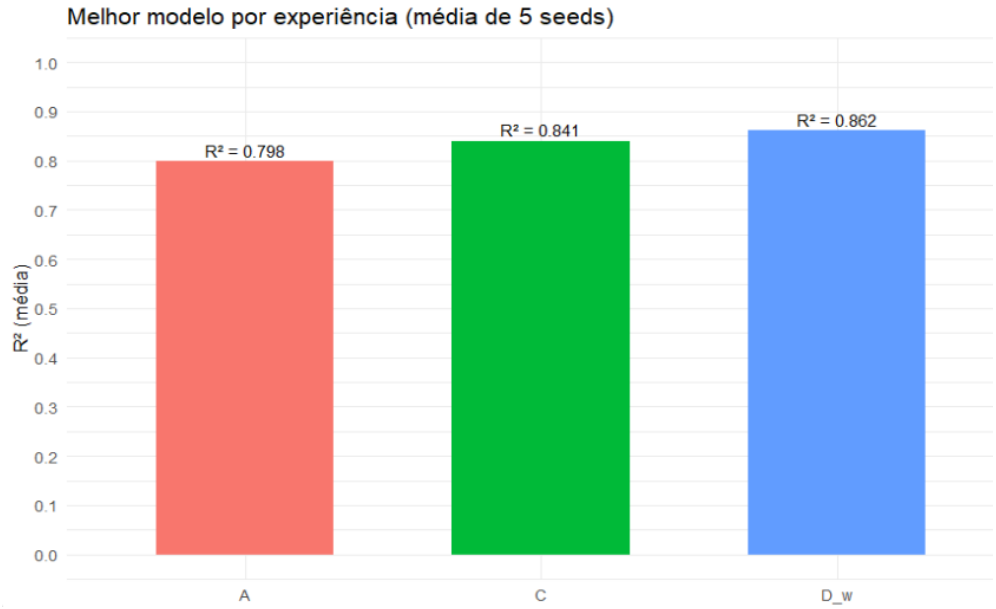


Figura C.3. Melhor variante (barras com R^2 médio).

C.2 Estudos de ablation (variáveis)

Tabela C.2. Ablations (R^2, RMSE, MAE).

Ablation	R^2	RMSE (M€)	MAE (M€)
a) sem club_mv_log	0.797	9.18	4.39
b) sem minutos (feature)	0.790	9.33	4.33
c) sem peak_pre_log	0.720	10.80	4.82

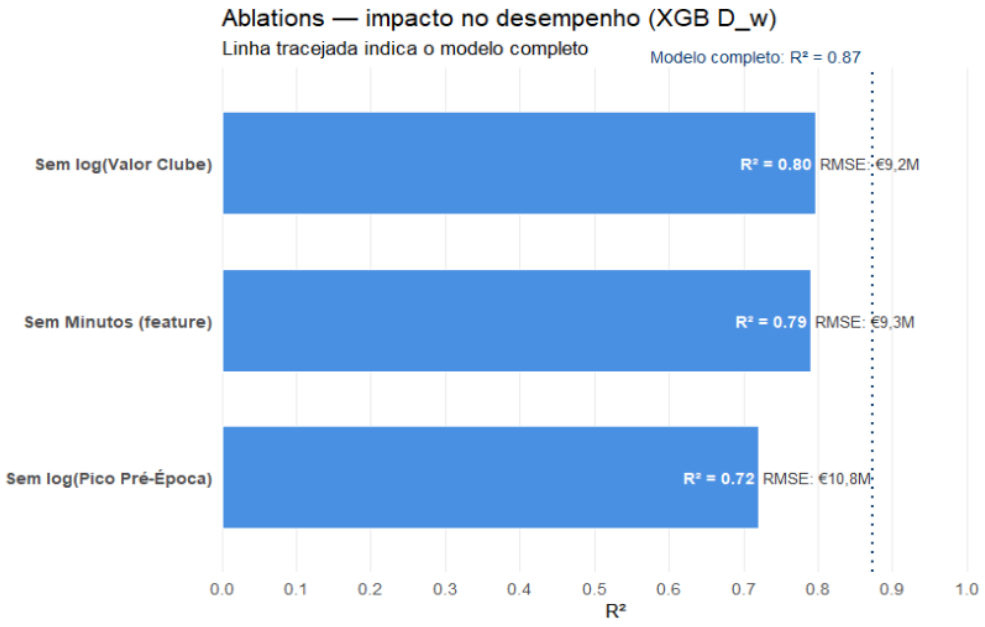


Figura C.4. Ablations: impacto em ².

C.3 Calibração e erro por segmentos

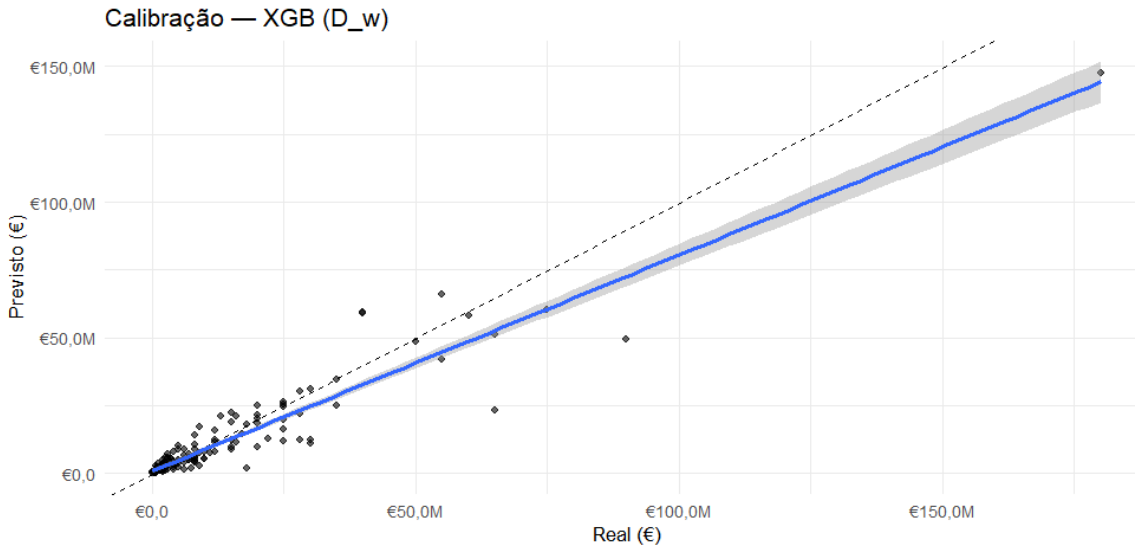


Figura C.5. Calibração (Real vs Previsto).

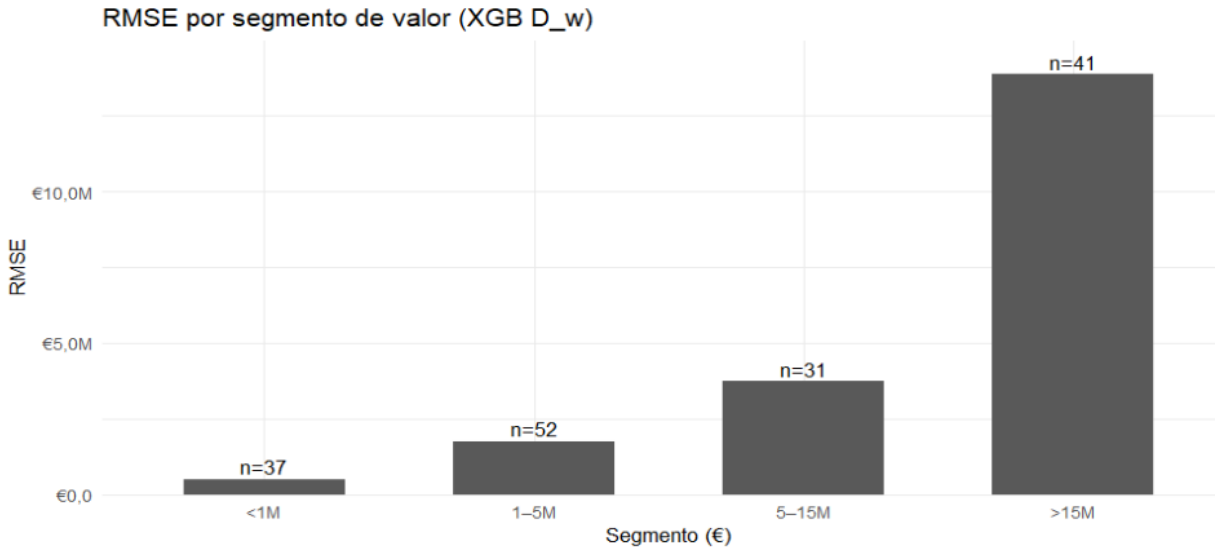


Figura C.6. RMSE por segmento de valor.

C.4 Interpretabilidade (SHAP)

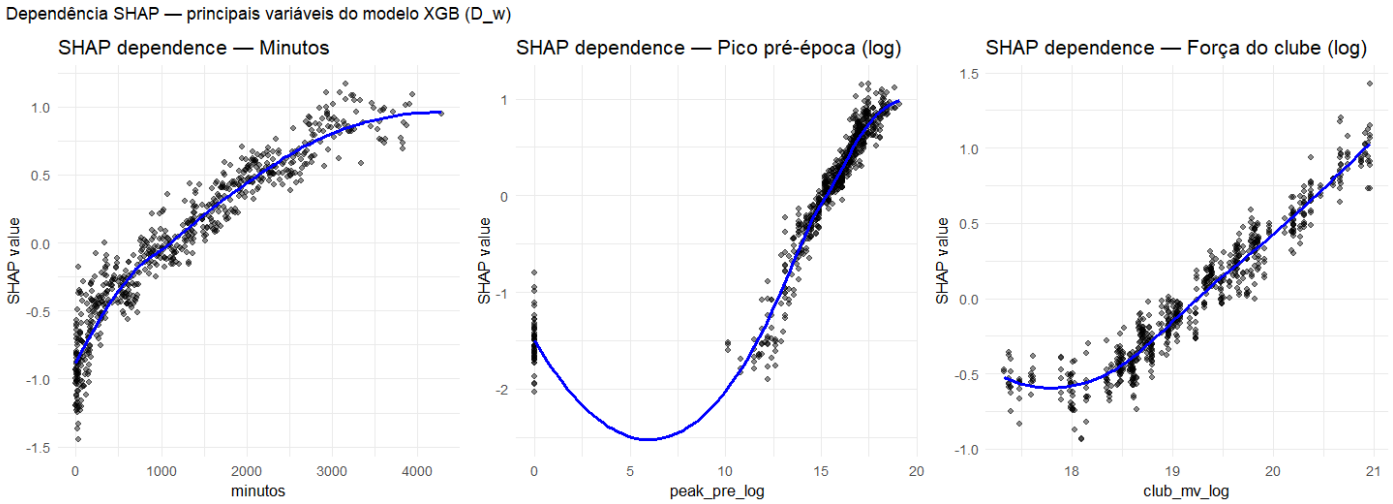


Figura C.7. Dependência SHAP das principais variáveis.

Anexo D – Configuração dos modelos e validação

Tabela D.1. Variáveis usadas no modelo final (XGB D_w).

Variável no modelo	Descrição	Transformação
idade	Idade do jogador na data do último valuation da época	—
idade2	Termo quadrático para não linearidade	Idade²
altura_cm	Altura (cm)	—
minutos	Minutos jogados em 2023/24	—
golos90	Golos por 90' em 2023/24	Winsor 99º pctl
ast90	Assistências por 90' em 2023/24	Winsor 99º pctl
club_mv_log	Força do clube (valor de plantel)	log(1+valor_clube)
peak_pre_log	Pico de valuation do jogador pré-época	log(1+pico_pré)
Target	Valor de mercado	ylog=log(1+valor)

Tabela D.2. Hiperparâmetros do XGBoost "D_w" (tuned + pesos)

Parâmetro	Valor	Nota
objective	reg:squarederror	Regressão em y_log
eval_metric	rmse	Métrica de treino
eta	0.03	learning rate
max_depth	4	profundidade da árvore
min_child_weight	10	robustez a ruído
gamma	1	penalização de splits
subsample	0.70	amostragem de linhas
colsample_bytree	0.90	amostragem de colunas
reg_lambda	5	L2
reg_alpha	1	L1
tree_method	hist	mais rápido
nrounds	620	iterações finais
sample_weight	√(minutos), min=1	mais peso a quem joga mais

Tabela D.3. Esquema de validação e robustez.

Item	Definição
Validação para A e C	5-fold CV, early_stopping_rounds = 40, nrounds máx 1500 (CV com seed=300)
Mini-grid (A/C)	eta ∈ {0.05, 0.03}; max_depth ∈ {4, 6}; min_child_weight ∈ {1, 5}
Pesos (C e D_w)	w_tr_sqrt = pmax(1, √minutos)
Robustez multi-seed	Seeds: 300, 101, 202, 303, 404 (5 repetições por variante)
Conj. treino/teste	Idêntico ao do Exp_2 (IDs fornecidos em test_ids.csv)

Tabela D.4 — Breusch–Pagan no LM (nível vs. log)

Modelo OLS	Estatística BP	gl	*p-value*	Conclusão
market_value_eur ~ minutos + golos + idade	409.79	3	< 2.2e-16	Rejeita H_0 → usar SE robustos (HC1)
log1p(market_value_eur) ~ minutos + golos + idade	149.74	3	< 2.2e-16	Rejeita H_0 → usar SE robustos (HC1)

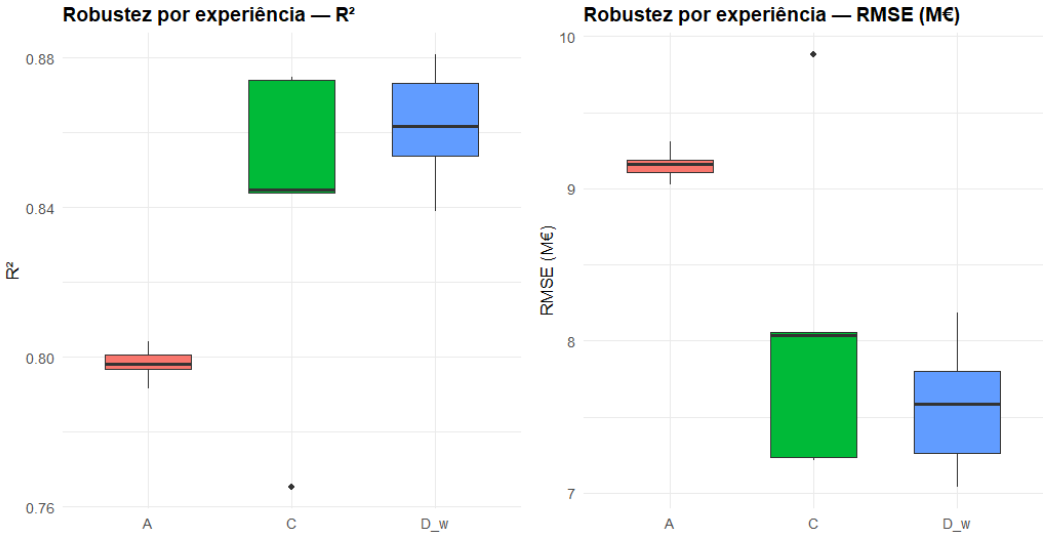


Figura D.1. Robustez por experiência.

Anexo E – Figuras complementares

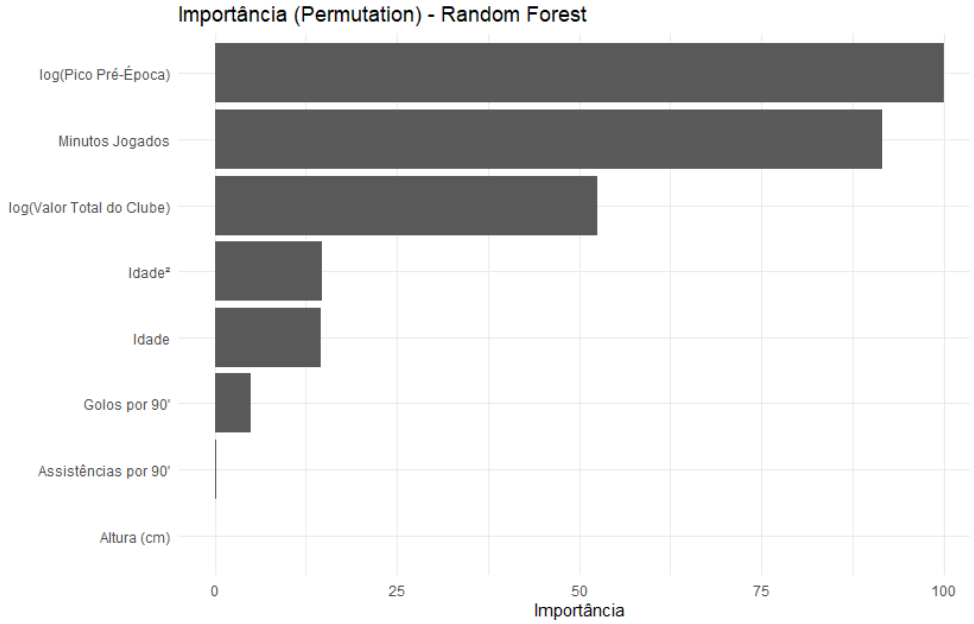


Figura E.1. Importância de variáveis.

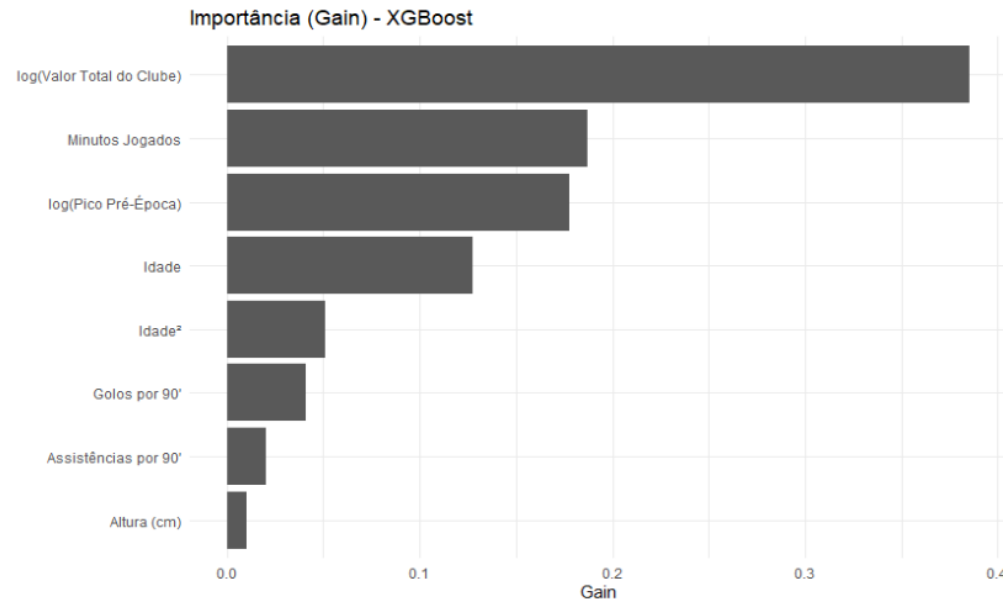


Figura E.2. Importância de variáveis — XGBoost.

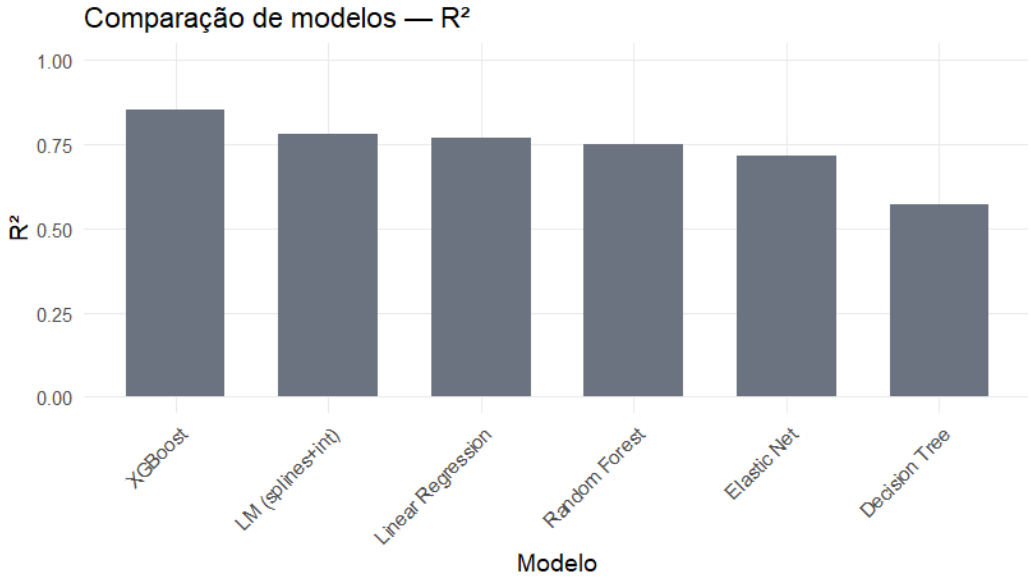


Figura E.3. Comparação global de modelos — R^2.

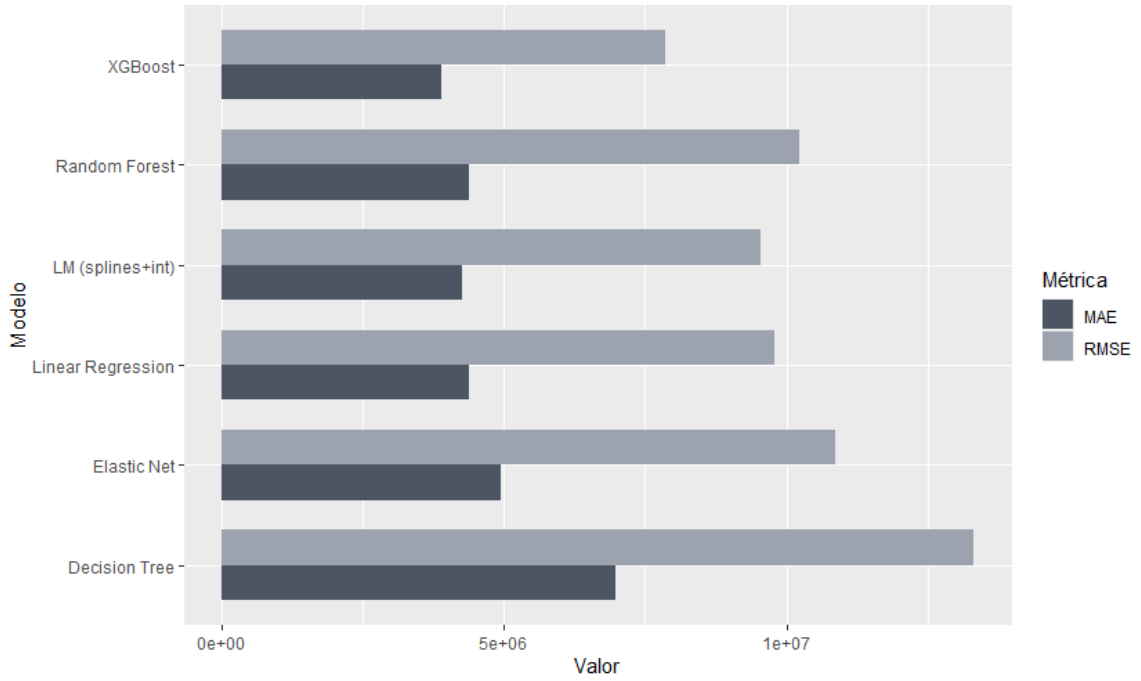


Figura E.4. Comparação de erros — RMSE & MAE.

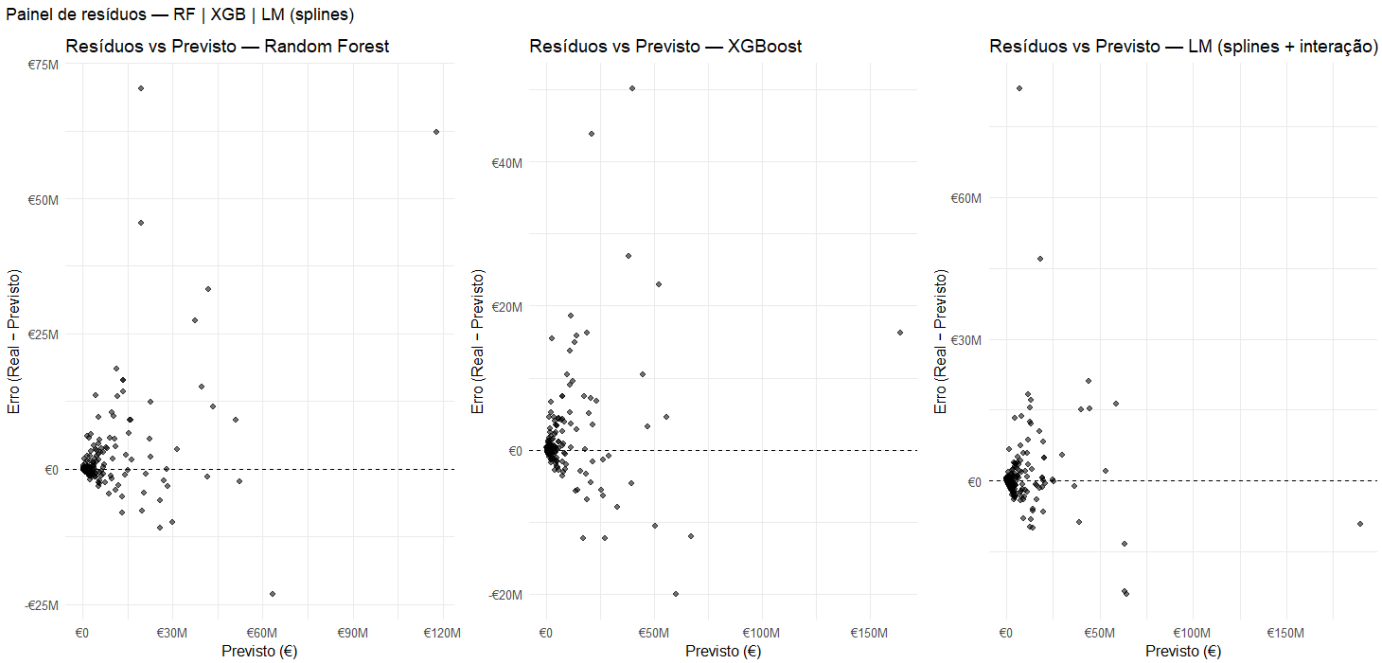


Figura E.5 Painel de resíduos (RF | XGB | LM-splines).

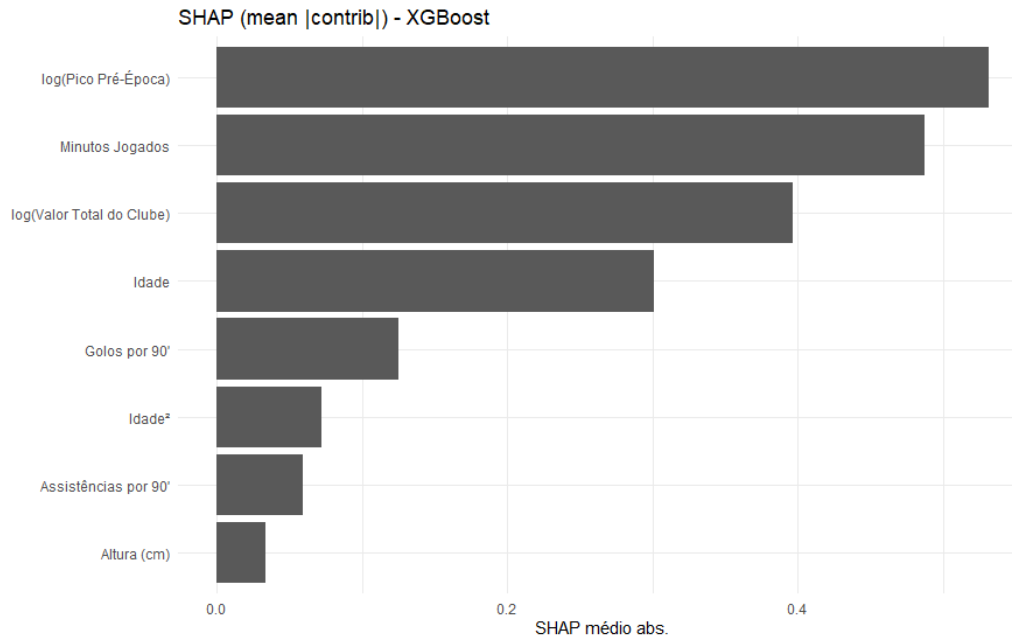


Figura E.6. SHAP — contributo médio absoluto (Top variáveis).

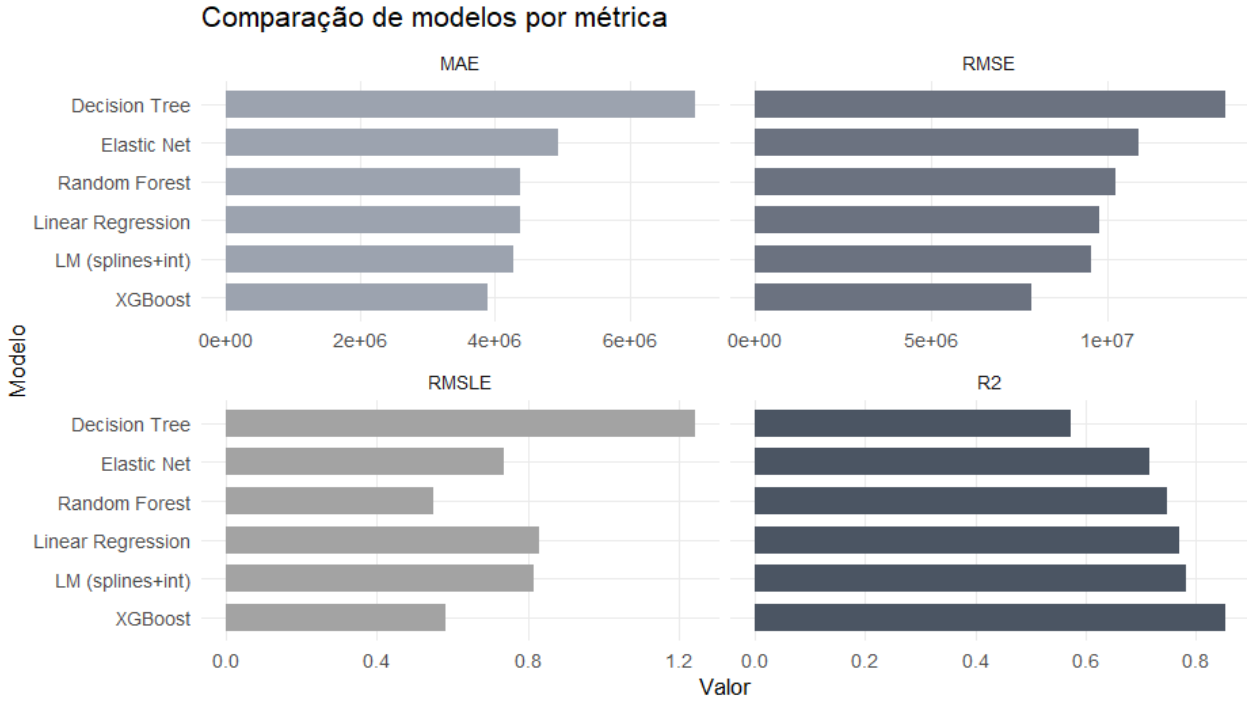


Figura E.7. Comparação de modelos por métrica (MAE, RMSE, RMSLE e R^2).