Estimativa de valor de jogadores de futebol no mercado de transferências brasileiro

Eduardo de Souza Cecconi[1]*; Walter Mesquita Filho[2]

[1] Analista de Desempenho. Rua Doutor Florêncio Ygartua, 73/509 – Bairro Moinhos de Vento; 90430-010 Porto Alegre, Rio Grande do Sul, Brasil
[2] Pecege. Professor Orientador. Parque Tecnológico R. Cezira Giovanoni Moretti, 600 - Santa Rosa; 13418-445 Piracicaba, São Paulo, Brasil
* autor correspondente: eduardocecconi@gmail.com

Estimativa de valor de jogadores de futebol no mercado de transferências brasileiro

Resumo

O mercado de transferências de jogadores de futebol movimenta milhões de euros anualmente. Entretanto, as negociações ainda são realizadas com base em avaliações subjetivas. O uso de modelos de machine learning supervisionados e não-supervisionados pode ser uma ferramenta poderosa de auxílio à tomada de decisão. Este estudo teve como objetivo criar um modelo preditivo capaz de estimar o valor de mercado de um jogador baseado em sua performance nas séries A ou B do Campeonato Brasileiro, além de avaliar as probabilidades de ele ser negociado por um valor considerado relevante (a partir de € 30 mil) e de se transferir para cada um entre seis conglomerados compradores definidos. Das 820 observações filtradas, em 663 (80.85%) não foi registrado nenhum recurso investido, o que levou à escolha da regressão Binomial Negativa para dados de contagem inflacionados de zeros. Os dados foram extraídos do site “Transfermarkt”, especializado no mercado de futebol, e do provedor “Wyscout”. O modelo identificou que jogadores jovens, de características ofensivas e com baixa amostragem no futebol profissional tendem a ser vendidos por valores acima do estimado, principalmente para grandes clubes europeus, em especial espanhóis, além de obter alta acurácia quanto à probabilidade de venda por valor relevante. Estes diagnósticos podem ser fundamentais para orientar os dirigentes dos clubes brasileiros não apenas na seleção dos jogadores com maior potencial de venda, como também dos mercados compradores para cada perfil, auxiliando na projeção de receitas.
Palavras-chave: clube; time; campeonato; janela.

Introdução

Entre as principais fontes de receita dos clubes de futebol do Brasil, a negociação de jogadores apresenta a maior queda recente - representou apenas 18% das receitas da Série A do Campeonato Brasileiro em 2021, interrompendo uma sequência de alta que foi de 20% a 27% nos três anos anteriores. No total, o valor arrecadado com transferências na Série A passou de aproximadamente € 234 milhões em 2020 para cerca de € 198 milhões em 2021 (XP Investimentos Corretora de Câmbio, Títulos e Valores Mobiliários [XP], 2022).

Esta queda pode ser atribuída em grande parte às consequências da pandemia de COVID-19 e às restrições necessárias ao seu combate, mas também à influência de contextos como o mecanismo de “fair-play” adotado pela “Union of European Football Associations” [UEFA] com a intenção de reduzir a distância entre o investimento dos clubes mais ricos e dos demais, e assim assegurar maior competitividade em campo.

Pela sua característica irregular, portanto, depender da receita com negociações é um risco. Driblar esta imprevisibilidade requer a avaliação criteriosa de cada jogador, assim como o diagnóstico dos perfis mais cobiçados e dos principais compradores.

A abordagem de mercado, muito utilizada para avaliação de empresas, propriedades ou ativos financeiros, oferece uma indicação de valor (“valuation”, em inglês) a partir da

comparação do objeto da análise com outros idênticos ou semelhantes que tenham informação de preço disponível (“International Valuation Standards Council” [IVSC], 2022).

Este conceito, aplicado ao futebol, serve de referência para a estimativa de valor de um jogador. Cabe ressaltar, entretanto, que os jogadores não são mercadorias, nem objetos. Qualquer referência teórica baseada em ações ou propriedades precisa ser interpretada como uma extrapolação para a implementação dos conceitos de “valuation” no cenário do futebol profissional, devido às características semelhantes entre os modelos de negócio.

Damodaran (2012) destaca que a estimativa de valor desempenha papel central na análise de aquisições. Através dela, o comprador deve definir qual o valor justo para fazer a oferta, assim como o vendedor precisa determinar o valor razoável sob a sua perspectiva, para decidir se aceita ou rejeita a proposta.

Em qualquer mercado, compradores e vendedores terão seus pontos de vista a respeito dos valores envolvidos, configurando uma reunião de interesses de indivíduos ou empresas, que estabelecem barganhas (Shapiro et al., 2013). Com a estimativa de valor baseada em dados de desempenho e perfil de jogadores já negociados, os clubes brasileiros podem identificar aqueles com maior potencial de valorização, assim como para quais mercados ofertá-los, assumindo papel proativo e aperfeiçoando a projeção de receitas.

O principal objetivo deste trabalho, baseado no contexto apresentado e com foco no suporte à tomada de decisão, é estimar o valor de mercado de jogadores de futebol a partir dos dados de desempenho/perfil coletados em partidas disputadas pelos Campeonatos Brasileiros das Séries A e B entre 2018 e 2022, reduzindo assim a imponderabilidade inerente à relação comprador-vendedor para chegar a um valor justo (Póvoa, 2008). Além disso, os objetivos secundários são estimar a probabilidade de que o jogador em análise seja negociado por um valor relevante (neste caso, a partir de € 30 mil, valor mínimo da base de dados), e as probabilidades de que o destino seja cada um entre os seis mercados identificados em análise de agrupamentos de 56 países compradores listados.

Material e Métodos

Data wrangling

Embora utilizando apenas as capacidades de processamento e armazenamento locais, sem recorrer a serviços de computação em nuvem, a consolidação de um único banco de dados para abastecer os modelos regressivos desejados seguiu os princípios do “Extract, Transform, Load” (ETL) – ou Extrair, Transformar e Carregar, em português. É nesta etapa

que os dados são extraídos de diferentes origens e transformados para adequação ao banco de dados projetado (Vida et al., 2013).

Cada sub etapa do processo de ETL local foi construída em diferentes “scripts” de códigos escritos na linguagem R (R Core Team, 2022), estabelecendo um fluxo desde a extração, passando pela transformação até chegar ao carregamento do “Dataset”. Além de armazenados em um disco rígido externo, todos os arquivos (um projeto na interface “R Studio”, em extensão “.Rproj”; 11 “scripts”, em extensão “.R”; 38 planilhas originais e 11 planilhas resultantes, em extensão “.xlsx”; e três modelos finais, em extensão “.rds”) estão disponíveis no repositório https://github.com/eduardocecconi/tcc.

A obtenção dos dados necessários à criação do “Dataset” se deu a partir de duas fontes distintas, o que também requisitou diferentes técnicas. Com o pacote “worldfootballR” (Zivkovic, 2023) foi possível fazer a raspagem dos dados de transferências (incluindo os valores das transações) direto do site “Transfermarkt” (https://www.transfermarkt.com/), sem a necessidade de construir um mecanismo de “web crawler” manualmente.

Foram utilizadas três funções deste pacote:

1) “tm_league_team_urls”, que atribui a um objeto os endereços das páginas dos clubes de uma determinada competição no “Transfermarkt”, em uma determinada temporada. Os dados da Série A do Campeonato Brasileiro foram obtidos através dos argumentos “country_name” e “start_year”, enquanto os dados da Série B através dos argumentos “league_url” e “start_year”;

2) “tm_team_transfers”, que atribui a um objeto a lista de transferências realizadas pelos clubes de uma determinada competição, em uma determinada temporada, também através de dois argumentos, “team_url” (preenchido com o nome do objeto anterior); e “transfer_window”, preenchido com “all” para obter dados tanto da janela de verão como da janela de inverno (referências sazonais sob a perspectiva do mercado europeu); e

3) “tm_player_bio”, que atribui a um objeto uma lista com dados de desempenho básicos e de perfil dos jogadores desejados, através do argumento “player_urls”, preenchido com o nome do objeto onde foram armazenadas as transferências.

Já os dados de desempenho utilizados na modelagem foram obtidos com o download de planilhas disponibilizadas pelo provedor “Wyscout” (https://wyscout.com/), cuja navegação permite filtrar a competição e as variáveis desejadas, e baixar os dados resultantes em um arquivo de extensão “.xlsx”, procedimento que foi realizado separadamente para as Séries A e B, e segmentado por ano.

A integração dos dados obtidos das duas fontes em um “Dataset” demandou rigorosa manipulação durante a técnica conhecida como “data wrangling”. Isso porque a grafia dos nomes dos atletas proveniente do “Transfermarkt” é diferente da grafia utilizada pelo

"Wyscout", assim como em muitos casos o clube de origem diverge, sem padrões comuns que permitissem automatizar a fusão. Os dados de cada temporada foram organizados e estruturados, incluindo a conferência de todos os nomes, com os necessários ajustes caso-a-caso. Embora exaustivo, o "data wrangling" permitiu a consolidação do banco de dados em formato "tidy", o que favorece a identificação das ferramentas apropriadas para a análise, devido à uniformidade existente (Wickham e Gerolemund, 2016). Diversas funções foram utilizadas no processo de limpeza, organização e estruturação do "Dataset", principalmente originárias da linguagem-base do "software" R e do pacote "tidyverse" (Wickham et al., 2019).

Variáveis dummies e análise de clusters

Para identificar características comuns aos clubes de origem, foi realizada uma análise de agrupamentos ("clusters"), com base em tabelas de frequência e sumarizações das idades dos jogadores negociados, dos valores totais adquiridos e do volume de vendas realizadas. Estes dados foram submetidos ao algoritmo não-supervisionado do método não-hierárquico "k-means", uma função do pacote "stats", que faz parte do "software" R (R Core Team, 2022), para agrupar os 49 clubes de origem em seis "clusters", quantidade determinada a partir da visualização de diferentes combinações com a função "fviz_cluster" do pacote "factoextra" (Kassambara e Mundt, 2020), o que originou a variável categórica Cluster Vendedor.

Da mesma forma, a identificação dos potenciais mercados de destino foi realizada com uma análise de agrupamentos baseada em tabelas de frequência e sumarizações das idades dos jogadores negociados, dos valores totais investidos e do volume de transferências realizadas por 56 países compradores, também através da função "k-means" com definição de seis "clusters" após análise visual de diferentes possibilidades (função "fviz_cluster"), resultando na variável categórica Cluster Comprador.

Na sequência, foram criadas variáveis binárias ("dummies") a partir de quatro variáveis categóricas originárias, seguindo o critério n – 1 (total de "dummies" igual à subtração de um do total de categorias da variável original):

1) Cluster Vendedor (6 categorias): do "Cluster" 1 ao "Cluster" 6;

2) Posição (10 categorias): Centroavante, Atacante, Extremo-Esquerdo, Extremo-Direito, Meia-Atacante, Meia, Volante, Lateral-Direito, Lateral-Esquerdo e Zagueiro. Os goleiros (65 observações, das quais apenas 8 com valor de venda) foram excluídos, assim como seis variáveis exclusivas desta função, porque formam um grupo que requer uma análise específica - o que não foi possível devido à baixa amostragem;

3) Liga (4 categorias): além das Séries A e B, objetos de análise, houve jogadores registrados em clubes da Série C (transferidos com esta origem ou destino em meio a uma

determinada temporada) ou Sem Série (transferidos de origens ou para destinos alheios às três divisões listadas anteriormente; e

4) Janela (duas categorias): Inverno e Verão, ambas referindo-se aos períodos de negociação assim denominados sob a perspectiva do mercado europeu de futebol.

Este procedimento se fez necessário para evitar a ponderação arbitrária (Fávero e Belfiore, 2017), e foi realizado com a função “dummy_columns” do pacote “fastDummies” (Kaplan, 2020), tendo como um dos argumentos a remoção das variáveis originais. Foram consideradas categorias de referência, e excluídas manualmente, aquelas com os maiores valores totais de vendas somados, critério estabelecido para enfatizar este aspecto como o preponderante na modelagem: na variável Posição, a referência foi o Extremo-Esquerdo; na Liga, a Série A; na Janela, o Verão; e na Cluster Vendedor, o “cluster” 3.

Variáveis de desempenho

Ainda durante os processos de manipulação para limpeza e organização do “Dataset”, foram excluídas variáveis que apresentaram altas correlações na análise da matriz de correlações de Pearson, gerada com a função “rcorr” do pacote “Hmisc” (Harrel Jr, 2021), o que pode ser atribuído principalmente à utilização de protocolos de coleta que registram a mesma ação em mais de uma variável. Por exemplo, um passe pode ser catalogado simultaneamente como Passe, Passe para a Frente e Passe Progressivo. Neste aspecto, a familiaridade do pesquisador com as características dos dados é fundamental para interpretar as correlações e excluir eventuais variáveis que contenham sobreposição de informação.

Garganta (2001) destaca a importância da criteriosa seleção das variáveis, quando se analisa o desempenho de jogadores, para a construção de uma matriz de referência. A utilização de sistemas informatizados na coleta e no tratamento de dados procura responder a questões como quem realiza a ação, de que forma, em que local do campo e com qual resultado (Teoldo et al., 2010). A abordagem de mercado também ressalta o papel relevante da escolha das variáveis, chamadas de unidades de comparação. Com informações de transações recentes é possível confrontá-las com outras exteriores ao “Dataset”, e assim estimar o valor de mercado, que será o mais provável e razoável preço praticado na data da avaliação (IVSC, 2022).

Seguindo estes critérios, foram selecionadas 29 variáveis preditoras específicas de desempenho ou perfil dos jogadores. A Tabela 1 apresenta cada uma delas, com a grafia utilizada no “Dataset” (sem acentuação, espaços ou caracteres especiais), na ordem em que aparecem, e com uma breve descrição. Foram consideradas apenas ações certas:

Tabela 1. Variáveis de desempenho e de perfil selecionadas

Variável	Definição
Idade	Idade do jogador na temporada de sua transferência
Jogos	Total de jogos disputados na temporada
Minutos_media	Média de minutos em campo por jogo disputado
Gols	Total de gols marcados
xG_media	Expectativa de gol (xG) criada por jogo
Conversao	Percentual de conversão da expectativa criada em gol
Chutes_media	Finalizações, de qualquer tipo, por jogo
Toques_Area_media	Toques na bola dentro da área adversária por jogo
Conducoes_media	Conduções (avanços com a bola dominada) por jogo
Dribles_media	Dribles (duelos com uso de habilidade) por jogo
Faltas_Recebidas_media	Faltas recebidas por jogo
Assistencias	Total de assistências (passes para gol) na temporada
Passes_Chave_media	Passes que resultaram em finalização por jogo
Pre_Assistencia_media	Passes que resultaram em assistência por jogo
Preparacao	Passes que resultaram em pré-assistência por jogo
Passes_media	Passes (de qualquer tipo) realizados por jogo
Distancia_Passes	Distância média (em metros) de todos os tipos de passes
Passes_Profundidade_media	Passes em profundidade por jogo
Passes_Progressivos_media	Passes que resultaram em progressão por jogo
Passes_Terco_Final_media	Passes direcionados ao terço ofensivo do campo por jogo
Passes_Area_media	Passes direcionados à área adversária por jogo
Cruzamentos_media	Lançamentos direcionados à área adversária por jogo
Duelos_Aereos_media	Disputas individuais com bola aérea por jogo
Carrinhos_media	Ações de contenção com deslize em campo por jogo
Bloqueios_media	Ações de contenção para bloquear finalizações por jogo
Interceptacoes_media	Rebatidas que não resultaram em troca de posse por jogo
Faltas_Cometidas_media	Faltas cometidas por jogo
Amarelos	Total de cartões amarelos recebidos na temporada
Vermelhos	Total de cartões vermelhos recebidos na temporada

Fonte: “Wyscout” (2023)

Desde a criação das “dummies”, passando pelo estudo das correlações e pela exclusão de diversas observações e variáveis que não atendiam aos requisitos estabelecidos para a pesquisa (como o filtro de no mínimo dois jogos com dados de desempenho registrados), resultaram 820 observações e 47 variáveis preditoras, as 29 de desempenho (numéricas) e as 18 “dummies”, caracterizando o “Dataset” definitivo para servir à modelagem. Com isso, o procedimento de extração, transformação e carregamento foi finalizado, tendo alcançado o objetivo de municiar os modelos regressivos com a base de dados estruturados.

Modelos supervisionados

Definidos os agrupamentos das variáveis categóricas Time e País Comprador, a modelagem preditiva teve início. O primeiro passo foi submeter o “Dataset” ao Teste de Cameron-Trivedi no “software” R, através da função “overdisp”, do pacote “overdisp” (Cameron e Trivedi, 2020). Este procedimento se fez necessário devido à distribuição da variável dependente (Valor), sem padrão definido, com grande acúmulo de zeros e alongamento da cauda provocado por “outliers”, o que foi observado a partir da visualização do seu histograma, através da função “hist”, do pacote “graphics” (R Core Team, 2022).

O resultado do teste, contextualizado às características da variável dependente (quantitativa, com valores discretos e não-negativos, e com unidade de exposição temporal - por ano), levou à escolha do modelo Binomial Negativo para dados de contagem (Gardner et al., 1995). Para executá-lo, foi utilizada a função “glm.nb” do pacote “MASS” (Venables e Ripley, 2002) no “software” R. Também foi aplicada a função “vif” do pacote “car” (Fox e Weisberg, 2019), para investigar a existência de multicolinearidade entre preditoras através do cálculo do Variance Inflated Factor (VIF), baseado na tolerância encontrada pela comparação do modelo com um modelo auxiliar (Ron, 1981). Quanto maior o valor de VIF, maior a tendência à multicolinearidade.

O valor do parâmetro theta encontrado no modelo Binomial Negativo foi utilizado como argumento para a construção de um novo modelo, desta vez a partir de outras duas funções: “glm”, do pacote “stats” (R Core Team, 2022); e “negative.binomial”, do pacote “MASS” (Venables e Ripley, 2002), que exige a definição do valor de theta e é declarada dentro do argumento “family” da função “glm”. Theta é um parâmetro de dispersão que captura o grau de afastamento da variância em relação à média da variável dependente (Hilbe, 2014).

O segundo modelo Binomial Negativo foi submetido ao procedimento “stepwise”, com a função “step” do pacote “stats” (R Core Team, 2022), tendo como argumento a definição do nível de significância em 5%. Seu resultado seria, na sequência, utilizado apenas como referência para confirmar a hipótese de inflação de zeros na variável dependente. O “stepwise”, aplicado nos Modelos Lineares Generalizados (ou GLM’s, “Generalized Linear

Models", em inglês), busca obter o melhor modelo possível a partir da retirada de uma variável a cada passo ("step", em inglês), comparando as medidas resultantes do indicador "Akaike Information Criterion" (AIC), e interrompendo o procedimento assim que não houver mais condições de melhora (Hastie e Chambers, 1991).

Três indicadores de desempenho foram salvos em uma tabela, para futura comparação com os resultados do modelo inflacionado de zeros: além do AIC, com a função "AIC", também foram obtidos o Logaritmo de Verossimilhança (conhecido por "log-likelihood", em inglês), extraído com a função "logLik", e o "Bayesian Information Criterion" (BIC), com a função "BIC", todas do pacote "stats" (R Core Team, 2022).

AIC é um critério que avalia quão ruim é o modelo, cujos parâmetros são estimados pelo método de máxima verossimilhança. BIC é um critério de avaliação de modelos definido em termos das suas posteriores probabilidades. E "log-likelihood" é uma função que busca maximizar o acerto do modelo (Konishi, 2008). Para AIC e BIC, quanto menores os valores, melhor é o ajuste do modelo aos dados; para o "log-likelihood", quanto maior o valor, melhor.

A quantidade excessiva de zeros na variável dependente pode criar vieses nos parâmetros estimados por modelos tradicionais de regressão para dados de contagem, como o Binomial Negativo utilizado inicialmente neste estudo, que não conseguem capturá-la. Os modelos inflacionados de zeros foram desenvolvidos para lidar com este fenômeno, através da combinação de um modelo para dados de contagem e outro para dados binários, com o objetivo de investigar não apenas que condições reduzem a média de eventos como também os motivos pelos quais o fenômeno ocorre (Lambert, 1992).

A fórmula resultante apresenta dois lados: uma regressão logística binária, que estuda os zeros estruturais (responsáveis pela inflação); e uma regressão para dados de contagem, que investiga os zeros amostrais (média de ocorrências do fenômeno). Os zeros estruturais referem-se aos valores zero das observações cuja resposta de contagem sempre será zero, em razão dos seus padrões de comportamento, enquanto os zeros amostrais (ou aleatórios) se originam de observações cujas contagens podem ou não ser maiores que zero, devido à variabilidade da amostragem (Tang et al., 2017).

O modelo Binomial Negativo inflacionado de zeros foi construído com a função "zeroinfl", do pacote "pscl" (Zeileis et al., 2008), tendo como argumentos a distribuição ("dist = negbin") e a função de ligação ("link = logit"), além da base de dados e da fórmula, que incluiu todas as 47 variáveis preditoras em seus dois lados (o de contagem e o de inflação de zeros).

Os parâmetros estimados servem à composição da expressão própria ao modelo de regressão Binomial Negativo inflacionado de zeros, como descrito por Fávero e Belfiore (2017) a partir do modelo proposto por Lambert (1992), para o cálculo do valor médio esperado, conforme segue:

$$\mu_{inflate} = \underbrace{\left\{1 - \frac{1}{1 + e^{-[\gamma + \delta_{1i}.W_{1i} + \delta_{2i}.W_{2i} + \cdots + \delta_{qi}.W_{qi}]}}\right\}}_{Modelo\ Log\acute{i}stico\ Bin\acute{a}rio\ (infla\c{c}\tilde{a}o)} \cdot \underbrace{\left\{e^{[\alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + \cdots + \beta_k.X_{ki}]}\right\}}_{Modelo\ Binomial\ Negativo\ (contagem)}$$

Na expressão, o termo μ é a taxa média estimada de incidência do evento durante dada exposição, no caso, o valor médio predito de um jogador em uma temporada do Campeonato Brasileiro. No lado Logístico Binário, γ é o intercepto, enquanto δ é o coeficiente estimado para cada uma das variáveis preditoras (W); da mesma forma, no lado Binomial Negativo o termo α é o intercepto, β é o coeficiente e X a variável preditora respectiva.

A realização do procedimento “stepwise” no modelo inflacionado de zeros foi possível com a função “be.zeroinfl” do pacote “mpath” (Wang, 2022), utilizando como argumentos o modelo inicial, o “Dataset”, a definição da distribuição (“dist=negbin”) e o nível de significância de 5% (“alpha=0.05”). Foram realizados ainda dois testes comparando os modelos binomiais negativos (com e sem inflação de zeros). Primeiro, o teste de razão de verossimilhança (“likelihood ratio test”, em inglês) com a função “lrtest” do pacote “lmtest” (Zeileis e Hothorn, 2002), a partir dos valores de “log-likelihood” e da significância estatística da diferença entre os modelos. A confirmação da escolha pelo modelo inflacionado de zeros se deu com o teste de Vuong, aplicado para definir se a quantidade de zeros é ou não excessiva (Vuong, 1989), através do valor da estatística Z e da significância estatística do resultado. Para isso, foi utilizada a função “vuongtest” do pacote “nonnest2” (Merkle e You, 2020).

Analisando-se apenas o lado logístico binário da equação do modelo inflacionado de zeros, é possível estimar a probabilidade de que o jogador analisado seja negociado por um valor relevante (a partir de € 30 mil). O cálculo pode ser feito manualmente a partir da fórmula exposta anteriormente (que investiga a inflação de zeros), ou através da construção específica de um modelo logístico binário para posterior submissão à função de predição no “software” R, o que foi realizado com a função “glm” do pacote “stats” (R Core Team, 2022), indicando a distribuição binomial no argumento “family”, e tendo a variável dependente Valor sido transformada em binária (indicando 1 para presença de valor, e 0 para ausência de valor).

A avaliação do modelo Logístico Binário foi possível com a elaboração da matriz de confusão, com a função “confusionMatrix” do pacote “caret” (Kuhn, 2021), estabelecendo-se um “cut-off” de 0.5 como argumento, padrão utilizado para direcionar os valores preditos para a forma binária (as probabilidades maiores ou iguais a 50% foram consideradas vendas com valor, enquanto as menores de 50%, vendas sem valor). A matriz de confusão é uma tabela de contingência (Pearson, 1904), ou seja, uma tabela em formato de matriz que apresenta a frequência da distribuição das variáveis, da qual podem ser extraídas medidas de avaliação

de um modelo de “machine learning”, como sensitividade, especificidade, acurácia e “Receiver Operating Characteristics” (ROC), entre outras (Powers, 2008).

Para uma variável binária, a matriz é formada pelo cruzamento de duas colunas com os valores reais (ocorrência do evento e ausência do evento) e de duas linhas com os valores preditos (ocorrência do evento e ausência do evento), o que resulta em quatro células, com as seguintes contagens: verdadeiros positivos (A) e falsos negativos (C), na primeira coluna; falsos positivos (B) e verdadeiros negativos (D), na segunda coluna, conforme a Tabela 2:

Tabela 2. Notação tradicional em matriz de confusão binária

Matriz	Positivo Real	Negativo Real
Positivo Predito	Verdadeiro Positivo (A)	Falso Positivo (B)
Negativo Predito	Falso Negativo (C)	Verdadeiro Negativo (D)

Fonte: Powers (2008)

A soma da primeira coluna da Tabela 2 apresenta os valores positivos reais (A + C), enquanto a soma da segunda coluna apresenta os valores negativos reais (B + D); da mesma forma, a soma da primeira linha apresenta os valores positivos previstos (A + B), e a soma da segunda linha apresenta os valores negativos previstos (C + D), enquanto a soma das quatro células (A + B + C + D) apresenta o total de observações (N) da amostra (Powers, 2008). Baseadas nesta configuração de matriz binária, as medidas de avaliação de modelos supervisionados de machine learning podem ser extraídas.

Sensitividade é a proporção de positivos reais corretamente preditos, definida pela equação:

$$Sensitividade = \frac{A}{(A + C)}$$

Especificidade é proporção de negativos reais corretamente preditos, como segue:

$$Especificidade = \frac{D}{(B + D)}$$

E a Acurácia mede a eficiência global do modelo, ou seja, a proporção de ambos os acertos em razão de toda a amostra:

$$Acurácia = \frac{(A + D)}{N}$$

Já a análise ROC se dá a partir de um gráfico formado pela sensitividade no eixo Y e pela proporção de falsos positivos (1 - especificidade) no eixo X, o que permite calcular a área sob uma curva trapezoide que se forma. No melhor cenário para a análise da área sob a curva ROC, a sensitividade deve ser de 100%, enquanto a proporção de falsos positivos deve ser igual a zero (Powers, 2008). Outro indicador de ajuste do modelo é o teste qui-quadrado, apropriado para aplicação em modelos logísticos (Hosmer e Lemesbow, 1980), cujo resultado e p-valor foram obtidos com a função “summ” do pacote “jtools” (Long, 2022).

No modelo complementar, cujo objetivo é prospectar os mercados com maior probabilidade de comprar o jogador em análise, a variável Valor foi substituída pela variável Cluster Comprador como dependente, mantidas as demais 47 preditoras, para a realização de uma regressão Logística Multinomial, através da função “multinom” do pacote “nnet” (Venables e Ripley, 2002). A distribuição multinomial é considerada uma extensão da distribuição binomial que testa a probabilidade de se estar em uma categoria, comparada às demais. É estabelecida uma categoria de referência e os parâmetros estimados são apresentados separadamente para cada uma das categorias de não-referência (Hilbe, 2009).

O “cluster” 3 permaneceu como referência, o que foi definido aplicando-se a função “relevel” do pacote “stats” (R Core Team, 2022) na variável dependente. Assim como nos modelos Binomial Negativo para dados de contagem e Logístico Binário, também foi utilizada a função “step” do pacote “stats” (R Core Team, 2022) para descartar as variáveis preditoras sem significância estatística ao nível de significância de 5%.

A regressão Logística Multinomial calcula individualmente tanto os coeficientes para cada categoria como o logaritmo natural da chance de ocorrência do evento (logito), através do critério (m – 1), onde m é o total de categorias. Ou seja, tendo seis níveis na variável dependente, foram calculados cinco logitos (Z_{i_m}), seguindo esta expressão genérica:

$$Z_{i_m} = \alpha + \beta_{1m.}X_{1i} + \beta_{2m.}X_{2i} + \cdots + \beta_{km.}X_{ki}$$

Z é o logito da categoria identificada pelo m; α é o intercepto da categoria; e β é o coeficiente estimado para cada variável X. Sendo assim, as probabilidades condicionais para cada um dos “clusters” 1, 2, 3 (referência), 4, 5 e 6, para dada observação i, foram calculadas através das funções logísticas que seguem, conforme exposto por Hilbe (2009):

$$\Pr(y = 3_i) = \frac{1}{1 + e^{[Z_{1i}]} + e^{[Z_{2i}]} + e^{[Z_{4i}]} + e^{[Z_{5i}]} + e^{[Z_{6i}]}}$$

$$\Pr(y = 1_i) = \frac{e^{[Z_{1i}]}}{1 + e^{[Z_{1i}]} + e^{[Z_{2i}]} + e^{[Z_{4i}]} + e^{[Z_{5i}]} + e^{[Z_{6i}]}}$$

$$\Pr(y = 2_i) = \frac{e^{[Z_{2i}]}}{1 + e^{[Z_{1i}]} + e^{[Z_{2i}]} + e^{[Z_{4i}]} + e^{[Z_{5i}]} + e^{[Z_{6i}]}}$$

$$\Pr(y = 4_i) = \frac{e^{[Z_{4i}]}}{1 + e^{[Z_{1i}]} + e^{[Z_{2i}]} + e^{[Z_{4i}]} + e^{[Z_{5i}]} + e^{[Z_{6i}]}}$$

$$\Pr(y = 5_i) = \frac{e^{[Z_{5i}]}}{1 + e^{[Z_{1i}]} + e^{[Z_{2i}]} + e^{[Z_{4i}]} + e^{[Z_{5i}]} + e^{[Z_{6i}]}}$$

$$\Pr(y = 6_i) = \frac{e^{[Z_{6i}]}}{1 + e^{[Z_{1i}]} + e^{[Z_{2i}]} + e^{[Z_{4i}]} + e^{[Z_{5i}]} + e^{[Z_{6i}]}}$$

Assim como no modelo Logístico Binário, foram extraídos os valores dos critérios de avaliação “log-likelihood”, AIC, BIC e teste qui-quadrado, com as mesmas funções utilizadas anteriormente. Também foi criada uma matriz de confusão para comparar os valores reais e preditos das seis categorias, através da sobreposição das funções “as.data.frame.matrix” e “table”, ambas da base do “software” R (R Core Team, 2022). Esta matriz, com seis colunas e seis linhas, permitiu o cálculo da acurácia, através da razão entre a soma das células que apresentam os acertos para cada categoria (diagonal da matriz) e a soma de todas as células.

As predições para valores externos à base de dados podem ser realizadas com a função “predict”, do pacote “stats” (R Core Team, 2022), que utiliza dois argumentos: o modelo final e um “data frame” com os valores obtidos pelo eventual jogador analisado nas variáveis preditoras selecionadas. Outra forma de fazer as predições é reproduzir manualmente as fórmulas e preenchê-las com os valores (das variáveis e dos coeficientes dos parâmetros).

Os três modelos foram salvos como arquivos de extensão “.rds”, com a função “readRDS” da base do “software” R (R Core Team, 2022), o que permite a reprodução dos resultados, assim como a realização de predições.

Resultados e Discussão

Clubes Vendedores

Além das variáveis de desempenho, investigar o impacto da procedência dos jogadores negociados na formação do valor de mercado médio foi considerado um aspecto

importante. Com isso, a variável Time, composta por 49 clubes de origem dos atletas (vendedores) passou pela análise de agrupamentos não-hierárquica, com a intenção de definir “clusters” a partir de três indicadores: volume total de transferências realizadas, valor total arrecadado com estas vendas, e média de idade dos jogadores negociados pelos clubes brasileiros no período entre 2018 e 2022. O resultado da divisão dos “clusters” pode ser observado na Figura 1:

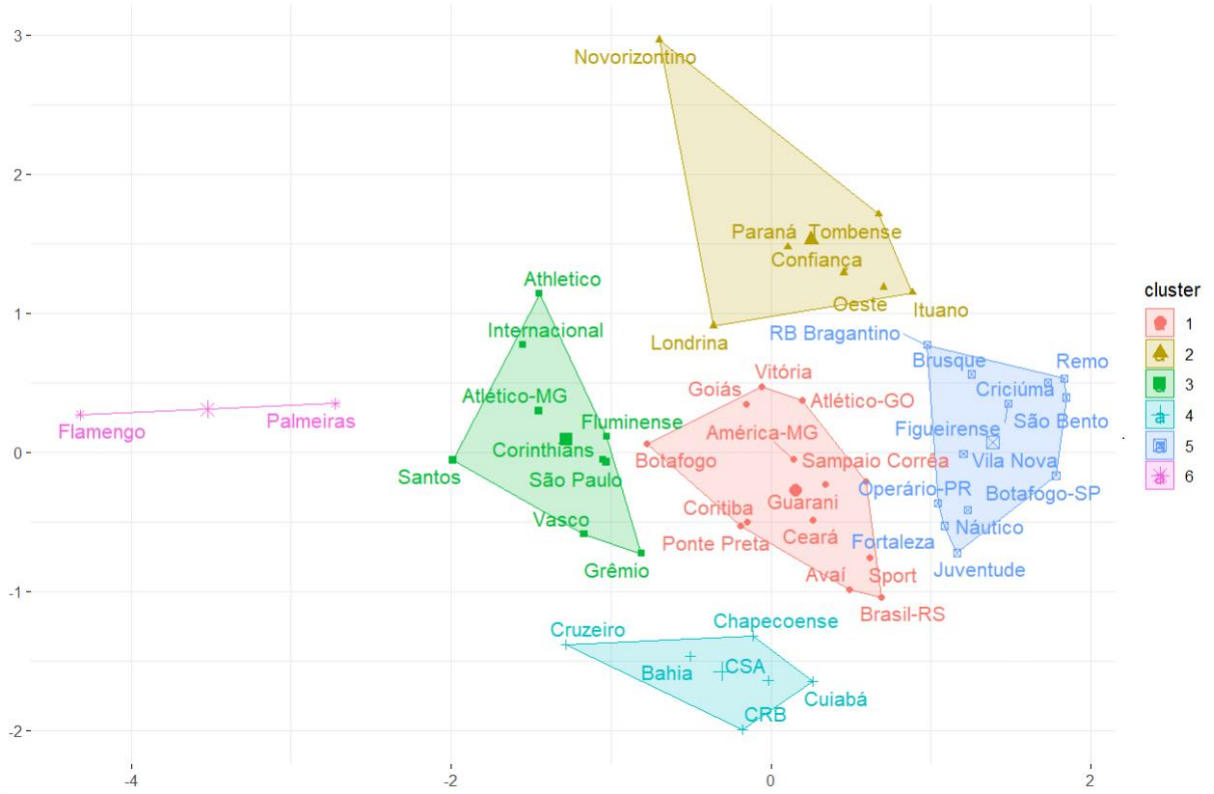

Figura 1. Análise de agrupamentos dos clubes com vendas registradas na base de dados
Fonte: Resultados originais da pesquisa

O “cluster” 6 reuniu os dois clubes de maior capacidade financeira do futebol brasileiro atual: Palmeiras e Flamengo. Segundo o Relatório Convocados (XP, 2022), ambos concentraram praticamente um terço das receitas de 26 clubes que disputaram a Série A do Campeonato Brasileiro na soma das temporadas 2020 e 2021. Enquanto o Palmeiras foi líder em direitos de transmissão neste período, e 3º em receitas com publicidade/marketing, o Flamengo foi 2º em direitos de transmissão e 1º em publicidade, poderio que se acentuou com as vendas de jogadores.

Em valores totais, somados seus nove integrantes (Santos, Grêmio, Athletico, Internacional, Corinthians, São Paulo, Atlético-MG, Fluminense e Vasco), o “cluster” 3 foi selecionado como referência no procedimento de criação das variáveis “dummies”. Este grupo

se caracteriza pela tradição dos clubes, não apenas histórica (conquistas, patrimônio, torcida), mas também como formadores de jogadores com potencial de grandes vendas.

O “cluster” 2 contrasta com os já mencionados 6 e 3. Composto por sete clubes de divisões inferiores (Paraná, Londrina, Confiança, Ituano, Novorizontino, Oeste e Tombense), cinco deles de interior, este grupo somou os menores valores e volumes de transferências (totais e médias), embora tenha negociado jogadores com a 2ª menor média de idade.

Os “clusters” 1 e 4 assemelham-se tanto em valor total como em volume total. Com menos integrantes, nesta comparação o “cluster” 4 fica à frente nas médias, composto por seis clubes, quatro deles classificados para a Série A em 2023 (Cruzeiro, Bahia, Chapecoense e Cuiabá, além de CRB e CSA). O “cluster” 1, com 13 integrantes (Botafogo, Coritiba, Ceará, Vitória, Ponte Preta, Atlético-GO, Sport, América-MG, Guarani, Goiás, Avaí, Brasil-RS e Sampaio Corrêa), mostrou-se heterogêneo em sua composição, e teve envolvimento apenas razoável com as movimentações de mercado.

Finalizando a relação de grupos da variável Time, o “cluster” 5 – também heterogêneo em sua formação - reuniu 12 clubes (RB Bragantino, Botafogo-SP, Brusque, Criciúma, Figueirense, Fortaleza, Juventude, Náutico, Operário-PR, Remo, São Bento e Vila Nova), caracterizando-se por volumes e valores baixos, e a maior média de idade dos jogadores negociados (superior a 30 anos). A Tabela 3 apresenta um resumo destas informações:

Tabela 3. Características dos agrupamentos da variável Time

“Cluster”	Nº de Clubes	Transferências (volume total)	Valor arrecadado (total, em euros)	Média de Idade
1	13	266	€ 32.791.000,00	28.11
2	7	39	€ 785.000,00	26.36
3	9	198	€ 505.295.000,00	26.86
4	6	204	€ 30.420.000,00	28.64
5	12	106	€ 12.000.000,00	30.13
6	2	48	€ 336.900.000,00	26.30

Fonte: Resultados originais da pesquisa

Países Compradores

Embora não com a finalidade de utilizá-la como preditora no modelo de estimativa de valor, a variável original País Comprador também passou pela análise de agrupamentos, seguindo os mesmos critérios utilizados para os agrupamentos da variável Time (volume total de compras realizadas, valor total investido, e média de idade dos jogadores adquiridos no período entre 2018 e 2022).

Ela seria, posteriormente, a variável dependente do modelo Logístico Multinomial complementar à pesquisa, com o objetivo de investigar as probabilidades de que o jogador em análise se transfira, baseado nas mesmas variáveis preditoras do modelo Binomial Negativo inflacionado de zeros, para cada um entre os seis mercados definidos pelo método não-hierárquico, cujo resultado se observa na Figura 2:

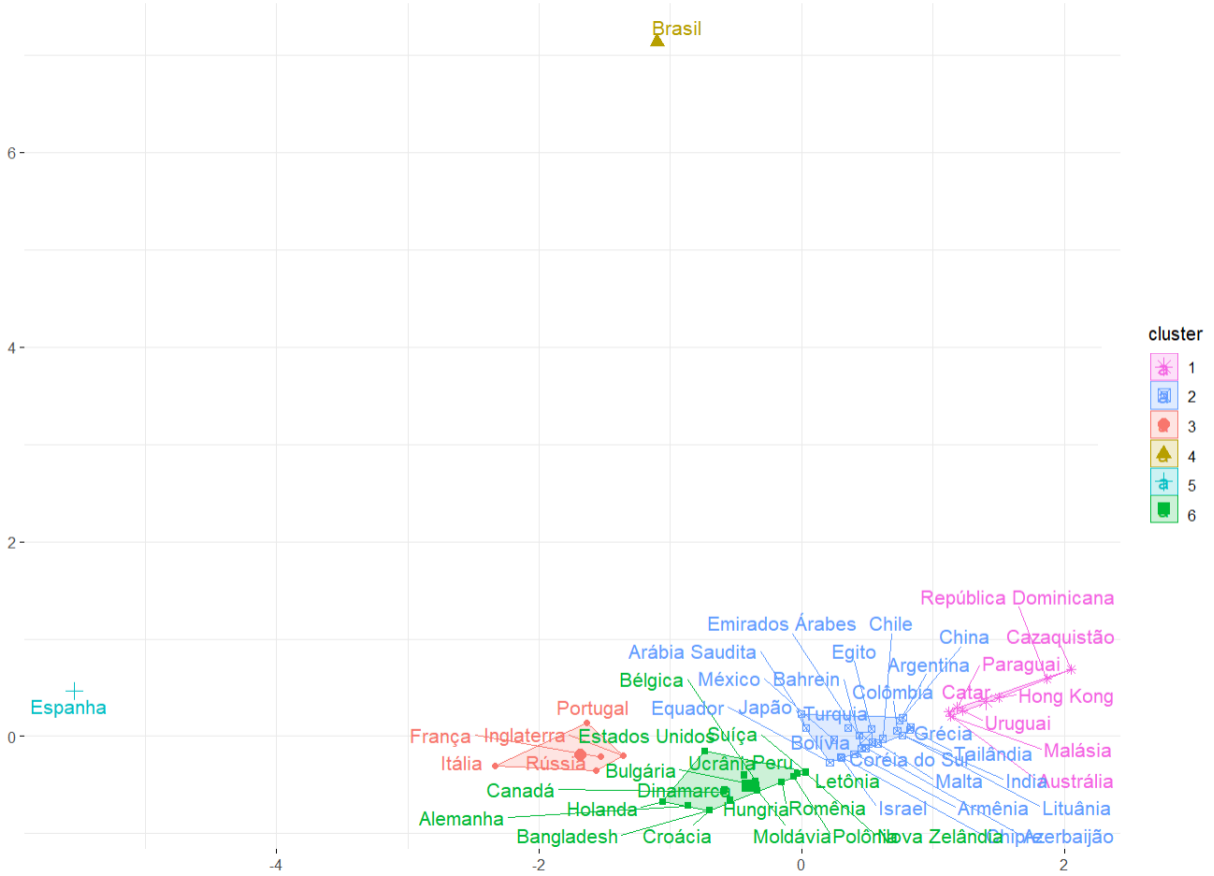

Figura 2. Análise de agrupamentos dos países compradores registrados na base de dados
Fonte: Resultados originais da pesquisa

Nota-se na Figura 2 que dois “clusters” resultaram em apenas um país cada, devido ao grande contraste de ambos com os demais, apresentando-se como “outliers” no gráfico. No grupo 4 ficou o Brasil, que concentrou no mercado interno 584 das 820 negociações registradas na base de dados. A média de idade dos jogadores foi a 2ª mais alta, e o valor total foi o 2º menor. Ou seja, foi neste grupo que se proliferou a inflação de zeros.

Por outro lado, o “cluster” 5 isolou a Espanha que, com apenas 17 compras, alcançou o 2º maior valor total, tendo como principais alvos os jogadores mais jovens e promissores (menor média de idade). O desempenho da Espanha foi fortemente influenciado pelo comportamento agressivo do clube Real Madrid no mercado de transferências.

Composto por cinco países europeus (Itália, Portugal, França, Inglaterra e Rússia), o "cluster" 3 caracteriza-se pela forte relação comercial com o mercado brasileiro, em especial devido ao comportamento de Portugal que, sozinho, teve o 2º maior volume de compras entre os 56 países registrados, atrás apenas do Brasil.

No lado oposto ficou o "cluster" 1, que teve os menores valores (total e médio) e o menor volume médio (2º menor total), além de prospectar os jogadores com a maior média de idade (superior a 32 anos). Seus oito países integrantes (Paraguai, Catar, Austrália, Cazaquistão, Hong Kong, Malásia, República Dominicana e Uruguai), portanto, assemelham-se pela baixa movimentação, focada em alvos "veteranos" e de baixo custo.

O "cluster" 2 foi o mais numeroso, com 23 países (Arábia Saudita, Japão, China, Coréia do Sul, Tailândia, Índia, Turquia, Israel, Armênia, Azerbaijão, Chipre, Grécia, Lituânia, Malta, Emirados Árabes, Egito, Bahrein, México, Argentina, Chile, Bolívia, Colômbia e Equador), o que dificultou a visualização de todos os rótulos dos pontos no gráfico da Figura 2. Por esse motivo, também foi o mais heterogêneo, combinando destinos de quase todos os continentes (oito europeus, seis asiáticos, seis latino-americanos e três africanos). Também por isso, obteve médias baixas de volume (3º menor) e valor (2º menor).

Já o "cluster" 6, encerrando as categorias da variável País Comprador, embora também numeroso, concentrou-se no continente europeu (13 dos 18 integrantes), incluindo seleções de tradição no futebol e destinos comuns a jogadores brasileiros, destacando-se nestes aspectos a Alemanha, principalmente. Formado por Alemanha, Holanda, Dinamarca, Bélgica, Ucrânia, Bulgária, Hungria, Croácia, Letônia, Moldávia, Polônia, Romênia, Suíça, Nova Zelândia, Estados Unidos, Canadá, Peru e Bangladesh, este grupo teve participação apenas razoável no mercado: 3º menor valor (tanto total como médio), 3º menor volume (sendo o 2º menor na média) e 3ª menor média de idade.

A Tabela 4 apresenta um resumo destas informações:

Tabela 4. Características dos agrupamentos da variável País Comprador

"Cluster"	Nº de Países	Transferências (volume total)	Valor investido (total, em euros)	Média de Idade
1	8	18	€ 595.000,00	32.48
2	23	121	€ 109.335.000,00	27.97
3	5	64	€ 372.330.000,00	22.65
4	1	584	€ 50.436.000,00	29.16
5	1	17	€ 281.250.000,00	20.88
6	18	57	€ 104.245.000,00	23.14

Fonte: Resultados originais da pesquisa

Estimativa de valor

Realizados os agrupamentos das variáveis Time e País Comprador, o primeiro passo da modelagem preditiva foi analisar as características da variável dependente Valor, cuja distribuição pode ser observada no histograma da Figura 3:

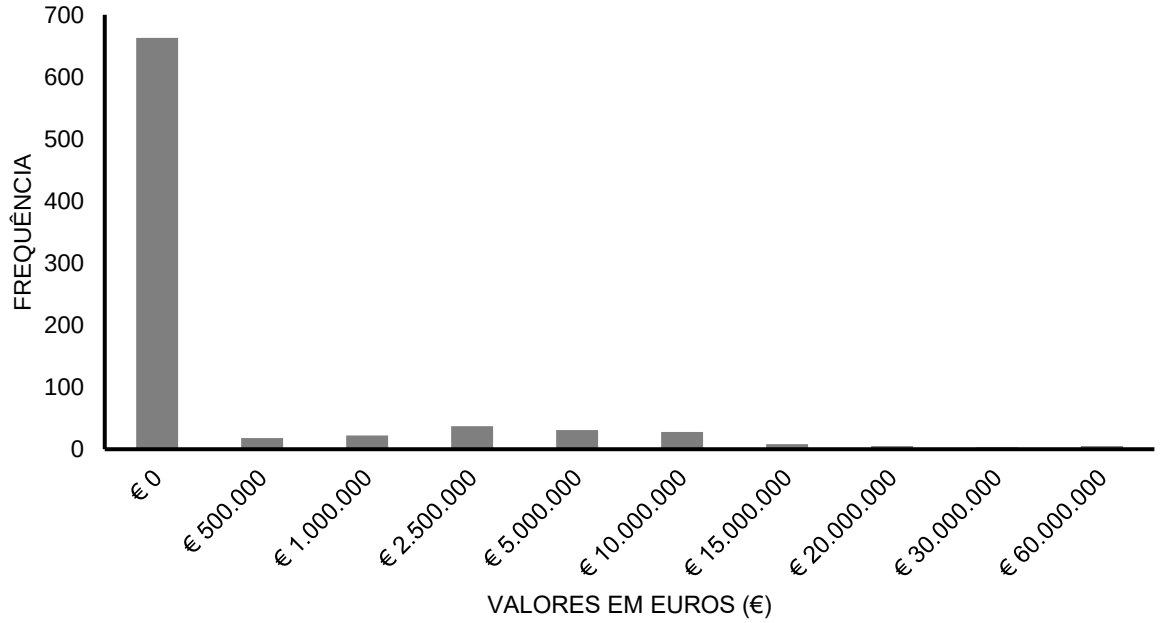

Figura 3. Valores de transferências de jogadores do Brasil registradas entre 2018 e 2022
Fonte: “Transfermarkt” (2023)

A simples visualização do histograma já seria suficiente para revelar fortes indícios de inflação de zeros e de alongamento da cauda, em um “Dataset” composto por 820 observações, das quais 663 têm valor 0 (80.85%). Diversos cenários podem explicar a expressiva ausência de registro de valor, que não necessariamente definam estas observações como “transferências gratuitas”. Algumas das hipóteses são: encerramento de contrato sem renovação (jogador livre para negociar com outro clube); trocas de jogadores em definitivo; perdão de dívidas pendentes do vendedor pelo comprador; valores baixos que não foram capturados pelo sistema de coleta do “Transfermarkt”; entre outras. Ou seja, estas observações devem ser interpretadas como valores considerados irrelevantes pelo provedor.

Ainda assim, os 663 zeros registrados são importantes para revelar padrões de comportamento e perfis de jogadores que não despertaram interesse comercial, mas que participaram do fluxo de negociações entre os clubes. Eles também permitem, ao estabelecer uma comparação, identificar os padrões de desempenho e de perfil do seleto grupo de jogadores envolvidos em transações milionárias.

Além do indicativo de inflação de zeros, a comparação entre a média e a variância da variável dependente reforçou a presença de cauda alongada: enquanto a média é de 1087428, a variância chega a 19716728298846, uma clara tendência à presença de superdispersão.

Esta hipótese foi confirmada com a aplicação do Teste de Cameron-Trivedi (Cameron e Trivedi, 2020), que resultou em p-valor = 0.02008. Sendo ele inferior ao nível de significância de 5%, foi rejeitada a hipótese nula (a variância se mostrou estatisticamente diferente da média), confirmando o direcionamento a um modelo Binomial Negativo para dados de contagem, indicado quando há superdispersão (Cameron e Trivedi, 1990).

O primeiro modelo Binomial Negativo não se ajustou satisfatoriamente aos dados. Todas as 47 variáveis preditoras obtiveram o mesmo p-valor (0.0000000000000002), com erros-padrão próximos de zero, o que pode ter sido provocado por existência de multicolinearidade entre variáveis, baixa amostragem de dados, alta quantidade de preditoras em comparação com o número de observações, ou influência de valores extremos com grande afastamento dos restantes (os “outliers”).

Para investigar a hipótese de multicolinearidade, foi calculado o Variance Inflated Factor (VIF). Seis variáveis apresentaram valores altos de VIF (maiores que 10): Passes_Progressivos_media, Passes_Area_media, Passes_media, Dribles_media, Toques_Area_media e Passes_Terco_Final_media. Como todas as variáveis apresentaram p-valor inferior a 0.05, o procedimento “stepwise” não conseguiria excluí-las.

A solução foi utilizar o valor do parâmetro theta (671432549) encontrado por este modelo como argumento para a criação de um segundo modelo Binomial Negativo, que conseguiu diferenciar a influência das preditoras, resultando em 15 variáveis com significância estatística ao nível de significância de 5%: Idade, Minutos_media, Passes_Chave_media, Pre_Assistencia_media, Passes_Terco_Final_media, Gols, Cruzamentos_media, Faltas_Cometidas_media, Cluster_Vendedor_1, Cluster_Vendedor_2, Cluster_Vendedor_4, Cluster_Vendedor_5, Cluster_Vendedor_6, Posicao_Lateral_Direito e Serie_B. Submetido ao teste de VIF, o novo modelo apresentou valores baixos em todas as variáveis selecionadas, ou seja, houve correção da multicolinearidade encontrada anteriormente.

Os três indicadores de desempenho extraídos do modelo, entretanto, apresentaram valores insatisfatórios, revelando grande espaço para melhora, o que viria a acontecer quando a inflação de zeros na variável dependente passasse a ser considerada.

Sendo assim, o modelo Binomial Negativo inflacionado de zeros foi construído. A execução do procedimento “stepwise” resultou em dez variáveis no lado de contagem e 16 no lado binário da equação, tendo seis variáveis comuns a ambos os lados (no total, 20 variáveis

preditoras com significância estatística, portanto). Houve expressiva melhora nos critérios de avaliação, conforme se observa na Tabela 5:

Tabela 5. Comparativo entre os critérios de avaliação dos modelos

Modelo	Log-Likelihood	AIC	BIC
Binomial Negativo	- 418487238	836974510	836974590
Binomial Negativo Inflacionado de Zeros	- 2697.25	5452.5	5589.069

Fonte: Resultados originais da pesquisa

O “likelihood ratio test” diagnosticou, ainda, que a diferença entre os dois modelos é estatisticamente significativa (qui-quadrado = 836969081 e p-valor = 0.00000000000000022), enquanto o teste de Vuong (Vuong, 1989) confirmou o melhor ajuste do modelo inflacionado de zeros aos dados, com valor positivo obtido no teste Z e significância estatística ao nível de 5% estabelecido (z = 10.881 e p-valor = 0.0000000000000002), rejeitando a hipótese nula de que os modelos teriam ajustes iguais.

A Tabela 6 apresenta as variáveis selecionadas e seus respectivos parâmetros, incluindo o intercepto, para o modelo de contagem da equação Binomial Negativa inflacionada de zeros:

Tabela 6. Coeficientes do modelo de contagem

Variável	Coeficiente Estimado	Erro-Padrão	p-valor
(intercepto)	17.465554	0.446424	0.000000002
Idade	-0.141605	0.015969	0.000000002
Jogos	-0.016089	0.006995	0.021455
Minutos_media	0.009066	0.002911	0.021455
Gols	0.091736	0.027575	0.000879
Dribles_media	0.180775	0.051619	0.000462
Passes_Terco_Final_media	0.094135	0.027248	0.000551
Cluster_Vendedor_1	-1.410694	0.166868	0.000000002
Cluster_Vendedor_2	-2.221173	0.553983	0.000060857
Cluster_Vendedor_4	-0.955199	0.191089	0.000000577
Cluster_Vendedor_6	0.731587	0.164951	0.000009199

Fonte: Resultados originais da pesquisa

Entre as dez variáveis selecionadas pelo modelo de contagem, seis têm origem nos dados de desempenho e/ou perfil (Idade, Jogos, Minutos_media, Gols, Dribles_media e

Passes_Terco_Final_media). Como era de se esperar, quanto menor a idade do jogador, maior a probabilidade de se realizar transferências com valor diferente de zero, o que se verifica pelo sinal negativo do coeficiente.

Chama a atenção que o total de jogos disputados também teve influência negativa na formação do valor médio (no caso, quanto menor, melhor), o que se deve à baixa exposição dos jogadores negociados por valores expressivos em partidas do Campeonato Brasileiro. Sendo eles jovens e promissores, fica evidente que os observadores internacionais especializados em prospecção de talentos os monitoram desde muito antes da profissionalização.

Quando estes atletas chegam às principais divisões do futebol brasileiro, em poucos jogos já são negociados. Não é exclusivamente pelos seus desempenhos nas séries A ou B que despertam a cobiça dos compradores, mas sim pelas observações realizadas enquanto estavam em categorias de base. A breve trajetória entre os profissionais serve apenas para validar o desfecho.

Os principais provedores de dados de futebol, entretanto, ainda negligenciam as principais competições de base brasileiras, devido a dificuldades estruturais, como a ausência de televisionamento de todas as partidas. Sediadas em outros países, estas empresas coletam dados remotamente, ou seja, sem transmissão das partidas ou captação própria das imagens pelos provedores, não há geração de dados, o que torna os “Datasets” destas competições insuficientes para submissão a modelos preditivos.

Embora consideradas como variáveis de desempenho, por serem originárias do banco de dados do “Wyscout”, Idade, Jogos e Minutos_media dizem respeito ao perfil/biografia do atleta. Na prática, foram selecionadas três variáveis verdadeiramente de desempenho no modelo de contagem, todas de viés ofensivo e com influência positiva na formação dos preços (quanto maiores, melhor): Gols, Dribles_media, e Passes_Terco_Final_media.

Ou seja, mesmo que nenhuma “dummy” com as posições dos jogadores tenha sobrevivido ao procedimento “stepwise”, a característica ofensiva é um claro indicativo que os jogadores jovens com funções de criação e finalização, além de habilidade individual, foram mais valorizados, em detrimento daqueles com atribuições defensivas.

As outras quatro variáveis selecionadas pelo modelo de contagem referem-se aos “clusters” dos clubes vendedores. Apenas o grupo 5 não teve o parâmetro aprovado. Entre os quatro selecionados, somente o “cluster” 6, composto por Flamengo e Palmeiras, teve impacto positivo nos preços em comparação com o grupo de referência (“cluster” 3), confirmando o que havia ficado nítido durante a análise de agrupamentos. Isso significa dizer que jogadores procedentes de Flamengo e Palmeiras obtiveram maior valorização na formação dos preços.

Os demais “clusters” selecionados (1, 2 e 4) influenciaram negativamente na formação do valor.

No lado Logístico Binário da fórmula, a Tabela 7 apresenta as variáveis selecionadas, e seus respectivos parâmetros - incluindo o intercepto - do modelo inflacionado de zeros:

Tabela 7. Coeficientes do modelo Logístico Binário

Variável	Coeficiente Estimado	Erro-Padrão	p-valor
(intercepto)	-6.455614	1.195920	0.0000000673754
Idade	0.375047	0.037692	0.000000000000002
Jogos	0.036338	0.015849	0.021859
Minutos_media	-0.052407	0.008098	0.0000000000971
Conversao	-0.025168	0.010676	0.018402
Toques_Area_media	-0.490028	0.150069	0.001093
Conducoes_media	-0.453605	0.154761	0.003379
Faltas_Recebidas_media	0.514610	0.188285	0.006273
Interceptacoes_media	0.233214	0.115490	0.043452
Cluster_Vendedor_1	1.699930	0.336554	0.0000004395212
Cluster_Vendedor_2	2.751237	0.831447	0.000936
Cluster_Vendedor_4	1.069667	0.392597	0.006438
Cluster_Vendedor_5	3.611366	1.156163	0.001787
Posicao_Meia	-1.863792	0.487640	0.000132
Posicao_Volante	-1.147080	0.413735	0.005563
Serie_B	1.014593	0.332264	0.002261
Serie_C	2.780882	1.200384	0.020527

Fonte: Resultados originais da pesquisa

O primeiro aspecto a se destacar é a inversão dos sinais dos coeficientes estimados das variáveis Idade, Jogos e Minutos_media, além do sinal do intercepto. Se, na formação do valor observada no modelo de contagem, Idade e Jogos tinham influência negativa (quanto menores, maiores os preços praticados), no modelo inflacionado de zeros ambas apresentam influência positiva (quanto maiores a idade do jogador e o número de partidas por ele disputadas, maiores as probabilidades de o valor ser nulo ou considerado irrelevante).

Já a variável Minutos_media passou de influência positiva na formação do valor para influência negativa na inflação de zeros. O mesmo aconteceu com o intercepto, ou seja, a ausência de uma ou mais das categorias de referência (Posição Extremo-Esquerdo, Liga

Série A, Cluster_Vendedor_3 e Janela Verão) aumentou as probabilidades de ocorrência de zero na variável dependente Valor.

Cinco variáveis específicas de desempenho técnico foram selecionadas no lado Logístico Binário da equação. Nenhuma delas constava no modelo de contagem: Conversao, Toques_Area_media e Conducoes, estas com influência negativa na inflação de zeros; Faltas_Recebidas_media e Interceptacoes_media, ambas com influência positiva. Mais uma vez o viés ofensivo se apresenta como determinante, afinal, os jogadores com maiores índices de conversão do perigo criado em gols, de toques na bola dentro da área adversária e de conduções com a bola dominada tiveram menor probabilidade de registrar valor zero, enquanto jogadores com mais interceptações (variável defensiva) e faltas recebidas (esta independente do viés) apresentaram maior probabilidade de registrar valor zero.

Novamente, quatro agrupamentos da variável Time (vendedor) foram aprovados, mas com uma substituição em comparação ao modelo de contagem. Como era de se esperar, o “cluster” 6 não aparece como um componente da inflação de zeros (Palmeiras e Flamengo contribuíram fortemente aos valores altos, como observado anteriormente), dando lugar ao “cluster” 5. Nota-se que todos os “clusters” selecionados têm influência positiva na inflação de zeros, ou seja, jogadores de origens diferentes dos “clusters” 3 (referência) e 6 apresentaram maior probabilidade de registrar valor zero.

Além disso, quatro “dummies” de variáveis que não haviam contribuído à construção do modelo de contagem foram incluídas na investigação da alta incidência de valores nulos pelo modelo Logístico Binário. Duas da variável Posição (Meia e Volante), ambas com parâmetro de sinal negativo, e duas da variável Liga (Serie_B e Serie_C), estas com sinal positivo. Ou seja, enquanto meias e volantes negociados ajudaram a amenizar a inflação, jogadores provenientes das competições Série B e Série C (principalmente) influenciaram na ocorrência do fenômeno (valores iguais a zero).

O modelo Logístico Binário executado separadamente e submetido ao procedimento “stepwise” permitiu que fossem extraídos critérios de avaliação, considerados satisfatórios: “log likelihood” = -193.7249; AIC = 421.4498; BIC = 501.5080; e estatística qui-quadrado = 413.4245. Com base na matriz de confusão, também foram obtidos 90.85% de acurácia, 70.06% de sensitividade, 95.78% de especificidade e 94.1% de área abaixo da curva ROC. A alta sensitividade indica que o modelo foi mais eficiente para capturar a ausência de valor.

Com base nos parâmetros estimados para as variáveis selecionadas nos dois lados do modelo Binomial Negativo inflacionado de zeros, a expressão para fins preditivos do valor médio pode ser escrita da seguinte forma:

$$\mu inflate = 1 - 1/(1 + \{e\text{^}(-[-6.455614 + 0.375047(Idade) + 0.036338(Jogos) - 0.052407(Minutos\_media) - 0.025168(Conversao) - 0.490028(Toques\_Area\_media) - 0.453605(Conducoes\_media) + 0.514610(Faltas\_Recebidas\_media) + 0.233214(Interceptacoes\_media) + 1.699930(Cluster\_Vendedor\_1) + 2.751237(Cluster\_Vendedor\_2) + 1.069667(Cluster\_Vendedor\_4) + 3.611366(Cluster\_Vendedor\_5) - 1.863792(Posicao\_Meia) - 1.147080(Posicao\_Volante) + 1.014593(Serie\_B) + 2.780882(Serie\_C)])\}.\{e\text{^}([17.465554 - 0.141605(Idade) - 0.016089(Jogos) + 0.009066(Minutos\_media) + 0.091736(Gols) + 0.180775(Dribles) + 0.094135(Passes\_Terco\_Final\_media) - 1.410694(Cluster\_Vendedor\_1) - 2.221173(Cluster\_Vendedor\_2) - 0.955199(Cluster\_Vendedor\_4) + 0.731587(Cluster\_Vendedor\_6)])\}$$

A comparação entre os valores previstos e os valores reais evidencia que o mercado europeu, em especial o espanhol, transigiu aos preços praticados pelos clubes brasileiros, concretizando aquisições acima do esperado, o que se observa na Tabela 8:

Tabela 8. Comparação entre valores reais e valores estimados

Jogador	Vendedor	Comprador	Valor real	Valor estimado
Endrick	Palmeiras	Real Madrid-ESP	€ 60.000.000	€ 17.574.451
Vinicius Júnior	Flamengo	Real Madrid-ESP	€ 45.000.000	€ 37.498.271
Rodrygo	Santos	Real Madrid-ESP	€ 45.000.000	€ 30.546.177
Lucas Paquetá	Flamengo	Milan-ITA	€ 38.400.000	€ 34.444.062
Arthur Melo	Grêmio	Barcelona-ESP	€ 31.000.000	€ 17.387.017
Reinier	Flamengo	Real Madrid-ESP	€ 30.000.000	€ 21.270.792
Yuri Alberto	Internacional	Zenit-RUS	€ 25.000.000	€ 11.138.925
Renan Lodi	Athletico	Atlético-ESP	€ 21.750.000	€ 8.608.349
Gérson	Flamengo	Olympique-FRA	€ 20.000.000	€ 10.101.645
Bruno Guimarães	Athletico	Lyon-FRA	€ 20.000.000	€ 8.547.685

Fonte: Resultados originais da pesquisa

Na primeira posição, a venda do “outlier” Endrick foi aproximadamente 3.4 vezes maior do que a estimativa de valor baseada em suas atuações pelo Palmeiras no Campeonato Brasileiro 2022, configurando-se em um dos exemplos de alvo prospectado ainda nas

categorias de base. Apesar da alta quantia investida, o Real Madrid sequer teve interesse em utilizá-lo imediatamente, mantendo-o no Palmeiras durante a temporada 2023 para aumentar sua experiência entre os profissionais antes de incluí-lo no elenco.

Sob a perspectiva dos valores estimados, entre as dez maiores vendas da base de dados quem desponta é Vinicius Júnior, que chegou a € 37.498.271 conforme a Tabela 8. Neste caso, a diferença entre predição e realidade foi menor. Em 2º lugar está Lucas Paquetá, com o qual o Milan obteve a menor margem entre valor estimado e valor real na lista das dez maiores transações, pagando apenas cerca de € 4 milhões a mais do que o esperado.

Chama a atenção a habilidade do Athletico nas negociações com os clubes europeus. Seus dois jogadores presentes na Tabela 8 (Renan Lodi em 8º e Bruno Guimarães em 10º) renderam € 20 milhões cada, embora seus valores estimados sejam inferiores a € 9 milhões, ou seja, resultaram em um lucro de aproximadamente € 23 milhões ao clube paranaense, o que reitera o conceito exposto por Shapiro et al. (2013) quanto aos pontos de vista diferentes entre compradores e vendedores, estabelecendo-se uma barganha - nestes casos, favorável ao Athletico.

Prospecção de mercado comprador

O modelo Logístico Multinomial, complementar à pesquisa, foi utilizado para investigar a probabilidade de que o jogador em análise seja vendido para um entre os seis mercados compradores identificados através da análise de agrupamentos. Após o procedimento “stepwise”, o modelo resultou em 23 variáveis preditoras. A Tabela 9 apresenta os coeficientes de cada variável selecionada, em cada categoria, incluindo o intercepto:

Tabela 9. Coeficientes do modelo Logístico Multinomial

(continua)

Variável	cluster 1	cluster 2	cluster 4	cluster 5	cluster 6
(intercepto)	-15.1041	-7.8209	-7.5592	11.3290	1.5332
Idade	0.513029	0.330455	0.395196	-2.91478	0.035705
Minutos_media	-0.01915	-0.01315	-0.03246	0.286759	-0.03209
xG_media	3.654003	-2.15829	-0.94557	83.10066	-1.93971
Conversao	0.027056	0.044518	0.026319	0.384703	0.018317
Faltas_Recebidas_media	-0.34713	0.094841	0.054233	-5.54379	-0.25707
Assistencias	-0.73507	-0.10638	-0.16931	-2.25428	0.11380
Passes_Chave_media	1.604479	-0.49469	-0.17707	9.187127	-1.37228

Tabela 9. Coeficientes do modelo Logístico Multinomial

(conclusão)

Variável	cluster 1	cluster 2	cluster 4	cluster 5	cluster 6
Passes_media	-0.09496	-0.00228	-0.02183	1.005759	0.007423
Passes_Progressivos_media	0.270556	0.018323	0.194473	-6.52145	-0.21326
Passes_Area_media	-0.57716	0.310599	0.044292	4.300304	1.494866
Carrinhos_media	0.065939	-0.00972	-0.48552	9.663772	-1.03273
Bloqueios_media	2.41878	2.304781	1.933285	-27.3188	3.317684
Cluster_Vendedor_1	1.196507	1.305554	1.942666	-11.8944	0.710165
Cluster_Vendedor_2	-18.3383	44.88672	46.04341	36.84028	44.58661
Cluster_Vendedor_4	0.122718	1.257511	1.728243	-24.8049	0.185053
Cluster_Vendedor_5	0.545400	0.523810	2.241001	-28.8763	0.743692
Cluster_Vendedor_6	-0.96528	0.359315	-1.00235	-4.05629	-2.74718
Posicao_Atacante	36.36326	34.23467	35.06283	61.84997	-43.5697
Posicao_Centroavante	0.572815	0.794781	0.448543	-91.3442	0.406943
Posicao_Lateral_Direito	1.313800	-1.86476	-0.36892	11.74802	-1.05136
Posicao_Lateral_Esquerdo	-45.3939	-0.59814	0.018695	25.32866	0.294509
Posicao_Zagueiro	0.406265	-1.49167	-0.85199	-23.1064	-2.08600
Janela_Inverno	1.747879	0.732361	1.170928	2.630965	0.691254

Fonte: Resultados originais da pesquisa

O modelo alcançou os seguintes valores nos critérios de avaliação, também considerados satisfatórios: “log likelihood” = -583.7385; AIC = 1407.477; BIC = 1972.594, qui-quadrado = 571.4246; e acurácia = 73%, também considerados satisfatórios. Com os coeficientes obtidos é possível construir as expressões dos cinco logitos, seguindo a forma genérica apresentada anteriormente, que servirão ao cálculo das probabilidades.

O primeiro aspecto a se destacar, na comparação com o modelo principal, é a inclusão de cinco “dummies” da variável Posição. Ou seja, enquanto no modelo de contagem nenhuma das nove diferentes posições dos jogadores alcançou significância estatística, em comparação com a categoria de referência (Extremo-Esquerdo), e no modelo Logístico Binário foram incluídas apenas duas (Meia e Volante), no modelo Logístico Multinomial passaram a ser relevantes as posições Atacante, Centroavante, Lateral-Direito, Lateral-Esquerdo e Zagueiro, com vieses positivos ou negativos conforme o mercado comprador.

Da mesma forma, a variável Janela, que não foi incluída em nenhum dos lados do modelo Binomial Negativo para dados de contagem inflacionados de zeros, desta vez obteve significância estatística, com a categoria Inverno influenciando positivamente as probabilidades em todos os “clusters”, na comparação com a categoria de referência (Verão).

Outra constatação é a grande diferença que algumas das variáveis selecionadas apresentam entre seus coeficientes. A Posicao_Centroavante, por exemplo, tem coeficientes positivos e próximos a zero em quatro grupos, enquanto no “cluster” 5 sua influência foi fortemente negativa (-91.3442). Ou seja, durante o período analisado, os centroavantes do futebol brasileiro tiveram reduzida probabilidade de ser negociados com clubes da Espanha.

Este é o tipo de análise que pode agregar valor à utilização dos modelos preditivos, sob a perspectiva dos dirigentes de clubes do Brasil. Saber quais os potenciais compradores para cada perfil auxilia a identificar não apenas seus jogadores mais valiosos, mas também para quais clubes oferecê-los, ou até mesmo em qual posição utilizá-los, vislumbrando chamar a atenção dos melhores compradores, e assim melhorar suas projeções de receitas.

Considerações Finais

A abordagem para estimar o valor de mercado de jogadores de futebol com dados do Campeonato Brasileiro das Séries A e B, proposta neste estudo, obteve resultados satisfatórios, e assim pode ser extrapolada para utilização em clubes, requerendo ajustes e adaptações conforme cada cenário.

O desenvolvimento da solução dependerá das características da base de dados disponíveis. Este estudo deparou-se com um “Dataset” inflacionado de zeros, tendo variável dependente com distribuição Binomial Negativa, o que foi preponderante para a escolha do modelo regressivo utilizado. Isto não significa que o fenômeno se repita diante de informações provenientes de outras fontes, em diferentes contextos.

Um bom exemplo é a aplicação da estimativa de valor para jogadores de base, um segmento com grande potencial de crescimento devido à carência de dados confiáveis, conforme mencionado anteriormente. Ao invés de depender de provedores externos, que em geral negligenciam competições alheias ao futebol profissional, o clube pode coletar dados próprios para formular um modelo capaz de estimar valor e probabilidade de venda dos seus jogadores com base nos preços e desempenhos daqueles formados e negociados por ele próprio, o que garantiria maior credibilidade e autonomia para a modelagem.

O primeiro passo é explorar o banco de dados disponível e identificar a variável dependente, e sua distribuição, para definir a abordagem estatística mais apropriada aos objetivos do projeto. Vale destacar, ainda, que na abordagem de mercado a estimativa de valor pode ser quantitativa, mas ela deixa espaço aberto a julgamentos subjetivos

(Damodaran, 2012). Como argumenta Póvoa (2008), o mais importante é entender a sensibilidade que o valor estimado tem às oscilações das variáveis preditoras, diante de um processo que é dinâmico.

Referências

XP Investimentos Corretora de Câmbio, Títulos e Valores Mobiliários [XP]. 2022. Convocados: Finanças, História e Mercado do Futebol Brasileiro em 2021. Disponível em: <https://conteudos.xpi.com.br/relatorio-futebol-2022/>. Acesso em: 11 out. 2022.

International Valuation Standards Council [IVSC]. 2022. International Valuation Standards (IVS). Disponível em: <https://www.rics.org/contentassets/542170a3807548a28aebb053152f1c24/ivsc-effective-31-jan-2022.pdf>. Acesso em: 11 out. 2022.

Póvoa, A. 2008. Valuation: como precificar ações. 2ed. Editora Globo, São Paulo, SP, Brasil.

Shapiro, E.; Mackmin, D.; Sams, G. 2013. Modern methods of valuation. 11ed. Routledge, Abingdon, Oxfordshire, Reino Unido.

Damodaran, A. 2012. Investment Valuation: tools and techniques for determining the value of any asset. 3ed. John Wiley & Sons, Hoboken, Nova Jersey, Estados Unidos.

Vida, E.; Alves, N.; Barboza, F.; Ferreira, R.; Gonçalves, P.; Marque, L.; Maschietto, L.; Oliveira, H.; Souza, D. 2021. Data Warehouse. 1ed. Sagah, Porto Alegre, RS, Brasil.

R Core Team. 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Disponível em: <http://www.R-project.org/>.

Zivkovic, J. 2023. worldfootballR: Extract and Clean World Football (Soccer) Data. Disponível em <https://github.com/JaseZiv/worldfootballR>.

Wickham, H; Gerolemund, G. 2016. R For Data Science: Import, Tidy, Transform, Visualize and Model Data. O'Reilly Media, Sebastopol, Califórnia, Estados Unidos.

Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; D'Agostino, M.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; Kuhn, M.; Pedersen, T.L.; Miller, E.; Bache, S.M.; Müller, K.; Ooms, J.; Robinson, D.; Seidel, D.P.; Spinu, V.; Takahashi, K.; Vaughan, D.; Wilke, C.; Woo, K.; Yutani, H. 2019. Welcome to the tidyverse. Journal of Open Source Software 4(43): 1686. Disponível em <https://doi.org/10.21105/joss.01686>.

Kassambara, A.; Mundt, F. 2020. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. Disponível em <https://CRAN.R-project.org/package=factoextra>.

Fávero, L.; Belfiore, P. 2017. Análise de Dados: estatística e modelagem multivariada com Excel, SPSS e Stata. 1ed. GEN LTC, Rio de Janeiro, RJ, Brasil.

Kaplan, J. 2020. fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. Disponível em: <https://CRAN.R-project.org/package=fastDummies>.

Harrell Jr, F.E. 2021. Hmisc: Harrell Miscellaneous. Disponível em: <https://CRAN.R-project.org/package=Hmisc>.

Garganta, J. 2001. Futebol e ciência. Ciência e futebol. Revista Digital Educación Física e Deportes. Disponível em <https://www.efdeportes.com/efd40/fcienc.htm>. Acesso em: 11 outubro 2022.

Teoldo, I.; Garganta, J.; Greco, P.; Mesquita, I. 2010. Análise e avaliação do comportamento tático no futebol. Revista da Educação Física/UEM 21: 443-455.

Cameron, A.C.; Trivedi, P.K. 2020. overdisp: Overdispersion in Count Data Multiple Regression Analysis. Disponível em: <https://CRAN.R-project.org/package=overdisp>.

Cameron, A.C.; Trivedi, P.K. 1990. Regression-based tests for overdispersion in the Poisson model. Journal of Econometrics 46:347-364

Gardner, W.; Mulvey, E.P.; Shaw, E.C. 1995. Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson, and Negative Binomial Models. Psychological bulletin. 118:392-404

Venables, W. N.; Ripley, B. D. 2002. Modern Applied Statistics with S. 4ed. Springer, Nova York, Estados Unidos.

Ron, S. 1981. Who Invented the Variance Inflation Factor? Discussion of How Cuthbert Daniel Invented the Variance Inflation Factor. Disponível em: <https://www.researchgate.net/publication/291808767_Who_Invented_the_Variance_Inflation_Factor>. Acesso em: 11 abr. 2023.

Fox, J.; Weisberg, S. 2019. An R Companion to Applied Regression, 3ed. Sage, Thousand Oaks, Califórnia, Estados Unidos.

Hilbe, J.M. 2014. Modeling Count Data. 1ed. Cambridge University Press, Nova York, Estados Unidos.

Hastie, T. J. and Pregibon, D. 1991. Statistical Models in S. 1ed. Routledge, Abingdon, Oxfordshire, Reino Unido.

Konishi, S.; Kitagawa, G. 2008. Information Criteria and Statistical Modeling. 1ed. Springer, Nova York, Estados Unidos.

Lambert, D. 1992. Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. Technometrics 34:1-14

Tang, W.; Wang, J.W.; Cheng, D.G. 2017. Untangle the Structural and Random Zeros in Statistical Modelings. J Appl Stat. 45(9):1714-1733

Zeileis, A.; Kleiber, C.; Jackman, S. 2008. Regression Models for Count Data in R. Journal of Statistical Software 27(8). Disponível em: <http://www.jstatsoft.org/v27/i08/>.

Zhu Wang (2022). mpath: Regularized Linear Models. Disponível em: <https://CRAN.R-project.org/package=mpath>.

Zeileis, A.; Hothorn, T. 2002. Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. Disponível em: <https://CRAN.R-project.org/doc/Rnews/>.

Vuong, Q.H. 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica 57(2):307-333

Merkle, E.; You, D. 2020. nonnest2: Tests of Non-Nested Models. Disponível em: <https://CRAN.R-project.org/package=nonnest2>.

Kuhn, M. 2021. caret: Classification and Regression Training. Disponível em: <https://CRAN.R-project.org/package=caret>.

Pearson, K. 1904. Mathematical Contributions to the Theory of Evolution. XIII. On the Theory of Contingency and Its Relation to Association and Normal Correlation. Drapers' Company Research Memoirs. Biometric Series I. Dulau and Co., Londres, Inglaterra.

Powers, D. 2008. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Mach. Learn. Technol. 2.

Hosmer, D.W; Lemesbow, S. 1980. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics - Theory and Methods 9(10): 1043-1069.

Long, J.A. 2022. jtools: Analysis and Presentation of Social Scientific Data. Disponível em: <https://cran.r-project.org/package=jtools>.

Hilbe, J.M. 2009. Logistic Regression Models. 1ed. Chapman & Hall/CRC, Boca Ratón, Flórida, Estados Unidos.