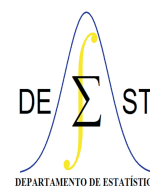




UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E
BIOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA



Pablo Vinícius dos Reis e Silva

Predição e Otimização para Escalação de Equipes no Cartola
FC por meio de Aprendizado de Máquina e Programação Inteira

Ouro Preto
2026

Pablo Vinícius dos Reis e Silva

Predição e Otimização para Escalação de Equipes no Cartola
FC por meio de Aprendizado de Máquina e Programação Inteira

Trabalho de conclusão de curso apresentado
ao curso de Estatística da Universidade Fe-
deral de Ouro Preto como requisito parcial
para obtenção do título de Bacharel em Es-
tatística.

Orientador: Dr. Tiago Martins Pereira

Ouro Preto
2026



FOLHA DE APROVAÇÃO

Pablo Vinícius dos Reis e Silva

Predição e otimização para escalação de equipes no Cartola FC por meio de aprendizado de máquina e programação inteira

Monografia apresentada ao Curso de Estatística da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística

Aprovada em 31 de julho de 2026

Membros da banca

Dr. Tiago Martins Pereira - Orientador (Universidade Federal de Ouro Preto)
Dra. Diana Campos de Oliveira (Universidade Federal de Ouro Preto)
Dr. Marcelo Carlos Ribeiro (Universidade Federal de Ouro Preto)

Prof. Dr. Tiago Martins Pereira, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 31/07/2026



Documento assinado eletronicamente por **Tiago Martins Pereira**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 01/08/2026, às 14:12, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Marcelo Carlos Ribeiro**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 03/08/2026, às 19:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1150740** e o código CRC **54849F84**.

Dedico este trabalho à minha família, que esteve ao meu lado em todas as etapas desta caminhada, oferecendo apoio, incentivo e confiança.

Em especial, à minha mãe, Tânia (*in memoriam*), minha maior força e inspiração. Esta conquista também pertence a vocês.

Agradecimentos

Agradeço primeiramente a Deus, por ter me dado força, saúde e perseverança para enfrentar os desafios encontrados ao longo desta caminhada.

À minha família, pelo apoio, pela compreensão e pelo incentivo durante toda a minha trajetória acadêmica. Foram anos de desafios, mudanças, aprendizados e decisões importantes, e ter pessoas ao meu lado acreditando na minha capacidade foi fundamental para que eu chegasse até aqui.

Ao meu orientador, Prof. Dr. Tiago Martins Pereira, agradeço pela orientação, pela disponibilidade e pelas contribuições ao desenvolvimento deste trabalho. Seus conhecimentos, suas sugestões e sua dedicação foram essenciais para a construção e o aprimoramento desta pesquisa.

Agradeço aos professores do Departamento de Estatística da Universidade Federal de Ouro Preto por todo o conhecimento compartilhado ao longo da graduação. Cada disciplina, trabalho e experiência contribuiu não apenas para minha formação profissional, mas também para o desenvolvimento de uma forma mais crítica e analítica de compreender os problemas.

Aos meus amigos de Ouro Preto e colegas de curso, agradeço pela convivência, pelas conversas, pelo apoio e pelos momentos compartilhados durante essa trajetória. A caminhada tornou-se mais leve graças às pessoas que estiveram presentes nos períodos de dificuldade e também nas conquistas.

Por fim, agradeço a todos que, de alguma forma, fizeram parte desta caminhada e contribuíram para que a conclusão desta etapa se tornasse possível.

“Em Deus nós confiamos; todos os outros devem trazer dados.”

Atribuída a W. Edwards Deming

RESUMO

O Cartola FC é um jogo de fantasia esportiva em que os participantes montam equipes sob restrições de orçamento, formação tática e quantidade de atletas por clube, buscando maximizar a pontuação em cada rodada do Campeonato Brasileiro. A elevada variabilidade do desempenho dos jogadores e o risco de não participação tornam essa decisão complexa. Este trabalho propõe uma metodologia de apoio à escalação que integra aprendizado de máquina e otimização matemática. Foi construída uma base histórica com dados públicos dos atletas, seguida de pré-processamento, engenharia de atributos e seleção de variáveis. Modelos Random Forest, XGBoost e *ensemble* foram desenvolvidos para estimar a pontuação futura dos jogadores. As previsões foram incorporadas a um modelo de programação inteira que seleciona automaticamente a equipe, respeitando orçamento, posições, formação e limite de atletas por clube. A avaliação considerou métricas de regressão, capacidade de ranqueamento, estratégias de escalação, escolha do capitão e desempenho das equipes nas 37 rodadas da temporada de 2025. O XGBoost apresentou o melhor desempenho preditivo individual, enquanto a estratégia baseada na previsão do *ensemble* produziu a maior pontuação média das equipes. As estratégias preditivas superaram aquelas baseadas apenas na média histórica ou no valor de mercado. Os resultados também indicaram relevância do desempenho recente, da consistência e da disponibilidade dos jogadores. Conclui-se que a integração entre aprendizado de máquina e programação inteira constitui uma abordagem quantitativa, sistemática e reproduzível para apoiar decisões de escalação em um problema sujeito a múltiplas restrições e elevada incerteza.

Palavras-chave: aprendizado de máquina, Cartola FC, otimização, previsão, programação inteira, Random Forest, XGBoost.

ABSTRACT

Cartola FC is a fantasy football game in which participants build teams under budget, tactical formation, and club-related constraints, seeking to maximize the score obtained in each round of the Brazilian Football Championship. The high variability of player performance and the risk of non-participation make this decision complex. This study proposes a team-selection support methodology that integrates machine learning and mathematical optimization. A historical database was built from publicly available player data, followed by preprocessing, feature engineering, and predictor selection. Random Forest, XGBoost, and *ensemble* models were developed to estimate players' future scores. The resulting predictions were incorporated into an integer programming model that automatically selects a team while respecting budget, position, formation, and club-limit constraints. The evaluation considered regression metrics, ranking ability, team-selection strategies, captain selection, and team performance over the 37 rounds of the 2025 season. XGBoost achieved the best individual predictive performance, whereas the strategy based on the *ensemble* prediction produced the highest average team score. Predictive strategies outperformed approaches based solely on historical averages or player market values. The results also indicated the relevance of recent performance, consistency, and player availability. The integration of machine learning and integer programming therefore provides a quantitative, systematic, and reproducible approach to supporting team-selection decisions in a problem characterized by multiple constraints and substantial uncertainty.

Keywords: Cartola FC, integer programming, machine learning, optimization, prediction, Random Forest, XGBoost.

Lista de figuras

- Figura 1 – Representação esquemática do processo de comunicação entre um script desenvolvido em Python e a API do Cartola FC. O sistema realiza requisições HTTP aos endpoints da API, recebendo respostas em formato JSON contendo informações sobre atletas, clubes, partidas, rodadas e estatísticas utilizadas na construção da base analítica. 17
- Figura 2 – Representação esquemática do algoritmo Random Forest. O método constrói múltiplas árvores de decisão a partir de amostras bootstrap da base de treino e combina suas previsões por meio da média, reduzindo a variância e aumentando a robustez do modelo. 26
- Figura 3 – Representação esquemática do algoritmo XGBoost. As árvores são ajustadas sequencialmente, de modo que cada nova árvore busca corrigir os erros residuais das previsões anteriores, produzindo ao final uma predição agregada e regularizada. 27
- Figura 4 – Variáveis originais de maior importância no modelo Random Forest. . . 41
- Figura 5 – Variáveis originais de maior importância no modelo XGBoost. 41

Lista de tabelas

Tabela 1 – Categorias de disponibilidade dos atletas fornecidas pela API do Cartola FC e interpretação utilizada para estimar a probabilidade de participação dos jogadores nas rodadas.	19
Tabela 2 – Conversão dos status de disponibilidade fornecidos pela API do Cartola FC em probabilidades de participação utilizadas na modelagem preditiva.	19
Tabela 3 – Comparação do desempenho dos modelos preditivos em relação ao modelo baseline.	37
Tabela 4 – Importância média das principais variáveis nos modelos Random Forest e XGBoost.	40
Tabela 5 – Pontuação real média dos atletas selecionados pelos diferentes critérios de ranqueamento.	45
Tabela 6 – Média da soma das pontuações reais nos grupos Top- k	46
Tabela 7 – Desempenho médio das estratégias de escalação.	49
Tabela 8 – Desempenho dos capitães escolhidos nas diferentes estratégias de escalação.	51
Tabela 9 – Frequência de bônus positivo e negativo dos capitães.	53
Tabela 10 – Resultados das estratégias de escalação na rodada 21.	54
Tabela 11 – Resultados das estratégias de escalação na rodada 9.	56

Sumário

1	Introdução	12
2	Aspectos Metodológicos	16
2.1	Coleta e Organização dos Dados	16
2.1.1	Fontes de Dados	16
2.1.2	Processo de Coleta	16
2.1.3	Construção de <i>Snapshots</i> Temporais	18
2.1.4	Variável de Disponibilidade dos Atletas	19
2.2	Engenharia de Atributos e Construção das Variáveis	20
2.2.1	Características Individuais	20
2.2.1.1	Variabilidade recente e regularização por posição	21
2.3	Controle de Vazamento de Informação	22
2.4	Organização da Base Final	23
2.5	Modelagem Preditiva	23
2.5.1	Divisão temporal das bases de treinamento e teste	25
2.5.2	Random Forest	25
2.5.3	XGBoost	26
2.5.4	Ensemble de Modelos	27
2.5.5	Modelo Baseline	28
2.5.6	Indicadores derivados para ranqueamento e otimização	29
2.5.7	Seleção Otimizada da Equipe	30
2.5.8	Seleção do Capitão	31

2.5.9	Avaliação dos Modelos	33
3	Resultados e Discussões	36
3.1	Avaliação dos Modelos Preditivos	36
3.2	Importância das variáveis	39
3.3	Avaliação da capacidade de ranqueamento dos atletas	45
3.4	Avaliação da estratégia de escalação otimizada	48
3.5	Avaliação da escolha do capitão	50
3.6	Estudo de caso	54
3.6.1	Rodada 21: cenário de elevado desempenho	54
3.6.2	Rodada 9: cenário de desempenho abaixo do esperado	55
3.6.3	Comparação entre os dois cenários	57
3.7	Discussão geral	58
4	Considerações finais	64
	Referências	67

1 Introdução

A crescente disponibilidade de dados esportivos estruturados e a popularização de competições virtuais baseadas em desempenho real têm impulsionado o desenvolvimento de sistemas de apoio à decisão fundamentados em técnicas estatísticas e de aprendizado de máquina. Nesse contexto, os jogos do tipo *Fantasy Sports* constituem um ambiente particularmente interessante para aplicações de Ciência de Dados, pois combinam previsão de eventos futuros, tomada de decisão sob incerteza e problemas de otimização com restrições. Entre esses jogos, o Cartola FC destaca-se como a principal plataforma de fantasy football do futebol brasileiro, reunindo milhões de usuários que competem semanalmente por meio da seleção de equipes formadas por atletas reais do Campeonato Brasileiro.

O funcionamento do Cartola FC baseia-se na premissa de que a pontuação obtida por cada atleta depende de seu desempenho efetivo em campo durante uma rodada. Assim, os participantes precisam selecionar, antes do fechamento do mercado, um conjunto de jogadores respeitando restrições de orçamento, formação tática e composição da equipe. A qualidade das escolhas realizadas depende diretamente da capacidade de antecipar o desempenho futuro dos atletas, tornando a previsão de pontuações um problema central para o sucesso dos participantes.

Sob a perspectiva da Ciência de Dados, esse problema pode ser formulado como uma tarefa de regressão supervisionada, em que o objetivo consiste em estimar a pontuação futura de um jogador a partir de informações históricas disponíveis antes da realização da rodada. Entretanto, diversos fatores tornam essa tarefa particularmente desafiadora. O desempenho esportivo apresenta elevada variabilidade, sendo influenciado por aspectos individuais, características da equipe, qualidade do adversário, mando de campo, condições físicas, suspensões, lesões e fatores contextuais frequentemente difíceis de quantificar. Além disso, a natureza dinâmica do futebol faz com que relações observadas em rodadas anteriores possam sofrer alterações ao longo da temporada.

A relevância desse problema tem motivado o desenvolvimento de diversas pesqui-

sas internacionais voltadas à previsão de desempenho em jogos do tipo Fantasy Premier League (FPL), principal *fantasy football* associado ao campeonato inglês. Um dos primeiros trabalhos de destaque foi apresentado por Matthews, Ramchurn e Chalkiadakis (2012), que modelaram a formação de equipes como um problema de decisão em ambientes parcialmente observáveis. Posteriormente, Stolyarov e Vasiliev (2017) propuseram uma abordagem baseada em aprendizado de máquina para previsão de desempenho e seleção sequencial de jogadores, estabelecendo uma das primeiras aplicações sistemáticas de técnicas preditivas ao contexto dos *fantasy sports*.

Entre os trabalhos mais relevantes na área, Gupta (2019) investigou a utilização de modelos de séries temporais para previsão de pontuação na *Fantasy Premier League*, demonstrando que informações históricas recentes dos atletas possuem capacidade preditiva significativa. Em uma direção complementar, Bonello et al. (2019) propuseram uma abordagem baseada na integração de múltiplas fontes de dados, combinando estatísticas esportivas e informações contextuais para melhorar a capacidade de previsão dos modelos. Esses estudos contribuíram para consolidar a importância da engenharia de atributos e da incorporação de diferentes horizontes temporais como componentes fundamentais em sistemas de previsão esportiva.

A evolução dos métodos de aprendizado de máquina também teve papel decisivo no avanço dessa área de pesquisa. O trabalho seminal de Breiman (2001) introduziu o algoritmo Random Forest, um método de ensemble baseado em múltiplas árvores de decisão que reduz a variância das previsões e aumenta a capacidade de generalização dos modelos. Posteriormente, Chen e Guestrin (2016) desenvolveram o algoritmo XGBoost (*Extreme Gradient Boosting*), considerado um dos métodos mais eficientes para problemas de regressão e classificação em bases tabulares. Ambos os algoritmos passaram a ser amplamente utilizados em aplicações esportivas devido à sua capacidade de modelar relações não lineares, interações complexas entre variáveis e estruturas de dados heterogêneas.

Outro aspecto relevante diz respeito ao ajuste dos hiperparâmetros dos modelos. Embora métodos como Random Forest e XGBoost apresentem elevado potencial preditivo, seu desempenho depende fortemente da escolha adequada de parâmetros de treinamento. Nesse contexto, Akiba et al. (2019) desenvolveram o framework Optuna, que automatiza processos de busca e otimização de hiperparâmetros, permitindo a obtenção de modelos mais robustos e eficientes. Atualmente, técnicas de otimização automática tornaram-se componentes essenciais em aplicações modernas de aprendizado de máquina.

Mais recentemente, Groos (2025) apresentou o OpenFPL, considerado um dos traba-

lhos mais relevantes da literatura contemporânea sobre previsão em *fantasy football*. O estudo demonstrou que modelos abertos, construídos exclusivamente com dados públicos, podem alcançar desempenho comparável ao de serviços comerciais especializados. A metodologia proposta baseia-se em modelos específicos por posição, utilização de informações históricas do jogador, da equipe e do adversário, além da combinação de previsões provenientes de algoritmos Random Forest e XGBoost por meio de estratégias de ensemble. Outro aspecto importante do OpenFPL é a incorporação explícita da disponibilidade dos atletas antes das partidas, reconhecendo que lesões, suspensões e incertezas de escalação constituem fontes relevantes de risco para as previsões. Os resultados obtidos evidenciaram que modelos transparentes e reproduzíveis podem competir com soluções proprietárias de alto desempenho.

Paralelamente ao problema de previsão, a literatura também tem dedicado atenção ao problema de seleção ótima de equipes. Nesse contexto, Venter e Vuuren (2024) formularam a escolha de jogadores na Fantasy Premier League como um problema de otimização matemática sujeito a restrições de orçamento e composição do elenco. Os autores demonstraram que técnicas de programação linear inteira podem ser utilizadas para transformar previsões individuais em decisões práticas de escalação, produzindo equipes com maior potencial de pontuação. Essa linha de pesquisa estabelece uma importante conexão entre aprendizado de máquina e pesquisa operacional, permitindo a construção de sistemas completos de suporte à decisão.

Apesar dos avanços observados na literatura internacional, ainda existe uma carência de estudos voltados especificamente ao contexto do Cartola FC. Grande parte dos trabalhos concentra-se na Fantasy Premier League, cuja estrutura de pontuação, mercado e disponibilidade de dados difere parcialmente daquela observada no campeonato brasileiro. Além disso, poucos estudos integram simultaneamente técnicas modernas de aprendizado de máquina, informações de disponibilidade dos atletas e métodos de otimização para seleção de equipes.

Diante desse cenário, surge a seguinte questão de pesquisa:

É possível construir um sistema baseado exclusivamente em dados públicos que seja capaz de prever, com precisão satisfatória, a pontuação futura dos jogadores do Cartola FC e utilizar essas previsões para selecionar automaticamente equipes com elevado potencial de desempenho?

Para responder a essa questão, esta pesquisa propõe o desenvolvimento de um sis-

tema integrado de previsão e otimização para o Cartola FC. Inicialmente, são coletados e organizados dados históricos dos jogadores e equipes do Campeonato Brasileiro. Em seguida, são construídos modelos preditivos específicos por posição utilizando técnicas de aprendizado de máquina, com destaque para algoritmos baseados em árvores, ensembles e procedimentos de otimização de hiperparâmetros. Posteriormente, as previsões geradas são incorporadas a um modelo de otimização inteira responsável por selecionar a escalação ideal, respeitando as regras do jogo.

A principal contribuição deste trabalho consiste na adaptação e integração de metodologias consolidadas na literatura internacional ao contexto do Cartola FC, incorporando informações de disponibilidade dos atletas, previsões probabilísticas e mecanismos de otimização de escalação. Espera-se que os resultados obtidos contribuam tanto para a compreensão do potencial das técnicas de Ciência de Dados aplicadas ao esporte quanto para o desenvolvimento de ferramentas transparentes e reproduzíveis de apoio à decisão em fantasy sports.

2 Aspectos Metodológicos

2.1 Coleta e Organização dos Dados

2.1.1 Fontes de Dados

A construção da base de dados utilizada neste trabalho foi realizada exclusivamente a partir de fontes públicas disponibilizadas pelo ecossistema do Cartola FC. A principal fonte de informação foi a API pública do Cartola, que fornece dados atualizados sobre atletas, clubes, rodadas, partidas e mercado de transferências. Essa API disponibiliza informações relevantes para modelagem preditiva, incluindo pontuações históricas, preços dos atletas, médias de desempenho, posição em campo, status de disponibilidade e demais atributos relacionados ao jogo.

Além dos dados disponibilizados diretamente pela API, foram utilizados registros históricos armazenados ao longo da temporada por meio de coletas periódicas automatizadas. Essa estratégia permitiu a construção de uma base longitudinal contendo a evolução temporal das características dos jogadores ao longo das rodadas do Campeonato Brasileiro.

A utilização de fontes públicas apresenta duas vantagens principais. Primeiramente, garante a reprodutibilidade da pesquisa, permitindo que outros pesquisadores obtenham os mesmos dados e reproduzam os resultados. Em segundo lugar, elimina a dependência de serviços pagos ou bases proprietárias, aproximando a metodologia da realidade dos usuários comuns do Cartola FC.

2.1.2 Processo de Coleta

A coleta dos dados foi implementada em linguagem Python por meio de requisições HTTP aos endpoints públicos disponibilizados pela API do Cartola FC. As informações foram obtidas em formato JSON e posteriormente convertidas para estruturas tabulares utilizando a biblioteca Pandas. Foram coletados dados relativos às temporadas de 2022,

2023, 2024, 2025 e 2026 do *fantasy game*.

Uma *Application Programming Interface* (API) pode ser definida como um conjunto de protocolos e regras que permite a comunicação estruturada entre diferentes sistemas computacionais. Em aplicações de Ciência de Dados, as APIs são amplamente utilizadas para automatizar o acesso a bases de dados, serviços web e plataformas digitais, permitindo a obtenção de informações atualizadas de forma padronizada e reproduzível. No contexto deste trabalho, a API pública do Cartola FC foi utilizada como principal fonte de dados, possibilitando a extração automática de informações sobre atletas, clubes, rodadas, preços de mercado, pontuações e status de disponibilidade. As requisições realizadas à API retornam os dados em formato JSON (*JavaScript Object Notation*), que posteriormente são convertidos para estruturas tabulares e integrados à base analítica utilizada nos modelos preditivos. Essa abordagem elimina a necessidade de coleta manual de dados e garante maior consistência, escalabilidade e reprodutibilidade ao processo de construção da base de dados.

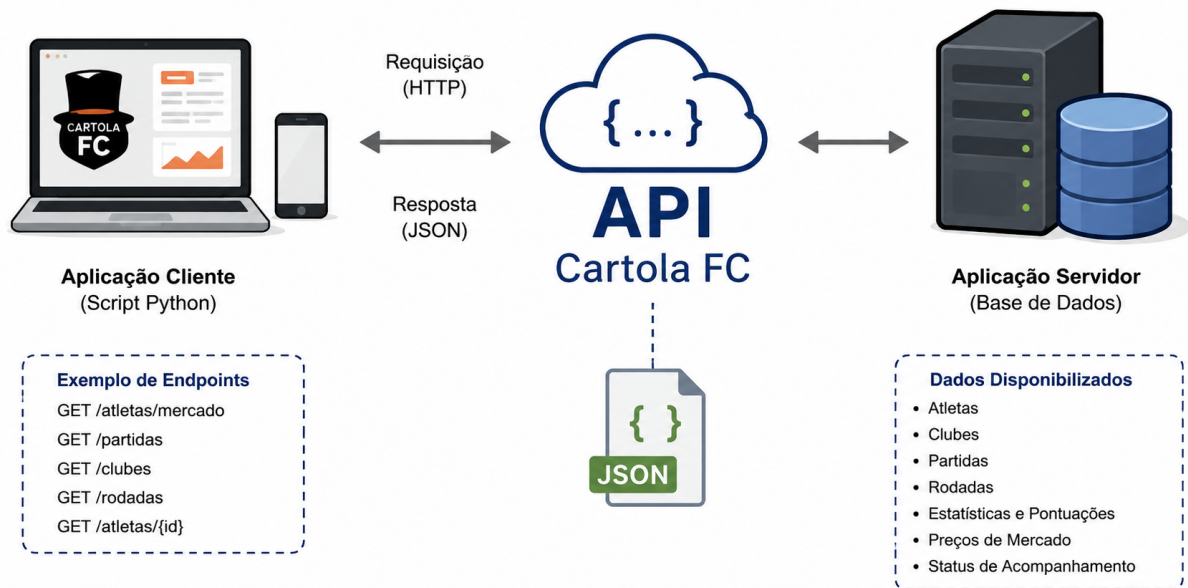


Figura 1: Representação esquemática do processo de comunicação entre um script desenvolvido em Python e a API do Cartola FC. O sistema realiza requisições HTTP aos endpoints da API, recebendo respostas em formato JSON contendo informações sobre atletas, clubes, partidas, rodadas e estatísticas utilizadas na construção da base analítica.

A Figura 1 apresenta o fluxo de comunicação entre o script de coleta e a API do Cartola FC, destacando o processo de requisição, resposta e transformação dos dados em informações utilizáveis pelos modelos de aprendizado de máquina.

Durante cada execução do processo de coleta, foram obtidos dados referentes a:

- atletas disponíveis no mercado;
- clubes participantes do campeonato;
- posições dos jogadores;
- status de disponibilidade;
- rodadas disputadas;
- partidas realizadas;
- pontuações individuais;
- preços de mercado;
- médias de pontuação.

Após a extração dos dados, foi realizada a padronização dos identificadores de atletas, clubes e posições, permitindo a integração das diferentes tabelas em uma única base analítica.

2.1.3 Construção de *Snapshots* Temporais

Um aspecto fundamental desta pesquisa foi a utilização de *snapshots* temporais do mercado.

Um *snapshot* pode ser definido como uma fotografia do estado do sistema em um determinado instante. Cada snapshot armazena todas as informações disponíveis sobre os jogadores antes do início de uma rodada específica.

Por exemplo, considere um atleta que apresente o status "Dúvida" durante a semana e seja posteriormente confirmado como "Provável" poucas horas antes do fechamento do mercado. Caso apenas a informação final fosse armazenada, perder-se-ia o histórico de incerteza existente durante o processo de tomada de decisão.

Dessa forma, foram armazenados *snapshots* periódicos contendo:

- preço do jogador;

- média de pontuação;
- clube;
- posição;
- status de disponibilidade;
- demais atributos disponíveis na API.

A utilização de *snapshots* garante que as previsões sejam construídas exclusivamente com informações conhecidas antes da realização da rodada.

2.1.4 Variável de Disponibilidade dos Atletas

Além das estatísticas tradicionais dos jogadores, foi incorporada uma variável relacionada à probabilidade de participação em uma rodada.

A API do Cartola disponibiliza categorias de status como:

Tabela 1: Categorias de disponibilidade dos atletas fornecidas pela API do Cartola FC e interpretação utilizada para estimar a probabilidade de participação dos jogadores nas rodadas.

Status	Interpretação
Provável	Deve atuar normalmente
Dúvida	Participação incerta
Suspenso	Não poderá atuar
Contundido	Lesionado
Nulo	Sem informação

As categorias apresentadas na Tabela 1 foram transformadas em uma variável quantitativa denominada `prob_jogar`, representando a probabilidade estimada de o atleta efetivamente entrar em campo, conforme mostrado na Tabela 2.

Tabela 2: Conversão dos status de disponibilidade fornecidos pela API do Cartola FC em probabilidades de participação utilizadas na modelagem preditiva.

Status	Probabilidade
Provável	0,90
Dúvida	0,60
Nulo	0,30
Contundido	0,05
Suspenso	0,00

Formalmente, a variável `prob_jogar_i` representa a probabilidade estimada de o atleta (i) participar da rodada.

A inclusão de informações relacionadas à disponibilidade dos atletas foi inspirada no trabalho OpenFPL (Groos, 2025), que utiliza o status pré-partida fornecido pela API da Fantasy Premier League como variável explicativa dos modelos preditivos. Como a API do Cartola FC apresenta categorias distintas, os status dos jogadores foram convertidos em valores entre zero e um, buscando representar diferentes níveis estimados de possibilidade de participação. Essa conversão constitui uma adaptação proposta neste trabalho para o contexto do Cartola FC.

Assim, além de prever o desempenho esperado de cada jogador, o modelo também considera o risco associado à sua participação efetiva na partida.

2.2 Engenharia de Atributos e Construção das Variáveis

Após a etapa de coleta e organização dos dados, foi realizada a construção das variáveis utilizadas pelos modelos preditivos. Essa etapa, frequentemente denominada engenharia de atributos (*feature engineering*), constitui um dos componentes mais importantes de projetos de Ciência de Dados, uma vez que a qualidade das informações fornecidas aos algoritmos influencia diretamente a capacidade de generalização dos modelos. No contexto do Cartola FC, o objetivo principal consiste em prever a pontuação futura dos jogadores utilizando apenas informações disponíveis antes do início da rodada. Dessa forma, as variáveis construídas devem representar adequadamente o histórico de desempenho dos atletas, seu valor de mercado, sua condição de participação e os riscos associados à sua utilização. A literatura recente em *fantasy football* demonstra que a utilização de atributos derivados frequentemente produz ganhos significativos de desempenho em relação ao uso exclusivo das estatísticas brutas. Trabalhos como Gupta (2019), Bonello et al. (2019) e Groos (2025) destacam a importância de combinar informações históricas, contexto competitivo e disponibilidade dos atletas para melhorar a qualidade das previsões.

2.2.1 Características Individuais

As características individuais têm como objetivo representar o histórico de desempenho esportivo dos atletas ao longo da competição. A hipótese subjacente é que o com-

portamento observado em rodadas anteriores contém informações relevantes sobre o desempenho futuro dos jogadores, constituindo uma importante fonte de informação para os modelos preditivos.

Entre as variáveis utilizadas, destaca-se a média histórica de pontuação, que busca sintetizar o desempenho acumulado do atleta ao longo da temporada. Essa variável é calculada como:

$$\text{Media}_i = \frac{\sum_{j=1}^n P_{ij}}{n}$$

em que P_{ij} representa a pontuação obtida pelo jogador i na rodada j , enquanto n corresponde ao número total de partidas disputadas pelo atleta até o momento da análise.

A média histórica constitui uma medida de tendência central capaz de resumir o desempenho geral do jogador ao longo da competição, reduzindo a influência de oscilações observadas em partidas isoladas. Dessa forma, atletas que apresentam desempenho consistentemente elevado tendem a possuir valores médios superiores, refletindo maior potencial de pontuação em rodadas futuras.

Além do histórico de pontuação, também foram incorporadas informações relacionadas à posição ocupada pelo atleta em campo, classificadas nas categorias goleiro, lateral, zagueiro, meio-campista, atacante e técnico. A inclusão dessa variável é particularmente importante porque os critérios de pontuação adotados pelo Cartola FC variam substancialmente entre posições. Enquanto atacantes e meio-campistas são fortemente influenciados por gols e assistências, goleiros e defensores dependem principalmente de ações defensivas, como defesas realizadas e saldo de gols. Consequentemente, espera-se que diferentes posições apresentem distribuições de pontuação distintas, justificando sua utilização como atributo relevante na modelagem.

2.2.1.1 Variabilidade recente e regularização por posição

Além das variáveis originalmente disponibilizadas pela API do Cartola FC, foram construídas variáveis derivadas com o objetivo de representar diferentes aspectos do desempenho esperado dos atletas. Seja x_{ij} a pontuação obtida pelo atleta na j -ésima das cinco últimas partidas.

Inicialmente, foi definida a variável *consistência*, destinada a mensurar a regularidade do desempenho recente dos atletas. Para isso, calculou-se o desvio-padrão das cinco

últimas pontuações,

$$s_i = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2}, \quad n = 5, \quad (2.1)$$

em que

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ij} \quad (2.2)$$

representa a média das cinco últimas pontuações do jogador. Posteriormente, essa medida foi regularizada (*shrinkage*) em relação aos atletas da mesma posição, produzindo o indicador de consistência

$$C_i = \lambda s_i + (1 - \lambda) \bar{s}_{g(i)}, \quad (2.3)$$

em que $\bar{s}_{g(i)}$ representa o desvio-padrão médio observado para os jogadores pertencentes à mesma posição do atleta i e λ corresponde ao parâmetro de regularização. Esse procedimento reduz a influência de valores extremos e de atletas com histórico recente limitado, produzindo estimativas mais estáveis da variabilidade individual.

2.3 Controle de Vazamento de Informação

Um dos principais desafios em problemas de previsão esportiva é evitar o chamado vazamento de informação (*data leakage*).

O vazamento ocorre quando informações futuras são utilizadas durante o treinamento ou avaliação do modelo. Nesse cenário, os resultados obtidos podem parecer artificialmente melhores do que aqueles observados em situações reais.

Para evitar esse problema, todas as variáveis utilizadas para prever a rodada (t) foram construídas apenas com informações disponíveis até a rodada ($t-1$).

De forma semelhante, os atributos relacionados à disponibilidade dos atletas foram obtidos a partir dos *snapshots* coletados antes do fechamento do mercado da rodada analisada.

Essa estratégia garante que o modelo opere em condições equivalentes às enfrentadas por um participante real do Cartola FC.

2.4 Organização da Base Final

Após a etapa de coleta e integração das informações, foi construída uma base longitudinal em que cada linha representa um atleta em uma determinada rodada do campeonato.

A estrutura final contém informações referentes a:

- identificação do atleta;
- clube;
- posição;
- rodada;
- preço;
- média de pontuação;
- pontuação observada;
- status de disponibilidade;
- probabilidade de participação;
- variáveis históricas derivadas.

Essa organização permite a construção de modelos preditivos supervisionados, nos quais as características observadas em rodadas anteriores são utilizadas para estimar a pontuação futura dos atletas.

Além disso, a estrutura resultante possibilita análises retrospectivas (*backtesting*), permitindo avaliar o desempenho do modelo em cenários históricos e medir seu potencial de aplicação prática na montagem de equipes do Cartola FC.

2.5 Modelagem Preditiva

A presente metodologia tem como objetivo construir um sistema preditivo para estimar a pontuação futura de jogadores no Cartola FC e, posteriormente, utilizar essas previsões para auxiliar a montagem de uma escalação otimizada. A proposta segue uma

abordagem semelhante à literatura recente sobre previsão em *fantasy football*, especialmente o estudo OpenFPL, que demonstra que modelos abertos, baseados em dados públicos, podem atingir desempenho competitivo quando combinam histórico de jogadores, informações das equipes, características dos adversários, status pré-rodada e modelos de aprendizado de máquina por posição.

O problema é tratado como uma tarefa de regressão supervisionada. Para cada atleta em uma rodada, busca-se prever sua pontuação na rodada seguinte. Assim, a variável resposta é definida como a pontuação futura do jogador, enquanto as variáveis explicativas são construídas a partir de informações disponíveis antes do início da rodada, como preço, média histórica, posição, clube, status, desempenho recente e demais atributos esportivos disponibilizados pelas fontes de dados. Essa estrutura evita que o modelo utilize informações posteriores à rodada prevista, reduzindo o risco de vazamento de informação.

A coleta dos dados é realizada por meio de consultas periódicas às bases públicas relacionadas ao Cartola FC. Para preservar a coerência temporal do estudo, são armazenados *snapshots* do mercado, isto é, registros do estado dos jogadores em determinado momento antes da rodada. Cada *snapshot* representa uma fotografia do mercado, contendo preço, status, clube, posição e demais informações conhecidas naquele instante. Essa estratégia é importante porque variáveis como status do jogador podem mudar durante a semana. Portanto, para prever a rodada (t), deve-se utilizar apenas o status observado antes do fechamento do mercado da rodada (t), e não informações atualizadas após a realização dos jogos.

A base final é organizada de modo longitudinal, em que cada linha representa um jogador em uma determinada rodada. A partir dessa estrutura, são criadas variáveis defasadas e agregadas, como médias recentes de pontuação, preço e desempenho. Essa decisão segue a lógica de estudos de previsão em *fantasy football*, nos quais o histórico do jogador, o contexto da equipe e as características do confronto são utilizados como sinais preditivos. Além disso, as variáveis são tratadas de maneira distinta conforme a posição do atleta, pois goleiros, defensores, meio-campistas, atacantes e técnicos possuem mecanismos diferentes de pontuação.

O modelo é ajustado separadamente por posição. Essa escolha metodológica é justificada pelo fato de que a distribuição dos pontos e a relevância dos atributos variam entre posições. Por exemplo, variáveis associadas a gols e finalizações tendem a ser mais importantes para atacantes e meio-campistas, enquanto variáveis defensivas podem ter maior relevância para goleiros, zagueiros e laterais. Essa estratégia também é adotada no

OpenFPL, no qual os modelos são treinados por posição para capturar padrões específicos de desempenho.

Foram utilizados modelos de aprendizado de máquina baseados em árvores, especialmente Random Forest e XGBoost.

2.5.1 Divisão temporal das bases de treinamento e teste

Para preservar a ordem temporal dos dados e avaliar a capacidade de generalização dos modelos em observações futuras, foi adotada uma divisão temporal entre as bases de treinamento e teste. Os dados das temporadas de 2022, 2023 e 2024 foram utilizados para o ajuste dos modelos, enquanto a temporada de 2025 foi reservada integralmente para teste e avaliação retrospectiva das estratégias propostas.

Essa divisão garante que os modelos sejam avaliados em uma temporada posterior àquelas utilizadas durante o treinamento, reproduzindo uma situação mais próxima da aplicação real. Dessa forma, nenhuma observação da temporada de 2025 foi utilizada no ajuste inicial dos modelos ou na seleção de seus hiperparâmetros.

Os dados de 2026 foram coletados para atualização e aplicação posterior do sistema, não fazendo parte dos resultados retrospectivos apresentados neste trabalho.

2.5.2 Random Forest

O algoritmo Random Forest, proposto por Breiman (2001), é um método de aprendizado de máquina baseado na combinação de múltiplas árvores de decisão. Em vez de ajustar uma única árvore aos dados, o método constrói um conjunto de árvores treinadas a partir de diferentes amostras da base original. Essas amostras são obtidas por meio de reamostragem com reposição, procedimento conhecido como **bootstrap**. Além disso, em cada divisão das árvores, apenas um subconjunto aleatório das variáveis explicativas é considerado, o que reduz a correlação entre as árvores e aumenta a diversidade do conjunto.

No contexto deste trabalho, o Random Forest foi utilizado como modelo de regressão para estimar a pontuação futura dos atletas do Cartola FC. Cada árvore produz uma previsão individual para a pontuação esperada de um jogador, e a predição final do modelo é obtida pela média das previsões geradas por todas as árvores. Essa estratégia tende a reduzir a variância do modelo e torna o método mais robusto a ruídos e oscilações presentes

nos dados esportivos.

A principal vantagem do Random Forest está em sua capacidade de capturar relações não lineares e interações complexas entre variáveis, como preço, média histórica, posição, status do jogador e probabilidade de participação. Além disso, por combinar várias árvores, o método geralmente apresenta melhor capacidade de generalização do que uma árvore de decisão isolada.

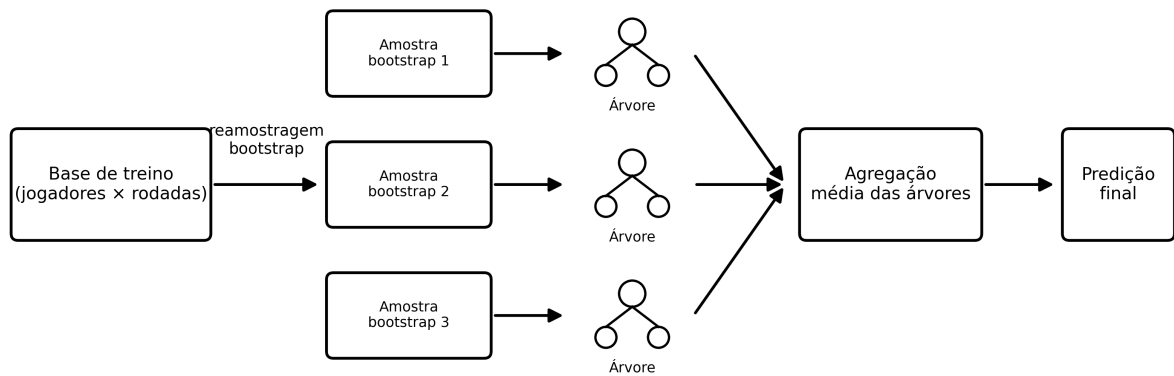


Figura 2: Representação esquemática do algoritmo Random Forest. O método constrói múltiplas árvores de decisão a partir de amostras bootstrap da base de treino e combina suas previsões por meio da média, reduzindo a variância e aumentando a robustez do modelo.

A estrutura geral do algoritmo Random Forest utilizada neste trabalho é apresentada na Figura 2. Observa-se que múltiplas árvores são construídas a partir de diferentes amostras bootstrap da base de treinamento e, posteriormente, suas previsões são combinadas por meio de uma operação de agregação. Esse mecanismo reduz a variabilidade das previsões individuais e aumenta a capacidade de generalização do modelo.

2.5.3 XGBoost

O XGBoost, proposto por Chen e Guestrin (2016), é um algoritmo de aprendizado de máquina pertencente à família dos métodos de gradient boosting. Diferentemente do Random Forest, em que as árvores são construídas de forma aproximadamente independente, o XGBoost ajusta árvores de decisão de maneira sequencial. Cada nova árvore é treinada para corrigir os erros cometidos pelo conjunto de árvores anteriores.

No problema de previsão da pontuação dos jogadores do Cartola FC, o XGBoost é utilizado como modelo de regressão. Inicialmente, o algoritmo produz uma predição aproximada para a pontuação futura. Em seguida, calcula os erros entre os valores observados e previstos, ajustando novas árvores para reduzir esses resíduos. Ao final do processo, a

previsão final é obtida pela soma ponderada das contribuições de todas as árvores ajustadas.

Uma característica importante do XGBoost é a presença de mecanismos de regularização, que ajudam a controlar a complexidade do modelo e reduzem o risco de sobreajuste. Além disso, o algoritmo permite ajustar hiperparâmetros como número de árvores, profundidade máxima, taxa de aprendizado e proporção de variáveis utilizadas em cada etapa. Essa flexibilidade torna o XGBoost especialmente adequado para dados tabulares com padrões não lineares, como aqueles observados em *fantasy sports*.

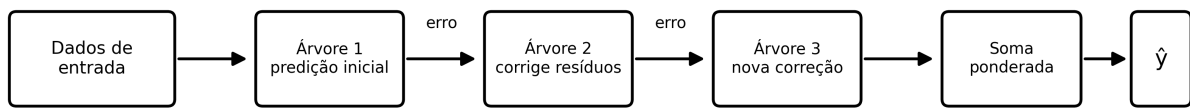


Figura 3: Representação esquemática do algoritmo XGBoost. As árvores são ajustadas sequencialmente, de modo que cada nova árvore busca corrigir os erros residuais das previsões anteriores, produzindo ao final uma previsão agregada e regularizada.

A Figura 3 ilustra o processo iterativo de construção do XGBoost. Diferentemente do Random Forest, as árvores não são ajustadas de forma independente, mas sim de maneira sequencial, em que cada nova árvore busca minimizar os erros residuais produzidos pelo conjunto de modelos já construídos. Essa estratégia permite capturar padrões complexos presentes nos dados e aumentar significativamente o desempenho preditivo.

2.5.4 Ensemble de Modelos

As previsões produzidas pelos modelos Random Forest e XGBoost foram combinadas por meio de um ensemble, isto é, pela agregação das estimativas geradas pelos dois modelos. Essa combinação busca reduzir a variabilidade das previsões e aumentar a estabilidade do sistema. No script desenvolvido, a variável `pred_ensemble` representa a previsão consolidada da pontuação esperada do jogador.

$$\text{pred_ensemble}_i = \frac{\text{PredRF}_i + \text{PredXGB}_i}{2}. \quad (2.4)$$

em que PredRF_i e PredXGB_i representam, respectivamente, as pontuações previstas para o jogador i pelos modelos Random Forest e XGBoost.

A variável `pred_ensemble` corresponde à previsão consolidada utilizada na avaliação do modelo ensemble. Sua qualidade preditiva foi avaliada por meio da comparação com a

pontuação real observada na rodada, utilizando métricas como MAE, RMSE e correlação de Pearson.

Além da previsão média, foi considerada uma variável de risco operacional denominada **prob_jogar**. Essa variável representa uma aproximação da probabilidade de o atleta efetivamente participar da rodada, construída a partir do status registrado antes do fechamento do mercado. Por exemplo, jogadores com status “Provável” recebem maior probabilidade de participação, enquanto atletas suspensos, contundidos ou em dúvida recebem valores menores.

A previsão ajustada pela probabilidade de participação é definida por:

$$\text{pred_ajustada_prob}_i = \text{pred_ensemble}_i \times \text{prob_jogar}_i. \quad (2.5)$$

Essa variável não substitui a previsão utilizada na avaliação estatística do ensemble, mas constitui um indicador voltado à tomada de decisão na montagem da equipe, pois penaliza atletas com maior risco de não participar da rodada.

2.5.5 Modelo Baseline

Para avaliar adequadamente o desempenho dos modelos de aprendizado de máquina desenvolvidos neste trabalho, foi adotado um modelo **baseline**, cuja finalidade consiste em fornecer uma referência simples de comparação para as abordagens mais sofisticadas. A utilização de um baseline é uma prática consolidada em problemas de modelagem preditiva, pois permite verificar se o ganho obtido por algoritmos mais complexos decorre efetivamente da capacidade de aprendizagem dos modelos ou se resultados semelhantes poderiam ser alcançados por estratégias mais simples.

Neste trabalho, o baseline foi construído utilizando a **média histórica de pontuação** de cada atleta até a rodada imediatamente anterior àquela que se deseja prever. Dessa forma, para cada jogador, a previsão da pontuação na rodada (t) corresponde simplesmente à média das pontuações observadas nas rodadas anteriores, sem considerar outras características como preço, posição, clube, adversário, status de disponibilidade ou variáveis derivadas. Formalmente, a previsão do modelo baseline é dada por

$$\hat{Y}_{i,t}^{(base)} = \frac{1}{n_i} \sum_{j=1}^{n_i} P_{ij},$$

em que $\hat{Y}_{i,t}^{(base)}$ representa a pontuação prevista para o atleta i na rodada t , P_{ij} corresponde à pontuação obtida pelo atleta i na rodada j e n_i representa o número de partidas disputadas até a rodada imediatamente anterior.

Embora essa abordagem seja bastante simples, ela apresenta uma característica importante: utiliza exclusivamente o comportamento histórico do próprio atleta como estimativa para seu desempenho futuro. Assim, o baseline representa a hipótese de que o melhor preditor da próxima pontuação é o desempenho médio previamente observado. As previsões produzidas pelos modelos Random Forest, XGBoost e pelo ensemble foram comparadas diretamente com esse modelo de referência por meio das métricas de avaliação adotadas na pesquisa. Dessa forma, foi possível quantificar objetivamente o ganho obtido pela incorporação de variáveis explicativas adicionais, pela modelagem de relações não lineares e pela utilização de técnicas modernas de aprendizado de máquina.

2.5.6 Indicadores derivados para ranqueamento e otimização

Para simplificar a notação, seja $\hat{P}_i$ a previsão do ensemble para o jogador i , correspondente à variável `pred_ensemble`, e seja p_i sua probabilidade estimada de participação, correspondente à variável `prob_jogar`.

Assim, a previsão ajustada pela probabilidade de participação, definida computacionalmente na Seção 2.5.4, pode ser representada por

$$P_i = \hat{P}_i p_i. \quad (2.6)$$

Em seguida, foi construída a variável *teto de pontuação*, cujo objetivo é incorporar à expectativa de pontuação uma medida do potencial de desempenho acima da média. Essa variável foi calculada por

$$T_i = P_i + 0.25 C_i. \quad (2.7)$$

Além disso, foi definido um indicador de retorno sobre investimento (*Return on Investment – ROI*), que busca relacionar a pontuação prevista ao custo do atleta em cartoletas. Para reduzir a penalização imposta aos jogadores de maior valor de mercado, utilizou-se a raiz quadrada do preço do atleta, obtendo-se

$$R_i = \frac{\hat{P}_i}{\sqrt{\text{Preço}_i}}. \quad (2.8)$$

Como as variáveis P_i , T_i e R_i apresentam escalas distintas, todas foram previamente normalizadas pelo método *min-max*, segundo

$$\tilde{X}_i = \frac{X_i - \min(X)}{\max(X) - \min(X)}, \quad (2.9)$$

em que X_i representa qualquer uma das variáveis utilizadas na composição do indicador final.

Por fim, foi proposta a variável *score híbrido*, utilizada como critério alternativo para o ranqueamento dos atletas na etapa de otimização da equipe. Esse indicador combina a expectativa de pontuação, o potencial de desempenho máximo e a relação entre desempenho e custo do jogador, sendo calculado por

$$S_i = 0.70 \tilde{P}_i + 0.20 \tilde{T}_i + 0.10 \tilde{R}_i. \quad (2.10)$$

Os pesos atribuídos aos três componentes foram definidos empiricamente, priorizando a capacidade preditiva do modelo (70%), sem desconsiderar o potencial de altas pontuações (20%) e a eficiência econômica do atleta (10%). Dessa forma, o *score híbrido* constitui um indicador multicritério capaz de equilibrar expectativa de desempenho, potencial de explosão e custo-benefício, sendo utilizado como critério complementar para o ranqueamento dos jogadores e posterior seleção da equipe por meio do modelo de programação inteira.

2.5.7 Seleção Otimizada da Equipe

Após a etapa preditiva, o problema de seleção do time é formulado como uma otimização inteira. Cada jogador recebe uma variável binária, indicando se será ou não escolhido para a escalação. O objetivo é maximizar a soma das previsões ajustadas dos jogadores selecionados, respeitando restrições típicas do Cartola FC, como esquema tático, número de jogadores por posição, orçamento máximo e limite de atletas por clube. Essa etapa transforma as previsões individuais em uma decisão prática de escalação.

A função objetivo consiste em maximizar a soma das pontuações previstas dos jogadores selecionados:

$$\max \sum_{i=1}^n x_i \cdot \text{pred_ajustada_prob}_i$$

em que:

$$x_i = \begin{cases} 1, & \text{se o atleta for selecionado} \\ 0, & \text{caso contrário} \end{cases}$$

O modelo está sujeito às seguintes restrições:

- formação tática;
- orçamento máximo;
- limite de atletas por clube;
- número obrigatório de jogadores por posição.

Essa formulação segue a linha proposta por Venter e Vuuren (2024), que demonstraram a eficiência da programação inteira na seleção de equipes de *fantasy football*.

Dessa forma, as previsões geradas pelos modelos de aprendizado de máquina são convertidas em decisões práticas de escalação, permitindo a construção automática de equipes com elevado potencial de pontuação.

2.5.8 Seleção do Capitão

Além da definição da escalação principal, um elemento estratégico fundamental no Cartola FC é a escolha do capitão da equipe. O capitão corresponde a um dos atletas previamente selecionados para a escalação e possui uma característica diferenciada dentro do sistema de pontuação do jogo: sua pontuação final recebe um bônus de 50% em relação aos demais jogadores da equipe. Dessa forma, o desempenho do capitão exerce influência significativa sobre o resultado final obtido pelo cartoleiro em cada rodada.

Formalmente, seja Y_i a pontuação obtida pelo jogador i em uma determinada rodada. Caso esse atleta seja escolhido como capitão, sua contribuição efetiva para a pontuação da equipe passa a ser calculada por:

$$Y_i^{(\text{cap})} = 1,5Y_i, \quad (2.11)$$

em que $Y_i^{(\text{cap})}$ representa a pontuação contabilizada para o atleta escolhido como capitão.

A existência desse mecanismo transforma a escolha do capitão em um problema de decisão distinto da simples seleção dos atletas. Enquanto a etapa de otimização da escalação busca identificar a combinação de jogadores capaz de maximizar a pontuação total esperada da equipe sob restrições de orçamento, formação tática e limite de atletas por clube, a escolha do capitão procura determinar qual dos jogadores selecionados possui maior potencial para gerar um ganho adicional de pontuação.

Do ponto de vista estatístico, a escolha do capitão envolve a avaliação simultânea de expectativa de desempenho e risco associado à previsão. Um atleta pode apresentar elevada média histórica e desempenho consistente ao longo da temporada, tornando-se uma alternativa segura para receber o bônus de pontuação. Por outro lado, alguns jogadores apresentam maior variabilidade de desempenho e podem alcançar pontuações excepcionalmente altas em determinadas rodadas, oferecendo maior potencial de retorno, mas também maior risco.

Nesse contexto, a literatura sobre *fantasy sports* frequentemente diferencia atletas de alta consistência daqueles que possuem elevado potencial de explosão (*ceiling*). Enquanto jogadores consistentes tendem a apresentar desempenho próximo ao valor esperado na maioria das partidas, atletas com alto potencial de explosão possuem maior probabilidade de produzir pontuações muito acima da média em cenários favoráveis. Essa distinção é particularmente relevante para a escolha do capitão, uma vez que o bônus de 50% amplifica tanto os benefícios de uma boa previsão quanto os impactos de uma previsão incorreta.

Neste trabalho, a seleção do capitão foi baseada nas previsões produzidas pelos modelos de aprendizado de máquina desenvolvidos para estimar a pontuação futura dos atletas. Inicialmente, cada jogador recebe uma previsão de pontuação denominada `pred_ensemble`, obtida pela combinação das estimativas geradas pelos modelos Random Forest e XGBoost. Em seguida, essa previsão é ajustada pela probabilidade estimada de participação do atleta na rodada, representada pela variável `prob_jogar`, resultando na medida:

$$\text{pred_ajustada_prob} = \text{pred_ensemble} \times \text{prob_jogar}$$

Essa correção é particularmente importante porque evita que jogadores com elevado potencial de pontuação, mas baixa probabilidade de entrar em campo, sejam selecionados como capitães.

Para a seleção do capitão foi utilizado o teto de pontuação T_i , definido na Seção 2.5.6.

Na implementação computacional, o indicador T_i corresponde à variável `teto_score`. Essa medida combina a previsão ajustada pela probabilidade de participação com uma parcela da variabilidade regularizada do atleta, buscando representar um cenário otimista, porém controlado, de desempenho.

Dessa forma, a escolha do capitão foi tratada como uma etapa específica do processo de otimização. Após a definição da equipe titular, os atletas selecionados são ordenados de acordo com seu potencial de desempenho, sendo escolhido como capitão aquele que apresenta o maior valor da métrica adotada pelo sistema. Quando o objetivo é privilegiar segurança e estabilidade, a seleção pode ser realizada utilizando a variável `pred_ajustada_prob`. Por outro lado, quando se busca maximizar o potencial de pontuação da equipe, a escolha pode ser baseada na variável `teto_score`.

Formalmente, considerando uma estratégia voltada para maximização do potencial de pontuação, o capitão é definido por:

$$\text{Capitão} = \arg \max_{i \in \text{Time}} (T_i) . \quad (2.12)$$

em que `Time` representa o conjunto de atletas selecionados para a escalação da rodada.

Sob a perspectiva da tomada de decisão, a figura do capitão pode ser interpretada como uma alocação diferenciada de importância sobre um único jogador da equipe. Em termos práticos, a escolha correta do capitão frequentemente representa um dos fatores mais relevantes para o sucesso em uma rodada, uma vez que seu desempenho recebe um acréscimo de 50% em relação aos demais atletas. Por esse motivo, a seleção do capitão foi incorporada à metodologia proposta como uma etapa independente da montagem da equipe, buscando maximizar o benefício proporcionado pelo mecanismo de bonificação disponibilizado pelo Cartola FC.

2.5.9 Avaliação dos Modelos

A avaliação da metodologia proposta foi realizada em duas etapas complementares. A primeira teve como objetivo analisar a capacidade preditiva dos modelos de aprendizado de máquina, comparando as pontuações previstas para os atletas com as pontuações efetivamente observadas nas rodadas avaliadas. A segunda etapa consistiu na avaliação da estratégia de seleção da equipe, verificando se as previsões individuais produzidas pelos modelos resultaram na construção de escalações competitivas quando submetidas

ao processo de otimização.

Para avaliar a qualidade das previsões individuais foram utilizadas três métricas amplamente empregadas em problemas de regressão: o Erro Médio Absoluto (*Mean Absolute Error* - MAE), a Raiz do Erro Quadrático Médio (*Root Mean Squared Error* - RMSE) e o coeficiente de correlação de Pearson.

O MAE mede o erro médio absoluto entre as pontuações previstas e observadas, sendo definido por

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|,$$

em que y_i representa a pontuação real do atleta i , $\hat{y}_i$ corresponde à pontuação prevista pelo modelo e n representa o número total de observações avaliadas. Essa métrica fornece uma interpretação direta do erro médio de previsão, sendo expressa na mesma unidade da variável resposta, isto é, pontos do Cartola FC.

Como segunda medida de desempenho foi utilizada a Raiz do Erro Quadrático Médio, definida por

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}.$$

Diferentemente do MAE, o RMSE atribui maior penalização a erros de grande magnitude devido ao uso do erro quadrático. Consequentemente, essa métrica é particularmente útil para avaliar a robustez dos modelos diante de previsões muito distantes dos valores observados.

Além das medidas de erro, foi calculado o coeficiente de correlação de Pearson entre as pontuações previstas e observadas, dado por

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}},$$

em que $\bar{y}$ e $\bar{\hat{y}}$ representam, respectivamente, as médias das pontuações observadas e previstas. Essa medida quantifica o grau de associação linear entre os valores reais e estimados, assumindo valores no intervalo $[-1, 1]$. Quanto mais próximo de 1 for o coeficiente, maior será a capacidade do modelo em preservar a ordenação relativa dos atletas em termos de

desempenho.

As previsões produzidas pelos modelos Random Forest, XGBoost e pelo ensemble foram comparadas com um modelo *baseline* baseado exclusivamente na média histórica de pontuação dos atletas. Essa comparação permitiu quantificar objetivamente o ganho obtido pela utilização de algoritmos de aprendizado de máquina e pela incorporação das variáveis explicativas propostas neste trabalho.

Após a avaliação das previsões individuais, procedeu-se à análise da estratégia de escalação. Nessa etapa, a qualidade do sistema foi avaliada comparando-se a pontuação total prevista para a equipe otimizada com a pontuação efetivamente obtida pelos atletas selecionados. Sempre que possível, essa estratégia também foi comparada com critérios alternativos de seleção, como a escolha dos jogadores de maior média histórica e o modelo *baseline* adotado na pesquisa.

Essa distinção entre avaliação preditiva e avaliação da escalação é fundamental, pois um modelo pode apresentar excelente desempenho na previsão individual das pontuações e, ainda assim, produzir equipes pouco competitivas caso não considere adequadamente as restrições impostas pelo Cartola FC, como orçamento disponível, formação tática, limite de atletas por clube e risco associado à disponibilidade dos jogadores.

Por fim, todas as etapas da metodologia, desde a coleta dos dados até a avaliação das equipes selecionadas, foram implementadas em linguagem Python, garantindo transparência, reprodutibilidade e facilidade de atualização do sistema. Essa estrutura permite que todo o procedimento seja reaplicado em novas rodadas e temporadas, possibilitando a incorporação contínua de novos dados e a reavaliação do desempenho dos modelos ao longo do tempo.

3 Resultados e Discussões

Este capítulo apresenta os resultados obtidos a partir da aplicação da metodologia proposta, bem como sua análise e discussão. Inicialmente, são avaliados os modelos preditivos desenvolvidos, comparando seu desempenho com o modelo *baseline* adotado na pesquisa por meio das métricas de avaliação definidas na metodologia. Em seguida, são analisadas as contribuições das variáveis utilizadas na previsão, investigando sua importância relativa na construção dos modelos.

Posteriormente, são apresentados os resultados do processo de otimização da equipe, avaliando a qualidade das escalações geradas e sua capacidade de selecionar atletas com elevado potencial de pontuação dentro das restrições estabelecidas pelo regulamento do Cartola FC. Sempre que possível, os resultados obtidos são comparados com estratégias alternativas de seleção, permitindo avaliar objetivamente os ganhos proporcionados pela abordagem proposta.

Por fim, os resultados são discutidos à luz dos objetivos da pesquisa, destacando as principais contribuições da metodologia, suas limitações e os aspectos que podem ser aprimorados em trabalhos futuros. Dessa forma, busca-se não apenas verificar a capacidade preditiva dos modelos desenvolvidos, mas também demonstrar sua aplicabilidade prática na construção de equipes competitivas para o *fantasy game*.

3.1 Avaliação dos Modelos Preditivos

A primeira etapa da análise dos resultados consistiu na avaliação da capacidade preditiva dos modelos desenvolvidos. Para isso, foram comparados o modelo *baseline*, o Random Forest, o XGBoost e o modelo *ensemble*, construído a partir da combinação das previsões dos dois algoritmos. Adicionalmente, foi avaliada uma versão do *ensemble* ajustada pela probabilidade de participação do atleta na rodada, denominada *Ensemble ajustado por prob_jogar*. O objetivo dessa análise foi verificar se a utilização de técnicas de aprendizado de máquina, aliada ao conjunto de variáveis elaborado neste trabalho, proporcionou

ganhos efetivos de desempenho em relação a uma estratégia simples baseada apenas na média histórica recente de pontuação dos atletas.

A comparação entre os modelos foi realizada utilizando as métricas apresentadas na metodologia: Erro Médio Absoluto (MAE), Raiz do Erro Quadrático Médio (RMSE) e coeficiente de correlação de Pearson. Enquanto o MAE fornece uma medida direta do erro médio de previsão, o RMSE penaliza com maior intensidade erros de grande magnitude, permitindo avaliar a ocorrência de previsões discrepantes. Já a correlação de Pearson quantifica o grau de associação linear entre as pontuações previstas e observadas, indicando a capacidade dos modelos de diferenciar atletas com maior e menor potencial de desempenho.

Além dessas métricas, foram analisadas a média das pontuações observadas, a média das pontuações previstas e o viés médio de previsão. Neste trabalho, o viés médio foi calculado pela diferença entre a pontuação observada e a pontuação prevista. Assim, valores positivos indicam tendência de subestimação das pontuações reais pelos modelos.

A Tabela 3 apresenta os resultados obtidos para as 7916 observações utilizadas na avaliação.

Tabela 3: Comparação do desempenho dos modelos preditivos em relação ao modelo baseline.

Modelo	MAE	RMSE	Correlação	Redução do MAE (%)
Baseline (últimas 5 rodadas)	3,235	4,425	0,176	–
Random Forest	2,999	4,370	0,248	7,3
XGBoost	2,938	4,255	0,265	9,2
Ensemble	2,963	4,306	0,261	8,4
Ensemble + Prob. de participação	2,997	4,395	0,261	7,3

Fonte: Elaborado pelo autor.

Os resultados evidenciam que o **XGBoost** apresentou o melhor desempenho global entre os modelos analisados, alcançando MAE igual a 2,938, RMSE de 4,255 e correlação de Pearson de 0,265. Esses valores correspondem ao menor erro absoluto médio, à menor ocorrência relativa de erros de grande magnitude e à maior associação entre as pontuações previstas e observadas. Em comparação com o modelo *baseline*, o XGBoost apresentou redução de aproximadamente 9,2% no MAE e de 3,8% no RMSE.

O modelo **Ensemble** apresentou desempenho bastante próximo ao do XGBoost, obtendo MAE de 2,963, RMSE de 4,306 e correlação de 0,261. Esses resultados representam uma redução de aproximadamente 8,4% no MAE em relação ao *baseline*. Embora a com-

binação das previsões tenha produzido resultados competitivos, ela não superou o melhor modelo individual. Esse comportamento indica que a combinação utilizada não acrescentou diversidade preditiva suficiente para reduzir os erros além do desempenho alcançado pelo XGBoost.

O **Random Forest** apresentou MAE de 2,999, RMSE de 4,370 e correlação de 0,248. Apesar de apresentar desempenho ligeiramente inferior ao XGBoost e ao *ensemble*, o modelo ainda superou o *baseline*, com redução aproximada de 7,3% no MAE. Esse resultado demonstra que a combinação de múltiplas variáveis explicativas permitiu capturar informações relevantes para a previsão que não são consideradas por uma abordagem baseada exclusivamente na média das pontuações anteriores.

A versão do *ensemble* ajustada pela probabilidade de participação do atleta apresentou MAE de 2,997, RMSE de 4,395 e correlação de 0,261. Em relação ao *ensemble* sem ajuste, verificou-se aumento tanto no MAE quanto no RMSE, enquanto a correlação permaneceu inalterada. Portanto, a ponderação pela variável `prob_jogar` não melhorou a precisão das previsões individuais segundo as métricas de erro utilizadas.

Esse resultado, entretanto, não implica que o ajuste pela probabilidade de participação seja desnecessário. Sua principal finalidade é reduzir o risco de selecionar atletas com baixa possibilidade de atuar, sendo mais relevante para a etapa de otimização da equipe do que para a minimização direta do erro preditivo. Ao multiplicar a previsão pela probabilidade de participação, o procedimento reduz deliberadamente a pontuação esperada de jogadores com maior risco de ausência, produzindo estimativas mais conservadoras.

O modelo **Baseline Last 5**, baseado exclusivamente na média das cinco últimas pontuações de cada atleta, apresentou o pior desempenho entre os métodos avaliados, com MAE de 3,235, RMSE de 4,425 e correlação de 0,176. Embora represente uma estratégia simples e intuitiva, os resultados mostram que o histórico recente de pontuação, quando utilizado isoladamente, não é suficiente para explicar adequadamente o desempenho futuro dos jogadores.

A correlação obtida pelo XGBoost foi aproximadamente 51% superior à observada no *baseline*, passando de 0,176 para 0,265. Essa diferença indica que o modelo de aprendizado de máquina apresentou maior capacidade de distinguir atletas com diferentes níveis de desempenho. Entretanto, os valores absolutos de correlação permaneceram relativamente baixos, evidenciando a elevada dificuldade inerente à previsão de pontuações individuais no futebol.

A análise das médias também revelou uma tendência sistemática de subestimação. A pontuação média observada foi de 3,798 pontos, enquanto as médias previstas foram de 2,280 para o XGBoost, 2,150 para o *ensemble*, 2,021 para o Random Forest, 1,935 para o *ensemble* ajustado e 3,011 para o *baseline*. Como consequência, todos os métodos apresentaram viés médio positivo, indicando que, em média, as pontuações reais foram superiores às estimativas produzidas.

O XGBoost apresentou viés médio de 1,518 ponto, enquanto o *ensemble* e o Random Forest apresentaram vieses de 1,648 e 1,777 ponto, respectivamente. O ajuste por `prob_jogar` aumentou essa diferença para 1,863 ponto, resultado coerente com o caráter conservador da ponderação pela probabilidade de participação. O *baseline*, embora tenha apresentado os maiores erros globais e a menor correlação, registrou o menor viés médio, igual a 0,787 ponto. Esse resultado mostra que ausência de viés, isoladamente, não representa necessariamente maior capacidade preditiva, pois um modelo pode apresentar média próxima à observada e, ainda assim, cometer erros elevados em nível individual.

De maneira geral, todos os modelos de aprendizado de máquina apresentaram desempenho superior ao *baseline* em termos de MAE, RMSE e correlação. Entre eles, o XGBoost apresentou os melhores resultados e demonstrou maior capacidade de capturar relações entre as variáveis explicativas e a pontuação futura dos atletas. O desempenho próximo do *ensemble* indica que a combinação dos modelos produziu previsões competitivas, embora sem superar o XGBoost isoladamente.

Os resultados também evidenciam que a previsão de pontuações no Cartola FC permanece um problema de elevada incerteza. Eventos como gols, assistências, cartões, substituições, lesões e decisões técnicas podem provocar variações acentuadas nas pontuações e não são completamente antecipados pelas informações disponíveis antes das partidas. Dessa forma, as métricas devem ser interpretadas não apenas pela magnitude absoluta dos erros, mas principalmente pelo ganho obtido em relação ao modelo de referência e pela capacidade do sistema de utilizar essas previsões na construção de escalações competitivas.

3.2 Importância das variáveis

Além da avaliação do desempenho preditivo, torna-se importante compreender quais variáveis exerceram maior influência sobre as previsões produzidas pelos modelos. Essa análise contribui para a interpretação dos algoritmos e permite verificar se os padrões aprendidos são coerentes com a dinâmica do Cartola FC e com o comportamento esperado

do desempenho dos atletas.

Os modelos Random Forest e XGBoost disponibilizam medidas de importância das variáveis, embora utilizem critérios distintos para sua obtenção. No Random Forest, a importância está associada à redução média da impureza proporcionada pelas divisões realizadas ao longo das árvores. No XGBoost, por sua vez, a importância está relacionada ao ganho obtido quando uma variável é utilizada nas divisões das árvores construídas sequencialmente.

Como os modelos foram ajustados separadamente para goleiros, laterais, zagueiros, meias, atacantes e técnicos, a importância final de cada variável foi calculada pela média dos valores obtidos nas seis posições. Também foi calculado o desvio-padrão entre as posições, permitindo avaliar a estabilidade da contribuição de cada atributo. As importâncias devem ser interpretadas separadamente dentro de cada algoritmo, uma vez que as escalas produzidas pelo Random Forest e pelo XGBoost não são diretamente comparáveis.

A Tabela 4 apresenta as principais variáveis identificadas pelos dois modelos. Os valores entre parênteses correspondem ao desvio-padrão da importância entre as seis posições avaliadas.

Tabela 4: Importância média das principais variáveis nos modelos Random Forest e XGBoost.

Variável	Random Forest	XGBoost
Status do atleta	0,2850 (0,0871)	0,4930 (0,0885)
Clube	0,0084 (0,0090)	0,1070 (0,0561)
Pontos — média de 5 rodadas	0,0627 (0,0212)	0,0491 (0,0430)
Pontos — média de 3 rodadas	0,0542 (0,0190)	0,0464 (0,0308)
Pontos — média de 10 rodadas	0,0523 (0,0165)	0,0319 (0,0202)
Consistência	0,0491 (0,0171)	0,0231 (0,0136)
Pontos — média de 38 rodadas	0,0403 (0,0089)	0,0181 (0,0152)
Desvio-padrão — últimas 5 rodadas	0,0375 (0,0126)	0,0100 (0,0032)
Preço atual	0,0326 (0,0072)	0,0205 (0,0093)
Variação de preço	0,0258 (0,0063)	0,0142 (0,0049)

Fonte: Elaborado pelo autor. Os valores representam a importância média, e os valores entre parênteses correspondem ao desvio-padrão entre as seis posições avaliadas.

A Figura 4 apresenta as variáveis originais de maior importância para o modelo Random Forest.

A Figura 5 apresenta os resultados correspondentes ao modelo XGBoost.

Os resultados apresentados na Tabela 4 e nas Figuras 4 e 5 mostram que o status do atleta foi a variável de maior importância nos dois modelos. No Random Forest, sua

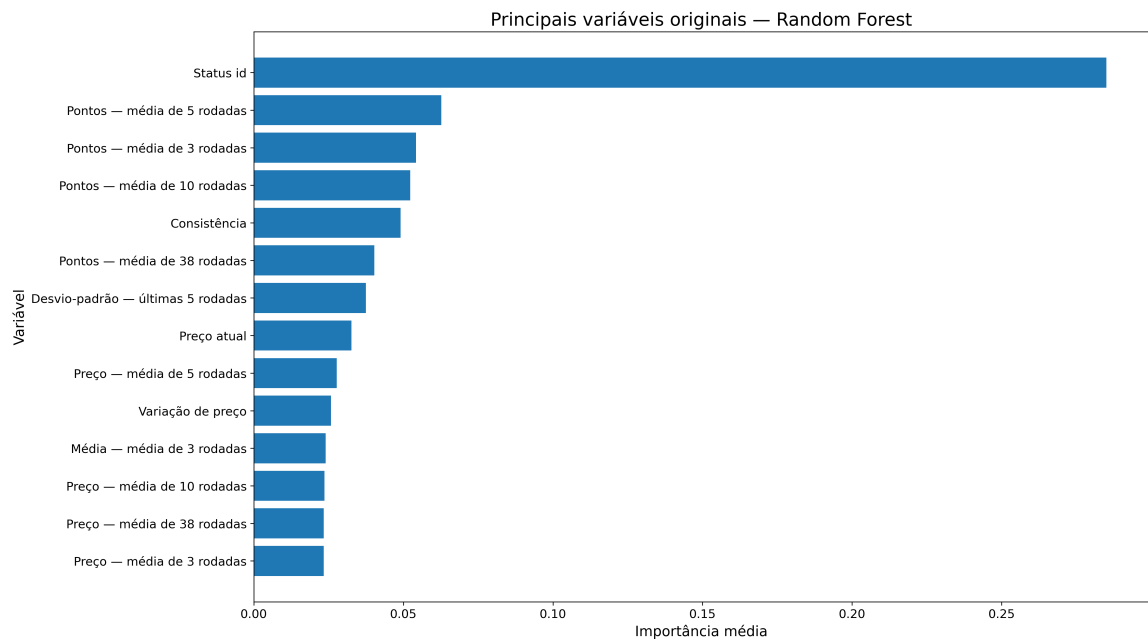


Figura 4: Variáveis originais de maior importância no modelo Random Forest.

Fonte: Elaborado pelo autor.

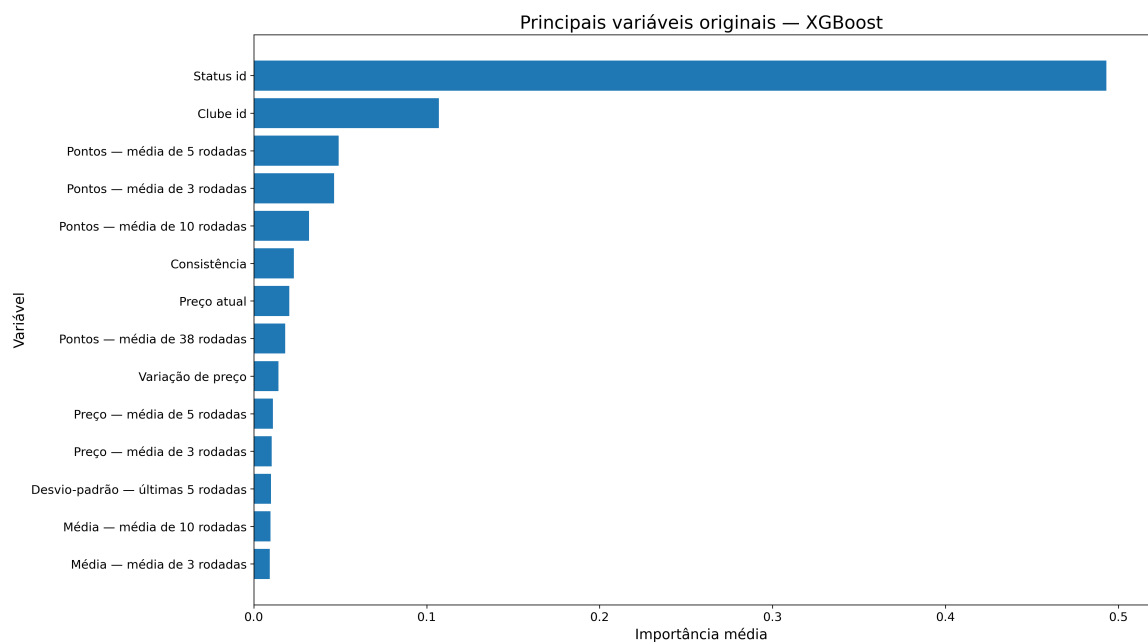


Figura 5: Variáveis originais de maior importância no modelo XGBoost.

Fonte: Elaborado pelo autor.

importância média foi igual a 0,2850, enquanto no XGBoost esse valor alcançou 0,4930. Além disso, essa característica apresentou importância diferente de zero nas seis posições analisadas, indicando que sua contribuição não ficou restrita a um grupo específico de jogadores.

A elevada importância do status é coerente com a estrutura do problema. A participação efetiva do atleta na partida é uma condição necessária para que ele possa pontuar. Assim, informações como provável, contundido, suspenso ou sem status definido ajudam os modelos a diferenciar jogadores disponíveis daqueles que apresentam maior risco de ausência.

No XGBoost, a variável clube apresentou a segunda maior importância, com média igual a 0,1070 e desvio-padrão de 0,0561. O valor máximo dessa importância entre as posições atingiu 0,2017. Esse resultado sugere que o XGBoost explorou de maneira mais intensa padrões associados às equipes, como qualidade coletiva, força ofensiva e defensiva, momento na competição e características específicas do contexto em que o atleta está inserido.

No Random Forest, por outro lado, a importância média da variável clube foi de apenas 0,0084, com desvio-padrão de 0,0090. Portanto, embora a variável tenha apresentado alguma contribuição nas seis posições, sua relevância média foi consideravelmente menor nesse algoritmo. Essa diferença demonstra que os modelos utilizaram as informações disponíveis de formas distintas. Enquanto o XGBoost concentrou parte relevante de sua capacidade preditiva no clube, o Random Forest distribuiu a importância entre um conjunto mais amplo de variáveis numéricas.

Após o status do atleta, as variáveis relacionadas ao histórico de pontuação foram as que apresentaram maior contribuição para o Random Forest. A média de pontos das últimas cinco rodadas alcançou importância média de 0,0627, seguida pela média das últimas três rodadas, com 0,0542, e pela média das últimas dez rodadas, com 0,0523.

No XGBoost, essas variáveis também figuraram entre as mais relevantes. A média de pontos das últimas cinco rodadas apresentou importância média de 0,0491, enquanto a média das últimas três e dez rodadas alcançaram valores de 0,0464 e 0,0319, respectivamente.

Esses resultados indicam que o desempenho recente exerce papel central na previsão da pontuação futura. As janelas de três e cinco rodadas permitem representar o momento atual do atleta, enquanto as médias de dez e trinta e oito rodadas fornecem informações

sobre seu comportamento em horizontes mais longos. A presença simultânea dessas variáveis mostra que os modelos combinaram informações de curto e longo prazo, em vez de dependerem exclusivamente de uma única janela temporal.

A média de pontos considerando até trinta e oito rodadas também apresentou contribuição relevante. Sua importância média foi de 0,0403 no Random Forest e de 0,0181 no XGBoost. Embora esses valores sejam inferiores aos obtidos pelas médias recentes, a variável de longo prazo permaneceu entre as principais características dos dois algoritmos. Esse comportamento sugere que o desempenho acumulado durante a temporada complementa as informações fornecidas pelas janelas menores.

A variável Consistência apresentou importância média de 0,0491 no Random Forest e de 0,0231 no XGBoost. Já o desvio-padrão das últimas cinco rodadas apresentou valores de 0,0375 e 0,0100, respectivamente. Essas duas variáveis estão relacionadas à regularidade do desempenho, mas representam aspectos complementares. Enquanto a consistência sintetiza a estabilidade geral do atleta, o desvio-padrão recente mede diretamente a dispersão de suas pontuações nas rodadas anteriores.

A relevância dessas medidas demonstra que os modelos não se basearam apenas na pontuação média. Atletas com médias semelhantes podem apresentar níveis distintos de risco: um deles pode pontuar de forma regular, enquanto outro alterna resultados muito altos e muito baixos. A incorporação de medidas de dispersão permite que os modelos diferenciem essas situações.

As variáveis relacionadas ao valor de mercado também apareceram entre as mais importantes. O preço atual apresentou importância média de 0,0326 no Random Forest e de 0,0205 no XGBoost. A variação de preço apresentou valores de 0,0258 e 0,0142, respectivamente. Além disso, as médias de preço das últimas três, cinco, dez e trinta e oito rodadas figuraram entre as principais variáveis do Random Forest.

Embora o preço não represente diretamente o desempenho esportivo, ele incorpora informações produzidas pela dinâmica do próprio jogo. Jogadores que apresentam boas atuações tendem a valorizar, enquanto desempenhos abaixo do esperado podem provocar desvalorização. Assim, o preço e sua variação funcionam como indicadores indiretos da trajetória recente do atleta e das expectativas acumuladas pelo mercado.

Entretanto, as variáveis relacionadas ao preço apresentaram importância inferior à das medidas de pontuação e do status do jogador. Esse resultado indica que o valor de mercado atua como uma informação complementar, mas não substitui os indicadores diretamente

associados ao desempenho em campo.

Também se observa diferença na concentração das importâncias. No XGBoost, o status e o clube concentram conjuntamente cerca de 60% da importância média. Já no Random Forest, embora o status também ocupe a primeira posição, as demais importâncias encontram-se distribuídas de maneira mais equilibrada entre histórico de pontuação, consistência, dispersão e preço.

Essa diferença é compatível com os mecanismos de aprendizagem dos algoritmos. O Random Forest combina árvores construídas de forma independente e tende a repartir a contribuição entre diversas variáveis correlacionadas. O XGBoost constrói árvores sequencialmente e pode concentrar maior importância nos atributos que proporcionam os maiores ganhos nas primeiras etapas do ajuste.

Os desvios-padrão apresentados na Tabela 4 também revelam diferenças entre as posições. No XGBoost, por exemplo, a média de pontos das últimas cinco rodadas apresentou importância média de 0,0491 e desvio-padrão de 0,0430, indicando variação considerável de sua contribuição entre goleiros, laterais, zagueiros, meias, atacantes e técnicos. O mesmo ocorreu com a variável clube, cujo desvio-padrão foi de 0,0561.

Em contraste, variáveis como preço atual e variação de preço apresentaram desvios-padrão menores nos dois modelos. Isso indica que, embora suas importâncias sejam mais modestas, sua contribuição foi relativamente mais estável entre as posições.

De maneira geral, os resultados demonstram que os modelos aprenderam padrões coerentes com a dinâmica do Cartola FC. A disponibilidade do atleta, seu histórico recente de pontuação, sua regularidade, o clube ao qual pertence e as informações de mercado constituíram as principais fontes de informação para a previsão.

A concordância entre os modelos quanto à relevância do status e do desempenho recente reforça a consistência da metodologia. Ao mesmo tempo, as diferenças observadas para a variável clube e para a distribuição das importâncias mostram que Random Forest e XGBoost capturaram aspectos complementares da base de dados, o que contribui para justificar a utilização conjunta de suas previsões no modelo *ensemble*.

3.3 Avaliação da capacidade de ranqueamento dos atletas

Embora as métricas de erro permitam avaliar a proximidade entre as pontuações previstas e observadas, elas não representam integralmente a utilidade prática dos modelos na construção de equipes. No Cartola FC, o objetivo principal não é necessariamente prever com exatidão a pontuação de todos os jogadores, mas identificar antecipadamente os atletas com maior potencial de desempenho em cada rodada.

Nesse contexto, torna-se relevante avaliar a capacidade de ranqueamento dos critérios utilizados pelo sistema. Um critério de seleção pode apresentar erros individuais relativamente elevados e, ainda assim, ser útil caso consiga posicionar entre os primeiros lugares os atletas que posteriormente obtenham boas pontuações. Essa característica é especialmente importante para o processo de otimização, que utiliza a ordenação produzida pelas previsões e pelos indicadores derivados para selecionar os jogadores da equipe.

A avaliação foi realizada em 37 rodadas, considerando os grupos formados pelos cinco, dez e vinte atletas mais bem classificados segundo seis critérios: previsão do *ensemble*, previsão ajustada pela probabilidade de participação, teto de pontuação, *score* híbrido, média histórica e preço do atleta. Para cada rodada e critério, foram calculadas a pontuação real média dos atletas pertencentes ao grupo Top- k e a soma de suas pontuações. Posteriormente, essas medidas foram agregadas entre as rodadas.

A Tabela 5 apresenta a pontuação real média dos atletas selecionados por cada critério. Os valores entre parênteses correspondem ao desvio-padrão da média entre as 37 rodadas.

Tabela 5: Pontuação real média dos atletas selecionados pelos diferentes critérios de ranqueamento.

Critério	Top 5	Top 10	Top 20
Predição do <i>ensemble</i>	6,489 (2,643)	5,944 (1,389)	5,683 (0,857)
Predição ajustada por <code>prob_jogar</code>	6,489 (2,643)	5,944 (1,389)	5,683 (0,857)
Teto de pontuação	6,918 (2,956)	6,286 (1,803)	5,815 (1,027)
<i>Score</i> híbrido	6,454 (2,395)	5,964 (1,411)	5,716 (0,874)
Média histórica	5,977 (2,745)	5,962 (2,003)	5,463 (1,484)
Preço	6,255 (3,224)	6,131 (2,553)	5,648 (1,451)

Fonte: Elaborado pelo autor. Os valores representam a média da pontuação real dos atletas selecionados ao longo de 37 rodadas; os desvios-padrão são apresentados entre parênteses.

Como complemento, a Tabela 6 apresenta a média da soma das pontuações reais dos atletas pertencentes a cada grupo Top- k .

Tabela 6: Média da soma das pontuações reais nos grupos Top- k .

Critério	Top 5	Top 10	Top 20
Predição do <i>ensemble</i>	32,445	59,444	113,663
Predição ajustada por prob_jogar	32,445	59,444	113,663
Teto de pontuação	34,591	62,862	116,308
<i>Score</i> híbrido	32,268	59,644	114,323
Média histórica	29,884	59,616	109,268
Preço	31,275	61,311	112,964

Fonte: Elaborado pelo autor.

Os resultados mostram que o critério baseado no **teto de pontuação** apresentou o melhor desempenho em todos os tamanhos de grupo analisados. No Top 5, os atletas selecionados por esse critério obtiveram pontuação real média de 6,918 pontos, superando a previsão pura do *ensemble*, com 6,489, o *score* híbrido, com 6,454, e a média histórica, com 5,977.

A superioridade do teto de pontuação permaneceu nos grupos maiores. No Top 10, o critério alcançou média de 6,286 pontos por atleta e soma média de 62,862 pontos. No Top 20, foram obtidos 5,815 pontos por atleta e soma média de 116,308 pontos. Em todos os casos, esses foram os maiores valores observados entre os critérios avaliados.

Em relação à média histórica, o teto de pontuação proporcionou ganho aproximado de 15,8% na pontuação média do Top 5, passando de 5,977 para 6,918 pontos. Para o Top 10, o ganho foi de aproximadamente 5,4%, enquanto, no Top 20, alcançou cerca de 6,4%. Esses resultados indicam que a incorporação de uma medida de potencial, formada pela previsão ajustada e pela variabilidade recente, permitiu identificar atletas com maior possibilidade de atingir pontuações elevadas.

A previsão pura do *ensemble* também apresentou desempenho superior à média histórica nos grupos Top 5 e Top 20. No Top 5, a pontuação média aumentou de 5,977 para 6,489, representando ganho de aproximadamente 8,6%. No Top 20, o ganho foi de aproximadamente 4,0%, com elevação de 5,463 para 5,683 pontos.

No Top 10, entretanto, a pontuação média obtida pela previsão do *ensemble*, igual a 5,944, ficou ligeiramente abaixo da média histórica, que alcançou 5,962. A diferença foi pequena e correspondeu a aproximadamente 0,3%. Esse resultado mostra que a vantagem de um critério pode variar conforme o número de atletas selecionados e reforça a necessidade de avaliar diferentes profundidades do ranking.

A previsão ajustada por **prob_jogar** apresentou exatamente os mesmos resultados da

previsão pura do *ensemble* nos três valores de k . Isso significa que, no conjunto utilizado para essa avaliação, o ajuste não alterou a composição dos grupos Top 5, Top 10 e Top 20. Esse resultado pode ocorrer quando os atletas considerados apresentam probabilidades de participação semelhantes ou quando a ponderação modifica os valores previstos sem alterar sua ordenação relativa. Assim, embora o ajuste seja relevante para reduzir o risco de selecionar atletas com baixa possibilidade de atuar, ele não produziu mudança no ranqueamento analisado.

O *score* híbrido apresentou comportamento competitivo. Sua pontuação média foi de 6,454 no Top 5, 5,964 no Top 10 e 5,716 no Top 20. No grupo Top 20, o *score* híbrido apresentou o segundo melhor desempenho, ficando atrás apenas do teto de pontuação. A soma média de 114,323 pontos superou os valores obtidos pela previsão pura, pela média histórica e pelo preço.

Esse resultado indica que a combinação da previsão ajustada, do teto de pontuação e do retorno esperado em relação ao preço consegue equilibrar diferentes aspectos da seleção. Entretanto, como o *score* híbrido também atribui peso ao custo-benefício, sua finalidade não é necessariamente selecionar exclusivamente os jogadores com maior pontuação potencial. Sua principal contribuição ocorre na etapa de otimização, em que o orçamento constitui uma restrição efetiva.

O preço dos atletas apresentou resultados relativamente elevados, especialmente no Top 10, em que alcançou média de 6,131 pontos e soma de 61,311. Esse valor foi superior aos obtidos pelo *ensemble*, pelo *score* híbrido e pela média histórica, embora ainda inferior ao teto de pontuação. Esse comportamento sugere que o preço contém informações relevantes sobre o potencial dos jogadores, uma vez que atletas valorizados frequentemente apresentam histórico de boas atuações.

Entretanto, o critério baseado no preço apresentou os maiores desvios-padrão entre as rodadas: 3,224 no Top 5, 2,553 no Top 10 e 1,451 no Top 20. Portanto, embora sua média tenha sido competitiva, o desempenho mostrou-se menos estável. Selecionar os atletas mais caros pode produzir resultados elevados em algumas rodadas, mas também está associado a maior variabilidade.

A média histórica apresentou o pior desempenho no Top 5 e no Top 20. Além disso, registrou desvios-padrão relativamente elevados, especialmente no Top 10 e no Top 20. Esse resultado indica que a utilização isolada do desempenho médio anterior não é suficiente para identificar de forma consistente os atletas com maior potencial de pontuação na rodada seguinte.

Observa-se ainda que as pontuações médias diminuem conforme o valor de k aumenta. Esse comportamento é esperado, pois os primeiros atletas do ranking representam as opções consideradas mais promissoras. À medida que o grupo é ampliado, passam a ser incluídos jogadores com menor pontuação esperada, reduzindo a média geral. Ao mesmo tempo, os desvios-padrão também tendem a diminuir, pois a média calculada sobre grupos maiores é menos sensível ao desempenho excepcional ou negativo de poucos atletas.

De maneira geral, os resultados fornecem evidências de que os indicadores construídos a partir dos modelos possuem utilidade prática para ordenar os atletas. O teto de pontuação apresentou o melhor desempenho em todos os grupos avaliados, demonstrando que a combinação entre expectativa de pontuação e variabilidade recente foi eficiente para identificar jogadores com maior potencial de desempenho.

O *score* híbrido também apresentou resultados relevantes, sobretudo no Top 20, e permanece especialmente adequado para a etapa de escalação, na qual pontuação esperada, risco e custo precisam ser avaliados conjuntamente. Assim, a análise de ranqueamento complementa as métricas preditivas e demonstra que os resultados dos modelos podem ser convertidos em critérios úteis para a seleção dos atletas.

3.4 Avaliação da estratégia de escalação otimizada

Após avaliar a capacidade dos modelos em prever a pontuação individual dos atletas e em ordená-los segundo seu potencial de desempenho, a etapa seguinte consistiu em verificar se essas informações seriam capazes de produzir equipes competitivas quando submetidas às restrições do Cartola FC.

Para isso, foi realizado um estudo retrospectivo utilizando as 37 rodadas da temporada de 2025. Em cada rodada foram construídas seis equipes distintas, empregando diferentes critérios de otimização: predição do *ensemble*, predição ajustada pela probabilidade de participação, teto de pontuação, *score* híbrido, média histórica e preço dos atletas.

Todas as equipes foram obtidas por meio do mesmo modelo de programação inteira, respeitando o orçamento máximo disponível, a formação tática escolhida, a quantidade de jogadores por posição e o limite máximo de atletas pertencentes ao mesmo clube. Dessa forma, as diferenças observadas entre as estratégias decorrem exclusivamente do critério utilizado para atribuir valor aos atletas durante o processo de otimização.

A Tabela 7 apresenta o desempenho médio das equipes obtidas pelas diferentes estra-

tégias ao longo das 37 rodadas avaliadas.

Tabela 7: Desempenho médio das estratégias de escalação.

Estratégia	Pontuação	Desvio	Erro absoluto	Custo médio
Predição pura	74,45	23,79	29,70	148,28
Predição ajustada	74,45	23,79	29,70	148,28
Teto de pontuação	72,00	22,75	27,53	143,98
Score híbrido	72,11	23,37	27,28	146,32
Mais caros	71,55	28,16	31,90	170,52
Maior média histórica	68,59	28,01	29,42	150,97

Fonte: Elaborado pelo autor.

Os resultados mostram que a estratégia baseada na predição do ensemble apresentou o melhor desempenho médio entre todas as abordagens avaliadas. As equipes construídas utilizando diretamente a pontuação prevista obtiveram média de 74,45 pontos por rodada. O mesmo resultado foi observado para a estratégia baseada na predição ajustada pela probabilidade de participação, indicando que, nas rodadas analisadas, a ponderação pela probabilidade de atuação não modificou a composição final das equipes selecionadas.

Em comparação com a estratégia baseada na média histórica dos atletas, a utilização das previsões proporcionou um ganho médio de aproximadamente 5,9 pontos por rodada, correspondente a um aumento de cerca de 8,5% na pontuação das equipes. Esse resultado demonstra que as informações aprendidas pelos modelos conseguem capturar padrões de desempenho que não são representados apenas pelas médias históricas dos jogadores.

Embora as estratégias baseadas no teto de pontuação e no score híbrido tenham produzido pontuações médias ligeiramente inferiores às obtidas pela predição direta, ambas apresentaram menor erro absoluto médio entre a pontuação prevista e a efetivamente observada. O score híbrido apresentou erro absoluto médio de 27,28 pontos, enquanto o teto de pontuação obteve 27,53 pontos, valores inferiores aos 29,70 pontos observados na estratégia baseada exclusivamente na previsão.

Esses resultados sugerem que a incorporação de medidas relacionadas ao risco e à variabilidade recente produz equipes com desempenho mais consistente, ainda que a pontuação média final seja ligeiramente inferior. Essa característica pode ser particularmente interessante para usuários que priorizam estabilidade ao longo da competição.

Outro aspecto relevante refere-se ao custo das equipes construídas. As estratégias baseadas no teto de pontuação e no score híbrido utilizaram, em média, 143,98 e 146,32 cartoletas, respectivamente, valores inferiores aos observados para a estratégia baseada

na previsão direta, cujo custo médio foi de 148,28 cartoletas. Já a estratégia formada simplesmente pelos atletas mais caros apresentou o maior custo médio de 170,52 cartoletas, e, ainda assim, obteve desempenho inferior ao das equipes construídas pelos modelos preditivos.

Esse resultado evidencia que atletas de maior valor de mercado não necessariamente produzem as melhores equipes quando consideradas simultaneamente as restrições orçamentárias e a necessidade de equilíbrio entre posições. O processo de otimização consegue explorar relações entre desempenho esperado, custo e composição da equipe que não são capturadas por estratégias baseadas exclusivamente no preço dos jogadores.

Observa-se também que as estratégias baseadas nos modelos preservaram parcela significativa do orçamento disponível. O saldo médio permaneceu superior a 100 cartoletas nas estratégias de predição, teto de pontuação e score híbrido, indicando que a maximização da pontuação não exigiu utilizar integralmente o orçamento disponível. Esse comportamento é consequência da existência de atletas com excelente relação entre desempenho esperado e custo, tornando desnecessária a aquisição dos jogadores mais caros para obter equipes competitivas.

De forma geral, os resultados demonstram que os modelos desenvolvidos não apenas apresentaram bom desempenho preditivo individual, mas também conseguiram produzir escalações competitivas quando incorporados ao processo de otimização. Entre as estratégias avaliadas, a utilização direta da previsão do ensemble apresentou a maior pontuação média, enquanto o score híbrido e o teto de pontuação produziram equipes ligeiramente mais estáveis e com menor erro médio de previsão. Esses resultados reforçam a aplicabilidade prática da metodologia proposta para apoiar a tomada de decisão na escalação de equipes do Cartola FC.

3.5 Avaliação da escolha do capitão

No Cartola FC, a escolha do capitão possui influência direta sobre a pontuação final da equipe, pois o atleta selecionado recebe uma bonificação sobre a pontuação obtida na rodada. Dessa forma, além de formar uma equipe competitiva, é necessário identificar, entre os jogadores escalados, aquele com maior potencial de produzir uma pontuação elevada.

Na metodologia proposta, o capitão foi escolhido entre os atletas pertencentes à equipe otimizada, utilizando como critério o maior valor do teto de pontuação. Essa medida com-

bina a previsão ajustada do atleta com uma parcela relacionada à variabilidade recente de seu desempenho, buscando identificar jogadores que, além de apresentarem boa expectativa de pontuação, possuam potencial para alcançar resultados superiores à média prevista.

É importante destacar que o critério utilizado para selecionar o capitão permaneceu o mesmo em todas as estratégias de escalação. Entretanto, como cada estratégia produziu uma composição de equipe diferente, o conjunto de atletas disponíveis para a escolha do capitão também variou. Assim, a análise permite verificar indiretamente quais estratégias foram capazes de incluir atletas com maior potencial para exercer essa função.

A Tabela 8 apresenta os resultados médios da escolha do capitão ao longo das 37 rodadas avaliadas.

Tabela 8: Desempenho dos capitães escolhidos nas diferentes estratégias de escalação.

Estratégia	Pontuação média	Desvio-padrão	Bônus médio	Bônus total
Mais caros	8,21	7,89	4,11	151,89
Maior média histórica	7,69	6,97	3,84	142,26
Predição pura	7,51	8,06	3,75	138,90
Predição ajustada	7,51	8,06	3,75	138,90
Score híbrido	7,47	8,05	3,74	138,20
Teto de pontuação	7,15	8,02	3,58	132,35

Fonte: Elaborado pelo autor. A pontuação média corresponde à pontuação real obtida pelo atleta escolhido como capitão, antes da aplicação da bonificação.

Os resultados mostram que a estratégia baseada nos atletas mais caros produziu os capitães com maior pontuação média, alcançando 8,21 pontos por rodada. Consequentemente, essa estratégia também apresentou o maior bônus médio, igual a 4,11 pontos, e o maior bônus acumulado ao longo da temporada, correspondente a 151,89 pontos.

Esse resultado sugere que os atletas de maior preço tendem a apresentar elevado potencial individual para exercer a função de capitão. O preço no Cartola FC incorpora, em parte, o desempenho histórico e as expectativas em relação aos jogadores, o que favorece a presença de atletas ofensivos e de destaque entre os mais valorizados.

Entretanto, esse resultado deve ser interpretado conjuntamente com a avaliação da equipe completa. Apesar de produzir o maior bônus de capitão, a estratégia baseada nos atletas mais caros não apresentou a maior pontuação total das escalações. Sua pontuação média com capitão foi de 71,55 pontos, inferior aos 74,45 pontos obtidos pelas estratégias baseadas na predição pura e na predição ajustada.

Portanto, a escolha de um capitão com desempenho elevado não foi suficiente para compensar uma composição de equipe menos eficiente. Esse resultado evidencia que o sucesso da escalação depende do equilíbrio entre a qualidade do conjunto de jogadores e a escolha individual do capitão.

A estratégia baseada na maior média histórica apresentou a segunda maior pontuação média dos capitães, com 7,69 pontos, além do menor desvio-padrão, igual a 6,97. O bônus médio foi de 3,84 pontos, totalizando 142,26 pontos ao longo das 37 rodadas. Assim, embora essa estratégia tenha apresentado desempenho inferior na formação da equipe, foi capaz de selecionar capitães relativamente produtivos e mais estáveis.

As estratégias baseadas na predição pura e na predição ajustada apresentaram resultados idênticos. Em ambas, os capitães obtiveram média de 7,51 pontos, gerando bônus médio de 3,75 pontos e bônus total de 138,90 pontos.

A igualdade dos resultados decorre do fato de as duas estratégias terem produzido as mesmas equipes nas rodadas avaliadas. Como o ajuste pela probabilidade de participação não alterou a composição das escalações, o conjunto de candidatos à função de capitão permaneceu o mesmo e, conseqüentemente, os atletas escolhidos também foram iguais.

A estratégia baseada no *score* híbrido apresentou pontuação média de 7,47 para o capitão e bônus médio de 3,74 pontos. Esses valores ficaram muito próximos dos resultados obtidos pelas estratégias de predição, indicando que a incorporação de informações relacionadas ao custo-benefício e ao risco pouco alterou a qualidade dos atletas selecionados para a função.

Já a estratégia baseada diretamente no teto de pontuação apresentou a menor pontuação média dos capitães, igual a 7,15 pontos, e bônus médio de 3,58 pontos. À primeira vista, esse resultado pode parecer contraditório, uma vez que o teto também foi utilizado para selecionar o capitão.

Entretanto, a otimização da equipe pelo teto de pontuação busca maximizar o valor agregado de todos os atletas escalados, e não necessariamente garantir a presença do jogador que posteriormente obterá a maior pontuação real. Além disso, o teto representa uma estimativa de potencial, e não uma previsão exata do desempenho máximo. Atletas com maior variabilidade recente podem receber valores elevados de teto, mas essa variabilidade também implica maior incerteza.

A Tabela 9 apresenta a frequência de rodadas em que o capitão gerou bônus positivo ou negativo.

Tabela 9: Frequência de bônus positivo e negativo dos capitães.

Estratégia	Bônus positivo (%)	Bônus negativo (%)
Mais caros	97,30	2,70
Maior média histórica	94,59	2,70
Predição pura	89,19	2,70
Predição ajustada	89,19	2,70
<i>Score</i> híbrido	89,19	0,00
Teto de pontuação	89,19	0,00

Fonte: Elaborado pelo autor. Rodadas com pontuação igual a zero não foram classificadas como bônus positivo nem negativo.

A estratégia baseada nos atletas mais caros apresentou bônus positivo em 97,30% das rodadas, o maior percentual entre as abordagens avaliadas. A maior média histórica também apresentou elevada frequência de resultados positivos, correspondente a 94,59%.

As estratégias de predição, teto e *score* híbrido apresentaram bônus positivo em 89,19% das rodadas. Apesar do percentual ligeiramente inferior, as estratégias de teto e *score* híbrido não registraram bônus negativo, enquanto as demais apresentaram resultado negativo em aproximadamente 2,70% das rodadas.

Esse comportamento indica que o teto e o *score* híbrido foram mais conservadores em relação à ocorrência de capitães com pontuação negativa. Por outro lado, a estratégia baseada nos atletas mais caros apresentou maior retorno médio e maior frequência de bônus positivo, embora tenha registrado maior variabilidade.

De forma geral, a análise mostra que a escolha do capitão contribuiu de maneira relevante para o desempenho das equipes em todas as estratégias. O bônus médio variou entre 3,58 e 4,11 pontos por rodada, representando, ao longo da temporada, um acréscimo acumulado entre 132,35 e 151,89 pontos.

Apesar de a estratégia baseada nos atletas mais caros ter produzido os capitães com melhor desempenho individual, as equipes construídas pelas previsões apresentaram maior pontuação total. Esse resultado reforça que a escolha do capitão deve ser considerada como uma etapa complementar à formação da equipe, e não como um substituto para uma estratégia de escalação globalmente eficiente.

Assim, o critério baseado no teto de pontuação mostrou-se adequado para identificar jogadores com potencial elevado e evitar, em algumas estratégias, capitães com pontuação negativa. Entretanto, os resultados também indicam a possibilidade de aperfeiçoamento dessa etapa, por meio da comparação futura entre diferentes critérios específicos para

capitão, como previsão direta, teto, consistência, potencial ofensivo e combinação entre retorno esperado e risco.

3.6 Estudo de caso

Com o objetivo de complementar a avaliação agregada das estratégias de escalação, foram selecionadas duas rodadas da temporada de 2025 para uma análise mais detalhada. A primeira corresponde a uma rodada de elevado desempenho das equipes otimizadas, enquanto a segunda representa uma situação em que a estratégia baseada nas previsões apresentou resultado abaixo do esperado.

A análise de rodadas específicas permite observar aspectos que não são completamente evidenciados pelas médias gerais, como a variabilidade do desempenho dos atletas, o impacto da escolha do capitão e as diferenças entre as equipes produzidas por cada critério de otimização.

3.6.1 Rodada 21: cenário de elevado desempenho

A rodada 21 foi selecionada por apresentar a maior pontuação obtida pela estratégia baseada na predição pura. Nessa rodada, a equipe construída a partir das previsões do *ensemble* alcançou 117,72 pontos antes da aplicação da bonificação do capitão e 134,27 pontos após a inclusão desse bônus.

A Tabela 10 apresenta os resultados das seis estratégias avaliadas nessa rodada.

Tabela 10: Resultados das estratégias de escalação na rodada 21.

Estratégia	Custo	Previsão	Real sem capitão	Real com capitão
Predição pura	151,12	43,10	117,72	134,27
Predição ajustada	151,12	43,10	117,72	134,27
Teto de pontuação	139,74	40,66	105,24	121,79
Score híbrido	151,12	43,10	117,72	134,27
Maior média histórica	160,18	37,14	169,71	176,46
Mais caros	169,78	39,32	162,21	178,76

Fonte: Elaborado pelo autor.

As estratégias de predição pura, predição ajustada e *score* híbrido produziram resultados idênticos nessa rodada. As três equipes apresentaram custo de 151,12 cartoletas, pontuação prevista de 43,10 pontos e pontuação real de 134,27 pontos após a aplicação da bonificação do capitão.

O capitão selecionado por essas estratégias foi Samuel Lino, que obteve 33,10 pontos. Como a bonificação aplicada correspondeu a 50% da pontuação do atleta, sua escolha acrescentou 16,55 pontos ao resultado da equipe. Portanto, a atuação do capitão teve participação importante no desempenho elevado observado nessa rodada.

A equipe baseada no teto de pontuação também selecionou Samuel Lino como capitão, mas apresentou composição distinta. Seu custo foi de 139,74 cartoletas e sua pontuação final foi de 121,79 pontos, aproximadamente 12,48 pontos abaixo do resultado obtido pela predição pura.

Apesar do bom desempenho das equipes preditivas, as estratégias baseadas na média histórica e no preço produziram os maiores resultados da rodada. A estratégia formada pelos atletas mais caros alcançou 178,76 pontos, enquanto a equipe baseada na maior média histórica obteve 176,46 pontos.

Na estratégia dos atletas mais caros, Samuel Lino também foi escolhido como capitão. Já na estratégia baseada na média histórica, o capitão foi Kaio Jorge, que marcou 13,50 pontos e produziu bônus de 6,75 pontos.

Esse resultado mostra que uma estratégia pode apresentar desempenho superior em determinada rodada mesmo quando sua média geral ao longo da temporada é inferior. A rodada 21 favoreceu especialmente os jogadores selecionados pelos critérios de preço e média histórica, os quais alcançaram pontuações excepcionalmente elevadas.

Também se observa uma diferença expressiva entre as pontuações previstas e realizadas. A equipe baseada na predição pura apresentou previsão total de 43,10 pontos, mas obteve 117,72 pontos antes da bonificação do capitão. Isso corresponde a uma diferença positiva de aproximadamente 74,62 pontos.

Essa discrepância indica que vários jogadores da equipe superaram simultaneamente suas expectativas individuais. Embora esse comportamento beneficie a escalação, ele também evidencia a dificuldade de estimar acontecimentos excepcionais, como gols, assistências, defesas de pênalti e grandes quantidades de desarmes em uma única rodada.

3.6.2 Rodada 9: cenário de desempenho abaixo do esperado

A rodada 9 foi escolhida para representar um cenário menos favorável à estratégia baseada nas previsões. Nessa rodada, a equipe construída pela predição pura apresentou pontuação prevista de 41,73 pontos, mas obteve apenas 30,20 pontos antes da bonificação

do capitão.

A Tabela 11 apresenta os resultados obtidos pelas diferentes estratégias.

Tabela 11: Resultados das estratégias de escalação na rodada 9.

Estratégia	Custo	Previsão	Real sem capitão	Real com capitão
Predição pura	154,91	41,73	30,20	32,00
Predição ajustada	154,91	41,73	30,20	32,00
Teto de pontuação	157,80	40,51	53,00	54,80
<i>Score</i> híbrido	149,95	41,68	33,20	35,00
Maior média histórica	158,59	34,65	56,52	58,32
Mais caros	174,30	35,05	52,30	54,10

Fonte: Elaborado pelo autor.

As estratégias de predição pura e predição ajustada produziram novamente a mesma equipe, com custo de 154,91 cartoletas. A pontuação final, considerando a bonificação do capitão, foi de apenas 32,00 pontos.

O capitão escolhido foi Kaio Jorge, que obteve 3,60 pontos e acrescentou somente 1,80 ponto à equipe. Embora sua pontuação tenha sido positiva, o retorno foi insuficiente para compensar o desempenho reduzido dos demais atletas escalados.

A estratégia baseada no *score* híbrido apresentou comportamento semelhante, alcançando 35,00 pontos com capitão. Apesar de possuir composição parcialmente distinta, seu resultado ficou próximo ao das equipes orientadas diretamente pela previsão.

Em contrapartida, a estratégia baseada na maior média histórica obteve o melhor desempenho da rodada, com 58,32 pontos. A equipe construída pelo teto de pontuação atingiu 54,80 pontos, enquanto a seleção dos atletas mais caros alcançou 54,10 pontos.

As seis estratégias escolheram Kaio Jorge como capitão. Portanto, as diferenças entre as pontuações finais decorreram essencialmente da composição do restante das equipes, e não da escolha do capitão.

A equipe baseada na predição pura apresentou erro de $-11,53$ pontos em relação à previsão, indicando que o desempenho real ficou abaixo do valor estimado. Ainda assim, o erro observado não foi extremamente elevado quando comparado às diferenças positivas registradas em rodadas de pontuações excepcionais.

O resultado da rodada 9 evidencia uma limitação inerente aos modelos preditivos utilizados. As estimativas representam valores esperados condicionados às informações disponíveis antes da rodada, mas não garantem que os atletas selecionados alcançarão

esse desempenho em uma realização específica.

Fatores como substituições precoces, cartões, gols sofridos, ausência de participação direta em jogadas ofensivas e baixo volume de ações contabilizadas podem fazer com que vários jogadores apresentem pontuações inferiores às previstas simultaneamente.

3.6.3 Comparação entre os dois cenários

Os dois estudos de caso ilustram a elevada variabilidade existente na pontuação do Cartola FC. Na rodada 21, a equipe baseada na predição pura superou sua previsão em 74,62 pontos antes da bonificação do capitão. Na rodada 9, por outro lado, o resultado ficou 11,53 pontos abaixo do previsto.

Essa assimetria mostra que os maiores erros absolutos podem ocorrer em rodadas nas quais os atletas apresentam desempenho muito superior ao esperado, e não somente em situações de fracasso da estratégia. Portanto, um erro elevado de previsão não significa necessariamente uma escalação ruim, pois pode estar associado a uma pontuação real excepcionalmente alta.

Também se verifica que nenhuma estratégia foi dominante em todas as rodadas. Embora a predição pura tenha apresentado a maior pontuação média ao longo da temporada, outras abordagens superaram seu desempenho nos dois casos analisados. Na rodada 21, os atletas mais caros apresentaram o melhor resultado, enquanto, na rodada 9, a maior média histórica ocupou a primeira posição.

Esse comportamento reforça a importância de avaliar as estratégias em um conjunto amplo de rodadas, evitando conclusões baseadas apenas em casos isolados. O melhor desempenho médio das equipes preditivas decorre de sua regularidade ao longo da temporada e não da capacidade de vencer todas as comparações rodada a rodada.

Os resultados também demonstram a influência da escolha do capitão. Na rodada 21, o bônus de Samuel Lino acrescentou 16,55 pontos à equipe, enquanto, na rodada 9, Kaio Jorge contribuiu com apenas 1,80 ponto adicional. Assim, o impacto do capitão depende fortemente da realização individual observada em cada rodada.

De maneira geral, os estudos de caso confirmam que a metodologia proposta é capaz de produzir equipes com pontuações elevadas, mas permanece sujeita à incerteza característica do futebol e do sistema de pontuação do Cartola FC. A análise agregada de toda a temporada, portanto, constitui uma medida mais adequada da qualidade das estratégias

do que o resultado de uma rodada isolada.

3.7 Discussão geral

Os resultados obtidos ao longo deste estudo mostram que a aplicação conjunta de técnicas de aprendizado de máquina e otimização combinatória constitui uma abordagem viável para apoiar a escalação de equipes no Cartola FC. A metodologia desenvolvida permitiu transformar previsões individuais de pontuação em decisões práticas de seleção de atletas, respeitando simultaneamente restrições de orçamento, formação tática, posição e limite de jogadores por clube.

Na etapa de modelagem preditiva, o XGBoost apresentou o melhor desempenho entre os modelos avaliados, com menor erro absoluto médio, menor raiz do erro quadrático médio e maior correlação entre as pontuações previstas e observadas. Em relação ao modelo de referência, houve redução aproximada de 9,2% no erro absoluto médio e aumento de cerca de 51% na correlação. Esses resultados indicam que os modelos foram capazes de identificar padrões relevantes associados ao desempenho futuro dos atletas, embora a capacidade preditiva tenha permanecido limitada pela elevada variabilidade inerente ao futebol.

A análise da importância das variáveis mostrou que o desempenho recente dos jogadores exerceu papel central nas previsões. Variáveis como média de pontos nas últimas três, cinco e dez partidas, consistência, desvio-padrão recente e desempenho acumulado apareceram entre os principais preditores. Além dessas medidas, o status do atleta apresentou elevada importância nos dois modelos analisados, evidenciando que a disponibilidade e a probabilidade de participação são fatores fundamentais para a construção de previsões úteis.

O preço dos jogadores também apresentou contribuição relevante, embora não tenha sido suficiente, isoladamente, para produzir as melhores escalações. Esse resultado é coerente com o funcionamento do Cartola FC, no qual o valor de mercado reflete parcialmente o desempenho histórico, a popularidade e as expectativas em relação aos atletas. Entretanto, a relação entre preço e pontuação futura não é determinística, o que explica o desempenho inferior da estratégia baseada exclusivamente nos jogadores mais caros.

A avaliação da capacidade de ranqueamento complementou a análise das métricas tradicionais de regressão. Enquanto o XGBoost apresentou o melhor desempenho global de previsão, a estratégia baseada no teto de pontuação foi superior na seleção dos atletas

mais bem posicionados nos conjuntos Top 5, Top 10 e Top 20. Esse resultado mostra que o modelo com menor erro médio não necessariamente produz o melhor ordenamento dos jogadores com maior potencial.

Essa distinção é importante no contexto do Cartola FC, pois a decisão prática não depende apenas da precisão das estimativas individuais. O principal interesse consiste em identificar um subconjunto reduzido de atletas com maior probabilidade de obter boas pontuações. Assim, métricas de ranqueamento e avaliações Top- k podem ser mais informativas do que medidas agregadas de erro quando o objetivo é selecionar jogadores para compor uma equipe.

A estratégia baseada no teto de pontuação apresentou a maior pontuação média nos diferentes conjuntos Top- k . Esse comportamento sugere que a incorporação da variabilidade recente permitiu identificar atletas com maior potencial de desempenho acima da média prevista. Entretanto, o mesmo critério não produziu a maior pontuação média quando incorporado diretamente à otimização da equipe completa.

Na avaliação das escalações, as estratégias baseadas na predição pura e na predição ajustada alcançaram a maior pontuação média, correspondente a 74,45 pontos por rodada. O teto de pontuação e o *score* híbrido apresentaram resultados ligeiramente inferiores, com médias próximas de 72 pontos, mas produziram menor erro absoluto entre a pontuação prevista e a observada.

Essa diferença mostra que uma estratégia pode ser eficiente para ranquear atletas individualmente e, ainda assim, não maximizar o desempenho da equipe completa. A otimização envolve a combinação simultânea de jogadores de diferentes posições e preços, de modo que o valor agregado de uma equipe depende da complementaridade entre seus integrantes e não apenas da presença dos atletas mais bem ranqueados.

A estratégia baseada na predição pura superou a escalação construída pela maior média histórica em aproximadamente 5,9 pontos por rodada. Esse ganho corresponde a um aumento de cerca de 8,5% na pontuação média, evidenciando que as previsões capturaram informações adicionais além do histórico simples de desempenho.

Esse resultado representa uma das principais evidências favoráveis à metodologia proposta. Caso as previsões apenas reproduzissem as médias históricas, não seria esperado um ganho consistente após a otimização. A superioridade observada indica que os modelos conseguiram combinar informações recentes, condições de participação, características dos clubes, preço e variabilidade de desempenho de forma útil para a seleção das equipes.

Também se verificou que as equipes construídas pelos modelos não utilizaram integralmente o orçamento disponível. Em média, as estratégias de predição, teto e *score* híbrido apresentaram custos inferiores aos da estratégia formada pelos jogadores mais caros. Esse comportamento evidencia que a otimização privilegiou atletas com melhor relação entre desempenho esperado e custo.

No Cartola FC, utilizar todo o patrimônio disponível não garante uma equipe superior. A presença de jogadores mais baratos com boas previsões permite liberar orçamento sem comprometer a pontuação esperada. A metodologia proposta, portanto, atuou como um mecanismo de identificação de custo-benefício, ainda que esse critério não tenha sido incluído de forma isolada na função objetivo de todas as estratégias.

A análise do capitão mostrou um resultado distinto daquele observado para a equipe completa. A estratégia baseada nos atletas mais caros produziu os capitães com maior pontuação média, igual a 8,21 pontos, e o maior bônus acumulado. Entretanto, essa vantagem não foi suficiente para fazer com que essa estratégia alcançasse a maior pontuação total das equipes.

Esse resultado evidencia que a qualidade da escolha do capitão não deve ser analisada separadamente da composição global da escalação. Um capitão de elevado desempenho pode aumentar significativamente a pontuação final, mas não consegue compensar sistematicamente uma equipe com seleção menos eficiente nas demais posições.

As estratégias de predição pura e ajustada apresentaram resultados idênticos tanto na formação das equipes quanto na escolha dos capitães. Isso indica que, no conjunto de rodadas avaliado, o ajuste pela probabilidade de participação não alterou a ordem relativa dos atletas de maneira suficiente para modificar as soluções do problema de otimização.

Essa igualdade não implica que a probabilidade de participação seja irrelevante. A variável status apresentou elevada importância nos modelos, mostrando que a disponibilidade dos atletas já estava parcialmente incorporada às previsões. Além disso, o ajuste pode se tornar mais relevante em rodadas com maior número de jogadores classificados como dúvidas ou com risco de não atuação.

Os estudos de caso evidenciaram a elevada variabilidade dos resultados entre as rodadas. Em determinadas situações, as equipes superaram amplamente suas previsões, enquanto, em outras, apresentaram desempenho inferior ao esperado. Na rodada 21, por exemplo, a equipe baseada na predição pura obteve pontuação muito superior à prevista. Já na rodada 9, seu desempenho ficou abaixo da estimativa inicial.

Essa variação decorre da natureza aleatória e dinâmica do futebol. Eventos como gols, assistências, cartões, substituições, lesões, pênaltis, defesas difíceis e alterações táticas podem modificar substancialmente a pontuação de um jogador em uma única partida. Muitos desses acontecimentos são de difícil previsão a partir das informações disponíveis antes do início da rodada.

A ocorrência simultânea de desempenhos excepcionais também pode produzir erros elevados, mesmo quando a escalação é bem-sucedida. Portanto, o erro entre a pontuação prevista e a observada não deve ser interpretado isoladamente como medida da qualidade prática de uma equipe. Uma previsão pode apresentar grande erro positivo porque os atletas selecionados superaram amplamente suas expectativas.

Outra evidência importante é que nenhuma estratégia foi superior em todas as rodadas. Embora as previsões tenham produzido o melhor desempenho médio ao longo da temporada, as estratégias baseadas em preço, média histórica, teto e *score* híbrido foram superiores em rodadas específicas.

Esse comportamento reforça a importância da avaliação em múltiplas rodadas. Resultados isolados podem favorecer determinada estratégia em razão de acontecimentos excepcionais, enquanto a análise agregada permite avaliar consistência, retorno médio e estabilidade ao longo do tempo.

De forma geral, os resultados permitem concluir que a combinação entre modelos preditivos e programação inteira foi mais eficiente do que abordagens baseadas exclusivamente em preço ou desempenho histórico. O ganho observado não se restringiu às métricas de previsão, mas se refletiu na pontuação efetivamente obtida pelas equipes otimizadas.

Apesar dos resultados favoráveis, algumas limitações devem ser consideradas. A primeira refere-se à própria incerteza da variável resposta. A pontuação dos atletas apresenta elevada dispersão e forte influência de eventos pouco frequentes, o que limita a capacidade de previsão dos modelos.

A segunda limitação está relacionada à disponibilidade das informações antes de cada rodada. Alterações de escalação, lesões, suspensões, mudanças táticas e notícias divulgadas próximas ao fechamento do mercado podem afetar significativamente a probabilidade de participação dos atletas.

Além disso, as variáveis utilizadas não representam integralmente o contexto de cada partida. Informações como adversário, mando de campo, desempenho ofensivo e defensivo das equipes, estilo de jogo, provável escalação, posição efetivamente exercida em campo e

dificuldade do confronto podem ampliar a capacidade de discriminação dos modelos.

Outra limitação refere-se ao critério utilizado para a escolha do capitão. Embora o teto de pontuação tenha sido empregado como medida de potencial, os resultados indicaram que outras estratégias produziram capitães com maior retorno médio. Isso sugere que a seleção do capitão pode ser tratada como um problema específico, distinto da otimização da equipe.

Em trabalhos futuros, pode-se desenvolver um modelo exclusivo para estimar a probabilidade de um atleta alcançar uma pontuação elevada. Em vez de utilizar apenas a previsão média, esse modelo poderia estimar quantis superiores, probabilidades de ultrapassar determinados valores ou medidas de retorno ajustado ao risco.

Também seria possível incorporar diferentes perfis de usuário na função objetivo. Uma estratégia conservadora poderia priorizar atletas com maior consistência e menor variabilidade, enquanto uma estratégia agressiva poderia favorecer jogadores com maior teto de pontuação. Uma abordagem intermediária poderia combinar retorno esperado, risco, custo-benefício e potencial de valorização.

Outra possibilidade consiste em construir portfólios de equipes, em vez de produzir apenas uma escalação por rodada. Nesse caso, seria possível gerar múltiplas equipes com níveis controlados de diversidade, reduzindo a dependência de atletas específicos e distribuindo o risco entre diferentes combinações.

A incorporação de informações textuais provenientes de notícias esportivas também representa uma alternativa promissora. Técnicas de processamento de linguagem natural poderiam ser utilizadas para identificar dúvidas, lesões, retornos, mudanças de posição e expectativas de escalação que não estejam plenamente representadas nas variáveis estruturadas.

Por fim, a metodologia pode ser atualizada continuamente à medida que novas rodadas são disputadas. A inclusão de observações recentes, o retreinamento periódico dos modelos e o ajuste dos hiperparâmetros podem reduzir a perda de desempenho causada por mudanças no comportamento das equipes e dos jogadores ao longo da temporada.

Em síntese, os resultados mostram que a metodologia proposta foi capaz de integrar previsão, ranqueamento e otimização em um único processo de apoio à decisão. Os modelos apresentaram ganhos em relação às referências utilizadas, e as equipes otimizadas com base nas previsões alcançaram o melhor desempenho médio entre as estratégias avaliadas.

Embora a imprevisibilidade do futebol impeça a obtenção de resultados superiores em

todas as rodadas, a abordagem desenvolvida mostrou-se consistente em análise agregada. Assim, o sistema não deve ser interpretado como um mecanismo capaz de garantir a melhor escalação em cada rodada, mas como uma ferramenta quantitativa para organizar informações, comparar alternativas e melhorar a qualidade das decisões ao longo de uma temporada.

4 Considerações finais

Este trabalho teve como objetivo desenvolver uma metodologia para apoiar a tomada de decisão na escalação de equipes do Cartola FC por meio da integração de técnicas de aprendizado de máquina e otimização matemática. Para isso, foram construídos modelos preditivos capazes de estimar a pontuação futura dos atletas, os quais foram posteriormente incorporados a um modelo de programação inteira responsável por selecionar a equipe ótima sob as restrições impostas pelo jogo.

Os resultados obtidos demonstraram que a abordagem proposta foi capaz de produzir previsões mais precisas do que um modelo de referência baseado na média histórica dos jogadores. Entre os algoritmos avaliados, o XGBoost apresentou o melhor desempenho nas métricas de regressão, obtendo menor erro absoluto médio, menor raiz do erro quadrático médio e maior correlação entre as pontuações previstas e observadas. Além disso, a análise da importância das variáveis mostrou que informações relacionadas ao desempenho recente, à consistência dos atletas e à sua disponibilidade para atuar constituem os fatores mais relevantes para a previsão das pontuações.

A avaliação da capacidade de ranqueamento evidenciou que diferentes critérios podem ser mais adequados dependendo do objetivo da análise. Enquanto os modelos preditivos apresentaram melhor desempenho global nas métricas de regressão, o critério baseado no teto de pontuação destacou-se na identificação dos atletas com maior potencial de desempenho entre os conjuntos Top- k . Esse resultado demonstra que a qualidade de um modelo não deve ser avaliada exclusivamente por métricas tradicionais de erro, mas também pela sua capacidade de ordenar corretamente os atletas quando o objetivo final é apoiar decisões de seleção.

Na etapa de otimização, verificou-se que as equipes construídas a partir das previsões obtiveram a maior pontuação média entre todas as estratégias avaliadas ao longo das 37 rodadas analisadas. Os resultados indicaram que a utilização dos modelos preditivos proporcionou ganho consistente em relação às estratégias baseadas exclusivamente na

média histórica ou no valor de mercado dos jogadores. Além disso, observou-se que a otimização foi capaz de identificar atletas com boa relação entre desempenho esperado e custo, produzindo equipes competitivas sem a necessidade de utilizar integralmente o orçamento disponível.

A análise da escolha do capitão mostrou que, embora a estratégia baseada nos atletas mais caros tenha produzido maior retorno médio para essa função, a qualidade da equipe como um todo exerceu influência muito mais significativa sobre a pontuação final das escalações. Esse resultado reforça que a seleção do capitão deve ser tratada como uma etapa complementar ao processo de otimização da equipe, e não como um fator isolado capaz de determinar o sucesso da estratégia.

Os estudos de caso evidenciaram a elevada variabilidade característica do futebol. Em algumas rodadas, as equipes superaram amplamente as previsões realizadas, enquanto em outras apresentaram desempenho inferior ao esperado. Esses resultados confirmam que, embora os modelos sejam capazes de capturar padrões relevantes presentes nos dados históricos, permanece um nível significativo de incerteza associado a eventos imprevisíveis, como gols, assistências, cartões, lesões, substituições e mudanças táticas ocorridas durante as partidas.

Apesar dessas limitações, a análise agregada das 37 rodadas demonstrou que a metodologia proposta apresentou desempenho consistente ao longo da temporada. Os resultados obtidos indicam que a combinação entre aprendizado de máquina e programação inteira constitui uma ferramenta promissora para apoiar decisões de escalação em ambientes caracterizados por restrições múltiplas e elevada incerteza.

Do ponto de vista científico, este trabalho contribui ao integrar diferentes etapas normalmente tratadas de forma independente na literatura. Em vez de limitar-se ao desenvolvimento de modelos preditivos, a metodologia proposta contempla todo o processo decisório, desde o tratamento dos dados e construção das variáveis até a previsão das pontuações, o ranqueamento dos atletas, a otimização da equipe e a avaliação retrospectiva das estratégias de escalação. Essa abordagem integrada permite avaliar não apenas a qualidade estatística dos modelos, mas também sua utilidade prática na tomada de decisão.

Além do contexto do Cartola FC, a metodologia desenvolvida pode ser adaptada para outros problemas de apoio à decisão que envolvam previsão e alocação ótima de recursos sob restrições, como seleção de portfólios financeiros, formação de equipes esportivas, escalonamento de profissionais, planejamento logístico e problemas de recomendação.

Como perspectivas para trabalhos futuros, recomenda-se a incorporação de novas fontes de informação capazes de representar de forma mais completa o contexto das partidas, incluindo estatísticas avançadas, características do adversário, mando de campo, informações climáticas, notícias esportivas e dados textuais obtidos por técnicas de processamento de linguagem natural. Também podem ser explorados modelos de aprendizado profundo, métodos probabilísticos capazes de estimar distribuições completas das pontuações e abordagens de aprendizado por reforço para atualização contínua das estratégias de escalação ao longo da temporada.

Outra possibilidade consiste em desenvolver um modelo específico para a escolha do capitão, considerando simultaneamente o retorno esperado e o risco associado ao desempenho dos atletas. Também podem ser investigadas formulações de otimização multiobjetivo, capazes de equilibrar a pontuação esperada, o custo, a consistência e o potencial de valorização dos jogadores. Além disso, a adoção de estratégias adaptativas, com atualização periódica dos modelos à medida que novas rodadas são disputadas, representa uma direção promissora para aumentar a robustez e a capacidade de adaptação do sistema.

Por fim, conclui-se que os objetivos propostos neste trabalho foram alcançados. A metodologia desenvolvida mostrou-se capaz de integrar previsão, ranqueamento e otimização em um único processo de apoio à decisão, produzindo equipes competitivas e resultados superiores às estratégias de referência consideradas. Embora a natureza imprevisível do futebol impeça a obtenção de previsões perfeitas, os resultados obtidos demonstram que a utilização de técnicas de Ciência de Dados e Pesquisa Operacional pode contribuir de forma significativa para melhorar a qualidade das decisões de escalação, oferecendo uma abordagem sistemática, quantitativa e reproduzível para um problema tradicionalmente baseado na experiência e na intuição.

Referências

- AKIBA, T. et al. Optuna: A next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2019. (KDD '19), p. 2623–2631.
- BONELLO, N. et al. Multi-stream data analytics for enhanced performance prediction in fantasy football. In: *Proceedings of the 27th Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2019)*. [S.l.]: CEUR-WS, 2019. (CEUR Workshop Proceedings, v. 2563), p. 284–292.
- BREIMAN, L. Random forests. *Machine Learning*, Springer, v. 45, n. 1, p. 5–32, 2001.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794.
- GROOS, D. Openfpl: An open-source forecasting method rivaling state-of-the-art fantasy premier league services. *CoRR*, abs/2508.09992, 2025.
- GUPTA, A. *Time Series Modeling for Dream Team in Fantasy Premier League*. 2019.
- MATTHEWS, T.; RAMCHURN, S. D.; CHALKIADAKIS, G. Competing with humans at fantasy football: Team formation in large partially-observable domains. In: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*. Toronto, Ontario, Canada: AAAI Press, 2012. p. 1394–1400.
- STOLYAROV, A.; VASILIEV, G. Predict to succeed: Optimal sequential fantasy football squad formation using machine learning tools. In: NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS (HSE). *Proceedings of the XVIII April International Academic Conference on Economic and Social Development*. Moscow, Russia, 2017.
- VENTER, V. Z.; VUUREN, J. H. van. An optimisation approach towards soccer fantasy premiere league team selection. *ORiON*, v. 40, n. 1, p. 69–107, 2024.