Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Monografia em Sistemas de Informação I

Relatório Final de Monografia

Caracterização e modelagem de público em jogos de futebol

Aluno: João Lucas Lage Gonçalves
Orientador: Wagner Meira Jr.

Julho de 2024

Sumário

1 Resumo

Nos últimos anos, a estrutura organizacional dos clubes de futebol no Brasil passou por uma transformação significativa, com a adoção do modelo de Sociedades Anônimas de Futebol (SAF). Esse movimento resultou em uma gestão mais profissional e focada no lucro, como evidenciado pelo aumento substancial das receitas de bilheteria. Em 2023, os clubes da Série A do Campeonato Brasileiro arrecadaram R$ 503.857.100,37 apenas com a venda de ingressos.

Nesse contexto, este trabalho visa desenvolver um modelo de previsão de público em estádios de futebol utilizando técnicas de aprendizado de máquina, com ênfase no algoritmo XGBoost, que demonstrou o melhor desempenho entre as abordagens testadas. O modelo considera variáveis preditoras como o time mandante, o período do dia, dia da semana, clima, rivalidade, a posição dos times na classificação, os resultados recentes das equipes e o momento do campeonato. No conjunto de teste, o modelo apresentou um Root Mean Squared Error (RMSE) de 7.773,53 e um coeficiente de determinação (R^2) de 0,7215.

O objetivo é fornecer uma ferramenta que ajude os clubes a entender melhor o comportamento do público e a prever a demanda de torcedores. Isso permitirá otimizar a gestão dos recursos, aumentar as receitas e reduzir os custos operacionais, maximizando os lucros por partida. O trabalho também aborda os desafios relacionados à precisão na quantificação de fatores subjetivos e na obtenção de dados precisos, ressaltando a complexidade da modelagem das interações entre diferentes variáveis contextuais.

2 Introdução

A estrutura organizacional dos clubes de futebol vem passando por uma evolução marcante nas últimas décadas. Surgindo inicialmente como associações lideradas por sócios e presidentes eleitos, muitos clubes transitaram para um modelo de propriedade privada no final do século XX. No contexto brasileiro, embora algumas instituições, como o Cuiabá Esporte Clube e o Red Bull Bragantino, já adotassem o formato de empresas antes de 2021, foi somente a partir de agosto desse ano que a transformação em Sociedades Anônimas de Futebol (SAF) foi legalmente viabilizada. Essa mudança, desencadeou uma tendência de profissionalização na gestão dos clubes, aproximando-os do modelo tradicionalmente adotado por empresas de outros setores, cujo foco principal é o lucro financeiro.

No ano de 2023, por exemplo, metade dos clubes participantes do Campeonato Brasileiro Série A já havia aderido a essa modalidade de gestão, indicando uma rápida aceitação do novo modelo em apenas dois anos. Nesse mesmo ano, o faturamento total dos 20 clubes da Série A do Campeonato Brasileiro alcançou a

expressiva marca de R$ 503.857.100,37 [3], proveniente exclusivamente da venda de ingressos. Isso equivale a uma média de R$ 25.192.855,02 por clube e R$ 1.369.176,90 por partida, considerando que 12 jogos foram realizados com portões fechados devido a punições. É importante ressaltar que esses valores se referem apenas à receita de bilheteria, não levando em conta os custos operacionais, o que significa que nem sempre os clubes obtiveram lucro significativo em cada partida.

Diante desse cenário, emerge a necessidade de estratégias mais sofisticadas de gestão, voltadas para a maximização dos lucros e o aumento da eficiência operacional dos clubes. Nesse contexto, o presente trabalho propõe retificar o desenvolvimento de um modelo de predição de público em estádios de futebol, utilizando como variáveis preditoras diversos fatores, tais como data e horário da partida, adversário, desempenho atual da equipe, importância do campeonato, preço médio do ingresso e previsão do clima para a data do jogo.

Este projeto não apenas aborda a relação entre os clubes e seus torcedores, mas também oferece uma ferramenta valiosa para aprimorar a gestão das instituições esportivas. Ao fornecer informações sobre o comportamento do público e permitir uma previsão da demanda, espera-se que os clubes possam não apenas aumentar suas receitas, mas também otimizar seus recursos e reduzir custos operacionais, maximizando, assim, seus lucros em cada partida.

Uma das principais dificuldades na modelagem desse problema reside na precisão da quantificação de fatores subjetivos, como o momento do clube, que se refere ao desempenho recente da equipe. Embora a pontuação obtida nos jogos seja uma medida simples para avaliar esse aspecto, ela nem sempre reflete com precisão a performance esportiva do time. Além disso, há outros desafios a serem enfrentados, como a obtenção de dados precisos, bem como a complexidade na modelagem das interações entre as diferentes variáveis contextuais.

Isso apresenta uma oportunidade de utilizar técnicas avançadas para capturar as interações complexas entre diversos fatores preditivos. Portanto, este trabalho propõe uma abordagem de aprendizado de máquina usando o algoritmo XGBoost aplicado à previsão de público em partidas de futebol. Foi desenvolvida uma estrutura preditiva abrangente que incorpora variáveis como data e horário da partida, adversário, desempenho da equipe, momento do campeonato, preço do ingresso e previsão do clima, permitindo que os clubes aprimorem sua gestão de receitas e eficiência operacional.

3 Referencial Teórico

A predição de público em eventos esportivos tem sido estudado sob várias perspectivas, entre as quais destacamos as seguintes.

O artigo [9] aborda o uso de regressão simbólica e programação genética (SR/GP), para determinar a função de previsão que melhor se ajusta às variáveis contextuais e independentes do painel em relação à presença em partidas de futebol. Para isso são considerados 5 temporadas de jogos no estádio Beira Rio, com jogos do campeonato brasileiro e estadual do Rio Grande do Sul. Além disso usa de especialista do meio do futebol para avaliar quais váriaveis independentes seriam usadas e suas interações. Este artigo se destaca pela identificação de fatores que influenciam na maior presença de público nos estádios, como se o jogo é um super clássico e a posição do time no campeonato sendo os principais.

O artigo [5] usa uma abordagem estatística para analisar a presença de público na primeira divisao do futebol inglês. O estudo constrói e estima um modelo levando em consideração influências comuns para aumentar a eficiência das estimativas. As variáveis independentes do modelo são divididas em três grupos principais, incluindo variáveis econômicas, demográficas/geográficas e relacionadas à atratividade do jogo. Utilizando a análise de regressão, o estudo examina a relação entre essas variáveis e as presenças nos jogos de futebol, destacando a complexidade e a variedade de fatores que influenciam as presenças. O estudo constatou que há influências comuns nas presenças nos jogos de futebol, como outras opções televisivas, condições climáticas e sazonalidade. Além disso, o estudo indica que a qualidade do time adversário também é um fator relevante para os torcedores.

O artigo [8] utiliza de técnicas de redes neurais como MLP, RNN e TLFN para modelar a presença em jogos de futebol como um exemplo de comportamento de grupo, caracterizado por um contexto complexo que só pode ser estimado por um conjunto limitado de fatores quantificáveis. O estudo compara a eficiência dessas diferentes arquiteturas de redes neurais utilizando dados de três equipes da segunda divisão do campeonato inglês de futebol. O trabalho demonstra que modelos neurais construídos a partir de conjuntos de dados individuais apresentam melhor desempenho do que um modelo neural generalizado construído a partir de dados agrupados, indentificando o momento do campeonato, a classificação do time mandante, o time rival e sua classificação como as principais variáveis relevantes para o modelo.

O artigo [7] tem como objetivo central identificar os motivos que influenciam a presença dos torcedores nos estádios de futebol. Para isso foi aplicado um questionário para uma amostra de 329 pessoas que já tinham ido a estádios de

futebol na cidade de São Paulo. Foi utlizada a Análise Fatorial Exploratória para analisar os dados e o coeficiente alfa de Cronbach para medir a confiabilidade dos fatores extraídos sendo identificados cinco fatores que representaram 73,89% da variância total: 1) motivações de entretenimento e sócio-psicológicos; 2) preço; 3) ações de marketing; 4) qualidade do futebol; e 5) prazer de assistir a uma partida de futebol no estádio.

O artigo [2] tem como objetivo analisar e prever o consumo de ingressos para jogos de futebol em estádios brasileiros, usando uma abordagem de três modelos de regressão: um modelo de regressão OLS usual com erros normalmente distribuídos; um modelo de regressão com variável dependente censurada (TOBIT); e modelos lineares generalizados (GLM) com distribuição gama para melhor ajustar a assimetria positiva da distribuição de consumo. Os modelos incluem variáveis explicativas relacionadas ao ambiente econômico, qualidade do produto e incentivos monetários e não monetários para comparecer aos eventos esportivos. Os resultados mostram que a maioria dessas variáveis é estatisticamente significativa para explicar a quantidade de pessoas que vão aos estádios. Dos 3 modelos ajustados, o modelo linear generalizado com distribuição gama apresentou melhores resultados para prever o consumo de ingressos para jogos do campeonato brasileiro, em comparação com o benchmark.

O artigo [10] utiliza uma modelagem matemática para precificar dinamicamente o valor dos ingressos se baseando na alteração de preços por meio de multiplicadores ao longo do tempo. O algoritmo de otimização busca encontrar os valores ótimos desses multiplicadores e utiliza o preço médio da temporada, combinando com um modelo de lógica difusa para predição de público. De acordo com os resultados do modelo de precificação dinâmica, a receita total gerada aumenta em 8,95% em comparação com a estratégia de precificação estática. Também é a primeira vez que este tipo de modelo matemático de precificação dinâmica para ingressos de jogos de futebol foi projetado.

O artigo [1] apresenta um modelo de previsão de público utilizando regressão linear para jogos do Botafogo de Futebol e Regatas. A pesquisa analisou uma série de variáveis, incluindo o desempenho do time, a qualidade do adversário, o horário da partida e contexto do campeonato. Os resultados indicam que o modelo de regressão linear pode prever de forma eficiente a presença de público nos jogos, destacando a relevância do desempenho do time e da atratividade do adversário como principais fatores influentes. Este trabalho contribui para a literatura ao demonstrar a aplicabilidade de métodos estatísticos simples para a previsão de público em eventos esportivos.

Dessa forma, este trabalho propõe o uso de técnicas de *boosting*, especificamente o algoritmo XGBoost, para a previsão de público em partidas de futebol. As

	DATA	PRECIPITACAO_TOTAL	TEMPERATURA_MAXIMA	TEMPERATURA_MEDIA	TEMPERATURA_MINIMA	CIDADE	LATITUDE	LONGITUDE
0	2012-01-01	7.4	28.5	23.225000	20.7	NIQUELANDIA	-14.469444	-48.485833
1	2012-01-02	19.8	23.3	22.108333	21.0	NIQUELANDIA	-14.469444	-48.485833
2	2012-01-03	19.2	24.2	21.187500	20.5	NIQUELANDIA	-14.469444	-48.485833
3	2012-01-04	NaN	25.0	NaN	19.4	NIQUELANDIA	-14.469444	-48.485833
4	2012-01-05	0.6	27.4	21.245833	18.6	NIQUELANDIA	-14.469444	-48.485833
5	2012-01-06	16.4	26.9	22.470833	20.0	NIQUELANDIA	-14.469444	-48.485833
6	2012-01-07	NaN	27.2	NaN	19.8	NIQUELANDIA	-14.469444	-48.485833
7	2012-01-08	0.2	28.0	22.963636	19.9	NIQUELANDIA	-14.469444	-48.485833
8	2012-01-09	0.0	27.1	22.286957	20.6	NIQUELANDIA	-14.469444	-48.485833
9	2012-01-10	NaN	NaN	NaN	NaN	NIQUELANDIA	-14.469444	-48.485833

Tabela 1: Tabela com dados climáticos históricos, incluindo precipitação total diária, temperaturas máxima, média e mínima diárias, coletados entre 2012 e 2024

técnicas de *boosting* [4] se destacam por sua capacidade de combinar vários modelos fracos para formar um modelo forte, resultando em previsões mais precisas e robustas.

4 Metodologia

4.1 Dados

Os dados contextuais, de público e de renda das partidas disputadas entre 2014 e 2023 dos 20 clubes que disputaram a Série A do Campeonato Brasileiro em 2023, totalizando 2480 jogos, foram coletados via *web scraping* do site srgoool.com.br. A coleta incluiu informações sobre o campeonato, rodada, classificação, e os detalhes dos jogos. Os dados foram organizados em dois conjunto de dados distintos: um contendo as informações detalhadas das partidas exibidos na tabela 2, e outro contendo as informações de classificação dos clubes ao longo das temporadas exibidos na tabela 3.

Além disso, para incluir dados contextuais, foram coletados dados históricos de clima a partir do Banco de Dados Meteorológicos do Instituto Nacional de Meteorologia (INMET), obtendo dados diários meteorológicos exibidos na tabela 1.

Após o tratamento dos dados, construção de características e exclusão de dados nulos, o conjunto de dados foi reduzido para 1709 partidas. Esse processamento garantiu a qualidade e a consistência dos dados, removendo entradas incompletas, além de criar novas variáveis que poderiam ser relevantes para a modelagem preditiva. Como resultado, os dados finais ficaram prontos para serem utilizados na análise e na construção dos modelos de predição de público e renda dos jogos.

4.2 Atributos

Para a escolha dos atributos finais, foram realizados vários testes utilizando como base nos atributos empregadaos pelos trabalhos relacionados. A partir dessa

	TEMPORADA	RODADA	MANDANTE	PLACAR	VISITANTE	DATA	HORARIO	ESTADIO	CIDADE	PUBLICO_PAGANTE	INGRESSOS_DISPONIVEIS	OCUPACAO	RENDA_LIQUIDA	RENDA_BRUTA	EVENTO
0	2023	38	Goiás-GO	1 x 0	América Mineiro-MG	06/12/2023	19h00	Hailé Pinheiro (Serrinha)	Goiânia	793	793	100	-72.885,55	19.000,00	NaN
1	2023	38	Fluminense-RJ	2 x 3	Grêmio-RS	06/12/2023	21h30	Maracanã	Rio de Janeiro	40.764	49.652	82	300.700,41	1.635.410,50	NaN
2	2023	38	Vasco-RJ	2 x 1	Bragantino-SP	06/12/2023	21h30	São Januário	Rio de Janeiro	19.729	20.708	95	374.435,91	1.009.864,00	NaN
3	2023	38	São Paulo-SP	1 x 0	Flamengo-RJ	06/12/2023	21h30	Morumbi	São Paulo	36.618	36.618	100	1.767.709,26	2.414.279,00	NaN
4	2023	38	Santos-SP	1 x 2	Fortaleza-CE	06/12	21h30	Vila Belmiro	Santos	14.130	14.130	100	250.972,81	708.607,50	Santos está rebaixado na Série A
5	2023	38	Cruzeiro-MG	1 x 1	Palmeiras-SP	06/12	21h30	Mineirão	Belo Horizonte	37.724	37.724	100	1.028.285,42	1.675.735,00	PALMEIRAS - CAMPEÃO DO BRASILEIRÃO 2023
6	2023	38	Internacional-RS	3 x 1	Botafogo-RJ	06/12/2023	21h30	Beira-Rio	Porto Alegre	27.844	27.844	100	809.908,62	1.148.623,00	NaN
7	2023	38	Coritiba-PR	0 x 2	Corinthians-SP	06/12/2023	21h30	Couto Pereira	Curitiba	NaN	NaN	NaN	NaN	NaN	Portões fechados
8	2023	38	Bahia-BA	4 x 1	Atlético Mineiro-MG	06/12/2023	21h30	Arena Fonte Nova	Salvador	27.743	27.743	100	254.723,85	838.345,00	NaN
9	2023	38	Cuiabá-MT	3 x 0	Athletico Paranaense-PR	06/12/2023	21h30	Arena Pantanal	Cuiabá	8.181	8.181	100	-185.267,33	101.735,00	NaN

Tabela 2: Tabela com dados históricos do Campeonato Brasileiro, incluindo todas as partidas disputadas entre 2012 e 2024

	TEMPORADA	RODADA	POSICAO	TIME	PONTOS	JOGOS	VITORIAS	EMPATES	DERROTAS	GP	GC	SG	APROVEITAMENTO
0	2023	38	1º	Palmeiras-SP	70	38	20	10	8	64	33	31	61,4
1	2023	38	2º	Grêmio-RS	68	38	21	5	12	63	56	7	59,6
2	2023	38	3º	Atlético Mineiro-MG	66	38	19	9	10	52	32	20	57,9
3	2023	38	4º	Flamengo-RJ	66	38	19	9	10	56	42	14	57,9
4	2023	38	5º	Botafogo-RJ	64	38	18	10	10	58	37	21	56,1
5	2023	38	6º	Bragantino-SP	62	38	17	11	10	49	35	14	54,4
6	2023	38	7º	Fluminense-RJ	56	38	16	8	14	51	47	4	49,1
7	2023	38	8º	Athletico Paranaense-PR	56	38	14	14	10	51	43	8	49,1
8	2023	38	9º	Internacional-RS	55	38	15	10	13	46	45	1	48,2
9	2023	38	10º	Fortaleza-CE	54	38	15	9	14	45	44	1	47,4

Tabela 3: Tabela com dados históricos de classificação do Campeonato Brasileiro rodada a rodada entre 2012 e 2024

análise, foram selecionadas e combinadas diferentes variáveis para identificar aquelas que apresentavam os melhores resultados nos algoritmos de *boosting*. Essa abordagem garantiu a inclusão de variáveis contextuais e específicas que capturam de maneira eficaz as interações complexas que influenciam a presença de público nos jogos de futebol. Os atributos selecionados foram:

1. **Time Mandante:** O time da casa na partida, codificado através de one-hot encoding para transformá-lo em uma variável categórica.
2. **Período do dia:** O horário da partida foi categorizado em manhã (horário de início antes das 13h), tarde (horário de início entre 13h e 18h30) e noite (horário de início a partir das 18h30). Essa variável foi transformada em categórica usando one-hot encoding, fornecendo um indicativo do contexto da partida.
3. **Fim de Semana:** Variável binária que indica se o jogo ocorreu durante o final de semana (sábado ou domingo). A expectativa é que partidas realizadas em finais de semana atraem maior público, já que muitas pessoas não trabalham nem estudam nesses dias, servindo como um indicativo de contexto da partida.
4. **Feriado:** Variável binária que indica se a partida ocorreu em um dia que é feriado, ou no dia anterior ou seguinte a um feriado. Essa variável serve como um indicativo de contexto da partida.
5. **Temperatura Média Normalizada:** Valor absoluto da temperatura média no dia da partida, normalizada entre os dados disponíveis.

Essa variável é um indicativo de contexto da partida.

6. **Chuva:** Variável binária que indica se houve chuva no dia da partida, servindo como um indicativo de contexto da partida.
7. **Rival Local:** Variável binária que indica se os dois clubes que disputam a partida são do mesmo estado. Essa variável indica a relação geográfica entre os clubes.
8. **Rival Nacional:** variável binária indicando se o clube adversário está na lista dos considerados "12 grandes"do futebol brasileiro, indicativo de relação entre os clubes
9. **Pts Diferença:** Diferença de pontos entre o time mandante e o time visitante, fornecendo uma métrica da vantagem competitiva.
10. **Libertadores Mandante:** Variável binária que indica se o clube mandante está na zona de classificação (primeiras 4 posições) para a Copa Libertadores. Essa variável é um indicativo do desempenho do clube.
11. **Libertadores Visitante:** Variável binária que indica se o clube visitante está na zona de classificação para a Copa Libertadores. Essa variável é um indicativo do desempenho do clube.
12. **Rebaixamento Mandante:** Variável binária que indica se o clube mandante está na zona de rebaixamento (últimas 4 posições) para a 2ª divisão do Campeonato Brasileiro. Essa variável é um indicativo do desempenho do clube.
13. **Rebaixamento Visitante:** Variável binária que indica se o clube visitante está na zona de rebaixamento para a 2ª divisão do Campeonato Brasileiro. Essa variável é um indicativo do desempenho do clube.
14. **Último Resultado:** Indica a quantidade de pontos obtidos pelo time visitante na última partida (3 pontos para vitória, 1 ponto para empate, 0 pontos para derrota). Essa variável serve como um indicativo do desempenho recente.
15. **3 Últimos Resultados:** Soma dos pontos obtidos nos últimos 3 jogos pelo time visitante. Essa variável fornece um indicativo do desempenho recente acumulado.
16. **PPG Mandante:** Número de pontos por jogo obtido pelo time mandante. Essa variável é um indicativo do desempenho geral do time mandante.

17. **Momento Campeonato:** Rodada a ser disputada normalizada entre 0 (primeira rodada) e 1 (última rodada).

4.3 Modelo

Para a modelagem da predição de público, foi escolhida o paradigma de *boosting* devido à sua eficácia em combinar múltiplos modelos fracos para formar um modelo forte e robusto. Os algoritmos de *boosting* funcionam iterativamente, ajustando erros de modelos anteriores e, assim, melhorando continuamente a precisão das previsões.

Foram testados vários algoritmos de *boosting*: XGBoost, AdaBoost, CatBoost e Random Forest. Esses algoritmos foram escolhidos devido à sua capacidade de combinar múltiplos modelos fracos para formar um modelo forte e robusto, melhorando iterativamente a precisão das previsões. Utilizou-se a técnica de Grid Search para selecionar os melhores parâmetros para cada algoritmo. A Grid Search é um método de busca exaustiva que testa todas as combinações possíveis de hiperparâmetros especificados para encontrar a configuração que proporciona o melhor desempenho do modelo. Após a otimização dos hiperparâmetros, os resultados de cada algoritmo foram comparados. O XGBoost se destacou, apresentando os melhores resultados tanto no Root Mean Squared Error (RMSE) quanto no coeficiente de determinação (R^2).

Para o XGBoost, os parâmetros testados incluíram a fração de colunas a serem amostradas para cada árvore (`colsample_bytree`), a taxa de aprendizado (`learning_rate`), a profundidade máxima das árvores (`max_depth`), o número de estimadores (`n_estimators`) e a fração de amostras utilizadas para treinar cada árvore (`subsample`). O grid search foi executado com uma validação cruzada de 5 vezes (*5-fold cross-validation*) e utilizando a métrica de erro quadrático médio negativo (*neg_mean_squared_error*) como critério de avaliação. Os melhores hiperparâmetros obtidos para o modelo XGBoost foram `colsample_bytree` igual a 0.8, `learning_rate` de 0.01, `max_depth` de 10, `n_estimators` de 1000 e `subsample` de 0.8.

5 Resultados

Os resultados obtidos pelo XGBoost na tarefa de predição de público mostraram que, apesar dos desafios envolvidos, o modelo conseguiu capturar algumas das relações complexas entre as diferentes características. No conjunto de teste, o modelo apresentou um *Root Mean Squared Error* (RMSE) de 7.773,53 e um coeficiente de determinação (R^2) de 0,7215. Para avaliar a eficácia do modelo de *boosting*, seus resultados foram comparados com os obtidos usando uma

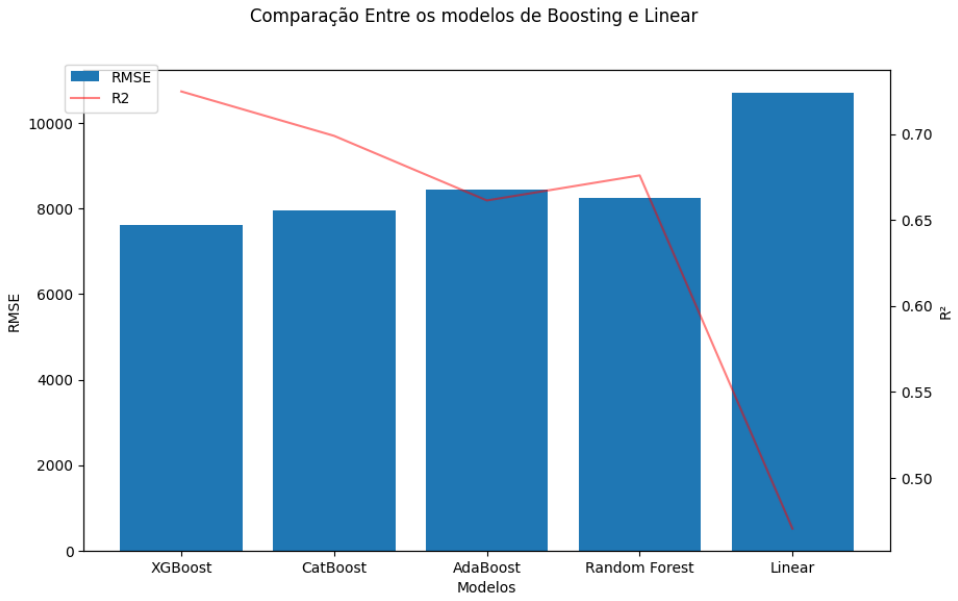

Figura 1: Comparação de desempenho entre modelos de *Boosting* (XGBoost, CatBoost, AdaBoost) e modelos Random Forest e Linear utilizando RMSE e R^2

regressão linear básica. O modelo de regressão linear apresentou um RMSE de 10.718,89 e um R^2 de 0,4705.

Esses resultados indicam que, embora a tarefa de predição de público seja complexa, o XGBoost superou a regressão linear básica em termos de precisão, demonstrando uma melhor capacidade de capturar as relações complexas entre as variáveis preditivas. No entanto, é importante notar que ambos os modelos têm espaço para melhorias, especialmente com a inclusão de mais e melhores dados. A tarefa de predição de público em jogos de futebol é particularmente desafiadora devido à variabilidade e à quantidade limitada de dados disponíveis. Muitos fatores influenciam a presença do público, incluindo aspectos econômicos, sociais, culturais e climáticos, e nem todos esses fatores são facilmente quantificáveis ou disponíveis para análise. Portanto, apesar dos resultados razoáveis obtidos com o XGBoost, a precisão das previsões pode ser significativamente melhorada com a expansão do conjunto de dados e a inclusão de variáveis adicionais que capturem melhor a complexidade do comportamento do público.

Para compreender melhor a importância e o impacto de cada feature na predição de público, foi utilizado o método SHAP (SHapley Additive exPlanations) [6]. O SHAP é uma técnica para a interpretação de modelos de aprendizado de máquina, que atribui um valor de importância a cada feature com base na teoria dos valores de Shapley, oriunda da teoria dos jogos. Esse método permite explicar as previsões do modelo ao distribuir de maneira justa o impacto de cada atributo em todas as possíveis combinações.

A característica mais significativa é o Preço Médio Ajustado, que apresenta uma ampla distribuição de valores SHAP tanto positivos quanto negativos, indicando

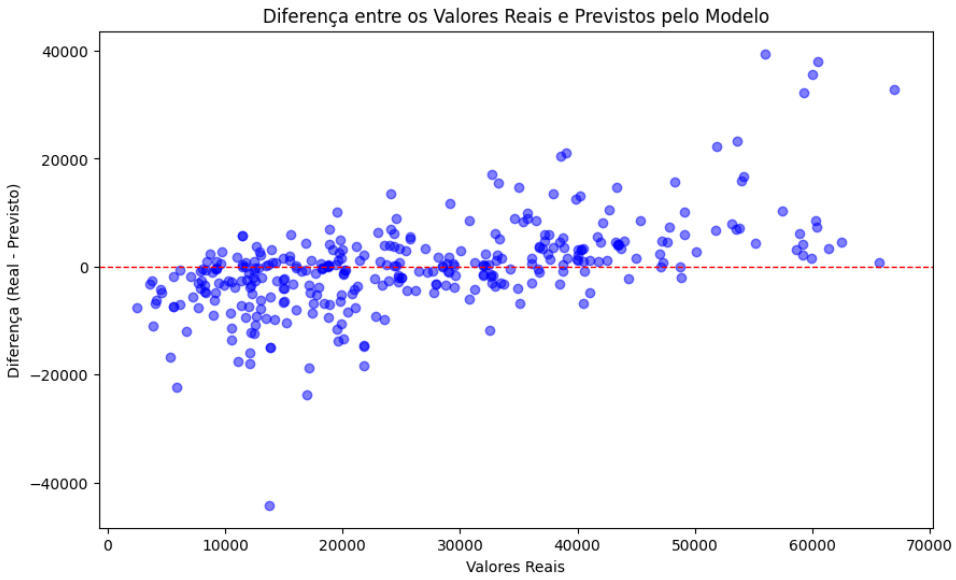

Figura 2: Diferença entre os valores reais e previstos pelo modelo nos dados de teste.

um impacto substancial e diversificado no modelo. Entre as características dos mandantes, Santos-SP e Flamengo-RJ tendem a influenciar positivamente as previsões, enquanto São Paulo-SP e Corinthians-SP exibem uma variação maior em ambos os sentidos.

Os fatores climáticos e contextuais, como Temperatura Média Normalizada e Chuva, junto com Noite, Fim de Semana e Momento do Campeonato, também têm impactos variados nas previsões. Essas características contextuais influenciam significativamente as previsões do modelo, refletindo condições extremas e diferentes momentos do campeonato, como finais de semana, horário dos jogos e condições climáticas. A presença de Rival Local e Rival Nacional é igualmente relevante, sugerindo que a rivalidade, tanto local quanto nacional, exerce um impacto importante nas previsões do modelo.

Quanto ao desempenho das equipes, características como Pontos Por Jogo (PPG), Pontos de Diferença, Último Resultado e Últimos 3 resultados demonstram relevância significativa, especialmente no que tange ao desempenho recente e a diferença de pontos, indicando a importância do estado de forma das equipes nas previsões.

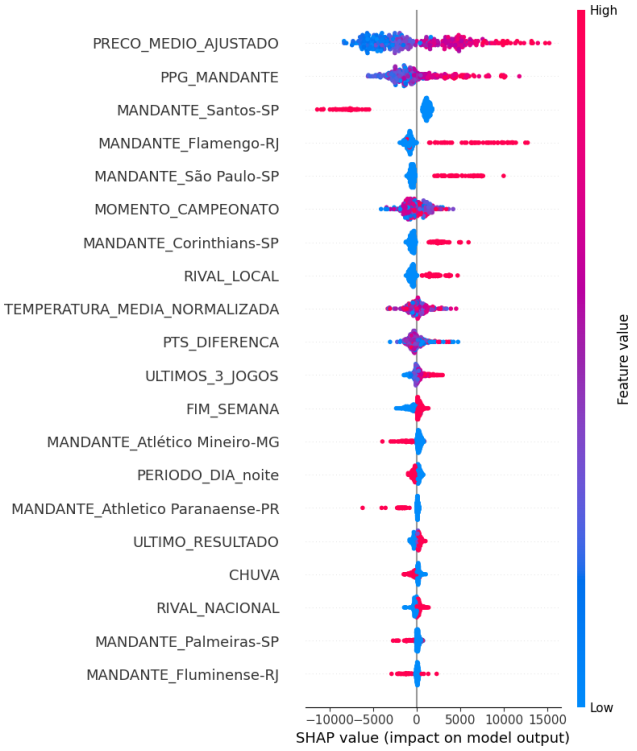

Figura 3: Importância dos atributos utilizadas pelo modelo usando o método SHAP.

6 Conclusão

Este trabalho apresentou um modelo de predição de público para jogos de futebol utilizando técnicas de aprendizado de máquina, com ênfase no algoritmo XGBoost. A abordagem proposta se mostrou eficaz em capturar as complexas relações entre as variáveis preditoras, tais como data e horário da partida, adversário, desempenho atual da equipe, importância do campeonato, preço médio do ingresso e previsão do clima. Os resultados indicam que o modelo desenvolvido superou a regressão linear básica. Isso sugere que a técnica de *boosting* é uma ferramenta poderosa para essa tarefa, podendo ser utilizada pelos clubes para otimizar a gestão dos recursos, aumentar as receitas e reduzir os custos operacionais.

Os atributos utilizados tiveram um impacto significativo na performance do modelo. O Preço Médio Ajustado foi a característica mais significativa, com uma ampla distribuição de valores SHAP, indicando um impacto substancial e diversificado no modelo. Fatores como Temperatura Média Normalizada, Chuva, Noite, Fim de Semana e Momento do Campeonato mostraram impactos variados, refletindo condições extremas e diferentes momentos do campeonato. A presença de Rival Local e Rival Nacional também exerceu um impacto importante nas previsões do modelo. Quanto ao desempenho das equipes, características como Pontos Por Jogo (PPG) e Último Resultado demonstraram relevância significativa, indicando a importância do estado de forma das equipes nas previsões.

Além dos benefícios práticos para a gestão dos clubes, este estudo também contribui para a literatura ao explorar a aplicação de técnicas de *boosting* no contexto de predição de público em eventos esportivos. Os desafios identificados, como a quantificação precisa de fatores subjetivos e a obtenção de dados acurados, ressaltam a complexidade da modelagem preditiva e a necessidade de contínua evolução e refinamento dos modelos.

Como extensão do trabalho atual, é importante testar a aplicação de novas técnicas e focar em dados de um único clube para criar um modelo mais específico. Ao concentrar a análise em um clube específico, é possível utilizar dados mais detalhados e contextualizados, o que pode resultar em previsões mais precisas. Essa abordagem permitirá uma modelagem mais detalhada e uma compreensão mais profunda dos fatores que influenciam a presença dos torcedores nos jogos, facilitando a identificação de estratégias eficazes para aumentar a frequência nos estádios.

De forma geral, para futuras pesquisas, o uso de dados mais detalhados, como o histórico de compra de ingressos, informações sobre sócios-torcedores e interações

em redes sociais, pode enriquecer significativamente o modelo. Além disso, a aplicação de técnicas de análise de sentimentos e processamento de linguagem natural para avaliar o impacto de notícias e eventos sobre a expectativa de público pode proporcionar informações valiosas. Dessa forma, buscar soluções personalizadas para cada clube, considerando suas características únicas e o comportamento de sua torcida, pode aumentar a precisão das previsões e oferecer uma compreensão mais profunda dos fatores que influenciam a presença nos estádios. Explorar essas novas variáveis pode ajudar a capturar nuances que não são evidentes nos dados agregados, melhorando a capacidade do modelo de prever a presença do público em diferentes contextos.

Em resumo, o desenvolvimento de um modelo preditivo de público em estádios de futebol, utilizando técnicas de aprendizado de máquina, mostrou-se viável e promissor, oferecendo uma ferramenta valiosa para a profissionalização e eficiência na gestão dos clubes de futebol no Brasil.

Referências

[1] Caio Fernando dos Santos Araujo. Previsão de público com modelo de regressão linear para jogos do botafogo de futebol e regatas. 2018.

[2] Adriana Bortoluzzo, Maurício Bortoluzzo, Sérgio Machado, Tatiana Melhado, Pedro Trindade, and Bruno Pereira. Ticket consumption forecast for brazilian championship games. *Revista de Administração*, 52, 10 2016.

[3] Pedro Melo Erich Beting. Exclusivo: Brasileirão 2023 faturou meio bilhão em bilheteria. *Máquina do Esporte*, 2023.

[4] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[5] Robert Hart, J Hutton, and Trevor Sharot. A statistical analysis of association football attendances. *Applied Statistics*, 24:308, jan 1975.

[6] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[7] Regina Madalozzo and Rodrigo Villar. Brazilian football: What brings fans to the game? *Journal of Sports Economics - J SPORT ECON*, 10:639–650, 12 2009.

[8] Damjan Strnad, Andrej Nerat, and Štefan Kohek. Neural network models for group behavior prediction: a case of soccer match attendance. *Neural Computing and Applications*, 28, 09 2015.

[9] Gabrielli H. Yamashita, Flavio S. Fogliatto, Michel J. Anzanello, and Guilherme L. Tortorella. Customized prediction of attendance to soccer matches based on symbolic regression and genetic programming. *Expert Systems with Applications*, 187:115912, 2022.

[10] Mehmet Şahin and Rizvan Erol. A dynamic ticket pricing approach for soccer games. *Axioms*, 6:31, 11 2017.