CODERHOUSE

ISADORA FERRAZ E FIGUEIREDO
PEDRO HENRIQUE GONÇALVES DA SILVA FERREIRA

Estratégia com Dados: Modelagem de Performance no Basquete

BRASIL
2024

SUMÁRIO

1 INTRODUÇÃO

O basquete é um dos esportes mais populares nos Estados Unidos, com a NBA (National Basketball Association) sendo uma das ligas esportivas mais lucrativas e influentes do mundo. Durante as partidas, uma vasta quantidade de dados sobre o desempenho dos jogadores é registrada, proporcionando uma base rica para análises estatísticas e aplicações em ciência de dados. Aproveitando esse contexto, este projeto tem como objetivo desenvolver um modelo preditivo para analisar o desempenho de jogadores da NBA, utilizando técnicas de ciência de dados e aprendizado de máquina.

A proposta inicial busca identificar as variáveis que influenciam o desempenho de um atleta de alta performance e, a partir dessa análise, criar um algoritmo capaz de estimar a pontuação dos jogadores em partidas futuras. Essa iniciativa combina métodos avançados de análise com uma abordagem baseada em dados, destacando a relevância de decisões orientadas por estatísticas no contexto esportivo.

2 CASE

O objetivo principal deste projeto é analisar o desempenho do jogador Stephen Curry, um dos maiores atletas da NBA, e criar um modelo preditivo capaz de estimar sua pontuação em partidas futuras. A análise baseia-se em uma rica base de dados que reúne informações detalhadas sobre mais de 900 jogos de sua carreira. Entre as variáveis consideradas estão assistências, rebotes, roubos de bola, bloqueios e minutos jogados, além de fatores externos como o adversário enfrentado, o local da partida e a data do jogo.

A NBA, organiza sua temporada em duas etapas distintas: a temporada regular e os playoffs. A temporada regular, com 82 rodadas, ocorre entre os meses de setembro e maio, sendo uma fase crucial para determinar as equipes classificadas para os playoffs. Os playoffs consistem em séries eliminatórias que definem o campeão da liga. Neste contexto, o desempenho individual de jogadores como Stephen Curry desempenha um papel essencial, influenciando diretamente o sucesso de sua equipe nas fases decisivas.

O foco deste estudo inicial é entender os padrões de desempenho de Curry ao longo do período regular das temporadas. A criação do modelo preditivo busca também fornecer uma ferramenta que permita análises estratégicas sobre o desempenho do atleta em diferentes condições de jogo.

2.1 Desafio

O principal desafio identificado neste projeto foi evitar a dependência circular entre as variáveis utilizadas no modelo e o evento que ele deveria prever. Variáveis diretamente ligadas à performance individual em uma partida, como assistências e rebotes, mostraram-se inadequadas, pois poderiam comprometer a capacidade de generalização e previsibilidade do modelo.

Para superar essa limitação, foram revisadas as variáveis utilizadas, priorizando aquelas independentes do evento específico. Além disso, foi aplicada feature engineering para enriquecer o conjunto de dados, garantindo maior eficácia do modelo preditivo.

Com essa estratégia, o modelo se torna menos dependente de dados de performance in-game, ganhando maior capacidade de prever pontuações futuras de maneira confiável e aplicável em diversas situações.

3 DATA ACQUISITION

A base de dados utilizada neste projeto foi adquirida na plataforma STATHEAD BASKETBALL, uma ferramenta confiável que oferece estatísticas detalhadas de partidas e jogadores da NBA. O link para acesso à plataforma está disponível em STATHEAD.

A partir dessa plataforma, foi gerado um arquivo no formato CSV, que serviu como ponto de partida para o projeto. Inicialmente, colunas diretamente relacionadas ao de-

sempenho individual dos jogadores em quadra foram removidas, como Game Started, Field Goal, 3 Points e Free Throw. A exclusão dessas colunas foi necessária para evitar redundâncias e dependência circular no modelo preditivo. A lista completa de colunas removidas inclui:

- GameStarted, FieldGoal, FieldGoalAttempt, FieldGoal%, 2Points, 2PointsAttempt, 2PointsAttempt%, 3Points, 3PointsAttempt, 3Points%, FreeThrow, FreeThrowAttempt, FreeThrow%, TrueShot%, OffensiveRebounds, DefensiveRebounds, PersonalFault, PTS.1, GameScore, BoxPointMin, +/- , Pos, Player-additional, Result.

Após a exclusão dessas colunas, a base de dados passou por um processo de tratamento detalhado, descrito na seção de Data Wrangling, para garantir sua qualidade e adequação às análises propostas. A Figura 1 apresenta os dados brutos antes do início do tratamento.

4 DATA WRANGLING

O processo de data wrangling incluiu etapas essenciais para garantir a qualidade e a adequação dos dados para análise. Inicialmente, foi realizado um levantamento para compreender o significado de cada variável, identificando quais precisavam de tratamento e quais poderiam ser descartadas.

	Rk	Player	PTS	Date	Age	Team	Unnamed: 6	Opp	Result	GS	...	STL	BLK	TOV	PF	PTS.1	GmSc	BPM	+/-	Pos.	Player-additional
0	1	Stephen Curry	62	2021-01-03	32-295	GSW	NaN	POR	W 137-122	*	...	0	0	5	0	62	46.8	18.6	20	G	curryst01
1	2	Stephen Curry	60	2024-02-03	35-326	GSW	@	ATL	L 134-141 (OT)	*	...	0	1	2	1	60	45.5	22.4	-1	G	curryst01
2	3	Stephen Curry	57	2021-02-06	32-329	GSW	@	DAL	L 132-134	*	...	1	0	3	3	57	43.8	18.3	7	G	curryst01
3	4	Stephen Curry	54	2013-02-27	24-350	GSW	@	NYK	L 105-109	*	...	3	0	4	3	54	46.1	23.0	-4	G	curryst01
4	5	Stephen Curry	53	2015-10-31	27-231	GSW	@	NOP	W 134-120	*	...	4	0	2	3	53	49.2	27.8	16	G	curryst01
5	6	Stephen Curry	53	2021-04-12	33-029	GSW	NaN	DEN	W 116-107	*	...	0	0	5	3	53	39.8	20.9	16	G	curryst01
6	7	Stephen Curry	51	2015-02-04	26-327	GSW	NaN	DAL	W 128-114	*	...	1	0	3	3	51	39.2	19.5	0	G	curryst01
7	8	Stephen Curry	51	2016-02-03	27-326	GSW	@	WAS	W 134-121	*	...	3	0	7	3	51	37.3	17.6	20	G	curryst01
8	9	Stephen Curry	51	2016-02-25	27-348	GSW	@	ORL	W 130-114	*	...	0	1	5	1	51	43.1	24.9	11	G	curryst01
9	10	Stephen Curry	51	2018-10-24	30-224	GSW	NaN	WAS	W 144-122	*	...	0	1	2	1	51	41.8	23.0	19	G	curryst01

10 rows × 38 columns

Figura 1 – Dados brutos

1. **Remoção de valores nulos e duplicados:** Garantiu a consistência e qualidade do conjunto de dados.
2. **Transformação da coluna de data:** Extraídas informações como mês
3. **Cálculo e ajuste da idade:** A idade do jogador, originalmente apresentada com anos e dias, foi arredondada para o número inteiro inferior.
4. **Renomeação e transformação da coluna 'Unnamed: 6':**
   - A coluna foi renomeada para 'Local' para maior clareza.
   - Transformada em uma variável booleana, onde:

- **0** representa jogos fora de casa.
- **1** representa jogos em casa.

	PTS	Date	Age	Local	Opp	MP	TRB	AST	STL	BLK	TOV
804	14	2009-10-28	21	0	HOU	36	2	7	4	0	2
855	12	2009-10-30	21	1	PHO	39	2	4	1	0	3
926	7	2009-11-04	21	0	MEM	28	5	9	2	0	1
938	5	2009-11-06	21	0	LAC	22	1	3	0	0	0
899	9	2009-11-08	21	1	SAC	31	4	6	0	0	5

Figura 2 – Dados pós tratamento

5 FEATURE ENGINEERING

No processo de feature engineering, foram criadas variáveis para enriquecer o conjunto de dados e capturar informações contextuais relevantes para o problema. As novas variáveis adicionadas incluem:

- **Temporada:** Identifica a temporada à qual cada jogo pertence, permitindo analisar padrões de desempenho ao longo dos anos.
- **Rodada:** Numeração sequencial dos jogos dentro de cada temporada, facilitando o acompanhamento da evolução do desempenho ao longo do tempo.
- **Mês:** Indica o mês em que o jogo foi realizado, ajudando a identificar padrões sazonais.
- **Dia da Semana:** Representa o dia da semana em que o jogo ocorreu, capturando possíveis variações de desempenho associadas ao calendário.
- **Parte:** Divide a temporada em duas fases:
  - **1ª parte:** Inclui jogos de setembro a dezembro.
  - **2ª parte:** Inclui jogos do início do ano até o final da temporada.

Essas variáveis foram criadas a partir da coluna data, com o objetivo de adicionar informações contextuais sem depender diretamente de métricas de desempenho específicas do jogador, garantindo maior robustez na análise e na modelagem preditiva.

```python
# Função para criar coluna temporada
def calcular_temporada(data):
    ano = data.year
    mes = data.month
    if mes >= 9:
        return (ano - 2009) + 1
    else:
        return (ano - 2009)

df['Temporada'] = df['Date'].apply(calcular_temporada)
```

Figura 3 – Criação da coluna Temporada

6 DICIONÁRIO DE DADOS

Abaixo segue um dicionário para a melhor compreensão do problema:

Variável	Tipo	Descrição
PTS	Numérica	Pontos marcados pelo jogador em um jogo
Date	Data	Data do jogo.
Age	Numérica	Idade do jogador no momento do jogo.
Local	Categórica	Indica se o jogo foi em casa (1) ou fora (0)
Opp	Categórica	Time adversário
MP	Numérica	Minutos jogados pelo jogador
TRB	Numérica	Total de rebotes do jogador no jogo
AST	Numérica	Assistências feitas pelo jogador
STL	Numérica	Roubos de bola pelo jogador
BLK	Numérica	Bloqueios feitos pelo jogador
TOV	Numérica	Número de turnovers cometidos pelo jogador
Temporada	Categórica	Temporada à qual o jogo pertence
Parte	Categórica	precisa achar uma forma de descrever
Rodada	Categórica	Número da rodada do jogo na temporada
Day_of_Week	Categórica	Dia da semana em que o jogo ocorreu
Month	Categórica	Mês em que o jogo ocorreu

7 ANÁLISE EXPLORATÓRIA DE DADOS (EDA)

A Análise Exploratória de Dados (EDA) é uma etapa essencial para compreender as principais características de um conjunto de dados. Por meio de visualizações gráficas, estatísticas descritivas e testes de hipóteses, a EDA identifica padrões, outliers, correlações e inconsistências, oferecendo insights que guiam as próximas etapas da análise ou modelagem.

Com base no entendimento inicial do problema, algumas hipóteses foram levantadas para direcionar a investigação e avaliar fatores relacionados ao desempenho do jogador:

- **Idade e Pontuação:** A pontuação média do jogador tende a diminuir com o avanço da idade devido ao desgaste físico acumulado.
- **Desempenho em Casa vs Fora de Casa:** O rendimento do jogador é superior em partidas realizadas em casa em comparação com as realizadas fora.
- **Intervalo entre Jogos e Desempenho:** Devido ao curto intervalo entre jogos, o jogador tende a se poupar em certas partidas ou apresentar variações de desempenho dependendo do dia da semana.
- **Diferenças entre Partes da Temporada:** A pontuação pode variar de forma significativa entre a primeira parte (setembro a dezembro) e a segunda parte (janeiro ao fim da temporada) devido a diferenças de ritmo e condições.
- **Adversário e Pontuação:** O desempenho do jogador pode estar associado à força ou ao estilo de jogo do adversário.
- **Minutos Jogados e Pontuação:** Existe uma relação direta entre os minutos jogados e a quantidade de pontos, indicando que mais tempo em quadra pode resultar em maior pontuação.

Essas hipóteses foram fundamentais para estruturar a análise e compreender os principais fatores que influenciam o desempenho do jogador. Cada uma delas foi explorada durante a EDA com o objetivo de validar ou refutar as suposições levantadas.

7.1 Idade e Pontuação

Diferente da hipótese inicial, a média de pontos do jogador aumentou ao longo dos anos, atingindo seu pico com 26. Foram observadas duas quedas significativas: aos 23 anos, devido a uma lesão, e aos 31 anos, durante a temporada impactada pela pandemia de COVID-19. Esses eventos destacam o impacto de fatores externos no desempenho, enquanto o crescimento geral sugere que a experiência e a adaptação compensaram possíveis limitações físicas relacionadas à idade.

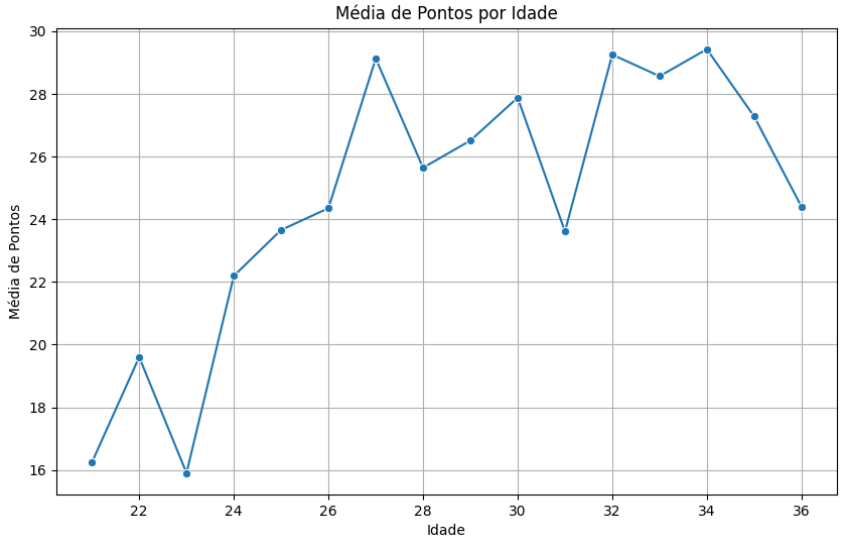

Figura 4 – Média de pontos por idade

7.2 Desempenho em Casa vs Fora de Casa

A hipótese inicial considerava que o jogador teria um rendimento superior em jogos realizados em casa, possivelmente devido ao apoio da torcida e à familiaridade com o ambiente. No entanto, os dados mostram que a pontuação total é semelhante em jogos dentro e fora de casa, sem uma diferença significativa entre os dois cenários. Essa consistência reflete a habilidade do jogador em manter seu desempenho independentemente do local da partida, evidenciando sua adaptabilidade e regularidade ao longo das temporadas.

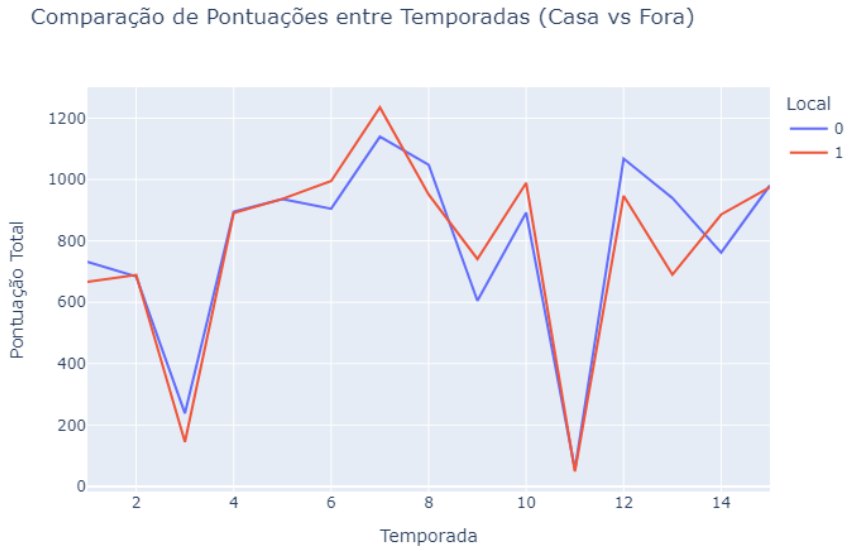

Figura 5 – Soma dos pontos dos Jogos em Casa e Fora de Casa

7.3 Intervalo entre Jogos e Desempenho

A hipótese de que o jogador se poupa devido ao curto intervalo entre jogos foi confirmada pelos gráficos de pontuação por rodada, que apresentaram um padrão claramente oscilatório em todas as temporadas. Essa variação indica que o atleta ajusta seu desempenho conforme a intensidade e as exigências de cada partida, possivelmente para preservar energia ao longo da temporada.

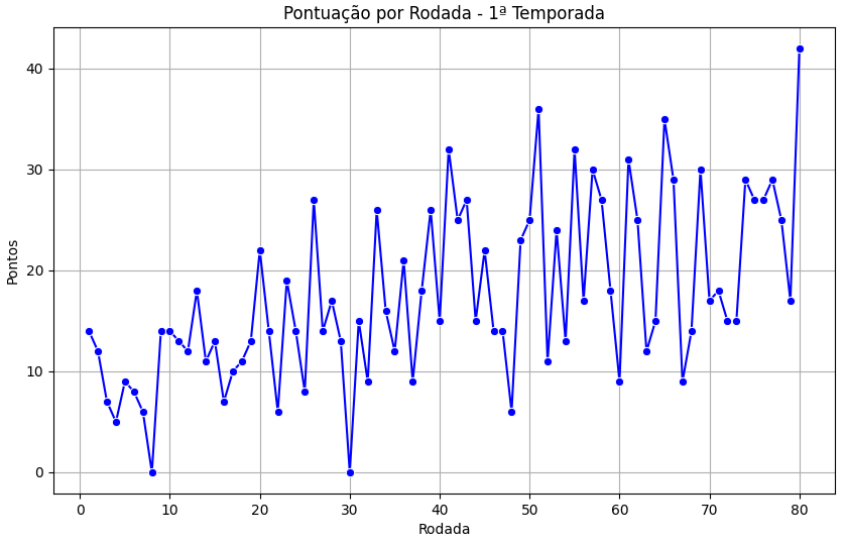

Figura 6 – Pontos por rodada da 1ª temporada

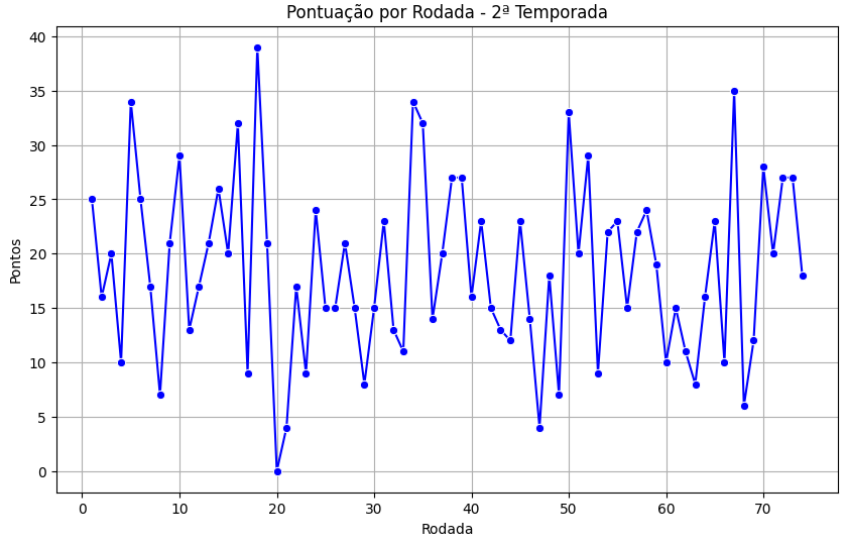

Figura 7 – Pontos por rodada da 2ª temporada

7.4 Diferenças entre Partes da Temporada

A hipótese levantada sugere que, após a virada do ano, o jogador precisa melhorar seu desempenho para garantir a classificação para os playoffs. A análise dos dados con-

firma essa suposição, evidenciando uma diferença significativa na pontuação entre as duas partes da temporada. Na maioria das temporadas, a segunda parte apresenta um desempenho muito superior à primeira, indicando que o jogador intensifica sua performance nos momentos decisivos da competição.

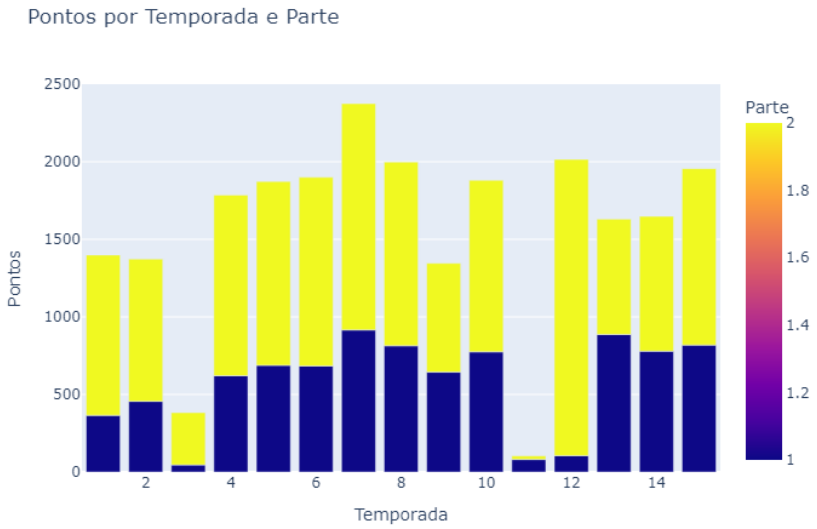

Figura 8 – Soma de pontos por parte

7.5 Média de Pontuação por Oponente

O gráfico confirma a hipótese de que o desempenho do jogador varia conforme o adversário. Observa-se que a média de pontos contra diferentes oponentes apresenta uma ampla variação, com alguns adversários permitindo médias significativamente mais altas, enquanto outros resultam em desempenhos mais baixos. Isso reflete a influência do estilo de jogo, estratégias defensivas e outros fatores específicos de cada equipe no rendimento do jogador.

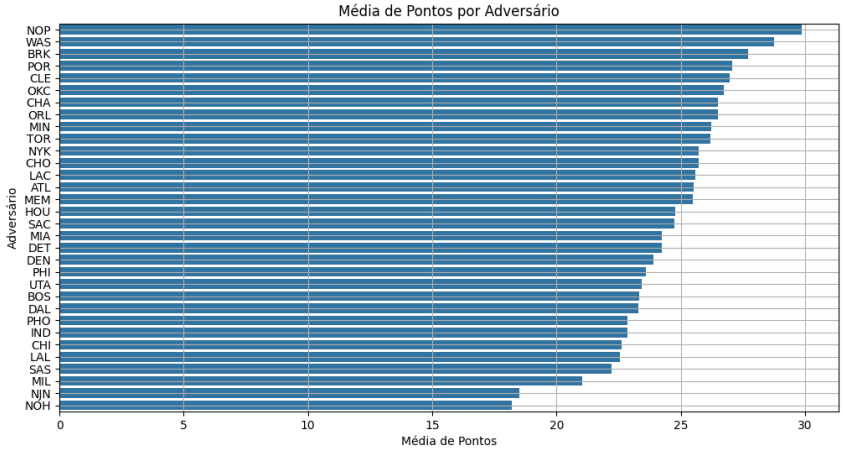

Figura 9 – média de pontos por oponente

7.6 Minutos Jogados e Pontuação

A hipótese de que o tempo em quadra está diretamente relacionado à pontuação do jogador foi confirmada pela análise do gráfico. Observa-se que a medida que os minutos jogados aumentam, a pontuação também cresce de forma consistente. Esse comportamento reflete que, quanto mais tempo o jogador permanece em quadra, maior é sua contribuição ofensiva para a equipe.

Esses resultados reforçam a importância do gerenciamento de tempo em quadra para maximizar o impacto do atleta.

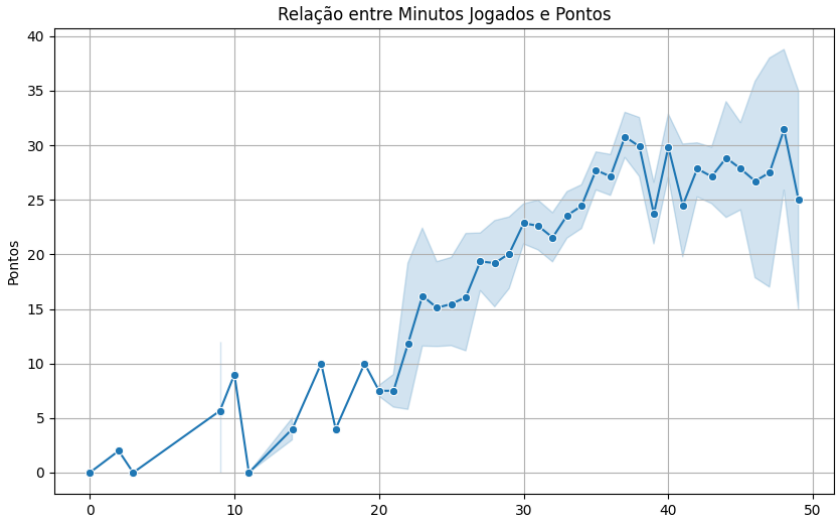

Figura 10 – Relação entre MP e PTS

7.7 Análises complementares

A distribuição das pontuações, próxima a uma curva normal, mostra que a maioria das partidas se concentra entre 20 e 30 pontos, representando o desempenho típico do jogador. O boxplot complementa essa análise, destacando que o intervalo de, 20 a 30 pontos, cobre a maior parte das pontuações. A média, 24.76 pontos, reforça essa consistência.

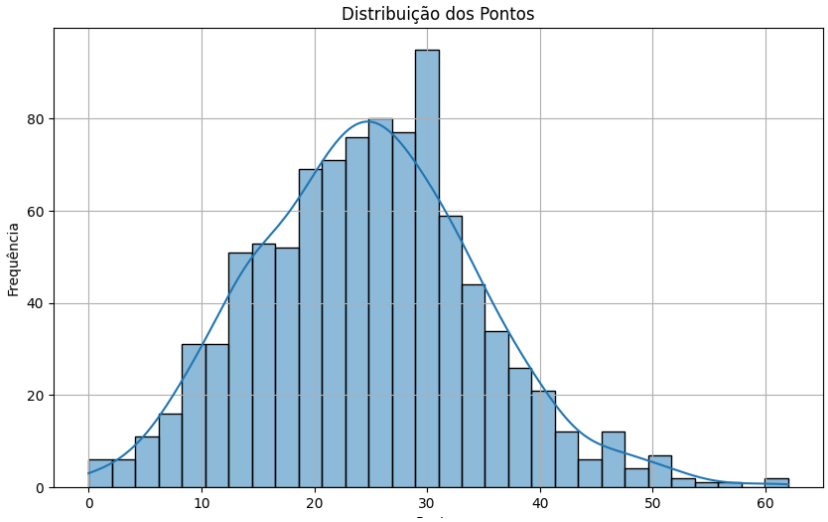

Figura 11 – Distribuição da frequência da pontos

Enquanto pontuações muito baixas ou muito altas são menos frequentes, o boxplot evidencia jogos excepcionais com pontuações acima de 50 pontos como outliers. Esses eventos raros elevam ligeiramente a média, mas o desempenho regular permanece concentrado no intervalo mais frequente, refletindo uma performance sólida e consistente do

jogador, com momentos ocasionais de destaque.

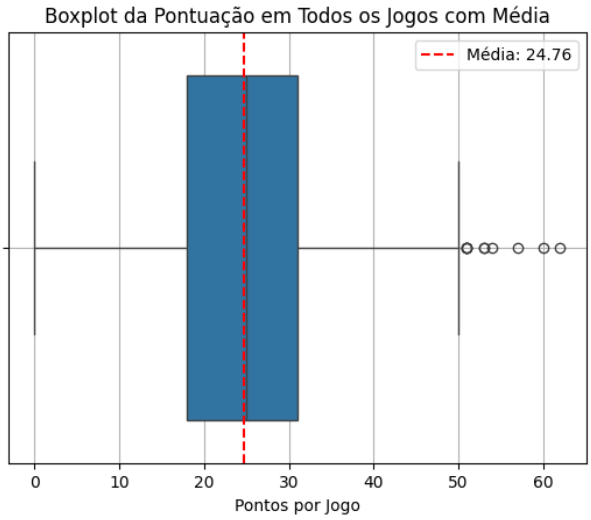

Figura 12 – Analise dos PTS a partir da média

O gráfico abaixo apresenta a soma dos pontos e minutos jogados por temporada, destacando de forma clara as conclusões já mencionadas. As métricas demonstram a evolução do jogador ao longo do tempo e a importância do tempo em quadra para seu rendimento.

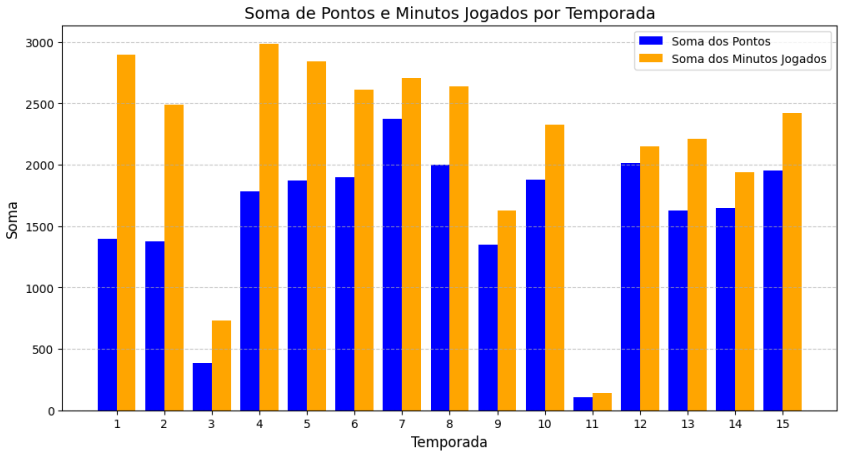

Figura 13 – Analise dos PTS e MP por Temporada

8 RELAÇÕES ENTRE VARIÁVEIS

A relação entre as variáveis e a target (PTS) representa como cada feature contribui para o número de pontos marcados. O mapa de calor e a ANOVA são ferramentas que ajudam a interpretar essa relação, oferecendo insights sobre correlação e significância estatística.

No mapa de calor, observa-se que a variável MP (minutos jogados) apresenta a maior correlação com a target (0.41), indicando que o tempo em quadra tem um impacto positivo moderado nos pontos marcados. As variáveis Age (0.33) e Temporada (0.33) também apresentam correlações positivas, embora mais fracas, sugerindo uma leve relação com o desempenho do jogador. Outras variáveis, como Month (0.01) e Day_of_Week (0.05), têm correlação quase nula, indicando pouca ou nenhuma conexão linear com os pontos.

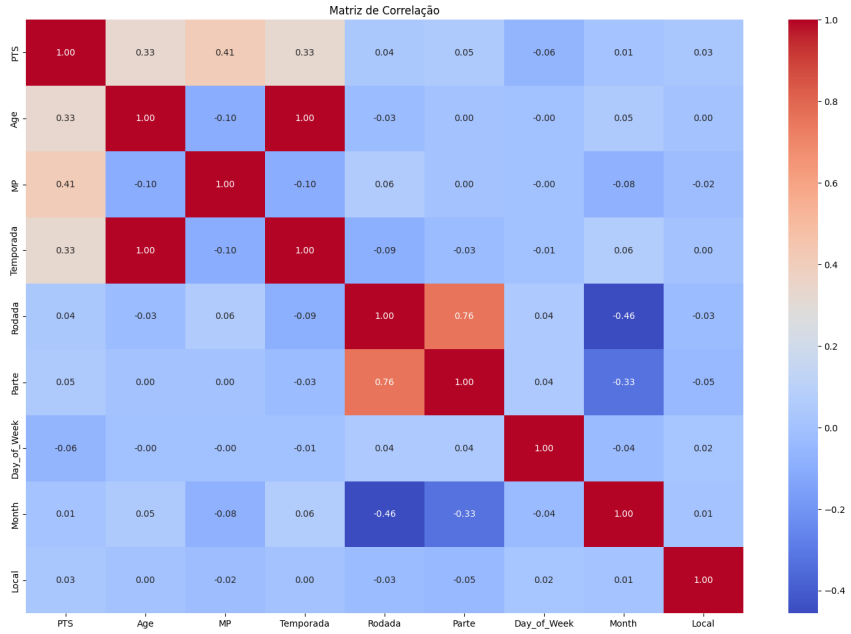

Figura 14 – Mapa de calor

A ANOVA complementa essa análise ao identificar a significância estatística das variáveis em relação à target. Temporada (p = 0.0000) e Month (p = 0.0000) apresentam impacto significativo, mesmo com correlações moderadas ou fracas no mapa de calor. Por outro lado, variáveis como Local (p = 0.4119) e Rodada (p = 0.7424) não mostram relevância estatística, corroborando suas baixas correlações.

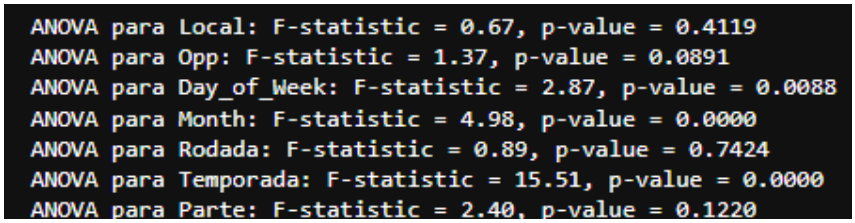

Figura 15 – ANOVA

Em resumo, variáveis como MP, Temporada e Month são as mais relevantes para prever a target, com base em sua correlação ou significância estatística. Outras, como Local e Rodada, têm pouco ou nenhum impacto no desempenho de pontos, sendo menos úteis para modelagem ou análise mais aprofundada.

9 MODELAGEM

Após o tratamento dos dados, aplicação de feature engineering e realização da Análise Exploratória de Dados (EDA), os dados foram organizados em ordem cronológica e divididos em dois conjuntos: 80% para treino e 20% para teste. Essa estrutura respeitou a sequência temporal dos jogos, garantindo uma avaliação consistente do modelo.

Com o objetivo de prever a pontuação dos jogadores, foram testados diferentes modelos de regressão, incluindo regressão linear, regressão logística, random forest e gradient boosting regressor. A avaliação de desempenho utilizou métricas de erro aplicadas aos conjuntos de treino e teste.

As variáveis foram selecionadas com base em análises exploratórias realizadas previamente, incluindo mapas de calor e testes de ANOVA. Foram testadas as variáveis mais relevantes identificadas por cada método separadamente e em conjunto. Os resultados indicaram que as variáveis destacadas pela ANOVA, como dia da semana, mês e temporada, apresentaram maior relevância estatística para o desempenho do jogador, direcionando sua inclusão no modelo.

Entre os modelos, o gradient boosting regressor apresentou os melhores resultados, destacando-se por sua capacidade de capturar padrões complexos e realizar correções iterativas de erro. A utilização das variáveis selecionadas com base na ANOVA contribuiu para a precisão do modelo, permitindo prever variações de desempenho em diferentes contextos de jogo e consolidando-o como a escolha mais adequada para o problema proposto.

9.1 Aprimoramento

No processo de aprimoramento do modelo, foi identificada a necessidade de reduzir o erro absoluto médio inicial, que era de 7,83, para melhorar a precisão das predições. Foram testados diferentes valores para os hiperparâmetros, utilizando validação cruzada (cross-validation) para identificar a combinação mais otimizada. Após essas iterações, o erro absoluto médio foi reduzido para 6,9928, com os seguintes parâmetros: n_estimators definido como 200, max_depth como 2, learning_rate como 0,01 e subsample como 0,6, resultando em um modelo mais preciso e ajustado.

Além desses ajustes, a proporção entre os conjuntos de treino e teste foi modificada, com o test_size reduzido de 20% para 15%, o que aumentou o volume de dados utilizados no treinamento e melhorou o desempenho preditivo. Outro refinamento incluiu a transformação da variável "idade"em uma variável numérica tipo float, substituindo sua representação categórica para agregar maior precisão ao modelo.

9.2 Resultados

Com essas otimizações, o erro absoluto médio caiu ainda mais, atingindo 6,6398 na predição de pontos. Diante desses resultados, nós decidimos por expandir o escopo do modelo, estendendo as predições para outras métricas de desempenho, como assistências, rebotes, roubos de bola, turnovers e bloqueios. Essas predições adicionais aumentam a aplicabilidade do modelo, permitindo uma análise mais ampla e completa do desempenho dos jogadores em múltiplos aspectos do jogo.

```
Random Forest - MAE: 7.29, RMSE: 2.70
Linear Regression - MAE: 7.83, RMSE: 2.80
```

Figura 16 – Resultado Random Forest e Regressão Linear para Pontuação

```
Métricas para o conjunto de treino - Pontos:
MAE (Treino): 4.7399
RMSE (Treino): 2.1771

Métricas para o conjunto de teste - Pontos:
MAE (Teste): 6.9982
RMSE (Teste): 2.6454
```

Figura 17 – Resultado Gradient Boost para Pontuação

```
key_value_updates = {
    "n_estimators": 200,
    "max_depth": 2,
    "subsample": 0.6,
    "learning_rate": 0.01
    }

gradient_params = redefine_target(gradient_params, df, features, 'PTS', key_value_updates)

print("Target - Pontos:")
y_pred_test_pts = gradient_boosting_regressor_custom(gradient_params)
```

```
Target - Pontos:
Métricas para o conjunto de treino:
MAE (Treino): 6.2395
RMSE (Treino): 2.4979

Métricas para o conjunto de teste:
MAE (Teste): 6.6398
RMSE (Teste): 2.5768
```

Figura 18 – Resultado Gradient Boost Pontuação otimizado

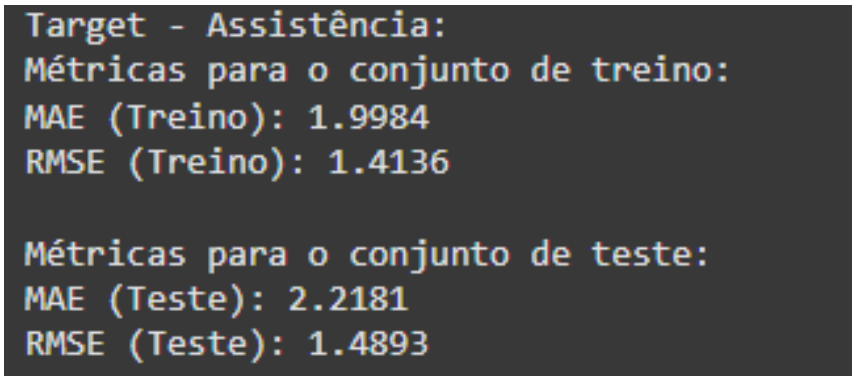

Figura 19 – Resultado Gadrient Boost Assistências

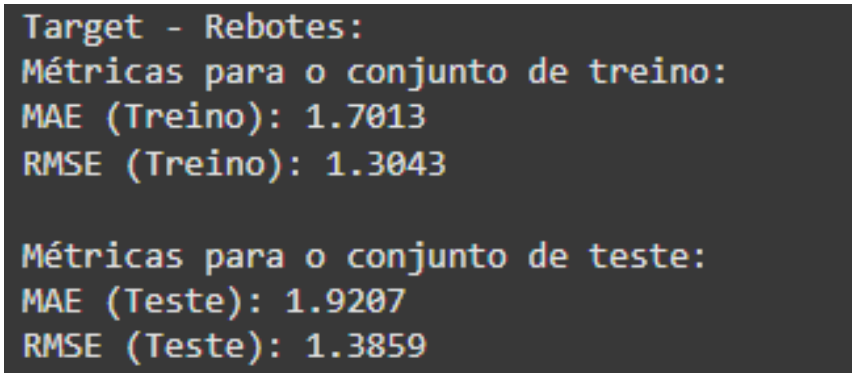

Figura 20 – Resultado Gadrient Boost Rebote

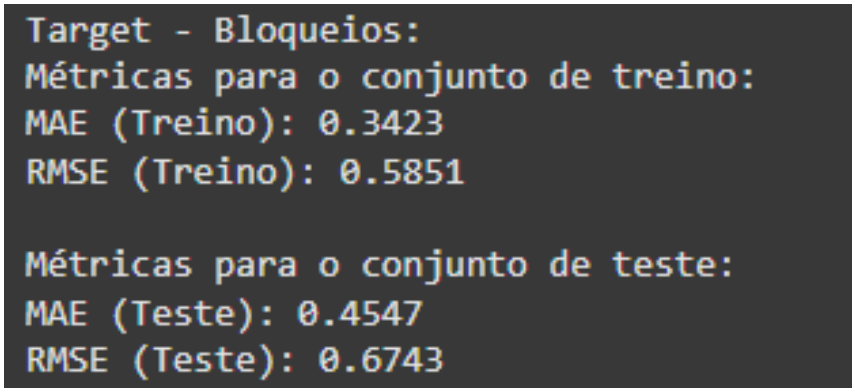

Figura 21 – Resultado Gadrient Boost Bloqueio

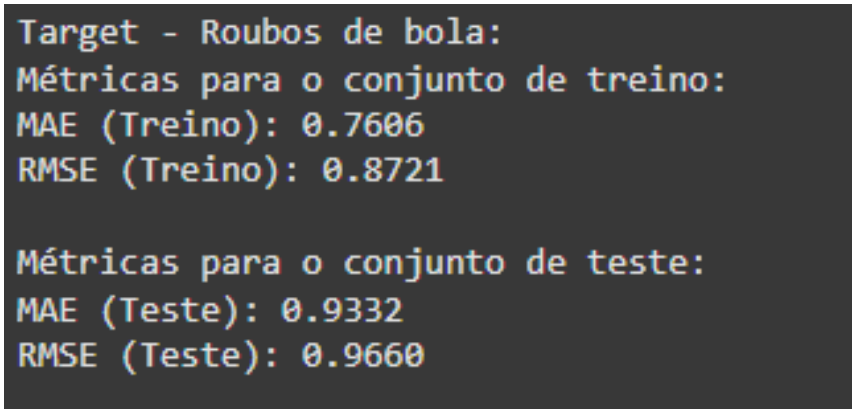

Figura 22 – Resultado Gadrient Boost Roubo de bola

10 CONCLUSÃO

Neste projeto, foram aplicadas técnicas de Machine Learning para o desenvolvimento de um modelo preditivo capaz de estimar a pontuação de jogadores da NBA, utilizando Stephen Curry como caso inicial de estudo. O pipeline abrangeu desde o tratamento inicial dos dados até a otimização de modelos, com destaque para o Gradient Boosting Regressor, que apresentou alta eficácia ao capturar relações complexas entre variáveis e reduzir significativamente o erro absoluto médio. A inclusão de variáveis contextuais, como idade, local da partida e sazonalidade, ampliou a precisão das predições e enriqueceu as análises realizadas.

A análise gráfica revelou oscilações consistentes no desempenho do jogador em diversas métricas, como assistências, roubos de bola, turnovers e bloqueios, permitindo a expansão do modelo para prever outros indicadores de performance. Os resultados destacam a ciência de dados como uma ferramenta estratégica no esporte, oferecendo insights valiosos para a montagem de equipes otimizadas e tomadas de decisão. Recomenda-se, como próxima etapa, a aplicação do modelo a todos os jogadores e a realização de análises a nível de equipe, visando prever resultados completos de partidas.

Por fim, a exploração de abordagens baseadas em séries temporais é sugerida como um possível aprimoramento, permitindo a captura de padrões sequenciais e maior precisão nas predições. Este trabalho evidencia o potencial da ciência de dados como recurso indispensável para a análise de desempenho esportivo e para a evolução de práticas estratégicas no esporte de alto rendimento.