Rodrigo Trindade Pedrosa

Modelos Preditivos Esportivos Aplicados a Dados da NBA

Niterói - RJ, Brasil

10 de maio de 2021

Universidade Federal Fluminense

Rodrigo Trindade Pedrosa

Modelos Preditivos Esportivos Aplicados a Dados da NBA

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Prof. Dr. Hugo Henrique Kegler dos Santos

Niterói - RJ, Brasil

10 de maio de 2021

Universidade Federal Fluminense

Rodrigo Trindade Pedrosa

Modelos Preditivos Esportivos Aplicados a Dados da NBA

Monografia de Projeto Final de Graduação sob o título *"Modelos Preditivos Esportivos Aplicados a Dados da NBA"*, defendida por Rodrigo Trindade Pedrosae aprovada em 10 de maio de 2021, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Folha de assinaturas

Prof. Dr. Hugo Henrique Kegler dos Santos
Departamento de Estatística – UFF

Prof. Dr. Luis Guillermo Coca Velarde
Departamento de Estatística – UFF

Prof. Dr. Rafael Santos Erbisti
Departamento de Estatística – UFF

Niterói, 10 de maio de 2021

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

P372m Pedrosa, Rodrigo Trindade
Modelos preditivos esportivos aplicados a dados da NBA / Rodrigo Trindade Pedrosa ; Hugo Henrique Kegler Dos Santos, orientador. Niterói, 2021.
47 f. : il.

Trabalho de Conclusão de Curso (Graduação em Estatística)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2021.

1. NBA. 2. Modelos Lineares Generalizados. 3. Poisson. 4. Previsão. 5. Produção intelectual. I. Dos Santos, Hugo Henrique Kegler, orientador. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

A NBA sempre foi e continuará sendo uma das maiores ligas esportivas do mundo, atraindo milhares de fãs ao redor do planeta e movimentando bilhões de dólares todos os anos. Com a o objetivo de ajudar as pessoas a entenderem melhor os jogos desse campeonato, a estatística se faz presente, sendo um dos seus métodos, o ajuste de modelos para prever resultados. Neste trabalho foram ajustados 7 modelos lineares generalizados utilizando a distribuição de Poisson, tendo a quantidade de pontos marcados por cada time como a variável resposta e as pontuações de fundamentos básicos do basquete como variáveis explicativas, além de duas variáveis indicando o local e em qual temporada a partida ocorreu. Com a previsão dos pontos para cada time, foram simuladas 1000000 de partidas para todos os últimos confrontos que ocorreram em quatro temporadas da NBA. Os modelos foram comparados por algumas medidas e tiveram certas dificuldades para predizer o vencedor de uma partida, com o melhor dos modelos prevendo corretamente, em média, 52,43% das partidas das simulações.

Palavras-chave: NBA, Modelos Lineares Generalizados, Poisson, LASSO, AIC, Previsão

Agradecimentos

O caminho até a conclusão deste trabalho foi árduo e cansativo, porém seria pior se não fosse pela ajuda e companherismo de algumas pessoas. Por isso, agradeço aos meus pais e minha irmã por todo amor que me deram não só no meu período acadêmico mas em toda a minha vida.

Agradeço a minha namorada pelo amor e paciência que me deu desde o início deste trabalho, sempre me encorajando a fazer o meu melhor.

Agradeço aos meus amigos pelo apoio e compreensão por ausências em determinados eventos.

Agradeço ao professor Hugo não só por toda ajuda e orientação neste trabalho e em toda minha passagem pela Universidade Federal Fluminense, mas pelas boas conversas de diversos assuntos.

Agradeço ao professor Guillermo por ter aceitado o convite para participar desta banca e pelas dicas na primeira apresentação deste trabalho, fazendo com que eu o melhorasse.

Agradeço gradeço ao professor Rafael por ter aceitado o convite para participar desta banca e por ter respondido diversas dúvidas durante a escrita deste trabalho.

Por fim, agradeço aos demais professores da Universidade Federal Fluminense por todo conhecimento que me passaram, ajudando na minha formação acadêmica.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Motivação

1.1.1 *National Basketball Assosciation*, a NBA

O esporte que será utilizado neste trabalho será o basquetebol. Mais especificamente, serão utilizados dados da *National Basketball Association (NBA)*, a liga de basquete profissional americana. A Associação Nacional de Basquetebol, em sua tradução livre para português, foi fundada em 6 de junho de 1946 com o nome de *Basketball Association of America* (BAA). Entretanto, ela não foi a primeira liga profissional de basquete americana, naquela época já existiam a ABL (*American Basketball League*), primeira liga profissional do país e a NBL (*National Basketball League*). A BAA apenas foi fundada porque um grupo de empresários do hóquei no gelo percebeu que existiam espaços no calendário da liga deste esporte, o que fazia com que suas majestosas arenas, na época, ficassem vazias. A BAA, então, entraria com um segundo esporte nessas arenas, o que preencheria os espaços e renderia mais dinheiro aos fundadores. Apesar de no início o plano não ter sido muito bem atingido (as arenas ficavam vazias e havia precariedade na mudança das quadras para os dois esportes), a BAA foi chamando a atenção por sediar jogos em acomadações bem maiores, em comparação às outras duas ligas, que usavam até mesmo ginásios de escolas do Ensino Médio. Com essa popoularidade crescendo, em 1947, o campeão da ABL se transferiu para BAA e, no ano seguinte, 4 times da ABL seguiram os mesmos passos, incluindo o último vencedor. E em 1949, a NBA, como é conhecida hoje em dia, finalmente surgiu após uma "fusão" da BAA com a NBL.[1]

Desde então diversos acontecimentos, jogadores e times entraram para a história da liga. Começando na década de 50 com o poderoso *Minneapolis Lakers* liderado por George Mikan, conquistando três vezes seguidas o título da liga, de 1952 a 1954. Nos anos 60 surgiam dois dos maiores jogadores de basquete que já existiram: Bill Russel, campeão onze vezes da liga com o *Boston Celtics* (nove vezes como jogador, sendo oito de forma

seguida, e duas vezes como treinador), além de ter sido cinco vezes o MVP (*Most Valuable Player*, prêmio concedido ao melhor jogador da temporada), e Wilt Chamberlain, eternizado por, em apenas uma partida, marcar 100 pontos e ter uma média de, aproximadamente, 50 pontos por partida em uma temporada (recordes não quebrados até hoje); além disso, também deve ser lembrado Jerry West, ídolo dos *Lakers*, que apesar de não ser confirmado, muitos dizem ser o jogador no logo da liga. A entrada de Bill Russel na liga foi importantíssima, pois foi o principal motivo que fez com que negros pudessem atuar de vez na liga. Na década de 70, pode ser destacada a entrada de Kareem Abdul-Jabbar, o maior "cestinha" da liga de todos os tempos, com 38.387 pontos, além disso foi campeão e MVP seis vezes com *Milwaukee Bucks* e *Los Angeles Lakers*. Nos anos 80, pode-se destacar a grande rivalidade entre Magic Johnson (*Los Angeles Lakers*) e Larry Bird (*Boston Celtics*) que foram campeões cinco e três vezes, respectivamente, e conquistaram mais de uma vez o prêmio de MVP; além dos *"Bad Boys"* do *Detroit Pistons* que ficaram lembrados por um estilo de jogo mais agressivo fisicamente que os fez ganhar a liga duas vezes. Em 1984, a liga recebe o que pra muitos é o melhor jogador de todos os tempos: Michael Jordan. [2]

Jordan brilhou nos anos 90 fazendo jogadas extraordinárias pelo *Chicago Bulls*, ao lado de Scottie Pippen e seu treinador, Phil Jackson. Eles foram campeões seis vezes ao todo de 1991 a 1993 e 1996 a 1998. Entre esses períodos o jogador ainda saiu do basquete para jogar beisebol e, mesmo assim, voltou conquistando o título três vezes seguidas com a adição de Dennis Rodman. Mas os anos 90 não foram só Michael Jordan e os *Chicago Bulls*, muitos jogadores foram dessa geração e marcaram época, como: Hakeem Olajuwon, que com *Houston Rockets* conquistou os títulos de 1993 e 1994, Charles Barkley, Patrick Ewing, a dupla do *San Antonio Spurs* formada por David Robinson e Tim Duncan, e treinada por Gregg Popovich, conhecida como "Torres Gêmeas", John Stockton, que foi o maior assistenciador da liga, Reggie Miller, Shaquille O'Neal, entre tantos outros. Esta década também presenciou uma das melhores classes dos *drafts* (processo de seleção de novos jogadores vindos de universidades ou outros países), a classe de 1996. Foram selecionados jogadores como Allen Iverson, Shareef Abdur-Rahim, Stephon Marbury, Ray Allen, Antoine Walker, Kobe Bryant, Peja Stojakovic, Steve Nash, Jermaine O'Neal e Zydrunas Ilgauskas, todos jogadores *all-star* (melhores jogadores da liga na temporada) posteriormente. E estes jogadores abriram o caminho para os anos 2000. [2]

Na década seguinte, Kobe Bryant foi o jogador que mais se destacou, anotando, inclusive, 81 pontos em uma partida em 2006, se consagrando como o jogador com a segunda maior pontuação em uma única partida. Ao lado de Shaquille O'Neal e treinado por Phil

Jackson (o mesmo treinador do *Chicago Bulls* nos anos 90), Kobe formou uma das maiores duplas da história do basquetebol, conquistando a NBA por 3 vezes seguidas entre 2000 e 2002 para o *Los Angeles Lakers*. Após o rompimento da dupla por desentendimentos pessoais, Kobe Bryant ainda foi campeão mais duas vezes com Phil Jackson, em 2009 e 2010. Além disso, outros fatos marcaram a década, como a entrada de LeBron James na liga, que também é considerado, por muitos, o maior jogador de todos os tempos, a conquista do primeiro título de Dwayne Wade com o *Miami Heat* em 2006, ao lado de Shaquille O'Neal, a volta de uma conquista da liga pelo *Boston Celtics* depois de 22 anos, entre tantos outros momentos. Na segunda década do século, pode-se dizer que LeBron desencantou e se consagrou como um dos melhores de todos os tempos. O jogador selecionado pelo *Cleveland Cavaliers* em 2003, se mudou para o *Miami Heat* em 2010, após apenas conquistar um vice-lugar com o time de Ohio, em 2007. Em Miami, LeBron foi treinado por Erik Spoelstra, formando com Dwayne Wade e Chris Bosh o *"The Big Three"*, um dos melhores trios da história da liga, para conquistar o título duas vezes: em 2012 e 2013. Voltou para Cleveland em 2015 e ajudou a conquistar o primeiro título da franquia na história, foi em cima do *Golden State Warriors* em 2016. O time de São Francisco é outro que merece atenção na década de 10, conquistou o campeonato três vezes em cinco aparições com a liderança dos *"Splash Brothers"*, dupla formada por Klay Thompson e Stephen Curry. Esse último é considerado um revolucionário do esporte fazendo com que os chutes de 3 pontos se tornassem a principal arma na última década. Neste período pode-se destacar também o primeiro e único título de Dirk Nowitzki com *Dallas Mavericks*, em 2011; a primeira vez que um time não estadunidense foi campeão da liga, o *Toronto Raptors* (Canadá) em 2019, a entrada de Giannis Antetokounmpo na liga em 2013, jogador grego que conquistou os dois últimos títulos de MVP e o último título de DPOY (*Defensive Player Of The Year*, o melhor jogador defensivo do ano), entre tantos outros. O início da década de 20 fica marcado por mais um título do *Los Angeles Lakers*, dessa vez liderado por Anthony Davis e LeBron James.

Todos esses dados foram retirados do site *Basketball Reference* que contém tanto informações dos jogadores e times atuais das ligas, quanto históricos de todas as temporadas. [3]

Hoje em dia, 29 times do Estados Unidos e 1 do Canadá disputam a liga, separados em 2 conferências (leste e oeste), os maiores ganhadores são os *Boston Celtics* e os *Los Angeles Lakers* com 17 títulos cada. Falando em em receitas e audiência, o valor médio de cada time é de a 1,9 bilhões de dólares, a receita total da liga chega a 8,76 bilhões de dólares; e na última final, só nos Estados Unidos, 7,5 milhões de pessoas estavam

assistindo a final.[4][5] Além de EUA e Canadá, a NBA já realizou jogos fora em mais de 15 países com o intuito de atrair o público e era, inclusive, a liga de esportes mais popular na China em 2018.[6][7] Com tanta história, ídolos e popularidade a NBA é umas das maiores ligas esportivas no mundo hoje, tendo importância não só nessa área mas como na área social também, e essa valorização ainda pode aumentar e muito, portanto é certo que existam trabalhos de previsão para os resultados das partidas da liga.

1.1.2 Estatística e o esporte

Os esportes sempre foram populares no mundo todo, afinal, é algo que traz alegria e descontração para todos. Com a crescente onda de profissionalização de atletas de diversas categorias no início do século XX, pouco demorou para que a indústria esportiva começasse a movimentar quantias extraordinárias em vários países e cada vez aumentasse mais; em 2019, por exemplo, a quantia estimada de dinheiro envolvida em apostas esportivas foi de 495 bilhões de dólares. Com tanto dinheiro envolvido, é claro que a estatística "entrou em campo" para ajudar a entender como os esportes funcionavam em números. Por exemplo, Charles Reep e Bernard Benjamin, em 1968, publicaram o artigo *"Skill and Chance in Assosciantion Football"*, em que, através de 15 anos de coletas de dados, conseguiram identificar padrões em uma partida de futebol, como um chute em cada oito terminar em gol. Outro trabalho, ainda mais famoso, é o de Billie Beane, ex-jogador e gerente geral de beisebol, que revolucionou todos os esportes, ao montar um time recrutando jogadores mais baratos baseado apenas em suas estatísticas e no *Sabermetrics*. O time foi de azarão a um dos melhores da liga de beisebol americana, chegando a participar da fase mata-mata da competição. O caso foi tão espetacular que virou o livro *Moneyball*, escrito por Michael Lewis, contando sua experiência, tornando-se *best-seller* e parando nas telas de cinemas mundiais.[8]

No basquetebol, que é o esporte deste trabalho, pode ser citado Dean Oliver, ex-jogador do esporte e atual treinador assistente do *Washington Wizards* na NBA, que foi um dos pioneiros da análise estatística na categoria e líder do movimento *APBRmetrics - Association for Professional Basketball Research Metrics* (Associação para Pesquisa de Métricas no Basquete Profissional, em tradução livre), baseado no *Sabermetrics*, e por isso é chamado por muitos de *"The Godfather of Advanced Basketball Analytics"* (O Padrinho da Análise Avançada no Basquetebol, em tradução livre).[9][10] Em 2004, Dean publicou o livro *Basketball On Paper* e nele apresentou pela primeira vez o termo *"Four Facts"*, os quatro fatores que influenciam sucesso, vitória no basquetebol; são eles: arremessos

(40%), perda de bola (25%), rebote (20%) e lances-livres (15%). Ao lado de cada fator está a influência deles em uma partida e podem ser aplicados pra ambos os times.[11]

Na seção seguinte, serão apresentados alguns outros projetos envolvendo, de fato, modelos preditivos que ajudaram na confecção deste trabalho.

1.2 Revisão Bibliográfica

Shanahan (1984), utilizando dados de partidas das temporadas de 1981-82 e 1982-83 dos times de basquetebol feminino e masculino da Universidade de Iowa, tentou demonstrar que a probabilidade de vencer um jogo de basquete poderia ser explicada por uma combinação de estatísticas relacionadas ao esporte, como quantidade de arremessos, quantidade de bloqueios, quantidade de rebotes, porcentagem de acerto em arremessos, etc; e também a significância de cada variável, individualmente, em ganhar ou não um jogo. Para isso, Shanahan utilizou Análise Fatorial e o método de Regressão Logística para as estatísticas dos times de ambos os gêneros. O primeiro foi usado para diminuir a quantidade de variáveis independentes agrupando-as quando conveniente, e com esses resultados foi verificado quais variáveis tinham as maiores comunalidades e correlações para serem eliminadas da análise, porém a autora também utilizou o conhecimento de treinadores do esporte para saber se as variáveis deveriam ser realmente eliminadas. No final, o modelo para a categoria feminina tinha as variáveis porcentagem de arremessoa certos, total de rebotes, total de faltas, porcentagem de rebotes, total de lances livres realizados, total de roubadas de bola, total de bloqueios e total de perdas de bola; destes, o total de rebotes foi o que mais tinha significância com o modelo. Além disso, a porcentagem de explicação do modelo foi de 64%, o que fez com que prevesse corretamente 90% dos jogos, sendo que 80% destes seriam vitórias para o time feminino em determinada temporada. Já para o time masculino, o modelo continha as variáveis porcentagem de arremessos certos, total de rebotes, total de faltas, total de roubadas de bola, total de arremessos bloqueados e total de erros forçados, sendo a terceira a que mais contribuiu para o modelo, que tinha uma porcentagem de explicação de 56% , prevendo 88% dos jogos corretamente, sendo que 93% deles eram vitórias. [12]

Lopez e Matthews (2015) ajustaram dois modelos para a previsão de partidas da Divisão I de basquetebol masculino da NCAA (*National Collegiate Athletic Association*, associação que organiza os esportes universitários nos EUA). O primeiro modelo usava regressão logística e retornava a probabilidade de um time vencer a partida utilizando,

como variável explicativa, o *"Las Vegas Point Spread"* que é uma predição realizada antes do jogo de quanto será a diferença de pontos entre o time mandante e o time visitante; por exemplo, se a *"spread"* for de -5,5, significa que o time mandante ganhará por 5,5 pontos de diferença. No segundo modelo também foi utilizado a regressão logística, na probabilidade de um time vencer uma partida com 5 variáveis explicativas, eficiência ofensiva ajustada (pontos marcados por 100 possessões de bola) para ambos os times, eficiência defensiva ajustada (pontos sofridos por 100 possessões de bola) para ambos os times e se a partida era realizada em um quadra neutra ou não. Eles, então, mediram pesos para minimizar a perda logarítmica, e mostraram que o segundo modelo era um pouco melhor que o primeiro. No final do trabalho os modelos foram unificados com seus respectivos pesos o que resultou em um modelo melhor que outros 400 em termos de perda logarítmica. [13]

Cheng *et al.* (2016) ajustaram um modelo que chamaram de NBAME e era baseado no Princípio da Máxima Entropia. Esse modelo tinha o objetivo de prever o resultado de um jogo da fase mata-mata da NBA, de um time qualquer, utilizando a média dos últimos seis jogos de cada um dos principais fundamentos de basquetebol. O NBAME passou por simulações com dados das temporadas de 2007-08 a 2014-15 e suas AUC's (Árra embaixo da Curva ROC) foram comparadas com as de outros modelos: (*Naive Bayes*, Regressão Logística, *BP Neural Network* e Floresta Aleatória). No final, o modelo só não apresentou os maiores valores nas temporadas 2012-13 e 2013-14. [14]

Além de trbalhos relacionados ao basquete, outros trabalhos no futebol também tiveram infulência na realização deste projeto. O primeiro deles é o artigo de Louzada *et al.* (2015) que propuseram uma simulação, utilizando métodos bayesianos e um modelo de Poisson para a quantidade de gols de dois times oponentes, com o objetivo de prever os resultados das Copas de 2010 e 2014. Para a construção da distribuição a priori dos parâmetros do modelo, o nível técnico das seleções, calculado pelo ranking da FIFA, e a opinião de especialistas foram utilizados. O nível técnico foi utilizado de forma que, por exemplo, os gols marcados por um time A contra um Time B são diretamente proporcionais ao nível técnico do time A e inversamente proporcional ao nível técnico do time B. Para a Copa de 2010 o modelo teve distâncias de DeFinnetti gerais de 0.568 e 0.570 para a fase de grupo e para a fase final, respectivamente. Além disso, o modelo foi escolhido o melhor em uma competição realizada pela Sociedade Brasileira de Pesquisa Operacional para a Copa do Mundo de 20210. [15]

Mahfuz (2009) utilizando dados da Copa do Mundo de 2002, também ajustou modelo

de regressão de Poisson para prever o resultados de partidas da competição, sendo a quantidade de gols marcados pelas seleções a variável resposta. Por sua vez, o ranking da FIFA para o time analisado e seu oponente foram utilizados como variávei explicativas, além da variável que mostra se o campo de jogo é neutro ou do time analisado. O modelo resultavanas médias de gols que cada seleção poderia marcar contra a outra, então esse resultado era utilizao como a média em ditribuição de Poisson para determinar as probabilidades de cada resultado. [16]

1.3 Objetivos

O principal objetivo deste trabalho é ajustar e comparar modelos que consigam prever quantos pontos os times marcarão em uma partida e com isso informar o ganhador da mesma.

1.4 Organização

Este trabalho está dividido em quatro capítulos, que são subdivididos em seções e subseções quando necessário. No capítulo 2, encontra-se a descrição da base de dados utilizada neste trabalho, bem como a metodologia utilizada. No capítulo 3, encontram-se resultados obtidos. Por fim, no capítulo 4 se encontra a conclusão do trabalho.

2 Materiais e Métodos

2.1 Base de Dados

A base de dados utilizada neste trabalho se refere às temporadas regulares (sem os jogos de "playoffs") de 2014-15 a 2017-18 da NBA. Ao todo são 9.840 observações que representam cada partida dos 30 times em todas as temporadas mencionadas acima e foi coletada em uma das páginas do site *Kaggle*.[17] Ela contém os valores de 40 variáveis que representam dados da partida e dos principais fundamentos de basquete do time principal da observação e seu oponente. O nome de cada time e sua sigla estão no quadro 16 do apêndice A. A seguir, são apresentados dois quadros com a definição de cada variável.

Quadro 1: Variáveis relacionadas ao time principal da observação

Variável	Definição
Team	O time principal da observação
Game	A ordem da partida analisada. Varia de 1 a 82
Date	A data da partida
Home	Se o time principal jogou em casa ou fora
Opponent	O oponente do time principal da observação
WINorLOSS	Se o time principal ganhou ou perdeu
TeamPoints	A quantidade de pontos do time principal na partida
OpponentPoints	A quantidade de pontos do time oponente na partida

Quadro 2: Variáveis relacionas ao time principal e seu oponente na observação

Variável	Definição
FieldGoals	A quantidade de arremessos certos
FieldGoalsAttempted	A quantidade de arremessos feitos
FieldGoalsPerc	A porcentagem de arremessos certos
X3PointShots	A quantidade de arremessos de 3 pontos certos
X3PointShotsAttempted	A quantidade de arremessos de 3 pontos feitos
X3PointShotsPerc	A porcentagem de arremessos de 3 pontos certos
FreeThrows	A quantidade de lances livre certos
FreeThrowsAttempted	A quantidade de lances livres feitos
FreeThrowsPerc	A porcentagem de lances livre certos
OffRebounds	A quantidade de rebotes ofensivos feitos
TotalRebounds	A quantidade total de rebotes feitos
Assists	A quantidade de assistências feitas
Steals	A quantidade de roubadas de bolas feitas
Blocks	A quantidade de tocos feitos
Turnovers	A quantidade total de vezes em que o time perdeu a bola para o adversário
TotalFouls	A quantidade total de faltas feitas

2.2 Modelos Lineares Generalizados (MLG)

Os modelos de regressão linear tradicionais são muito utlizados em trabalhos de diversas áreas com o intuito de explicar a relação entre uma variável resposta (dependente)

e as variáveis explicativas (independentes). O grande obstáculo é esse tipo de modelo pressupor que a variável resposta segue uma distribuição normal, o que limita muito o estudo.

Em 1972, Nelder e Wedderburn apresentaram os Modelos Lineares Generalizados (MLG) como sendo uma forma de escapar da limitação dos modelos de regressão linear simples. Alguns desses modelos já eram estudados separadamente, como, por exemplo, os modelos log-lineares para dados de contagens (Birch,1963) e os modelos de regressão para análise de sobrevivência (Feigl e Zelen, 1965; Zippin e Armitage, 1966; Glasser, 1967). Os MLG's possuem três componentes:

(i) Componente aleatório do modelo (variável resposta), que segue uma distribuição pertencente à família exponencial;

(ii) Componente sistemático do modelo, que são as variáveis explicativas em uma estrutura de modelo linear;

(iii) Função de ligação, que é a ligação entre as componentes do modelo através de uma função.

Nas subseções seguintes serão apresentados temas que são necessarios para o compreendimento na utilização dos MLG's. Mais informações podem ser encontrados no livro Modelos Lineares Generalizados e Extensões (2013), de Gauss Moutinho Cordeiro e Clarice G.B. Demétrio.[18]

2.2.1 Família Exponencial

A Família Exponencial de distribuições é a classe na qual os modelos lineares generalizados são aplicados.

Uma distribuição pertence a Família Exponencial se sua função de probabilidade pode ser escrita da seguinte maneira: considerando o vetor $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, onde $k \in \mathbb{N}$ como os parâmetros desconhecidos:

$$f(y|\boldsymbol{\theta}) = a(\theta_1, \ldots, \theta_k)b(y) \exp\left\{\sum_{i=1}^{k} c_i(\theta_1, \ldots, \theta_k)d_i(y)\right\}, \tag{2.1}$$

sendo $a(\boldsymbol{\theta})$, $b(y)$, $c_i(\boldsymbol{\theta})$ e $d_i(y)$, com i = 1,..., k, são funções que assumem valores reais quaisquer. A seguir, como exemplo, serão apresentados 3 distribuições que pertencem à família exponencial: Normal, Binomial e Gama.[18]

Exemplo 2.2.1.1. Seja uma distribuição Normal com parâmetros μ (conhecido) e σ^2 (desconhecido) > 0 e função de densidade:

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \times \boldsymbol{I}_{(-\infty,\infty)}(y).$$

A equação acima pode ser reescrita da seguinte maneira:

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \times \boldsymbol{I}_{(-\infty,\infty)}(y) \times \exp\left\{\frac{-1}{2\sigma^2}(y-\mu)^2\right\}.$$

Observa-se que $\mathrm{a}(\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}$, $b(y) = \boldsymbol{I}_{(-\infty,\infty)}(y)$, $c(\sigma^2) = \frac{-1}{2\sigma^2}$ e $d(y) = (y-\mu)^2$. Logo, a distribuição Normal pertence à Família Exponencial.

Exemplo 2.2.1.2. Seja uma distribuição binomial, com parâmetros n e p, onde n é conhecido e $n \in \mathbb{N}$, e $0 < p < 1$ (parâmetro procurado), e sua função de probabilidade:

$$\begin{aligned}
f(y|n,p) &= \binom{n}{y} p^y (1-p)^{n-y} \times \boldsymbol{I}_{(0,\ldots,n)}(y) \\
&= \binom{n}{y} \times \boldsymbol{I}_{(0,\ldots,n)}(y) \times p^y \frac{(1-p)^n}{(1-p)^y} \\
&= \binom{n}{y} \times \boldsymbol{I}_{(0,\ldots,n)}(y) \times (1-p)^n \exp\left\{\log\left(\left(\frac{p}{(1-p)}\right)^y\right)\right\} \\
&= (1-p)^n \binom{n}{y} \times \boldsymbol{I}_{(0,\ldots,n)}(y) \exp\left\{\log\left(\frac{p}{(1-p)}\right) y\right\}.
\end{aligned}$$

Nota-se que $a(p) = (1-p)^n$, $b(y) = \binom{n}{y} \times \boldsymbol{I}_{(0,\ldots,n)}(y)$, $c(p) = log(\left(\frac{p}{(1-p)}\right))$ e $d(y) = y$. Portanto, a distribuição Binomial pertence à Família Exponencial.

Exemplo 2.2.1.3. Seja uma Distribuição Gama(α,β) com α e $\beta > 0$ (desconhecidos) e a seguinte distribuição de densidade:

$$\begin{aligned}
f(y|\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\{-\beta y\} \times \boldsymbol{I}_{(0,\infty)}(y) \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \times \boldsymbol{I}_{(0,\infty)}(y) \times \exp\{(\alpha-1)\log(y)\} \exp\{-\beta y\} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \times \boldsymbol{I}_{(0,\infty)}(y) \times \exp\{(\alpha-1)\log(y) - \beta y\}
\end{aligned}$$

Sendo $a(\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}$, $b(y) = \boldsymbol{I}_{(0,\infty)}(y)$, $c_1(\alpha,\beta) = \alpha - 1$, $c_2(\alpha,\beta) = -\beta$, $d_1(y) = log(y)$ e $d_2(y) = y$, conclui-se que a distribuição Gama pertence à Família Exponencial.

2.2.1.1 Propriedades da Família Exponencial

Os valores de esperança e variância de $d(y)$ podem ser obtidos utilizando os resultados abaixo, que podem ser aplicados para qualquer função de densidade de probabilidade em que a ordem de derivação e integração podem ser mudadas.[19]

Sabe-se que para a uma função de densidade de probabilidade qualquer:

$$\begin{aligned} \int_{\mathbb{R}} f(y|\theta)dy &= 1 \implies \\ \implies \frac{d}{d\theta}\int_{\mathbb{R}} f(y|\theta)dy &= \frac{d1}{d\theta} \implies \\ \implies \int_{\mathbb{R}} \frac{df(y|\theta)}{d\theta}dy &= 0 \end{aligned} \tag{2.2}$$

Sabe-se que a equação 2.1, considerando k = 1, pode ser reescrita na forma:

$$f(y|\theta) = \exp\left\{c(\theta)d(y) + s(\theta) + t(y)\right\}, \tag{2.3}$$

sendo $s(\theta) = \exp\{a(\theta)\}$ e $t(x) = \exp\{b(x)\}$.

Isso implica em

$$\frac{df(y|\theta)}{d\theta} = [c'(\theta)d(y) + s'(\theta)]f(y|\theta). \tag{2.4}$$

Aplicando 2.2 em 2.4:

$$\begin{aligned} \int_{\mathbb{R}} \frac{df(y|\theta)}{d\theta}dy &= \int_{\mathbb{R}} [c'(\theta)d(y) + s'(\theta)]f(y|\theta)dy \\ &= \int_{\mathbb{R}} c'(\theta)d(y)f(y|\theta)dy + \int_{\mathbb{R}} s'(\theta)f(y|\theta)dy \\ &= c'(\theta)\int_{\mathbb{R}} d(y)f(y|\theta)dy + s'(\theta)\int_{\mathbb{R}} f(x|\theta)dy \\ &= c'(\theta)E[d(y)] + s'(\theta) = 0. \end{aligned}$$

Com isso, a esperança de $d(y)$ pode ser facilmente encontrada:

$$\begin{aligned} c'(\theta)E[d(y)] + s'(\theta) &= 0 \implies \\ \implies c'(\theta)E[d(y)] &= -s'(\theta) \implies \\ \implies E[d(y)] &= -\frac{s'(\theta)}{c'(\theta)} \end{aligned} \tag{2.5}$$

Além disso, a variância também pode ser determinada. Derivando 2.3 mais uma vez, verifica-se que:

$$\int_{\mathbb{R}} \frac{d^2 f(y|\theta)}{d\theta^2} dy = 0.$$

Logo, considerando $f(y|\theta) = \exp\{c(\theta)d(y) + s(\theta) + t(y)\}$, observa-se que:

$$\frac{d^2 f(y|\theta)}{d\theta^2} dy = [c''(\theta)d(y) + s''(\theta)]f(y|\theta) + [c'(\theta)d(y) + s'(\theta)]^2 f(y|\theta).$$

Por 2.5 e integrando a equação 2.6, obtém-se:

$$\begin{aligned} \int_{\mathbb{R}} \frac{d^2 f(y|\theta)}{d\theta^2} dy &= \int_{\mathbb{R}} [c''(\theta)d(y) + s''(\theta)]f(y|\theta) + [c'(\theta)]^2[d(y) - E[d(y)]]^2 f(y|\theta)dx \\ &= c''(\theta)E[d(y)] + s''(\theta) + [c'(\theta)]^2 Var[d(y)] = 0. \end{aligned} \quad (2.6)$$

Aplicando-se, então, 2.5 em 2.6, o valor de Var$[d(x)]$ é encontrado:

$$Var[d(y)] = \frac{c''(\theta)s'(\theta) - s''(\theta)c'(\theta)}{[c'(\theta)]^3}.$$

2.2.2 Função Escore

Com o objetivo de fazer inferências sobre os parâmetros dos MLG, pode ser usado a estatística *score (U)*,[19] que pode ser definida pela derivada da função log-verossimilhança, $l(\theta|y)$,

$$U(\theta|y) = \frac{dl(\theta|y)}{d\theta}$$

Aplicando a função log-verossimilhnça em uma função de probabilidade de uma distribuição pertencente à Família Exponencial (2.3) e então derivando-a, encontra-se:

$$\begin{aligned} U(\theta|y) &= \frac{dl(\theta|y)}{d\theta} \\ &= \frac{d\{c(\theta)d(y) + s(\theta) + t(y)\}}{d\theta} \\ &= c'(\theta)d(y) + s'(\theta). \end{aligned} \quad (2.7)$$

Portanto, o valor esperado da estatística score pode ser encontrado aplicando 2.5:

$$\begin{aligned} E[U] &= E[c'(\theta)d(y) + s'(\theta)] \\ &= c'(\theta)E[d(y)] + s'(\theta) \\ &= c'(\theta)\left\{\frac{-s'(\theta)}{c'(\theta)}\right\} + s'(\theta) \\ &= 0. \end{aligned}$$

O mesmo pode ser feito para encontrar a $Var[U]$, que é conhecida por informação. Logo:

$$\begin{aligned} Var[U] &= Var[c'(\theta)d(y) + s'(\theta)] \\ &= [c'(\theta)]^2]Var[d(y)], \end{aligned}$$

que por, 2.7 pode ser reescrita como

$$\begin{aligned} [c'(\theta)]^2]Var[d(y)] &= [c'(\theta)]^2]\left\{\frac{c''(\theta)s'(\theta) - s''(\theta)c'(\theta)}{[c'(\theta)]^3}\right\} \\ &= \frac{c''(\theta)s'(\theta) - s''(\theta)c'(\theta)}{c'(\theta)} \\ &= \frac{c''(\theta)s'(\theta)}{c'(\theta)} - s''(\theta). \end{aligned} \tag{2.8}$$

O que também pode ser constatado com $Var[U]$ é:

$$Var[U] = E[U^2] = -E[U'] \tag{2.9}$$

A primeira igualdade é facilmente demonstrada pela definição de variância:

$$Var[U] = E[U^2] - (E[U])^2.$$

Como $E[U] = 0$, isso implica em $(E[U])^2 = 0$, logo $Var[U] = E[U^2]$.

Para demonstrar a segunda equação, primeiro deriva-se U, em relação a θ:

$$U' = \frac{dU}{d\theta} = c''(\theta)d(y) + s''(\theta).$$

Logo, a esperança de U' é dada por:

$$\begin{aligned} E[U'] &= E[c''(\theta)d(y) + s''(\theta)] \\ &= c''(\theta)E[d(y)] + s''(\theta) \\ &= c''(\theta)\left\{\frac{-s'(\theta)}{c'(\theta)}\right\} + s''(\theta) \\ &= -Var[U]. \end{aligned}$$

2.2.3 Modelo de Poisson

O MLG que será utilizado neste trabalho será o Modelo de Poisson. A distribuição de Poisson é uma das indicadas quando a variável é discreta. Primeiro, uma variável aleatória Y tem distribuição Poisson, P(λ), sendo $\lambda > 0$, se sua função de probabilidade pode ser escrita na forma:

$$f(y|\lambda) = \frac{\exp\{-\lambda\}\lambda^y}{y!},$$

sendo $y \in \mathbb{N}$. O valor de sua esperança e de sua variância é o próprio parâmetro λ.

Colocando a função de probabilidade na forma da família exponencial, equação 2.3, observa-se que:

$$\begin{aligned} f(y|\lambda) &= \frac{\exp\{-\lambda\}\lambda^y}{y!} \\ &= \frac{\exp\{\log\{\exp\{-\lambda\}\}\}\exp\{\log\{\lambda^y\}\}}{\exp\{\log\{y!\}\}} \\ &= \exp\left\{\log\{\exp\{-\lambda\}\} + \log\{\lambda^y\} - \log\{y!\}\right\} \\ &= \exp\left\{y\log\{\lambda\} - \lambda - \log\{y!\}\right\}. \end{aligned} \tag{2.10}$$

Como $a(y)$ na equação 2.10 está em sua forma canônica, ou seja, está escrita em sua forma mais simplificada, por definição, $\log\{\lambda\}$ é o candidato a função de ligação entre a variável resposta e a combinação linear das variáveis explicativas com seus parâmetros, o que implica no Modelo de Poisson é dado por:

$$\log\{\lambda_i\} = \boldsymbol{\beta}\boldsymbol{X} = \eta_i, \tag{2.11}$$

sendo $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ o vetor de parâmetros do modelo, $\boldsymbol{X} = (x_{i1}, \ldots, x_{ik})^T$, o vetor

de variáveis explicativas, e k o número de variáveis explicativas.

O Modelo de Poisson é um bom modelo para variáveis que assumem valores discretos, o que é o caso deste trabalho, já que não existe 0,5 ponto, por exemplo; é esta a principal diferença desse modelo para outro MLG, o modelo normal, que se utiliza de dados contínuos. Além disso, o modelo apresenta poucas restrições.[18]

2.2.4 Estimação dos Parâmetros do MLG via Máxima Verossimilhança

A função log-verossimilhança de uma variável $\boldsymbol{Y} = (y_1, \ldots, y_n)$, com vetor de médias $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ e parâmetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, que tem distribuição pertencente a família exponencial e é discreta é dada por:

$$L = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \log\{f(\boldsymbol{Y}|\boldsymbol{\theta})\} = \sum_{i=1}^{n} \{y_i c(\theta_i) + s(\theta_i) + t(y_i)\} \tag{2.12}$$

Para encontrar, então, o estimador de máximo verossimilhança de cada β_j (sendo j= 1,. . . ,p; e p o número de variáveis explicativas), deve-se derivar 2.12 com relação ao próprio β_j, o que resulta em:

$$\frac{dL}{d\beta_j} = U_j = \sum_{i=1}^{n} \left\{ \frac{dl_i}{d\theta_i} . \frac{d\theta_i}{d\mu_i} . \frac{d\mu_i}{d\beta_j} \right\}, \tag{2.13}$$

sendo U_j o j-ésimo elemento do vetor escore $\boldsymbol{U}$.

As estimativas dos parâmetros são encontradas igualando cada U_j a 0. Porém, a equação 2.13 pode ser melhor desenvolvida. Primeiro, tendo em mente a equação 2.5, olha-se para a primeira derivada dessa equação:

$$\frac{dl_i}{d\theta_i} = y_i c'(\theta_i) + s'(\theta_i) = b'(\theta_i)(y_i - \mu_i).$$

Para segunda derivação, tendo em mente o encontrado em 2.8, verifica-se que:

$$\begin{aligned}\frac{d\theta_i}{d\mu_i} &= \frac{1}{\frac{d\mu_i}{d\theta_i}} \\ &= \frac{1}{\frac{s''(\theta_i)c'(\theta_i)-s'(\theta_i)c''(\theta_i)}{\{c'(\theta_i)\}^2}} \\ &= \frac{1}{c'(\theta_i)Var[Y_i]}.\end{aligned}$$

Já a terceira derivação, pode ser obtida a partir da equação 2.11, resultando em:

$$\frac{d\mu_i}{d\beta_j} = \frac{d\mu_i}{d\eta_i}.\frac{d\eta_i}{d\beta_j} = \frac{d\mu_i}{d\eta_i}x_{ij}.$$

Logo, pode-se dizer que

$$U_j = \sum_{i=1}^{n}\left\{\frac{(y_i-\mu_i)}{Var[Y_i]}.\frac{d\mu_i}{d\eta_i}.x_{ij}\right\}.$$

A matriz de variância-covariância (τ, formada pelos termos τ_{jk}) de $\boldsymbol{U}$ pode ser dada a partir da equação acima:

$$\begin{aligned}\tau_{jk} = E[U_jU_k] &= E\left[\sum_{i=1}^{n}\left\{\frac{(y_i-\mu_i)}{Var[Y_i]}.\frac{d\mu_i}{d\eta_i}.x_{ij}\right\}.\sum_{l=1}^{n}\left\{\frac{(y_l-\mu_l)}{Var[Y_l]}.\frac{d\mu_l}{d\eta_l}.x_{lk}\right\}\right] \\ &= \sum_{i=1}^{n}\left\{\frac{E[(y_i-\mu_i)^2]x_{ij}x_{ik}}{(Var[Y_i)^2]}.\left(\frac{d\mu_i}{d\eta_i}\right)\right\} \\ &= \sum_{i=1}^{n}\left\{\frac{x_{ij}x_{ik}}{Var[Y_i]}.\left(\frac{d\mu_i}{d\eta_i}\right)\right\}.\end{aligned}$$

Com todas essas informações, pode ser encontrado o vetor de estimativas, $\boldsymbol{\beta}$, dos parâmetros do modelo, pela equação de estimação via método de máxima verossimilhança, utilizando a função escore. Ela utiliza o método iterativo de Newton-Raphson e é escrito da seguinte forma:

$$\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m-1)} + (\boldsymbol{\tau}^{(m-1)})^{-1}\boldsymbol{U}^{(m-1)},$$

sendo $\hat{\boldsymbol{\beta}}^{(m)}$ o vetor de estimativas dos parâmetros $\boldsymbol{\beta}$ na m-ésima iteração. São feitas então diferenças entre as consecutivas aproximações de $\hat{\boldsymbol{\beta}}^{(m-1)}$ e $\hat{\boldsymbol{\beta}}^{(m)}$, quando uma dessas diferenças é suficientemente pequena, o vetor $\hat{\boldsymbol{\beta}}^{(m)}$ é dito o vetor de estimativa dos

parâmetros via máximo verossimilhança. Mais informações podem ser encontradas no livro *An Introduction to Generalized Linear Models*, de Annette J. Dobson.[19]

2.3 Seleção de variáveis

Além dos métodos que serão apresentados nesta seção, também será utilizado o conhecimento prévio dos fundamentos de basquetebol. Essa visão pode encaixar variáveis que mesmo não sendo numericamente importantes, podem ser significativas no jogo.

2.3.1 Teste de Significância Individual de Wald

O Teste de Significância Individual de Wald verifica qual a real relação que a variável explicativa tem com o modelo, ou seja, se tal variável explicativa é significativa ou não. Para isso, é feito um teste com as seguintes hipóteses:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0, \end{cases}$$

sendo β_i o parâmetro relativo a i-ésima variável. Ou seja, se $\beta_i \neq 0$, significa dizer que a variável é estatisticamente significante para o modelo.

Para chegar a essa conclusão, é necessário utilizar analisar o p-valor obtido com a seguinte estatística de teste, sob H_0, verdadeiro:

$$W = \frac{\hat{\beta}_i^2}{V\hat{a}r[\hat{\beta}_i]} \sim \chi_1^2.$$

Caso o p-valor seja menor do o nível de significância $\alpha = 0,05$ (será usado em este valor em todo o trabalho), rejeita-se a hipótese nula, logo a j-ésima variável é estatisticamente significante para o modelo.[20]

2.3.2 Método LASSO

O método LASSO (Least Absolute Shrinkage and Selection Operator) é um método de redução de dimensão de modelos, mas como zera os coeficientes, é útil na seleção de variáveis. Para isso, ele estima os parâmetros, regularizando-os para tenderem a 0. Ele

pode ser usado em diversos tipos de modelo, mas nos MLG's sua fórmula é dada por:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{min} \quad \frac{1}{N}\sum_{i=1}^{N} w_i l(y_i, \eta_i) + \lambda \sum_{j=1}^{p} |\beta_j|,$$

onde $l(y_i, \eta_i)$ é a função log-verossimilhança negativa para a observação i, que é usada pois minimizá-la é equivalente a maximizar a função log-verossimilhança, o que permite utilizar a penalização. E essa é dada por $\lambda \sum_{j=1}^{p} |\beta_j|$, que tem efeito na estimação dos parâmetros, permitindo viés porém reduzindo a variância. Onde $\lambda \geq 0$ é um parâmetro de diminuição da quantidade de parâmetros do modelo, ou seja, quanto menor λ, maior a quantidade de estimativas dos parâmetros iguais a zero, implicando em menos parâmetros. Mais informações podem ser encontradas em Tibshirani (1996) e Friedman *et al.* (2010).[21][22]

Após a verificação de diversos valores para as estimativas de $\boldsymbol{\beta}$, o melhor modelo será escolhido por validação cruzada. Para realizar este processos, será utilizado o pacote *"glmnet"*.[23]

2.4 Métodos de comparação e avaliação dos modelos

2.4.1 Teste de Comparabilidade de Modelos

O teste de comparabilidade de modelos é feito comparando dois modelos 0 e 1 que possuem a mesma distribuição de probabilidade, mesma função de ligação e com quantidades p e q, respectivamente, distintas de variáveis explicativas, mas que as variáveis em 0 sejam parte das variáveis em 1, ou seja $q < p < N$, sendo N o tamanho da amostra. Logo as hipóteses da comparação são dadas por:

$$\begin{cases} H_0 : \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix} \\ \\ H_1 : \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \\ \vdots \\ \beta_p \end{bmatrix} \end{cases}.$$

A estatística utilizada neste teste é a estatística *deviance*, também conhecida como estatística log-verossimilhança. Ela consiste em:

$$D = 2[l(\hat{\boldsymbol{\beta}}_{max}|\boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}|\boldsymbol{y})],$$

sendo $\hat{\boldsymbol{\beta}}_{max}$ o vetor com todos os N parâmetros que o modelo pode assumir. Este modelo é denominado de saturado.

Aplicando a diferença das estatísticas *deviance* para os modelos 0 e 1, obtém-se:

$$\begin{aligned} \Delta D = D_0 - D_1 &= 2[l(\hat{\boldsymbol{\beta}}_{max}|\boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}_0|\boldsymbol{y})] - 2[l(\hat{\boldsymbol{\beta}}_{max}|\boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}_1|\boldsymbol{y})] \\ &= 2[l(\hat{\boldsymbol{\beta}}_1|\boldsymbol{y}) - l(\hat{\boldsymbol{\beta}}_0|\boldsymbol{y})]. \end{aligned}$$

Sabendo que $D_0 \sim \chi^2_{(N-q)}$ e $D_1 \sim \chi^2_{(N-p)}$, implica que $\Delta D \sim \chi^2_{(p-q)}$, com a condição de independência garantida. Logo, pode ser realizada a mesma aplicação através do p-valor: caso o p-valor encontrado seja menor que α, rejeita-se a hipótese nula, logo o modelo 1 é mais indicado que o modelo 0 para descrever os dados. Se for maior, a hipótese nula é aceita e então o modelo 0 é o mais indicado. [19]

2.4.2 Critério de Informação de Akaike (AIC)

O AIC foi desenvolvido por Hirotugu Akaike em 1974 e tem como principal função medir a qualidade de um modelo ajustado. Para isso, esse critério utiliza a função de log-verossimilhança e sua fórmula é dada por:

$$AIC = -2log(l(\hat{\boldsymbol{\beta}})) + 2p,$$

onde $L(\hat{\boldsymbol{\beta}})$ é a função de log-verossimilhança com as estimativas dos parâmetros do modelo e p é a quantidade de parâmetros que o modelo possui. A medida funciona de modo que quanto menor o AIC, melhor ajustado é o modelo em questão.[24]

3 Resultados e Análises

3.1 Análise Descritiva dos Dados

Nesta seção será realizada uma breve análise dos dados utilizados neste trabalho, com o objetivo de entender como se comportam. Primeiro, será analisada a variável de interesse desta monografia, a quantidade de pontos marcados pelos times.

Pelas Figura 1 e Tabela 3 abaixo, percebe-se que a maior quantidade de pontos marcada foi de 149 pontos, feito realizado por *Golden State Warriors* e *Miami Heat* em 2016 e 2018, respectivamente. Já a menor quantidade, foi marcada pelo *Dallas Mavericks* também em 2016. No período em questão, a média de pontos marcados foi de 103,7 pontos com uma variância de 148,5. Além disso, apenas 25% das partidas tiveram o time principal da observação marcando menos de 95 pontos e 25% marcando mais que 112 pontos, o que indica pontuações centradas entre esses valores nas temporadas 2014-15 e 2017-18 da NBA.

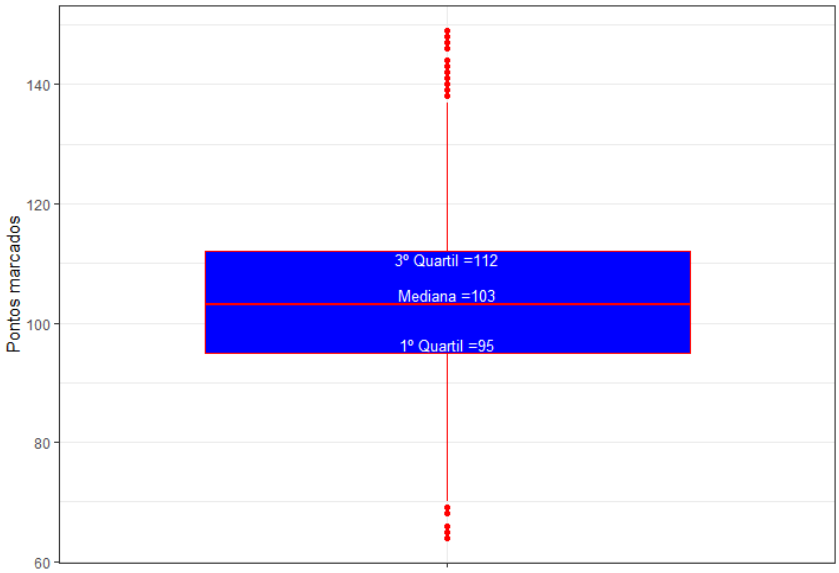

Figura 1: Boxplot da quantidade de pontos

Tabela 3: Medidas de tendência central com relação aos pontos

Mínimo	1º Quartil	Mediana	Média	Variância	3º Quartil	Máximo
64	95	103	103,7	148,5	112	149

Focando nos dados do times individualmente, na Tabela 4, observa-se que o campeão de três das quatro temporadas analisadas, *Golden State Warriors*, foi o que teve a maior pontuação média com 113,5 pontos por partida. O *Cleveland Cavaliers*, campeão da outra temporada, ficou em terceiro lugar com uma média de 107,2 pontos por partida. Na última posição aparece o *Memphis Grizzlies* com apenas 99,3 pontos por partida.

Tabela 4: Média de pontos de cada time por partida

Conferência Oeste		Conferência Leste	
Time	Média de pontos	Time	Média de pontos
MEM	99,3	NYK	99,8
UTA	99,4	MIA	100,3
DAL	101,9	PHI	100,4
LAL	102,1	ORL	100,6
SAC	102,4	DET	101,4
PHO	103,7	MIL	101,7
SAS	103,7	CHI	102,1
MIN	103,8	BRK	102,3
NOP	104,5	IND	102,5
POR	105,4	CHO	102,7
DEN	106,3	ATL	103,0
LAC	107,2	WAS	104,6
OKC	107,2	BOS	104,8
HOU	109,5	TOR	106,3
GSW	113,5	CLE	107,2

Com relação às vitórias e derrotas nas temporadas e levando em consideração que ocorreram 328 partidas no período em questão, a Figura 2 abaixo mostra que o time com maior porcentagem de vitórias (80,8%) e consequentemente, a menor porcentagem de derrotas (19,2%) foi o *Golden State Warriors*. Já o time com maior porcentagem de derrotas (69,8%) e menor de vitórias (30,2%) foi o *Los Angeles Lakers*. A Tabela 17 no apêndice B contém os valores apresentados no gráfico para melhor visualização.

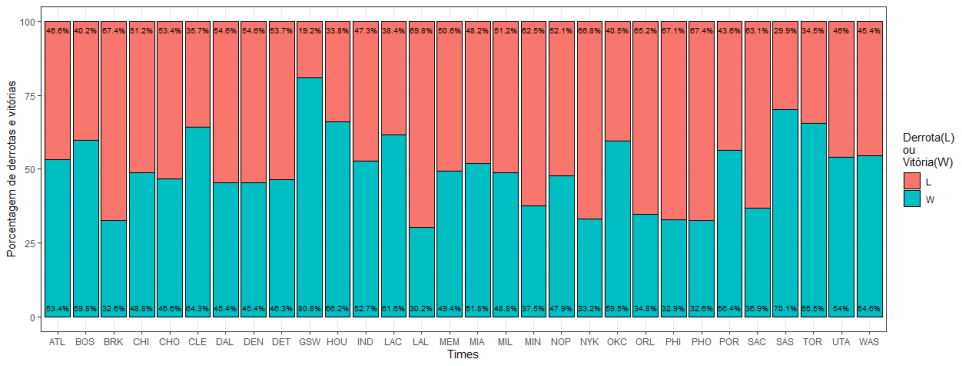

Figura 2: Gráfico para as porcentagens de vitória ou derrota de cada time

A nível de interesse a Tabela 5 abaixo contém as médias por partida das variáveis de fundamentos do basquete.

Tabela 5: Médias das variáveis de fundamentos do basquete por partida

Variável	Média	Variável	Média
Quantidade de arremessos certos (FieldGoals)	38,6	Porcentagem de lances livres certos (FreeThrowsPerc)	76,24%
Quantidade de arremessos realizados (FieldGoalsAttempted)	84,9	Quantidade de rebotes ofensivos (OffRebounds)	10,3
Porcentagem de arremessos certos (FieldGoalsPerc)	45,57%	Quantidade de rebotes (TotalRebounds)	43,5
Quantidade de arremesos de 3 pontos certos (X3PointShots)	9,1	Quantidade de assistências (Assists)	22,5
Quantidade de arrmessos de 3 pontos realizados (X3PointShotsAttempted)	25,6	Quantidade de roubos de bola (Steals)	7,7
Porcentagem de arremessos de 3 pontos certos (X3PointShotsPerc)	35,43%	Quantidade de tocos (Blocks)	4,8
Quantidade de lances livres certos (FreeThrows)	17,3	Quantidade de perdas de bola (Turnovers)	13,6
Quantidade de lances livres realizados (FreeThrowsAttempted)	22,7	Quantidade total de faltas (TotalFouls)	20,1

3.2 Organização da base de dados

Para a construção dos modelos, algumas alterações e organizações foram feitas nas bases de dados. A primeira delas foi a exclusão das variáveis “Game” e “Date” pois estas não tinham nenhum tipo de importância para este trabalho.

Foi criada a variável “Temporada” que indica em qual das quatro temporadas, a que a

base de dados pertence, a partida foi realizada. Essa criação foi feita pois os times podem mudar muito de uma temporada para outra: como a adição de novos jogadores, perda de outros, mudanças na comissão técnica, etc. Isto pode fazer com que os times melhorem ou piorem seu rendimento entre as temporadas.

Por fim, a base foi dividida em bases de treino e teste. A base de treino seria utilizada para ajustar os modelo e base de teste, como o nome já diz, para testar os modelos. Elas foram criadas de forma que 75% dos primeiros jogos de cada time em cada temporada foram alocados na base de treino, e os 25% restantes foram alocados na base de teste.

3.3 Modelos

Ao todo foram ajustados sete modelos que utilizavam a base de dados treino criada e a variável "TeamPoints" (pontos marcados) como variável resposta. As variáveis categóricas "Home" e "Temporada" foram transformadas em variáveis do tipo *dummy* e suas categorias referências foram "Away" e "Primeira", respectivamente. A seguir, todos os modelos serão apresentados com tabelas para as estimativas dos parâmetros.

O primeiro modelo, chamado de Modelo 1 é o mais completo de todos, possuindo todas as variáveis de fundamentos do basquete, tanto para o time principal quanto para o time oponente, mais as variáveis "Home" e "Temporada". As estimativas para este modelo podem ser encontradas na Tabela 6 abaixo, além dos p-valores provenientes dos testes de Wald para a significância das variáveis.

Tabela 6: Estimativas dos parâmetros para o Modelo 1

Variáveis	Estimativas	P-valor	Variáveis	Estimativas	P-valor
Intercepto	3,341e+00	<0,001	Turnovers	-1,187e-04	0,8609
HomeHome	6,856e-05	0,9768	TotalFouls	-6,874e-06	0,9884
TemporadaQuarta	8,238e-04	0,8258	Opp.FieldGoals	-1,244e-03	0,6802
TemporadaSegunda	8,127e-04	0,8069	Opp.FieldGoalsAttempted	3,827e-04	0,7917
TemporadaTerceira	5,862e-04	0,8681	Opp.FieldGoalsPerc	9,992e-02	0,6913
FieldGoals	1,319e-02	<0,001	Opp.3PointShots	-2,087e-04	0,9019
FieldGoalsAttempted	2,747e-03	0,0584	Opp.3PointShotsAttempted	8,175e-05	0,8978
FieldGoalsPerc	5,327e-01	0,0337	Opp.3PointShotsPerc	3,683e-03	0,9281
X3PointShots	7,852e-03	<0,001	Opp.FreeThrows	-1,616e-04	0,9120
X3PointShotsAttempted	6,266e-04	0,3286	Opp.FreeThrowsAttempted	4,321e-05	0,9703
X3PointShotsPerc	3,920e-02	0,3361	Opp.FreeThrowsPerc	3,001e-03	0,9257
FreeThrows	9,062e-03	<0,001	Opp.OffRebounds	7,658e-05	0,9182
FreeThrowsAttempted	4,413e-04	0,7043	Opp.TotalRebounds	5,274e-06	0,9918
FreeThrowsPerc	1,481e-02	0,6481	Opp.Assists	-1,697e-05	0,9567
OffRebounds	6,248e-05	0,9331	Opp.Steals	4,427e-05	0,9427
TotalRebounds	-1,179e-05	0,9816	Opp.Blocks	3,078e-05	0,9523
Assists	-7,606e-05	0,8063	Opp.Turnovers	-6,193e-05	0,9272
Steals	-3,313e-05	0,9571	Opp.TotalFouls	-3,800e-05	0,9357
Blocks	-5,367e-05	0,9162			

Os próximos modelos são modelos encaixados ao Modelo 1, ou seja, suas variáveis

explicativas se encontram dentro do conjunto de variáveis explicativas ao Modelo 1. O primeiro deles, chamado de Modelo 2, tem apenas duas variáveis a menos: "Home" e "Temporada". Ou seja, é um modelo com apenas as variáveis quantitativas. Isto foi feito para verificar se o fato de o time jogar ou não em casa tem melhora ou não o modelo. O mesmo pensamento vale para a temporada. Suas estimativas e p-valores referentes ao teste de Wald se encontram na Tabela 7 abaixo.

Tabela 7: Estimativas dos parâmetros para o Modelo 2

Variáveis	Estimativas	P-valor	Variáveis	Estimativas	P-valor
Intercepto	3,341e+00	<0,0001	Opp.FieldGoals	-1,220e-03	0,6856
FieldGoals	1,321e-02	<0,0001	Opp.FieldGoalsAttempted	3,728e-04	0,7969
FieldGoalsAttempted	2,738e-03	0,0591	Opp.FieldGoalsPerc	9,897e-02	0,6940
FieldGoalsPerc	5,318e-01	0,0340	Opp.3PointShots	-2,157e-04	0,8985
X3PointShots	7,844e-03	<0,0001	Opp.3PointShotsAttempted	9,370e-05	0,8820
X3PointShotsAttempted	6,392e-04	0,3151	Opp.3PointShotsPerc	3,850e-03	0,9249
X3PointShotsPerc	3,938e-02	0,3336	Opp.FreeThrows	-1,668e-04	0,9090
FreeThrows	9,056e-03	<0,0001	Opp.FreeThrowsAttempted	4,743e-05	0,9673
FreeThrowsAttempted	4,463e-04	0,7008	Opp.FreeThrowsPerc	3,265e-03	0,9191
FreeThrowsPerc	1,509e-02	0,6418	Opp.OffRebounds	6,607e-05	0,9293
OffRebounds	5,172e-05	0,9445	Opp.TotalRebounds	1,109e-05	0,9827
TotalRebounds	-5,637e-06	0,9912	Opp.Assists	-2,367e-05	0,9391
Assists	-8,098e-05	0,7924	Opp.Steals	4,160e-05	0,9461
Steals	-3,683e-05	0,9522	Opp.Blocks	3,701e-05	0,9425
Blocks	-4,580e-05	0,9282	Opp.Turnovers	-5,793e-05	0,9319
Turnovers	-1,148e-04	0,8653	Opp.TotalFouls	-3,855e-05	0,9345
TotalFouls	-9,000e-06	0,9847			

Os modelos seguintes foram ajustados tendo como variáveis explicativas, as mais significativas de acordo com o teste de Wald, realizado a partir do Modelo 1, e as variáveis qualitativas "Home" e "Temporada", estando presentes de maneira alternada. Considerando um nível de significância de 5% e analisando os p-valores da Tabela 6, as variáveis mais significativas são: FieldGoals, FieldGoalsPerc, X3PointShots e FreeThrows. Pensando de maneira lógica, faz sentido estas variáveis serem as mais significativas, afinal, são as variáveis relacionadas aos arremessos feitos pelo time que originam os pontos marcados.

O Modelo 3 possui, além das significativas, a variável "Home" que pode ter influência nos resultados das partidas. Suas estimativas e p-valores referentes ao teste de Wald se encontram na Tabela 8 abaixo. O Modelo 4, que tem suas estimativas e p-valores na Tabela 9, possui a variável "Temporada" no lugar da variável "Home". O Modelo 5 (Tabela 10) possui ambas as variáveis categóricas e o Modelo 6 (Tabela 11), nenhuma delas.

Tabela 8: Estimativas dos parâmetros para o Modelo 3

Variáveis	Estimativas	P-valor
Intercepto	3,6348310	<0,0001
HomeHome	0,0002050	0,9290
FieldGoals	0,0190061	<0,0001
FieldGoalsPerc	0,0331099	0,3710
X3PointShots	0,0094594	<0,0001
FreeThrows	0,0095697	<0,0001

Tabela 9: Estimativas dos parâmetros para o Modelo 4

Variáveis	Estimativas	P-valor
Intercepto	3,6345853	<0,0001
TemporadaQuarta	0,0007219	0,8300
TemporadaSegunda	0,0007329	0,8230
TemporadaTerceira	0.0006359	0,8480
FieldGoals	0,0189980	<0,0001
FieldGoalsPerc	0,0303173	0,3610
X3PointShots	0,0094439	<0,0001
FreeThrows	0,0095680	<0,0001

Tabela 10: Estimativas dos parâmetros para o Modelo 5

Variáveis	Estimativas	P-valor
Intercepto	3,6345805	<0,0001
HomeHome	0,0002111	0,9270
TemporadaQuarta	0,0007251	0,8300
TemporadaSegunda	0,0007345	0,8230
TemporadaTerceira	0,0006396	0,8470
FieldGoals	0,0189971	<0,0001
FieldGoalsPerc	0,0302179	0,3630
X3PointShots	0,0094437	<0,0001
FreeThrows	0,0095669	<0,0001

Tabela 11: Estimativas dos parâmetros para o Modelo 6

Variáveis	Estimativas	P-valor
Intercepto	3,6348351	<0,0001
FieldGoals	0,0190070	<0,0001
FieldGoalsPerc	0,0297447	0,3690
X3PointShots	0,0094595	<0,0001
FreeThrows	0,0095708	<0,0001

Por fim, o Modelo 7 foi ajustado utilizando o Método LASSO de seleção de variáveis, utilizando o valor de $\lambda = 0,02882747$, que minimiza a deviance. Utilizando a validação cruzada foram selecionadas as variáveis: FieldGoals, FieldGoalsAttempted, FieldGoalsPerc, X3PointShots, X3PointShotsAttempted, X3PointShotsPerc, FreeThrows, FreeThrowsPerc, OffRebounds, Turnovers, TotalFouls e Opp.FieldGoalsAttempted. Suas estimativas estão na Tabela 12. Além das variáveis relacionadas aos arremessos, a presença das outras variáveis também pode ser explicada pensando no basquetebol em si. Quando você tem um rebote ofensivo, você tem uma nova chance de arremessar a bola, portanto pode ter influência nos pontos. Para as perdas de bola, você perde a chance arremessar e marcar pontos, o que pode influenciar de maneira negtiva a quantidade de pontos marcadas.

3.4 Comparação entre os modelos e predições

3.4.1 Aplicação do AIC e comparabilidade dos modelos

A Tabela 13 abaixo apresenta os valores do Critério de Informação de Akaike. Percebe-se que todos os modelos tiveram valores de AIC muito próximos uns aos outros. Porém,

Tabela 12: Estimativas dos parâmetros para o Modelo 7

Variáveis	Estimativas
Intercepto	3,624029e+00
FieldGoals	1,875049e-02
FieldGoalsAttempted	5,821002e-05
FieldGoalsPerc	4,848485e-02
X3PointShots	8,346260e-03
X3PointShotsAttempted	3,971953e-04
X3PointShotsPerc	2,614594e-02
FreeThrows	9,491394e-03
FreeThrowsPerc	2,847364e-03
OffRebounds	8,422852e-05
Turnovers	-1,555736e-04
Opp.FieldGoalsAttempted	-1,810319e-05

os modelos que possuem a maior quantidade de variáveis (modelos 1 e 2) foram os que tiveram os maiores valores. O menor valor, e portanto, o modelo que melhor se adequa aos dados segundo o AIC, foi o Modelo 6.

Tabela 13: Valores de AIC para os modelos ajustados

Modelos	AIC
Modelo 1	48287
Modelo 2	48279
Modelo 3	48233
Modelo 4	48237
Modelo 5	48239
Modelo 6	48231
Modelo 7	48243

Com relação à adequabilidade dos modelos, foram calculados algumas medidas e realizado o teste de comparabilidade de modelos, que, como falado anteriormente, utiliza a *deviance* dos modelos.

Primeiro foram realizadas as comparações entre os modelos 2, 3, 4, 5, 6 e 7 com o Modelo 1, pois todos eram encaixados ao primeiro modelo. Todos os p-valores foram maiores que 0,9, que, ao nível de significância de 5%, sugere que os modelos supracitados se adequam melhor aos dados do que o Modelo 1. Após, os modelos 6 e 7 foram comparados com o Modelo 2, e também foi indicado que os modelos menores eram melhor que o maior modelo. Em seguida foram comparados os modelos 3, 4 e 5, sendo os dois primeiros encaixados ao último. Mais uma vez os modelos com dimensão menor foram considerados os possuidores de melhor ajuste aos dados, sendo seus p-valores também maiores que 0,9. Por fim, compararam-se os modelos 2, 3, 4, 5, 6 e 7, sendo o Modelo 6 encaixado aos demais. O Modelo 6 foi considerado o que se melhor ajusta aos dados em todas as comparações.

Com todos os resultados, os teste de comparabilidade mostram que os modelos com menos variáveis são os de melhor ajuste, sendo o menor deles (Modelo 6), o que melhor se ajustou em comparação com os demais. A seguir, encontra-se a Tabela 14 com os p-valores de todas as comparações feitas com o teste de comparabilidade de modelos. Ela está dividida em 4 "níveis" de comparação: nas primeira, segunda e terceira partes, os modelos eram encaixados aos modelos 1, 2 e 5, respectivamente; na última o Modelo 6 era encaixado aos demais.

Tabela 14: P-valores dos testes de comparabilidade de modelos aplicados

Modelo 1	
Modelos	P-valor
Modelo 2	0,9994
Modelo 3	1
Modelo 4	1
Modelo 5	1
Modelo 6	1
Modelo 7	1
Modelo 2	
Modelos	P-valor
Modelo 6	1
Modelo 7	1
Modelo 5	
Modelos	P-valor
Modelo 3	0,9955
Modelo 4	0,9267
Modelo 6	
Modelos	P-valor
Modelo 3	0,9287
Modelo 4	0,9956
Modelo 5	0,9993
Modelo 7	0,9046

3.4.2 Predição

A última tarefa deste trabalho foi realizar predições para algumas partidas e verificar o quão boas eram essas predições. Foram realizadas previsões para todas as últimas partidas de cada time, em cada temporada. Para realizar a previsão de quantos pontos os times marcariam na partida, as médias das 19 partidas anteriores (que estavam na base teste), além do local da partida e em qual temporada aconteceu, foram utilizados como valores para as variáveis explicativas dos 7 modelos ajustados. Alguns times (identificados por um asterisco nas tabelas do apêndice C) participaram dos últimos jogos de dois outros

times, e por isso, para essa partidas, foram utilizadas as médias dos 18 últimos jogos para a partida que ocorreu primeiro.

Com esses valores definidos, as fórmulas dos 7 modelos foram aplicadas, e o resultado foi utilizado como a média de uma distribuição Poisson para gerar amostras de 1000 valores para cada time. Após, foram comparados cada valor de cada amostra dos times, gerando assim 1000000 de resultados possíveis para cada partida. Com as pontuações definidas, foram verificadas quantas dessas partidas previram corretamente a vitória ou derrota dos times. As Tabelas 18, 19, 20 e 21 do anexo C indicam as porcentagens de acerto para cada modelo em cada uma das partidas. A Tabela 15 mostra as médias das porcentagens de acerto de cada simulação para cada modelo, em cada uma das temporadas, além da média geral.

Tabela 15: Porcentagem de acerto dos resultados

Temporadas	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7
Primeira temporada	59,01%	58,98%	58,99%	59,00%	59,01%	58,98%	58,93%
Segunda temporada	51,18%	51,13%	51,22%	51,19%	51,22%	51,18%	51,16%
Terceira temporada	51,01%	51,01%	51,09%	51,01%	51,04%	51,03%	51,01%
Quarta temporada	48,39%	48,41%	48,42%	48,42%	48,43%	48,42%	48,42%
Médias	52,40%	52,38%	52,43%	52,41%	52,43%	52,40%	52,38%

Os modelos tiveram suas médias das predições muito próximas, estando entre 52,38% (Modelo 2 e 7) e 52,43% (Modelos 5 e 3). Algo que chama a atenção foi a grande dificuldade dos modelos em predizer corretamente algumas partidas, estando algumas vezes abaixo de 30%.

Analisando especificamente dois confrontos distintos (*Utah Jazz* contra *Los Angeles Lakers* na segunda temporada e *Cleveland Cavaliers* contra *Toronto Raptors* na terceira temporada), nota-se que o *Utah Jazz* ganhou 12 das 19 partidas anteriores, tendo uma média de 97,05 pontos marcados e 91,42 pontos sofridos, enquanto o *Los Angeles Lakers* ganhou apenas 4 partidas com médias de 95,58 e 106,4 pontos marcados e sofridos, respectivamente. Olhando para a outra partida, com o placar de 98 a 83 para o time canadense, apesar de o *Cleveland Cavaliers* ter apenas ganhado 9 jogos e o *Toronto Raptors*, 13 jogos, o time de *Ohio* teve médias de 110,4 pontos marcados e 109,8 pontos sofridos, enquanto seu adversário, 103,2 e 99,05, respectivamente. Isso mostra que os modelos tiveram dificuldades em predizer os resultados de jogos considerados "surpresa", pois informavam que o time com o melhor histórico marcaria mais pontos e teria uma chance maior de vencer a partida.

Porém, alumas predições foram interessantes de uma boa maneira, como, por exemplo, a previsão de vitória na quarta temporada do *Utah Jazz* em cima do *Golden State*

Warriors, que participou de todas as finais das temporadas utilizadas neste trabalho, sendo campeão em 3 dela, além de em uma dessas temporadas ter estabelecido o recorde de vitórias em uma só temporada.

Baseado nas comparações dos modelos via teste de comparabilidade de modelos e AIC, além dos resultados das predições, pode-se dizer que entre os 7 modelos ajustados aquele que teve uma performance ligeiramente melhor, foi o Modelo 3, que possui as variáveis mais significativas e a variável indicando o local da partida. Apesar de ser considerado inferior ao Modelo 6 pelo teste de comparabilidade dos modelos e ter o segundo menor AIC, seu valor foi muito próximo ao menor com uma diferença de apenas 2 unidades e teve a maior predição média de todos os modelos.

4 Conclusão

Sem dúvidas, a NBA sempre foi uma das maiores ligas esportivas do mundo, tendo diversos personagens conhecidos, craques, times lendários e momentos marcantes durante toda sua história. Essa popularidade se manteve até os dias atuais e a tendência é crescer no futuro, atraindo fãs de todos os cantos do mundo e movimentando bilhões de dólares todos os anos. Isso faz com que a procura por entender melhor como essa liga funciona, seja por motivo de lazer, profissional, ou até mesmo para a realização de apostas, é muito grande. Uma das ferramentas que ajuda nesse entendimento é a estatística.

Como mencionado anteriormente, a estatística sempre esteve presente no mundo esportivo e não seria diferente para um grande campeonato. Com o intuito de realizar previsões para os jogos da NBA foram ajustados 7 modelos lineares generalizados que tinham como variável resposta a quantidade de pontos marcado por um time em uma determinada partida, sendo essa variável pertencente a uma distribuição de Poisson. Os modelos continham variáveis de fundamentos básicos como variáveis respostas, além de variáveis indicando quando o jogo ocorreu e em que local. A partir dos resultados encontrados com os modelos foram simulados 1000000 de partidas e verificado o resultado de vitória ou derrota em cada uma.

Os modelos tiveram certas dificuldades de prever os resultados, principalmente em jogos em que o time que tinha o melhor histórico, perdia o jogo. As médias de porcentagem de acerto encontradas pelos modelos não foram tão altos, sendo a mais alta de 52,43%. Assim como grande parte dos esportes, o basquete é uma modalidade difícil de predizer a quantidade de pontos, pois muitas são as reviravoltas que podem ocorrer durante a temporada, e até mesmo durante uma partida.

Em geral, o objetivo deste trabalho foi alcançado: foram ajustados e comparados 7 modelos diferentes e definido qual deles era o melhor. Para trabalhos futuros, seria interessante a utilização de outros MLG's, não necessariamente para a previsão dos pontos marcados, mas talvez para o resultado entre vitória ou derrota. Além disso, a utilização

de algumas outras variáveis que pudessem dizer, por exemplo, através de análise de especialistas, quem é o favorito ao jogo (como fizeram Louzada *et al.* (2015)[15]) , ou até mesmo variáveis que quantificassem o histórico do clube durante a temporada.

Referências

1 CAMARGO, V. L. *Era de Gigantes: A História do Basquete Profissional Norte-Americano no Século XX*. 1. ed. [S.l.]: Amazon, 2019.

2 ESPN. *Vinte anos atrás, um dos melhores drafts da história mudou o rumo da NBA*. 2016. ⟨http://www.espn.com.br/blogs/nbanaespn/608583_vinte-anos-atras-um-dos-melhores-drafts-da-historia-mudou-o-rumo-da-nba⟩. Acesso em: 14 nov. 2020.

3 SPORTS REFERENCE. *Basketball Reference*. ⟨https://www.basketball-reference.com/⟩. Acesso em: 14 nov. 2020.

4 BADENHOUSEN, K. *NBA Team Values 2019: Knicks On Top At $4 Billion*. 2019. ⟨https://www.forbes.com/sites/kurtbadenhausen/2019/02/06/nba-team-values-2019-knicks-on-top-at-4-billion/?sh=6685584e6671⟩. Acesso em: 25 nov. 2020.

5 GOUGH, C. *National Basketball Association (NBA) - Statistics & Facts*. 2020. ⟨https://www.statista.com/topics/967/national-basketball-association/⟩. Acesso em: 25 nov. 2020.

6 KOZLOWSKI, J. *How Many World Countries Have Hosted NBA Games?* 2019. ⟨https://www.sportscasting.com/how-many-countries-have-hosted-nba-games/⟩. Acesso em: 25 nov. 2020.

7 SAIIDI, U. *The NBA is China's most popular sports league. Here's how it happened*. 2018. ⟨https://www.cnbc.com/2018/11/20/the-nba-is-chinas-most-popular-sports-league-heres-how-it-happened.html⟩. Acesso em: 25 nov. 2020.

8 COSTA Ígor Barbosa da; PIRES, C. E. S.; MARINHO, L. B. Sports analytics: Mudando o jogo. In: *Tópicos em Gerenciamento de Dados e Informações 2017*. Uberlândia, Brasil: [s.n.], 2017. p. 30–61. ISBN 978-85-7669-400-7.

9 NBASTUFFER. *Dean Oliver*. ⟨https://www.nbastuffer.com/analytics101/dean-oliver/⟩. Acesso em: 14 nov. 2020.

10 PRICE, M. *Advanced Stats in Basketball: An Explainer Series*. 2016. ⟨https://medium.com/@mrprice33/advanced-stats-in-basketball-an-explainer-series-f5fdaf8f1c39⟩. Acesso em: 14 nov. 2020.

11 OLIVER, D. *Basketball on Paper: Rules and Tools for Performance Analysis*. Brassey's, Incorporated, 2004. ISBN 978-15-7488-687-0. Disponível em: ⟨https://books.google.com.br/books?id=Xh2iSGCqJJYC⟩.

12 SHANAHAN, K. J. *A Model for Predicting the Probability of a win in Basketball*. Dissertação (Mestrado) — Universidade de Iowa, Iowa, Estados Unidos da América, 1984.

13 LOPEZ, M. J.; MATTHEWS, G. J. Building an ncaa men's basketball predictive model and quantifying its success. *Journal of Quantitative Analysis in Sports*, v. 11, 2015.

14 CHENG, G. et al. Predicting the outcome of nba playoffs based on the maximum entropy principle. *Entropy*, v. 18, n. 12, 2016. ISSN 1099-4300. Disponível em: ⟨https://www.mdpi.com/1099-4300/18/12/450⟩.

15 LOUZADA, F. et al. SIMULATION-BASED METHODOLOGY FOR PREDICTING FOOTBALL MATCH OUTCOMES CONSIDERING EXPERTS' OPINIONS: THE 2010 AND 2014 FOOTBALL WORLD CUP CASES. *Pesquisa Operacional*, scielo, v. 35, p. 577 – 598, 12 2015. ISSN 0101-7438. Disponível em: ⟨http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-74382015000300577&nrm=iso⟩.

16 MAHFUZ, T. O. *Modelagem do número de gols marcados usando a Regressão de Poisson*. 2009. Monografia (Graduação em Ciências Econômicas), IBMEC São Paulo, São Paulo, Brasil.

17 KELEPOURIS, I. *NBA Team Game Stats from 2014 to 2018*. 2016. ⟨https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018⟩. Acesso em: 18 ago. 2019.

18 CORDEIRO, G. M.; DEMéTRIO, C. G. *Modelos Lineares Generalizados e Extensões*. [S.l.: s.n.], 2008.

19 DOBSON, A. J. *An Introduction to Generalized Linear Models, Second Edition*. [S.l.]: Taylor & Francis, 2010. (Chapman & Hall/CRC Texts in Statistical Science). ISBN 9781420057683.

20 TURKMAN, M. A. A.; SILVA, G. L. Modelos lineares generalizados-da teoria à prática. *SPE Edition, Lisbon*, 2000.

21 TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, [Royal Statistical Society, Wiley], v. 58, n. 1, p. 267–288, 1996. ISSN 00359246. Disponível em: ⟨http://www.jstor.org/stable/2346178⟩.

22 FRIEDMAN, J. H.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, v. 33, n. 1, p. 1–22, 2010. ISSN 1548-7660. Disponível em: ⟨https://www.jstatsoft.org/v033/i01⟩.

23 HASTIE, T.; QIAN, J.; TAY, K. *An Introduction to glmnet*. [S.l.], 2021.

24 EMILIANO, P. C. *Fundamentos e aplicações dos critérios de informação: Akaike e Bayesiano*. 2009. Monografia (Pós-Graduação em Estatística e Experimentação Agropecuária), Universidade Federal de Lavras, Minas Gerais, Brasil.

APÊNDICE A – Nome dos times e siglas

Quadro 3: Siglas e seus respectivos times

Conferência Oeste		Conferência Leste	
Sigla	Time	Sigla	Time
DAL	*Dallas Mavericks*	ATL	*Atlanta Hawks*
DEN	*Denver Nuggets*	BOS	*Boston Celtics*
GSW	*Golden State Warriors*	BRK	*Brooklyn Nets*
HOU	*Houston Rockets*	CHI	*Chicago Bulls*
LAC	*Los Angeles Clippers*	CHO	*Charlotte Hornets*
LAL	*Los Angeles Lakers*	CLE	*Cleveland Cavaliers*
MEM	*Memphis Grizzlies*	DET	*Detroit Pistons*
MIN	*Minnesota Timberwolves*	IND	*Indiana Pacers*
NOP	*New Orleans Pelicans*	MIA	*Miami Heat*
OKC	*Oklahoma City Thunder*	MIL	*Milwaukee Bucks*
PHO	*Phoenix Suns*	NYK	*New York Knicks*
POR	*Portland Trail Blazers*	ORL	*Orlando Magic*
SAC	*Sacramento Kings*	PHI	*Philadelphia 76ers*
SAS	*San Antonio Spurs*	TOR	*Toronto Raptors*
UTA	*Utah Jazz*	WAS	*Washington Wizards*

APÊNDICE B – Tabela para melhor visualização dos dados

Tabela 17: Porcentagem de de vitórias e derrotas de cada time

Conferência Oeste			Conferência Leste		
Time	Vitória (W)	Derrota (L)	Time	Vitória (W)	Derrota (L)
DAL	45,4%	54,6%	ATL	53,4%	46,6%
DEN	45,4%	54,6%	BOS	59,8%	40,2%
GSW	80,8%	19,2%	BRK	32,6%	67,4%
HOU	66,2%	33,8%	CHI	48,8%	51,2%
LAC	61,6%	38,4%	CHO	46,6%	53,4%
LAL	30,2%	69,8%	CLE	64,3%	35,7%
MEM	49,4%	50,6%	DET	46,3%	53,7%
MIN	37,5%	62,5%	IND	52,7%	47,3%
NOP	47,9%	52,1%	MIA	51,8%	48,2%
OKC	59,5%	40,5%	MIL	48,8%	51,2%
PHO	32,6%	67,4%	NYK	33,2%	66,8%
POR	56,4%	43,6%	ORL	34,8%	65,2%
SAC	36,9%	63,1%	PHI	32,9%	67,1%
SAS	70,1%	29,9%	TOR	65,5%	34,5%
UTA	54,0%	46,0%	WAS	54,6%	45,4%

APÊNDICE C – Porcentagens de acerto

Tabela 18: Porcentagem de acerto referente à primeira temporada

Partidas	Vencedor Real	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7
HOU X UTA	HOU	75,44%	75,16%	75,10%	75,14%	75,15%	75,06%	74,83%
ATL X CHI	CHI	37,49%	36,94%	37,30%	37,55%	37,42%	37,26%	36,90%
BOS X MIL	BOS	59,86%	59,60%	59,98%	60,03%	59,98%	60,03%	60,11%
BRK X ORL	BRK	66,28%	66,47%	66,47%	66,23%	66,45%	66,47%	66,48%
CHO X TOR	TOR	76,46%	76,58%	76,18%	76,01%	76,14%	76,08%	76,13%
CLE X WAS	CLE	65,95%	65,88%	66,51%	66,47%	66,51%	66,47%	66,33%
DAL X POR	DAL	49,70%	49,83%	50,85%	50,58%	50,73%	50,83%	50,23%
DEN X GSW	GSW	59,03%	59,29%	59,50%	59,57%	59,61%	59,45%	59,37%
DET X NYK	DET	68,98%	69,12%	68,47%	68,31%	68,26%	68,48%	68,80%
IND X MEM	MEM	31,35%	31,35%	31,44%	31,61%	31,60%	31,41%	31,52%
LAC X PHO	LAC	84,47%	84,45%	84,42%	84,22%	84,20%	84,47%	84,41%
LAL X SAC	SAC	66,64%	66,55%	66,28%	66,39%	66,27%	66,32%	66,19%
MIA X PHI	MIA	45,47%	45,34%	45,63%	45,79%	45,82%	45,74%	45,83%
MIN X OKC	OKC	75,79%	75,79%	74,70%	74,89%	74,85%	74,75%	74,94%
NOP X SAS	NOP	22,22%	21,97%	21,99%	22,22%	22,20%	21,82%	21,82%
Médias		59,01%	58,98%	58,99%	59,00%	59,01%	58,98%	58,93%

Tabela 19: Porcentagem de acerto referente à segunda temporada

Partidas	Vencedor Real	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7
HOU X SAC	HOU	56,30%	56,47%	56,30%	56,12%	56,16%	56,24%	56,24%
UTA X LAL	LAL	44,21%	44,09%	44,47%	44,41%	44,48%	44,33%	44,32%
ATL X WAS	WAS	59,11%	59,05%	59,20%	59,28%	59,28%	59,17%	59,15%
CHI X PHI	CHI	52,05%	51,98%	52,33%	52,17%	52,32%	52,27%	52,11%
BOS X MIA	BOS	39,90%	39,67%	39,99%	39,87%	39,94%	39,94%	39,86%
MIL X IND	IND	57,77%	57,79%	57,97%	58,34%	57,96%	58,02%	58,06%
BRK X TOR	TOR	48,32%	48,46%	48,37%	48,39%	48,66%	48,42%	48,41%
ORL X CHO	CHO	48,91%	49,05%	48,50%	48,40%	48,52%	48,44%	48,52%
CLE X DET	DET	36,87%	37,11%	37,48%	37,65%	37,58%	37,51%	37,69%
DAL X SAS*	SAS	44,87%	44,87%	44,23%	44,35%	44,30%	44,28%	44,28%
POR X DEN	POR	67,74%	66,74%	66,71%	66,45%	66,50%	66,63%	66,33%
GSW X MEM	GSW	83,04%	83,11%	83,93%	83,74%	83,86%	83,86%	83,53%
NYK X IND*	IND	61,60%	61,70%	61,96%	61,89%	62,08%	61,89%	61,83%
LAC X PHO	PHO	43,53%	43,56%	43,70%	43,77%	43,78%	43,68%	43,60%
MIN X NOP	MIN	53,60%	53,38%	53,96%	53,84%	53,67%	53,81%	53,77%
OKC X SAS	SAS	21,07%	21,11%	20,46%	20,44%	20,49%	20,41%	20,89%
Médias		51,18%	51,13%	51,22%	51,19%	51,22%	51,18%	51,16%

Tabela 20: Porcentagem de acerto referente à terceira temporada

Partidas	Vencedor Real	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7
HOU X MIN	HOU	70,81%	70,80%	70,54%	70,43%	70,64%	70,49%	70,44%
UTA X SAS	UTA	49,51%	49,38%	49,03%	48,99%	49,05%	49,09%	49,16%
ATL X IND	IND	55,68%	55,69%	55,63%	55,64%	55,60%	55,57%	55,71%
CHI X BRK	CHI	37,04%	37,06%	37,11%	37,08%	37,02%	37,09%	37,07%
BOS X MIL	BOS	71,97%	71,99%	71,99%	71,99%	72,01%	72,02%	71,94%
ORL X DET	ORL	65,51%	65,51%	65,09%	65,05%	65,17%	65,02%	65,25%
CHO X ATL*	ATL	37,03%	36,97%	37,50%	37,46%	37,49%	37,48%	37,13%
TOR X CLE	TOR	29,71%	29,68%	29,43%	29,54%	29,49%	29,54%	29,62%
WAS X MIA	MIA	33,91%	34,02%	34,75%	33,72%	33,76%	33,73%	33,76%
DAL X MEM	DAL	45,78%	45,83%	46,28%	46,33%	46,28%	46,32%	46,09%
POR X NOP	NOP	51,14%	51,12%	51,58%	51,59%	51,52%	51,59%	51,54%
DEN X OKC	DEN	69,68%	69,73%	69,92%	69,94%	69,90%	69,96%	69,88%
GSW X LAL	GSW	60,85%	60,84%	60,81%	60,86%	61,03%	60,95%	60,71%
NYK X PHI	NYK	36,70%	36,68%	36,72%	36,67%	36,73%	36,68%	36,86%
LAC X SAC	LAC	69,56%	69,56%	69,43%	69,27%	69,44%	69,38%	69,21%
PHO X SAC*	SAC	31,25%	31,22%	31,58%	31,54%	31,58%	31,55%	31,74%
Médias		51,01%	51,01%	51,09%	51,01%	51,04%	51,03%	51,01%

Tabela 21: Porcentagem de acerto referente à quarta temporada

Partidas	Vencedor Real	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7
HOU X SAC	SAC	21,40%	21,34%	21,66%	21,62%	21,66%	21,46%	21,90%
UTA X POR	POR	44,71%	44,73%	44,17%	44,14%	44,17%	44,14%	44,38%
ATL X PHI	PHI	82,52%	82,69%	82,87%	82,87%	82,83%	82,99%	82,90%
CHI X DET	DET	54,91%	54,93%	55,13%	55,29%	55,12%	55,17%	55,14%
BOS X BRK	BOS	33,34%	33,36%	33,19%	33,01%	33,07%	33,12%	33,16%
MIL X PHI*	PHI	56,74%	56,83%	57,39%	57,51%	57,57%	57,50%	57,11%
ORL X WAS	ORL	25,60%	25,62%	25,40%	25,26%	25,33%	25,34%	25,38%
CHO X IND	CHO	74,67%	74,73%	75,21%	75,23%	75,28%	75,23%	75,08%
TOR X MIA	MIA	44,77%	44,72%	44,91%	44,73%	44,89%	44,88%	44,75%
CLE X NYK	NYK	24,66%	24,57%	24,75%	24,88%	24,85%	24,77%	24,92%
DAL X PHO	PHO	44,23%	44,28%	44,34%	44,39%	44,33%	44,44%	44,20%
DEN X MIN	MIN	31,94%	31,80%	31,16%	31,14%	31,18%	31,11%	31,48%
GSW X UTA*	UTA	53,70%	53,81%	52,93%	52,87%	52,92%	52,89%	52,74%
MEM X OKC	OKC	78,74%	78,80%	78,54%	78,47%	78,53%	78,51%	78,46%
LAC X LAL	LAL	41,68%	41,75%	41,64%	41,99%	41,84%	41,84%	41,95%
NOP X SAS	NOP	60,68%	60,66%	61,36%	61,30%	61,36%	61,31%	61,12%
Médias		48,39%	48,41%	48,42%	48,42%	48,43%	48,42%	48,42%