Universidade de Brasília
Departamento de Estatística

Modelagem Probabilística em Apostas Esportivas

José Vítor Barreto Porfírio

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

José Vítor Barreto Porfírio

Modelagem Probabilística em Apostas Esportivas

Orientador(a): Prof. Guilherme Souza Rodrigues

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Agradecimentos

Aos meu pais, Cleide e José Paulo, que me criaram com muito amor, apoio e as condições para chegar onde estou hoje, sem medir esforços para garantir minha saúde e bem estar.

À minha irmã mais velha, Carol, que sempre trouxe os melhores diálogos na mesa de jantar, dos mais engraçados aos mais profundos e filosóficos, sempre agregando muito à minha vida.

À minha namorada, Maria Emília, por sempre me dar amor, carinho e me motivar a estudar em direção aos meus objetivos.

Aos amigos que fiz ao longo da graduação e proporcionaram bons momentos, dentro e fora da universidade. Em especial, Stefan, por se mostrar um amigo presente em momentos importantes e sempre me dar carona de volta pra casa.

Ao meu orientador, Guilherme, que não só viabilizou esse TCC, mas também por ser um professor que extrai o melhor de seus alunos e apresentar um comportamento exemplar em cuidar de seus alunos.

"Sorte é o que acontece quando a preparação encontra a oportunidade"
(Sêneca)

Resumo

A modelagem probabilística de partidas de futebol pode impactar diversos públicos, entre eles estão os apostadores. As casas de apostas fornecem os odds para diversos cenários das partidas, esses odds podem ser analisados como um conjunto de equações que descreve uma superfície de probabilidade do jogo, ao serem transformados para probabilidade e normalizados pelo método de Shin. Um modelo paramétrico, apelidado de Distribuição Placar Ajustada, foi ajustado às equações por meio de métodos iterativos e da divergência de Kullback-Leibler. Esse modelo se mostrou decentemente efetivo, ao permitir a reconstrução das partições da superfície de probabilidade. Em média as casas de apostas apresentam probabilidades calibradas, embora ainda haja uma margem para que os apostadores possam se beneficiar de regiões não calibradas em quase todos os mercados.

Palavras-chaves: Apostas. Calibração. Distribuição Placar Ajustada. Divergência de Kullback-Leibler. Superfície de Probabilidade. Odds.

Abstract

Probabilistic modeling of football matches may impact several people, specifically the bettors. Bookmakers provide odds for a significant number of match outcomes. These odds can be examined as a set of equations that describe the probability surface of the match, as they are transformed into probabilities and normalized by Shin's method. A parametric model, nicknamed Adjusted Score Distribution, was fitted to the set of equations through iterative means and Kullback-Leibler Divergence. This model proved to be quite effective, as it was able to reconstruct the partitions of the probability surface. On average, bookmakers do share calibrated probabilities, although there is still a margin for bettors to take advantage of uncalibrated regions in almost every market.

Keywords: Adjusted Score Distribution. Bet. Calibration. Kullback-Leibler Divergence. Probability Surface. Odds.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A predição de resultados em eventos esportivos foi tópico de pesquisa em diversos esportes e foi exercida por meio de diversas técnicas. Caso essas predições para os resultados do evento sejam robustas, então essa é uma informação crucial para os atletas e técnicos, uma vez que pode alterar suas estratégias. Ademais, vale destacar também a importância para os apostadores, que tem expectativas de faturar dinheiro ao fornecer bons palpites nos resultados dos eventos (LU et al., 2021).

Sob essa ideia, estudos passados sobre futebol focaram mais em dados sobre o time, como as estatísticas de gols, de faltas e de cartões. Ao dispor desse tipo de estatísticas, alguns estudos utilizaram técnicas de modelagem bayesiana (RAZALI et al., 2017; RAHMAN et al., 2018), de aprendizado não supervisionado (GAUB, 2022) e de machine learning (SAMBA, 2019).

Outra técnica no ramo estatístico também já explorada, no contexto de análise probabilística em partidas de futebol, é a modelagem via séries temporais. Ao adicionar ao modelo variáveis explicativas sobre o time, torcida e até mesmo o local onde o jogo ocorreu, para melhorar seu desempenho preditivo, pesquisadores conseguiram adequar modelos de séries temporais a dados de partidas de futebol e realizar predições (YIANNAKIS et al., 2006; JOSEPH, 2022).

Não obstante, existem fatores que não podem ser transmitidos com clareza por estatísticas dos times, estatísticas individuais dos jogadores, a condição do time jogar em casa, etc. Um exemplo desses fatores é a situação em que um time já garantiu sua classificação para a próxima fase do campeonato, permitindo a execução de novas estratégias ao mudar a escolha dos jogadores que começam em campo.

Contudo, existe um grupo de indivíduos que detém conhecimento notório sobre o quadro de um ou mais times e das partidas que estão por vir, esses indivíduos são torcedores, fãs que acompanham o esporte, apostadores e precificadores de casas de apostas. Em outras palavras, pode ser interessante tentar avaliar o conhecimento desse grupo acerca da partida para modelagens sobre os resultados dos jogos.

A métrica que as casas de aposta fornecem e expressa o posicionamento dos apostadores são os *odds*, que indicam quanto um apostador irá ganhar caso dê um palpite correto em sua aposta, por exemplo se os *odds* são de 1,24 para vitória do time da casa e o apostador coloca 100 reais "na linha" a favor do time da casa, caso seu palpite seja correto, ele receberá 124 reais.

Os apostadores são um grupo que se expandiu junto à popularidade do setor de apostas, setor esse que cresceu significativamente em 2022 nos Estados Unidos (American Gaming Association, 2023). Entretanto, nem sempre as apostas esportivas foram permitidas nos Estados Unidos, e nem no Brasil. Em maio de 2018 a Suprema Corte Americana considerou o Ato de Proteção do Esporte Profissional e Amador (PASPA) como inconstitucional, que proíbia os estados americanos de legalizarem apostas nos esportes (DAN, 2023), desde então o setor de apostas esportivas vem crescendo e em 2022 apresentou uma receita anual de mais de 7 bilhões de dólares para as casas de aposta (American Gaming Association, 2023).

Em acordo à lei que proibe os jogos de azar no Brasil (Brasil, 1946), apostas esportivas não seriam uma prática legalizada, porém desde a publicação da lei que alterou algumas práticas sobre apostas (Brasil, 2018) foi permitida a operação de casas de aposta no Brasil, contanto que seus sites estejam hospedados em domínios internacionais. Desde então também houve crescimento do setor de apostas esportivas no Brasil, como indicado pelo mapa de patrocínio dos times de futebol e a crescente inserção dos nomes e endereços eletrônicos de casas de apostas nos uniformes dos jogadores (iGaming Brazil, 2023).

Haja vista o crescente mercado de apostas esportivas citado, é nítido que existem diversos indivíduos apostando como um passatempo ou até mesmo como fonte de renda, compartilhando entre eles o desejo de ganhar. Utilizando-se as definições probabilísticas dos *odds* e os valores fornecidos pelas casas de aposta de acordo com a colocação das apostas, é viável estimar a probabilidade dos eventos possíveis em uma partida, porque os *odds* são preditores acurados sobre a margem de vitória (KAIN; LOGAN, 2014), já se mostraram superiores a previsão de especialistas ou de modelos estatísticos em estudos passados (SONG; BOULIER; STEKLER, 2007) e já foram apontados como uma variável importante a ser considerada em estudos futuros para melhoria de modelos já propostos na literatura (JOSEPH, 2022).

Conforme as apostas em uma partida são feitas por meio das casas de aposta, os *odds* serão alterados refletindo a inclinação dos apostadores, portanto, este estudo visa utilizar os *odds* das casas de aposta como o ponto de partida para determinação probabilística dos possíveis resultados para uma dada partida, aliado à proposição de um novo modelo paramétrico que descreve toda a superfície de probabilidade dos placares de uma partida de futebol.

Obter estimativas de probabilidade para cada resultado possível de uma partida de futebol permite encontrar oportunidades de apostas que paguem mais que a quantidade adequada e induzir cenários de *arbitrage*, favorecendo o apostador. Esses pontos são

os pontos mal calibrados em termos de probabilidade pelas casas de apostas e foram apontados na seção 6.

Na seção 2 é fornecido um contexto geral sobre apostas esportivas, o funcionamento dos mercados de apostas e dos *odds*. Na seção 3 foram levantadas as técnicas passadas sobre a modelagem probabilística dos placares partidas de futebol. Na seção 4 estão dispostas as técnicas estatísticas e matemáticas pré-existentes necessárias para realizar esse estudo. Na seção 5 está apresentada a distribuição Placar Ajustada e o método para seu ajuste, tal que essa distribuição modela as superfícies de probabilidade. Na seção 6 estão apresentadas algumas comparações sobre os ajustes obtidos e as distribuições cuja origem é das casas de apostas.

2 Contexto de Apostas Esportivas

2.1 *Odds*

O termo *odd* é um neologismo vindo da língua inglesa e no contexto de probabilidade se traduz como "chance", expressando quantas vezes um resultado é mais provável que o seu complementar. De acordo com Giolo (2017) tem-se a Definição 1

Definição 1 *A chance de ocorrência de um evento "A" de interesse (ou chance de sucesso) é dada por*

$$\mathcal{O}_A = \frac{P(A)}{1 - P(A)} = \frac{\textit{Probabilidade do evento A ocorrrer}}{\textit{Probabilidade do evento A não ocorrer}}$$

Para as casas de apostas, a relação entre odds e probabilidade é inversa à Definição 1, esse ponto é muito importante para que a casa de apostas proteja seu lucro ao pagar quantias menores para os resultados mais prováveis, além disso os odds das casas de apostas são acrescidos de 1 para garantir que o apostador premiado receba ao menos a quantia apostada.

No contexto de partidas de futebol, o time que tem a maior probabilidade de ganhar apresentará os menores *odds* para vitória, de forma que o evento de interesse para calcular esse *odd* é o tal time não ganhar. Isto é, seja $Casa$ o evento que descreve a vitória do time da casa, ao adaptar a Definição 1, tem-se que o *odd* a favor do time da casa é dado por

$$\begin{aligned}\mathcal{O}_{Casa} &= 1 + \frac{1 - P(Casa)}{P(Casa)} \\ &= \frac{P(Casa) + 1 - P(Casa)}{P(Casa)} \\ &= \frac{1}{P(Casa)}\end{aligned}$$

em que $P(Casa)$ é a probabilidade do time da casa ganhar a partida. Logo, os odds das casas de apostas são dados pela Definição 2.

Definição 2 *A chance (odd) para a ocorrência de um evento "A" de interesse, fornecida pelas casas de aposta, é dada por*

$$\mathcal{O}_A = \frac{1}{P(A)}$$

2.2 *Arbitrage*

Arbitrage trata-se de uma estratégia utilizada em apostas esportivas que consiste em apostar nos casos em que há certeza de ganho, devido à variabilidade entre as probabilidades fornecidas pelas diversas casas de aposta, aproveitada ao apostar em resultados diferentes em mais de uma casa de aposta.

Exemplo 1 *Suponha que um apostador deseja fazer apostas em partidas de futebol, onde há as possibilidades de vitória do time da casa, vitória do time visitante e empate. Em geral, segue-se que a soma das probabilidades desses 3 eventos citados deve totalizar 100%, uma vez que descrevem todos os cenários possíveis, caso o jogo ocorra normalmente.*

Suponha também que uma casa de aposta forneça os seguintes odds

$$\mathcal{O}_{Casa} = 3,1; \quad \mathcal{O}_{Empate} = 2,5; \quad \mathcal{O}_{Visitante} = 3,25$$

em que $\mathcal{O}_{Casa}$, $\mathcal{O}_{Visitante}$ e $\mathcal{O}_{Empate}$ são os odds para a vitória do time da casa, vitória do time visitante e empate, respectivamente.

Sabe-se que $P(Casa) + P(Empate) + P(Visitante) = 1$, não obstante, ao tomar a Definição 2, obtém-se

$$\begin{aligned}\frac{1}{\mathcal{O}_{Casa}} + \frac{1}{\mathcal{O}_{Empate}} + \frac{1}{\mathcal{O}_{Visitante}} &= \frac{1}{\frac{1}{P(Casa)}} + \frac{1}{\frac{1}{P(Empate)}} + \frac{1}{\frac{1}{P(Visitante)}} \\ &= P(Casa) + P(Empate) + P(Visitante) = 1\end{aligned}$$

Porém, ao substituir os odds fornecidos pela casa de apostas deste exemplo, observa-se

$$P(Casa) + P(Empate) + P(Visitante) = \frac{1}{3,1} + \frac{1}{2,5} + \frac{1}{3,25} \approx 1,03 \neq 1 \quad (2.2.1)$$

indicando que a casa de aposta tem uma margem de aproximadamente 3% em cima dos resultados das apostas.

Entretanto, ao consultar outra casa de aposta, encontrou-se

$$\mathcal{O}_{Casa} = 3,7; \quad \mathcal{O}_{Empate} = 2,3; \quad \mathcal{O}_{Visitante} = 3,05$$

$$P(Casa) + P(Empate) + P(Visitante) = \frac{1}{3,7} + \frac{1}{2,3} + \frac{1}{3,05} \approx 1,03 \neq 1$$

ou seja, essa segunda casa também tem uma margem de 3% sobre os odds, entretanto o odd para a vitória do time da casa é maior que na primeira casa. Ao substituir o maior valor disponível de $\mathcal{O}_{Casa}$ *em (2.2.1), obtém-se*

$$P(Casa) + P(Empate) + P(Visitante) = \frac{1}{3,7} + \frac{1}{2,5} + \frac{1}{3,25} \approx 0,978 \neq 1,$$

indicando que o apostador tem uma margem por volta de 2% em cima dos resultados das apostas e está diante de uma oportunidade de arbitrage.

Sob esse cenário, caso o apostador disponha de R$100,00 para fazer suas apostas, é possível calcular valores estratégicos utilizando pesos baseados nas probabilidades, isto é,

$$\frac{P(Casa)}{P(Casa) + P(Empate) + P(Visitante)} \cdot 100 \approx \frac{0,27}{0,978} \cdot 100 \approx 27,64$$

e, de forma análoga para as demais possibilidades, apostar como na Tabela **??**,

Caso ocorra a vitória do time da casa, então o apostador irá dispor de R$102,25 na casa de apostas 2 e R$0,00 na casa de apostas 1, caso não ocorra vitória do time da casa, então o apostador irá dispor de R$102,25 na casa de apostas 1 e R$0,00 na casa de apostas 2. Em outras palavras, o apostador sairá de um montante de R$100,00 para um montante de R$102,25, totalizando um lucro de R$2,25, independentemente do resultado da partida.

2.3 Mercados de Apostas

2.3.1 Listagem

Em termos de apostas esportivas existem diversos mercados para se fazer apostas sobre a partida. Devido ao escopo deste estudo, os mercados selecionados são apenas aqueles que contribuem diretamente para a reconstrução da superfície de probabilidade dos resultados possíveis de cada partida, isto é:

1. *Head to Head* (h2h): odds para o resultado da partida (vitória do time da casa,

vitória do time visitante ou empate);

2. *Spread*: odds para o resultado da partida com desvantagem para um dos times, "spread +x/-x" significa o cenário fictício em que o time da casa já começa a partida com "x" gols de vantagem, então é possível apostar na vitória do time da casa, empate e vitória do time visitante, sempre sob esse cenário fictício;

3. *Over/Under*: odds para a soma dos gols do time da casa e visitante, em que aposta-se em uma soma maior ou menor que a fixada pela casa de aposta;

4. *Correct Score*: odds para o resultado exato da partida, aposta-se no placar da partida como "1 a 1", "2 a 1", etc.

5. *Both Score*: odds para ambos times marcarem pelo menos um gol na partida ou pelo menos um dos times não marcar nenhum gol na partida.

2.3.2 Submercados e Cenários

Alguns dos mercados apresentados se relacionam diretamente com os odds para um dado cenário da partida, isto é, sob o mercado h2h, as casas de apostas apresentam odds para o cenário de vitória do time casa. Esse estudo apelidou estes casos de cenários dos mercados de aposta.

Alguns dos mercados apresentados só se relacionam com os odds para um dado cenário da partida por meio de alguma espécie de "submercado", por exemplo o *spread* primeiro se relaciona com a vantagem ou desvantagem para um dos times, em seguida as casas de apostas apresentam odds para o cenário de vitória do time casa. Esse estudo apelidou estes casos de submercados, então os submercados são as somas de gols do *Over/Under* e as vantagens e desvantagens concedidas pelo *Spread*.

2.3.3 Propriedades Probabilísticas

Como já explicado na seção 2.2, tem-se que os eventos descritos pelo mercado *Head to Head* descrevem todos os resultados possíveis em uma partida de futebol (vitória do time da casa, empate e vitória do time visitante), ou seja, se os odds desse mercado tem uma correspondência probabilística, então essas probabilidades deveriam somar exatamente 100%, mas para isso devem ser normalizadas, devido às casas de apostas terem uma margem de lucro sobre os odds.

Convenientemente todos os mercados compartilham dessa propriedade de que suas probabilidades correspondentes aos odds devem somar 100%, incluindo o *Over/Under*, porque os placares fixados são sempre entre dois número inteiros, como 2.5, 3.5, etc. Em outras palavras, ao estudar um mercado por vez, os odds do mercado em estudo podem ser analisados como uma distribuição de probabilidade discreta.

2.4 Normalização

Estudos passados apontam que as estimativas de probabilidade fornecidas pelas casas de apostas por meio dos odds não satisfazem que a soma das probabilidades de todos os eventos possíveis seja igual a 1, pois as casas de aposta incluem suas margens de lucro em cima dos próprios odds (KONING; ZIJM, 2023). Portanto, foram propostas técnicas de normalização para que essa soma seja 1, a técnica adotada por este estudo é o método de Shin que está explicado em detalhes na seção 4.1.

3 Revisão de literatura

Embora alguns estudos já empregaram os odds como parte de seus modelos para predição dos resultados de partidas de futebol (ODACHOWSKI; GREKOW, 2013; SAMBA, 2019), em geral se limitam ao uso de poucos mercados de apostas como apenas o *Head to Head*, uma vez que esse mercado de apostas descreve as apostas para vitória do time da casa, empate ou vitória do time visitante.

Se as probabilidades dos possíveis resultados de uma partida são conhecidas e entram em desacordo com as casas de aposta, é viável montar um conjunto de apostas com altas probabilidades de ganho, um exemplo próximo a esse cenário é uma estratégia conhecida como *arbitrage*. A estratégia consiste na alocação de recursos em apostas considerando os melhores *odds* em diversas casas de aposta, de forma que não há chance do apostador perder dinheiro, pois a diferença entre as probabilidades estimadas para um mesmo resultado compensa as perdas.

Os principais problemas de se limitar à *arbitrage* são a baixa margem de lucro obtida pelo apostador, o usuário depende das casas de aposta oferecerem odds que criem a oportunidade de *arbitrage* e, por ser uma estratégia amplamente conhecida, as casas de aposta e seus times de analistas de dados já desenvolverem métodos para identificação e potencial suspensão dos usuários que fazem essa prática, embora o usuário não esteja praticando nada que esteja previsto como passível de punição (KAUNITZ; ZHONG; KREINER, 2017).

As combinações de todos os placares possíveis para o time da casa e o time visitante compõem uma superfície de probabilidade discutida a fundo na seção 5.2. Estudos passados propuseram métodos para estimar essa superfície de probabilidade, Maher (1982) analisa o placar de partidas de futebol considerando distribuições de Poisson independentes para os gols do time da casa e do time visitante, em extensão, o estudo de Dixon e Coles (1997) propõe a modelagem dos placares de partidas de futebol por meio de distribuições de Poisson mais sofisticadas, com a inserção de dependência de forma indireta.

O estudo de Karlis e Ntzoufras (2003) endereçou os problemas computacionais e de interpretação ao considerar modelos que incluam correlação entre o número de gols do time da casa e do time visitante. O estudo investigou o ajuste da família de distribuições de Poisson bivariadas e concluiu que houve melhora no ajuste ao introduzir o uso da correlação e a aplicação de inflação das probabilidades de empate.

O estudo de Mchale e Scarf (2006) aponta que é comum a presença de correlação negativa entre o número de gols do time da casa e do time visitante em ligas de futebol internacional, entretanto, é mostrado que a distribuição de Poisson bivariada não pode cumprir esse propósito.

Mchale e Scarf (2006) propõem a utilização de cópulas com distribuições marginais Poisson e Binomial Negativa, uma vez que as cópulas tem a flexibilidade necessária para lidar com a correlação negativa e a Binomial Negativa é uma boa alternativa para lidar com problemas de sob e sobredisperção, que geram incompatibilidade dos dados com a distribuição de Poisson.

Em termos de apostas esportivas, o estudo de Dixon e Coles (1997) apresenta a possibilidade de obter um modelo que forneça estimativas mais acuradas que as casas de aposta, entretanto, o estudo analisou apenas a English Premier, o que pode comprometer a generalização desse resultado para mais ligas de futebol.

Em contrapartida, o estudo de Kaunitz, Zhong e Kreiner (2017) propõe uma abordagem muito mais simples sobre o assunto, em que os odds das casas de aposta são utilizados para compor uma estratégia de apostas que analiza o quão justas são as apostas e a expectativa de retorno, sem a elaboração de um modelo sofisticado, mas utilizando as médias das probabilidades obtidas a partir dos odds.

Ou seja, existe um interesse em conhecer não só a predição dos casos de vitória do time da casa, empate ou vitória do time visitante, mas detalhadamente as probabilidades desses eventos e na verdade para todos os placares possíveis de uma partida, uma vez que esse conhecimento pode guiar as apostas e abre mais oportunidades para apostar em múltiplos mercados de aposta (KAUNITZ; ZHONG; KREINER, 2017).

4 Referencial Teórico

4.1 Normalização

O estudo de Koning e Zijm (2023) aponta dois métodos de normalização: o método multiplicativo, que consiste em dividir as probabilidades encontradas por sua soma, e o método de Shin proposto por Shin (1992) e simplificado por Clarke, Kovalchik e Ingram (2017) para uma forma equivalente apresentado pelas equações (4.1.1) e (4.1.2).

O método de Shin consiste em assumir uma proporção de *insiders*, indivíduos que sempre fazem a aposta correta, e aplicar a normalização considerando essa proporção. Usa-se a equação (4.1.1) para ajustar a proporção de *insiders* z de forma numérica, tal que z satisfaça a equação (4.1.1), em seguida substitui-se o valor de z encontrado na equação (4.1.2) para obter as probabilidades normalizadas.

$$\frac{\sum_{j=1}^{n} \sqrt{z^2 + 4(1-z)\frac{\pi_j^2}{\pi}} - 2}{n-2} - z = 0 \tag{4.1.1}$$

$$p_i = \frac{\sqrt{z^2 + 4(1-z)\frac{\pi_i^2}{\pi}} - z}{2(1-z)} \tag{4.1.2}$$

em que π_i é o inverso dos odds, para $i = 1, 2, ..., n$ e $\pi = \sum_{i=1}^{n} \pi_i$. O i-ésimo evento é um dos cenários descritos pelo mercado em análise, por exemplo, $i = 1, 2$ para o mercado *Over/Under 2.5* e os eventos são soma de gols acima ou abaixo de 2.5, de forma que p_i é a probabilidade normalizada para $i = 1, 2$. O processo é análogo para os demais mercados, cada um dos mercados tem um respectivo z ajustado.

A vantagem da Normalização Shin se da por conta do *Favourite-longshot bias*, um fenômeno no mundo das apostas esportivas em que um dos times tem menos chance de ganhar e recebe mais apostas do que condiz com sua real probabilidade de ganhar, analogamente, o time favorito recebe menos apostas do que condiz com sua real probabilidade de ganhar, esse comportamento leva a casa de apostas a modificar a oferta dos odds (KONING; ZIJM, 2023).

4.2 Distribuição Normal

Uma variável aleatória X cuja distribuição é normal de média μ e variância σ^2 é denotada por

$$X \sim N(\mu, \sigma^2)$$
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

de forma que se

$$X \sim N(0, 1)$$

então X é dita uma variável aleatória cuja distribuição é normal padrão. Especificamente no caso da distribuição normal padrão denota-se suas funções de distribuição acumulada e função quantil por Φ e Φ^{-1} respectivamente. Nem Φ, nem Φ^{-1} possuem forma fechada, portanto só podem ser calculadas por métodos numéricos.

4.3 Normal Bivariada

Se o vetor $\boldsymbol{X} = (X, Y)$ tem distribuição conjunta normal bivariada, denota-se

$$(X, Y) \sim \boldsymbol{\mathcal{N}(\mu, \Sigma)}$$

em que $\boldsymbol{\mu}$ é o vetor de médias e $\boldsymbol{\Sigma}$ é a matriz de covariâncias e sua densidade é dada por

$$f_{\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = (2\pi)^{-k/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right),$$

Especificamente no caso bivariado, a sua distribuição acumulada pode ser implementada pelo método de Drezner e Wesolowsky (1990) que propõe a utilização de quadraturas gaussianas com 5 pontos. Agca e Chance (2003) aponta que esse foi um dos métodos computacionalmente mais rápidos para calcular a distribuição acumulada de uma normal bivariada. Ademais, esse método dispõe da implementação de Robitzsch (2020) em C++ para ser exceutado em R.

4.4 Binomial Negativa

Se X é uma variável aleatória, que modela o número de fracassos até o r-ésimo sucesso em uma sequência de Ensaios de Bernoulli, denota-se

$$X \sim Binomial\ Negativa(r, p)$$

em que p é a probabilidade de ocorrer um sucesso em um Ensaio de Bernoulli. Sua função de probabilidade é dada por

$$P(X = x) = \begin{cases} p^r(1-p)^x \frac{\Gamma(x+r)}{\Gamma(r)x!} & , \text{ se } x = 0, 1, 2, ... \\ 0 & , \text{ caso contrário} \end{cases}$$

$$\mu = E(X) = \frac{r(1-p)}{p}; \quad \sigma^2 = Var(X) = \frac{r(1-p)}{p^2}$$

4.5 Poisson

A distribuição de Poisson é uma distribuição discreta de probabilidade, em que se X é uma variável aleatória que segue distribuição de Poisson, então

$$X \sim Poisson(\alpha)$$

$$P(X = x) = \begin{cases} \frac{e^{-\alpha}\alpha^x}{x!} & , \text{ se } x = 1, 2, ... \\ 0 & , \text{ caso contrário} \end{cases}$$

4.6 Mistura de Distribuições de Probabilidade

Se X e Y são variáveis aleatórias cujas funções de probabilidade ou densidade são dadas respectivamente por f_1 e f_2, então pode-se definir uma variável aleatória Z a partir da mistura de X e Y, cuja função de probabilidade ou densidade é dada por

$$f_Z(z) = \epsilon f_X(z) + (1 - \epsilon) f_Y(z)$$

em que $0 < \epsilon < 1$.

4.7 Teste de Hipóteses

Estudos passados apontam que existe uma correlação entre o número de gols do time da casa e do time visitante, embora outros estudos tenham desconsiderado essa correlação (DIXON; COLES, 1997; KARLIS; NTZOUFRAS, 2003). Entende-se que o teste de hipóteses para a correlação pode ajudar a sanar questões sobre a correlação entre o número de gols do time da casa e do time visitante.

Seja ρ o parâmetro de correlação, então tem-se as hipotéses

$$H_0)\ \rho = 0;\ H_1)\ \rho \neq 0$$

A estatística do teste é a estatística t, tal que

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n^2-2} \sim T_{n-2}$$

em que r é a correlação estimada, n é o número de observações e a estatística do teste segue distribuição t-student com n-2 graus de liberdade.

A correlação r entre duas variáveis X e Y pode ser calculada por

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

em que σ_X e σ são os respectivos desvios-padrão de X e Y. $Cov(X,Y)$ é a covariância entre X e Y e é dada por

$$Cov(X,Y) = \sum_x \sum_y f(x,y)(x-\mu_X)(y-\mu_Y)$$

em que $f(x,y)$ é a frequência relativa do par (x,y), μ_X e μ_Y são respectivamente as médias de X e Y.

Um nível de significância α deve ser estabelecido, então a estatística do teste é calculada e obtém-se o p-valor

$$\text{p-valor} = P(T_{n-2} \geq |t|)$$

caso o p-valor seja inferior ao nível de significância α, rejeita-se a hipótese nula, do contrário não rejeita-se a hipótese nula.

4.8 Cópula Gaussiana

Cópulas, em teoria de probabilidade, são distribuições que surgem com o objetivo de gerar um conjunto de observações correlacionadas. Pela evidência de estudos passados sobre a correlação entre o número de gols do time da casa e do time visitante, espera-se um ganho de flexibilidade ao incluir a correlação na representação das probabilidades para esses eventos, como proposto por Mchale e Scarf (2006).

Arbitrariamente, escolheu-se partir da Cópula Gaussiana, obtida após gerar uma amostra aleatória (x, y) da distribuição normal bivariada com correlação não nula entre as variáveis, médias nulas e variâncias unitárias. Sobre essa amostra, aplica-se a função de distribuição acumulada da distribuição normal padrão, isto é, obtém-se a amostra

$$(u, v) = (\Phi(x), \Phi(y))$$

4.9 Divergência de Kullback-Leibler (D_{KL})

A Divergência de Kullback-Leiber é uma medida de entropia que quantifica uma divergência não simétrica entre duas distribuições de probabilidade P e Q. No caso discreto, se a distribuição P tem suporte $\mathcal{X}$, então

$$D_{KL}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \geq 0$$

Nos pontos do suporte $\mathcal{X}$ em que $P(x)$ assume valores maiores, então a contribuição para a divergência também será maior, ou seja, as maiores reduções da D_{KL} são nos casos em que a distribuição Q consegue se aproximar mais da distribuição P nos pontos em que $P(x)$ é mais acentuada.

Uma vez que a métrica D_{KL} é sempre não negativa, então o procedimento de ajustar os parâmetros da distribuição Q com objetivo de minimizar essa divergência é um possível método para tornar as distribuições P e Q mais similares com respeito às suas funções de probabilidade.

Este estudo adotou a notação simplificada D_{KL}, em que a distribuição P sempre se refere à distribuição de probabilidade provinda das casas de apostas, já a distribuição Q sempre se refere à distribuição proposta para modelar a distribuição P.

4.10 Calibração

No contexto de modelos de classificação, a calibração de um modelo está relacionada com capacidade do modelo de predizer uma certa probabilidade para um resultado e aquele resultado realmente ocorrer com aquela proporção.

Isto é, suponha um problema de classificação binária com resultado 0 ou 1, então ajusta-se um modelo preditivo que lança predições entre 0 e 1, para cada observação toma-se o valor predito $\hat{p}$ entre 0 e 1, obtém-se também a predição binarizada $\hat{y}$ igual a 0 ou 1, caso o valor real observado y seja igual ao predito $\hat{y}$, então conta-se que houve 1 caso de acerto da predição $\hat{p}$ arredondado para 2 casas demais para simplificações, caso $y \neq \hat{y}$ então conta-se que houve 1 caso de erro da predição $\hat{p}$.

Sob esse procedimento, há uma contagem de acertos e erros de todas as predições possíveis $\hat{p}$, ou seja, dividindo a contagem de acertos pela soma da quantidade de acertos e da quantidade de erros, espera-se que esse cociente seja igual a $\hat{p}$, para um modelo bem calibrado.

4.11 Regressão Isotônica

Regressão isotônica é uma técnica estatística que se ajusta as observações por meio do ponto mais próximo a cada grupo, sendo que a linha de regressão ajustada deve ser não descrente ou não crescente em toda a sua extensão e tem forma livre, podendo se comportar como degraus de uma escada, tal como está apresentado no exemplo com dados fictícios da Figura 1.

A regressão isotônica pode ser ajustada pelo método dos mínimos quadrados, com a restrição de ser sempre não decrescente (ou não crescente), isto é, considere um conjunto de dados que apresenta pares (x_i, y_i) para $i \in \{1, 2, ..., n\}$, então escolhe-se uma função não descrente (ou não crescente) f, em que $f(x_i) = \hat{y}_i$ é a predição da regressão e satisfaz a equação

$$\min \sum_{i=1}^{n} (\hat{y}_i - y_i)^2, \text{ tal que } \hat{y}_i \leq \hat{y}_j \text{ se } x_i \leq x_j$$

Uma das utilidades da regressão isotônica é conseguir fazer uma melhor calibração de um modelo pré-existente, fazendo com que probabilidades obtidas pelo modelo original se tornem mais parecidas com as probabilidades empíricas apresentadas pelos dados, isto

é, quando se está lidando com um problema de classificação (SANGANI, 2022).

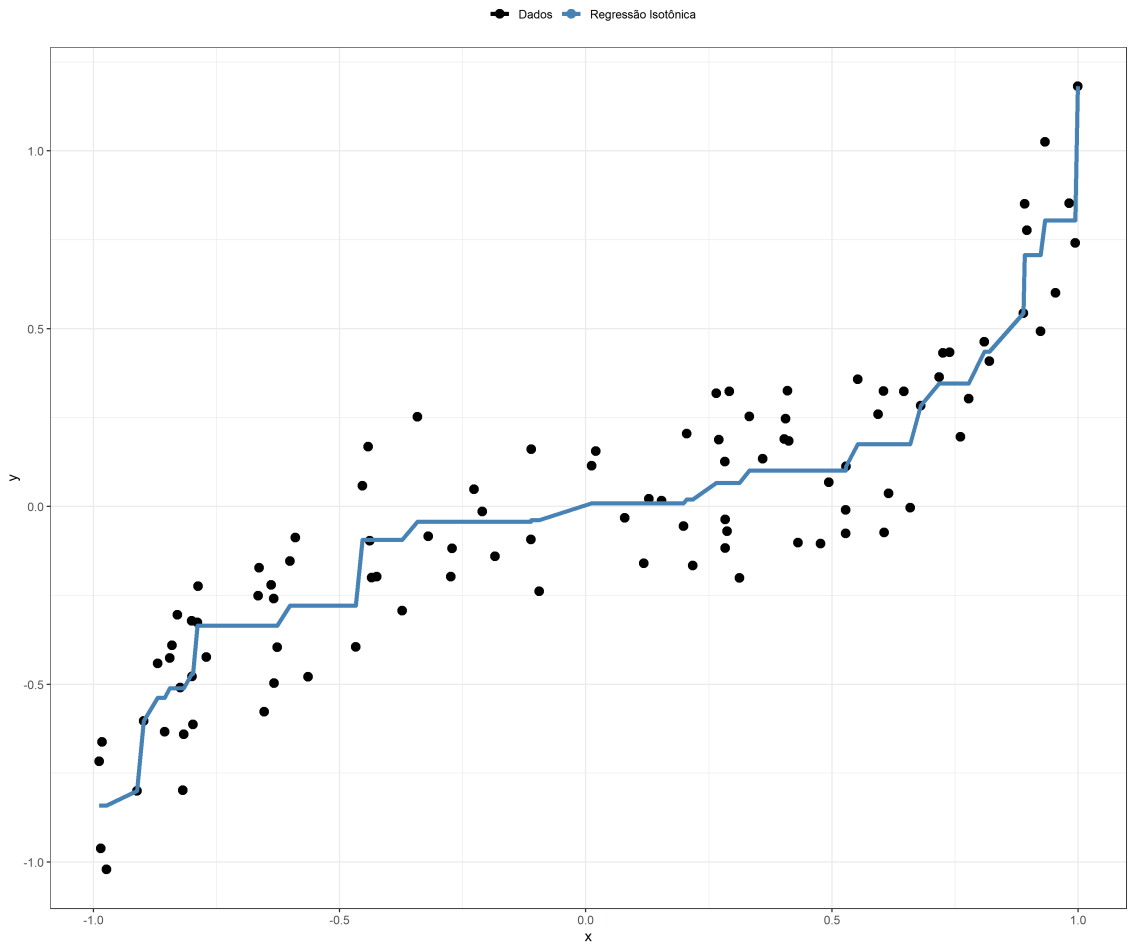

Figura 1: Ajuste de uma regressão isotônica à dados fictícios para um polinômio de terceiro grau.

4.12 Função vec

A função vec é uma função que permite transformar uma matriz $\mathbb{M}$ qualquer em um vetor $\mathbb{V}$ de forma que

$$\mathbb{M} = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1m} \\ M_{21} & M_{22} & \cdots & M_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{nm} \end{bmatrix}$$

$$\text{vec}(\mathbb{M})^T = \mathbb{V}^T = [M_{11}, M_{12}, ..., M_{1m}, M_{21}, ..., M_{nm}]$$

ou seja, obtém-se uma representação vetorial para os elementos de uma matriz. A função vec tem um uso particular na seção 5.3.4.

4.13 Matriz Aumentada

Caso hajam n vetores $\mathbb{V}_1, \mathbb{V}_2, ..., \mathbb{V}_n$ contendo a mesma quantidade m de elementos em cada um, então é possível compor uma matriz aumentada $\mathbb{M}$ com n linhas e m colunas na forma

$$\mathbb{M} = \begin{bmatrix} \mathbb{V}_1 \\ \mathbb{V}_2 \\ \vdots \\ \mathbb{V}_n \end{bmatrix}$$

em que a i-ésima linha da matriz contém os valores do vetor $\mathbb{V}_i$ em suas colunas, na mesma ordem apresentada pelo vetor $\mathbb{V}_i$. Tipicamente esse processo de criar uma matriz aumentada é feito concatenando as colunas e não as linhas da matriz.

5 Metodologia

5.1 Conjunto de dados

O conjunto de dados foi organizado em 2 arquivos json, um arquivo organiza as informações sobre as partidas por liga e o outro organiza os odds. Tanto as informações sobre as partidas como os odds apresentam os dados disponíveis para os anos 2019, 2020, 2021 e 2022 e foram coletados por meio de *web scrapping* da plataforma https://oddspedia.com/br.

5.1.1 Coleta

Para a coleta, a ferramenta utilizada foi o Selenium, uma biblioteca para linguagens de programação que emula um navegador e pode ser controlada programaticamente.

Para utilizar o Selenium, foi elaborado um script em Python que faz requisições à API do site https://oddspedia.com/br. Para tornar esse procedimento mais claro, os próximos parágrafos dessa subseção descrevem e exemplificam como acessar a API e os parâmetros mais importantes para determinar as requisições.

No caso das informações sobre os jogos de cada liga, as requisições foram feitas configurando os parâmetros da requisição para o caminho "getMatchList" e os parâmetros da busca configurados para abrangir as ligas pelo parâmetro "league" e por datas no formato "YYYY-MM-DD T hh:mm:ss Z" sem os espaços. A url a seguir exemplifica uma requisição das partidas da Série A do Brasileirão no período de 30 de outubro de 2023 a 6 de novembro de 2023. Ademais o parâmetro "excludeSpecialStatus" foi sempre escolhido como 1, para que partidas que foram adiadas, canceladas, ou outra situação incomum não fossem retornadas.

"https://oddspedia.com/api/v1/getMatchList?geoCode=BR&sport=futebol

&category=brasil&league=brasileirao-serie-a&seasonId=101053&popularLeaguesOnly=0

&excludeSpecialStatus=0&status=all&sortBy=default

&startDate=2023-10-30T03%3A00%3A00Z&endDate=2023-11-06T02%3A59%3A59Z

&round=&page=1&perPage=100&perPageDefault=100&language=br"

Já no caso dos odds, é necessário optar pela opção "getMatchOdds" e os parâmetros a serem configurados são "matchKey" que foi obtida juntamente às informações da par-

tida e é um identificador único de cada jogo e o parâmetro "oddGroupId" que é um código para cada mercado disponível. A url a seguir demonstra uma requisição de um jogo cujo id é 4122 para o mercado h2h. Ademais o parâmetro "inplay" foi sempre escolhido como 0, para que apenas os odds antes da partida começar fossem retornados.

$$\begin{cases} oddGroupId = 1 \Rightarrow \text{mercado} = h2h \\ oddGroupId = 4 \Rightarrow \text{mercado} = over/under \\ oddGroupId = 6 \Rightarrow \text{mercado} = spread \\ oddGroupId = 8 \Rightarrow \text{mercado} = exact\ score \\ oddGroupId = 11 \Rightarrow \text{mercado} = both\ score \end{cases}$$

"https://oddspedia.com/api/v1/getMatchOdds?wettsteuer=0&geoCode=BR

&bookmakerGeoCode=BR&bookmakerGeoState=&matchKey=4122

&oddGroupId=1&inplay=0&language=br"

5.1.2 Informações

O arquivo de informações sobre as partidas contém os nomes dos times que jogaram, o número de gols de cada um na partida, o momento do campeonato em rodadas e a data e hora da partida, de forma que essas informações podem ser acessadas por meio de identificadores únicos de cada partida.

5.1.3 Odds

O arquivo de odds das partidas contém os mesmos identificadores únicos das partidas do arquivo de informações, porém os identificadores das partidas levam às repectivas casas de apostas, que por sua vez levam aos seus respectivos mercados disponíveis para aquela partida e os mercados levam aos odds.

5.1.4 Ligas

As ligas cujos odds foram coletados para este estudo são algumas dentre as ligas mais populares no Brasil e no mundo. São elas:

1. Brasileirão Série A

2. Brasileirão Série B
3. Bundesliga
4. CONMEBOL Libertadores
5. Campeonato Baiano
6. Campeonato Carioca A
7. Campeonato Gaúcho
8. Campeonato Mineiro
9. Copa América
10. Copa Sul-Americana
11. Copa Verde
12. Copa do Brasil
13. Copa do Nordeste
14. Euro Classificação
15. Eurocopa
16. Liga Conferência
17. Liga Europa
18. Liga dos Campeões
19. Ligue 1
20. Premier League
21. Primeira Divisão
22. Série A (Itália)

5.2 Superfície de Probabilidade

5.2.1 Conceito

Para cada partida no banco de dados, há um interesse especial nos odds *Correct Score*, uma vez que a Definição 2 permite obter estimativas das probabilidades para cada placar possível para a partida. Dispor das estimativas das probabilidades do resultado da partida permite a investigação da consistência dos odds entre as diferentes casas de aposta.

Para superfícies de probabilidade em que o eixo X representa os gols do time da casa e Y os gols do time visitante, com o valor dos eixos truncados em 6, em geral, a soma das probabilidades estimadas já soma mais do que 1, refletindo a margem de lucro da casa de aposta, uma vez que as casas de aposta inserem artificialmente uma margem de lucro sobre os odds. Portanto, as superfícies serão analisadas com agrupamento, de forma que X e Y variam em $\{0, 1, 2, 3, 4, 5, 6+\}^2$, em que "6+" significa "6 ou mais gols".

Em específico, este estudo se propõe a investigar uma família de modelos capaz de fornecer uma descrição concisa de toda a superfície de probabilidade. Em outras palavras, o trabalho passa a ser a estimação dos parâmetros do modelo ao invés de estimar cada um dos placares possíveis para a partida.

Uma vez que nem todas as casas de aposta disponibilizam os odds para os placares exatos, tem-se que os demais mercados listados na seção 2.3 fornecem partições da superfície de probabilidade, portanto também contribuem para a reconstrução da superfície.

5.2.2 Partições

Cada um dos mercados de apostas informados na secção 2.3 foi coletado com o intuito de fornecer estimativas para as probabilidades de cada par $(x, y) \in \{0, 1, 2, 3, 4, 5, 6+\}^2$ possível.

Por exemplo, o resultado de empate descreve os pares (x, y) tais que $x = y$, logo, os odds para empate descrevem a soma das probabilidades ao longo da diagonal da superfície, quando $x = y$, conforme a Definição 2. Outro caso é a vitória do time da casa, que descreve os pares (x, y) tais que $x > y$, ou seja, os odds para a vitória do time da casa descrevem a soma de todas as probabilidades das caselas da superfície abaixo da diagonal do empate.

De forma análoga para os demais mercados de apostas, a Figura 2 ilustra em azul quais partições da superfície são descritas pelos odds para cada mercado disponível. O total de gols escolhido foi de 2,5 e o *spread* foi de 2 gols de desvantagem para o time da casa.

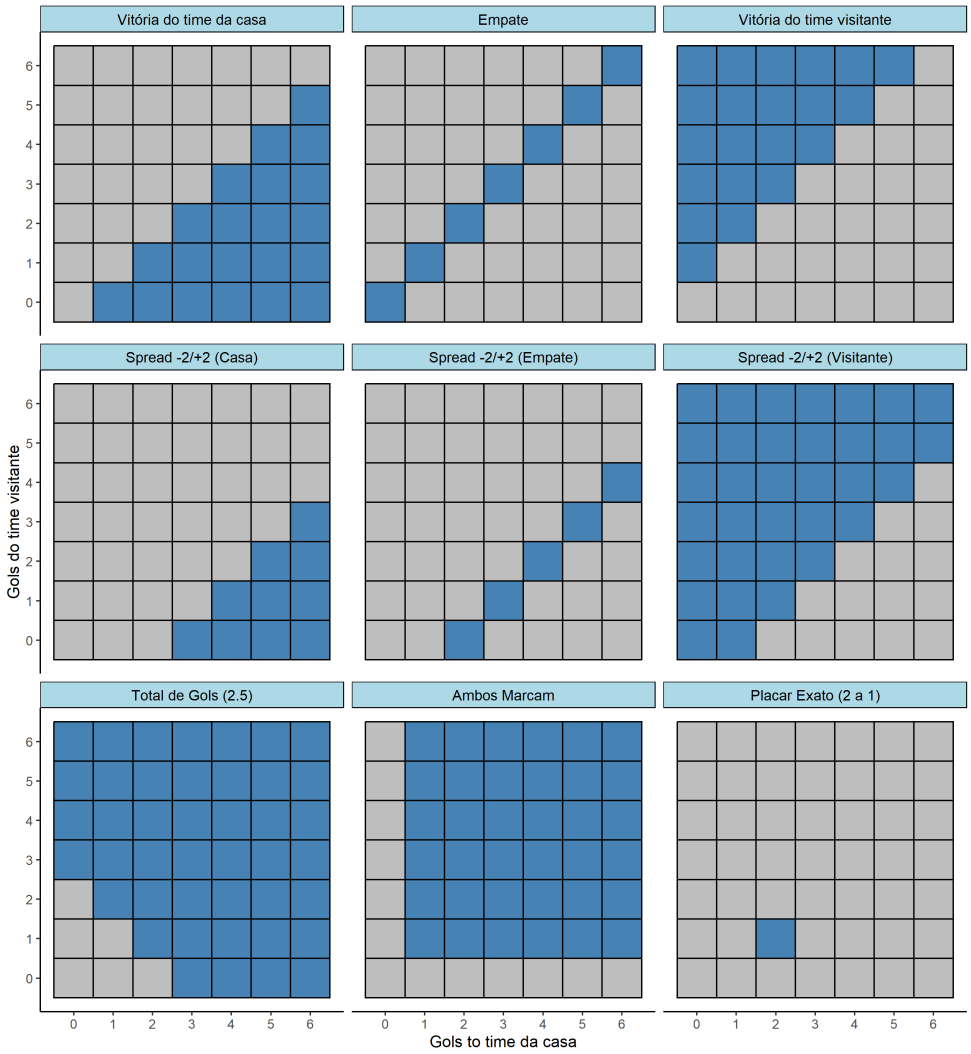

Figura 2: Conjunto de resultados cujas probabilidades são descritas pelos odds de cada mercado de apostas, caso o resultado positivo se concretize.

Cada um dos mercados de aposta pode ser interpretado como uma ou mais equações que ajudam a descrever a superfície de probabilidade. Sob a notação em que p_{ij} é a probabilidade de ocorrer i gols do time da casa e j gols do time visitante e $\mathcal{O}_{Mercado}$

é o odd para um dado mercado de aposta, segue da Figura 2 que

$$\begin{cases} \frac{1}{\mathcal{O}_{Empate}} & = \sum_{i=0}^{6} p_{ii} \\ \frac{1}{\mathcal{O}_{Casa}} & = \sum_{i=0}^{6} \sum_{j=0}^{6} p_{ij} \cdot I_{(i>j)} \\ \frac{1}{\mathcal{O}_{Visitante}} & = \sum_{i=0}^{6} \sum_{j=0}^{6} p_{ij} \cdot I_{(j<i)} \\ \frac{1}{\mathcal{O}_{Over}} & = \sum_{i=0}^{6} \sum_{j=0}^{6} p_{ij} \cdot I_{(j+i>T)} \\ \frac{1}{\mathcal{O}_{Under}} & = \sum_{i=0}^{6} \sum_{j=0}^{6} p_{ij} \cdot I_{(j+i<T)} \\ \frac{1}{\mathcal{O}_{Spread_{Casa}}} & = \sum_{i=0}^{6} \sum_{j=0}^{6} p_{ij} \cdot I_{(i+S>j)} \\ \frac{1}{\mathcal{O}_{Spread_{Empate}}} & = \sum_{i=0}^{6} \sum_{j=0}^{6} p_{ij} \cdot I_{(i+S=j)} \\ \frac{1}{\mathcal{O}_{Spread_{Visitante}}} & = \sum_{i=0}^{6} \sum_{j=0}^{6} p_{ij} \cdot I_{(i+S<j)} \\ \frac{1}{\mathcal{O}_{(Sim)Ambos\ \ Marcam}} & = \sum_{i=1}^{6} \sum_{j=1}^{6} p_{ij} \\ \frac{1}{\mathcal{O}_{(Nao)Ambos\ \ Marcam}} & = \sum_{j=1}^{6} p_{0j} + \sum_{i=0}^{6} p_{i0} \end{cases}$$

em que T é o total de gols fixado (2,5 para a Figura 2), S é a vantagem ou desvantagem do Spread para o time da casa (2 gols contra o time da casa, para a Figura 2) e $I_{(desigualdade)}$ é a função Indicadora, apresentada na equação (5.2.1). Aliadas às probabilidades fornecidas pelos odds para cada casela pelo mercado *Correct Score*, tem-se que todas as equações contribuem para a reconstrução da superfície.

$$I_{(desigualdade)} = \begin{cases} 1, \text{ se a desigualdade é satisfeita} \\ 0, \text{ caso contrário} \end{cases} \tag{5.2.1}$$

5.2.3 Representação Vetorial

A Figura 2 apresenta as superfícies de probabilidade em planos cartesianos, que sugerem que a superfície poderia ser representada por uma matriz. Cada mercado pode ser representado por uma matriz contendo apenas valores 0 ou 1, 0 se o mercado não diz respeito a uma certa probabilidade e 1 caso contrário. Por exemplo, partindo da Figura 2, a matriz $\boldsymbol{M}$ do cenário "vitória do time da casa" é dada por

$$\boldsymbol{M} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Considere a representação dos elementos de uma matriz M por M_{ij}, em que $i, j \geq 0$, então faria sentido que o índice M_{10} represente o placar 1 a 0 para o time da casa, dessa forma obtém-se que a matriz $\mathbb{M}$ para o cenário "vitória do time da casa" é dada por

$$\mathbb{M} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Todavia, é possível eliminar a bidimensionalidade da representação matricial e tomar uma representação vetorial em que os placares seguem 0 a 0, 0 a 1, 0 a 2,..., 1 a 0, 1 a 1, ..., 6+ a 5, 6+ a 6+. Ou seja, o cenário "vitória to time da casa" pode ser presentado pelo vetor $\mathbb{V}$ da forma

$$\mathbb{V} = \text{vec}\left(\mathbb{M}^T\right)$$

A partir dessa seção, o vetor correspondente a um dado cenário de um dado mercado será denotado pela notação $\mathbb{V}_{[cenario]}$, de forma que a representação matricial do mercado é composta por todos os seus cenários, por exemplo, para o mercado h2h tem-se a vitória do time da casa, do time visitante ou empate, então sua representação matricial é dada por

$$\mathbb{M}_{[h2h]} = \begin{bmatrix} \mathbb{V}^T_{[casa]} \\ \mathbb{V}^T_{[empate]} \\ \mathbb{V}^T_{[visitante]} \end{bmatrix}$$

já mercados com submercados, como *spread*, é preciso específicar o submercado, como por

exemplo $spread+2/-2$, então

$$\mathbb{M}_{[spread(+2/-2)]} = \begin{bmatrix} \mathbb{V}^T_{[casa]} \\ \mathbb{V}^T_{[empate]} \\ \mathbb{V}^T_{[visitante]} \end{bmatrix}$$

em ambos os casos os vetores $\mathbb{V}$ são respectivos ao respectivo mercado a qual a matriz $\mathbb{M}$ corresponde.

5.3 Modelo Paramétrico

5.3.1 Parametrização

Embasado pelos estudos anteriores e pelos dados disponíveis, tem-se o conhecimento de uma estrutura de dependência entre o número de gols do time da casa e do time visitante. Portanto, este estudo propõe a pesquisa por um modelo capaz de se adequar a essa estrutura de dependência por meio do uso de cópulas, para gerar observações correlacionadas, em conjunto com uma mistura entre as famílias de distribuições Binomial Negativa, responsável pelo modelo geral da superfície, e de Poisson, responsável por lidar com a potencial subestimação das probabilidades para resultados de empate.

Conforme estudos passados (MCHALE; SCARF, 2006), a escolha da Binomial Negativa para as marginais permite maior flexibilidade para comportamento do modelo, pois há a presença de 2 parâmetros para cada uma dar marginais, já para a inflação da diagonal, tem-se que tem-se que a distribuição de Poisson parece ser flexível o suficiente (KARLIS; NTZOUFRAS, 2003).

O modelo é composto pelos parâmetros

1. r_X, p_X = Parâmetros da Binomial Negativa para o número de gols do time da casa;
2. r_Y, p_Y = Parâmetros da Binomial Negativa para o número de gols do time visitante;
3. ρ = Parâmetro de correlação entre as distribuições marginais;
4. ϵ = Parâmetro que regula a mistura de distribuições ao incluir a distribuição que inflaciona a diagonal da superfície.
5. α = Parâmetro da distribuição de Poisson, que descreve o inflacionamento as probabilidades da diagonal da superfície (empates);

5.3.2 Processo gerador de dados

Uma vez que os parâmetros do modelo são conhecidos, então o modelo proposto é obtido a partir do processo gerador de dados

Passo 1: Gere uma amostra de Z, tal que $Z \sim Bernoulli(\epsilon)$

$$\text{Se Z} = 0 \begin{cases} \text{Passo 2: Gere } W, \text{ tal que } W \sim Poisson(\alpha) \\ \text{Passo 3: Obtenha } \boldsymbol{W} = (W, W) \end{cases}$$

$$\text{Se } Z = 1 \begin{cases} \text{Passo 2: Gere o vetor } (X, Y), \text{ tal que } (X, Y) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \\ \text{Passo 3: Obtenha } (U, V) = (\Phi(X), \Phi(Y)) \\ \text{Passo 4: Obtenha } \boldsymbol{W} = (W_1, W_2) = (F^{-1}(U|r_X, p_X), F^{-1}(V|r_Y, p_Y)) \end{cases}$$

Último passo: $\boldsymbol{W} = (\min\{W_1, 6\}, \min\{W_2, 6\})$

em que $\boldsymbol{W}$ é o vetor obtido como amostra e F^{-1} é a função quantil da Binomial Negativa descrita na seção 4.4.

O modelo proposto é uma distribuição de probabilidade conjunta $P(W_1 = w_1, W_2 = w_2)$, em que o suporte dessa distribuição de probabilidade é dado por

$$(W_1, W_2) \in \{0, 1, 2, 3, 4, 5, 6+\}^2$$

Ou seja, o suporte da distribuição dos placares da partida sob o modelo proposto envolve apenas 49 pares, com probabilidade não nula.

5.3.3 Distribuição Placar

Sabe-se que o processo gerador de dados apresentado é uma mistura entre uma distribuição de Poisson e uma outra distribuição cuja forma não é óbvia ao pesquisador, mas que será apelidada de distribuição Placar. Analogamente a mistura com a distribuição de Poisson será apelidada de distribuição Placar Ajustada.

O processo gerador de dados proposto na seção 5.3.2 é computacionalmente pouco eficiente para o método de ajuste dos parâmetros a ser discutido na seção 5.3.4.

Para melhorar a eficiência computacional cogitou-se escrever uma forma analítica para a distribuição Placar. O fato do suporte da distribuição Placar envolver apenas 49 pares permite pensar em expressar a sua função de probabilidade em termos de uma

distribuição Normal Bivariada.

Considerando apenas a distribuição Placar, suponha que uma iteração do processo gerador de dados resultou em um vetor (w_1, w_2), então sabe-se que esse vetor foi obtido a partir da função quantil da distribuição Binomial Negativa aplicada em certos valores (u, v), mas não é possível determinar os valores do vetor (u, v), porque a função quantil da distribuição Binomial Negativa não é uma função injetora.

Por outro lado, é possível determinar o menor valor u_1 para o qual $F^{-1}(u_1|r_X, p_X) = w_1$ e analogamente o maior valor u_2 para o qual $F^{-1}(u_2|r_X, p_X) = w_1$. Analogamente para w_2, tem-se para $(w_1, w_2) \in \{0, 1, 2, 3, 4, 5, 6+\}$ que

$$\begin{aligned}
&P(W_1 = w_1, W_2 = w_2) = \\
&= P(u_1 < U < u_2, v_1 < V < v_2) \\
&= P(\Phi^{-1}(u_1) < \Phi^{-1}(U) < \Phi^{-1}(u_2), \Phi^{-1}(v_1) < \Phi^{-1}(V) < \Phi^{-1}(v_2)) \\
&= P(\Phi^{-1}(u_1) < X < \Phi^{-1}(u_2), \Phi^{-1}(v_1) < Y < \Phi^{-1}(v_2))
\end{aligned}$$

Uma vez que W foi definido a partir da função quantil da distribuição Binomial Negativa, é possível determinar (u_1, u_2) a partir da inversa da função quantil, isto é, a função de distribuição acumulada aplicada na forma

$$\begin{cases} u_1 = P(W_1 \leq w_1 - 1|r_X, p_X) \\ u_2 = P(W_1 \leq w_1|r_X, p_X) \end{cases}$$

o processo para determinar (v_1, v_2) é análogo.

Quando alguma das componentes do vetor (w_1, w_2) é 0 ou 6, deve-se tomar cuidado com a função de distribuição acumulada da Binomial Negativa, portanto, convém-se definir as variáveis aleatórias G e H tais que

$$H \sim Binomial\ Negativa(r, p)$$

$$P(G \leq g|r, p) = \begin{cases} P(H \leq g|r, p) & ,\ \text{se } g < 6 \\ 1 & ,\ \text{se } g \geq 6 \end{cases}$$

Ou seja, a tarefa de escrever a distribuição Placar de forma analítica não foi concluída, não obstante, alcançou-se uma expressão em termos da distribuição Normal Bivariada, o que é computacionalmente mais eficiente que o processo gerador de dados da seção 5.3.2.

Retomando à distribuição Placar Ajustada, para a distribuição de Poisson, responsável pela inflação da diagonal, convém definir as variáveis aleatórias C e D, tais que

$$C \sim Poisson(\alpha)$$

$$P(D = k|\alpha) = \begin{cases} P(C = k|\alpha), \text{ se } k \neq 6 \\ P(C \geq 6|\alpha), \text{ se } k = 6 \end{cases}$$

Finalmente é possível obter a função de probabilidade da distribuição Placar Ajustada para cada par possível $(w_1, w_2) \in \{0, 1, 2, 3, 4, 5, 6+\}$ por meio do Algoritmo 1

Algoritmo 1: Cálculo da Distribuição Placar Ajustada

Input:
- $\mathbf{\Theta} = (r_X, r_Y, p_X, p_Y, \rho, \epsilon, \alpha)$;
- Número máximo de gols M, nesse estudo foi utilizado sempre como 6.

Output:
- Vetor de probabilidades com M^2 entradas, incrementando os placares primeiro do time visitante e depois do time da casa.

Compute:

$gols = \{0, 1, 2, ..., M\}$
$probabilidades[(M+1)^2] = \{\}$
for $w_1 \in gols$ **do**
  for $w_2 \in gols$ **do**
    $u_1, u_2 = (P(G \leq w_1 - 1|r_X, p_X), P(G \leq w_1|r_X, p_X))$
    $v_1, v_2 = (P(G \leq w_2 - 1|r_Y, p_Y), P(G \leq w_2|r_Y, p_Y))$
    $x_1, x_2 = (\Phi^{-1}(u_1), \Phi^{-1}(u_2))$
    $y_1, y_2 = (\Phi^{-1}(v_1), \Phi^{-1}(v_2))$
    $P_{XY} = P(x_1 < X < x_2,\ y_1 < Y < y_2|\rho)$
    $P_D = P(D = w_1|\alpha) I_{(w_1 = w_2)}$
    $probabilidades[7w_1 + w_2] = \epsilon P_{XY} + (1 - \epsilon) P_D$
return *probabilidades*

O Algoritmo 1 evidencia que as probabilidades do modelo proposto podem ser obtidas a partir da mistura de uma distribuição Normal Bivariada e da distribuição de Poisson. A implementação do Algoritmo 1 se mostrou mais de 1000 vezes computacionalmente mais eficiente que a implementação do processo gerador de dados descrito na seção 5.3.2.

5.3.4 Estimação dos parâmetros do modelo

Por meio da propriedade probabilística dos mercados de apostas descrita na seção 2.3.3 e das equações descritas no sistema de equações 5.2.2, segue-se que é possível utilizar as equações de um mercado para obter a probabilidade correspondente fornecida pelo modelo proposto, assim obtém-se uma distribuição de probabilidade fornecida pelos odds da casa de apostas e uma distribuição de probabilidade fornecida pelo modelo proposto.

Ambas as distribuições de probabilidade obtidas fornecem probabilidades para os mesmos eventos de interesse, então é possível comparar tais distribuições de probabilidade ao calcular a Divergência de Kullback-Leiber (D_{KL}), por meio de derivadas numéricas é possível saber como deve-se ajustar os parâmetros do modelo para que a D_{KL} seja reduzida, após diversas iterações espera-se que as probabilidades fornecidas pelo modelo sejam muito similares às encontradas por meio dos odds.

A função de probabilidade resultante do Algoritmo 1 descreve as probabilidade para 49 casos possíveis, é possível alocar todas as 49 probabilidades em um vetor $\mathbb{T}$, de forma análoga ao caso da seção 5.2.3. A vantagem dessa representação é que para um dado mercado, tem-se que as probabilidades correspondentes a cada cenário são dadas por

$$\mathbb{M}_{[mercado]} \cdot \mathbb{T} = \begin{bmatrix} \text{Probabilidade do cenário 1} \\ \text{Probabilidade do cenário 2} \\ \vdots \\ \text{Probabilidade do cenário n} \end{bmatrix} = \mathbb{Q}_{[mercado]}$$

aloque as probabilidades obtidas a partir dos odds, de uma dada casa de apostas, para um dado mercado, no vetor $\mathbb{P}_{[mercado]}$, então

$$D_{KL} = \sum_{i=1}^{n} \mathbb{P}_{[mercado]i} \log\left(\frac{\mathbb{P}_{[mercado]i}}{\mathbb{Q}_{[mercado]i}}\right)$$

em que $\mathbb{V}_i, i = 1, 2, 3, ..., n$ é a i-ésima componente de um vetor $\mathbb{V}$ qualquer.

Fixando a casa de apostas, é possível calcular a D_{KL} para todos os mercados que a casa de apostas forneceu e obter a média $\bar{D}_{KL}$ dentre todas as divergências obtidas, então o processo de estimação dos parâmetros do modelo segue de forma análoga ao descrito para um único mercado, calculando derivadas numéricas e minimizando a métrica de distância $\bar{D}_{KL}$ entre a distribuição do modelo e da casa de apostas.

Em termos matriciais, seja o suporte dos mercados dado por

$$\mathbb{C} = \{h2h, spread(+1/-1), spread(+2/-2), ..., over/under(0.5), ..., both\ score, exact\ score\}$$

podendo ser reduzido caso algum mercado não esteja disponível, então

$$\bar{D}_{KL} = \frac{1}{\|\mathbb{C}\|} \sum_{mercado \in \mathbb{C}} \sum_{i=1}^{n} \mathbb{P}_{[mercado]i} \log \left(\frac{\mathbb{P}_{[mercado]i}}{\mathbb{Q}_{[mercado]i}} \right)$$

Esse processo de otimização é descrito pelos Algoritmos 2 e 3. A função optim da linguagem R foi utilizada especificamente para viabilizar o Algoritmo 3, lidando com as derivadas numéricas e o processo de minimização da Divergência de Kullback-Leibler, em geral.

Algoritmo 2: Cálculo da Divergência Média de Kullback-Leibler

Dependências:
- Função $P(\boldsymbol{\Theta}, M)$, que executa o Algoritmo 1;
- Operador $\bigodot$, para multiplicação matricial.

Input:
- $\boldsymbol{\Theta} = (r_X, r_Y, p_X, p_Y, \rho, \epsilon, \alpha)$;
- Número máximo de gols M, nesse estudo foi utilizado sempre como 6;
- $shin_probs$, um dicionário (Python) em que os mercados ou submercados levam nas probabilidades normalizadas pelo método de Shin em uma casa de apostas;
- $mascara$, um dicionário (Python) contendo como chave os mercados e submercados e como valor um vetor binário, como descrito na seção 5.2.3;

Output:
- Um escalar que aponta a Divergência média de Kullback-Leibler entre a distribuição Placar Ajustada proposta e a casa de apostas.

Compute:

$probabilidades = P(\boldsymbol{\Theta}, M)$
$D_{KL} = 0$
for $mercado \in shin\_probs$ **do**
  $probs\_mercado = probabilidades \bigodot mask$
  for $i \in \{0, 1, ..., length(shin\_probs[mercado])\}$ **do**
    $D_{KL} \mathrel{+}= shin\_probs[mercado][i] \cdot \log \left(\frac{shin\_probs[mercado][i]}{probs\_mercado[i]} \right)$
$mean\_D_{KL} = \frac{D_{KL}}{length(odds)}$
return $mean\_D_{KL}$

Algoritmo 3: Ajuste da Distribuição Placar Ajustada

Dependências:

- Função $P(\boldsymbol{\Theta}, M)$, que executa o Algoritmo 1;
- Função $MDKL(\boldsymbol{\Theta}, shin\_probs, mascara)$, que retorna o resultado Algoritmo 2 a partir das probabilidades já calculadas por $P(\boldsymbol{\Theta}, M)$;
- Função $N(odds)$, que fornece probabilidades normalizadas pelo método de Shin a partir dos odds de um mercado ou submercado;
- Função $\nabla(\boldsymbol{\Theta}, Fun)$, em que Fun é uma função avaliada no conjunto de parâmetros $\boldsymbol{\Theta}$. A função retorna um gradiente numérico da função Fun;
- Função $Patch(\boldsymbol{\Theta}, Fun, grad\_Fun)$, que retorna um conjunto de parâmetros que minimiza a função Fun entre as opções fornecidas pelo gradiente $grad\_Fun$;

Input:

- Chute inicial qualquer $\boldsymbol{\Theta} = (r_X, r_Y, p_X, p_Y, \rho, \epsilon, \alpha)$;
- Número máximo de gols M, nesse estudo foi utilizado sempre como 6;
- *odds*, um dicionário (Python) em que os mercados ou submercados levam nos odds de uma casa de apostas;
- *mascara*, um dicionário (Python) contendo como chave os mercados e submercados e como valor um vetor binário, como descrito na seção 5.2.3;
- Tolerância de convergência $\epsilon > 0$. • *maxit*, o critério de parada do algoritmo, pode ser por exemplo número de atualizações sem melhora.

Output:

- Vetor de 7 parâmetros que descrevem a superfície de probabilidade reconstruída por meio da distribuição Placar Ajustada.

Compute:

$min\_D_{KL} = \infty$
$probabilidades = P(\boldsymbol{\Theta}, M)$
$shin\_probs = N(odds)$
while $no\_improv < maxit$ **do**
  $mean\_D_{KL} = MDKL(probabilidades, shin\_probs, mascara)$
  if $mean\_D_{KL} < min\_D_{KL}$ **then**
    $min\_D_{KL} = mean\_D_{KL}$
    $no\_improv = 0$
  else
    $no\_improv \mathrel{+}= 1$
  $grad\_Fun = \nabla(\boldsymbol{\Theta}, MDKL)$
  $\boldsymbol{\Theta} = Patch(\boldsymbol{\Theta}, MDKL, grad\_Fun)$
return $\boldsymbol{\Theta}$

5.4 Modelo Médio

Ao ajustar modelos como na seção 5.3.3 tem-se que cada jogo, cada casa de apostas apresentará um conjunto de 7 parâmetros que descreve as probabilidades segundo seus odds disponibilizados, entretanto, espera-se que os parâmetros de cada casa de apostas sejam similares, pois descrevem probabilidades para os placares do mesmo jogo.

O estudo de Kaunitz, Zhong e Kreiner (2017) propõe, em um escopo mais simples, que tome-se a média das probabilidades de cada casa de apostas para um mesmo resultado, essa média é chamada de "probabilidade consensual" e permite a tomada de decisão para a alocação das apostas. Baseando-se nessa ideia de tomar a média das medidas obtidas para a tomada de decisão, esse estudo propõe a construção de "superfícies consensuais" a partir de um "modelo médio" entre os obtidos para as casas de apostas disponíveis.

Se existem n casas de apostas disponíveis para um dado jogo, então obtêm-se os parâmetros estimados

$$\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n$$

em que θ representa qualquer um dos 7 parâmetros do modelo. Supondo que exista uma superfície de probabilidade que realmente descreve as probabilidades dos placares acontecerem, que também pode ser descrita pelo modelo proposto neste estudo e que as superfícies normalizadas obtidas a partir das casas de apostas se aproximam dessa superfície real, então uma ideia para estimar o parâmetro real θ para um dado jogo, é tomar o estimador

$$\bar{\hat{\theta}} = \frac{\sum_{i=1}^{n} \hat{\theta}_i}{n} \approx \theta$$

em que θ representa qualquer um dos parâmetros do model, ou seja, define-se o conjunto de parâmetros estimados

$$\hat{\Theta} = \left(\bar{\hat{r}}_X, \bar{\hat{r}}_Y, \bar{\hat{p}}_X, \bar{\hat{p}}_Y, \bar{\hat{\rho}}, \bar{\hat{\epsilon}}, \bar{\hat{\alpha}}\right)$$

e as probabilidades da superfície consensual podem ser obtidas como no Algoritmo 1, para cada par possível $(w_1, w_2) \in \{0, 1, 2, 3, 4, 5, 6+\}$, calcula-se

$$P\left(W_1 = w_1, W_2 = w_2 | \hat{\Theta}\right)$$

5.5 Modelo Preditivo

Ao obter os modelos médios para cada jogo disponível, há apenas uma representação média obtida a partir das casas de apostas para os possíveis resultados da par-

tida. Pode-se considerar prontamente utilizar esse resultado para alocação de apostas, de forma similar ao estudo de (KAUNITZ; ZHONG; KREINER, 2017).

Em contrapartida, este estudo considera que pode-se obter um modelo mais robusto, por meio da adição de novas características que podem ser obtidas antes da partida ou talvez apenas abstraindo relações não consideradas pelo modelo médio, especialmente para os casos em que o modelo médio não se comporta tão bem.

O estudo de Karlis e Ntzoufras (2003) aponta que modelos anteriores subestimam os empates, por esta razão o modelo proposto na seção 5.3.3 dispõe de parâmetros que regulam uma inflação da diagonal dos empates, todavia, podem existir pontos descalibrados não considerados no modelo.

Uma forma de lidar com situações mal calibradas pelo modelo médio é realizar a recalibração das probabilidades via regressão isotônica, de forma que no caso do resultado do mercado ser ternário, como é no h2h, tem-se a regressão isotônica ajustada sobre a probabilidade do cenário ocorrer supervisionada pela ocorrência do cenário ou não no resultado real da partida.

Já para incluir informações adicionais no modelo que possam contribuir com o ajuste das probabilidades é o ajuste de uma rede neural sobre as informações adicionais e o vetor de 49 probabilidades do modelo médio, supervisionado pelos resultados reais da partida, de modo que a rede neural deve apresentar dois objetivos, um para a quantidade de gols do time da casa e um para a quantidade de gols do time visitante, dessa forma a rede não será penalizada por errar o placar apenas de um dos times.

Ademais, para a rede neural, o problema deve ser tratado como uma regressão ordinal, visando reduzir a perda caso a predição do placar seja próxima ao placar real, isto é, caso o placar real seja de 2 gols para o time da casa, então as predições 1 e 5 gols para o time da casa estão ambas erradas, porém são penalizadas de forma diferente, pois 1 gol é muito mais próximo do placar real.

Quanto a adição de novas características, deseja-se utilizar características que poderiam alterar a precificação das casas de apostas e torná-la menos fidedigna às probabilidades reais. O uso da normalização de Shin já considera o *favourite longshot bias*.

Nas diversas ligas em estudo o formato da competição pode fazer com que ocorram jogos com baixa importância, em situações que nem a vitória, empate ou derrota pode alterar a situação dos times no campeonato, portanto, espera-se que jogos desse tipo apresentem menor apelo do público, uma ideia para quantificar esse apelo é utilizar a API do Google Trends para consultar o número de buscas por aquele time nas últimas 24h

antes do jogo.

Suspeita-se que situações extremas também possam levar a má precificação pelas casas de apostas, cenários como alta pluviometria, temperaturas muito elevadas ou muito baixas, locais com pressão atmosférica muito baixa por conta de sua posição geográfica, entre outros cenários. Essas informações podem ser obtidas ao conhecer o estádio, a cidade, data e hora em que o jogo ocorreu ou ocorrerá, há institutos de monitoramento que realizam esse tipo de medição.

6 Resultados

6.1 Visualização das Superfícies

Como levantado na seção 2.2, a aplicação da *arbitrage* e de outras técnicas de apostas esportivas motivam o estudo das probabilidades dos potenciais resultados em um evento esportivo. Uma vez conhecidas as probabilidades, é possível alocar os recursos financeiros de forma que o apostador obtenha uma margem de perda mitigada ou até mesmo lucro garantido sobre as apostas.

Na seção 2.3 os mercados de apostas esportivas disponíveis foram apontados, entre eles há um interesse particular pelos odds do mercado *Correct Score*, uma vez que esse mercado permite obter as superfícies de probabilidade estimadas para cada casa de aposta de forma imediata ao fornecer os odds para cada placar possível da partida. Aplicando a Definição 2, obtém-se a Figura 3

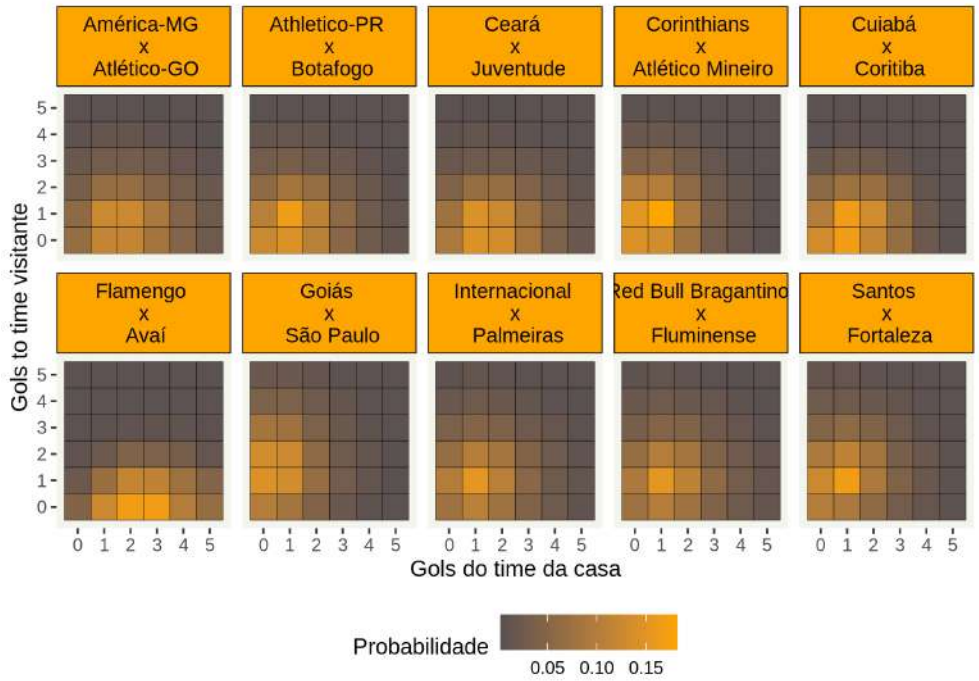

Figura 3: Superfícies de probabilidade fornecidas pela 1xBet para os jogos da 38ª rodada da Série A do Brasileirão de 2022.

que evidencia quais eram os placares mais prováveis de acordo a estimação provinda pela 1xBet para cada um dos jogos da última rodada da Série A do Brasileirão de 2022.

Analogamente obtém-se a Figura 4, que por sua vez tem o papel de apresentar a existência de variabilidade entre as estimativas provindas pelas diversas casas de aposta e

consequentemente fornece subsídio para se acreditar que deve haver um modelo específico para cada casa de aposta. Com bastante atenção, percebe-se que a etimativa para o placar 2 a 0 a favor do time da casa (Flamengo) segundo a casa "Mr Green" difere da estimativa da "NetBet", aproximadamente 15% e 18% respectivamente.

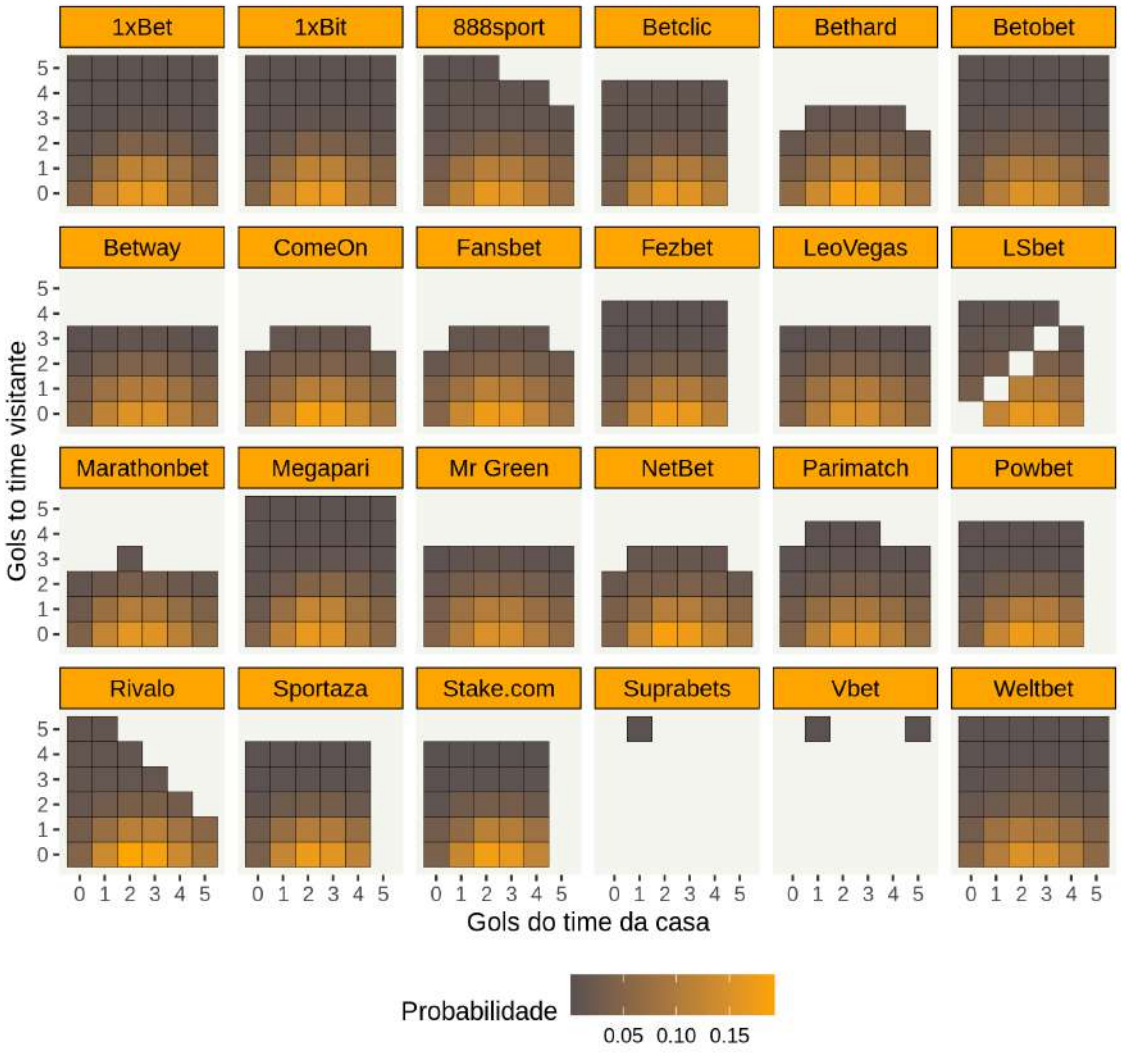

Figura 4: Superfíecies de probabilidade forneceidas por diversas casas de aposta para o jogo entre Flamengo e Avaí pela 38ª da Série A do Brasileirão de 2022

Outro ponto importante a ser notado da Figura 4 é que algumas casas de apostas não abrem muitas possibilidades para o mercado de Correct Score e podem abrir apenas apostas para resultados não muito relevantes.

Especificamente na partida escolhida para a Figura 4 tem-se que os resultados mais significativos se concentravam para o time visitante marcar 2 gols ou menos e o time da casa marcar 5 gols ou menos, entretanto, as casas Suprabets e Vbet forneceram apostas para os placares "1x5" e "5x5", ou seja, apenas um ruído na especificação da superfície,

portanto, casos como esse foram retirados do banco.

Formalmente, caso fossem fornecidos menos de 10 placares exatos, então esse mercado seria descartado da análise dessa casa de apostas. O mercado *spread* também apresentou incosistências, alguns odds não vieram padronizados nos submercados na forma "+k/-k" ou "-k/+k", com k inteiro, então o mercado seria descartado caso "k" informado fosse não inteiro. Exceto *Correct Score*, caso qualquer mercado não apresentasse odds para todos os cenários possíveis da partida, então seriam descartados.

6.2 Testes de Correlação

Estudos passados apontam que existe correlação entre o número de gols do time da casa e o número de gols to time visitante (DIXON; COLES, 1997) e (KARLIS; NTZOUFRAS, 2003), portanto, uma vez que o banco de dados obtido para este estudo tem um fator diferencial dos estudos passados ao considerar ligas brasileiras de futebol, foram calculas as proporções amostrais para cada placar possível para quase todas as ligas do banco de dados, por sua vez essas proporções formam superfícies de probabilidades que estão apresentadas na Figura 5

Da Figura 5 é possível notar uma distribuição particular do número de gols em cada liga, uma possibilidade para explicar esse fenômeno é o conjunto de regras de cada liga, por exemplo, caso o saldo de gols seja importante para a classificação dos times no campeonato, então é provável que os times disputem as partidas não só com intenção de ganhar, mas de marcar o maior número de gols possível.

Para dar prosseguimento à análise da correlação entre o número de gols do time da casa e o número de gols do time visitante, vale notar que cada superfície de probabilidade obtida na Figura 5 descreve as probabilidades conjuntas do número de gols de cada um dos times, logo, pode-se calcular as probabilidades marginais, esperanças, variâncias, covariâncias e correlações.

Uma vez que as correlações foram calculadas e o tamanho da amostra é conhecido para cada uma das ligas, obtém-se a Tabela 2, que indica os resultados do teste de correlação, para o nível de significância de 5%, pelas estatísticas do teste e seus p-valores, cujas hipóteses são

$$H_0)\ \rho = 0;\ H_1)\ \rho \neq 0$$

em que ρ é o coeficiente de correlação entre o número de gols do time da casa e número de gols do time visitante.

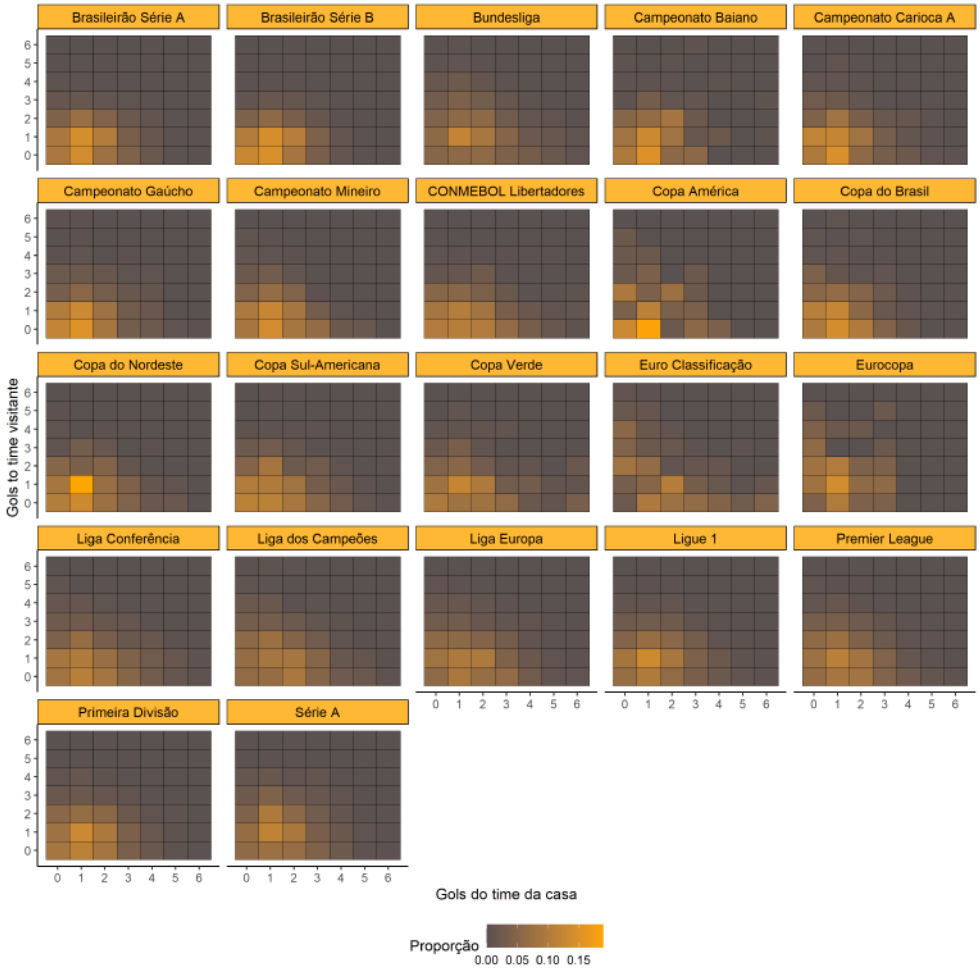

Figura 5: Proporção empírica dos placares das partidas de futebol para os anos de 2019 a 2022 por liga.

Da Tabela 2 tem-se que diversas ligas rejeitaram a hipótese de independência entre o número de gols do time da casa e do time visitante, por exemplo o Campeonato Mineiro apresentou um p-valor menor que 0.1%, muito inferior ao nível de significância de 5% pré-configurado, isto é, para o Campeonato Carioca e ligas em situação análoga existem evidências significativas de que há correlação entre o número de gols do time da casa e do time visitante.

6.3 Modelagem

Ciente das implicações de não poder assumir que a correlação entre o número de gols do time da casa e do time visitante é nula para diversas ligas, então seguiu-se com o ajuste do modelo paramétrico proposto na seção 5.3.3.

Para cada jogo, cada casa de aposta teve um modelo ajustado de acordo com os seus mercados de apostas disponíveis após passarem pela Normalização de Shin. O

Tabela 2: Resultados dos testes de ausência de correlação entre os gols do time da casa e do time visitante

Liga	n	r	Estatística	p-valor
Brasileirão Série A	1520	0.019	0.745	> 0.1
Brasileirão Série B	1518	0.052	2.036	0.042
Bundesliga	1213	-0.199	-7.059	< 0.001
Campeonato Baiano	200	0.129	1.830	0.069
Campeonato Carioca A	492	-0.046	-1.013	> 0.1
Campeonato Gaúcho	286	-0.008	-0.143	> 0.1
Campeonato Mineiro	299	-0.245	-4.348	< 0.001
CONMEBOL Libertadores	619	-0.087	-2.177	0.03
Copa América	54	-0.149	-1.088	> 0.1
Copa do Brasil	480	-0.098	-2.150	0.032
Copa do Nordeste	284	-0.091	-1.533	> 0.1
Copa Sul-Americana	524	-0.099	-2.266	0.024
Copa Verde	115	-0.138	-1.483	> 0.1
Euro Classificação	476	-0.482	-11.985	< 0.001
Eurocopa	51	-0.123	-0.865	> 0.1
Liga Conferência	789	-0.116	-3.276	0.001
Liga dos Campeões	818	-0.208	-6.083	< 0.001
Liga Europa	1234	-0.149	-5.293	< 0.001
Ligue 1	1422	-0.069	-2.617	0.009
Premier League	1491	-0.159	-6.232	< 0.001
Primeira Divisão	1501	-0.024	-0.935	> 0.1
Série A	1480	-0.068	-2.610	0.009

ajuste numérico do modelo levou a uma redução da D_{KL} em relação ao chute inicial dado pelas distribuições empíricas das ligas, para uma análise visual desse resultado a Figura 6 foi elaborada, em que há a superfície fornecida pelo mercado *Exact Score* de algumas casas de apostas e a superfície correspondente aos parâmetros ajustados, para o jogo entre Juventude e Coritiba pela 35ª rodada da Série A do Brasileirão de 2022.

As figuras 6 e 7 mostram uma grande similaridade entre as superfícies provindas das casas de apostas e as superfícies ajustadas, isso indica que o ajuste do modelo foi um sucesso com respeito à capacidade de descrever as probabilidades correspondentes aos odds fornecidos pelas casas de apostas por meio de uma distribuição de probabilidade com os 7 parâmetros propostos.

Embora visualmente as superfícies sejam muito similares, ao calcular a diferença entre as probabilidades das mesmas é possível notar diferenças aparecendo na terceira ou quarta casa decimal. O diagnóstico para tal diferença é que, devido à inconsistência dos odds entre os mercados e a métrica a ser otimizada ser a média das Divergências

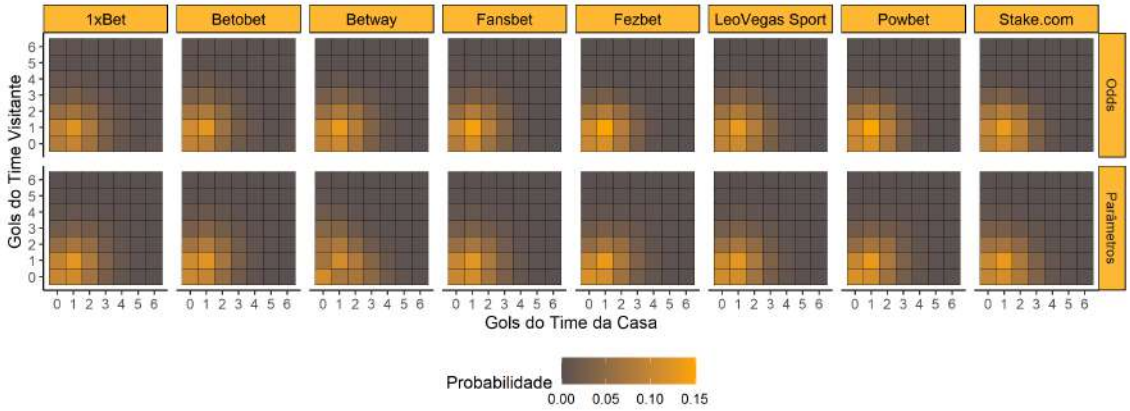

Figura 6: Comparações entre algumas superfícies obtidas por meio dos odds para *Exact Score* e superfícies obtidas por meio dos parâmetros estimados para a respectiva casa de apostas para o jogo entre Juventude e Coritiba pela 35ª rodada da Série A do Brasileirão de 2022.

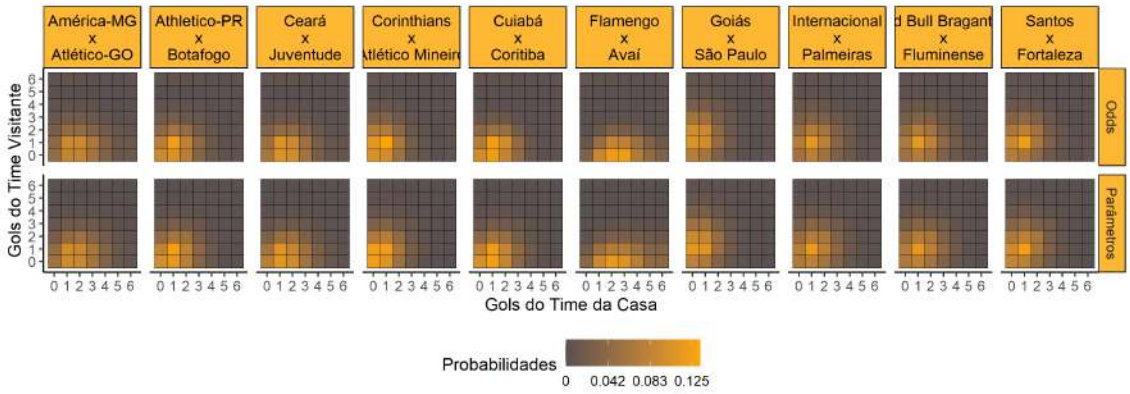

Figura 7: Comparações entre algumas superfícies obtidas por meio dos odds para *Exact Score* e superfícies obtidas por meio dos parâmetros estimados para a respectiva casa de apostas para os últimos 10 jogos da Série A do Brasileirão de 2022, apenas para a casa de apostas 1xBet.

de Kullback-Leibler, o ajuste não pode se adequar perfeitamente ao mercado dos placares exatos e nem a nenhum mercado em particular, mas deve ser uma representação parcimoniosa de todos os mercados simultaneamente.

Para visualizar o que ocorre com todos os demais mercados, cada partição da superfície de probabilidade teve suas respectivas probabilidades somadas e fez-se o gráfico das partições obtidas por meio dos odds após normalização e por meio da distribuição Placar Ajustada.

Um jogo da Série A do Brasileirão foi escolhido aleatoriamente e a casa de apostas escolhida foi a 1xBet, por dispor de todos os mercados, o resultado está disposto na Figura 8 e evidencia o comportamento parcimonioso do ajuste da superfície ao modelar diversas partições da superfície.

Figura 8: Partições da superfície de probabilidade obtida após normalização dos odds e avaliação da distribuição Placar Ajustada, para o entre Botafogo e Atlético Mineiro pela 17ª rodada da Série A do Brasileirão de 2022.

6.4 Superfície Consensual

As estimativas dos parâmetros de cada casa de apostas para cada jogo foram agregadas por meio de uma média simples para cada parâmetro, isto possibilita uma interpretação de que a superfície obtida por estes parâmetros médios representa o consenso de todas as casas de apostas disponíveis para um dado jogo.

A Figura 9 foi elaborada para compreender se a superfície consensual fornece estimativas calibradas da probabilidade dos placares. Se a superfície é bem calibrada, então espera-se que nos resultados em que a superfície consensual indica uma probabilidade de x% de fato ocorra x% das vezes, do contrário a superfície estaria descalibrada.

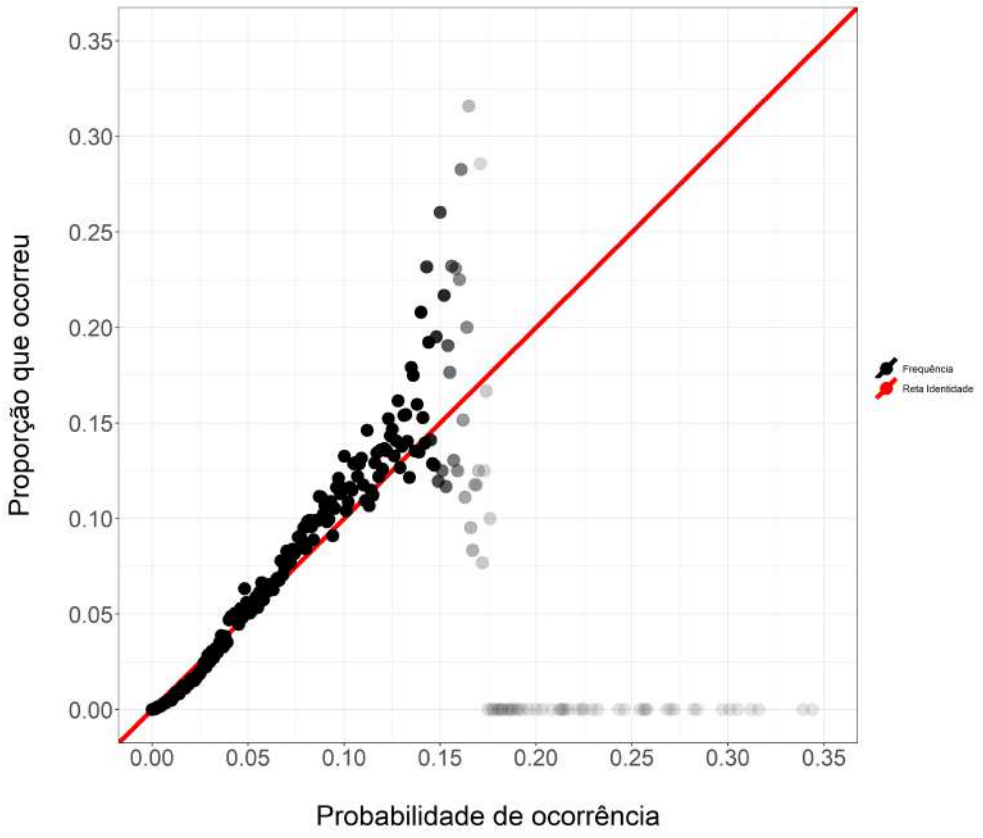

Figura 9: Gráfico de dispersão da calibração da superfície consensual, em que probabilidades com pouca frequência mudam na escala de transparência.

Da Figura 9 observa-se os pontos parecem próximos à reta identidade até cerca de 15% e em seguida começam a manifestar uma dispersão evidente. Entretanto, nos primeiros 15% pode-se perceber que resultados que foram apontados entre cerca de 1% a 3% apresentam uma leve subestimação em relação à proporção empírica, já entre 8% e 15% parece haver uma tendência de superestimação da proporção empírica, por conseguinte, é possível que um apostador munido desse conhecimento consiga se beneficiar ao considerar

que uma oportunidade de aposta com 13% de probabilidade segundo a casa de apostas pode estar pagando mais que deveria, enquanto uma aposta de 2% de probabilidade pode estar pagando menos que deveria, tendo em vista o mercado *Correct Score*.

Além dos 15% de probabilidade as estimativas ocorrem com pouca proporção, isto é, dificilmente a superfície consensual aponta que um dado placar apresenta 21% de probabilidade de ocorrência, por isso alguns pontos estão mais transparentes que outros na Figura 9.

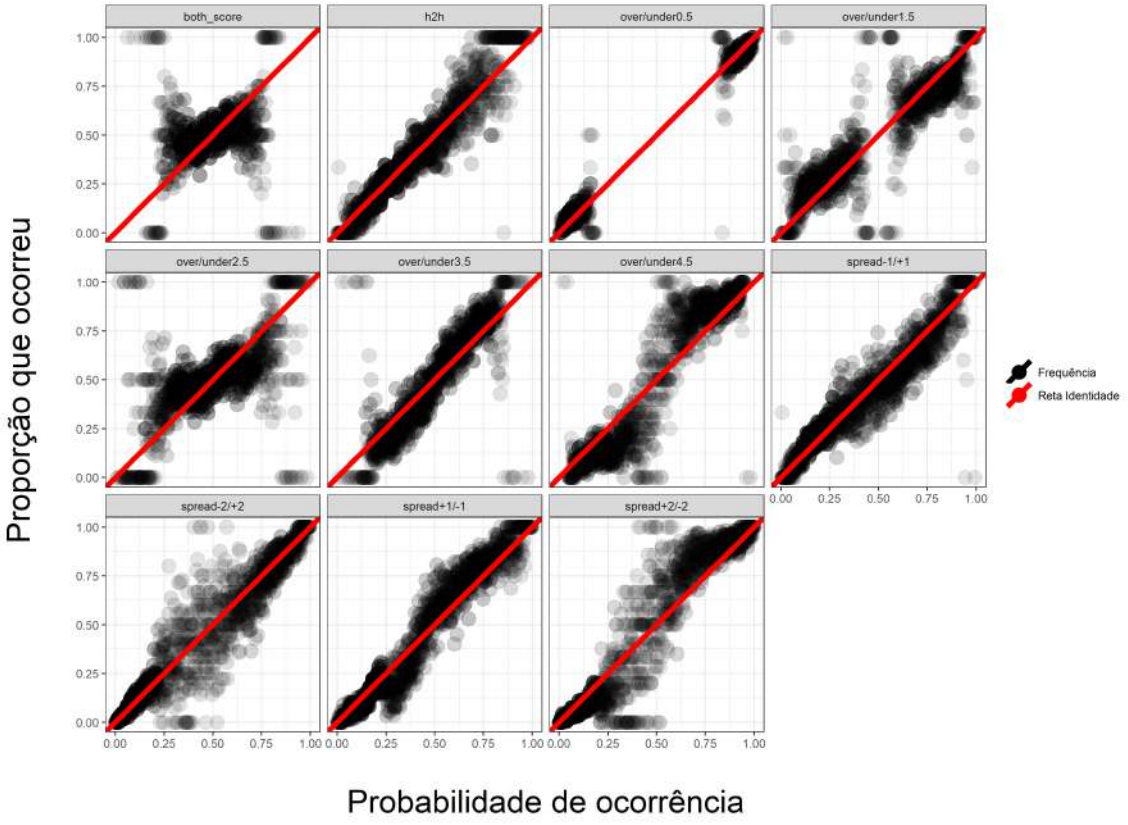

Figura 10: Gráficos de dispersão da calibração da superfície consensual considerando os mercados de aposta, em que probabilidades com pouca frequência mudam na escala de transparência.

De forma análoga, pode-se fazer a análise para a calibração dos demais mercados, ao agregar as probabilidades condizentes com as respectivas partições da superfície de probabilidade para cada jogo, o resultado está apresentado na Figura 10. Em específico, foi escolhido para a análise os mercados que tiveram melhor comportamento no gráfico de calibração, em geral, para mercados muito extremos como "*over/under* 6.5", "*spread* +6/-6", entre outros, as probabilidades se acumulam em valores muito extremos próximos de 0 ou 1. Para fins de exemplo dos gráficos com comportamento muito extremo, o gráfico

do "*over/under* 0.5" foi mantido na Figura 10.

Da Figura 10 é possível notar que alguns mercados apresentam regiões de maior descalibração, ou seja, regiões com oportunidades de apostas vantajosas, alguns gráficos estão forma de "S" ou tem inclinação diferente da reta identidade como *Both Score*, *Spread +2/-2*, entre outros. Aparentemente o mercado mais calibrado é o *Head to Head*, os pontos estão muito bem comportados em torno da reta identidade, exceto para valores próximos de 1.

6.5 Recalibração

Sob o intuito de recalibrar as probabilidades apresentadas pela superfície consensual na Figura 10 foi ajustada uma regressão isotônica em cada um dos mercados e submercados, cujas probabilidades obtidas estão dispostas na Figura 11.

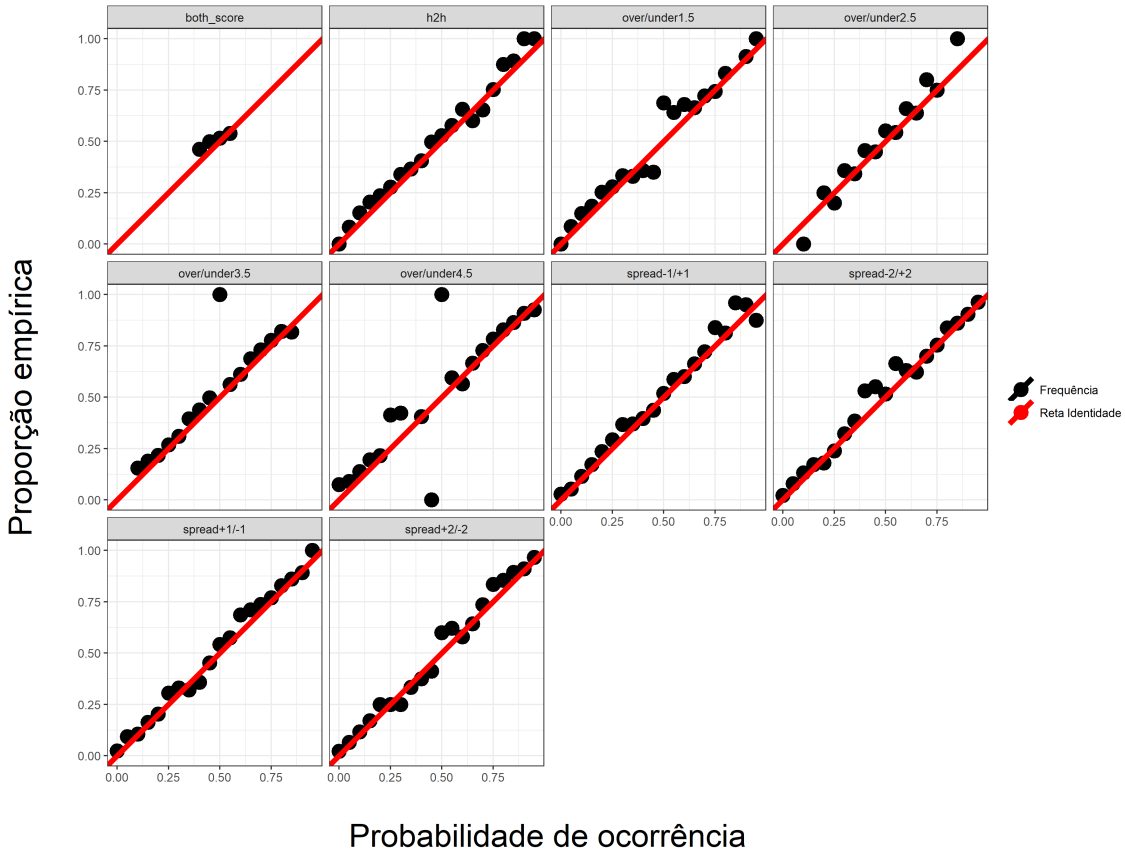

Figura 11: Gráficos de dispersão da calibração da regressão isotônica considerando os mercados de aposta.

A Figura 11 mostra que o comportamento das regressões isotônicas obtidas é de

probabilidades bem mais próximas às retas identidades para cada mercado, indicando que o ajuste foi bem sucedido em recalibrar as probabilidades da superfície consensual.

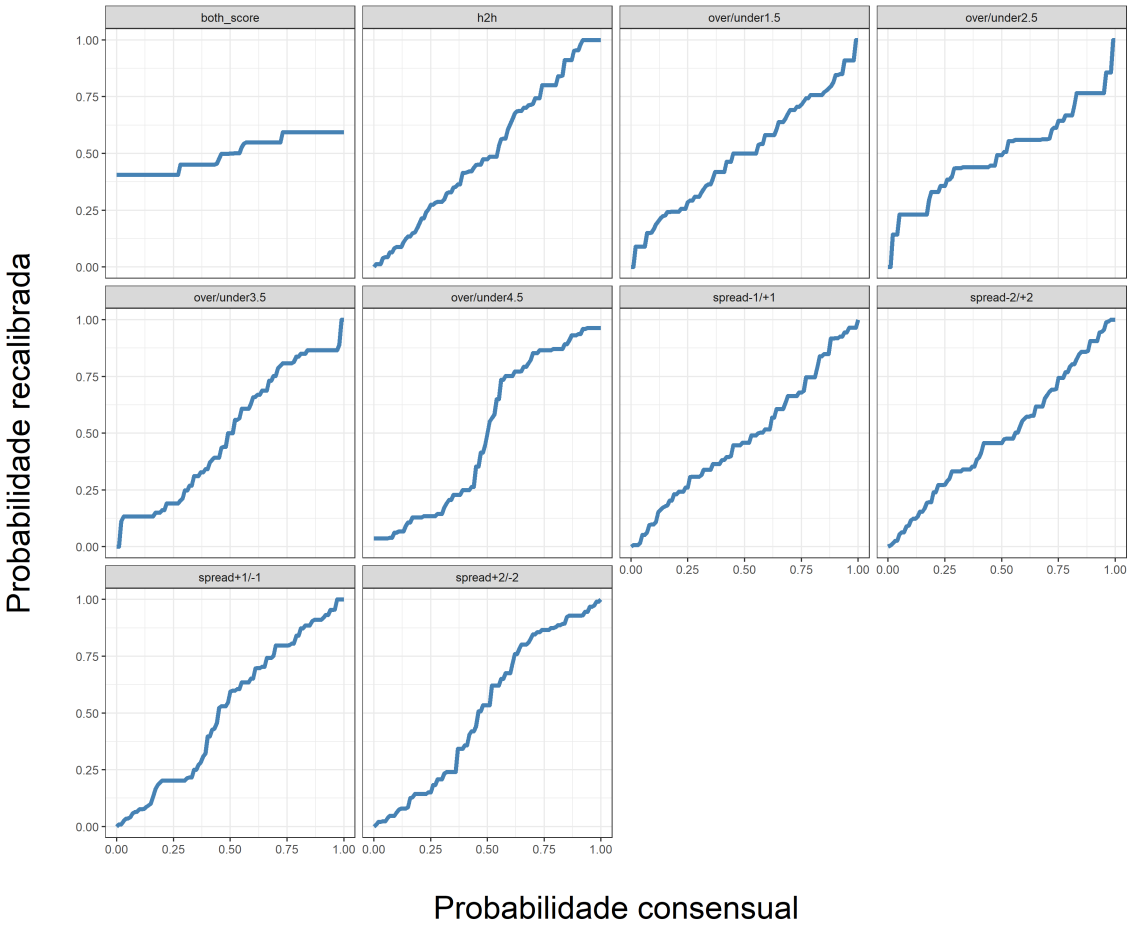

Figura 12: Gráficos de linhas da calibração via regressão isotônica sobre as probabilidades consensuais, considerando os mercados de aposta.

As probabilidades podem ser mapeadas como nas funções apresentadas pela Figura 12 para cada mercado, incluir esse mapeamento em um algoritmo permitiria automatizar a recalibração e possivelmente traria uma vantagem no momento da alocação das apostas.

6.6 Rede Neural

O ajuste de uma rede neural sobre as probabilidades obtidas pela superfície consensual não apresentaram melhorias notórias sobre a calibração do ajuste e tornaram as as probabilidades mais dispersas, entretanto, ainda seria possível melhorar o desempenho da rede ao adicionar novas informações sobre a partida para o treinamento da rede, portanto, seguiu-se com a elaboração de um script em Python para a coleta das buscas pelos times

nas 24h precedentes ao jogo pela API do Google Trends.

Todavia, os retornos obtidos a partir da API do Google Trends não foram satisfatórios, há diversos casos de inconsistências em que a busca não retorna nenhum valor pela API, mas realizando a busca pelo site obtém-se resultados. Em alguns casos uma alteração de 1h na busca poderia fazer com que o resultado que antes era vazio retornasse diferente.

O ajuste da rede neural sem as informações adicionais da partida não aparentava valer a pena em relação à sua complexidade e resultados obtidos, agregado ao fato de que as requisições à API do Google Trends foram frequentemente recusadas, mesmo a política de coleta automatizada permitindo-a, optou-se por cessar esse procedimento de coleta de mais dados e o ajuste da rede neural foi descartado do estudo.

7 Conclusão

Os resultados deste estudo apontam que a distribuição Placar Ajustada apresenta capacidade adequada de modelar as superfícies de probabilidade descritas pelos odds das casas de apostas em seus vários mercados. A métrica $\bar{D}_{KL}$ entre a superfície proposta e os odds obtidos a partir dos mercados de apostas parece ter sido um sucesso, devido à clara semelhança entre as superfícies propostas e as partições que vieram das casas de apostas.

As superfícies consensuais obtidas apresentaram casos de boa e má calibração nas probabilidades mais frequentes, a depender do mercado em análise. Os casos de má calibração apresentaram melhorias visualmente perceptíveis ao serem calibrados via regressão isotônica, indicando que, em média, as probabilidades obtidas ao normalizar as informações das casas de apostas e recalibrar as probabilidades consensuais podem apontar para probabilidades confiáveis. Embora que não tenha sido projetado um algoritmo capaz de identificar as casas de apostas com probabilidades mais confiáveis e nem utilizar uma medida de dispersão para agregar os valores na superfície consensual.

Esse estudo trás uma contribuição computacional ao propor a distribuição Placar, tem-se que a eficiência computacional foi mais de mil vezes maior em termos de tempo decorrido quando comparado com o processo gerador de dados da seção 5.3.2. Do mesmo modo que a distribuição Placar foi obtida em termos da Normal Bivariada, outras distribuições discretas provindas de cópulas também podem se beneficiar desse resultado e formalizar uma forma mais eficiente de calcular suas probabilidades.

Referências

AGCA, S. enay; CHANCE, D. M. Speed and accuracy comparison of bivariate normal distribution approximations for option pricing. 2003.

American Gaming Association. *2022 Commercial Gaming Revenue Tops $60B, Breaking Annual Record for Second Consecutive Year*. 2023. Disponível em: ⟨https://www.americangaming.org/new/2022-commercial-gaming-revenue-tops-60b-breaking-annual-record-for-second-consecutive-year/⟩.

Brasil. 1946. Disponível em: ⟨https://www.planalto.gov.br/ccivil_03/decreto-lei/del9215.htm⟩.

Brasil. 2018. Disponível em: ⟨https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13756.htm⟩.

CLARKE, S.; KOVALCHIK, S.; INGRAM, M. Adjusting Bookmaker's Odds to Allow for Overround. *American Journal of Sports Science*, v. 5, n. 6, p. 45, 2017. ISSN 2330-8559. Disponível em: ⟨http://www.sciencepublishinggroup.com/journal/paperinfo?journalid=155{&}doi=10.11648/j.ajss.20170506⟩.

DAN, P. States Where Sports Betting Is Legal. feb 2023. Disponível em: ⟨https://www.forbes.com/betting/guide/legal-states/#:~:text=states\%20considering\%20legalization.-,2018\%20U.S.\%20Supreme\%20Court's\%20Role\%20in\%20Sports\%20Betting,Protection\%20Act\%20(PASPA)\%20unconstitutional.⟩

DIXON, M. J.; COLES, S. G. Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Royal Statistical Society*, 1997.

DREZNER, Z.; WESOLOWSKY, G. O. On the computation of the bivariate normal integral. *Journal of Statistical Computation and Simulation*, Taylor Francis, v. 35, n. 1-2, p. 101–107, 1990. Disponível em: ⟨https://doi.org/10.1080/00949659008811236⟩.

GAUB, G. Prediction Of English Premier League Soccer Matches, Based On Player Data Using Supervised Learning. 2022. Disponível em: ⟨https://diglib.uibk.ac.at/ulbtirolhs/download/pdf/8282083?originalFilename=true⟩.

GIOLO, S. R. *Introdução á análise de dados categóricos com aplicações*. [S.l.]: Blucher, 2017.

iGaming Brazil. *Mapa do Patrocínio no futebol brasileiro é divulgado e aponta aumento de acordos com casas de apostas*. 2023. Disponível em: ⟨https://igamingbrazil.com/aposta-esportiva/2023/01/27/mapa-do-patrocinio-no-futebol-brasileiro-e-divulgado-e-aponta-aumento-de-acordos-com-casas-de-aposta⟩.

JOSEPH, L. D. Time series approaches to predict soccer match outcome. *National College of Ireland - School of Computing*, 2022.

KAIN, K. J.; LOGAN, T. D. Are Sports Betting Markets Prediction Markets? *Journal of Sports Economics*, v. 15, n. 1, p. 45–63, feb 2014. ISSN 1527-0025. Disponível em: ⟨http://journals.sagepub.com/doi/10.1177/1527002512437744⟩.

KARLIS, D.; NTZOUFRAS, I. Analysis of sports data by using bivariate poisson models. *The Statistician*, 2003.

KAUNITZ, L.; ZHONG, S.; KREINER, J. Beating the bookies with their own numbers - and how the online sports betting market is rigged. In: . [S.l.: s.n.], 2017.

KONING, R. H.; ZIJM, R. Betting market efficiency and prediction in binary choice models. *Annals of Operations Research*, v. 325, n. 1, p. 135–148, jun 2023. ISSN 0254-5330. Disponível em: ⟨https://link.springer.com/10.1007/s10479-022-04722-3⟩.

LU, C.-J. et al. Improving Sports Outcome Prediction Process Using Integrating Adaptive Weighted Features and Machine Learning Techniques. *Processes*, v. 9, n. 9, p. 1563, sep 2021. ISSN 2227-9717. Disponível em: ⟨https://www.mdpi.com/2227-9717/9/9/1563⟩.

MAHER, M. J. Modelling association football scores. *Statistica Neerlandica*, 1982.

MCHALE, I.; SCARF, P. Forecasting international soccer match results using bivariate discrete distributions. 01 2006.

ODACHOWSKI, K.; GREKOW, J. *Using bookmaker odds to predict the final result of football matches*. [S.l.: s.n.], 2013.

RAHMAN, M. H. A. A. et al. Bayesian approach to classification of football match outcome. *International Journal of Integrated Engineering*, 2018.

RAZALI, N. et al. Predicting football matches results using bayesian networks for english premier league. *IOP Conference Series: Materials Science and Engineering*, 2017.

ROBITZSCH, A. *pbv: Probabilities for Bivariate Normal Distribution*. [S.l.], 2020. R package version 0.4-22. Disponível em: ⟨https://CRAN.R-project.org/package=pbv⟩.

SAMBA, S. Football result prediction by deep learning algorithms. *Tilburg University*, 2019.

SANGANI, R. A Comprehensive Guide on Model Calibration: What, When, and How. *Towards Data Science*, 2022.

SHIN, H. S. Prices of State Contingent Claims with Insider Traders, and the Favourite-Longshot Bias. *The Economic Journal*, v. 102, n. 411, p. 426–435, 03 1992. ISSN 0013-0133. Disponível em: ⟨https://doi.org/10.2307/2234526⟩.

SONG, C.; BOULIER, B. L.; STEKLER, H. O. The comparative accuracy of judgmental and model forecasts of American football games. *International Journal of Forecasting*, v. 23, n. 3, p. 405–413, jul 2007. ISSN 01692070. Disponível em: ⟨https://linkinghub.elsevier.com/retrieve/pii/S0169207007000672⟩.

YIANNAKIS, A. et al. Forecasting in sport: The power of social context — a time series analysis with english premier league soccer. *International Review for the Sociology of Sport*, v. 41, n. 1, p. 89–115, 2006. Disponível em: ⟨https://doi.org/10.1177/1012690206063508⟩.