Previsões de Resultados em Partidas do Campeonato Brasileiro de Futebol

Joao Marcos Amorim dos Santos

Fundação Getúlio Vargas
Escola de Matemática
Programa de Pós-Graduação em Matemática aplicada

Rio de Janeiro
2019

Joao Marcos Amorim dos Santos

Previsões de Resultados em Partidas do Campeonato Brasileiro de Futebol

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Matemática Aplicada da FGV, como parte dos requisitos para a obtenção do título de Mestre em Modelagem Matemática.

Área de concentração: Matemática aplicada

Orientador: Moacyr Alvim Horta Barbosa da Silva
Coorientador: Rodrigo dos Santos Targino

Rio de Janeiro
2019

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pelo Sistema de Bibliotecas/FGV

Santos, João Marcos Amorim dos
Previsões de resultados em partidas do Campeonato Brasileiro de Futebol / João Marcos Amorim dos Santos. – 2019.
74 f.

Dissertação (mestrado) -Fundação Getulio Vargas, Escola de Matemática Aplicada.
Orientador: Moacyr Alvim Horta Barbosa da Silva.
Coorientador: Rodrigo dos Santos Targino
Inclui bibliografia.

1. Campeonato Brasileiro (Futebol) – Métodos estatísticos. 2. Futebol – Brasil - Modelos matemáticos. 3. Futebol – Brasil – Métodos estatísticos. I. Silva, Moacyr Alvim Horta Barbosa da. II. Targino, Rodrigo dos Santos. III. Fundação Getulio Vargas. Escola de Matemática Aplicada. IV. Título.

CDD – 796.334015195

Elaborada por Márcia Nunes Bacha – CRB-7/4403

JOÃO MARCOS AMORIM DOS SANTOS

"PREVISÕES DE RESULTADOS EM PARTIDAS DO CAMPEONATO BRASILEIRO DE FUTEBOL".

Dissertação apresentado(a) ao Curso de Mestrado em Modelagem Matemática do(a) Escola de Matemática Aplicada para obtenção do grau de Mestre(a) em Modelagem Matemática.

Data da defesa: 29/04/2019

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

Moacyr Alvim Horta Barbosa da Silva
Orientador(a)

Rodrigo dos Santos Targino
Membro Interno

Paulo Cezar Pinto Carvalho
Membro Interno

Samy Dana
Membro Externo

Ainda que a minha mente e o meu corpo enfraqueçam, Deus é a minha força, Ele é tudo o que eu sempre preciso.
Salmos 73:26

Agradecimentos

Agradeço à Deus, pois até aqui me ajudou o Senhor;

aos meus pais, Marcos Antônio e Sandra Amorim, pois me deram todo amor e suporte para chegar até aqui e ir além;

à minha irmã, Juliana Amorim, pois sempre me amou e me incentivou;

a todos meus parentes e amigos, pois cada um deles soube compreender minhas ausências devido aos meus estudos, em especial a Mariana Neves, pois me levantava sempre que eu desanimava;

a cada um dos amigos que fiz no mestrado. Em especial, ao "Sombra", "Aranha", "Fexu", "Impa", Gabriel Jardim e Thiago Trabach . Cada um deles tornou o mestrado mais leve e divertido;

aos meus amigos que fiz durante a faculdade, Arthur Schilithz, Fernando Alencar, Igor Pinto, mesmo estando longe continuamos nos ajudando e mantendo nossa jogatina viva;

Ao meu orientador, Moacyr Alvim, pois além de professor foi um amigo e incentivador;

Ao meu co-orientador, Rodrigo Targino, por cada minuto dedicado a tirar minhas dúvidas e me forçar a pensar.

aos professores e amigos que fiz no projeto Esporte em números, Paulo César, Walter Sande, Asla Sá, Antônio Neto e Rodrigo Escorcio. Pois desse projeto surgiu a motivação do tema da minha dissertação;

a todos professores da Escola de Matemática Aplicada (EMAP) - FGV, pois tiveram atenção e dedicação em passar seus conhecimentos, muitos deles sendo muito mais que professores;

à EMAP - FGV, pela bolsa de estudo e infraestrutura oferecida durante o mestrado;

Aos meus professores do jardim, ensino fundamental, ensino médio, faculdade, mestrado, pois sem os professores nenhum país iria pra frente;

a tantas outras pessoas que mesmo indiretamente fizeram parte deste momento.

"O Brasil ficou entre os 8 melhores do mundo no futebol e ficou triste. É 85º em educação e não há tristeza".

Cristóvam Buarque

Resumo

Prever resultados de partidas de futebol é um problema que vem sendo explorado há décadas. Tais resultados podem ser vistos por dois pontos de vista, prever o placar ou apenas prever o resultado: vitória, empate ou derrota. Quando se tem modelos que buscam prever a quantidade de gols marcados por cada uma das equipes, ambos pontos de vista do resultado de uma partida podem ser contemplados, placar e resultado.

Desde 1950, diferentes abordagens e modelos foram propostos com intuito de modelar a quantidade de gols marcadas por cada time em uma partida. Uma das abordagens mais exploradas foi caracterizar a quantidade de gols marcados por cada uma das equipes como uma variável que segue a distribuição de Poisson. Desde os primeiros trabalhos, uma hipótese trabalhada foi que a quantidade de gols marcados pelo time mandante e visitante seriam independentes. Porém, alguns autores utilizaram abordagens que consideram correlação no placar das duas equipes, sejam elas através do uso da Poisson Bivariada ou da adaptação do modelo independente. Contudo, a grande maioria desses trabalhos esteve limitada a usar como informação apenas os times participantes das partidas e a quantidade de gols marcados e sofridos por cada uma delas.

Esta dissertação tem por objetivo explorar a capacidade preditiva de diferentes modelos de Poisson propostos na literatura para prever a quantidade de gols marcados por cada uma das equipes em uma partida, além de fazer uso de mais variáveis explicativas, tais como número de finalizações, número de finalizações certas, roubadas de bola, variáveis essas provenientes do Cartola FC. Cada um dos modelos explorados é analisado tanto do ponto de vista de acertar o verdadeiro placar da partida quanto acertar o verdadeiro resultado da partida, vitória, empate ou derrota.

Palavras-chave: Modelagem de gols, Distribuição de Poisson, Poisson Bivariada, Distribuição Binomial, Campeonato Brasileiro, Regressão de Poisson..

Abstract

predicting football (soccer) results is a problem that has been explored for decades. The results can be seen from two points of view, predict the score or just to predict the result: win, draw or defeat. When we modeling the number of goals from each team, both points of view can be contemplated, score and result.

Since 1950, many approaches have been proposed in order to model the number of goals scored by each team in a match. One of the most explored approaches considers the number of goals scored by each team as a variable following a Poisson distribution. From the first works, a underlying hypothesis was that the number of goals scored by the home team and away team was independent. However, some authors have used approaches that consider correlation in the score of the two teams, either through the use of Bivariate Poisson or the adaptation of the independent model. However, the vast majority of these works were limited to the data about the teams playing the matches and the number of goals scored and concede only.

This thesis aims to explore the predictive capacity of different Poisson models proposed in the literature to predict the number of goals scored by each of the teams in a match, in addition to making use of more explanatory variables, such as number of shots, number of shots on target, tackles, all those variables coming from Cartola FC. Each one of the explored models was analyzed from the point of view to correct the true scoreboard of the game, as well as to correct the true result of the match, win, draw or defeat.

Keywords: Goal modeling, Poisson Distribution, Bivariate Poisson, Binomial Distribution, Brasileirão, Poisson Regression..

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

1.1 O futebol

O futebol é um dos esportes mais praticados mundialmente (RUSSELL, 2017). Com o passar dos anos o esporte vem atraindo cada vez mais adeptos e apreciadores. Tal magnitude pode ser observada quando se fala dos valores envolvidos anualmente em apostas esportivas, investimentos em contratações e estrutura por parte dos clubes, patrocínios aos clubes futebolísticos, e a entidade máxima do futebol, FIFA, tendo lucros bilionários mesmo nos momentos de recessão econômica (SPORT, 2017). Um dos fatores principais que influenciam tais investimentos é o quão bem um clube pretende ir em um campeonato, resumido por suas vitórias e conquistas. As vitórias atraem patrocinadores, torcedores, audiência, aumentando a receita do clube. No entanto, muitos clubes mesmo gastando quantias elevadas se comparadas aos seus adversários, nem sempre tem um bom desempenho durante um campeonato.

Dado todo investimento feito pelos clubes para disputa de campeonatos, os clubes relacionam tais quantias ao objetivo final em cada campeonato, seja um título, uma classificação para competição internacional ou manter-se em alguma competição. Os pontos necessários para se alcançar um título, se classificar para competições internacionais ou não ser rebaixado seguem uma distribuição gaussiana (ARTUSO, 2008), e os clubes buscam prever o desempenho necessário para alcançar tais metas. Contudo gerar tais previsões é um grande problema, devido a quase imprevisibilidade dos jogos de futebol, em que um clube que é considerado favorito pode perder para um clube considerado inferior.

Diante da necessidade por modelos com bom desempenho preditivo pelos clubes, mídia esportiva, torcedores e apostadores, estudos de tais modelos vêm sendo desenvolvidos há décadas. Alguns modelos estimam as chances de vitória de cada equipe se baseando na distribuição de gols dentro de uma partida, assumindo que tal evento segue uma distribuição de Poisson (MAHER, 1982) (LEE, 1997) (KARLIS; NTZOUFRAS, 2003).

Todavia as informações relevantes para tais modelos são apenas o desempenho do time nos jogos anteriores e a quantidade de gols sofridos e marcados. Neste trabalho, busca-se através do avanço computacional e maior quantidade de dados, utilizar modelos que levem em consideração outros fatores para prever tais resultados levando em consideração outras características como o número de finalizações, faltas cometidas e roubadas de bola. Os modelos propostos por (MAHER, 1982) e (KARLIS; NTZOUFRAS, 2003) não levam em consideração tais tipos de variáveis. Os dados considerados neste trabalho são do campeonato brasileiro de futebol de 2014 a 2018 e do Cartola FC dos respectivos anos.

Além de explorar a modelagem fazendo uso de outras informações referentes à partida, o

presente trabalho pretende avaliar diferentes tipos de modelagem de gols em uma partida. Os modelos analisados são: o modelo proposto por (LEE, 1997), que utiliza apenas os gols marcados e sofridos por cada equipe, sendo modelado por duas Poisson independentes; o modelo proposto por (KARLIS; NTZOUFRAS, 2003), que utiliza Poisson Bivariada, considerando uma correlação entre os gols marcados pelas equipes; o modelo proposto por (DIXON; COLES, 1997) que utiliza o modelo proposto por (LEE, 1997), mas ele atribui maior probabilidade para os placares com menos de 2 gols; e o modelo proposto por (STENERUD, 2015), que modela as chances de gol de cada equipe através da distribuição de Poisson e em seguida modela os gols pela distribuição Binomial condicionada as chances criadas e a taxa de conversão de chance em gol de cada equipe.

1.2 Campeonato brasileiro de futebol

O campeonato brasileiro de futebol, também conhecido como campeonato brasileiro ou Brasileirão, é a principal liga de futebol profissional entre clubes do Brasil. O Brasileirão atualmente é organizado pela Confederação Brasileira de Futebol (CBF). Esse campeonato passou por diversas mudanças em seu formato desde sua primeira edição (1959), mudando desde o sistema de disputa, assim como as regras e o número de participantes. Dentre os vários formatos já adotados incluem-se sistema eliminatório (1959-1968) e sistemas mistos de grupos (1967-2002). A fórmula de disputa do campeonato foi padronizada somente em 2003, quando foi adotado o sistema de pontos corridos com todas as equipes se enfrentando em turno e returno.

O regulamento atual do Brasileirão continua sendo o de pontos corridos, sendo disputado por vinte clubes. A temporada ocorre de maio a dezembro, cada clube joga duas vezes contra os outros dezenove, uma vez em seu estádio e a outra no de seu adversário, em um total de 38 jogos por equipe. As equipes recebem três pontos por vitória, um por empate e nenhum em caso de derrota. (CBF, 2018)

As equipes são classificadas pelos pontos acumulados, número de vitórias, saldo de gols e, em seguida, pelos gols marcados. Em caso de empate entre dois ou mais clubes, os critérios de desempate são os seguintes: maior número de vitórias; maior saldo de gols; maior número de gols pró; confronto direto; menor número de cartões vermelhos recebidos; menor número de cartões amarelos recebidos. (CBF, 2018)

O primeiro colocado ao final do campeonato conquista o título. Os seis times que mais pontuaram ao final do campeonato se qualificam para a Copa Libertadores, sendo os quatro melhores se classificando para a fase de grupos da competição, e o quinto e sexto colocado participam de uma fase preliminar a fase de grupos. Os outros seis melhores colocados eliminando os classificados para a Libertadores, se classificam para o Copa Sul-Americana. No caso de os vencedores da Copa do Brasil, da Copa Libertadores e/ou da

Copa Sul-Americana estiverem na zona de classificação, aquele lugar vai para a próxima equipe melhor colocada no campeonato. As quatro últimas equipes na classificação do campeonato são rebaixadas a série B. (CBF, 2018)

1.3 Cartola FC

O Cartola FC®é um *fantasy* game em que a cada rodada os participantes escalam seus times com os jogadores reais do Brasileirão. Participantes do game (cartoleiros) criam seus times e a cada rodada do campeonato brasileiro escalam seus times no cartola, escolhendo uma escalação, 3-5-2, 3-4-3, 4-4-2, 4-3-3, 4-5-1, 5-3-2 e, em seguida, escolhendo seus respectivos jogadores de cada posição, mais o treinador. As opções de escolhas são os jogadores do campeonato brasileiro, sendo as principais limitações na escolha de um jogador seu valor em cartoletas e sua posição uma vez que um jogador que seja um atacante não pode ser escalado como zagueiro. Um jogador só pode ser escalado em sua posição original em seu clube. Cada time criado no Cartola começa com $100 cartoletas, moeda do jogo.

Durante cada rodada do campeonato os jogadores pontuam de acordo com suas estatísticas na partida. Cada tipo de estatística tem uma determinada pontuação, sendo positiva ou negativa, por exemplo, fazer gol é um fator positivo na pontuação e ser expulso um fator negativo. Um jogador pode terminar a rodada com pontuação positiva ou negativa. Estas estatísticas também são chamadas de scouts dentro do jogo. Os scouts e suas respectivas pontuações são:

Positivos (+): roubada de bola (1,7); gol (8); assistência (5); jogo sem sofrer gol (5); falta sofrida (0,5); finalização para fora (0,7); finalização defendida (1); finalização na trave (3,5); defesa difícil (3), defesa de pênalti (7).

Negativos (-): gol contra (6); cartão vermelho (5); cartão amarelo (2); gol sofrido (5); pênalti perdido (3,5); falta cometida (0,5); impedimento (0,5); passe errado (0,3).

1.4 Objetivos

Este trabalho tem o objetivo de explorar a construção de modelos preditivos; a fim de prever o resultado de partidas do campeonato brasileiro. Dentre os modelos explorados encontram-se: Modelos Poisson independente, Poisson Dixon e Coles, Binomial-Poisson e Poisson Bivariado. Cada um dos modelos é utilizado para prever o número de gols feito por cada equipe em uma partida, em seguida prever a probabilidade de vitória, empate e derrota do time mandante. Além de criar os modelos utilizando apenas as informações dos resultados passados, o trabalho visa explorar modelos que utilizem informações provenientes do Cartola FC e comparar a capacidade preditiva dos diferentes modelos.

Os modelos analisados serão avaliados por 4 medidas; *Rank Probability Score*, medida de De Finetti, proporção de acertos e taxa de acerto de placar, medidas essas que serão explicada mais adiante.

1.5 Estrutura do trabalho

Este trabalho está dividido na seguinte estrutura: No capítulo 1 é apresentado a introdução, contextualização do problema abordado no trabalho e são apresentados os objetivos do trabalho. O capítulo 2 apresenta uma breve revisão da literatura e dos modelos já explorados na finalidade de previsão de resultados de partida de futebol. Também é apresentado o referencial teórico para construção dos modelos e das métricas. São apresentadas as distribuições pertinente a cada um dos modelos apresentados. O capítulo 3 apresenta os resultados da análise de cada uma das bases, e o resultado apresentado por cada modelo. O capítulo 4 apresenta as conclusões e considerações finais do trabalho.

2 Materiais e Métodos

2.1 Trabalhos Correlatos

Há décadas, o futebol vem atraindo o público pela sua emoção e imprevisibilidade. Tal imprevisibilidade atraiu olhares de estudiosos que tentaram criar modelos preditivos e entender os diversos fatores envolvidos em uma partida de futebol. Duas vertentes têm sido usadas para prever resultado em partidas de futebol: A primeira, modelando os gols marcados e sofridos por cada time; a segunda, modelando diretamente os resultados de vitória, empate e derrota. Uma semelhança em ambas abordagens são os fatores (variáveis independentes) utilizados nos modelos. Três fatores foram amplamente explorados, sendo eles: mando de campo, poder ofensivo e defensivo de cada time (LEE, 1997; MAHER, 1982; KARLIS; NTZOUFRAS, 2003; FARIAS, 2008). Entretanto muitos outros autores exploraram outros fatores em seus respectivos modelos, além de diferentes abordagem.

Alguns dos autores exploraram a vantagem do mando de campo em diferentes ligas e esportes. Alguns autores (COURNEYA; CARRON, 1992) definem vantagem de jogar em casa como o termo utilizado para descrever que os times que jogam em casa tendem a ganhar mais de 50% dos jogos jogando em casa. Já (POLLARD, 1986) observou que a vantagem de jogar em casa foi estabelecida para todos os principais esportes de equipe profissional na Inglaterra e na América do Norte. A vantagem foi maior no futebol, com a equipe da casa obtendo cerca de 64 % dos pontos ganhos na Liga Inglesa de Futebol nos anos de 1981 a 1984. Em (NEVILL; NEWELL; GALE, 1996) observou-se que a vantagem de jogar em casa estava presente nas oito principais divisões das ligas de futebol inglesas e escocesas de 1992-1993. E a vantagem de jogar em casa foi significante nas oito divisões estudadas.

Sabendo da influência do mando de campo nos resultados, pesquisadores usaram esse fato nos modelos preditivos. (MAHER, 1982) analisa as 4 divisões do campeonato inglês de 1971, 1972 e 1973, considerando que os gols marcado pelos times em uma partida seguem duas Poisson independentes, com médias que refletem o poder de ataque e defesa de cada um dos dois times. Além desse modelo foi proposto o uso da Poisson bivariada considerando uma possível correlação entre os gols marcados por cada equipe. (LEE, 1997) também considerou os gols de uma partida seguindo a distribuição de Poisson e havendo independência entre os gols marcados pelo time da casa e pelo time visitante. Diante disso, ele usou modelos lineares generalizados de Poisson para gerar previsão para a *Barclays Premier League* de 1996/1997 considerando no modelo, poder de ataque, defesa e mando de campo. (DIXON; COLES, 1997) realiza uma modelagem de Poisson semelhante aos dois trabalhos anteriores, mas inclui outras informações não consideradas nos anteriores, tais como decaimento de importância dos jogos mais antigos, retrospecto dos últimos

jogos, além que os parâmetros de força das equipes se alteram no decorrer da temporada.

Usando uma abordagem um pouco diferente da apresentada acima alguns autores utilizaram prioris bayesianas para estimação dos parâmetros de força das equipes. (SARAIVA et al., 2016) usa a abordagem com Poisson independente, usando priori bayesianas para estimação dos parâmetros do modelo para a *Barclays Premier League* de 2012-2013 e para o Brasileirão de 2015. (OLIVIERI FILHO et al., 2017) utiliza o mesmo modelo proposto por (SARAIVA et al., 2016), para a *Barclays Premier League* de 2012-2013, modificando apenas as prioris dos parâmetros. (FARIAS, 2008) faz uso de modelagem bayesiana dinâmica, permitindo atualização dos parâmetros por fatores auto-regressivos e compara com modelos dinâmicos proposto por (LEE, 1997), utilizando os dados do Campeonato Brasileiro de 2008. (STENERUD, 2015) utiliza a modelagem Poisson para as chances de gol (finalizações no gol) criadas por cada equipe e modela os gols feitos por cada equipe baseado na distribuição binomial condicionada as chances estimadas e a um parâmetro de conversão de chances em gol. Além disso, verifica diferentes efeitos relacionados a criação de chances de gol, tais como faltas, cartões amarelo, cartões vermelhos e escanteios.

Uma outra abordagem da distribuição Poisson foi assumida por alguns autores, a Poisson Bivariada (PB). (ARRUDA, 2000) propõe o uso do modelo de Poisson Bivariada para modelagem dos placares de uma partida, os dados foram do Campeonato Brasileiro de 1998 e Copa Rio-São Paulo de 99. Foram utilizados fatores de ataque, defesa e se as equipes eram do mesmo Estado. Além disso, ele compara uso de 4 modelos, o proposto por (LEE, 1997) utilizando duas Poisson independentes e utilizando Poisson Bivariada, fazendo uso destes dois modelos com estimação baseada apenas nos dados e com estimação através de Soma e Diferença (SD) e uma outra versão dos mesmos modelo mas baseando-se em uma priori bayesiana. (SUZUKI et al., 2007) dá continuidade no modelo proposto por (ARRUDA, 2000), diferenciando seu trabalho pela inclusão da variável de incidência de crise na equipe, e por prioris bayesianas diferentes, além de observar se o fator casa melhoraria a predição, utilizando os dados do Campeonato Brasileiro de 2005 e 2006. (SILVA et al., 2014) faz um comparativo entre o modelo Poisson Bivariado (PB) e Poisson Duplo (DP) para o Campeonato Brasileiro de 2012, mas não faz uso de nenhuma priori bayesiana para o valor dos parâmetros e considera como fatores explicativos, força de ataque, defesa e efeito de jogar em casa. (KARLIS; NTZOUFRAS, 2003) também comparam o modelo Poisson Bivariado *vs* Poisson Duplo, além de fazer uma adaptação no modelo Poisson Bivariado, inflando a probabilidade de empate em 0-0 e 1-1. Os fatores utilizados foram, fator casa, força de ataque e defesa. (BAIO; BLANGIARDO, 2010) usa modelagem hierárquica bayesiana para prever resultados, fazendo também uso da Poisson Bivariada para modelagem de gols, e usando mistura para tratar o *over-shrinkage* gerado pela modelagem hierárquica bayesiana, e testa ambos modelos para o campeonato italiano de 2007-2008.

Usando uma abordagem diferente alguns autores deixaram à modelagem dos gols das partidas de lado para modelar a variável tricotômica resultado da partida, vitória mandante, empate ou vitória visitante. (ALVES et al., 2011) utiliza o modelo logístico multinomial para prever os resultados do campeonato brasileiro de 2007, utilizando como variáveis explicativas a força das equipes. (GODDARD, 2005) compara a capacidade dos modelos Poisson Bivariado *vs* a regressão logística multinomial da variável tricotômica, vitória, empate e derrota. (DINIZ et al., 2018) utiliza duas versões da modelagem multinomial de Dirichlet, baseando-se apenas no número de vitórias, empates e derrota de cada equipe como base de dados para prever a probabilidade de vitória, empate e derrota em cada partida, os dados foram os Campeonatos Brasileiro de 2006 a 2014.

A fim de medir a qualidade das previsões, e comparar os modelos, diferentes metodologias foram utilizadas, sendo mais amplamente utilizada, teste de razão de verossimilhança, medida De Finetti, taxa de acertos. (ARRUDA, 2000) (SUZUKI et al., 2007) comparam a medida de De Finetti dos seus modelos *vs* 2/3, considerando um bom modelo medida abaixo de 2/3. A medida De Finetti é 2/3 quando é atribuída probabilidade de 1/3 para vitória do mandante, 1/3 para o empate e 1/3 para vitória do visitante. A taxa de acertos é dada pela proporção de rodadas que a previsão estava correta, não tendo um valor genérico a se comparar, podendo apenas comparar com outros modelos.

2.2 Materiais

Para o presente trabalho foram consideradas duas bases de dados, uma com o histórico de resultados do campeonato brasileiro de 2014, 2015, 2016, 2017 e 2018, e uma base do cartola dos mesmos anos.

2.2.1 Base do Brasileirão

Desde 2006 o brasileirão é disputado por 20 equipes que se enfrentam em partidas de turno e returno, totalizando 19 rodadas no primeiro turno e 19 rodadas no segundo turno. Um ponto importante de salientar é que cada time enfrenta um determinado adversário apenas duas vezes dentro do mesmo campeonato. Para ilustrar tomemos o seguinte exemplo: o time A enfrenta o time B na quinta rodada do primeiro turno, então na quinta rodada do segundo turno o time B vai enfrentar o time A. Sendo assim os dois times não se enfrentam em mais nenhuma outra rodada do campeonato. A base de dados do campeonato brasileiro deste trabalho possui os seguintes dados:

- Data
- Rodada

- Time mandante
- Gols marcados pelo time mandante
- Time visitante
- Gols marcados pelo time visitante

2.2.2 Base do Cartola FC

O Cartola FC é um *fantasy* game em que a cada rodada os participantes escalam seus times com os jogadores reais do Brasileirão. Os jogadores pontuam no Cartola de acordo com suas estatísticas (*scouts)* em cada partida do campeonato, pontuando quando, fazem gol, sofrem falta, dão assistência, e pontuam negativamente quando, cometem falta, recebem cartão amarelo ou vermelho, fazem gol contra.

Para o presente trabalho foi coletada informações do *game* dos anos de 2014-2018. Cada base possui as seguintes informações:

- Posições
  - ID: ID da posição
  - Nome: Nome da posição
  - Abreviação: Abreviação da posição
- Status
  - ID: ID do status
  - Nome: Nome do status
- Clubes
  - ID: ID do clube
  - Nome: Nome do clube
  - Abreviacao: Abreviação do clube
  - Slug: Slug do clube
- Partidas
  - ID: ID da partida
  - Rodada: Rodada em que a partida ocorreu
  - CasaID: ID do clube mandante

  - VisitanteID: ID do clube visitante
  - PlacarCasa: Placar do clube mandante
  - PlacarVisitante: Placar do clube visitante
  - Resultado: Resultado final da partida [‘"Casa"‘, ‘"Visitante"‘, ‘"Empate"‘]

- Atletas

  - ID: ID do atleta
  - Apelido: Apelido do atleta
  - ClubeID: ID do clube do atleta
  - PosicaoID: ID da posição do atleta

- Scouts

  - Rodada: Rodada em que o scout ocorreu
  - ClubeID: ID do clube do atleta
  - AtletaID: ID do atleta
  - Participou: ‘TRUE‘ se o atleta participou do jogo, ‘FALSE‘ se ficou no banco
  - Pontos: Pontuação do atleta nesta rodada
  - PontosMedia: Média de pontos do atleta até está rodada (inclui rodada atual)
  - Preco: Preço do atleta nesta rodada
  - PrecoVariacao: Variação do preço da rodada passada para está
  - FS: Faltas sofridas
  - PE: Passes errados
  - A: Assistências
  - FT: Finalizações na trave
  - FD: Finalizações defendidas
  - FF: Finalizações para fora
  - G: Gols
  - I: Impedimentos
  - PP: Pênaltis perdidos
  - RB: Roubadas de bola
  - FC: Faltas cometidas
  - GC: Gols contras

- CA: Cartões Amarelos
- CV: Cartões Vermelhos
- SG: Jogo sem sofrer gols
- DD: Defesas difíceis
- DP: Defesa de pênaltis
- GS: Gols sofridos

2.3 Metodologia

Neste capítulo será apresentado parte do embasamento teórico necessário para construção do modelos propostos.

2.3.1 Distribuição Poisson

Uma das principais hipóteses para a construção dos modelos apresentados neste trabalho é modelar os gols de uma partida seguindo a distribuição de Poisson.

A distribuição de Poisson foi nomeada pelo matemático francês Siméon Denis Poisson. Ela é caracterizada por ser uma distribuição discreta que modela o número de ocorrências de um determinado evento em um espaço de tempo condicionado a uma taxa de ocorrência. Os eventos ocorrem com uma taxa constante e independente do último evento.

2.3.1.1 Propriedades da distribuição de Poisson

A distribuição de Poisson é dada pela seguinte fórmula:

$$P(\mathbf{X} = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad ; \quad x = 0, 1, 2, 3, \cdots \quad ; \quad \lambda \in \mathbb{R}_+ \tag{1}$$

$$E(\mathbf{X}) = \lambda \qquad ; \quad Var(\mathbf{X}) = \lambda$$

Um exemplo da contextualização do problema de estimação de gols marcados por um time A e a distribuição de Poisson é definir:

X = Número de gols marcados por um time A em 90 minutos de jogo.
Sabendo que a média de gols marcadas pelo time em 90 minutos é λ_A.
Então pode-se calcular a probabilidade do time A marcar x gols em 90 minutos, $x = 0, 1, \cdots$
Por exemplo, a probabilidade do time A marcar 3 gols seria,

$$P(\mathbf{X} = 3|\lambda_A) = \frac{e^{-\lambda_A}\lambda_A^3}{3!}$$

Considerando uma partida entre equipe A x B, e considerando o número de gols marcados pela equipe A independente do adversário, fica intuitivo perceber o cálculo da probabilidade de vitória do time A, de empate e de vitória do time B.

Seja Y = Número de gols marcados pelo time B em 90 minutos de jogo.
Sabendo que a média de gols marcadas pelo time B em 90 minutos é λ_B.
Considere X e Y independentes.

Segue que,

- Probabilidade de vitória do time A

$$P(X > Y|\lambda_A, \lambda_B) \tag{2}$$

- Probabilidade de empate

$$P(X = Y|\lambda_A, \lambda_B) \tag{3}$$

- Probabilidade de vitória do time B

$$P(Y > X|\lambda_A, \lambda_B) \tag{4}$$

Uma maneira de calcular tais probabilidades (2, 3, 4) é através da distribuição de Skellam, descoberta pelo estatístico John Gordon Skellam. A distribuição de Skellam é derivada da seguinte forma. Seja X uma variável aleatória de Poisson com média λ_1 e Y uma variável aleatória de Poisson com média λ_2. Sendo X e Y independentes, então a variável aleatória $Z = X - Y$ segue a distribuição de Skellam com parâmetros λ_1 e λ_2.

A distribuição de Skellan é dada pela seguinte fórmula:

$$P(\mathbf{W} = w|\lambda_1, \lambda_2) = e^{-(\lambda_1+\lambda_2)} \left(\frac{\lambda_1}{\lambda_2}\right)^{w/2} I_{|w|}\left(2\sqrt{\lambda_1\lambda_2}\right) \tag{5}$$

$$\mathbf{W} \in \mathbb{N} \quad ; \quad \lambda_1 \mathrm{e} \lambda_2 > 0$$

Em que$I_{|w|}(.)$ é a função de Bessel modificada do primeiro tipo.

2.3.2 Distribuição Poisson Bivariada

Quando tem-se um vetor bivariado de variáveis aleatórias (X,Y) com suporte $\in \mathbb{N}_0^2$, isso é, $N_0 = 0, 1, 2, \cdots$, diz-se que o vetor (X, Y) segue a distribuição de Poisson Bivariada quando:

1. $\sum_x \sum_y f_{X,Y}(x, y) = 1$;
2. $f_{X,Y}(x, y) \geq 0 \forall (x, y) \in \mathbb{N}_0$;

3. $f(X;\lambda_x) = \sum_y f_{X,Y}(x,y) = \dfrac{e^{-\lambda_x}\lambda_x^x}{x!}; \quad x \in \mathbb{N}_0; \quad \lambda_x > 0;$

4. $f(Y;\lambda_y) = \sum_x f_{X,Y}(x,y) = \dfrac{e^{-\lambda_y}\lambda_y^y}{y!}; \quad y \in \mathbb{N}_0; \quad \lambda_y > 0.$

Dentre as diferentes estruturas da distribuição Poisson Bivariada (PB) uma das que mais se destaca é a da classe de Holgate. A distribuição Poisson Bivariada da classe de Holgate foi proposta por Holgate em 1964. Ela é atribuída para os eventos em que se tem um vetor bivariado (X,Y), sendo X e Y duas variáveis aleatórias de Poisson. A construção da distribuição PB é derivada do caso em que X e Y não são independentes. Holgate utilizou o método de redução trivariada para construção desta distribuição, sendo construída da seguinte forma:

Considere as variáveis aleatórias Z_i, $i = 1,2,3$, que são distribuição de Poisson independentes com parâmetro $\lambda_i > 0$. Seja $X = Z_1 + Z_3$ e $Y = Z_2 + Z_3$, então,

$$\begin{aligned}
P(X=x, Y=y) &= P(Z_1+Z_3=x, Z_2+Z_3=y) \\
&= \sum_{i=0}^{\infty} P(Z_1+Z_3=x, Z_2+Z_3=y|Z_3=i)P(Z_3=i) \\
&= \sum_{i=0}^{min(x,y)} P(Z_1=x-i, Z_2=y-i|Z_3=i)P(Z_3=i) \\
&= \sum_{i=0}^{min(x,y)} P(Z_1=x-i, Z_2=y-i)P(Z_3=i) \\
&= \sum_{i=0}^{min(x,y)} P(Z_1=x-i)P(Z_2=y-i)P(Z_3=i) \\
&= \sum_{i=0}^{min(x,y)} \frac{e^{-\lambda_1}\lambda_1^{x-i}}{(x-i)!}\frac{e^{-\lambda_2}\lambda_2^{y-i}}{(y-i)!}\frac{e^{-\lambda_3}\lambda_3^{i}}{(i)!} \\
&= e^{-(\lambda_1+\lambda_2+\lambda_3)} \sum_{i=0}^{min(x,y)} \frac{\lambda_1^{x-i}}{(x-i)!}\frac{\lambda_2^{y-i}}{(y-i)!}\frac{\lambda_3^{i}}{(i)!} \qquad \text{☺}.
\end{aligned}$$

Dado o processo de construção, é fácil ver que:

$$\begin{aligned}
E[X] &= \lambda_1 + \lambda_3; \qquad E[Y] = \lambda_2 + \lambda_3 \\
Var[x] &= \lambda_1 + \lambda_3; \qquad Var[Y] = \lambda_2 + \lambda_3
\end{aligned}$$

e de fácil demonstração que $\text{cov}(X,Y) = \lambda_3$,

$$\begin{aligned}
\text{cov}(X,Y) &= \text{cov}(Z_1+Z_3, Z_2+Z_3) \\
&= \text{cov}(Z_1,Z_2) + \text{cov}(Z_1,Z_3) + \text{cov}(Z_3,Z_2) + \text{cov}(Z_3,Z_3) \\
&= \cancelto{0}{\text{cov}(Z_1,Z_2)} + \cancelto{0}{\text{cov}(Z_1,Z_3)} + \cancelto{0}{\text{cov}(Z_3,Z_2)} + \text{var}(Z_3) \\
&= \lambda_3.
\end{aligned}$$

Sabendo que o valor da $\text{cov}(X, Y) = \lambda_3$, pode-se observar um detalhe importante da PB da classe de Holgate, o fato da correlação entre as variáveis X e Y terem de ser não-negativa.

$$\begin{aligned}\rho_{X,Y} &= \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} \\ &= \frac{\lambda_3}{\sqrt{(\lambda_1+\lambda_3)(\lambda_2+\lambda_3)}}\end{aligned}$$

Outras classes da Poisson Bivariada foram desenvolvidas para lidar com o caso da correlação ser negativa, no entanto para o problema a ser tratado neste trabalho a PB da classe de Holgate é a mais adequada . Segundo (ARRUDA, 2000) 3 características tornam a classe de Holgate a mais adequada, são elas:

1. As distribuições marginais devem ser Poisson;
2. As distribuições conjunta deve possuir suporte pleno ao menos perto da origem;
3. A distribuição conjunta e as marginais devem ser infinitamente divisíveis.

Pode-se perceber que o caso em que $\lambda_3 = 0$ a distribuição PB recaí sobre a distribuição Poisson Dupla (PD), em que X e Y são independentes com parâmetros λ_1 e λ_2. No caso da PB também pode-se observar que a diferença das variáveis aleatórias X e Y também segue a distribuição de Skellan 5,

$$W = X - Y = Z_1 + Z3 - (Z_2 + Z_3) = Z_1 - Z_2$$

Uma das complicações a serem apresentadas para PB é a estimação dos parâmetros via função verossimilhança, necessitando de um algoritmo de otimização para solução.
A função verossimilhança da distribuição PB é dada pela seguinte fórmula:

$$\mathcal{L}\left((\lambda_1,\lambda_2,\lambda_3);(x,y)\right) = \prod_{i=1}^{n} e^{-(\lambda_1+\lambda_2+\lambda_3)} \sum_{i}^{min(x,y)} \frac{\lambda_1^{x-i}}{(x-i)!}\frac{\lambda_2^{y-i}}{(y-i)!}\frac{\lambda_3^{i}}{(i)!} \quad \therefore \tag{6}$$

$$ln\mathcal{L}\left((\lambda_1,\lambda_2,\lambda_3);(x,y)\right) = \sum_{i=1}^{n}\left[-(\lambda_1+\lambda_2+\lambda_3) + ln\left(\sum_{i}^{min(x,y)} \frac{\lambda_1^{x-i}}{(x-i)!}\frac{\lambda_2^{y-i}}{(y-i)!}\frac{\lambda_3^{i}}{(i)!}\right)\right] \tag{7}$$

Tal complicação é dada pelo cálculo da derivada da $\mathcal{L}\left((\lambda_1,\lambda_2,\lambda_3);(x,y)\right)$ em relação a cada um dos parâmetros. Diferentes autores propuseram diferentes métodos numéricos de obter a solução. Pode ser observado em (KAWAMURA, 1984) um desses métodos.

2.3.3 Distribuição Binomial

A distribuição binomal, $\text{Bin}(n, p)$, é uma distribuição de probabilidade discreta caracterizada por modelar o número de sucessos em n ensaios independentes com probabilidade p de sucesso.

Em contextualização com o objetivo de prever os gols marcador por uma equipe em uma rodada do campeonato, pode-se dizer que: Considerando X_i como o número de gols marcados por uma equipe na i-ésima partida, tal que $X_i \sim \text{Bin}(n, p)$, seja n o número de finalizações certas que um time faz no jogo e p a proporção de gols por finalizações certas na partida. Pode-se então calcular a probabilidade do time fazer x gols, para $x \in \mathbb{N}$.

2.3.4 Modelos Poisson Independente

Com o objetivo de prever o resultado de uma partida, ou seja, a quantidade de gols que cada time fará em uma partida, pode-se abordar tais previsões através da modelagem Poisson, baseando-se nas seguintes hipóteses; Em um campeonato com n times, tal como o campeonato brasileiro, seja $X_{im} \sim Poi(\lambda_{im})$ o número de gols que o time mandante faz na i-ésima partida e seja $Y_{iv} \sim Poi(\lambda_{iv})$ o número de gols que o time visitante faz na i-ésima partida, em que X e Y são independentes. Dado tais hipóteses e o fato da distribuição Poisson pertencer a família exponencial, pode-se modelar tal problema através de Modelos Lineares Generalizados (MLG), no caso, através do modelo log-linear de Poisson.

Com o intuito de modelar o número de gols marcados pelo time mandante (X_{im}) na i-ésima partida e os gols marcados pelo time visitante (Y_{iv}) é intuitivo associar os parâmetros (λ_{im} e λ_{iv}) através de fatores que estão relacionados com a quantidade de gols marcados, no caso, poder de ataque e defesa de ambos times. Utilizando a função de ligação logarítmica, tem se:

$$\lambda_i = e^{\boldsymbol{W_i \beta_i}} \tag{8}$$

$$\log(\lambda_i) = \mathbf{W_i} \boldsymbol{\beta_i} \tag{9}$$

Diante de tal relação podemos construir diferentes modelos com diferentes variáveis explicativas.

2.3.4.1 Modelo 1 - Lee

O modelo 1 baseia-se no modelo proposto por (LEE, 1997), em que as variáveis explicativas do modelo estão limitadas ao fator casa, gol marcados e gols sofridos. Para estimar o número de gols marcados pelo time mandante (m) contra o time visitante (v) na i-ésima partida tem se:

$$\log(\lambda_{mi}) = \beta_{0i} + \beta_{1mi} + \beta_{2vi} \tag{10}$$

O fator β_{0i} está associado ao efeito casa, β_{1mi} está associado ao poder de ataque do time mandante e β_{2vi} está associado ao poder defensivo do time visitante.

Para estimar o número de gols marcados pelo time visitante (v) contra o time mandante (m) na i-ésima partida tem se:

$$\log(\lambda_{vi}) = \beta_{1vi} + \beta_{2mi} \tag{11}$$

Como o time visitante não se beneficia do fator casa, o modelo não possui o fator β_{0i}. O β_{1vi} está associado ao poder de ataque do time visitante e β_{2mi} está associado ao poder defensivo do time mandante.

Como pode-se observar o modelo 1, não tem identificabilidade garantida. Para garantir tal identificabilidade faz-se necessário o uso de restrições no conjuntos dos parâmetros. Diferentes tipos de restrições podem ser utilizadas, tais como fixar o parâmetros de uma das equipes como 1, soma zero, dentre outras restrições. A equação 12 exemplifica o uso da restrição soma zero. Com o uso dela temos a identificabilidade do modelo garantida.

$$\sum_{j=1}^{n} \beta_{1ij} = 0; \qquad \sum_{j=1}^{n} \beta_{2ij} = 0 \tag{12}$$

2.3.4.2 Modelo 2 - Cartola

O modelo 2 possui estrutura semelhante àquela apresentada no modelo 1, diferindo em relação à quantidade de variáveis explicativas. Como o modelo 1 limita-se na quantidade de variáveis explicativas, sendo apenas os gols sofridos e marcados, o modelo 2 faz uso de mais variáveis explicativas, variáveis essas provenientes do Cartola FC. Diferentes variáveis do Cartola FC foram analisadas no intuito de observar quais geravam os melhores resultados nas previsões.

Para estimar o número de gols marcados pelo time mandante (m) contra o time visitante (v) na i-ésima partida tem-se:

$$\log(\lambda_{mi}) = \beta_{0i} + \beta_{1mi} + \beta_{2vi} + x_{1mi}\beta_{3mi} + x_{2mi}\beta_{4vi} \tag{13}$$

O fator β_{0i} está associado ao efeito casa, β_{1mi} está associado ao poder de ataque do time mandante, β_{2vi} está associado ao poder defensivo do time visitante, x_{1mi} é a quantidade de finalizações do time mandante na i-ésima rodada, β_{3mi} está associado a quantidade de finalizações do time mandante, x_{2vi} é a quantidade de roubadas de bola feitas pelo time visitante na i-ésima rodada, β_{4vi} está associado a quantidade de bolas roubadas pelo time visitante.

Para estimar o número de gols marcados pelo time visitante (v) contra o time mandante (m) na i-ésima partida tem se:

$$\log(\lambda_{vi}) = \beta_{1vi} + \beta_{2mi} + x_{1vi}\beta_{3vi} + x_{2mi}\beta_{4mi} \tag{14}$$

O β_{1vi} está associado ao poder de ataque do time visitante, β_{2mi} está associado ao poder defensivo do time mandante, x_{1vi} é a quantidade de finalizações do time visitante na i-ésima rodada, β_{3vi} está associado a quantidade de finalizações do time visitante, x_{2mi} é a quantidade de roubadas de bola feitas pelo time mandante na i-ésima rodada, β_{4mi} está associado a quantidade de bolas roubadas pelo time mandante.

Como pode ser observado, quando necessitarmos fazer uma previsão de uma rodada que ainda não ocorreu não teremos o número de finalizações e roubadas de bola para cada time. Para solucionar tal problema será considerado a média de finalizações e a média de roubada de bolas de cada um dos times até aquela rodada.

2.3.4.3 Modelo 3 - Cartola 2

O modelo 3 possui estrutura semelhante à estrutura do modelo 2, o ponto que diferencia ambos modelos é o fato do modelo 3 levar em consideração finalizações certas em vez do total de finalizações, todos os demais atributos são os mesmos já apresentados no modelo 2.

2.3.4.4 Modelo 4 - Cartola 3

O modelo 4 pode ser visto como um modelo aninhado do modelo 2, pois as variáveis explicativas utilizadas no modelo 4 são as mesmas do modelo 2, exceto pela exclusão da variável explicativa roubadas de bola. Sendo então explicado pelas variáveis time casa, time visitante, efeito casa e finalizações .

2.3.4.5 Modelo 5 - Cartola 4

O modelo 5 pode ser visto como um modelo aninhado do modelo 3, pois as variáveis explicativas utilizadas no modelo 5 são as mesmas do modelo 3, exceto pela exclusão da variável explicativa roubadas de bola. Sendo então explicado pelas variáveis time casa, time visitante, efeito casa e finalizações certas.

2.3.5 Modelo 6 - Binomial - Poisson

Proposto por (STENERUD, 2015), o modelo 6 também possui o intuito de prever gols marcados por cada um dos times em uma partida, porém diferencia-se dos modelos apresentados por (LEE, 1997) e (MAHER, 1982). O modelo 6 é feito em duas etapas, a primeira modela o número de finalizações certas de um time na partida e a segunda modela o número de gols marcados por cada time como uma distribuição binomial condicionada

as chances estimadas pela etapa anterior. Assim como o modelo proposto por (LEE, 1997) e (MAHER, 1982), que fazem uso da suposição que a variável resposta, gols marcados, seguem uma distribuição de Poisson, o modelo 6 também faz tal suposição, contudo a variável resposta aqui considerada é o número de finalizações certas de um time na partida. A modelagem do número de finalizações certas segue a mesma estrutura apresentada no modelo 1, diferenciando-se apenas no fato da variável resposta ser o número de finalizações certas. Apesar da hipótese do número de finalizações certas em uma partida seguir a distribuição de Poisson parecer grosseira, a figura 7 mostra que tal aproximação é bem razoável. O segundo parâmetro da distribuição binomial, p, que é traduzido como a taxa de conversão de finalizações certas em gol é definida pela função 15.

$$\hat{p}_{mi} = \frac{\sum_{j=1}^{i-1} g_{jmi}}{\sum_{j=1}^{i-1} FC_{jmi}} \tag{15}$$

Em que, $\sum_{j=1}^{i-1} g_{jmi}$ é o total de gols marcados pelo time mandante até a i-ésima partida, e $\sum_{j=1}^{i-1} FC_{jmi}$ é o total de finalizações certas do time mandante até i-ésima partida.

Portanto, temos que o número de chances criadas pela equipe mandante e visitante na i-ésima partida é dado por $F_{imv} \sim \text{Poi}(\lambda_{imv})$ e $F_{ivm} \sim \text{Poi}(\lambda_{ivm})$. Então o número estimado de gols marcados pela equipe mandante e visitante é dado por $\hat{X}_{imv} \sim \text{Bin}(\hat{F_{imv}}, \hat{p_{im}})$ e $\hat{Y}_{ivm} \sim \text{Bin}(\hat{F_{ivm}}, \hat{p_{iv}})$.

2.3.6 Modelo 7 - Dixon e Coles

(DIXON; COLES, 1997) propuseram uma adaptação do modelo proposto por (LEE, 1997; MAHER, 1982). Enquanto Lee e Maher consideram os gols marcados pela equipe mandante e visitante sendo independentes, Dixon e Coles fazem uma adaptação na distribuição de Poisson para considerar uma possível correlação para quantia de gols menores que dois. i.e, 0x0, 1x0, 0x1 e 1x1. A adaptação proposta pode ser vista na equação 16.

$$\tau_{\lambda,\mu}(x,y) == \begin{cases} 1-\lambda\mu\rho & \text{se} \quad x=y=0 \\ 1+\lambda\rho & \text{se} \quad x=0, y=1 \\ 1+\mu\rho & \text{se} \quad x=1, y=0 \\ 1-\rho & \text{se} \quad x=y=1 \\ 1 & \text{se} \quad \text{c.c} \end{cases}$$

$$\lambda = \exp(\alpha_i\beta_j\gamma)\ ; \qquad \mu = \exp(\alpha_j\beta_i)\ ; \qquad max(-1/\lambda, -1/\mu) \leq \rho \leq min(1/\lambda\mu, 1)$$

$$P(X=x, Y=y) = \tau_{\lambda,\mu}(x,y)\frac{\lambda^x exp(-\lambda)}{x!}\frac{\mu^y exp(-\mu)}{y!} \tag{16}$$

É fácil perceber que ao considerar $\rho = 0$ recaímos sobre o proposto por (LEE, 1997; MAHER, 1982)

Com o intuito de considerar os jogos mais recentes mais informativos, (DIXON; COLES, 1997) também propuseram uma função de esquecimento para os jogos mais antigos. Contudo, a função de esquecimento apresentada por Dixon e Coles não foi utilizada no modelo 7. A função de esquecimento utilizada no modelo 7 está apresentada na função 17.

$$W(a, f, k) = \left(\frac{a}{f}\right)^{\frac{1}{k}}; \tag{17}$$

Em que a é rodada que a partida foi jogada, f é última rodada antes da rodada prevista e k é um fator de escolha da importância dos jogos passados. Um exemplo pode ser visto na tabela 1 e no gráfico 1, em que o objetivo era prever os jogos da 20º rodada.

Tabela 1 – Exemplo da função de esquecimento

Rodada atual = 19	**Rodada jogada**						
$W(a, f, k)$	1	2	7	11	15	18	19
$W(a, 19, 2)$	0.229	0.324	0.607	0.761	0.889	0.973	1.0
$W(a, 19, 3)$	0.375	0.472	0.717	0.833	0.924	0.982	1.0
$W(a, 19, 6)$	0.612	0.687	0.847	0.913	0.961	0.991	1.0

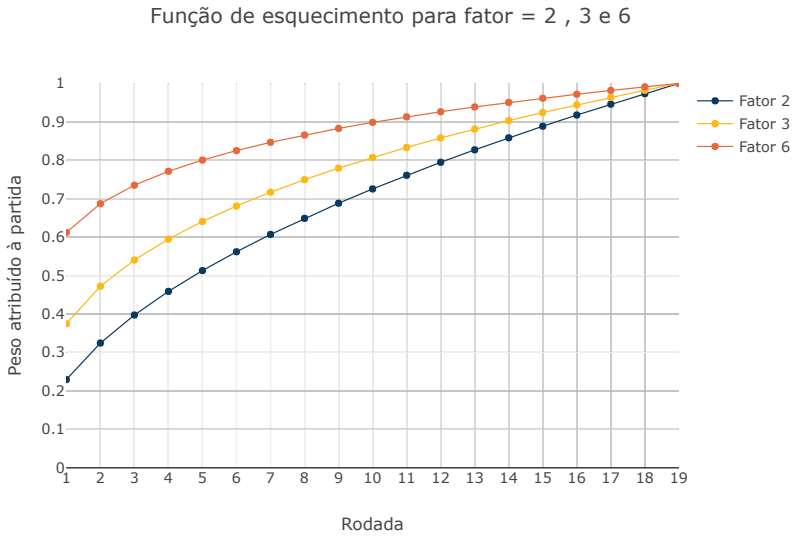

Figura 1 – Função de esquecimento na 19º rodada

Como pode ser observado na tabela 1 e no gráfico 1, quanto maior o fator k, maior será a importância dada aos jogos mais antigos.

As variáveis explicativas utilizadas no modelo 7 são as mesmas já apresentadas no modelo 1, a principal diferença entre os dois modelos é o fato da adaptação da distribuição

Poisson e a função esquecimento 17. Considerando as adaptações feitas no modelo proposto por Dixon e Coles, temos que a função verossimilhança para os parâmetros a serem estimados é dada pela função 18. Um fato importante a ser notado é que o modelo de Dixon e Coles tem um parâmetro a mais a ser estimado, no caso o ρ.

$$\mathcal{L}(\theta|x,y,a,f,k) = \left(\tau_{\lambda,\mu}(x,y)\frac{\lambda^x exp(-\lambda)}{x!}\frac{\mu^y exp(-\mu)}{y!}\right)^{W(a,f,k)} \tag{18}$$

2.3.7 Modelo Poisson Bivariada

Considerar a quantidade de gols marcados pelo time mandante e visitante em uma partida como independes pode gerar um certo desconforto, uma vez o time que sofre o primeiro gol na partida tende a atacar mais em busca do empate. A modelagem através da Poisson Bivariada vem sendo trabalhada a fim de incorporar a correlação entre as duas quantidades.

Como apresentada anteriormente em 2.3.2, a distribuição PB construída através da redução trivariada, tem se duas variáveis aleatórias (X,Y) dependentes e que marginalmente seguem uma distribuição de Poisson.

$$\begin{aligned} X &\sim Poi(\lambda_1 + \lambda_3) \\ Y &\sim Poi(\lambda_2 + \lambda_3) \\ (X,Y) &\sim PB(\lambda_1, \lambda_2, \lambda_3) \end{aligned}$$

Como X e Y são marginalmente Poisson, usaremos a mesma estrutura, MLG, para estimar a quantidade de gols marcadas pelo time mandante e visitante na i-ésima partida. Seja X_{mi} a quantidade de gols marcados pelo time mandante (m) na i-ésima partida, seja Y_{vi} a quantidade de gols marcadas pelo time visitante (v) na i-ésima partida. Então, é possível associar cada um dos parâmetros λ_{1i}, λ_{2i} e λ_{3i} a um modelo de regressão log-linear.

$$\begin{aligned} \lambda_{ji} &= e^{\boldsymbol{W_{ji}\beta_{ji}}} \\ \log(\lambda_{ji}) &= \boldsymbol{W_{ji}\beta_{ji}}; \qquad j = 1,2,3 \end{aligned}$$

É valido notar que os fatores explicativos não precisam ser os mesmos para cada modelo. Por exemplo, podemos ter associado ao λ_{1i} fatores relacionados à intensidade de gols do time mandante, associado ao λ_{2i} fatores relacionados à intensidade de gols do time visitante e associados ao λ_{3i} fatores relacionados às características do campeonato e da partida, tal como o árbitro escalado para a partida, clima e horário da partida.

Assim como foi feito em (SILVA et al., 2014) o parâmetro λ_3 foi considerado igual para todos os times.

2.3.7.1 Modelo 8 - Poisson Bivariada

O modelo 8 apresenta estrutura semelhante àquela apresentada no modelo 1, diferindo pelo acréscimo do parâmetro λ_{3i}.

Tendo X_i e Y_i sendo a quantidade de gols marcados pelo time mandante (m), e pelo time visitante (v), na i-ésima partida. Para estimar tais quantidades temos que;

$$\log(\lambda_{1mi}) = \beta_{0i} + \beta_{1mi} + \beta_{2vi} \tag{19}$$

$$\log(\lambda_{2vi}) = \beta_{1vi} + \beta_{2mi} \tag{20}$$

$$\log(\lambda_{3i}) = k \tag{21}$$

O modelo 8 possui o parâmetro λ_3 que está associado tanto a o número de gols marcados pelo time mandante como pelo time visitante, este parâmetro é considerado igual para todos os times.

2.3.8 Notas sobre os modelos e estimação dos parâmetros

Em contextualização com o Campeonato Brasileiro, os índices apresentados nos modelos acima estão compreendidos nos seguintes intervalos:

- $i = 1, \cdots, 38 \rightarrow$ indicando a rodada do campeonato
- $m = 1, \cdots, 20 \rightarrow$ indicando o índice do time mandante;
- $v = 1, \cdots, 20 \rightarrow$ indicando o índice do time visitante.

Os modelos aqui propostos são estimados através da função verossimilhança. Dado a complexidade do cálculo para alguns dos modelos a estimação dos parâmetros é obtida através de análise numérica. O algoritmo utilizado para encontrar as soluções das derivadas foi o Limited-Memomory- Boxed-Broyden–Fletcher–Goldfarb–Shanno (L-BFGS-B). O método LBFGS se destaca pelo fato de permitir que mais de uma restrição (BYRD et al., 1995), no caso que a soma de cada um dos parâmetros seja zero, como apresentado na equação 12.

Com relação ao uso do fator de esquecimento, foi considerado o fator sendo igual a dois. Tal parâmetro é de total escolha do usuário, baseando-se principalmente em suas premissas com relação à importância dos jogos mais antigos.

2.3.9 Medidas de Comparabilidade

Diante dos diferentes modelos propostos, uma pergunta que surge é qual possui o melhor desempenho, e para responder essa pergunta diferentes métricas foram analisadas. Quatro métricas são utilizadas neste trabalho para comparação dos modelos. Três delas comparam a capacidade preditiva do modelo no ponto de vista de acertar o resultado da partida (vitória, derrota, empate), são elas: *Rank Probability Score* , Medida de de Finetti, proporção de acertos. Para comparar a capacidade dos modelos em acertar o placar da partida é utilizada a taxa de acerto de placar. Como o intuito do trabalho é buscar o modelo com melhor desempenho preditivo, não foram utilizadas métricas como *Akaike Information Criterion* (AIC), *Bayesian Criterion Information*.

2.3.10 Rank Probability Score (RPS)

A medida RPS é uma das muitas medidas existentes para avaliar o desempenho de modelos de previsão probabilístico (CONSTANTINOU; FENTON, 2012), tais como os propostos neste trabalho. O RPS é descrito como uma medida útil para avaliar modelos probabilístico ranqueados. O termo ranqueado refere-se ao fato do resultado empate ser mais perto do resultado vitória do que o resultado vitória ser perto do resultado derrota. Se um time está ganhando uma partida por um gol de diferença, então o time visitante necessita apenas de um gol para mover o resultado de vitória do mandante para empate e necessita de dois gols para mover o resultado de vitória do mandante para vitória do visitante. Sendo assim em um jogo que o resultado foi vitória para mandante a medida RPS penalizará mais a probabilidade atribuída à vitória do visitante que a probabilidade atribuída ao empate. (CONSTANTINOU; FENTON, 2012) A medida RPS para um único jogo é definida por:

$$\text{RPS} = \frac{1}{r-1}\sum_{i=1}^{r}\left(\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} e_j\right)^2 \tag{22}$$

onde r is o número de possíveis resultados, p_j é a previsão na posição j e e_j é o resultado observado na posição j. A equação 22 representa a diferença entre a distribuição acumulada das previsões e da distribuição acumulada do resultado observado. Para critério de comparação entre os modelos é calculado a média do valor de RPS de cada jogo previsto.

Usando a medida RPS, o modelo considerado com melhor desempenho será aquele que tiver a menor média de RPS.

2.3.10.1 Medida de de Finetti

A medida de de Finetti é utilizada como uma medida de qualidade em previsões de vetores de probabilidade tricotômica. Tal medida consiste na média das distâncias de de Finetti. A distância de de Finetti é vista como a distância euclidiana quadrática entre o

vetor de probabilidades previsto e o vetor do vértice do resultado observado. E a medida de de Finetti é definida como a média das distâncias de de Finetti, também conhecida como *Score de Brier*.

Considerando o problema de atribuir probabilidades aos três possíveis resultados de uma partida de futebol, os vértices da medida de de Finetti podem assumir três possíveis valores, (1,0,0), (0,1,0) e (0,0,1), correspondendo respectivamente à vitória, empate e derrota. Para exemplificação, consideremos um jogo em que as probabilidades previstas são $\hat{p} = (0.45, 0.25, 0.30)$, correspondendo respectivamente à probabilidade de vitória, empate e derrota. Sabendo que este jogo terminou empatado temos que o vértice correspondente seria igual a (0,1,0). Temos então a distância de de Finetti = $(0.45-0)^2 + (0.25-1)^2 + (0.3-0)^2$. Para um conjunto de n jogos a medida de de Finetti é definida pela média das distâncias de de Finetti atribuída para o conjunto de n jogos.

Um ponto de referência da medida de de Finetti é quando se considera os possíveis resultados de uma partida sendo equiprováveis, atribuindo probabilidade de 1/3 para cada um dos resultados, tendo como resultado da medida de de Finetti 0.66. Modelos que possuam medida de de Finetti inferior a 0.66 podem ser ditos que possuem capacidade preditiva superior a do modelo equiprovável. Neste trabalho outro ponto da medida de de Finetti foi considerado para comparabilidade. Considerando todas as partidas ocorridas no Brasileirão de 2014 - 2018, 51% terminaram com vitória para equipe mandante, 26% terminaram empatadas e 23% com vitória para equipe visitante. Atribuindo a probabilidade $p = (0.51, 0.26, 0.23)$ a todos os jogos do campeonato podemos comparar a capacidade preditiva do modelo vs um palpite ingênuo.

Pelo aspecto da medida de de Finetti o modelo escolhido é aquele com a menor medida de de Finetti.

2.3.10.2 Proporção de acertos

A proporção de acertos mede a proporção de jogos que o modelo acertou o verdadeiro resultado. Diz-se que um modelo acertou o resultado quando o vetor de probabilidades tricotômico atribui a maior probabilidade ao resultado que de fato ocorreu. A proporção de acertos é definida como: Seja $w_i = 1$ se o modelo acerta o resultado da i-ésima partida e $w_i = 0$ se o modelo erra o resultado da i-ésima partida. Seja n a quantidade de jogos previstos. Então,

$$\text{Proporção de acertos} = \sum_{i=1}^{n} \frac{W_i}{n} \tag{23}$$

2.3.10.3 Taxa de acerto de placar

A taxa de acerto de placar para um único jogo é definida como a probabilidade que o modelo atribui para o placar que de fato ocorreu. Tal medida é vista como o produto interno das probabilidades atribuída a cada placar com vetor que possui 1 no placar ocorrido e 0 nos demais placares. Em um conjunto de n jogos a taxa de acerto de placar é definida como o a média da taxa de acerto de placar de cada jogo.

3 Análise dos Resultados

Neste capítulo pode-se observar os resultados das análises descritiva das base de dados e os resultados gerado por cada um dos modelos.

3.1 Manipulação e análise descritiva das bases

Gostaríamos de começar este capítulo com duas notas:

1. Para melhor entendimento do texto apresentado neste capítulo, leia-se o resultado derrota como vitória do time visitante. Então uma partida terá três possíveis resultados, vitória (vitória mandante), empate e derrota (vitória visitante).

2. Todos códigos de manipulação, análise e modelagem estão no endereço https://github.com/joaomamorim/brasileirao-modelagem

A fim de melhor conhecer as informações do Campeonato Brasileiro, foi realizada a análise descritiva da base com o histórico de resultados de 2014 a 2018. Os principais aspectos estão em observar a distribuição de gols marcados pelos times mandante e visitante, proporção de vitórias, empates e derrotas, além da proporção de placares observados.

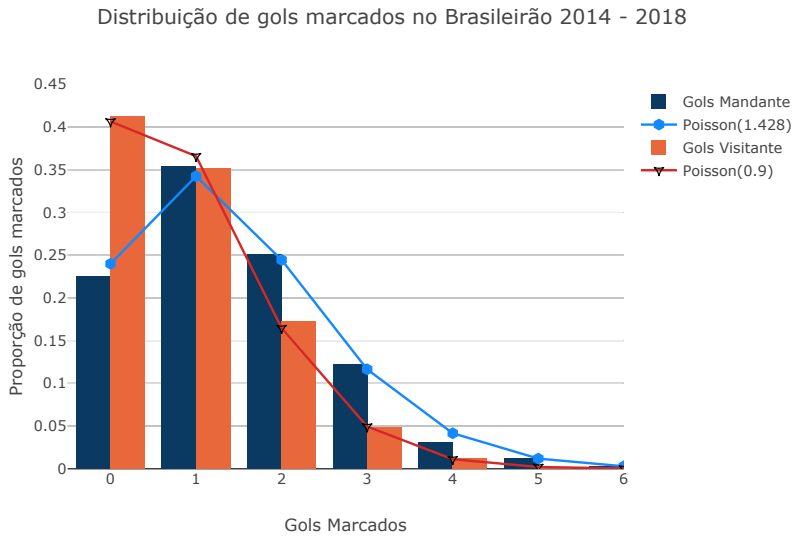

Figura 2 – Distribuição de gols marcados no Brasileirão 2014-2018

Como pode ser observado na figura 2, a suposição que os gols marcados pela equipe mandante e visitante seguem cada uma delas uma distribuição de Poisson com parâmetro igual a média da variável aleatória, no caso, média de gols marcados pela equipe mandante

igual a 1.428 e média de gols marcados pela equipe visitante igual a 0.9, não aparenta ser uma suposição grosseira. Outro ponto importante na suposição das variáveis seguirem a distribuição de Poisson é o fato da média ser igual a variância, e tais estatísticas descritivas podem ser observadas na tabela 2. Como pode ser observado, de fato, a média e a variância possuem valores bem próximos. A figura 2 também auxilia observar que de fato

Tabela 2 – Estatísticas descritiva dos gols marcados

Estatísticas	**Média**	**Variância**	**Min**	**Max**	**p25**	**Mediana (p50)**	**p75**
Gols mandante	1,43	1,34	0	6	1	1	2
Gols visitante	0,9	0,91	0	6	0	1	1

a distribuição de gols marcados pela equipe mandante e visitante tem alguma diferença, tal diferença sendo possivelmente explicada pelo 'efeito casa'.

Outra figura que auxilia notar a diferença entre a distribuição de gols marcados pela equipe mandante e visitante é o box plot. Como pode-se observar na figura 3 as duas variáveis possuem comportamento diferentes, a linha pontilhada no boxplot está representando a média de cada uma das variáveis. Um fato interessante é no caso do primeiro quartil (p25) ser igual ao segundo quartil (p50) no caso da gols marcados pelos mandantes, e o terceiro quartil (p75) ser igual ao segundo quartil (p50) para os gols marcados pela equipe visitante.

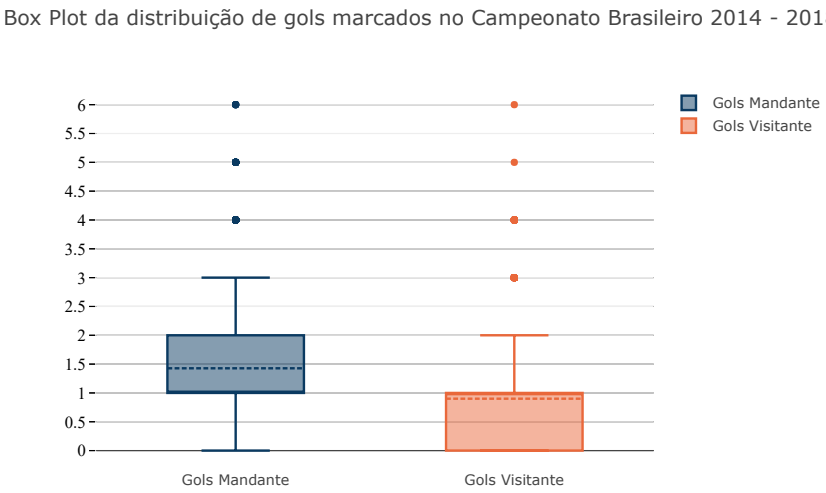

Figura 3 – Box plot da distribuição de gols marcados no Brasileirão 2014-2018

Para conhecermos a distribuição de placares durante os campeonatos de 2014-2018 podemos ver a figura 4. Como pode ser constatado na figura 4, o placar mais observado foi 1x0 para equipe mandante, 1x1, 2x0 para equipe mandante, 2x1 para equipe mandante, 0x0 e 1x0 para equipe visitante. Pode-se perceber que de fato o time da casa costuma sair

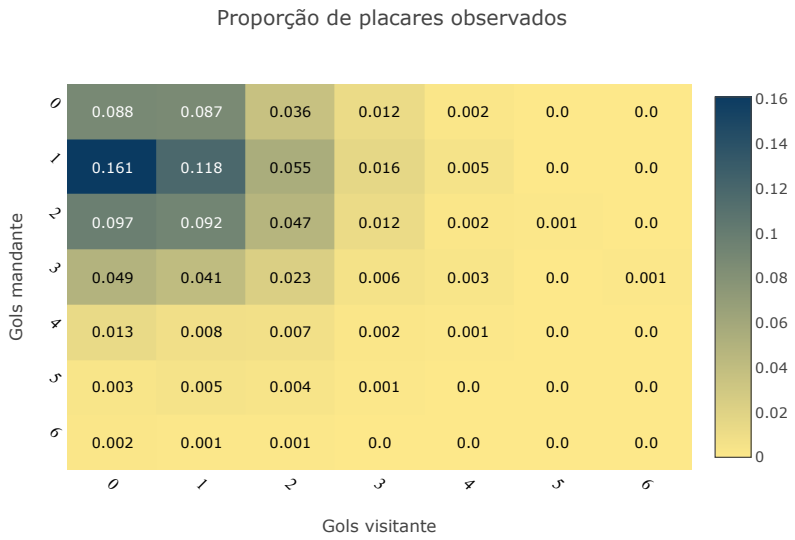

Figura 4 – Heatmap dos placares do Brasileirão 2014-2018

com a vitória. A figura 5 auxilia observar que tal efeito se manteve ao longo dos anos, no qual das partidas jogadas no Campeonato Brasileiro 2014 até 2018, em 51% terminou com vitória do time mandante, em 23% delas terminou empatada e em 21% delas com vitória do time visitante.

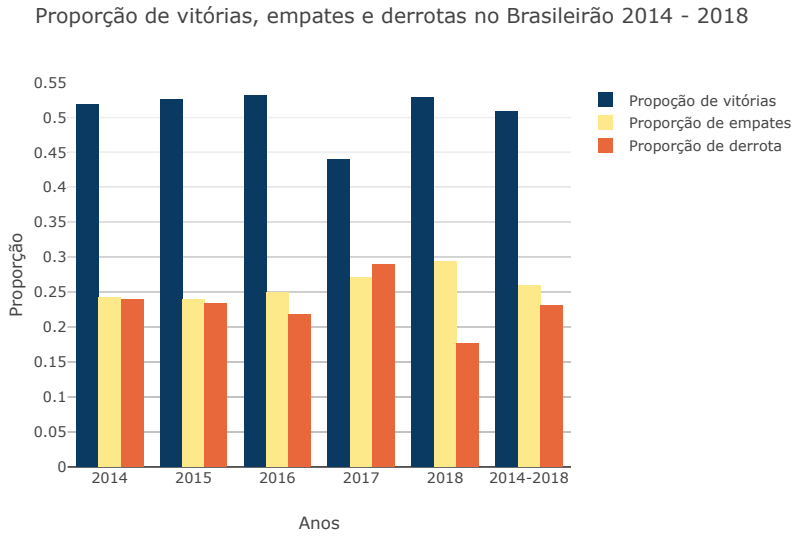

Figura 5 – Proporção de vitória, empate e derrota ao longo dos Campeonatos Brasileiro

Com a análise das 4 figuras apresentadas até aqui, podemos observar os seguintes pontos:

1. Não seria um absurdo considerar a distribuição de gols marcados através de uma

distribuição Poisson;

2. De fato o time da casa tende a se sair melhor em suas partidas;

3. A vantagem do time da casa se manteve ao longo dos anos.

Como a base do Campeonato Brasileiro apresenta apenas rodada, ano, time mandante, time visitante e placar, vamos analisar a base de scouts do Cartola FC. Antes de começar apresentar os resultados da base do Cartola FC quero apresentar parte das manipulações feitas na base e suas motivações.

Apesar da grande quantidade de informações provenientes nas bases do Cartola FC, nem todos jogos do Campeonato Brasileiro estavam no Cartola. Partidas que são adiadas ou com problema de calendário não possuem informações de scouts, ficando assim sem qualquer informação de scouts. Para lidar com esse problema limitamos a quantidade de variáveis disponíveis, ficando apenas com : clube, rodada, ano, faltas sofridas, finalizações na trave, finalizações defendida, finalizações para fora, roubadas de bola, faltas cometidas e cartões amarelos. O segundo passo para lidar com o problema foi preencher os scouts da partidas faltantes com informações provenientes de sites de scouts. O Site utilizado foi o https://whoscored.com, devido a riqueza de informações e o fato de os scouts nele apresentados serem bem semelhantes aos observados no Cartola.

Para o ano de 2015, 2017 e 2018 os scouts apresentados para cada jogador estava sendo acumulado ao longo das rodadas, porém para algumas rodadas e jogadores a informação deixava de acumular e só contava da rodada corrente. Para resolver tal problema foi necessário identificar quais jogadores e rodadas apresentavam tal problema. Após identificar tais jogadores e rodadas foi possível gerar a informação de cada jogador por rodada em vez de acumular seu scouts ao longo das rodadas.

Ao fim desses dois processos foram geradas duas variáveis:

- Finalizações certas = Finalizações na trave + Finalizações defendidas

- Finalizações total = Finalizações certas + Finalizações para fora

Para finalizar o pré-processamento da base do Cartola, as informações de um clube na rodada foi gerada como a soma dos scouts dos jogadores daquele clube na respectiva rodada. Por exemplo, os scouts da equipe do Ceará na sétima rodada do Campeonato Brasileiro de 2018, foi a soma dos scouts dos jogadores do Ceará que jogaram a sétima rodada do Campeonato Brasileiro de 2018. Após isso tínhamos os scouts por time, rodada e ano. Uma amostra da base final pode ser vista na figura 6.

Com a disponibilidade de outras variáveis além dos gols marcados, ou seja, aquelas apresentadas no scouts do Cartola, pode-se fazer uso de modelos que utilizem tais variáveis

Clube	Rodada	Ano	FS	FT	FD	FF	RB	FC	CA	FinC	Finalizacoes
Corinthians	18	2017	18.000	0.000	2.000	3.000	16.000	12.000	1.000	2.000	5.000
América-MG	28	2016	13.000	0.000	0.000	6.000	10.000	15.000	2.000	0.000	6.000
Coritiba	5	2014	12.000	0.000	2.000	1.000	13.000	12.000	1.000	2.000	3.000
Cruzeiro	23	2017	15.000	0.000	5.000	2.000	10.000	17.000	1.000	5.000	7.000
Chapecoense	7	2018	20.000	1.000	4.000	5.000	19.000	14.000	1.000	5.000	10.000
São Paulo	30	2018	13.000	2.000	5.000	5.000	20.000	20.000	3.000	7.000	12.000
Palmeiras	34	2016	18.000	0.000	2.000	3.000	10.000	9.000	0.000	2.000	5.000
Chapecoense	24	2017	6.000	0.000	4.000	2.000	20.000	18.000	5.000	4.000	6.000

Figura 6 – Amostra da base final do Cartola FC

em busca de uma melhor capacidade preditiva. Como relatado nos modelos 3, 5 e 6 que fazem uso da informação do número de Finalizações certas, observou-se através da figura 7 o comportamento de tal variável para a equipe mandante e visitante. Como pode ser

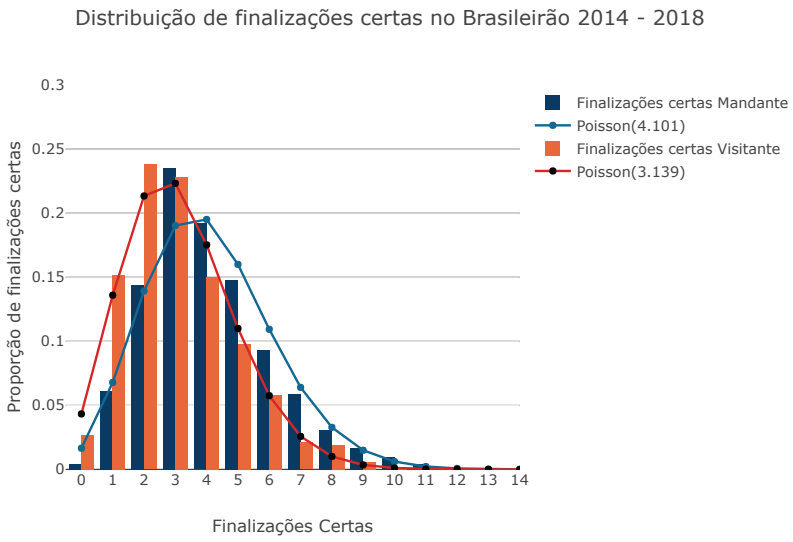

Figura 7 – Distribuição do número de finalizações certas no Brasileirão 2014 - 2018

observado na figura 7 de fato o número de finalizações certas pode ser aproximado por uma distribuição Poisson com o parâmetro λ sendo a média da variável aleatória. Outro ponto importante é observar se a média e a variância da variável aleatória são próximas, uma vez que a distribuição de Poisson possui média igual a variância, tais estatísticas podem ser observadas na tabela 3.

Assim como foi possível observar na figura 2, também é possível perceber na figura 7 que a distribuição do time da casa e do time visitante tem uma certa diferença. A figura 7 mostra que o time da casa tende gerar um número maior de finalizações certas na partida.

Tabela 3 – Estatísticas descritivas do número de finalizações certas

Estatísticas	Média	Variância	Min	Max	p25	Mediana (p50)	p75
Finc mandante	4.101	4.076	0.0	3.0	4.0	5.0	14.0
Finc visitante	3.139	3.389	0.0	2.0	3.0	4.0	14.0

Finc: Finalizações certas

Outra variável utilizada foi o total de finalizações, como pode ser observado na figura 8, poderíamos acreditar que tal variável também possa ser modelada pela distribuição de Poisson, como feito no modelo 6, no entanto a média e variância são bem diferentes, como pode ser visto na tabela 4.

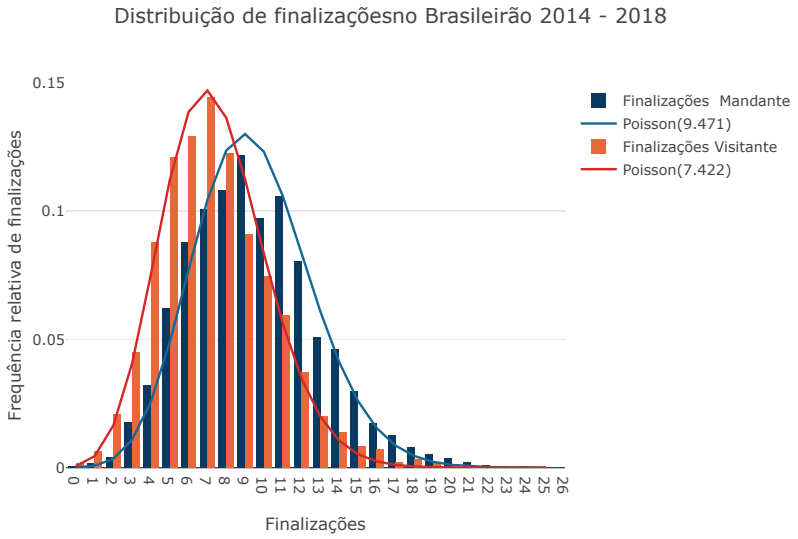

Figura 8 – Distribuição do número total de finalizações no Brasileirão 2014 - 2018

Tabela 4 – Estatísticas descritivas do número total de finalizações

Estatísticas	Média	Variância	Min	Max	p25	Mediana (p50)	p75
Fina mandante	9.472	12.184	0.0	25.0	7.0	9.0	12.0
Fina visitante	7.422	9.386	0.0	22.0	5.0	7.0	9.0

Fina: Total de finalizações

Como a variável roubada de bola também foi utilizada como variável explicativa nos modelos 2 e 3, pode ser observado na figura 9 que aparentemente não é possível detectar um comportamento marcante seja pro time mandante ou visitante.

3.2 Análise dos modelos

Ao todo foram apresentados oito modelos, cada um deles foi avaliado segundo os critérios apresentados na subseção 2.3.9. Vale ressaltar que um modelo pode ter o melhor

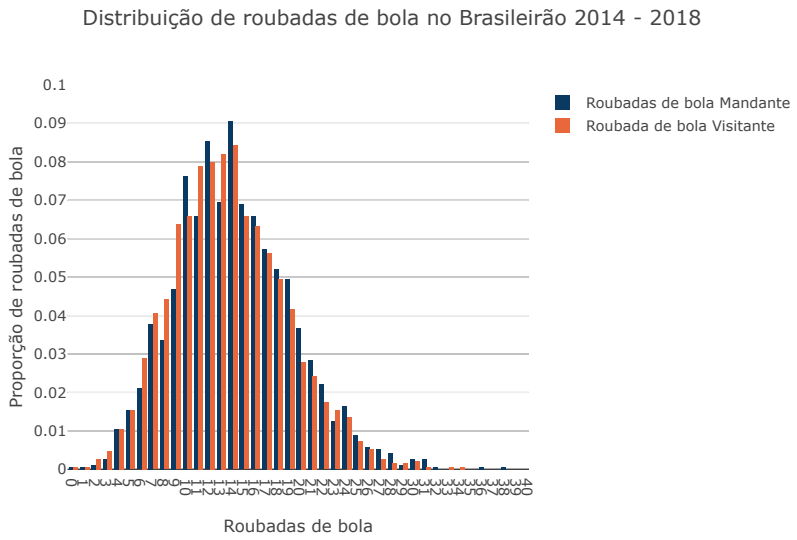

Figura 9 – Distribuição do número de roubadas de bola no Brasileirão 2014 - 2018

resultado em uma medida e não ser o melhor em outra. Cada medida observa um comportamento do modelo, não sendo necessário concordarem na escolha do modelo.

Foram analisados dois tipos de previsões, h-passos e 1-passo. A previsão h-passos está relacionada ao intuito de prever uma grande quantidade de rodadas, como por exemplo, na 25ª rodada prever todas as demais rodadas faltantes. Tal tipo de previsão é necessária no intuito de prever chances de classificação de um time. A previsão 1-passo tem o intuito de prever apenas a rodada seguinte a última rodada que se tem informação. Por exemplo, para prever um jogo da 20ª rodada, são utilizados todos dados até 19ª rodada para o ajuste do modelo. A abordagem de previsões 1-passo é útil, por exemplo, para apostadores. Convenhamos que fazer uma aposta baseada em uma previsão feita 3 meses antes não seria uma boa estratégia, uma vez que o desempenho dos times envolvidos podem ter variado muito durante os 3 meses.

A previsão h-passos foi realizada considerando 3 pontos distintos: No primeiro o modelo era ajustado com os dados até 19ª rodada e previsto todas as rodadas faltantes, no segundo o modelo era ajustado com os dados até a 25ª rodada e previstas as 13 rodadas faltantes, e no terceiro ponto o modelo era ajustado considerando as informações até a rodada 33 e previstas as 5 rodadas faltantes. Os pontos foram escolhidos ao acaso, exceto o ponto da 19ª rodada, pois configura-se a rodada em que todos os times já se enfrentaram.

A previsão 1-passo foi feita a partir da 19ª rodada. Para prever os jogos da 20ª rodada o modelo era ajustado com informações até a 19ª rodada, para prever a 21ª rodada o modelo era reajustado com informações até a 20ª rodada, assim por diante.

As medidas de comparabilidade foram calculadas apenas para os jogos posteriores a

última rodada usada no ajuste dos modelos. A tabela 5 apresenta as medidas RPS e de de Finetti para as previsões 1-passo. Foram simulados e utilizados nos cálculos das medidas todos jogos do segundo turno do campeonato. A tabela 6 apresenta as mesmas medidas considerando a previsão h-passos no ponto 19, sendo usado como base de treino todos jogos até a 20ª rodada. As médias proporção de acertos e taxa de acerto de placar são mostradas na tabela 7 para o modelo 1-passo e na tabela 8 para o modelo h-passos. Como em ambas abordagem foram previsto todo segundo turno, fica fácil perceber se a abordagem 1-passo, que tem mais trabalho computacional, trouxe alguma melhora na predição dos modelos. As tabelas que possuem a linha "Palpite Bra"leia-se como o desempenho do modelo ingênuo, em que são atribuídos probabilidade $(0.51, 0.26, 0.23)$ para todos os jogos. A motivação dessas probabilidades foi apresentada na subseção 2.3.10.1. As tabelas considerando a abordagem h- passos no ponto 25 e 33 podem ser encontradas no apêndice deste trabalho.

Tabela 5 – Tabela de comparação das medidas RPS e de Finetti 1 - passo

Modelo	**Medida**									
	RPS					**de Finetti**				
	Ano									
	2014	**2015**	**2016**	**2017**	**2018**	**2014**	**2015**	**2016**	**2017**	**2018**
Modelo 1	0.2069	0.2246	0.1997	0.2421	0.1777	0.5873	0.6264	0.5842	0.6899	0.5595
Modelo 2	**0.2062**	0.2245	0.1984	0.2415	0.1769	**0.5854**	0.6264	0.5813	0.6896	0.5575
Modelo 3	0.2077	0.2257	0.2009	0.2428	0.1791	0.5889	0.6294	0.5861	0.6911	0.5625
Modelo 4	0.2063	0.2241	**0.1981**	**0.2413**	**0.1761**	0.5856	0.6253	**0.5811**	**0.6892**	**0.5560**
Modelo 5	0.2078	0.2254	0.2003	0.2425	0.1784	0.5893	0.6286	0.5853	0.6904	0.5611
Modelo 6	0.2103	**0.2233**	0.2051	0.2421	0.1797	0.5916	**0.6218**	0.5949	0.6921	0.5700
Modelo 7	0.2206	0.2292	0.2038	0.2445	0.1985	0.6101	0.6346	0.5924	0.6956	0.6065
Modelo 8	0.2107	0.2237	0.1987	0.2435	0.1780	0.5988	0.6248	0.5850	0.6918	0.5576
Palpite Bra*	0.2135	**0.2209**	0.2109	**0.2244**	0.1953	0.5956	0.6181	0.6055	**0.6554**	0.6023

Tabela 6 – Tabela de comparação das medidas RPS e de Finetti h - passos (19ª Rodada)

Modelos	**Medida**									
	RPS					**de Finetti**				
	Ano									
	2014	**2015**	**2016**	**2017**	**2018**	**2014**	**2015**	**2016**	**2017**	**2018**
Modelo 1	0.2041	0.2311	0.2047	0.2505	0.1851	0.5848	0.6402	0.5934	0.7092	0.5783
Modelo 2	**0.2031**	0.2299	0.2034	0.2490	0.1861	**0.5823**	**0.6376**	0.5906	0.7072	0.5802
Modelo 3	0.2041	0.2321	0.2059	0.2517	0.1881	0.5847	0.6426	0.5957	0.7114	0.5848
Modelo 4	**0.2031**	0.2304	0.2032	0.2486	**0.1840**	0.5824	0.6386	0.5902	0.7064	**0.5757**
Modelo 5	0.2041	0.2324	0.2054	0.2507	0.1863	0.5848	0.6432	0.5950	0.7095	0.5809
Modelo 6	0.2153	0.2314	**0.2008**	**0.2392**	0.2034	0.6079	0.6407	**0.5839**	**0.6866**	0.6190
Modelo 7	0.2197	0.2411	0.2059	0.2571	0.2063	0.6098	0.6586	0.5954	0.7222	0.6260
Modelo 8	0.2066	**0.2312**	0.2030	0.2577	0.1883	0.5974	0.6407	0.5930	0.7234	0.5849
Palpite br	0.2135	0.2209	0.2109	**0.2244**	0.1953	0.5956	**0.6181**	0.6055	**0.6554**	0.6023

Como pode ser constatado na tabela 5 na maioria dos anos, o modelo 4 teve melhor desempenho que os demais modelos em ambas medidas. O modelo 2 foi aquele que teve um desempenho próximo do modelo 4, podemos assim perceber a contribuição da variável "Finalizações"para o ajuste do modelo. Um detalhe importante a se destacar é o baixo desempenho dos modelos nos anos de 2015 e 2017, em que o "Palpite Bra"teve um melhor desempenho, porém ao observar as previsões realizadas pelos modelos notou-se que os modelos atribuíram probabilidades compatíveis com as opiniões de torcedores. Como foram previsto 190 jogos e isso geraria uma tabela enorme neste trabalho, tais previsões podem ser vista no github (SANTOS, 2019) https://github.com/joaomamorim/brasileirao-modelagem/. Outro ponto interessante de ser observado é o baixo desempenho do modelo proposto por Dixon e Coles. O modelo foi proposto pensando nas características da Barclays Premier League daquela década, em que ocorria um maior percentual de empates, possivelmente esse pode ser o fato do modelo não ter tido um bom desempenho no campeonato brasileiro.

Com relação a tabela 6 podemos constatar que assim como na abordagem 1 - passo , o modelo 4 também teve o melhor desempenho comparado aos demais modelos, porém para o ano de 2016 o modelo 6 foi o de melhor desempenho. Quando comparamos os resultados da tabela 5 com os da tabela 6 observa-se que de fato o uso de uma maior quantidade de dados e dados mais recentes melhorou a capacidade preditiva dos modelos, porém esse resultado não foi observado para o ano de 2014, mas sendo verdade nos demais anos.

Tabela 7 – Tabela de comparação das proporções de acerto e taxas de acerto de placar 1 - passo

Modelo	**Medida**									
	Proporção de acertos					**Taxa de acerto de placar**				
	Ano									
	2014	**2015**	**2016**	**2017**	**2018**	**2014**	**2015**	**2016**	**2017**	**2018**
Modelo 1	0.5211	0.4842	0.5368	0.3579	0.5526	0.0894	0.0779	0.0874	0.0793	0.1006
Modelo 2	0.5211	0.4947	0.5263	0.3737	0.5632	0.0894	0.0776	**0.0881**	0.0799	**0.1011**
Modelo 3	0.5211	0.4895	0.5421	0.3474	0.5526	0.0893	0.0775	0.0875	0.0792	0.1005
Modelo 4	0.5158	**0.5**	0.5368	0.3789	0.5632	**0.0895**	**0.0781**	0.0879	0.0799	0.1011
Modelo 5	0.5211	0.4895	0.5316	0.3474	0.5526	0.0894	0.0778	0.0874	0.0793	0.1005
Modelo 6	**0.5316**	**0.5**	**0.5474**	**0.4105**	0.4737	0.0847	0.0749	0.0869	**0.081**	0.0944
Modelo 7	**0.5316**	0.4737	0.5211	0.3632	0.4895	0.0796	0.0707	0.0829	0.0783	0.0853
Modelo 8	0.5	0.4842	0.5421	0.3474	**0.5842**	0.0887	0.0776	0.0884	0.0803	**0.1043**

Tabela 8 – Tabela de comparação das proporções de acertos e taxas de acerto de placar h - passos

Modelo	**Medida**									
	Proporção de acertos					**Taxa de acerto de placar**				
	Ano									
	2014	**2015**	**2016**	**2017**	**2018**	**2014**	**2015**	**2016**	**2017**	**2018**
Modelo 1	0.5457	0.4737	0.5053	0.3737	0.5316	0.0909	0.0784	0.0840	0.0764	0.0974
Modelo 2	0.5632	**0.5053**	0.5211	0.3789	0.5263	**0.0910**	0.0785	**0.0844**	0.0772	0.0976
Modelo 3	0.5421	0.4684	0.5000	0.3789	0.5158	0.0909	0.0781	0.0841	0.0763	0.0970
Modelo 4	**0.5632**	0.5000	0.5158	0.3789	**0.5368**	**0.0910**	**0.0786**	0.0843	0.0772	0.0978
Modelo 5	0.5474	0.4632	0.5000	0.3737	0.5105	0.0909	0.0782	0.0841	0.0762	0.0971
Modelo 6	0.5316	0.4947	**0.5526**	**0.3947**	0.4737	0.0855	0.0763	0.0839	**0.0786**	0.0913
Modelo 7	0.5053	0.4158	0.4842	0.3737	0.4737	0.0812	0.0707	0.0846	0.0772	0.0839
Modelo 8	0.4947	0.5000	0.5211	0.3668	0.5263	0.0903	0.0788	0.0792	0.0755	**0.0988**

Quando olhamos para a proporção de acertos, vemos na tabela 7 que o modelo 6 teve o melhor desempenho comparado aos demais modelos, exceto pelo ano de 2018 no qual o modelo 8 apresentou o de melhor desempenho. Quando olhamos para a abordagem h-passos, observamos na tabela 7 que o modelo 6 foi o de melhor desempenho nos anos de 2016 e 2017 e que nos demais anos o modelo 4 foi o de melhor desempenho. Assim como quando observamos a medida de de Finetti e RPS, a proporção de acertos para o ano de 2014 não teve melhora em usar os dados mais recentes. Possivelmente em 2014 os times oscilaram mais.

Com base na tabela 7 percebe-se que a média da probabilidade atribuída aos placares que de fato ocorreram foi bem próxima entre os modelos apresentados, mas pode ser notado que o modelo 4 foi sutilmente melhor que os demais, exceto para o ano de 2016 e 2018 que o modelo 8 foi o de melhor resultado. Observando a taxa de acerto de placar na abordagem h-passos pode ser visto que as médias das probabilidades atribuída ao placar da partida também foram muito semelhantes, sendo o modelo 4 sutilmente superior aos demais no ano de 2014, o modelo 6 sutilmente superior no ano de 2017 e o modelo 8 sutilmente superior nos anos de 2015, 2016 e 2018.

Apesar da comparação dos modelos ser baseada nas médias de cada uma das medidas, outras estatísticas das medidas podem ser apresentadas, como primeiro quartil, mediana, terceiro quartil. Para verificar o comportante da distribuição das medidas foi gerado o bloxplot de cada uma delas para os oito modelos apresentados. A figura 10 mostra o boxplot da distribuição da distância de de Finetti para cada um dos oito modelos considerando a abordagem 1-passo para o ano de 2018, semelhantemente a figura 11 mostra o boxplot para a distribuição do RPS de cada um dos oito modelos, e a figura 12 apresenta o boxplot para distribuição da taxa de acerto de placar. No geral a abordagem 1-passo obteve um desempenho melhor que a h-passos, os boxplot são apresentados para apenas a abordagem 1-passo e para o ano de 2018. Os demais anos podem ser visto em (SANTOS, 2019).

Através da figura 10 é possível perceber que a distribuição da distância de de Finetti dos modelos é bem semelhante, além disso nota-se que os modelos 6 e 7 em 50% dos casos tem a distância de de Finetti maior que 0.5. Outro ponto importante notado é que apesar do modelo 8 ter a média da distância de de Finetti bem distante da média do modelo 4, ambos modelos possuem comportamento e mediana bem semelhante.

Na figura 11 é constatado que assim como na figura anterior onde a maioria dos modelos possuíam distribuição bem semelhante, aqui tal fato também pode ser notado, sendo novamente o modelo 6 e 7 os que tem um comportamento diferente dos demais.

A figura 12 mostra que apesar dos modelos 6 e 7 terem um comportamento diferente dos demais, eles são os que tem os dados mais concentrados, entretanto são os dois que possuem a menor mediana.

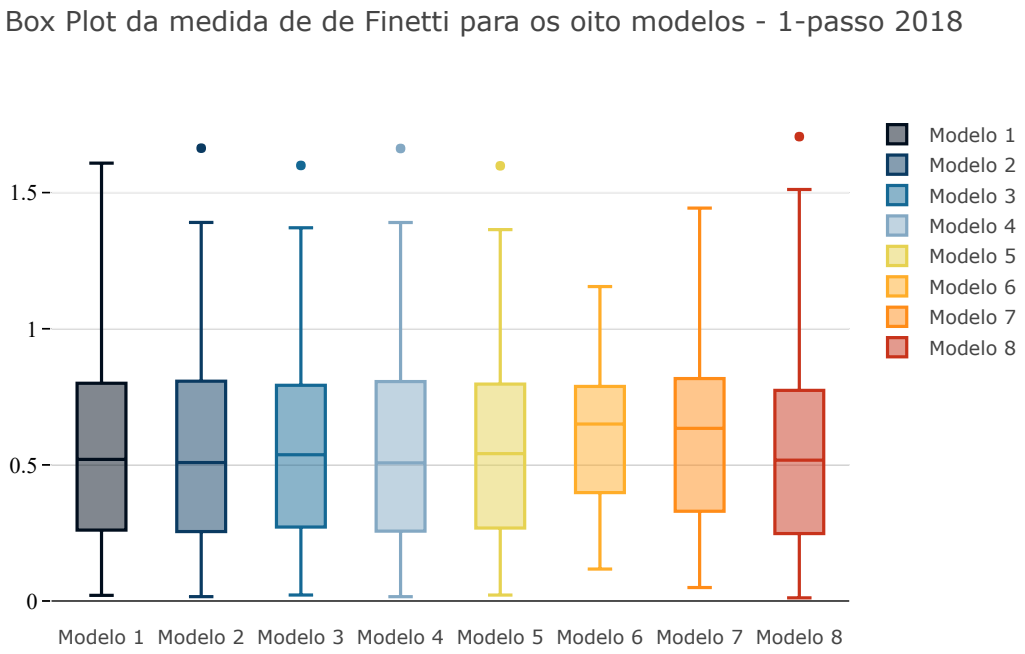

Figura 10 – Boxplot da distribuição da distancia de de Finetti de cada modelo

Como pode ser constatado nas figuras 10, 11 e 12, o modelo 8 possui comportamento muito parecido aos modelos 1, 2 ,3 e 4. Tal semelhança entre os modelos pode ser dada pelo fato que ao considerarmos os modelos via Poisson independente, trata-se de um caso particular da Poisson bivarida, cujo parâmetro $\lambda_3 = 0$.

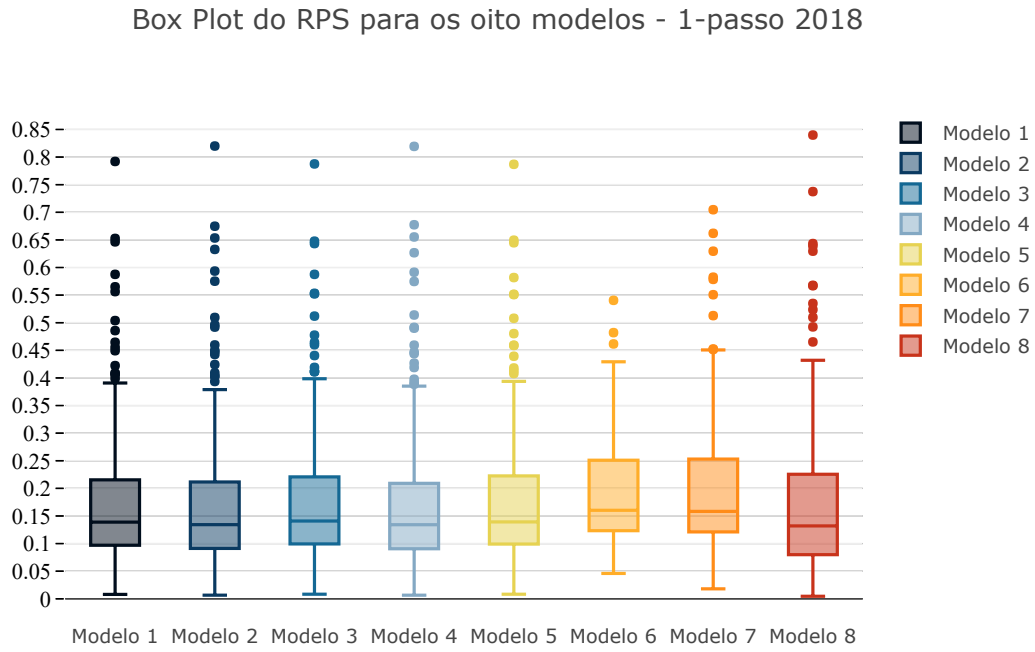

Figura 11 – Boxplot da distribuição do RPS de cada modelo

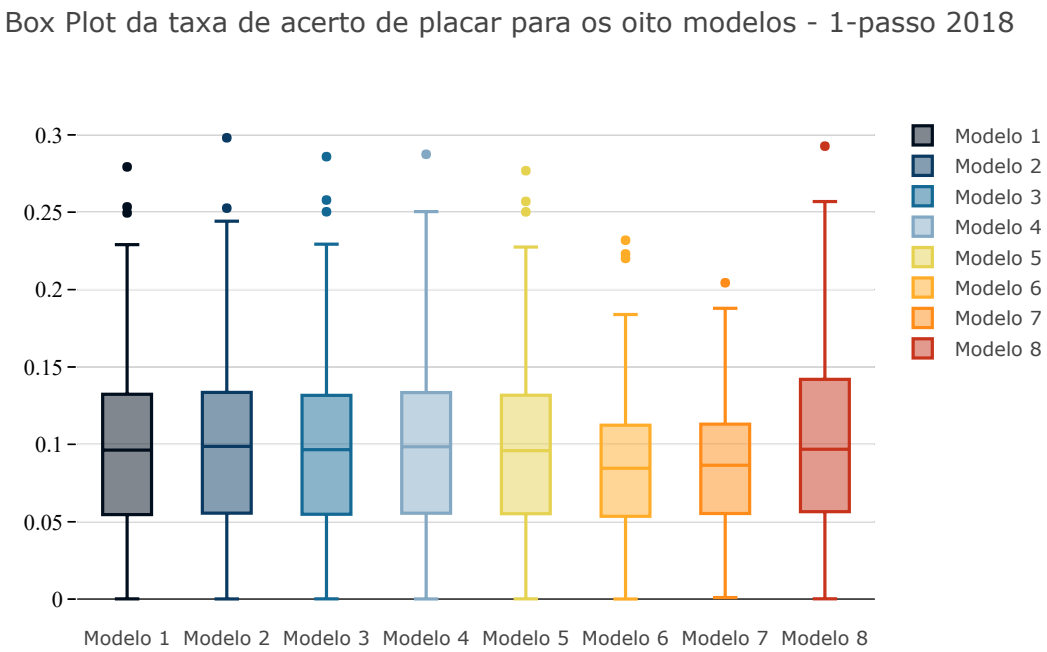

Figura 12 – Boxplot da distribuição da taxa de acerto de placar para os modelo

4 Conclusão

A modelagem de resultados de partida de futebol vem sendo exploradas há décadas. Nesses anos uma das principais abordagens que amplamente explorada foi modelar o número de gols marcados por cada uma das equipes como variável aleatória pertencente à distribuição Poisson. Na maioria dos casos os modelos tinham seus dados limitados aos clubes participante da partida e o placar da partida. Este trabalho teve por objetivo comparar o desempenho preditivo de diferentes abordagens propostas na literatura, além de tentar melhorar a capacidade preditiva dos modelos usando informações provenientes do Cartola FC, tais como finalizações e roubadas de bola.

Ao todo neste trabalho foram apresentados oito modelos. Os principais questionamentos era se o uso de variáveis como finalizações melhoraria a capacidade preditiva do modelo independente, podendo ser observado na comparação do modelo 1 (LEE, 1997) *vs* modelo 2, 3 e 4. Outra questão era se a modelagem Poisson bivariada, modelo 8, possuía vantagem em relação ao modelo Poisson independe, ou se o modelo proposto por (DIXON; COLES, 1997) se sairia melhor.

Com os resultados apresentados foi possível notar que o modelo 4 foi o que teve o melhor desempenho na maioria das vezes. Também foi possível observar uma leve vantagem do modelo 2 e 3 quando comparado ao modelo 1. Apesar de ser notada uma vantagem em fazer o uso da variável roubadas de bola, o desempenho foi melhor quando o modelo não levou em consideração tal variável, levando em consideração apenas as finalizações.

Apesar de inicialmente acreditarmos que o número de finalizações certas seria melhor que o número total de finalizações, como visto em (STENERUD, 2015), o modelo que fazia o uso do número total de finalizações, modelo 4, teve desempenho melhor que o modelo que fazia uso das finalizações certas, modelo 5.

O modelo 6 foi proposto considerando uma previsão em dois passos, dependendo de prever o número de finalizações certas e posteriormente calcular a probabilidade de um time marcar x gols em uma partida, baseando-se na distribuição binomial. Pode ser visto que o modelo não teve o melhor desempenho baseado nas medidas de de Finetti e RPS, porém teve um bom desempenho na proporção de acertos. Uma outra abordagem que poderia ser explorada no modelo 6 é em vez de usar o número de finalizações certas, basear no número total de finalizações, porém tal variável não pode ser modelada com a distribuição Poisson, pois tem média muito diferente da variância, podendo então ser pensado o uso de outra distribuição. Outro ponto importante a ser destacado no modelo 6 é possivelmente modelar o parâmetro p, através de um modelo de mistura Beta-Bernoulli, pois é o que mais se assemelha a proporção de gols por finalizações no Brasileirão.

Assim como foi modelado o número de finalizações certas no modelo 6 através da

distribuição Poisson independente, inicialmente pensava-se em modelar o número de finalizações através da distribuição Poisson bivariada, porém a distribuição Poisson bivariada da classe de Holgate, necessita que a correlação entre as duas variáveis, "Finalizações certa time da casa"e "Finalizações certas time visitante", tenham correlação positiva. Para os dados aqui analisados foi observado que a correlação entre as variáveis era negativa, tanto para finalizações certas quanto para finalizações totais, o que limita o uso da Poisson bivariada neste caso.

Quando comparamos os modelos 1 com o modelo 8, que compara o uso da priori que a quantidade de gols marcadas por ambos times é independente *vs* considerar a quantidades de gols marcadas ambos times correlacionadas, percebe-se que os modelos tiveram pouca diferença na capacidade preditiva. Tal fato pode ser explicado pelo fato do parâmetro λ_3 ter sido muito próximo de zero. A correlação entre gols marcados pelo time mandante com gols marcados pelo time visitante foi menor que 0.005, o que pode explicar tal semelhança.

Apesar da conclusão baseada na medida de de Finetti e na proporção de acertos serem diferentes, cabe o usuário ver a medida de acurácia que melhor se ajusta a sua necessidade.

4.1 Trabalhos Futuros

Como foi observado neste trabalho o uso de informações como o número de finalizações certas e finalizações totais contribuíram para a capacidade dos modelos. Apesar disso foi percebido que nos anos de 2015 e 2017 os 8 modelos não tiveram um bom desempenho. Uma possível explicação é o fato do futebol gerar surpresas, em que nem sempre o time favorito ganha, as vezes o primeiro colocado perde para o último colocado. Tal tipo de situação penaliza os modelos quando olhamos para medida de de Finetti e RPS.

De fato é muito complicado prever todos esses percalços que ocorrem nas partidas, mas possíveis melhorias que podemos explorar em trabalhos futuros é o uso de outras variáveis como posse de bola dos times, percentual de pontos ganhos até a rodada. Neste trabalho podemos observar que o modelo 8 não fez uso de informações como o número de finalizações, uma possível melhoria no modelo 8 seria fazer uso de tal variável, dentre outras que possam explicar o resultado em uma partida.

Uma outra abordagem que podemos explorar é considerar a modelagem do número de gols através de outros tipos de modelagem para o parâmetro λ da distribuição Poisson, como considerar um processo Gaussiano, onde não necessitaríamos de uma relação linear entre as variáveis explicativas. Outro ponto a ser considerado é usar dados dos anos anteriores na previsão. Apesar dos elenco dos times mudarem de um ano para o outro pode ser relevante utilizar informações do ano anterior, para tal uma questão a ser pensada é como tratar os times que foram promovidos a divisão superior e os que foram rebaixados. Uma outra abordagem que vem sendo explorada que pode ser ainda mais explorada é o uso

de *machine learning*. Explorando técnicas como *Random Forest*, *Decision Trees*, *Artificial neural network*. Apesar de tais técnicas não possuirem interpretabilidade nas variáveis dependentes, ainda assim podem ter um bom resultado preditivo.

Referências

ALVES, A. M. et al. Logit models for the probability of winning football games. **Pesquisa Operacional**, SciELO Brasil, v. 31, n. 3, p. 459–465, 2011.

ARRUDA, M. L. d. **Poisson, Bayes, Futebol e DeFinetti**. Tese (Doutorado) — Universidade de São Paulo, 2000.

ARTUSO, A. R. Distribuição gaussiana dos resultados do campeonato brasileiro de futebol: um modelo para estimar classificações em campeonatos de modalidades coletivas. **Revista Brasileira de Ciências do Esporte**, v. 30, n. 1, 2008.

BAIO, G.; BLANGIARDO, M. Bayesian hierarchical model for the prediction of football results. **Journal of Applied Statistics**, Taylor & Francis, v. 37, n. 2, p. 253–264, 2010.

BYRD, R. H. et al. A limited memory algorithm for bound constrained optimization. **SIAM Journal on Scientific Computing**, SIAM, v. 16, n. 5, p. 1190–1208, 1995.

CBF. **Regulamento Específico da Competição Campeonato Brasileiro da Série A 2018**. [S.l.]: 2018, 2018.

CONSTANTINOU, A. C.; FENTON, N. E. Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. **Journal of Quantitative Analysis in Sports**, De Gruyter, v. 8, n. 1, 2012.

COURNEYA, K. S.; CARRON, A. V. The home advantage in sport competitions: A literature review. **Journal of Sport and Exercise Psychology**, v. 14, n. 1, p. 13–27, 1992.

DINIZ, M. A. et al. Comparing probabilistic predictive models applied to football. **Journal of the Operational Research Society**, Taylor & Francis, p. 1–13, 2018.

DIXON, M. J.; COLES, S. G. Modelling association football scores and inefficiencies in the football betting market. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 46, n. 2, p. 265–280, 1997.

FARIAS, F. F. Análise e previsão de resultados de partidas de futebol. **Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro**, 2008.

GODDARD, J. Regression models for forecasting goals and match results in association football. **International Journal of forecasting**, Elsevier, v. 21, n. 2, p. 331–340, 2005.

KARLIS, D.; NTZOUFRAS, I. Analysis of sports data by using bivariate poisson models. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 52, n. 3, p. 381–393, 2003.

KAWAMURA, K. Direct calculation of maximum likelihood estimator for the bivariate poisson distribution. **Kodai Math. J.**, Tokyo Institute of Technology, Department of Mathematics, v. 7, n. 2, p. 211–221, 1984. Disponível em: <https://doi.org/10.2996/kmj/1138036908>.

LEE, A. J. Modeling scores in the premier league: is manchester united really the best? **Chance**, Taylor & Francis Group, v. 10, n. 1, p. 15–19, 1997.

MAHER, M. J. Modelling association football scores. **Statistica Neerlandica**, Wiley Online Library, v. 36, n. 3, p. 109–118, 1982.

NEVILL, A. M.; NEWELL, S. M.; GALE, S. Factors associated with home advantage in english and scottish soccer matches. **Journal of Sports Sciences**, Taylor & Francis Group, v. 14, n. 2, p. 181–186, 1996.

OLIVIERI FILHO, C. et al. Uma abordagem bayesiana para previsÃo de resultados de jogos de futebol: Uma aplicaÇÃo ao campeonato inglÊs. **REVISTA BRASILEIRA DE BIOMETRIA**, v. 35, n. 1, p. 76–97, 2017. ISSN 1983-0823. Disponível em: <http://www.biometria.ufla.br/index.php/BBJ/article/view/296>.

POLLARD, R. Home advantage in soccer: A retrospective analysis. **Journal of sports sciences**, Taylor & Francis Group, v. 4, n. 3, p. 237–248, 1986.

RUSSELL, B. **top 10 Most Popular Sports | Most Followed Sports | Most Watched Sports**. 2017. Accessed: 2010-09-30. Disponível em: <https://sporteology.com/top-10-popular-sports-world/>.

SANTOS, J. **Previsões de Resultados em Partidas do Campeonato Brasileiro de Futebol**. [S.l.]: GitHub, 2019. <https://github.com/joaomamorim/brasileirao-modelagem>.

SARAIVA, E. F. et al. Predicting football scores via poisson regression model: applications to the national football league. **Communications for Statistical Applications and Methods**, Korean Statistical Society, v. 23, n. 4, p. 297–319, 2016.

SILVA, W. B. d. et al. Distribuição de poisson bivariada aplicada à previsão de resultados esportivos. Universidade Federal de São Carlos, 2014.

SPORT, G. **Premier League clubs spend £1.4bn to break summer transfer record**. Guardian News and Media, 2017. Disponível em: <https://www.theguardian.com/football/2017/sep/01/transfer-window-deadline-day-record-spend>.

STENERUD, S. G. **A study on soccer prediction using goals and shots on target**. Dissertação (Mestrado) — NTNU, 2015.

SUZUKI, A. K. et al. Modelagem estatística para a determinação de resultados de dados esportivos. Universidade Federal de São Carlos, 2007.

Apêndices

.1 Resultados auxiliares

Tabela 9 – Tabela de comparação das proporções de acertos e taxas de acerto de placar h - passos (25º rodada)

Modelos	**Medida**									
	Proporção de acertos					**Taxa de acerto de placar**				
	Ano									
	2014	**2015**	**2016**	**2017**	**2018**	**2014**	**2015**	**2016**	**2017**	**2018**
Modelo 1	0.5461	0.4923	0.5538	0.3769	0.5462	0.0891	**0.0765**	0.0885	0.0781	0.0952
Modelo 2	0.5462	0.4846	0.5385	0.3846	0.5462	0.0889	0.0762	0.0891	0.0788	0.0952
Modelo 3	0.5462	0.4846	0.5462	0.3769	0.5154	0.0892	0.0764	0.0886	0.0780	0.0950
Modelo 4	0.5462	**0.5000**	0.5462	0.3846	0.5308	0.0890	0.0762	0.0889	0.0788	0.0953
Modelo 5	0.5462	0.4923	0.5538	0.3769	0.5231	**0.0892**	**0.0765**	0.0885	0.0780	0.0952
Modelo 6	**0.5615**	0.4846	0.5077	**0.4538**	0.4846	0.0878	0.0762	0.0857	**0.0819**	0.0921
Modelo 7	0.5461	0.4538	0.5462	0.3846	0.4923	0.0817	0.0692	0.0847	0.0773	0.0816
Modelo 8	0.5077	0.4769	**0.5615**	0.4000	**0.5385**	0.0870	0.0751	**0.0906**	0.0793	**0.0990**

Tabela 10 – Tabela de comparação das proporções de acertos e taxas de acerto de placar h - passos (33º rodada)

Modelos	**Medida**									
	Proporção de acertos					**Taxa de acerto de placar**				
	Ano									
	2014	**2015**	**2016**	**2017**	**2018**	**2014**	**2015**	**2016**	**2017**	**2018**
Modelo 1	0.4000	0.4600	0.5400	0.3000	0.5800	0.0864	0.0766	0.0884	0.0765	0.1157
Modelo 2	0.4000	0.4800	0.5400	0.3600	0.6200	0.0863	0.0757	0.0894	0.0770	0.1163
Modelo 3	0.4000	0.4600	0.5400	0.3000	0.5800	0.0866	0.0764	0.0882	0.0766	0.1156
Modelo 4	0.3800	0.4800	0.5400	0.3600	0.6200	0.0863	0.0764	0.0894	0.0770	0.1163
Modelo 5	0.4000	0.4600	0.5400	0.3000	0.5800	0.0865	0.0769	0.0882	0.0765	0.1156
Modelo 6	0.6000	0.4400	0.5400	0.4200	0.5200	0.0894	0.0730	0.0889	0.0803	0.1071
Modelo 7	0.5400	0.4800	0.5200	0.3400	0.5600	0.0820	0.0695	0.0817	0.0764	0.0942
Modelo 8	0.3800	0.4600	0.5400	0.3400	0.5800	0.0842	0.0765	0.0914	0.0776	0.1195

Tabela 11 – Tabela de comparação das medidas RPS e de Finetti h - passos (25º Rodada)

Modelos	**Medida**									
	RPS					**de Finetti**				
	Ano									
	2014	**2015**	**2016**	**2017**	**2018**	**2014**	**2015**	**2016**	**2017**	**2018**
Modelo 1	0.2069	**0.2320**	0.1918	0.2322	0.1905	0.58513	0.6485	0.5711	0.679121	0.5805
Modelo 2	0.2064	0.2327	0.1910	0.2303	0.1910	**0.58385**	**0.6500**	0.5689	0.676314	0.5818
Modelo 3	0.2076	0.2327	0.1929	0.2330	0.1919	0.58663	0.6506	0.5732	0.68024	0.5835
Modelo 4	0.2065	0.2323	0.1907	**0.2302**	**0.1900**	0.58412	0.6489	0.5687	0.676144	**0.5796**
Modelo 5	0.2077	0.2325	0.1922	0.2328	0.1910	0.58692	0.6502	0.5720	0.67983	0.5815
Modelo 6	**0.2043**	0.2328	**0.2011**	0.2330	0.1964	**0.58090**	0.6463	0.5914	**0.683185**	0.6013
Modelo 7	0.2106	0.2408	0.1914	0.2367	0.2066	0.58713	0.6654	0.5718	0.687541	0.6159
Modelo 8	0.2109	0.2349	**0.1880**	0.2342	0.1933	0.59570	0.6566	**0.5657**	0.681796	0.5841
Palpite Bra	0.2260	**0.2220**	0.2094	**0.2205**	0.2054	0.61934	**0.6223**	0.6082	**0.652581**	0.6150

Tabela 12 – Tabela de comparação das medidas RPS e de Finetti h - passos (33º Rodada)

Modelos	**Medida**									
	RPS					**de Finetti**				
	Ano									
	2014	**2015**	**2016**	**2017**	**2018**	**2014**	**2015**	**2016**	**2017**	**2018**
Modelo 1	0.2224	0.2408	0.1838	0.2421	0.1681	0.6379	0.6756	0.5696	0.7014	0.5272
Modelo 2	0.2220	0.2402	0.1821	0.2390	0.1675	**0.6371**	0.6740	0.5649	0.6965	0.5262
Modelo 3	0.2230	0.2432	0.1861	0.2444	0.1688	0.6395	0.6812	0.5748	0.7050	0.5286
Modelo 4	0.2222	0.2397	**0.1807**	0.2386	**0.1670**	0.6372	0.6730	**0.5623**	0.6957	**0.5251**
Modelo 5	0.2234	0.2430	0.1845	0.2437	0.1683	0.6400	0.6809	0.5715	0.7037	0.5276
Modelo 6	**0.1971**	**0.2358**	0.1862	**0.2186**	0.1855	**0.5802**	**0.6576**	0.5791	**0.6491**	0.5584
Modelo 7	0.2155	0.2471	0.1928	0.2400	0.1851	0.6188	0.6866	0.5981	0.6972	0.5648
Modelo 8	0.2306	0.2424	0.1810	0.2373	0.1681	0.6564	0.6792	0.5637	0.6915	0.5287
Palpite Bra	0.2383	**0.2161**	0.2000	**0.2000**	0.1941	0.6593	**0.6150**	0.6116	**0.6116**	0.5806

Anexos