MODELAGEM ESTATÍSTICA E DE APRENDIZADO DE MÁQUINA: PREVISÃO DO CAMPEONATO BRASILEIRO SÉRIE A 2017

Hugo Ribeiro Santana
Universidade Federal da Bahia
hugo.nabity@hotmail.com

Francisco Louzada
Universidade de São Paulo
louzada@icmc.usp.br

Paulo Henrique Ferreira
Universidade Federal da Bahia
paulohenri@ufba.br

Adriano Kamimura Suzuki
Universidade de São Paulo
suzuki@icmc.usp.br

Anderson Ara
Universidade Federal da Bahia
anderson.ara@ufba.br

RESUMO

Prever o resultado de uma partida de futebol depende de diversos fatores um tanto inesperados. Inúmeras vezes, equipes com grandes investimentos e jogadores de alto nível são derrotadas por equipes consideradas menores e que não possuem jogadores tão renomados. É de interesse de muitos, saber quantos gols serão marcados numa partida ou ainda, ser capaz de acertar qual será o placar final de um jogo. Este trabalho está dividido em duas partes. A primeira parte refere-se a uma modelagem de classificação atemporal e utiliza as técnicas de regressão logística politômica, máquinas de vetores suporte e florestas aleatórias para estimar as probabilidades de cada resultado possível (Vitória Mandante, Empate ou Vitória Visitante) dos jogos do Campeonato Brasileiro de Futebol Série A 2017. Nota-se, no entanto, que o modelo obtido utilizando a técnica de florestas aleatórias apresentou previsões de má qualidade (Brier Score $= 0,7706$). A segunda parte remete a uma modelagem de regressão temporal, com o intuito de estimar o número de gols marcados por cada equipe numa partida, mais precisamente, de predizer qual será o placar final desse jogo, considerando a dependência serial dos jogos e a superdispersão, características comumente presentes em dados de futebol. O modelo Poisson Auto-Regressivo com Covariáveis Exógenas (PARX) foi utilizado para modelar o número de gols marcados pelas equipes. Observa-se que os modelos PARX apresentaram resultados satisfatórios, de acordo com diagnósticos feitos com o auxílio dos gráficos PIT e de calibração marginal. Os resultados preditivos foram similares aos obtidos com a aplicação das duas técnicas de classificação (regressão logística politômica e máquinas de vetores suporte).

ABSTRACT

Predicting the outcome of a soccer match depends on many factors that are somewhat unexpected. Often, teams with high investments and high-level players are defeated by teams considered smaller and with players less known. It is in the interest of many people to know how many goals will be scored in a match, as well as to be able to guess what the final score will be in a game. This paper is divided into two parts. The first one refers to (timeless) classification modeling using the multinomial logistic regression, support vector machines and random forests techniques, in order to estimate the probabilities of each possible outcome (home team win, draw or away team win) of the matches of the Brazilian Series A Championship 2017, also known as Campeonato Brasileiro or Brasileirão 2017. We noticed, however, that for this season, the classification model based on the random forests method produced poor predictions (Brier Score $= 0.7706$). The second part of this work refers to time series regression modeling using the Poisson autoregression with exogenous covariates (PARX) model, in order to estimate the number of goals scored by each team in a match. More precisely, to predict the final score of that match considering the serial dependence of the games and the overdispersion, which are characteristics commonly found in football data. It was observed that the PARX models presented satisfactory results, according to diagnostics performed with the help of PIT and marginal calibration plots. Besides, their prediction results were similar to those obtained with the application of the two classification techniques (multinomial logistic regression and support vector machines).

Palavras-chave: Aprendizado de Máquina, Distribuição de Poisson, Futebol, Previsão.

1 INTRODUÇÃO

O futebol é um esporte popular em todo o planeta. Presente em vários cantos do mundo, a paixão por esse esporte surge desde tenra idade. Países como Brasil, Alemanha e Itália são as principais escolas mundiais, isto é, as grandes potências do futebol, pois juntos possuem treze títulos de vinte e uma edições da Copa do Mundo. A seleção brasileira de futebol, por sua vez, é a maior vencedora de títulos mundiais com cinco conquistas.

No Brasil, o futebol chegou por intermédio de Charles W. Miller, que trouxe duas bolas ao país no ano de 1894 (Oliveira [1]). Com regras claras e objetivas, o esporte foi conquistando adeptos no país, primeiramente praticado apenas pela elite e, depois, chegando às classes mais pobres. Em 1950, o Brasil sediou a Copa do Mundo, porém a seleção brasileira foi derrotada na decisão do torneio pela seleção do Uruguai (2×1). Nesse jogo, ocorrido no estádio do Maracanã, no Rio de Janeiro, aproximadamente 200.000 pessoas estavam presentes. Essa derrota marcante provocou diversas mudanças no futebol brasileiro que, depois dessa década, começou a revelar grandes craques, dentre os quais, destacam-se Garrincha, Zagallo e Pelé.

O Campeonato Brasileiro de Futebol da Primeira Divisão, mais conhecido como Brasileirão Série A, é o principal torneio entre times de futebol do Brasil. Em 2017, foi considerado como o 3^{o} campeonato nacional de futebol mais difícil do mundo, segundo a Federação Internacional de Histórias e Estatísticas no Futebol (IFFHS, do inglês *"International Federation of Football History & Statistics"*), que adotou como critério o desempenho dos cinco melhores times de cada país em torneios nacionais e internacionais, no período entre 01 de janeiro e 31 de dezembro de 2017. A partir de 2003, o Brasileirão adotou o sistema de pontos corridos, ou seja, todos os clubes participantes se enfrentam entre si, em turno e returno, e a equipe que fizer mais pontos é considerada campeã do torneio. A competição

é organizada pela Confederação Brasileira de Futebol (CBF) e, a partir de 2016, classifica doze equipes para competições continentais. Os seis primeiros colocados, ao final do campeonato, são contemplados com uma vaga na Copa Libertadores da América do ano seguinte, sendo que os quatro primeiros têm vaga assegurada na fase de grupos e o 5^{o} e 6^{o} colocados têm que disputar a fase pré-Libertadores. Além disso, os clubes que terminam entre a 7^{a} e a 12^{a} colocações têm acesso garantido na Copa Sul-Americana do próximo ano; e os quatro últimos colocados são rebaixados à Série B do Campeonato Brasileiro do ano seguinte.

A expressão "o futebol é uma caixinha de surpresas" é utilizada por muitos do meio esportivo. Ela foi empregada pela primeira vez pelo jornalista brasileiro Benjamin Wright. Ele não estava errado quando utilizou tal expressão, afinal, no futebol há um alto grau de aleatoriedade, em que as partidas se decidem nos detalhes. O resultado de uma partida de futebol depende de diversos fatores um tanto inesperados. Inúmeras vezes, equipes com grandes investimentos e jogadores de alto nível são derrotadas por equipes consideradas menores e que não possuem jogadores de renome. É a famosa "zebra" que, segundo muitos especialistas, torna o futebol tão fascinante e imprevisível.

Há diversos trabalhos na literatura que utilizam procedimentos matemáticos, estatísticos ou computacionais para prever resultados de partidas de futebol. Por exemplo, Farias [2] propõe, para o número de gols marcados pelas equipes do Brasileirão 2006, o uso de dois modelos Bayesianos dinâmicos com coeficientes auto-regressivos de evolução, bem como realiza uma comparação com um modelo estático e um dinâmico, propostos por Rue & Salvesen [3] e utilizados por Souza Jr. & Gamerman [4] para analisar dados do Brasileirão 2002 e 2003.

Brillinger [5], por sua vez, modela diretamente as probabilidades de vitória, empate e derrota dos jogos do Brasileirão 2006 considerando um modelo de regressão trinomial, com interesse em modelar a tabela de classificação do campeonato depois que um certo número de rodadas foram concluídas.

Suzuki *et al.* [6] propõem uma metodologia Bayesiana para predição dos jogos da Copa do Mundo 2006, utilizando opiniões de especialistas e ranking da FIFA (sigla do francês *"Fédération Internationale de Football Association"*) na distribuição a priori. O método aplicado permite calcular as probabilidades de vitória, empate e derrota de cada time em cada partida, bem como simular toda a competição a fim de estimar as probabilidades de classificação na fase de grupos, de alcançar as oitavas-de-final, de ser campeão, entre outras.

Alves *et al.* [7] propõem o ajuste de dois modelos logísticos ordinais para prever a chance de uma equipe do Brasileirão 2008 se classificar para a Copa Libertadores da América, ou seja, terminar o campeonato entre os quatro primeiros colocados, ou ainda, a chance dessa equipe ser rebaixada, isto é, terminar entre os quatro piores colocados.

Tavares & Suzuki [8] utilizam dados do Brasileirão 2013 para aplicar uma metodologia baseada no método Soma e Diferença, proposta por Arruda [9], visando calcular as probabilidades de interesse, tais como: time campeão, times rebaixados, melhor time mandante, melhor time visitante, dentre outras. A metodologia apresentada pelos autores assume uma distribuição univariada de Poisson para o número de gols marcados em uma partida. Além disso, considera modelos lineares que expressam a soma e a diferença de gols marcados em função de quatro covariáveis: a média de gols em uma partida, a vantagem do time mandante, o poder ofensivo da equipe e o poder defensivo do adversário.

Angelini & De Angelis [10] aplicam o modelo Poisson Auto-Regressivo com covariáveis exógenas (PARX), proposto por Agosto *et al.* [11], aos dados da Premier League 2013-14 e 2014-15. Tal proposta visa a modelagem em séries temporais de contagens para predizer qual será o placar final de um jogo, levando em consideração a dependência serial dos jogos e a superdispersão, que são características comumente presentes em dados de futebol.

Filho *et al.* [12] assumem que os números de gols marcados pelos times em um jogo são independentes e seguem uma distribuição de Poisson, em que a média reflete a força do

ataque, da defesa e da casa. Com aplicação aos dados da Premier League 2012-13, antes do início de cada rodada do returno foram calculadas as probabilidades de vitória, empate e derrota dos times em cada uma das partidas. Além disso, por meio de um procedimento de simulação, obtiveram a probabilidade de um determinado time sagrar-se campeão, de ser rebaixado para a segunda divisão ou de se classificar para a Liga dos Campeões da UEFA.

O presente trabalho tem por objetivo principal comparar a performance preditiva de algumas técnicas de classificação atemporal (regressão logística politômica, máquinas de vetores suporte e florestas aleatórias) e de regressão temporal (modelo PARX), quando aplicadas aos dados do Campeonato Brasileiro (ou Brasileirão) 2017.

Este artigo está organizado da seguinte forma. A Seção 2 descreve a metodologia (técnicas de aprendizado de máquina e de modelagem estatística) empregada neste trabalho. A Seção 3 apresenta os resultados e discussões após aplicação da metodologia proposta aos dados do Campeonato Brasileiro 2017. Finalmente, a Seção 4 apresenta as considerações finais.

2 Métodos

Nesta seção são apresentadas as técnicas de classificação atemporal (Seção 2.1) e de regressão temporal (Seção 2.2) utilizadas neste trabalho, bem como as medidas empregadas para avaliar a capacidade preditiva da modelagem (Seção 2.3).

2.1 Técnicas de Classificação Atemporal

Nesta subseção são descritos os métodos de classificação supervisionada que foram empregados neste trabalho, a saber: regressão logística politômica (Seção 2.1.1), máquinas de vetores suporte (Seção 2.1.2) e florestas aleatórias (Seção 2.1.3). A primeira técnica foi escolhida por possuir boa interpretabilidade, enquanto as duas últimas por sua maior flexibilidade. Tais técnicas consistem, basicamente, em apresentar um conjunto de exemplos de treinamento previamente conhecidos e rotulados com as respectivas classes para que possa ser construído um modelo preditivo para um novo conjunto de exemplos fornecidos e com classes desconhecidas.

2.1.1 Regressão Logística Politômica

A regressão logística politômica é um modelo de regressão linear aplicado quando a variável dependente é nominal e possui mais de dois níveis (Hosmer & Lemeshow [13]). É a generalização do método de regressão logística binária (Cox [14]).

Neste caso, a variável resposta Y assume três categorias: Vitória Mandante (VM), Empate (E) e Vitória Visitante (VV). Logo, o modelo logístico terá duas funções *logit*: a razão entre $Y = VM$ e $Y = E$ e a razão entre $Y = VV$ e $Y = E$. Neste caso, a categoria $Y = E$ foi assumida como referência. Assim, temos:

$$g_1(\mathbf{x}) = \log\left(\frac{P(Y = VM\,|\,\mathbf{x})}{P(Y = E\,|\,\mathbf{x})}\right) = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1p}x_p,$$

$$g_2(\mathbf{x}) = \log\left(\frac{P(Y = VV\,|\,\mathbf{x})}{P(Y = E\,|\,\mathbf{x})}\right) = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2p}x_p.$$

A partir das funções lineares $g_i(\mathbf{x})$, $i = 1, 2$, cujos parâmetros $\beta_{10}, \cdots, \beta_{1p}$, $\beta_{20}, \cdots, \beta_{2p}$ são estimados por máxima verossimilhança, é possível calcular as probabilidades condicionais de ocorrência de cada categoria da variável resposta Y dado um vetor de variáveis explicativas (ou covariáveis) x (Bittencourt [15]), conforme segue:

$$P(Y = E\,|\,\mathbf{x}) = \frac{1}{1 + \exp\{g_1(\mathbf{x})\} + \exp\{g_2(\mathbf{x})\}}, \tag{1}$$

$$P(Y = VM\,|\,\mathbf{x}) = \frac{\exp\{g_1(\mathbf{x})\}}{1 + \exp\{g_1(\mathbf{x})\} + \exp\{g_2(\mathbf{x})\}}, \tag{2}$$

$$P(Y = VV\,|\,\mathbf{x}) = \frac{\exp\{g_2(\mathbf{x})\}}{1 + \exp\{g_1(\mathbf{x})\} + \exp\{g_2(\mathbf{x})\}}. \tag{3}$$

As probabilidades descritas em (1), (2) e (3) são utilizadas para se estabelecer uma regra para discriminação das classes. A regra de classificação para alocar uma observação $\mathbf{x}^*$ numa das classes é:

$$\begin{aligned}
\mathbf{x}^* \in E \quad &\text{se} \quad P(Y = E\,|\,\mathbf{x}^*) > [P(Y = VM\,|\,\mathbf{x}^*) \text{ e } P(Y = VV\,|\,\mathbf{x}^*)],\\
\mathbf{x}^* \in VM \quad &\text{se} \quad P(Y = VM\,|\,\mathbf{x}^*) > [P(Y = E\,|\,\mathbf{x}^*) \text{ e } P(Y = VV\,|\,\mathbf{x}^*)],\\
\mathbf{x}^* \in VV \quad &\text{se} \quad P(Y = VV\,|\,\mathbf{x}^*) > [P(Y = E\,|\,\mathbf{x}^*) \text{ e } P(Y = VM\,|\,\mathbf{x}^*)].
\end{aligned}$$

Para a realização de testes de significância para os parâmetros β_{ij}, $i = 1, 2$ e $j = 0, 1, \ldots, p$, pode ser utilizada a estatística de *Wald*, definida como o quadrado da razão entre a estimativa de máxima verossimilhança para o parâmetro e a estimativa do respectivo erro-padrão [$\widehat{\text{EP}}(\hat{\beta}_{ij})$]. Neste teste, a hipótese nula é de que o valor de um parâmetro específico é zero, isto é, $H_0 : \beta_{ij} = 0$. Assintoticamente, essa estatística (W) converge em distribuição para uma variável aleatória de qui-quadrado com 1 grau de liberdade, isto é,

$$W = \left(\frac{\hat{\beta}_{ij}}{\widehat{\text{EP}}(\hat{\beta}_{ij})}\right)^2 \xrightarrow{D} \chi^2_1.$$

A regressão logística politômica (RLP) foi aplicada aos dados do Brasileirão 2017 no *software* estatístico R (R Core Team [16]) utilizando a função *multinom(.)* do pacote "nnet" (Venables e Ripley [17]).

2.1.2 MÁQUINAS DE VETORES SUPORTE

O método de Máquinas de Vetores Suporte (SVM, do inglês "*support vector machines*") foi prosposto por Vapnik (Vapnik & Lerner [18]; Cortes & Vapnik [19]), sendo um método não-probabilístico direcionado à classificação.

Neste método, Y assume valores em $C = \{-1, +1\}$, em que $+1$ representa uma classe e -1 representa as outras duas classes. Diante da suposição que é possível construir um hiperplano que separe perfeitamente todas as observações do conjunto de treinamento segundo a(s) classe(s) a que pertencem, um hiperplano separador possui a seguinte propriedade:

$$\begin{aligned}
\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} > 0 \quad &\text{se} \quad y_i = +1,\\
\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} < 0 \quad &\text{se} \quad y_i = -1.
\end{aligned}$$

O conceito por trás deste método está em decidir qual é o hiperplano separador que será utilizado. A escolha mais simples é o hiperplano de margem máxima (ou hiperplano ótimo), sendo este o hiperplano criado entre a maior distância entre as observações de treinamento. Essa margem é definida unicamente pelos vetores suporte, que são os exemplos da base de treinamento mais próximos do hiperplano. Na ilustração da Figura 1, observa-se que três observações de treino são equidistantes do hiperplano de margem máxima e situam-se ao longo das linhas tracejadas, indicando a largura da margem.

Quando o conjunto de dados não é linearmente separável, a solução é projetar os dados em um espaço de maior dimensão, denominado espaço característico (ou ainda, espaço de características), no qual com alta probabilidade os dados são linearmente separáveis. Para isso, é necessário que a dimensão do espaço característico seja suficientemente alta e que a transformação seja não-linear. As funções associadas a essas transformações são chamadas de funções de mapeamento ϕ. Tal transformação é ilustrada na Figura 2.

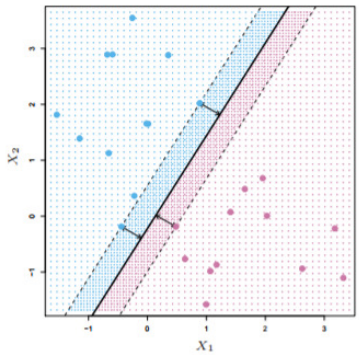

FIGURA 1: Ilustração de um hiperplano de margem máxima em $\mathbb{R}^2$. Fonte: James *et al.* [20].

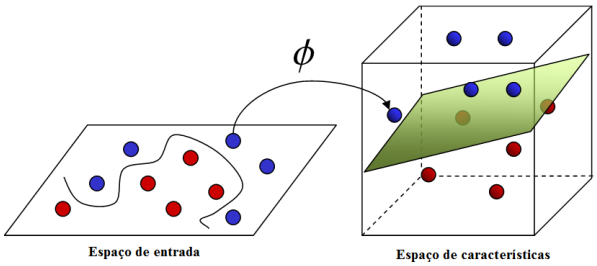

FIGURA 2: À esquerda, um conjunto não-separável linearmente é projetado em um espaço de maior dimensão (espaço de características), à direita, por meio de uma função de mapeamento ϕ. Fonte: Nalini & Palanivel [21].

De uma forma geral, um *kernel* K é uma função que recebe dois pontos $\mathbf{x}_i$ e $\mathbf{x}_j$ do espaço de entradas e calcula o produto escalar no espaço de características (Lorena & Carvalho [22]). Tem-se então:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j).$$

Os *kernels* mais utilizados na prática são listados na Tabela 1. Sendo que o *kernel* Gaussiano (ou RBF, do inglês *"radial-basis function"*) foi aplicado com sucesso por Vlastakis *et al.* [23].

TABELA 1: Funções *Kernel* mais comuns. Fonte: adaptado de Lorena & Carvalho [22].

Tipo de *Kernel*	Função $K(\mathbf{x}_i, \mathbf{x}_j)$	Parâmetro
Polinomial	$(\delta(\mathbf{x}_i \cdot \mathbf{x}_j) + \kappa)^d$	δ, κ e d
Gaussiano	$\exp\left\{-\eta\|\|\mathbf{x}_i - \mathbf{x}_j\|\|^2\right\}$	η
Sigmoidal	$\tanh\left(\delta(\mathbf{x}_i \cdot \mathbf{x}_j) + \kappa\right)$	δ e κ

Como a variável resposta assume três níveis, ajusta-se então três SVMs, cada vez comparando uma das três classes com as outras duas remanescentes. A partir daí, sejam $\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk}$ os parâmetros resultantes de um ajuste SVM comparando a k-ésima classe (codificada como $+1$) com as outras (codificadas como -1) e $\mathbf{x}^* = \left(1, x_1^*, \ldots, x_p^*\right)'$ uma observação-teste. Atribui-se a observação à classe para a qual $\beta_{0k} + \beta_{1k}x_1^* + \cdots + \beta_{pk}x_p^*$ é maior, pois isso equivale a um alto nível de confiança que a observação de teste pertence à k-ésima classe e não a qualquer outra das classes. Essa abordagem é conhecida como *"one-versus-all"* (i.e. um-contra-todos) (James *et al.* [20]) e é ilustrada na Figura 3.

A técnica SVM foi aplicada aos dados do Brasileirão 2017 no *software* R utilizando a função *svm(.)* do pacote "e1071" (David [24]).

2.1.3 FLORESTAS ALEATÓRIAS

A Floresta Aleatória (RF, do inglês *"random forest"*) foi proposta por Breiman [25] e consiste em uma técnica de agregação de classificadores do tipo árvore, construídos de

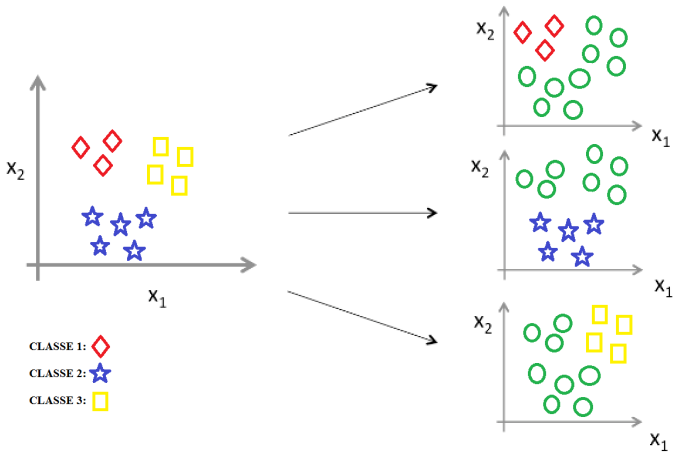

FIGURA 3: Ilustração da abordagem *"one-versus-all"*.

forma que a sua estrutura seja composta de maneira aleatória (Ghosal *et al.* [26]), ou seja, trata-se de uma floresta (conjunto de árvores preditoras) em que cada árvore necessita de um vetor aleatório amostrado independentemente e todas as árvores da floresta têm a mesma distribuição.

Cada nó da árvore representa um teste e as bordas representam os possíveis resultados do teste. Para classificar uma amostra com tal árvore, é apresentado o primeiro nó. O resultado do teste determina qual será o nó subsequente, e assim por diante, até o exemplo chegar a uma folha. A partir daí, rotula-se a amostra com o rótulo (ou classe) mais comum na folha. Esse objeto é conhecido como uma árvore de decisão (ou ainda, árvore de classificação).

A Figura 4 mostra um conjunto de dados bidimensionais (2D) simples com três classes, ao lado de uma árvore de decisão que classificará corretamente pelo menos os dados de treinamento.

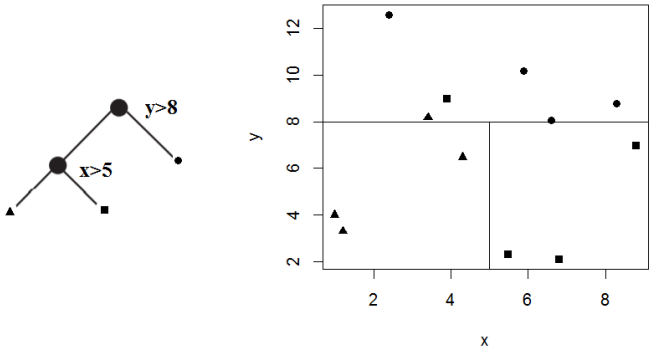

FIGURA 4: Árvore de decisão ilustrada de duas maneiras. À esquerda, são dadas diretamente as regras de cada decisão; à direita, são mostrados os pontos em duas dimensões e a estrutura que a árvore produz no espaço.

Para determinar a classe de uma instância, um conjunto de árvores passa por um mecanismo de votação (*bagging*). A classe que obtém o maior número de votos ganha.

O método RF foi aplicado aos dados do Brasileirão 2017 no *software* R mediante o uso da função *randomForest(.)* do pacote "randomForest" (Liaw & Wiener [27]).

2.2 TÉCNICAS DE REGRESSÃO TEMPORAL

Muitas vezes, o interesse na modelagem de dados de futebol reside em saber quantos gols serão marcados numa partida. Ou ainda, ser capaz de acertar qual será o placar final de um jogo. Boldrin [28] desenvolveu um modelo baseado na distribuição de Poisson para

prever os resultados e as melhores seleções de apostas da Premier League na temporada 2016-17.

Nenhum dos métodos/modelos apresentados na Seção 2.1 é direcionado a prever os placares dos jogos, neste sentido estimam as probabilidades de vitória, empate ou derrota dos times. Por isso, nesta subseção é descrita uma técnica de modelagem estatística capaz de estimar o número de gols marcados por cada equipe numa partida, mais precisamente, de predizer qual será o placar final desse jogo. Essa técnica, que pode ser usada para modelar séries temporais de contagem, permite considerar a dependência serial dos jogos e a superdispersão, que são características comumente presentes em dados de futebol. Trata-se do modelo PARX, proposto por Agosto *et al.* [11] e aplicado aos dados da Premier League 2013-14 e 2014-15 por Angelini & De Angelis [10]. O modelo é descrito a seguir.

2.2.1 MODELO PARX

Para encontrar as probabilidades associadas a cada possível resultado de um jogo de futebol, as principais características da distribuição dos gols marcados por um time devem ser levadas em consideração: a dependência com jogos anteriores e a sobredispersão (ou superdispersão). Os modelos PARX permitem o ajuste a dados que mostram dependência serial e são capazes de capturar a sobredispersão em distribuições marginais. Além disso, permitem incluir covariáveis exógenas na especificação do modelo, no intuito de melhorar a precisão dos placares previstos.

Se Y_t denota o número de gols marcados por um time no instante t, em que $t = 1, 2, \cdots, T$, o modelo PARX com intensidade $\lambda_t > 0 \ \forall t$ pode ser especificado como:

$$Y_t | F_{t-1} \sim \texttt{Poisson}(\lambda_t), \tag{4}$$

$$\lambda_t = \omega + \sum_{i=1}^{p} \alpha_i \lambda_{t-i} + \sum_{i=1}^{q} \beta_i y_{t-i} + \gamma \mathbf{x}_{t-1}, \tag{5}$$

em que F_{t-1} é o conjunto de informações disponíveis no tempo $(t-1)$, isto é, $F_{t-1} = \{y_{t-m}, \mathbf{x}_{t-m} : m \geq 1\}$ e $\mathbf{x}_{t-1}$ denota um vetor de m covariáveis exógenas (não-negativas). Os parâmetros $\omega > 0$ e $\alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q, \gamma \geq 0$ são invariantes no tempo e asseguram que a distribuição de $Y_t | F_{t-1}$ seja uma Poisson não degenerada ($\lambda_t \neq 0$) e com variância não negativa ($\lambda_t > 0$). Mais especificamente, quando $\gamma = 0$, o modelo PARX se reduz ao modelo PAR considerado em Fokianos *et al.* [29].

Agosto *et al.* [11] afirma que, ao incluir valores passados da resposta, bem como covariáveis em $\mathbf{x}_{t-1}$, os modelos PARX são capazes de gerar uma sobredispersão na distribuição marginal, característica que é predominante em muitas séries temporais de contagem. Logo, $Var[Y_t] \geq E[Y_t]$.

O modelo PARX foi ajustado aos dados do Brasileirão 2014-17 no *software* R utilizando a função *tsglm(.)* do pacote "tscount" (Liboschik *et al.* [30]).

2.2.2 ESTIMAÇÃO, SELEÇÃO DE MODELOS E DIAGNÓSTICO

Considere o modelo para Y_t, especificado em (4)-(5), como:

$$\lambda_t(\boldsymbol{\theta}) = \omega + \sum_{i=1}^{p} \alpha_i \lambda_{t-i}(\boldsymbol{\theta}) + \sum_{i=1}^{q} \beta_i y_{t-i} + \gamma \mathbf{x}_{t-1},$$

em que $\boldsymbol{\theta} = (\omega, \alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q, \gamma)'$.

O logaritmo da função de verossimilhança condicional é dado por:

$$\ell_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} y_t \log \lambda_t(\boldsymbol{\theta}) - \lambda_t(\boldsymbol{\theta}),$$

em que foram omitidos os termos constantes.

O estimador de máxima verossimilhança (EMV) de $\boldsymbol{\theta}$ é dado por:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ \ell_T(\boldsymbol{\theta}). \tag{6}$$

O problema de maximização da Equação (6) está sujeito às seguintes restrições:

1. $\omega > 0$, $\alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q, \gamma \geq 0$, para garantir que $\lambda_t > 0$;
2. $\sum_{i=1}^{\max\{p,q\}} (\alpha_i + \beta_i) < 1$, para garantir que o processo seja estacionário (estacionariedade fraca ou de segunda ordem).

As duas restrições anteriores são discutidas com mais detalhes em Agosto *et al.* [11].

A seleção dos parâmetros p e q do modelo PARX pode ser realizada de acordo com um critério de informação. Um dos critérios de informação mais utilizados é o AIC (do inglês "*Akaike information criterion*"), definido por Akaike [31] como:

$$AIC(\hat{\boldsymbol{\theta}}) = 2(c - \ell_T(\hat{\boldsymbol{\theta}})),$$

em que c denota o número de parâmetros do modelo. Deve-se selecionar o modelo cujo critério de informação calculado seja mínimo.

Uma ferramenta para avaliar a qualidade do ajuste é a transformação integral de probabilidade (PIT, do inglês "*probability integral transform*"), a qual utiliza a distribuição uniforme contínua (0,1) para dizer se a distribuição preditiva está correta. Para dados de contagem, Czado *et al.* [32] propõem uma versão não-aleatorizada da PIT, definida por:

$$F_t(u|y) = \begin{cases} 0, & \text{se } u \leq P_t(y-1), \\ \frac{u - P_t(y-1)}{P_t(y) - P_t(y-1)}, & \text{se } P_t(y-1) < u < P_t(y), \\ 1, & \text{se } u \geq P_t(y), \end{cases}$$

em que $P_t(y) = P(Y_t \leq y | F_{t-1})$ é a função de distribuição acumulada.

A PIT média é dada por:

$$\bar{F}(u) = \frac{1}{n} \sum_{t=1}^{n} F_t(u|y), \quad 0 \leq u \leq 1,$$

em que n é o número de jogos analisados.

Para comparar $\bar{F}(u)$ com a função de distribuição acumulada da uniforme, Czado *et al.* [32] propõem traçar um histograma PIT não-aleatorizado, que pode ser interpretado de forma diagnóstica.

Outra ferramenta para diagnóstico é o diagrama de calibração marginal. Para a sua construção, é preciso definir o que Gneiting *et al.* [33] propõem como a calibração marginal (CL): a diferença entre a função de distribuição acumulada preditiva média e a função de distribuição acumulada empírica das observações, ou seja,

$$\mathtt{CL} = \frac{1}{n} \sum_{t=1}^{n} P_t(y) - \frac{1}{n} \sum_{t=1}^{n} I(y_t \leq y).$$

Na prática, traçamos a calibração marginal para os valores y no intervalo das observações originais. Valores discrepantes no diagrama sugerem motivos para falhas de previsão e deficiências do modelo.

No entanto, vale ressaltar que a calibração avaliada por um histograma PIT ou um gráfico de calibração marginal é uma condição necessária, mas não suficiente, para que um modelo preditivo seja ideal (maiores detalhes em Gneiting *et al.* [33]).

2.2.3 PREDIÇÃO

Para prever os placares das partidas são estimados dois modelos PARX: um para a equipe mandante e outro para a equipe visitante. Em termos de notação, M irá denotar a equipe mandante e V denotará a equipe visitante, em uma determinada partida. Daí, temos que a distribuição do número de gols marcados pela equipe mandante jogando em casa segue uma Poisson (condicional), ou seja, $Y_t^{\mathrm{M}}|F_{t-1} \sim \texttt{Poisson}(\lambda_t^{\mathrm{M}})$, assim como a distribuição do número de gols marcados pela equipe visitante jogando fora de casa segue uma Poisson (condicional) diferente, isto é, $Y_t^{\mathrm{V}}|F_{t-1} \sim \texttt{Poisson}(\lambda_t^{\mathrm{V}})$. Essas distribuições são estimadas usando o modelo PARX, definido nas Equações (4)-(5).

As previsões segundo o modelo PARX mostram-se muito semelhantes às obtidas por meio dos modelos GARCH (Hansen *et al.* [34]). No caso do futebol, o interesse está em prever o número de gols marcados por uma equipe na próxima partida dado as informações disponíveis no tempo T, ou seja, $y_{T+1}^j|F_T$, para $j = \mathrm{M}, \mathrm{V}$. Portanto, é necessário calcular o parâmetro da distribuição de Poisson (condicional) associada, λ_{T+1}^j, para obter uma previsão do número de gols na próxima partida. Esse valor é obtido por meio do processo a seguir, que leva em consideração as informações disponíveis no tempo T e o vetor de parâmetros $\boldsymbol{\theta}$:

$$\lambda_{T+1|T}^j(\boldsymbol{\theta}) = \omega + \sum_{i=1}^{p} \alpha_i \lambda_{T+1-i}^j(\boldsymbol{\theta}) + \sum_{i=1}^{q} \beta_i y_{T+1-i}^j + \gamma \mathbf{x}_T, \quad j = \mathrm{M}, \mathrm{V}.$$

A partir daí, pode-se encontrar a distribuição de $y_{T+1}^j|F_T$:

$$\hat{P}(Y_{T+1}^j = y^j|F_T) = \frac{\lambda_{T+1|T}^j(\boldsymbol{\theta})^{y^j} \exp\{-\lambda_{T+1|T}^j(\boldsymbol{\theta})\}}{y^j!}, \quad y^j \in \{0, 1, 2, \cdots\}.$$

Assumindo independência entre as distribuições previstas das equipes mandante e visitante, obtém-se a distribuição conjunta da seguinte forma:

$$\hat{P}(Y_{T+1}^{\mathrm{M}} = y^{\mathrm{M}}, Y_{T+1}^{\mathrm{V}} = y^{\mathrm{V}}|F_T) = \hat{P}(Y_{T+1}^{\mathrm{M}} = y^{\mathrm{M}}|F_T) \cdot \hat{P}(Y_{T+1}^{\mathrm{V}} = y^{\mathrm{V}}|F_T).$$

A expressão anterior indica qual a probabilidade do resultado final da partida ter o placar de y^{M} gols para a equipe mandante e y^{V} gols para a equipe visitante. Por exemplo, para saber qual a probabilidade de um determinado jogo terminar sem gols, basta fazer $\hat{P}(Y_{T+1}^{\mathrm{M}} = 0, Y_{T+1}^{\mathrm{V}} = 0|F_T)$.

A respeito das covariáveis que são incluídas no modelo, Angelini & De Angelis [10] sugerem que seja a média dos gols concedidos pela equipe adversária devido ao fato de que uma equipe com boa defesa tende a conceder menos gols do que uma equipe com uma defesa ruim. Esta foi a única covariável considerada no modelo, logo, $\mathbf{x}_{t-1} = x_{t-1}$.

2.3 AVALIAÇÃO DA CAPACIDADE PREDITIVA

Após ajustar um modelo preditivo, deve-se avaliar quantitativamente o desempenho do mesmo, isto é, analisar a qualidade das previsões realizadas pelo modelo. Nesta subseção são apresentadas as principais medidas da capacidade preditiva que podem ser obtidas/calculadas a partir da chamada matriz de confusão (Seção 2.3.1), bem como uma métrica geral de desempenho conhecida como Brier Score (Seção 2.3.2).

2.3.1 MEDIDAS DE PERFORMANCE

A capacidade preditiva de um modelo de classificação pode ser avaliada computando o número de observações corretamente reconhecidas como pertencentes à classe (verdadeiros-positivos), o número de observações corretamente reconhecidas que não pertencem à classe (verdadeiros-negativos) e observações que foram incorretamente atribuídas à classe (falsos-positivos) ou que não foram reconhecidas como pertencentes à classe (falsos-negativos). A

Tabela 2, chamada de matriz de confusão, é uma forma simples de se estabelecer e visualizar o cálculo dessas medidas.

TABELA 2: Matriz de confusão para classificação binária.

	Real	
Predição	Positivo	Negativo
Positivo	Verdadeiro-Positivo (vp)	Falso-Positivo (fp)
Negativo	Falso-Negativo (fn)	Verdadeiro-Negativo (vn)

Seja k o número de classes. No caso do futebol, em que $k = 3$, são construídas três matrizes de confusão para classificação binária:

1. Vitória Mandante em relação à Não-Vitória Mandante;
2. Empate em relação à Não-Empate;
3. Vitória Visitante em relação à Não-Vitória Visitante.

A Tabela 3 exemplifica como seriam essas matrizes de confusão para classificações binárias no futebol.

TABELA 3: Exemplos da matriz de confusão para classificações binárias no futebol.

	Real	
Predição	VM	Não-VM
VM	vp_1	fp_1
Não-VM	fn_1	vn_1

	Real	
Predição	E	Não-E
E	vp_2	fp_2
Não-E	fn_2	vn_2

	Real	
Predição	VV	Não-VV
VV	vp_3	fp_3
Não-VV	fn_3	vn_3

A partir daí, as medidas de performance podem ser obtidas. Dentre elas, merecem destaque: precisão média, taxa de erro, micro precisão, micro recall, micro F-score, macro precisão, macro recall e macro F-score. Tais medidas, propostas por Sokolova & Lapalme [35], são calculadas como segue:

- **Precisão Média:** A eficácia média por classe, de um classificador, é definida por:

$$\texttt{PM} = \frac{\sum_{i=1}^{k} \frac{vp_i + vn_i}{vp_i + fn_i + fp_i + vn_i}}{k}.$$

- **Taxa de Erro:** O erro médio de classificação por classe é definido por:

$$\texttt{TE} = \frac{\sum_{i=1}^{k} \frac{fp_i + fn_i}{vp_i + fn_i + fp_i + vn_i}}{k}.$$

- **Micro Precisão:** A precisão (também chamada de valor preditivo positivo), considerando a frequência de cada rótulo, é definida por:

$$\texttt{Precisão}_{\mu} = \frac{\sum_{i=1}^{k} vp_i}{\sum_{i=1}^{k} (vp_i + fp_i)}.$$

- **Micro Recall:** A revocação (*"recall"* em inglês, também conhecida como sensibilidade), considerando a frequência de cada rótudo, é definida por:

$$\texttt{Recall}_{\mu} = \frac{\sum_{i=1}^{k} vp_i}{\sum_{i=1}^{k} (vp_i + fn_i)}.$$

- **Micro F-Score:** A média ponderada da precisão e revocação é definida por:

$$\texttt{FS}_{\mu} = \frac{(\psi^2+1)(\texttt{Precisão}_{\mu} \cdot \texttt{Recall}_{\mu})}{\psi^2\ \texttt{Precisão}_{\mu} + \texttt{Recall}_{\mu}}.$$

  O parâmetro ψ determina o peso da precisão na pontuação combinada. Mais comumente, considera-se $\psi = 1$, que foi o valor utilizado neste trabalho.

- **Macro Precisão:** A média aritmética das precisões parciais de cada classe é definida por:

$$\texttt{Precisão}_{M} = \frac{\sum_{i=1}^{k} \frac{vp_i}{vp_i+fp_i}}{k}.$$

- **Macro Recall:** A eficácia média por classe de um classificador, para identificar rótulos de classe, é definida por:

$$\texttt{Recall}_{M} = \frac{\sum_{i=1}^{k} \frac{vp_i}{vp_i+fn_i}}{k}.$$

- **Macro F-Score:** A relação entre os rótulos positivos dos dados e aqueles obtidos por um classificador, com base em uma média por classe, é definida por:

$$\texttt{FS}_{M} = \frac{(\psi^2+1)(\texttt{Precisão}_{M} \cdot \texttt{Recall}_{M})}{\psi^2\ \texttt{Precisão}_{M} + \texttt{Recall}_{M}}.$$

  Mais comumente, considera-se $\psi = 1$.

As medidas que envolvem as palavras "precisão" e/ou "revocação" (Precisão Média, Micro Precisão, Micro Recall, Micro F-Score, Macro Precisão, Macro Recall e Macro F-Score), quanto maiores melhor a qualidade do modelo preditivo. Por outro lado, a medida Taxa de Erro é interpretada da seguinte forma: quanto menor, melhor o poder preditivo.

2.3.2 BRIER SCORE

Proposto por Brier [36], o Brier Score serve para medir a precisão das previsões probabilísticas. Quando aplicado às previsões de várias classes, o Brier Score é definido como:

$$\texttt{BS} = \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{k}(p_{ti} - o_{ti})^2,$$

em que k é o número de classes, n é o número de instâncias, p_{ti} é a probabilidade prevista da t-ésima instância pertencer à i-ésima classe, e o_{ti} é 1 se a classe atual y_t é igual a i, e 0 se a classe y_t é diferente de i.

No exemplo do futebol, em que $k = 3$, tem-se:

p_{t1} é a probabilidade de vitória da equipe mandante na t-ésima partida;
p_{t2} é a probabilidade de empate na t-ésima partida;
p_{t3} é a probabilidade de vitória da equipe visitante na t-ésima partida;
$o_{t1} = 1$ se a equipe mandante venceu a t-ésima partida e 0 caso contrário;
$o_{t2} = 1$ se a t-ésima partida terminou empatada e 0 caso contrário;
$o_{t3} = 1$ se a equipe visitante venceu a t-ésima partida e 0 caso contrário.

Logo, em cada partida (t) em que alguma equipe mandante vencer, a contribuição ao Brier Score é dada por:

$$[P(VM)_t - 1]^2 + [P(E)_t - 0]^2 + [P(VV)_t - 0]^2.$$

Analogamente, a contribuição ao Brier Score de cada partida que terminar empatada é dada por:

$$[P(VM)_t - 0]^2 + [P(E)_t - 1]^2 + [P(VV)_t - 0]^2.$$

E finalmente, a contribuição ao Brier Score de cada partida em que o visitante vencer é dada por:

$$[P(VM)_t - 0]^2 + [P(E)_t - 0]^2 + [P(VV)_t - 1]^2.$$

Com isso, é fácil ver que o Brier Score é uma medida que pode assumir valores entre 0 e 2. Será 0 (zero) quando o previsor atribuir sempre probabilidades 1 para o resultado e 0 para os demais, e o resultado observado for aquele a que se atribuiu probabilidade 1. E será 2 (dois) se sempre for observado algum dos outros dois resultados (Arruda [9]).

Em termos de previsão, pode-se adotar um valor de referência útil para a medida Brier Score. Trata-se da atribuição equiprovável de probabilidades: $P(VM)_t = P(E)_t = P(VV)_t = 1/3$, isto é, atribui-se a mesma probabilidade a cada resultado da partida. Com essa atribuição, o Brier Score é sempre igual a $(1/3-1)^2 + (1/3-0)^2 + (1/3-0)^2 = 2/3$. Isso significa dizer que o mínimo que se espera de um bom modelo estatístico ou computacional é que sua medida Brier Score seja menor do que $2/3$. Por outro lado, são considerados modelos com previsões de má qualidade aqueles que apresentam medidas maiores que $2/3$.

Pode-se ainda utilizar a medida de número de acertos, ou seja, considera-se que um método acerta o resultado de um jogo quando o resultado com maior probabilidade estimada coincide com o verdadeiro resultado dessa partida. Por exemplo, se em um determinado jogo o time visitante vencer e a probabilidade estimada para a vitória do time visitante for maior que as probabilidades de vitória do time mandante e de empate, considera-se um acerto para essa previsão, caso contrário, um erro. O mesmo vale para quando acontecer empate e vitória do time mandante. No entanto, vale ressaltar que não é correto associar a eventos com alta probabilidade uma certeza de ocorrência, ou a eventos com baixa probabilidade uma certeza de não-ocorrência. Aqui, o objetivo não é afirmar que uma dada previsão é correta ou não, mas sim construir uma métrica para o conjunto das previsões (Filho *et al.* [12]).

3 BRASILEIRÃO 2017

Nesta seção é feita uma aplicação da metodologia descrita na Seção 2 a um banco de dados referente ao Campeonato Brasileiro da Série A de 2017, ou simplesmente, Brasileirão 2017. Na Tabela 4 são apresentadas as vinte equipes participantes, bem como as respectivas siglas utilizadas na base de dados e no decorrer deste trabalho.

TABELA 4: Equipes participantes do Brasileirão 2017.

Equipe	Sigla	Equipe	Sigla	Equipe	Sigla	Equipe	Sigla
Atético-GO	ACG	Botafogo	BOT	Flamengo	FLA	Santos	SAN
Atlético-MG	CAM	Chapecoense	CHA	Fluminense	FLU	São Paulo	SAO
Atlético-PR	CAP	Corinthians	COR	Grêmio	GRE	Sport	SPT
Avaí	AVA	Coritiba	CTB	Palmeiras	PAL	Vasco da Gama	VAS
Bahia	BAH	Cruzeiro	CRU	Ponte Preta	PON	Vitória	VIT

Para o levantamento dos dados referentes ao Brasileirão 2017, foi necessário acessar o site "ogol", disponível gratuitamente em: http://www.ogol.com.br, onde estão os resultados de todas as partidas disputadas. Vinte equipes participaram desta edição do campeonato, totalizando 380 jogos. As variáveis coletadas estão listadas na Tabela 5. Outras variáveis potencialmente relevantes, como crise política e/ou financeira, jogadores lesionados ou poupados, condições meteorológicas e árbitro da partida, não foram consideradas neste trabalho. Isso será objeto de trabalhos futuros.

Houve ainda a inclusão de mais uma variável: *V31 - Palpite FNV*, que se trata de palpites de um site destinado ao esporte (futebol), chamado "Futebol na Veia" (http://www.futebolnaveia.com.b Para todas as partidas do Brasileirão 2017, por exemplo, o site opinou sobre qual equipe sairia de campo vitoriosa ou se ocorreria um empate. Tal variável pode ser considerada como uma potencial indicadora da equipe "favorita" no confronto.

TABELA 5: Variáveis coletadas do Brasileirão 2017.

Variável	Variável
V1 - Jogo Clássico	V16 - Vitórias do Mandante
V2 - Equipe Mandante	V17 - Empates do Mandante
V3 - Equipe Visitante	V18 - Derrotas do Mandante
V4 - Vitórias do Mandante em Casa	V19 - Gols Marcados pelo Mandante
V5 - Empates do Mandante em Casa	V20 - Gols Sofridos pelo Mandante
V6 - Derrotas do Mandante em Casa	V21 - Pontos Ganhos pelo Mandante
V7 - Gols Marcados pelo Mandante em Casa	V22 - Vitórias do Visitante
V8 - Gols Sofridos pelo Mandante em Casa	V23 - Empates do Visitante
V9 - Pontos Ganhos pelo Mandante em Casa	V24 - Derrotas do Visitante
V10 - Vitórias do Visitante Fora de Casa	V25 - Gols Marcados pelo Visitante
V11 - Empates do Visitante Fora de Casa	V26 - Gols Sofridos pelo Visitante
V12 - Derrotas do Visitante Fora de Casa	V27 - Pontos Ganhos pelo Visitante
V13 - Gols Marcados pelo Visitante Fora de Casa	V28 - Posição do Mandante
V14 - Gols Sofridos pelo Visitante Fora de Casa	V29 - Posição do Visitante
V15 - Pontos Ganhos pelo Visitante Fora de Casa	**V30 - Resultado da Partida**

O gráfico de setores da Figura 5 mostra o desempenho dos times nos jogos do Brasileirão 2017. O que mais chama a atenção nesse gráfico é que o amplo domínio das equipes mandantes, fato constatado em edições anteriores da competição, foi amenizado pelo fato dos visitantes terem vencido 29% do total de jogos. O desempenho dos visitantes aponta para um nivelamento ainda maior do Campeonato Brasileiro da Série A. As equipes que mais venceram jogando fora de casa foram: Corinthians (9 triunfos), Vitória (8) e Grêmio (8). Por outro lado, os times com mais vitórias jogando dentro de casa foram: Corinthians, Santos e Palmeiras, cada um com doze triunfos.

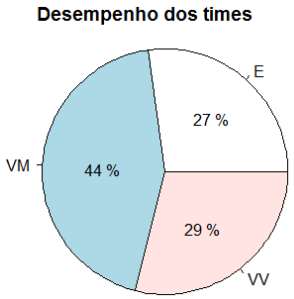

FIGURA 5: Gráfico de setores dos jogos do Brasileirão 2017.

No dia 03 de dezembro de 2017, o Brasileirão 2017 terminou e a classificação final do torneio encontra-se na Figura 9 do Apêndice, na qual observa-se, dentre outros, que a equipe campeã foi o Corinthians; o Palmeiras ficou com o vice-campeonato; Corinthians, Palmeiras, Santos, Grêmio e Flamengo se classificaram diretamente à fase de grupos da Copa Libertadores da América do ano seguinte, além do Cruzeiro, campeão da Copa do Brasil 2017, enquanto que Vasco e Chapecoense, sétimo e oitavo colocados, respectivamente, se classificaram para a disputa da fase Pré-Libertadores; e, por fim, as equipes do Coritiba, Avaí, Ponte Preta e Atlético-GO foram rebaixadas à Série B do Campeonato Brasileiro do ano seguinte.

3.1 RESULTADOS - MODELAGEM DE CLASSIFICAÇÃO ATEMPORAL

Os modelos preditivos apresentados na Seção 2.1 foram aplicados à base de dados do Brasileirão 2017, sendo que as 35 primeiras rodadas do torneio foram utilizadas como um conjunto de treinamento e as rodadas 36 - 38 como um conjunto de teste.

As covariáveis selecionadas (dentre as listadas na Tabela 5), por significância estatística (nível de 10%), para compôr o modelo de classificação final utilizando a técnica da RLP foram: número de gols marcados pela equipe visitante fora de casa (V13), número de

vitórias da equipe mandante (V16), pontos ganhos da equipe mandante (V21), número de derrotas da equipe visitante (V24), número de gols marcados pela equipe visitante (V25), pontos ganhos da equipe visitante (V27) e posição da equipe visitante (V29). As estimativas dos parâmetros do modelo final são apresentadas na Tabela 6.

TABELA 6: Modelo final de RLP para os dados do Brasileirão 2017. E.P. = Erro-Padrão.

Variável	Estimativa	E.P.	Escore z	Valor-p
VM: Intercepto	0,337	0,436	0,772	0,440
VM: Gols Marcados pelo Visitante Fora De Casa	-0,102	0,060	-1,696	0,090
VM: Vitórias do Mandante	-0,604	0,290	-2,084	0,037
VM: Pontos Ganhos pelo Mandante	0,234	0,099	2,365	0,018
VM: Derrotas do Visitante	-0,244	0,085	-2,867	0,004
VM: Gols Marcados pelo Visitante	0,096	0,047	2,040	0,041
VM: Pontos Ganhos pelo Visitante	-0,069	0,034	-2,006	0,045
VM: Posição do Visitante	0,061	0,042	1,457	0,145
VV: Intercepto	-0,391	0,492	-0,794	0,427
VV: Gols Marcados pelo Visitante Fora De Casa	-0,003	0,063	-0,051	0,960
VV: Vitórias do Mandante	-0,270	0,312	-0,865	0,387
VV: Pontos Ganhos pelo Mandante	0,073	0,105	0,692	0,489
VV: Derrotas do Visitante	-0,106	0,089	-1,200	0,230
VV: Gols Marcados pelo Visitante	0,020	0,050	0,395	0,693
VV: Pontos Ganhos pelo Visitante	0,000	0,036	-0,005	0,996
VV: Posição do Visitante	0,082	0,045	1,819	0,069

As duas funções *logit* estimadas foram as seguintes:

$$g_1(\mathbf{x}) = 0,337 - 0,102 * \texttt{V13} - 0,604 * \texttt{V16} + 0,234 * \texttt{V21} - 0,244 * \texttt{V24} + 0,096 * \texttt{V25} - 0,069 * \texttt{V27} + 0,061 * \texttt{V29},$$

$$g_2(\mathbf{x}) = -0,391 - 0,003 * \texttt{V13} - 0,270 * \texttt{V16} + 0,073 * \texttt{V21} - 0,106 * \texttt{V24} + 0,020 * \texttt{V25} + 0,000 * \texttt{V27} + 0,082 * \texttt{V29}.$$

Considerando uma observação $\mathbf{x}^* = (1, 26, 13, 51, 11, 52, 62, 3)'$, obtém-se as seguintes probabilidades:

$$P(Y = VM\,|\,\mathbf{x}^*) \approx 0,340,$$
$$P(Y = VV\,|\,\mathbf{x}^*) \approx 0,300,$$
$$P(Y = E\,|\,\mathbf{x}^*) \approx 0,360.$$

Neste caso, uma partida com tais características seria classificada como *Empate*, porque a maior probabilidade está associada à classe $Y = E$. Essa partida é, justamente, o confronto entre Atlético-MG e Grêmio pela última rodada do Brasileirão 2017, que terminou com vitória da equipe mandante (CAM) por 4 a 3. Repare, no entanto, que as probabilidades de pertencimento às classes estão muito próximas entre si. Assim, o modelo ajustado não é tão eficiente em prever o resultado final desse confronto, isto é, ele fica em dúvida, seja porque a partida em questão é realmente difícil de ser prevista ou devido ao alto número de gols.

As previsões para as partidas das três últimas rodadas do Brasileirão 2017, obtidas a partir do modelo de RLP ajustado, encontram-se na Tabela 7. A medida Brier Score associada a essas previsões é 0,6406 e o total de acertos foi de doze para os trinta jogos previstos (40%).

A classificação a partir da 35^a rodada, completada com os resultados previstos utilizando a técnica da RLP (ver Tabela 16 do Apêndice) [1], conseguiu acertar todos os times

[1]Esse tipo de visualização permite tirar conclusões que não eram possíveis de serem feitas ao olhar somente para os resultados previstos por rodada. Caso duas ou mais equipes tenham o mesmo número de pontos ao final da classificação prevista, será utilizado como critério de desempate o número de vitórias de cada equipe

que garantiram vaga diretamente na fase de grupos da Copa Libertadores do ano seguinte, que foram: Corinthians, Palmeiras, Santos e Flamengo. Vale ressaltar que as equipes do Grêmio e do Cruzeiro já tinham garantido essa vaga por terem sido os campeões da Copa Libertadores e da Copa do Brasil de 2017, respectivamente. No entanto, o modelo conseguiu cravar apenas sete posições corretas se comparado com a classificação real. Na parte do meio da tabela, o modelo trocou as classificações para competições internacionais de duas equipes: o previsto para a Chapecoense era uma vaga na Copa Sul-Americana mas, na verdade, a equipe conseguiu uma classificação para a pré-Libertadores de 2018; já a equipe do Botafogo, que pelas previsões ficaria com a vaga na fase preliminar da Copa Libertadores, conseguiu, na realidade, apenas a classificação para a Copa Sul-Americana de 2018. Quando se trata da parte de baixo da tabela prevista, o modelo também apresentou inconsistência, rebaixando para a Série B do Campeonato Brasileiro a equipe do Sport que, no final do campeonato, conseguiu livrar-se do rebaixamento e colocar a equipe do Coritiba para disputar uma competição de nível inferior no ano seguinte.

As rodadas 36 - 38 do Brasileirão 2017 também foram preditas pelo modelo de classificação baseado na técnica SVM (ver Tabela 7). O *kernel* utilizado foi o Gaussiano e os melhores parâmetros tiveram: função de custo $cost = 1$, $\eta = 0,033$ e 325 vetores suporte. Pode-se observar que esse método de previsão sempre creditou a maior probabilidade para a equipe mandante. Logo, a taxa de acertos foi de 46,67% e a medida Brier Score foi 0,6606, esta última ainda menor do que 2/3. A classificação final prevista encontra-se na Tabela 16 do Apêndice, a partir da qual constata-se que o modelo acertou apenas seis posições/colocações. Além disso, tal tabela prevista não foi condizente com os times do Vasco, Chapecoense e Sport, pois os dois primeiros conseguiram uma classificação para a fase preliminar da Copa Libertadores, mas o modelo previu uma classificação para a Copa Sul-Americana; enquanto que a equipe do Sport, de acordo com a classificação prevista por SVM, seria rebaixada para a Série B, porém conseguiu permanecer na elite do futebol brasileiro. Por outro lado, a tabela final cravou os quatro times que foram direto para a fase de grupos da Copa Libertadores de 2018.

Com respeito às previsões para o Brasileirão 2017 usando a técnica de RF (ver Tabela 7), os parâmetros selecionados que apresentaram maiores valores de acurácia no conjunto de treinamento (35 primeiras rodadas) e utilizados no modelo final foram: número de variáveis amostradas aleatoriamente como candidatas em cada divisão $mtry = 1$; e número de árvores utilizadas nas florestas $ntree = 1000$. A Figura 6 mostra que esses parâmetros foram os melhores dentre uma variedade de combinações consideradas. Quando aplicado ao conjunto de teste (3 últimas rodadas), o número de acertos foi muito pequeno: apenas oito de trinta jogos previstos ($26,67\%$). Para essas previsões, a medida Brier Score foi de 0,7706 (maior do que 2/3). Portanto, as predições desse modelo para os dados do Brasileirão 2017 são de má qualidade. Por conta disso, não foi simulada a classificação final do campeonato usando essa técnica, pois não tem lógica insistir num modelo que não é eficiente; afinal de contas, se alguém fechar os olhos e atribuir probabilidades (1/3, 1/3, 1/3) para todos os jogos, obterá resultados melhores, além de ser presumivelmente mais barato. Porém, vale ressaltar que, numa aplicação das três técnicas de classificação supervisionada (RLP, SVM e RF) aos dados do Brasileirão 2016 (resultados não apresentados), a técnica RF foi a que obteve melhor performance preditiva. O que deixa claro que cada campeonato tem uma história e não deve ser generalizado por meio de um modelo único.

Por fim, as medidas de capacidade preditiva, definidas na Seção 2.3.1 e calculadas para os três modelos ajustados, estão disponíveis na Tabela 8. Observa-se que o modelo que utiliza a técnica de RF apresentou performance preditiva inferior, se comparado aos demais modelos (RLP e SVM).

no campeonato. Quem tiver o maior número de vitórias ficará à frente na tabela de classificação. Caso haja equipes empatadas em pontos ganhos e número de vitórias, será adotado como critério de desempate a equipe que tiver o melhor saldo de gols na 35^a rodada, já que as técnicas descritas na Seção 2.1 não modelam o número de gols marcados.

TABELA 7: Previsões (probabilidades estimadas, em %) para as rodadas 36 - 38 do Brasileirão 2017, usando as técnicas de RLP, SVM e RF. Em negrito, os "acertos".

		RLP			SVM			RF		
Rodada	Jogo	VM	E	VV	VM	E	VV	VM	E	VV
36	FLA 3 X 0 COR	**47,54**	25,46	27,00	**44,81**	24,54	30,65	**49,40**	23,20	27,40
36	SAO 0 X 0 BOT	39,52	35,47	25,01	44,76	24,28	30,96	55,00	25,80	19,20
36	SPT 1 X 0 BAH	**48,23**	23,68	28,08	**42,52**	26,80	30,68	35,80	25,60	38,60
36	VIT 1 X 1 CRU	31,27	**36,18**	32,56	45,00	28,83	28,18	58,90	17,30	23,80
36	ACG 1 X 1 CHA	11,78	**51,03**	37,20	43,54	26,59	29,86	40,60	28,30	31,10
36	SAN 1 X 0 GRE	**38,48**	36,13	25,39	**45,46**	23,92	30,62	**48,80**	24,90	26,30
36	CAM 3 X 0 CTB	**38,89**	28,06	33,06	**46,41**	24,93	28,67	42,90	44,00	13,10
36	CAP 3 X 1 VAS	28,50	39,74	31,76	**46,81**	24,44	28,75	**53,80**	30,10	16,10
36	FLU 2 X 0 PON	**53,20**	19,23	27,56	**43,94**	25,54	30,51	39,80	46,70	13,50
36	AVA 2 X 1 PAL	**37,23**	28,25	34,52	**38,32**	29,08	32,60	34,20	23,90	41,90
37	FLU 1 X 2 SPT	53,52	18,22	28,26	45,89	24,93	29,18	44,20	41,20	14,60
37	COR 2 X 2 CAM	50,03	33,29	16,68	45,04	24,30	30,66	67,00	22,60	10,40
37	CRU 0 X 1 VAS	41,50	32,77	25,73	46,26	23,27	30,47	53,20	29,70	17,10
37	CTB 1 X 2 SAO	43,35	25,94	30,72	46,52	24,51	28,97	45,80	42,20	12,00
37	GRE 1 X 1 ACG	24,96	**49,41**	25,63	44,85	24,27	30,87	55,40	28,60	16,00
37	PON 2 X 3 VIT	22,86	35,18	**41,96**	46,65	24,80	28,55	54,70	33,10	12,20
37	AVA 1 X 0 CAP	**38,82**	27,57	33,61	**37,85**	28,88	33,28	33,20	25,20	41,60
37	FLA 1 X 2 SAN	46,53	29,83	23,64	47,17	22,96	29,87	76,00	12,70	11,30
37	BAH 0 X 1 CHA	31,75	35,02	33,23	45,58	24,46	29,96	48,30	40,10	11,60
37	PAL 2 X 0 BOT	40,34	41,12	18,53	**43,87**	24,96	31,16	**62,80**	22,60	14,60
38	BOT 2 X 2 CRU	40,49	34,62	24,89	46,99	24,34	28,68	53,00	20,90	26,10
38	VAS 2 X 1 PON	**51,79**	24,13	24,08	**42,88**	25,55	31,57	**39,80**	36,80	23,40
38	SAN 1 X 1 AVA	44,92	28,77	26,31	48,09	22,92	28,98	54,70	27,90	17,40
38	SAO 1 X 1 BAH	56,96	21,66	21,38	44,99	24,34	30,67	46,00	34,20	19,80
38	CAM 4 X 3 GRE	34,06	35,95	29,99	**45,14**	26,09	28,77	**41,80**	25,80	32,40
38	CAP 3 X 0 PAL	31,53	41,15	27,32	**44,18**	26,21	29,62	**41,00**	33,10	25,90
38	SPT 1 X 0 COR	30,44	35,65	33,91	**43,41**	26,92	29,67	40,10	15,90	44,00
38	VIT 1 X 2 FLA	46,68	28,61	24,72	44,04	26,79	29,17	58,90	19,10	22,00
38	ACG 1 X 1 FLU	29,49	28,66	41,84	46,41	25,60	28,00	52,30	22,20	25,50
38	CHA 2 X 1 CTB	29,84	37,09	33,07	**46,46**	24,07	29,47	**51,00**	32,70	16,30

TABELA 8: Medidas de performance das técnicas de classificação aplicadas aos dados do Brasileirão 2017.

	Medida de Performance							
Técnica	PM	TE	$\texttt{Precisão}_\mu$	$\texttt{Recall}_\mu$	$\texttt{FS}_\mu$	$\texttt{Precisão}_M$	$\texttt{Recall}_M$	$\texttt{FS}_M$
RLP	0,600	0,400	0,400	0,400	0,400	0,415	0,349	0,379
SVM	0,644	0,356	0,467	0,467	0,467	0,156	0,333	0,212
RF	0,511	0,489	0,267	0,267	0,267	0,111	0,191	0,140

3.2 RESULTADOS - MODELAGEM DE REGRESSÃO TEMPORAL

A partida Coritiba *versus* Cruzeiro, válida pela 29^{a} rodada do Brasileirão 2017, foi tomada para exemplificação das técnicas apresentadas na Seção 2.2. O duelo ocorreu no dia 18 de outubro de 2017 e teve como resultado final (verdadeiro) o placar de Coritiba 1×0 Cruzeiro. Os dados que foram considerados estão resumidos na Tabela 9.

TABELA 9: Exemplo de conjunto de dados da partida entre Coritiba (equipe mandante) e Cruzeiro (equipe visitante), ocorrida em 18 de outubro de 2017.

Coritiba				Cruzeiro			
Data	Time Adversário	y_t^{M}	x_{t-1}	Data	Time Adversário	y_t^{V}	x_{t-1}
25/10/2014	Grêmio	1	0,67	25/10/2014	Figueirense	1	1,13
08/11/2014	Fluminense	1	1,13	16/11/2014	Santos	1	0,63
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
24/09/2017	Botafogo	2	1,24	24/09/2017	Atlético-GO	2	1,23
15/10/2017	Grêmio	0	1,13	11/10/2017	Grêmio	1	0,61
18/10/2017	Cruzeiro	?	1,15	18/10/2017	Coritiba	?	0,80

O valor de $x_{t-1} = 1,15$, localizado na última linha da quarta coluna da Tabela 9, indica que a equipe do Cruzeiro, da temporada 2014 até antes de 18 de outubro de 2017, conce-

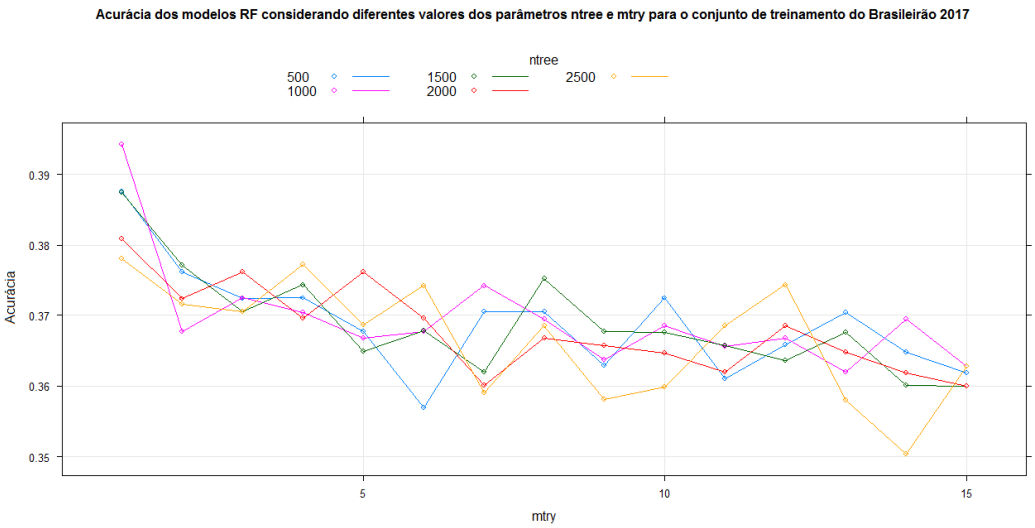

FIGURA 6: Seleção do melhor modelo de RF para aplicar no conjunto de teste do Brasileirão 2017.

deu, em média, 1,15 gols aos seus adversários quando jogou fora de casa. Analogamente, o valor de $x_{t-1} = 0,80$, localizado na última linha e coluna da Tabela 9, refere-se à média de gols concedidos pela equipe do Coritiba quando jogou dentro de casa no mesmo período.

A Figura 7 mostra a série temporal dos gols marcados pelo Coritiba como mandante e pelo Cruzeiro como visitante. A média de gols marcados pelo Coritiba em casa é de 1,17 gols e a variância é de 1,24 gols^2. Já o número médio de gols marcados pelo time do Cruzeiro jogando fora de casa é de 1,05 gols com variância de 1,09 gols^2.

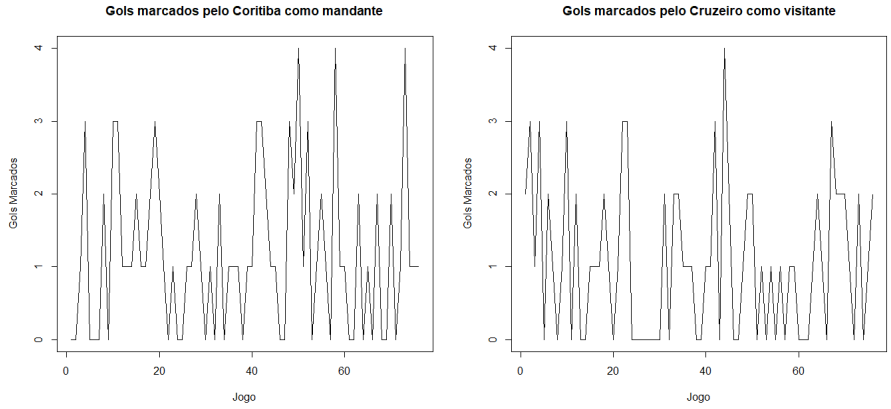

FIGURA 7: Série temporal dos gols marcados pelo Coritiba como mandante e pelo Cruzeiro como visitante, durante as temporadas 2014-2017.

A seleção do modelo PARX de previsão (ou predição) é baseada na abordagem descrita na Seção 2.2.2. Assim, o modelo selecionado para explicar a distribuição dos gols do Coritiba como mandante foi o PARX(0,1), cujo AIC foi 158,5461 (ver Tabela 10). Enquanto o modelo PARX selecionado via AIC para representar a distribuição dos gols do Cruzeiro como visitante também foi o PARX(0,1), com AIC igual a 149,1381 (ver Tabela 11).

Os parâmetros estimados pelos dois modelos são apresentados na Tabela 12. A distribuição prevista dos gols marcados pelo Coritiba (equipe mandante) contra o Cruzeiro (equipe visitante) é $\hat{P}(y^{\mathrm{M}}_{T+1} = y^{\mathrm{M}}|F_T) = \texttt{Poisson}(y^{\mathrm{M}}|\hat{\lambda}^{\mathrm{M}}_{T+1|T} = 0,8890)$, enquanto que a distribuição prevista dos gols marcados pelo Cruzeiro (equipe visitante) contra o Coritiba (equipe mandante) é $\hat{P}(y^{\mathrm{V}}_{T+1} = y^{\mathrm{V}}|F_T) = \texttt{Poisson}(y^{\mathrm{V}}|\hat{\lambda}^{\mathrm{V}}_{T+1|T} = 0,7416)$.

Como o valor esperado de uma distribuição de Poisson é igual ao parâmetro de intensidade, o número esperado de gols marcados pelo Coritiba contra o Cruzeiro é 0,8890. Por

TABELA 10: Valores de AIC para seleção do melhor modelo PARX(p, q) que explique a distribuição dos gols marcados pelo Coritiba como mandante.

		p				
		0	1	2	3	4
q	0	-	159,1011	161,1011	163,1012	165,1028
	1	**158,5461**	160,5462	162,5462	164,5463	166,5474
	2	159,3407	161,3389	163,3387	165,3421	168,0132
	3	161,3402	163,3371	165,3407	167,3402	170,5663
	4	163,3367	165,3571	167,5154	169,3403	173,6687

TABELA 11: Valores de AIC para seleção do melhor modelo PARX(p, q) que explique a distribuição dos gols marcados pelo Cruzeiro como visitante.

		p				
		0	1	2	3	4
q	0	-	150,1943	152,1891	154,1891	156,2050
	1	**149,1381**	151,1383	153,1446	155,1402	157,1612
	2	151,1383	153,1381	155,1383	157,1615	159,1384
	3	153,1381	155,1385	157,1634	159,1420	161,2673
	4	155,0912	157,2324	159,1523	161,3645	163,7292

TABELA 12: Resultados dos dois modelos PARX estimados para o exemplo da partida Coritiba *versus* Cruzeiro, ocorrida em 18/10/2017. Erro-padrão entre parênteses.

	Coritiba	Cruzeiro
Parâmetro	PARX(0,1)	PARX(0,1)
ω	0,3745 (0,568)	0,0001 (0,514)
β_1	0,0943 (0,135)	0,1402 (0,134)
γ	0,4558 (0,374)	0,9791 (0,589)

outro lado, espera-se que o Cruzeiro marque 0,7416 gols contra o Coritiba.

A distribuição conjunta, apresentada na Tabela 13, permite obter previsões para diferentes placares. Observa-se que o resultado mais provável é Coritiba 0 $\times$ 0 Cruzeiro, com probabilidade 0,1958. Vale ressaltar que o resultado final (verdadeiro) foi Coritiba 1 $\times$ 0 Cruzeiro e a probabilidade prevista associada a esse resultado é 0,1741.

TABELA 13: Distribuição de probabilidade conjunta dos gols marcados na partida Coritiba *versus* Cruzeiro. Em negrito, a probabilidade associada ao verdadeiro resultado da partida.

	y_t^{V}						
y_t^{M}	0	1	2	3	4	5	6
0	0,1958	0,1452	0,0538	0,0133	0,0025	0,0004	0,0000
1	**0,1741**	0,1291	0,0479	0,0118	0,0022	0,0003	0,0000
2	0,0774	0,0574	0,0213	0,0053	0,0010	0,0001	0,0000
3	0,0229	0,0170	0,0063	0,0016	0,0003	0,0000	0,0000
4	0,0051	0,0038	0,0014	0,0003	0,0001	0,0000	0,0000
5	0,0009	0,0007	0,0002	0,0001	0,0000	0,0000	0,0000
6	0,0001	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000

Ainda da Tabela 13, podem ser obtidas as probabilidades de vitória do mandante, empate e vitória do visitante. Para encontrar a probabilidade do jogo terminar empatado, basta somar os valores da diagonal principal, ou seja, $\hat{P}(y_{T+1}^{\mathrm{M}} = y_{T+1}^{\mathrm{V}}|F_T) = 0,3478$. Para obter a probabilidade de vitória da equipe mandante, basta somar os valores da matriz triangular inferior, isto é, $\hat{P}(y_{T+1}^{\mathrm{M}} > y_{T+1}^{\mathrm{V}}|F_T) = 0,3679$. Analogamente, somando os valores da matriz triangular superior, obtém-se a probabilidade de vitória da equipe visitante, isto é, $\hat{P}(y_{T+1}^{\mathrm{M}} < y_{T+1}^{\mathrm{V}}|F_T) = 0,2843$.

Todo esse procedimento, desde a seleção dos modelos PARX até a obtenção das proba-

bilidades de VM, E e VV, foi feito não só para a partida Coritiba *versus* Cruzeiro, mas sim para todos os jogos da 36ª, 37ª e 38ª rodadas do Brasileirão 2017, no intuito de calcular as medidas definidas na Seção 2.3.1, as quais avaliam a capacidade preditiva do método. As probabilidades preditas usando a modelagem de regressão temporal são mostradas na Tabela 14.

De trinta jogos, o modelo de regressão temporal acertou doze, o que corresponde a 40% de acerto no geral. A medida Brier Score associada a essas previsões é igual a 0,6618, que é menor do que 2/3.

TABELA 14: Previsões (probabilidades estimadas, em %) para as rodadas 36 - 38 do Brasileirão 2017, usando a modelagem de regressão temporal (modelos PARX).

Rodada	Jogo	VM	E	VV
36	FLA 3 X 0 COR	**44,92**	26,94	28,15
	SAO 0 X 0 BOT	59,20	22,69	18,09
	SPT 1 X 0 BAH	**46,32**	25,98	27,69
	VIT 1 X 1 CRU	44,16	27,13	28,70
	ACG 1 X 1 CHA	41,14	29,59	29,27
	SAN 1 X 0 GRE	**59,67**	21,52	18,79
	CAM 3 X 0 CTB	**65,25**	19,26	15,44
	CAP 3 X 1 VAS	**47,42**	30,49	22,09
	FLU 2 X 0 PON	**55,06**	25,93	19,01
	AVA 2 X 1 PAL	18,12	24,17	57,70
37	FLU 1 X 2 SPT	51,01	25,07	23,29
	COR 2 X 2 CAM	41,11	23,35	35,53
	CRU 0 X 1 VAS	57,91	23,98	18,11
	CTB 1 X 2 SAO	43,66	31,03	25,31
	GRE 1 X 1 ACG	44,12	25,71	30,18
	PON 2 X 3 VIT	44,31	26,62	29,07
	AVA 1 X 0 CAP	**37,18**	33,66	29,17
	FLA 1 X 2 SAN	47,42	26,93	25,65
	BAH 0 X 1 CHA	47,32	24,17	28,50
	PAL 2 X 0 BOT	**63,81**	23,39	12,79
38	BOT 2 X 2 CRU	46,65	25,36	27,97
	VAS 2 X 1 PON	**41,74**	33,76	24,50
	SAN 1 X 1 AVA	63,60	22,42	13,96
	SAO 1 X 1 BAH	43,14	30,17	26,69
	CAM 4 X 3 GRE	**66,07**	21,12	12,72
	CAP 3 X 0 PAL	36,85	23,98	39,16
	SPT 1 X 0 COR	**51,83**	24,99	23,17
	VIT 1 X 2 FLA	45,03	31,59	23,38
	ACG 1 X 1 FLU	51,83	26,89	21,28
	CHA 2 X 1 CTB	**45,13**	28,27	26,60

A Tabela 16 do Apêndice exibe uma comparação entre a classificação ao final do campeonato e uma classificação a partir da 35ª rodada completada com os resultados previstos pelo modelo de regressão temporal. Nota-se que o modelo conseguiu ser relativamente bom na parte de cima da tabela, pois conseguiu acertar o vice-campeão do torneio, além de predizer corretamente quais times iriam para a fase de grupos da Copa Libertadores 2018. Entretanto, o modelo foi incoerente nas partes do meio e final da tabela, pois não foi capaz de prever corretamente quais times, ao final do campeonato, estariam classificados para a fase preliminar da Copa Libertadores, além de apresentar também inconsistências nos times classificados para a Copa Sul-Americana 2018 e prever um possível rebaixamento do time do Sport para a segunda divisão do Campeonato Brasileiro, fato este que não ocorreu na realidade porque a equipe conseguiu terminar na 15ª colocação. Ademais, é importante observar que todos os modelos ajustados (RLP, SVM e PARX) identificaram corretamente quais times permaneceram nas posições 1, 5, 6 e 16.

As medidas da capacidade preditiva da modelagem de regressão temporal para os dados do Brasileirão 2017 estão disponíveis na Tabela 15. Observa-se um desempenho preditivo similar ao das técnicas de classificação supervisionada consideradas neste estudo, em

especial a técnica de RLP (ver Tabela 8).

TABELA 15: Medidas de performance da modelagem de regressão temporal (modelos PARX) para o Brasileirão 2017.

Medida	Valor	Medida	Valor
PM	0,600	$\texttt{FS}_{\mu}$	0,400
TE	0,400	$\texttt{Precisão}_{M}$	0,429
$\texttt{Precisão}_{\mu}$	0,400	$\texttt{Recall}_{M}$	0,286
$\texttt{Recall}_{\mu}$	0,400	$\texttt{FS}_{M}$	0,343

Na Figura 8 são mostrados os gráficos utilizados para avaliar a qualidade do ajuste dos modelos PARX. Os histogramas PIT parecem estar próximos da uniformidade e os diagramas de calibração marginal são satisfatórios, pois não apresentam valores discrepantes.

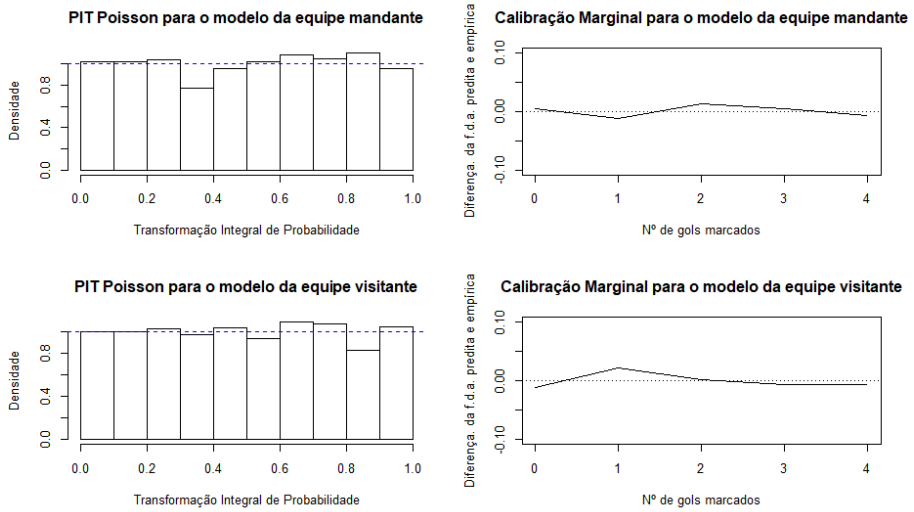

FIGURA 8: Diagnóstico dos modelos PARX ajustados.

4 CONSIDERAÇÕES FINAIS

Ao analisar diferentes temporadas do Campeonato Brasileiro da Série A (ou Brasileirão) separadamente, percebe-se, imediatamente, diferenças no desempenho dos times. Por exemplo, no Brasileirão 2016, as equipes mandantes venceram 53% do total de partidas, destacando assim que o fator casa foi uma das características desse campeonato. Por outro lado, um resultado atípico foi identificado no Brasileirão 2017: as equipes visitantes venceram 29% do total dos jogos e, ainda mais, houve mais vitórias do time que joga fora de casa do que empates em todo o campeonato. Essas análises servem para dizer que o presente trabalho tem suas particularidades e limitações. Os modelos ajustados para a temporada 2017 são diferentes dos modelos das temporadas anteriores, pois cada campeonato tem uma história e não deve ser representado por meio de um modelo único. Ou seja, nenhum modelo irá ser bom o suficiente para prever qualquer torneio. Não são os dados que devem se adequar a um particular modelo, mas sim o modelo que deve ajustar bem os dados. Nota-se que, para o Brasileirão 2017, os modelos de classificação baseados nas técnicas de RLP e SVM produzem previsões de qualidade minimamente aceitáveis por terem medidas Brier Score associadas às previsões menores do que 2/3, enquanto que o modelo usando a técnica de RF apresentou previsões de má qualidade (Brier Score = 0,7706).

A abordagem atemporal deste trabalho apresenta limitações no sentido de que não considera a possível dependência entre observações (isto é, entre partidas, principalmente

aquelas de rodadas sucessivas envolvendo a mesma equipe), bem como o comportamento não-decrescente de algumas variáveis, como "Gols Marcados pelo Mandante", "Vitórias do Mandante", etc. No caso do tratamento da dependência serial dos jogos e a superdispersão, foram aplicados os modelos de regressão temporal PARX, que produziram predições satisfatórias tanto com o Brier Score (menor do que 2/3) quanto com os diagnósticos feitos com o auxílio dos gráficos PIT e de calibração marginal. Porém, existem limitações devido ao fato da análise gráfica apresentar conclusões subjetivas. Para ser mais objetivo, recomenda-se utilizar um teste estatístico apropriado que mensure a bondade do ajuste desse modelo. Além disso, pretende-se utilizar em trabalhos futuros o modelo BNARX, que utiliza a distribuição Binomial Negativa ao invés da Poisson. Apesar do modelo PARX poder se ajustar bem a dados com superdispersão, espera-se que o modelo BNARX apresente resultados melhores, pois existe um parâmetro exclusivo que mensura o grau de superdispersão existente nos dados. Por fim, vale ressaltar que modelar o Campeonato Brasileiro é bastante complicado devido ao nivelamento das equipes participantes e o seu alto grau de aleatoriedade nas partidas, em que o resultado final é decidido nos mínimos detalhes.

Referências

[1] A. F. Oliveira, "Origem do futebol na Inglaterra no Brasil," *Revista Brasileira de Futsal e Futebol*, vol. 4, pp. 170–174, 2012.

[2] F. F. Farias, "Análise e previsão de resultados de partidas de futebol," Master's thesis, Universidade Federal do Rio de Janeiro, 2008.

[3] H. Rue and O. Salvesen, "Prediction and retrospective analysis of soccer matches in a league," *Journal of the Royal Statistical Society: Series D*, vol. 49, pp. 399–418, 2000.

[4] O. G. Souza-Jr and D. Gamerman, "Previsão de partidas de futebol usando modelos dinâmicos," in *Anais do XXXVI SBPO*, 2004.

[5] D. R. Brillinger, "Modelling game outcomes of the Brazilian 2006 Series A championship as ordinal-valued," *Brazilian Journal of Probability and Statistics*, vol. 22, pp. 89–104, 2008.

[6] A. K. Suzuki, L. E. B. Salasar, J. G. Leite, and F. Louzada-Neto, "A Bayesian approach for predicting match outcomes: The 2006 (Association) Football World Cup," *Journal of the Operational Research Society*, vol. 61, pp. 1530–1539, 2009.

[7] A. M. Alves, J. C. C. B. S. de Mello, T. G. Ramos, and A. P. Sant'anna, "Logit models for the probability of winning football games," *Pesquisa Operacional*, vol. 31, pp. 459–465, 2011.

[8] L. Tavares and A. K. Suzuki, "Modelagem estatística para previsão esportiva: Uma aplicação no futebol," *Matemática e Estatística em Foco*, vol. 3, pp. 32–47, 2015.

[9] M. L. Arruda, "Poisson, Bayes, Futebol e DeFinetti," Master's thesis, Universidade de São Paulo, 2000.

[10] G. Angelini and L. de Angelis, "PARX model for football matches predictions," *Journal of Forecasting*, vol. 36, pp. 795–807, 2017.

[11] A. Agosto, G. Cavaliere, D. Kristensen, and A. Rahbek, "Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX)," *Journal of Empirical Finance*, vol. 38, pp. 640–663, 2016.

[12] C. A. O. Filho, A. K. Suzuki, F. Louzada, E. F. Saraiva, and L. E. B. Salasar, "Uma abordagem Bayesiana para previsão de resultados de jogos de futebol: Uma aplicação ao Campeonato Inglês," *Revista Brasileira de Biometria*, vol. 35, pp. 76–97, 2017.

[13] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. John Wiley & Sons, 1989.

[14] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B*, vol. 20, pp. 215–242, 1958.

[15] H. R. Bittencourt, "Regressão logística politômica: revisão teórica e aplicações," *Acta Scientiae*, vol. 5, pp. 77–86, 2003.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[17] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*. Springer, 2002.

[18] V. Vapnik and A. J. Lerner, "Generalized portrait method for pattern recognition," *Automation and Remote Control*, vol. 24, pp. 774–780, 1963.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[20] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013.

[21] N. J. Nalini and S. Palanivel, "Music emotion recognition: The combined evidence of MFCC and residual phase," *Egyptian Informatics Journal*, vol. 17, pp. 1–10, 2016.

[22] A. C. Lorena and A. C. P. L. F. Carvalho, "Introdução às máquinas de vetores suporte (support vector machines)," tech. rep., Universidade de São Paulo, 2003.

[23] N. Vlastakis, G. Dotsis, and R. N. Markellos, "Nonlinear modelling of European football scores using support vector machines," *Applied Economics*, vol. 40, pp. 111–118, 2008.

[24] M. David, "e1071: Misc Functions of the Department of Statistics, Probability Theory Group," 2017.

[25] L. Breiman, "Random forests," *Journal of Machine Learning*, vol. 45, pp. 5–32, 2001.

[26] V. Ghosal, P. Tikmani, and P. Gupta, "Face classification using gabor wavelets and random forest," in *Canadian Conference on Computer and Robot Vision*, 2009.

[27] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, pp. 18–22, 2002.

[28] B. Boldrin, "Predicting the result of English Premier League Soccer Games with the use of Poisson models," Master's thesis, Stetson University, 2017.

[29] K. Fokianos, A. Rahbek, and D. Tjostheim, "Poisson autoregression," *Journal of the American Statistical Association*, vol. 104, pp. 1430–1439, 2009.

[30] T. Liboschik, K. Fokianos, and R. Fried, "tscount: An R package for analysis of count time series following generalized linear models," *Journal of Statistical Software*, vol. 82, pp. 1–51, 2017.

[31] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[32] C. Czado, T. Gneiting, and L. Held, "Predictive model assessment for count data," *Biometrics*, vol. 65, pp. 1254–1261, 2009.

[33] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society: Series B*, vol. 69, pp. 243–268, 2007.

[34] P. R. Hansen, Z. Huang, and H. W. Shek, "Realized GARCH: A joint model for returns and realized measures of volatility," *Journal of the Applied Econometrics*, vol. 27, pp. 877–906, 2012.

[35] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, pp. 427–437, 2009.

[36] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, pp. 1–3, 1950.

APÊNDICE

TABELA 16: Previsão da classificação final do Brasileirão 2017 usando a modelagem de RLP, SVM e PARX.

	Classificação Real			Classificação Prevista								
				RLP			SVM			PARX		
Posição	Equipe	PG	V	Equipe	PG	V	Equipe	PG	V	Equipe	PG	V
1	COR	72	21	COR	75	22	COR	74	22	COR	74	22
2	PAL	63	19	GRE	63	18	GRE	64	19	PAL	69	21
3	SAN	63	17	PAL	62	18	PAL	63	19	GRE	64	19
4	GRE	62	18	SAN	62	17	SAN	62	17	SAN	62	17
5	CRU	57	15	CRU	59	16	CRU	58	16	CRU	58	16
6	FLA	56	15	FLA	56	15	FLA	56	15	FLA	56	15
7	VAS	56	15	BOT	55	15	BOT	54	15	BOT	54	15
8	CHA	54	15	VAS	54	14	CAM	53	14	CAM	53	14
9	CAM	54	14	FLU	52	13	VAS	53	14	VAS	53	14
10	BOT	53	14	SAO	51	14	BAH	52	14	BAH	52	14
11	CAP	51	14	CAM	51	13	SAO	51	14	SAO	51	14
12	BAH	50	13	BAH	50	13	CAP	51	14	CHA	50	14
13	SAO	50	13	CHA	50	13	CHA	50	13	FLU	49	12
14	FLU	47	11	CAP	47	12	FLU	49	12	CAP	48	13
15	SPT	45	12	CTB	47	12	CTB	46	12	CTB	46	12
16	VIT	43	11	VIT	46	12	VIT	45	12	VIT	45	12
17	CTB	43	11	AVA	42	10	PON	42	11	PON	42	11
18	AVA	43	10	SPT	40	10	SPT	42	11	SPT	42	11
19	PON	39	10	PON	39	10	AVA	42	10	ACG	39	11
20	ACG	36	9	ACG	35	9	ACG	39	11	AVA	39	9

		P	J	V	E	D	GP	GC	SG
1	Corinthians	72	38	21	9	8	50	30	+20
2	Palmeiras	63	38	19	6	13	61	45	+16
3	Santos	63	38	17	12	9	42	32	+10
4	Grêmio	62	38	18	8	12	55	36	+19
5	Cruzeiro	57	38	15	12	11	47	39	+8
6	Flamengo	56	38	15	11	12	49	38	+11
7	Vasco	56	38	15	11	12	40	47	-7
8	Chapecoense	54	38	15	9	14	47	49	-2
9	Atlético Mineiro	54	38	14	12	12	52	49	+3
10	Botafogo	53	38	14	11	13	45	42	+3
11	Atlético Paranaense	51	38	14	9	15	45	43	+2
12	Bahia	50	38	13	11	14	50	48	+2
13	São Paulo	50	38	13	11	14	48	49	-1
14	Fluminense	47	38	11	14	13	50	53	-3
15	Sport	45	38	12	9	17	46	58	-12
16	Vitória	43	38	11	10	17	50	58	-8
17	Coritiba	43	38	11	10	17	42	51	-9
18	Avaí	43	38	10	13	15	29	48	-19
19	Ponte Preta	39	38	10	9	19	37	52	-15
20	Atlético Goianiense	36	38	9	9	20	38	56	-18

FIGURA 9: Classificação final do Brasileirão 2017. P - Pontos Ganhos; J - Jogos; V - Vitórias; E - Empates; D - Derrotas; GP - Gols Pró; GC - Gols Contra; SG - Saldo de Gols.
Fonte: http://www.ogol.com.br