A relação entre o sucesso de uma seleção nacional de futebol e sua performance em partidas das Copas do Mundo de 2014, 2018 e 2022

Palavras-Chave: Curva ROC, Copa do mundo, Regressão logística

Autores:

MATHEUS RUBIO WODEWOTZKI, IMECC – UNICAMP

RAFAEL PIMENTEL MAIA, IMECC - UNICAMP

INTRODUÇÃO:

Neste projeto, a intenção é investigar como a análise estatística pode contribuir para a identificação de padrões de desempenho e como esses padrões se refletem nos resultados das partidas, utilizando como referência os eventos das Copas do Mundo de 2014, 2018 e 2022. A aplicação de métricas específicas, como posse de bola, eficácia nos passes, finalizações e comportamento tático, será fundamental para desenvolver modelos estatísticos robustos que transcendam a mera observação superficial do jogo..

METODOLOGIA:

Uma vez determinados os objetivos desse projeto, passamos para a obtenção de um conjunto de dados que nos permita realizar análises e extrair informações relevantes. Nesse contexto, optou-se por utilizar o site *"whoscored.com*" como fonte primária para obtenção dos dados (Liu et al, 2015), devido à sua confiabilidade e abrangência no fornecimento de estatísticas detalhadas relacionadas ao nosso campo de estudo. Foram selecionadas informações de 174 partidas de futebol referentes às copas do mundo dos anos de 2014, 2018 e 2022, oferecendo uma visão detalhada do desempenho de 47 seleções diferentes.

Para uma análise mais precisa, foram desconsiderados os jogos que ultrapassaram o tempo padrão de 90 minutos (Castellano et al, 2012), isto é, jogos da fase de mata-mata que precisaram de prorrogação (tempo extra) para sua conclusão. No total, foram 18 partidas eliminadas. Foram selecionadas ao total 18 variáveis que serviriam de base para o estudo e realizados tratamentos de dados convenientes para os objetivos da pesquisa através do auxílio do *software* de programação RStudio.

Para observar como o comportamento e as ações realizadas por uma equipe durante uma partida de futebol podem influenciar em seu sucesso, foram utilizadas algumas métricas e ferramentas estatísticas focadas em discriminar entre as seleções com os melhores resultados contra as que não obtiveram tanto êxito (Lago-Peñas et al, 2010).

A abordagem inicial foi realizada através de uma análise descritiva do banco de dados, buscando enxergar padrões e comportamentos das variáveis explicativas, segregando as observações em grupos semelhantes e diferenciando as respostas. A obtenção de estatísticas sumárias para as variáveis estudadas retorna valores capazes de carregar possíveis tendências para as classes, além de auxiliar

na visualização das distribuições das informações ao longo das amostras, possibilitando a identificação de médias e possíveis valores discrepantes.

Durante a organização do banco de dados, foi criada uma variável adicional y cujo valor é igual a 1 se a equipe venceu e 0 caso contrário, e foi ajustado um modelo de regressão logística utilizando a função $gamlss()$ (Stasinopoulos et al, 2007) do pacote gamlss do R, definindo a binomial como a família de distribuição, retornando as probabilidades de vitória de cada equipe em suas respectivas partidas.

O modelo de regressão logística (Cox et al, 1989) é dado por

$$\log\left\{\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right\} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

onde $\pi(\boldsymbol{x})$ representa a probabilidade de sucesso dado os valores $\boldsymbol{x} = \left(1, x_1, \ldots, x_{p-1}\right)^T$ de variáveis explicativas.

Após a definição do modelo, foi realizada a análise das predições. Sendo a variável de interesse binária, os valores obtidos variaram entre 0 e 1, representando as chances de vitória de cada equipe em suas respectivas partidas. Com base nisso, foi realizado um exercício de classificação utilizando a curva ROC.

Há uma série de estimadores estatísticos que são empregados na análise do desempenho de modelos classificatórios, e um dos mais utilizados é a curva ROC (*receiver operating characteristic*), que consiste em uma representação gráfica da performance de um modelo de dados quantitativos segundo sua taxa de sensibilidade (fração dos verdadeiros positivos) e a fração dos falsos positivos (1-especificidade) (Polo et al, 2020). Nesse estudo, a técnica da curva ROC foi aplicada para a obtenção de um ponto de corte, onde valores preditos de nossa variável resposta acima desse ponto foram classificados como equipes que obtiveram sucesso em seu respectivo jogo e seleções abaixo desse ponto não obtiveram sucesso. O experimento serve como uma espécie de validação, uma vez que a partida já ocorreu e queremos descobrir as chances de um time ter saído vitorioso da mesma de acordo com seu desempenho nela.

RESULTADOS E DISCUSSÃO:

As Tabelas 1 e 2 são resultados da análise descritiva do banco de dados estudado, contendo as estatísticas sumárias das variáveis analisadas para as equipes que obtiveram sucesso em seus respectivos jogos e para aquelas que saíram derrotadas.

O ajuste do modelo foi feito através da função do pacote gamlss do R chamada de $stepGAIC()$ que realiza a seleção de modelo utilizando a técnica de ***stepwise*** usando o critério de informação de akaike generalizado.

Posteriormente, foi testado a qualidade do modelo definido através de testes dos resíduos. O teste de Shapiro-Wilks retornou um p-valor igual a 0.274, indicando normalidade dos resíduos. A estatística do teste de Durbwin-Watson resultou em 1.9, valor que por ser próximo de 2, indica ausência de correlação entre os resíduos e, por fim, o teste de Breusch-Pagan rejeitou a hipótese de homocedasticidade, entretanto, o gráfico desses resíduos não apresentou clara tendência de variação da variância, portanto, o modelo foi considerado adequado.

	Média	**Desvio Padrão**	**Mediana**	**Mínimo**	**Máximo**	**1º Quartil**	**3º Quartil**
Gols	2.34	1.24	2	1	7	2	3
Finalizações	13.17	4.93	13	3	32	10	15
Passes certos (%)	81.12	7.08	82	58	94	78	86

Dribles	8.67	4.58	8	0	22	5	12
Duelos aéreos vencidos	15.62	6.67	15	2	38	10	19
Tackles	17.73	5.24	17	6	30	14	21
Cartões amarelos	1.36	1.18	1	0	6	0	2
Cartões vermelhos	0.03	0.16	0	0	1	0	0
Finalizações no alvo	5.14	2.54	5	1	13	3	6
Passes	456.95	143.58	440	218	1045	348	544
Cruzamentos	15.69	7.48	15	2	42	10	21
Passes cruciais	9.75	4.10	10	1	26	7	12
Interceptações	10.41	4.78	10	2	29	7	12
Clearances	22.11	9.35	21	4	47	16	28
Faltas	13.15	4.65	12	4	31	10	16
Defesas	2.66	1.95	3	0	9	1	4
Perdas de posse de bola	22.77	5.92	22	11	47	19	26

Tabela 1: Estatísticas sumárias das equipes que venceram suas respectivas partidas

	Média	**Desvio Padrão**	**Mediana**	**Mínimo**	**Máximo**	**1º Quartil**	**3º Quartil**
Gols	10	0.67	0.0	0	3	0.00	1.00
Finalizações	11.40	5.18	10.0	0	28	8.00	14.00
Passes certos (%)	80.79	6.05	81.0	58	92	78.00	85.00
Dribles	8.15	4.43	8.0	0	22	5.00	10.75
Duelos aéreos vencidos	14.45	6.28	13.0	3	32	9.00	19.00
Tackles	16.07	4.88	16.0	5	37	12.00	19.75
Cartões amarelos	1.80	1.20	2.0	0	6	1.00	2.00
Cartões vermelhos	0.08	0.26	0.0	0	1	0.00	0.00
Finalizações no alvo	3.25	2.06	3.0	0	9	2.00	4.00
Passes	437.22	125.63	434.0	30	1058	359.25	509.25
Cruzamentos	18.24	7.80	18.0	2	46	13.00	23.00
Passes cruciais	8.49	4.39	7.5	0	24	6.00	11.00
Interceptações	10.58	4.80	10.0	2	26	7.00	13.00
Clearances	19.55	9.50	18.0	2	47	12.25	25.00
Faltas	13.51	4.32	14.0	3	24	10.00	16.00
Defesas	2.84	2.06	2.0	0	10	1.00	4.00
Perdas de posse de bola	24.12	5.66	25.0	11	38	20.00	27.00

Tabela 2:Estatísticas sumárias das equipes que perderam suas respectivas partidas.

Parâmetros	**Estimado**	**EP**	**Estatística t**	**p-valor**
Intercepto	-2.779	0.790	-3.52	0.00049

Chutes certos	0.499	0.075	6.67	0
Cruzamentos	-0.100	0.021	-4.74	0
Rank	-0.039	0.009	-4.20	0
Clearances	0.055	0.016	3.42	0.00070
Duelos Aéreos	0.054	0.022	2.52	0.01234
Tackles	0.075	0.028	2.71	0.00715
Interceptações	-0.042	0.028	-1.49	0.13842

Tabela 3: Parâmetros do modelo

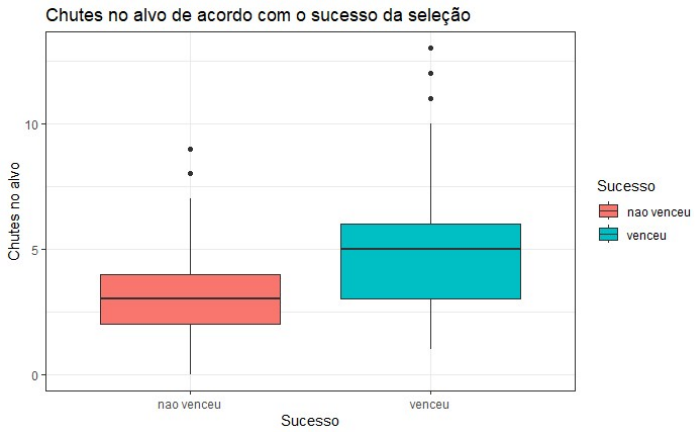

Figura 1: Boxplot que compara a quantidade de chutes no alvo de acordo com o sucesso da seleção.

Através da Tabela 3, verifica-se que a variável que se demonstrou mais significativa no modelo foi a variável de chutes no alvo e, observando o *boxplot* que compara o comportamento dessa estatística para as equipes vencedoras e para as perdedoras, aparentemente equipes que apresentarem melhor pontaria nas finalizações, possuirão maiores chances de vencer a partida.

Ainda sobre o modelo ajustado, percebe-se que variáveis como o número de cruzamentos, o *rank* da equipe e a quantidade de interceptações afetam negativamente na resposta, enquanto que além do número de chutes certos, *Clearances,* Duelos Aéreos e *Tackles* influenciam positivamente no modelo.

Além disso, é importante pontuar que todas as variáveis selecionadas se mostraram significativas para o modelo, com exceção da variável “Interceptações”, que foi mantida por motivos de uma melhor predição.

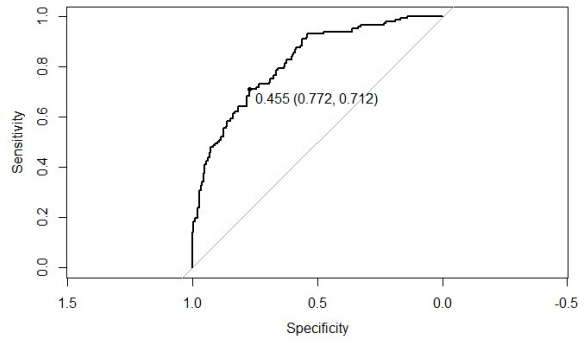

Figura 2: Curva ROC - Ponto Crítico

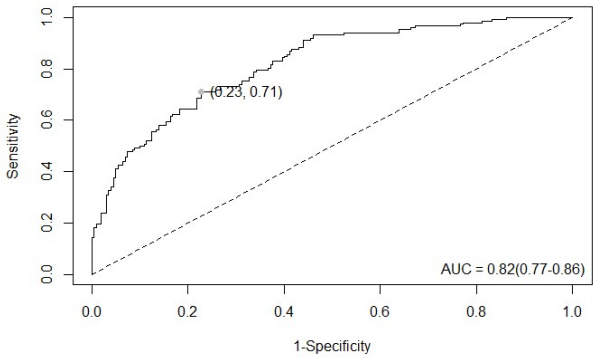

Figura 3: Curva ROC - AUC

Sobre a validação do modelo, as Figuras 2 e 3 apresentam alguns detalhes importantes. A aplicação da técnica da curva ROC retornou um ponto crítico que atinge o melhor equilíbrio entre a sensibilidade (verdadeiros positivos) e a especificidade (falsos positivos) igual a 0.455, enquanto que a área abaixo da curva que mede o poder discriminatório do modelo resultou um valor de 0.82, indicando boa capacidade de discriminação do modelo.

CONCLUSÕES:

O modelo se comportou bem em relação aos resíduos e pareceu conseguir discriminar bem as observações de acordo com a curva ROC, conseguindo identificar aproximadamente 75% das equipes vencedoras. Além disso, a análise identificou como relevante para o estudo as seguintes variáveis: chutes no alvo, cruzamentos, *rank* (posição da seleção no *ranking* da FIFA na época que foi disputada a copa), *clearances*, duelos aéreos, *tackles* e interceptações.

O estudo se mostrou satisfatório, entretanto para conclusões mais sólidas e relevantes, uma amostra maior juntamente com um maior leque de variáveis explicativas possibilitaria uma análise mais fundamentada e com resultados mais concisos.

BIBLIOGRAFIA

Castellano J, Casamichana D, Lago C. The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams. *J Hum Kinet*. 2012

COX, D. R.; SNELL, E. J. Analysis of binary data. 2. ed. Filadélfia, PA, USA: Chapman & Hall/CRC, 1989.

Lago-Peñas C, Lago-Ballesteros J, Dellal A, Gómez M. Game-Related Statistics that Discriminated Winning, Drawing and Losing Teams from the Spanish Soccer League. J Sports Sci Med. 2010

Liu H, Gomez MÁ, Lago-Peñas C, Sampaio J. Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup. J Sports Sci. 2015

Polo TCF, Miot HA. Aplicações da curva ROC em estudos clínicos e experimentais. J Vasc Bras. 2020;19: e20200186.

Rigby, R.A. and Stasinopoulos, D.M. (2004) Smooth Centile Curves for Skew and Kurtotic Data Modelled Using the Box-Cox Power Exponential Distribution. Statistics in Medicine, 23, 3053-3076.

Stasinopoulos, D. M., and Rigby, R. A. (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. Journal of Statistical Software, 23(7), 1–46.