RBFF
Revista Brasileira de Futsal e Futebol

MODELO PREDITIVO APLICADO AO FUTEBOL BRASILEIRO

Bruno Melo da Silva[1]

RESUMO

Estimar o resultado do campeonato brasileiro de 2021, através de um modelo de regressão de 2003 a 2020. Analisando através do software Excel 365 e R-Studio, o segundo responsável por achar a melhor equação dentre as variáveis analisadas com um p-valor (0,05) e conseguiu um R^2=91,73%, que foram obtidas através do Kaggle. Utilizando da regressão linear para obter uma equação para tentar prever o resultado do brasileiro 21 e posteriormente utilizamos o método stepwise para encontrar uma equação mais curta e eficiente, pode-se observar a tabela estimada, onde se deve levar em consideração que o trabalho é feito por médias e uma evolução pode alterar sua posição estimada. Desta forma foi obtida duas tabelas de previsão onde a primeira com todas as variáveis e a segunda com três, assim acertando todas as posições e com 35% de acerto da pontuação.

Palavras-chave: Futebol. Coleta de dados. Ciência de dados. Análise de dados.

ABSTRACT

Predictive model applied to brazilian football

Estimate the result of the 2021 Brazilian championship, through a regression model from 2003 to 2020. Analyzing through Excel 365 and R-Studio software, the second responsible for finding the best equation among the variables analyzed with a p-value (0.05) and achieved an R^2=91.73%, which were obtained through Kaggle. Using linear regression to obtain an equation to try to predict the result of the Brazilian 21 and then use the stepwise method to find a shorter and more efficient equation, we can observe the estimated table, where it must be taken into account that the work is done by averages and an evolution can change its estimated position. In this way, two prediction tables were obtained, where the first one with all the variables and the second one with three, thus getting all the positions right and with 35% of accuracy of the score.

Key words: Football. Data collect. Data Science. Data analysis.

1 - Universidade Federal de Sergipe, Brasil.

E-mail do autor:
brunomeloufs@hotmail.com

RBFF
Revista Brasileira de Futsal e Futebol

INTRODUÇÃO

Na perspectiva nacional, é usual que aos domingos as famílias parem em frente a tv para assistir o futebol, porque este é o esporte mais visto pelos brasileiros e que gera uma euforia diante dos campeonatos em que seus times participam.

É fato que por este motivo o futebol virou um dos maiores entretenimento da família brasileira, assim movimentando milhões de pessoas neste país e contribuindo para a economia brasileira.

No Brasil o futebol tem um potencial enorme de gerar emprego e contabiliza-se: 300 mil empregos direto, 30 milhões de praticantes (formal e não forma) e 580 mil participantes de formais em 13 mil times Leoncini, Silva, 2004.

Neste contexto é obvio que a indústria do futebol movimenta muito dinheiro e segundo relatório do Plano de Modernização de Futebol Brasileiro (Leoncini, Silva,2004), realizado pela Fundação Getúlio Vargas em parceria com a Confederação Brasileira de Futebol (2017), estima-se que o futebol mundial movimente em média 250 bilhões de dólares anuais, mas podemos imaginar que atualmente estes valores sejam bem maiores.

Em 2000/2001 a transferência mais cara de um jogador segundo o site transfermarkt foi do jogador português Luís Figo para o Real Madrid por cerca de € 60 milhões de Euros na cotação atual isso seria por volta de R$ 330 milhões de reais, valor do euro em relação ao real é de R$ 5,49, enquanto em 2017/2018 foi realizada a maior transferência de um jogador na história. O jogador brasileiro, Neymar, foi contratado pelo PSG por € 220 que na cotação atual R$ 1.218 por um único jogador.

Acredito que com estes relatos já é entendível o motivo e a necessidade de pesquisas voltada ao esporte que mais apaixonante do mundo, e já é possível perceber o aumento de pesquisas nesse sentido como também o movimento do mercado de clubes por colocarem centros de inteligência dentro de suas sedes.

Lá em 2019 quando o Liverpool se consagrou campeão da Liga dos Campeões o site (Época Negócios, 2002) soltou a matéria “Como o Liverpool usou dados e tecnologia para vencer a Champions League” e que contém relatos de como o time usou a inteligência de dados para tentar diminuir a distância financeira dos outros clubes europeus e que deu bons frutos por sinal.

Isso acontece porque cada passe, chute, gol, cartão e entre outros dados do jogo são coletados e juntos podem ser utilizados para inferir estatísticas. Enquanto você provavelmente assiste com o celular na mão e não percebe todos os detalhes do jogo, tem alguém coletando esses números para utilizar para tomar decisões em seus clubes por exemplo.

O cenário mais fácil de entender a importância desses dados é a coleta de pênaltis batidos por um jogador, existe jogadores que só batem de um lado e se existe alguém em sua equipe que analisa esses dados pode contar ao goleiro e tentar diminuir a chance deste jogador converter o pênalti, é muito comum durante a transmissão dos jogos aparecer os locais que o jogador geralmente bate. Ou seja, é um número já presente no nosso cotidiano e serve de exemplo para que você possa entender mais sobre essa ciência que é a estatística.

O valor do dado estatístico, na prática desportiva sempre foi colocada por especialistas como um grande avanço de qualidade no esporte (Vendite, Vendite, Moraes, 2005).

Entendendo a paixão do brasileiro nesse esporte, me incluo nisso, este trabalho tem como objetivo tentar prever o resultado do campeonato brasileiro de futebol masculino de 2021 através de um modelo de regressão linear múltipla com dados de todos os anos (2003-2020) do brasileirão desde que se tornou nesse modelo de pontos corridos.

O campeonato, de hoje conta, com 20 times, onde todos se enfrentam totalizando 38 rodadas, 19 jogadas com o seu mando, o clube que recebe escolhe o seu campo, as outras 19 fora de casa. Finalizado o campeonato ganha quem atingir o maior número de pontos.

Caso empate tem os critérios de desempates, como saldo de gols, gols feitos menos os gols recebidos, entre outros.

MATERIAIS E MÉTODOS

Será exposta uma análise estatística feita entre 2003 e 2020, que através de uma análise de regressão aplicada a estes anos pretende-se estimar o resultado do brasileirão 2021 e comparar com o resultado real.

RBFF
Revista Brasileira de Futsal e Futebol

Os dados foram retirados Kaggle que é uma subsidiária da Google LLC, é uma comunidade on-line de cientistas de dados e profissionais de aprendizado de máquina (Wikipedia, 2022).

Lá você pode fazer o download de alguns bancos de dados e até competir por soluções em que os donos do banco de dados premiam em dinheiro a melhor solução para o caso dela.

Foram levantados os seguintes dados: Pontuação (PTS), Vitórias (V), Derrotas (D), Empates (E), Gols Contra (GC), Gols a Favor/Gols Pró (GP) e Saldo de Gols (SG).

De todos os dados foram tiradas as médias, por exemplo, dividindo o número de gols totais pelo número de partidas, porque não faria sentido utilizar o total já que lá em 2003 o campeonato é disputado por 24 clubes e no formato atual são 20.

Além do fato, pretendo testar esse modelo para o campeonato brasileiro de 2022 e utilizando os totais só seria possível com a finalização do campeonato e deste modo ele será realizado rodada a rodada.

A análise de regressão estuda a relação entre uma variável chamada dependente e outras variáveis chamada independentes. Elas são representadas por um modelo matemático, que as associam. Pode-se dizer que a dependente é a resposta e independente é a explicativa.

A regressão tem por objetivo explicar a intensidade das variáveis X (passes, gols sofridos...) sobre o Y(pontos). Estimando o intercepto (β0) e seu fatores (β1... βn), explicando melhor seria:

$$Y = \beta 0 + \beta 1X1 + \cdots + \beta nXn$$

Onde esse n significa o número de fatores analisados, assim pode-se observar um Xn após cada beta, imaginemos que o X1 é o número de gols e o beta obteve um resultado de 0,5, pode-se dizer que a cada 2 gols o time aumenta um ponto.

Depois será feito uma análise da melhor função, separando as variáveis mais relevantes para conseguir os melhores resultados (pontos), ou seja, o mais próximo possível com o menor número de elementos.

Existe métodos para isso e com o auxílio do software R-studio conseguirá encontrar este sistema, onde a ideia é ir testando como os pontos se comportam adicionando e retirando as variáveis, ao final obtendo o que se pode chamar de modelo ideal.

O nome desse teste é Regressão stepwise, ela é uma ferramenta automática usada nos estágios exploratórios da construção de modelos para identificar um subconjunto útil de preditores. O processo adiciona sistematicamente a variável mais significativa ou remove a variável menos significativa durante cada etapa (Suporte Minitab, 2022).

RESULTADOS

Inicialmente faremos a análise de regressão comum, com todos os dados e observaremos o resultado para tomar as decisões diante do modelo. Para isso vamos importar a tabela par o software R.

```
Brasileirao_historico
< − read_excel(C:/Users/bruno.silva/Desktop/Dash
    _Previsão_BR/Brasileirao_R.xlsx, sheet
                = Ranking)
```

Para realizar o experimento, como já abordamos acima, temos que tirar a média de todos as variáveis, então:

```
Brasileirao_historico$Pontuação_Média
= (Brasileirao_historico$Pontuação
/ Brasileirao_historico$Partidas)
```

Deste modo criamos nossas colunas que iremos usar para o modelo de regressão, o próximo passo é retirar os dados de 2021 da tabela.

```
Br
= Brasileirao_historico[Brasileirao_historico$Ano !
= 2021, ]
```

Também devemos retirar as outras colunas que não serão necessárias para o modelo em questão, então:

```
Br_2 = Br[, c(4,12: 18)]
```

Finalmente podemos rodar o modelo de regressão com os dados e verificar o resultado e já sabemos que um valo maior que 0,05 será estatisticamente insignificante.

RBFF
Revista Brasileira de Futsal e Futebol

$$lm\_Br = lm(Pontuação \sim ., Br\_2)$$

Residuals:

Min	1Q	Median	3Q	Max
-5.0735	-2.1687	-1.0271	0.2476	13.5892

Coefficients: (2 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-8.530	3.285	-2.597	0.00979	**
Pontuação_Média	31.995	6.204	5.157	4.13e-07	***
Vitórias_Média	20.770	19.535	1.063	0.28840	
Empates_Média	7.135	7.562	0.943	0.34605	
Derrotas_Média	NA	NA	NA	NA	
SG_Média	-3.544	1.456	-2.434	0.01543	*
GP_Média	6.902	1.012	6.821	3.78e-11	***
GC_Média	NA	NA	NA	NA	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.633 on 364 degrees of freedom
Multiple R-squared: 0.9183, Adjusted R-squared: 0.9171
F-statistic: 817.7 on 5 and 364 DF, p-value: < 2.2e-16

Perceba que somente Pontuação_Média, SG_Média e GP_Média obtiveram resultados aceitáveis em seu p-valor e por este motivo devemos rodar o stepwise para que ele ache o melhor modelo com essas variáveis.

Mas para contextualizar o que aconteceria com esse modelo, vamos aplicar a equação aos dados de 2021 e comparar com o resultado real. Onde tiver NA vamos considerar zero, logo ele não precisa ser calculado já que seria zero qualquer número que multiplique por ele.

$$PTS = -8.53 + 31.99 * Pontuação_{Média} + 20.77 Vitórias_{Média} + 7.13 * Empates_{Média} + NA * Derrotas_{Média} - 3.54 * SG_{Média} + 6.90 * GP_{Média} + NA * GC_{Média}$$

Na tabela 1 se nota que ele teve uma acertabilidade de 70%, acertando 14 posições embora que os 6 erros de posição é com o time acima, como no caso do Santos e Ceará em que o modelo acertou a pontuação de ambos que foi 50 pontos mas diante do empate acredito que o Excel ajustou por ordem alfabética e eu contei como um "erro" do modelo que não é real pois deveríamos utilizar o critério utilizado no regulamento da CBF que em caso de empate o desempate acontece pelo número de vitórias em seu primeiro critério e em caso de continuar o empate compara outras variáveis.

CBF, 2021 - Art. 14 - Em caso de empate em pontos ganhos entre 2 (dois) ou mais clubes ao final do CAMPEONATO, o desempate, para efeito de classificação final, será efetuado observando-se os critérios abaixo. 1º) maior número de vitórias; 2º) maior saldo de gols; 3º) maior número de gols pró; 4º) confronto direto; 5º) menor número de cartões vermelhos recebidos; 6º) menor número de cartões amarelos recebidos; 7º) sorteio.

§ 1º – Para efeito do quarto critério (confronto direto), considera-se o resultado dos jogos de ida e volta somados, ou seja, o resultado do "jogo de 180 (cento e oitenta) minutos".

§ 2º – No caso de empate entre mais de 2 (dois) clubes, não será considerado o quarto critério.

Deste modo o Santos está à frente do Ceará em quesito de desempate, isso cabe um ajuste manual já que não tem como dizer para o modelo que se os times empatarem vale o critério x ou y.

Já no caso do Athletico e São Paulo não tem jeito, pois o modelo estimou este com um ponto a menos que o real e o time paranaense tem duas vitórias a mais que o paulista, não conseguindo contornar, outro erro foi os dois pontos a mais que ele estimou para o bragantino e aumentando sua pontuação em relação ao Corinthians.

Neste momento vamos aplicar a stepwise e verificar o que o software entende como o melhor modelo que obtenha um melhor resultado com uma menor quantidade de variáveis e o resultado pode ser observado a seguir

RBFF
Revista Brasileira de Futsal e Futebol

```
step_Br = step(lm_Br)
Residuals:
    Min      1Q  Median      3Q     Max
-5.0263 -2.1597 -1.0415  0.2497 13.1348

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -7.6933     2.5883  -2.972  0.00315
**
Pontuação_Média   38.3438     1.7132  22.381
< 2e-16 ***
SG_Média          -3.1452     1.3154  -2.391
0.01731 *
GP_Média           6.9165     0.9624   7.187 3.75e-
12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Residual standard error: 3.629 on 366 degrees
of freedom
Multiple R-squared:  0.918,    Adjusted R-
squared:  0.9173
F-statistic:  1366 on 3 and 366 DF, p-value: <
2.2e-16
```

O resultado do Adjusted R-squared: 0.9173 é superior ao obtido com o modelo anterior por 0,02%, foram resultados muito próximo, mas com um menor número de variáveis, o que facilita os cálculos. A equação ficou com o seguinte resultado:

$$\text{PTS} = -7.6933 + \text{Pontuação}_{\text{Média}} * 38.3438 - \text{SG}_{\text{Média}} * 3.14 + \text{GP}_{\text{Média}} * 6.91$$

A equação foi reduzida drasticamente de 7 variáveis mais o intercepto para somente 3, e demonstrando um r-quadrado ajustado um pouco maior, por este fato o stepwise é de suma importância para que consigamos encontrar um modelo sucinto com resultado igual ou superior ao atual.
Agora vamos comparar o resultado da nossa segunda regressão com o resultado real do brasileirão 2021.

Tabela 1 - Comparativo Estimado X Real Regressão 1.

Posição Estimada	Estimado	Estimado = Real	Real	Posição Real	Erro
Atlético/MG	87	Verdadeiro	84	Atlético/MG	-3
Flamengo/RJ	74	Verdadeiro	71	Flamengo/RJ	-3
Palmeiras/SP	68	Verdadeiro	66	Palmeiras/SP	-2
Fortaleza/CE	59	Verdadeiro	58	Fortaleza/CE	-1
Bragantino/SP	58	Falso	57	Corinthians/SP	0
Corinthians/SP	57	Falso	56	Bragantino/SP	-2
Fluminense/RJ	54	Verdadeiro	54	Fluminense/RJ	0
América/MG	53	Verdadeiro	53	América/MG	0
Atlético/GO	52	Verdadeiro	53	Atlético/GO	1
Ceará/CE	50	Falso	50	Santos/SP	0
Santos/SP	50	Falso	50	Ceará/CE	0
Internacional/RS	49	Verdadeiro	48	Internacional/RS	-1
Atlético/PR	47	Falso	48	São Paulo/SP	1
São Paulo/SP	47	Falso	47	Atlético/PR	0
Cuiabá/MT	46	Verdadeiro	47	Cuiabá/MT	1
Juventude/RS	46	Verdadeiro	46	Juventude/RS	0
Grêmio/RS	44	Verdadeiro	43	Grêmio/RS	-1
Bahia/BA	44	Verdadeiro	43	Bahia/BA	-1
Sport/PE	36	Verdadeiro	38	Sport/PE	2
Chapecoense/SC	16	Verdadeiro	15	Chapecoense/SC	-1

RBFF
Revista Brasileira de Futsal e Futebol

Tabela 2 - Comparativo Estimado X Real Regressão 2+

Posição Estimada	Estimado	Estimado = Real	Real	Posição Real	Erro
Atlético/MG	87	Verdadeiro	84	Atlético/MG	-3
Flamengo/RJ	74	Verdadeiro	71	Flamengo/RJ	-3
Palmeiras/SP	68	Verdadeiro	66	Palmeiras/SP	-2
Fortaleza/CE	59	Verdadeiro	58	Fortaleza/CE	-1
Corinthians/SP	58	Verdadeiro	57	Corinthians/SP	-1
Bragantino/SP	57	Verdadeiro	56	Bragantino/SP	-1
Fluminense/RJ	54	Verdadeiro	54	Fluminense/RJ	0
América/MG	53	Verdadeiro	53	América/MG	0
Atlético/GO	52	Verdadeiro	53	Atlético/GO	1
Santos/SP	50	Verdadeiro	50	Santos/SP	0
Ceará/CE	50	Verdadeiro	50	Ceará/CE	0
Internacional/RS	49	Verdadeiro	48	Internacional/RS	-1
São Paulo/SP	48	Verdadeiro	48	São Paulo/SP	0
Atlético/PR	47	Verdadeiro	47	Atlético/PR	0
Cuiabá/MT	46	Verdadeiro	47	Cuiabá/MT	1
Juventude/RS	46	Verdadeiro	46	Juventude/RS	0
Grêmio/RS	44	Verdadeiro	43	Grêmio/RS	-1
Bahia/BA	44	Verdadeiro	43	Bahia/BA	-1
Sport/PE	36	Verdadeiro	38	Sport/PE	2
Chapecoense/SC	16	Verdadeiro	15	Chapecoense/SC	-1

Como pode ser observado acima o modelo acertou todas as posições do campeonato brasileiro, trazendo um resultado excelente e acima do esperado.

É evidente que o objetivo é a acertabilidade na pontuação dos times e por consequência que vem a colocação, mas visto que a somatória dos erros é -9, considero um bom resultado diante dos dados que podem ser tirado de qualquer tabela do brasileirão disponível na internet, pois estes dados é o padrão mais visto ao que concerne a tabela de pontos corridos do futebol profissional.

CONCLUSÃO

Levando em consideração todas as análises feitas, pode-se afirmar que por mais que ele tenha acertado 100% das posições ele teve uma acertabilidade de 35% nos pontos, assumindo assim que um pouco mais de 1/3 dos dados são explicado pelo modelo encontrado.

Deve ser destacada, também, a suposição de que uma temporada está relacionada a outra, pois tem-se 16 dos 20 clubes do campeonato anterior no atual, apesar da variação de elenco nos clubes, o que está sendo calculado são médias que eles desempenharam ao decorrer dos jogos.

Diante de toda a análise, será considerada a segunda equação, pois foi obtida uma função menor com um r quadrado muito próximo ao outro cálculo, 0.9173 para 0.9171.

De acordo com a tabela encontrada, o Atlético Mineiro seria o campeão, Flamengo, Palmeiras-SP e Fortaleza-CE completam o grupo dos quatros, enquanto o Grêmio-RS, Bahia-BA, Sport-PE e Chapecoense/SC seriam os rebaixados de acordo com o desempenho de suas médias, resultado este que foi constatado ao fim do campeonato.

Apesar do 35% parecer um número muito baixo, eu acredito piamente que a análise foi um sucesso quando observamos as posições estimadas para os clubes.

Posteriormente irei aplicar para o ano vigente este modelo, desta vez com os dados de 2021 incluso e convido você a acompanha o dashboard que vou deixar o link ao final.

Vale ressaltar que os resultados se alteram de acordo com as médias dos clubes, partindo do pressuposto de que pelo menos um dos times do campeonato vigente aumente seus números, ou seja, melhore seu desempenho no campeonato, subirá de posição, fazendo com que a tabela se modifique rodada a rodada.

REFERNCIAS

1-Confederação Brasileira de Futebol. Disponível em<https://www.cbf.com.br>.24/10/2017.

2-CBF. Regulamento Específico da Competição Brasileirão Assaí - 2021. Fonte: CBF: https://conteudo.cbf.com.br/cdn/202104/20210414232127_16.pdf 14/04/2021.

3-Época Negócios. 2002. Fonte: Globo: https://epocanegocios.globo.com/Empresa/noticia/2019/06/como-o-liverpool-usou-dados-e-tecnologia-para-chegar-final-da-champions-league.html. 21/03/2022.

4-kaggle. https://www.kaggle.com/datasets/rczekster/matches-brazilian-football-from-2003-to-2019/metadata. kaggle. 21/03/2022.

5-Leoncini, M. P.; Silva, M. T. Entendendo o futebol como um negócio: um estudo exploratório. Gestão e Produção. p. 11-23. 2004.

6-Suporte Minitab. 2022. Fonte: Minitab: https://support.minitab.com/pt-br/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/basics-of-stepwise-regression/. 03/03/2022.

7-Wikipedia. Fonte: wikipedia: https://en.wikipedia.org/wiki/Kaggle. 21/03/2021.

8-Vendite, C.C.; Moraes, A.C.; Vendite A.C. Scout no Futebol: Uma ferramenta Para a Imprensa Esportiva. 2005. Disponível em <www.portcom.intercom.org.br/pdfs/29839791442711236695040612710072498671.pdf>. 24/10/2017.

Recebido para publicação em 21/03/2022
Aceito em 01/06/2022