CENTRO UNIVERSITÁRIO DE BELO HORIZONTE
Graduação - Ciência da Computação

JOSUÉ FILIPE PONTES NERY - 119122537
MARCELA CRISTYNE - 11827313
THAIS CORDEIRO PEREIRA - 11821085

UC: ANÁLISE DE DADOS E BIG DATA

Prática 08 - Estatística Preditiva (Regressão Logística)

Belo Horizonte
2º Semestre de 2021

Objetivo: Vocês deverão obter os dados dos times rebaixados mais o primeiro time fora da zona de rebaixamento, que disputaram a série A do campeonato Brasileiro de 2010 a 2020. Os dados de interesse são: pontuação final, número de gols marcados, número de gols sofridos e, se possível, o quanto esses times viajaram, seja por total de km's viajados ou total de horas viajadas, ambos podendo ser valores aproximados. Logo, este último parâmetro não é obrigatório, embora ele deixa a análise mais interessante. Vocês deverão agregar à base de dados, uma coluna que representa se o time foi rebaixado ou não. Esta será nossa variável de resposta (desfecho).
Usem o RStudio para os itens abaixo. Não esqueçam de **interpretar os resultados**.

1) Descrever a fonte para a coleta dos dados.
Foi usado como referência os dados encontrados nos sites Transfermarkt, RSSSF Brazil e Wikipédia.

TRANSFERMARKT GMBH & CO. KG. **TRANSFERMARKT**. [S. l.], entre [2000 e 2021]. Disponível em: https://www.transfermarkt.com.br/serie-a/startseite/wettbewerb/BRA1. Acesso em: 10 nov. 2021.

DIOGO, Julio Bovi; REC.SPORT.SOCCER STATISTICS FOUNDATION; RSSSF BRAZIL. **Brazil - List of Champions.** [S. l.], entre [1997 e 2021]. Disponível em: https://rsssfbrasil.com/tablesae/brcamp.htm. Acesso em: 10 nov. 2021.

CAMPEONATO BRASILEIRO DE FUTEBOL DE 2010 - SÉRIE A. In: WIKIPÉDIA, a enciclopédia livre. Flórida: Wikimedia Foundation, 2021. Disponível em: <https://pt.wikipedia.org/w/index.php?title=Campeonato_Brasileiro_de_Futebol_de_2010_-_S%C3%A9rie_A&oldid=61282876>. Acesso em: 10 nov. 2021.

---

2) Criar o modelo final de Regressão Logística. Se alguma variável foi excluída, explique o motivo da exclusão para o modelo final. Apresente os valores ’p-value’, ’teste z’, ’erro padrão' e ’estimados’. Considere p-value significativo como 0.05.

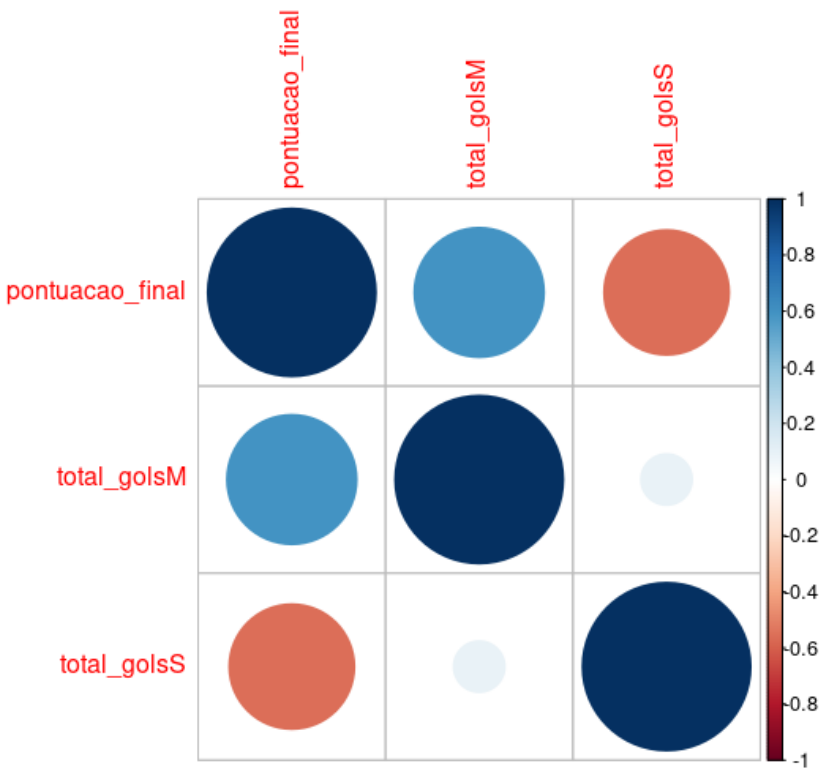
Nosso modelo possui três variáveis, sendo elas, pontuação final, total de gols marcados e total de gols sofridos. Temos os seguintes valores:

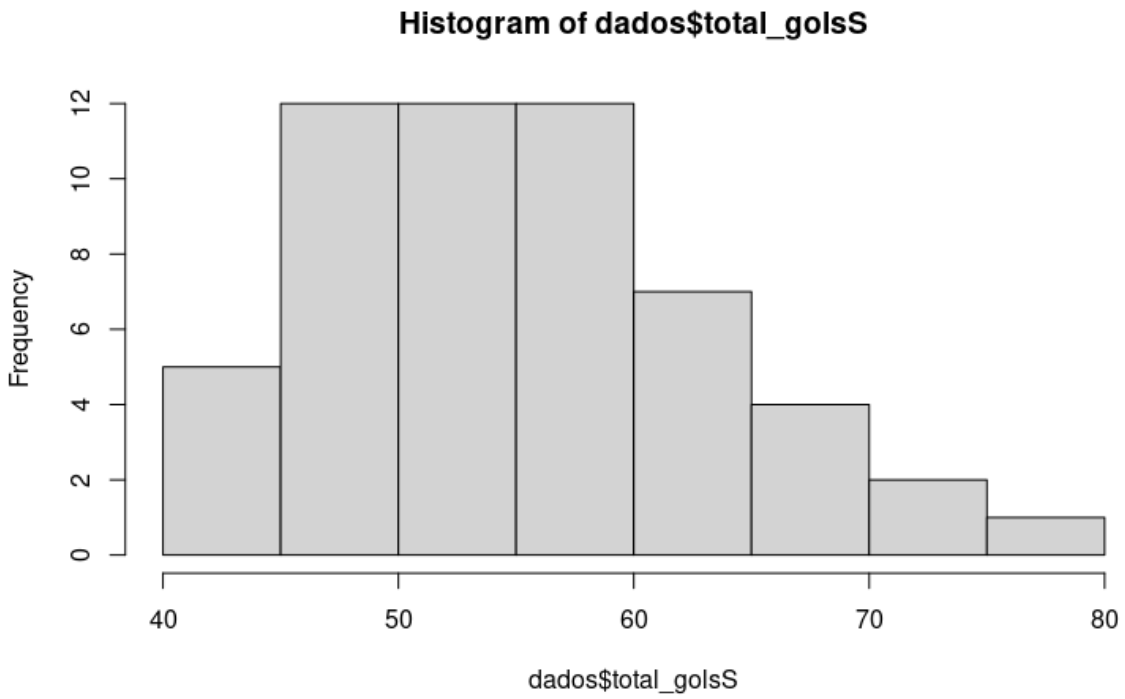
```
Call:
glm(formula = dados$rebaixado ~ dados$pontuacao_final + dados$total_golsM +
    dados$total_golsS, family = binomial, data = dados)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.34795   0.01913   0.19046   0.48769   1.94652

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          13.15562    9.31412   1.412    0.158
dados$pontuacao_final -0.34840    0.21248  -1.640    0.101
dados$total_golsM     -0.09637    0.08472  -1.137    0.255
dados$total_golsS      0.11437    0.08309   1.376    0.169

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.044  on 54  degrees of freedom
Residual deviance: 32.831  on 51  degrees of freedom
AIC: 40.831

Number of Fisher Scoring iterations: 7
```

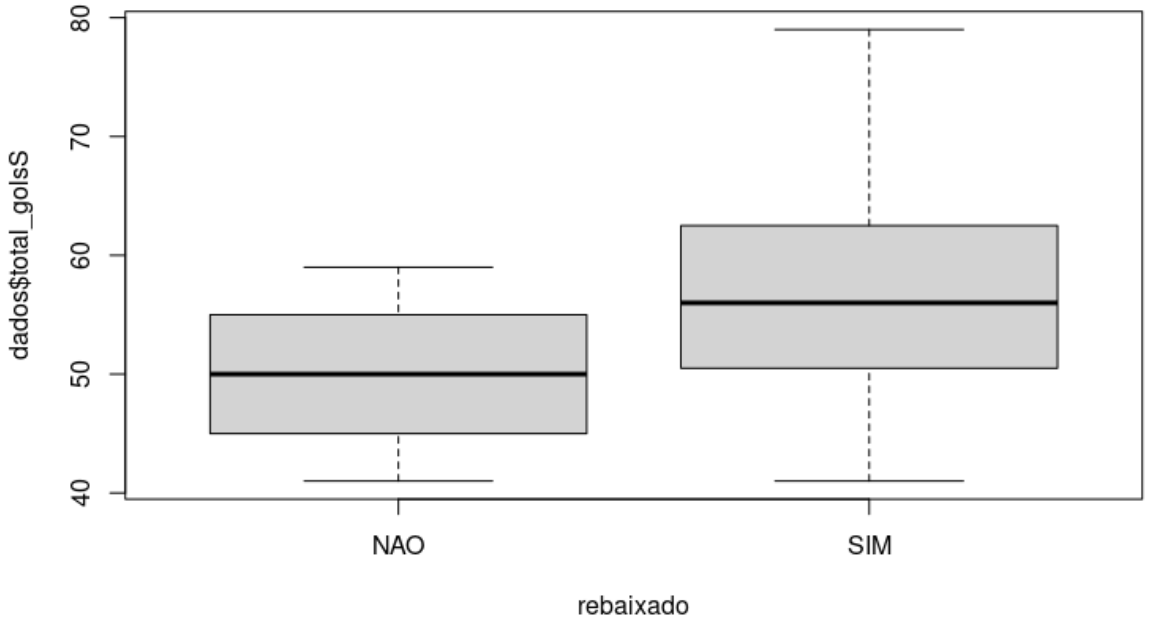
Como podemos perceber pela imagem acima, não temos a presença de algum erro padrão (Std. Error) significativo, assim como não temos “z valores” com forte intensidade do impacto na variedade de resposta, todos bem próximos de zero.

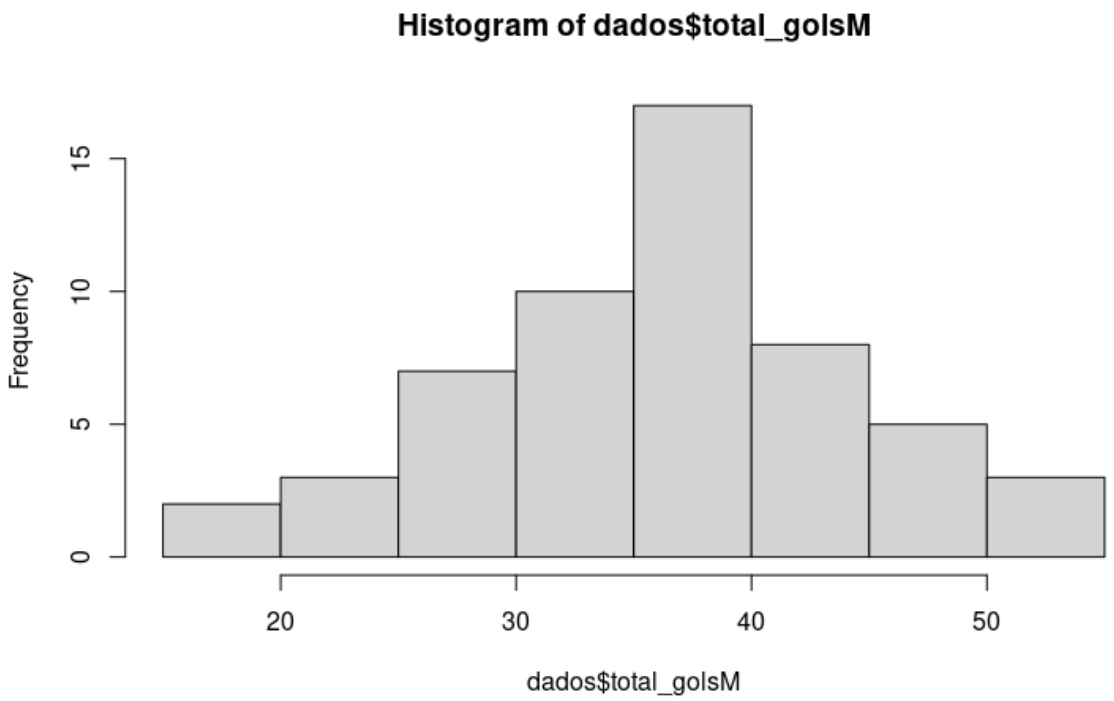
Outra importante observação é o P valor, que está sendo considerado significativo como 0.05, mas o nosso modelo obteve valores acima em todas as variáveis, ou seja, nenhuma variável possui valor significativo para nossa variável resposta.

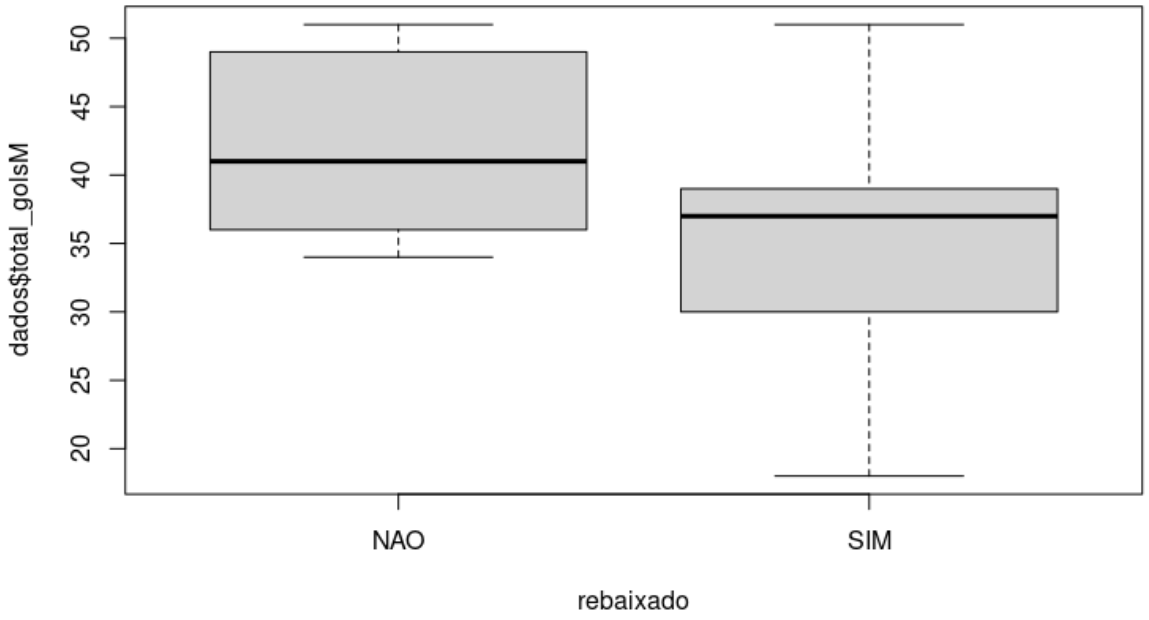
Foi pensado então em retirar algumas variáveis para ver o impacto no modelo e para chegar a essa conclusão plotamos os gráficos abaixo, sendo que podemos observar que existem interações significativas entre as variáveis auxiliares, tanto o total de gols marcados quanto o total de gols sofridos possuem correlações com a variável pontuação final, positiva e negativa respectivamente, mas não possuem correlação, **significativa,** entre si.

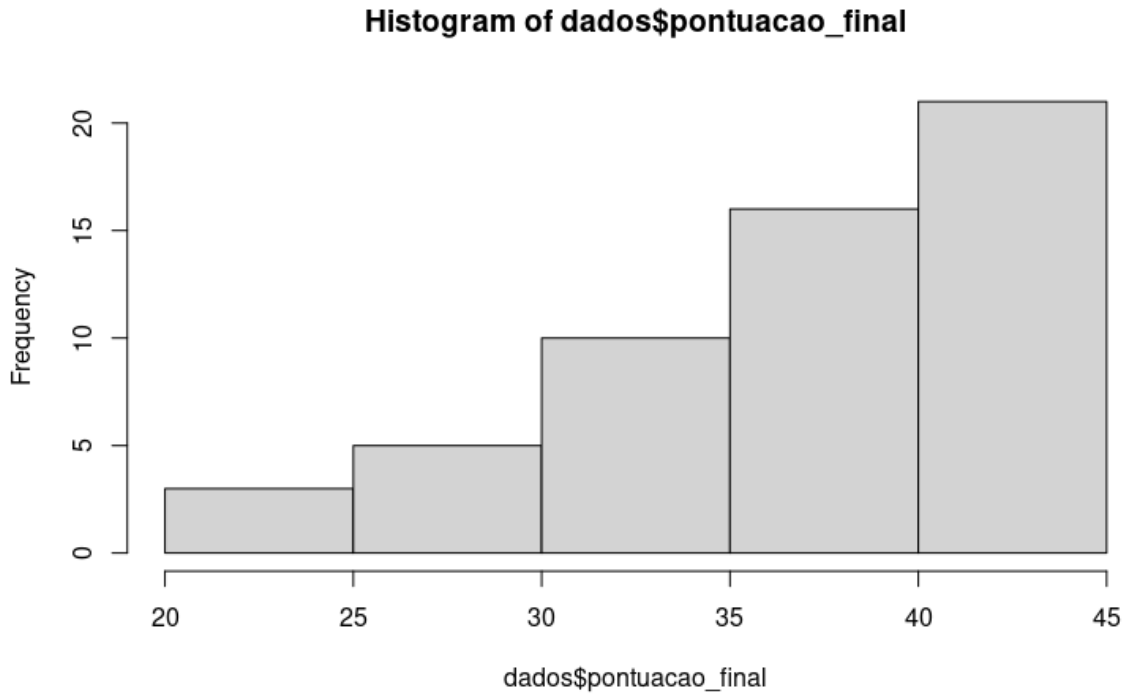

pontuacao_final
total_golsM
total_golsS
pontuacao_final
total_golsM
total_golsS
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

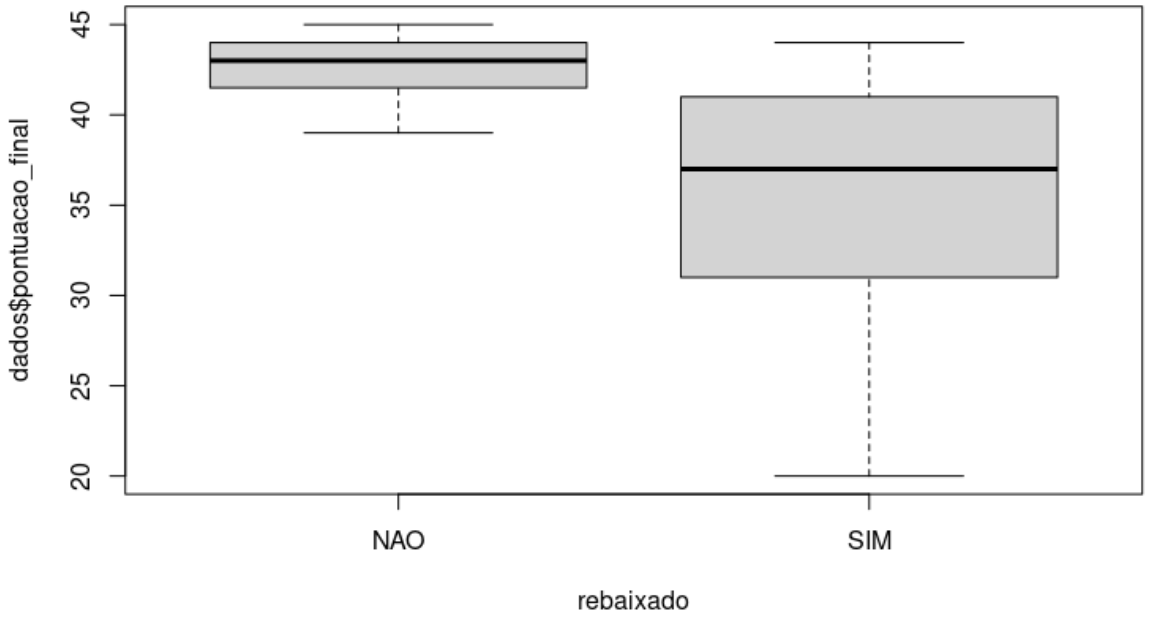

Histogram of dados$total_golsS
Frequency
0
2
4
6
8
10
12
40
50
60
70
80
dados$total_golsS


dados$total_golsS
40
50
60
70
80
NAO
SIM
rebaixado


Histogram of dados$total_golsM
Frequency
0
5
10
15
20
30
40
50
dados$total_golsM


dados$total_golsM
20
25
30
35
40
45
50
NAO
SIM
rebaixado


Histogram of dados$pontuacao_final
Frequency
0
5
10
15
20
20
25
30
35
40
45
dados$pontuacao_final



3) Apresente a Matriz de confusão, o limiar usado para a classificação, a taxa de acerto em média, a curva ROC e o valor AUC.
Foi utilizado o limiar de 0.7 (70%) e o valor de AUC foi 0.9607438, a média de acerto foi de 88,63636%.

pred	**0**	**1**
0	42	8
1	2	36