Análise NBA

Paola de Oliveira Prado

19/04/2021

Objetivo

O objetivo desta análise é encontrar um modelo estatístico capaz de identificar atletas como possíveis promissores a uma carreira ou não na NBA, com base em algumas de suas estatísticas no seu primeiro ano de NBA.

Base de Dados

A base é composta por estatísticas de 1340 atletas obtida durante seu ano de estreia na NBA. A base possui 21 variáveis, e 1 delas é a variável resposta que indica 1 se o jogador ficou mais de 5 anos ou mais na liga, ou caso 0 caso contrário. Mais informações pode ser encontrada no link https://data.world/lrdegeest/stats250 (https://data.world/lrdegeest/stats250).

```
library(kernlab)
library(caret)
library(gbm)

base = read.csv("https://query.data.world/s/6uhvcowok54gcj5jz3i2ccwldlb6rl",
                header=TRUE, stringsAsFactors=FALSE)
```

Pré-Processamento

A primeira parte do pré-processamento foi verificar a existência de linhas duplicadas, e foram encontradas 12 linhas. Essas linhas foram retiradas da base de dados.

```
duplicados = duplicated(base,fromLast = TRUE)
which(duplicados) #12 linhas duplicadas
base=base[!duplicados,]

dados = base[,-1] # o nome do jogador não é relevante para avaliar se ele será promissor ou não

dados$TARGET_5Yrs=as.factor(dados$TARGET_5Yrs)
```

O segundo passo foi verificar a existência de valores nulos, foram verificados 10 valores nulos na variável X3P. Os valores nulos foram tratados pelo método *k-Nearest Neighbors*, kNN, que consiste em substituir os dados faltantes pela média dos seus *k* vizinhos mais próximos, nesse caso, considerou-se 5 vizinhos. Após, foi verificada a correlação entre as variáveis, e verificou-se que 10 variáveis possuem alta correlação, ou seja, correlação acima de 0,75. Dessa forma, as variáveis:“PTS”, “MIN”, “FGM”, “FGA”, “FTM”, “FTA”, “REB”, “DREB”, “STL”, “X3PA”, foram retiradas da base. No final do pré-processamento, a base de dados ficou com 1328 observações e 10 variáveis.

```
# Tratando Na's
sum(is.na(base$x3p.))
apply(dados,2, function(x) any(is.na(x)))
preproc_NA = preProcess(dados,method = "knnImpute",k=5)
dados_na = predict(preproc_NA,dados)

# correlação ponto de corte 0.75

var_num= c(names(Filter(is.numeric,dados_na)))
descrCor =  cor(dados_na[var_num])

#Quais variaveis tem alta correlacao?
findCorrelation(descrCor, cutoff = .75, verbose=T)

#Novo banco sem var. com alta correlacao
highCor=findCorrelation(descrCor, cutoff = .75,names=T)

dados_cor = dplyr::select(dados_na,-highCor)
```

Regressão Logística

A regressão logística foi o método escolhido para criar uma modelo capaz de prever os atletas do NBA com futuro promissor. O modelo será calculado com base em 9 variáveis que são estatísticas desses atletas referentes ao seu primeiro ano de NBA. As variáveis são: “GP”, “FG.”, “X3P.Made”, “X3P.”, “FT.”, “OREB”, “AST”, “BLK”, “TOV”.

A base de dados foi separada em 75% para treino e 25% para teste.

```
set.seed(2021)

inTrain = createDataPartition(dados_cor$TARGET_5Yrs,p=0.75,list=F)
treino = dados_cor[inTrain,]
teste  = dados_cor[-inTrain,]
```

Foi utilizado o método de reamostragem bootstrap que consistem gerar uma nova amostra do mesmo tamanho da amostra original a partir de uma seleção aleatória com reposição de seus elementos. Para esse caso, foi realizado o método com 3 repetições.

```
library("VGAM")
library("e1071")

ctrl = trainControl(method="boot", number=3)

model = train(TARGET_5Yrs~., data=treino, trControl=ctrl, method="vglmAdjCat")

predicao = predict(model,teste)
```

```
confusionMatrix(predicao, teste$TARGET_5Yrs, positive = '1')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  65  35
##          1  61 171
##
##                Accuracy : 0.7108
##                  95% CI : (0.6588, 0.759)
##     No Information Rate : 0.6205
##     P-Value [Acc > NIR] : 0.0003446
##
##                   Kappa : 0.3604
##
##  Mcnemar's Test P-Value : 0.0107244
##
##             Sensitivity : 0.8301
##             Specificity : 0.5159
##          Pos Pred Value : 0.7371
##          Neg Pred Value : 0.6500
##              Prevalence : 0.6205
##          Detection Rate : 0.5151
##    Detection Prevalence : 0.6988
##       Balanced Accuracy : 0.6730
##
##        'Positive' Class : 1
##
```

Após rodar o modelo, foi feita a predição da variável Target na base de teste. Como resultado temos que o modelo possui uma taxa de sensitividade de 83,01% que corresponde a porcentagem de atletas promissores que o modelo classificou corretamente. Enquanto a taxa de especificidade foi de 51,59% que corresponde a porcentagem de atletas que não foram promissores que o modelo conseguiu classificar corretamente. A acurácia do classificador, ou seja, a taxa de acerto foi de 71,08%.