NBA finals prediction model

The final of the National Basketball Association is a combination of 7 matches/games played by the two opposite teams. The NBA has collects 2 types data for every single match: team-wise and player-wise. We will use the team-wise data of only the finals to predict whether the team wins or not.

**Our data consists** of 24 variables:

#Year: year the corresponding team won the championship

#Team: the team that won the championship

#Game: 7 games are played in the NBA finals, thisis the nth no. of game being played

#Win: if the team won or not

#Home: if the game was played at home court or not

#minutes: minutes for which the game was played

#FG: field goals are baskets scored apart from free throws. this is the number of field goals scored

#FGA: field goals attempted

#FGP: field goals percentage (FG/FGA)

#TP: three pointers scored

#TPA: three pointers attempted

#TPP: three pointers percentage (TP/TPA)

#FT: free throws made

#FTA: free throws attempted

#FTP: free throws percentage (FT/FTA)

#ORB: offensive rebounds - rebounds collected in opponents’ side of court

#DRB: defensive rebounds - rebounds collected in own side of court

#TRB: total rebounds

#AST: total assists made

#STL: total steals

#BLK: total blocks made

#TOV: turnovers in the game

#PF: personal fouls

#PTS: points made by the winning team

We will predict “Win” by using logistic regression and lasso tool as it consists of binary values:

1-win

2-lose

**The prediction model code is as follows**:

b<-read.csv("C:/Users/voraa/Desktop/championsdata.csv")

names(b)

str(b)

dim(b)

sum(is.na(b))

a<-na.omit(b)

str(a)

a$Game<-as.factor(a$Game)

a$FGP<-scale(a$FGP)

a$TPP<-scale(a$TPP)

a$FTP<-scale(a$FTP)

a$ORB<-scale(a$ORB)

a$DRB<-scale(a$DRB)

a$AST<-scale(a$AST)

a$STL<-scale(a$STL)

a$BLK<-scale(a$BLK)

a$TOV<-scale(a$TOV)

a$PF<-scale(a$PF)

a$PTS<-scale(a$PTS)

library(caTools)

library(glmnet)

set.seed(123)

split = sample.split(a$Win,SplitRatio = 0.6)

train = subset(a,split == T)

test = subset(a,split == F)

Y = a$Win

X = model.matrix(Win~.,a)[,-1]

l1 <- glmnet(X,Y,alpha = 1)

X

plot(l1)

l1.cv <- cv.glmnet(X,Y,alpha = 1,nfolds = 10)

plot(l1.cv)

lambdamin = l1.cv$lambda.min

predict(l1,type="coefficients",s=lambdamin)[1:40,]

m1 <- glm(Win~.-Year-Home-minutes-FG-TP-FTA-ORB-BLK-AST-PTS,data = train,family = "binomial")

m1

summary(m1)

pred <- predict(m1,newdata = test,type = "response")

pred

pred\_win <- ifelse(pred>0.5,1,0)

pred\_win

table(test$Win,pred\_win)

m2 <- glm(Win~.-Year-Home-minutes-FG-TP-FTA-ORB-BLK-AST-PTS-TPA-FT-DRB-PF-Team,data = train,family = "binomial")

summary(m2)

pred1 <- predict(m2,newdata = test,type = "response")

pred1

pred\_win1 <- ifelse(pred1>0.5,1,0)

pred\_win1

table(test$Win,pred\_win1)

The above code predicts the outcome of the game with a 92.9% accuracy, thus, making the prediction a safe bet.

Question arises when we realize that a few data variables are directly recorded during the match. So, it compels us to make use of the average data of the complete season which the NBA records on a daily basis.

The **final model** consists of 7 variables: FGA,FGP,TPP,FTP,TRB,STL,TOV

Initially, we applied lasso to decide which variables to be considered. The first model, m1 consists of those variables. There were still some less significant variables which were discarded in the final model, leaving us with all the variables being significant and an accuracy of **92.9%.**

**Interpretation:**

After preparing this model, certain interpretations changed the way we look at the game.

Firstly, our model does not consist of the final points scored by the winning team. Never had we imagined that points would be less significant compared to other smaller factors.

It also provides us a good prediction irrespective whether a game is being played at home court or not. This is understandable because when it comes to the finals there is hardly any difference in every players’ focus level.

On the other hand, the most important factor that will help predict the winner is the FGP (field goal percentage). In other words, the number of baskets converted by the team matters the most. Some teams may score higher points but if their ‘scored: attempted’ ratio is low, they are less likely to win.