SAIL Application Paper

Abigail Mabe

2022-12-15

What are some of the most important factors that impact game-time strategies when considering success in the March Madness Tournament? And to what extent do the same factors play into certain college basketball regular season strategies?

library(readr)  
library(bestglm)

## Loading required package: leaps

library(leaps)  
  
  
BBall <- read\_csv('cbb19.csv', show\_col\_types = FALSE)  
  
head(BBall)

## # A tibble: 6 × 23  
## TEAM CONF G W ADJOE ADJDE BARTHAG EFG\_O EFG\_D TOR TORD ORB  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Gonzaga WCC 37 33 123. 89.9 0.974 59 44.2 14.9 19 31.5  
## 2 Virginia ACC 38 35 123 89.9 0.974 55.2 44.7 14.7 17.5 30.4  
## 3 Duke ACC 38 32 119. 89.2 0.965 53.6 45 17.5 19.4 35.6  
## 4 North Car… ACC 36 29 120. 91.4 0.958 52.9 48.9 17.2 18.3 35.3  
## 5 Michigan B10 37 30 115. 85.6 0.966 51.6 44.1 13.9 18 24.7  
## 6 Michigan … B10 39 32 120. 91 0.960 55.2 43.9 18.5 14.9 33.9  
## # … with 11 more variables: DRB <dbl>, FTR <dbl>, FTRD <dbl>, `2P\_O` <dbl>,  
## # `2P\_D` <dbl>, `3P\_O` <dbl>, `3P\_D` <dbl>, ADJ\_T <dbl>, WAB <dbl>,  
## # POSTSEASON <chr>, SEED <dbl>  
## # ℹ Use `colnames()` to see all variable names

The variables:

TEAM: Division 1 college basketball school

CONF: The Athletic conference in which the school is in

G: Number of Games played

W: Number of Games won

ADJOE: Adjusted Offensive Efficiency (An estimate of defensive efficiency (points allowed per 100 possessions) a team would have against an average Division 1 offense)

ADJDE: Adjusted Defensive Efficiency (An estimate of the defensive efficiency (points allowed per 100 possessions) a team would have against the average Division 1 offense)

BARTHAG: Power Rating (Chance of beating an average Division 1 team)

EFG\_O: Effective Field Goal Percentage Shot

EFG\_D: Effective Field Goal Percentage Allowed

TOR: Turnover Percentage Allowed (Turnover Rate)

TORD: Turnover Percentage Committed (Steal Rate)

ORB: Offensive Rebound Rate

DRB: Offensive Rebound Rate Allowed

FTR: Free Throw Rate (How often the given team shoots Free Throws)

FTRD: Free Throw Rate Allowed

2P\_O: Two-Point Shooting Percentage

2P\_D: Two-Point Shooting Percentage Allowed

3P\_O: Three-Point Shooting Percentage

3P\_D: Three-Point Shooting Percentage Allowed

ADJ\_T: Adjusted Tempo (An estimate of the tempo (posessions per 40 minutes) a team would have against the team that wants to play at an average Division 1 tempo)

WAB: Wins Above Bubble

POSTSEASON: Round where given team was eliminated

SEED: Seed in March Madness

YEAR: Season

BBallTour <- subset(BBall, is.na(POSTSEASON) == FALSE)  
  
BBallTour$FINAL4 <- ifelse(BBallTour$POSTSEASON == 'F4' | BBallTour$POSTSEASON == '2ND' | BBallTour$POSTSEASON == 'Champions',1,0)  
  
#head(BBallTour)

To begin this analysis, we observe the Final Four teams of the 2019 March Madness Tournament. I created a new variable called FINAL4 as our predictor of success (1 if the team made it to the Final4, 0 if the team did not). Teams that did not make the tournament are not included in this first level of analysis.

Since we have chosen a binary variable as our response, we can then use a binary logistic model to estimate which predictors are the most important when considering tournament success (which is determined by whether or not the team made the Final 4). Below I created a new data frame without the the categorical variables since these cannot be included when determining the best binary logistic model. These are also the predictors that are least likely to have an effect on tournament success, which supports that they can be excluded from our tested data frame.

BBallTour.1 = within(BBallTour,   
 {TEAM = NULL  
 CONF = NULL  
 POSTSEASON = NULL})  
  
BBallTour.1 = as.data.frame(BBallTour.1)

To find which predictors out of the ones remaining in the data frame above are the most important when predicting tournament success, the Bayesian Information Criterion (BIC) is useful to look at. This type of model selection searches for a correct model given the current parameters.

bestglm(BBallTour.1, IC = 'BIC')

## BIC  
## BICq equivalent for q in (0.0756657587521878, 0.629738583428767)  
## Best Model:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -3.19304925 0.57317596 -5.570801 5.215887e-07  
## G 0.06179192 0.01186025 5.210000 2.086093e-06  
## X3P\_O 0.03148081 0.01024388 3.073132 3.093955e-03

As seen above, the number of games played and three-point shooting percentage are the predictors that led to the success of the Final 4 teams in 2019 (which is where we gathered data to being with). However, since BIC focuses on finding a current corrective model rather than a predictive model, we are more interested in the Akaike Information Criterion (AIC) without discounting these predictors. The AIC model selection method chooses the best model for predicting future observations.

bestglm(BBallTour.1, IC = 'AIC')

## AIC  
## BICq equivalent for q in (0.629738583428767, 0.756044637172223)  
## Best Model:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -1.88789146 1.153331633 -1.636903 1.069744e-01  
## W 0.03122600 0.008391876 3.720979 4.446761e-04  
## ADJOE 0.07088339 0.013001358 5.451999 1.027673e-06  
## ADJDE -0.04472235 0.014317151 -3.123691 2.767742e-03  
## BARTHAG -1.59493967 0.484908173 -3.289158 1.697631e-03  
## TOR 0.08174893 0.021911132 3.730931 4.307176e-04  
## ORB -0.03224281 0.008676098 -3.716280 4.514170e-04  
## X2P\_O -0.03806828 0.010881857 -3.498325 8.971132e-04  
## SEED 0.04327147 0.015132879 2.859434 5.858417e-03

According to the AIC results, we can observe that the Number of Games won, Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, Power rating, Turnover Percentage allowed, Offensive Rebound rate, Two-Point shooting percentage, and Seed in March Madness are the factors that predict success in the March Madness tournament.

Now we must make logical remarks about the above predictors in the context of game strategy. Seeds in March Madness are pre-set before the tournament, so other than the mental factor of that predictor, there is no evidence that the pre-set seed determines tournament success. There have only been 90 No. 1 seeds out of 160 that have made the Final 4 (or in our context only 56.3% of No. 1 seeds have found success). Therefore, we can take this predictor out of our final model as it does not determine game-time strategy.

The power rating is also something that is compiled from regular season success and is therefore not a true measure of tournament success. This is not a factor that can be changed to improve success in a certain game, it only represents a prediction of success against an average Division 1 basketball team. Therefore, this predictor can be taken out of our final model as it does not impact tournament success for a certain team, nor does it impact a team’s game-time strategy.

The same conclusion can be made for the number of games won during the regular season. This could offer a sense of comfort to the team and an opportunity for the team to refine their plays but other than that, wins in the regular season typically do not affect tournament success (nor do they affect game-time strategy). Therefore, we can take this predictor out of our final model.

Other than these predictors, the others would logically make sense in the realm of tournament game-time strategies, so we will keep those when determining our final model. Now we need to logically incorporate the top BIC predictors if they are significant to the current model. Similarly to the logical remarks about the number of games won during the regular season, the same could be said about the number of games played during the regular season. The number of games could help refine plays for the team, however when it comes to tournament success, the number of games played during the regular season does not have a large importance (as it does not affect game-time strategy).

However, three-point shooting percentage could have a viable impact on tournament success, so we will include it in our final predictive model shown below.

TourMod = glm(BBallTour.1$FINAL4~BBallTour.1$ADJOE+BBallTour.1$ADJDE+ BBallTour.1$TOR +BBallTour.1$ORB+BBallTour.1$'2P\_O'+BBallTour.1$'3P\_O')  
summary(TourMod)

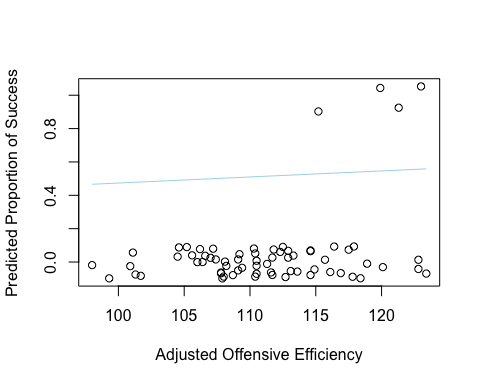
##   
## Call:  
## glm(formula = BBallTour.1$FINAL4 ~ BBallTour.1$ADJOE + BBallTour.1$ADJDE +   
## BBallTour.1$TOR + BBallTour.1$ORB + BBallTour.1$"2P\_O" +   
## BBallTour.1$"3P\_O")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.27813 -0.11219 -0.03755 0.03614 0.73651   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.333731 1.092923 -1.220 0.2270   
## BBallTour.1$ADJOE 0.015865 0.008639 1.836 0.0712 .  
## BBallTour.1$ADJDE -0.009299 0.006395 -1.454 0.1510   
## BBallTour.1$TOR 0.030793 0.023617 1.304 0.1972   
## BBallTour.1$ORB -0.009856 0.010126 -0.973 0.3342   
## BBallTour.1$"2P\_O" -0.011180 0.011484 -0.974 0.3341   
## BBallTour.1$"3P\_O" 0.024413 0.014383 1.697 0.0947 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.04710298)  
##   
## Null deviance: 3.7647 on 67 degrees of freedom  
## Residual deviance: 2.8733 on 61 degrees of freedom  
## AIC: -6.1799  
##   
## Number of Fisher Scoring iterations: 2

Based on the analysis shown above, we can determine the most influential factors that impact tournament success are Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, Turnover Percentage allowed, Offensive Rebound rate, Two-Point shooting percentage, and Three-Point shooting percentage.

Overall, to increase chances of tournament success, a team should have increased efficiency (opening more opportunities to take shots on the offensive side, and blocking opportunities to take shots on the defensive side).

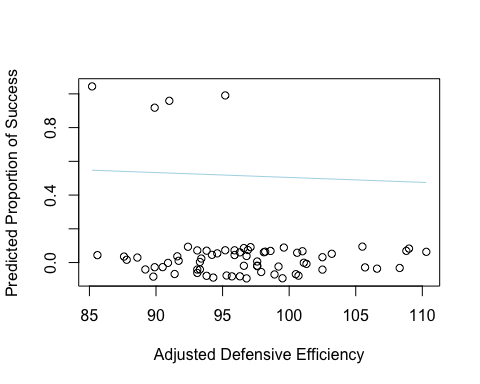
Both the adjusted offensive efficiency and adjusted defensive efficiency can impact the other predictors that determine tournament success, so observing these variables can help determine the overall success rate. For example, offensive rebound rate, two-point shooting percentage, and three-point shooting percentage can all be increased by increasing the offensive efficiency of a team. There is also a correlation between turnover percentage allowed and the adjusted defensive efficiency of a team.

logit = function(B0, B1, x)  
 {  
 exp(B0+B1\*x)/(1+exp(B0+B1\*x))  
}  
  
TourMod.1 = glm(BBallTour.1$FINAL4~BBallTour.1$ADJOE)  
plot(jitter(FINAL4, amount = 0.1)~ADJOE, data = BBallTour.1, xlab = 'Adjusted Offensive Efficiency', ylab = 'Predicted Proportion of Success')  
  
B0 = summary(TourMod.1)$coef[1]  
B1 = summary(TourMod.1)$coef[2]  
  
curve(logit(B0,B1,x), col = 'light blue', add = TRUE)



The above graph demonstrates the adjusted offensive efficiency versus success rate in the tournament (according to our 2019 tournament statistics). It can be viewed above that as the adjusted offensive efficiency of a team increases, so does the rate of success. Logically, this makes sense since the adjusted offesnive efficiency of a team is the amount of points a team would score per 100 possessions, or trips down the floor with the basketball (providing more opportuniteis to score and have offensive rebounds, which are all important factors when determining tournmanet success).

TourMod.2 = glm(BBallTour.1$FINAL4~BBallTour.1$ADJDE)  
plot(jitter(FINAL4, amount = 0.1)~ADJDE, data = BBallTour.1, xlab = 'Adjusted Defensive Efficiency', ylab = 'Predicted Proportion of Success')  
  
B0 = summary(TourMod.2)$coef[1]  
B1 = summary(TourMod.2)$coef[2]  
  
curve(logit(B0,B1,x), col = 'light blue', add = TRUE)



The above graph demonstrates the adjusted defensive efficiency versus success rate in the tournament (according to our 2019 tournament statistics). The adjusted defensive efficiency of a team is the amount of points a team would allow per 100 possessions, so as it is viewed above, the lower the adjusted defensive efficiency, the more successful the team seems to be in the tournament. The adjusted defensive efficiency controls turnover percentage allowed, which is another important variable when determining tournament success.

Now we have to compare these factors to regular season success. To begin this analysis of regular season success, we will use our original basketball data from 2019 (including all teams this time). I will also determine regular season success as winning 70% of regular season games.

In order to determine the predictors that are most responsible for regular season success, I created another binary response variable called REGSZN which will have a value of 1 for teams who won 70% or more of their regular season games and a value of 0 for teams who won less than 70% of their regular season games.

BBall$REGSZN <- ifelse(BBall$G\*.7 <= BBall$W,1,0)

We can use the same BIC and AIC methods as last time to determine the best predictors of regular season success (assuming this is another binary logistic model). This time, we have to refine our data frame and delete unnecessary categorical predictors as well as any postseason predictors before we can begin analysis, since we are only performing analysis on regular season success. We can also delete the variables G, W, and BARTHAG, because as we discussed previously, these are not variables that can be changed by a team’s game-time strategy, and therefore should not be included in predicting regular season success.

BBall.1 = within(BBall,   
 {TEAM = NULL  
 CONF = NULL  
 POSTSEASON = NULL  
 SEED = NULL  
 WAB = NULL  
 G = NULL  
 W = NULL  
 BARTHAG = NULL})  
  
BBall.1 = as.data.frame(BBall.1)

bestglm(BBall.1, IC = 'BIC')

## BIC  
## BICq equivalent for q in (0.397213100602503, 0.694747175809856)  
## Best Model:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -0.39031439 0.594361775 -0.656695 5.118136e-01  
## EFG\_O 0.04434235 0.006149416 7.210823 3.525772e-12  
## EFG\_D -0.03523738 0.006445789 -5.466728 8.792304e-08  
## TOR -0.03064828 0.008625581 -3.553184 4.334044e-04  
## TORD 0.03009179 0.008505049 3.538109 4.579909e-04  
## ORB 0.02102206 0.004455669 4.718048 3.461203e-06  
## DRB -0.01622373 0.006300534 -2.574977 1.043957e-02

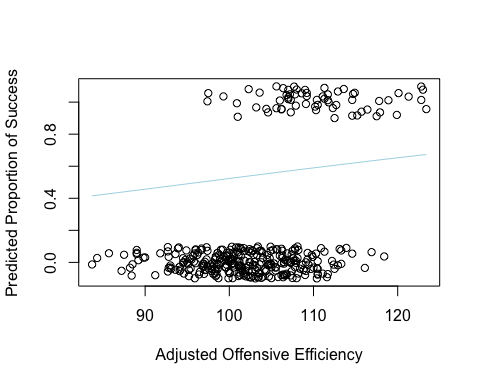
bestglm(BBall.1, IC = 'AIC')

## AIC  
## BICq equivalent for q in (0.804336028178494, 0.884339987242288)  
## Best Model:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -0.547622230 0.604036520 -0.9066045 3.652524e-01  
## ADJDE 0.013098974 0.005671110 2.3097726 2.149406e-02  
## EFG\_D -0.056681118 0.011239789 -5.0428989 7.443689e-07  
## TOR -0.028701391 0.009056827 -3.1690339 1.667060e-03  
## TORD 0.050600925 0.011065630 4.5728012 6.730039e-06  
## ORB 0.023552859 0.004489289 5.2464560 2.720928e-07  
## DRB -0.024921580 0.006981269 -3.5697779 4.082338e-04  
## FTRD -0.008287515 0.004214238 -1.9665512 5.004091e-02  
## X2P\_O 0.019932446 0.005666986 3.5172920 4.946277e-04  
## X3P\_O 0.034216170 0.007093097 4.8238686 2.121507e-06

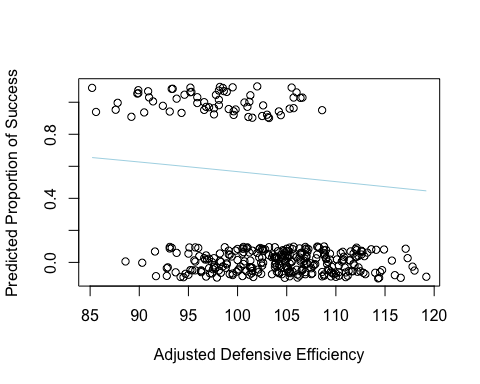
Again, we can assume the model produced through AIC methods is the most accurate for predicting. However, we should add EFG\_O since this would effectively combine the two model selection methods and offer us with the most accurate current and predictive model.

Therefore, the most important variables when predicting regular season success are Adjusted Offensive Efficiency, Adjusted Defensive Efficiency, Effective Field Goal Percentage Shot, Effective Field Goal Percentage Allowed, Turnover Percentage Committed, Turnover Percentage Allowed, Offensive Rebound Rate, Defensive Rebound Rate, Free Throw Rate, Two-Point shooting Percentage and Three-Point shooting Percentage.

BBallmod = glm(BBall.1$REGSZN~BBall.1$ADJOE)  
plot(jitter(REGSZN, amount = 0.1)~ADJOE, data = BBall.1, xlab = 'Adjusted Offensive Efficiency', ylab = 'Predicted Proportion of Success')  
  
B0 = summary(BBallmod)$coef[1]  
B1 = summary(BBallmod)$coef[2]  
  
curve(logit(B0,B1,x), col = 'light blue', add = TRUE)



BBallmod1 = glm(BBall.1$REGSZN~BBall.1$ADJDE)  
plot(jitter(REGSZN, amount = 0.1)~ADJDE, data = BBall.1, xlab = 'Adjusted Defensive Efficiency', ylab = 'Predicted Proportion of Success')  
  
B0.1 = summary(BBallmod1)$coef[1]  
B1.1 = summary(BBallmod1)$coef[2]  
  
curve(logit(B0.1,B1.1,x), col = 'light blue', add = TRUE)



Looking at the above data and graphs, we can assume multiple similarities between regular season and tournament strategies. Regular season success seems to be heavily reliant on adjusted offensive and defensive efficiency (similar to tournament success), focusing on rebound rates, field goal percentages, free throws, and two/three-point shooting percentages.

However, we do see differences between the two. Regular season strategies seem to be more reliant on effective field goal percentages and turnover percentages, indicating a more aggressive offensive strategy than a team in the tournament season might suggest. The rebound rate (both offensively and defensively) is highlighted in regular season play, also indicating more aggressive plays offensively and defensively. A more aggressive game-time strategy could hypothetically lead to increased fouls on both sides, which is why there is also more focus on the free throw rate than in tournament play.

A team in the tournament season might also rely on a greater defensive focus, especially when it comes to turnover percentage allowed. However, efficiency is still the main focus of a team in tournament play, and running the clock is not a typical option in tournament play. The goal of teams who seek success in such tournaments should try to get back and forth down the court as quickly as possible, raising efficiency and overall odds for success.