Predicting the Trajectory of College Basketball

Ryan Smith

01/15/2023

Introduction

Every year, the goal of a collegiate basketball is to win as many games as possible for as long as possible. Every coach, player, and fan has their own opinion on what is most important aspect of the game of basketball and leads to the most wins, but analytics and data are the way to truly explain the game of basketball and what can lead to more wins. This analytic report's goal is to look at the game of basketball across a decade through statistics and analytics to see how the game continues to change and what a team must do to stay ahead of these new trends.

Methodology

The data used in this report came from Andrew Sundberg's College Basketball Dataset on Kaggle along with his source, barttorvik.com. While Sundberg had seasons 2013-19 scraped and cleaned, seasons 2020-23 were had been scraped by "Bart Torvik" and was cleaned within R to be merged with Sundberg's dataset. These combined datasets included 2013-present day statistics on collegiate basketball teams as a whole. The variables used are listed below with descriptions:

*TEAM: The Division I college basketball school

*CONF: The Athletic Conference in which the school participates in (A10 = Atlantic 10, ACC = Atlantic Coast Conference, AE = America East, Amer = American, ASun = ASUN, B10 = Big Ten, B12 = Big 12, BE = Big East, BSky = Big Sky, BSth = Big South, BW = Big West, CAA = Colonial Athletic Association, CUSA = Conference USA, Horz = Horizon League, Ivy = Ivy League, MAAC = Metro Atlantic Athletic Conference, MAC = Mid-American Conference, MEAC = Mid-Eastern Athletic Conference, MVC = Missouri Valley Conference, MWC = Mountain West, NEC = Northeast Conference, OVC = Ohio Valley Conference, P12 = Pac-12, Pat = Patriot League, SB = Sun Belt, SC = Southern Conference, SEC = South Eastern Conference, Slnd = Southland Conference, Sum = Summit League, SWAC = Southwestern Athletic Conference, WAC = Western Athletic Conference, WCC = West Coast Conference)

*G: Number of games played

*W: Number of games won

*ADJOE: Adjusted Offensive Efficiency (An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average Division I defense)

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

*BARTHAG: Power Rating (Chance of beating an average Division I 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 (TWOP_O): Two-Point Shooting Percentage

*2P_D (TWOP_D): Two-Point Shooting Percentage Allowed

*3P_O (THREEP_O): Three-Point Shooting Percentage

*3P_D (THREEP_D): Three-Point Shooting Percentage Allowed

*ADJ_T: Adjusted Tempo (An estimate of the tempo (possessions per 40 minutes) a team would have against the team that wants to play at an average Division I tempo)

*WAB: Wins Above Bubble (The bubble refers to the cut off between making the NCAA March Madness Tournament and not making it)

*POSTSEASON: Round where the given team was eliminated or where their season ended (R68 = First Four, R64 = Round of 64, R32 = Round of 32, S16 = Sweet Sixteen, E8 = Elite Eight, F4 = Final Four, 2ND = Runner-up, Champion = Winner of the NCAA March Madness Tournament for that given year)

*SEED: Seed in the NCAA March Madness Tournament

*YEAR: Season

*W_PERCENT: Total Win Percentage

After cleaning, the variable W_PERCENT was created to represent the success of a team's season. A correlation matrix was created to find any relationships between the response variables, SEED, W_PERCENT, BARTHAG, and POSTSEASON. There were no strong relationships but W_PERCENT was chosen as the primary response variable because it applied to all teams in the dataset (as SEED and POSTSEASON were in relation to March Madness teams).

Modeling:

To best analyze how the game of basketball has changed within the past decade, linear models were made for each season before the COVID-19 season, 2020, and the coefficients were compared. Stepwise functions in both directions were performed for each season to create the best linear model to predict W_PERCENT based on ADJOE, ADJDE, EFG_O, EFG_D, TOR, TORD, ORB, DRB, FTR, FTRD, ADJ_T, TWOP_O, TWOP_D, THREEP_O, and THREEP_D.

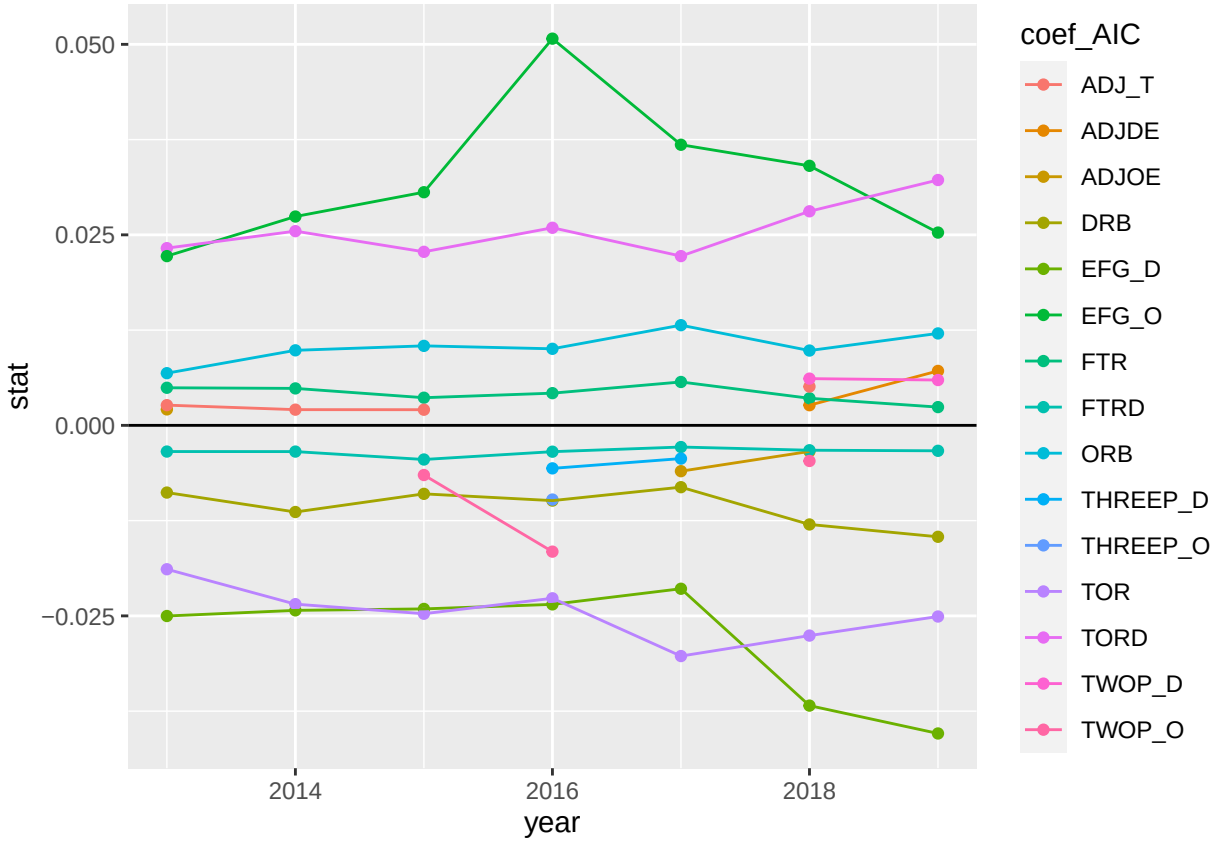
The coefficients for each season's model were separated by statistic to create a linear model predicting the coefficient for a new model that could be used in 2023 or any future year. These models of coefficients would be used to predict the impact (coefficient) of each statistic towards the total win percentage in years where there was either incomplete or nonexistent data.

Data

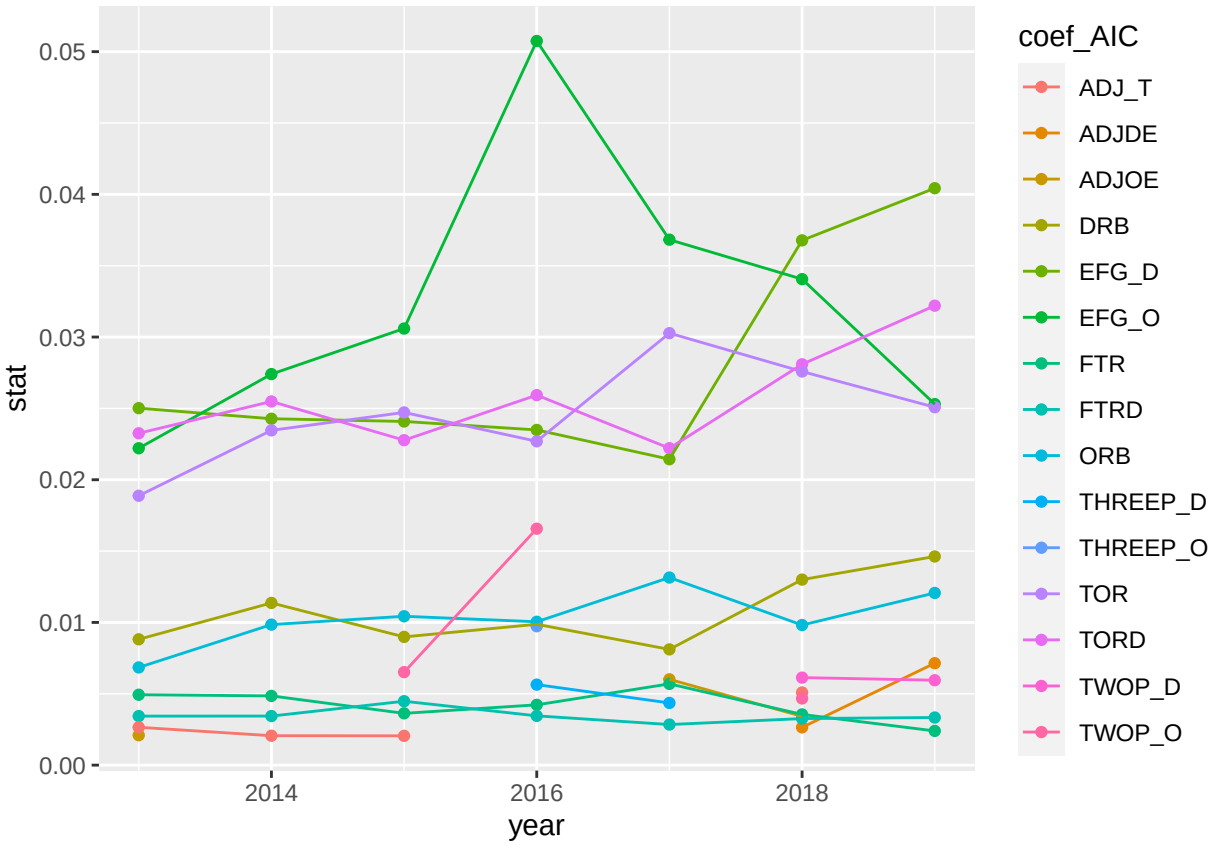
As shown below, each season's individual step model had an R-Squared value greater than 0.8 showing high levels of correlation across the models.

```
## 2013 R-Squared value: 0.8908704

## 2014 R-Squared value: 0.8843349

## 2015 R-Squared value: 0.8978132

## 2016 R-Squared value: 0.8933141

## 2017 R-Squared value: 0.8453668

## 2018 R-Squared value: 0.8586521

## 2019 R-Squared value: 0.8375774
```

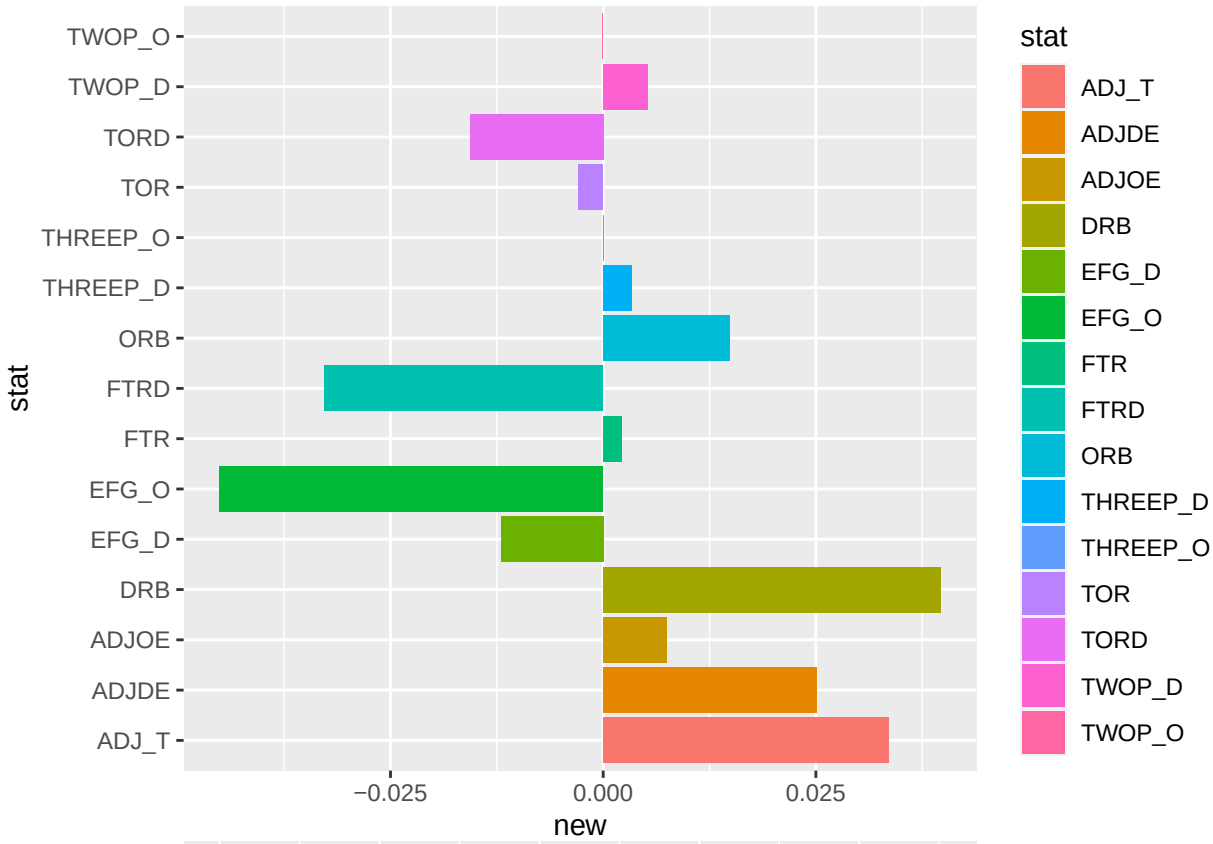
The below graphic shows the coefficients of each variable across the years 2013-2019. The two most positively significant variables are EFG_O and TORD, while the two most negatively significant variables are EFG_D and TOR.

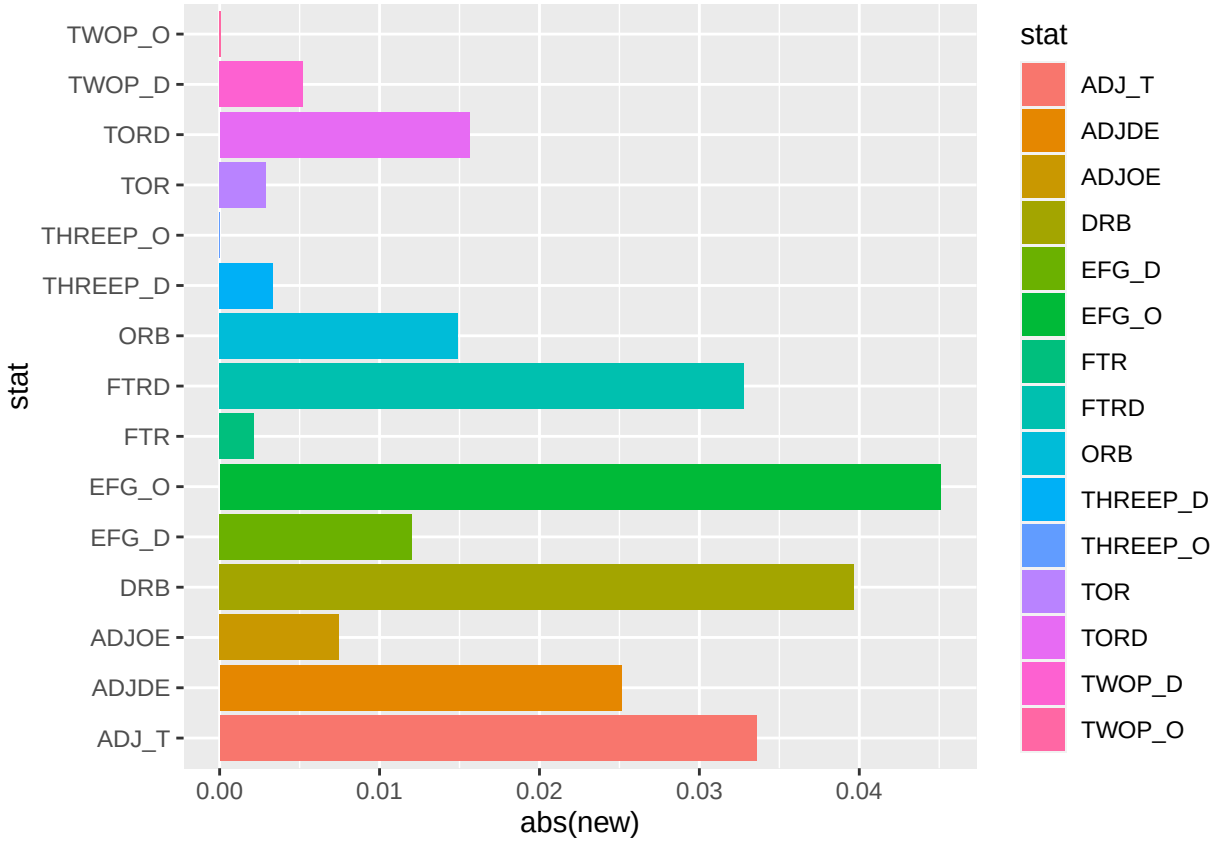


The graphic below shows the absolute value of all coefficients to show that while EFG_O was the most significant variable for a majority of the seasons, EFG_D and TORD have been showing increases since 2017 that have surpassed the significance of EFG_O. Following 2017, the data also show increases in DRB and decreases in TOR which may foreshadow changes in how college basketball is played.



Finally, the final models were used with 2023 as the input to predict the significance, through coefficients, of each variable in the linear model predicting W_PERCENT. The first graphic below shows each variable's predicted coefficient for 2023 and the second graphic shows the absolute value of each coefficient to best compare significance. EFG_O remains the most significant variable with DRB being second most significant, and FTRD and ADJ_T showing themselves to be significant as well.


stat
TWOP_O
TWOP_D
TORD
TOR
THREEP_O
THREEP_D
ORB
FTRD
FTR
EFG_O
EFG_D
DRB
ADJOE
ADJDE
ADJ_T
−0.025
0.000
0.025
new
stat
ADJ_T
ADJDE
ADJOE
DRB
EFG_D
EFG_O
FTR
FTRD
ORB
THREEP_D
THREEP_O
TOR
TORD
TWOP_D
TWOP_O


stat
TWOP_O
TWOP_D
TORD
TOR
THREEP_O
THREEP_D
ORB
FTRD
FTR
EFG_O
EFG_D
DRB
ADJOE
ADJDE
ADJ_T
0.00
0.01
0.02
0.03
0.04
abs(new)
stat
ADJ_T
ADJDE
ADJOE
DRB
EFG_D
EFG_O
FTR
FTRD
ORB
THREEP_D
THREEP_O
TOR
TORD
TWOP_D
TWOP_O

Conclusion

In the previous graphic, FTRD is likely to be more significant than ADJ_T because ADJ_T was removed from many models and was left with an NA value which did not weaken the significance of this model due to the methodology of this report. In the future, a different methodology for model creation may be used to better represent weak correlations. College basketball seems to maintain having EFG_O be a major priority for teams to yield victorious seasons, more points mean more wins, and large DRB and FTRD values show the importance of defense against free throws and rebounds. To further expand upon this study, more seasons of data may help to see larger trends in variables. More variables such as players, coaches, home vs away games, and more could explain more about seasons outside of on-the-court and overall team statistics.