Incorporating High School Recruitment Ratings and Statistics in Predictive Models for Collegiate Basketball Success

Ted Henson and Mario Giacomazzo

11/19/2019

Introduction

The goal of this project is to predict collegiate win shares from ESPN recruiting ratings [1] and high school statistics provided by Prep Circuit [2] and AAUStats [3]. As it stands, there is little research on predicting individual players' collegiate performance. In 2010, Jamie McNeilly used recruiting ranking quartiles to predict PER and other barometers of success [3]; however, the models presented did not consider high school statistics as an input, nor did they consider predicting win shares, which is a more authentic measurement of how a player contributes to overall team success as shown by basketball reference [5].

Accusations of many NCAA coaches paying high profile recruits hundreds of thousands of dollars to single recruits catalyized this analysis. If one player can cause decorated coaches to potentially resort to unethical methods, then programs should explore every possible avenue of predicting college performance, especially programs with smaller budgets and less recognition.

In addition to benefiting collegiate programs, the methods in this paper could benefit NBA front office decision making. Some high-profile recruits have had mediocre freshman collegiate performances (Harry Giles), or hardly any at all (Michael Porter Jr., Thon Maker), and are

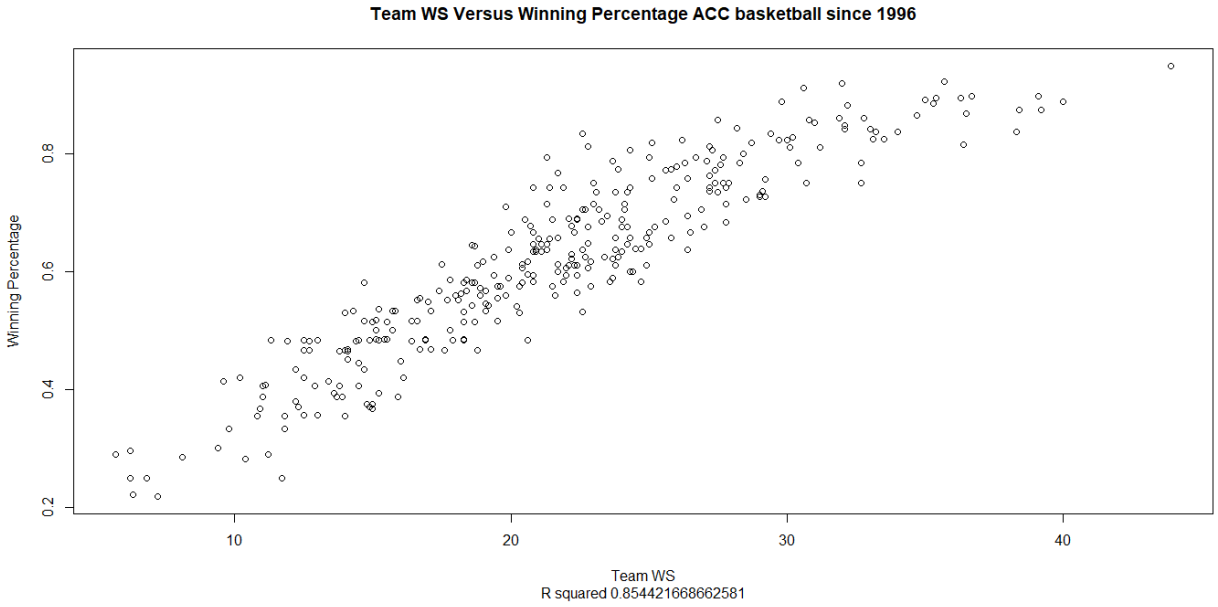
still selected in the first round based on their high school evaluations. As NBA teams are investing millions of dollars on players with little to no collegiate data, the methods and data presented in this paper could be used to model NBA performance in conjunction with their collegiate performance.

The model with the best out-of-sample error incorporating ESPN Ratings and the two sources of high school statistics. All other models had similar error rates. Future research adjusting for players' strength of schedule and teamates could significantly improve the models presented in this paper.

Data

Basketball Reference

Every statistic listed on a player's college basketball reference page was collected; however, only a player's first season playing in the NCAA was used in the modeling process in order to fairly evaluate a player's true production out of high school. Due to its all- encompassing nature, win shares represents the dependent variable. Below is a graph of ACC teams' sum of player win shares plotted against their season winning percentage.



College win shares have a weaker relationship with winning than WAR in baseball and in the NBA partially due to large differences in league competition; nonetheless, it is a strong predictor of team success as shown by the above R-squared and basketball reference's analysis [5]. Other popular basketball metrics such as PER and BPM were ploted against wins as well and had a much lower correlation value than win shares. Additionally, the top players in terms of win shares aligned with the concencus best players over the past few seasons more so than those with the top PER or BPM. Below are the top 10 players in terms of win shares in our data.

group	player.id	ws	Season
Prep and AAU	zion-williamson	8.3	2019
Prep and AAU	deandre-ayton	7.6	2018
Only Prep	marvin-bagleyiii	6.9	2018
Only Prep	lonzo-ball	6.8	2017
Prep and AAU	wendell-carterjr	5.9	2018
Only Prep	malik-monk	5.8	2017
Only Prep	tj-leaf	5.8	2017
Prep and AAU	trae-young	5.7	2018
Prep and AAU	tyler-herro	5.4	2019
Neither	omari-spellman	5.2	2018

From a basketball perspective, these players had some of the best freshman seasons over the past few years. Zion in particular has been widely regarded as having the best season from a statistical and basketball perspective. This gives more confidence and validitiy to win shares as an overall barometer of success.

ESPN

The ESPN data gathered contained players' overall rating from 55 to 100. Only the classes from 2016 to 2018 were used in this analysis due to the lack of Prep Circuit data before the 2016 high school season. In terms of grabbing the basketball reference data, the ESPN data played a critical role. There was no feasible or swift way to accurately gather a high school player's collegiate win shares without knowing where he went to college, which was not in the Prep Circuit or AAU data. Also gathered from ESPN were players' height, weight, and position.

Prep Circuit

The high school statistics gathered from Prep Circuit contained regular season averages and totals from box score statistics such as points, points per game, assists, etc. The data is fairly encompassing; however, there appear to be some inaccuracies in the data. For example, Lonzo Ball had 31 games where points were tracked, 4 games for minutes, 22 games for assists steals and turnovers, and 21 games for rebounds. One explanation is that Prep Circuit does not keep track of all statistics for every game. The other hypothesis was that if a player did not log a statistic in a given table, Prep Circuit did not count that towards your game total for that statistic. Upon futher inspection, it appeared that the most reliable statistics were the given per game statistics.

AAUStats

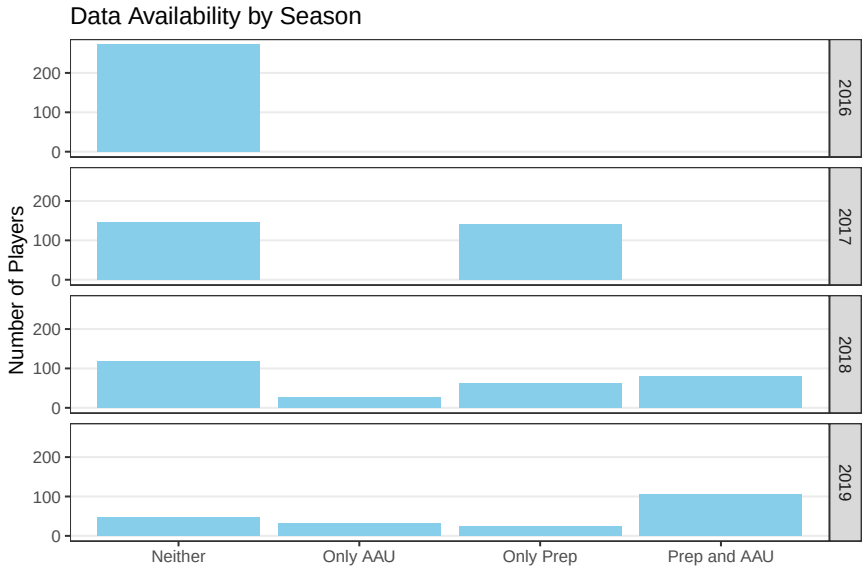
AAU data was gathered from aaustats.com [3]. This data contained box scores from Nike, Adidas, and Under Armour circuits from the 2017 to 2019 seasons. There is potential to create more advanced metrics with these box scores, or scaling based on the quality of a player's team or opponent; however, for this analysis, common per game box score statistics akin to Prep Circuit's data were created based on player's final season in all AAU circuits.

Prior analyses used k nearest neighbor imputation to deal with players missing a given statistic, such as rebounds in Prep Circuit. Below is a table of the number of Prep Circuit players that had points per game, but were missing another per game statistics.

Per Game Statistic	Number of Players with NAs
mpg.prep	126
reb.prep	147
blk.prep	58
spg.prep	58
tov.prep	58

Although this knn imputation was effective for seasons past, it will not be the best method going forward, as the amount and quality of data has improved drastically over the past few years. Below is a table and histogram of the number of players that have each of our data sources by season.

Season	Neither	Only AAU	Only Prep	Prep and AAU
2016	272	0	0	0
2017	147	0	141	0
2018	117	27	61	79
2019	46	32	25	106



As shown by the graph above, in 2016, the data sources presented did not even exist. By 2019, most ESPN rated players had both AAU and Prep Circuit statistics. This includes players that are rated poorly. Additionally, by the time this project will complete, the 2019-2020 college basketball season will be over. Their data could be incorporated into the training set or held out as a test set.

Models

In order to assess the predictive value of high school statistics, several different models will be constructed using different sources of information. As shown by the graph above, not every model can be considered on every player as not all players have all sources of information. Therefore, for a given model, only players with complete information will be considered. The models are the following:

$$M_{AAU} := \hat{ws} = f_{AAU}(X_{AAU}) + \epsilon$$

$$M_{PREP} := \hat{ws} = f_{PREP}(X_{PREP}) + \epsilon$$

$$M_{ESPN} := \hat{ws} = f_{ESPN}(X_{ESPN}) + \epsilon$$

$$M_{AAU.ESPN} := \hat{ws} = f_{ESPN}(X_{AAU}, X_{ESPN}) + \epsilon$$

$$M_{PREP.ESPN} := \hat{ws} = f_{ESPN}(X_{PREP}, X_{ESPN}) + \epsilon$$

$$M_{AAU.PREP} := \hat{ws} = f_{ESPN}(X_{AAU}, X_{PREP}) + \epsilon$$

$$M_{FULL} := \hat{ws} = f_{FULL}(X_{AAU}, X_{PREP}, X_{ESPN}) + \epsilon$$

where

- $\hat{ws}$ *is the predicted win shares of a given player*,
- $X_{AAU},\ X_{PREP},$ *and* X_{ESPN} *are the data matrices for each data source*,
- $f_{AAU}, f_{Prep}, f_{ESPN}, and f_{FULL} are functions that output predicted win shares base each data source$,
- ϵ *is a random error*

Methods

All models were trained through leave one out cross validation. Each player's predicted win shares was created through fitting on all other players, and then predicted on that player. In order to speed up these computations, hyperparameters for some of the different methods were trained through 5 fold cross validation for all players. These values were chosen based on the parameters yielding the smallest out of sample error. So for example, in lasso regression, all models used the same optimal lambda value during the fitting process.

In addition to predicted raw win shares, a seperate set of predictions were also generated by predicting a players offensive and defensive win shares seperately. The final predictions were a weighted sum of these two quantities. The linear weights were calculated through leave one out cross validation. A weighted sum was used as this significantly decreased out of sample accuracy compared to adding the two together.

Results

Leave One Out Win Share RMSEs

Table 4: Table continues below

X1	espn n=185	prep n=185	aau n=185	full n=185	aau.espn n=185
pls	NA	1.37	1.41	1.33	1.35
lasso	1.36	1.36	1.38	1.32	1.35
ridge	1.36	1.36	1.38	1.3	1.33
rf	NA	1.42	1.42	1.35	1.36
earth	1.34	1.4	1.43	1.3	1.32
svm.radial	1.28	1.43	1.4	1.37	1.34
xgbDART	1.56	1.69	1.66	1.55	1.59
stacked	1.33	1.41	1.41	1.25	1.3

prep.espn n=185	aau.prep n=185
1.33	1.36
1.32	1.33
1.32	1.32
1.34	1.37
1.3	1.53
1.32	1.39
1.66	1.54
1.3	1.31

Leave One Out Offensive + Defensive Win Share RMSEs

Table 6: Table continues below

X1	espn n=185	prep n=185	aau n=185	full n=185	aau.espn n=185
pls	NA	1.36	1.41	1.33	1.35
lasso	1.36	1.36	1.38	1.32	1.34
ridge	NA	1.36	1.38	1.3	1.33
rf	NA	1.42	1.41	1.34	1.35
earth	1.31	1.39	1.43	1.31	1.32
svm.radial	1.29	1.42	1.41	1.35	1.35
xgbDART	1.56	1.64	1.57	1.49	1.49
stacked	1.31	1.4	1.42	1.26	1.31

prep.espn n=185	aau.prep n=185
1.33	1.37
1.32	1.33
1.32	1.32
1.33	1.39
1.3	1.45
1.34	1.38
1.63	1.57
1.3	1.33

Important Variables

Conclusion

Appendix

References

[1] "ESPN Basketball Recruiting - Player Rankings", ESPN. [Online]. Available: http://www.espn.com/collegesports/basketball/recruiting/playerrankings/_/class/2016/order/true. [Accessed: 26- Mar- 2019].

[2] "Statistics - 2015-2016 Regular Season - HS Circuit", Prep Circuit. [Online]. Available: https://www.prepcircuit.com/stats/league_instance/34558?subseason=245525. [Accessed: 01- Apr- 2019].

[3]"AAUStats", Aaustats.com, 2019. [Online]. Available: http://aaustats.com/. [Accessed: 13- Dec- 2019].

[4] J. McNeilly, "Prediction Versus Production: Examining the Relationship Between NCAA Division I Ranked Recruits and their Ensuing Athletic Production in College", Epublications.marquette.edu, 2010. [Online]. Available: https://epublications.marquette.edu/cgi/viewcontent.cgi?article=1013&context=cps_profe ssional. [Accessed: 04- Apr- 2019].

[5] “Calculating Win Shares”, Sports Reference. [Online]. Available: https://www.sportsreference.com/cbb/about/ws.html. [Accessed: 04- Apr- 2019].