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

Ted Henson and Mario Giacomazzo

11/19/2019

First Take

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 catalyzed 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.

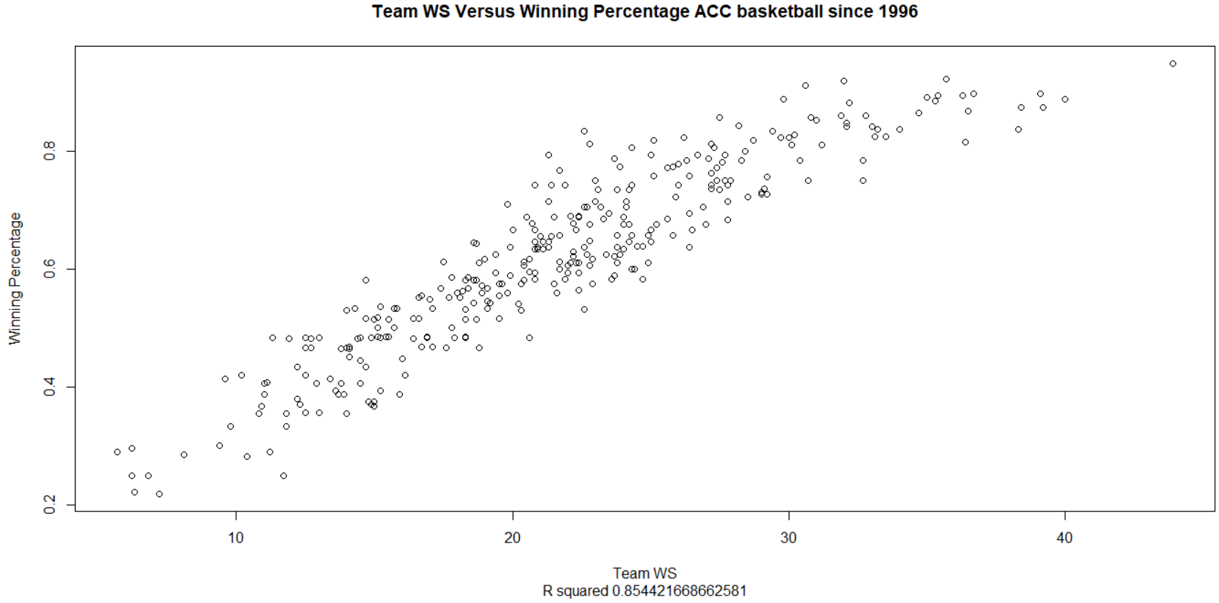
In addition to benefiting collegiate programs, the methods in this paper could benefit NBA front office decision making. Some high-profile recruits have had mediocre freshman collegiate performances (Harry Giles), or hardly any at all (Michael Porter Jr., Thon Maker), and are still selected in the first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

The Players

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 plotted 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 validity 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 further inspection, it appeared that the most reliable statistics were the given per game statistics.

AAUStats

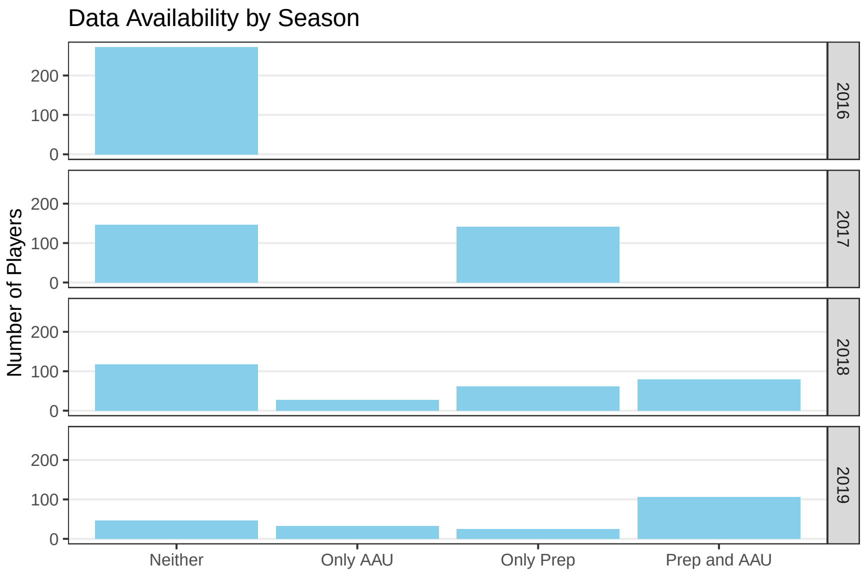
AAU data was gathered from AAUStats.com [3]. This data contained box scores from Nike, Adidas, and Under Armour circuits from the 2017 to 2019 seasons. There is potential to create more advanced metrics with these box scores, or scaling based on the quality of a player's team or opponent; however, for 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. As shown by the graph above, not every model can be considered on every player as not all players have all sources of information. Using all possible players unfairly favored ESPN models since their sample of players contains many players not talented enough to play in the competitive AAU Circuits or high profile high schools in Prep Circuit. Therefore, only player's with complete information will be considered. Additionally, since many player's only had partial Prep Circuit data, only points per game and the number of games was used in fitting. Other variables were used in some models, but the elimination of some players lacking say blocks per game increased the error rates. In total, there were 185 players used in the models.

Game Plan

In order to assess the predictive value of high school statistics, models will be constructed using all possible input combinations. 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

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

All models were trained through leave one out cross validation. Each player's predicted win shares was outputted from a function fitted on all other players. In order to speed up these computations, hyper parameters for some of the different methods were trained through 10 fold cross validation. These values were chosen based on the parameters yielding the smallest out of sample error. The methods and subsequent hyper parameters will be discussed briefly below

- Lasso
  - Fits a standard linear regression, but shrinks the sum of the absolute value of the coefficients by a value lambda chosen through cross validation.
- Ridge
  - Fits a standard linear regression, but shrinks the sum of squares of the coefficients by a value lambda chosen through cross validation.
- Averaged Neural Network
  - Assigns one or more linear weights to each of the variables depending on the number of input layers (number of variables). The hidden layers are comprised of a linear combination of the input variables. The number of hidden layers is somewhat arbitrarily chosen, in this case, 1/2 of the number of input layers. The final output is a linear combination of the hidden layers, and is then converted into a predicted value through some function, in this case a linear weight was chosen. 200 networks were constructed as such. The final prediction for a player was an average of all such networks.
- Earth
  - Models non lineareaties and interactions by creating one or more hinge functions constructed as h(x-a), where a is the cutoff value for variable x. When x is below a, it is multiplied by a constant b1. When x is greater than a, it is multiplied by a separate constant b2, and is added to b1. The number of these hinge functions is controlled by a hyper parameter as is the degree (number of interactions and higher order terms).
- Support Vector Machine

  - Builds a hyper plane (a plane or line with dimension of our predictor matrix) that attempts to minimize the distance from the response value. Rather than minimize the total sum of the epsilon's, the hyper plane is fit so that all epsilon's are less than a given cost value, C, found through cross validation. In this case, the non linear predictor space had a lower out of sample error.

- Random Forest
  - Builds n regression trees and creates a prediction based on the average of all of the trees. To reduce correlation between the trees, m random predictors are chosen to construct the tree. M is found through cross validation.
- Gradient Boosted Trees
  - Builds n "weak" regression trees and creates a prediction based on the average of all of the "weak" trees. The trees are pruned by cutting them at a specified "max depth" found through cross validation.
- Stacked
  - A full linear regression is performed on all prior predictions using leave one out cross validation.

Box Score

There were many variables to consider for these models. Many statistics in the prep data set were not always valid values, and many in both high school data sets or were randomly inaccurate, leading to poor out of sample prediction. Methods were trained through many different combinations of the inputs. The best models across the board only used the most reliable and predictive statistics. A full linear regression on all variables used in the models is shown below. ## Full Linear Regression

```
## 
## Call:
## lm(formula = ws ~ ., data = full.2)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.8728 -0.7624 -0.1667  0.7021  5.1933 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)   
## (Intercept)      -2.74824    1.31798  -2.085  0.03854 * 
## espn.rating       0.03796    0.01812   2.095  0.03766 * 
## Points.prep       0.04857    0.01745   2.783  0.00599 **
## GamesPlayed.prep  0.03659    0.01448   2.527  0.01240 * 
## Minutes.aau      -0.02972    0.01969  -1.509  0.13307   
## Rebounds.aau      0.09856    0.07118   1.385  0.16797   
## Blocks.aau        0.33595    0.24591   1.366  0.17368   
## Steals.aau        0.44626    0.23673   1.885  0.06111 . 
## Points.aau        0.01097    0.01396   0.785  0.43336   
## GamesPlayed.aau  -0.02708    0.02597  -1.043  0.29861   
## Position.BasicPF -0.55909    0.35825  -1.561  0.12047   
## Position.BasicPG -0.33086    0.52347  -0.632  0.52819   
## Position.BasicSF -0.78491    0.40205  -1.952  0.05254 . 
## Position.BasicSG -0.37776    0.45602  -0.828  0.40861   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.263 on 171 degrees of freedom
## Multiple R-squared:  0.3879, Adjusted R-squared:  0.3414
## F-statistic: 8.337 on 13 and 171 DF,  p-value: 6.445e-13
```

The most interesting result is that the AAU points per game was not highly significant, but blocks and steals were. This may seem surprising, but many AAU teams are super teams which may lead to individual star players sharing the ball more. In contrast, these same star players may be the only capable scorer on their school teams, leading to high and predictive Prep Circuit points per game. Although the full linear regression performed close to the other methods, it was the worst method across all input combinations. Therefore it will not be discussed in the final results. Below are the cross validated R^2s for the different input combinations and models.

$Leave\ One\ Out\ Win\ Shares\ R^2$

Method	espn n=185	prep n=185	aau n=185
nnet	NA%	17.2%	12.8%
lasso	NA%	23.3%	21.2%
ridge	NA%	23%	21.4%
rf	NA%	16.7%	16.5%
earth	28.9%	20.8%	15%
svm.radial	33.1%	18.1%	21.6%
xgboost	NA%	6.9%	6.8%
stacked	32%	16.6%	17.3%

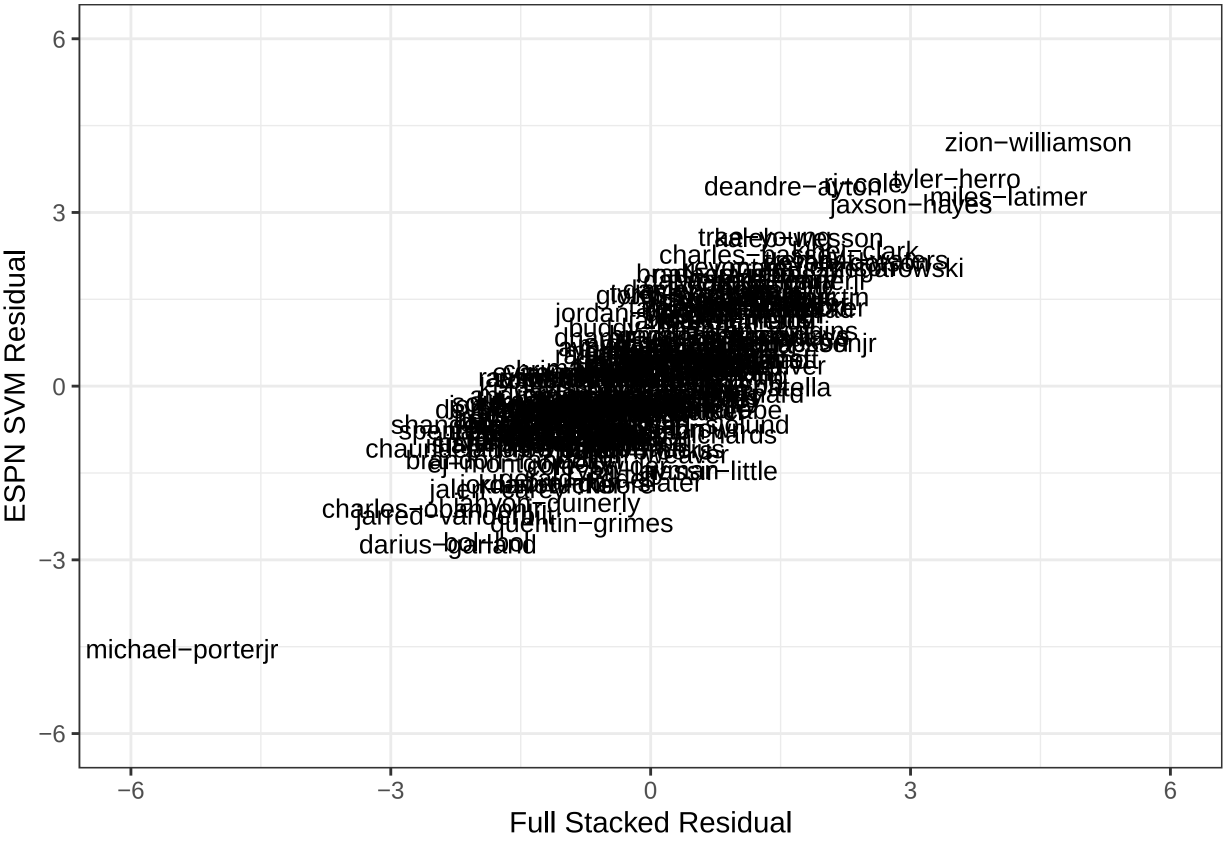
full n=185	aau.espn n=185	prep.espn n=185	aau.prep n=185
23.4%	18.7%	23.1%	19.2%
28.1%	25.3%	27.4%	26.9%

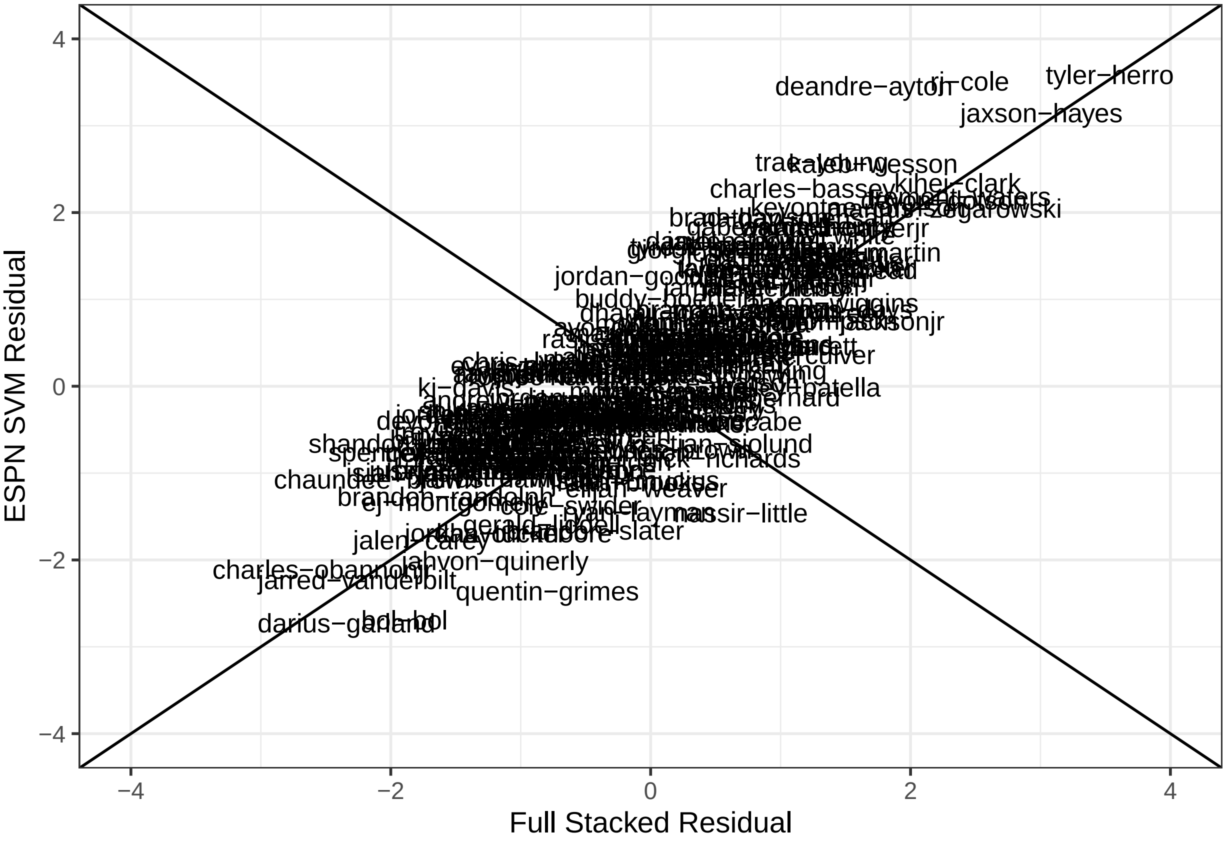
full n=185	aau.espn n=185	prep.espn n=185	aau.prep n=185
29.5%	26.2%	27.3%	27.7%
24.4%	23.2%	25.6%	21.7%
30.1%	28.5%	30.1%	14.4%
24.9%	27.7%	28.7%	23%
14.8%	14.7%	8.8%	13.4%
36.6%	32.4%	29.7%	21.3%

Linear Weights for Full Stacked Model

```
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5218 -0.8201 -0.0173  0.7226  4.0877
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.2433     0.4691  -2.650  0.00877 **
## nnet          0.9665     0.4178   2.314  0.02184 *
## lasso        -1.4766     0.8689  -1.699  0.09101 .
## ridge         2.5444     0.9437   2.696  0.00769 **
## rf           -1.2868     0.4416  -2.914  0.00403 **
## earth         1.2093     0.1828   6.616 4.22e-10 ***
## svm.radial    0.0953     0.2643   0.361  0.71887
## xgboost      -0.2606     0.1646  -1.584  0.11507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.177 on 177 degrees of freedom
## Multiple R-squared:  0.4492, Adjusted R-squared:  0.4274
## F-statistic: 20.62 on 7 and 177 DF,  p-value: < 2.2e-16
```

Best ESPN Model Versus Best Stacked Model





Last Take

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].