Predicting Collegiate Basketball Win Shares from ESPN Ratings, High School Statistics, and AAU Data

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

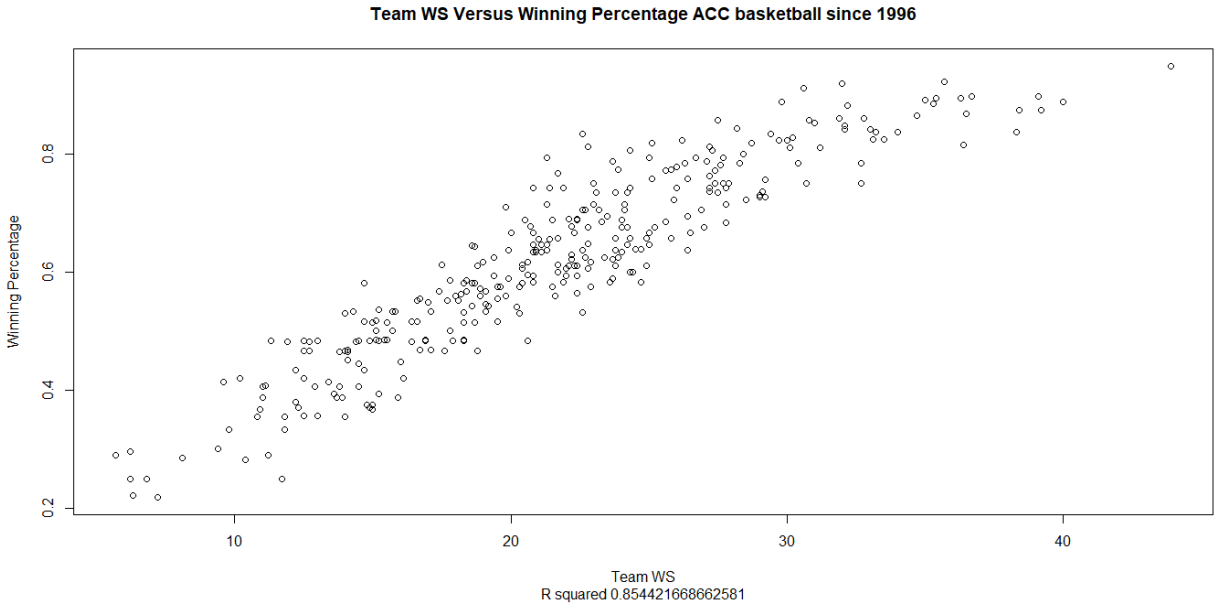
presented in this paper could be used to model NBA performance in conjunction with their collegiate performance.

The model with the best out-of-sample mean absolute error included both ESPN ratings and Prep Circuit statistics. The models with only ESPN rating and only Prep Circuit statistics followed suit. 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]. To further justify the use of win shares as the response, below are the top 10 players in terms of win shares in our data.

group	Name	ws	Season
Prep and AAU	Zion Williamson	7.9	2019
Only Prep	Lonzo Ball	6.8	2017
Only Prep	Malik Monk	5.8	2017
Only Prep	T.J. Leaf	5.8	2017
Prep and AAU	Trae Young	5.7	2018
Prep and AAU	Tyler Herro	5.3	2019
Neither	Omari Spellman	5.2	2018
Only Prep	De'Aaron Fox	5.1	2017
Prep and AAU	Gary Trent Jr.	5.1	2018
Neither	Ivan Rabb	5.1	2016

From a basketball perspective, these players arguably had some of the best seasons over the past few years, and 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 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.

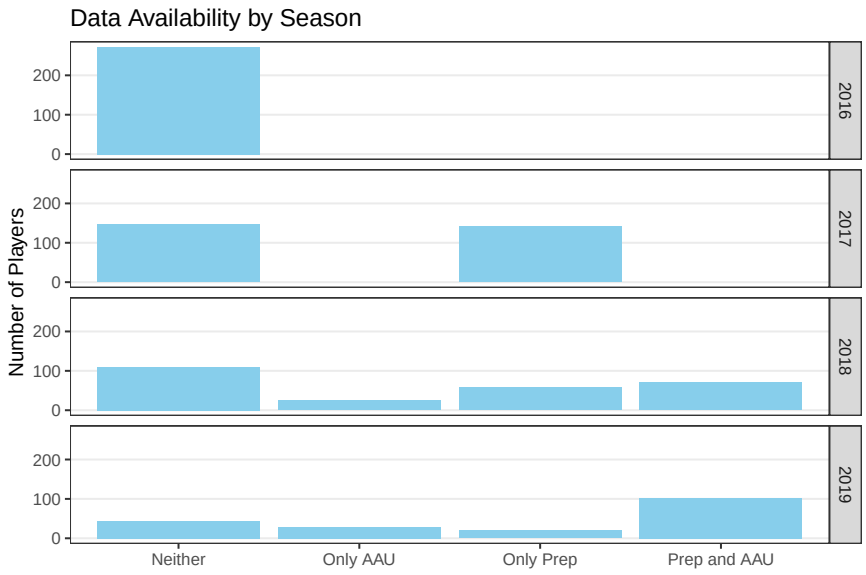
AAUStats

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. Quitting from lines 65-67 (Predicting-Collegiate-Basketball-Win-Shares-from-ESPN-Ratings-and-High-School-Statistics.rmd) Error in pander(group.freq) : object 'group.freq' not found Calls: ... inline_exec -> hook_eval -> withVisible -> eval -> eval -> pander In addition: Warning message: Missing column names filled in: 'X1' [1]



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:

$$M_{AAU} := \hat{ws} = f_{AAU}(X_{AAU}) * gp_{bball-ref} + \epsilon$$

$$M_{PREP} := \hat{ws} = f_{PREP}(X_{PREP}) * gp_{bball-ref} + \epsilon$$

$$M_{ESPN} := \hat{ws} = f_{ESPN}(X_{ESPN}) * gp_{bball-ref} + \epsilon$$

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

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*,
- $f_{AAU}, f_{Prep}, f_{ESPN}, and f_{FULL} are functions that output predicted win shares per game using one or all thre$
- $gp_{bball-ref}$ *is the number of games played in a player's first season*,
- ϵ *is a random error*

Methodology

Basic Linear Regression

Many methods will be considered for each model, but to start, linear models will be considered as they are the most interpretable. First and foremost, deciphering which variables are of

interest, and where the variation in the data lies will lead to better predictive modeling. To evaluate the predictive value of the data and each variable, a full linear regression was trained using leave one out cross validation using only either complete AAU data or Prep Circuit statistics. Win shares per game was the response variable. The predictions by the number of games played by the player. This allows for players who suffered injuries to be used in modeling, but not decrease the accuracy of the results. Below are the summary statistics for the linear M_{AAU}.

AAU Leave One Out CV Linear Regression (Predicting Win Shares Per Game)

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.068880 -0.022912 -0.005513  0.018711  0.188988
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.0793850  0.1114272   0.712 0.476962
## Points.aau       0.0014772  0.0003357   4.401  1.7e-05 ***
## Minutes.aau     -0.0016791  0.0004941  -3.398 0.000808 ***
## Rebounds.aau     0.0050225  0.0017983   2.793 0.005693 **
## Blocks.aau       0.0143403  0.0060914   2.354 0.019459 *
```

```
## Steals.aau       0.0172521  0.0057245   3.014 0.002889 **
## Turnovers.aau    0.0039808  0.0040101   0.993 0.321976
## GamesPlayed.aau  0.0001738  0.0006321   0.275 0.783554
## Position         0.0003367  0.0026352   0.128 0.898440
## Position.BasicF -0.0221148  0.0100913  -2.191 0.029487 *
## Position.BasicG -0.0096806  0.0151035  -0.641 0.522235
## Height          -0.0009043  0.0013726  -0.659 0.510715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03448 on 216 degrees of freedom
## Multiple R-squared:  0.3517, Adjusted R-squared:  0.3187
## F-statistic: 10.65 on 11 and 216 DF,  p-value: 1.384e-15
```

Minutes, rebounds, points, and steals were significant at the .01 level, and blocks and whether or not a player was a forward were significant at the .05 level. The significance of the minutes variable was somewhat suprising. Minutes may be indicative that a player who plays a lot of minutes has an advantage logging good per game statistics. Below are the confidence intervals for the coefficients.

```
##                        2.5 %       97.5 %
## (Intercept)     -0.140238863  0.299008819
## Points.aau       0.000815602  0.002138840
## Minutes.aau     -0.002653026 -0.000705123
## Rebounds.aau     0.001477973  0.008567041
## Blocks.aau       0.002334215  0.026346465
```

```
## Steals.aau       0.005969037  0.028535154
## Turnovers.aau   -0.003923180  0.011884683
## GamesPlayed.aau -0.001071943  0.001419612
## Position        -0.004857339  0.005530813
## Position.BasicF -0.042004844 -0.002224787
## Position.BasicG -0.039449613  0.020088501
## Height          -0.003609805  0.001801165
```

As shown above, the variables that had highly significant p values all had either strictly positive or negative confidence intervals. Notably, assists are not tracked at all in the AAU stats. They may have elected to not track it due to difficulties with accuracy. Prep Circuit, however, does track assists. This may not be reliable based on AAU's decision to not track it, and the appearance of some star players with zero assists, such as Zion Williamson. It was not included in the regression as it appeared to be inaccurate. The results from M_{PREP} differed significantly. Below are the same summary statistics for M_{PREP}.

Prep Circuit Leave One Out CV Linear Regression (Predicting Win Shares Per Game)

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.094141 -0.024897 -0.004828  0.022879  0.121413
```

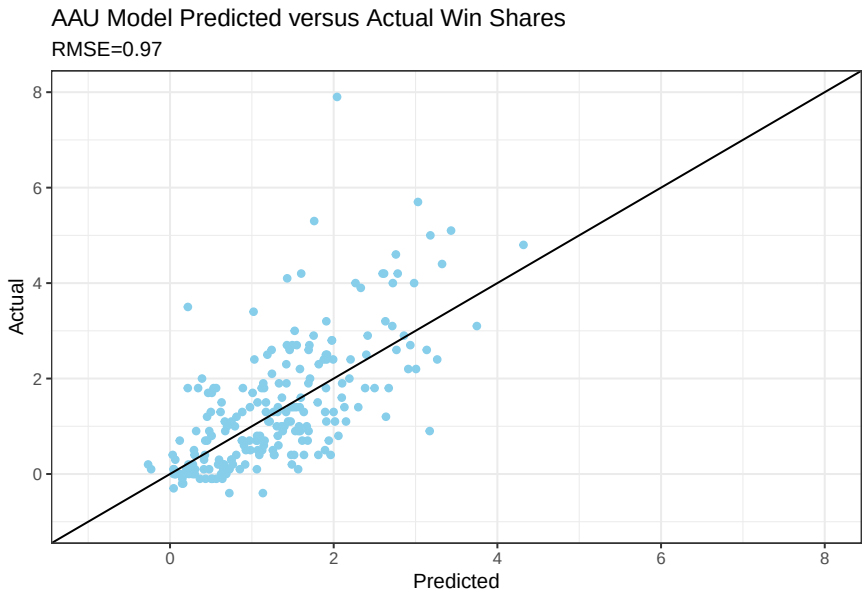
```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.455e-01  1.300e-01  -1.119  0.26454
## Points.prep       2.964e-03  5.430e-04   5.458  1.6e-07 ***
## Minutes.prep      3.569e-05  3.200e-04   0.112  0.91132
## Rebounds.prep    -1.961e-03  1.526e-03  -1.285  0.20055
## Blocks.prep       8.537e-03  2.888e-03   2.955  0.00355 **
## Steals.prep       5.714e-03  3.614e-03   1.581  0.11560
## Turnovers.prep   -1.458e-03  3.372e-03  -0.432  0.66608
## GamesPlayed.prep  1.183e-03  4.001e-04   2.956  0.00354 **
## Position         -4.218e-03  3.406e-03  -1.238  0.21719
## Position.BasicF   5.069e-03  1.298e-02   0.391  0.69662
## Position.BasicG   8.659e-03  1.861e-02   0.465  0.64228
## Height            1.806e-03  1.616e-03   1.118  0.26525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03922 on 177 degrees of freedom
## Multiple R-squared:  0.3255, Adjusted R-squared:  0.2836
## F-statistic: 7.766 on 11 and 177 DF,  p-value: 6.581e-11
```

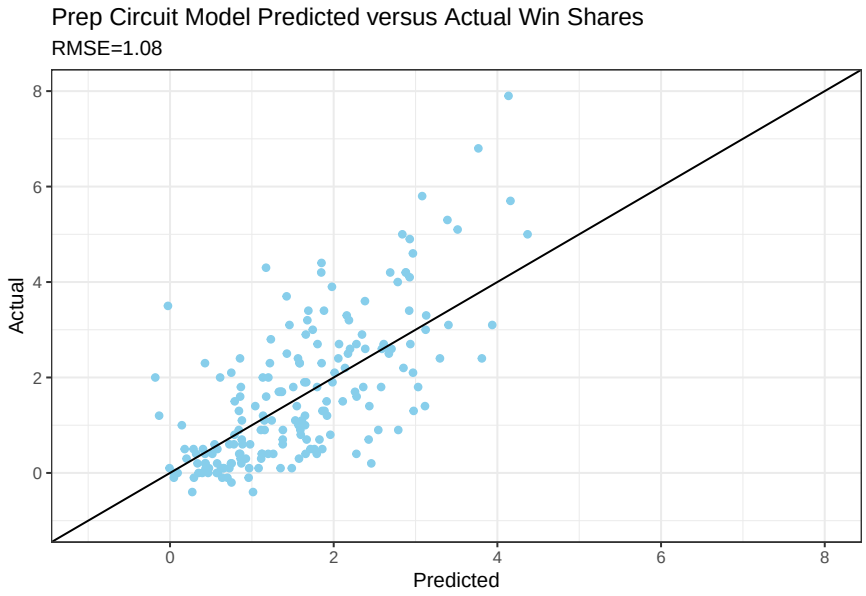
For M_{PREP}, points was by far the most significant, as was the number of games played. The only other significant variable was the number of blocks per game, which was significant in both M_{AAU} and M_{PREP}. Minutes were not significant at all in M_{PREP}, whereas the opposite was true in M_{AAU}. Steals and rebounds were also not significant in this case (although steals

were almost significant at the .05 level), whereas they were significant at the .01 level in the M_{AAU}. As shown below, the signifcant variables had confidence intervals that did not contain zero.

```
##                           2.5 %       97.5 %
## (Intercept)      -0.4020734491 0.1110485934
## Points.prep       0.0018920945 0.0040350615
## Minutes.prep     -0.0005959006 0.0006672902
## Rebounds.prep    -0.0049724866 0.0010510376
## Blocks.prep       0.0028364973 0.0142366493
## Steals.prep      -0.0014172176 0.0128454075
## Turnovers.prep   -0.0081123366 0.0051971009
## GamesPlayed.prep  0.0003933070 0.0019725908
## Position         -0.0109403474 0.0025035984
## Position.BasicF  -0.0205466739 0.0306845413
## Position.BasicG  -0.0280634640 0.0453805059
## Height           -0.0013827271 0.0049939827
```

These results support some hypotheses generated during web scraping and exploratory analysis: the AAU data is more accurate and has more significant variables across the board, whereas the Prep Circuit data is highly erratic, but still achieves a similar RMSE. Below are the plots of predicted versus fitted values for both models.





Principal Component Analysis

Principal component analysis also revealed some interesting insights into the two data sources as well. Below are the principal components after removing the categorial position and season variables and scaling the data.

AAU PC Scores

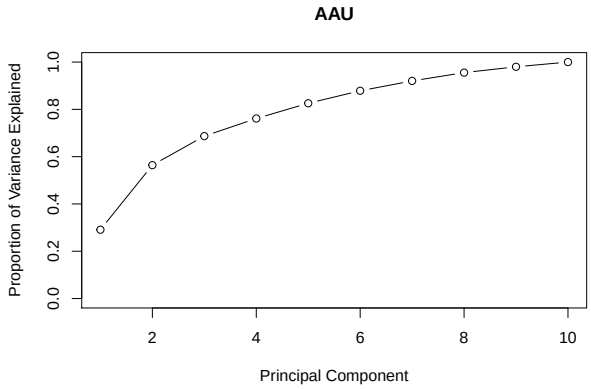
```
##                          PC1         PC2
## ws.per.game     -0.307449192 -0.20243242
## Points.aau      -0.448983879 -0.02529063
## Minutes.aau     -0.434323744 -0.08393322
## Rebounds.aau    -0.235701025 -0.48347356
## Blocks.aau       0.001669132 -0.52542983
## Steals.aau      -0.437586237  0.15552116
## Turnovers.aau   -0.451938651 -0.01600279
## GamesPlayed.aau -0.044082256  0.10480323
## Position        -0.177342258  0.40034191
## Height           0.174842262 -0.49579557
```

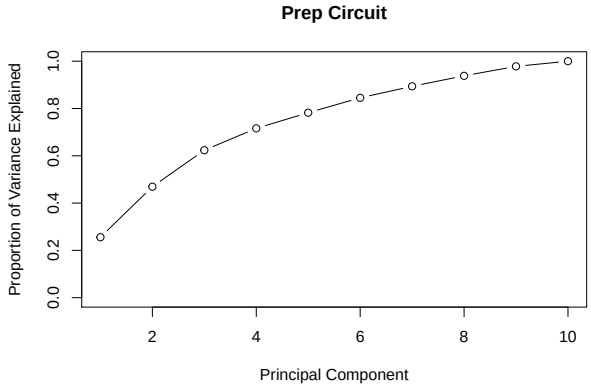
Prep Circuit PC Scores

```
##                           PC1         PC2
## ws.per.game      -0.108683945  0.48660793
## Points.prep       0.075270968  0.49519166
## Minutes.prep      0.007975080 -0.29557459
## Rebounds.prep    -0.450421864  0.25464890
## Blocks.prep      -0.498382775  0.12261304
## Steals.prep       0.203352646  0.27126518
## Turnovers.prep    0.008654123 -0.27372900
## GamesPlayed.prep  0.094572467  0.42559771
## Position          0.465240983  0.09608104
```

```
## Height            -0.514172070 -0.10897817
```

For the first principle component, the weight on win shares per game, points, and blocks, are nearly identical between the Prep Circuit and AAU model. Although points was not highly significant for the AAU model, the first principal component shows that the majority of the variation in the AAU data is in the same direction as the Prep Circuit data. Scree plots also showed that the overall variation in the data can be captured by a similar number of principal components. Below are the cumulative scree plots for the two models.





Regularized Linear Methods

Tree Based Methods

Neural Networks

Support Vector Machines

Ensemble Methods

Results

Conclusion

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].