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

Ted Henson and Mario Giacomazzo

University of North Carolina at Chapel Hill

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]. 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].

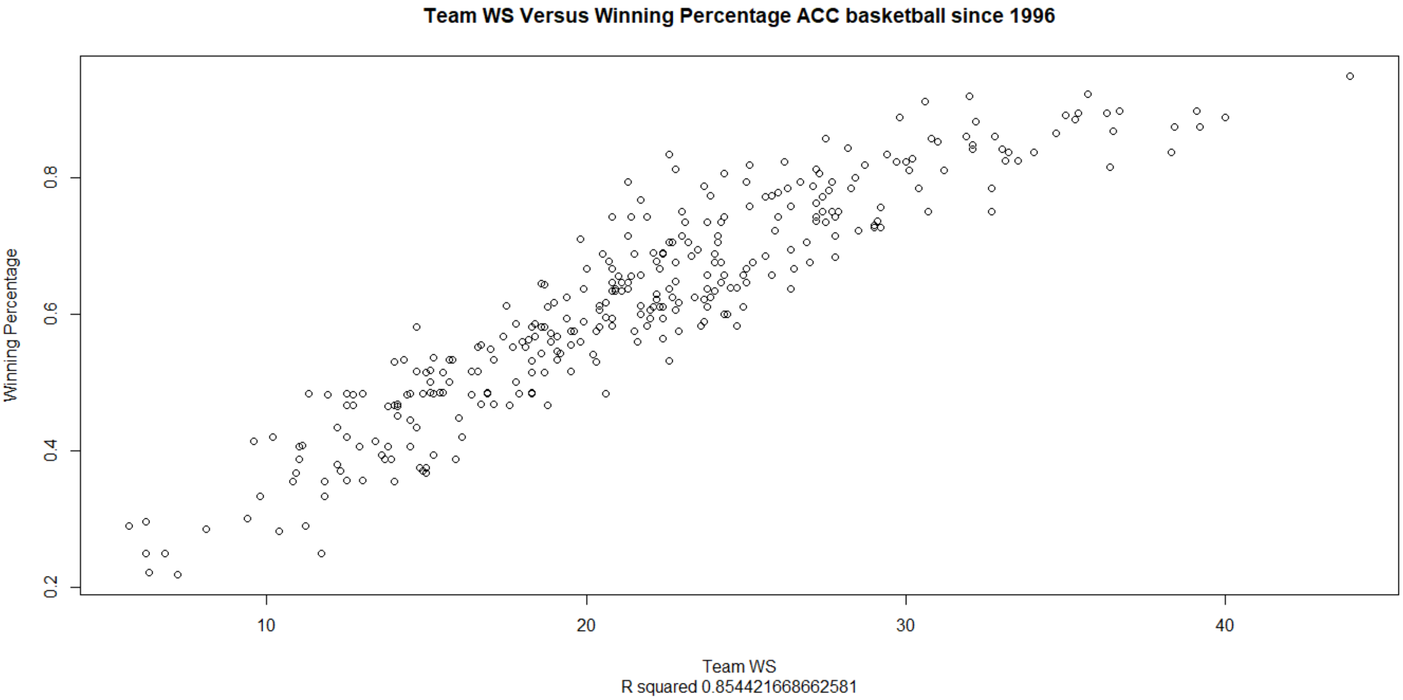
Accusations of many NCAA coaches paying high profile recruits hundreds of thousands of dollars to single recruits was the catalyst for this analysis. If one player can cause decorated coaches to potentially use unethical methods, then programs should explore every possible avenue of predicting college 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 incorporating AAU statistics and adjusting for players' strength of schedule could significantly improve the models presented in this paper.

Data

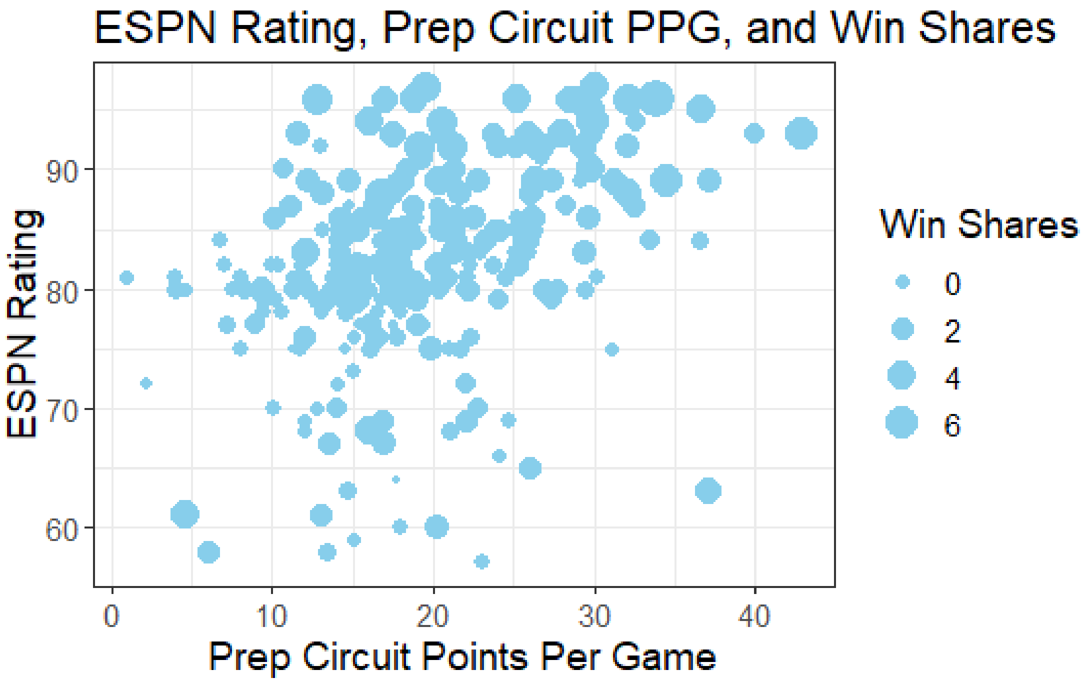
Every statistic listed on a player's college basketball reference page was collected; however, only a player's first season playing in the NCAA was used in the modeling process in order to fairly evaluate a player's true production out of high school. In addition, we only selected players who played more than 10 games to control for injuries. 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 [4].

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. The ESPN dataset had hundreds of players; however, many of these players did not have Prep Circuit data or had missing values in most of the tables. As a result, this analysis only considered players with at least Prep Circuit points data. In all models, 273 players were in the training set, and 66 were in the test set.

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. This data did not include any AAU games. 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. This explanation was rejected by noticing that other Chino Hills players had assists logged in games that Lonzo's table says they did not. Due to these apparent inaccuracies in the number of games logged, the true number of games played was estimated by the max of all games logged for an individual statistic. This measure was then used to recreate the per game statistics. The average number of games logged was also added as a reliability index. The full model included both the given Prep Circuit per game statistics, totals, and the transformed per game statistics. In addition, if a player had missing values for a given statistic, k nearest neighbor was trained and then applied to estimate the missing value in the training set. The same preprocessing method was applied to the test set. The most predictive variable was a player's points per game. Below is a scatter plot of ESPN Rating and Prep Circuit Points Per Game. The sizes of the dots represent a player's win shares.



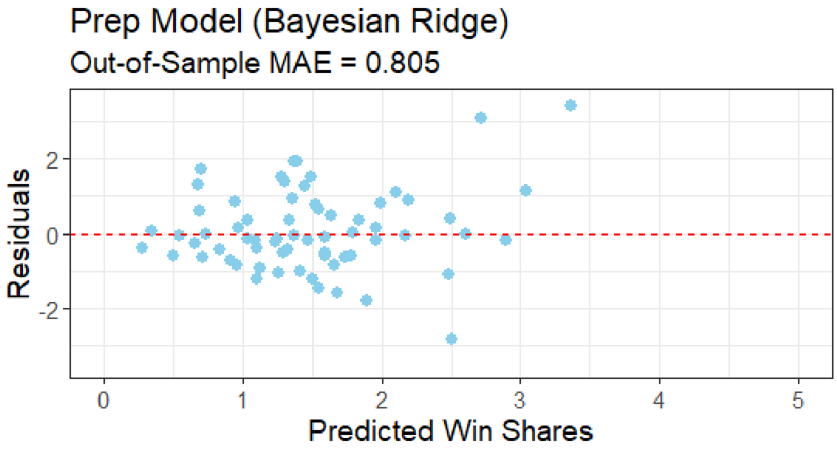
Methods and Results

In order to assess the added value of the two different data sources, two models were built with only one data source, and a third model was built combining the two. Four models, two linear and two nonlinear, were trained for each of these input combinations. The two linear models utilize identical information but are estimated using two forms of regularization, lasso and ridge penalties. A neural network and random forest were chosen to detect non linearities and interactions between variables. Additionally, a linear stacked regression model was built using the predictions of the previous four.

Exploratory data analysis showed that the ESPN rating may have a nonlinear relationship with win shares. Due to this discovery, orthogonal polynomial terms up to degree 3 of the ESPN rating were added to account for this nonlinear nature of the ESPN variable, which increased model performance. Below is a table of each models' out-of-sample mean absolute error:

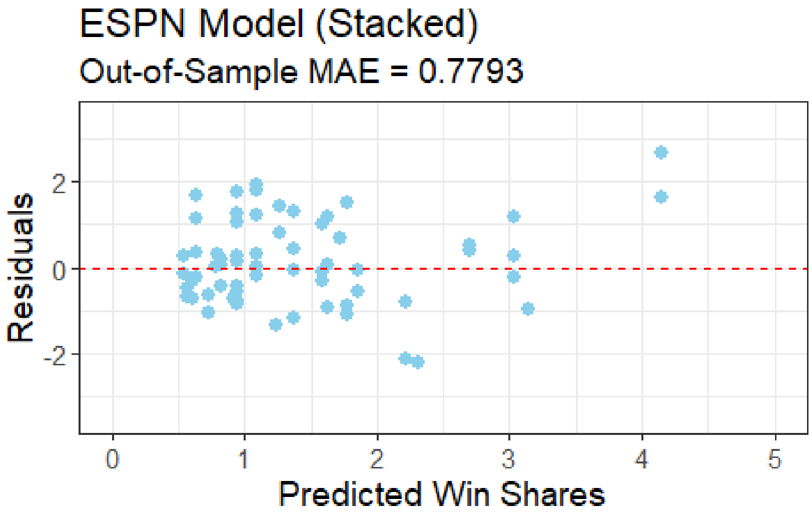
Model Type	**ESPN Model**	**Prep Model**	**Full Model**
Bayesian Ridge	0.789985	0.804995	0.745034
Bayesian Lasso	0.792723	0.817524	0.740465
Random Forest	0.868088	0.879499	0.841297
Neural Network	0.79568	0.809442	0.749521
Stacked	0.779274	0.839882	0.737103

This table revealed three important takeaways. First, the ESPN models performed marginally better than the prep models across each model type. The only corresponding difference in MAE's greater than about .02 was between the stacked models. Second, the random forest models predicted significantly worse than the other four model types within each input combination. Third, the full models performed better than the ESPN and prep models across each model type. Below are the plots of fitted values versus the residuals for each input combination's best model:



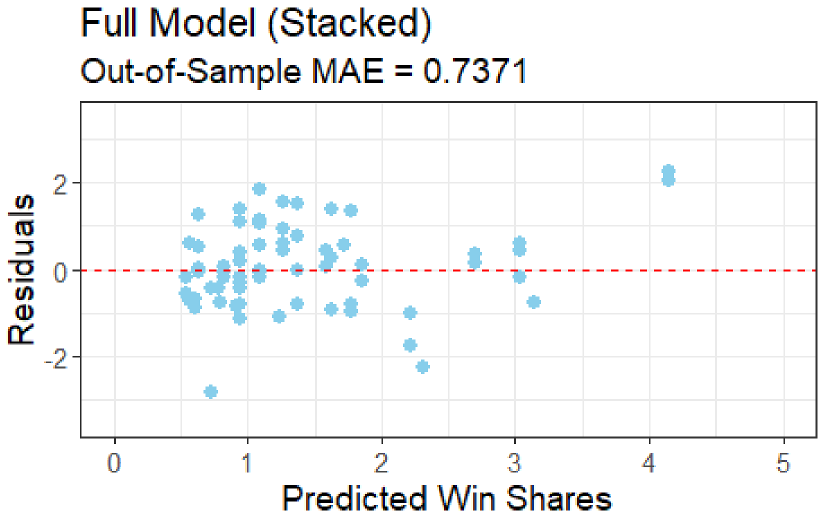
Prep Model (Bayesian Ridge)

Out-of-Sample MAE = 0.805



ESPN Model (Stacked)

Out-of-Sample MAE = 0.7793



Full Model (Stacked)

Out-of-Sample MAE = 0.7371

Conclusion

Although the model involving only Prep Circuit data performed worse than the ESPN model on the test set, its out-of-sample MAE was only about .02 worse despite not controlling for quality of competition. Future models could benefit tremendously by creating a quality of competition index using Prep Circuit's game logs and incorporating Prep Circuits AAU Statistics (specifically Nike EYBL), which has been tracked in greater detail for the class of 2019.

It is important to judge the data and its predictive accuracy with some perspective. In a project that lasted less than 4 months, most of which revolved around data collection, high school statistics improved and nearly matched a rating system that has been improving upon itself for 13 years. This finding should encourage more analysis and collection of high school data not only at the collegiate level, but at the professional level, where one draft pick can change a franchise for decades. If this data holds predictive value at the collegiate level, there is reason to believe it can assist an NBA projection system, particularly in cases where a high school superstar falters and a dark horse emerges.

References

[1] "ESPN Basketball Recruiting - Player Rankings", *ESPN*. [Online]. Available: http://www.espn.com/college-sports/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] 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_professional. [Accessed: 04- Apr- 2019]

[4] "Calculating Win Shares", *Sports Reference*. [Online]. Available: https://www.sports-reference.com/cbb/about/ws.html. [Accessed: 04- Apr- 2019]