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

Ted Henson and Mario Giacomazzo

The goal of this paper was to predict collegiate basketball success from recruiting ratings and high school statistics. Accusations of many NCAA coaches paying high profile recruits hundreds of thousands of dollars 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. Combining linear and nonlinear models, we predicted collegiate win shares using high school statistics from Prep Circuit and recruitment rankings from ESPN.

We chose Prep Circuit over similar websites for its reputation, reliability, and consistency. For our analysis, we only considered collegiate athletes where at least points per game were recorded. This criterion seemed to be a good indicator that the data on Prep Circuit was accurate and valid for an individual. When necessary, imputation via K Nearest Neighbors filled in missing high school information. Furthermore, we engineered a reliability index to assess the quality of the Prep Circuit information on each player. Since juniors lacked data, we only gathered high school statistics for seniors. ESPN's recruiting database assigns grades to high school basketball players based on scouting reports. Since many universities rely on expert opinion from scouts, it is reasonable to conclude that recruitment grades can be used to predict collegiate success. Prep Circuit and ESPN's recruitment database are two of only a few sources of data measuring high school performance of basketball players.

Win shares was chosen as the response variable due to its all-encompassing nature of a player's contribution to team success. Based on the current "one-and-done" climate, universities are looking for players to contribute immediately; therefore, we only extracted win shares for the freshman year. The table below shows out-of-sample mean absolute error (MAE) of 15 different models. Three different input combinations were chosen using identical modeling techniques in order to assess the value of the data: ESPN only, Prep Circuit only, and a full model.

Model Type	**ESPN Model**	**Prep Circuit 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

We highlight two interesting results. First, ESPN models consistently out-performed Prep Circuit models indicating the power of expert opinion. Second, incorporating Prep Circuit data with the ESPN data led to further reduction of out-of-sample MAE.

It is important to judge these results with some perspective. Original modeling with high school statistics improved and nearly matched a rating system that has been improving upon itself for 13 years. This finding will encourage more analysis and collection of high school data for collegiate and professional purposes. Future models will improve through more advanced methods and incorporating AAU statistics which has been tracked in greater detail for more recent seasons.

Word Count: 484