University of California, Berkeley

Data 100 - Principles and Techniques of Data Science

Spring 2020

College Type: Does it Matter for the NBA

Authors:

Gary Kwong

Vincent Chiang

Abstract

This report demonstrates whether or not private schools produce better NBA players in terms of stats, (i.e. points, rebounds, assists) than state schools. Given the pool of player data from 2012 to 2018, we are restricted to using relatively newer player data and hence our results and conclusions reflect those of the newer player base. Our final model(s) may or may not predict the best fit but gives us insight into whether or not public schools and private schools are different in how they mold players. The model ideally helps demonstrate the difference(s) if there is any between private and public university players currently in the NBA.

Table of Contents

List of Figures, Equations, & Tables

Introduction

Basketball as a sport has evolved over the years since its creation in 1891. It has grown from a basic conditioning "event" to now being a widely enjoyed sport. As a result, the game has grown and expanded from amateur leagues to professional leagues over the world. With so many prospective players wanting to play professionally in the NBA, leagues such as the G-league and college league (NCAA) were implemented as a way to enjoy the sport as well as funnel talent into the NBA. With around 20,000 NCAA basketball players and only 60 players being chosen each year in the draft, the quality of the picks is of utmost importance to NBA teams. Hence the production of talent between two "different" school systems (public (state) vs private) are two key factors to examine. The common perception is that private schools have more freedom to allocate funds towards their sports programs, hence have more intricate methods to recruit better players (from high school) and better facilities would entice better players to grow within their environment. Though there have been star NBA players who have come from both types of schools (Stephen Curry - Davidson (private), Michael Jordan - UNC (public), we do not have a set way to look at how these schools may operate. Furthermore with the NCAA now allowing players to earn money off of their image this raises the practical question, will private colleges gain an edge (or extend a lead) over public colleges in terms of talent acquisition and creation for the NBA?

Data Cleaning and Method

Initially, we were given a set of CSV files and we selected to use the college.csv and player_box_score.csv files which provided us with data in regards to a player's career in the NBA and NCAA. The player box score data contains the NBA player's statistics from the years 2012 to 2018. Given this smaller subset of players than what appears in the college dataset as the college dataset ranges from the 1960s to 2010s, we have to be more generous in our cleaning techniques. Given that a typical NBA game consists of 48 minutes in regulation we felt that if a player played at least 10 minutes in a game then their stats would be included as playing 10 minutes would indicate some form of outside factors that are not measured (stamina, etc.). In addition, for each of the games and players, those who played at least 10 minutes made up of about 86.67% of the data given in player box score. Moreover, we also decided to drop players that did not play over 34 games over the course of the 2012-2018 seasons. This is to account for the rookies in the dataset who played at most 1 season, which is 82 games (not including postseason games), hence roughly 70% of the data played at least 34 games. We then scraped data off of Wikipedia (https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_institutions) to assign each school to a specific type (public (state) or private). Next, we merged the 3 datasets by player name and college name, this way we get a table that contains the type of college and the player's NBA statistics. Furthermore, we grouped by the player's display name (as the display names are different), and a simple google search would show that currently, no two NBA players have the same first and last name. After selecting the basic statistics, player points, assists, rebounds, blocks, height, weight. We further simplified the model by making the statistics (points, assists, rebounds, blocks) per game statistics. We also dropped players listed as G or F as we followed the 5 position layout (PG, SG, SF, PF, C) and there were only 3 listed guards at the end of the above cleaning. Lastly, we checked if any null values existed, but there were none. At the end of cleaning the data, we were left with 416 players composed of 296 from public schools and 120 from private schools. (71% and 29%).

Given the state of the data, we decided to go with a Logistic Model since we have two classes (Public or Private that we wanted to predict). Using the cleaned table we decided to fit all the features first, using height, weight, points per game, rebounds per game, assists per game, blocks per game, rebounds per game, and player position. From the data points, we can visualize the difference between public and private schools in regard to points per position (Figure 1). Public schools have higher average points per game (ppg) in two positions while private schools have higher average ppg for two other positions, this does not give us a clear distinction between the two types and this becomes a trend for the other statistics as well.

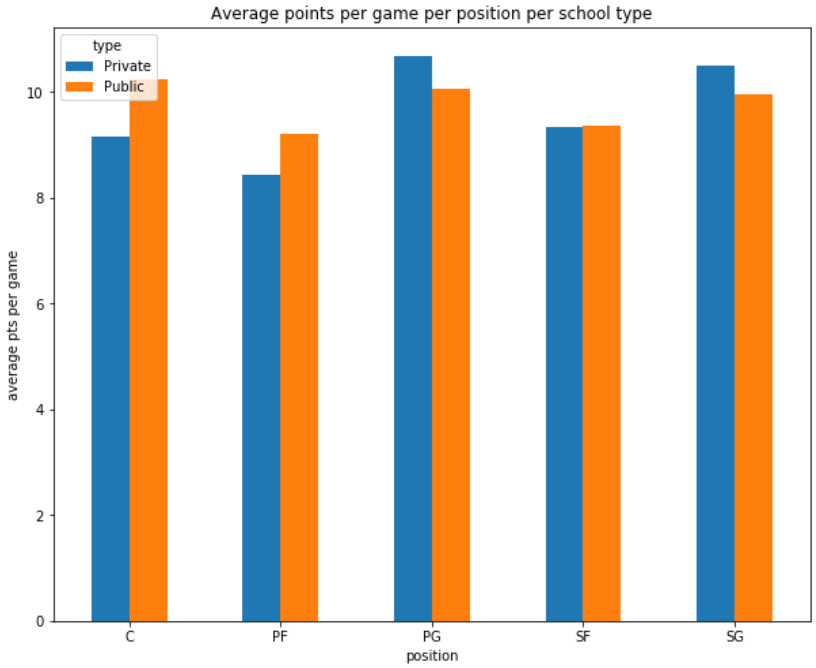

Figure 1: Points and Position separated by School type

Using a train-test split of the data we fitted the train portion of the split with the Logistic model. The initial accuracy (r^2) of the model was 0.71, hence we needed to use cross-validation to see which of the above features were going to produce the best Logistic model. Using this method we find the cross-validation (cv) accuracy starts decreasing when the “height” feature is added (Figure 2). Furthermore, with the first four features, we get the same training accuracy and cv accuracy, hence we chose to be more cautious about overfitting. We selected the best model as the one that fitted points and assists per game as features.

Figure 2: Training and Cross-Validation accuracy for select features

Applying this "best" model to the test set we got an accuracy of roughly .74. This value does not seem really good as we will discuss in the conclusion, hence this prompted us to run a decision tree model and a random forest model to model our train and test data. Plotting the two features against each other relative to the school type we can see there is not much of a split between the two features per type (Figure 3). Hence a decision tree model may not be the best model to compare against. We find similar progress using a random forest model because of the classes (Private vs Public) overlaps.

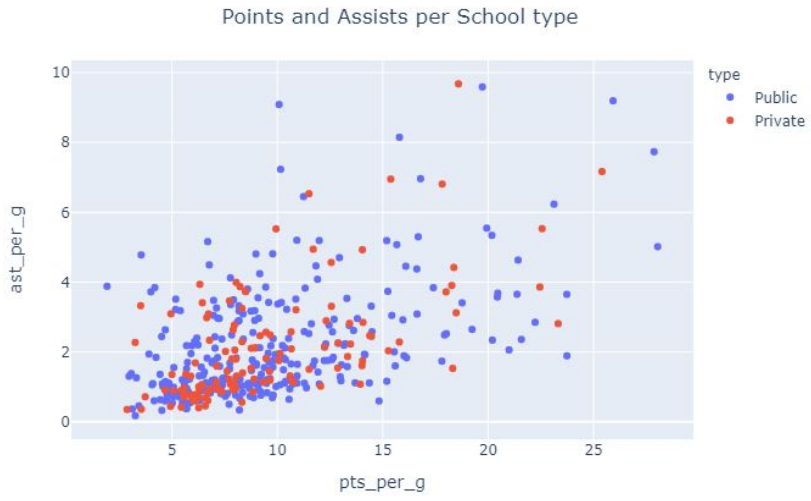

Figure 3: Points and Assists per School type

Summary and Discussion

The model created helps us understand whether or not private schools are different from public schools in terms of their basketball talent production. Initially, we came into this project assuming that on average private schools would have a slight advantage over public schools in terms of talent production (alumni success in the NBA). With our initial model, we made a logistic model to classify a school's status based on selected features from players in the NBA as mentioned in the data methods section. We found that our best model used a small number of features of our selected features (Figure 2). We had calculated that if we built a model based on just the proportion of public schools we would expect to guess around 220 schools which is .587 of the schools correctly. However with our model we have an accuracy of .73 which is not entirely bad but is nowhere near good enough to claim we can construct a good model to tell whether a school is private or public from basic player statistics alone. Hence, we built a random forest and a decision tree model to see if we could go any further with this analysis. Using the two best features we got a training accuracy of .986 but a test accuracy of .62, this would mean that we were somehow overfitting on the data or we didn't have enough data points to support using this method (Figure 4). This shows how the decision tree saw the data, and it shows that there were no "decisive cuts" (besides at around 10 assits per game) which contained no points.

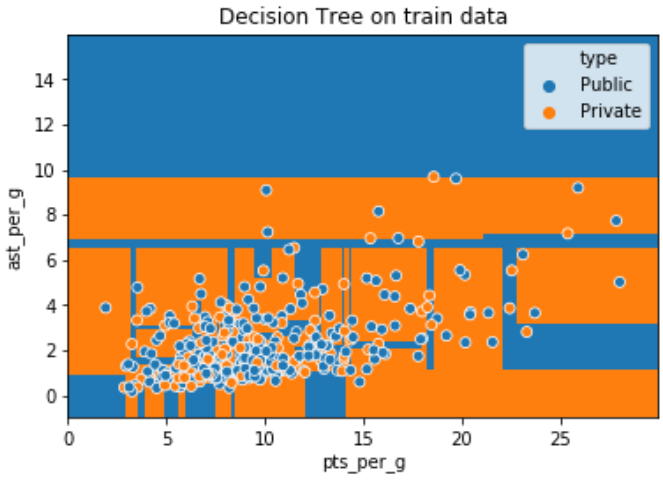

Figure 4: Decision Tree model on the training data

A forest model showed about the same conclusion, heavily predicting points as private (Figure 5). This gave us a training accuracy of .98 but a testing accuracy of .71, adding more features mainly overfit the model hurting both the training and testing accuracy.

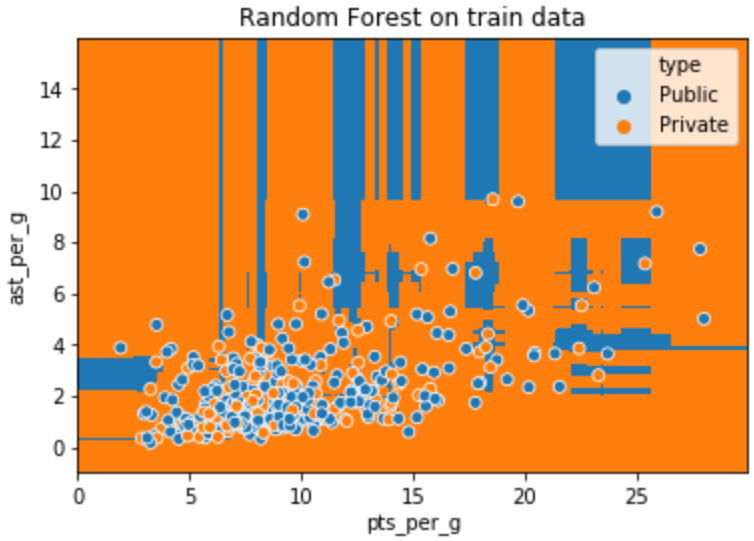

Figure 5: Random Forest model on training data

Features such as points, assists, blocks, and rebounds per game were all included in our model due to the fact that these were the only reliable stats that differentiate individual players from others in the league. We thought a player's position would help us define a player's identity further, but it failed to improve the model. After further discussion, we concluded that due to different higher institutions are more likely to produce the same amount of players for each position due to the low team retaining period of at most four years, so a player's position will serve no purpose in differentiating a player's likely college type from the others. The main limitation in working with the data we are given was the lack of representation from people from different divisions in the NCAA. There was only one single player in the league that's not from a Division-1 school that played in the 2018 season, so any difference in the quality of players from private and public institutions outside of D1 was not explorable for us. It was also the same for players from Division-1 schools that didn't make it into the NBA. Our data is ultimately biased since the players in the NBA had already been cherry-picked by all the team through NBA drafts.

As discussed above, one ethical dilemma when working with this data was excluding players that did not make it into the NBA. Another ethical dilemma that we faced was whether to weight star players more than below-average players when modeling. We decided against this idea as players, good or not, represent their university equally, and they all had their exceptional performance during college for them to be selected into the NBA. Additional data that can help us further improve our model will be the salary and endorsement of every individual player. The amount of money they earned, can serve a large part when it comes to determining how successful a player is since players can be highly valued as a game winner or a highly marketable athlete. The concern about including financial data into our model was a potential of devaluing a player's value due to having a relatively low-earning contract or simply being a loyal player to stay with the same team instead of going after huge money signings. We could address this by looking at their current market value instead, but this will only lead to older and potentially highly successful to have lower weight in the model. Thus, we cannot say with great certainty that this model can predict whether or not a college is public or private, but from the models tried above we can infer that public schools are much more close to private schools in terms of talent (through player points, rebounds, and other stats) acquisition or molding for the NBA than we previously thought.