Final Report: From High School Hype to NBA Stardom

Isaac Ke

1. Introduction

LeBron James. Stephen Curry. What do these household names have in common? Surely, they both have been named the NBA's Most Valuable Player (MVP) on multiple occasions and have led their teams to numerous world championships. However, only LeBron James was ranked the number one best player coming out of high school. In fact, Steph Curry was *unranked* by recruiters in high school which meant that he went so unnoticed that he couldn't even crack the bottom of the list. Is future stardom really that hard to predict? Or was Steph Curry a once-in-a-blue-moon type of phenom?

In this age of social media and video streaming, the hype surrounding top high school and college players can blow-up quite literally in an instant. Hoop mixtapes posted on YouTube or highlight reels circulating around Twitter are just a few of the efficient ways for young athletes to catch the eyes of college or NBA scouts.

But how does a player's success at the high school or collegiate level impact how they will fair at the professional level? By examining and modeling athletes from their high school basketball days through their NBA career, we can gain an understanding on how a young athlete's potential translates to future stardom, if at all.

1.1 Importance of Study

A better understanding of how professional success can be predicted based on key high school, collegiate, and advanced NBA statistics will allow recruiters to re-focus their scouting programs to bring in the best athletes to their program while minimizing resource expenditure. Having more "box-office" athletes pass through college programs and NBA team arenas will certainly boost revenue and spur on the overall economy of the city.

Similarly, influential sport companies such as Nike or Adidas will be able to offer endorsement deals to young blossoming athletes with less risk and a higher rate of return. Coaches and NBA executives will be able to sign players to more valuable contracts which would help them build winning franchises that pack the stands with fans.

Furthermore, we can redefine false stigmas surrounding young athletes, especially those who are discouraged by the seemingly daunting obstacles they face early on in their professional basketball careers.

1.2 Data

Data was gathered from data-world [1]. A comprehensive data dictionary can be found at the provided link. The data contains information on high school basketball players who played between 1998-2013. Some information included are the player's name, their high school rank, draft information, advanced NBA statistics, total seasons played, and their highest NBA level reached. For this particular dataset, the high school rank reported is the RSCI (Recruiting Services Consensus Index) [2]. This ranking is an objective ranking that combines rankings from experts such as ESPN, MaxPreps, and Rivals. It is important to note that for rankings and draft position, a value closer to 1 (numerically lower) is considered better. A summary table containing the number of players for each category of "highest level reached" is shown in Table 1.

Highest Level Reached	Number of players
High school	32
College	1058
Draft	31
Rookie (2 years in NBA)	122
Bad (in NBA)	416
Good (in NBA)	130
Great (in NBA)	46
All-star (in NBA)	38

Table 1: The number of players in each category of "highest level reached"

1.3 Aims

My first aim is to test my hypothesis that a player’s high school rank and draft position have a strong positive association with NBA career longevity. I aim to uncover what the true relationship is between a player’s pre-professional rankings and how many seasons they last in the NBA.

My second aim is to test my related hypothesis that the better a player’s high school rank and draft position are, the higher the player’s maximum level reached will be (draft, rookie, bad, good, great, all-star). My goal is to assess which factors are good predictors of peak performance.

2. Methods

2.1 Exploratory Data Analysis

In the exploratory data analysis phase, I investigated the “college funnel” and other aspects of the basketball journey to the pros. The college funnel is the phenomenon where a large number of athletes play basketball at the collegiate level, but only a small percentage go on to play professionally in the NBA. First, I assessed the prevalence of missing data and then decided if it was reasonable to remove applicable observations, or if imputation was necessary. I also created boxplots to check for outliers.

Furthermore, I used histograms and bar charts to look at the distribution of NBA performance across different colleges and time. I also analyzed the conditional distribution of high school rank and draft pick given an NBA player’s performance level. Even more, I observed the distribution of the number of seasons played across different levels of high school rank, draft number, and NBA statistical ratings.

In addition, scatterplots and a feature correlation matrix were generated in order to investigate how a player’s stance in one part of the NBA journey, in terms of their ranking, related to their success at the next subsequent level (high school transitioning to college and college transitioning to NBA). Lastly, I investigated which covariates were highly correlated with one another and with the target responses of interest.

2.2 Gradient Boosting Machine

2.2.1 Gradient Boosting Machine Overview

In order to address my first aim of testing my hypothesis that a player’s high school rank and draft position have a strong positive association with their NBA career longevity, I chose to implement gradient boosting machines (GBM). GBM is a powerful machine learning algorithm that uses decision trees. It is an ensemble method which constructs a group of successive shallow and weak trees by learning from the error of the previous iteration. Shallow refers to having relatively few splits in the decision tree, and weak means that the error rate is only slightly better than random guessing.

The principle behind boosting algorithms is to do ensemble formation sequentially. At each iteration, a new weak learner (in my case a regression tree) is trained in context of the error of the whole ensemble. First, an initial decision tree is fit to the data, then the next decision tree is fit to the residuals of the previous tree, and then that tree is added to the model. This continues on until the error rate stops changing significantly. The “gradient” part in GBM refers to the optimization algorithm of gradient descent. In essence, regression trees are fit by minimizing the loss function. Gradient descent measures the local gradient of the loss function given a set of parameters and moves in the direction that has the steepest gradient. In the end, the final model is a stepwise additive model made up of individual regression trees.

There are many reasons why I chose to implement GBMs. One, tree-based methods have been shown to perform well on unprocessed data (without scaling, normalizing, centering, etc). With a bunch of shallow and weak trees, the algorithm has great speed and accuracy. This is because weak learners require less computation and are able to learn slowly, making detailed adjustments in key areas to improve the error rate. With small incremental improvements at each iteration, overfitting can be efficiently detected with cross validation, and the algorithm can be stopped immediately. Lastly, the algorithm handles missing values, so imputation is not a required step.

For my use case, I used the *gbm* package, which was the first implementation of gradient boosting in R [3]. It is an implementation of Freund and Schapire's Adaboost algorithm and Friedman's gradient boosting machine.

2.2.2 Data Preparation & Split

First, I removed all observations which had missing values for the variable *total seasons*, which was my target response of interest. 943 out of 1,885 players had no data for *total seasons*, so these observations were removed. This left 942 players, or about 50% of the original data. Next, I did a random split of my data into 80% training and 20% testing.

2.2.3 Model Training & Hyperparameter Tuning

To tune the parameters of the model, I used grid search where I defined a set of ranges of parameters to search over. I used 5-fold cross-validation to tune the number of trees in the model, the depth of each tree, the learning rate of the gradient descent, the minimum number of observations in terminal nodes, and the sampling percentage for the stochastic gradient descent.

2.2.4 Model Selection Criteria

For model selection, the root mean square error (RMSE) loss function was used. Root mean square error is the square root of the average squared difference between the actual response and the fitted response across multiple iterations. Through 5-fold cross-validation, the model that had the lowest RMSE was the model that was selected.

2.3 Proportional Odds Model

2.3.1 Proportional Odds Model Overview

In order to address my second aim of testing my hypothesis that a player's high school rank and draft position have a strong positive association with their maximum performance level reached, I fit a proportional odds model (also called ordered logistic regression) to my data. Ordered logistic regression is an extension of binary logistic regression that applies to a categorical response variable with two or more *ordered* levels. One key difference between logistic regression and ordered logistic regression is that logistic regression assigns a probability that a variable will *equal* a specific value while ordered logistic regression assigns a probability that a variable will be *less than or equal* to a certain value.

Ordered logistic regression models the log odds as a linear function of the covariates. The coefficients in the model cannot be estimated using ordinary least squares, so maximum likelihood estimation is used instead. These estimates are obtained using iteratively reweighted least squares.

The key assumption of the proportional odds model is stated in its name: "proportional odds". The proportional odds assumption says that the relationship between each pair of response levels is the same. Simply put, ordinal logistic regression assumes that the coefficients that describe the relationship between two response categories are the same for any other pair of response levels. In terms of log odds, the number added to the log odds ratio to get from, say the 1st response level to the 2nd response level, is the same for each step. This assumption is also called the parallel regression assumption because mathematically we obtain j-1 different regression lines, where j is the number of levels in the ordered response variable. Each line has the same slope but different intercepts. Due to this assumption, there is only one set of coefficients.

For my particular aim, a proportional odds model is appropriate because I have multiple discrete numeric covariates with a response variable that is both categorical and ordered.

2.3.2 Data Preparation

To begin, all observations that had any missing values for the variables *high school rank*, *draft pick*, or *highest level* were removed. 1,471 observations were omitted, leaving 411 observations with complete data for *high school rank*, *draft pick*, and *highest level*.

2.3.3 Model Selection Criteria

In order to gain an understanding of which coefficients should be included in my model, I calculated the p-value of each estimate of the coefficients and intercepts (for each level). This p-value was obtained using a one-sided test assuming normality (since my sample size was 411). The reason for doing a one-sided test was because it is expected for the coefficients to be strictly less than 0 due to an inverse relationship. When one's ranking goes numerically up (gets worse), one would also expect the odds of him being able to reach the next performance level to go down.

Keeping my specific aim in mind, it is preliminarily beneficial to include both *high school rank* and *draft pick* in the model in order to do prediction on hypothetical players who have data in both of these categories. Thus, both the value of the coefficient as well as the related p-value were taken into consideration when deciding which variables to keep in the model, as will be discussed in section 3.3.1.

3. Results

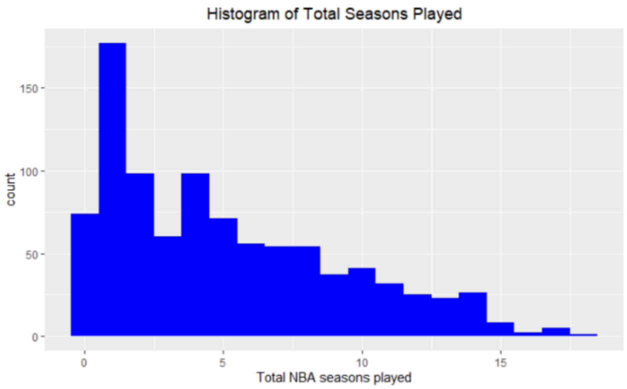
3.1 Exploratory Data Analysis

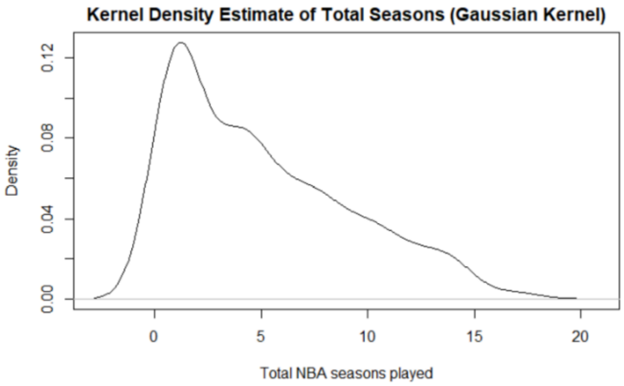
3.1.1 Discussion on Missing Data

In my dataset, there were many missing values, but I chose not to remove any observations upfront because the pattern of the missing data was very revealing; the data was not missing at random. For example, there were 1,515 players with no data for their NBA statistics and rankings (win shares, value over replacement, plus minus, and wins added). Because my data set was from 1,885 *high school* basketball players, this missing data was indicative that those players did not make it to the NBA or did not last long enough in the league to accumulate these statistics.

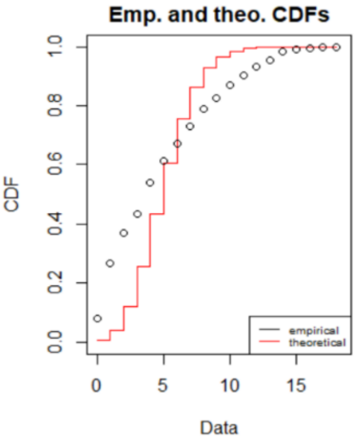
3.1.2 Distribution of Variables

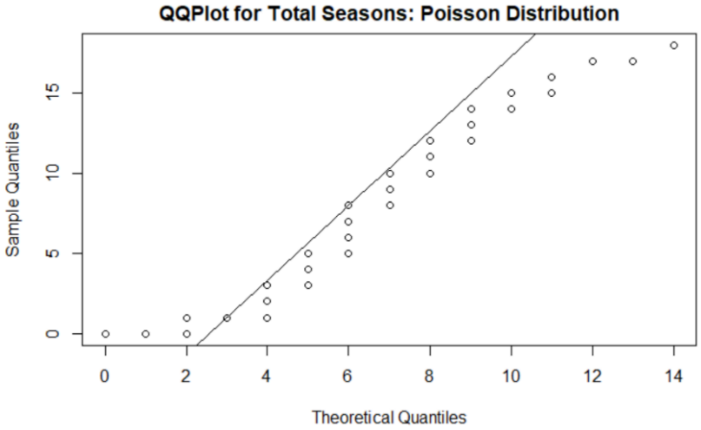
First, I generated various plots to analyze the distribution of key variables. In the following two plots, we see the distribution of total seasons played.

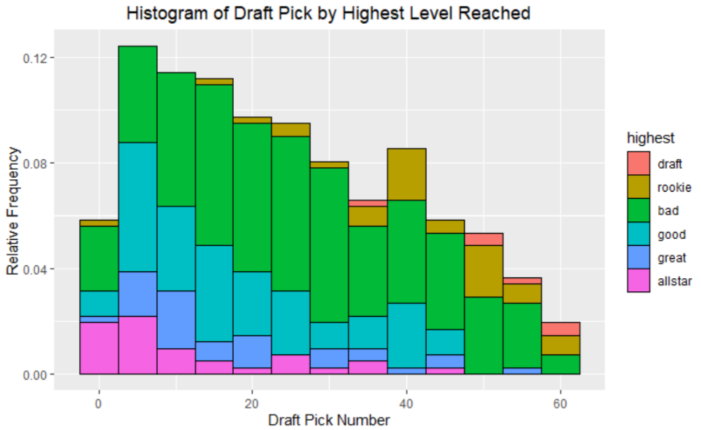




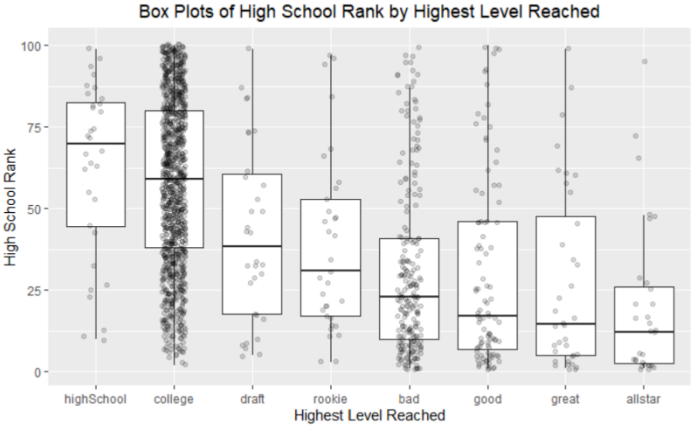
From the histogram and kernel density estimate, I conjectured that the variable had a Poisson distribution because it was discrete and right-skewed. The MLE of the Poisson parameter was λ=5.05 with a standard error of 0.073. From the following two plots of the empirical versus theorical cumulative distribution function (cdf) along with the Poisson quantile plot, I concluded that the Poisson distribution is a good fit for the random variable *total seasons*. The empirical and theoretical values match up reasonably well.





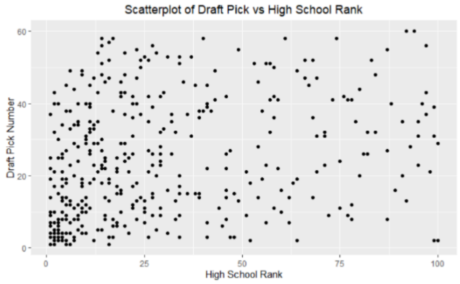


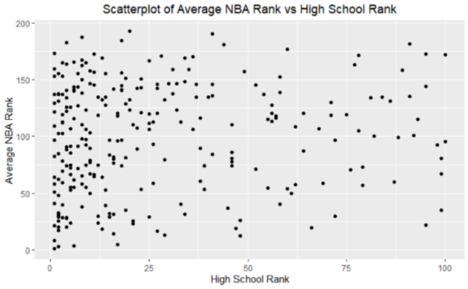
In the histogram to the right of draft number by highest NBA level reached, we can see that more players who become all-stars (pink) are picked higher in the draft (number closer to one). Interestingly, the number of “bad” players (green) seem to be clustered in the center of the draft rather than toward the end of the draft. Surprisingly, there is still a handful of “bad” NBA players who were drafted very high. Players like these who do not live up to their hype are called “busts”. Overall, we see the trend that the higher the draft pick, the higher one’s performance ceiling is in the NBA.

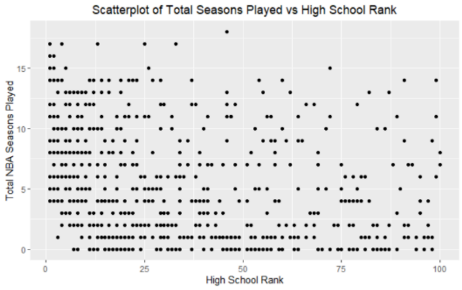


Going back one level from college to high school, we notice the similar trend we saw in the previous histogram. In the box plots to the right, we see that the median high school rank goes up (numerically decreases) the higher up one goes in NBA level. We also see that the majority of players in our dataset hit a ceiling in college, which is explained by the aforementioned college funnel. Interestingly, the distribution from high school to all-star also moves from being left-skewed to right-skewed. This can be attributed to the fact that there seems to be a presence of both underdogs (players who rank low but perform high) and busts (players who rank high but perform low) at each level that skew the distribution.

Next, to analyze how a player’s ranking translates from high school through the different stages of the NBA journey, I generated the following three scatterplots to show how a player’s high school rank compares to their draft pick after college, their ranking in the NBA, and how long they ultimately last in the NBA.





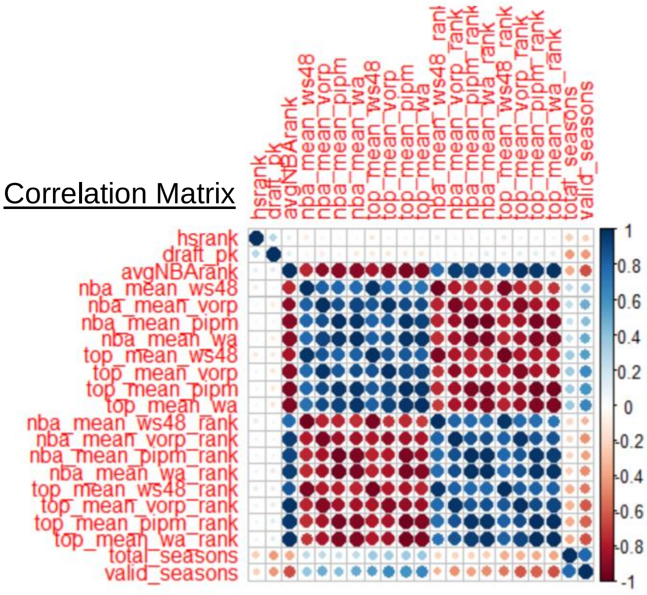


In all three of these plots, there is a noticeable clustering of points in the section of the graph corresponding to better high school rank and favorable draft number, NBA ranking, and total seasons played. This suggests that a player’s success at a young age in high school does indeed positively impact their success at the NBA level. Furthermore, it provides preliminary evidence that success in one stage does in fact translate to further success in the subsequent stage. This can be attributed to the fact that increased success of a player gives that athlete more exposure to scouts at the next level, especially with today’s access to the internet and online ranking systems. Thus, there are more opportunities available for better players to take their game to the next level.

3.1.3 Association Between Variables

Lastly, I created a correlation matrix with all of the numeric variables in my data set. The pattern is very distinct. Since lower numbers are considered a better ranking and higher values are considered better for NBA statistics, there are visible squares of positive and negative correlation formed in the matrix. The Pearson correlation *r* across different NBA statistics is strong, greater than |0.7|. This is expected because when a player is producing at, say a high level in the stat category of *win shares per 48 minutes*, they will most likely also be producing well in the area of *total wins added*. What stands out is that the correlation between high school rank and draft pick is evident, but not too strong, with *r* = 0.28. Furthermore, the association between these two rankings with various NBA statistics is fairly weak (as indicated by the small faint circles in these entries). However, the correlation between total and valid seasons with high school rank, draft pick, and various NBA statistics is much more prominent, with 0.3 < | *r* | < 0.7 for these associations. This shows that there is a stronger association between a player's NBA *career length* with their high school, college, and NBA rankings compared to the association between a player's NBA *statistics* and their high school and college rank. So, it can be seen that a top-ranked high school player is more likely to have a *longer NBA career* rather than averaging *better stats* in the NBA.

Correlation Matrix



3.2 Gradient Boosting Machine

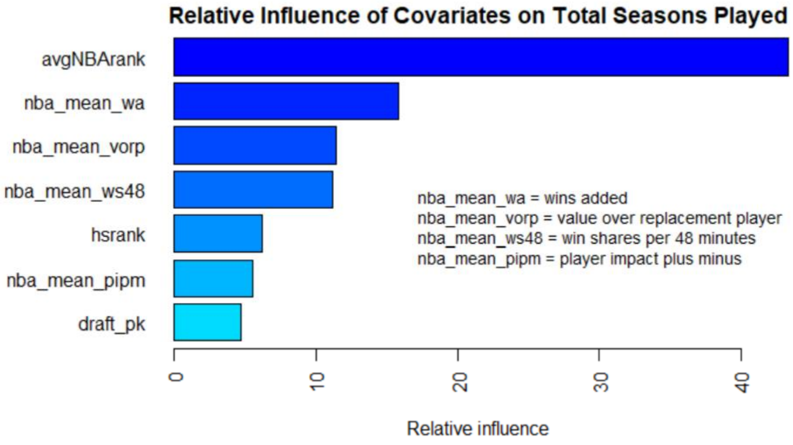
3.2.1 Selected Model

The final model that had the lowest RMSE of 2.624 was the model that had 36 trees, a learning rate of 0.1, a tree depth of 7, a minimum of 10 observations per terminal node, and a 0.8 sampling rate for stochastic gradient descent. From these parameters, we can conclude the following. The use of stochastic gradient descent over regular gradient descent showed that there might have been local minima and plateaus in the loss function gradient. Also, the fact that the tree depth is greater than one shows that there are likely some important interactions that require deeper trees to uncover.

3.2.2 Variable Importance

Looking further into the final tuned model, I investigated which covariates had the most impact on the model. In the variable importance plot to the right, the relative influence of each covariate is shown. At each split in the tree, the improvement in mean square error (MSE) is computed. The relative influence of each variable is the average decrease in MSE across all trees that used that variable.

Average NBA rank has the most importance in modeling total seasons played. This makes sense since players that perform better in the NBA and thus rank higher are more valuable to teams. Thus, when this player's contract ends, they are more likely to sign an extension or move to another team that offers another long-term contract. Higher performance means



more winning, more winning means more revenue for teams, and more revenue and success means a team wants to keep their players for a longer time.

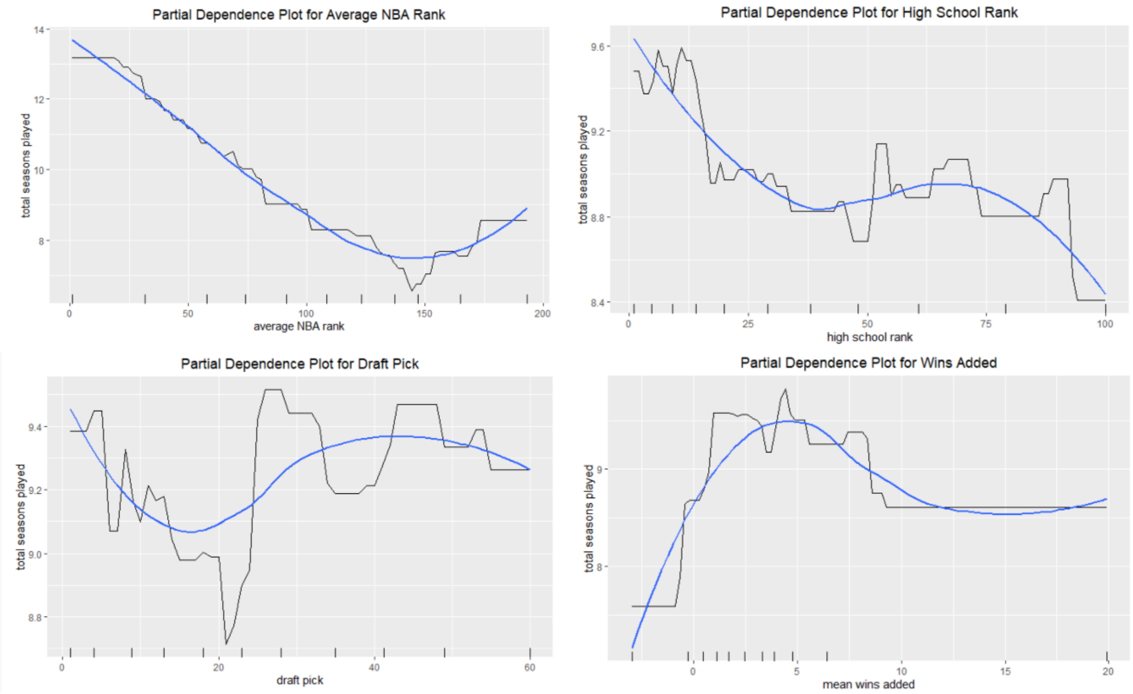
Interestingly, high school rank and draft pick do not have as much of an effect on career longevity, about 15% of the influence, compared to average NBA rank. Even so, high school rank has 25% more relative influence than draft pick. These two observations indicate that a player's performance in their young basketball career (high school and college) do not have as much of an effect on NBA career longevity when compared to their performance *in* the NBA. Furthermore, a high school player's ranking is just slightly more indicative of total NBA seasons played compared to draft pick. This can be due to the fact that high school is where a player first becomes popular and gains national attention from professional scouts.

One more variable that stands out is a player's mean wins added. This variable has the second most impact on our model which supports our claim that the more a player contributes to a winning franchise, the more valuable they are and thus the more seasons they play in the league.

3.2.3 Partial Dependence

To analyze the specific impact each variable has on our target, I created the following partial dependence plots (PDPs). PDPs plot the average change in the response when one variable is varied while the others are held constant.

In the PDP of average NBA rank, we see that there is a positive trend between a higher NBA rank (numerically lower) and a player's total seasons. If a player moves up 25 spots in NBA ranking, he can expect to last one extra season in the NBA, on average. There is also an interesting curve upward in this plot past the 150th rank. This is very peculiar, but can possibly be explained by the occurrence of "underdog" players who are unfairly ranked very low but perform very well on special occasions.

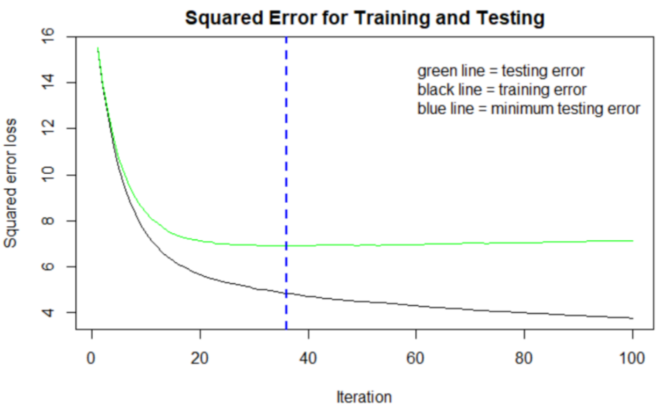


For the PDPs of high school rank and draft pick, we see the trend is not that strong compared to average NBA rank. For high school rank, the overall trend is that the higher one is ranked, the longer a player will last in the league. Interestingly, there is no clear association between draft pick and total seasons played, as seen in their PDP. There is an overall positive trend from draft picks 1-20, but after that the relationship to total seasons played spikes up and fluctuates without an apparent trend. Perhaps this is because players picked high out of college play more in the league, but for players that fall to the late first round or second round of the draft, their longevity in the league is left up to other factors besides draft pick. Players that do not get a head start in the league (players who are drafted low), must work harder to prove the critics wrong.

One of these factors is their mean wins added. In this particular PDP, there is another strong positive association between total seasons played and the NBA stat of wins added. The upward trend plateaus past wins added = 10. This might be explained by the fact that a player's wins added statistic has a maximum effect it has on a player’s total seasons played. Once a player reaches a certain level, the statistic of wins added cannot adequately convey a player’s performance level as well.

3.2.4 Prediction on Test Data

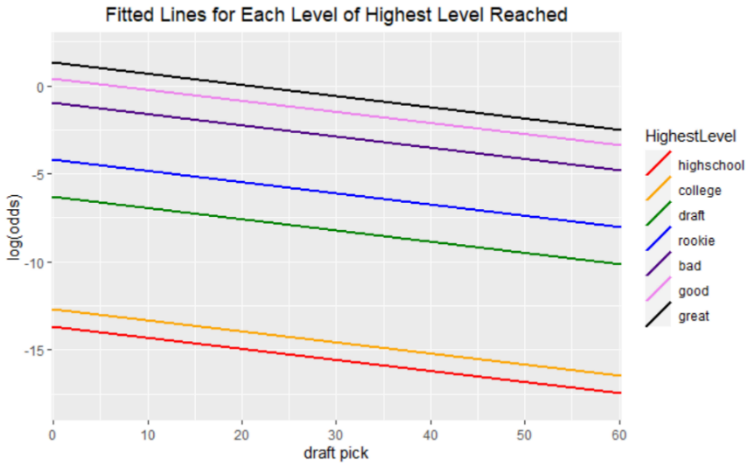
Finally, I did prediction on my held-out 20% test data set. The resulting RMSE was 2.627. This was very close to my training result, which had an RMSE of 2.624. The fact that my error stayed low and also consistent from training to testing showed that my model was not overfitting, which is desired. In the plot to the right, the training error hits a minimum and then steadies out (instead of rising again). This is further evidence that my GBM is not overfitting to the data.



3.3 Proportional Odds Model

3.3.1 Final Model

In the final model, six of the seven intercepts were statistically significant at the 95% confidence level, so overall the intercepts for the response variable were concluded as being significant. The coefficient of *high school rank* was -0.000996 and the coefficient of *draft pick* was -0.06328. *Draft pick* was highly significant at the 95% confidence level, but *high school rank* was not, having a p-value of 0.39. This was indicative that *high school rank* does not have a significant effect on modeling the odds of reaching certain levels in the NBA. This makes intuitive sense since an athlete’s high school basketball career is temporally further away from the NBA compared to his career at the collegiate level (which is the stage *right before* the NBA). Thus, a lot can change about a player in between high school and the NBA, so high school rank becomes a less reliable predictor of NBA performance. The fitted lines for each performance level modeled by draft pick (the significant variable) are shown in the figure to the right.

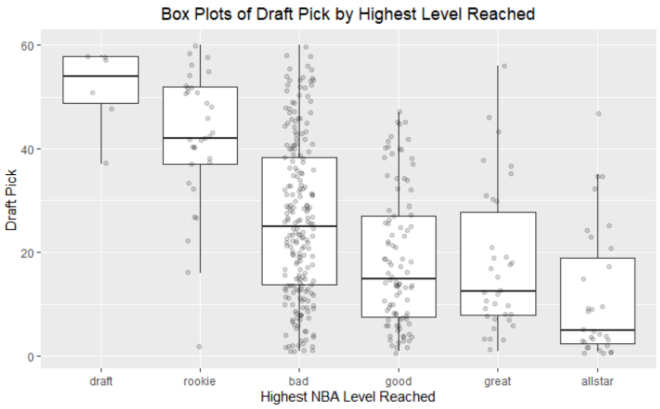


Although *high school rank* was not statistically significant, I chose to leave it in my model because it provided valuable information on the relationship between young athletes and their professional careers. Moreover, prediction (in section 3.3.4) *with* this variable included also provided key insight into my aim.

3.3.2 Goodness of Fit

In order to check the proportional odds assumption, I created box plots of highest level reached for *draft pick* (the significant variable), shown to the right. Looking at the median lines, it is reasonable to assume that the difference between any two adjacent pairs of NBA levels is the same for all combinations of levels.

Logically, it makes sense that the relationship between each pair of response levels is the same. It is safe to assume that the relationship between a “bad” player and a “good” player is the same for a “good” player and a “great” player.



3.3.3 Model Interpretation

The signs of both coefficients for *high school rank* and *draft pick* are negative, which align with our intuition about the nature of the NBA. When a player moves up to a better ranking or draft pick (numerically moves down closer to 1), he can expect his odds of making the next highest level in the NBA to increase. For our specific model, with all other factors held constant, when a player moves up one high school rank (numerically lower), his odds of reaching the next performance tier in the NBA increase by 0.09959%. Similarly, with each draft position a player moves up (numerically closer to 1), his odds of reaching the next performance level in the NBA increase by 6.13261%. This shows that an improvement in draft position boosts a player's odds of performing better in the NBA more so than an improvement in their high school rank.

Furthermore, the intercepts increase as the performance level increases, as expected. Of particular interest is the difference between the intercepts of adjacent levels. In the plot of fitted lines in section 3.3.1, the gap between lines is roughly the same except for the jump between college and draft. This shows how difficult it is for an athlete to get over the college hump. Numerous players around the country make it to the college level, but only a small fractional percentage rise the ranks high enough to be drafted into the NBA. This goes to show that getting drafted into the NBA is an extremely difficult task, but once one makes the NBA, it is a relatively steady climb to ascend to better performance tiers.

3.3.4 Prediction

Lastly, prediction was done on three cases of interest. For a player with a high school rank of 100 (100 being the lowest high school rank recorded) and a draft position of 60 (the very last pick in the draft), he has a 52.7% chance of ending up as a "bad" player and only a 0.5% chance of becoming an all-star. On the other hand, a player ranked number one in high school and also drafted number one has a 27.4% chance of being "bad" and a 20% chance of being an all-star. For someone in the middle of the pack ranked 50th in high school and picked 30th in the draft, this athlete interestingly has a 63.4% chance of ending up "bad" and a 3.7% chance of reaching all-star level.

These numbers are very revealing. For one, there is an obvious increase in one's chances of reaching higher levels given they have more stellar high school and draft rankings. Surprisingly, for the very best athlete ranked number one, although his chances of elite NBA performance increases, there is still only a 20% probability he becomes an all-star. This goes to show that being great or a superstar in the NBA is truly a daunting task, one that requires elite dedication to the game.

Even more, it appears that players ranked in the middle of the standings have higher chances of performing poorly compared to players ranked at the very back of the pack. This is a very special phenomenon, and it can be explained by the fact that players who are ranked very badly are more fueled to prove the doubters wrong and ascend the ranks. In other words, when a player is ranked at the very bottom, they have nowhere to go but up. And so, for people in the middle of the ranking, they do not have as much pressure to constantly improve so they can become complacent with their performance. As a result, with other lower-ranked players putting in work to move on up, these mediocre players in the middle of the rankings end up putting up sub-par performances.

4. Conclusion

4.1 Future Work

In order to do a deeper dive into the journey from high school to the pros, future work can revolve around statistics and rankings *within* career milestones rather than the transition between them, the latter of which was the focus of my analyses. This would require gathering a data set that has individual game and season statistics on a player's performance *during* high school and *while* he was in college. Certain well-rounded measures of performance such as offensive rating, defensive rating, and usage rate can be compared across high school, collegiate, and professional levels. This fine-combed analysis would uncover more about the specific constituent aspects of a player's game that translate to the next level.

4.2 Takeaways

In retrospective, there are many key points to note. High school rank is a *slightly* better indicator of NBA career longevity compared to draft pick. Draft position is a *much better* indicator of maximum NBA performance level compared to high school rank. Also, getting drafted into the NBA is extremely difficult, but once one makes it, it is a relatively easier climb to ascend to better performance tiers.

In general, high school and college success do indeed provide an advantage in the NBA, but nothing is guaranteed. In fact, mediocre players are at an increased risk of under-performing in the NBA. On the bright side, being at the bottom of the rankings does not imply success is out of reach.

All of these conclusions point to one theme: When it comes to sports, nothing is given; everything is earned. It is a stark reality that the majority of athletes who pursue professional careers in basketball end up falling short. However, with the right amount of undeterred perseverance, unconditional support, and a helpful amount of luck, one's dreams of making it to the pros might just become a reality.

References

1. Data Source & Data Dictionary
2. Recruiting Services Consensus Index
3. Gradient Boosting in R
4. Introduction to Gradient Boosting Algorithms
5. Ordinal Logistic Regression in R
6. Fitting and Interpreting a Proportional Odds Model
7. R Packages Used: caret, dplyr, ggplot2, ggplots, MASS, rsample, gbm, pdp, qualityTools, fitdistrplus, DescTools, corrplot, heatmaply