Using Machine Learning to Predict NBA Success

Aidan Gibbons
University of Toronto
April 4, 2019

Abstract

The problem of predicting the future performance of college athletes to decide how draft selections should be made in professional sports leagues has been an ongoing issue ever since amateur drafts have first been used. This paper uses machine learning techniques to create an ensemble prediction model to predict the future performance of NBA players drafted out of college. Using college data for players selected in the NBA draft from 1996 to 2012, the ensemble model developed consists of a weighted average of a random forest model, a lasso regression model, and a linear regression model. The order of the predictions obtained from this ensemble model is then compared to how the players were actually selected in the NBA draft. This analysis finds that there is no apparent change in the expected future performance of the players selected using the ensemble predictions compared to real draft selection order. However, the ensemble model does appear to be more risk averse, selecting a worse performing player less than 45% of the time. If NBA decision makers value risk aversion, this model could potentially serve as a better alternative for how college basketball players are selected in the draft.

1 Introduction

In professional sports leagues, if you have a superstar player, you will typically do everything you can to make sure they stay on your team. Therefore, acquiring such top end talent is difficult to do through free agency or trades as these kinds of players are rarely available. The best way (and for some teams, the only way) to acquire these franchise changing players is through the draft. While there is such an importance on the draft, it remains a difficult selection process. In the National Basketball Association (NBA), the majority of players selected in the draft were previously playing basketball in college. Teams are essentially tasked with predicting how these college athletes will perform in the professional league through their 20s and 30s. The difficulty in these selection decisions are clear with the frequency of highly drafted players underperforming based on expectations as well as players picked later in the draft developing into highly coveted superstars. While this problem has been an ongoing issue since the draft process was first used, there have been major strides in predictive models such as machine learning techniques. This begs the question, can machine learning techniques be used to predict NBA player performance using college statistics and can these predictions improve how NBA teams select college prospects in the NBA draft?

There have been numerous papers looking into the prediction problem that accompanies deciding how to select players in the draft for professional sports leagues. Many papers have discovered a potential inefficiency as the variables that best predict a the future performance of a player are not the same as those that best predict where a player gets selected in the draft. While this result has been found for different leagues and using different measures of success, the predictions have not used machine learning techniques and more importantly, these papers have not developed an alternative method of selecting players that would do a better job compared to how professional sports teams actually select prospects. Developing such an alternative method of draft selections is the main focus of this paper.

For this analysis, data on NBA draft order from 1996-2012 was taken from Basketball-Reference.com as well as data on NBA statistics, age, and where a player played college basketball (if at all). Then from Sports-Reference.com, college statistics (including advanced statistics measures) of a player as well as their listed height, weight, and position in college was collected. For a measure of NBA success, Peak Win Shares is calculated as a per game average of a player's Win Shares from the seasons when they were age 24, 25, and 26. This average is then extrapolated over a standard 82 game season to obtain Peak Win Shares, which is used as the main variable of interest in this paper.

This paper employs an ensemble method for prediction. This includes a weighted average of a random forest model, a lasso regression model, and a linear regression model. The first two of these are empirically tuned individually for optimal prediction, then the weighted averages are similarly adjusted to achieve the final ensemble prediction model. Prediction performance is initially measured as minimiz-

ing the root mean square error of the predictions, as has been done in similar papers such as Mulholland & Jensen (2014). Finally, these predictions are then sorted and used as alternative order for selections in the NBA draft to see if the predictions can improve upon how the players were actually selected.

While the predictions made by the ensemble model do not definitively improve the order in which players are selected in terms of future performance, the model does seem to be fairly risk averse, overall selecting a worse player less than 45% of the time, and doing so a majority of the time on only a quarter of selections.

A key limitation in this analysis is the assumption that NBA decision makers are primarily intending to choose players with the highest Peak Win Shares. While this may not be the case, further analysis into this issue reveals that draft selection best predicts the future performance measure of Peak Win Shares, more so than looking at a certain percentile of players. This suggests that either NBA selections are made with poor accuracy with the intention to acquire an elite talent more than any other factor, or selection decisions primarily align with the Peak Win Shares measure.

Despite the discovered difference between what best predicts future performance and what best predicts draft selection in the NBA, Berri, Brook, and Fenn (2011) like many others do not lay out an alternative method for determining draft selection order that would improve upon this apparent inefficiency. Therefore, building on the findings of previous works that the draft selection process in sports is something that could be improved upon, this paper goes about developing a model that could potentially achieve this.

2 Background

Ichniowski and Preston (2017) look at performance in the NCAA "March Madness" college basketball tournament, over the span of 1997-2010 to see how it affects teams' draft decisions. Their results find that unexpected performances in this tournament (both teams wins and player scoring) does affect draft decisions. Having a player's team win one more game than expected and that player scoring an additional 4 points more than expected results in a 4.7 spot improvement in their draft selection slot. Interestingly, this small sample of high media attention games (relative to the collegiate season as a whole) is not overweighed by NBA personnel. In fact, their evidence suggests that unexpected performance in the March Madness tournament should warrant more weight it appears to be predictive of professional success in the NBA, in the player's first year in the league and over their entire career, both in the regular season and in the playoffs. This could potentially be a result of the media attention the tournament receives as well as the lose-and-go-home format of March Madness providing similar pressures to that of NBA games.

Berri and Simmons (2011) find that in the NFL, quarterbacks are drafted on variables such as height, intelligence (as measured by the Wonderlic test), and speed (40 yard dash), but there is no observed evidence that these factors impact their future NFL performance. Overall, where a quarterback is selected in the draft was found to be not a significant predictor of professional performance in the NFL.

In a similar study, Mulholland and Jensen (2014) employ different models to predict both draft selection results and National Football League (NFL) success based on pre-draft data on college football tight ends (one particular position in football). They find that the variables used in prediction of draft selection are different than those that best predict professional performance in the NFL. These predictions are done using both decision tree models and linear regression models. The pre-draft data includes college statistics, combine performance (ex. 40 yard dash time, bench press, broad jump, etc.), and physical measures (height, weight, BMI). Their findings suggest that NFL teams tend to favour strength (as measured by bench press results), whereas NFL performance was best predicted by athletic explosiveness (as measured by the broad jump). Variables such as forty yard dash times were predictive of both NFL career performance and where the players were selected in the draft.

The most similar literature to this paper is work by Berri, Brook, and Fenn (2011) in which they look at players selected in the NBA draft who were previously playing in the National Collegiate Athletic Association (NCAA)[1]. The first part of this paper uses three different models to predict where a player gets selected in the NBA draft using their college statistics, position, conference they played in, height, and age. The three different models used for this prediction are OLS, Poisson, and a negative binomial model. No variable was significant at the 5% level across all 3 models, however points, assists, blocks, personal fouls, 2 point field goal percentage, relative height, appearing in the Final Four (the semifinals of the NCAA playoff tournament), playing on an NCAA championship team, age, many conference indicator variables, and if a player played shooting guard were all significant at least at the 10% level across all three of these models. Next, the authors use a model using the same explanatory variables, but now using a measure of NBA performance as the dependant variable. They model NBA success for a players 2nd, 3rd, 4th, and 5th year in the league separately[2]. It is found that mostly different variables are significant in these models compared to the models for where a player was selected. They also find that draft position alone explains only a small portion of career performance (between 1% and 7%). The authors remark that this discontinuity between what best predicts draft selection and what

[1]Players who were drafted out of high school or from other countries outside of the United States are excluded. This gives a more consistent data set and almost 80% of the players drafted over the span of their analysis (1995-2009) had previous experience in the NCAA.

[2]It is worth noting that this analysis only includes players who logged an average of 500 minutes per season. As a result, this model only looks at those who played a significant amount of time in the NBA, excluding those who either never ended up playing in the NBA or those who were out of the league after their first few years. In addition, the models for this analysis have an adjusted R^2 measure ranging from 0.30 to 0.36. For these reasons, the problem of predicting NBA performance prior to the draft would seem to be more difficult that the conclusion of this paper suggests.

best predicts NBA performance suggests that NBA decision makers are not selecting college prospects optimally. Variables such as points scored in college and if a player made it to the Final Four seem to be given too much emphasis for determining where to select a player in the draft. The authors state that "the problem is not that data is unavailable or that performance is difficult to predict", but rather it is this apparent overemphasis on the wrong factors.

These previous works illustrate that professional sports teams may not be making optimal draft decisions since the variables that best predict draft selection are not the same as those which predict future professional success once drafted into the league. One of the differences in this paper compared to this previous literature is that it uses machine learning for the predictions, while the papers mentioned above do not[3]. This paper will also attempt to take this one step further and develop an alternative selection process which could be used to potentially make better decisions on draft day. Similar to the procedure used by Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan (2017) where a prediction model is developed using machine learning, then the predictions are used to show if and how they can improve decision making.

3 Data

For the data used in this analysis, the first step was gathering info on player selection in the NBA draft. In this case, for the draft years from 1996-2012[4]. This data was obtained from Basketball-Reference.com and shows not only the players selected, but also the college they were drafted from if they played collegiate basketball in the National Collegiate Athletic Association (NCAA). This analysis only looks at players who did play college basketball, excluding those who either entered the draft straight out of high school or played in a league outside the United States. For these former college players, their age as well as their games played and win shares in their 24, 25, and 26 year old seasons were collected.

3.1 Peak Win Shares

Win shares was originally developed by Bill James (one of the pioneers of sports analytics) for baseball as a measure of attributing team wins to players. Basketball-Reference.com has developed a similar measure for basketball. This win shares measure roughly splits up a team's wins for a season among the players of that team. For example, if a team has 50 wins in a season, the sum of the players win

[3]Mulholland & Jensen (2014) do use decision trees, but there is no element of empirical tuning. Only single trees with one set of parameters are used.

[4]The reasoning for this span is twofold. First, any more recent draft years would include players who have not yet played most of their "peak" seasons. Second, college data prior to 1996 does not include win shares data. This measure was found to have significantly high predictive power, so to keep this variable in the data set, draft years before 1996 were not used. Without win shares or defensive win shares measures, there lacks any reliable measure of the defensive performance of a player.

shares would be approximately 50. This measure can be negative (if a player's performance is so poor that they are essentially taking away wins from their team) and is made up of defensive and offensive win shares. As a result, it quantifies the impact of a player's performance in both the offensive and defensive aspects of the game. To get the measure of Peak Win Shares that is used throughout this paper as the variable to be predicted, the sum of win shares accumulated by a player in the regular seasons (so not including playoffs) in which they were age 24-26 is taken[5], then divided by the total number of regular season games played over that span to obtain an average win shares per game measure. This is then multiplied by 82 (the number of games in a standard regular season) to arrive at the Peak Win Shares measure. Extrapolating over an 82 game season allows for the measure to more easily be contextualised. For example, the player with the highest Peak Win Shares in the data set is Kevin Durant with 18.61. This means that over the "peak" of his career, Durant contributed an average of 18.61 wins to his team each year. Peak Win Shares serves as the measure of NBA performance in this analysis. The rationale being that how well a player can perform during the "peak" of their careers may be the most important factor to teams in deciding which player to select during the draft. By looking at average win shares per game over a range of 3 years, this minimizes the impact of an injury that may cost a player some or most of the season. Additionally, it should account for if a player has an unusually good or unusually poor season compared to their standards and give a more representative measure of their performance during this period of their career compared to just considering a single year.

To determine if this Peak Win Shares measure if accurate in quantifying player success, we can look at how it compares to other traditional metrics for success in the NBA. Figures 1 and 2 orders all the players in the data set by highest Peak Win Shares to lowest, then looks at the accumulation of NBA accolades such as number of times a player is an All Star, the number of times a player is named to the the All-NBA team, and the number of times a player leads the league in one of the 5 major statistical categories (points, rebounds, assists, steals, blocks). As these graphs show, Peak Win Shares does a good job overall in awarding players with these accolades larger Peak Win Shares than others. It is worth noting the nature of the two most significant outliers in Figure 1. The furthest big spike at bin 110 is caused by Steve Nash who could be considered a late bloomer. Nash didn't make the first of his 8 All Star appearances until age 27 and proceeded to win back to back MVP awards at age 30 and 31. On the other hand, the player who is the furthest outlier (in bin 380) is Derek Rose who became the youngest MVP in league history at age 22. However, a major injury caused Rose to miss his entire 24 year old season and he then only managed to play 10 games in his 25 year old season. Even the year after, Rose did not perform to the same level as he had before his injury. These two outliers highlight a limitation of the Peak Win Shares measure, namely that it only covers a 3 year span of a player's career, and may not capture the whole story of a player. However, these cases are both extremes and

[5]Data collection took place February-March of 2019, so some of the players statistics for the 2019 season (if this was part of their age 24, 25, or 26 year old seasons) does not include data for the whole season, but for the majority of the season.

increasing the span of the measure to try and capture this would create more issues and also further limit the size of the data set. Overall, Peak Win Shares does seem to be able to properly capture the performance of a player during the peak of their career.

3.2 Independent Variables

The variables used for prediction are mostly taken from Sports-Reference.com. From this website, the college statistics of a player are obtained for the previous season right before they were drafted. These statistics include both traditional statistics (games played, baskets (made, attempted and percentage broken down by 2s, 3s, and free throws), rebounds, assists, steals, blocks, points, strength of schedule) and advanced statistic measures (true shooting percentage, effective field goal percentage, 3 point attempt rate, free throw attempt rate, win shares (offensive, defensive, and total)). There is also the listed height, weight (and from this a height/weight ratio), and position. Variables such as weight and position can change over the course of their professional careers, so by using these measures as they were listed in college means that the data used is the same as what would have been available to NBA decision makers at the time of the draft. Lastly, the age (which was obtained from Basketball-Reference.com) is calculated as the player's age in June of the year in which they were drafted, so their age at the time of them being drafted since the NBA draft typically takes place in June each year. After removing some players from the data set due to missing data, this leaves data on 683 players selected in the 17 drafts analysed. To put this into context, over this 17 year span, exactly 1000 players were selected in the NBA draft. Of those 1000, 782 of them played college basketball the season before being drafted. This means that the data set includes 68.3% of all players drafted over this span, and 87.3% of those drafted from the NCAA. The descriptive statistics for all variables used can be seen in Figure 3.

4 Empirical Strategy

First, the ensemble method was developed to predict Peak Win Shares using the variables described in the previous section. This ensemble method consists of a weighted average of 3 different prediction models: random forest, lasso regression, and linear regression. In each of these cases, the models learn on the training set data, then their performance is measured using the test set data.

4.1 Splitting the Data

The data is first divided, with the data for the draft years 2011 and 2012 set aside as the "lock box" set, similar to Kleinberg et al. (2017), where this set will be used following the empirical tuning of the final ensemble model as an additional robustness test of the results. Next, the remaining data is split into training and test sets. Due to the relatively small sample size (compared to what would be ideal for machine learning analysis), and part of the evaluation of prediction needing to be for specific years

(as it is impossible to compare where a player was drafted in one draft class compared to how another player was drafted in a different draft class), the results vary significantly depending on how the data is split up. To account for this, years of the data are randomly selected as test sets until at least 20% of the data is selected[6]. This division of data is then used for training and testing the prediction models. Then there is another random draw for splitting the data again. This process is repeated 100 times in each analysis[7]. This repetition was intended to make the results more robust rather than a product of a particular random split of the data.

4.2 Random Forest Model

The first prediction model developed was a random forest. This was implemented using the scikit-learn package for Python. Each time the model was run, 1000 trees were created for each random forest[8]. Empirical tuning for this model as well as all following models (other than linear regression for which empirical tuning is not required) was measured using the root mean square error (RMSE) of the predictions for the test sets. This RMSE measure for prediction performance is similarly used by Mulholland & Jensen (2014). For the empirical tuning to maximize the prediction power of the model, three main parameters were tuned. First is the number of maximum features. This determines how many variables the algorithm goes through before deciding which to use to make a split. It was found that setting this parameter to the maximum (so that all variables are considered before a split is made) always obtained the best measure for prediction (the lowest RMSE). The next parameter tuned was the max depth of the trees. This determines the maximum number of splits a tree can have along a given branch. As Figure 4 shows, the RMSE is the lowest at max depths of 4, 5, and 6. Given this result, these three levels of max depth were then further tuned at various levels of the final tuning parameter: minimum split size. The minimum split size determines how large a sample must be at a node for a split to occur. The results of varying this parameter for the 3 different levels of max depth is shown in Figure 5. This shows that the optimal parameter set for the random forest model is setting maximum features to the maximum, setting maximum depth to 6, and setting minimum split size to 25. An example of such a decision tree achieved using these parameters is shown in Figure 6. This creates trees which are not overly simplistic as to be poor at prediction, but also not overly complex as to overfit the training data. While it is difficult to draw conclusions from the variable used in prediction, this random forest model determines that the variables Age and Win Shares have a significantly higher "importance" than the others. This importance measure is correlated with the probability of reaching a node where the variable is used as the split decision. Therefore, Age and Win Shares are often one of the first splits in

[6]On average, the test set size was about 24% of the data, and was never more than 27%.

[7]This was coded so that the same random seed is used at the beginning of running the analysis. This ensures that each time the analysis is repeated, the same 100 splits are performed for consistency and repeatability of the results.

[8]For each data split, the random forest model was run 5 times. This is because the inherent nature of the random forest randomly developing the trees results in different trees populating the forest each time, even for the same training set data. By performing 5 runs each time and taking averages of the results, this theoretically reduces the random noise that comes from developing the random forest for a given training set

the trees. After these, the other variables that have the next highest importance for this model are Steals Per Game, Defensive Win Shares, Offensive Win Shares, Strength of Schedule, and Games Played, in that order.

4.3 Lasso Regression Model

The next model used is a lasso regression. This model is similar to OLS, but with a penalty term for the number of variables used (refer to Figure 7). The weight of this penalty term, α, is empirically tuned. Due to the randomness of the draws for the training set and test set, Figure 8 shows the results of this empirical tuning, averaged over 5 runs of the model to get a better idea of the optimal value for the α parameter. At large values for alpha, the prediction gets considerably worse. This is due to the high weight on the penalty term resulting in too many variables being removed from the model. At very small values for α, the predictive power is good, but the variance seems to be larger. Additionally, with smaller and smaller values of α, the model starts to resemble a linear regression model. Since a linear regression is already present a component of the ensemble model, an α of 0.13 was selected as the optimal. It could be argued that this value could be lower to slightly improve predictive performance, but because of the higher variance and similarity to a linear regression model at very small values for α, this value was chosen. To give a sense of what variables remain in the model given this level of α, refer to Figure 9 to see the results from 5 runs of the model. As was the case in the random forest model, the variables Age, Win Shares, Steals Per Game, Strength of Schedule, and Games Played all appear to have high predictive power for Peak Win Shares. The variables for Year and Weight appear each time even though these had low importance in the random forest model. The variables Defensive Win Shares and Offensive Win Shares do not appear to show up in this model, despite being in the top 5 for importance in the random forest model. In terms of the direction of effect of these variables on the prediction of Peak Win Shares, two variables stand out in particular as different than one might expect. Games Played and 2-Point Field Goal Attempts Per Game both have a negative impact on the prediction of Peak Win Shares. A reason for this may be efficiency. Since Win Shares seems to have high predictive power and is an accumulative measure (compared to say, if it were Win Shares Per Game), players who can accumulate the same number of Win Shares in less games or in less shots would be considered more efficient players.

4.4 Linear Regression Model

The third and final component for the ensemble model is a linear regression model. This model is a standard OLS regression run with all of the variables included. As a result, none of the obtained beta values are relevant for interpretation since many of the variables are components of others and as a result, there will be significant multi-collinearity. However, this model is simply used as a component for prediction, not for interpretation of the beta values obtained.

4.5 Ensemble Model

With the three models empirically tuned (with the exception of the linear regression model which requires no such tuning), the ensemble method is obtained as a weighted average of these models. To figure out the optimal weights for each model, a number of different combinations were tested. These results can be seen in Figure 10. The weight for the linear regression model was held fixed at 1 with the weights for the other two models adjusted to determine which combination of weights minimized the RMSE. It is also worth noting that taking pairs of the models was also analysed, but a combination of all 3 models outperformed any pair. As Figure 10 shows, the optimal weights for the 3 models was found to be a weight of 2 for the random forest model, 3 for the lasso regression, with a weight of 1 for the linear regression model. This means that the optimal ensemble model takes a weighted average of 1/3 from the random forest, 1/2 from the lasso regression, and 1/6 from the linear regression model. Given the optimal parameters for each model, the prediction performance of each model, as described by the R^2 value, is presented in Figure 11. As these results show, the ensemble model does better than any other model individually. The reason for this may be a result of the data as well as how each model performs predictions. The random forest model is good at capturing integrations between the variables through branches of the trees. However, about one third of the players in the data set have a Peak Win Shares value of 0, and the nature of the random forest predictions makes it so that the model rarely ever predicts a value of zero, and likely never predicts a negative value. On the other hand, the lasso and linear regression models do not contain any interaction terms, but they can (and often do) predict negative values and values close to zero. Therefore, the ensemble method seems to be benefiting from this combination of different models to make up for the ways that each model performs poorly individually.

5 Results

With the ensemble model developed and empirically tuned, we now arrive at the question of how well these predictions can select the order the players. This is to see if by selecting players based on the predictions made by the ensemble model, are these selections better than how the players were drafted in reality. The first way to analyse this is to look at the difference between the average Peak Win Shares of the players selected in reality and the average Peak Win Shares of the players chosen by the ensemble model, the results to which are shown in Figure 12. As these results show, the model does not appear to significantly perform better in terms of Peak Win Shares of draft selections over the span, but it also does not appear to perform significantly worse either. At some spots the ensemble model makes better selections, at others it makes worse. However, most of the selections are within 1 Peak Win Share of the actual outcomes.

Another way to look at how the ensemble model selects players is the frequency with which the

model selects a worse player. These results are shown in Figure 13. As this figure shows, for most of the selections, the model doesn't select a worse player a majority of the time. This means that more often than not, the ensemble model is selecting a player with a Peak Win Shares value higher or the same as the player who was actually selected. Only for about a quarter of the selections does the ensemble model select a worse player with a frequency of over 50%. Overall, the model makes a worse selection 44.8% of the time. These results are interesting because when coupled with the previous figure, it appears that while the model does not seem to significantly alter the expected value of the players selected, it does seem that the model is more risk averse in the selections made. Therefore, the ensemble model appears that it could be a better alternative for draft selection choices if NBA decision makers are wanting to make less risky selections.

Finally, the lock box data is analysed using the same ensemble model as an additional check to determine the performance of the predictions out of sample and consistency of the results. Figure 14 shows similarly mixed results that appear to be not significantly better or worse compared to the Peak Win Shares of how players were actually selected. For the frequency of worse selections, Figure 15 also appears to be close to the previous results. Again, we observe a similar proportion of one quarter of the selections resulting in a worse player being chosen a majority of the time. With this data, the model only makes a worse selection 38.5% of the time. Both Figures 14 and 15 show that the analysis on the lock box set has higher variance, but this is to be expected since this set has a smaller sample size. However, it does appear that the main results hold, that being while the ensemble model does not appear to significantly make any better or worse selections overall, the model does seem to offer a more risk averse alternative for how to select players in the NBA draft.

6 Robustness

The crucial assumption for this analysis is that teams are trying to draft players based on who will have the best career performance during the peak of their career. The issue here is that we cannot observe the exact factors that NBA decision makers are considering when making their selections of these college prospects. It could be that rather than selecting the player with the highest expected performance, teams instead are trying to draft franchise players, the best of the best. Selections may be based on attempting to acquire these high-end talented players, even if it means making a potentially riskier selection. To attempt to better understand how teams are trying to select players, a simple regression model is run to see which measures of player performance are best explained by where they are selected in the draft. Figure 16 shows the results for these tests. When the entire dataset is considered, Log Peak Win Shares is best explained by draft position (22.7%), with Peak Win Shares slightly less (21.8%). A player being above the median in Peak Win Shares, as well as players in the 75th, 90th, 95th, and 99th percentile are all explained less by draft selection, with the explanatory power decreasing

for higher percentiles (16.1%, 15.3%, 9.4%, 5.7%, and 1.8% respectively). This seems to indicate one of two possibilities. Either NBA decision makers are selecting players primarily based on trying to select the player with the highest Peak Win Shares (log or linear) or they are trying to select these top tier players, but are not accurate in doing so. If the sample is restricted to just looking at players who have a Peak Win Shares measure above 0 (similar to the restriction done by Berri, Brook, and Fenn(2011) when they look at the explanatory power of draft selection on future performance), Peak Win Shares is best explained by draft position (13.0%). In this restricted sample, a player being in the 75th percentile of Peak Win Shares is now better explained by draft position (9.7%) than Log Peak Win Shares (9.2%). Overall, it appears that draft position best explains the variation in Peak Win Shares compared to other measures of future player performance.

These results seem to suggest that while Peak Win Shares may not be the primary factor used in determining where a player is drafted, but when compared to other potential factors, it is the variable best explained by draft position. If NBA decision makers are just trying to select top-tier talent above all else, draft history reflects that they are not accurate in trying to select these types of players. Rather, the draft history would seem to suggest that how a player performs during the peak of their career is most in line with how NBA teams select college prospects.

7 Conclusion

Overall, the ensemble model is able to predict future performance of college basketball prospects more accurately than any of the three individual components alone. However, even with this level of predictive power, the ensemble model does not clearly improve the Peak Win Shares of the selections made when comparing drafting using the model's predictions to how the players were drafted in reality. While the model does not significantly improve the quality of those selected overall, it also does not seem to perform significantly worse either. The key finding is that while the model does not appear to alter the expected value of the player selected, it does tend to make less risky picks. The ensemble model makes a worse selection than reality less than 45% of the time, and only at about one quarter of the draft selections does the model make a worse pick more than 50% of the time. Analysis on the "lock box" set obtains similar findings, demonstrating the prediction ability of the model out of sample as well as the overall consistency of the findings.

While the literature shows that the variables that best predict draft order are not the same as those that best predict future performance, this does not necessarily mean that developing an alterative method for selecting college athletes is straightforward. In previous papers, as well as this one, prediction models alone are easy enough to develop. However, there is still much of the variation in player performance that cannot be explained by using data available before a player is drafted. While the

ensemble model developed in this paper does not clearly show improvement in the performance levels of players selected, it also does not appear to be significantly worse either. Looking at the selections by a different metric, we can see that the ensemble model's predictions appear to be less risk averse. If decision makers in the NBA wish to make less risky draft selections, this model could potentially be beneficial in helping to achieve that goal.

Future research similar to this paper could use college data along with data from the first few years of a player's career to predict performance in the later seasons. These predictions, if accurate, could then be used to aide in asset management decisions such as contract offers and trade deals. Further extensions of this paper could add more variables such as more advanced statistics measures (such as usage rate), combine data (for vertical leap and shuttle run measures), could break up college performance into regular season and playoff (NCAA March Madness) statistics, include mock draft data, and could also attempt to quantify qualitative characteristics of a player such as leadership, work ethic, shot mechanics, etc. by seeing if these qualities are mentioned in scouting reports. Perhaps some of these variables are highly predictive of future NBA performance and could be used to improve the prediction power of the model to that point that it unambiguously improves the NBA draft selection process, both in terms of expected value and risk aversion.

While we are unable to determine exactly all of the factors that NBA teams consider when making a draft selection, selecting players based on expected Peak Win Shares (intended or otherwise) seems to be the outcome that occurs as draft position best explains this measure for future performance more than indicators of a top-tier, superstar type player. It is likely that a multitude of different factors are considered before drafting a player such as a team's positional needs, marketability, as well as how well a player will perform during the peak of their career. Narrowing down all these potential factors for how teams draft into a single measure is likely impossible, but the ensemble model developed in this paper does seem to be able to offer improvement in one such factor, risk.

References

Berri, D., Brook, S., & Fenn, A. (2011). From college to the pros: Predicting the NBA amateur player draft. *Journal of Productivity Analysis, 35*, 25-35.

Berri, D., & Simmons, R. (2011). Catching a draft: On the process of selecting quarterbacks in the National Football League amateur draft. *Journal of Productivity Analysis, 35*, 37-49.

Ichniowski, C., & Preston, A. (2017). Does March Madness lead to irrational exuberance in the NBA draft? High-value employee selection decisions and decision-making bias. *Journal of Economic Behaviour & Organization, 142*, 105-119.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human Decisions and Machine Predictions. NBER Working Paper 23180. National Bureau of Economic Research.

Mulholland, J., & Jensen, S. (2014). Predicting the draft and career success of tight ends in the National Football League. *Journal of Qualitative Analysis in Sports, 10*(4), 381-396.

Appendix

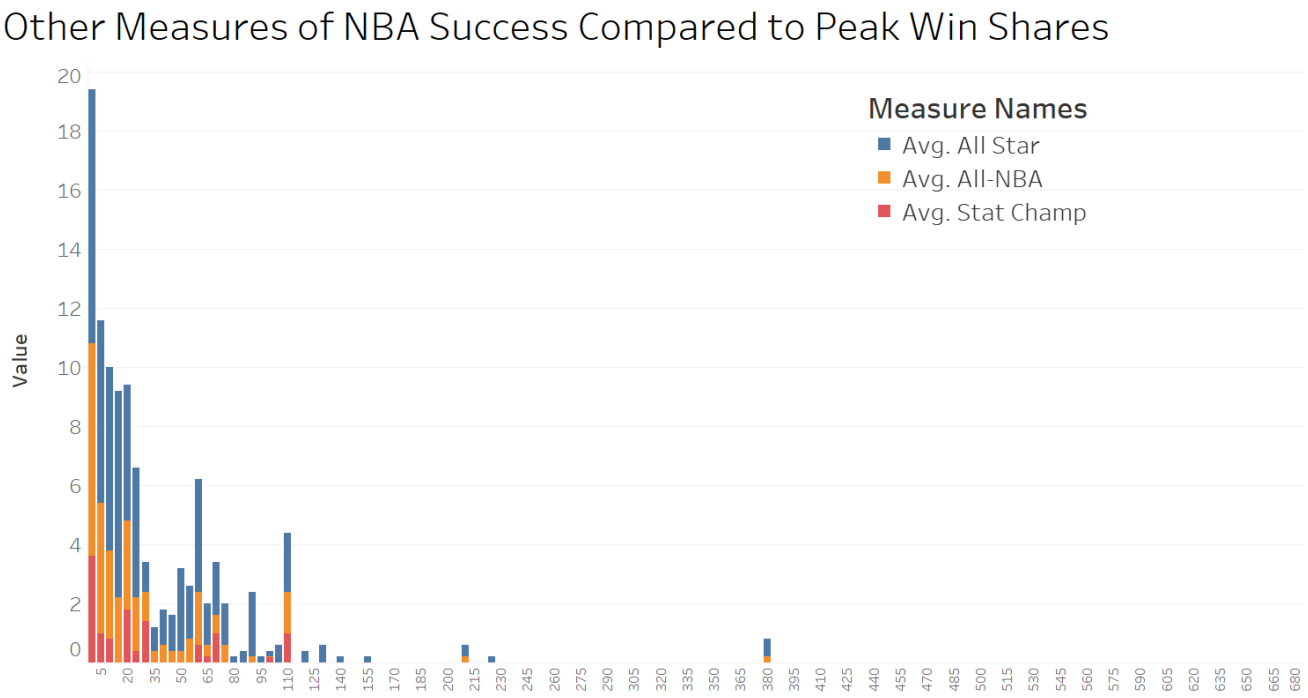

Figure 1: Ordering all the players in the data set by descending Peak Win Shares and looking at their accumulation of other measures of NBA success (average for each bin of 5)

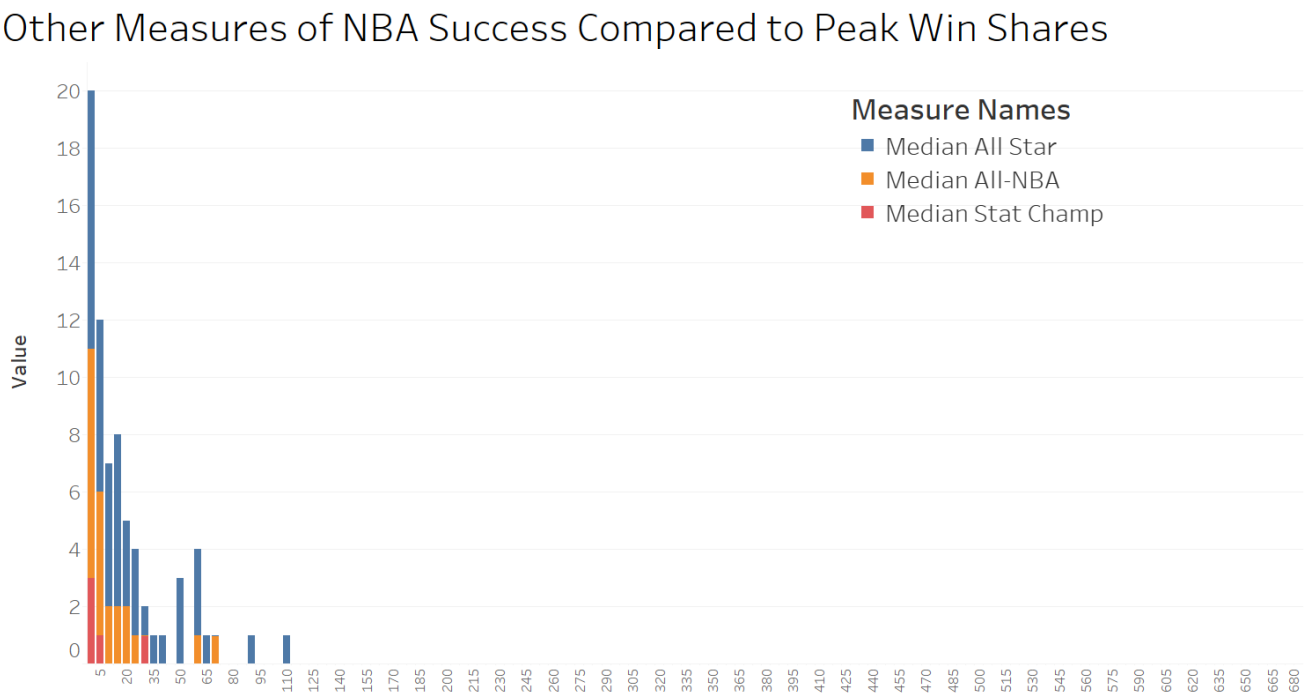

Figure 2: Ordering all the players in the data set by descending Peak Win Shares and looking at their accumulation of other measures of NBA success (median for each bin of 5)

Variables	Mean	SD	Min	Max
Dependent Variable				
Peak Win Shares	2.86	3.30	-1.94	18.61
Independent Variables				
Draft Information				
Draft Year	2004.30	5.09	1996	2012
Pick	26.96	16.39	1	60
Player Characteristics (at time of draft)				
Age	21.74	1.26	18.75	25.17
Height (cm)	200.34	8.65	173	221
Weight (lbs)	220.11	26.70	160	300
Height/Weight Ratio	0.92	0.08	0.71	1.16
Position: Center (indicator variable)	0.18	0.39	0	1
Position: Forward (indicator variable)	0.43	0.49	0	1
Position: Guard (indicator variable)	0.39	0.49	0	1
College Statistics (in final year)				
Games Played	32.53	4.47	10	41
Field Goals Per Game	5.73	1.43	1.30	11.50
Field Goal Attempts Per Game	11.73	3.13	2.40	22.10
Field Goal Percentage	0.50	0.06	0.29	0.69
2-Point Field Goals Per Game	4.67	1.46	0.80	9.90
2-Point Field Goal Attempts Per Game	8.84	2.59	1.60	17.30
2-Point Field Goal Percentage	0.53	0.05	0.35	0.73
3-Point Field Goals Per Game	1.07	0.95	0.00	3.90
3-Point Field Goal Attempts Per Game	2.90	2.42	0.00	9.90
3-Point Field Goal Percentage	0.25	0.17	0.00	0.57
Free Throws Per Game	3.67	1.33	0.70	7.80
Free Throw Attempts Per Game	5.05	1.67	1.40	11.20
Free Throw Percentage	0.72	0.09	0.30	0.97
Total Rebounds Per Game	6.57	2.51	1.60	14.70
Assists Per Game	2.54	1.68	0.20	9.80
Steals Per Game	1.25	0.58	0.20	3.60
Blocks Per Game	1.03	1.00	0.00	6.40
Points Per Game	16.21	4.17	3.40	31.70
Strength of Schedule	6.74	3.45	-11.83	12.71
True Shooting Percentage	0.57	0.04	0.39	0.72
Effective Field Goal Percentage	0.54	0.05	0.30	0.71
3 Point Attempt Rate	0.23	0.18	0.00	0.81
Free Throw Attempt Rate	0.45	0.15	0.16	1.19
Offensive Win Shares	2.94	1.31	-0.70	7.30
Defensive Win Shares	2.34	1.05	0.20	6.70
Win Shares	5.29	1.74	0.10	11.30

Observations: 683
Notes: If a player played less than 500 minutes total in their age 24, 25, and 26 year old seasons, their Peak Win Shares measure was set to 0. Only 4 players were drafted too old to have a 24 year old season. If a player had less than 1.00 3-Point Field Goal Attempts Per Game, their 3-Point Field Goal Percentage was set to 0.
Sources: Basketball-Reference.com and Sports-Reference.com

Figure 3: Descriptive statistics for the dependent and independent variables

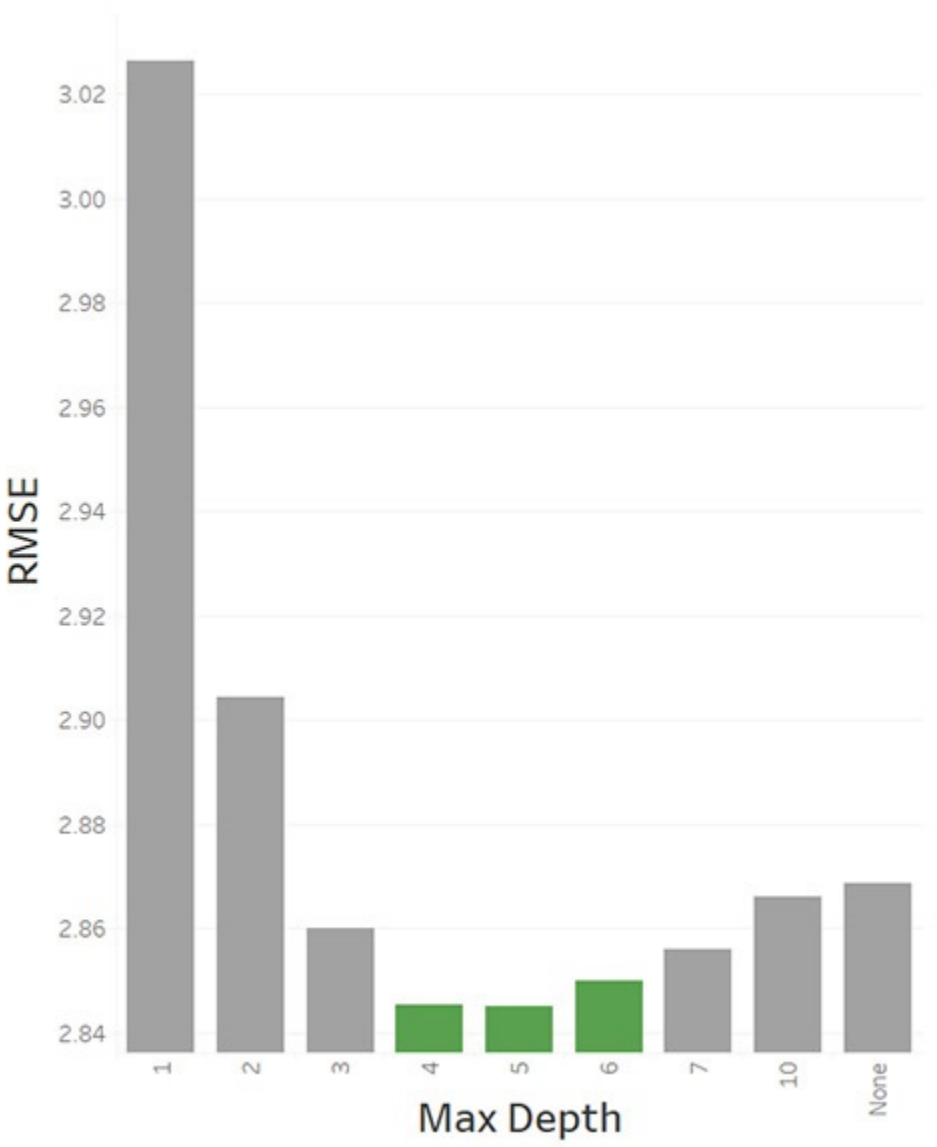

Figure 4: Results of empirical tuning the maximum depth parameter for the random forest model

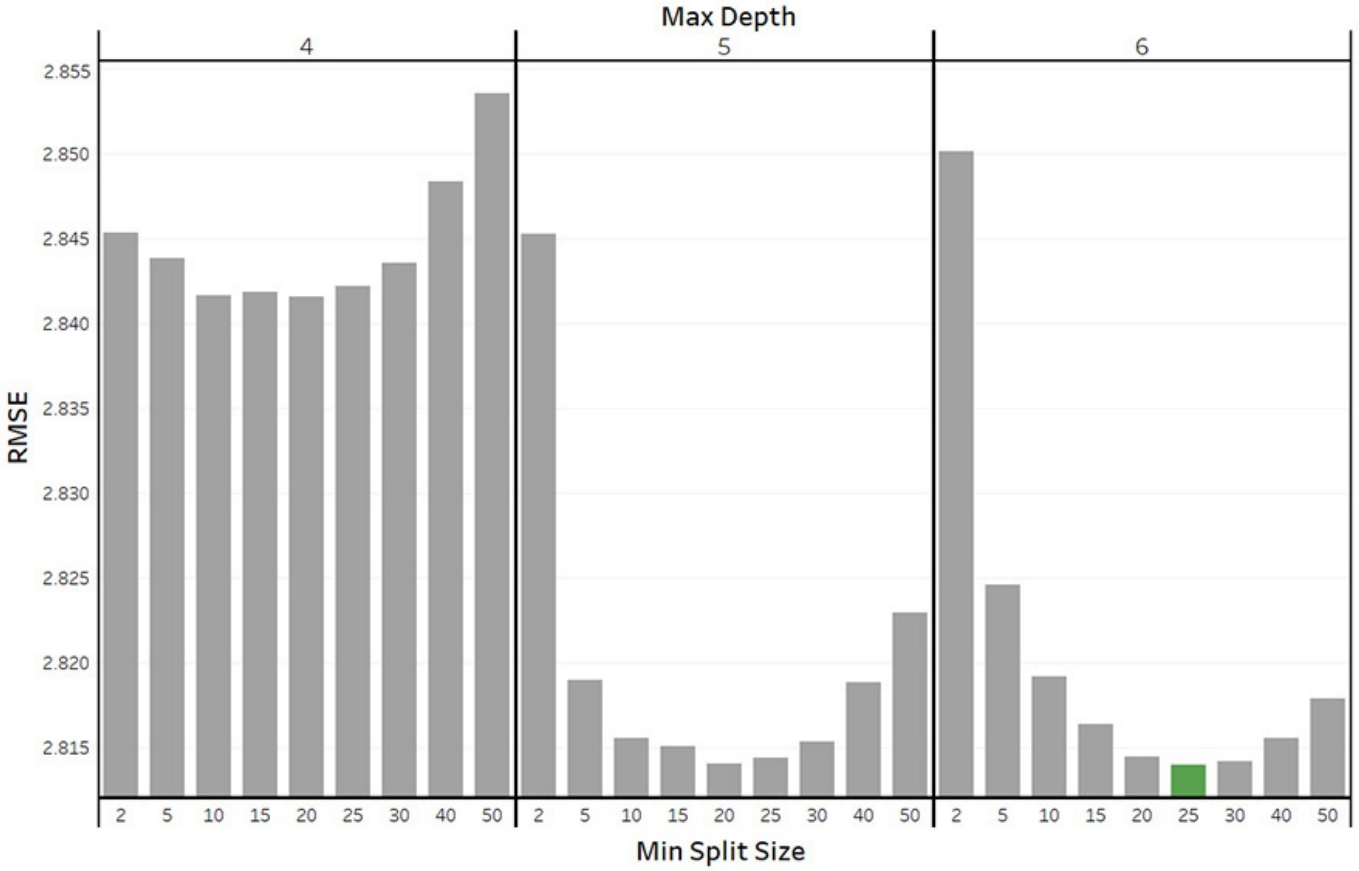

Figure 5: Results of empirical tuning the minimum split size parameter for the random forest model at the three best max depths

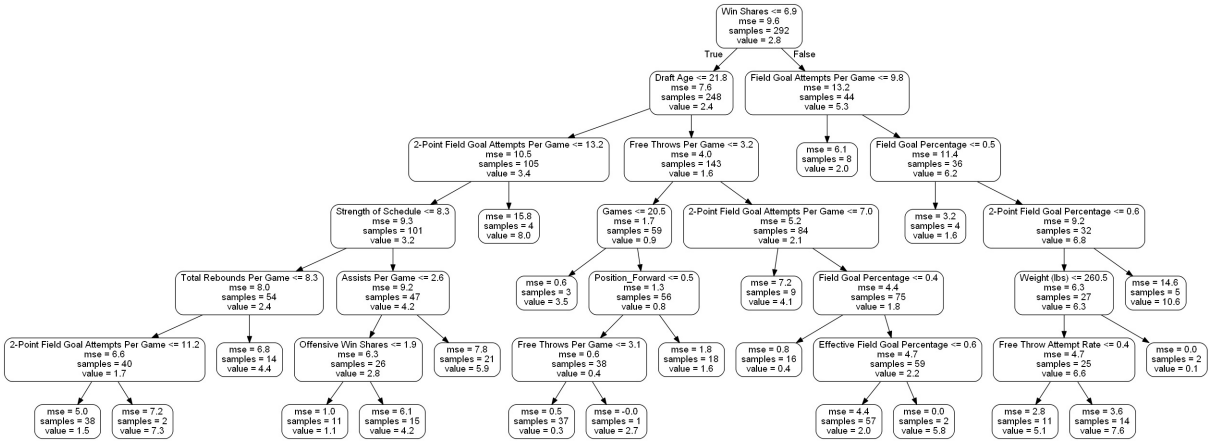

Figure 6: One of the trees created using the optimal random forest parameters: maximum features = maximum, max depth = 6, minimum split size = 25

$$min_\beta L(\beta) = E[Y - X\beta]^2 + \alpha \left(\sum_{j=1}^{k} \beta_j \right)$$

Figure 7: Lasso regression equation where α is the weight of the penalty term

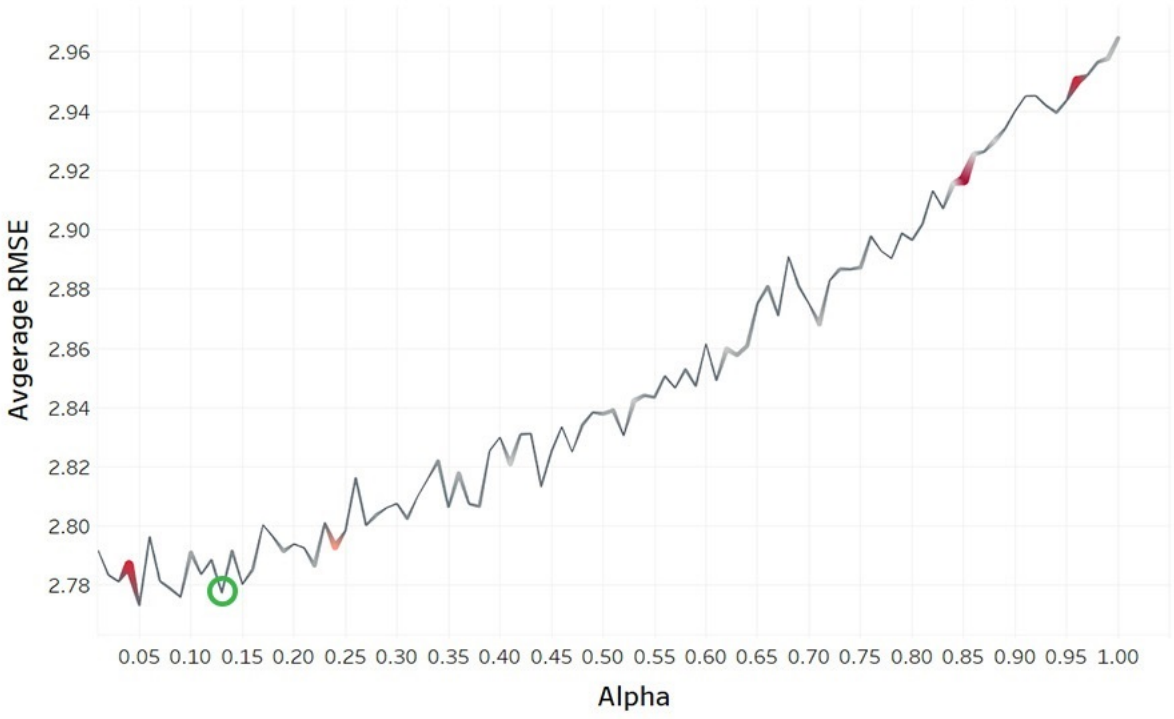

Figure 8: Results of empirical tuning α for the lasso regression model*

*Thicker/red portions of the lines indicate higher vairance. The green circle indicates where $\alpha = 0.13$, the value chosen as the "optimal" level of α

Variable	Frequency	Effect
Age	1.0	-
Year	1.0	-
Weight	1.0	-
Games Played	1.0	-
Steals Per Game	1.0	+
Strength of Schedule	1.0	+
Win Shares	1.0	+
2-Point Field Goal Attempts Per Game	0.8	-
Total Rebounds Per Game	0.8	+
Assists Per Game	0.8	+
Height	0.4	-
Field Goal Attempts Per Game	0.4	-
Free Throw Attempts Per Game	0.2	+

Figure 9: Variables that remain in 5 runs of the lasso regression model, how often they remain in the model, and their direction of effect on the prediction of Peak Win Shares

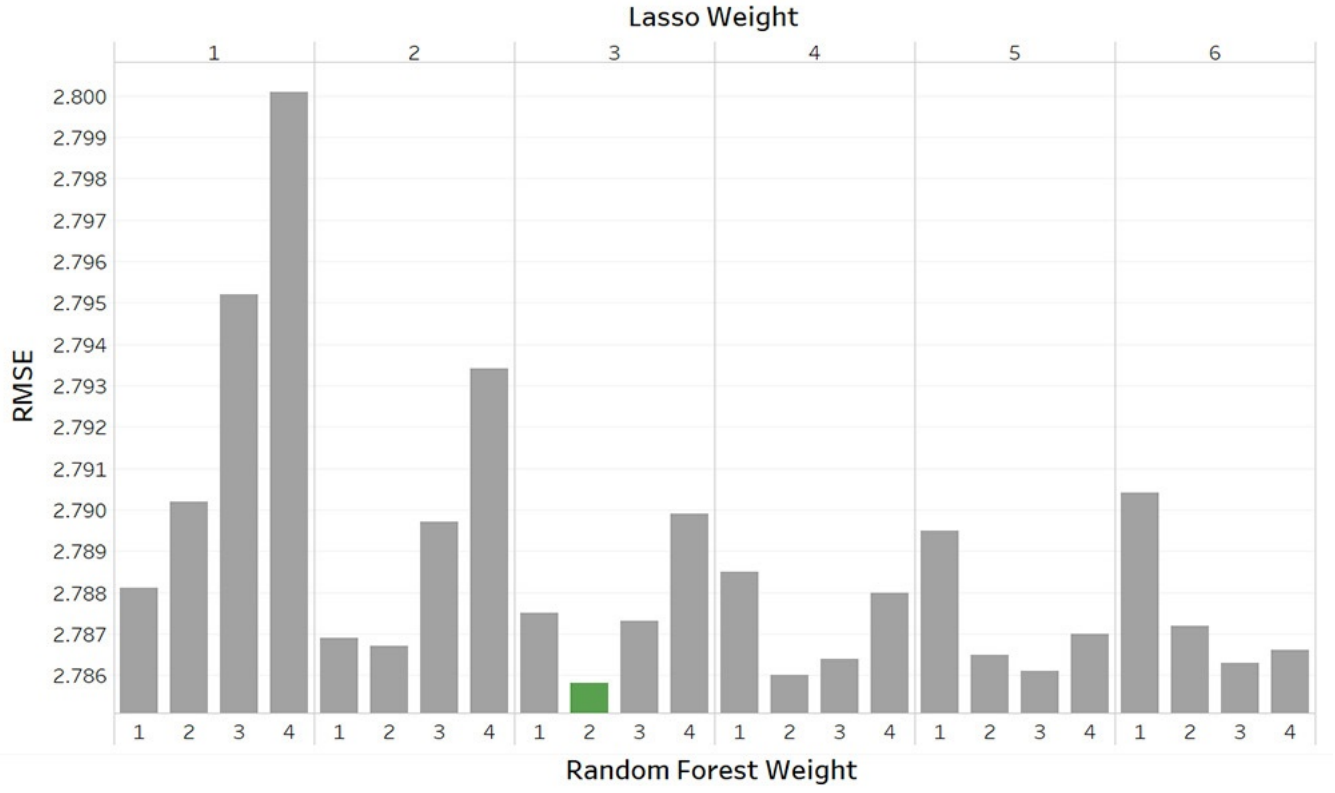

Figure 10: The results of tuning the weights of the different models to obtain the ensemble model (weight of the linear regression model is held fixed at 1)

Model	R^2
Random Forest	0.2383
Lasso Regression	0.2619
Linear Regression	0.2307
Ensemble	0.2691

Figure 11: A comparison of the predictive power of each model

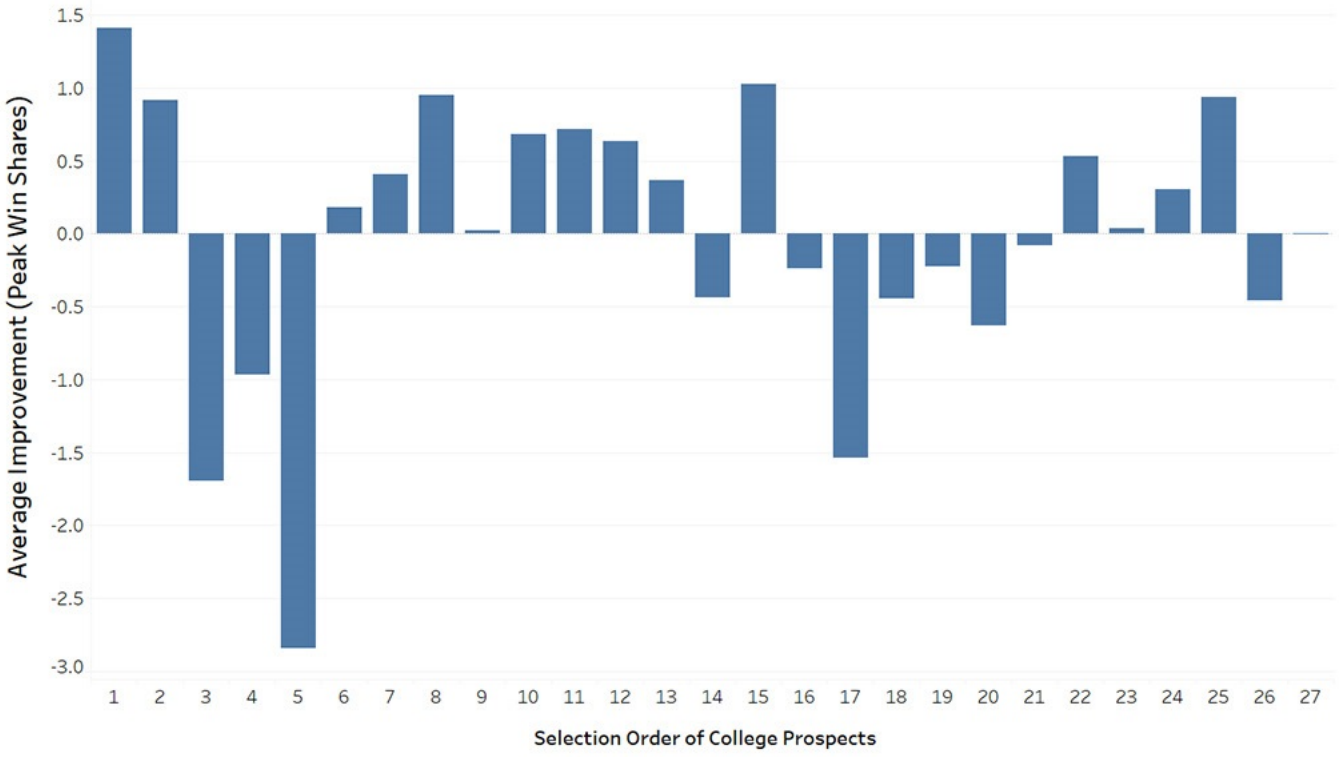

Figure 12: The average improvement in Peak Win Shares of the ensemble model compared to the actual NBA draft selections based on order of players selected

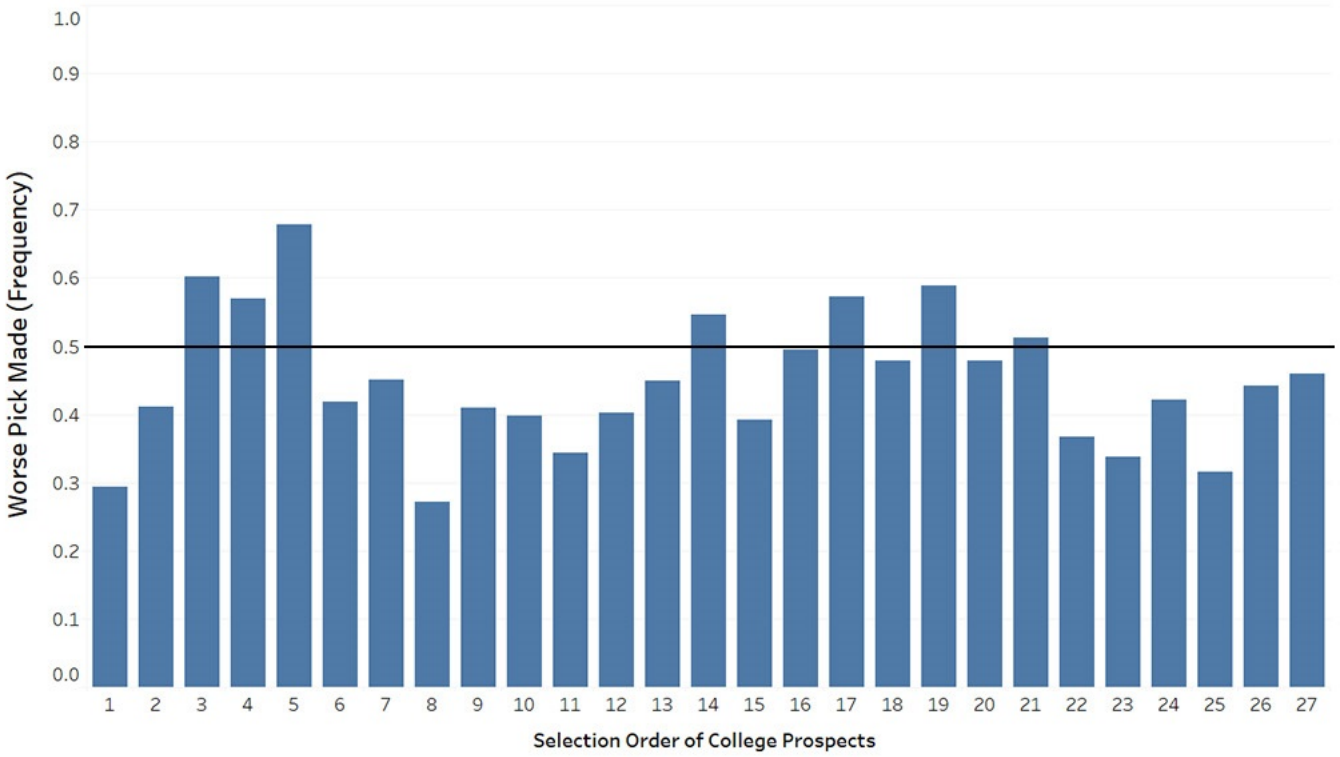

Figure 13: The frequency that the ensemble model selected a player with a lower Peak Win Shares than the actual order of NBA draft selections

*The horizontal black line indicates the 50% mark. The average for this graph is 44.8%.

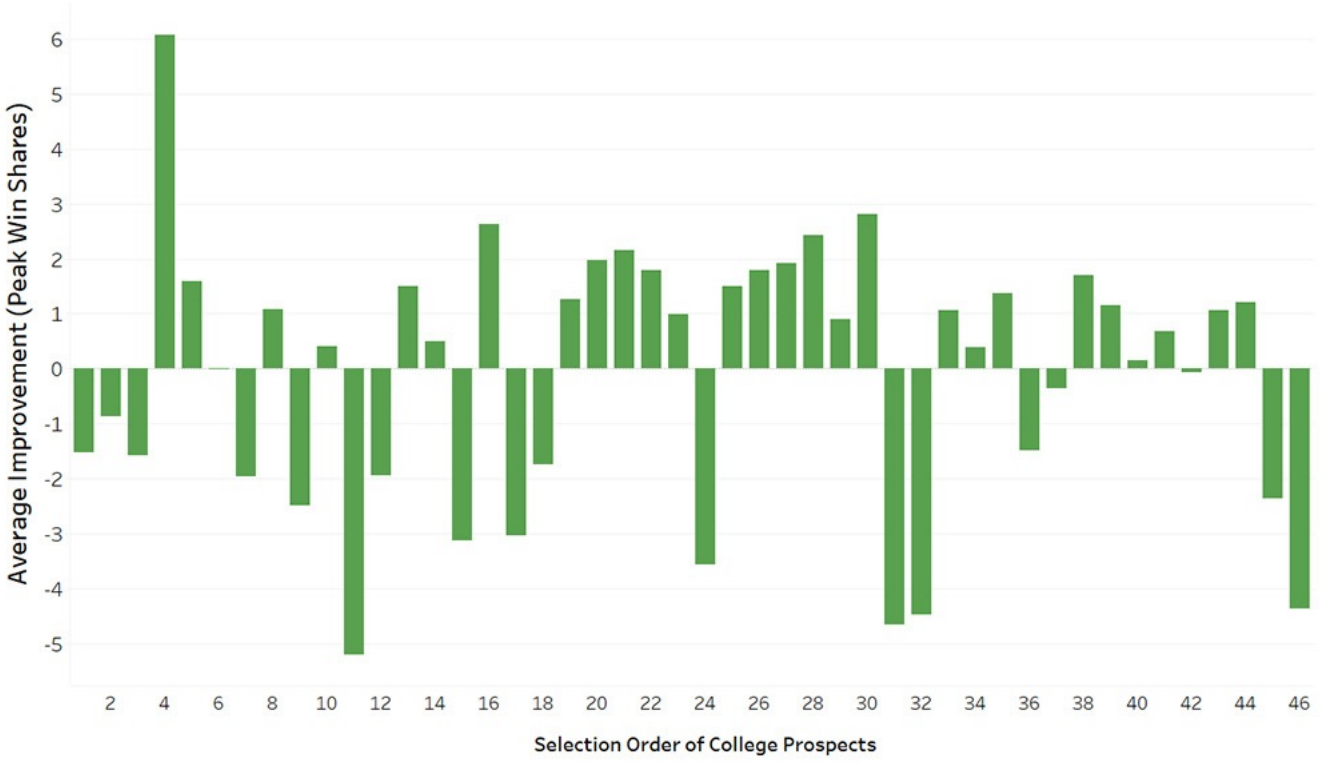

Figure 14: The average improvement in Peak Win Shares of the ensemble model compared to the actual NBA draft selections based on order of players selected (Lock box data set)

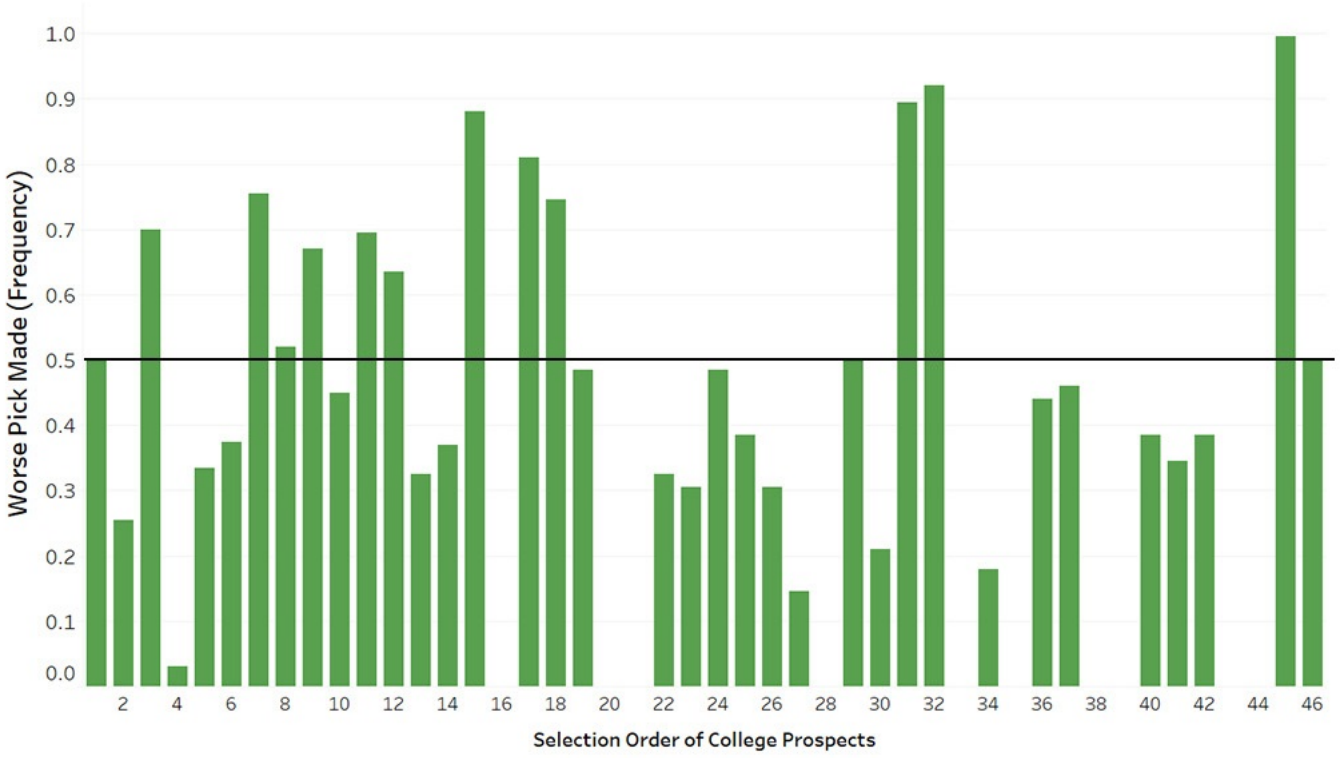

Figure 15: The frequency that the ensemble model selected a player with a lower Peak Win Shares than the actual order of NBA draft selections (Lock box data set)

*The horizontal black line indicates the 50% mark. The average for this graph is 38.5%.

Variable	Observations	Percent Explained by Draft Position
Full Data Set		
Peak Win Shares	683	0.218
Log Peak Win Shares*	683	0.227
50th Percentile of Peak Win Shares (indicator variable)	683	0.161
75th Percentile of Peak Win Shares (indicator variable)	683	0.153
90th Percentile of Peak Win Shares (indicator variable)	683	0.094
95th Percentile of Peak Win Shares (indicator variable)	683	0.057
99th Percentile of Peak Win Shares (indicator variable)	683	0.018
Players with Peak Win Shares > 0		
Peak Win Shares	452	0.130
Log Peak Win Shares**	452	0.092
50th Percentile of Peak Win Shares (indicator variable)	452	0.079
75th Percentile of Peak Win Shares (indicator variable)	452	0.097
90th Percentile of Peak Win Shares (indicator variable)	452	0.045
95th Percentile of Peak Win Shares (indicator variable)	452	0.045
99th Percentile of Peak Win Shares (indicator variable)	452	0.013

*Since this sample includes zeros and negative values for Peak Win Shares, this variable is specifically log[Peak Win Shares $- min$(Peak Win Shares) $+ 1$] which shifts the data so that after the transformation, the smallest value in the data set is 0.

**By restricting the data set to players who have a positive value for Peak Win Shares, this allows the log of this measure to be calculated normally. This restriction is similar to Berri, Brook, and Fenn (2011) where they look at how much of career performance can draft position explain in the NBA. In doing so, their data set "only includes players who logged an average of 500 min per season". The authors claim that this restriction "likely overstates the explanatory power of draft position".

Figure 16: Robustness checks to see which variables of NBA performance are best explained by where a player is selected in the draft