NBA Draft Prospect Data Analysis

Micaleb Johnson, Elias Cheleuitte, Chris Williams

1. Importance and Aims

1.1 Importance

The NBA, and sports in general, have a unifying effect on society bringing people who otherwise have nothing in common together. This has led the NBA to become a billion dollar industry where the primary goal of each team is to win. Each year the NBA holds a draft where each team is able to select the most promising prospects to add to their roster in hopes of developing these players into ones who contribute to a winning team. Given the limited number of draft picks, NBA teams spend a lot of time and money, each year, on scouting prospective players for the upcoming NBA Draft. Despite the large amount of resources that are put into finding out which players to draft, it is still an imperfect science. Oftentimes players who scouts deem as a future franchise player, fall out of the league within a few years and other prospects who fly under the radar end up becoming some of the league's top stars.

1.2 Aims

Our primary aim is to accurately predict the success of an NBA draft prospects career, measured by the career win shares gained by the prospects, using data of players' performances prior to when they are drafted as well as their draft positions. We hope to determine which of the pre-draft factors have the greatest and least effect on models predicting draft prospects' NBA success. We also aim to determine the relative value of each of the first round draft selection spots. In essence, we are interested in whether a top five pick really is worth significantly more than a bottom five pick or even any other pick outside of the top five. This will help in determining how successful teams are in utilizing their draft picks and developing their prospects, by comparing how their selected player performed relative to the expected value of the draft pick as well as the projected performance of the player. Furthermore, this will aid in evaluating the value of draft picks, which are commonly traded for established players by teams looking for immediately impactful players.

1.3 Hypotheses

We expect that the higher a draft selection is, i.e. closer to the start of the draft, the more value that the pick will possess. Similarly we anticipate pick position to have an extremely high effect on any predictive model, as well as age, points per game, and what competition in which the prospects played prior to being drafted, i.e. NCAA, highschool, or international.

2. Data and methods

2.1 Data Collection

We collected data from a variety of sources, we scraped and downloaded data and merged them together. Our initial dataset contained 623 players drafted in the first round of the NBA Draft in the date range 1995-2015. Data on each player included their physical stats (age, height, weight), NCAA career statistics, draft year, college, position, pick, and career NBA win shares. However, a large portion of the players, primarily those who had most recently played in high school or internationally, had missing statistical data. Thus, we had to manually search for the missing data one player at a time. For players

drafted before 2000 the data was still quite sparse so we removed all players drafted before 2000. This still left us with a dataset of over 450 players and 16 explanatory variables with one response variable.

2.2 Statistical Methods

Since our primary aim was to project the career success of an NBA draft prospect based on the continuous response variable career win shares, and all of our subsequent aims were based on the results we would get from the first aim; we decided that we should use two types of regression models. Our first model was chosen to be a parametric regression model to predict which factors from our dataset of draft prospects were the most influential when predicting nba career success. For our second model we wanted to use a non parametric regression method. Initially we considered Quasi-Poisson, Zero-Inflated Poisson, and Zero-Inflated Negative Binomial models, after grouping our NBA data, for our second model attempt, we decided to instead use the non-parametric Random Forest regression model. This was due to a number of factors, including some of our predictors being relatively highly correlated, the presence of a few outliers that we determined should be kept in the data set, and the complexity of any potential parametric models. We believe both models will allow us to achieve both aims in analyzing both the value of each draft position and a prospects expected success.

3. Exploratory Data Analysis

3.1 Data Cleaning/Preprocessing

3.1.1 Missing Data

Despite our best efforts to find accurate statistics for each player in our dataset, we still had large quantities of missing data in the for certain players and the variable turnovers per game. Thus, since our dataset was already quite large we decided to remove the turnover variable and the players who had too much missing data to be beneficial to our investigation. We also found that certain players had a null value for the variable 3pt percentage, however, this was simply due to the fact that these players hadn't attempted any three pointer shots. We decided to fill in these missing values with 0, as players who don't attempt three pointers are likely to not be very bad three point shooters, so assigning this value seemed reasonable.

3.1.2 Outliers

Some higher outliers were present in our response variable, career win shares. However, we had no interest in removing these outlier players from our data as they are the star players, who have a huge impact on a team's success. These players represent the potential upside that a prospect could one day become, so there was no point in removing these star players. Furthermore, after running a preliminary regression model with all variables included and all players included, we did not see any player who had an abnormally high Cook's distance or leverage. This backed up our intuition that these players simply had very successful careers and should not be removed from our dataset.

3.1.3 Transformations

The variable measuring player height required some adjustment because the original height format wouldn't have been usable in R programs, therefore we converted it from the format (6,8) to height in inches (80). The format for the college and position variable would have also been difficult for an R program to accurately interpret as there were many different levels to each variable, thus we turned each variable into a factor with 3 levels. The college factor is broken down into 3 factors, 1 being high school, 2 being international, and 3 being NCAA college. The position factor was broken down into the three primary basketball positions, G (guard), F (forward), and C (center).

3.2 Variable Associations

To begin looking at variable associations for our career win shares, we created a correlation matrix with each variable(Figure 3A). From the matrix you can see that pick appears to have the strongest correlation with win shares with a negative correlation. This follows our hypothesis that as the pick number increases there should be a decrease in win shares. The matrix shows that the other variables do not have strong correlations with win shares. To further explore the correlation with pick and win shares we created a graph that had a boxplot for each pick level to the variability in pick values (Figure 3B). The boxplot shows the mean player win shares decline from the first few picks and steadies out around the tenth pick. The graph also shows the variability of win shares, particularly the upward variability, decreases after the first few picks. Indicating that in general most superstars are taken within the first 5 picks.

Figure 3A: Correlation Plot of Variables

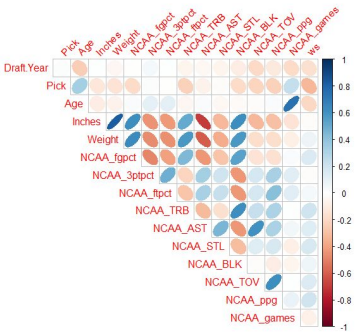
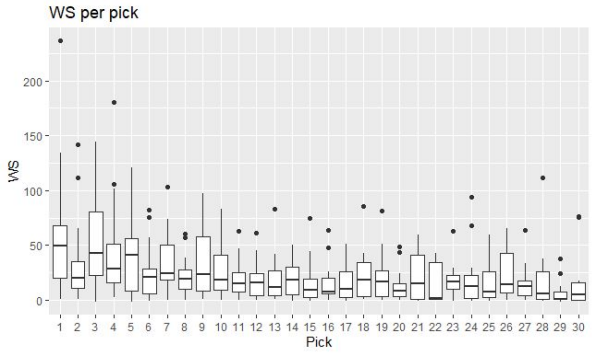


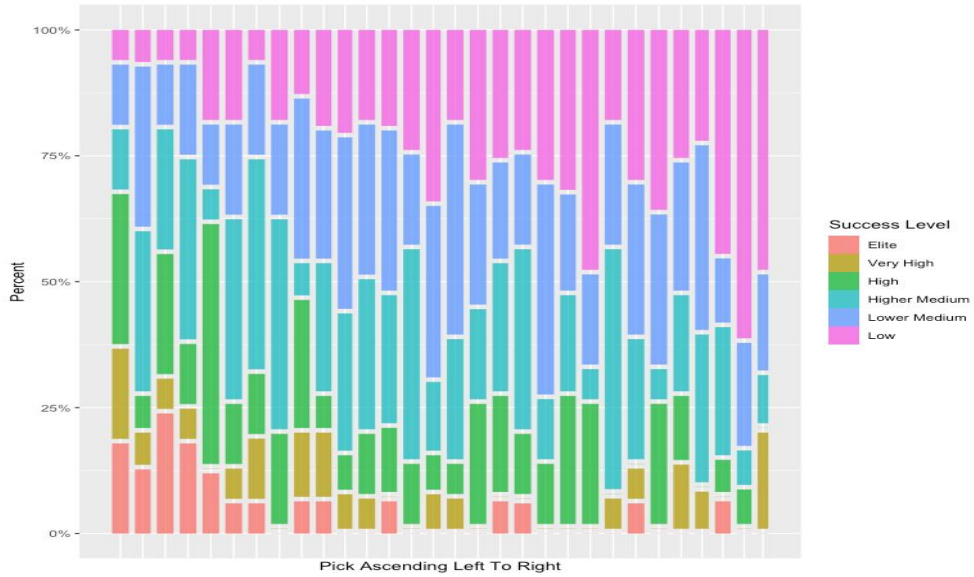
Figure 3B: Box plots of WS per Pick



3.3 Additional Visualizations

Another way to visualize the varying career success levels drafted at each pick was to create a mosaic plot (Figure 3C). The plot shows the proportion of players of the varying success levels drafted at each pick. Starting from the 1st pick going all the way to the 30th pick.The success levels were determined as follows, low = the bottom quartile of the total career win shares, lower medium = players who fall between the 25th and 50th percentile, higher medium = players who fall between the 50th and 75th percentile, High = 75th and 90th percentile, Very high = 90th to 95th percentile, elite = 95th percentile and above. Several interesting things popped out to us. For one, there is rarely a player who is drafted outside the top 5 picks who becomes an elite player. This indicates the importance of a top 5 pick as you have a much higher chance of drafting a franchise altering player than outside the top 5 where it is quite uncommon, The proportion of players who are deemed to have a success level greater than or equal to “high” stays roughly the same from picks 10-30. For most picks in this range, the proportion of players who have a success level lower than high is greater than 75%. Also, the second pick appears to be much different than the other top, with an abnormally low proportion of players with high or greater success levels. We thought that this likely is due to the randomness of selecting successful.

Figure 3C: Mosaic Plot of Proportions Success levels per pick



4. Linear Model

4.1 Model Selection

We randomly split our original dataset by rows into train and test sets, with 80% of our data being put into the training set and the remaining 20% into the testing set. Now working only with the training dataset, we automated the selection of the optimal regression model based on the AIC and BIC for all possible model subsets starting from a full model with all 15 covariates down to an intercept only model. As a result, a six variable model consisting of Pick, Age, Height (inches), Weight, pre NBA draft field goal percentage (The number of shots a player makes / The number of shots a player attempts), and pre NBA draft steals per game, looked to be the best model based on our metrics. All of these covariates had p-values less than the significance level of .05 and thus all are likely meaningful to the model. Another model with seven variables including the same covariates as the previous six variable model plus pre NBA free throw percentage also looked promising, with all but one of the variables appearing to be significant by their p-values. We also looked for any interaction effects to include but none appeared to drastically improve the models performance without greatly increasing the complexity of the model thus none were added.

Figure X: Model Selection Criteria of Top Two Models

Number of Variable	AIC	BIC	Adjusted R^2
6	3459.8047	3490.9819	0.1854
7	3458.6545	3493.7289	0.1898

Six Model Variable	Intercept	Pick	Age	Height (in)	Weight	College fgpct	College steals

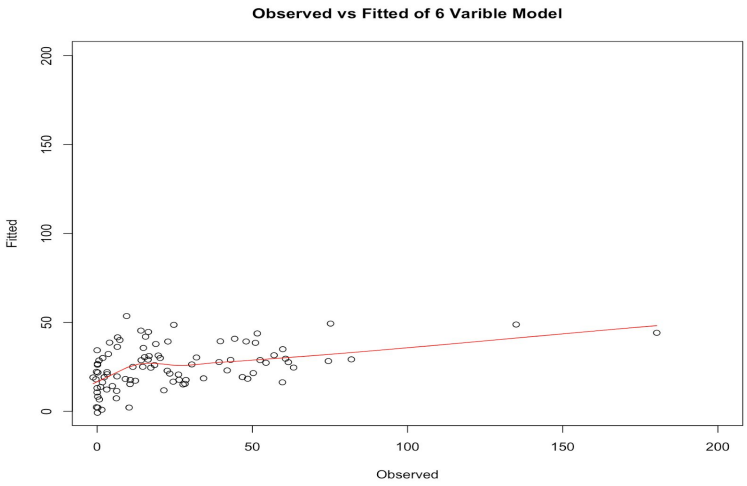
Coefficients	196.0927	-.8997	-2.6013	-2.7692	0.2795	86.9358	8.9628

Seven Variable Model	Intercept	Pick	Age	Height (in)	Weight	College fgpct	College steals	College steals
Coefficients	164.5176	-.8583	-2.8831	-2.7156	0.3039	96.8072	31.6313	8.6381

4.2 Significant Figures

When we looked at the adjusted r-squared values for both models we saw that there was little difference between them. The six variable model had an adjusted r-squared of .1854 whereas the seven variable model's was only slightly higher at .1898, while these values seem low it is to be expected when analyzing a problem with as many complicating factors as we are. We wanted to verify that there wasn't a significant difference between the two models. So we performed an ANOVA test between the two models on our training data. Our ANOVA test between the two models provided p value 0.08755. Based on this result we failed to reject Ho, so we concluded that the more complex model wasn't a significantly better fit on the train data than the less complex six variable model, thus we should use the six variable model. The Six Variable Model was used to predict the career WS on our test data, and plot our predicted ws versus our observed win shares (Figure 4A).

Figure 4A: Observed Win Share vs. Predicted Win Shares



5. Random Forest Model

5.1 Data Pre-Processing

We proceeded to split our modified NBA dataset (as we did in the previous linear model) with the 15 covariates and the response variable into training and testing sets with 80% of the original data going into the training set and 20% into the testing set. First, a default random forest model to predict win shares was fitted on the training dataset. It created 500 different forests each with a different number of trees within them, ranging from 1 to 500, each picking from a random sample of 5 of the covariates at each tree split, and resampling with replacement from all of the sample data. This resulted in a model with an out-of-bag mean squared error (MSE) of 663.392 and explained 17.48 % of the variance.

5.2 Model Fitting and Tuning Parameters

While random forests can provide accurate predictions with little to no tuning, using a standard set of parameters, in order to improve our random forest model to obtain optimal performance and accurate predictions of NBA win shares we proceeded to evaluate different values of our tuning parameters. The tuning parameters of primary focus were the sample size of the predictors which would be considered at each tree split, minimum node size (the lowest number of observations that could be at the end of any one branch of a decision tree, and the size of the bootstrapping sample on which each decision tree would be fitted.

To begin the tuning process a large grid search was performed to predict the best hyperparameters for our model. Specifically, we created a grid that would evaluate 96 different models using an mtry that ranged from 2 to 12 by a sequence of 2, minimum node size with a sequence of 1 to 7 by 2, and a sample size of 0.55, 0.623, 0.70, or 0.80. The grid search generated candidates for each hyperparameter from the sequence constraints and generated combinations of the hyperparameter values. With the generated combinations of parameter values we ran a random forest model and used the OOB RSME to evaluate each model. The 5 best models and their hyperparameter values were recorded as you can see in Figure Y below. From this, our highest performing model can be seen to have an OOB RSME of 25.5754 along with the following hyperparameter values, an mtry of 12, node size of 1 and sample size of 0.80.

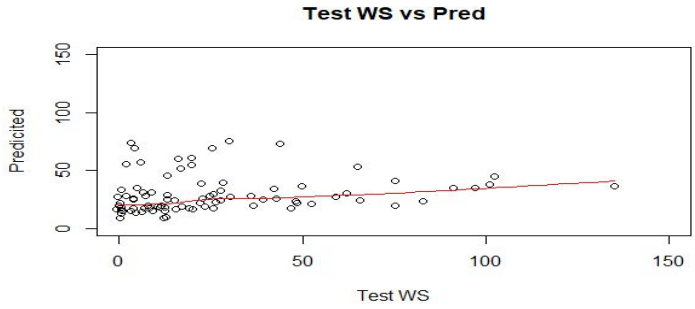
Figure Y: Most optimal hyperparameter Values

Mtry	Node Size	Sample Size	OOB RMSE
12	1	0.80	25.57545
6	1	0.55	25.58738
8	3	0.70	25.62135
12	3	0.80	25.62392
6	1	0.70	25.65105

5.3 Significant Figures

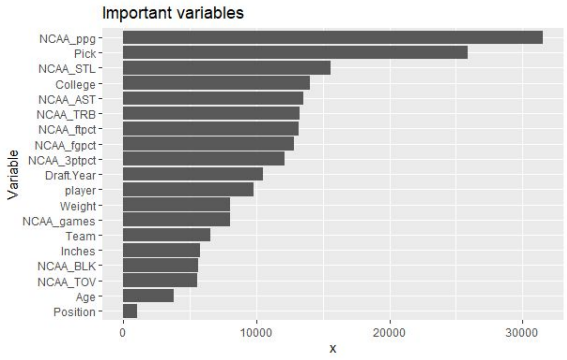
The resulting optimal model with our training data produced an output with an out-of-bag mean squared error (MSE) of 655.0384 and explained 18.77% of the variance. This model is more optimal than the default random forest model, the optimal model has a higher explained variance and a lower OOB MSE While this is not a great R-squared value our data still follows the win share trend fairly accurately outside of lowest extremes of win shares and the highest extremes of win share totals. Our model fails to take into account the variability of win share values. The highest win share values are underpredicted and the lowest win share values are overpredicted. This is likely due to the fact that random forest models are robust to outliers.

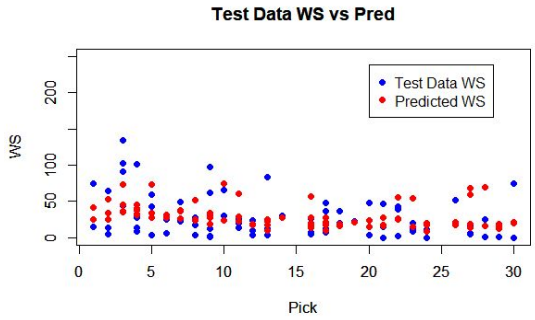
Figure 5A: Plot of Test Datas Observed Win Share vs. Predicted Win Share



The most optimal random forest model allowed us to see which variables were most important to predicting a player's career success via career win shares explained by pre-draft statistics which helps us answer the question for our Aim 2 (which factors are most and least influential in predicting NBA success). Pre-draft points per game and pick value being the most influential while age and position are the least influential. The model also provides evidence in our first aim/hypothesis that pick is a highly influential factor and has a negative effect on player value.

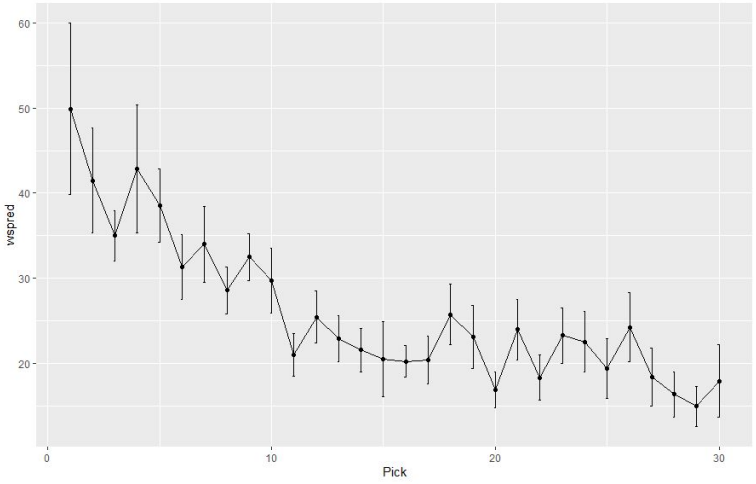
Figure 5B: Important Variable from Optimal Model Figure 5C: Obs and Pred WS by Pick





Using the optimal model we created confidence intervals for each draft pick to predict the value of each draft pick relative to the win share and create a plot showing the confidence intervals (Figure 5D). As predicted, the lower the draft pick the higher the predicted value

Figure 5D: Expected number of career win shares with corresponding 95% confidence interval for each pick, predicted from optimal random forest model



6. Conclusion

6.1 Decisions

From both models we can see that certain factors seemed to be important in predicting a player's career success. For instance, where a player was picked seemed to be telling on how successful a player's career was . However, other variables such as Age greatly differed in importance between the two models. In our linear regression model age was a highly significant variable whereas in the random forest model it was the second least important variable. When it came to selecting which model performed better at predicting NBA career success, Occam's razor rang true in our decision as both of our models performed similarly. Thus, we preferred the simpler multiple linear regression model over the random forest model.

We found the random forest model to be more appropriate for calculating the draft pick values. As the model did a good job of projecting win shares for each pick number (Figure 5C). Thereby, we estimated the expected draft pick values as the projected win shares at each pick from the random forest model. From (Figure 5D) It can be seen that as our hypothesis suggested the expected win shares for each pick declines as the pick number increases.

6.2 Reflections & Future Possibilities

In retrospect, this was a very hard problem to predict, indicated by teams who diligently scout each prospect and still end up making bad picks. There are many immeasurable factors and random things that can determine how successful a player is in their career, injuries for example. Situational factors such as the team a player is drafted to also have an effect on the success of a player. Certain teams are better at developing players than others and thus, the success a player can have in their career depends on this.

One thought we had upon completing our models was that our data was somewhat incomplete as players drafted in 2000, the first draft year in our data, had their entire careers to accumulate win shares whereas players drafted in 2015 have only had 4 seasons in the NBA, far from a full career for the more successful players. So, we should have only included win share counts for a player's first four seasons to better account for this.

New technologies are being developed that are providing teams with more and more information on players and it is vital that teams are able to utilize this data correctly to help form a successful team. One such technology that the NBA has also become heavily invested into is in-game player tracking devices, which can record all types of new information about players. This data can then be used to analyze player performance among many other things. In the future, it is possible that the NCAA could begin to use similar technologies, thus providing NBA teams with brand new types of data on upcoming prospects and possibly introducing new ideas and methods on how to correctly evaluate a prospect.

7. References

[1] https://www.basketball-reference.com/draft/

[2] https://data.world/bgp12/nbancaacomparisons/workspace/file?filename=players.csv

[3] https://data.world/gmoney/nba-drafts-2016-1989/workspace/file?filename=NBA+Drafts.xlsx

[4] https://github.com/SidTheKid007/NBARookieAnalysis/blob/master/CollegeRookieStatLog4.csv

[5] https://uc-r.github.io/random_forests