From Draft Night to Career Longevity: A Data Mining Analysis of NBA Draft Picks (1989–2021)

Abstract

The following project examines the National Basketball Association's (NBA) annual player draft data from 1989 to 2021, with the primary goal of understanding how draft-related factors influence long-term and/or productive career outcomes. By combining association analysis (Apriori), correlation analysis (Pearson's and Spearman's), multiple linear regression, and classification techniques such as decision trees and Random Forest, this analysis aims to identify relationships between draft positioning, draft team and conference, and educational background, and the resulting career outcomes, including career length and productivity. The results of this analysis show that draft position is the most consistent and strongest predictor of career success, while other factors, such as educational background and draft team and conference, demonstrate weaker associations with career success. Overall, these findings provide insight into how draft-related factors impact the career outcomes of NBA players over the 32-year period encompassed by the data.

Related Work

Prior research has demonstrated the increasing application of machine learning technique in sports-related areas. An October 2021 paper by Marin Fotache, Irina-Cristina Cojocariu, and

Armand Bertea titled “High-Level Machine Learning Framework for Sports Events Ticket Sales Prediction” demonstrates the implementation of supervised learning techniques like Random Forest and gradient boosting (i.e., XGBoost) in an entirely R-based framework to predict the match ticket sales for a football/soccer club that will be competing in a national championship in light of life potentially returning to normal during the height of the COVID-19 Pandemic. Despite the small sample size of 83 matches, the R^2 values were around 82%, which is very promising and could potentially lay the foundation for reproducing in a larger-scale environment (Fotache et al.).

While the project outlined in this paper shares similarities with the work of Fotache et al. in its use of an entirely R-based framework and supervised learning techniques, such as Random Forest, it differs in both scope and objective. Rather than focusing on short-term, event-based prediction on association football, this project examines player career longevity and productivity in professional basketball through the analysis of draft-day data, highlighting a distinct application of machine learning methods within sports analytics.

Data

The dataset used in this analysis was the “NBA Draft Basketball Player Data 1989-2021,” a Kaggle-based dataset created by Kaggle Grandmaster Matt Hill in 2021, a clinical data and AI Researcher at the University of Pennsylvania Perelman School of Medicine (Hill’s LinkedIn Profile). Hill sourced the data from Basketball-Reference, a comprehensive online database maintained by Sports Reference, a sports statistics company. Each row in this dataset represents a

player drafted by an NBA team between the 1989 NBA Draft and the 2021 NBA Draft. The description that Hill provides for the dataset is “The dataset contains all NBA Draft picks from 1989-2021. Dataset consists of year, overall pick and player data (Hill).” The database contains twenty-four attributes, including overall pick, team, player name, college, years active, total stats (e.g., games, minutes, points, rebounds, and assists), rate stats (e.g., field goal, three point, and free throw percentage), averages (points, rebounds, and assists per game), and advanced analytics (win shares, win shares per 48 minutes, box plus minus, and value over replacement).

The dataset required numerous preprocessing and cleaning steps in order for it to be ready for the different analyses discussed in this project. Firstly, the dataset’s columns needed to be manipulated. Columns that are useless for the sake of this project, like id and rank, were removed. In addition, most columns were renamed for standardization and ease of use. For example, some of the advanced analytics contained within the dataset were renamed to their well-known abbreviations (i.e., BPM, VORP). Secondly, players who were drafted in a particular draft but never once played in the NBA (as of 2021) were removed from the dataset. These players were identified by the NAs or empty strings in the years active column. This allows players who have actually played NBA minutes to remain in the dataset, removing unnecessary noise and enabling more accurate analysis later on. Thirdly, a few team abbreviations needed to be standardized to their present-day counterparts, as different abbreviations representing the same franchise were present in the database. This is due to the fact that, within the 32 years contained in the dataset, different NBA franchises relocated to different cities, thereby changing their abbreviations. A few well-known examples include the Brooklyn Nets, formerly known as the New Jersey Nets until their relocation to Brooklyn in 2012, and the New Orleans Pelicans, formerly known as the “original” Charlotte Hornets until their relocation to New Orleans in

2002. Standardizing team abbreviations allowed for consistency, as well as better traceability and identification of which players in the past got drafted by present-day NBA franchises. Fourth, the addition of new features was required to address holes in the dataset. The first of these was creating a team conference column, where each team would be mapped to the conference it belongs to. In the NBA, there are two conferences, Eastern and Western, each with fifteen teams that correspond to the rough geographic bounds of each conference within the contiguous 48 states of the U.S., as well as Canada. To accomplish this, two vectors corresponding to the abbreviations of each team in each conference were created, which were then used to map each team to a conference based on the vector to which it belonged. The second feature added to this dataset was "educational prestige." This new feature maps each NBA draft pick to one of three educational pathways: Power Schools, Other Colleges, and Non-Collegiate. For the purposes of this dataset only, a power school is defined as one of the top ten college basketball programs historically in the United States, as determined th a thorough analysis led by CBS Sports Writer Matt Norlander, published on November 19, 2020. In his analysis, titled "The Greatest College Basketball Programs Ever: Ranking the top teams of all time," Norlander was driven to research the history of college basketball in the United States, fueled by the long college basketball offseason of 2020 due to the COVID-19 Pandemic. In support of his drive to research, Norlander stated, "In this longest of offseasons, I was sparked to research the history of the sport, and in doing so began to assemble a list of the greatest schools. From there, I wanted to provide a ranking built much more on statistics and achievements than broad perception or subjective rankings (Norlander et al.)." In order to build this list of the greatest college basketball programs ever, Norlander looked to a diverse set of sources, including "…the NCAA record book, Sports Reference, RealGM.com, CollegePollArchive.com, Stats Inc., school record books,

KenPom.com, and ESPN's College Basketball Encyclopedia (Norlander et al.).” From there, Norlander developed a system to assign points to a college basketball program based on its achievements throughout the program’s history. Notable inclusions in this system include NCAA Tournament championships (20 points), Final Four appearances without a national title (10 points), and NBA draft picks, ranging from Top 10 picks (5 points), picks 11-30 (3 points), and picks 31-60 (1 point). Ultimately, Norlander was able to compile a definitive list of the top 25 college basketball programs of all time. For this project, I adopted Norlander’s top 10 programs as the power schools. The top 10 programs are Kentucky, North Carolina, Duke, UCLA, Kansas, Louisville, Indiana, UConn, Villanova, and Cincinnati. While notable programs like St. John’s (#17), Michigan (#18), and Georgetown (#20) were right outside the top 15, the top 10 schools from this list were chosen for simplicity and to ensure the power schools represented in the dataset are truly the cream of the crop. In addition to the power schools, drafted players are also represented by other colleges (schools not included as power schools) and non-collegiate pathways (high school, international, G League Ignite, etc). Once compiled, these tags were mapped to each drafted NBA player to ensure a greater level of distinguishability between each player’s educational journey. The final features added to the original dataset included binning attributes, such as overall pick and total games played, and categorizing them into buckets. Binning these attributes is useful for the association analysis that will be performed on the dataset, as it will aid the search for strong association rules through running the Apriori algorithm. These attributes were categorized based on real-life scenarios. For the overall NBA Draft, the process typically consists of two rounds, each containing 60 picks. The top 14 picks are known as “lottery picks,” as they are awarded to the teams with the worst records in the previous season through a lottery with varying odds depending on the team’s record at the end of

the season. Picks 15 through 30 make up the rest of the first round, while picks 31 through 60 make up the second round. A break for the top five picks was made on top of lottery picks, as the top five picks in any given draft are very lucrative and are where the most coveted draft prospects will be drafted. For total games played, the bucket was based on the average length of an NBA career, which is typically 4 to 5 years, or approximately 328 to 410 games. This is, of course, assuming the typical 82-game regular-season schedule and assuming a player plays all 82 games in the regular season.

Methodology

This project employed a combination of various analytical and statistical methods, including association analysis (Apriori), correlation analysis (Pearson's and Spearman's), multiple linear regression, and classification techniques such as decision trees and Random Forest. These methods were selected to address specific questions about the dataset, aiming to achieve the primary goal of understanding how draft-related factors influence long-term and/or productive career outcomes. It is important to note that for each analysis, a copy of the cleaned dataset was made to ensure the preservation of the cleaned version of the original dataset. To ensure that these methods could be properly executed in R, a few libraries had to be installed and imported. These include arules, arulesviz, dpylr, caret, rpart, rpart.plot, and randomForest. These libraries come with all the necessary tools needed to not only execute these methods properly, but also, in the case of the dyplr library, to manipulate the dataset, such as renaming and selecting/deleting different columns within a dataset.

The first analytical method utilized in this project was association analysis, also known as association rule mining or association rule learning. According to the online learning platform Geeks for Geeks, "Association rules are a fundamental concept used to find relationships, correlations or patterns within large sets of data items… Association rules originated from market basket analysis and help retailers and analysts understand customer behavior by discovering item associations in transaction data (Geeks for Geeks)." Association analysis relies on three key evaluation metrics: support, confidence, and lift. Support simply refers to the frequency with which an itemset appears in a dataset. Confidence is the probability that transactions containing an item 'X' will also contain item 'Y'. Finally, lift measures the strength of an association rule by showing how much more likely items 'X' and 'Y' are to be coupled together than expected if they were independent of each other (Geeks for Geeks). In the context of this project, association analysis was employed to identify frequent combinations of draft position, educational background, and team conference with varying career outcomes. The aim was to answer the question of which combinations of draft positioning, college/educational background, and team drafted by are most frequently associated with players who played at least 400 games. As seen previously, continuous variables such as total career games played and overall pick were binned in the original dataset in preparation for association analysis to uncover the relationships between draft-related attributes and career outcomes. During the association analysis phase, the dataset was copied and modified to perform association analysis using the Apriori algorithm. These modifications included converting dataset items into factors as a list, forcing the dataset back into a data frame, and then creating a transaction list for the Apriori algorithm based on the modified dataset. According to Geeks for Geeks, "Apriori Algorithm is a basic method used in data analysis to find groups of items that often appear together in large sets

of data. It helps to discover useful patterns or rules about how items are related which is particularly valuable in market basket analysis (Geeks for Geeks).” A total of three iterations of the Apriori algorithm were run, starting with a 5% minimum support threshold and 40% minimum confidence threshold. Each successive iteration increased the minimum support and confidence threshold, up to 20% minimum support and 50% minimum confidence. This was done in order to extract the strongest association rules possible from the dataset. Additionally, subsets of the transaction list were created to filter the right-hand side (or the “consequent”) and search for career bins representing short, average, and long careers. This was done to analyze the draft-related factors that lead to one of the three binned career outcomes in terms of total games played.

The second method employed to analyze this dataset was correlation analysis, specifically using Pearson’s and Spearman’s correlation coefficients (also known as Spearman’s Rank Correlation). The aim of this analysis is to determine how strongly a draft position correlates with a player's career longevity and productivity, and how strongly a higher Box Plus Minus (BPM) or Value Over Replacement (VORP) correlates with a player's draft position. According to Geeks for Geeks, “Pearson Correlation Coefficient (PCC) is used for measuring the strength and direction of a linear relationship between two variables. It is important in fields like data science, finance, healthcare, and social sciences, where understanding relationships between different factors is important (Geeks for Geeks).” In addition, Geeks for Geeks demonstrates that Spearman’s Correlation Coefficient “…is a statistical measure of the strength and direction of the monotonic relationship between two continuous variables. Therefore, these attributes are ranked or put in order of preference. It is denoted by the symbol "rho" (ρ) and can take values between -1 to +1. A positive value of rho indicates that there exists a positive relationship between the two

variables, while a negative value of rho indicates a negative relationship. A rho value of 0 indicates no association between the two variables (Geeks for Geeks).” While similar, the primary difference between Pearson’s and Spearman’s is that Pearson’s is used for linear relationships, whereas Spearman’s is used for monotonic, not necessarily linear relationships. While one or the other could’ve been chosen for this project, the decision was made to include both to observe the differences between each in terms of measuring relationships between variables, regardless of the linearity in these relationships. For correlation analysis, attributes such as overall pick, totals, per-game averages, and advanced metrics were retained, while the rest were discarded from the dataset. Next, a correlation matrix was run on the dataset using both Pearson’s and Spearman’s correlation coefficients. Finally, two important scatter plots were generated: one showing overall pick by career games played and another showing overall pick by win shares. Each plot was fitted with a line of best fit, with the line itself modified to be a red color and wide enough to distinguish within each scatter plot.

The third method used to analyze this dataset was multiple linear regression. The goal of this analysis was to predict total career games, total minutes played, and minutes per game (dependent variables) based on the draft-related attributes (overall pick, pick bucket, educational prestige, and team conference), which served as the independent variables. The goal of this analysis is to accurately predict the number of career games played based on different draft-related attributes, and to determine the extent to which draft order predicts the total minutes a player plays per game and throughout their entire career. According to Matt Hayes of Investopedia, a leading online financial education resource, “Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of MLR is to model the linear

relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable (Hayes).” In essence, because of the existence of more than one draft-related attribute, multiple linear regression was the most sensible choice for the third step of the analysis of this dataset. For this analysis, it was only necessary to keep the draft attributes, total games played, total minutes played, and minutes per game. Three MLR models were created, where total games played, total minutes played, and minutes per game served as the dependent variables in their respective models. Then, each model was summarized using the base R summary method to examine the results produced by each model.

The fourth and final method used to analyze this dataset was binary classification, specifically decision trees and Random Forest. The goal of this analysis was to predict whether NBA players will be successful or unsuccessful, have long or short careers, and other similar outcomes. The critical question that seeks an answer is whether it is possible to accurately label which draft picks will have long-term careers. With regard to decision trees, Geeks for Geeks describes them as something that “…helps us make decisions by showing different options and how they are related. It has a tree-like structure that starts with one main question called the root node which represents the entire dataset. From there, the tree branches out into different possibilities based on features in the data (Geeks for Geeks).” On the other hand, Eda Kavlakoglu, who works in AI Business Development + Partnerships at IBM’s Research Division, describes the random forest classifier as “…a commonly-used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles

both classification and regression problems (Kavlakoglu). ” Given the nature of decision trees and random forests, it is safe to say that decision trees are able to give a good performance baseline, while random forests can compound the effect and deliver much better results.

For classification, only draft-based factors were kept in the dataset, while career length attributes were kept temporarily. This was done to ensure that binary class labels represented “long careers” and “not long careers”, the latter of which encompasses both short and average careers. The labels were then factorized, followed by the deletion of the now useless career length attributes. Finally, it was verified that factorization had been completed properly, and a table was created showing the total number of long and non-long careers. For both decision trees and random forests, the decision was made to try iterations with different training splits of 70% training and 30% testing, and 80% training and 20% testing. This was done to see if there would be an increase in the performance of each classifier. To top it all off, the predictions for each classifier were executed successfully, resulting in the creation of confusion matrices and plots to illustrate the results in full color. For random forest in particular, a variable importance plot was created for each training/test split in order to visualize which of the draft-related attributes were the most important and thus “driving” the results of the model the most. Geeks for Geeks briefly explains that “…Feature Importance in Random Forests measures how much each feature contributes to the model’s prediction accuracy. It helps in identifying the most influential input variables, improving performance, interpretability and computational efficiency (Geeks for Geeks).”

Results

For the first question of which combinations of draft positioning, college/educational background, and team drafted by are most frequently associated with players who played at least 400 games, the Apriori algorithm used a minimum support of 5% and a minimum confidence of 40% to generate the following association rules for both long careers and short careers on the right-hand side, as seen below:

```
> inspect(sort(rules_long_career, by = "confidence")[1:3])
    lhs                                                  rhs                          support    confidence coverage   lift     count
[1] {pick_bucket=Top5}                                => {career_bucket=LongCareer} 0.06770521 0.6848485  0.09886159 2.055777 113
[2] {pick_bucket=Lottery}                             => {career_bucket=LongCareer} 0.08687837 0.4898649  0.17735171 1.470476 145
[3] {educational_prestige=OtherCollege, pick_bucket=Lottery} => {career_bucket=LongCareer} 0.05512283 0.4742268  0.11623727 1.423533  92
```

Rules when the right-hand side is sorted by long career at 5% support/40% confidence

```
> #Inspect new top 10 rules by confidence
> inspect(sort(rules_short_career, by = "confidence")[1:10])
    lhs                                                                          rhs                           support   confidence coverage  lift     count
[1]  {educational_prestige=OtherCollege, pick_bucket=SecondRound, conference=Western} => {career_bucket=ShortCareer} 0.1012582 0.6954733  0.1455962 1.501610 169
[2]  {educational_prestige=OtherCollege, pick_bucket=SecondRound}                    => {career_bucket=ShortCareer} 0.2067106 0.6900000  0.2995806 1.489793 345
[3]  {educational_prestige=OtherCollege, pick_bucket=SecondRound, conference=Eastern} => {career_bucket=ShortCareer} 0.1054524 0.6848249  0.1539844 1.478619 176
[4]  {pick_bucket=SecondRound, conference=Western}                                   => {career_bucket=ShortCareer} 0.1408029 0.6811594  0.2067106 1.470705 235
[5]  {pick_bucket=SecondRound}                                                       => {career_bucket=ShortCareer} 0.2810066 0.6767677  0.4152187 1.461223 469
[6]  {pick_bucket=SecondRound, conference=Eastern}                                   => {career_bucket=ShortCareer} 0.1402037 0.6724138  0.2085081 1.451822 234
[7]  {educational_prestige=OtherCollege, conference=Eastern}                         => {career_bucket=ShortCareer} 0.1695626 0.4887737  0.3469143 1.055321 283
[8]  {educational_prestige=OtherCollege}                                             => {career_bucket=ShortCareer} 0.3301378 0.4854626  0.6800479 1.048172 551
[9]  {educational_prestige=OtherCollege, conference=Western}                         => {career_bucket=ShortCareer} 0.1605752 0.4820144  0.3331336 1.040727 268
[10] {conference=Eastern}                                                            => {career_bucket=ShortCareer} 0.2342720 0.4710843  0.4973038 1.017128 391
```

Rules when the right-hand side is sorted by short career at 5% support/40% confidence

Based on the generated association rules, several rules stand out as strong association rules due to their high confidence and lift. The rule "{pick_bucket = Top5} ⇒ {LongCareer}" stands out amongst the pack with a confidence over 68%, and a lift over 2, which indicates a very strong rule that indicates that players drafted in the top 5 of any given draft class are twice as likely to have strong careers compared to any other player drafted outside the top 5. This means that the top 5 picks of any given draft are highly coveted and lucrative. The rule "{pick_bucket = Lottery} ⇒ {LongCareer}" also stands tall with a confidence near 49% and a lift of about 1.47. While not as strong as the previous rule, it is still a significant rule that demonstrates players drafted between 5 and 14 are nearly 1.5 times more likely to have strong careers than players

drafted after the lottery. In terms of rules, where the right-hand side is sorted for short careers, most rules convey the idea that if a player is drafted in the second round (between picks 31-60) and goes to a college outside of the power schools, they are more likely to have short careers. Both the Western and Eastern conferences are represented in this ruleset, indicating that the team a player is drafted to or the conference they are drafted into has little impact on the likelihood of a short career being extended into a long one if a player already attended a non-power school and was drafted in the second round. During the second iteration of Apriori, the support was set to 10% with the same confidence of 40%. As a result, only 12 association rules with short careers on the right-hand side were generated, with none from long careers. These 12 rules matched the results seen in the previous iteration. Finally, for the third iteration, the support was increased to 20%, and the confidence increased to 50%. This generated only two association rules with short careers on the right-hand side, as seen below:

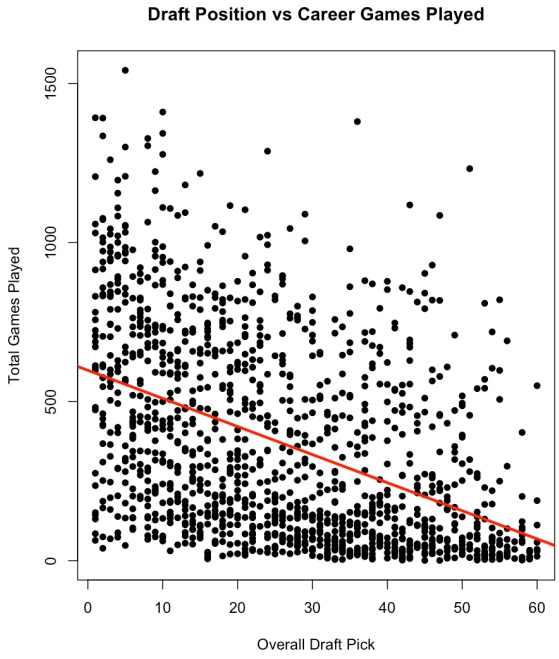
```
> inspect(sort(rules_short_career3, by = "confidence")[1:2])
    lhs                                                    rhs                          support   confidence coverage  lift     count
[1] {educational_prestige=OtherCollege, pick_bucket=SecondRound} => {career_bucket=ShortCareer} 0.2067106 0.6900000  0.2995806 1.489793 345
[2] {pick_bucket=SecondRound}                              => {career_bucket=ShortCareer} 0.2810066 0.6767677  0.4152187 1.461223 469
```

Rules when the right-hand side is sorted by short career at 20% support/50% confidence

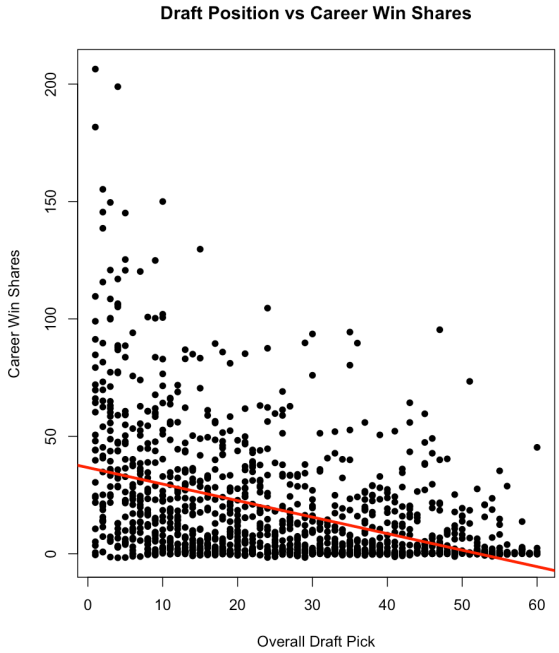
The two rules generated above don't generate anything newer than what's been known for decades: second-round picks in the NBA typically have very short careers, especially those who didn't attend a power school, and many don't even make it to an NBA roster at all. The confidence of over 67% and a lift of over 1.46 for both solidify this conclusion.

For the second and third questions of how strongly a draft position correlates with a player's career longevity and productivity, and how strongly a higher Box Plus Minus (BPM) or Value Over Replacement (VORP) correlates with a player's draft position, the correlation analysis consists of both Pearson's and Spearman's correlation coefficient between each attribute

in the dataset. In addition, the analysis led to the generation of two scatter plots that visualize the relationship between draft position and career games, as well as draft position and win shares:



Scatter plot of Draft Position vs Career Games Played with red line of best fit



Scatter plot of Draft Position vs Career Win Shares with red line of best fit

	overall_pick	total_games_played	total_minutes_played	total_points	total_rebounds	total_assists	mpg	ppg	rpg	apg	win_shares	ws_per_48	bpm	vorp
overall_pick	1.0000000	-0.4445080	-0.4605023	-0.4544591	-0.4339013	-0.3545715	-0.5580782	-0.5393377	-0.4686739	-0.3540784	-0.3947948	-0.2124969	-0.3028719	-0.3212033
total_games_played	-0.4445080	1.0000000	0.9618342	0.8836041	0.8482185	0.7431653	0.7380997	0.6494436	0.5690414	0.4843710	0.8546047	0.3816557	0.4876670	0.6682114
total_minutes_played	-0.4605023	0.9618342	1.0000000	0.9606697	0.8717571	0.8219563	0.7914839	0.7410813	0.5890667	0.5697912	0.9244661	0.3657285	0.4930924	0.7878512
total_points	-0.4544591	0.8836041	0.9606697	1.0000000	0.8367307	0.8211576	0.7720416	0.8120560	0.5726132	0.5871031	0.9413768	0.3581825	0.4921378	0.8627457
total_rebounds	-0.4339013	0.8482185	0.8717571	0.8367307	1.0000000	0.5733817	0.6746865	0.6331232	0.8033335	0.3322848	0.8819415	0.3774270	0.4477732	0.7248861
total_assists	-0.3545715	0.7431653	0.8219563	0.8211576	0.5733817	1.0000000	0.6568858	0.6490752	0.3150450	0.8170881	0.7907804	0.2779919	0.4325334	0.7947493
mpg	-0.5580782	0.7380997	0.7914839	0.7720416	0.6746865	0.6568858	1.0000000	0.9195119	0.6982088	0.7016757	0.6928008	0.3952907	0.5637342	0.5823059
ppg	-0.5393377	0.6494436	0.7410813	0.8120560	0.6331232	0.6490752	0.9195119	1.0000000	0.6544041	0.6964776	0.7065205	0.3991625	0.5657953	0.6582608
rpg	-0.4686739	0.5690414	0.5890667	0.5726132	0.8033335	0.3150450	0.6982088	0.6544041	1.0000000	0.2505629	0.6157203	0.4308592	0.4605560	0.4789766
apg	-0.3540784	0.4843710	0.5697912	0.5871031	0.3322848	0.8170881	0.7016757	0.6964776	0.2505629	1.0000000	0.5199537	0.2314235	0.4422823	0.5631976
win_shares	-0.3947948	0.8546047	0.9244661	0.9413768	0.8819415	0.7907804	0.6928008	0.7065205	0.6157203	0.5199537	1.0000000	0.3982438	0.4999754	0.9298994
ws_per_48	-0.2124969	0.3816557	0.3657285	0.3581825	0.3774270	0.2779919	0.3952907	0.3991625	0.4308592	0.2314235	0.3982438	1.0000000	0.9061562	0.3343481
bpm	-0.3028719	0.4876670	0.4930924	0.4921378	0.4477732	0.4325334	0.5637342	0.5657953	0.4605560	0.4422823	0.4999754	0.9061562	1.0000000	0.4649890
vorp	-0.3212033	0.6682114	0.7878512	0.8627457	0.7248861	0.7947493	0.5823059	0.6582608	0.4789766	0.5631976	0.9298994	0.3343481	0.4649890	1.0000000

Results of Pearson's Coefficient Matrix

	overall_pick	total_games_played	total_minutes_played	total_points	total_rebounds	total_assists	mpg	ppg	rpg	apg	win_shares	ws_per_48	bpm	vorp
overall_pick	1.0000000	-0.4967772	-0.5355165	-0.5525265	-0.5399758	-0.5038487	-0.5600458	-0.5561892	-0.4976593	-0.3789518	-0.4646162	-0.2738492	-0.3694010	-0.2671223
total_games_played	-0.4967772	1.0000000	0.9838990	0.9647278	0.9567032	0.9210234	0.7953845	0.7533139	0.6627774	0.5869471	0.9269421	0.5993165	0.6879402	0.5464172
total_minutes_played	-0.5355165	0.9838990	1.0000000	0.9909167	0.9660840	0.9525303	0.8872626	0.8437907	0.7160251	0.6700582	0.9403517	0.6090589	0.7291244	0.5882270
total_points	-0.5525265	0.9647278	0.9909167	1.0000000	0.9565515	0.9527622	0.9043419	0.8933743	0.7233954	0.6907213	0.9375345	0.6211946	0.7506563	0.6082337
total_rebounds	-0.5399758	0.9567032	0.9660840	0.9565515	1.0000000	0.8758433	0.8436687	0.8032225	0.8418387	0.5461550	0.9390342	0.6702727	0.7186540	0.5637302
total_assists	-0.5038487	0.9210234	0.9525303	0.9527622	0.8758433	1.0000000	0.8871361	0.8510352	0.5985846	0.8429789	0.8736483	0.5348064	0.7389880	0.6032998
mpg	-0.5600458	0.7953845	0.8872626	0.9043419	0.8436687	0.8871361	1.0000000	0.9556485	0.7571709	0.7866117	0.8253207	0.5400681	0.7164928	0.5774584
ppg	-0.5561892	0.7533139	0.8437907	0.8933743	0.8032225	0.8510352	0.9556485	1.0000000	0.7257775	0.7641404	0.7973133	0.5562939	0.7309202	0.5807501
rpg	-0.4976593	0.6627774	0.7160251	0.7233954	0.8418387	0.5985846	0.7571709	0.7257775	1.0000000	0.3613769	0.7421189	0.6488380	0.6070575	0.4449988
apg	-0.3789518	0.5869471	0.6700582	0.6907213	0.5461550	0.8429789	0.7866117	0.7641404	0.3613769	1.0000000	0.5730042	0.3024742	0.6074214	0.5041587
win_shares	-0.4646162	0.9269421	0.9403517	0.9375345	0.9390342	0.8736483	0.8253207	0.7973133	0.7421189	0.5730042	1.0000000	0.7925536	0.8301901	0.7138212
ws_per_48	-0.2738492	0.5993165	0.6090589	0.6211946	0.6702727	0.5348064	0.5400681	0.5562939	0.6488380	0.3024742	0.7925536	1.0000000	0.8546047	0.6947106
bpm	-0.3694010	0.6879402	0.7291244	0.7506563	0.7186540	0.7389880	0.7164928	0.7309202	0.6070575	0.6074214	0.8301901	0.8546047	1.0000000	0.8399475
vorp	-0.2671223	0.5464172	0.5882270	0.6082337	0.5637302	0.6032998	0.5774584	0.5807501	0.4449988	0.5041587	0.7138212	0.6947106	0.8399475	1.0000000

Results of Spearman's Coefficient Matrix

The scatter plots reveal a moderate negative correlation between draft position and career games, as well as between draft position and career win shares. This indicates that players drafted are given more opportunities to play in their careers, which leads them to accumulate win shares over time. However, both plots show high variability between individual career outcomes, especially amongst high draft picks. Overall, the scatter plots indicate that draft position is a significant but imperfect predictor of career longevity and productivity. About the correlation matrices above, there wasn't much of a difference at all between Pearson's and Spearman's. The draft position in both is shown to have a moderate negative correlation with every other attribute in the dataset, except for itself, which supports the claims expressed by the aforementioned scatter plots. Strong positive correlations between a few variables do exist, however, especially between career volume and production metrics, such as total minutes played and the advanced analytics of BPM, VORP, Win Shares, and Win Shares per 48 minutes. Overall, these findings suggest that career longevity and productivity complement one another, and that draft order influences career outcomes through playing time and long-term opportunities, which is something that top draft picks can claim for themselves.

For the fourth and fifth questions of accurately predicting the number of career games played from different draft-related attributes and the extent draft order predicts how many

minutes a player played per game and in total over their whole career, multiple linear regression delivered results from three different models, where total minutes played, total games played, and minutes per game were the dependent variables for each of their own models. The results are below:

```
> summary(model_total_games)

Call:
lm(formula = total_games_played ~ overall_pick + pick_bucket +
    educational_prestige + team_conference, data = nbaDraftRegressionGamesPlayed)

Residuals:
    Min      1Q  Median      3Q     Max
-621.31 -198.56  -88.96  190.86 1139.15

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     678.372     29.367  23.100  < 2e-16 ***
overall_pick                     -4.721      1.198  -3.940 8.48e-05 ***
pick_bucketLottery             -121.930     29.268  -4.166 3.26e-05 ***
pick_bucketMidtoLateFirst      -196.716     34.719  -5.666 1.72e-08 ***
pick_bucketSecondRound         -258.775     54.231  -4.772 1.99e-06 ***
educational_prestigeOtherCollege  -26.236     20.371  -1.288    0.198
educational_prestigePowerSchool   -26.394     25.196  -1.048    0.295
team_conferenceWestern             17.611     14.183   1.242    0.215
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 288.4 on 1661 degrees of freedom
Multiple R-squared:  0.2151,    Adjusted R-squared:  0.2118
F-statistic: 65.04 on 7 and 1661 DF,  p-value: < 2.2e-16
```

Summary of the MLR model for predicting total games played

```
> summary(model_total_minutes)

Call:
lm(formula = total_minutes_played ~ overall_pick + pick_bucket +
    educational_prestige + team_conference, data = nbaDraftRegressionMinPlayed)

Residuals:
   Min     1Q Median     3Q    Max
-19110  -5014  -2432   3628  37990

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      20628.51     868.78  23.744  < 2e-16 ***
overall_pick                      -126.74      35.45  -3.575  0.00036 ***
pick_bucketLottery               -6110.25     865.84  -7.057 2.49e-12 ***
pick_bucketMidtoLateFirst        -9109.12    1027.09  -8.869  < 2e-16 ***
pick_bucketSecondRound          -10567.12    1604.34  -6.587 6.02e-11 ***
educational_prestigeOtherCollege   -863.51     602.64  -1.433  0.15208
educational_prestigePowerSchool    -976.54     745.38  -1.310  0.19033
team_conferenceWestern              434.23     419.58   1.035  0.30086
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8533 on 1661 degrees of freedom
Multiple R-squared:  0.252,     Adjusted R-squared:  0.2489
F-statistic: 79.96 on 7 and 1661 DF,  p-value: < 2.2e-16
```

Summary of the MLR model for predicting total minutes played

```
> summary(model_MPG)

Call:
lm(formula = mpg ~ overall_pick + pick_bucket + educational_prestige +
    team_conference, data = nbaDraftRegressionMPG)

Residuals:
     Min       1Q   Median       3Q      Max
-19.2059  -5.3153  -0.3314   4.9660  23.3041

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     29.77968    0.71824  41.462  < 2e-16 ***
overall_pick                    -0.18065    0.02931  -6.164 8.86e-10 ***
pick_bucketLottery              -4.75166    0.71581  -6.638 4.29e-11 ***
pick_bucketMidtoLateFirst       -7.40581    0.84911  -8.722  < 2e-16 ***
pick_bucketSecondRound          -8.23413    1.32634  -6.208 6.76e-10 ***
educational_prestigeOtherCollege -0.40821    0.49821  -0.819    0.413
educational_prestigePowerSchool   0.32215    0.61622   0.523    0.601
team_conferenceWestern           -0.03467    0.34687  -0.100    0.920
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.054 on 1661 degrees of freedom
Multiple R-squared:  0.3464,    Adjusted R-squared:  0.3437
F-statistic: 125.8 on 7 and 1661 DF,  p-value: < 2.2e-16
```

Summary of the MLR model for predicting minutes per game

Overall, the multiple linear regression analysis reveals that draft position remains the strongest and most reliable predictor of NBA career outcomes. Across all three models, overall draft pick and draft pick bucket were statistically significant, with earlier selections associated with longer careers, greater cumulative playing time, and higher average minutes per game. In contrast, educational prestige and team conference were not statistically significant predictors in any of the models once draft position was controlled for, suggesting that collegiate pedigree and conference provide minimal additional explanatory power beyond draft capital. This is consistent with the other findings in this project so far. Model performance improved when R^2 jumped

from 21.5% and 25.2% in the first two models to 34.6% in the third model, indicating that draft position most strongly influences how consistently players are utilized on a minutes-per-game basis rather than solely how long they remain in the league. Overall, these findings continue to reinforce the central role of draft capital in shaping professional basketball careers, especially for high draft picks.

For the fifth and final question of the possibility of accurately labeling which draft picks will have long-term careers, the constructed decision tree and random forest classification models provide the following results:

```
> confusionMatrix(prediction_30, test_30$long_career_flag)
Confusion Matrix and Statistics

               Reference
Prediction      LongCareer NotLongCareer
  LongCareer            48            46
  NotLongCareer        118           287

               Accuracy : 0.6713
                 95% CI : (0.6282, 0.7124)
    No Information Rate : 0.6673
    P-Value [Acc > NIR] : 0.4454

                  Kappa : 0.1694

 Mcnemar's Test P-Value : 2.954e-08

            Sensitivity : 0.28916
            Specificity : 0.86186
         Pos Pred Value : 0.51064
         Neg Pred Value : 0.70864
             Prevalence : 0.33267
         Detection Rate : 0.09619
   Detection Prevalence : 0.18838
      Balanced Accuracy : 0.57551

       'Positive' Class : LongCareer
```

Decision Tree 70/30 Split Matrix

```
> confusionMatrix(prediction_20, test_20$long_career_flag)
Confusion Matrix and Statistics

               Reference
Prediction      LongCareer NotLongCareer
  LongCareer            39            33
  NotLongCareer         72           189

               Accuracy : 0.6847
                 95% CI : (0.6318, 0.7343)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.2625713

                  Kappa : 0.2222

 Mcnemar's Test P-Value : 0.0002086

            Sensitivity : 0.3514
            Specificity : 0.8514
         Pos Pred Value : 0.5417
         Neg Pred Value : 0.7241
             Prevalence : 0.3333
         Detection Rate : 0.1171
   Detection Prevalence : 0.2162
      Balanced Accuracy : 0.6014

       'Positive' Class : LongCareer
```

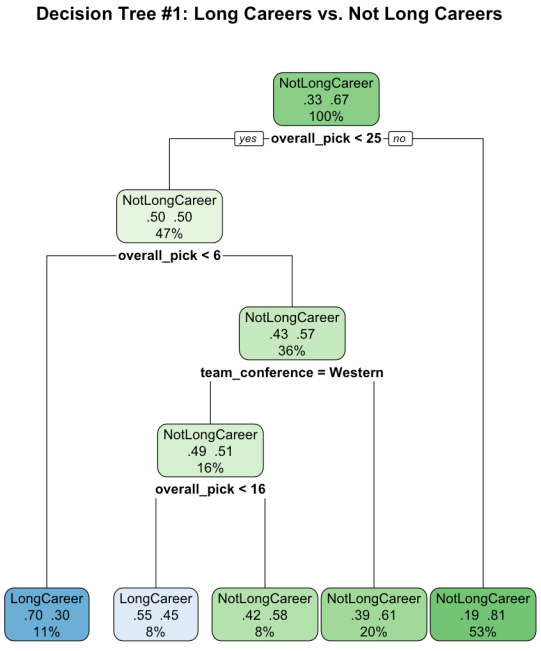
Decision Tree 80/20 Split Matrix

```
> confusionMatrix(prediction_30_rf, test_30$long_career_flag)
Confusion Matrix and Statistics

               Reference
Prediction      LongCareer NotLongCareer
  LongCareer            44            52
  NotLongCareer        122           281

               Accuracy : 0.6513
                 95% CI : (0.6077, 0.6931)
    No Information Rate : 0.6673
    P-Value [Acc > NIR] : 0.7909

                  Kappa : 0.1218

 Mcnemar's Test P-Value : 1.687e-07

            Sensitivity : 0.26506
            Specificity : 0.84384
         Pos Pred Value : 0.45833
         Neg Pred Value : 0.69727
             Prevalence : 0.33267
         Detection Rate : 0.08818
   Detection Prevalence : 0.19238
      Balanced Accuracy : 0.55445

       'Positive' Class : LongCareer
```

Random Forest 70/30 Split Matrix

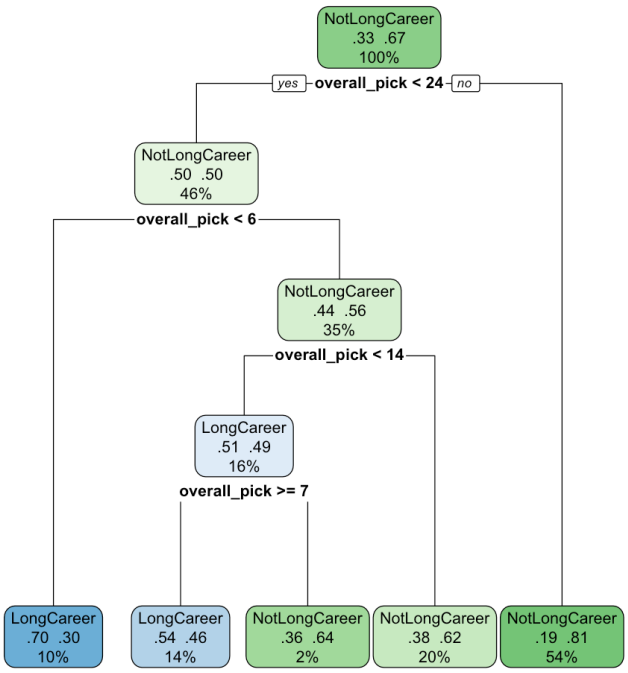
```
> confusionMatrix(prediction_20_rf, test_20$long_career_flag)
Confusion Matrix and Statistics

               Reference
Prediction      LongCareer NotLongCareer
  LongCareer            30            30
  NotLongCareer         81           192

               Accuracy : 0.6667
                 95% CI : (0.6132, 0.7171)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.5257

                  Kappa : 0.1527

 Mcnemar's Test P-Value : 2.077e-06

            Sensitivity : 0.27027
            Specificity : 0.86486
         Pos Pred Value : 0.50000
         Neg Pred Value : 0.70330
             Prevalence : 0.33333
         Detection Rate : 0.09009
   Detection Prevalence : 0.18018
      Balanced Accuracy : 0.56757

       'Positive' Class : LongCareer
```

Random Forest 80/20 Split Matrix

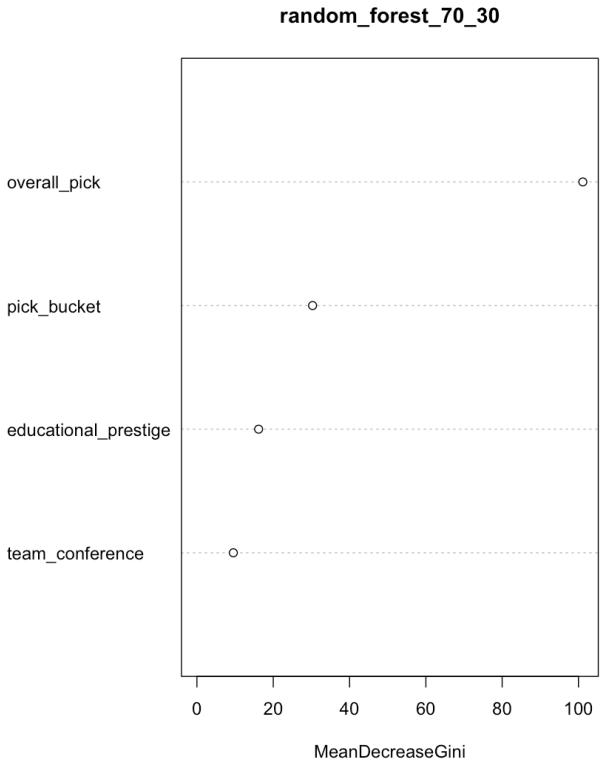


Decision Tree # 1 on 70/30 Split

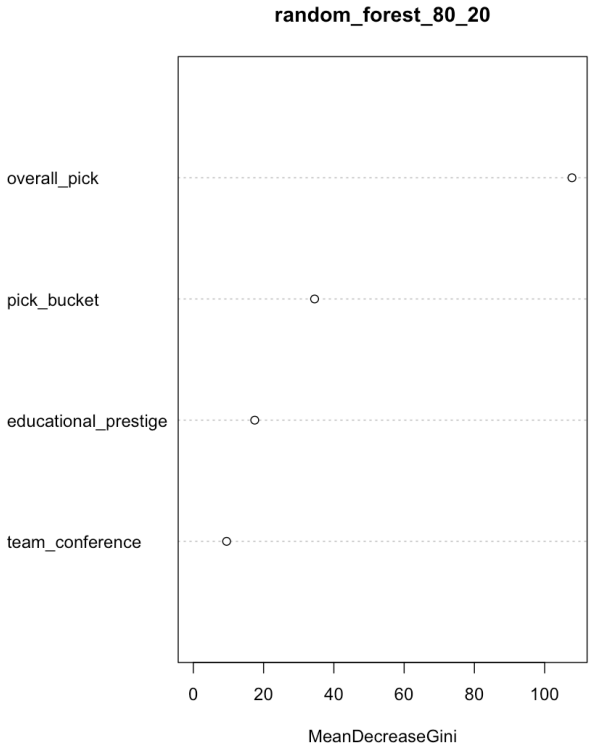
Decision Tree #2: Long Careers vs. Not Long Careers



Decision Tree # 1 on 80/20 Split



Random Forest Variable Importance Plot on 70/30 Split



Random Forest Variable Importance Plot on 80/20 Split

Across all four iterations of classification (70/30 split and 80/20 split each), model accuracy ranged between 65% and 68%, which is slightly above the baseline no-information rate of 67%. Additionally, Cohen's kappa values are extremely low across the board. This means that the models are just slightly better than guessing the majority class every time, which is "notLong Career." All models showed a high range of specificity, ranging around 85%, while displaying a low sensitivity at the same time, between 26% and 35%. The models had a great time identifying players that weren't likely to have long careers, but struggled mightily when it came to identifying players that were likely to have long careers. The decision tree plots and random forest variable importance plots continue to show how overall draft position is the highest and

most dominant predictor of a long career, with the other draft-related attributes providing very minimal contributions.

In conclusion, while draft positioning has proven to be the strongest and most dominant predictor of long-term NBA careers in this dataset, the classification analysis revealed that, by itself, it is insufficient to predict long-term NBA careers accurately. In addition, this project also revealed that the other draft-related attributes, such as educational prestige and the team that drafts a player, provide minimal contribution and are completely overshadowed by draft positioning. In the future, further work can be done on this project through a Python-based implementation, as opposed to an R-based implementation, and incorporating additional context, such as draft-day trades, and considering additional variables like injuries to improve model performance beyond just draft-day data alone.

Works Cited

"Apriori Algorithm." *GeeksforGeeks*, 21 Nov. 2025, www.geeksforgeeks.org/machine-learning/apriori-algorithm/.

"Association Rule." *GeeksforGeeks*, 8 Sept. 2025, www.geeksforgeeks.org/machine-learning/association-rule/.

"Decision Tree." *GeeksforGeeks*, GeeksforGeeks, 30 June 2025, www.geeksforgeeks.org/machine-learning/decision-tree/.

"Feature Importance with Random Forests." *GeeksforGeeks*, 11 Nov. 2025, www.geeksforgeeks.org/machine-learning/feature-importance-with-random-forests/.

Fotache, Marin, et al. "High-Level Machine Learning Framework for Sports Events Ticket Sales Prediction." *ACM Digital Library*, 7 Oct. 2021, dl.acm.org/doi/10.1145/3472410.3472426.

Hayes, Adam. "Multiple Linear Regression (MLR): Definition, Formula, and Example." *Investopedia*, Investopedia, 14 Apr. 2025, www.investopedia.com/terms/m/mlr.asp.

Hill, Matt. "NBA Draft Basketball Player Data 1989-2021." *Kaggle*, 26 May 2022, www.kaggle.com/datasets/mattop/nba-draft-basketball-player-data-19892021/data.

Kavlakoglu, Eda. "What Is Random Forest?" *IBM*, www.ibm.com/think/topics/random-forest. Accessed 20 Dec. 2025.

"Matt Hill - University of Pennsylvania Perelman School of Medicine." *LinkedIn*, www.linkedin.com/in/matt-hill-ds/. Accessed 20 Dec. 2025.

"Matt Op's Profile." *Kaggle*, www.kaggle.com/mattop. Accessed 20 Dec. 2025.

Norlander, Matt, et al. "The Greatest College Basketball Programs Ever: Ranking the Top Teams of All Time." *CBS Sports*, 19 Nov. 2020, www.cbssports.com/college-basketball/news/the-greatest-college-basketball-programs-ever-ranking-the-top-teams-of-all-time/.