Beyond the Hype: Predicting the 2023 NBA Draft Class

Caleb Miller

September 7, 2023

Executive Summary:

This project utilizes web scraping, data analysis, and machine learning to predict each player's Player Efficiency Rating (PER) from the 2023 NBA Draft Class. To accomplish this, data sets containing NBA player statistics were scraped from the internet and cleaned. This data was analyzed, revealing:

1. College players with higher field-goal percentages, blocks, and rebounds typically have higher NBA Career PERs.
2. On the contrary, college players with higher free-throw percentages and three-point percentages typically have lower NBA Career PERs.

The same data was used to train and test different machine-learning models. Three of the models were capable of predicting greater than 26% of the variation in PER. While these are not the strongest, lower predictive power is expected with models dealing with human behavior. Finally, college data from the 2023 NBA Draft Class was fitted to these models to provide predictions. The findings from these models are:

1. Many college prospects were drafted far too late or went undrafted, thus deeming them undervalued. Specifically, Trayce Jackson-Davis, Drew Timme, and Jalen Slawson were undervalued in the draft but were predicted to be top-5 players according to the models' PER predictions.
2. On the contrary, many college prospects are overvalued. Specifically, Brandon Miller and Jordan Hawkins were highly valued in the draft, but the models project they will not live up to the hype that comes along with being a lottery pick.

The data used for training and testing of these models includes 898 observations of 9 variables including "Points", "Assists", and "Steals". The athletes played between the 1982-83 and 2022-23 NBA seasons.

The following sections will guide the data collection, cleaning, analysis, and training/testing processes while highlighting flaws with the model and reinforcing the true value of utilizing data to derive valuable insights for scouting.

Data Collection:

The data used to train, test, and fit the models was collected from https://www.basketball-reference.com/ and https://basketball.realgm.com/ using various web scrapers. This data includes all NBA athletes who played in the 1982-83, 1992-93, 2002-03, 2012-13, and 2022-23 seasons. These seasons are included in 10-year intervals to represent different eras of the NBA and represent a more diverse pool of players. The first season included is 1983 because modern statistics were first tracked in the 1979-80 season.

This project utilized Python's Selenium, Requests, and BeautifulSoup4 packages to collect the NBA Player's data from https://www.basketball-reference.com/. First, the names, profile URLs, and ages of all NBA players who played in the seasons mentioned above were collected from the season's player directory page. Next, the following college statistics of those athletes were collected from their individual player statistics page:

- Games Played, Minutes, Points, Offensive Rebounds, Total Rebounds, Assists, Steals, Blocks, Field-Goals Made, Field-Goals Attempted, Field-Goal Percentage, Three-Pointers Made, Three-Pointers Attempted, Three-Point Percentage, Free-Throws Made, Free-Throws Attempted, and Free-Throw Percentage

The college statistics collected were from each player's Freshman year. Lastly, the athlete's NBA career PERs were collected. Athletes who played internationally or entered the NBA following high school without attending college are not included in this data.

To collect statistics for the 2023-2024 NBA draft class, Python's requests and beautifulsoup4 packages were used. The same statistics collected for NBA Players were also collected for prospects from https://basketball.realgm.com/. Prospect statistics were gathered using Requests since it is a more efficient collection method, and the data was not loaded via JavaScript. Prospects who played in the G-League are not included in this dataset.

Data Cleaning and Preparation:

The NBA Player data needed cleaning and was merged into one dataset. To accomplish this, the four PER files were concatenated into one, the four College Statistics files were concatenated into one, and players were removed from the data if they were duplicates or played college basketball when three-pointers or offensive rebounds were not tracked. Then, the PER and College stats files were merged into a single file. The data was cleaned, and the profile URL, age, and name columns were dropped due to being unnecessary at this point. Additionally, outliers and unrealistic data entries were excluded from the data (22 players – i.e. Stanley Umude played 2 minutes in the NBA and had the highest PER of the dataset at 65.6). Lastly, any entries with missing values were omitted leaving the cleaned dataset with 898 rows of 17 columns.

NBA PER Tier Definition:

For the purpose of this project, these are the tiers of success used for evaluation:

Career PER	Type of Player
Less than 9	Won't stick around in the NBA
9 - 11	Fringe roster player
11 - 13	Will get a roster spot
13 - 15	Rotation player
15 - 17	Good role player
17 – 22	All-Star
22+	Superstar

Summary statistics were helpful when determining tier levels. The average PER aligns with a 'rotation player'. Additionally, the lower and upper quartiles of the data coincide with a 'fringe roster player' and a 'good role player'. Each season, 24 players become NBA All-Stars. With approximately 450 active players, this club is exclusive as only 5% of NBA athletes get this honor each season. However, many players do not get this honor every year due to fierce competition, voter fatigue, injuries, and other factors. Therefore, the 'all-star' and 'superstar' tiers will make up approximately 15% of the data.

Data rows by PER tier:

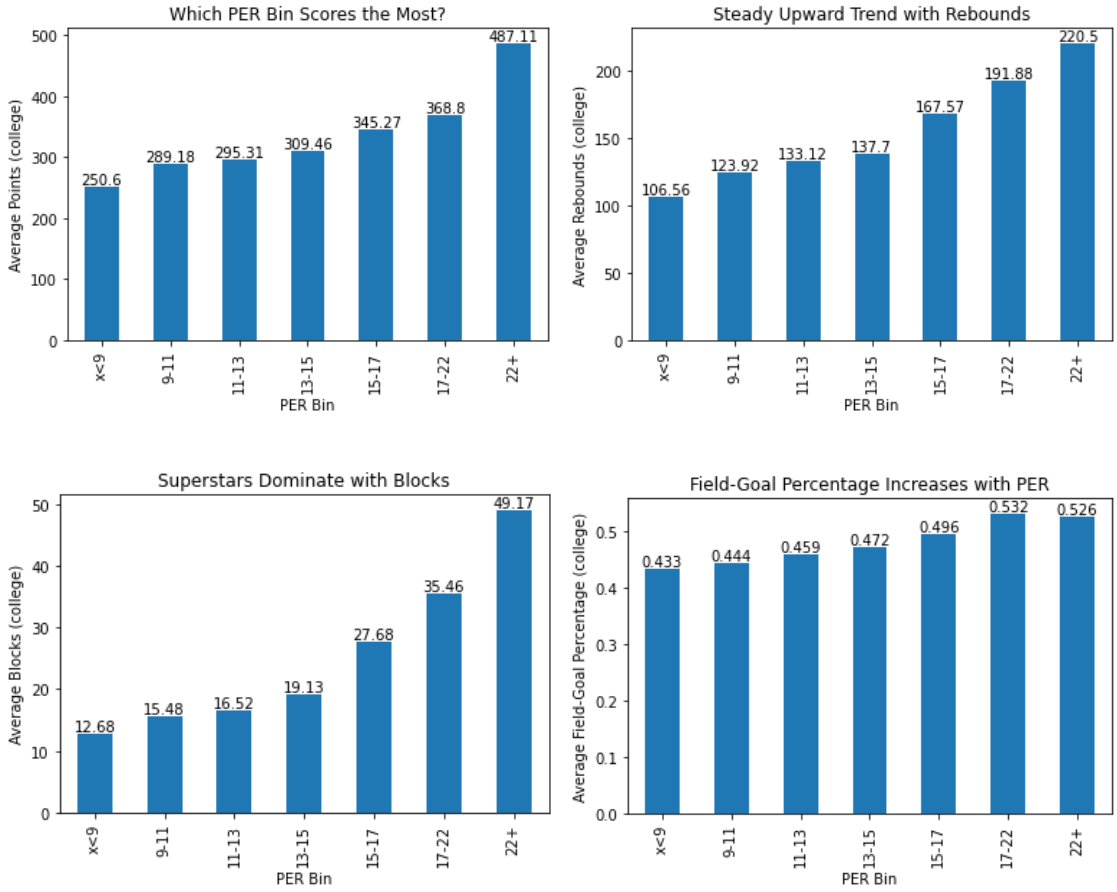
Type of Player	Data Rows
Won't stick around in the NBA	85
Fringe roster player	126
Will get a roster spot	189
Rotation player	197
Good role player	131
All-Star	118
Superstar	17

Data Analysis:

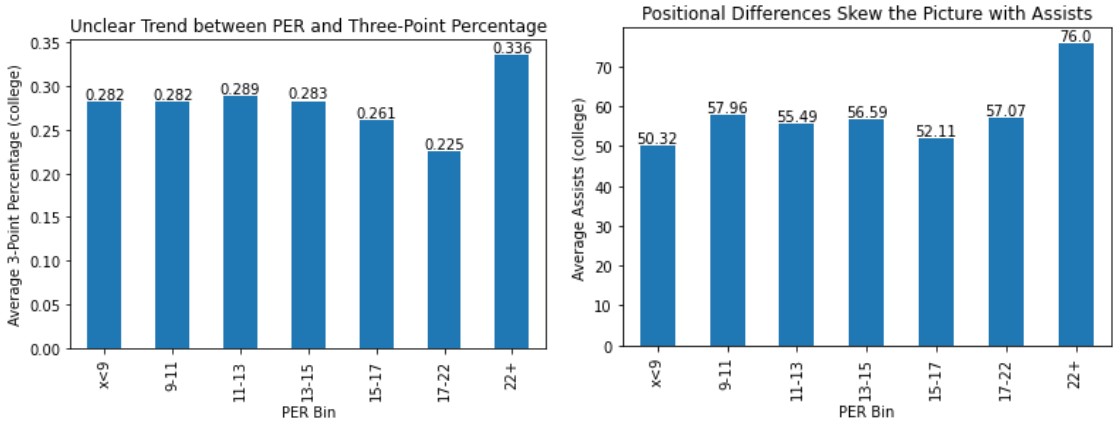
The median PER of the data is 13.2 while the mean is 13.29 indicating a normal distribution. An average player from this dataset is categorized as a 'rotation player' according to the tier definitions. For comparison, the 25th percentile PER is 11.1 which is categorized as a 'player that will get a roster spot'. Additionally, the 75th percentile PER is 15.6 which is categorized as a 'good role player'.

A correlation matrix was created to uncover strong relationships between variables. Field-goal percentage has the highest correlation with PER at 0.4. Coming in second is blocks (0.36) and third strongest is rebounds (0.35). The strongest negative correlations with PER include three-pointers made (-0.1), three-pointers attempted (-0.12), three-point percentage (-0.082) (which is extremely logical considering the first two), and Free-Throw Percentage (-0.039).

Players were then assigned to bins to analyze statistics over different levels of PER:



A deeper evaluation of variables with strong positive correlations with PER reveals that the trend between points and PER is positive. Most notably, the bin of ‘Superstar’ players scored 29.64% more points than the second-closest bin. Equally as clear is the positive trend between PER and rebounds. Similarly, ‘Superstar’ players grabbed 14.92% more rebounds than the second-closest bin. Although there are large increases at the highest level, one will notice that there is a constant upward trajectory. The trend of ‘Superstar’ player domination becomes even more apparent between blocks and PER. The bin of ‘Superstar’ players had the largest increase bin over bin at 38.66%. Field-goal percentage has the highest correlation with PER and holds strong predictive power. The trend between these variables is clearly positive with the exception of ‘Superstar’ players. This is likely due to ‘superstar’ athletes taking more shots which typically lowers field-goal percentage.



Some relationships are less clearly positive or negative. For example, the relationship between three-point percentage and PER is negative until three-point Percentage trends upward with ‘Superstar’ players. Another example is between assists and PER where there is no clear trend. The data shifts erratically, and it can be inferred that this is due to positional needs. Guards typically account for the vast majority of assists whereas forwards and centers do not.

In testing, it was discovered that minutes, field-goals made, field-goals attempted, free-throws made, free-throws attempted, three-pointers made, and three-pointers attempted statistics need to be dropped due to multi-collinearity. Therefore, the models were trained to predict PER from games, points, rebounds, assists, steals, blocks, field-goal percentage, three-point percentage, and free-throw percentage.

Model Training and Performance:

Model Name	R Squared	Root Mean Squared Error	Mean Absolute Error
Quantile	0.272857	3.346237	2.582522
Lasso	0.272589	3.346854	2.583629
Linear Regression	0.262390	3.370237	2.598229
XGBoost	0.241699	3.417178	2.621969
Bayesian Ridge	0.236643	3.428552	2.643920
Random Forest	0.224650	3.455381	2.633166
KNN	0.150318	3.617222	2.713222
Decision Tree	0.116342	3.688832	2.869624
SVR	0.016237	3.892173	2.994967
Neural Network MLP	-0.035017	3.992278	2.988547

The data is split into two datasets. One holds 75% of the total data and will be used to train these models. The other holds the remaining 25% of the data and will be used to test the models. To find the best-fitting model for the data, 10 models were trained using the Sklearn package within Python. Three different metrics were used to measure model fit: R-squared, Root Mean Squared Error, and Mean Absolute Error. Some models performed abysmally such as the Neural Network MLP and SVR. However, predictive power increases above 10% of variation in models such as Decision Tree and K-Nearest Neighbors Regression Models. Ultimately, six of the models are able to predict more than 20% of the variation in PER. The top three performing models are:

- Quantile Regression: Predicts 27.29% of the variation in PER

- Lasso Regression: Predicts 27.26% of the variation in PER

- Linear Regression: Predicts 26.24% of the variation in PER

All of the model's predictive power is too weak to be considered for a scientific study. However, R-squared values of less than 50% are common in studies that attempt to predict human behavior. These models can still be used to derive insights into what to look for in an NBA draft prospect.

NBA Rookie Career PER Prediction:

The top 3 performing models were fit with prospect data to predict the athlete's NBA Career PER:

- Quantile Regression:

Ranking in Draft Class	Drafted	Name	Predicted Career PER
1	57	Trayce Jackson-Davis	19.39
2	1	Victor Wembanyama	18.69
3	12	Dereck Lively II	17.80
4	Undrafted	Drew Timme	17.25
5	54	Jalen Slawson	17.24
23	9	Taylor Hendricks	15.08
28	6	Anthony Black	14.57
30	8	Jarace Walker	14.39
35	7	Bilal Coulbaly	14.22
41	2	Brandon Miller	13.97
42	10	Cason Wallace	13.94
71	13	Gradey Dick	12.90
86	11	Jett Howard	12.05
90	14	Jordan Hawkins	11.97

- Lasso Regression:

Ranking in Draft Class	Drafted	Name	Predicted Career PER
1	57	Trayce Jackson-Davis	19.39
2	1	Victor Wembanyama	19.00
3	12	Dereck Lively II	17.56
4	54	Jalen Slawson	17.12
5	Undrafted	Drew Timme	17.09
21	9	Taylor Hendricks	15.17
27	6	Anthony Black	14.57
28	8	Jarace Walker	14.49
36	2	Brandon Miller	14.11
39	7	Bilal Coulbaly	14.09
43	10	Cason Wallace	13.90
72	13	Gradey Dick	12.91
87	11	Jett Howard	12.15
90	14	Jordan Hawkins	12.07

- Linear Regression:

Ranking in Draft Class	Drafted	Name	Predicted Career PER
1	57	Trayce Jackson-Davis	19.37
2	1	Victor Wembanyama	18.58
3	12	Dereck Lively II	17.89
4	Undrafted	Drew Timme	17.28
5	54	Jalen Slawson	17.28
23	9	Taylor Hendricks	15.06
28	6	Anthony Black	14.56
32	8	Jarace Walker	14.35
35	7	Bilal Coulbaly	14.25
41	10	Cason Wallace	13.95
42	2	Brandon Miller	13.92
71	13	Gradey Dick	12.90
86	11	Jett Howard	12.03
89	14	Jordan Hawkins	11.94

Interestingly, the highest predicted PER is less than 20. This would categorize the highest-ranked players as 'all-stars'. This differs from public opinion as the 2023 NBA draft class has been hyped and even called the best draft class since 2003 (The draft class with Lebron James, Carmelo Anthony, Dwayne Wade, Chris Bosh, and many other notable players). However, this could be indicative that this draft class was overhyped. For the analysis of the predicted PER results, one should pay less attention to the actual predicted PER, and more attention to the athlete's ranking among the draft class. The models predict that many players were drafted too late if they were drafted at all:

- Trayce Jackson-Davis stood out among the draft class and ranked number 1 in all three of our best predictive models. He was projected to be a second-round pick and got drafted 57th overall. The predictive models say that he is a steal of a draft pick.

- Drew Timme ranked number 4 in two of the models and number 5 in the other. He was projected to be a late second-round pick and went undrafted. This is another instance where the models revealed much greater value in a player than scouts.

- Jalen Slawson ranked number 4 in one of the models and number 5 in the other two. He was a projected second-round pick and got drafted 54th overall. Jalen Slawson's value is incredibly high according to these models.

These models also predict that many players were drafted too high:

- Brandon Miller ranked surprisingly low at 36th, 41st, and 42nd among draft prospects. He was a projected top-3 pick and was drafted 2nd overall. Brandon Miller's value is low for where he was drafted according to the models.

- Jordan Hawkins consistently ranked last among lottery selections in predicted career PER (90th in two models and 89th in the other). He was a projected first-round pick with the potential to go at the end of the lottery and was drafted 14th overall. According to these models, Jordan Hawkins' value is very low for a lottery pick.

Flaws with the Models:

It is highly probable that the presence of unquantifiable statistics such as work ethic and overall potential lead to flaws in these predictive models. This is a common issue when predicting human behavior. Many players had to be removed from the dataset because their college did not track statistics such as 'Three-Point Percentage', 'Minutes', and 'Steals'. Additionally, there were players whose statistics were unavailable on basketballreference.com because they entered the NBA straight from high school or played professionally overseas or in the G-League. A better data source would likely increase the models' predictive power.

Key Takeaways:

All models hold predictive power and can aid with difficult decisions. However, since these models' predictive power is not extremely strong, these takeaways are to be taken with a grain of salt. The key takeaways of the analysis of this data are:

- The strongest positive correlations with PER are field-goal percentage, blocks, and rebounds.

- The strongest negative correlations with PER are three-point percentage and free-throw percentage.

The key takeaways of the training of these models are:

- Trayce Jackson-Davis, Drew Timme, and Jalen Slawson project to be top-5 players from the 2023 NBA Draft class. This was unforeseen as these athletes were drafted nowhere near the top 5. (54^{th}, 57^{th}, and Undrafted)

- Victor Wembanyama will be great. However, these models predict he won't be the best player in this draft class.

- Brandon Miller was drafted far too high, as these models predict he will be in the middle of this draft class PER-wise.

- Of the lottery picks from the 2023 NBA Draft, Jordan Hawkins consistently ranked last in predicted Player Efficiency Rating.

Trayce Jackson-Davis might not end up being the best player from this draft class - realistically, he won't. That title is probably going to be Victor Wembanyama's, but the fact that the models predict him to be number one implies he could be the steal of the draft. Other significant picks include Drew Timme and Jalen Slawson who could be diamonds in the rough. Obtaining one of these players could greatly impact the success of a team. One of the most famous examples of this was when the Denver Nuggets selected Nikola Jokic 41st overall in the 2014 NBA Draft. This pick paid off as he has collected tons of accolades and one championship. Not everyone can become league MVP, however, so a more realistic example is Draymond Green. He was drafted 35th overall and has been integral to the success of the Warriors dynasty collecting accolades and championships along the way. Additionally, Brandon Miller and Jordan Hawkins might not end up being busts, but these models predict they may not live up to the expectations of the number 2 and 14 overall picks. Their career trajectory could follow that of Markelle Fultz who was drafted 1st overall in 2017 and has not lived up to the hype. Fultz is still a quality NBA player but clearly did not warrant the 1st overall pick. Time tells all, but machine learning models are useful tools for predicting NBA success beforehand and can be utilized by executives to drive team success.

Work Cited:

RealGM LLC. (n.d.). *Basketball News, rumors, scores, stats, analysis, depth charts, forums.* Basketball.RealGM.com. https://basketball.realgm.com/

Sports Reference LLC. (n.d.). *Basketball Statistics & History.* Basketball-Reference.com. https://www.basketball-reference.com/