Drafting Success: Predicting NBA Success Using College Performance Metrics

Stephen Yu : 405570842[a]

[a] *University of California, Los Angeles,*

Abstract

This study explores the predictive power of college basketball statistics in forecasting NBA All-Star selections. Utilizing a comprehensive dataset encompassing various collegiate performance metrics, I employed four supervised learning models: Logistic Regression, K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA). The objective was to ascertain which statistics, if any, serve as significant predictors for identifying future NBA All-Stars. My results revealed that out of the original 26 collegiate statistics, 9 were found to be significant. Solely using these 9 independent variables, I was able to predict whether a drafted player would become an NBA All-Star with up to 92% accuracy.

1. Introduction

On May 22nd, 2009 the Minnesota Timberwolves, an NBA franchise that has seen little success, hired David Kahn as general manager. Kahn's first task was to draft a guard in the upcoming 2009 NBA draft to pair with star forward Kevin Love. This was a momentous moment that could not only make or break his future as GM, but shift the paradigm of the league. Luckily for him, the Timberwolves not only owned their pick at 6 but also had drafting rights to Washington's pick at 5. With Minessota on the clock, and with the rare opportunity to get back to back chances for success Kahn selected Spain guard Ricky Rubio and Syracuse guard Jonny Flynn. One pick later, Golden State chose Davidson guard, Steph Curry.

Rubio and Flynn would never get close to stardom, while Curry became a future first-ballot Hall of Famer, 4 time champion, 2 time league MVP, and made the NBA's top 75 players of all time list. After 3 losing seasons in Minnesota, David Kahn was fired, and since 2009 the Timberwolves have embodied their role as the league's ”poverty” franchise while the Warriors have ascended into a dynasty.

Owners, general managers, coaches, players, and fan bases have all suffered from poor draft choices. The magnitude of making the correct selection while your team is on the clock cannot be overstated. Yet, much like a company's stock price, true player potential seems nearly impossible to predict. So many metrics go into an NBA player and balancing all of these traits is arduous. NBA scouts approach this by watching hundreds of games, identifying player habits, and measuring strengths and weaknesses in their game to construct a thorough draft portfolio. This paper approaches this task quantitatively with the aim being to predict NBA success solely from collegiate stats.

2. Literature Search

To be able to predict an NBA player success from their collegiate stats is the holy grail of drafting. For this reason, a diverse range of approaches emerge from the literature.

The first article from DraftExpress [4] provides a narrative-style analysis on specific players. This article focuses on pre-draft statistics and their correlation with NBA performance. However, its approach is less systematic and more anecdotal, lacking the academic rigor typically associated with predictive sports analytics. It seems more oriented towards a general audience and does not delve deeper into statistical methodologies that would be of interest to an academic audience in sports analytics. Despite it being completely non-academic, its similarities to this paper's goals highlights the wide range of interested parties in this particular field.

The paper from the SMU Data Science Review, ”Predicting National Basketball Association Success: A Machine Learning Approach,” [6] shares similarities with my research in its use of statistical models to predict NBA success. However, a key difference lies in their inclusion of draft position as a predictor. This inclusion can be seen as counterintuitive to pre-draft analysis, as it introduces post-draft information into the model. Not surprisingly, they found that draft position was the leading predictor in NBA success. My study focuses exclusively on college performance data without the influence of draft position.

3. Data Collection and Pre-Processing

This study has a clear and quantitative measure for NBA success which is defined as any player that has been selected to at least 1 all star team. The data set is broken up into players that fall under this category and players that do not. While various measurements of success could be used, I personally believe that being selected to 1 all star team is the cutoff between a star and non-star.

It is important to note that the data set only encompasses players that have played at least 1 game in the NBA, and have played at least 1 game for a recognized NCAA team. While there are many players in the NBA, even some who become all-stars, that come from backgrounds such as oversees, G League, or straight from high school, including these players statistics would taxing and disingenuous. Every pre-NBA league is different and the stats between these vastly different leagues should not be considered as equivalent because of difference of competition level, different rules, and different pace of game.

The primary source of collegiate data was the exhaustive repository available at Basketball-Reference.com [1], renowned for its detailed and extensive statistical records. I scraped career totals of all available categories for each player for a total of 4577 players.

To determine All-Star selections, I utilized Wikipedia's ”List of NBA All-Stars” page [2]. This list provided a definitive record of players who achieved this accolade, allowing me to effectively label our dataset with the response variable – whether a player became an NBA All-Star (1) or not (0).

3.1. MICE

The data set generated had a significant problem because it had many incomplete entries. More specifically, older players, players that did not shoot 3 pointers, and players coming from lesser known colleges all had missing values in at least 1 statistic. Keeping only the complete rows cut the data set from 4577 to 815 observations. I decided to strike a middle ground, by keeping mostly filled rows and using imputation to fill out the remaining missing values. This lead to a data set containing 2490 players.

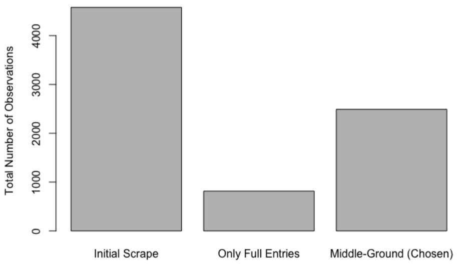

Figure 1: Comparison of the Three Data Set Options

The imputation method was Multiple Imputation by Chained Equations (MICE) [7], a robust approach that leverages machine learning models to predict missing values. MICE works under the premise of creating multiple imputations for missing data, which involves iteratively filling in missing values using predictive models based on the observed data.

This approach is particularly suited for complex data sets with interdependencies among variables. In the first step, the algorithm generates initial guesses for missing values, typically using simple methods like mean imputation. Following this, it cycles through each variable with missing data, treating it as a dependent variable predicted by all other variables. These steps are repeated for several iterations, refining the imputation with each cycle [9]. By adopting the MICE methodology, I was able to mitigate potential biases and data loss that could have arisen from excluding players with incomplete statistics.

3.2. SMOTE

When one class dominates the other class, there are concerns about model becoming biased toward the majority class because of overexposure. Our dataset suffers from class imbalance, indicated by the barplot below:

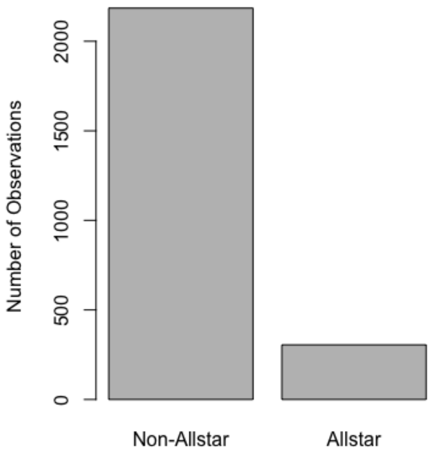

Figure 2: Count of All-Stars to Non-All-Stars in the Data (non-SMOTE)

This imbalance is because the majority of NBA basketball players do not become All-Stars. To address this imbalance, I implemented Synthetic Minority Oversampling Technique (SMOTE) [3].

SMOTE works by creating synthetic samples from the minority class (in this case, players who became All-Stars) instead of creating copies. This technique helps in balancing the dataset by generating new instances that are a convex combination of neighboring instances. SMOTE is particularly beneficial as it goes beyond simple under- or over-sampling methods, which can lead to model overfitting or the exclusion of important instances, respectively [5].

Implementing SMOTE lead to 2085 synthetic data points being added to the minority class leading to a total of 2390 added data points for all-star selected players. Below is the updated class balance:

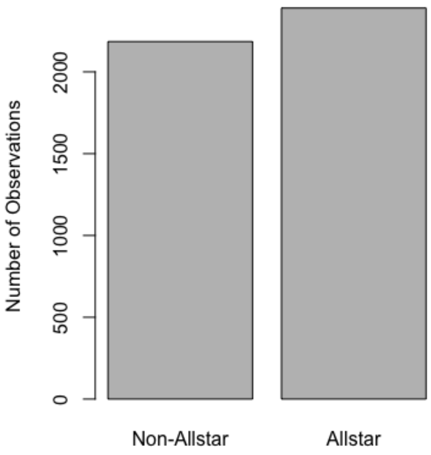

Figure 3: Count of All-Stars to Non-All-Stars in the Data (SMOTE)

Comparing Figure 3 to the Figure 2, we can see that the data imbalance is largely gone. However, while SMOTE is a powerful tool, it does come with certain drawbacks that must be taken into account. One significant issue is overgeneralization. SMOTE may blindly generalize the minority area without regard to the majority class. Furthermore, SMOTE can be less effective in high-dimensional feature spaces as it might fail to capture more complex patterns [8]. To mitigate these concerns and ensure the robustness of my results, this paper will include findings for both the original (non-SMOTE) and the SMOTE-enhanced datasets. Below is a comparison of the total number of observations in the two data sets I will use moving forward:

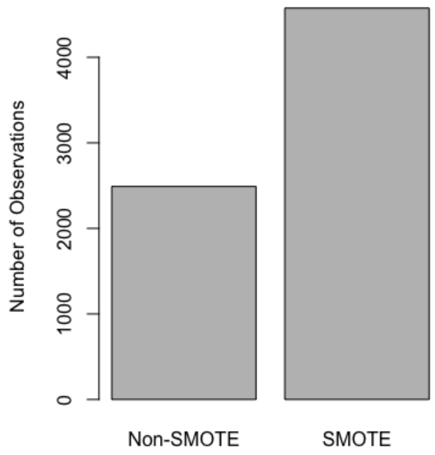

Figure 4: Comparison of non-SMOTE and SMOTE data sets

4. Dimension Reduction

When running a logistic regression model on the full data set few significant variables were identified. Since the predictors are highly correlated to each other (a player that makes more field goals per game also makes more points per game etc.), the lack of significant variables can be attributed to multicollinearity. Shown below is the Variable Inflation Factor (VIF) of each of the 26 original predictors; for context any number above a 5 is considered highly correlated with the rest of the predictors:

Table 1: Variance Inflation Factor - Full Model

	x
games	6.3847
gs	7.1444
mpg	6.5956
fg	1685.2444
fga	3513.1905
fgp	19.2977
two	883.8785
twoa	2459.3417
twop	15.0984
three	478.1427
threea	1629.7317
threep	1.3210
ft	301.4056
fta	66.3413
ftp	5.8151
orb	87.0021
drb	213.1932
trb	498.0868
apg	4.6012
spg	2.1730
bpg	2.0982
tpg	3.8499
pfpg	1.8380
ppg	2970.1681
sos	1.1245

It is clear from these results that dimension reduction is needed. In my study, I diverged from traditional dimension reduction methods like PCA and LASSO regression, choosing instead to apply stepwise Bayesian Information Criterion (BIC). BIC is known for its stringent penalties on models with more predictors. This was beneficial not only to reduce multicollinearity, but also because I want the results to be interpretive.

Out of the 26 initial variables, only a handful were identified as significant - games, field goals, field goal percentage, free throws per game, defensive rebounds per game, assists per game, steals per game, turnovers per game, and the SOS score. Interestingly, turnovers per game emerged as the only predictor with a negative coefficient, a logical outcome reflecting its inverse relation to a player's likelihood of becoming an All-Star. The VIF of the 9 predictors is shown below:

Table 2: Variance Inflation Factor - Reduced Model

	x
games	1.0526
fg	2.7122
fgp	1.2795
ft	2.2988
drb	1.6372
apg	2.8674
spg	2.2037
tpg	2.3316
sos	1.0730

After dimension reduction, all variables become significant and have little to no signs of multicolliniarity. Moving forward, these 9 predictors will be the only ones taken into account by the models. Below are plots of 2 of the significant variables:

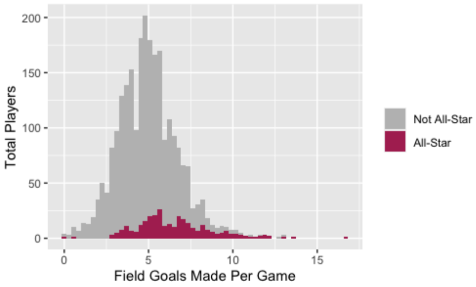

Figure 5: Comparing Field Goals Per Game between future All-Stars and future Non-All-Stars

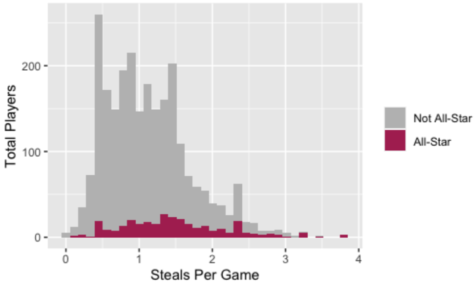

Figure 6: Comparing Steals Per Game between future All-Stars and future Non-All-Stars

5. Experiment

To recap, I have chosen to use the MICE data set. However, the MICE data set will be broken up into 2 data sets: the SMOTE-enhanced and non-SMOTE data sets. Also, all data sets will only be taking into account the 9 significant predictor variables along with the response variable. Moving forward into the experimental phase of my research, the 2 data sets will be labeled non-SMOTE and SMOTE.

For the experimental analysis, I will utilize four supervised learning algorithms: Logistic Regression (LR), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA). I have split the both the non-SMOTE and SMOTE data sets into training and testing subsets (80/20) and all results shown are the measurements of the predictive power that the model trained on the training data set has for the testing data set. Logistic regression will serve as my primary results, and I will go more into depth on the process there, while KNN, LDA, and QDA will serve as my secondary results in which accuracy metrics will just be shown to illustrate that there is no model bias and that the patterns exist in the data itself.

5.1. Logistic Regression

Logistic regression analysis is a statistical technique to evaluate the relationship between various predictor variables (either categorical or continuous) and an outcome which is binary. There are assumptions that must be met in order to properly implement a logistic regression model. The first, which has already been addressed is multicolliniearity. The second and third are the linearity assumption and influential observation assumptions which are shown below:

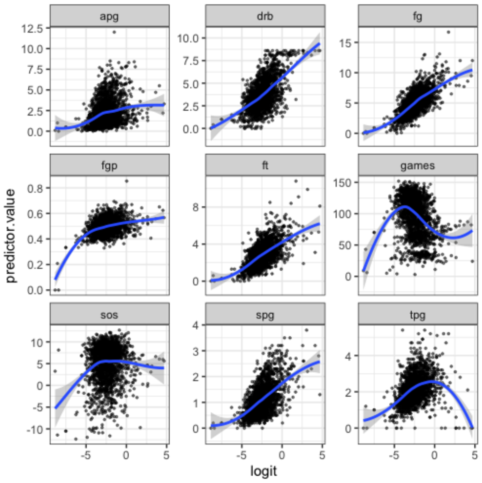

Figure 7: Linearity Assumption

In the ideal world, the linearity assumptions are met when all of the charts shown above have a linear slope. The plots are not perfect but do indicate that the linearity assumption, after logit transformation, has been met for most of the predictor variables.

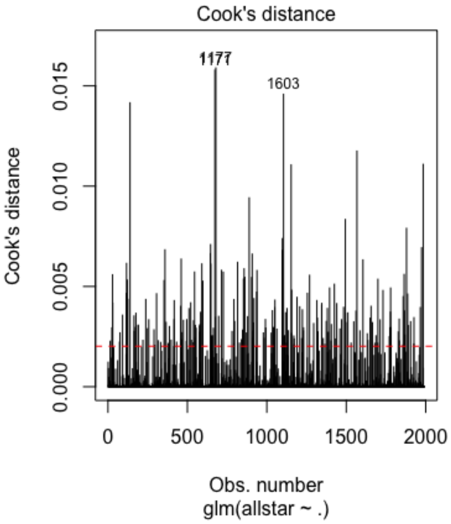

Figure 8: Cook's Assumption

In general, we want most points to be below Cook's distance, indicated by the red line, to ensure not many influential points. These results indicate that there are some but not many influential points. The coefficients for the logistic regression model on the training data set are shown below:

Table 3: Coefficients on All Predictor Variables

	x
(Intercept)	-9.0225516
games	-0.0107885
fg	0.2301487
fgp	8.2860414
ft	0.2015378
drb	0.2785286
apg	0.2562168
spg	0.9337275
tpg	-0.5642581
sos	0.0754953

While all predictors are found to be significant, Field Goal Percentage seems to have the largest predictive power over a player becoming an All-Star.

The results of logistic regression on the non-SMOTE data set are shown below:

Table 4: Non-SMOTE Logistic Regression

	Precision	Recall	F1-Score
False	0.9284	0.9888	0.9577
True	0.7826	0.3462	0.4800
Accuracy	NA	NA	0.9217

It performs quite well, in fact its accuracy of 92% will be the highest of any non-SMOTE or SMOTE model. However, notably, it struggles with identifying the positive class. Below is the results of the SMOTE model:

Table 5: SMOTE Logistic Regression

	Precision	Recall	F1-Score
False	0.7476	0.7302	0.7388
True	0.7657	0.7814	0.7735
Accuracy	NA	NA	0.7574

A trend that will be seen moving forward is the SMOTE data set performing much stronger on true metrics such as true F1-Score. This can be attributed to the aforementioned class imbalance issue with the minority class being biased against when modeling.

5.2. *K-Nearest Neighbor*

K-Nearest Neighbors is simple approach to supervised machine learning that creates a non-linear boundary by measuring zones in which one response category dominates another. It is a non-parametric model and there are no assumptions being made about the underlying distribution of the data. It generally performs well on large data sets which is a disadvantage for my research since I have no more than 5000 samples in the any data set. Below is the results for the non-SMOTE data set:

Table 6: Non-SMOTE K-Nearest Neighbor

	Precision	Recall	F1-Score
False	0.9097	0.9933	0.9496
True	0.7273	0.1538	0.2540
Accuracy	NA	NA	0.9056

Below is the results for the SMOTE data set.

Table 7: SMOTE K-Nearest Neighbor

	Precision	Recall	F1-Score
False	0.7927	0.7116	0.7500
True	0.7656	0.8351	0.7988
Accuracy	NA	NA	0.7770

5.3. *Linear Discriminate Analysis* & *Quadratic Discriminate Analysis*

Linear Discriminant Analysis and Quadratic Discriminant Analysis differ from the previous models in two major ways.

First, LDA and QDA are generative models, meaning that they capture the joint probability of their response variable and its predictors. This differs from Logistic Regression and KNN because those are discriminate models that capture the conditional probability of response given the predictors. While generative models are more flexible than discriminate models, they are more sensitive to outlier.

The second key difference is that, LDA and QDA make the assumption that the data comes from multivariate normal distribution. Similarly to KNN, LDA and QDA are both non-parametric models which is a main cause for their aforementioned flexibility. LDA is a special case of QDA because its boundary is strictly linear while QDA often has a quadratic boundary and differing covariance matrices for both classes. Below are the results for both LDA and QDA on the non-SMOTE and SMOTE data set:

Table 8: Non-SMOTE Linear Discriminate Analysis

	Precision	Recall	F1-Score
False	0.9296	0.9776	0.9530
True	0.6552	0.3654	0.4691
Accuracy	NA	NA	0.9137

Table 9: SMOTE Linear Discriminate Analysis

	Precision	Recall	F1-Score
False	0.7324	0.7512	0.7417
True	0.7743	0.7567	0.7654
Accuracy	NA	NA	0.7541

Table 10: Non-SMOTE Quadratic Discriminate Analysis

	Precision	Recall	F1-Score
False	0.9276	0.9484	0.9379
True	0.4524	0.3654	0.4043
Accuracy	NA	NA	0.8876

Table 11: SMOTE Quadratic Discriminate Analysis

	Precision	Recall	F1-Score
False	0.6483	0.8186	0.7235
True	0.7903	0.6062	0.6861
Accuracy	NA	NA	0.7060

6. Model Comparison

Comparing the models, we see 2 noticeable trends. First, is that all models performed significantly more accurately on the non-SMOTE data set but had significantly worse True Recall. Second, is that Quadratic Discriminate Analysis greatly under performed the other models across the board. The results of the non-SMOTE model accuracy comparison are shown below:

Table 12: Non-SMOTE Accuracy

	Accuracy
LR	0.9217
KNN	0.9056
LDA	0.9137
QDA	0.8876

The results of the non-SMOTE true precision comparison are shown below:

Table 13: Non-SMOTE True Precision

	T Precision	T Recall	T F1-Score
LR	0.7826	0.3462	0.4800
KNN	0.7273	0.1538	0.2540
LDA	0.6552	0.3654	0.4691
QDA	0.4524	0.3654	0.4043

The results of the non-SMOTE false precision comparison are shown below:

Table 14: Non-SMOTE False Precision

	F Precision	F Recall	F F1-Score
LR	0.9284	0.9888	0.9577
KNN	0.9097	0.9933	0.9496
LDA	0.9296	0.9776	0.9530
QDA	0.9276	0.9484	0.9379

It is not surprising that KNN performed well on false metrics and performed poorly on true metrics because KNN gets severly hampered when there is data imbalance. Basically, it identified the domain as a false region. Moving on, the results of the SMOTE accuracy comparison are shown below:

Table 15: SMOTE Accuracy

	Accuracy
LR	0.7574
KNN	0.7770
LDA	0.7541
QDA	0.7060

The results of the SMOTE true precision comparison are shown below:

Table 16: SMOTE True Precision

	T Precision	T Recall	T F1-Score
LR	0.7657	0.7814	0.7735
KNN	0.7656	0.8351	0.7988
LDA	0.7743	0.7567	0.7654
QDA	0.7903	0.6062	0.6861

The results of the SMOTE false precision comparison are shown below:

Table 17: SMOTE False Precision

	F Precision	F Recall	F F1-Score
LR	0.7476	0.7302	0.7388
KNN	0.7927	0.7116	0.7500
LDA	0.7324	0.7512	0.7417
QDA	0.6483	0.8186	0.7235

We notice that in the SMOTE data set the metrics almost average out with the non-SMOTE's false and true precision metrics. The non-SMOTE had true F1-Scores around .4 and had false F1-Scores at around .95 while the SMOTE has both true and false F1-Scores around .75. This indicates that the SMOTE models seem to be more flexible both ways.

7. Exploration

Treating the raw results of the logistic model as a probability, I will be exploring both the test set and the 2023 Rookie Class (brand new data).

7.1. Test Set

For the test set, I will only be displaying the top 10 in terms of odds by name and doing some informal analysis on how each of these 10 players fared in their career. I also will add what draft pick they were selected with.

Table 18: Top Ten Players in Testing Data Set

Player	All-Star Percentage
Elgin Baylor	97.852
Wilt Chamberlain	97.1827
Austin Carr	94
Larry Bird	92.7771
Kareem Abdul-Jabbar	86.1288
Dave Bing	86.1072
Paul Arizin	82.7361
Mookie Blaylock	78.721
Michael Beasley	78.4602
Dave Cowens	77.9438

This list is quite interesting and even indicates potential that my models go beyond the scope of my paper. To begin, 9 out of the 10 players shown went on to become All-Stars, with Michael Beasley being the exception. But even more impressively, 7 of the 10 identified went on to become Hall of Famers. With a total of 443 all-stars in NBA history and 110 Hall of Famers in NBA history, these results suggest that the predictors for All-Stars have potential to be extrapolated out to predict all of famers.

Table 19: Results of the Top Ten Players in Test Set

Player	Allstar	Hall of Fame
Elgin Baylor	Yes	Yes
Wilt Chamberlain	Yes	Yes
Austin Carr	Yes	No
Larry Bird	Yes	Yes
Kareem Abdul-Jabbar	Yes	Yes
Dave Bing	Yes	Yes
Paul Arizin	Yes	Yes
Mookie Blaylock	Yes	No
Michael Beasley	No	No
Dave Cowens	Yes	Yes

In 1996, the NBA created a comprehensive list of the 50 greatest players of all time. Elgin Baylor, Wilt Chamberlain, Larry Bird, Kareem Abdul-Jabbar, Dave Bing, Paul Arizin, and Dave Cowens all made the list. The 50 greatest player ceremony picture depicting the aforementioned players is shown below:

Figure 9: Group Photo Depicting the Top 50 Players of All Time

7.2. 2023 Rookie Class

For the 2023 Rookie Class, I will display the top 5 picks and look at each of their corresponding probabilities. As this paper is being written in December of 2023, I can only add a few months worth of informal analysis.

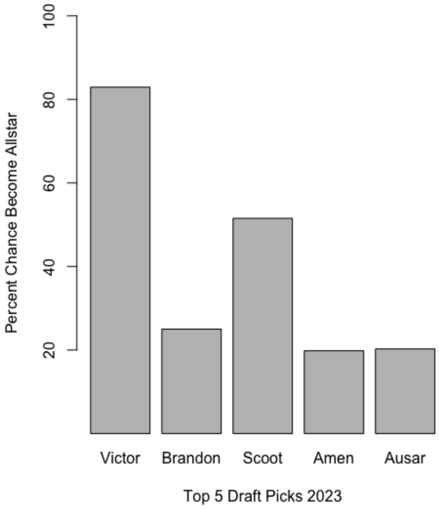

Figure 10: Top 5 Picks in 2023 Draft

The results are interesting as only French pheonom Victor Wembanyama is considered to have an above 50% chance of becoming a future All-Star. These results are in line with what they have shown so far as Brandon, Amen, and Ausar looking like they will turn out to be solid role players but not All-Stars while Victor will likely take home Rookie of the Year. Interestingly, Scoot Henderson, has the 2nd highest potential but has struggled as of late in the league.

8. Conclusion

This study was personally fun and interesting to me as a basketball fanatic. Seeing the results were pleasing and I believe that they can be taken further. For one, I did not use as many advanced metrics as predictors. Secondly, I believe using different measurement of success, perhaps a continuous variable, would provide interesting results and expand on my research.

Acknowledgements

I am deeply grateful to Michael Tsiang for his guidance and mentorship throughout this research project. His insights and expertise have been crucial in shaping both the direction and success of this study.

References

[1] , 2023. College basketball players and statistics. URL: https://www.basketball-reference.com.

[2] , 2023. List of nba all-stars. URL: https://en.wikipedia.org/wiki/List$_o f_N BA_A ll-$ $Stars.on\ Computer\ Science, B., 2023.Howtohandleunbalanceddataw$ https://www.baeldung.com.

[3][4] DraftExpress, 2021. Protrade: Analyzing future nba success through college stats. DraftExpress URL: www.draftexpress.com.

[5] GENESIS, F.T., 2023. Smote (synthetic minority oversampling technique). URL: https://www.fromthegenesis.com.

[6] Patton, A., Scott, M., Walker, N., Ottenwess, A., Power, P., Cherukumudi, A., Lucey, P., 2021. Predicting nba talent from enormous

amounts of college basketball tracking data. MIT Sloan Sports Analytics Conference URL: https://www.sloansportsconference.com/research-papers/predicting-nba-talent-from-enormous-amounts-of

[7] Plus, M.L., 2023. Mice imputation - how to predict missing values using machine learning in python. URL: https://www.machinelearningplus.com.

[8] Science, P.D., 2023. How to use smote for imbalanced classification. URL: https://www.practicaldatascience.co.uk.

[9] Spotintelligence, 2023. Imputation of missing values comprehensive & practical guide. URL: https://www.spotintelligence.com.