Project Progress Report
MGT 6203 – Spring 2023
Team 77

Predicting NBA Success from NCAA Stats

Francisco Sweet, Nivedita Minjur, Tao Hu, and Yunle Huang

Problem Statement

The NBA All-Star game showcases the 24 top players in the league. The players are selected from a combination of fan, media, and coach voting. Becoming an All-Star indicates a high level of skill and is very valuable to the franchise. However, not all the top picks drafted go on to become NBA All-Stars.

The ability to properly evaluate college basketball players is important because it allows professional teams to optimize their player budget. One way to evaluate the success of a player is if the player went on to play in the NBA All-Star game. Thus, team managers might be interested in predicting which college players will become an All-Star.

NBA All-Stars are top players in the league and have a huge financial impact on franchises that draft them. Many recent NBA All-Stars were relatively low draft picks, whereas many top draft choices never reached success in the NBA. Better prediction of future All-Stars could mean millions of dollars of extra revenue for NBA teams.

Literature Review

A handful of prior studies used historical NBA data to predict All-Star players. Cameron Porteous, a data scientist, and former GT MSCS student published an article in 2021 on how he created a model to predict NBA All-Star players from the current year's NBA roster. He started by scraping NBA player statistics using BeautifulSoup and Selenium. This data included traditional stats such as points, rebounds, and assists, as well as NBA popularity stats which are much harder to obtain for NCAA players.

After exploratory data analysis, Porteous split the dataset into test, train, and validation subsets. He then used XGBoost, an implementation of gradient-boosted decision trees similar to a random forest. After a few tweaks, the model looked promising, with an area under the receiver operating characteristic curve (AUC) close to 1. When tested on 2020 All-Star data, the model accurately selected 91% of the 2020 NBA All-Star players.

Yasmine Zakaria, a graduate of German University in Cairo, wrote her bachelor's thesis on a similar approach of identifying NBA All-Star players based on current year statistics. She used logistic regression and K Nearest Neighbor and showed that both models are promising, but logistic regression has slightly higher accuracy. Porteous and Zakaria show that decision trees, logistic regression, and classification algorithms all have promise in predicting NBA All-Stars.

Data Wrangling and Preliminary Analysis

Data for this project was scraped from three websites. Basketball-reference.com provides a list of historical NBA All-Stars. Sports-reference.com provides historical career statistics for college basketball players. WoodenAward.com provides lists of college players who have received the Wooden Award, indicating establishment recognition of achievements at the collegiate level. We considered each college player a data point and pulled the 24 basic performance statistics available from the sites for each player. The dependent variable for our training set, IsAllStar, was derived by matching the college players to the list of All-Stars. A screenshot of the source data on Sports-Reference.com is found in Figure 1.

Per Game Share & Export ▾ Glossary

Season | Conference

Season	School	Conf	G	GS	MP	FG	FGA	FG%	2P	2PA	2P%	3P	3PA	3P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	SOS
2020-21	Houston	AAC	26	2	9.9	1.3	3.0	.455	1.2	2.0	.604	0.1	0.9	.125	0.5	0.6	.750	0.1	1.0	1.1	1.5	0.8	0.2	1.0	0.8	3.3	5.37
2021-22	Houston	AAC	38	32	31.0	3.7	9.1	.405	2.7	5.9	.462	0.9	3.2	.298	1.7	2.1	.802	0.6	2.4	3.0	5.8	1.6	0.2	2.0	2.4	10.0	6.47
2022-23	Houston	AAC	31	31	32.3	3.8	9.1	.420	2.6	5.4	.485	1.2	3.7	.328	1.0	1.5	.689	0.6	2.5	3.2	5.4	1.8	0.1	1.8	2.0	9.9	4.25
Career	**Houston**		**95**	**65**	**25.6**	**3.1**	**7.4**	**.416**	**2.3**	**4.7**	**.488**	**0.8**	**2.7**	**.295**	**1.1**	**1.5**	**.761**	**0.5**	**2.1**	**2.5**	**4.5**	**1.5**	**0.2**	**1.7**	**1.8**	**8.1**	**5.36**

Figure 1

We scraped this data using Python scripts as shown in the jupyter notebooks available in the code folder of the accompanying GitHub repository. Restrictions on the source sites allowed for scraping of only 20 pages per minute. It took a total of about 5 days to scrape the data for all records. Once complete, we gathered historical records for 149,594 college basketball players and identified 515 all-stars.

We encountered a few challenges with the scraping. In addition to circumventing the anti-bot protection, we discovered that the college player data set with the All-Star data set based on player name was inaccurate. Multiple players share the same name, which caused some of the players to be incorrectly flagged as an All-Star. We solve this by re-scraping the data to identify each player based on the player key, a unique identifier for players.

Additional attributes were added in secondary scrapes in an attempt to improve poor-performing models. The Wooden Award indicator was added to indicate additional recognition for players and date-specific fields were added to identify how soon players achieved NBA status after joining the NBA.

The final list of attributes for the dataset is as follows.

Dependent Variable: IsAllStar, defined as a college player who played on at least one All-Star team during their NBA career.

Independent Variables

Stat	Description		Stat	Description		Stat	Description
G	Games		3P%	3Point %		TOV	Turnovers
GS	Games Started		FT	Free Throws Made		PF	Personal Fouls
MP	Minutes Played per game		FTA	Free Throw Attempts		Pts	Total Points
2P	2 Point Field Goals		FT%	Free Throw %		FG%	Field Goal Percentage
2PA	2 Point Attempts		TRB	Rebounds		FG3 %	3 Point Field Goal Percentage
2P%	2 Point %		AST	Assists		FT%	Free Throw Percentage
3P	3 Point Field Goals		STL	Steals		eFG %	FG% Weighted for points
3PA	3 Point Attempts		BLK	Blocks		SOS	Strength of Schedule
Wooden winner	John Wooden Award winner		LYC	Last Year of College		FYA	First-year as All-Star

Analysis/Code

Logistic Regression

Preliminary data analysis found correlations between many of the independent variables. To gain a better understanding, a correlation matrix was created using the 6 top-level performance indicators (Points, Blocks, Steals, Assists, Rebounds, and Strength of Schedule). Positive correlations were found between average points per game and all other factors except the strength of schedule. Points have the best correlation with a player's all-star potential, regardless of position. No college player in our dataset became an All-Star with less than 10 points per game on average. Results are visualized in Figure 2.

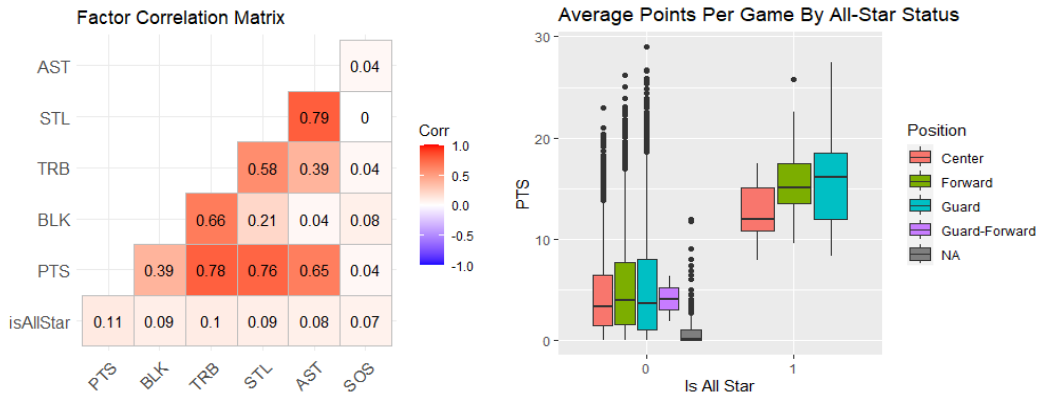

Figure 2

A logistic regression model was created using a player's points per game and strength of schedule to predict the probability of the player becoming an all-star within 10 years of leaving college. The model produced solid results with AUC=0.9798. Given a threshold of 17%, we were able to achieve a positive prediction rate of **37.5%** and an overall accuracy of **99.6%**. Results are visualized in Figure 3.

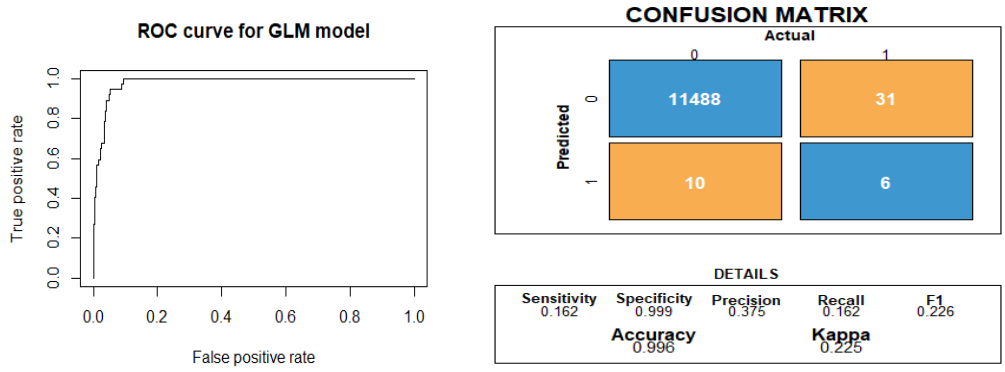

Figure 3

XGBoost

Another model that we test was XGBoost, an ensemble learning algorithm based on the decision tree model. It has some advantages over other algorithms such as improved performance, regularization to prevent overfitting, and fast training speed. We used cross-validation to train and validate the XGBoost models.

With the first attempt at using XGBust, which used our original data set, the model had poor performance. It had an overall accuracy of about 58.6% and a sensitivity of 58.5%. We made a few adjustments to improve the performance. First, we scraped a new predictor for whether or not the player was a winner of the John Wooden Award, an annual award given to outstanding men's and women's college basketball players. We believe that this award could act as a surrogate for valuable information such as popularity and playing ability. Second, we added additional filters to the data to make the dataset more relevant such as only including men in the analysis, removing players with zero game count, removing variables that wouldn't be available at the time of prediction, and performing the data merge using a unique player identifier key instead of player name. Third, we tried the XGBoost model on two different types of data provided by the websites. Sports-reference.com provides one dataset that shows normalized statistics that they call "advanced" data. They also provide a dataset without normalization. We found that the dataset without normalization performed better.

Another major challenge we encountered was that there were a lot of missing values. We relied on XG boost's built-in imputation capabilities to correct this. In addition, another challenge was that our data had very imbalanced classes. We saw that there were roughly 300 times more non-All-Star players than there were All-Star players (Table 1). This can potentially lead to misleading models that have high accuracy but aren't useful. We corrected the class imbalance in two ways. First, we used XGBoost's parameter, `scale_pos_weight`, to assign weights to the

two classes. This allowed us to place a greater emphasis on the All-Star player observations. Second, we used the area under the precision-recall curve (PRAUC) as the evaluation metric. From our research, we found that PRAUC was a more appropriate evaluation metric compared to the traditional area under the receiver operating characteristic curve. Maximizing PRAUC places a heavier emphasis on finding positives (the All-Star players).

Table 1

Classification	N
Non-All Star player	90,721
All-Star player	283

The overall PRAUC for all 5 folds was 0.081, and the overall accuracy was 90.6%. The overall sensitivity, specificity, positive predictive value (precision), and negative predictive value are below:

Table 2

Metric	Value
Sensitivity	91.2%
Specificity	90.7%
Positive predictive value	0.03%
Negative predictive value	100%

The high sensitivity suggests that out of all the All-Star players, our model was effective at finding them. The very high negative predictive value suggests that it was highly effective at ruling out players who could not be All-Star players. Unfortunately, we have a very low positive predictive value, which means there is still a lot of uncertainty when the model makes a positive prediction. The low PRAUC and accuracy probably result from the low positive predictive value.

K Nearest Neighbor

K Nearest Neighbor is a non-parametric, supervised classifier that makes predictions on groupings of an individual data point. We used k values from 1-5 to train the models.

The major challenge of using KNN is that it requires predicting variables to be all numeric and without any missing values (NA). Hence, we dropped some columns that might be useful in making a prediction such as a school name and conference name. Moreover, due to the non-NA limitation of KNN, we dropped all the rows that contain at least 1 null value, which reduces our dataset size to around 30000 records. The data is split into training and testing sets following the 7:3 ratio.

As a result, using a k value of 1 on the dataset of advanced players played in conferences yields the best result with an accuracy rate of 99.7%. However, the high accuracy rate is caused by data imbalance – 99% of the players in our dataset are not all-stars, so the model did a good job in classifying non-all-star players but did poorly in predicting who will be an all-star player as illustrated by the confusion matrix shown in table 3.

Table 3

	Not all-star	All-star
Not all-star	11514	19
All-star	16	2

Support Vector Machine

A support vector machine (SVM) is a classification algorithm commonly used for 2 group classification approaches. In this case, our model predicts if an NCAA player is an all-star or not.

Similarly to KNN, one of the main issues with SVM is that all predictive variables must be numeric (i.e. not categorical or null). Therefore, to maximize the number of data points available, the model looks at games played, total points, assists, and rebounds only to predict future all-stars. After filtering out null values for these fields and filtering out data points not applicable to the model (ex: women players, players with no games played), our model has 70,550 data points. These were split into 80% training data and 20% test data.

An additional issue with SVM is addressing the class imbalance. When the model is initially run, it records a very high accuracy by predicting that all players will not become all-stars. This issue is addressed by assigning class weights to the SVM model. This allows the model to assign more importance to classifying all-star players.

The final model has an accuracy of 89.55%. It classifies all-star players reasonably well, but as a result, erroneously classifies many non-all-stars as all-stars. The confusion matrix below shows this.

Table 4

	Not all-star (Actual)	All-star (Actual)
Not all-star (Prediction)	12485	3
All-star (Prediction)	1458	38

Conclusion and future steps

Achieving NBA All-Star status for a college basketball player is a rare event. Rare event analysis like this makes it easy to predict negative outcomes, but hard to find positive ones without excessive false positives. We want a high positive prediction rate with a low false positve rate.

Of the four models developed, Logistic Regression achieved the best results, achieving a positive prediction rate of 37.5% while keeping false positves to 62% using a threshold of 17%. Although these results are promising and can help in deciding which players to focus on during

the draft, this isn't enough confidence to fully commit all funds to a particular player. During our research, it was discovered that less tangible factors, such as injuries and personal issues may have a strong influence on player achievement.

As an anecdote on these other factors, we submit the case of Michael Beasley. Based on our logistic model of his college career, Michael Beasley had a 77% probability of becoming an all-star during his NBA career. His college statistics were impressive, but as documented in this 2013 article by Grant Huhges, there were off-the-court issues that were amplified in the NBA.

Future models could incorporate off-court factors which could impact the NBA performance of college athletes. For now, we'll use our logistic regression model to identify the top three prospects of the current crop of college athletes. Be on the lookout for these three players during the next NBA draft.

- **Brandon Miller**, a forward from Alabama with a 30% probability of becoming an All-Star.
- **Trayce Jackson-Davis**, a Forward from Indiana with a 23% probability of becoming an All-Star.
- **Hunder Dickenson**, a Center from Michigan with a 22% probability of becoming an All-Star.

Works Cited

1. Zakaria, Yasmine. "Identifying All-Star Players in NBA." *German University in Cairo*, 2019.
2. Porteous, Cameron. "Using Machine Learning to Predict NBA All-Stars." *Medium*, Towards Data Science, 17 Sept. 2020, https://towardsdatascience.com/using-machine-learning-to-predict-nba-all-stars-part-1-data-collection-9fb94d386530.
3. Hughes, Grant. "Complete Timeline of Michael Beasley's Downward Spiral in the NBA", BleacherReport, 7 Aug. 2013., https://bleacherreport.com/articles/1730852-complete-timeline-of-michael-beasleys-downward-spiral-in-the-nba