What Makes an All-NBA Player?

By Stephen Chen and Ryan Alvarez-Cohen

Bruin Sports Analytics, 2020

Contents

Abstract

As new All-NBA teams are named at the culmination of each season, there are always players who are considered confident selections to make the teams, just like there are players that might be surprise members and those whom some might consider snubs from the picking. But why are the players who are chosen to these teams the ones selected? Looking beyond a player's holistic impact on the court, could there be specific stats that are key indicators for those who make these teams? Furthermore, could these stats then be used to predict the set of players to make the All-NBA teams? We set out to answer these questions by creating a model to output the probability of a player to make an All-NBA team by observing their most heavily-influential season stats. In order to train our model, we gathered data from the last 10 seasons from both NBA.com and Basketball-Reference.com. We explored a variety of variable selection methods to identify the most influential predictors and used them in a logistic regression model. Our final model contained 8 predictors and was able to predict All-NBA players with a 57% accuracy. The eight predictors utilized in our final model were minutes per game, games played, points, free-throws attempted, number of triple doubles on the season, player efficiency rating (PER), win shares, and true-shooting percentage. Future studies we would like to conduct include not only predicting if a player will make an All-NBA team, but which specific team they will be placed on.

Introduction

Every year, a select group of the best NBA players are chosen for one of three All-NBA teams, with each team consisting of five players comprising two guards, two forwards, and two centers. The first team is supposedly made up of the highest tiered players, with the second and third teams having correspondingly second and third tiered players. All-NBA teams have been selected every year dating back to the NBA's initial season, when only two All-NBA teams were selected. The players are named to these teams based on the cumulative votes from a committee of sports analysts made up of broadcasters and writers throughout North America. Analysts give players five, three, or one point(s) depending on whether they vote them into the first, second, or third team respectively. Those with the five highest points make the first team, the five with the second highest make the second team, and so on. The players are judged on their performance purely from the regular season, without taking the playoffs into consideration ("All NBA Team"). Veteran players with at least 8 years in the league serve to gain a salary increase - up to 35% of the salary cap - from making an All-NBA team. They can do this by making an All-NBA team in their most recent season (8th) or by making two All-NBA teams in the previous three years. Without this accolade on their resume, players can only make up to 30% of the salary cap from their veteran contracts. This 5% decrease can result in a difference greater than $5 million in a single year compared to a contract making them 35% of the salary cap (Kent). Thus, the incentive for players to make an All-NBA team is clear as they serve to gain a significant amount in their career by doing so.

In previous research studies, many have created a variety of models to predict various NBA season accolades such as MVPs, All Stars, and All-NBA players using different techniques. In one example, Peter Li created a model to predict MVPs by utilizing a custom value formula for each player which took into account a player's general win contribution to their team in addition to a breakdown of their individual stats. Fantasy scoring was used to weigh the individual stats from a player. With respect to fantasy scoring, we found it to be a decent predictor for one of our models, which will be discussed further later on. Li's model correctly predicted 9 out of the previous 10 MVPs, displaying fairly high accuracy (Li).

In another study conducted by Tal Boger, he created models to predict players making one of the All-NBA teams. The data for these models consisted of basic counting stats (points per game, assists per game, total rebounds, field goal percentage), advanced stats (win shares, value over replacement player) and team stats (seed, wins). He used four different classification models to make player predictions: support vector machine, random forest, k-nearest neighbors, and a deep neural network classifier. The support vector classifier was overall the most accurate model, correctly classifying close to 84% of NBA players as All-NBA or not from the testing dataset. Interestingly enough, in this study, defensive stats were purposefully ignored. Boger's reasoning was that defensive stats are largely random in nature and struggle to truly quantify a player's defensive quality in today's modern game. He further mentioned that the quality of defensively oriented big men who typically make All-NBA teams is exhibited through holistic stats such as value over replacement player (VORP), win shares, and total rebounds, all of which were collected in this study. Similarly, we found in our research that the quality of a player's defense as exhibited by shutting down the opposing team's best player, or making a game clinching steal in the final minutes, often goes overlooked when assessing value for that player. Whether we are measuring a player by salary, their minutes played, or awards such as All-NBA, All Star or even MVP, the accolades by which we gauge players often undervalue defensive-minded players in comparison to their more offensive minded counterparts (Boger). We observed this trend of "defensive snubbing" in our own models, which we will also further discuss later on.

One final piece of background research worth mentioning was conducted by Griffin H. He created a decision tree classification model to predict All Stars using the *scikit-learn* package in Python. In the creation of our models, we used the same package to utilize a select k-best classifier to determine predictors for logistic regression. In Griffin's case, however, the decision tree classifier determined the following 10 stats to be the best predictors: FGM, FTA, Points, FGA, Mins, Rebounds, Assists, TS%, Blocks, EFG% (H.). We will see how these 10 compare to the predictors our model outputted later on.

The goal of our research was to build a logistic regression model to predict the All-NBA team for the 2019-2020 season. To build this model, we had to determine which stats were the best predictors for a player to make an All-NBA team. We hypothesized that points, assists, PER, and team record would be the stats

that most heavily-influenced whether a player would be selected to an All-NBA team.

Methods

To obtain our dataset, we scraped data from nba.com and basketball-reference.com. Our goal was to analyze the numbers from the modern NBA, so we decided to gather the most recent ten complete seasons, starting from the 2009-2010 season to 2018-2019 season. In these seasons, we scraped traditional box-score statistics (such as points, rebounds, assists, etc.), advanced statistics (such as true shooting percentage, player efficiency rating, win shares, etc.), defensive statistics (such as rebound percentage, defensive win shares, etc.), and players selected in the All-NBA team. In all, we scraped 65 distinct variables. Scraping across the two websites created inconsistencies in player names. We found that nba.com sometimes used players' nicknames while basketball-reference.com used players' legal names. Additionally, our scraper struggled to translate accented characters common in European, Asian, and African names into plain text. To merge datasets together, we corrected all inconsistent and incorrect names.

To analyze the core question and reduce extreme outliers in different measurements, we wanted to examine players who had played more than twelve minutes a game and more than twenty games a season, effectively removing players who did not play a meaningful role on their team and only played in occasional, situational moments. After cleaning our dataset, 3501 players remained.

We explored several methods to identify important predictors. Our first method was forward stepwise selection. Forward stepwise selection is a regression method from the R package *leaps* that fits models by starting with 0 predictors and sequentially adding predictors to the models until reaching a specified number of predictors. This function identifies all of the best predictors for each k-predictor model. Based on a graph of the residual sum of squares from the 1- to 30-predictor models, the curve flattens around the 7- to 10-predictor models. Using this function, we identified all the best predictors from the 7, 10, and 15-predictor models and ran logistic regression models.

Our second method in identifying important predictors was select k-best. Select k-best is a function accessible from the *scikit-learn* package in Python. It returns the k features in the dataset that contribute most heavily to the target variable. This contribution can be determined by a variety of methods. We used the f_classif method which functions by performing an ANOVA test between each predictor variable and the target variable, computing a corresponding F-value for each instance. To obtain this F-value, the numeric predictors are grouped by the categorical target variable and the grouped means from each predictor are observed; if the means within each group are significantly different, then that predictor is dependent on the target variable, whereas if the means do not differ significantly, then the predictor is independent of the target variable. The k variables with the greatest dependency, as conveyed by the F-value, are returned. The variable k is specified by the user and determines the number of features or predictors to be returned. We created models with output from the select k-best function at k values of 10, 20 and 30. One of the primary advantages of select k-best is its reduction of overfitting and its limiting of redundant variables (thus, using it to return 30 variables may have been somewhat counterintuitive, but interesting to observe nonetheless). We chose the 10-variable model, selecting points per game, free-throws attempted, free-throws made, NBA fantasy points, player efficiency rating, win shares, offensive win shares, player impact estimate, personal fouls drawn, and number of double-doubles.

We also implemented ridge regression on our more complex, higher-variable models using the glmnet() function from the *glmnet* package in R. Ridge regression aims to offset the problems in complex and high-predictor models, such as multicollinearity and large variances, by adding bias to regression estimates. We applied ridge regression on models with 20 and 30 variables.

For all of these methods, we applied these sets of variables in logistic regression models, predicting the binary outcome whether or not a player would be selected as an All-NBA player. To assess the performance of our models, we employed a validation method. We trained these models on 6 seasons:the 2009-2010, 2010-2011, 2013-14, 2014-15, 2016-17, and 2018-2019 seasons. We used the trained models to predict which players would be selected as All-NBA in the other 4 seasons: the 2011-12, 2012-13, 2015-16, and 2017-18 seasons.

We aimed to make the training and testing datasets contain a balanced mix of older and newer seasons in the decade.

To validate our models, we aimed to minimize collinearity and overfitting. To minimize collinearity, we computed the variance inflation factor (VIF) for our models using the vif() function from the *car* library in R. We also used the findCorrelation() from the *caret* library in R, which returns the predictors to remove based on high pairwise correlations between variables. To minimize overfitting, we employed 5-fold cross validation to make sure our models stayed consistent and were not highly variable to inputs (and thus not overfitting).

Predictions based on logistic regression models output probabilities for each player. We determined that a player would have made an All-NBA team if the player's predicted probability of making an All-NBA team was above 0.5. Accuracies among these models were determined with confusion matrices, labeling the number of correctly predicted non-All-NBA players, correctly predicted All-NBA players, incorrectly predicted players as All-NBA, and incorrectly predicted players as non-All-NBA.

Results

We determined the accuracy of our models based on the number of actual All-NBA players the models were able to predict on the testing dataset. Our testing dataset contained 4 seasons, so there were 60 All-NBA players in the testing dataset. If a model could predict more actual All-NBA players than other models, we deemed that model more accurate.

For our forward stepwise variable selection method, we chose the number of predictors to use in our logistic regression model based on its residual sum of squares graph, where we determined that there was very minimal change in residual sum of squares in the 7-15 predictor models. We thus concluded that this range of predictors fitted the observations in the training dataset well. We utilized 7-predictor, 10-predictor, 20-predictor, and 30-predictor results into our logistic regression model. We also used ridge regression on higher-predictor models.

For the select k-best variable selection method, we utilized 10-predictor, 20-predictor, and 30-predictor results into our logistic regression model. We also used ridge regression on higher-predictor models. Below is the accuracy of these models.

Table 1: Prediction Accuracies of Each Model

Model	Accuracy
Forward stepwise, 7-predictor	55%
Forward stepwise, 10-predictor	60%
Forward stepwise, 20-predictor	68.33%
Forward stepwise, 20-predictor w/ ridge regression	53.33%
Forward stepwise, 30-predictor	76.67%
Forward stepwise, 30-predictor w/ ridge regression	68.33%
Select k-best, 10-predictor	60%
Select k-best, 20-predictor	68.33%
Select k-best, 20-predictor w/ ridge regression	58.33%
Select k-best, 30-predictor	71.67%
Select k-best, 30-predictor w/ ridge regression	58.33%
Final Model: 8-predictor	56.67%

Since the high-variable models had high VIF scores for the variables and had many highly correlated pairwise variables, we deemed that these models were not viable for analysis and had too many unimportant

variables. Thus, we decided to move forward with a 10-variable model. Because the forward stepwise 10-predictor model had fewer false positives, we decided to refine the variable selection process with this model.

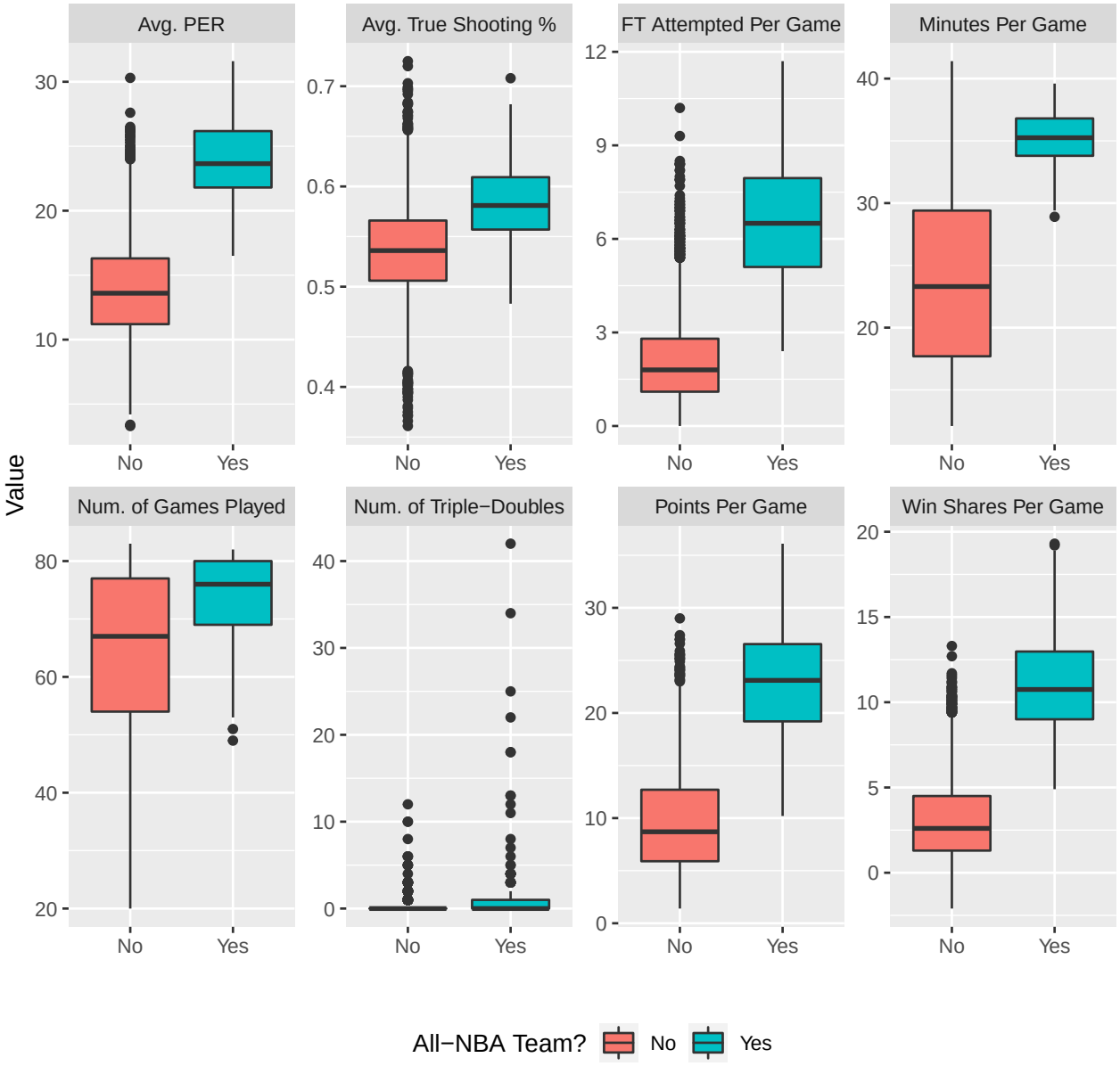
After using forward stepwise variable selection and refining the variable selection, our best model was an 8-predictor logistic regression model. The predictors in this model were minutes played per game, games played, points per game, free throws attempted per game, number of triple-doubles, player efficiency rating, win shares, and true shooting percentage. The 5-fold cross validation error rate for our final model was 2.20%, and had a near-0 variance, which demonstrated that our model was not overfitting to our training dataset. Note that the error rate was very low during cross-validation due to the very low number of All-NBA players within the data of each fold.

Table 2: Summary Output of Final Model

	Estimate	Odds Ratio	CI (lower)	CI (upper)	Std. Error	z value	Pr(>|z|)	
(Intercept)	6.3376404	565.4604998	0.0001385	4.126310e+07	6.9221653	0.9155575	0.36	
MIN	-0.1469942	0.8632990	0.7117244	1.071078e+00	0.1035539	-1.4194946	0.156	
GP	-0.0391766	0.9615809	0.9060533	1.029301e+00	0.0325610	-1.2031743	0.229	
PTS	0.2365397	1.2668579	1.0738623	1.507102e+00	0.0862079	2.7438288	0.006	**
FTA	0.0868099	1.0906893	0.8117061	1.478517e+00	0.1522442	0.5702017	0.569	
TD3	0.2150571	1.2399327	1.0195765	1.602435e+00	0.1308072	1.6440775	0.1	
PER	-0.0034697	0.9965363	0.7813100	1.300921e+00	0.1297264	-0.0267463	0.979	
WS	1.1303817	3.0968384	2.0645458	4.711008e+00	0.2111291	5.3539839	<0.001	***
TS_PCT	-26.0237177	0.0000000	0.0000000	2.310000e-05	7.9140085	-3.2883105	0.001	**

Visual 1 below shows a distribution of the final 8 predictors between All-NBA players and non-All-NBA players.

Visual 1: Boxplots of 8 Predictors in Final Model



After constructing the confusion matrix, this model predicted 34 of 60 All-NBA players in the testing dataset.

Table 3: Confusion Matrix - Accuracy of Test Predictions

	Actual-No	Actual-Yes
Prediction-No	1339	26
Prediction-Yes	9	34

This model was able to predict many All-NBA players at the point guard, shooting guard, small forward, and power forward positions, but it struggled to predict many centers.

Table 4: Position Accuracy Table

Position	Accuracy
PG	66.67%
SG	50%
SF	63.64%
PF	75%
C	25%

This model was also able to predict most of the First Team and Second Team All-NBA players, but it struggled to predict many Third Team All-NBA players.

Table 5: All-NBA Team Accuracy Table

All-NBA Team	Accuracy
First Team	80%
Second Team	65%
Third Team	25%

Predicting on our testing dataset, the model correctly predicted 34 All-NBA players.

Table 6: Correctly-Predicted Players

Player	Season	All-NBA Team	Position	Probability
Chris Paul	2011-12	First Team	PG	94.84%
Kevin Durant	2011-12	First Team	SF	96.03%
LeBron James	2011-12	First Team	SF	99.76%
Kevin Love	2011-12	Second Team	PF	85.37%
Chris Paul	2012-13	First Team	PG	96.25%
Kevin Durant	2012-13	First Team	SF	100%
Kobe Bryant	2012-13	First Team	SG	93.07%
LeBron James	2012-13	First Team	PF	100%
Blake Griffin	2012-13	Second Team	PF	59.33%
Carmelo Anthony	2012-13	Second Team	PF	85.84%
Marc Gasol	2012-13	Second Team	C	53.14%
Russell Westbrook	2012-13	Second Team	PG	97.04%
Dwyane Wade	2012-13	Third Team	SG	50.25%
James Harden	2012-13	Third Team	SG	97.48%
Kawhi Leonard	2015-16	First Team	SF	96.91%
LeBron James	2015-16	First Team	SF	99.5%
Russell Westbrook	2015-16	First Team	PG	99.99%
Stephen Curry	2015-16	First Team	PG	99.99%
Chris Paul	2015-16	Second Team	PG	94.92%
Damian Lillard	2015-16	Second Team	PG	59.3%
Draymond Green	2015-16	Second Team	PF	84.96%
Kevin Durant	2015-16	Second Team	SF	99.61%
Kyle Lowry	2015-16	Third Team	PG	84.04%
LaMarcus Aldridge	2015-16	Third Team	PF	54.73%
Anthony Davis	2017-18	First Team	PF	99.37%
Damian Lillard	2017-18	First Team	PG	97.81%
James Harden	2017-18	First Team	SG	99.98%
Kevin Durant	2017-18	First Team	SF	69.93%
LeBron James	2017-18	First Team	PF	99.97%
DeMar DeRozan	2017-18	Second Team	SG	64.66%
Giannis Antetokounmpo	2017-18	Second Team	PF	95.75%
LaMarcus Aldridge	2017-18	Second Team	C	85.99%
Russell Westbrook	2017-18	Second Team	PG	99.95%
Karl-Anthony Towns	2017-18	Third Team	C	90.75%

However, the model did not predict 26 All-NBA players in our testing dataset.

Table 7: Missed Players

Player	Season	All-NBA Team	Position	Probability
Dwight Howard	2011-12	First Team	C	14.16%
Kobe Bryant	2011-12	First Team	SG	24.98%
Andrew Bynum	2011-12	Second Team	C	5.88%
Blake Griffin	2011-12	Second Team	PF	44.19%
Russell Westbrook	2011-12	Second Team	PG	38.75%
Tony Parker	2011-12	Second Team	PG	11.18%
Carmelo Anthony	2011-12	Third Team	SF	19.55%
Dirk Nowitzki	2011-12	Third Team	PF	17.17%
Dwyane Wade	2011-12	Third Team	SG	34.53%
Rajon Rondo	2011-12	Third Team	PG	2.08%
Tyson Chandler	2011-12	Third Team	C	0.37%
Tim Duncan	2012-13	First Team	C	20.98%
Tony Parker	2012-13	Second Team	PG	37.7%
David Lee	2012-13	Third Team	PF	16.81%
Dwight Howard	2012-13	Third Team	C	3.17%
Paul George	2012-13	Third Team	SF	20.54%
DeAndre Jordan	2015-16	First Team	C	20.6%
DeMarcus Cousins	2015-16	Second Team	C	18.68%
Andre Drummond	2015-16	Third Team	C	12.91%
Klay Thompson	2015-16	Third Team	SG	5.85%
Paul George	2015-16	Third Team	SF	47.73%
Joel Embiid	2017-18	Second Team	C	7.43%
Jimmy Butler	2017-18	Third Team	SG	29.57%
Paul George	2017-18	Third Team	SF	20.69%
Stephen Curry	2017-18	Third Team	PG	27.2%
Victor Oladipo	2017-18	Third Team	SG	17.72%

Our model also predicted 9 players who weren't voted in as All-NBA players.

Table 8: Snubs

Player	Season	Position	Probability
Deron Williams	2012-13	PG	53.07%
Stephen Curry	2012-13	PG	65.99%
DeMar DeRozan	2015-16	SG	74.95%
Isaiah Thomas	2015-16	PG	61.79%
James Harden	2015-16	SG	99.5%
Kemba Walker	2015-16	PG	51.13%
Ben Simmons	2017-18	PG	67.44%
Chris Paul	2017-18	PG	51.56%
Nikola Jokic	2017-18	C	86.02%

For prediction purposes, we also used our model to predict the 2019-20 All-NBA season using data gathered before the NBA shutdown in March.

Table 9: Predicted All-NBA Players for 2019-20 Season and Next 5 in

Player	Position	Probability
James Harden	SG	99.72%
Giannis Antetokounmpo	PF	98.82%
Luka Doncic	PG	97.91%
LeBron James	PG	97.64%
Anthony Davis	PF	87.14%
Nikola Jokic	C	79.54%
Damian Lillard	PG	74.75%
Kawhi Leonard	SF	54.53%
Jimmy Butler	SF	44.41%
Russell Westbrook	PG	19.18%
Trae Young	PG	13.41%
Domantas Sabonis	C	9.21%
Joel Embiid	C	6.76%
Bradley Beal	SG	6.51%
Khris Middleton	SF	6.5%
Jayson Tatum	PF	5.2%
Bam Adebayo	PF	4.8%
Chris Paul	PG	4.58%
Kyle Lowry	PG	4.3%
Montrezl Harrell	C	3.59%

Conclusions

Analysis

In this 8-predictor logistic regression model, a majority of the predictors identified were offensively-focused. This is an interesting but plausible result because most NBA media attention favors players who are extremely productive offensively (Anderson). Based on the summary output, points per game, win shares, and true shooting percentage were the only statistically significant predictors, further establishing this narrative. The selection of the variable player efficiency rating suggests that these players heavily impact the game offensively and efficiently. The selection of the variable free throws attempted per game suggests that these players would be usual recipients of foul-calls, possibly due to reputation or style of play. In addition, the high thresholds for games played and minutes per game indicate that, naturally, these players must also have a significant on-court role and stay healthy throughout the season. In all, based on the predictors chosen and highlighted, an All-NBA player is one that plays a high number of games, plays heavy minutes, scores a lot of points, and shoots at a high percentage. In other words, this is a player who is a healthy starter that has an integral role in the offensive scheme as a prolific and efficient scorer.

We settled on our final model by considering various factors among each of our models. We observed accuracy scores based on the confusion matrices and considered multicollinearity levels by looking at VIF scores and other inter-correlation measures, all while trying to minimize model complexity. The final 8-predictor model was the one in which we were able to use the fewest predictors while maintaining relatively decent accuracy and VIF levels. Of course, the accuracy at which our model was able to predict All-NBA players is not spectacular. However, our model's predictions can give interesting arguments and perspectives for players that our model predicted as All-NBA players but actually were not (false positives), and players that our model predicted as non All-NBA players but actually were (false negatives) based on the predictors that were chosen in our model. If each season's All-NBA selections can be a loose measure of the top-15 players in a season, our model roughly represents this idea. Analysts and fans alike can very possibly agree on the top 5 players in a season, but they may not be able to fully agree on the next 10 best players. Similarly, our model was able to correctly predict almost all of the First Team All-NBA players per season and a majority of the Second Team All-NBA players but struggled to predict Third Team All-NBA players.

According to Visual 1, it can be seen that the medians between All-NBA and non-All-NBA players are significantly different from each other, and that there are far fewer lower and upper outliers, with the exception of the number of triple doubles category. Across these categories All-NBA players have consistently sizeable increases compared to non-All-NBA players. The number of triple doubles was an interesting variable selected by our selection method, and upon closer look, it may seem as though the medians are very similar with All-NBA and non-All-NBA players. However, there are many and more extreme, upper outliers in this category. Given that triple doubles are generally very difficult to achieve in a game, it can be understood why this predictor is influential.

Another impactful predictor was the 'number of games played' variable. In Visual 1, we can see that 50% of All-NBA players play 70-80 games, and very few All-NBA players play less than 60 games. However in the 2011-12 season, players did not play the full 82 regular season games and instead played a reduced 66 regular season games, due to the 161-day lockout. Thus, we found that players like Carmelo Anthony, Dwayne Wade, Kobe Bryant, Rajon Rondo, and Tony Parker in the 2011-12 season all played 60 games or less in that season but were not predicted by our model as All-NBA players, even though they were selected that year. The reduced regular season games can explain why the model only predicted 4 of 15 players that season. This predictor also had major implications for predictions on the 2019-20 season that we explain in the later paragraphs below.

On the other hand, there were instances in which our model predicted certain players to make the All-NBA team to a high degree of probability, when in fact they did not. Players that fell into this false positive grouping can be considered 'snubs' from the All-NBA team according to our model. They likely should have made the team based on their season statistics, but were not voted in. The two most obvious mentions of this snub category are James Harden in the 2015-16 season and Nikola Jokic in the 2017-18 season. Our model predicted these players to make the All-NBA teams with probabilities of 99.5% and 86% respectively,

further supporting the theory that James Harden not making any of the All-NBA teams in 2016 is one of, if not the biggest, snub in All-NBA history. That season, Harden led the league in total points, minutes, and free throws. He played all 82 games and maintained a season average of 29 points, 7.5 assists and 6.1 rebounds. The only other players to ever average 29-7-6 in 70 or more games are LeBron James, Michael Jordan and Oscar Robertson. This is impressive company for any player to be in, and conveys how shocking it was for Harden to not even have made the All-NBA Third team (Rothstein). Given all of this, Harden was not voted onto an All-NBA team supposedly because of his highly questionable defense for the majority of that season. While it is true that his defense was certainly not good, to put both Klay Thompson and Kyle Lowry, each of whose offensive stats hardly compared to Harden's, on the teams instead of him was dubious (Rothstein). This outcome also plays directly into how we know our model struggles to consider players' defensive qualities, for if it did consider Harden's poor defense that season, maybe it would not have given him such a high probability for making an All-NBA team.

The snub of Jokic in 2017-18 is not as extreme as that of Harden in 2016, but it is in a similar vein. He had a highly productive season, and played a crucial role in a Denver team that struggled when he was not on the floor. While Joel Embiid and DeMar DeRozan both had strong seasons, they were debatably not as valuable to their teams as Jokic was to his team when considering each of their respective supporting casts (Huff). Embiid had the Rookie of the Year and breakout star in Ben Simmons beside him, while DeRozan had a current All-Star and experienced veteran in Kyle Lowry. Conversely, Jokic mostly had to rely on a young, still developing Jamal Murry.

With regards to our model's All-NBA predictions for the 2019-20 season, we can see that our model can predict the first 9 All-NBA players with high certainty. There was a fairly severe dropoff to the player with the next highest probability, which was Russell Westbrook at 19.18%. After him, it continues to drop rapidly, reaching under 5% within the next 7 players. However, this observation is consistent with the predictions from the testing dataset. The most surprising result was Trae Young listed 11th on this list, in terms of probability, who completed his second NBA season with the Atlanta Hawks. Although Young did not have a bad season offensively, he likely will not make any of this year's All-NBA teams, due to the subliminal notion that he is a poor defender and because he leads a team that has one of the worst records in the Eastern Conference. However, based on what we know about our model and the type of players that it favors, it is unsurprising to see a high volume shooter and prolific scorer like Trae Young this high on the list. We are expecting that a more seasoned veteran, like Chris Paul or Kyle Lowry, or a more all-around talent, like Ben Simmons, will be voted over Trae Young at the point guard position once the actual All-NBA teams are released. Given this probable inaccuracy in our model's predictions, it is likely that we will see our predicted top 9 players make it on to one of the All-NBA teams.

Shortcomings

However, there were some glaring faults in our procedures. Each All-NBA team consists of 2 guards, 2 forwards, and 1 center. When building our models and manufacturing predictions, we did not account for positions. So whereas each All-NBA team has an even mix of guards and forwards, our model did not account for this ratio and based predictions entirely off of player stats, without considering position. As mentioned previously in our introduction when referencing other studies, most stats, including the ones in our final model, typically do not give an accurate representation of the value that big men bring to the floor. Rather, these stats skew strongly towards primarily scoring-minded players. The few big men that are predicted to be All-NBA players from our model are ones that have a heavy offensive presence on their team and are prolific scorers (Nikola Jokic, Karl-Anthony Towns, LaMarcus Aldridge). Defensive-minded big men who do not have a large offensive focus would often be overlooked by our model since it does not really value their defensive worth.

The paradigm of this notion was Tyson Chandler, who won the Defensive Player of the Year award and was named to the All-NBA Third Team all in the 2011-2012 season. Of all players who were actual All-NBA players in our testing dataset, he was the least likely player (0.37%) to make an All-NBA team of all seasons according to our model. Against our model with offensively-focused predictors and the overall lack of any defensively-focused predictors, it is easy to see why his chances were extremely slim and how difficult it is for

our model, in general, to identify prototypical big men as All-NBA players. Today's star big men do much more than rebound, screen, post-up, and defend the paint; many also have developed an outside jumpshot and are expected to facilitate the offense. For these reasons, it makes sense that our model had difficulty identifying so many traditional All-NBA centers in the past decade.

However as the league moves towards positionless basketball, it will be difficult to predict All-NBA players based on position. For example, in the 2019-20 season, Jayson Tatum is listed as a power forward, but he plays more like a small forward. Domantas Sabonis is listed on the roster as a center; although he plays many minutes at the center position, he is listed as a power forward in starting lineups on a game-to-game basis. Anthony Davis and Bam Adebayo are listed as a power forward but both play significant minutes at the center position. In the future, it may be futile to separate All-NBA players by position as these players likely will likely produce similar statistical performances on the court.

We also did not account for predicting within individual seasons. It may have been more relevant to predict 15 All-NBA players per season and compare against the real results instead of predicting All-NBA-caliber players from all the 4 seasons. More specifically, we assigned a value between 0 and 1 representing the likelihood that a player from any of the four seasons from our testing set would make an All-NBA team. If the player received a probability greater than 50%, then we labeled them as an All-NBA player. However, instead of using this as a cutoff, we could have simply taken the players with the 15 highest scores from each of the four seasons respectively and labeled them as All-NBA players or not, based on our model. This would have been a more faithful method and could have improved the overall accuracy of our model. We also explored the alternative option of labeling the top 60 players from the four seasons in the testing dataset as All-NBA players, but it presented mixed results. Although this method correctly predicted significantly more All-NBA players, it also found significantly more false positives. We concluded that while this alternative method could identify more actual All-NBA players than our final method, there seemed to be a large group of players with similar, low probabilities incidentally being labeled as All-NBA players when many of them were not, which meant that this alternative method simply could not discern players accurately.

When predicting the currently unknown 2019-20 All-NBA team, one of the largest factors to consider was the fewer number of regular games played this season, similar to the 2011-12 lockout. The league has already announced that the regular season awards would be based on players' performance up until the league's shutdown in March. Because of this implication, the "number of games played" predictor in our model becomes somewhat detrimental for predicting players from the 2019-20 season, as 67 games was the maximum number of games played in the 2019-20 season and our final model was trained on normal seasons with 82 regular season games. Given that it was one of the predictors in our final model and has a sizable difference in medians compared to All-NBA and non-All-NBA players, there were clear effects on the predictions. For one, Kawhi Leonard is generally regarded among analysts and fans as one of the top-5 players in the league, and he notoriously did not play several regular season games to maintain body health. As a result, he only played 52 games in the 2019-2020 season, which is far lower than the median number (67) of games played across the dataset from the 2009-2019 seasons. Despite having relatively high measures in the other categories, our model predicted that he had only a 54.53% chance of making an All-NBA team, which is much lower than his peers in this season and the 96.91% chance he had in the 2015-16 season. Thus, many of the players from the 2019-20 season with fairly low prediction probabilities could likely have been a result of having played significantly fewer games.

Future Research

One future step we would like to take with our model, as mentioned in the introduction, is predicting All-NBA players at a more granular level and actually predicting the specific All-NBA teams that players would make. We would likely do this using a nested regression technique in which on the outer level, players are predicted to either make any of the All-NBA teams or none of them using logistic regression in the way we have already done. Once a group of predicted All-NBA players is selected, we would run an ordinal regression to predict which of the 3 All-NBA teams each of the players would be placed on.

Another direction we would like to explore is the potential impact that big market teams have on their players being selected to an All-NBA team. Teams like the Los Angeles Lakers and the New York Knicks are

located in large cities in the country, and they garner a majority of media attention even when their teams perform poorly, compared to small-market teams like the Charlotte Hornets and the Cleveland Cavaliers. Could this influence have an effect on voters' decisions when choosing between players? Furthermore, this idea can analogously be applied to individual players. There were many very high-profile names in the 2011-2012 season that were not predicted as All-NBA players, despite a shortened season. Dirk Nowitzki had just become the NBA Finals MVP in the prior season. Kobe Bryant and Andrew Bynum were NBA champions 2 seasons prior. Dwayne Wade, Rajon Rondo, Blake Griffin, Carmelo Anthony and Dwight Howard were all either established All-Stars or highly reputable players in their respective conferences. In the 2012-13 season, Tim Duncan and Tony Parker were selected into the First Team and Second Team All-NBA and led their San Antonio Spurs to the NBA finals, but both were not predicted as All-NBA players in our model. It is possible that voters could be enticed to pick veteran, big-name players, especially those who have won several accolades throughout their careers, over younger, less-established players with statistically better performances.

Additionally, we would like to explore other variable-selection processes beyond forward stepwise regression and select k-best, such as random forests. We are confident that there are other existing variable-selection processes that could more accurately predict All-NBA players than our final model. Other prediction methods could be explored as well. For the purposes of prediction methods, we utilized ridge regression almost exclusively on models with a high number of predictors as a way to minimize the effects of overfitting in highly complex results. Compared to models that didn't utilize ridge regression, models with ridge regression predicted slightly fewer actual All-NBA players but also predicted significantly fewer false positives, especially in higher-predictor models. In future analysis, we would definitely like to continue exploring the impact of ridge regression in terms of prediction methods.

References

"All-NBA Team." *Legit Gambling Sites*, www.legitgamblingsites.com/online-betting/nba/all-nba-team/.

Boger, Tal. "Predicting the 2019 All-NBA Teams with Machine Learning." *Dribble Analytics*, WordPress, 13 Oct. 2019, dribbleanalytics.blog/2019/03/ml-all-nba-predict/.

H., Griffin. "Using Machine Learning to Predict the 2017 NBA All Star Rosters." *Medium*, Medium, 29 Jan. 2017, medium.com/@griffinhoopes/using-machine-learning-to-predict-the-2017-nba-all-star-rosters-bbdd500f7ea5.

Huff, Mathew. "Denver Nuggets: Nikola Jokic Snubbed from All NBA Teams." *Nugg Love*, FanSided, 26 May 2018, nugglove.com/2018/05/26/denver-nuggets-nikola-jokic-snubbed/.

Kent, Austin. "The Financial Impact of All-NBA Teams." *SLAM*, SLAM Media, Inc., 23 May 2019, www.slamonline.com/nba/the-financial-impact-of-all-nba-nods/.

Li, Peter. "NBA MVP Prediction Model." *Medium*, Towards Data Science, 20 Apr. 2019, towardsdatascience.com/nba-mvp-predictor-c700e50e0917.

Rothstein, Ethan. "James Harden's Snub from the All-NBA Teams Is a Historic Blunder." *The Dream Shake*, The Dream Shake, 26 May 2016, www.thedreamshake.com/2016/5/26/11790016/james-harden-all-nba-snub-michael-jordan-lebron-james-klay-thompson.

Link to GitHub repository: https://github.com/skchen999/BSA_Research2020_All-NBA